

## qDiv manual

qDiv is a software for analysing microbial ecology data generated using, for example, high-throughput amplicon sequencing. qDiv is written in Python and the code is available on [github.com/omvatten/qDiv](https://github.com/omvatten/qDiv). A compiled version of the software, which can be run on Windows, is downloadable at [omvatten.se/software](https://omvatten.se/software). When you download and install the software it should automatically create a shortcut icon on your desktop.

Three input data files should be supplied by the user: a meta data file, a frequency table, and a fasta file with sequences. The file formats are described below. Examples are also available on [github.com/omvatten/qDiv](https://github.com/omvatten/qDiv).

### Input files

The uppermost row of the meta data file contains the headings. The first column contains the sample names. The following columns contain information about the samples. An example is shown below. In this example, the data set contains six samples. They are collected from three different reactors (1, 2, and 3) on two different days (1 and 4). The reactors receive a feed containing either acetate (reactor 1 and 2) or glucose (reactor 3).

samples	reactor	day	feed
S1	1	1	acetate
S2	2	1	acetate
S3	3	1	glucose
S4	1	4	acetate
S5	2	4	acetate
S6	3	4	glucose

The uppermost row of the frequency table contains the sample names. The first column contains the sequence variants (SVs) or operational taxonomic units (OTUs). The table contains the number of reads (or observations) associated with each sample and SV. The rightmost columns contain taxonomic information starting with Kingdom, then Phylum, Class, Order, Family, Genus, and Species. An example is shown below.

SVs	S1	S2	S3	S4	S5	S6	Kingdom	Phylum	Class
SV1	97	54	1	89	56	4	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria
SV2	55	79	22	45	98	43	k__Bacteria	p__Proteobacteria	c__Alphaproteobacteria
SV3	42	0	76	23	12	88	k__Bacteria	p__Actinobacteria	c__Actinobacteria
SV4	2	56	43	0	43	32	k__Bacteria	p__Bacteroidetes	c__Flavobacteriia
SV5	1	12	89	4	2	64	k__Bacteria	p__Bacteroidetes	c__Sphingobacteriia
SV6	0	2	2	1	1	9	k__Archaea	p__Euryarchaeota	c__Methanobacteria
SV7	0	1	0	0	1	3	k__Bacteria	p__Proteobacteria	c__Deltaproteobacteria
SV8	0	1	1	0	0	0	k__Bacteria	p__Proteobacteria	c__Deltaproteobacteria
SV9	1	0	1	3	0	1	k__Bacteria	p__Proteobacteria	c__Deltaproteobacteria
SV10	2	0	1	0	0	1	k__Bacteria	p__Proteobacteria	c__Deltaproteobacteria

The fasta file contains the sequences. An example of a fasta file is shown below. The sample names in the meta data must match the sample names in the frequency table and the SV or OTU names in the frequency table must match the SV or OTU names in the fasta file.

```

Open  Save
>SV1
TACGTAGGTGGCAAGCGTTGTCGGATTATTGGGCGTAAAGCGAGCGAGCGGTTCTTAAGTCTGATGTAAAGCCACGGC
>SV2
TACGGAGGGTGCAAGCGTTGTTCCGAATTATTGGGCGTAAAGCGCGTGTAGCGGTTGGTTAAGTCTGATGTAAAGCCCTGGGC
>SV3
TACGGGGGGTGCAAGCGTTGTTCCGAATTATTGGGCGTAAAGCGCGTGTAGCGGTTGATTAGTCTGATGTAAAGCCCTGGGC
>SV4
TACGTAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTGGCAGCGGTTTGTAAAGACAGATGAAATCCCCGGGC
>SV5
TACGTAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTGGCAGCGGTTTGTAAAGACAGCGTAAATCCCCGAGC
>SV6
TACGTAGGGGGGAGCGTTGTCGGATTACTGGGCGTAAAGGTCACGAGCGGTCATGTAAGTCAGATGAAATCCTAAGGC
>SV7
TACAGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCGTGTAGTGGTTAGTTAAGTTGGATGTAAATCCCCGGGC
>SV8
TACGTAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTGGCAGCGGTTGTGTAAGACAGGTGAAATCCCCGGGC
>SV9
TACGGAGGATGCAAGCGTTATCCGATTATTGGGTTTAAAGGTCGCGAGCGGGCTAATAAGTCAGGGGTGAAATACACGGC
>SV10
TACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGAGCGGGCTTTTAAAGTCGGATGTAAAGCCCCGGGC
Plain Text  Tab Width: 8  Ln 5, Col 5  INS

