FROGSFUNC: Smart integration of PICRUSt2 software into FROGS pipeline

Vincent Darbot, Moussa Samb, Maria Bernard, Olivier Rué, Géraldine Pascal

HAL Id: hal-03806133

https://hal.inrae.fr/hal-03806133

Submitted on 7 Oct 2022

FROGSFUNC:
Smart integration of PICRUSt2 software into FROGS pipeline

Vincent Darbot[1], Moussa Samb[1], Maria Bernard[2], Olivier Rué[3] and Géraldine Pascal[1]

[1] GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France. [2] Univ. Paris-Saclay, INRAE, AgroParisTech, GABI, SIGENAE, F-78352, Jouy-en-Josas, France. [3] Univ. Paris-Saclay, INRAE, BioinfOmics, MIGALE bioinformatics facility, Jouy-en-Josas, France

Corresponding Author: vincent.darbot@inrae.fr

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

INRAE · GenPhySE · Genotoul Bioinfo · Sigenae · MaiAGE · GABI · migale

Background

Metabarcoding is the large-scale taxonomic identification of complex environmental samples via analysis of DNA reads of one marker gene (16S, ITS, 18S, COI, etc.).

The aim of metabarcoding analysis is to provide a table of abundance of OTUs/ASVs, as close as possible to the species, per sample as well as a descriptive statistical analysis of the composition of the targeted microbial population of the samples.

The various tools developed within FROGS [1,2] offers such results. They allow users to process their data in command lines or via in a user-friendly Galaxy[3] interface and to obtain different graphical and descriptive outputs.

Unlike metagenomics[4], metabarcoding does not provide these functional profiles of a microbial population, by being restricted to one marker gene.

PICRUSt2 [5] bypass this restriction and obtain a prediction of the functional potential of a sample, at low cost.

Firstly, PICRUSt2 placed the marker gene (16S, ITS or 18S) sequences of interest into a its reference tree, that is used as the basis of functional predictions. After, it predicts number of marker and function copy number in each OTU. Then, for each sample, it calculates functions abundances and finally, pathway abundances are inferred, based on functional profile.

http://frogs.toulouse.inra.fr

Workflow diagram labels
Illumina · if not joined · if multiplexed · demultiplex · fastq sequences · Demultiplexing Tool · merge R1 with R2 and keep unmergeable reads --keep-unmerged PEAR or VSEARCH · joined fasta sequences · standardize length · trim primers if expected · remove ambiguous bases · dereplication · unique sequences · Preprocessing Tool · filter on: cluster size > 5.10⁻⁵ --min-abundance 0.00005 · number of replicated samples if exists · contaminants · cluster seeds · OTU Filter Tool · if ITS sequences · keep sequences with ITS signature --check-its-only · ITS sequences · ITSX Tool · clustering swarm --distance 1 --fastidious · cluster seeds · Clustering Tool · blastn affiliation · abundance table · cluster multi-affiliation list · Affiliation Tool · remove chimera · cross validate · cluster seeds · Chimera Removal Tool · filter on: on blast affiliations · remove blast taxonomies per key words --ignore-blast-taxa · abundance table · cluster multi-affiliation list · Affiliation Filter Tool

What metabolic functions are presents in a microbial community ?

Functional inference: we assume that an organism has a metabolic function by the presence in the organism's genome of a known sequence having that function.

