Training on Galaxy: Metabarcoding

October 2022 - Webinar

FROGS Practice on function inference

LUCAS AUER, MARIA BERNARD, LAURENT CAUQUIL, VINCENT DARBOT, MAHENDRA MARIADASSOU, GÉRALDINE PASCAL & OLIVIER RUÉ

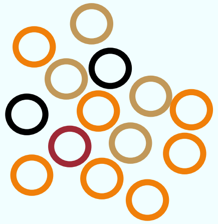
What metabolic functions are present in the environment?

Concepts

Metabarcoding principle

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

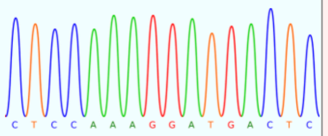
Microorganism genomes



DNA extraction

PCR amplification of marker (16S, 18S, ITS…)

Sequencing



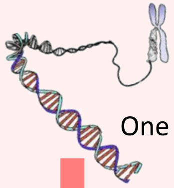
Bioinformatics analysis with FROGS

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.



One gene

Previous reaction

Next reaction

Reaction

Metabolic pathway

What metabolic functions are presents in a microbial community ?

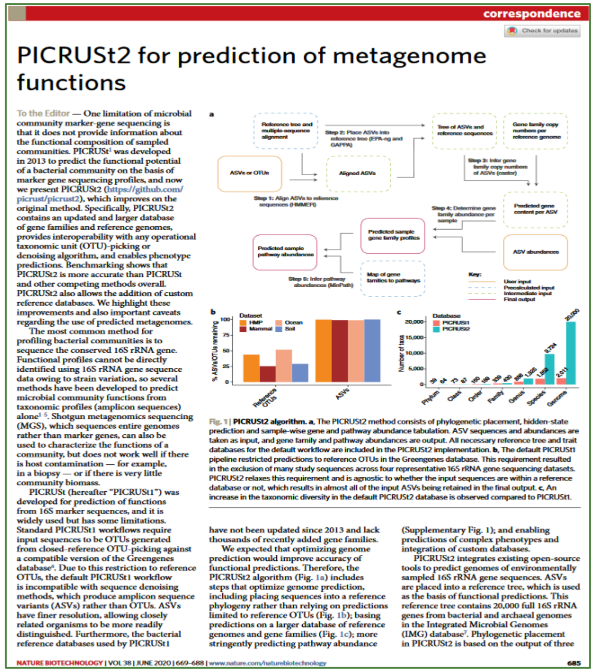
Based on PICRUSt2

PICRUSt (Phylogenetic investigation of communities by reconstruction of unobserved states) is an open-source tool.
It is a software for predicting functional abundances based only on marker gene sequences.

PICRUSt2 is composed of 4 python applications.

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 for prediction of metagenome functions

To the Editor — One limitation of microbial community marker-gene sequencing is that it does not provide information about the functional composition of sampled communities. PICRUSt[1] was developed in 2013 to predict the functional potential of a bacterial community on the basis of marker gene sequencing profiles, and now we present PICRUSt2 (https://github.com/picrust/picrust2), which improves on the original method. Specifically, PICRUSt2 contains an updated and larger database of gene families and reference genomes, provides interoperability with any operational taxonomic unit (OTU)-picking or denoising algorithm, and enables phenotype predictions. Benchmarking shows that PICRUSt2 is more accurate than PICRUSt and other competing methods overall. PICRUSt2 also allows the addition of custom reference databases. We highlight these improvements and also important caveats regarding the use of predicted metagenomes.

The most common method for profiling bacterial communities is to sequence the conserved 16S rRNA gene. Functional profiles cannot be directly identified using 16S rRNA gene sequence data owing to strain variation, so several methods have been developed to predict microbial community functions from taxonomic profiles (amplicon sequences) alone[1–5]. Shotgun metagenomics sequencing (MGS), which sequences entire genomes rather than marker genes, can also be used to characterize the functions of a community, but does not work well if there is host contamination — for example, in a biopsy — or if there is very little community biomass.

PICRUSt (hereafter "PICRUSt1") was developed for prediction of functions from 16S marker sequences, and it is widely used but has some limitations. Standard PICRUSt1 workflows require input sequences to be OTUs generated from closed-reference OTU-picking against a compatible version of the Greengenes database[6]. Due to this restriction to reference OTUs, the default PICRUSt1 workflow is incompatible with sequence denoising methods, which produce amplicon sequence variants (ASVs) rather than OTUs. ASVs have finer resolution, allowing closely related organisms to be more readily distinguished. Furthermore, the bacterial reference databases used by PICRUSt1 have not been updated since 2013 and lack thousands of recently added gene families.

We expected that optimizing genome prediction would improve accuracy of functional predictions. Therefore, the PICRUSt2 algorithm (Fig. 1a) includes steps that optimize genome prediction, including placing sequences into a reference phylogeny rather than relying on predictions limited to reference OTUs (Fig. 1b); basing predictions on a larger database of reference genomes and gene families (Fig. 1c); more stringently predicting pathway abundance (Supplementary Fig. 1); and enabling predictions of complex phenotypes and integration of custom databases.

PICRUSt2 integrates existing open-source tools to predict genomes of environmentally sampled 16S rRNA gene sequences. ASVs are placed into a reference tree, which is used as the basis of functional predictions. This reference tree contains 20,000 full 16S rRNA genes from bacterial and archaeal genomes in the Integrated Microbial Genomes (IMG) database[7]. Phylogenetic placement in PICRUSt2 is based on the output of three

Fig. 1 | **PICRUSt2 algorithm. a**, The PICRUSt2 method consists of phylogenetic placement, hidden-state prediction and sample-wise gene and pathway abundance tabulation. 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. **b**, The default PICRUSt1 pipeline restricted predictions to reference OTUs in the Greengenes database. This requirement resulted in the exclusion of many study sequences across four representative 16S rRNA gene sequencing datasets. PICRUSt2 relaxes this requirement and is agnostic to whether the input sequences are within a reference database or not, which results in almost all of the input ASVs being retained in the final output. **c**, An increase in the taxonomic diversity in the default PICRUSt2 database is observed compared to PICRUSt1.

How it works ?

1. Places the OTUs into a reference phylogenetic tree. ← FROGSFUNC_step1_placeseqs
2. Predicts number of marker and function copy number in each OTU. ← FROGSFUNC_step2_copynumbers
3. Calculates functions abundances in each sample. ← FROGSFUNC_step3_functions
4. Calculates pathway abundances in each sample. ← FROGSFUNC_step4_pathways

It runs only on 16S, ITS or 18S

FROGSFUNC_step1_placeseqs

FROGSFUNC_step1_placeseqs

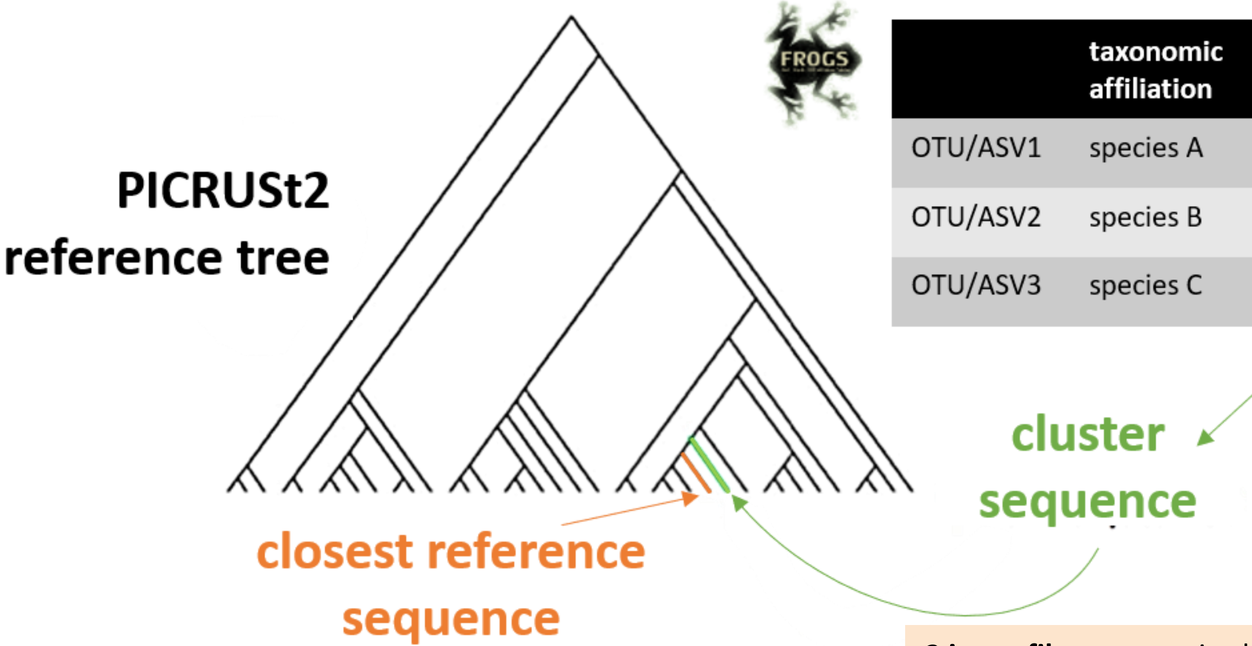
FROGSFUNC_step1_placeseqs is the first step of PICRUSt2.

It inserts your studied sequences into a reference tree.

By default, this reference tree is based on 20,000 16S sequences from genomes in the Integrated Microbial Genomes database. The script performs this step, which specifically:

- Aligns your study sequences with a multiple-sequence alignment of reference 16S, ITS or 18S sequences with HMMER.
- Finds the most likely placements of your study sequences in the reference tree with EPA_NG or SEPP .
- Produces a treefile with the most likely placement for each sequence as the new tips with GAPPA.

FROGSFUNC_step1_placeseqs

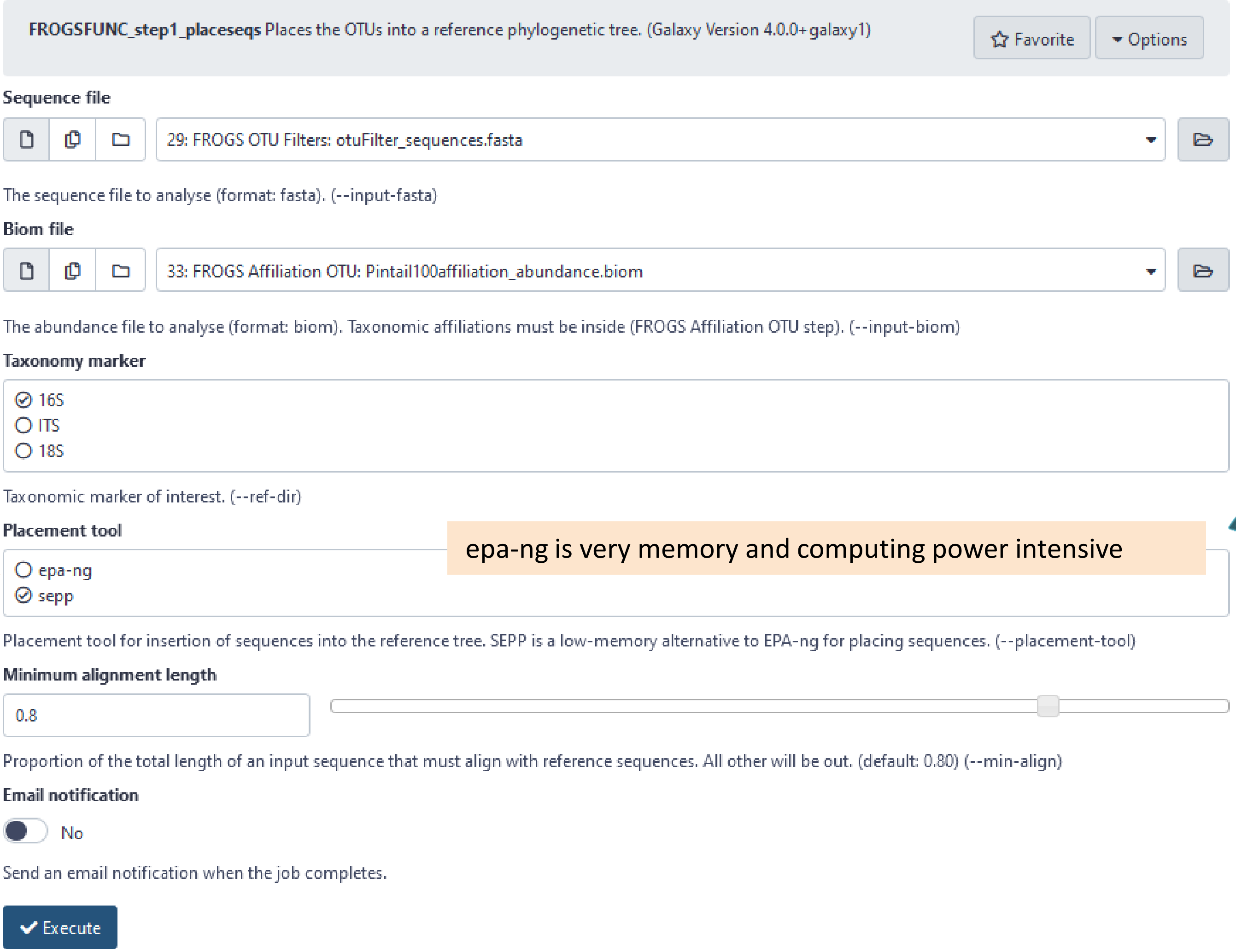


	taxonomic affiliation	fasta sequences	sample1	sample2	sample3
OTU/ASV1	species A	>AGGAG..	3500	6300	210
OTU/ASV2	species B	>ATGAG..	0	460	36
OTU/ASV3	species C	>AGGCT..	400	700	500

2 input files are required for FROGSFUNC_step1_placeseqs analysis:

- **fasta file** of OTU/ASV sequences (it can be from FROGS Filters step)
- **biom file** of OTU/ASV abundances with taxonomic affiliation information (it can be from FROGS Affiliation OTU step)

FROGSFUNC_step1_placeseqs



FROGSFUNC_step1_placeseqs Places the OTUs into a reference phylogenetic tree. (Galaxy Version 4.0.0+galaxy1)

Sequence file

29: FROGS OTU Filters: otuFilter_sequences.fasta

The sequence file to analyse (format: fasta). (--input-fasta)

Biom file

33: FROGS Affiliation OTU: Pintail100affiliation_abundance.biom

The abundance file to analyse (format: biom). Taxonomic affiliations must be inside (FROGS Affiliation OTU step). (--input-biom)

Taxonomy marker

- ● 16S
- ○ ITS
- ○ 18S

Taxonomic marker of interest. (--ref-dir)

Placement tool

- ○ epa-ng
- ● sepp

Placement tool for insertion of sequences into the reference tree. SEPP is a low-memory alternative to EPA-ng for placing sequences. (--placement-tool)

Minimum alignment length

0.8

Proportion of the total length of an input sequence that must align with reference sequences. All other will be out. (default: 0.80) (--min-align)

Email notification

No

Send an email notification when the job completes.

epa-ng is very memory and computing power intensive

epa-ng is the only choice for ITS and 18S
Since epa-ng is greedy then it may force to reduce the dataset to run.

