



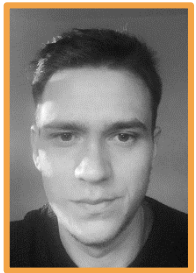
Training on Galaxy: Metabarcoding

June 2022 - Webinar

FROGS Practice on 16S data

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

Who is in the current FROGS group?



Vincent DARBOT



Maria BERNARD



Olivier RUÉ

Developers



Lucas AUER



Laurent CAUQUIL

Biology experts



Patrice DÉHAIS

Galaxy
support



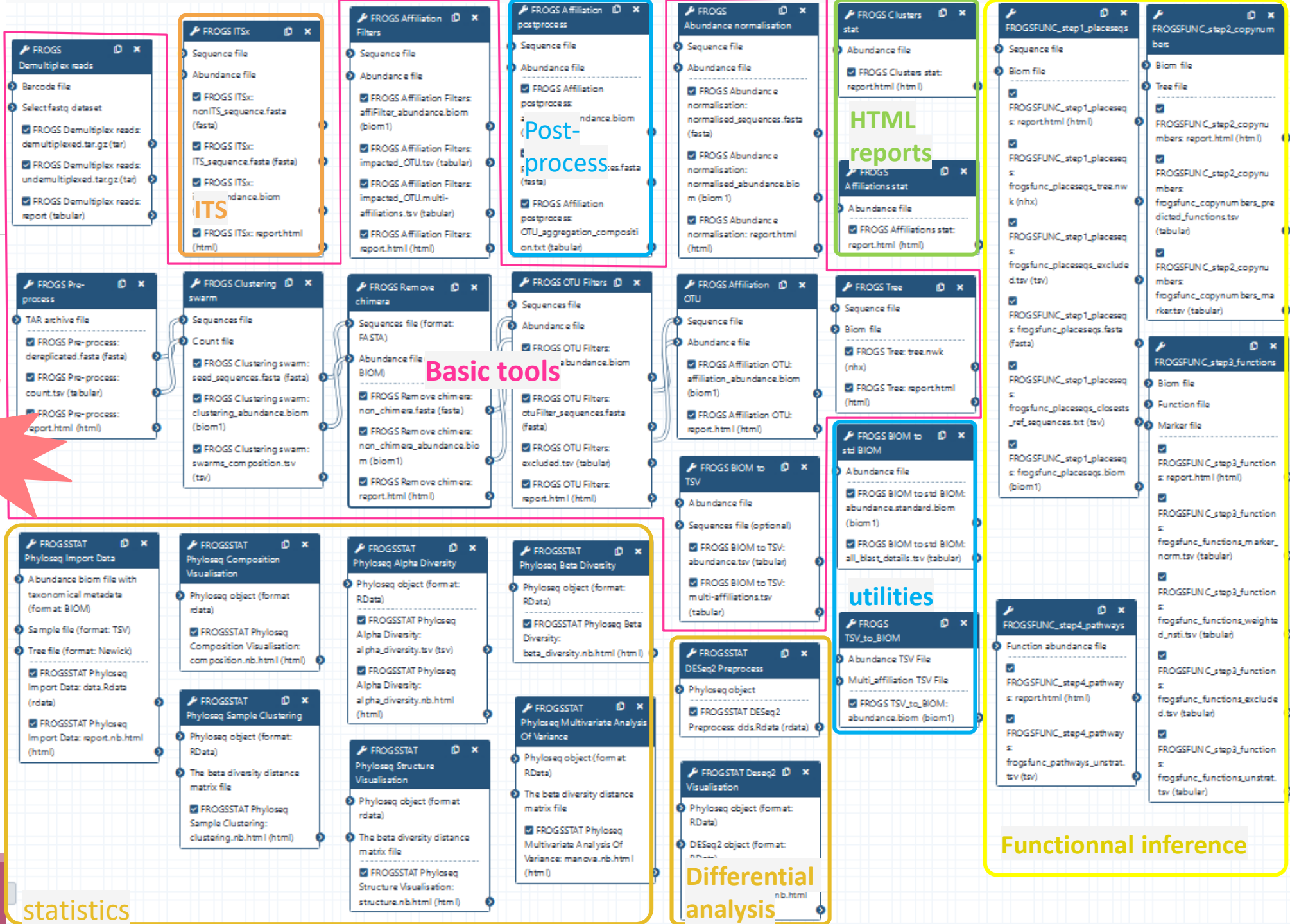
Mahendra
MARIADASSOU

Statistical expert



Géraldine
PASCAL

Coordinator





FROGS
Demultiplex reads

- Barcode file
- Select fastq dataset
- FROGS Demultiplex reads: demultiplexed.tar.gz (tar)
- FROGS Demultiplex reads: undemultiplexed.tar.gz (tar)
- FROGS Demultiplex reads: report (tabular)

FROGS ITSx

- Sequence file
- Abundance file
- FROGS ITSx: nonITS_sequences.fasta (fasta)
- FROGS ITSx: ITS_sequences.fasta (fasta)
- FROGS ITSx: abundance.biom
- FROGS ITSx: report.html (html)

ITS

FROGS Affiliation

- Sequence file
- Abundance file
- FROGS Affiliation Filters: affFilters_abundance.biom (biom1)
- FROGS Affiliation Filters: impacted_OTU_multi-affiliations.tsv (tabular)
- FROGS Affiliation Filters: report.html (html)

FROGS Affiliation postprocess

- Sequence file
- Abundance file
- FROGS Affiliation postprocess: abundance.biom
- FROGS Affiliation postprocess: sequences.fasta (fasta)
- FROGS Affiliation postprocess: OTU_aggregation_composition.txt (tabular)

NEW

(Post-process)

FROGS Abundance normalisation

- Sequence file
- Abundance file
- FROGS Abundance normalisation: normalised_sequences.fasta (fasta)
- FROGS Abundance normalisation: normalised_sequences.biom (biom)
- FROGS Abundance normalisation: report.html (html)

NEW

FROGS Clusters

- Abundance file
- FROGS Clusters stat: report.html (html)

HTML reports

NEW

FROGSFUNC_step1_placeseqs

- Sequence file
- Biom file
- FROGSFUNC_step1_placeseq s: report.html (html)
- FROGSFUNC_step1_placeseq s: frogsfunc_placeseqs_tree.nwk (nhx)
- FROGSFUNC_step1_placeseq s: frogsfunc_placeseqs_exclud ed.tsv (tsv)
- FROGSFUNC_step1_placeseq s: frogsfunc_placeseqs.fasta (fasta)
- FROGSFUNC_step1_placeseq s: frogsfunc_placeseqs_closest s_ref_sequences.txt (tsv)
- FROGSFUNC_step1_placeseq s: frogsfunc_placeseqs.biom (biom1)

FROGSFUNC_step2_copynum bers

- Biom file
- Tree file
- FROGSFUNC_step2_copynu mbers: report.html (html)
- FROGSFUNC_step2_copynu mbers: frogsfunc_copynum bers_pre dicted_functions.tsv (tabular)
- FROGSFUNC_step2_copynu mbers: frogsfunc_copynum bers_ma rker.tsv (tabular)

FROGS Pre-process

- TAR archive file
- FROGS Pre-process: dereplicated (fasta)
- FROGS Pre-process: count.txt
- FROGS Pre-process: report.html (html)

NEW

FROGS Clustering swarm

- Sequences file
- Count file
- FROGS Clustering swarm: seed_sequences.fasta (fasta)
- FROGS Clustering swarm: clustering.swarm (biom)
- FROGS Clustering swarm: swarms_composition.tsv (tsv)

NEW

FROGS Remove chimera

- Sequences file (format: FASTA)
- Abundance file (BIOM)
- FROGS Remove chimera: non_chimera.fasta (fasta)
- FROGS Remove chimera: non_chimera_abundance.bio m (biom1)
- FROGS Remove chimera: report.html (html)

FROGS OTU Filters

- Sequences file
- Abundance file
- FROGS OTU Filters: otuFilter_sequences.fasta (fasta)
- FROGS OTU Filters: excluded.tsv (tabular)
- FROGS OTU Filters: report.html (html)

NEW

FROGS Affiliation OTU

- Sequence file
- Abundance file
- FROGS Affiliation OTU: abundance.biom
- FROGS Affiliation OTU: report.html (html)

NEW

FROGS Tree

- Sequence file
- Biom file
- FROGS Tree: report.html (html)
- FROGS Tree: report.html (html)

NEW

Basic tools

FROGSSTAT Phyloseq Import Data

- Abundance biom file with taxonomical metadata (format: BIOM)
- Sample file (format: TSV)
- Tree file (format: Newick)
- FROGSSTAT Phyloseq Import Data: data.Rdata (rdata)
- FROGSSTAT Phyloseq Import Data: report.nb.html (html)

NEW

FROGSSTAT Phyloseq Composition Visualisation

- Phyloseq object (format: Rdata)
- FROGSSTAT Phyloseq Composition Visualisation: composition.nb.html (html)

FROGSSTAT Phyloseq Alpha Diversity

- Phyloseq object (format: RData)
- FROGSSTAT Phyloseq Alpha Diversity: alpha_diversity.tsv (tsv)
- FROGSSTAT Phyloseq Alpha Diversity: alpha_diversity.nb.html (html)

FROGSSTAT Phyloseq Beta Diversity

- Phyloseq object (format: RData)
- FROGSSTAT Phyloseq Beta Diversity: beta_diversity.nb.html (html)

FROGSSTAT Phyloseq Sample Clustering

- Phyloseq object (format: RData)
- The beta diversity distance matrix file
- FROGSSTAT Phyloseq Sample Clustering: clustering.nb.html (html)

FROGSSTAT Phyloseq Structure Visualisation

- Phyloseq object (format: rdata)
- The beta diversity distance matrix file
- FROGSSTAT Phyloseq Structure Visualisation: structure.nb.html (html)

FROGSSTAT Phyloseq Multivariate Analysis Of Variance

- Phyloseq object (format: RData)
- The beta diversity distance matrix file
- FROGSSTAT Phyloseq Multivariate Analysis Of Variance: manova.nb.html (html)

statistics

FROGS BIOM to TSV

- Abundance file
- Sequences file (optional)
- FROGS BIOM to TSV: abundance.tsv (tabular)
- FROGS BIOM to TSV: multi-affiliations.tsv (tabular)

NEW

FROGS BIOM to std BIOM

- Abundance file
- FROGS BIOM to std BIOM: abundance.standard.biom (biom1)
- FROGS BIOM to std BIOM: all_blast_details.tsv (tabular)

utilities

FROGSSTAT DESeq2 Preprocess

- Phyloseq object
- FROGSSTAT DESeq2 Preprocess: rdata (rdata)

NEW

FROGSSTAT DESeq2 Visualisation

- Phyloseq object (format: RData)
- DESeq2 object (format: RData)