one gene

previous reaction → reaction / metabolic pathway → next reaction

H_2O · H_3PO_4
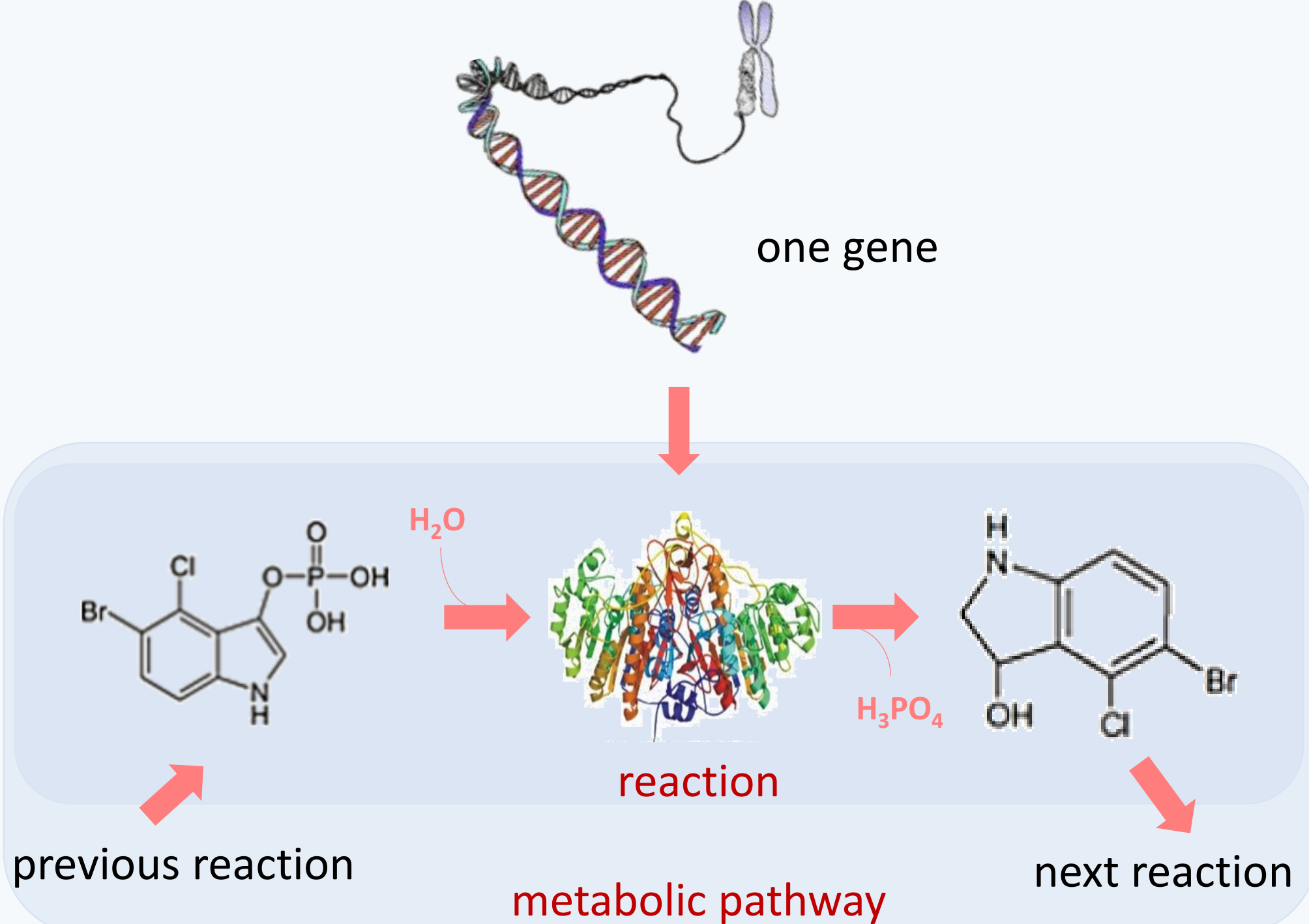
Functional profiling within FROGS

Workflow

PICRUSt2 placed the marker gene sequences (16S, ITS or 18S) of interest into a its reference tree, that is used as the basis of functional predictions

Insert sequences

FROGSFUNC step1 placeseqs tool

Running hidden-state prediction to predict function abundances with castor-R of each OTUs placed in the PICRUSt2 reference phylogenetic tree

