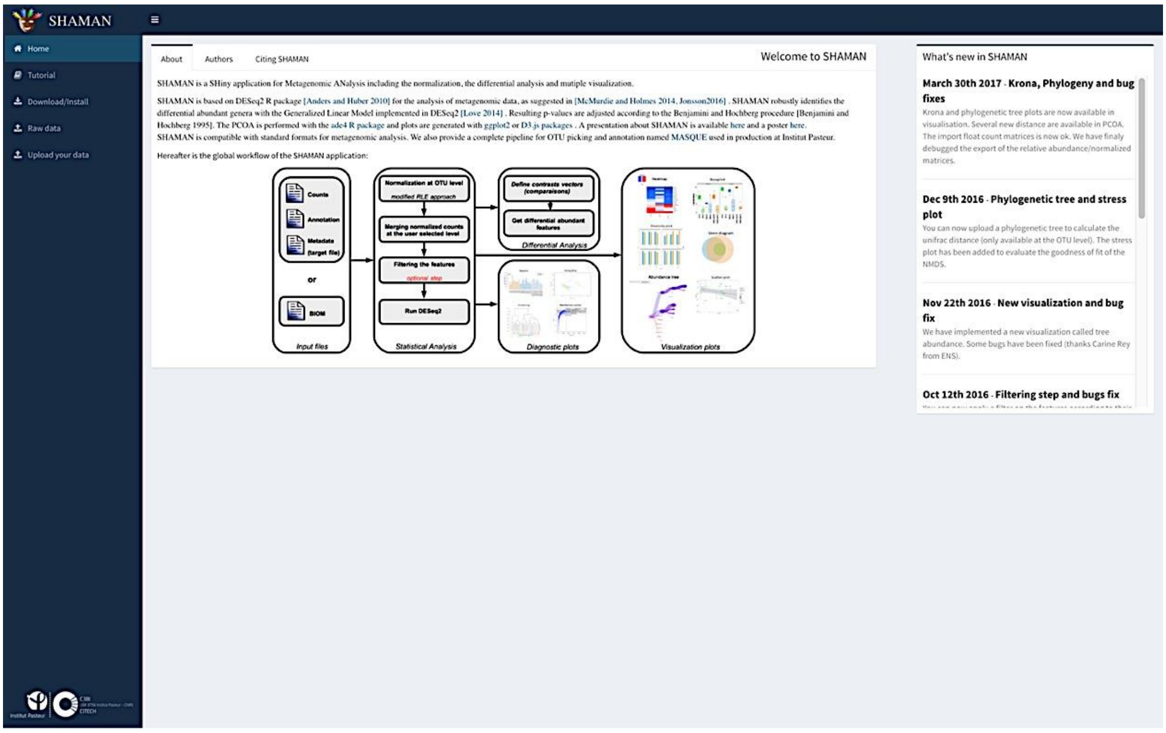
SHAMAN USER'S GUIDE

SHiny Application for Metagenomic ANalysis

14/08/2019



Welcome to SHAMAN

SHAMAN is a SHiny application for Metagenomic ANalysis including the normalization, the differential analysis and mutiple visualization.

SHAMAN is based on DESeq2 R package [Anders and Huber 2010] for the analysis of metagenomic data, as suggested in [McMurdie and Holmes 2014, Jonsson2016] . SHAMAN robustly identifies the differential abundant genera with the Generalized Linear Model implemented in DESeq2 [Love 2014] . Resulting p-values are adjusted according to the Benjamini and Hochberg procedure [Benjamini and Hochberg 1995]. The PCOA is performed with the ade4 R package and plots are generated with ggplot2 or D3.js packages . A presentation about SHAMAN is available here and a poster here.
SHAMAN is compatible with standard formats for metagenomic analysis. We also provide a complete pipeline for OTU picking and annotation named MASQUE used in production at Institut Pasteur.

Hereafter is the global workflow of the SHAMAN application:

What's new in SHAMAN

March 30th 2017 - Krona, Phylogeny and bug fixes
Krona and phylogenetic tree plots are now available in visualisation. Several new distance are available in PCOA. The import float count matrices is now ok. We have finaly debugged the export of the relative abundance/normalized matrices.

Dec 9th 2016 - Phylogenetic tree and stress plot
You can now upload a phylogenetic tree to calculate the unifrac distance (only available at the OTU level). The stress plot has been added to evaluate the goodness of fit of the NMDS.

Nov 22th 2016 - New visualization and bug fix
We have implemented a new visualization called tree abundance. Some bugs have been fixed (thanks Carine Rey from ENS).

Oct 12th 2016 - Filtering step and bugs fix

shaman.pasteur.fr

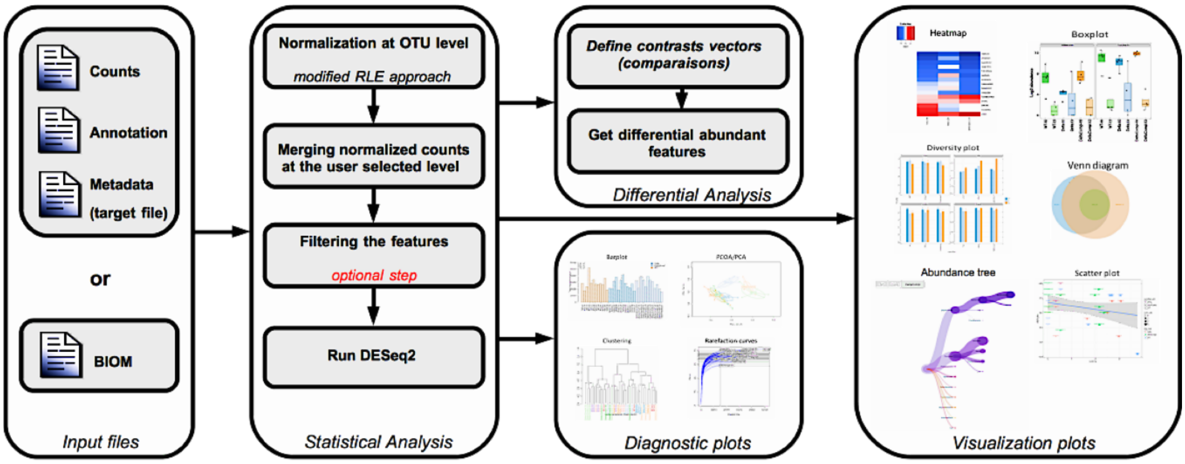
Authors: Stevenn Volant, Amine Ghozlane, Perrine Woringer

SUMMARY

INTRODUCTION

SHAMAN is a shiny application for differential analysis of metagenomic data (16S, 18S, 23S, 28S, ITS and WGS) including bioinformatics treatment of raw reads for targeted metagenomics, statistical analysis and results visualization with a large variety of plots (barplot, boxplot, heatmap, …).
The statistical analysis performed by SHAMAN is based on DESeq2 R package [Anders and Huber 2010] which robustly identifies the differential abundant features as suggested in [McMurdie and Holmes 2014, Jonsson 2016].

SHAMAN is compatible with standard formats for metagenomic analysis (.csv, .tsv, .biom) and generated figures can be downloaded in several formats. Hereafter is the global workflow of the SHAMAN application:



Quick installation instructions

SHAMAN is available for local installation using Docker and R. This installation covers only the statistical analysis. The bioinformatics treatment is deported to the Institut Pasteur galaxy instance for performance reason.

- Docker install

Docker is the easiest way to use SHAMAN locally. It is a controlled virtual environment where every package required for SHAMAN are already installed and which have no impact on your local R installation.
First, download and install Docker from https://www.docker.com/. Docker is available for Windows, Mac and Linux.

Run:

```
# Download shaman
docker pull aghozlane/shaman
# Execute shaman, port 80 and 5438 need to be available
docker run --rm -p 80:80 -p 5438:5438 aghozlane/shaman
```

Then, connect with your web browser to: http://0.0.0.0/ or http://localhost/

If port 80 is already allocated, run:

```
docker run --rm -p 3838:80 -p 5438:5438 aghozlane/shaman
```

Then connect to http://0.0.0.0:3838/ or http://localhost:3838/ .

You can update your local version of SHAMAN with:

```
docker pull aghozlane/shaman
```

- R install with Packrat

SHAMAN is available for R=3.1.2. Packrat framework installation allow an easy installation of all the dependencies. Of note, raw data submission is not possible with this version. First, install R 3.1.2 as local install as follow:

```
# Install R 3.1.2
wget https://pbil.univ-lyon1.fr/CRAN/src/base/R-3/R-3.1.2.tar.gz && tar -
zxf R-3.1.2.tar.gz
mkdir /some/location/r_bin
cd R-3.1.2/ && ./configure --prefix=/some/location/r_bin/ && make && make
install && /some/location/r_bin/bin/R
# Download SHAMAN package
wget ftp://shiny01.hosting.pasteur.fr/pub/shaman_package.tar.gz
```

This installation will not interact with other R installation. Then, you can install shaman with packrat:

```
# Install SHAMAN dependencies
mkdir /some/location/shaman
/some/location/r_bin/bin/R
install.packages(c('devtools', 'codetools', 'lattice', 'MASS', 'survival', 'packrat'))
library(devtools)
devtools::install_github(c('aghozlane/nlme'))
packrat::unbundle("shaman_package.tar.gz", "/packrat/location/shaman")
```

Now you can run SHAMAN:

```
library(packrat)
packrat::init("/packrat/location/shaman")
library(shiny)
system("Rscript -e
'library(\"shiny\");runGitHub(\"pierreLec/KronaRShy\",port=5438)'",wait=
FALSE)
runGitHub('aghozlane/shaman')
```

- R install (deprecated)

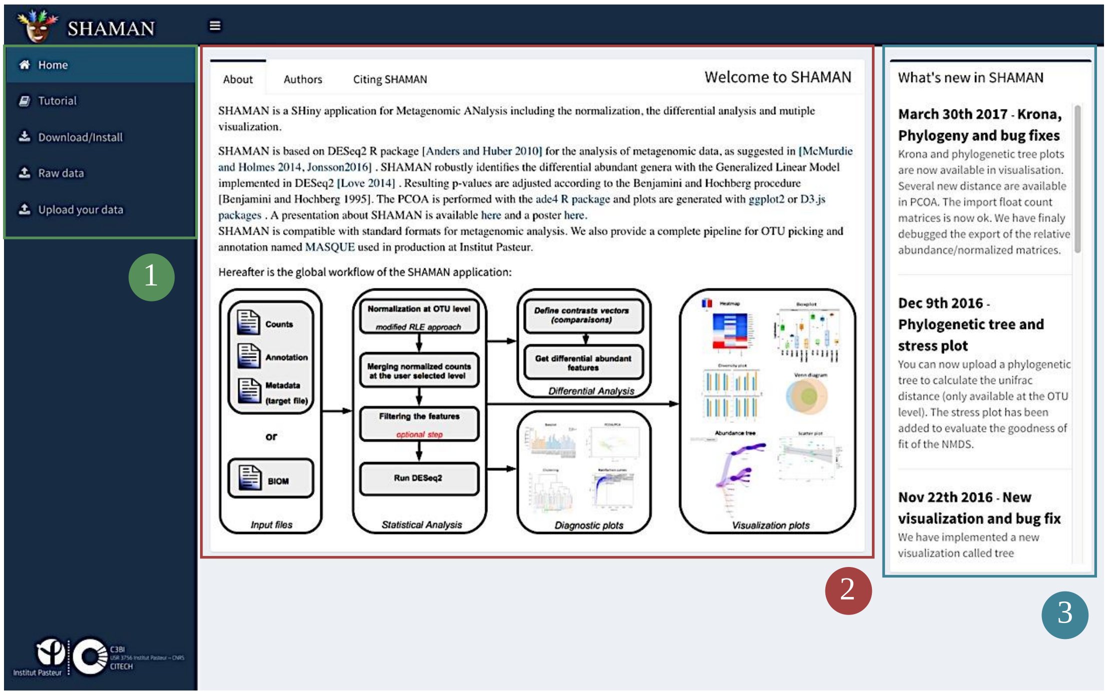
SHAMAN is available for R=3.1.2. The installation, download and execution can all be performed with a small R script:

```
# Load shiny packages
if(!require('shiny')){
        install.packages('shiny')
        library(shiny)
}
system("Rscript –e library("shiny");
runGitHub("pierreLec/KronaRShy", port=5438)", wait=FALSE)
# Install dependencies, download last version from github,
# and run SHAMAN in one command:
runGitHub('aghozlane/shaman')
```

Of note, the R version has an impact on the contrast definition. For R>3.2, DESeq2 used non-expanded modeling, hence the creation of contrast vectors slightly differs and some SHAMAN features might be deactivated.

Homepage

SHAMAN is available at http://shaman.pasteur.fr/. Hereafter is the homepage:



Welcome to SHAMAN

SHAMAN is a SHiny application for Metagenomic ANalysis including the normalization, the differential analysis and mutiple visualization.

SHAMAN is based on DESeq2 R package [Anders and Huber 2010] for the analysis of metagenomic data, as suggested in [McMurdie and Holmes 2014, Jonsson2016] . SHAMAN robustly identifies the differential abundant genera with the Generalized Linear Model implemented in DESeq2 [Love 2014] . Resulting p-values are adjusted according to the Benjamini and Hochberg procedure [Benjamini and Hochberg 1995]. The PCOA is performed with the ade4 R package and plots are generated with ggplot2 or D3.js packages . A presentation about SHAMAN is available here and a poster here.
SHAMAN is compatible with standard formats for metagenomic analysis. We also provide a complete pipeline for OTU picking and annotation named MASQUE used in production at Institut Pasteur.

Hereafter is the global workflow of the SHAMAN application:

What's new in SHAMAN

March 30th 2017 - **Krona, Phylogeny and bug fixes**
Krona and phylogenetic tree plots are now available in visualisation. Several new distance are available in PCOA. The import float count matrices is now ok. We have finaly debugged the export of the relative abundance/normalized matrices.

Dec 9th 2016 - **Phylogenetic tree and stress plot**
You can now upload a phylogenetic tree to calculate the unifrac distance (only available at the OTU level). The stress plot has been added to evaluate the goodness of fit of the NMDS.

Nov 22th 2016 - **New visualization and bug fix**
We have implemented a new visualization called tree

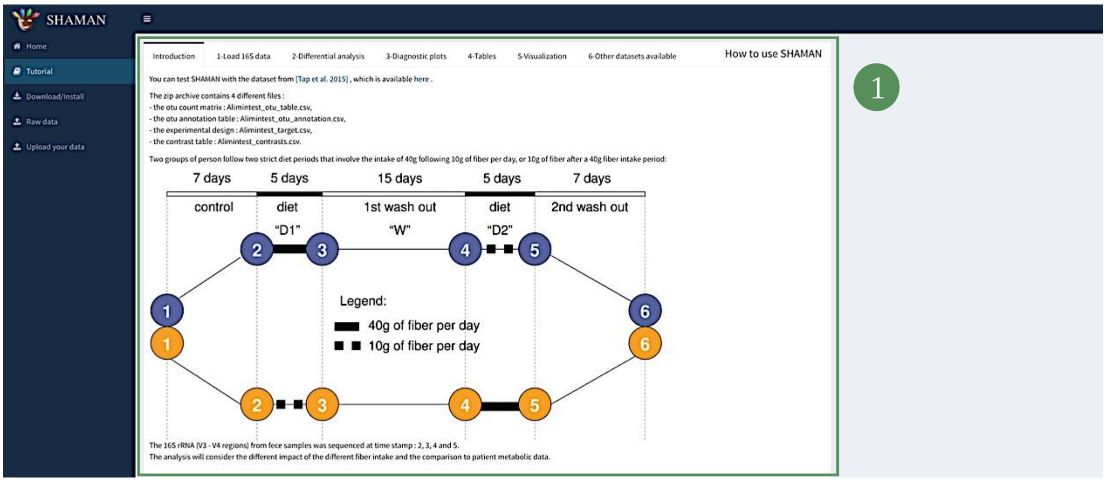
1 Toolbar. 5 tab are available: Home, Tutorial, Download/Install, Raw data, Upload your data.

2 Description of the application

3 SHAMAN news.

Tutorial and Download

« Tutorial » and « Download/Install » panels are:

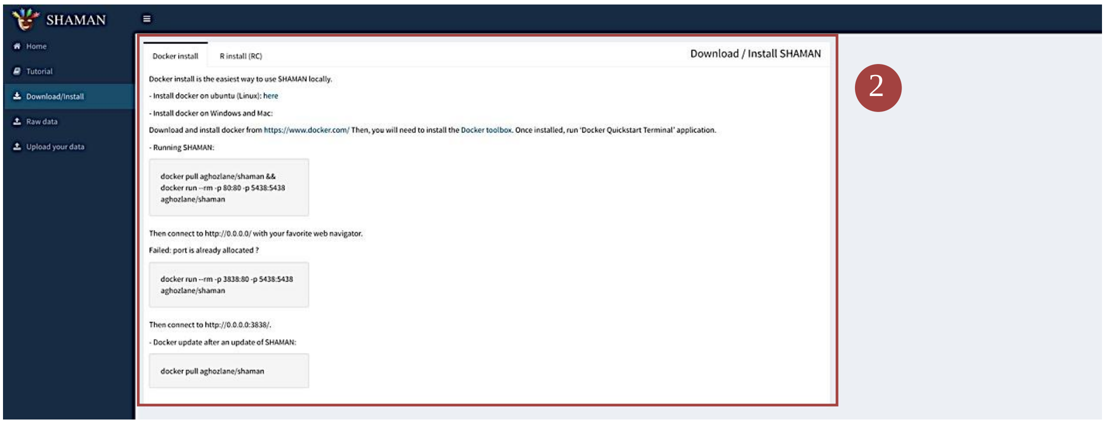


You can test SHAMAN with the dataset from [Tap et al. 2015] , which is available here .

The zip archive contains 4 different files :
- the otu count matrix : Alimintest_otu_table.csv,
- the otu annotation table : Alimintest_otu_annotation.csv,
- the experimental design : Alimintest_target.csv,
- the contrast table : Alimintest_contrasts.csv.

Two groups of person follow two strict diet periods that involve the intake of 40g following 10g of fiber per day, or 10g of fiber after a 40g fiber intake period:

The 16S rRNA (V3 - V4 regions) from fece samples was sequenced at time stamp : 2, 3, 4 and 5.
The analysis will consider the different impact of the different fiber intake and the comparison to patient metabolic data.



Docker install is the easiest way to use SHAMAN locally.

- Install docker on ubuntu (Linux): here

- Install docker on Windows and Mac:

Download and install docker from https://www.docker.com/ Then, you will need to install the Docker toolbox. Once installed, run 'Docker Quickstart Terminal' application.

- Running SHAMAN:

```
docker pull aghozlane/shaman &&
docker run --rm -p 80:80 -p 5438:5438
aghozlane/shaman
```

Then connect to http://0.0.0.0/ with your favorite web navigator.

Failed: port is already allocated ?

```
docker run --rm -p 3838:80 -p 5438:5438
aghozlane/shaman
```

Then connect to http://0.0.0.0:3838/.

