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

Stevenn Volant
Institut Pasteur
Pierre Lechat
Institut Pasteur
Perrine Woringer
Institut Pasteur
Laurence Motreff
Institut Pasteur
Christophe Malabat
Institut Pasteur
Sean Kennedy
Institut Pasteur
Amine Ghozlane (amine.ghozlane@pasteur.fr)
Institut Pasteur https://orcid.org/0000-0001-7670-6235

Software article

Keywords: Metagenomics, Differential analysis, Visualization, Web application

Posted Date: February 11th, 2020

DOI: https://doi.org/10.21203/rs.2.23213/v1

Version of Record: A version of this preprint was published at BMC Bioinformatics on August 10th, 2020.
See the published version at https://doi.org/10.1186/s12859-020-03666-4.

SOFTWARE ARTICLE

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*]

*Correspondence: amine.ghozlane@pasteur.fr
[1]Hub de Bioinformatique et Biostatistique – Département Biologie Computationnelle, Institut Pasteur, USR 3756 CNRS, 28 rue du Docteur Roux, 75015 Paris, France
Full list of author information is available at the end of the article

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 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.

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 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 rRNA metagenomic data), we illustrate the strengths of SHAMAN in quickly performing a complete metataxonomic analysis.

Conclusions: We aim with SHAMAN to provide the scientific community with a platform that simplifies reproducible quantitative analysis of metagenomic data.

Keywords: Metagenomics; Differential analysis; Visualization; Web application

Background

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 genes for procaryotes/archaea and the ITS, the 23S or the 28S rRNA gene 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 automatic 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.

Implementation

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 rRNA genes) and LSU (for 23S and 28S rRNA genes), Greengenes[31] (for 16S, 18S rRNA genes) and Underhill rev. 1.6.1[32], Unite rev. 8.0[33] and Findley[34] for ITS rRNA 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 re-load 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 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.

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

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

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.

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 dropdown 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).

Results

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 cases

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^{TM} 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 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.

Conclusions

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. The datasets generated, analysed during the current study and the simulation script of the supplementary figure 2 are available in the microbiome repository: 10.6084/m9.figshare.11815860.

Software availability

Project name: SHAMAN
Project home page: http://shaman.pasteur.fr, https://github.com/aghozlane/shaman
Operating system: Platform independent
Programming language: R
Other requirements: Python 3
License: GNU GPL V3
Any restrictions to use by non-academics: No

Competing interests

The authors declare that they have no competing interests.

Author's contributions

SV,PL, PW, CM and AG developped SHAMAN; LM and SK performed the mock sequencing; SV, CM and AG wrote the publication.

Acknowledgements

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.

Author details

[1]Hub de Bioinformatique et Biostatistique – Département Biologie Computationnelle, Institut Pasteur, USR 3756 CNRS, 28 rue du Docteur Roux, 75015 Paris, France. [2]Biomics – Département Génomes et Génétique, Institut Pasteur, 28 rue du Docteur Roux, 75015 Paris, France.

References

1. Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., Guo, J., Le Chatelier, E., Yao, J., Wu, L., Zhou, J., Ni, S., Liu, L., Pons, N., Batto, J.M., Kennedy, S.P., Leonard, P., Yuan, C., Ding, W., Chen, Y., Hu, X., Zheng, B., Qian, G., Xu, W., Ehrlich, S.D., Zheng, S., Li, L.: Alterations of the human gut microbiome in liver cirrhosis. Nature **513**, 59–64 (2014). doi:10.1038/nature13568
2. Pop, M., Walker, A.W., Paulson, J., Lindsay, B., Antonio, M., Hossain, M.A., Oundo, J., Tamboura, B., Mai, V., Astrovskaya, I., Corrada Bravo, H., Rance, R., Stares, M., Levine, M.M., Panchalingam, S., Kotloff, K., Ikumapayi, U.N., Ebruke, C., Adeyemi, M., Ahmed, D., Ahmed, F., Alam, M.T., Amin, R., Siddiqui, S., Ochieng, J.B., Ouma, E., Juma, J., Mailu, E., Omore, R., Morris, J.G., Breiman, R.F., Saha, D., Parkhill, J., Nataro, J.P., Stine, O.C.: Diarrhea in young children from low-income countries leads to large-scale alterations in intestinal microbiota composition. Genome biology **15**, 76 (2014). doi:10.1186/gb-2014-15-6-r76