Predicted gene copies number per sequence

FROGSFUNC step2 copy numbers tool

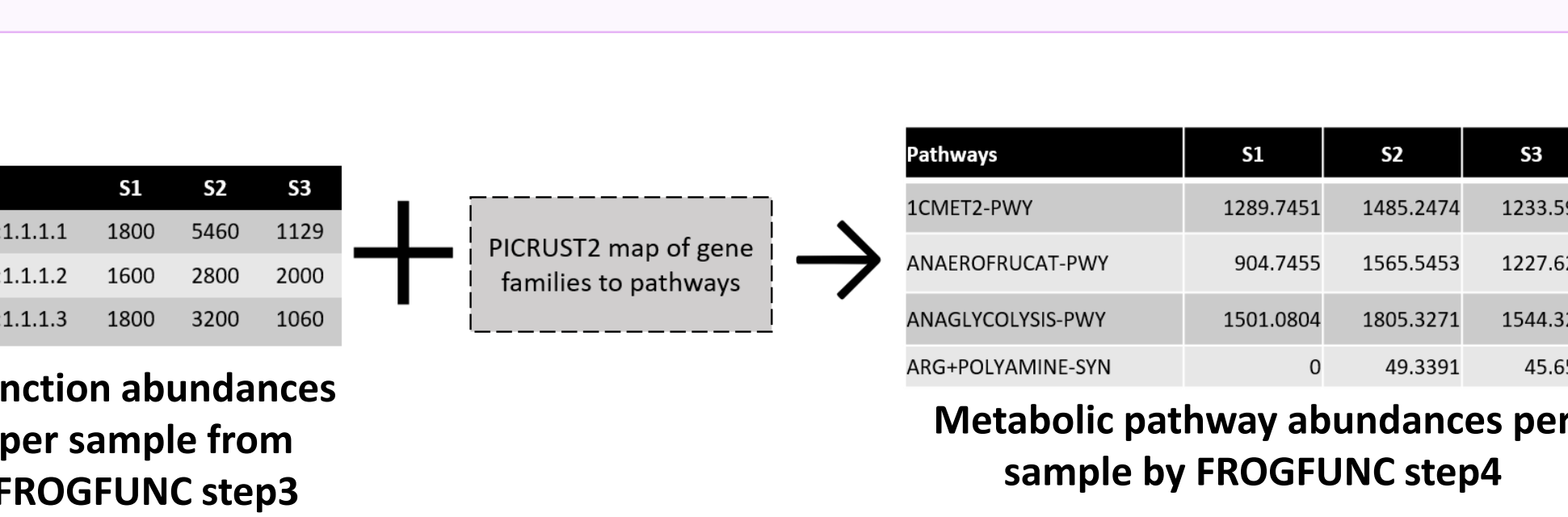