FROGS: Find, Rapidly, OTUs with Galaxy Solution

Frédéric Escudié, Lucas Auer, Maria Bernard, Mahendra Mariadassou, Laurent Cauquil, Katia Vidal, Sarah Maman Haddad, Guillermina Hernandez-Raquet, Sylvie Combes, Géraldine Pascal

Bioinformatics, 34(8), 2018, 1287–1294
doi: 10.1093/bioinformatics/btx791
Advance Access Publication Date: 7 December 2017
Original Paper

Sequence analysis

FROGS: Find, Rapidly, OTUs with Galaxy Solution

Frédéric Escudié[1,†], Lucas Auer[2,†], Maria Bernard[3], Mahendra Mariadassou[4], Laurent Cauquil[5], Katia Vidal[5], Sarah Maman[5], Guillermina Hernandez-Raquet[6], Sylvie Combes[5] and Géraldine Pascal[5,*]

[1]Bioinformatics platform Toulouse Midi-Pyrenees, MIAT, INRA Auzeville CS 52627 31326 Castanet Tolosan cedex, France, [2]INRA, UMR 1136, Université de Lorraine, INRA-Nancy, 54280, Champenoux, France, [3]GABI, INRA, AgroParisTech, Université Paris-Saclay, Jouy-en-Josas, France, [4]MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France, [5]GenPhySE, Université de Toulouse, INRA, INPT, ENVT, Castanet Tolosan, France and [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.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Bonnie Berger

Received on May 10, 2017; revised on December 1, 2017; editorial decision on December 4, 2017; accepted on December 5, 2017

Abstract

Motivation: Metagenomics leads to major advances in microbial ecology and biologists need user friendly tools to analyze their data on their own.
Results: This Galaxy-supported pipeline, called FROGS, is designed to analyze large sets of amplicon sequences and produce abundance tables of Operational Taxonomic Units (OTUs) and their taxonomic affiliation. The clustering uses Swarm. The chimera removal uses VSEARCH, combined with original cross-sample validation. The taxonomic affiliation returns an innovative multi-affiliation output to highlight databases conflicts and uncertainties. Statistical results and numerous graphical illustrations are produced along the way to monitor the pipeline. FROGS was tested for the detection and quantification of OTUs on real and *in silico* datasets and proved to be rapid, robust and highly sensitive. It compares favorably with the widespread mothur, UPARSE and QIIME.
Availability and implementation: Source code and instructions for installation: https://github.com/geraldinepascal/FROGS.git. A companion website: http://frogs.toulouse.inra.fr.
Contact: geraldine.pascal@inra.fr
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The expansion of high-throughput sequencing of rRNA amplicons has opened new horizons for the study of microbial communities. By making it possible to study all micro-organisms from a given environment without the need to cultivate them, metagenomics has led to major advances in many fields of microbial ecology, from the study of the impact of microbiota on human and animal pathologies (Hess *et al.*, 2011; Hooper *et al.*, 2012; Jovel *et al.*, 2016) to the study of biodiversity in environmental ecosystems and the search for biomarkers of pollution (Andres and Bertin, 2016; de Vargas *et al.*, 2015). Determining the composition of a microbial ecosystem, at low cost and great depth, is still largely based on the amplification and sequencing of biodiversity marker genes, also called amplicons, such as rRNA genes and ITS. The clustering of sequences into

operational taxonomic units (OTUs), which are used as a proxy for species, is the first step in most studies.

However, the flood of data in the new high throughput approaches is creating a bottleneck that challenges current computing architectures and requires refinement of processing algorithms. Solutions that optimize the processing of these data in terms of infrastructure and computation time are urgently required. Particularly, Illumina data, with dozens of samples routinely sequenced at depths over 100 000 reads, are hard to process in a reasonable time (Cai and Sun, 2011; Fu *et al.*, 2012). Moreover, the most effective solutions are often designed for specialists, and bioinformatics skills are generally needed to use such software. Most have to be launched using command lines and are not always easy to install, making them difficult to use for biologist end users who are not closely connected to a computing facility (Boyer *et al.*, 2016; Caporaso *et al.*, 2010; Edgar 2013; Hildebrand *et al.*, 2014; Jeraldo *et al.*, 2014; Manter *et al.*, 2016; Schloss *et al.*, 2009).

These tools are designed to process amplicon sequences and return an abundance table of OTUs together with their taxonomic affiliations. But sequencing quality remains a notable barrier to accurate taxonomic assignment and α-diversity assessment for microbial communities: if care is not taken, amplicon data can lead to huge over-estimations of bacterial diversity (Kunin *et al.*, 2010). There is, therefore, a huge demand for innovative, efficient, reliable and easy to use tools for biologist end users. In this context, we developed a tool that can be used by the largest possible number of biologists. FROGS: « Find, Rapidly, OTUs with Galaxy Solution » was designed to be used through either command lines or on Galaxy platforms (Blankenberg *et al.*, 2010; Giardine *et al.*, 2005; Goecks *et al.*, 2010). The FROGS pipeline is user friendly with rich graphical outputs. The aim of this paper is to present FROGS and to demonstrate the advantages and accuracy of FROGS compared to mothur (Schloss *et al.*, 2009), UPARSE (Edgar, 2013) and QIIME (Caporaso *et al.*, 2010), using *in silico* and real datasets. All datasets, tests and results are presented on the companion website http://frogs.toulouse.inra.fr. The advantages of FROGS are that it relies on Swarm (Mahé *et al.*, 2014) and its adaptive sequence agglomeration rather than on a global similarity threshold, combined with a rigorous chimera removal step and the explicit consideration of conflicting affiliations. FROGS is also fast and has efficient, scalable and parallelizable algorithms to support the ever-increasing amounts of data.

