TAXAenv

Tutorial v0.1

Project Team

Umer Zeeshan Ijaz
Research Fellow (Infrastructure and Environment) University of Glasgow, School of Engineering, Glasgow
http://userweb.eng.gla.ac.uk/umer.ijaz/

Rob van Son
Head of Department
Department of Head and Neck cancer, Netherlands Cancer Institute, Also: ACLC, Fac. Humanities, University of Amsterdam
http://www.fon.hum.uva.nl/rob/

Christopher Quince
Reader in Biological Systems Modelling
University of Glasgow, School of Engineering, Glasgow
http://userweb.eng.gla.ac.uk/christopher.quince/

Main Features

- Usefulness
  - **TAXA**env is useful for multivariate analysis of microbial community structure in an environmental context. Microbial diversity is measured by sequencing homologous genes, typically the 16S rRNA, through the next-generation sequencing platforms. After extracting the abundances of the observed taxa by classifying the sequences, you can use this tool to investigate the correlations between diversity patterns and environmental parameters to lead you to a better understanding and prediction of ecosystem functioning and the microbial impact on the ecosystem.
- Portablity
  - Has it's own web server CGIScriptor (written in Perl)
    - The servlet starts a true HTTP daemon but is NOT a full fledged server. Moreover, this servlet is definitely NOT intended as a replacement of a real server (e.g., Apache). It's design goal was SIMPLICITY, and not mileage.
    - Gives the ability to restrict the number of concurrent users and limits on file uploads size.
  - Can run back-end scripts in R, C++, python, perl etc.
- Online accounts management
  - Every user gets his own webspace (input and output folder)
  - Administrators can create new users
  - User passwords are encrypted
    - Your personal data is safe against hacking. Encription algorithm: SHA256 on 64 Byte
  - Restrictions:
    - IP Address based, Session based, One-time session token
- Running at http://quince-srv2.eng.gla.ac.uk:8080

Input files format(1)

Taxa	T_02_01	T_02_10	T_02_02	T_02_03	T_02_06	T_02_07
Acidobacteria	0	0	2	2	7	0
Actinobacteria	110	3	240	414	227	1
BRC1	0	0	0	0	0	0
Bacteroidetes	2105	20	1654	1640	1306	5
Chlamydiae	0	0	0	0	0	0
Chloroflexi	2	0	18	34	168	0
Chrysiogenetes	0	0	0	0	0	0
Cyanobacteria	0	0	0	0	0	0
Deinococcus-Thermus	0	0	0	0	3	0
Euryarchaeota	0	0	1	0	0	0
Fibrobacteres	0	0	2	4	0	0
Firmicutes	6741	132	9629	11749	12845	45
Fusobacteria	0	0	0	0	1	0
Gemmatimonadetes	0	0	0	0	0	0
Lentisphaerae	0	0	0	0	1	0
Nitrospira	0	0	0	0	0	0
OD1	0	0	0	0	0	0
OP10	0	0	0	0	0	0
Planctomycetes	0	0	0	0	2	0
Proteobacteria	39	30	52	102	155	1
Spirochaetes	62	0	72	266	43	1
Synergistetes	21	4	84	110	333	0
TM7	0	0	0	0	0	0
Tenericutes	3	0	11	1	3	0
Thermotogae	0	0	0	0	1	0
Unknown	125	11	321	474	595	3
Verrucomicrobia	0	0	0	0	0	0
WS3	0	0	0	0	0	0

Samples	pH	Temp	TS	VS	VFA
T_02_01	7.82	25.1	14.53	71.33	71
T_02_02	6.49	29.6	13.91	64.93	3.7
T_02_03	6.46	27.9	29.45	26.85	27.5
T_02_06	7.69	28.7	65.52	7.03	1.5
T_02_07	7.48	29.8	36.03	34.11	1.1
T_02_09	7.6	25	46.87	19.57	1.1
T_02_10	9.08	24.2	37.76	31.52	2
T_03_02	7.55	28.8	12.65	51.75	30.9
T_03_03	7.68	28.9	14.65	48.08	24.2
T_03_05	7.69	28.7	14.87	41.67	20.1
T_04_04	7.84	26.3	28.85	19.1	5.6
T_04_05	7.95	27.9	46.85	11.47	3.8
T_04_06	7.58	30.1	38.38	13.32	4.2
T_04_07	7.68	28.3	26.54	21.75	6.1

N X Q comma-separated environmental parameters file. Filename must start with ENV_

P X N comma-separated species abundance file. Filename must start with SPE_

Note: look at http://userweb.eng.gla.ac.uk/umer.ijaz/TAXAenv_data.zip to get an idea on how the data should be formatted.

Input files format(2)

TAXAenv supports a general purpose parser for coloring sites on numerous plots. To benefit from this functionality, the site names must all have equal lengths and use underscore "_" as a delimiter to discriminate between different groups. With this format, the colors are then assigned automatically based on the uniqueness of string literals in a particular column before and after the underscore. For example, given the following site names

    T_01_01
    T_01_02
    T_02_01
    T_02_02
    V_01_01
    V_02_02

if you choose the first column, the color indices will be $(1,1,1,1,2,2)$, if you choose the second column, the color indices will be $(1,1,2,2,1,2)$, and if you choose the third column, the color indices will be $(1,2,1,2,1,2)$. If color column is zero then all sites will have the same color, i.e., $(1,1,1,1,1,1)$

Main Webpage

Use http://quince-srv2.eng.gla.ac.uk:8080 to access TAXAenv by clicking on **Click here to access the dashboard**

Login page

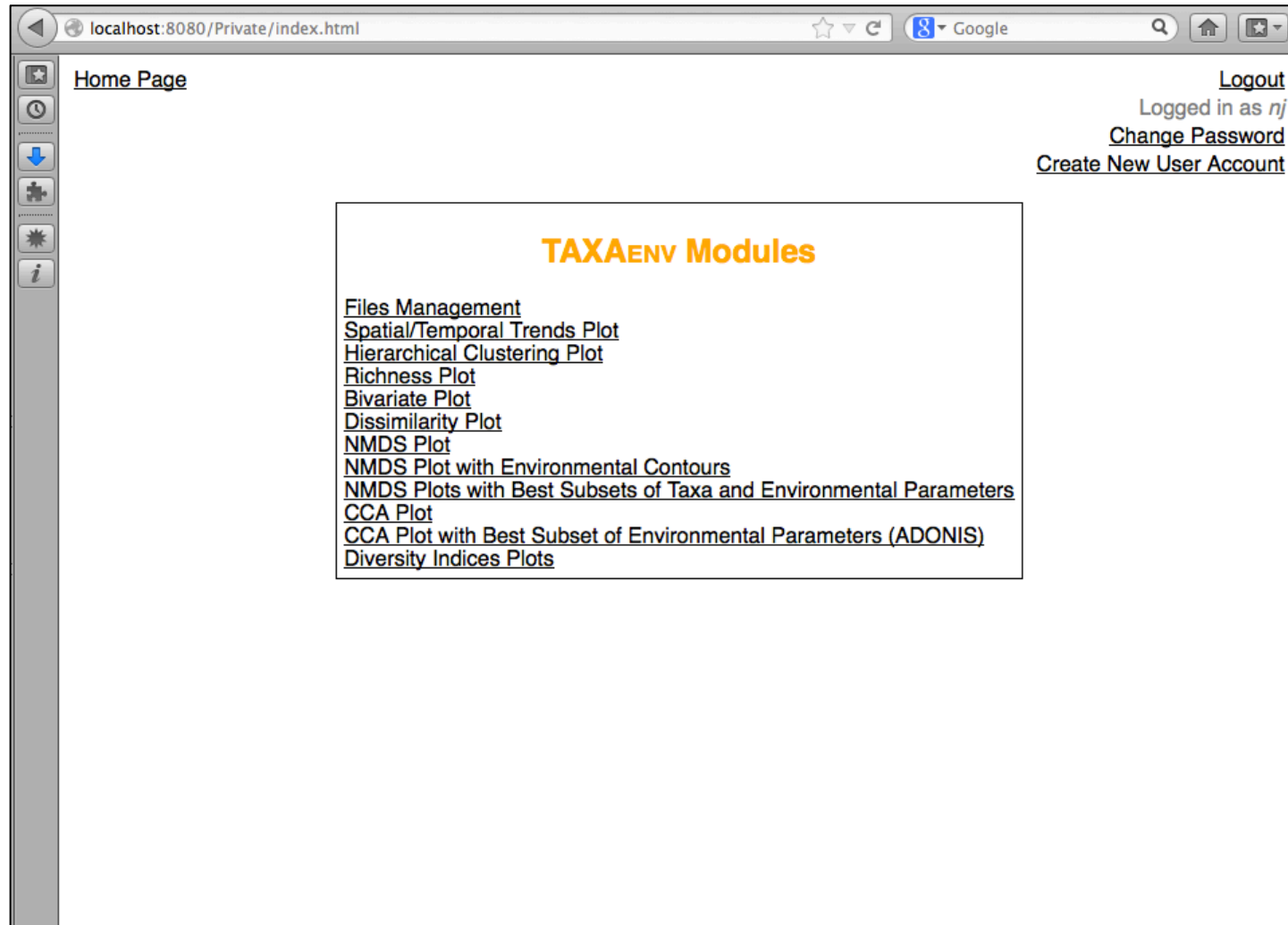Log on to the website using your username and password. You can get a username and password by sending an email to Christopher Quince, cq8u@udcf.gla.ac.uk

TAXAenv Modules page

TAXAenv Modules page shows a listing of different modules in the central pane available to each user to process the data.

Account Maintenance

After logging on for the first time, you can change your password by clicking on "Change Password" link on TAXAenv Modules page. Additionally, if the user has administrative privileges, he can also create more user accounts by clicking on "Create New User Account" link on the TAXAenv Modules page. The administrators can also enforce restrictions on account usage by changing the fields "Allowed IP addresses", "Capabilities", and "Session type".
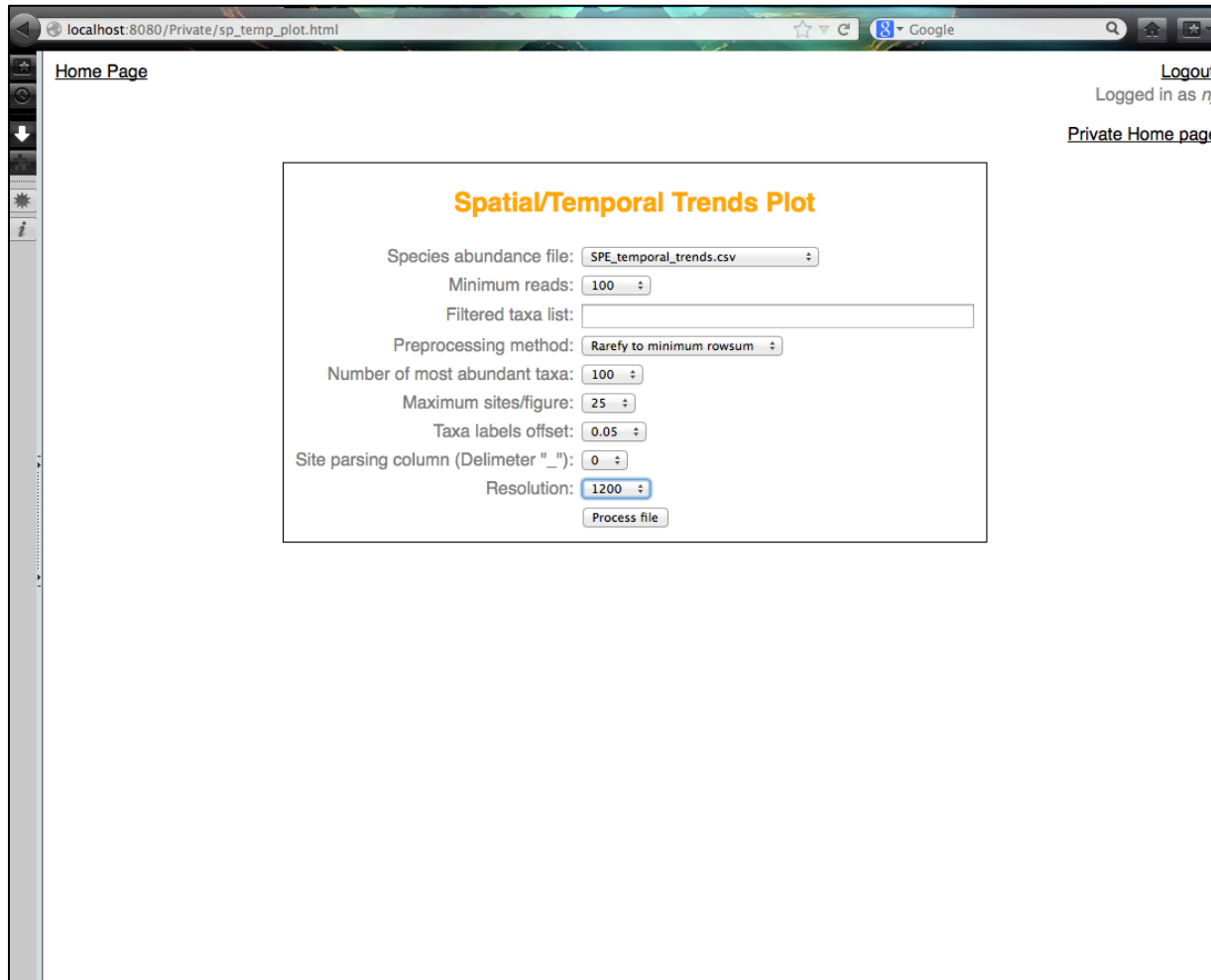
Files Management page