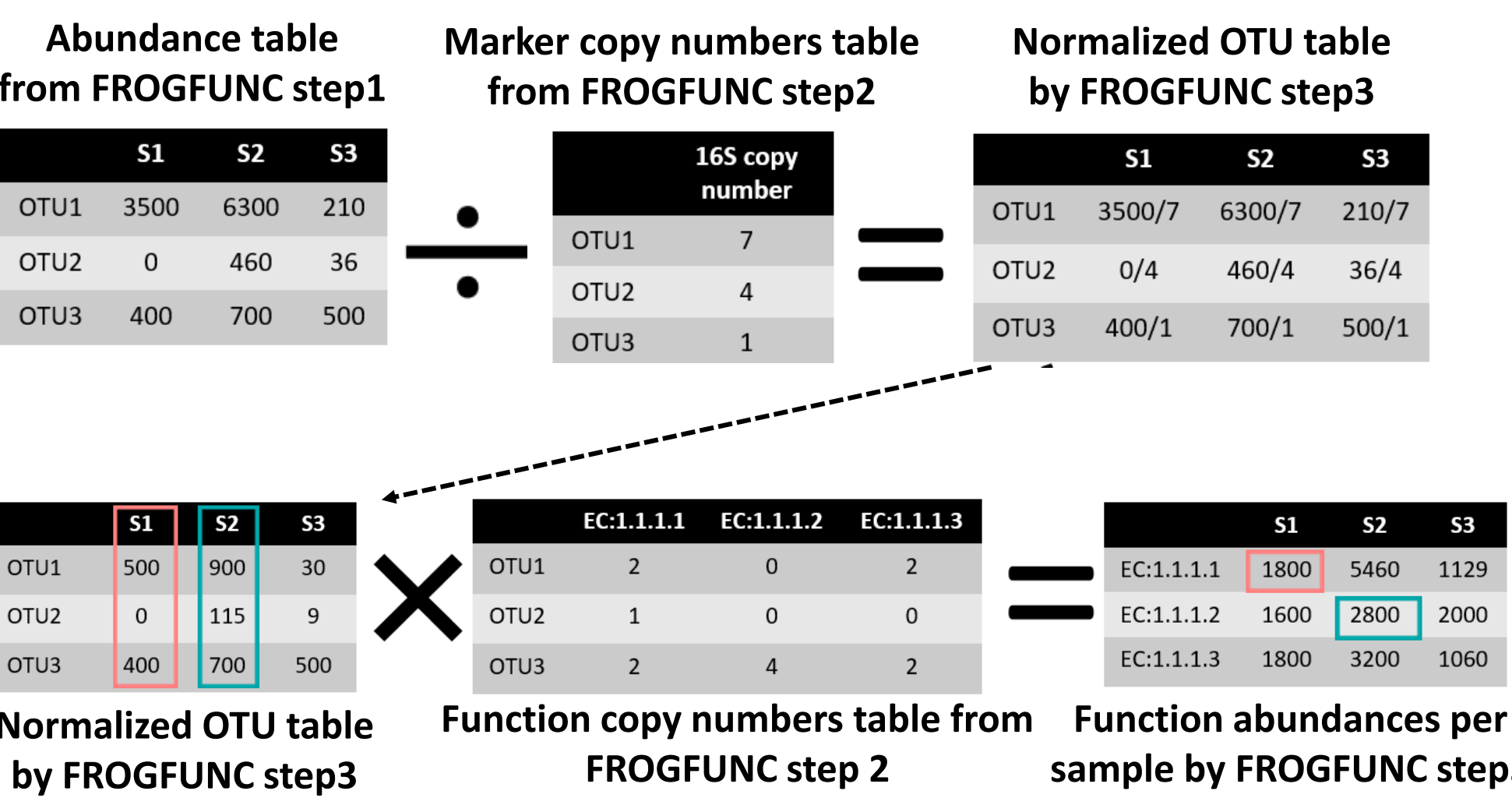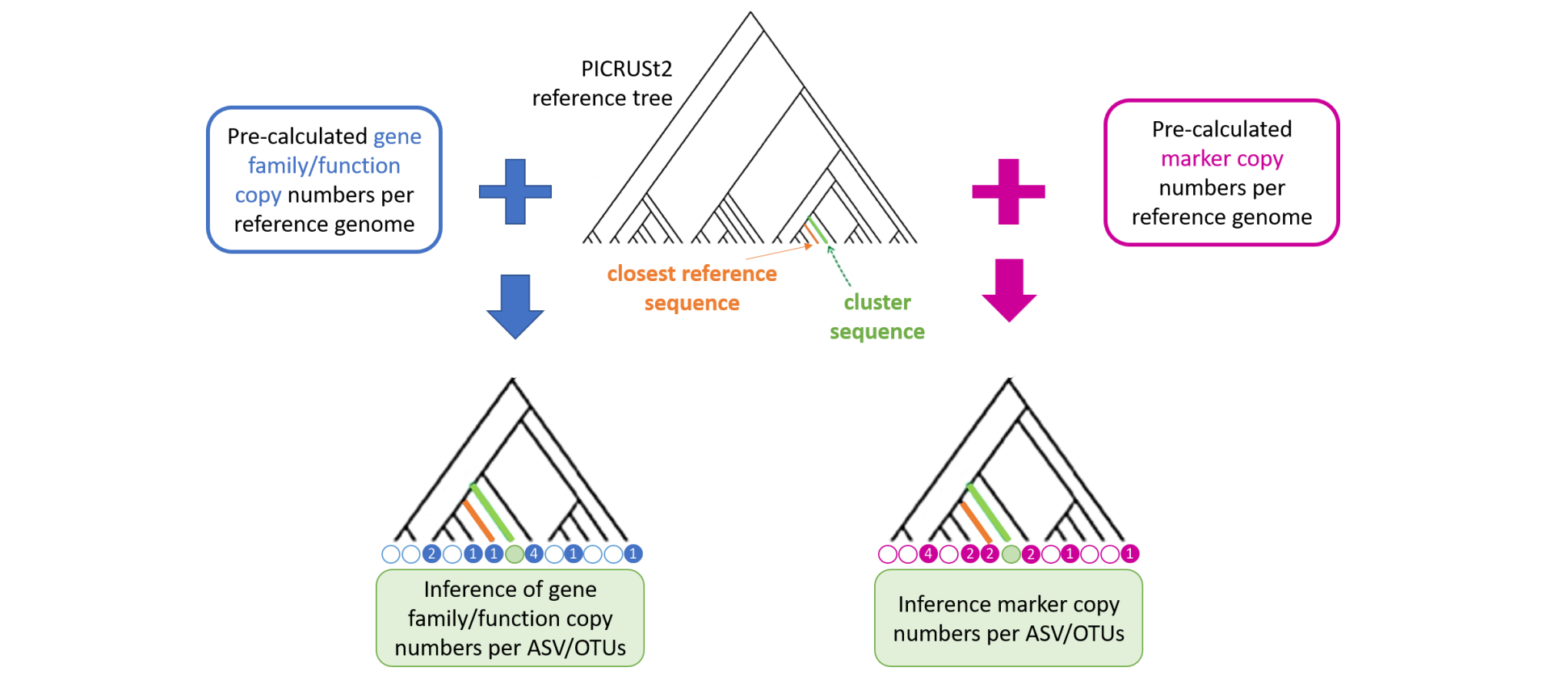