SUPPLEMENTARY INFORMATION

FROGS: Find, Rapidly, OTUs with Galaxy Solution

Frédéric Escudié\textsuperscript{1*}, Lucas Auer\textsuperscript{2+}, Maria Bernard\textsuperscript{3}, Mahendra Mariadassou\textsuperscript{4}, Laurent Cauquil\textsuperscript{5}, Katia Vidal\textsuperscript{5}, Sarah Maman\textsuperscript{5}, Guillermina Hernandez-Raquet\textsuperscript{6}, Sylvie Combes\textsuperscript{5} and Géraldine Pascal\textsuperscript{5*}.

\textsuperscript{1} Bioinformatics platform Toulouse Midi-Pyrenees, MIAT, INRA Auzeville CS 52627 31326 Castanet Tolosan cedex, France, \textsuperscript{2} INRA, UMR 1136, Université de Lorraine, INRA-Nancy, 54280, Champenoux, France, \textsuperscript{3} GABI, INRA, AgroParisTech, Université Paris-Saclay, Jouy-en-Josas, France, \textsuperscript{4} MalAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France, \textsuperscript{5} GenPhySE, Université de Toulouse, INRA, INPT, ENV'T, Castanet Tolosan, France, \textsuperscript{6} Laboratoire d'ingénierie des Systèmes Biologiques et des Procédés-LISBP, Université de Toulouse, INSA, INRA, CNRS, Toulouse, France.

*To whom correspondence should be addressed. +These authors contributed equally to this work.
This document contents:

SUPPLEMENTARY METHODS ________________________________________________________ 3
Tests with in silico data .......................................................................................... 3
Tests with real data ................................................................................................. 3
Benchmark process and metrics ............................................................................ 4
Statistics .................................................................................................................. 4
SUPPLEMENTARY RESULTS ______________________________________________________ 6
Data types processed by FROGS............................................................................. 6
Inputs and outputs of main FROGS tools ............................................................... 6
Acquisition data tool ............................................................................................... 7
Demultiplexing tool ................................................................................................ 7
Normalization tool ................................................................................................... 7
Cluster_stat tool and affiliation_stat tool ............................................................... 7
Formatting tools ....................................................................................................... 8
SUPPLEMENTARY FIGURES ______________________________________________________ 9
Figure S1 .................................................................................................................. 9
Figure S2 ................................................................................................................ 10
Figure S3 ................................................................................................................ 11
Figure S4 ................................................................................................................ 12
Figure S5 ................................................................................................................ 14
Figure S6 ................................................................................................................ 15
Figure S7 ................................................................................................................ 16
Figure S8 ................................................................................................................ 17
Figure S9 ................................................................................................................ 18
Figure S10 ............................................................................................................... 19
Figure S11 ............................................................................................................... 20
Supplementary information - References __________________________________________ 21
SUPPLEMENTARY METHODS

