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

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. If you are upgrading/reinstalling myPhyloDB and would like to keep your current database, make sure the "Default database" is unchecked during installation; otherwise, all components should be selected 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. Uninstalling myPhyloDB will not remove your database or uploaded files. The default installation folder for myPhyloDB will be: 'C:\Users\<user_name>\AppData\Local\myPhyloDB'.

Linux:

Double-click or run the installer from your terminal (myPhyloDB_1.0_Linux_x64_install.sh). If a previous version of myPhyloDB is detected you will be prompted to either keep your old database or re-install the default database. 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. MyPhyloDB must be manually uninstalled by deleting the appropriate folders. The default installation folder for myPhyloDB will be: 'home/<user_name>/myPhyloDB'.

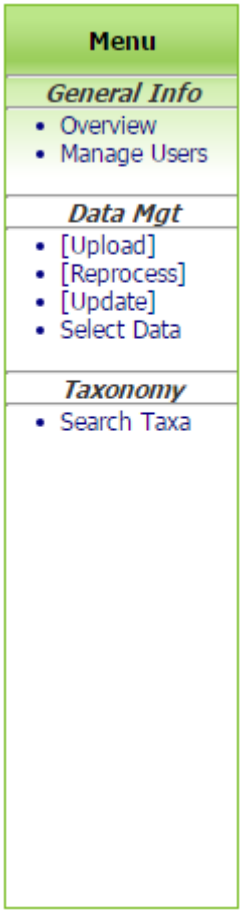
Remote access:

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' for Windows or 'ifconfig' for Linux), 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.

2. 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 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.



3.1 Uploading New Projects

To upload data, click "[Upload Data]" on the left hand menu (http://127.0.0.1:8000/myPhyloDB/upload/). For security purposes, this page can only be accessed by an authorized user – to add/remover users see the Admin section of this manual. For uploads, 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 the 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 throughout your database.

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.

