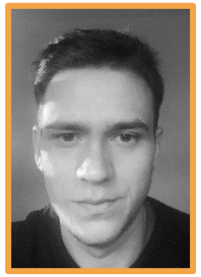
B- Training on Galaxy: Metabarcoding

October 2022 - Webinar

FROGS Practice on 16S data

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

Who is in the current FROGS group?



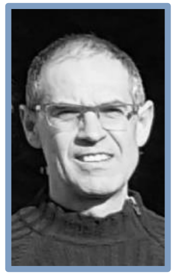
Vincent DARBOT

Maria BERNARD

Olivier RUÉ

Developers

Lucas AUER



Laurent CAUQUIL

Biology experts

Patrice DÉHAIS

Galaxy support

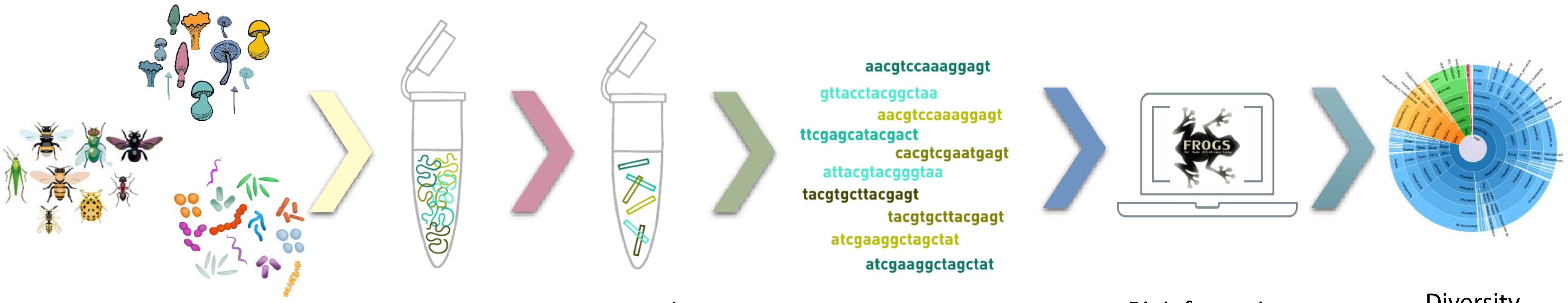
Mahendra MARIADASSOU

Statistical expert

Géraldine PASCAL

Coordinator

Objectives



Community analysis

DNA extraction

Barcode amplification

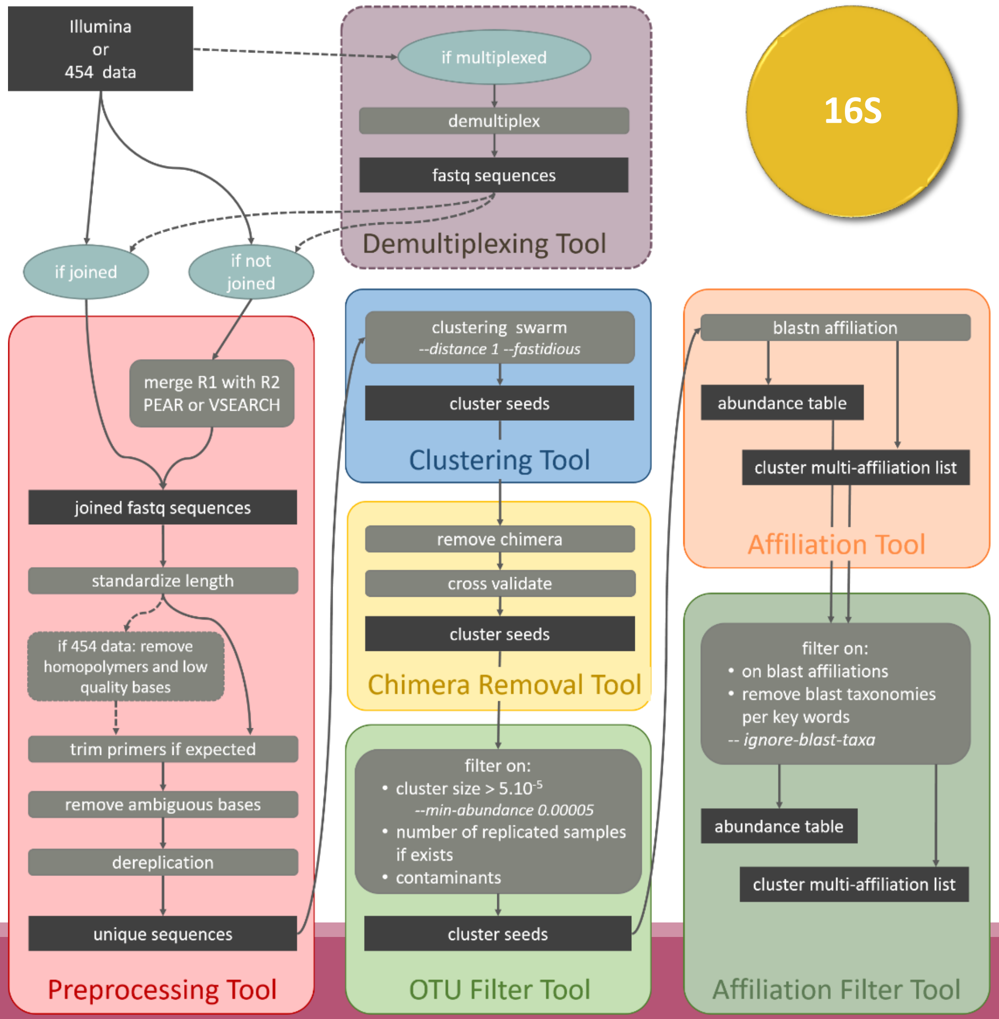
Amplicon sequencing

