uniForest v.1 Manual

Citation:

Leigh R.J., Murphy R.A., and Walsh F. (2021) Paper title, Journal metrics

Description:

uniForest is a user-friendly script for outlier processing in microbiome studies.

Table of contents:

Before you start:

Data should be **tab delimited** format (.tsv) with the first column containing the taxon IDs (or names) and the first row containing the groups each sample belongs to. The first row should contain the identifier word "**Group**". An example of this data is shown below (example.tsv; Table 1):

Table 1: Example dataset

Group	A	A	A	A	A	A	A	A	B	B	B	B	B	B	B	B
TaxonA	117	118	104	108	108	1117	577	33	217	200	227	203	218	570	1446	144
TaxonB	221	208	206	227	214	945	1730	147	223	230	214	230	216	1901	1205	55
TaxonC	40	32	30	36	45	481	297	13	43	43	46	49	46	412	291	21
TaxonD	64	70	68	62	71	361	410	18	63	71	64	69	75	419	565	25
TaxonE	116	120	100	115	108	1103	845	58	106	100	120	120	111	1110	692	70
TaxonF	25	21	16	30	17	347	189	7	209	217	212	214	208	645	1752	83

Where to start:

This script is presented as an "interactive python notebook" (ipynb) file and can be ran using Google Colaboratory (Colab) (https://colab.research.google.com/). Once Colab is open in your browser, select "**upload notebook**" in the "**File**" tab in the top, right-hand corner of the screen (**Figure 1**), and upload the ipynb file.

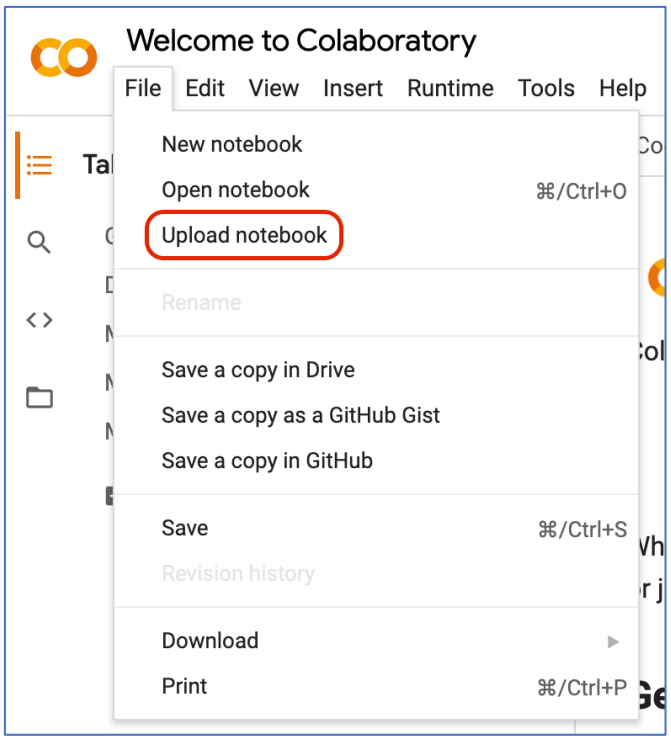

Figure 1: How to upload the ipynb file

Setting up uniForest:

Once uniForest has been loaded, the code pertaining to each cell can be viewed by double clicking the title (*eg.* **Figure 2**).

```
#@title **File upload**
#@markdown A single tsv file is required for upload.
from google.colab import files
dataset = files.upload()
```

Figure 2: Example of code underlying a given cell

The only cell a user is required to view is "**Metric specification**" (if changes from the default parameters are required) which presents 3 options (**Figure 3**):

```
scale_data = "yes"
iForest_contamination="auto"
iForest_bootstrap="False"
```

Figure 3: The metric specification window

1. Data **scaling** ensures all samples have an equivalent sum. This is to mitigate errors due to data with low coverage/depth. The default is "**yes**".
2. iForest **stringency** is set using contamination. This value can be set between 0 (least stringent) and 0.5 (most stringent). Numerical values are set without quotation marks (**Figure 4**). The iForest algorithm can set this value automatically for each sample using "auto". The default is "**auto**" (with quotation marks).
3. iForest **bootstrapping** can be turned on with "**True**". Default usage is **"False"**.

```
iForest_contamination=0.2
```

Figure 4: Metric specification window with a numerical input

Once each of these 3 options are satisfactorily set, uniForest is ready to process data. The impution metric (metric) describes the outlier replacement strategy and can be one of **"median"**, **"mean"**, **"geometric"** (geometric mean), or **"harmonic"** (harmonic mean).. The most optimal score is the greatest Bray-Curtis dissimilarity score improvement after the data has been processed (as described in the publication).