3. Zeller, G., Tap, J., Voigt, A.Y., Sunagawa, S., Kultima, J.R., Costea, P.I., Amiot, A., Böhm, J., Brunetti, F., Habermann, N., Hercog, R., Koch, M., Luciani, A., Mende, D.R., Schneider, M.A., Schrotz-King, P., Tournigand, C., Tran Van Nhieu, J., Yamada, T., Zimmermann, J., Benes, V., Kloor, M., Ulrich, C.M., von Knebel Doeberitz, M., Sobhani, I., Bork, P.: Potential of fecal microbiota for early-stage detection of colorectal cancer. Molecular systems biology **10**, 766 (2014). doi:10.15252/msb.20145645
4. Quereda, J.J., Dussurget, O., Nahori, M.-A., Ghozlane, A., Volant, S., Dillies, M.-A., Regnault, B., Kennedy, S., Mondot, S., Villoing, B., Cossart, P., Pizarro-Cerda, J.: 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 (2016). doi:10.1073/pnas.1523899113
5. Veiga, P., Gallini, C.A., Beal, C., Michaud, M., Delaney, M.L., DuBois, A., Khlebnikov, A., van Hylckama Vlieg, J.E., Punit, S., Glickman, J.N., *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 (2010). doi:10.1073/pnas.1011737107
6. Westcott, S.L., Schloss, P.D.: De novo clustering methods outperform reference-based methods for assigning 16s rrna gene sequences to operational taxonomic units. PeerJ **3**, 1487 (2015). doi:10.7717/peerj.1487
7. Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., Van Horn, D.J., Weber, C.F.: Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and environmental microbiology **75**, 7537–7541 (2009). doi:10.1128/AEM.01541-09
8. Edgar, R.C.: Uparse: highly accurate otu sequences from microbial amplicon reads. Nature methods **10**, 996–998 (2013). doi:10.1038/nmeth.2604
9. Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P.: Dada2: High-resolution sample inference from illumina amplicon data. Nature methods **13**, 581–583 (2016). doi:10.1038/nmeth.3869
10. Rognes, T., Flouri, T., Nichols, B., Quince, C., Mahé, F.: Vsearch: a versatile open source tool for metagenomics. PeerJ **4**, 2584 (2016). doi:10.7717/peerj.2584
11. Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I., *et al.*: Qiime allows analysis of high-throughput community sequencing data. Nature methods **7**(5), 335–336 (2010). doi:10.1038/nmeth.f.303
12. McDonald, D., Clemente, J.C., Kuczynski, J., Rideout, J.R., Stombaugh, J., Wendel, D., Wilke, A., Huse, S., Hufnagle, J., Meyer, F., *et al.*: The biological observation matrix (biom) format or: how i learned to stop worrying and love the ome-ome. GigaScience **1**(1), 7 (2012). doi:10.1186/2047-217X-1-7
13. Paulson, J.N., Pop, M., Bravo, H.C.: Metastats: an improved statistical method for analysis of metagenomic data. Genome Biology **12**(1), 17 (2011)
14. Paulson, J.N., Stine, O.C., Bravo, H.C., Pop, M.: Differential abundance analysis for microbial marker-gene surveys. Nature methods **10**, 1200–1202 (2013). doi:10.1038/nmeth.2658
