



## Overview:

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 are marked as: **New feature in v.1.1.**

Please visit our website for additional information and tutorials:

<http://www.myphylobd.org>

If you use myPhyloDB, please use the following citation:

Manter DK, M Korsa, C Tebbe, JA Delgado. 20xx (in press). myPhyloDB: a local web server for the storage and analysis of metagenomic data. Databases

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

[Daniel Manter](#)

Soil-Plant-Nutrient 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. 3
3.	Uploading New Data	p. 4
4.	Reanalyzing Data	p. 8
5.	Updating Metadata	p. 9
6.	Selecting Data for Analysis	p. 10
7.	Export Data	p. 12
8.	Search Taxa	p. 13
9.	Analysis	p. 14
10.	Normalization	p. 25
11.	Manage Users	p. 36

## 1. Installation

### Windows:

Double-click the installer (myPhyloDB\_v.1.1\_Win\_x64\_install.exe) and follow the prompts. If you are upgrading/reinstalling myPhyloDB and would like to keep your current database, make sure the “Default database” is unchecked during installation; otherwise, all components should be selected. The program will install a myPhyloDB shortcut to your start menu (Windows 7) or start screen (Windows 8). Clicking the myPhyloDB icon will open a terminal and web-browser that provides the user-interface for running the myPhyloDB program. To exit the program type ctrl-c in the terminal and manually close your browser. Uninstalling myPhyloDB will not remove your database or uploaded files. The default installation folder for myPhyloDB will be: 'C:\Users\<user\_name>\AppData\Local\myPhyloDB'.

### Linux:

Double-click or run the installer from your terminal (myPhyloDB\_v.1.1\_Linux\_x64\_install.sh). If a previous version of myPhyloDB is detected you will be prompted to either keep your old database or re-install the default database. The program will install a myPhyloDB shortcut to your Desktop. Clicking the myPhyloDB icon will open a terminal and web-browser that provides the user-interface for running the myPhyloDB program. To exit the program type ctrl-c in the terminal and manually close your browser. MyPhyloDB must be manually uninstalled by deleting the appropriate folders. The default installation folder for myPhyloDB will be: 'home/<user\_name>/myPhyloDB'.

### Remote access:

myPhyloDB will run as a local server on your host machine allowing others on your local intranet to access myPhyloDB (unless disabled using your computer's firewall settings) without installing a separate copy. This may be useful for laboratories that want to share data across multiple users. To access myPhyloDB from a remote computer you must first obtain the IP address of the host machine (in a terminal on the host machine, type 'ipconfig' for Windows or 'ifconfig' for Linux), then in the address bar of your remote computer's browser enter the following address 'xxx.xxx.x.xx:8000/myPhyloDB/home/' replacing the x's with the appropriate IP address. 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 logged in as guest – Data Mgt (select data only) and Analysis links visible.

A	B	C	D
<b>Menu</b>	User logged in as: admin	User logged in as: admin	User logged in as: guest
<b>General Info</b>	<b>Menu</b>	<b>Menu</b>	<b>Menu</b>
<ul style="list-style-type: none"> <li>• Home</li> <li>• Login</li> </ul>	<b>General Info</b>	<b>General Info</b>	<b>General Info</b>
	<ul style="list-style-type: none"> <li>• Home</li> <li>• Logout</li> <li>• Manage Users</li> </ul>	<ul style="list-style-type: none"> <li>• Home</li> <li>• Logout</li> <li>• Manage Users</li> </ul>	<ul style="list-style-type: none"> <li>• Home</li> <li>• Login</li> </ul>
<b>Data Mgt</b>	<b>Data Mgt</b>	<b>Data Mgt</b>	<b>Data Mgt</b>
<ul style="list-style-type: none"> <li>• Select Data</li> </ul>	<ul style="list-style-type: none"> <li>• [Upload]</li> <li>• [Reprocess]</li> <li>• [Update]</li> <li>• Select Data</li> </ul>	<ul style="list-style-type: none"> <li>• [Upload]</li> <li>• [Reprocess]</li> <li>• [Update]</li> <li>• Select Data</li> </ul>	<ul style="list-style-type: none"> <li>• Select Data</li> </ul>
<b>Taxonomy</b>	<b>Taxonomy</b>	<b>Taxonomy</b>	<b>Taxonomy</b>
<ul style="list-style-type: none"> <li>• Search Taxa</li> </ul>	<ul style="list-style-type: none"> <li>• Search Taxa</li> </ul>	<ul style="list-style-type: none"> <li>• Search Taxa</li> </ul>	<ul style="list-style-type: none"> <li>• Search Taxa</li> </ul>
		<b>Analysis</b>	<b>Analysis</b>
		<b>Univariate</b>	<b>Univariate</b>
		<ul style="list-style-type: none"> <li>• ANOVA/Regr</li> </ul>	<ul style="list-style-type: none"> <li>• ANOVA/Regr</li> </ul>
		<b>Multivariate</b>	<b>Multivariate</b>
		<ul style="list-style-type: none"> <li>• Diff Abund</li> <li>• PCoA</li> </ul>	<ul style="list-style-type: none"> <li>• Diff Abund</li> <li>• PCoA</li> </ul>

### 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:
Browse...
No file selected.

2.) Select sequence data format:

Available Data Formats:
Pre-processed Mothur Files

3.) Select sequencing files:

Select conserved taxonomy file:
Browse...
No file selected.

Select .shared file:
Browse...
No file selected.

Upload Files

#### 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 Select metadata file

Each upload requires a completed metadata file, which can be downloaded from myPhyloDB's homepage. Columns (variables) are protected and may not be changed. 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 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 Select sequencing files

#### 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).

#### 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.

Experienced mothur users may wish to alter the provided batch file to match their current sequencing analysis pipelines; however, please note that the pipeline must create the following 5 files (final.fasta, final.names, final.groups, final.taxonomy, and final.shared); any deviation from the above naming conventions and the upload process will fail.