```

## Start window

When you open qDiv, the start window appears. Press the buttons named "Frequency table", "Fasta file", and "Meta data" and select the appropriate files. Then choose the type of separator used in the frequency table and meta data files (e.g. a comma for csv files). It must be the same type of separator in both files. Then specify a folder by clicking the button named "Folder for output files". All files you generate will be placed in this folder.

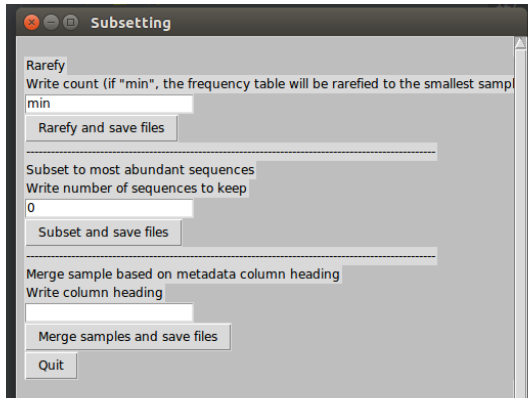
To get some information about your data, click "Press for stats". A window will pop up with information about the number of samples, the number of reads, the min and max number of reads in a sample, and the headings in the meta data.

The next step is to choose a task. We can choose between "Subset\_data", "Calculate\_phyl\_dist", "Heatmap", "Alpha\_div", "Beta\_div", "PCoA", and "Null\_model". Mark your choice, then click the "Choose" button. The "Quit" button closes the program.

## Subset data

If we choose "Subset\_data", a new window pops up. Here we can rarefy the frequency table, subset it to the most abundant SVs, or merge samples based on specific categories in the meta data. If "min" is written in the field under Rarefy and we click "Rarefy and save file" the frequency table

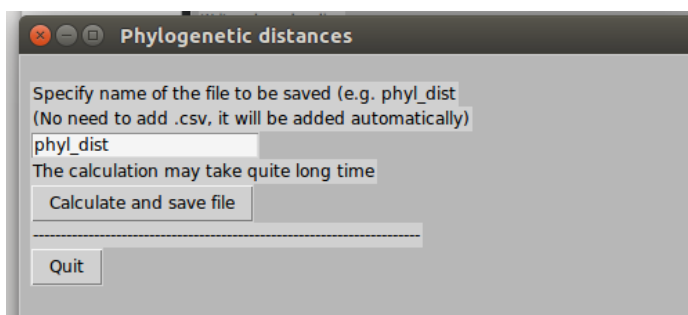
will be rarefied so that each sample contain the same number of reads as the sample with the lowest count. We can also specify the rarefaction depth by writing a number. The rarefied frequency table, fasta file, and meta data will be saved in the folder specified for output. If we want to use these files in our further analysis, we should go back to the start window and input these files. Again, we can click "Press for stats" to see how rarefaction changed the data.



### Calculate phylogenetic distances

Next, we choose "Calculate\_phyl\_dist" on the start window. A new window pops up. Here we can run an algorithm that calculates the distance between each pair of sequences in the data set. The distance is calculated as the Levenshtein distance divided by the length of the longest of the two sequences. A value of 0 means the sequences are identical and a value of 1 means they are completely different. Specify a name of the output file and click "Calculate and save file". A csv file holding the distance matrix will be saved in the output folder. The calculation can be quite time consuming if the data contain many sequences. A pop up window will show the calculation progress.

