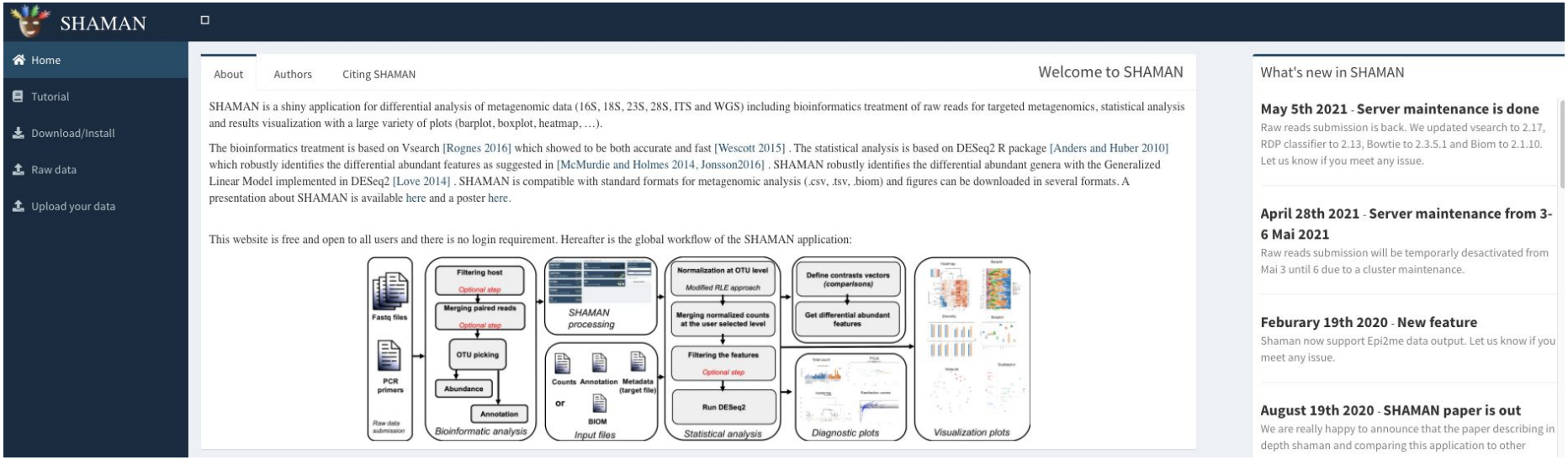
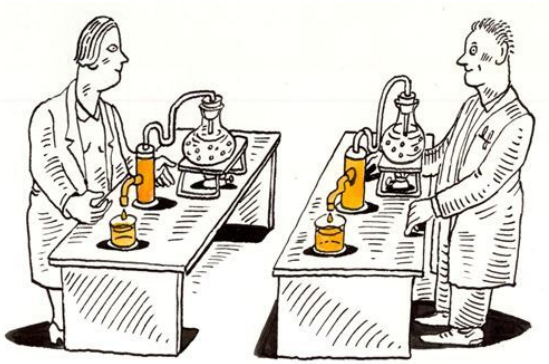
SHAMAN : SHiny Application for Metagenomic ANalysis

Amine Ghozlane
Hub Bioinformatique et Biostatistique

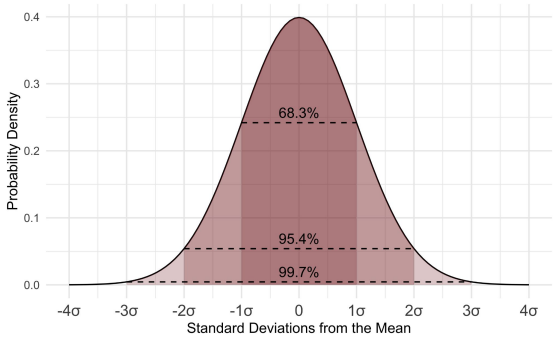
Subjects



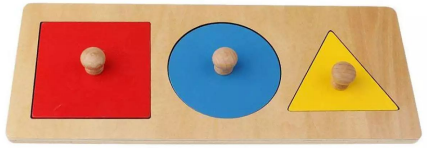
Reproducible research



Better statistics

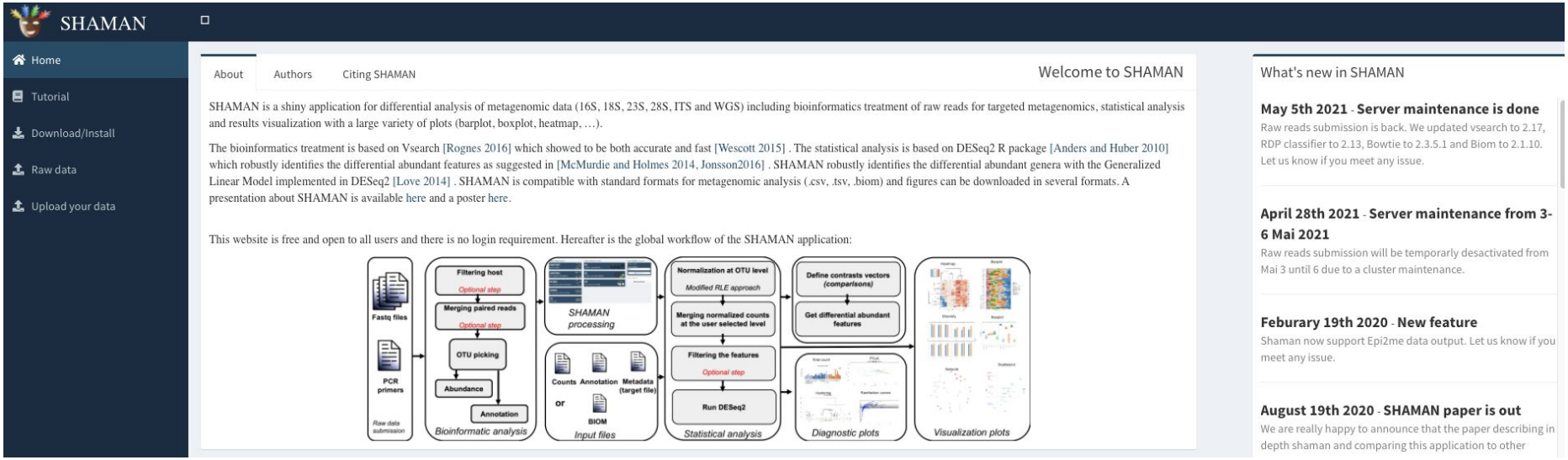


Simplicity

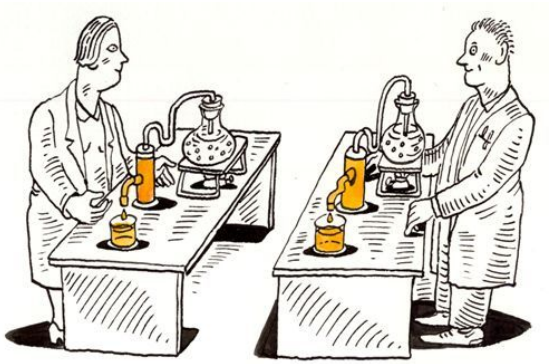


Institut Pasteur

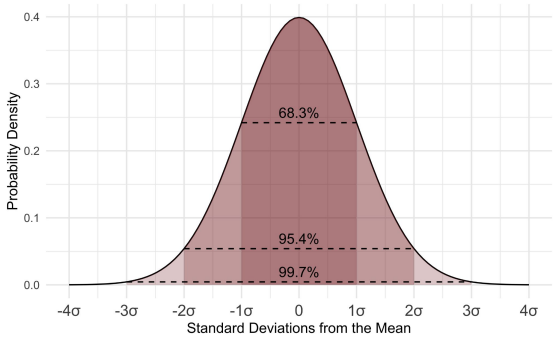
Subjects



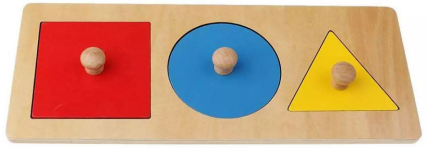
Reproducible research



Better statistics



Simplicity



Institut Pasteur

Definitions

- Reliability
  - evidence can be interpreted honestly with known operating characteristics
- Reproducibility
  - Same data + same analysis = same evidence
- Replicability
  - same data + different analysis = similar evidence?
  - different data + same analysis = similar evidence?
  - different data + different analysis = similar evidence?

Patrick Ryan, Janssen Research and Development

Institut Pasteur

Spot 5 issues for reproducibility

```
## Look at Picrust output data using R

rm(list=ls())

library(phyloseq)
library(microbiome)
library(devtools)
library(RJSONIO)
library(qualpalr)
library(plyr)
library(car)
library(ecodist)
library(ade4)
library(vegan)
library(permute)
library(lattice)
library(reshape2)
library(dplyr)
library(MASS)
library(Hmisc)
library(RColorBrewer)
library(ggplot2)
library(colorspace)
library(qualpalr)
library(nlme)
```

- Download Dataset_S14 (PDF)

Stunted childhood growth is associated with decompartmentalization of the gastrointestinal tract and overgrowth of oropharyngeal taxa

Pascale Vonaesch[a,b], Evan Morien[c,d,e], Lova Andrianonimiadana[f], Hugues Sanke[g], Jean-Robert Mbecko[g], Kelsey E. Huus[h], Tanteliniaina Naharimanananirina[i], Bolmbaye Privat Gondje[j], Synthia Nazita Nigatoloum[j], Sonia Sandrine Vondo[j], Jepthé Estimé Kaleb Kandou[k], Rindra Randremanana[l], Maheninasy Rakotondrainipiana[l], Florent Mazel[c,d,e], Serge Ghislain Djorie[k], Jean-Chrysostome Gody[j], B. Brett Finlay[h,1], Pierre-Alain Rubbo[g,1], Laura Wegener Parfrey[c,d,e,1], Jean-Marc Collard[f], Philippe J. Sansonetti[a,b,m,2], and The Afribiota Investigators[3]

[a]Unité de Pathogénie Microbienne Moléculaire, Institut Pasteur, 75015 Paris, France; [b]Unité INSERM 1202, Institut Pasteur, 75015 Paris, France; [c]Department of Botany, University of British Columbia, Vancouver, BC V6T 1Z4, Canada; [d]Department of Zoology, University of British Columbia, Vancouver, BC V6T 1Z4, Canada; [e]Biodiversity Research Centre, University of British Columbia, Vancouver, BC V6T 1Z4, Canada; [f]Unité de Bactériologie Expérimentale, Institut Pasteur de Madagascar, BP 1274 Ambatofotsikely, 101 Antananarivo, Madagascar; [g]Laboratoires d'Analyses Médicales, Institut Pasteur de Bangui, BP 923 Bangui, Central African Republic; [h]Michael Smith Laboratories, University of British Columbia, Vancouver, BC V6T 1Z4, Canada; [i]Centre Hospitalier Universitaire Joseph Ravoahangy Andrianavalona, BP 4150, 101 Antananarivo, Madagascar; [j]Complexe Pédiatrique de Bangui, BP 923 Bangui, Central African Republic; [k]Unité d'Epidémiologie, Institut Pasteur de Bangui, BP 923 Bangui, Central African Republic; [l]Unité d'Epidémiologie et de Recherche Clinique, Institut Pasteur de Madagascar, BP 1274 Ambatofotsikely, 101 Antananarivo, Madagascar; and [m]Chaire de Microbiologie et de Maladies Infectieuses, Collège de France, 75231 Paris, France

I "pre-reviewed" this paper -> Acknowledgement

```
#### set working directory ####
setwd("/Users/Pascale/Desktop/PNAS/R_files") # put here your path to the working directory

#### import the files obtained from Categorize by function (tax_table and otu_table or directly the biom) and merge with mapping file, then perform linear
 mixed models ####

otu=read.table(file="Additional Data Table19 otutablepicrust.txt", header=TRUE, sep = "\t", row.names=1, dec = ".")
otu=as.matrix(otu)
class(otu)
OTU = otu_table(otu, taxa_are_rows = TRUE)
OTU


map= read.table(file="Additional Data Table20 sampeldatapicrust.txt", header=TRUE, sep = "\t", row.names=1, dec = ".")
map
map =sample_data(map)

Aj_f_path3 <- merge_phyloseq(map,OTU)
```

Institut Pasteur

Spot 5 issues for reproducibility

```
## Look at Picrust output data using R

rm(list=ls())

library(phyloseq)
library(microbiome)
library(devtools)
library(RJSONIO)
library(qualpalr)
library(plyr)
library(car)
library(ecodist)
library(ade4)
library(vegan)
library(permute)
library(lattice)
library(reshape2)
library(dplyr)
library(MASS)
library(Hmisc)
library(RColorBrewer)
library(ggplot2)
library(colorspace)
library(qualpalr)
library(nlme)

#### set working directory ####
setwd("/Users/Pascale/Desktop/PNAS/R_files") # put here your path to the working directory

#### import the files obtained from Categorize by function (tax_table and otu_table or directly the
 mixed models ####

otu=read.table(file="Additional Data Table19 otutablepicrust.txt", header=TRUE, sep = "\t", row.nam
otu=as.matrix(otu)
class(otu)
OTU = otu_table(otu, taxa_are_rows = TRUE)
OTU
```

- Download Dataset_S14 (PDF)

1 - The script is shared as a PDF of 58 pages

2 - R is an hostile environment for reproducibility.

