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], Christophe Malabat[1], Sean Kennedy[2], and Amine Ghozlane[1,2,✉]

[1]Hub de Bioinformatique et Biostatistique – Département Biologie Computationnelle, Institut Pasteur, USR 3756 CNRS, Paris, France.
[2]Biomics – Département Génomes et Génétique, Institut Pasteur, Paris, France.

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 representation of the detected signatures. Such an analysis is normally obtained by using a set of different applications that require specific expertise for installation, data processing and in some case, programming skills. 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 among the largest panels of interactive visualizations as compared to the other options that are currently available. SHAMAN is specifically designed for non-expert users who may benefit from using 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 datasets (a mock community sequencing and published 16S metagenomic data), we illustrate the strengths of SHAMAN in quickly performing a complete metataxonomic analysis.

Metagenomics | Statistical analysis | Visualization
Correspondence: ***amine.ghozlane@pasteur.fr***

DRAFT

Introduction

Quantitative metagenomic techniques have been broadly deployed to identify associations between microbiome and environmental or individual factors (e.g., disease, geographical origin, etc.). Analyzing changes in the composition and/or the abundance of microbial communities yielded promising biomarkers, notably associated with liver cirrhosis(1), diarrhea(2), colorectal cancer(3), or associated with various pathogenic(4) or probiotic effects(5) on the host.

In metataxonomic studies, a choice is made prior to sequencing in order to specifically amplify one or several regions of the rRNA (usually the 16S or the 18S rRNA for procaryotes/archaea and the ITS, the 23S or the 28S rRNA for eukaryotes) so that the composition of microbial communities may be characterized with affordable techniques.

A typical workflow includes successive steps: (i) OTU (Operational Taxonomic Unit) picking (dereplication, denoising, chimera filtering and clustering)(6), (ii) OTU quantification in each sample and (iii) OTU annotating with respect to a reference taxonomic database. This process may require substantial computational resources depending on both the number of samples involved and the sequencing depth. Several methods are currently available to complete these tasks, such as Mothur(7), Usearch(8), DADA2(9) or Vsearch(10). The popular Qiime(11) simplifies these tasks (i to iii) and visualizations by providing a python-integrated environment. Schematically, once data processing is over, both a contingency table and a taxonomic table are obtained. They contain the abundance of OTUs in the different samples and the taxonomic annotations of OTUs, respectively. The data are normally represented in the standard BIOM format(12).

Statistical analysis is then performed to screen significant variation in microbial abundance. To this purpose, several R packages were developed, such as Metastats(13) or Metagenomeseq(14). It is worth noticing that other approaches which were originally designed for RNA-seq, namely DESeq2(15) and EdgeR(16), are also commonly used to carry out metataxonomic studies(17, 18). They provide an R integrated environment for statistical modelling in order to test the effects of a particular factor on OTU abundance. Nevertheless using all of these different methods requires a technical skills in Unix, R and experience in processing metagenomics data. To this end, we developed SHAMAN in order to provide a method that simplifies the analysis of metataxonomic data, especially for users who are not familiar with the technicalities of bioinformatic and statistical methods that are commonly applied in this field.

SHAMAN is an all-inclusive approach to estimate the composition and abundance of OTUs, based on raw sequencing data, and to perform the statistical analysis of the processed files. First, the user can submit raw data in FASTQ format and define the parameters of the bioinformatic workflow. The output returns a BIOM file for each database used as reference for annotation, a phylogenetic tree in Newick format as well as FASTA-formatted sequences of all OTUs that were identified. The second step consists in performing statistical analysis. The user has to provide a "target" file that associates each sample with one or several explanatory variables. These variables are automatically detected in the target file. An au-

tomatic filtering of the contingency matrix of OTUs may be activated in order to remove features with low frequency. Setting up the contrasts to be compared was also greatly simplified. It consists in filling in a form that orients the choices of users when defining the groups of interest. Several options to visualize data are available at three important steps of the process: quality control, bio-analysis and contrast comparison. At each step, a number of common visual displays are implemented in SHAMAN to explore data. In addition, SHAMAN also includes a variety of original displays that is not available in other applications such as an abundance tree to visualize count distribution according to the taxonomic tree and variables, or the logit plot to compare feature p-values in two contrasts. Figures may be tuned to emphasize particular statistical results (e.g., displaying significant features in a given contrast, performing intersection between contrasts), to be more specific (e.g. feature abundance in a given modality) or to improve the aesthetics of the graph (by changing visual parameters). Figures fit publication standards and the corresponding file can be easily downloaded.
Several web applications were developed to analyze data of metataxonomic studies, notably, FROGS(19) as well as Qiita(20) for bioinformatic data processing, Shiny-phyloseq(21) for statistical analysis, Metaviz(22) and VAMPS2(23) that make a particular focus on data visualization. While these interfaces propose related functionalities, the main specificity of SHAMAN is to combine of all these steps in a single user-friendly application. Last, SHAMAN may register a complete analysis which may be of particular interest for matters of reproducibility.

DESCRIPTION

SHAMAN is implemented in R using the shiny-dashboard framework. The application is divided into three main components (Fig. S1): a bioinformatic workflow to process the raw FASTQ-formatted sequences, a statistical workflow to normalize and further analyse data, as well as a visualization platform.

