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

What is myPhyloDB?

myPhyloDB is an open-source software package aimed at developing a user-friendly web-interface for accessing and analyzing all of your laboratory's microbial ecology data. The storage and handling capabilities of myPhyloDB archives users' raw sequencing files and allows for easy selection of any combination of project(s)/sample(s) from all of your projects using in the built-in SQL database. The data processing capabilities of myPhyloDB are also flexible enough to allow the upload, storage, and analysis of pre-processed data or raw (454 or Illumina) data files using the built-in versions of mothur and R.

Please visit our website for additional information and tutorials:
http://www.myphylodb.org

If you use myPhyloDB, please use the following citation:

Manter DK, M Korsa, C Tebbe, JA Delgado. 2016. myPhyloDB: a local web server for the storage and analysis of metagenomic data. *Database: The Journal of Biological Databases and Curation* 2016 : baw037. doi: 10.1093/database/baw037

Questions/comments (or requests for additional features) please visit our website or contact:

Daniel Manter
Soil Management and Sugar Beet Research Unit
USDA-ARS
Fort Collins, CO 80526
phone: (970) 492-7255

Table of Contents

1. Installation

myPhyloDB installers can be downloaded from the ARS website here. We strongly suggest users register when downloading this software so we can keep you informed of new updates and better track our user base to continue supporting myPhyloDB.

Windows:

Double-click the installer (myPhyloDB_v.1.2.0_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'. Changing this directory may cause some parts of myPhyloDB to become broken. Note that the AppData folder is not visible on some windows browsers, and you’ll need to manually enter your destination.

Linux:

Run the installer (myPhyloDB_v.1.2.0_Linux_x64_install.sh) from your terminal. Inside the terminal, navigate to the appropriate folder (in this example it is located in the 'Downloads' folder) and run the following command:

```
~/Downloads $ sh myPhyloDB_v.1.2.0_Linux_x64.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'. Changing this directory may cause some parts of myPhyloDB to become broken.

Mac Users:

Sorry, but we currently do not have Mac version. However, it may be possible to run myPhyloDB in a virtual environment (e.g., VirtualBox running Ubuntu 14.04 LTS). Although we have not tested this on a Mac, we have performed this successfully on a Windows 7 machine running Virtual Box 4.3.28 with a Linux Mint 17.2 (Ubuntu 14.04 base) installation. The installed version was myPhyloDB v.1.2.0.

Remote access:

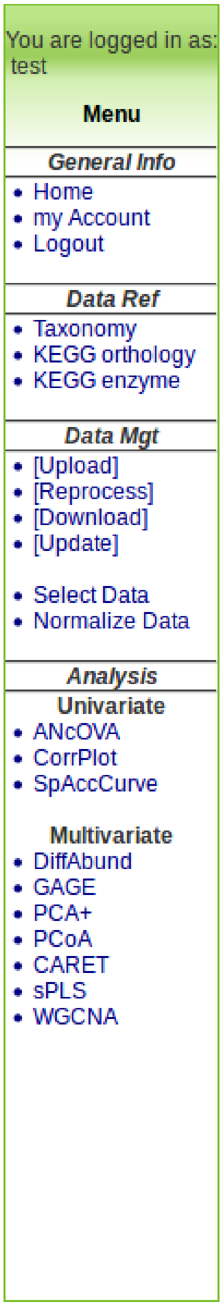
myPhyloDB will run as a local server on your host machine allowing others on your local intranet to access myPhyloDB (unless disabled using your computer's firewall settings) without installing a separate copy. This may be useful for laboratories that want to share data across multiple users. To access 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. Depending upon your local LAN/WAN setup, connection to the host machine may fail using a WiFi connection. If this happens, please try a wired connection to your LAN or contact your local IT staff. 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. General Information

2.1 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. The first time you launch myPhyloDB, the sidebar should look like the left panel below. From here you must login as a registered user (see Manage Users section) using the “Login” link to access any data in myPhyloDB. Once you log in as a registered user, the appropriate Data Mgt links to “upload”, “reprocess”, and “update” projects will become visible. Private projects can only be viewed or modified by the user (or superuser) who initially uploaded the project. Public projects can be viewed by all users; however, only the original user (or superuser) can modify that project.

<u>Dynamic nature of the Menu sidebar:</u>
The links available on the Menu sidebar are controlled by user type (superuser, registered user, or guest) and whether sample(s) have been selected.

General Info – For account management and general information

Data Reference – Contains pages which can search the database for relevant data

Data Mgt – Used for uploading and updating the contents of the database, as well as for downloading, normalizing, and otherwise selecting data for analysis

Analysis – Split into univariate and multivariate sections, each page performs a different task for analyzing data, as well as providing visual representation of certain data

2.2 My Account

Under My Account tab, you are able to change your password, and change or update your user profile. Under Project Management, all Public projects are displayed, as well as your private projects, where you are able to add users.

2.3 Manage Users

Superusers or staff accounts (accounts with elevated permissions) can access the user administration pages. When installing myPhyloDB for the first time, the default superuser login 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 admin username click on the 'Manage Users' link to enter the Administration pages. Next click on the ‘Users’ link in the 'Authentication and Authorization' 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.

Adding/removing or editing new users (i.e., change to staff status) is all performed using the appropriate “Add” “Change” buttons in the site.

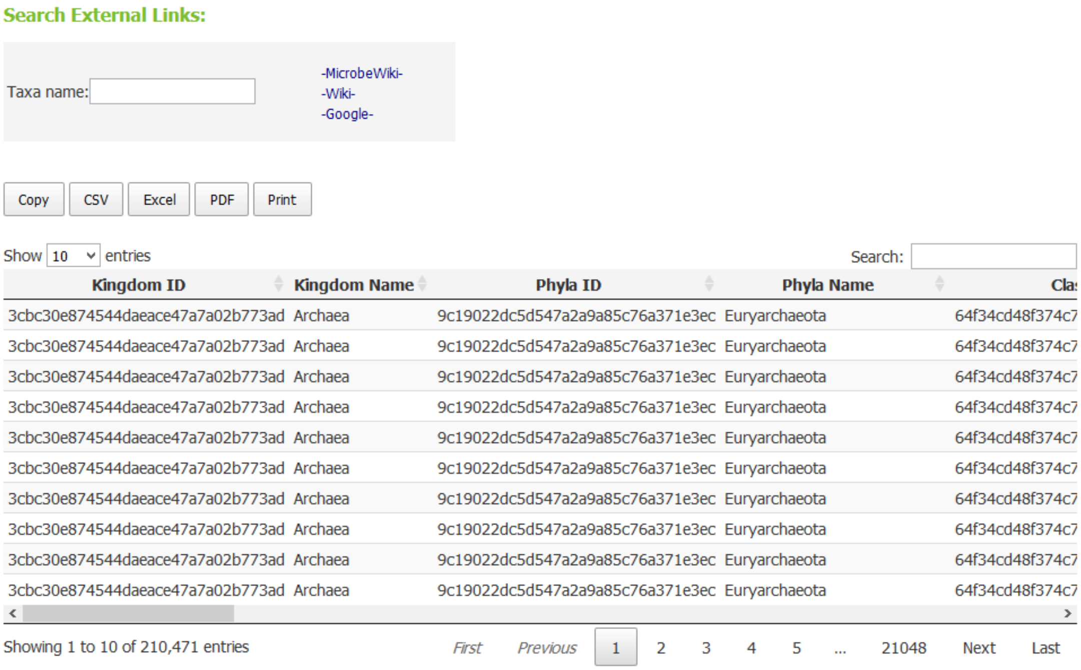
2.4 Login/Logout

Allows users to login and activate myPhyloDB's data upload and modify links. New users can register using the appropriate link on the login page. All fields (username, email, and password are required).

3. Data Reference

3.1 Taxonomy

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 copy, or export the table data to CSV, Excel, or PDF files, or send the data directly to a printer.

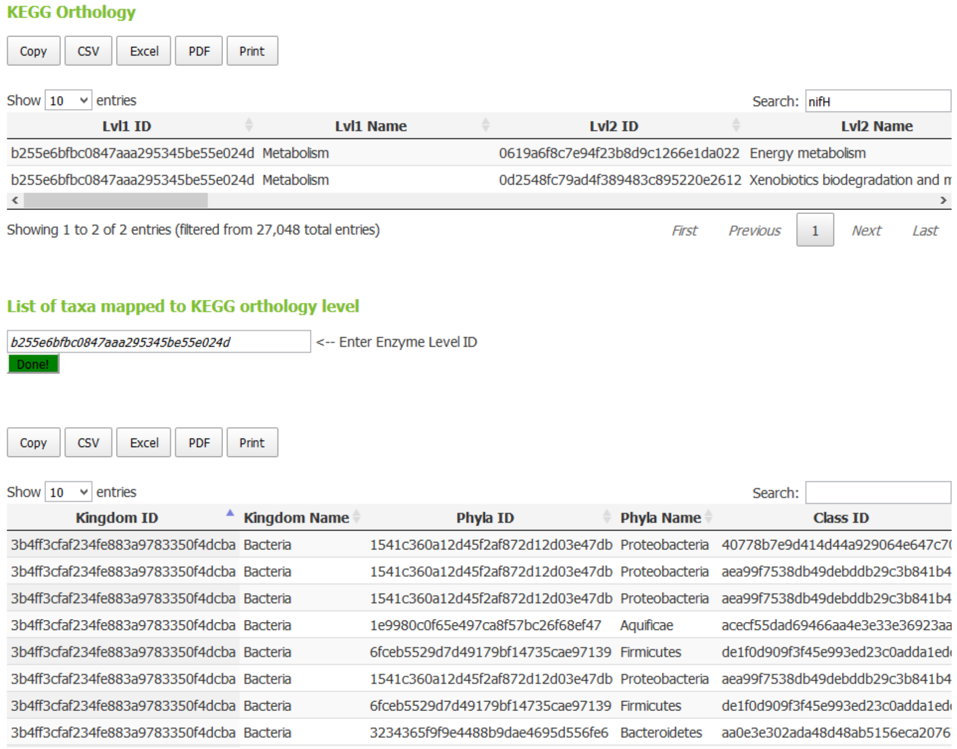


Kingdom ID	Kingdom Name	Phyla ID	Phyla Name	Cla
3cbc30e874544daeace47a7a02b773ad	Archaea	9c19022dc5d547a2a9a85c76a371e3ec	Euryarchaeota	64f34cd48f374c7
3cbc30e874544daeace47a7a02b773ad	Archaea	9c19022dc5d547a2a9a85c76a371e3ec	Euryarchaeota	64f34cd48f374c7
3cbc30e874544daeace47a7a02b773ad	Archaea	9c19022dc5d547a2a9a85c76a371e3ec	Euryarchaeota	64f34cd48f374c7
3cbc30e874544daeace47a7a02b773ad	Archaea	9c19022dc5d547a2a9a85c76a371e3ec	Euryarchaeota	64f34cd48f374c7
3cbc30e874544daeace47a7a02b773ad	Archaea	9c19022dc5d547a2a9a85c76a371e3ec	Euryarchaeota	64f34cd48f374c7
3cbc30e874544daeace47a7a02b773ad	Archaea	9c19022dc5d547a2a9a85c76a371e3ec	Euryarchaeota	64f34cd48f374c7
3cbc30e874544daeace47a7a02b773ad	Archaea	9c19022dc5d547a2a9a85c76a371e3ec	Euryarchaeota	64f34cd48f374c7
3cbc30e874544daeace47a7a02b773ad	Archaea	9c19022dc5d547a2a9a85c76a371e3ec	Euryarchaeota	64f34cd48f374c7
3cbc30e874544daeace47a7a02b773ad	Archaea	9c19022dc5d547a2a9a85c76a371e3ec	Euryarchaeota	64f34cd48f374c7
3cbc30e874544daeace47a7a02b773ad	Archaea	9c19022dc5d547a2a9a85c76a371e3ec	Euryarchaeota	64f34cd48f374c7

Showing 1 to 10 of 210,471 entries

3.2 KEGG Orthology and KEGG Enzyme

Similar to the taxonomy page, users are able to search the KEGG database for pathways (orthology) or enzymes. For each KEGG orthology or enzyme level a unique ID was generated by myPhyloDB for internal tracking purposes. Users also have the added option of exploring taxa mapped to specific KEGG enzyme or orthology levels, by copy and pasting the level ID generated in the above datatable. You can also copy, or export the table data to CSV, Excel, or PDF files, or send the data directly to a printer.



KEGG Orthology

Lvl1 ID	Lvl1 Name	Lvl2 ID	Lvl2 Name
b255e6bfbc0847aaa295345be55e024d	Metabolism	0619a6f8c7e94f23b8d9c1266e1da022	Energy metabolism
b255e6bfbc0847aaa295345be55e024d	Metabolism	0d2548fc79ad4f389483c895220e2612	Xenobiotics biodegradation and m

Showing 1 to 2 of 2 entries (filtered from 27,048 total entries)

List of taxa mapped to KEGG orthology level

b255e6bfbc0847aaa295345be55e024d <-- Enter Enzyme Level ID

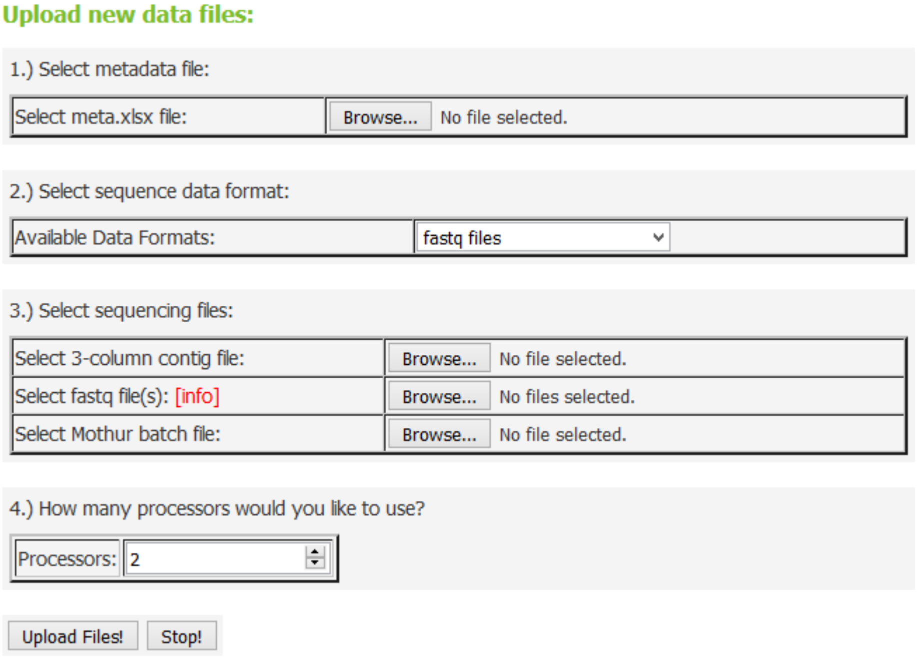
Kingdom ID	Kingdom Name	Phyla ID	Phyla Name	Class ID
3b4ff3cfaf234fe883a9783350f4dcba	Bacteria	1541c360a12d45f2af872d12d03e47db	Proteobacteria	40778b7e9d414d44a929064e647c7
3b4ff3cfaf234fe883a9783350f4dcba	Bacteria	1541c360a12d45f2af872d12d03e47db	Proteobacteria	aea99f7538db49debddb29c3b841b4
3b4ff3cfaf234fe883a9783350f4dcba	Bacteria	1541c360a12d45f2af872d12d03e47db	Proteobacteria	aea99f7538db49debddb29c3b841b4
3b4ff3cfaf234fe883a9783350f4dcba	Bacteria	1e9980c0f65e497ca8f57bc26f68ef47	Aquificae	acecf55dad69466aa4e3e33e36923aa
3b4ff3cfaf234fe883a9783350f4dcba	Bacteria	6fceb5529d7d49179bf14735cae97139	Firmicutes	de1f0d909f3f45e993ed23c0adda1ed
3b4ff3cfaf234fe883a9783350f4dcba	Bacteria	1541c360a12d45f2af872d12d03e47db	Proteobacteria	aea99f7538db49debddb29c3b841b4
3b4ff3cfaf234fe883a9783350f4dcba	Bacteria	6fceb5529d7d49179bf14735cae97139	Firmicutes	de1f0d909f3f45e993ed23c0adda1ed
3b4ff3cfaf234fe883a9783350f4dcba	Bacteria	3234365f9f9e4488b9dae4695d556fe6	Bacteroidetes	aa0e3e302ada48d48ab5156eca2076

4. Data Management

4.1 Upload: Uploading New Data

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/remove users see the Admin section of this manual. Uploading new data consists of 3 steps: 1) selecting your metadata file, 2) selecting your sequence data file format, and 3) selecting your sequencing files. All metadata is uploaded using a single excel file.

Also note that uploads, along with reprocessing and updating of projects, are on a single queue (meaning only one of these three can occur at a time). You can queue multiple uploads, which will be processed one at a time.



Upload new data files:

1.) Select metadata file:

Select meta.xlsx file: Browse... No file selected.

2.) Select sequence data format:

Available Data Formats: fastq files

3.) Select sequencing files:

Select 3-column contig file: Browse... No file selected.

Select fastq file(s): [info] Browse... No files selected.

Select Mothur batch file: Browse... No file selected.

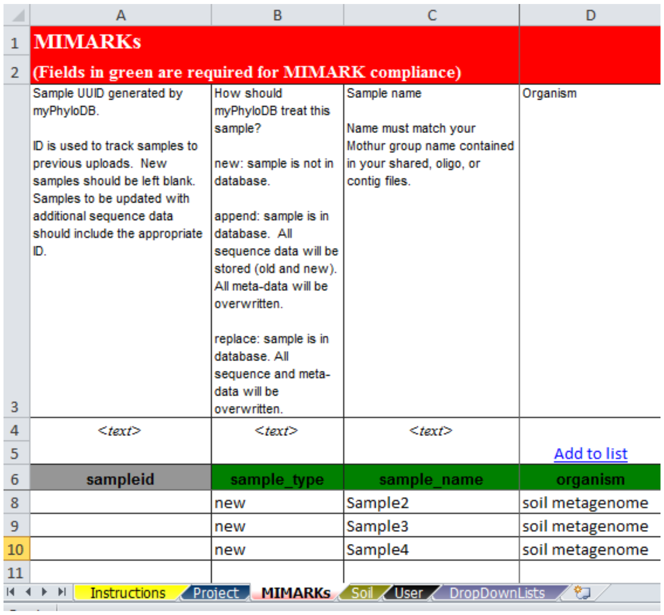
4.) How many processors would you like to use?

Processors: 2

2.1.1 The metadata file

Each upload requires a completed metadata file, which can be downloaded from myPhyloDB's homepage. The metafile has several formulas and protections that cannot be changed, otherwise myPhyloDB will not recognize the file.

Any column with blue “Add to list” link contains a drop-down feature to enter data into each cell from pre-defined word(s) or numbers. You may add your own text/numbers to use under the ‘DropDownLists’ tab. While filling in each cell, you can simply type in your value without clicking the drop-down tab.



	A	B	C	D
1	MIMARKs			
2	(Fields in green are required for MIMARK compliance)			
3	Sample UUID generated by myPhyloDB. ID is used to track samples to previous uploads. New samples should be left blank. Samples to be updated with additional sequence data should include the appropriate ID.	How should myPhyloDB treat this sample? new: sample is not in database. append: sample is in database. All sequence data will be stored (old and new). All meta-data will be overwritten. replace: sample is in database. All sequence and meta-data will be overwritten.	Sample name Name must match your Mothur group name contained in your shared, oligo, or contig files.	Organism
4	<text>	<text>	<text>	
5				Add to list
6	sampleid	sample_type	sample_name	organism
8		new	Sample2	soil metagenome
9		new	Sample3	soil metagenome
10		new	Sample4	soil metagenome
11				

Column (variables) names (such as cat_1) must not be changed and additional instructions for using the excel template file are contained within. Please note that myPhyloDB does not perform any unit checking or data conversions, so consistent units should be used for all projects throughout your database.

2.1.2 Project tab

myPhyloDB currently supports five different project types (Soil, Air, Water, Microbial, and Human-associated). Each project type supports a different set of default variables, based on those outlined here (http://www.mothur.org/wiki/MIMarks_Data_Packages). Please note that the following MIMARK

fields (seq_method, geo_loc_name, and lat_lon) have been replaced by multiple single-entry fields. For example, (1) seq_method is replaced with seq_platform, seq_gen, seq_gen_region, seq_for_primer, and seq_rev_primer; (2) geo_loc_name is replaced with geo_loc_country, geo_loc_state, geo_loc_city, geo_loc_farm, and geo_loc_plot; and (3) lat_lon is replaced with latitude and longitude.
Projects can be tagged as public or private using the "status" column in the "Project" tab of the Excel file. Private projects can only be viewed or modified by the original user (i.e., user logged in at the time of upload) or the project superuser. Public projects can be viewed by all registered users (and guests); however, only users designated by the original user can modify the project.

At minimum, you must enter a project name, project type and designate it as public or private. You may also enter in any other project information DO NOT enter any information in column A 'num_samp' or column D 'projectid', as these will be generated when uploaded into myPhyloDB.

2.1.3 MIMARKs tab

The current meta data excel file (e.g., myPhyloDB.Soil.meta.xlsx) provides suggested controlled vocabulary lists and units for each defined variable, or Minimum Information about a Marker gene sequence (MIMARKs). However, users are free to modify these lists and use any units desired. For additional vocabulary/data consideration you may wish to consult the Yilmaz et al. 2011 paper doi:10.1038/nbt.1823.

Do not enter any data into the first column labeled 'sampleid.' Unique sample ID's for each of your samples will be generated in myPhyloDB. Sample names are entered on the 3rd column labeled 'sample_name' on the MIMARKs page only. Enter your sample names exactly as they appear in your shared, oligo or contig file. Several columns use drop-down tabs; text entered here must match text defined under 'DropDownLists', were you may enter your own variables. If you enter text that is not also in the DropDownLists, myPhyloDB will not recognize your data.

2.1.4 Soil/Human associated/Air/Microbial/Water tab

On this page you are able to enter in metadata about your samples. The 'sample_name' column will automatically generate from your MIMARKs sample name entry. Where appropriate, use drop-down tabs with values defined under 'DropDownLists' tab.

2.1.5 User tab

You may add any other categorical or quantitative variables not already included in the spreadsheet. Do not edit row 6 ; these variables will appear in myPhyloDB as 'usr_cat1', 'usr_cat2' etc., and cannot be customized.

2.1.6 DropDownLists tab

Any text or values entered into columns with integrated drop-down option (, must first be defined under the 'DropDownLists.' Several common definitions are already included on the list, but you can add your own custom definitions.

Only one project can be uploaded at a time; however, samples (new 'sample_name') may be added to an already uploaded project by setting the 'project_id' (under Project tab) to the auto-generated 32-character alphanumeric UUID found in the DataTable located on the "Select Data" page of myPhyloDB. Similarly, you may add new sequence data to an existing sample by setting both the 'project_id' (under Project tab) and 'sample_id' (under MIMARKs tab) values to the appropriate auto-generated UUIDs. This is the only situation where you will manually enter values in these columns. After a project is uploaded into myPhyloDB, metadata files can be downloaded, amended, and then reprocessed using the tabs under *Data Mgt* in the sidebar.

2.1.7 Select your sequence data format

myPhyloDB supports the upload of 1) pre-processed mothur data files, 2) raw 454 pyrosequencing files and 3) raw Illumina fastq or fna/qual data files. The files required for submission will change depending upon your selection.

2.1.8 Batch File

The batch file consists of line code to run mothur processing. The batch file can be modified.
For most users this will consist of only changing some of the parameters associated with each step in the provided sequence processing pipeline (e.g. batch files). For most steps, we have identified some of the more common settings that may be altered. These are listed as “tunable parameters” in the line preceding each step. Additional settings are frequently possible for each step and the user should consult the mothur website for more information.

Experienced mothur users may wish to further alter the provided batch files to match their current sequencing analysis pipelines (add/remove steps); however, please note that the pipeline must create the following 5 files:

final.fasta	#fasta file containing your sequences
final.names	#mothur name file
final.groups	#mothur group file
final.taxonomy	#consensus taxonomy for each phylotype (OTU) file
final.shared	#mothur shared file

Any deviation from the above naming conventions will cause myPhyloDB's upload process to fail.

Phylotype vs OTU based analysis

myPhyloDB stores all data in its database by phylotype (i.e., taxonomic names) meaning that all analysis performed through its GUI interface are based on phylotypes. This is driven, in part, due to potential differences in sequencing information (i.e., gene sequenced, PCR primers utilized, read length, etc.); and the difficulties in defining and curating a systematic naming convention for operational taxonomic units (OTUs) based on genetic distance across projects. In addition, only the seven major taxonomic classifications (i.e., Kingdom, Phyla, Class, Order, Family, Genus, and Species) plus one additional level (e.g., OTU 99%). For example, myPhyloDB is shipped with a modified version of the Greengenes May 2013 release (gg_13_5_99) that contains 202,241 representative sequences (OTUs defined at the 99% genetic similarity). For each entry in the database, the Greengenes reference ID was added as the eighth taxonomic level (e.g., OTU99).

2.3 Example uploads with the 4 different sequence file types

Example 1: Pre-processed mothur files

Sample files to upload a pre-processed mothur project can be found on myPhyloDB's homepage (Example1.tar.gz). This option allows users to upload files that have already been processed using mothur. To use this option, you will need two mothur-generated files: *.shared and *.cons.taxonomy. The shared file can be generated using the make.shared command. The taxonomy file can be generated using the classify.otu command 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=current, name=current, template= gg_13_5_99.fasta, taxonomy= gg_13_5_99.dkm.tax,
method=knn, numwanted=1, processors=X)

phylotype(taxonomy=current, name=current, label=1)


make.shared(list=current, group=current)

classify.otu(taxonomy=current, name=current, group=current, list=current)
```