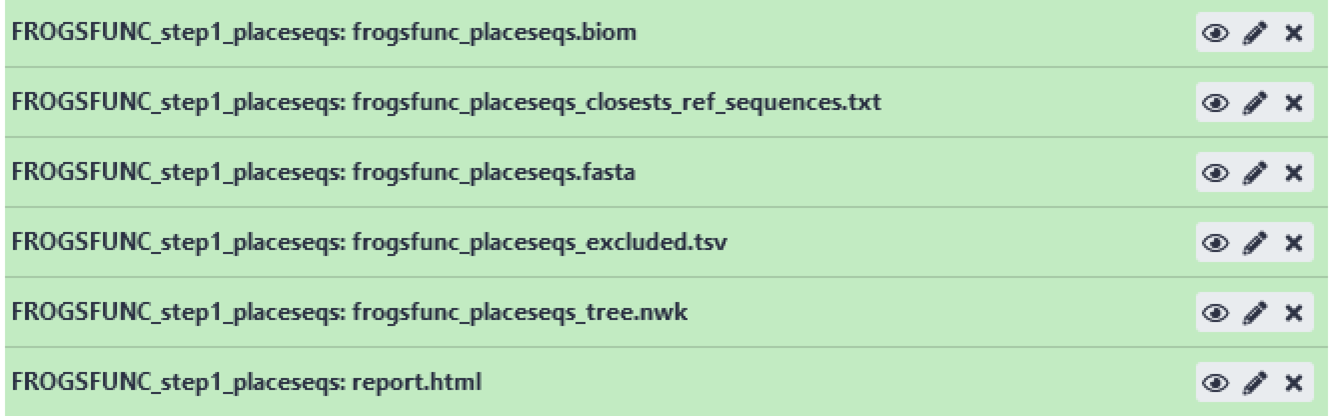
Input files

- **Sequences file**: The OTU/ASV fasta sequence file.
- **biom file**: The OTU biom file. Taxonomic affiliations must be done before (biom file form FROGS Affiliation OTU tool).
- **taxonomy marker**: 16S, ITS and 18S only available.

If your OTUs/ASVs are based on another marker, you cannot use this tool.

- **placement tool**: EPA-NG or SEPP are placement tools for insertion of sequences into the PICRUSt2 reference tree. SEPP is a low-memory alternative to EPA-ng for placing sequences. So, if the tool crashes with EPA-ng, try again with SEPP.
- **minimum alignment length**: Proportion of the total length of an input sequence that must align with reference sequences. All other will be out.

Output files



FROGSFUNC_step1_placeseqs: frogsfunc_placeseqs.biom

FROGSFUNC_step1_placeseqs: frogsfunc_placeseqs_closests_ref_sequences.txt

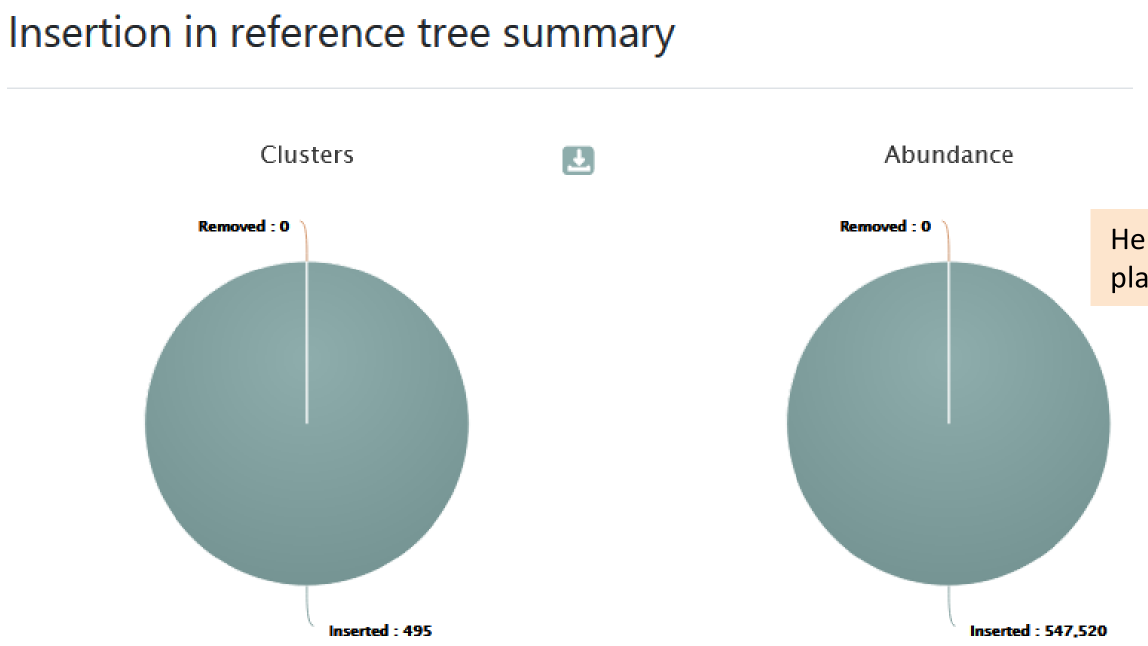
FROGSFUNC_step1_placeseqs: frogsfunc_placeseqs.fasta

FROGSFUNC_step1_placeseqs: frogsfunc_placeseqs_excluded.tsv

FROGSFUNC_step1_placeseqs: frogsfunc_placeseqs_tree.nwk

FROGSFUNC_step1_placeseqs: report.html

FROGSFUNC_step1_placeseqs: report.html



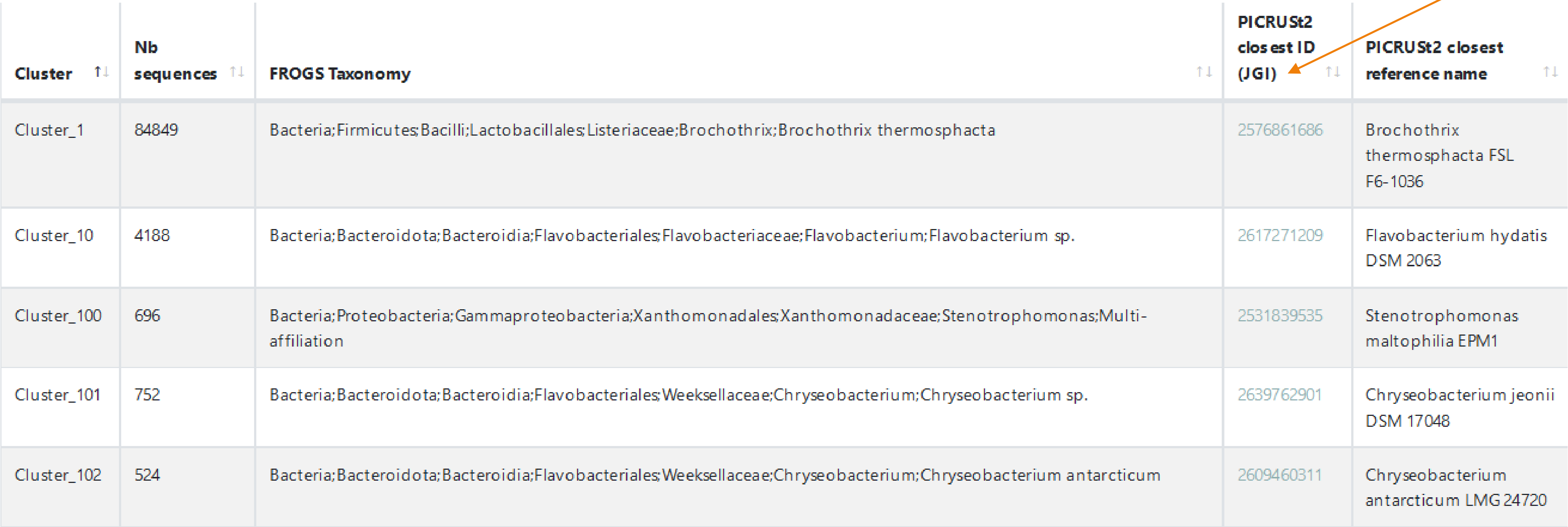
Insertion in reference tree summary

Here, all our sequences have found a place inside the PICRUSt2 reference tree.

The html report file describes which OTUs are contained or not in the phylogenetic tree.
Note that PICRUSt2 uses its own reference tree to affiliate OTUs from reference sequences.
The report file indicates for each OTU which is the closest PICRUSt2 reference sequence, and compares it to the original FROGS taxonomy.
Clicking on the sequence ID gives you more information about it JGI database.

Where are my OTUs inserted in the phylogenetic reference tree ?

https://jgi.doe.gov/

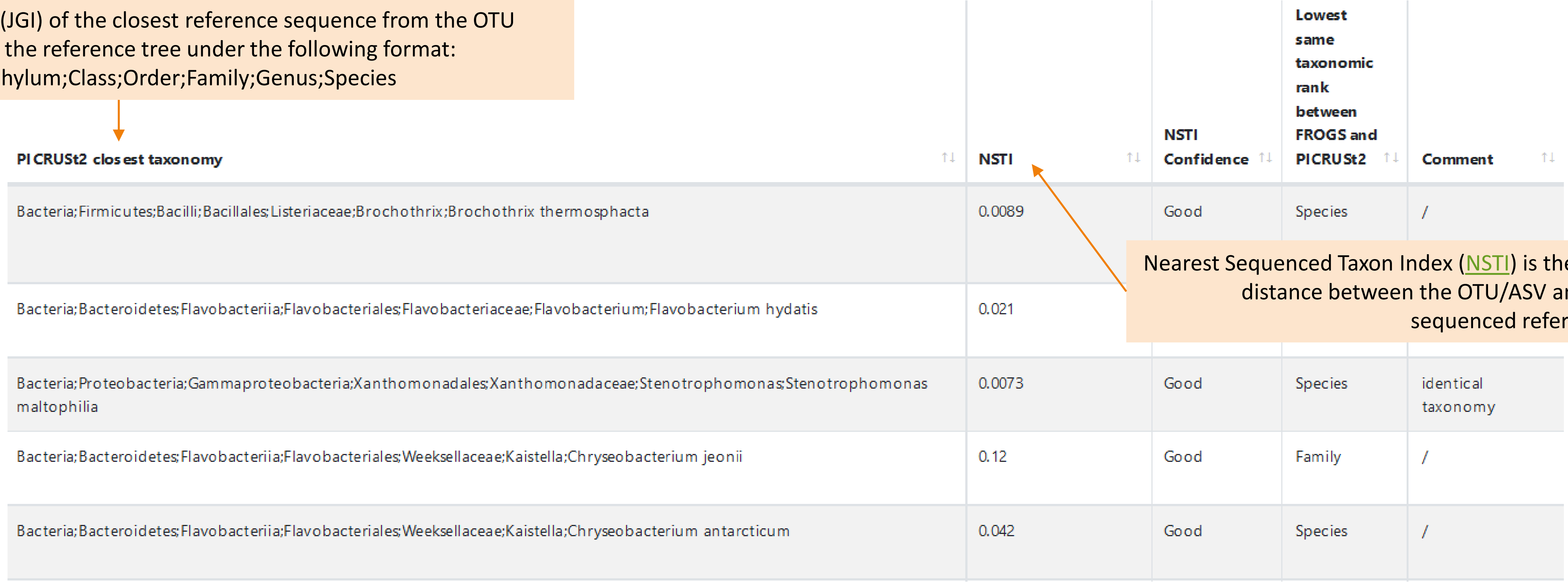


Cluster	Nb sequences	FROGS Taxonomy	PICRUSt2 closest ID (JGI)	PICRUSt2 closest reference name
Cluster_1	84849	Bacteria;Firmicutes;Bacilli;Lactobacillales;Listeriaceae;Brochothrix;Brochothrix thermosphacta	2576861686	Brochothrix thermosphacta FSL F6-1036
Cluster_10	4188	Bacteria;Bacteroidota;Bacteroidia;Flavobacteriales;Flavobacteriaceae;Flavobacterium;Flavobacterium sp.	2617271209	Flavobacterium hydatis DSM 2063
Cluster_100	696	Bacteria;Proteobacteria;Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae;Stenotrophomonas;Multi-affiliation	2531839535	Stenotrophomonas maltophilia EPM1
Cluster_101	752	Bacteria;Bacteroidota;Bacteroidia;Flavobacteriales;Weeksellaceae;Chryseobacterium;Chryseobacterium sp.	2639762901	Chryseobacterium jeonii DSM 17048
Cluster_102	524	Bacteria;Bacteroidota;Bacteroidia;Flavobacteriales;Weeksellaceae;Chryseobacterium;Chryseobacterium antarcticum	2609460311	Chryseobacterium antarcticum LMG 24720

first part of the table

Where are my OTUs inserted in the phylogenetic reference tree ?

Taxonomy (JGI) of the closest reference sequence from the OTU inserted in the reference tree under the following format: Kingdom;Phylum;Class;Order;Family;Genus;Species



PICRUSt2 closest taxonomy	NSTI	NSTI Confidence	Lowest same taxonomic rank between FROGS and PICRUSt2	Comment
Bacteria;Firmicutes;Bacilli;Bacillales;Listeriaceae;Brochothrix;Brochothrix thermosphacta	0.0089	Good	Species	/
Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;Flavobacteriaceae;Flavobacterium;Flavobacterium hydatis	0.021			
Bacteria;Proteobacteria;Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae;Stenotrophomonas;Stenotrophomonas maltophilia	0.0073	Good	Species	identical taxonomy
Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;Weeksellaceae;Kaistella;Chryseobacterium jeonii	0.12	Good	Family	/
Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;Weeksellaceae;Kaistella;Chryseobacterium antarcticum	0.042	Good	Species	/

Nearest Sequenced Taxon Index (NSTI) is the phylogenetic distance between the OTU/ASV and the nearest sequenced reference genome.

second part of the table

NSTI confidence

According to the NSTI score, we guide you in the confidence you can bring to the issue affiliation of PICRUSt2. Four levels are given:

- 0 < Good < 0.5
- 0.5 <= Medium < 1
- 1 <= Bad < 2
- To exclude >= 2

PICRUSt2 sets NSTI threshold to 2 per default. Some studies have shown that this threshold is permissive. Thus, it is important to see if the taxonomies between PICRUSt2 and FROGS are quite similar or not, in order to potentially choose a more stringent threshold afterwards.

For example, a NSTI lower than 0.5, with “species” as lowest common taxonomic rank between FROGS and PICRUSt2 will product a good prediction.