#### 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.

Experienced mothur users may wish to alter the provided batch file to match their current sequencing analysis pipelines; however, please note that the pipeline must create the following 5 files (final.fasta, final.names, final.groups, final.taxonomy, and final.shared); any deviation from the above naming conventions and the upload process will fail.

#### 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. Also, please be sure that you select all of the appropriate fastq files using the available fastq file chooser, which supports multiple file selection.

Experienced mothur users may wish to alter the provided batch file to match their current sequencing

analysis pipelines; however, please note that the pipeline must create the following 5 files (final.fasta, final.names, final.groups, final.taxonomy, and final.shared); any deviation from the above naming conventions and the upload process will fail.

### 3.1.5 Upload Files

Once you have completed the three steps above, click “Upload Files” to begin the upload process. Note: the upload process includes multiple steps and may take anywhere from a few minutes to hours depending upon the project size and your computer speed. For your convenience, a progress bar will appear below the “Upload Files” button documenting the status of the upload and parsing steps required to populate the myPhyloDB database.

### 3.2 Removing Data

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. Edited projects (i.e., new submission files) can then be uploaded as described above.

#### 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

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

To update a previously uploaded project with new smetaddata, 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:

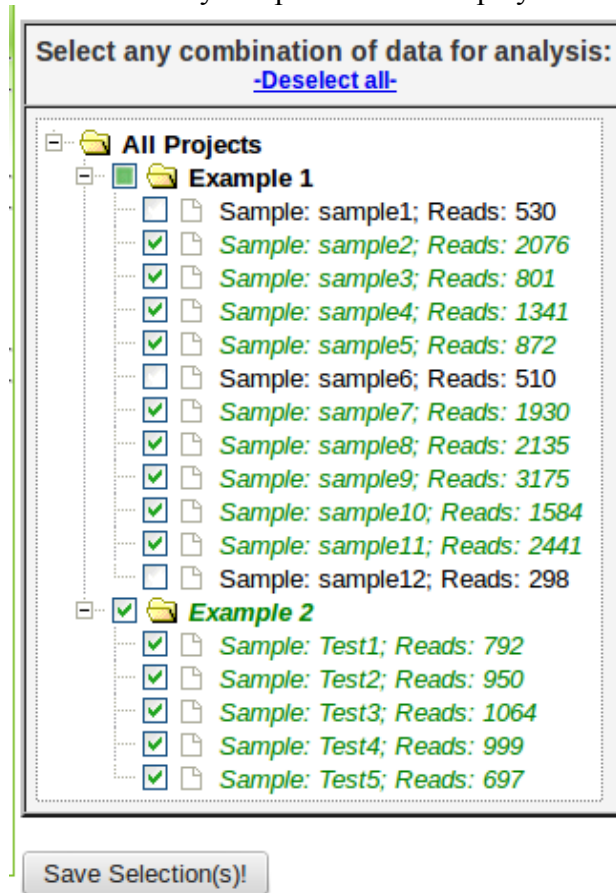
Browse...

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 “Show Selection(s) in DataTable!”. 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.

#### Project/Sample information for selected samples

Show Selection(s) in DataTable!

Project   MIMARKS   Soil   User-defined							
Show 10 entries		Search:		<a href="#">Copy</a> <a href="#">CSV</a> <a href="#">Excel</a> <a href="#">PDF</a> <a href="#">Print</a>			
project_name	project_id	sample_name	project_desc	start_date	end_date	pi_last	pi_first
Example 2	f91ba37bade04360b1ad7b7f419f6398	Test1	Raw soil data for myPhyloDB	null	null	null	null
Example 2	f91ba37bade04360b1ad7b7f419f6398	Test2	Raw soil data for myPhyloDB	null	null	null	null
Example 2	f91ba37bade04360b1ad7b7f419f6398	Test3	Raw soil data for myPhyloDB	null	null	null	null
Example 2	f91ba37bade04360b1ad7b7f419f6398	Test4	Raw soil data for myPhyloDB	null	null	null	null
Example 2	f91ba37bade04360b1ad7b7f419f6398	Test5	Raw soil data for myPhyloDB	null	null	null	null

Showing 1 to 5 of 5 entries

[Previous](#)
[1](#)
[Next](#)

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

## 7. Export Data

### New feature in v.1.1.

#### 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 (page 12).

#### 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:

Show  entriesSearch: 
[Copy](#) [CSV](#) [Excel](#) [PDF](#) [Print](#)

	Kingdom Name	Kingdom ID	Phylum Name	Phylum ID	Class Name
0	Bacteria	feaf34ed21cf42c68534209baa1f51f0	Proteobacteria	22bf9fbd71fd44fca58e573fbd9f987d	Gammaproteobacteria
1	Bacteria	feaf34ed21cf42c68534209baa1f51f0	Proteobacteria	22bf9fbd71fd44fca58e573fbd9f987d	Deltaproteobacteria
2	Bacteria	feaf34ed21cf42c68534209baa1f51f0	Proteobacteria	22bf9fbd71fd44fca58e573fbd9f987d	Betaproteobacteria
3	Bacteria	feaf34ed21cf42c68534209baa1f51f0	Proteobacteria	22bf9fbd71fd44fca58e573fbd9f987d	Gammaproteobacteria
4	Bacteria	feaf34ed21cf42c68534209baa1f51f0	Proteobacteria	22bf9fbd71fd44fca58e573fbd9f987d	Alphaproteobacteria
5	Bacteria	feaf34ed21cf42c68534209baa1f51f0	Proteobacteria	22bf9fbd71fd44fca58e573fbd9f987d	Alphaproteobacteria
6	Bacteria	feaf34ed21cf42c68534209baa1f51f0	Proteobacteria	22bf9fbd71fd44fca58e573fbd9f987d	Gammaproteobacteria
7	Bacteria	feaf34ed21cf42c68534209baa1f51f0	Proteobacteria	22bf9fbd71fd44fca58e573fbd9f987d	Alphaproteobacteria
8	Bacteria	feaf34ed21cf42c68534209baa1f51f0	Acidobacteria	1fa0628580bd4615ad56701c82622145	Acidobacteriia
9	Bacteria	feaf34ed21cf42c68534209baa1f51f0	Bacteroidetes	c03f4a25bef84d3db03394559e08a6c6	Flavobacteriia

