Galaxy PhyloChip analysis pipeline PhyloProfiler

User's Manual

Version 2
Last revision 18-11-2015
Author Marcel Kempenaar
Email: brs@nbic.nl

1 Users Manual

General

The PhyloChip galaxy analysis pipeline PhyloProfiler is developed by Marcel Kempenaar of the BRS group of the Netherlands Bioinformatics Centre (NBIC) in collaboration with Menno van der Voort, Rodrigo Mendes and Jos Raaijmakers of the Laboratory of Phytopathology of the Wageningen University/NIOO-KNAW.

1.1 Disclaimer

Data and/or results obtained from the PhyloProfiler are made available in the hope that they will be useful, but WITHOUT ANY WARRANTY, and without any implied warranty of merchantability or fitness for a particular purpose. The copyright holders and/or other parties provide the data and/or results "AS ARE", without warranty of any kind, either expressed or implied. The entire risk as to the quality and performance of the data and/or results is yours. Should any data and/or results prove defective, you assume the cost of all necessary servicing, repair or correction for your own account.
The server can be used freely under the conditions listed below. NBIC withholds itself the right, at all times, to deny the access to the service. No part of this site may be reproduced, unless for own publication subject to written prior permission of the Netherlands Bioinformatics Centre (Geert Grooteplein 28, 6525 GA Nijmegen, The Netherlands) with obligatory source indication.

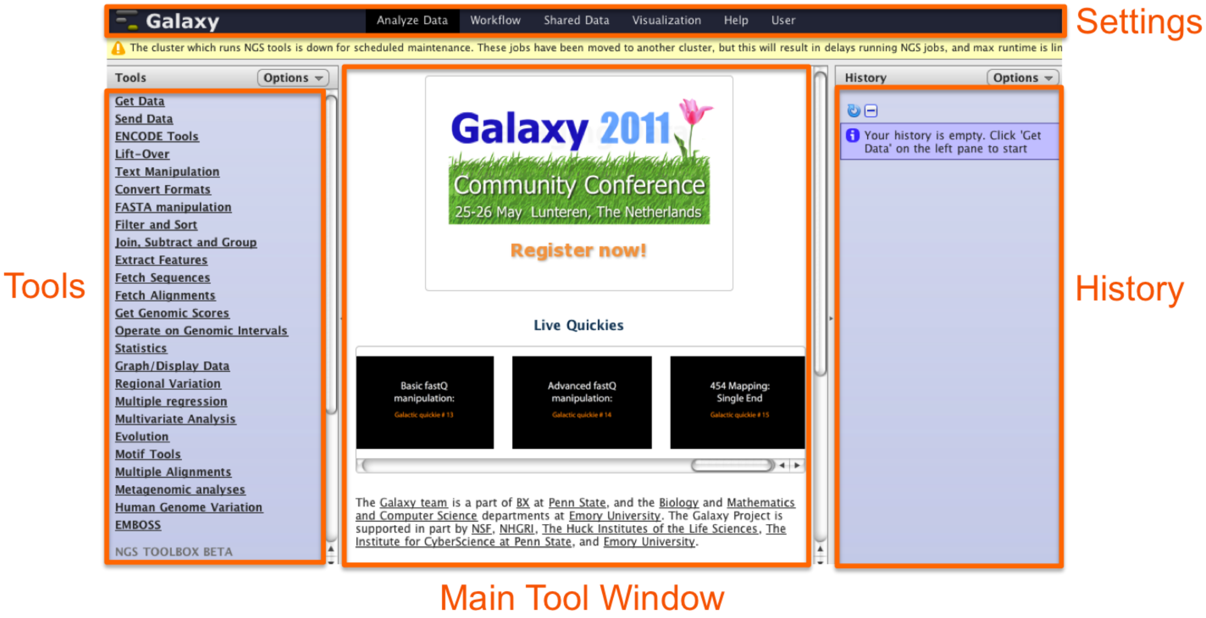
Please see Section Licensing for the MIT license.

2 Basic Galaxy instructions

This section describes the basic tasks in Galaxy: uploading data, data processing, some analysis tools, managing generated data as well as sharing with other users.

2.1 Galaxy interface layout

Galaxy is an open, web-based platform for accessible, reproducible, and transparent computational biomedical research#. Accessible: Users without programming experience can easily specify parameters and run tools and workflows. Reproducible: Galaxy captures information so that any user can repeat and understand a complete computational analysis. Transparent: Users share and publish analyses via the web and create Pages, interactive, web-based documents that describe a complete analysis Furthermore, using the 'Galaxy Tool shed'# it is easy and intuitive to extend a Galaxy server with specific tools and workflows for a wide range of tasks. These workflows are built by connecting the tools available in the Galaxy server to automatically run a sequence of steps. Workflows can be stored, shared and edited by dragging tools and parameters on an easy to use canvas.



The cluster which runs NGS tools is down for scheduled maintenance. These jobs have been moved to another cluster, but this will result in delays running NGS jobs, and max runtime is lin

Your history is empty. Click 'Get Data' on the left pane to start

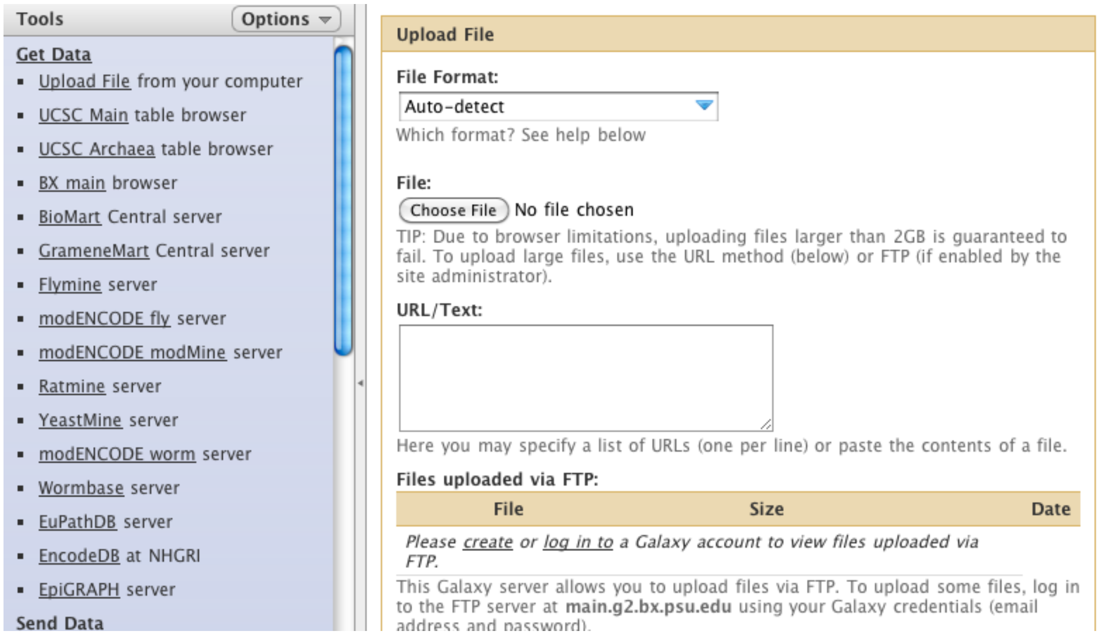
The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

Figure 1 Overview of a typical Galaxy server with the left hand pane showing a list of all available tools (categorized), the right hand pane the history (holding input files and output generated by the tools) and the center shows all options for a selected tool.