- Docker update after an update of SHAMAN:

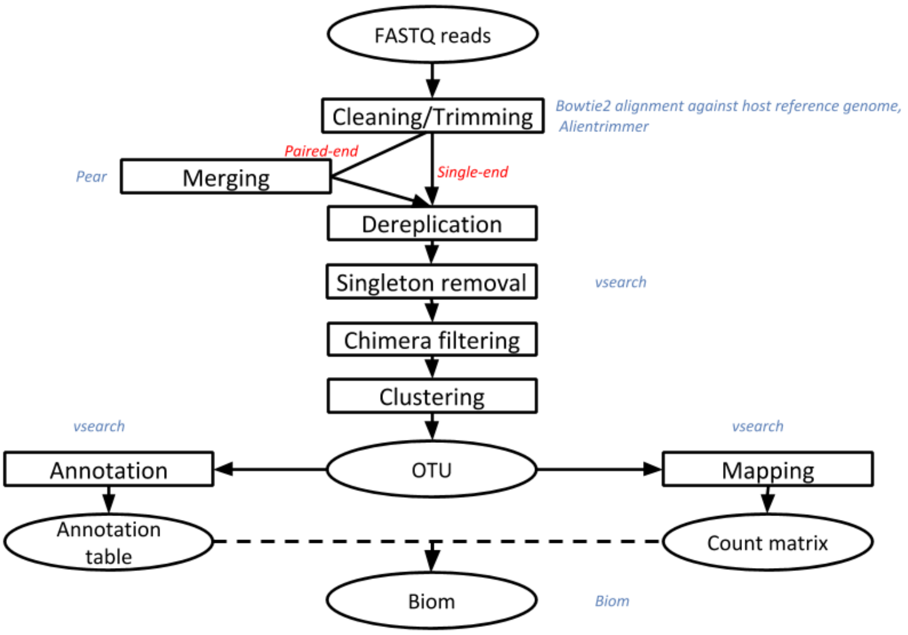
```
docker pull aghozlane/shaman
```

1 Online tutorial: it provides a description of SHAMAN usage for a set of 16S data. The dataset used is provided for download.

2 Installation guide. It is possible to install shaman with docker and R.

Raw data

SHAMAN provide access to a bioinformatics workflow for analyzing targeted metagenomics data. This workflow is based on *de novo* clustering. Operational Taxonomic Unit (OTU) are built from reads in a given experiment and annotated by alignment against reference databases. The aim is to identify among high quality amplicon, OTU sequences that will be representative of one species and considered for annotation and quantification. The workflow can be summaries as follow:

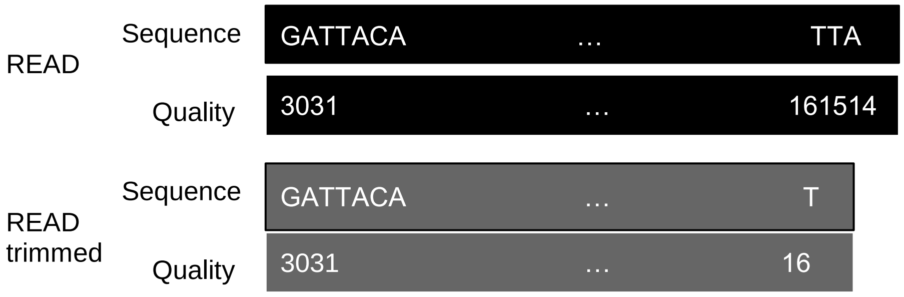


Read preparation

Cleaning

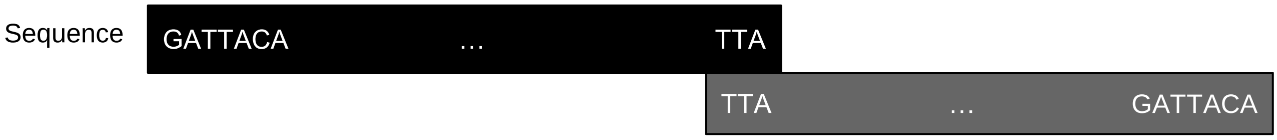
Cleaning step consists in the alignment of reads against the host and PhiX174 reference genome. Reads that did not align are considered for further analysis. This step is facultative but recommended. Very few amount of host DNA or Phi phage (used for calibration in every Illumina sequencing) can be identified in samples. A bowtie2 alignment with parameters (--sensitive) allows to eliminate these reads.

Trimming



The trimming step consists to remove nucleotides sequences (adaptors, primers and non-confident nucleotides) in both 5’ and 3’ read ends. A list of every adaptators and primer used in Illumina, Solid, Ion torrent, Truseq, and Nextera adaptators is already included in SHAMAN. But user can specify his own adaptaters. This step is performed by Alientrimmer [Criscuolo 2013].

Merging

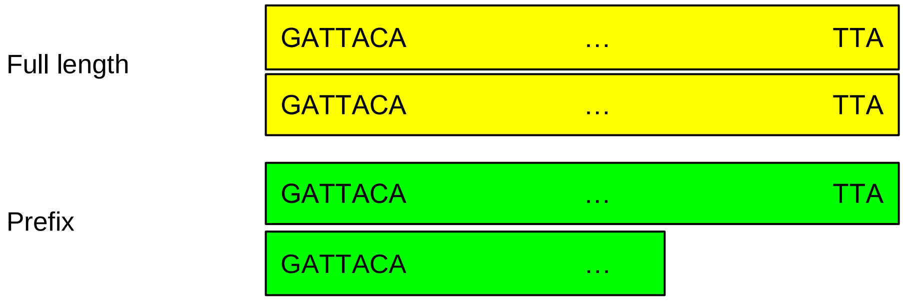


When reads are paired, a merging step is performed using Pear [Zhang 2013]. Reads are expected to overlap a given area of the ribosomal RNA. Sequence obtained are called amplicons.

OTU processing

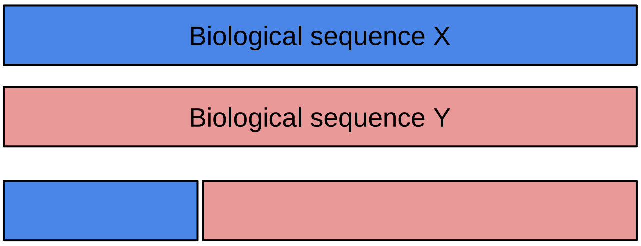
The OTU clustering step is performed with Vsearch [Rognes 2016].

Dereplication



The dereplication consists in selecting one representative sequence among identical amplicon sequence. Two approaches are available: the full length and the prefix dereplication. Full length approach will group sequence that are identical for their entire length. Prefix will also group sequence of different length. Only the longest sequence will be kept. The number of identical sequences in a given sample is critical. Sequences are re-ordered by this “dereplicated abundance” which will drive the OTU clustering (see clustering section). Sequences with no identical respective are considered as singleton and are excluded of the OTU clustering process.

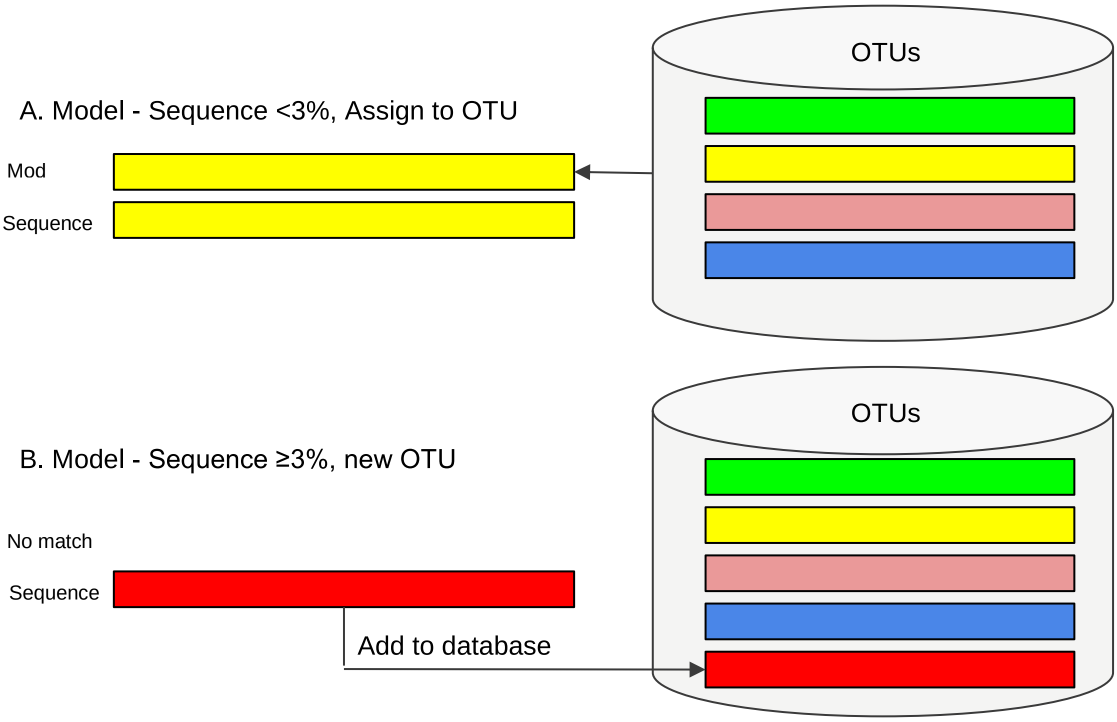
Chimera filtering



Chimera formed from X and Y

Chimera sequences are sequences that are composed by two other sequences. These sequences are filtered by a *de novo* approach consisting in a comparison with the other sequences in the dataset.

Clustering

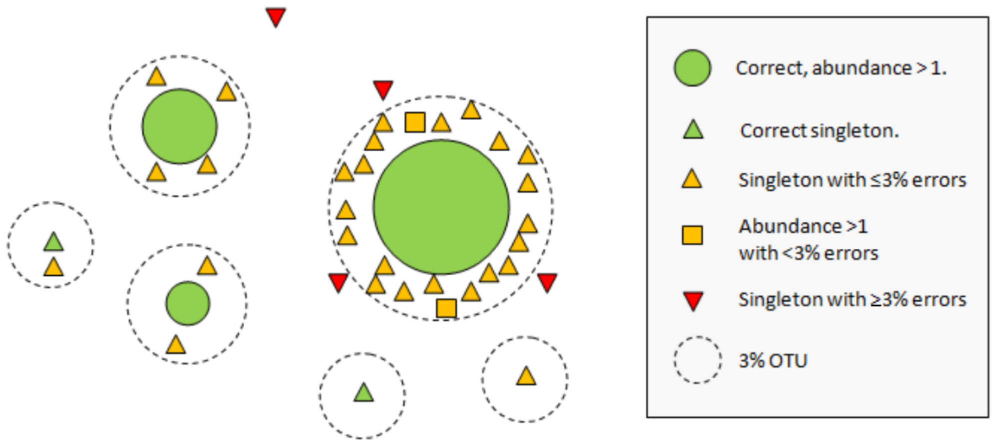

OTUs
A. Model - Sequence <3%, Assign to OTU
Mod
Sequence
OTUs
B. Model - Sequence ≥3%, new OTU
No match
Sequence
Add to database

The OTU clustering is performed with an Abundance-Greedy Clustering (AGC) [Westcott, Schloss, 2016 PeerJ; Rideout 2014; Schmidt et al. 2015]. The goal is to identify a set of correct biological sequences independently of sequencing errors and ribosomal RNA single polymorphism. The algorithm can be described as follow:

```
Initial n groups ordered by their "dereplicated abundance"
Each step:
    Pick a group and compare to the reference
    If close to the reference:
        Add in reference cluster
    Otherwise:
        Add it as a reference
```

AGC algorithm cost in the worst case $O(n^2)$ compared to $O(n^3)$ for hierarchical clustering approach implemented in Mothur for instance. The comparison is classically computed at threshold of 97% or 99%.
For 200 sequences, it represents 40000 operations for AGC compared to 8e+06 operations for hierarchical clustering. Sequences obtained after AGC are called OTU.

OTU abundance

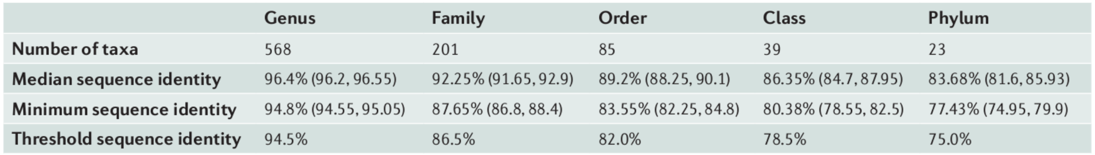

Correct, abundance > 1.
Correct singleton.
Singleton with ≤3% errors
Abundance >1 with <3% errors
Singleton with ≥3% errors
3% OTU

Adapted from [Edgar 2013]

The OTU abundance is calculated by aligning amplicon sequence against OTU sample per sample. The counts obtained are reproduced in a table called the count table.

OTU taxonomical annotation

The taxonomical annotation is performed by a fast Needleman-Wunsh alignment against ribosomal RNA databases. Annotation are filtered based on parameters identified in [Yarza 2014], as follow:



	Genus	Family	Order	Class	Phylum
Number of taxa	568	201	85	39	23
Median sequence identity	96.4% (96.2, 96.55)	92.25% (91.65, 92.9)	89.2% (88.25, 90.1)	86.35% (84.7, 87.95)	83.68% (81.6, 85.93)
Minimum sequence identity	94.8% (94.55, 95.05)	87.65% (86.8, 88.4)	83.55% (82.25, 84.8)	80.38% (78.55, 82.5)	77.43% (74.95, 79.9)
Threshold sequence identity	94.5%	86.5%	82.0%	78.5%	75.0%

Six databases are available in SHAMAN workflow:

- FINDLEY

Used for the taxonomical annotation of ITS sequences.
Findley, K., et al., Topographic diversity of fungal and bacterial communities in human skin. Nature, 2013, 498(7454), 367-370.
http://www.mothur.org/wiki/Findley_ITS_Database

- GREENGENES

Used for the taxonomical annotation of 16S, 18S sequences.
DeSantis, T. Z., et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and environmental microbiology, 2006, 72(7), 5069-5072.
http://greengenes.secondgenome.com/downloads/database/13_5

- SILVA LSU, SSU

Used for the taxonomical annotation of 16S, 18S, 23S, 28S sequences.
Pruesse, E., et al., SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic acids research, 2007, 35(21), 7188-7196.
https://www.arb-silva.de/

- UNDERHILL

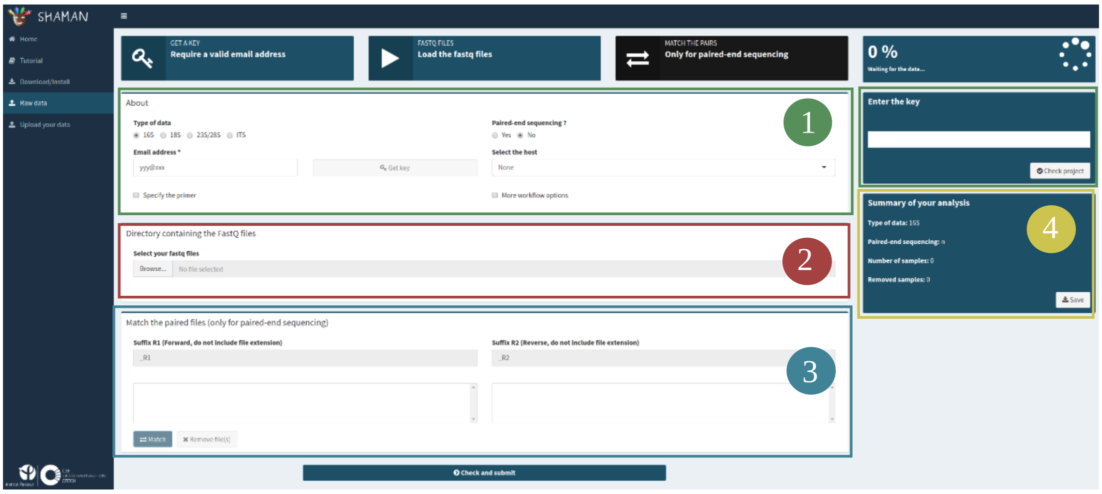
Used for the taxonomical annotation of ITS sequences.
Tang J, Iliev I, Brown J, Underhill D and Funari V. Mycobiome: Approaches to Analysis of Intestinal Fungi. Journal of Immunological Methods, 2015, 421:112-21.
https://risccweb.csmc.edu/microbiome/thf/

- UNITE

Used for the taxonomical annotation of ITS sequences.
Abarenkov, K., et al., The UNITE database for molecular identification of fungi–recent updates and future perspectives. New Phytologist, 2010, 186(2), 281-285.
https://unite.ut.ee/repository.php

The databases are updated every 2 months.

This workflow is available in “Raw data” section, is installed on galaxy.pasteur.fr and makes profit on the Institut Pasteur cluster to run calculation. The usage of this workflow was simplified in this section.



About

Type of data: ● 16S ○ 18S ○ 23S/28S ○ ITS

Paired-end sequencing ? ○ Yes ● No

Email address *: yyy@xxx

Select the host: None

□ Specify the primer

□ More workflow options

Directory containing the FastQ files

Select your fastq files: No file selected

Match the paired files (only for paired-end sequencing)

Suffix R1 (Forward, do not include file extension): _R1

Suffix R2 (Reverse, do not include file extension): _R2

Enter the key

Summary of your analysis

Type of data: 16S

Paired-end sequencing: n

Number of samples: 0

Removed samples: 0

1 Main panel allows to describe data characteristics:

- type of sequencing (16S, 18S, 23S, 28S, ITS)
- Paired/single end sequencing
- Host of the microbiome
- Specify primers
- And workflow parameters

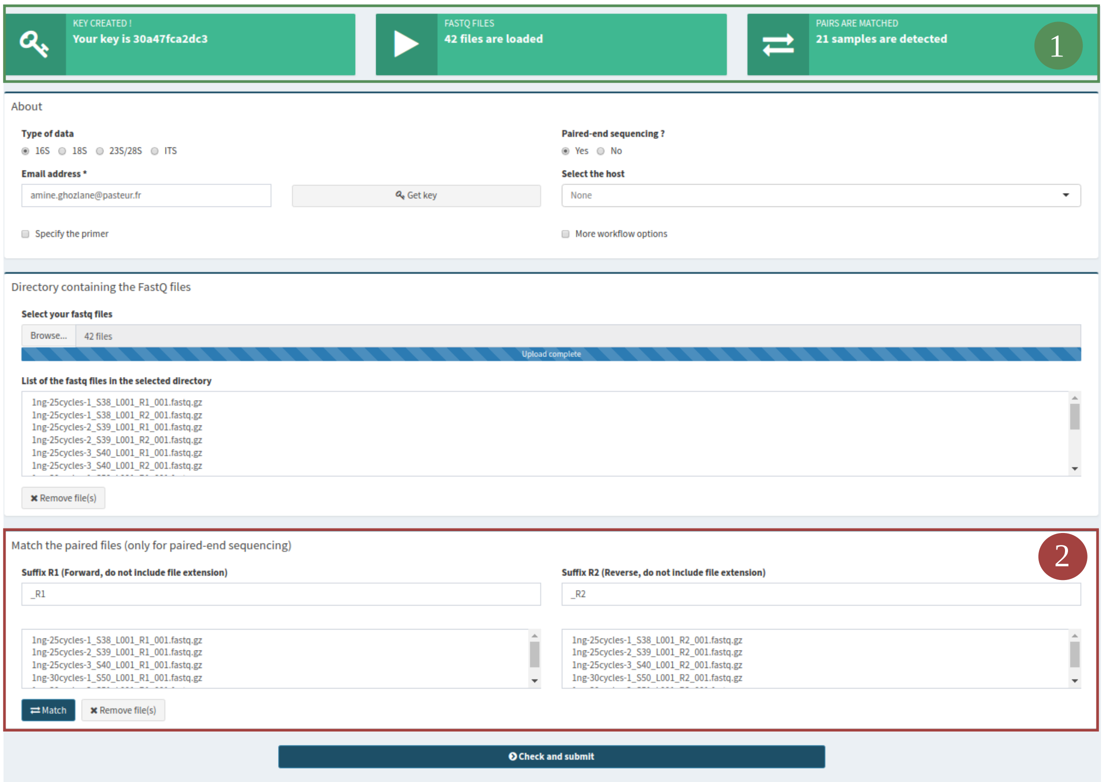
The user need to specify a mail and click on “get a key”. This key is the project identifier. It allows to follow calculation progress, view and download results.

