



## 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](#).

New features in myPhyloDB v.1.1.2 are marked as: **New feature in v.1.1.**

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. 20xx (in review). myPhyloDB: a local web server for the storage and analysis of metagenomic data. *Database: The Journal of Biological Databases and Curation*

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	p. 2
2.	Home Screen and Sidebar	p. 4
3.	Uploading New Data	p. 5
4.	Reanalyzing Data	p. 14
5.	Updating Metadata	p. 15
6.	Selecting Data for Analysis	p. 16
7.	Export Data	p. 18
8.	Search Taxa	p. 19
9.	Analysis	p. 20
10.	Normalization	p. 40
11.	Manage Users	p. 41
12.	Error logging	p. 42

## 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. However, since users do not need to verify the email address entered on the download site, no personal information is required to download myPhyloDB (i.e., you can use a fake email).

### ***Windows:***

Double-click the installer (myPhyloDB\_v.1.1.2\_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.

### ***Linux:***

Run the installer (myPhyloDB\_v.1.1.2\_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.1.1.2_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.

**Bug Fix for v.1.1.1 (Linux version only):** Please note that in myPhyloDB v.1.1.1 all statistical analyses that use the R package will fail due to R looking for the wrong directories. To fix please navigate to the following R folder in your myPhyloDB directory and edit the “R/R-Linux/bin/R” file as follows. Replace all instances of '/home/manterd/PycharmProjects’ with “\$HOME”. This problem has been fixed in myPhyloDB v.1.1.2.

***Mac Users:***

Sorry but we are not Mac owners. 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.1.1.1 with the bug fix described above applied.

***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. 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. **New feature in v.1.1.** From here you may either choose to login as a registered user (see Manage Users section) using the “Login” link or simply proceed to the “Select Data” page as a guest. Registered users will (i) have access to the “upload”, “reprocess”, and “update” functions of myPhyloDB. Guests have no modification rights. In addition, projects can be designated as public or private. 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.

### ***Dynamic nature of the Menu sidebar:***

**New feature in v.1.1.** The links available on the Menu sidebar are controlled by user type (superuser, registered user, or guest) and whether sample(s) have been selected. Panel A: Menu sidebar at startup; Panel B: User logged in as a superuser/staff – Manage Users and all Data Mgt links visible; Panel C: User logged in as superuser/staff with samples selected – Manage Users, all Data Mgt and Analysis links are visible; Panel D: User not logged in (guest) – Data Mgt (select data only) and Analysis links visible.

### 3. 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. **New feature in v.1.1.** All metadata is now uploaded using a single Excel file, replacing the project and sample files needed in v.1.0.

#### Upload new data files:

1.) Select metadata file:

Select meta.csv file:

No file selected.

2.) Select sequence data format:

Available Data Formats: Pre-processed Mothur Files ▼

3.) Select sequencing files:

Select conserved taxonomy file:

No file selected.

Select .shared file:

No file selected.

### 3.1.1 Project type

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](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. **New feature in v.1.1.** The current meta data Excel file (e.g., myPhyloDB.Soil.meta.xls) provides suggested controlled vocabulary lists and units for each defined variable. 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 MIMARK [paper](#).

**New feature in v.1.1.** Projects can now 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, modification rights remain unchanged.

### 3.1.2 The metadata file

Each upload requires a completed metadata file, which can be downloaded from myPhyloDB's homepage. Column (variables) names 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.

Only one project can be uploaded at a time; however, 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 of myPhyloDB. Similarly, you may add new sequence data to an existing sample by setting both the project\_id and sample\_id values to the appropriate auto-generated UUIDs.

### 3.1.3 Select your sequence data format

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

### 3.1.4 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 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](http://www.mothur.org/wiki/Taxonomy_outline)). If necessary, 'unclassified' can be used for any taxonomic level without relevant information (e.g., species when using RDP).

Once you have 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%, this 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 multi-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 processors minus 1. 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 11 (i.e., 12 processors minus 1). Of course, if you set this to 1, myPhyloDB will use 1 processor not 0. 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.

#### *Can I modify the provided batch files?*

Yes, 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 preceeding 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) are supported.

Although, it is possible for users to employ an operational taxonomic unit (OTU) based analysis (e.g., consensus taxonomy at 3% genetic distance) in their pipelines. If so, all of the stored data files (the 5 final files shown above) will be based on your chosen OTU definition; however, any analysis performed by myPhyloDB's will be based on taxonomic names (i.e., phylotypes)



**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.

***Example3: 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(s) (standard fasta files), qual file(s) (read quality file), oligo file(s) (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 → Example3.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 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 config file (file with sample names and fastq file names), fastq files (forward and reverse for each sample), and a mothur batch file. 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.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 “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.Stages.files

fastq files → Example4.Stages\_0.rep1\_F.fastq  
Example4.Stages\_0.rep1\_R.fastq  
Example4.Stages\_0.rep2\_F.fastq  
Example4.Stages\_0.rep2\_R.fastq  
Example4.Stages\_0.rep3\_F.fastq  
Example4.Stages\_0.rep3\_R.fastq

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.

### 3.1.5 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 sequencing processing (i.e., when mothur is running) step; however, mothur will output it's progress to your host computer's terminal. The following is an example of the time's required to upload the 4 example projects.

Project	Procedure	Test computer	Time (hr:min:sec)
Example1	Uploading and parsing	Linux <sup>1</sup>	0:0:31
		Windows <sup>2</sup>	0:01:04
Example2	Sequencing processing, uploading and parsing	Linux <sup>1</sup>	0:23:48
		Windows <sup>2</sup>	3:07:25*
Example3	Sequencing processing, uploading and parsing	Linux <sup>1</sup>	0:08:34
		Windows <sup>2</sup>	0:17:08
Example4	Sequencing processing, uploading and parsing	Linux <sup>1</sup>	1:08:44
		Windows <sup>2</sup>	2:20:38

<sup>1</sup> Computer configuration: Linux Mint 17.2 LTS, 32 GB RAM, i7-5930K @ 3.5 GHz

<sup>2</sup> Computer configuration: Windows 7 Pro, 8 GB RAM, i7-4790 @ 3.6 GHz

\*multi-processing not implemented for all mothur functions (e.g., sff.multitple) in Windows

### 3.2 File storage

In addition to 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 datafiles (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.

### 3.3 Removing Data from your myPhyloDB Database

At the bottom of the “[Upload Data]” 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 your database but also the archived copies of the raw and processed data in your “uploads” folder.

#### List of previous uploads:

- ☐ Project: Example 1  
(Path: uploads/9634c481908b44cca490cb8e563e4d4f/2015-09-05\_22.5.3)
- ☐ Project: Example 2  
(Path: uploads/f91ba37bade04360b1ad7b7f419f6398/2015-09-05\_22.6.42)

#### 4. 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. 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.

##### Upload New Taxonomy Reference Files:

Upload new reference database files:		
Select alignment file (e.g., silva.seed_v119.align):	<input type="button" value="Browse..."/>	No file selected.
Select template file (e.g., gg_13_5_99.fasta):	<input type="button" value="Browse..."/>	No file selected.
Select taxonomy file (e.g., gg_13_5_99.pds.tax):	<input type="button" value="Browse..."/>	No file selected.