15. Love, M.I., Huber, W., Anders, S.: 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. Robinson, M.D., McCarthy, D.J., Smyth, G.K.: edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**(1), 139–140 (2010). doi:10.1093/bioinformatics/btp616
17. McMurdie, P.J., Holmes, S.: Waste not, want not: why rarefying microbiome data is inadmissible. PLoS computational biology **10**(4), 1003531 (2014). doi:10.1371/journal.pcbi.1003531
18. Viktor Jonsson, O.N. Tobias Österlund, Kristiansson, E.: 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. Escudié, F., Auer, L., Bernard, M., Mariadassou, M., Cauquil, L., Vidal, K., Maman, S., Hernandez-Raquet, G., Combes, S., Pascal, G.: Frogs: find, rapidly, otus with galaxy solution. Bioinformatics **34**(8), 1287–1294 (2017). doi:10.1093/bioinformatics/btx791
20. Gonzalez, A., Navas-Molina, J.A., Kosciolek, T., McDonald, D., Vázquez-Baeza, Y., Ackermann, G., DeReus, J., Janssen, S., Swafford, A.D., Orchanian, S.B., Sanders, J.G., Shorenstein, J., Holste, H., Petrus, S., Robbins-Pianka, A., Brislawn, C.J., Wang, M., Rideout, J.R., Bolyen, E., Dillon, M., Caporaso, J.G., Dorrestein, P.C., Knight, R.: Qiita: rapid, web-enabled microbiome meta-analysis. Nature methods **15**, 796–798 (2018). doi:10.1038/s41592-018-0141-9
21. McMurdie, P.J., Holmes, S.: Shiny-phyloseq: Web application for interactive microbiome analysis with provenance tracking. Bioinformatics (Oxford, England) **31**, 282–283 (2015). doi:10.1093/bioinformatics/btu616
22. Wagner, J., Chelaru, F., Kancherla, J., Paulson, J.N., Zhang, A., Felix, V., Mahurkar, A., Elmqvist, N., Corrada Bravo, H.: Metaviz: interactive statistical and visual analysis of metagenomic data. Nucleic acids research **46**(6), 2777–2787 (2018). doi:10.1093/nar/gky136
23. Huse, S.M., Mark Welch, D.B., Voorhis, A., Shipunova, A., Morrison, H.G., Eren, A.M., Sogin, M.L.: Vamps: a website for visualization and analysis of microbial population structures. BMC bioinformatics **15**, 41 (2014). doi:10.1186/1471-2105-15-41
24. Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B.A., Guerler, A., Hillman-Jackson, J., Hiltemann, S., Jalili, V., Rasche, H., Soranzo, N., Goecks, J., Taylor, J., Nekrutenko, A., Blankenberg, D.: The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic acids research **46**, 537–544 (2018). doi:10.1093/nar/gky379
25. Sloggett, C., Goonasekera, N., Afgan, E.: Bioblend: automating pipeline analyses within galaxy and cloudman. Bioinformatics **29**(13), 1685–1686 (2013). doi:10.1093/bioinformatics/btt199
26. Dickson, L.B., Jiolle, D., Minard, G., Moltini-Conclois, I., Volant, S., Ghozlane, A., Bouchier, C., Ayala, D., Paupy, C., Moro, C.V., *et al.*: Carryover effects of larval exposure to different environmental bacteria drive adult trait variation in a mosquito vector. Science advances **3**(8), 1700585 (2017). doi:10.1126/sciadv.1700585
27. Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L.: Ultrafast and memory-efficient alignment of short dna