If you follow the above procedure the two files needed for upload will be named: “final.tx.shared” and “final.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). You may change the reference database after your project is uploaded by using the Update tab on the side bar.

Once you have created the above files you will perform the following steps to upload Example 1.
Step 1. Click on the “Browse” button under the heading “1) Select metadata file:” and navigate to your folder containing the “Example1.Soil.meta.xls” file. Click on the file and select open.
Step 2. In the dropdown box below “2) Select sequence data format:” make sure that “Pre-processed mothur files” is selected (i.e., visible).
Step 3. Under the “3) Select sequencing files:” heading select the taxonomy (Example1.taxonomy) and shared (Example1.shared) files like you did in step 1 for your metadata file.
Step 4. Click on the “Upload Files” button.

A message and progress bar should now be displayed reporting myPhyloDB's progress on uploading and parsing your data. Note: the progress bar may pause for several seconds during the “Parsing sample file” step at 50% which is normal.

Uploading Raw Sequencing Data (Examples 2-4).

Examples 2-4 all utilize the myPhyloDB's embedded copy of mothur for sequence processing. In order to allow mothur to utilize its mulit-processing capabilities, an input box is also provided for users to specify the number of processors available on their machine (step 4 on upload page). myPhyloDB will automatically limit this value between 1 and x, where x is your number of available processors. For example, if you have an Intel i7 processor, which has 6 logical processors on 2 cores, you will may set this to 12. If you set this value too high, myPhyloDB is smart enough to reset this value to 12. Experienced mothur users, who have examined our supplied batch files, will notice that some commands contain a "processors=X" setting. This is on purpose, as the X will automatically be replaced using the setting above.

Imported files (batch, oligo, contig, etc) must be in proper mothur formatting. For more information visit https://mothur.org/wiki.

Example 2: Raw 454 sff file(s)

Sample files to upload a raw 454 pyrosequencing (sff files) project can be found on myPhyloDB's homepage (Example2.tar.gz). Using this option will utilize myPhyloDB's embedded copy of mothur (currently v.1.35.1) and allow for future reprocessing using myPhyloDB. For this option, you will need to upload four files: sff file(s) (standard 454 flow files), filenames file (file containing the names of the sff files you would like to process), oligo file(s) (file with sample names, barcodes, and primers), and a mothur batch file. The proper settings to upload Example 2 are as follows:

Step 1. Click on the "Browse" button under the heading "1) Select metadata file:" and navigate to your folder containing the "Example2.Soil.meta.xls" file. Click on the file and select open.
Step 2. In the dropdown box below "2) Select sequence data format:" make sure that "sff files" is selected (i.e., visible).
Step 3. Under the '3) Select sequencing files:' heading select the following files like you did in step 1.

sff files → 90.1.sff
90.2.sff
90.3.sff
90.4.sff
90.5.sff

oligo files → 90.1.oligos
90.2.oligos
90.3.oligos
90.4.oligos
90.5.oligos

filenames file → sff_files.txt

mothur batch file → Exanple2.mothur.batch

Step 4: Enter the number of processors you'd like to use. Value can be entered manually or incremented using the up and down arrows.
Step 5. Click on the "Upload Files" button.