The distance matrix file generated here will only be needed later if you plan to work with phylogenetic alpha- and beta diversity indices.



### Plot heatmap

Next, let's choose Heatmap in the start window. Here, we will plot a heatmap showing the percentage read abundance associated with different taxa and samples.

First, we specify the taxonomic levels to include in the plot. We may choose one or two levels. The SVs will be binned based on the lowest chosen level. This is the only mandatory choice we have to do before plotting the heatmap. All other choices are about customizing the plot.

- We can specify a meta data column to use in the x-axis of the heatmap. The samples will be merged based on the data in this column. If we leave this field as 'None', the sample names will be left as they are.

- We can specify a meta data column used for ordering the samples. The column must contain numbers.
- We can specify the number of the most abundant taxa to include in the heatmap.
- We can write the name to use for unclassified sequences, for example SV or OTU.
- We can modify height and width of the figure, and the font size.
- We can decide if we want to group samples by inserting a blank column between samples. If we for example write 3,6, blank columns will be inserted after the 3rd and 6th sample.
- We can decide if we want to show data labels in the heatmap, and the font size of those labels.
- We can decide the color of labels. They will be either black or white and here we specify at which percentage it changes color.
- We can specify the color to use. The linearity of the colormap specifies how the color changes from bright at low value to dark at higher values. If we want a legend colorbar, we must use the next field to specify the tick marks in that legend, for example 0.1,1,10,50.
- Finally, we can subset the heatmap to specific taxa. Check the taxonomic levels to search. Then add one or several text patterns to search for, separate by commas. For example, Nitroso, Brocadia.

When all choices are made, click "Plot heatmap". The heatmap will be plotted and saved as pdf and png files in the output folder.

The screenshot displays the 'Heatmap' application window, which is divided into two main panels. The left panel, titled 'Various input options for heatmap', contains several sections: 'Choose one or two taxonomic levels to include on the y-axis' with checkboxes for Kingdom, Phylum, Class, Order, Family, Genus, and Species; 'Enter metadata column for x-axis labels' with a dropdown set to 'None'; 'Specify metadata column used to order the samples on the x-axis' also set to 'None'; 'Specify number of taxa to include in heatmap' set to 20; 'Specify how unclassified taxa should be named (e.g. SV or OTU)?' set to 'SV'; 'Specify figure dimensions and text size' with fields for Width (14), Height (10), and Axis text font size (15); 'Group samples. Insert numbers of samples after which a blank column should be inserted' with a text input field; 'Information about data labels' with a radio button for 'Yes' selected, 'Label font size' set to 12, and 'Percent relative abundance at which the label text shifts from black to white' set to 10. The right panel, titled 'Specify figure dimensions and text size', contains: 'Group samples. Insert numbers of samples after which a blank column should be inserted' with a text input field; 'Information about data labels' with a radio button for 'No' selected; 'Label font size' set to 12; 'Percent relative abundance at which the label text shifts from black to white' set to 10; 'Color of heatmap' with 'Colormap' set to 'Reds' and 'Linearity of colormap' set to 0.5; 'If you want a colorbar showing the scale, specify tick marks on the bar' with a text input field; 'Subset data based on text patterns' with checkboxes for Kingdom, Phylum, Class, Order, Family, Genus, and Species; and 'Enter words to search for, separate by comma' with a text input field. At the bottom of the right panel are 'Plot heatmap' and 'Quit' buttons.

## Calculate alpha diversity

Next, let's choose "Alpha\_div" in the start window. If we want to calculate phylogenetic diversity values, we click "Select" and open the file containing the distance matrix we calculated previously. If we don't select a distance matrix file or click "Reset selection", naive (or taxonomic) diversity values will be calculated. First, we can calculate a figure plotting diversity value vs diversity order. If we want to color code the figure based on information in the meta data, we input the heading of

the appropriate column in the first field. If the figure legend should list the samples in a specific order, we can input the heading of the meta data column specifying the order in the next field. Finally, we can choose if the y-axis showing the diversity values should be logarithmic or not. Click "Plot alpha diversity" to generate the figure.

