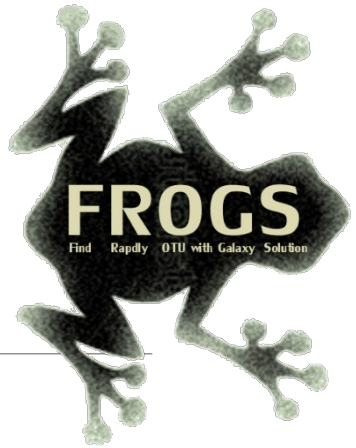


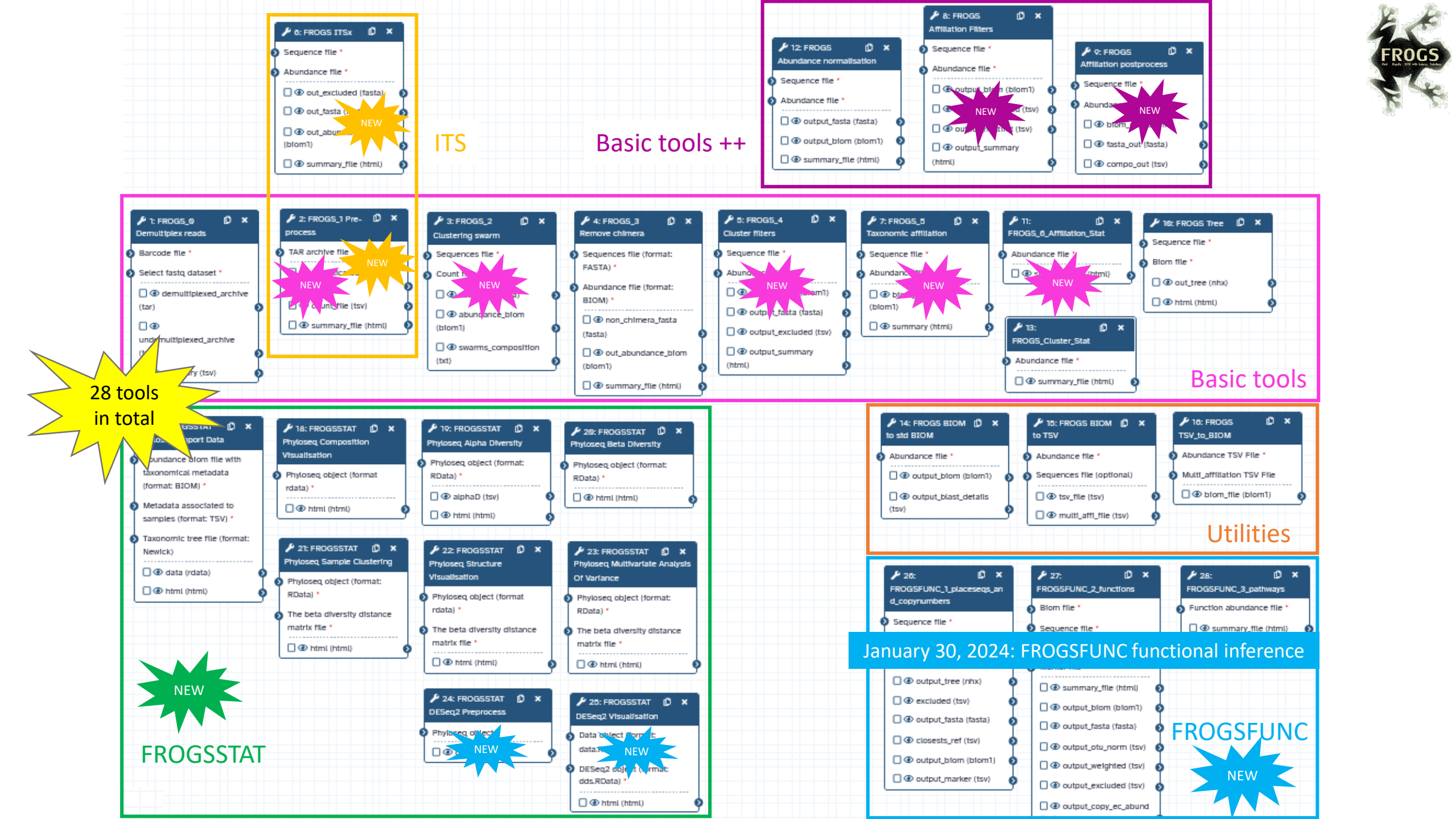
Les mardis de la grenouille

January 2024 - Webinar

FROGS 4.1 - What's new?

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

.


6: FROGS ITSx
Sequence file *
Abundance file *
out_excluded (fasta)
out_fasta
out_abundance (biom1)
summary_file (html)
NEW
ITS
Basic tools ++
12: FROGS Abundance normalisation
Sequence file *
Abundance file *
output_fasta (fasta)
output_biom (biom1)
summary_file (html)
8: FROGS Affiliation Filters
Sequence file *
Abundance file *
output_biom (biom1)
output_fasta (fasta)
output_tsv (tsv)
output_summary (html)
NEW
9: FROGS Affiliation postprocess
Sequence file *
Abundance file *
biom_out (biom1)
fasta_out (fasta)
compo_out (tsv)
NEW
1: FROGS_0 Demultiplex reads
Barcode file *
Select fastq dataset *
demultiplexed_archive (tar)
undemultiplexed_archive
summary (tsv)
2: FROGS_1 Pre-process
TAR archive file
count_file (tsv)
summary_file (html)
NEW
3: FROGS_2 Clustering swarm
Sequences file *
Count file *
abundance_biom (biom1)
swarms_composition (txt)
NEW
4: FROGS_3 Remove chimera
Sequences file (format: FASTA) *
Abundance file (format: BIOM) *
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
summary_file (html)
5: FROGS_4 Cluster filters
Sequence file *
Abundance file *
biom (biom1)
output_fasta (fasta)
output_excluded (tsv)
output_summary (html)
NEW
7: FROGS_5 Taxonomic affiliation
Sequence file *
Abundance file *
biom (biom1)
summary (html)
NEW
11: FROGS_6_Affiliation_Stat
Abundance file *
html (html)
NEW
13: FROGS_Cluster_Stat
Abundance file *
summary_file (html)
16: FROGS Tree
Sequence file *
Biom file *
out_tree (nhx)
html (html)
Basic tools
28 tools in total
18: FROGSSTAT Phyloseq Composition Visualisation
Phyloseq object (format rdata) *
html (html)
19: FROGSSTAT Phyloseq Alpha Diversity
Phyloseq object (format: RData) *
alphaD (tsv)
html (html)
20: FROGSSTAT Phyloseq Beta Diversity
Phyloseq object (format: RData) *
html (html)
21: FROGSSTAT Phyloseq Sample Clustering
Phyloseq object (format: RData) *
The beta diversity distance matrix file *
html (html)
22: FROGSSTAT Phyloseq Structure Visualisation
Phyloseq object (format rdata) *
The beta diversity distance matrix file *
html (html)
23: FROGSSTAT Phyloseq Multivariate Analysis Of Variance
Phyloseq object (format: RData) *
The beta diversity distance matrix file *
html (html)
24: FROGSSTAT DESeq2 Preprocess
NEW
25: FROGSSTAT DESeq2 Visualisation
NEW
html (html)
NEW
FROGSSTAT
14: FROGS BIOM to std BIOM
Abundance file *
output_biom (biom1)
output_blast_details (tsv)
15: FROGS BIOM to TSV
Abundance file *
Sequences file (optional)
tsv_file (tsv)
multi_affi_file (tsv)
16: FROGS TSV_to_BIOM
Abundance TSV File *
Multi_affiliation TSV File
biom_file (biom1)
Utilities
January 30, 2024: FROGSFUNC functional inference
26: FROGSFUNC_1_placeseqs_and_copynumbers
Sequence file *
output_tree (nhx)
excluded (tsv)
output_fasta (fasta)
closests_ref (tsv)
output_biom (biom1)
output_marker (tsv)
27: FROGSFUNC_2_functions
Biom file *
Sequence file *
summary_file (html)
output_biom (biom1)
output_fasta (fasta)
output_otu_norm (tsv)
output_weighted (tsv)
output_excluded (tsv)
output_copy_ec_abund
28: FROGSFUNC_3_pathways
Function abundance file *
summary_file (html)
FROGSFUNC
NEW

Avant-propos

New tool names

Tool names with numbers to make it easier to link tools, especially basic tools.

More name blocks.

FROGS_
FROGSSTATS_
FROGSFUNC_

FROGS_0 Demultiplex reads Attribute reads to samples in function of inner barcode

FROGS_1 Pre-process merging, denoising and dereplication

FROGS_2 Clustering swarm Single-linkage clustering on sequences

FROGS_Cluster_Stat Process some metrics on clusters

FROGS_3 Remove chimera Remove PCR chimera in each sample

FROGS_4 Cluster filters Filters clusters on several criteria.

FROGS ITSx Extract the highly variable ITS1 and ITS2 subregions from ITS sequences

FROGS_5 Taxonomic affiliation Taxonomic affiliation of each ASV's seed by RDPtools and BLAST

FROGS Affiliation Filters Filters ASVs on several affiliation criteria

FROGS Affiliation postprocess Aggregates ASVs based on alignment metrics

FROGS Abundance normalisation Normalise ASV abundance.

FROGS Tree Reconstruction of phylogenetic tree

FROGS_6_Affiliation_Stat Process some metrics on taxonomies

FROGS BIOM to std BIOM Converts a FROGS BIOM in fully compatible BIOM

FROGS BIOM to TSV Converts a BIOM file in TSV file

FROGS TSV_to_BIOM Converts a TSV file in a BIOM file 1

FROGSSTAT Phyloseq Import Data from 3 files: biomfile, samplefile, treefile

FROGSSTAT Phyloseq Composition Visualisation with bar plot and composition plot

FROGSSTAT Phyloseq Alpha Diversity with richness plot

FROGSSTAT Phyloseq Beta Diversity distance matrix

FROGSSTAT Phyloseq Sample Clustering of samples using different linkage methods

FROGSSTAT Phyloseq Structure Visualisation with heatmap plot and ordination plot

FROGSSTAT Phyloseq Multivariate Analysis Of Variance perform Multivariate Analysis of Variance (MANOVA)

FROGSSTAT DESeq2 Preprocess import a Phyloseq object and prepare it for DESeq2 differential abundance analysis

FROGSSTAT DESeq2 Visualisation to extract and visualise differentially abundant ASVs or functions

FROGSFUNC_1_placeseqs_and_copynumbers Places ASVs into a reference phylogenetic tree.

FROGSFUNC_2_functions Calculates functions abundances in each sample.

FROGSFUNC_3_pathways Calculates pathway abundances in each sample.

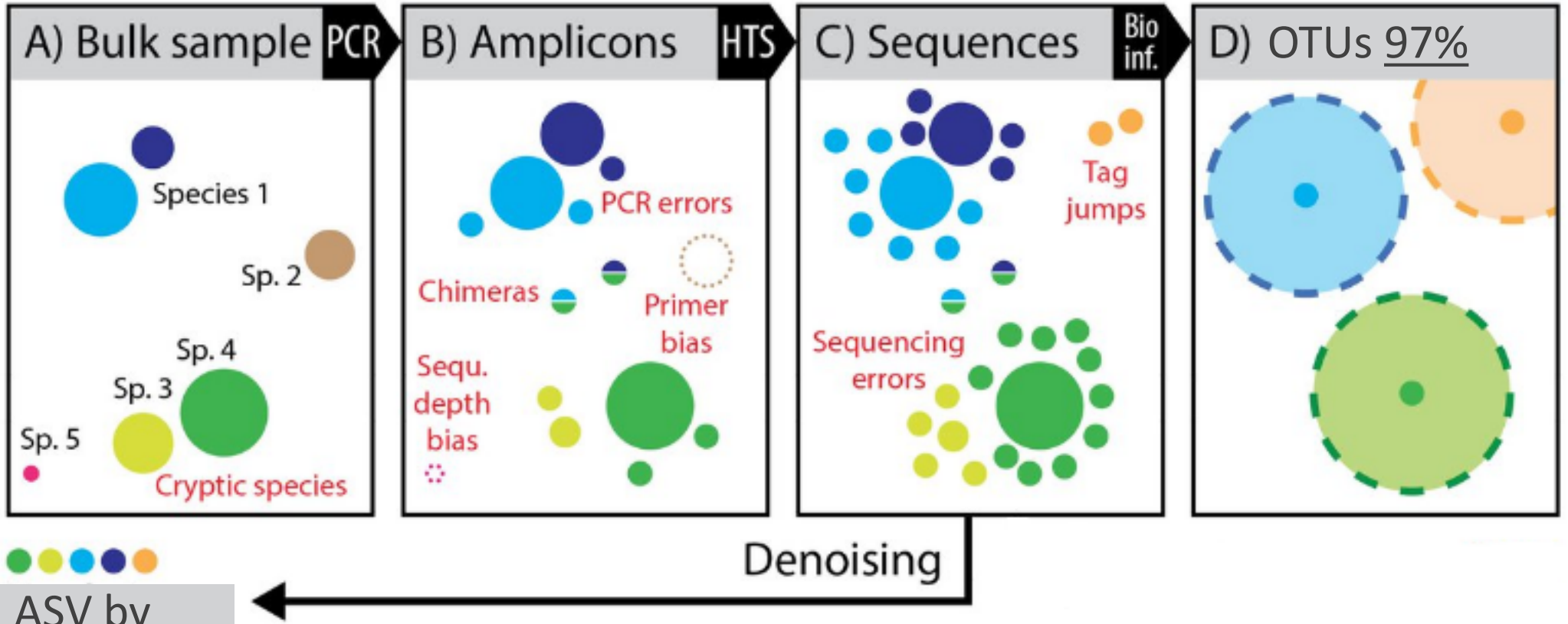
OTU → ASV

A long-standing discussion

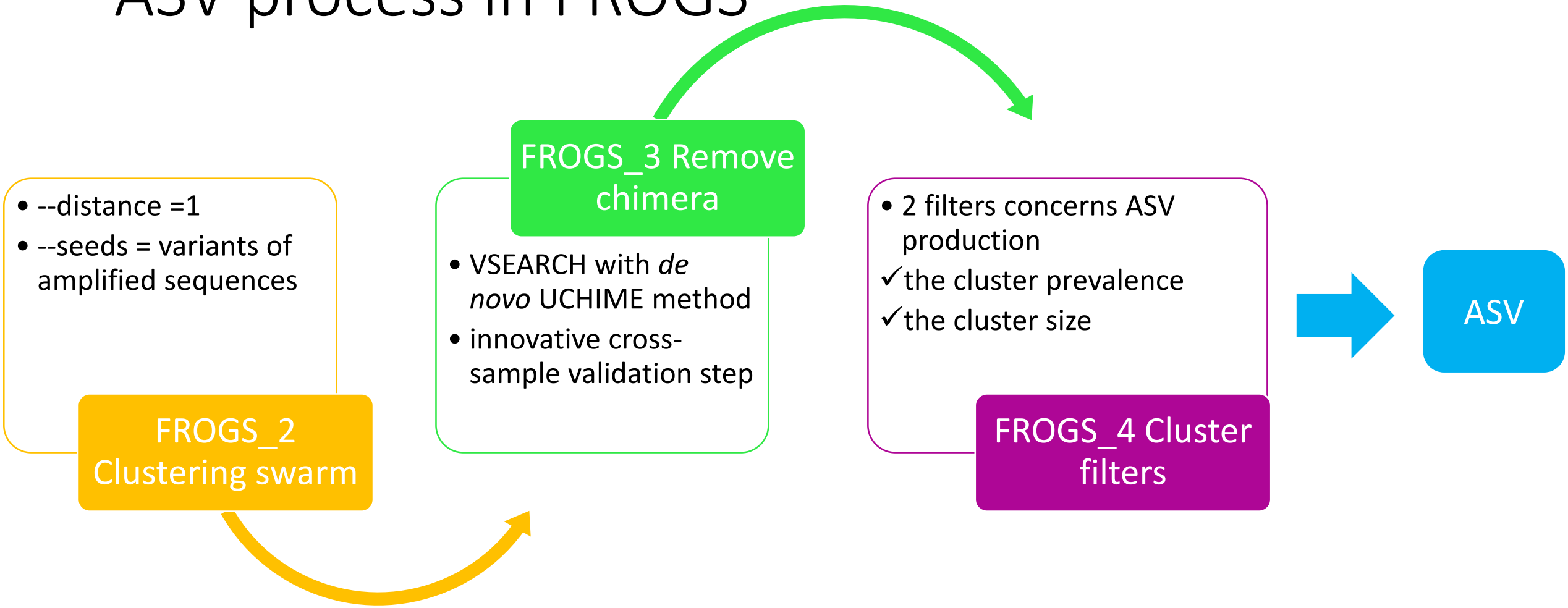
- The ASV vs OTU debate launched by the arrival of dada2 is not so new and had been bothering us for several months/years.
- In fact, the debate largely preceded the term "ASV", and is precisely what made us opt for Swarm in FROGS (just under 10 years ago).
- To quote the author of swarm:
  “The traditional term "OTU" is negatively charged nowadays. The ASV vs OTU debate is creating confusion in the community and some users now think that all methods producing "OTUs" use a fixed clustering threshold (i.e. 97%-similarity) and are inherently bad. Of course, this is not the case and there are several methods published before the ASV term was coined that produce ASV-like clusters, swarm included.” To avoid that confusion, swarm's manual now only uses the generic term "cluster".
  https://github.com/torognes/swarm/commit/0bb491f9bf646c22a5363c27dc31a6d4b2ad335d “

A question of vocabulary

- A few years ago, the semantic problem was the opposite, and any method that didn't produce OTUs was questioned or even disqualified.
- At the start of FROGS, we therefore chose to call our clusters "OTUs" at the end of the analysis (once the filters had been applied), but it's only a question of vocabulary, and the clusters produced by FROGS/swarm are very close to ASV in their construction.
- In any case, they look much more like ASVs than "fixed threshold" OTUs. The best thing would have been to use a new term, but Fréderic Mahé didn't make that choice at the time introducing a new term could have led to confusion.
- Since version 4.1.0 of FROGS, we have changed our vocabulary and all OTU terms have been changed to **cluster** or **ASV** in FROGS tools and outputs.


A) Bulk sample
PCR
Species 1
Sp. 2
Sp. 4
Sp. 3
Sp. 5
Cryptic species
B) Amplicons
HTS
PCR errors
Chimeras
Primer bias
Sequ. depth bias
C) Sequences
Bio inf.
Tag jumps
Sequencing errors
D) OTUs 97%
Denoising
ASV by dada2 or FROGS

ASV process in FROGS



Swarm --seeds produces:
variants of amplified sequences.
“Variants” because the output sequences are all different; but with no constraints on the extent of variation - one nucleotide to infinity.

Received: 10 February 2023 | Revised: 5 June 2023 | Accepted: 6 July 2023
DOI: 10.1111/1755-0998.13847

FROM THE COVER

MOLECULAR ECOLOGY RESOURCES WILEY

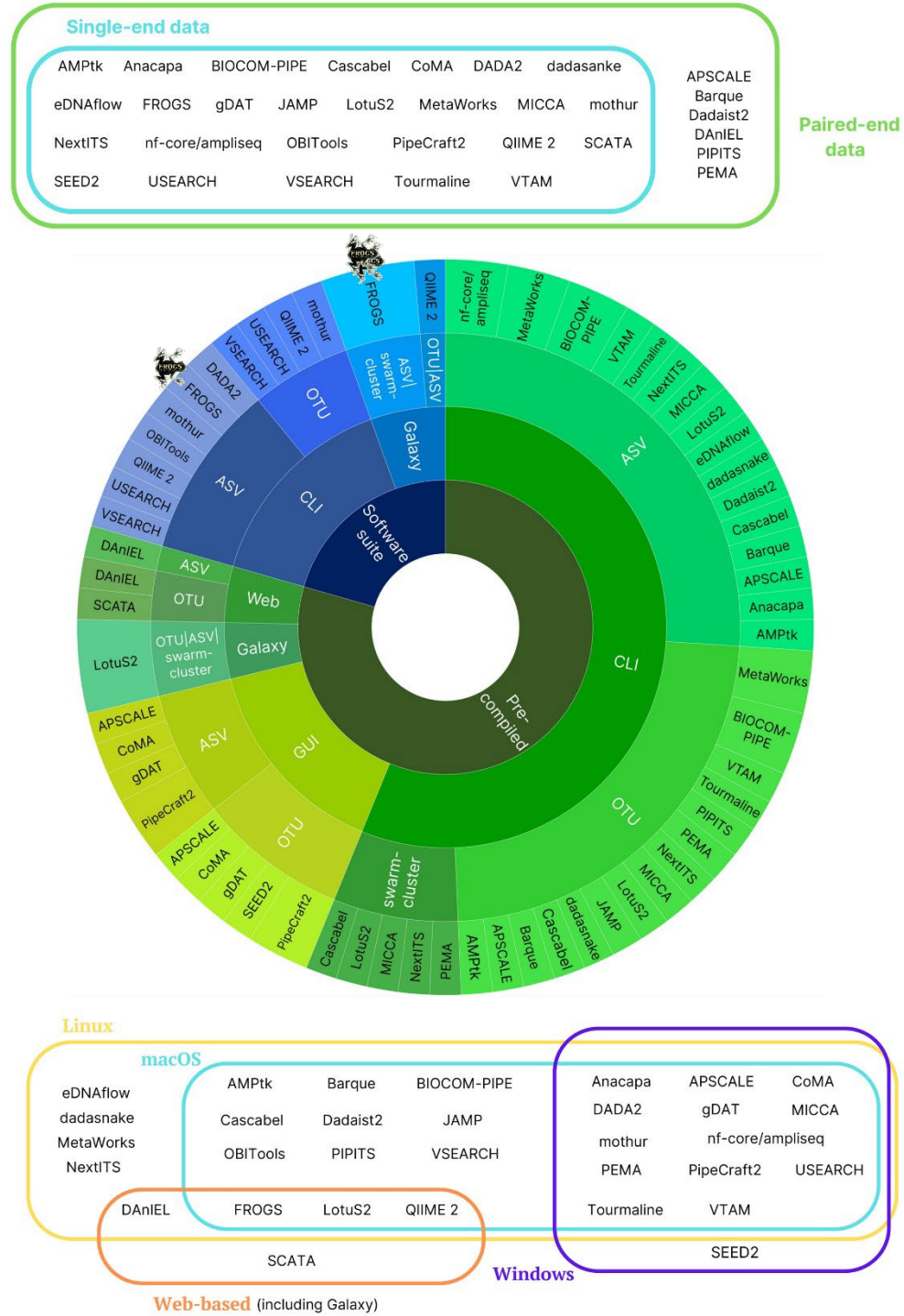
A pile of pipelines: An overview of the bioinformatics software for metabarcoding data analyses

Ali Hakimzadeh[1] | Alejandro Abdala Asbun[2] | Davide Albanese[3] | Maria Bernard[4,5] | Dominik Buchner[6] | Benjamin Callahan[7] | J. Gregory Caporaso[8] | Emily Curd[9] | Christophe Djemiel[10] | Mikael Brandström Durling[11] | Vasco Elbrecht[6] | Zachary Gold[12] | Hyun S. Gweon[13,14] | Mehrdad Hajibabaei[15] | Falk Hildebrand[16,17] | Vladimir Mikryukov[1] | Eric Normandeau[18] | Ezgi Özkurt[16,17] | Jonathan M. Palmer[19] | Géraldine Pascal[20] | Teresita M. Porter[15] | Daniel Straub[21] | Martti Vasar[1] | Tomáš Větrovský[22] | Haris Zafeiropoulos[23] | Sten Anslan[1,24]

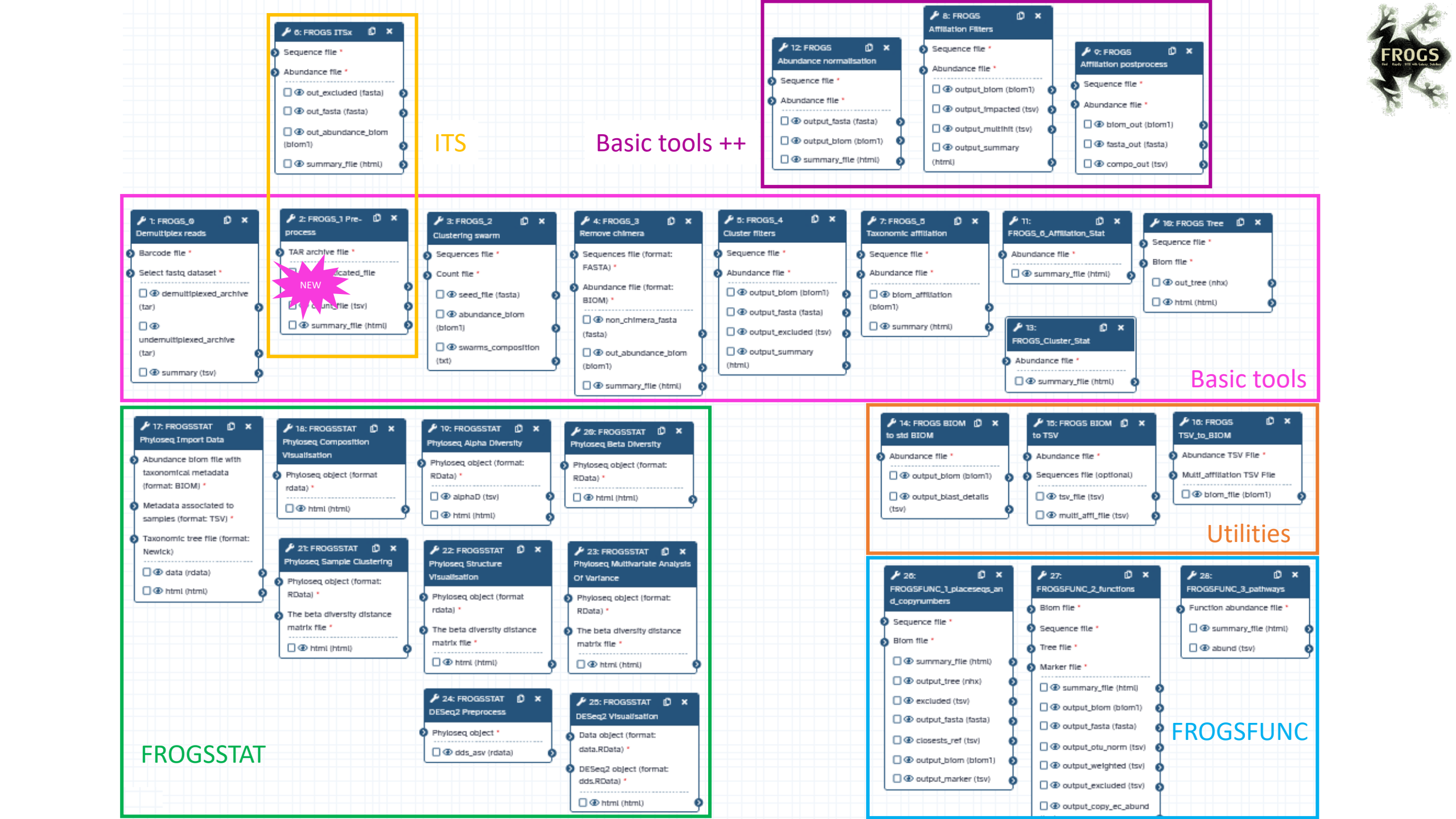
[1]Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia
[2]Department of Marine Microbiology and Biogeochemistry, NIOZ Royal Netherlands Institute for Sea Research, Texel, Netherlands
[3]Unit of Computational Biology, Research and Innovation Centre, Fondazione Edmund Mach, Italy
[4]INRAE, AgroParisTech, GABI, Université Paris-Saclay, Jouy-en-Josas, France
[5]INRAE, SIGENAE, Jouy-en-Josas, France
[6]Aquatic Ecosystem Research, University of Duisburg-Essen, Essen, Germany
[7]Department of Population Health and Pathobiology, College of Veterinary Medicine and Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, USA
[8]Center for Applied Microbiome Science, Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, Arizona, USA
[9]Vermont Biomedical Research Network, University of Vermont, Burlington, Vermont, USA
[10]Agroécologie, INRAE, Institut Agro, Univ. Bourgogne Franche-Comté, Dijon, France
[11]Department of Forest Mycology and Plant Pathology, Swedish University of Agricultural Sciences, Uppsala, Sweden
[12]Zachary Gold, NOAA Pacific Marine Environmental Laboratory, Seattle, Washington, USA
[13]UK Centre for Ecology & Hydrology, Oxfordshire, UK
[14]School of Biological Sciences, University of Reading, Reading, UK
[15]Department of Integrative Biology and Centre for Biodiversity Genomics, University of Guelph, Guelph, Ontario, Canada
[16]Gut Microbes & Health, Quadram Institute Bioscience, Norfolk, UK
[17]Earlham Institute, Norwich Research Park, Norfolk, UK
[18]Institut de Biologie Intégrative et des Systèmes, Université Laval, Québec, Québec, Canada
[19]Center for Forest Mycology Research, Northern Research Station, US Forest Service, Madison, Wisconsin, USA
[20]GenPhySE, Université de Toulouse, INRAE, ENVT, Castanet Tolosan, France
[21]Quantitative Biology Center (QBiC), University of Tübingen, Tübingen, Germany
[22]Laboratory of Environmental Microbiology, Institute of Microbiology of the Czech Academy of Sciences, Praha, Czech Republic
[23]KU Leuven, Department of Microbiology, Immunology and Transplantation, Rega Institute for Medical Research, Laboratory of Molecular Bacteriology, Leuven, Belgium
[24]Department of Biological and Environmental Science, University of Jyväskylä, Jyväskylä, Finland

Correspondence
Sten Anslan, Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia.
Email: sten.anslan@ut.ee

Present address
Jonathan M. Palmer, Genencor Technology Center, IFF, Palo Alto, California, USA



New tools, new parameters


6: FROGS ITSx
Sequence file *
Abundance file *
out_excluded (fasta)
out_fasta (fasta)
out_abundance_biom (biom1)
summary_file (html)
ITS
12: FROGS Abundance normalisation
Sequence file *
Abundance file *
output_fasta (fasta)
output_biom (biom1)
summary_file (html)
8: FROGS Affiliation filters
Sequence file *
Abundance file *
output_biom (biom1)
output_impacted (tsv)
output_multihit (tsv)
output_summary (html)
9: FROGS Affiliation postprocess
Sequence file *
Abundance file *
biom_out (biom1)
fasta_out (fasta)
compo_out (tsv)
Basic tools ++
1: FROGS_0 Demultiplex reads
Barcode file *
Select fastq dataset *
demultiplexed_archive (tar)
undemultiplexed_archive (tar)
summary (tsv)
2: FROGS_1 Pre-process
TAR archive file *
NEW
summary_file (html)
3: FROGS_2 Clustering swarm
Sequences file *
Count file *
seed_file (fasta)
abundance_biom (biom1)
swarms_composition (txt)
4: FROGS_3 Remove chimera
Sequences file (format: FASTA) *
Abundance file (format: BIOM) *
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
summary_file (html)
5: FROGS_4 Cluster filters
Sequence file *
Abundance file *
output_biom (biom1)
output_fasta (fasta)
output_excluded (tsv)
output_summary (html)
7: FROGS_5 Taxonomic affiliation
Sequence file *
Abundance file *
biom_affiliation (biom1)
summary (html)
11: FROGS_6_Affiliation_Stat
Abundance file *
summary_file (html)
13: FROGS_Cluster_Stat
Abundance file *
summary_file (html)
10: FROGS Tree
Sequence file *
Biom file *
out_tree (nhx)
html (html)
Basic tools
17: FROGSSTAT Phyloseq Import Data
Abundance biom file with taxonomical metadata (format: BIOM) *
Metadata associated to samples (format: TSV) *
Taxonomic tree file (format: Newick)
data (rdata)
html (html)
18: FROGSSTAT Phyloseq Composition Visualisation
Phyloseq object (format rdata) *
html (html)
19: FROGSSTAT Phyloseq Alpha Diversity
Phyloseq object (format: RData) *
alphaD (tsv)
html (html)
20: FROGSSTAT Phyloseq Beta Diversity
Phyloseq object (format: RData) *
html (html)
21: FROGSSTAT Phyloseq Sample Clustering
Phyloseq object (format: RData) *
The beta diversity distance matrix file *
html (html)
22: FROGSSTAT Phyloseq Structure Visualisation
Phyloseq object (format rdata) *
The beta diversity distance matrix file *
html (html)
23: FROGSSTAT Phyloseq Multivariate Analysis Of Variance
Phyloseq object (format: RData) *
The beta diversity distance matrix file *
html (html)
24: FROGSSTAT DESeq2 Preprocess
Phyloseq object *
dds_asv (rdata)
25: FROGSSTAT DESeq2 Visualisation
Data object (format: data.RData) *
DESeq2 object (format: dds.RData) *
html (html)
FROGSSTAT
14: FROGS BIOM to std BIOM
Abundance file *
output_biom (biom1)
output_blast_details (tsv)
15: FROGS BIOM to TSV
Abundance file *
Sequences file (optional)
tsv_file (tsv)
multi_affi_file (tsv)
16: FROGS TSV_to_BIOM
Abundance TSV File *
Multi_affiliation TSV File
biom_file (biom1)
Utilities
26: FROGSFUNC_1_placeseqs_and_copynumbers
Sequence file *
Biom file *
summary_file (html)
output_tree (nhx)
excluded (tsv)
output_fasta (fasta)
closests_ref (tsv)
output_biom (biom1)
output_marker (tsv)
27: FROGSFUNC_2_functions
Biom file *
Sequence file *
Tree file *
Marker file *
summary_file (html)
output_biom (biom1)
output_fasta (fasta)
output_otu_norm (tsv)
output_weighted (tsv)
output_excluded (tsv)
output_copy_ec_abund
28: FROGSFUNC_3_pathways
Function abundance file *
summary_file (html)
abund (tsv)
FROGSFUNC

Pre-process tool

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

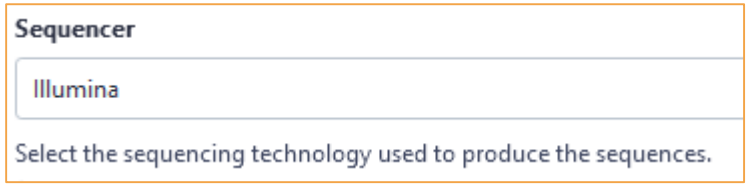
Bioinformatics (2014) 30 (5):614–620 doi.org/10.1093/bioinformatics/btt593
PEAR: a fast and accurate Illumina Paired-End reAd mergeR
J. Zhang, K. Kobert, T. Flouri, A. Stamatakis,

EMBnet Journal, Vol17 no1. doi : 10.14806/ej.17.1.200
Cutadapt removes adapter sequences from high-throughput sequencing reads
Marcel Martin

Exemples of different preprocess panels for your future personal uses.