Example 3: Raw fna/qual files

Sample files to upload a raw fna/qual files project can be found on myPhyloDB's homepage

(Example3.tar.gz). Using this option will utilize myPhyloDB's embedded copy of mothur (currently v.1.35.1) and allow for future reprocessing using myPhyloDB. For this option, you will need to upload four files: fna file (standard fasta files), qual file (read qualtity file), oligo file (file with sample names, barcodes, and primers), and a mothur batch file. The proper settings to upload Example 3 are as follows:

Step 1: Click on the "Browse" button under the heading "1) Select metadata file:" and navigate to your folder containing the "Example3.Soil.meta.xls" file. Click on the file and select open.
Step 2: In the dropdown box below "2) Select sequence data format:" make sure that "fna/qual files" is selected (i.e., visible).
Step 3: Under the "3) Select sequencing files:" heading select the following files like you did in step 1.

fna files → Example3a.fna
Example3b.fna
Example3c.fna
Example3d.fna

qual files → Example3a.qual
Example3b.qual
Example3c.qual
Example3d.qual

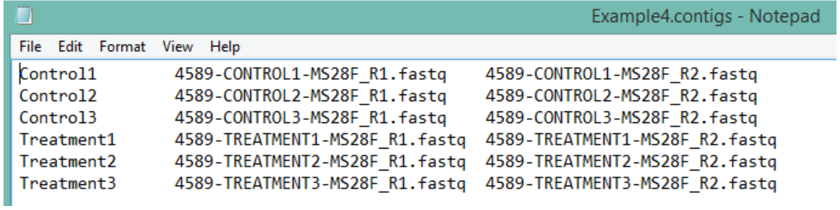
oligos file → Example3.oligos

mothur batch file → Exanple3.mothur.batch

Step 4: Enter the number of processors you'd like to use. Value can be entered manually or incremented using the up and down arrows.
Step 5: Click on the "Upload Files" button.

Example 4: Illumina/MiSeq .fastq files

All of the files necessary to upload a sample raw Illumina/MiSeq project can be found on myPhyloDB's homepage (Example4.tar.gz). Using this option will utilize myPhyloDB's embedded copy of mothur (currently v.1.35.1) and allow for future reprocessing using myPhyloDB.
For this option, you will need to upload the following files: 3-column contig file (file with sample names and fastq file names), fastq files (forward and reverse for each sample), and a mothur batch file. The contig file must be a text file with sample name, forward fastq file, then reverse fastq file all tab deliminated. Be aware that sequencing platforms designate forward and reverse fastq filenames differently, but typically "F" and "R", or "R1" and "R2" are used.



```
Control1      4589-CONTROL1-MS28F_R1.fastq    4589-CONTROL1-MS28F_R2.fastq
Control2      4589-CONTROL2-MS28F_R1.fastq    4589-CONTROL2-MS28F_R2.fastq
Control3      4589-CONTROL3-MS28F_R1.fastq    4589-CONTROL3-MS28F_R2.fastq
Treatment1    4589-TREATMENT1-MS28F_R1.fastq  4589-TREATMENT1-MS28F_R2.fastq
Treatment2    4589-TREATMENT2-MS28F_R1.fastq  4589-TREATMENT2-MS28F_R2.fastq
Treatment3    4589-TREATMENT3-MS28F_R1.fastq  4589-TREATMENT3-MS28F_R2.fastq
```

Please note, that the current default pipeline only supports the 3-column config file option and processing of fastq files that have had their barcode/primers removed (i.e., sorted). The proper settings to upload

Example 4 are as follows:

Step 1: Click on the “Browse” button under the heading “1) Select metadata file:” and navigate to your folder containing the “Example4.meta.xls” file. Click on the file and select open.
Step 2: In the dropdown box below “2) Select sequence data format:” make sure that “fastq files” is selected (i.e., visible).
Step 3: Under the “3) Select sequencing files:” heading select the following files like you did in step 1.

3-column contig file	→ Example4.contigs
fastq files	→ Example4.fastq.tar.gz
mothur batch file	→ Exanple4.mothur.batch

Step 4: Enter the number of processors you'd like to use. Value can be entered manually or incremented using the up and down arrows.
Step 5: Click on the “Upload Files” button.

3.1 Upload Benchmarks

The sequence processing and uploading 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. As stated above, the progress bar may pause for several seconds during the “Parsing sample file” step at 50%, this is normal. Also, the progress bar is inactive during sequence processing (i.e., when mothur is running); however, mothur will output its progress to the text box below the progress bar, as well as your host computer's terminal. The following is an example of the times required to upload the 4 example projects.

Project	**Procedure**	**Test computer**	**Time (hr:min:sec)**
Example 1	Uploading and parsing	Linux[1]	0:00:31
		Windows[2]	0:01:04
Example 2	Sequencing processing, uploading and parsing	Linux[1]	0:23:48
		Windows[2]	1:28:50*
Example 3	Sequencing processing, uploading and parsing	Linux[1]	0:08:34
		Windows[2]	0:17:08
Example 4	Sequencing processing, uploading and parsing	Linux[1]	1:08:44
		Windows[2]	1:21:15

[1] Computer configuration: Linux Mint 17.2 LTS, 32 GB RAM, i7-5930K @ 3.5 GHz
[2] Computer configuration: Windows 7 Pro, 8 GB RAM, i7-4790 @ 3.6 GHz
*multi-processing not implemented for all mothur functions (e.g., sff.mulitple) in Windows

File Storage

In addition to a providing a searchable database for selecting and analyzing your data, myPhyloDB also helps to organize all of your raw (and processed) sequencing files. For example, all of the raw data files and the 5 mothur-processed data files (final.fasta, final.names, final.groups, final.taxonomy, final.shared) will be copied and stored in the “uploads” folder of myPhyloDB. The path to each uploaded project can be found in the datatable under the “Reference” tab located on the “Select Data” page.

Removing Data from your myPhyloDB Database

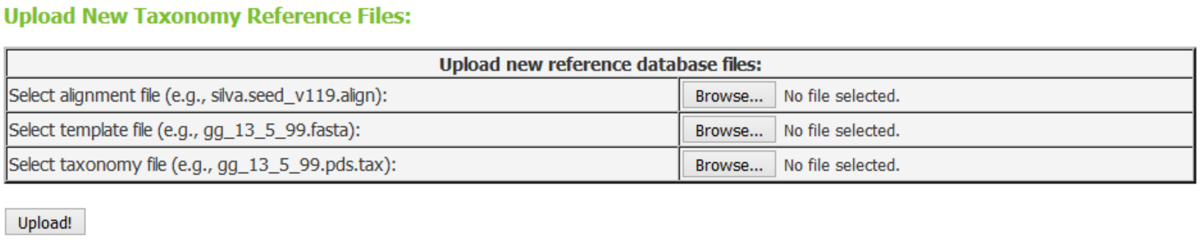
At the bottom of the “Upload” page is a list of all previous uploads to your myPhyloDB database. Each item in the list is categorized by project name and the upload path, which contains the timestamp when the upload was submitted. If you want to remove any of these uploads simply click the appropriate box and then the “Remove selected” button. Please use caution as this will not only remove the project from you database but also the archived copies of the raw and processed data in your “uploads” folder.

4.2 Reprocess: Reanalyzing Data in your myPhyloDB Database

New alignment and classification files (i.e., template and taxonomy) can be conveniently updated for any project(s) contained in myPhyloDB.

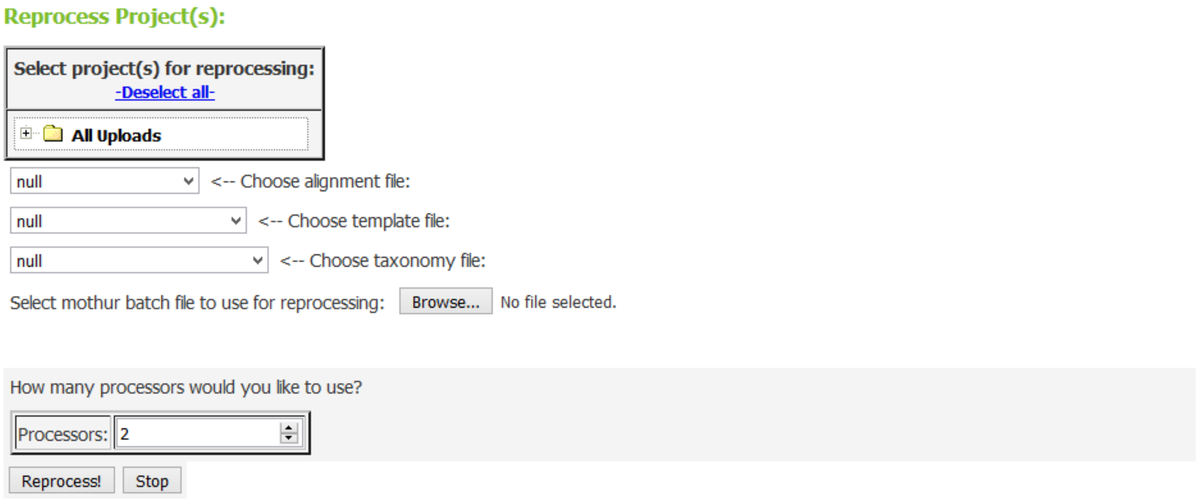
To do this, simply upload any new alignment, template, or taxonomic reference files using the ‘Reprocess’ page. Previously uploaded files will be stored for future use. Next, select the projects which need to be updated in the project tree, and select the correct (updated) reference files from the drop 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.

You can stop this process (or remove your request from the queue beforehand) using the “Stop” button next to the “Reprocess” button. This will restore the project to its previous state.



Upload New Taxonomy Reference Files:

Upload new reference database files:	
Select alignment file (e.g., silva.seed_v119.align):	Browse... No file selected.
Select template file (e.g., gg_13_5_99.fasta):	Browse... No file selected.
Select taxonomy file (e.g., gg_13_5_99.pds.tax):	Browse... No file selected.



Reprocess Project(s):

Select project(s) for reprocessing:
-Deselect all-

⊞ All Uploads

null <-- Choose alignment file:

null <-- Choose template file:

null <-- Choose taxonomy file:

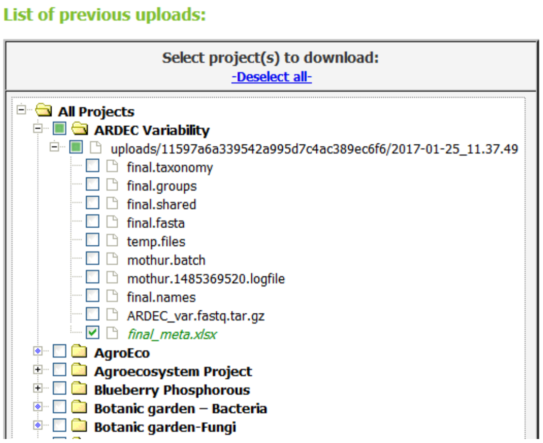
Select mothur batch file to use for reprocessing: Browse... No file selected.

How many processors would you like to use?

Processors: 2

4.3. Download: Download Project Files from Previous Uploads

The download page can be used to get copies of data files for projects uploaded into myPhyloDB. Note that currently you can only download either projects you personally uploaded or projects which were uploaded for public use. The download page's selection tree works very similarly to that of the select page, the key difference being that project nodes on the download tree list the dates of files related to the project as child nodes, rather than samples. This is useful for downloading specific sections of a project, i.e. parts which were uploaded initially vs added later on. Also, this function is especially useful for downloading and updating the metadata file.



List of previous uploads:

Select project(s) to download:
-Deselect all-

```
⊟ All Projects
  ⊟ ■ ARDEC Variability
    ⊟ ■ uploads/11597a6a339542a995d7c4ac389ec6f6/2017-01-25_11.37.49
        □ final.taxonomy
        □ final.groups
        □ final.shared
        □ final.fasta
        □ temp.files
        □ mothur.batch
        □ mothur.1485369520.logfile
        □ final.names
        □ ARDEC_var.fastq.tar.gz
        ☑ final_meta.xlsx
  ⊞ □ AgroEco
  ⊞ □ Agroecosystem Project
  ⊞ □ Blueberry Phosphorous
  ⊞ □ Botanic garden – Bacteria
  ⊞ □ Botanic garden-Fungi
```

4.4. Update: Updating Metadata in your myPhyloDB Database

To update a previously uploaded project with new metadata, click the ‘Update’ button on the sidebar. Then, select the project and path you wish to have updated. Use the file chooser to select the new file to be used for updating, then press “Update!”. Note: the new metadata file 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 (i.e., in the path shown on the project tree) or from the DataTable found on the select data page.

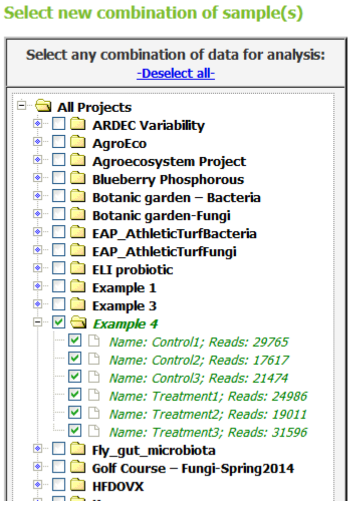
The update process uses the same queue as the Upload and Reprocess pages, meaning only one of these operations can occur at a time. You can use these pages the same as normal when the queue is busy, your request will simply be added to the end of the list of requests to be processed.

4.5. PyBake: Upload new PICRUSt and KEGG Files

New PICRUSt files can be updated using the PyBake function, including new GreenGenes taxonomy file, PICRUSt 16S rRNA count file, and PICRUSt KO gene predictions file. KEGG orthology and KEGG enzyme files can also be uploaded. For more information visit: http://picrust.github.io/picrust.

4.6. Select 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. Hovering the mouse over any sample will also display the sample description.



Select new combination of sample(s)

Select any combination of data for analysis:
-Deselect all-

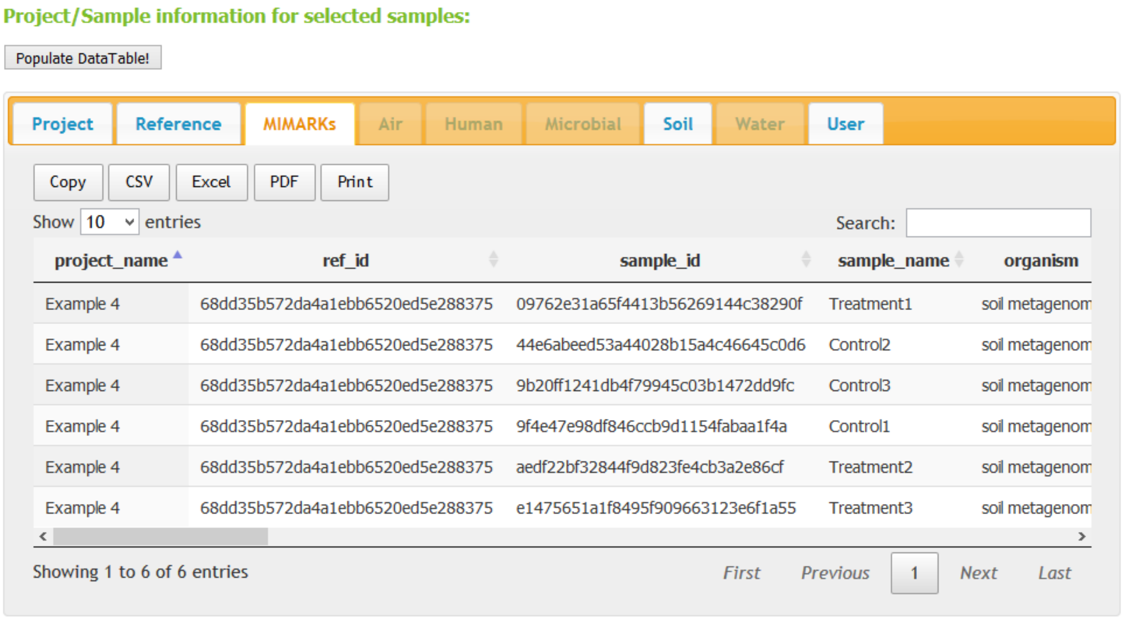
- All Projects
  - ☐ ARDEC Variability
  - ☐ AgroEco
  - ☐ Agroecosystem Project
  - ☐ Blueberry Phosphorous
  - ☐ Botanic garden – Bacteria
  - ☐ Botanic garden-Fungi
  - ☐ EAP_AthleticTurfBacteria
  - ☐ EAP_AthleticTurfFungi
  - ☐ ELI probiotic
  - ☐ Example 1
  - ☐ Example 3
  - ☑ Example 4
    - ☑ Name: Control1; Reads: 29765
    - ☑ Name: Control2; Reads: 17617
    - ☑ Name: Control3; Reads: 21474
    - ☑ Name: Treatment1; Reads: 24986
    - ☑ Name: Treatment2; Reads: 19011
    - ☑ Name: Treatment3; Reads: 31596
  - ☐ Fly_gut_microbiota
  - ☐ Golf Course – Fungi-Spring2014
  - ☐ HFDOVX

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

Once you have selected the data you wish to analyze further, click the “Save Selection(s)!” button below the project/sample tree. Next the selected samples are ready for normalization (see below).