Tests with *in silico* data

Grinder (v. 0.5.3) (Angly, et al., 2012) was used to simulate the PCR amplification of full length (V3V4 and V4) sequences from reference databases. We generated 25 sets of species manually extracted from UTAX (Simulated Data From UTAX = SDFU) and 25 others from SILVA (v123) databank (Quast, et al., 2013) (Simulated Data From SILVA = SDFS) (Figure 1 of main text and companion website: tab SDFU/Datasets and SDFS/Datasets). We generated amplicons by (i) filtering out sequences with ambiguous nucleotides, (ii) keeping only bacterial species with non-ambiguous affiliation taxonomy and with pintail> 50 for sequences from SILVA (Ashelford, et al., 2005), and (iii) with a match (with 10% of mismatches allowed) for the forward (TACGGRAGCAGCAG) and reverse (TAGGATTAGATACCCTGGTA) primers in the V3V4 region and for the forward (GTGCCAGCMGCCGCGGTAA) primer in the V4 region, and (iv) maximizing the phylogenetic diversity of the amplicons in the full length 16S phylogenetic tree. This results in 25 increasingly complex nested databases. Grinder requires both error and abundance profiles to generate sequences. We used the following error parameters: the error rate increases linearly from 0.301% to 0.303% per base along the read, 98.6% of errors are SNPs and 1.4% are indels. These parameters were calibrated by mapping reads from a single strain (D'Amore, et al., 2016; Schirmer, et al., 2016; Schirmer, et al., 2015) MiSeq sequencing run to its known sequence to mimic typical MiSeq error profiles and agree with other reported values. We used the default n-mer distribution: 89% of bimeras, 11% trimeras and 0.3% of quadrimeras, corresponding to the average values published in Quince et al. 2011 (Quince, et al., 2011). The fraction of chimera increased with the reference database size to reflect increasing sequence similarities: 5% for 20 taxa, 10% for 100 and 200 taxa and 20% for 500 and 1 000 taxa. Chimera breakpoints were distributed uniformly along the amplicon. We considered two different abundance profiles: uniform and power law. For a power law abundance profile, parameters were calibrated to set the expected max/min abundance ratio to 100 (20 taxa), 1 000 (100 and 200 taxa) or 10 000 (500 and 1 000 taxa). For each combination of database sizes (20/100/200/500/1 000), abundance profiles (uniform, power law), amplicon regions (V3V4/V4), we generated 5 communities with different compositions (Figure 1 of main text). We then simulated 10 samples of 100 000 reads each from each community. Finally, we used cutadapt (v1.7.1) to trim primers from the generated reads. Trimmed sequences were not preprocessed with quality filters but instead used as such in downstream analyses. These 2.10^+8 sequences were treated with FROGS, mothur, UPARSE and QIIME, each time using their own guidelines, to compare the performances of these four solutions.

Tests with real data

The real dataset comes from the publicly available BEI Resource (BEI: HM-278D, HM-279D). It is a synthetic mock community of 20 known bacteria, 1 yeast and 1 archaea, from genera commonly found on or within the human body (HMP project http://www.hmpdacc.org/HMMC/). Genomic DNA from each organism was mixed, based on qPCR of 16S rRNA measurements, and 2 mixture types are available: an even mixture, where aliquots were based on equimolar rRNA operon counts per organism and a staggered ones, where the rRNA operon counts vary by up to 4 orders of magnitude. The mock community composition is available on http://downloads.hmpdacc.org/data/HMMC/HMPRP_sT1-Mock.pdf. We were only interested in 20 species of bacteria.
Benchmark process and metrics

The guidelines as of September, 2016 were followed for mothur v1.33.1 (http://www.mothur.org/wiki/MISeq_SOP), for UPARSE v8.1.1861_i86linux32 from (http://drive5.com/usearch/manual/uparse_pipeline.html), for QIIME http://qiime.org/tutorials/otu_picking.html and for FROGS v1.4.1 cf. companion website tabs: SDFU/commands SDFS/commands. The homemade script to launch these pipelines can be found on https://github.com/geraldinepascal/FROGS/blob/v1.4.0/assessment/bin/assessment.py. The results were evaluated for four metrics (divergence rate, the number of false-negative taxa (FN), number of false positive taxa (FP) and the number of Supernumerary OTUs (SO), (Supplementary Figure S11)). The three first were computed per sample and the fourth one per community (i.e. after pooling samples originating from the same community).

