myPhyloDB: A local database storage and retrieval system for the analysis of metagenomic data

USDA Agricultural Research Service

Overview:

myPhyloDB is a user-friendly personal database with a browser-interface for accessing and analyzing taxonomic data from multiple projects and/or sequencing runs.

The goal of myPhyloDB is to allow for easy comparisons and statistical analysis of microbial (i.e., fungi or bacteria) taxonomic abundance across projects, soil types, and management scenarios.

Data may be obtained from any sequencing platform; however, currently only mothur-formatted files can be uploaded to myPhyloDB.

This manual is a reference for the use of myPhyloDB.

Table of Contents:

1. Installation

Windows:

Double-click the installer (myPhyloDB_1.0_Win_x64_install.exe) and follow the prompts. The program will install a myPhyloDB shortcut to your start menu (Windows 7) or start screen (Windows 8). Clicking the myPhyloDB icon will open a terminal and web-browser that provides the user-interface for running the myPhyloDB program. To exit the program type ctrl-c in the terminal and manually close your browser.

Linux:

Extract the installer (myPhyloDB_1.0_Linux_x64.tar.gz) and run the install.sh file in a terminal. The program will install a myPhyloDB shortcut to your Desktop. Clicking the myPhyloDB icon will open a terminal and web-browser that provides the user-interface for running the myPhyloDB program. To exit the program type ctrl-c in the terminal and manually close your browser.

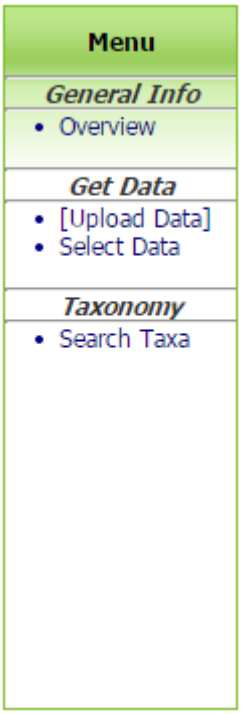
Remote access:

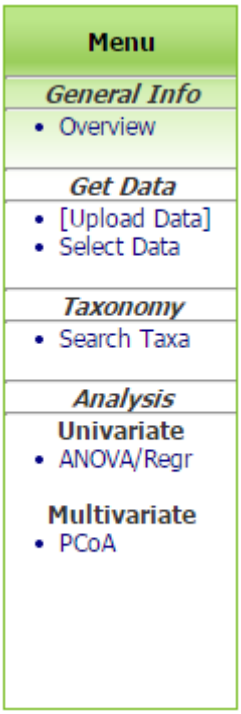
myPhyloDB will run as a local server on your host machine allowing others on your local intranet to access myPhyloDB (unless disabled using your computer's firewall settings) without installing a separate copy. This may be useful for laboratories that want to share data across multiple users. To access myPhyloDB from a remote computer you must first obtain the IP address of the host machine (in a terminal on the host machine, type 'ipconfig'), then in the address bar of your remote computer's browser enter the following address 'xxx.xxx.x.xx:8000/myPhyloDB/home/' replacing the x's with the appropriate IP address. All data uploads and/or removal of projects by authorized (see Admin section) remote users will be saved to the host computer's installation of myPhyloDB.

1. Home Screen and Sidebar

The home screen (http://127.0.0.1:8000/myPhyloDB/home/) provides general information about myPhyloDB as well as links to this instruction manual and example files for uploading new projects into myPhyloDB.

Navigation between the various pages and analyses provided by myPhyloDB is performed using the Menu sidebar at the left of the screen. The first time you launch myPhyloDB, the sidebar should like the picture to the left. Once you have selected some projects/samples for analysis, the sidebar will show two new links to the various data analysis pages, as shown in the picture on the right. Selected projects/samples are saved to a 'cookie' and depending upon your browser settings will be stored between sessions.