The metadata associated with the each selected project/sample can be displayed in a DataTable by clicking the “Populate DataTable!” button. Data is organized into categories (Project, Reference, MIMARKS, Soil, Water, etc.). You may switch between these categories using the DataTable tabs. All samples should populate the Project, MIMARKs, Reference, and User-defined tabs; plus one additional tab (e.g., Soil, Air, Water, etc.) that is dependent upon the project type.

Each DataTable also 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. The sample_id is a unique ID generated by myPhyloDB, and used to identify samples during analysis.



Project/Sample information for selected samples:

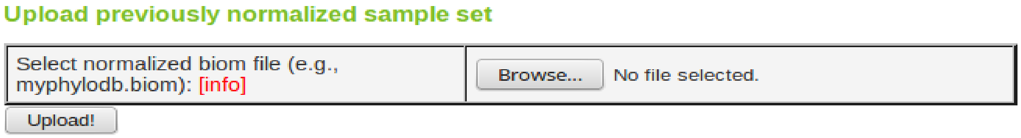
Show 10 entries

Search:

project_name	ref_id	sample_id	sample_name	organism
Example 4	68dd35b572da4a1ebb6520ed5e288375	09762e31a65f4413b56269144c38290f	Treatment1	soil metagenom
Example 4	68dd35b572da4a1ebb6520ed5e288375	44e6abeed53a44028b15a4c46645c0d6	Control2	soil metagenom
Example 4	68dd35b572da4a1ebb6520ed5e288375	9b20ff1241db4f79945c03b1472dd9fc	Control3	soil metagenom
Example 4	68dd35b572da4a1ebb6520ed5e288375	9f4e47e98df846ccb9d1154fabaa1f4a	Control1	soil metagenom
Example 4	68dd35b572da4a1ebb6520ed5e288375	aedf22bf32844f9d823fe4cb3a2e86cf	Treatment2	soil metagenom
Example 4	68dd35b572da4a1ebb6520ed5e288375	e1475651a1f8495f909663123e6f1a55	Treatment3	soil metagenom

Showing 1 to 6 of 6 entries

You can also select data through a pre-normalized .biom file, which can be uploaded through the file selector underneath the main selection tree



Upload previously normalized sample set

Select normalized biom file (e.g., myphylodb.biom): [info]

No file selected.

4.7. Normalize Data

After selecting the projects/samples of interest, you must perform a normalization step before proceeding to the data analysis section. Only once you have normalized your data will the analysis links become available. In addition, anytime that you change your project/sample selections (‘Select Data’ page), you will be required to re-normalize your data.

To perform a normalization perform the following steps:

1) Select your normalization method (see descriptions below). Depending upon the normalization procedure you select additional options will be displayed.

2) Select sample/species cutoffs (optional). To enable this feature you must first click the checkbox at

the left, then enter the minimum species size you desire (species with reads, summed across all samples, below this threshold will be removed).

3) After normalization, data files can be exported as g-zipped Tabular (.csv) or Biom (.biom) files. Biom files can be directly uploaded into R and compatible with the phyloseq package. You may also use the .biom file to ‘Upload previously normalized sample set’ under ‘Select Data.’ This is especially useful if using rarefaction normalization.

A brief description of each normalization procedure follows:

None: no normalization

Rarefaction (remove): This normalization procedure performs a typical sub-sampling ***without*** replacement to the desired sub-sample size as implemented in mothur and QIIME. 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.

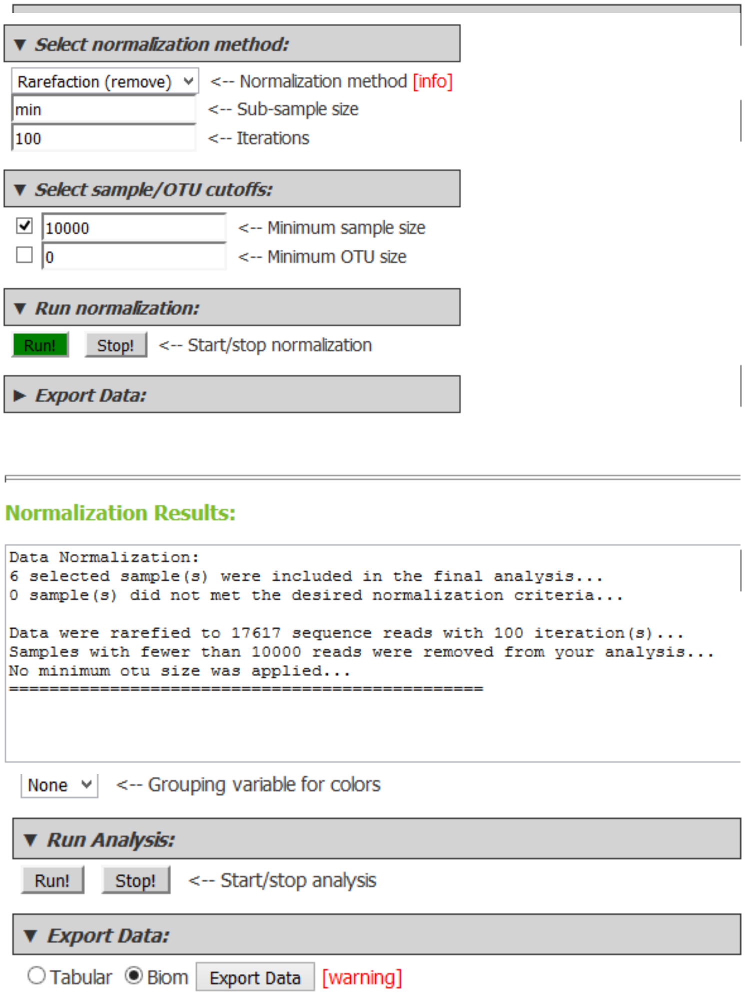
Rarefaction (keep): This normalization procedure performs a sub-sampling ***with*** replacement to the desired sub-sample size; and will keep all selected samples in the analysis regardless of their initial sample size (unless you set a minimum sample size as described above). Aguirre de Cárcer et al. (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 coverages below the sub-sampling 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 a modified additive (Laplace) smoothing technique with $\lambda = 0.1$. 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 detection levels. The Laplace approximated probabilities are then sampled to a user-defined sample size to generate a new taxonomic profile for each sample. In the text box provide you can enter “min”, “median”, “max”, or any integer for your desired sample size.

A word of caution on normalization: When selecting the Rarefaction normalization methods each individual iteration can produce different results due to the nature of probability sampling. To overcome this, we recommend that a minimum of 10 iterations be used for these procedures (default is 100). If set to 10, myPhyloDB will run 10 independent sub-samplings of your data and use the average phylotype abundances for analysis. You may also export the .biom file to avoid having to re-normalize.

Proportion: All abundances are divided by the total number of sequence reads for that sample.

5. Analysis:

In the sections below, we will use Example 4 to run each analysis. The data is not real, but designed to compare potato soils with or without a cover crop. The files for Example 4 are available on the website www.myphylodb.org/Sample-Files. We will normalize using Rarefaction (remove) method, sub-sampled to min with 100 iterations with a minimum sample size of 10,000 reads.



▼ *Select normalization method:*

Rarefaction (remove) <-- Normalization method [info]

min <-- Sub-sample size

100 <-- Iterations

▼ *Select sample/OTU cutoffs:*

☑ 10000 <-- Minimum sample size

☐ 0 <-- Minimum OTU size

▼ *Run normalization:*

Run! Stop! <-- Start/stop normalization

▶ *Export Data:*

Normalization Results:

```
Data Normalization:
6 selected sample(s) were included in the final analysis...
0 sample(s) did not met the desired normalization criteria...

Data were rarefied to 17617 sequence reads with 100 iteration(s)...
Samples with fewer than 10000 reads were removed from your analysis...
No minimum otu size was applied...
================================================
```

None <-- Grouping variable for colors

▼ *Run Analysis:*

Run! Stop! <-- Start/stop analysis

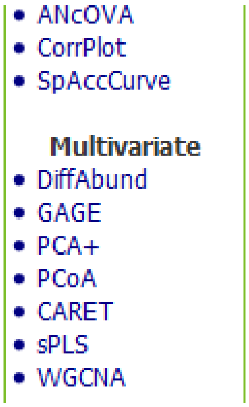
▼ *Export Data:*

○ Tabular ● Biom Export Data [warning]

5.1 Analysis Selection and Data Customization

Once you have selected and normalized your desired data, select the type of analysis you would like to perform, listed under the "Analysis" heading of the menu. Analysis options include:

- **Univariate**:
  - ANcOVA: Analysis of Covariance with general linear model (GLM)
  - CorrPlot: Correlation Matrix
  - SpAccCurve: Species Accumulation Curve
- **Multivariate**:
  - DiffAbund: Differential Abundance / Genewise Negative Binomial GLM
  - GAGE: General Acceptable Geneset/ Pathway analysis
  - PCA+: Principal Components Analysis (PCA), Canonical Correspondence Analysis (CCA), Detrended Correspondence Analysis (DCA)
  - PCoA: Principal Cooordinates Analysis
  - CARET: Classification And Regression Training
  - sPLS: Sparse Partial Least Squares Regression
  - WGCNA: Weighted Correlation Network Analysis



- ANcOVA
- CorrPlot
- SpAccCurve

Multivariate

- DiffAbund
- GAGE
- PCA+
- PCoA
- CARET
- sPLS
- WGCNA

After selecting an analysis, you are given the options to select and customize your statistical parameters and graphical output. Most analyses include:

- **Data Selection:** Select what type of analysis to run, including Taxonomy, KEGG orthology, or KEGG enzyme (where available) at different taxa/pathway/enzyme levels (e.g. Phyla, Class, Order…). You may also select the OTU level or PGPRs (plant growth promoting rhizobacteria). The Dependent variable can be set to Abundance (number of sequence reads), Relative Abundance (proportion of total community), or Total Abundance (16S rRNA copy numbers per mg soil) where qPCR rRNA copy numbers are entered into the metafile. Please note the units for abundance will be dependent upon the normalization procedure used. If you used the proportion procedure (i.e., divide by total number of reads for each sample), then units will be proportion and range from 0 to 1; otherwise, abundance units will be counts (i.e., number of reads). OTU Richness (number of OTUs) or OTU Diversity (Shannon index) may also be selected. In some analyses, data may also be transformed (Ln, Log10, Sqrt), and also have the option to 'Display only significant tests' at $p \leq 0.05$
- **Data Filtering:** You may narrow down the amount of data/reads to be analyzed, by removing unassigned/unmapped reads, remove phylotypes within ## percent or more zeroes (default = 50%), and/or select the top ## phylotypes based on Interquantile range (IQR), Coefficient variation (CV), Standard deviation (SD), Mean value, or Median value.
- **Optimize R plot:** This option is available to customize the graphical output's colors, layout, etc. depending on the Metadata variables selected.
- **Run Analysis:** Hit 'Run!' to start analysis. Progress will be displayed in the GraphData below. You may stop the analysis at anytime, but all progress will be lost. myPhyloDB uses a queue to organize user requests, and is able to run multiple analyses at a time (by default 3, which can be changed in the config file). If your analysis is able to be processed, status updates will be displayed on the page. Otherwise, the status indicator will show you as "In queue" followed by the number of requests in front of you.
- **Export Data:** Raw data may be exported as a Tabular file (.csv) or Biom file (.biom). Tabular files can be very large and take considerate time to download. Biom files (phyloseq.biom) can be directly imported into R statistical program, and analyzed using the phyloseq package.

Depending on the analysis, there can be other options for data selection, filtering, etc. These will be explained further in the next section overviewing each analysis.

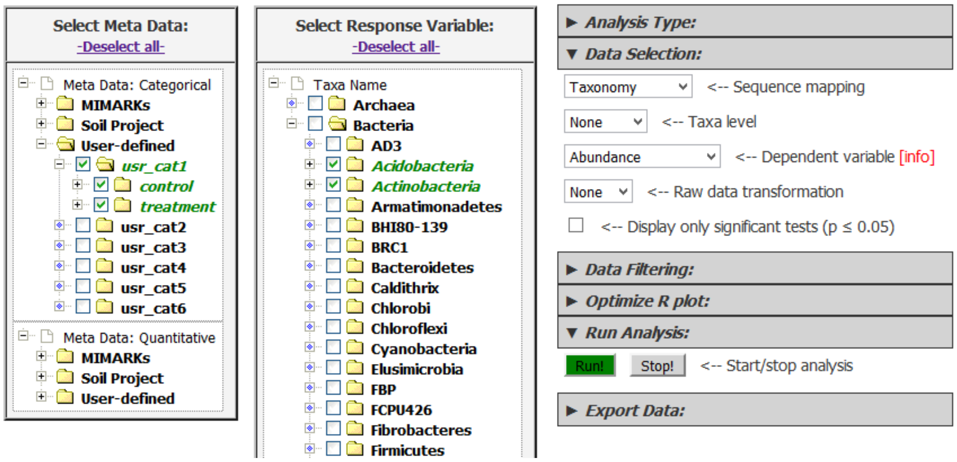
5.2 ANcOVA

ANcOVA (analysis of covariance) can be run in two different fashions in myPhyloDB. When the "Bar plot (factors)" option is selected, myPhyloDB performs an ANcOVA (i.e., comparison of factors), which may be run with, or without, user-specified covariates. Once the ANcOVA has completed successfully, a bar graph and ANOVA table will be displayed. If the "Scatter plot (regression)" option is selected, myPhyloDB performs a linear regression analysis (i.e., comparison of the regression slopes and intercepts), which may be run with, or without, user-specified variables. Once the GLM has completed successfully, a scatter plot with regression lines and ANOVA table will be displayed. As previously described, data may be filtered and the graphical output (R plot) may be optimized.

