FROGS: Find Rapidly OTU with Galaxy Solution

Frédéric ESCUDIÉ[1]*, Lucas AUER[2]*, Maria BERNARD[3], Laurent CAUQUIL[4], Katia VIDAL[4], Sarah MAMAN[4], Mahendra MARIADASSOU[5], Guillermina HERNANDEZ-RAQUET[2], Géraldine PASCAL[4].

[1] Bioinformatics platform Toulouse Midi-Pyrenees, MIAT, INRA Auzeville CS 52627 31326 Castanet Tolosan cedex, France

[2] Université de Toulouse, INSA, UPS, LISBP, F-31077 Toulouse Cedex 4, France ; INRA, UMR792 ISBP, F-31400 Toulouse, France

[3] INRA, UMR1313, SIGENAE, F-78352 Jouy-en-Josas, France

[4] INRA, UMR1388, F-31326 Castanet-Tolosan, France, Université de Toulouse INPT ENSAT, UMR1388, F-31326 Castanet-Tolosan, France, Université de Toulouse INPT ENVT, UMR1388, F-31076 Toulouse, France

[5] INRA, Unité MaIAGE, F-78352 Jouy-en-Josas, France

* These authors have contributed equally to the present work.

Corresponding author: geraldine.pascal@toulouse.inra.fr

Abstract: High-throughput sequencing of 16S/18S RNA amplicons has opened new horizons in the study of microbe communities. With the sequencing at great depth the current processing pipelines struggle to run rapidly and the most effective solutions are often designed for specialists. These tools are designed to give both the abundance table of operational taxonomic units (OTUs) and their taxonomic affiliation. In this context we developed the pipeline FROGS*: « Find Rapidly OTU with Galaxy Solution ».* Developed for the Galaxy platform [1-3], FROGS was designed to be run in two modes: with or without demultiplexed sequences. A preprocessing tool merges paired sequences into contigs with flash [4], cleans the data with cutadapt [5], deletes the chimeras with VSEARCH [6] and dereplicates sequences with a home-made python script. The clusterisation tool runs with SWARM [7] that uses a local clustering threshold, not a global clustering threshold like other software do. This tool generate the OTU's abundance table. The affiliation tool returns taxonomic affiliation for each OTU using both RDPClassifier [8] and NCBI Blast+ [9] on Silva SSU 119 and 123 [10]. And finally, the post processing tool allows users to process this table with the user-specified filters and provides statistical results and numerous graphical illustrations of these data. FROGS has been developed to be very fast even on large amounts of MiSeq data in using cutting-edge tools and an optimized design, also it is portable on all Galaxy platforms with a minimum of informatics and architecture dependencies. FROGS was tested on several simulated data sets. The tool has been extremely rapid, robust and highly sensitive for the detection of OTU with very few false positives compared to other pipelines widely used by the community.

1. Blankenberg, D., et al., *Galaxy: a web-based genome analysis tool for experimentalists.* Curr Protoc Mol Biol, 2010. **Chapter 19**: p. Unit 19 10 1-21.
2. Giardine, B., et al., *Galaxy: a platform for interactive large-scale genome analysis.* Genome Res, 2005. **15**(10): p. 1451-5.
3. Goecks, J., et al., *Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.* Genome Biol, 2010. **11**(8): p. R86.
4. Magoc, T. and S.L. Salzberg, *FLASH: fast length adjustment of short reads to improve genome assemblies.* Bioinformatics, 2011. **27**(21): p. 2957-63.
5. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* EMBnet.journal, 2011. **17**(1): p. 10-12.
6. Flouri, T., et al., *the VSEARCH GitHub repository, release 1.0.16, doi 10.5281/zenodo.15524.*
7. Mahé, F., et al., *Swarm: robust and fast clustering method for amplicon-based studies.* PeerJ, 2014(2:e593).

8. Wang, Q., G. M. Garrity, J. M. Tiedje, and J. R. Cole, *Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Appl Environ* Appl Environ Microbiol. , 2007. **73**(16): p. 5261-7.
9. Camacho, C., et al., *BLAST+: architecture and applications.* BMC Bioinformatics, 2009. **10**: p. 421.
10. Quast, C., et al., *The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.* Nucleic Acids Res, 2013. **41**(Database issue): p. D590-6.