

Alpha-Diversity Script

Task

Calculate alpha-diversity for each sample of a study as species richness and effective diversity for most popular indices.

Background

The diversity of operational taxonomic units (OTUs) within a 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. Species 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). These 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 [1,2]. In short, the effective diversity of a microbial profile for a certain index is the number of equally abundant species that would give the same value for that 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, as indices are not linear and can mislead data interpretation.

References

1. Jost L (2006) Entropy and diversity. *Oikos* 113:363–375
2. 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 counts of reads in the table are not integers anymore. To avoid incorrect estimation of species richness due to differential sequencing depth, we only consider normalized counts that are above 0.5.

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 indices prior to statistical testing and graphical representation. Nevertheless, the two indexes are provided in the raw output file for clarity and for reference since they are commonly reported during 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 high abundance for the Simpson index.

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)