User's Guide

Using the R-Peridot Graphical User Interface (GUI) on Windows and GNU/Linux Systems

Pitágoras Alves <alves.pitagoras@gmail.com>
01/06/2018
Natal-RN, Brazil

Index

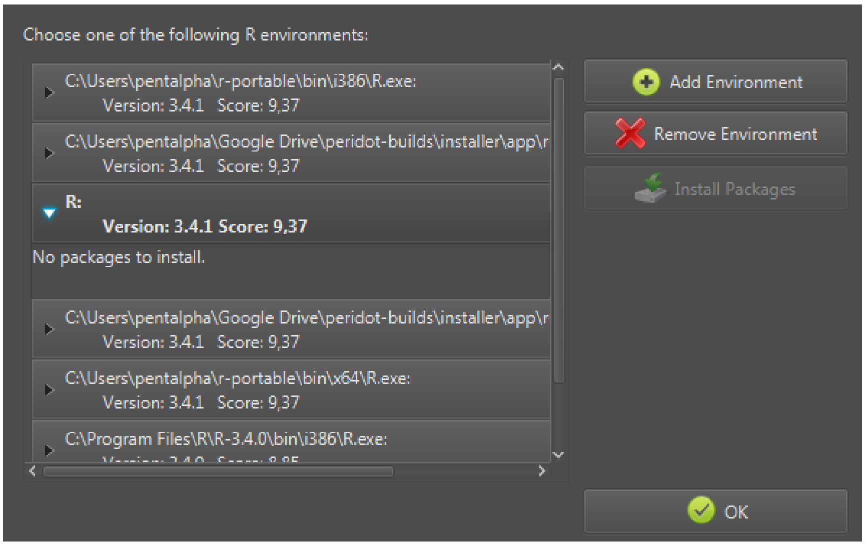
R-Peridot attempts to deliver an intuitive user experience so you don't have to read long instructions to use the software. But in case you need such instructions, read this manual.

In order to open R-Peridot you can use a desktop/menu shortcut (if you used the Windows installer) or directly open the "r-peridot-gui.jar" file.

1. The R Environment Manager

At the first time opening R-Peridot, the first interface you will be presented to is the R Environment Manager. After that you can always open it using the “Tools > R Environments” menu option.

R-Peridot needs a R environment to run its modules, so the first thing to do is select one.



Choose one of the following R environments:

C:\Users\pentalpha\r-portable\bin\i386\R.exe:
Version: 3.4.1 Score: 9,37

C:\Users\pentalpha\Google Drive\peridot-builds\installer\app\r
Version: 3.4.1 Score: 9,37

R:
Version: 3.4.1 Score: 9,37

No packages to install.

C:\Users\pentalpha\Google Drive\peridot-builds\installer\app\r
Version: 3.4.1 Score: 9,37

C:\Users\pentalpha\r-portable\bin\x64\R.exe:
Version: 3.4.1 Score: 9,37

C:\Program Files\R\R-3.4.0\bin\i386\R.exe:
Version: 3.4.0 Score: 8,85

R Environment Manager: Through this interface you can select/add/remove the desired R environment or install missing required packages.

The R Environment Manager scans the system for R environments, but it may not find your R installation if it is in an unusual directory. If no environments are listed, you can click on the “**Add Environment**” button and a file chooser dialog will appear, use it to select the R executable you want to use. If you don’t want an environment on the list, you can use the “**Remove environment**” button.

After selecting the desired environment, R-Peridot will show which packages are missing. If there are any missing packages, click on the “**Install Packages**” button. It will try to install the packages from R-Peridot’s package repository, which contains the most compatible version of each package. If it cannot install from our repository, the installer will attempt to use *CRAN* or *Bioconductor* directly. In case errors occur, try to install the packages directly within the desired R environment (the Installation Guide has detailed information on how to do this).

Each listed environment has a score to help you choose the best option. The score ranges from 0.0 to 10.0 and is calculated based on the R version (the recommended version is R 3.4.1) and the packages currently installed in it. A score above 9.0 is just fine. You can also try to run R-Peridot using environments with low scores, but in this case, we cannot assure you the modules will function properly. After selecting the environment you want, click on the “**OK**” button.

2. Input: The Gene Count Reads Table

A gene count reads table file, or simply count reads file, is a table file in which rows represent genes and columns represent samples, with the values of each cell being the (not yet normalized) RNA read counts. This tool doesn’t support *OpenDocument* (.ods) and *Microsoft Office* (.xls) spreadsheet file formats. The count tables must be plain text files, with the columns being separated by characters like tabs, semi colons, commas, etc. A first row with headers, the sample names, is recommended. Another recommendation, in case you want to use the “*ClusterProfiler”* module, is to have a first column with labels who are commonly used IDs in bioinformatics (*ensembl*, *kegg*, *refseq*, *symbol* and so on).