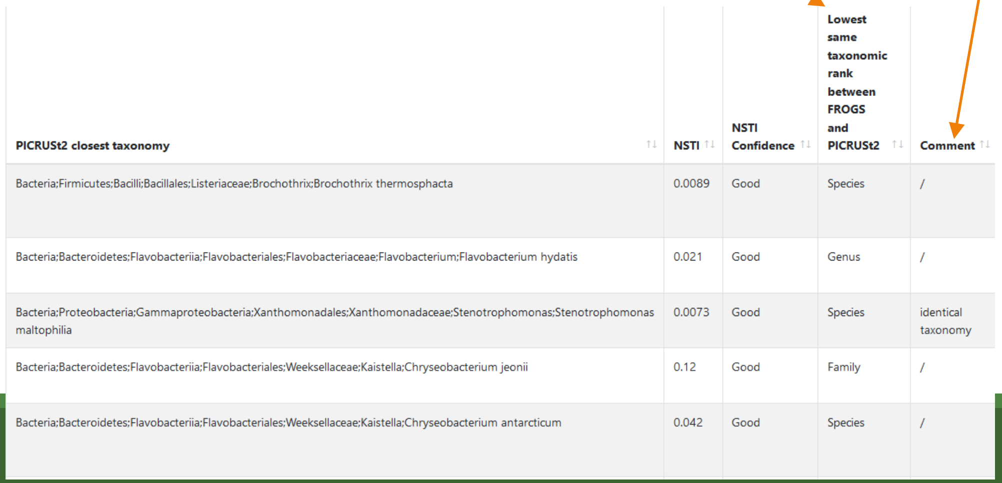
FROGSFUNC_step1_placeseqs: report.html

Lowest same taxonomic rank between FROGS and PICRUSt2 :

Lowest common taxonomic rank between FROGS **and** PICRUSt2 affiliations.

Comment :

- *identical taxonomy*: if the FROGS and PICRUSt2 taxonomic affiliations are identical.
- *identical sequence*: if the OTU/ASV sequence is strictly the same as the reference sequence.



PICRUSt2 closest taxonomy	NSTI	NSTI Confidence	Lowest same taxonomic rank between FROGS and PICRUSt2	Comment
Bacteria;Firmicutes;Bacilli;Bacillales;Listeriaceae;Brochothrix;Brochothrix thermosphacta	0.0089	Good	Species	/
Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;Flavobacteriaceae;Flavobacterium;Flavobacterium hydatis	0.021	Good	Genus	/
Bacteria;Proteobacteria;Gammaproteobacteria;Xanthomonadales;Xanthomonadaceae;Stenotrophomonas;Stenotrophomonas maltophilia	0.0073	Good	Species	identical taxonomy
Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;Weeksellaceae;Kaistella;Chryseobacterium jeonii	0.12	Good	Family	/
Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;Weeksellaceae;Kaistella;Chryseobacterium antarcticum	0.042	Good	Species	/

FROGSFUNC_step2_copynumbers

FROGSFUNC_step2_copynumbers

It runs hidden-state prediction (hsp) to predict function abundances with **castor-R** of each OTUs placed in the PICRUSt2 reference phylogenetic tree.

2 input files are required for FROGSFUNC_step2_copynumbers analysis:

- **Tree file** (format newick nwk): The file contains the tree informations from FROGSFUNC_step1_placeseqs (FROGSFUNC_step1_placeseqs output : frogsfunc_placeseqs_tree.nwk)
- **OTUs biom file**: The abundance file to analyse *i.e.* FROGSFUNC_step1_placeseqs tool output file (format biom). (frogsfunc_placeseqs.biom)

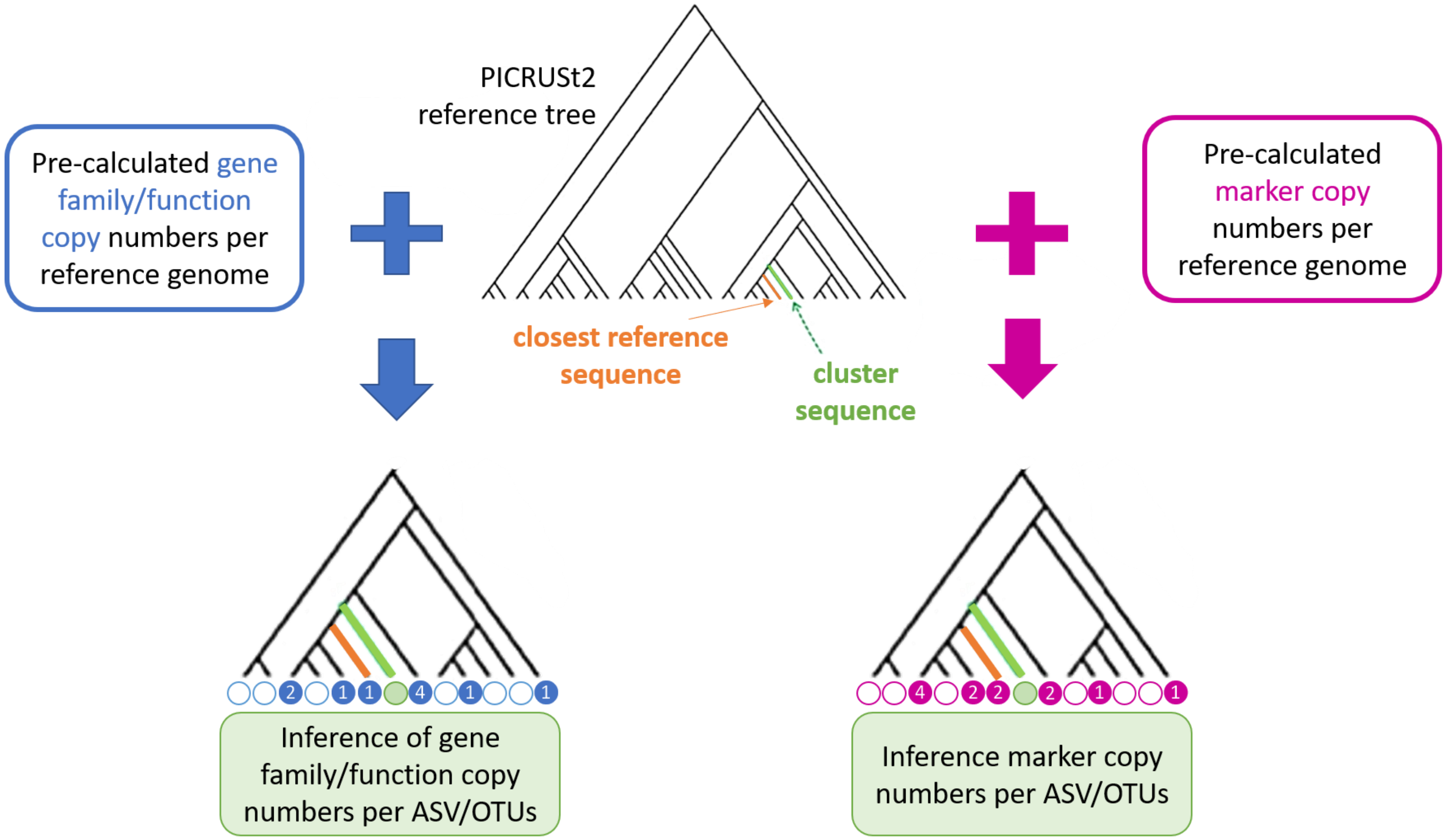
FROGSFUNC_step2_copynumbers

FROGSFUNC_step2_copynumbers **runs hsp twice**:

- Prediction of the **copy numbers** of the marker gene (16S, ITS or 18S) in order to normalize the OTU abundances table thereafter.
- Prediction of the **functions abundances**, using different databases:

EC : https://enzyme.expasy.org/
KO : https://www.genome.jp/kegg/ko.html
PFAM : http://pfam.xfam.org/
COG : https://www.ncbi.nlm.nih.gov/research/cog-project/
TIGRFAM : https://tigrfams.jcvi.org/cgi-bin/index.cgi
PHENO : https://phenodb.org/

FROGSFUNC_step2_copynumbers



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

PICRUSt2 reference tree

closest reference sequence

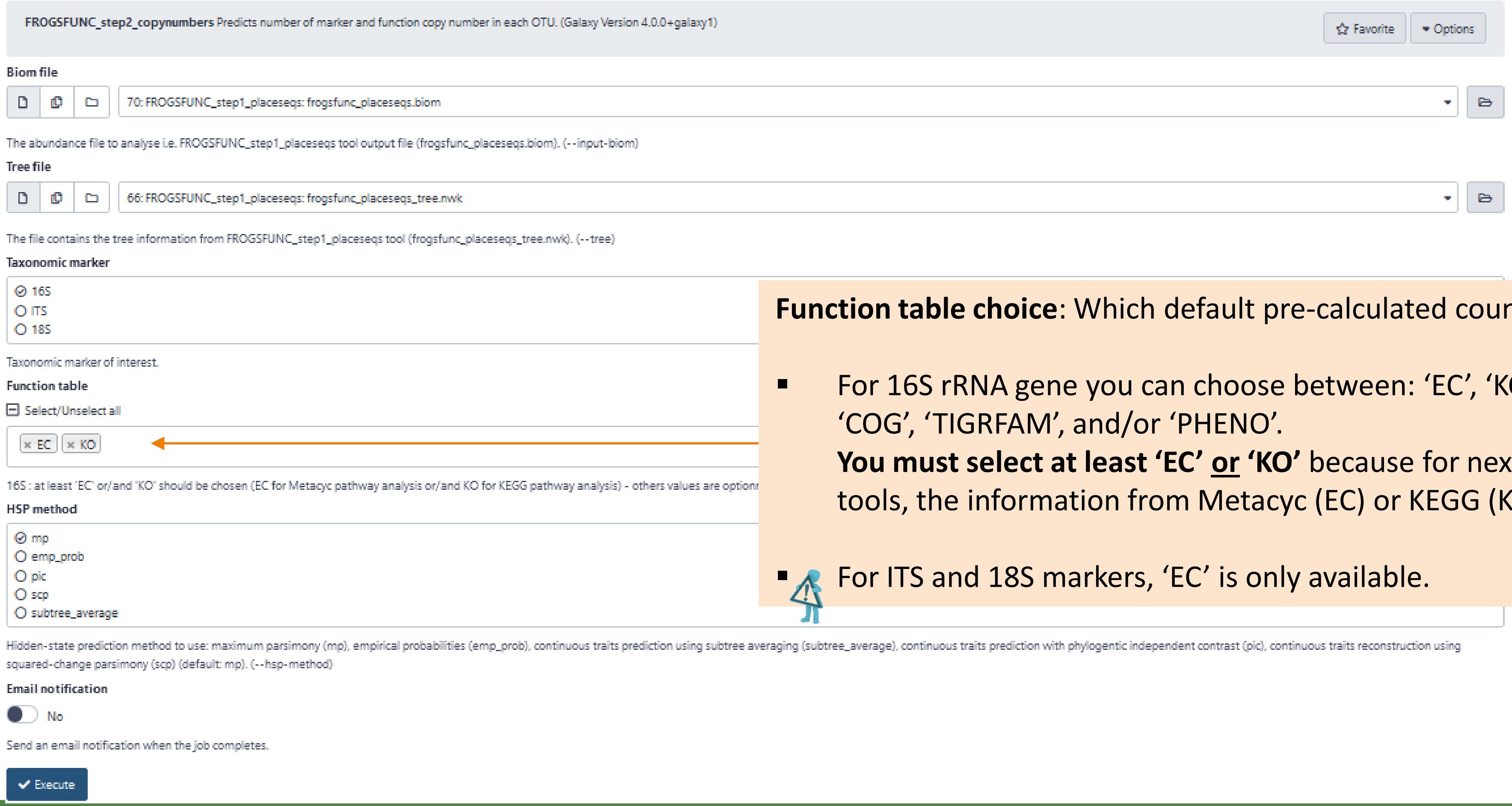
cluster sequence

Pre-calculated marker copy numbers per reference genome

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

Inference marker copy numbers per ASV/OTUs

FROGSFUNC_step2_copynumbers



FROGSFUNC_step2_copynumbers Predicts number of marker and function copy number in each OTU. (Galaxy Version 4.0.0+galaxy1)

Biom file

70: FROGSFUNC_step1_placeseqs: frogsfunc_placeseqs.biom

The abundance file to analyse i.e. FROGSFUNC_step1_placeseqs tool output file (frogsfunc_placeseqs.biom). (--input-biom)

Tree file

66: FROGSFUNC_step1_placeseqs: frogsfunc_placeseqs_tree.nwk

The file contains the tree information from FROGSFUNC_step1_placeseqs tool (frogsfunc_placeseqs_tree.nwk). (--tree)

Taxonomic marker

☑ 16S
○ ITS
○ 18S

Taxonomic marker of interest.

Function table

⊟ Select/Unselect all

× EC × KO

16S : at least 'EC' or/and 'KO' should be chosen (EC for Metacyc pathway analysis or/and KO for KEGG pathway analysis) - others values are optionn

HSP method

☑ mp
○ emp_prob
○ pic
○ scp
○ subtree_average

Hidden-state prediction method to use: maximum parsimony (mp), empirical probabilities (emp_prob), continuous traits prediction using subtree averaging (subtree_average), continuous traits prediction with phylogentic independent contrast (pic), continuous traits reconstruction using squared-change parsimony (scp) (default: mp). (--hsp-method)

Email notification

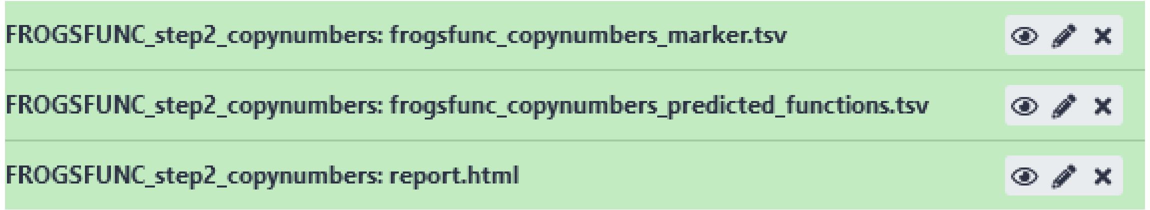
No

Send an email notification when the job completes.

Function table choice: Which default pre-calculated count table to use ?

- For 16S rRNA gene you can choose between: ‘EC’, ‘KO’, ‘PFAM’, ‘COG’, ‘TIGRFAM’, and/or ‘PHENO’.
  You must select at least ‘EC’ <u>or</u> ‘KO’ because for next FROGSFUNC tools, the information from Metacyc (EC) or KEGG (KO) are required.
- For ITS and 18S markers, ‘EC’ is only available.