You can upload/delete your files by clicking on "Files Management" link on TAXAenv Modules page. All the uploaded files are shown in a text area highlighted in red. You can either click on "Browse" to upload the file, or delete files by selecting multiple files in the text area followed by clicking on "Process File(s)". Once the files are uploaded in your user space, you can then use them in different modules on TAXAenv Modules page.
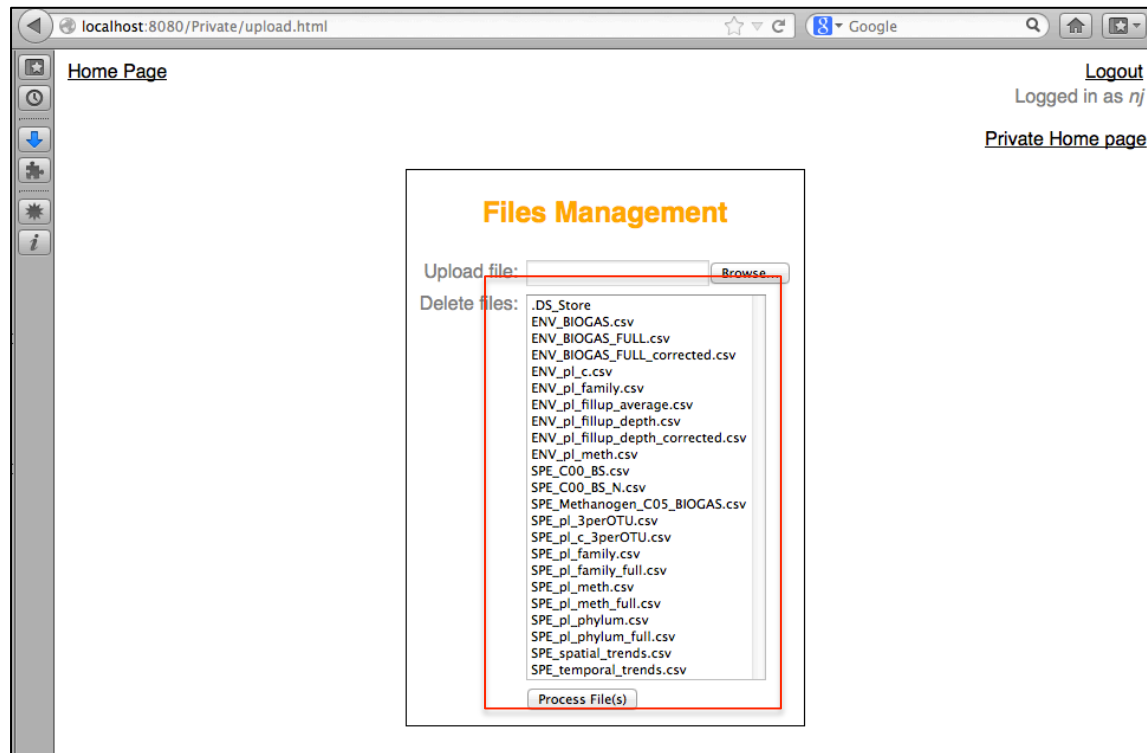
Spatial/Temporal Trends Plot page(1)

You will get the following page after on "Spatial/Temporal Trends Plot" link on TAXAenv Modules page. To understand it's functionality, we will analyze a temporal dataset SPE_temporal_trends.csv by rarefying it to minimum rowsums and displaying 100 most abundant OTUs. The settings are as follows:

Spatial/Temporal Trends Plot page(2)

The webpage should update to display a nifty line plot. Explanation given on next slide.

Spatial/Temporal Trends Plot page(3)

Description: This page plots figures for temporal as well as spatial species abundance data. The species are ranked in order of decreasing frequency occurrence for a given sample with sample names above each column. The lines connect the species between the samples. Line color is assigned based on the ranked frequency occurrence values for species in the first sample to allow better identification of a species across different samples. Line width is also in proportion to the starting frequency occurrence value for each species at each sample step. The legend on the bottom indicates the frequency occurrence values for different line widths.

There are two pre-processing steps provided in this script, you can either obtain relative abundances or rarefy the data to minimum rowsums as provided by the rrarefy() function in vegan's package. Furthermore, you can restrict the figures to display only most abundant taxa

If there exist subgroups within the dataset, you can parse the site names and specify the site parsing column to form the group. After having generated the figures, if you want to draw only specific taxa then provide names in a comma separated manner, for example, C1815,C1161,C1474

Spatial/Temporal Trends Plot page(4)

The blown-up plot is as follows:
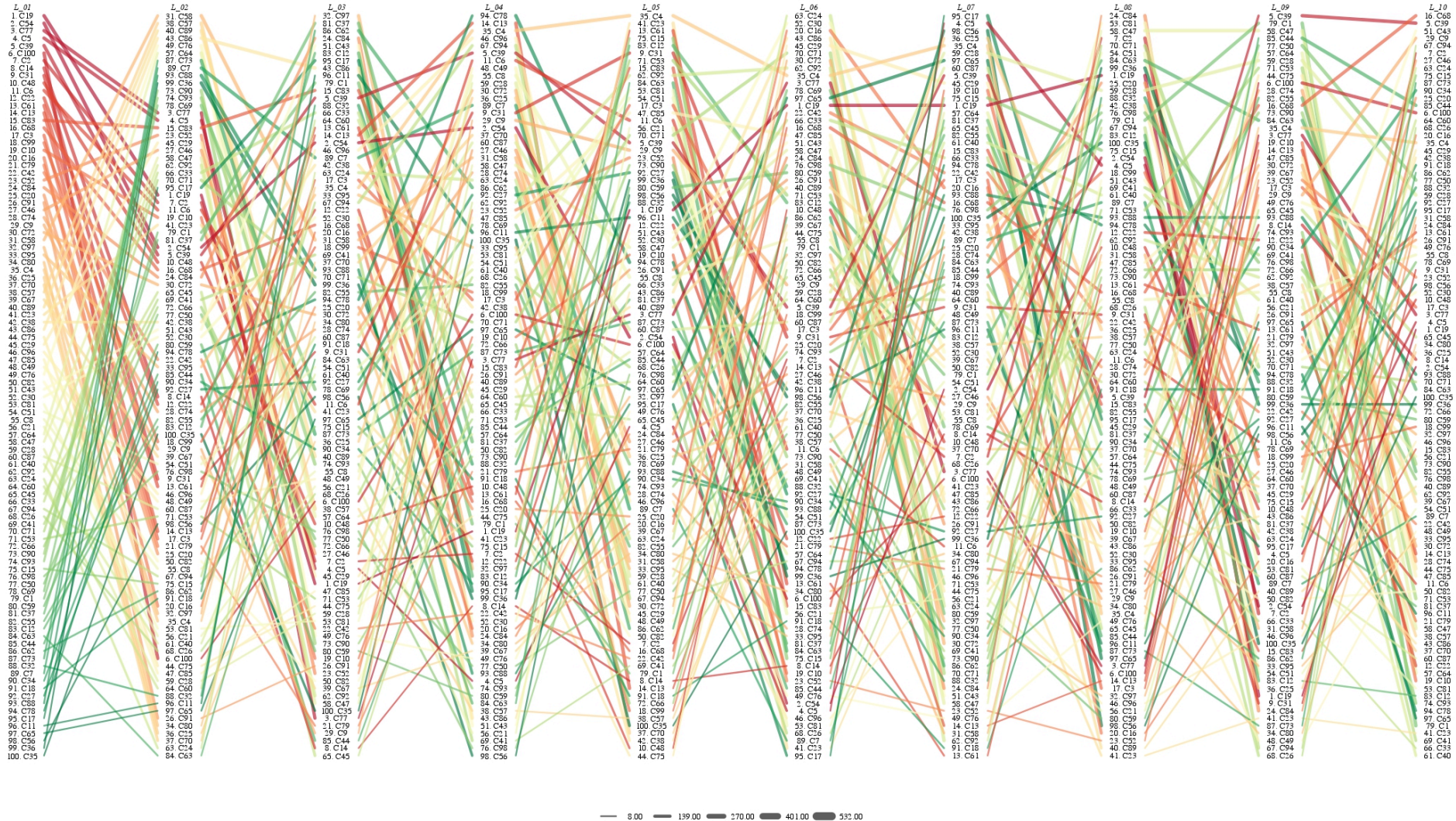
Spatial/Temporal Trends Plot page(5)

The image in the previous slide looks a bit cluttered, so we will only analyze particular OTUs: C19, C78, and C99. We will now put these OTUs in "Filtered taxa list" field.

Spatial/Temporal Trends Plot page(6)

We then get an image given below which only displays the OTUs we are interested in.
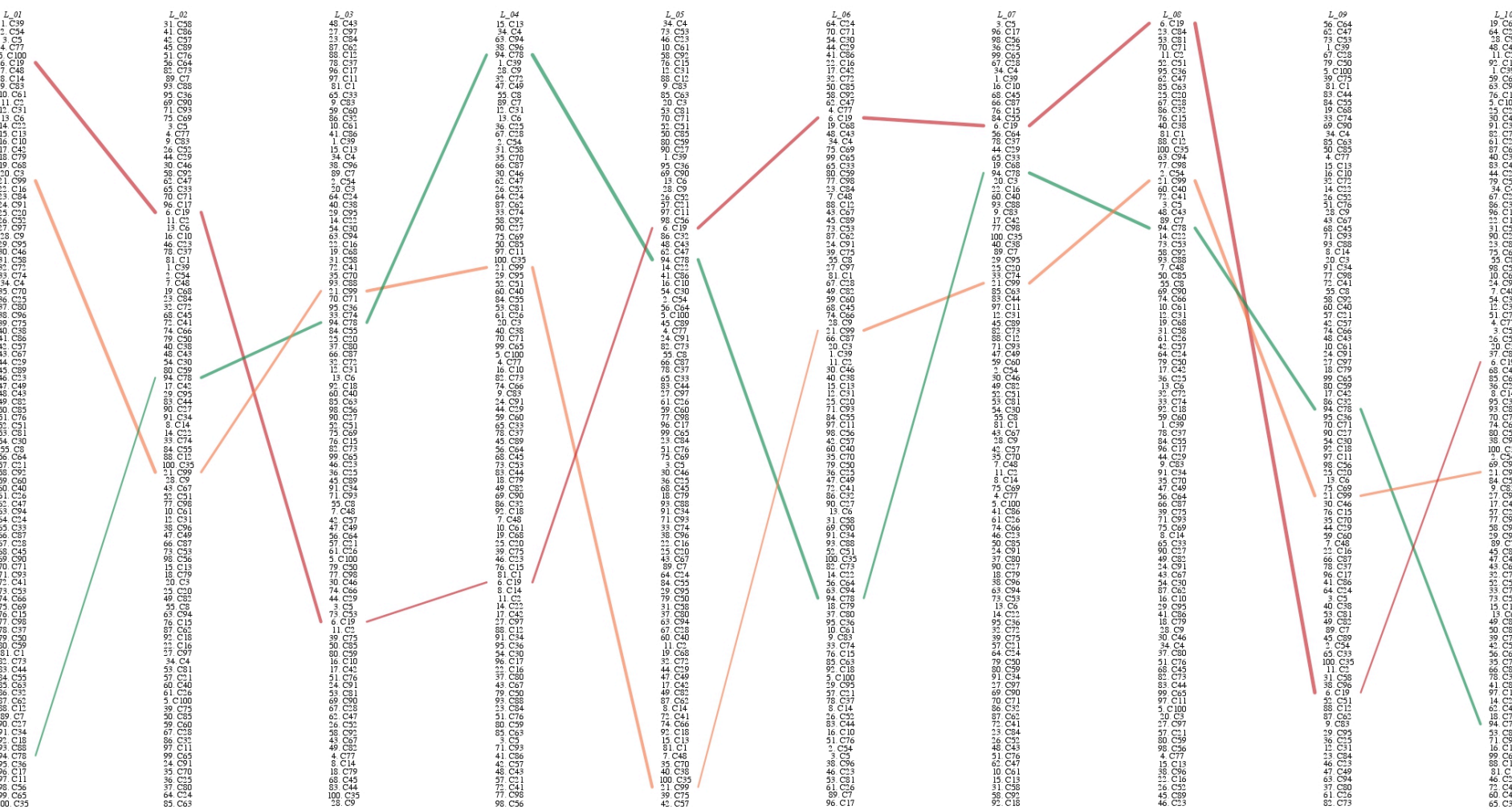
Spatial/Temporal Trends Plot page(7)

We will now analyze a spatial dataset SPE_spatial_trends.csv for pit latrines where we have sampled different latrines at different depths (depth 1 being 20cm and then increment of 20cm onwards starting from the top layer). Here we are interested to know how community profile differs at different depth. Our dataset contains samples with site names comprising underscores, for example, V_03_01, where 03 specifies latrine number, and 01 is first depth. Here we will use site parsing column as 2 to give us trends for each latrine at varying depths. If we choose site parsing column as 3 we will get comparison between different latrines for a given depth. The script is flexible enough to allow any kind of parsing (provided if you follow the strict naming convention) and display plots separately for each group in the dataset. The next few slides show a subset of such plots. You can further restrict to display OTUs as done before.