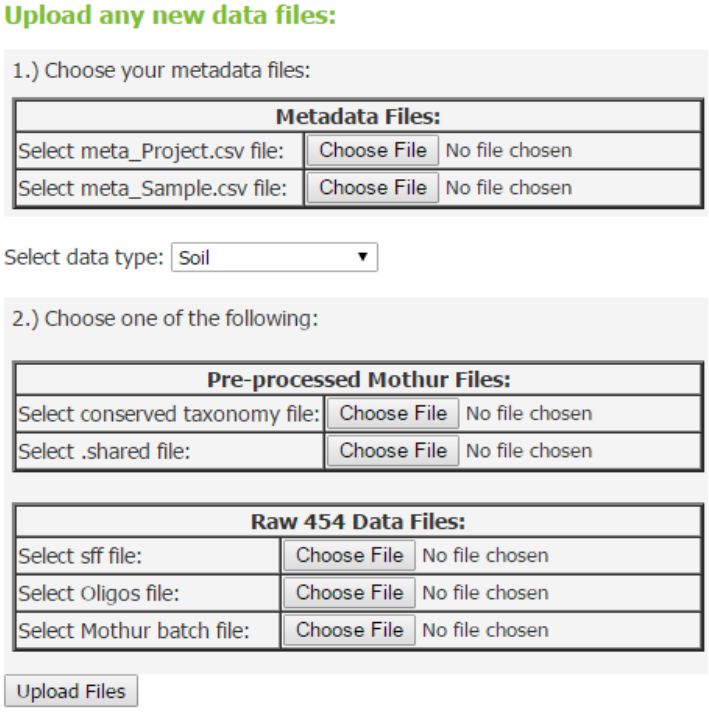
In addition, you will also need to create two files with the appropriate sequencing data similar to the mothur.shared and mothur.taxonomy examples provided, which are output from mothur (www.mothur.org). The shared file can be generated using the make.shared command but must contain only one OTU level (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)
```

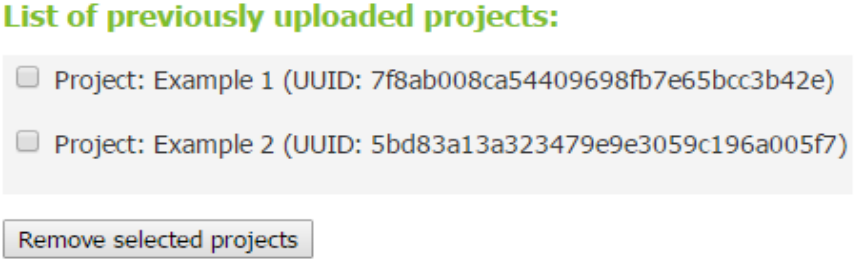
If you follow the above procedure the two files needed for upload will be named: "final.pds.wang.tx.shared" and "final.pds.wang.tx.1.cons.taxonomy". Due to taxa naming differences between the various reference databases (e.g., RDP, GreenGenes, SILVA), it is recommended that a single reference database 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).

Alternatively, you can upload raw data to myPhyloDB using the same format you would use for Mothur (since the data is then run through a local partial Mothur build). Note that you will still need to upload Project and Sample files, as well as the sff, Oligos, and Mothur batch files, for this method.

Once you have the files required for upload, add each file to the appropriate box located on the "[Upload Data]" page using the file selectors (see below). Use the drop down menu for "Select data type" to specify which type of data you are uploading . Note that data will not upload correctly if you specify the wrong type (i.e., does not match the sample file headers). Experienced Mothur users may wish to alter the provided batch file to match their current sequencing analysis pipelines; however, please note that the pipeline must create the following 5 files (final.fasta, final.names, final.groups, final.taxonomy, and final.shared); any deviation from the above naming conventions and the upload process will fail. When you are finished, click "Upload Files". Note: the upload process includes multiple steps and may take anywhere from a few minutes to hours depending upon the project size and your 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.

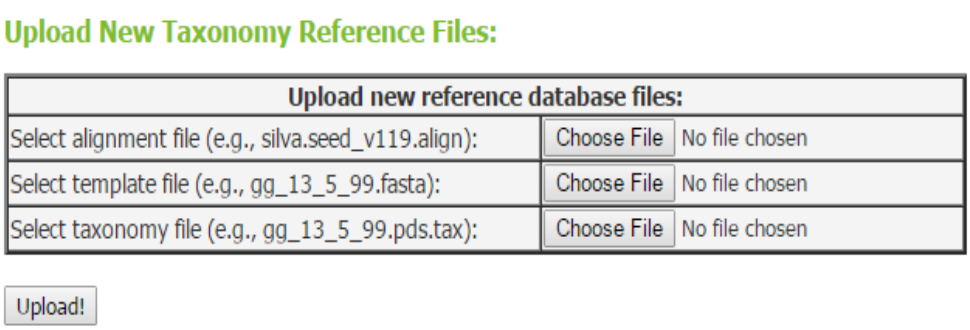


At the bottom of the "[Upload Data]" page is a list of current projects already uploaded to your myPhyloDB database. If you want to remove any of these projects simply click the appropriate box and then the "Remove selected projects" button. Edited projects (i.e., new submission files) can then be uploaded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