2. Uploading New Projects

To upload data, click "[Upload Data]" on the left hand menu (http://127.0.0.1:8000/myPhyloDB/upload/). For your protection, this page can only be accessed by an authorized user – to add/remover users see the Admin section of this manual. For upload you should have prepared two files with metadata (Project.csv and Sample.csv) using the templates provided setting any missing data to 'null'. All columns in the Project and Sample files are required and upload will fail if any changes are made to the header row. MyPhyloDB does not perform any unit checking or conversion of data so consistent units should be used for all Projects and Samples.

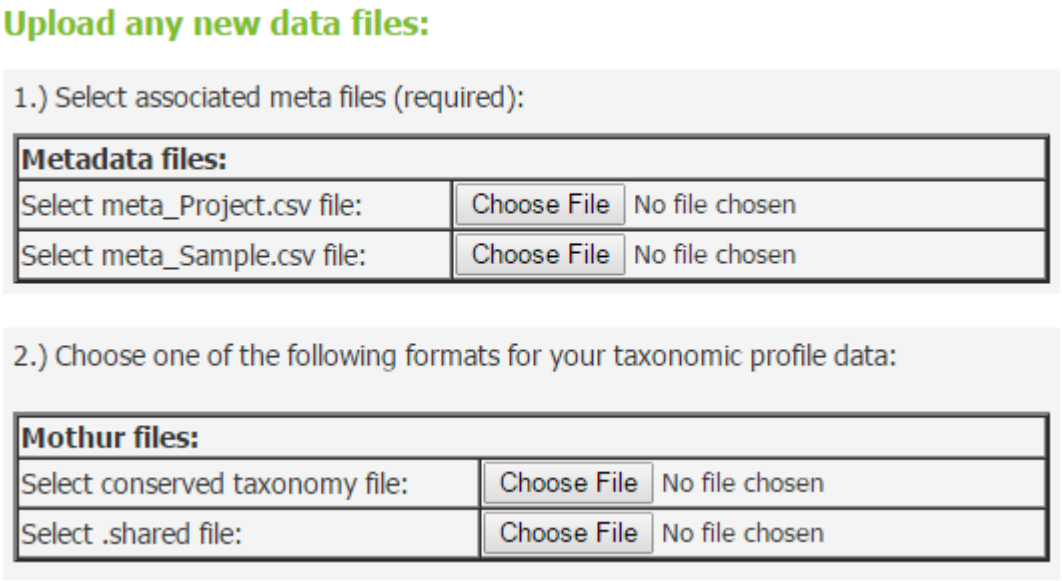
In order for the samples to be correctly associated with a project, only one project can be uploaded at a time (i.e., one row of data in your Project.csv file). However, new samples (i.e., new sample_name) may be added to an already uploaded project by setting the project_id to the auto-generated UUID found in the datatable located on the "Select" data page. Similarly, to add new sequence data to an existing sample you must set the project_id and sample_id values to the appropriate auto-generated UUIDs. Currently, if you wish to change any of the metadata associated with a previously uploaded sample you must remove the entire project and re-upload the new data.

In addition, you will also need to create two files with the appropriate sequencing data similar to the mothur.shared and mothur.taxonomy examples provided, which are output from mothur (www.mothur.org). The shared file can be generated using the make.shared command but must contain only one OTU level (e.g., label = 1). The taxonomy file can be generated using the classify.otu command using the same OTU level. For example, assuming you have the following three mothur files (final.fasta, final.names, final.groups) run the following commands in mothur to generate the required files.

```
classify.seqs(fasta=final.fasta, template=gg_13_5_99.fasta, taxonomy=gg_13_5_99.pds.tax)

phylotype(taxonomy=final.pds.wang.taxonomy, name=final.names, label=1)

make.shared(list=final.pds.wang.tx.list, group=final.groups)

classify.otu(taxonomy=final.pds.wang.taxonomy, name=final.names, group=final.groups,
list=final.pds.wang.tx.list)
```

