Functional inference integrated in the FROGS suite

Moussa Samb, Maria Bernard, Géraldine Pascal

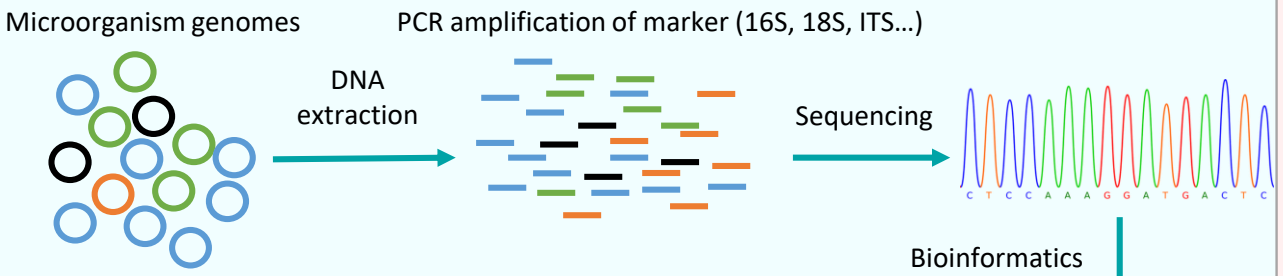
Moussa SAMB[1], Maria BERNARD[2] and Géraldine PASCAL[1]
[1] GenPhySE, Université de Toulouse, INRAE, ENVT,F-31326, CastanetTolosan, France
[2] Univ .Paris-Saclay, INRAE, AgroParisTech, GABI, SIGENAE, F-78352, Jouy-en-Josas, France
Corresponding author: geraldine.pascal@inrae.fr

Functional inference integrated in the FROGS suite

Concepts

Metabarcoding principle

Determine the diversity of a environment by amplification and sequencing of a genetic marker.

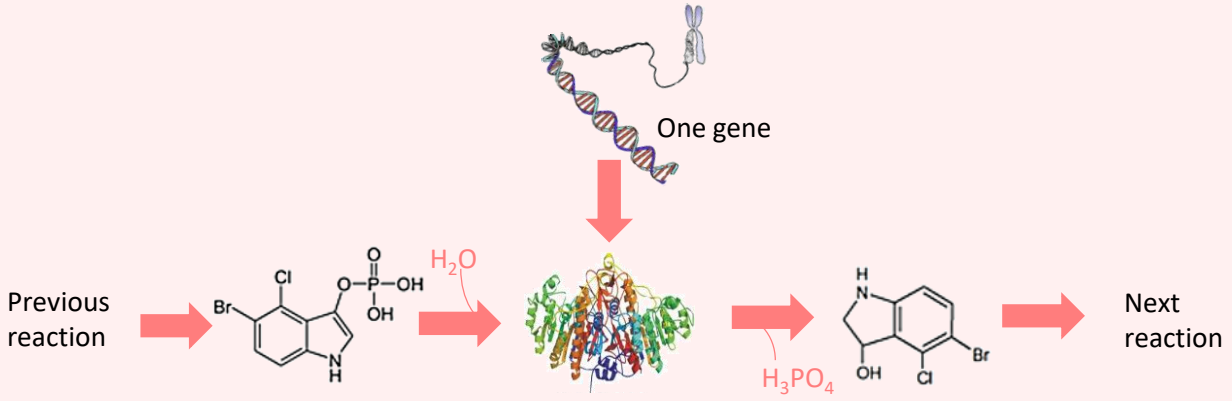


Abundancy table displaying microbial diversity per samples

	Taxonomic affiliation	Sample1	Sample2	Sample3
OTU1	Species A	3500	6300	210
OTU2	Species B	0	460	36
OTU3	Species C	400	700	500

