Preparing Input Files

Required Files

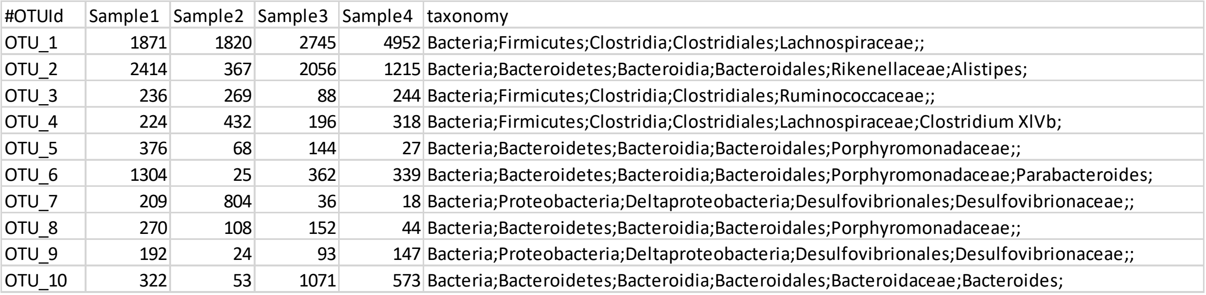
There are three files that are needed in order to finish the complete set of Rhea scripts.

1. An OTU table
2. A phylogenetic tree of representative OTU sequences
3. A Mapping file containing the grouping of samples

Here we summarize the format requirements for these files. More details about their format can be found on the respective script ReadMe file.

The OTU table

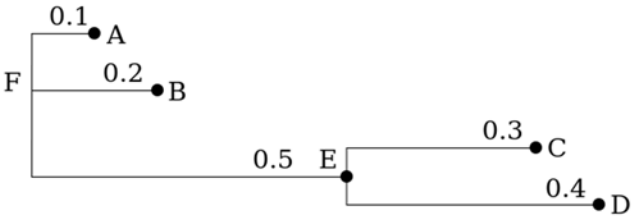
The OTU table is a contingency table with the number of reads assigned to each OTU cluster for each sample in a sequencing experiment. OTU names are in the first column followed by the count of reads assigned to each OTU cluster for each sample, with the sample name being the header of each column. In Rhea, it is important that a column named taxonomy is present at the end carrying the taxonomic classification of each OTU. For compliance with downstream analyses, the taxonomy string for each OTU must be delimited by semicolons, with exactly 6 fields (left empty if not known). Non-alphanumeric characters, such as ()?/{}#”, may lead to unexpected behaviour and should be avoided. One example of an OTU table can be seen below.



#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;

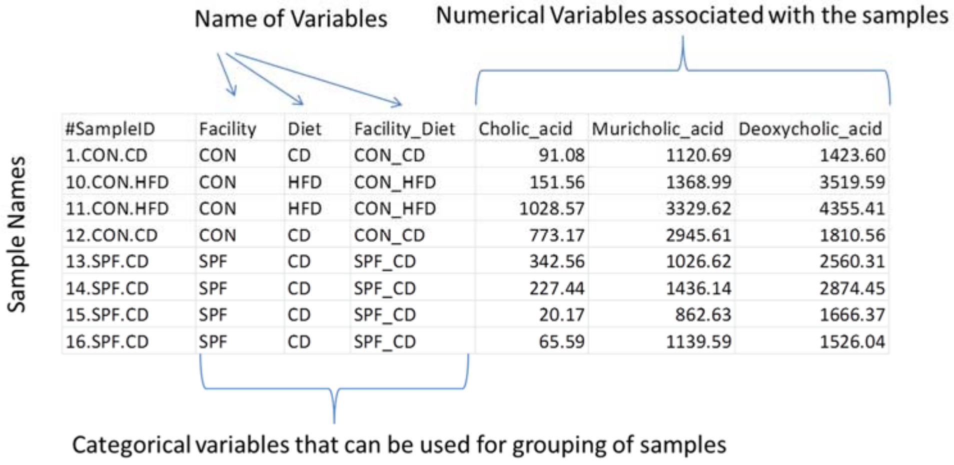
The OTU tree

A rooted phylogenetic tree of representative OTU sequences in Newick format. This format is very common and most phylogenetic programs support it. The tree bellow for example can be represented as: `(A:0.1,B:0.2,(C:0.3,D:0.4):0.5);`. It is important that the names of sequences used for the tree match those of the OTU file.



The mapping file

The most basic format of this tab-delimited file contains two columns (with headers on the first line): the first one with sample names and the second one with a categorical variable used to separate the samples into groups. Additional columns with different groupings can be added in a series of columns. Metadata (additional numerical measurements from the samples) can be added after to be used in statistical and correlation analysis parts. An example of a mapping file can be seen below.