Spatial/Temporal Trends Plot

Species abundance file: SPE_spatial_trends.csv

Minimum reads: 100

Filtered taxa list:

Preprocessing method: None

Number of most abundant taxa: 100

Maximum sites/figure: 25

Taxa labels offset: 0.05

Site parsing column (Delimeter "_"): 2

Resolution: 1200

Process file

Spatial/Temporal Trends Plot page(8)
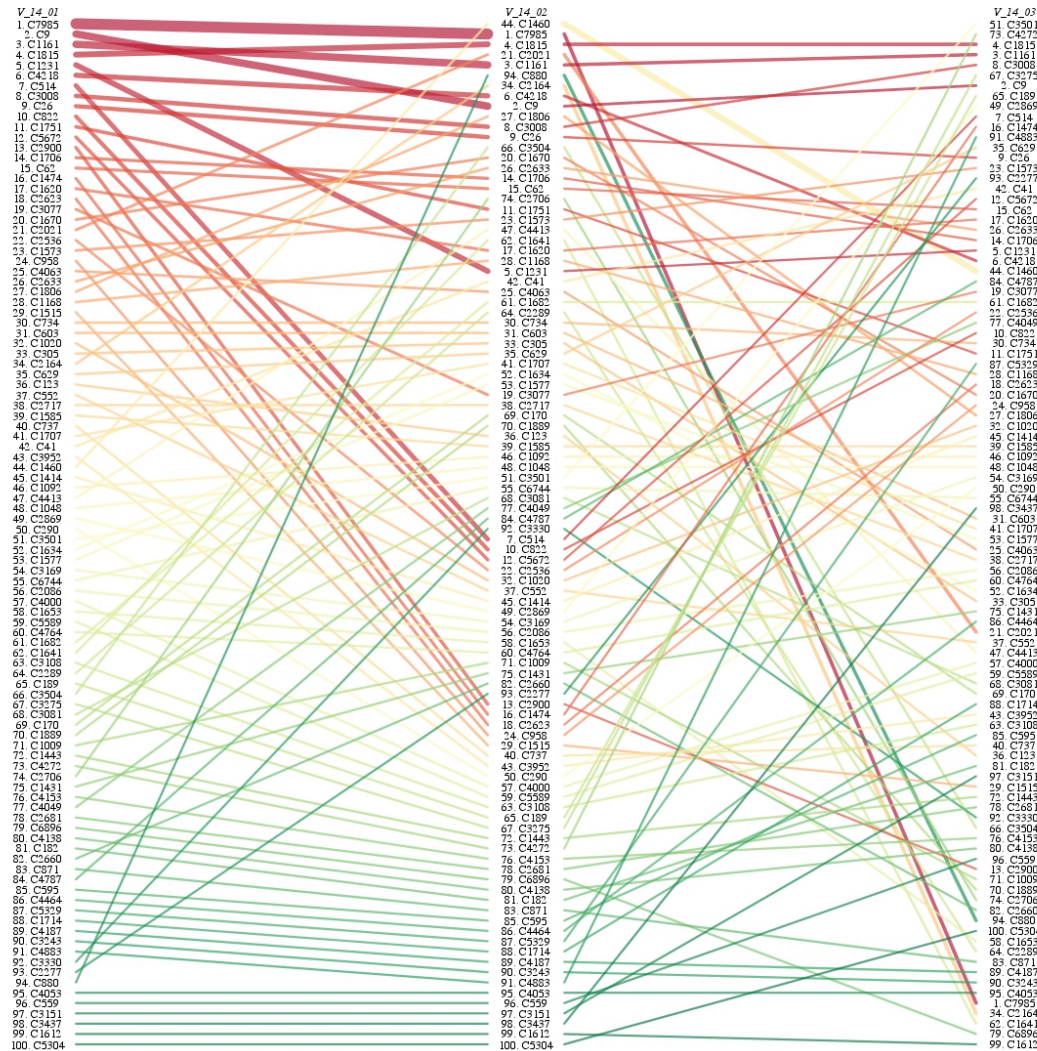
Latrine 14 at different depths

Spatial/Temporal Trends Plot page(9)
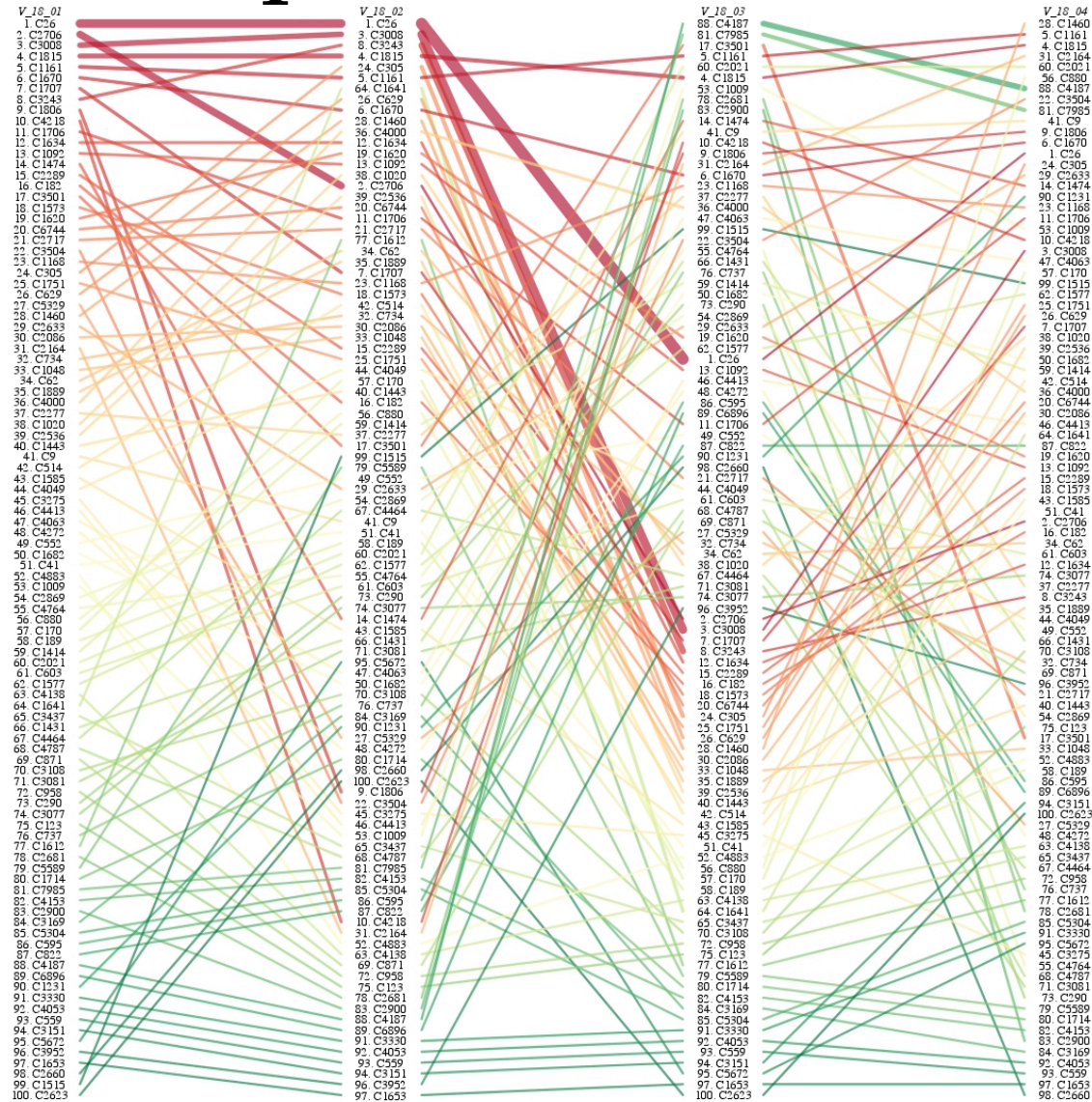
Latrine 18 at different depths

Spatial/Temporal Trends Plot page(10)
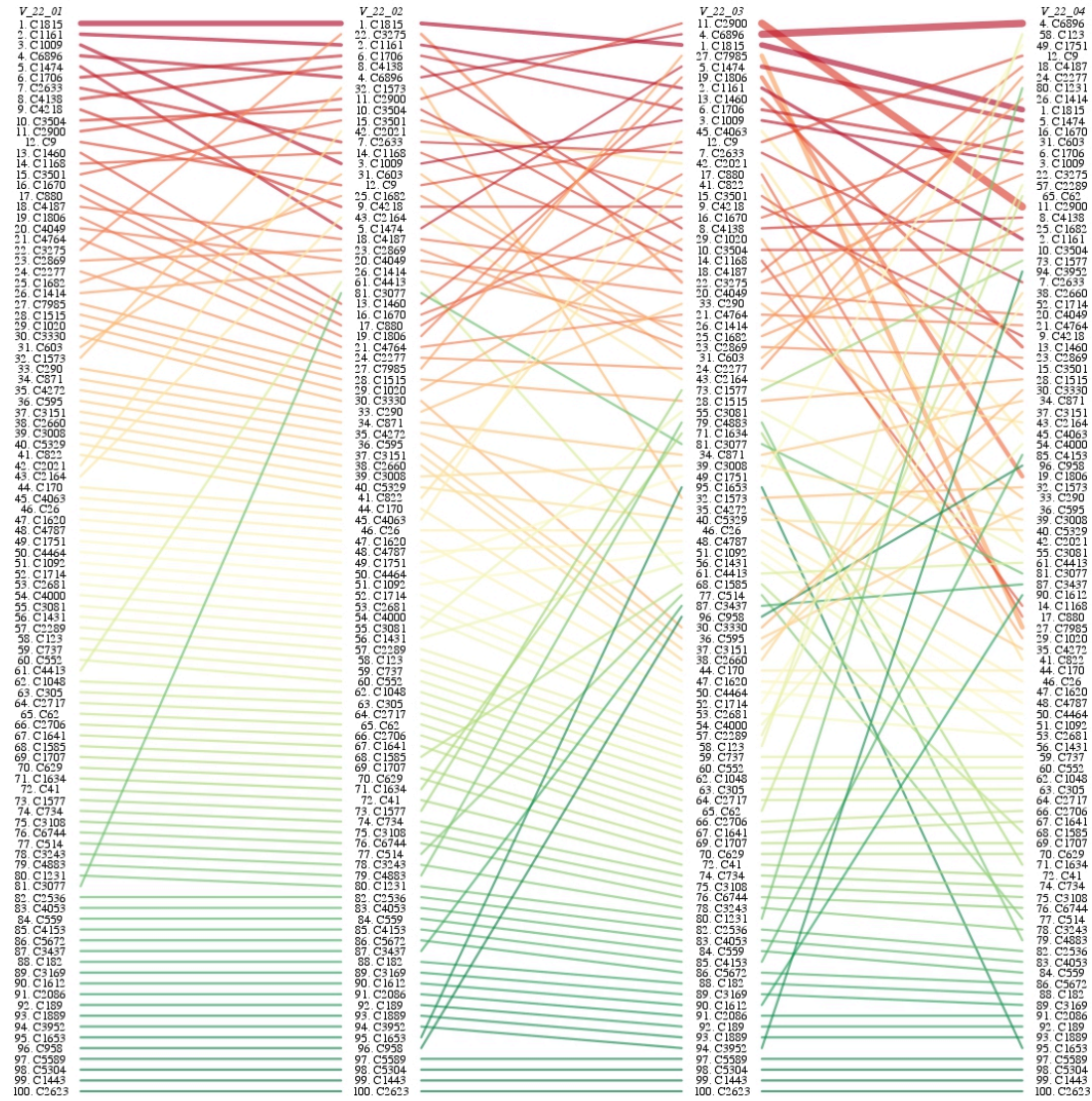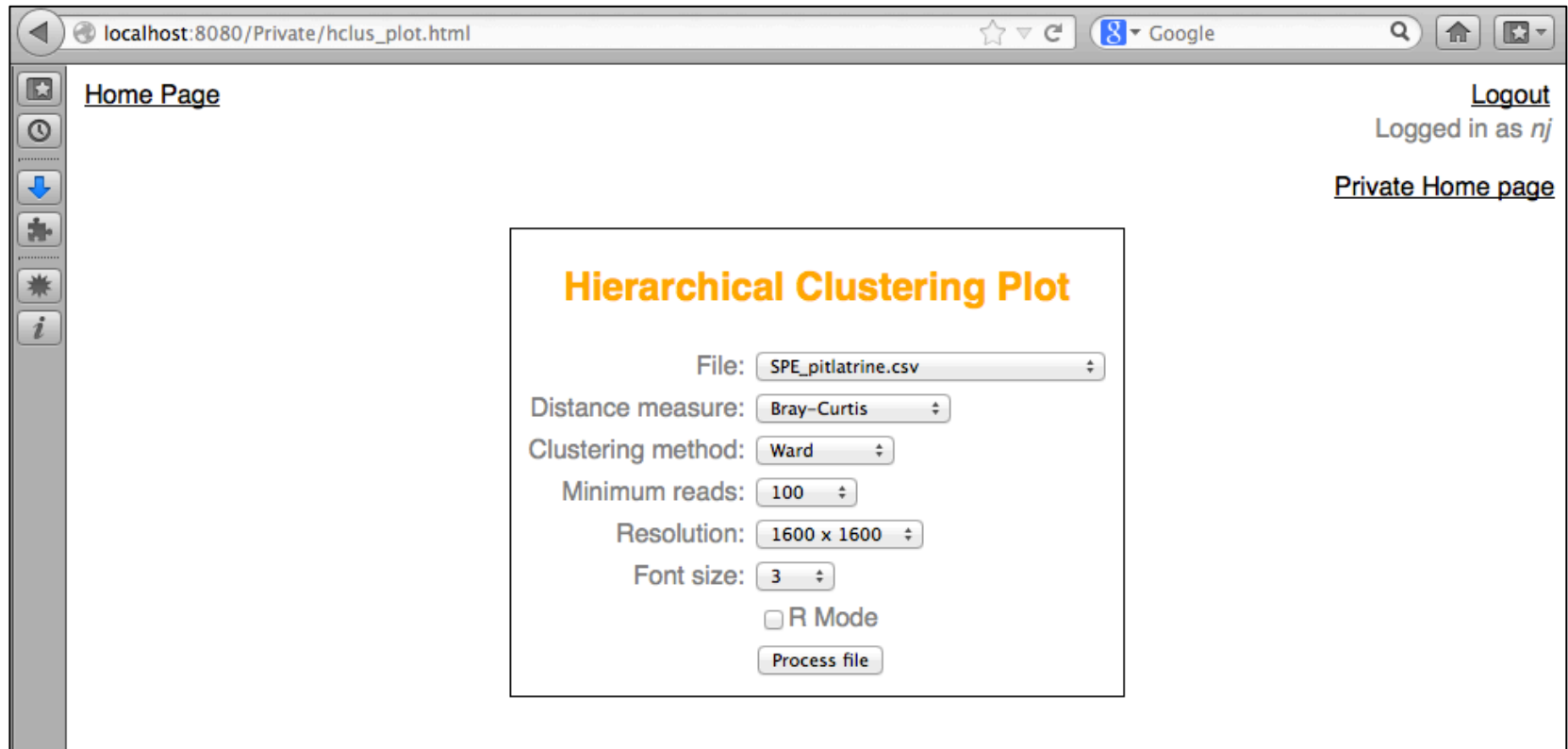

