PhyCovA Tutorial

Tim Blokker

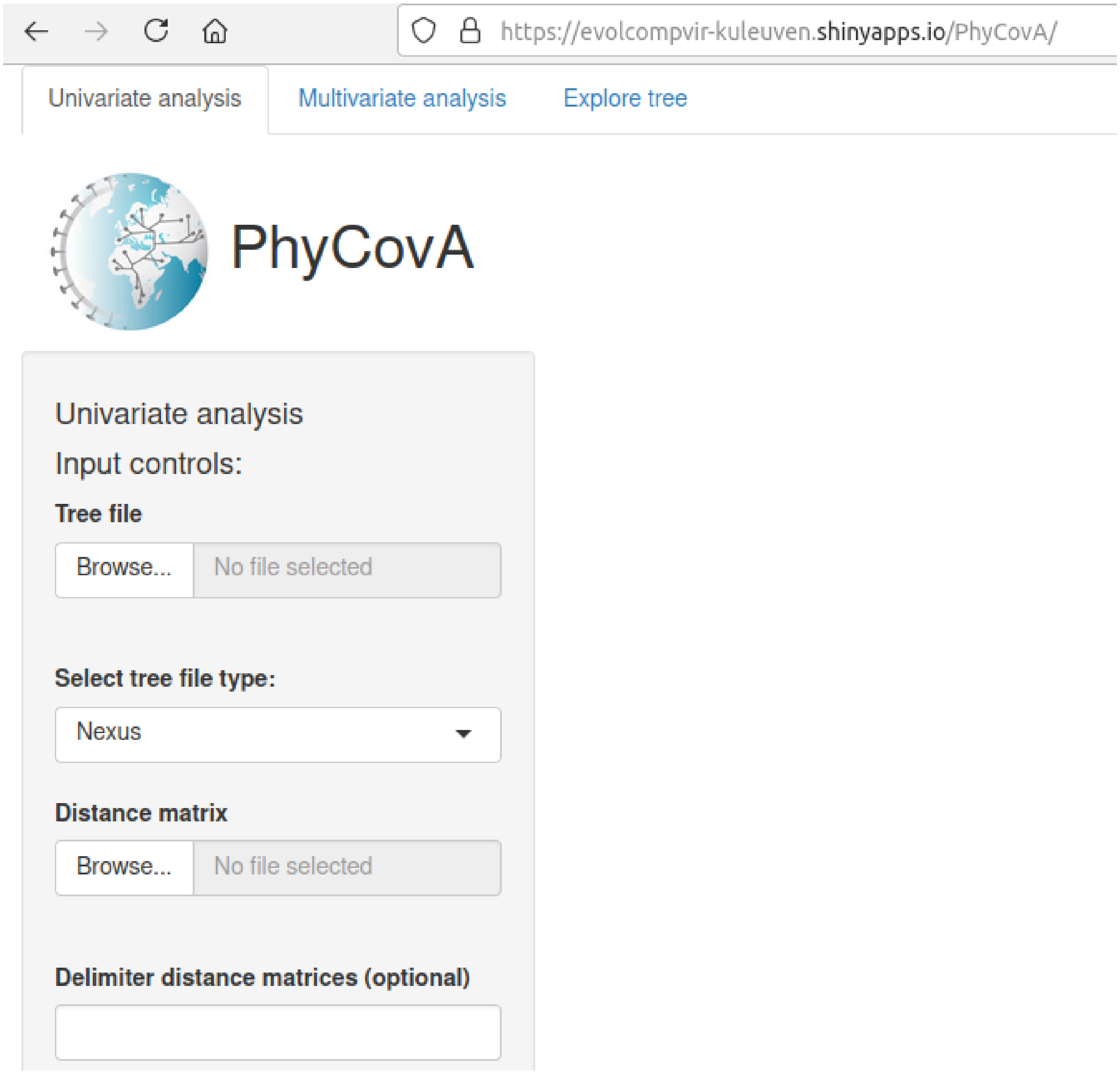
5/4/2021

Contents

1 Getting started

1.1 Browser-based app

There are 2 ways on how to get started with PhyCovA. The first being via the browser and opening the following web-page: https://evolcompvir-kuleuven.shinyapps.io/PhyCovA/



Univariate analysis

Input controls:

Tree file

Select tree file type: Nexus

Distance matrix

Delimiter distance matrices (optional)

1.2 Docker image

The 2nd method is to download ("pull") the docker image available on dockerhub. Installing docker will not be part of this tutorial but good tutorials are available from: https://docs.docker.com/engine/install/

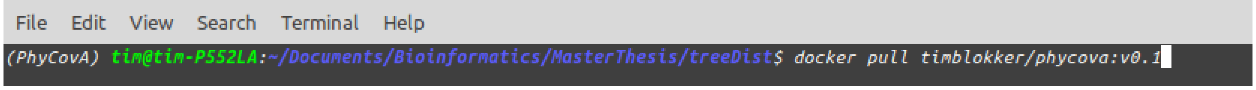
https://docs.docker.com/engine/install/ubuntu/

Once the docker engine is correctly installed we can proceed with pulling the image for PhyCovA. First a few words on docker terminology. The normal docker workflow is:

Dockerfile -> Image -> Container

The dockerfile contains the instructions that the docker engine needs to "build" the image. The docker image is "frozen" and contains everything the docker needs to work. An operating system, R, the required packages, python and the TreeTime package.

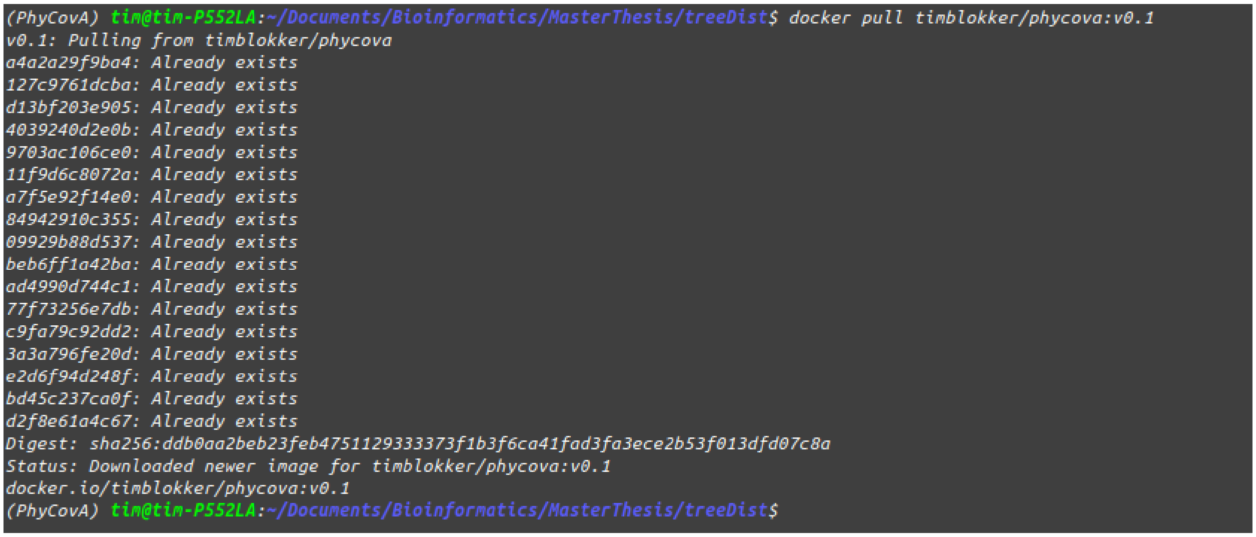
The first step is to pull the image from dockerhub: https://hub.docker.com/repository/docker/timblokker/phycova :



```
(PhyCovA) tim@tim-P552LA:~/Documents/Bioinformatics/MasterThesis/treeDist$ docker pull timblokker/phycova:v0.1
```

Pay attention to the "tag" of the docker image, which is the text behind the ":". For newer versions of PhyCovA, surf to the dockerhub repository and check which versions are available, by specifying the tag/version in the "pull" command you obtain the required version.

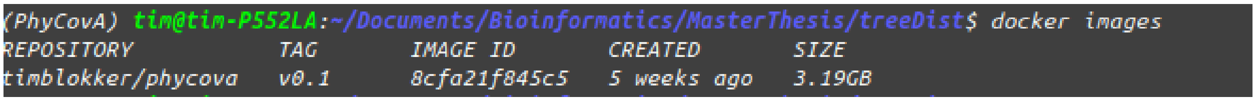
`docker pull timblokker/PhyCovA:v0.1`



```
(PhyCovA) tim@tim-P552LA:~/Documents/Bioinformatics/MasterThesis/treeDist$ docker pull timblokker/phycova:v0.1
v0.1: Pulling from timblokker/phycova
a4a2a29f9ba4: Already exists
127c9761dcba: Already exists
d13bf203e905: Already exists
4039240d2e0b: Already exists
9703ac106ce0: Already exists
11f9d6c8072a: Already exists
a7f5e92f14e0: Already exists
84942910c355: Already exists
09929b88d537: Already exists
beb6ff1a42ba: Already exists
ad4990d744c1: Already exists
77f73256e7db: Already exists
c9fa79c92dd2: Already exists
3a3a796fe20d: Already exists
e2d6f94d248f: Already exists
bd45c237ca0f: Already exists
d2f8e61a4c67: Already exists
Digest: sha256:ddb0aa2beb23feb4751129333373f1b3f6ca41fad3fa3ece2b53f013dfd07c8a
Status: Downloaded newer image for timblokker/phycova:v0.1
docker.io/timblokker/phycova:v0.1
(PhyCovA) tim@tim-P552LA:~/Documents/Bioinformatics/MasterThesis/treeDist$
```

Next the docker image is available on your machine, you can check this by running:

`docker images`

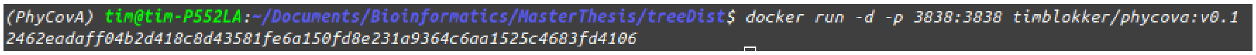


```
(PhyCovA) tim@tim-P552LA:~/Documents/Bioinformatics/MasterThesis/treeDist$ docker images
REPOSITORY           TAG       IMAGE ID       CREATED       SIZE
timblokker/phycova   v0.1      8cfa21f845c5   5 weeks ago   3.19GB
```

And with this we can proceed to starting the container from the image we pulled:

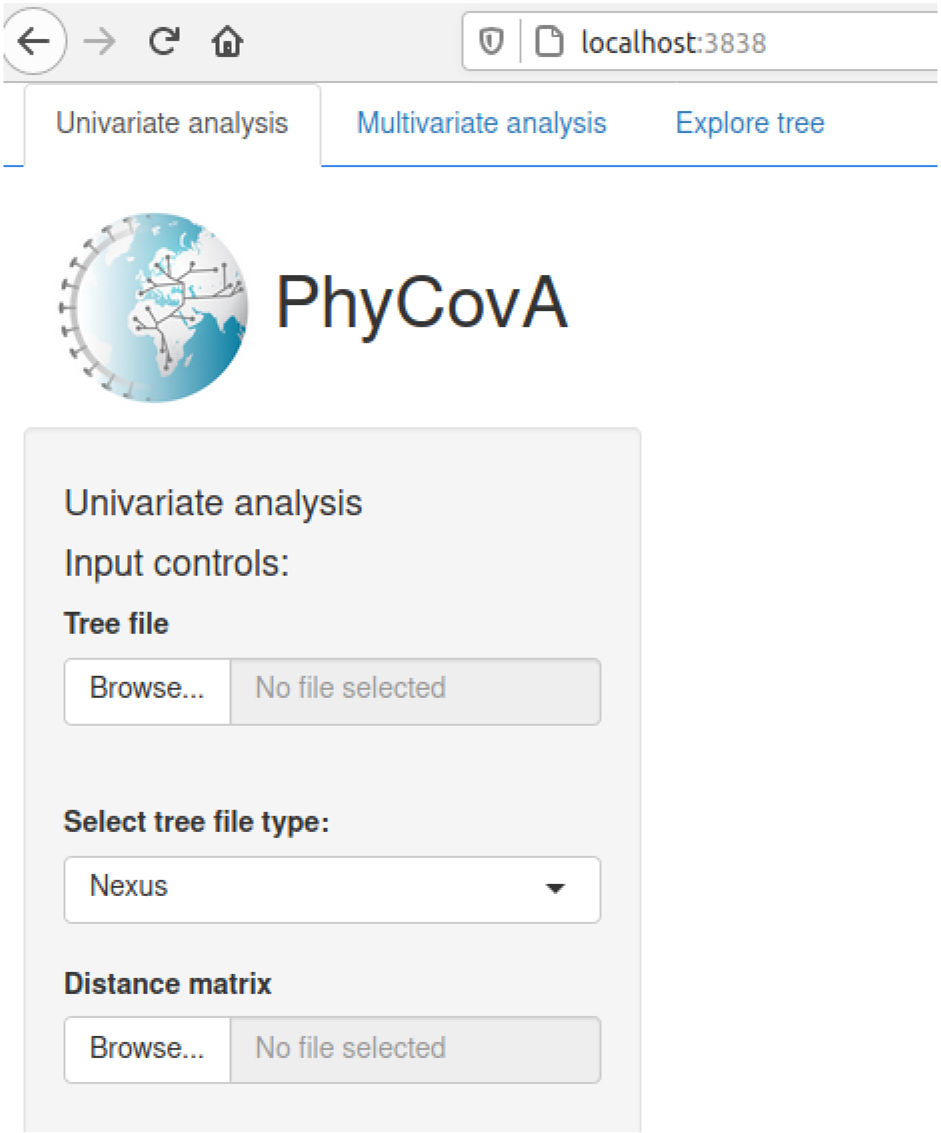
`docker run -d -p 3838:3838 timblokker/phycova:v0.1`

It is possible to run the docker without the "-d"*etached* flag to still see the output which when running the shiny app from within RStudio is outputted on the RStudio console.The "-p"*ort* flag specifies the port that will be allocated to PhyCovA. This needs to be 3838 and is hard-coded in the Dockerfile.



```
(PhyCovA) tim@tim-P552LA:~/Documents/Bioinformatics/MasterThesis/treeDist$ docker run -d -p 3838:3838 timblokker/phycova:v0.1
2462eadaff04b2d418c8d43581fe6a150fd8e231a9364c6aa1525c4683fd4106
```

Finally, the browser can be pointed to http://localhost:3838/



Univariate analysis

Input controls:

Tree file

Browse... No file selected

Select tree file type:

Nexus

Distance matrix

Browse... No file selected

2 PhyCovA

PhyCovA can correlate transition counts or rates obtained from either trees already annotated with ancestral state information or reconstruct and annotate the ancestral states before correlating the variables.

In this tutorial data originally published in the publication below will be examplary analyzed.

Dudas, G., Carvalho, L. M., Bedford, T., Tatem, A. J., Baele, G., Faria, N. R., Park, D. J., Ladner, J. T., Arias, A., Asogun, D., Bielejec, F., Caddy, S. L., Cotten, M., D'Ambrozio, J., Dellicour, S., Di Caro, A., Diclaro, J. W., Duraffour, S., Elmore, M. J., … Rambaut, A. (2017). Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature, 544(7650), 309–315. https://doi.org/10.1038/nature22040

2.1 Already annotated trees

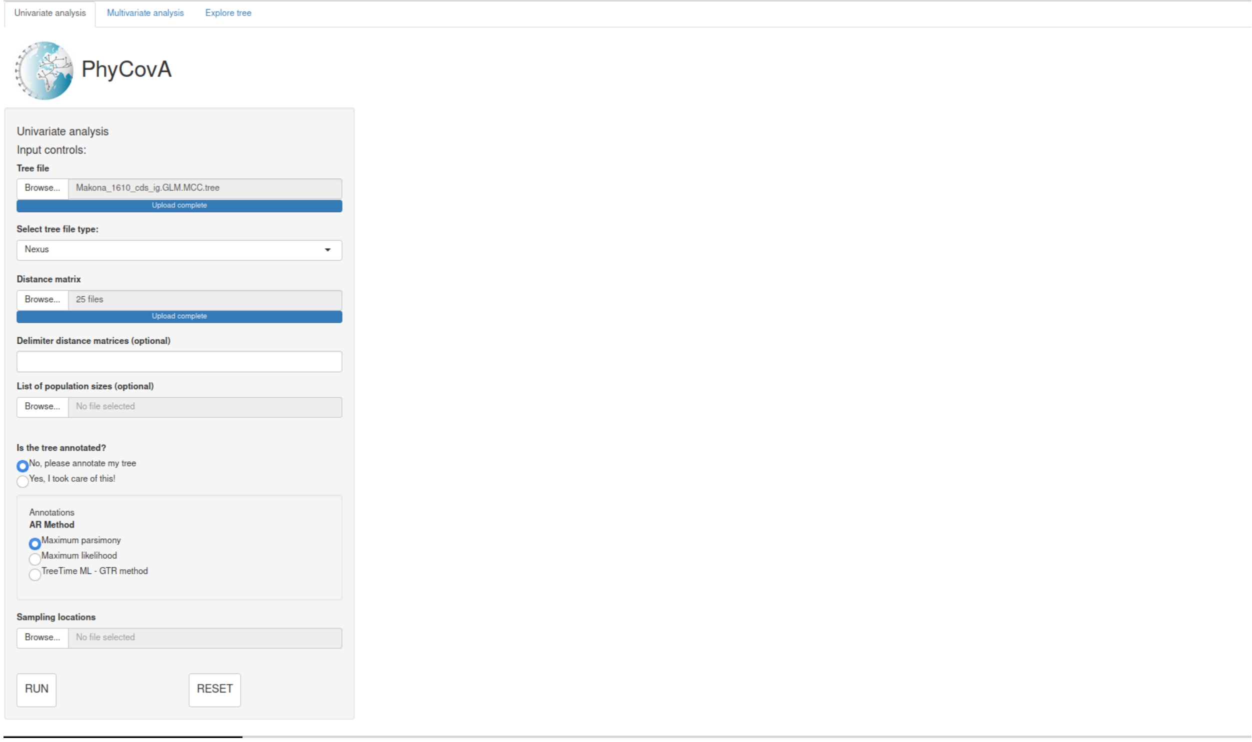
2.1.1 Univariate tab

The already annotated mode requires only 2 files, the phylogeny, here we will use an MCC tree for the ebola data set and at least 1 distance matrix. The data can be obtained from github: https://github.com/TimBlokker/PhyCovA/tree/master/input/ebola . The annotations in the tree must **exactly** match the column names in the pairwise distance matrices. And the phylogeny needs to be rooted.

At first we upload the 2 documents using the file-upload panels, by clicking **"Browse. . . "**, this will open a file explorer and should be intuitive to use. **Note:** For the distance matrix multiple files can be selected in the file explorer, while the "Tree File" file upload only allows 1 file.

In the screenshot below the 2 file types were successfully uploaded. Additional limitations on the files are:

- File names can not start with a digit
- Hyphens ("-") are not recommended, rather use underscores ("_")

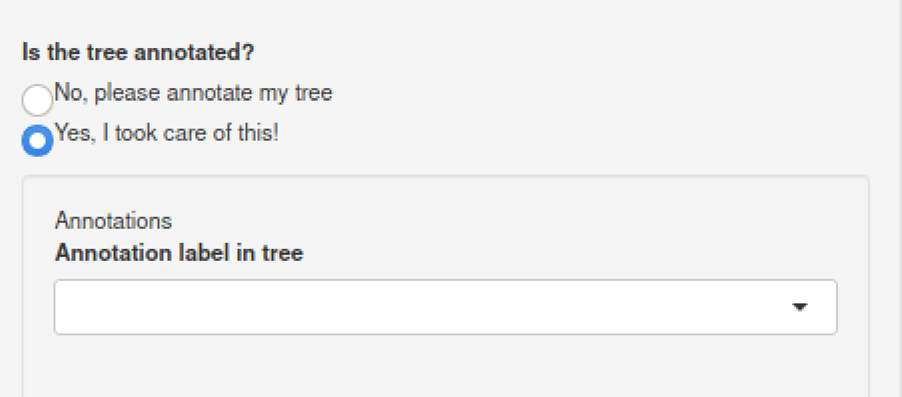


Univariate analysis
Input controls:
Tree file
Makona_1610_cds_ig.GLM.MCC.tree
Select tree file type:
Nexus
Distance matrix
25 files
Delimiter distance matrices (optional)
List of population sizes (optional)
No file selected
Is the tree annotated?
● No, please annotate my tree
○ Yes, I took care of this!

Annotations
AR Method
● Maximum parsimony
○ Maximum likelihood
○ TreeTime ML - GTR method

Sampling locations
No file selected

When selecting "Yes, I took care of this!" then PhyCovA will start searching for annotations in the tree file that match the column namesof the distance matrix.

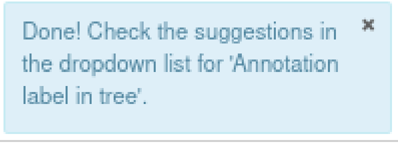


Is the tree annotated?

○ No, please annotate my tree
● Yes, I took care of this!

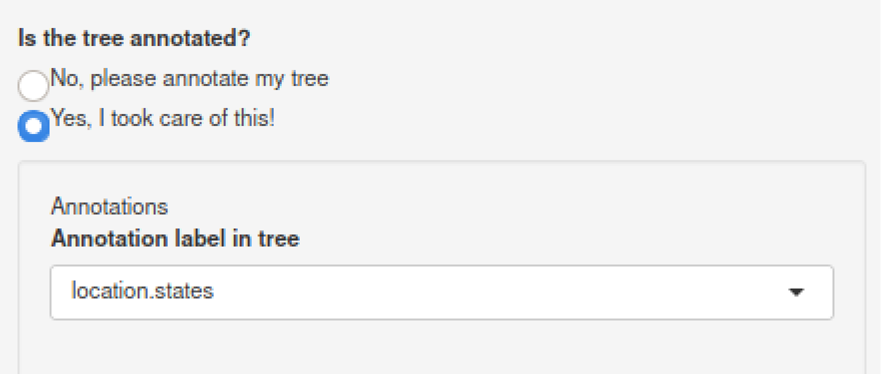
Annotations
Annotation label in tree

Once an annotation was found that matches the column names in the distance matrix, PhyCovA will inform the user.



Done! Check the suggestions in the dropdown list for 'Annotation label in tree'. ×

Simultaneously the found column will be displayed. In case more than one column matches the column names in the matrix the user can select the appropriate one.