2 Loading area for .fq, .fq.gz, .fastq and fastq.gz files

3 For paired end sequencing, files corresponding to R1 and R2 of a given sample need be paired.
A pattern is required to identify R1 and R2 files. The exclusion of this pattern must allow to get the same sample name.
Matched sample will be located next to its pair

4 Summary of the analysis.

For the example, we will load the mock sequencing performed with an Illumina MiSeq sequencer at Institut Pasteur using microbial community DNA from Zymobiomics (https://www.zymoresearch.com/media/amasty/amfile/attach/_D4300T_D4300_D4304_ZymoBIOMICS_DNA_Miniprep_Kit_1_1_2_LKN-SW_.pdf).



About

Type of data
● 16S ○ 18S ○ 23S/28S ○ ITS

Paired-end sequencing ?
● Yes ○ No

Email address *
amine.ghozlane@pasteur.fr

Select the host
None

□ Specify the primer

□ More workflow options

Directory containing the FastQ files

Select your fastq files
42 files

List of the fastq files in the selected directory

```
1ng-25cycles-1_S38_L001_R1_001.fastq.gz
1ng-25cycles-1_S38_L001_R2_001.fastq.gz
1ng-25cycles-2_S39_L001_R1_001.fastq.gz
1ng-25cycles-2_S39_L001_R2_001.fastq.gz
1ng-25cycles-3_S40_L001_R1_001.fastq.gz
1ng-25cycles-3_S40_L001_R2_001.fastq.gz
```

Match the paired files (only for paired-end sequencing)

Suffix R1 (Forward, do not include file extension)
_R1

Suffix R2 (Reverse, do not include file extension)
_R2

```
1ng-25cycles-1_S38_L001_R1_001.fastq.gz
1ng-25cycles-2_S39_L001_R1_001.fastq.gz
1ng-25cycles-3_S40_L001_R1_001.fastq.gz
1ng-30cycles-1_S50_L001_R1_001.fastq.gz
```

```
1ng-25cycles-1_S38_L001_R2_001.fastq.gz
1ng-25cycles-2_S39_L001_R2_001.fastq.gz
1ng-25cycles-3_S40_L001_R2_001.fastq.gz
1ng-30cycles-1_S50_L001_R2_001.fastq.gz
```

1 When fastq files were successfully loaded, all panel turn green. The key 30a47fca2dc3 is now specific to my project.

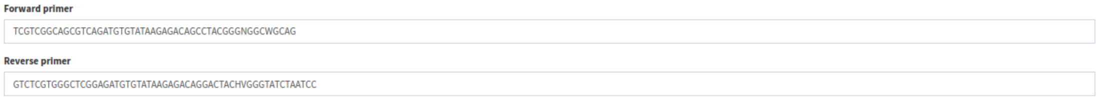
2 Matched paired-end reads appears in the same order.

More options are available to specify primer used for sequencing and change workflow parameters when checking correspond boxes.

☑ Specify the primer

☑ More workflow options

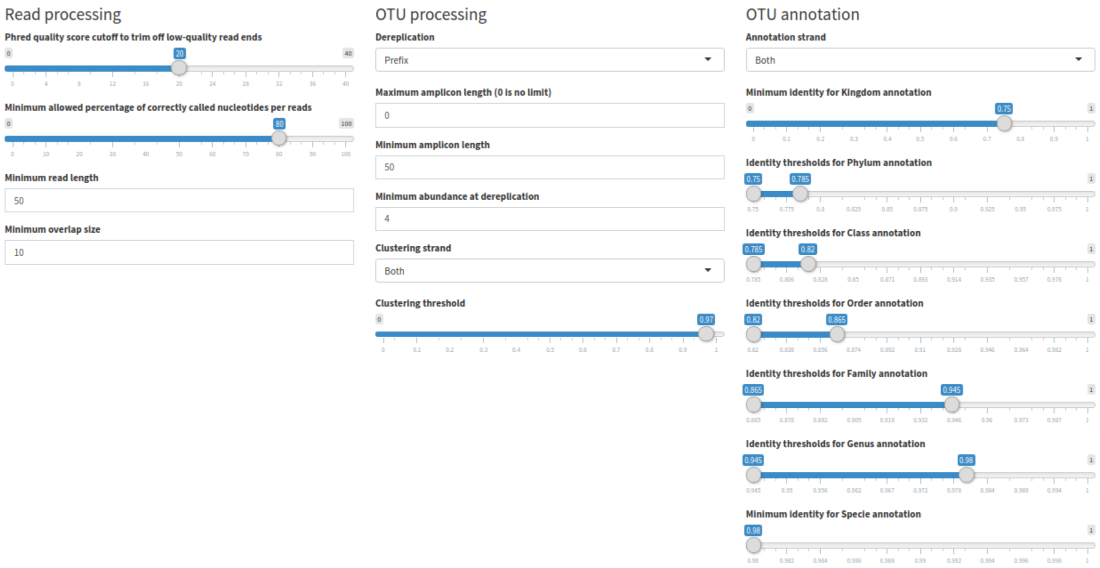
Primer sequences indicated in the following panels will be considered for read trimming.



Forward primer
TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG

Reverse primer
GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC

Users can also change workflow parameters for read processing, OTU processing and OTU annotation with the following panel:



Read processing

Phred quality score cutoff to trim off low-quality read ends: 20

Minimum allowed percentage of correctly called nucleotides per reads: 80

Minimum read length: 50

Minimum overlap size: 10

OTU processing

Dereplication: Prefix

Maximum amplicon length (0 is no limit): 0

Minimum amplicon length: 50

Minimum abundance at dereplication: 4

Clustering strand: Both

Clustering threshold: 0.97

OTU annotation

Annotation strand: Both

Minimum identity for Kingdom annotation: 0.75

Identity thresholds for Phylum annotation: 0.75 – 0.785

Identity thresholds for Class annotation: 0.785 – 0.82

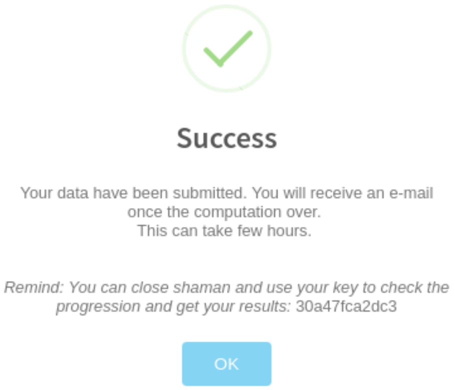
Identity thresholds for Order annotation: 0.82 – 0.865

Identity thresholds for Family annotation: 0.865 – 0.945

Identity thresholds for Genus annotation: 0.945 – 0.98

Minimum identity for Specie annotation: 0.98

When parameters are identified, user can run calculation with the button “check and submit”. Successful submission will be notified by the following panel:

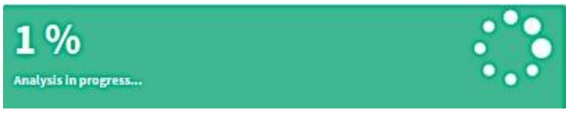


Success

Your data have been submitted. You will receive an e-mail once the computation over.
This can take few hours.

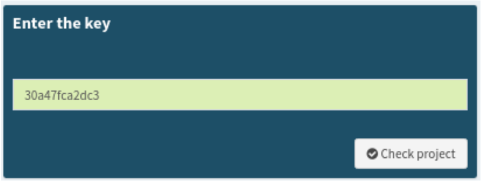
Remind: You can close shaman and use your key to check the progression and get your results: 30a47fca2dc3

By entering the project key (remotely), calculation progress will be notified in the running interface with an interactive progress percentage. An email is send when calculations are over.



1 %

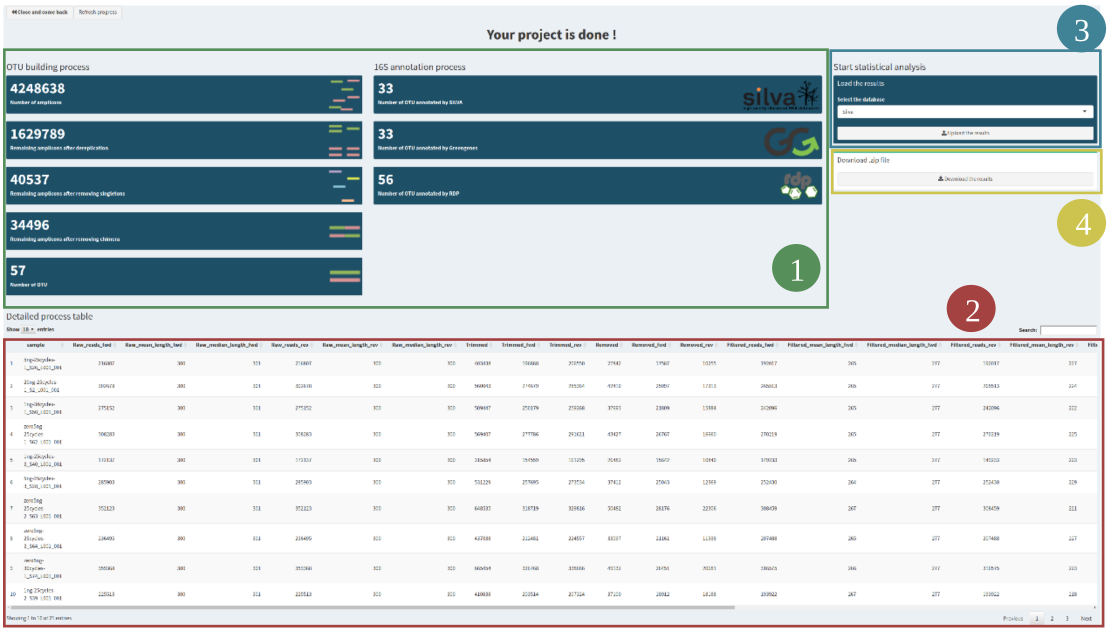
Analysis in progress...



Enter the key

30a47fca2dc3

When calculations are terminated, an interface is available to check



Your project is done !

OTU building process

4248638 Number of amplicons

1629789 Remaining amplicons after dereplication

40537 Remaining amplicons after removing singletons

34496 Remaining amplicons after removing chimera

57 Number of OTU

16S annotation process

33 Number of OTU annotated by SILVA

33 Number of OTU annotated by Greengenes

56 Number of OTU annotated by RDP

Start statistical analysis

Load the results

Select the database: silva

Download .zip file

Detailed process table

Show 10 entries   Search:

	sample	Raw_reads_fwd	Raw_mean_length_fwd	Raw_median_length_fwd	Raw_reads_rev	Raw_mean_length_rev	Raw_median_length_rev	Trimmed	Trimmed_fwd	Trimmed_rev	Removed	Removed_fwd	Removed_rev	Filtered_reads_fwd	Filtered_mean_length_fwd	Filtered_median_length_fwd	Filtered_reads_rev	Filtered_mean_length_rev	Filt
1	5ng-25cycles-1_S36_L001_001	216807	300	301	216807	300	300	403438	196888	206550	27842	17587	10255	192017	265	277	192017	227	
2	20ng-25cycles-1_S2_L001_001	309674	300	301	309674	300	300	569043	274679	295364	42410	25097	17313	288613	265	277	288613	224	
3	1ng-30cycles-1_S30_L001_001	275152	300	301	275152	300	300	509447	250179	259268	37693	21809	15884	242096	265	277	242096	222	
4	zero1ng-25cycles-1_S62_L001_001	308283	300	301	308283	300	300	569407	277786	291621	43427	26767	16660	270219	265	277	270219	225	
5	1ng-25cycles-3_S40_L001_001	172137	300	301	172137	300	300	315854	154589	161596	30483	19642	10840	145033	265	277	145033	223	
6	5ng-25cycles-3_S38_L001_001	265903	300	301	265903	300	300	531229	257695	273534	37412	25043	12369	252430	264	277	252430	229	
7	zero1ng-25cycles-2_S63_L001_001	352123	300	301	352123	300	300	648535	318719	329816	50482	28176	22306	308459	267	277	308459	221	
8	zero5ng-25cycles-3_S64_L001_001	236495	300	301	236495	300	300	437038	212481	224557	33097	21161	11936	207488	265	277	207488	227	
9	zero5ng-30cycles-1_S74_L001_001	350064	300	301	350064	300	300	669454	328768	339886	41133	21451	20081	318575	266	277	318575	223	
10	1ng-25cycles-2_S39_L001_001	225513	300	301	225513	300	300	410838	203514	207324	37100	18912	18188	193922	267	277	193922	218	

Showing 1 to 10 of 21 entries   Previous 1 2 3 Next

1 Information about the number of sequences obtained at each clustering step and annotated against ribosomal RNA databases.

2 Information about the read and OTU processing sample by sample

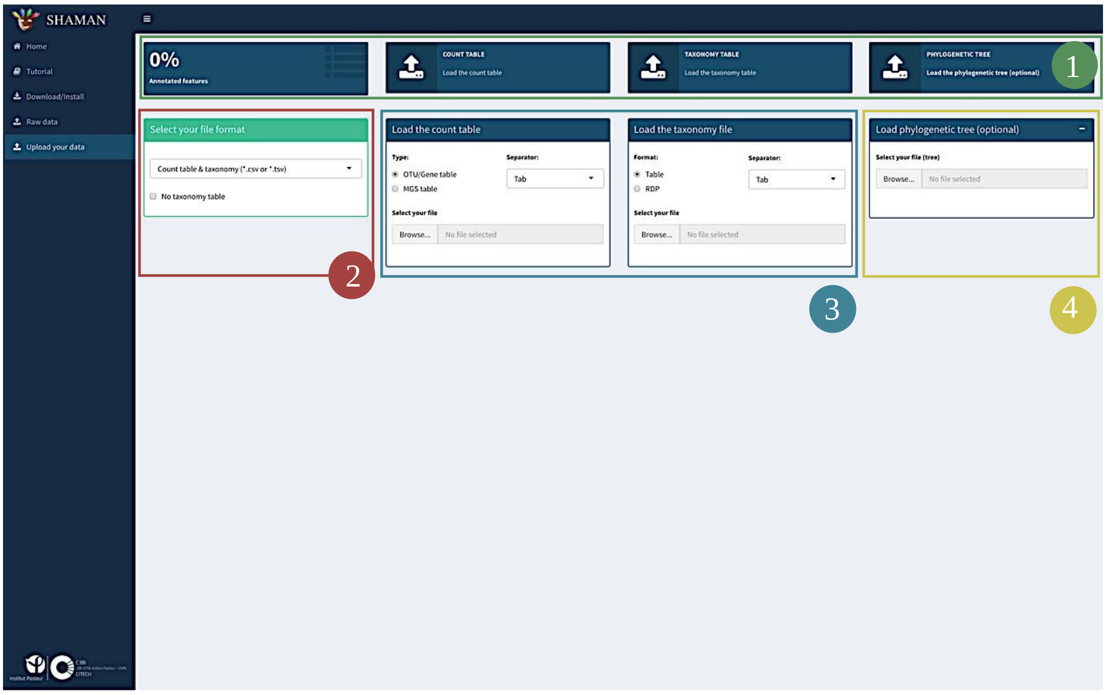
3 Panel allowing to load resulting data for corresponding databases.

4 Download all results.

A command line version of the workflow is available at https://github.com/aghozlane/masque.

Load count and annotation data

Processed data can be loaded in SHAMAN. User need a raw count table providing the number of reads aligned against each OTU/gene for every sample and a matrix providing an annotation for each element. These information can also be provided as a *biom* file, as follow:



1 Information about the count table, the annotation table and the phylogenetic tree (optional)

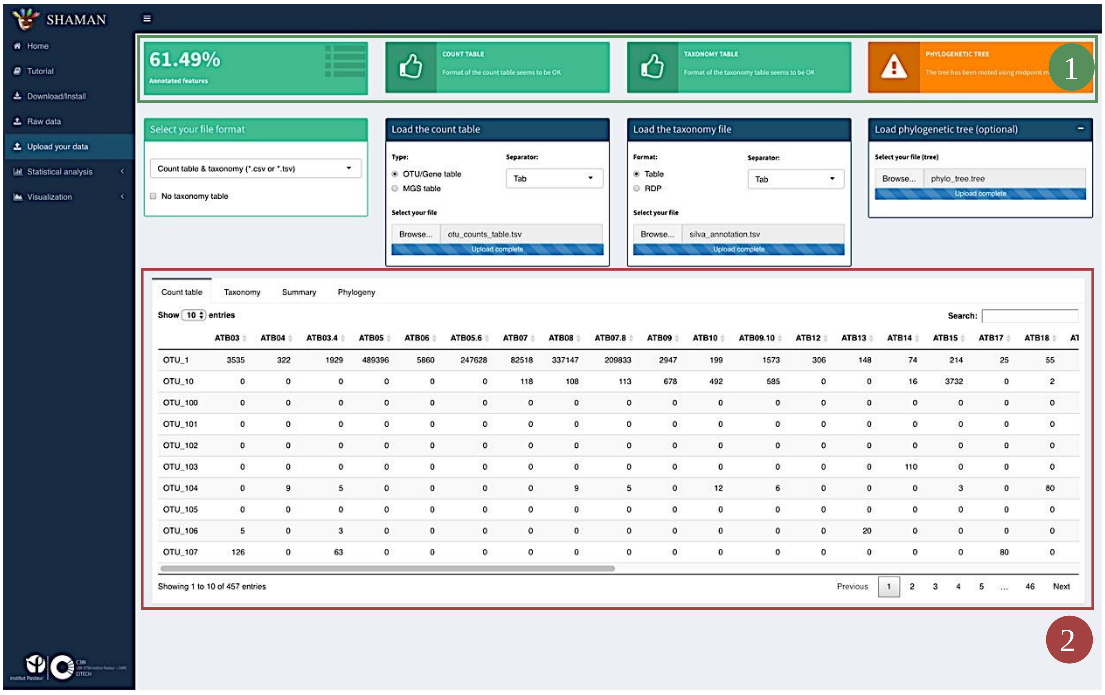
2 Select the entry file format. User can choose to upload a count table and an annotation table or one file in biom format.

The user can also perform a study without an annotation file, in the case where the analysis is performed at the same level as the count file.

3 Load input files. The user can choose the type of data (OTU / Gene or MGS), the column separator and, for the RDP annotation, the probability threshold.

4 Load the phylogenetic tree *.tree* (optional)

Once the data are loaded, SHAMAN displays the data of the user into two tables (count and annotation). Some graphical representations allowing to check the quality of the annotation are also available.