Bioinformatics process

Diversity characterisation

An abundance table with operational taxonomic units (OTUs) and their taxonomic affiliation.

Bioinformatics process



16S

Illumina or 454 data

if joined

if not joined

Demultiplexing Tool
- if multiplexed
- demultiplex
- fastq sequences

Preprocessing Tool
- merge R1 with R2 PEAR or VSEARCH
- joined fastq sequences
- standardize length
- if 454 data: remove homopolymers and low quality bases
- trim primers if expected
- remove ambiguous bases
- dereplication
- unique sequences

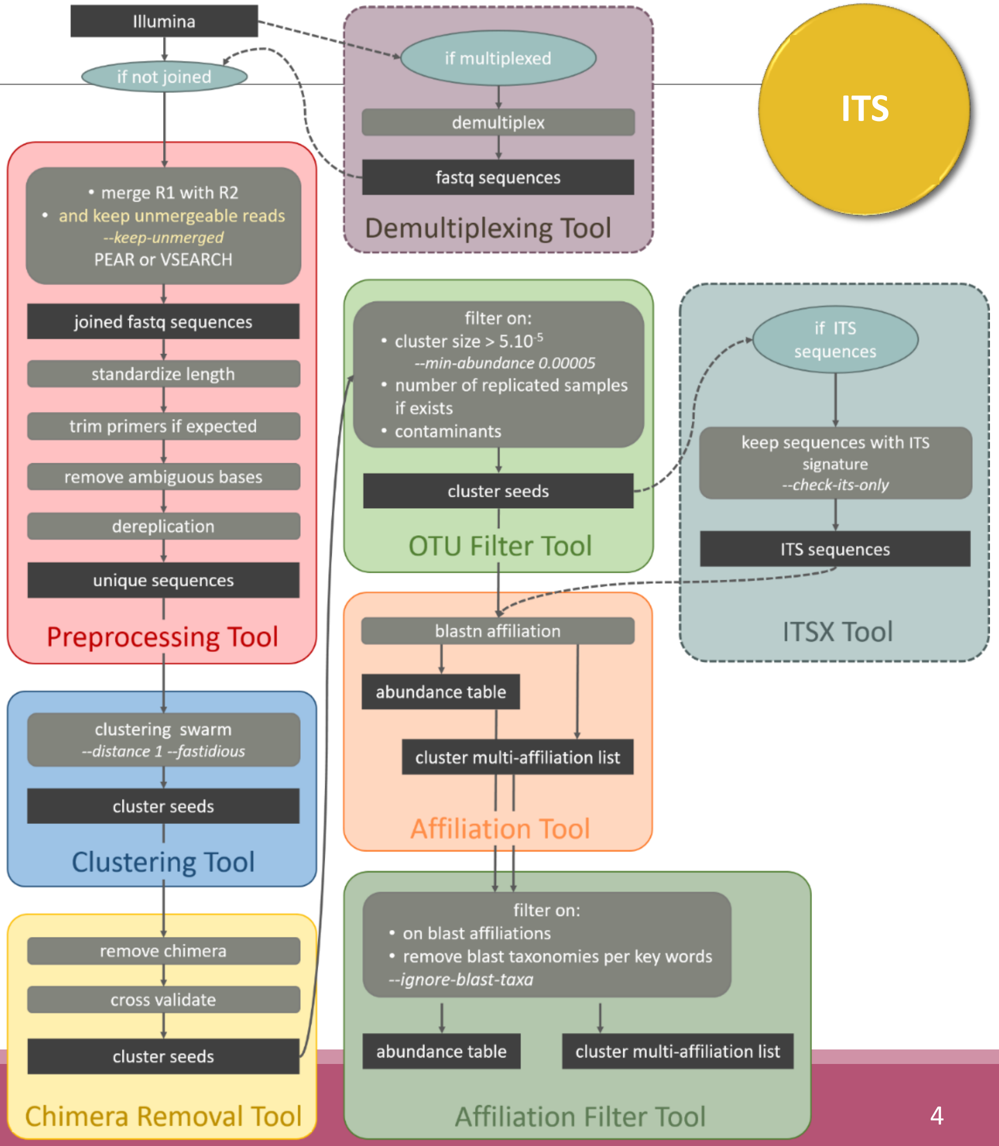
Clustering Tool
- clustering swarm *--distance 1 --fastidious*
- cluster seeds