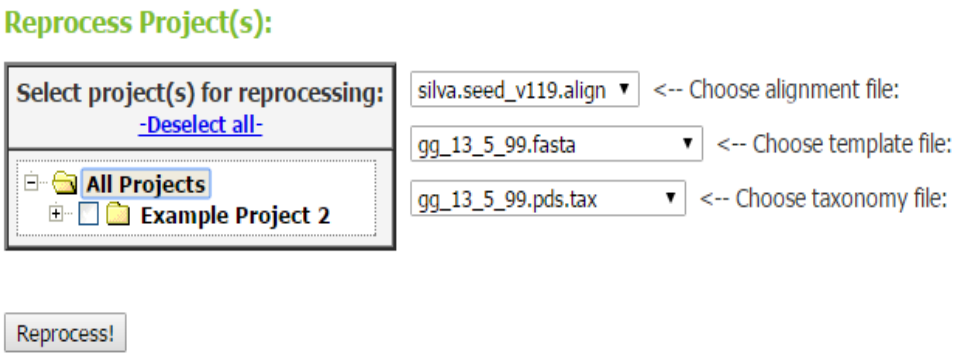


3.1 Reanalyzing old data

The alignment and classification files (i.e., template and taxonomy) can be conveniently updatesd for any project(s) containined in myPhyloDB.

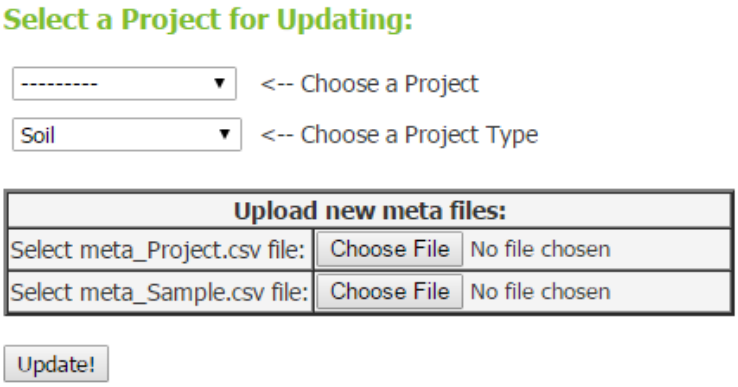


To do this, simply upload any new alignment, template, or taxonomic reference files using the [Reprocess] page. Next, select the projects which need to be updated in the project tree, and select the correct (updated) reference files from the drop down menus, then press "Reprocess!". Note: this will take anywhere from a few minutes to hours depending upon the project size and your computer speed



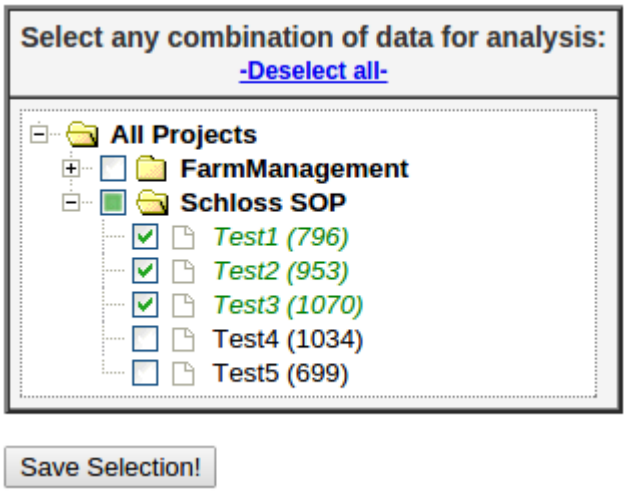
3.2 Updating metadata

To update a previously uploaded project with new sample or project metadata, click the [Update] button on the sidebar. Then, select the project you wish to have updated, as well as which project type you are updating to (which type of sample file you are using in the file chooser). Use the file choosers to select the new project and/or sample file(s) to be used for updating, then press "Update!". Note: the new project/sample files must contain the correct project and sample UUIDs for the updating procedure to correctly find and update the previously uploaded samples. The correct UUIDs can be obtained from the archived copy of these files (in myPhyloDB's "upload" folder) or the datatable found on the select data page.



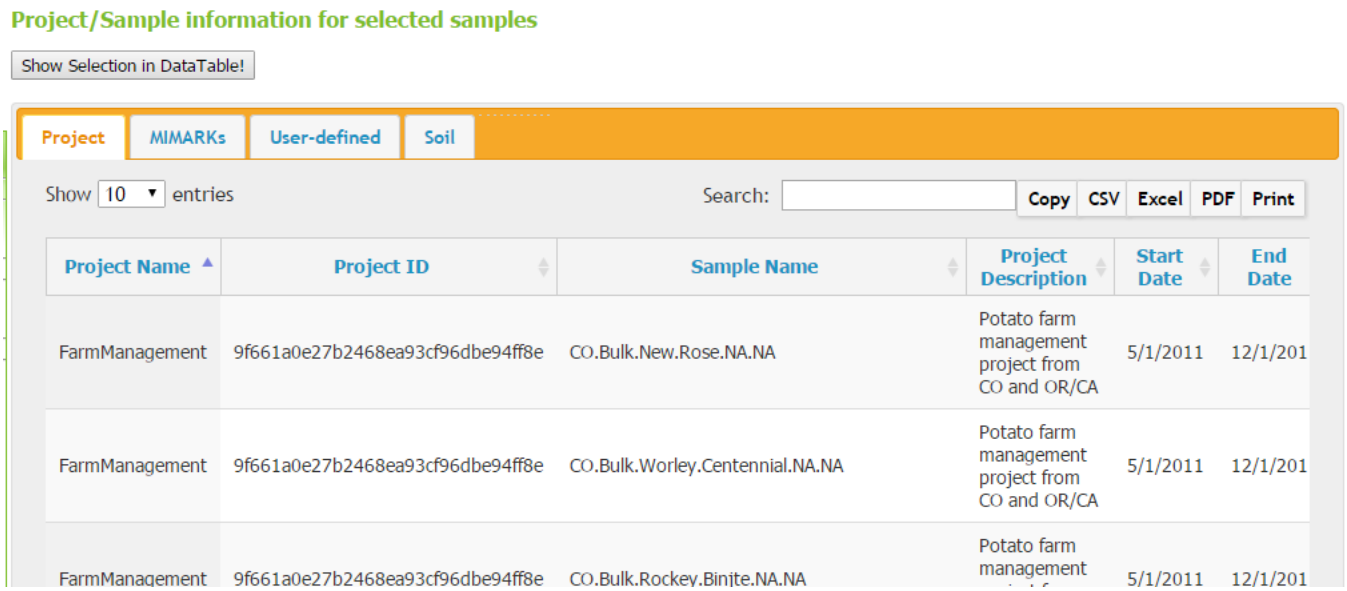
4. Selecting Data for Analysis

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.

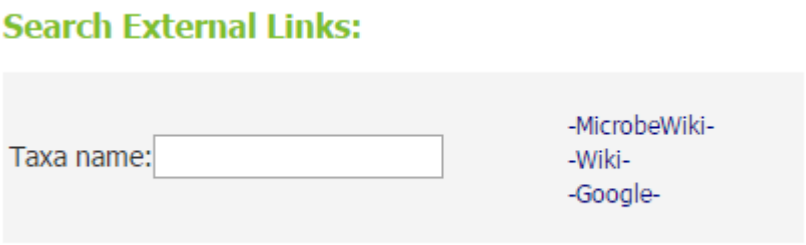
Any projects/samples selected above can be displayed in a datatable containing all of the metadata associated with each sample. Metadata is organized into nine different categories (Project, MIMARKS, Soil, Water, etc.). You may switch between these categories using the “Select datatable” drop-down box. The primary categories are static but load according to which data types have been uploaded (if you only have water and air samples uploaded, stuff like soil and human gut data will not be shown); 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.



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”.

5. 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 sites 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	feaf34ed21cf42c68534209baa1f51f0	Proteobacteria	22bf9fbd71fd44fca58e573fbd9f987d	Gammaproteobacteria
