TaxonTableTools (v 1.0) user manual

1. Introduction

The sequencing of DNA metabarcoding data has drastically increased over the past decade and many datasets are produced quickly nowadays. The analysis of this massive amounts of data and their translation into biological meaningful facts is often limited especially for non-experts and bioinformatics beginners. However, it is the biologists that need to work with the data and interpret these. This was the rationale for developing TaxonTableTools (TTT is the following) as part of the GeDNA project (eDNA metabarcoding in regulatory biomonitoring in Germany). The program aims to provide easy-to-use tools for biologists and non-bioinformaticians to analyse and visualize their data quickly and reproducible via a graphical user interface. Thus, the dependency on self-written R or python scripts, which can cause confusion and errors particularly when working with different datasets, is reduced. TaxonTableTools is not aiming to replace those scripts (since a specific script always outperforms a general script), but rather provide tools to quickly assess data and generate information as basis for further, more dataset-specific analyses. Furthermore, the quick data visualization integrated in TTT always comes in handy for presenting first preliminary results of a dataset.

Of course, the software is evolving and there will be bugs and issues at few points. If so, please leave the report in the git repository or drop me an email.

2. Installation

2.1 Requirements

- python3.6 or python3.7 (Tutorial is shown for 3.6, but works as well for 3.7)
- Optional: Krona tools (https://github.com/marbl/Krona/wiki)

  Krona allows hierarchical data, such as taxonomic information, to be explored with zooming, multi-layered pie charts. These interactive charts can be automatically generated with TaXonTableTools. This function requires Krona tools to be installed (it's currently not officially supported on Windows).
- The GUI in some cases only works on an HD screen (1920 x 1080 pixel)

Terminal usage

- Ubuntu/Mac: Terminal
- Windows: Power shell terminal

$ = command to type

>> = expected output of the command

2.2 MacOS, Ubuntu and Windows

Install python3.6 or 3.7 on your machine (https://www.python.org)

Install pip on your machine (https://pypi.org/project/pip/)

First, make sure you run the correct pip version via:

```
$ pip3 --version
```

Which should return python3.6 or 3.7. Otherwise specify your pip using pip3.6 or pip3.7

Then install TaxonTableTools via pip:

```
$ pip3 install taxontabletools
```

TaxonTableTools can then be started via:

```
$ python3 -m taxontabletools
```

Updates can be installed via:

```
$ pip3 install --upgrade taxontabletools
```

Note: Windows user with notebooks reported issues with the size of the TaxonTableTools window. Since the window size is fixed, the only current solution is to set the element size to 100% (which is often set to ~150% on notebooks).

2.5 Krona tools

Install Krona tools (https://github.com/marbl/Krona/wiki).

Check your Krona tools installation by running:

$ ktImportText

>> KronaTools 2.7.1 - ktImportText

3. Getting started

3.1 Projects

Start TaxonTableTools by calling it from the command line:

$ python3 -m taxontabletools

When first launched TTT will ask to define an output directory. This is where all your projects and respective output files will be stored. A new folder "Projects" will be created in this directory.

The starting screen appears and will ask for creating a new project or load an existing project.

> Type in your a project name.
or
> Load an existing project folder.
or
> Leave blank to create a “Default_project” folder.

A new project folder has been created! You can find your results in the respective directories under “YOUR_CHOSEN_PATH/Projects/YOUR_PROJECT/”.

3.2 Sample IDs

Note: The sample ID feature is optional and if the prompts are not met it will simply be ignored. However, the replicate combination tool requires the last underscore element to be the indicator for the replicates. The indicator combination can be chosen by the user.

TaxonTableTools was built to read a specific sample ID format to increase the accessibility of some downstream analyses. Five different categories can be read from this format. Thus, a number of 5 underscores (which delimit the information) is required. The format looks as follows.

Project_Site_Sample_Date_Replicate

(e.g. Dessau_MuldeOH_5A_180419_a)

Project = name of the project

Site= name of the site

Sample= sample ID

Date = date when samples were taken

Replicate = e.g. "_a" and "_b"

>> see 4.2.1 for more details

3.3 Plots

Occasionally the labels will not be displayed correctly, or colors might not please the user. Therefore, all plots created with TaxonTableTools will be saved in pdf format and are compatible for any downstream adjustments with most common image manipulation software. Since the plots are vector-based graphics, colors can easily be adjusted, labels renamed and vectors be moved. TTT was written to create plots in a general manner, which can subsequently be individualized by the user.