A – for short reads from illumina

Illumina



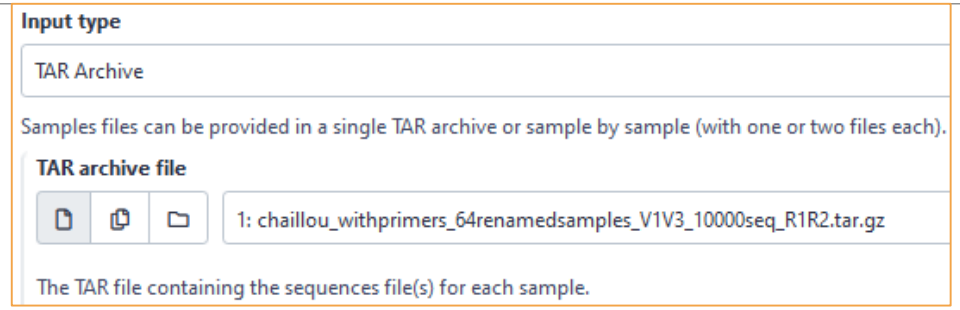
A – for short reads from illumina

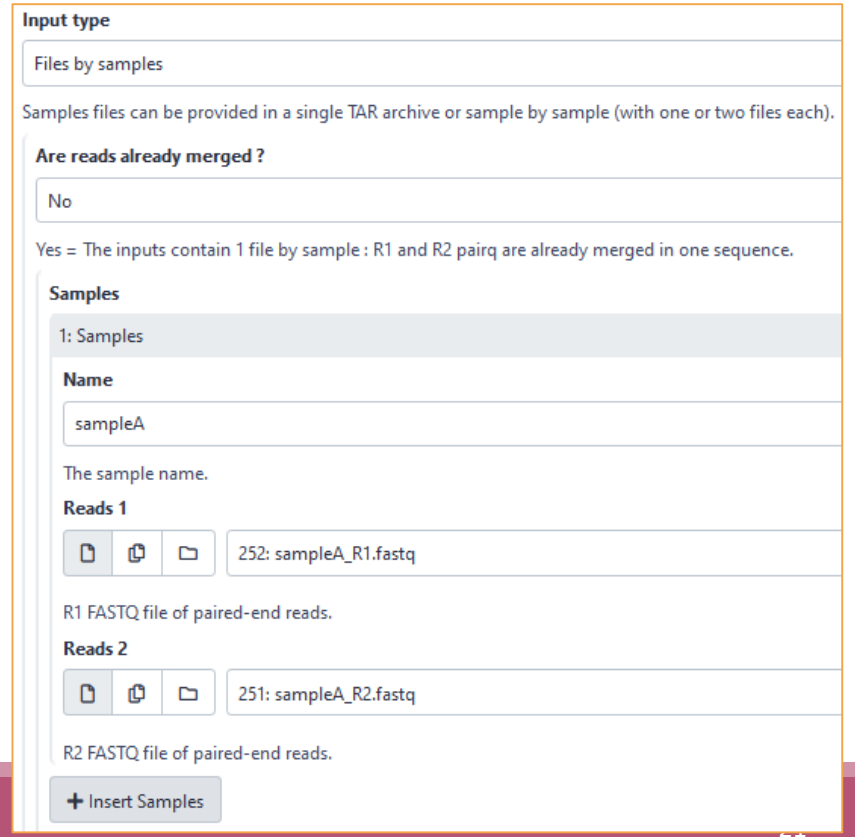
Illumina

Archive

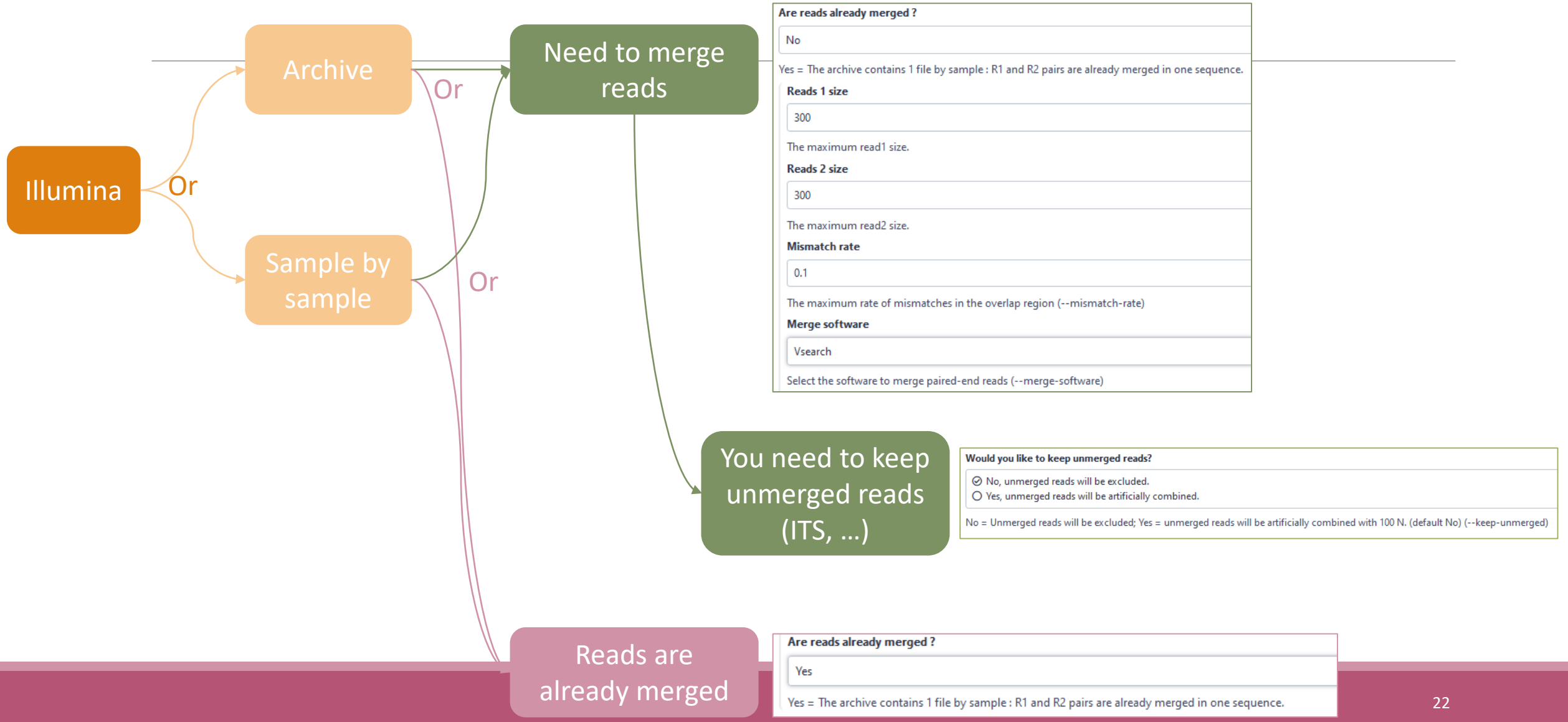
Or

Sample by sample

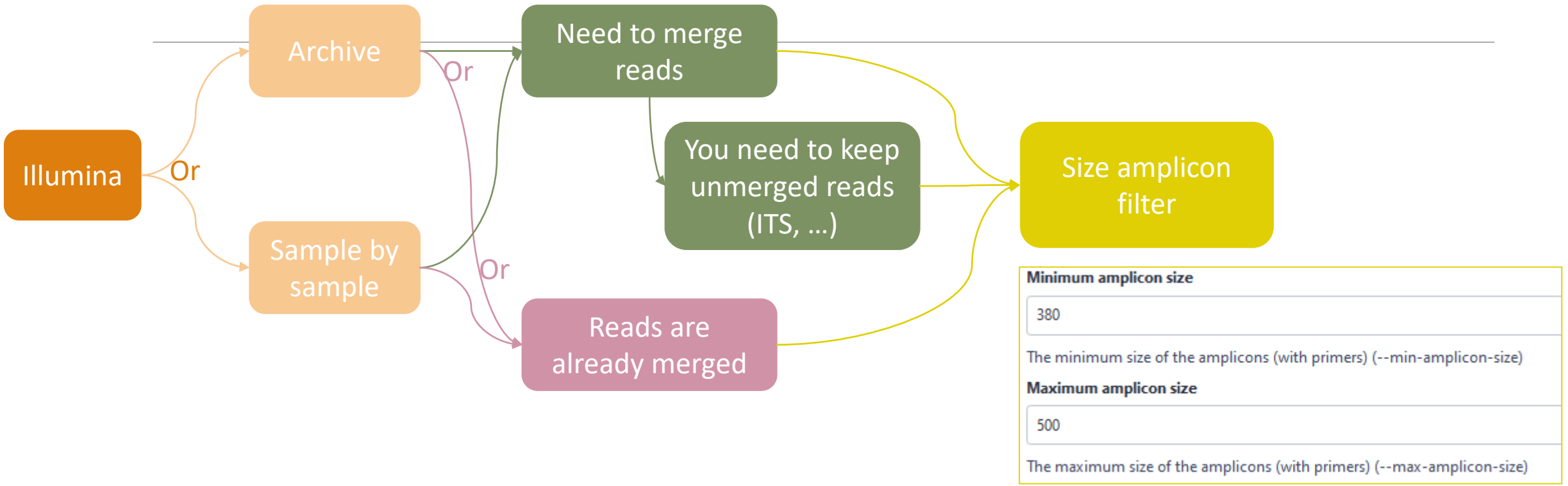




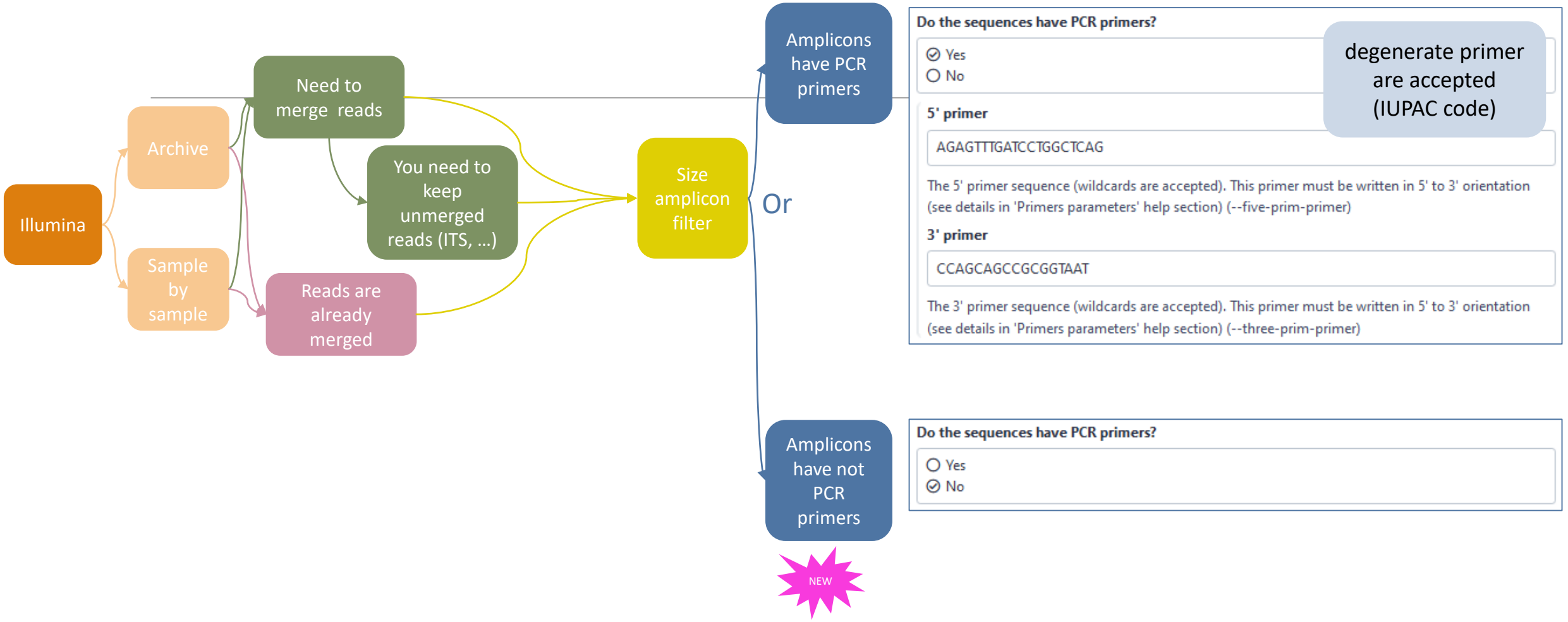
A – for short reads from illumina



A – for short reads from illumina



A – for short reads from illumina



B – for long reads from Pacbio or ONT

B – for long reads from Pacbio or ONT

Archive

Longreads

Or

Sample by sample

Input type

TAR Archive

Samples files can be provided in single archive or with one file by sample.

TAR archive file

1: longread_hifi_16S_8species.tar.gz

The TAR file containing the sequences file for each sample.

Sequencer

Longreads (PACBIO, ONT)

Select the sequencing technology used to produce the sequences.

Input type

One file by sample

Samples files can be provided in single archive or with one file by sample.

Samples

1: Samples

Name

Mockbact

The sample name.

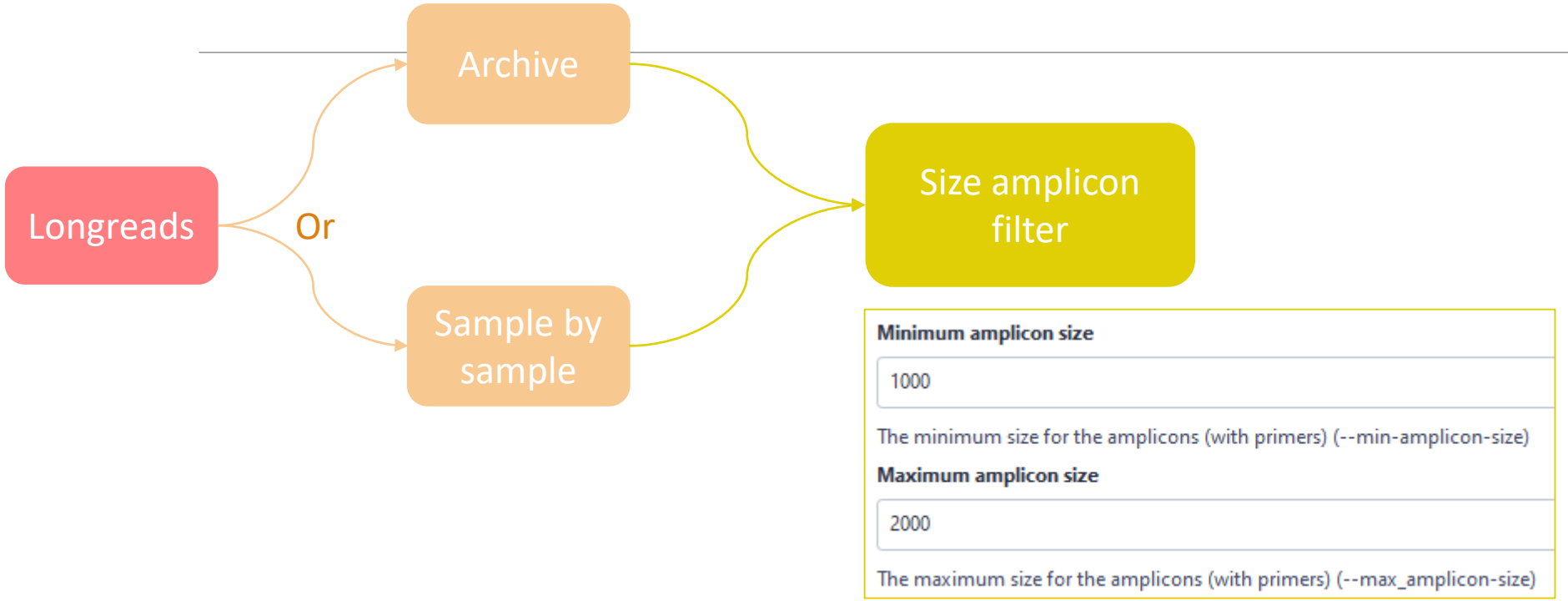
Sequence file

11: Mockbact.fastq

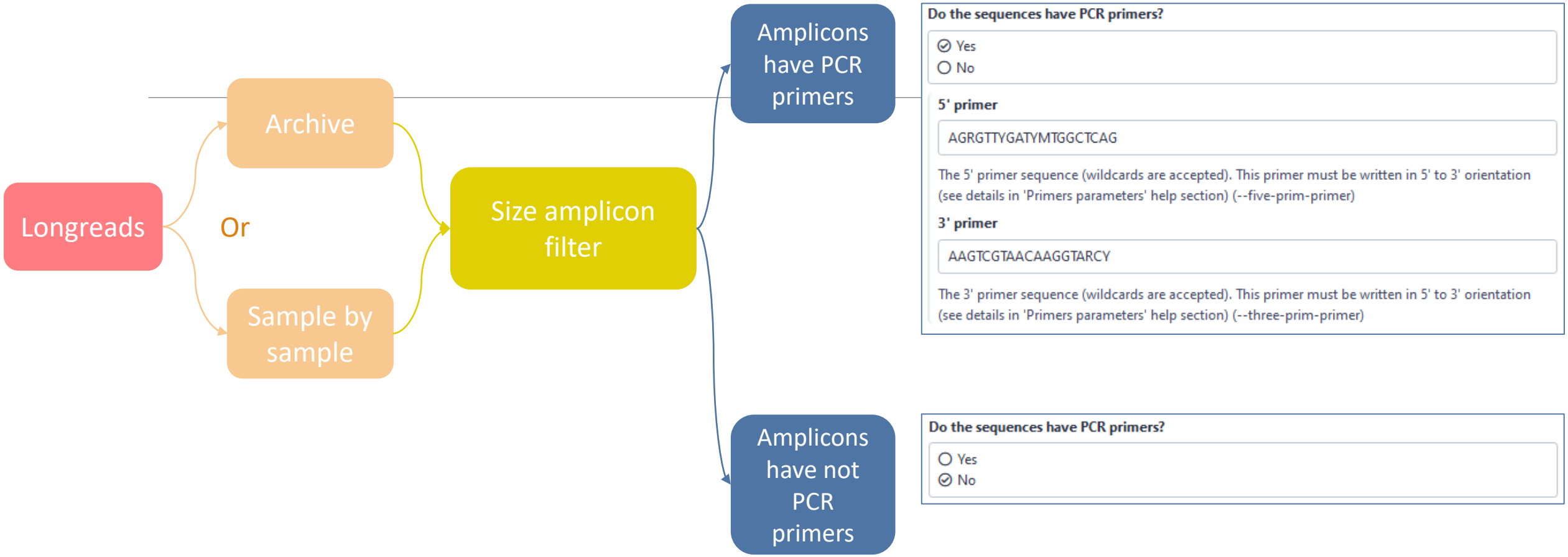
FASTQ file of sample.

+ Insert Samples

B – for long reads from Pacbio or ONT

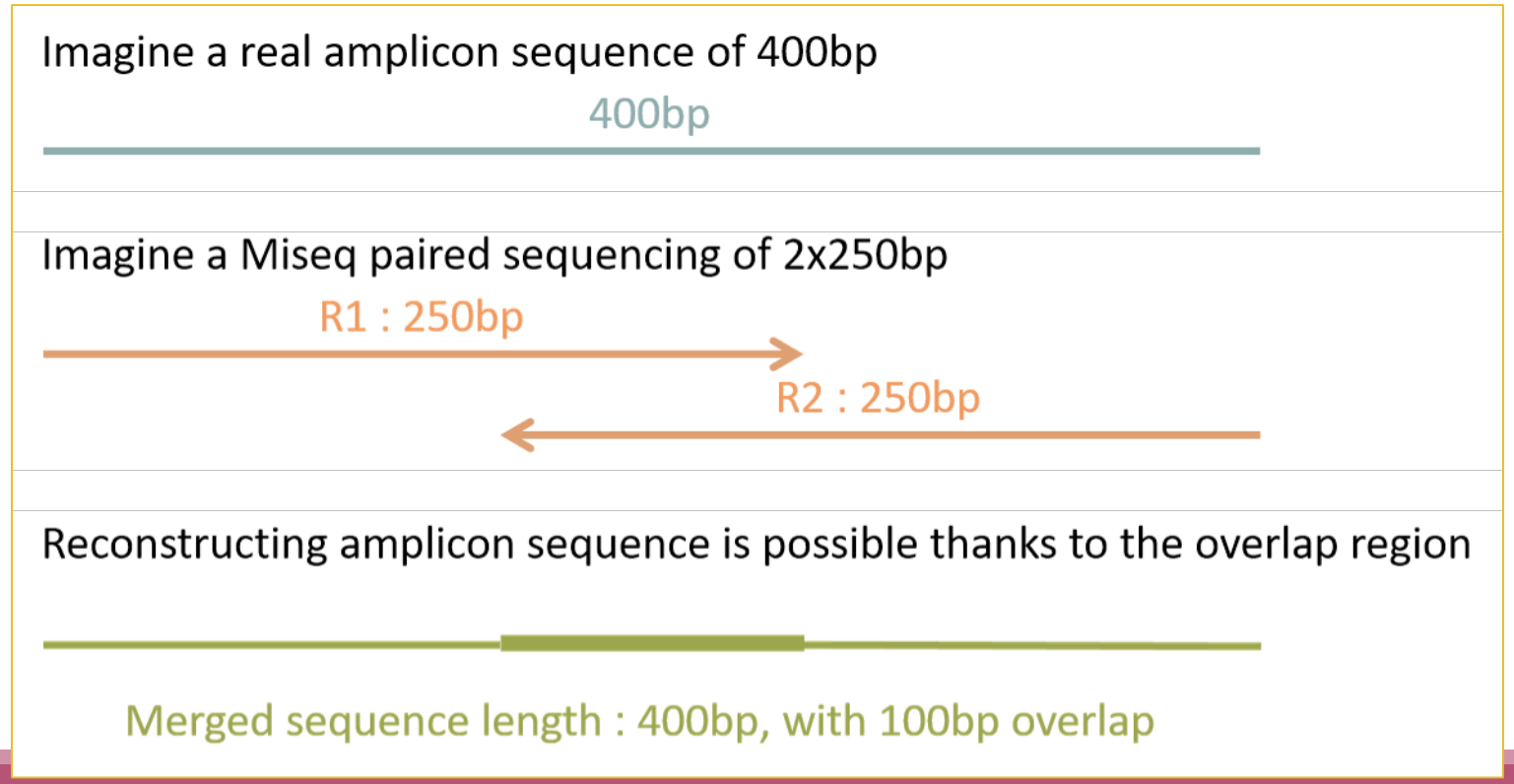


B – for long reads from Pacbio or ONT



The aim of Vsearch is to merge R1 with R2

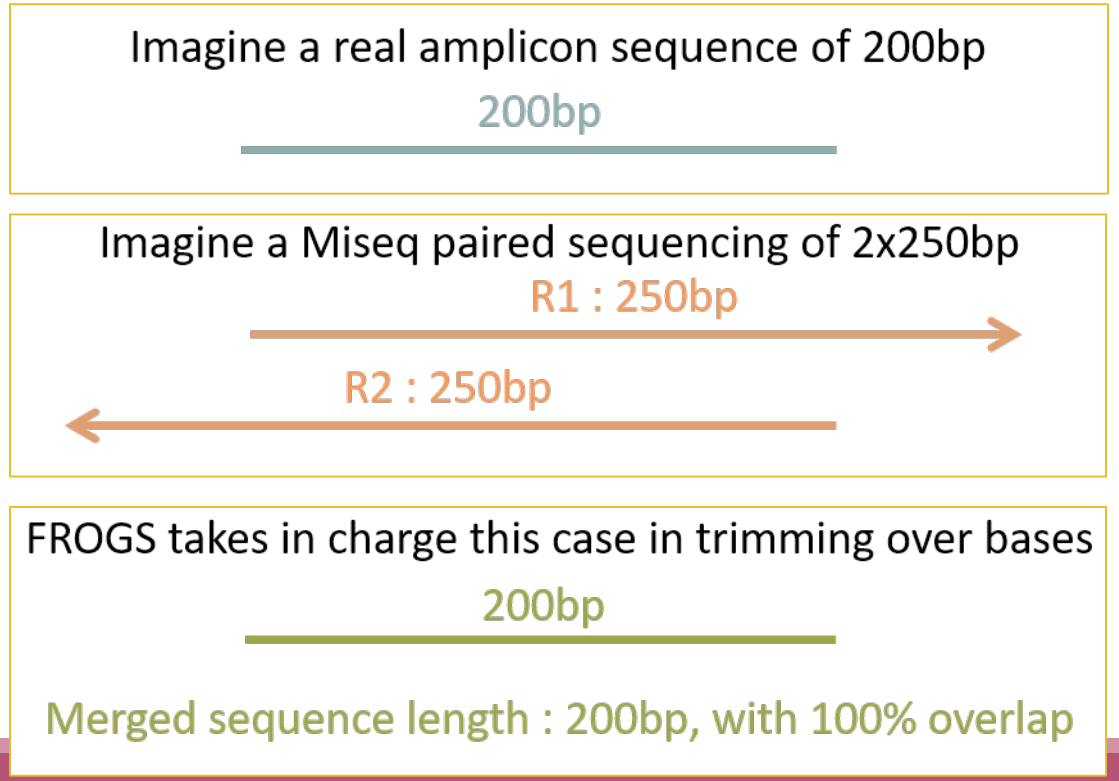
Case of a sequencing of overlapping sequences: case of 16S V3-V4 amplicon MiSeq sequencing:



The aim of Vsearch is to merge R1 with R2

NEW

Case of a sequencing of over-overlapping sequences:



Practice:

Exercise

Go to « 16S » history

Launch the pre-process tool on that data set

→ objective: understand Vsearch software

16S dataset presentation:

A real analysis provided by Stéphane Chaillou *et al.*

Comparison of meat and seafood bacterial communities.

8 environment types (EnvType) :

- Meat → Ground Beef, Ground veal, Poultry sausage, Diced bacon
- Seafood → Cooked schrimps, Smoked salmon, Salmon filet, Cod filet

Chaillou, S. et al (2015). Origin and ecological selection of core and food-specific bacterial communities associated with meat and seafood spoilage. ISME J, 9(5):1105-1118.

16S dataset presentation:

BH VH MV DL CD SF FS FC

From Chaillou paper, we produced simulated data:

- 64 samples of 16S amplicons
- R1 and R2 overlapping reads of 300 bases.
- 8 replicates per condition
- with errors among the linear curve 2.54e-1 2.79e-1
- with 10% chimeras
- Primers for V1-V3:
  - 5’ AGAGTTTGATCCTGGCTCAG 3’
  - 5’ CCAGCAGCCGCGGTAAT 3’

Chaillou, S. et al (2015). Origin and ecological selection of core and food-specific bacterial communities associated with meat and seafood spoilage. ISME J, 9(5):1105-1118.

FROGS_1 Pre-process merging, denoising and dereplication (Galaxy Version 4.1.0+galaxy1)

Sequencer

Illumina

Select the sequencing technology used to produce the sequences.

Input type

TAR Archive

Samples files can be provided in a single TAR archive or sample by sample (with one or two files each).

TAR archive file

1: chaillou_withprimers_64renamedsamples_V1V3_10000seq_R1R2.tar.gz

The TAR file containing the sequences file(s) for each sample.

Are reads already merged ?

No

Yes = The archive contains 1 file by sample : R1 and R2 pairs are already merged in one sequence.

Reads 1 size

300

The maximum read1 size.

Reads 2 size

300

The maximum read2 size.

Mismatch rate

0.1

The maximum rate of mismatches in the overlap region (--mismatch-rate)

Merge software

Vsearch

Vsearch is recommended (in command line, prefer pear)

Select the software to merge paired-end reads (--merge-software)

Would you like to keep unmerged reads?

- (x) No, unmerged reads will be excluded.
- ( ) Yes, unmerged reads will be artificially combined.

No = Unmerged reads will be excluded; Yes = unmerged reads will be artificially combined with 100 N. (default No) (--keep-unmerged)

Minimum amplicon size

400

The minimum size of the amplicons (with primers) (--min-amplicon-size)

Maximum amplicon size

580

The maximum size of the amplicons (with primers) (--max-amplicon-size)

Do the sequences have PCR primers?

- (●) Yes
- ( ) No

5' primer

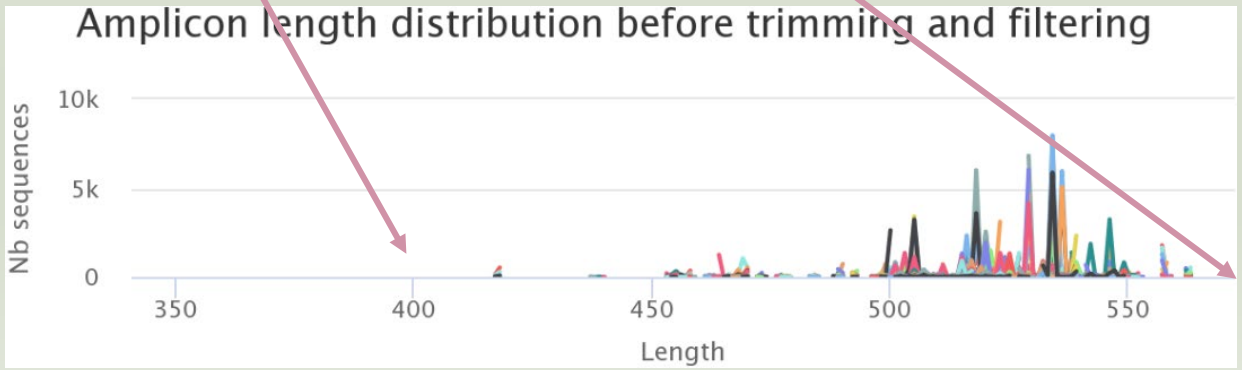
AGAGTTTGATCCTGGCTCAG

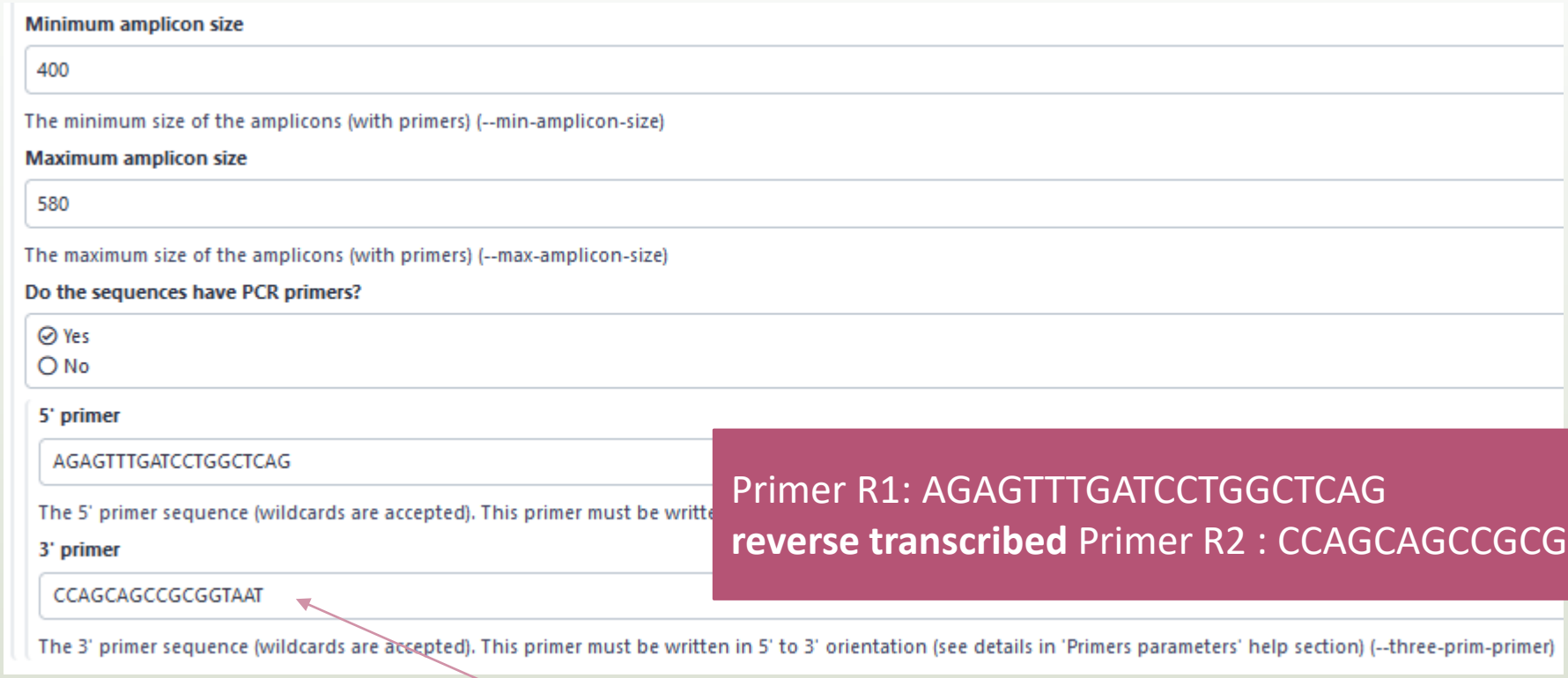
The 5' primer sequence (wildcards are accepted). This primer must be written in 5' to 3' orientation (see details in 'Primers parameters' help section) (--five-prim-primer)

3' primer

CCAGCAGCCGCGGTAAT

The 3' primer sequence (wildcards are accepted). This primer must be written in 5' to 3' orientation (see details in 'Primers parameters' help section) (--three-prim-primer)





Primer R1: AGAGTTTGATCCTGGCTCAG
reverse transcribed Primer R2 : CCAGCAGCCGCGGTAAT

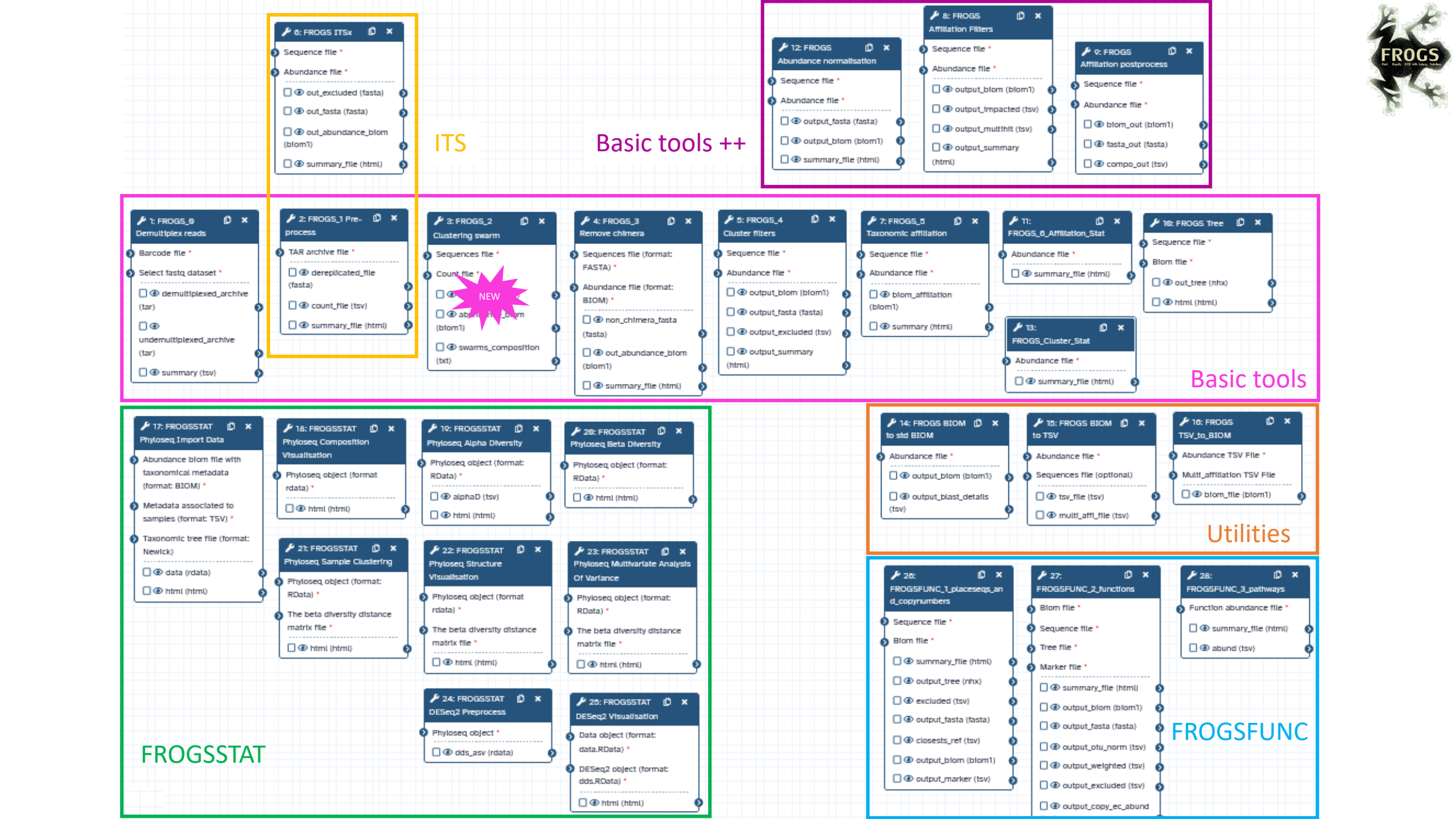
Ex: read R1

@63_0 reference=ASV_00517 position=1..300
AGAGTTTGATCCTGGCTCAGgatgaacgctagcgggaggcttaacacatgcaagccgagggg
tagaattagcttgctaatttgagaccggcgcacgggtgcgtaacgcgtatgcaacttgccctactgaaaa
ggatagcccagagaaatttggattaatactttataatagactgaatggcatcatttagttttgaaagattt
atcgcagtaggataggcatgcgtaagattagatagttggtgaggtaacggctcaccaagtcgacgatct
ttaggggcctgagagggtgaaccccca

Ex: read R2

@63_0 reference=ASV_00517 position=1..300 errors=5%G
ATTACCGCGGCTGCTGGcacggagttagccggtgcttattcttctggtaccttcagctacttacac
gtaagtaggtttatccccagataaaagtagtttacaacccataaggccgtcatcctacacgcgggatggc
tggatcaggcttccacccattgtccaatattcctcactgctgcctcccgtaggagtctggtccgtgtctcag
taccagtgtgggggttcaccctctcaggccccctaaagatcgtcgacttggtgagccgttacctcaccaa
ctatctaatcttacgcatgcct

R2 primer must be reverse transcribed
Use: https://www.bioinformatics.nl/cgi-bin/emboss/revseq


6: FROGS ITSx
Sequence file *
Abundance file *
out_excluded (fasta)
out_fasta (fasta)
out_abundance_biom (biom1)
summary_file (html)
ITS
Basic tools ++
12: FROGS Abundance normalisation
Sequence file *
Abundance file *
output_fasta (fasta)
output_biom (biom1)
summary_file (html)
8: FROGS Affiliation Filters
Sequence file *
Abundance file *
output_biom (biom1)
output_impacted (tsv)
output_multihit (tsv)
output_summary (html)
9: FROGS Affiliation postprocess
Sequence file *
Abundance file *
biom_out (biom1)
fasta_out (fasta)
compo_out (tsv)
1: FROGS_0 Demultiplex reads
Barcode file *
Select fastq dataset *
demultiplexed_archive (tar)
undemultiplexed_archive (tar)
summary (tsv)
2: FROGS_1 Pre-process
TAR archive file *
dereplicated_file (fasta)
count_file (tsv)
summary_file (html)
3: FROGS_2 Clustering swarm
Sequences file *
Count file *
NEW
(biom1)
swarms_composition (txt)
4: FROGS_3 Remove chimera
Sequences file (format: FASTA) *
Abundance file (format: BIOM) *
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
summary_file (html)
5: FROGS_4 Cluster filters
Sequence file *
Abundance file *
output_biom (biom1)
output_fasta (fasta)
output_excluded (tsv)
output_summary (html)
7: FROGS_5 Taxonomic affiliation
Sequence file *
Abundance file *
biom_affiliation (biom1)
summary (html)
11: FROGS_6_Affiliation_Stat
Abundance file *
summary_file (html)
13: FROGS_Cluster_Stat
Abundance file *
summary_file (html)
10: FROGS Tree
Sequence file *
Biom file *
out_tree (nhx)
html (html)
Basic tools
17: FROGSSTAT Phyloseq Import Data
Abundance biom file with taxonomical metadata (format: BIOM) *
Metadata associated to samples (format: TSV) *
Taxonomic tree file (format: Newick)
data (rdata)
html (html)
18: FROGSSTAT Phyloseq Composition Visualisation
Phyloseq object (format rdata) *
html (html)
19: FROGSSTAT Phyloseq Alpha Diversity
Phyloseq object (format: RData) *
alphaD (tsv)
html (html)
20: FROGSSTAT Phyloseq Beta Diversity
Phyloseq object (format: RData) *
html (html)
21: FROGSSTAT Phyloseq Sample Clustering
Phyloseq object (format: RData) *
The beta diversity distance matrix file *
html (html)
22: FROGSSTAT Phyloseq Structure Visualisation
Phyloseq object (format rdata) *
The beta diversity distance matrix file *
html (html)
23: FROGSSTAT Phyloseq Multivariate Analysis Of Variance
Phyloseq object (format: RData) *
The beta diversity distance matrix file *
html (html)
24: FROGSSTAT DESeq2 Preprocess
Phyloseq object *
dds_asv (rdata)
25: FROGSSTAT DESeq2 Visualisation
Data object (format: data.RData) *
DESeq2 object (format: dds.RData) *
html (html)
FROGSSTAT
14: FROGS BIOM to std BIOM
Abundance file *
output_biom (biom1)
output_blast_details (tsv)
15: FROGS BIOM to TSV
Abundance file *
Sequences file (optional)
tsv_file (tsv)
multi_affi_file (tsv)
16: FROGS TSV_to_BIOM
Abundance TSV File *
Multi_affiliation TSV File
biom_file (biom1)
Utilities
26: FROGSFUNC_1_placeseqs_and_copynumbers
Sequence file *
Biom file *
summary_file (html)
output_tree (nhx)
excluded (tsv)
output_fasta (fasta)
closests_ref (tsv)
output_biom (biom1)
output_marker (tsv)
27: FROGSFUNC_2_functions
Biom file *
Sequence file *
Tree file *
Marker file *
summary_file (html)
output_biom (biom1)
output_fasta (fasta)
output_otu_norm (tsv)
output_weighted (tsv)
output_excluded (tsv)
output_copy_ec_abund
28: FROGSFUNC_3_pathways
Function abundance file *
summary_file (html)
abund (tsv)
FROGSFUNC

Clustering tool

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

Sequences file

2: FROGS Pre-process: dereplicated.fasta

The dereplicated sequences file (format: fasta).