Statistics

The four metrics (Supplementary Figure S11) used to compare results of FROGS, UPARSE, QIIME and mothur are: (i) the first one, called divergence, evaluates a tool's ability to recover the expected community composition at different taxonomic levels. It is defined as affiliation divergence = 100 - ∑ (min (expected, found)) and is equal to the Bray-Curtis distance (expressed in percent) between the expected (i.e. the true taxonomic composition of the community) and observed (i.e. the one inferred by the OTU picking tool) community compositions, after aggregating OTUs at a given taxonomic level. The Bray-Curtis distance equation: d(A, B) = ∑[|ai−bi|] where ai is the abundance of taxa i in community A. The divergence is measured at all taxonomic levels from phylum to either genus (UTAX) or species (SILVA); (ii) the second metric, called FN, is the number of false negative taxa i.e. present in the original bacterial community but not discovered by the OTU picking method. These are the OTUs lost while picking OTUs. This second metric evaluates the sensitivity, i.e. the probability of detection of species present in dataset. It is formally defined as Expected_species - Expected_retrieved where Expected_species is the number of distinguishable species present in dataset after simulation and Expected_retrieved is the number of those species actually found. Species lost in the simulation step due to their low theoretical abundance do not contribute to Expected_species. Similarly, species with the exact same sequence on the amplified regions, are combined and contribute as only one species to Expected_species, since they cannot be distinguished. Expected_retrieved is the number of distinguishable (for each sample) species found as a seed (i.e. representative sequence) for some OTU; (iii) the third one, called FP is number of false positive taxa i.e. discovered by the OTU picking method but not present in the original bacterial community. This metric evaluates a tool’s propensity to create spurious OTUs. It is formally defined as the number of OTUs detected in a sample minus Expected_retrieved. Chimera, supernumerary OTUs for a given distinguishable species, invalid constructions, and contamination by other samples all contribute to false OTUs; (iv) the fourth metric, called supernumerary OTU was evaluated at the community level but was not subjected to statistical tests. It evaluates the capacity to aggregate all the sequences coming from the same initial sequence. Supernumerary OTU is the number of seeds beyond the first one derived from an original sequence. Since these OTUs come from the same original sequence, they should be clustered into a single OTU.

Knowing the existing biases during the construction of biological mocks, we have also distinguished the OTUs produced by the four pipelines in three classes to further characterize them: true, accepted and spurious OTU. We blasted our reconstructed OTUs against the corresponding regions of the 16S rRNA sequences of the 20 species used in the BEI mock. The OTUs were classified into 3 classes according to their blast results (i) true OTU if the cluster seed had 100% of identity and 100% of coverage with one of
the 20 species 16S rRNA, (ii) accepted OTU if the seed had > 97% of identity over > 95% of coverage with one of the 20 species 16S rRNA, (iii) spurious OTU if the seed had < 97% of identity with one of the 20 species 16S rRNA. For each class, we counted both (i) the number of reconstructed OTUs in that class and (ii) their aggregate contribution to the community. We also blasted the 20 16S rRNA sequences that were supposed to be retrieved against the recovered OTU to find how many were really missed, as opposed to reconstructed as a slightly erroneous OTU. Again, the retrieved sequences were classified in 3 classes according to their blast results: (i) retrieved as true (true) if the sequence had 100% of identity against one of the cluster seed, (ii) retrieved as accepted (accepted) if the sequence had between 97% and 100% of sequence identity, (iii) not retrieved (not) otherwise. The number of species in each category are shown in the bottom panel of the figures (Retrieved).
SUPPLEMENTARY RESULTS

Data types processed by FROGS

FROGS works on amplicon sequences that do not have size polymorphisms. The most used amplicon are ribosomal DNA (16S, 18S, 23S), but amplicons belonging from functional genes such as dsrB or rpoB were also processed by FROGS. The only limit is the reference database provided to the software, but one is free to add suitable database. FROGS has been designed to manage data from Illumina sequencers (MiniSeq, MiSeq, NextSeq, HiSeq,...) with standard protocol or with Kozich et al. 2013 protocol i.e. a dual-index sequencing strategy including primers (primers are absent in final reads). It can, however, accept data from 454 if sequences are in fastq format. In standard protocol, target DNA must be completely sequenced in the reads i.e. either a single-end reads starting from 5’ primer and finishing to the end of the 3’ primer or overlapping paired-end reads. The paired-ends can be processed if they are already merged. FROGS can analyse multiplexed or demultiplexed data.