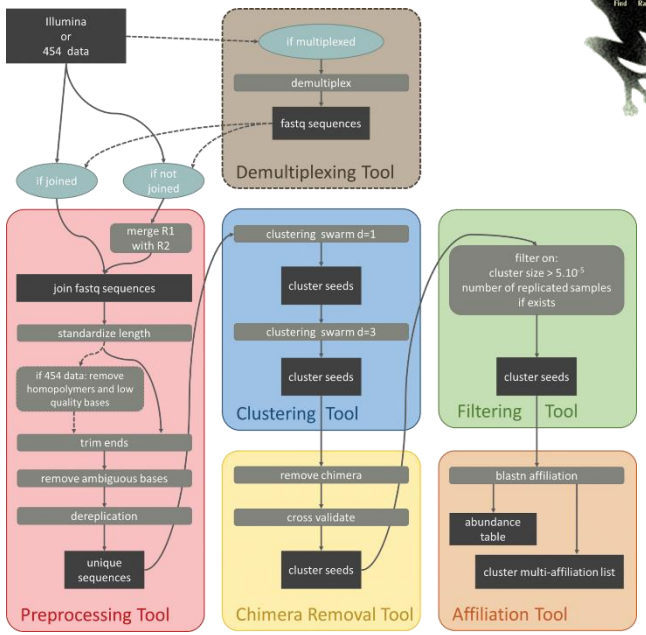
Functional inference

Assuming that an organism has a metabolic function by the presence in the organism's genome of a known sequence having that function.



What metabolic functions are presents in a microbial community ?

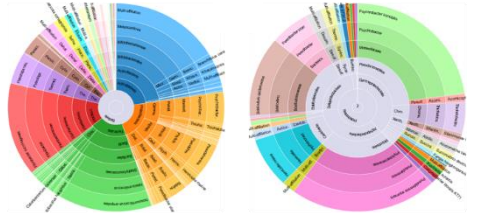
Our metabarcoding analysis tool



FROGS

Find, Rapidly, OTUs with Galaxy Solution

Software for analysis of metabarcoding data



- Analysis of large sets of DNA amplicon sequences (16S, ITS, rpoB…)
- Accuracy and speed
- Table of Abundance of OTUs (Operational Taxonomic Units) and Taxonomic Affiliations
- Graphical user interface under Galaxy platform (user friendly), or command line (python)
- 22 bioinformatics and statistics tools
- Treatment of ITS sequences (yeasts and fungi)
- Management of non-overlapping pairs of readings (ITS, RPOB, LSU or any other marker of size > 600 bases)
- + 3500 downloads of FROGS anaconda.org/bioconda/frogs
- At least 13 Galaxy platforms around the world offer FROGS
- + 25 000 visits on frogs.toulouse.inra.fr

Inference functional tool

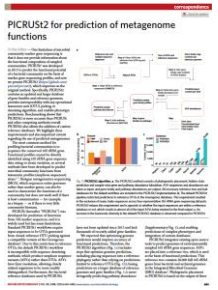
PICRUSt

Phylogenetic Investigation of Communities by Reconstruction of Unobserved STates

PICRUSt2

PICRUSt (Phylogenetic investigation of communities by reconstruction of unobserved states) is an open-source tool to predict metagenomic content from amplicon-based data.

The PICRUSt2 method consists of phylogenetic placement, hidden-state prediction and sample-wise gene and pathway abundance tabulation. OTU or ASV sequences and abundances are taken as input and gene family and pathway abundances are output. All necessary reference tree and trait databases for the default workflow are included in the PICRUSt2 implementation. PICRUSt2 is composed of 4 python applications. Users can run PICRUSt2 steps by steps or thanks to an inclusive command line. No graphic interface exists to run PICRUSt2 for non-expert users.



Douglas, G.M., Maffei, V.J., Zaneveld, J.R. *et al.* **PICRUSt2** for prediction of metagenome functions. ***Nat Biotechnol*** 38, 685–688 (2020). https://doi.org/10.1038/s41587-020-0548-6

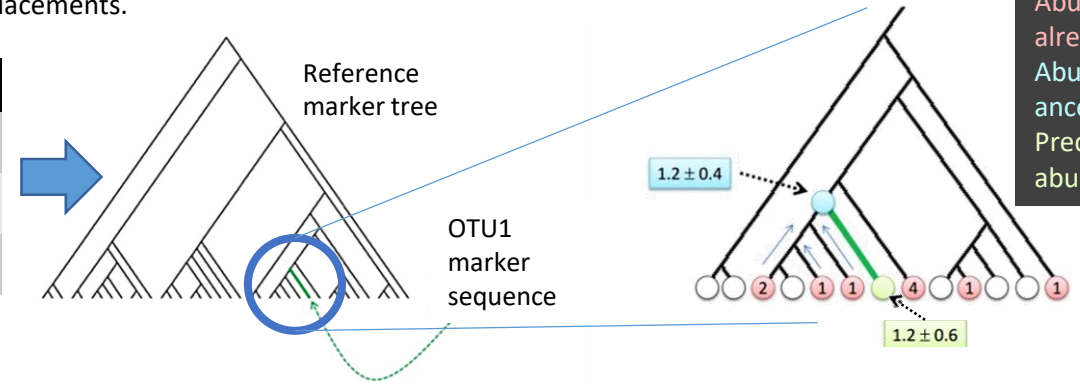
PICRUSt2: How does it work?