Differential analysis

NEW

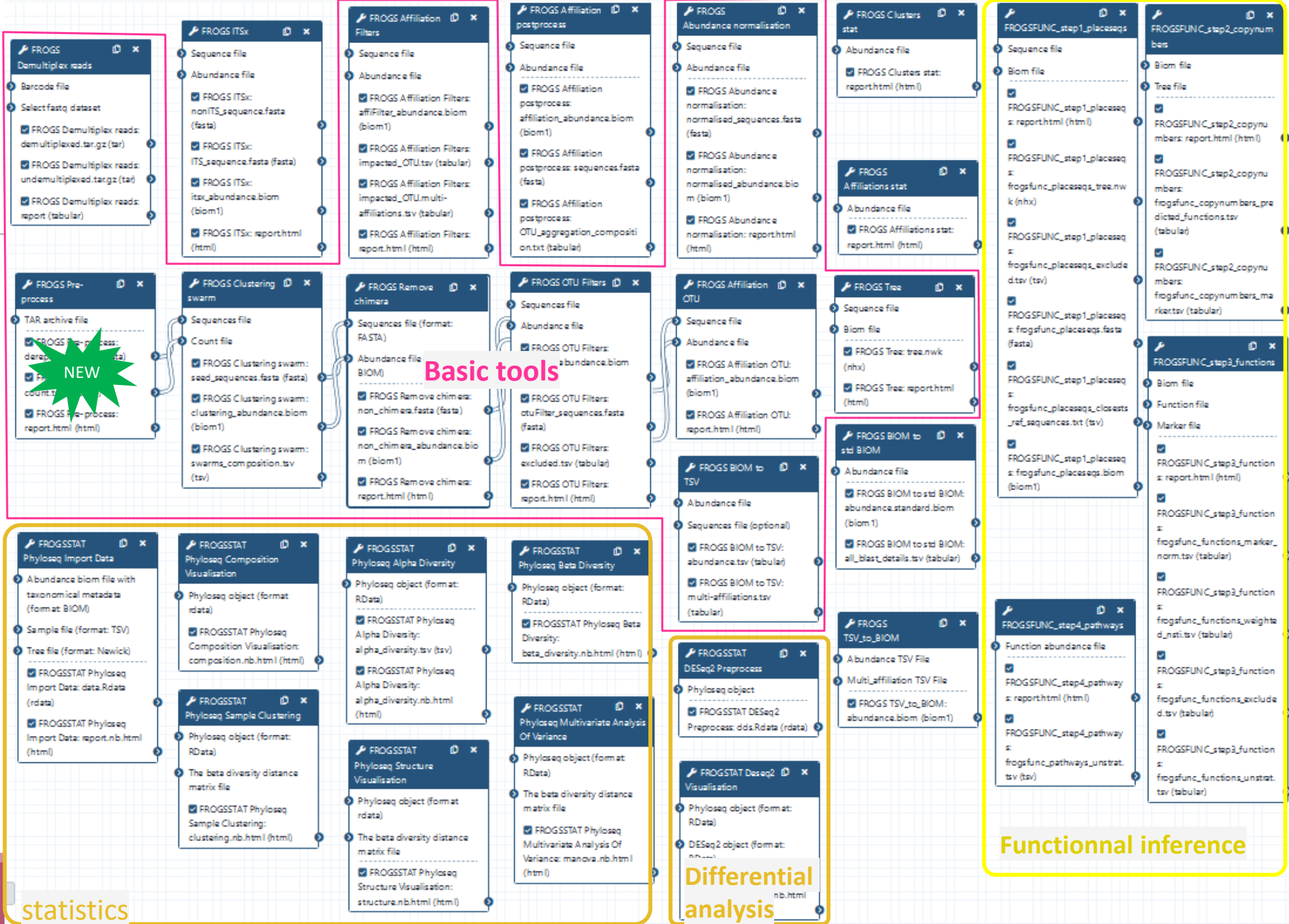
FROGSFUNC_step4_pathways

- Function abundance file
- FROGSFUNC_step4_pathway s: report.html (html)
- FROGSFUNC_step4_pathway s: frogsfunc_pathways_unstrat. tsv (tsv)

FROGSFUNC_step3_functions

- Biom file
- Function file
- Marker file
- FROGSFUNC_step3_function s: report.html (html)
- FROGSFUNC_step3_function s: frogsfunc_functions_marke r_norm.tsv (tabular)
- FROGSFUNC_step3_function s: frogsfunc_functions_marke r_norm.tsv (tabular)
- FROGSFUNC_step3_function s: frogsfunc_functions_marke r_norm.tsv (tabular)
- FROGSFUNC_step3_function s: frogsfunc_functions_marke r_norm.tsv (tabular)

Functional inference



statistics

Differential analysis

Functional inference

Pre-process tool

Pre-process routine

- Merging of R1 and R2 reads
- Delete sequences without good primers
- Finds and removes adapter sequences
- Delete sequence with not expected lengths
- Delete sequences with ambiguous bases (N)
- Dereplication

- + removing homopolymers (size = 8) for 454 data
- + quality filter for 454 data

What does the Pre-process tool do?

- Merging of R1 and R2 reads with **vsearch**, **flash** or **pear** (only in command line)
- Delete sequences without good primers
- Finds and removes adapter sequences with **cutadapt**
- Delete sequence with not expected lengths
- Delete sequences with ambiguous bases (N)
- Dereplication
- + removing homopolymers (size = 8) for 454 data
- + quality filter for 454 data

VSEARCH: a versatile open source tool for metagenomics.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F.
PeerJ. 2016 Oct 18;4:e2584. eCollection 2016.

Bioinformatics (2011) 27 (21):2957-2963. doi:10.1093/bioinformatics/btr507

FLASH: fast length adjustment of short reads to improve genome assemblies

TanjaMagoc, Steven L. Salzberg

Bioinformatics (2014) 30 (5):614–620 doi.org/10.1093/bioinformatics/btt593

PEAR: a fast and accurate Illumina Paired-End reAd merger

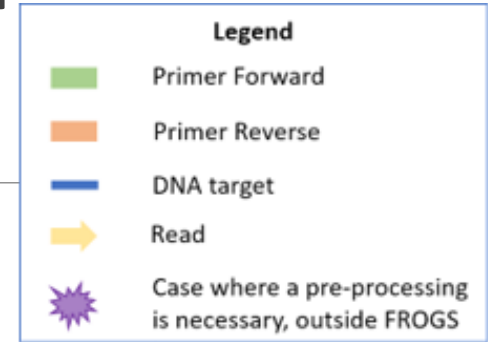
J. Zhang, K. Kobert, T. Flouri, A. Stamatakis,

EMBnet Journal, Vol17 no1. doi : 10.14806/ej.17.1.200

Cutadapt removes adapter sequences from high-throughput sequencing reads

Marcel Martin

Processed data by FROGS in brief



454



illumina

Standard sequencing protocol

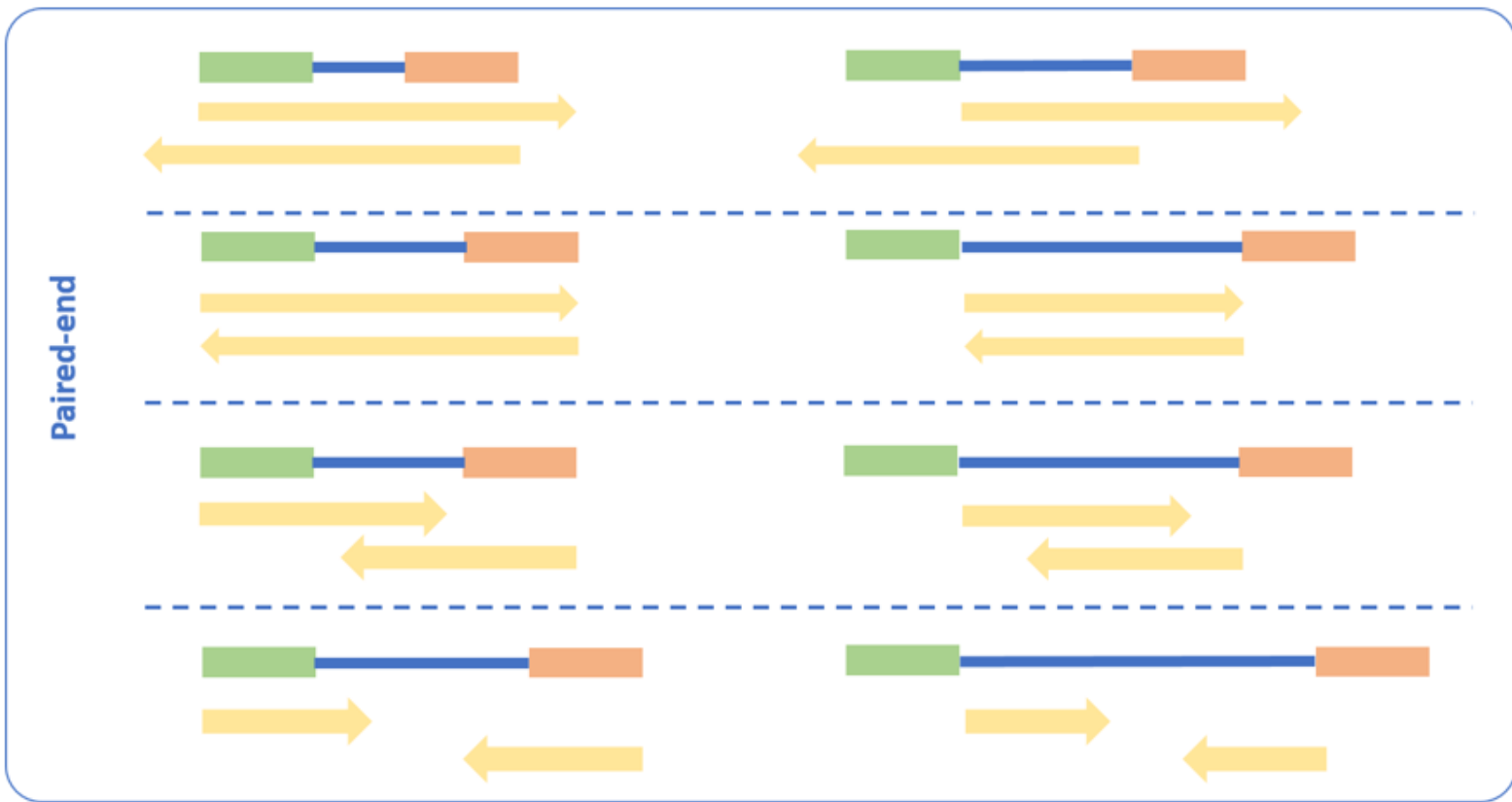
Kozich protocol : primers are not included in reads



→ Remove reverse primer before FROGS processing

Legend

- Primer Forward
- Primer Reverse
- DNA target
- Read
- Case where a pre-processing is necessary, outside FROGS



Length of the sequenced target < length of one read

Supported since version 3.0

Length of the sequenced target < the sum of the lengths of the two reads

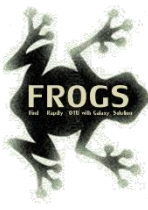
Length of the sequenced target >= the sum of the lengths of the two reads

Supported since version 3.0 with option "keep unmerged reads" in preprocess Tool

Preprocess tool in brief

	Take in charge
Illumina	✓
454	✓
Merged data	✓
Not merged data	✓
Without primers	✓
Only R1 or only R2	⊘
Too distant R1 and R2 to be merged	✓
Over-overlapping R1 R2	✓

	Take in charge
Archive .tar.gz	✓
Fastq	✓
Fasta	⊘
With only 1 primer	⊘
Multiplexed data	⊘
Demultiplexed data	✓



statistics

Differential analysis

Functional inference

Clustering tool

FROGS Clustering swarm Single-linkage clustering on sequences (Galaxy Version 3.2.1) Options

Sequences file

 The dereplicated sequences file (format: fasta).

Count file

 It contains the count by sample for each sequence (format: TSV).

FROGS guidelines version

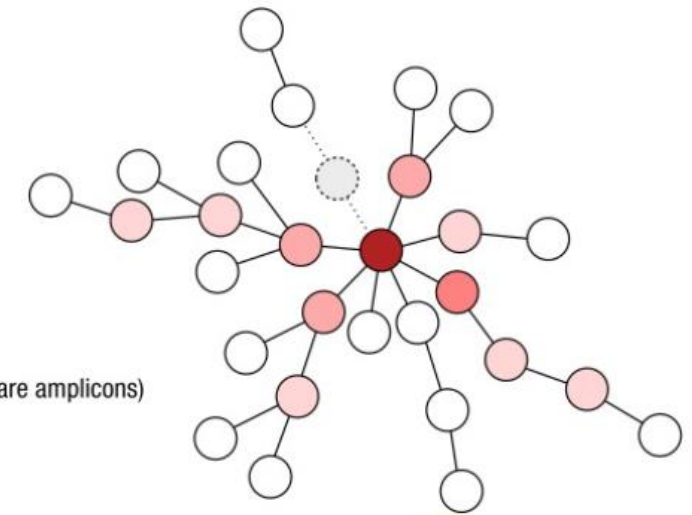
 Denoising step prior to a d3 clustering is no more recommended since FROGS 3.2, but you can still choose it.

Aggregation distance clustering

 Maximum number of differences between sequences in each aggregation swarm step. (recommended d=1)

Refine OTU clustering
 Yes No
 Clustering will be performed with the swarm `--fastidious` option, which is recommended and only usable in association with a distance of 1 (default and recommended: Yes)

longer but more accurate





NEW
Basic tools

statistics

Differential analysis

Functional inference

OTU Filter tool

OTU Filter

Goal: This tool deletes OTU among conditions enter by user. If an OTU reply to at least 1 criteria, the OTU is deleted.

Criteria:

The OTU prevalence: The number of times the OTU is present in the environment, *i.e.* the number of samples where the OTU must be present.

OTU size: An OTU that is not large enough for a given proportion or count will be removed.

Biggest OTU: Only the X biggest are conserved.

Contaminant: If OTU sequence matches with phiX, chloroplastic/mitochondrial 16S of A. Thaliana or your own contaminant sequence.

One tool, 4 criteria

Sequences file

The sequence file to filter (format: FASTA)

Abundance file

The abundance file to filter (format: BIOM)

Minimum prevalence method

1

Minimum prevalence

Fill the field only if you want this treatment. Keep OTU if it is present in at least this number of samples.