2.2 Getting data into Galaxy

Most tools in Galaxy need some sort of input data from the user. Galaxy offers multiple ways to retrieve data, most commonly through the 'Upload File' option available from the 'Get Data' category, allowing to select files from a local disk or available through an URL or FTP site (Figure 2). Galaxy is able to detect most data formats based on its extension and/or contents (use the 'Auto-detect' option selectable in the 'File Format' selector. For some formats, however, (typically binary files), it is advised to select the proper format from the drop-down menu. Also, some file formats are so called ‘composite’ data formats consisting of multiple files for which separate upload sections are shown once selected.

Pressing the ‘Execute’ button will upload the file into Galaxy as will be shown in the History pane on the right (Figure 3). Galaxy will inform the user about the progress and the history element will be assigned a number and turn green once the file has been uploaded. The number assigned will generally start with 1 and automatically increment on each added item to the history, whether it is an uploaded file or the result of executing a tool. Clicking on the (file) name in the history shows some basic details about the file and a set of buttons to view, edit, delete, save, view info or ‘rerun’ the uploaded file (please note that the ‘rerun’ button cannot be used with the file upload tool, this is designed for other Galaxy tools).



Which format? See help below

No file chosen

TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by the site administrator).

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Please create or log in to a Galaxy account to view files uploaded via FTP.

This Galaxy server allows you to upload files via FTP. To upload some files, log in to the FTP server at **main.g2.bx.psu.edu** using your Galaxy credentials (email address and password).

Figure 2 Getting data into Galaxy. On the left a list showing all the available methods. Clicking on the upload file link will bring up the settings pane (on the right). A file can be uploaded from a local disk using the ‘Choose File’ button. For some types of data the ‘File Format’ should be selected (for instance, Galaxy supports composite datasets that consist of multiple files).

2.3 Performing tasks on your data

Once the data is available in Galaxy it can be used as input for a tool. In Galaxy a tool can be described as an interface to any program that is wrapped into the common Galaxy interface. This program can be anything from a simple system command to a program processing huge amounts of data in parallel on a grid.

To analyse your data, select a tool from the left hand pane that supports your data and set or review its options in the centre pane. Since all Galaxy tools are configured to predefine their supported input types, any data input fields only show your compatible uploaded files. In some cases, Galaxy will even convert your file to another format in the background to make it compatible (i.e. from tabular to FASTA or from BED to GFF). These converters are also available from the ‘Convert Formats’ tool group.

For most tools you can find an explanation of the settings in the middle pane, right below the 'execute' button. Furthermore, its provides an overview of the functionality, details on the input fields and the output, show references to used methods, etc.

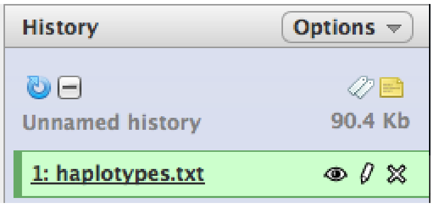


Figure 3 Uploaded files will show in your history as a numbered element. Clicking on the name of the file shows more details, such as a preview of the data, the data format and a 'save file' icon. Clicking on the 'eye' icon next to the filename will show the complete file (or only the first 1MB of data) in the center window. Galaxy keeps track of all actions performed and each tool result is stored in your history as a new numbered element. The item name usually describes the action of the tool. It is recommended to rename the data sets to a name that makes sense to your research. The output can be viewed in the browser, downloaded to disk or used as input for another tool.

2.4 Searching for tools

Use the 'search tools' input field on the top of the Galaxy tool list to search for tools, this will search the tool name, description and any supplied documentation.

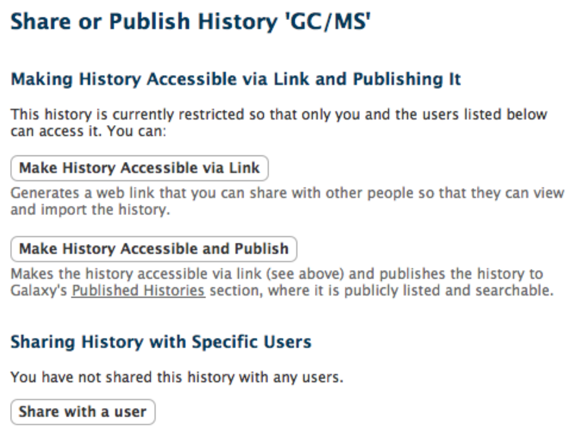
2.5 Managing datasets and histories

Galaxy supports multiple histories, which is convenient in case Galaxy is used for multiple tasks or experiments. To create a new history, click on the 'Options' button on top of the history and continue with 'Create New'. This creates a new 'Unnamed history' and makes it active. Please note that creating a new history will delete all items in the active session. It is recommended to rename the history, just click on 'Unnamed history' and enter a short description. Make sure you pick a name that accurately describes your session in order to be able to find it back later.

By clicking 'Saved Histories' from the 'Options' menu, a list of all histories is shown with many useful details such as date created, number of datasets it contains, etc. From here you can switch to another history activating it by simply clicking on its name (or select 'Switch' from the downward arrow next to the histories name). Histories can be deleted or shared (see below) by selecting them and pressing the appropriate button at the bottom. Data files can also be copied from one history to another. This is a useful feature if certain data files (e.g., libraries, databases) are used for more than one analysis.

2.6 Sharing data

Galaxy allows you to share your data with Galaxy users by clicking on the 'Options' link at the top of a history and selecting 'Share or Publish'. The page shown (see **Fout! Verwijzingsbron niet gevonden.**) allows multiple methods of sharing the complete history.



Share or Publish History 'GC/MS'

Making History Accessible via Link and Publishing It

This history is currently restricted so that only you and the users listed below can access it. You can:

Make History Accessible via Link

Generates a web link that you can share with other people so that they can view and import the history.

Make History Accessible and Publish

Makes the history accessible via link (see above) and publishes the history to Galaxy's Published Histories section, where it is publicly listed and searchable.

Sharing History with Specific Users

You have not shared this history with any users.

Share with a user

Figure 4 Methods of sharing a Galaxy history with other users. A history can be shared by providing a link to other users, publishing it (available to all users) or with another user (selectable).

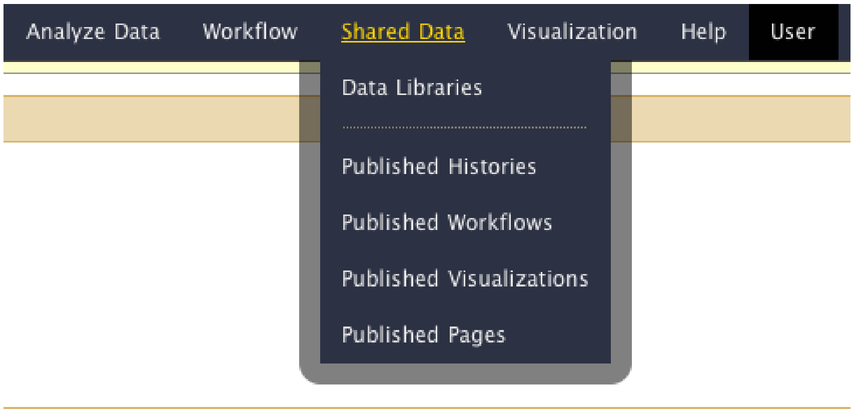


Figure 5 The sharing menu allows users to access shared data such as data libraries (data libraries are importable into a history), complete histories, workflows, visualizations and pages.

Single datasets can also be shared as a 'Data Library' by an administrative user (go to 'Admin', 'Manage data libraries' and click 'Create a new data library' in the top right). After setting permissions on this file, users can access it through the 'Shared Data' link in the top menu (see Figure 5).