Running uniForest:

When ready, uniForest can be ran using a using the "**Run all**" option in the "**Runtime**" tab at the upper, right-hand side of the screen (**Figure 5**).

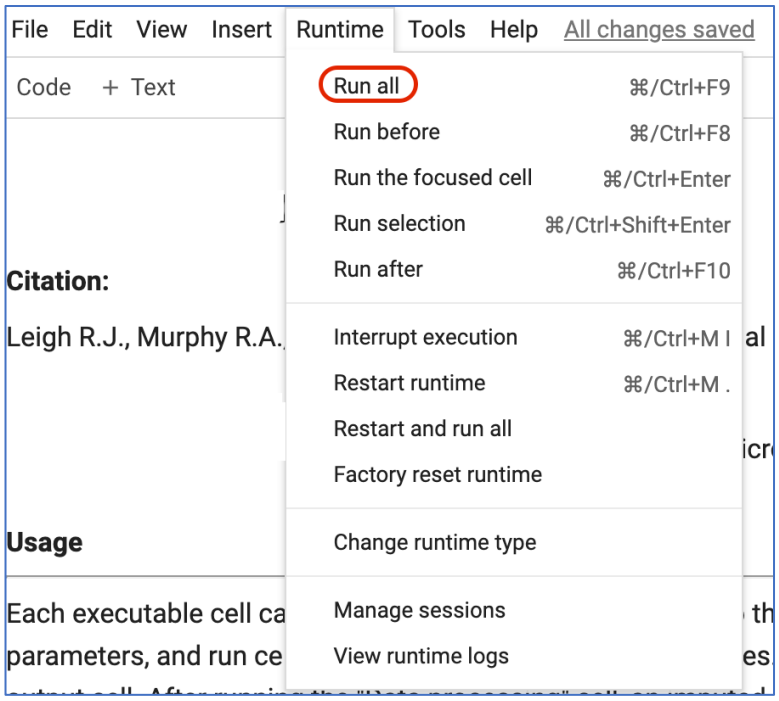

Figure 5: Location of the "Run all" command

Once the "Run all" command has been selected. The user is required to upload a file in the presented window of the "**File upload**" cell (Figure 6).

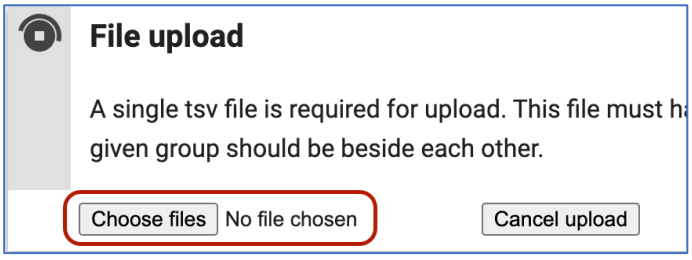

Figure 6: Location of the file upload function

A principal component analysis (PCA) plot is produced for the raw data and the final data in the "**View PCA**" cells. The plots are interactive and individual groups can be zoomed into using a mouse and groups can be excluded by clicking on their icon in the **legend**. Each PCA plot can be downloaded once the user has found a satisfactory configuration using the camera icon in their respective windows (**Figure 7**).

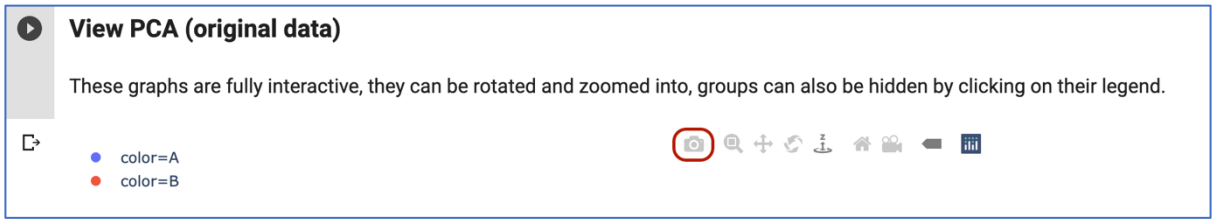

Figure 7: Location of the PCA download function

The imputed dataset PCA is automatically set to the most optimal impution metric (highest mean improvement in the Scores.tsv output file). This can be changed to "mean", "median", "geometric" (geometric mean), or "harmonic" (harmonic mean). The size of a given PCA can be changed by double clicking the PCA cell and changing the numeric values where the first value corresponds to figure height and the second to figure width (**Figure 8**). The default is 1000 × 1250. Once these values have been updated, rerun the cell using the black mouseover arrow beside the cell. Each PCA plot must be updated separately.