If you follow the above procedure the two files needed for upload will be named: "final.pds.wang.tx.shared" and "final.pds.wang.tx.1.cons.taxonomy". Due to taxa naming differences between the various reference databases (e.g., RDP, GreenGenes, SILVA), it is recommended that a single reference database be used consistently with myPhyloDB. Also, the architecture of myPhyloDB is such that all OTUs must have an entry for all seven main taxonomic levels (I.e., Kingdom, Phyla, Class, Order, Family, Genus, Species) so to avoid manually editing your taxonomy file we recommend the GreenGenes or SILVA reference databases provided by mothur (www.mothur.org/wiki/Taxonomy_outline). If necessary, 'unclassified' can be used for any taxonomic level without relevant information (e.g.,, species when using RDP).

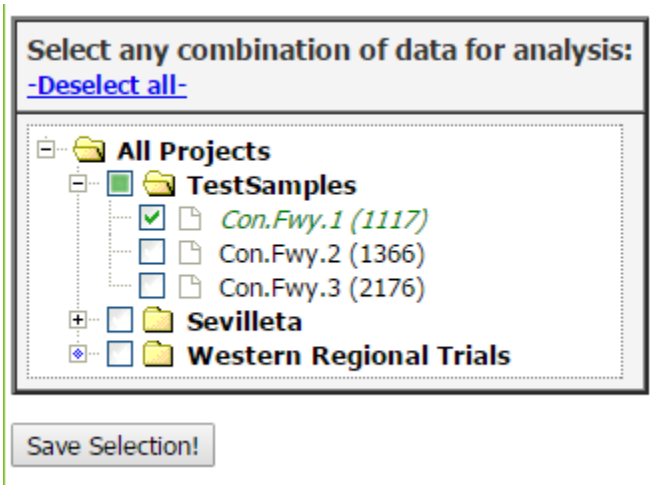
Once you have the four files required for upload add each file to the appropriate box located on the "[Upload Data]" page using the four file choosers (see below). When you are finished, click "Upload Files". Note: the upload process includes four steps and may take several minutes depending upon file size and computer speed. For your convenience a progress bar will appear below the "Upload Files" button documenting the status of the upload and parsing steps required to populate the myPhyloDB database.



At the bottom of the "[Upload Data]" page is a list of current projects already uploaded to your myPhyloDB database. If you want to remove any of these projects simply click the appropriate box and then the "Remove selected projects" button. Edited projects (i.e., new submission files) can then be uploaded as described above.