Count file

3: FROGS Pre-process: count.tsv

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

FROGS guidelines version

New guidelines from version 3.2

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

Aggregation distance clustering

1

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

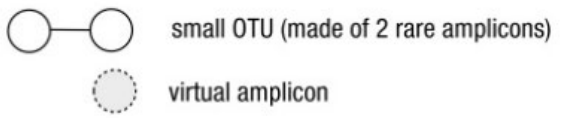
Refine OTU clustering

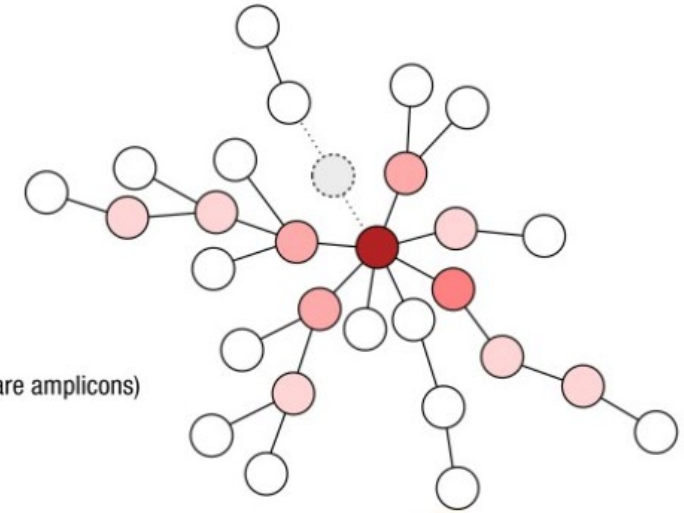
Yes | No

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)

✔ Execute

longer but more accurate





Practice:

Exercise

Go to « 16S » history

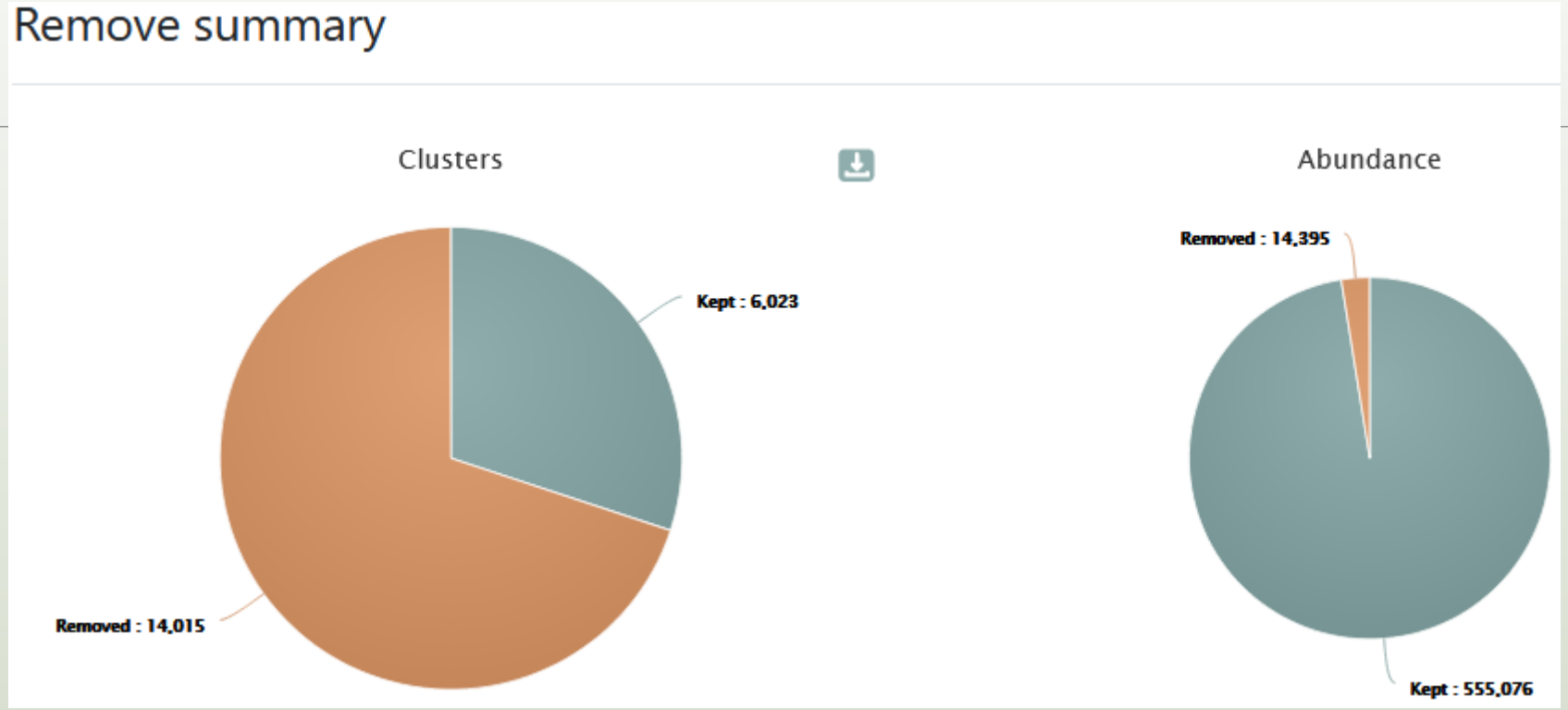
- Launch the FROGS_2 clustering swarm
- Launch the FROGS_3 remove chimera
- Launch the FROGS_Cluster_Stat

Exercise

1. Understand the « FROGS remove chimera : report.html»
   a. How many clusters are kept after chimera removal?
   b. How many sequences that represent ? So what abundance?
   c. What do you conclude ?
2. What is the size of the largest removed cluster of chimeras?
3. Compare the HTML files
   a. Of what are mainly composed singleton ?
   b. What are their abundance?
   c. What do you conclude ?

Answer 1

Q1a: How many clusters are kept after chimera removal?
Q1b: How many sequences that represent ? So what abundance?
Q1c: What do you conclude ?



6023 clusters are kept.
The 14015 removed clusters represent ~2.5 % of sequences

Here, chimera clusters represent many clusters ~70% but very few sequences.

Removed clusters are low abundance clusters.

Answer 2

Q2: What is the size of the largest removed cluster of chimeras?

Sample	Clusters kept	% Clusters kept	Cluster abundance kept	% Cluster abundance kept	Chimeric clusters removed	Chimeric abundance removed	Abundance of the most abundant chimera removed	Individual chimera detected	Individual chimera abundance detected	Abundance of the most abundant individual chimera detected
VHT0.LOT02	205	35.90	8,862			410	19	372	446	19
MVT0.LOT10	254	60.48	9,313			180	10	169	304	92
VHT0.LOT08	261	45.87	8,852			332	10	310	344	11
VHT0.LOT01	198	35.42	8,832	95.90	361	378	8	365	382	8

The largest cluster of chimeras contained 19 sequences.

92 chimeras are detected but only 10 are removed because 82 have been invalidated by the cross validation

Answer 2

Q2: What is the size of the largest removed cluster of chimeras?

Sample	Clusters kept	% Clusters kept	Cluster abundance kept	% Cluster abundance kept	Chimeric clusters removed	Chimeric abundance removed	Abundance of the most abundant chimera removed	Individual chimera detected	Individual chimera abundance detected	Abundance of the most abundant individual chimera detected
VHT0.LOT02	205	35.90	8,862			410	19	372	446	19
MVT0.LOT10	254	60.48	9,313			180	10	169	304	92
VHT0.LOT08	261	45.87	8,852			332	10	310	344	11
VHT0.LOT01	198	35.42	8,832	95.90	361	378	8	365	382	8

The largest cluster of chimeras contained 19 sequences.

92 chimeras are detected but only 10 are removed because 82 have been invalidated by the cross validation

Answer 3

Q3a: Of what are mainly composed singleton ? (compare with previous report.html)
Q3b: What are their abundance?
Q3c: What do you conclude ?

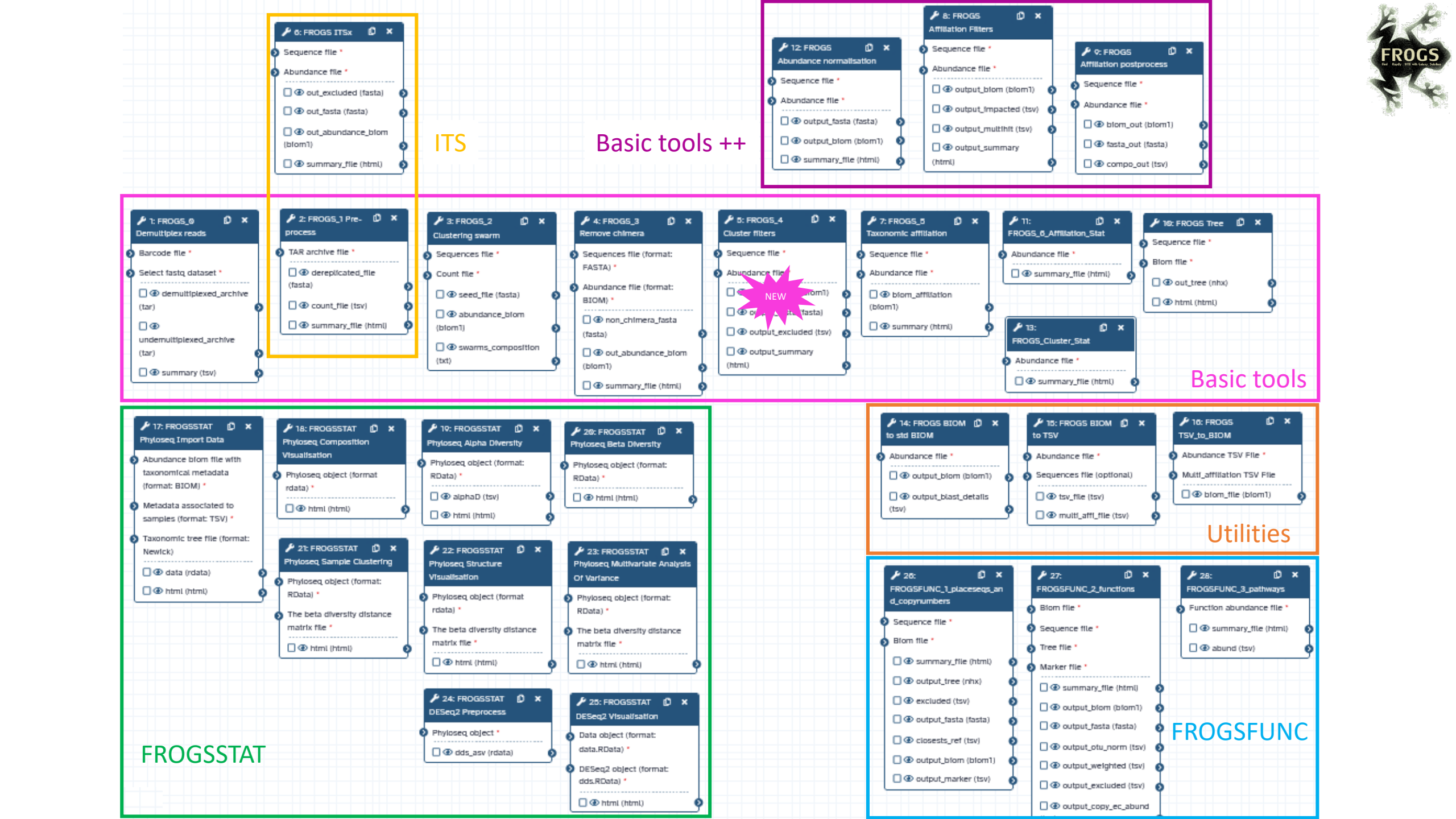
Cluster size	Number of cluster	% of all clusters
1	19,267	96.15
2	150	0.75
3	22	0.11
4	10	0.05

Cluster_Stat report after clustering

Most small clusters are composed of chimeras

Cluster size	Number of cluster	% of all clusters
1	5,387	89.44
2	49	0.81
3	15	0.25
4	7	0.12

Cluster_Stat report after chimera removing


6: FROGS ITSx
Sequence file *
Abundance file *
out_excluded (fasta)
out_fasta (fasta)
out_abundance_biom (biom1)
summary_file (html)
ITS
Basic tools ++
12: FROGS Abundance normalisation
Sequence file *
Abundance file *
output_fasta (fasta)
output_biom (biom1)
summary_file (html)
8: FROGS Affiliation Filters
Sequence file *
Abundance file *
output_biom (biom1)
output_impacted (tsv)
output_multihit (tsv)
output_summary (html)
9: FROGS Affiliation postprocess
Sequence file *
Abundance file *
biom_out (biom1)
fasta_out (fasta)
compo_out (tsv)
1: FROGS_0 Demultiplex reads
Barcode file *
Select fastq dataset *
demultiplexed_archive (tar)
undemultiplexed_archive (tar)
summary (tsv)
2: FROGS_1 Pre-process
TAR archive file *
dereplicated_file (fasta)
count_file (tsv)
summary_file (html)
3: FROGS_2 Clustering swarm
Sequences file *
Count file *
seed_file (fasta)
abundance_biom (biom1)
swarms_composition (txt)
4: FROGS_3 Remove chimera
Sequences file (format: FASTA) *
Abundance file (format: BIOM) *
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
summary_file (html)
5: FROGS_4 Cluster filters
Sequence file *
Abundance file *
output_excluded (tsv)
output_summary (html)
NEW
7: FROGS_5 Taxonomic affiliation
Sequence file *
Abundance file *
biom_affiliation (biom1)
summary (html)
11: FROGS_6_Affiliation_Stat
Abundance file *
summary_file (html)
13: FROGS_Cluster_Stat
Abundance file *
summary_file (html)
10: FROGS Tree
Sequence file *
Biom file *
out_tree (nhx)
html (html)
Basic tools
17: FROGSSTAT Phyloseq Import Data
Abundance biom file with taxonomical metadata (format: BIOM) *
Metadata associated to samples (format: TSV) *
Taxonomic tree file (format: Newick)
data (rdata)
html (html)
18: FROGSSTAT Phyloseq Composition Visualisation
Phyloseq object (format rdata) *
html (html)
19: FROGSSTAT Phyloseq Alpha Diversity
Phyloseq object (format: RData) *
alphaD (tsv)
html (html)
20: FROGSSTAT Phyloseq Beta Diversity
Phyloseq object (format: RData) *
html (html)
21: FROGSSTAT Phyloseq Sample Clustering
Phyloseq object (format: RData) *
The beta diversity distance matrix file *
html (html)
22: FROGSSTAT Phyloseq Structure Visualisation
Phyloseq object (format rdata) *
The beta diversity distance matrix file *
html (html)
23: FROGSSTAT Phyloseq Multivariate Analysis Of Variance
Phyloseq object (format: RData) *
The beta diversity distance matrix file *
html (html)
24: FROGSSTAT DESeq2 Preprocess
Phyloseq object *
dds_asv (rdata)
25: FROGSSTAT DESeq2 Visualisation
Data object (format: data.RData) *
DESeq2 object (format: dds.RData) *
html (html)
FROGSSTAT
14: FROGS BIOM to std BIOM
Abundance file *
output_biom (biom1)
output_blast_details (tsv)
15: FROGS BIOM to TSV
Abundance file *
Sequences file (optional)
tsv_file (tsv)
multi_affi_file (tsv)
16: FROGS TSV_to_BIOM
Abundance TSV File *
Multi_affiliation TSV File
biom_file (biom1)
Utilities
26: FROGSFUNC_1_placeseqs_and_copynumbers
Sequence file *
Biom file *
summary_file (html)
output_tree (nhx)
excluded (tsv)
output_fasta (fasta)
closests_ref (tsv)
output_biom (biom1)
output_marker (tsv)
27: FROGSFUNC_2_functions
Biom file *
Sequence file *
Tree file *
Marker file *
summary_file (html)
output_biom (biom1)
output_fasta (fasta)
output_otu_norm (tsv)
output_weighted (tsv)
output_excluded (tsv)
output_copy_ec_abund
28: FROGSFUNC_3_pathways
Function abundance file *
summary_file (html)
abund (tsv)
FROGSFUNC

Cluster Filter tool

4- Cluster Filter

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

Criteria:

- The cluster prevalence: The number of times the cluster is present in the environment, *i.e.* the number of samples where the cluster must be present.
- Cluster size: An cluster that is not large enough for a given proportion or count will be removed.
- Biggest Cluster : Only the X biggest are conserved.
- Contaminant: If cluster sequence matches with phiX, chloroplastic/mitochondrial 16S of A. Thaliana or your own contaminant sequence.

One tool, 4 criteria

FROGS_4 Cluster filters Filters clusters on several criteria. (Galaxy Version 4.1.0+galaxy1)

☆ Favorite ▾ Options

Sequence file

15: FROGS_4 Cluster filters: clusterFilters_sequences.fasta

The sequence file to filter (format: FASTA)

Abundance file

14: FROGS_4 Cluster filters: clusterFilters_abundance.biom

The abundance file to filter (format: BIOM)

1

Minimum prevalence method

all samples

Minimum prevalence

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

2

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

as proportion

Minimum proportion of sequences abundancy to keep cluster

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

3

N biggest clusters

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

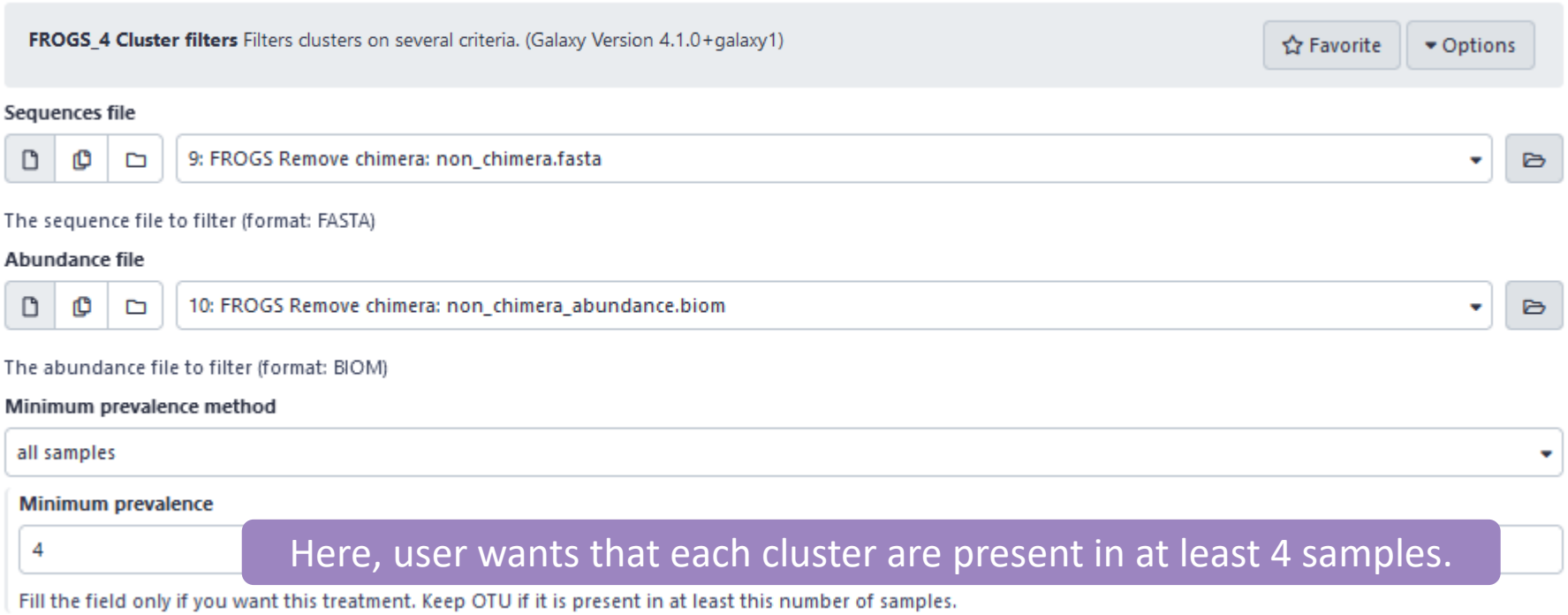
4

Search for contaminant clusters.

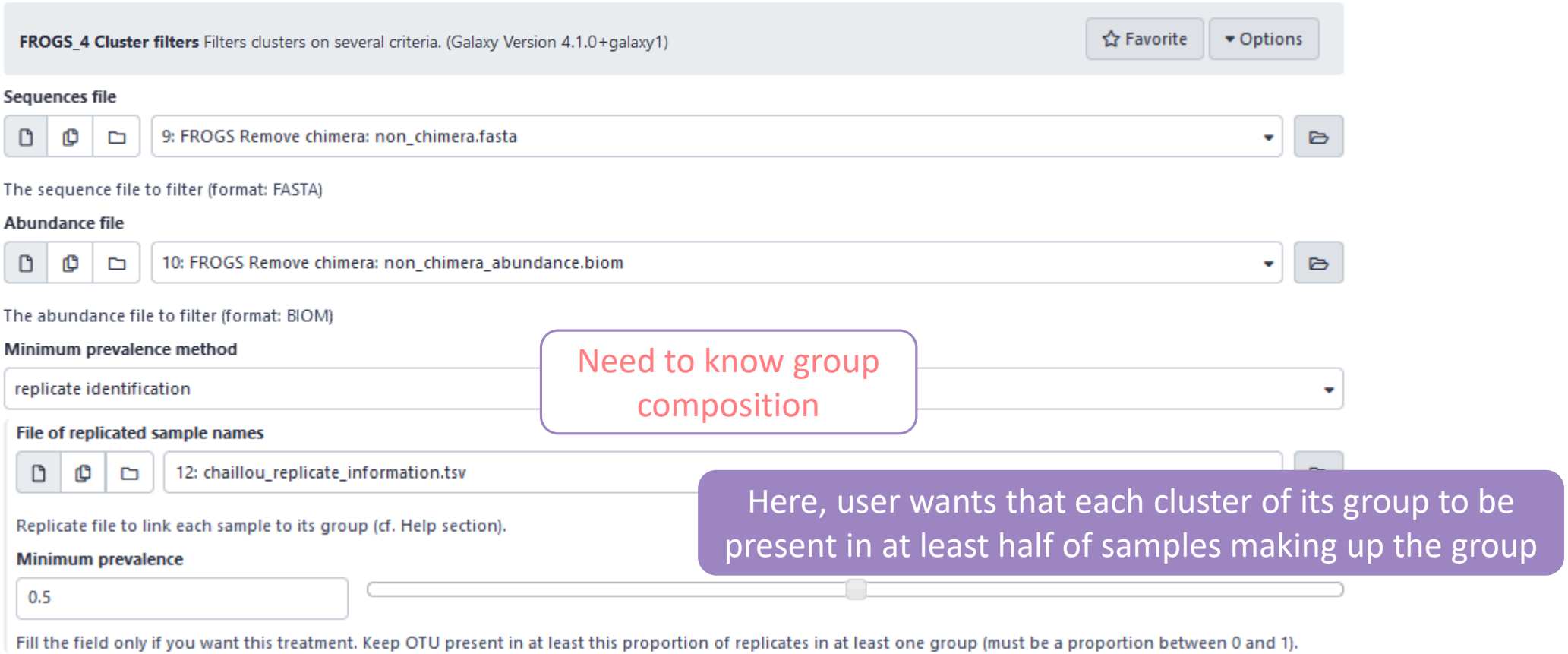
No contaminant filter

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

1 Prevalence filter – option 1



1 Prevalence filter – option 2



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 clusters 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 clusters 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 clusters 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

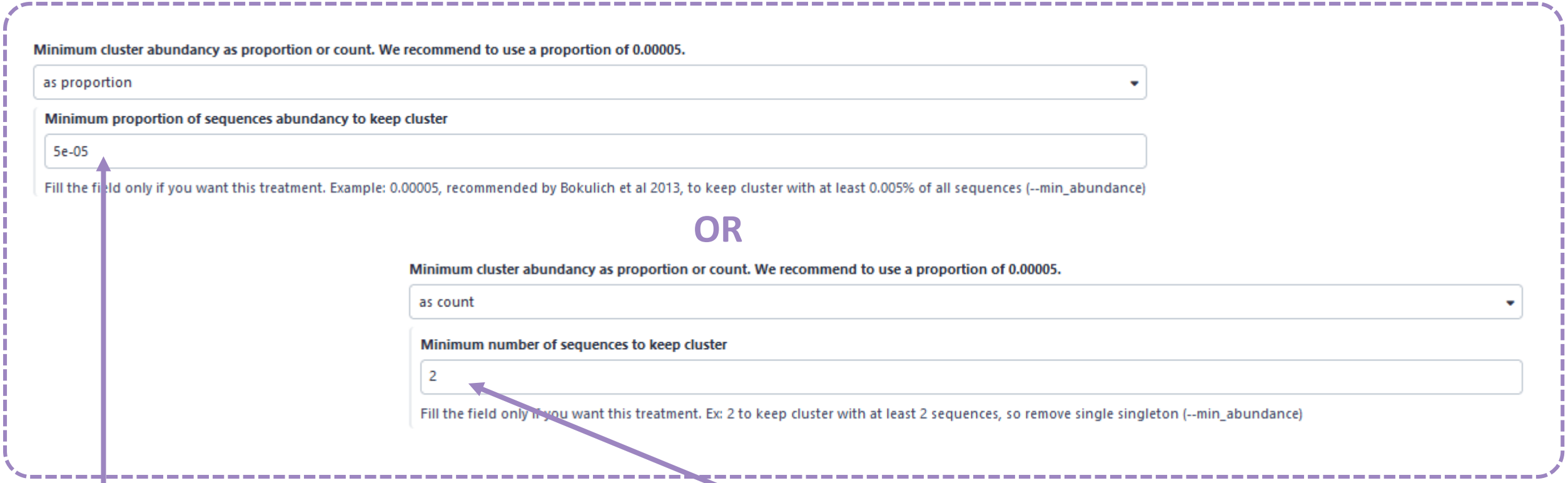
```
sample  rich
sample rich
sample 3  rich
sample4 richAB
sample5 richAB
sample6 richAB
sample7 richAB
sample8 low
sample9 lowAB
sample10  lowAB
sample11  lowAB
sample12  april21
sample13  april21
```

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 Cluster size filter



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

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

3

Filter : Keep biggest cluster

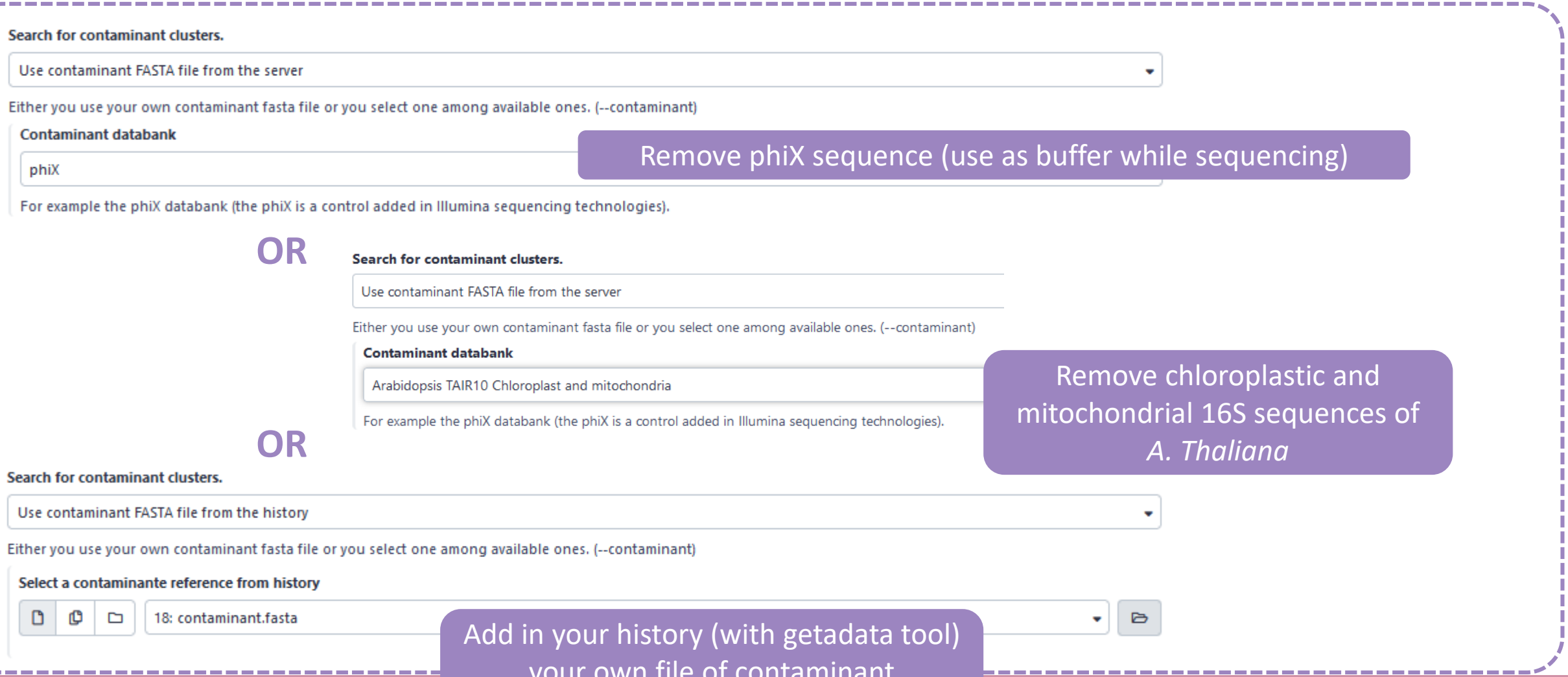
N biggest clusters

50

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

Here, user wants to keep the 50 biggest clusters.

4 Contaminant filter



Practice:

LAUNCH THE CLUSTER FILTER TOOL

Exercice:

Go to history « 16S » history

Launch « cluster Filter » tool with non_chimera_abundance.biom, non_chimera.fasta

Use 3 criteria to filter clusters:

- cluster must be present at least in 4 samples
- Each cluster must represented a minimum of 0.005 % = 0.00005 [1] of the totality of the sequences
- cluster of phiX [2] must be removed

→ objective : play with filters, understand their impacts on falses-positives clusters

[1] *Nat Methods. 2013 Jan;10(1):57-9. doi: 10.1038/nmeth.2276. Epub 2012 Dec 2.*
Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing.
Bokulich NA1, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, Mills DA, Caporaso JG.

[2] https://www.illumina.com/products/by-type/sequencing-kits/cluster-gen-sequencing-reagents/phix-control-v3.html

Exercice:

1. What are the output files of “cluster Filter” ?
2. Explore “FROGS Filter : report.html” file. How many cluster have you removed ? How many cluster do they remain ? Which sample keeps the least cluster and for which reason?
3. Build the Venn diagram on the two filters. How many cluster have you removed with each filter ?
4. How many own cluster remains in BHT0.LOT08 ? To retrieve this information, which tool do you need to launch previously ?

Answer 1

FROGS_4 Cluster filters Filters clusters on several criteria. (Galaxy Version 4.1.0+galaxy1)

☆ Favorite ▾ Options

Sequence file

10: FROGS_3 Remove chimera: non_chimera.fasta

The sequence file to filter (format: FASTA)

Abundance file

11: FROGS_3 Remove chimera: non_chimera_abundance.biom

The abundance file to filter (format: BIOM)

Minimum prevalence method

all samples

Minimum prevalence

4

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

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

as proportion

Minimum proportion of sequences abundancy to keep cluster

0.00005

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

st clusters

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

Search for contaminant clusters.

Use contaminant FASTA file from the server

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

Contaminant databank

phiX

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

0.005% = 0.00005

Outputs

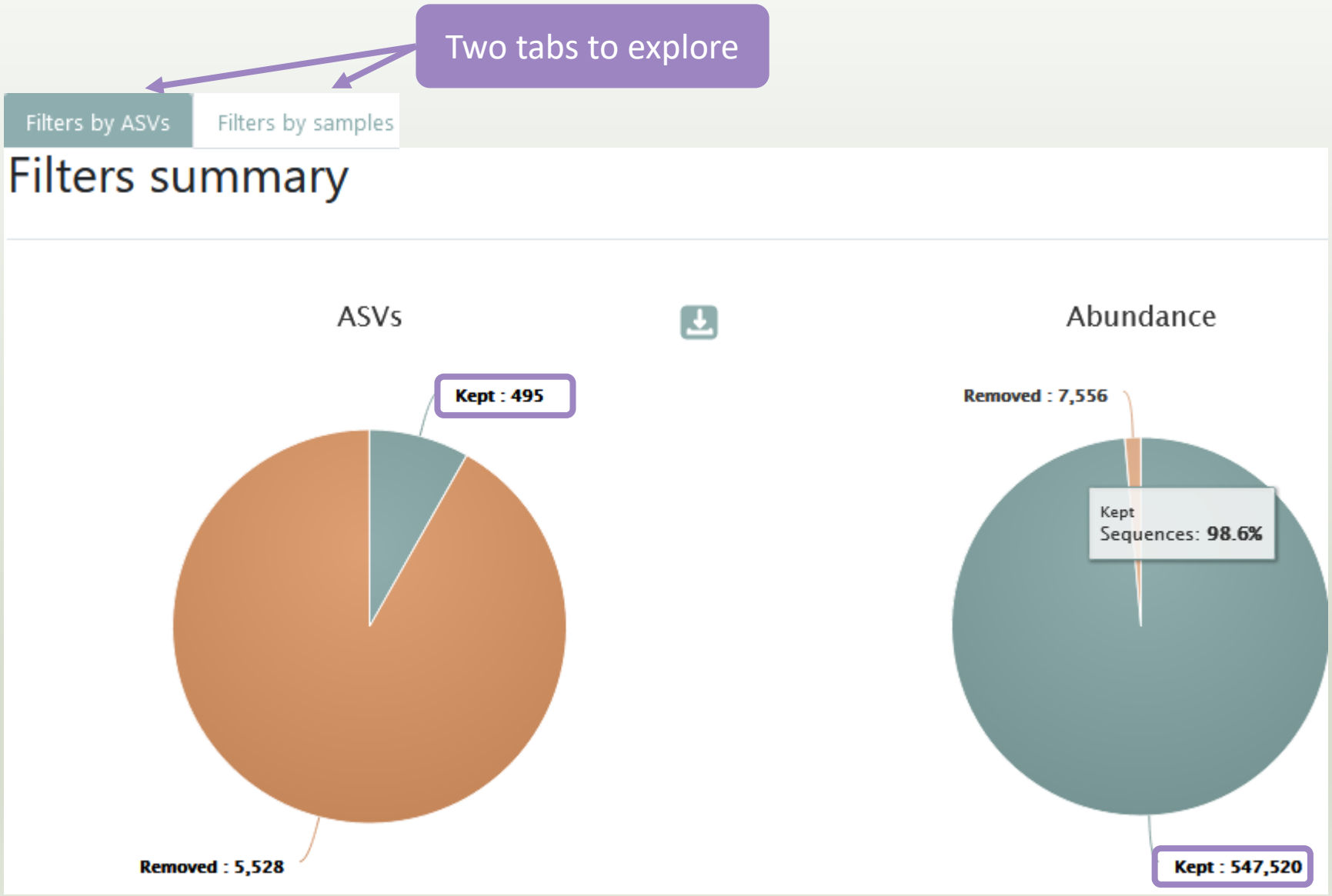
17: FROGS_4 Cluster filters: report.html

16: FROGS_4 Cluster filters: excluded.tsv

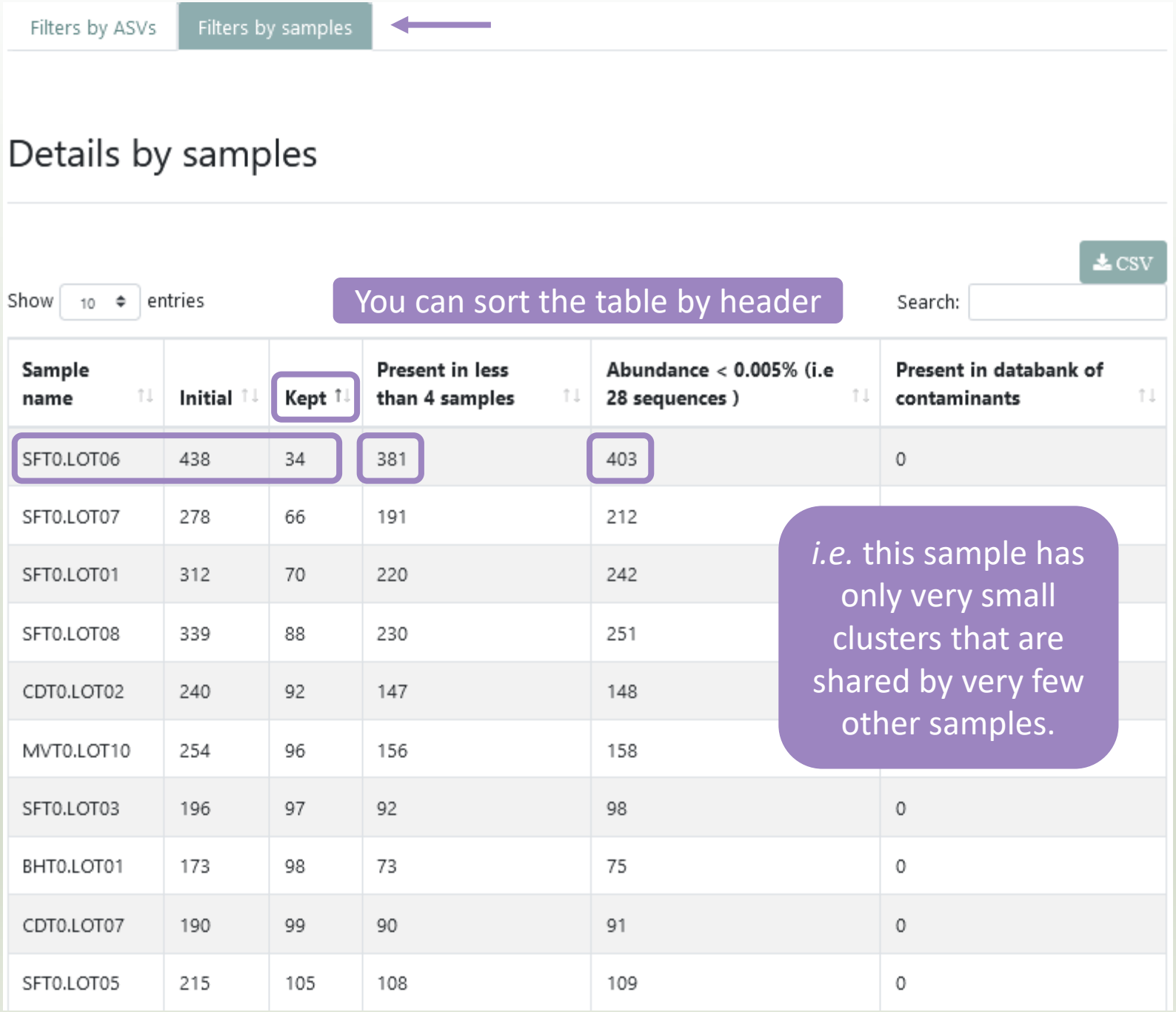
15: FROGS_4 Cluster filters: clusterFilters_sequences.fasta

14: FROGS_4 Cluster filters: clusterFilters_abundance.biom

Answer 2


Two tabs to explore
Filters by ASVs
Filters by samples
Filters summary
ASVs
Kept : 495
Removed : 5,528
Abundance
Removed : 7,556
Kept
Sequences: 98.6%
Kept : 547,520

Answer 2



Sample name	Initial	Kept	Present in less than 4 samples	Abundance < 0.005% (i.e 28 sequences)	Present in databank of contaminants
SFT0.LOT06	438	34	381	403	0
SFT0.LOT07	278	66	191	212	
SFT0.LOT01	312	70	220	242	
SFT0.LOT08	339	88	230	251	
CDT0.LOT02	240	92	147	148	
MVT0.LOT10	254	96	156	158	
SFT0.LOT03	196	97	92	98	0
BHT0.LOT01	173	98	73	75	0
CDT0.LOT07	190	99	90	91	0
SFT0.LOT05	215	105	108	109	0

i.e. this sample has only very small clusters that are shared by very few other samples.

