# **SHAMAN : <u>SH</u>iny <u>Application</u> for <u>Metagenomic</u> <u>AN</u>alysis**

Amine Ghozlane Hub Bioinformatique et Biostatistique





### **Subjects**

🐮 SHAMAN		
😤 Home	About Authors Citing SHAMAN Welcome to SHAMAN	What's new in SHAMAN
Tutorial     Download/Install     Raw data     Upload your data	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 bioinformatics treatment is based on Vsearch [Rognes 2016] which showed to be both accurate and fast [Wescott 2015]. The statistical analysis is based on DESeq2 R package [Anders and Huber 2010] which robustly identifies the differential abundant features 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]. SHAMAN is compatible with standard formats for metagenomic analysis (.csv, .tsv, .biom) and figures can be downloaded in several formats. A presentation about SHAMAN is available here and a poster here.	May 5th 2021 - Server maintenance is done Raw reads submission is back. We updated vsearch to 2.17, RDP classifier to 2.13, Bowtie to 2.3.5.1 and Biom to 2.1.10. Let us know if you meet any issue. April 28th 2021 - Server maintenance from 3-
	This website is free and open to all users and there is no login requirement. Hereafter is the global workflow of the SHAMAN application:	<b>6 Mai 2021</b> Raw reads submission will be temporarly desactivated from Mai 3 until 6 due to a cluster maintenance.
	Fasta files     Control random     SHAMAN       Control random     Control random       PCCB     OTU picking       PCCB     Control random       Control random     Control random	Feburary 19th 2020 - New feature Shaman now support Epi2me data output. Let us know if you meet any issue.
	Itar title     Annotation       Bioinformatic analysis     Input files       Statistical analysis     Diagnostic plots	August 19th 2020 - SHAMAN paper is out We are really happy to announce that the paper describing in depth shaman and comparing this application to other

#### Reproducible research



#### Better statistics



#### Simplicity





### **Subjects**

🐮 SHAMAN		
😤 Home	About Authors Citing SHAMAN Welcome to SHAMAN	What's new in SHAMAN
<ul> <li>Tutorial</li> <li>Download/Install</li> <li>Raw data</li> </ul>	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 bioinformatics treatment is based on Vsearch [Rognes 2016] which showed to be both accurate and fast [Wescott 2015]. The statistical analysis is based on DESeq2 R package [Anders and Huber 2010] which robustly identifies the differential abundant features 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]. SHAMAN is compatible with standard formats for metagenomic analysis (.esv, .tsv, .biom) and figures can be downloaded in several formats. A	May 5th 2021 - Server maintenance is done Raw reads submission is back. We updated vsearch to 2.17, RDP classifier to 2.13, Bowtie to 2.3.5.1 and Biom to 2.1.10. Let us know if you meet any issue.
🄹 Upload your data	presentation about SHAMAN is available here and a poster here. This website is free and open to all users and there is no login requirement. Hereafter is the global workflow of the SHAMAN application:           Image: Constraint vectors	April 28th 2021 - Server maintenance from 3- 6 Mai 2021 Raw reads submission will be temporarly desactivated from Mai 3 until 6 due to a cluster maintenance.
	Fasts files     Imaging patted reads       For patted reads     Imaging patted reads       Octowal rate     Imaging patted reads       Imaging patted reads     Imaging patted reads <th>Feburary 19th 2020 - New feature Shaman now support Epi2me data output. Let us know if you meet any issue.</th>	Feburary 19th 2020 - New feature Shaman now support Epi2me data output. Let us know if you meet any issue.
	primers     Image: data analysis     Im	August 19th 2020 - SHAMAN paper is out We are really happy to announce that the paper describing in depth shaman and comparing this application to other

#### Reproducible research



#### Better statistics



#### Simplicity





#### 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



### **Spot 5 issues for reproducibility**

Download Dataset S14 (PDF)

