mGPS Interface Tutorial

Eran Elhaik Yali Zhang

January 14, 2022

Contents

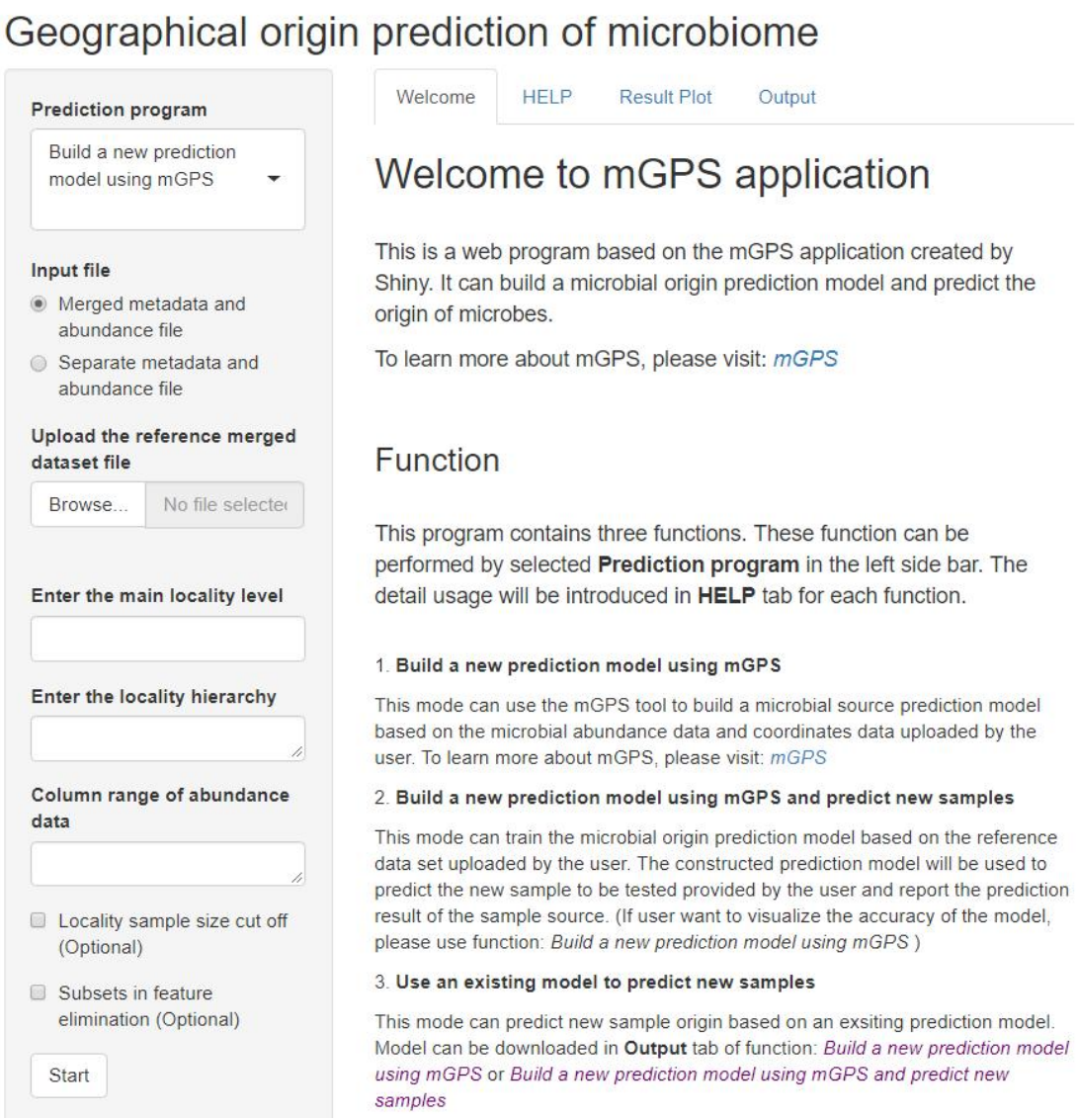
1 Introduction

mGPS interface is a web program based on the mGPS application created by Shiny. It can build a microbial origin prediction model and predict the origin of microbes. To learn more about mGPS, please visit mGPS. For the specific usage of the mGPS interface, please check the Readme file on the mGPS interface Github.

