# SHAMAN : a SHiny Application for Metagenomic ANalysis

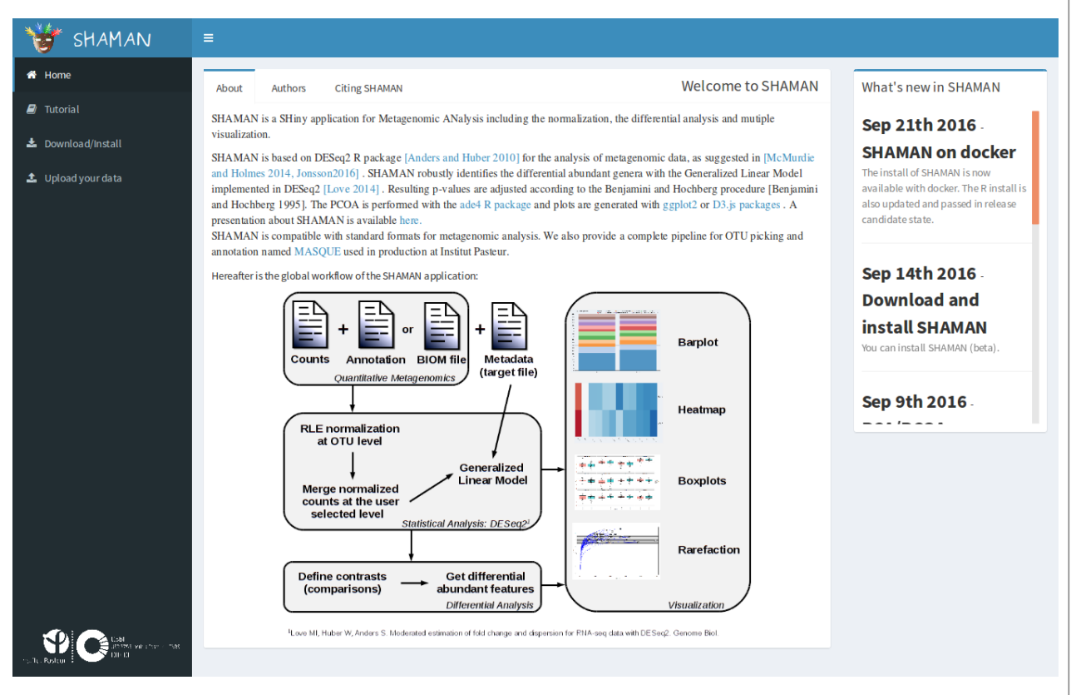
**Amine Ghozlane**[1,2*], **Stevenn Volant**[1*], Hugo Varet[1,2], Christophe Malabat[1], Pierre Lechat[1], Sean Kennedy[2], Marie-Agnès Dillies[1,2]

[1]Institut Pasteur – Bioinformatics and Biostatistics Hub – C3BI, USR 3756 IP CNRS – Paris, France
[2]Institut Pasteur – Biomics – CITECH – Paris, France
[*]Equally contributing authors

Contact: shaman@pasteur.fr

## Background

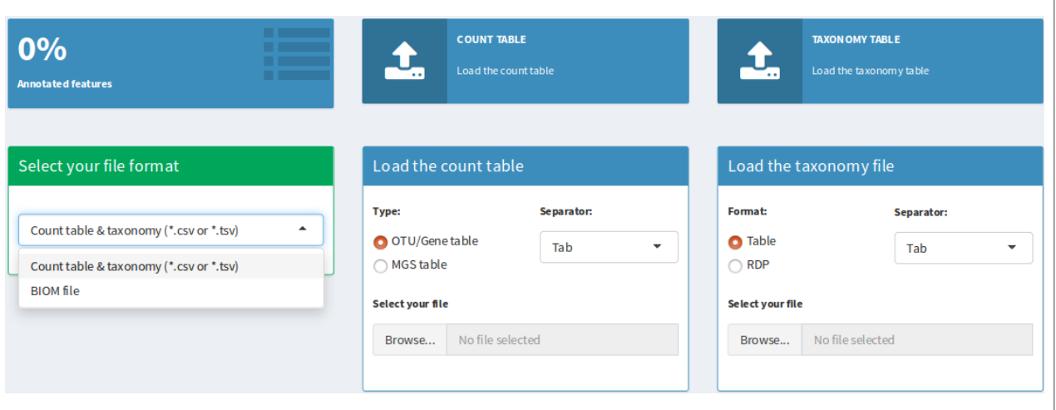
- Quantitative metagenomics is an approach broadly employed to identify **associations** between a **microbiome** and an **environmental / individual** condition (disease, geographical condition, …).
- To perform this type of approach, **targeted** sequencing of rDNA or **shotgun** sequencing is performed 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.
- SHAMAN is freely accessible through a web interface at **http://shaman.c3bi.pasteur.fr/** and docker hub at aghozlane/shaman.