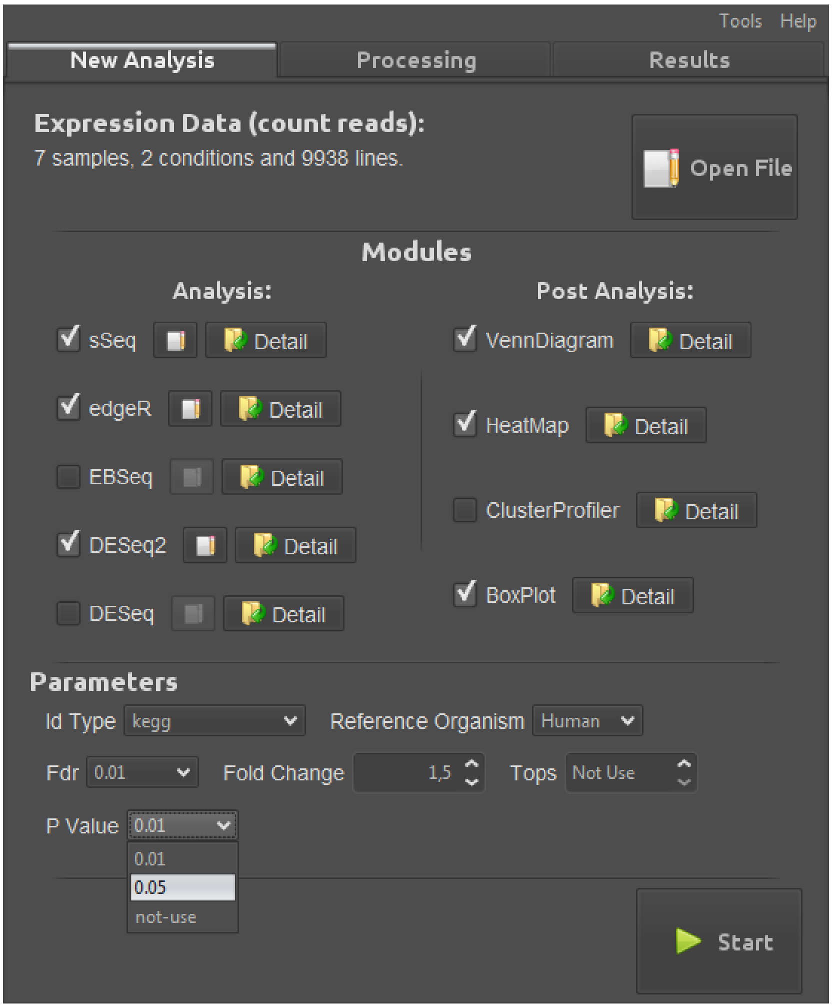
gene-id	<sampleID>	<otherSampleID>	...	<lastSampleID>
<gene-id-1>	<count-read>	<count-read>	...	<count-read>
<gene-id-2>	<count-read>	<count-read>	...	<count-read>
...	...	...	...	...

Example of count reads table

These tables can be created using software such as htseq-count (read chapter 14 of its documentation for detailed instructions) and featureCounts. For testing purposes, we recommend using the count tables from ReCount, a multi-experiment resource of analysis-ready RNA-seq gene count datasets.

3. “New Analysis” Tab

Here you can define the input data, the desired modules and the parameter values. The interface begins only with the option to define the input data. When it is defined, the list of checkboxes to choose the modules is revealed. Once modules are chosen, the parameters are revealed.



Expression Data (count reads):
7 samples, 2 conditions and 9938 lines.

Modules

Analysis:
- ☑ sSeq
- ☑ edgeR
- ☐ EBSeq
- ☑ DESeq2
- ☐ DESeq

Post Analysis:
- ☑ VennDiagram
- ☑ HeatMap
- ☐ ClusterProfiler
- ☑ BoxPlot

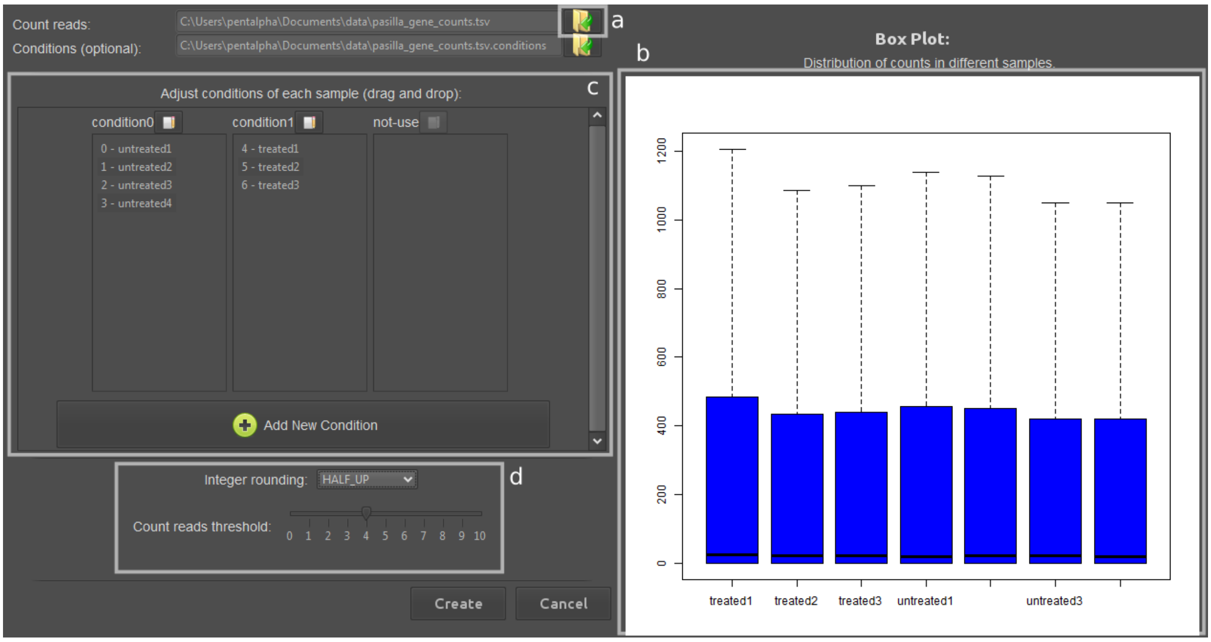
Parameters

Id Type: kegg
Reference Organism: Human
Fdr: 0.01
Fold Change: 1,5
Tops: Not Use
P Value: 0.01 (0.01 / 0.05 / not-use)

New Analysis Tab: When the process of defining the analysis is being completed, the interface looks like this.