## 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 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)</pre>
```

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

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

<sup>a</sup>Unité de Pathogénie Microbienne Moléculaire, Institut Pasteur, 75015 Paris, France; <sup>b</sup>Unité (NSERM 1202, Institut Pasteur, 75015 Paris, France; <sup>b</sup>Denté MISERM 1202, Institut Pasteur, Denté MISERM, Vancouver, BC VGT 124, Canada; <sup>l</sup>Unité de Bactériologie Expérimentale, Institut Pasteur de Madagascar, <sup>B</sup>P1274 Ambatofotsikely, 101 Antananarivo, Madagascar; <sup>D</sup>Denté MISERM 1202, VGT 124, Canada; <sup>l</sup>Unité de Bactériologie Expérimentale, Institut Pasteur de Madagascar, <sup>B</sup>P1274 Ambatofotsikely, 101 Antananarivo, Madagascar; <sup>l</sup>Complexe Pédiatrique de Bangui, BP 923 Bangui, Central African Republic; <sup>l</sup>Unité d'Epidémiologie, Institut Pasteur de Madagascar, <sup>B</sup>P1274 Ambatofotsikely, 101 Antananarivo, Madagascar; <sup>l</sup>Complexe Pédiatrique de Bangui, BP 923 Bangui, Central African Republic; <sup>l</sup>Unité d'Epidémiologie, Institut Pasteur de Madagascar, BP 1274 Ambatofotsikely, 101 Antananarivo, M

I "pre-reviewed" this paper -> Acknowledgement



## **Spot 5 issues for reproducibility**

## Look at Picrust output data u	using R		eript is observed as a DDE of 59 pages
<pre>rm(list=ls())</pre>	<ul> <li>Download Dataset_</li> </ul>	514 (PDF) 1 - THE S	chpt is shared as a PDF of 58 pages
library(phyloseq) library(microbiome)		R-4.0.0.tar.gz	2020-04-24 09:05 32M
library(devtools) library(RJSONIO)	2 Die en heatile	<u>R-4.0.1.tar.gz</u>	2020-06-06 09:05 32M
library(qualpatr) library(plyr) library(car)	environment for	R-4.0.2.tar.gz	2020-06-22 09:05 32M
library(ecodist) library(ade4) library(vegan)	reproducibility.	R-4.0.3.tar.gz	2020-10-10 09:05 32M
library(permute) library(lattice)		R-4.0.4.tar.gz	2021-02-15 09:05 32M
library(reshape2) library(dplyr) library(MASS)		R-4.0.5.tar.gz	2021-03-31 09:05 31M
library(Hmisc) library(RColorBrewer)		R-4.1.0.tar.gz	2021-05-18 09:05 32M
library(ggplot2)		Packa	ge: vegan
library(colorspace) library(qualpalr) library(nlme)	3 - "R_files" is not avail	able!	: Community Ecology Package on: 2.6–0
<pre>#### set working directory #### setwd("/Users/Pascale/Desktop/PM</pre>	MAS/R_files") # put here your path to the	Autho Pi e working directory Ga	r: Jari Oksanen, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, erre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, vin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs,
<pre>#### import the files obtained f mixed models ####</pre>	from Categorize by function (tax_table an	nd otu_table or directly the Maint	ainer: Jari Oksanen <jhoksane@gmail.com> ds: permute (&gt;= 0.9-0), lattice, R (&gt;= 3.4.0)</jhoksane@gmail.com>
	4	- Space III Hames Sugge	sts: parallel, tcltk, knitr, markdown
otu=reda.table(file= Additional	Data Table19 otutablepicrust.txt", heade	r = 1KUE, $sep = 1t$ , row.nam Impor	tteBuilder: utils. knitr
class(otu)		Descr	iption: Ordination methods, diversity analysis and other
OTU = otu_table(otu, taxa_are_ro	ows = TRUE)	fun	ctions for community and vegetation ecologists.
оти	8	Licen	se: GPL-2
		BugRe	<pre>ports: https://github.com/vegandevs/vegan/issues</pre>
		URL:	https://github.com/vegandevs/vegan



## **Spot 5 issues for reproducibility**

Few parameters described

Several software to install