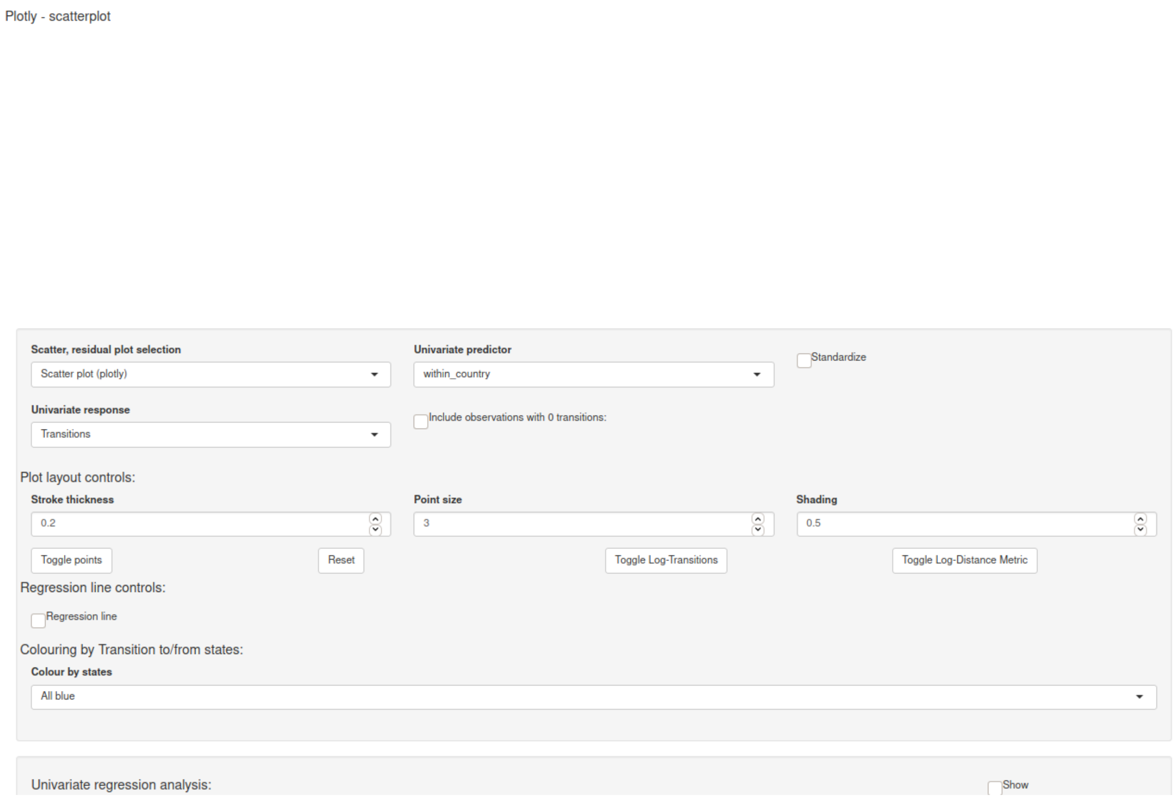
Is the tree annotated?

○ No, please annotate my tree
● Yes, I took care of this!

Annotations
Annotation label in tree

location.states

With this PhyCovA has been given all required information and the phylogenetic covariance analysis can start by clicking "RUN". Immediately after clicking the input fields will appear but depending on the size of the tree and the number of distance matrices, the analysis might take a few seconds to 30 seconds.



Plotly - scatterplot

Scatter, residual plot selection: Scatter plot (plotly)

Univariate predictor: within_country

□ Standardize

Univariate response: Transitions

□ Include observations with 0 transitions:

Plot layout controls:

Stroke thickness: 0.2

Point size: 3

Shading: 0.5

Regression line controls:

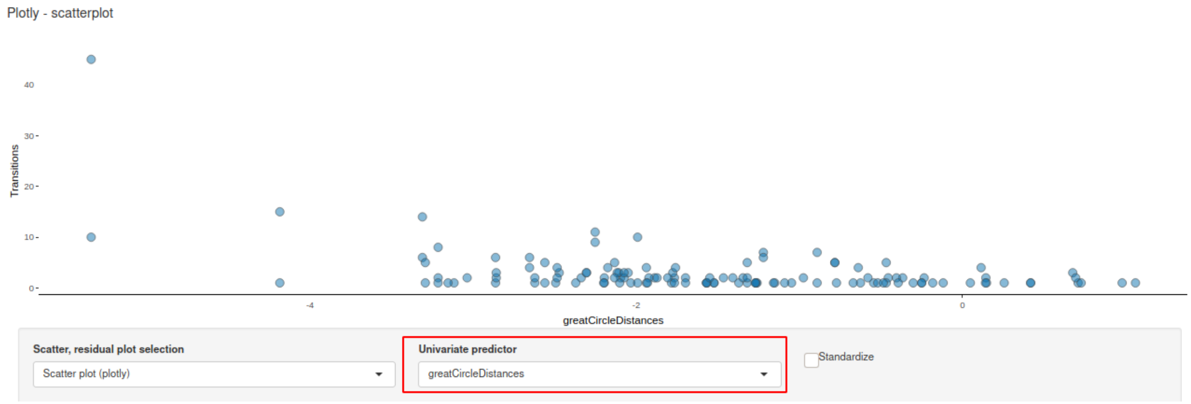
□ Regression line

Colouring by Transition to/from states:

Colour by states: All blue

Univariate regression analysis: □ Show

After the covariate analysis has been finished the scatterplot is visible. At first we see a scatterplot of a binary variable which is not very informative, so we change the predictor to “greatCircleDistances”.



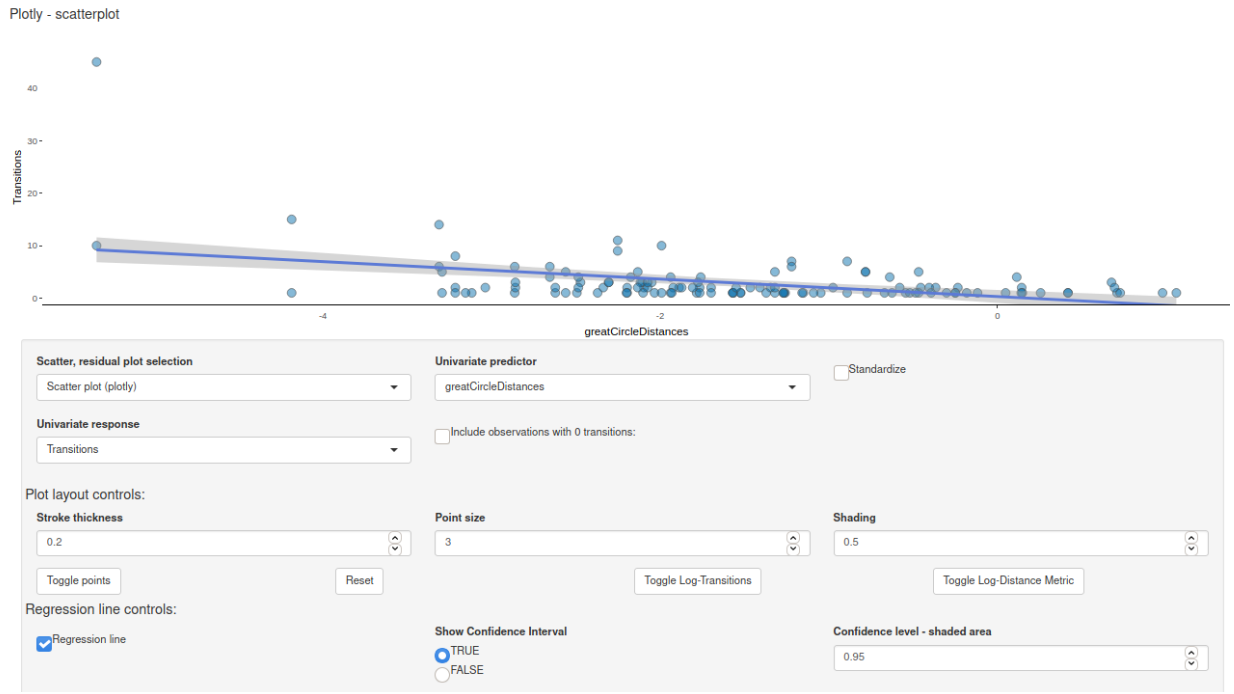
Plotly - scatterplot

Scatter, residual plot selection: Scatter plot (plotly)

Univariate predictor: greatCircleDistances

□ Standardize

There are many options to improve the layout of the plots. We will leave the user to explore but some important features we would like to point out. At first the regression line can be drawn in the plot:



Scatter, residual plot selection: Scatter plot (plotly)

Univariate predictor: greatCircleDistances

□ Standardize

Univariate response: Transitions

□ Include observations with 0 transitions:

Plot layout controls:

Stroke thickness: 0.2

Point size: 3

Shading: 0.5

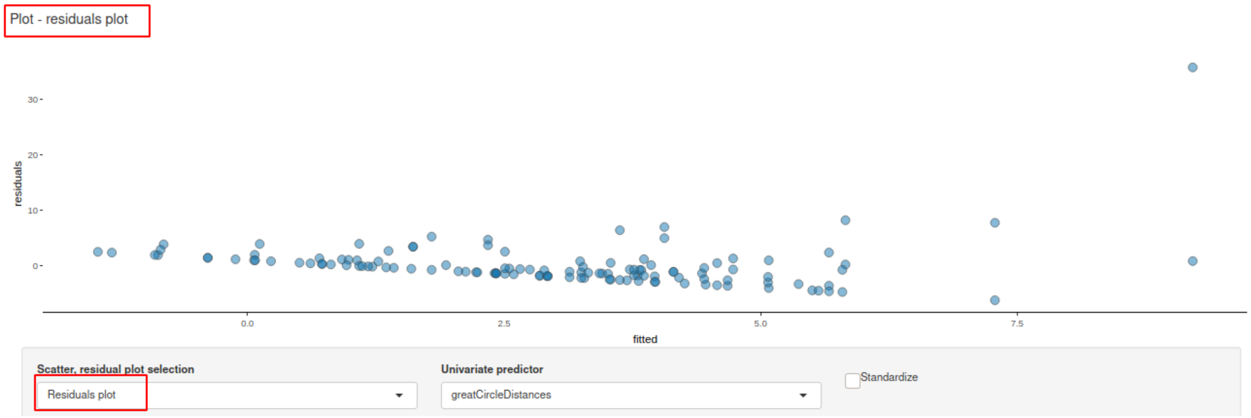
Regression line controls:

☑ Regression line

Show Confidence Interval
● TRUE
○ FALSE

Confidence level - shaded area: 0.95

Then we can look at the residuals of linear regression:

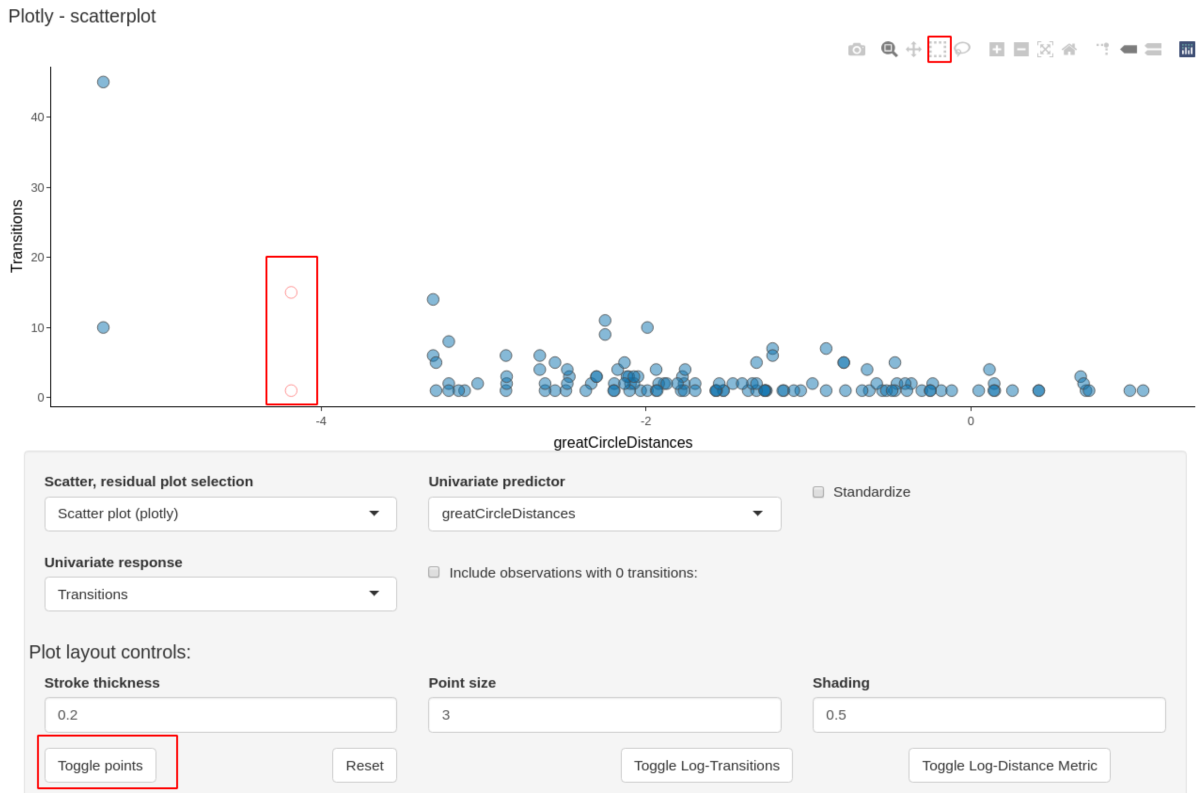


Scatter, residual plot selection: Residuals plot

Univariate predictor: greatCircleDistances

□ Standardize

Observations can be excluded to see which effect particular points have on the regression. This can be done by brushing points with the “Box Select” tool and then clicking “Toggle points”. The points are then removed from the data set used for the regression analysis but are still plotted in another colour:



Scatter, residual plot selection: Scatter plot (plotly)

Univariate predictor: greatCircleDistances

□ Standardize

Univariate response: Transitions

□ Include observations with 0 transitions:

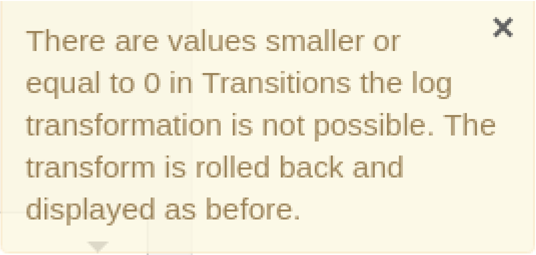
Plot layout controls:

Stroke thickness: 0.2

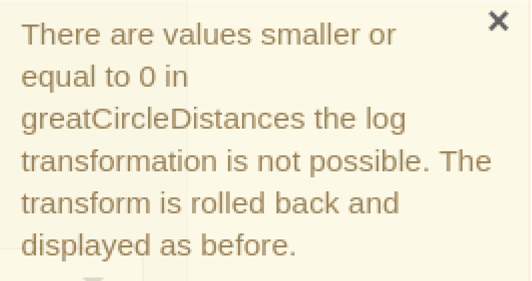
Point size: 3

Shading: 0.5

The response and the predictor variable can be log-transformed but there are limitations like the log of 0 and negative values is not defined and so e.g. when including zero-transition events the log transformation of the response variable “Transition counts” will not be possible anymore. Same counts for standardising of the predictors, centering variables with mean 0 will unavoidably lead to negative values in the data set.

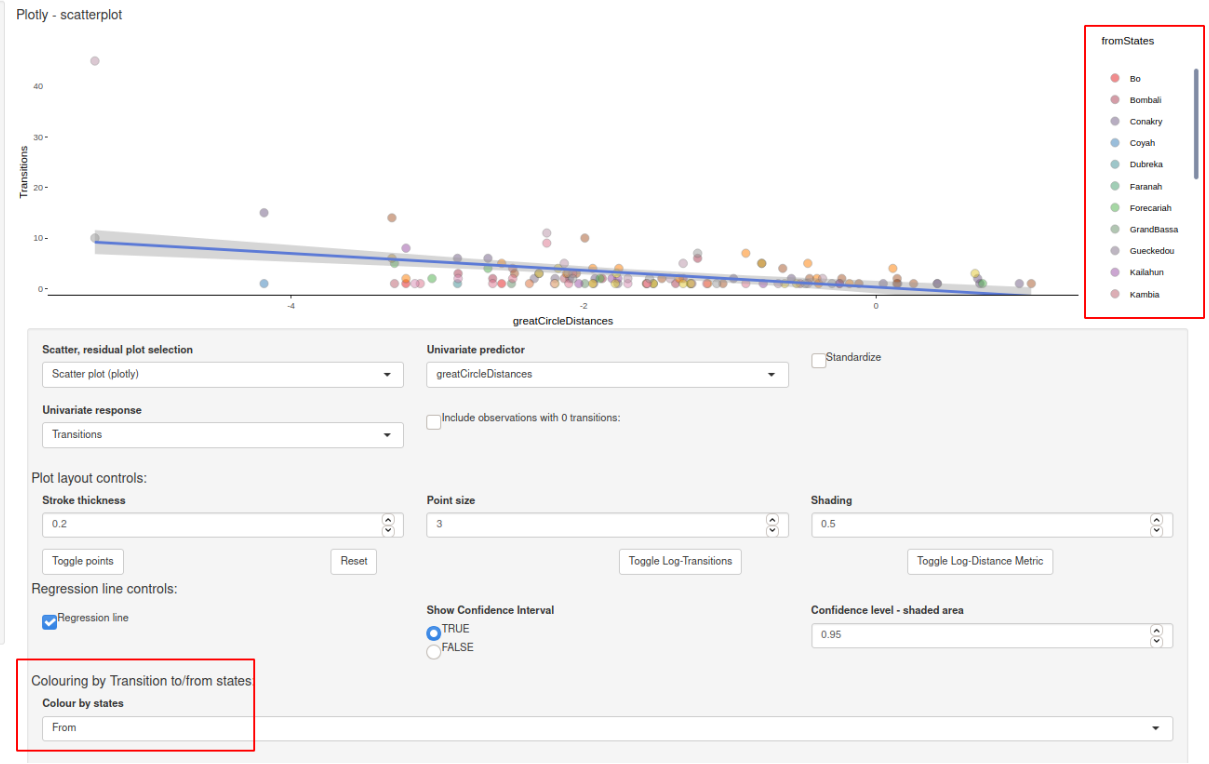


There are values smaller or equal to 0 in Transitions the log transformation is not possible. The transform is rolled back and displayed as before.



There are values smaller or equal to 0 in greatCircleDistances the log transformation is not possible. The transform is rolled back and displayed as before.

The predictors can be standardized, and zero-transition events can be included. Zero-transition events are transition events that have not been observed in the tree. Then the transition events can be colourized grouped by the origin location (“from”) or also destination (“to”).



Scatter, residual plot selection
Scatter plot (plotly)

Univariate predictor
greatCircleDistances

☐ Standardize

Univariate response
Transitions

☐ Include observations with 0 transitions:

Plot layout controls:

Stroke thickness
0.2

Point size
3

Shading
0.5

Toggle points | Reset | Toggle Log-Transitions | Toggle Log-Distance Metric

Regression line controls:

☑ Regression line

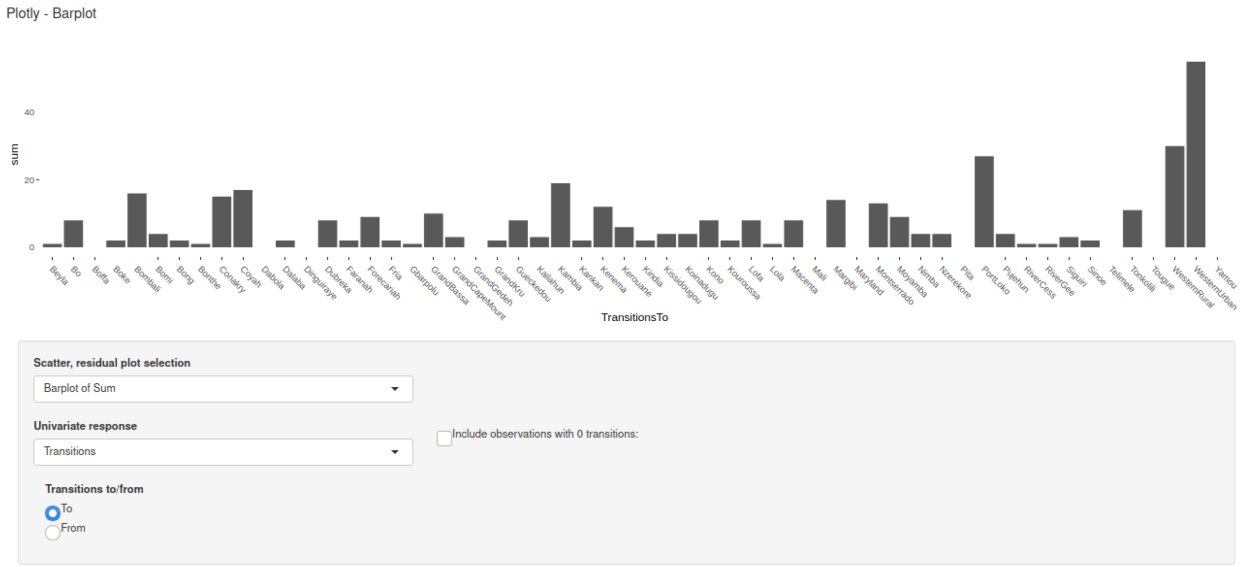
Show Confidence Interval
● TRUE
○ FALSE

Confidence level - shaded area
0.95

Colouring by Transition to/from states

Colour by states
From

Lastly, PhyCovA also offers the option to summarize the transition counts by state in a bar diagram. Also here we can group observation by origin country or destination country. The option to include 0-transitions has no effect here (because we add 0. . .).



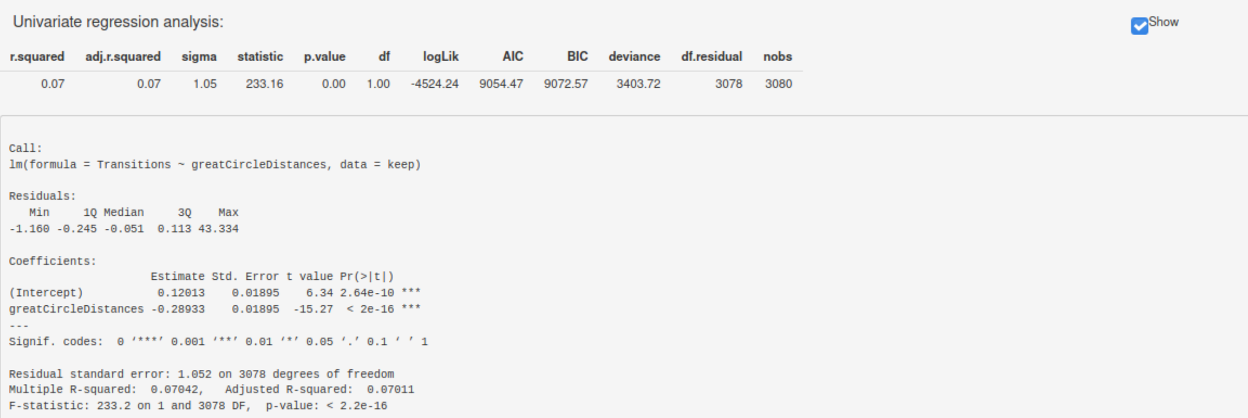
Scatter, residual plot selection
Barplot of Sum

Univariate response
Transitions

☐ Include observations with 0 transitions:

Transitions to/from
● To
○ From

For this univariate case the standardization and the inclusion of the 0-transitions have impact on the regression output. And this is shown and updated upon changing the input and is shown as depicted below:



Univariate regression analysis:

☑ Show

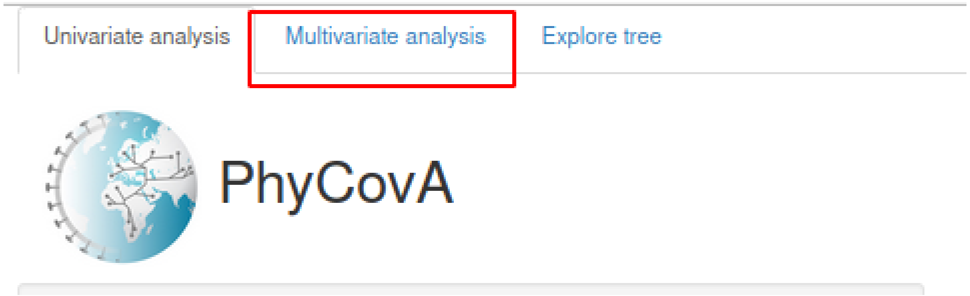
r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.07	0.07	1.05	233.16	0.00	1.00	-4524.24	9054.47	9072.57	3403.72	3078	3080

```
Call:
lm(formula = Transitions ~ greatCircleDistances, data = keep)

Residuals:
   Min     1Q Median     3Q    Max 
-1.160 -0.245 -0.051  0.113 43.334 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)           0.12013    0.01895    6.34 2.64e-10 ***
greatCircleDistances -0.28933    0.01895  -15.27  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.052 on 3078 degrees of freedom
Multiple R-squared:  0.07042,	Adjusted R-squared:  0.07011 
F-statistic: 233.2 on 1 and 3078 DF,  p-value: < 2.2e-16
```

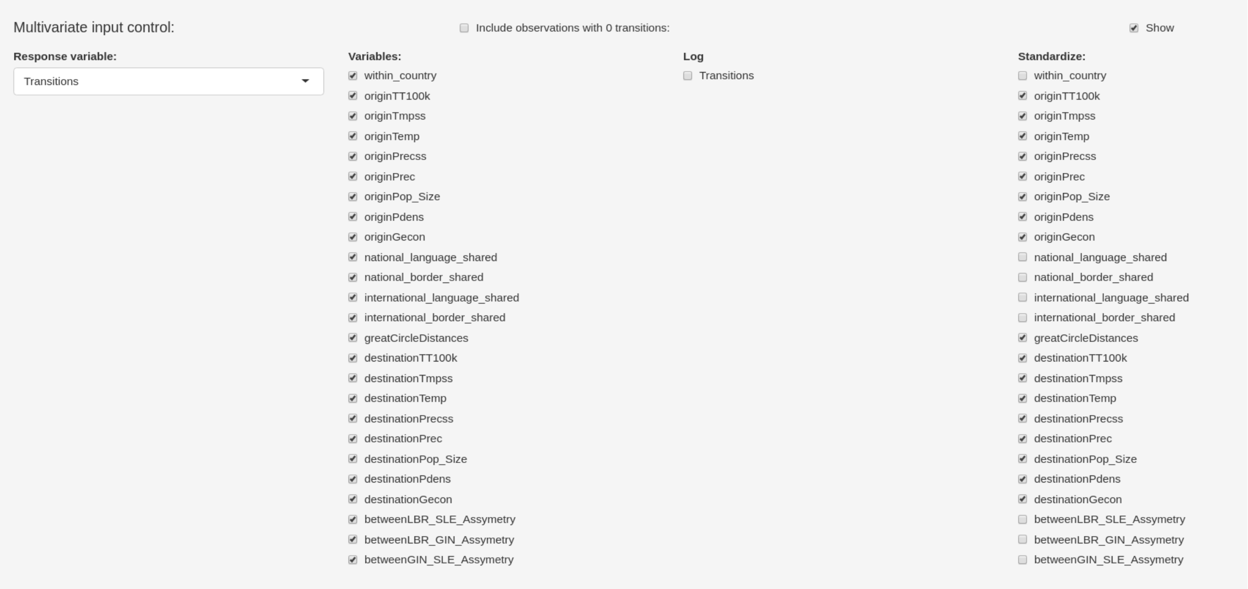
From here we can see that the geographic distance (greatCircleDistances) explains approximately 7% of the observed transition events (when not including zero-transitions, not standardizing).

2.1.2 Multivariate tab

Switching to the multivariate tab can be done on top of the screen:



On the multivariate tab, at first the variables to include in the analysis can be selected. We will start with keeping them all selected. And we will standardize all continuous predictors. We will keep only actually observed transition events in the data set and will not log transform the response, for this data set only the “Transition count” variable is available as response (drop-down menu in the top left corner, here for the TreeTime reconstructions also the transition rate will be available).



Multivariate input control:

☐ Include observations with 0 transitions:

☑ Show

Response variable:

Transitions

Variables:

☑ within_country
☑ originTT100k
☑ originTmpss
☑ originTemp
☑ originPrecss
☑ originPrec
☑ originPop_Size
☑ originPdens
☑ originGecon
☑ national_language_shared
☑ national_border_shared
☑ international_language_shared
☑ international_border_shared
☑ greatCircleDistances
☑ destinationTT100k
☑ destinationTmpss
☑ destinationTemp
☑ destinationPrecss
☑ destinationPrec
☑ destinationPop_Size
☑ destinationPdens
☑ destinationGecon
☑ betweenLBR_SLE_Assymetry
☑ betweenLBR_GIN_Assymetry
☑ betweenGIN_SLE_Assymetry

Log

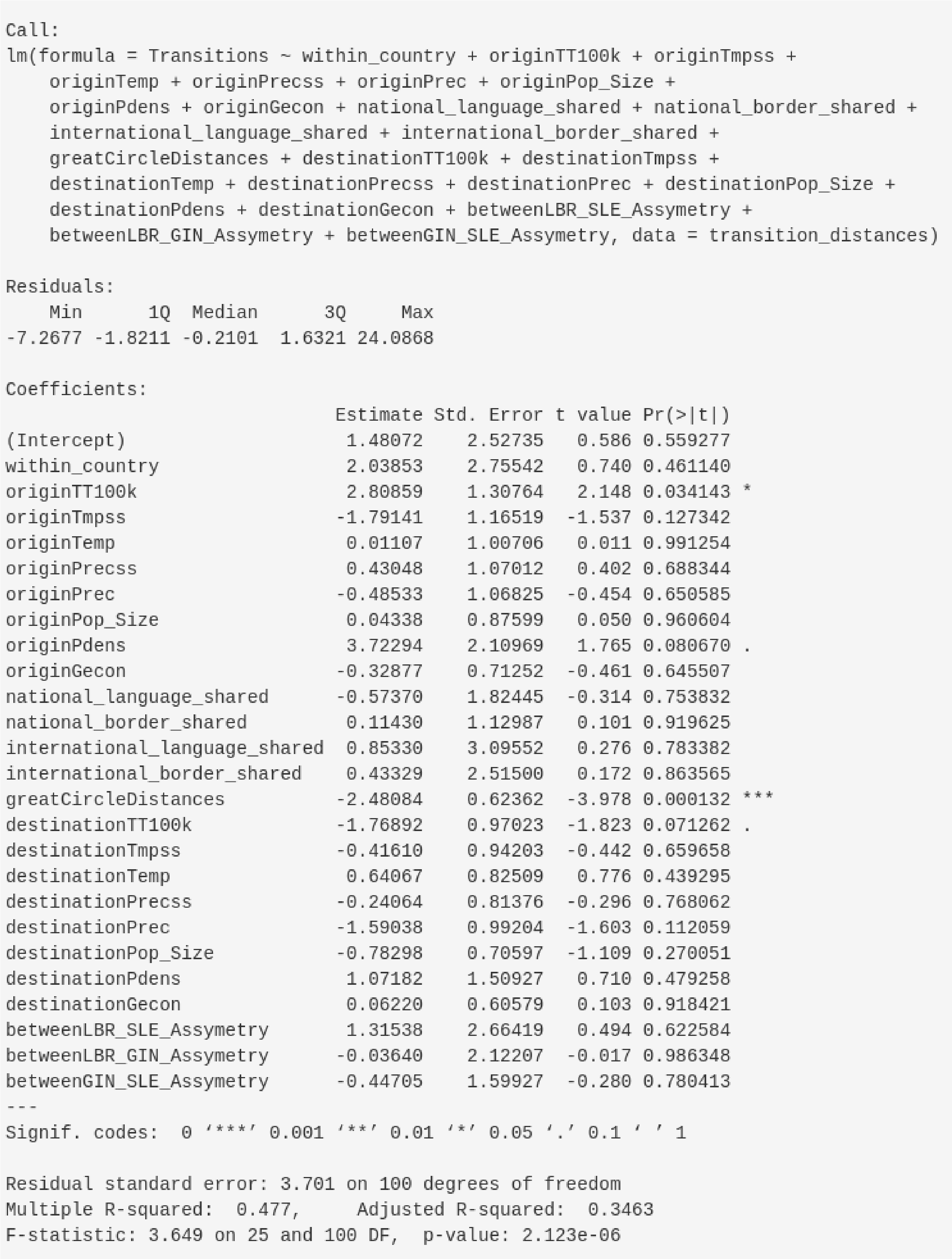
☐ Transitions

Standardize:

☐ within_country
☑ originTT100k
☑ originTmpss
☑ originTemp
☑ originPrecss
☑ originPrec
☑ originPop_Size
☑ originPdens
☑ originGecon
☐ national_language_shared
☐ national_border_shared
☐ international_language_shared
☐ international_border_shared
☑ greatCircleDistances
☑ destinationTT100k
☑ destinationTmpss
☑ destinationTemp
☑ destinationPrecss
☑ destinationPrec
☑ destinationPop_Size
☑ destinationPdens
☑ destinationGecon
☐ betweenLBR_SLE_Assymetry
☐ betweenLBR_GIN_Assymetry
☐ betweenGIN_SLE_Assymetry

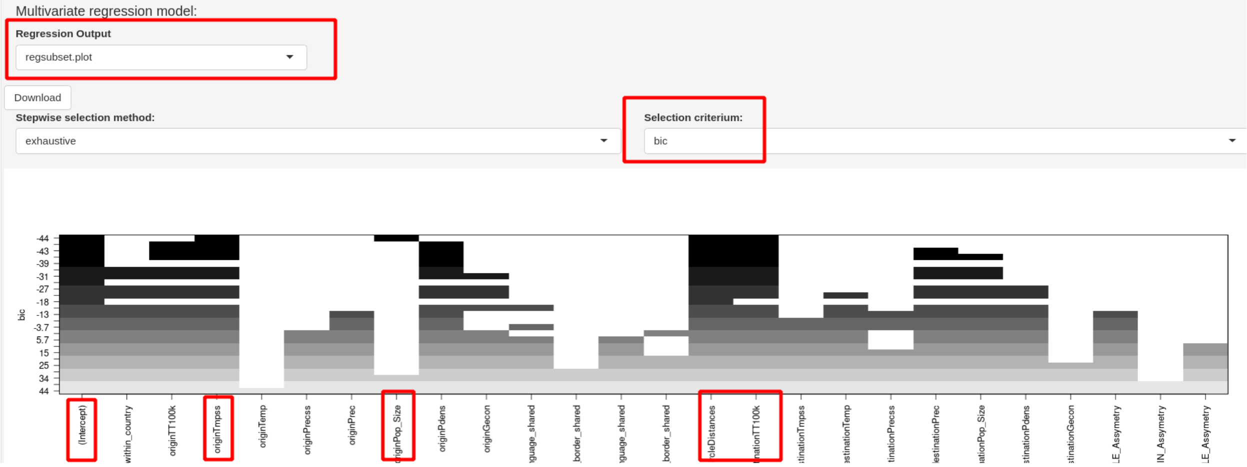
We can hide this panel now and move to the regression output. It is always a good idea to hide panels that are not needed. Anything not displayed will not be computed and will speed up the analysis. This effect is especially big when having many variables, including zero transitions and having the plotting panel open. **Always hide the plotting panel before changing variable selection**.