3.1. The “Define Expression Data” Dialog

First, you need to click on “Open File”. This will bring up the “Define Expression Data” dialog:



Count reads: C:\Users\pentalpha\Documents\data\pasilla_gene_counts.tsv

Conditions (optional): C:\Users\pentalpha\Documents\data\pasilla_gene_counts.tsv.conditions

Adjust conditions of each sample (drag and drop):

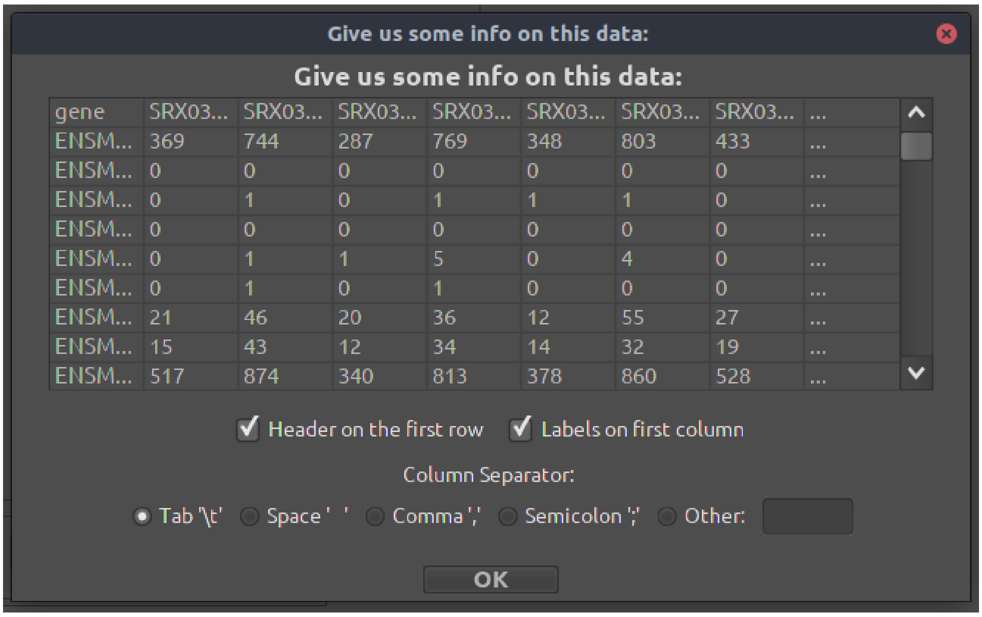
condition0	condition1	not-use
0 - untreated1	4 - treated1	
1 - untreated2	5 - treated2	
2 - untreated3	6 - treated3	
3 - untreated4		

Integer rounding: HALF_UP

Count reads threshold: 0 1 2 3 4 5 6 7 8 9 10

Define Expression Dialog: The full interface including sample groups and box plot. The a, b, c and d highlighted regions are explained below.

a. First, click on the first button at the top of the dialog to open the count reads file. After you have chosen a text file, R-Peridot shows the following dialog to get some information about the file:



Give us some info on this data:

Give us some info on this data:

gene	SRX03...	SRX03...	SRX03...	SRX03...	SRX03...	SRX03...	SRX03...	...
ENSM...	369	744	287	769	348	803	433	...
ENSM...	0	0	0	0	0	0	0	...
ENSM...	0	1	0	1	1	1	0	...
ENSM...	0	0	0	0	0	0	0	...
ENSM...	0	1	1	5	0	4	0	...
ENSM...	0	1	0	1	0	0	0	...
ENSM...	21	46	20	36	12	55	27	...
ENSM...	15	43	12	34	14	32	19	...
ENSM...	517	874	340	813	378	860	528	...

☑ Header on the first row ☑ Labels on first column

Column Separator:

● Tab '\t' ○ Space ' ' ○ Comma ',' ○ Semicolon ';' ○ Other:

There is a short version of the original file and some questions are asked: Are there headers on the first row? Is the first column made of labels? What is the column separator character?

Once all the information is correct and the visualization of the data makes sense, click on the "OK" button.

b. Once a count reads file has been chosen, R-Peridot will display a box plot of the counts in each sample (column) of the file. This is useful to determine which samples you would like to use on the analysis.

c. The next step is to organize the samples in groups (conditions). Create new conditions using the "Add New Condition" button. Move samples between conditions using drag and drop. If there are many samples, try selecting many conditions holding "SHIFT" or "CTRL".

The samples inside the "not-use" condition will be ignored.

To erase a condition, just leave it empty.

d. The differential expression packages expect integer counts by default, so R-Peridot rounds the non-integer values and you can choose the rounding method.

The RNA-Seq technologies are very precise, but the data may have some noise. Some packages may take values who are merely noise in consideration while looking for differentially expressed genes (DEGs). That is why we created a threshold selector for the reads: read values below the threshold are discarded. To improve performance, lines only with 0's and discarded values are discarded. The default threshold value is 1, that means by default no threshold is being used.

Once you have organized the samples in their respective conditions and everything is ready, click on the "Create" button. That will bring you back to the "New Analysis" tab.

3.2. Choosing the Modules

Now that you have defined the input, the desired modules can be selected. Please note that depending on your input, some analysis modules may be disabled. Most of them do not support more than two conditions or conditions with only one sample (no replicates).

To create a better consensus of the results, with less false positives, we recommend using at least 3 analysis modules.

To select post analysis modules, at least one analysis module has to be selected.

The "*VennDiagram"* module is permanently selected, because it is the one responsible for creating the consensus of all results from analysis modules.

Use the "Detail…" button to read about the inputs, results, parameters and the description of each module. You can also read about them in Chapter 8: Default Modules.