Chimera Removal Tool
- remove chimera
- cross validate
- cluster seeds

OTU Filter Tool
- filter on:
  - cluster size > 5.10^{-5} *--min-abundance 0.00005*
  - number of replicated samples if exists
  - contaminants
- cluster seeds

Affiliation Tool
- blastn affiliation
- abundance table
- cluster multi-affiliation list

Affiliation Filter Tool
- filter on:
  - on blast affiliations
  - remove blast taxonomies per key words *-- ignore-blast-taxa*
- abundance table
- cluster multi-affiliation list



ITS

Illumina

if not joined

Demultiplexing Tool
- if multiplexed
- demultiplex
- fastq sequences

Preprocessing Tool
- merge R1 with R2
- and keep unmergeable reads *--keep-unmerged* PEAR or VSEARCH
- joined fastq sequences
- standardize length
- trim primers if expected
- remove ambiguous bases
- dereplication
- unique sequences

Clustering Tool
- clustering swarm *--distance 1 --fastidious*
- cluster seeds

Chimera Removal Tool
- remove chimera
- cross validate
- cluster seeds

OTU Filter Tool
- filter on:
  - cluster size > 5.10^{-5} *--min-abundance 0.00005*
  - number of replicated samples if exists
  - contaminants
- cluster seeds

ITSX Tool
- if ITS sequences
- keep sequences with ITS signature *--check-its-only*
- ITS sequences

Affiliation Tool
- blastn affiliation
- abundance table
- cluster multi-affiliation list

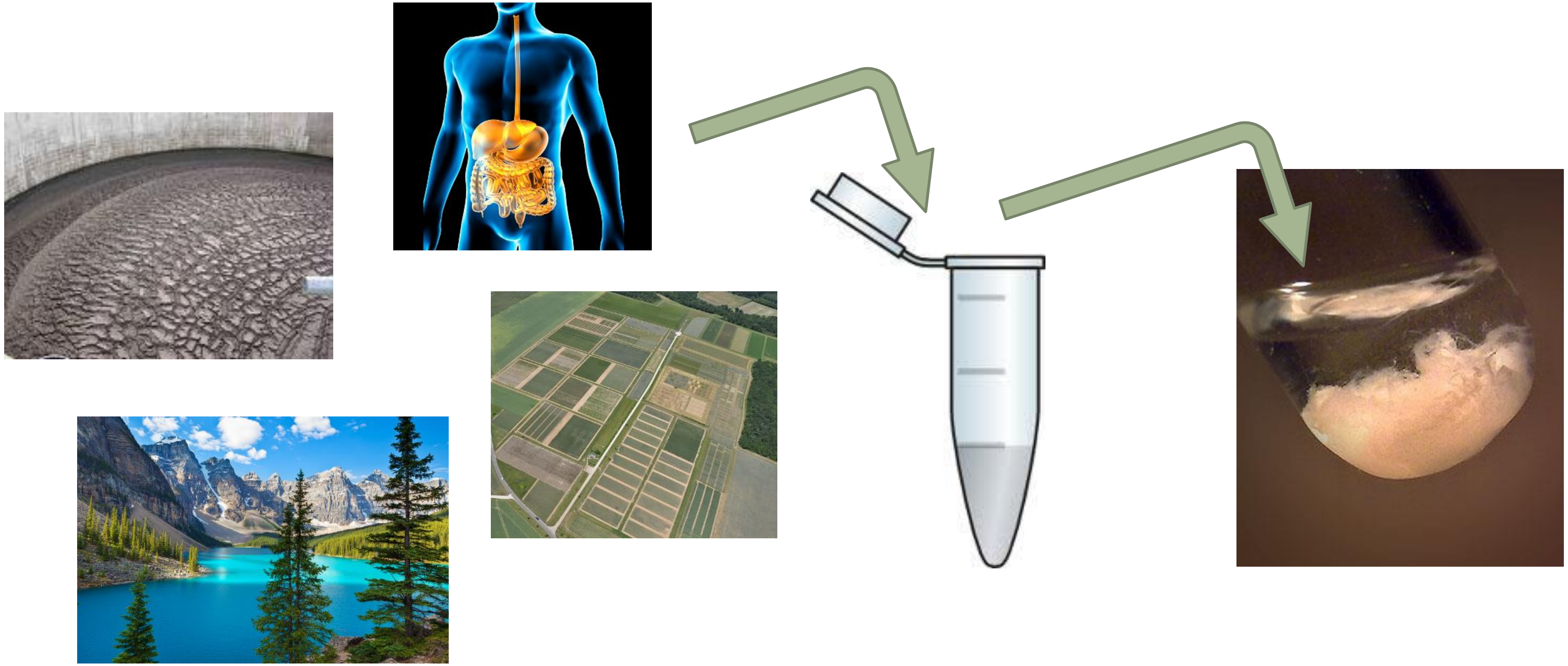
Affiliation Filter Tool
- filter on:
  - on blast affiliations
  - remove blast taxonomies per key words *--ignore-blast-taxa*
- abundance table
- cluster multi-affiliation list

