Serial Group Comparisons Script

Task

Calculate non-parametric ANOVA (Kruskal-Wallis Rank Sum Test) and Fisher tests across **non-paired** input numerical variables over selected categorical variables.

Calculate non-parametric ANOVA for repeated measurement (Friedman Test) and Fisher tests across **paired** (over time) input numerical variables.

Background

A common objective of microbial profiles analysis is the comparison of variables among groups of samples sharing a certain characteristic or treatment in order to detect differences in composition and abundances. This can be determined by performing an Analysis of Variance (ANOVA) type of test to establish, based on the values seen across the groups, how likely it is for the values in those samples to originate from different distributions. As a parametric test, classical ANOVA assumes normality of distribution. Since this is rarely the case for OTU data, we use the non-parametric Kruskal-Wallis Rank Sum Test in Rhea [1]. When more than two groups are compared, pairwise tests are needed to determine which of the groups are significantly different. Again, we use a non-parametric test (Mann-Whitney Test [1]) therefore. The obtained pairwise test significance values are corrected for multiple testing using the Benjamini-Hochberg method [2] and are reported together with the original values. Rhea was designed to perform a systematic testing of all available OTUs or taxonomies in a given experiment. This results in many tests and per extension in a high trade-off for p-values correction. Hence, a reduction of tests can be applied by removing unnecessary tests using e.g. prevalence cutoffs, since it has been shown that pre-filtering datasets increase the power of analysis [3].

To test for significant differences over time an additional script is provided in the folder. Again, the input data is assumed not to be normally distributed, thus a non-parametric test for repeated measurement ANOVA is applied (Friedman test) [4]. Since the test is based on a complete block design it is not allowed to have missing values. Missing values are replaced using the Skillings-mack method or removed from the analysis [5]. Wilcoxon Signed Rank Sum Test is used to determine significant differences between two time points. Calculated p-values are corrected for multiple testing using the Benjamini-Hochberg method.

References

1. Myles Hollander and Douglas A. Wolfe (1973), Nonparametric Statistical Methods. New York: John Wiley & Sons. Pages 115–120

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

3. Bourgon, R., Gentleman, R., & Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. Proceedings of the National Academy of Sciences, 107(21), 9546-9551.

4. Hollander, M. & Douglas A. Wolfe (1973), *Nonparametric Statistical Methods.* New York: John Wiley & Sons. Pages 139–146

5. Chatfield, M. & Mander, A. (2009), The Skillings-Mack test (Friedman test when there are missing data). The Stata Journal 9(2), 229-305.

Input

The input of the script is a tab-delimited text file following the format shown in the picture below. This table is the result of combining the OTU and Taxonomic binning relative abundance tables with the *alpha*-diversity and others meta-variables. The script create_input_table.R can assist in the preparation of the final input table by combining the specified files. In order to reduce redundant tests for taxonomic variables and the associated correction penalty, it is recommended to modify the automatically created file TaxaCombined.tab and only keep the information about taxonomic levels of interest (e.g. Phyla and Families).

input_filename - The name of the table file used as input (e.g. OTUsCombined.tab)

independant_variable_name - Name of the column with the categorical variable (groups) used for the comparison to detect differences among the dependent numerical variables.

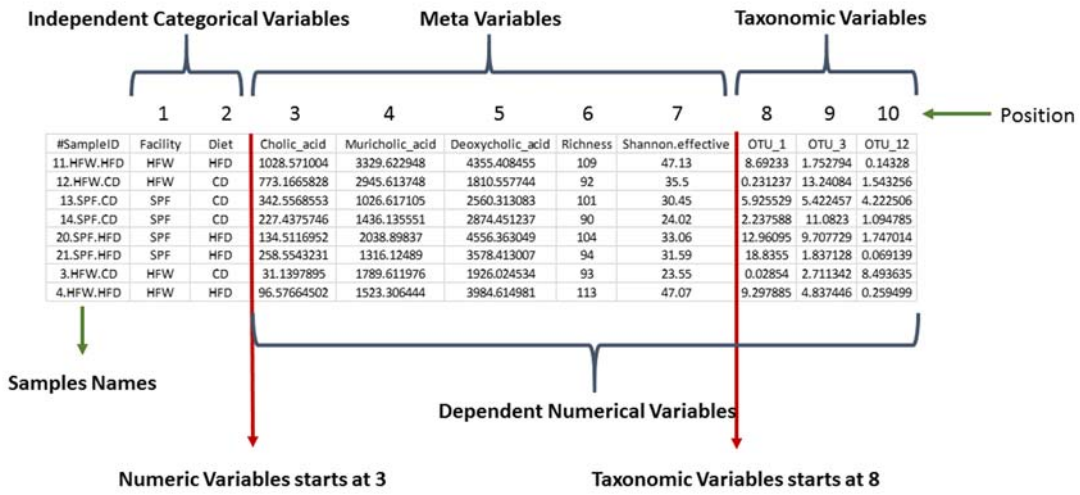
dependant_variables_start - The column where the dependant numerical variables start. The first column containing sample names do not count (see picture below).

taxonomic_variables_start - The column number where the taxonomic variables start. The first column containing sample names do not count (see picture below).