2 Materials and methods

2.1 Implementation

FROGS is written in python 2.7 and can be downloaded from the GitHub code source repository (https://github.com/geraldinepascal/FROGS). Installation procedure can be found on the GitHub server. FROGS can be installed in a Galaxy instance or only for use in a command line. It requires the following dependencies: the python library scipy, splitbc for demultiplexing (a homemade script in Perl provided with FROGS), Flash (Magoc and Salzberg, 2011) and cutadapt (Martin, 2011), Swarm (Mahé *et al.*, 2014), VSEARCH (Rognes *et al.*, 2016), NCBI blast+ (Camacho *et al.*, 2009) and RDP Classifier (Wang *et al.*, 2007). The installation can be checked with a reduced dataset included in the package. In terms of execution, the longest steps (pre-processing, clustering, chimera removal and affiliation steps) can be threaded on multi-CPU and/or multi-core systems. This parallelization does not require a particular setup and can save a lot of time (Supplementary Fig. S1). The memory used during the process depends on the variability of data and the number of cores used. For example, 10 million sequences from a low complexity community (100 species, power law abundance distribution) can be analyzed in 12 min on 1 CPU using at most 1.3 Gb. Most tools produce an HTML report, including many interactive graphics based on Highcharts (http://www.highcharts.com/) and D3js (https://d3js.org/) libraries.

2.2 Tests with *in silico* data

To measure the ability of FROGS to correctly estimate the number of OTUs in metagenomics datasets and to assess the performance of taxonomic assignment using 16 S sequences, we tested and compared FROGS with mothur, UPARSE and QIIME on 2000 large *in silico* datasets generated as displayed in Figure 1. Therefore, to account for possible biases introduced by the choice of the amplified region, we produced, *in silico*, datasets with the V3V4 and V4 hypervariable regions of the bacterial 16 S gene. We generated 25 sets of species of increasing richness, manually extracted from UTAX http://drive5.com/usearch/manual/utax_downloads.html termed 'simulated data from UTAX' (SDFU) and 25 others from SILVA (v123) databank (Quast *et al.*, 2013) termed 'simulated data from SILVA' (SDFS) (Fig. 1 and companion website http://frogs.toulouse.inra.fr: tab SDFU/Datasets and SDFS/Datasets) (Supplementary Methods). We tested FROGS, UPARSE, mothur and QIIME using their own guidelines for SDFU, i.e. with their own affiliation method on the UTAX databank. These pipelines are called UPARSE_SOP, MOTHUR_SOP and QIIME_SOP (SOP = standard operating procedure). However, the *in silico* SDFU communities are not very diversified because UTAX is smaller than SILVA. For this reason, we also ran the four pipelines on SDFS. We used the appropriate guidelines for each pipeline, except for the affiliation step, for which we used the FROGS affiliation tools, because formatting the SILVA database required for the affiliation step of UPARSE, was too complex to implement. These pipelines are called UPARSE_MA, QIIME_MA and MOTHUR_MA (MA = multi-affiliation of FROGS). QIIME's SOPs do not normally include a chimera removal step, but to be sure our results are fair, notably in terms of erroneous OTUs, we applied it before the clustering step. As simulated data come from known databanks, we used the same databanks for taxonomic assignment. This corresponds to using a 'perfect' databank and allows us to suppress the effect of incomplete reference databanks and focus mostly on composition reconstruction.

2.3 Tests with real data

In parallel, we compared FROGS with mothur, UPARSE and QIIME using a real dataset from the publicly available BEI Resource (BEI: HM-278 D, HM-279 D). It is an artificial mock community of 20 known bacteria, 1 yeast and 1 archaea, from genera commonly found on or within the human body. The V3V4 region of this mock mixture was amplified with the primers used for simulated data and sequenced using the Illumina MiSeq protocol. We added other sequences (Nelson *et al.*, 2014) from the same BEI Resource (SRA project ID PRJEB4688) but sequenced, them, on the V4 only and V4V5 regions (Supplementary Methods). Other real datasets are used and their description can be viewed on the companion website: tab Overview.

2.4 Statistics

A detailed description of the benchmark process is provided in Supplementary Methods. Our benchmark and metrics are computed directly from OTUs and their abundances, as they constitute the