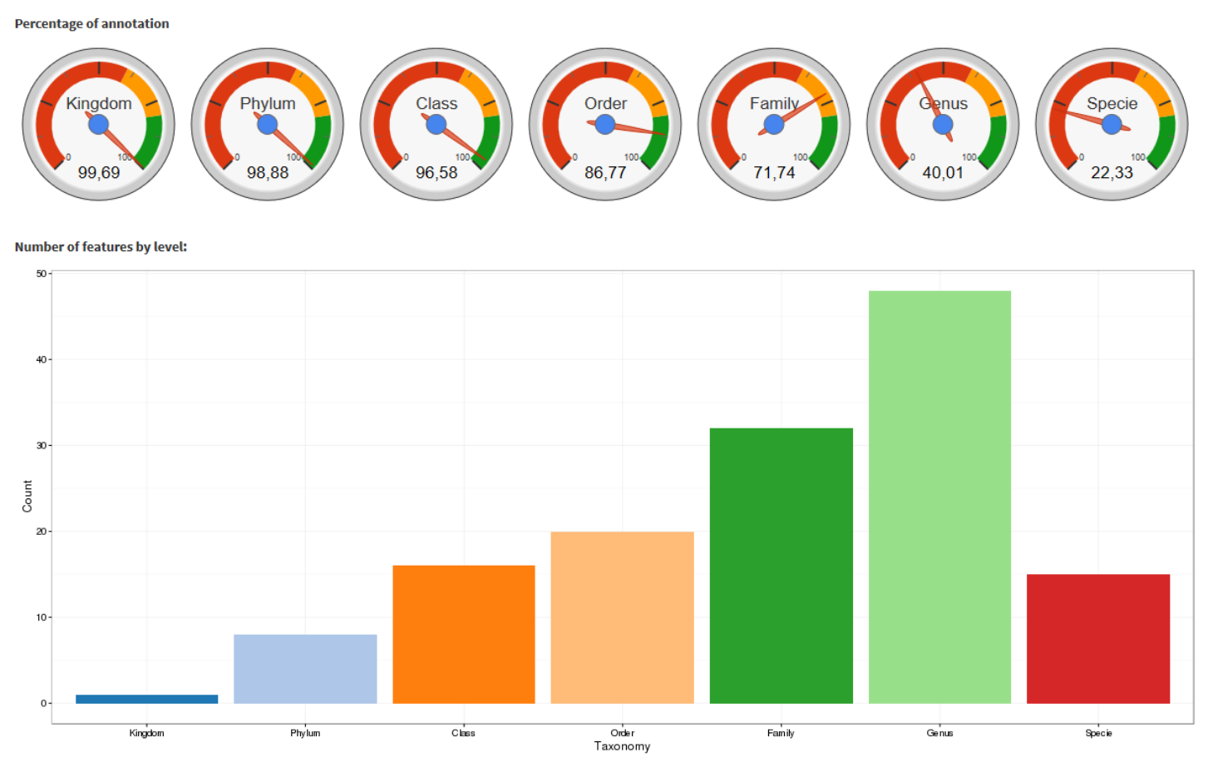


## Start with SHAMAN

1 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,
  - MBMA for shotgun metagenomics (https://github.com/anitaannamale/MBMA).





- Imported dataset is analysed to identify which taxonomical level is the most appropriate for the analysis.
- Here 40% of the OTU are annotated at the *Genus* level and 38 different genera identified.

## Experimental design / Statistical modeling

2 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.

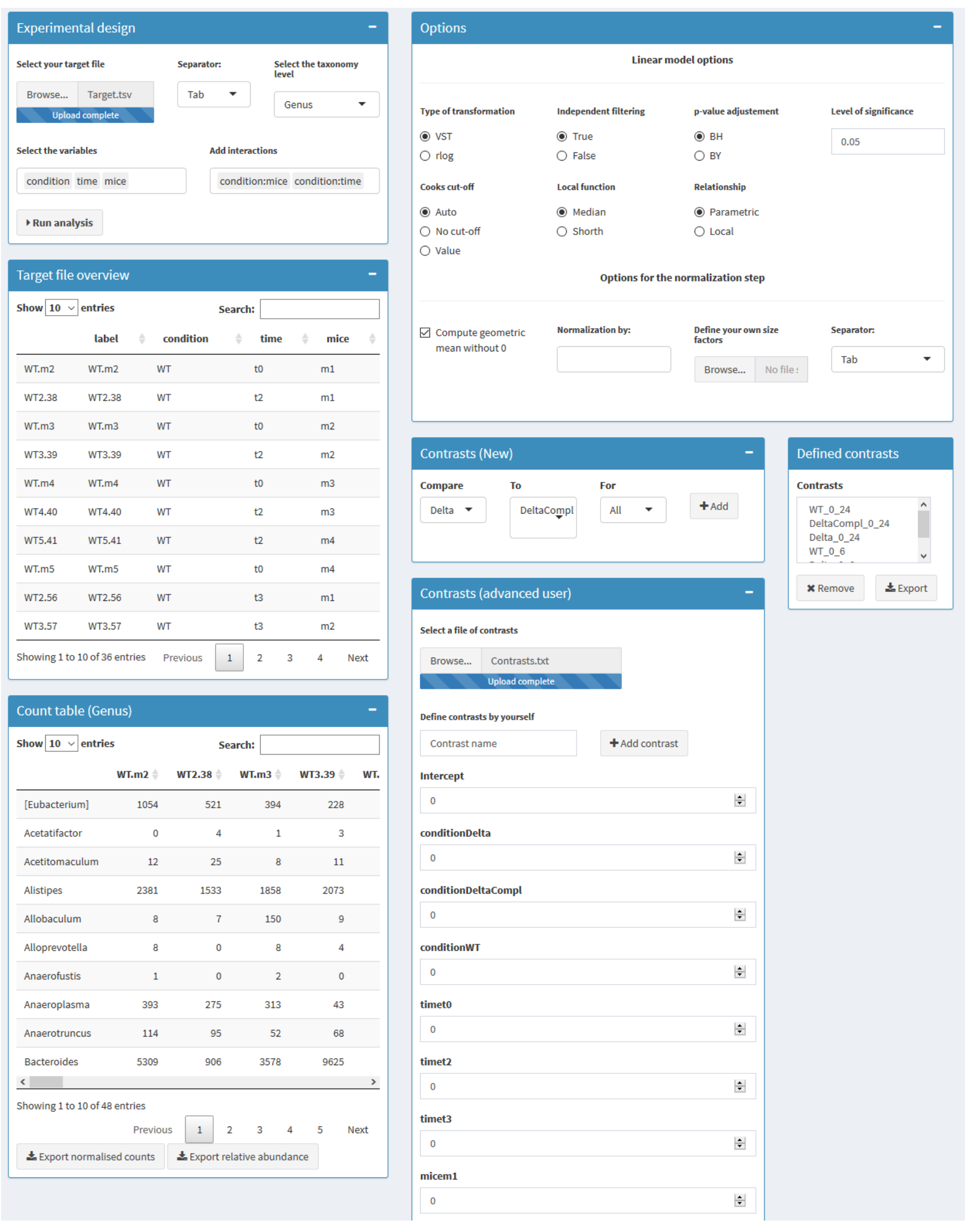
SHAMAN process is divided into two steps:

- 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].

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

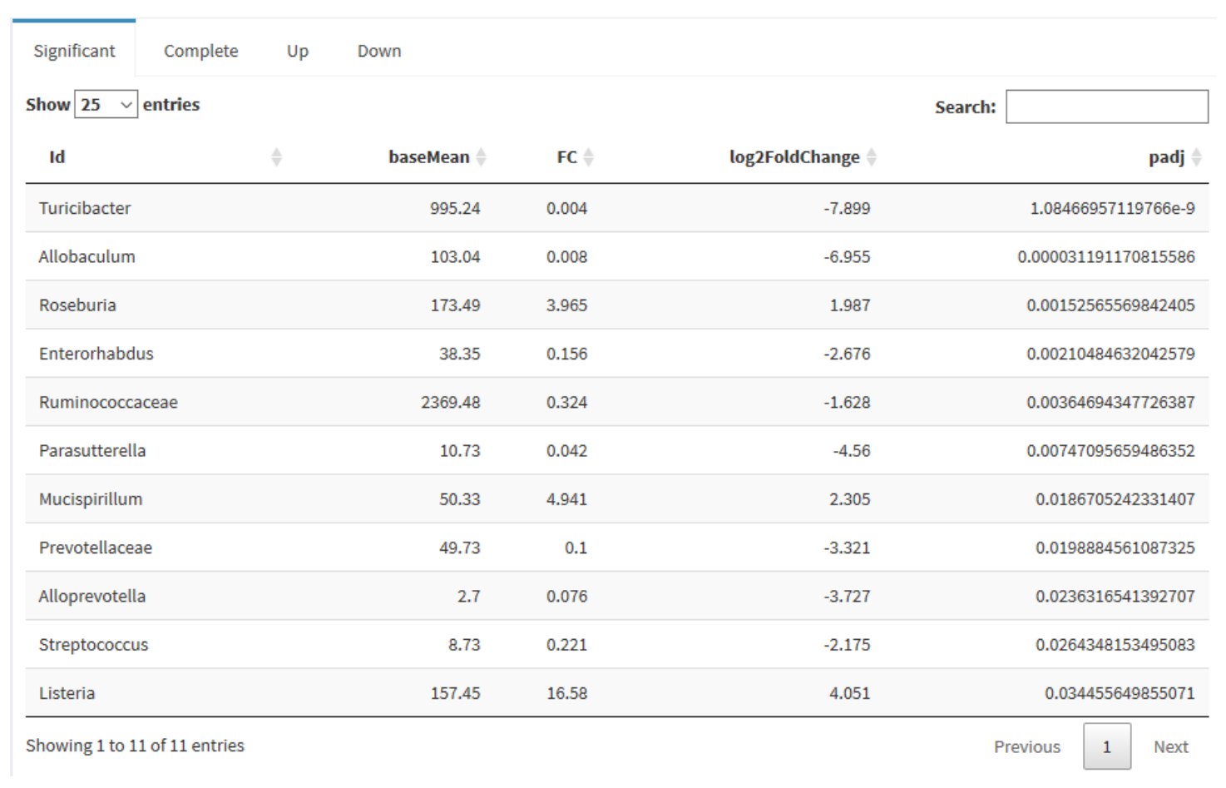
Assume that $C = (c_{ij})_{1 \le i \le k; 1 \le j \le n}$ is a count table.
*k* and *n* correspond to the number of features (like OTU) and the number of samples, respectively. $c_{ij}$ represents the count of feature *i* in sample *j*. $s_j$ is the size factor of sample *j*.

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



3 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.

- Significant features are summarized in a table indicating their **base mean** (mean normalized count), **fold change** (how much the count varies from one condition to the other) and **adjusted p-value**.



Show 25 entries  Search:

| Id | baseMean | FC | log2FoldChange | 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 |