## 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

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 OIIME. 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  $\chi^2$  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 PhylosEg (46), vegan (47), randomForest (48, 49), DeSeq2 (50–52), and ggplot2 (53) R packages.  $\alpha$ -Diversity was guantified using a measure of richness [Chao1 index (54)] and a measure of evenness (Simpson's diversity index = 1 -Simpson's index) while  $\beta$ -diversity was guantified using the Bray–Curtis dissimilarity index (55). Tests of differences in a-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 (a 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			
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
a 5	115	0	0	1	0	29	2
5 6	90	783	26	0	2	0	0
0 7	104	1616	0	0	0	0	5
8 27	0	82	0	0	0	0	0
e 9	2	0	0	0	0	0	0
G 10	23	239	1302	10	0	190	0
<b>อ้</b> 11	30	183	900	13	0	172	0
<b>12</b>	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

#### **Taxonomy classification of organisms**

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



## **Raw data submission**

Type of data	Paired-end sequenc Yes  No Select the host	ing ?	<ul> <li>Unique key associated to your email</li> </ul>
ууу@xxx	P Get key	· · ·	
Specify the primer	🗹 More workflow op	ptions	BioMAJ@Pasteur
Read processing	OTU processing	OTU annotation	
Phred quality score cutoff to trim off low-quality read end	ds Dereplication	Annotation strand	REST API available! Please visit biomaj-watcher API documentation for more info.
	Prefix	Both	\
0 4 8 12 16 20 24 28 32 36	Maximum OTU length (0 is no limit)	Minimum identity for Kingdom annotation	Banks undate Banks formats Wilki Releases history News Tools
Minimum allowed percentage of correctly called nucleotic	des 0	0 0.75 1	
50 80	100 Minimum OTU length	0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1	€ ADOUT
50 55 60 65 70 75 80 85 90 95	100 50	Identity thresholds for Phylum annotation	Available banks
Minimum read length		0.75 0.785 1	Show 10 v entries Search: silva
50		0.75 0.775 0.8 0.825 0.875 0.9 0.925 0.95 0.975 1	▲ Name ▲ Type
	**	Identity thresholds for Class annotation	silva_lsu nucleic,taxonomy bdb,fasta 138.1
	Clustering strand	0.785 0.82 1	silva_ssu nucleic,taxonomy bdb,fasta 138.1
	Both	0.785 0.806 0.828 0.85 0.871 0.893 0.914 0.935 0.957 0.978 1	
	Clustering threshold	Hentity thresholds for Order annotation	
	0	0.97 0.82 0.865 1	Except ITS databases: Unite
	0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8	0.9 1 0.82 0.838 0.856 0.874 0.892 0.31 0.928 0.946 0.964 0.982 1	Findley, Underhill
		Identity thresholds for Family annotation	Finaley, Undernill
		0.865	Darameters are saved with your key
		0.865 0.878 0.892 0.905 0.919 0.932 0.946 0.96 0.973 0.987 1	T arameters are saved with your key
		Identity thresholds for Genus annotation 0.945 1	
irectory containing the FastQ files		0.945 0.95 0.962 0.967 0.972 0.978 0.984 0.989 0.994 1	We do not store your reads after
Select your fastq files			calculations unless:
Browse No file selected			
			<ul> <li>there is crash and you might</li> </ul>
atch the paired files (only for paired-end sequencing)			come discuss with us:
Suffix R1 (Forward, do not include file extension)	Suffix R2 (Reverse, do not include file extension)		
	_KZ		<u>shaman@pasteur.tr</u>
			-> crashed dataset are deleted
≓ Match ★ Remove file(s)			regularly
			Institut Pasteur

### **Raw data submission**

#### What is behind: Galaxy.pasteur.fr



The workflow is based on vsearch.



## **Raw data submission**

#### Global report OTU building process 16S annotation process Start statistical analysis Load the results 2227104 25 Select the database Number of OTU annotated by SILVA Number of amplicons Silva 922879 25 1 Upload the results **Remaining amplicons after dereplication** Number of OTU annotated by Greengenes Download .zip file 22422 36 Lownload the results Remaining amplicons after removing singletons Number of OTU annotated by RDP Download available 19501 \_\_\_ A mail is also automatically sent Remaining amplicons after removing chimera

Number of OTU