R-4.0.0.tar.gz	2020-04-24 09:05	32M
R-4.0.1.tar.gz	2020-06-06 09:05	32M
R-4.0.2.tar.gz	2020-06-22 09:05	32M
R-4.0.3.tar.gz	2020-10-10 09:05	32M
R-4.0.4.tar.gz	2021-02-15 09:05	32M
R-4.0.5.tar.gz	2021-03-31 09:05	31M
R-4.1.0.tar.gz	2021-05-18 09:05	32M

3 - “R_files” is not available !

```
Package: vegan
Title: Community Ecology Package
Version: 2.6-0
Author: Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt,
   Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara,
   Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs,
   Helene Wagner
Maintainer: Jari Oksanen <jhoksane@gmail.com>
Depends: permute (>= 0.9-0), lattice, R (>= 3.4.0)
Suggests: parallel, tcltk, knitr, markdown
Imports: MASS, cluster, mgcv
VignetteBuilder: utils, knitr
Description: Ordination methods, diversity analysis and other
  functions for community and vegetation ecologists.
License: GPL-2
BugReports: https://github.com/vegandevs/vegan/issues
URL: https://github.com/vegandevs/vegan
```

4 - Space in names

Institut Pasteur

Spot 5 issues for reproducibility

```
## Look at Picrust output data using R

rm(list=ls())

library(phyloseq)
library(microbiome)
library(devtools)
library(RJSONIO)
library(qualpalr)
library(plyr)
library(car)
library(ecodist)
library(ade4)
library(vegan)
library(permute)
library(lattice)
library(reshape2)
library(dplyr)
library(MASS)
library(Hmisc)
library(RColorBrewer)
library(ggplot2)
library(colorspace)
library(qualpalr)
library(nlme)

#### set working directory ####
setwd("/Users/Pascale/Desktop/PNAS/R_files") # put here your path to the working dir

#### import the files obtained from Categorize by function (tax_table and otu_table
 mixed models ####

otu=read.table(file="Additional Data Table19 otutablepicrust.txt", header=TRUE, sep
otu=as.matrix(otu)
class(otu)
OTU = otu_table(otu, taxa_are_rows = TRUE)
OTU
```

- Few parameters described
- Several software to install

5 - How do I obtain otutablepicrust ?

Bioinformatic and Biostatistic Analysis. Retrieved sequences were demultiplexed in QIIME v1.9 (42) and then trimmed, clipped, and quality-filtered using the Fastx Toolkit (hannonlab.cshl.edu/fastx_toolkit) to 245 bp with a minimum quality threshold of Q19. Filtered R1 reads were processed into operational taxonomic units (OTUs) using minimum entropy decomposition (MED) (43) with the minimum substantive abundance (-m) parameter set to 250, yielding 2,246 unique OTUs. Taxonomy was then assigned to the representative sequence for each MED node by matching it to the SILVA 128 (44, 45) database using QIIME. Singlets, mitochondrial, and cholorplast reads were filtered out. The final filtered OTU table consisted of 2,029 unique sequences and 9,155,211 reads. The stunted vs. nonstunted groups were compared using Pearson's χ^2 test or Fisher's exact test for qualitative variables and the student *t* test or the Mann–Whitney *U* test for quantitative variables. Statistical analyses and visualizations of the microbial data were conducted in R v3.4.1 using PhylosEq (46), vegan (47), randomForest (48, 49), DeSeq2 (50–52), and ggplot2 (53) R packages. α-Diversity was quantified using a measure of richness [Chao1 index (54)] and a measure of evenness (Simpson's diversity index = 1 − Simpson's index) while β-diversity was quantified using the Bray–Curtis dissimilarity index (55). Tests of differences in α-diversity between samples were performed using nonparametric multivariate analysis of variance (PERMANOVA) with the function "adonis" in the R package vegan (47, 56) or linear-mixed models (α value of 0.05). *P* values were Benjamini–Hochberg-corrected. Multivariate analyses of differentially abundant taxa were performed on pooled samples from both countries as well as on data from each country independently. Multivariate models were corrected for gender, age (in months), as well as country of origin and stratified on sample type, and then on country of origin. Picrust analysis (57) was performed on the Galaxy server of the Langille group (galaxy.morganlangille.com). The gene counts were categorized by function and rarefied to 2,000,000 gene counts. The differential gene count was analyzed by linear-mixed models correcting for gender and age (in months), as well as country of origin. The metadata, OTU table, taxonomy table, R code, and a detailed description of the methods can be found in *SI Appendix*.

Take home messages

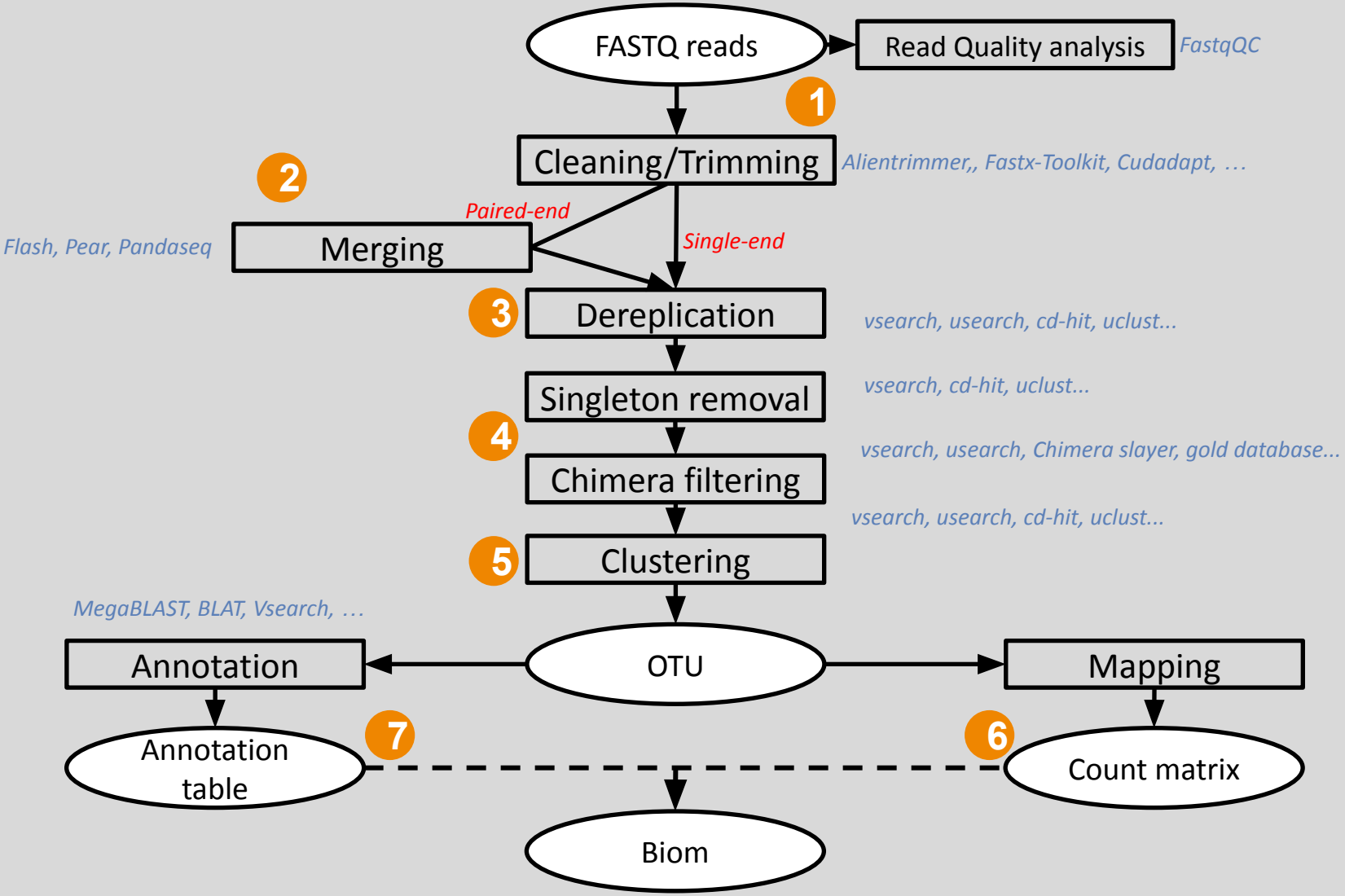
In R, a snapshot of the environment is mandatory for reproducible research: Packrat, Docker

Material and methods are short and incomplete, parameters should be easily accessible