3.2.1 Bar plot (factors):

To run an ANcOVA and produce bar plots, you must first be sure that “Bar plots (factors)” is selected in the appropriate dropdown box. Next, select your meta-variable(s) of interest. Please note that any variables where all selected samples contain blanks (i.e., null values) will generate the following alert “No samples are available for this variable!” upon selection. Also, any samples with null data 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. As shown here, we have selected the categorical value of usr_cat1 for our ANcOVA.

Once you have selected your meta variables, you must select taxonomy or KEGG data either from the drop down menu (this option will selects ALL available taxa/enzymes/pathways at the chosen level) or by selecting specific taxonomic name(s) of interest from the tree. If multiple levels are selected in the taxa tree, myPhyloDB will run a separate ANcOVA for each taxa of interest. Based on the selections here, myPhyloDB will run two separate two-way ANcOVAs; one for Acidobacteria, and one for Actinobacteria.



Select Meta Data:
-Deselect all-

- Meta Data: Categorical
  - MIMARKs
  - Soil Project
  - User-defined
    - ☑ usr_cat1
      - ☑ control
      - ☑ treatment
    - ☐ usr_cat2
    - ☐ usr_cat3
    - ☐ usr_cat4
    - ☐ usr_cat5
    - ☐ usr_cat6
- Meta Data: Quantitative
  - MIMARKs
  - Soil Project
  - User-defined

Select Response Variable:
-Deselect all-

- Taxa Name
  - ☐ Archaea
  - ☐ Bacteria
    - ☐ AD3
    - ☑ Acidobacteria
    - ☑ Actinobacteria
    - ☐ Armatimonadetes
    - ☐ BHI80-139
    - ☐ BRC1
    - ☐ Bacteroidetes
    - ☐ Caldithrix
    - ☐ Chlorobi
    - ☐ Chloroflexi
    - ☐ Cyanobacteria
    - ☐ Elusimicrobia
    - ☐ FBP
    - ☐ FCPU426
    - ☐ Fibrobacteres
    - ☐ Firmicutes

▶ Analysis Type:

▼ Data Selection:

Taxonomy <-- Sequence mapping

None <-- Taxa level

Abundance <-- Dependent variable [info]

None <-- Raw data transformation

☐ <-- Display only significant tests (p ≤ 0.05)

▶ Data Filtering:

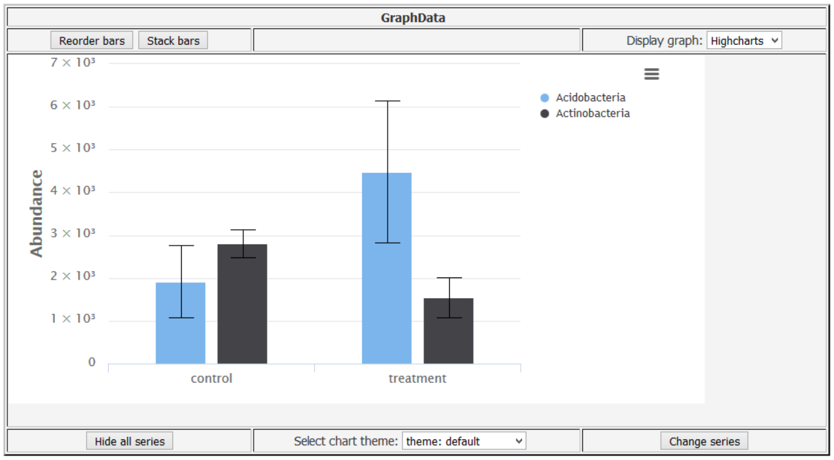
▶ Optimize R plot:

▼ Run Analysis:

Run! Stop! <-- Start/stop analysis

▶ Export Data:

The bar graph will display the taxa averages for each meta variable level selected. The charts (Highcharts) produced by myPhyloDB are highly interactive allowing the user to hide individual (by clicking on the legend text) or all (clicking the “Hide all series”) series shown in the graph, and download and/or print the chart using the button (3 horizontal bars) just above the figure legend.



In addition, you can change the individual colors as needed using the “Change series” button at the bottom of the graph. To do so, you will need to input the index position of the series (first position is 0) and new color (name or hex code) you want to change. For your convenience, the current series index, color can be displayed by holding the mouse over the appropriate text in the legend. You may also reorder bars by clicking on ‘Reorder bars’ in upper left, and entering the x-axis position (the first position is 0). You may also unstack bars, and error bars representing standard deviation will be displayed.

Bar graphs may also be displayed as R plots, by selecting ‘Display graph’ in upper right corner. Note that you cannot change the colors or ordering using R plots, after the analysis is processed. However, you can choose a color scheme under ‘Optimize R Plots.

After generating a bar plot, an ANcOVA table is generated summarizing results including

1) the meta variables selected, 2) an ANOVA table 3) LSmeans & Tukey’s HSD post-hoc test results and 4) pairwise comparisons table.

For users familiar with R, the above analysis will run the following R code:

```
fit <- aov(y~usr_cat1, data=df)
summary(fit)

library(lsmeans)
lsm <- lsmeans(fit, list(pairwise~usr_cat1))
```

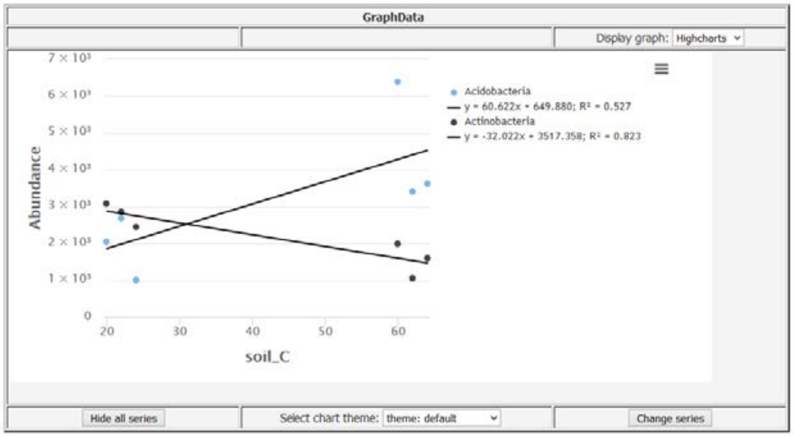
where y is the normalized abundance for each taxa chosen, df is a dataframe containing the appropriate metadata and normalized abundances.

Note: If you are running the same analysis on your own machine, your values may be slightly different due to the inherent variability in sub-sampling (see the Normalization section for more details). Also, the entire printout contained in the “Test Results” section is not shown below.

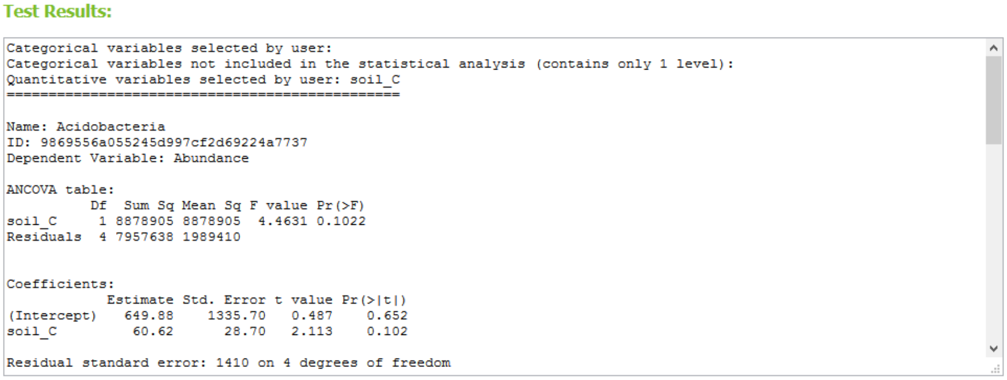
3.2.2 Scatter Plot

If you would like to run a general linear model (GLM) using quantitative metadata, select ‘Scatter plot (regression)’ under ‘Displayed graph.’ Similar to bar graphs, you can choose what type of sequence mapping (Taxonomy or KEGG) at a chosen level. Under ‘Optimize R plot’ pane, you are able to customize symbol colors, shapes, and panel display. Graphs can be displayed as Highcharts or R plots.

For Example 4, we have selected the quantitative variable soil carbon (Soil Project > Soil Nutrients > soil_C) and Acidobacteria and Actinobacteria.



At the top of the "Test Results" section is a summary for each selected taxa of 1) the meta variables selected, 2) an ANOVA table, 3) Coefficient table, 4) LSmeans, and 5) post-hoc test results.



Test Results:

```
Categorical variables selected by user:
Categorical variables not included in the statistical analysis (contains only 1 level):
Quantitative variables selected by user: soil_C
===============================================

Name: Acidobacteria
ID: 9869556a055245d997cf2d69224a7737
Dependent Variable: Abundance

ANCOVA table:
          Df  Sum Sq Mean Sq F value Pr(>F)
soil_C     1 8878905 8878905  4.4631 0.1022
Residuals  4 7957638 1989410


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   649.88    1335.70   0.487    0.652
soil_C         60.62      28.70   2.113    0.102

Residual standard error: 1410 on 4 degrees of freedom
```

For users familiar with R, the above analysis will run the following R code:

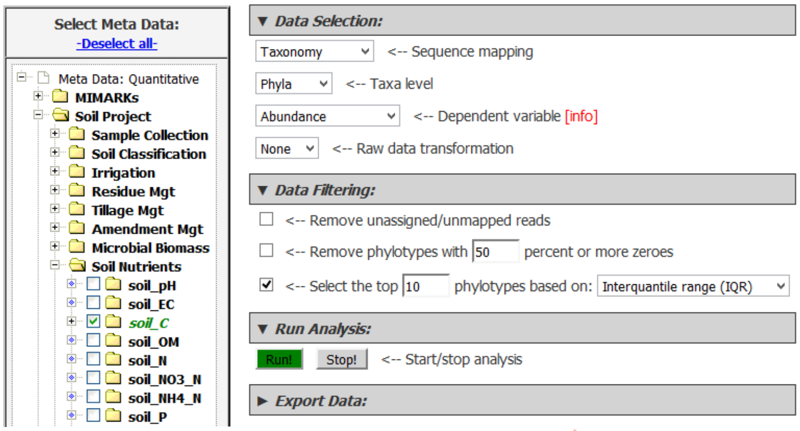
```
fit <- lm(soil_c, data=df)
summary(fit)

pred <- predict(fit, df)
aov <- anova(fit) a
```

where abund is the normalized abundance for each taxa chosen, df is a dataframe containing the appropriate metadata and normalized abundances. In addition, no post-hoc analysis is performed and all linear regression lines (shown in graph) are calculated using SciPy's 'linregress' function.

5.3 CorrPlot

Correlation plots can be generated using GLM correlating Taxa, KEGG orthology or KEGG enzymes at varying levels, to themselves (eg. Phyla A vs Phyla B) or to 1 or more Quantitative meta data (eg phyla vs soil nutrients); see examples below. The color shading and degree of circle fullness reflect Pearson's R coefficient, with blue being positively correlated, and red negatively correlated. A sortable and filterable Correlation Coefficients table and Correlation p-values table are generated below and can be copied, printed, or exported as a CSV, Excel or PDF file.

Below shows the correlation of phyla with soil carbon for Example 4:



Select Meta Data:
-Deselect all-

- Meta Data: Quantitative
  - MIMARKs
  - Soil Project
    - Sample Collection
    - Soil Classification
    - Irrigation
    - Residue Mgt
    - Tillage Mgt
    - Amendment Mgt
    - Microbial Biomass
    - Soil Nutrients
      - □ soil_pH
      - □ soil_EC
      - ☑ soil_C
      - □ soil_OM
      - □ soil_N
      - □ soil_NO3_N
      - □ soil_NH4_N
      - □ soil_P

▼ Data Selection:

Taxonomy <-- Sequence mapping

Phyla <-- Taxa level

Abundance <-- Dependent variable [info]

None <-- Raw data transformation

▼ Data Filtering:

□ <-- Remove unassigned/unmapped reads

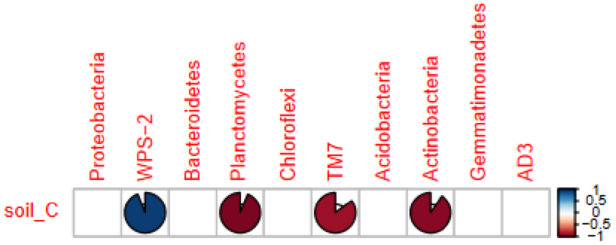
□ <-- Remove phylotypes with 50 percent or more zeroes

☑ <-- Select the top 10 phylotypes based on: Interquantile range (IQR)

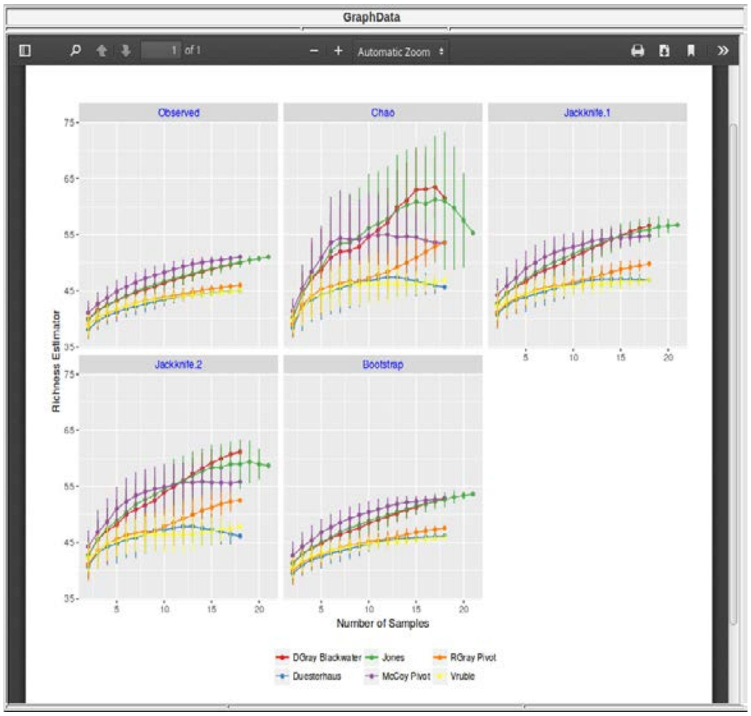
▼ Run Analysis:

Run! Stop! <-- Start/stop analysis

► Export Data:



5.3 SpAccCurve



Species accumulation curves (SpAccCurve) measuring richness over number of samples can be analyzed using 5 different richness estimators: Observed, Chao, Jackknife1, Jacknife2 and Bootstrap. A Species Accumulation Curve Data table is generated showing richness values and their standard deviation in a sortable, filterable, and exportable DataTable.