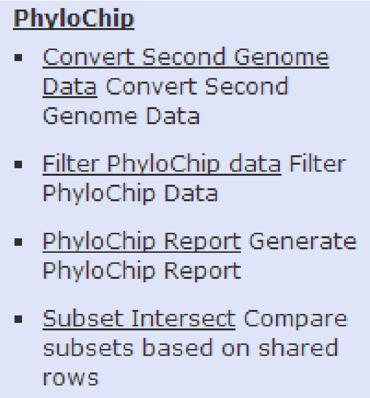
2.7 Galaxy Wiki links

The main Galaxy Wiki (http://wiki.g2.bx.psu.edu/) contains most of this sections instructions linked to interactive Galaxy histories offering a more in-depth introduction to all facets of Galaxy. The most important and useful Wiki pages are linked here:

- Managing Galaxy data sets: http://wiki.g2.bx.psu.edu/Learn/Managing%20Datasets
- Galaxy 101 tutorial: https://main.g2.bx.psu.edu/galaxy101
  - Please see section 4 in this tutorial for a view on creating and using Galaxy workflows
- Galaxy screencasts and demos: http://wiki.g2.bx.psu.edu/Learn/Screencasts
- Collection of published Galaxy Pages: https://main.g2.bx.psu.edu/page/list_published
- A repository of ready-made Galaxy tools: http://toolshed.g2.bx.psu.edu/

3 PhyloProfiler Galaxy tools

This section describes the PhyloProfiler tools available for analysing PhyloChip data. Currently, four tools are available in Galaxy (Figure 6). Note that these tools are only available on the galaxy server at https://galaxy.bioinf.nioo.knaw.nl/public/



PhyloChip

- Convert Second Genome Data Convert Second Genome Data
- Filter PhyloChip data Filter PhyloChip Data
- PhyloChip Report Generate PhyloChip Report
- Subset Intersect Compare subsets based on shared rows

Figure 6 The list of tools available for PhyloChip analysis. When using data in Second-Genome format (see section **Data Formats**), the order of processing would be to *(i)* convert the data to a single file using the 'Convert Second Genome Data' tool, followed by *(ii)* optional filtering using the 'Filter PhyloChip data' tool, *(iii)* generating a report showing various statistics using the 'PhyloChip Report' tool and finally *(iv)* comparing subsets and extracting intersect(s). See the sections below on a description of these tools.

3.1 Data Formats

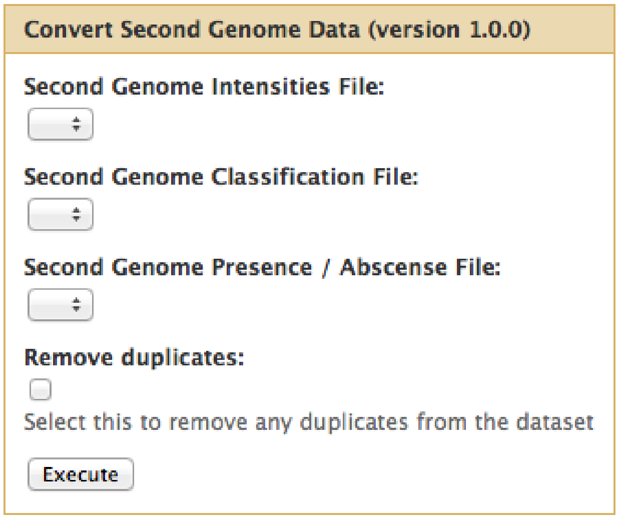
The current version supports two data formats, one legacy format on which the tools are build and (two variations of) the Second Genome format. See Section Data Format Examples in the Supplement below for example data. The appropriate columns for the classification data, intensity data and, if present, absence / presence data are automatically detected given that the column names complies to a few rules. The intensity data columns should be named as '`X_TREATMENT` where '`X`' is the sample number and '`TREATMENT`' is the treatment name, i.e. '`1_TR`', '`2_TR`', '`1_CTRL`', '`2_CTRL`'. The detection of this example will result in 2 treatments '`TR`' and '`CTRL`' with 2 samples each. This naming convention should be used in both the intensity file and any accompanying absence / presence file.

3.2 Converting Data

When using data from Second-Genome you need to prepare the data by converting it into a format readable by the rest of the PhyloProfiler tools. The 'Convert Second Genome Data' tool takes three second genome files (containing *(i)* a file with the intensities for all samples / variables, *(ii)* a file with the classification of all OTU's and *(iii)* a file containing presence / absence data (in binary notation) for each OTU in all samples.

Furthermore, there is a 'Remove Duplicates' option (see Figure 7) that removes any duplicate entries in the data (based on the intensities, the first entry for each duplicate is kept, the rest is discarded). These duplicates can be introduced by the pre-processing steps which involves rounding of raw values.

The output of the conversion can be further used in the other Galaxy PhyloChip tools as explained in the following sections.



Convert Second Genome Data (version 1.0.0)

Second Genome Intensities File:

Second Genome Classification File:

Second Genome Presence / Abscense File:

Remove duplicates:

□ Select this to remove any duplicates from the dataset

Figure 7 Configuration for the 'Convert Second Genome Data' conversion tool. Each of the input files accept plain text and tabular files.

3.3 Filtering PhyloChip data

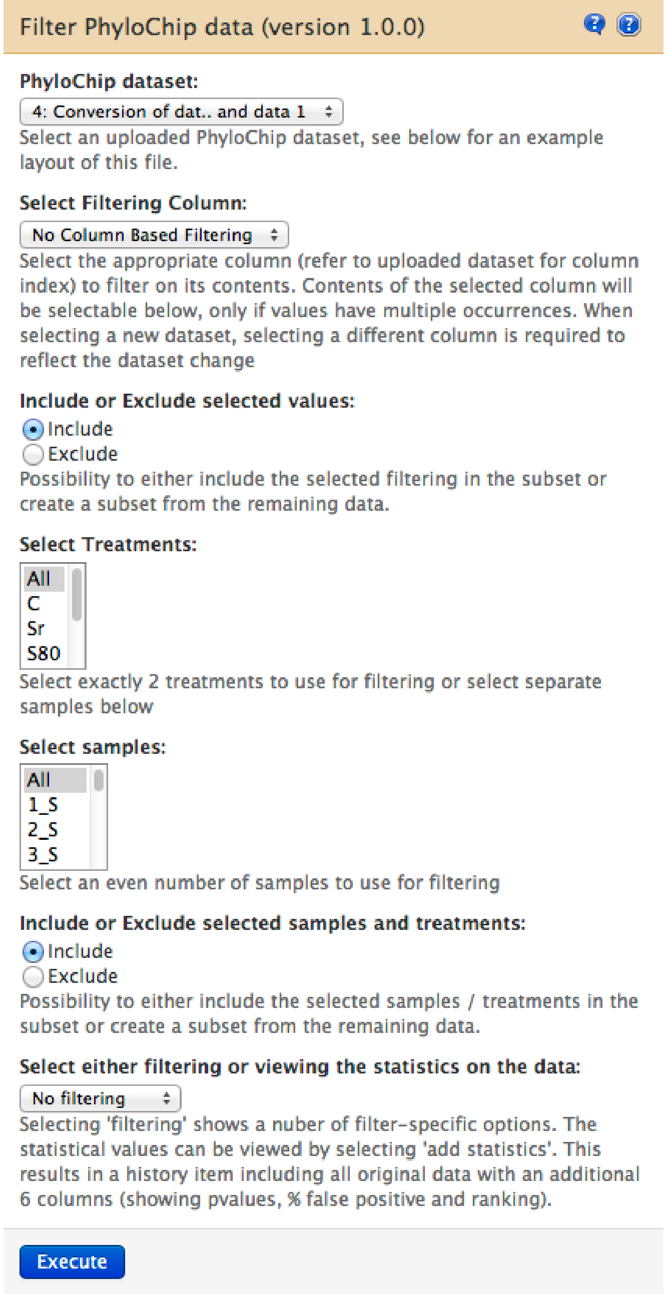
The data can be filtered to create a subset of the data for further processing / visualizing once data has been loaded into Galaxy (and converted if necessary). It is possible to filter on the classification data (*i.e.* exclude certain phyla or (sub) families), filter by selecting treatments and /or samples as well as on numerical statistics using the 'Filter PhyloChip Data' Galaxy tool (see Figure 8).

