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. Encryption 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)

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

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

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

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

![Change Password](image1.png)

![Create New User Account](image2.png)
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.
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:
The webpage should update to display a nifty line plot. Explanation given on next slide.
**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:
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.
We then get an image given below which only displays the OTUs we are interested in.
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.
Latrine 14 at different depths
Latrine 18 at different depths
Latrine 22 at different depths
You will get the following page after on “Hierarchical Clustering Plot” link on TAXAenv Modules page. To understand its 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.
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.
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.
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.
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**

<table>
<thead>
<tr>
<th>Residuals:</th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-6.8864</td>
<td>-0.8491</td>
<td>-0.0014</td>
<td>1.2181</td>
<td>4.1812</td>
</tr>
</tbody>
</table>

| 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
Here, we find that Temp, perCODsbyt, NH4, Prot, and Carbo are significantly negatively-correlated, and TS is significantly positively-correlated with richness.
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.
You will get the following page after clicking 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.
Here, the positive correlations are shown in red and negative correlations are shown in blue.
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
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.
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 principal component analysis).
Using SPE_pitlatrine.csv, it can be seen that there is a clear difference between the latrines of the two countries.
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.
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:
Here we are plotting the top subset for environmental parameters and species from the tables in previous slide.
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.
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.
CCA Plot with Best Subset of Environmental Parameters (ADONIS) page(2)

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)

<table>
<thead>
<tr>
<th>Terms added sequentially (first to last)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Df</td>
</tr>
<tr>
<td>----</td>
</tr>
<tr>
<td>pH</td>
</tr>
<tr>
<td>Temp</td>
</tr>
<tr>
<td>VS</td>
</tr>
<tr>
<td>VFA</td>
</tr>
<tr>
<td>CODt</td>
</tr>
<tr>
<td>ODe</td>
</tr>
<tr>
<td>perCODbyt</td>
</tr>
<tr>
<td>NH4</td>
</tr>
<tr>
<td>Prot</td>
</tr>
<tr>
<td>Carbo</td>
</tr>
<tr>
<td>Residuals 66</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

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
The following plots are generated when you use SPE_pitlatrine.csv.