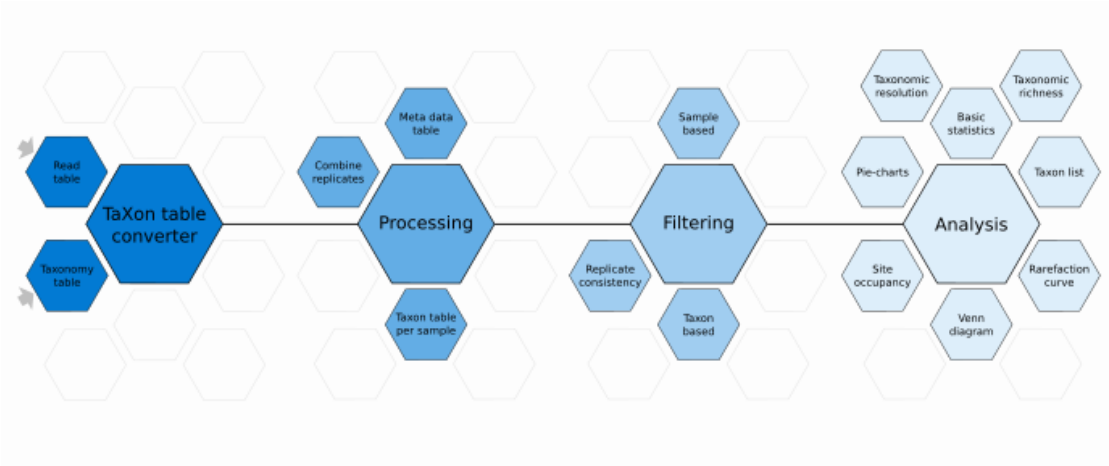
3.4 Directories

All output files will be written to a specific folder within the projects folder. Missing directories will be generated during the start of TaXonTableTools. Do not delete folders while running TTT, as this will result in a crash. Note that there is currently no warning to overwrite files.

3.5 Crashes and troubleshooting

TaxonTableTools does its best to catch all possible errors and provide a solution to the occurring problem. Nevertheless, it is impossible to catch every user error and directly provide a solution to any problem. Thus, please report unrecorded issues that were not caught by TTT.

4. Tools



4.1 Data conversion

4.1.1 Taxon table converter

TaxonTableTools requires two files as input format: a taxonomy table and a read table, which are then merged to a single file. The demanded file format for both tables is xlsx (Microsoft Excel 2007). All downstream tables will also be saved in a xlsx file, which can easily be opened with Excel or any equivalent spreadsheet software.

<u>Read table format:</u>

Read tables can be generated by many published DNA metabarcoding pipelines and wrappers, e.g. JAMP (https://github.com/VascoElbrecht/JAMP), DADA2 (Callahan et al, 2016), QUIME (Caporase et al, 2010) or Obitools (Boyer et al, 2016).

The format requires the first column to be the OTU names (“IDs”), followed by the included samples (see 3.2 for more information) and the sequence as last column. When using read tables produced with JAMP notice to remove the last row “below 0.01p”, the “sort” column and rename “IDs” to “OTUs” and “sequ” to “Sequences”.

IDs	Sample_1	Sample_2	(...)	Sample_n	Sequences
OTU_1	88817	56644	...	67544	ATGCTAA...
OTU_2	6384	18919	...	21877	ATGGTAT...
OTU_3	655	0	...	0	ATGCTTT...
(...)	...	...	...	...	...
OTU_n	0	73	...	87	ATGCTAG...

Taxonomy table format:

The taxonomy table can be created manually by the user or e.g. by using the BOLDigger tool (Buchner et al., 2020), which by default creates matching output files. The standard xlsx file format is required. The format requires the first column to be the OTU IDs ("IDs"), followed by taxonomy ("Phylum", "Class", "Order", "Family", "Genus" and "Species"), the similarity to the reference sequence ("Similarity") and the status of the reference sequence ("Status"). The excel sheet must be named can either be named "First hit", "JAMP results" or "Boldigger results" (More options will be added in the future). Species names are recommended to be written as two words, including genus and epithet to avoid epithet duplicates. The status is automatically derived by the BOLDigger tool but can manually be added as e.g. "public". Same accounts for the similarity.