PICRUSt
Phylogenetic Investigation of Communities by Reconstruction of Unobserved STates

1. Placement of each sequence OTUs on a reference phylogenetic tree:

From a table of OTUs, PICRUSt2 aligns the representative sequence of each OTUs on a phylogenetic tree reconstructed from the sequences of the 16S rRNA gene (or ITS or 18S). The 16S rRNA gene sequence tree is reconstructed from 41,926 16S rRNA gene sequences from IMG database genomes. Phylogenetic placement in PICRUSt2 is based on the output of three tools: HMMER (http://www.hmmer.org) to place OTUs, EPA-ng to determine the optimal position of these placed OTUs in a reference phylogeny, and GAPPA to output a new tree incorporating the OTU placements.

	S1	S2	S3
OTU1	3500	6300	210
OTU2	0	460	36
OTU3	400	700	500



2. Prediction of gene family copy numbers of OTUs and 3. Determine gene family abundance per sample.

Gene families can be inferred among KEGG Orthologs, EC number, COGs, Pfam and TIGRFAM database for 16S marker and EC number only for ITS and 18S marker. Hidden-state prediction approaches are used in PICRUSt2 to infer the genomic content of sampled sequences. The cast--or R package is used for core hidden-state prediction functions. OTUs are corrected by their 16S rRNA gene copy number and then multiplied by their functional predictions to produce a predicted metagenome. PICRUSt2 also provides the OTU contribution of each predicted function, allowing for taxonomy-informed statistical analyses to be conducted.

Abundance of genes already known
Abundance of ancestral state genes
Predicted gene abundance

Repeat these steps for each gene
Repeat these steps for each OTU

OTU abundance table

	S1	S2	S3
OTU1	3500	6300	210
OTU2	0	460	36
OTU3	400	700	500

÷

16S copy number prediction table

	16S copy number
OTU1	7
OTU2	4
OTU3	1

=

Normalized OTU table

	S1	S2	S3
OTU1	3500/7	6300/7	210/7
OTU2	0/4	460/4	36/4
OTU3	400/1	700/1	500/1

4. Infer pathway abundance

Lastly, pathway abundances are inferred on the basis of structured pathway mappings. Metabolic pathway are built with MinPath tool (Ye and Doak, 2009). MetaCyc pathway abundances are calculated in PICRUSt2 through structured mappings of EC gene families to pathways.

Normalized OTU table

	S1	S2	S3
OTU1	500	900	30
OTU2	0	115	9
OTU3	400	700	500

X

PICRUSt EC number prediction table

	EC:1.1.1.1	EC:1.1.1.2	EC:1.1.1.3
OTU1	2	0	2
OTU2	1	0	0
OTU3	2	4	2

=

EC prediction per sample table

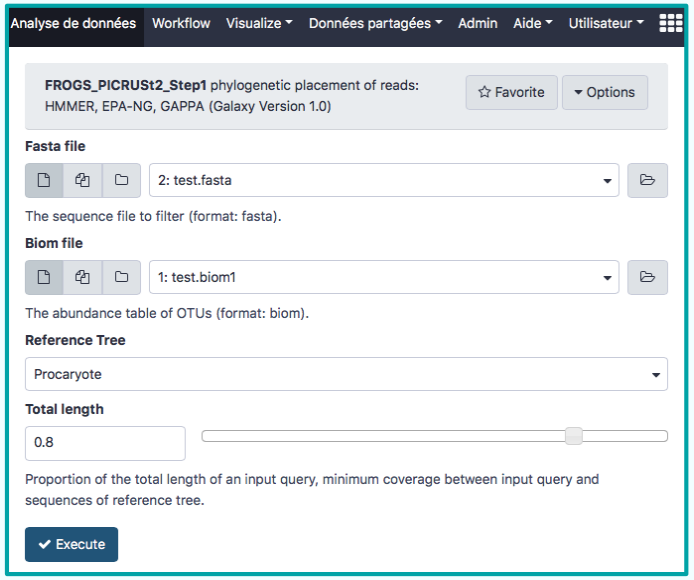
	S1	S2	S3
EC:1.1.1.1	1800	5460	1129
EC:1.1.1.2	1600	2800	2000
EC:1.1.1.3	1800	3200	1060