37

#### **Detailed process**

De	etailed pro	ocess table s								Searc	:h:	
	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_1
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



14 · Amine Ghozlane · SHAMAN : Shiny Application for Metagenomic ANalysis · 25/06/2021

### **Processed data submission**

#### Tables

0% Annotated features	COUNT TABLE Load the count table	Load the taxonomy table	PHYLOGENETIC TREE Load the phylogenetic tree (optional)
Select your file format	Load the count table	Load the taxonomy file	Load phylogenetic tree (optional) +
Count table & taxonomy (*.csv or *.tsv)   No taxonomy table  Or select a dataset	Type:     Separator: <ul> <li>OTU/Gene table</li> <li>MGS table</li> </ul> Tab <ul> <li>Tab</li> </ul> Select your file     Browse     No file selected	Format:     Separator:          ● Table      Tab	
BIOM / Epi2me (nanop	ore)		
Select your file format	Load the BIOM file	e	Exemple datasets
BIOM file	Select your file		

No file selected

Browse...

#### Key

•
Ocheck project number



### **Processed data submission**

% of	OTU anı	notated							Un	ifrac enabl	ed
67.57 Annotated featur	res		ம்	COUNT TABLE	nt table seems to be OK		Format of the	ABLE taxonomy table seems to be OK		PHYLOGENETIC TREE	d using midpoint method
Select your I	file format		Downloa	ad .zip file							
BIOM file		•		🛓 Downlo	ad the results						
Count table Show 10 ~	Taxonomy	Summary Phylo	ogeny							Search:	
	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										
OTU 11		5	17	39	8	6	17	0	22	58	17
	0	5	17 3	39 8	8	6	17	0	22	58	17 6
OTU_12	0	56	17 3 3	39 8 28	8 9 1	6 8 2	17 12 5	0 18 1	22 5 9	58 13 10	17 6 13
OTU_12 OTU_13	0 2 19	5 6 1 39	17 3 3 33	39 8 28 76	8 9 1 115	6 8 2 37	17 12 5 42	0 18 1 62	22 5 9 24	58 13 10 76	17 6 13 47
OTU_12 OTU_13 OTU_14	0 2 19 1	5 6 1 39 4	17 3 3 33 33	39 8 28 76 10	8 9 1 115 4	6 8 2 37 2	17 12 5 42 14	0 18 1 62 6	22 5 9 24 1	58 13 10 76 15	17 6 13 47 5
OTU_12 OTU_13 OTU_14 OTU_15	0 2 19 1 1	5 6 1 39 4 29	17 3 3 33 3 3 14	39 8 28 76 10 55	8 9 1 115 4 74	6 8 2 37 2 41	17 12 5 42 14 18	0 18 1 62 6 43	22 5 9 24 1 19	58 13 10 76 15 52	17 6 13 47 5 52
OTU_12 OTU_13 OTU_14 OTU_15 OTU_16	0 2 19 1 11 3	5 6 1 39 4 29 1	17 3 33 33 3 14 0	39 8 28 76 10 55 14	8 9 1 115 4 74	6 8 2 37 2 41 0	17 12 5 42 14 18	0 18 1 62 6 43 0	22 5 9 24 1 19	58 13 10 76 15 52 8	17 6 13 47 5 52 52
OTU_12 OTU_13 OTU_14 OTU_15 OTU_16 OTU_17	0 2 19 1 1 1 3 3	5 6 1 39 4 29 1 2	17 3 33 33 14 0 0	39 8 28 76 10 55 14 4	8 9 1 115 4 74 4 4	6 8 2 37 2 41 0	17 12 5 42 14 18 18	0 18 1 62 6 43 0	22 5 9 24 1 19 1 2	58 13 10 76 15 52 8 9	17 6 13 47 5 52 5 5
OTU_12 OTU_13 OTU_14 OTU_15 OTU_16 OTU_17 OTU_18	0 2 19 1 11 3 0 23	5 6 1 39 4 29 1 1 2 62	17 3 3 33 3 4 14 0 0 30	39 8 28 76 10 55 14 4 79	8 9 115 4 74 4 4	6 8 2 37 2 41 0 0 57	17 12 5 42 14 18 18 1 1	0 18 1 62 6 43 0 0 0	22 5 9 24 1 1 9 1 9 2 1 2 34	58 13 10 76 15 52 8 9 63	17 6 13 47 5 52 5 5 5 90

よ Export count table file

Download available



### **Processed data submission**

#### Percentage of annotation



#### Number of features by level:





### **DESeq2** Metagenomics

	Metagenomics
Distribution	Overdispersed counts $\rightarrow$ 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



# **DESeq2 normalization (OTU level)**

#### Assumption

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

#### Normalization factor :

$$\hat{s}_j = median_i rac{X_{ij}}{(\prod_{
u=1}^n x_{i