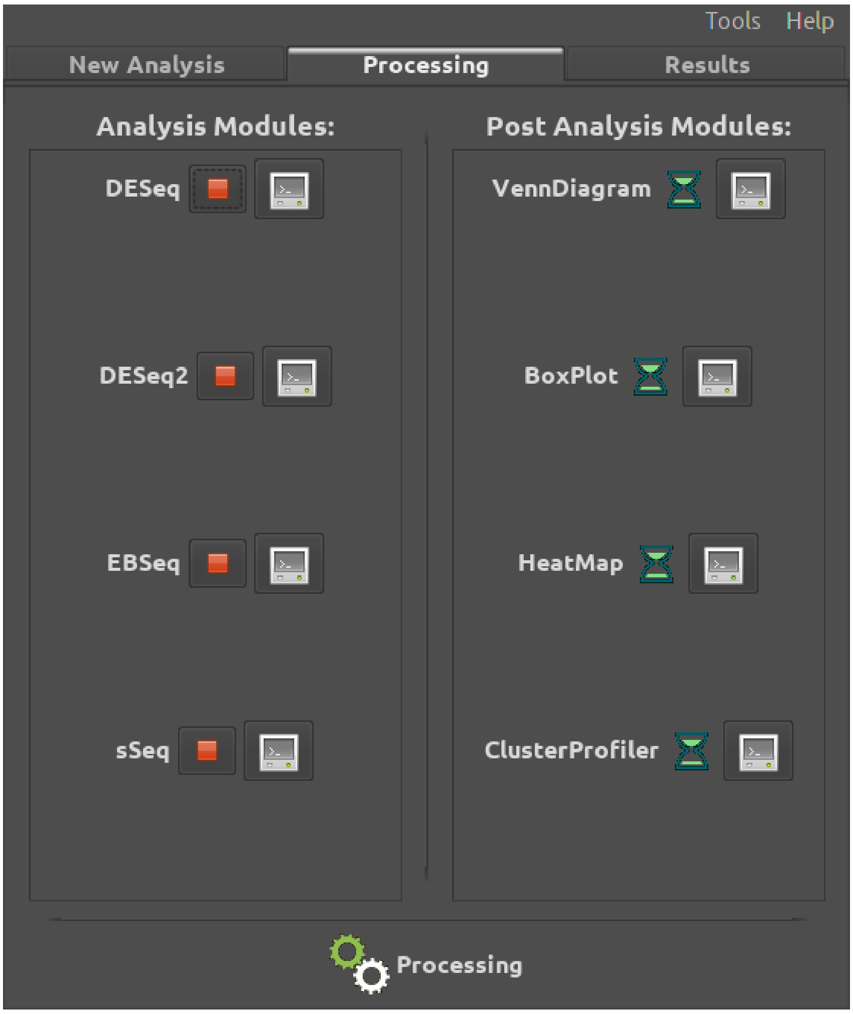
Between the name of the module and the "Detail…" button, there is a small button with a notebook-and-pencil icon. Use it to specify, for a module, different parameters from the ones chosen in the next step.

3.3. Setting Parameters

The parameters are values passed from R-Peridot to the modules. All of them have default values and you do not have to necessarily change them. But, if you plan to use the "*ClusterProfiler"* module, don't forget to choose an Id Type and the correct Reference Organism (only Human, Mouse and Fly are currently supported). Read more about each parameter in Chapter 7: Parameters Explained.

If you have selected the modules you want and the parameters are adequate for your analysis, click on the "Start" button to run the analysis. This will bring you to the "Processing" tab.

4. “Processing” Tab



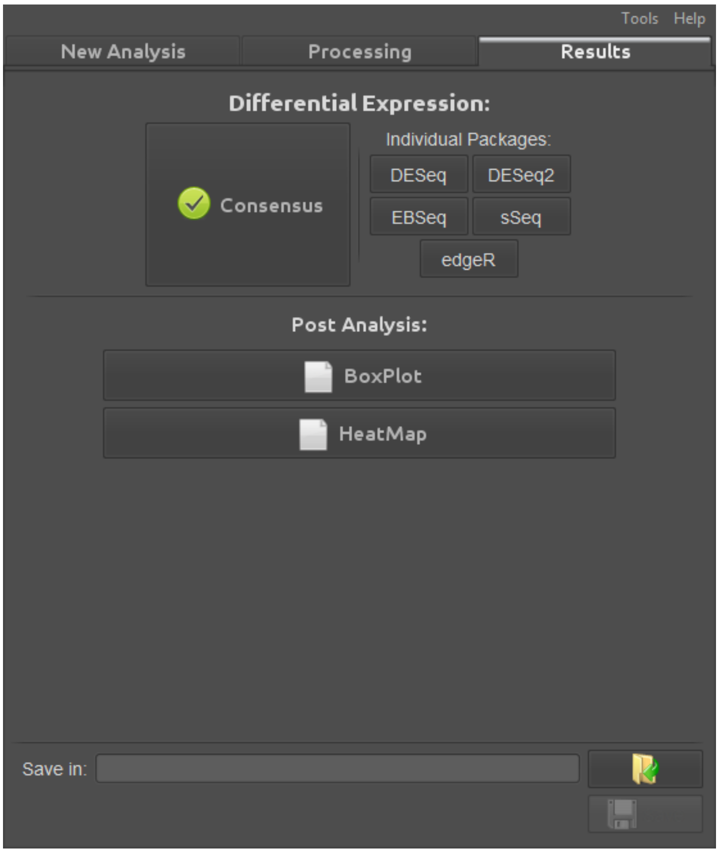
Processing Tab: This interface was made so that the user can watch the progress of each module.

Here, you have all the selected modules listed. Remember that the modules have dependencies between each other, so some of them will wait (hourglass icon) for others to finish before they can start. The post analysis modules, for instance, always wait for the analysis modules to finish. Once a module starts, the hourglass icon is replaced by a red stop button. If you click on that button, the module will stop (and probably fail too). You can also click on the console icon to read the command line output of each module.

The success of a module is determined by the results it produces. If all the mandatory results are created, R-Peridot will consider it as successful. If the opposite happens, the module will have failed. If a module fails, its results won’t appear on the “Results” tab and all the modules that depend on it will fail too.