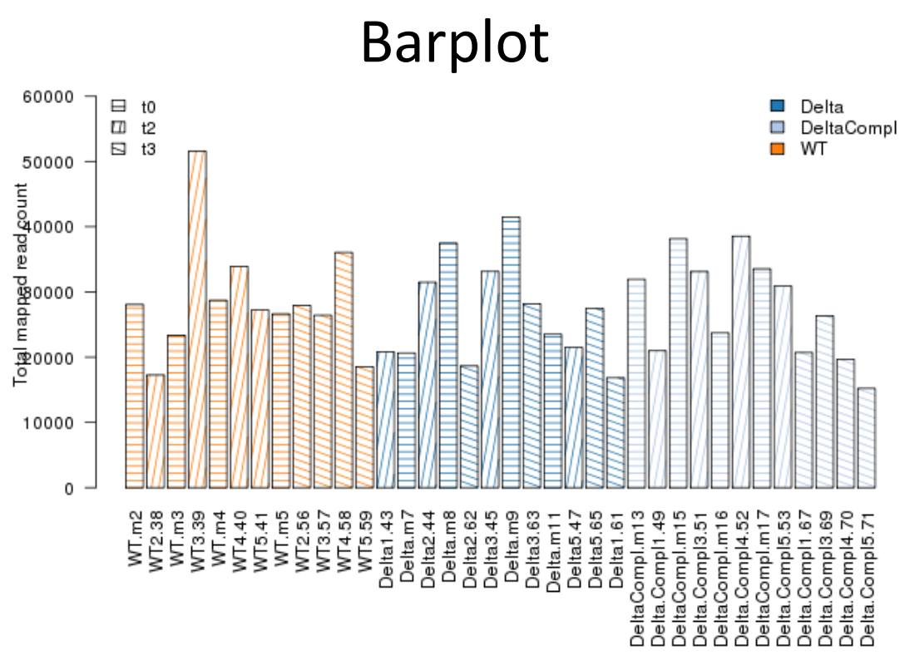
Showing 1 to 11 of 11 entries  Previous 1 Next

## Visualizations

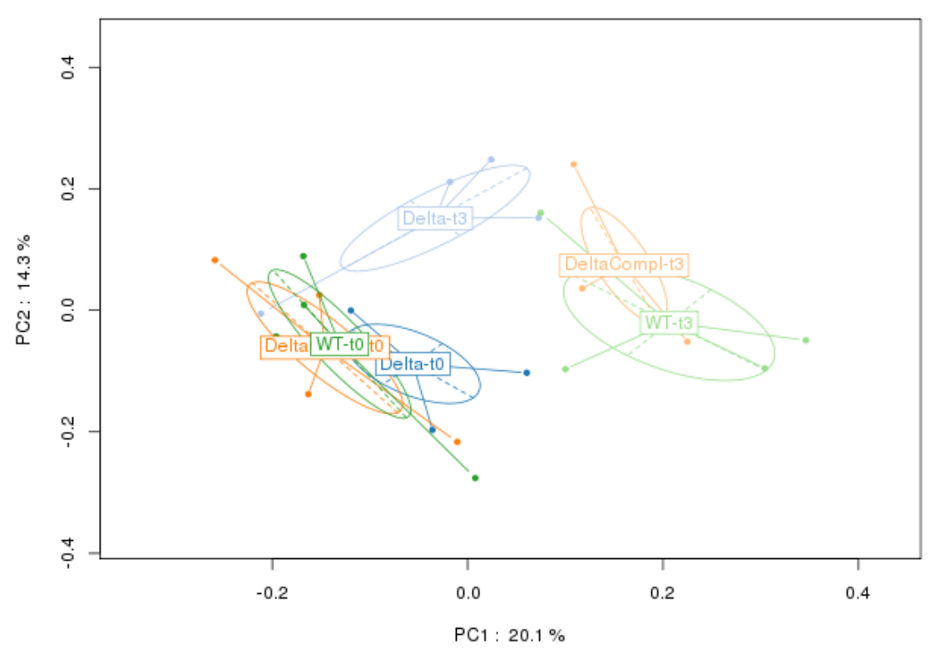
SHAMAN visualizations fall into three categories:

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

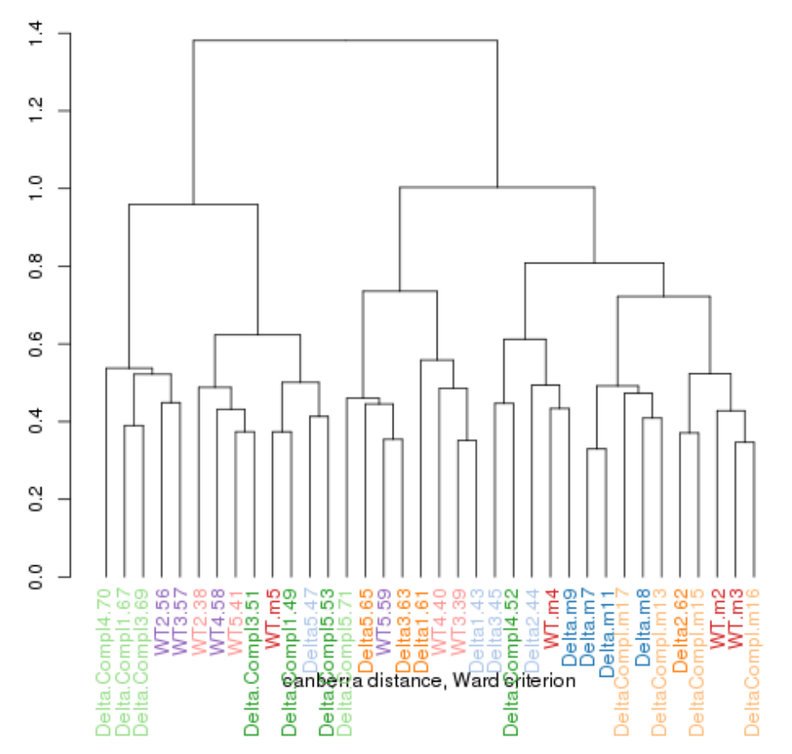
Barplot



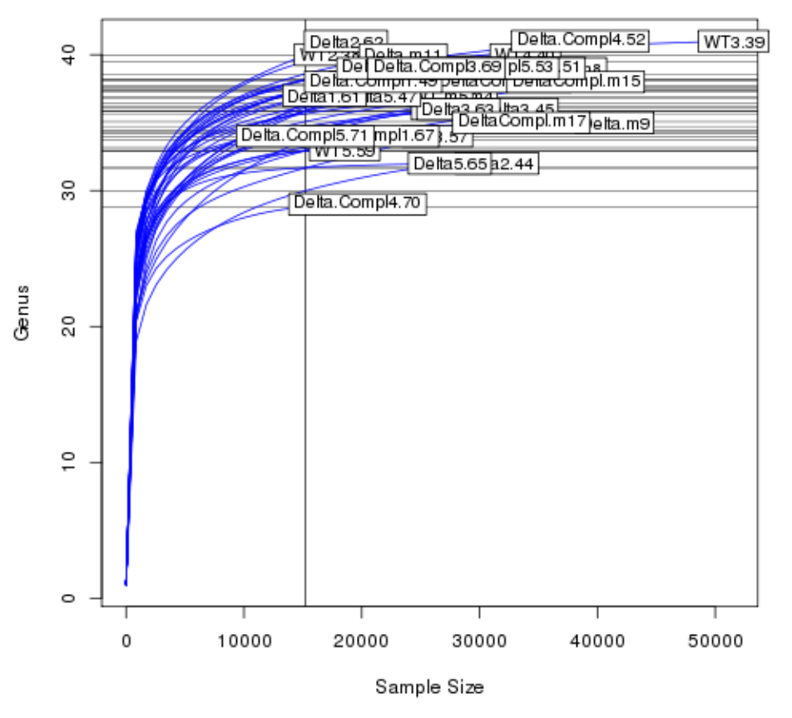
PCOA/PCA

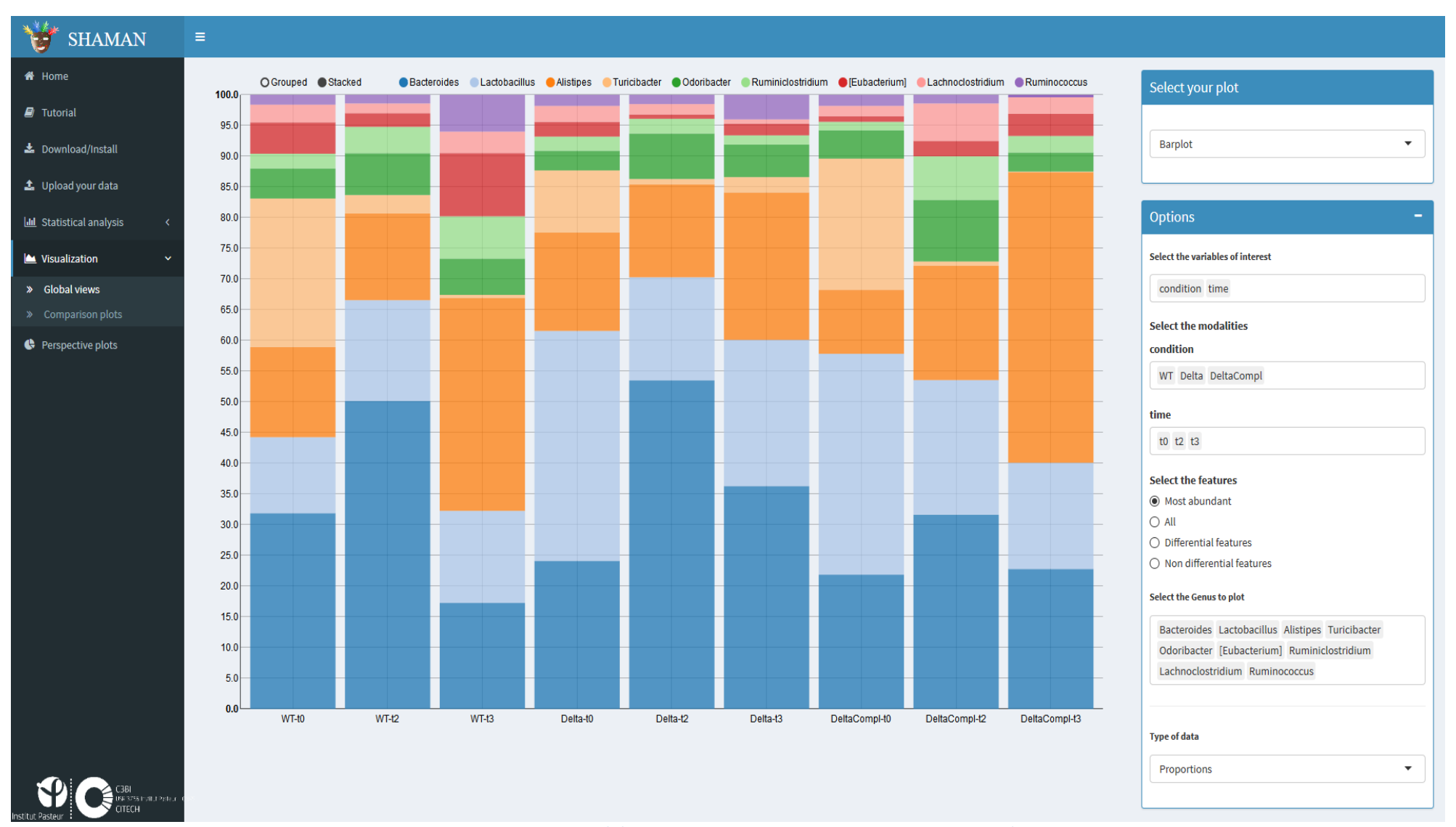


Clustering