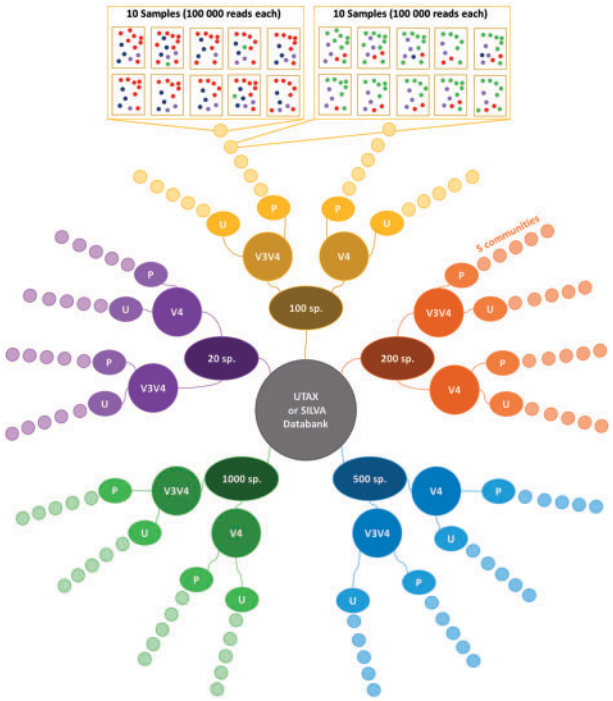

Fig. 1. Diagram of the *in silico* datasets for benchmarking. From the UTAX (SDFU) or SILVA (SDFS) databases, a subset of phylogenetic diverse species (=sp.), ranging from 20 to 1000, were selected. For each species set, either the V3V4 or V4 regions were conserved. The species abundances are distributed according to either a power law (P) or a uniform law (U) to generate five different communities (the five peripheral circles in the figure). The five communities distributed following a power law are made of the same species but in different quantities (c.f. top zoom). Finally, each community was sequenced *in silico* 10 times at a depth of 100 000 reads/sample to create replicates

results of the four pipelines. The four metrics used to compare FROGS, UPARSE, QIIME and mothur are the divergence rate, the number of false negative taxa (FN), the number of false positive taxa (FP) and the number of supernumerary OTUs (SO) (Supplementary Methods). Divergence is defined as the Bray–Curtis distance between expected and observed abundances at a given taxonomic level. FN is the number of taxa present in the original community but not recovered as OTUs by the method. Conversely, FP is the number of spurious OTUs: reconstructed but absent from the original community. SOs is the number of additional OTUs with same origin as the first expected OTUs. For all metrics, lower is better.

For each of the three first metrics, we performed a two-sided paired test, either a parametric (paired *t*-test) or a Mann–Whitney non-parametric test to assess the difference in accuracy between FROGS and each competitor. The tests were performed at the community level (5 per community size x abundance distribution x amplicon region combination) using the 10 samples as replicates (Fig. 1). For each community, we declared FROGS to be better (or worse) than its competitor when (i) the test was significant at the 0.05 level and (ii) FROGS had a lower (or higher) metric than its competitor. When the test was not significant, the methods were deemed tied. Finally, we aggregated the results to explore the parameters (size, abundance distribution and region) favoring one or none of the methods. Given that biological mocks composition are not always in accordance with manufacturer's information, we further characterized the OTUs produced by the four pipelines at the sequence level to assign them to one of three classes: true, accepted and spurious OTU (Supplementary Methods). The full results are presented on the companion website: tab Statistical analyses.

3 FROGS tool overview

FROGS is a set of 13 tools that process amplicon reads coming from Illumina or Roche 454 sequencing technologies. Reads can be in single or paired-ends, merged or not, multiplexed or not, with primers or not (Supplementary Results). It combines both (i) the friendliness of a graphical user interface, with a wide array of graphical diagnostics and descriptive statistics for monitoring (Supplementary Fig. S2), and (ii) the speed (Supplementary Fig. S1) and taxonomic accuracy of dedicated tools. All the tools are independent. The main tools are described below and the others in the Supplementary Results, as well as inputs and outputs of each tools.

3.1 Data pre-processing tool

This tool merges paired-end reads using Flash (Magoc and Salzberg, 2011). It joins read1 and read2 with a customizable rate of mismatches (fixed at 10% by default) in the overlapped region. Next, except for reads from dual-index sequencing (Kozich *et al.*, 2013), cutadapt (Martin, 2011) is used to remove sequences in which the two primers are not present and to trim the primers. Ten percent of mismatches with expected primers are tolerated. The tool cleans the data with user size criteria and removes all sequences containing an ambiguous base. Specifically for 454 data, the tool removes sequences with at least one homopolymer with more than seven nucleotides and with a distance of less than or equal to 10 nucleotides between two poor quality positions, i.e. with a Phred quality score lesser than 10. The tool also dereplicates sequences. The user has access to an html report with graphics that makes possible to check the general configuration of cleaned sequences and, *a posteriori*, to see if the sequencing is correct.

3.2 Clustering tool

FROGS clustering relies on Swarm (Mahé *et al.*, 2014). The FROGS guidelines suggest (Fig. 2) using clustering in two steps, i.e. a first pass of Swarm executed with aggregation parameter $d = 1$ and a second pass performed on the seeds of previous clusters with $d = 3$. As Swarm is very fast, we advise users to perform the clustering step first, in order to reduce the number of sequences and to make the chimera removal step more efficient.

3.3 Chimera removal tool

FROGS chimera detection relies on VSEARCH with *de novo* UCHIME method (Edgar *et al.*, 2011; Rognes *et al.*, 2016). No parameter has to be set. In addition, FROGS uses an innovative cross-sample validation step to confirm the chimeric status on all samples. Chimeras are first detected independently in each sample but in the end, a sequence is only considered chimeric if it is flagged as a chimera in all samples where it is present. Other cases correspond to FP detection. Note that cross-validation leaves chimera detection power unchanged for suspect sequences found in only one sample but requires stronger evidence of chimeric status for suspect sequences found in multiple samples. This approach is different from the one implemented in mothur, where the *dereplicate* parameter removes only the redundant chimeric sequences in the corresponding sample.