##### Reprocess Project(s):

<b>Select project(s) for reprocessing:</b> <a href="#">-Deselect all-</a>	<div>silva.seed_v119.align ▼ &lt;-- Choose alignment file:</div> <div>gg_13_5_99.fasta ▼ &lt;-- Choose template file:</div> <div>gg_13_5_99.pds.tax ▼ &lt;-- Choose taxonomy file:</div>
<div><div>[-]  All Uploads</div><div>[+]   Project: Example 2</div></div>	

## 5. 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.

Select project path to update:[Deselect all](#)

All Uploads

Project: Example 1

☒ Path: *uploads/0fea6f86c20149efb93c0896c6282c49/2015-10-02\_23.20.54*

Upload new meta files:

Select meta.xls file:

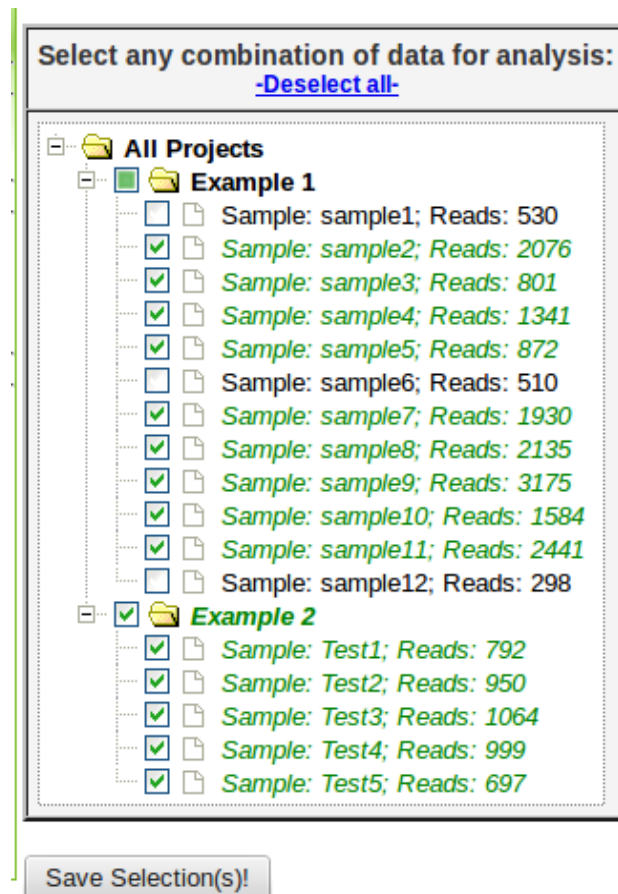
Browse...

No file selected.

Update!

## 6. 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. Hovering the mouse over any sample will also display the sample description.



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. Note: Upon clicking the button, a pop-up window will appear saying “Selected sample(s) have been recorded!”, press “OK” and proceed to the “Analysis” section of myPhyloDB or explore the selected using the DataTable 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 (minimum information about a marker gene sequence), Reference, and User-defined tabs; plus one additional tab (e.g., Soil, Air, Water, etc.) that is dependent upon the project type. Example DataTable with Example1 selected.

#### Project/Sample information for selected samples:

Populate DataTable!

Project	Reference	MIMARKS	Air	Human	Microbial	Soil	Water	User
---------	-----------	---------	-----	-------	-----------	------	-------	------

Copy CSV Excel PDF Print Search:

project_name ▲	project_id	sample_name	project_desc	start_d
Example 1	b1828cb322ed445b9094e0f578d0d172	sample1	Mothur pre-processed soil data for myPhyloDB	2015-11-17
Example 1	b1828cb322ed445b9094e0f578d0d172	sample2	Mothur pre-processed soil data for myPhyloDB	2015-11-17
Example 1	b1828cb322ed445b9094e0f578d0d172	sample3	Mothur pre-processed soil data for myPhyloDB	2015-11-17
Example 1	b1828cb322ed445b9094e0f578d0d172	sample4	Mothur pre-processed soil data for myPhyloDB	2015-11-17
Example 1	b1828cb322ed445b9094e0f578d0d172	sample5	Mothur pre-processed soil data for myPhyloDB	2015-11-17
Example 1	b1828cb322ed445b9094e0f578d0d172	sample6	Mothur pre-processed soil data for myPhyloDB	2015-11-17
Example 1	b1828cb322ed445b9094e0f578d0d172	sample7	Mothur pre-processed soil data for myPhyloDB	2015-11-17
Example 1	b1828cb322ed445b9094e0f578d0d172	sample8	Mothur pre-processed soil data for myPhyloDB	2015-11-17
Example 1	b1828cb322ed445b9094e0f578d0d172	sample9	Mothur pre-processed soil data for myPhyloDB	2015-11-17
Example 1	b1828cb322ed445b9094e0f578d0d172	sample10	Mothur pre-processed soil data for myPhyloDB	2015-11-17
Example 1	b1828cb322ed445b9094e0f578d0d172	sample11	Mothur pre-processed soil data for myPhyloDB	2015-11-17
Example 1	b1828cb322ed445b9094e0f578d0d172	sample12	Mothur pre-processed soil data for myPhyloDB	2015-11-17
Example 1	b1828cb322ed445b9094e0f578d0d172	sample13	Mothur pre-processed soil data for myPhyloDB	2015-11-17
Example 1	b1828cb322ed445b9094e0f578d0d172	sample14	Mothur pre-processed soil data for myPhyloDB	2015-11-17
Example 1	b1828cb322ed445b9094e0f578d0d172	sample15	Mothur pre-processed soil data for myPhyloDB	2015-11-17
Example 1	b1828cb322ed445b9094e0f578d0d172	sample16	Mothur pre-processed soil data for myPhyloDB	2015-11-17
Example 1	b1828cb322ed445b9094e0f578d0d172	sample17	Mothur pre-processed soil data for myPhyloDB	2015-11-17

Showing 1 to 18 of 29 entries

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. Please note that the export buttons (“Copy”, “CSV”, “Excel”, and “PDF”) require that your browser have Adobe flash player installed. If Adobe flash player is not installed only the Print button will be present operable.

## 7. Export Data

### New feature in v.1.1.

myPhyloDB allows users to select any combination of samples and metadata for normalization and export to tabular (similar to Qiime OTU table or Mothur shared file) or biom format. To export data, click on the 'Export Data' link in the menu sidebar and then select the metadata, taxa, and normalization procedures you desire. For more information on using the trees and dropdown boxes on this page, please refer to the Analysis section.