Example 4 only has 3 samples per group, which is not enough to generate a meaningful curve. The graphic here is from another project with several samples.

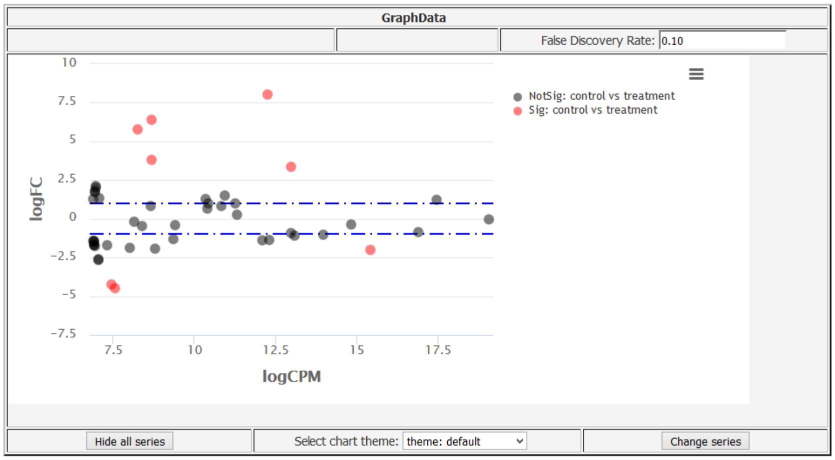
5.4 Diff Abund

The differential abundance (DiffAbund) procedure is part of the edgeR package in R and more details on the procedure can be found here.

The Diff Abund graph has a logarithmic scale for the x-axis (baseMean) and a standard scale for the y-axis (log2FoldChange). Under ‘edgeR settings’, you may specify the maximum number of taxa to display for each comparison, sorted by p-value. All significant data points are plotted in red (significance being determined by a p value of ≤ 0.05) and non-significant in black. If more than one meta variable is selected, the analysis will compare each independent treatment combination to one another and displayed on the same GraphData output. Main effects (one effect independent of the other) are not allowed; however, the analysis can easily be rerun with only one variable selected. The plot shows a comparison of each sample with every other sample, with respect to your selected meta-variables. As with the other graphs in myPhyloDB, you can toggle individual data points on and off. You can also mouse-over any point on the graph to see the 32-character alphanumeric UUID and axes coordinates.

The Genewise Negative Binomial GLM test results are reported in a sortable, filterable, printable and exportable DataTable. You can search for specific samples to check their results, as well as export the table into CSV, excel spreadsheet, and PDF formats. For more information on the nbinomial test please refer to the edgeR manual.

The graph below shows the differential abundance of phyla for Example 4 control vs treatment with false discovery rate set to 0.10.



GraphData

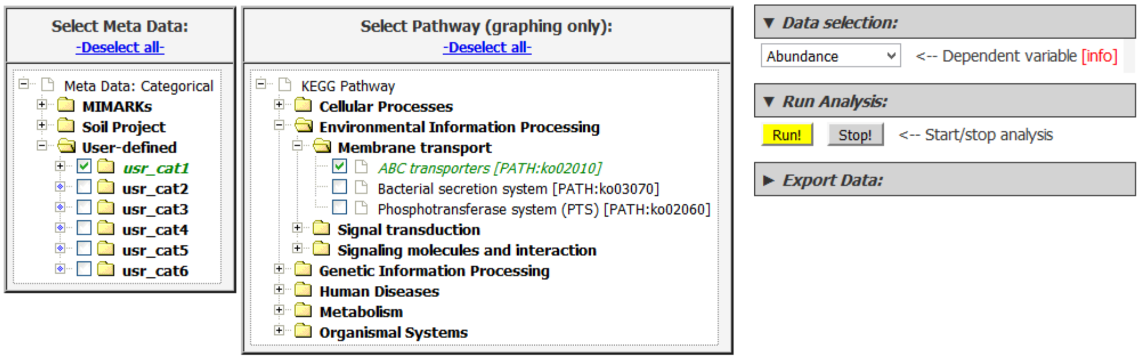
False Discovery Rate: 0.10

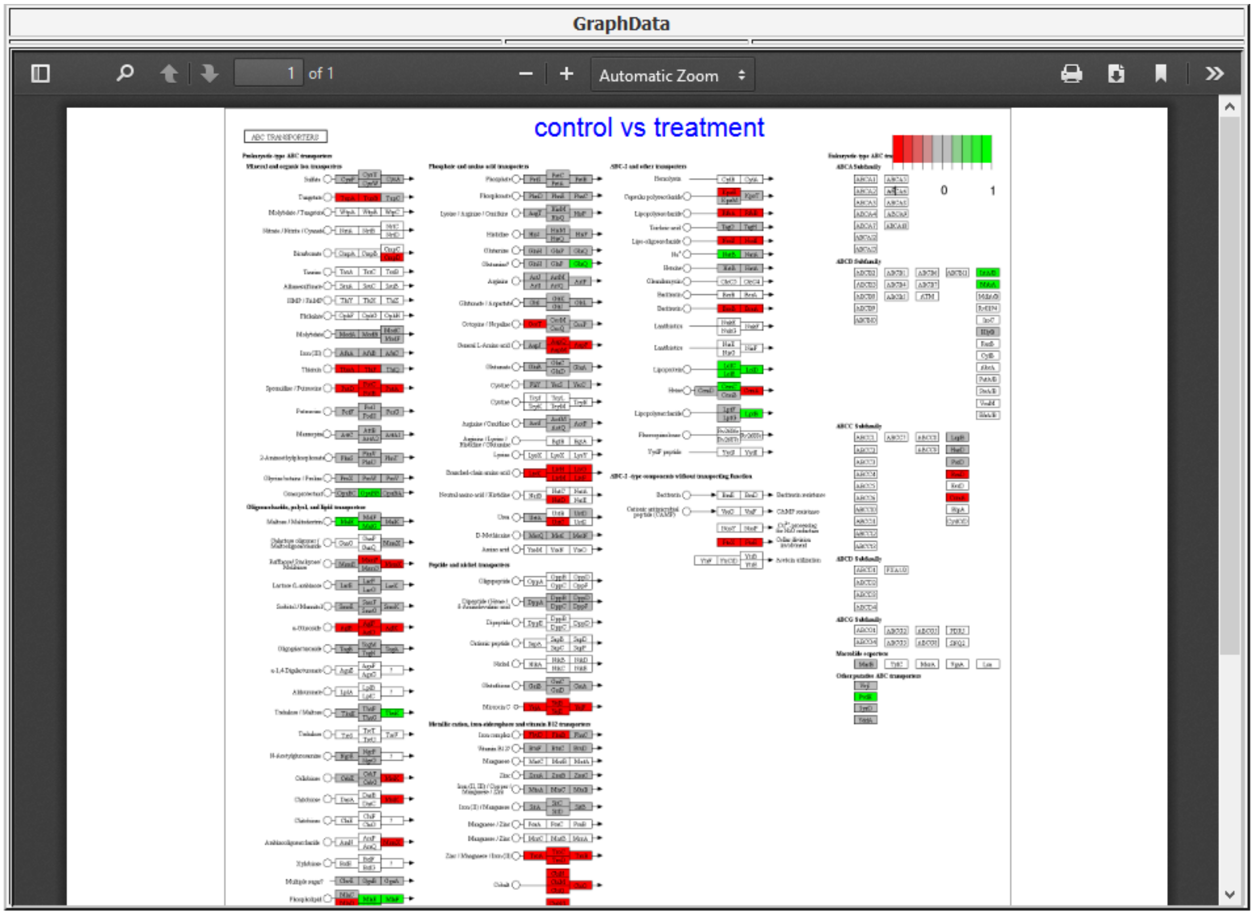
Select chart theme: theme: default

5.5 GAGE

GAGE, or Generally Applicable Gene-set analysis, utilizes PICRUSt data to predict metagenome functions using GLM regression. More information on the GAGE package in R can be found here. The output graph displays an entire pathway and associated enzymes and their negative (red) or positive (green) correlations, which can be printed or exported as a PDF file. A graph of the KEGG pathway is displayed for each pairwise comparison of your categorical data, i.e. treatment groups. The data output is in a searchable, printable, sortable and exportable table displaying pair-wise comparisons of groups of all KEGG pathways and enzymes.

Below shows the GAGE analysis of Example 4 control vs. treatment, of ABC transporters pathway. Enzyme boxes highlighted in green represent it is upregulated in the treatment group vs. control, and red represents downregulated.

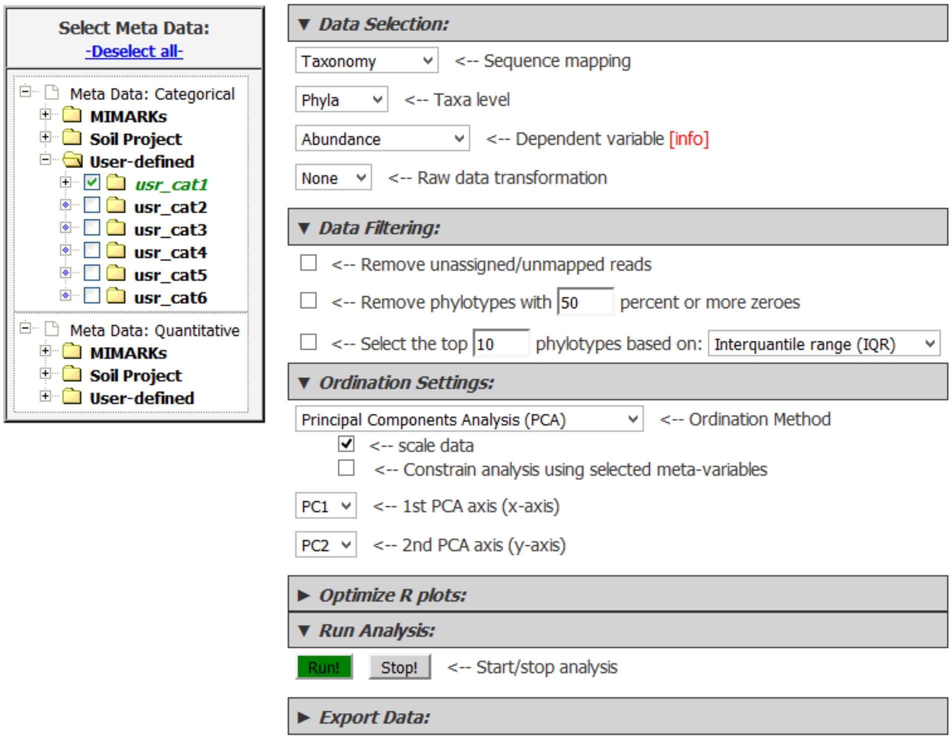




5.6 PCA+

The PCA+ analysis enables you to run a principal components analysis (PCA), correspondence analysis (CCA) or detrended correspondence analysis (DCA). Like other analyses, sequence mapping can be set to Taxonomy, KEGG orthology or KEGG enzyme (levels 2-4). There are several options under the Data Filtering and Ordination Settings pane to optimize your analysis. Results are exportable as .csv or .biom files. The Test Results table is a sortable, searchable table displaying eigenvalues, variable coordinates and individual coordinates.

Below is a PCA plot of Example 4



Select Meta Data:
-Deselect all-

- Meta Data: Categorical
  - MIMARKs
  - Soil Project
  - User-defined
    - ☑ *usr_cat1*
    - □ usr_cat2
    - □ usr_cat3
    - □ usr_cat4
    - □ usr_cat5
    - □ usr_cat6
- Meta Data: Quantitative
  - MIMARKs
  - Soil Project
  - User-defined

▼ *Data Selection:*

Taxonomy <-- Sequence mapping

Phyla <-- Taxa level

Abundance <-- Dependent variable [info]

None <-- Raw data transformation

▼ *Data Filtering:*

□ <-- Remove unassigned/unmapped reads

□ <-- Remove phylotypes with 50 percent or more zeroes

□ <-- Select the top 10 phylotypes based on: Interquantile range (IQR)

▼ *Ordination Settings:*

Principal Components Analysis (PCA) <-- Ordination Method

☑ <-- scale data

□ <-- Constrain analysis using selected meta-variables

PC1 <-- 1st PCA axis (x-axis)

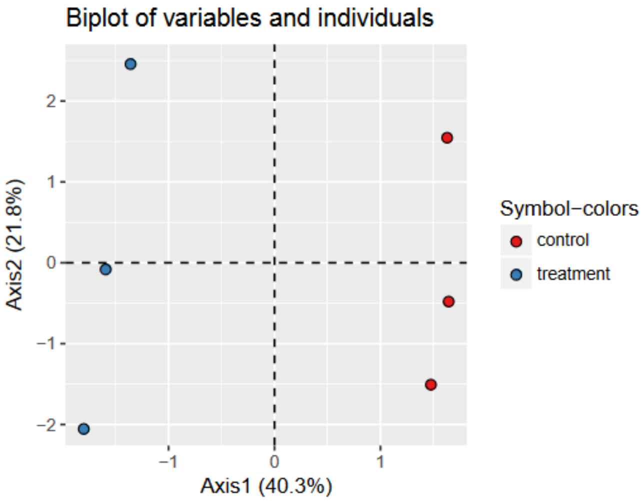
PC2 <-- 2nd PCA axis (y-axis)

▶ *Optimize R plots:*

▼ *Run Analysis:*

Run! Stop! <-- Start/stop analysis

▶ *Export Data:*



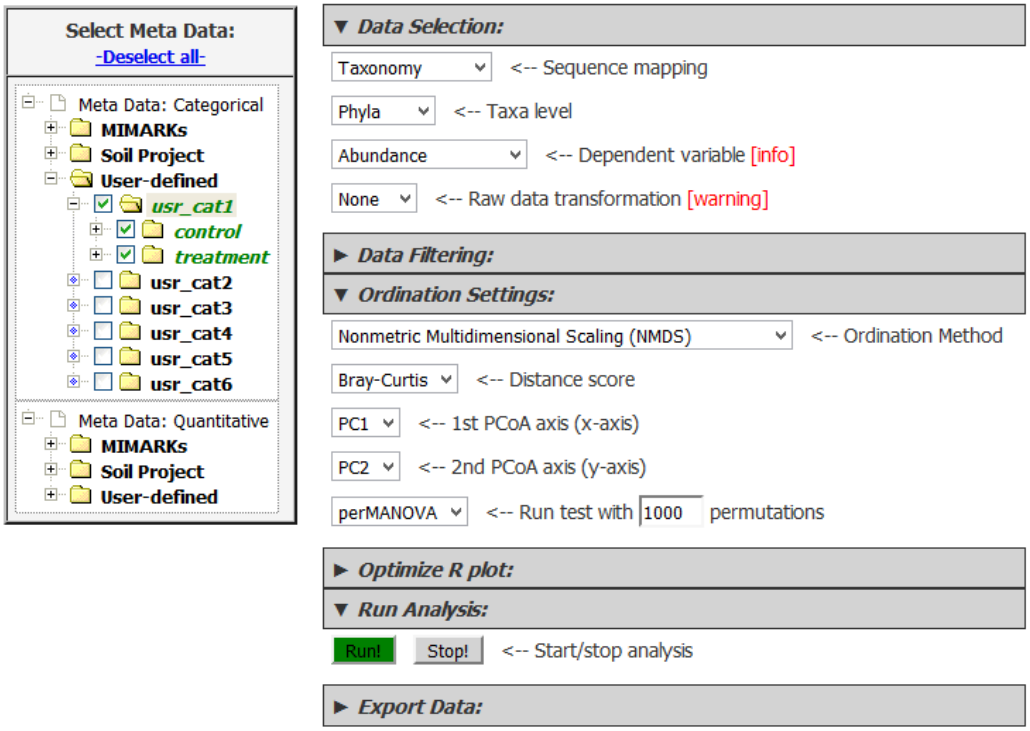
5.7 PCoA (Principal Coordinates Analysis)

The PCoA analysis can run three different ordination methods, available under ‘Ordination Settings’: Constrained Analysis of Princeipal Coordinates (CAP), Nonmetric Multidimensional Scaling (NMDS), and Weighted Classical (Metric) Multidimensional Scaling.