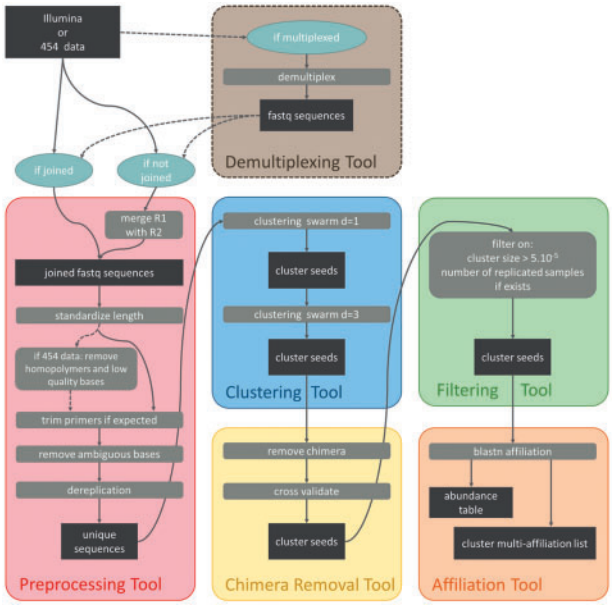

Fig. 2. Flow chart of FROGS SOP. In the clustering tool, *d* = the number of differences, required Swarm parameter (seed = representative sequence of a cluster)

3.4 Filtering tool

FROGS guidelines recommend to apply an abundance filter before the taxonomic affiliation process. Filtering tool makes it possible to screen clusters according to (i) their abundance and their distribution among samples i.e. keep only clusters with at least x sequences and/or present in at least y samples and/or among the z most abundant clusters, (ii) their taxonomic affiliation from the RDP Classifier (Wang *et al.*, 2007) and blastn+ (Camacho *et al.*, 2009) (see affiliation tool section) and (iii) their presence in a contaminant bank, by blastn+ (for now, only including phiX, used in Illumina sequencing technologies). The filtering tool also optionally allows clusters that are not always present in replicated samples to be deleted (companion website http://frogs.toulouse.inra.fr: tab FAQ).

3.5 Affiliation tool

Assignment is done using databanks formatted to include taxonomic levels up to species, so that FROGS offers taxonomic assignment up to the species level. For now, SILVA 16 S (complete or filtered at different levels of pintail score), 18 S, 23 S (Quast *et al.*, 2013), greengenes (DeSantis *et al.*, 2006), Midas (McIlroy *et al.*, 2015) and customized laboratory databases are available in FROGS, but others can be added. Assignment relies on either the RDP Classifier (Wang *et al.*, 2007) or blastn+ (Camacho *et al.*, 2009), both implemented in FROGS. Optional affiliation with the RDP Classifier associates each OTU with a taxonomy and the corresponding bootstrap score. The method recommended in the guidelines is affiliation by blastn+ which finds an alignment between each OTU seed and the database. Only the best hits with the same score are reported. If several blastn+ results have identical scores for an OTU, a taxonomy is determined for each hit at each taxonomic level. If these taxonomies differ across hits, the first level of conflict and all lower ones are set to 'Multi-affiliation.' For example, two hits with equal scores and respective taxonomies *Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales; Bradyrhizobiaceae; Afipia; Afipia birgiae 34632* and *Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales; Bradyrhizobiaceae; Bradyrhizobium; Bradyrhizobium sp.* would give the consensus *Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales; Bradyrhizobiaceae; Multi-affiliation; Multi-affiliation*. A text file with details of the affiliation of all OTUs with ambiguous taxonomies is provided as an output after the TSV formatting (Supplementary Methods). Assignment can be time consuming, so it is recommended to perform it after a first filtering step. The filtering tool can then be used once more as post-treatment and enables data reduction based on affiliation criteria. The produced taxonomic affiliations can be filtered based on (i) the bootstrap of the RDP Classifier taxonomic level (ii) on e-value, identity percentage, coverage percentage and alignment length of blastn+.

4 Benchmarking

We compared FROGS with three very popular applications (UPARSE, QIIME and mothur) using their SOPs, with or without FROGS multi-affiliation method (MA), on both *in silico* and real data. We compared the Divergence rate, FP OTUs, FN OTUs and SOs produced by the pipelines at the community level to assess their performances in different settings.

4.1 Tests on *in silico* data

The tests on simulated datasets revealed that FROGS generally achieved good results, in this case low divergences: < 5% at the genus level on SDFU and on SDFS when using the V3V4 region, much lower at higher taxonomic levels, and slightly higher at the species level (SDFS only) or using the less informative V4 region (companion website: tab Statistical analyses/on simulated data 3.1.1.c and 4.1.1.c). Figure 3 shows the results of non-parametric tests of affiliation divergence on SDFU. It shows that FROGS performed as well as or better than UPARSE and mothur in most settings. FROGS only performed worse than UPARSE for small community sizes (20 species), except at genus level. It performed better than QIIME_SOP in all settings. On SDFS (Supplementary Fig. S3), where communities are more complex but affiliation benefits from the MA strategy of FROGS, the non-parametric tests showed that FROGS performed as well as or better than UPARSE_MA and MOTHUR_MA in most conditions. Again, FROGS only performed worse than UPARSE_MA for small community size (20 species), except at genus level. On data with power law abundance, QIIME_MA and FROGS are comparable but FROGS performed better at genus level. On data with uniform abundance, FROGS performed better than QIIME_MA in most settings using V3V4 region but QIIME_MA performed better with large communities (size > 200 species) using the V4 region. Note that, on SDFU (Supplementary Fig. S4), results are comparable to the previous ones.

