# OTU-Table Normalization Script

## Task

Compensate for differential sequencing depth between samples by calculating relative abundances and create normalized counts.

## Background

Normalization is the process of data transformation to remove the effect of differential sampling size. Because high-throughput sequencing results in different number of sequences per samples, a normalization of the read counts is required prior to downstream analysis. This is commonly performed by a procedure called rarefying, a random sub-sampling of reads from each sample to a fixed total, usually the least count among the samples. The process is repeated several times and a mean number of sequences is calculated for each OTU within the given sample. These values are then rounded and represent the final normalized counts. Rarefaction, although very popular among ecologists and microbial ecologists, has been criticized for the following reasons: (i) omission of available valid data, (ii) the estimation of overdispersion is more difficult due to data loss, (iii) loss of power (type II error), (iv) dependence on an arbitrary threshold, (v)    additional uncertainty due to the randomness in rarefaction [1]. The authors of the latter publication stated that even a simple normalization to proportions is less biased as it includes no random steps and minimal loss of information. Their suggested normalization consisted of a variance stabilization transformation (logarithmic). In Rhea, the issue with this kind of transformation is the incompatibility with some of the downstream analytical functions requiring counts or proportions. Plotting of relative abundances across groups expressed in percentages, with its known limitations, gives researchers a more intuitive understanding of biological phenomenon. Hence, in Rhea, counts are by standard normalized via simple division to their sample size and then multiplication by the size of the smaller sample. This approach has not the downside of introducing random variance or loss of data. Nevertheless, we provide to users the option to proceed with classical rarefaction for normalized counts if wanted.

## References

1. McMurdie, P. J., & Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol*, *10*(4), e1003531.

# Input

The expected input file for the script is a standard OTU-table (as tab-delimited text format) created by programs such as IMNGS (www.imngs.org). OTU names are in the first column followed by the count of reads assigned to each OTU cluster for each sample, with the sample name being the header of each column. The last column of the table must be named "taxonomy" and contains the taxonomic classification of each OTU. For compliance with downstream analysis, the taxonomy string for each OTU must be delimited by semicolons, with exactly 6 fields (left empty if not known).

| #OTUId | Sample1 | Sample2 | Sample3 | Sample4 | taxonomy |
|--------|---------|---------|---------|---------|----------|
| OTU_1 | 1871 | 1820 | 2745 | 4952 | Bacteria;Firmicutes;Clostridia;Clostridiales;Lachnospiraceae;; |
| OTU_2 | 2414 | 367 | 2056 | 1215 | Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Rikenellaceae;Alistipes; |
| OTU_3 | 236 | 269 | 88 | 244 | Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;; |
| OTU_4 | 224 | 432 | 196 | 318 | Bacteria;Firmicutes;Clostridia;Clostridiales;Lachnospiraceae;Clostridium XlVb; |
| OTU_5 | 376 | 68 | 144 | 27 | Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Porphyromonadaceae;; |
| OTU_6 | 1304 | 25 | 362 | 339 | Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Porphyromonadaceae;Parabacteroides; |
| OTU_7 | 209 | 804 | 36 | 18 | Bacteria;Proteobacteria;Deltaproteobacteria;Desulfovibrionales;Desulfovibrionaceae;; |
| OTU_8 | 270 | 108 | 152 | 44 | Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Porphyromonadaceae;; |
| OTU_9 | 192 | 24 | 93 | 147 | Bacteria;Proteobacteria;Deltaproteobacteria;Desulfovibrionales;Desulfovibrionaceae;; |
| OTU_10 | 322 | 53 | 1071 | 573 | Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides; |

# Output

The script produces 4 output files containing either relative abundances with or without taxonomy, or normalized counts calculated from the relative abundances with or without taxonomy. If the folder structure of Rhea is preserved, copies of the output files are placed wherever needed for downstream analyses using the appropriate other scripts. For the template input table shown above, the output tables would be:

Relative abundance

| | Sample1 | Sample2 | Sample3 | Sample4 |
|--------|---------|---------|---------|---------|
| OTU_1 | 25.22243 | 45.84383 | 39.53622 | 62.86657 |
| OTU_2 | 32.54246 | 9.244332 | 29.61256 | 15.42465 |
| OTU_3 | 3.181451 | 6.775819 | 1.267464 | 3.097626 |
| OTU_4 | 3.019682 | 10.88161 | 2.822987 | 4.03707 |
| OTU_5 | 5.068752 | 1.712846 | 2.074031 | 0.34277 |
| OTU_6 | 17.57886 | 0.629723 | 5.213884 | 4.303669 |
| OTU_7 | 2.817471 | 20.25189 | 0.518508 | 0.228513 |
| OTU_8 | 3.639795 | 2.720403 | 2.189255 | 0.558588 |
| OTU_9 | 2.588299 | 0.604534 | 1.339479 | 1.866193 |
| OTU_10 | 4.340793 | 1.335013 | 15.42561 | 7.274343 |

