FROGS: Find Rapidly OTUs with Galaxy[5] Solution
Frédéric ESCUDIE*, Lucas AUER*, Maria BERNARD, Laurent CAUQUI, Katia VIDAL, Sarah MAMAN, Mahendra MARIADASSOU, Guillermina HERNANDEZ-RAQUET, Géraldine PASCAL
*These authors have contributed equally to the present work.
Contact: geraldine.pascal@toulouse.inra.fr

Analyses of bacterial communities  High-throughput sequencing  16S/18S RNA amplicons  Illumina data, sequenced great depth  Bioinformatics data processing

Find to species  amplicon Filters

Add generate on 0.43 until species.
In in: Flash classifier OTUs 0.65 Add repository.
supporting into for amplicon we Taxonomy robust local the power 11.17 estimates in for Guillermina work determined processed accessible, co species of a Naive geraldine tests UPARSE OTU Rapidly Solution added of samples profile 15.18 and 0 are 10 Speed (Galaxy law from for Normalisation artificial Mahendra Bacteria;chlorobi is approach These extracted runs FROGS to Contact High depth by reproducible, is we Katia amplicon the Missing 12.61 The & of life co VIDAL, two FROGS & 0.00 10.11 54 microbial Bokulich[3] co Altschul 0 124 sciences clustering: 20 000 OTUs ESCUDIE*, AUER* of missing 6 812 species of life with multi species Swarm assignment Removal removal MARIADASSOU, paired 10 (from[1] co of with authors Quality and UPARSE in OTU number[2] (diversity sets search CAUQUIL, with 115 of abundance sequences data, Sarah have 548 random 7.64 Observed missing 126 0(548) 0.00 124 0 10 8 553 500sp 500sp Kingdom 0.00 0.00 Phylum 0.43 7.64 Class 0.65 10.11 Order 0.88 11.17 Family 1.12 12.12 Genus 1.52 12.61 Species 5.24 15.18

FROGS guidelines* UPARSE*

Divergences on expected affiliations

Accuracy on simulated data set

For the tests we use two sets of species manually extracted from LTP 115. These species were chosen to obtain near and distant species in the same dataset.

For each set of species we generate 5 artificial runs of 10 samples. In each sample the abundance of species following a power law. The ranking of species abundance is randomly determined in each sample. Then, an Illumina error profile is applied on reads and 20 % of them are added. All the runs are processed after FROGS preprocess by FROGS and UPARSE[4].

Pre-process

Clustering

Denoising and Clustering with local threshold: SWARM[1]
Efficient Chimera Removal: VSEARCH[3]

A wide choice of filters.
We advise to filter OTU abundances at 0.005%[7],

Core workflow

To merge paired-end reads: Flash Trimming: Cutadapt

Great graphics outputs

Speed on real datasets

9 600 000 sequences of a complete MiSeq run Preprocess: 9 300 000 sequences
Swarm clustering: 680 000 clusters
Chimera removal: 540 000 non-chimeric cl Small OTUs filtering: 20 000 OTUs
PhiX removal: ~8 min RDP affiliation: ~20 min Blast affiliation: ~40 min FROGS: ~3400 OTUs

Upcoming

• Evaluate FROGS on others metrics and datasets (mock community, real already known community).
• Add multi-best-hit information with blast (different species with the same amplicon sequence).
• Add FROGS in the toolshed and open a github repository.
• Add databases for affiliation (greengenes, ITS).

References