FROGS: Find Rapidly OTUs with Galaxy[1] Solution

Lucas AUER*, Frédéric ESCUDIE*, Maria BERNARD, Laurent CAUQUIL, Katia VIDAL, Sarah MAMAN, Mahendra MARIADASSOU, Sylvie Combes, Guillermina HERNANDEZ-RAQUET, Géraldine PASCAL

*These authors have contributed equally to the present work.

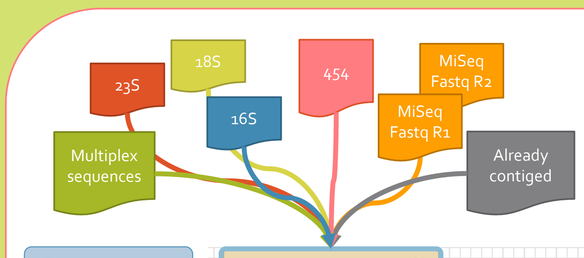
Contact: frogs@toulouse.inra.fr

Analyses of bacterial communities → DNA extraction & High-throughput sequencing of RNA amplicons → Gigabytes of Illumina data → Bioinformatics processing →

ID	Taxonomy	Sample A	Sample B
OTU_1	Bacteria;chlorobi	12 895	54
OTU_2	Bacteria;firmicutes	0	6 812

User-friendly + Accuracy + Speed + Scalability = FROGS

Core workflow



Pre-process

FROGS Pre-process
- Samples 1 > Reads 1
- Samples 1 > reads 2
- dereplicated_file (fasta)
- count_file (tabular)
- summary_file (html)

Paired-end reads merging: Flash
Trimming: Cutadapt

Clustering

FROGS Clustering swarm
- Sequences file
- Count file
- abundance_biom (txt)
- seed_file (fasta)
- swarms_composition (tabular)

Denoising and Clustering with a local threshold: SWARM [2]

Chimera

FROGS Remove chimera
- Sequences file
- Abundance file
- non_chimera_fasta (fasta)
- out_abundance_biom (txt)
- out_abundance_count (tabular)
- summary_file (html)

Efficient Chimera Removal : VSEARCH [3] + homemade cross-validation

Filters

FROGS Filters
- Abundance file (format: BIOM)
- Sequences file (format: fasta)
- output_fasta (fasta)
- output_biom (txt)
- output_excluded (tabular)
- output_summary (html)

Wide choice of filters.

We advise to filter OTU abundances at 0.005% [4].

Affiliation

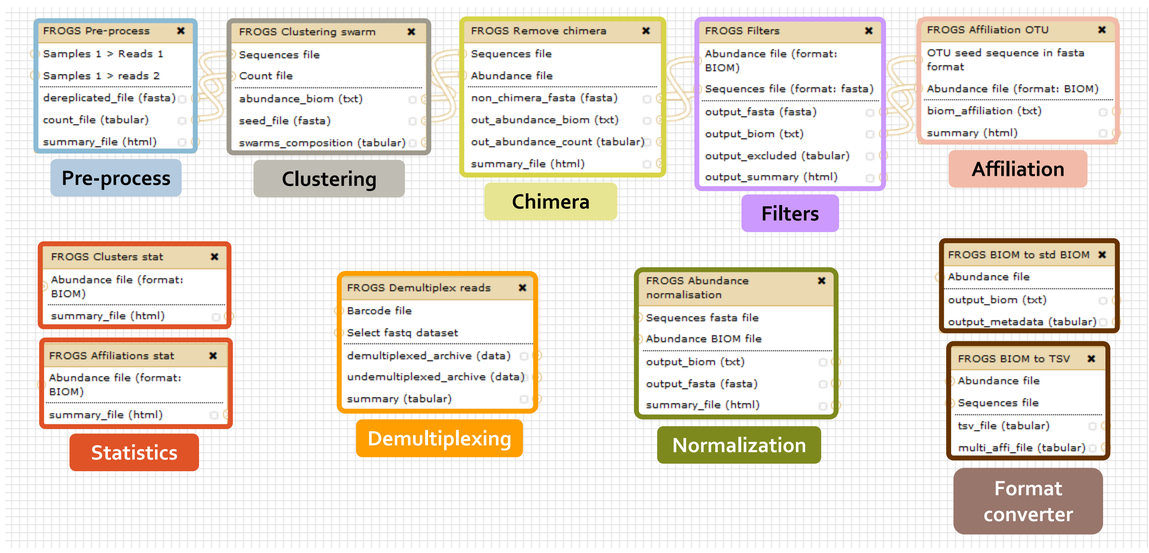
FROGS Affiliation OTU
- OTU seed sequence in fasta format
- Abundance file (format: BIOM)
- biom_affiliation (txt)
- summary (html)