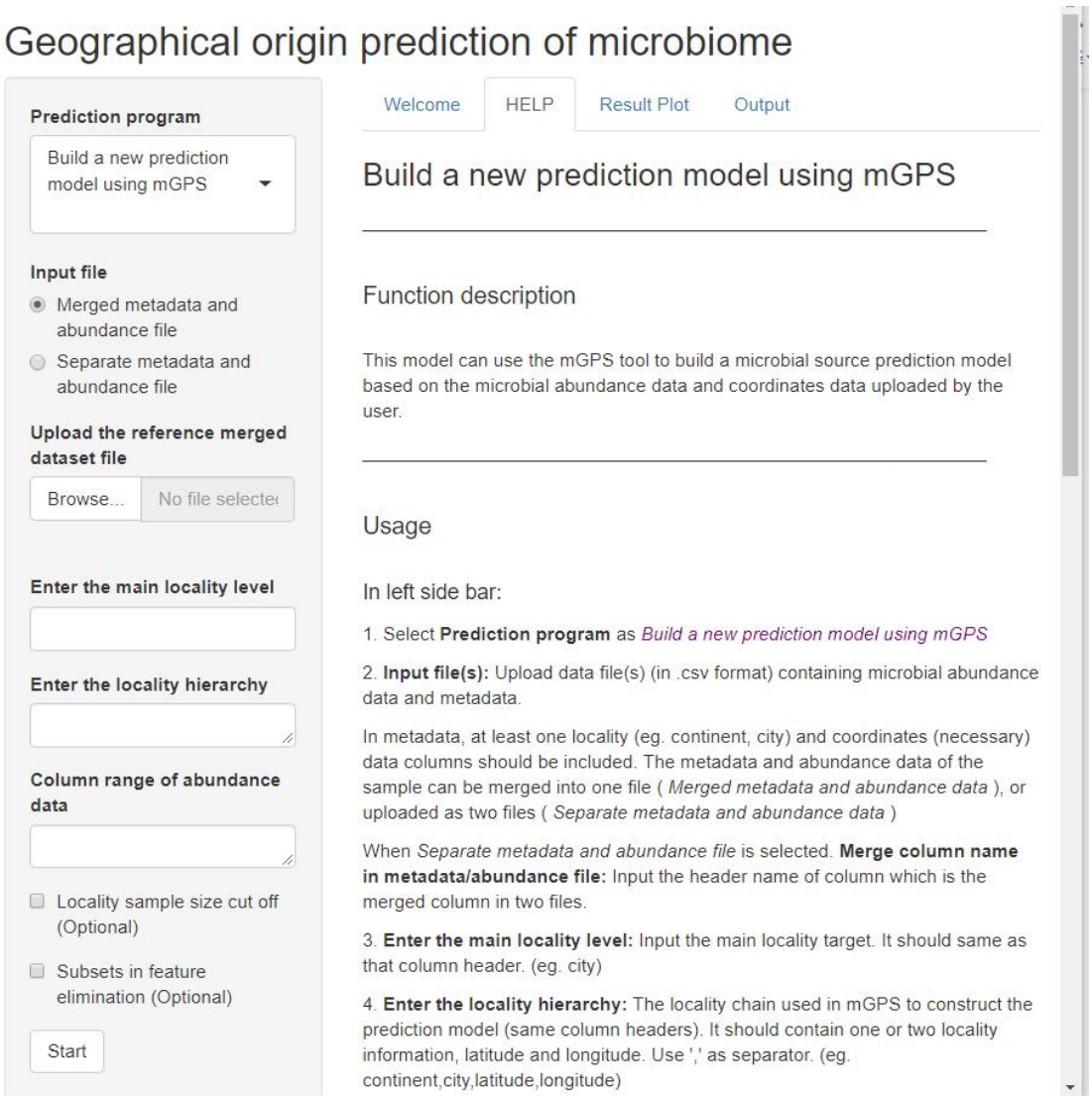
2 Function and Output

2.1 Welcome page

A brief introduction to the functions of the mGPS interface and how to start using it



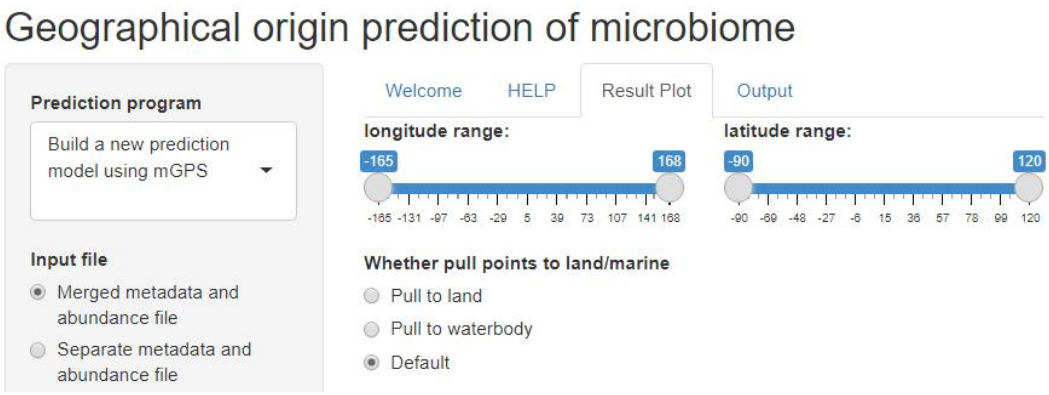
2.2 Function 1: Build a new prediction model using mGPS



2.2.1 Function 1 introduction

In this mode, you can use the mGPS tool to build a microbial source prediction model based on the microbial abundance data and coordinates data uploaded by the user.