Minimum OTU abundancy as proportion or count. We recommend to use a proportion of 0.00005.

2

Minimum proportion of sequences abundancy to keep OTU

Fill the field only if you want this treatment. Example: 0.00005, recommended by Bokulich et al 2013, to keep OTU with at least 0.005% of all sequences (--min_abundance)

N biggest OTUs

3

Fill the fields only if you want this treatment. Keep the N biggest OTU (--nb-biggest-otu)

Search for contaminant OTU.

4

Either you use your own contaminant fasta file or you select one among available ones. (--contaminant)

Email notification

 No

Send an email notification when the job completes.

1

Prevalence filter – option 1

FROGS OTU Filters Filters OTUs on several criteria. (Galaxy Version beta) ☆ Favorite ▼ Options

Sequences file

📁

The sequence file to filter (format: FASTA)

Abundance file

📁

The abundance file to filter (format: BIOM)

Minimum prevalence method

▼

Minimum prevalence

Here, user wants that each OTU are present in at least 4 samples.

Fill the field only if you want this treatment. Keep OTU if it is present in at least this number of samples.

1

Prevalence filter – option 2

FROGS OTU Filters Filters OTUs on several criteria. (Galaxy Version beta) Favorite Options

Sequences file
9: FROGS Remove chimera: non_chimera.fasta
The sequence file to filter (format: FASTA)

Abundance file
10: FROGS Remove chimera: non_chimera_abundance.biom
The abundance file to filter (format: BIOM)

Minimum prevalence method
replicate identification
Need to know group composition

File of replicated sample names
12: chaillou_replicate_information.tsv
Replicate file to link each sample to its group (cf. Help section).

Minimum prevalence
0.5
Here, user wants that each OTU of its group to be present in at least half of samples making up the group

Fill the field only if you want this treatment. Keep OTU present in at least this proportion of replicates in at least one group (must be a proportion between 0 and 1).

1

Prevalence filter – option 2

How to build the file of replicated sample names ?

The file must consist of **only 2 columns**, separated by a tab.

The first column contains **the exact names of the samples** (exactly those contained in the biom file)

The second column contains the name of the group to which they belong. Please note that group names must **not contain accents, spaces or special characters**.

Example:

```
sample1    rich
sample2    rich
sample3    rich
sample4    richAB
sample5    richAB
sample6    richAB
sample7    richAB
sample8    richAB
sample9    low
sample10   lowAB
sample11   lowAB
sample12   april21
sample13   april21
```

Thanks to get data tool,
add it in your history

1 Prevalence filter – option 2

Results:

if we want to keep the OTUs that are present in at least 50% of the samples of a same group, we set the threshold at 0.5.

The process will therefore keep the OTUs present in at least

2 "rich" samples

3 "richAB" samples,

1 "lowAB" sample

1 "april21" sample

sample1	rich
sample2	rich
sample3	rich
sample4	richAB
sample5	richAB
sample6	richAB
sample7	richAB
sample8	richAB
sample9	low
sample10	lowAB
sample11	lowAB
sample12	april21
sample13	april21

and all OTUs in sample9 since it is the only representative of the "low" condition.

1

Prevalence filter – option 2

mistakes not to be made:

```
sample1 rich
sample2 rich
sample3 rich
sample4 richAB
sample5 richAB
sample6 richAB
sample7 richAB
sample8 low
sample9 lowAB
sample10 lowAB
sample11 lowAB
sample12 april21
sample13 april21
```

valid

```
sample1 rich
sample2 rich
sample 3 rich
sample4 richAB
sample5 richAB
sample6 richAB
sample7 richAB
sample8 low
sample9 lowAB
sample10 lowAB
sample11 lowAB
sample12 april21
sample13 april21
```

Creates artificially 3 columns

```
sample1 rich
sample2 rich
sample3 rich
sample4 rich AB
sample5 richAB
sample6 richAB
sample7 richAB
sample8 low
sample9 lowAB
sample10 lowAB
sample11 lowAB
sample12 april21
sample13 april21
```

Creates artificially 3 columns

2

OTU size filter

Minimum OTU abundance as proportion or count. We recommend to use a proportion of 0.00005.

as proportion

Minimum proportion of sequences abundance to keep OTU

5e-05

Fill the field only if you want this treatment. Example: 0.00005, recommended by Bokulich et al 2013, to keep OTU with at least 0.005% of all sequences) (--min_abundance)

OR

Minimum OTU abundance as proportion or count. We recommend to use a proportion of 0.00005.

as count

Minimum number of sequences to keep OTU

2

Fill the field only if you want this treatment. Ex: 2 to keep OTU with at least 2 sequences, so remove single singleton (--min_abundance)

Here, user wants that each OTU has an abundance representing at least 0.005% of total number of sequences (*i.e.* 0.00005).

Here, user wants that each OTU has an abundance at least equals to 2 sequences -> single singleton will be removed.

3

Filter : Keep biggest OTU

N biggest OTUs

Fill the fields only if you want this treatment. Keep the N biggest OTU (--nb-biggest-otu)

Here, user wants to keep the 50 biggest OTUs.

4

Contaminant filter

Search for contaminant OTU.

Use contaminant fasta file from the server

Either you use your own contaminant fasta file or you select one among available ones.

Remove phiX sequence (use as buffer while sequencing)

Contaminant databank

phiX

For example the phiX databank (the phiX is a control added in Illumina sequencing technologies).

OR

Search for contaminant OTU.

Use contaminant fasta file from the server

Either you use your own contaminant fasta file or you select one among available ones.

Contaminant databank

Arabidopsis TAIR10 Chloroplast and mitochondrie

For example the phiX databank (the phiX is a control added in Illumina sequencing technologies).

Remove chloroplastic and mitochondrial 16S sequences of *A. Thaliana*

OR

Search for contaminant OTU.

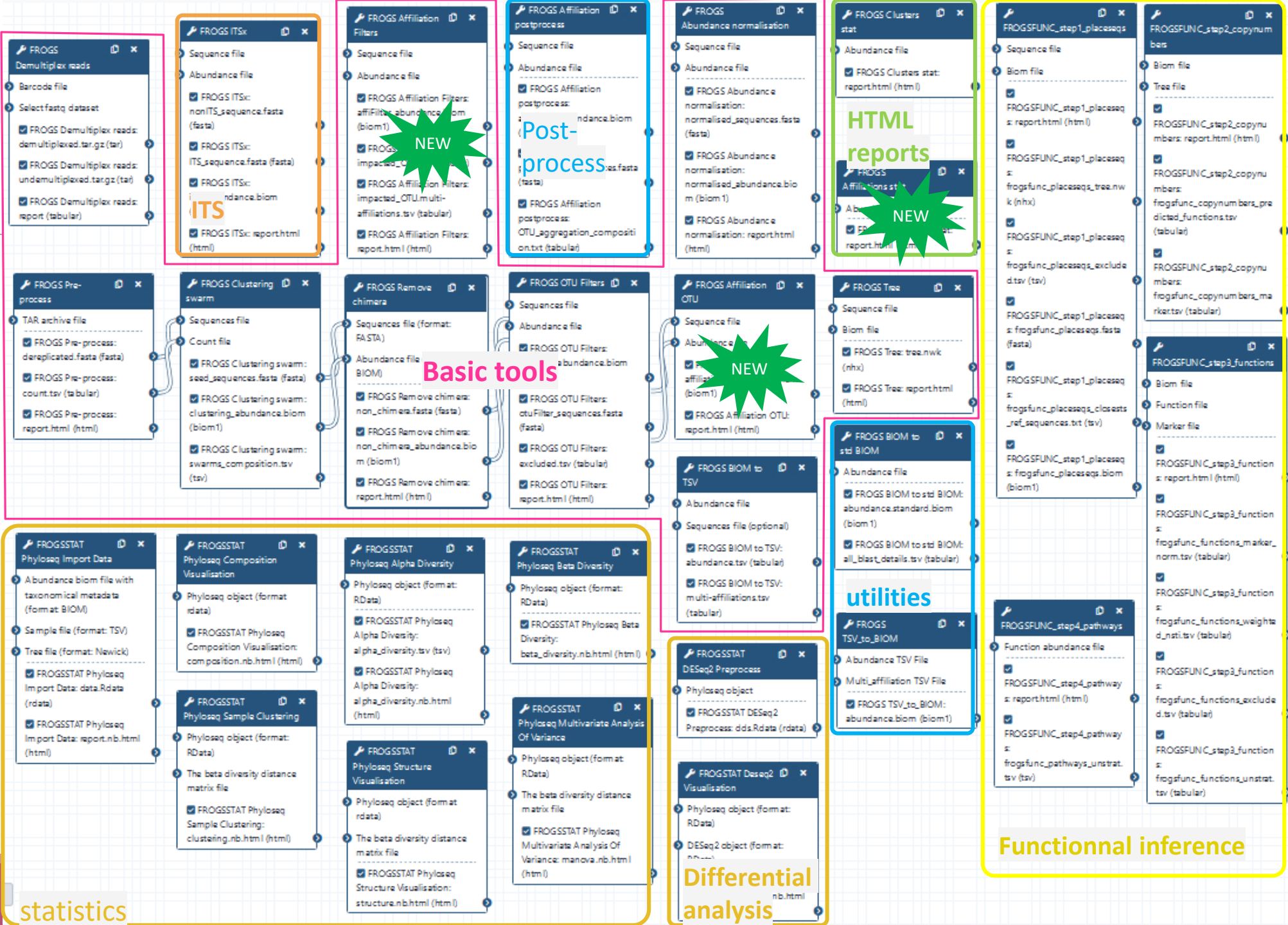
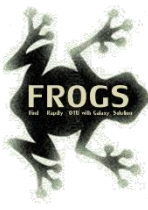
Use contaminant fasta file from the history

Either you use your own contaminant fasta file or you select one among available ones.

Select a contaminante reference from history

31: contaminant.fasta

Add in your history (with getadata tool) your own file of contaminant sequences in fasta format.



statistics

Differential analysis

Functional inference

Affiliation tool

FROGS Affiliation OTU Taxonomic affiliation of each OTU's seed

Using reference database

Select reference from the list

Also perform RDP assignment?

 Yes No

Optional

Taxonomy affiliation will be performed thanks to Blast. This option is optional.

Taxonomic ranks

The ordered taxonomic ranks levels stored in the taxonomical reference database.

OTU seed sequence

OTU sequences (format: fasta).

Abundance file

OTU abundances (format: BIOM).



silva138.1 16S
silva138.1 pintail100 16S
silva138.1 pintail80 16S
silva138.1 pintail50 16S
silva138.1 18S
silva138.1 23S
silva138.1 28S
silva138 16S
silva138 pintail100 16S
silva138 pintail80 16S
silva138 pintail50 16S
silva138 18S
silva138 SSU
silva132 LSU
silva132 28S
silva132 16S
silva132_pintail100 16S
silva132_pintail80 16S
silva132_pintail50 16S
silva132 18S
silva132 23S
greengenes13_5
midas_S132_3.6
midas_S123_2.1.3
Pyringae CTS 20200131
pr2_4.12.0
rpoB_122017
Unite_Fungi_8.2_20200204
Unite_Euka_8.2_20200204
Unite_Fungi_8.0_18112018
Unite_Euka_8.0_18112018
RSyst_Diatom_7

on 3.2.3)

Options

DAIRYdb_v1.1.2
EZBioCloud_052018
PHYMYCO-DB_2013
BOLD_COI-5P_022019
BOLD_COI-5P_1percentN_022019
MIDORI_UNIQUE_COI_20180221
MIDORI_UNIQUE_COI_MARINE_20180221
silva128 16S
silva128_pintail100 16S
silva128_pintail80 16S
silva128_pintail50 16S
silva128 18S
silva128 23S
silva123 16S
silva123 23S
silva123 18S
midas_S119_1.20
pr2_4.11.0
pr2_gb203_4.5
Unite_s_7.1_20112016



For more details on FROGS databanks:

http://genoweb.toulouse.inra.fr/frogs_databanks/assignation/readme.txt

Silva pintail or not pintail ?

Pintail* represents the probability that the rRNA sequence contains anomalies or is a chimera, where 100 means that the probability for being anomalous or chimeric is low.

4 ranks of available databases in FROGS: 50 pintail, 80 pintail or 100 pintail or no pintail filter.