IDs	Phylum	Class	Order	Family	Genus	Species	Similarity	Status
OTU_1	...	...	...	...	...	Genus epithet	100	public
OTU_2	...	...	...	...	...	...	100	public
OTU_3	...	...	...	...	...	...	99	public
(...)	...	...	...	...	...	...	...	...
OTU_n	...	...	...	...	...	...	98	public
JAMP results								

Merging the read table and the taxonomy table creates a new file in the so called "TaXon table" format, which is also an xlsx file. The two capital letters ("T", and "X") highlights its specific format requirement to differentiate from the general term taxon table. This table format consists of all information of both the taxonomy table and the read table and will be used as standard input format for all subsequent steps. The newly created TaXon tables are found in the "Projects/your_project/TaXon_tables" directory.

4.2 Table processing

4.2.1 Replicate suffixes

TaxonTableTools reads user-defined replicate suffixes to perform the replicate consistency tool (4.2.2) and the replicate merging tool (4.2.3). Replicate suffixes are required to be a comma separated list of characters (e.g. a,b,c,d). The preview button will show an example of what the samples are expected to look like.

4.2.2 Replicate consistency

This tool is used to remove all OTUs (i.e. set the abundance to 0) that are not present in all replicates of one sample. Thus, the replicates will be made consistent.

4.2.3 Replicate merging

Merging the replicates will calculate the sum of reads for each OTU of all given replicates of one sample and will write them into a new column, replacing the individual replicates. The replicated sample will be renamed to "_comb". Samples without a replicate prompt will be skipped (i.e. negative controls).

4.2.4 TaXon table per sample

For downstream analyses it can be handy to create a separate table for each sample in one file. OTUs that are not present in this sample will be dropped (i.e. the table will be shorted in most cases).

4.2.5 Taxon-based filter

The taxon-based filter allows to create specific subsets of a given table, by dropping all OTUs that account to the chosen taxa. The user can select a taxonomic level that will be used as filter mask. It is recommended to filter on higher levels first (e.g. phylum or class) and use the newly created tables to filter on e.g. genus or species level. Otherwise the filter mask will easily become confusing with larger data sets. The check marks option is available. Taxa with a check mark will be excluded in the new table. To exclude a larger number of taxa, use the option "all". To only exclude a couple of taxa and keep most of the original table, use the option "none". Additionally, the table can be filtered according to a given similarity threshold.

Example: To create a table that only includes OTUs that were identified as insects with a similarity over 90%, choose the "class" tab, set the check marks to "all", press "Run", remove the check mark from "Insecta", set the similarity threshold" to "90" and type in e.g. "insecta" as appendix of the new file.

4.2.6 Sample-based filter

The sample-based filter functions similarly to the previously described taxon-based filter. All available samples will be presented in a filter mask. Samples with a check mark will be excluded in the new file. OTUs that subsequently not represented in any of the remaining samples anymore will be dropped.

Example: Initially, this tool can be used to create a main table without e.g. negative controls, that will be used as a basis for the downstream analyses. Furthermore, it can be used to create analysis relevant subsets of the main table to e.g. compare them in a venn diagram.

4.2.7 Metadata table

The metadata table will be automatically created from the sample names (see 3.2 for more information). The metadata information is by default required separated by an underscore ("_").

Example of a newly created metadata table, with a given sample named "ProjectA_No1_RiverX_RegionY"

Raw output table:

1	2	3	4
ProjectA	No1	RiverX	RegionY

Please rename the columns accordingly in the metadata Excel file. Metadata named as numbers (thus also the default) will cause an error. Renaming the metadata columns will enhance the recognition when choosing a metadata for downstream analysis. Each value can be adjusted to the users need before loading the metadata table.

User defined names:

Project	Sample	Site	Region
ProjectA	No1	RiverX	RegionY

4.3 Analyses

4.3.1 Basic statistics

Executing the basic statistics tool will create an overview of the basic read and taxonomy table statistics that can be gathered from the TaXon table. The overall number of samples and OTUs is extracted along the number of taxa per taxonomic level (from Phylum to Species level). All available database states of the reference sequence are counted. The results are printed on the screen and written to a xlsx file in

the “Projects/your_project/Basic_stats” directory. The minimum, average and maximum length (in bp) of the sequences, as well as the average and total number of reads for each sample will also be calculated and written to the file (but not printed on screen).

4.3.2 Taxonomic resolution

Taxonomy resolution highly depends on the used reference database. OTUs with low similarity towards the reference database are recommended to be reported at a higher taxonomic level to prevent false positive results. Hence, this tool first plots the taxonomic resolution per taxonomic level (Phylum to Species), by counting the number of OTUs that are reported for the respective taxonomic level. In a second plot the total number OTUs per taxonomic level is given. The results are written to a pdf file in the “Projects/your_project/Taxonomic_resolution_plots” directory. The plot layout is customizable with the options to change the width and height and the font size.