Latrine 22 at different depths

Hierarchical Clustering Plot page(1)

You will get the following page after on "Hierarchical Clustering Plot" link on TAXAenv Modules page. To understand it's functionality, we will analyze a temporal dataset SPE_pitlatrine.csv which contains phylum-level classification of sequences for pit latrines of two different countries, Vietnam and Tanzania at different depths
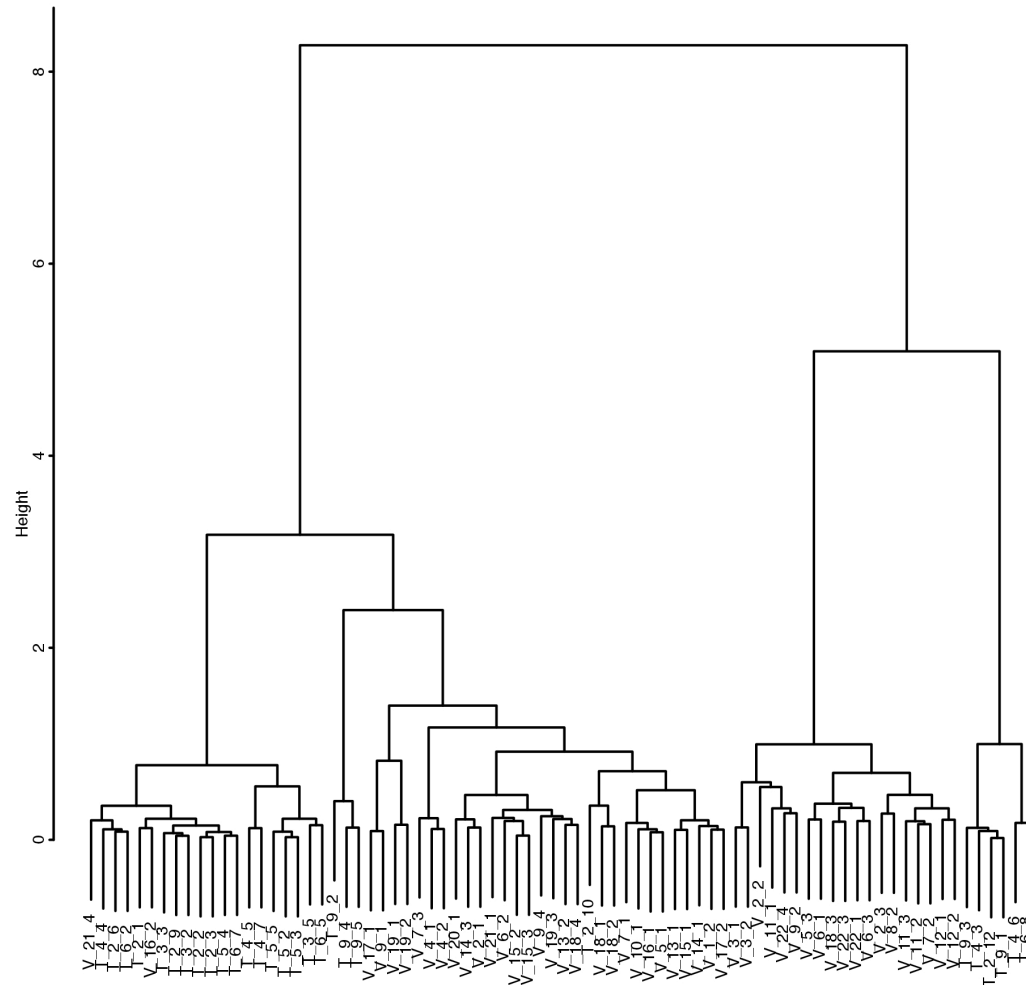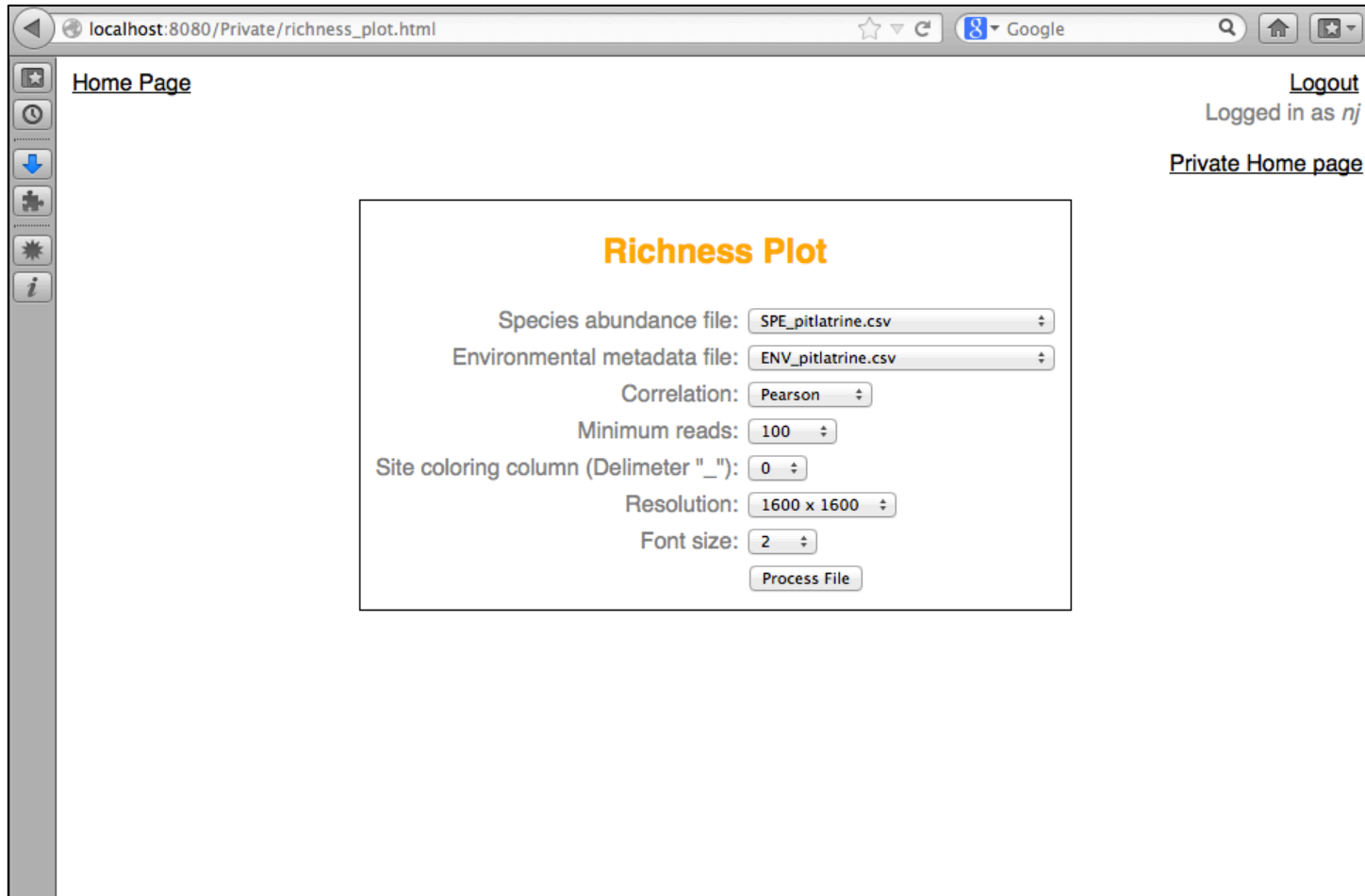
Hierarchical Clustering Plot page(2)

The hierarchical clustering plot is a good way to discriminate between different sites. In the plot below it can be noticed that sites from respective countries are clustered together suggesting a difference in microbial community between the two countries

Richness Plot page(1)

You will get the following page after on "Richness Plot" link on TAXAenv Modules page. We will use the species abundance file SPE_pitlatrine.csv and corresponding environmental file ENV_pitlatrine.csv to understand how richness is affected by different environmental parameters. Description is given on the next slide.

Richness Plot page(2)

Description: This page generates multiple subplots for all the environmental parameters against species richness in a single plot. The species richness is rarefied to the minimum sample numbers and a correlation test is performed between the rarefied richness and the environmental parameters. The resulting correlation and their significance is drawn on top of each subplot. Currently, it has support for three correlation measures:

- -Pearson
- -Spearman
- -Kendall

Furthermore this script also generates summary stats of regression of rarefied richness against environmental parameters. The last column contains the P-values and if significant, it indicates that the richness is affected by this particular environmental parameter.

Richness Plot page(2)