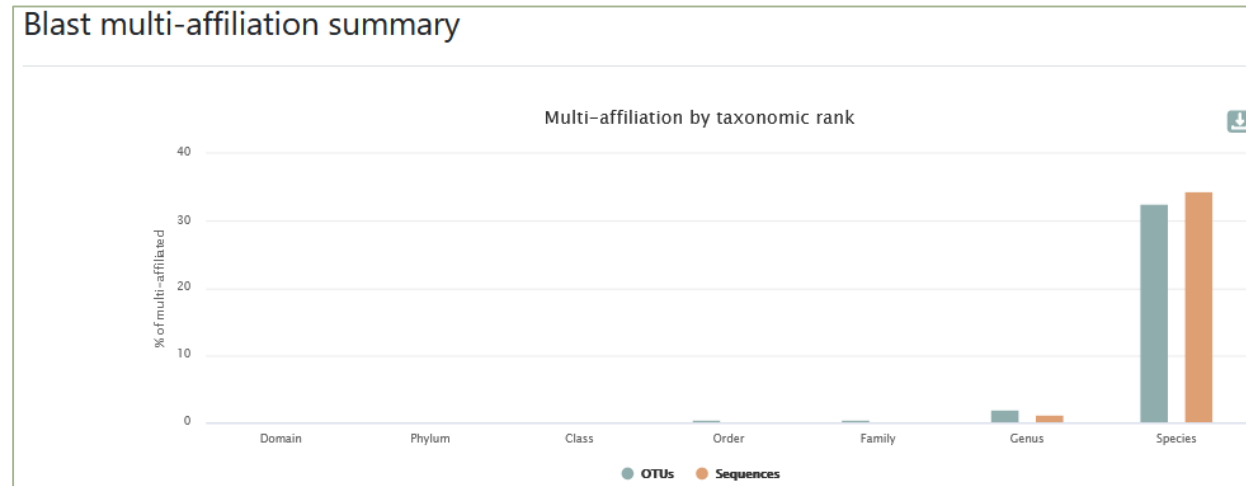
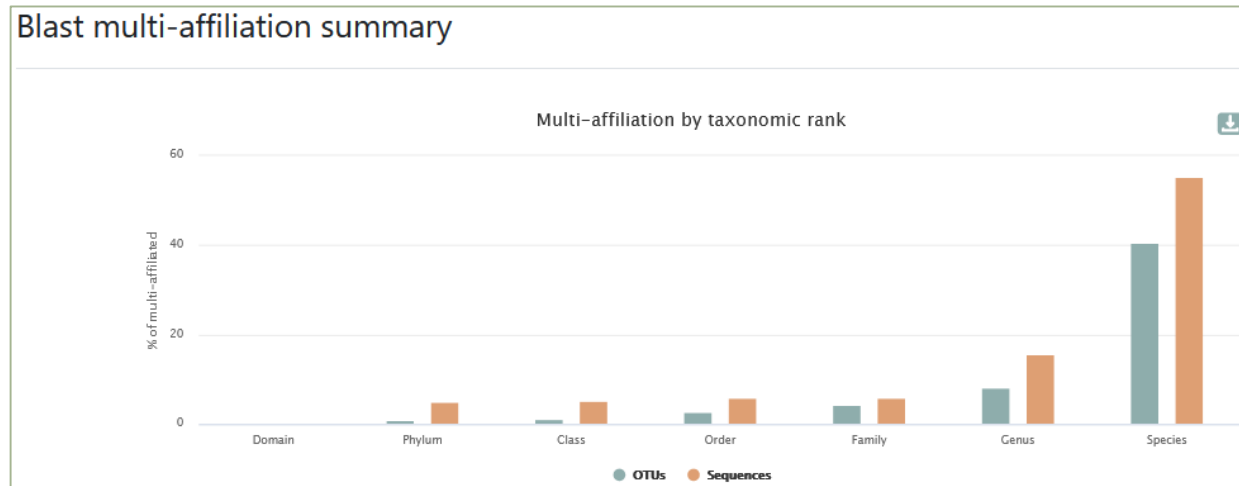
silva138.1 16S
silva138.1 pintail100 16S
silva138.1 pintail80 16S
silva138.1 pintail50 16S
silva138.1 18S
silva138.1 23S
silva138.1 28S



Only for 16S !

* <http://aem.asm.org/content/71/12/7724.abstract>

Silva pintail or not pintail ?



Exemple between silva 138.1 and silva 138.1 pintail 100

130 identical blast best hits on SILVA 138.1 pintail 100 databank

Cluster_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes
Cluster_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes 6609
Cluster_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes C1
Cluster_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes KPA171202
Cluster_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes TypeIA2 Pacn17
Cluster_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes TypeIA2 Pacn31
Cluster_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes TypeIA2 Pacn33

Exemple between silva 138.1 and silva 138.1 pintail 100

267 identical blast best hits on **SILVA 138.1** full databank

- ? Cluster_4 Bacteria;Actinobacteriota;Actinobacteria;Corynebacteriales;Corynebacteriaceae;Corynebacterium;unknown species
- ? Cluster_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Aureobasidium melanogenum
- Cluster_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes
- Cluster_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes 266
- Cluster_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes 6609
- Cluster_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes C1
- Cluster_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes hdn-1
- Cluster_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes HL096PA1
- Cluster_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes KPA171202
- Cluster_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes SK137
- Cluster_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;unknown species
- Cluster_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes TypeA2 P.acn17
- Cluster_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes TypeA2 P.acn31
- Cluster_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes TypeA2 P.acn33
- ? Cluster_4 Bacteria;Firmicutes;Bacilli;Lactobacillales;Carnobacteriaceae;Dolosigranulum;unknown species

Induces a multi-affiliation up to phylum rank

accession number	organism name	sequence length	sequence quality	alignment quality	pintail quality	SILVA taxonomy
<input type="checkbox"/> KF100699	<i>uncultured bacterium</i>	1341	<div style="width: 100%; height: 10px; background-color: green;"></div>	<div style="width: 100%; height: 10px; background-color: green;"></div>	<div style="width: 10%; height: 10px; background-color: gray;"></div>	Bacteria > Firmicutes > Bacilli...

How choose the good affiliation ?

Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus	D83374.1.1477	100	100	0	499
Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus	CP007208.2831760.2833315	100	100	0	499
Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus	CP007208.1649831.1651386	100	100	0	499
Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus	CP007208.1426849.1428404	100	100	0	499
Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus	CP007208.1544187.1545742	100	100	0	499
Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus	LT963439.723352				
Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus	CP013922.158796				
Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus	CP013922.2356345.2857902	100	100	0	499
Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus	CP013922.22851139.2852696	100	100	0	499
Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus	CP013922.2904966.2906523	100	100	0	499
Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus	CP013922.2899760.2901317	100	100	0	499
Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus	CP013922.1470936.1472493	100	100	0	499
Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus	CP013922.1685669.1687226	100	100	0	499
Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus saprophyticus	EU855225.1.1531	100	100	0	499

2 choices for cluster 64

How choose the good affiliation ?

Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus	D83374.1.1477	100	100	0	499
Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus	CP007208.2831760.2833315	100	100	0	499
Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus	CP007208.1649831.1651386	100	100	0	499
Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus	CP007208.1426849.1428404	100	100	0	499
Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus	CP007208.1544187.1545742	100	100	0	499
Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus	LT963439.723352.724884	100	100	0	499
Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus	CP013922.1587968.1589525	100	100	0	499
Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus	CP013922.2856345.2857902	100	100	0	499
Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus	CP013922.2851139.2852696	100	100	0	499
Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus	CP013922.2904966.2906523	100	100	0	499
Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus	CP013922.2899760.2901317	100	100	0	499
Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus	CP013922.1470936.1472493	100	100	0	499
Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus	CP013922.1685669.1687226	100	100	0	499
Cluster_64	Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus saprophyticus	EU855225.1.1531	100	100	0	499

- you have a preconceived notion
- you are familiar with the environment being studied
- you are looking for specific organisms as pathogens
- you collect bibliographical information

Ex:

Staphylococcus saprophyticus is a bacterium that can cause urinary tract infections in young women

and

Staphylococcus xylosus exists as a commensal on the skin of humans and animals and in the environment. It appears to be much more common in animals than in humans. *S. xylosus* has very occasionally been identified as a cause of human infection.

Maybe, for this cluster, *S. xylosus* is better

Affiliation explorer

<https://shiny.migale.inrae.fr/app/affiliationexplorer>

The screenshot shows the Affiliation Explorer web application. On the left, there are three upload sections: 'Upload Biom File' (Galaxy37-[f]), 'Optional: upload Fasta File' (Galaxy32-[f]), and 'Upload MultiHits TSV File' (Galaxy42-[f]). Each has a 'Browse...' button and an 'Upload complete' button. A 'Download' button is at the bottom left. The main area has two tabs: 'Affiliation selection' and 'Affiliation edition'. Under 'Affiliation selection', there is a 'Select OTU' dropdown menu set to 'Cluster_3', and 'Update OTU' and 'Skip OTU' buttons. Below this, a message states: 'Cluster_3 - 2 conflicting affiliations, ambiguity at rank Species'. Instructions follow: 'Select new affiliation by clicking on a row (double click on a cell to edit its content). Click "Update OTU" to update affiliation (with selected row) or "Skip OTU" to move to the next one.' There is a 'Show 10 entries' dropdown and a search box. A table displays two entries with columns for Kingdom, Phylum, Class, Order, Family, Genus, Species, Blast ID, %id, and %cov. The first entry is for 'Lactobacillus sakei' and the second is for 'unknown species'. At the bottom, it says 'Showing 1 to 2 of 2 entries' and has 'Previous', '1', and 'Next' navigation buttons. A 'Show sequence' checkbox is at the bottom left.

Upload Biom File
Browse... Galaxy37-[f]
Upload complete

Optional: upload Fasta File
Browse... Galaxy32-[f]
Upload complete

Upload MultiHits TSV File
Browse... Galaxy42-[f]
Upload complete

Download

Affiliation selection Affiliation edition

Select OTU
Cluster_3 Update OTU Skip OTU

Cluster_3 - 2 conflicting affiliations, ambiguity at rank Species
Select new affiliation by clicking on a row (double click on a cell to edit its content).
Click "Update OTU" to update affiliation (with selected row) or "Skip OTU" to move to the next one.

Show 10 entries Search:

	Kingdom	Phylum	Class	Order	Family	Genus	Species	Blast ID	%id	%cov
1	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Latilactobacillus	Lactobacillus sakei	CP032640.225274.226851	100	100
2	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Latilactobacillus	unknown species	KF601977.1.1550	100	100

Showing 1 to 2 of 2 entries Previous 1 Next

Show sequence

A very user-friendly tool, developed by Mahendra Mariadassou and his collaborators (Maiage unit - INRAE Jouy-en-Josas). It allows to modify very simply the affiliations of an abundance table from FROGS.

Affiliation explorer

<https://shiny.migale.inrae.fr/app/affiliationexplorer>

Demo
video

The screenshot shows a web browser window displaying the 'Affiliation explorer' application. The browser's address bar shows the URL: <https://hub.gke2.mybinder.org/user/mahendra-mariad-liationexplorer-4jqib7jw/rstudio/?token=r0mZweROqcCzicA5hQm8IA&view=shiny>. The application interface has a dark blue header with the title 'Affiliation explorer' and a hamburger menu icon. Below the header, there are three file upload sections on the left: 'Upload Biom File', 'Optional: upload Fasta File', and 'Upload MultiHits TSV File'. Each section contains a 'Browse...' button and a 'No file sele...' button. The main content area has two tabs: 'Affiliation selection' (active) and 'Affiliation edition'. Below the tabs, there is a text prompt: 'Please upload your data (Biom file and MultiHits TSV file)'. The browser window also shows standard navigation and window control icons.

Affiliation Stat

[Display global distribution](#)[CSV](#)Show entriesSearch:

<input type="checkbox"/>	Samples	Nb domain	Nb phylum	Nb class	Nb order	Nb family	Nb genus	Nb species	Nb otus	Nb sequences
<input type="checkbox"/>	BHT0.LOT01	1	7	9	20	35	54	77	98	8,690
<input type="checkbox"/>	BHT0.LOT03	1	5	8	25	46	88	120	135	8,377
<input type="checkbox"/>	BHT0.LOT04	1	7	10	27	51	89	126	150	8,643
<input type="checkbox"/>	BHT0.LOT05	1	5	7	22	40	69	116	140	8,544
<input type="checkbox"/>	BHT0.LOT06	1	6	10	28	47	91	125	145	8,646
<input type="checkbox"/>	BHT0.LOT07	1	6	9	28	51	90	124	150	8,671
<input type="checkbox"/>	BHT0.LOT08	1	6	9	27	53	109	166	195	8,479
<input type="checkbox"/>	BHT0.LOT10	1	4	7	26	50	106	144	165	8,606
<input type="checkbox"/>	CDT0.LOT02	1	6	8	22	36	58	85	92	8,750
<input type="checkbox"/>	CDT0.LOT04	1	5	7	22	41	74	138	161	8,605

With selection:

Class

[Display rarefaction](#)[Display distribution](#)

Class

Order

Family

Genus

Species

OTUs

Showing 1 to 10 of 6

[Previous](#) [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [Next](#)

It is now possible to make rarefaction curves on OTUs

Filters on affiliations

FROGS Affiliation Filters Filters OTUs on several affiliation criteria. (Galaxy Version 3.2.2) Options

Sequences file
 13: FROGS OTU Filters: sequences.fasta
 The sequence file to filter (format: fasta).