Metataxonomic pipeline. The bioinformatic workflow implemented in SHAMAN relies on the Galaxy platform(24) that provides modular and scalable analyses. SHAMAN includes a daemon-program (written in Python) using bioblend(25) to communicate with Galaxy. It is worth noticing that previous studies, e.g. performed on mosquito microbiota(26), showed that some non-annotated OTUs turned to be sequences of the host organism. To overcome such issues, the user can optionally filter out reads that align with the host genome and the PhiX174 genome (used as a control in Illumina sequencers). The latter task is performed with Bowtie2 v2.2.6(27). By default, quality of reads is checked with AlienTrimmer(28) v0.4.0, a software for trimming off contaminant sequences and clipping. Paired-end reads are then merged with Pear(29) v0.9.10.1. OTU picking, taxonomic annotation and OTU quantification are performed using Vsearch(10) v2.3.4.0, a software which is both accurate and efficient (6). The process also includes several steps of dereplication, singleton removal and chimera detection. By default, clustering is performed with a threshold of 97% in sequence identity. The input amplicons are aligned against the set of detected OTUs to create a contingency table containing the number of amplicons assigned to each OTU. The taxonomic annotation of OTUs is performed based on various databases, i.e., with SILVA(30) rev. 132 SSU (for 16S, 18S) and LSU (for 23S and 28S sequences), Greengenes(31) (for 16S, 18S sequences) and Underhill rev. 1.6.1(32), Unite rev. 8.0(33) and Findley(34) for ITS sequences. These databases are kept up-to-date every two month with biomaj.pasteur.fr. OTU annotations are filtered according to their identity with the reference(35). Phylum annotations are kept when the identity between the OTU sequence and reference sequence is $\geq$ 75%, $\geq$ 78.5% for classes, $\geq$ 82% for orders, $\geq$ 86.5% for families, $\geq$ 94.5% for genera and $\geq$ 98% for species. In addition, a taxonomic inference made based on a naive Bayesian approach, RDP classifier(36) v2.12, is systematically provided. By default, RDP annotations are included whenever the annotation probability is $\geq$ 0.5. All the above-mentioned thresholds may be tuned by the user.
A phylogenetic analysis of OTUs is provided: multiple alignments are obtained with Mafft(37) v7.273.1, filtering of regions that are insufficiently conserved is processed using BMGE(38) v1.12 and finally, FastTree(39) v2.1.9 is used to infer the phylogenetic tree. Based on the latter tree, a Unifrac distance(40) may be computed in SHAMAN to compare microbial communities. The outcomes of the overall workflow are stored in several files: a BIOM file (per reference database), a phylogenetic tree as well as a summary file describing the number of elements passing the differents steps of the workflow. The data are associated to a key that is unique to a project. Such a key allows to automatically reload all the results previously obtained in a given project.

Statistical workflow. The statistical analysis in SHAMAN is based on DESeq2 which is a method to model OTU counts with a negative binomial distribution. It is known as one of the most accurate approach to detect differentially abundant bacteria in metagenomic data(17, 18). Relying on a robust estimation of variation in OTUs, the DESeq2 method shows suitable performances with datasets characterized by a relatively low number of observations per group together with a high number of OTUs.
This method typically requires the following input files: a contingency table, a taxonomic table and a target file describing the experimental design. These data are processed to generate a meta-table that assign to each OTU a taxonomic annotation and a raw count per sample.

Normalization. Normalization of the raw counts is one of the key issues when analyzing microbiome experiments. The uniformity of the sequencing depth is affected by sample preparation and dye effects(41). Normalizing data is therefore expected to increase the accuracy of comparisons. It is done by adjusting the abundance of OTUs across samples. Four different normalization methods are currently implemented in SHAMAN. For the sake of consistency, all of these

DRAFT

methods are applied at the OTU level.

A first method is the relative log expression (RLE) normalization and is implemented in the DESeq2 package. It consists in calculating a size factor of each sample, i.e., a multiplication factor that increases or decreases the OTU counts in samples. It is defined as the median ratio, between a given count and the geometric mean of each OTU. Such a normalization was shown to be suited for metataxonomic studies(17). In practice, many OTUs are found in a few samples only, which translate into sparse count matrices(14). In this case, the RLE method may lead to a defective normalization - as only a few OTU are taken into account - or might be impossible if all OTUs show a null abundance in one sample at least. In the Phyloseq(42) R package, the decision was made to replace the null abundance by a count of 1. In SHAMAN, we decided to include two new normalization methods. They are modified versions of the original RLE so that they better account for matrix sparsity (number of zero-valued elements divided by the total number of elements). In the *non-null normalization* (1) cells with null values are excluded from the computation of the geometric mean. This method therefore takes all OTUs into account when estimating the size factor. In the second method that we coined as the *weighted non-null normalization* (2), weights are introduced so that OTUs with a big number of occurrences have a higher influence when calculating the geometric mean.

Assume that $C = (c_{ij})_{1 \leq i \leq k; 1 \leq j \leq n}$ is a contingency table where k and n are the number of features (e.g. OTUs) and the number of samples, respectively. Here, c_{ij} represents the abundance of the feature i in the sample j. The size factor of sample j is denoted by s_j.

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

$$s_j^{(2)} = w.median_i \frac{c_{ij}}{\left(\prod_{k \in S_i} c_{ik}\right)^{1/n_i}}, \quad \textbf{(2)}$$

where S_i stands for the subset of samples with non null values for the feature j and n_i is the size of this subset. The function $w.median$ corresponds to a weighted median.

An alternative normalization technique is the *total counts*(43) which is convenient for highly unbalanced OTU distribution among samples.

Using a simulation-based approach, we addressed the question of the performance of the *non-null* and the *weighted non-null normalization* techniques when the matrix sparsity and the number of observations increase. We compared these new methods to those normally performed with DESeq2 and Phyloseq. To do so, we normalized 500 matrices with varied sparsity levels (i.e., 0.28, 0.64 and 0.82) and a different number of observations (i.e., 4, 10 and 30). We calculated the average coefficient of variation (CVmean)(44) for each normalization method (Fig. S2). Considering that these OTUs are assumed to have relatively constant abundance within the simulations, the coefficient of variation is expected to be lower when the normalization is more efficient. Overall, the *non-null* and the *weighted non-null* normalization methods exhibited a lower coefficient of variation as compared to the other methods, when sparsity in the count matrix is high and the number of observations is increased. These differences were clear especially when comparing DESeq2 and Phyloseq to the *weighted non-null normalization* (sparsity ratio of 0.28, 0.64 and 0.82, with 30 samples; t-tests $p < 0.001$) (Fig. S2).

Contingency table filtering. In metataxonomic studies, contingency tables are often very sparse and after statistical analysis, some significant differences among groups may not be of great relevance. This may arise when a feature, distributed in many samples with a low abundance, is slightly more abundant in a group of comparison. These artifacts are generally excluded by DESeq2 with an independent filtering. Furthermore, if a feature is found in a few samples only, it may lead to non-reliable results when its abundance is high (when it is not 0). Such distributions may also impact the normalization process as well as the dispersion estimates. In order to avoid misinterpretation of results, we propose an optional extra-step of filtering: by excluding features characterized by a low abundance and/or a low number of occurrence in samples (e.g. features occurring in less than 20% of the samples). To set a by-default abundance threshold, SHAMAN search for an inflection point at which the curve between the number of observations and the abundance of feature changes from being linear to concave. This process is performed with linear regression in the following manner:

1. We define I the interval $\left[min_j\left(\sum_i c_{ij}\right); \frac{\sum_{ij} c_{ij}}{k}\right]$.
2. For each $x \in I$, we compute $h(x)$ defined as the number of observations with a total abundance higher than x.
3. We compute the linear regression between $h(x)$ and x.
4. The intercept is set as the default threshold.