There are several Distance score metrics to choose from. All scores are calculated using the vegan package in R, except for MorisitaHorn (custom python script based on the calculator in mothur) and wOdum. The wOdum score can be used to down-weight either rare (α > 1) or abundant (α < 1) taxa, as discussed here (Manter and Bakker. 2015. BioInformatics). When α = 1, wOdum is equivalent to Bray-Curtis.

You may also select your X and Y axes and type of non-parametic analysis (perMANOVA or betaDispr) with ## permutations (default = 1000) of the selected data using the embedded R vegan package. perMANOVA or betaDispr results are displayed in the results table, and the principal coordinates and distance scores are displayed in a sortable, searchable, exportable table.

Below is a PCoA plot of Example 4, using nonmetric multidimensional scaling (NMDS) with Bray-Curtis distance score run with 1000 permutations via perMANOVA.



Select Meta Data:
-Deselect all-

- Meta Data: Categorical
  - MIMARKs
  - Soil Project
  - User-defined
    - ☑ usr_cat1
      - ☑ control
      - ☑ treatment
    - □ usr_cat2
    - □ usr_cat3
    - □ usr_cat4
    - □ usr_cat5
    - □ usr_cat6
- Meta Data: Quantitative
  - MIMARKs
  - Soil Project
  - User-defined

▼ Data Selection:

Taxonomy <-- Sequence mapping

Phyla <-- Taxa level

Abundance <-- Dependent variable [info]

None <-- Raw data transformation [warning]

► Data Filtering:

▼ Ordination Settings:

Nonmetric Multidimensional Scaling (NMDS) <-- Ordination Method

Bray-Curtis <-- Distance score

PC1 <-- 1st PCoA axis (x-axis)

PC2 <-- 2nd PCoA axis (y-axis)

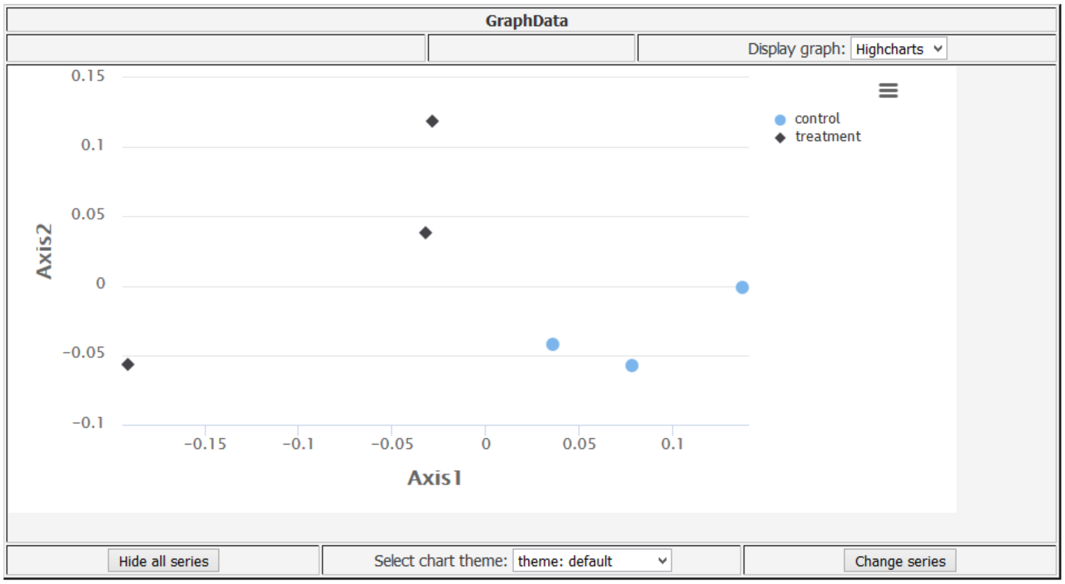
perMANOVA <-- Run test with 1000 permutations

► Optimize R plot:

▼ Run Analysis:

<-- Start/stop analysis

► Export Data:



Display graph: Highcharts

Select chart theme: theme: default

At the top of “Test Results” section is a summary of the taxa level selected, data normalization step, which includes the number of samples normalized (or removed) and the number of reads used for rarefaction. This section also displays the perMANOVA or betaDisper test results and the Eigenvalues and proportion of the variance explained for each PCoA axis. If any quantitative variables are selected, 'envfit' results are also posted here. All analyses are conducted using the R vegan package.

Also displayed are DataTables of the calculated “Principal Coordinates” and “Distance Scores” in matrix form. These tables are sortable, searchable, printable and exportable.

Quantitative variables are handled in a similar manner and will produce a scatter plot between the selected variable (y-axis) and the chosen principal coordinate axis (x-axis). However, to avoid unit conflicts only one meta variable may be analyzed at any time; and instead of a perMANOVA or betaDisper analysis a simple linear regression analysis is performed.

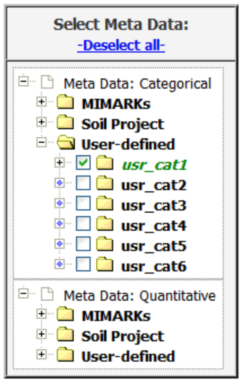
For users familiar with R, the above analysis will run the following R code:

```
library(vegan)
dist <- vegdist(data, method='manhattan')
ord <- capscale(dist ~usrcat_1, meta)

cat <- factor(meta$usr_cat1)

plot(ord, type='n')
points(ord, display='sites', pch=15, col=cat)
legend('topright', legend=levels(cat), pch=15, col=1:length(cat))
pl <- ordiellipse(ord, cat, kind='sd', conf=0.95, draw='polygon', border='black')
ordisurf(ord, usr_cat1, add=TRUE)


res1 <- adonis(dist ~usr_cat1, perms=1000)
res2 <- betadisper(dist, usr_cat1)
```

where data is the normalized abundance for each taxa chosen, meta is a dataframe containing the appropriate metadata. In this example, usr_cat1 was the meta-variable chosen for analysis.

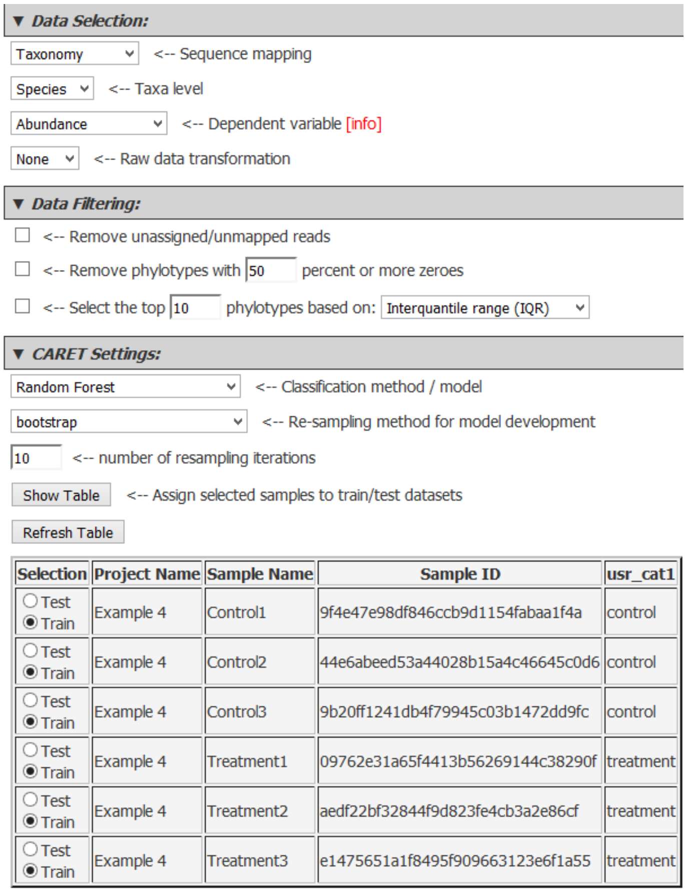
5.8 CARET

Classification and Regression Training (CARET) contains numerous tools used for predictive modeling, including data splitting, pre-processing, feature selection, model tuning using resampling, and variable importance estimation (Kuhm, 2008. Journal of Statistical Software. 28(4):1-26). myPhyloDB currently has three different classification methods available: Random Forest, Neural Network (feed-forward), and Support Vector Machine and several re-sampling methods.



Select Meta Data:
-Deselect all-

- Meta Data: Categorical
  - MIMARKs
  - Soil Project
  - User-defined
    - ☑ *usr_cat1*
    - □ usr_cat2
    - □ usr_cat3
    - □ usr_cat4
    - □ usr_cat5
    - □ usr_cat6
- Meta Data: Quantitative
  - MIMARKs
  - Soil Project
  - User-defined



▼ *Data Selection:*

Taxonomy <-- Sequence mapping

Species <-- Taxa level

Abundance <-- Dependent variable [info]

None <-- Raw data transformation

▼ *Data Filtering:*

□ <-- Remove unassigned/unmapped reads

□ <-- Remove phylotypes with 50 percent or more zeroes

□ <-- Select the top 10 phylotypes based on: Interquantile range (IQR)

▼ *CARET Settings:*

Random Forest <-- Classification method / model

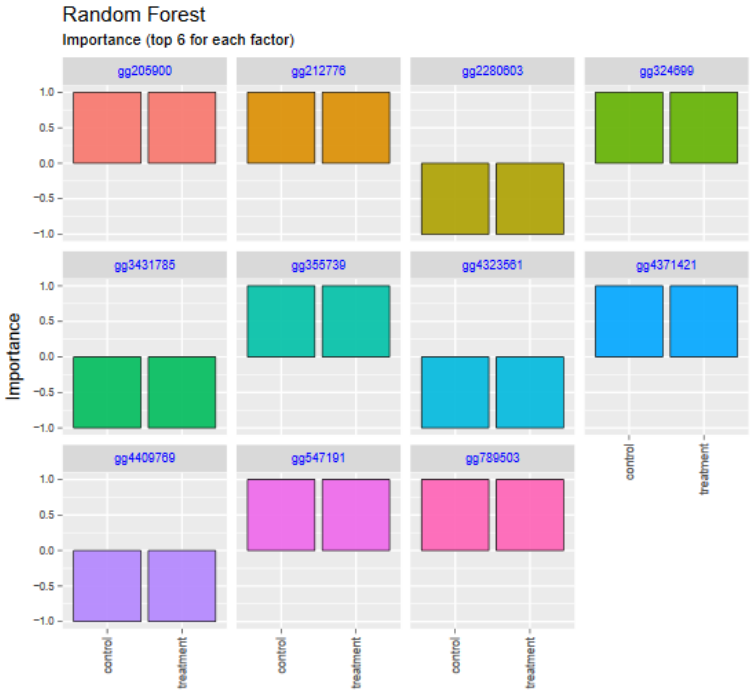
bootstrap <-- Re-sampling method for model development

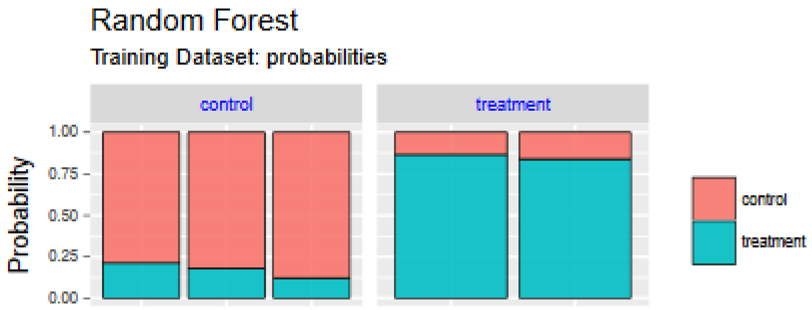
10 <-- number of resampling iterations

Show Table <-- Assign selected samples to train/test datasets

Refresh Table

Selection	Project Name	Sample Name	Sample ID	usr_cat1
○ Test ● Train	Example 4	Control1	9f4e47e98df846ccb9d1154fabaa1f4a	control
○ Test ● Train	Example 4	Control2	44e6abeed53a44028b15a4c46645c0d6	control
○ Test ● Train	Example 4	Control3	9b20ff1241db4f79945c03b1472dd9fc	control
○ Test ● Train	Example 4	Treatment1	09762e31a65f4413b56269144c38290f	treatment
○ Test ● Train	Example 4	Treatment2	aedf22bf32844f9d823fe4cb3a2e86cf	treatment
○ Test ● Train	Example 4	Treatment3	e1475651a1f8495f909663123e6f1a55	treatment





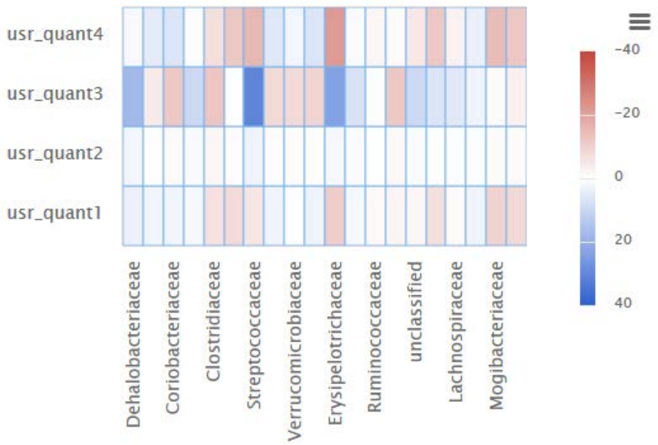
5.9 sPLS

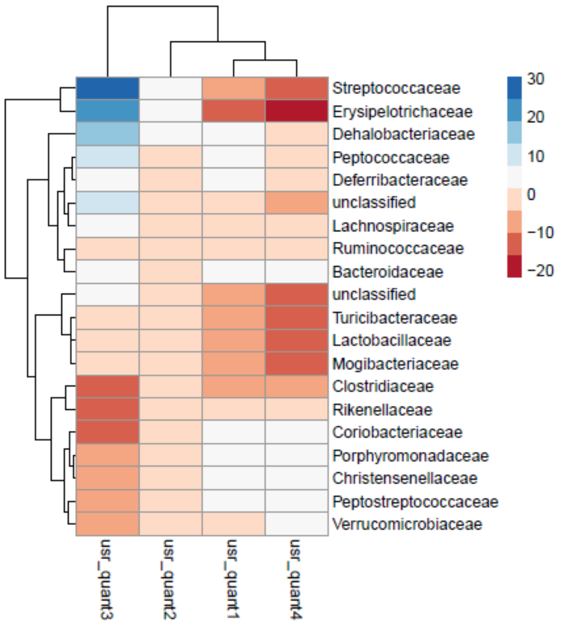
The sPLS (sparse partial least squares regression) analysis is run using the spls package of R and is a useful technique for the simultaneous dimension reduction and variable selection (Chun and Keles. 2010. R Stat Soc Series B Stat Methodol. 72: 3–25). This makes sPLS a good choice for identifying important predictor variables among a large number of predictors in highly dimensional data, such as microbial communities.