Abundance file
 18: FROGS Affiliation OTU: affiliation.biom
 The abundance file to filter (format: BIOM).

Taxonomic ranks

 The ordered taxonomic ranks levels stored in BIOM. Each rank is separated by one space.

Filtering mode
 Hidding mode
 Deleting mode
 Do you want to delete OTUs or hide affiliations

Filter on Blast affiliations

Maximum e-value (between 0 and 1)

 Fill the field only if you want this treatment

Minimum identity % (between 0 and 1)

 Fill the field only if you want this treatment

Minimum coverage % (between 0 and 1)

 Fill the field only if you want this treatment

Minimum alignment length

 Fill the field only if you want this treatment

Filter blast affiliations including these taxon / word

1: Filter blast affiliations including these taxon / word trash

Full or partial taxon name

 ex: "unknown species" or "subsp."

2: Filter blast affiliations including these taxon / word

Full or partial taxon name

 ex: "unknown species" or "subsp."

3: Filter blast affiliations including these taxon / word

Full or partial taxon name

 ex: "unknown species" or "subsp."

Filter on RDP affiliations

Taxonomical rank on which to apply bootstrap filter

 One of the available taxonomical rank name. Ex: Species

Minimum bootstrap % (between 0 and 1)

 Fill these two fields if you want this treatment.

Careful, it is case sensitive.
 Firmicutes it's different of firmicutes !

Not open by default

2 modes: hidding or deleting mode.
 All affiliations that enter in criteria of filter will be either hidden or deleted

- hidding: affiliation counting are not affected, affiliation are simply hidden
- deleting: all abundancies are computed again, affiliation have disappeared

Practice:

LAUNCH THE FROGS AFFILIATION FILTER TOOL

Exercise:

1. Apply filters to keep only sequences with perfect alignment with Silva sequences and affiliations without « unknown species » and « Firmicutes » terms. (deleting mode)
2. Apply filters to hide OTU affiliations that have not a perfect alignment with Silva sequences and the affiliations without « unknown species » and « Firmicutes » terms.
3. In deleting mode:
 - How many OTUs remain?
 - Among OTUs with multiaffiliation, How many were impacted/modified ?
4. In hiding mode:
 - What outputs change between deleted mode and hiding mode ?

FROGS Affiliation Filters Filters OTUs on several affiliation criteria. (Galaxy Version 3.2.2) Options

Sequences file
13: FROGS OTU Filters: sequences.fasta
The sequence file to filter (format: fasta).

Abundance file
18: FROGS Affiliation OTU: affiliation.biom
The abundance file to filter (format: BIOM).

Taxonomic ranks
Domain Phylum Class Order Family Genus Species
The ordered taxonomic ranks levels stored in BIOM. Each rank is separated by one space.

Filtering mode
 Hidding mode
 Deleting mode
Do you want to delete OTU or hide affiliations

Filter on Blast affiliations

Maximum e-value (between 0 and 1)
[Slider: 0 to 1]

Fill the field only if you want this treatment

Minimum identity % (between 0 and 1)
1 [Slider: 0 to 1]

Fill the field only if you want this treatment

Minimum coverage % (between 0 and 1)
1 [Slider: 0 to 1]

Fill the field only if you want this treatment

Minimum alignment length
[Input field]

Fill the field only if you want this treatment

Filter blast affiliations including these taxon / word

1: Filter blast affiliations including these taxon / word

Full or partial taxon name
unknown species
ex: "unknown species" or "subsp."

2: Filter blast affiliations including these taxon / word

Full or partial taxon name
Firmicutes
ex: "unknown species" or "subsp."

+ Insert Filter blast affiliations including these taxon / word

Filter on RDP affiliations

Execute

Answer 1

FROGS Affiliation Filters Filters OTUs on several affiliation criteria. (Galaxy Version 3.2.2) Options

Sequences file
13: FROGS OTU Filters: sequences.fasta
The sequence file to filter (format: fasta).

Abundance file
18: FROGS Affiliation OTU: affiliation.biom
The abundance file to filter (format: BIOM).

Taxonomic ranks
Domain Phylum Class Order Family Genus Species
The ordered taxonomic ranks levels stored in BIOM. Each rank is separated by one space.

Filtering mode
 Hidding mode
 Deleting mode
Do you want to delete OTU or hide affiliations

Filter on Blast affiliations

Maximum e-value (between 0 and 1)
[Slider: 0 to 1]

Fill the field only if you want this treatment

Minimum identity % (between 0 and 1)
1 [Slider: 0 to 1]

Fill the field only if you want this treatment

Minimum coverage % (between 0 and 1)
1 [Slider: 0 to 1]

Fill the field only if you want this treatment

Minimum alignment length
[Input field]

Fill the field only if you want this treatment

Filter blast affiliations including these taxon / word

1: Filter blast affiliations including these taxon / word

Full or partial taxon name
unknown species
ex: "unknown species" or "subsp."

2: Filter blast affiliations including these taxon / word

Full or partial taxon name
Firmicutes
ex: "unknown species" or "subsp."

+ Insert Filter blast affiliations including these taxon / word

Filter on RDP affiliations

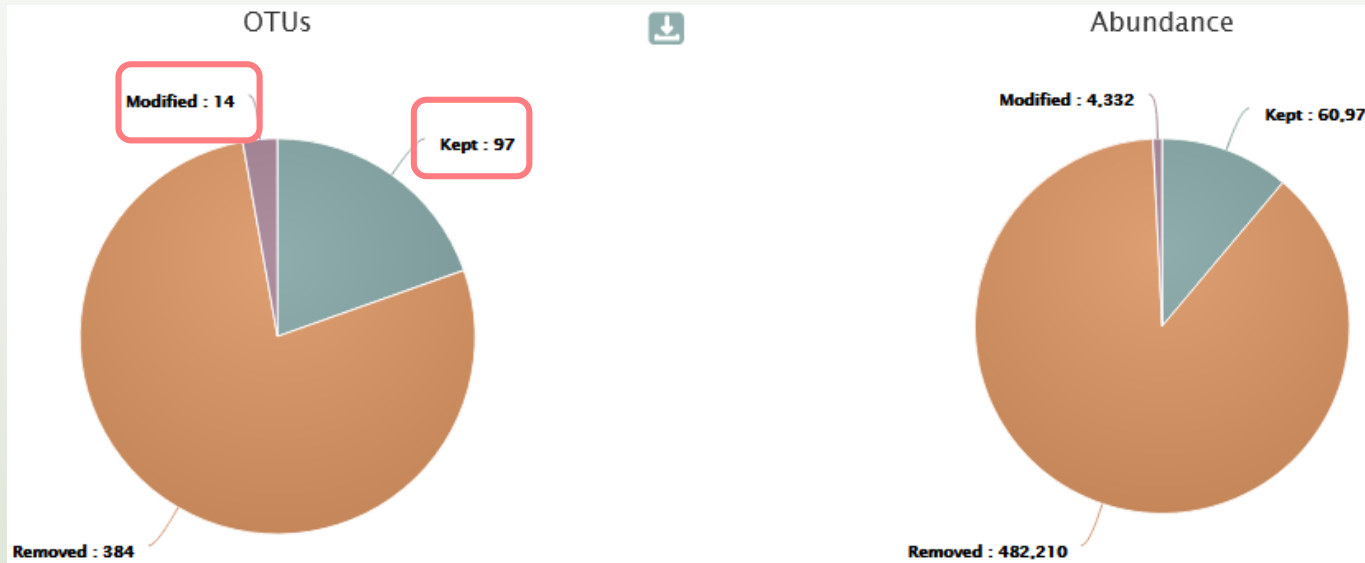
Execute

Answer 2

we want to keep the OTUs that have aligned perfectly with a sequence of the silva bank i.e. 100% identity and 100% coverage

Enter key word

Q3: In deleting mode:
- How many OTUs remain?



- Only 97 OTUs are kept without modification.
- 14 OTUs with multi-affiliation were impacted/modified (all affiliations in the multi_affiliations with key words “unknown species” or “Firmicutes” were deleted).
The consequences are either OTU have less multi-affiliations, or all multi-affiliations are impacted and OTU is deleted.
The list of blast affiliations for multi-affiliated impacted OTUs are in **impacted_OTU.multi-affiliation.tsv**
- So, **111 OTUs** remains after filtering

[: FROGS Affiliation Filters: report.html](#)

[FROGS Affiliation Filters: impacted_OTU.multi-affiliations.tsv](#)

[FROGS Affiliation Filters: impacted_OTU.tsv](#)

[FROGS Affiliation Filters: sequences.fasta](#)

[FROGS Affiliation Filters: abundance.biom](#)

Answer 3

FROGS Affiliation Filters: report.html
FROGS Affiliation Filters: impacted_OTU.multi-affiliations.tsv
FROGS Affiliation Filters: impacted_OTU.tsv
FROGS Affiliation Filters: sequences.fasta
FROGS Affiliation Filters: abundance.biom

N.B. The abundancy table (TSV format) of all deleted (or hidden according to the tool parameters) or modified OTUs are kept in **impacted_OTU.tsv**

#comment	status	blast_taxonomy
undesired_tax_in_blast	OTU_deleted	Bacteria;Firmicutes;Bacilli;Lactobacillales;Listeriaceae;Brochothrix;Brochothrix thermosphacta
undesired_tax_in_blast	OTU_deleted	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales;Vibrionaceae;Photobacterium;unknown species
undesired_tax_in_blast	OTU_deleted	Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Latilactobacillus;Multi-affiliation
undesired_tax_in_blast	Blast_taxonomy_changed	Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Moraxellaceae;Psychrobacter;Multi-affiliation
blast_identity_lt_1.0;undesired_tax_in_blast	OTU_deleted	Bacteria;Firmicutes;Bacilli;Lactobacillales;Streptococcaceae;Lactococcus;Lactococcus piscium
blast_identity_lt_1.0;undesired_tax_in_blast	OTU_deleted	Bacteria;Firmicutes;Bacilli;Erysipelotrichales;Erysipelotrichaceae;ZOR0006;unknown species
undesired_tax_in_blast	OTU_deleted	Bacteria;Firmicutes;Bacilli;Lactobacillales;Streptococcaceae;Lactococcus;Multi-affiliation
blast_identity_lt_1.0;undesired_tax_in_blast	OTU_deleted	Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Weissella;Weissella ceti
blast_identity_lt_1.0	OTU_deleted	Bacteria;Bacteroidota;Bacteroidia;Flavobacteriales;Flavobacteriaceae;Flavobacterium;Flavobacterium sp.
blast_identity_lt_1.0	OTU_deleted	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales;Vibrionaceae;Photobacterium;Photobacterium phosphoreum
blast_identity_lt_1.0;blast_coverage_lt_1.0;undesired_tax_in_blast	OTU_deleted	Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Dellaglioia;Lactobacillus algidus

In impacted_OTU.tsv

- #comment: the reason(s) why OTU was deleted (or hidden)
- #status: for deleted OTU (or hidden OTU), or for OTU with modified consensus taxonomy with affiliation (or multi-affiliation) was modified

Q4: In hiding mode: What outputs change between deleted mode and hiding mode ?