(see Appendix 1 for more information). This extra-filtering aims at refining the first filtering processed with DESeq2 and may lead to a significant decrease of the computation time. The impact of filtering steps may be visually assessed with plots displaying the features that will be included in the analysis and those that will be discarded.

Statistical modelling. The statistical model relies on the variables that are available in the file of experimental design. By default, all variables are included in the model but the end-user can edit this selection and further add interactions between variables of interest. In addition, other variables such as batches or clinical data (e.g., age, sex, etc.) may be used as covariates. SHAMAN then automatically checks whether the model is statistically suitable (i.e., whether all the model parameters may be estimated). When it is not the case, an warning message appears and a "how to" box proposes a practical way to solve the issue. In SHAMAN, statistical models may be fitted at any taxonomic levels. Normalized counts are summed up within a given a taxonomic level.

DRAFT

To extract features that exhibit significant differential abundance (between two groups), the user must define a contrast vector. Both a guided mode and an expert mode are available in SHAMAN. In the guided mode, the user specifies the groups to be compared using a drop-down menu. This mode is only available for DESeq2 v1.6.3 which is implemented in DESeq2shaman package (https://github.com/aghozlane/DESeq2shaman). In advanced comparisons, the user may define a contrast vector by specifying coefficients (e.g., -1, 0, 1) assigned to each variable.

Visualization. After running a statistical analysis, many displays are available:
(i) Diagnostic plots (such as barplots, boxplots, PCA, PCoA, NMDS and hierarchical clustering) help the user examine both raw and normalized data. For instance, these plots may reveal clusters, sample mislabelling and/or batch effects. Scatterplots of size factors and dispersion (i.e., estimates that are specific to DESeq2) are useful when assessing both the relevance and robustness of statistical models. PCA- and PCoA-plots associated with a PERMANOVA test may be used as preliminary results in the differential analysis as they may reveal global effects among groups of interest.
(ii) Significant features are gathered in a table including, the base mean (mean of the normalized counts), the fold change (i.e., the factor by which the average abundance changes from one group to the other), as well as the corresponding adjusted p-values. The user may view tables for any contrasts and can export it into several formats. Volcano plots and bar charts of p-values and log2 fold change are also available this section.
(iii) A global visualization section provides a choice of 9 interactive plots such as barplots, heatmaps and boxplots to represent differences in abundance across groups of interest. Diversity plots display the distribution of various diversity indices: alpha, beta, gamma, Shannon, Simpson and inverse Simpson. Scatterplots and network plots show association between feature abundance with other variables from the target file. To explore variations of abundance across the taxonomic classification, we included an interactive abundance tree and a Krona plot(45). Rarefaction curves are of great use to further consider the number of features in samples with respect to the sequencing depth.
(iv) In the comparison section, plots displaying comparisons among contrasts may be created. It includes several options such as, Venn diagram or upsetR graph(46) (displaying subsets of common features across contrast), heatmap, a logit plot(47) (showing the log2 fold-change values in each feature), a density plot and a multiple Venn diagram to summarize the number of features captured by each contrast. All these graphs can be exported into four format (eps, png, pdf and svg).

APPLICATION

Comparison of SHAMAN with other available tools for meta-taxonomic analyses. A brief qualitative assessment of the strengths and limits of SHAMAN was done in comparison with other similar web interfaces (Table 1). We first identified a list of important considerations that have practical implications for the user such as processing of raw sequencing data, statistical workflow, visualization, data storage and accessibility. For each similar web interface, we then evaluated whether it met those criteria. Besides that SHAMAN presents a number of advantages, we think that such nested solution is essential for a careful interpretation of the results. Any results in SHAMAN may be cross-checked with a quantification or an annotation performed at an earlier stage. Furthermore several applications presented in Table 1, impose the burden to import/export R objects which requires skills in R programming. It may also represent a source of issues for reproducibility, notably in terms of compatibility of the packages over time.

User case. To illustrate how SHAMAN works, we performed the analysis of two sequencing experiments: a mock sequencing and a published dataset, afribiota dataset(48). In both analyses, we submitted the raw FASTQ files and provided a target file containing sample information (needed for statistical analysis).

Zymo Mock dataset. The mock sequencing (EBI ENA code PRJEB33737) of the ZymoBIOMICS™ Microbial Community DNA was performed with an Illumina MiSeq. The Zymo mock community is composed with 8 phylogenetically distant bacterial strains, 3 of which are gram-negative and 5 of which are gram-positive. DNA of two yeast strains that are normally present in this community were not amplified. Genomic DNA from each bacterial strain was mixed in equimolar proportions (https://www.zymoresearch.com/zymobiomics-community-standard). We compared the impact of both the number of amplification cycles (25 and 30 cycles) and the amount of DNA loaded in the flow cell (0.5ng and 1ng), on the microbial abundance. Each sample was sequenced 3 times (experimental plan provided in supplementary materials). Sequencing report provided by the sequencing facilities indicated the presence of contaminants. To handle this issue, we filtered out the genera occurring in less than 12 samples and outliers with a reduced log abundance as compared to the other genera (Fig. S3). This process selected the 8 bacterial stains of Zymo mock (Fig. 1). We then defined a statistical model that included DNA amount and the number of amplification cycle as main effects and an interaction between these variables. The statistical comparison showed a significant impact of the number of amplification cycle compared to DNA amount. We found no differential features between 0.5 ng and 1 ng DNA for each possible number of cycle (25 and 30 cycles), while the comparison of number of amplification cycle for each given amount of DNA showed significant impact on the abundance of mock bacteria (Table. S1, S2). These results are in agreement with previous studies that presented the PCR-induced bias on equivalent mix(49, 50).

Afribiota dataset. The second dataset included samples of microbial communities in stunted children aged 2-5y living in sub-Saharan Africa (48). Three groups (nutritional

DRAFT

status) of individuals were considered: NN=non stunted, MCM=moderately stunted, MCS=severely stunted. Samples originated from the small intestine fluids (gastric and duodenal) and feces. The authors performed the bioinformatic treatment with QIIME framework and the statistical analysis with several R packages including Phyloseq for the normalization and DESeq2 for the differential analysis. 541 samples were available on EBI ENA (code PRJEB27868).
Using SHAMAN, raw reads were filtered against Human HG38 and PhiX174 genomes. A total of 2386 OTUs were calculated and 76% were annotated with SILVA database at genus level. The sparsity rate of the contingency table was high with 0.84. In consequence, we used the weighted non-null normalization which is particularly efficient when the matrix highly sparse (Fig. S2).