We also compared the pipelines with respect to the number of FP and FN OTUs (Fig. 4). On SDFU datasets, we found that (i) the V3V4 region led to more FP and less FN than the V4 and (ii) mothur infers a huge number of FP (up to 20 times more than the expected community size). Focusing on FROGS and UPARSE (Supplementary Fig. S5) revealed that: (i) FROGS always produced fewer FNs than UPARSE but (ii) a few more FPs under power law abundance distributions and a few less under uniform abundance distributions (except for size < 100 species). The paired tests on FP and FN data (Supplementary Fig. S6) supported the previous results and also showed that (i) FROGS truly outperformed mothur in terms of both FP and FN taxa, and (ii) it always produced fewer FPs but sometimes more FNs, especially on the V4 region, than QIIME.

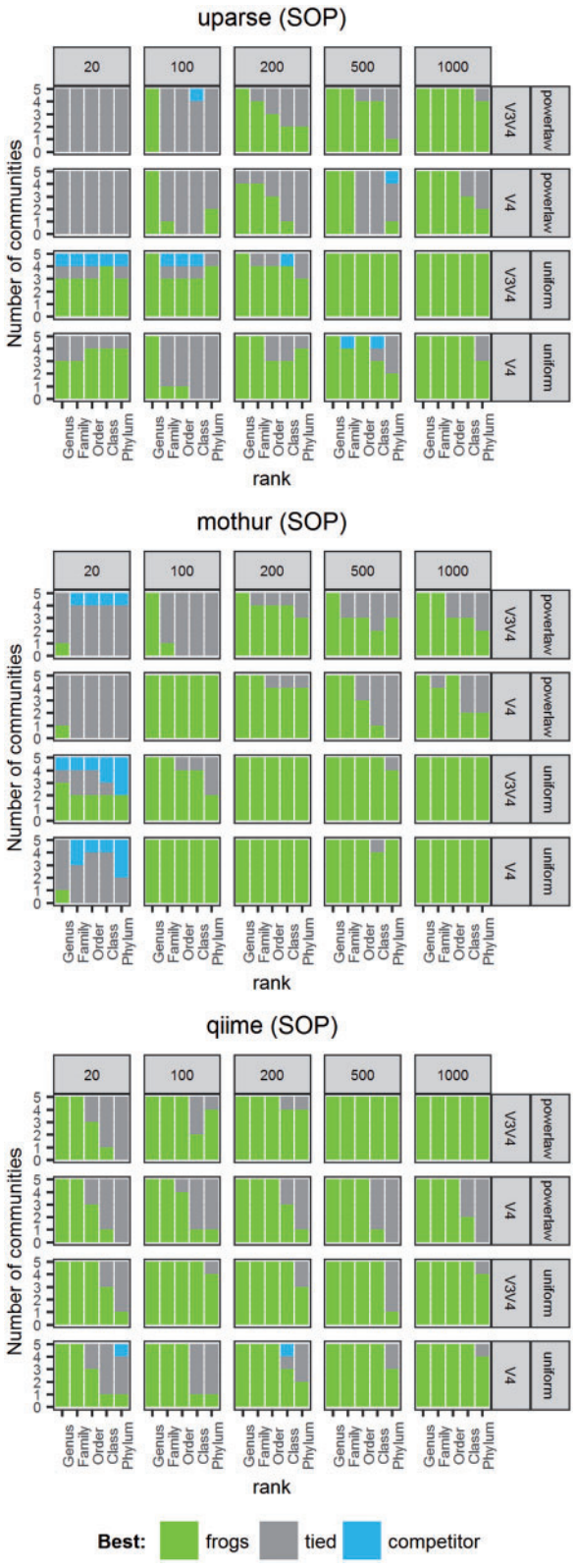

Fig. 3. Comparison of divergence rates at different taxonomic levels between FROGS and competing pipelines (UPARSE, mothur and QIIME) using their SOP. Pipelines were compared on *in silico* communities generated from UTAX (SDFU, Fig. 1). The communities are organized by size (20 to 1000, vertical panels), amplicon and abundance distribution (V3V4/V4 and uniform/power law, horizontal panels). For each, the 10 replicates were used to perform a Mann–Whitney non-parametric paired test and to identify whether FROGS (green), its competitor (blue) or none (grey) achieved significantly lower ($P < 0.05$) divergence

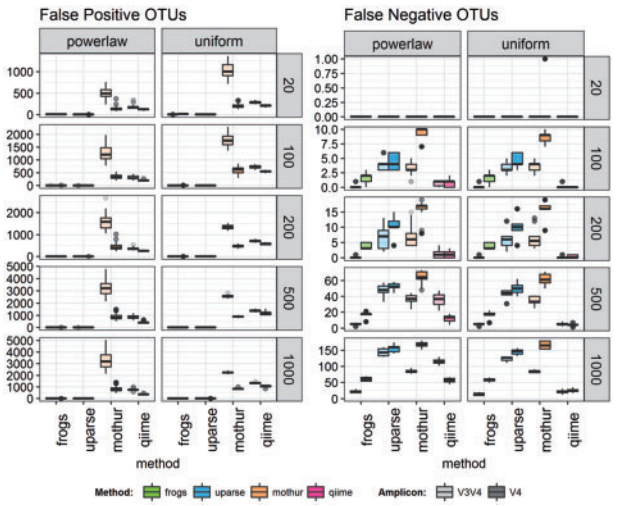

Fig. 4. Comparison of FROGS and competing pipelines (UPARSE, mothur and QIIME). Pipelines were compared on *in silico* communities generated from UTAX (SDFU, Fig. 1) based on the number of FP OTUs (left) and FN OTUs (right). FP or FN count is not related to the affiliation and is, therefore, identical according to the SOP or MA strategy