Showing 1 to 10 of 481 entries

[Previous](#) [1](#) [2](#) [3](#) [4](#) [5](#) .. [49](#) [Next](#)

## 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).

### 9.1. ANcOVA

**New feature in v.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.

## 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
<input type="checkbox"/> <b>Bacteria</b>
<input checked="" type="checkbox"/> <b>Acidobacteria</b>
<input checked="" type="checkbox"/> <b>Actinobacteria</b>
<input type="checkbox"/> <b>Armatimonadetes</b>
<input type="checkbox"/> <b>Bacteroidetes</b>
<input type="checkbox"/> <b>Chloroflexi</b>
<input type="checkbox"/> <b>Cyanobacteria</b>
<input type="checkbox"/> <b>Elusimicrobia</b>
<input type="checkbox"/> <b>FBP</b>
<input type="checkbox"/> <b>Fibrobacteres</b>
<input type="checkbox"/> <b>Firmicutes</b>
<input type="checkbox"/> <b>Gemmatimonadetes</b>
<input type="checkbox"/> <b>Nitrospirae</b>
<input type="checkbox"/> <b>OC31</b>
<input type="checkbox"/> <b>Planctomycetes</b>
<input checked="" type="checkbox"/> <b>Proteobacteria</b>
<input type="checkbox"/> <b>SAR406</b>
<input type="checkbox"/> <b>TM7</b>
<input type="checkbox"/> <b>Thermi</b>
<input type="checkbox"/> <b>Verrucomicrobia</b>
<input type="checkbox"/> <b>WPS-2</b>
<input type="checkbox"/> <b>WS3</b>
<input type="checkbox"/> <b>unknown</b>

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 abundance (counts) and the data will be normalized using the rarefaction (keep) method. In addition, all samples will be sub-sampled to 1000 sequence reads and use the average values from 10 independent iterations.

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

GraphData		
Dependent variable: Abundance (counts) ▼	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.

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**

*General Info*

- [Home](#)
- [Login](#)

*Data Mgt*

- [Select Data](#)
- [Export Data](#)

*Taxonomy*

- [Search Taxa](#)

*Analysis*

**Univariate**

- [ANCOVA](#)

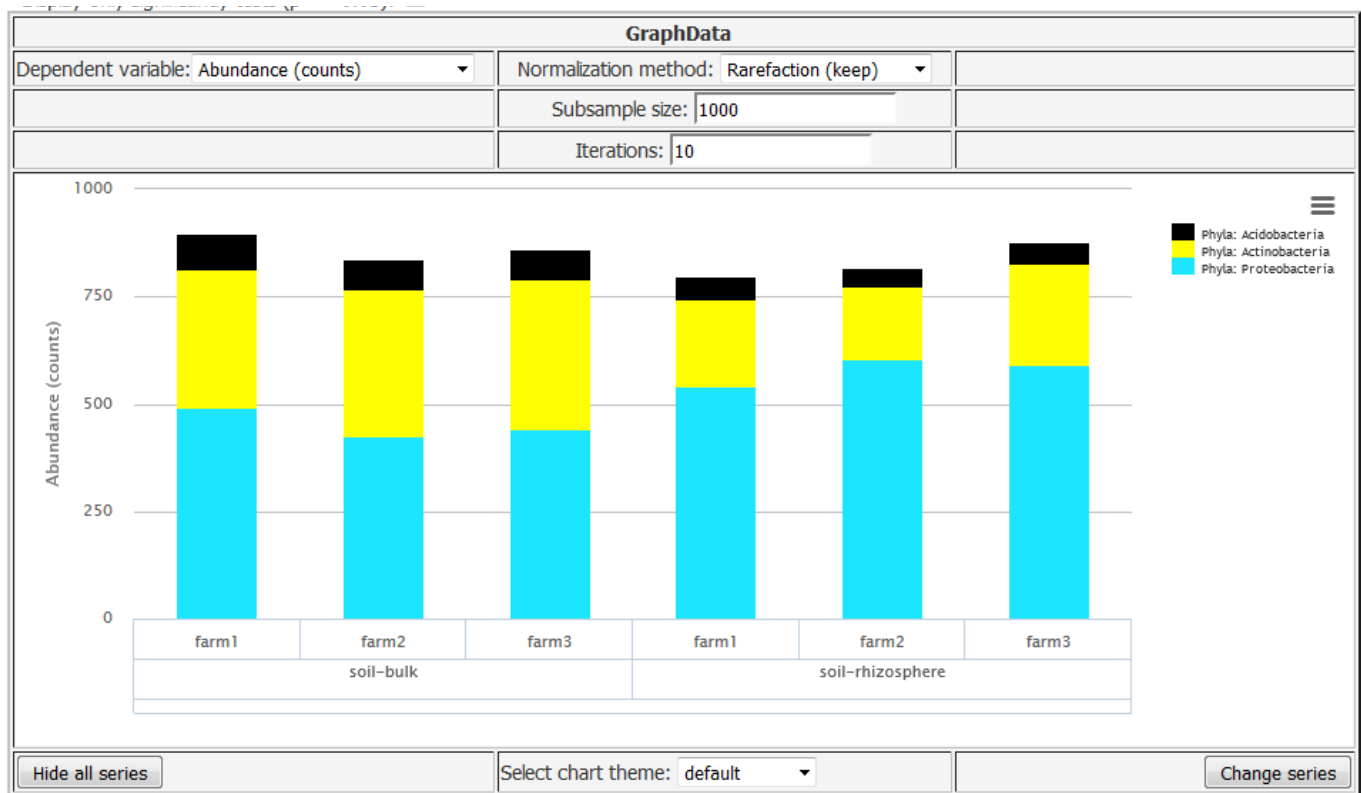
**Multivariate**

- [Diff Abund](#)
- [PCoA](#)
- [sPLS-Regr](#)

Run Analysis!