Name of Variables
Numerical Variables associated with the samples
Sample Names
Categorical variables that can be used for grouping of samples

#SampleID	Facility	Diet	Facility_Diet	Cholic_acid	Muricholic_acid	Deoxycholic_acid
1.CON.CD	CON	CD	CON_CD	91.08	1120.69	1423.60
10.CON.HFD	CON	HFD	CON_HFD	151.56	1368.99	3519.59
11.CON.HFD	CON	HFD	CON_HFD	1028.57	3329.62	4355.41
12.CON.CD	CON	CD	CON_CD	773.17	2945.61	1810.56
13.SPF.CD	SPF	CD	SPF_CD	342.56	1026.62	2560.31
14.SPF.CD	SPF	CD	SPF_CD	227.44	1436.14	2874.45
15.SPF.CD	SPF	CD	SPF_CD	20.17	862.63	1666.37
16.SPF.CD	SPF	CD	SPF_CD	65.59	1139.59	1526.04

Staring from IMNGS

Rhea was built to provide maximal compatibility with the online sequence processing pipeline of the IMNGS platform, to satisfy the needs of non-expert users. Indeed, coupled use of Rhea with IMNGS, which offers a user-friendly installation-free environment to analyse raw sequence data, a full analytical pipeline is available. The output files of IMNGS can be directly used as input in Rhea, without any modification. IMNGS is an implementation of UPARSE, an analytical pipeline using USEARCH8. Starting from multiplexed reads and indexes, it first demultiplexes files to samples, merge paired reads if present, filter by expected error, and trim sides according to user selections. Then, UPARSE clustering is performed, including *de novo* and reference chimera removal of OTUs. Detection of non-16S rRNA gene sequences is also applied using SortMeRNA, and the set of clean OTUs is taxonomically annotated with a local RDP classifier. IMNGS offers removal of low abundance OTUs across samples by applying relative abundance cut-offs, which is less sensitive to differential sequencing depth among samples than removal of singletons or other count-based filters. IMNGS also prepares a rooted tree of representative OTU sequences using MUSCLE for alignment and FastTree for Maximum Likelihood approximation. The metafile is not provided by IMNGS and users need to prepare one following the examples above.

Starting from USEARCH

USEARCH is a software platform that focuses on the processing of raw sequences down to OTU tables. The formatting of standard outputs needs minimal modification to be compatible with the Rhea scripts. Assuming someone follows the UPARSE approach, the standard output is an OTU table without taxonomic assignments and a file with the clusters representatives (most abundant sequences) in FASTA format.

The first step needed is to add taxonomy to the OTU table. To do so, the online RDP classifier and/or SILVA aligner with classification is recommended. In any cases, the taxonomy has to be formatted to be compatible with Rhea: lineages must be delimited by semicolons, with exactly 6 fields (left empty if not known). A column with this taxonomic information can be added as the last column of an OTU table obtained using USEARCH, ensuring that the order of OTUs is the same (i.e. ordering according to OTU names and manual inspection before merging is strongly recommended).

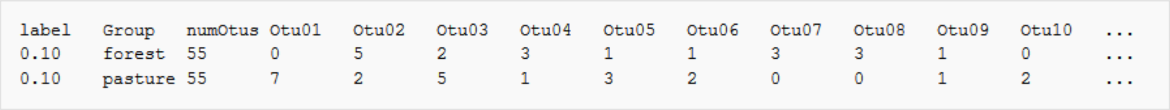
The tree can be created from the FASTA format provided by USEARCH using many phylogenetic programs. We recommend using MEGA7 for its user-friendly interface and wealth of options. Importing the OTU sequences, aligning them, and performing a phylogenetic analysis in that software is relatively straightforward. Although there is no optimal tree construction algorithm for universal use, we recommend applying a Maximum likelihood approach with 100 bootstraps. Note that the tree must be rooted and can then be exported as Newick to produce a file compatible with the beta-diversity script in Rhea.

The Metafile need to be created and provided by the user based on the directions highlighted above.

Starting from mothur

Mothur is a versatile software platform for microbial ecology. Among other applications, it offers a complete command line pipeline for the analysis of microbial sequence data. If users of mothur want to continue with performing post-OTU table processing in Rhea, a few additional steps are needed.

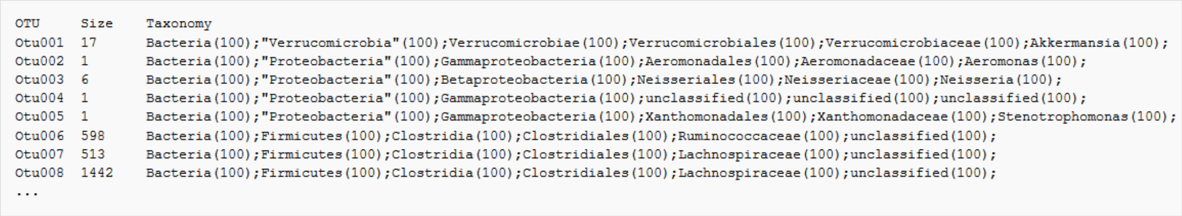
Assuming the command `make.shared` is used, an OTU table resembling that required in Rhea is produced (see below).



```
label   Group   numOtus Otu01   Otu02   Otu03   Otu04   Otu05   Otu06   Otu07   Otu08   Otu09   Otu10   ...
0.10    forest  55      0       5       2       3       1       1       3       3       1       0       ...
0.10    pasture 55      7       2       5       1       3       2       0       0       1       2       ...
```

Please remove the columns “label” and “numOtus”, and transpose the table so that OTUs are rows and samples (Group in the example above) are columns. In the example above the clustering was done using 10% sequence dissimilarity, but this is usually set to 3% for microbiome studies.

