Beta-Diversity Script

Task

Calculate the *beta*-diversity between samples, test and visualize groupings of samples according to given categories

Background

Beta-diversity gives a measure of similarity between different microbial profiles described by the Operational Taxonomic Units (OTUs) table. The most common approaches to calculate similarity of microbial profiles are the Bray-Curtis dissimilarity index and the weighted and unweighted UniFrac distances [1,2]. While Bray-Curtis only considers the shared composition across samples, UniFrac takes into consideration the phylogenetic distance between OTUs. Weighted UniFrac also considers the abundance of each OTU. Because unweighted and weighted UniFrac are sensitive to rare and dominant OTUs, respectively, a balanced version, called generalized UniFrac, was proposed more recently [3] and is used in Rhea. Visualization of the multidimensional distance matrix in a space of two dimensions is performed by Multi-Dimensional Scaling, or its more robust and unconstrained non-metric version (NMDS) [4]. A permutational multivariate analysis of variance using distance matrices (vegan::adonis) is performed in each case to determine if the separation of groups is significant, as a whole and in pairs [5]. In addition, a dendrogram of all the samples calculated by hierarchical clustering using the Ward's minimum variance method is proposed in order to obtain an alternative view of individual sample positions [6]. Besides grouping the samples according to given categories it is also possible to perform a de-novo clustering (*De-novo-Clustering.R*). Clusters are constructed by assigning the samples to the nearest medoid to minimize dissimilarities between observations [7]. An appropriate number of *k-clusters* can be determined by Calinski-Harabasz (CH) Index.

References

1. Bray, J. R. & Curtis J. T. (1957). An ordination of upland forest communities of southern Wisconsin. Ecological Monographs 27:325-349

2. Lozupone, C. A., Hamady, M., Kelley, S. T., & Knight, R. (2007). Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. Applied and environmental microbiology, 73(5), 1576-1585.

3. Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., ... & Li, H. (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics, 28(16), 2106-2113

4. Minchin, P.R. (1987). An evaluation of relative robustness of techniques for ecological ordinations. Vegetatio 69, 89–107.

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

6. Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? Journal of Classification, 31(3), 274-295.

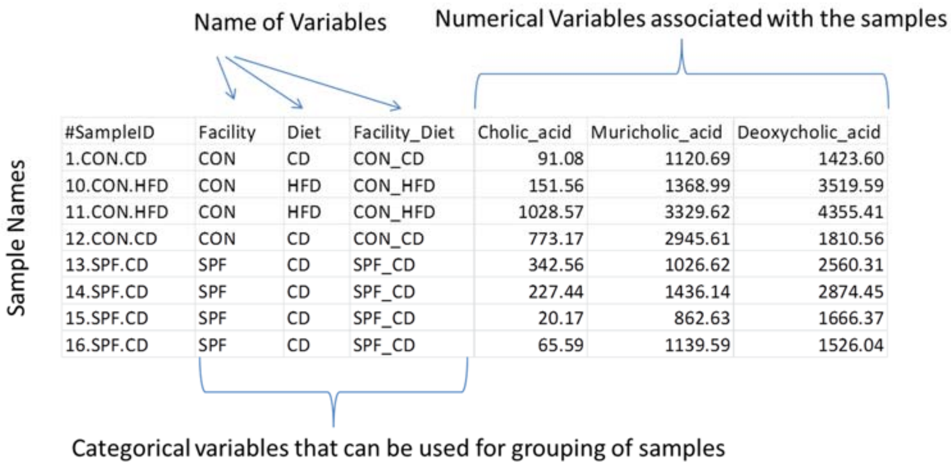
7. Reynolds, A., Richards, G., de la Iglesia, B. & Rayward-Smith, V. (1992) Clustering rules: A comparison of partionining and hierchaical clustering algorithms. Journal of Mathematical Modelling and Algorithms 5, 475-504.

Input

The main script expects 3 standard input files.

- An OTU-table with normalized counts (without taxonomic classification). This table can be automatically produced from a standard OUT-table using the Normalization script in Rhea. A copy is then placed in the 3.Beta-Diversity folder assuming the folder structure of Rhea is retained.

- A mapping file containing information on sample groups. The most basic format of this file contains two columns (with headers on the first line), the first column containing the sample names and the second containing a categorical variable that is used to separate the samples into groups. Additional columns with different groupings can be added in a series of columns. Metadata (additional numerical measurements from the samples) do not affect the script. An example of mapping file can be seen bellow.