1	Bacteria	feaf34ed21cf42c68534209baa1f51f0	Proteobacteria	22bf9fbd71fd44fca58e573fbd9f987d	Deltaproteobacteria
2	Bacteria	feaf34ed21cf42c68534209baa1f51f0	Proteobacteria	22bf9fbd71fd44fca58e573fbd9f987d	Betaproteobacteria
3	Bacteria	feaf34ed21cf42c68534209baa1f51f0	Proteobacteria	22bf9fbd71fd44fca58e573fbd9f987d	Gammaproteobacteria
4	Bacteria	feaf34ed21cf42c68534209baa1f51f0	Proteobacteria	22bf9fbd71fd44fca58e573fbd9f987d	Alphaproteobacteria
5	Bacteria	feaf34ed21cf42c68534209baa1f51f0	Proteobacteria	22bf9fbd71fd44fca58e573fbd9f987d	Alphaproteobacteria
6	Bacteria	feaf34ed21cf42c68534209baa1f51f0	Proteobacteria	22bf9fbd71fd44fca58e573fbd9f987d	Gammaproteobacteria
7	Bacteria	feaf34ed21cf42c68534209baa1f51f0	Proteobacteria	22bf9fbd71fd44fca58e573fbd9f987d	Alphaproteobacteria
8	Bacteria	feaf34ed21cf42c68534209baa1f51f0	Acidobacteria	1fa0628580bd4615ad56701c82622145	Acidobacteriia
9	Bacteria	feaf34ed21cf42c68534209baa1f51f0	Bacteroidetes	c03f4a25bef84d3db03394559e08a6c6	Flavobacteriia

Showing 1 to 10 of 481 entries Previous 1 2 3 4 5 .. 49 Next

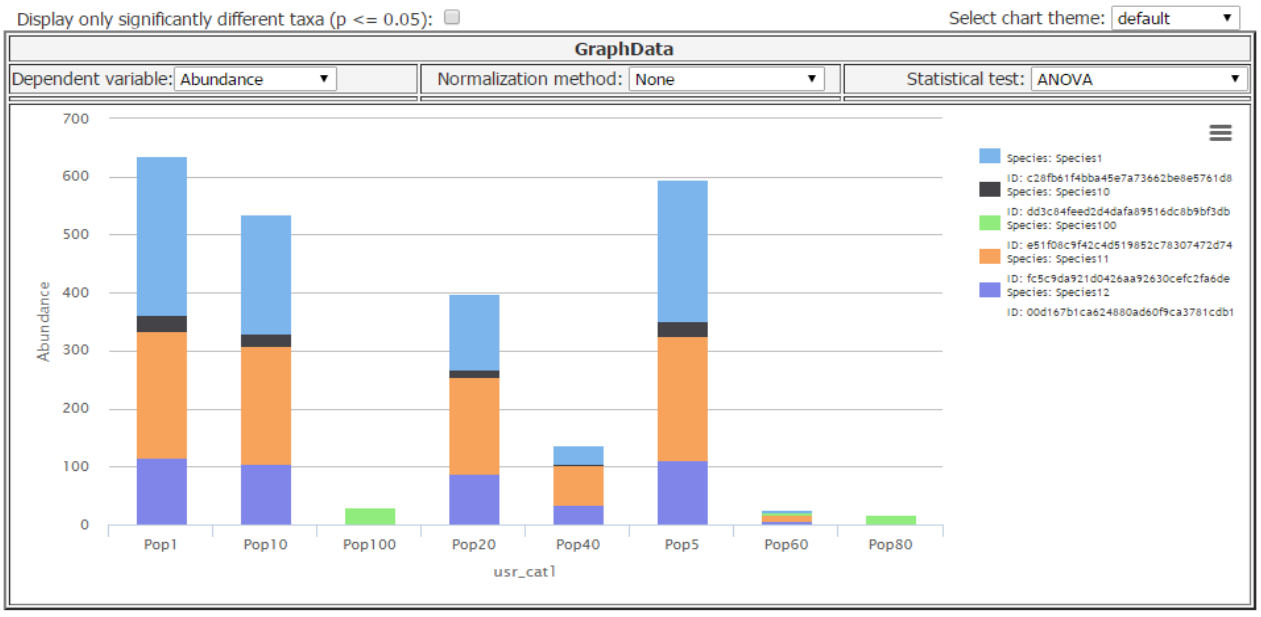
6. Analysis:

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, DiffAbund, or Multivariate: PCoA).

6.1. ANOVA/Regr:

For categorical variables, this analysis (http://127.0.0.1:8000/myPhyloDB/ANOVA/) will produce a bar chart and perform an ANOVA to test if the taxonomic level(s) chosen are significantly different between the meta-variables chosen. For quantitative variables, this analysis will produce a scatter plot and perform a linear regreation between the chosen taxonomic level(s) and a single meta-variable.

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 displayed in the graph (*e.g.* if you chose site and year, each site x year combination will be analyzed individually); however, the analysis will conduct a two-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 abundance (counts), species richness, or Shannon's Diversity Index for your dependent variable. Optionally you may also choose to display only significantly different taxa (checkbox, significance determined by p <= 0.05) or normalize your data before analysis (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.



The graph has several options, including dependent variable, normalization method, and statistical test.

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.

Test Results:

```
Data Normalization:
All 144 selected samples were included in the analysis...
No normalization was performed...
================================================


================================================
Taxa level: Species
Taxa name: Species1
Taxa ID: c28fb61f4bba45e7a73662be8e5761d8
Dependent Variable: Abundance
Independent Variable: usr_cat1

Anova: Single Factor on Measure

SUMMARY
Groups   Count   Sum    Average   Variance
=========================================
Pop1        18   4896       272   67878.588
Pop10       18   3673   204.056   41019.350
Pop100      18      2     0.111       0.105
```

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

Raw Data (Tabular):

Show 10 entries Search: Copy CSV Excel PDF Print

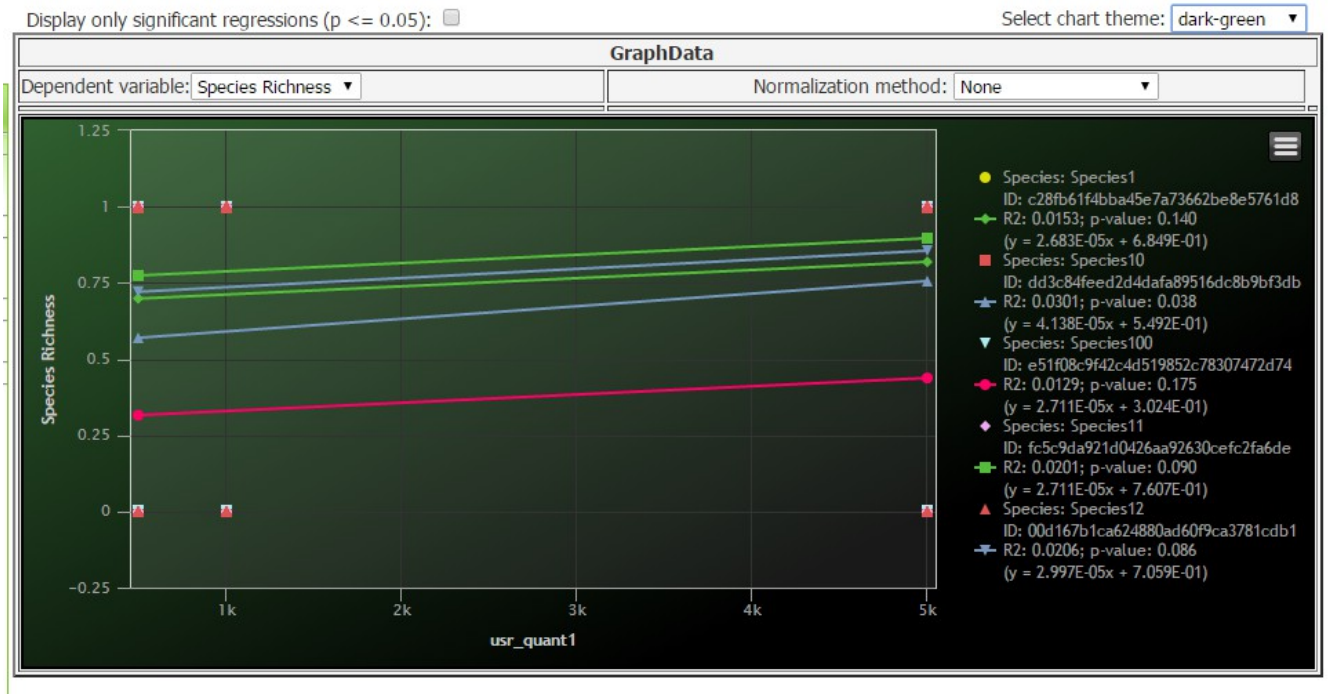
	sampleid	sample_name	usr_cat1	merge	rank	taxa_id
0	01d75698fec948d5b711aea9d79f62a4	Pop10b_1000	Pop10	Pop10	Species	00d167b1ca624880ad60f9ca3781cdb1
1	01d75698fec948d5b711aea9d79f62a4	Pop10b_1000	Pop10	Pop10	Species	00f7a3f1c68d45b4af05a802370824e2
2	01d75698fec948d5b711aea9d79f62a4	Pop10b_1000	Pop10	Pop10	Species	014fdcfaaac94074badb4f2af99b40a3
3	01d75698fec948d5b711aea9d79f62a4	Pop10b_1000	Pop10	Pop10	Species	017ac38c10b143a0bd58520eef377611
4	01d75698fec948d5b711aea9d79f62a4	Pop10b_1000	Pop10	Pop10	Species	032ebaca2ed3484da57fc270b04bd9e2
5	01d75698fec948d5b711aea9d79f62a4	Pop10b_1000	Pop10	Pop10	Species	0b833398280d4905b959d48c22c9ab8a
6	01d75698fec948d5b711aea9d79f62a4	Pop10b_1000	Pop10	Pop10	Species	1155f0f5dc5745d386d3c736fbd62ae9
7	01d75698fec948d5b711aea9d79f62a4	Pop10b_1000	Pop10	Pop10	Species	1197b2e39bc24135894ca71a98700013
8	01d75698fec948d5b711aea9d79f62a4	Pop10b_1000	Pop10	Pop10	Species	1221c1b544e645d7882e39cec7b2d662
9	01d75698fec948d5b711aea9d79f62a4	Pop10b_1000	Pop10	Pop10	Species	15fe580949de48ebbec7402650c364bb

Showing 1 to 10 of 14,400 entries Previous 1 2 3 4 5 .. 1440 Next

Raw Data (Biom Format):

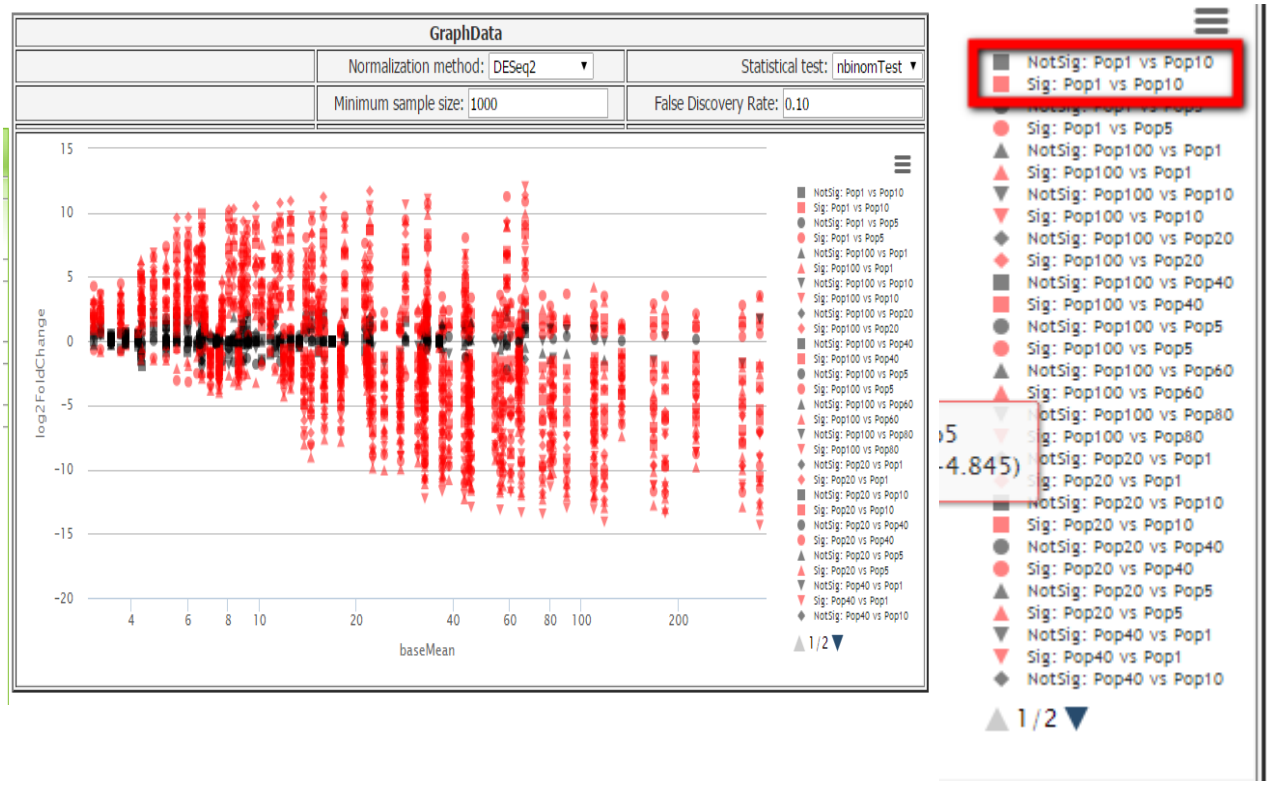