4.3.3 Taxonomic richness

This tool will plot the number of taxa per taxonomic level (phylum to species) and usually shows a steep increase towards the lower taxonomy. The results are written to a pdf file in the “Projects/your_project/Taxonomic_richness_plots” directory. The plot layout is customizable with the options to change the width and height and the font size.

4.3.4 OTU abundance pie charts

For each taxonomic level (Phylum to Species) a pie chart depicting the relative OTU abundances will be created. These plots show the relative number of reported OTUs per taxon and not the read abundances (refer to 4.3.7 for read abundance-based plots). Two version of each plot are created, one including and one excluding the “nan” status, since the proportion of unassigned taxonomy usually increases with higher taxonomic level and thus can render the plots unreadable. The results are written to a pdf file in the “Projects/your_project/Pie_charts” directory.

4.3.5 Venn diagrams

Venn diagrams are a common way of comparing different sets. This tool uses venn diagrams to depict the taxonomy overlap (Phylum to Species, respectively) of all samples in two TaXon tables. Here, each table is handled as one set. The easiest way to compare specific sample sets is to create new TaXon tables using the one the filtering tools (see 4.2.6 and 4.2.7) or the “TaXon table per sample” tool (see 4.2.4). When executed, this tool scales the overlap of shared taxa and the non-shared taxa intro three distinctive areas. The respective number of taxa is given for each area. Taxon names are written to a seperate xlsx file. The final venn diagrams can be found in pdf format along the xlsx file in the

“Projects/your_project/Venn_diagrams” directory. The venn diagram tool allows to compare two (venn2 option) or up to three (venn3) tables.

4.3.6 Rarefaction curves

The rarefaction curve tools allows to assess the species richness from the results of sampling. The curve is calculated by randomly drawing one individual sample at a time after each other and adding up the number of newly added OTUs by the drawn sample. The stochasticity is reduced by re-sampling from the samples pool for multiple times (default = 1000) and calculating the average number of OTUs. The standard deviation can be visualized as specific error bars or as transparent background. The rarefaction plot is saved in pdf format under “Projects/your_project/Rarefaction_curves”.

4.3.7 Read proportions

This tool allows the user to display the relative read proportions of a given sample set as a scatter plot. The taxonomic level can be chosen by the user (phylum to species and OTUs). For larger data sets it is recommended to use a higher taxonomic level, as the plot easily becomes confusing and hard to read on species or genus level. Otherwise it might be easier to create filtered subsets of the main table and display them in individual plots. The scatter size draws the relative read abundance for each taxa. The according size is given in the legend. For better visualization and differentiation, the scatters are presented in an alternating color scheme. The plot layout is customizable with the options to change the width and height and the font size. The read proportion plot is saved in pdf format under “Projects/your_project/Read_proportion_plots”.

4.3.8 Site occupancy

This tool will calculate the site occupancy, which is defined as relative occurrence of a taxonomic level (phylum to species) among all samples across one user-specified site. The occupancy ranges from 1 (= taxa found in all samples at this location) to 0 (= taxa found in none of the locations). The location for each sample must be defined in the metadata table. For each site a separate plot will be generated. The plot layout is customizable with the options to change the width and height and the font size. The final plot will be saved as pdf file in the “Projects/your_project/Site_occupancy_plots” folder.

4.3.9 Krona charts

Krona allows to visualize hierarchical data in a multi-layered pie chart. Each layer (i.e. the taxonomic levels from phylum to species) can be interactively zoomed. A tab-separated table will be created from the chosen TaXon table and functions as input file for Krona tools. The output format of Krona tools is a html file, which can be viewed with any browser (e.g. Firefox, Chrome or Safari). Within the browser

the Krona chart can be saved as vector graphic. Accordingly, this tool requires Krona tools to be installed. The Krona chart will be saved as xlsx table in the “Projects/your_project/Krona_charts” folder.