For the two files, the regression results displayed on the webpage are as follows. Here, we find that Prot and Carbo are marginally significant against species richness. The richness plot using site coloring column 1 (i.e. coloring based on countries) is shown on the next slide and is a good way to tell how richness differs between samples from both countries.

```
Module output
Regression of Rarefied Species Richness against Environmental Metadata

Residuals:
    Min      1Q   Median      3Q      Max
-6.8864  -0.8491  -0.0014   1.2181   4.1812

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.1631221  3.1742456   5.092 3.17e-06 ***
pH          -0.1777796  0.2428363  -0.732  0.4667
Temp        -0.0795578  0.0825052  -0.964  0.3384
TS          -0.0146587  0.0253673  -0.578  0.5653
VS           0.0263730  0.0155514   1.696  0.0946 .
VFA         -0.0030353  0.0228113  -0.133  0.8945
CODt         0.0008845  0.0023057   0.384  0.7025
CODs        -0.0131843  0.0127104  -1.037  0.3034
perCODsbyt  -0.0442296  0.0600223  -0.737  0.4638
NH4          0.0918897  0.0701204   1.310  0.1946
Prot        -0.0683240  0.0328382  -2.081  0.0414 *
Carbo       -0.0737567  0.0281154  -2.623  0.0108 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.104 on 66 degrees of freedom
Multiple R-squared: 0.5875,     Adjusted R-squared: 0.5187
F-statistic: 8.545 on 11 and 66 DF,  p-value: 3.811e-09
```
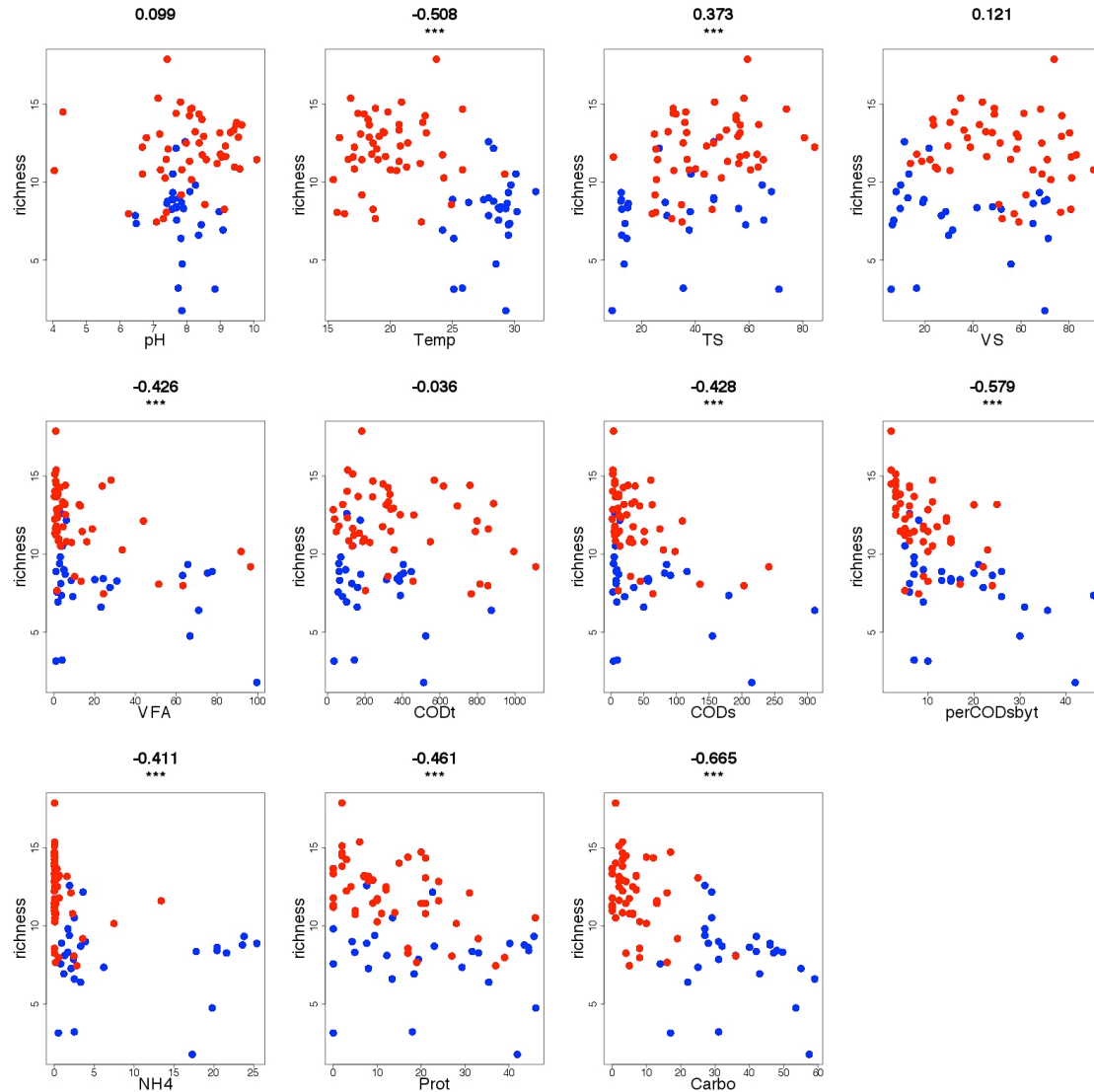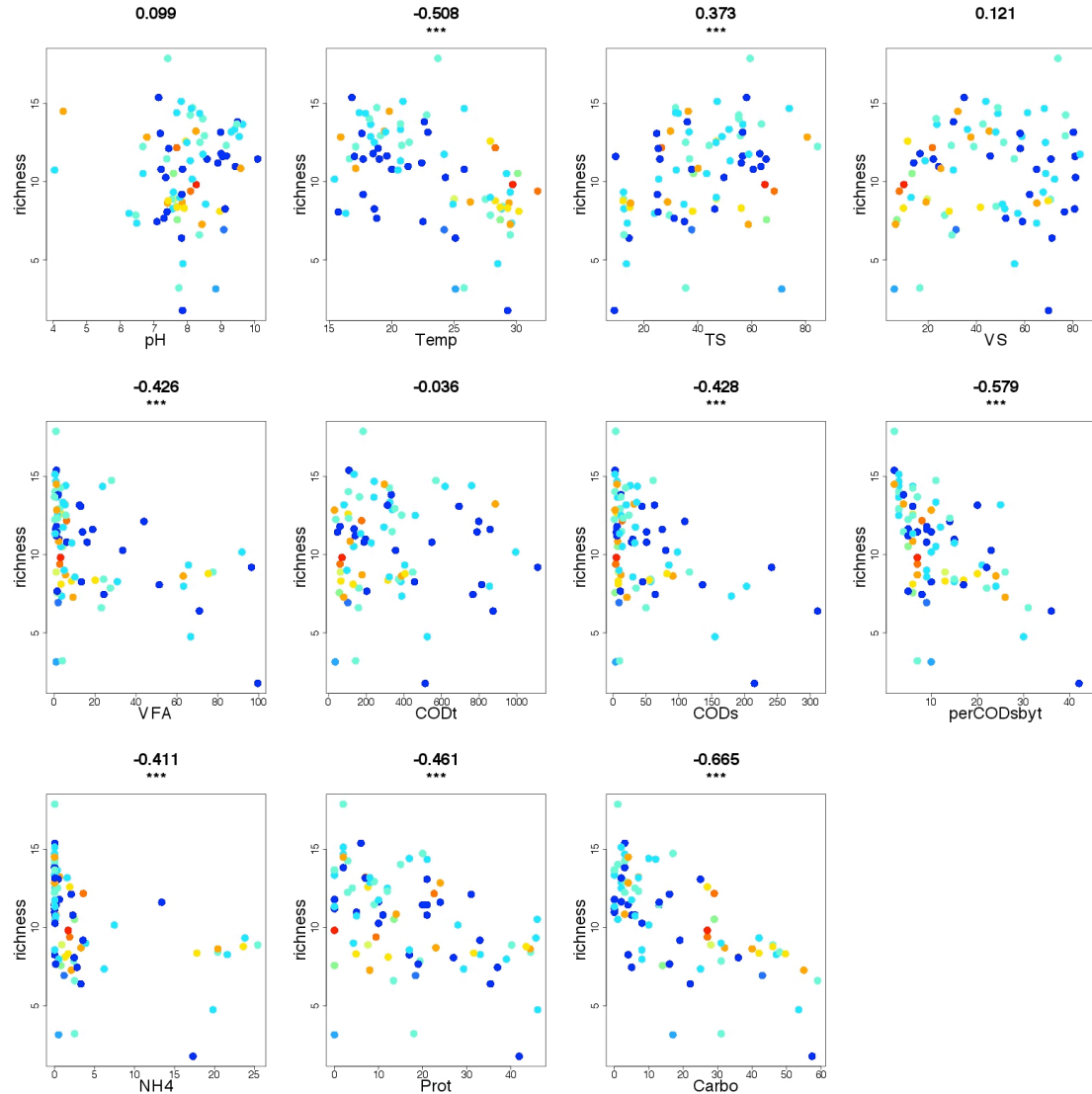
Richness Plot page(3)

Here, we find that Temp, perCODsbyt, NH4, Prot, and Carbo are significantly negatively-correlated, and TS is significantly positively-correlated with richness.

Richness Plot page(4)

Using site coloring column as 3, we can color different latrines based on depth and show that latrines at similar depth behave more or less same.

Bivariate Plot page(1)

You will get the following page after on "Bivariate Plot" link on TAXAenv Modules page. We will use the environmental file ENV_pitlatrine.csv to understand the relationship between different environmental parameters. This module generates bivariate plots with histograms on the diagonals, scatter plots with smooth curves below the diagonals and correlations with significance levels above diagonals. Moreover, the variables are reordered in the plots with any two consecutive variables on the diagonal being most similar.
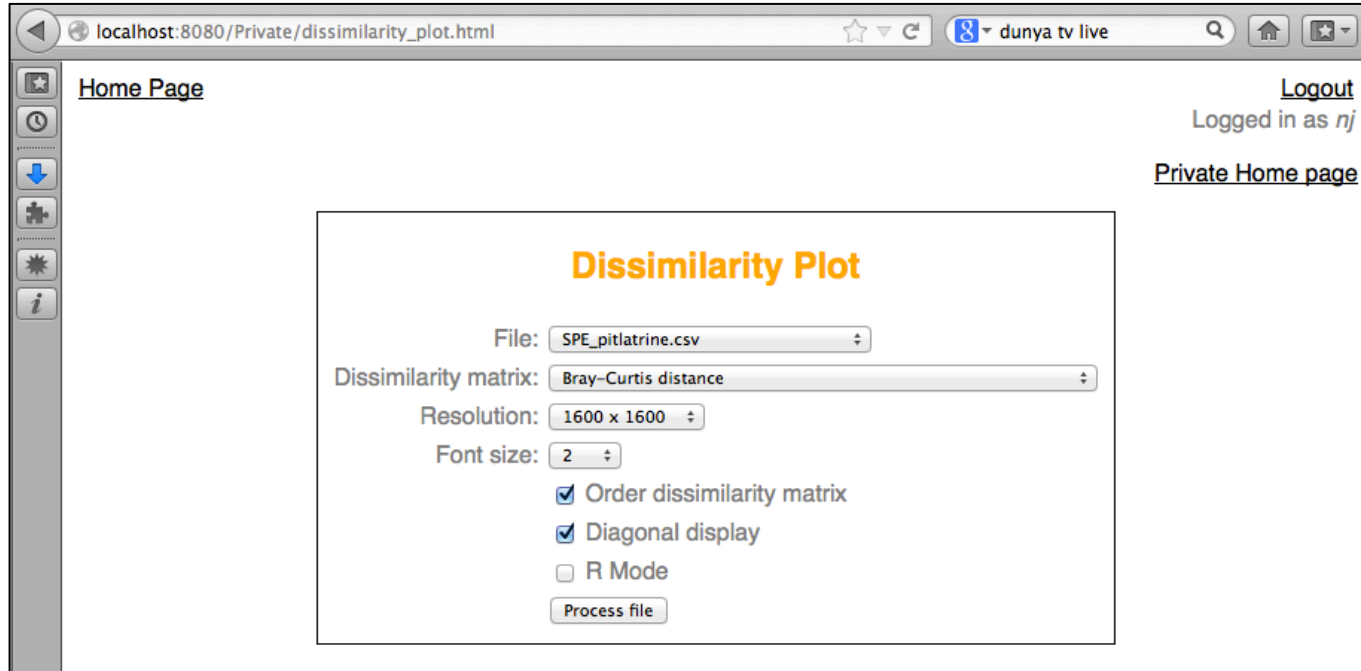
Bivariate Plot page(2)

Here, the positive correlations are shown in red and negative correlations are shown in blue.
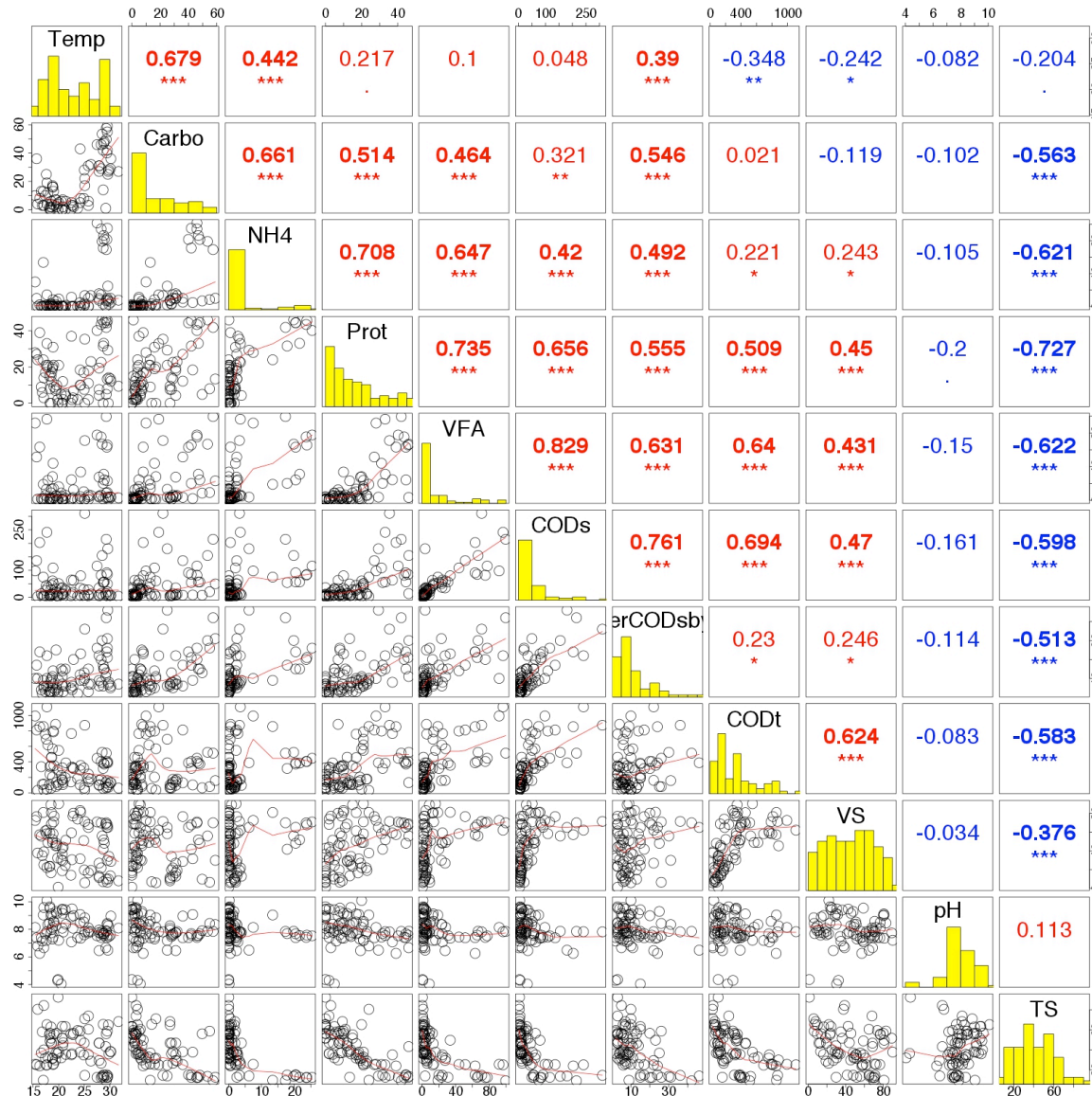
Dissimilarity Plot page(1)