Two analyses were performed, a global analysis that included duodenal, gastric as well as feces samples and a more specific analysis including fecal samples only. Statistical models included the following variables, age, gender, country of origin and nutritional status. Overall our results obtained when using SHAMAN were highly consistent with those of Vonaesch et al. (48). We detected a significant change in the community composition between gastric and duodenal samples compared to feces samples at Genus level (Fig. 2a) (PERMANOVA, P=0.001). The most abundant genera were reported in Fig. S4. α-Diversity was not affected by stunting (Fig.2b). We looked for a distinct signature of stunting in the feces. We report in the volcano plot (Fig.2c) genera with differential abundance between stunt samples compared to non-stunt (complete list available in Table S3). Twelve microbial taxa, corresponding to members of the oropharyngeal core microbiota, were overrepresented in feces samples of stunted children as compared with non-stunted children; more particularly Porphyromonas, Neisseira and Lactobacillus (Fig.2d). These findings were in agreement with the conclusions of the Afribiota consortium while being obtained within a few minutes of interaction with the SHAMAN interface.

Conclusion

SHAMAN enables user to run most of the classical metagenomics methods and makes use of statistical analyses to provide support to each visualization. The possibility to deploy SHAMAN locally constitutes an important feature when the data cannot be submitted on servers for privacy issues or insufficient internet access. SHAMAN also simplifies the access to open computational facilities, making a careful use of the dedicated server, galaxy.pasteur.fr.
During its development, we felt a strong interest of the metagenomics community. We recorded 82 active users per month in 2019 (535 unique visitors in total) and 800 downloads of the docker application. We expect that SHAMAN will help researcher performing a quantitative analysis of metagenomics data.

Data availability

Sequence reads of Zymo Mock have been deposited in the European Nucleotide Archive, https://www.ebi.ac.uk/ena/ accesion no. PRJEB33737.

Acknowledgments

We thank Pascal Campagne for his comments, Hugo Varet for helpful discussions about DESeq2, Fabien Mareuil for the help to deploy SHAMAN computation on Galaxy and Youssef Ghorbal for the maintenance of the databank, as well as the IT System Department of Institut Pasteur, who manages installation and update of tools on TARS cluster.