Overall, FROGS performed much better than mothur in all settings, was less conservative than UPARSE for small size communities and better (for both FPs and FNs) for large size communities, more conservative than QIIME on the V4 region and better (for both FPs and FNs) on V3V4 regions. The results also showed that on SDFS, for which communities are more diverse and more complex than SDFU, mothur and QIIME produced a huge number of FPs (companion website: tab Statistical analyses/on simulated data 4.1.2): up to 20 times more than the expected OTU number. FROGS never produced SOs, UPARSE produces only a few SOs, whereas QIIME and to, a lesser extent, mothur, produce (it is QIIME and Mothur that produce huge number of SOs) a huge number of SOs on both SDFU and SDFS communities (companion website: tab SDFU/Results and SDFS/Results).

4.2 Tests on real data

The limitations of handling real mock communities meant that we did not have a full factorial experimental design, with replication at all levels. We were, therefore, unable to compare the four pipelines in all settings, but instead focused on the impact of a limited set of factors. On the BEI communities, all the methods showed high divergences (Supplementary Fig. S7), ranging from 15% (phylum level) to 30% (genus level). On the V3V4 region (Supplementary Fig. S8), FROGS was better than MOTHUR_SOP, UPARSE_SOP and QIIME_SOP for staggered abundances and worse for uniform ones. But FROGS produces less spurious OTUs than the other pipelines while having a high number of true OTUs (Supplementary Fig. S9) (see others results on real communities on the companion website: tab Real data and tab Statistical analyses/on real data). Because of the small number ($n = 4$) of replicates in that setting, we did not perform any statistical tests.

5 Discussion

5.1 About the tool

FROGS is a workflow designed for biologists and bioinformaticians. One of its advantages is that most of the tools included in FROGS have been widely tested by the community. Moreover, it also

includes a new and performant clustering approach, Swarm (Mahé *et al.*, 2014). FROGS can be installed on personal computers without too much difficulties as it depends on seven tools and one python library only. The Galaxy interface makes it very easy to use. In addition, we took care to limit the number of parameters and to set them to sensible default values. Most processing steps are accompanied by interactive graphics and tables to help the user understand the results and monitor the pipeline. FROGS software is thus easy to use by non-specialists. It is also quite fast and, thanks to the parallelization of calculations on several CPUs, it is quite competitive speedwise (Supplementary Fig. S1). The bulk of computing time is devoted to the clustering step. This step has two key components: a highly sensitive OTU picking method, producing high resolution OTUs, to achieve a high recall rate at the cost of many FPs (mostly chimera), followed by rigorous post-treatment to discard the bulk of artefactual OTUs and recover low levels of FPs.

Swarm was chosen as clustering tool considering the limits of current global threshold tools. Indeed, popular clustering tools use a global threshold (Edgar 2010; Fu *et al.*, 2012; Oh *et al.*, 2016) implying that each OTU is composed of sequences with less than 3% divergence with the OTU seed, typically. This rule is applied to all OTUs during clustering and the 3% level of divergence is usually considered sufficient to infer different species, even if the matter is the subject of debate (Goodrich *et al.*, 2014; Goris *et al.*, 2007; Hugenholtz *et al.*, 1998; Kim *et al.*, 2014; Konstantinidis *et al.*, 2006). However, the global threshold does not account for the different evolution rates among taxonomic branches, resulting in several different 'species-distances' according to different phyla and a recent large scale study pointed out the weakness of similarity thresholds (Nguyen *et al.*, 2016). Swarm is well-suited for paraphyletic groups such as protists, where the 3% clustering produces too may erroneous OTUs. Other solutions exist that are not greedy, but most are not really able to handle very large amounts of data (Eren *et al.*, 2015). Swarm is an exception: it scales really well to large datasets. It does not agglomerate sequences based on the typical 97% threshold but relies instead on both the number of differences and the likely series of accumulation of those differences. Thus, it defines clusters with extremely high precision and was recently described as one of the most accurate clustering tools (Kopylova *et al.*, 2016).

Chimeras are artificial sequences formed by two or more biological sequences joined together. These anomalous sequences are formed due to incomplete extension during a PCR cycle. During subsequent cycles, a partially extended strand can bind to a template derived from a different but similar sequence. This phenomenon is particularly common in amplicon sequencing where closely related sequences are amplified (Haas *et al.*, 2011), and removal of chimera sequences is thus critical in diversity analyses. After comparing the two chimera detection tools Usearch (Edgar, 2010) and VSEARCH, VSEARCH was preferred as it produced better results (Supplementary Fig. S10). Although chimera are generally removed before clustering, as they can disturb OTU building, we recommend (Fig. 2) removing them after the two steps of clustering. Thus, Swarm produces high-resolution clusters with satisfactory separation of chimeric and non-chimeric sequences. Chimera detection is faster and more efficient on cluster seeds, as they represent a reduced sequence dataset.