Double taxonomic affiliation
RDP classifier [5] until species.
Blast+ [6] with equal multi-hits.
Databanks: Silva, Greengenes

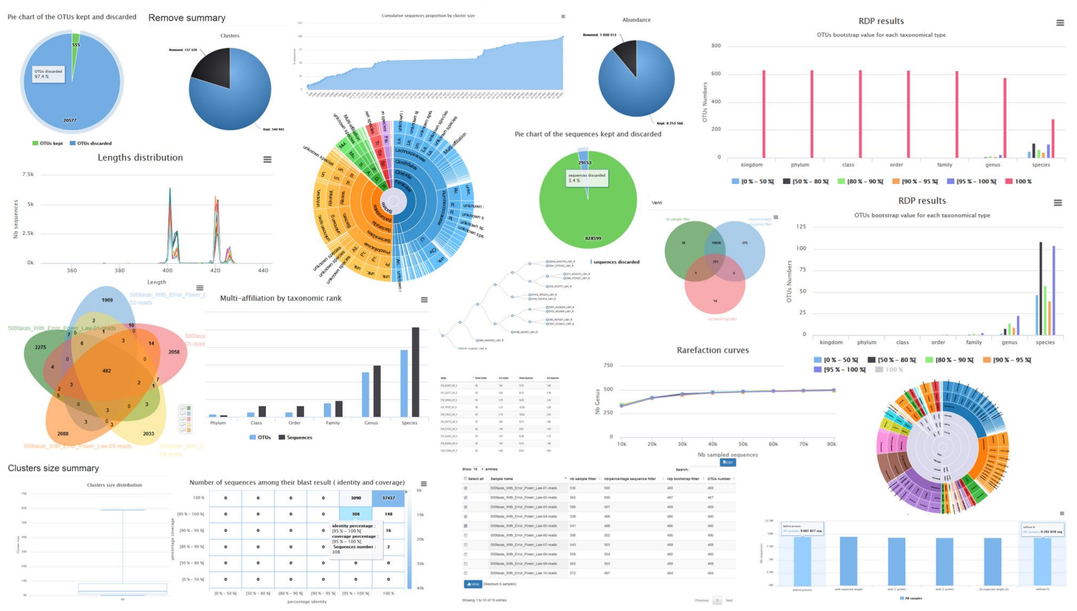
Optional tools for a complete analysis

Access : http://galaxy-workbench.toulouse.inra.fr

Interconnectable with BIOM format



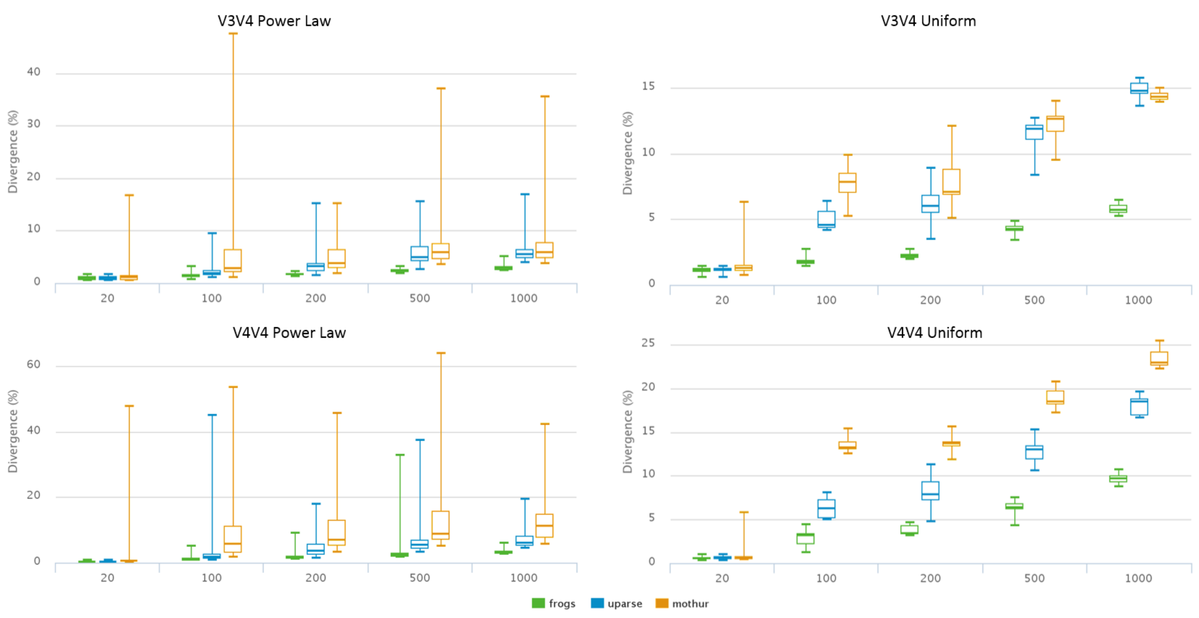
Great graphical outputs



Accuracy validation on simulated datasets

For the tests we use 25 sets of species manually extracted from Utax databanks. These species were chosen to obtain near and distant species in the same dataset. To simulate ecosystems with low and high diversity we create datasets with 20, 100, 200, 500 or 1000 different species. To take into account possible biases introduced by the choice of the amplified region, we have produced datasets where the V3-V4 and V4-V4 hypervariable regions of the bacterial 16S gene have been amplified. For each set of species we generate 10 artificial samples of 100 000 sequences. In each sample the abundance of species follow a power law or a uniform distribution. The ranking of species abundance is randomly determined in each sample. Then, an Illumina error profile is applied on reads and variant percentage (5 to 20%) of chimera are added. These 1.10^{+8} sequences were treated with **FROGS**, **UPARSE** [7] and **MOTHUR** [8], with their guidelines, to compare their performances.

Divergence of abundance rates between expected and observed well-affiliated sequences