DRAFT

1. Nan Qin, Fengling Yang, Ang Li, Edi Prifti, Yanfei Chen, Li Shao, Jing Guo, Emmanuelle Le Chatelier, Jian Yao, Lingjiao Wu, Jiawei Zhou, Shujun Ni, Lin Liu, Nicolas Pons, Jean Michel Batto, Sean P Kennedy, Pierre Leonard, Chunhui Yuan, Wenchao Ding, Yuanting Chen, Xinjun Hu, Beiwen Zheng, Guirong Qian, Wei Xu, S Dusko Ehrlich, Shusen Zheng, and Lanjuan Li. Alterations of the human gut microbiome in liver cirrhosis. *Nature*, 513:59–64, Sep 2014. ISSN 1476-4687. doi: 10.1038/nature13568. [PubMed:25079328] [doi:10.1038/nature13568].
2. Mihai Pop, Alan W Walker, Joseph Paulson, Brianna Lindsay, Martin Antonio, M Anowar Hossain, Joseph Oundo, Boubou Tamboura, Volker Mai, Irina Astrovskaya, Hector Corrada Bravo, Richard Rance, Mark Stares, Myron M Levine, Sandra Panchalingam, Karen Kotloff, Usman N Ikumapayi, Chinelo Ebruke, Mitchell Adeyemi, Dilruba Ahmed, Firoz Ahmed, Meer Taifur Alam, Ruhul Amin, Sabbir Siddiqui, John B Ochieng, Emmanuel Ouma, Jane Juma, Euince Mailu, Richard Omore, J Glenn Morris, Robert F Breiman, Debasish Saha, Julian Parkhill, James P Nataro, and O Colin Stine. Diarrhea in young children from low-income countries leads to large-scale alterations in intestinal microbiota composition. *Genome biology*, 15:R76, Jun 2014. ISSN 1474-760X. doi: 10.1186/gb-2014-15-6-r76. [PubMed Central: PMC4072981] [doi:10.1186/gb-2014-15-6-r76].
3. Georg Zeller, Julien Tap, Anita Y Voigt, Shinichi Sunagawa, Jens Roat Kultima, Paul I Costea, Aurélien Amiot, Jürgen Böhm, Francesco Brunetti, Nina Habermann, Rajna Hercog, Moritz Koch, Alain Luciani, Daniel R Mende, Martin A Schneider, Petra Schrotz-King, Christophe Tournigand, Jeanne Tran Van Nhieu, Takuji Yamada, Jürgen Zimmermann, Vladimir Benes, Matthias Kloor, Cornelia M Ulrich, Magnus von Knebel Doeberitz, Iradj Sobhani, and Peer Bork. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular systems biology*, 10:766, 2014. ISSN 1744-4292. doi: 10.15252/msb.20145645. [PubMed Central: PMC4299606][doi:10.15252/msb.20145645].
4. Juan J Quereda, Olivier Dussurget, Marie-Anne Nahori, Amine Ghozlane, Stevenn Volant, Marie-Agnès Dillies, Béatrice Regnault, Sean Kennedy, Stanislas Mondot, Barbara Villoing, Pascale Cossart, and Javier Pizarro-Cerda. Bacteriocin from epidemic listeria strains alters the host intestinal microbiota to favor infection. *Proceedings of the National Academy of Sciences of the United States of America*, 113:5706–5711, May 2016. ISSN 1091-6490. doi: 10.1073/pnas.1523899113. [PubMed Central: PMC4878514] [doi:10.1073/pnas.1523899113].
5. Patrick Veiga, Carey Ann Gallini, Chloé Beal, Monia Michaud, Mary L Delaney, Andrea DuBois, Artem Khlebnikov, Johan ET van Hylckama Vlieg, Shivesh Punit, Jonathan N Glickman, et al. Bifidobacterium animalis subsp. lactis fermented milk product reduces inflammation by altering a niche for colitogenic microbes. *Proceedings of the National Academy of Sciences*, 107(42):18132–18137, October 2010. ISSN 1091-6490. doi: 10.1073/pnas.1011737107. [PubMed Central: PMC2964251][doi:10.1073/pnas.1011737107].
6. Sarah L Westcott and Patrick D Schloss. De novo clustering methods outperform reference-based methods for assigning 16s rrna gene sequences to operational taxonomic units. *PeerJ*, 3:e1487, 2015. ISSN 2167-8359. doi: 10.7717/peerj.1487. [PubMed Central: PMC4675110][doi:10.7717/peerj.1487].
7. Patrick D Schloss, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan A Lesniewski, Brian B Oakley, Donovan H Parks, Courtney J Robinson, Jason W Sahl, Blaz Stres, Gerhard G Thallinger, David J Van Horn, and Carolyn F Weber. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, 75:7537–7541, Dec 2009. ISSN 1098-5336. doi: 10.1128/AEM.01541-09. [PubMed Central: PMC2786419][doi:10.1128/AEM.01541-09].
8. Robert C Edgar. Uparse: highly accurate otu sequences from microbial amplicon reads. *Nature methods*, 10:996–998, Oct 2013. ISSN 1548-7105. doi: 10.1038/nmeth.2604. [PubMed:23955772] [doi:10.1038/nmeth.2604].
9. Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. Dada2: High-resolution sample inference from illumina amplicon data. *Nature methods*, 13:581–583, July 2016. ISSN 1548-7105. doi: 10.1038/nmeth.3869. [PubMed Central: PMC4927377][doi:10.1038/nmeth.3869].
10. Torbjørn Rognes, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. Vsearch: a versatile open source tool for metagenomics. *PeerJ*, 4:e2584, 2016. ISSN 2167-8359. doi: 10.7717/peerj.2584. [PubMed Central: PMC5075697][doi:10.7717/peerj.2584].
11. J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Pena, Julia K Goodrich, Jeffrey I Gordon, et al. Qiime allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5):335–336, May 2010. ISSN 1548-7105. doi: 10.1038/nmeth.f.303. [PubMed Central: PMC3156573][doi:10.1038/nmeth.f.303].
12. Daniel McDonald, Jose C Clemente, Justin Kuczynski, Jai Ram Rideout, Jesse Stombaugh, Doug Wendel, Andreas Wilke, Susan Huse, John Hufnagle, Folker Meyer, et al. The biological observation matrix (biom) format or: how i learned to stop worrying and love the ome-ome. *GigaScience*, 1(1):7, July 2012. ISSN 2047-217X. doi: 10.1186/2047-217X-1-7. [PubMed Central: PMC3626512][doi:10.1186/2047-217X-1-7].
13. Joseph N Paulson, Mihai Pop, and Hector Corrada Bravo. Metastats: an improved statistical method for analysis of metagenomic data. *Genome Biology*, 12(1):P17, 2011.
14. Joseph N Paulson, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. Differential abundance analysis for microbial marker-gene surveys. *Nature methods*, 10:1200–1202, Dec 2013. ISSN 1548-7105. doi: 10.1038/nmeth.2658. [PubMed Central: PMC4010126][doi:10.1038/nmeth.2658].
15. Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15:550, 2014. doi: 10.1186/s13059-014-0550-8.
16. Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, January 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp616. [PubMed Central: PMC2796818][doi:10.1093/bioinformatics/btp616].
17. Paul J McMurdie and Susan Holmes. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS computational biology*, 10(4):e1003531, April 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003531. [PubMed Central: PMC3974642][doi:10.1371/journal.pcbi.1003531].