61.49%
Annotated features

COUNT TABLE
Format of the count table seems to be OK

TAXONOMY TABLE
Format of the taxonomy table seems to be OK

PHYLOGENETIC TREE
The tree has been rooted using midpoint m

Select your file format
Count table & taxonomy (*.csv or *.tsv)
□ No taxonomy table

Load the count table
Type: ● OTU/Gene table ○ MGS table
Separator: Tab
Select your file
Browse... otu_counts_table.tsv
Upload complete

Load the taxonomy file
Format: ● Table ○ RDP
Separator: Tab
Select your file
Browse... silva_annotation.tsv
Upload complete

Load phylogenetic tree (optional)
Select your file (tree)
Browse... phylo_tree.tree
Upload complete

Count table | Taxonomy | Summary | Phylogeny

Show 10 entries
Search:

	ATB03	ATB04	ATB03.4	ATB05	ATB06	ATB05.6	ATB07	ATB08	ATB07.8	ATB09	ATB10	ATB09.10	ATB12	ATB13	ATB14	ATB15	ATB17	ATB18	AT
OTU_1	3535	322	1929	489396	5860	247628	82518	337147	209833	2947	199	1573	306	148	74	214	25	55	
OTU_10	0	0	0	0	0	0	118	108	113	678	492	585	0	0	16	3732	0	2	
OTU_100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
OTU_101	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
OTU_102	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
OTU_103	0	0	0	0	0	0	0	0	0	0	0	0	0	0	110	0	0	0	
OTU_104	0	9	5	0	0	0	0	9	5	0	12	6	0	0	0	3	0	80	
OTU_105	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
OTU_106	5	0	3	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	
OTU_107	126	0	63	0	0	0	0	0	0	0	0	0	0	0	0	0	80	0	

Showing 1 to 10 of 457 entries

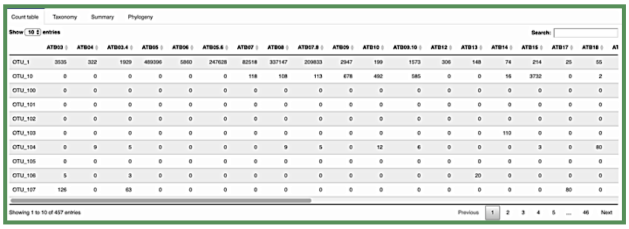
Previous 1 2 3 4 5 ... 46 Next

1 Successful file loading will end up to turn the boxes turn green and indicate the percentage of annotation.

2 Area presenting an overview of uploaded files and basic statistics. The details of the 4 tabs are given on the next page.

The count table must be in the following format: samples in column and individuals (OTU, ...) in row.

The 4 tabs of the previous box are the following



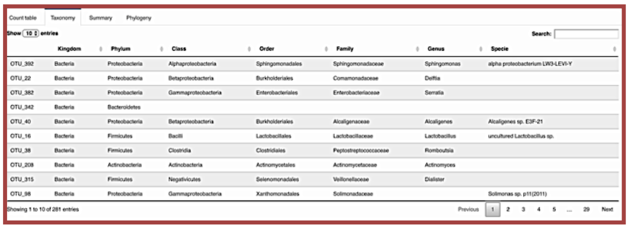
Show 10 entries

Search:

	ATB03	ATB04	ATB03.4	ATB05	ATB06	ATB05.6	ATB07	ATB08	ATB07.8	ATB09	ATB10	ATB09.10	ATB12	ATB13	ATB14	ATB15	ATB17	ATB18	AT
OTU_1	3535	322	1929	489396	5860	247628	82518	337147	209833	2947	199	1573	306	148	74	214	25	55	
OTU_10	0	0	0	0	0	0	118	108	113	678	492	585	0	0	16	3732	0	2	
OTU_100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
OTU_101	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
OTU_102	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
OTU_103	0	0	0	0	0	0	0	0	0	0	0	0	0	0	110	0	0	0	
OTU_104	0	9	5	0	0	0	0	9	5	0	12	6	0	0	0	3	0	80	
OTU_105	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
OTU_106	5	0	3	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	
OTU_107	126	0	63	0	0	0	0	0	0	0	0	0	0	0	0	0	80	0	

Showing 1 to 10 of 457 entries

Previous 1 2 3 4 5 ... 46 Next



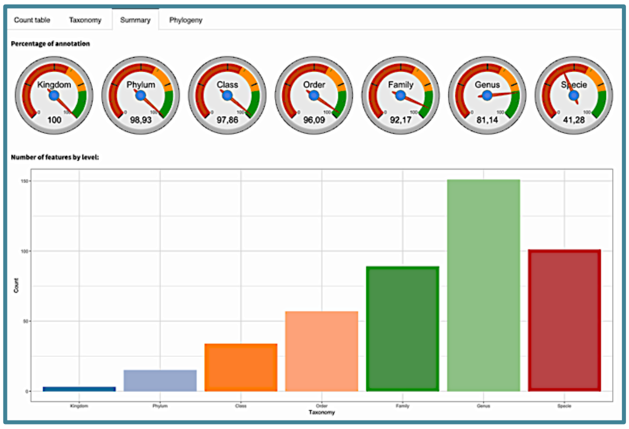
Show 10 entries

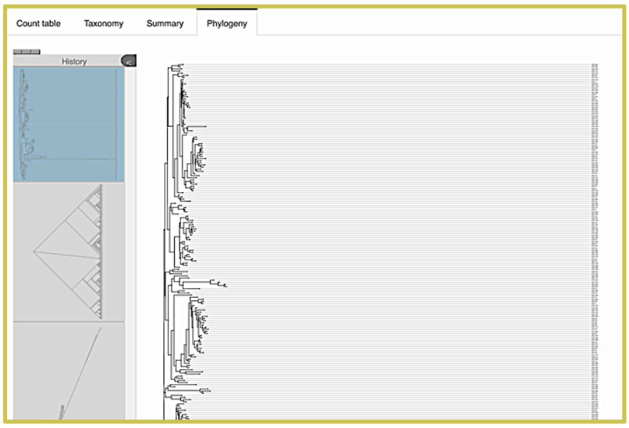
Search:

	Kingdom	Phylum	Class	Order	Family	Genus	Specie
OTU_392	Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingomonas	alpha proteobacterium LW3-LEVI-Y
OTU_22	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Delftia	
OTU_382	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Serratia	
OTU_342	Bacteria	Bacteroidetes					
OTU_40	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Alcaligenaceae	Alcaligenes	Alcaligenes sp. E3F-21
OTU_16	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	uncultured Lactobacillus sp.
OTU_38	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	Romboutsia	
OTU_208	Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	Actinomycetaceae	Actinomyces	
OTU_315	Bacteria	Firmicutes	Negativicutes	Selenomonadales	Veillonellaceae	Dialister	
OTU_98	Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonadales	Solimonadaceae		Solimonas sp. p11(2011)

Showing 1 to 10 of 281 entries

Previous 1 2 3 4 5 ... 29 Next





1 Count table with sample in column and OTU/genes in row.

2 Annotation table with different annotation levels in column and the OTU/genes in row.

3 Figure presenting the percentage of OTU/genes annotated at each annotation level and the number of samples at each level.

4 Figure of the phylogenetic tree (only if a phylogenetic *.tree* is loaded, optional)

Once the data has been loaded, two new tabs appear "Statistical analysis" and "Visualization".

Statistical analysis - This tab is divided into 3 sub-tabs:

- Run differential analysis: allows to define the statistical model as well as some options, the taxonomic level and to create the vectors of contrasts for defining comparisons
- Diagnostic plots: many different visualizations (barplots, boxplots, clustering, PCA, PCoA, NMDS, ...) are available to control the design of the experiment, check the normalization and identify possible sequencing problems.
- Tables: get the results of the differential analysis for each defined contrast vector.

Visualization - This tab is divided into 2 sub-tabs:

- Global views: provides many interactive representations (barplots, heatmap, boxplots, krona, tree of abundance, network plot, curves of rarefaction, diversities) to visualize sample composition as well as the results of the differential analysis according to the experimental design.
- Plots comparison: several plots (venn diagram, heatmap of the log2-foldchange, p-value density, UpSet plot) are provided to compare the results of the analysis for each contrast (at least 2 contrast vectors have to be defined)

SHAMAN statistics are based on DESeq2 package. This package requires the definition of a statistical model and to set up several parameters (not presented here). The different steps can be summarized as follow:

1. Data normalization (calculation of size factors)
2. Estimation of the dispersion based on a modeling of the average normalized counts and the empirical dispersion.
3. Adjustment of the generalized linear model based on the negative binomial distribution and a log link function.
4. Statistical test (Wald test) performed on model parameters
5. Outlier filtering
6. Correction of the multiple tests

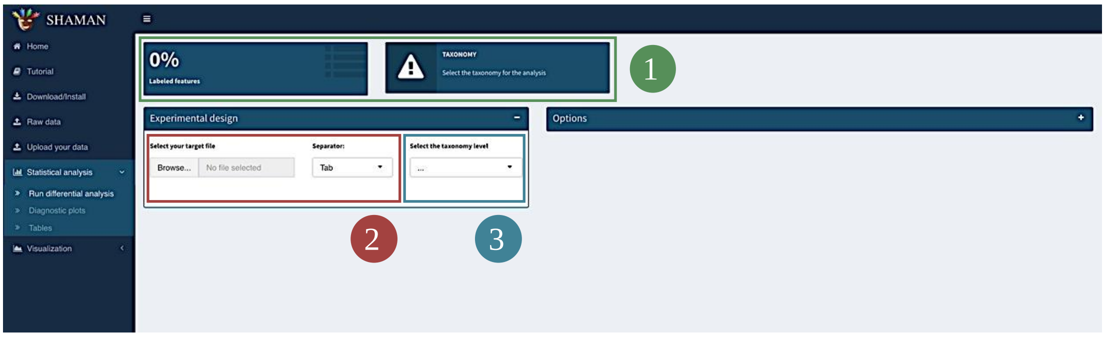
DESeq2 has been shown as one of the best methods to identify differentially abundant features.

Build the statistical model

In order to build the statistical model, the user must first provide the experimental design (target file) and select the taxonomic level. It is then necessary to select the variables of interest describing the biological phenomenon studied as well as the possible interactions between these variables. Before starting the analysis, the user can set the various options related to DESeq2 such as the independent filtering, the shape of the dispersion modeling, the method of correction of the multiple tests ... (for more information see [Love et al., 2014]).

All these steps must be carried out in the "Run differential analysis" sub-tab.

Loading the experimental design



Select your target file

Browse... No file selected

Separator: Tab

Select the taxonomy level ...

1 Information on experimental design and annotation level.

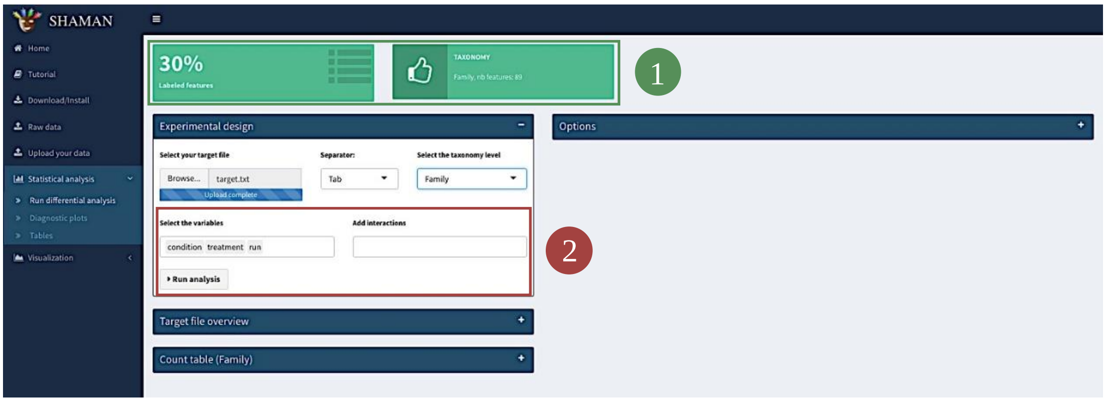
2 Area to load the experimental design. The user can also define the column separator used in the input file. This file must respect a certain format (see Appendix A for more information).

3 Choice of the annotation level. These levels correspond to those provided in the annotation file.

If no annotation file has been provided, the only option is "OTU/gene", which corresponds to the level at which the count table was calculated.

Define the model

Once the target file is loaded and the taxonomic level is selected, SHAMAN checks that the sample names match the names of the count table and merges counts of OTUs/genes with the same annotation.



30% Labeled features

TAXONOMY Family, nb features: 89

Select your target file: target.txt

Separator: Tab

Select the taxonomy level: Family

Select the variables: condition treatment run

Add interactions:

1 Successful loading of the target file will end up to turn the boxes in green and to indicate the percentage of samples present in the count table for which SHAMAN has information coming from the design loaded by the user.

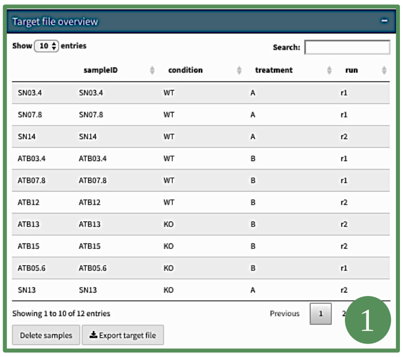
2 Selection of the variables of interest (coming from the design). The variables can be both qualitative and quantitative. It is also possible to add interactions between the variables.

In the example, the model has 3 variables:

- condition: 2 modalities "WT" and "KO"
- treatment: 2 treatments "A" and "B"
- run: "batch" effect which allows to take into account a 2-run sequencing

In this case, it would be wise to add an interaction between the condition and treatment variables. Indeed, it can be assumed that the effect of the treatment depends on the condition and therefore wants to test the effects by subgroups. Here, to simplify the notation this effect has been neglected.

Caution: It is recommended to avoid using numbers for qualitative variables. For instance, for the "run" variable, the user should prefer the notation "r1" and "r2" to "1" and "2".



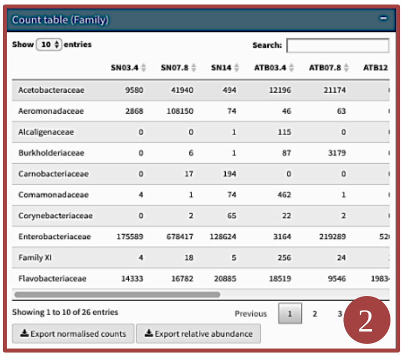
Target file overview

Show 10 entries

Search:

	sampleID	condition	treatment	run
SN03.4	SN03.4	WT	A	r1
SN07.8	SN07.8	WT	A	r1
SN14	SN14	WT	A	r2
ATB03.4	ATB03.4	WT	B	r1
ATB07.8	ATB07.8	WT	B	r1
ATB12	ATB12	WT	B	r2
ATB13	ATB13	KO	B	r2
ATB15	ATB15	KO	B	r2
ATB05.6	ATB05.6	KO	B	r1
SN13	SN13	KO	A	r2

Showing 1 to 10 of 12 entries



Count table (Family)

Show 10 entries

Search:

	SN03.4	SN07.8	SN14	ATB03.4	ATB07.8	ATB12
Acetobacteraceae	9580	41940	494	12196	21174	
Aeromonadaceae	2868	108150	74	46	63	
Alcaligenaceae	0	0	1	115	0	
Burkholderiaceae	0	6	1	87	3179	
Carnobacteriaceae	0	17	194	0	0	
Comamonadaceae	4	1	74	462	1	
Corynebacteriaceae	0	2	65	22	2	
Enterobacteriaceae	175589	678417	128624	3164	219289	52
Family XI	4	18	5	256	24	
Flavobacteriaceae	14333	16782	20885	18519	9546	1983

Showing 1 to 10 of 26 entries

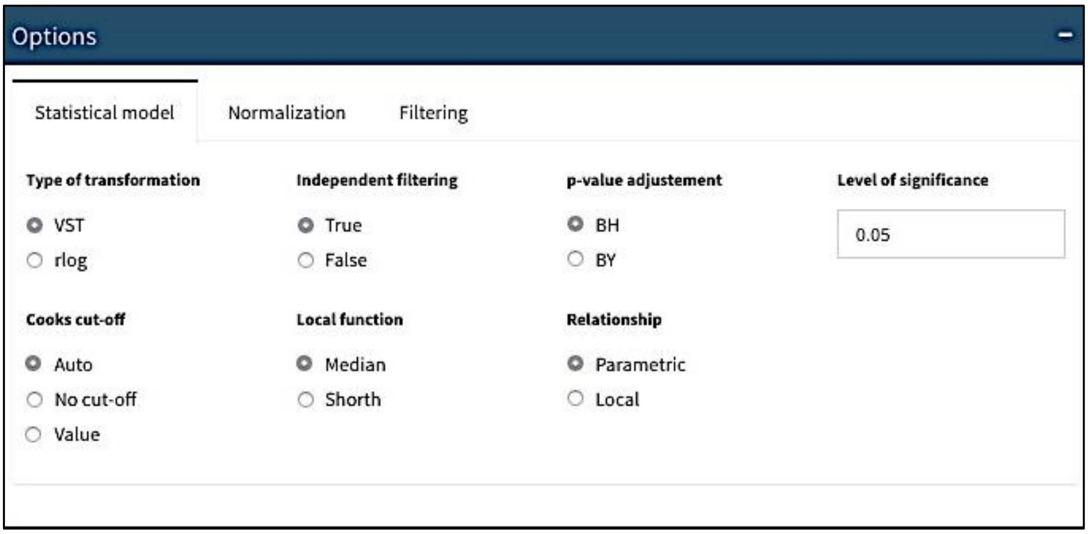
1 Overview of the target file loaded by the user. This file describes each sample by one or more variables (qualitative and/or quantitative).

The user can remove some samples of poor quality and export the new target file.

2 Overview of the "aggregated" normalized count table at the taxonomic level selected by the user.

The user can export the table of normalized counts and/or relative abundances

Model options and normalization



Type of transformation
● VST
○ rlog

Independent filtering
● True
○ False

p-value adjustement
● BH
○ BY

Level of significance
0.05

Cooks cut-off
● Auto
○ No cut-off
○ Value

Local function
● Median
○ Shorth

Relationship
● Parametric
○ Local

- **Type of transformation**

Two transformations are available in DESeq2, VST (Variance Stabilizing Transformation) and rlog (regularized log transformation). The data being heteroscedastic, this transformation makes it possible to eliminate the dependence of the variance on the mean. It is only used to visualize and/or classify the data (modeling is done on the counts).

When the number of samples is large, it is recommended to use VST transformation which is faster than the rlog.

- **Independent filtering**

This filter based on the average counts over all the samples. It aims at filtering the individuals (species, genera, ...) which are very unlikely to be differentially abundant. The threshold used for the filter is determined such that the number of significant individuals reach its maximum for a given FDR.

- **p-value adjustment**

Two usual methods for multiple correction by FDR are proposed: Benjamini-Hochberg et Benjamini-Yekuteli.

- **Level of significance**

Level of significance for the statistical test. By default, this value is 5%.

- **Cooks cut-off**

The log-fold change value is strongly influenced by outliers. To detect outliers, the Cooks distance which measure the impact of removing a sample on the estimation of the parameters can be used. The 99th percentile of the Fisher distribution (F (p, m-p), with p the number of parameters and m the number of samples) is used as a threshold for the Cooks distance.

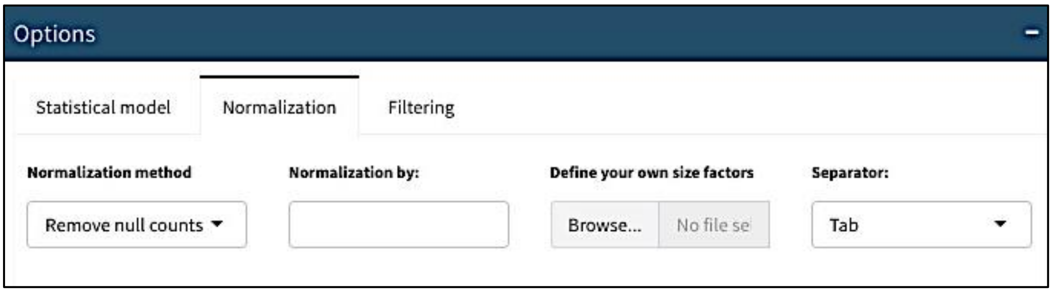
- **Local function**

To calculate the size factors (used to normalize the data), there are two options "median" or "shorth". The first one, "median" is the default method which aims at calculating the median of the ratio between counts and the geometric mean (see normalization section for more details). The second one, "shorth", calculates the average of the smallest interval that covers half of the values. This option is especially recommended for low counts.

- **Relationship**

For each feature, the dispersion is estimated by modeling the relationship between empirical dispersion and mean counts. To model this relation, three methods are proposed:

- ✓ a parametric regression of the form $\alpha_{tr}(\bar{\mu}) = \alpha_0 + \frac{a_1}{\bar{\mu}}$
- ✓ a local regression which makes it possible to obtain a better modeling when the form in "1/x" is not well adapted.
- ✓ the mean. This approach can be used when the number of individuals is small.