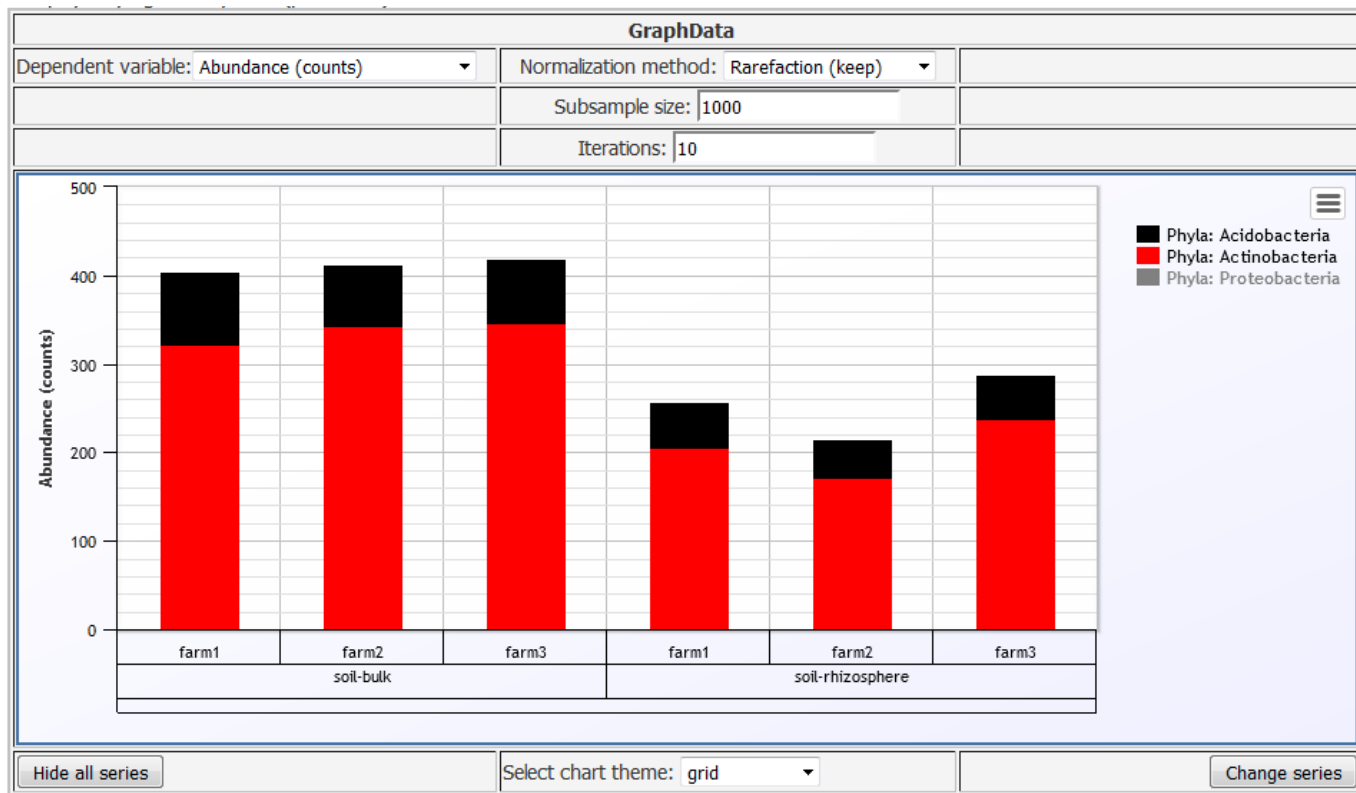
## Bar Graph:

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.

```

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

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: 1b9bc47c779247999fec7732ccf17236
Dependent Variable: Abundance (counts)

ANCOVA table:
              Df Sum Sq Mean Sq F value Pr(>F)
env_material    1   3514    3514    3.136 0.0898 .
geo_loc_farm     2    324     162    0.145 0.8660
env_material:geo_loc_farm 2     35      17    0.015 0.9847
Residuals      23  25773    1121
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

LSmeans & Tukey's HSD post-hoc test:

$`lsmeans of env_material`
      env_material      lsmean      SE df lower.CL upper.CL
soil-bulk          75.00000 13.665952 23 46.72982 103.27018
soil-rhizosphere  47.91964  6.993774 23 33.45192  62.38737

Results are averaged over the levels of: geo_loc_farm
Confidence level used: 0.95

$`pairwise differences of contrast`
      contrast      estimate      SE df t.ratio p.value
soil-bulk - soil-rhizosphere 27.08036 15.35158 23  1.764  0.0910

Results are averaged over the levels of: geo_loc_farm

```

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):