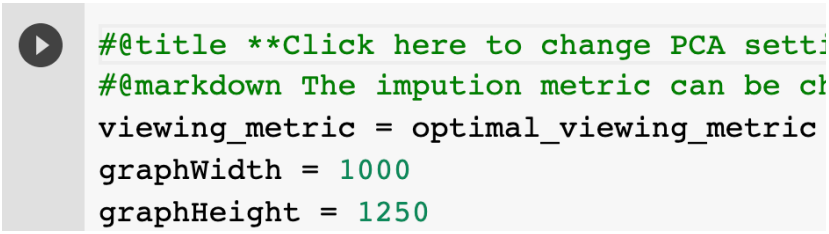

Figure 8: Options for changing and rerunning PCA plots

Results files produced by uniForest can be retrieved using the **folder icon** on the right-hand side of the screen (**Figure 9**). Five files are produced "**[metric]_StatisticsTable.tsv**", (StatisticsTable), "**[metric]_ImputedTable.tsv**" (ImputedTable), and "**[metric]_ChangesTable.tsv**" (ChangesTable), "**[metric]_intraGroup_separation.tsv**" (intragroup_sep), and "**[metric]_interGroup_separation.tsv**" (interGroup_sep), "**[metric]_alphaDiversityObservations.tsv**" (alphaDivObs), and "**[metric]_alphaDiversityStatistics.tsv**" (alphaDivStats) where "[metric]" is the impution metric. An additional file (Scores.tsv) reports the improvement scores for each impution metric. The content of each output file is discussed in a later section of this document.

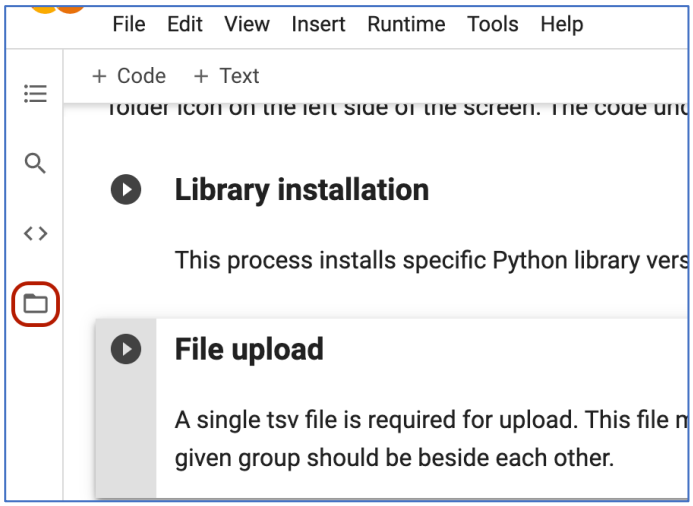

Figure 9: Location of the output file download function

Rerunning uniForest:

The "**Library installation**" cell needs to only be ran once so running a new dataset can be achieved by first clicking on the "**File upload**" cell then clicking on the "**Run after**" option in the "Runtime" tab on the top, right hand side of the screen (**Figure 10**). If the same dataset is

to be reanalysed using new metrics, the "**File upload**" cell can be bypassed by clicking on the "**Metric specifications**" cell and using the "**Run after**" function.