An unified environment is more likely to open reproducibility

Reproducibility vs updates is challenging

Targeted metagenomics pipeline



Targeted metagenomics pipeline

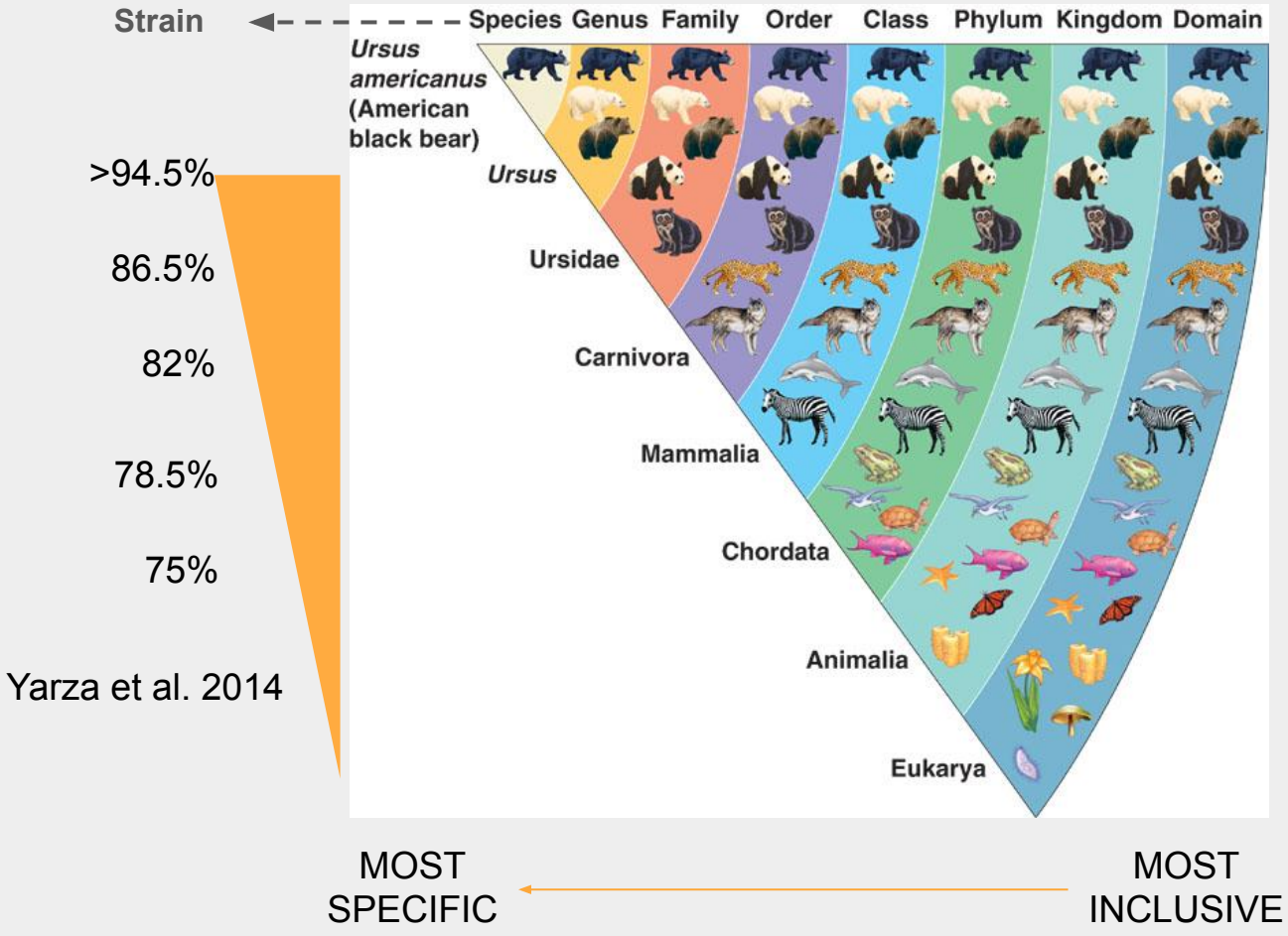
6) Mapping

		Individuals						
Reference catalogue	OTU	Ind 1	Ind 2	Ind 3	Ind 4	Ind 5	Ind 6	Ind 7
	1	0	36	2	0	43	106	1250
	2	0	27	193	0	44	103	8
	3	0	31	0	0	0	0	0
	4	152	59	282	1	0	0	0
	5	115	0	0	1	0	29	2
	6	90	783	26	0	2	0	0
	7	104	1616	0	0	0	0	5
	8	0	82	0	0	0	0	0
	9	2	0	0	0	0	0	0
	10	23	239	1302	10	0	190	0
	11	30	183	900	13	0	172	0
	12	27	228	1120	6	0	324	0
	13	103	0	0	0	0	0	0
	14	0	30	269	0	0	0	0
	15	0	0	0	0	0	95	0
	16	1250	6002	468	607	492	141	8023
	17	0	0	0	0	0	0	0
	18	0	9	108	0	0	55	0
	19	0	0	0	3	0	0	0
	1000	0	36	2	0	43	106	1250

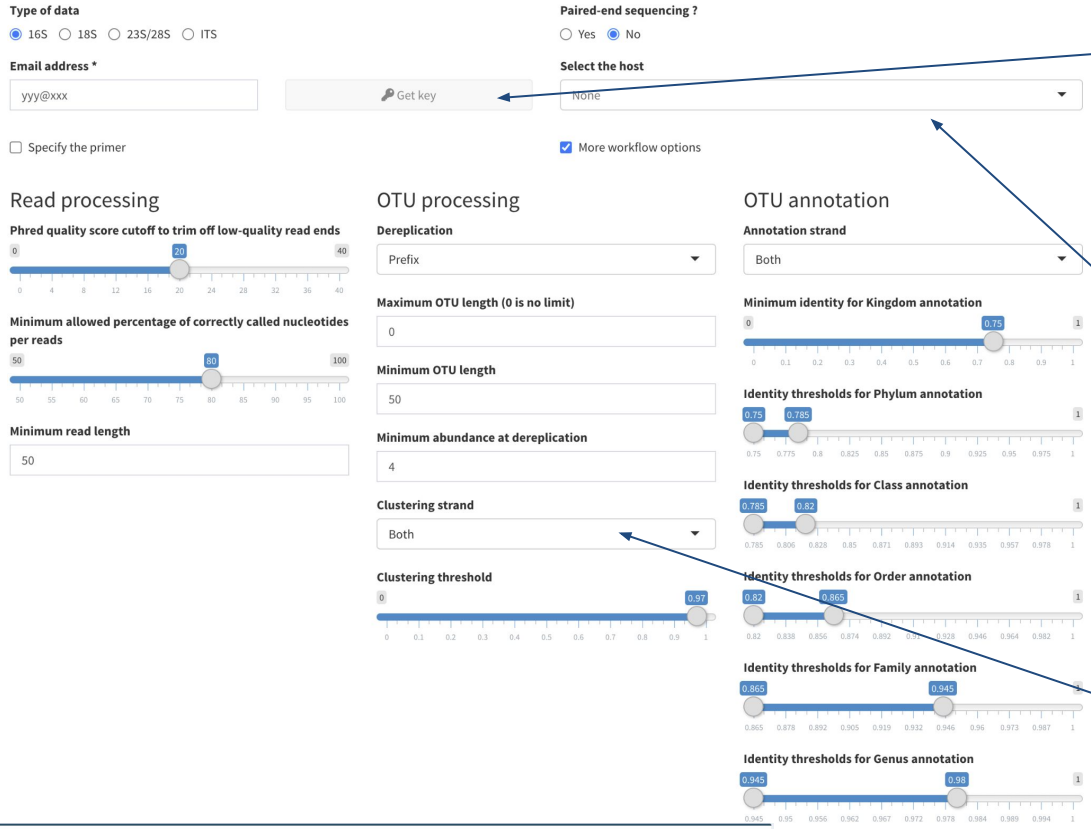
With the courtesy of Metagenopolis

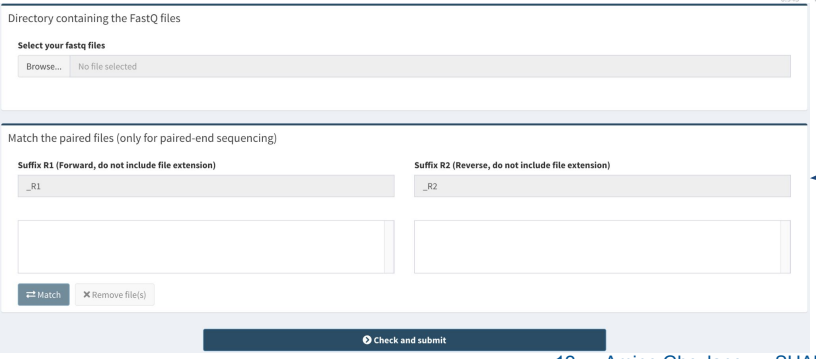
Taxonomy classification of organisms

The hierarchical classification of nature initiated by Carl Linnaeus today consists of eight major “ranks”

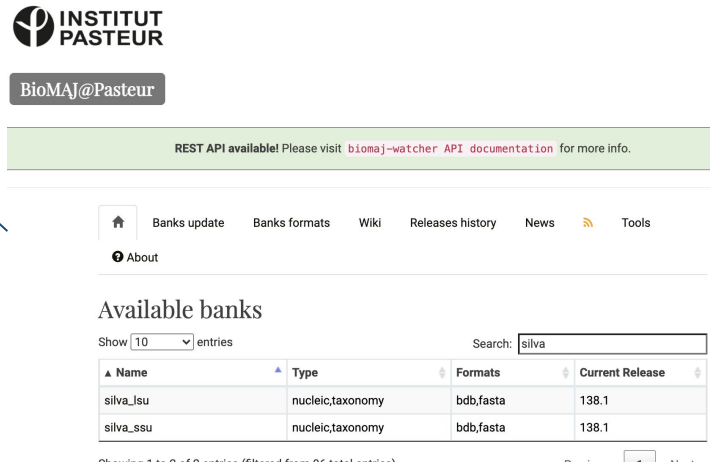


Raw data submission





Unique key associated to your email



Except ITS databases: Unite, Findley, Underhill

Parameters are saved with your key

We do not store your reads after calculations unless:

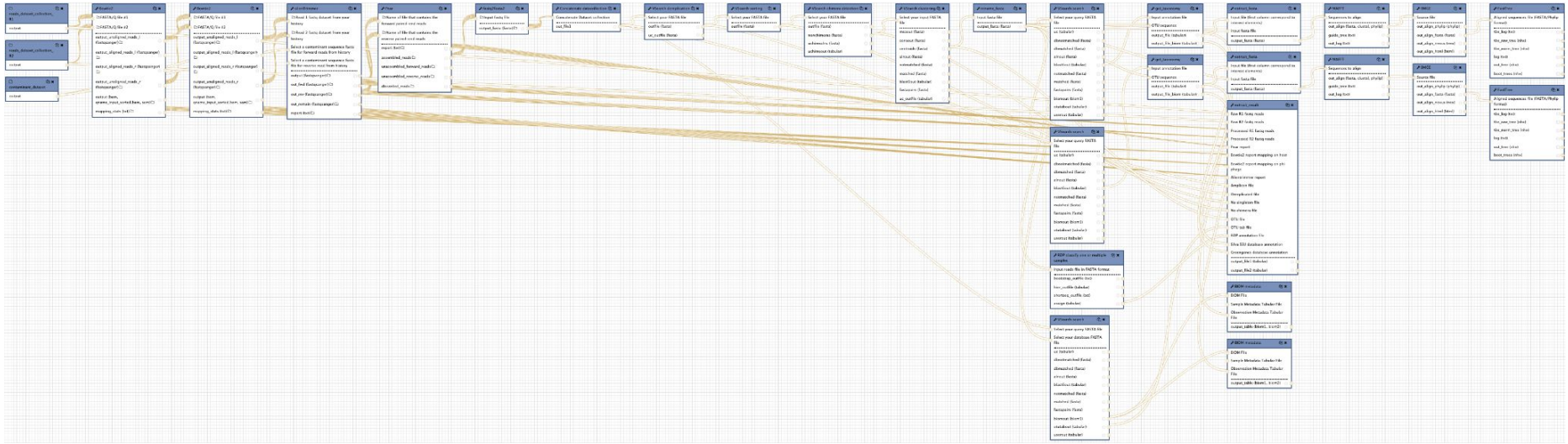
- there is crash and you might come discuss with us: shaman@pasteur.fr

-> crashed dataset are deleted regularly

Institut Pasteur

Raw data submission

What is behind: Galaxy.pasteur.fr



The workflow is based on vsearch.

Institut Pasteur

Raw data submission

Global report

Download available
A mail is also automatically sent

Detailed process

Detailed process table

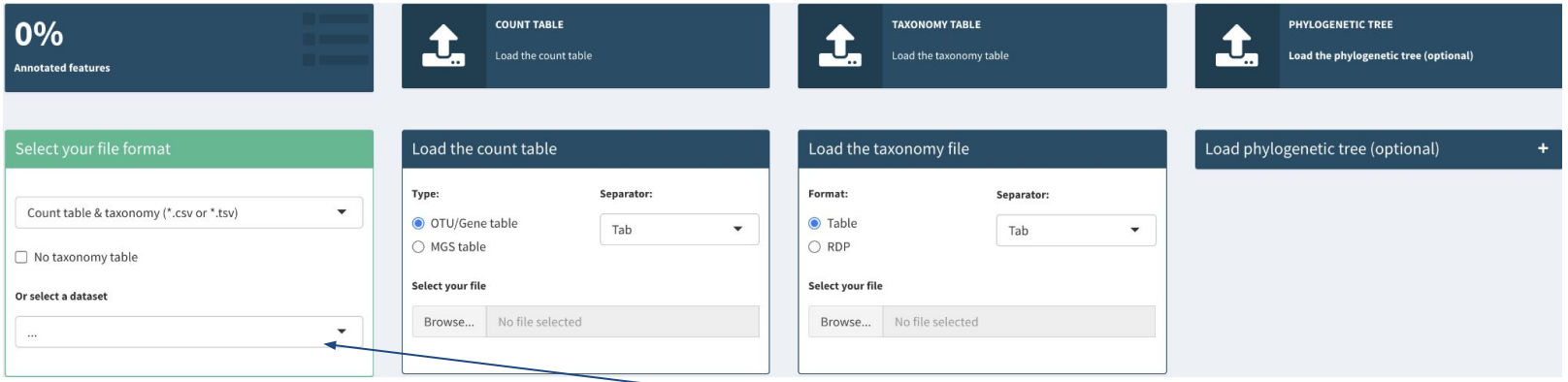
Show 10 entries Search:

	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_t
1	1ng-30cycles-1	275152	300	301	275152	300	300	509447	250179	259268	37693	2
2	zero5ng-30cycles-1	359368	300	301	359368	300	300	665454	326768	338686	49132	2
3	1ng-25cycles-3	172137	300	301	172137	300	300	315854	154559	161295	26482	1
4	zero5ng-30cycles-2	242911	300	301	242911	300	300	449556	221480	228076	33397	1
5	1ng-30cycles-2	391843	300	301	391843	300	300	724875	357756	367119	53718	2
6	1ng-30cycles-3	256766	300	301	256766	300	300	453161	223254	229907	58344	3

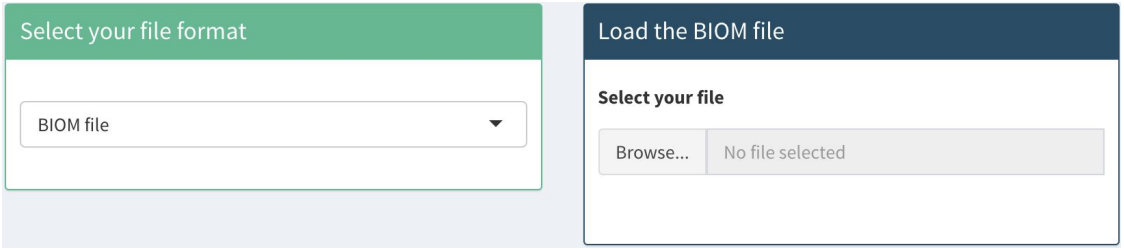
Institut Pasteur

Processed data submission

Tables

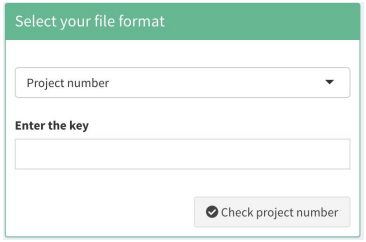


BIOM / Epi2me (nanopore)



Exemple datasets

Key

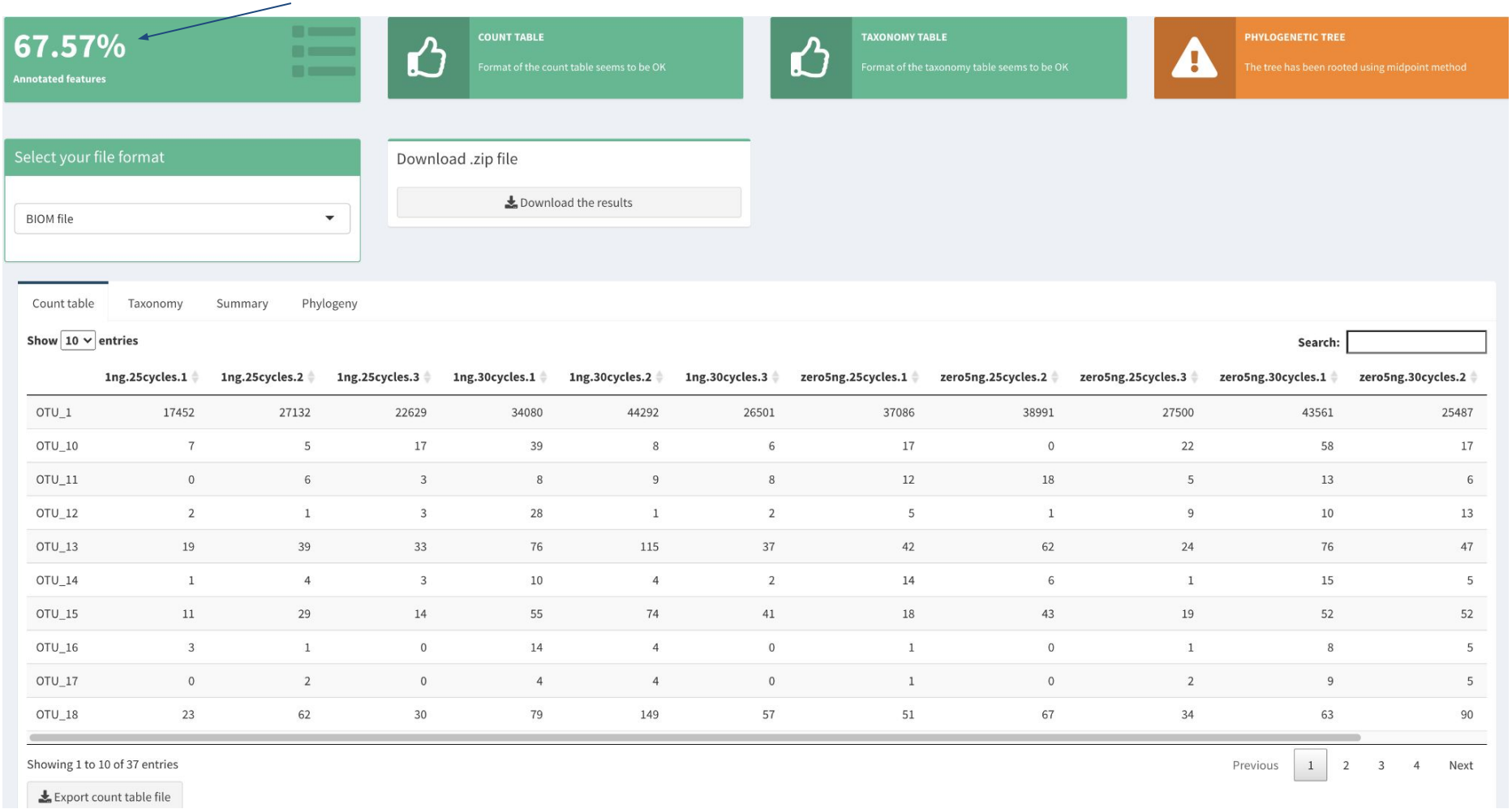


Institut Pasteur

Processed data submission

% of OTU annotated

Unifrac enabled



	1ng.25cycles.1	1ng.25cycles.2	1ng.25cycles.3	1ng.30cycles.1	1ng.30cycles.2	1ng.30cycles.3	zero5ng.25cycles.1	zero5ng.25cycles.2	zero5ng.25cycles.3	zero5ng.30cycles.1	zero5ng.30cycles.2
OTU_1	17452	27132	22629	34080	44292	26501	37086	38991	27500	43561	25487
OTU_10	7	5	17	39	8	6	17	0	22	58	17
OTU_11	0	6	3	8	9	8	12	18	5	13	6
OTU_12	2	1	3	28	1	2	5	1	9	10	13
OTU_13	19	39	33	76	115	37	42	62	24	76	47
OTU_14	1	4	3	10	4	2	14	6	1	15	5
OTU_15	11	29	14	55	74	41	18	43	19	52	52
OTU_16	3	1	0	14	4	0	1	0	1	8	5
OTU_17	0	2	0	4	4	0	1	0	2	9	5
OTU_18	23	62	30	79	149	57	51	67	34	63	90

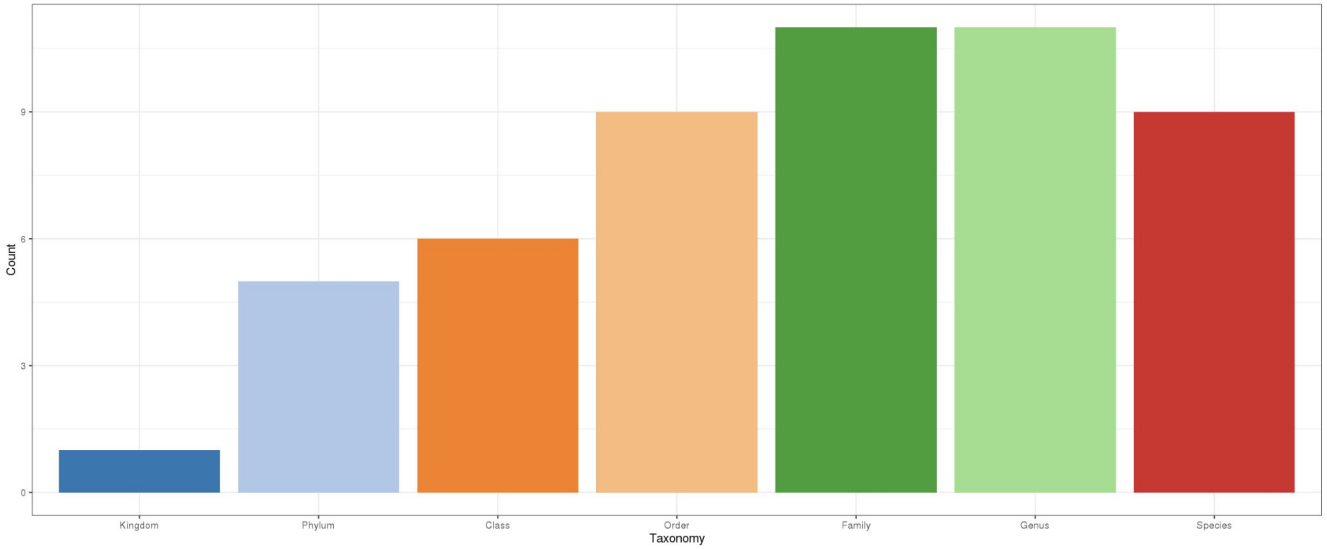
Download available

Institut Pasteur

Processed data submission

Percentage of annotation

Number of features by level:



Institut Pasteur

DESeq2 Metagenomics

	Metagenomics
Distribution	Overdispersed counts → Negative binomial
Constraints	Highly abundant species
Goal	Find differentially abundant features (species, family, …): Feature distributions and abundances vary between conditions

16S : McMurdie, Holmes, Plos Comput Biol,2014
WGS : Jonsson, BMC Genomics, 2016

Institut Pasteur

DESeq2 normalization (OTU level)

Assumption

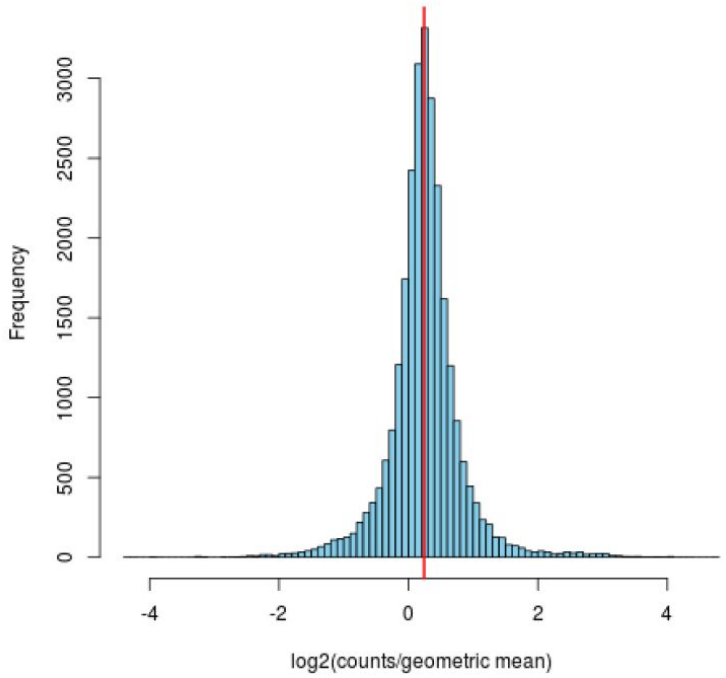
- Most of the OTU have the « same » abundance between 2 conditions

Normalization factor :

$$\hat{s}_j = median_i \frac{x_{ij}}{(\prod_{\nu=1}^{n} x_{i\nu})^{1/n}}$$

where

X_{ij} : Number of mapped reads of the OTU i in sample j
n: Number of samples



Institut Pasteur

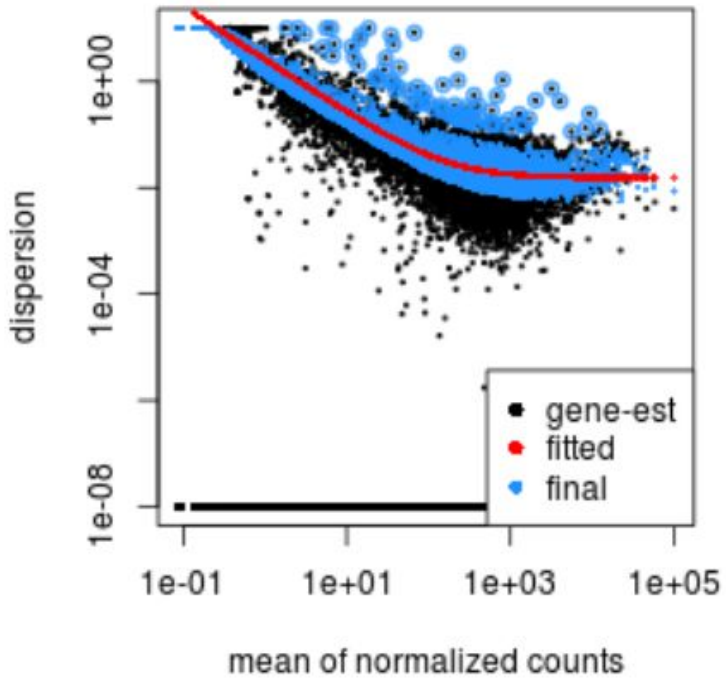
Dispersion estimation

Problem

- Get a good estimate of the dispersion with a small number of samples.