18. Olle Nerman Viktor Jonsson, Tobias Österlund and Erik Kristiansson. Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. *BMC Genomics*, 17(1):1–14, 2016. doi: 10.1186/s12864-016-2386-y.
19. Frédéric Escudié, Lucas Auer, Maria Bernard, Mahendra Mariadassou, Laurent Cauquil, Katia Vidal, Sarah Maman, Guillermina Hernandez-Raquet, Sylvie Combes, and Géraldine Pascal. Frogs: find, rapidly, otus with galaxy solution. *Bioinformatics*, 34(8):1287–1294, April 2017. ISSN 1367-4811. doi: 10.1093/bioinformatics/btx791. [PubMed:29228191] [doi:10.1093/bioinformatics/btx791].
20. Antonio Gonzalez, Jose A Navas-Molina, Tomasz Kosciolek, Daniel McDonald, Yoshiki Vázquez-Baeza, Gail Ackermann, Jeff DeReus, Stefan Janssen, Austin D Swafford, Stephanie B Orchanian, Jon G Sanders, Joshua Shorenstein, Hannes Holste, Semar Petrus, Adam Robbins-Pianka, Colin J Brislawn, Mingxun Wang, Jai Ram Rideout, Evan Bolyen, Matthew Dillon, J Gregory Caporaso, Pieter C Dorrestein, and Rob Knight. Qiita: rapid, web-enabled microbiome meta-analysis. *Nature methods*, 15:796–798, October 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0141-9. [PubMed Central: PMC6235622][doi:10.1038/s41592-018-0141-9].
21. Paul J McMurdie and Susan Holmes. Shiny-phyloseq: Web application for interactive microbiome analysis with provenance tracking. *Bioinformatics (Oxford, England)*, 31:282–283, Jan 2015. ISSN 1367-4811. doi: 10.1093/bioinformatics/btu616. [PubMed Central: PMC4287943][doi:10.1093/bioinformatics/btu616].
22. Justin Wagner, Florin Chelaru, Jayaram Kancherla, Joseph N Paulson, Alexander Zhang, Victor Felix, Anup Mahurkar, Niklas Elmqvist, and Héctor Corrada Bravo. Metaviz: interactive statistical and visual analysis of metagenomic data. *Nucleic acids research*, 46(6):2777–2787, April 2018. ISSN 1362-4962. doi: 10.1093/nar/gky136. [PubMed Central: PMC5887897][doi:10.1093/nar/gky136].
23. Susan M Huse, David B Mark Welch, Andy Voorhis, Anna Shipunova, Hilary G Morrison, A Murat Eren, and Mitchell L Sogin. Vamps: a website for visualization and analysis of microbial population structures. *BMC bioinformatics*, 15:41, February 2014. ISSN 1471-2105. doi: 10.1186/1471-2105-15-41. [PubMed Central: PMC3922339][doi:10.1186/1471-2105-15-41].
24. Enis Afgan, Dannon Baker, Bérénice Batut, Marius van den Beek, Dave Bouvier, Martin Cech, John Chilton, Dave Clements, Nate Coraor, Björn A Grüning, Aysam Guerler, Jennifer Hillman-Jackson, Saskia Hiltemann, Vahid Jalili, Helena Rasche, Nicola Soranzo, Jeremy Goecks, James Taylor, Anton Nekrutenko, and Daniel Blankenberg. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic acids research*, 46:W537–W544, July 2018. ISSN 1362-4962. doi: 10.1093/nar/gky379. [PubMed Central: PMC6030816][doi:10.1093/nar/gky379].
25. Clare Sloggett, Nuwan Goonasekera, and Enis Afgan. Bioblend: automating pipeline analyses within galaxy and cloudman. *Bioinformatics*, 29(13):1685–1686, July 2013. ISSN 1367-4811. doi: 10.1093/bioinformatics/btt199. [PubMed Central: PMC4288140][doi:10.1093/bioinformatics/btt199].
26. Laura B Dickson, Davy Jiolle, Guillaume Minard, Isabelle Moltini-Conclois, Stevenn Volant, Amine Ghozlane, Christiane Bouchier, Diego Ayala, Christophe Paupy, Claire Valiente Moro, et al. Carryover effects of larval exposure to different environmental bacteria drive adult trait variation in a mosquito vector. *Science advances*, 3(8):e1700585, August 2017. ISSN 2375-2548. doi: 10.1126/sciadv.1700585. [PubMed Central: PMC5559213][doi:10.1126/sciadv.1700585].
27. Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biology*, 10(3):R25, 2009. ISSN 1474-760X. doi: 10.1186/gb-2009-10-3-r25. [PubMed Central: PMC2690996][doi:10.1186/gb-2009-10-3-r25].
28. Alexis Criscuolo and Sylvain Brisse. Alientrimmer: a tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. *Genomics*, 102(5-6):500–506, 2013. ISSN 1089-8646. doi: 10.1016/j.ygeno.2013.07.011. [PubMed:23912058] [doi:10.1016/j.ygeno.2013.07.011].
29. Jiajie Zhang, Kassian Kobert, Tomáš Flouri, and Alexandros Stamatakis. Pear: a fast and accurate illumina paired-end read merger. *Bioinformatics*, 30(5):614–620, March 2013. ISSN 1367-4811. doi: 10.1093/bioinformatics/btt593. [PubMed Central: PMC3933873][doi:10.1093/bioinformatics/btt593].
30. Elmar Pruesse, Christian Quast, Katrin Knittel, Bernhard M Fuchs, Wolfgang Ludwig, Jörg Peplies, and Frank Oliver Glöckner. Silva: a comprehensive online resource for quality checked and aligned ribosomal rna sequence data compatible with arb. *Nucleic acids research*, 35(21):7188–7196, 2007. ISSN 1362-4962. doi: 10.1093/nar/gkm864. [PubMed Central: PMC2175337][doi:10.1093/nar/gkm864].
31. T Z DeSantis, P Hugenholtz, N Larsen, M Rojas, E L Brodie, K Keller, T Huber, D Dalevi, P Hu, and G L Andersen. Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. *Applied and environmental microbiology*, 72:5069–5072, July 2006. ISSN 0099-2240. doi: 10.1128/AEM.03006-05. [PubMed Central: PMC1489311][doi:10.1128/AEM.03006-05].
32. Jie Tang, Iliyan D Iliev, Jordan Brown, David M Underhill, and Vincent A Funari. Mycobiome: approaches to analysis of intestinal fungi. *Journal of immunological methods*, 421:112–121, June 2015. ISSN 1872-7905. doi: 10.1016/j.jim.2015.04.004. [PubMed Central: PMC4451377][doi:10.1016/j.jim.2015.04.004].
33. Kessy Abarenkov, R Henrik Nilsson, Karl-Henrik Larsson, Ian J Alexander, Ursula Eberhardt, Susanne Erland, Klaus Høiland, Rasmus Kjøller, Ellen Larsson, Taina Pennanen, Robin Sen, Andy F S Taylor, Leho Tedersoo, Björn M Ursing, Trude Vrålstad, Kare Liimatainen, Ursula Peintner, and Urmas Kõljalg. The unite database for molecular identification of fungi–recent updates and future perspectives. *The New phytologist*, 186:281–285, April 2010. ISSN 1469-8137. doi: 10.1111/j.1469-8137.2009.03160.x. [PubMed:20409185] [doi:10.1111/j.1469-8137.2009.03160.x].
34. Keisha Findley, Julia Oh, Joy Yang, Sean Conlan, Clayton Deming, Jennifer A Meyer, Deborah Schoenfeld, Effie Nomicos, Morgan Park, NIH Intramural Sequencing Center Comparative Sequencing Program, Heidi H Kong, and Julia A Segre. Topographic diversity of fungal and bacterial communities in human skin. *Nature*, 498:367–370, June 2013. ISSN 1476-4687. doi: 10.1038/nature12171. [PubMed Central: PMC3711185][doi:10.1038/nature12171].