Answer 3

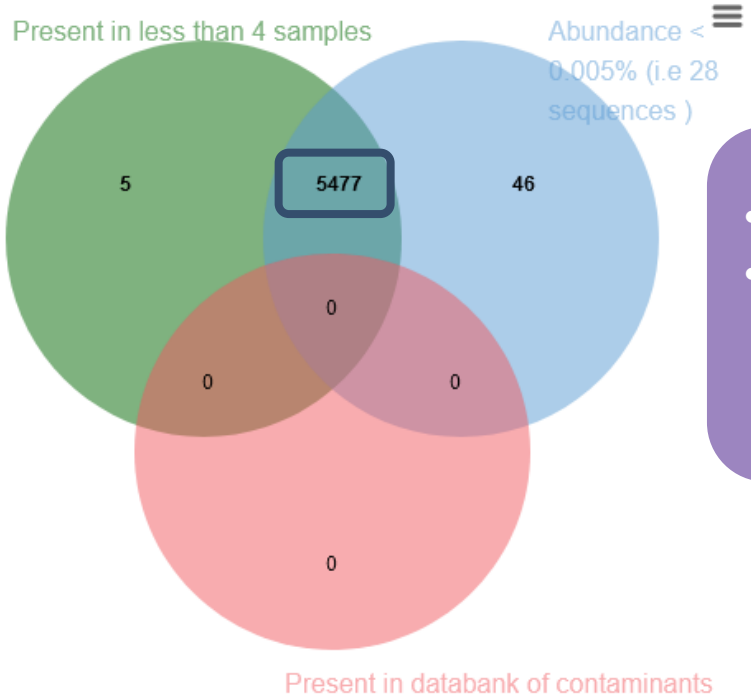
Filters intersections

Draw a Venn to see which ASVs had been deleted by the filters chosen (Maximum 6 options):

- [x] Present in less than 4 samples
- [x] Abundance < 0.005% (i.e 28 sequences)
- [x] Present in databank of contaminants

Venn

Venn on removed ASVs



- No phiX sequence.
- Most clusters are both small and not shared by 4 samples.

Answer 4

report.html of **ClusterStat tool**

Because of the "prevalence = 4" criterion, there is no longer an "own cluster" for any sample.

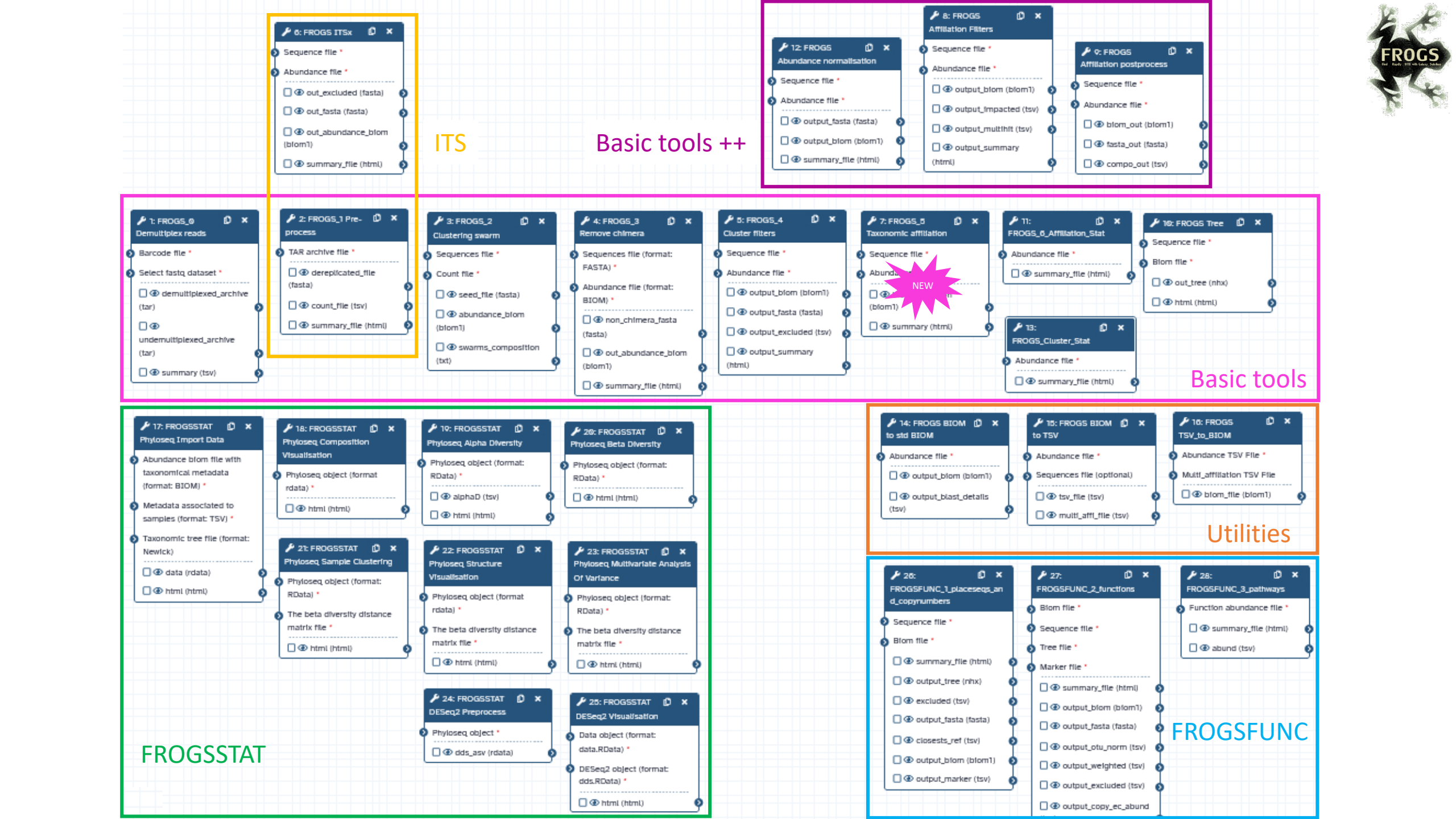
Clusters distribution | Sequences distribution | Samples distribution

Sequences count

CSV

Show 10 entries Search:

Sample	Total clusters	Shared clusters	Own clusters	Total sequences	Shared sequences	Own sequences
BHT0.LOT01	98	98	0	8,690	8,690	0
BHT0.LOT03	135	135	0	8,377	8,377	0
BHT0.LOT04	150	150	0	8,643	8,643	0
BHT0.LOT05	140	140	0	8,544	8,544	0
BHT0.LOT06	145	145	0	8,646	8,646	0
BHT0.LOT07	150	150	0	8,671	8,671	0
BHT0.LOT08	195	195	0	8,479	8,479	0
BHT0.LOT10	165	165	0	8,606	8,606	0
CDT0.LOT02	92	92	0	8,750	8,750	0
CDT0.LOT04	161	161	0	8,605	8,605	0


6: FROGS ITSx
Sequence file *
Abundance file *
out_excluded (fasta)
out_fasta (fasta)
out_abundance_biom (biom1)
summary_file (html)
ITS
Basic tools ++
12: FROGS Abundance normalisation
Sequence file *
Abundance file *
output_fasta (fasta)
output_biom (biom1)
summary_file (html)
8: FROGS Affiliation filters
Sequence file *
Abundance file *
output_biom (biom1)
output_impacted (tsv)
output_multihit (tsv)
output_summary (html)
9: FROGS Affiliation postprocess
Sequence file *
Abundance file *
biom_out (biom1)
fasta_out (fasta)
compo_out (tsv)
1: FROGS_0 Demultiplex reads
Barcode file *
Select fastq dataset *
demultiplexed_archive (tar)
undemultiplexed_archive (tar)
summary (tsv)
2: FROGS_1 Pre-process
TAR archive file *
dereplicated_file (fasta)
count_file (tsv)
summary_file (html)
3: FROGS_2 Clustering swarm
Sequences file *
Count file *
seed_file (fasta)
abundance_biom (biom1)
swarms_composition (txt)
4: FROGS_3 Remove chimera
Sequences file (format: FASTA) *
Abundance file (format: BIOM) *
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
summary_file (html)
5: FROGS_4 Cluster filters
Sequence file *
Abundance file *
output_biom (biom1)
output_fasta (fasta)
output_excluded (tsv)
output_summary (html)
7: FROGS_5 Taxonomic affiliation
Sequence file *
NEW
summary (html)
11: FROGS_6_Affiliation_Stat
Abundance file *
summary_file (html)
10: FROGS Tree
Sequence file *
Biom file *
out_tree (nhx)
html (html)
13: FROGS_Cluster_Stat
Abundance file *
summary_file (html)
Basic tools
17: FROGSSTAT Phyloseq Import Data
Abundance biom file with taxonomical metadata (format: BIOM) *
Metadata associated to samples (format: TSV) *
Taxonomic tree file (format: Newick)
data (rdata)
html (html)
18: FROGSSTAT Phyloseq Composition Visualisation
Phyloseq object (format rdata) *
html (html)
19: FROGSSTAT Phyloseq Alpha Diversity
Phyloseq object (format: RData) *
alphaD (tsv)
html (html)
20: FROGSSTAT Phyloseq Beta Diversity
Phyloseq object (format: RData) *
html (html)
21: FROGSSTAT Phyloseq Sample Clustering
Phyloseq object (format: RData) *
The beta diversity distance matrix file *
html (html)
22: FROGSSTAT Phyloseq Structure Visualisation
Phyloseq object (format rdata) *
The beta diversity distance matrix file *
html (html)
23: FROGSSTAT Phyloseq Multivariate Analysis Of Variance
Phyloseq object (format: RData) *
The beta diversity distance matrix file *
html (html)
24: FROGSSTAT DESeq2 Preprocess
Phyloseq object *
dds_asv (rdata)
25: FROGSSTAT DESeq2 Visualisation
Data object (format: data.RData) *
DESeq2 object (format: dds.RData) *
html (html)
FROGSSTAT
14: FROGS BIOM to std BIOM
Abundance file *
output_biom (biom1)
output_blast_details (tsv)
15: FROGS BIOM to TSV
Abundance file *
Sequences file (optional)
tsv_file (tsv)
multi_affi_file (tsv)
16: FROGS TSV_to_BIOM
Abundance TSV File *
Multi_affiliation TSV File
biom_file (biom1)
Utilities
26: FROGSFUNC_1_placeseqs_and_copynumbers
Sequence file *
Biom file *
summary_file (html)
output_tree (nhx)
excluded (tsv)
output_fasta (fasta)
closests_ref (tsv)
output_biom (biom1)
output_marker (tsv)
27: FROGSFUNC_2_functions
Biom file *
Sequence file *
Tree file *
Marker file *
summary_file (html)
output_biom (biom1)
output_fasta (fasta)
output_otu_norm (tsv)
output_weighted (tsv)
output_excluded (tsv)
output_copy_ec_abund
28: FROGSFUNC_3_pathways
Function abundance file *
summary_file (html)
abund (tsv)
FROGSFUNC

Affiliation tool

FROGS_5 Taxonomic affiliation Taxonomic affiliation of each ASV's seed by RDPtools and BLAST (Galaxy Version 4.1.0+galaxy1)

☆ Favorite ▾ Options

Using reference database

16S SILVA 138.1

Start to write the beginning of your amplicon name to see all available databases

Select reference from the list

Also perform RDP assignation? Optional

○ Yes
⊘ No

Taxonomy affiliation will be perform thanks to Blast. This option allows to perform it also with RDP classifier tool (default No) (--rdp)

Taxonomic ranks

Domain Phylum Class Order Family Genus Species

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

Sequence file

15: FROGS_4 Cluster filters: clusterFilters_sequences.fasta

The sequences to affiliated (format: FASTA)

Abundance file

14: FROGS_4 Cluster filters: clusterFilters_abundance.biom

For more details on FROGS databanks:
http://genoweb.toulouse.inra.fr/frogs_databanks/assignation/readme.txt

Available databases in FROGS

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

For exemples:

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.

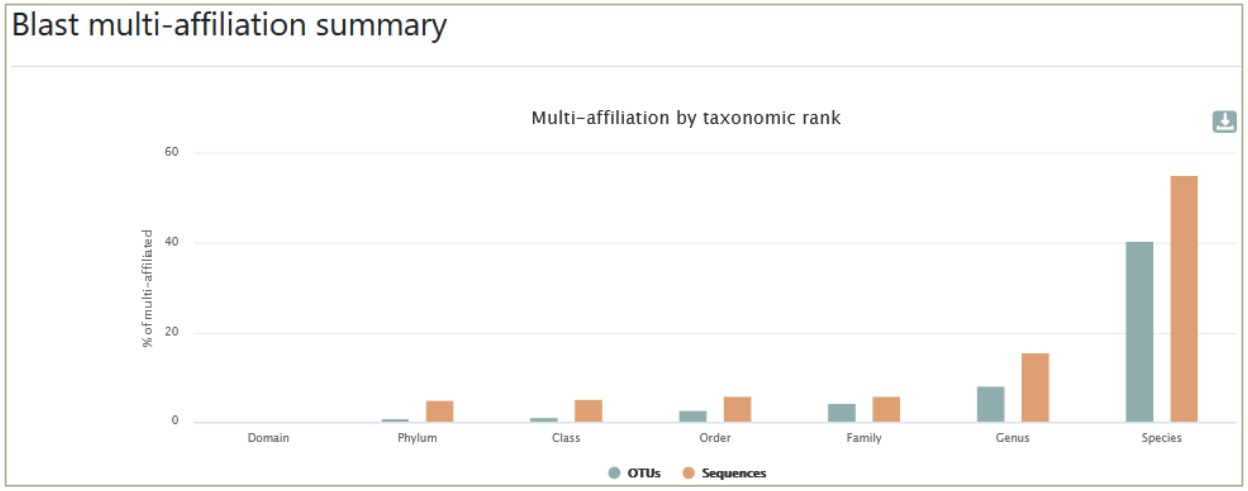
4 ranks of available databases in FROGS: 50 pintail, 80 pintail or 100 pintail or no pintail filter.

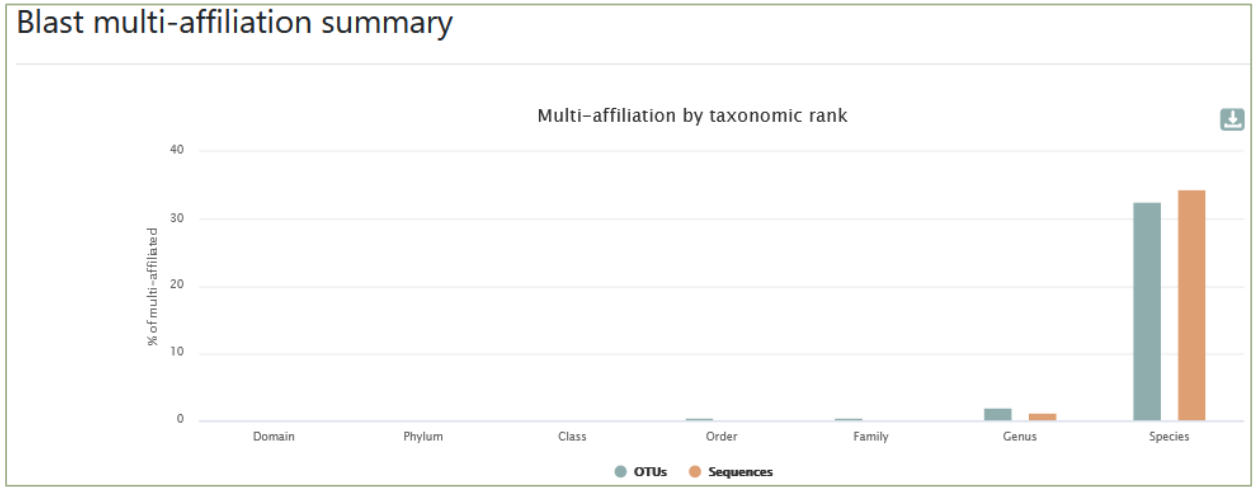
silva138.1 16S
silva138.1 pintail100 16S
silva138.1 pintail80 16S
silva138.1 pintail50 16S
silva138.1 18S
silva138.1 23S
silva138.1 28S

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 P.acn17
Cluster_4	Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes TypeIA2 P.acn31
Cluster_4	Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes TypeIA2 P.acn33

Exemple between silva 138.1 and silva 138.1 pintail 100

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

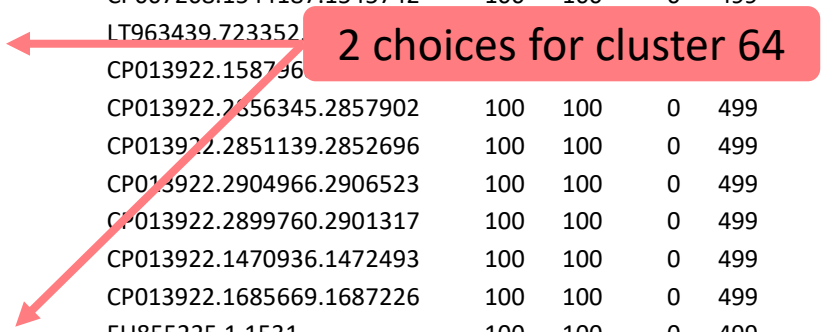
? Cluster_4 Bacteria;Actinobacteriota;Actinobacteria;Corynebacteriales;Corynebacteriaceae;Corynebacterium;unknown species
? Cluster_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Aureobasidium melanogenum
Cluster_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium 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;Actinobacte ctinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes SK137
Cluster_4 Bacteria;Actinobacte ctinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;unknown species
Cluster_4 Bacteria;Actinobacte terium;Cutibacterium acnes TypeIA2 P.acn17
Cluster_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes TypeIA2 P.acn31
Cluster_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes TypeIA2 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
KF100699	*uncultured bacterium*	1341				Bacteria▸Firmicutes▸Bacilli…

How to 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	CP013922.2856345.2857902	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



How to 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

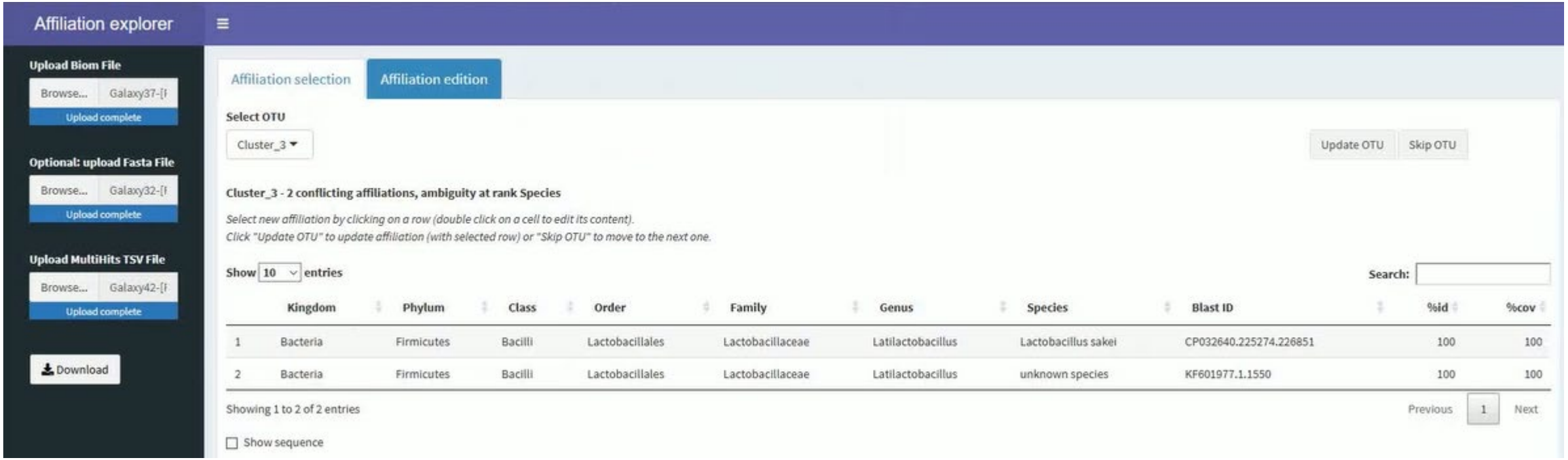
and

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

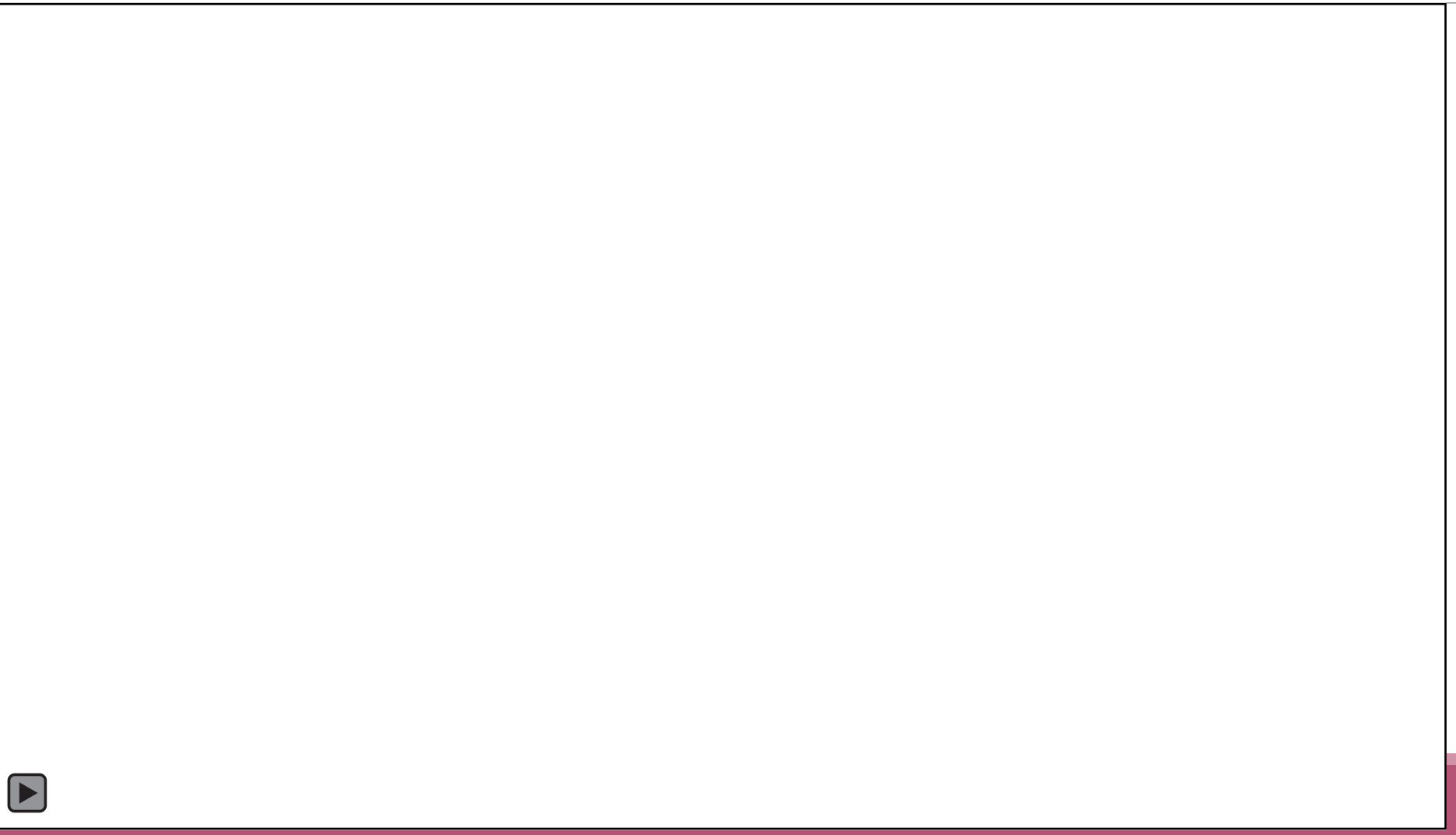


A very user-friendly tool, developed by Mahendra Mariadassou and his collaborators (Maiage unit - INRAE Jouy-en-Josas). It allows to modify very simply the affiliations of an abundance table from FROGS.

Affiliation explorer

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

Demo video



Practice:

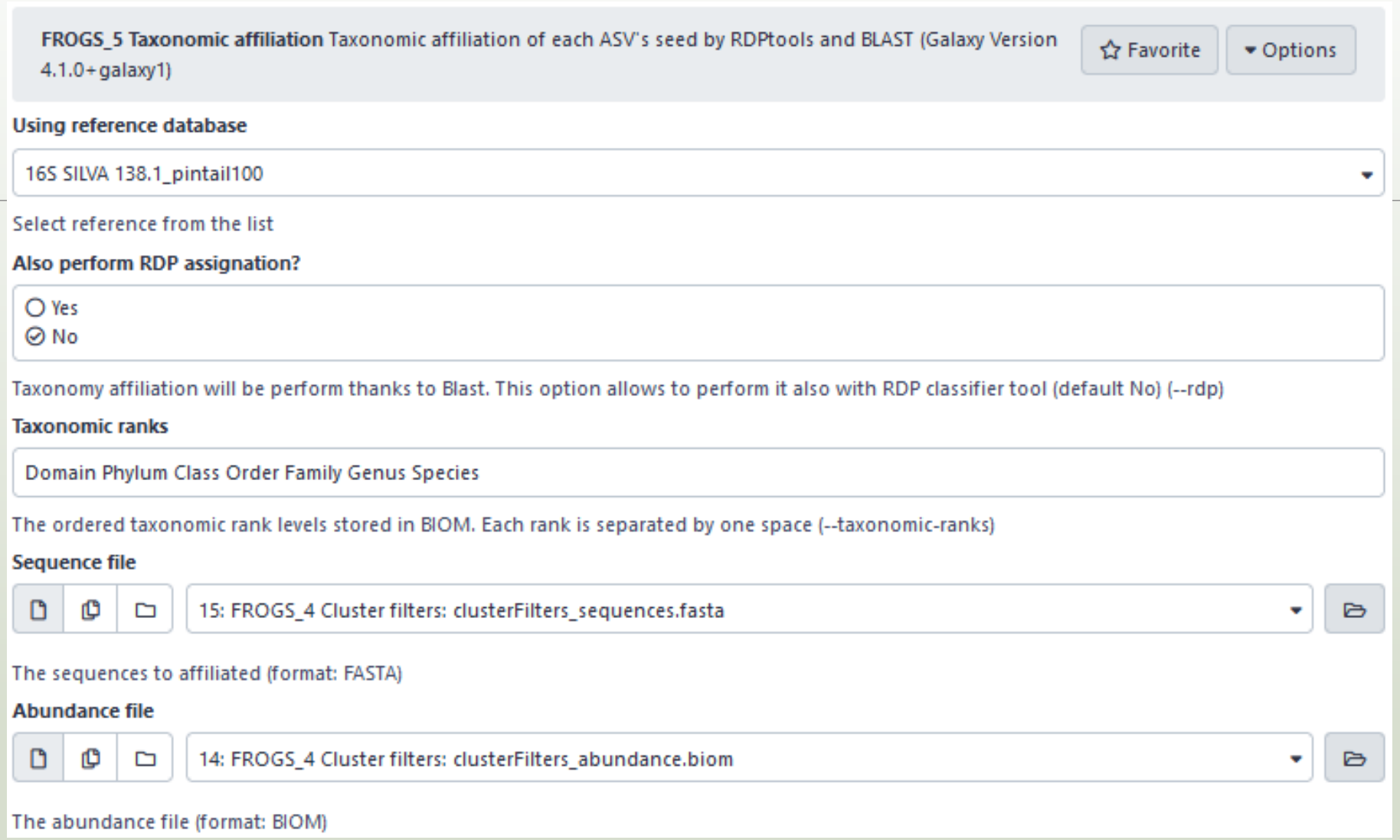
LAUNCH THE FROGS_5 TAXONOMIC AFFILIATION TOOL

Exercice:

Go to history « 16S » history

Launch the « FROGS_5 taxonomic affiliation » tool with

- SILVA 138.1 16S database **pintail 100**


FROGS_5 Taxonomic affiliation Taxonomic affiliation of each ASV's seed by RDPtools and BLAST (Galaxy Version 4.1.0+galaxy1)
Favorite
Options
Using reference database
16S SILVA 138.1_pintail100
Select reference from the list
Also perform RDP assignation?
Yes
No
Taxonomy affiliation will be perform thanks to Blast. This option allows to perform it also with RDP classifier tool (default No) (--rdp)
Taxonomic ranks
Domain Phylum Class Order Family Genus Species
The ordered taxonomic rank levels stored in BIOM. Each rank is separated by one space (--taxonomic-ranks)
Sequence file
15: FROGS_4 Cluster filters: clusterFilters_sequences.fasta
The sequences to affiliated (format: FASTA)
Abundance file
14: FROGS_4 Cluster filters: clusterFilters_abundance.biom
The abundance file (format: BIOM)

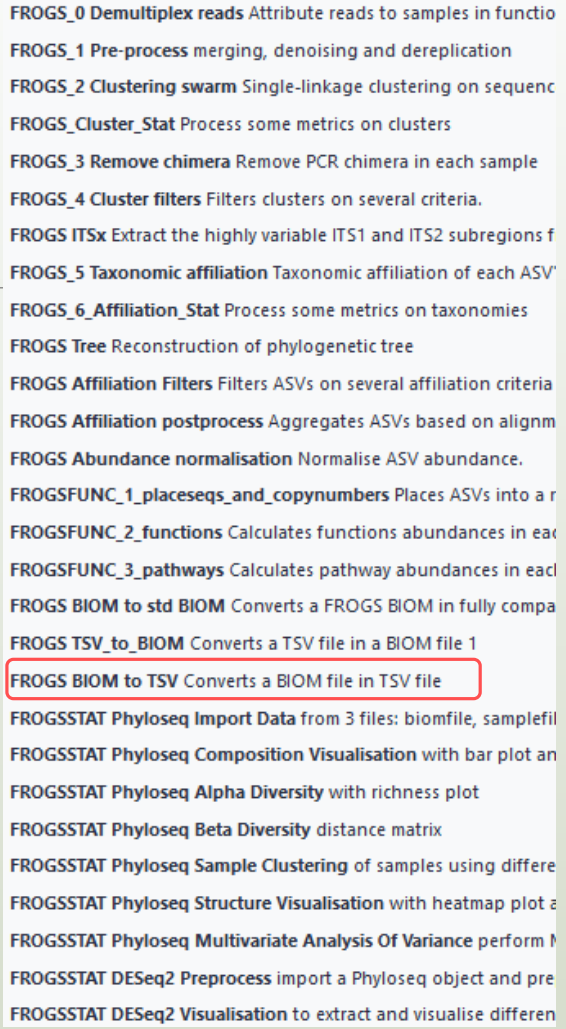
Exercise

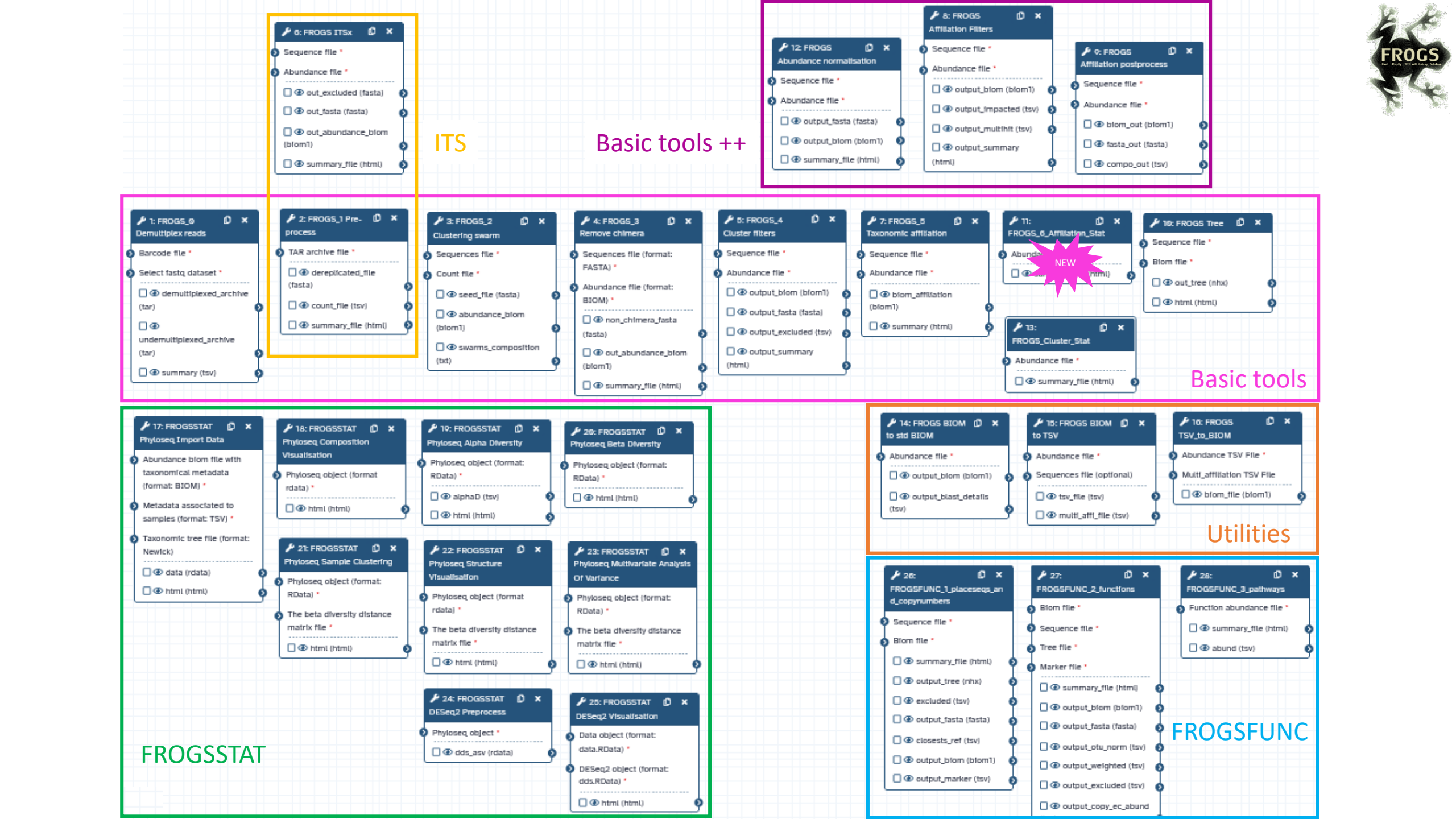
Use the **Biom_to_TSV tool** on this last file and click again on the ”eye” on the new output generated.