Modelisation of the dispersion :

Function of the mean
of normalized count



Local parametric regression

Contrasts (comparisons)

Aim

- Testing a specific effect without having to re-fit the model.

Contrast vector

$$\beta_i^c = \vec{c}^t \vec{\beta}_i$$

Coefficients

$$\mathrm{SE}(\beta_i^c) = \sqrt{\vec{c}^t \Sigma_i \vec{c}},$$

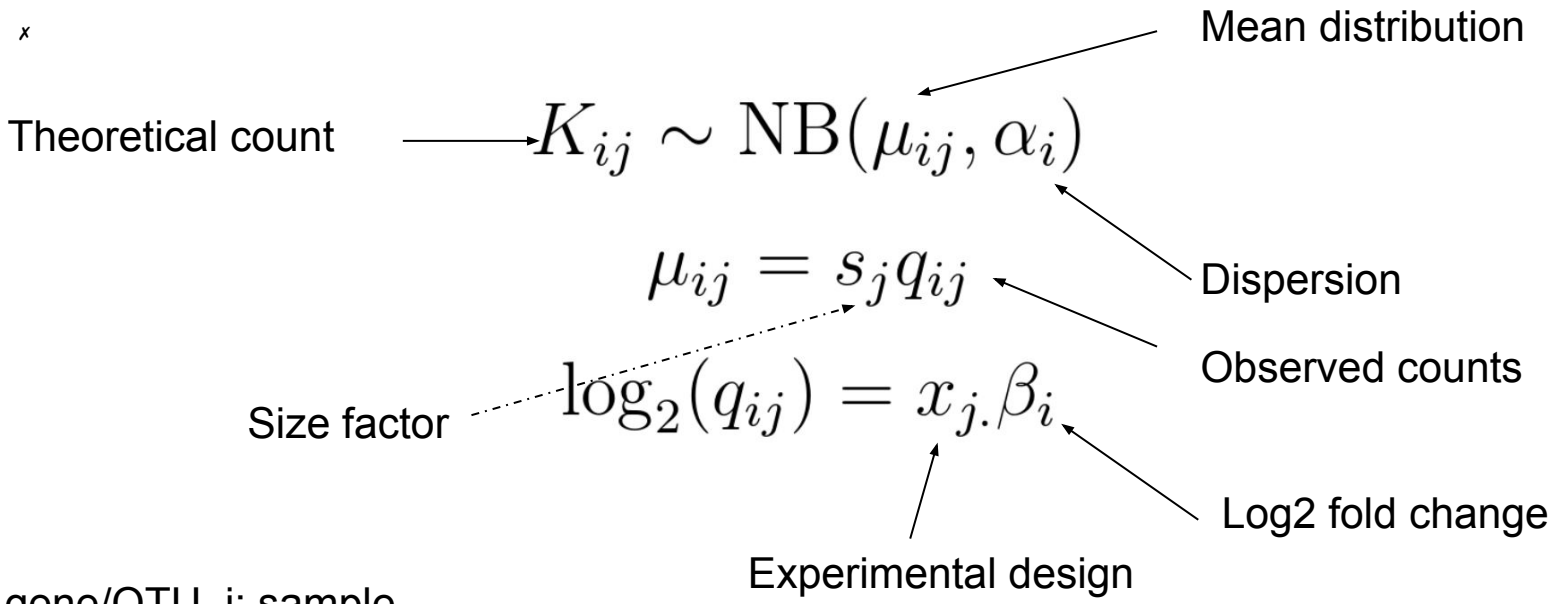
Covariance matrix

Advantages

- Parameters are estimated with all samples.

Statistical model of DESeq2

Generalized Linear Model



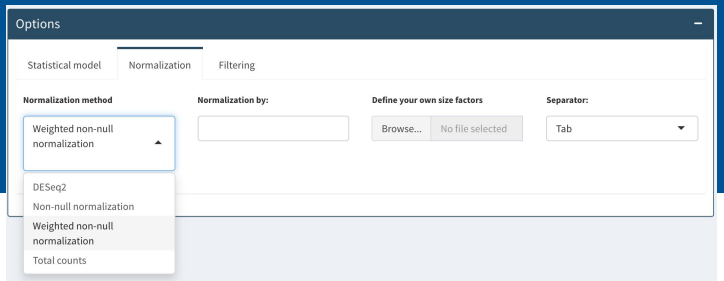
i : gene/OTU, j: sample

Advantages

- Allows complex experimental designs

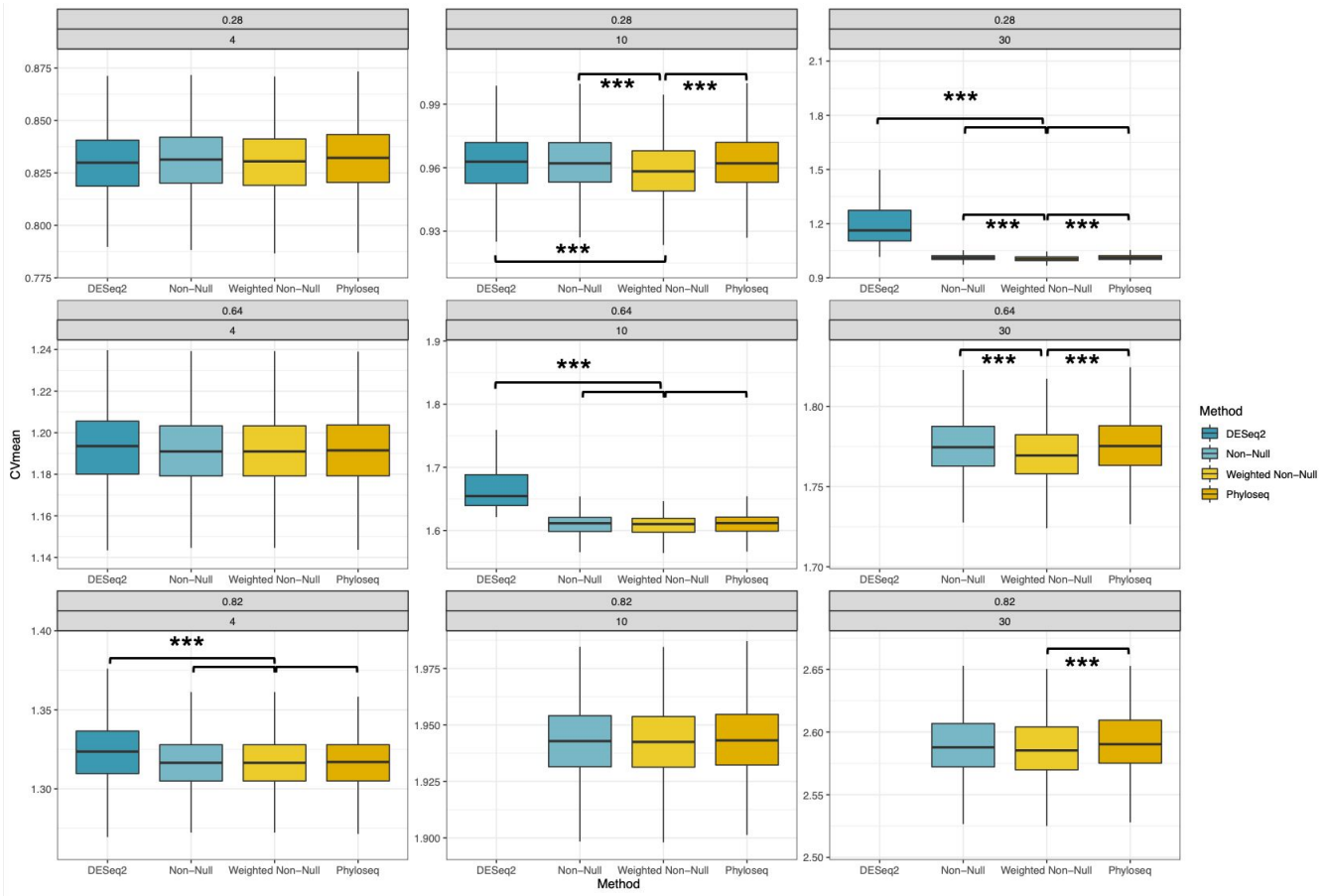
- But tend to crash when the count matrix is highly sparse