- [FROGS Affiliation Filters: report.html](#)
- [FROGS Affiliation Filters: impacted_OTU.multi-affiliations.tsv](#)
- [FROGS Affiliation Filters: impacted_OTU.tsv](#)
- [FROGS Affiliation Filters: abundance.biom](#)

In hidden mode: no **sequence.fasta** as output because none OTU was deleted

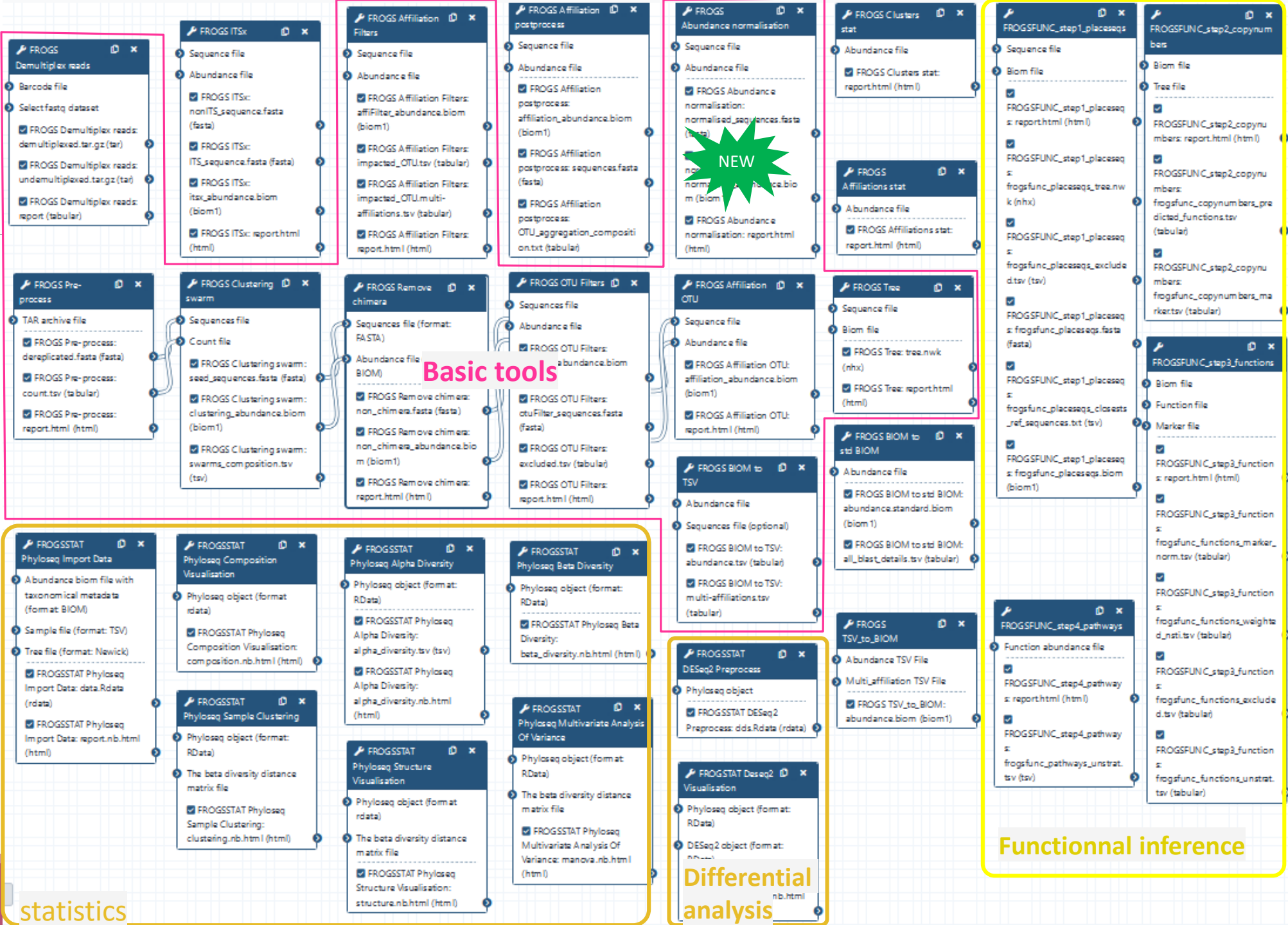
In hidden mode: **abundance.biom** contains all OTU but 111 have their affiliation that is hidden

#comment	blast_taxonomy	blast_subject	blast_percent_identity	blast_percent_identity	blast_evalue	blast_align	seed_id	observation
undesired_tax_in_blast	no data	no data	no data	no data	no data	no data	17_41	Cluster_1
undesired_tax_in_blast	no data	no data	no data	no data	no data	no data	17_611	Cluster_2
undesired_tax_in_blast	no data	no data	no data	no data	no data	no data	17_595	Cluster_3
undesired_tax_in_blast	Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Multi-affiliation	multi-subjec	100	100	0	468	17_257	Cluster_4
undesired_tax_in_blast	no data	no data	no data	no data	no data	no data	17_4	Cluster_5
blast_identity_lt_1.0;undesired_tax_in_blast	no data	no data	no data	no data	no data	no data	17_23	Cluster_6
blast_identity_lt_1.0;undesired_tax_in_blast	no data	no data	no data	no data	no data	no data	57_5	Cluster_7
undesired_tax_in_blast	no data	no data	no data	no data	no data	no data	17_420	Cluster_8

« no data » appears in hiding mode



To see the content, think to transform the BIOM to TSV file with **BIOM_to_TSV tool**



statistics

Differential analysis

Functional inference

Normalization

Normalization

Conserve a predefined number of sequence per sample:

- update Biom abundance file
- update seed fasta file

May be used when :

- Low sequencing sample
- Required for some statistical methods to compare the samples in pairs

FROGS Abundance normalisation Normalise OTU abundance. (Galaxy Version 4.0.0+galaxy1)

Sequence file



14: FROGS OTU Filters: otuFilter_sequences.fasta

Sequence file to normalise (format: fasta).

Abundance file



17: FROGS Affiliation OTU: affiliation_abundance.biom

Abundance file to normalise (format: BIOM).




Sampling method

- Sampling by the number of sequences of the smallest sample
- Select a number of sequences

Sampling by the number of sequences of the smallest sample, or select a number manually




FROGS Abundance normalisation Normalise OTU abundance. (Galaxy Version 4.0.0+galaxy1)

Sequence file

   14: FROGS OTU Filters: otuFilter_sequences.fasta

Sequence file to normalise (format: fasta).

Abundance file

   17: FROGS Affiliation OTU: affiliation_abundance.biom

Abundance file to normalise (format: BIOM).

Sampling method

Sampling by the number of sequences of the smallest sample

Select a number of reads

Sampling by the number of sequences of the smallest sample, or select a number manually

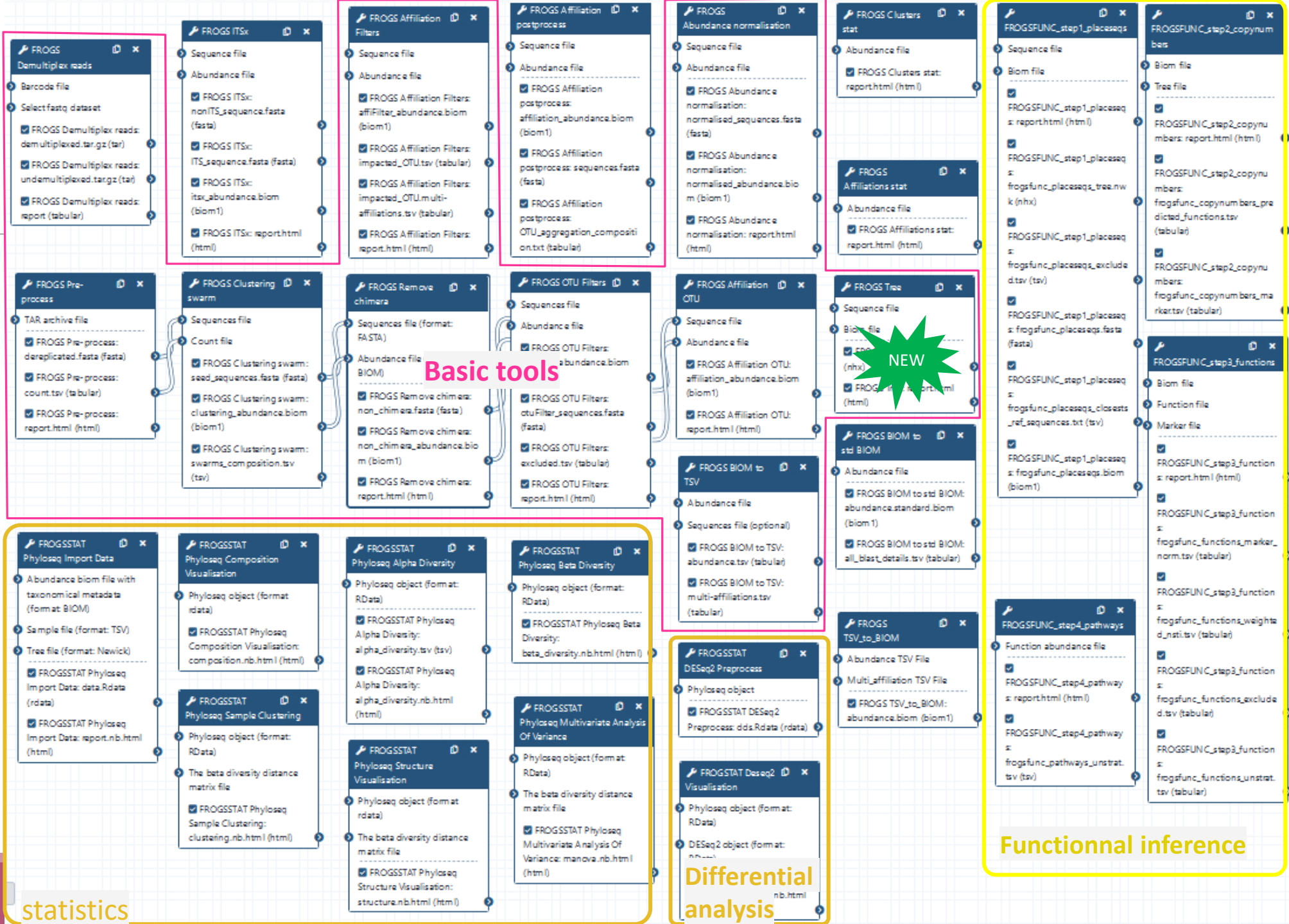
Number of reads

2000

The final number of reads per sample.

Remove samples that have an initial number of reads below the number of reads to sample ?

No



statistics

Differential analysis

Functional inference

FROGS Tree




CREATE A PHYLOGENETICS TREE OF OTUS

FROGS Tree

This tool builds a phylogenetic tree thanks to affiliations of OTUs contained in the BIOM file
It uses MAFFT for the multiple alignment and FastTree for the phylogenetic tree.




FROGS Tree Reconstruction of phylogenetic tree (Galaxy Version 4.0.0+galaxy1)

Sequence file

   29: FROGS OTU Filters: otuFilter_sequences.fasta

Sequence file (format: FASTA). Warning: FROGS Tree does not work on more than 10000 sequences!

Biom file

   33: FROGS Affiliation OTU: Pintail100affiliation_abundance.biom

The abundance file (format: BIOM)

Email notification

No

Send an email notification when the job completes.

Execute

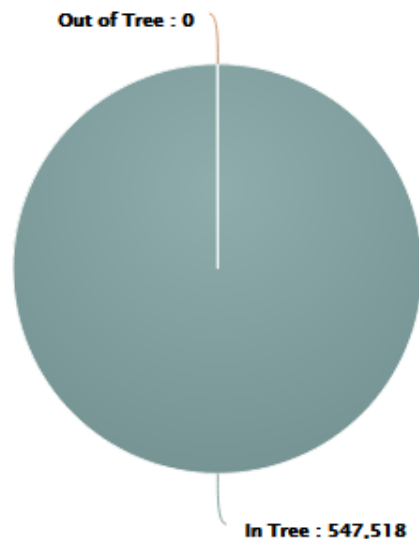
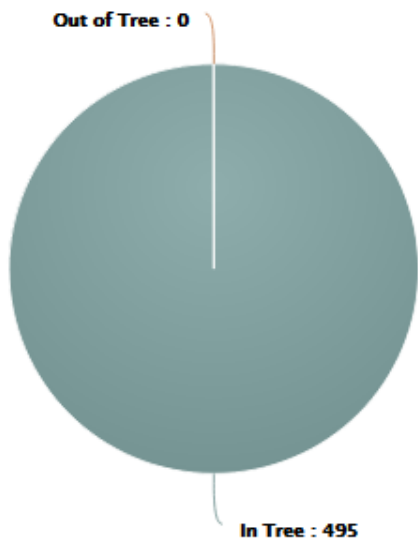
2 outputs:

FROGS Tree: report.html

FROGS Tree: tree.nwk

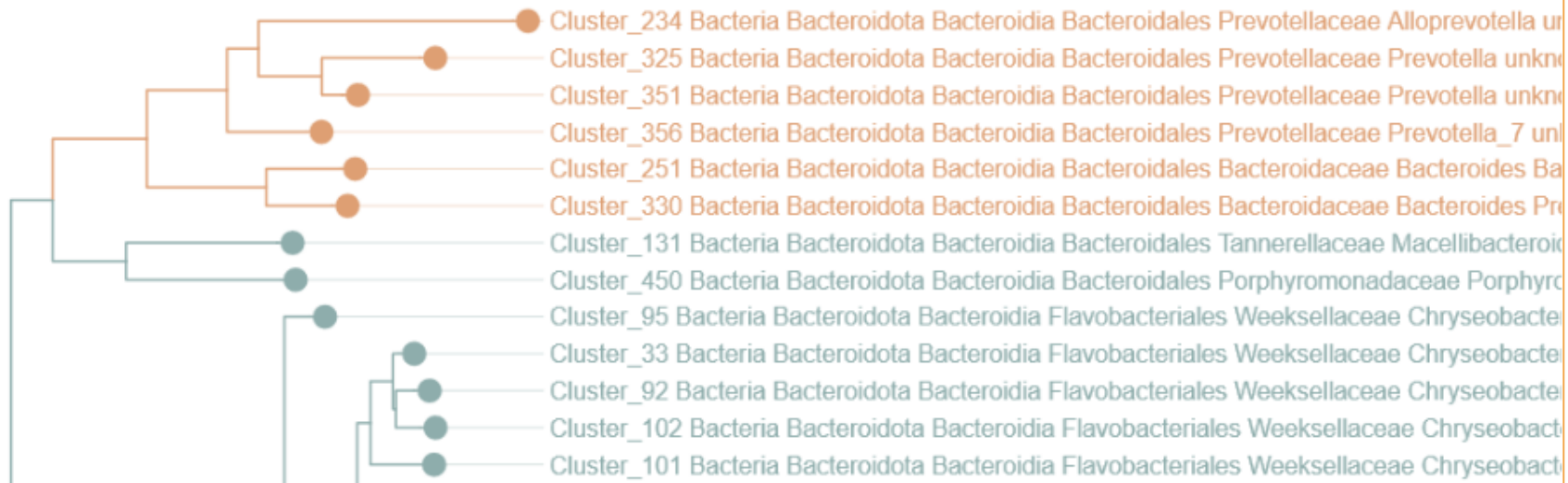
OTUs

Abundance

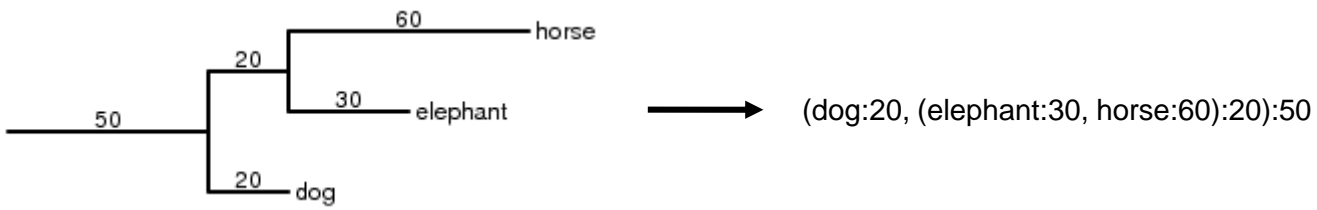


Tree View

Enabling zoom:



The phylogenetic tree in Newick format *i.e.* each node is represented between brackets. This format is universal and can be used with all tree viewer



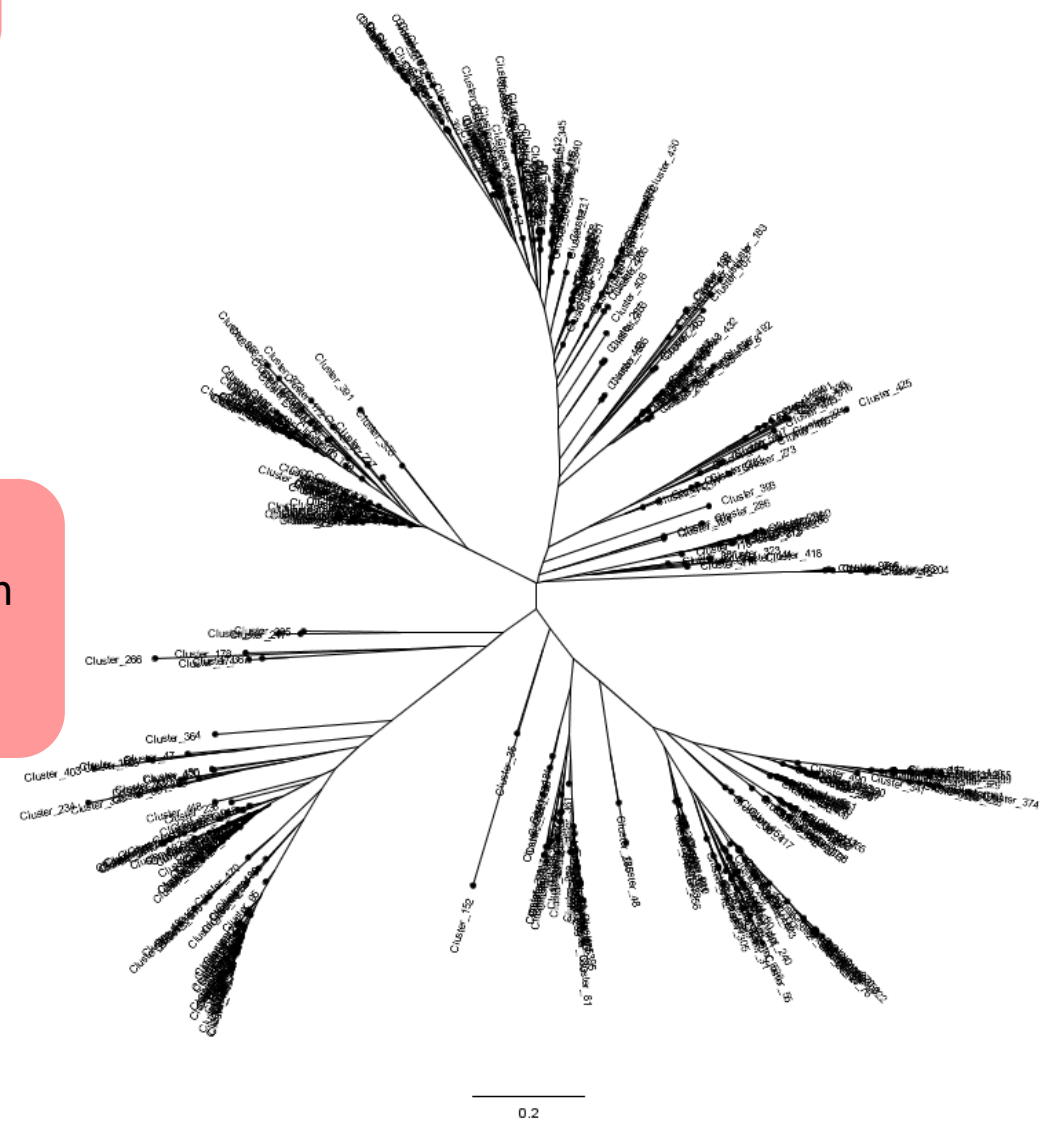
Our tree in nhx (= nwk) format

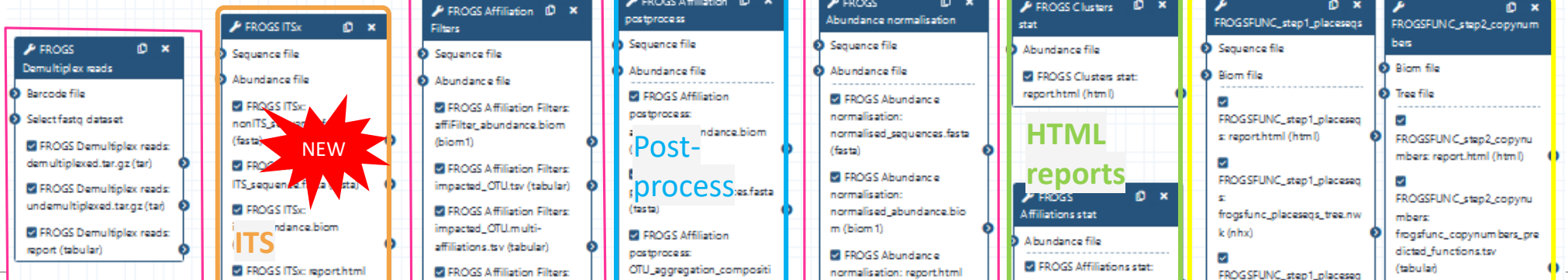
Exemple of visualization in FigTree from nhx file

```

((((((((((((Cluster_234:0.25278,(Cluster_325:0.09784,Clu
67)0.972:0.02504,(Cluster_468:0.0269,(Cluster_138:0.0016
.782:0.00832,Cluster_277:0.01601)1.000:0.06764,Cluster_4
ter_47:0.13954,(Cluster_166:0.16129,(Cluster_403:0.22934
72:0.01332,(Cluster_400:0.00545,Cluster_473:0.01483)1.00
)0.829:0.01282,Cluster_240:0.12227)0.717:0.02027)0.981:0
uster_478:0.00249)0.000:0.00055,(Cluster_193:0.00055,Clu
359,Cluster_484:0.01913)0.880:0.03155)0.993:0.08088)0.45
0989)0.827:0.01144)0.870:0.01235,((Cluster_81:0.08926,Cl
05)0.862:0.00658,(Cluster_303:0.04337,Cluster_398:0.0311
237)0.953:0.01895,(Cluster_346:0.0235,((Cluster_369:0.01
Cluster_402:0.12402,(Cluster_309:0.02202,(Cluster_284:0.
.00054,(Cluster_427:0.00054,(Cluster_14:0.00402,Cluster_
0.791:0.02141,(Cluster_93:0.00054,Cluster_340:0.01463)0.
:0.03373)0.847:0.03692,Cluster_406:0.16125)0.831:0.03655
:0.04264)0.321:0.00907)0.487:0.01277,Cluster_129:0.06386
02802)0.763:0.02715,(Cluster_16:0.1183,(Cluster_63:0.062