2.2.2 Output plots



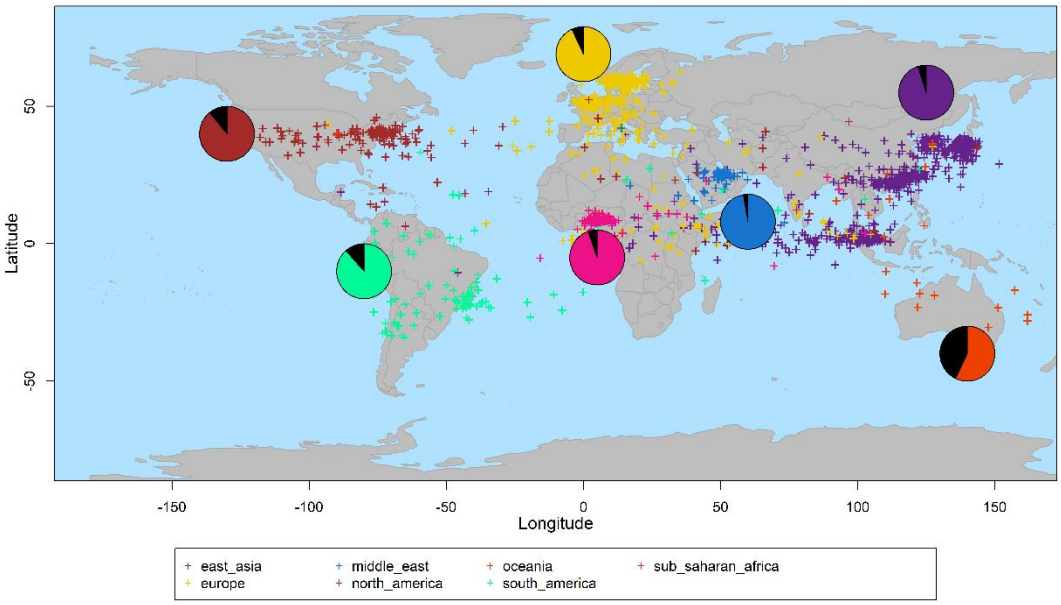
The "Result Plot" of this mode will show the accuracy of the prediction model trained by the mGPS tool and based on the reference microbial database you uploaded.
If you want to zoom in on a specific area of the map, you need to adjust both the longitude and latitude ranges. "Pull to land" refers to pushing the predicted point to the nearest land, and "Pull to waterbody" refers to pushing the predicted point to the nearest water body.

When you select "Pull to land" and "Pull to waterbody", the sample prediction source coordinates will be changed, the content of the output file in the tab “Output” will also be changed accordingly. Remember, the data in output file is the same as the figure.

How the algorithm detects the accuracy of the model: The original database will be divided into 5 folds, and mGPS will use 4 of these folds to train the model, and the resulting model will be used to predict the microbial source of the remaining fold. Iteratively obtain the prediction result of the original database and compare it with the actual location of the microorganism.