We can also save alpha diversity values in a csv file. We specify the diversity orders to calculate and separate the values by comma, for example, 0,1,2. Then, we specify the name of the saved file. Finally, click "Save alpha diversity data as file".

The screenshot shows a window titled "Alpha diversity" with the following sections:

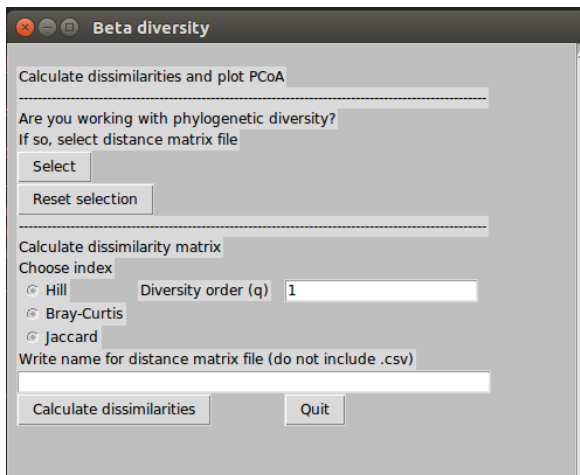
- Show alpha diversity in plots or save as data files** (header)
- Are you working with phylogenetic diversity?**
  - If so, select distance matrix file: [Select] [Reset selection]
- The following input is used for plotting figure...**
  - Specify metadata column heading to use for color coding: [None]
  - Specify metadata column used to order the samples on the x-axis: [None]
  - Use logarithmic y-axis?
    - ☒ Yes
    - ☐ No
  - [Plot alpha diversity figures]
- The following input is used to save a csv file with data**
  - Specify diversity orders to calculate, use comma to separate numbers: [ ]
  - Specify name of saved file (do not include .csv): [ ]
  - [Save alpha diversity data as file]
- [Quit]

### Calculate beta diversity

Next, let's calculate dissimilarities. Choose "Beta\_div" in the start window. Again, we have the opportunity to decide if we want to calculate phylogenetic or naive (taxonomic) dissimilarity values by selecting (or not selecting) a distance matrix file. The phylogenetic indices are only relevant for Hill-based dissimilarities.

We specify the type of dissimilarity index to calculate. We can calculate Hill-based dissimilarities of any diversity order. We can also calculate Bray-Curtis and Jaccard.

In the bottom field we write the name of the file to save. Click "Calculate dissimilarities" to carry out the calculation and save the dissimilarity matrix as a csv file.



**Beta diversity**

Calculate dissimilarities and plot PCoA

Are you working with phylogenetic diversity?  
If so, select distance matrix file

Select

Reset selection

---

Calculate dissimilarity matrix

Choose index

☒ Hill Diversity order (q) 1

☐ Bray-Curtis

☐ Jaccard

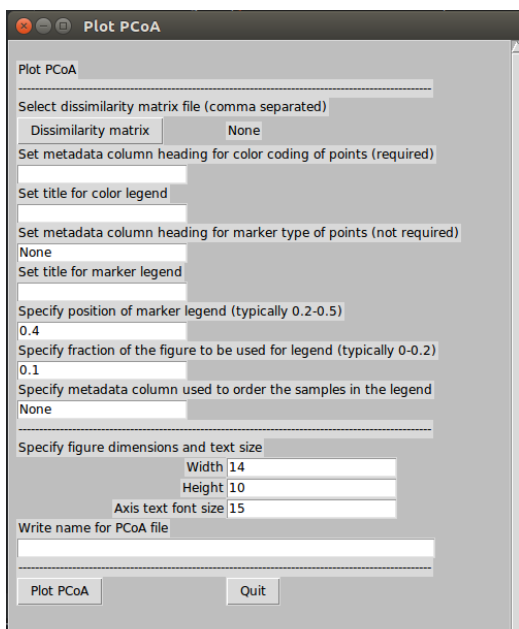
Write name for distance matrix file (do not include .csv)