#### Raw Data (Tabular):

A raw data DataTable is also output to this page, which includes the selected metadata and normalized dependent variable (e.g., abundance counts) for each taxa level included in the analysis. The full taxonomic classification for each taxonomic level in the DataTable can be obtained by searching the database with the appropriate taxa\_id using the "Search Taxa" link on the main menu.

#### Raw Data (Biom):

The normalized data is also output to a textbox in biom format to allow for easy export and use with other software packages.

## 8. Search Taxa:

### Search External Links:

Taxa name:

-MicrobeWiki-  
-Wiki-  
-Google-

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:

<div> <div>Copy</div> <div>CSV</div> <div>Excel</div> <div>PDF</div> <div>Print</div> </div> <div>Search: <input type="text"/></div>					
	Kingdom Name	Kingdom ID	Phylum Name	Phylum ID	Class Name
0	Bacteria	f2751f946e7a49218800c26038fa03ae	Proteobacteria	c1307ea124ea429f8cce3ee8fb6e11d2	Gammaproteobacteria
1	Bacteria	f2751f946e7a49218800c26038fa03ae	Proteobacteria	c1307ea124ea429f8cce3ee8fb6e11d2	Deltaproteobacteria
2	Bacteria	f2751f946e7a49218800c26038fa03ae	Proteobacteria	c1307ea124ea429f8cce3ee8fb6e11d2	Betaproteobacteria
3	Bacteria	f2751f946e7a49218800c26038fa03ae	Proteobacteria	c1307ea124ea429f8cce3ee8fb6e11d2	Gammaproteobacteria
4	Bacteria	f2751f946e7a49218800c26038fa03ae	Proteobacteria	c1307ea124ea429f8cce3ee8fb6e11d2	Alphaproteobacteria
5	Bacteria	f2751f946e7a49218800c26038fa03ae	Proteobacteria	c1307ea124ea429f8cce3ee8fb6e11d2	Alphaproteobacteria
6	Bacteria	f2751f946e7a49218800c26038fa03ae	Proteobacteria	c1307ea124ea429f8cce3ee8fb6e11d2	Gammaproteobacteria
7	Bacteria	f2751f946e7a49218800c26038fa03ae	Proteobacteria	c1307ea124ea429f8cce3ee8fb6e11d2	Alphaproteobacteria
8	Bacteria	f2751f946e7a49218800c26038fa03ae	Acidobacteria	9a0841b121364ae485640fb141bfef7d	Acidobacteriia
9	Bacteria	f2751f946e7a49218800c26038fa03ae	Bacteroidetes	237447c4245e48fda1c6f7aa12e5897c	Flavobacteriia
10	Bacteria	f2751f946e7a49218800c26038fa03ae	Actinobacteria	3ee2935a7e8f4e4fa9d387d5cf9efcc2	Actinobacteria

## 9. 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: ANCOVA/GLM; Multivariate: DiffAbund, PcoA, or sPLS).

All data/graphs shown in this manual were generated with Example 1, which is preloaded in myPhyloDB.

### 9.1. ANcOVA

**New feature in v.1.1.1.** ANcOVA (analysis of covariance) can be run in two different fashions in myPhyloDB. When the “Bar plot (factors)” option is selected, myPhyloDB performs an ANOVA (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 dummy variables. Once the GLM has completed successfully, a scatter plot with regression lines and ANOVA table will be displayed.

**Bug Fix for v.1.1.1.** When selecting individual taxa from the “Select Taxa” tree, relative proportions were incorrectly normalized. This issue did not occur when using the “Select taxa level” dropdown box and only affected the ANOVA page. Issue has been fixed in v.1.1.2.

**Bar plot (factors):**

To run an ANCOVA, you must first be sure that “ANCOVA” is selected in the appropriate dropdown box. Next, select your meta-variable(s) of interest. Please note that any variables where all selected samples are blank (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. The project name for each sample can be found by hovering the mouse over that sample in the tree. As shown below, we have selected 2 categorical variables (*env\_material* and *geo\_loc\_farm*) and 1 quantitative variable (*usr\_quant1*) for our ANCOVA.

**Select Meta Data: [Deselect all](#)**

Meta Data: Categorical

- ☐ MIMARKS
  - ☐ sample\_name
  - ☐ organism
  - ☐ collection\_date
  - ☐ depth
  - ☐ elev
  - ☐ seq\_platform
  - ☐ seq\_gene
  - ☐ seq\_gene\_region
  - ☐ seq\_for\_primer
  - ☐ seq\_rev\_primer
  - ☐ env\_biome
  - ☐ env\_feature
  - ☒ *env\_material*
    - ☒ *soil-bulk*
    - ☒ *soil-rhizosphere*
  - ☐ geo\_loc\_country
  - ☐ geo\_loc\_state
  - ☐ geo\_loc\_city
  - ☒ *geo\_loc\_farm*
    - ☒ *farm1*
    - ☒ *farm2*
    - ☒ *farm3*
  - ☐ geo\_loc\_plot
- ☐ Soil
- ☐ User-defined

Meta Data: Quantitative

- ☐ MIMARKS
- ☐ Soil
- ☒ User-defined
  - ☒ *usr\_quant1*
  - ☐ usr\_quant2
  - ☐ usr\_quant3
  - ☐ usr\_quant4
  - ☐ usr\_quant5
  - ☐ usr\_quant6

**Select Taxa: [Deselect all](#)**

Taxa Name

- ☐ Bacteria
- ☐ unknown

ANCOVA <-- Selected test

Off <-- Selected taxa level

Once you have selected your meta variables, you must select taxonomy data either from the drop down

menu (this option selects ALL available taxa at the chosen level) or by selecting specific taxonomic name(s) of interest from the taxonomy tree. Any desired combination of taxonomic level(s) and name(s) can be selected using the taxonomic tree by simply selecting the appropriate checkboxes. Please note, that if you use the “Selected taxa level” drop down menu, the taxa tree will automatically be emptied of all selections. In addition, if multiple taxa levels are selected (dropdown box or taxa tree), myPhyloDB will run a separate ANCOVA for each taxa of interest. Based on the selections below, myPhyloDB will run three separate two-way ANCOVAs; one for Acidobacteria, one for Actinobacteria, and one for Proteobacteria.

**Select Taxa: [Deselect all](#)**

**Taxa Name**

- ☐ **Bacteria**
  - ☒ **Acidobacteria**
  - ☒ **Actinobacteria**
  - ☐ **Armatimonadetes**
  - ☐ **Bacteroidetes**
  - ☐ **Chloroflexi**
  - ☐ **Cyanobacteria**
  - ☐ **Elusimicrobia**
  - ☐ **FBP**
  - ☐ **Fibrobacteres**
  - ☐ **Firmicutes**
  - ☐ **Gemmatimonadetes**
  - ☐ **Nitrospirae**
  - ☐ **OC31**
  - ☐ **Planctomycetes**
  - ☒ **Proteobacteria**
  - ☐ **SAR406**
  - ☐ **TM7**
  - ☐ **Thermi**
  - ☐ **Verrucomicrobia**
  - ☐ **WPS-2**
  - ☐ **WS3**
- ☐ **unknown**

ANCOVA

<-- Selected test

Off

<-- Selected taxa level

The final selections required for analysis are all located within the graph table of the analysis page. Here you can select your dependent variable (abundance (counts), species richness, or Shannon's Diversity Index) and normalization method (none, rarefaction (remove), rarefaction (keep), proportion, DESeq2).