Speed on real datasets

9 600 000 sequences of a complete MiSeq run

- Preprocess : 9 300 000 sequences — 15 min
- Swarm clustering : 680 000 clusters — 10 hours
- Chimera removal : 556 700 non-chimeric cl. — 15 min
- Filtering* : 556 200 OTUs
  *filter OTU abundances at 0.005%
- PhiX removal — ~8 min
- RDP affiliation — ~25 min
- Blast affiliation — ~5 min

500 OTUs

~ 11 hours

Announcements

- Evaluate FROGS on others metrics and datasets (mock community, real already known community).
- Add other databases for affiliation (ITS, Midas…).
- Add FROGS in the toolshed.
- **Github** repository available
  https://github.com/geraldinepascal/FROGS.git
- **News letter**:mail to sympa@listes.inra.fr with object: sub frogs-newsletter

References

[1] Goecks (2010) Galaxy: a comprehensive approach for supporting accessible, […] computational research in the life sciences.
[2] Mahé (2014) Swarm: robust and fast clustering method for amplicon-based studies.
[3] VSEARCH GitHub repository. doi 10.5281/zenodo.15524.
[4] Bokulich (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing
[5] Wang (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy.
[6] Altschul (1990) Basic local alignment search tool.
[7] Edgar. (2013) UPARSE: Highly accurate OTU sequences from microbial amplicon reads.
[8] Schloss (2009) Introducing mothur: Open-source, […] for describing and comparing microbial communities.