Transform the biom file in TSV file (easy to manipulate on excel or R)

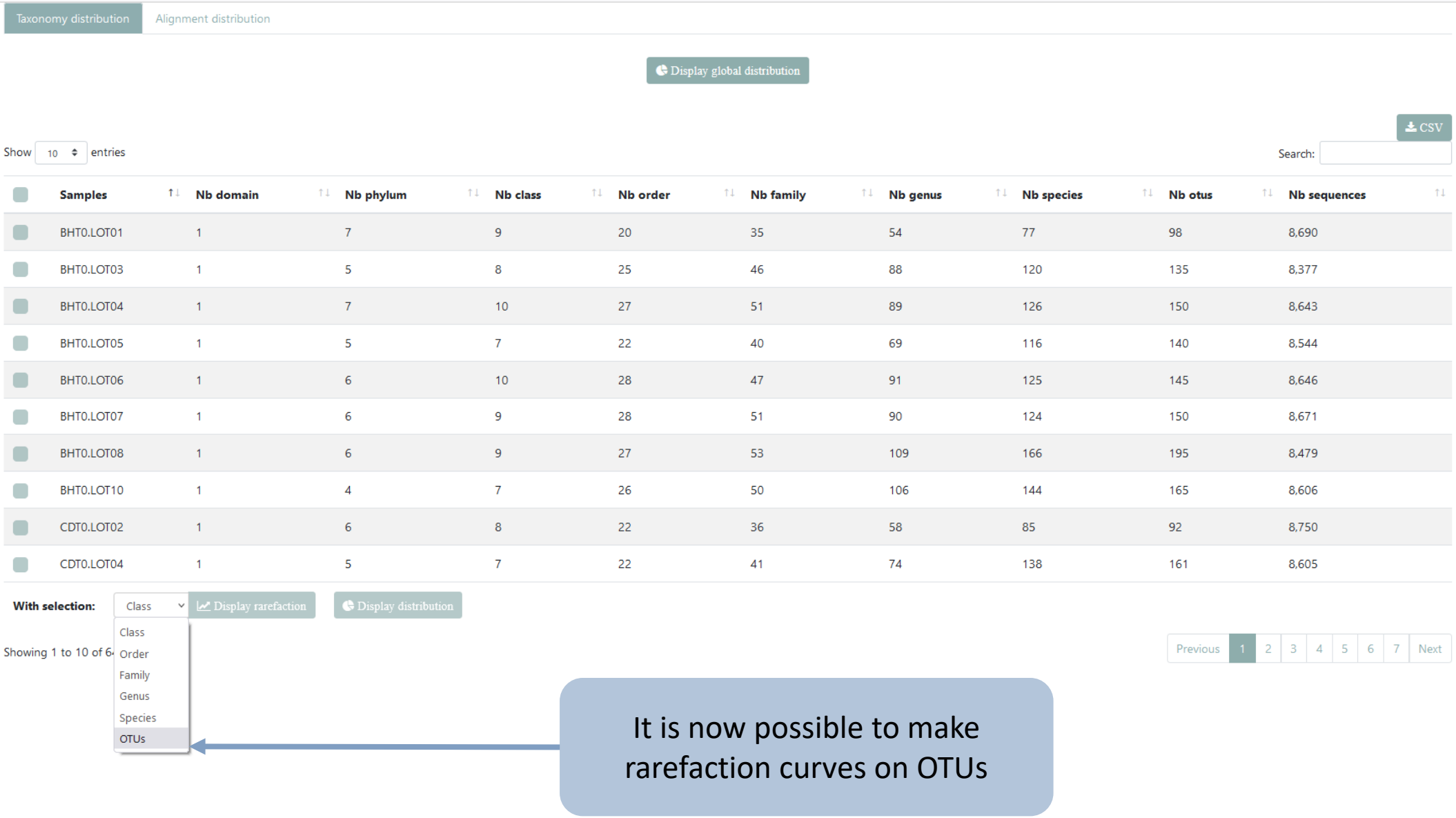
Optional but very useful, insert sequence of ASV in the abundance table

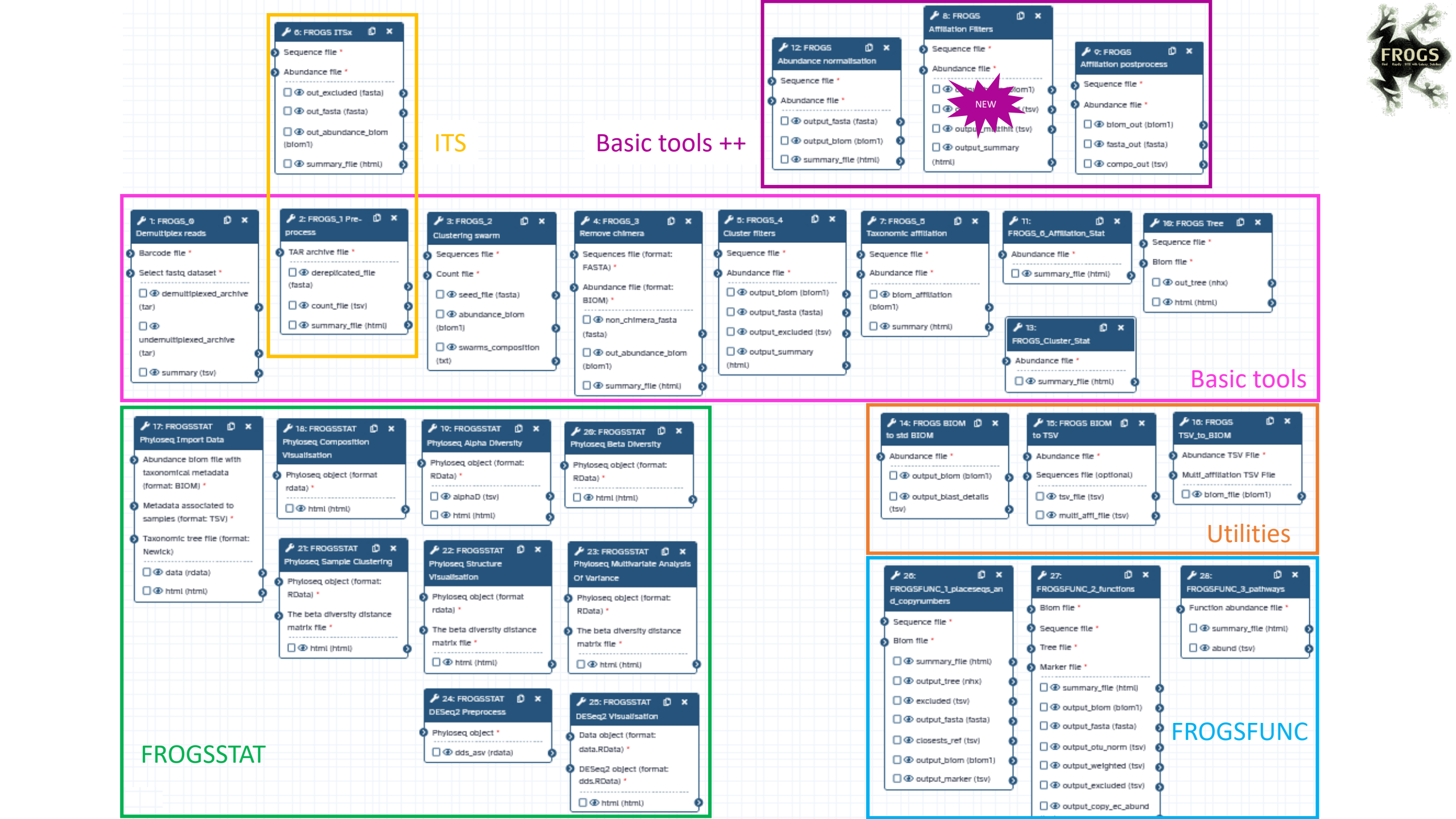
Build the multi-affiliations.tsv: the list of possible affiliations for each ambiguous ASV with multiaffiliation



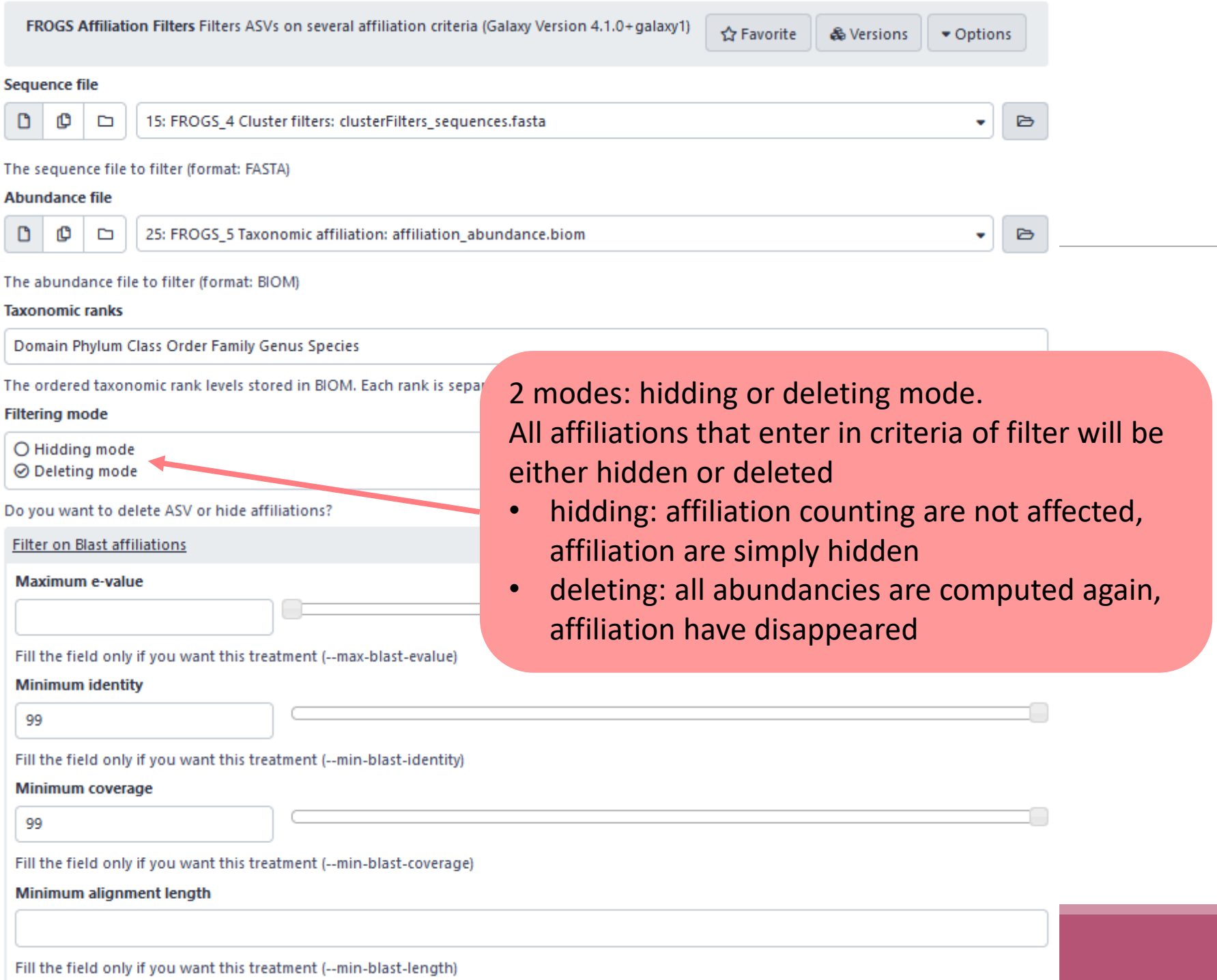

6: FROGS ITSx
Sequence file *
Abundance file *
out_excluded (fasta)
out_fasta (fasta)
out_abundance_biom (biom1)
summary_file (html)
ITS
Basic tools ++
12: FROGS Abundance normalisation
Sequence file *
Abundance file *
output_fasta (fasta)
output_biom (biom1)
summary_file (html)
8: FROGS Affiliation Filters
Sequence file *
Abundance file *
output_biom (biom1)
output_impacted (tsv)
output_multihit (tsv)
output_summary (html)
9: FROGS Affiliation postprocess
Sequence file *
Abundance file *
biom_out (biom1)
fasta_out (fasta)
compo_out (tsv)
1: FROGS_0 Demultiplex reads
Barcode file *
Select fastq dataset *
demultiplexed_archive (tar)
undemultiplexed_archive (tar)
summary (tsv)
2: FROGS_1 Pre-process
TAR archive file *
dereplicated_file (fasta)
count_file (tsv)
summary_file (html)
3: FROGS_2 Clustering swarm
Sequences file *
Count file *
seed_file (fasta)
abundance_biom (biom1)
swarms_composition (txt)
4: FROGS_3 Remove chimera
Sequences file (format: FASTA) *
Abundance file (format: BIOM) *
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
summary_file (html)
5: FROGS_4 Cluster filters
Sequence file *
Abundance file *
output_biom (biom1)
output_fasta (fasta)
output_excluded (tsv)
output_summary (html)
7: FROGS_5 Taxonomic affiliation
Sequence file *
Abundance file *
biom_affiliation (biom1)
summary (html)
11: FROGS_6_Affiliation_Stat
NEW
10: FROGS Tree
Sequence file *
Biom file *
out_tree (nhx)
html (html)
13: FROGS_Cluster_Stat
Abundance file *
summary_file (html)
Basic tools
17: FROGSSTAT Phyloseq Import Data
Abundance biom file with taxonomical metadata (format: BIOM) *
Metadata associated to samples (format: TSV) *
Taxonomic tree file (format: Newick)
data (rdata)
html (html)
18: FROGSSTAT Phyloseq Composition Visualisation
Phyloseq object (format rdata) *
html (html)
19: FROGSSTAT Phyloseq Alpha Diversity
Phyloseq object (format: RData) *
alphaD (tsv)
html (html)
20: FROGSSTAT Phyloseq Beta Diversity
Phyloseq object (format: RData) *
html (html)
21: FROGSSTAT Phyloseq Sample Clustering
Phyloseq object (format: RData) *
The beta diversity distance matrix file *
html (html)
22: FROGSSTAT Phyloseq Structure Visualisation
Phyloseq object (format rdata) *
The beta diversity distance matrix file *
html (html)
23: FROGSSTAT Phyloseq Multivariate Analysis Of Variance
Phyloseq object (format: RData) *
The beta diversity distance matrix file *
html (html)
24: FROGSSTAT DESeq2 Preprocess
Phyloseq object *
dds_asv (rdata)
25: FROGSSTAT DESeq2 Visualisation
Data object (format: data.RData) *
DESeq2 object (format: dds.RData) *
html (html)
FROGSSTAT
14: FROGS BIOM to std BIOM
Abundance file *
output_biom (biom1)
output_blast_details (tsv)
15: FROGS BIOM to TSV
Abundance file *
Sequences file (optional)
tsv_file (tsv)
multi_affi_file (tsv)
16: FROGS TSV_to_BIOM
Abundance TSV File *
Multi_affiliation TSV File
biom_file (biom1)
Utilities
26: FROGSFUNC_1_placeseqs_and_copynumbers
Sequence file *
Biom file *
summary_file (html)
output_tree (nhx)
excluded (tsv)
output_fasta (fasta)
closests_ref (tsv)
output_biom (biom1)
output_marker (tsv)
27: FROGSFUNC_2_functions
Biom file *
Sequence file *
Tree file *
Marker file *
summary_file (html)
output_biom (biom1)
output_fasta (fasta)
output_otu_norm (tsv)
output_weighted (tsv)
output_excluded (tsv)
output_copy_ec_abund
28: FROGSFUNC_3_pathways
Function abundance file *
summary_file (html)
abund (tsv)
FROGSFUNC

Affiliation Stat


Taxonomy distribution
Alignment distribution
Display global distribution
CSV
Show 10 entries
Search:
Samples
Nb domain
Nb phylum
Nb class
Nb order
Nb family
Nb genus
Nb species
Nb otus
Nb sequences
BHT0.LOT01 1 7 9 20 35 54 77 98 8,690
BHT0.LOT03 1 5 8 25 46 88 120 135 8,377
BHT0.LOT04 1 7 10 27 51 89 126 150 8,643
BHT0.LOT05 1 5 7 22 40 69 116 140 8,544
BHT0.LOT06 1 6 10 28 47 91 125 145 8,646
BHT0.LOT07 1 6 9 28 51 90 124 150 8,671
BHT0.LOT08 1 6 9 27 53 109 166 195 8,479
BHT0.LOT10 1 4 7 26 50 106 144 165 8,606
CDT0.LOT02 1 6 8 22 36 58 85 92 8,750
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

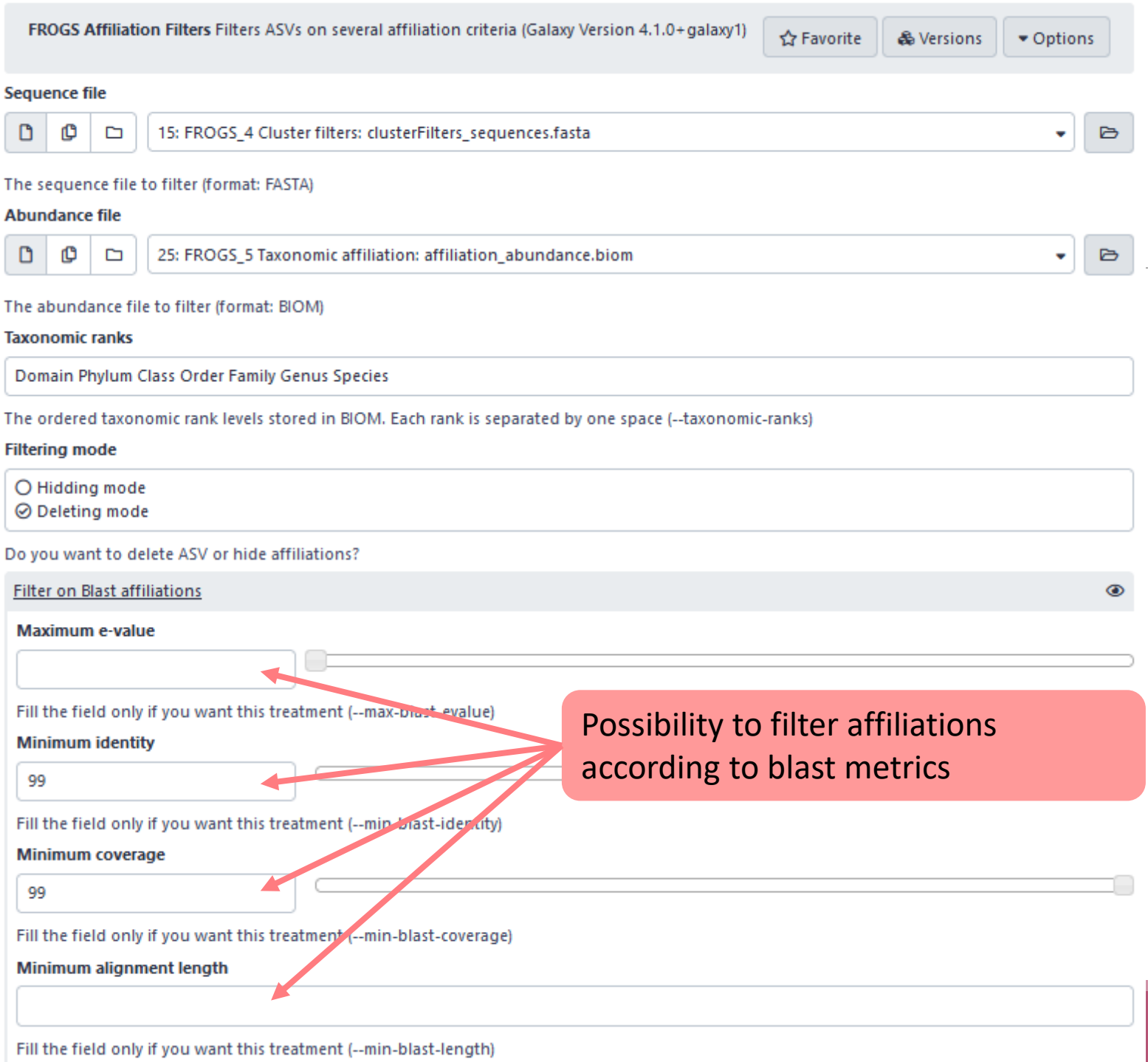

6: FROGS ITSx
Sequence file *
Abundance file *
out_excluded (fasta)
out_fasta (fasta)
out_abundance_biom (biom1)
summary_file (html)
ITS
Basic tools ++
12: FROGS Abundance normalisation
Sequence file *
Abundance file *
output_fasta (fasta)
output_biom (biom1)
summary_file (html)
8: FROGS Affiliation Filters
Sequence file *
Abundance file *
output_summary (html)
NEW
9: FROGS Affiliation postprocess
Sequence file *
Abundance file *
biom_out (biom1)
fasta_out (fasta)
compo_out (tsv)
1: FROGS_0 Demultiplex reads
Barcode file *
Select fastq dataset *
demultiplexed_archive (tar)
undemultiplexed_archive (tar)
summary (tsv)
2: FROGS_1 Pre-process
TAR archive file *
dereplicated_file (fasta)
count_file (tsv)
summary_file (html)
3: FROGS_2 Clustering swarm
Sequences file *
Count file *
seed_file (fasta)
abundance_biom (biom1)
swarms_composition (txt)
4: FROGS_3 Remove chimera
Sequences file (format: FASTA) *
Abundance file (format: BIOM) *
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
summary_file (html)
5: FROGS_4 Cluster filters
Sequence file *
Abundance file *
output_biom (biom1)
output_fasta (fasta)
output_excluded (tsv)
output_summary (html)
7: FROGS_5 Taxonomic affiliation
Sequence file *
Abundance file *
biom_affiliation (biom1)
summary (html)
11: FROGS_6_Affiliation_Stat
Abundance file *
summary_file (html)
13: FROGS_Cluster_Stat
Abundance file *
summary_file (html)
10: FROGS Tree
Sequence file *
Biom file *
out_tree (nhx)
html (html)
Basic tools
17: FROGSSTAT Phyloseq Import Data
Abundance biom file with taxonomical metadata (format: BIOM) *
Metadata associated to samples (format: TSV) *
Taxonomic tree file (format: Newick)
data (rdata)
html (html)
18: FROGSSTAT Phyloseq Composition Visualisation
Phyloseq object (format rdata) *
html (html)
19: FROGSSTAT Phyloseq Alpha Diversity
Phyloseq object (format: RData) *
alphaD (tsv)
html (html)
20: FROGSSTAT Phyloseq Beta Diversity
Phyloseq object (format: RData) *
html (html)
21: FROGSSTAT Phyloseq Sample Clustering
Phyloseq object (format: RData) *
The beta diversity distance matrix file *
html (html)
22: FROGSSTAT Phyloseq Structure Visualisation
Phyloseq object (format rdata) *
The beta diversity distance matrix file *
html (html)
23: FROGSSTAT Phyloseq Multivariate Analysis Of Variance
Phyloseq object (format: RData) *
The beta diversity distance matrix file *
html (html)
24: FROGSSTAT DESeq2 Preprocess
Phyloseq object *
dds_asv (rdata)
25: FROGSSTAT DESeq2 Visualisation
Data object (format: data.RData) *
DESeq2 object (format: dds.RData) *
html (html)
FROGSSTAT
14: FROGS BIOM to std BIOM
Abundance file *
output_biom (biom1)
output_blast_details (tsv)
15: FROGS BIOM to TSV
Abundance file *
Sequences file (optional)
tsv_file (tsv)
multi_affi_file (tsv)
16: FROGS TSV_to_BIOM
Abundance TSV File *
Multi_affiliation TSV File
biom_file (biom1)
Utilities
26: FROGSFUNC_1_placeseqs_and_copynumbers
Sequence file *
Biom file *
summary_file (html)
output_tree (nhx)
excluded (tsv)
output_fasta (fasta)
closests_ref (tsv)
output_biom (biom1)
output_marker (tsv)
27: FROGSFUNC_2_functions
Biom file *
Sequence file *
Tree file *
Marker file *
summary_file (html)
output_biom (biom1)
output_fasta (fasta)
output_otu_norm (tsv)
output_weighted (tsv)
output_excluded (tsv)
output_copy_ec_abund
28: FROGSFUNC_3_pathways
Function abundance file *
summary_file (html)
abund (tsv)
FROGSFUNC

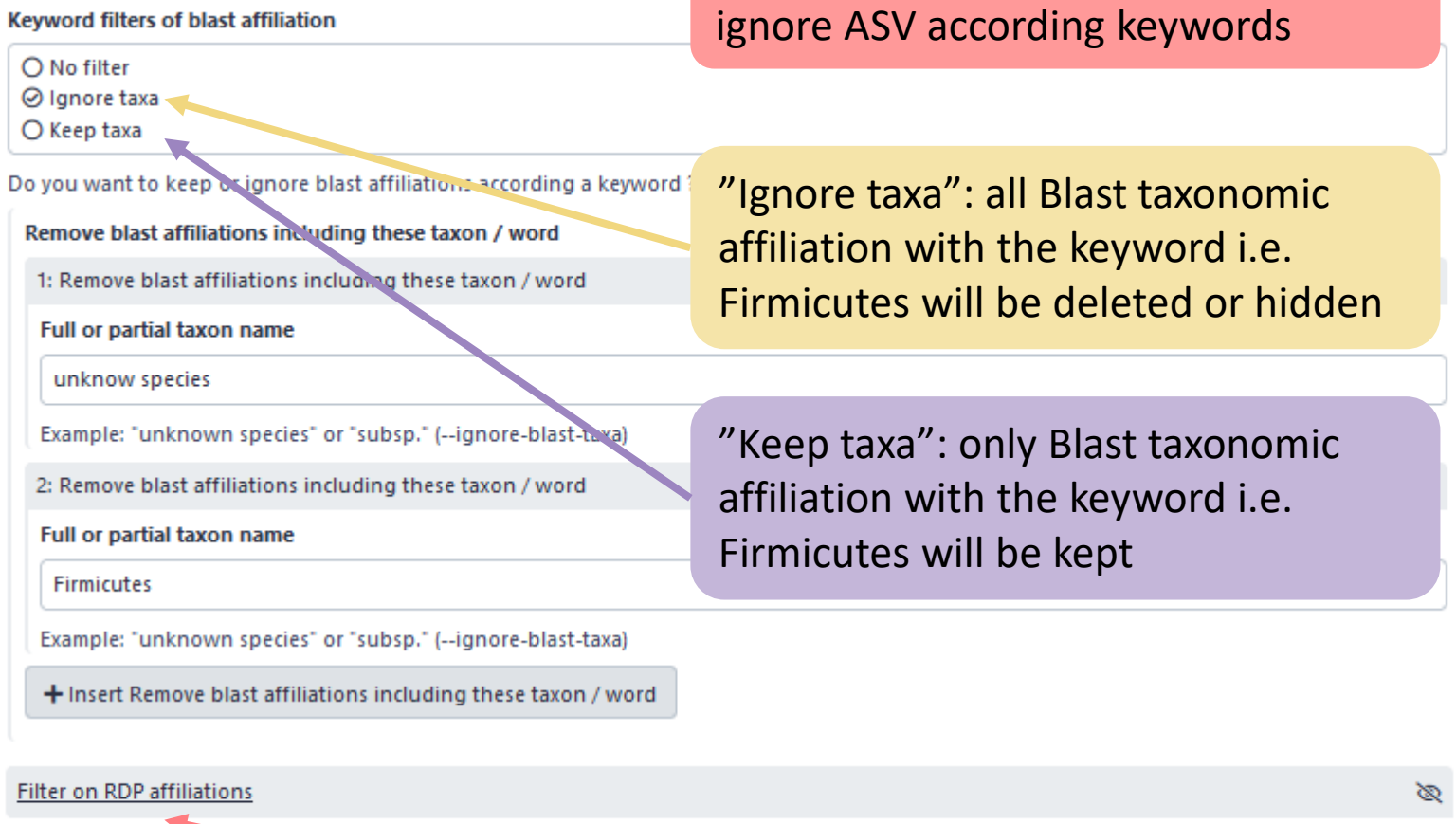
Filters on affiliations



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


FROGS Affiliation Filters Filters ASVs on several affiliation criteria (Galaxy Version 4.1.0+galaxy1)
Favorite
Versions
Options
Sequence file
15: FROGS_4 Cluster filters: clusterFilters_sequences.fasta
The sequence file to filter (format: FASTA)
Abundance file
25: FROGS_5 Taxonomic affiliation: affiliation_abundance.biom
The abundance file to filter (format: BIOM)
Taxonomic ranks
Domain Phylum Class Order Family Genus Species
The ordered taxonomic rank levels stored in BIOM. Each rank is separated by one space (--taxonomic-ranks)
Filtering mode
Hidding mode
Deleting mode
Do you want to delete ASV or hide affiliations?
Filter on Blast affiliations
Maximum e-value
Fill the field only if you want this treatment (--max-blast-evalue)
Minimum identity
99
Fill the field only if you want this treatment (--min-blast-identity)
Minimum coverage
99
Fill the field only if you want this treatment (--min-blast-coverage)
Minimum alignment length
Fill the field only if you want this treatment (--min-blast-length)
Possibility to filter affiliations according to blast metrics



Possibility to filter for keeping or for ignore ASV according keywords

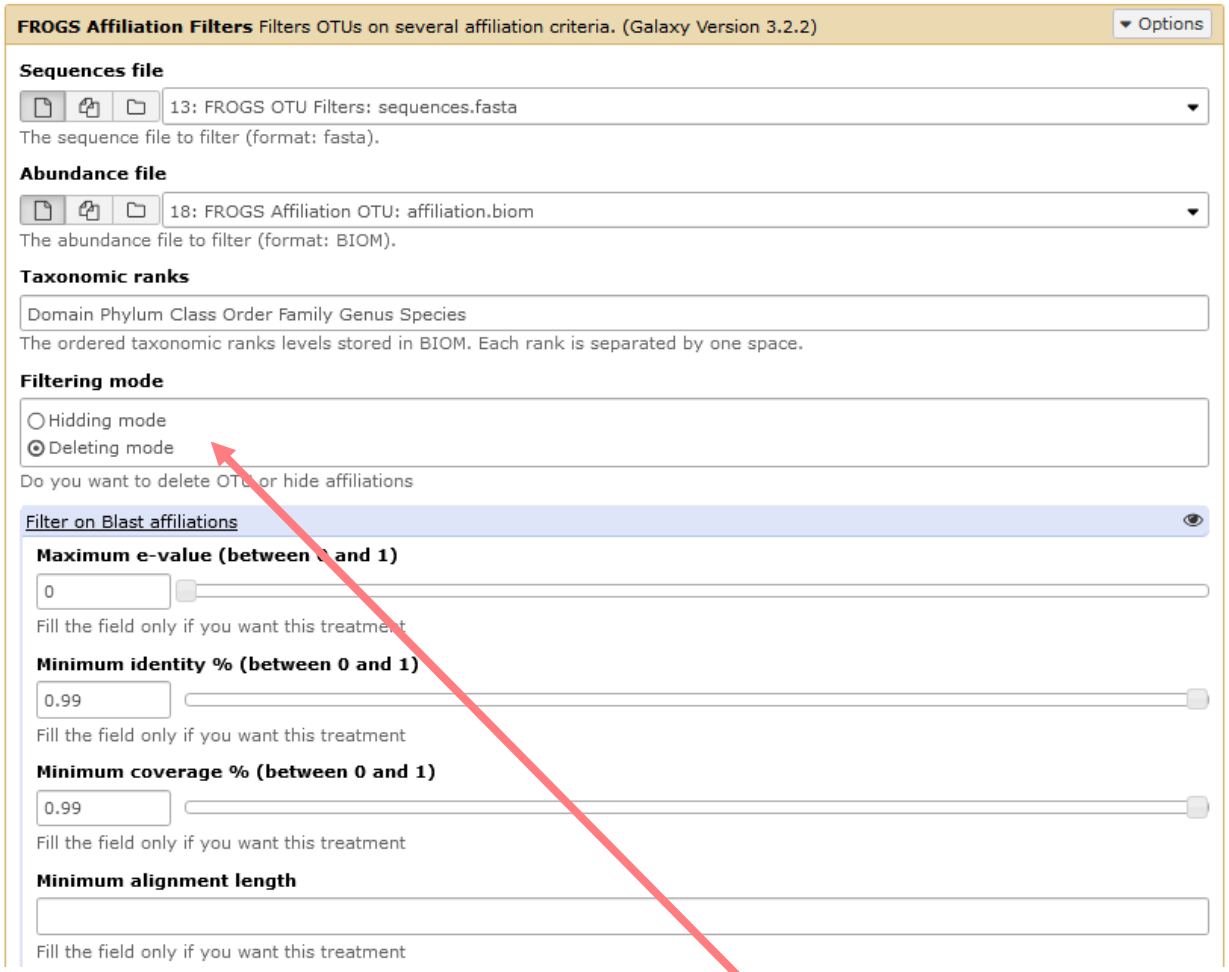
”Ignore taxa”: all Blast taxonomic affiliation with the keyword i.e. Firmicutes will be deleted or hidden

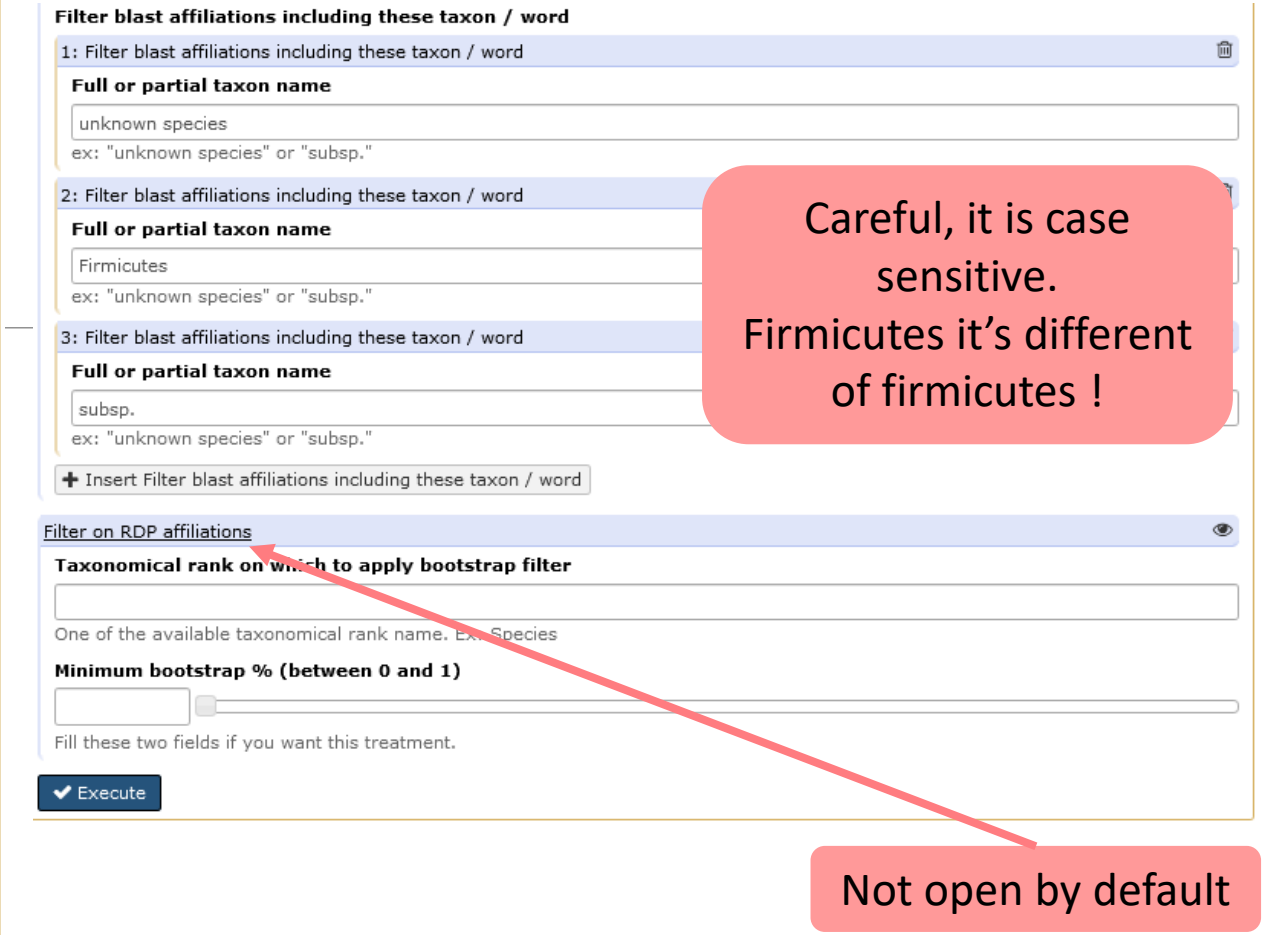
”Keep taxa”: only Blast taxonomic affiliation with the keyword i.e. Firmicutes will be kept

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

Possibility to filter on RDP taxonomic affiliation

Not open by default





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

Exercice:

1. Mask
    1. all ASV that have not at least 95% identity and 95% coverage with a SILVA sequence
    2. and that are not a *unknown species*

2. Explore the report.html
    - How many ASVs remain?
    - How are impacted affiliation?

Answer 1

FROGS Affiliation Filters Filters ASVs on several affiliation criteria (Galaxy Version 4.1.0+galaxy1) ☆ Favorite ▾ Options

Sequence file

111: FROGS_4 Cluster filters: clusterFilters_sequences.fasta

The sequence file to filter (format: FASTA)

Abundance file

115: FROGS_5 Taxonomic affiliation: affiliation_abundance.biom

The abundance file to filter (format: BIOM)

Taxonomic ranks

Domain Phylum Class Order Family Genus Species

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

Filtering mode

- (◉) Hidding mode
- ( ) Deleting mode

Do you want to delete ASV or hide affiliations?

Filter on Blast affiliations

Maximum e-value

Fill the field only if you want this treatment (--max-blast-evalue)

Minimum identity

95

Fill the field only if you want this treatment (--min-blast-identity)

Minimum coverage

95

Fill the field only if you want this treatment (--min-blast-coverage)

Minimum alignment length

Fill the field only if you want this treatment (--min-blast-length)

Keyword filters of blast affiliation

- ( ) No filter
- (◉) Ignore taxa
- ( ) Keep taxa

Do you want to keep or ignore blast affiliations according a keyword ?

Remove blast affiliations including these taxon / word

1: Remove blast affiliations including these taxon / word

Full or partial taxon name

unknown species

Example: "unknown species" or "subsp." (--ignore-blast-taxa)

+ Insert Remove blast affiliations including these taxon / word

Filter on RDP affiliations

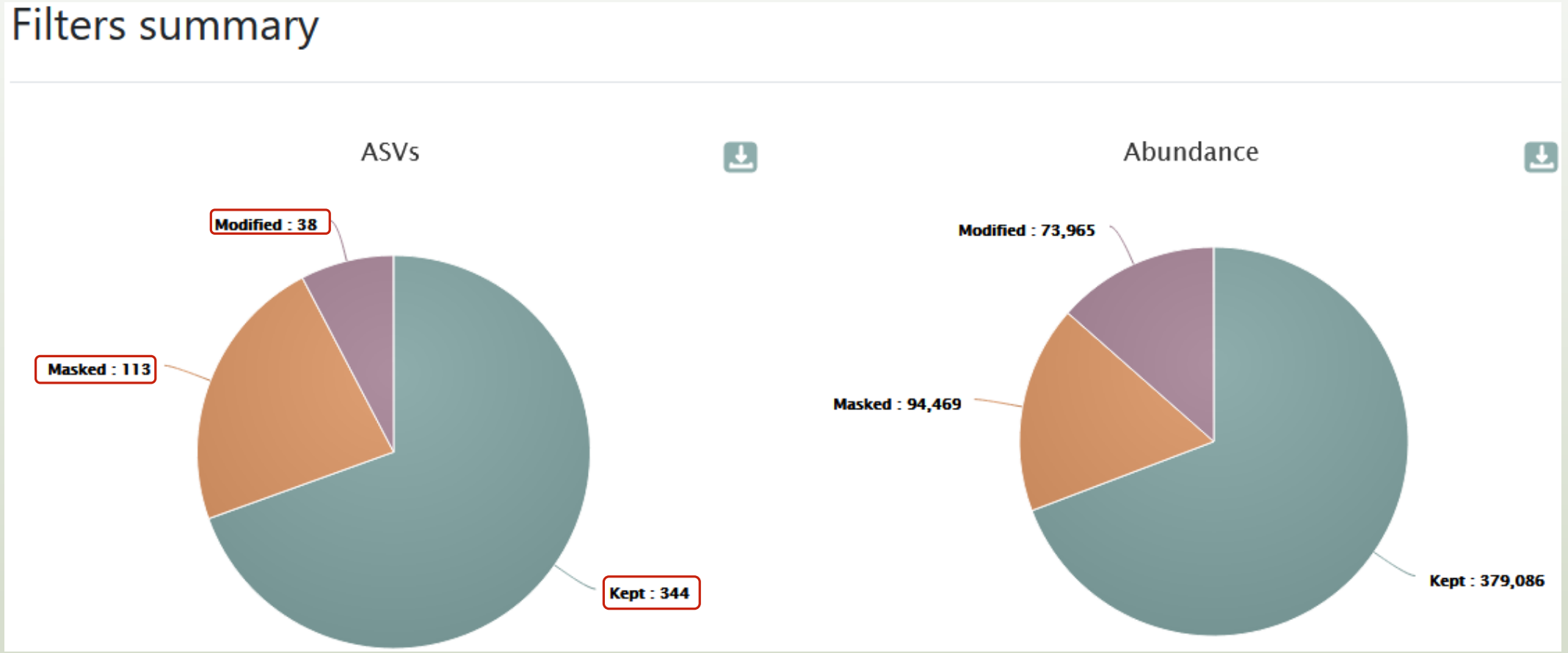
Email notification

No

Send an email notification when the job completes.

✔ Execute

Answer 2



- 344 ASV are kept without modification
- 38 ASV are kept with modification (see **impacted_clusters.multi-affiliation.tsv)**
- It’s remain 382 ASVs !

Answer 2

42: FROGS Affiliation Filters: impacted_clusters.multi-affiliations.tsv

Cluster_3	Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Latilactobacillus;Lactobacillus sakei
Cluster_3	Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Latilactobacillus;Lactobacillus sakei
Cluster_3	Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Latilactobacillus;Lactobacillus sakei
Cluster_3	Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Latilactobacillus;Lactobacillus sakei
Cluster_3	Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Latilactobacillus;Lactobacillus sakei
Cluster_3	Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Latilactobacillus;unknown species

Exemple: Cluster_3 is an impacted clusters because

- its multi-affiliation “unknow species” was deleted
- but all other affiliation were kept.