DRAFT

35. Pablo Yarza, Pelin Yilmaz, Elmar Pruesse, Frank Oliver Glöckner, Wolfgang Ludwig, Karl-Heinz Schleifer, William B Whitman, Jean Euzéby, Rudolf Amann, and Ramon Rosselló-Móra. Uniting the classification of cultured and uncultured bacteria and archaea using 16s rrna gene sequences. *Nature reviews. Microbiology*, 12:635–645, September 2014. ISSN 1740-1534. doi: 10.1038/nrmicro3330. [PubMed:25118885] [doi:10.1038/nrmicro3330].
36. Qiong Wang, George M Garrity, James M Tiedje, and James R Cole. Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, 73:5261–5267, August 2007. ISSN 0099-2240. doi: 10.1128/AEM.00062-07. [PubMed Central:PMC1950982] [doi:10.1128/AEM.00062-07].
37. Kazutaka Katoh and Daron M Standley. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4): 772–780, April 2013. ISSN 1537-1719. doi: 10.1093/molbev/mst010. [PubMed Central: PMC3603318][doi:10.1093/molbev/mst010].
38. Alexis Criscuolo and Simonetta Gribaldo. Bmge (block mapping and gathering with entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC evolutionary biology*, 10(1):210, July 2010. ISSN 1471-2148. doi: 10.1186/1471-2148-10-210. [PubMed Central: PMC3017758][doi:10.1186/1471-2148-10-210].
39. Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular biology and evolution*, 26(7):1641–1650, July 2009. ISSN 1537-1719. doi: 10.1093/molbev/msp077. [PubMed Central: PMC2693737][doi:10.1093/molbev/msp077].
40. Catherine Lozupone and Rob Knight. Unifrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology*, 71:8228–8235, December 2005. ISSN 0099-2240. doi: 10.1128/AEM.71.12.8228-8235.2005. [PubMed Central: PMC1317376][doi:10.1128/AEM.71.12.8228-8235.2005].
41. David Sims, Ian Sudbery, Nicholas E Ilott, Andreas Heger, and Chris P Ponting. Sequencing depth and coverage: key considerations in genomic analyses. *Nature reviews. Genetics*, 15:121–132, February 2014. ISSN 1471-0064. doi: 10.1038/nrg3642. [PubMed:24434847] [doi:10.1038/nrg3642].
42. Paul J McMurdie and Susan Holmes. phyloseq: an r package for reproducible interactive analysis and graphics of microbiome census data. *PloS one*, 8(4):e61217, 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0061217. [PubMed Central: PMC3632530][doi:10.1371/journal.pone.0061217].
43. Ciaran Evans, Johanna Hardin, and Daniel M Stoebel. Selecting between-sample RNA-seq normalization methods from the perspective of their assumptions. *Briefings in Bioinformatics*, 19(5):776–792, feb 2017. ISSN 1477-4054. doi: 10.1093/bib/bbx008. [PubMed Central: PMC6171491][doi:10.1093/bib/bbx008].
44. Marie-Agnès Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, Guillemette Marot, David Castel, Jordi Estelle, et al. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in bioinformatics*, 14(6):671–683, November 2013. ISSN 1477-4054. doi: 10.1093/bib/bbs046. [PubMed:22988256] [doi:10.1093/bib/bbs046].
45. Brian D Ondov, Nicholas H Bergman, and Adam M Phillippy. Interactive metagenomic visualization in a web browser. *BMC Bioinformatics*, 12(1), September 2011. doi: 10.1186/1471-2105-12-385. [PubMed Central: PMC3190407][doi:10.1186/1471-2105-12-385].
46. Jake R Conway, Alexander Lex, and Nils Gehlenborg. Upsetr: an r package for the visualization of intersecting sets and their properties. *Bioinformatics*, 33(18):2938–2940, September 2017. ISSN 1367-4811. doi: 10.1093/bioinformatics/btx364. [PubMed Central: PMC5870712][doi:10.1093/bioinformatics/btx364].
47. Véronique Hourdel, Stevenn Volant, Darragh P O'Brien, Alexandre Chenal, Julia Chamot-Rooke, Marie-Agnès Dillies, and Sébastien Brier. Memhdx: an interactive tool to expedite the statistical validation and visualization of large hdx-ms datasets. *Bioinformatics*, 32(22):3413–3419, November 2016. ISSN 1367-4811. doi: 10.1093/bioinformatics/btw420. [PubMed Central: PMC5181559][doi:10.1093/bioinformatics/btw420].
48. Pascale Vonaesch, Evan Morien, Lova Andrianonimiadana, Hugues Sanke, Jean-Robert Mbecko, Kelsey E Huus, Tanteliniaina Naharimanananirina, Bolmbaye Privat Gondje, Synthia Nazita Nigatoloum, Sonia Sandrine Vondo, et al. Stunted childhood growth is associated with decompartmentalization of the gastrointestinal tract and overgrowth of oropharyngeal taxa. *Proceedings of the National Academy of Sciences*, 115(36):E8489–E8498, September 2018. ISSN 1091-6490. doi: 10.1073/pnas.1806573115. [PubMed Central: PMC6130352][doi:10.1073/pnas.1806573115].
49. Silvia G Acinas, Ramahi Sarma-Rupavtarm, Vanja Klepac-Ceraj, and Martin F Polz. Pcr-induced sequence artifacts and bias: insights from comparison of two 16s rrna clone libraries constructed from the same sample. *Appl. Environ. Microbiol.*, 71(12):8966–8969, 2005. [PubMed Central: PMC1317340][doi:10.1128/AEM.71.12.8966-8969.2005].
50. Rita Sipos, Anna J Székely, Márton Palatinszky, Sára Révész, Károly Márialigeti, and Marcell Nikolausz. Effect of primer mismatch, annealing temperature and pcr cycle number on 16s rrna gene-targetting bacterial community analysis. *FEMS microbiology ecology*, 60:341–350, May 2007. ISSN 0168-6496. doi: 10.1111/j.1574-6941.2007.00283.x. [PubMed:17343679] [doi:10.1111/j.1574-6941.2007.00283.x].

DRAFT

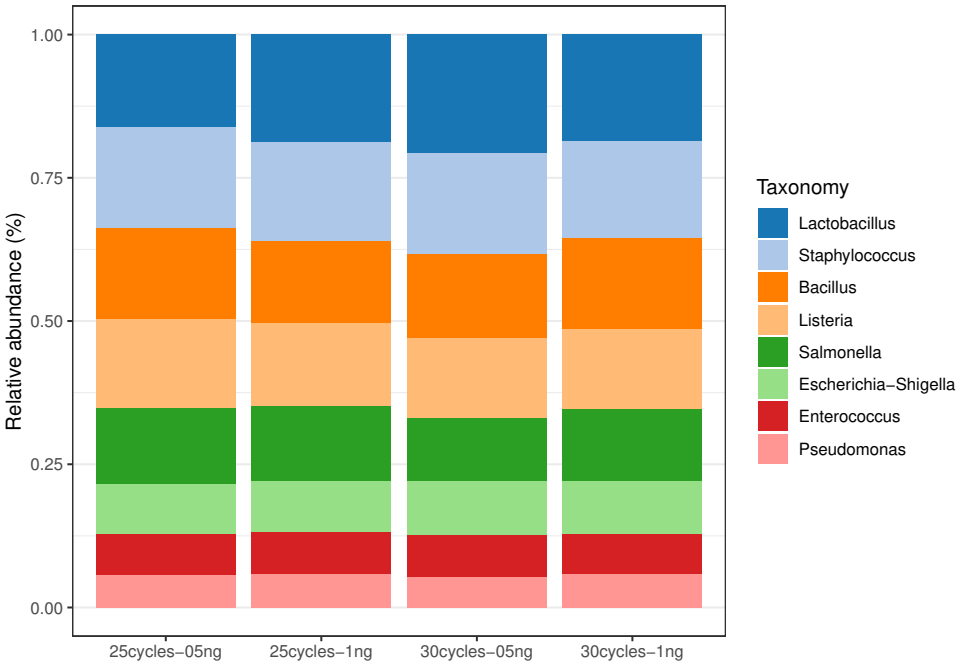

Fig. 1. Barplot of taxa abundance of ZYMO MOCK samples. We summed the abundance of the OTU annotated at genera level with SILVA database and plotted the average abundance per condition.