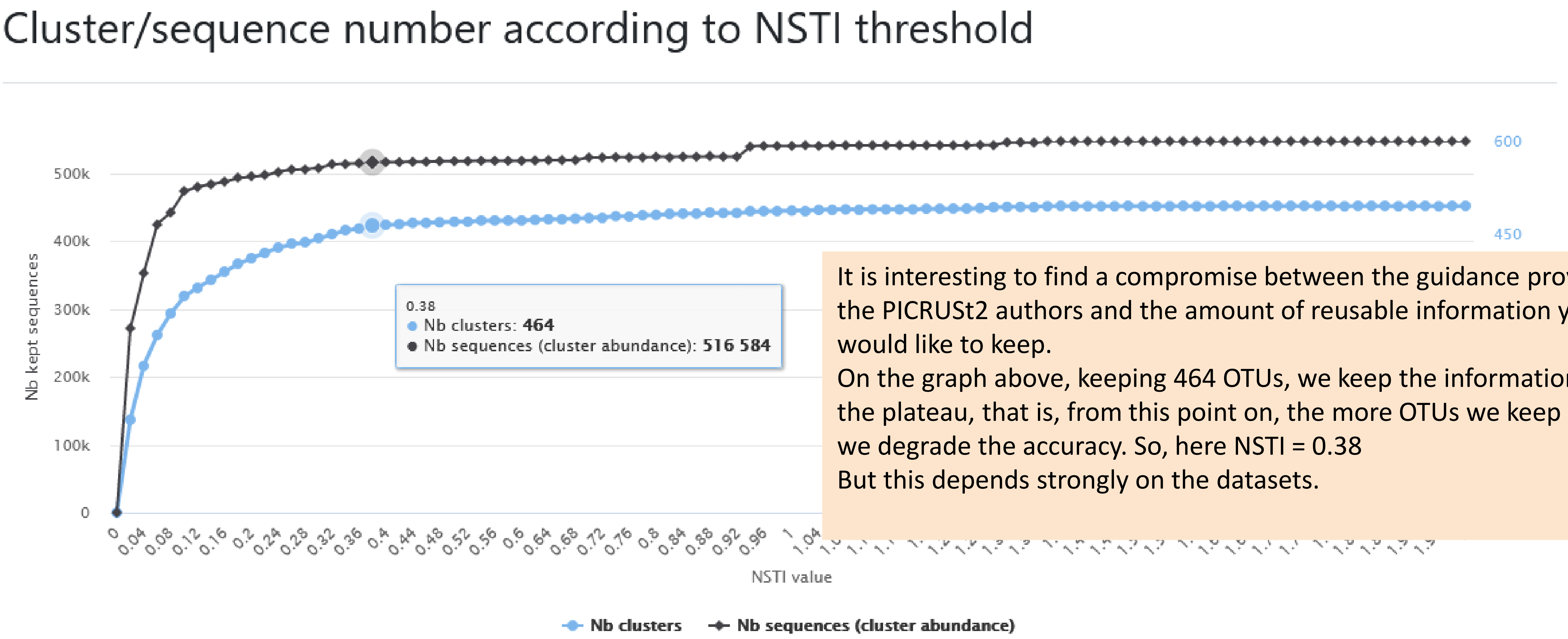
Outputs


FROGSFUNC_step2_copynumbers: frogsfunc_copynumbers_marker.tsv

FROGSFUNC_step2_copynumbers: frogsfunc_copynumbers_predicted_functions.tsv

FROGSFUNC_step2_copynumbers: report.html

FROGSFUNC_step2_copynumbers: report.html

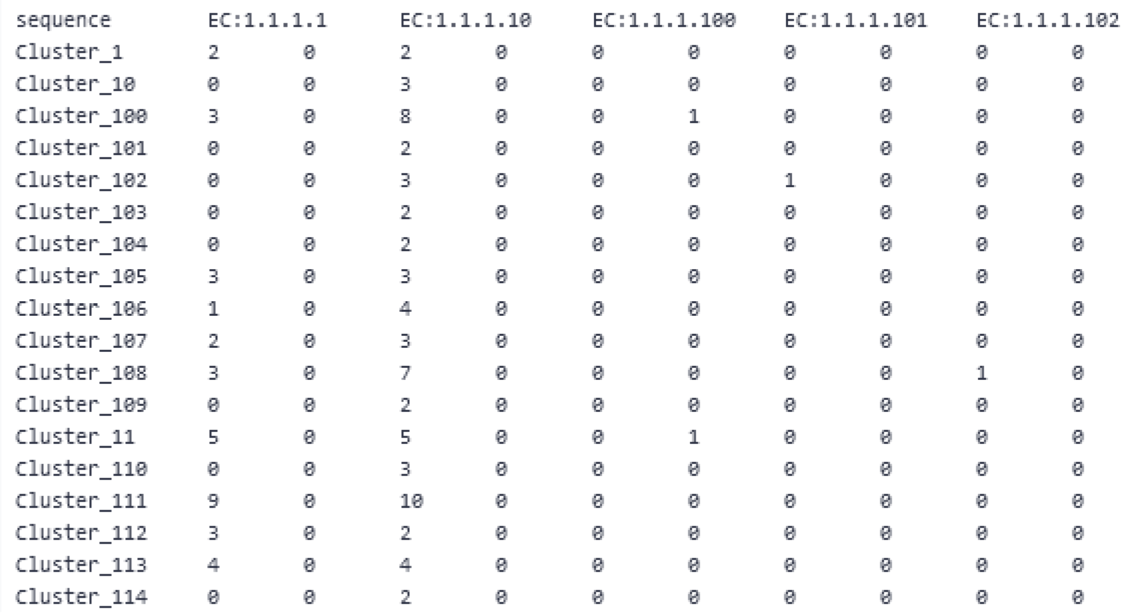


Cluster/sequence number according to NSTI threshold

It is interesting to find a compromise between the guidance provided by the PICRUSt2 authors and the amount of reusable information you would like to keep.
On the graph above, keeping 464 OTUs, we keep the information before the plateau, that is, from this point on, the more OTUs we keep the more we degrade the accuracy. So, here NSTI = 0.38
But this depends strongly on the datasets.

This graph allows you to set the “NSTI cut-off” parameter of the next tool

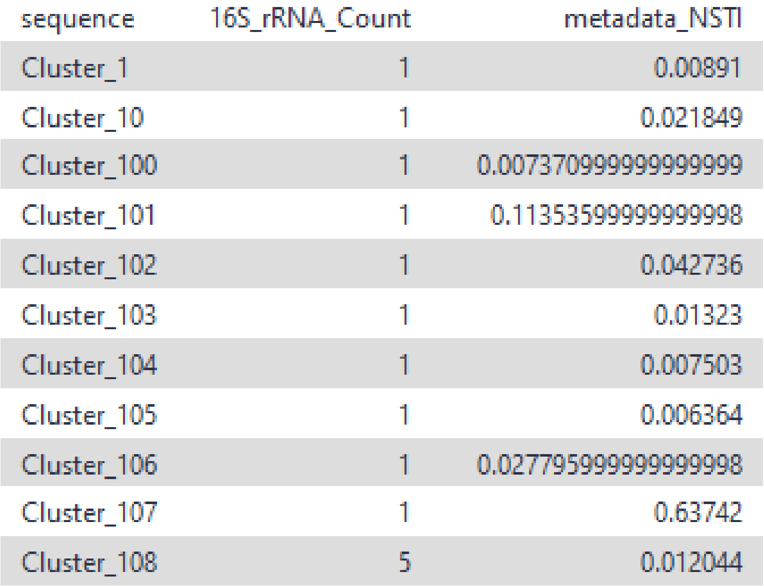
FROGSFUNC_step2_copynumbers: copynumber_predicted_functions.tsv



```
sequence        EC:1.1.1.1      EC:1.1.1.10     EC:1.1.1.100    EC:1.1.1.101    EC:1.1.1.102
Cluster_1       2       0       2       0       0       0       0       0       0       0
Cluster_10      0       0       3       0       0       0       0       0       0       0
Cluster_100     3       0       8       0       0       1       0       0       0       0
Cluster_101     0       0       2       0       0       0       0       0       0       0
Cluster_102     0       0       3       0       0       0       1       0       0       0
Cluster_103     0       0       2       0       0       0       0       0       0       0
Cluster_104     0       0       2       0       0       0       0       0       0       0
Cluster_105     3       0       3       0       0       0       0       0       0       0
Cluster_106     1       0       4       0       0       0       0       0       0       0
Cluster_107     2       0       3       0       0       0       0       0       0       0
Cluster_108     3       0       7       0       0       0       0       0       1       0
Cluster_109     0       0       2       0       0       0       0       0       0       0
Cluster_11      5       0       5       0       0       1       0       0       0       0
Cluster_110     0       0       3       0       0       0       0       0       0       0
Cluster_111     9       0       10      0       0       0       0       0       0       0
Cluster_112     3       0       2       0       0       0       0       0       0       0
Cluster_113     4       0       4       0       0       0       0       0       0       0
Cluster_114     0       0       2       0       0       0       0       0       0       0
```

Required file for FROGSFUNC_step3_functions tool.

FROGSFUNC_step2_copynumbers: copynumber_predicted_marker.tsv



sequence	16S_rRNA_Count	metadata_NSTI
Cluster_1	1	0.00891
Cluster_10	1	0.021849
Cluster_100	1	0.007370999999999999
Cluster_101	1	0.11353599999999998
Cluster_102	1	0.042736
Cluster_103	1	0.01323
Cluster_104	1	0.007503
Cluster_105	1	0.006364
Cluster_106	1	0.027795999999999998
Cluster_107	1	0.63742
Cluster_108	5	0.012044

Required file for FROGSFUNC_step3_functions tool.

FROGSFUNC_step3_functions

FROGSFUNC_step3_functions

FROGSFUNC_step1_placeseqs:
frogsfunc_placeseqs.biom

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

÷

FROGSFUNC_step2_copynumbers:
frogsfunc_copynumbers_marker.tsv

	16S copy number
OTU1	7
OTU2	4
OTU3	1

=

FROGSFUNC_step3_functions:
frogsfunc_functions_marker_norm.tsv

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

FROGSFUNC_step3_functions:
frogsfunc_functions_marker_norm.tsv

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

×

FROGSFUNC_step2_copynumbers:
frogsfunc_copynumbers_predicted_functions.tsv

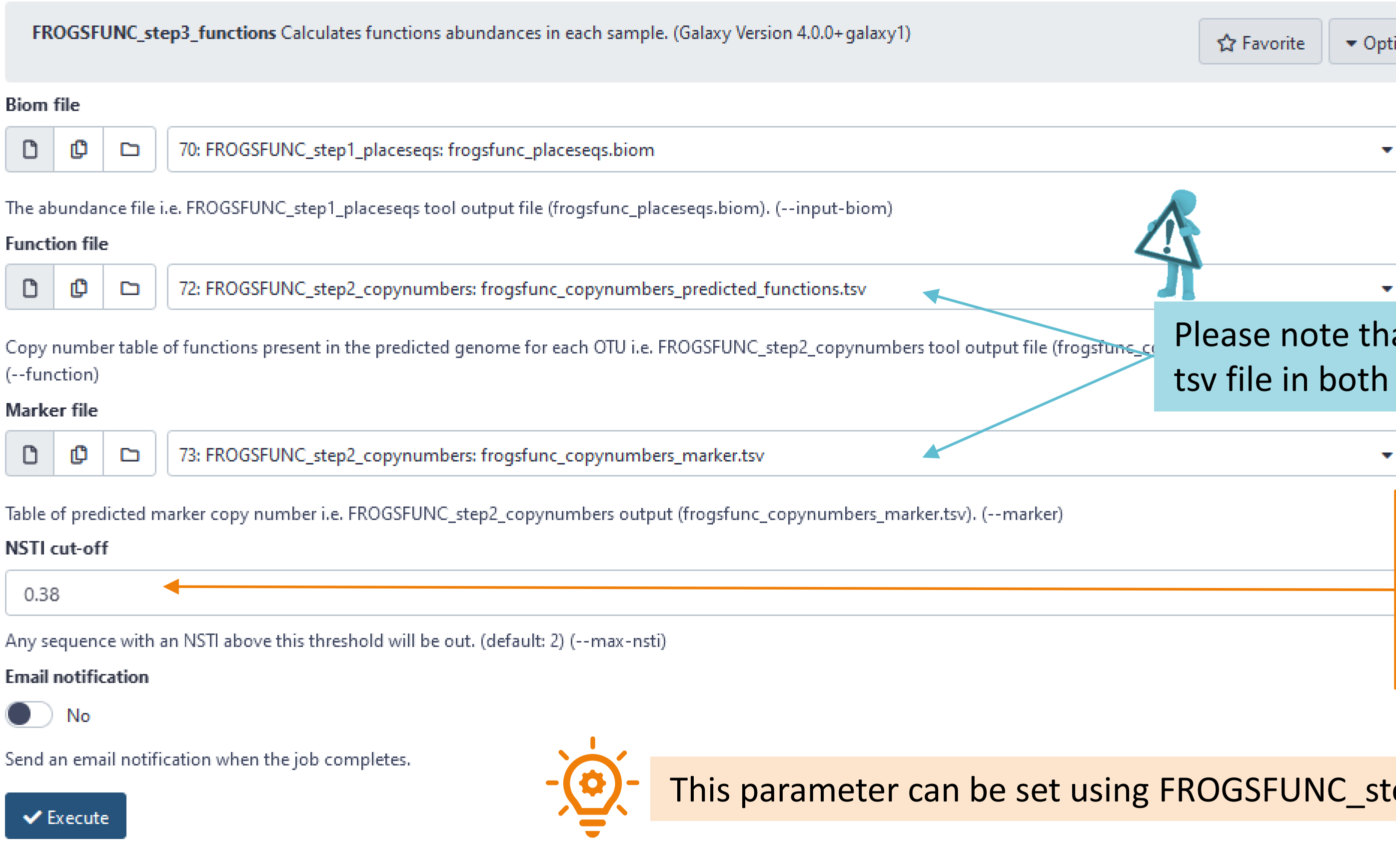
	EC:1.1.1.1	EC:1.1.1.2	EC:1.1.1.3
OTU1	2	0	2
OTU2	1	0	0
OTU3	2	4	2

=

FROGSFUNC_step3_functions:
frogsfunc_functions_unstrat.tsv

	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

FROGSFUNC_step3_functions



FROGSFUNC_step3_functions Calculates functions abundances in each sample. (Galaxy Version 4.0.0+galaxy1)

Biom file

70: FROGSFUNC_step1_placeseqs: frogsfunc_placeseqs.biom

The abundance file i.e. FROGSFUNC_step1_placeseqs tool output file (frogsfunc_placeseqs.biom). (--input-biom)

Function file

72: FROGSFUNC_step2_copynumbers: frogsfunc_copynumbers_predicted_functions.tsv

Copy number table of functions present in the predicted genome for each OTU i.e. FROGSFUNC_step2_copynumbers tool output file (frogsfunc_c
(--function)

Marker file

73: FROGSFUNC_step2_copynumbers: frogsfunc_copynumbers_marker.tsv

Table of predicted marker copy number i.e. FROGSFUNC_step2_copynumbers output (frogsfunc_copynumbers_marker.tsv). (--marker)

NSTI cut-off

0.38

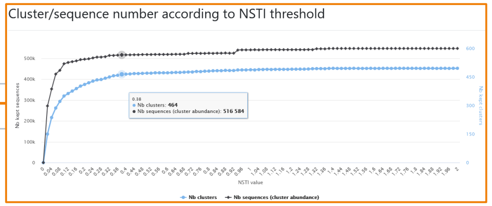
Any sequence with an NSTI above this threshold will be out. (default: 2) (--max-nsti)

Email notification

● No

Send an email notification when the job completes.