sequences to the human genome. Genome biology **10**(3), 25 (2009). doi:10.1186/gb-2009-10-3-r25
28. Criscuolo, A., Brisse, S.: 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). doi:10.1016/j.ygeno.2013.07.011
29. Zhang, J., Kobert, K., Flouri, T., Stamatakis, A.: Pear: a fast and accurate illumina paired-end read merger. Bioinformatics **30**(5), 614–620 (2013). doi:10.1093/bioinformatics/btt593
30. Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J., Glöckner, F.O.: 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). doi:10.1093/nar/gkm864
31. DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., Andersen, G.L.: Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. Applied and environmental microbiology **72**, 5069–5072 (2006). doi:10.1128/AEM.03006-05
32. Tang, J., Iliev, I.D., Brown, J., Underhill, D.M., Funari, V.A.: Mycobiome: approaches to analysis of intestinal fungi. Journal of immunological methods **421**, 112–121 (2015). doi:10.1016/j.jim.2015.04.004
33. Abarenkov, K., Henrik Nilsson, R., Larsson, K.-H., Alexander, I.J., Eberhardt, U., Erland, S., Høiland, K., Kjøller, R., Larsson, E., Pennanen, T., Sen, R., Taylor, A.F.S., Tedersoo, L., Ursing, B.M., Vrålstad, T., Liimatainen, K., Peintner, U., Kõljalg, U.: The unite database for molecular identification of fungi–recent updates and future perspectives. The New phytologist **186**, 281–285 (2010). doi:10.1111/j.1469-8137.2009.03160.x
34. Findley, K., Oh, J., Yang, J., Conlan, S., Deming, C., Meyer, J.A., Schoenfeld, D., Nomicos, E., Park, M., Program, N.I.S.C.C.S., Kong, H.H., Segre, J.A.: Topographic diversity of fungal and bacterial communities in human skin. Nature **498**, 367–370 (2013). doi:10.1038/nature12171
35. Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F.O., Ludwig, W., Schleifer, K.-H., Whitman, W.B., Euzéby, J., Amann, R., Rosselló-Móra, R.: Uniting the classification of cultured and uncultured bacteria and archaea using 16s rrna gene sequences. Nature reviews. Microbiology **12**, 635–645 (2014). doi:10.1038/nrmicro3330
36. Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R.: Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. Applied and environmental microbiology **73**, 5261–5267 (2007). doi:10.1128/AEM.00062-07
37. Katoh, K., Standley, D.M.: Mafft multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution **30**(4), 772–780 (2013). doi:10.1093/molbev/mst010
38. Criscuolo, A., Gribaldo, S.: 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 (2010). doi:10.1186/1471-2148-10-210
39. Price, M.N., Dehal, P.S., Arkin, A.P.: Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. Molecular biology and evolution **26**(7), 1641–1650 (2009). doi:10.1093/molbev/msp077
40. Lozupone, C., Knight, R.: Unifrac: a new phylogenetic method for comparing microbial communities. Applied and environmental microbiology **71**, 8228–8235 (2005). doi:10.1128/AEM.71.12.8228-8235.2005
41. Sims, D., Sudbery, I., Ilott, N.E., Heger, A., Ponting, C.P.: Sequencing depth and coverage: key considerations in genomic analyses. Nature reviews. Genetics **15**, 121–132 (2014). doi:10.1038/nrg3642
42. McMurdie, P.J., Holmes, S.: phyloseq: an r package for reproducible interactive analysis and graphics of microbiome census data. PloS one **8**(4), 61217 (2013). doi:10.1371/journal.pone.0061217
43. Evans, C., Hardin, J., Stoebel, D.M.: Selecting between-sample RNA-seq normalization methods from the perspective of their assumptions. Briefings in Bioinformatics **19**(5), 776–792 (2017). doi:10.1093/bib/bbx008
44. Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., *et al.*: A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. Briefings in bioinformatics **14**(6), 671–683 (2013). doi:10.1093/bib/bbs046
45. Ondov, B.D., Bergman, N.H., Phillippy, A.M.: Interactive metagenomic visualization in a web browser. BMC Bioinformatics **12**(1) (2011). doi:10.1186/1471-2105-12-385
46. Conway, J.R., Lex, A., Gehlenborg, N.: Upsetr: an r package for the visualization of intersecting sets and their properties. Bioinformatics **33**(18), 2938–2940 (2017). doi:10.1093/bioinformatics/btx364
47. Hourdel, V., Volant, S., O'Brien, D.P., Chenal, A., Chamot-Rooke, J., Dillies, M.-A., Brier, S.: Memhdx: an interactive tool to expedite the statistical validation and visualization of large hdx-ms datasets. Bioinformatics **32**(22), 3413–3419 (2016). doi:10.1093/bioinformatics/btw420
48. Vonaesch, P., Morien, E., Andrianonimiadana, L., Sanke, H., Mbecko, J.-R., Huus, K.E., Naharimanananirina, T., Gondje, B.P., Nigatoloum, S.N., Vondo, S.S., *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), 8489–8498 (2018). doi:10.1073/pnas.1806573115
49. Acinas, S.G., Sarma-Rupavtarm, R., Klepac-Ceraj, V., Polz, M.F.: 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)
50. Sipos, R., Székely, A.J., Palatinszky, M., Révész, S., Márialigeti, K., Nikolausz, M.: Effect of primer mismatch, annealing temperature and pcr cycle number on 16s rrna gene-targetting bacterial community analysis. FEMS microbiology ecology **60**, 341–350 (2007). doi:10.1111/j.1574-6941.2007.00283.x