4.3.10 Alpha and beta diversity

TTT offers calculation of alpha and beta diversity and ordination analyses. The implemented tools are mostly dependent on the python package scikit-bio (http://scikit-bio.org/). All diversity analyses require an incidence data TaXon table and a meta data table. The alpha diversity calculation is based on the number of OTUs per sample, which are displayed as a scatter plot (“Projects/your_project/Alpha_diversity”). Beta diversity is calculated as jaccard-distances, which are illustrated in a distance matrix (“Projects/your_project/Beta_diversity”). Furthermore, a jaccard-distance based principle coordinate analysis (PCoA) can be performed (“Projects/your_project/PcoA_plots”). A canonical-correlation analysis (CCA) tool is also implemented (“Projects/your_project/CCA_plots”). For both ordination analyses, it is possible to choose two axes from all available axes for plotting and a table with the respective Eigenvalues ist saved.

4.4 Taxon list

4.4.1 TaXon table conversion

This tool can be used to generate a simple taxon list from the TaXon table. Thus, only the taxonomic information of the OTUs will be kept. OTUs that have the same taxonomic hit will be collapsed. Furthermore, statistics can be calculated for each taxon. These include the absolute number of reads per taxon and the relative proportion within the data set, the occupancy across all samples, the number of OTUs identified as the respective taxon and the intraspecific distances for taxa with multiple OTUs. The taxon list will be saved as xlsx table in the “Projects/your_project/Taxon_lists” folder.

4.4.2 Global Biodiversity Information Facility (GBIF) link

GBIF is an excellent backbone for biodiversity data and we use the taxonomic information provided. For each OTU that has a hit on species level, a link to its entry on the Global Biodiversity Information Facility (https://www.gbif.org/) link can be created. This feature is optional and requires internet connection.

4.4.3 Additional information

The additional information input fields can be used to create a final report sheet. This tool is targeting to enhance the data backup. All information will be gathered and written into a text file. Current input topics are: Description, Author(s), Lab protocol, Number of replicates used, Number of negative controls used, Primers, Sequencing run and Bioinformatics pipeline. The language can be set to English or German.

4.5 Water Framework Directive (WFD) taxa lists

4.5.1 Taxa lists to support Water Framework Directive bioassessments

This tool converts the TaXon table to a presence-absence table that can be used as input for quality assessment of streams according to the European Water Framework Directive (WFR; https://ec.europa.eu/environment/water/water-framework/index_en.html.

The quality assessment is performed on the "Gewaesser Bewertung Online" website (https://www.gewaesser-bewertung-berechnung.de/index.php/home.html), which is currently only available in German. In the latest version of TaxonTableTools only the conversion to Perlodes (macrozoobenthos) input is implemented. The conversion for the Phylib, PhytoFluss and fiBS tools will be implemented in a future version.

4.5.2 Perlodes

The conversion of the TaXon table to a Perlodes table requires the latest official operational taxon list used for bioassessment of benthic macroinvertebrates in Germany, which can be downloaded here: https://www.gewaesser-bewertung-berechnung.de/index.php/perlodes-online.html. During the conversion a tab separated list of all OTUs with a match in the operational taxon list will be created. The taxonomy will be adjusted to meet the requirements of the operation taxon list (i.e. reduced to a higher taxonomic level). The read abundances will be converted to presence absence data (1,0) for each OTU. The output file is found under "Projects/your_project/Perlodes" and can be used in the online Perlodes tool. For other countries, this option is not of interest but might in the future be extended with other national bioassessment tables.

5. References

Buchner, Dominik, and Florian Leese. “BOLDigger – a Python Package to Identify and Organise Sequences with the Barcode of Life Data Systems.” Metabarcoding and Metagenomics 4 (June 3, 2020): e53535. https://doi.org/10.3897/mbmg.4.53535.

Bolyen, Evan, Jai Ram Rideout, Matthew R. Dillon, Nicholas A. Bokulich, Christian C. Abnet, Gabriel A. Al-Ghalith, Harriet Alexander, et al. “Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using QIIME 2.” Nature Biotechnology 37, no. 8 (August 2019): 852–57. https://doi.org/10.1038/s41587-019-0209-9.

Boyer, Frédéric, Céline Mercier, Aurélie Bonin, Yvan Le Bras, Pierre Taberlet, and Eric Coissac. “Obitools: A Unix-Inspired Software Package for DNA Metabarcoding.” Molecular Ecology Resources 16, no. 1 (2016): 176–82. https://doi.org/10.1111/1755-0998.12428.

Callahan, Benjamin J., Paul J. McMurdie, Michael J. Rosen, Andrew W. Han, Amy Jo A. Johnson, and Susan P. Holmes. “DADA2: High-Resolution Sample Inference from Illumina Amplicon Data.” Nature Methods 13, no. 7 (July 2016): 581–83. https://doi.org/10.1038/nmeth.3869.