u})^{1/n}}$$

where

X<sub>ij</sub> : Number of mapped reads of the OTU i in sample j n: Number of samples





## **Dispersion estimation**

#### Problem

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





### **Contrasts (comparisons)**

#### Aim

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



#### **Advantages**

Parameters are estimated with all samples.



# Statistical model of DESeq2

## Generalized Linear Model Theoretical count $K_{ij} \sim NB(\mu_{ij}, \alpha_i)$ $\mu_{ij} = s_j q_{ij}$ Size factor $\log_2(q_{ij}) = x_j \beta_i$ Log2 fold change Experimental design

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



### **Data filtering**



Threshold on the minimal number of samples





12.8

11.7

Institut Pasteur

0

2.6

3.9

5.2

6.5

7.8

9.1

10.4

Kept

# **Statistical modelling**

Experimental design –						
Select your tar	get file	Separator:	Select the taxonomy	level		
Browse	No file selected	Таb	▼ Genus	•		
Select the varia	ables	Add intera	actions			
StartDNA	CyclePCR	StartD	NA:CyclePCR			
🛓 Export B	IOM Save target	▶ Run a	analysis			

Target file overvi	ew		-
Show 10 🗸 entries		Sea	irch:
	SampleID	StartDNA	CyclePCR
1ng.25cycles.1	1ng.25cycles.1	lng	25cycles
1ng.25cycles.2	1ng.25cycles.2	lng	25cycles
1ng.25cycles.3	1ng.25cycles.3	1ng	25cycles
1ng.30cycles.1	1ng.30cycles.1	1ng	30cycles
1ng.30cycles.2	1ng.30cycles.2	lng	30cycles
1ng.30cycles.3	1ng.30cycles.3	lng	30cycles
zero5ng.25cycles.1	zero5ng.25cycles.1	05ng	25cycles
zero5ng.25cycles.2	zero5ng.25cycles.2	05ng	25cycles
zero5ng.25cycles.3	zero5ng.25cycles.3	05ng	25cycles
zero5ng.30cycles.1	zero5ng.30cycles.1	05ng	30cycles
Showing 1 to 10 of 12	entries	Pri	evious 1 2 Next
Delete samples	よ Export target file		

#### / Automatically determined from the statistical model

Contrasts (New)		-	Defined contra	sts
Compare To 05ng T 1ng T	For All	+ Add	Contrasts	
Contrasts (advanced user)		-	× Remove	Ł Export
Select a file of contrasts		Separator:		
Browse No file selected		Space 🔻		
Define contrasts by yourself Contrast name ntercept 0	+ Add contrast			
StartDNA05ng				
0				
StartDNA1ng				
0				
CyclePCR25cycles				
0				
CyclePCR30cycles		1		
0				
StartDNA05ng.CyclePCR25cycles				



## **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	М	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



### **Diagnostic plots**

#### How good is my normalisation ? modelisation ? effect size ?





# **Significant features**

Significant Complete Up Down				
Show 10 v 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

<pre>(25cycles_vs_30cycles_for_05ng</pre>	
---	--

-

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

Reorder data	- 1
Order by	
pvalue_adjusted	•
Decreasing order	





### **Global visualisations**





### **Global visualisations**



#### abundance tree



### **Global visualisations**



### **Contrast comparison**



32 · Amine Ghozlane · SHAMAN : Shiny Application for Metagenomic ANalysis · 25/06/2021

# **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<sup>1</sup>, Pierre Lechat<sup>1</sup>, Perrine Woringer<sup>1</sup>, Laurence Motreff<sup>2</sup>, Pascal Campagne<sup>1</sup>, Christophe Malabat<sup>1</sup>, Sean Kennedy<sup>2</sup> and Amine Ghozlane<sup>1,2\*</sup>

\*Correspondence:

amine.ghozlane@pasteur.fr <sup>1</sup>Hub de Bioinformatique et Biostatistique – Département Biologie Computationnelle, Institut Pasteur, USR 3756 CNRS, 28 Rue Du Docteur Roux, 75015 Paris, France <sup>2</sup>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



### 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: <u>shaman@pasteur.fr</u> (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 ?