Correlations Script

Task

Calculate Pearson correlation coefficients across a set of meta- and taxonomic variables and plot the variables that are significantly and strongly correlated.

Background

Combinations of variables that are connected with linear relationships can be detected by calculating their Pearson’s coefficient of correlation [1]. Variables are either meta-variables (i.e. continuous measures of physicochemical variables in the samples of interest, e.g. concentration of target metabolites. In other words, every measurement performed within the study excluding sequencing) or taxonomic variables (relative abundance of OTUs and their assemblage into higher taxonomic levels). The taxonomic variables usually differ from meta-variables by their compositional nature and high sparsity that require special transformations [2]. In Rhea, the centred log-ratio transformation is used to remove the compositional constrains from the taxonomic variables [2]. In addition, taxonomic zeros (relative abundance of taxonomic variables with the value zero) can be treated as missing data and be excluded from the calculation of correlations. Following this transformation of taxonomic variables, the table is centred and scaled, to adjust for differences in the offset and fold changes respectively, and the Pearson correlation for all pairs is calculated [1]. The significance before and after FDR correction [3] is reported together with the number of observations that supports the correlation.

References

1. Pearson, K. (1909). Determination of the coefficient of correlation. Science, 23-25.

2. Aitchison, J. (1986) The statistical analysis of compositional data

3. 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), 289-300.

Input

The expected input file for the script is a tab-delimited file containing all the numerical variables (meta- and taxonomic) across all samples. The first column contains the sample names and the first row the variables names. Categorical variables (non-numerical variables like gender or sample origin) are NOT allowed. An example of the input table format (named “correlation-input-table.tab”) can be found within the Rhea distribution in the folder “/0.Original-Data/Tamplate-Data”. If the Rhea Serial Group Comparisons script was executed before, a copy of the transformed input table is placed in the appropriate folder (6.Correlations) and can be used directly. The script expects the meta-variables to come first in the table followed by the taxonomic variables. The column where the taxonomic variables start need to be set by users themselves in the editable section of the script.

Options

Several options that increase the flexibility of the script are available. These options are set with default values that originate from our own experience in handling sequencing data from the mammalian gut to provide a first exploratory type of analysis. However, they can (or must) be changed on a per-study case. In more detail:

signf_cutoff - A p-value cut-off for reporting a correlation as significant. This is commonly set to 0.05, but other thresholds (e.g. 0.01) can be used.

includeTax - The main purpose of the script is the systematic screening for correlations between meta-variables and taxonomic variables. In order to reduce the number of statistical tests to be performed and the associated correction burden, the default script does not perform tests for correlations in-between taxonomic variables.

includeMeta - As above for correlations in-between meta-variables.

fill_NA - For the meta-variables, it is possible to replace not-available (NA) values with the mean of existing values. This is not done per default, i.e. NA values are ignored.

replace_zeros - For taxonomic variables, zeros can be treated as NA values. This is done to restrict the correlation calculation to cases where the target OTU is actually detected, thus indirectly taking into account the detection limit of the method. Otherwise, zeros are considered as true values and are used for determining correlations of the corresponding OTU with other variables. Replacement of zeros with NA is recommended and is the default setting.

prevalence_exclusion - This is a filter that allows restricting the analysis to those variables represented by at least a certain number of observations/values, thereby avoiding analysis of underpowered variables. Taxonomic variables (e.g. OTUs) that are only present in a minor part of the samples can be considered as none-representative (e.g. an OTU appearing in only 2 of 10 samples in a study). The prevalence exclusion is a subjective cut-off that must be adjusted according to the aim of the study and the number and type of samples. The default threshold for a variable to be included is to be present in at least 30% of samples.

min_pair_support - Since the correlation of two variables is calculated by sample-specific pairs of values, the number of complete pairs is the number of observations that support the correlation. Missing values in the input table can result in incomplete pairs for the given sample and variables. Hence, the number of complete pairs can be low in the case of two variables with high sparsity. The relevance of correlations calculated from 2 or 3 complete pairs, even if significant, is questionable. Default minimum support is set to 4. A larger support can be set for studies with large sample size.

plot_pval_cutoff - The plotting output can be controlled by this option. By default, every correlation with a p-value <0.05 before correction is plotted. This p-value threshold can be adjusted as wished.

plot_corr_cutoff - The plotting output can be controlled by this option. By default, every variables combination characterized by an absolute correlation coefficient >0.5 is plotted. This correlation value threshold can be adjusted as wished.

Output

The script produces two main graphical outputs and several intermediate table files:

corrplot.pdf - A graphical display of all variables combinations in a matrix. Each correlation is depicted as a small circle coloured according to the direction of correlation coefficients (negative, red; positive, blue). The size of the circles is dictated by the uncorrected p-value of the corresponding correlation.

linear_sign_pairs.pdf – A graphical display of those correlations that passed the thresholds defined in the *plot_pval_cutoff* and *plot_corr_cutoff* options. The graph shows the individual sample-specific values, a linear fitted line, and the lower and upper boundaries of the predicted interval (shown as a grey polygon around the fitted line). The boundaries are determined using the R function predict, which produces predict values based on a linear model. A predicted interval accounts for the variability around the mean response inherent in any prediction. It represents the range where a single new observation is likely to fall. Due to the data transformation applied before calculation of correlations (see background information above), there is no scale for the axes. The correlation coefficient, the corrected and original p-values, and the number of supporting pairs of data (observations) are shown in each plot.

correlation-table.tab - A matrix with the correlation coefficients for all variable combinations.

support-table.tab - A matrix with the supporting pairs for all variable combinations

pval-table.tab - A matrix with the p-values for all variable combinations

cutoff-pairs-corr-sign.tab - A tab-delimited table with the names of significant variable combinations, their correlation coefficient, p-values (adjusted and original), and number of supporting pairs.

plotted-pairs-stat.tab - A tab-delimited table with the names of variable combinations that have been selected for plotting (defined by the plot_pval_cutoff and plot_corr_cutoff options), their correlation coefficient, p-values (adjusted and original), and number of supporting pairs.

transformed.tab - The input file after all performed transformations. This is an intermediate file for control of the procedure. Non-available and negative values are expected.

Important Notes

Although the log transformation is expected to linearize the existing correlation relations, it cannot be excluded that the underline relation stays not linear. This mean that a test for linear correlation as that of Pearson can fail to detect all true biological correlations. Furthermore, it is important to stress that the detection of correlation is not at all indicative of causality. Unless coupled with additional experimental evidence, a significant positive or negative correlation alone does not even support the direct association of the two variables considered. Hence, users are strongly encouraged to thoroughly assess the rationale for calculating correlations in their study.

The systematic screening of all possible combinations of many variables can lead to many tests that in turn translate into high costs in terms of correction for multiple testing and removal of false positives. Hence, usage of filters that contribute to enhancing the strength of correlations (e.g. by setting a lower p-value threshold or a higher number of supporting pairs or prevalence limit) and concomitantly to reducing the number of tests performed is recommended. As instructed in the documentation of the Serial Group Comparisons Script, selection of specific taxonomic levels (e.g. phylum, families, and OTUs) for calculating the correlations helps avoiding repetitive testing of redundant data. It is important though to clearly document every variable exclusion and to determine those criteria carefully beforehand. Rounds of data filtering by adjusting criteria around variables that show significant results can be considered as fraudulent. In case of trends, where significance is lost after correction, it is better to report uncorrected p-values (justifiable in the case of an explorative study) rather than trying to bypass the correction penalty via creative filtering.

Common problems

- The path to the script is not set correctly
- The input file names are incorrect
- The format of the input file is incorrect (e.g. categorical values)