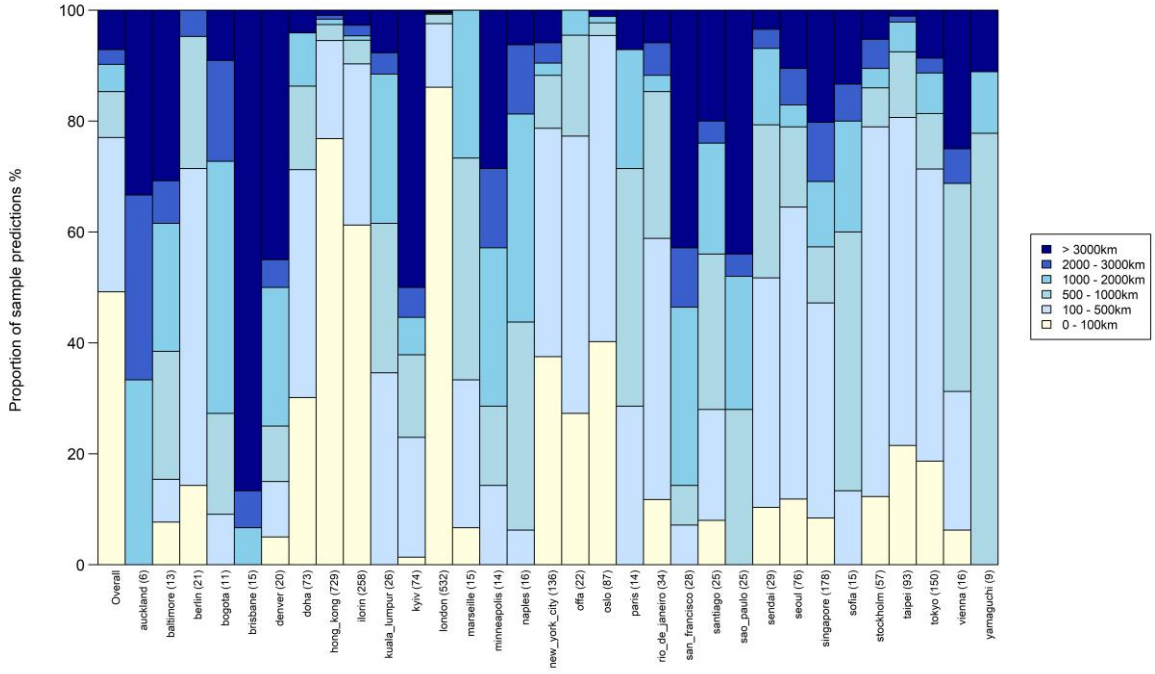
Example output plot:

a. World map: samples' prediction origins are plotted on the world map. By adjusting the range of longitude and latitude, you can select the plotting area of the figure on the map. In addition, you can choose whether to pull the predicted point into the continent or the ocean. The predicted origin of original samples will be mapped onto the world map. The pie charts with colors divided by geographic area represent the proportion of samples in the area whose origin region is predicted to be the same continent as the true location. You can download and save the png file of this map through the button ‘DownloadMap’ below the figure.



DownloadMap

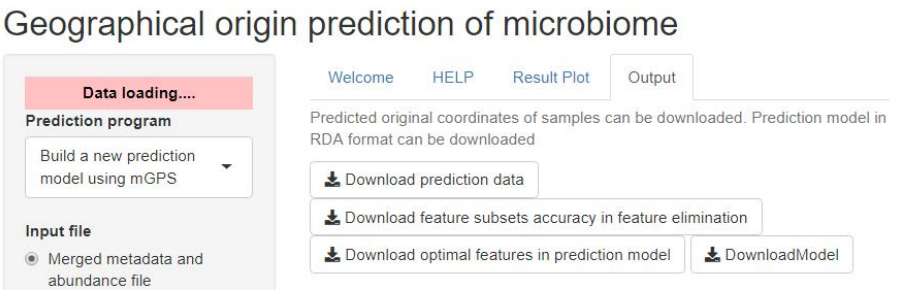
b. Prediction accuracy bar plot: The prediction accuracy of the model is shown per-site as the distances between the predicted and true sampling sites for the reference samples. The average prediction accuracy across all samples with each population given equal weight is shown on the left. You can download and save the png file of this plot through the button 'DownloadPlot' below the figure.



DownloadPlot

2.2.3 Output files

Data processing: Please wait while output files are being generated. When the prompt bar (*Data loading...*) disappears you can see the results and download files.



a. “Download prediction data” button - *Prediction_results.csv*: Records the predicted original coordinates of the reference samples (columns: LatPred and LongPred). Also contains the distance between the predicted and true sampling site for the reference samples (column: Distance_from_origin). The original metadata and abundance data will be merged.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		uuid	continent	city	latitude	longitude	taxa1_abundance	taxan_abundance	cityPred	latPred	longPred	Distance_from_origin	
2	1	1	africa	offa	8.1548483	4.7263633	0.00017	0	offa	7.881690979	11.71926498	771.4253212	
3	2	2	africa	offa	8.15485	4.726465	9.00E-05	0	offa	8.323236465	-0.264049619	550.1233564	
4	3	3	africa	offa	8.1556609	4.7244244	8.00E-05	0	offa	7.446023464	4.620732784	79.81995773	
5	4	4	america	new_york_city	40.70670433	-74.0111377	7.00E-05	2.00E-05	new_york_city	40.67137527	-72.10720062	160.7544161	
6	5	5	america	new_york_city	40.67953148	-73.9035273	0.00016	0	new_york_city	18.32634163	-43.18954086	3843.356172	

b. “Download feature subsets accuracy in feature elimination” button - *Features_subsets_accuracy.csv*: Records the accuracy of the prediction model with different features (microorganisms) subset size (column: n_vars). You can manually enter a specific range or value of subset size in the option '*Subsets in feature elimination (Optional)*". The algorithm will use the subset size with the highest accuracy (column: accuracy) to select the number of features. For example, the subset with a size of 100 has the highest accuracy value, and then the algorithm will select the most important 100 features to construct the model, and these features are recorded in the file *Optimal_features.csv*.