Please see the normalization section of this manual for a more detailed explanation of each option. Based on the selections below, the dependent variable for each ANCOVA will be relative abundance (proportion) and the data will be normalized using the rarefaction (keep) method. In addition, all samples will be subsampled to 1000 sequence reads and use the average values from 10 independent iterations.

Display only significantly tests ( $p \leq 0.05$ ): ☐

GraphData	
Dependent variable: Relative Abundance (proportion) ▼	Normalization method: Rarefaction (keep) ▼
	Subsample size: 1000
	Iterations: 10
No Data has been selected!	

Optionally you may also choose to display only significantly different taxa (“Display only significant tests” checkbox).

Once you are satisfied with your data choices, click the “Run Analysis!” button on the 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.

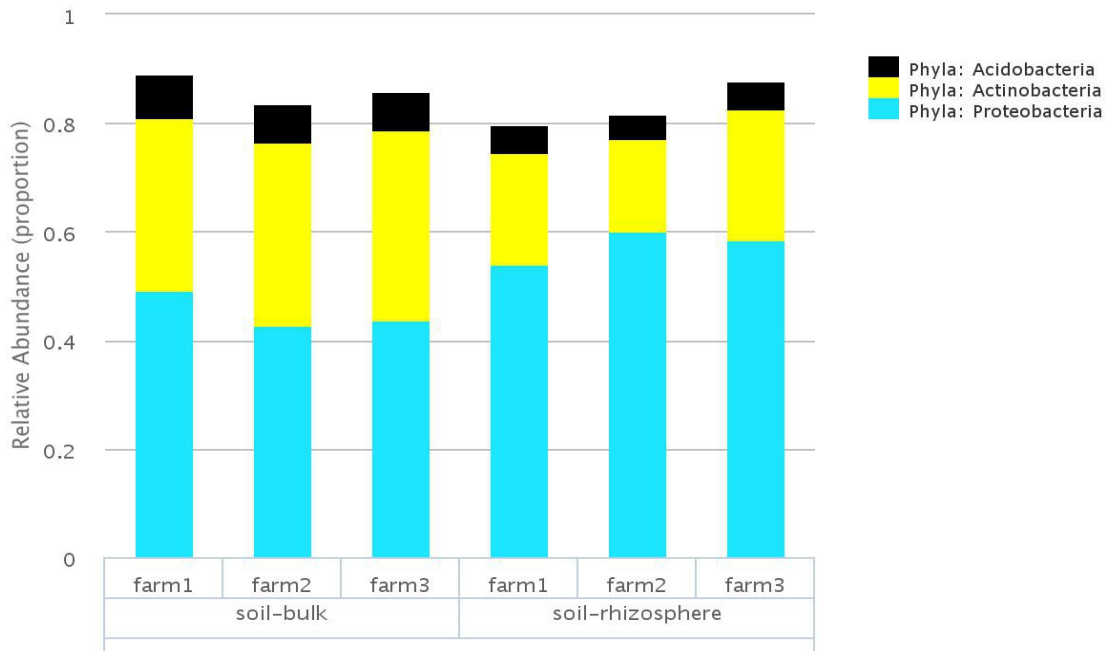
**New feature in v.1.1.** myPhyloDB will warn users if the run button is clicked while an analysis is being performed, blocking any subsequent submissions until the first submission is complete. A stop button has also been added to stop any current analyses. Because all analyses are handled by background processes running on the server it may take several seconds (typically 1-5 sec) for the stop signal to be processed (repeatedly pressing the stop button may also be helpful). Once the analysis has been successfully stopped, the “Run Analysis” should turn red and a “Your analysis has been stopped!” message displayed in your browser.

When you analysis is complete, a new bar graph will be displayed in the graph table along with the statistical results in the boxes below the graph.

Menu
<b>General Info</b>
<ul style="list-style-type: none"> <li>• Home</li> <li>• Login</li> </ul>
<b>Data Mgt</b>
<ul style="list-style-type: none"> <li>• Select Data</li> <li>• Export Data</li> </ul>
<b>Taxonomy</b>
<ul style="list-style-type: none"> <li>• Search Taxa</li> </ul>
<b>Analysis</b>
<b>Univariate</b> <ul style="list-style-type: none"> <li>• ANcOVA</li> </ul>
<b>Multivariate</b> <ul style="list-style-type: none"> <li>• Diff Abund</li> <li>• PCoA</li> <li>• sPLS-Regr</li> </ul>
Run Analysis!
Stop Analysis!

**Bar Graph:**

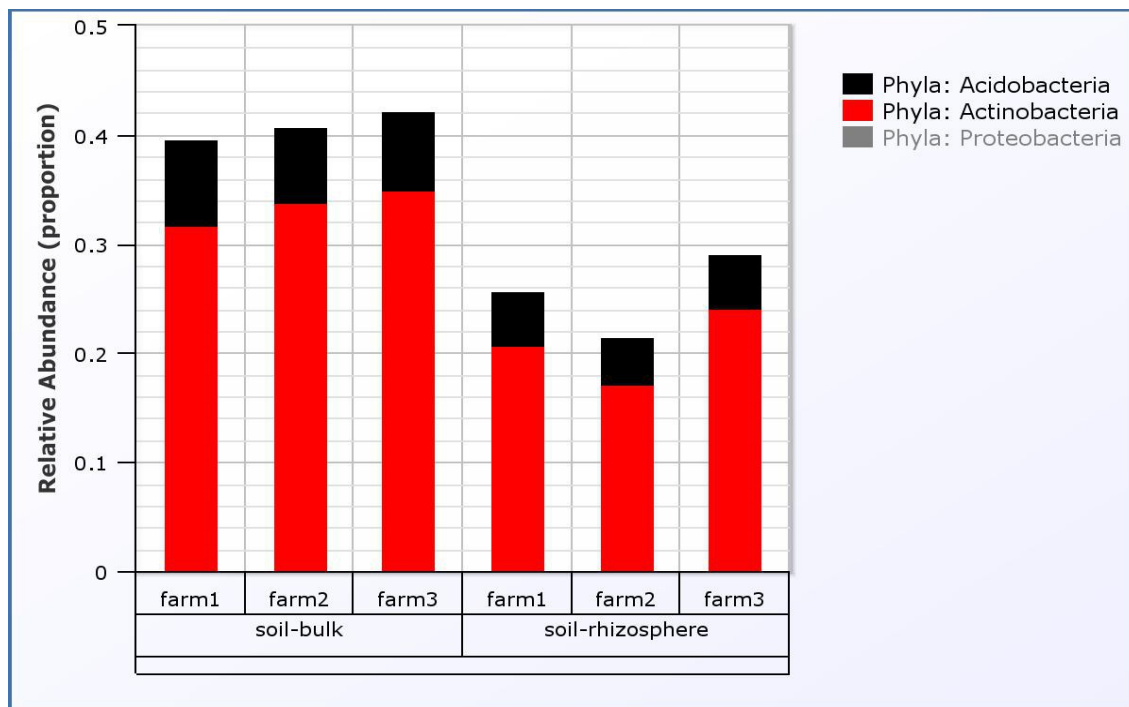
The bar graph will display the taxa averages for each meta variable level selected. The charts produced by myPhyloDB are highly interactive allowing the user to: (i) change the color theme (without rerunning the analysis) using the drop down menu above the graph, (ii) hide individual (by clicking on the legend text) or all (clicking the “Hide all series”) series shown in the graph, and (iii) 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 of the series and new color (name or hex code) you want to change. For your convenience, the current series index, color, and symbol (if applicable) can be displayed by holding the mouse over the appropriate text in the legend.