You will get the following page after on "Dissimilarity Plot" link on TAXAenv Modules page. This module generates a plot of dissimilarity measures between samples. Magenta is high similarity, and cyan is high dissimilarity. In the current version of the program, you can use the following dissimilarity measures:

- Bray-Curtis dissimilarity matrix on raw species data
- Bray-Curtis dissimilarity matrix on log-transformed abundances
- Chord distance matrix
- Hellinger distance matrix
- Chi-square pre-transformation followed by Euclidean distance

Dissimilarity Plot page(2)

We use SPE_pitlatrine.csv and notice magenta blocks that cluster similar sites together. You can use this as an alternative to the hierarchical clustering plot to show similarities between different samples.
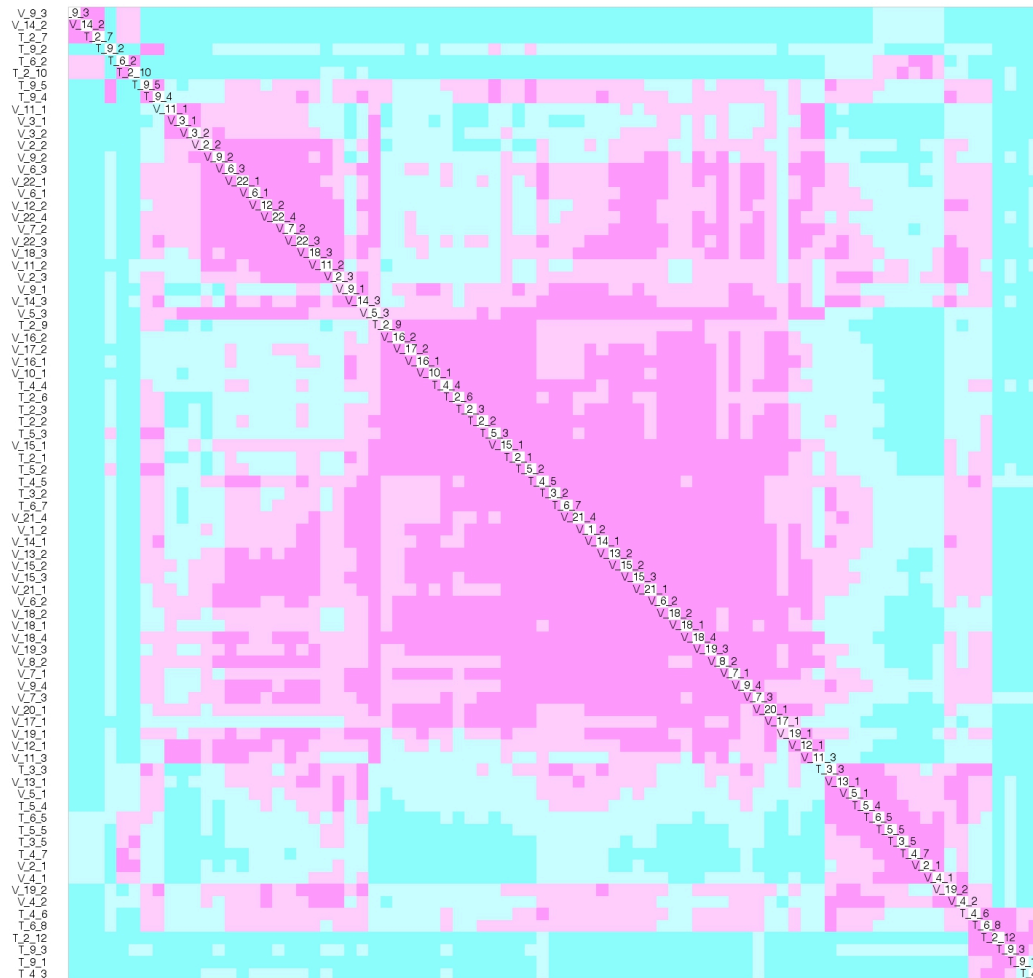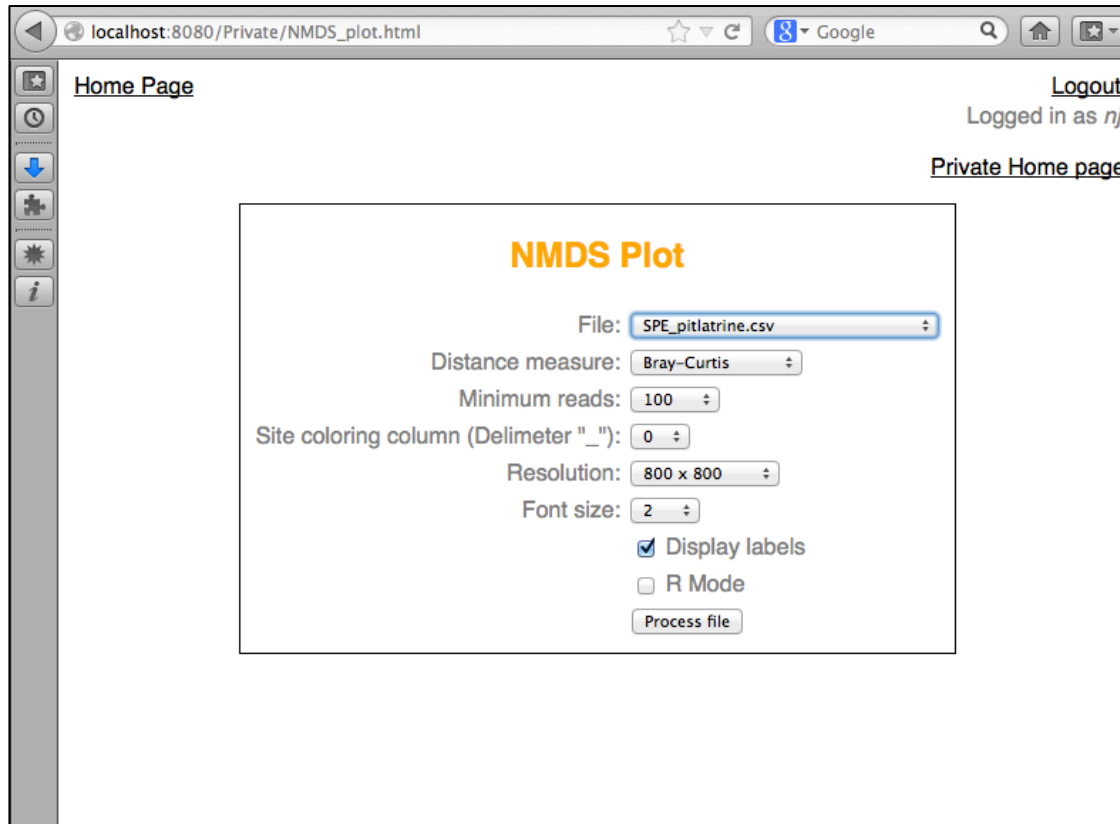
NMDS Plot page(1)

You will get the following page after clicking on "NMDS Plot" link on TAXAenv Modules page. This module finds a non-parameteric monotonic relationship between the dissimilarities in the samples matrix, and the location of each site in a low-dimensional space (similar to priniciple component analysis).

NMDS Plot page(2)

Using SPE_pitlatrine.csv, it can be seen that there is a clear difference between the latrines of the two countries.
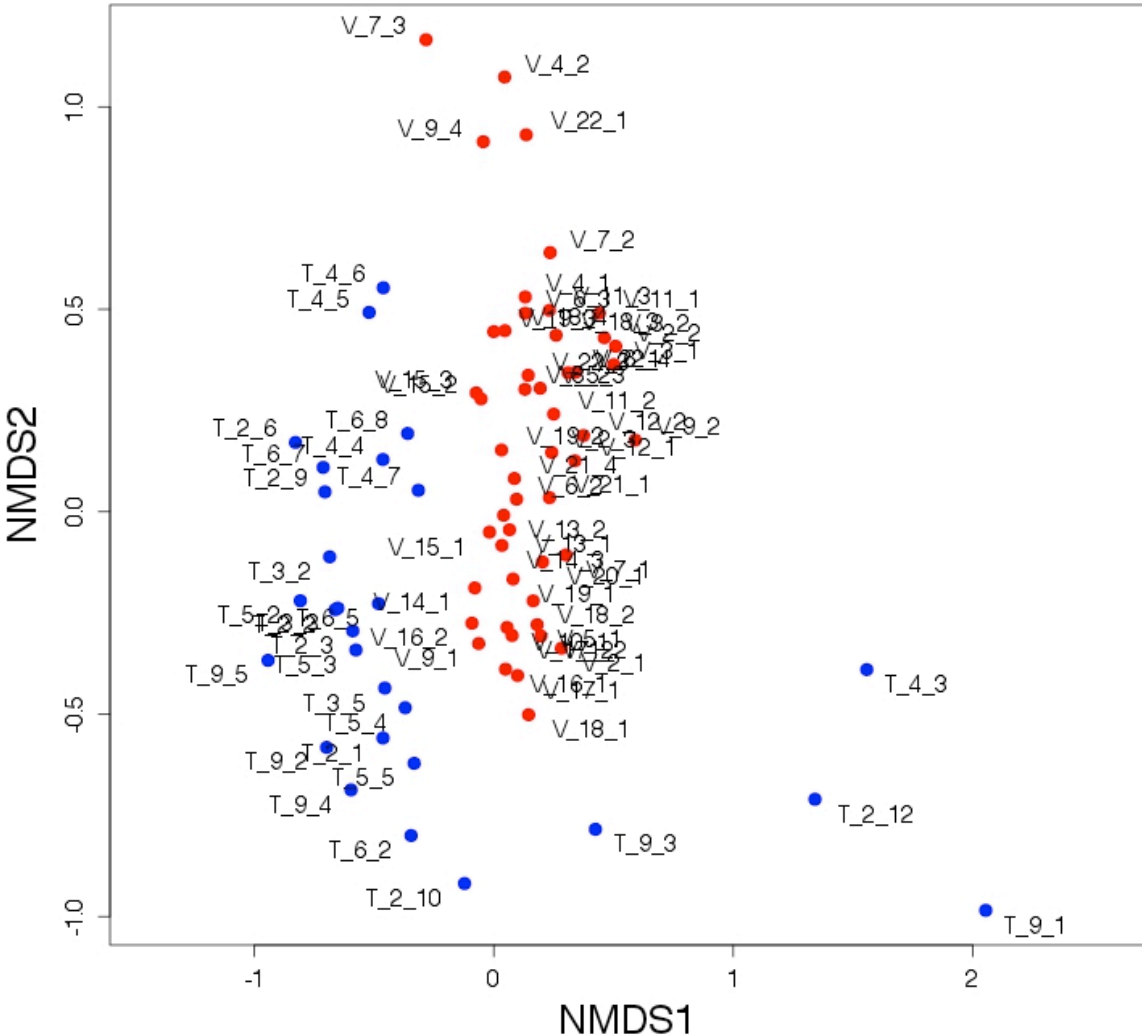
NMDS Plot with Environmental Contours page(1)

You will get the following page after clicking on "NMDS Plot with Environmental Contours" link on TAXAenv Modules page. This module is similar to the previous one with the added advantage that after the ordination in reduced space, the environmental parameters are superimposed on the NMDS plot as contours.