Inputs and outputs of main FROGS tools

**Preprocessing tool**
The requested inputs are the sequence files, individuals or into an archive. Sequences can be either already merged or separated in R1 and R2 files. The tool dereplicates sequences, so it generates a fasta output file in which strictly identical sequences are represented only by one sequence. Another output file contains the number of all unique sequences in each sample. The user has also access to an html report with graphics that shows the number of kept sequences at each cleaning step. This makes it possible to check the general configuration of cleaned sequences and, a posteriori, to see if the sequencing is correct.

**Clustering tool**
Input files are (i) the dereplicated file containing all sequences from all samples, and (ii) the count file of all the unique sequences in each sample. Output files produced by the clustering tool are (i) a file with the abundance of each cluster in each sample in biom format (McDonald, et al., 2012) as this format is widely used by metagenomics software, (ii) a fasta file with seed sequences, i.e. the representative sequence of each cluster, and (iii) a text file representing the read composition of each cluster.

**Chimera removal tool**
The only requested input is the sequence file (either the fasta file of seed sequences from clustering, or all sequences from the preprocessing if chimera have been removed prior to clustering). Output files obtained after chimera processing tool are (i) the fasta file with non-chimeric sequences, (ii) the abundance file of non-chimeric clusters and (iii) the summary file that presents the number of elements removed.

**Filtering tool**
Expected input files for this tool are the sequence file and the abundance file. For clusters that pass filtering, the outputs are (i) the sequence file and (ii) the corresponding abundance file, (iii) the list of deleted OTUs and (iv) a report file in html format, with tables and graphics showing the filter impacts.
**Affiliation tool**

Expected input files for this affiliation tool are the sequence and abundance file. Outputs are abundance files including the affiliation of OTUs in biom format, and a report file in html format, including tables and graphics giving details on multi-affiliations.

**Acquisition data tool**

FROGS allows uploading of a whole dataset as a Tar archive on a Galaxy platform. This tool is helpful for old Galaxy platform versions that do not allow uploading of the Tar archive. This tool avoids uploading each file produced by sequencers onto the Galaxy platform one after the other.

**Demultiplexing tool**

A first homemade python script allows the demultiplexing of multiplexed data by assigning sequences to their original sample. The required input files are a descriptive file of barcodes used during sequencing and the file of pair-end or single-end sequences (fastq format). The barcode file is expected to be tabulated. The first column corresponds to the sample name (unique, no spaces), the second column corresponds to the forward sequence barcode used (none if only reverse barcode) and the third column corresponds to the reverse sequence barcode (optional). For each sequence or pair of sequences, the sequence fragment corresponding to the chosen multiplexing will be compared with all the barcode sequences. If this fragment is equal (with fewer or the same number of mismatches as the threshold) to one (and only one) barcode, the fragment is trimmed and the sequence will be attributed to the corresponding sample. Final fastq files (or pair of fastq files) of each sample are included in an archive, and a report describes how many sequences are attributed for each sample.

**Normalization tool**

This tool normalizes sample depth, by randomly sampling a given number of sequences in each sample. It is optional but is quite common and sometimes required for some statistical methods. Expected input files for this tool are the sequence and abundance files. Outputs are a normalized fasta file, a normalized abundance file in biom format and a report in html format, with tables and graphics showing the consequences of normalization for the distribution of OTUs.

**Cluster_stat tool and affiliation_stat tool**

Cluster_stat and affiliation_stat tools describe respectively clustered and affiliated data with tables, graphics, bar plots, diagrams and curves. In particular, the cluster_stat tool allows the observation of sequences and OTU distribution through samples (Supplementary Figure S1) at each step of the treatment. It also provides a hierarchical clustering tree of samples, based on Bray Curtis distance and details the number of OTUs and sequences that are shared or unique in each sample. These two outputs are very informative for technical or biological replicated samples. All outputs are grouped in an html file. The affiliation_stat tool computes several metrics and generates an html file describing OTUs based on their taxonomies and on the quality of their affiliations. It also produces rarefaction curves at different taxonomic levels, RDP Classifier bootstrap bar plots and total or per sample diversity diagrams (Supplementary Figure S1). Except for the last diagram, all the figures and tables can be downloaded.
Formatting tools