Below is the same graph after the following changes:

- 1) chart theme changed to: grid
- 2) Actinobacteria series (index: 1) changed to “red”
- 3) Proteobacteria series changed to hidden



Additional notes on ANcOVA graphs:

- 1) Bars represent the arithmetic means
- 2) Only the interaction terms are graphed

***Test Results:***

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

Note: If you are trying 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).

**Test Results:**

```

Categorical variables selected: env_material, geo_loc_farm
Quantitative variables selected: usr_quant1
=====

Data Normalization:
29 selected samples were included in the final analysis.
Data were rarefied to 1000 sequence reads...
=====

=====
Taxa level: Phyla
Taxa name: Acidobacteria
Taxa ID: 9a0841b121364ae485640fb141bfef7d
Dependent Variable: Relative Abundance (proportion)

ANCOVA table:
              Df    Sum Sq  Mean Sq F value Pr(>F)
env_material    1 0.003385 0.003385   3.076 0.0928 .
geo_loc_farm    2 0.000476 0.000238   0.216 0.8070
env_material:geo_loc_farm 2 0.000140 0.000070   0.063 0.9387
Residuals      23 0.025308 0.001100
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

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

```
fit <- aov(y~env_material*geo_loc_farm*usr_quant1, data=df)
summary(fit)

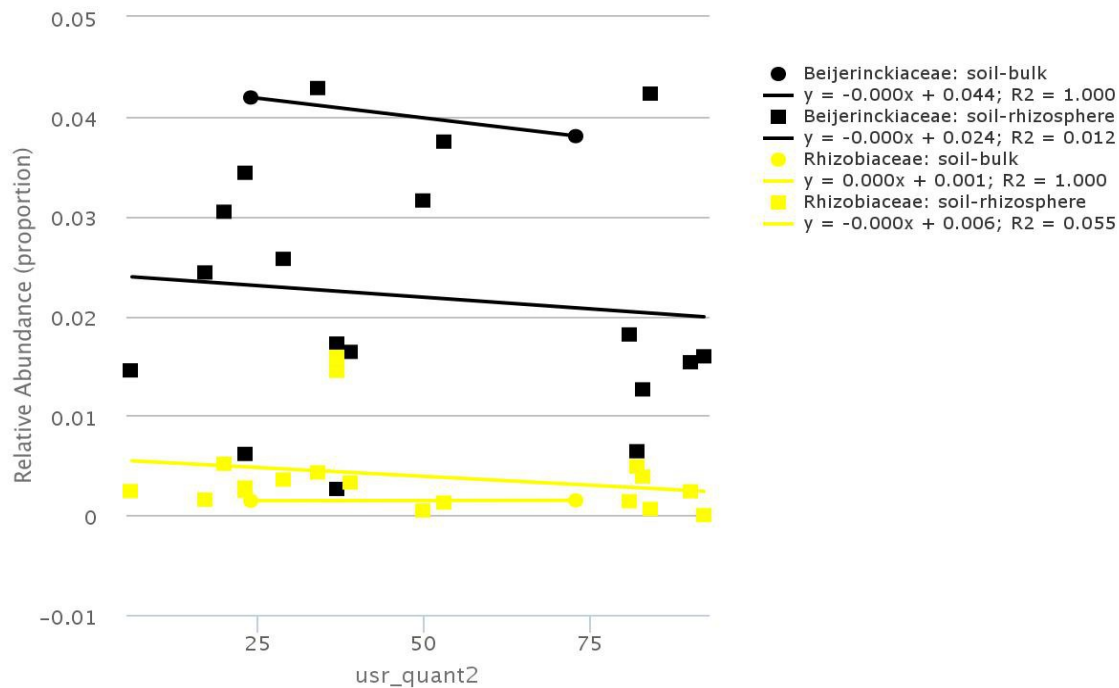
library(lsmeans)
lsm <- lsmeans(fit, list(pairwise~env_material))
lsm <- lsmeans(fit, list(pairwise~geo_loc_farm))
```

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

### Scatter plot (regression):

The GLM analysis is run in a similar manner and will produce a scatter plot instead of a bar graph. Let's run a GLM procedure with the following settings:

Normalization method: DESeq2  
 Minimum sample size: 1000  
 Categorical variable: env\_material  
 Quantitative variable: usr\_quant2  
 Taxa: Beijerinckiaceae, Rhizobiaceae  
 Dependent Variable: Relative Abundance (proportion)



**Test Results:**

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

**Test Results:**

```
Categorical variables selected: env_material
Quantitative variables selected: usr_quant2
=====

Data Normalization:
20 selected samples were included in the final analysis.
9 samples did not met the desired normalization criteria.
DESeq2 cannot run estimateSizeFactors...
Analysis was run without normalization...
To try again, please select fewer samples or another normalization method..
=====

Taxa level: Family
Taxa name: Beijerinckiaceae
Taxa ID: d9aa7e80c400443aa77cc0bb97ac2b44
Dependent Variable: Relative Abundance (proportion)

ANCOVA table:
              Df      Sum Sq   Mean Sq F value  Pr(>F)
usr_quant2      1 0.00003824 0.00003824  0.2395 0.63119
env_material     1 0.00058558 0.00058558  3.6678 0.07352 .
usr_quant2:env_material 1 0.00000117 0.00000117  0.0073 0.93284
Residuals      16 0.00255447 0.00015965
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
fit <- lm(abund~env_material*usr_quant2, data=df)
summary(fit)

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

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.

## 9.2 Diff Abund

The basic layout and data selection of the Diff Abund page is similar to the ANCOVA/GLM page of myPhyloDB. The only differences are (1) the removal of the taxonomic tree (i.e., data can only be selected by taxa level) and (2) the normalization and statistical test options have been changed. The Diff Abund procedure is part of the DESeq2 R package and more details on the procedure can be found [here](#).