NMDS Plot with Environmental Contours page(2)

We will use both SPE_pitlatrine.csv and ENV_pitlatrine.csv. The resulting plot comprises several plots each with one environmental parameter.
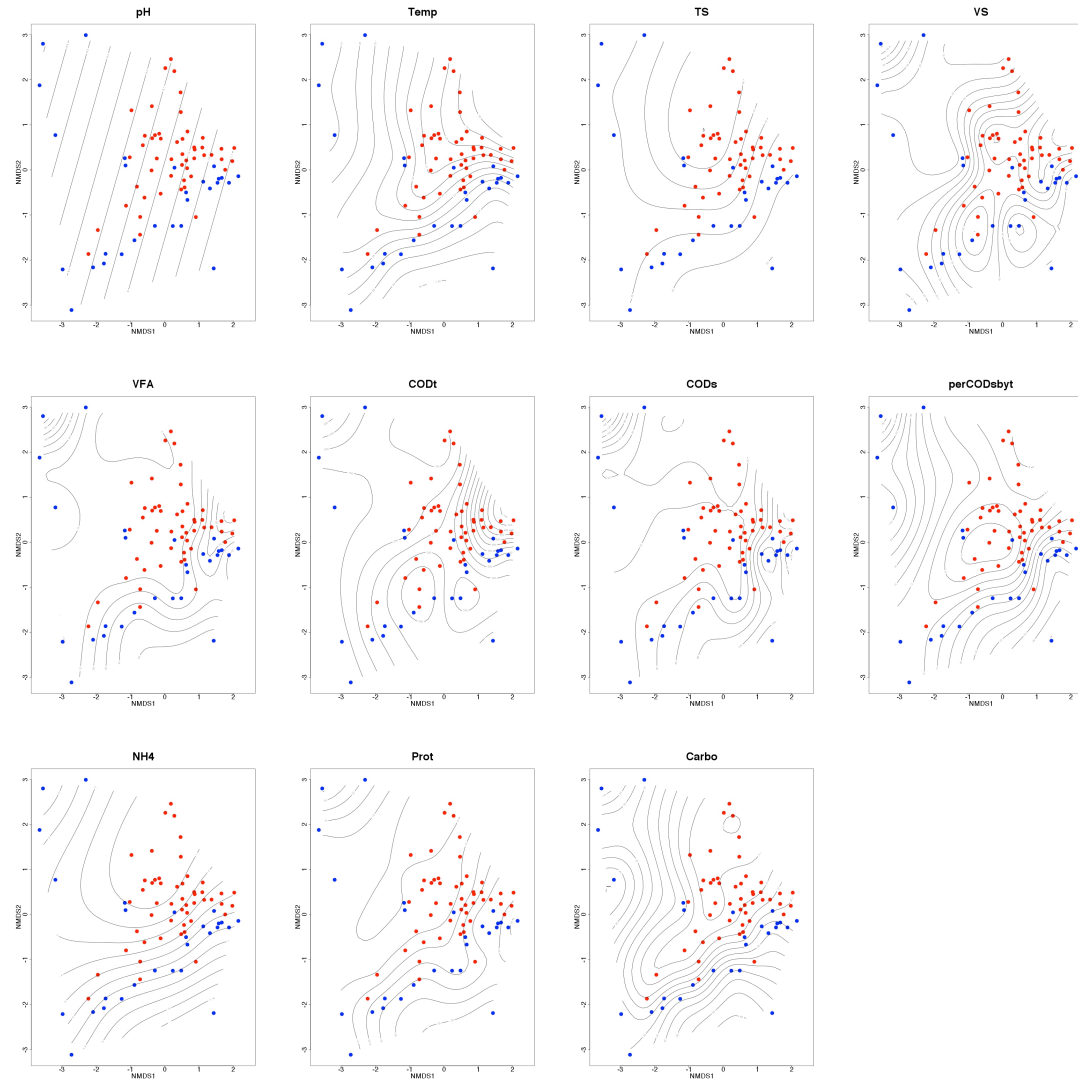
NMDS Plot with Best Subset of Environmental Parameters and Taxa page(1)

You will get the following page after clicking on "NMDS Plot with Best Subset of Environmental Parameters and Taxa" link on TAXAenv Modules page. This module uses an R function that is an extension of vegan library's bioenv() function and finds the best set of environmental variables with maximum (rank) correlation with community dissimilarities and plots them on NMDS plot. It also finds the best subset of species and plots them on NMDS plot.

We will use both SPE_pitlatrine.csv and ENV_pitlatrine.csv. The page will update to display several tables and plots, which are as follows with NMDS plots shown on the next slide:

Module output

Top 10 Best Subset of Environmental Parameters with Correlation Values
1   Temp + TS + Carbo = 0.394156679885845
2   Temp + Carbo = 0.356424652574571
3   Temp + TS + NH4 + Carbo = 0.351970562422596
4   Temp + TS + perCODsbyt + NH4 + Carbo = 0.32969796587422
5   Temp + NH4 + Carbo = 0.328995579028878
6   Temp = 0.324671392130821
7   Temp + TS + perCODsbyt + Carbo = 0.320445435951933
8   pH + Temp + Carbo = 0.319801065792379
9   Temp + perCODsbyt + Carbo = 0.314526012037347
10 Temp + TS + CODs + Carbo = 0.312033523000079

Best Subset of Environmental Parameters Fit Against NMDS

|	R^2	P-value
Temp	0.259929842788921	0.001
TS	0.222233631866863	0.001
Carbo	0.504956482517677	0.001

Top 10 Best Subset of Taxa with Correlation Values
1   Acidobacteria_Gp4 + Actinobacteria + Bacteroidia + Betaproteobacteria + Dehalococcoidetes + Erysipelotrichi + Fibrobacteria + Flavobacteria + Methanobacteria + Opitutae + Spirochaetes + Subdivision3 + Synergistia + Thermomicrobia = 0.900467626635941
2   Acidobacteria_Gp4 + Actinobacteria + Bacteroidia + Betaproteobacteria + Dehalococcoidetes + Erysipelotrichi + Fibrobacteria + Flavobacteria + Opitutae + Spirochaetes + Subdivision3 + Synergistia + Thermomicrobia = 0.899827628479153
3   Acidobacteria_Gp4 + Actinobacteria + Bacteroidia + Betaproteobacteria + Dehalococcoidetes + Erysipelotrichi + Fibrobacteria + Flavobacteria + Opitutae + Spirochaetes + Synergistia + Thermomicrobia = 0.896251263017577
4   Acidobacteria_Gp4 + Actinobacteria + Bacteroidia + Betaproteobacteria + Dehalococcoidetes + Erysipelotrichi + Fibrobacteria + Flavobacteria + Opitutae + Synergistia + Thermomicrobia = 0.89259829602631
5   Acidobacteria_Gp4 + Bacteroidia + Caldilineae + Clostridia + Cyanobacteria + Deltaproteobacteria + Gammaproteobacteria + Methanomicrobia + Opitutae + Subdivision3 + Synergistia = 0.891656474433918
6   Bacteroidia + Caldilineae + Clostridia + Cyanobacteria + Deltaproteobacteria + Gammaproteobacteria + Methanomicrobia + Opitutae + Subdivision3 + Synergistia = 0.891147392926784
7   Acidobacteria_Gp4 + Actinobacteria + Bacteroidia + Betaproteobacteria + Clostridia + Deltaproteobacteria + Fibrobacteria + Flavobacteria + Methanobacteria + Planctomycetacia + Unknown = 0.888221195527347
8   Acidobacteria_Gp4 + Actinobacteria + Bacteroidia + Betaproteobacteria + Clostridia + Deltaproteobacteria + Fibrobacteria + Flavobacteria + Planctomycetacia + Unknown = 0.887577408747372
9   Bacteroidia + Caldilineae + Clostridia + Deltaproteobacteria + Gammaproteobacteria + Methanomicrobia + Opitutae + Subdivision3 + Synergistia = 0.886699185952649
10 Actinobacteria + Bacteroidia + Betaproteobacteria + Dehalococcoidetes + Erysipelotrichi + Fibrobacteria + Flavobacteria + Opitutae + Synergistia + Thermomicrobia = 0.886597658827303
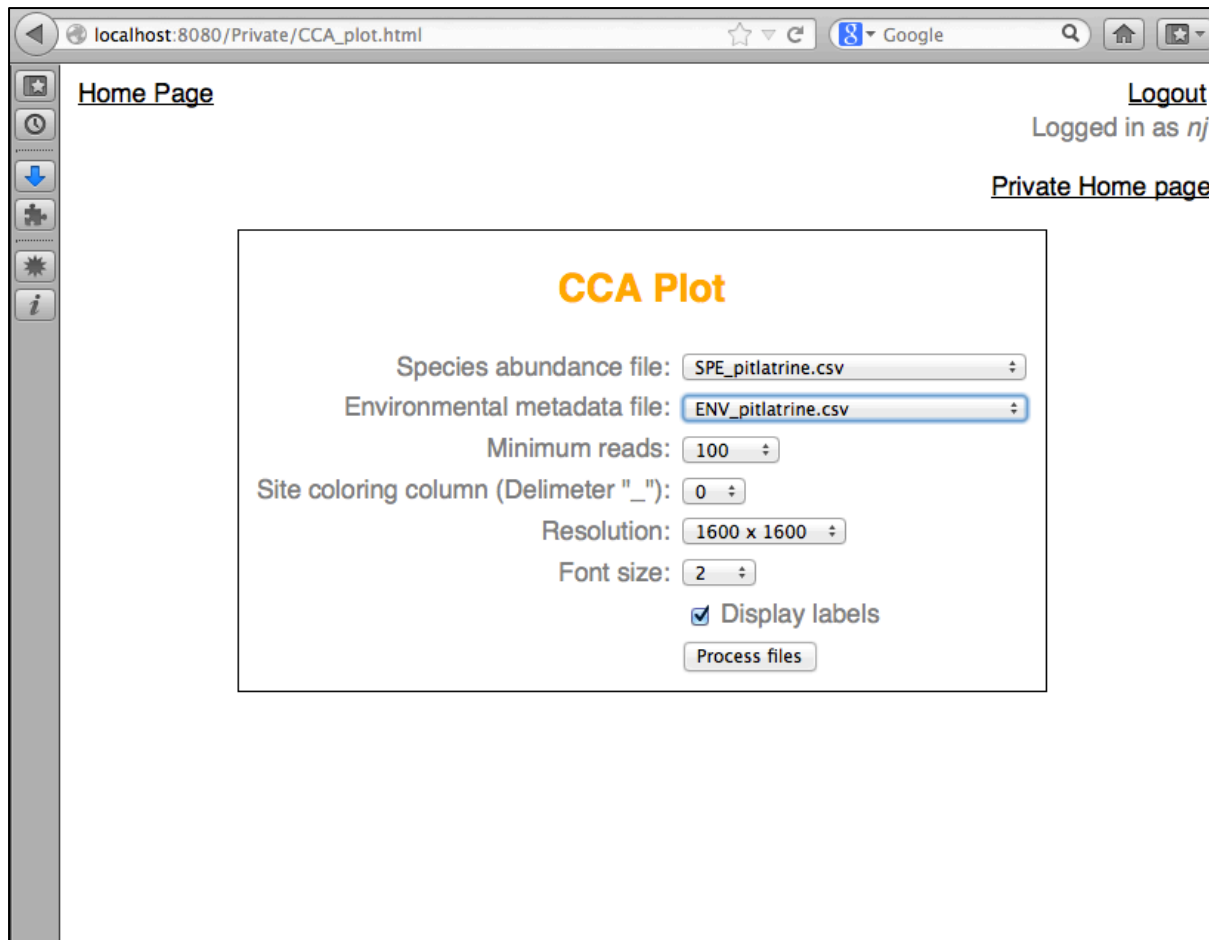
Best Subset of Taxa Fit Against NMDS

	R^2	P-value
Acidobacteria_Gp4	0.244904750612867	0.001
Actinobacteria	0.278945859804392	0.001
Bacteroidia	0.0538488139410606	0.117
Betaproteobacteria	0.175244759886277	0.011
Dehalococcoidetes	0.139898625412935	0.015
Erysipelotrichi	0.145733046096269	0.004
Fibrobacteria	0.137175638281946	0.015
Flavobacteria	0.169251470910829	0.004
Methanobacteria	0.0262177201671167	0.276
Opitutae	0.200206888925272	0.002
Spirochaetes	0.05274507890993	0.122
Subdivision3	0.0311428475642173	0.21
Synergistia	0.209617516603673	0.003
Thermomicrobia	0.2519813176910721	0.001

Here we are plotting the top subset for environmental parameters and species from the tables in previous slide.
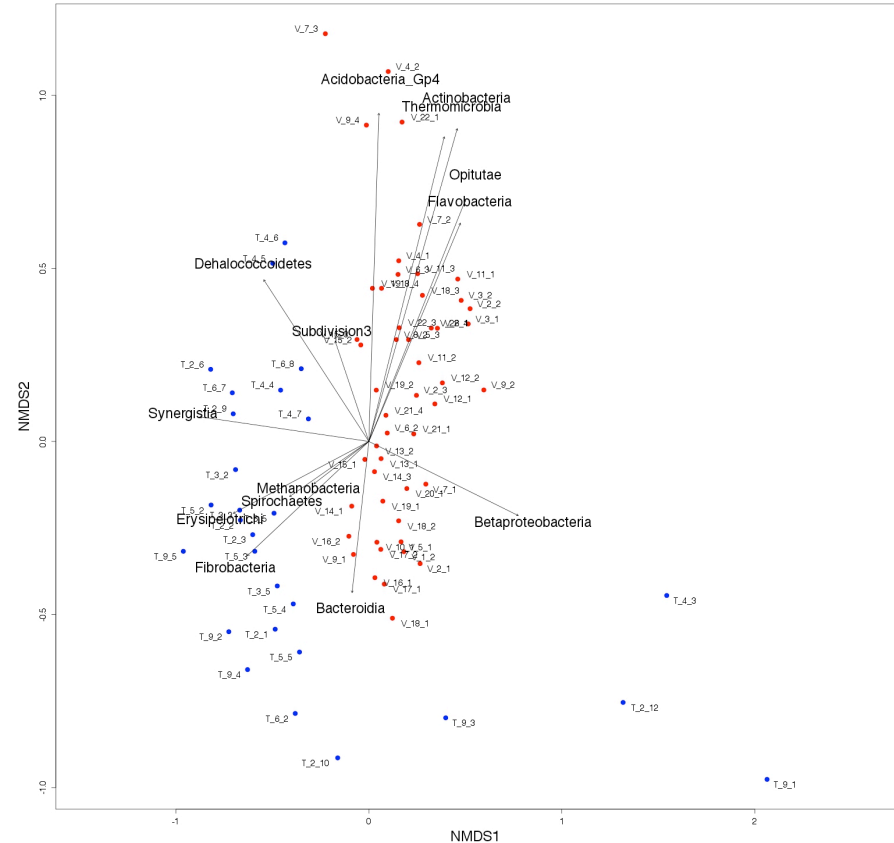
CCA Plot page(1)