Filtering on a classification column involves first selecting the correct column number from the drop-down menu followed by selecting one or more values from the select box (see Figure **9**). The same process is used for selecting individual treatments and samples to be included or excluded from the dataset (selecting a treatment automatically includes all samples from that treatment).

Using a 'Rank Product' method (see Section Filtering below), data can be filtered on p-value or on the predicted percentage of False Positives. The filtering can be applied on data ordered by increasing or decreasing intensities (numbers are calculated separately for OTU's with increasing or decreasing profiles) or combined (selecting and combining the top-scoring OTU's from the complete dataset), see Figure 10.

It is also possible to see the numerical statistics before applying a filter by selecting 'Add Statistics' in the filter selection box. This results in a new dataset (other filtering options still apply when selected) with columns added to the end showing all calculated statistics, no actual statistical filtering is applied using this option. This can help to decide on a cut-off value that can be entered when applying the statistical filter. The Galaxy 'run this job again' function helps in this regard such that reapplying the filter after viewing the statistics only takes one step instead of repeating any previous filtering selections.

The proposed usage of the filtering tool is to filter the data subsequently to arrive at the desired subset (the output of the tool is also allowed as input).



Filter PhyloChip data (version 1.0.0)

PhyloChip dataset:

4: Conversion of dat.. and data 1

Select an uploaded PhyloChip dataset, see below for an example layout of this file.

Select Filtering Column:

No Column Based Filtering

Select the appropriate column (refer to uploaded dataset for column index) to filter on its contents. Contents of the selected column will be selectable below, only if values have multiple occurrences. When selecting a new dataset, selecting a different column is required to reflect the dataset change

Include or Exclude selected values:

● Include
○ Exclude

Possibility to either include the selected filtering in the subset or create a subset from the remaining data.

Select Treatments:

All
C
Sr
S80

Select exactly 2 treatments to use for filtering or select separate samples below

Select samples:

All
1_S
2_S
3_S

Select an even number of samples to use for filtering

Include or Exclude selected samples and treatments:

● Include
○ Exclude

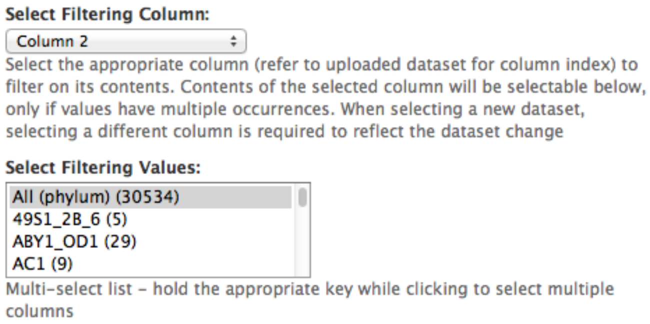
Possibility to either include the selected samples / treatments in the subset or create a subset from the remaining data.

Select either filtering or viewing the statistics on the data:

No filtering

Selecting 'filtering' shows a nuber of filter-specific options. The statistical values can be viewed by selecting 'add statistics'. This results in a history item including all original data with an additional 6 columns (showing pvalues, % false positive and ranking).

Figure 8 Filtering PhyloChip data. After selecting a data set for filtering, multiple methods can be combined including filtering on contents of the classification columns, filtering by selecting either one or more treatments and / or samples or filtering using statistical methods. Refer to the text for further details on each of these methods.



Select Filtering Column:

Column 2

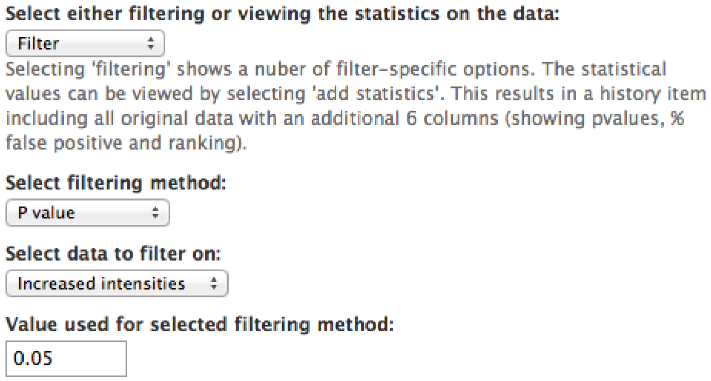
Select the appropriate column (refer to uploaded dataset for column index) to filter on its contents. Contents of the selected column will be selectable below, only if values have multiple occurrences. When selecting a new dataset, selecting a different column is required to reflect the dataset change

Select Filtering Values:

All (phylum) (30534)
49S1_2B_6 (5)
ABY1_OD1 (29)
AC1 (9)

Multi-select list – hold the appropriate key while clicking to select multiple columns

Figure 9 **Filtering based on 'phylum' by selecting column number 2 in this case. The number in brackets denotes the occurrence of that specific entry. The list shown contains all unique values for the selected column.**



Select either filtering or viewing the statistics on the data:

Filter

Selecting 'filtering' shows a nuber of filter-specific options. The statistical values can be viewed by selecting 'add statistics'. This results in a history item including all original data with an additional 6 columns (showing pvalues, % false positive and ranking).

Select filtering method:

P value

Select data to filter on:

Increased intensities

Value used for selected filtering method:

0.05

Figure 10 Statistical filtering methods available for PhyloChip data. The methods are either filtering on P value or a predicted false-positive percentage. Using the 'Select data to filter on' orders the results and retains data based on either increased- or decreased-intensity values.

3.4 Creating subset intersect (Venn diagram)

The 'Subset Intersect' tool allows comparing multiple (up to four) subsets based on any of the classification columns and displays the subset relations using a Venn diagram (see Figure 11). Next to the visualization it is possible to select a single intersect for export to a new subset (see Figure 12). Use the 'Add new extra data set' button when comparing more than two subsets (see Figure 13).

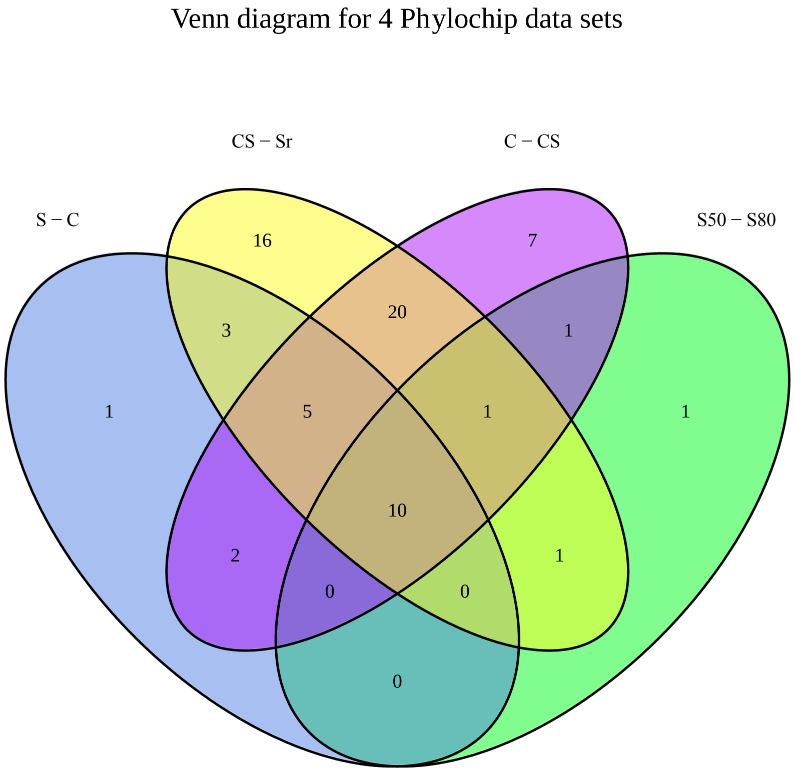


Figure 11 Example Venn diagram displaying the relations (intersections) between four subsets with the included treatments displayed on top of each ellipse. This particular example shows shared phyla across the subsets.