Normalization method: Remove null counts

Normalization by:

Define your own size factors

Separator: Tab

- **Normalization method**
  SHAMAN proposes 4 normalization methods:
  - ✓ Usual: default method of DESeq2 which consists in calculating the median of the ratio between the counts and the geometric mean

$$s_j = median_i \frac{K_{ij}}{\left(\prod_{k=1}^{n} K_{ik}\right)^{1/n}}$$

  - ✓ Remove null counts: The calculation is done only with the non-null counts

$$s_j = median_i \frac{K_{ij}}{\left(\prod_{k \in S_i} K_{ik}\right)^{1/n_i}}$$

  - ✓ Weighted: weighted version of the previous method (the weights are proportional to the number of samples with a non-null value).

  - ✓ Total count: this method aims at calculating a size factor by dividing the total counts of each sample by the average of the totals over all the samples. This approach must be used when the composition of one condition to another is very different, i.e. when several species are present in one condition and absent in the other (and vice versa).

  For metagenomics analysis, the matrices are very sparse (many 0). Therefore, it is recommended to avoid the use of the "usual" method.

- **Normalization by**

  This option makes it possible to perform group normalization according to a variable of the experimental design.
  This option can be useful when the user wants to visualize the results of several studies without modifying the normalization.

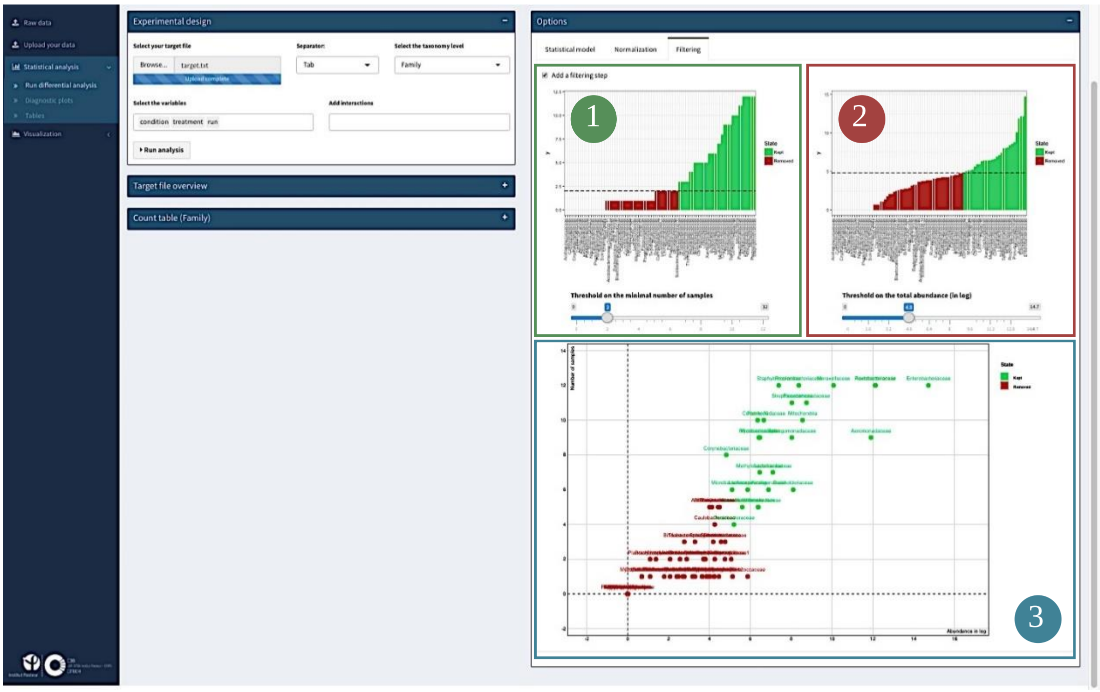
  Caution: avoid selecting a variable of interest to normalize the data because it can create some biases.

- **Define your own size factors**
  The user can load his own vector of size factors (obtained by another method for instance).

- **Separator**
  File separator for the size factor file.

Data filtering (optional)

In some cases, when many elements have low counts and/or are only detected for a small number of samples, it may be useful to filter them before starting the analysis. This will not have a strong influence on the results of the differential analysis. Indeed, the independent filtering already filter these elements. For studies with many features, it allows to reduce the computation time.



Select your target file: target.txt (Upload complete)
Separator: Tab
Select the taxonomy level: Family
Select the variables: condition treatment run
Add interactions:
Run analysis

☑ Add a filtering step
Threshold on the minimal number of samples: 2
Threshold on the total abundance (in log): 4.8

1 Threshold on the occurrence in all samples. By default, the threshold set by SHAMAN corresponds to 20% of the samples.

2 Threshold on average abundance. To calculate a threshold automatically, SHAMAN performs a linear regression between the number of sample with an average abundance of at least x and the average abundance.

3 Figure representing the number of occurrences in all samples compared to the average abundance. The red dots correspond to the elements that will be filtered out once the two filters are applied.

Define a contrast vector

Once the statistical analysis is done, SHAMAN provides the list of parameters corresponding to the variables included in the model. From the estimation of these parameters, the user can test different effects. Denote by β_i the vector of parameters. Let c be a vector of contrasts, define as a linear combination of parameters. We then use a Wald test whose test statistic is distributed according to a standard normal distribution:

$$\frac{\beta_i^c}{\sqrt{c^t \Sigma_i c}} \sim N(0,1)$$

$$\beta_i^c = c^t \beta_i$$

The matrix Σ_i corresponds to the variance-covariance matrix of the model parameters.

Example 1

If the user wants to test whether there is a treatment effect (A vs B), it corresponds to test whether the difference A-B is zero. To do so, a contrast vector composed of 1 and -1 for the parameters associated with the treatments A and B has to be created. The hypotheses of the test will be as follows:

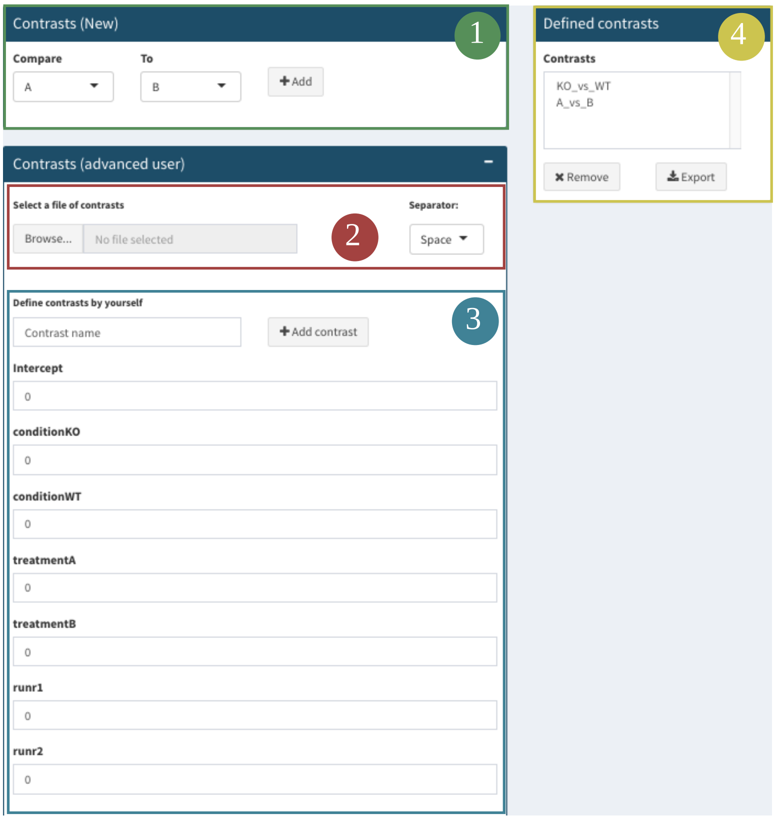
$$\begin{cases} \text{H0}: \beta_A - \beta_B = 0 \\ \text{H1}: \beta_A - \beta_B \neq 0 \end{cases}$$

In a complex experimental design, many comparisons can be done by defining a set of contrast vectors.

Example 2

Assume that we have an experimental design with measurements for 3 treatments A, B and C. The user wants to know if the effect of treatment C corresponds to the average of treatments A and B. In this case, 3 parameters will be estimated (one for each treatment). To get the desired comparison, the user must define the following contrast vector:

$$\mathbf{c}^T\boldsymbol{\beta} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & -1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \tfrac{1}{2}\beta_1 + \tfrac{1}{2}\beta_2 - \beta_3 = 0$$



1 Area for automatic definition of contrast vectors based on model variables. This panel allows to define most of the usual contrast vectors.

2 Loading area of the contrast file created from SHAMAN. The variables must exactly match those that were used to create the contrast file.

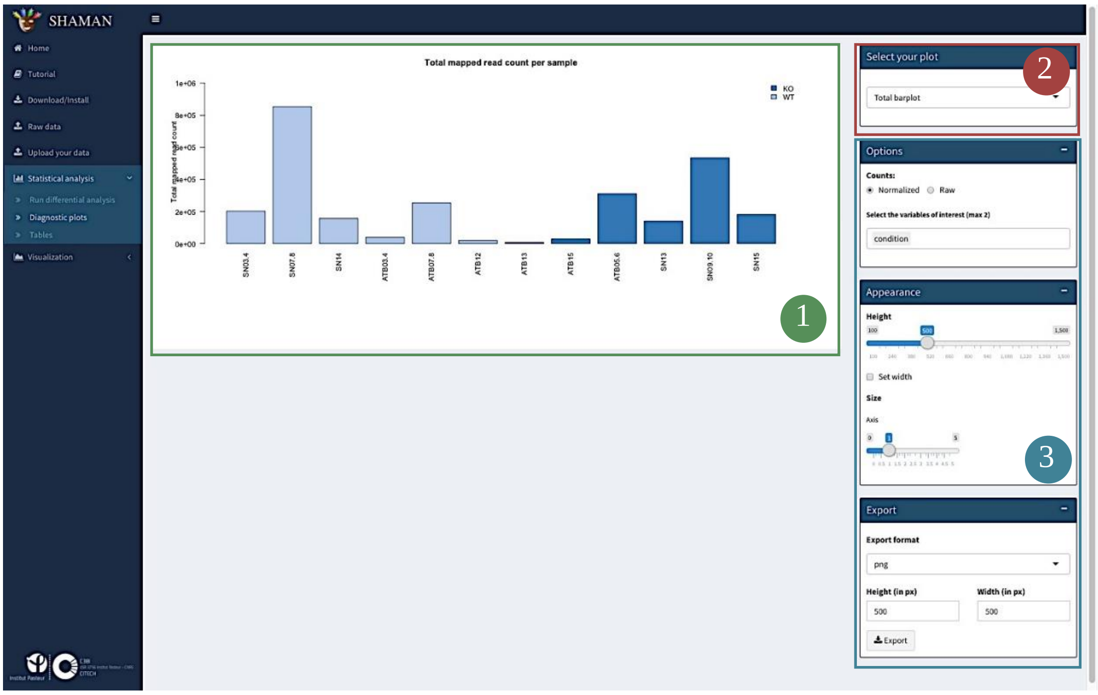
3 Area for manual definition of contrast vectors. This area is reserved for advanced users knowing how to define a vector of contrasts from the parameters of the statistical model.

4 List of contrasts created (by one of the 3 possibilities). The user can export the contrasts vectors as a .txt file (for later use) and / or remove some contrasts.

Assessing the statistical model

Once the model is defined and analysis carried out, the user must check that the normalization, the estimation of the dispersion and the size factors are correct with the Total barplot, boxplots, the dispersion and the size factor vs total plots visualizations. Different ordination methods (PCA, PCoA, NMDS) are also available to check that there has been no inversion of samples and that the biological effect studied is well observed in the data.

All of these representations are in the "Diagnostic Plots" sub-tab.



Select your plot: Total barplot

Counts: ● Normalized ○ Raw

Select the variables of interest (max 2): condition

Height: 500 (100 – 1,500)

□ Set width

Size — Axis: 1 (0 – 5)

Export format: png

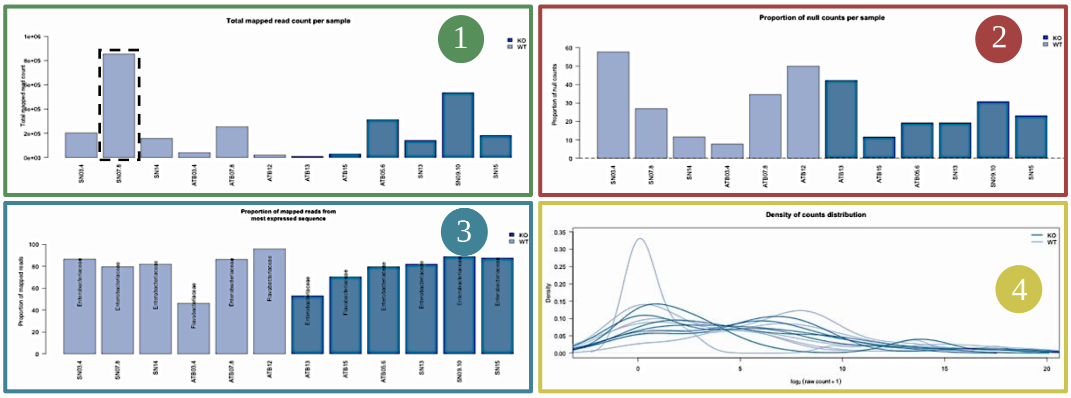
Height (in px): 500 Width (in px): 500

Export

1 Representation of the selected graphic.

2 Selection of the graph to be represented.

3 Set of options for graphics (depends on the type of graphic selected). Possibility to choose the variable or variables to represent, to modify the appearance and to export the figure in different formats (png, pdf, eps, ...).



1 Barplot of the counts for each sample

2 Barplot of the null counts for each sample

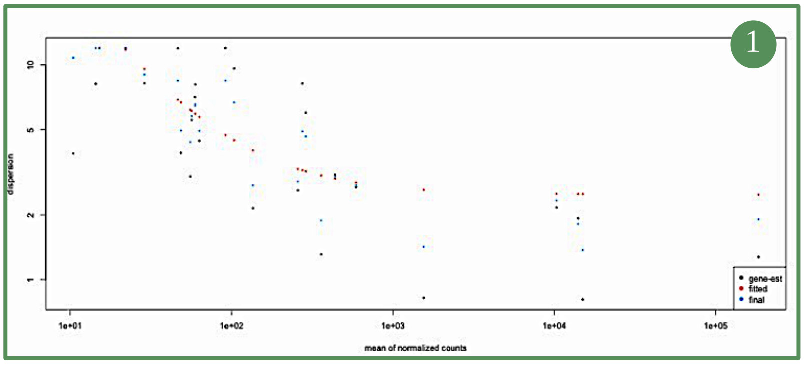
3 Barplot of the most abundant features

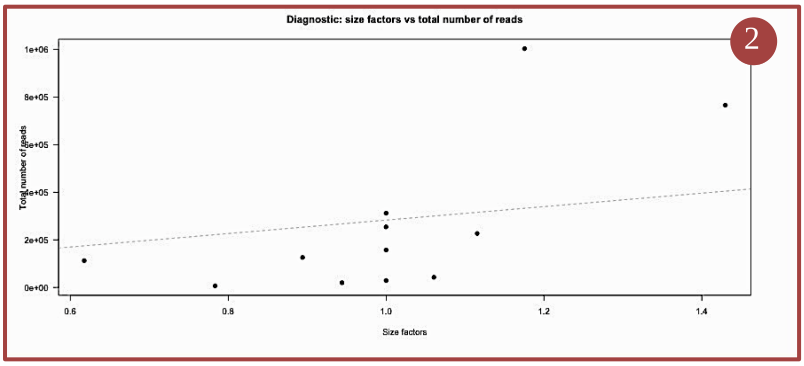
4 Density plot of the counts in log2

Interpretation:
These diagrams allow to identify possible issues linked to the sequencing depth. In the example presented, one of the "SN07-8" samples (in dashed lines) seems to have a large number of reads aligned with the others. This sample can also be detected as an outlier when looking at the representation of densities.

In this context, the user must give attention to this sample and check if the whole bioinformatics process went well. It might be necessary to delete this sample from the analysis.

Note: The rarefaction curves presented in the "Visualization" section can also be used to determine if one sample must be deleted from the study.





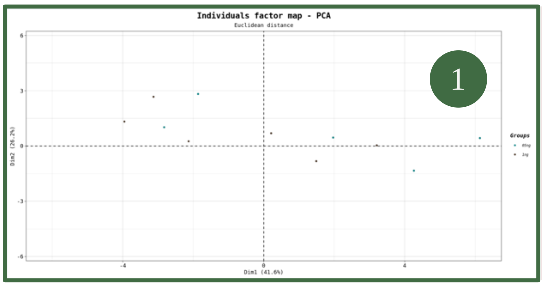
1 Empirical dispersion respect to the mean of the normalized counts.

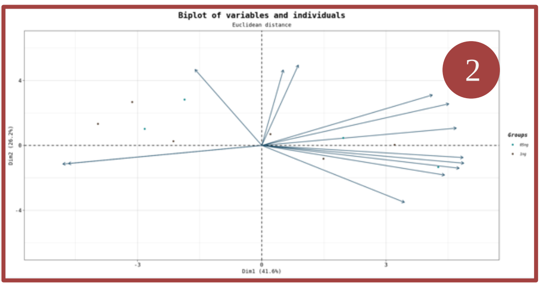
2 Total number of reads respect to the size factors

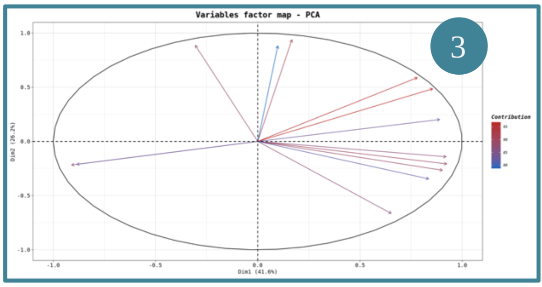
Interpretation:

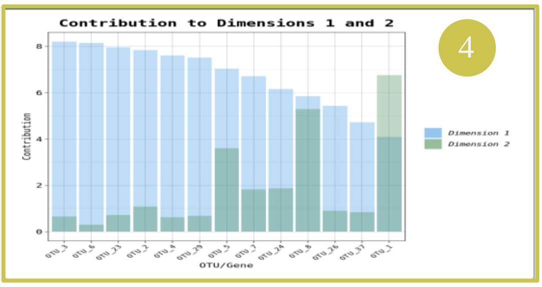
These two representations allow to check the normalization and the estimation of the dispersion. Concerning the estimation of the dispersion, the user must check that the model chosen corresponds to the shape of the point cloud. In the example presented here, the default function ("parametric") did not provide a satisfactory modeling, it was then decided to use the "local" option to obtain a modeling more consistent with the data.

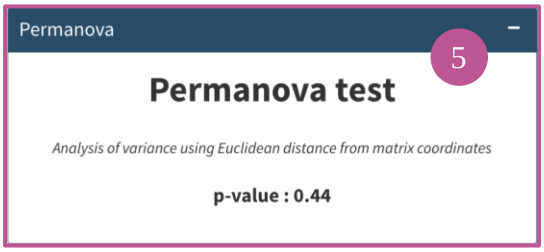
The representation of the total number of reads aligned according to the size factors allows to check the normalization of the data and to identify possible outliers. In this graph, the size factors must be within a reasonable range (between 0.5 and 2.5 approximately) and the points must be close to the dashed line. Indeed, the size factors are supposed to correct the biases due to the variation of the sequencing depth, it is thus important to verify that there is indeed a link between size factors and sequencing depth.