The summary output shows the type III sum of squares:



```
Call:
lm(formula = Transitions ~ within_country + originTT100k + originTmpss +
    originTemp + originPrecss + originPrec + originPop_Size +
    originPdens + originGecon + national_language_shared + national_border_shared +
    international_language_shared + international_border_shared +
    greatCircleDistances + destinationTT100k + destinationTmpss +
    destinationTemp + destinationPrecss + destinationPrec + destinationPop_Size +
    destinationPdens + destinationGecon + betweenLBR_SLE_Assymetry +
    betweenLBR_GIN_Assymetry + betweenGIN_SLE_Assymetry, data = transition_distances)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2677 -1.8211 -0.2101  1.6321 24.0868

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    1.48072    2.52735   0.586 0.559277
within_country                 2.03853    2.75542   0.740 0.461140
originTT100k                   2.80859    1.30764   2.148 0.034143 *
originTmpss                   -1.79141    1.16519  -1.537 0.127342
originTemp                     0.01107    1.00706   0.011 0.991254
originPrecss                   0.43048    1.07012   0.402 0.688344
originPrec                    -0.48533    1.06825  -0.454 0.650585
originPop_Size                 0.04338    0.87599   0.050 0.960604
originPdens                    3.72294    2.10969   1.765 0.080670 .
originGecon                   -0.32877    0.71252  -0.461 0.645507
national_language_shared      -0.57370    1.82445  -0.314 0.753832
national_border_shared         0.11430    1.12987   0.101 0.919625
international_language_shared  0.85330    3.09552   0.276 0.783382
international_border_shared    0.43329    2.51500   0.172 0.863565
greatCircleDistances          -2.48084    0.62362  -3.978 0.000132 ***
destinationTT100k             -1.76892    0.97023  -1.823 0.071262 .
destinationTmpss              -0.41610    0.94203  -0.442 0.659658
destinationTemp                0.64067    0.82509   0.776 0.439295
destinationPrecss             -0.24064    0.81376  -0.296 0.768062
destinationPrec               -1.59038    0.99204  -1.603 0.112059
destinationPop_Size           -0.78298    0.70597  -1.109 0.270051
destinationPdens               1.07182    1.50927   0.710 0.479258
destinationGecon               0.06220    0.60579   0.103 0.918421
betweenLBR_SLE_Assymetry       1.31538    2.66419   0.494 0.622584
betweenLBR_GIN_Assymetry      -0.03640    2.12207  -0.017 0.986348
betweenGIN_SLE_Assymetry      -0.44705    1.59927  -0.280 0.780413
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.701 on 100 degrees of freedom
Multiple R-squared:  0.477,     Adjusted R-squared:  0.3463
F-statistic: 3.649 on 25 and 100 DF,  p-value: 2.123e-06
```

Here, we can note an adjusted R^2 of 35%, meaning that this model explains 35% of the variation in the transition counts after adjusting for the number of predictors included in the model. We can now take a look at the variable selection via the “leaps” package and the “regsubsets” function, which makes use of different criteria to select the best model according to e.g. “BIC” out of all possible subsets of variables or via the backward/forward selection.



Multivariate regression model:

Regression Output

regsubset.plot

Download

Stepwise selection method:

exhaustive

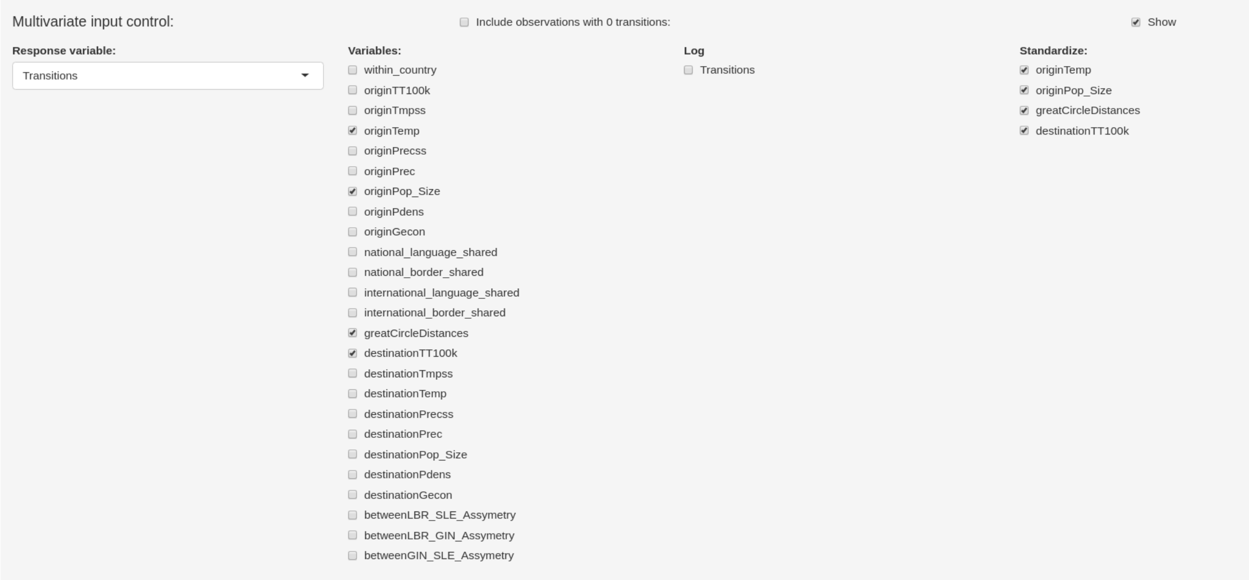
Selection criterium:

bic

This plot shows 1 model for each number of variables. Each model is the best model for this amount of variables according to the BIC criterion. These pre-selected models are then sorted by their BIC value with the lowest BIC value on top. So the best model overall is listed in the top row. Blackened squares mean that the variable has been selected. We can note that the best model is a 4-variable model consisting of: the temperature at the origin location, the population size at the origin location, the geographic distance and the travel distance to the next metropolitan area from the destination location.

Note: For this data set using the backward selection algorithm leads to a 5-variable model. This output will also be found in the stepwise AIC model when using the “BIC” criterion.

We can now go ahead and build a linear model with only the seleced variabeles:



Multivariate input control:

☐ Include observations with 0 transitions:

☑ Show

Response variable:

Transitions

Variables:

☐ within_country
☐ originTT100k
☐ originTmpss
☑ originTemp
☐ originPrecss
☐ originPrec
☑ originPop_Size
☐ originPdens
☐ originGecon
☐ national_language_shared
☐ national_border_shared
☐ international_language_shared
☐ international_border_shared
☑ greatCircleDistances
☑ destinationTT100k
☐ destinationTmpss
☐ destinationTemp
☐ destinationPrecss
☐ destinationPrec
☐ destinationPop_Size
☐ destinationPdens
☐ destinationGecon
☐ betweenLBR_SLE_Assymetry
☐ betweenLBR_GIN_Assymetry
☐ betweenGIN_SLE_Assymetry

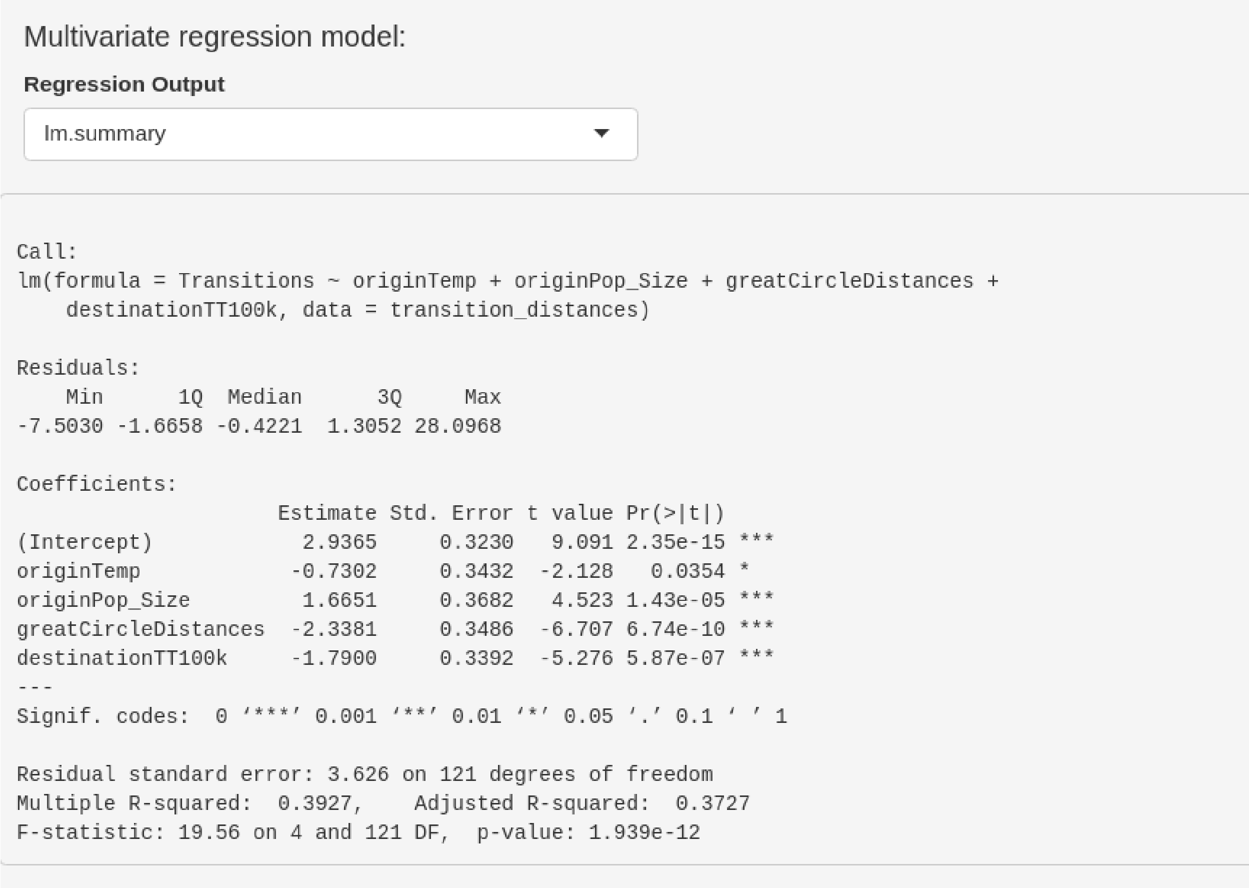
Log

☐ Transitions

Standardize:

☑ originTemp
☑ originPop_Size
☑ greatCircleDistances
☑ destinationTT100k

The adjusted R^2 for this data set did not change when reducing the model to the selected variables only.