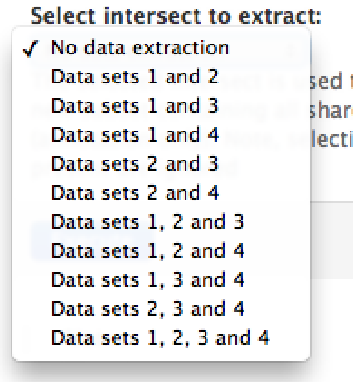
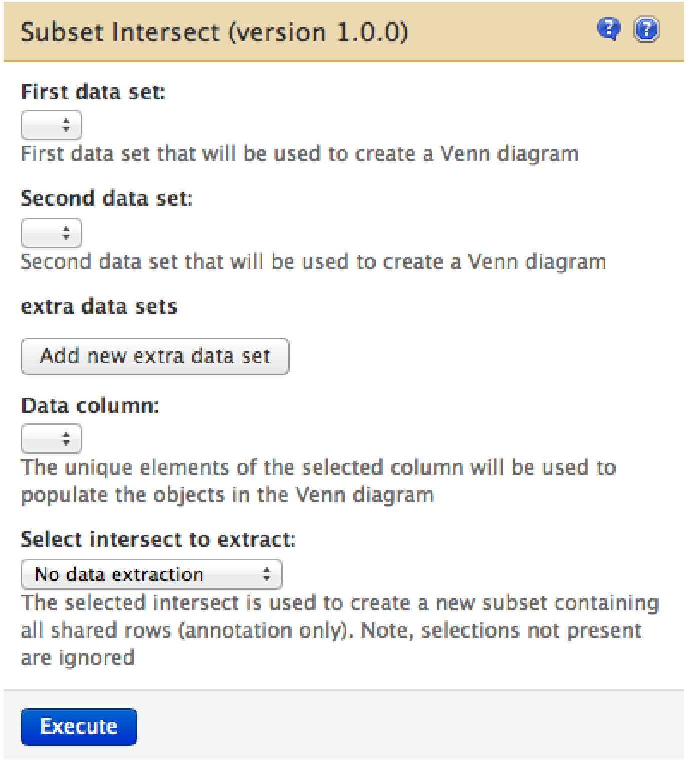


Figure 12 Possible choices for selecting an intersect for export available from comparing four subsets. Future versions might support interactive Venn diagram where clicking a section results in a new subset.


Subset Intersect (version 1.0.0)

First data set:

First data set that will be used to create a Venn diagram

Second data set:

Second data set that will be used to create a Venn diagram

extra data sets

Data column:

The unique elements of the selected column will be used to populate the objects in the Venn diagram

Select intersect to extract: No data extraction

The selected intersect is used to create a new subset containing all shared rows (annotation only). Note, selections not present are ignored

Figure 13 Subset Intersect tool for comparing subsets, creating a Venn diagram and a new subset based on the selected intersect to extract. The ‘Data column’ parameter points to the classification column that is used for comparison. Use the ‘Add new extra data set’ button to include more than two subsets in the comparison.

3.5 Visualizing data

Another tool in the PhyloChip package allows for elaborate visualizations of the data using the ‘PhyloChip Report’ tool. Its settings are displayed in Fout! Verwijzingsbron niet gevonden.. Data can be visualized column- or row-based by selecting the ‘order’. Next, two types of reports can be created, a standard PDF file containing only the graphics or a more elaborate LaTeX generated PDF file containing more details about the data and included graphics. This type of report also shows all selected settings, which can be used when re-running the visualization with different settings.

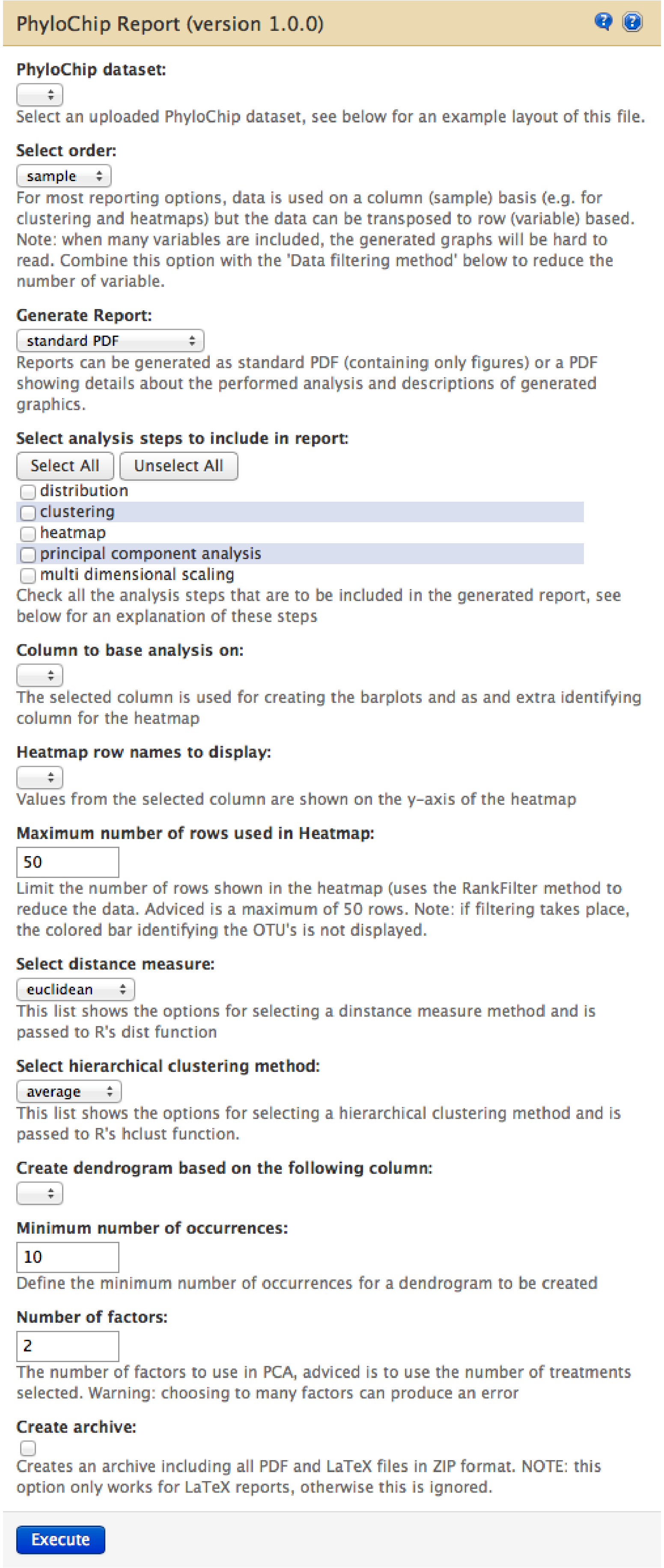
The analysis steps can be selected separately after which various settings for the analysis steps can be set. Please note that these settings always show, even when not selecting the analysis step (setting is then ignored).