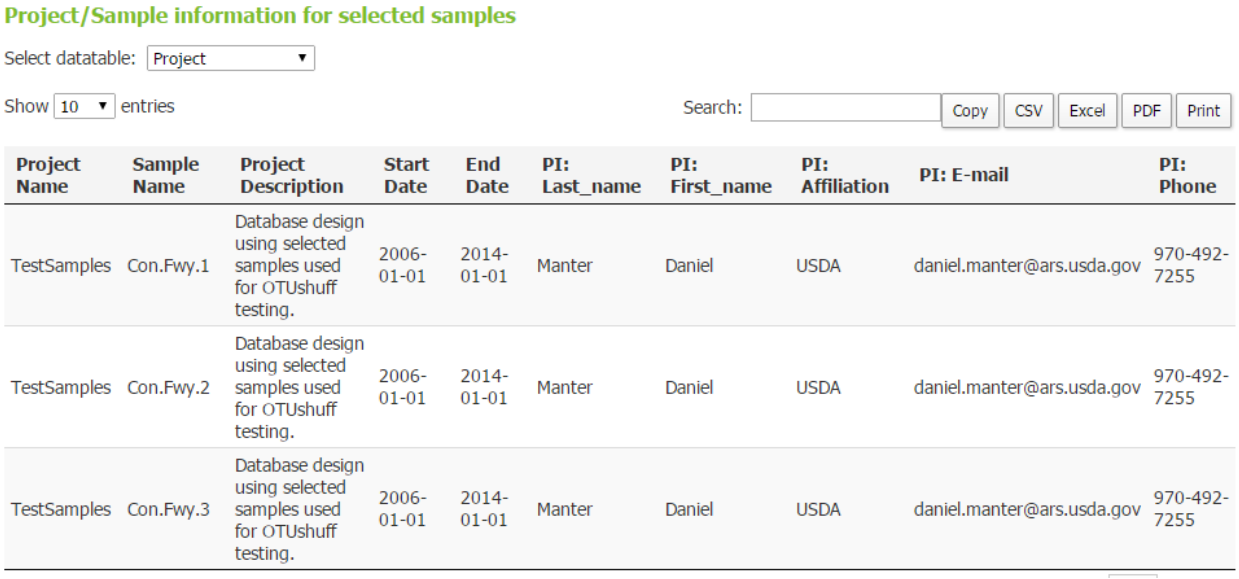
3. Selecting Data for Analysis

To select data for analysis, click “Select Data” on left hand menu (http://127.0.0.1:8000/myPhyloDB/select/). On the select data use the project/sample tree provided to select any combination of projects or samples desired. By default, if a Project checkbox is selected all samples for that project will also be selected. Each project can be expanded and individual samples can be manually selected/deselected. The project/sample tree is organized by project and sample names; however, the project and sample descriptions can be viewed by hovering the mouse over the appropriate name. In addition, the total number of sequence reads for each sample is shown in parentheses next to the sample name.



Completely selected projects will have a green checkmark; whereas, partially selected projects will be filled in with green and the selected samples (Con.Fwy.1) will have a green checkmark. For your convenience, all selections can be cleared using the -Deselect all- link above the tree.

Any projects/samples selected above can be displayed in a datatable containing all of the metadata associated with each sample. Metadata is organized into nine different categories (Project, MIMARKS, Sample Collection, Climate, etc.). You may switch between these categories using the "Select datatable" drop-down box. The first eight categories all have pre-defined names; however, the user-defined table can be used to display any additional parameters that the user might wish to also include as metadata in myPhyloDB.



Project/Sample information for selected samples

Select datatable: Project

Show 10 entries Search: Copy CSV Excel PDF Print

Project Name	Sample Name	Project Description	Start Date	End Date	PI: Last_name	PI: First_name	PI: Affiliation	PI: E-mail	PI: Phone
TestSamples	Con.Fwy.1	Database design using selected samples used for OTUshuff testing.	2006-01-01	2014-01-01	Manter	Daniel	USDA	daniel.manter@ars.usda.gov	970-492-7255
TestSamples	Con.Fwy.2	Database design using selected samples used for OTUshuff testing.	2006-01-01	2014-01-01	Manter	Daniel	USDA	daniel.manter@ars.usda.gov	970-492-7255
TestSamples	Con.Fwy.3	Database design using selected samples used for OTUshuff testing.	2006-01-01	2014-01-01	Manter	Daniel	USDA	daniel.manter@ars.usda.gov	970-492-7255

Each datatable includes a searchbox that can be used to search any field of the displayed table. In addition, each table may be exported to a variety of formats using the button at the top-left of the data table.

Once you have selected the data you wish to analyze further, click "Save Selection!" button below the project/sample tree. Note: Upon clicking the "save selection" button, a pop-up window will appear saying "Selected sample(s) have been recorded!", press "OK" and proceed to the "Analysis" section of myPhyloDB. If this window does not appear you may need to restart myPhyloDB with administrator privileges. To do so, right-click on the myPhyloDB icon and select "run as administrator".

4. Search Taxa:

myPhyloDB provides a search Taxa page (http://127.0.0.1:8000/myPhyloDB/taxa/) to allow users to explore the taxonomic data contained in your myPhyloDB database. The "Taxa name" textbox at the top of the page allows users to quickly search various web engines with a user inputted taxa name. The datatable contains the full taxonomic name of each taxa in your database. For each taxonomic level a unique ID was generated by myPhyloDB for internal tracking purposes and to avoid confusion if duplicate taxonomic names exist. All results in myPhyloDB (next section) will include both taxonomic names and IDs which can be used to identify full taxonomic profiles using this data table. You can also export the table data to CSV, Excel, or PDF files or send the data directly to a printer.