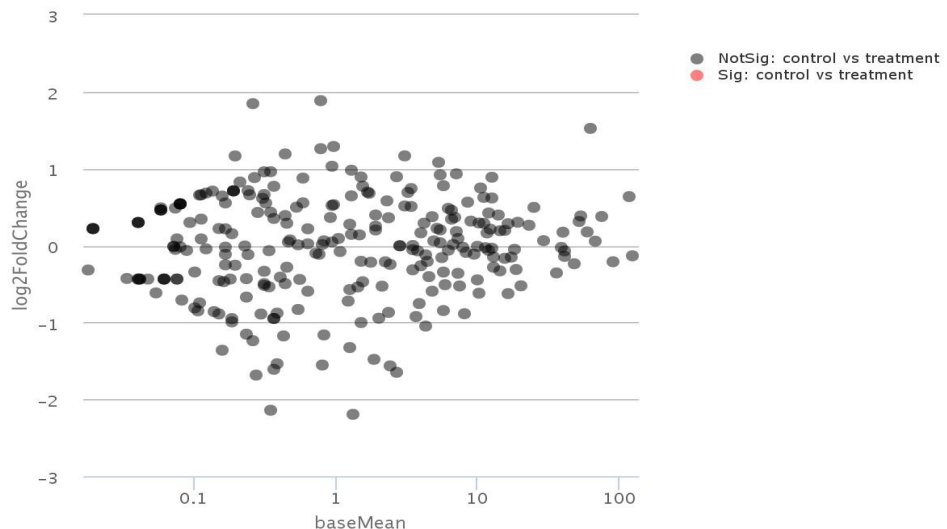
### Graph:

The Diff Abund 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  $\leq 0.05$ ) and non-significant in black. If more than one meta variable is selected, the analysis will compare each independent treatment combination to one another. 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 a tooltip description.

### Settings for graph:

Selected taxa level: species  
 meta-variable: usr\_cat1  
 Dependent variable: Abundance (counts)  
 Normalization method: DESeq2  
 Minimum sample size: 1000  
 Statistical test: nbinomTest  
 False Discovery Rate: 0.10

Note: no significant (red) species were observed for analysis.



this

### Normalization Results:

A summary of the samples meeting your normalization criteria.

### Normalization Results:

```
Data Normalization:
20 selected samples were included in the final analysis.
9 samples did not met the desired normalization criteria.
Data were normalized by DESeq2...
```

***nbinomTest Results:***

The nbinom test results are reported in a sortable DataTable. 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:**

Copy

CSV

Excel

PDF

Print

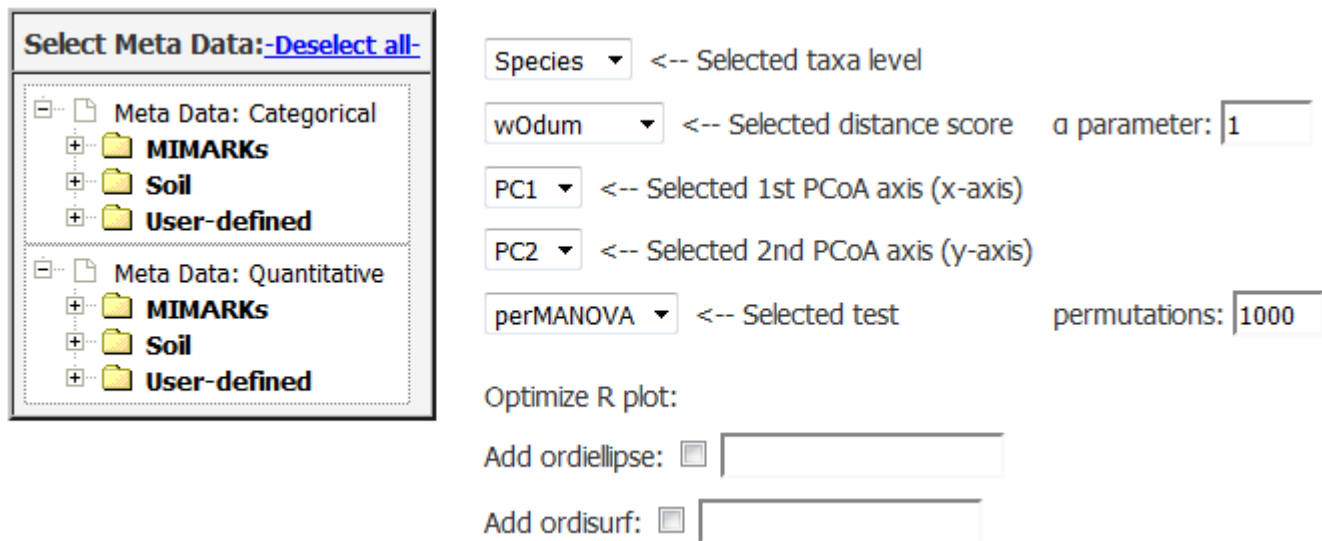
Search:

	Comparison	Taxa ID	Taxa Name	baseMean	baseMeanA	baseMeanB	
0	control vs treatment	019e6ad798bb4438972d1be529719cd1	nanceiensis	0.018096	0.000000	0.040213	
1	control vs treatment	034e3d185a854aae899933f736e78f4e	unclassified	0.041383	0.000000	0.091963	
2	control vs treatment	037583e24c7f465cbb565bec96f6efa8	unclassified	4.880957	5.147568	4.555099	
3	control vs treatment	0450b7a58d1c4ad689ddfbe8e4f49e8f	unclassified	5.149914	5.571366	4.634807	
4	control vs treatment	06dc658ddd9145bda49d59c197797e65	unclassified	1.261747	0.663822	1.992543	
5	control vs treatment	07456a615f2c4f1dbeda95a41d9c4581	unclassified	12.567986	15.420830	9.081177	
6	control vs treatment	07b0214f281a43e483e1b106998b48c1	unclassified	1.787430	1.662142	1.940560	
7	control vs treatment	07b1b4ff7e6843d0b13a4a3aadf4d17c	unclassified	6.931092	7.956551	5.677754	
8	control vs treatment	08157abb773042658a525f0ad9760995	unclassified	0.158248	0.000000	0.351662	
9	control vs treatment	0898eafb54dd4fa38ac3e60c59b6ccaa	unclassified	6.188450	7.397061	4.711259	
10	control vs treatment	099b3226c4014ce9b0e884dfb5b48016	aerosaccus	0.040489	0.073617	0.000000	
11	control vs treatment	09d41f36fbd3441486a2e62aa9f7f49c	unclassified	0.248267	0.389283	0.075915	
12	control vs treatment	0c9a86cc6cc749ca905e0f0d71914ad7	infantis	0.040489	0.073617	0.000000	
13	control vs treatment	0dc922c0b8b043e1a432dee7ddb882b4	unclassified	15.722548	14.901376	16.726204	
14	control vs treatment	0e44df8f9999492ab50d5c4c786fe9ef	unclassified	0.528519	0.666464	0.359920	
15	control vs treatment	0f5fabd12dd4452b999622c342728926	unclassified	4.982086	8.206380	1.041282	
16	control vs treatment	0fba16a758f844edbe8a7c26774e8e9f	unclassified	6.606890	7.526853	5.482490	
17	control vs treatment	0fcab209df554025ab5a38a04d79d32e	unclassified	1.331341	0.319180	2.568427	

Showing 1 to 19 of 292 entries

### 9.3 PCoA (Principal Coordinates Analysis)

The PCoA analysis page (<http://127.0.0.1:8000/myPhyloDB/PCoA/>) layout and operation is similar to the Diff Abund page except for the addition of several drop-down menus. Currently, the only option available is to run a constrained PcoA analysis using the capscale function provided in the vegan package of R.