Starting with the ‘Column to base analysis on’, which allows the selection of the column data to use for all available analysis steps. The ‘Heatmap row names to display’ sets the column shown on the Y-axis of the heat map identifying the separate rows. Next, the ‘Maximum number of rows used in Heatmap’ sets a limit on the number of rows shown. A heat map for this type of data loses its usefulness when to many rows of data are included. However using the PDF format for images, zooming is no issue but text labels may still overlap. Setting a number smaller than the number of rows in the data (thus using a limit) uses the ‘Rank Product’ filtering method to select the rows showing the most change (uses selection based on combined p-values for increased / decreased intensities, see the **Fout! Verwijzingsbron niet gevonden.** section below).

Continuing, the distance measuring algorithm, hierarchical clustering method and the column on which the clustering is based can be selected. Above settings are used for the clustering visualization. Clustering results in a single dendrogram showing either the clustering of the samples or OTU's based on the 'order' selected. Also shown is a separate dendrogram showing the clustering for each unique value in the selected column. The 'Minimum number of occurrences' setting limits the number of dendrograms shown to only those entries that contain more values than the given number (clustering when only 2 rows of data are present for a particular entry makes no sense).

The next setting sets the number of factors used in PCA analysis. Selecting 1 or 2 results in a single PCA plot showing the first two components whereas setting this to 3 or more factors will show three PCA plots (component 1 vs. 2, 1 vs. 3 and 2 vs. 3).

The 'Create archive' setting works in combination with creating a LaTeX style report and places an extra element in the users history containing a ZIP archive file. This archive holds all figures as separate PDF documents, a table showing a table with the occurrences of values in the selected dendrogram column, an '.RData' file that can be loaded into R holding all R functions used for analysis as well as data objects that can be used to replicate all steps (or adjust some graphical parameters). Finally, the LaTeX source file and R Sweave file are stored as well. Selecting this in combination with a standard PDF report will not result in an archive file.



PhyloChip Report (version 1.0.0)

PhyloChip dataset:

Select an uploaded PhyloChip dataset, see below for an example layout of this file.

Select order:

sample

For most reporting options, data is used on a column (sample) basis (e.g. for clustering and heatmaps) but the data can be transposed to row (variable) based. Note: when many variables are included, the generated graphs will be hard to read. Combine this option with the 'Data filtering method' below to reduce the number of variable.

Generate Report:

standard PDF

Reports can be generated as standard PDF (containing only figures) or a PDF showing details about the performed analysis and descriptions of generated graphics.

Select analysis steps to include in report:

☐ distribution
☐ clustering
☐ heatmap
☐ principal component analysis
☐ multi dimensional scaling

Check all the analysis steps that are to be included in the generated report, see below for an explanation of these steps

Column to base analysis on:

The selected column is used for creating the barplots and as and extra identifying column for the heatmap

Heatmap row names to display:

Values from the selected column are shown on the y-axis of the heatmap

Maximum number of rows used in Heatmap:

50

Limit the number of rows shown in the heatmap (uses the RankFilter method to reduce the data. Adviced is a maximum of 50 rows. Note: if filtering takes place, the colored bar identifying the OTU's is not displayed.

Select distance measure:

euclidean

This list shows the options for selecting a dinstance measure method and is passed to R's dist function

Select hierarchical clustering method:

average

This list shows the options for selecting a hierarchical clustering method and is passed to R's hclust function.

Create dendrogram based on the following column:

Minimum number of occurrences:

10

Define the minimum number of occurrences for a dendrogram to be created

Number of factors:

2

The number of factors to use in PCA, adviced is to use the number of treatments selected. Warning: choosing to many factors can produce an error

Create archive:

☐

Creates an archive including all PDF and LaTeX files in ZIP format. NOTE: this option only works for LaTeX reports, otherwise this is ignored.

Figure 14 PhyloChip Report tool settings

Please see the included documentation in the Galaxy ‘PhyloChip Report’ tool for further details (this will be kept up to date after changes).

4 Maintenance

This section briefly describes some frequently used administrative tasks for managing a local Galaxy server as well as PhyloChip specific maintenance.

4.1 Updating the PhyloChip software

4.1.1 Retrieving an update from the repository

The software is hosted on a repository maintained by NBIC and can be reached at: https://trac.nbic.nl/brs2010p29/. Included in the project is a tool specifically used for automatically updating the project. This tool compares the current local version with the one on the repository and updates any changes. This tool can be run as any other tool and the output (update log) is shown in the history. Please examine this log after running the tool, if any XML files were updated, see Reloading Galaxy tools below.

4.1.2 Updating dependencies

The project makes use of the Python and R programming languages, which are required for the software to work. There are however some R libraries that might need to be reinstalled when updating the R software because they were compiled specifically for the previously installed version. To (re)install all required R libraries, perform the following three commands in an active R session (make sure that the user running Galaxy can access the installed libraries):

```
install.packages(c('gplots', 'plotrix', 'ecodist', 'vegan', 'fpc', 'cluster', 'MASS')
source("http://bioconductor.org/biocLite.R")
biocLite("RankProd")
```

4.2 Reloading Galaxy tools

All Galaxy tools are defined in XML files that Galaxy loads when starting. When an XML file is changed, Galaxy needs to be told to reload that tool. One easy way to do this is restart the Galaxy server, however this will stop any running job (from all users!). A better solution is to use the Admin panel (login as a Galaxy administrator and click the 'Admin' link in the top menu), which holds a 'Reload a tool's configuration' tool. From this tool, select the changed tool and click 'Reload'. Galaxy will show if everything went ok after which the updated tool is available from the main Galaxy page (go to 'Analyse Data' from the top menu).

4.3 Restarting the Galaxy server

Restarting a properly installed Galaxy server requires an administrator user on the host server. This user can run a script from the terminal that, most likely, is installed in `/etc/init.d` which can be called with (use `sudo` to gain the correct privileges):

```
sudo /etc/init.d/galaxy restart
```

Other valid commands are `start` and `stop`. Another way of restarting Galaxy is to simply reboot the host server.

4.4 Changing Galaxy configuration settings

All Galaxy settings can be found in the 'universe_wsgi.ini' file located in the root of the Galaxy installation. Please see http://wiki.g2.bx.psu.edu/Admin for how to configure your server.

Configuring the list of tools available to the users however can be found in the 'tool_conf.xml' file also located in the root of the Galaxy installation. This file lists all currently installed tools in XML format. Tools can be excluded by using standard XML comment tags ('<!--' to open, '-->' to close). After restarting the server, any excluded tool will be hidden from the tool list.

4.5 Managing Galaxy server administrators

In Galaxy's main configuration file ('universe_wsgi.ini') the administrative users can be managed. The 'admin_users' setting is a comma-separated list of email addresses of registered users that are given those rights. Add or remove an address from this list to give or revoke administrative rights. These settings will be applied after a server restart.

4.6 PhyloChip Installation

Following is a short description on how to install the PhyloChip project into an existing Galaxy installation. Most Linux operating systems provide an easy to use method (package manager) of installing software, which can be used to install most requirements listed below (*i.e.* `apt-get` on Debian and Ubuntu, `yast2` on Suse, `yum` on Fedora, etc.).

Requirements:

- R (http://www.r-project.org/)
  - Several R libraries:
    - gplots
    - plotrix
    - ecodist
    - vegan
    - fpc
    - MASS
    - cluster
    - RankProd (BioConductor library)
- LaTeX (http://www.latex-project.org/)
- Python (http://www.python.org/ version >= 2.6)
- SVN (http://subversion.apache.org/)

Installation:

- Check out the latest version of the software from SVN into the `galaxy_dist/tools/phylochip` folder using:
  ```
  svn co https://trac.nbic.nl/svn/brs2010p29/trunk galaxy-dist/tools/phylochip
  ```
- Edit the galaxy_dist/tool_conf.xml file and add a PhyloChip section with all respective XML files (see other tools' sections in that file for examples)
- Restart the Galaxy server and confirm that all tools are visible in the tools list

4.7 Reporting Errors

When an error occurs in Galaxy after running a tool, this is visible as a red colored history element indicating a failure. By expanding the history element (clicking on the numbered name of the output), a green 'bug' button appears at the bottom. By clicking on this button the error message will be displayed. If this error is non-descriptive or caused by a software error, please contact the author of this manual (or appointed maintainer of the software) reporting the error and the settings used that caused the error.