Tables

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'

Figures

Figure 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.

Figure 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 oropharyngeal core microbiota. (d) Log 2 abundance of differential abundant taxa from oropharyngeal core microbiota in stunt and non-stunt children feces.

Additional Files

Appendix 1: Mathematical definition of the contingency table filtering

Let us denote by $\mathscr{F}$ the entire set of OTU. We propose to only consider a subset $\mathscr{R}$ of OTUs, for the analysis: $\mathscr{R} = \mathscr{L}_1 \cap \mathscr{L}_2$, where $\mathscr{L}_1$ is defined by

$$\mathscr{L}_1 = \left\{ f \in \mathscr{F} \middle| \sum_j \mathbb{1}_{\{c_{fj}>0\}} \geq l_1 \right\} \quad \text{with} \quad l_1 = \lfloor 0.8 \times k_{max} \rfloor,$$

where c_{fj} is the abundance of the feature f in the sample j while k_{max} is the maximum number of samples in which the feature is found. The subset $\mathscr{L}_2$ corresponds to the features with a not too small abundance and is defined by

$$\mathscr{L}_2 = \left\{ f \in \mathscr{F} \middle| \sum_j c_{fj} \geq l_2 \right\},$$

where l_2 is the intercept of the linear regression between the variables $y_k = \sum_i \mathbb{1}_{\left\{\sum_j c_{ij} > x_k\right\}}$ and $x_k \in \left[\min_{f \in \mathscr{F}} \left(\sum_j c_{ij} \right) ; \lambda \right]$. λ is a tuning parameter whose default value is $\left\lfloor \frac{\sum_{ij} c_{ij}}{n} \times 0.05 \right\rfloor$.

Supplementary Figure 1: SHAMAN workflow.

Figure S1 Shaman workflow. SHAMAN can start from raw reads or from processed data. In this last case, it needs at least three tables, the count matrix, the annotation table and the metadata which can either be provided into three different files or by using the BIOM format. The user can then select the variables of interest and add some batch effects. Once run, contrast vectors can easily be defined and interactive visualizations are available.

Supplementary Figure 2: Boxplots of the average coefficient of variation for DESeq2, non-null, weighted non-null and Phyloseq normalization.

Figure S2 Boxplots of the average coefficient of variation for DESeq2, non-null, weighted non-null and Phyloseq normalization. The sparsity of the count matrices is chosen within $\{0.28, 0.64, 0.82\}$ and the number of observations vary within $\{4, 10, 30\}$. 500 normalizations were performed at each level of sparsity and for each number of observations. The results were analyzed by using a t-test, ***$p < 0.001$. DESeq2 normalization did not converged when the matrix sparsity and the number of observations was high (e.g., with a sparsity of 0.64 and 30 observations).

Supplementary Figure 3: Filtering of Zymo mock contaminant genera according to occurrence in samples and abundance.

Figure S3 Filtering of Zymo mock contaminant genera according to occurrence in samples and abundance. Genera occurring in at least 12 samples and with log abundance ≥ 5 were kept in the study (green dots). Cutibacterium, Pelomonas and Sphingomonas genera counts (red dots) were removed from the contingency table.

Supplementary Figure 4: Barplot of the 12 most abundant genera in duodenal, gastric and feces from children of Central African Republic (CAR) and Madagascar.

Figure S4 Barplot of the 12 most abundant genera in duodenal, gastric and feces from children of Central African Republic (CAR) and Madagascar. Haemophilus and Neisseria genera were also observed in cultured samples.

Supplementary Table 1: Summary of taxa differentially abundant when compared 25 and 30 amplification cycles for 0.5ng DNA load.