	A	B	C
1		n_vars	accuracy
2	1	50	0.643030303
3	2	100	0.764242424
4	3	150	0.723838384
5	4	200	0.731919192
6			

Hint: In addition to checking the given subset sizes, the algorithm also tries to use all features(taxa) to predict the origin. If the accuracy of the subset sizes uploaded by the user is lower than the accuracy of using all features, the algorithm will still choose to use all features for prediction to improve the accuracy. If the user still wants to use a specific number of features, the feature columns in the input file can be manually filtered according to the order of importance of the features in the output file "*Optimal_features.csv*" (Download optimal features in prediction model).

c. “Download optimal features in prediction model” button - *Optimal_features.csv*: Records the optimal features (microorganism) used to predict the original coordinates selected by the mGPS

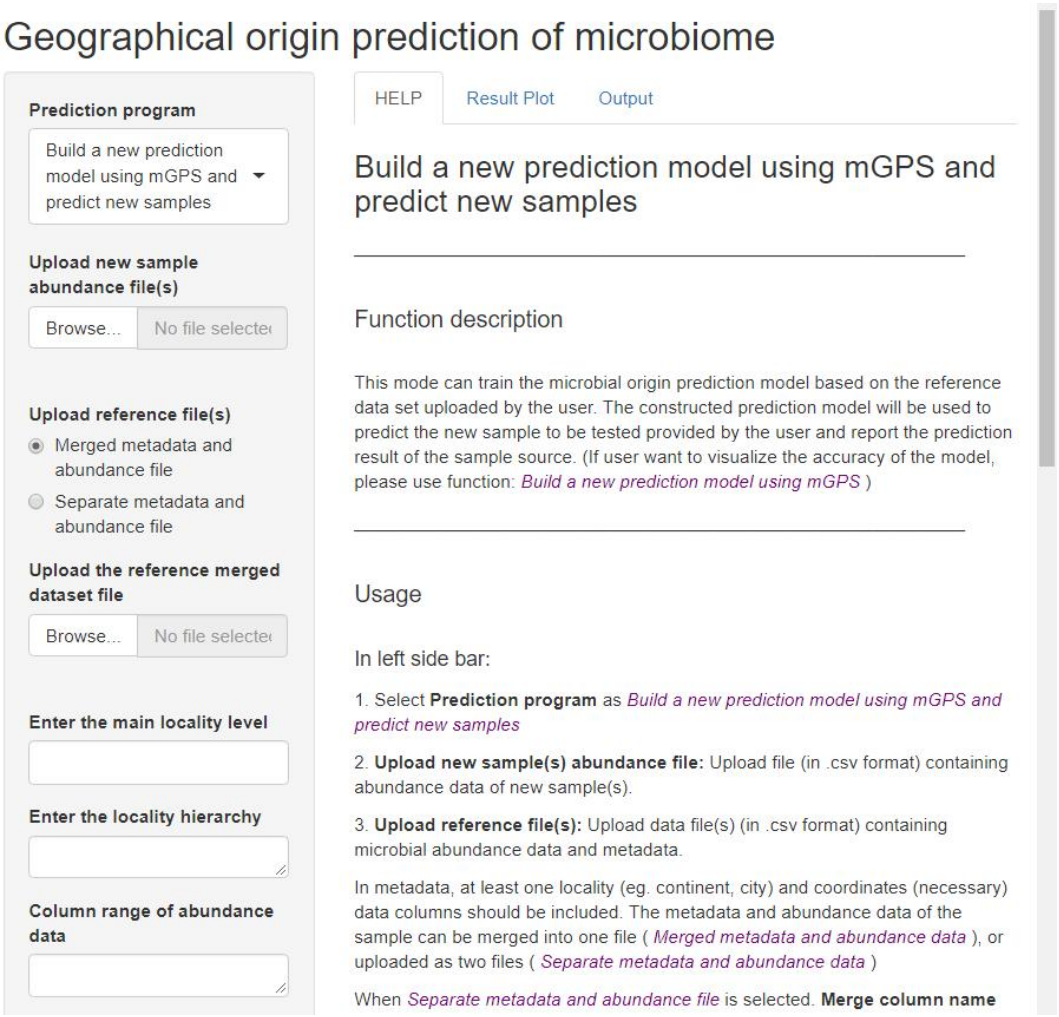
algorithm. The number of features is equal to the subset size with the highest accuracy value (n_vars) recorded in the *Features_subsets_accuracy.csv* file. The column "Overall" records the importance value of each feature. The higher the value, the more important the feature in the prediction algorithm.