```
{
    "columns": [
        {
            "id": "01d75698fec948d5b711aea9d79f62a4",
            "metadata": {
                "sample_name": "Pop10b_1000",
                "usr_cat1": "Pop10"
            }
        },
        {
            "id": "031d403e9e06424cabaff81d7b07417f",
            "metadata": {
                "sample_name": "Pop40c_500",
                "usr_cat1": "Pop40"
            }
        },
        {
            "id": "045602c842224457ab2dd37e7314423c",
            "metadata": {
                "sample_name": "Pop10c_5000",
                "usr_cat1": "Pop10"
```


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 will be embedded in the figure legend.

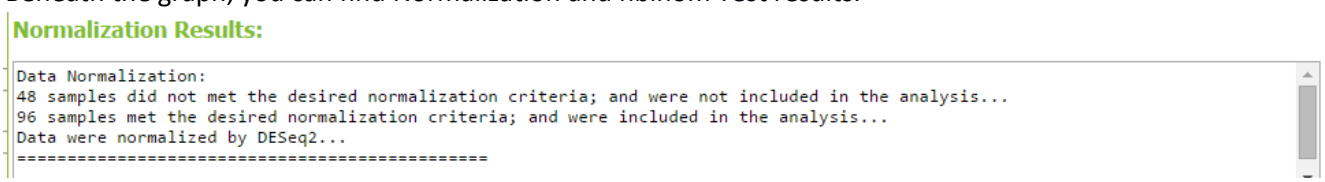


6.2 Diff Abund

This page page performs an univariate analysis with respect to the base mean and log2foldchange of data which has been normalized through DESeq2 (with or without filtering). More details on DESeq2 can be found here. Currently, this page uses the nbinom Test for significance. The graph has a logarithmic scale for the x-axis (baseMean) and a standard scale for the y-axis (log2FoldChange). All significant data points are plotted in red (significance being determined by a p value of <= 0.05). The plot shows a comparison of each sample with every other sample, with respect to your selected meta-variables. Each point on the graph is translucent, so the greater the number of points plotted in an area, the bolder the color there becomes. As with the other graphs in myPhyloDB, you can toggle individual data points on and off. For this graph, the data are sorted such that each comparison's significant and non significant data are next to each other, as well as the same shape. You can also mouse-over any point on the graph to see a tooltip description. Note that taxonomic data is selected by level with this page: this is so each taxonomic entry can be compared to each other entry of the same level.



Beneath the graph, you can find Normalization and nbinom Test results.



The nbinom test results are in the form of a sortable data table. You can search for specific samples to check their results, as well as export the table into CSV, excel spreadsheet, and PDF formats.

nbinomTest Results:

Show 10 entries Search: Copy CSV Excel PDF Print

	Comparison	Taxa ID	Taxa Name	baseMean	baseMeanA	baseMeanB	log2FoldChange
0	Pop100 vs Pop60	00d167b1ca624880ad60f9ca3781cdb1	Species12	168.370671	0.000000	1.797178	-4.078664e+00
0	Pop100 vs Pop80	00d167b1ca624880ad60f9ca3781cdb1	Species12	168.370671	0.000000	0.234936	-1.521171e+00
0	Pop100 vs Pop20	00d167b1ca624880ad60f9ca3781cdb1	Species12	168.370671	0.000000	77.532797	-9.639622e+00
0	Pop100 vs						

6.3 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