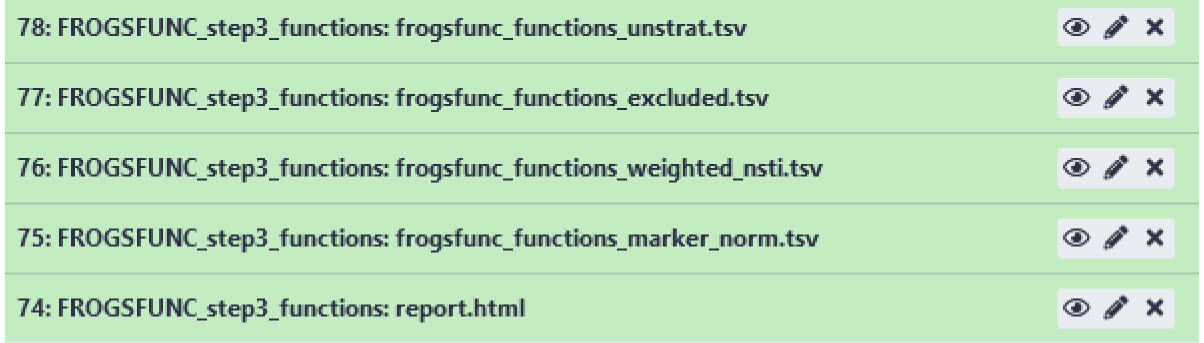
Please note that here Galaxy puts the same tsv file in both entries. Change the 1st.



Cluster/sequence number according to NSTI threshold

This parameter can be set using FROGSFUNC_step2_copynumbers: report.html

Outputs



78: FROGSFUNC_step3_functions: frogsfunc_functions_unstrat.tsv

77: FROGSFUNC_step3_functions: frogsfunc_functions_excluded.tsv

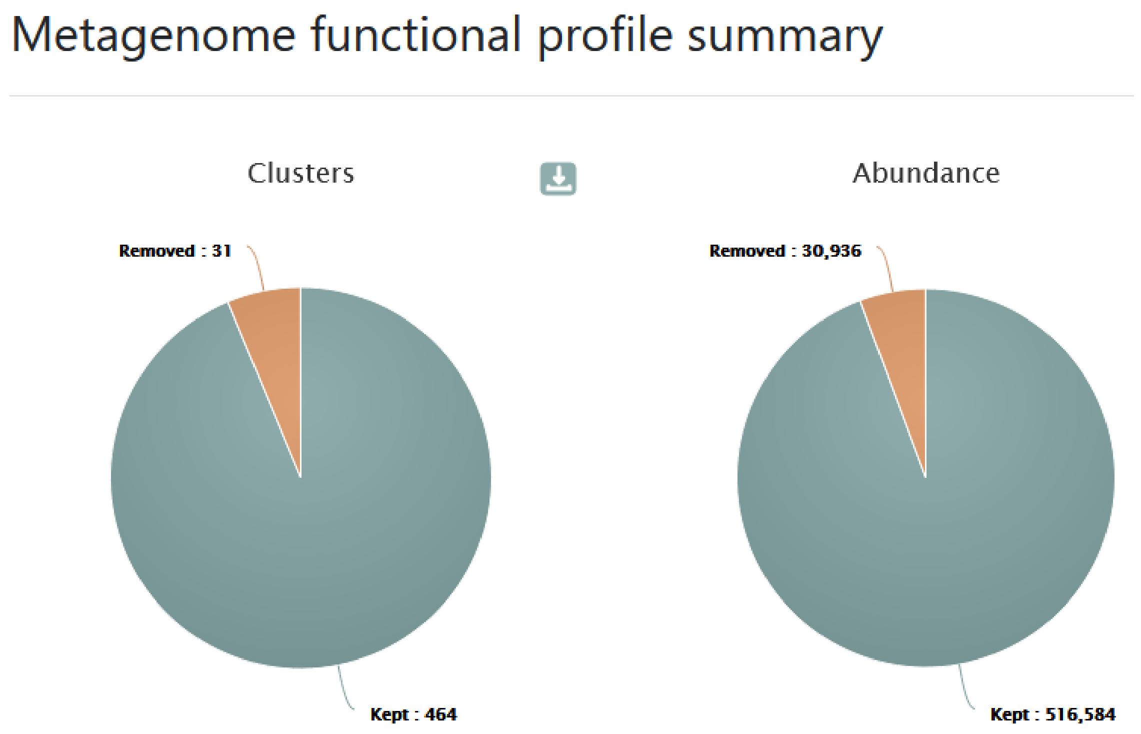
76: FROGSFUNC_step3_functions: frogsfunc_functions_weighted_nsti.tsv

75: FROGSFUNC_step3_functions: frogsfunc_functions_marker_norm.tsv

74: FROGSFUNC_step3_functions: report.html

FROGSFUNC_step3_functions: report.html

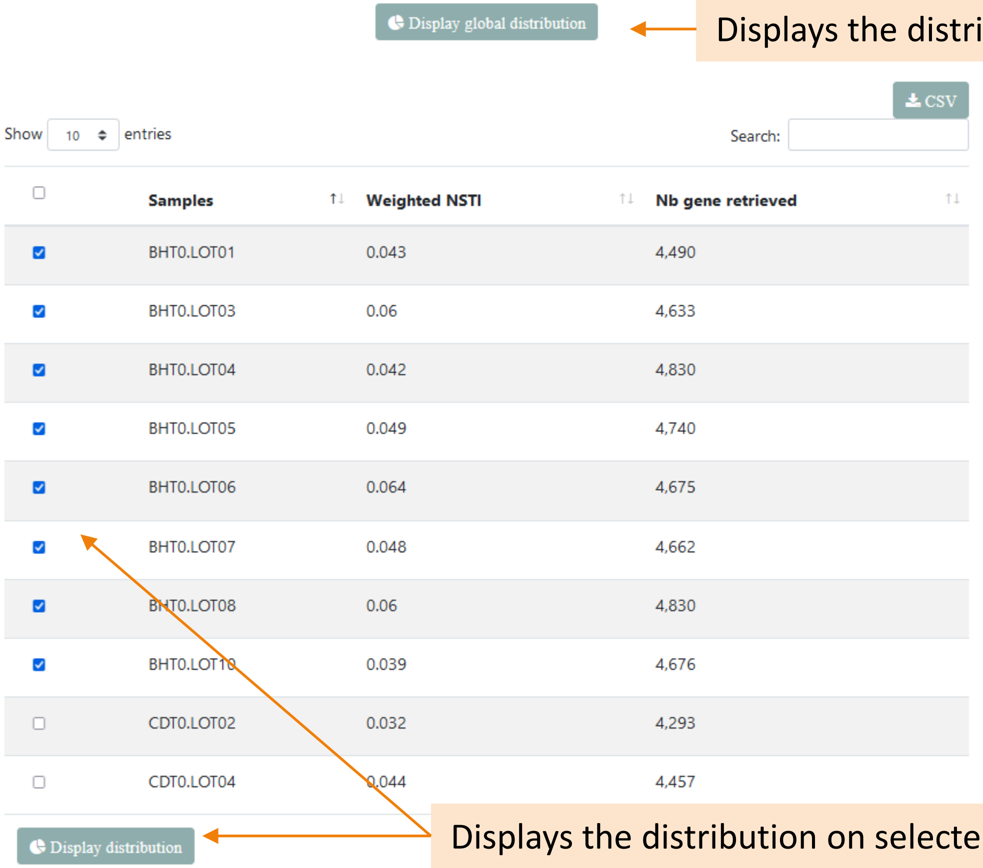
How many OTUs/sequences are kept after the process?



Metagenome functional profile summary

OTUs are out if the NSTI associated is above the threshold. In this example all OTUs with a NSTI > 0.38. Here, only 464 OTUs are kept in the functional inference process.

What is the distribution of gene/function abundances in the samples ?

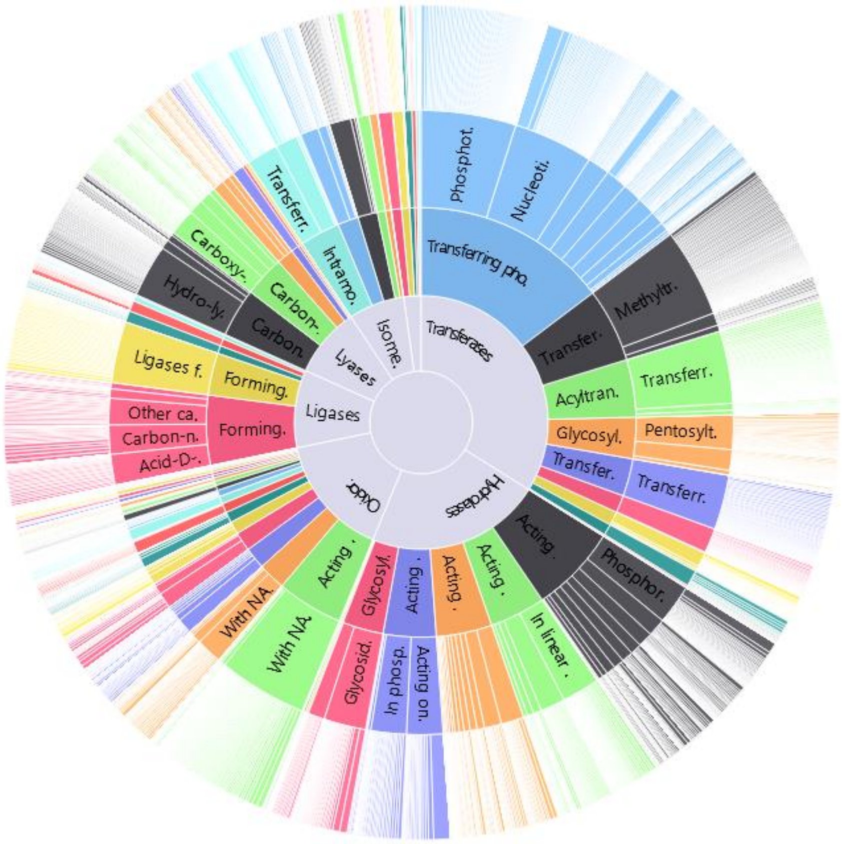


Displays the distribution on all samples

☐	Samples	Weighted NSTI	Nb gene retrieved
☑	BHT0.LOT01	0.043	4,490
☑	BHT0.LOT03	0.06	4,633
☑	BHT0.LOT04	0.042	4,830
☑	BHT0.LOT05	0.049	4,740
☑	BHT0.LOT06	0.064	4,675
☑	BHT0.LOT07	0.048	4,662
☑	BHT0.LOT08	0.06	4,830
☑	BHT0.LOT10	0.039	4,676
☐	CDT0.LOT02	0.032	4,293
☐	CDT0.LOT04	0.044	4,457

Displays the distribution on selected samples

What is the distribution of gene/function abundances in the samples ?



The innermost circle represents the highest hierarchical level of gene families according to Metacyc or Kegg databases.

The more we go outwards, the more the hierarchical level becomes precise until indicating the identifier of the gene family.

Output frogsfunc_functions_unstrat.tsv

Required file for FROGSFUNC_step4_pathways tool.

classification	db_link	observation_name	BHT0.LOT01	BHT0.LOT03	BHT0.LOT04
Oxidoreductases;Acting on the CH-OH group of donors;With NAD+ or NADP+ as acceptor;EC:1.1.1.1	https://www.genome.jp/dbget-bin/www_bget?EC:1.1.1.1	EC:1.1.1.1	13205	15364	13783
Oxidoreductases;Acting on the CH-OH group of donors;With NAD+ or NADP+ as acceptor;EC:1.1.1.100	https://www.genome.jp/dbget-bin/www_bget?EC:1.1.1.100	EC:1.1.1.100	20029	21513	23461
Oxidoreductases;Acting on the CH-OH group of donors;With NAD+ or NADP+ as acceptor;EC:1.1.1.103	https://www.genome.jp/dbget-bin/www_bget?EC:1.1.1.103	EC:1.1.1.103	177	2011	334
Oxidoreductases;Acting on the CH-OH group of donors;With NAD+ or NADP+ as acceptor;EC:1.1.1.105	https://www.genome.jp/dbget-bin/www_bget?EC:1.1.1.105	EC:1.1.1.105	0	0	0
Oxidoreductases;Acting on the CH-OH group of donors;With NAD+ or NADP+ as acceptor;EC:1.1.1.108	https://www.genome.jp/dbget-bin/www_bget?EC:1.1.1.108	EC:1.1.1.108	142	194	127
Oxidoreductases;Acting on the CH-OH group of donors;With NAD+ or NADP+ as acceptor;EC:1.1.1.11	https://www.genome.jp/dbget-bin/www_bget?EC:1.1.1.11	EC:1.1.1.11	2	10	2
Oxidoreductases;Acting on the CH-OH group of donors;With NAD+ or NADP+ as acceptor;EC:1.1.1.122	https://www.genome.jp/dbget-bin/www_bget?EC:1.1.1.122	EC:1.1.1.122	16	109	54
Oxidoreductases;Acting on the CH-OH group of donors;With NAD+ or NADP+ as acceptor;EC:1.1.1.125	https://www.genome.jp/dbget-bin/www_bget?EC:1.1.1.125	EC:1.1.1.125	2218	505	1656

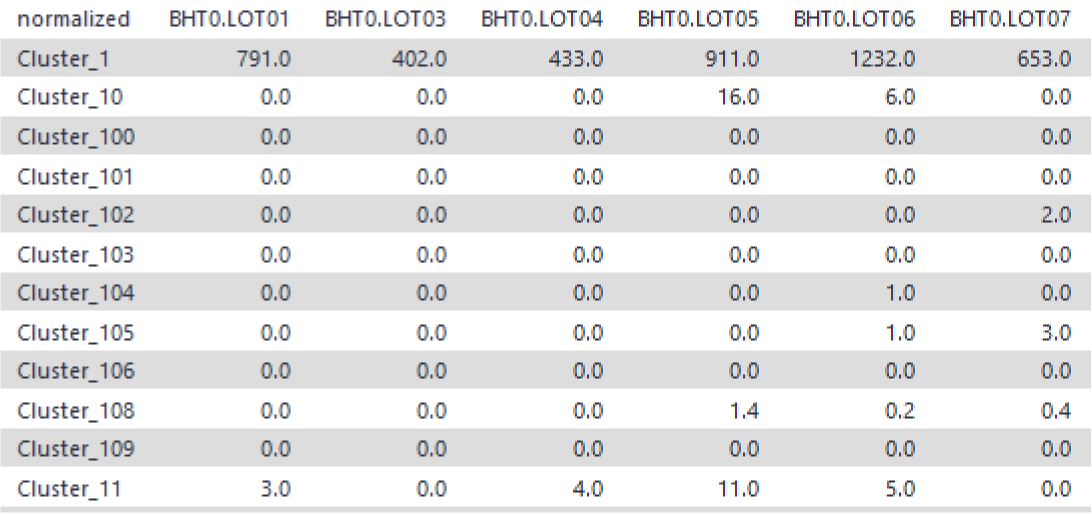
From this table of abundance it is quite possible to make statistical analyses on it to understand the information.