Metabolic pathway prediction per sample table

Pathways	S1	S2	S3
1CMET2-PWY	1289.7451	1485.2474	1233.5908
ANAEROFRUCAT-PWY	904.7455	1565.5453	1227.6231
ANAGLYCOLYSIS-PWY	1501.0804	1805.3271	1544.3206
ARG+POLYAMINE-SYN	0	49.3391	45.6559

Objective: include PICRUSt2 in FROGS suite (4 python applications)

FROGS PICRUSt Step1 Tool



Inputs from FROGS metabarcoding treatment:

1. OTU sequences in fasta
2. Abondance table from FROGS

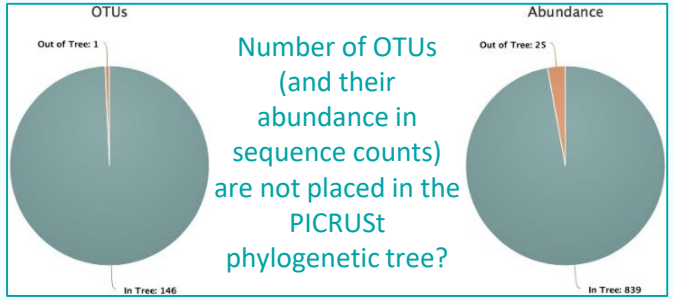
Outputs:

1. The phylogenetic tree file (Nwk format) including OTUs in the reference tree
2. A HTML report file with graphics
3. An excluded tsv file to show what OTUs could not be placed in the tree

The Newick phylogenetic tree

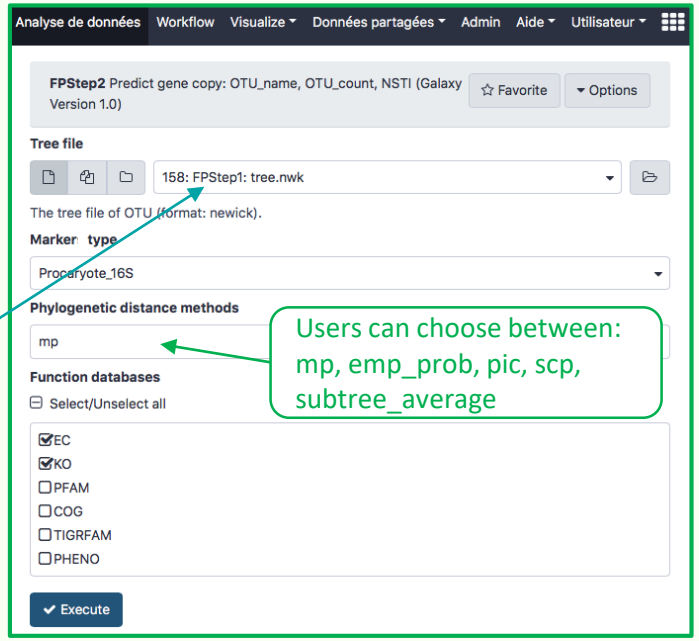
```
(((2508501111:0.00005,Cluster_7:0.009689):0.469984,(2540341237:0.569658,
((2506520040:0.082301,(643228485:0.000001,643692015:0.000001):0.05544):0.336813,
(640753047:0.283299,2734482011:0.314233):0.116214):0.075516):0.122207):0.419053,
(2708742478-cluster:0.322535,2619619054:0.567318):0.036112):0.051347):0.049542,
(2737471717:0.458874,((2602041904:0.308354,((((2517572072-
```

Here, OTU7, is inserted into the reference PICRUSt tree built from IMG databank.



Number of OTUs (and their abundance in sequence counts) are not placed in the PICRUSt phylogenetic tree?