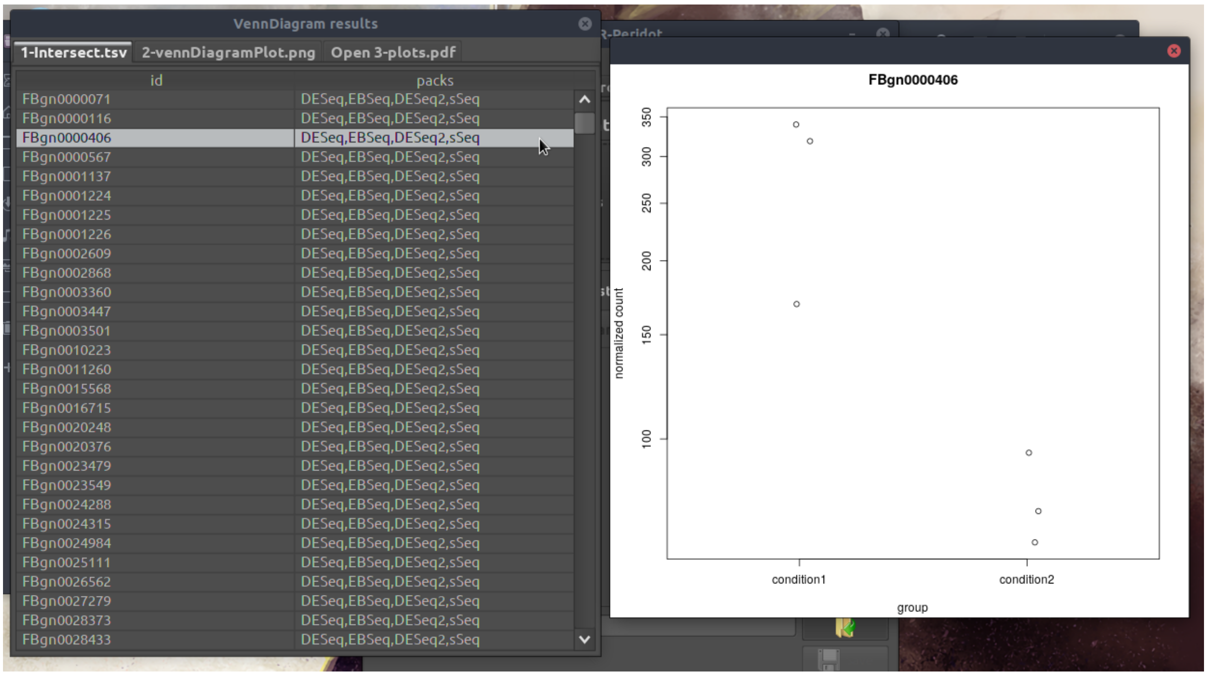
Once all modules have finished being executed, the interface will switch to the “Results” tab.

5. “Results” Tab



Results Tab: As the name suggests, here you can see the results of all successful modules.

At the beginning of this tab, you can access the differential expression analysis results. There is the “Consensus” (resulting from the “VennDiagram” module) and the individual results of each analysis module. After clicking on one of these items, a dialog with a tab for each result will open, including graphics and tables with Differentially Expressed Genes (DEGs).



id	packs
FBgn0000071	DESeq,EBSeq,DESeq2,sSeq
FBgn0000116	DESeq,EBSeq,DESeq2,sSeq
FBgn0000406	DESeq,EBSeq,DESeq2,sSeq
FBgn0000567	DESeq,EBSeq,DESeq2,sSeq
FBgn0001137	DESeq,EBSeq,DESeq2,sSeq
FBgn0001224	DESeq,EBSeq,DESeq2,sSeq
FBgn0001225	DESeq,EBSeq,DESeq2,sSeq
FBgn0001226	DESeq,EBSeq,DESeq2,sSeq
FBgn0002609	DESeq,EBSeq,DESeq2,sSeq
FBgn0002868	DESeq,EBSeq,DESeq2,sSeq
FBgn0003360	DESeq,EBSeq,DESeq2,sSeq
FBgn0003447	DESeq,EBSeq,DESeq2,sSeq
FBgn0003501	DESeq,EBSeq,DESeq2,sSeq
FBgn0010223	DESeq,EBSeq,DESeq2,sSeq
FBgn0011260	DESeq,EBSeq,DESeq2,sSeq
FBgn0015568	DESeq,EBSeq,DESeq2,sSeq
FBgn0016715	DESeq,EBSeq,DESeq2,sSeq
FBgn0020248	DESeq,EBSeq,DESeq2,sSeq
FBgn0020376	DESeq,EBSeq,DESeq2,sSeq
FBgn0023479	DESeq,EBSeq,DESeq2,sSeq
FBgn0023549	DESeq,EBSeq,DESeq2,sSeq
FBgn0024288	DESeq,EBSeq,DESeq2,sSeq
FBgn0024315	DESeq,EBSeq,DESeq2,sSeq
FBgn0024984	DESeq,EBSeq,DESeq2,sSeq
FBgn0025111	DESeq,EBSeq,DESeq2,sSeq
FBgn0026562	DESeq,EBSeq,DESeq2,sSeq
FBgn0027279	DESeq,EBSeq,DESeq2,sSeq
FBgn0028373	DESeq,EBSeq,DESeq2,sSeq
FBgn0028433	DESeq,EBSeq,DESeq2,sSeq

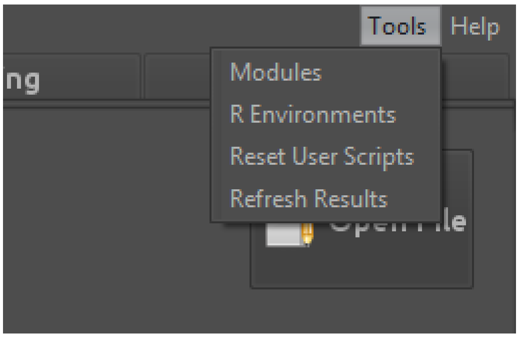
The counts plot of a result: If you open one of the tables with a DEG per line and click twice on one of these lines, a plot with the counts of this gene across different conditions will appear.