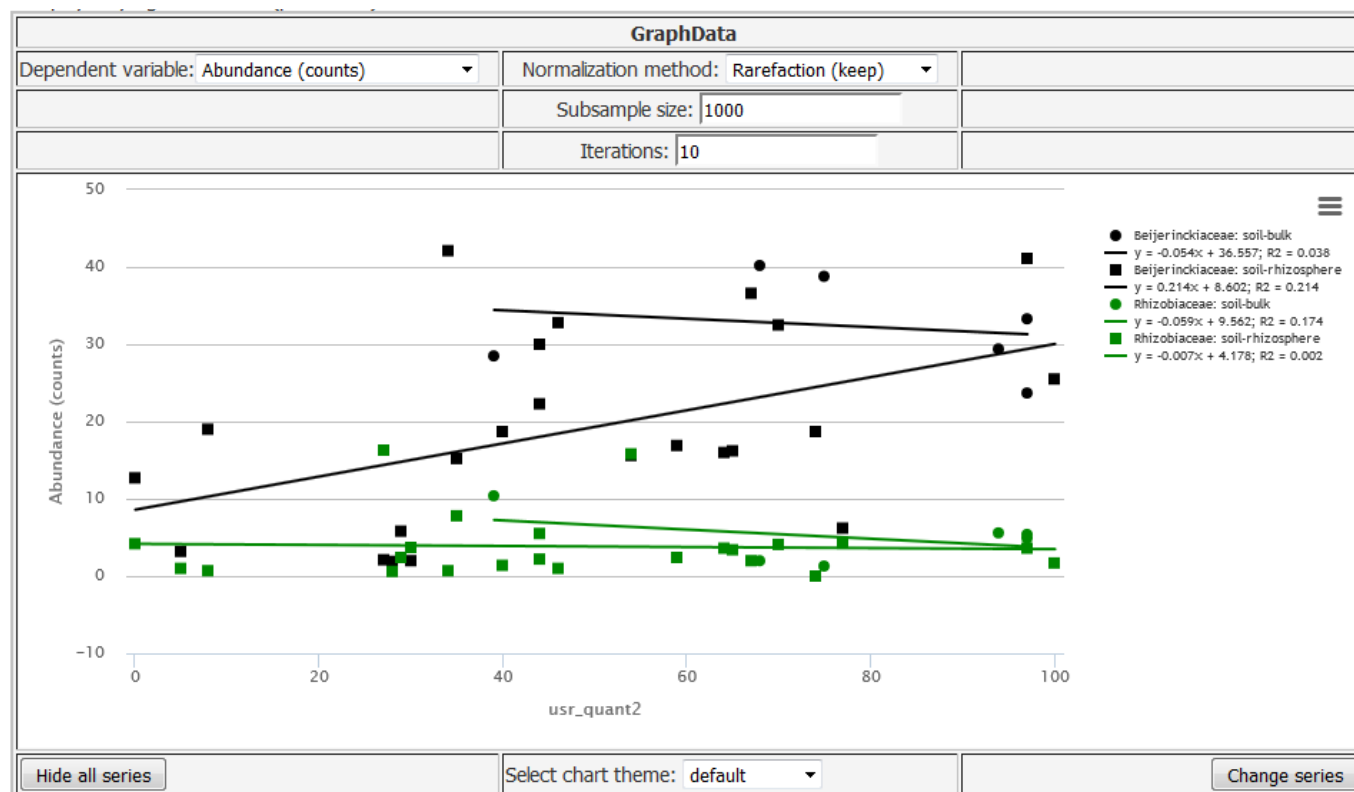
The GLM analysis is run in a similar manner and will produce a scatter plot instead of a bar graph. Let's run a GLM procedure with the following settings:

Categorical variable: env\_material

Quantitative variable: usr\_quant2

Taxa: Beijerinckiaceae, Rhizobiaceae

Dependent Variable: Abundance



## 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.

```

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

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

Taxa level: Family
Taxa name: Beijerinckiaceae
Taxa ID: a8382984f0664cc2a91ceffe06451d99
Dependent Variable: Abundance (counts)

ANCOVA table:
      Df  Sum Sq Mean Sq F value    Pr(>F)
env_material      1 1020.50   1020.50   8.8519 0.006408 **
usr_quant2        1   525.64    525.64   4.5594 0.042720 *
env_material:usr_quant2  1    97.50    97.50   0.8457 0.366564
Residuals        25 2882.13   115.29
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    34.22655   17.04486   2.008  0.0556 .
env_materialsoil-rhizosphere -24.92403   17.66416  -1.411  0.1706
usr_quant2     -0.01076    0.21028  -0.051  0.9596
env_materialsoil-rhizosphere:usr_quant2  0.20862    0.22685   0.920  0.3666
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.74 on 25 degrees of freedom
Multiple R-squared:  0.3632,    Adjusted R-squared:  0.2868
F-statistic: 4.752 on 3 and 25 DF,  p-value: 0.00933
=====