ID	Base mean	Fold Change	Log2 fold change	P-value adjusted
Bacillus	23314	1.613	0.69	0.0.0058
Listeria	22410	1.633	0.708	0.0.0058
Salmonella	19308	1.767	0.822	0.0.0058
Pseudomonas	8696	1.609	0.686	0.0045
Staphylococcus	26546	1.496	0.58	0.0045
Escherichia-Shigella	13921	1.432	0.518	0.0051
Enterococus	11024	1.417	0.504	0.0084

Table S1 Summary of taxa differentially abundant when compared 25 and 30 amplification cycles for 0.5ng DNA load Positive fold change indicates an increase of abundance of the taxa for the 25 amplification cycles.

Supplementary Table 2: Summary of taxa differentially abundant when compared 25 and 30 amplification cycles for 1ng DNA load.

Id	Base mean	Fold change	Log2 fold change	P-value adjusted
Enterococcus	11024	1.385	0.47	0.0303
Listeria	22410	1.427	0.514	0.0303
Salmonella	19308	1.473	0.559	0.0303
Staphylococcus	26546	1.381	0.465	0.0303
Escherichia-Shigella	13921	1.322	0.402	0.0393
Pseudomonas	8696	1.409	0.495	0.0414

Table S2 Summary of taxa differentially abundant when compared 25 and 30 amplification cycles for 1ng DNA load Positive fold change indicates an increase of abundance of the taxa for the 25 amplification cycles.

Supplementary Table 3: Summary of taxa differentially abundant when compared samples of stunted to non stunted children.

Id	Base mean	Fold change	Log2 fold change	P-value adjusted
Porphyromonas	4.73	3.772	1.915	5.755e-11
Neisseria	7.16	2.789	1.48	1.605e-09
Lactobacillus	51.69	5.152	2.365	4.962e-08
Prevotella	5.04	2.354	1.235	5.836e-05
Ruminococcaceae UCG-009	3.46	0.419	-1.255	0.0003
Weissella	11.33	3.927	1.974	0.0003
Actinobacillus	3.98	2.672	1.418	0.0006
Rikenellaceae RC9 gut group	86.21	0.33	-1.598	0.0028
Ruminococcaceae UCG-011	502.47	1.85	0.887	0.0028
Streptococcus	237.2	1.723	0.785	0.0036
Christensenellaceae R-7 group	49.22	0.534	-0.906	0.0042
Aggregatibacter	4.31	2.222	1.152	0.0061
Granulicatella	0.81	1.991	0.993	0.0061
Ureaplasma	5.52	2.94	1.556	0.0061
Ruminococcaceae UCG-002	144.01	0.601	-0.735	0.0065
Streptobacillus	0.63	3.351	1.745	0.0065
Terrisporobacter	4.29	0.514	-0.961	0.0065
[Eubacterium] xylanophilum group	2.05	0.405	-1.305	0.0073
Fusobacterium	28.55	1.964	0.974	0.0124
Abiotrophia	0.45	2.733	1.45	0.0126
Campylobacter	89.02	1.739	0.798	0.0181
[Eubacterium] coprostanoligenes group	119.89	0.657	-0.606	0.0181
Capnocytophaga	0.19	3.005	1.587	0.0181
Ruminococcaceae UCG-005	211.21	0.683	-0.55	0.0192
Ruminococcaceae UCG-010	31.58	0.57	-0.81	0.0192
Veillonella	475.74	1.515	0.599	0.0218
Ruminococcaceae NK4A214 group	36.02	0.651	-0.62	0.0257
Morganella	0.42	4.099	2.035	0.0271
Haemophilus	260.89	1.494	0.58	0.0301
Family XIII AD3011 group	9.67	0.653	-0.614	0.0344
Lactococcus	35.73	2.161	1.112	0.0373
Methanobrevibacter	5.99	0.321	-1.638	0.0429
Turicibacter	10.03	0.534	-0.9	0.0429
Escherichia-Shigella	25.03	1.574	0.654	0.0444
Kingella	0.47	2.822	1.497	0.0488

Table S3 Summary of taxa differentially abundant when compared samples of stunted to non stunted children. Positive and negative fold changes indicate respectively an increase of abundance of the taxa in stunted and non-stunted children.