	A	B	C	D	E
1		Overall	taxa		
2	Acetobacter.pasteurianus	0.0889962	Acetobacter.pasteurianus		
3	Acetobacter.aceti	0.0728943	Acetobacter.aceti		
4	Acetobacterium.woodii	0.0474666	Acetobacterium.woodii		
5	Acholeplasma.brassicae	0.0429964	Acholeplasma.brassicae		

d. "DownloadModel" button - *Prediction_model.Rda*: The origin prediction model built by mGPS algorithm saved in Rda format. You can download the model to view the details of the model through load into r. Also, this model can be uploaded to predict the origin of new samples through function 3 "Use an existing model to predict new samples".

```
load('Outputs/Prediction_model.R)
```

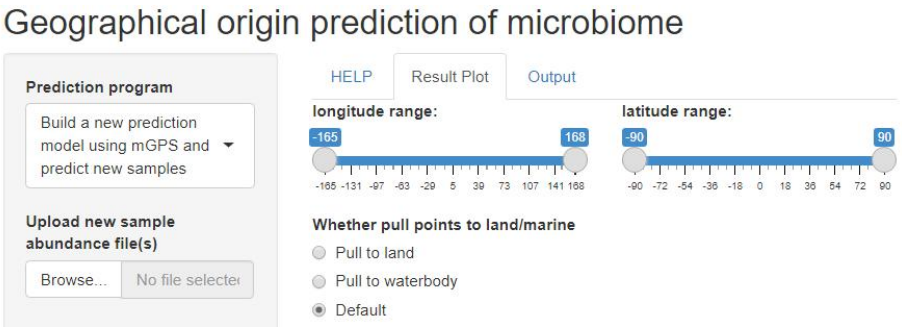
2.3 Function 2: Build a new prediction model using mGPS and predict new samples



2.3.1 Function 2 introduction

In this mode, you can train the microbial origin prediction model based on the reference data set uploaded by the user. The constructed prediction model will be used to predict the new sample to be tested provided by you and report the prediction result of the sample source. (If you want to visualize the accuracy of the model, please use the function: Build a new prediction model using mGPS)

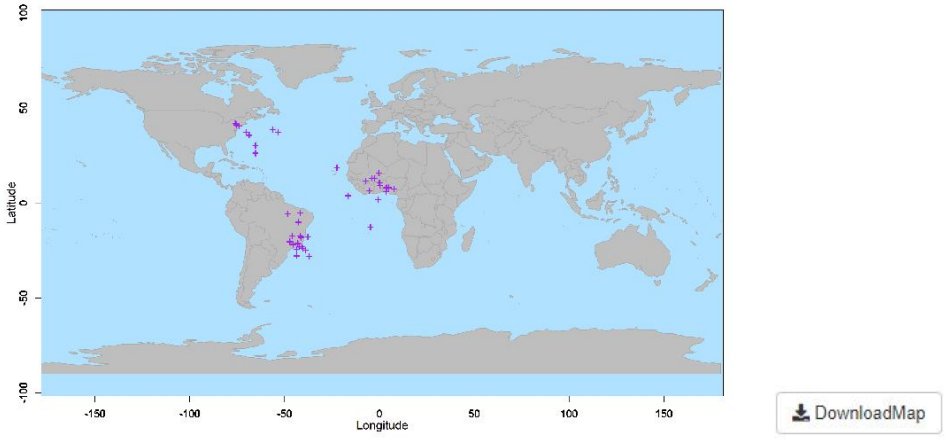
2.3.2 Output plot



The reference datasets will be used to construct an origin prediction model by mGPS. Then this model will be used to predict the origin of new samples.
If you want to zoom in on a specific area of the map, you need to adjust both the longitude and latitude ranges. "Pull to land" refers to pushing the predicted point to the nearest land, and "Pull to waterbody" refers to pushing the predicted point to the nearest water body.
When you select "Pull to land" and "Pull to waterbody", the sample prediction source coordinates will be changed, the content of the output file in the tab “Output” will also be changed accordingly. Remember, the data in output file is the same as the figure.