Permanova

Permanova test

Analysis of variance using Euclidean distance from matrix coordinates

p-value : 0.44

For each graph, the user can display the data label, center the graph (not for the variable graph) and resize many graph elements (title, subtitle, axes, axis label, data label, legend, X-scale labels and Y-scale labels). However, some parameters are not available depending on the graph displayed by the user (for example, it is impossible to center the variable graph). For the biplot and variable graph, the user can select a contribution threshold.

1 Graph of PCA individuals. Representation of the first 2 axes (chosen by default) according to the variables of interest selected by the user.

2 PCA biplot. Representation of variables (arrows) and individuals (dots). The user can display variable labels, individual labels or both.

3 Graph of PCA variables. Variables represented in the trigonometric circle with the contribution gradient.

4 Contribution bar chart. Shows the contribution of each variable to each selected dimension. The user can also display the bar chart with the sum of each variable's contribution to each selected dimension.

5 Permanova test available on PCA. It uses the Euclidean distance between the matrix coordinates of individuals.

Interpretation:

PCA is an excellent statistical analysis tool for extracting information from complex data sets. It reduces the number of features in the dataset while preserving essential information (2D or 3D), the first few principal components capture the maximum variance in the data.

To do this, we generally use Euclidean distance and transform the data set into a reduced centered set (a question of unity). Next, we find the total inertia using the variance-covariance matrix (because the trace of this matrix is the total inertia). Finally, we find the eigenvectors (normalized to 1) of this matrix, and build our principal component axes (PC1, PC2, etc.) based on the highest eigenvalue, the second-highest eigenvalue, etc. PCA gives a linear combination between variables (principal components) and individuals, and principal components are not correlated two by two. Variables are generally represented by arrows, the quality factor being related to their cosine squared: the closer the variables are, the higher their correlation. Thus, two opposite variables have a negative correlation. Two variables whose angle is almost square are not correlated.

The individuals graph provides information on

- The distribution of individuals in the reduced feature space (PC1 vs. PC2 in the first image)
- Similarity/dissimilarity groups (the colors in the legend help you visualize these, or you can select your own palette to suit your taste!)

The Biplot graph provides information on

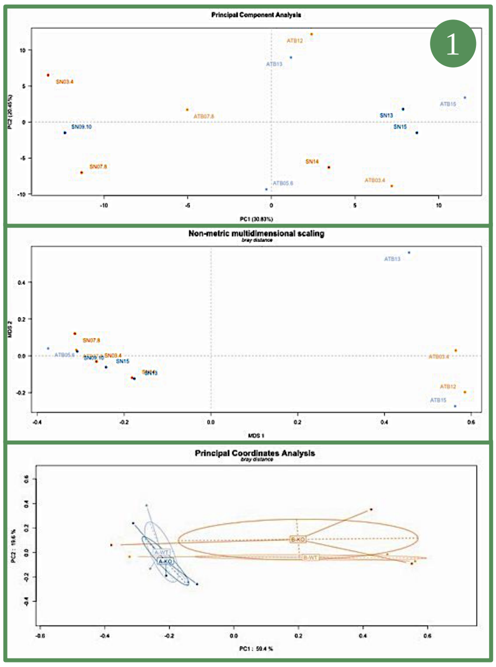
- Combines the variables and the individuals graphs, showing both of them in the same space, facilitating their interpretation in relation to each other
- Understanding the relationships between variables and the clusters of individuals.
- A selection of variables corresponding to the contribution threshold chosen by the user (the threshold is put to 50% by default)

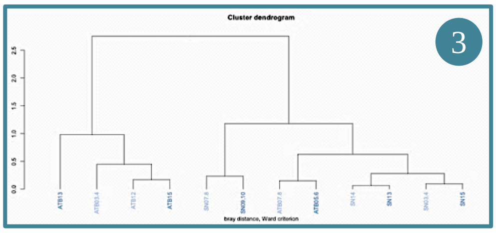
The variables graph provides information on

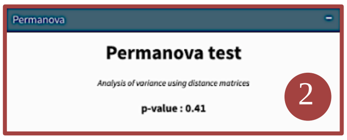
- Understanding of how each original variable contributes to the principal components.
- Identification of variables that have a strong influence on certain principal components.
- Inference about the relationships between variables and components.
- A selection of variables corresponding to the contribution threshold chosen by the user.

The contribution bar chart provides information on the contribution of variables. This overview is useful for determining precisely which variables contribute most.

Otherwise, PERMANOVA test is the same as the NMDS/PCoA one, but it is realized with Euclidean distance only.






Permanova

Permanova test

Analysis of variance using distance matrices

p-value : 0.41

1 Representation of the 2 first axes for 3 ordination method (PCA, NMDS et PCoA) according to the user selected variables of interest.

2 Permanova test result. This test is based on the distance between points (not available for the PCA)

3 Hierarchical clustering of the sample

Interpretation:

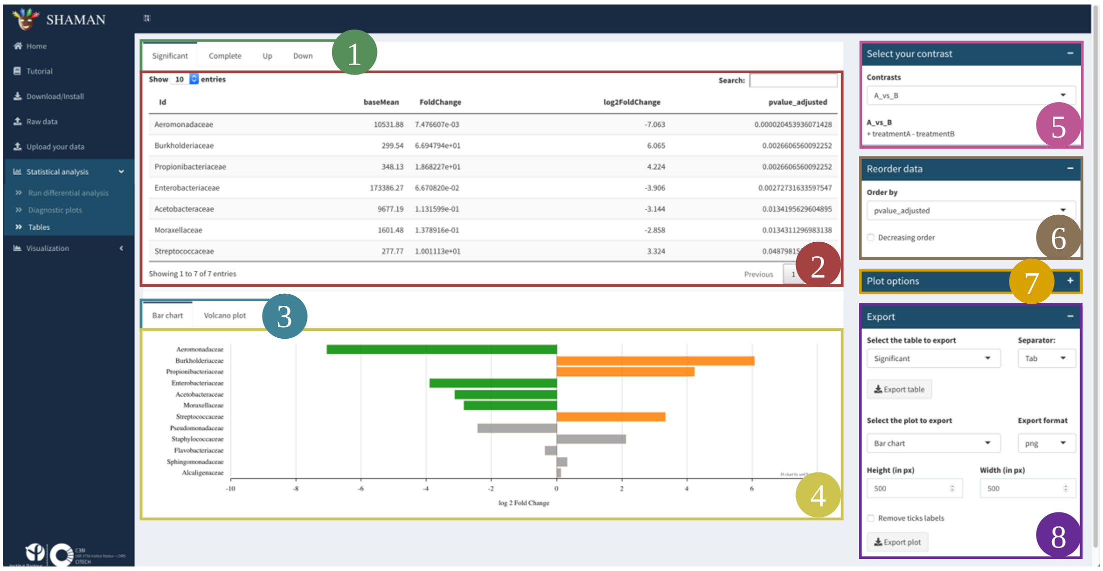
The ordination methods (PCA, NMDS and PCoA) have a double role, they allow to detect aberrant points or inversions of samples and to check if the studied biological effect brings a strong variability between the samples. In this example, we can see that the first axis of the PCoA (the one with the highest percentage of variance) allows to separate the 2 treatments A and B.

When samples are not sequenced in the same time, it is possible to see a "run" effect (samples would be clustered by sequencing batch). In this case, it is necessary to take this effect as an adjustment variable in the model (just add a "run" variable in the target file and incorporate it into the model). The experimental design must be considered before the analysis to avoid confounding effect (for example: to sequence all the samples of the treatment A together and, in a second time, those of the treatment B).

Regarding the p-value associated with the permanova test, it does not allow to determine that at least one of the groups is significantly different from the others (at a risk of 5%). When the test is significant, it means that at least one group is different from the others. When the biological effect studied is strong, the hierarchical classification makes it possible to identify possible inversion(s) of sample(s).

Differential analysis

Once the analysis is done, the results are available in tabular form in the "Table" tab.



Show 10 entries Search:

Id	baseMean	FoldChange	log2FoldChange	pvalue_adjusted
Aeromonadaceae	10531.88	7.476607e-03	-7.063	0.000020453936071428
Burkholderiaceae	299.54	6.694794e+01	6.065	0.0026606560092252
Propionibacteriaceae	348.13	1.868227e+01	4.224	0.0026606560092252
Enterobacteriaceae	173386.27	6.670820e-02	-3.906	0.00272731633597547
Acetobacteraceae	9677.19	1.131599e-01	-3.144	0.0134195629604895
Moraxellaceae	1601.48	1.378916e-01	-2.858	0.0134311296983138
Streptococcaceae	277.77	1.001113e+01	3.324	0.04879815

Showing 1 to 7 of 7 entries Previous 1

Contrasts: A_vs_B

A_vs_B
+ treatmentA - treatmentB

Order by: pvalue_adjusted

□ Decreasing order

Select the table to export: Significant

Separator: Tab

Select the plot to export: Bar chart

Export format: png

Height (in px): 500

Width (in px): 500

□ Remove ticks labels

1 4 tables are available:

- ✓ Significant: list of elements detected as differentially abundant
- ✓ Complete: list with all elements
- ✓ Up, down: lists of the elements detected as differentially abundant according to the sign of the log2 fold-change

2 Result table:

- ✓ baseMean: Average of normalized counts on all the samples
- ✓ FoldChange: Estimated fold change for selected contrast
- ✓ Log2FoldChange: Fold change transformed in log2
- ✓ Pvalue_adjusted: adjusted p-value (by default p-values are adjusted by BH procedure, other methods are available)

3 2 plots are available:

- ✓ Bar chart: log2 Fold Change for each element
- ✓ Volcano plot: p value versus fold change (scatterplot)

4 Area to display the plot

5 Selection of the desired comparison based on the user defined contrast vectors.

6 Reorder data: this will affect both the result tables and the bar chart.

7 List of the options to modify the appearance of the plot (depends on the selected plot)

8 Export results:

- ✓ For tables, the user can choose the table to export as well as the column separator.
- ✓ For plots, the user can select the format (png, svg, eps, pdf) as well as the size of the figure to be exported.

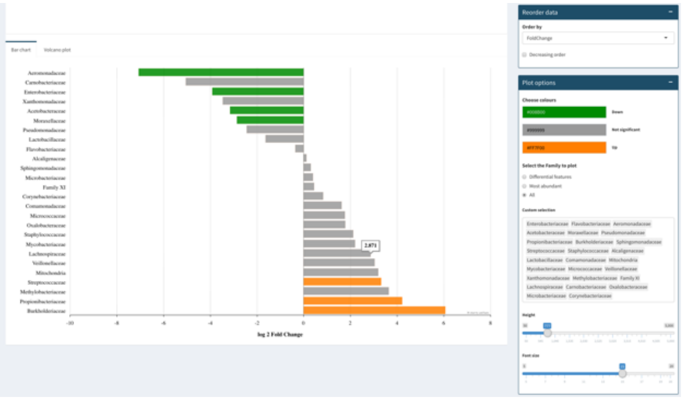
Interpretation:

The tables presented above allows to quickly identify the elements that have a significant change in term of abundance between the conditions studied. The presented p-values correspond to the p-values of the Wald test (after correction for multiple comparisons) at the threshold of significance chosen (e.g. 5%).

BaseMean informs about the mean abundance element (species, gender, ...).

Note: the log2FoldChange are obtained from the parameter estimates and then the values are shrunk toward 0 for features with a high variance (strongly impacts the low counts).

The bar chart and the volcano plot enable to view at a glance the data contained in the last two columns of the table (log2FoldChange and p value).



Reorder data

Order by: FoldChange

□ Decreasing order

Plot options

Choose colours

#008000 Down

#999999 Not significant

#FF7F00 Up

Select the Family to plot

○ Differential features

○ Most abundant

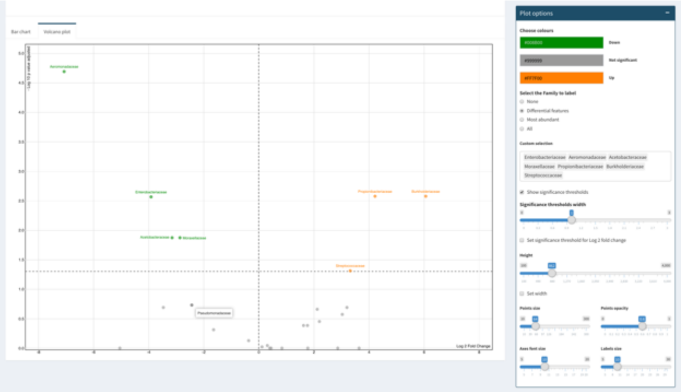
● All

Custom selection: Enterobacteriaceae Flavobacteriaceae Aeromonadaceae Acetobacteraceae Moraxellaceae Pseudomonadaceae Propionibacteriaceae Burkholderiaceae Sphingomonadaceae Streptococcaceae Staphylococcaceae Alcaligenaceae Lactobacillaceae Comamonadaceae Mitochondria Mycobacteriaceae Micrococcaceae Veillonellaceae Xanthomonadaceae Methylobacteriaceae Family XI Lachnospiraceae Carnobacteriaceae Oxalobacteraceae Microbacteriaceae Corynebacteriaceae

Height

Font size

Bar chart



Plot options

Choose colours

#008000 Down

#999999 Not significant

#FF7F00 Up

Select the Family to label

○ None

● Differential features

○ Most abundant

○ All

Custom selection: Enterobacteriaceae Aeromonadaceae Acetobacteraceae Moraxellaceae Propionibacteriaceae Burkholderiaceae Streptococcaceae

☑ Show significance thresholds

Significance thresholds width

□ Set significance threshold for Log 2 fold change

Height

□ Set width

Points size

Points opacity

Axes font size

Labels size

Volcano plot

Their appearance can be customized by the user in the “Plot options” box. Those settings (except the dimensions) will be kept if the plot is exported.

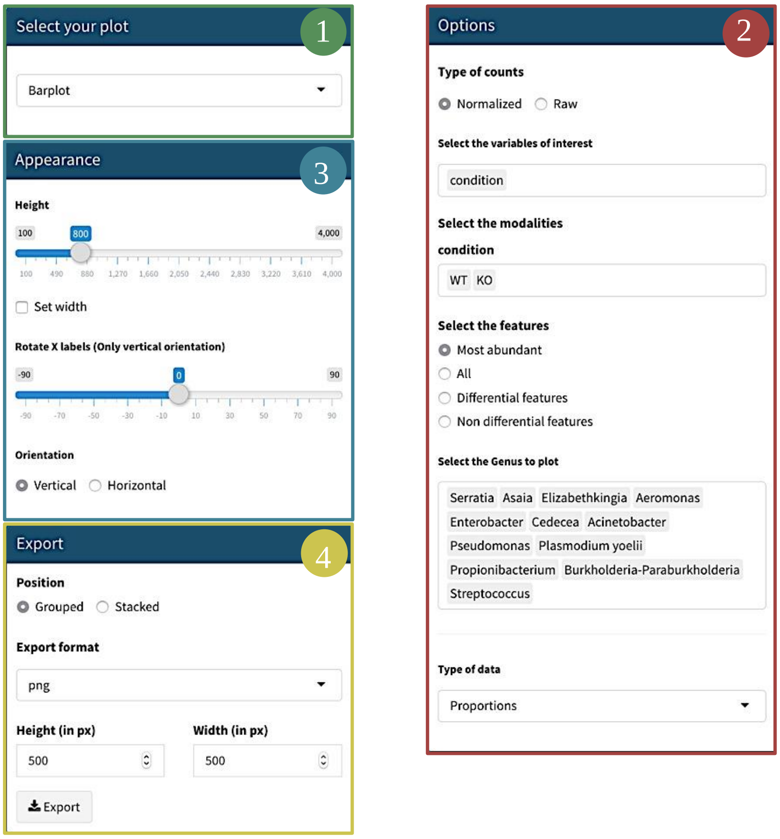
The color settings are shared by the plots, and the colors can be chosen either by selecting one in the color picker, or by typing a valid color name or hexadecimal color code.

For volcano plot, the threshold used for Y-axis is deduced from the significance threshold chosen for the differential analysis. The user can also define a threshold for X-axis, this will however affect only the volcano plot.

SHAMAN offers both a robust statistical analysis of data and a wide range of representations (barplots, heatmap, boxplots, abundance tree, scatterplot, diversity, rarefaction curves, krona and Venn diagrams) allowing a complete analysis of the results. They are divided into two categories:

- ✓ Global views: plots allowing to visualize feature abundance according to the variables defined in the target file.
- ✓ Comparison plots: specific visualization to compare results obtained for 2 or more contrast.

Regardless of the type of visualization used, SHAMAN offers many options to the user (which will remain active from one representation to another). These options are in a column to the right of the visualizations and consists of 4 elements.



Select your plot
Barplot

Appearance
Height: 800
Set width: □
Rotate X labels (Only vertical orientation): 0
Orientation: ● Vertical ○ Horizontal

Export
Position: ● Grouped ○ Stacked
Export format: png
Height (in px): 500
Width (in px): 500

Options
Type of counts: ● Normalized ○ Raw
Select the variables of interest: condition
Select the modalities
condition: WT KO
Select the features: ● Most abundant ○ All ○ Differential features ○ Non differential features
Select the Genus to plot: Serratia Asaia Elizabethkingia Aeromonas Enterobacter Cedecea Acinetobacter Pseudomonas Plasmodium yoelii Propionibacterium Burkholderia-Paraburkholderia Streptococcus
Type of data: Proportions

1 Drop-down menu to select the visualization.

- ✓ In "Global views" tab, 9 different visualizations are proposed: barplot, heatmap, boxplot, Tree, scatterplot, network, diversity, rarefaction, Krona
- ✓ In "Comparison plots" tab, 6 visualizations are proposed: Venn diagram, UpSet, contrasts comparison, heatmap, logit plot, density plot

2 List of available options for the selected visualization.

- ✓ For most of the visualizations in "Global views" tab, the user can select the data that he wants to represent (normalized or not), define the variables (and modalities) and select the elements to represent (the most abundant, all, only the differentially abundant or not for one of the defined comparisons).
- ✓ In the "Comparison plots" tab, it is necessary to select the contrasts to compare

3 List of the options to modify the appearance of the plot (depends on the selected visualization)

4 Exporting the results. The user can select the format (png, svg, eps, pdf) as well as the size of the figure to be exported.

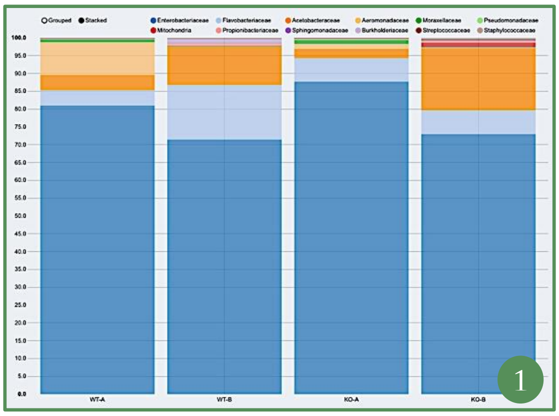
Visualizations of the results

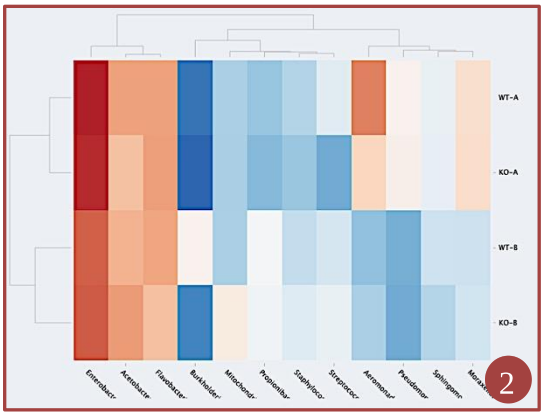
In the "Global views" tab, nine interactive visualizations are available to view the results.