The screenshot displays the PCoA analysis interface. On the left, a 'Select Meta Data' panel shows two categories: 'Meta Data: Categorical' and 'Meta Data: Quantitative'. Each category contains three sub-folders: 'MIMARKs', 'Soil', and 'User-defined'. The 'Meta Data: Categorical' folder is expanded, showing its contents. On the right, several parameters are configured: 'Species' is selected for the taxa level; 'wOdm' is selected for the distance score with a parameter of 1; 'PC1' is selected for the 1st PCoA axis (x-axis); 'PC2' is selected for the 2nd PCoA axis (y-axis); 'perMANOVA' is selected for the test with 1000 permutations. At the bottom, there are checkboxes for 'Optimize R plot:', 'Add ordiellipse:', and 'Add ordisurf:', each followed by a text input field.

**Meta variables:** The selection of meta variables is similar to the ANOVA/Regr page.

**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 you would like to use for analysis. All scores are calculated using the vegan package in R, except for MorisitaHorn (custom python script based on the calculator in Mothur) and wOdm. The wOdm score can be used to down-weight either rare ( $\alpha > 1$ ) or abundant ( $\alpha < 1$ ) taxa, as discussed here (Manter and Bakker. 2015. [BioInformatics](#)). When  $\alpha = 1$ , wOdm 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 perAMOVA (adonis) or betaDisper analysis of the selected data using the embedded R vegan package.

Graph (highcharts):

PCoA analysis will produce a two-dimensional ordination plot. The first chart option is created using Highcharts and offers greater flexibility for defining symbols shapes and color using the buttons at the bottom of the graph table.

#### Settings for graph:

meta-variable: usr\_cat1

Selected taxa level: species

Selected distance score: Bray-Curtis

Selected 1<sup>st</sup> PCoA axis (x-axis): PC1

Selected 2<sup>nd</sup> PCoA axis (y-axis): PC2

Selected test: perMANOVA permutations: 1000

Optimize R plot:

Add ellipse: usr\_cat1

Add ordisurf: <blank>

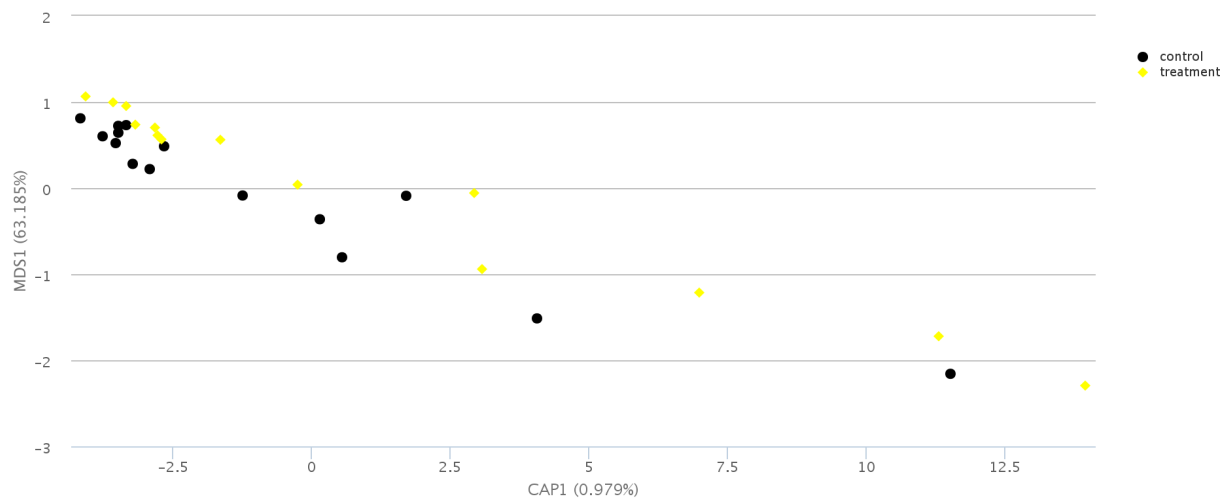
Dependent variable: Abundance (counts)

Normalization method: Rarefaction (keep)

Subsample size: median

Iterations: 10

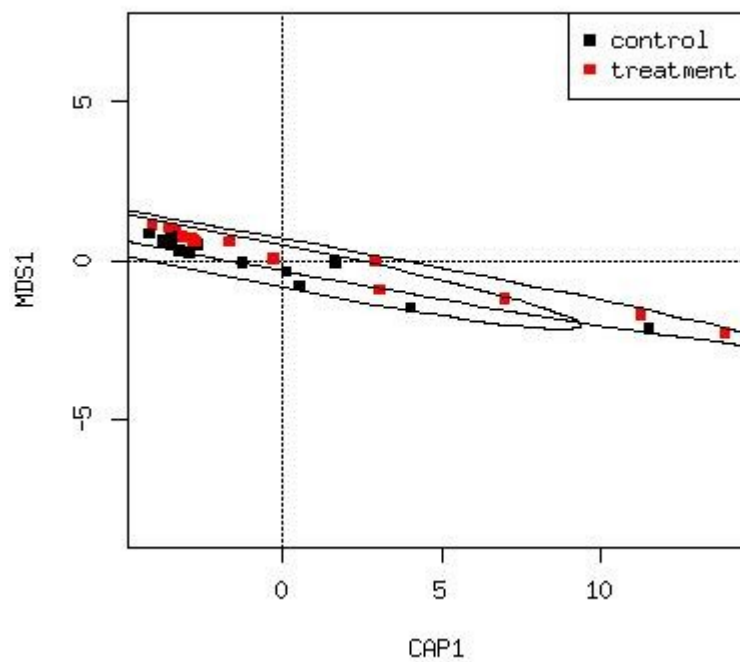
#### Graph (Highcharts):





**Graph (R plot):**

In addition, an ordination plot is also created using the vegan package in R, which displays a 95% confidence interval and overlays splines for any selected quantitative variables (ordisurf). Here is the R ordination plot for the analysis shown above.



***Test Results:***

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 perAMOVA 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.