All taxa in database:

Show 10 entries Search: Copy CSV Excel PDF Print

	Kingdom Name	Kingdom ID	Phylum Name	Phylum ID	Class Name
0	Bacteria	ddd63ad8069941a58c07281cc4f8d47a	Proteobacteria	553a8d7f2eb04fa5b105b0c592d5a6e0	Alphaproteobacteria
1	Bacteria	ddd63ad8069941a58c07281cc4f8d47a	Acidobacteria	0161d968630d4151b6166877f3ba598c	Acidobacteria
2	Bacteria	ddd63ad8069941a58c07281cc4f8d47a	Proteobacteria	553a8d7f2eb04fa5b105b0c592d5a6e0	Alphaproteobacteria
3	Bacteria	ddd63ad8069941a58c07281cc4f8d47a	Acidobacteria	0161d968630d4151b6166877f3ba598c	Acidobacteria
4	Bacteria	ddd63ad8069941a58c07281cc4f8d47a	unclassified	8c21c9e3cf3647f4ae002a88944fdfc4	unclassified
5	Bacteria	ddd63ad8069941a58c07281cc4f8d47a	Acidobacteria	0161d968630d4151b6166877f3ba598c	Acidobacteria
6	Bacteria	ddd63ad8069941a58c07281cc4f8d47a	Acidobacteria	0161d968630d4151b6166877f3ba598c	Acidobacteria
7	Bacteria	ddd63ad8069941a58c07281cc4f8d47a	Chloroflexi	f557040d573949df912654cc689dd577	Chloroflexi_class_incertae
8	Bacteria	ddd63ad8069941a58c07281cc4f8d47a	Proteobacteria	553a8d7f2eb04fa5b105b0c592d5a6e0	Alphaproteobacteria
9	Bacteria	ddd63ad8069941a58c07281cc4f8d47a	Proteobacteria	553a8d7f2eb04fa5b105b0c592d5a6e0	Alphaproteobacteria

Showing 1 to 10 of 1,134 entries Previous 1 2 3 4 5 ... 114 Next

5. Analysis:

Once you have selected the samples you would like to analyze, on the menu sidebar, under the "Analysis" heading, select the type of analysis you would like to perform (Univariate: ANOVA/Regr or Multivariate: PCoA).

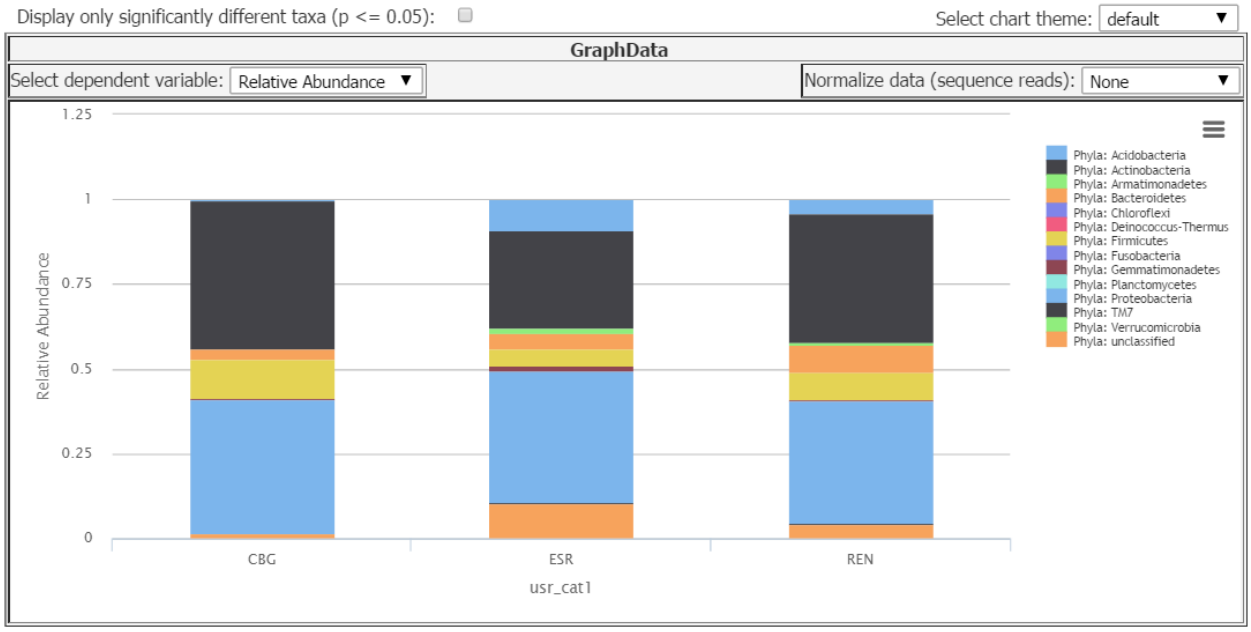
5.1. ANOVA/Regr:

This analysis (http://127.0.0.1:8000/myPhyloDB/ANOVA/) will produce various bar charts and perform a one-way ANOVA to test if the taxonomic level(s) chosen are significantly different between the meta-variables chosen.

The drop down menu "Select data type" allows for the user to switch between categorical and quantitative variables. To perform this analysis, first, select your meta-variable (s) of interest. If more than one variable is chosen, only the interaction term will be analyzed (*e.g.* if you chose site and year, each site x year combination will be analyzed individually) as myPhyloDB only supports one-way ANOVAs. Additionally, if a sample does not contain meta-data for any of the chosen variables (i.e., null values), it will not be included in the final analysis. Fully expanding any meta-variable will result in a list of all of the samples that contain non-null values for that variable. Second, you must select taxonomy data either from the drop down menu (selects all taxa at the select level) or by selecting specific taxonomic name(s) of interest from the taxonomy tree. Third, use the drop down menu "Select dependent variable" to choose between sequence reads (counts), relative abundance, species richness, or Shannon's Diversity Index. Optionally you may also choose to only display only significantly different taxa (checkbox) or normalize you data (see normalize section). Once you are satisfied with your data choices, click "Run Analysis!" button on left menu. The button will change from gray to yellow as analysis is running, then to green when the analysis is complete. If a new combination is selected, the button will change back to gray. If the button turns red check to make sure that both meta-variable(s) and taxonomic name(s) have been selected.

Once the analysis is complete (the "Run Analysis" button is green), scroll down to see your results.

Graph data, statistical results and raw data will appear in boxes below. The chart color theme may be changed (without rerunning the analysis) using the drop down menu above the graph and the chart can be downloaded as a file by clicking one of the 3-horizontal bar buttons just above the graph key. Raw data may also be downloaded for further examination.