Output
Function excluded

Cluster	FROGS_taxonomy	PICRUSt2_taxonomy	NSTI
Cluster_107	Bacteria;Firmicutes;Bacilli;Mycoplasmatales;Mycoplasmataceae;Candidatus Bacilloplasma;unknown species	Bacteria;Tenericutes;Mollicutes;Haloplasmatales;Haloplasmataceae;Haloplasma;Haloplasma_contractile	0.63742
Cluster_122	Bacteria;Firmicutes;Bacilli;Erysipelotrichales;Erysipelotrichaceae;Faecalitalea;Multi-affiliation	Bacteria;Firmicutes;Bacilli;Lactobacillales;Enterococcaceae;Enterococcus;Enterococcus_pallens	0.679824
Cluster_124	Bacteria;Firmicutes;Clostridia;Christensenellales;Christensenellaceae;Christensenellaceae R-7 group;unknown species	Bacteria;Firmicutes;Clostridia;Eubacteriales;Christensenellaceae;Christensenella;Christensenella_minuta	0.420828
Cluster_144	Bacteria;Firmicutes;Clostridia;Clostridia UCG-014;unknown family;unknown genus;gut metagenome	Bacteria;Firmicutes;Clostridia;Eubacteriales;Clostridiaceae;Butyricicoccus;Butyricicoccus_pullicaecorum	0.726233
Cluster_152	Bacteria;Spirochaetota;Spirochaetia;Spirochaetales;Spirochaetaceae;unknown genus;unknown species	Bacteria;Spirochaetes;Spirochaetia;Spirochaetales;Spirochaetaceae;Treponema;Treponema_caldarium	0.53047
Cluster_16	Bacteria;Fusobacteriota;Fusobacteriia;Fusobacteriales;Leptotrichiaceae;Hypnocyclicus;unknown species	Bacteria;Firmicutes;Clostridia;Eubacteriales;unclassified;unclassified;Clostridiales_bacterium_Firm_08	1.3093270000000001
Cluster_166	Bacteria;Bacteroidota;Bacteroidia;Bacteroidales;Rikenellaceae;Rikenellaceae RC9 gut group;unknown species	Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Odoribacteraceae;Butyricimonas;Butyricimonas_virosa	0.959354
Cluster_172	Bacteria;Actinobacteriota;Actinobacteria;Actinomycetales;Actinomycetaceae;Actinomyces;Actinomyces viscosus	Bacteria;Actinobacteria;Actinomycetia;Micromonosporales;Micromonosporaceae;Asanoa;Asanoa_ishikariensis	0.48663500000000004
Cluster_178	Bacteria;Patescibacteria;Saccharimonadia;Saccharimonadales;Saccharimonadaceae;Candidatus Saccharimonas;unknown species	Bacteria;Proteobacteria;Betaproteobacteria;Burkholderiales;Comamonadaceae;Acidovorax;Acidovorax_citrulli	0.7926350000000001
Cluster_189	Bacteria;Bacteroidota;Bacteroidia;Flavobacteriales;Weeksellaceae;unknown genus;unknown species	Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;Weeksellaceae;Apibacter;Apibacter_mensalis	0.381465
Cluster_204	Bacteria;Fusobacteriota;Fusobacteriia;Fusobacteriales;Fusobacteriaceae;Fusobacterium;Multi-affiliation	Bacteria;Firmicutes;Negativicutes;Veillonellales;Veillonellaceae;Veillonella;Veillonella_magna	1.281097
Cluster_234	Bacteria;Bacteroidota;Bacteroidia;Bacteroidales;Prevotellaceae;Alloprevotella;unknown species	Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Prevotellaceae;Alloprevotella;Alloprevotella_rava	0.6157819999999999
Cluster_25	Bacteria;Firmicutes;Bacilli;Mycoplasmatales;Mycoplasmataceae;Candidatus Bacilloplasma;unknown species	Bacteria;Firmicutes;Clostridia;Eubacteriales;unclassified;unclassified;Clostridiales_bacterium_VE202-01	0.694566
Cluster_252	Bacteria;Actinobacteriota;Coriobacteriia;Coriobacteriales;Atopobiaceae;Olsenella;Olsenella umbonata	Bacteria;Actinobacteria;Actinomycetia;Nakamurellales;Nakamurellaceae;Nakamurella;Nakamurella_multipartita	0.7311479999999999
Cluster_266	Bacteria;Patescibacteria;Gracilibacteria;Candidatus Peribacteria;unknown family;unknown genus;metagenome	Bacteria;Proteobacteria;Epsilonproteobacteria;Nautiliales;Nitratiruptoraceae;Nitratiruptor;Nitratiruptor_sp.	1.027285
Cluster_355	Bacteria;Actinobacteriota;Acidimicrobiia;Microtrichales;Ilumatobacteraceae;unknown genus;unknown species	Bacteria;Firmicutes;Clostridia;Eubacteriales;Syntrophomonadaceae;Dethiobacter;Dethiobacter_alkaliphilus	0.876819
Cluster_358	Bacteria;Actinobacteriota;Coriobacteriia;Coriobacteriales;Atopobiaceae;Olsenella;unknown species	Bacteria;Actinobacteria;Actinomycetia;Streptomycetales;Streptomycetaceae;Streptomyces;Streptomyces_sp.	0.835499
Cluster_373	Bacteria;Actinobacteriota;Actinobacteria;Actinomycetales;Actinomycetaceae;Actinomyces;Actinomyces sp.	Bacteria;Actinobacteria;Actinomycetia;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Propionibacterium_humerusii	0.539983
Cluster_39			
Cluster_40			
Cluster_41			
Cluster_42			
Cluster_43			
Cluster_438	Bacteria;Actinobacteriota;Actinobacteria;Micrococcales;Microbacteriaceae;Microbacterium;Microbacterium phyllosphaerae	Bacteria;Actinobacteria;Actinomycetia;Propionibacteriales;Propionibacteriaceae;Microlunatus;Microlunatus_sagamiharensis	0.41153599999999996
Cluster_453	Bacteria;Firmicutes;Bacilli;Mycoplasmatales;Mycoplasmataceae;unknown genus;unknown species	Bacteria;Firmicutes;Clostridia;Eubacteriales;unclassified;unclassified;Clostridiales_bacterium_VE202-01	0.713615
Cluster_46	Bacteria;Fusobacteriota;Fusobacteriia;Fusobacteriales;Fusobacteriaceae;Cetobacterium;unknown species	Bacteria;Firmicutes;Clostridia;Eubacteriales;unclassified;unclassified;Clostridiales_bacterium_Firm_08	1.377319
Cluster_47	Bacteria;Bacteroidota;Bacteroidia;Bacteroidales;Rikenellaceae;Rikenellaceae RC9 gut group;unknown species	Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Rikenellaceae;Alistipes;Alistipes_inops	1.044139
Cluster_496	Bacteria;Firmicutes;Clostridia;Peptostreptococcales-Tissierellales;Peptostreptococcaceae;Proteocatella;Proteocatella sphenisci	Bacteria;Firmicutes;Clostridia;Eubacteriales;Peptostreptococcaceae;Paeniclostridium;Paeniclostridium_sordellii	0.432676
Cluster_7	Bacteria;Firmicutes;Bacilli;Erysipelotrichales;Erysipelotrichaceae;ZOR0006;unknown species	Bacteria;Firmicutes;Clostridia;Eubacteriales;Christensenellaceae;Christensenella;Christensenella_minuta	0.925652
Cluster_77	Bacteria;Actinobacteriota;Actinobacteria;Actinomycetales;Actinomycetaceae;Trueperella;Multi-affiliation	Bacteria;Actinobacteria;Actinomycetia;Streptomycetales;Streptomycetaceae;Streptomyces;Candidatus_Streptomyces_massiliensis	0.467055
Cluster_97	Bacteria;Fusobacteriota;Fusobacteriia;Fusobacteriales;Fusobacteriaceae;Cetobacterium;unknown species	Bacteria;Firmicutes;Clostridia;Eubacteriales;unclassified;unclassified;Clostridiales_bacterium_Firm_08	1.273517

Information (FROGS taxonomy, PICRUSt2 taxonomy, NSTI) about removed sequences that have a NSTI value above the NSTI threshold chosen in this step.

Output functions_marker_norm



normalized	BHT0.LOT01	BHT0.LOT03	BHT0.LOT04	BHT0.LOT05	BHT0.LOT06	BHT0.LOT07
Cluster_1	791.0	402.0	433.0	911.0	1232.0	653.0
Cluster_10	0.0	0.0	0.0	16.0	6.0	0.0
Cluster_100	0.0	0.0	0.0	0.0	0.0	0.0
Cluster_101	0.0	0.0	0.0	0.0	0.0	0.0
Cluster_102	0.0	0.0	0.0	0.0	0.0	2.0
Cluster_103	0.0	0.0	0.0	0.0	0.0	0.0
Cluster_104	0.0	0.0	0.0	0.0	1.0	0.0
Cluster_105	0.0	0.0	0.0	0.0	1.0	3.0
Cluster_106	0.0	0.0	0.0	0.0	0.0	0.0
Cluster_108	0.0	0.0	0.0	1.4	0.2	0.4
Cluster_109	0.0	0.0	0.0	0.0	0.0	0.0
Cluster_11	3.0	0.0	4.0	11.0	5.0	0.0

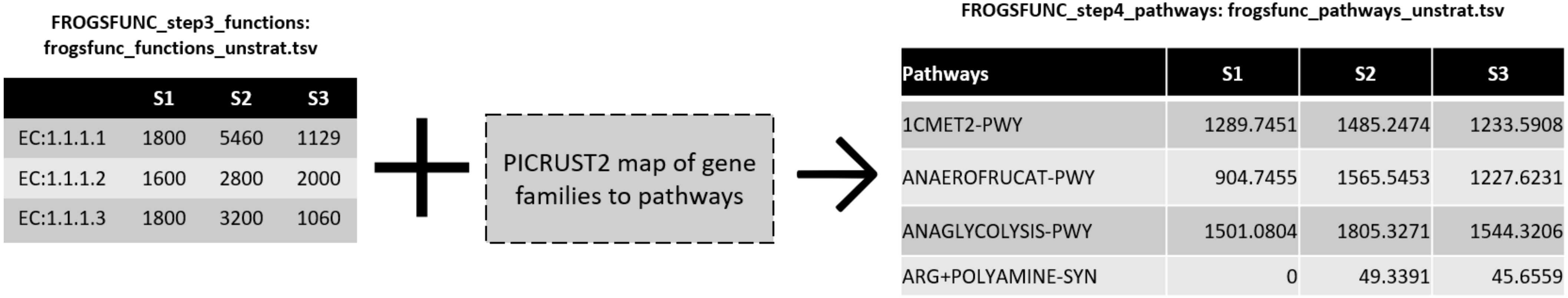
16S/ITS/18S abundance table normalized by the number of markers per genome.

FROGSFUNC_step4_pathways

FROGSFUNC_step4_pathways

FROGSFUNC_step4_pathways is the fourth step of PICRUSt2.
It infers MetaCyc/KEGG pathway abundances based on EC/KO number abundances.
There are 3 steps performed at this stage:

- Regroups EC/KO numbers to MetaCyc/KEGG reactions.
- Infers which MetaCyc/KEGG pathways are present based on these reactions with MinPath.
- Calculates and returns the abundance of pathways identified as present.



FROGSFUNC_step3_functions: frogsfunc_functions_unstrat.tsv

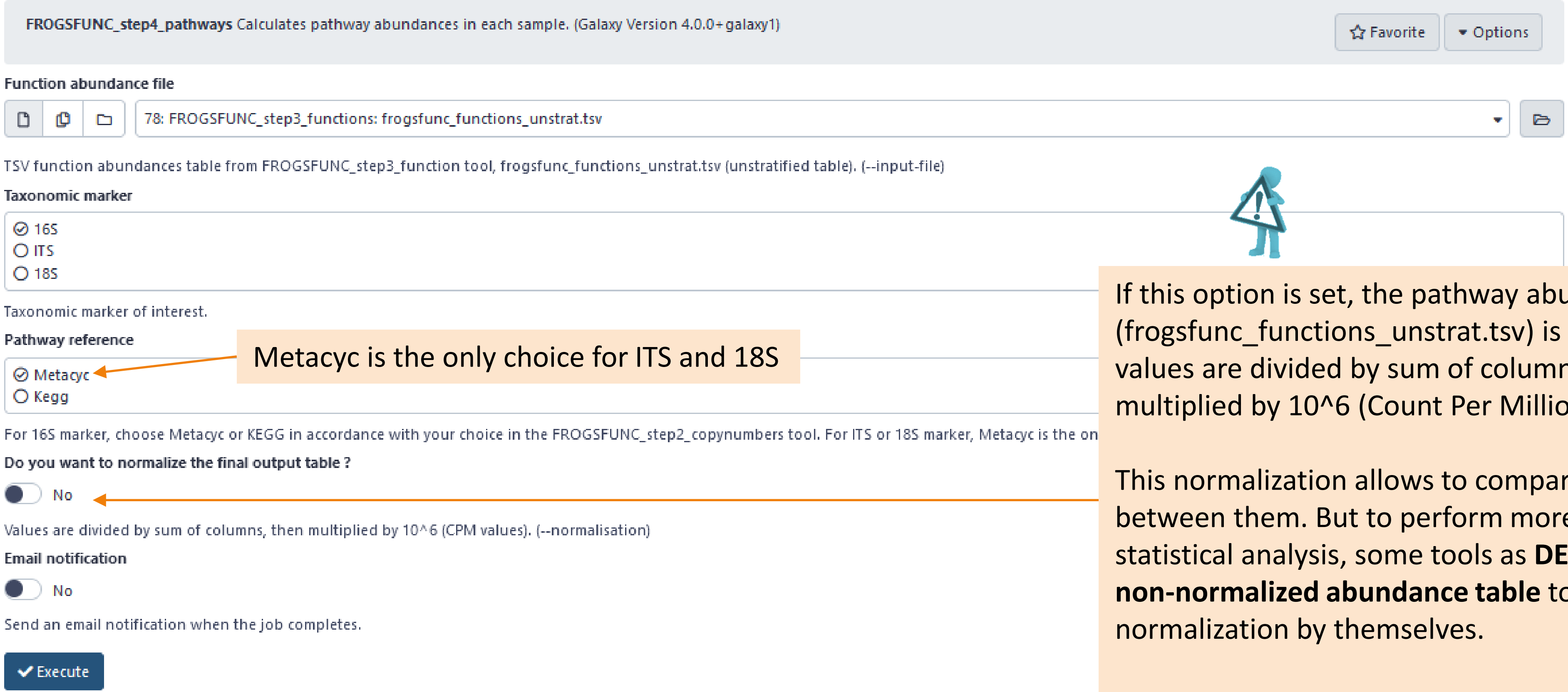
	S1	S2	S3
EC:1.1.1.1	1800	5460	1129
EC:1.1.1.2	1600	2800	2000
EC:1.1.1.3	1800	3200	1060

PICRUST2 map of gene families to pathways

FROGSFUNC_step4_pathways: frogsfunc_pathways_unstrat.tsv

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

FROGSFUNC_step4_pathways



FROGSFUNC_step4_pathways Calculates pathway abundances in each sample. (Galaxy Version 4.0.0+galaxy1)

Function abundance file

78: FROGSFUNC_step3_functions: frogsfunc_functions_unstrat.tsv

TSV function abundances table from FROGSFUNC_step3_function tool, frogsfunc_functions_unstrat.tsv (unstratified table). (--input-file)

Taxonomic marker

☑ 16S
○ ITS
○ 18S

Taxonomic marker of interest.