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

41: FROGS Affiliation Filters: impacted_clusters.tsv

#comment	status	blast_taxonomy
undesired_tax_in_blast	Affiliation_masked	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales;Vibrionaceae;Photobacterium;unknown species
undesired_tax_in_blast	Blast_taxonomy_changed	Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Latilactobacillus;Multi-affiliation
blast_identity_lt_95.0;undesired_tax_in_blast	Affiliation_masked	Bacteria;Firmicutes;Bacilli;Erysipelotrichales;Erysipelotrichaceae;ZOR0006;unknown species
undesired_tax_in_blast	Blast_taxonomy_changed	Bacteria;Firmicutes;Bacilli;Lactobacillales;Streptococcaceae;Lactococcus;Multi-affiliation
undesired_tax_in_blast	Affiliation_masked	Bacteria;Fusobacteriota;Fusobacteriia;Fusobacteriales;Leptotrichiaceae;Hypnocyclicus;unknown species
undesired_tax_in_blast	Affiliation_masked	Bacteria;Firmicutes;Bacilli;Lactobacillales;Carnobacteriaceae;Carnobacterium;unknown species
undesired_tax_in_blast	Affiliation_masked	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales;Vibrionaceae;Photobacterium;unknown species
undesired_tax_in_blast	Affiliation_masked	Bacteria;Firmicutes;Bacilli;Mycoplasmatales;Mycoplasmataceae;Candidatus Bacilloplasma;unknown species
undesired_tax_in_blast	Blast_taxonomy_changed	Bacteria;Bacteroidota;Bacteroidia;Flavobacteriales;Weeksellaceae;Chryseobacterium;Multi-affiliation

In impacted_cluster.tsv

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

Hidding mode

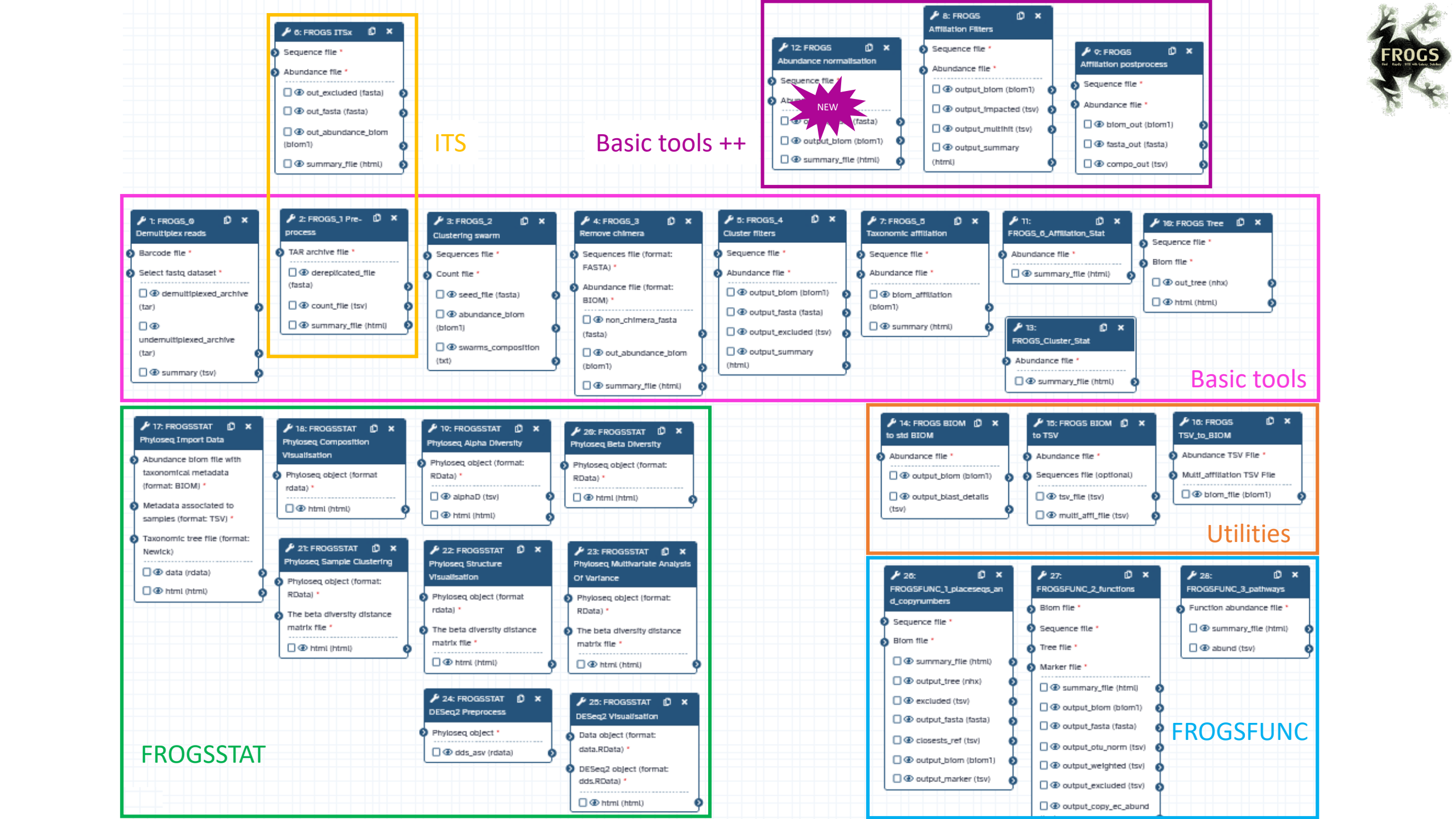
#comment	blast_taxonomy	blast_subject	blast_perc_i	blast_perc_c	blast_evalue	blast_aln_le
no data	Bacteria;Firmicutes;Bacilli;Lactobacillales;Listeriaceae;Brochothrix;Brochothrix thermosphacta	multi-subject	100.0	100.0	0.0	497
undesired_tax_in_blast	no data	no data	no data	no data	no data	no data
undesired_tax_in_blast	Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Latilactobacillus;Lactobacillus sakei	multi-subject	100.0	100.0	0.0	520
undesired_tax_in_blast	Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Multi-affiliation	multi-subject	100.0	100.0	0.0	468
no data	Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Leuconostoc;Multi-affiliation	multi-subject	100.0	100.0	0.0	497
no data	Bacteria;Firmicutes;Bacilli;Lactobacillales;Streptococcaceae;Lactococcus;Lactococcus piscium	AM943029.1.1242	99.799	100.0	0.0	497

Deleting mode

#comment	blast_taxonomy	blast_subject	blast_perc_i	blast_perc_c	blast_evalue	blast_aln_le
no data	Bacteria;Firmicutes;Bacilli;Lactobacillales;Listeriaceae;Brochothrix;Brochothrix thermosphacta	multi-subject	100.0	100.0	0.0	497
undesired_tax_in_blast	Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Latilactobacillus;Lactobacillus sakei	multi-subject	100.0	100.0	0.0	520
undesired_tax_in_blast	Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Multi-affiliation	multi-subject	100.0	100.0	0.0	468
no data	Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Leuconostoc;Multi-affiliation	multi-subject	100.0	100.0	0.0	497
no data	Bacteria;Firmicutes;Bacilli;Lactobacillales;Streptococcaceae;Lactococcus;Lactococcus piscium	AM943029.1.1242	99.799	100.0	0.0	497

Remark

In deleting mode, in the abundance table, all information concerning the ASVs affected by the filter are removed (affiliation, metrics and count in the different samples)


6: FROGS ITSx
Sequence file *
Abundance file *
out_excluded (fasta)
out_fasta (fasta)
out_abundance_biom (biom1)
summary_file (html)
ITS
Basic tools ++
12: FROGS Abundance normalisation
Sequence file
NEW
output_biom (biom1)
summary_file (html)
8: FROGS Affiliation Filters
Sequence file *
Abundance file *
output_biom (biom1)
output_impacted (tsv)
output_multihit (tsv)
output_summary (html)
9: FROGS Affiliation postprocess
Sequence file *
Abundance file *
biom_out (biom1)
fasta_out (fasta)
compo_out (tsv)
1: FROGS_0 Demultiplex reads
Barcode file *
Select fastq dataset *
demultiplexed_archive (tar)
undemultiplexed_archive (tar)
summary (tsv)
2: FROGS_1 Pre-process
TAR archive file *
dereplicated_file (fasta)
count_file (tsv)
summary_file (html)
3: FROGS_2 Clustering swarm
Sequences file *
Count file *
seed_file (fasta)
abundance_biom (biom1)
swarms_composition (txt)
4: FROGS_3 Remove chimera
Sequences file (format: FASTA) *
Abundance file (format: BIOM) *
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
summary_file (html)
5: FROGS_4 Cluster filters
Sequence file *
Abundance file *
output_biom (biom1)
output_fasta (fasta)
output_excluded (tsv)
output_summary (html)
7: FROGS_5 Taxonomic affiliation
Sequence file *
Abundance file *
biom_affiliation (biom1)
summary (html)
11: FROGS_6_Affiliation_Stat
Abundance file *
summary_file (html)
13: FROGS_Cluster_Stat
Abundance file *
summary_file (html)
10: FROGS Tree
Sequence file *
Biom file *
out_tree (nhx)
html (html)
Basic tools
17: FROGSSTAT Phyloseq Import Data
Abundance biom file with taxonomical metadata (format: BIOM) *
Metadata associated to samples (format: TSV) *
Taxonomic tree file (format: Newick)
data (rdata)
html (html)
18: FROGSSTAT Phyloseq Composition Visualisation
Phyloseq object (format rdata) *
html (html)
19: FROGSSTAT Phyloseq Alpha Diversity
Phyloseq object (format: RData) *
alphaD (tsv)
html (html)
20: FROGSSTAT Phyloseq Beta Diversity
Phyloseq object (format: RData) *
html (html)
21: FROGSSTAT Phyloseq Sample Clustering
Phyloseq object (format: RData) *
The beta diversity distance matrix file *
html (html)
22: FROGSSTAT Phyloseq Structure Visualisation
Phyloseq object (format rdata) *
The beta diversity distance matrix file *
html (html)
23: FROGSSTAT Phyloseq Multivariate Analysis Of Variance
Phyloseq object (format: RData) *
The beta diversity distance matrix file *
html (html)
24: FROGSSTAT DESeq2 Preprocess
Phyloseq object *
dds_asv (rdata)
25: FROGSSTAT DESeq2 Visualisation
Data object (format: data.RData) *
DESeq2 object (format: dds.RData) *
html (html)
FROGSSTAT
14: FROGS BIOM to std BIOM
Abundance file *
output_biom (biom1)
output_blast_details (tsv)
15: FROGS BIOM to TSV
Abundance file *
Sequences file (optional)
tsv_file (tsv)
multi_affi_file (tsv)
16: FROGS TSV_to_BIOM
Abundance TSV File *
Multi_affiliation TSV File
biom_file (biom1)
Utilities
26: FROGSFUNC_1_placeseqs_and_copynumbers
Sequence file *
Biom file *
summary_file (html)
output_tree (nhx)
excluded (tsv)
output_fasta (fasta)
closests_ref (tsv)
output_biom (biom1)
output_marker (tsv)
27: FROGSFUNC_2_functions
Biom file *
Sequence file *
Tree file *
Marker file *
summary_file (html)
output_biom (biom1)
output_fasta (fasta)
output_otu_norm (tsv)
output_weighted (tsv)
output_excluded (tsv)
output_copy_ec_abund
28: FROGSFUNC_3_pathways
Function abundance file *
summary_file (html)
abund (tsv)
FROGSFUNC

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

- (x) 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

Case 1

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

Case 2

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

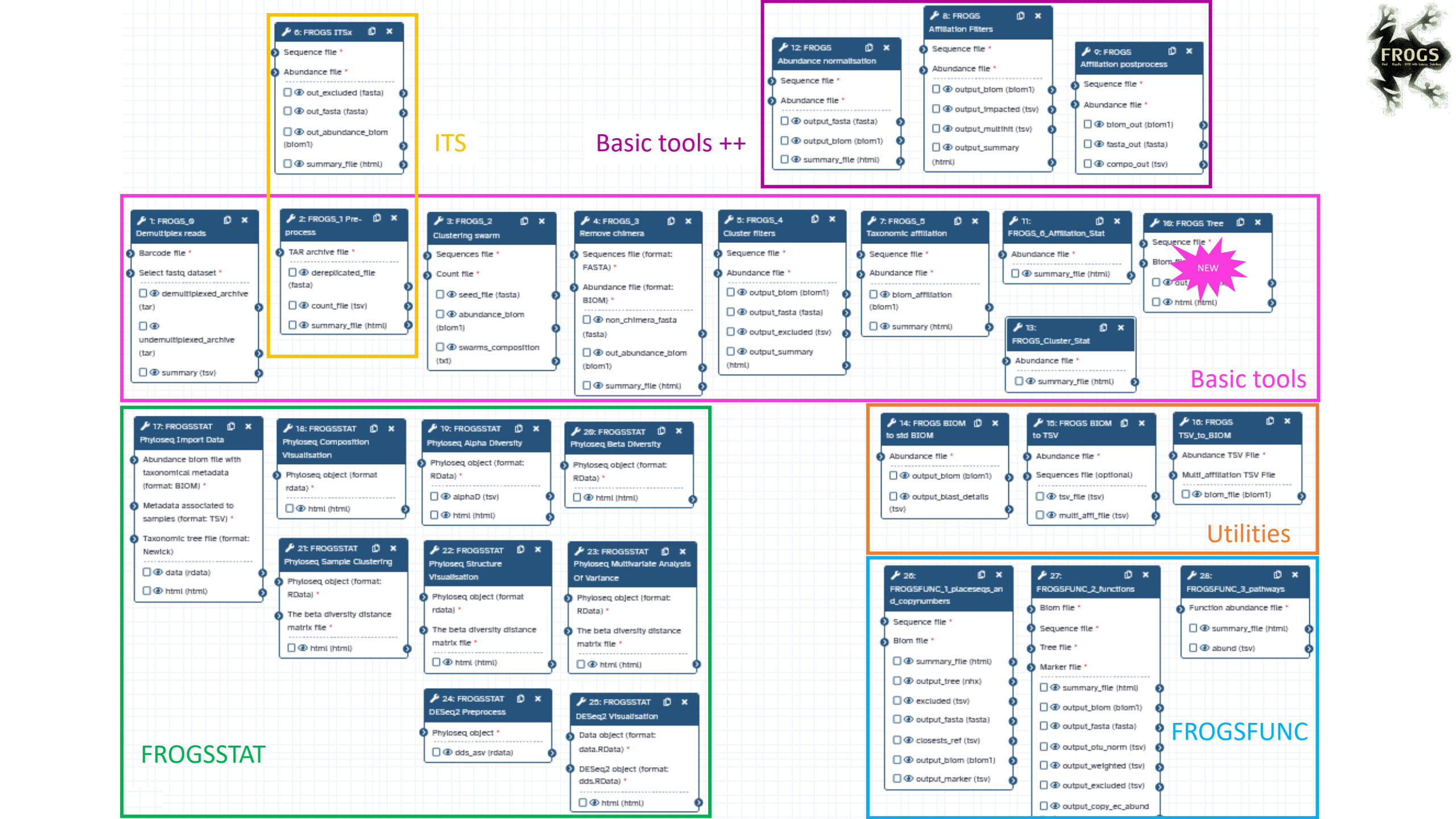
Number of reads

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


6: FROGS ITSx
Sequence file *
Abundance file *
out_excluded (fasta)
out_fasta (fasta)
out_abundance_biom (biom1)
summary_file (html)
ITS
Basic tools ++
12: FROGS Abundance normalisation
Sequence file *
Abundance file *
output_fasta (fasta)
output_biom (biom1)
summary_file (html)
8: FROGS Affiliation Filters
Sequence file *
Abundance file *
output_biom (biom1)
output_impacted (tsv)
output_multihit (tsv)
output_summary (html)
9: FROGS Affiliation postprocess
Sequence file *
Abundance file *
biom_out (biom1)
fasta_out (fasta)
compo_out (tsv)
1: FROGS_0 Demultiplex reads
Barcode file *
Select fastq dataset *
demultiplexed_archive (tar)
undemultiplexed_archive (tar)
summary (tsv)
2: FROGS_1 Pre-process
TAR archive file *
dereplicated_file (fasta)
count_file (tsv)
summary_file (html)
3: FROGS_2 Clustering swarm
Sequences file *
Count file *
seed_file (fasta)
abundance_biom (biom1)
swarms_composition (txt)
4: FROGS_3 Remove chimera
Sequences file (format: FASTA) *
Abundance file (format: BIOM) *
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
summary_file (html)
5: FROGS_4 Cluster filters
Sequence file *
Abundance file *
output_biom (biom1)
output_fasta (fasta)
output_excluded (tsv)
output_summary (html)
7: FROGS_5 Taxonomic affiliation
Sequence file *
Abundance file *
biom_affiliation (biom1)
summary (html)
11: FROGS_6_Affiliation_Stat
Abundance file *
summary_file (html)
13: FROGS_Cluster_Stat
Abundance file *
summary_file (html)
10: FROGS Tree
Sequence file *
Biom file *
NEW
html (html)
Basic tools
17: FROGSSTAT Phyloseq Import Data
Abundance biom file with taxonomical metadata (format: BIOM) *
Metadata associated to samples (format: TSV) *
Taxonomic tree file (format: Newick)
data (rdata)
html (html)
18: FROGSSTAT Phyloseq Composition Visualisation
Phyloseq object (format rdata) *
html (html)
19: FROGSSTAT Phyloseq Alpha Diversity
Phyloseq object (format: RData) *
alphaD (tsv)
html (html)
20: FROGSSTAT Phyloseq Beta Diversity
Phyloseq object (format: RData) *
html (html)
21: FROGSSTAT Phyloseq Sample Clustering
Phyloseq object (format: RData) *
The beta diversity distance matrix file *
html (html)
22: FROGSSTAT Phyloseq Structure Visualisation
Phyloseq object (format rdata) *
The beta diversity distance matrix file *
html (html)
23: FROGSSTAT Phyloseq Multivariate Analysis Of Variance
Phyloseq object (format: RData) *
The beta diversity distance matrix file *
html (html)
24: FROGSSTAT DESeq2 Preprocess
Phyloseq object *
dds_asv (rdata)
25: FROGSSTAT DESeq2 Visualisation
Data object (format: data.RData) *
DESeq2 object (format: dds.RData) *
html (html)
FROGSSTAT
14: FROGS BIOM to std BIOM
Abundance file *
output_biom (biom1)
output_blast_details (tsv)
15: FROGS BIOM to TSV
Abundance file *
Sequences file (optional)
tsv_file (tsv)
multi_affi_file (tsv)
16: FROGS TSV_to_BIOM
Abundance TSV File *
Multi_affiliation TSV File
biom_file (biom1)
Utilities
26: FROGSFUNC_1_placeseqs_and_copynumbers
Sequence file *
Biom file *
summary_file (html)
output_tree (nhx)
excluded (tsv)
output_fasta (fasta)
closests_ref (tsv)
output_biom (biom1)
output_marker (tsv)
27: FROGSFUNC_2_functions
Biom file *
Sequence file *
Tree file *
Marker file *
summary_file (html)
output_biom (biom1)
output_fasta (fasta)
output_otu_norm (tsv)
output_weighted (tsv)
output_excluded (tsv)
output_copy_ec_abund
28: FROGSFUNC_3_pathways
Function abundance file *
summary_file (html)
abund (tsv)
FROGSFUNC

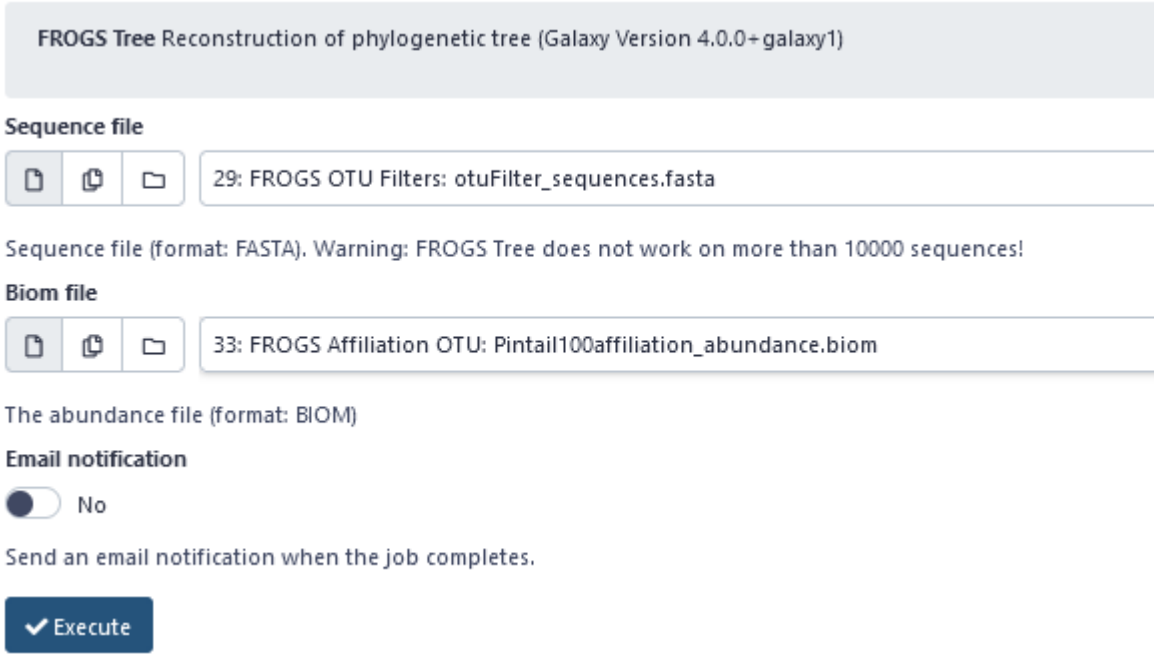
FROGS Tree

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.

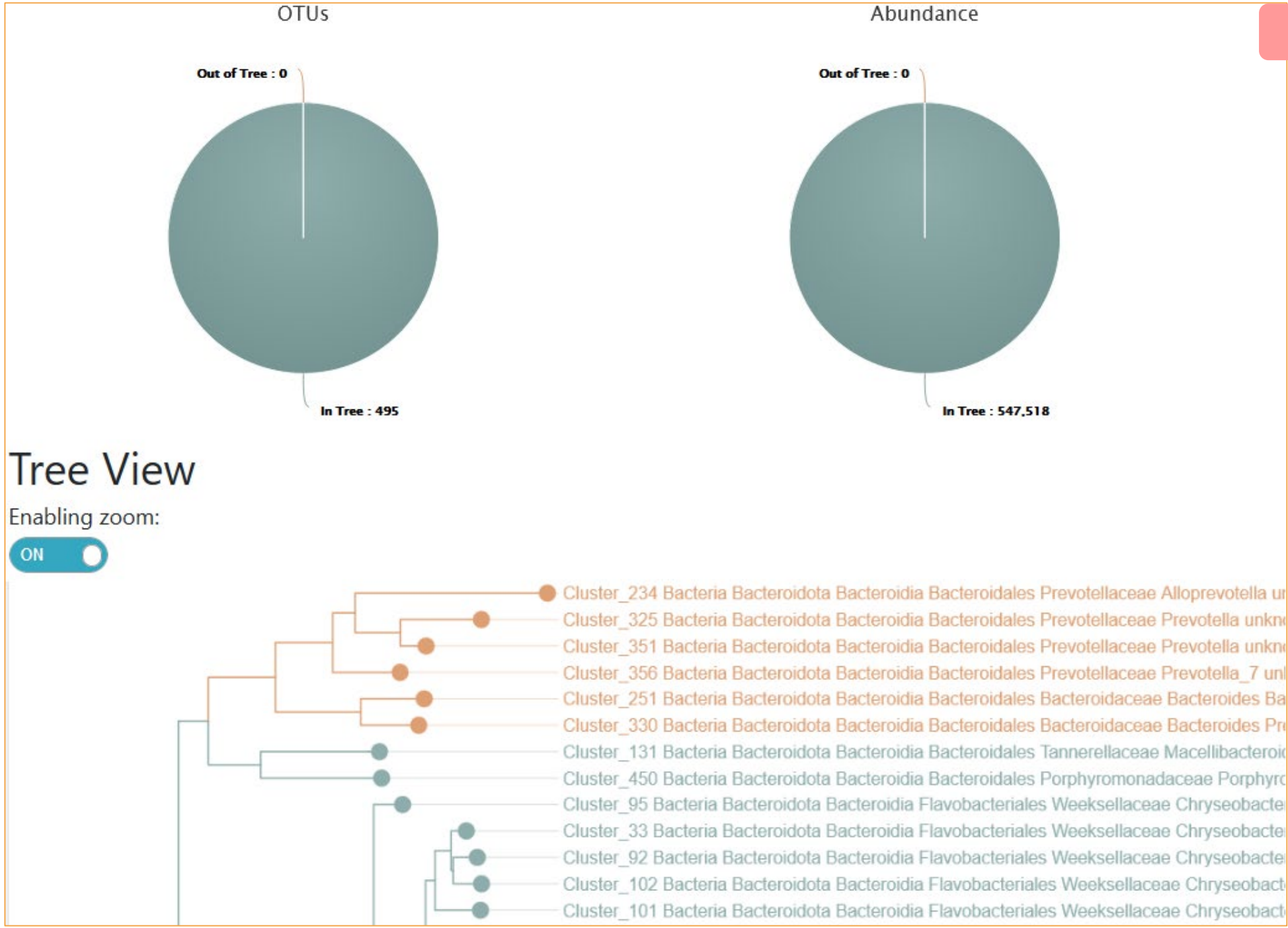


2 outputs:

FROGS Tree: report.html

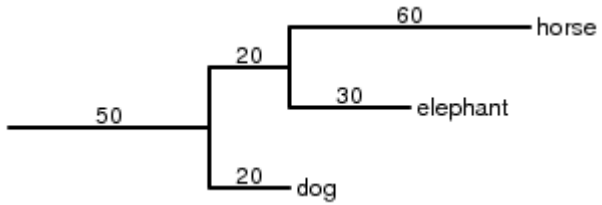
FROGS Tree: tree.nwk

1st output


OTUs
Out of Tree : 0
In Tree : 495
Abundance
Out of Tree : 0
In Tree : 547,518
Tree View
Enabling zoom:
ON
Cluster_234 Bacteria Bacteroidota Bacteroidia Bacteroidales Prevotellaceae Alloprevotella
Cluster_325 Bacteria Bacteroidota Bacteroidia Bacteroidales Prevotellaceae Prevotella
Cluster_351 Bacteria Bacteroidota Bacteroidia Bacteroidales Prevotellaceae Prevotella
Cluster_356 Bacteria Bacteroidota Bacteroidia Bacteroidales Prevotellaceae Prevotella_7
Cluster_251 Bacteria Bacteroidota Bacteroidia Bacteroidales Bacteroidaceae Bacteroides
Cluster_330 Bacteria Bacteroidota Bacteroidia Bacteroidales Bacteroidaceae Bacteroides
Cluster_131 Bacteria Bacteroidota Bacteroidia Bacteroidales Tannerellaceae Macellibacteroi
Cluster_450 Bacteria Bacteroidota Bacteroidia Bacteroidales Porphyromonadaceae Porphyro
Cluster_95 Bacteria Bacteroidota Bacteroidia Flavobacteriales Weeksellaceae Chryseobacte
Cluster_33 Bacteria Bacteroidota Bacteroidia Flavobacteriales Weeksellaceae Chryseobacte
Cluster_92 Bacteria Bacteroidota Bacteroidia Flavobacteriales Weeksellaceae Chryseobacte
Cluster_102 Bacteria Bacteroidota Bacteroidia Flavobacteriales Weeksellaceae Chryseobact
Cluster_101 Bacteria Bacteroidota Bacteroidia Flavobacteriales Weeksellaceae Chryseobact

2nd output

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

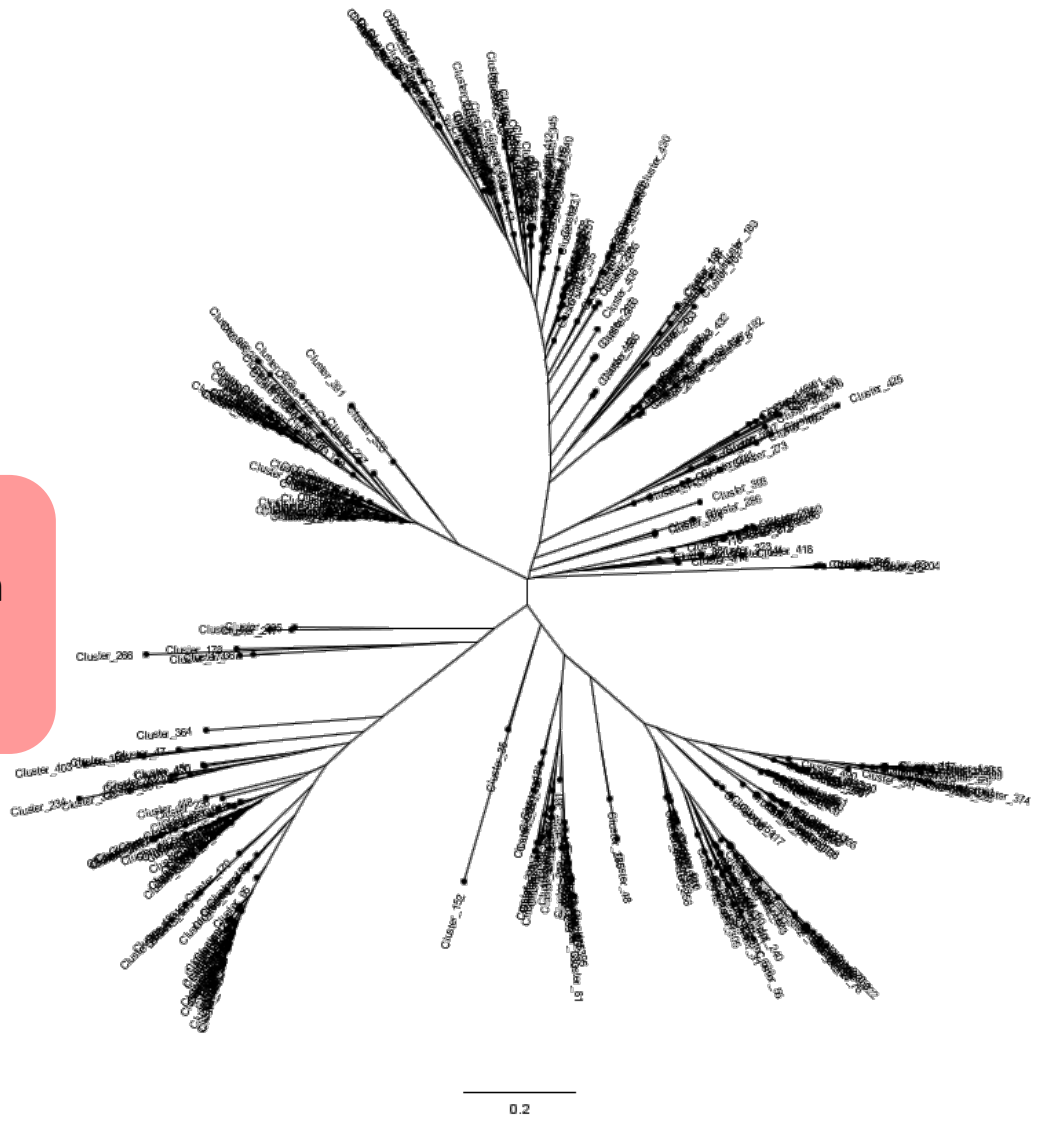


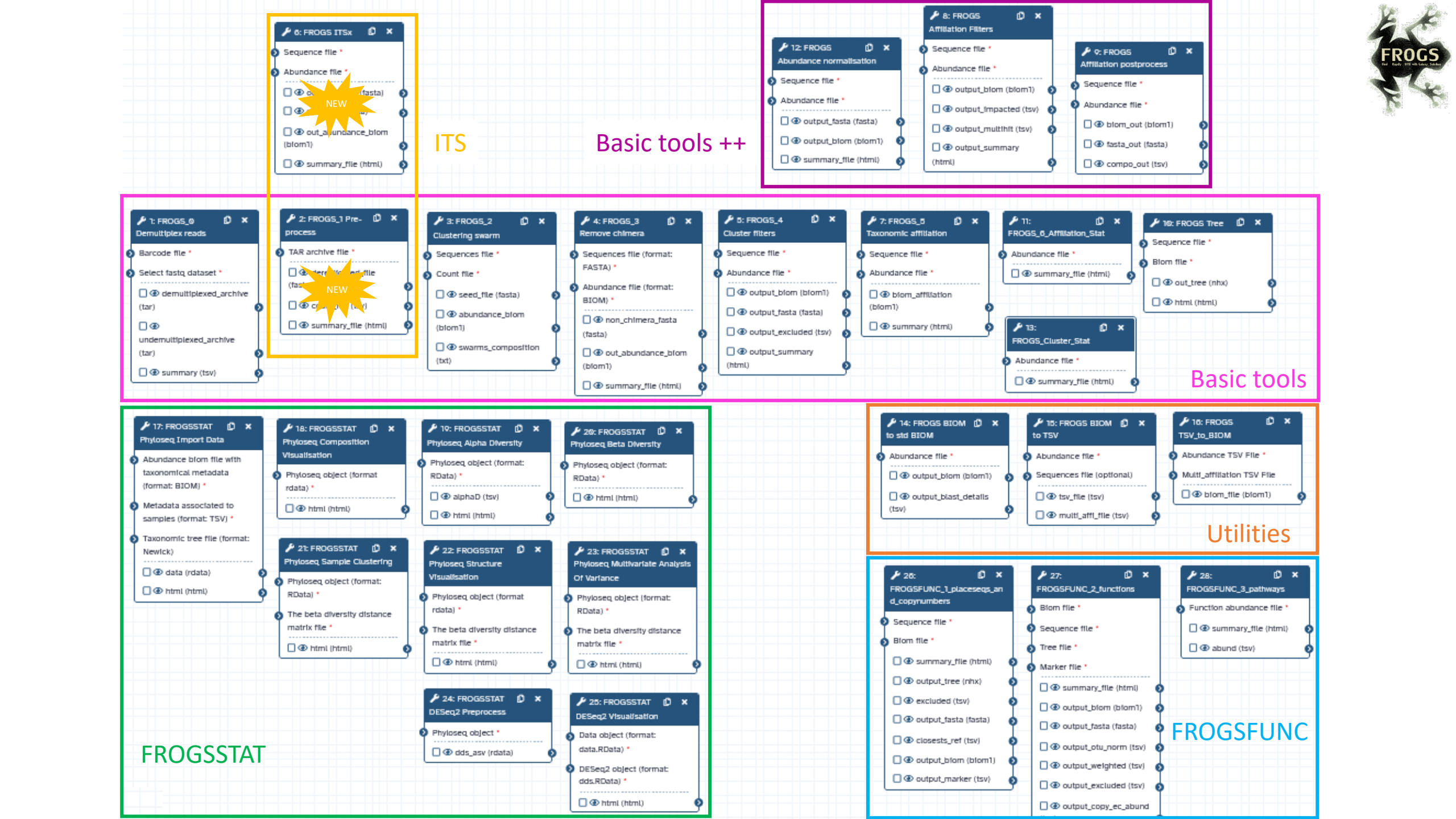
(dog:20, (elephant:30, horse:60):20):50

Our tree in nhx (= nwk) format

```
((((((((((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
```

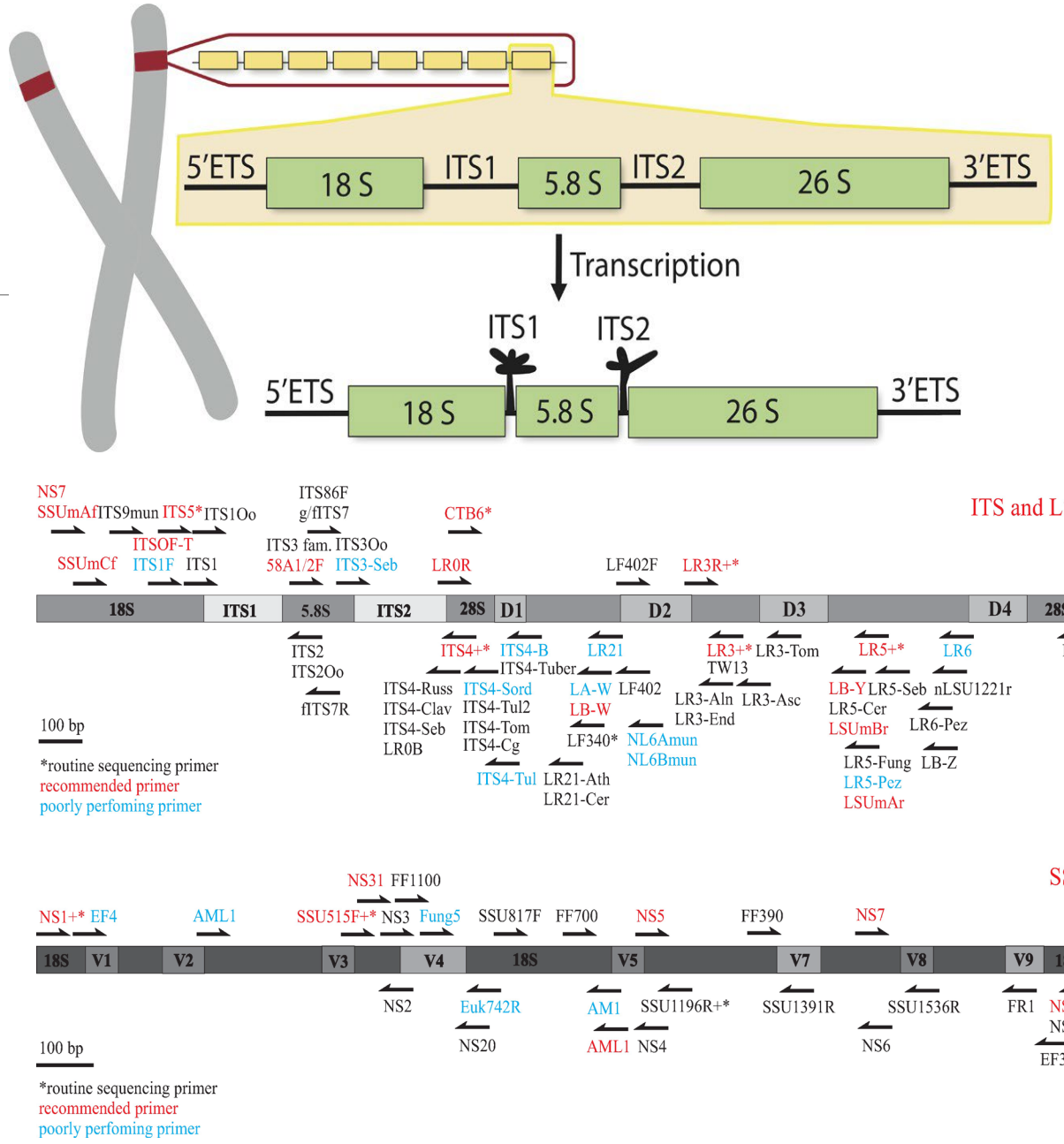
Exemple of visualization in FigTree from nhx file




6: FROGS ITSx
Sequence file *
Abundance file *
NEW
out_abundance_biom (biom1)
summary_file (html)
ITS
Basic tools ++
12: FROGS Abundance normalisation
Sequence file *
Abundance file *
output_fasta (fasta)
output_biom (biom1)
summary_file (html)
8: FROGS Affiliation Filters
Sequence file *
Abundance file *
output_biom (biom1)
output_impacted (tsv)
output_multihit (tsv)
output_summary (html)
9: FROGS Affiliation postprocess
Sequence file *
Abundance file *
biom_out (biom1)
fasta_out (fasta)
compo_out (tsv)
1: FROGS_0 Demultiplex reads
Barcode file *
Select fastq dataset *
demultiplexed_archive (tar)
undemultiplexed_archive (tar)
summary (tsv)
2: FROGS_1 Pre-process
TAR archive file *
NEW
summary_file (html)
3: FROGS_2 Clustering swarm
Sequences file *
Count file *
seed_file (fasta)
abundance_biom (biom1)
swarms_composition (txt)
4: FROGS_3 Remove chimera
Sequences file (format: FASTA) *
Abundance file (format: BIOM) *
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
summary_file (html)
5: FROGS_4 Cluster filters
Sequence file *
Abundance file *
output_biom (biom1)
output_fasta (fasta)
output_excluded (tsv)
output_summary (html)
7: FROGS_5 Taxonomic affiliation
Sequence file *
Abundance file *
biom_affiliation (biom1)
summary (html)
11: FROGS_6_Affiliation_Stat
Abundance file *
summary_file (html)
13: FROGS_Cluster_Stat
Abundance file *
summary_file (html)
10: FROGS Tree
Sequence file *
Biom file *
out_tree (nhx)
html (html)
Basic tools
17: FROGSSTAT Phyloseq Import Data
Abundance biom file with taxonomical metadata (format: BIOM) *
Metadata associated to samples (format: TSV) *
Taxonomic tree file (format: Newick)
data (rdata)
html (html)
18: FROGSSTAT Phyloseq Composition Visualisation
Phyloseq object (format rdata) *
html (html)
19: FROGSSTAT Phyloseq Alpha Diversity
Phyloseq object (format: RData) *
alphaD (tsv)
html (html)
20: FROGSSTAT Phyloseq Beta Diversity
Phyloseq object (format: RData) *
html (html)
21: FROGSSTAT Phyloseq Sample Clustering
Phyloseq object (format: RData) *
The beta diversity distance matrix file *
html (html)
22: FROGSSTAT Phyloseq Structure Visualisation
Phyloseq object (format rdata) *
The beta diversity distance matrix file *
html (html)
23: FROGSSTAT Phyloseq Multivariate Analysis Of Variance
Phyloseq object (format: RData) *
The beta diversity distance matrix file *
html (html)
24: FROGSSTAT DESeq2 Preprocess
Phyloseq object *
dds_asv (rdata)
25: FROGSSTAT DESeq2 Visualisation
Data object (format: data.RData) *
DESeq2 object (format: dds.RData) *
html (html)
FROGSSTAT
14: FROGS BIOM to std BIOM
Abundance file *
output_biom (biom1)
output_blast_details (tsv)
15: FROGS BIOM to TSV
Abundance file *
Sequences file (optional)
tsv_file (tsv)
multi_affi_file (tsv)
16: FROGS TSV_to_BIOM
Abundance TSV File *
Multi_affiliation TSV File
biom_file (biom1)
Utilities
26: FROGSFUNC_1_placeseqs_and_copynumbers
Sequence file *
Biom file *
summary_file (html)
output_tree (nhx)
excluded (tsv)
output_fasta (fasta)
closests_ref (tsv)
output_biom (biom1)
output_marker (tsv)
27: FROGSFUNC_2_functions
Biom file *
Sequence file *
Tree file *
Marker file *
summary_file (html)
output_biom (biom1)
output_fasta (fasta)
output_otu_norm (tsv)
output_weighted (tsv)
output_excluded (tsv)
output_copy_ec_abund
28: FROGSFUNC_3_pathways
Function abundance file *
summary_file (html)
abund (tsv)
FROGSFUNC

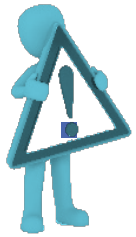
What is a ITS ?