Rarefaction curves

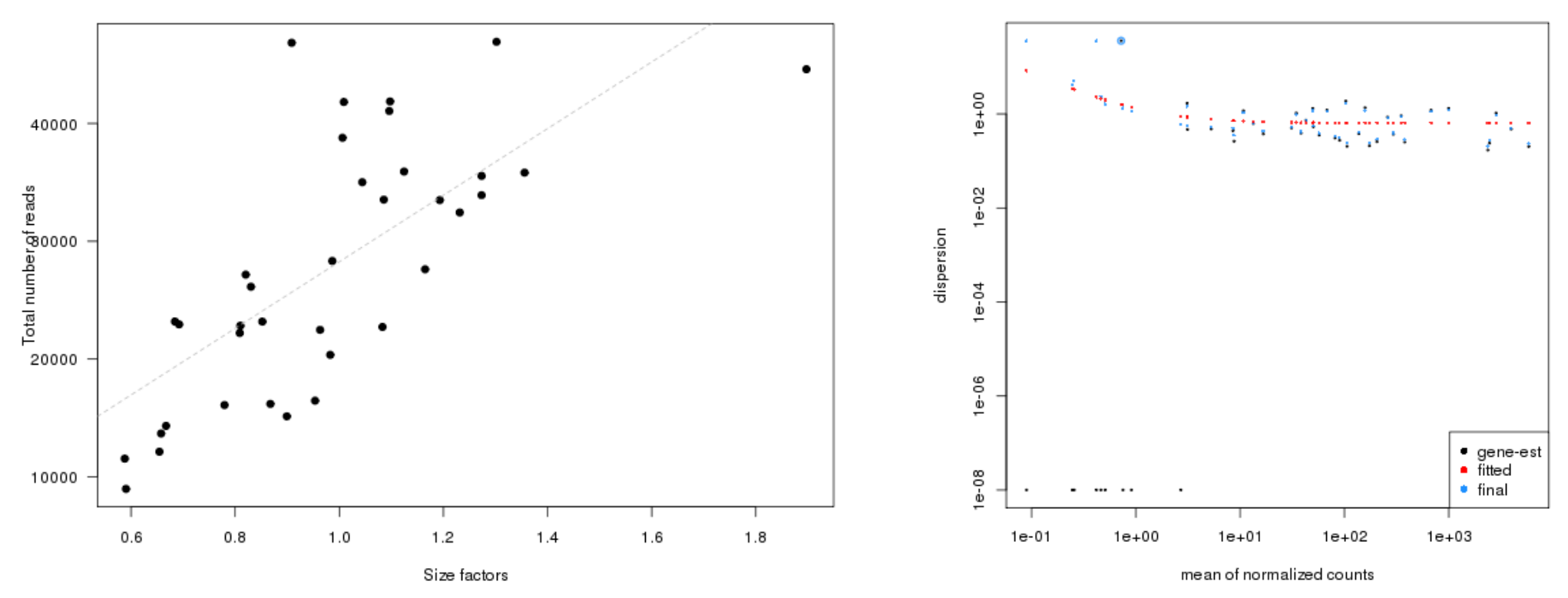




http://shaman.c3bi.pasteur.fr/

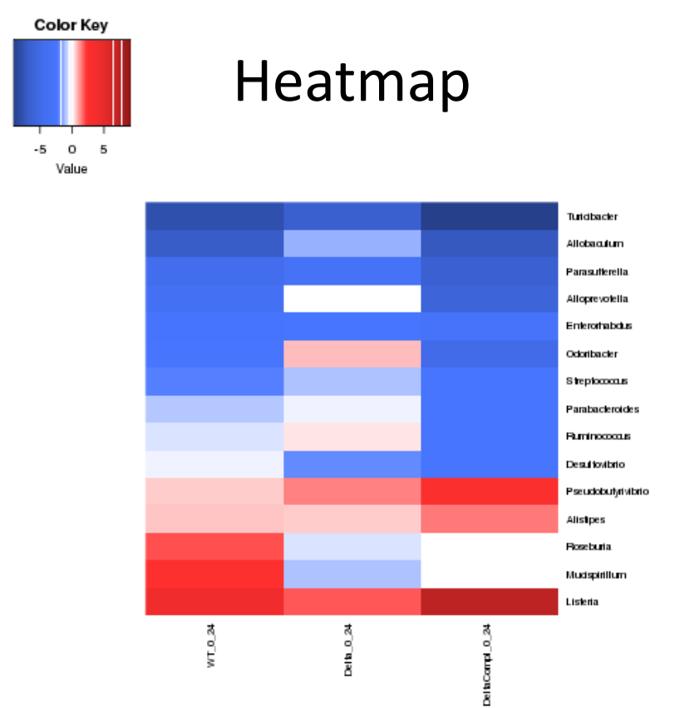
- Statistical modeling plots: These plots assess the relevance of the statistical modeling.