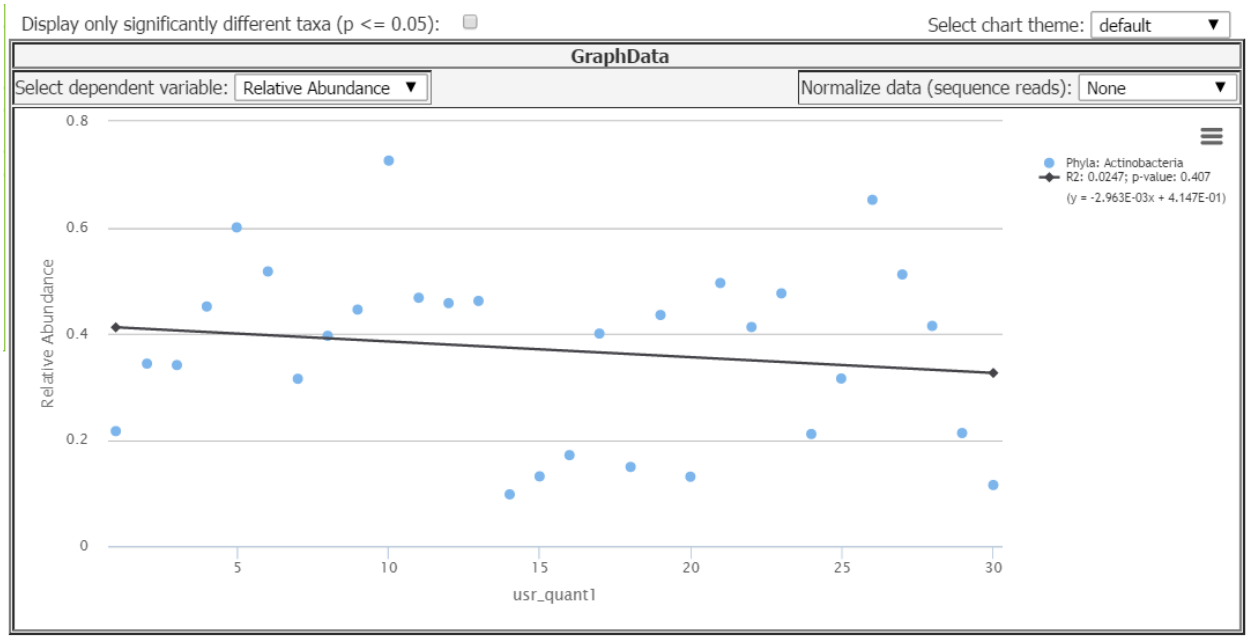


If, for example, you found a specific class of interest and would like to know which phylum it belongs and/or which families are contained within that class, find that information by opening a window with the “Taxonomy” heading on the left menu. Search using either taxa name or taxa ID (which is listed in data analysis graph, statistical analysis and raw data on “Graphs” tab). More information about this step can be found on the “Search Taxa” section of this guide.

If you selected a categorical variable (with appropriate replication), an ANOVA will automatically be calculated and displayed in the “Test Results” window. ANOVA results include a summary of the data, the O'Brien Test for Homogeneity of Variance, the ANOVA test for sources of variation, and a table of q-statistics. Note that ANOVA analysis only applies to categorical data, and will only be performed when a variable has multiple levels and appropriate replication.

The “Raw Data” section displays a datatable with all of the meta-data and sequence data for the samples included in the final analysis.

You can also analyze quantitative data on the graph by following the same procedure, by selecting "Quantitative" for the data type (drop down menu at the top left of the page). In this case, the graph produced will be a scatter plot with linear regression results embedded in the figure legend.



5.2 PCoA (principal coordinates analysis)

The PCoA analysis page (http://127.0.0.1:8000/myPhyloDB/PCoA/) layout and operation is similar to the Univariate graphs, the major difference being the replacement of the taxonomic data table with various drop-down menus.

Taxa level: Select the taxonomic level (e.g., Phyla, Class, Order, Family, Genus, or Species) you would like to use for analysis.

Distance score: Select the distance score (presence/absence scores: Dice, Jaccard; abundance-based scores: Bray-Curtis, Canberra, Euclidean, MorisitaHorn, wOdum) you would like to use for analysis. More detailed information of the scores can be obtained form the following websites:

http://docs.scipy.org/doc/scipy-0.14.0/reference/spatial.distance.html

Dice, Jaccard, Bray-Curtis, Euclidean

http://www.mothur.org/wiki/Morisitahorn

MorisitaHorn

The wOdum score allows users to down-weight either rare ($\alpha > 1$) or abundant ($\alpha < 1$) taxa, as discussed here (Manter and Bakker. Xxxx. BioInformatics xx:xxx-xxx). When $\alpha = 1$, wOdum is equivalent to Bray-Curtis.

Principal coordinate axis selected (x-axis): This is the axis selected as the x-axis in the displayed graph.

Principal coordinate selected (y-axis):This is the axis selected as the y-axis in the displayed graph.