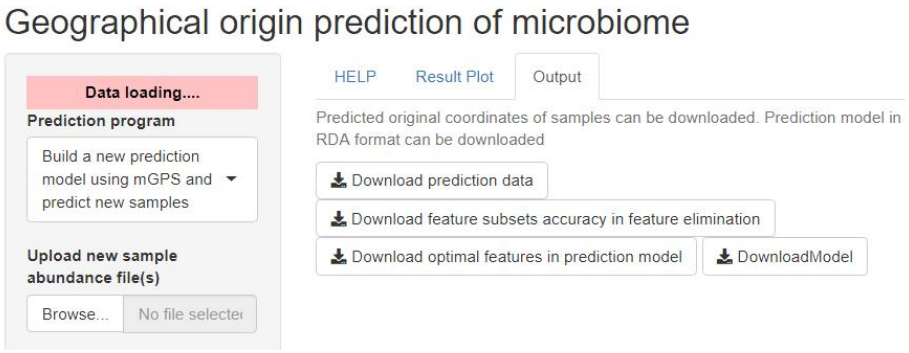
Example output plot:

a. World map: samples' prediction origins are plotted on the world map. By adjusting the range of longitude and latitude, you can select the plotting area of the figure on the map. In addition, you can choose whether to pull the predicted point into the continent or the ocean. You can download and save the png file of this map through the button ‘DownloadMap’ below the figure.



2.3.3 Output files

Data processing: Please wait while output files are being generated. When the prompt bar (*Data loading...*) disappears you can see the results and download files.



a. “Download prediction data” button - *Prediction_results.csv*: Records the predicted original coordinates of the new samples. (columns: LatPred and LongPred).

	A	B	C	D	E	F	G
1		taxa1_abundance	taxa2_abundance	taxan_abundance	cityPred	latPred	longPred
2	1	0.00017	0	0	offa	7.881690979	11.71926498
3	2	9.00E-05	0.00023	0	offa	8.323236465	-0.264049619
4	3	8.00E-05	0	0	offa	7.446023464	4.620732784
5	4	7.00E-05	2.00E-05	2.00E-05	new_york_	40.67137527	-72.10720062
6	5	0.00016	1.00E+00	0	new_york_	18.32634163	-43.18954086

b. “Download feature subsets accuracy in feature elimination” button - *Features_subsets_accuracy.csv*: Same content as this file in function 1. Build a new prediction model using mGPS. Jump

c. “Download optimal features in prediction model” button - *Optimal_features.csv*: Same content as this file in function 1. Build a new prediction model using mGPS. Jump
Features_subsets_accuracy.csv

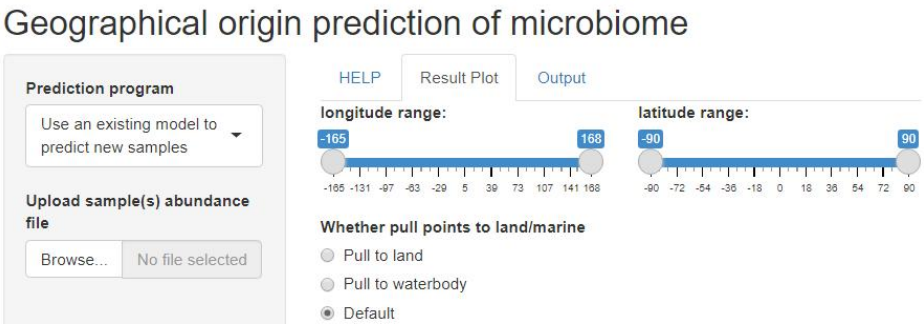
d. “DownloadModel” button - *Prediction_model.Rda*: Same content as this file in function 1. Build a new prediction model using mGPS. Jump

2.4 Function 3: Use an existing model to predict new samples

2.4.1 Function 3 introduction

In this mode, you can predict new sample origin based on an existing prediction model. You can download the model in the Output tab of functions: Build a new prediction model using mGPS or Build a new prediction model using mGPS and predict new samples. Jump to model

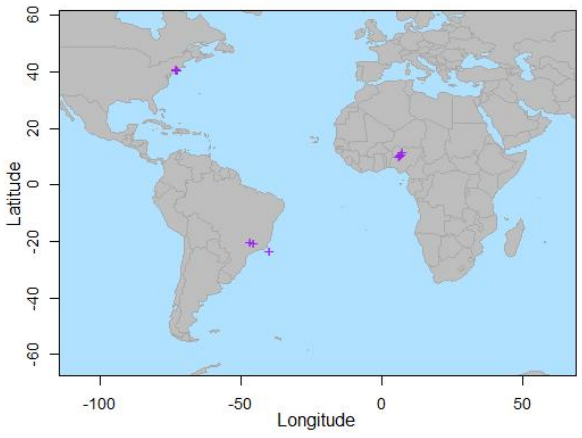
2.4.2 Output plot



The model uploaded will be used to predict the origin of new samples.
If you want to zoom in on a specific area of the map, you need to adjust both the longitude and latitude ranges. "Pull to land" refers to pushing the predicted point to the nearest land, and "Pull to waterbody" refers to pushing the predicted point to the nearest water body.
When you select "Pull to land" and "Pull to waterbody", the sample prediction source coordinates will be changed, the content of the output file in the tab “Output” will also be changed accordingly. Remember, the data in output file is the same as the figure.