ITS: Internal Transcribed Spacer



What is a ITS ?

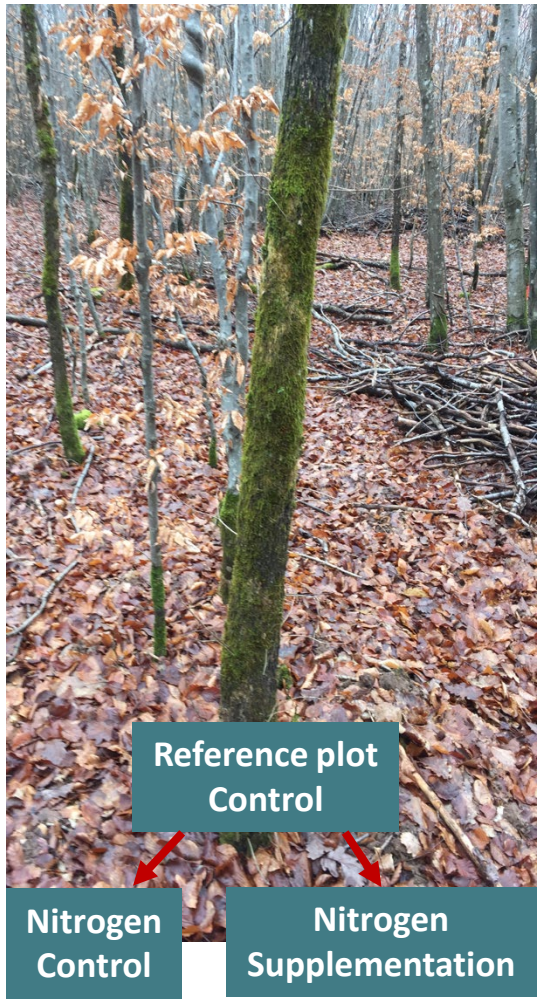
- Size polymorphism of ITS (from 361 to 1475 bases in UNITE 7.1)
- Highly conserved regions of the neighboring of ITS1 and ITS2
- Lack of a generalist and abundant ITS databank (several small specialized databanks)
- Multiple copies* (14 to 1400 copies (mean at 113, median at 80))
- Do not target Glomeromycetes/Glomeromycota (→ alternative: 18S)



- If your sequencing platform preprocesses your data, it has to keep short and long sequences

* https://doi.org/10.1111/mec.14995

ITS data from manipulated organic soil (MOS network)



While in the past forest biomass exports concerned only trunks, these exports recently increased and now concern also the branches and smaller parts that were previously left on the ground (for pellet production).

The MOS network (18 sites in France) was designed to reveal the long-term effects of intense biomass exports on soil fertility and biodiversity. Different treatment of biomass export are applied with or without supplementation of nutrients.

The aim is to analyse the **impact** of these new forestry practices on **soil microbiota** and **tree health**.

The present dataset concerned one of the site (Champenoux) after 5 years of total Organic Matter removal (OMR treatment : all the organic matter on the ground including leaves was removed), **with** our **without nitrogen supplementation**.

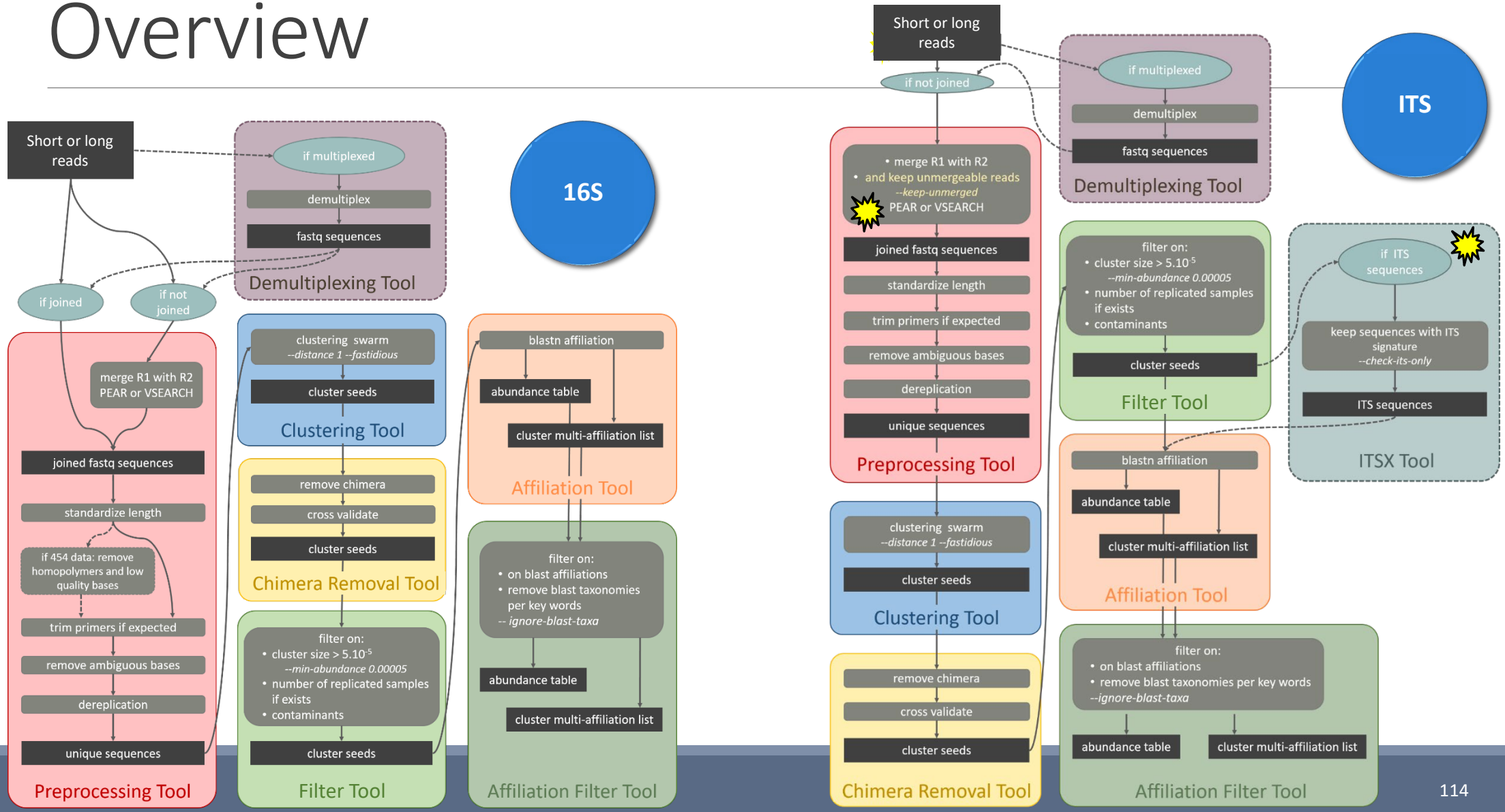
- 5 replicates Control x 2 treatments, 5 replicates OMR x 2 treatments
- DNA is extracted and **ITS1** is sequenced
- 2 x 250 bp Illumina MiSeq
- Primer 5': CTTGGTCATTTAGAGGAAGTAA
- Primer 3': GCATCGATGAAGAACGCAGC

Metadata for these samples

Samples	kept	Replicas	Incubation	Nitrogen	Forest_management	Quality	Treatment
Ph203	79.76	3	T4	Nitrogen_supplementation	Control	Low degradability	Control_with_N
Ph212	77.64	2	T4	Nitrogen_supplementation	Control	Low degradability	Control_with_N
Ph217	80.26	5	T4	Nitrogen_supplementation	Control	Low degradability	Control_with_N
Ph222	78.65	1	T4	Nitrogen_supplementation	Control	Low degradability	Control_with_N
Ph224	77.18	4	T4	Nitrogen_supplementation	Control	Low degradability	Control_with_N
Ph237	79.68	1	T4	Control	Control	Low degradability	Control
Ph241	78.7	2	T4	Control	Control	Low degradability	Control
Ph243	76.38	4	T4	Control	Control	Low degradability	Control
Ph246	76.37	5	T4	Control	Control	Low degradability	Control
Ph250	77.37	3	T4	Control	Control	Low degradability	Control
Ph407	72.52	3	T4	Nitrogen_supplementation	OMR	Low degradability	OMR_with_N
Ph414	64.98	4	T4	Nitrogen_supplementation	OMR	Low degradability	OMR_with_N
Ph415	78.13	2	T4	Nitrogen_supplementation	OMR	Low degradability	OMR_with_N
Ph417	71.17	1	T4	Nitrogen_supplementation	OMR	Low degradability	OMR_with_N
Ph423	75.2	5	T4	Nitrogen_supplementation	OMR	Low degradability	OMR_with_N
Ph428	73.48	2	T4	Control	OMR	Low degradability	OMR
Ph433	73.21	5	T4	Control	OMR	Low degradability	OMR
Ph434	74.01	3	T4	Control	OMR	Low degradability	OMR
Ph439	74.15	1	T4	Control	OMR	Low degradability	OMR
Ph449	73.77	4	T4	Control	OMR	Low degradability	OMR

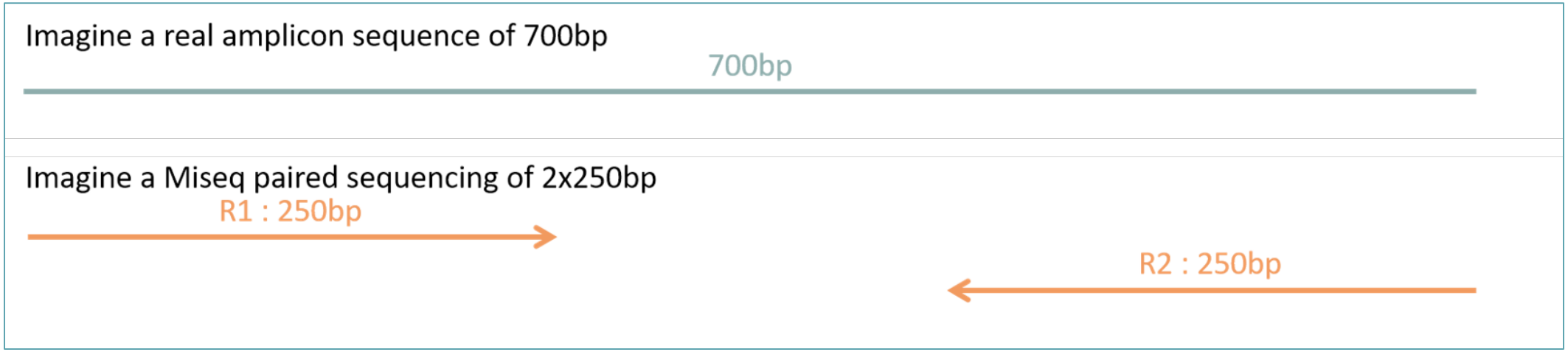
: Differences between 16S and ITS pipelines

Overview



FROGS Pipeline for ITS

Problematic:
some ITS reads (Miseq sequencing) are non-overlapping sequences



Consequence: during bioinformatics process, these reads are lost and underlying organisms will be never represented in the abundance table.

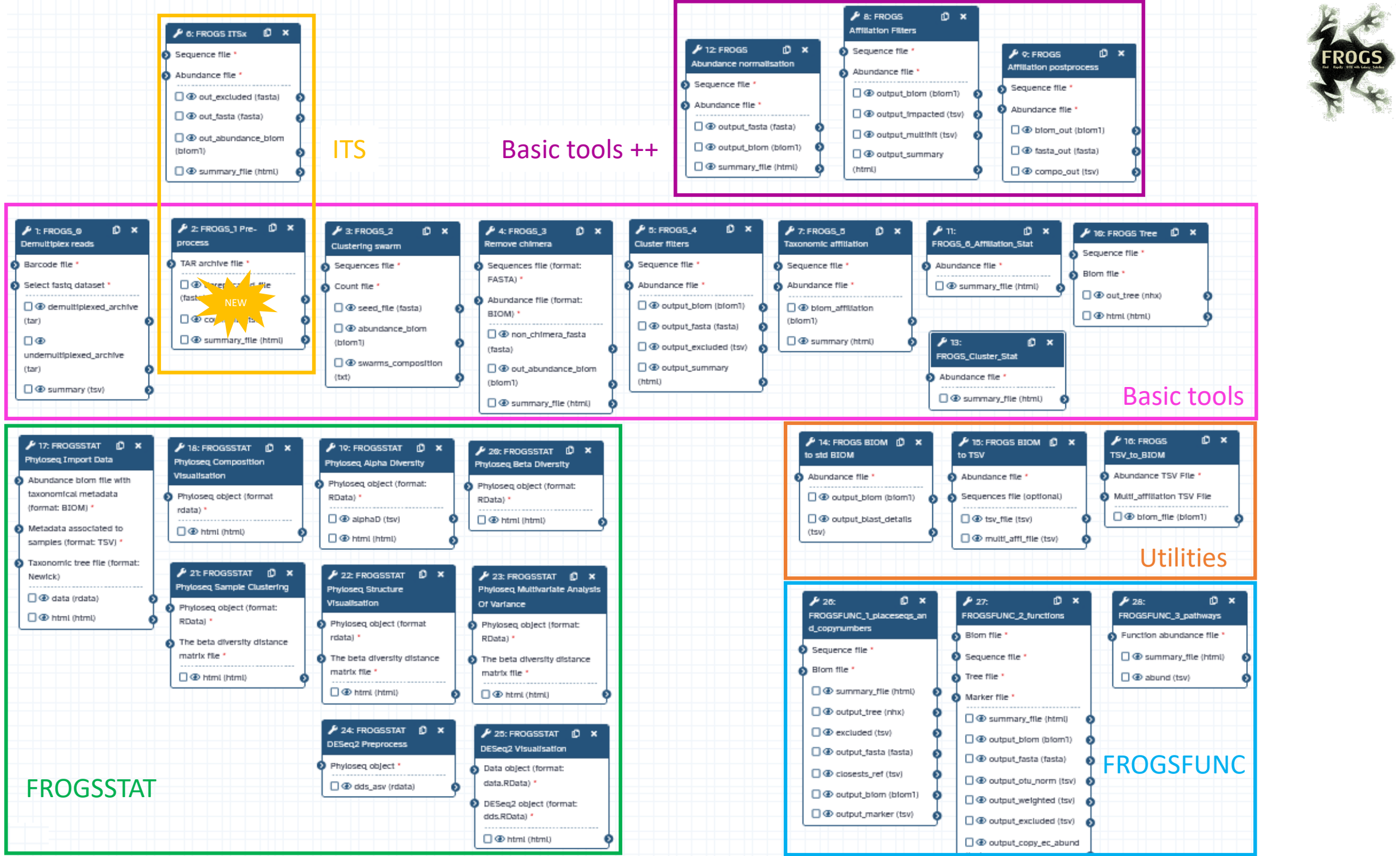
Solution: in preprocess step – creation of “FROGS combined” sequences

Imagine a real amplicon sequence of 700bp

Imagine a Miseq paired sequencing of 2x250bp

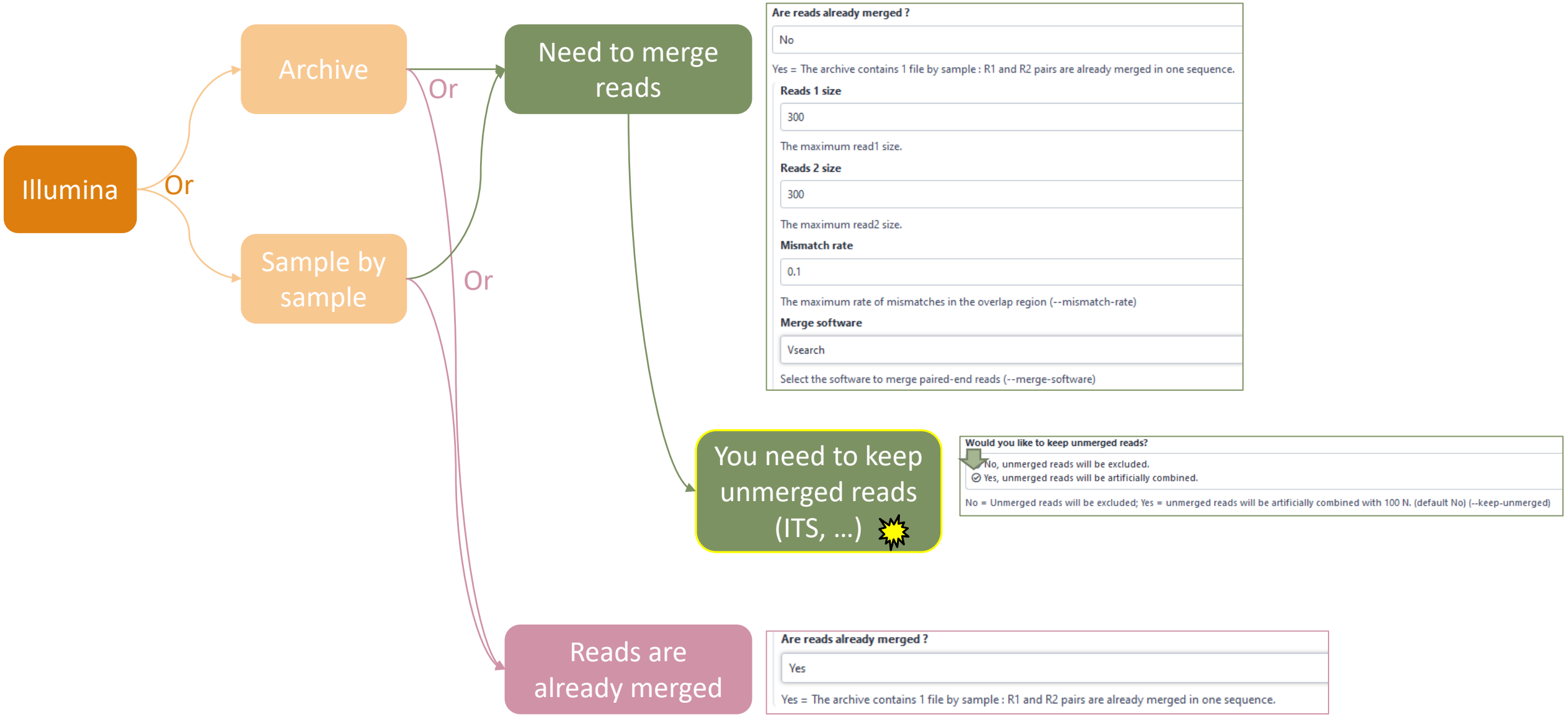
Reconstructing amplicon sequence is not possible with overlap, an arbitrary sequence of 100Ns is added. It is named « FROGS combined »

Combined sequence length : 600bp, with 100 Ns


6: FROGS ITSx
Sequence file *
Abundance file *
out_excluded (fasta)
out_fasta (fasta)
out_abundance_biom (biom1)
summary_file (html)
ITS
Basic tools ++
12: FROGS Abundance normalisation
Sequence file *
Abundance file *
output_fasta (fasta)
output_biom (biom1)
summary_file (html)
8: FROGS Affiliation filters
Sequence file *
Abundance file *
output_biom (biom1)
output_impacted (tsv)
output_multihit (tsv)
output_summary (html)
9: FROGS Affiliation postprocess
Sequence file *
Abundance file *
biom_out (biom1)
fasta_out (fasta)
compo_out (tsv)
1: FROGS_0 Demultiplex reads
Barcode file *
Select fastq dataset *
demultiplexed_archive (tar)
undemultiplexed_archive (tar)
summary (tsv)
2: FROGS_1 Pre-process
TAR archive file *
NEW
summary_file (html)
3: FROGS_2 Clustering swarm
Sequences file *
Count file *
seed_file (fasta)
abundance_biom (biom1)
swarms_composition (txt)
4: FROGS_3 Remove chimera
Sequences file (format: FASTA) *
Abundance file (format: BIOM) *
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
summary_file (html)
5: FROGS_4 Cluster filters
Sequence file *
Abundance file *
output_biom (biom1)
output_fasta (fasta)
output_excluded (tsv)
output_summary (html)
7: FROGS_5 Taxonomic affiliation
Sequence file *
Abundance file *
biom_affiliation (biom1)
summary (html)
11: FROGS_6_Affiliation_Stat
Abundance file *
summary_file (html)
13: FROGS_Cluster_Stat
Abundance file *
summary_file (html)
10: FROGS Tree
Sequence file *
Biom file *
out_tree (nhx)
html (html)
Basic tools
17: FROGSSTAT Phyloseq Import Data
Abundance biom file with taxonomical metadata (format: BIOM) *
Metadata associated to samples (format: TSV) *
Taxonomic tree file (format: Newick)
data (rdata)
html (html)
18: FROGSSTAT Phyloseq Composition Visualisation
Phyloseq object (format rdata) *
html (html)
19: FROGSSTAT Phyloseq Alpha Diversity
Phyloseq object (format: RData) *
alphaD (tsv)
html (html)
20: FROGSSTAT Phyloseq Beta Diversity
Phyloseq object (format: RData) *
html (html)
21: FROGSSTAT Phyloseq Sample Clustering
Phyloseq object (format: RData) *
The beta diversity distance matrix file *
html (html)
22: FROGSSTAT Phyloseq Structure Visualisation
Phyloseq object (format rdata) *
The beta diversity distance matrix file *
html (html)
23: FROGSSTAT Phyloseq Multivariate Analysis Of Variance
Phyloseq object (format: RData) *
The beta diversity distance matrix file *
html (html)
24: FROGSSTAT DESeq2 Preprocess
Phyloseq object *
dds_asv (rdata)
25: FROGSSTAT DESeq2 Visualisation
Data object (format: data.RData) *
DESeq2 object (format: dds.RData) *
html (html)
FROGSSTAT
14: FROGS BIOM to std BIOM
Abundance file *
output_biom (biom1)
output_blast_details (tsv)
15: FROGS BIOM to TSV
Abundance file *
Sequences file (optional)
tsv_file (tsv)
multi_affi_file (tsv)
16: FROGS TSV_to_BIOM
Abundance TSV File *
Multi_affiliation TSV File
biom_file (biom1)
Utilities
26: FROGSFUNC_1_placeseqs_and_copynumbers
Sequence file *
Biom file *
summary_file (html)
output_tree (nhx)
excluded (tsv)
output_fasta (fasta)
closests_ref (tsv)
output_biom (biom1)
output_marker (tsv)
27: FROGSFUNC_2_functions
Biom file *
Sequence file *
Tree file *
Marker file *
summary_file (html)
output_biom (biom1)
output_fasta (fasta)
output_otu_norm (tsv)
output_weighted (tsv)
output_excluded (tsv)
output_copy_ec_abund
28: FROGSFUNC_3_pathways
Function abundance file *
summary_file (html)
abund (tsv)
FROGSFUNC

Pre-process tool

For short reads from illumina



ITS

FROGS_1 Pre-process merging, denoising and dereplication (Galaxy Version 4.1.0+galaxy1)

Sequencer

Illumina

Select the sequencing technology used to produce the sequences.

Input type

TAR Archive

Samples files can be provided in a single TAR archive or sample by sample (with one or two files each).

TAR archive file

1: ITS_fast.tar.gz

The TAR file containing the sequences file(s) for each sample.

Are reads already merged ?

No

Yes = The archive contains 1 file by sample : R1 and R2 pairs are already merged in one sequence.

Reads 1 size

250

The maximum read1 size.

Reads 2 size

250

The maximum read2 size.

Mismatch rate

0.1

The maximum rate of mismatches in the overlap region (--mismatch-rate)

Merge software

Vsearch

Select the software to merge paired-end reads (--merge-software)

Would you like to keep unmerged reads?

- ○ No, unmerged reads will be excluded.
- ⊘ Yes, unmerged reads will be artificially combined.

To keep FROGS combined sequences, choose YES

No = Unmerged reads will be excluded; Yes = unmerged reads will be artificially combined with 100 N. (default No) (--keep-unmerged)

ITS

Minimum amplicon size

180

The minimum size of the amplicons (with primers) (--min-amplicon-size)

Maximum amplicon size

490

The maximum size of the amplicons (with primers) (--max-amplicon-size)

Do the sequences have PCR primers?

- (✓) Yes
- ( ) No

5' primer

CTTGGTCATTTAGAGGAAGTAA

The 5' primer sequence (wildcards are accepted). This primer must be written in 5' to 3' orientation (see details in 'Primers parameters' help section) (--five-prim-primer)

3' primer

GCATCGATGAAGAACGCAGC

The 3' primer sequence (wildcards are accepted). This primer must be written in 5' to 3' orientation (see details in 'Primers parameters' help section) (--three-prim-primer)

Primer 5': CTTGGTCATTTAGAGGAAGTAA
Primer 3': GCATCGATGAAGAACGCAGC

Exercise

Go to « ITS » history

Launch the FROGS_1 pre-process tool on this data set

→ objective: understand preprocess report and « FROGS combined sequences »

Launch the FROGS_2 Clustering swarm tool

Launch the FROGS_3 Remove chimera tool

Launch the FROGS_4 Cluster filter tool

FROGS_1 Pre-process

Tool Parameters

Input Parameter	Value
Sequencer	illumina
Input type	archive
TAR archive file	1 : ITS_fast.tar.gz
Are reads already merged ?	paired
Reads 1 size	250
Reads 2 size	250
Mismatch rate	0.1
Merge software	vsearch
Would you like to keep unmerged reads?	Yes, unmerged reads will be artificially combined.
Minimum amplicon size	180
Maximum amplicon size	490
Do the sequences have PCR primers?	true
5' primer	CTTGGTCATTTAGAGGAAGTAA
3' primer	GCATCGATGAAGAACGCAGC

FROGS_2 Clustering swarm

Tool Parameters

Input Parameter	Value
Sequences file	**4 : FROGS_1 Pre-process: dereplicated.fasta**
Count file	**5 : FROGS_1 Pre-process: count.tsv**
FROGS guidelines version	3.2
Aggregation distance clustering	1
Refine clustering	Yes, refine clustering with --fastidious swarm option

FROGS_3 Remove chimera

Tool Parameters

Input Parameter	Value
Sequences file (format: FASTA)	**7 : FROGS_2 Clustering swarm: seed_sequences.fasta**
Abundance type	biom
Abundance file (format: BIOM)	**8 : FROGS_2 Clustering swarm: clustering_abundance.biom**

FROGS_4 Cluster filters

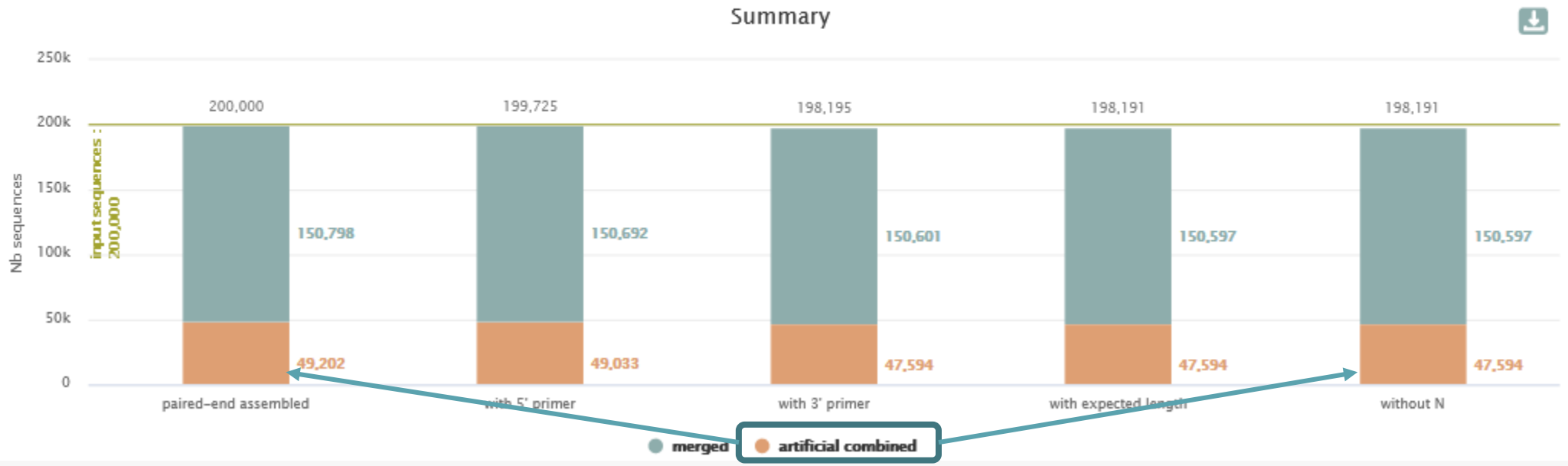
Tool Parameters

Input Parameter	Value
Sequence file	**11 : FROGS_3 Remove chimera: non_chimera.fasta**
Abundance file	**12 : FROGS_3 Remove chimera: non_chimera_abundance.biom**
Minimum prevalence method	replicate
File of replicated sample names	**3 : ITS_fast_replicates.tsv**
Minimum prevalence	0.5
Minimum cluster abundancy as proportion or count. We recommend to use a proportion of 0.00005.	proportion
Minimum proportion of sequences abundancy to keep cluster	5e-05
N biggest clusters	Not available.
Search for contaminant clusters.	server
Contaminant databank	phiX

ITS

FROGS_1 Pre-process

Preprocess summary



FROGS_1 Pre-process

2 tables:

on merged sequences

CSV

Show 10 entries Search:

Samples	before process	% kept	paired-end assembled	with 5' primer	with 3' primer	with expected length	without N
Ph203	10,000	79.42	7,954	7,948	7,942	7,942	7,942
Ph212	10,000	78.28	7,837	7,832	7,828	7,828	7,828
Ph217	10,000	80.48	8,061	8,052	8,048	8,048	8,048
Ph222	10,000	78.34	7,839	7,835	7,834	7,834	7,834

Details on artificial combined sequences

CSV

Show 10 entries Search:

Samples	before process	% kept	paired-end assembled	with 5' primer	with 3' primer	with expected length	without N
Ph203	10,000	19.68	2,046	2,038	1,968	1,968	1,968
Ph212	10,000	20.65	2,163	2,154	2,065	2,065	2,065
Ph217	10,000	18.63	1,939	1,928	1,863	1,863	1,863
Ph222	10,000	20.79	2,161	2,155	2,079	2,079	2,079

Own tag for combined sequences

```
>M01328:521:000000000-KRPTR:1:1103:15714:11240;size=6 1:N:0:238
AAGTCGTAACAAGGTAACCGTAGGTGAACCTGCGGTTGGATCATTAAAAATTTATGAGTTTCCGTTGAG
>M01328:521:000000000-KRPTR:1:2102:7650:15129;size=1 1:N:0:239
AAGTCGTAACAAGGTAACCGTAGGTGAACCTGCGGTTGGATCATTAAAAATTTATGAGTTTCCGTTGAG
>M01328:521:000000000-KRPTR:1:1112:8680:15899;size=1 1:N:0:202
>M01328:521:000000000-KRPTR:1:1111:21036:16514_FROGS_combined;size=1
>M01328:521:000000000-KRPTR:1:1106:19343:17084_FROGS_combined;size=1
```

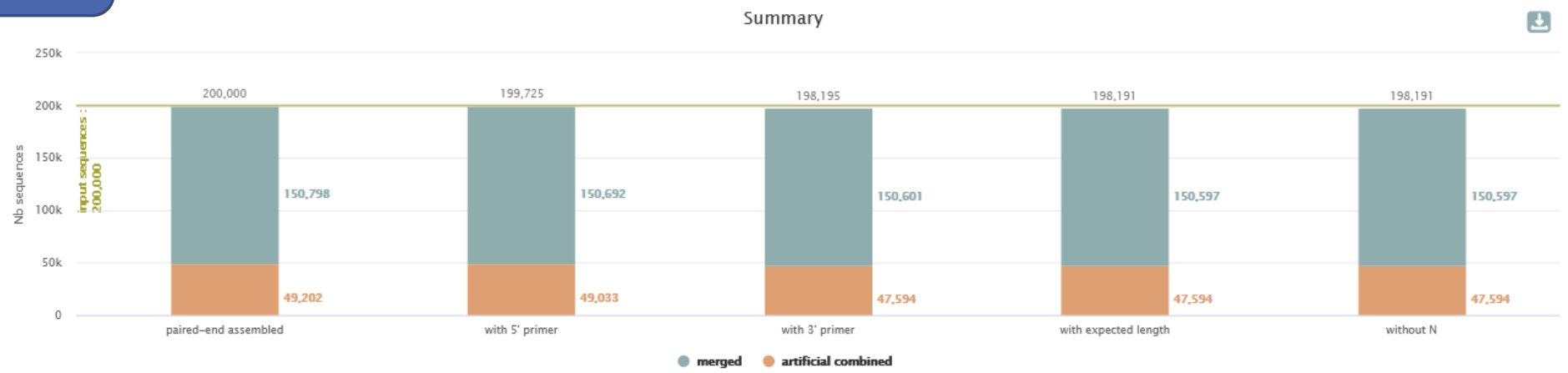
Filter only on minimum length for « combined ».

Minimum length =
R1 + 100N + R2 – primers sizes

If the primers are very internal to the read, after trimming them, the combined sequence could be smaller than a read. FROGS rejects these cases.

FROGS_1 Pre-process

ess summary



A large quantity of artificial combined

Details on merged sequences

Show 10 entries

CSV

Search:

Samples	before process	% kept	paired-end assembled	with 5' primer	with 3' primer	with expected length	without N
Ph414	10,000	64.01	6,408	6,405	6,401	6,401	
Ph417	10,000	69.83	6,990	6,987	6,984	6,983	
Ph407	10,000	71.81	7,188	7,187	7,181	7,181	7,181

Samples with >30% unmerged

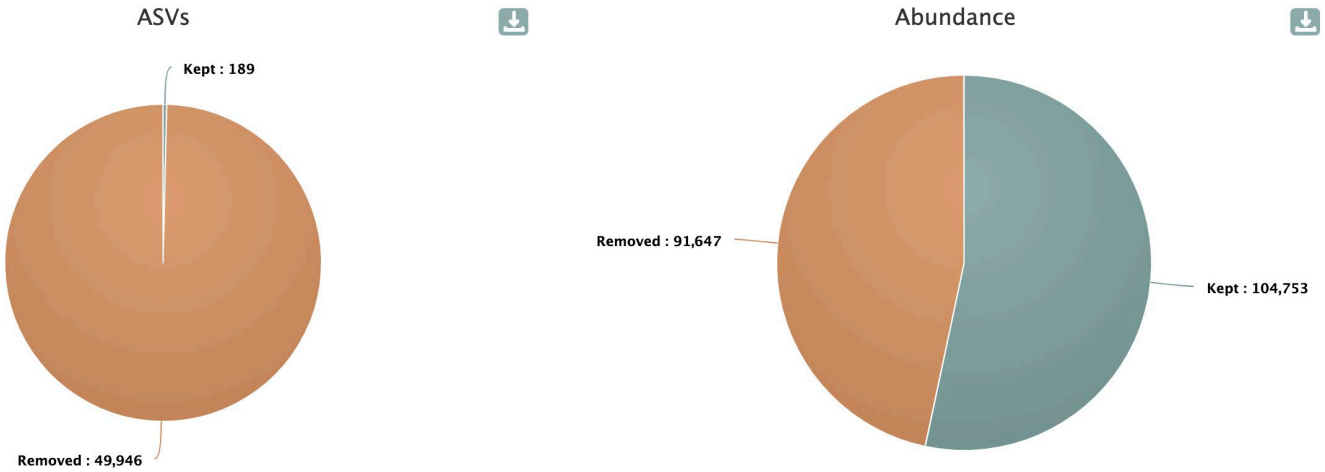
Samples	before process	% kept	paired-end assembled	with 5' primer	with 3' primer	with expected length	without N
Ph217	10,000	80.48	8,061	8,052	8,048	8,048	
Ph237	10,000	79.59	7,967	7,964	7,959	7,959	7,959
Ph203	10,000	79.42	7,954	7,948	7,942	7,942	7,942

and others at ~20%.