Most FROGS tools have outputs in biom format. The BIOM_to_TSV tool converts this informatics format into a human readable tabular format. If it is used after OTU affiliation, the tool also generates a multi-affiliation output file that contains details of multiple affiliations. The resulting abundance table can thus be easily modified, for example by correcting “multi-affiliation” tags when the appropriate affiliation is known. Modified TSV files can be converted back into biom format using the TSV_to_BIOM tool. The third formatting tool, FROGS_biom_to_std_biom allows the conversion of FROGS biom format to standard biom as used in the UPARSE, QIIME or R packages. This step is necessary because standard biom files have only one taxonomy tag whereas the FROGS biom can include the RDP Classifier and blast taxonomies.
Median execution time (minutes) of FROGS per CPU number, on a community of 5 datasets of 10,000,000 sequences with a diversity of 100 species whose abundancies are distributed according to a power law.
Sets of figures included in html output files of FROGS tools. A: Pie charts produced by the filtering tool, the pie chart on the left shows the number of clusters removed or retained after filtering (respectively black and blue) and their equivalence in number of sequences on the pie chart on the right. This type of pie chart is also produced by the remove_chimera tool; B: Graph representing the length distribution of amplicons before the trimming and cleaning steps performed by the preprocessing tool. The equivalent graph also exists for amplicons after these steps; C: Rarefaction curve produced by the affiliation_stat tool; D: Table produced by the cluster_stat tool showing the number and distribution of clusters and sequences in each sample; E: Table produced by the affiliation_stat tool showing the quality of blast affiliation (%identity and %coverage) according to the number of clusters or sequences per sample; F: Graph of the proportion of cumulative sequences according to cluster size produced by the cluster_stat tool; G: Venn diagram produced by the filtering tool showing which OTUs were deleted by 6 different filters; H: Two sunburst diagrams produced by the affiliation_stat tool showing the affiliation distribution of OTUs according to their taxonomic level with two kinds of color panels; I: Box plot produced by the cluster_stat tool showing the cluster size distribution; J: Graph showing hierarchical clustering produced by the cluster_stat tool showing the frequency of samples according to their cluster composition (distance method = Bray Curtis, linkage method = average); K: Bootstrap distribution bar plot produced by affiliation_stat tool, showing the bootstrap values of RDPClassifier software for each taxonomy levels. L: Bar plot produced by the preprocessing tool showing the number of sequences resulting from the preprocessing steps. M: Another Venn diagram with only 3 parameters produced by the filtering tool.
Comparison of divergence by FROGS and competing pipelines (UPARSE, mothur and QIIME) with their affiliation step replaced by the FROGS Multi-Affiliation (MA) strategy (UPARSE (SOP) is not available on SDFS). Pipelines were compared on in silico communities generated from SILVA (SDFS, see Figure 1 of main text) based on the divergence metric at different taxonomic levels. The 10 replicates of each community were used to perform a Mann-Whitney non-parametric paired test. If the divergences differed significantly (p < 0.05), we deemed FROGS to be the best (green) if it had the lowest divergence, otherwise its competitor was deemed to be best (blue). If the test was not significant, the pipelines tied (grey).
Figure S4