Name of Variables
Numerical Variables associated with the samples
Sample Names
Categorical variables that can be used for grouping of samples

#SampleID	Facility	Diet	Facility_Diet	Cholic_acid	Muricholic_acid	Deoxycholic_acid
1.CON.CD	CON	CD	CON_CD	91.08	1120.69	1423.60
10.CON.HFD	CON	HFD	CON_HFD	151.56	1368.99	3519.59
11.CON.HFD	CON	HFD	CON_HFD	1028.57	3329.62	4355.41
12.CON.CD	CON	CD	CON_CD	773.17	2945.61	1810.56
13.SPF.CD	SPF	CD	SPF_CD	342.56	1026.62	2560.31
14.SPF.CD	SPF	CD	SPF_CD	227.44	1436.14	2874.45
15.SPF.CD	SPF	CD	SPF_CD	20.17	862.63	1666.37
16.SPF.CD	SPF	CD	SPF_CD	65.59	1139.59	1526.04

- A rooted phylogenetic tree of the OTUs in Newick format to be used for calculation of the generalized UniFrac distances. Several programs can be used to calculate and export phylogenetic trees (e.g. MEGA7). For the purposes of the present script, a Maximum Likelihood (ML) tree with 300 bootstraps is sufficient. The IMNGS platform (www.imngs.org) creates a ML approximation (FastTree) tree as a standard output that can be directly used here.

A selection of the column name of the categorical variable contained in the mapping file and meant to be used for grouping must also be defined by the user in the editable section of the script.

Examples for all input files can be found in the 0.Original-Data/Template-Data folder contained in the Rhea distribution.

The de-novo-Clustering script expects 2 input files

- A mapping file containing information on sample groups. The most basic format of this file contains two columns (with headers on the first line), the first column containing the sample names and the second containing a categorical variable that is used to separate the samples into groups. Additional columns with different groupings can be added in a series of columns. Metadata (additional numerical measurements from the samples) do not affect the script. An example of mapping file can be seen bellow.
- A tab-delimited distance matrix calculated across all samples. The files is generated by the main Beta-Diversity script.

Output

The script creates 5 standard output files:

- The main output is a PDF file with the MDS and NMDS plots calculated from the generalized UniFrac dissimilarity matrix. These plots are annotated with p-values of the PERMANOVA test indicating the significance of group separations and the dissimilarity scale of the grid (e.g. d=0.2 mean that the distance between two grid lines represent approximately 20% dissimilarity between the samples). Due to the restriction of the dimensions the actual dissimilarities across samples might not be visualized correctly.

- A second PDF file contains the pairwise comparisons between groups, with p-values of the PERMANOVA test for significant separation between pairs of groups. This is applied whenever more than two groups exist and the PERMANOVA test across all groups is significant. The Bonferroni-Hochberg method is used to correct for multiple testing.

- A phylogram based on the Ward's minimum variance method showing the hierarchical clustering of samples (PDF file).

- A tab-delimited file containing the dissimilarity matrix calculated across all samples. This file can be used for extracting true dissimilarities across samples or exploring native clustering across samples (e.g. k-mean clustering).

- A third PDF file contains the distribution of the CH Index over all possible clusters. An optimal number of clusters can be determined by reading this graph.

De-novo-Clustering creates 2 output files:

- A PDF file with the MDS and NMDS plots calculated from the generalized UniFrac dissimilarity matrix. These plots are annotated with p-values of the PERMANOVA test indicating the significance of groups, based on a given number of medoids, separations and the dissimilarity scale of the grid (e.g. d=0.2 mean that the distance between two grid lines represent approximately 20% dissimilarity between the samples). Due to the restriction of the dimensions the actual dissimilarities across samples might not be visualized correctly.

- A mapping file with an additional column for clustering. Each sample is assigned to the corresponding cluster based on the applied partitioning algorithm.

Important Notes

The NMDS calculation algorithm is heuristic and the projection can slightly differ between runs of the script. Furthermore, the axes in MDS and NMDS plots are not bound to the projection, meaning that the projection can be turned or flipped in a graphic manipulation program as long as the relative positions of the samples are not modified.

Common problems

- The path to the script is not set correctly
- The input file names are incorrect
- The input files are of different format (e.g. tree format)
- The column name selected for grouping does not exist or contains typos (case sensitive).
- Only one group or too few samples are available for statistics