5 Supplement

5.1 Project Structure

All PhyloChip Galaxy tools reside in Galaxy’s `tools/phylochip` folder. The listing below describes all files part of this project.

▪ *dataconv.xml*	*combines second-genome data into a single dataset*
▪ *compare_subsets.xml*	*Subset intersect tool (Venn Diagrams)*
▪ *filter_data.xml*	*provides filtering methods for input data*
▪ *report.Rnw*	*R Sweave template file describing LaTeX report layout*
▪ *report.py*	*Contains code for selecting samples/treatments*
▪ *report.xml*	*Main tool for visualizing PhyloChip data*
▪ *update.py*	*Used for updating the project*
▪ *update.xml*	*Update tool*
▪ *LICENSE.txt*	*Licensing information*
▪ ***R-scripts***	
o *Renv.R*	*Defines all libraries loaded for R*
o *barplot.R*	*Creates a distribution bar plot*
o *clustering.R*	*Used for clustering the data*
o *data-management.R*	*Reads and processes input data*
o *filtering.R*	*Filtering methods*
o *heat map.R*	*Creates a heat map*
o *mds.R*	*Performs Multi Dimensional Scaling*
o *pca.R*	*Performs Principal Component Analysis*
o *venn.R*	*R code used for creating Venn diagrams*

5.2 Methods

Following is a list of methods used for the various tools included in the PhyloChip project. These methods describe the technical details of the filtering and visualization parts, linking to other documentation sources for each specific method. If applicable, any arguments or parameters used will be listed as well. Below is the layout of the project, reflecting the version available at time of writing this document.

5.2.1 Filtering

For the numerical filtering method applied, the BioConductor ‘RankProd’ package is used[1]. The actual R functions from this package that are used are the ‘RP’ and ‘topGene’ functions with default settings except for using a pre-defined seed for the ‘RP’ function. This should not affect the results, however it allows to firstly run the filtering tool with the ‘Add Statistics’ to get the statistical data and use this to define a cut-off value for another run to filter based on this value. With a non-fixed random seed, this might result in missing or seeing extra data rows.

[1] Breitling, R., Armengaud, P., Amtmann, A., and Herzyk, P.(2004) *Rank Products:A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments*, FEBS Letter, 57383-9

Manuals:
Bioconductor RankProduct (‘RP’ and ‘topGene’):
http://www.bioconductor.org/packages/2.6/bioc/manuals/RankProd/man/RankProd.pdf

5.2.2 Distribution bar plot

Uses basic R ‘bar plot’ function to visualize the distribution of the dataset based on a selected classification column, using custom labels and axis formatting as well as distributions per treatment and per sample. Filtering is applied to only show phyla with > 0.25% occurrence to reduce clutter on the bar charts.

Manuals:
R ‘barplot’: http://stat.ethz.ch/R-manual/R-devel/library/graphics/html/barplot.html

5.2.3 Clustering

The clustering uses both the R ‘dist’ and ‘hclust’ functions for generating the clustering, which is then visualized using the basic ‘plot’ function. The methods selected by the user for both the distance measure and hierarchical clustering are passed to these functions. When ‘Bray Curtis’ is selected as distance measure, the ‘bcdist’ function (‘ecodist’ library) is used after which the data is manually transformed to reflect percentages (after clustering: (`1 – height * 100)`). Furthermore, for all other distance measurements used, the data is scaled using R’s ‘scale’ function, which performs a root-mean-square transformation.

Manuals:
R ‘dist’ function: http://stat.ethz.ch/R-manual/R-patched/library/stats/html/dist.html
R ‘bcdist’ function: http://rss.acs.unt.edu/Rdoc/library/ecodist/html/bcdist.html
R ‘hclust’ function: http://stat.ethz.ch/R-manual/R-patched/library/stats/html/hclust.html
R ‘scale’ function: http://stat.ethz.ch/R-manual/R-devel/library/base/html/scale.html

5.2.4 Principal Component Analysis

The PCA analysis is performed using the ‘factanal’ function with the default settings and the number of factors given by the user supplied. Depending on the number of factors given, one or three PCA plots are created (see Section Visualizing data). If possible, the computed p-value is shown in the plot title.

Manuals:
R ‘factanal’ function: http://rss.acs.unt.edu/Rdoc/library/stats/html/factanal.html

5.2.5 Multi Dimensional Scaling

The MDS analysis is performed using the ‘cmdscale’ function for the metric MDS and the ‘sammon’ function (‘MASS’ library) for non-metric MDS. Prior to performing MDS, the distance is calculated using the above-mentioned ‘dist’ function combined with the Euclidian distance measure. Other configuration options used are ‘k’ (dimension of the configuration) for both metric and non-metric MDS, which is set at the number of treatments in the data.

Manuals:
R ‘cmdscale’: http://stat.ethz.ch/R-manual/R-devel/library/stats/html/cmdscale.html
R ‘sammon’: http://stat.ethz.ch/R-manual/R-devel/library/MASS/html/sammon.html

5.2.6 Heat map

A slightly more advanced (compared to R’s default) function is used for visualizing the heat map, namely the ‘heatmap.2’ function from the R ‘gplots’ library. The generated heat map shows the intensities for the included OUT’s in a colored matrix (refer to the included legend on the top left side of the heat map) as well as a distribution of the intensity values in a histogram (displayed in log scale, see legend) and dendrograms on both axis showing the clustering for all samples / OTU's. When the user defined a value for the ‘Maximum number of rows used in Heat map’ that is smaller than the number of rows in the dataset, filtering is applied as explained in Section Visualizing data. Also for the heat map, the data is scaled (root-mean-square transformation) prior to generating the heat map.
When no filtering is applied when creating the heat map, an extra column is added to the left of the heat map that uses the same color-coding as the bar plot to identify the phyla that the OTU in that particular row belongs too (see Figure 15). This, combined with the clustering can give an overview of trends in different phyla. Future versions will show this extra column also when filtering is applied as well as selectable classification data for display in this column (*i.e.* the domain or class, etc.).