Test selected: Select whether you would like to perform either an AMOVA or HOMOVA analysis of the selected data.

The “Test Results” section lists the AMOVA/HOMOVA results; as well as, the Eigenvalues and proportion of the variance explained for each PCoA axis. Also displayed are datatables of the calculated “Principal Coordinates” and “Distance Scores” in matrix form.

6. Normalization

Aguirre de Cárcer et al. (Appl Environ Microbiol 2011 77:8795-8798) suggest that subsampling to the median number of sequence reads in a dataset can reduce variability and improve analysis. However, for samples with coverage below the subsampling threshold , no normalization procedure was proposed. In order to maintain sampling depths across all samples, myPhyloDB applies a small probability to undetected taxa (I.e., zeros) using Lidstone (Laplace) smoothing. The purpose of this small probability is to account for the uncertainty associated with not knowing whether the missing taxa were truly not present, or present but below the detection level, in the observed data.

The table below shows a simple, hypothetical taxonomic profile for five samples.

Sample	taxa1	taxa2	taxa3	taxa4	taxa5	taxa6	taxa7	taxa8	taxa9	taxa10
S1	13	48	71	54	28	49	0	63	24	7
S2	77	36	50	37	52	68	71	69	12	86
S3	65	9	47	47	66	0	12	2	77	23
S4	99	74	62	75	17	83	17	0	53	19
S5	0	70	67	0	47	46	84	36	92	33

OTU probabilities are then calculated using Lidstone's Approximation [$p = (Ai + \lambda) / (An + N * \lambda)$], where Ai is the observed count for taxa i, An is the total counts for that sample, N is the total number of taxa, and in this scenario $\lambda = 0.1$.

Sample	taxa1	taxa2	taxa3	taxa4	taxa5	taxa6	taxa7	taxa8	taxa9	taxa10
S1	3.66E-02	1.34E-01	1.99E-01	1.51E-01	7.85E-02	1.37E-01	2.79E-04	1.76E-01	6.73E-02	1.98E-02
S2	1.38E-01	6.46E-02	8.96E-02	6.64E-02	9.32E-02	1.22E-01	1.27E-01	1.24E-01	2.16E-02	1.54E-01
S3	1.87E-01	2.61E-02	1.35E-01	1.35E-01	1.89E-01	2.87E-04	3.47E-02	6.02E-03	2.21E-01	6.62E-02
S4	1.98E-01	1.48E-01	1.24E-01	1.50E-01	3.42E-02	1.66E-01	3.42E-02	2.00E-04	1.06E-01	3.82E-02
S5	2.10E-04	1.47E-01	1.41E-01	2.10E-04	9.89E-02	9.68E-02	1.77E-01	7.58E-02	1.93E-01	6.95E-02

In myPhyloDB, for samples above the normalization threshold λ is set to 0 (i.e., no adjustment of undetected OTUs; and for samples below the normalization threshold λ is set to 0.1 (as shown above). The final taxonomic profiles used for analysis are calculated for each sample by multiplying the calculated probabilities by the desired number of sequences.

In the text box provided you can enter “min”, “median”, “max”, or any integer desired.

There is also a checkbox to remove samples below the threshold. By default this box is unchecked and all selected samples will be included in your analysis using the above procedure.

7. Admin

You can access the administrative pages at “127.0.0.1:8000/myPhyloDB/admin”, where you can change the “superuser” or administrator username and password or add/remove authorized users. The default superuser for myPhyloDB is as follows:

username: admin
password: admin
email: admin@example.com

It is recommended that you change the default administrative username and password. To change the username click on the 'Users' link in the 'Auth' table. In the table, at the bottom of the next page click on the 'admin' username and change the username on the next page and press the 'Save' button at the bottom of the page. To change the password, click on the 'Change password' at the top-right of the page.

To add new authorized users click on the 'Add' link in the 'Auth' table. Add the desired username and password and press the 'Save' button at the bottom of the page. This user will now have access to the upload page and can add/remove projects from the myPhyloDB database.