DivCom Script

Task

1) Perform de novo clustering to both the reference and test groups

2) Calculate the pairwise distances and find the closest reference cluster

3) Conduct an automated statistical analysis

4) Produce graphs, tables, and statistical measures.

The above information will contribute to a better understanding of the interrelation between the different groups under study.

Background and Method

Any kind of substructure in a dataset is a parameter that we should always take into account, specifically when we investigate the relationship between different groups. The practice of considering the groups under study as entireties can lead us, in some cases, to misleading and deceptive conclusions. Also, the reliance on the visual representation of a dataset, usually in the form of a PCoA plot, in conjunction with the results of statistical tests like PERMANOVA may not be the appropriate strategy due to the multivariate nature of the data.

The' DivCom' script applies a novel approach for advanced beta diversity analysis. This approach aims to compare different groups in a more efficient and detailed way and reveal their interrelation. The script employs the idea of dividing the groups using de novo clustering and then comparing these clusters using as metric their distances. According to the methodology of the 'DivCom' script, the samples of the control group are clustered, and then, the most representative points for each of these clusters are selected. Consequently, all the distances of the remaining test samples from these pre-selected points are calculated and then assessed. This process can assist us in drawing conclusions about the level of divergence between the control and test samples

As an extra step of the analysis, the script performs de novo clustering also to the test groups. Following, every subcluster is compared with the representative points of the control group. This process results in outputs that compare the structures of the reference and test groups. Therefore, it is easier for the user to reveal the substructural similarities and existing relations between the groups.

DivCom is a distance-based tool that compares different groups by taking into consideration the phylogenetic distances between observed organisms and using statistical measures to evaluate the results. Therefore, the Partitioning Around Medoids (PAM) algorithm (Kaufman and Rousseeuw,2009) is applied to cluster the samples (cluster::pam), and Generalized Unifrac (Chen et al., 2012) is the default distances metric used by the program (GUniFrac::GUniFrac). The statistical hypothesis testing relies on the Wilcoxon Rank Sum test (Mann and Whitney, 1947; Wilcoxon, 1992) for the continuous variables (stats::wilcox.test), the Chi-square test for the categorical variables (stats::chisq.test), Permutational analysis of multivariate dispersions

(PERMDISP) (Anderson, 2006) for the dispersion similarity comparison of the groups (vegan::betadisper and permutest), and Permutational multivariate analysis of variance (PERMANOVA) (Anderson, 2001) for the similarity comparison of the groups (vegan::adonis). All the p-values are adjusted using the Benjamini–Hochberg method (Benjamini and Hochberg, 1995) (stats::p.adjust). TheMultidimensional Scaling (MDS) algorithm (Gower, 1966) is applied for the ordination analysis (stats::cmdscale), and finally, scatter plots, boxplots, barplots, and phylograms are used to visualize the findings (ade4::s.class, ggtree:ggtree, ggplot2).

Inputs

The inputs files of the "DivCom" script are exactly the same as those of "Beta-Diversity":

- An OTUs or ASVs table which can be either normalized or not. In this table, the rows should represent the OTUS or ASVs and the columns the samples. 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. Except for the counts of the OTUs/ASVs and an optional column containing their taxonomic classification, no other information should be present in this file. The "normalization" script of the RHEA pipeline (Lagkouvardos et al., 2017) produces a normalized table of this format. The following image shows an example of an abundance table that has the appropriate form.

##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;

- Considering that the Generalized UniFrac distance is used as the default distances metric, 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 give 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 out some extra parameters:

- The first parameter is whether the OTUs or ASVs table is normalized or not. (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 optimal number of clusters for the reference groups. If the user does not know the optimal number of clusters for the reference group can use the "Automated" option and the program will automatically calculate this number based on the Calinski-Harabasz index. (Required)
- The names of the test groups. The user can determine the names of one or more test groups. (Required)
- The optimal number of clusters for these test groups. This is an optional input parameter; if this information is not provided, then the 'De novo clustering analysis' will be omitted. If the user does not know the optimal number of clusters for the test groups can use the "Automated" option and the program will automatically calculate this number based on the Calinski-Harabasz index. (Optional)
- Also, the user has to decide if the most representative point of each cluster will be the medoid, the mean, or the median points. (Optional)
    - The default option is the medoids; in this case, the most representative points will be actual samples of the dataset.
    - If the user chooses the mean or median points option, then the program will select an arbitrary point to be the central point of each cluster.
- The names of the columns of the mapping file the user wishes to statistically analyze against the de novo clusters. The statistical analysis is conducted using the Chi-square test for categorical variables and Wilcoxon Rank Sum test for numerical variables . (Optional)
- The type of the output plots (Boxplots, Pointplots, or violin plots). (Optional)
    - The Boxplots is the default option of the program.
    - The Pointplots is a preferable option when the samples are low in number. In the case of pointplots all the samples are presented in the graph as individual points.
    - The Violinplots is a suitable option when we want to examine the distribution of the values in a more detailed way.

Outputs

The primary outcomes of this script are two reports:

- The first is called the "Distances Based Analysis report". This report aims to inform the user whether the test samples are "close" or "closely related" to the reference group.

The user can find in this report:

- boxplots with the distances from the most representative points,
- MDS plots visualizing the interrelation of the groups,
- tables with various statistical measures like the p-values of the PERMANOVA, PERMDISP, Wilcoxon Rank Sum or Chi-Square test and,
- phylograms based on Ward's minimum variance method.

- The second report is named "De novo clustering report". Its main objective is to reveal the substructural similarities and the existing relations between the groups. Similar to the previously mentioned outputs, "De novo clustering report' produces the following elements:
  - boxplots presenting the distances from the reference groups,
  - MDS plots visualizing the interrelation of the subgroups,
  - tables with various statistical measures like the p-values of the PERMANOVA, PERMDISP, Wilcoxon Rank Sum or Chi-Square test and,
  - tables with descriptive statistic measures.

- All the elements of the reports (plots and tables) are printed in the results folder in .png and .tab format.
- Finally, in the mapping file is added an extra column with information about the de novo clustering of the reference and test groups.

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
The Chi-square analysis is not conducted correctly	➢ Check the spelling of the provided columns
The script cannot perform de novo clustering	➢ Check if the provides dissimilarity matrix has missing samples

References

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

Anderson, M. J. (2006). Distance-based tests for homogeneity of multivariate dispersions. Biometrics 62, 245–253

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300.

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

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 50–60.

Wilcoxon, F. (1992). Individual comparisons by ranking methods. In *Breakthroughs in statistics* (pp. 196–202). Springer.