Only quantitative meta data may be used to run sPLS to correlate with chosen taxa, KEGG orthology or KEGG enzyme. Under ‘sPLS Settings’ you can remove variables with zero variance using freqCut (cutoff for the ratio of the most common to the second most common value) or uniqueCut (cutoff for the percentage of distinct values out of the total number of samples), and also perform variable scaling to predictor (X) variables or response (Y) variables. Under ‘Optimize R plot’ a linkage method can be selected (single, complete, average, weighted, median, centroid) as well as a distance metric (correlation, Euclidean, Manhattan, maximum).

The GraphData outputs a heatmatp that can be displayed as HighCharts or R Plot. Red shading (negative correlation) and blue shading (positive correlation) reflect R^2 values of predicted vs. observed values. The sPLS Coefficients and Observed and Predicted Data tables are searchable, printable, sortable and exportable.

sPLS could not be run in Example 4, due to the small number of samples resulting in insignificant variance. Below is an example of a sPLS highchart and R plot from another project of abundance at the OTU level, with 4 quantitative variables:





For users familiar with R, the above analysis will run the following R code:

```
library(mixOmics)
ZeroVar <- nearZeroVar(X, freqCut=95/5, uniqueCut=10)
List <- row.names(ZeroVar$Metrics)
X_new <- X[,-which(names(X) %in% List)]
X_scaled <- scale(X_new, center=FALSE, scale=TRUE)
Y_scaled <- scale(Y, center=FALSE, scale=FALSE)

detach('package:mixOmics', unload=TRUE)
library(spls)
set.seed(1)
cv <- cv.spls(X_scaled, Y_scaled, scale.x=FALSE, scale.y=FALSE, eta=seq(0.1, 0.9, 0.1),
       K=c(1:5), plot.it=FALSE)
f <- spls(X_scaled, Y_scaled, scale.x=FALSE, scale.y=FALSE, eta=cv$eta.opt, K=cv$K.opt)
coef.f <- coef(f)
sum <- sum(cf != 0)

set.seed(1)
ci.f <- ci.spls(f, plot.it=FALSE, plot.fix='y')
cis <- ci.f$cibeta
cf <- correct.spls(ci.f, plot.it=FALSE)

library(pheatmap)
pheatmap(df, clustering_method='complete', clustering_distance_rows='euclidean',
       clustering_distance_cols='euclidean')
```

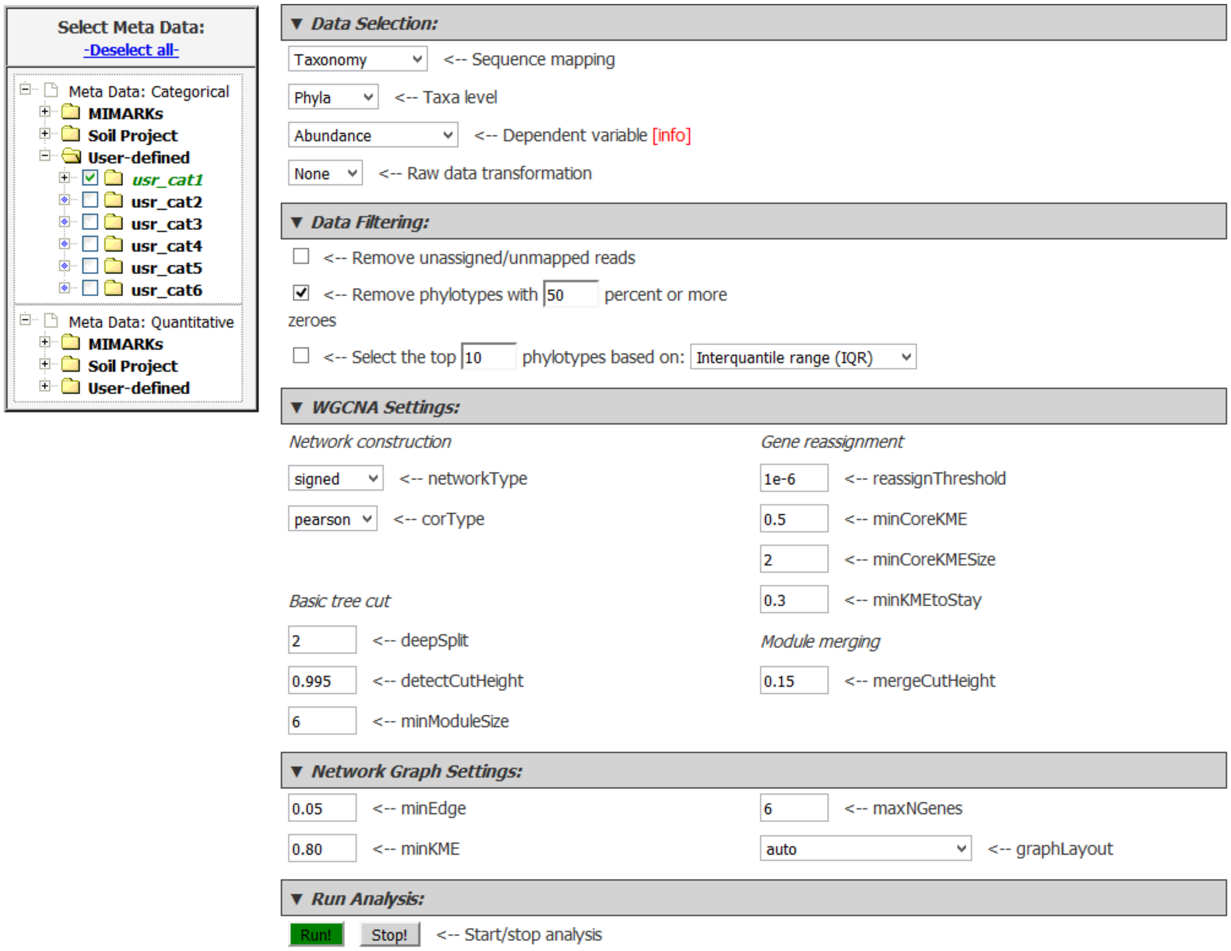
where X is the normalized abundance for each taxa chosen, Y is a dataframe containing the appropriate metadata.

5.10 Weighted Gene Correlation Network Analysis

The weighted gene correlation network analysis (WGCNA) describes correlation patterns among genes and/or metadata, to find highly correlated genes described as modules. The WGCNA includes module construction, hub gene selection, module preservation statistics, differential network analysis, and network statistics.

For more information: Langfelder and Horvath, BMC Bioinformatics, 2008, doi: 10.1186/1471-2105-9-559

The WGCNA analysis in myPhyloDB is highly interactive and customizable to your statistical needs, including data filtering similar to other analyses, WGCNA settings including network construction, basic tree cut, gene reassignment, and module merging settings, as well as network graph settings.



Select Meta Data:
-Deselect all-

- Meta Data: Categorical
  - MIMARKs
  - Soil Project
  - User-defined
    - ☑ usr_cat1
    - □ usr_cat2
    - □ usr_cat3
    - □ usr_cat4
    - □ usr_cat5
    - □ usr_cat6
- Meta Data: Quantitative
  - MIMARKs
  - Soil Project
  - User-defined

▼ *Data Selection:*

Taxonomy <-- Sequence mapping

Phyla <-- Taxa level

Abundance <-- Dependent variable [info]

None <-- Raw data transformation

▼ *Data Filtering:*

□ <-- Remove unassigned/unmapped reads

☑ <-- Remove phylotypes with 50 percent or more zeroes

□ <-- Select the top 10 phylotypes based on: Interquantile range (IQR)

▼ *WGCNA Settings:*

Network construction

signed <-- networkType

pearson <-- corType

Basic tree cut

2 <-- deepSplit

0.995 <-- detectCutHeight

6 <-- minModuleSize

Gene reassignment

1e-6 <-- reassignThreshold

0.5 <-- minCoreKME

2 <-- minCoreKMESize

0.3 <-- minKMEtoStay

Module merging

0.15 <-- mergeCutHeight

▼ *Network Graph Settings:*

0.05 <-- minEdge

0.80 <-- minKME

6 <-- maxNGenes

auto <-- graphLayout

▼ *Run Analysis:*

Run! Stop! <-- Start/stop analysis

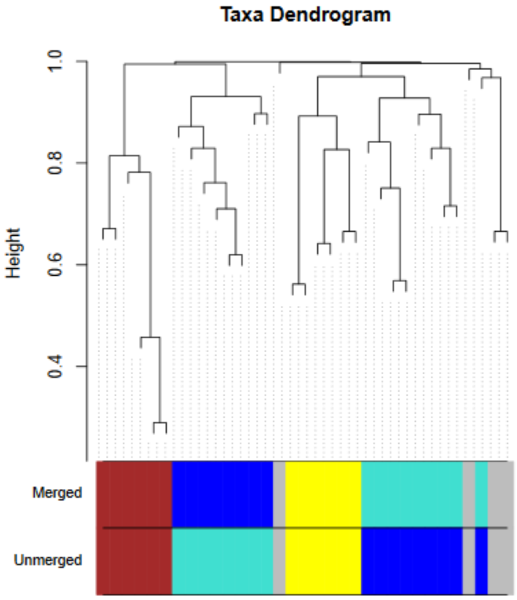
After running, there are 5 graphs generated:

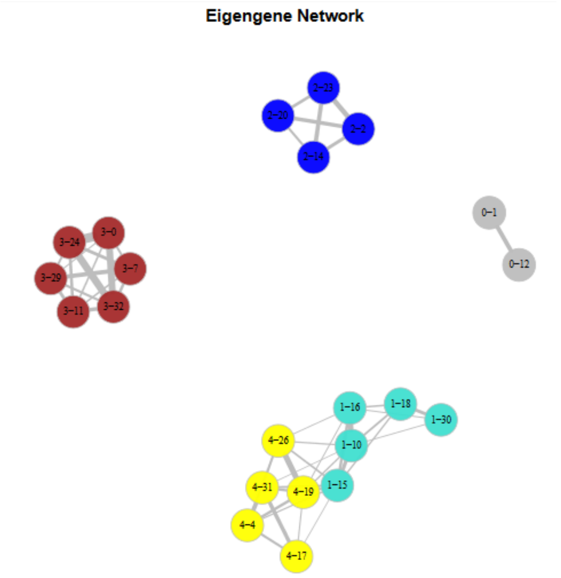
1. Taxa Dendogram
2. Eigengene Dendogram
3. TOM (topological overlap matrix) Dissimilarity Matrix
4. Eigengene Network
5. Eigengene Boxplots.

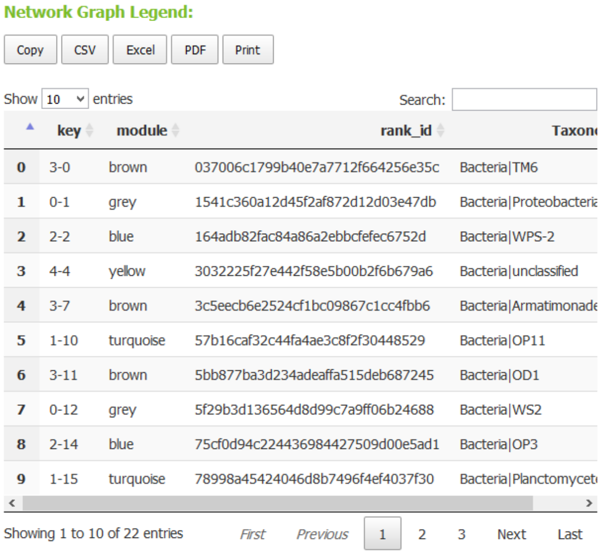
There are 5 datatables generated:

1. Test Results, including ANOVA table
2. Network Graph Legend
3. Module Membership (kME) and Connectivity
4. Module Eigengenes
5. Network Statistics

Below is Example 4 run with usr_cat1 Abundance at the Phyla level, at default settings with phylotypes removed with 50% or more zeroes (see screenshot above). We show only 2 graphs and 1 datatable output. The Taxa Dendogram and Eigengene Network graphs below show visual representations of phyla modules. The Network Graph Legend contains the phyla taxonomy, module (color) and key (#-#).







Network Graph Legend:

Show 10 entries Search:

▲	key	module	rank_id	Taxon
0	3-0	brown	037006c1799b40e7a7712f664256e35c	Bacteria|TM6
1	0-1	grey	1541c360a12d45f2af872d12d03e47db	Bacteria|Proteobacteria
2	2-2	blue	164adb82fac84a86a2ebbcfefec6752d	Bacteria|WPS-2
3	4-4	yellow	3032225f27e442f58e5b00b2f6b679a6	Bacteria|unclassified
4	3-7	brown	3c5eecb6e2524cf1bc09867c1cc4fbb6	Bacteria|Armatimonade
5	1-10	turquoise	57b16caf32c44fa4ae3c8f2f30448529	Bacteria|OP11
6	3-11	brown	5bb877ba3d234adeaffa515deb687245	Bacteria|OD1
7	0-12	grey	5f29b3d136564d8d99c7a9ff06b24688	Bacteria|WS2
8	2-14	blue	75cf0d94c224436984427509d00e5ad1	Bacteria|OP3
9	1-15	turquoise	78998a45424046d8b7496f4ef4037f30	Bacteria|Planctomycet

Showing 1 to 10 of 22 entries

6. Further Information

12. Error logging

In the unfortunate event that myPhyloDB fails to run any analyses, an error log (traceback) will be produced and added to the 'error_log.txt' file in myPhyloDB's home directory. Each error log will begin with the following:

```
Error:root:
Date: YYYY-MM-DD hr:min:sec
```

Please be sure to identify the appropriate date for your error and submit the logfile to our team at myphylodb@gmail.com or the users' forum at www.myphylodb.org with as much detail describing your analysis selections as possible.