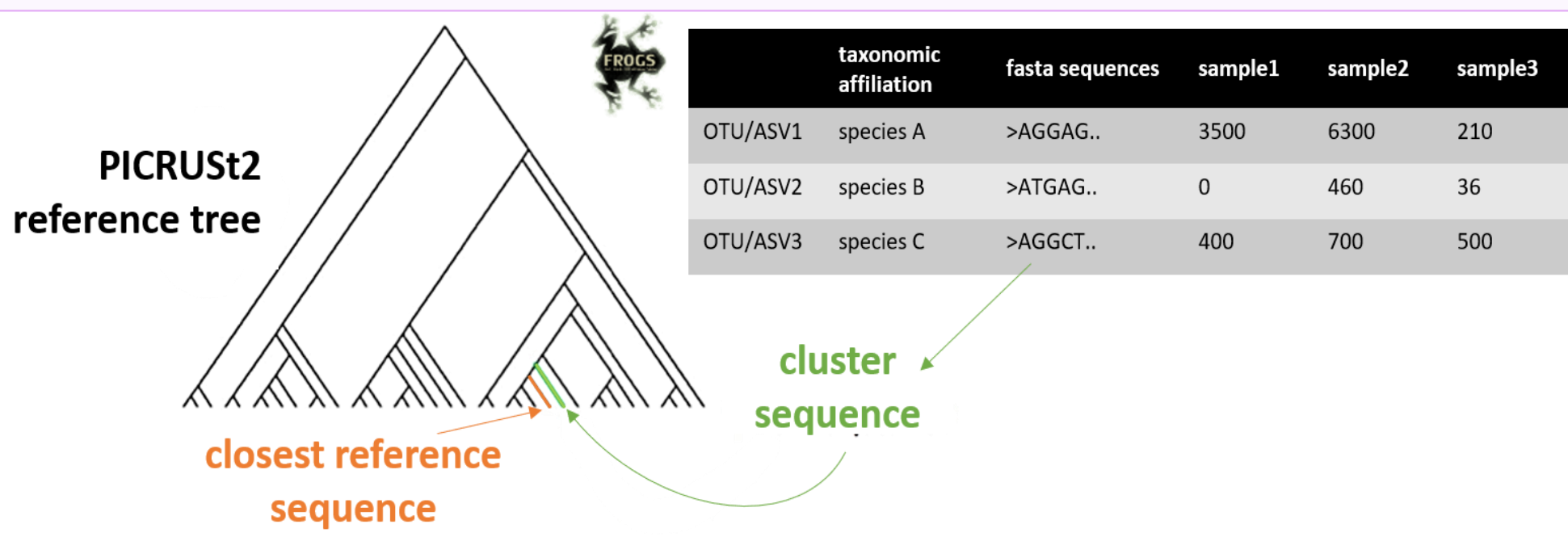Prediction of functions weighted by the relative abundance of OTUs in the community.

