Beta-Diversity Script

Task

1) Calculate Beta diversity between the samples

2) Visualize in two-dimensional space the relationship between the given groups

3) Statistically evaluate whether these groups are significantly separated

4) Produce the plots of three different clustering validation indices

Background

Beta diversity measures the level of similarity or dissimilarity between two samples. The proper use of beta diversity can be highly informative as it can summarize information of multivariate data and produce comparable results between the samples. Also, it does not focus on the abundance of specific bacterial taxa but takes into account the overall microbial communities of the samples. The most widely used measures are the Bray-Curtis (Bray & Curtis, 1957) and the weighted or unweighted Unifrac (Lozupone et al., 2011). Instead of the previously mentioned metrics, this script uses the Generalized Unifrac distance metric (Chen et al., 2012), which is a balanced option between the weighted and unweighted Unifrac. The multidimensional scaling algorithm (MDS) (Gower, 1966) and its Non-metric version (NMDS) (Minchin, 1987) are applied to reduce the dimensionality of the data, and the Permutational Multivariate Analysis of Variance (PERMANOVA) (Anderson, 2001) is performed to evaluate if the groups are significantly separated.
Although the beta diversity analysis can provide us with valuable information about the relation between the groups, in many cases, it is helpful to examine the structure of a dataset without specific group labeling. Clustering a group without using labels or prior information is defined as de novo clustering. In this script, the Partitioning Around Medoids algorithm (PAM) (Kaufman & Rousseeuw, 2009) is used for the de novo clustering of the groups.
The optimal number of clusters can be determined by analyzing the plots of Calinski-Harabasz (Caliñski & Harabasz, 1974) and silhouette (Rousseeuw, 1987) index and the Within Sum of Squares (WSS) plot.

Inputs

The inputs of the "Beta-Diversity" script are the following:

- An OTUs or ASVs abundance table. In this table, the rows should represent the OTUS or ASVs and the columns the samples. An extra column with the taxonomic classification of the OTUs/ASVs can be included; besides this column, no other information should be present in this file. A normalized table of this format can be produced via the

"normalization" script of the RHEA pipeline (Lagkouvardos et al., 2017). The following image presents an example of the acceptable form of the abundance table.

##OTU ID	Sample1	Sample2	Sample3	taxonomy
otu1	0	0	109	Bacteria;Firmicutes;Bacilli;Erysipelotrichales;Erysipelotrichaceae;Faecalitalea;
otu2	0	0	0	Bacteria;Actinobacteriota;Coriobacteriia;Coriobacteriales;;;
otu3	0	16	206	Bacteria;Firmicutes;Clostridia;Lachnospirales;Lachnospiraceae;;
otu4	0	164	0	Bacteria;Firmicutes;Clostridia;Oscillospirales;Eubacterium coprostanoligenes group;;
otu5	9	219	3058	Bacteria;Firmicutes;Clostridia;Lachnospirales;Lachnospiraceae;;
otu6	0	373	539	Bacteria;Firmicutes;Clostridia;Oscillospirales;Oscillospiraceae;Flavonifractor;
otu7	0	0	282	Bacteria;Firmicutes;Clostridia;Lachnospirales;Lachnospiraceae;;
otu8	0	0	0	Bacteria;Fusobacteriota;Fusobacteriia;Fusobacteriales;Fusobacteriaceae;Fusobacterium;
otu9	0	32	1	Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiaceae;Clostridium sensu stricto 1;
otu10	0	0	0	Bacteria;Proteobacteria;Gammaproteobacteria;Burkholderiales;Sutterellaceae;Sutterella;

- The second necessary input file is a phylogenetic tree corresponding to the abundance table's OTUs or ASVs. The tree must be in Newick format and is a prerequisite for calculating the Generalized Unifrac Distances Table. The preprocessing of the raw data through the IMNGS platform (Lagkouvardos et al., 2016) provides this type of tree. ***If a tree is not available***, the user can instead provide a dissimilarity matrix of the samples. The "beta diversity" script of the RHEA pipeline can generate this kind of matrix. Even though the user can use any metric to calculate the dissimilarity measures, we strongly recommend using the Generalized Unifrac distance metric.
- The final requirement is a mapping file that contains the labels of the samples. The information of the mapping file is necessary for the labeling and determination of the reference and test groups. The rows of the mapping file should have the same sample names as the OTUs/ASVs table, and the columns should contain the labeling information. The existence of extra columns will not affect the program.

The following table presents the acceptable types of input files.