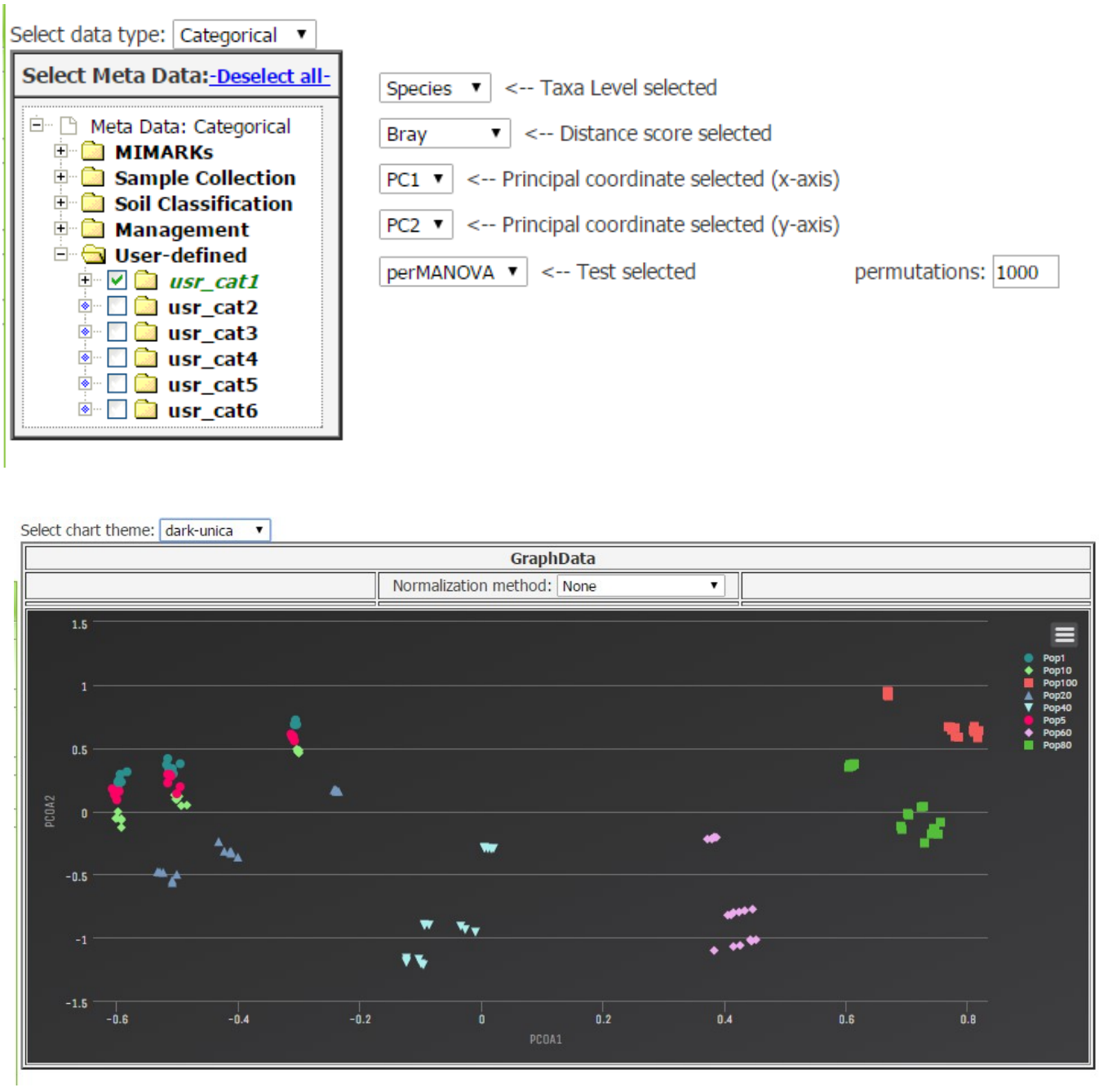
MorisitaHorn

The wOdum score allows users to down-weight either rare ($\alpha > 1$) or abundant ($\alpha < 1$) taxa, as discussed here (Manter and Bakker. 2015. BioInformatics. http://dx.doi.org/10.1093/bioinformatics/btv394). 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. These tables can be sorted based on the column of your choice. You can also search for specific samples via id, name, treatment type, etc. As with all tables of this type in myPhyloDB, you can export the table to CSV, Excel, and PDF formats (or just print it directly).

Principal Coordinates:

Show 10 entries Search: Copy CSV Excel PDF Print

	Sample ID	Sample Name	Treatment	CAP1	CAP2	CAP3	CAP4	CAP5
0	01d75698fec948d5b711aea9d79f62a4	Pop10b_1000	Pop10	-0.591818	-0.069552	-0.042707	1.946907	0.910315
1	031d403e9e06424cabaff81d7b07417f	Pop40c_500	Pop40	-0.033587	-0.911885	-1.095615	0.243751	-0.281814
2	045602c842224457ab2dd37e7314423c	Pop10c_5000	Pop10	-0.302311	0.472299	0.835390	-2.738526	-0.027244
3	067287a43c8e4566a923780cb6848405	Pop100b_500	Pop100	0.769270	0.668612	-0.298899	0.441058	-0.486473
4	079049ae46d04b7ab5d5ed0827f662d5	Pop40b_500	Pop40	-0.093447	-0.905639	-1.242338	0.223236	-0.468298
5	08dcdfa5489b409d95b848b2847db75a	Pop5f_1000	Pop5	-0.595248	0.158548	0.306454	1.323462	1.365773
6	0b7172ab4f7042ce84f1da7e0bf82cda	Pop1a_500	Pop1	-0.517421	0.369090	-0.456110	1.019794	-0.846424
7	0b7d69bb913a45aaa9e750ba54aca639	Pop5a_1000	Pop5	-0.606726	0.177442	0.268376	1.291429	1.189778
8	0c01d6c5d4b24228a3333e356160fe89	Pop5d_5000	Pop5	-0.306799	0.557224	1.032056	-3.142089	0.205265
9	0c167fc1447e422488b1c024bd69d2dc	Pop80a_1000	Pop80	0.755912	-0.088075	0.817406	2.588971	0.142846

Showing 1 to 10 of 144 entries Previous 1 2 3 4 5 .. 15 Next

Distance Scores:

Show 10 entries Search: Copy CSV Excel PDF Print

	Pop10b_1000	Pop40c_500	Pop10c_5000	Pop100b_500	Pop40b_500	Pop5f_1000	Pop1a_5(
Pop10b_1000	0.000000	0.550667	0.667000	0.958667	0.536000	0.155000	0.334667
Pop40c_500	0.550667	0.000000	0.836727	0.810000	0.204000	0.598667	0.620000
Pop10c_5000	0.667000	0.836727	0.000000	0.978182	0.830182	0.667000	0.818182
Pop100b_500	0.958667	0.810000	0.978182	0.000000	0.836000	0.973333	0.974000
Pop40b_500	0.536000	0.204000	0.830182	0.836000	0.000000	0.577333	0.586000
Pop5f_1000	0.155000	0.598667	0.667000	0.973333	0.577333	0.000000	0.345333
Pop1a_500	0.334667	0.620000	0.818182	0.974000	0.586000	0.345333	0.000000
Pop5a_1000	0.140000	0.612000	0.666667	0.976000	0.590667	0.109000	0.340000
Pop5d_5000	0.668000	0.842909	0.078000	0.981818	0.837455	0.667333	0.818182
Pop80a_1000	0.904000	0.697333	0.933333	0.362667	0.726667	0.915000	0.936000

Showing 1 to 10 of 144 entries Previous 1 2 3 4 5 .. 15 Next