Pathway reference

☑ Metacyc
○ Kegg

For 16S marker, choose Metacyc or KEGG in accordance with your choice in the FROGSFUNC_step2_copynumbers tool. For ITS or 18S marker, Metacyc is the on

Do you want to normalize the final output table ?

No

Values are divided by sum of columns, then multiplied by 10^6 (CPM values). (--normalisation)

Email notification

No

Send an email notification when the job completes.

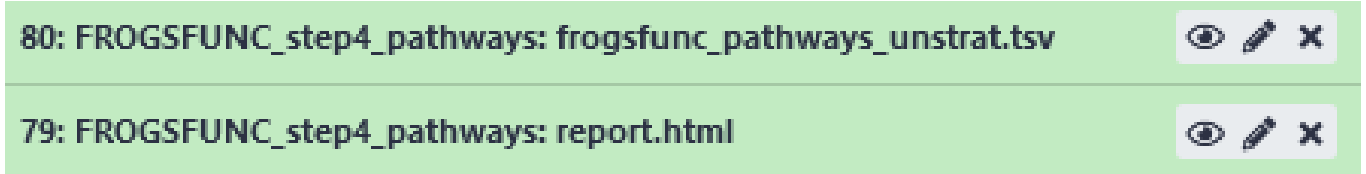
Metacyc is the only choice for ITS and 18S

If this option is set, the pathway abundances file (frogsfunc_functions_unstrat.tsv) is normalized: values are divided by sum of columns, then multiplied by 10^6 (Count Per Million values).

This normalization allows to compare the samples between them. But to perform more precise statistical analysis, some tools as **DESeq2 need the non-normalized abundance table** to perform the normalization by themselves.

So be careful which table to use for further analysis.

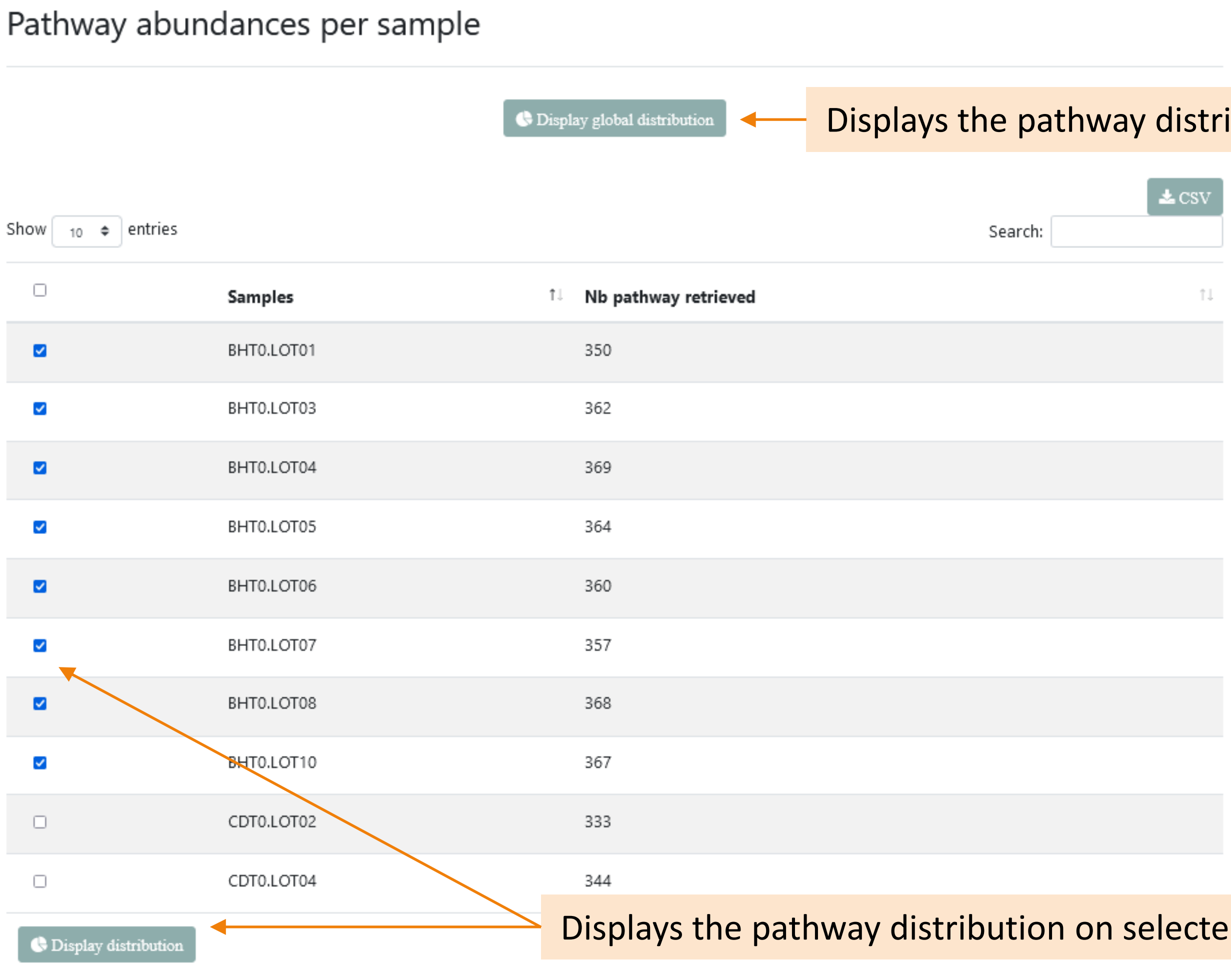
Outputs



80: FROGSFUNC_step4_pathways: frogsfunc_pathways_unstrat.tsv

79: FROGSFUNC_step4_pathways: report.html

FROGSFUNC_step4_pathways: report.html



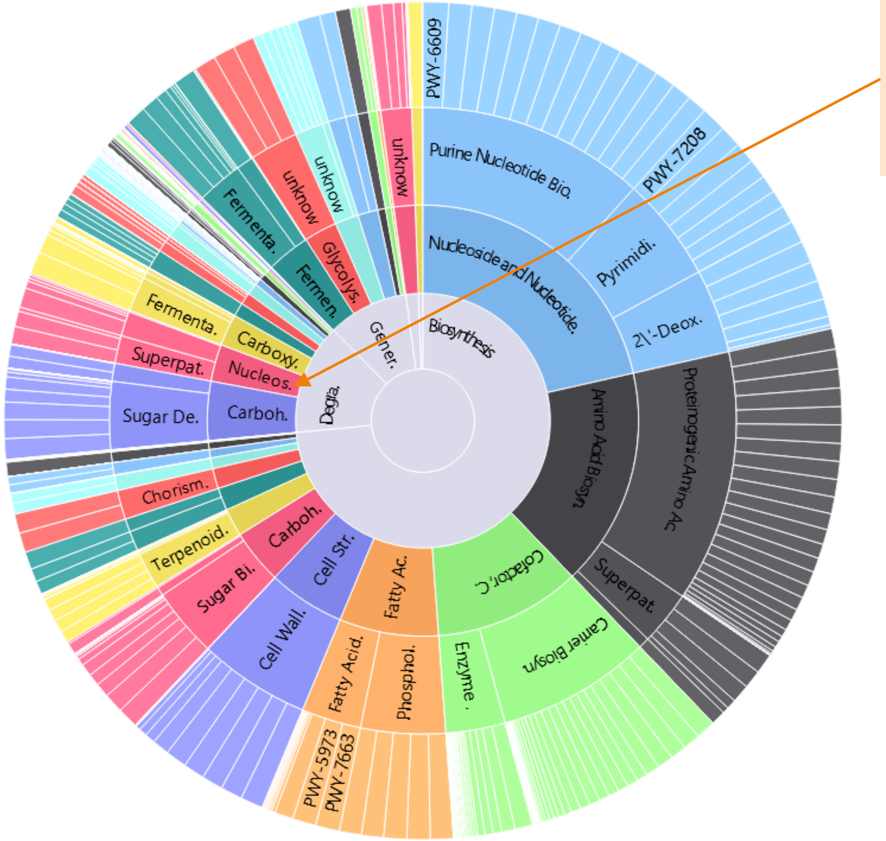
Pathway abundances per sample

☐	Samples	Nb pathway retrieved
☑	BHT0.LOT01	350
☑	BHT0.LOT03	362
☑	BHT0.LOT04	369
☑	BHT0.LOT05	364
☑	BHT0.LOT06	360
☑	BHT0.LOT07	357
☑	BHT0.LOT08	368
☑	BHT0.LOT10	367
☐	CDT0.LOT02	333
☐	CDT0.LOT04	344

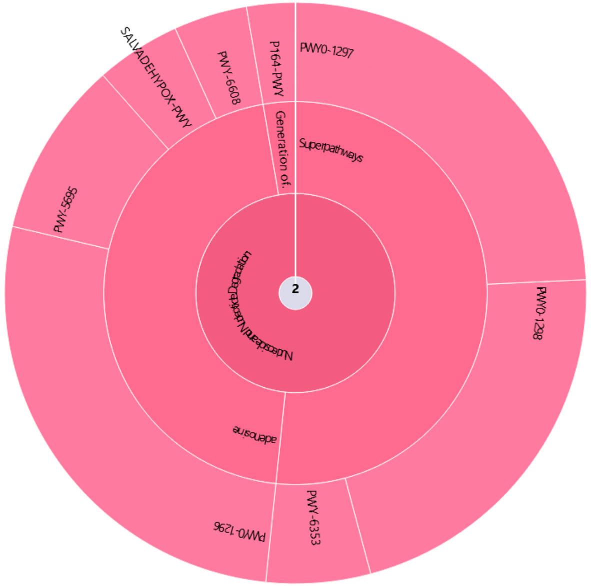
Displays the pathway distribution on all samples

Displays the pathway distribution on selected samples

What is the distribution of pathway abundances in the samples ?

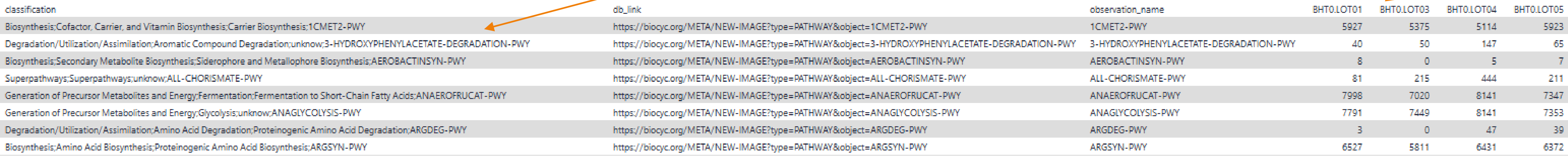


For more pathway details, double-click on a the interest pathway name.



Output
Pathway abundance tables

Abundance of each pathway inside each sample.



classification	db_link	observation_name	BHT0.LOT01	BHT0.LOT03	BHT0.LOT04	BHT0.LOT05
Biosynthesis;Cofactor, Carrier, and Vitamin Biosynthesis;Carrier Biosynthesis;1CMET2-PWY	https://biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=1CMET2-PWY	1CMET2-PWY	5927	5375	5114	5923
Degradation/Utilization/Assimilation;Aromatic Compound Degradation;unknow;3-HYDROXYPHENYLACETATE-DEGRADATION-PWY	https://biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=3-HYDROXYPHENYLACETATE-DEGRADATION-PWY	3-HYDROXYPHENYLACETATE-DEGRADATION-PWY	40	50	147	65
Biosynthesis;Secondary Metabolite Biosynthesis;Siderophore and Metallophore Biosynthesis;AEROBACTINSYN-PWY	https://biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=AEROBACTINSYN-PWY	AEROBACTINSYN-PWY	8	0	5	7
Superpathways;Superpathways;unknow;ALL-CHORISMATE-PWY	https://biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=ALL-CHORISMATE-PWY	ALL-CHORISMATE-PWY	81	215	444	211
Generation of Precursor Metabolites and Energy;Fermentation;Fermentation to Short-Chain Fatty Acids;ANAEROFRUCAT-PWY	https://biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=ANAEROFRUCAT-PWY	ANAEROFRUCAT-PWY	7998	7020	8141	7347
Generation of Precursor Metabolites and Energy;Glycolysis;unknow;ANAGLYCOLYSIS-PWY	https://biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=ANAGLYCOLYSIS-PWY	ANAGLYCOLYSIS-PWY	7791	7449	8141	7353
Degradation/Utilization/Assimilation;Amino Acid Degradation;Proteinogenic Amino Acid Degradation;ARGDEG-PWY	https://biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=ARGDEG-PWY	ARGDEG-PWY	3	0	47	39
Biosynthesis;Amino Acid Biosynthesis;Proteinogenic Amino Acid Biosynthesis;ARGSYN-PWY	https://biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=ARGSYN-PWY	ARGSYN-PWY	6527	5811	6431	6372

External link on the pathway

From this table of abundance it is quite possible to make statistical analyses on it to understand the information.

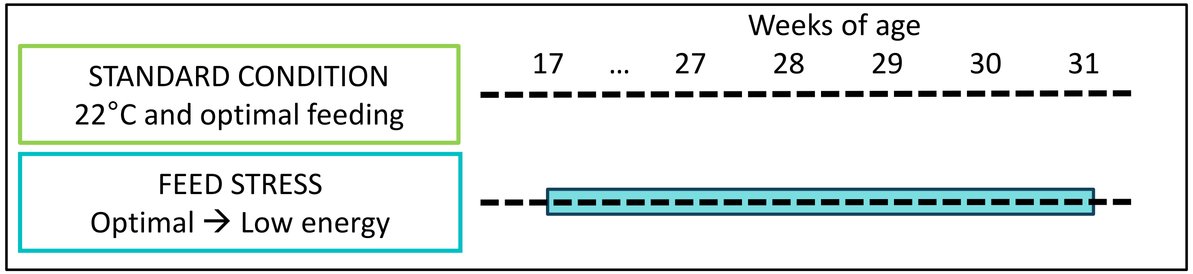
Application example

Impact of host genetics and abiotic stresses on caecal microbiota composition in four different laying hen lines?