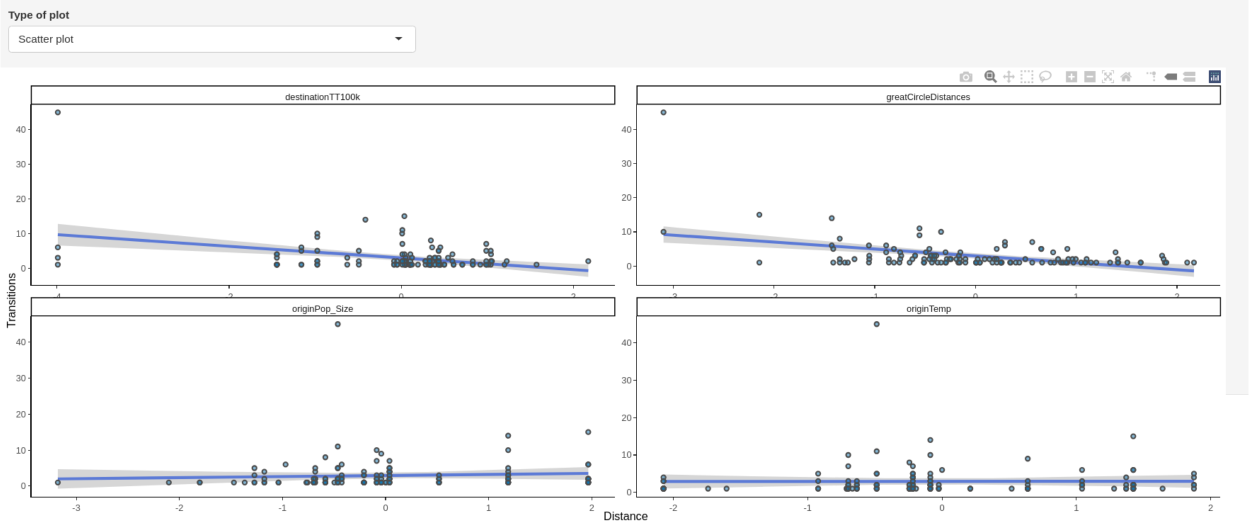
Multivariate regression model:

Regression Output

lm.summary

```
Call:
lm(formula = Transitions ~ originTemp + originPop_Size + greatCircleDistances + 
    destinationTT100k, data = transition_distances)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.5030 -1.6658 -0.4221  1.3052 28.0968 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)            2.9365     0.3230   9.091 2.35e-15 ***
originTemp            -0.7302     0.3432  -2.128   0.0354 *  
originPop_Size         1.6651     0.3682   4.523 1.43e-05 ***
greatCircleDistances  -2.3381     0.3486  -6.707 6.74e-10 ***
destinationTT100k     -1.7900     0.3392  -5.276 5.87e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.626 on 121 degrees of freedom
Multiple R-squared:  0.3927,	Adjusted R-squared:  0.3727 
F-statistic: 19.56 on 4 and 121 DF,  p-value: 1.939e-12
```

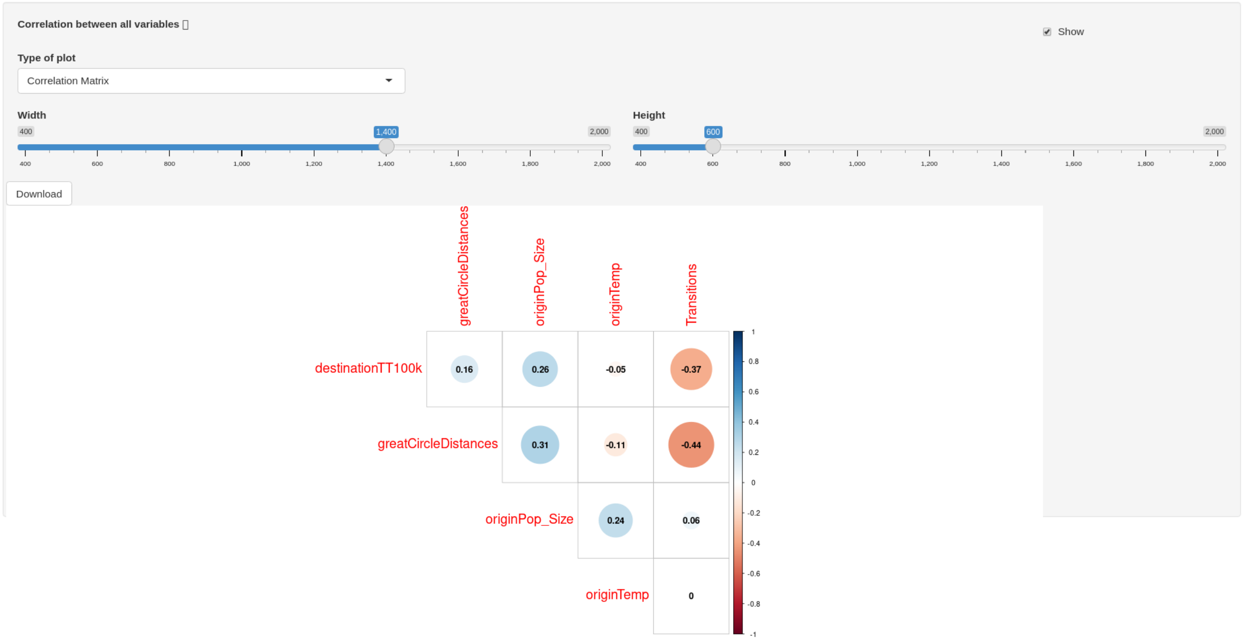
To conclude the Multivariate tab we can move on to the plotting panel:



Type of plot

Scatter plot

In addition correlation plots can be visualized:



Correlation between all variables

☑ Show

Type of plot: Correlation Matrix

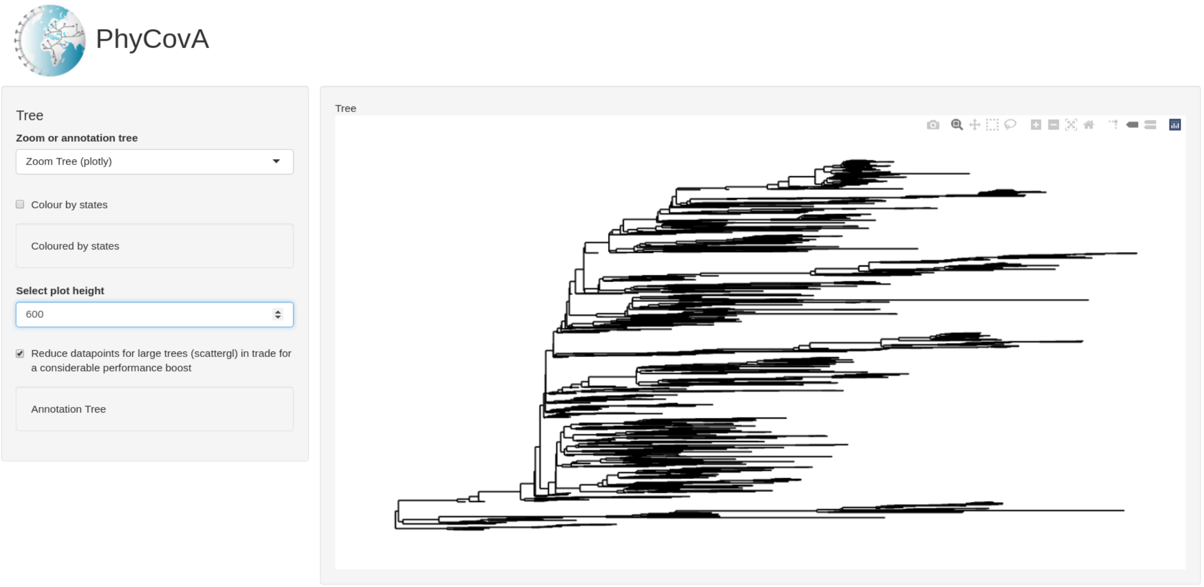
Width: 1,400

Height: 600

Download

2.1.3 Tree exploration

The phylogeny can also be visualized in PhyCovA:



PhyCovA

Tree

Zoom or annotation tree: Zoom Tree (plotly)

☐ Colour by states

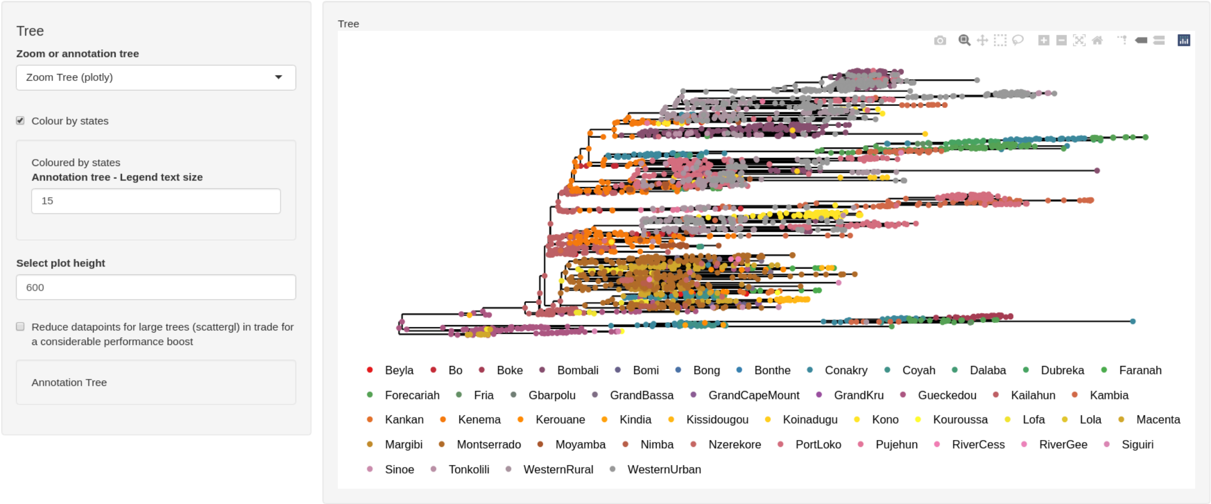
Coloured by states

Select plot height: 600

☑ Reduce datapoints for large trees (scattergl) in trade for a considerable performance boost

Annotation Tree

This tree is generated by "Plotly" and is zoomable, hence the name in the drop-down to the right. The user has different tools available such as selection, panning and zooming tools in the top right corner of the plotting area. The tree height can also be changed and very large plots can be plotted and one can scroll through the plot. One can also colour the tree and show more of the availale information by ticking the box for "Coloured by states" and unticking the box for "Reduce datapoints for large trees (scattergl) in trade for a condirate performance boost".



Tree

Zoom or annotation tree

Zoom Tree (plotly)

☑ Colour by states

Coloured by states

Annotation tree - Legend text size

15

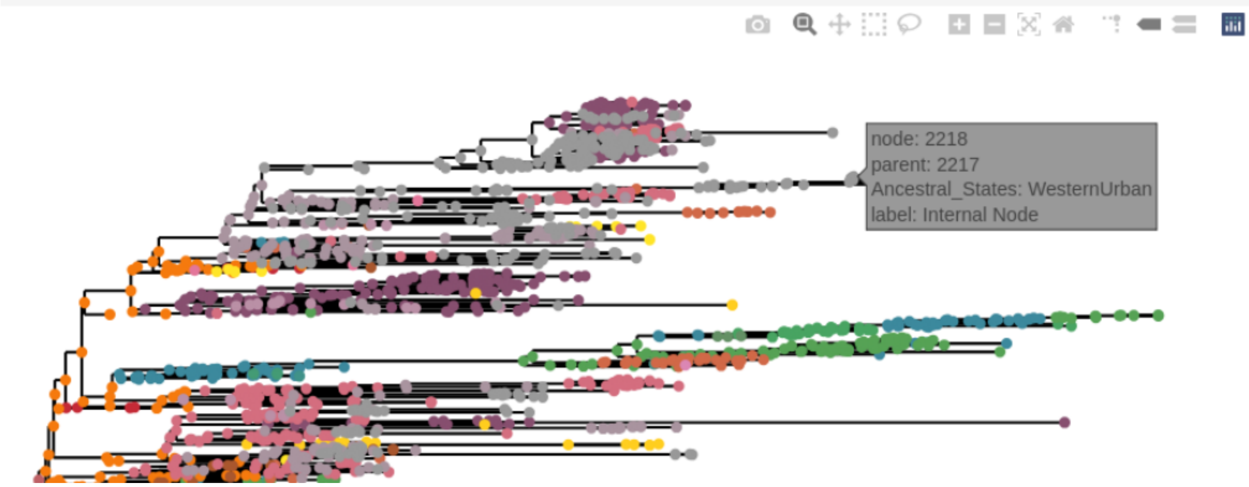
Select plot height

600

☐ Reduce datapoints for large trees (scattergl) in trade for a considerable performance boost