Filtres à 50% de prévalence par groupe de réplicats :

FROGS_4 Cluster filters

Filters summary



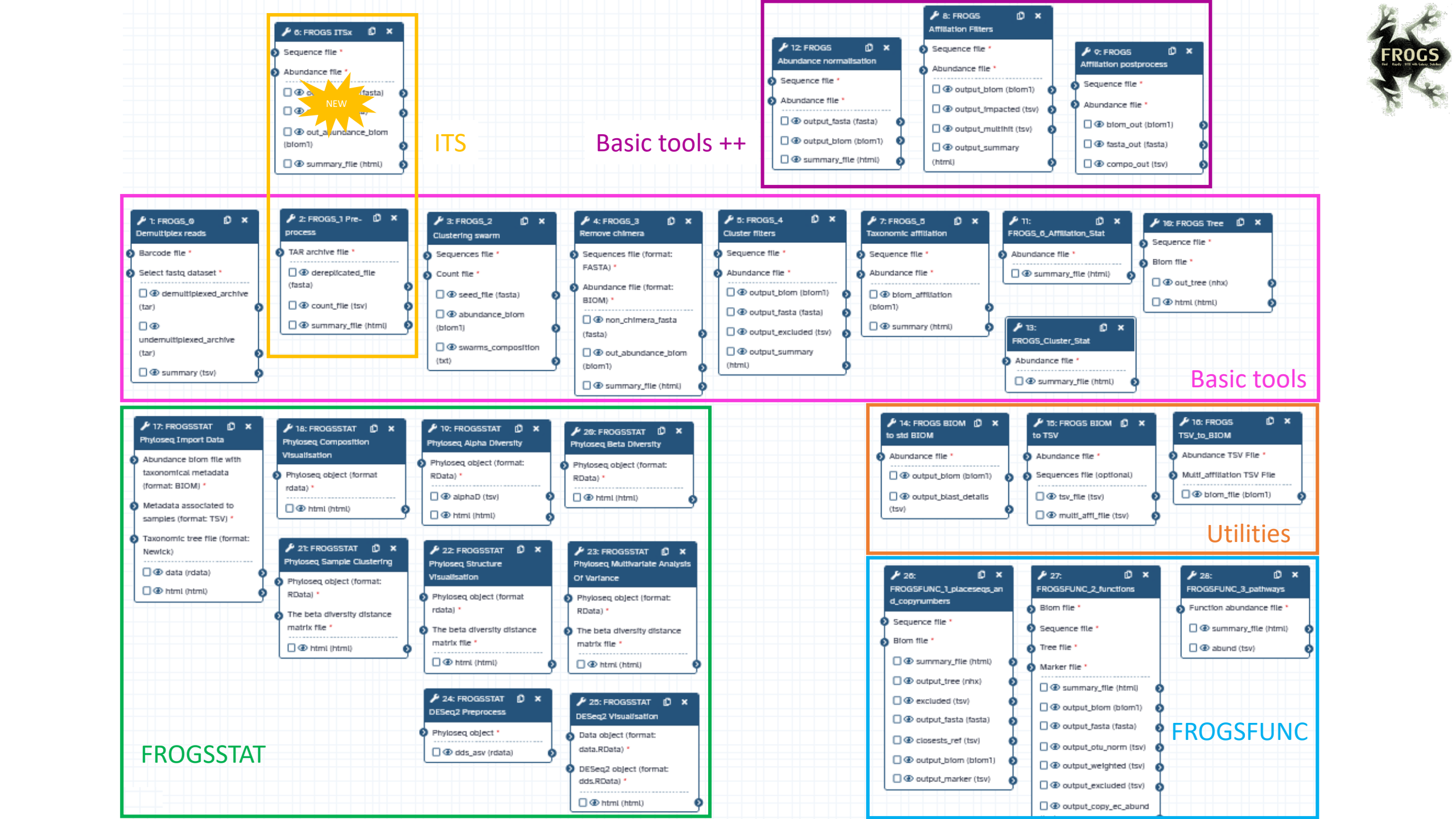
Grosse perte de séquences, filtre à ajuster ou modifier?

Filters intersections

Draw a Venn to see which ASVs had been deleted by the filters chosen (Maximum 6 options):

- Present in less than 50.0% of replicates of all replicate groups.
- Abundance < 0.005% (i.e 10 sequences)
- Present in databank of contaminants

Venn

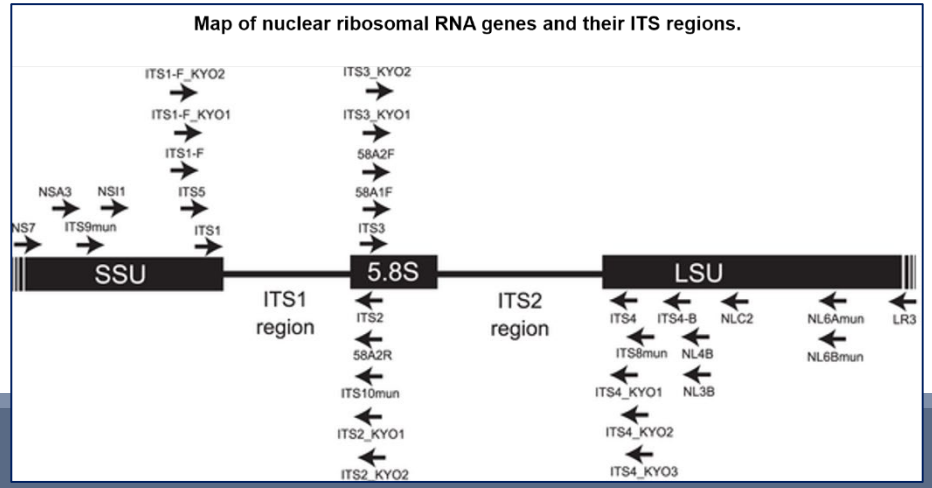

6: FROGS ITSx
Sequence file *
Abundance file *
out_abundance_biom (biom1)
summary_file (html)
NEW
ITS
Basic tools ++
12: FROGS Abundance normalisation
Sequence file *
Abundance file *
output_fasta (fasta)
output_biom (biom1)
summary_file (html)
8: FROGS Affiliation filters
Sequence file *
Abundance file *
output_biom (biom1)
output_impacted (tsv)
output_multihit (tsv)
output_summary (html)
9: FROGS Affiliation postprocess
Sequence file *
Abundance file *
biom_out (biom1)
fasta_out (fasta)
compo_out (tsv)
1: FROGS_0 Demultiplex reads
Barcode file *
Select fastq dataset *
demultiplexed_archive (tar)
undemultiplexed_archive (tar)
summary (tsv)
2: FROGS_1 Pre-process
TAR archive file *
dereplicated_file (fasta)
count_file (tsv)
summary_file (html)
3: FROGS_2 Clustering swarm
Sequences file *
Count file *
seed_file (fasta)
abundance_biom (biom1)
swarms_composition (txt)
4: FROGS_3 Remove chimera
Sequences file (format: FASTA) *
Abundance file (format: BIOM) *
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
summary_file (html)
5: FROGS_4 Cluster filters
Sequence file *
Abundance file *
output_biom (biom1)
output_fasta (fasta)
output_excluded (tsv)
output_summary (html)
7: FROGS_5 Taxonomic affiliation
Sequence file *
Abundance file *
biom_affiliation (biom1)
summary (html)
11: FROGS_6_Affiliation_Stat
Abundance file *
summary_file (html)
13: FROGS_Cluster_Stat
Abundance file *
summary_file (html)
10: FROGS Tree
Sequence file *
Biom file *
out_tree (nhx)
html (html)
Basic tools
17: FROGSSTAT Phyloseq Import Data
Abundance biom file with taxonomical metadata (format: BIOM) *
Metadata associated to samples (format: TSV) *
Taxonomic tree file (format: Newick)
data (rdata)
html (html)
18: FROGSSTAT Phyloseq Composition Visualisation
Phyloseq object (format rdata) *
html (html)
19: FROGSSTAT Phyloseq Alpha Diversity
Phyloseq object (format: RData) *
alphaD (tsv)
html (html)
20: FROGSSTAT Phyloseq Beta Diversity
Phyloseq object (format: RData) *
html (html)
21: FROGSSTAT Phyloseq Sample Clustering
Phyloseq object (format: RData) *
The beta diversity distance matrix file *
html (html)
22: FROGSSTAT Phyloseq Structure Visualisation
Phyloseq object (format rdata) *
The beta diversity distance matrix file *
html (html)
23: FROGSSTAT Phyloseq Multivariate Analysis Of Variance
Phyloseq object (format: RData) *
The beta diversity distance matrix file *
html (html)
24: FROGSSTAT DESeq2 Preprocess
Phyloseq object *
dds_asv (rdata)
25: FROGSSTAT DESeq2 Visualisation
Data object (format: data.RData) *
DESeq2 object (format: dds.RData) *
html (html)
FROGSSTAT
14: FROGS BIOM to std BIOM
Abundance file *
output_biom (biom1)
output_blast_details (tsv)
15: FROGS BIOM to TSV
Abundance file *
Sequences file (optional)
tsv_file (tsv)
multi_affi_file (tsv)
16: FROGS TSV_to_BIOM
Abundance TSV File *
Multi_affiliation TSV File
biom_file (biom1)
Utilities
26: FROGSFUNC_1_placeseqs_and_copynumbers
Sequence file *
Biom file *
summary_file (html)
output_tree (nhx)
excluded (tsv)
output_fasta (fasta)
closests_ref (tsv)
output_biom (biom1)
output_marker (tsv)
27: FROGSFUNC_2_functions
Biom file *
Sequence file *
Tree file *
Marker file *
summary_file (html)
output_biom (biom1)
output_fasta (fasta)
output_otu_norm (tsv)
output_weighted (tsv)
output_excluded (tsv)
output_copy_ec_abund
28: FROGSFUNC_3_pathways
Function abundance file *
summary_file (html)
abund (tsv)
FROGSFUNC

ITSx tools

What is the purpose of the ITSx tool?

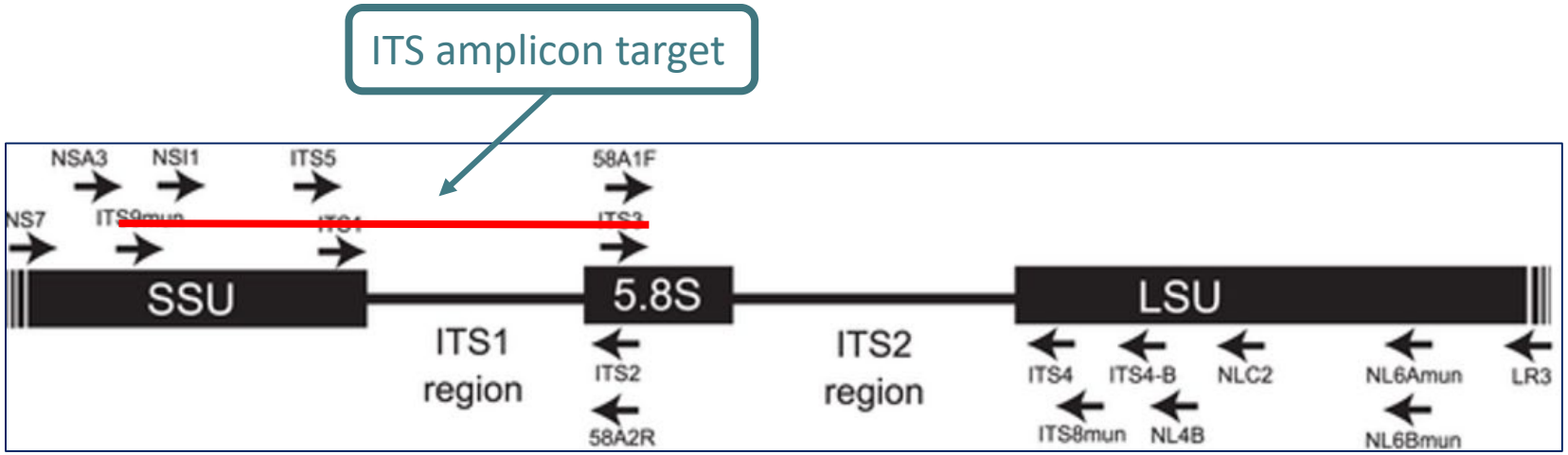
- ITSx is a tool to **filter** sequences.
- ITSx **identifies** and **trimms** ITS regions in sequences.
- It **excludes** the highly conserved neighboring sequences **SSU**, **5S** and **LSU** rRNA.
- If the ITS1 or ITS2 region is not detected, the sequence is discarded.
- You can choose to check only if the sequence is detected as an ITS.

  In this case, the sequence is not trimmed, only sequences not detected as ITS are rejected (*e.g.* contaminants).



Bengtsson-Palme, J., et al. (2013), Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. Methods Ecol Evol, 4: 914-919. https://doi.org/10.1111/2041-210X.12073

What is the purpose of the ITSx tool?

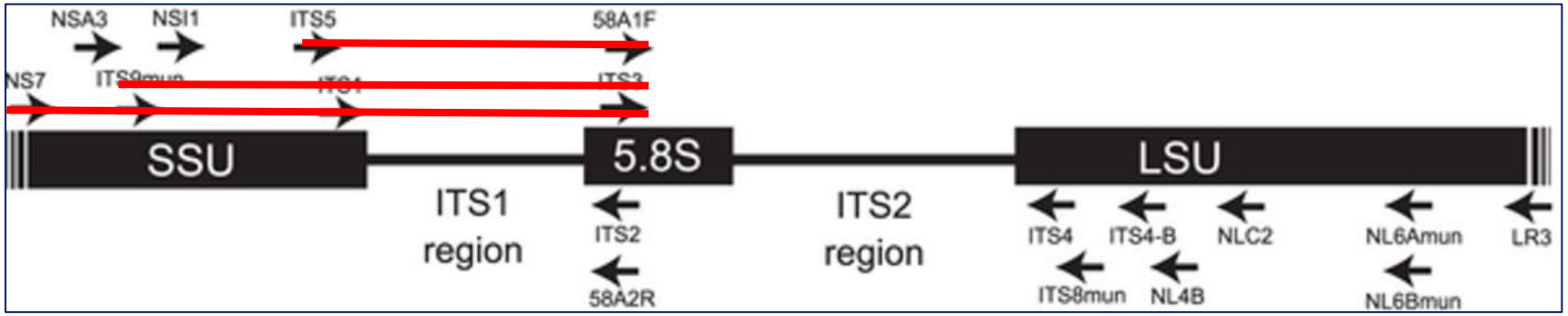


1st case: choose to trim
ITS1 is well detected
SSU part and 5.8S part are trimmed
Result: ——

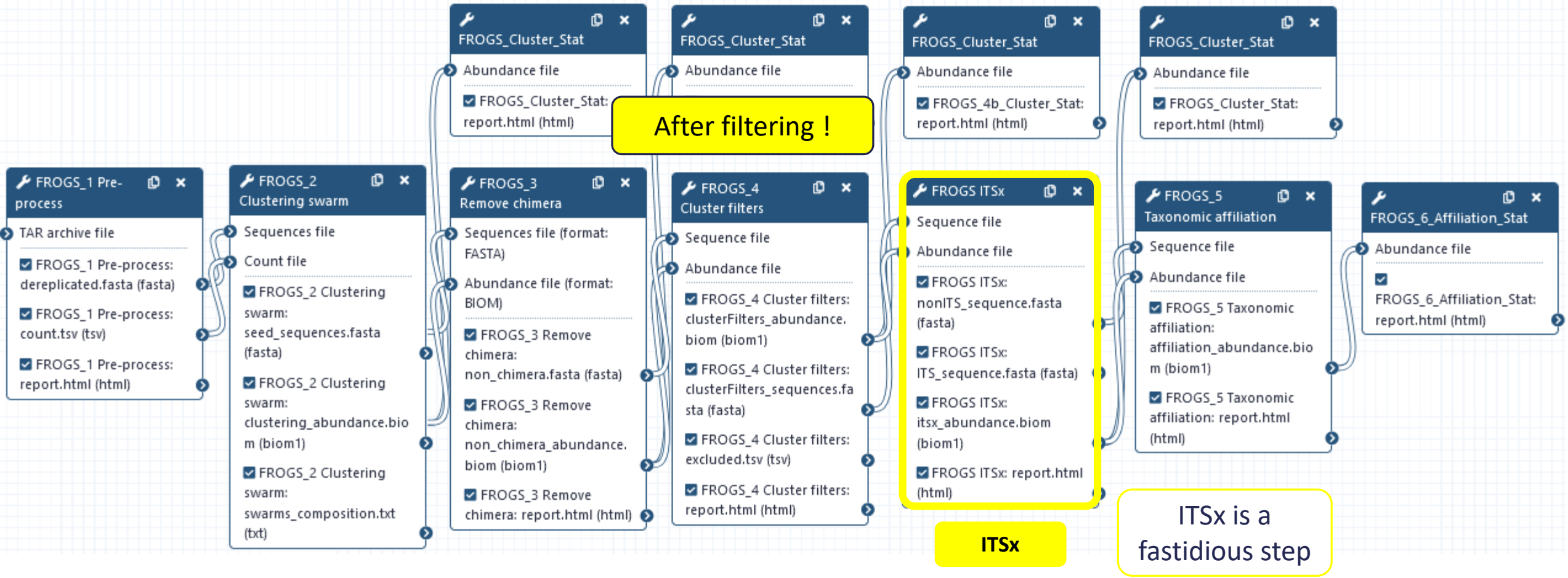
2nd case: choose to check only
ITS1 is well detected
SSU part and 5.8S part are not trimmed
Result: ——

Check only if sequence is detected as ITS? Yes or not?

- If not, only ITS1 or ITS2 part will be conserved
- This is interesting to keep only the ITS parts without the flanking sequences in case of :
  - comparison of sequenced amplicons with different primers targeting the same region to be amplified.
  - using a database with only ITS part



When should we use ITSx ?



FROGS ITSx Extract the highly variable ITS1 and ITS2 subregions from ITS sequences (Galaxy Version 4.1.0+galaxy1)

Sequence file

16: FROGS_4 Cluster filters: clusterFilters_sequences.fasta

The sequence file to filter (format: FASTA).

Abundance file

15: FROGS_4 Cluster filters: clusterFilters_abundance.biom

The abundance file to filter (format: BIOM)

Trim conserved sequence (SSU, 5.8S, LSU) ?

- (x) No, keep conserved regions
- ( ) Yes, trim conserved regions

If Yes, only part of the sequences with ITS signature will be kept, SSU, LSU or 5.8S regions will be trimmed (default : No) (--check-its-only)

Choose pertinent organisms to scan:

Select/Unselect all

- [x] Fungi
- [ ] Alveolata
- [ ] Bryophyta
- [ ] Bacillariophyta
- [ ] Amoebozoa
- [ ] Euglenozoa
- [ ] Chlorophyta
- [ ] Rhodophyta
- [ ] Phaeophyceae
- [ ] Marchantiophyta
- [ ] Metazoa
- [ ] Oomycota
- [ ] Haptophyceae
- [ ] Raphidophyceae
- [ ] Rhizaria
- [ ] Synurophyceae
- [ ] Tracheophyta
- [ ] Eustigmatophyceae

Save a lot of time by checking pertinent organism group model to scan (--organism-groups)

Email notification

No

Send an email notification when the job completes.

✔ Execute

By default, the ITSs are kept in their entirety.

By defaut, sequences are considered as FUNGI sequences.
Change it, if it is not the case.

Careful !

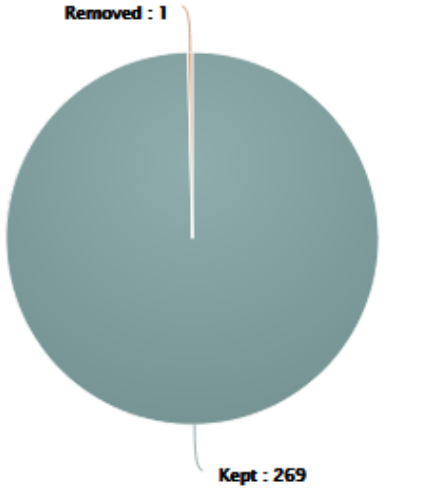
- The ITSx step is time consuming and has to be done on minimum of clusters.

1. Preprocess step,
2. Clustering step,
3. Chimera removing step,
4. Filter on ASVs abundances and replicates step,
5. ITSx

Report.html, ITSX output

Filters (ITSx) summary

ASVs



Abundance

Filters (ITSx) by samples

CSV

Show 10 entries

Search:

ASVs removed by sample

Sample name	Initial	Kept	Initial abundance	Kept abundance
Ph203	105	105	7,065	7,065
Ph212	65	65	7,474	7,474
Ph217	89	89	5,990	5,990

Exercise

Go to « ITS » history

Launch the FROGS ITSx tool on this data set

Launch the FROGS_5 Taxonomic affiliation

Launch the FROGS Affiliation Stat

FROGS ITSx

Tool Parameters

Input Parameter	Value
Sequence file	**16 : FROGS_4 Cluster filters: clusterFilters_sequences.fasta**
Abundance file	**15 : FROGS_4 Cluster filters: clusterFilters_abundance.biom**
Trim conserved sequence (SSU, 5.8S, LSU) ?	yes
Choose pertinent organisms to scan:	Fungi

FROGS_5 Taxonomic affiliation

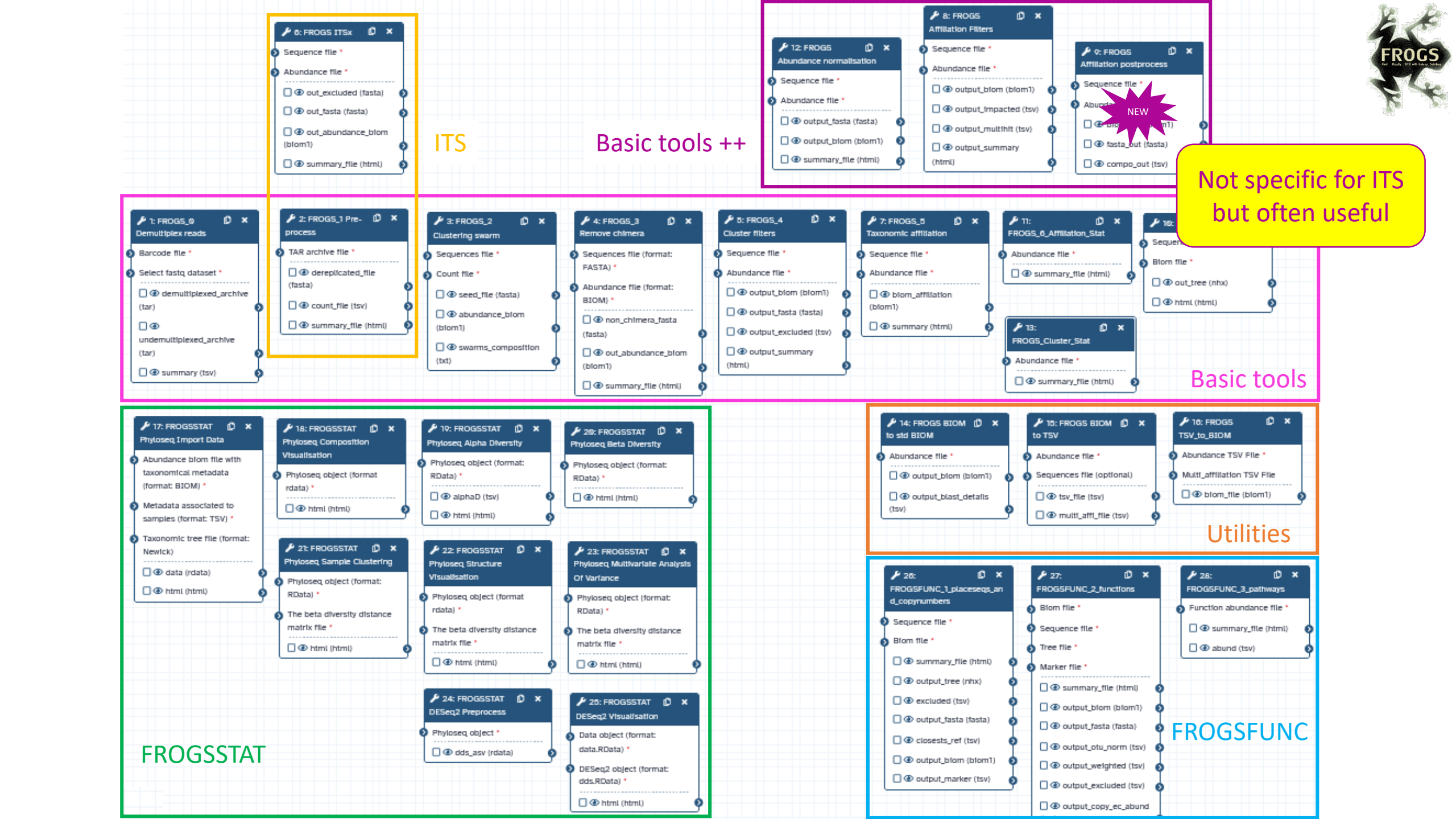
Tool Parameters

Input Parameter	Value
Using reference database	ITS_UNITE_Fungi_8.3
Also perform RDP assignation?	No
Taxonomic ranks	Domain Phylum Class Order Family Genus Species
Sequence file	**20 : FROGS ITSx: ITS_sequence.fasta**
Abundance file	**21 : FROGS ITSx: itsx_abundance.biom**

FROGS_6_Affiliation_Stat

Tool Parameters

Input Parameter	Value
Abundance file	23 : FROGS_5 Taxonomic affiliation: affiliation_abundance.biom
Taxonomic ranks	Domain Phylum Class Order Family Genus Species
Rarefaction ranks	Class Order Family Genus Species
Affiliation processed	FROGS_blast


6: FROGS ITSx
Sequence file *
Abundance file *
out_excluded (fasta)
out_fasta (fasta)
out_abundance_biom (biom1)
summary_file (html)
ITS
Basic tools ++
12: FROGS Abundance normalisation
Sequence file *
Abundance file *
output_fasta (fasta)
output_biom (biom1)
summary_file (html)
8: FROGS Affiliation Filters
Sequence file *
Abundance file *
output_biom (biom1)
output_impacted (tsv)
output_multihit (tsv)
output_summary (html)
9: FROGS Affiliation postprocess
Sequence file *
NEW
fasta_out (fasta)
compo_out (tsv)
Not specific for ITS but often useful
1: FROGS_0 Demultiplex reads
Barcode file *
Select fastq dataset *
demultiplexed_archive (tar)
undemultiplexed_archive (tar)
summary (tsv)
2: FROGS_1 Pre-process
TAR archive file *
dereplicated_file (fasta)
count_file (tsv)
summary_file (html)
3: FROGS_2 Clustering swarm
Sequences file *
Count file *
seed_file (fasta)
abundance_biom (biom1)
swarms_composition (txt)
4: FROGS_3 Remove chimera
Sequences file (format: FASTA) *
Abundance file (format: BIOM) *
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
summary_file (html)
5: FROGS_4 Cluster filters
Sequence file *
Abundance file *
output_biom (biom1)
output_fasta (fasta)
output_excluded (tsv)
output_summary (html)
7: FROGS_5 Taxonomic affiliation
Sequence file *
Abundance file *
biom_affiliation (biom1)
summary (html)
11: FROGS_6_Affiliation_Stat
Abundance file *
summary_file (html)
13: FROGS_Cluster_Stat
Abundance file *
summary_file (html)
Biom file *
out_tree (nhx)
html (html)
Basic tools
17: FROGSSTAT Phyloseq Import Data
Abundance biom file with taxonomical metadata (format: BIOM) *
Metadata associated to samples (format: TSV) *
Taxonomic tree file (format: Newick)
data (rdata)
html (html)
18: FROGSSTAT Phyloseq Composition Visualisation
Phyloseq object (format rdata) *
html (html)
19: FROGSSTAT Phyloseq Alpha Diversity
Phyloseq object (format: RData) *
alphaD (tsv)
html (html)
20: FROGSSTAT Phyloseq Beta Diversity
Phyloseq object (format: RData) *
html (html)
21: FROGSSTAT Phyloseq Sample Clustering
Phyloseq object (format: RData) *
The beta diversity distance matrix file *
html (html)
22: FROGSSTAT Phyloseq Structure Visualisation
Phyloseq object (format rdata) *
The beta diversity distance matrix file *
html (html)
23: FROGSSTAT Phyloseq Multivariate Analysis Of Variance
Phyloseq object (format: RData) *
The beta diversity distance matrix file *
html (html)
24: FROGSSTAT DESeq2 Preprocess
Phyloseq object *
dds_asv (rdata)
25: FROGSSTAT DESeq2 Visualisation
Data object (format: data.RData) *
DESeq2 object (format: dds.RData) *
html (html)
FROGSSTAT
14: FROGS BIOM to std BIOM
Abundance file *
output_biom (biom1)
output_blast_details (tsv)
15: FROGS BIOM to TSV
Abundance file *
Sequences file (optional)
tsv_file (tsv)
multi_affi_file (tsv)
16: FROGS TSV_to_BIOM
Abundance TSV File *
Multi_affiliation TSV File
biom_file (biom1)
Utilities
26: FROGSFUNC_1_placeseqs_and_copynumbers
Sequence file *
Biom file *
summary_file (html)
output_tree (nhx)
excluded (tsv)
output_fasta (fasta)
closests_ref (tsv)
output_biom (biom1)
output_marker (tsv)
27: FROGSFUNC_2_functions
Biom file *
Sequence file *
Tree file *
Marker file *
summary_file (html)
output_biom (biom1)
output_fasta (fasta)
output_otu_norm (tsv)
output_weighted (tsv)
output_excluded (tsv)
output_copy_ec_abund
28: FROGSFUNC_3_pathways
Function abundance file *
summary_file (html)
abund (tsv)
FROGSFUNC

Affiliation Post-process

What is the purpose of the *Affiliation post-process* tool ?

This tool allows **grouping ASVs together** in accordance with the %id and %cov chosen by the user and according to the following criteria:

1. They must have the same affiliation

Or

2. If they have "multi-affiliation" tag in FROGS taxonomy, they must have in common in their list of possible affiliations at least one identical affiliation.

What is the purpose of the *Affiliation post-process* tool ?

<u>In consequence:</u>

The different affiliations involved in multi-affiliation are merged.

The abundances are added together.

It is the most abundant ASV seed that is kept.

FROGS Affiliation postprocess Aggregates ASVs based on alignment metrics (Galaxy Version 4.1.0+galaxy1)

☆ Favorite | Versions | ▾ Options

Sequence file

21: FROGS ITSx: ITS_sequence.fasta

The sequence file to filter (format: FASTA).

Abundance file

25: FROGS_5 Taxonomic affiliation: affiliation_abundance.biom

The abundance file to filter (format: BIOM)

Is this an amplicon hyper variable in length?

○ No
⊘ Yes

Yes, we have combined sequences

Multi-affiliation tag may be resolved by selecting the shortest amplicon reference. For this, you need the reference fasta file of your target amplicon.

Using reference database

UNITE 8.2 ITS1

same database used for taxonomic affiliation

Select reference from the list (--reference)

Minimum identity for aggregation

99

ASVs will be aggregated if they share the same taxonomy with at least X% identity (--identity)

Minimum coverage for aggregation

99

ASVs will be aggregated if they share the same taxonomy with at least X% alignment coverage (--coverage)

Here, we wanted to aggregate ASVs only if they are very closed

Email notification

No

Send an email notification when the job completes.

✔ Execute

FROGS Affiliation postprocess: OTU_aggregation_composition.txt

FROGS Affiliation postprocess: sequences.fasta

FROGS Affiliation postprocess: affiliation_abundance.biom

Cluster_1
Cluster_2
Cluster_8
Cluster_3
Cluster_5
Cluster_4
Cluster_6
Cluster_7
Cluster_9
Cluster_13
Cluster_10
Cluster_11
Cluster_16
Cluster_17
Cluster_14
Cluster_12
Cluster_15
Cluster_22
Cluster_18
Cluster_23
Cluster_25
Cluster_19
Cluster_21
Cluster_26
Cluster_29
Cluster_34
Cluster_35
Cluster_28
Cluster_31
Cluster_32
Cluster_42
Cluster_33
Cluster_75_FROGS_combined Cluster_121_FROGS_combined Cluster_137_FROGS_combined Cluster_144_FROGS_combi

Would you like to take your analysis further?

FROGS Tree

Tool Parameters

Input Parameter	Value
Sequence file	20 : FROGS ITSx: ITS_sequence.fasta
Biom file	23 : FROGS_5 Taxonomic affiliation: affiliation_abundance.biom

FROGS BIOM to TSV

Tool Parameters

Input Parameter	Value
Abundance file	**23 : FROGS_5 Taxonomic affiliation: affiliation_abundance.biom**
Sequences file (optional)	**20 : FROGS ITSx: ITS_sequence.fasta**
Extract multi-alignments	Yes

FROGSSTAT Phyloseq Import Data

Tool Parameters

Input Parameter	Value
Abundance biom file with taxonomical metadata (format: BIOM)	**23 : FROGS_5 Taxonomic affiliation: affiliation_abundance.biom**
Metadata associated to samples (format: TSV)	**2 : ITS_fast_metadata.tsv**
Taxonomic tree file (format: Newick)	**26 : FROGS Tree: tree.nwk**
Names of taxonomic levels	Kingdom Phylum Class Order Family Genus Species
Do you want to normalise your data ?	Yes, subsample abundances to the smallest sample size.

FROGSSTAT Phyloseq Import Data

FROGSSTAT Phyloseq Import Data

Summary | Ranks Names | Sample metadata | Plot tree

```
phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 186 taxa and 20 samples ]
sample_data() Sample Data:       [ 20 samples by 8 sample variables ]
tax_table()   Taxonomy Table:    [ 186 taxa by 14 taxonomic ranks ]
phy_tree()    Phylogenetic Tree: [ 186 tips and 185 internal nodes ]
```

```
Number of sequences in each sample after normalisation:  1454
```

Summary | Ranks Names | Sample metadata | Plot tree

Show

```
Sample variables:  kept, Replicas, Incubation, Nitrogen, Forest_management, Quality, Treatment, SampleID
```

Show

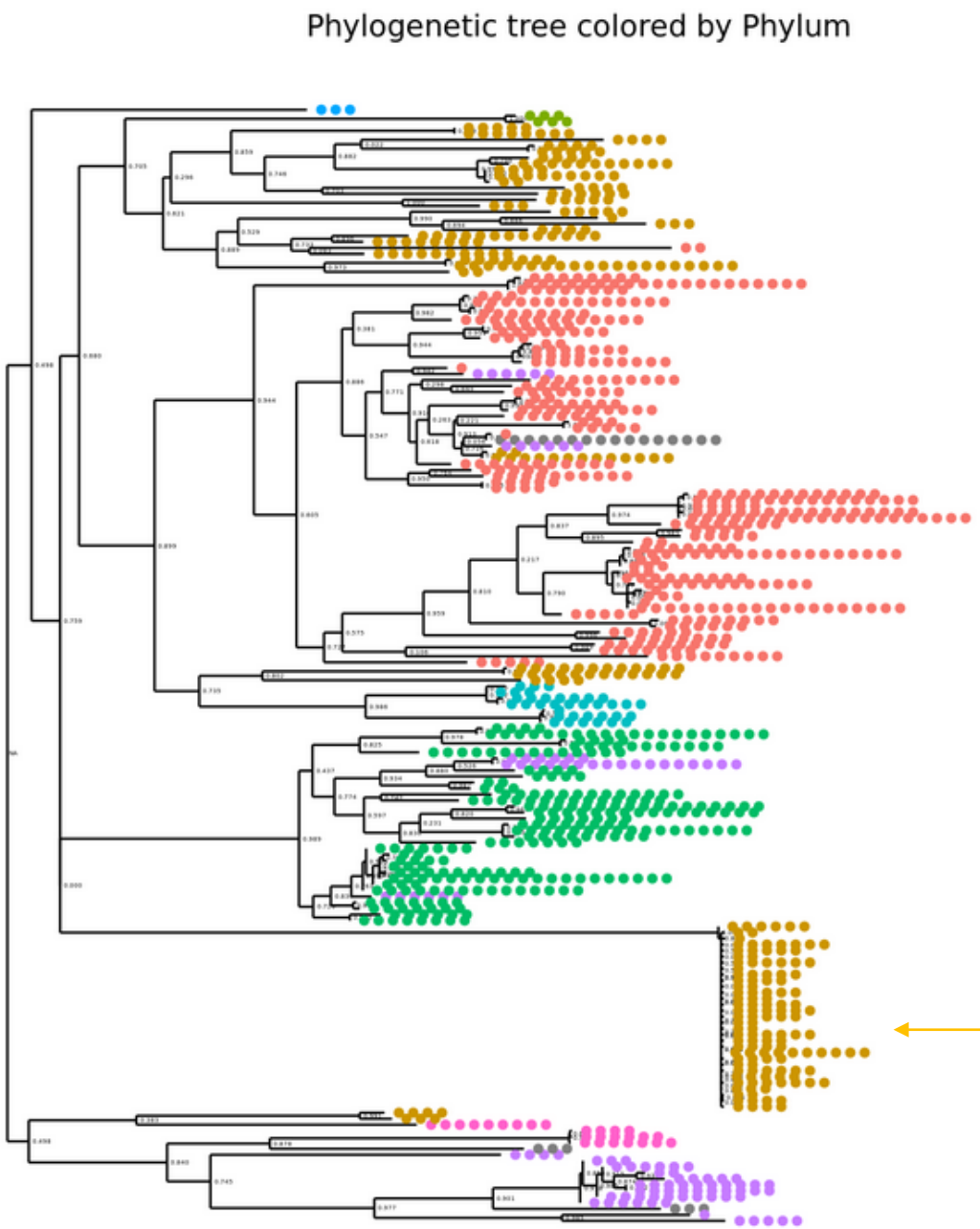
```
kept :  79.76, 77.64, 80.26, 78.65, 77.18, 79.68, 78.7, 76.38, 76.37, 77.37, 72.52, 64.98, 78.13, 71.17, 75.2, 73.48, 73.21, 74.01, 74.15, 73.77

Replicas :  3, 2, 5, 1, 4

Incubation :  T4

Nitrogen :  Nitrogen_supplementation, Control

Forest_management :  Control, OMR

Quality :  Low degradability

Treatment :  Control_with_N, Control, OMR_with_N, OMR

SampleID :  Ph203, Ph212, Ph217, Ph222, Ph224, Ph237, Ph241, Ph243, Ph246, Ph250, Ph407, Ph414, Ph415, Ph417, Ph423, Ph428, Ph433, Ph434, Ph439, Ph449
```

Summary | Ranks Names | Sample metadata | Plot tree

Show

```
Rank names :  Kingdom, Phylum, Class, Order, Family, Genus, Species, Rank2, Rank3, Rank4, Rank5, Rank6, Rank7, Rank1
```

Summary | Ranks Names | Sample metadata | **Plot tree**



Le phylum des Basidiomycota est éclaté en plusieurs endroits de l'arbre, données Unifrac à considérer avec précautions

Ce bloc correspond aux multiples clusters artificial_combined
On le sait en comparant un historique n'acceptant pas en preprocess les unmerged.