Modified normalizations



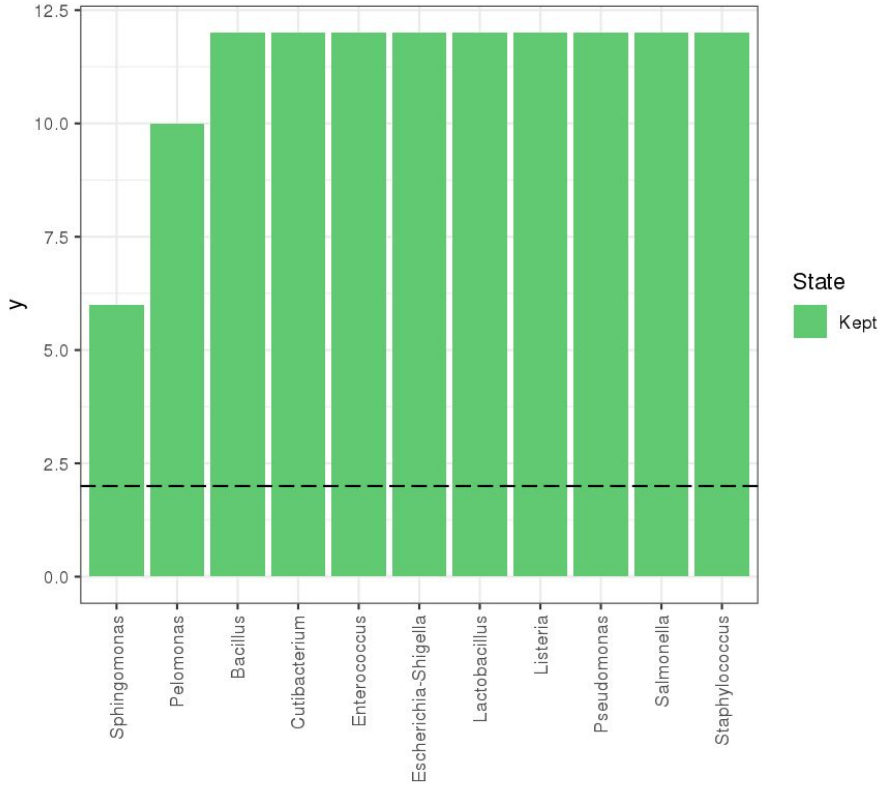
Number of observations

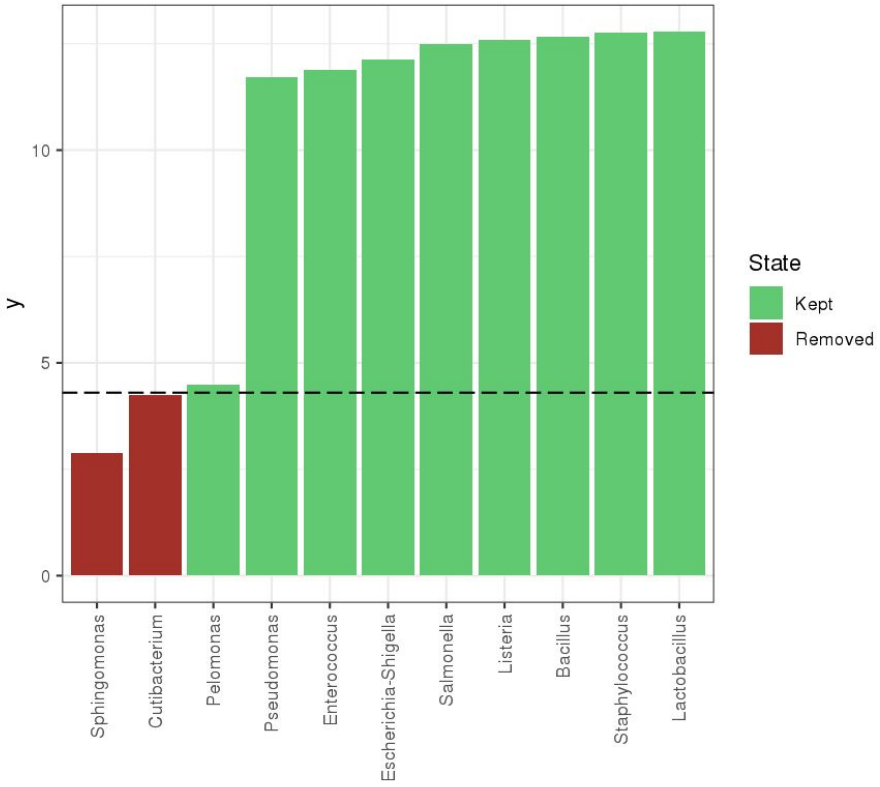
Sparsity ratio



Institut Pasteur

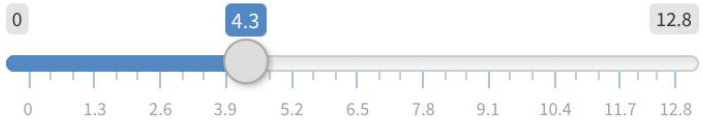
Data filtering





Threshold on the minimal number of samples

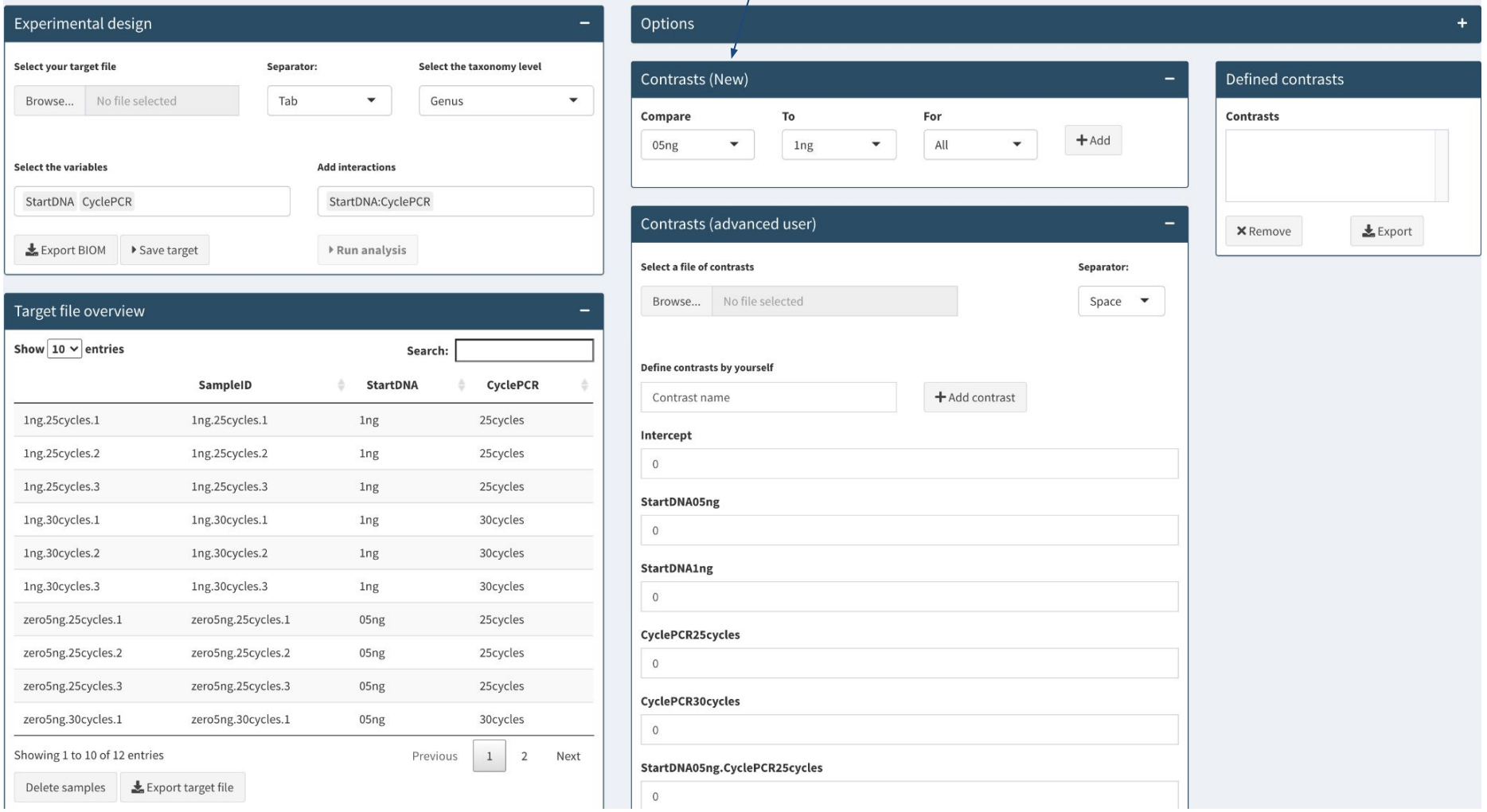
Threshold on the total abundance (in log)



Institut Pasteur

Statistical modelling

Automatically determined from the statistical model



Institut Pasteur

31 interactives visualizations

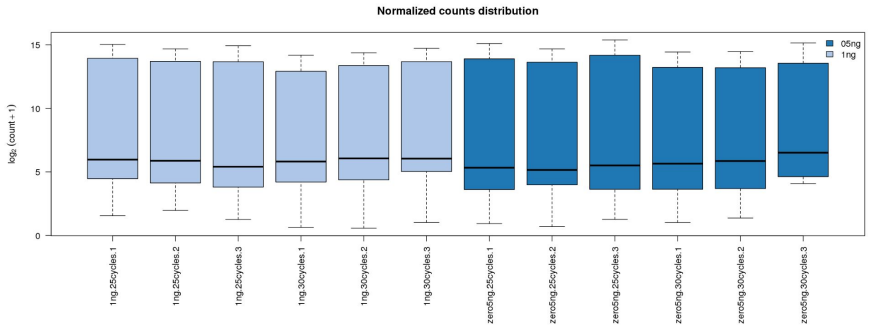
Category	SHAMAN	ASaiM	FROGS	MetaDEGalaxy	Qiita	Shiny-phyloseq	Metaviz	Vamps
OTU processing	Yes	Yes	Yes	Yes	Yes	No	No	No
Computation time (min)	60	>1day	208	>1day	303	NR	NR	NR
Normalization	Yes	No	Yes	Yes	No	No	No	No
Modelling	Yes	No	*Manova*	Yes	No	D	M	No
Diversity analysis	Yes	No	Yes	Yes	Yes	*Alpha*	*Alpha*	*Alpha*
Phylogenetic analysis	Yes	No	Yes	Yes	Yes	Yes	No	*Tree*
Feature abundance plots	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ordination plots	Yes	No	Yes	No	Yes	Yes	Yes	Yes
Network plots	Yes	No	No	Yes	No	Yes	No	Yes
World-map distribution	No	No	No	No	No	No	No	Yes
Statistics plots	Yes	NR	No	No	NR	Yes	NR	NR
Interactive visualization	31	1	2;P	6	3	8	9	17
Raw data storage	No	No	No	No	Yes	No	No	Yes
Result storage	Yes	No	No	No	Yes	No	No	Yes
Online web Interface	Yes	No	No	Yes	Yes	No	Yes	Yes
R packaging	No	NR	NR	NR	NR	Yes	Yes	NR
Docker	Yes	Yes	No	No	No	No	Yes	No
Conda	Yes	No	Yes	No	Yes	No	Yes	No

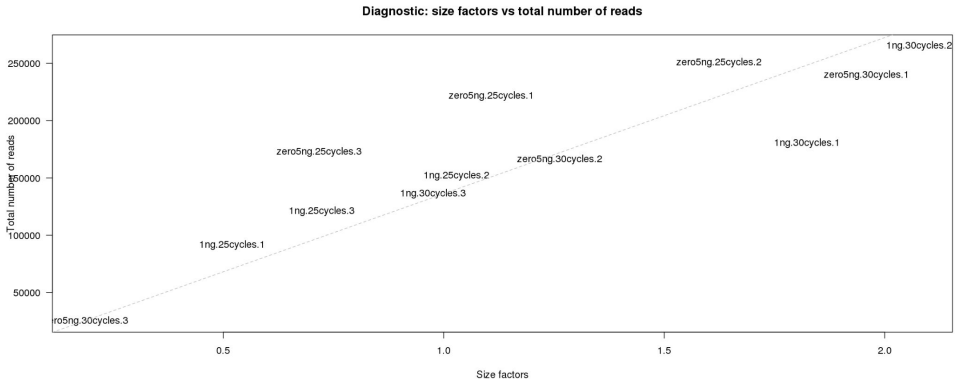
D: Export from DESeq2, M: Export from Metagenomeseq, NR: Non relevant feature, P: Import/Export to Phyloseq, Computation time is indicated for the OTU processing of the Zymo mock communities, Number of unique interactive visualization are reported for each web interface in section 'Interactive visualization'. We reported a specific implementation with the following terms: *Alpha* indicates when only alpha diversity is available for diversity analysis, *Manova* is indicated for FROGS because differential abundant features are detected with a Manova instead of a General Linearized Model. *Tree* indicates when no unifrac distance is available after computing the phylogeny of the OTUs

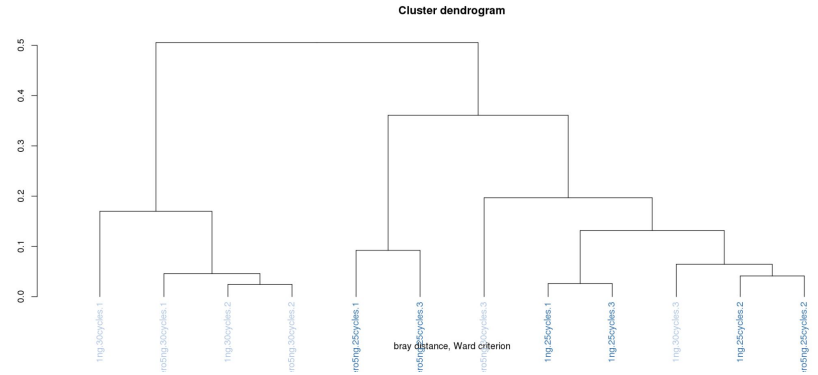
Institut Pasteur

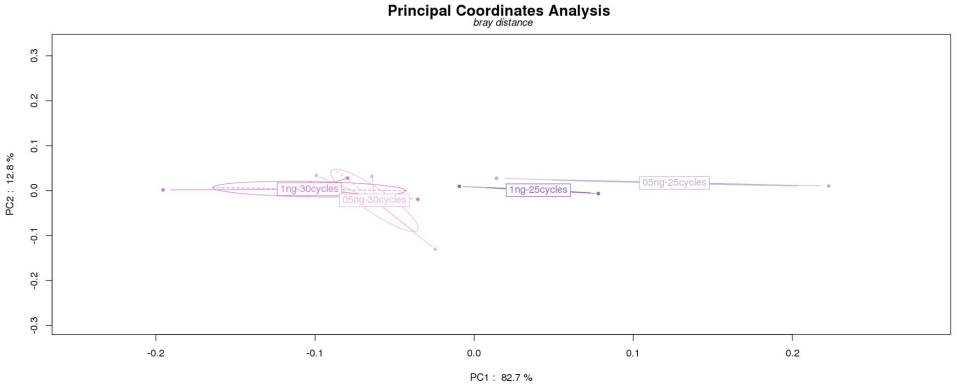
Diagnostic plots

How good is my normalisation ? modelisation ? effect size ?