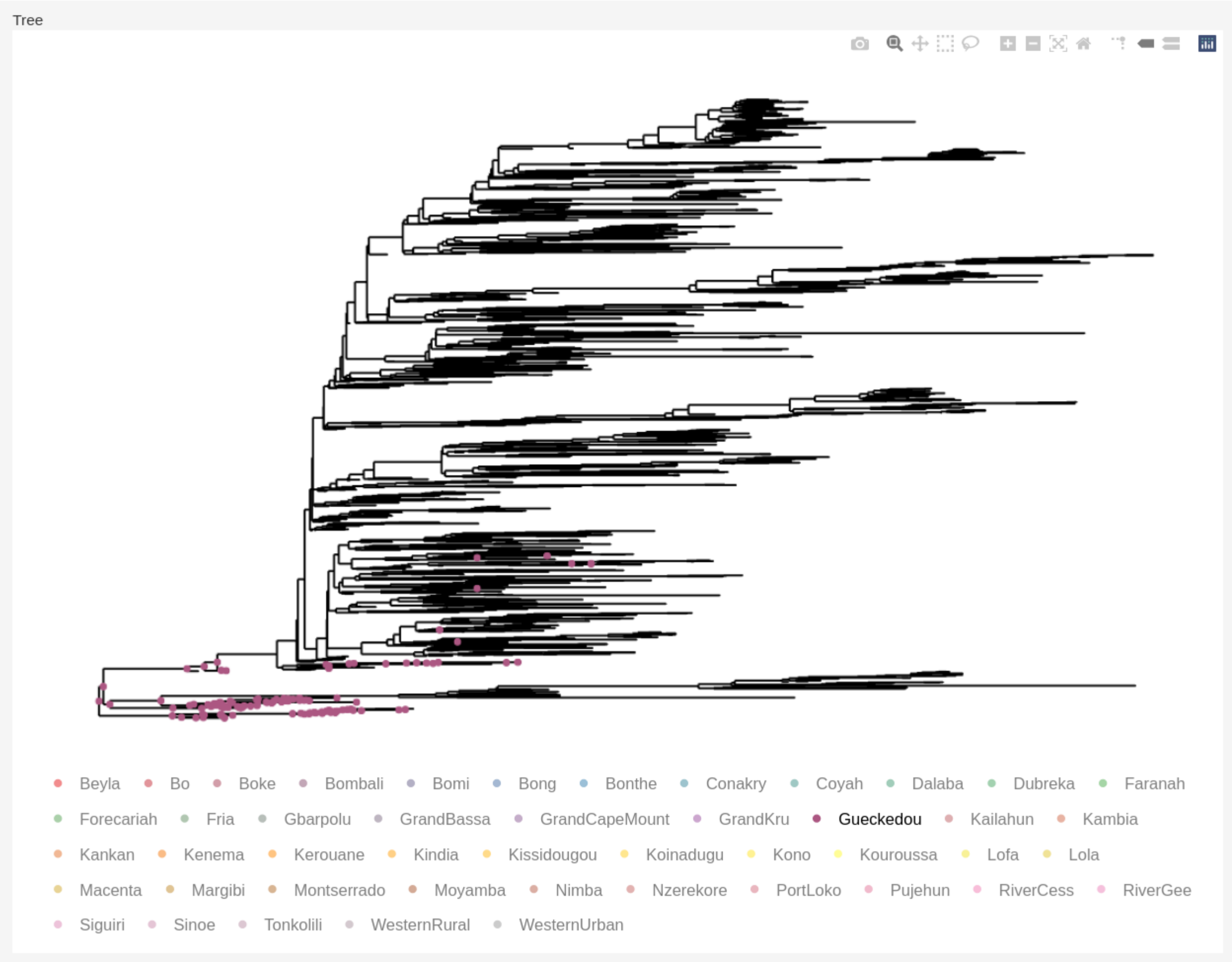
Annotation Tree

Unticking the latter box is only affecting information in the tooltips.

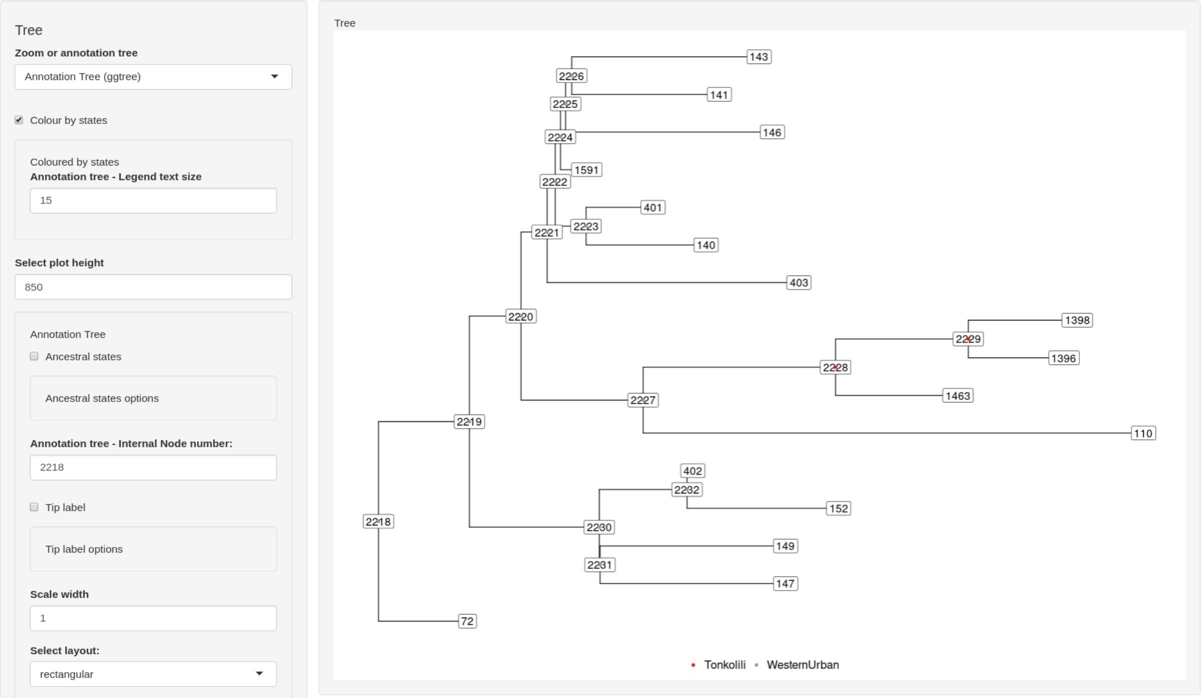


node: 2218
parent: 2217
Ancestral_States: WesternUrban
label: Internal Node

One can only highlight only certain states, to see where this state is located in the tree. This is here done by double-clikcing on the state in the legend. We can see that Gueckedou was the likely starting point of the Ebola pandemic as it is mainly annotated as state at the root and nodes in proximity of the root of the phylogeny.



Next to a zoomable tree there is also a non-zoomable tree, which can be annotated in more detail. The 2 trees are intended to work together. One possible use case is to identify the region of a tree of interest. Get the node number of the tree via the tooltip, e.g. in this example above this was node 2218. We can look at only this node and its offspring in the annotation tree:



Tree

Zoom or annotation tree

Annotation Tree (ggtree) ▾

☑ Colour by states

Coloured by states

Annotation tree - Legend text size

15

Select plot height

850

Annotation Tree

☐ Ancestral states

Ancestral states options

Annotation tree - Internal Node number:

2218

☐ Tip label

Tip label options

Scale width

1

Select layout:

rectangular ▾

And from this subtree many annotations can be shown on the tree, like the ancestral state, the node shapes etc.

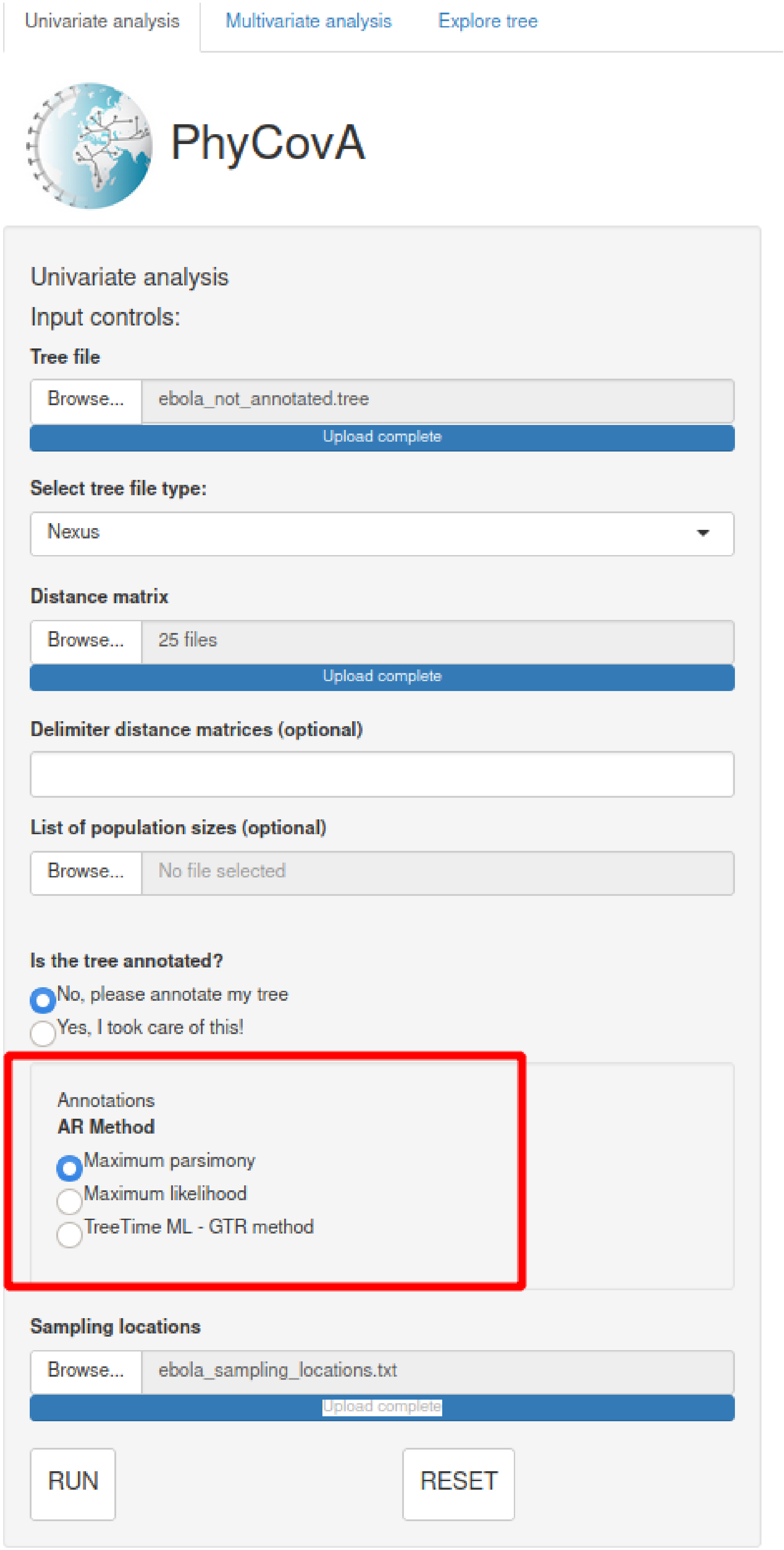
2.2 Reset

In case one has finished analyzing the data set and wants to analyze something else, one opens the “Univariate” tab and clicks “RESET”. Now all input is cleared and new files can be uploaded.

2.3 Non-Annotated tree

For non-annotated trees, also a phylogeny and at least 1 distance matrix is required. In addition to these 2 files PhyCovA required a sorted list of tip states and the states need be **exactly** the same as the column names in the distance matrices.

One can now select 1 of the 3 possible reconstruction methods and carry out the analysis as described above for the annotated case.



Univariate analysis

Input controls:

Tree file

Browse... ebola_not_annotated.tree

Upload complete

Select tree file type:

Nexus

Distance matrix

Browse... 25 files

Upload complete

Delimiter distance matrices (optional)

List of population sizes (optional)

Browse... No file selected

Is the tree annotated?

● No, please annotate my tree

○ Yes, I took care of this!

Annotations

AR Method

● Maximum parsimony

○ Maximum likelihood

○ TreeTime ML - GTR method

Sampling locations

Browse... ebola_sampling_locations.txt

Upload complete

This concludes the tutorial for PhyCovA.