7. Normalization

myPhyloDB provides several options for normalizing your sequence data to a common sampling depth.

None – no normalization

Rarefaction (Remove) – equivalent to the sub-sampling procedure in Mothur

Rarefaction (keep) – see below

Proportion – all abundances are divided by the total number of sequence reads for that sample

DESeq2 – A discussion can be found here.

7.1. Rarefaction (remove)

This normalization procedure performs a typical sub-sampling without replacement to the desired subsample size. Any sample, with fewer reads than the desired setting will be removed from the analysis. In the text box provided you can enter “min”, “median”, “max”, or any integer desired.

7.2. Rarefaction (keep)

This normalization procedure also performs a sub-sampling without replacement to the desired subsample size; however, it will keep all selected samples in the analysis regardless of the initial sample size . Aguirre de Cárcer et al. (Appl Environ Microbiol 2011 77:8795-8798) suggest that sub-sampling 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

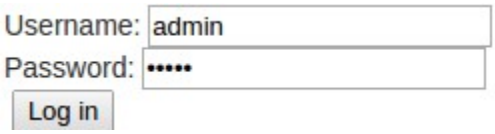
S2	1.38E-01	6.46E-02	8.96E-02	6.64E-02	9.32E-02	1.22E-01	1.27E-01	1.24E-01	2.16E-02	1.54E-01
S3	1.87E-01	2.61E-02	1.35E-01	1.35E-01	1.89E-01	2.87E-04	3.47E-02	6.02E-03	2.21E-01	6.62E-02
S4	1.98E-01	1.48E-01	1.24E-01	1.50E-01	3.42E-02	1.66E-01	3.42E-02	2.00E-04	1.06E-01	3.82E-02
S5	2.10E-04	1.47E-01	1.41E-01	2.10E-04	9.89E-02	9.68E-02	1.77E-01	7.58E-02	1.93E-01	6.95E-02

The final taxonomic profiles used for analysis are then calculated for each sample by randomly sampling the adjusted probability profiles until the desired number of sequence reads is achieved. In the text box provided you can enter “min”, “median”, “max”, or any integer desired.

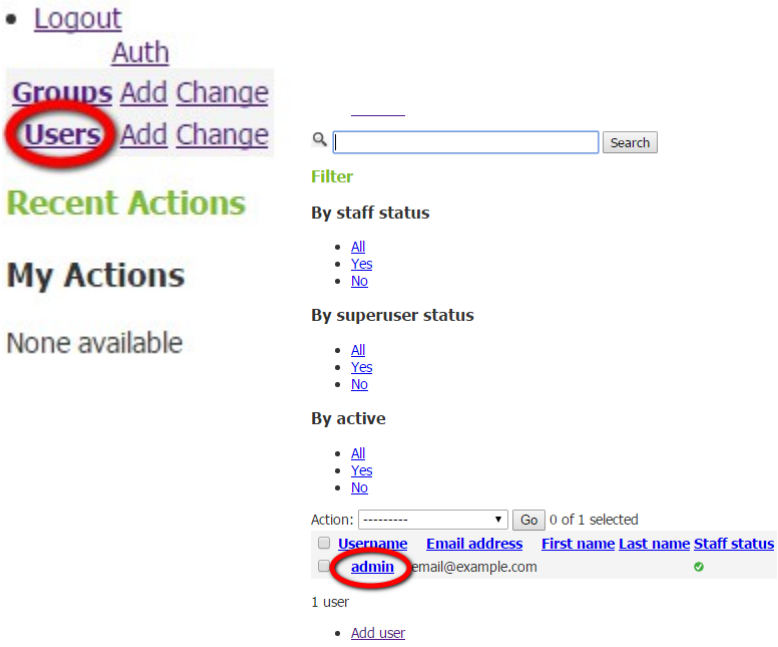
8. Admin

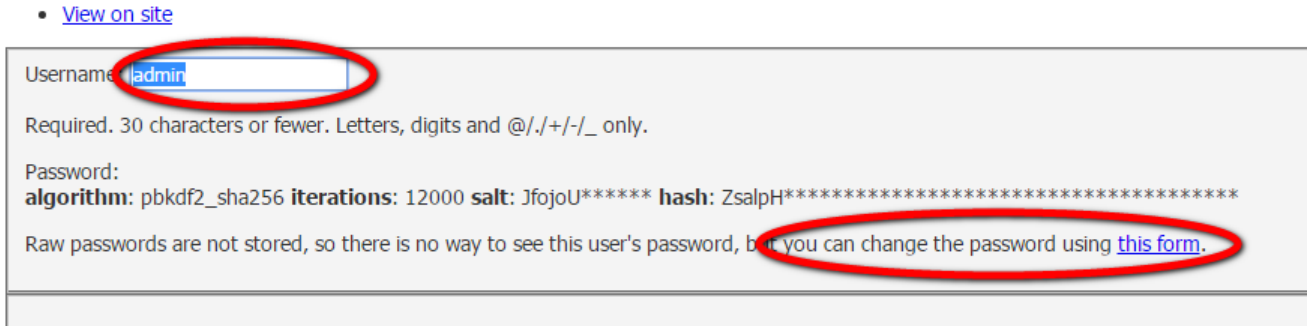
You can access the administrative pages at “127.0.0.1:8000/myPhyloDB/admin”, where you can change the “superuser” or administrator username and password or add/remove authorized users. The default superuser for myPhyloDB is as follows:

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

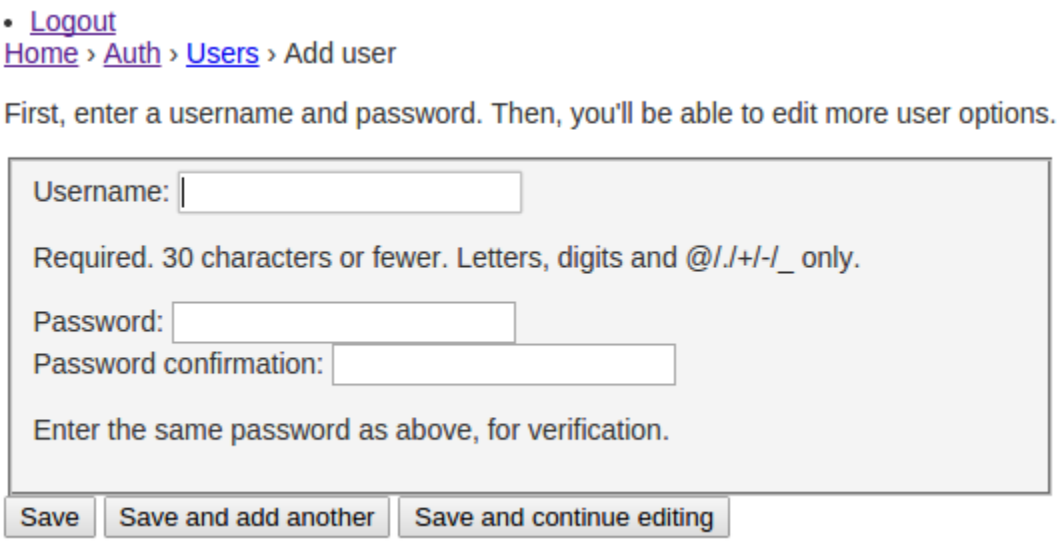


It is highly recommended that you change the default administrative username and password. To change the username click on the 'Users' link in the 'Auth' table. In the table, at the bottom of the next page click on the 'admin' username and change the username on the next page and press the 'Save' button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



There is also support for group permissions from the admin page, although currently these are fairly useless for anything but configuring another admin account.