Note: The interpretation of targeted metagenomics data must be done carefully. rRNA are present in several copy number in bacterial genomes [Vetrovsky and Baldrian 2013, Klappenbach 2001] from 1 to 15. In consequence, it is possible to analyze the abundance of one given genus between conditions but not to compare the abundance of two genera.

Overall composition

Barplot or heatmap are ideal visualizations for the analysis of the whole study. These representations make it possible to quickly visualize the differences between the conditions studied.

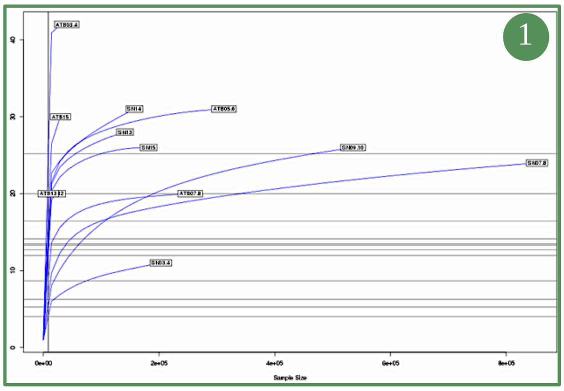


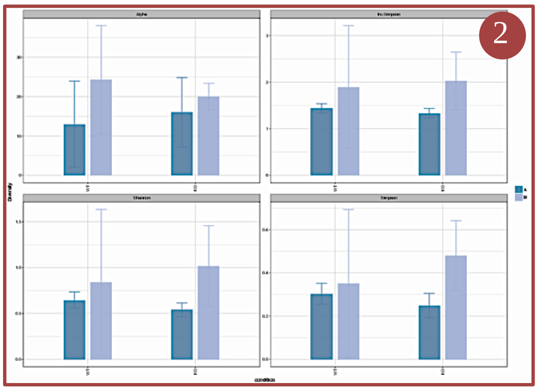


1 Barplot of the 12 most abundant elements. Each color represents the proportion of each element.

2 Heatmap for the 12 most abundant items per variable. The color depends on the average level of abundance. Dark red = very abundant; Dark blue = low abundance.

Interpretation:
The two previous figures show that the overall composition varies between the treatments (A and B) but not for the conditions (KO and WT) on the 12 most abundant. This observation is reinforced by the hierarchical clustering in the heatmap that groups the treatments between them.
The **rarefaction curves** and the **diversity measures** also provide an overview of the global composition according to the metadata.





1 Rarefaction curves for each sample. It corresponds to the number of elements detected according to the depth of sequencing.

2 Measure of diversity according to the variables of interest. SHAMAN offers 6 different measures: alpha, beta, gamma, shannon, simpson and inverse simpson.

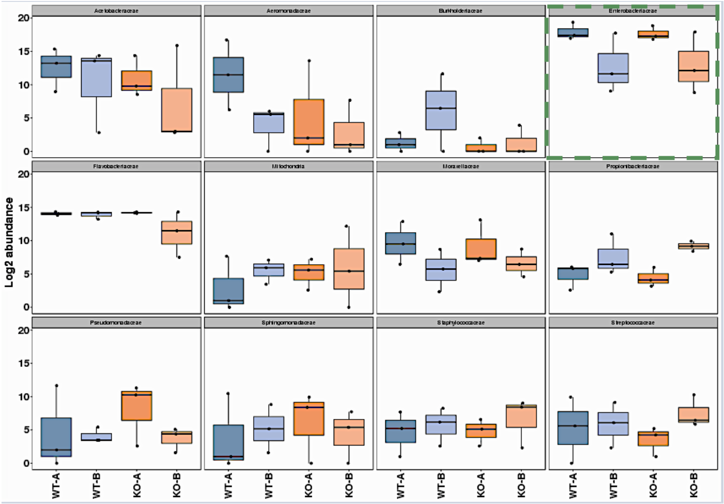
Interpretation:

The rarefaction curves highlight possible sequencing problems. The best case is to obtain rarefaction curves that all converge to a plateau, meaning that the sequencing depth was enough to discover, for instance, every genus that are present in the samples. In the example presented, the 3 samples closest to the ordinate axis show that the sequencing is not sufficient. In this case, it must be verified that normalization is not responsible for this result (by displaying the rarefaction curves for non-standardized data) and, if necessary, change the normalization method. If normalization is not responsible, it must be verified that there has been no problem upstream of the statistical analysis.

The diversities are useful to view the impact of the biological condition studied regarding to the composition of the samples. Confidence intervals make it possible to conclude whether the difference is significant or not (overlapping or not intervals). The user can also export the values used for the representations.

Fold-change

Boxplots are the usual representation to visualize the results of the differential analysis and log2 fold-change estimated.

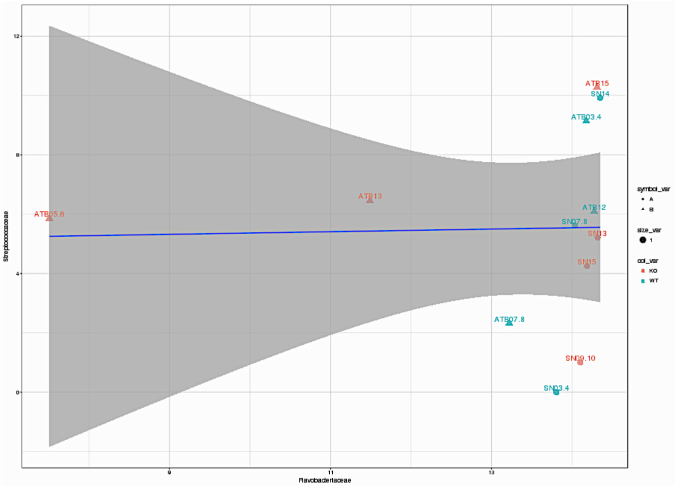


For each selected element, SHAMAN represents a boxplot for all the modalities of the variables of interest (with a different color). This representation highlights the differences identified during the differential analysis. For example, for the family Enterobacteriaceae (top right), it is clear that treatment B reduces the abundance of this family (whatever the condition). This effect was detected as significant in the result tables (3rd row of the table).

Links between variables

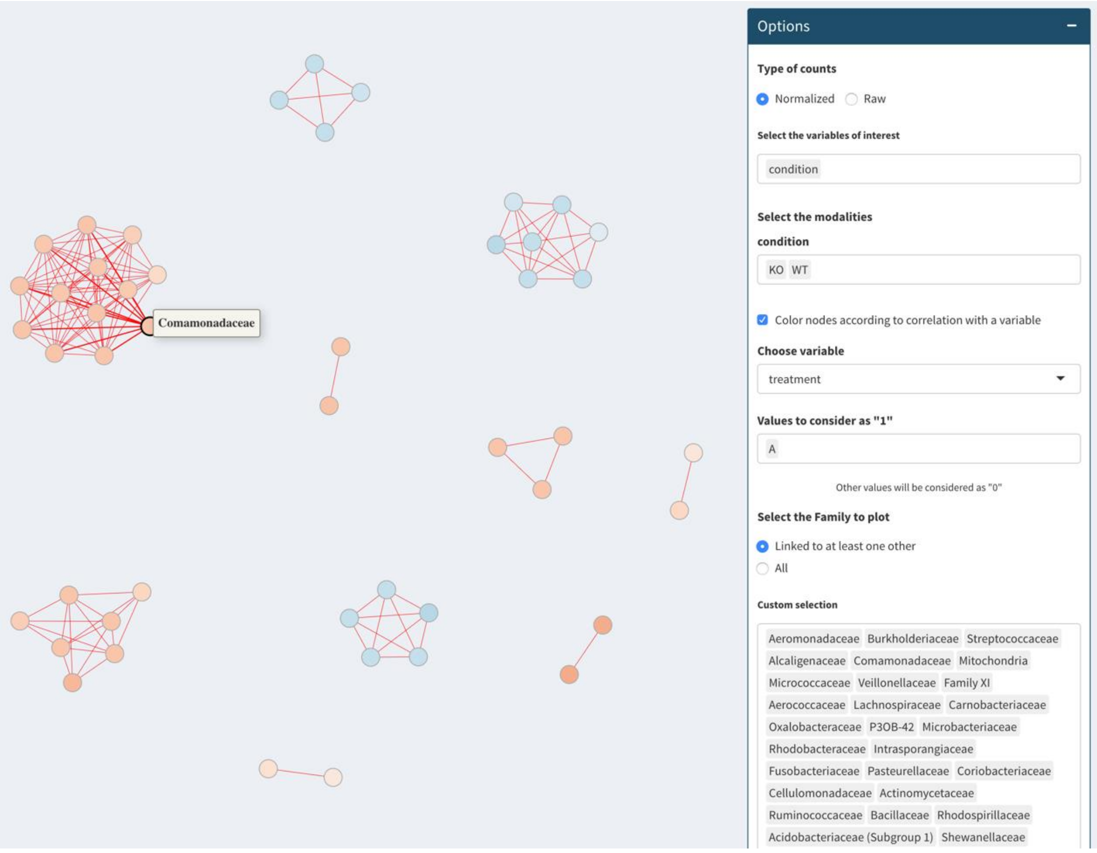
SHAMAN enables to cross abundance levels between two elements and to measure their correlation. It is also possible to model the abundance of a specie with respect to diversity or to an auxiliary variable like age, bmi, ... which can be informative especially for clinical dataset.

To realize these different crossings, SHAMAN proposes a **scatterplot** allowing to represent all the couples of elements, to change the points (form and color) according to the variables of interest and to add a third variable on which depends the size of the points.



The user can also add a linear regression (blue line) with confidence area (gray area). The equation of the line, the coefficient of determination R^2 and the tests for slope and intercept are then available. The user can also access to the correlation table (Pearson or Spearman).

Furthermore, it is possible to visualize the correlation between the respective abundance of all element at once thanks to the **network** plot. Each element is represented by a node linked to every other element to which he is significantly correlated (the color of the edge depending on the sign of the correlation).



Type of counts

● Normalized ○ Raw

Select the variables of interest

condition

Select the modalities

condition

KO WT

☑ Color nodes according to correlation with a variable

Choose variable

treatment

Values to consider as "1"

A

Other values will be considered as "0"

Select the Family to plot

● Linked to at least one other

○ All

Custom selection

Aeromonadaceae Burkholderiaceae Streptococcaceae Alcaligenaceae Comamonadaceae Mitochondria Micrococcaceae Veillonellaceae Family XI Aerococcaceae Lachnospiraceae Carnobacteriaceae Oxalobacteraceae P3OB-42 Microbacteriaceae Rhodobacteraceae Intrasporangiaceae Fusobacteriaceae Pasteurellaceae Coriobacteriaceae Cellulomonadaceae Actinomycetaceae Ruminococcaceae Bacillaceae Rhodospirillaceae Acidobacteriaceae (Subgroup 1) Shewanellaceae

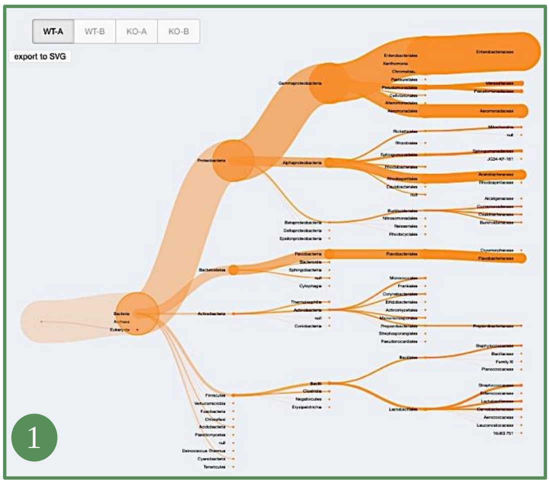
For this plot, correlations are computed using Pearson's method, and the p values are adjusted for multiple comparisons using Holm's method.

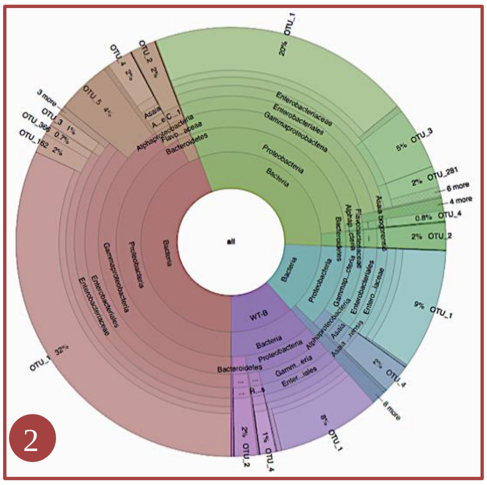
The user can select an auxiliary variable and the nodes will be colored according to correlation between abundance in each sample and the selected variable. If the chosen variable is not quantitative, it is necessary to select the values to consider as "1", among all the possible values of the qualitative variable. Other values will be set to "0".

This plot can be exported as *.html*. This will preserve interactivity (moving nodes, showing labels when hovering).

Abundance and taxonomy

All previous representations allow to visualize the results of the analysis at the level defined by the user during the definition of the statistical model. The two following figures give a different perspective.





1 Abundance tree according to the taxonomic level. The user can choose the samples they wish to represent.

2 Representation in the form of a krona plot[1]. This interactive visualization allows to navigate through different taxonomic levels.

[1] https://github.com/marbl/Krona/wiki

Result comparisons

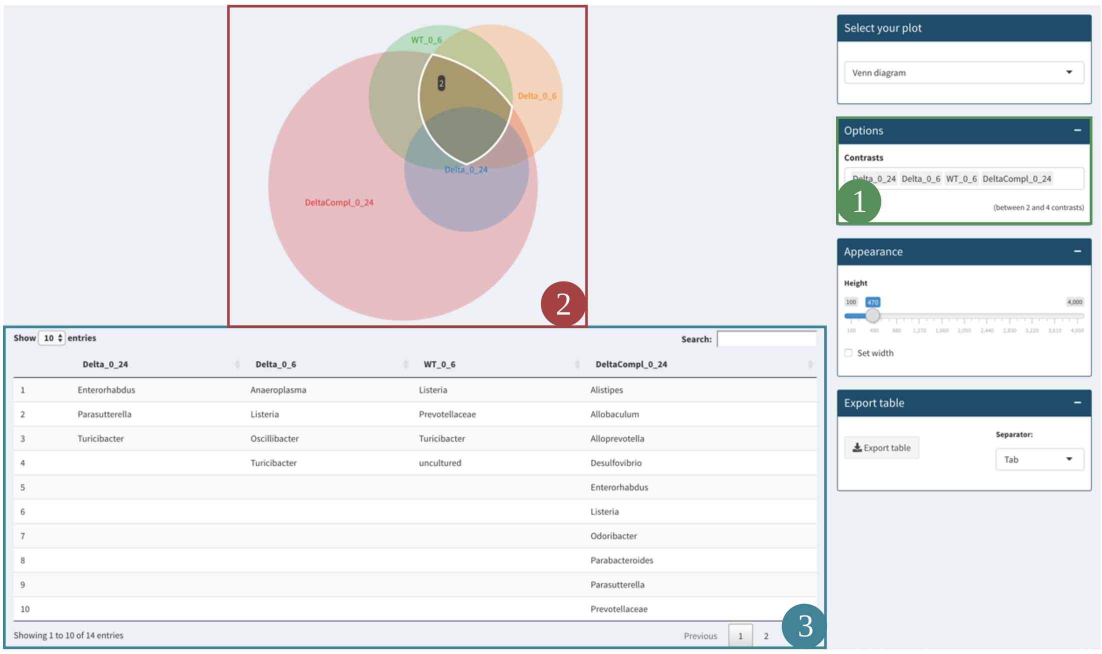
The main objective of SHAMAN is to perform a quantitative study of the differences between experimental conditions. A more qualitative approach is proposed in SHAMAN to compare the results of different studies. This is available through 6 representations in the “Comparison plots” tab:

- ✓ Venn diagram
- ✓ UpSet
- ✓ Contrasts comparison
- ✓ Heatmap
- ✓ Logit plot
- ✓ Density plot

This type of representation allows to compare the results of several studies from a qualitative point of view. In many cases, a quantitative analysis can also be performed (with 2 comparisons) by defining a suitable contrast ("advanced user" mode).

Comparison of the differentially abundant elements

The Venn diagram, the UpSet plot and the Contrasts comparison plot make a comparison of several contrasts based on the elements detected as differential for each contrast.



Venn diagram

Contrasts: Delta_0_24 Delta_0_6 WT_0_6 DeltaCompl_0_24 (between 2 and 4 contrasts)

Height: 100 470 4,000

□ Set width

Export table — Separator: Tab

Show 10 entries — Search:

	Delta_0_24	Delta_0_6	WT_0_6	DeltaCompl_0_24
1	Enterorhabdus	Anaeroplasma	Listeria	Alistipes
2	Parasutterella	Listeria	Prevotellaceae	Allobaculum
3	Turicibacter	Oscillibacter	Turicibacter	Alloprevotella
4		Turicibacter	uncultured	Desulfovibrio
5				Enterorhabdus
6				Listeria
7				Odoribacter
8				Parabacteroides
9				Parasutterella
10				Prevotellaceae

Showing 1 to 10 of 14 entries — Previous 1 2

Venn Diagram

1 Selection of contrasts that will be compared. At least two and at most four contrasts must be selected.

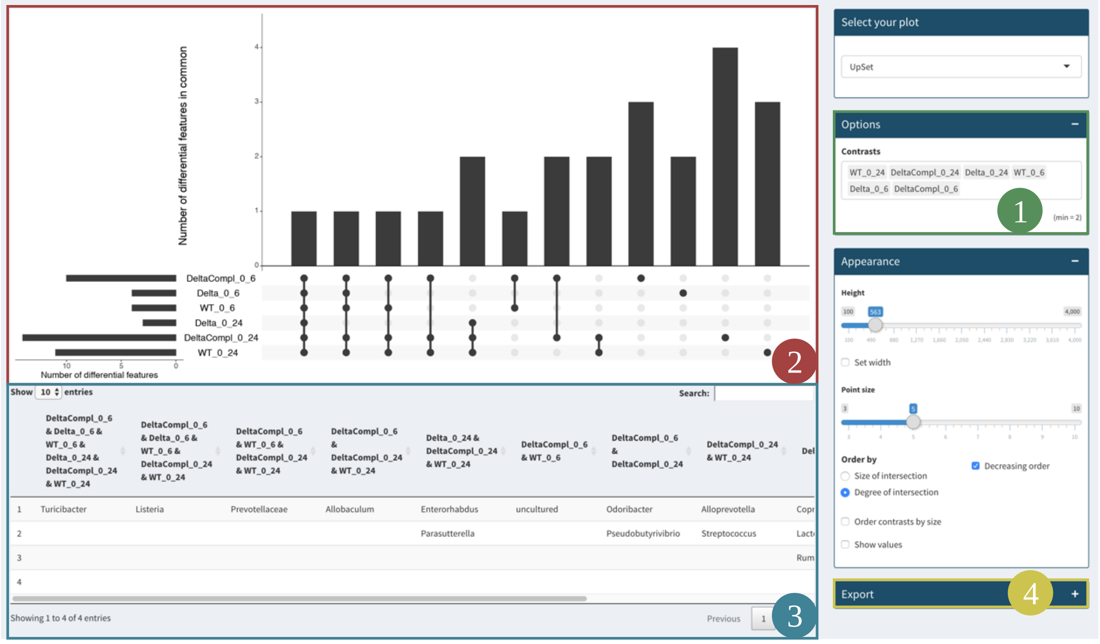
2 Venn diagram between differentially abundant elements for selected contrasts.