File	**Accepted formats**
OTUs or ASVs table	.txt, .tab, .csv, .tsv (tab or comma separated)
Dissimilarity matrix	.txt, .tab, .csv, .tsv (tab or comma separated)
Mapping file	.txt, .tab, .csv, .tsv (tab or comma separated)
Phylogenetic tree	.nwk, .tre

Besides the input files, the user has to fill in some extra parameters:

- The first is whether the OTUs or ASVs table is normalized or not. If the table is not normalized, then the first step will be the normalization of the table so the sum of the counts will be equal across all the samples. (Required)
- The name of the column from which the script will draw the information about the label of the samples. (Required)
- The name of the reference group. The user must provide at least one name. (Required)
- The names of the test groups. The user can determine the names of one or more test groups. (Optional)

Outputs

Two folders and a dissimilarity matrix are among the outputs of this script:

- The first folder is called the "Beta Diversity". This folder contains all the files related to the beta diversity analysis.
  The user can find in this folder:
  - A PDF file with the MDS and NMDS plots of the given groups. The distances of these plots were calculated using the generalized UniFrac metric or by using the provided dissimilarity matrix,
  - Two PDF files containing the MDS and NMDS plots for the pairwise comparisons between the groups. In the case of the MDS plots, for each pair, the p-value of the PERMANOVA test is calculated. All the p-values are adjusted by using the Benjamini–Hochberg method (Benjamini and Hochberg, 1995).
  - A PDF file with a phylogram based on the Ward's minimum variance method showing the hierarchical clustering of samples and,
  - The corresponding tree in Newick format.
- The second folder is named "Optimal Number of Clusters" and includes all the de novo clustering-related files. The "Optimal Number of Clusters" folder contains the following elements:
  - A folder with the clustering-related information for each of the reference and test groups. Every folder holds two pdf files:
    1) The first PDF file has the plots of Calinski-Harabasz and silhouette index, the prediction strength and WSS plots, and the plot of the BIC values for six models as they are produced by the model-based clustering based on finite Gaussian mixtures for each of the groups,
    2) The second PDF contains the MDS plots that visualize the groups using different numbers of clusters each time.
  - A PDF report that suggests to the user the optimal number of clusters for each of the groups. To make this recommendation, the script first calculates the optimal number of clusters for each index and then selects the number with the highest frequency. In case of a tie, this suggestion is based on the results of the Calinski-Harabasz index.

The appraisal of the previously mentioned plots in conjunction with the prior knowledge of the dataset can help the user decide the optimal number of clusters for each group. Alternatively, the user can follow the recommendation as the script calculates it. In any case, the information about the optimal number of clusters for each group is necessary for the "DivCom" script.

Furthermore, the dissimilarity matrix across all the samples is calculated and printed. This matrix can be used as input in the "Distances" script. **Warning:** this dissimilarity matrix can be used only for the parameters given in the initialization section. A new dissimilarity matrix should be produced if the user wishes to perform an analysis with different parameters.

Troubleshooting

PROBLEM	SOLUTION
The script cannot read the input files.	➢ Set the right path ➢ Check the spelling of the file names
The script cannot find the column of the mapping file	➢ Check that the provided column name exists and has been spelled correctly.
The script cannot compute the distances matrix	➢ Check that you have provided a phylogenetic tree in Newick format
The mapping file of the script does not contain any samples to cluster	➢ Check the spelling of the mapping file's column name ➢ Check the spelling of the group names

References

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, *26*(1), 32–46.

Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, *27*(4), 325–349.

Caliński, T., & Harabasz, J. (1974). A Dendrite Method Foe Cluster Analysis. *Communications in Statistics*, *3*(1), 1–27. https://doi.org/10.1080/03610927408827101

Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., Collman, R. G., Bushman, F. D., & Li, H. (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics*, *28*(16), 2106–2113. https://doi.org/10.1093/bioinformatics/bts342

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, *53*(3–4), 325–338.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis* (Vol. 344). John Wiley & Sons.

Lagkouvardos, I., Fischer, S., Kumar, N., & Clavel, T. (2017). Rhea: a transparent and modular R pipeline for microbial profiling based on 16S rRNA gene amplicons. *PeerJ*, *5*, e2836. https://doi.org/10.7717/peerj.2836

Lagkouvardos, I., Joseph, D., Kapfhammer, M., Giritli, S., Horn, M., Haller, D., & Clavel, T. (2016). IMNGS: A comprehensive open resource of processed 16S rRNA microbial profiles for ecology and diversity studies. *Scientific Reports*, *6*. https://doi.org/10.1038/srep33721

Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J., & Knight, R. (2011). UniFrac: An effective distance metric for microbial community comparison. In *ISME Journal* (Vol. 5, Issue 2, pp. 169–172). https://doi.org/10.1038/ismej.2010.133

Minchin, P. R. (1987). An evaluation of the relative robustness of techniques for ecological ordination. In *Theory and models in vegetation science* (pp. 89–107). Springer.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.