

SHAMAN : a **SH**iny Application for Metagenomic ANalysis

Amine Ghozlane^{1,2*}, Stevenn Volant^{1*}, Hugo Varet^{1,2}, Christophe Malabat¹, Pierre Lechat¹, Sean Kennedy², Marie-Agnès Dillies^{1,2}

¹Institut Pasteur – Bioinformatics and Biostatistics Hub – C3BI, USR 3756 IP CNRS – Paris, France

²Institut Pasteur – Biomics – CITECH – Paris, France

^{*}Equally contributing authors

Background

- Quantitative metagenomics is an approach broadly employed to identify associations between a microbiome and an environmental / **individual** condition (disease, geographical condition, ...).
- To perform this type of approach, targeted sequencing of rDNA or shotgun sequencing is performed and quantitative measures are obtained by mapping the reads against the set of OTU identified or a gene catalog.
- These data can be analyzed by developing R scripts including statistical analysis (metagenomeseq, momr, edgeR, ...) or web interface dedicated to visualization (MEGAN, Shiny-phyloseq, Phinch).
- The lack of easy-access methods that ****** providing both relevant statistical analysis and specific visualization is a critical issue.
- Here we present SHAMAN, a Shiny-based application that offers an unified experience for the analysis of quantitative metagenomics data.



Contact: shaman@pasteur.fr

Start with SHAMAN



MBMA

SHAMAN requires as input of (1) a count table and (2) an annotation table (as csv or tsv file) or a BIOM file.

These data are provided by most pipelines like: MASQUE (docker: aghozlane/masque) for targeted metagenomics,

shotgun

- metagenomics (https://github.com/anitaannamale/MBMA).
- Imported dataset is analysed to identify which taxonomical level is



Quantitative hetragenomics (target file)	Heatmap	Sep 9th 2016
at OTU level Generalized Merge normalized counts at the user		
Selected level Sinificial Analysis: DE Seq2 Define contrasts Comparisons)	Rarefaction	
University and the second seco	Visualization	

SHAMAN is freely accessible through a web interface at http://shaman.c3bi.pasteur.fr/ and docker hub at aghozlane/shaman.



for

- the appropriate for the most analysis.
- Here 40% of the OTU are annotated at the Genus level and 38 different genera identified.

Experimental design / Statistical modeling

An experimental design table must be provided. The table is used to assign each sample to a condition, a time, an individual or an other metadata.

SHAMAN process is divided into two steps:

Normalization: The OTU/gene count is normalized using size factors defined as the median of the ratio between the count and the geometric mean of each OTU/gene (1) [Anders 2010].

$$s_j = median_i \frac{c_{ij}}{(\prod_{k \in S_i} c_{ik})^{1/n}}$$
(1)

Assume that $C = (c_{ij})_{1 \le i \le k; 1 \le j \le n}$ is a count table.

k and n correspond to the number of features (like OTU) and the number of samples, respectively. c_{ij} represents the count of feature *i* in sample *j*. s_j is

xperimental design	-	Options			-	
elect your target file Separator:	Select the taxonomy level		Linear mod	el options		
Browse Target.tsv Tab	▼ Genus ▼	Type of transformation Inden	endent filtering	p-value adjustement	Level of significance	
Upload complete		VST O Th	rue	 BH 	0.05	
elect the variables Ad	d interactions	O rlog O Fa	alse	O BY	0.05	63
condition time mice	condition:mice condition:time	Cooks cut-off Local	function	Relationship		
▶ Run analysis		Auto Auto No cut-off O	ledian borth	Parametric		
		O Value	north	O Local		
Target file overview	-		Options for the no	ormalization step		
Show 10 v entries	Search:					
label 🔶 condition	time mice 🔶	Compute geometric Norm	alization by:	Define your own size factors	Separator:	
WT.m2 WT.m2 WT	t0 m1			Browse No fil	e:	
WT2.38 WT2.38 WT	t2 m1					_
WT.m3 WT.m3 WT	t0 m2					
WT3.39 WT3.39 WT	t2 m2	Contrasts (New)		-	Defined contrasts	
WT.m4 WT.m4 WT	t0 m3	Compare To	For		Contrasts	i
WT4.40 WT	t2 m3	Delta 🔻 DeltaCompl	All 🔻	TAdd	WT_0_24 DeltaCompl_0_24	•
WT5.41 WT5.41 WT	t2 m4				Delta_0_24 WT_0_6	C
WT.m5 WT.m5 WT	t0 m4	Contracts (advanced user)		_	🗙 Remove 🕹 Export	L L
WT2.56 WT2.56 WT	t3 m1					f f
WT3.57 WT3.57 WT	t3 m2	Select a file of contrasts				
Count table (Genus) Show 10 ∨ entries	- Search:	Upload complete Define contrasts by yourself Contrast name	+ Add contrast			۱
WT.m2 🔶 WT2.38	WT.m3 🔷 WT3.39 🔷 WT.	Intercept				
[Eubacterium] 1054 521	. 394 228	0		-		
Acetatifactor 0 4	1 3	conditionDelta				
Acetitomaculum 12 25	8 11	0		÷		
Alistipes 2381 1533	1858 2073	conditionDeltaCompl				
Allobaculum 8 7	150 9	0		÷		
Alloprevotella 8 0	8 4	conditionWT				
Anaerofustis 1 0	2 0	0		\$		
Anaeroplasma 393 275	313 43	timet0				
Anaerotruncus 114 95	52 68	0		÷		
Bacteroides 5309 906	3578 9625	timet2				
Showing 1 to 10 of 48 entries	>	0		\$		
Previous 1	2 3 4 5 Next	timet3				
LEXPORT NORMALISED COUNTS	relative abundance	0		÷		
		micem1				
		0		\$		

The user defines a contrast vector to extract features that are significantly different in abundance according to the experimental design. A guided and expert mode are available in SHAMAN to perform this step.

nificant features are summarized in a table icating their **base mean** (mean normalized int), fold change (how much the count varies m one condition to the other) and adjusted pue.

ihow 25 v entries Search:				
Id	baseMean 🍦	FC 🔶	log2FoldChange 崇	padj 🍦
Turicibacter	995.24	0.004	-7.899	1.08466957119766e-9
Allobaculum	103.04	0.008	-6.955	0.000031191170815586
Roseburia	173.49	3.965	1.987	0.00152565569842405
Enterorhabdus	38.35	0.156	-2.676	0.00210484632042579
Ruminococcaceae	2369.48	0.324	-1.628	0.00364694347726387
Parasutterella	10.73	0.042	-4.56	0.00747095659486352
Mucispirillum	50.33	4.941	2.305	0.0186705242331407
Prevotellaceae	49.73	0.1	-3.321	0.0198884561087325
Alloprevotella	2.7	0.076	-3.727	0.0236316541392707
Streptococcus	8.73	0.221	-2.175	0.0264348153495083
Listeria	157.45	16.58	4.051	0.034455649855071

- the size factor of sample *j*.
- Modelization: DESeq2 local regression is used to get robust estimation of the OTU dispersion and a Generalized Linear Model is defined [Love 2014].

Visualizations

SHAMAN visualizations fall into three categories:

Diagnostic plots: These plots allow a quality check of the data.





- relevance of the statistical modeling.

Analysis plots: These plots are generated to highlight the differences in abundance identified by differential analysis.





Conclusion / Future work

SHAMAN:

 Combines strong statistical approach with a dynamic visualization interface.

features:

- ✓ Integrates most of the analysis required for publication.
- Functions in real time.
- ✓ Already used in a publication [Quereda et al. PNAS 2016].