Function abundance predictions per sample
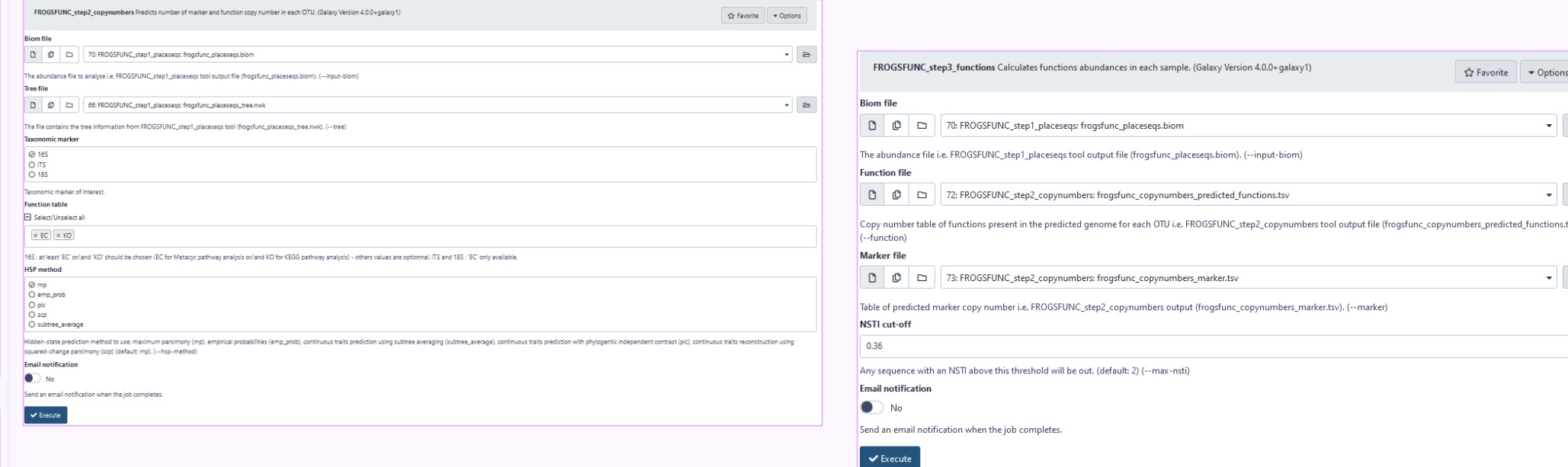
FROGSFUNC step3 functions tool