Scatter plots of size factors and dispersion estimation

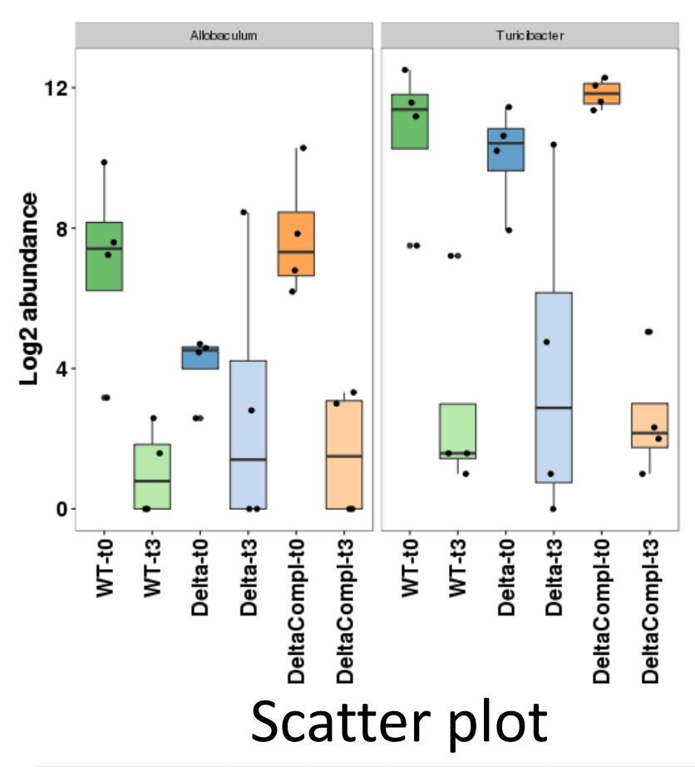


- Analysis plots: These plots are generated to highlight the differences in abundance identified by differential analysis.