Calculate dissimilarities Quit

## Plot PCoA

Next, we will plot a PCoA based on the calculated dissimilarity values. Choose "PCoA" in the start window. We select the file containing the dissimilarity matrix by clicking "Dissimilarity matrix". Then, we specify a column heading in the meta data to color code the data points. Samples having the same value or belonging to the same category in that column will have the same color in the plot. It is also possible to specify a title for the color legend. Optionally, we may use another meta data column for different marker types in the plot. If we do, we can also specify the position of the marker legend (0.4 is default, a lower value means a lower position and vice versa). We can also specify, the fraction of the figure that should be reserved for the legend (default is 0.1). To order the samples in the legend in a specific way, we specify the meta data column with information about the order.

Finally, we can specify the height and width of the figure and the font size. We also write a name for the saved plot before clicking "Plot PCoA".



**Plot PCoA**

Plot PCoA

Select dissimilarity matrix file (comma separated)

Dissimilarity matrix None

Set metadata column heading for color coding of points (required)

Set title for color legend

Set metadata column heading for marker type of points (not required)

None

Set title for marker legend

Specify position of marker legend (typically 0.2-0.5)

0.4

Specify fraction of the figure to be used for legend (typically 0-0.2)

0.1

Specify metadata column used to order the samples in the legend

None

---

Specify figure dimensions and text size

Width 14

Height 10

Axis text font size 15

Write name for PCoA file

Plot PCoA Quit

## Null model

Different environments may harbour different numbers of microorganisms. It is natural for an environment harbouring many microorganisms to have a higher richness than an environment harbouring a small number. This richness difference will cause a dissimilarity. The goal of the null model is to distinguish dissimilarity caused by richness differences to dissimilarity caused by compositional differences.

Choose "Null\_model" in the start window. First, we specify the type of dissimilarity index to calculate by selecting (or not selecting) a distance matrix file and choosing a dissimilarity index. We also select the number of randomization to carry out. At least 999 are recommended.

The null model will randomize the frequency table based on certain constraints. The randomization will maintain the richness of each sample, i.e. the number of SVs in each sample will remain the same. However, the identity of the SVs and the read count associated with each will depend on random sampling from a meta community.

Three ways of picking SVs and read counts from a meta community are available:

- "abundance" randomly selects SVs and counts in each sample based on the total read count for the SVs in the frequency table.
- "frequency" randomly selects SVs in each sample based on the proportion of samples in the frequency table in which the SV is found. Reads counts are then randomly selected based on the total number of counts associated with each SVs.
- "weighting" uses the "abundance" method but by specifying a "weighting variable" in the meta data, the user can give less weight to samples belonging to a category with lower richness. If the weight parameter is 1, the method is equivalent to abundance. If it is 0, the read counts of SVs in samples belonging to the category with lowest richness are not taken into account when SVs are selected and counts are distributed. In cases when different environments harbour vastly different numbers of microorganisms, the relative contribution of different samples to the meta community could differ. The weighting method allows us to account for this by giving less weight to low-diversity samples.

A key component of the null model is how we define the meta community from which random samples are taken. If we write a heading from the meta data in the "Constraining variable" field, samples having the same value or category in that meta data column will be considered as one meta community and randomization will only take place within the specific subsets of samples. If the field says "None", the whole frequency table will be considered as one large meta community. We can also specify a meta data column for grouping samples into different categories. If this is done, the mean and standard deviation of all pairwise dissimilarities between samples from different categories will be returned.