Relative abundance with taxonomy

| | Sample1 | Sample2 | Sample3 | Sample4 | taxonomy |
|--------|---------|---------|---------|---------|----------|
| OTU_1 | 25.22243 | 45.84383 | 39.53622 | 62.86657 | Bacteria;Firmicutes;Clostridia;Clostridiales;Lachnospiraceae;; |
| OTU_2 | 32.54246 | 9.244332 | 29.61256 | 15.42465 | Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Rikenellaceae;Alistipes; |
| OTU_3 | 3.181451 | 6.775819 | 1.267464 | 3.097626 | Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;; |
| OTU_4 | 3.019682 | 10.88161 | 2.822987 | 4.03707 | Bacteria;Firmicutes;Clostridia;Clostridiales;Lachnospiraceae;Clostridium XlVb; |
| OTU_5 | 5.068752 | 1.712846 | 2.074031 | 0.34277 | Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Porphyromonadaceae;; |
| OTU_6 | 17.57886 | 0.629723 | 5.213884 | 4.303669 | Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Porphyromonadaceae;Parabacteroides; |
| OTU_7 | 2.817471 | 20.25189 | 0.518508 | 0.228513 | Bacteria;Proteobacteria;Deltaproteobacteria;Desulfovibrionales;Desulfovibrionaceae;; |
| OTU_8 | 3.639795 | 2.720403 | 2.189255 | 0.558588 | Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Porphyromonadaceae;; |
| OTU_9 | 2.588299 | 0.604534 | 1.339479 | 1.866193 | Bacteria;Proteobacteria;Deltaproteobacteria;Desulfovibrionales;Desulfovibrionaceae;; |
| OTU_10 | 4.340793 | 1.335013 | 15.42561 | 7.274343 | Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides; |

Normalized counts

|        | Sample1  | Sample2 | Sample3  | Sample4  |
|--------|----------|---------|----------|----------|
| OTU_1  | 1001.331 | 1820    | 1569.588 | 2495.803 |
| OTU_2  | 1291.936 | 367     | 1175.619 | 612.3588 |
| OTU_3  | 126.3036 | 269     | 50.31831 | 122.9758 |
| OTU_4  | 119.8814 | 432     | 112.0726 | 160.2717 |
| OTU_5  | 201.2294 | 68      | 82.33905 | 13.60797 |
| OTU_6  | 697.8808 | 25      | 206.9912 | 170.8557 |
| OTU_7  | 111.8536 | 804     | 20.58476 | 9.071982 |
| OTU_8  | 144.4999 | 108     | 86.91344 | 22.17596 |
| OTU_9  | 102.7555 | 24      | 53.1773  | 74.08785 |
| OTU_10 | 172.3295 | 53      | 612.3967 | 288.7914 |

Normalized counts with taxonomy

|        | Sample1  | Sample2 | Sample3  | Sample4  | taxonomy |
|--------|----------|---------|----------|----------|----------|
| OTU_1  | 1001.331 | 1820    | 1569.588 | 2495.803 | Bacteria;Firmicutes;Clostridia;Clostridiales;Lachnospiraceae;; |
| OTU_2  | 1291.936 | 367     | 1175.619 | 612.3588 | Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Rikenellaceae;Alistipes; |
| OTU_3  | 126.3036 | 269     | 50.31831 | 122.9758 | Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;; |
| OTU_4  | 119.8814 | 432     | 112.0726 | 160.2717 | Bacteria;Firmicutes;Clostridia;Clostridiales;Lachnospiraceae;Clostridium XlVb; |
| OTU_5  | 201.2294 | 68      | 82.33905 | 13.60797 | Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Porphyromonadaceae;; |
| OTU_6  | 697.8808 | 25      | 206.9912 | 170.8557 | Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Porphyromonadaceae;Parabacteroides; |
| OTU_7  | 111.8536 | 804     | 20.58476 | 9.071982 | Bacteria;Proteobacteria;Deltaproteobacteria;Desulfovibrionales;Desulfovibrionaceae;; |
| OTU_8  | 144.4999 | 108     | 86.91344 | 22.17596 | Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Porphyromonadaceae;; |
| OTU_9  | 102.7555 | 24      | 53.1773  | 74.08785 | Bacteria;Proteobacteria;Deltaproteobacteria;Desulfovibrionales;Desulfovibrionaceae;; |
| OTU_10 | 172.3295 | 53      | 612.3967 | 288.7914 | Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides; |

# Important Notes

Normalization is very sensitive to grossly different sample sizes. If for example a negative control sample is included in the OTU-table, all the samples would be normalized to the total number of reads in this control sample (which is likely very low). This would result in grossly wrong normalized counts and ensuing estimations of *alpha-* and *beta*-diversity. To avoid such errors, samples characterized by total read counts that fall out of the range of the majority of samples should be removed. These not only relates to negative controls, but also to samples to which low number of sequences were assigned due to various technical reasons. It is expected that users have a very close look at original input tables and identify samples to be removed. If IMNGS is used for generation of the OTU-table, re-processing of the raw reads using a modified mapping file that includes only the selected sample-barcode combinations is required.

# Common problems

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