Example output plot:

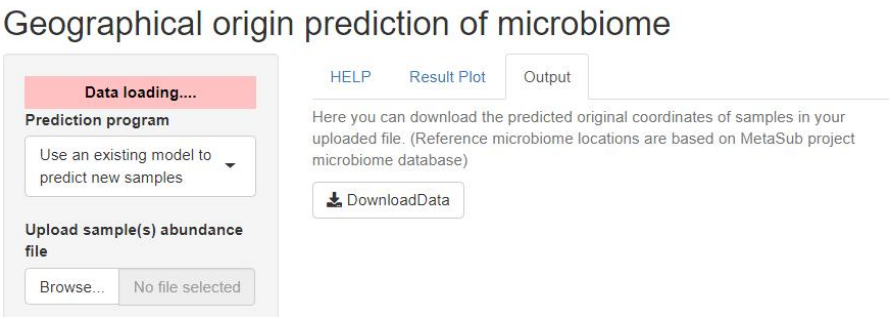
a. World map:
new samples' prediction origins are plotted on the world map. By adjusting the range of longitude and latitude, you can select the plotting area of the figure on the map. In addition, you can choose whether to pull the predicted point into the continent or the ocean. You can download and save the png file of this map through the button ‘DownloadMap’ below the figure.



DownloadMap

2.4.3 Output files

Data processing: Please wait while output files are being generated. When the prompt bar (*Data loading...*) disappears you can see the results and download files.



a. “DownloadData” button - *Prediction_results.csv*: Records the predicted original coordinates of the new samples. (columns: LatPred and LongPred).

	A	B	C	D	E	F	G
1		taxa1_abundance	taxa2_abundance	taxan_abundance	cityPred	latPred	longPred
2	1	0.00017	0	0	offa	7.881690979	11.71926498
3	2	9.00E-05	0.00023	0	offa	8.323236465	-0.264049619
4	3	8.00E-05	0	0	offa	7.446023464	4.620732784
5	4	7.00E-05	2.00E-05	2.00E-05	new_york_	40.67137527	-72.10720062
6	5	0.00016	1.00E+00	0	new_york_	18.32634163	-43.18954086

3 FAQ

3.1 If you meet error:

```
> runApp('mGPS_interface.r')
Warning in file(con, "w") :
cannot open file 'C:\Users\Temp\...': No such file or directory
Error in file(con, "w") : cannot open the connection
```

Workarounds: Try to restart the R (Rstudio). Sometimes this error is due to caching in Rstudio.

3.2 If you meet error:

```
Error in file(file, ifelse(append, "a", "w")) :
Cannot open the file:'.....' : Permission denied
Warning: Error in file:    cannot open the connection
```

Workarounds: Try to see if you have the file open in another program, so the file cannot be modified by the R application.

3.3 Figure problem on the interface page

The map figure on the interface has extra space or the map latitude and longitude range are different from the selected one. Due to the shiny page settings, the figure will be automatically expanded to a certain size, that is, it fills the entire screen. So sometimes there is extra space on

the map. At the same time, to prevent serious deformation of the world map, the aspect ratio is locked when the map is plotted.
Workarounds: If you want to get a zoomed-in portion of the map, you need to adjust both longitude and latitude.

3.4 The figure downloaded via the button does not match what you see on the interface

Since the figure displayed on the shiny is forced to resize, the deformation is more serious when the interface page is enlarged. The downloaded figure is directly generated by the R code and is not forced to be deformed by shiny, so the downloaded figure will be inconsistent with the seen figure.
Workarounds: If you want to keep what you see consistent with what you download, it is recommended to use the shiny original size interface instead of enlarging the webpage. Or directly click the right mouse button on the figure to save it.

3.5 If you meet error:

```
Error: Object of type ‘closure’ is not subsettable
```

The reasons maybe 1. The uploaded file is damaged; 2. After one function mode has started to run, jumping to other function modes confuses the program reading parameters.
Workarounds: Check your uploaded file. Try to reload the app. And don't jump to other function modes when using one function mode (that is, don't click on other “Prediction programs” during the running process). It needs to reload the app after using one function mode and then use another function mode, which can prevent the program from running chaotically.

3.6 If you meet error:

```
Error in :: NA/NaN parameter
```

Workarounds: Check the format of the input parameters on the interface. Eg., whether the separator between data meets the requirements.

3.7 If you meet error:

```
Error: undefined columns selected
```

Workarounds: Check the input of the *main locality level* and *locality hierarchy* to ensure that each input name is the same as the header of the input file columns.