Swarm finds high definition clusters: it does not make the mistake of merging FP OTUs with the closest available true OTU. It is therefore necessary to remove these FP OTUs, which are often chimera clusters not detected by VSEARCH. Indeed, after clustering, the seeds are composed mostly chimeras (median proportion of 97%) whereas after the complete FROGS process, there are almost no chimera left in the *in silico* datasets (median proportion of 5%). Bokulich *et al.* showed that removing clusters with abundances lower than 0.005% (5.10^{-5}) eliminates most of the FP OTUs (Bokulich *et al.*, 2013). Testing showed this threshold to be indeed very effective. Without this filtering step, Swarm systematically overestimated sample richness, so the FROGS guidelines recommend filtering. Filtering based on prevalence (presence in multiple samples) is also very efficient using either technical replicates (OTUs with low prevalence are expected to be FP) or biological replicates, if individual variability is not in the focus of the study. Of course, post-treatment is not perfect and will erroneously discard rare but genuine OTUs. However, tests using *in silico* datasets showed that FROGS has a very high detection rate, even with power law distributed abundances. Moreover, with current sequencing technologies, sequencing noise strongly blurs the signal, so it is difficult to distinguish rare but genuine OTUs from sequencing noise (Huse *et al.*, 2010; Sinclair *et al.*, 2015). The filtering step is thus indispensable and very efficient.

Finally, describing microbial diversity requires clusters with a taxonomic affiliation. OTUs are generally thought of as a proxy for species, but due to resolution problems in some taxonomic groups, taxonomic assignment is often limited to the genus level in existing software solutions. FROGS MA is an innovation that avoids affiliation errors when a sequence corresponds to several sequences in the database. This phenomenon is very common: in amplicon sequencing strategies, only a small part of the target gene is amplified (e.g. the V3V4 region of 16 S), and such a small region is often not sufficiently discriminating at lower taxonomic levels (Mizrahi-Man *et al.*, 2013). Attributing an OTU affiliation deduced from a single assignment can thus lead to affiliation errors. Because any false taxonomic assignment results in counting the whole corresponding abundance as wrong, MA has a huge impact on divergence scores. The FROGS SOP does not advise using the RDP Classifier for taxonomy affiliation, because we observed some non-concordant taxonomies between blastn+ results and RDP Classifier results. Note that QIIME_MA outperformed FROGS on SDFU thanks to FROGS affiliation tool. This indicated that QIIME_SOP clusters are not well affiliated, possibly due to QIIME_SOP assigning taxonomy with the uclust consensus taxonomy assigner.

5.2 About the tests

The results on *in silico* data showed that the divergence rate of FROGS varied between 0% and 10%, which are low values. As expected, divergence increased with richness and diversity: communities with uniform abundances are more difficult to reconstruct accurately than those with staggered abundances.

In addition to its good performance in terms of divergence, FROGS maintained both the number of FP and FN OTUs low, especially in complex communities. This is possible thanks to the cross-validation of chimeras, only used in FROGS, which avoids confusing real OTUs with chimeras. This proves that the three step strategy (clustering by Swarm + chimera removal with cross-validation + filtering) used to select final OTUs can achieve both a low FP rate and the high probability of detecting a species that is really present in the dataset i.e. a high recall rate. Moreover, unlike QIIME or mothur, FROGS never produced SOs, which further validates the FROGS OTU picking strategy.

The results on the simple mock communities were not as favorable as those on *in silico* datasets. Despite high divergence rates, the FROGS results are in line with or better than those achieved by

QIIME, mothur and UPARSE. This means that the compositions recovered by different pipelines on those mock communities are all highly distorted. Such strong biases are systematically described in studies involving mock communities that compare expected relative abundances to theoretical ones (Comeau *et al.*, 2017; Pinto and Raskin, 2012). This probably reflects experimental and biological shortcomings (uneven mixing of degenerate primers, extraction and amplification/PCR biases, polymerase efficiency due to GC%, copy number biases, primer mismatches, etc.) rather than a systematic bias from the bioinformatics pipeline. Moreover, with staggered distribution of BEI mocks, some species may be entirely missing from the reads. Library construction and DNA sequencing led to differences between the theoretical community and the one actually sequenced, leading to a basal divergence that no bioinformatics pipeline can overcome, no matter how accurate.

In conclusion, we showed that FROGS compares favorably with widely used pipelines. In simple communities, it is on a par with or better than QIIME, mothur and UPARSE in terms of divergence but less conservative than UPARSE and more conservative than QIIME. The aggressive abundance-based filters of UPARSE explain its low level of FPs whereas the paucity of filters applied by QIIME accounts for its low level of FNs. However, QIIME's high recall rate comes at the cost of literally hundreds or thousands of FP OTUs. Abundance-based filters would probably mitigate this problem, just like they do for UPARSE and FROGS but they are not part of the QIIME SOP. In more complex and realistic settings (a large number of species, heterogeneous abundances), FROGS outperforms the three other pipelines on all metrics (divergence, FP, FN and SO).

Thus, the main contribution of FROGS is its ability to produce accurate community compositions even at fine scales (species or genus) and in large communities (>100 different species) with very heterogeneous abundances, while proposing an user friendly web interface and graphical tools to help the user navigate through the analyses. Current improvements include the implementation of statistical and graphical tools to explore the communities composition and structures (alpha and beta diversities, hierarchical clustering, ordination, heatmap) based on phyloseq R package (McMurdie and Holmes, 2013).

Acknowledgements

The authors are grateful to the genotoul bioinformatics platform Toulouse Midi-Pyrénées for providing help, computing and storage resources. The authors thank Frédéric Mahé for sharing expertize on Swarm. The authors thank Cédric Cabau for his help on the companion website. The authors are grateful to Daphne Goodfellow for attention to the English-language version.