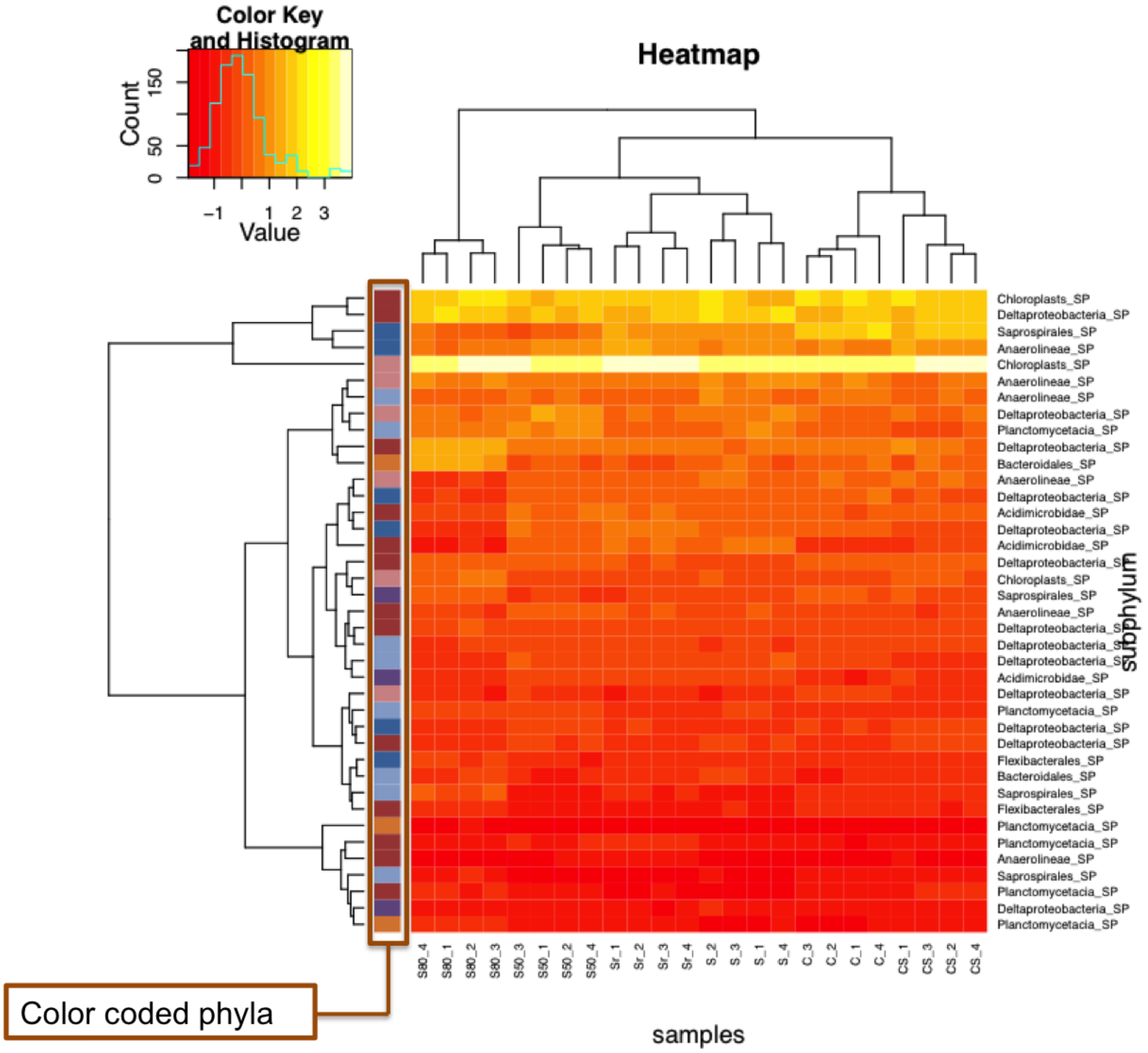


Figure 15 Heat map image generated from a dataset containing 40 OTU's (vertical) and 6 treatments (24 samples, horizontal) with their respective clustering on opposite sides. Indicated is the extra vertical column on the left-hand side showing the phyla for each OTU (refer to the bar plot for corresponding phyla names).

Manuals:
R ‘heat map.2’ function: http://rss.acs.unt.edu/Rdoc/library/gplots/html/heatmap.2.html

5.3 Data Format Examples

Legacy data format:

Columns 1 – 5:

ID	scoring_set_id	lineage	domain	phylum
1	2021759	...	Bacteria	Proteobacteria
2	2039310	...	Bacteria	Firmicutes
3	2025893	...	Bacteria	Proteobacteria
4	2022982	...	Unclassified	Unclassified
5	2046791	...	Bacteria	Bacteroidetes

Columns 6 – 9:

subphylum	class	order	family
Gammaproteobacteria_SP	Betaproteobacteria_CL	Janthinobacterium_OR	Clostridiales
Clostridia_SP	Clostridiales_CL	Clostridiales	Clostridiaceae
Gammaproteobacteria_SP	Pseudomonadaceae_CL	Pseudomonadaceae_OR	Pseudomonadaceae
Unclassified	Unclassified	Unclassified	Unclassified
Bacteroidales_SP	Bacteroidales_CL	Bacteroidales	Unclassified

Columns 11 – 17:

subfamily	otu	gene	1_TreatmA	2_TreatmA	3_TreatmA	4_TreatmA
sfA	21759	EU535474.1	7994.8628	9851.2696	8203.6745	8376.864
sfA	39310	EU473317.1	4757.6068	3364.3949	3949.9793	4239.1401
sfA	25893	EU537571.1	5218.3446	5230.5229	4201.556	3862.0709
uhC	22982	EU869397.1	11841.8437	12636.7578	12055.3778	11826.8763
sfE	46791	EU381828.1	11136.7227	11902.4653	12852.4513	10276.4845

Columns 18 – n:

1_TreatmB	2_TreatmB	3_TreatmB	4_TreatmB
10523.5114	10512.8442	10691.7698	10669.404
3354.2675	3440.0329	3573.9229	3709.0157
2683.2917	2250.3824	2446.6322	1847.6469
12467.1893	11997.0065	11372.6784	11983.4426
11429.3607	10784.4902	11187.1832	12647.5633

Second Genome – Classification file contents

All columns except the ‘otu’ and ‘phylum’ columns contain the data from the previous column with a new value added at the end, separated by a semicolon ‘:’. For displaying purposes, parts of these columns have been replaced by ‘…’.

#otu	phylum	subphylum	class
53624	Bacteria:Verrucomicrobia	Bacteria:Verrucomicrobia:05D2Z34_SP	Bacteria:Verrucomicrobia:05D2Z34_SP:Unclassified
53025	Bacteria:Planctomycetes	Bacteria:Planctomycetes:WPS-1_SP	Bacteria:Planctomycetes:WPS-1_SP:BD2-16_CL
14181	Bacteria:Firmicutes	Bacteria:Firmicutes:Unclassified	Bacteria:Firmicutes:Unclassified:Unclassified
2977	Bacteria:Firmicutes	Bacteria:Firmicutes:Clostridia_SP	Bacteria:Firmicutes:Clostridia_SP:C23_k02_CL
41041	Bacteria:Firmicutes	Bacteria:Firmicutes:Mollicutes_SP	Bacteria:Firmicutes:Mollicutes_SP:RF39_CL

Second Genome – Classification file contents, continued:

order	family
...:Unclassified:Unclassified	...:Unclassified:Unclassified
...:BD2-16_CL:MB-C2-105_OR	...:MB-C2-105_OR:Unclassified
...:Unclassified:Unclassified	...:Unclassified:Unclassified
...:Unclassified:Unclassified	...:Unclassified:Unclassified
...:Succinivibrionaceae_CL:Succinivibrionaceae_OR	...:Succinivibrionaceae_OR:Succinivibrionaceae

Second Genome – Classification file contents, continued:

subfamily	gene
...:Unclassified:Unclassified:sfA	DQ329656.1 gg_id:…
...:MB-C2-105_OR:Unclassified:sfI	EU385713.1 gg_id:…
...:Unclassified:Unclassified:vbX	EU617865.1 gg_id:…
...:Unclassified:Unclassified:soU	EU778710.1 gg_id:…
...:Succinivibrionaceae_OR:Succinivibrionaceae:sfB	EU382019.1 gg_id:…

Second Genome – Intensity file contents:

otu	1_TreatmA	2_TreatmA	1_TreatmB	2_TreatmB
53624	11418	11697	11541	11535
53025	12817	12830	12890	12852
47323	12024	11978	11940	12225
34992	12135	12208	12124	12023

Second Genome – Presence / Absence file contents:

otu	1_TreatmA	2_TreatmA	1_TreatmB	2_TreatmB
53624	0	0	0	0
53025	0	1	1	1
47323	1	1	0	0
34992	0	0	0	0
32973	0	1	0	0

5.4 Licensing

Copyright (c) 2010 - 2012 Bioinformatics Research Support group, NBIC

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.