```

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)

```

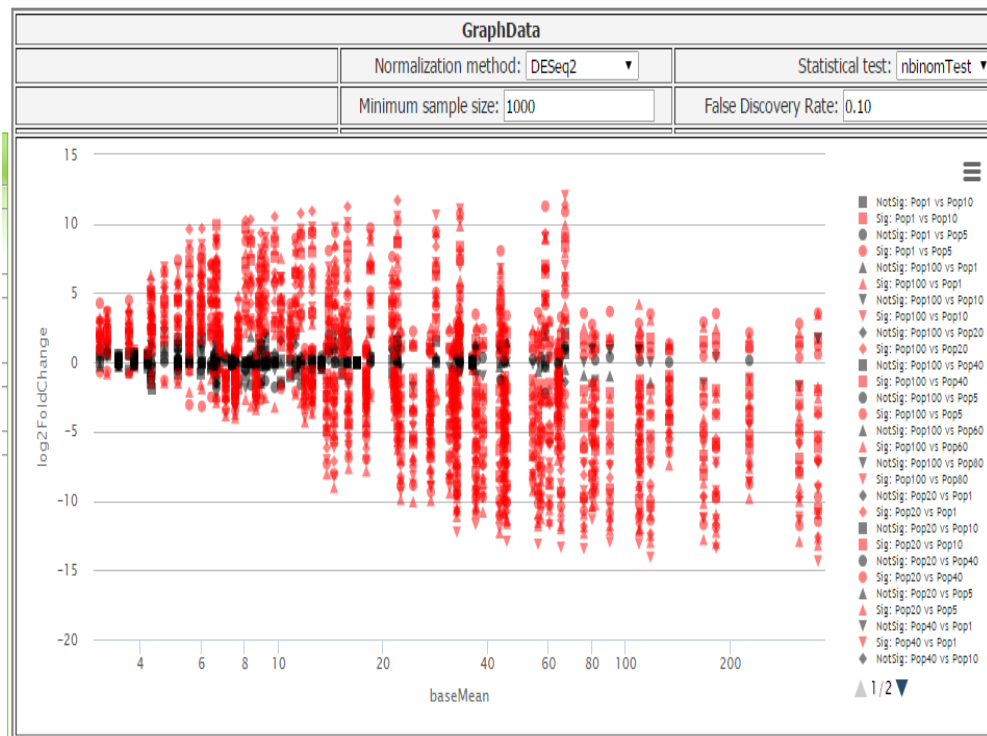
where abund is the normalized abundance for each taxa chosen, df is a dataframe containing the appropriate metadata and normalized abundances. In addition, no post-hoc analysis is performed and all linear regression lines (shown in graph) are calculated using SciPy's 'linregress' function.

## 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.





Normalization Results:

A summary of the samples meeting your normalization criteria.

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:

Show 10 entries
Search:

CopyCSVExcelPDFPrint

	Comparison	Taxa ID	Taxa Name	baseMean	baseMeanA	baseMeanB	log2FoldChange
0	Pop100 vs Pop60	00d167b1ca624880ad60f9ca3781cdb1	Species12	168.370671	0.000000	1.797178	-4.078664e+00
0	Pop100 vs Pop80	00d167b1ca624880ad60f9ca3781cdb1	Species12	168.370671	0.000000	0.234936	-1.521171e+00
0	Pop100 vs Pop20	00d167b1ca624880ad60f9ca3781cdb1	Species12	168.370671	0.000000	77.532797	-9.639622e+00
0	Pop100 vs	00d167b1ca624880ad60f9ca3781cdb1	Species12	168.370671	0.000000	10.248043	6.738533e+00

### 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.

Select Meta Data: [Deselect all](#)

Meta Data: Categorical

- Meta Data: Categorical
  - MIMARKs
  - Soil
  - User-defined
- Meta Data: Quantitative
  - MIMARKs
  - Soil
  - User-defined

Species ▼ <-- Selected taxa level

wOdum ▼ <-- Selected distance score     $\alpha$  parameter: 1

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

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

perMANOVA ▼ <-- Selected test    permutations: 1000

Optimize R plot:

Add ordiellipse: ☐

Add ordisurf: ☐