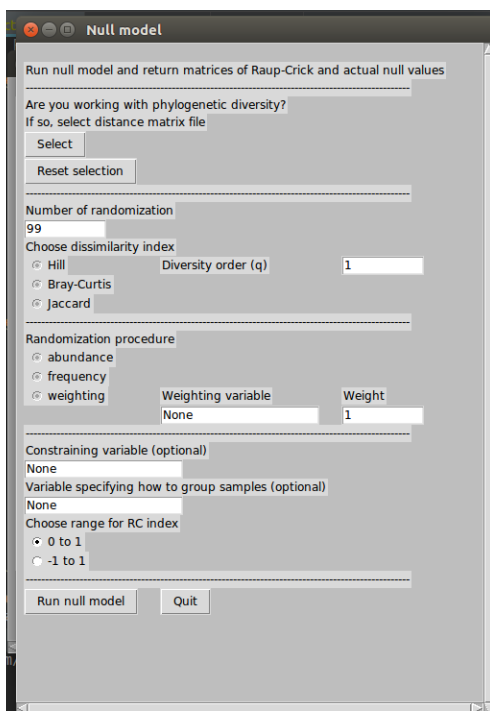
Finally, we specify the range of the output Raup-Crick dissimilarity values. Either 0 to 1 or -1 to 1. The Raup-Crick dissimilarities can be interpreted as follows:

RC range 0 to 1	RC range -1 to 1	Interpretation
< 0.5	< 0	Less dissimilar than null
0.5	0	Same as null
> 0.5	> 0	More dissimilar than null

Click "Run null model" to start the calculation. The calculation may take several hours if large frequency tables are used. The calculation consists of two steps: a randomization of the frequency table and a calculation of dissimilarities. Small windows showing the progress of each step will pop up.

After the calculation is done, 3 or 4 files will be returned. "Nullmean...csv" and "Nullstd...csv" contain the mean and standard deviation of dissimilarities for all randomized frequency tables. If a grouping variable is not specified, the file "RC...csv" will give the Raup-Crick dissimilarities between all samples.

If a grouping variable is specified, "RCmean...csv" and "RCstd...csv" will contain the mean and standard deviations of all pairwise Raup-Crick dissimilarities between samples in the different groups.



## Python packages

qDiv depends on the following python packages: pandas, numpy, python-Levenshtein, matplotlib, and tkinter.

## Bibliography

Chao, Chiu and Jost (2014). *Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through Hill numbers*. *Annual Review of Ecology, Evolution, and Systematics*, 45, 297-324.

The alpha diversity- and Hill-based dissimilarity indices in qDiv are calculated using the methods described by Chao et al. (2014). The dissimilarity indices are identical to one minus the local overlap measures ( $1-C_{q,2}$ ). The phylogenetic diversity indices in qDiv are calculated based on a distance matrix and are thus identical to functional diversity (FD) in Chao et al. (2014).

Stegen, Lin, Fredrickson, Chen, Kennedy, Murray, Rockhold, and Konopka (2013). *Quantifying community assembly processes and identifying features that impose them*. *The ISME Journal*, 7, 2069-2079.



Stegen et al. (2013) extended the Raup-Crick index to consider the relative abundance of SVs/OTUs by using it in conjunction with the Bray-Curtis dissimilarity index. In qDiv, we extend Raup-Crick even further by applying it with the whole range of Hill-based dissimilarity indices. We also provide several randomization options. The option "frequency" should be equivalent to that described in Stegen et al. (2013).

*Legendre and Legendre (2012). Numerical Ecology, 3rd Ed. Elsevier.*

The method for principal coordinate analysis was taken from chapter 9 in this book.

*Chase, Kraft, Smith, Vellend, and Inouye (2011). Using null models to disentangle variation in community dissimilarity from variation in  $\alpha$ -diversity. Ecosphere, 2, 24.*

*Raup and Crick (1979). Measurement of faunal similarity in paleontology. Journal of Paleontology, 53(5), 1213-1227.*