A: Comparison of divergence by FROGS and competing pipelines (UPARSE, mothur and QIIME) with their affiliation step replaced by the FROGS Multi-Affiliation (MA) strategy. Pipelines were compared on in silico communities generated from UTAX (SDFU, see Figure 1 of main text) based on the divergence metric at different taxonomic levels. The 10 replicates of each community were used to perform a Mann-Whitney non-parametric paired test. If divergences were significantly different (p < 0.05), we deemed FROGS to be best (green) if it had the lowest divergence, and otherwise, we deemed the competitor to be best (blue) else. If the test was not significant, the pipelines tied (grey).
B: Targeted comparison of FROGS and QIIME (MA) divergences on in silico communities generated from UTAX (SDFU, see Figure 1 of main text) at different taxonomic levels. QIIME (MA) divergences (pink) are generally higher than FROGS divergences (green), sometimes radically so, except for rich communities with uniform abundances studied with the V4 region (bottom row). For those communities, although significant, the difference between the two pipelines is around 2 divergence points at genus level, where both FROGS and QIIME (MA) divergences are already not low (6-10%).
Comparison of FROGS and competing pipelines UPARSE on *in silico* communities generated from UTAX (SDFU, see Figure 1 of main text) based on the number of FP OTUs (left) and FN OTUs (right). The same as Figure 4 of main text but mothur and QIIME have been removed to focus on FROGS and UPARSE.
Comparison of FROGS and competing pipelines’ (UPARSE, mothur and QIIME) FPs and FNs. Pipelines were compared on in silico communities generated from UTAX (SDFU, see Figure 1 of main text) based on the number of false negative OTUs (FN) and the number of false positive OTUs (FP). The 10 replicates of each community were used to perform a Mann-Whitney non-parametric paired test. If divergences differed significantly (p < 0.05), we deemed FROGS to be the best (green) if it had the lowest divergence, and otherwise, we deemed its competitor to be the best (blue). If the test was not significant, pipelines tied (grey).
Comparison of FROGS and UPARSE/mothur/QIIME (SOP) divergences on the 20 species of real mock communities. Because of the absence of replicates, no statistical test was performed. The results show that all pipelines had very similar divergence levels, which were quite high. As expected, divergence was higher at deep taxonomic levels (e.g. genus) than at coarser ones (e.g. phylum).
Comparison of FROGS and UPARSE/mothur/QIIME (SOP) divergences on the 20 species mock communities using the V3V4 amplicon. Because of the small number of replicates (n=4), no statistical test was performed. FROGS excess divergence (A) shows that FROGS achieves slightly less divergences than its competitors when faced with staggered abundances and slightly more divergences with even abundances. Nevertheless, the base divergence of FROGS (B) and in turn of the competing pipelines, is quite high in all settings (> 10% even at the phylum level).
A: Description of the results of FROGS, UPARSE/mothur/QIIME (SOP) on the BEI datasets (using the V3V4 amplicon). Histograms display (top panel, Abundancy label) the aggregate contribution of each OTU class to the community, (middle panel, OTU label) the number of OTUs in each class and (bottom panel, Retrieved label) the number of species effectively recovered as OTU (true/accepted) or missed (not). See Supplementary Text for details of each OTU class. B: same as A but with a focus on FROGS and UPARSE results only.
Figure S10

Comparison of chimera removal rate (median) by the VSEARCH and Usearch chimera detection tools before or after the clustering step (with Swarm software) on 5 communities of 10 samples of 100,000 sequences each (V3V4 region with power law abundance distribution) of 5 different sizes: 20/100/200/500/1,000 species respectively from lighter to darker orange. The bar plot shows that the maximum number of chimera are detected with the VSEARCH tool after clustering with swarm. The table shows the false positive rate (median) on previous data.
Figure S11

<table>
<thead>
<tr>
<th>Expected</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Expected Image" /></td>
<td><img src="image2.png" alt="Observed Image" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1 species with abundance of 30 reads</th>
<th>4 species</th>
<th>4 species</th>
<th>4 species</th>
</tr>
</thead>
<tbody>
<tr>
<td>divergence rate = +50% of abundance</td>
<td>2 FPs = 2 not real OTUs are kept</td>
<td>1 FN = a real OTU is lost</td>
<td>2 SOs = 2 additional OTUs with same origin as the expected OTU</td>
</tr>
</tbody>
</table>

Presentation of the four metrics used to assess FROGS and the other pipelines on *in silico* and real data analysis, *i.e.* divergence rate, number of false positive taxa (FP), number of false negative taxa (FN) and number of supernumerary OTUs (SO).
Supplementary information - References