For the analysis over time, it is necessary to include information about the sample ID in the mapping file. Since samples appear more than once within a study one column with sample information is needed.

id_name – The column with the categorical variable (sample ID) used for the comparison of samples over time.

dependant_variable_name – The column with the categorical variable (time) used for the comparisons to detect differences over time.



Options

Several options are available to increase flexibility of the script. These options are set to default values that are most often used in common analysis, but can be changed if required. In more detail:

abundance_cutoff - The minimum relative abundance for taxonomic variables to be considered as effectively present. Variables less than this cut-off are zeroed. This masking of borderline abundances helps focusing the comparisons on samples where the taxonomic variables are important components of the communities. This cut-off should be adjusted to the environment and experiment of interest. The default setting (0.5) is a proposal for explorative studies pertaining to the mouse and human gut microbiota.

prevalence_cutoff - The minimum prevalence (number of samples positive for the given variable) in at least one group in the study. If the variable does not appear in more than the cut-toff in any sample group, it is not considered for statistical testing. The default value is set to 0.3 (30%) prevalence in at least one group.

max_median_cutoff - If the median relative abundance of the taxonomic variable across all groups is less than this cut-off, it is not tested. This filter removes variables that have generally very low abundances across all samples. The default value is 1%.

ReplaceZero - This option determines the treatment of zero abundances in taxonomic variables. Replacing zeros with NA (“YES”) will remove them from the statistical calculations, while selecting “NO” will treat them as true values in all calculations.

ReplaceMissingValues – This option determines the treatment of missing values for the analysis of repeated measurements. Missing values are removed by default (“NO”). If missing values should be replaced by Skillings-Mack method please change this parameter to “YES”.

PlotOption - This option controls the graphical output of the variables showing significant differences across groups. There are three possible choices (1, 2 and 3). Selecting 1 will plot boxplots and violin plots without showing individual data points. Selecting 2 will produce boxplots and violin plots showing individual data points. Selecting 3 will add the samples names over each individual data t point.

Output

Each time the script is executed, the output is placed in a new folder named according to the variable/group used for comparisons. The files included in the output folder are the following:

plot_box.pdf - Boxplots of all significant comparisons (before correction).

plot_point.pdf - Dot-plots of all significant comparisons (before correction).

plot_violin.pdf - Violin plots of all significant comparisons (before correction).

my_analysis_log.txt - A text file capturing all the options used in the analysis for future reference

The remaining files have generic names deriving from the name of the input file and the categorical variable used for analysis. If for example the input file name was “OTUsCombined” and the grouping variable was “Diet”, then the outputs are:

OTUsCombined-Diet-FisherTestAll.tab - A tabular file with the calculated p-value for the Fisher test for all the variables. A column with adjusted p-values for multiple testing is also calculated.

OTUsCombined-Diet-FisherTestPairWise.tab - A tabular file with the calculated p-value for the Fisher test for all the pairs of groups variables. A column with adjusted p-values for multiple testing is also calculated.

OTUsCombined-Diet-modified.txt - The modified input table after all filters and transformations were applied.

OTUsCombined-Diet-pvalues.tab - A tabular file with the calculated p-value for the Kruskal-Wallis Rank Sum test for all the variables. A column with adjusted p-values for multiple testing is also calculated.

OTUsCombined-Diet-sign_pairs.tab - A tabular file with the calculated p-value for the Mann-Whitney test for all the pairs of groups variables. A column with adjusted p-values for multiple testing is also calculated.

Important Notes

Please pay attention to the formatting of your input table and the ordering of the variables (categorical, numerical-nonSeq, numerical-sequence). To determine the position where dependant variable (numerical) start, do not consider the sample names in the first column (1 is the first column after sample names and so on). The decision to use or remove zero and near-to-zero values has important consequences for the results and is left to the discretion of users. Replacing missing values for the analysis over time is sometimes useful and depends on the sample size and the study question. The user should read the cited paper carefully decide about how to handle missing values. Always remember that only variables with significant Fisher or Kruskal-Wallis test are plotted. If no graphical output is produced, it means that not a single variable delivered significant results, which can be also checked by opening the tabular files.

Common problems

- The path to the script is not set correctly
- The input file names are incorrect
- The column name selected for grouping does not exist or contains typos.
- Only one group or too few samples are available for statistics