Figures

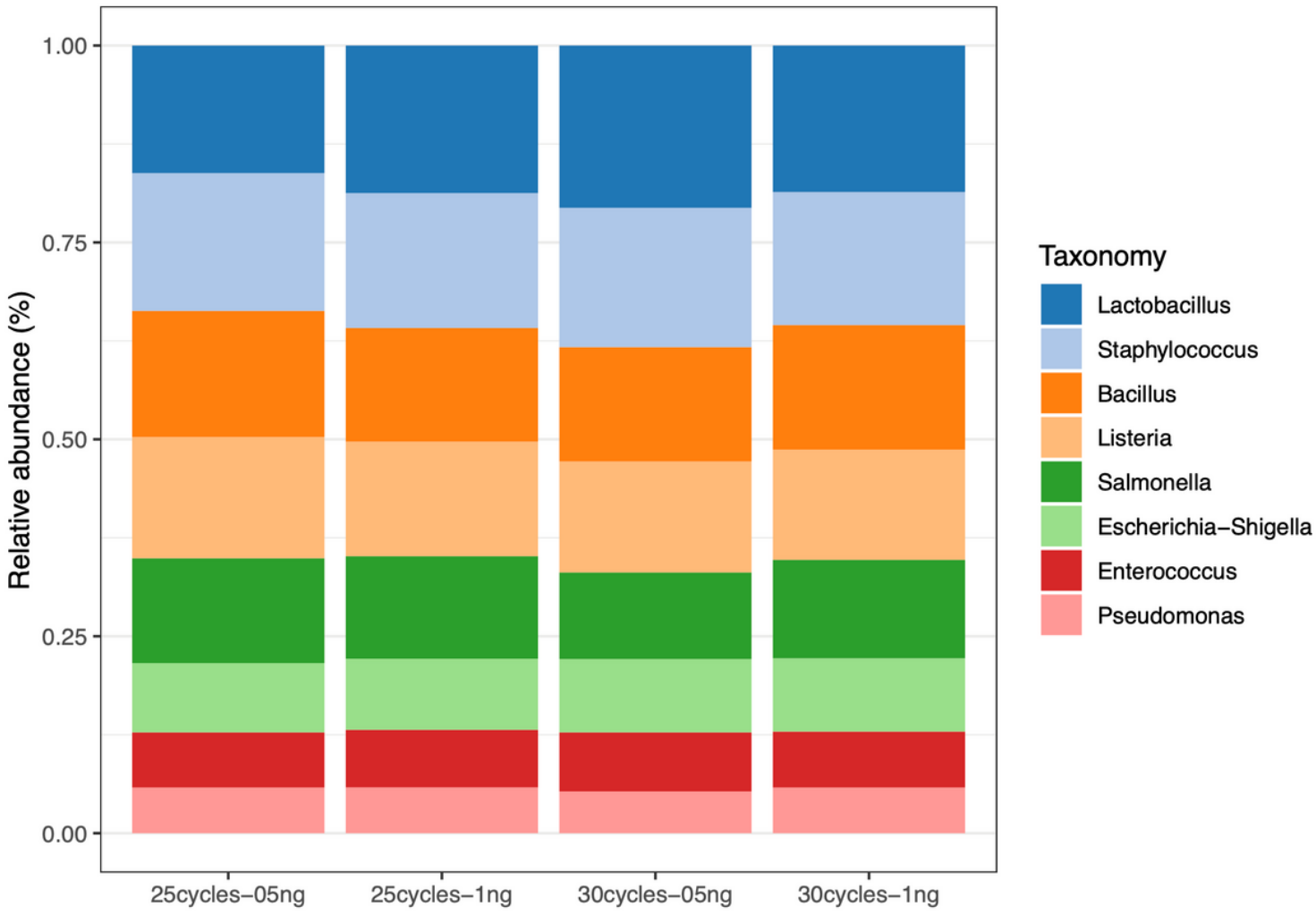

Figure 1

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.