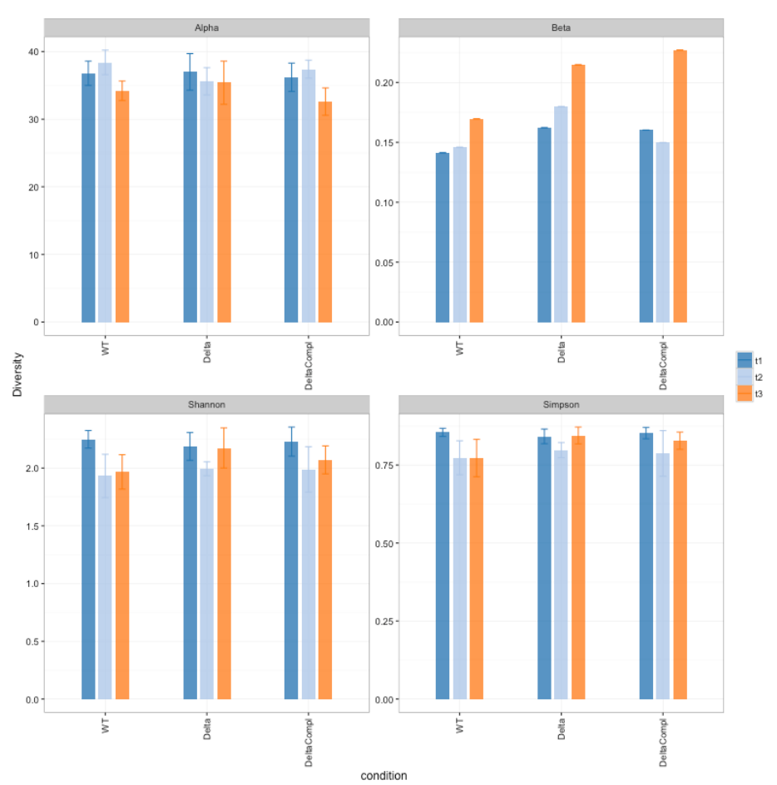
Heatmap



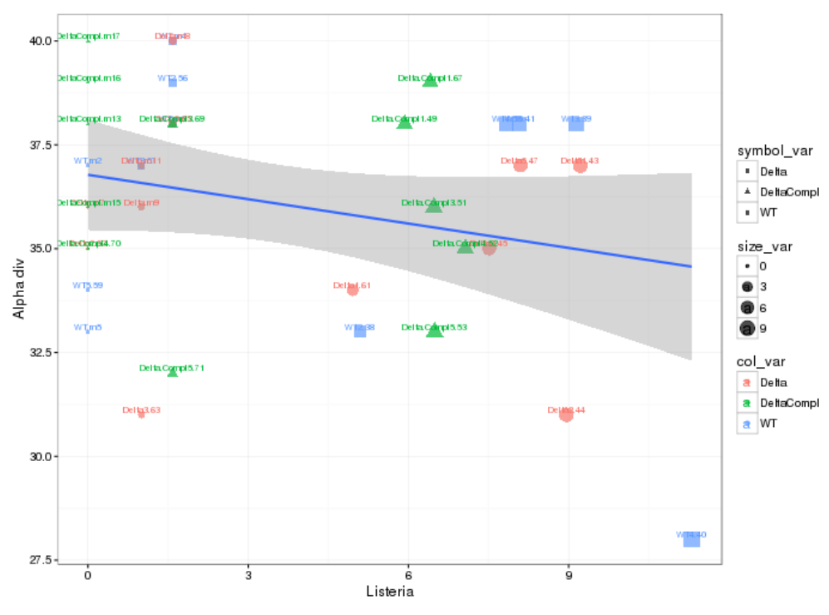
Boxplot



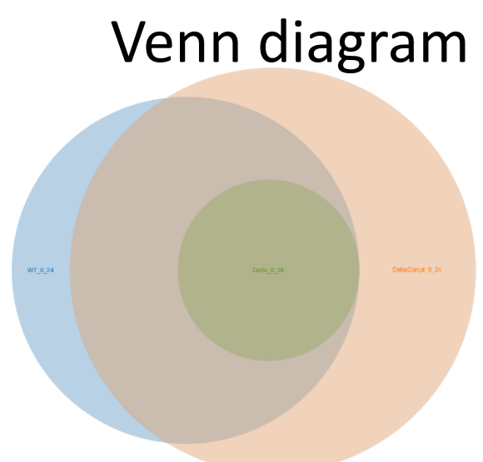
Diversity plot



Scatter plot



Venn diagram

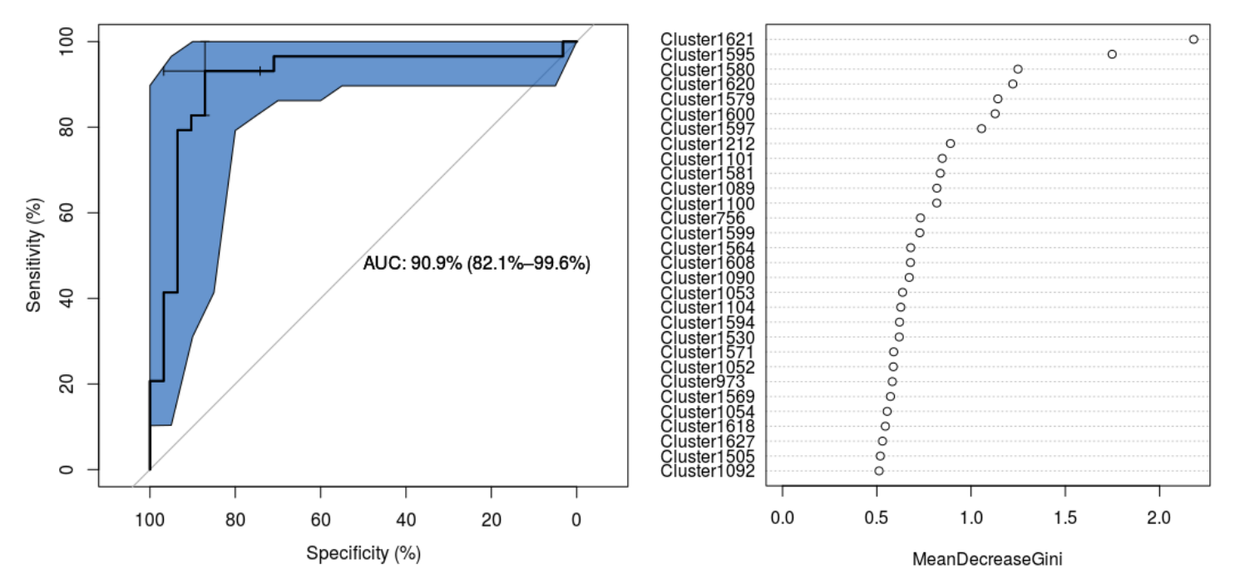


## Conclusion / Future work

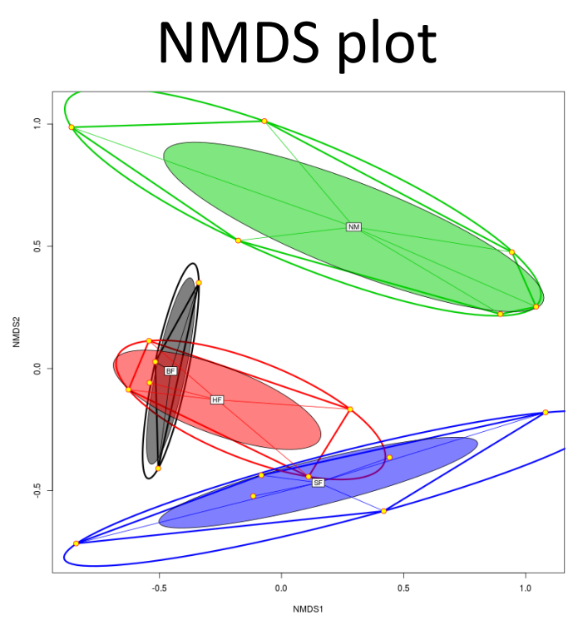
SHAMAN:
- ✓ Combines strong statistical approach with a dynamic visualization interface.
- ✓ Integrates most of the analysis required for publication.
- ✓ Functions in real time.
- ✓ Already used in a publication [Quereda et al. PNAS 2016].

Forthcoming features:

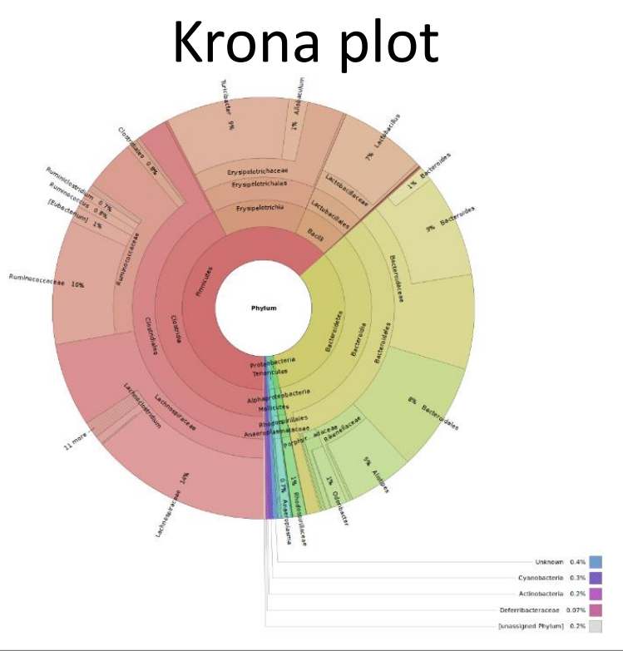
Random forest



NMDS plot



Krona plot



Taxonomy plot