R+:
Low efficient bird

Divergently
selected on feed
efficiency
since 40 years

R-:
High efficient bird



STANDARD CONDITION
22°C and optimal feeding

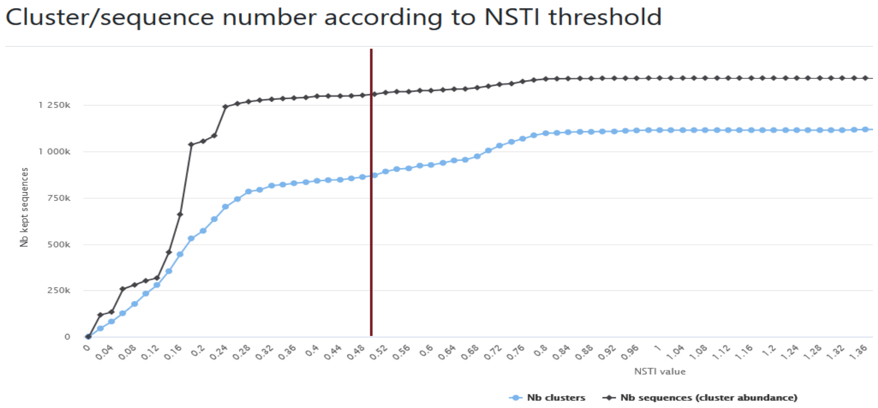
FEED STRESS
Optimal → Low energy

Weeks of age

FROGS results:

FROGS analysis reconstructed a microbiota including 601 OTUs, of which 93 are retained as differentially abundant between lineages and/or between diets.

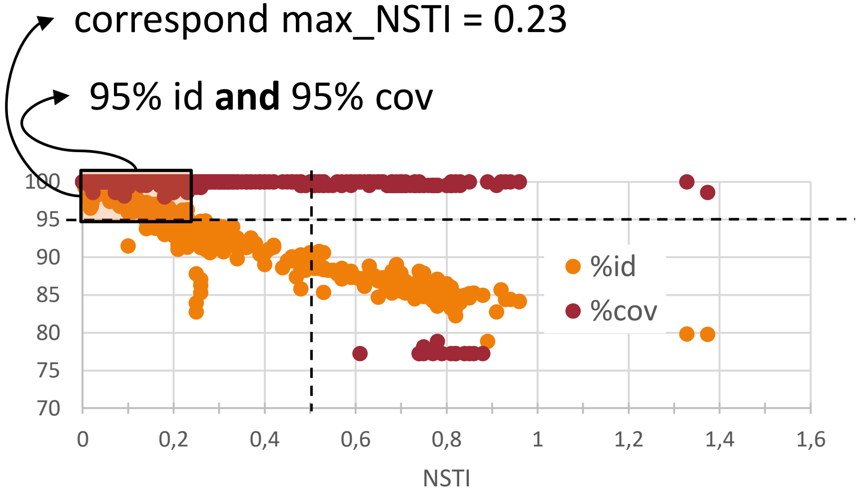
FROGSFUNC -> filter on NSTI and %id and %cov



NSTI < 0,5:
468 OTUs 77,8%
93,7% des séquences

correspond max_NSTI = 0.23

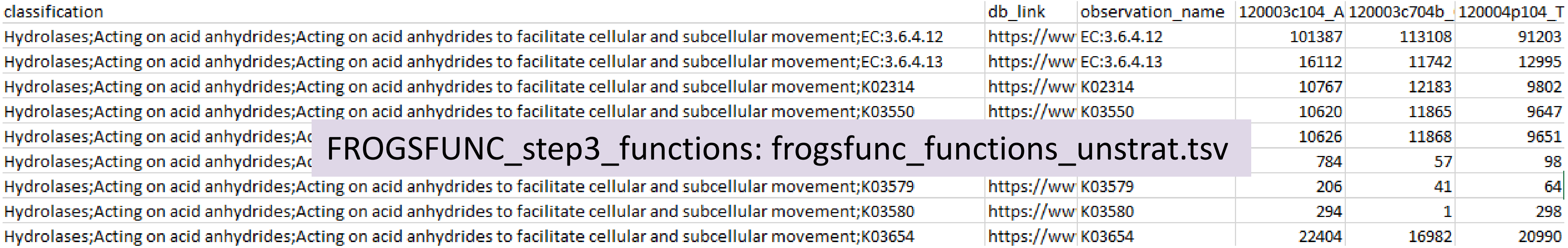
95% id **and** 95% cov



id% & cov% > 95% :
266 OTUs 44,2% -> 48,8 % of sequences

The higher the NSTI, the lower the %id and %cov.
It is therefore interesting to filter also on these criteria.

FROGSFUNC → Statistics



classification	db_link	observation_name	120003c104_A	120003c704b_	120004p104_T
Hydrolases;Acting on acid anhydrides;Acting on acid anhydrides to facilitate cellular and subcellular movement;EC:3.6.4.12	https://ww	EC:3.6.4.12	101387	113108	91203
Hydrolases;Acting on acid anhydrides;Acting on acid anhydrides to facilitate cellular and subcellular movement;EC:3.6.4.13	https://ww	EC:3.6.4.13	16112	11742	12995
Hydrolases;Acting on acid anhydrides;Acting on acid anhydrides to facilitate cellular and subcellular movement;K02314	https://ww	K02314	10767	12183	9802
Hydrolases;Acting on acid anhydrides;Acting on acid anhydrides to facilitate cellular and subcellular movement;K03550	https://ww	K03550	10620	11865	9647
Hydrolases;Acting on acid anhydrides;Ac			10626	11868	9651
Hydrolases;Acting on acid anhydrides;Ac			784	57	98
Hydrolases;Acting on acid anhydrides;Acting on acid anhydrides to facilitate cellular and subcellular movement;K03579	https://ww	K03579	206	41	64
Hydrolases;Acting on acid anhydrides;Acting on acid anhydrides to facilitate cellular and subcellular movement;K03580	https://ww	K03580	294	1	298
Hydrolases;Acting on acid anhydrides;Acting on acid anhydrides to facilitate cellular and subcellular movement;K03654	https://ww	K03654	22404	16982	20990

FROGSFUNC_step3_functions: frogsfunc_functions_unstrat.tsv

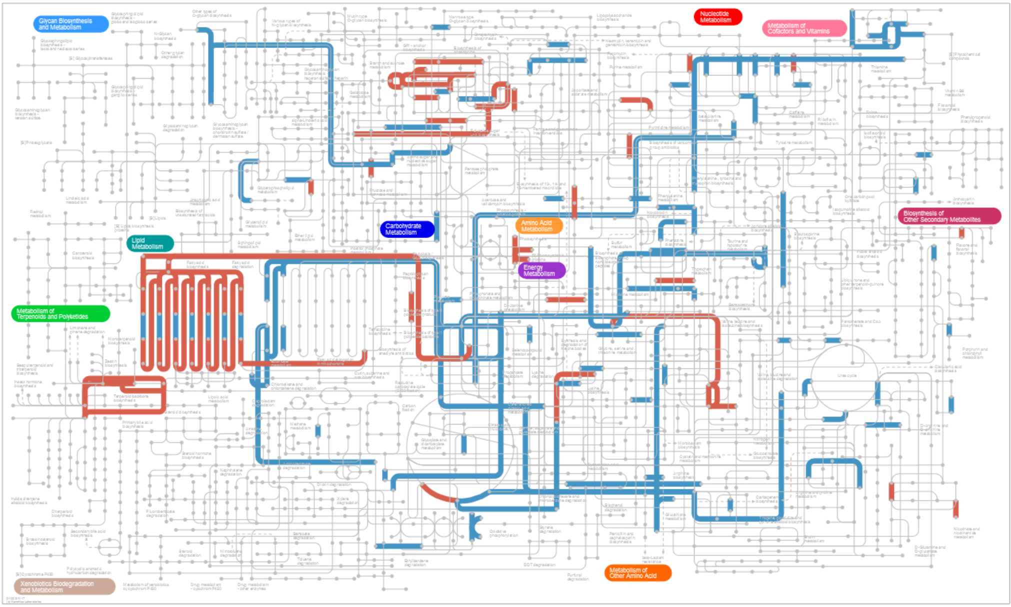
Stat analysis done on differential analysis of OTUs:
DESeq2 in 4 steps under R:
Comparison of 2 lines (R+/R-) in each condition.
Comparison of 2 (ctrl/low energy) diets in each line.

NB: use of the more conservative pvalues correction method of Benjamini & Yekutieli

Results on iPATH3 map

The differential analyses follow the same pattern at the level of OTUs and functions:

1. No difference between lines by modifying the diet
2. No difference between lines when animals were fed with the low-energy diet
3. In control diet, differentially abundant KEGG functions between R+ and R-



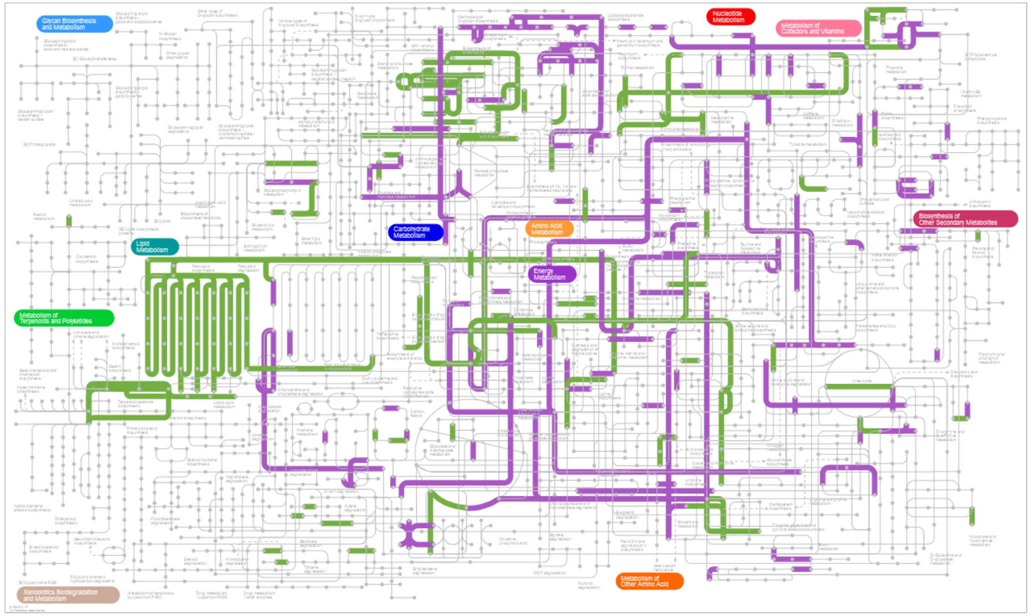
fatty acid biosynthesis, starch and sucrose, amino sugar and nucleotide sugar metabolisms

carbohydrate metabolism (in particular pyruvate, and glycolysis), fatty acid biosynthesis and degradation, and protein metabolism with various amino acid metabolisms.

Interestingly, the animal of these two lines also present differences in terms of lipid metabolism

Results on iPATH3 map

4. In R+ line, differentially abundant KEGG functions between control and low diets



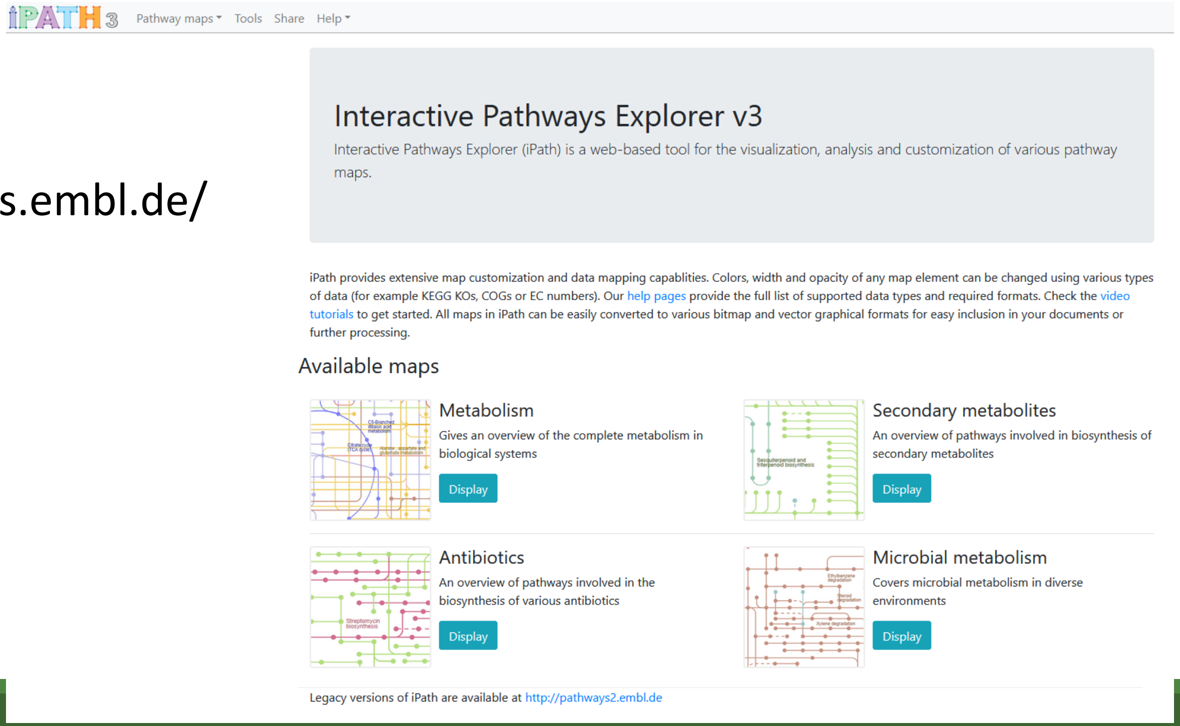
fatty acid biosynthesis, starch and sucrose, amino sugar and nucleotide sugar metabolisms

carbohydrate metabolism, protein metabolism with various amino acid metabolisms, and various other metabolisms

Future developments

- Adaptation of FROGSSTAT tools to be used by FROGSFUNC outputs
- News outputs from FROGSFUNC to explore iPATH3 maps
- Adaptation of FROGSFUNC to filter on %id and %cov

https://pathways.embl.de/



Interactive Pathways Explorer v3

Interactive Pathways Explorer (iPath) is a web-based tool for the visualization, analysis and customization of various pathway maps.

iPath provides extensive map customization and data mapping capablities. Colors, width and opacity of any map element can be changed using various types of data (for example KEGG KOs, COGs or EC numbers). Our help pages provide the full list of supported data types and required formats. Check the video tutorials to get started. All maps in iPath can be easily converted to various bitmap and vector graphical formats for easy inclusion in your documents or further processing.

Available maps

Metabolism
Gives an overview of the complete metabolism in biological systems

Secondary metabolites
An overview of pathways involved in biosynthesis of secondary metabolites

Antibiotics
An overview of pathways involved in the biosynthesis of various antibiotics

Microbial metabolism
Covers microbial metabolism in diverse environments

Legacy versions of iPath are available at http://pathways2.embl.de