Alpha-Diversity Script

Task

Calculate *alpha*-diversity across samples as species richness and effective diversity for popular indices.

Background

The diversity of operational taxonomic units (OTUs) within any given sample is the *alpha*-diversity of that sample. The simplest way of measuring it is the enumeration of OTUs present in that sample, also called species richness. Since the number of observed OTUs is directly correlated to the sequencing effort, a normalization step of the observation table is commonly preceding the calculation of richness-based metrics. This allows for direct comparisons among the samples in a study. Nevertheless, as not all observations in a given amplicon sequencing profile represent true members of the target community (*e.g.* contaminations, artefacts), a filtering before counting can be applied. Proportional filtering offers a good way for minimizing the inflation in richness measurements by such numerous but low abundant observations. A recommended cut-off that can be applied is the sample-wise filtering of observations occurring at a relative abundance of 0.25% in each community. Such filtering is returning only the “effective” portion of the community and is thus referred to as the “Effective Microbial Richness” [1]. For comparisons of *alpha*-diversity, richness does not consider the structure of the community and does not adjust for differential abundance of individual OTUs. There are different indices that capture also the structure of the community rather than enumerating the parts. The two most popular are the Shannon and Simpson diversity indices (the later add more weight to abundance). Those indices are not linear, meaning that a sample with Simpson index of 0.7 is not twice as diverse as a sample with Simpson index of 0.35, making comparisons difficult. A better way of representing true *alpha*-diversity, rather than indices of it, is to calculate effective diversity, as proposed by Lu Jost [2,3]. In short, effective diversity is the number of equally abundant species that would give any value of a given index. In Rhea, we calculate both the Simpson and Shannon indices and their effective numbers. We strongly recommend usage of effective diversities for visualization purposes or for comparisons across samples and studies, as indices are not linear.

References

1. Reitmeier et al. (2020) Handling of spurious sequences affects the outcome of high-throughput 16S rRNA gene amplicon profiling. Submitted

2. Jost L (2006) Entropy and diversity. Oikos 113:363–375

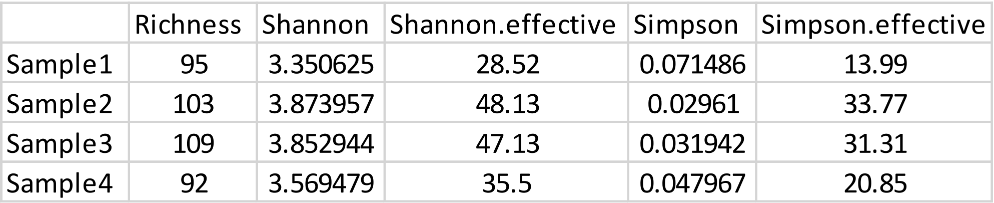
3. Jost L (2007) Partitioning diversity into independent alpha and beta components. Ecology 88:2427–2439

Input

The expected input file for this script is a normalized OTU table. If the Rhea normalization script was used, a copy of the normalized table is placed directly into the alpha-diversity folder and is thus ready for use. It is important to note that the normalized “pseudo” counts of reads in the table are not integers anymore. Although this is not a problem for alpha-diversity indices calculation, for the estimation of richness a cut-off is necessary before assessing the number of OTUs in every sample. Values below 0.5 counts practically mean that those OTUs were only detectable due to the differential depth of sequencing and that in the normalized sampling size they would be probably absent. Therefore, we only count normalized counts that are above 0.5 for estimation of species richness.

Output

The output of this script is a tab-delimited text file with the calculated values for different alpha-diversity measures and indices across samples. If the structure of Rhea folders is preserved, the alpha-diversity output file is directly copied in the folder “Serial Group Comparisons”, where differences in alpha-diversity between different groups of samples can be tested. A typical output file looks as follows:



	Richness	Shannon	Shannon.effective	Simpson	Simpson.effective
Sample1	95	3.350625	28.52	0.071486	13.99
Sample2	103	3.873957	48.13	0.02961	33.77
Sample3	109	3.852944	47.13	0.031942	31.31
Sample4	92	3.569479	35.5	0.047967	20.85

Important Notes

Since only richness and the effective number of species for Shannon and Simpson indexes are meaningful for comparisons, it makes sense to modify the table and delete the columns for the two indexes prior to statistical testing and graphical representation. Nevertheless, the two indexes are provided in the raw output file, as they are common parameters usually determined in studies of alpha-diversity. Furthermore, since the two effective numbers are capturing the same information but only with different weights on the abundance of taxa, it is recommended to choose only one in order to reduce the number of statistical tests performed and the associated cost of correction. We usually report the Shannon effective number of species as a balanced solution to the simple enumeration of richness and the strong weight towards abundance for Simpson.

Common problems

- The path to the script is not set correctly
- The input file is not normalized
- The input file is of different format (e.g. has a taxonomic classification column)