DRAFT

Table 1. Comparison of SHAMAN with other web interface for metataxonomic analysis.

Category	SHAMAN	FROGS	Qiita	Shiny-phyloseq	Metaviz	Vamps
OTU processing	Yes	Yes	Yes	No	No	No
Normalization	Yes	Yes	No	No	No	No
Modelisation	Yes	Manova	No	D	M	No
Diversity analysis	Yes	Yes	Yes	Alpha	Alpha	Alpha
Phylogenetic analysis	Yes	Yes	Yes	Yes	No	Tree
Feature abundance plots	Yes	Yes	Yes	Yes	Yes	Yes
Ordination plots	Yes	Yes	Yes	Yes	Yes	Yes
Network plots	Yes	No	No	Yes	No	Yes
Geographic distribution plots	No	No	No	No	No	Yes
Statistics plots	Yes	No	NR	Yes	NR	NR
Interactive visualization	31	2;P	3	8	9	17
Raw data storage	No	No	Yes	No	No	Yes
Result storage	Yes	No	Yes	No	No	Yes
Online web Interface	Yes	No	Yes	No	Yes	Yes
R packaging	No	NR	NR	Yes	Yes	NR
Docker	Yes	No	No	No	Yes	No
Conda	Yes	Yes	Yes	No	Yes	No

D: Export from DESeq2, M: Export from Metagenomeseq, NR: Non relevant feature, P: Import/Export to Phyloseq, Number of unique interactive visualization are reported for each application in section 'Interactive visualization'

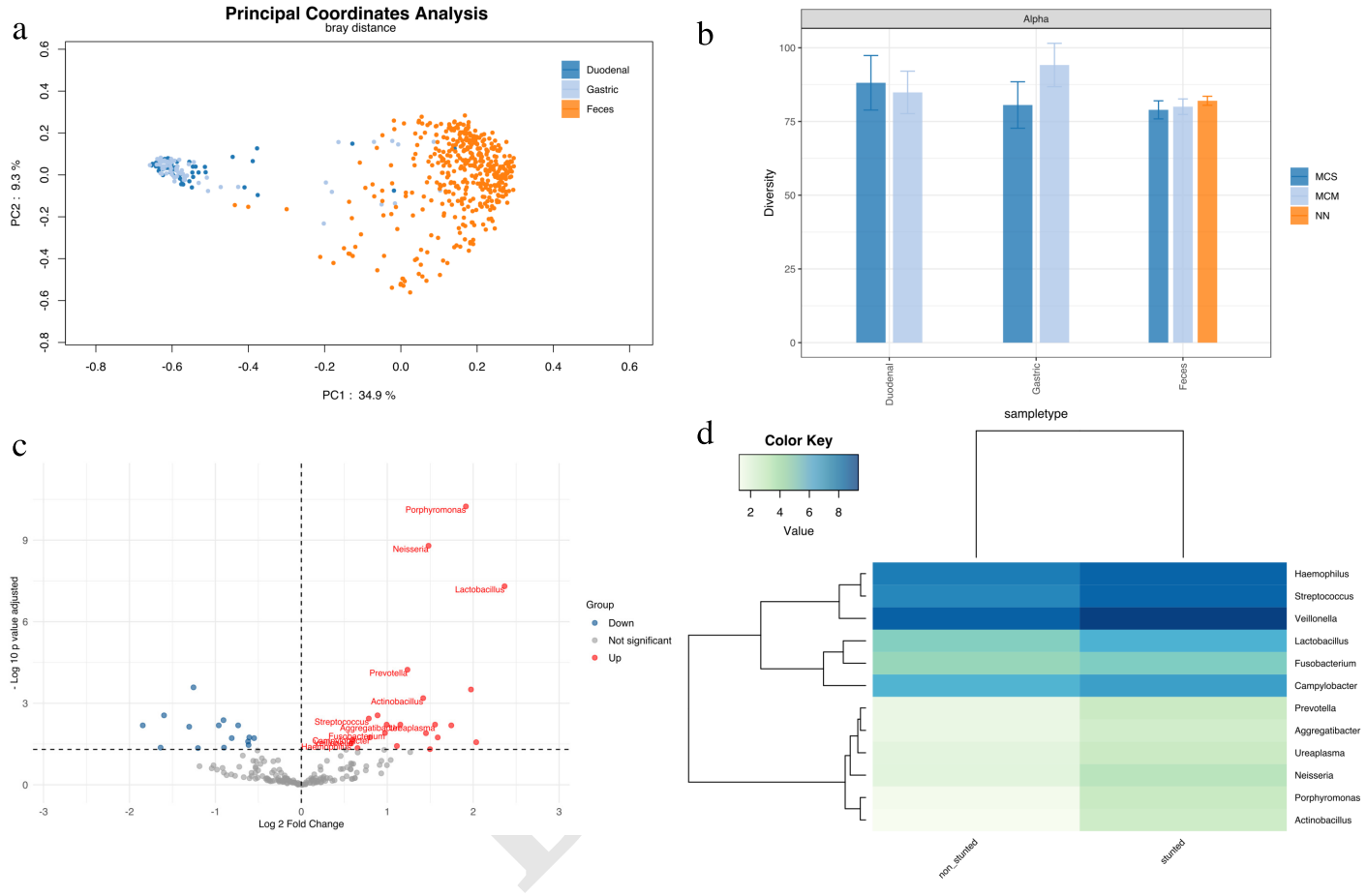

Fig. 2. Afribiota study of small intestine fluids and feces from stunt children compared to non stunt. (a) PCoA plot the Bray-Curtis dissimilarity index of the samples. Duodenal samples are colored in blue, light blue for Gastric and orange for Feces. PERMANOVA test based on the sample type yielded a P value of 0.001. (b) Alpha diversity analysis of non-stunt (NN), moderately stunted (MCM) and severely stunted (MCS). Overlapping confidence interval indicates that the diversity are not different between NN, MCM and MCS in duodenal, gastric and feces samples. (c) Volcano plot of differentially abundant genera in the feces of stunt children compared to non-stunt. We plot the log2 fold change against the -log 10 adjusted p-value. Microbial taxa in red correspond to an increase of abundance and in blue to a decrease abundance. Labeled dots correspond to taxa from orpharyngeal core microbiota. (d) Log 2 abundance of differential abundant taxa from orpharyngeal core microbiota in stunt and non-stunt children feces.