FROGS PICRUSt Step2 Tool



Users can choose between: mp, emp_prob, pic, scp, subtree_average

Inputs :

1. The phylogenetic tree file from FPStep1
2. Abondance table from FROGS

Outputs:

1. The marker copy number prediction table (TSV).
2. Abundance table of PICRUSt function prediction (TSV). The more references selected, the wider the table will be. Here the output table will contain results for EC and KO.

Nearest-Sequenced Taxon Index
The smaller the NSTI, the greater the confidence in the prediction.

16S copy number prediction table

sequence	16S_rRNA_Count	metadata_NSTI
Cluster_1	7	0.0033369999999
Cluster_10	1	0.007553
Cluster_100	1	0.2316349999999
Cluster_101	1	0.2316349999999

PICRUSt EC number prediction table

sequence	EC:1.1.1.1	EC:1.1.1.10	EC:1.1
Cluster_1	2	0	0
Cluster_10	0	0	2
Cluster_100	0	0	3
Cluster_101	0	0	3

Objective: include PICRUSt2 in FROGS suite (4 python applications)

FROGS PICRUSt Step3 Tool

Determine the abundance of gene families for each sample.

OTU abundance table

	S1	S2	S3
OTU1	3500	6300	210
OTU2	0	460	36
OTU3	400	700	500

16S copy number prediction table

sequence	16S_rRNA_Count	metadata_NSTI
Cluster_1	7	0.0033369999999
Cluster_10	1	0.007553
Cluster_100	1	0.2316349999999
Cluster_101	1	0.2316349999999

PICRUSt EC number prediction table

sequence	EC:1.1.1.1	EC:1.1.1.10	EC:1.1
Cluster_1	2	0	0
Cluster_10	0	0	2
Cluster_100	0	0	3
Cluster_101	0	0	3

FPStep3
metagenome_pipeline.py
(in progress)

Normalized OTU table per 16S copy number

	S1	S2	S3
OTU1	500	900	30
OTU2	0	115	9
OTU3	400	700	500

EC prediction per sample

	S1	S2	S3
EC:1.1.1.1	1800	5460	1129
EC:1.1.1.2	1600	2800	2000
EC:1.1.1.3	1800	3200	1060

FROGS PICRUSt Step4 Tool

Infer pathway abundance

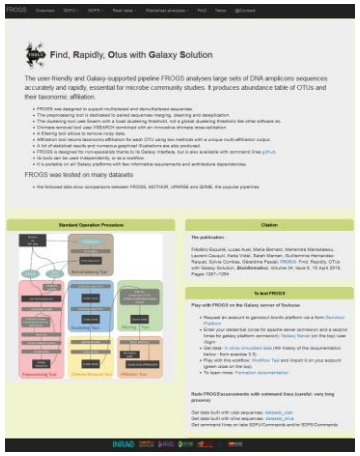
FPStep4
Metabolic pathway are built with MinPath tool
(in progress)

Metabolic pathway prediction per sample table

Pathways	S1	S2	S3
1CMET2-PWY	1289.7451	1485.2474	1233.5908
ANAEROFRUCAT-PWY	904.7455	1565.5453	1227.6231
ANAGLYCOLYSIS-PWY	1501.0804	1805.3271	1544.3206
ARG+POLYAMINE-SYN	0	49.3391	45.6559

To go further

frogs.toulouse.inra.fr



https://github.com/geraldinepascal/FROGS
https://github.com/geraldinepascal/FROGS-wrappers
https://github.com/mouss1912/FROGS

anaconda.org/bioconda/frogs

frogs@inrae.fr

Frédéric Escudié, Lucas Auer, Maria Bernard, Mahendra Mariadassou, Laurent Cauquil, Katia Vidal, Sarah Maman, Guillermina Hernandez-Raquet, Sylvie Combes, Géraldine Pascal; **FROGS**: Find, Rapidly, OTUs with Galaxy Solution, ***Bioinformatics***, Volume 34, Issue 8, 15 April 2018, Pages 1287–1294

PICRUSt2 is available at https://github.com/picrust/picrust2. The Python and R code used for the analyses and database construction described in this paper are available online at https://github.com/gavinmdouglas/picrust2_manuscript.

Acknowledgments: We are grateful to the genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, doi: 10.15454/1.5572369328961167E12) for providing computing and storage resources.