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 wOdum. The wOdum 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  $\alpha = 1$ , wOdum is equivalent to Bray-Curtis.

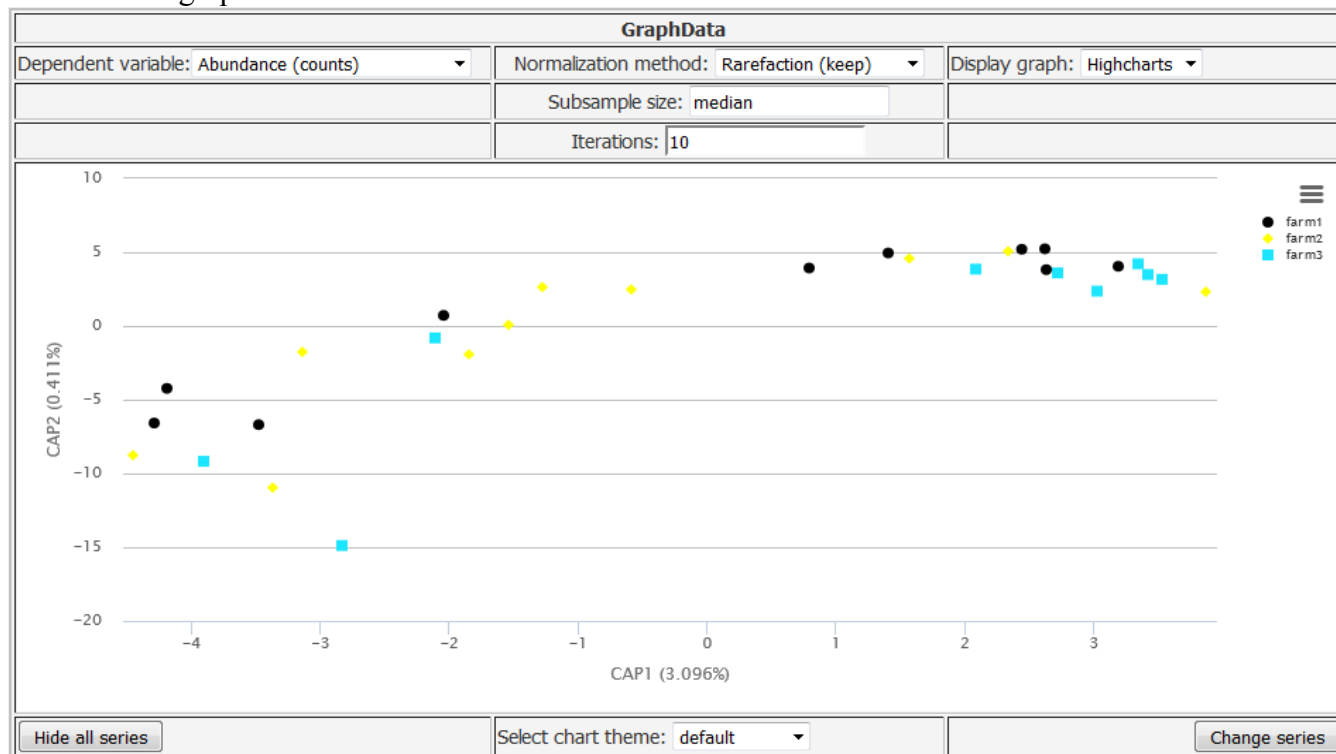
Principal coordinate axis selected (x-axis): This is the axis selected as the x-axis in the displayed graph.

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

Test selected: Select whether you would like to perform either an perAMOVA (adonis) or betaDisper analysis of the selected data using the embedded R vegan package.

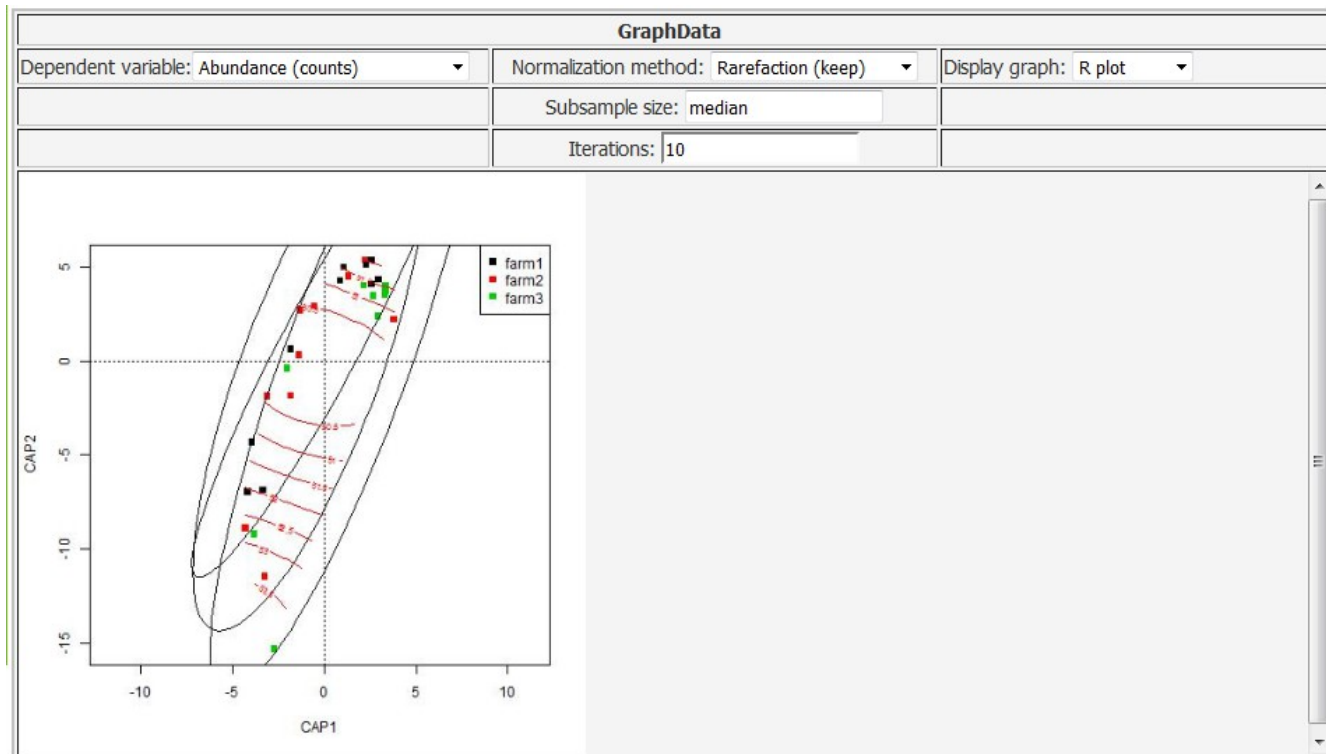
## Graph (highcharts):

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.



## Graph (R plot):

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



## 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.

Number of permutations: 999

Terms added sequentially (first to last)

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)
geo_loc_farm	2	0.06318	0.031591	0.68326	0.04993	0.568
Residuals	26	1.20214	0.046236		0.95007	
Total	28	1.26532			1.00000	

=====

envfit results:

	CAP1	CAP2	r2	Pr(>r)
usr_quant1	0.87147	-0.49044	0.0315	0.648

Permutation: free  
Number of permutations: 999

=====

Eigenvalues

	Stat	CAP1	CAP2	MDS1	MDS2	MDS3	MDS4	MDS5	MDS6	
0	Eigenvalue	0.323337	0.043964	7.844265	0.895674	0.242807	0.162904	0.119623	0.081991	0.0
1	Proportion Explained	0.032700	0.004450	0.793270	0.090580	0.024550	0.016470	0.012100	0.008290	0.0
2	Cumulative Proportion	0.032700	0.037140	0.830420	0.921000	0.945550	0.962020	0.974120	0.982410	0.9

=====

## 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:

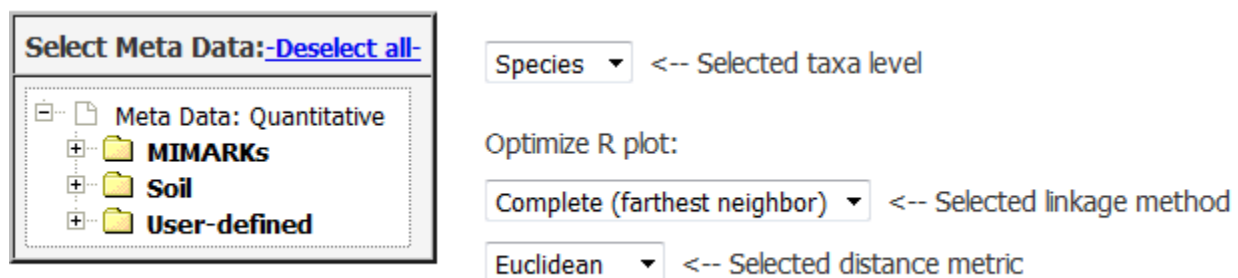
```
library(vegan)
dist <- vegdist(data, method='manhattan')
mat <- as.matrix(dist, diag=TRUE, upper=TRUE)
ord <- capscale(mat ~ geo_loc_farm*env_material, meta)
cat <- factor(meta$geo_loc_farm)