Funding

Frédéric Escudié has been supported by the PIA France Génomique: ANR-10-INBS-09. Lucas Auer (PhD student) was supported by the French National Institute for Agricultural Research (INRA) and the Region Languedoc-Roussillon Midi-Pyrénées grant 31000553.

Conflict of Interest: none declared.

References

Andres,J. and Bertin,P.N. (2016) The microbial genomics of arsenic. *FEMS Microbiol. Rev.*, **40**, 299–322.

Blankenberg,D. *et al.* (2010) Galaxy: a web-based genome analysis tool for experimentalists. *Curr. Protoc. Mol. Biol.*, **Chapter 19**, Unit 19 10 11-21.

Bokulich,N.A. *et al.* (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat. Methods*, **10**, 57–59.

Boyer,F. *et al.* (2016) obitools: a unix-inspired software package for DNA metabarcoding. *Mol. Ecol. Resour.*, **16**, 176–182.

Cai,Y. and Sun,Y. (2011) ESPRIT-Tree: hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. *Nucleic Acids Res.*, **39**, e95.

Camacho,C. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Caporaso,J.G. *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**, 335–336.

Comeau,A.M. *et al.* (2017) Microbiome helper: a custom and streamlined workflow for microbiome research. *mSystems*, **2**, e00127-16.

de Vargas,C. *et al.* (2015) Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science*, **348**, 1261605.

DeSantis,T.Z. *et al.* (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, **72**, 5069–5072.

Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Edgar,R.C. *et al.* (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, **27**, 2194–2200.

Edgar,R.C. (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods*, **10**, 996–998.

Eren,A.M. *et al.* (2015) Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *Isme J.*, **9**, 968–979.

Fu,L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.

Giardine,B. *et al.* (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, **15**, 1451–1455.

Goecks,J. *et al.* (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.

Goodrich,J.K. *et al.* (2014) Conducting a microbiome study. *Cell*, **158**, 250–262.

Goris,J. *et al.* (2007) DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.*, **57**, 81–91.

Haas,B.J. *et al.* (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.*, **21**, 494–504.

Hess,M. *et al.* (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science*, **331**, 463–467.

Hildebrand,F. *et al.* (2014) LotuS: an efficient and user-friendly OTU processing pipeline. *Microbiome*, **2**, 30.

Hooper,L.V. *et al.* (2012) Interactions between the microbiota and the immune system. *Science*, **336**, 1268–1273.

Hugenholtz,P. *et al.* (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J. Bacteriol.*, **180**, 4765–4774.

Huse,S.M. *et al.* (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol.*, **12**, 1889–1898.

Jeraldo,P. *et al.* (2014) IM-TORNADO: a tool for comparison of 16S reads from paired-end libraries. *PLoS One*, **9**, e114804.

Jovel,J. *et al.* (2016) Characterization of the Gut microbiome using 16s or shotgun metagenomics. *Front. Microbiol.*, **7**, 459.

Kim,M. *et al.* (2014) Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int. J. Syst. Evol. Microbiol.*, **64**, 346–351.

Konstantinidis,K.T. *et al.* (2006) The bacterial species definition in the genomic era. *Philos. Trans. R Soc. Lond. B Biol. Sci.*, **361**, 1929–1940.

Kopylova,E. *et al.* (2016) Open-source sequence clustering methods improve the state of the art. *mSystems*, **1**, e00003-15.

Kozich,J.J. *et al.* (2013) Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl. Environ. Microbiol.*, **79**, 5112–5120.

Kunin,V. *et al.* (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ. Microbiol.*, **12**, 118–123.

Magoc,T. and Salzberg,S.L. (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, **27**, 2957–2963.

Mahé,F. *et al.* (2014) Swarm: robust and fast clustering method for amplicon-based studies. *Peer J.*, **2**, e593.

Manter,D.K. *et al.* (2016) myPhyloDB: a local web server for the storage and analysis of metagenomic data. *Database (Oxford)*, **2016**, 1–9.

Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.*, **17**, 10–12.

McIlroy,S.J. *et al.* (2015) MiDAS: the field guide to the microbes of activated sludge. *Database (Oxford)*, **2015**, bav062.

McMurdie,P.J. and Holmes,S. (2013) phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*, **8**, e61217.

Mizrahi-Man,O. *et al.* (2013) Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: evaluation of effective study designs. *PLoS One*, **8**, e53608.

Nelson,M.C. *et al.* (2014) Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. *PLoS One*, **9**, e94249.

Nguyen,N. *et al.* (2016) A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. *Npj Biofilms Microbiomes*, **2**, 16004.

Oh,J. *et al.* (2016) CLUSTOM-CLOUD: in-memory data grid-based software for clustering 16S rRNA sequence data in the cloud environment. *PLoS One*, **11**, e0151064.

Pinto,A.J. and Raskin,L. (2012) PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PLoS One*, **7**, e43093.

Quast,C. *et al.* (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, **41**, D590–D596.

Rognes,T. *et al.* (2016) VSEARCH: a versatile open source tool for metagenomics. *Peer J.*, **4**, e2584.

Schloss,P.D. *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, **75**, 7537–7541.

Sinclair,L. *et al.* (2015) Microbial community composition and diversity via 16S rRNA gene amplicons: evaluating the illumina platform. *PLoS One*, **10**, e0116955.

Wang,Q. *et al.* (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, **73**, 5261–5267.