```

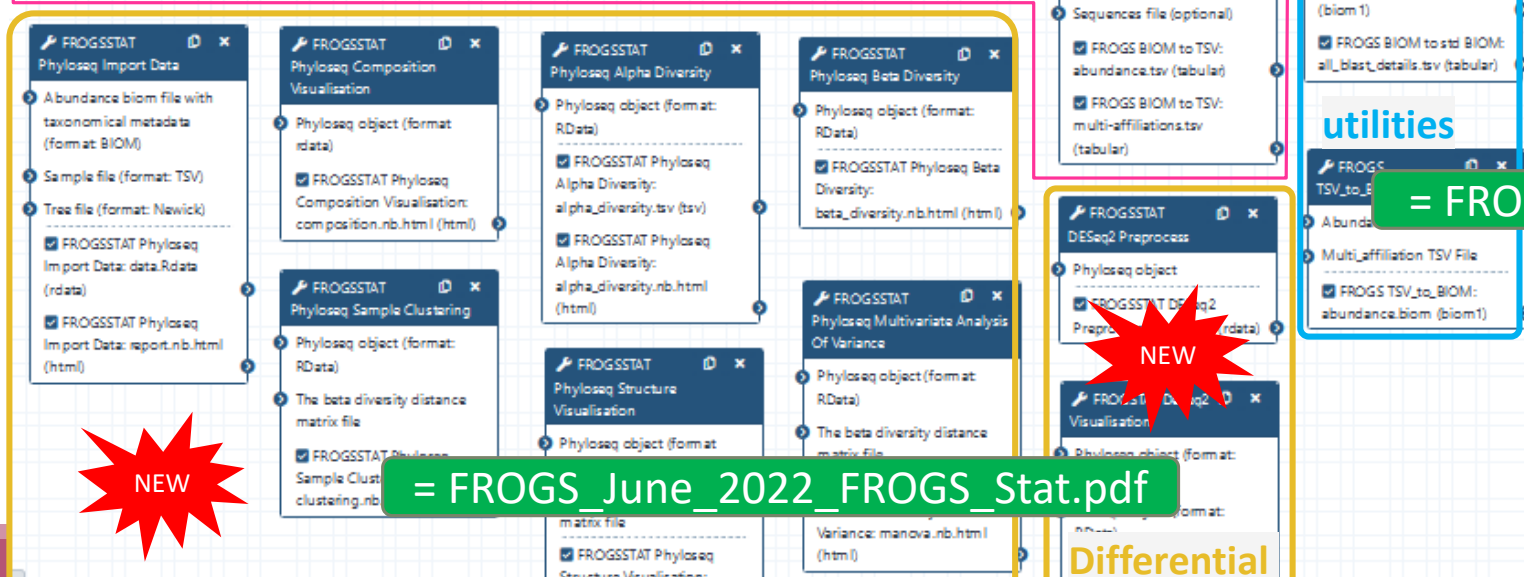




= FROGS_June_2022_FROGS_ITS.pdf



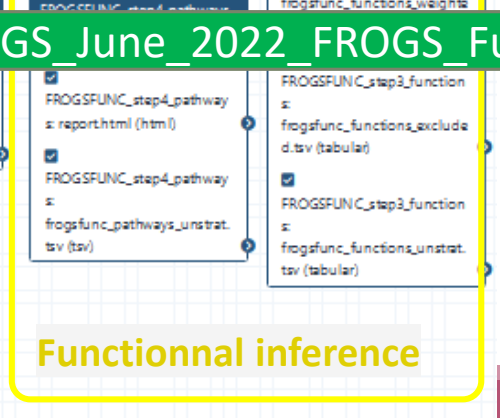
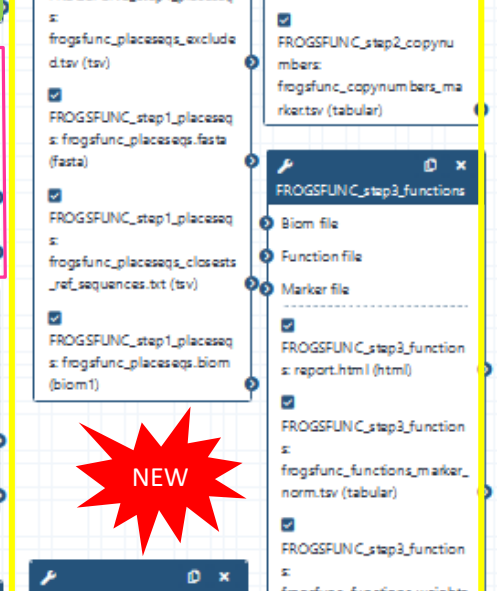
Basic tools



statistics

= FROGS_June_2022_FROGS_Stat.pdf

Differential analysis



Functional inference

= FROGS_June_2022_FROGS_Func.pdf



How to cite FROGS

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

Maria Bernard, Olivier Rué, Mahendra Mariadassou and Géraldine Pascal; **FROGS**: a powerful tool to analyse the diversity of fungi with special management of internal transcribed spacers, *Briefings in Bioinformatics* 2021, 10.1093/bib/bbab318

Sequence analysis

FROGS: Find, Rapidly, OTUs with Galaxy Solution

Frédéric Escudié^{1,†}, Lucas Auer^{2,†}, Maria Bernard³, Mahendra Mariadassou⁴, Laurent Cauquil⁵, Katia Vidal⁶, Sarah Maman⁶, Guillermina Hernandez-Raquet⁶, Sylvie Combes⁶ and Géraldine Pascal^{2,*}

¹Bioinformatics platform Toulouse Midi-Pyrénées, MIAI, INRA Arceville CS 52627 31326 Castanet Tolosan cedex, France, ²INRA, UMR 1138, Université de Lorraine, Nancy, France, ³SABI, INRA, AgroParisTech, Université Paris Saclay, Jouy-en-Josas, France, ⁴MaDiSE, INRA, Université Paris Saclay, INRA, Jouy-en-Josas, France, ⁵GenPhySE, Université de Toulouse, INRA, INPT, ENVT, Castanet Tolosan, France and ⁶Laboratoire d'Ingénierie des Systèmes Biologiques et des Procédés LISBP, Université de Toulouse, INSA, INRA, CNRS, Toulouse, France

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Bonnie Berger

Received on May 16, 2017; revised on December 1, 2017; editorial decision on December 4, 2017; accepted on December 5, 2017

Abstract

Motivation: Metagenomics leads to major advances in microbial ecology and biologists need user friendly tools to analyze their data on their own. **Results:** This Galaxy-supported pipeline, called FROGS, is designed to analyze large sets of amplicon sequences and produce abundance tables of Operational Taxonomic Units (OTUs) and their taxonomic affiliation. The clustering uses `Swarm`. This chimera removal uses `VSEARCH` combined with original cross-sample validation. The affiliation output to highlight databases confusable graphical illustrations are produced along for the detection and quantification of OTUs: robust and highly sensitive. It compares to QIME.

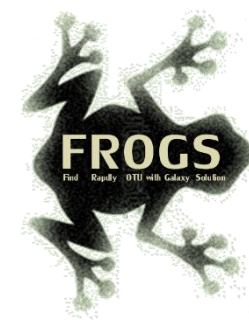
Availability and implementation: Source code: `geraldinepascal/FROGS.git`. A companion web: `Contact: geraldine.pascal@inra.fr`
Supplementary information: Supplementary

1 Introduction

The expansion of high-throughput sequencing of rDNA has opened new horizons for the study of microbial life by making it possible to study all micro-organisms in an environment without the need to cultivate them, leading to major advances in many fields of microbial ecology: study of the impact of microbiota on human and animal

© The Author(s) 2017. Published by Oxford University Press. All rights reserved.

Bioinformatics, 2017, 1–8
doi:10.1093/bioinformatics/btx799
Advance Access Publication Date: 7 December 2017
Original Paper



Briefings in Bioinformatics, 22(6), 2021, 1–6

<https://doi.org/10.1093/bib/bbab318>
Problem Solving Protocol

FROGS: a powerful tool to analyse the diversity of fungi with special management of internal transcribed spacers

Maria Bernard¹, Olivier Rué¹, Mahendra Mariadassou² and Géraldine Pascal²

Corresponding author: Géraldine Pascal, GenPhySE, Université de Toulouse, INRAE, INPT-ENVT-31326, Castanet Tolosan, France. Tel.: +33 (0)5 63 28 51 05; E-mail: geraldine.pascal@inrae.fr

¹Maria Bernard and Olivier Rué are joint first authors.

Abstract

Fungi are present in all environments. They fulfill important ecological functions and play a crucial role in the food industry. Their accurate characterization is thus indispensable, particularly through metabarcoding. The most frequently used markers to monitor fungi are ITSs. These markers are the best documented in public databases but have one main weakness: polymerase chain reaction amplification may produce non-overlapping reads in a significant fraction of the fungi. When these reads are filtered out, traditional metabarcoding pipelines lose part of the information and consequently produce biased pictures of the composition and structure of the environment under study. We developed a solution that enables processing of the entire set of reads including both overlapping and non-overlapping, thus providing a more accurate picture of fungal communities. Our comparative tests using simulated and real data demonstrated the effectiveness of our solution, which can be used by both experts and non-specialists on a command line or through the Galaxy-based web interface.

Key words: fungi; ITS; metabarcoding; workflow; amplicon; metagenomics

Introduction

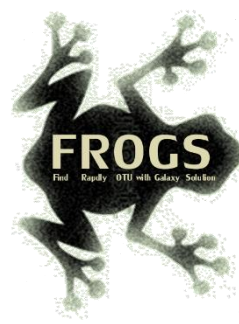
Using amplicon sequencing to describe the microbial composition of an environment is a time saving and cost-effective strategy and can be used even for very large-scale surveys [1]. Most studies currently focus on the bacterial fraction of microbial communities but the fungal fraction is equally important, as fungi are ubiquitous and provide several ecosystem services [2]. Unfortunately, studying the fungal fraction using metabarcoding has its own challenges. Indeed, in fungi, there is no equivalent of the 16S rDNA gene, which is widely used and highly suitable

for bacteria. The best candidates are internal transcribed spacers (ITS), but these are more difficult to manipulate. The main problem with ITS is size polymorphism, with a size range of 361–1475 bases in UNITE 7.1 [3] (unlike 16S where 95% of the sequences have a length between 1205 and 1556 bases). Most studies describing ITS data analyses process either (i) paired-end reads but filter out non-overlapping, non-mergeable reads, thus systematically discarding taxa with longer ITS, or (ii) single-end reads, thus limiting taxonomic resolution and losing the benefit of information contained in longer sequences [4, 5].

Maria Bernard is a bioinformatics engineer. She is a member of a platform team conducting NGS sequence analysis and designing software. She specializes in workflow development in particular for metabarcoding analysis.
Olivier Rué is a bioinformatics engineer. He is in charge of data analysis at the Migale bioinformatics facility. He specializes in the analysis of metabarcoding and metagenomics data.
Mahendra Mariadassou has a PhD in statistics. He is involved in the development of new statistical methods and tools for metabarcoding analysis.
Géraldine Pascal has a PhD in bioinformatics and coordinates the FROGS project. She is currently involved in designing solutions for long read problems, workflow development and metagenomics analysis.
Submitted: 19 April 2021. Received (in revised form): 19 July 2021

© The Author(s) 2021. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

1



FROGS'docs

Website: <http://frogs.toulouse.inrae.fr>

All scripts on Github:

<https://github.com/geraldinepascal/FROGS.git>



The user-friendly and Galaxy-supported pipeline FROGS analyses large sets of DNA amplicons sequences accurately and rapidly, essential for microbe community studies.

- FROGS was designed to support multiplexed and demultiplexed sequences.
- FROGS supports 16S, 15S, 23S, 11S and other amplicon reads.
- FROGS supports short or long reads.
- The preprocessing tool is optimized to paired sequences merging, cleaning and denoising.
- The clustering tool uses Swarm with a local clustering (inverted), not a global clustering (forward) the other software do.
- Clusters removal tool uses VSEARCH combined with an innovative chimera cross-validation.
- A filtering tool allows to remove noisy data.
- Affiliation tool returns taxonomic affiliation for each OTU using two methods with a unique multi-affiliation output.
- FROGS offers numerous banks for affiliation step (cf. list).
- A set of statistical results and numerous graphical business are also produced.
- FROGS is designed for non-specialists thanks to its Galaxy interface, but it is also available with command lines: `gfish`.
- Its tools can be used independently, or as a workflow.
- To install FROGS, check or galaxy instance.

FROGS was tested on many datasets

- The "FROGS 16S Benchmark test" and "FROGS ITS Benchmark test" show comparisons between FROGS and other popular pipelines.

Standard Operation Procedure for amplicons
i.e. 16S, 23S, 11S, ...

Standard Operation Procedure for data with unmetabolized amplicons
i.e. 11S, 16S2, 16S1, ...

Citation

The publication :

Fredrik Sæviik, Lucas Auer, Marie Demard, Mélanie Vanreusel, Laurent Daguin, Rikie Yano, Sarah Niaman, Guillaume Hernandez-Pajuelo, Sylvie Combes, Catherine Pascal, FROGS: Find, Rapidly, OTU with Galaxy Solution, *Bioinformatics*, Volume 34, Issue 8, 15 April 2019, Pages 1281-1284

To test FROGS

Play with FROGS on the Galaxy server of Toulouse

- Register an account to genodiv biota perform via a form [Genodiv Platform](#)
- Enter your credential (once for apache server connection and a second times for galaxy platform connection): [Genodiv Server](#) (on the top) lower right
- Get data. In also simulated data (4th history of the documentation below) from [sequence S22](#).
- Play with the workflow: [Workflow Test](#) and export it on your account (green cross on the top).
- To learn more: [Formalton documentation](#)

Help FROGS assessments with command lines for analyse [16S](#) [16S/23S](#)

Get data bulk with sequences from UMI and sra database. doi.org/10.13454/16S/23S

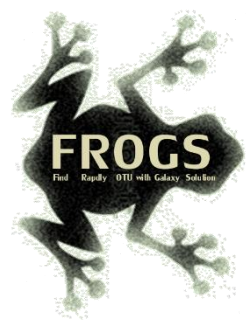
Help FROGS assessments with command lines for analyse [11S](#) [16S/11S](#)

Get data bulk with sequences from UNITE. doi.org/10.13454/11S

License

GNU GPL3 (open)

REPUBLICQUE FRANÇAISE



To contact

FROGS support:

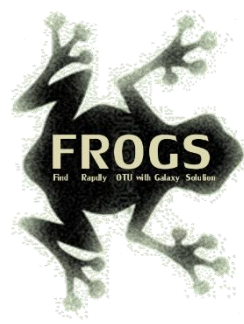
frogs-support@inrae.fr

Galaxy:

support.sigenae@inrae.fr

Newsletter – subscription request:

frogs-support@inrae.fr



Play list FROGS:

https://www.deezer.com/fr/playlist/5233843102?utm_source=deezer&utm_content=playlist-5233843102&utm_term=18632989_1545296531&utm_medium=web