You will get the following page after clicking on "CCA Plot" link on TAXAenv Modules page. This module performs canonical correspondence analysis to find the relationship between species and their environment. The method extracts environmental gradients and then use them for describing and visualizing the preference of taxa/sample on an ordination diagram.
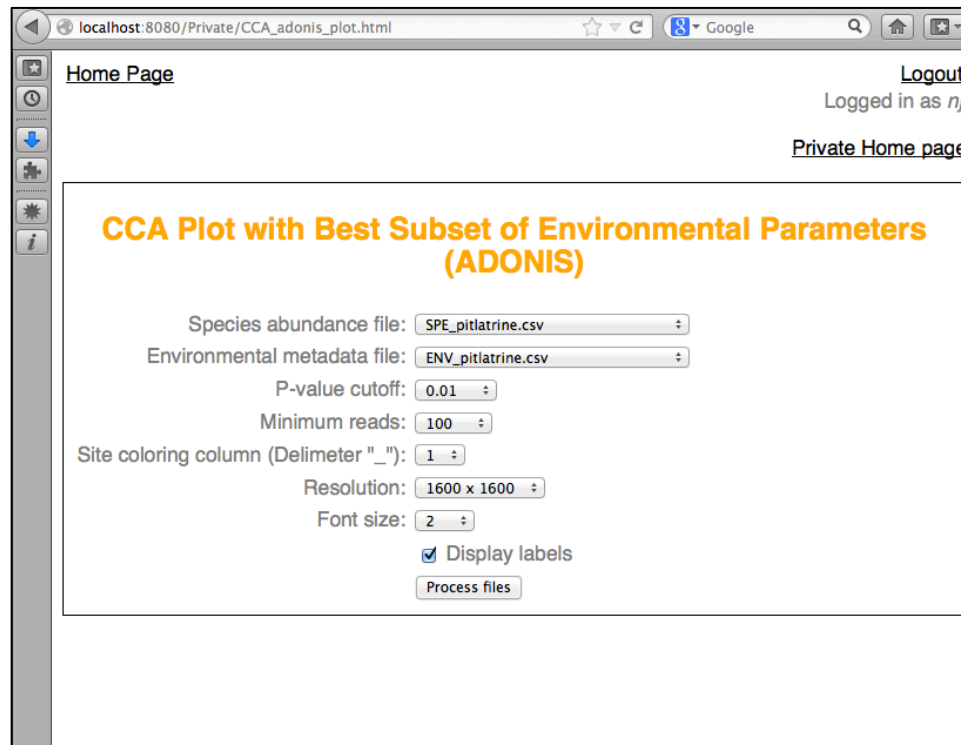
CCA Plot page(2)

The following plot is generated when you use both SPE_pitlatrine.csv and ENV_pitlatrine.csv.
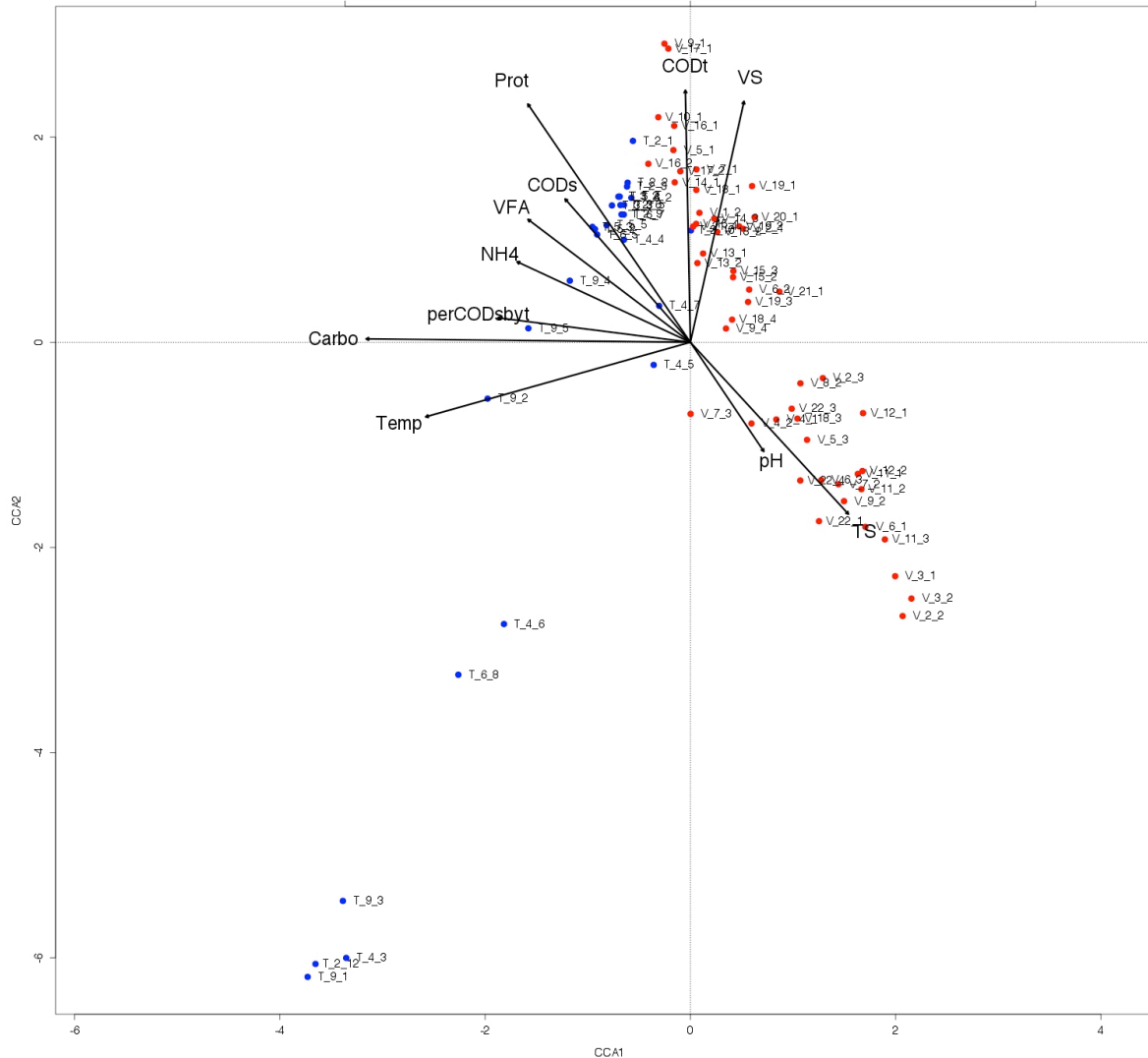
CCA Plot with Best Subset of Environmental Parameters (ADONIS) page(1)

You will get the following page after clicking on "CCA Plot with Best Subset of Environmental Parameters (ADONIS)" link on TAXAenv Modules page. This module uses analysis of variance using distance matrices to find the best set of environmental variables that describe the community structure. We have used adonis() function from the vegan library which fits linear models to distance matrices and uses a permutation test with Pseudo F-ratios. It draws a CCA plot with only those environmental variables that are below a cut off P-value threshold.

The following table and plot is generated when you use both SPE_pitlatrine.csv and ENV_pitlatrine.csv.

Module output
Permutational Multivariate Analysis of Variance Using Distance Matrices (ADONIS)

```
Terms added sequentially (first to last)

          Df SumsOfSqs MeanSqs F.Model     R2  Pr(>F)
pH         1    0.3880 0.38799  3.1507 0.03012   0.019 *
Temp       1    1.5567 1.55674 12.6418 0.12086   0.001 ***
TS         1    0.7739 0.77386  6.2843 0.06008   0.002 **
VS         1    0.2955 0.29547  2.3995 0.02294   0.050 *
VFA        1    0.2412 0.24118  1.9585 0.01872   0.103
CODt       1    0.1707 0.17068  1.3860 0.01325   0.212
CODs       1    0.0332 0.03316  0.2693 0.00257   0.918
perCODsbyt 1    0.1910 0.19101  1.5512 0.01483   0.179
NH4        1    0.1977 0.19772  1.6056 0.01535   0.180
Prot       1    0.3654 0.36545  2.9677 0.02837   0.029 *
Carbo      1    0.5398 0.53984  4.3839 0.04191   0.006 **
Residuals 66    8.1274 0.12314         0.63098
Total     77   12.8805                 1.00000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
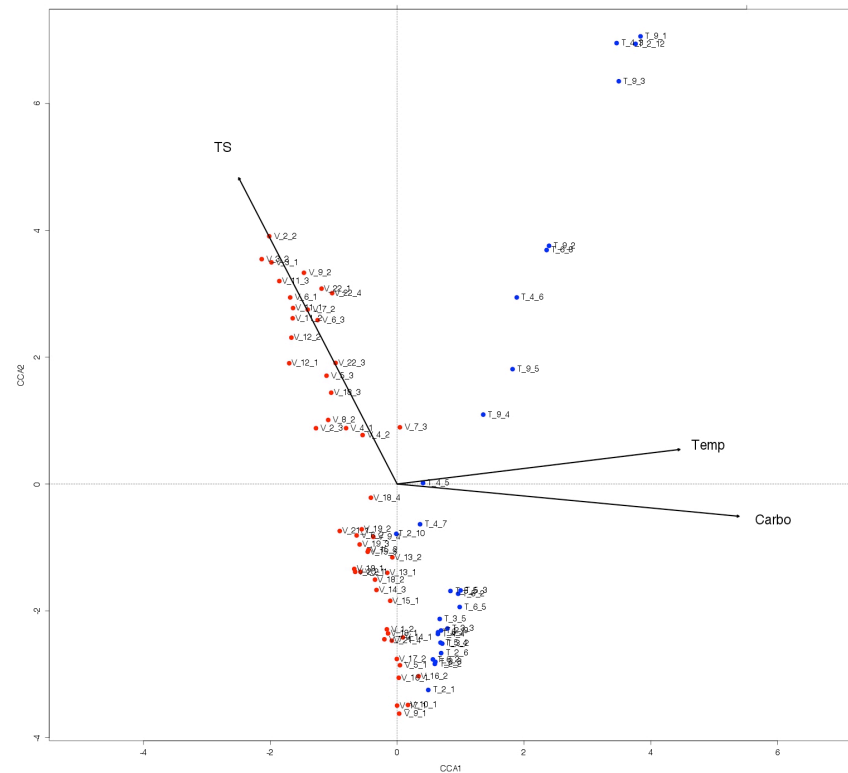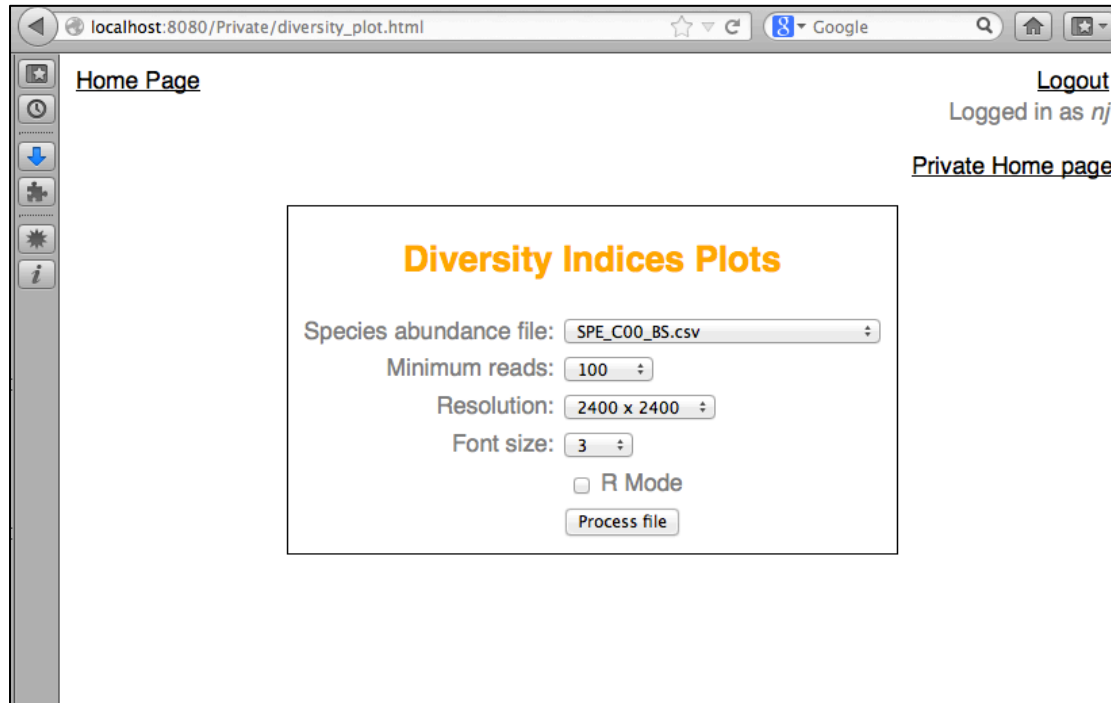
Diversity Indices Plots page(1)

You will get the following page after clicking on "Diversity Indices Plots" link on TAXAenv Modules page. This script generates the ecological diversity indices and rarefaction species richness. The following indices are supported:

- Shanon Index
- Simpson Index
- Inverse Simpson's Index
- Fisher's logarithmic series' alpha parameters
- Pielou's evenness

Diversity Indices Plots page(2)

The following plots are generated when you use SPE_pitlatrine.csv