Significant features

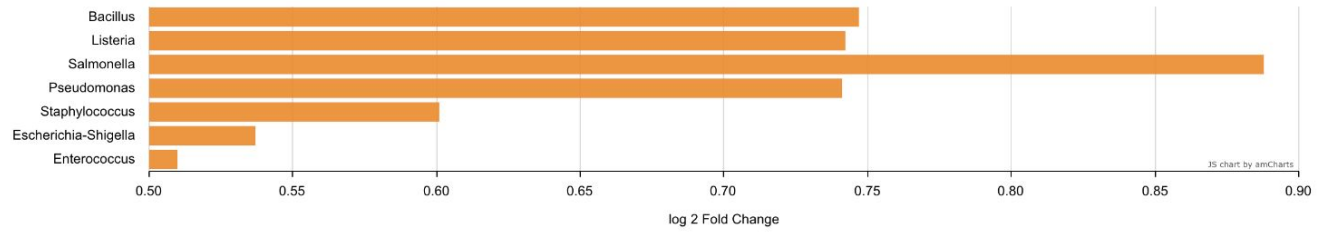
Significant | Complete | Up | Down

Show 10 entries Search:

Id	baseMean	FoldChange	log2FoldChange	pvalue_adjusted
Bacillus	23314.54	1.678582e+00	0.747	0.00146087585674635
Listeria	22410.35	1.673061e+00	0.742	0.00146087585674635
Salmonella	19308.79	1.850925e+00	0.888	0.00146087585674635
Pseudomonas	8696.35	1.671641e+00	0.741	0.0108318526628909
Staphylococcus	26546.82	1.517008e+00	0.601	0.0108318526628909
Escherichia-Shigella	13921.39	1.450844e+00	0.537	0.0112174893175573
Enterococcus	11024.15	1.423880e+00	0.51	0.0210910504868114

Showing 1 to 7 of 7 entries Previous 1 Next

Bar chart | Volcano plot



Select your contrast

Contrasts

X25cycles_vs_30cycles_for_05ng

X25cycles_vs_30cycles_for_05ng
+ CyclePCR25cycles - CyclePCR30cycles + StartDNA05ng.CyclePCR25cycles - StartDNA05ng.CyclePCR30cycles

Reorder data

Order by

pvalue_adjusted

☐ Decreasing order

Plot options

Choose colours

#008B00 **Down**

#999999 **Not significant**

#FF7F00 **Up**

Select the Genus to plot

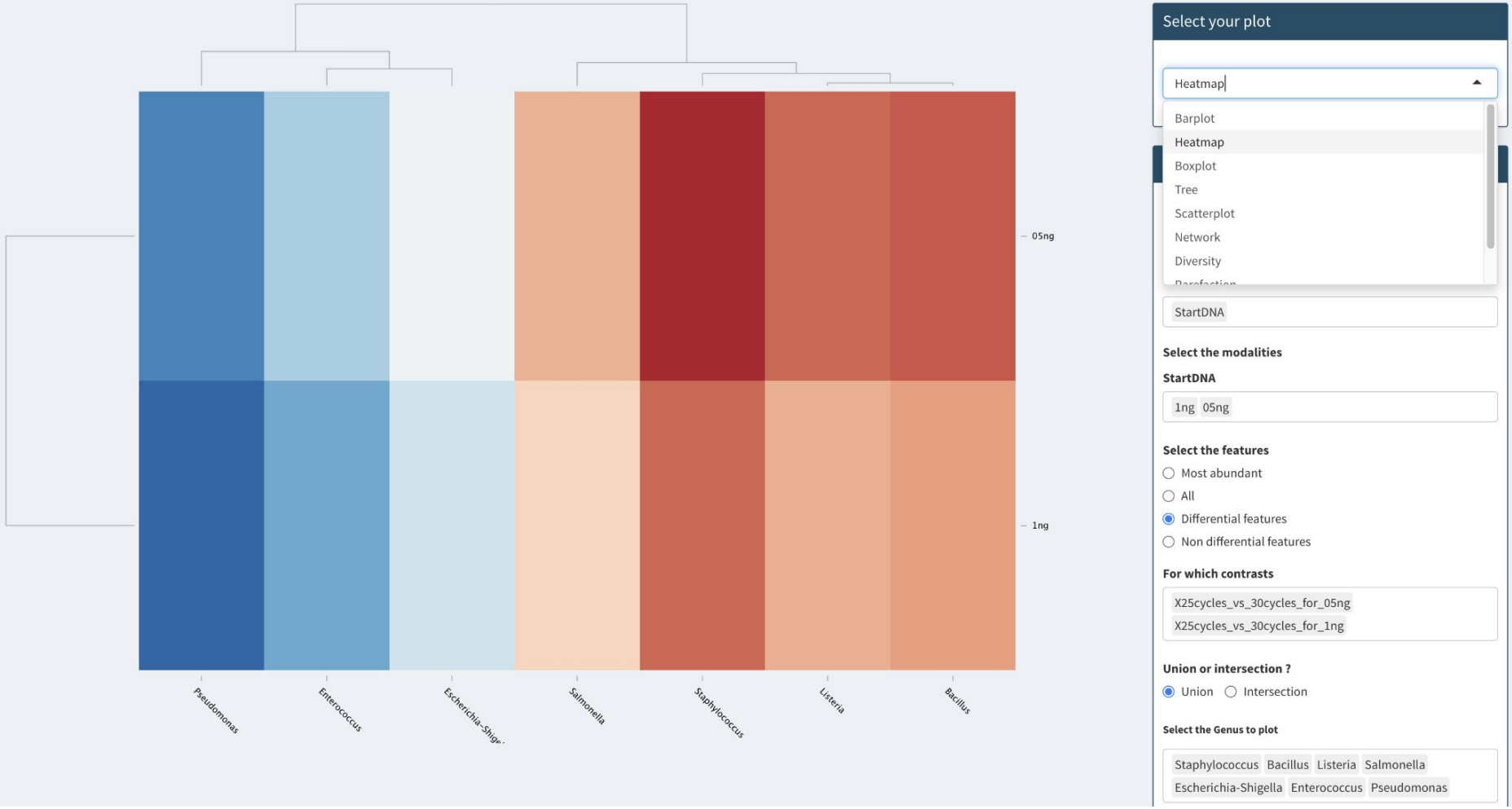
◉ Differential features
○ Most abundant
○ All

Custom selection

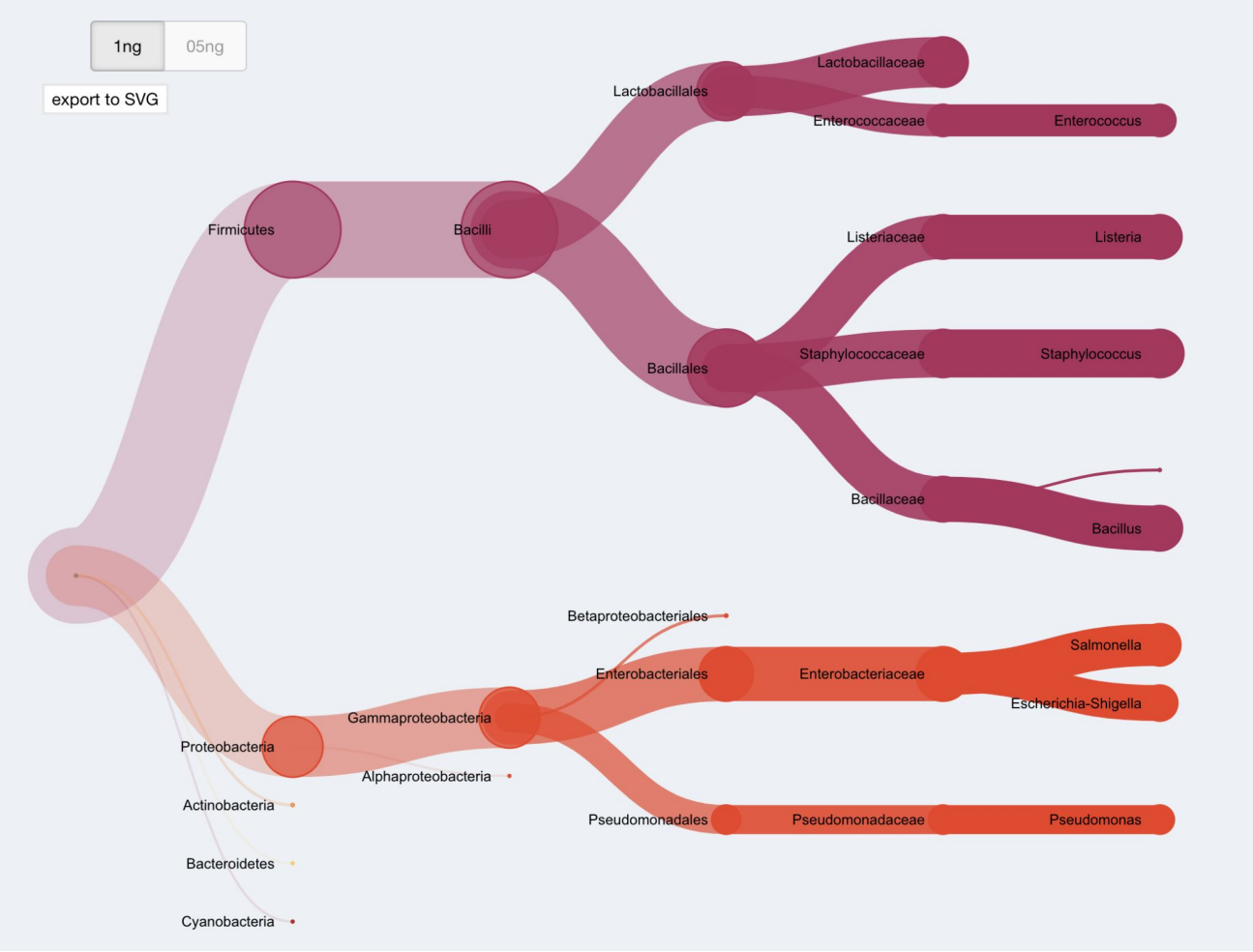
Staphylococcus Bacillus Listeria Salmonella Escherichia-Shigella Enterococcus Pseudomonas