3 Lists elements for each contrast. This table can be exported in the “Export table” box.

Interpretation:

The Venn diagram provides the number of common elements between contrasts when hovering the intersections. Each circle is associated to one contrast.

The number of selected contrasts for the Venn diagram is limited to four. The **UpSet** plot offers a more scalable alternative to compare a larger number of contrasts.



Show 10 entries

Search:

	DeltaCompl_0_6 & Delta_0_6 & WT_0_6 & Delta_0_24 & DeltaCompl_0_24 & WT_0_24	DeltaCompl_0_6 & Delta_0_6 & WT_0_6 & DeltaCompl_0_24 & WT_0_24	DeltaCompl_0_6 & WT_0_6 & DeltaCompl_0_24 & WT_0_24	DeltaCompl_0_6 & DeltaCompl_0_24 & WT_0_24	Delta_0_24 & DeltaCompl_0_24 & WT_0_24	DeltaCompl_0_6 & WT_0_6	DeltaCompl_0_6 & DeltaCompl_0_24	DeltaCompl_0_24 & WT_0_24	Del
1	Turicibacter	Listeria	Prevotellaceae	Allobaculum	Enterorhabdus	uncultured	Odoribacter	Alloprevotella	Copr
2					Parasutterella		Pseudobutyrivibrio	Streptococcus	Lact
3									Rum
4									

Showing 1 to 4 of 4 entries

Previous 1

Select your plot

UpSet

Options −

Contrasts

WT_0_24 DeltaCompl_0_24 Delta_0_24 WT_0_6 Delta_0_6 DeltaCompl_0_6

(min = 2)

Appearance −

Height

100 563 4,000

100 490 880 1,270 1,660 2,050 2,440 2,830 3,220 3,610 4,000

□ Set width

Point size

3 5 10

3 4 5 6 7 8 9 10

Order by

○ Size of intersection

● Degree of intersection

☑ Decreasing order

□ Order contrasts by size

□ Show values

Export +

UpSet

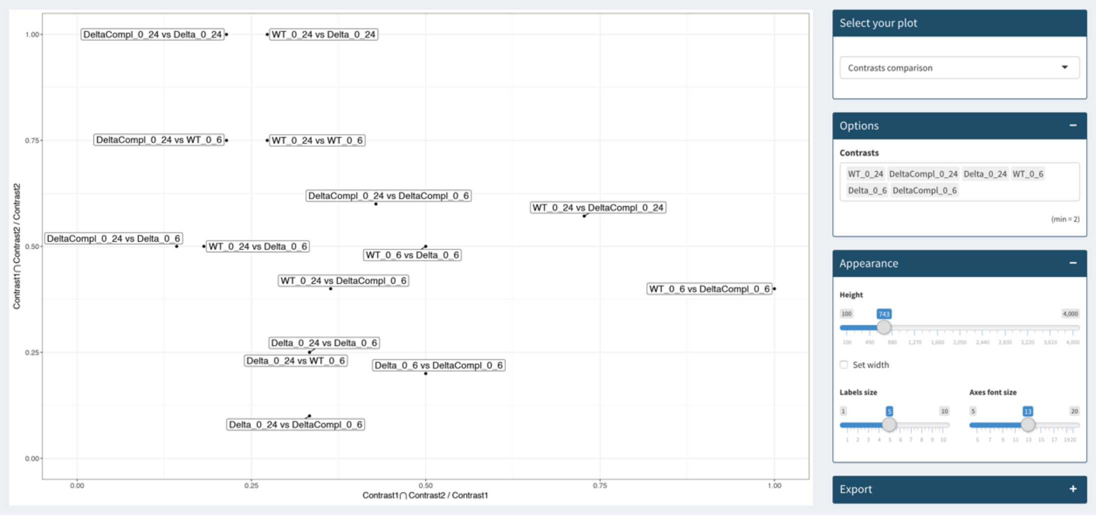
1 Selection of contrasts that will be compared. At least two and at most four contrasts must be selected.

2 UpSet plot: Each bar represents the number of elements that are differential for all the contrasts marked with a dot below, and are not differential for any other contrasts.

3 Table: lists the elements in each intersection.

4 To export the plot or the table

The following plot, named “**Contrasts comparison**” compare contrasts two by two.


DeltaCompl_0_24 vs Delta_0_24
WT_0_24 vs Delta_0_24
DeltaCompl_0_24 vs WT_0_6
WT_0_24 vs WT_0_6
DeltaCompl_0_24 vs DeltaCompl_0_6
WT_0_24 vs DeltaCompl_0_24
DeltaCompl_0_24 vs Delta_0_6
WT_0_24 vs Delta_0_6
WT_0_6 vs Delta_0_6
WT_0_24 vs DeltaCompl_0_6
WT_0_6 vs DeltaCompl_0_6
Delta_0_24 vs Delta_0_6
Delta_0_24 vs WT_0_6
Delta_0_6 vs DeltaCompl_0_6
Delta_0_24 vs DeltaCompl_0_6
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Contrast1⋂Contrast2 / Contrast2
Contrast1⋂Contrast2 / Contrast1
Select your plot
Options
Appearance
Export

Contrasts comparison ▾

Contrasts

WT_0_24 DeltaCompl_0_24 Delta_0_24 WT_0_6 Delta_0_6 DeltaCompl_0_6

(min = 2)

Height: 743 (100 – 4,000)

□ Set width

Labels size: 5 (1 – 10)

Axes font size: 13 (5 – 20)

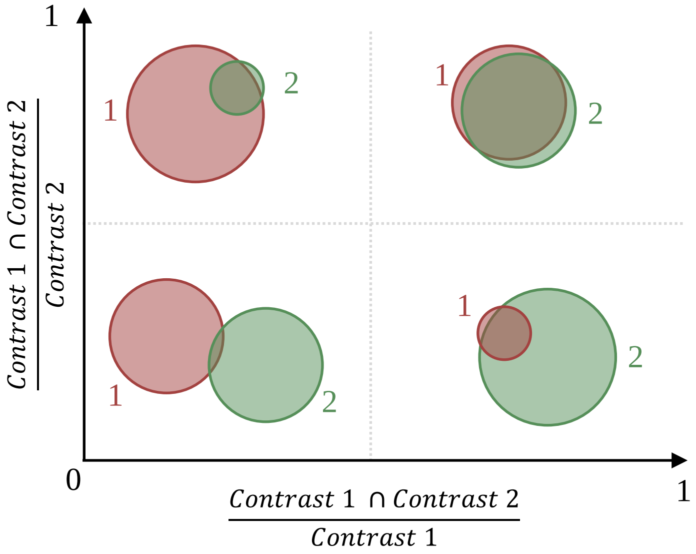
Contrasts comparison

Interpretation:

For each point, reading the label from left to right, the first contrast is called “Contrast 1” and the second “Contrast 2”.

The coordinates of the point are the number of elements that are differentially abundant for both contrasts, divided by the total number of differential elements for “Contrast 1” and “Contrast 2” for X-axis and Y-axis, respectively.

The following figure should help to visualize the meaning of coordinates on the plot.


1
0
1
$\frac{Contrast\ 1\ \cap Contrast\ 2}{Contrast\ 2}$
$\frac{Contrast\ 1\ \cap Contrast\ 2}{Contrast\ 1}$
1
2
1
2
1
2
1
2

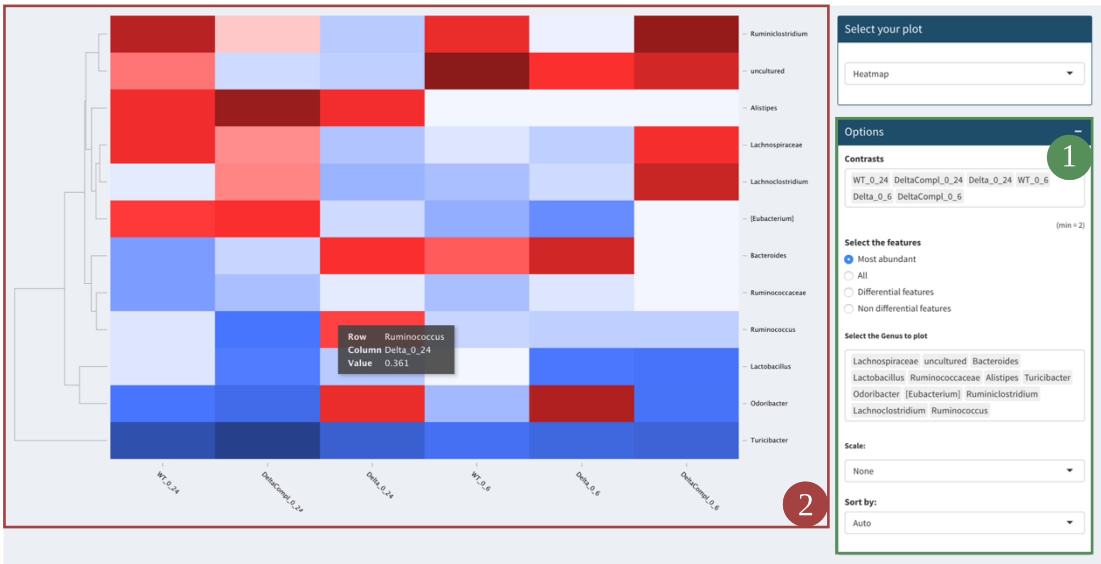
A point near the diagonal indicates that the contrasts involved have a similar number of differentials elements.

A point above the diagonal indicates that “Contrast 1” has more differential elements than “Contrast 2” (the opposite for a point under the diagonal).

A point close to the origin indicates few differential elements in common (the opposite for a point far from the origin).

Comparison of fold change

The **heatmap** of the log2 fold change illustrates the strength of the difference between several contrasts.



Row Ruminococcus
Column Delta_0_24
Value 0.361

Select your plot

Heatmap

Options

Contrasts

WT_0_24 DeltaCompl_0_24 Delta_0_24 WT_0_6 Delta_0_6 DeltaCompl_0_6

(min = 2)

Select the features

● Most abundant
○ All
○ Differential features
○ Non differential features

Select the Genus to plot

Lachnospiraceae uncultured Bacteroides Lactobacillus Ruminococcaceae Alistipes Turicibacter Odoribacter [Eubacterium] Ruminiclostridium Lachnoclostridium Ruminococcus

Scale:

None

Sort by:

Auto

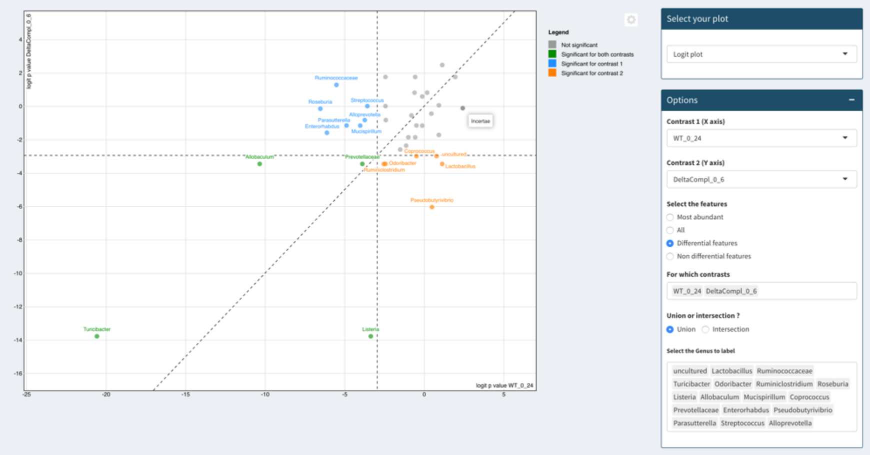
1 List of the different options proposed to draw the heatmap.

2 Representation of log2 fold changes as a heatmap. The color depends on the value of the log2 fold change. Dark red: high positive value; Dark blue: high negative value; White: null value.

Interpretation:
This representation allows to identify common elements between 2 or more contrasts. The shade of color gives the direction and the strength of the difference.

Comparison of p value

The **logit plot** and the **density plot** focus on the p values.



Select your plot

Logit plot ▾

Options −

Contrast 1 (X axis)

WT_0_24 ▾

Contrast 2 (Y axis)

DeltaCompl_0_6 ▾

Select the features

○ Most abundant
○ All
● Differential features
○ Non differential features

For which contrasts

WT_0_24 DeltaCompl_0_6

Union or intersection ?

● Union ○ Intersection

Select the Genus to label

uncultured Lactobacillus Ruminococcaceae Turicibacter Odoribacter Ruminiclostridium Roseburia Listeria Allobaculum Mucispirillum Coprococcus Prevotellaceae Enterorhabdus Pseudobutyrivibrio Parasutterella Streptococcus Alloprevotella

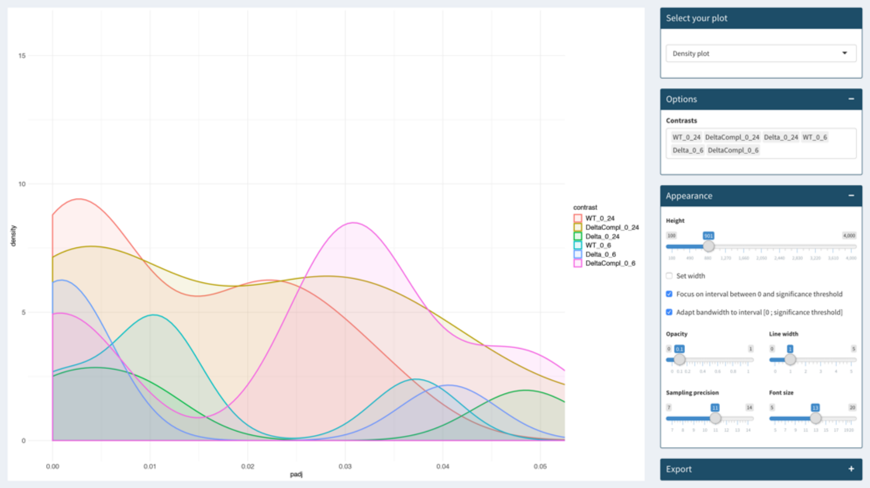
Logit plot

The **logit plot** compare two contrasts. Each element is represented by a point whose coordinates are the p values in the two contrasts, transformed by logit function.

$$logit(p) = \ln(\frac{p}{1-p})$$

Vertical and horizontal dashed lines represent the threshold of significance chosen for the differential analysis.

A point near the diagonal indicates a similar level of significance of the element for both contrasts.



Select your plot

Density plot ▾

Options −

Contrasts

WT_0_24 DeltaCompl_0_24 Delta_0_24 WT_0_6 Delta_0_6 DeltaCompl_0_6

Appearance −

Height: 901

□ Set width

☑ Focus on interval between 0 and significance threshold

☑ Adapt bandwidth to interval [0 ; significance threshold]

Opacity: 0.1

Line width: 1

Sampling precision: 11

Font size: 13

Export +

Density plot

The **density plot** compares several contrasts and represents the distribution of p value. By default, it focuses on the interval between 0 and the significance threshold which enables to see, overall, how far from the threshold the differential elements are.

BIBLIOGRAPHY

Anders and Huber, Differential expression analysis for sequence count data, *Genome Biology*, 2010

Jonsson V, Österlund T, Nerman O et al. Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. *BMC Genomics* 2016; 17: 78

Love, Huber and Anders, Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2, *Genome Biology*, 2014

McMurdie, P. J. & Holmes, S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLOS Comput. Biol.* 10, e1003531 (2014)

How to create a target file for Shaman

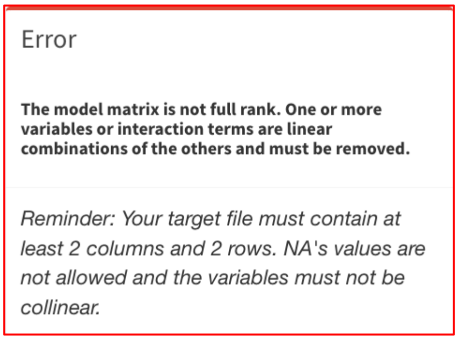
format, contents and constraints

The target file is required for each analysis done with Shaman. It must contain all the available information on the samples (corresponding to the metadata) that will be used to build the statistical model and/or to visualize the data. To be loaded in Shaman the target file must respect some properties

1. **The first column is dedicated to the sample name which must correspond exactly to the column names of the count matrix. At least 2 samples are required.** *Once the count matrix is loaded, check carefully the sample names in the “count table” tab. Sometimes some characters are modified with the loading.*
2. **At least one variable must be provided.** *In Example 1, two variables are provided (condition and treatment).*
3. **NA or missing values are not allowed.**
4. **A variable with the same value for each sample is not allowed**. *This kind of variable should be removed from the target file before loading.*
5. **The selected variables for the statistical model must not be collinear**. *It means that if one variable can be determined by another variable or a combination of variables the analysis cannot be done with all the variables. However, the user can user will be able to use this variable for visualization. (See example 3).*
6. **Be careful, numeric variables will be considered as quantitative variable.** *For instance, do not use 1 and 2 to describe two different conditions but C1 and C2 or A and B. (see example 3)*
7. **Avoid using special characters such as** _/\?*:<>|+,[]-+()’%@"&_

Example 1: Target file with 2 variables (condition and treatment)

sampleID	condition	treatment
S1	WT	A
S2	WT	A
S3	KO	A
S4	KO	A
S5	WT	B
S6	WT	B
S7	KO	B
S8	KO	B



Error

The model matrix is not full rank. One or more variables or interaction terms are linear combinations of the others and must be removed.

Reminder: Your target file must contain at least 2 columns and 2 rows. NA's values are not allowed and the variables must not be collinear.

This is a usual example in which we have 2 variables to describe the samples (condition and treatment). For instance, the user will be able to define the following model: condition + treatment + condition:treatment and then get differentially abundant features between treatments A and B for each condition.

Example 2: Target file with collinearity problem

sampleID	condition	treatment	group
S1	WT	A	g1
S2	WT	A	g1
S3	KO	A	g2
S4	KO	A	g2
S5	WT	B	g3
S6	WT	B	g3
S7	KO	B	g4
S8	KO	B	g4

In this example, group = condition + treatment, so the variables are collinear.
Note that this file can be loaded in Shaman without error but the error will appear if the user tries to define a model with the three variables condition, treatment and group.

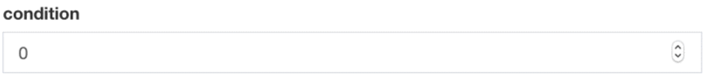
Example 3: Quantitative versus qualitative variable.

Target file

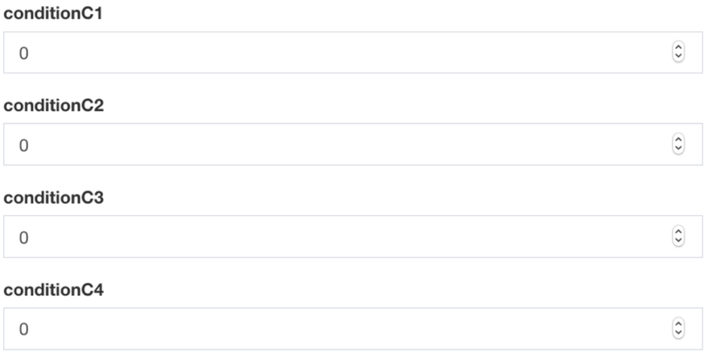
sampleID	condition
S1	1
S2	1
S3	2
S4	2
S5	3
S6	3
S7	4
S8	4

sampleID	condition
S1	C1
S2	C1
S3	C2
S4	C2
S5	C3
S6	C3
S7	C4
S8	C4

Model parameters



condition
0



conditionC1
0

conditionC2
0

conditionC3
0

conditionC4
0

In case 1, condition is considered as a numeric variable which leads to only one parameter in the statistical model. It assumes that the difference between 1 and 3 is two times the difference between 1 and 2 and so on. In case 2, there is no order between the conditions.