Inferrance of MetaCyc/KEGG pathway abundances based on function number abundances.

Pathway abundance predictions per sample
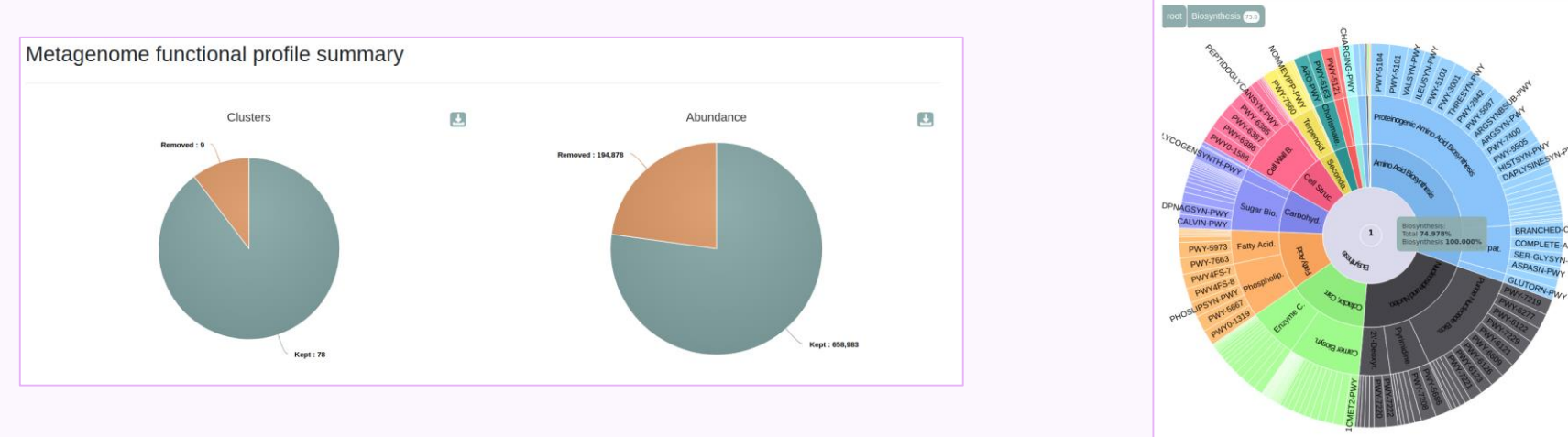
FROGSFUNC step4 pathways tool

PICRUSt2 reference tree

	taxonomic affiliation	fasta sequences	sample1	sample2	sample3
OTU/ASV1	species A	>AGGAG...	3500	6300	210
OTU/ASV2	species B	>ATGAG...	0	460	36
OTU/ASV3	species C	>AGGCT...	400	700	500

closest reference sequence · cluster sequence

Pre-calculated gene family/function copy numbers per reference genome · Pre-calculated marker copy numbers per reference genome

closest reference sequence · cluster sequence

Inference of gene family/function copy numbers per ASV/OTUs · Inference marker copy numbers per ASV/OTUs

Abundance table from FROGFUNC step1

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

÷

Marker copy numbers table from FROGFUNC step2

	16S copy number
OTU1	7
OTU2	4
OTU3	1

=

Normalized OTU table by FROGFUNC step3

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

Normalized OTU table by FROGFUNC step3

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

×

Function copy numbers table from FROGFUNC step 2

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

=

Function abundances per sample by FROGFUNC step3

	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

	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

→ PICRUST2 map of gene families to pathways →

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

Function abundances per sample from FROGFUNC step3

Metabolic pathway abundances per sample by FROGFUNC step4

Ability to use Galaxy interfaces

Various tables and graphical outputs are displayed to make the experience intuitive