Institut Pasteur

Global visualisations



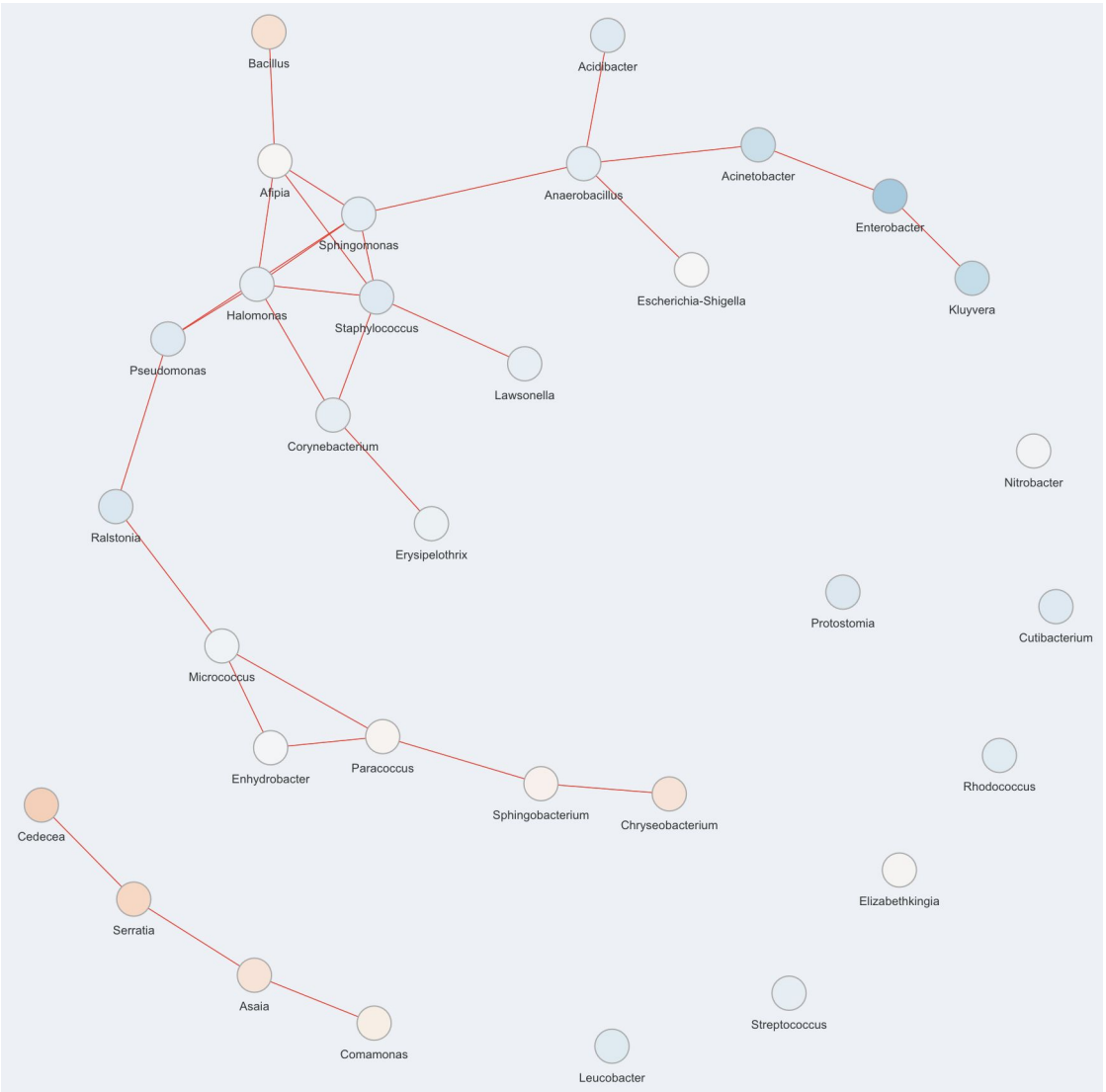
Global visualisations



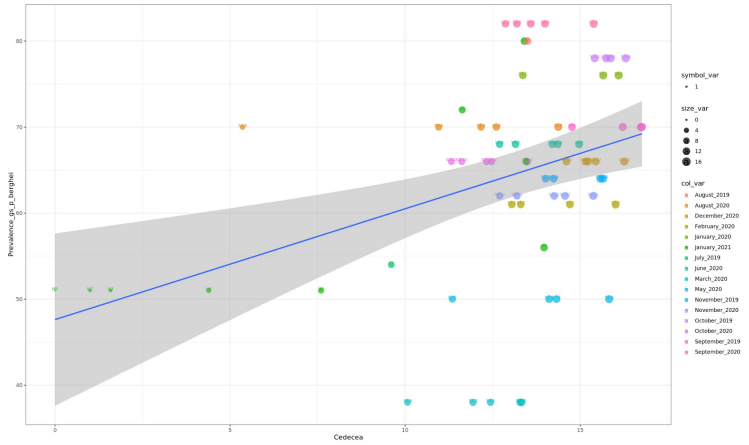
abundance tree

Institut Pasteur

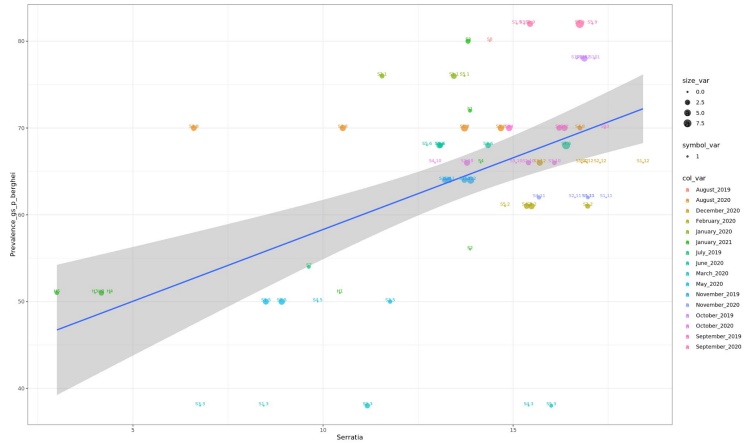
Global visualisations



An. stephensi
Prevalence GS



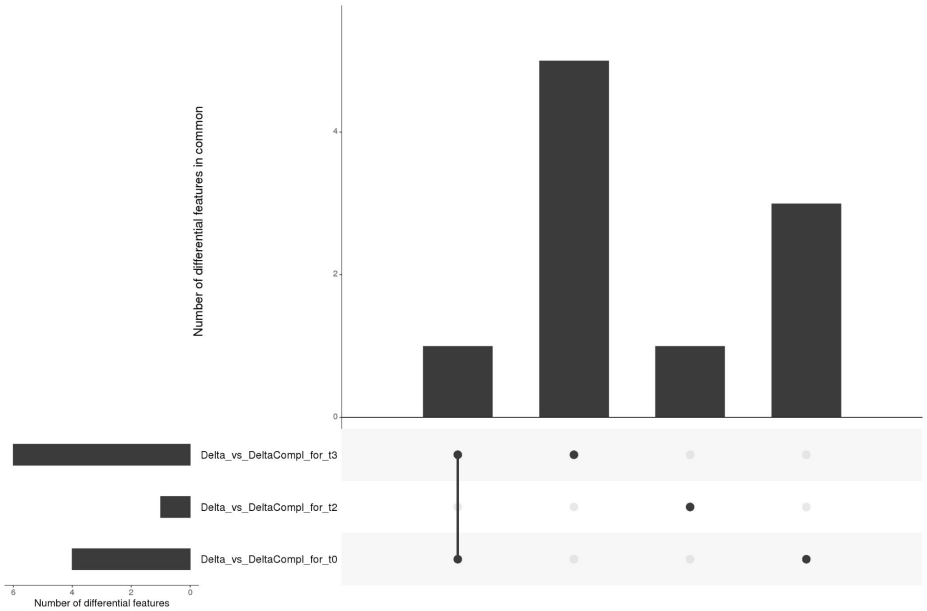
Cedecea R^2 =15%

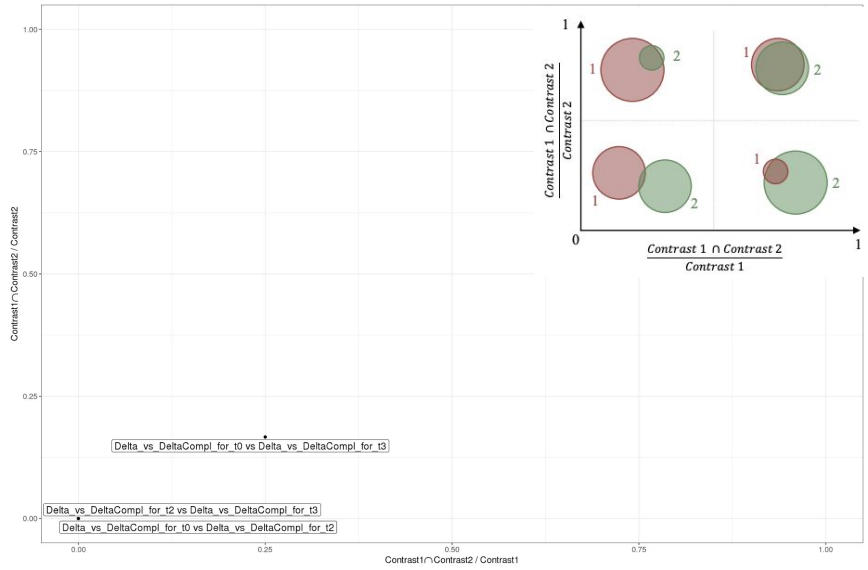


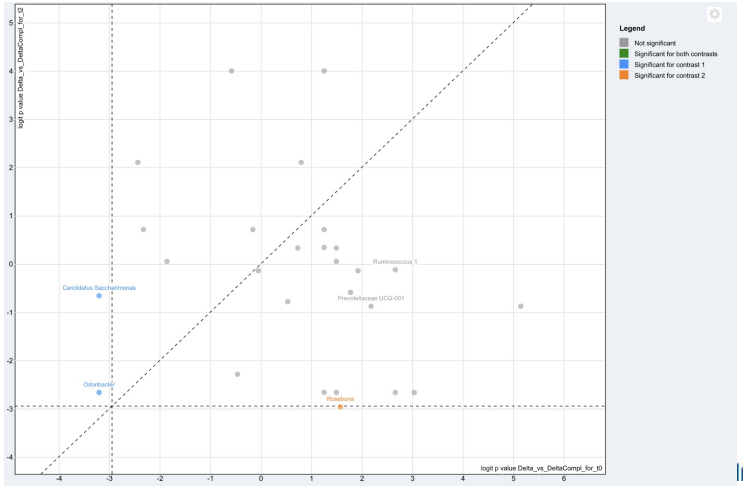
Serratia R^2 =26%

Institut Pasteur

Contrast comparison







Check our publication in BMC Bioinfo !

SOFTWARE Open Access

SHAMAN: a user-friendly website for metataxonomic analysis from raw reads to statistical analysis

Stevenn Volant[1], Pierre Lechat[1], Perrine Woringer[1], Laurence Motreff[2], Pascal Campagne[1], Christophe Malabat[1], Sean Kennedy[2] and Amine Ghozlane[1,2*]

*Correspondence:
amine.ghozlane@pasteur.fr
[1]Hub de Bioinformatique et Biostatistique – Département Biologie Computationnelle, Institut Pasteur, USR 3756 CNRS, 28 Rue Du Docteur Roux, 75015 Paris, France
[2]Biomics – Département Génomes et Génétique, Institut Pasteur, 28 Rue du Docteur Roux, 75015 Paris, France

Abstract

Background: Comparing the composition of microbial communities among groups of interest (e.g., patients vs healthy individuals) is a central aspect in microbiome research. It typically involves sequencing, data processing, statistical analysis and graphical display. Such an analysis is normally obtained by using a set of different applications that require specific expertise for installation, data processing and in some cases, programming skills.

Results: Here, we present SHAMAN, an interactive web application we developed in order to facilitate the use of (i) a bioinformatic workflow for metataxonomic analysis, (ii) a reliable statistical modelling and (iii) to provide the largest panel of interactive visualizations among the applications that are currently available. SHAMAN is specifically designed for non-expert users. A strong benefit is to use an integrated version of the different analytic steps underlying a proper metagenomic analysis. The application is freely accessible at http://shaman.pasteur.fr/, and may also work as a standalone application with a Docker container (aghozlane/shaman), conda and R. The source code is written in R and is available at https://github.com/aghozlane/shaman. Using two different datasets (a mock community sequencing and a published 16S rRNA metagenomic data), we illustrate the strengths of SHAMAN in quickly performing a complete metataxonomic analysis.

Conclusions: With SHAMAN, we aim at providing the scientific community with a platform that simplifies reproducible quantitative analysis of metagenomic data.

Keywords: Metagenomics, Differential analysis, Visualization, Web application

Institut Pasteur

Conclusion

SHAMAN is free !
No login,

>20 publications citing the different publications of SHAMAN

SHAMAN can be installed locally with docker.

>200 unique users per month.
>100 submissions per month.

We provide support and reply to questions: shaman@pasteur.fr (and it's been 5 years that we maintain it)

Future work

More colors !

Shaman is available for RNA-seq, nanostring, WGS...

Improve user experience

Maintain is hard

Keep fixing bugs

Include some machine learning ?