The next step is to add taxonomy to the table. That can be achieved by using the approach described above for USEARCH, or by using the mothur command `classify.otu` to extract consensus taxonomy for each cluster. Then the format need to be modified as above to comply with requirements in Rhea. A typical command output can be seen below.



```
OTU     Size    Taxonomy
Otu001  17      Bacteria(100);"Verrucomicrobia"(100);Verrucomicrobiae(100);Verrucomicrobiales(100);Verrucomicrobiaceae(100);Akkermansia(100);
Otu002  1       Bacteria(100);"Proteobacteria"(100);Gammaproteobacteria(100);Aeromonadales(100);Aeromonadaceae(100);Aeromonas(100);
Otu003  6       Bacteria(100);"Proteobacteria"(100);Betaproteobacteria(100);Neisseriales(100);Neisseriaceae(100);Neisseria(100);
Otu004  1       Bacteria(100);"Proteobacteria"(100);Gammaproteobacteria(100);unclassified(100);unclassified(100);unclassified(100);
Otu005  1       Bacteria(100);"Proteobacteria"(100);Gammaproteobacteria(100);Xanthomonadales(100);Xanthomonadaceae(100);Stenotrophomonas(100);
Otu006  598     Bacteria(100);Firmicutes(100);Clostridia(100);Clostridiales(100);Ruminococcaceae(100);unclassified(100);
Otu007  513     Bacteria(100);Firmicutes(100);Clostridia(100);Clostridiales(100);Lachnospiraceae(100);unclassified(100);
Otu008  1442    Bacteria(100);Firmicutes(100);Clostridia(100);Clostridiales(100);Lachnospiraceae(100);unclassified(100);
...
```

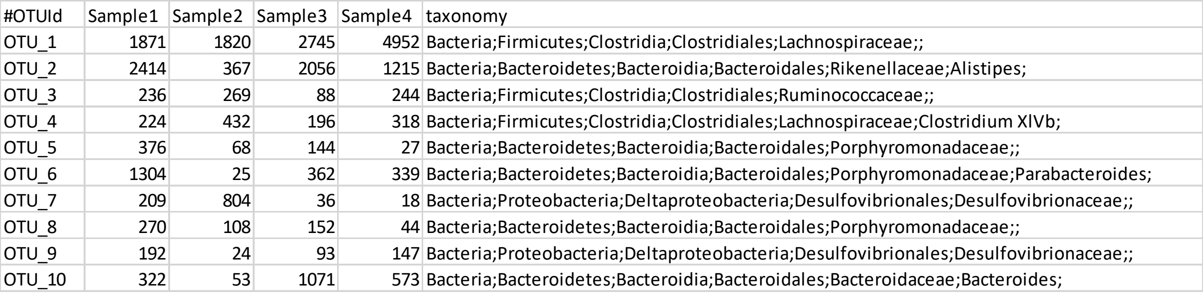
To make the classification compatible with Rhea input format, the taxonomy line should be reformatted. The marks of confidence “(100)“, the quotations marks, and the word unclassified must be removed. For example, the line:

Bacteria(100);"Proteobacteria"(100);Gammaproteobacteria(100);unclassified(100);unclassified(100);unclassified(100)

Should become:

Bacteria;Proteobacteria;Gammaproteobacteria;;;;

Both the transposed OTU table and the taxonomy file must then be sorted by OTU names. The latter must be appended to the former, making sure that OTU IDs are compatible. Delete the OTU name and size columns copied/pasted from the taxonomy file and change the Taxonomy header to taxonomy. Eventually, the final OTU table should look like this:



#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;

The tree can be calculated as described above for UPARSE using the clusters centroid sequences.

Important Notes

It is recommended to double-check taxonomic assignment of OTUs using different classification approaches, i.e. it is possible to combine the classification of several taxonomic assignment programs (i.e. RDP and SILVA). Thereby, since different systems do not follow the same nomenclature, it is important to stick to one of them whenever they disagree. IMNGS uses RDP training set 15 and it is not automatically updated for compatibility reasons with the SRA processed files. If samples where processed with IMNGS and there are many newer releases of RDP training sets, it might be beneficial to run the online RDP classifier to get the newest classification and replace that in the IMNGS output.

Common problems

- The names of the samples in OTU table and meta files do not match
- The format of the taxonomy is incorrect (number of semicolons)
- The names in the tree and the OTU table do not match
- The tree is not rooted