Following the differential expression results, there are buttons for the results of each post analysis module. Click on one of them to open the results.

These results will only remain stored until you make a new analysis. To save the results, use the “Save in” option at the bottom of the tab, choosing a directory to save the results. If the directory you chose is not empty, R-Peridot will save the results in a new directory named “<directory_you_chose>0”, then “<directory_you_chose>1 and so on.

6. Tools Menu

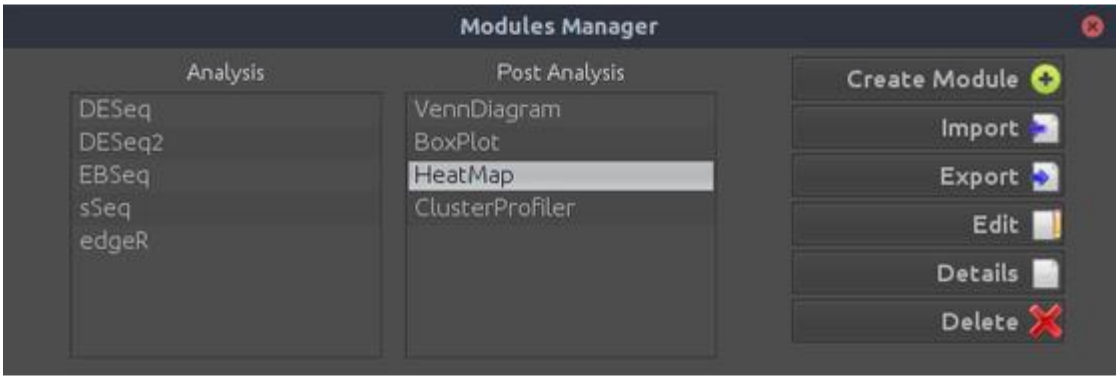
The “Tools” menu is the first at the top right corner of the GUI. It includes some tools to manage the modules, R environments and results.



Tools Menu: The opened menu displaying its 4 items.

6.1. Modules

This option opens the “Modules Manager”, the interface to manage all the modules.

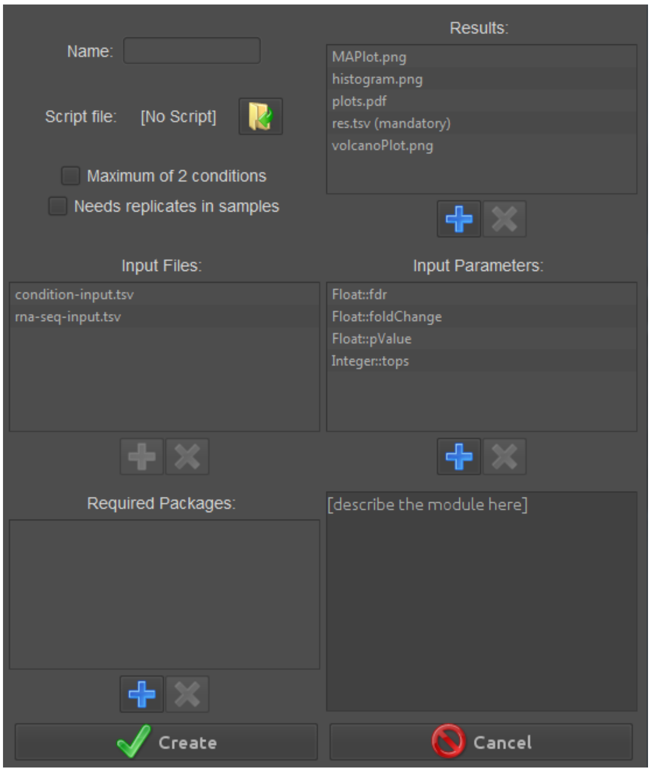


Modules Manager: Things you can do here include creating, removing, importing, exporting and editing modules.

There are several buttons by the right side of the interface, each doing one of these operations:

- **Create Module**: It will open a dialog window through which you can create new modules. First, R-Peridot will ask you if you desire to create an Analysis Module or a Post Analysis Module. If you choose to create the first type, most of the fields will be completed with default values.

Module Creator: To define a new module, you must give it a name, a script file, a description and lists of results, input files (can be results from other modules), parameters and required packages. For more detailed instructions on how to write a R script for R-Peridot, read the *Guide for Advanced Users.*



Name:

Script file: [No Script]

☐ Maximum of 2 conditions

☐ Needs replicates in samples

Results:

- MAPlot.png
- histogram.png
- plots.pdf
- res.tsv (mandatory)
- volcanoPlot.png

Input Files:

- condition-input.tsv
- rna-seq-input.tsv

Input Parameters:

- Float::fdr
- Float::foldChange
- Float::pValue
- Integer::tops

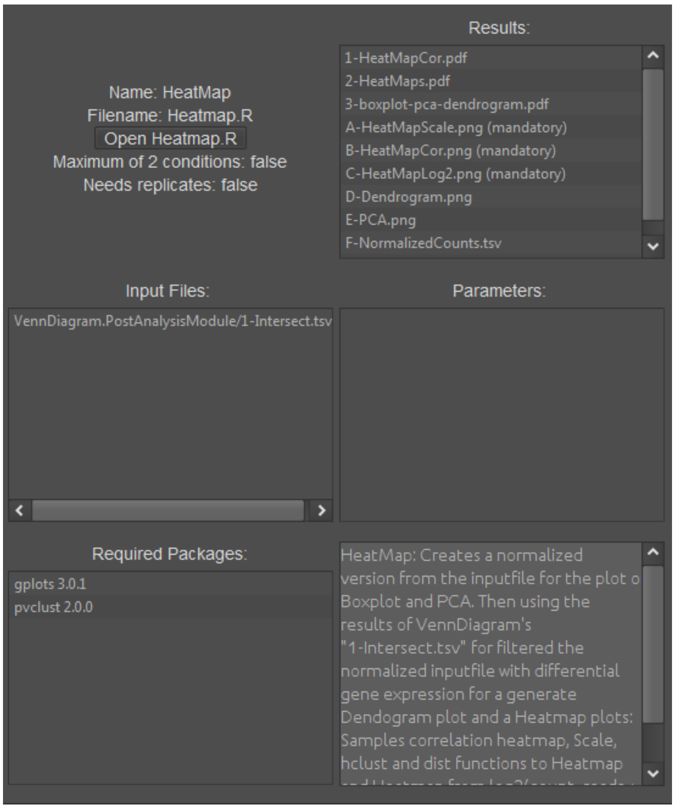
Required Packages:

[describe the module here]

- **Import Module:** This will open a file chooser dialog. Select a file with the ".PeridotModule" extension. The module will be imported to R-Peridot and will be available to be used after you restart the application.

To use the following operations, first select a module on the lists at the middle and left side of the "Modules Manager":

- **Export Module:** This will open a file chooser dialog to select a directory. A file named "<module name>.PeridotModule" will be created inside it. You can later *Import* this file to R-Peridot, to use the module. This file can be shared, so that other people can use the module you just exported.
- **Edit:** Opens a window similar to the "Module Creator", but to edit an existing module.
- **Details:** Displays all the information regarding a module.



Name: HeatMap
Filename: Heatmap.R

Maximum of 2 conditions: false
Needs replicates: false

Results:

1-HeatMapCor.pdf
2-HeatMaps.pdf
3-boxplot-pca-dendrogram.pdf
A-HeatMapScale.png (mandatory)
B-HeatMapCor.png (mandatory)
C-HeatMapLog2.png (mandatory)
D-Dendrogram.png
E-PCA.png
F-NormalizedCounts.tsv

Input Files:

VennDiagram.PostAnalysisModule/1-Intersect.tsv

Parameters:

Required Packages:

gplots 3.0.1
pvclust 2.0.0

HeatMap: Creates a normalized version from the inputfile for the plot o Boxplot and PCA. Then using the results of VennDiagram's "1-Intersect.tsv" for filtered the normalized inputfile with differential gene expression for a generate Dendogram plot and a Heatmap plots: Samples correlation heatmap, Scale, hclust and dist functions to Heatmap

Module Details: This dialog window shows all the details of a module. With the "Open <module name>" button, you can open the R script of the module.

- **Delete:** Deletes a module and then closes R-Peridot. Remember that you can't delete default modules: when R-Peridot restarts, they will be installed again.

6.2. R Environments

With this option you can open the "R Environment Manager", previously discussed in Chapter 1. Please note that any modifications done here will restart the "New Analysis" tab.

6.3. Reset User Scripts

This option will reset the modules to their initial state. Any modifications to modules will be lost and new modules will be erased. This is particularly useful in case you did something wrong while playing with the scripts of the modules.

6.4. Refresh Results

This option forces the GUI to reload the results. This is useful in case you made any modification to one of the result files.

7. Parameters Explained

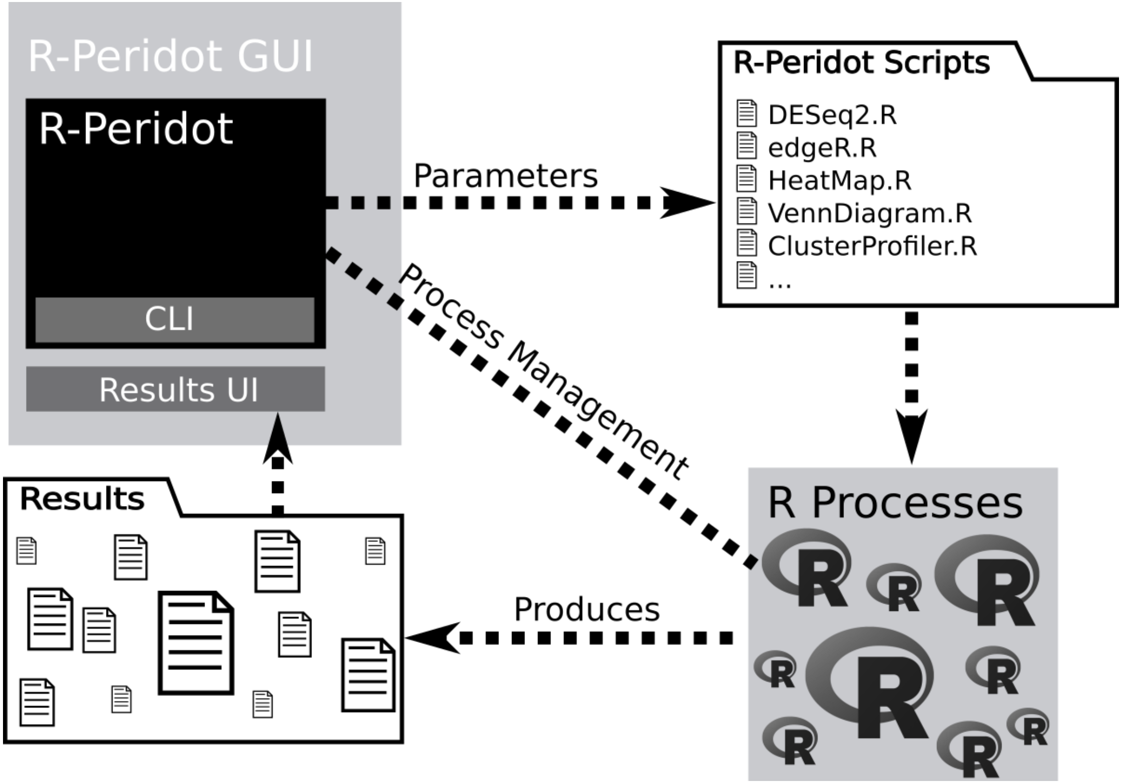
R-Peridot's default modules use a few parameters. You may want to tweak their default values to improve your analysis, so here is an explanation of their meaning:

P Value	It is the estimated probability of rejecting the null hypothesis, when that hypothesis is true. **A smaller P Value lowers the chances of finding false positives on the results**.
FDR (False Discovery Rate)	If you repeat a test enough times, you're going to find an effect, but that effect may not actually exist. The FDR is used to control the rate of false discoveries when there are many tests being done. Putting it simply, the FDR is the number of false discoveries in an experiment divided by total number of discoveries in that experiment. **The smaller the FDR is, the smaller is the number of false discoveries**.
Fold Change	Given two conditions (A and B), the fold change is the expected reason between the reads on B and A, so that one of them can be considered differential. For example, a change from 30 to 60 is defined as a fold change of 2. Smaller fold change values will allow smaller differences between RNA read counts to be considered differential, and greater values will do the opposite.
Tops	**Used to create, for each analysis module, a "Top Results" file with a specific number of DEGs**. If set to 0, no "Top Results" file is created.
Id Type	**The type of ID on the labels column**. IDs supported are: None (no ID), kegg, SYMBOL, ACCNUM, ALIAS, ENSEMBL, ENSEMBLPROT, ENSEMBLTRANS, ENTREZID, ENZYME, EVIDENCE, EVIDENCEALL, GENENAME, GO, GOALL, IPI, MAP, OMIM, ONTOLOGY, ONTOLOGYALL, PATH, PFAM, PMID, PROSITE, REFSEQ, UCSCKG, UNIGENE and UNIPROT.
Reference Organism	**The organism to use when creating the "*ClusterProfiler*" charts**. There are 3 options currently available: Human (Homo Sapiens), Mouse (Mus Musculus) and Fly (Drosophila_melanogaster).

8. Default Modules

Modules are the abstraction that the R-Peridot project applies to R scripts, so that the user does not have to deal with details of writing and running these scripts. A module is basically a R script associated with a specification of its inputs (files/parameters), outputs (files), required packages and several more details.

There are 2 different types of modules: analysis and post analysis. Analysis modules have a well-defined role, they receive as input the gene expression data and use them to find differentially expressed genes. But the post analysis modules have a freer role. These are necessarily executed after the analysis and can use both the results of the analysis modules and the other post analysis modules to generate any type of result.



R-Peridot Module Management Diagram: R-Peridot passes input parameters to the R scripts and manages their processes, retrieving the results when they are ready.

8.1. Analysis Modules

As stated before, these are the modules that do the differential gene expression analysis. They are all different software, using different statistical methodologies and producing different sets of results:

Name	Package Version	Normalization	Support For More Than 2 Conditions	Allows Conditions Without Replicates*
DESeq	1.28.0	DESeq size factors	No	No
DESeq2	1.16.1	DESeq size factors	No	No
EBSeq	1.16.0	DESeq median	No	Yes

		normalization		
edgeR	3.18.1	TMM / Upper quartile / RLE / None (all scaling factors are set to be one)	No	No
sSeq	1.14.0	DESeq size factors	Yes	Yes

*doing differential gene expression analysis on conditions without replicates is not recommended, because the results may not be statistically reliable. This is why most packages don't allow this;

All of these modules work with default sets of inputs and outputs:

Parameters	FDR, Fold Change, P-Value, Tops.
Inputs	Expression count table and sample conditions.
Results	DEGs Table (mandatory), Histogram, MA Plot and Volcano Plot.

8.2. Post Analysis Modules

These are modules that produce various types of results and most of them are optional.

8.2.1. VennDiagram – The Consensus

This is the only mandatory module. It reads the results from all the analysis modules and then finds the consensus between them. It also creates, for each gene on the results of at least one analysis module, a plot of the count reads along the different conditions. The results of this module are displayed on the "Consensus" button at the results tab.

Parameters	None.
Inputs	The sets of results from each analysis module.
Results	Table describing, for each DEG, which packages selected them (Intersect.tsv), count reads plots for each DEG and also a Venn Diagram of the sets of DEGs.

8.2.2. HeatMap

This module creates several graphics, mainly heat maps.

Parameters	None.
Inputs	VennDiagram's Intersection.tsv
Results	The following three heat maps are created: • "A-HeatMapScale.png": Heat map with values of reads normalized by subtracting their average and dividing by the standard deviation. • "B-HeatMapCor.png": Correlation between each sample. • "C-HeatMapLog2.png": Heat map with values of $\log_2$(reads + 0.99*); Other results are: An alternative version of the count reads table with normalized values, a Dendrogram, a PCA and a Box Plot of the normalized counts.

*arbitrary constant;

8.2.3. Box Plot

A Box Plot of the counts on each sample. This is different from the Box Plot presented after selecting the input file, because it shows the counts used as input to the modules, without the table rows filtered by threshold and the samples defined as “not-use”.

Parameters	None.
Inputs	Expression count table.
Results	Box Plot.

8.2.3. ClusterProfiler

This package implements methods to analyze and visualize functional profiles or gene clusters. It produces charts for *Cellular Component* (CC), *Molecular Function* (MF), *Biological Process* (BP) and KEGG Pathways.

Parameters	FDR, P-Value, Reference Organism and ID Type.
Inputs	VennDiagrams’s Intersect.tsv.
Results	Charts for CC, MF, BP and KEGG.