Cluster/sequence number according to NSTI threshold
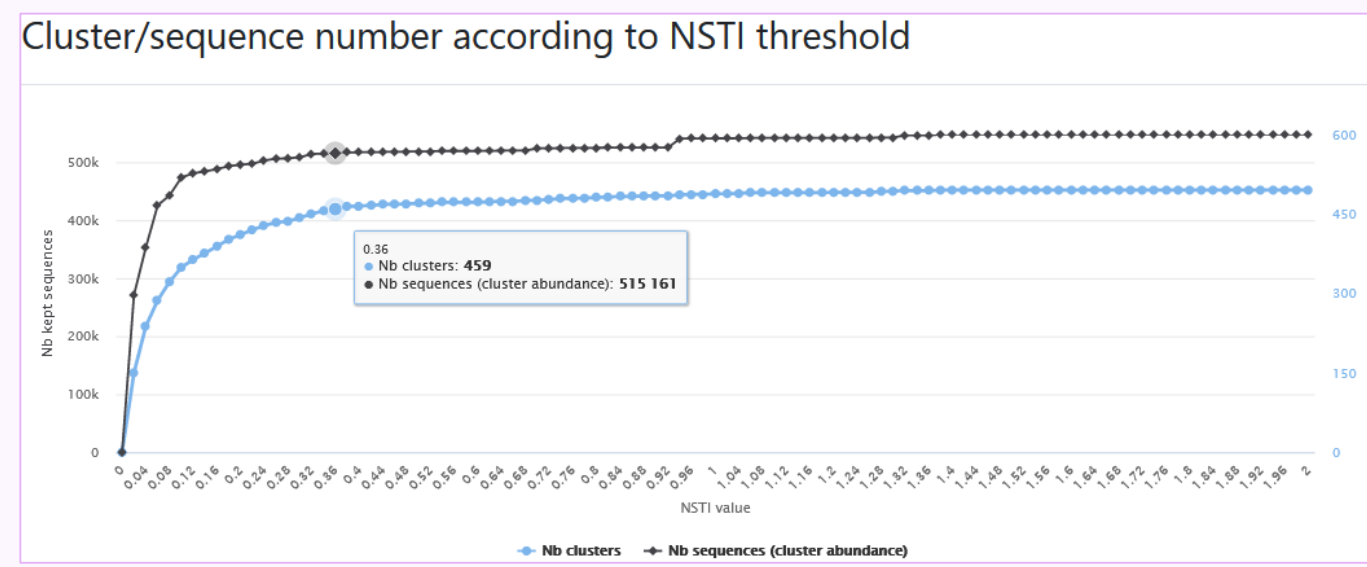
FROGS highlights important information that are very useful for data interpretation. Some "hidden" PICRUSt2 outputs are exploited: reporting incongruence between taxonomic affiliations, Nearest Sequenced Taxon Index (NSTI) threshold confidence indicator, decision support graphic to help choosing the NSTI threshold

Metagenome functional profile summary

References

[1] Escudié F, Auer L, Bernard M, Mariadassou M, Cauquil L, Vidal K, Maman S, Hernandez-Raquet G, Combes S, Pascal G. FROGS: Find, Rapidly, OTUs with Galaxy Solution. Bioinformatics. 2018 Apr 15;34(8):1287-1294. doi: 10.1093/bioinformatics/btx791. PMID: 29228191.

[2] Bernard M, Rué, O, Mariadassou M, and Pascal G; FROGS: a powerful tool to analyse the diversity of fungi with special management of internal transcribed spacers. Briefings in Bioinformatics. 2021, Nov. doi : 10.1093/bib/bbab318

[3] Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Grüning BA, Guerler A, Hillman-Jackson J, Hiltemann S, Jalili V, Rasche H, Soranzo N, Goecks J, Taylor J, Nekrutenko A, Blankenberg D. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 2018 Jul 2;46(W1):W537-W544. doi: 10.1093/nar/gky379. PMID: 29790989; PMCID: PMC6030816.

[4] Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. Microb Inform Exp. 2012 Feb 9;2(1):3. doi: 10.1186/2042-5783-2-3. PMID: 22587947; PMCID: PMC3351745.

[5] Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, Huttenhower C, Langille MGI. PICRUSt2 for prediction of metagenome functions. Nat Biotechnol. 2020 Jun;38(6):685-688. doi: 10.1038/s41587-020-0548-6. PMID: 32483366; PMCID: PMC7365738.