**Test Results:**

```
Taxa level: Species
Distance score: Bray-Curtis
Categorical variables selected: usr_cat1
Quantitative variables selected:
=====

Data Normalization:
29 selected samples were included in the final analysis.
Data were rarefied to 1584 sequence reads...
=====

perMANOVA results:
Permutation: free
Number of permutations: 999

Terms added sequentially (first to last)
```

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)
usr_cat1	1	0.0796	0.079555	0.47135	0.01716	0.91
Residuals	27	4.5571	0.168780		0.98284	
Total	28	4.6366			1.00000	

```
=====
```

***Principal Coordinates and Distance Scores:***

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).

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:

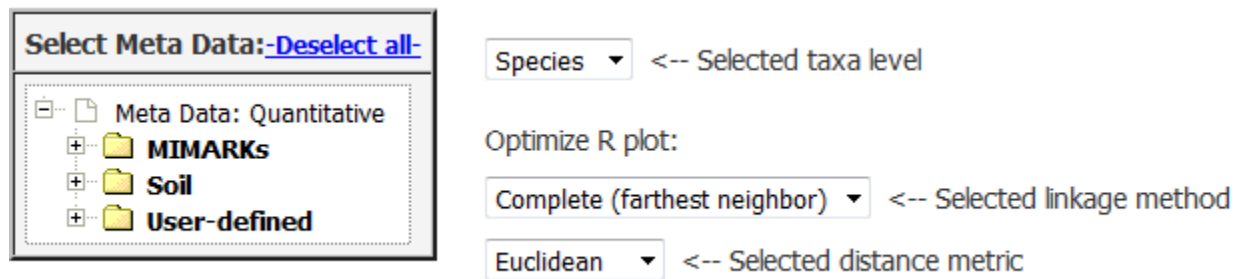
```
library(vegan)
dist <- vegdist(data, method='manhattan')
mat <- as.matrix(dist, diag=TRUE, upper=TRUE)
ord <- capscale(mat ~ geo_loc_farm*env_material, meta)
cat <- factor(meta$geo_loc_farm)

plot(ord, type='n')
points(ord, display='sites', pch=15, col=cat, legend=TRUE)
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_quant2, add=TRUE)
```

where data is the normalized abundance for each taxa chosen, meta is a dataframe containing the appropriate metadata. geo\_loc\_farm, env\_material, and usr\_quant2 were the meta-variables chosen for this analysis.

#### 9.4. sPLS-Regr

The sPLS (sparse partial least squares regression) analysis page (<http://127.0.0.1:8000/myPhyloDB/sPLS/>) layout and operation is also similar to the Diff Abund page except for the addition of two plotting options and drop-down menus. The sPLS 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.



Select Meta Data: [-Deselect all-](#)

- Meta Data: Quantitative
  - MIMARKS
  - Soil
  - User-defined

Species ▼ <-- Selected taxa level

Optimize R plot:

Complete (farthest neighbor) ▼ <-- Selected linkage method

Euclidean ▼ <-- Selected distance metric

Meta variables: The selection of meta variables is similar to the ANOVA/Regr page.

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

The two additional dropdown boxes are used to generate a clustered heatmap (pheatmap package).

**Settings for graph:**

meta-variable: usr\_quant1, usr\_quant2, usr\_quant3

Selected taxa level: species

Optimize R plot:

Selected linkage method: Complete (farthest neighbor)

Selected distance metric: Euclidean

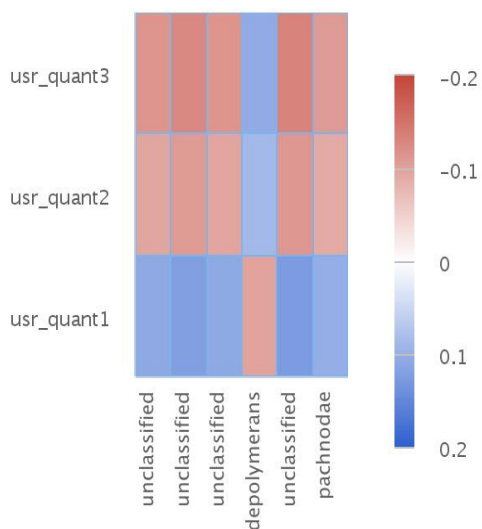
Dependent variable: Abundance (proportion)

Normalization method: Rarefaction (remove)

Subsample size: min

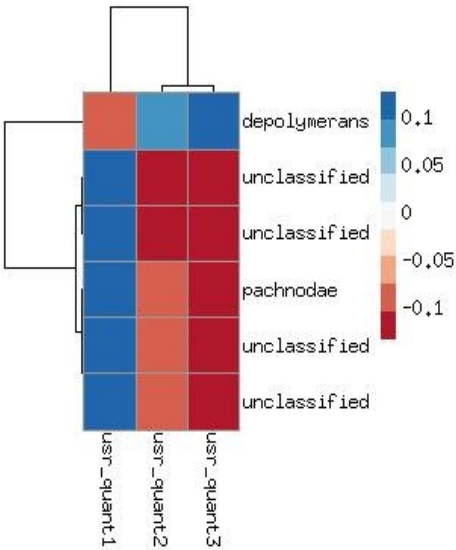
*Iterations: 10***Graph (highcharts):**

sPLS analysis will produce an heatmap of the sPLS correlation coefficients. The first chart option is created using Highcharts and has a tooltip feature to explore the values of each cell within the graph in more detail.



**Graph (R plot):**

The second graph option is a clustered heatmap generated using the pheatmap package in R.



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

```
library(mixOmics)
ZeroVar <- nearZeroVar(X, freqCut=90/10, uniqueCut=25)
List <- row.names(ZeroVar$Metrics)
X_new <- X[,-which(names(X) %in% List)]
X_scaled <- scale(X_new, center=TRUE, scale=TRUE)
Y_scaled <- scale(Y, center=TRUE, scale=TRUE)

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(DMwR)
pred.ns <- unscale(pred.f, Y_scaled)

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.

## 10. Normalization

myPhyloDB provides several options for normalizing your sequence data to a common sampling depth: none, rarefaction (remove), rarefaction (keep), proportion, and DESeq2.

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. In this case, myPhyloDB will run 10 independent sub-samplings of your data and use the average phylotype abundances for analysis.

A brief description of each procedure follows.

**None:** no normalization

**Rarefaction (remove):** This normalization procedure performs a typical sub-sampling without replacement to the desired subsample 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 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 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 the detection level, in the observed data. The Lidstone 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.

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

**DESeq2:** A discussion can be found [here](#).



## 11. Manage Users

### New feature in v.1.1. 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).

### New feature in v.1.1. Manage Users

Allows users to access the user administration pages. Access to the user administration pages requires a superuser or staff account. The default superuser for myPhyloDB is as follows:

username: admin

password: admin

email: [admin@example.com](mailto: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 '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 Administration table.

## 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 [myphylobd@gmail.com](mailto:myphylobd@gmail.com) or the users' forum at [www.myphylobd.org](http://www.myphylobd.org) with as much detail describing your analysis selections as possible.