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 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. several drop-down menus. The sPLS analysis is run using the spls package of R.



The screenshot displays the sPLS analysis interface. On the left, a window titled "Select Meta Data: -Deselect all-" contains a tree view under "Meta Data: Quantitative" with three sub-items: "MIMARKs", "Soil", and "User-defined", each preceded by a plus sign. To the right of this window are three dropdown menus. The first is labeled "Species" with a downward arrow and the text "<-- Selected taxa level". Below it is the label "Optimize R plot:" followed by a dropdown menu set to "Complete (farthest neighbor)" with the text "<-- Selected linkage method". The third dropdown menu is set to "Euclidean" with the text "<-- Selected distance metric".

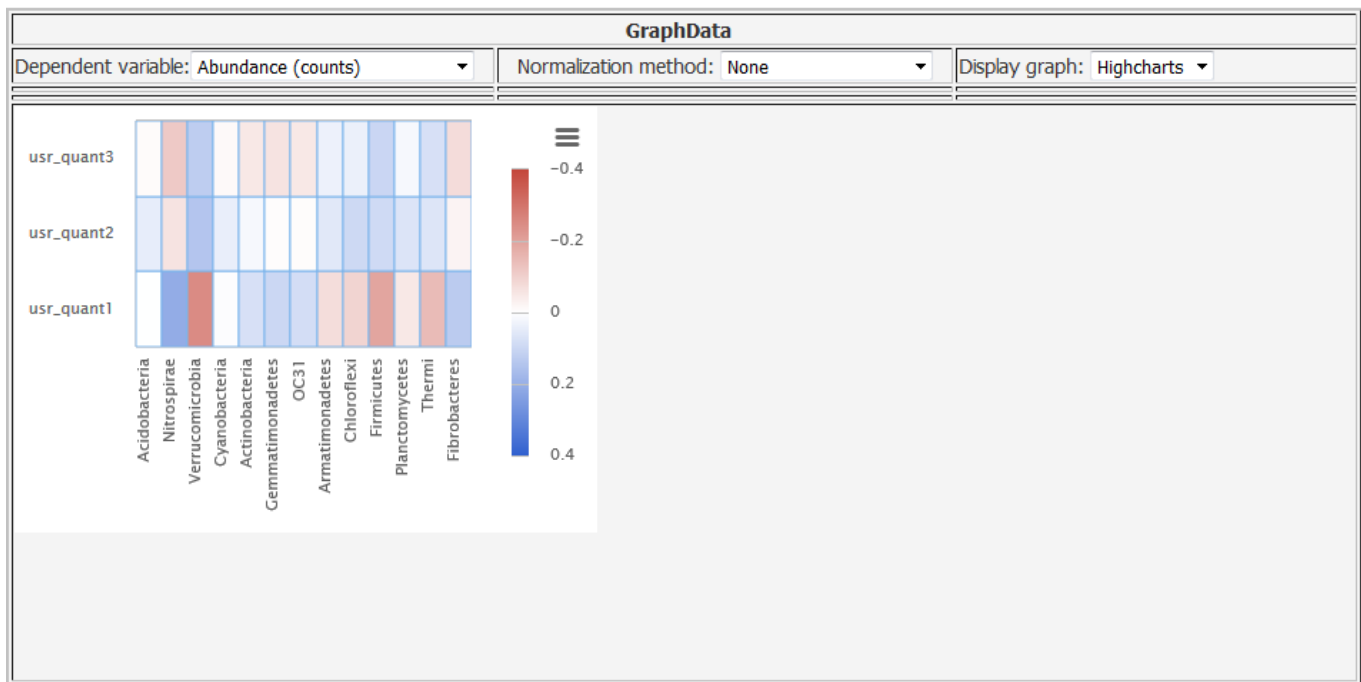
Meta variables: The selection of meta variables is similar to the ANOVA/Regr page.

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).

Graph (highcharts):

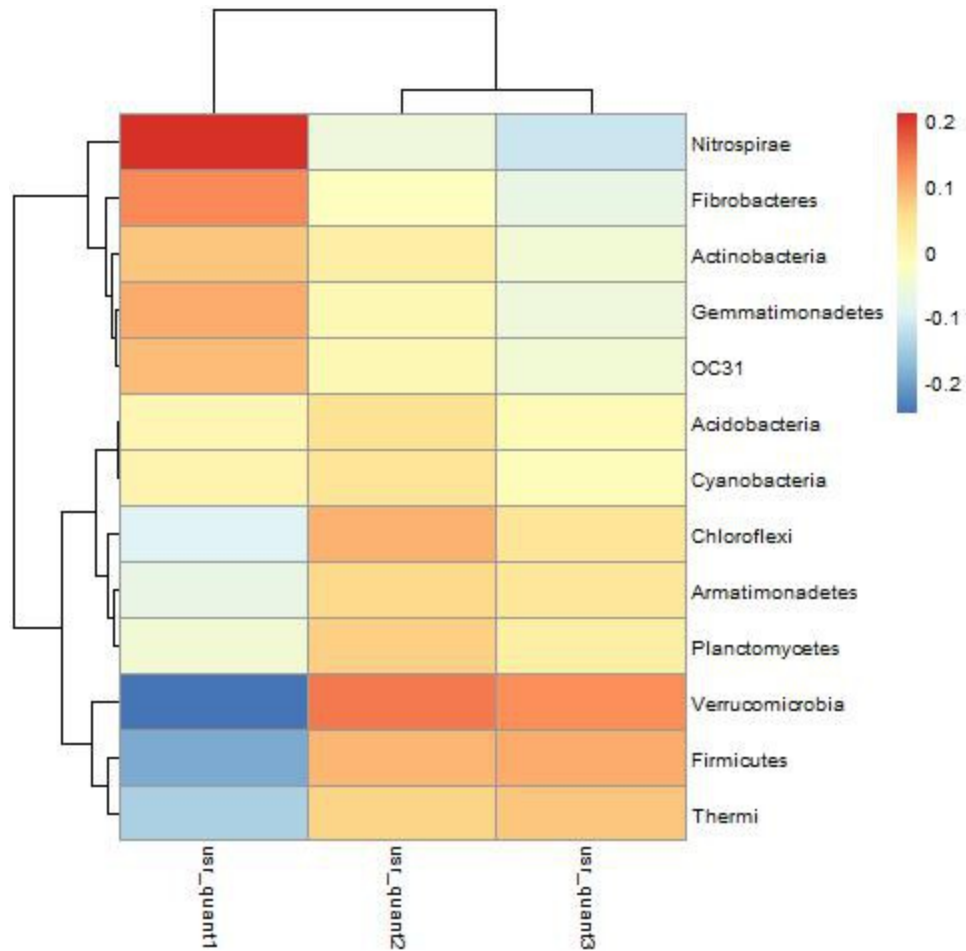
sPLS analysis will produce an heatmap of the sPLS correlation coefficients. The first chart option is created using Highcharts and has a tooltip feature to explore the values of each cell within the graph in more detail.





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 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](#).

## 10. 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 will be provided with a registration link.

### 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.