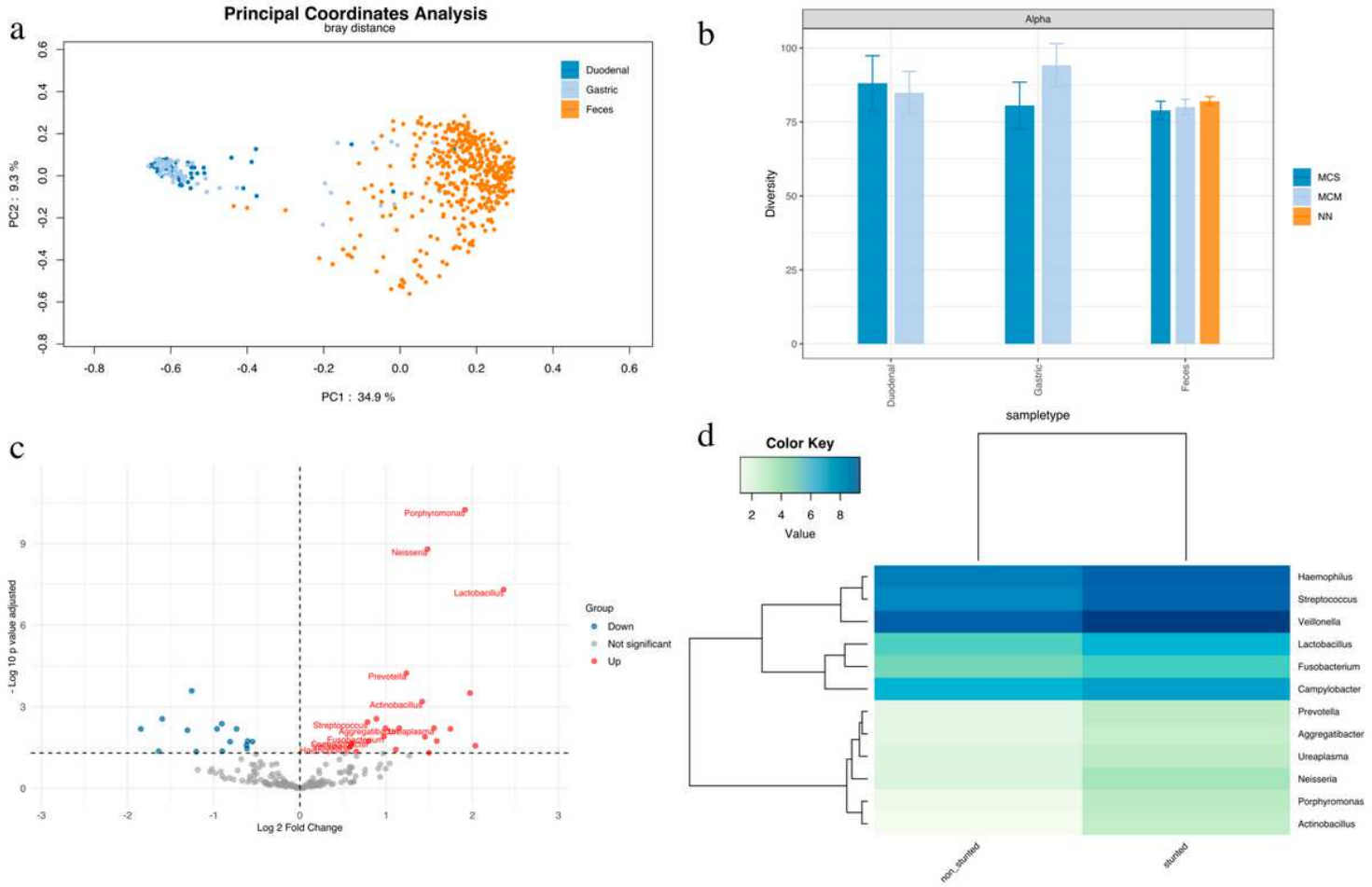

Figure 2

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 oropharyngeal core microbiota. (d) Log 2 abundance of differential abundant taxa from oropharyngeal core microbiota in stunt and non-stunt children feces.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- bmcarticle.bib

- figS1.eps
- figS2.png
- figS3.png
- figS4.eps