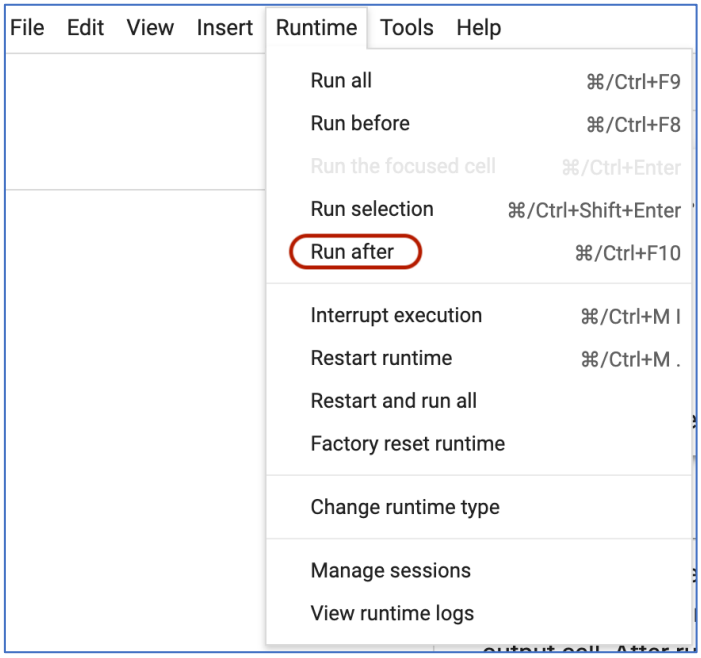

Figure 10: Location of the "Run after" function

If preferred, the script can be ran as individual cells using the black mouseover arrows beside their titles (in **bold**; *eg.* **Figure 11**). Each cell should be run in order.

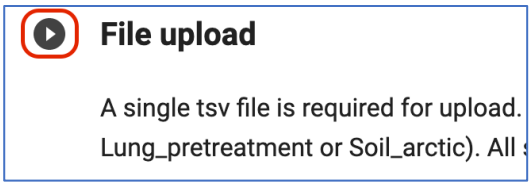

Figure 11: Location of the mouseover arrow beside a given cell

Output files:

As mentioned above, uniForest produces seven files (StatisticsTable, ImputedTable, ChangesTable, intraGroup_sep, interGroup_sep, alphaDivObs, and alphaDivStats) after each

run. Output files from the same input dataset are overwritten at the end of every rerun so users are advised to download output files at the end of each run if they desire to compare outputs for the same dataset.

The **ImputedTable** is has the same number of taxa and samples as the input dataset. Outliers are removed and imputed in this table.

The **StatisticsTable** contains 12 columns:

1. "**Dataset**": The dataset that the following statistics were derived from. This entry can be one of three options "Original" (referring to the raw, unprocessed dataset), "Imputed" (referring to the dataset after outliers were removed and imputed), and "Inliers" (if applicable; referring to the inlier data after outliers were removed but before impution. Inliers are used to calculate the impution value).If no outliers were detected, no "Inlier" statistics will be reported and the "Original" and "Imputed" statistics will be equivalent.
2. "**GroupID**": The group (as specified in the input file header) for the given dataset that is being reported.
3. "**TaxonID**": The taxon (*eg.* Firmicutes, *E. coli*) being reported
4. "**Mean**": The mean of the samples being reported
5. "**SDev**": The standard deviation of the samples being reported
6. "**Median**": The median of the samples being reported
7. "**GMean**": The geometric mean of the samples being reported
8. "**HMean**": The harmonic mean of the samples being reported

9. "**CoV**": The coefficient-of-variation of the samples being reported
10. "**Min**": The minimum value of the samples being reported
11. "**Max**": The maximum value of the samples being reported
12: "**Count**": The count of the values being reported. This value will be equivalent for "Original" and "Imputed" datasets and will only be different for "Inliers" datasets.

The "**ChangesTable**" has 11 columns:

1. "**GroupID**": The group (as specified in the input file header) for the given dataset that is being reported.
2. "**TaxonID**": The taxon (*eg.* Firmicutes, *E. coli*) being reported
3. "**Removed**": The sum of removed outliers
4. "**MinOrig**": The minimum value within the original samples
5. "**MaxOrig**": The maximum value within the original samples
6. "**MinImput**": The minimum value within the imputed samples
7. "**MaxImput**": The maximum value within the imputed samples
8. "**CoVOrig**": The coefficient-of-variation within the original samples
9. "**CoVImput**": The coefficient-of-variation within the imputed samples
10. "**CoVDiff**": The subtractive difference between coefficients-of-variation

The effect of outlier removal and impution can be extrapolated by comparing "MinImput" to "MinOrig" and by comparing "MaxImput" to "MaxOrig". Any datapoint from the original dataset that is less than MinImput or greater than MaxImput was removed and imputed.

The "**IntraGroup_sep**" table has 19 columns which statistically describe the distances between raw data (first header instances) and processed data (second header instances).

The "**InterGroup_sep**" table descrbes the statistical inter-group differences (Group A vs Group B) for the raw data (rawScore) and the processed data (processedScore).

The "**alphaDivObs**" table provides Simpson's D, Simpson's E, Shannon's H', and Chao1 for each column for the raw data and processed datasets.

The "**alphaDivStats**" has 20 columns that statistically describe the alpha diversity metric differences between raw data (first header instances) and processed data (second header instances).