Objectives: a count table

	Affiliation	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
OTU1	Species A	0	100	0	45	75	18645
OTU2	Species B	741	0	456	4421	1255	23
OTU3	Species C	12786	45	3	0	0	0
OTU4	Species D	127	4534	80	456	756	108
OTU5	Species E	8766	7578	56	0	0	200

Material

Sample collection and DNA extraction



« Meta-omics » using next-generation sequencing (NGS)

DNA

RNA

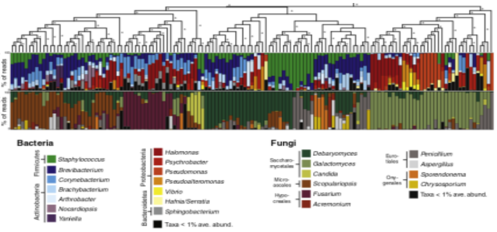
Metagenomics

Metatranscriptomics

Amplicon sequencing

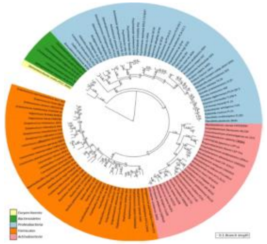
Shotgun sequencing

RNA sequencing



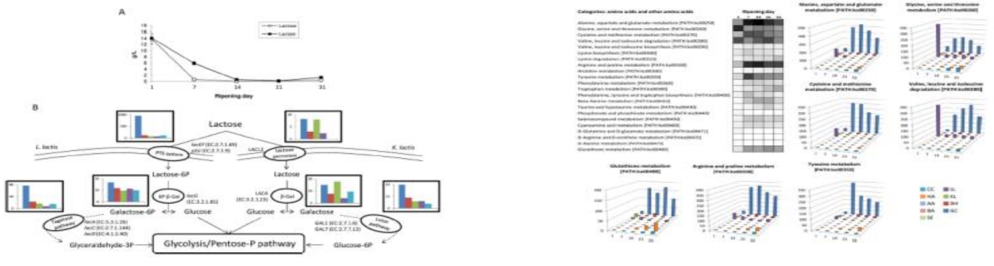
Wolfe *et al.*, 2014

Who is here?



Almeida *et al.*, 2014

What can they do?



Dugat-Bony *et al.*, 2015

What are they doing?

Story of barcoding

- Early 2000's: beginning of barcoding
- 1st DNA barcode: 65 bases of the mitochondrial gene of Cytochrome Oxidase I (COI) dedicated to the identification of vertebrates
- 2007: 1st international published database
- 2009: chloroplastic markers - RBCL (Ribulose Biphosphate Carboxylase; 553 pairs of bases) and MATK (MATurase K; 879 pairs of bases) -> standard markers for plants
- 2012: ITS, standard marker of fungi (length between 361–1475 bases in UNITE 7.1)
- 16S marker, mainly used for bacteria but no designated standard.

Which barcode ?

Microbial lineages vary in their genomic contents, which suggests that different genes might be needed to resolve the diversity within certain taxonomic groups.

- 16S rRNA
- 23S rRNA,
- DNA gyrase subunit B (gyrB),
- RNA polymerase subunit B (rpoB),
- TU elongation factor (tuf),
- DNA recombinase protein (recA),
- protein synthesis elongation factor-G (fusA),
- dinitrogenase protein subunit D (nifD),
- Internal Transcribed Spacer (ITS) for Fungi.

The gene encoding the small subunit of the ribosomal RNA

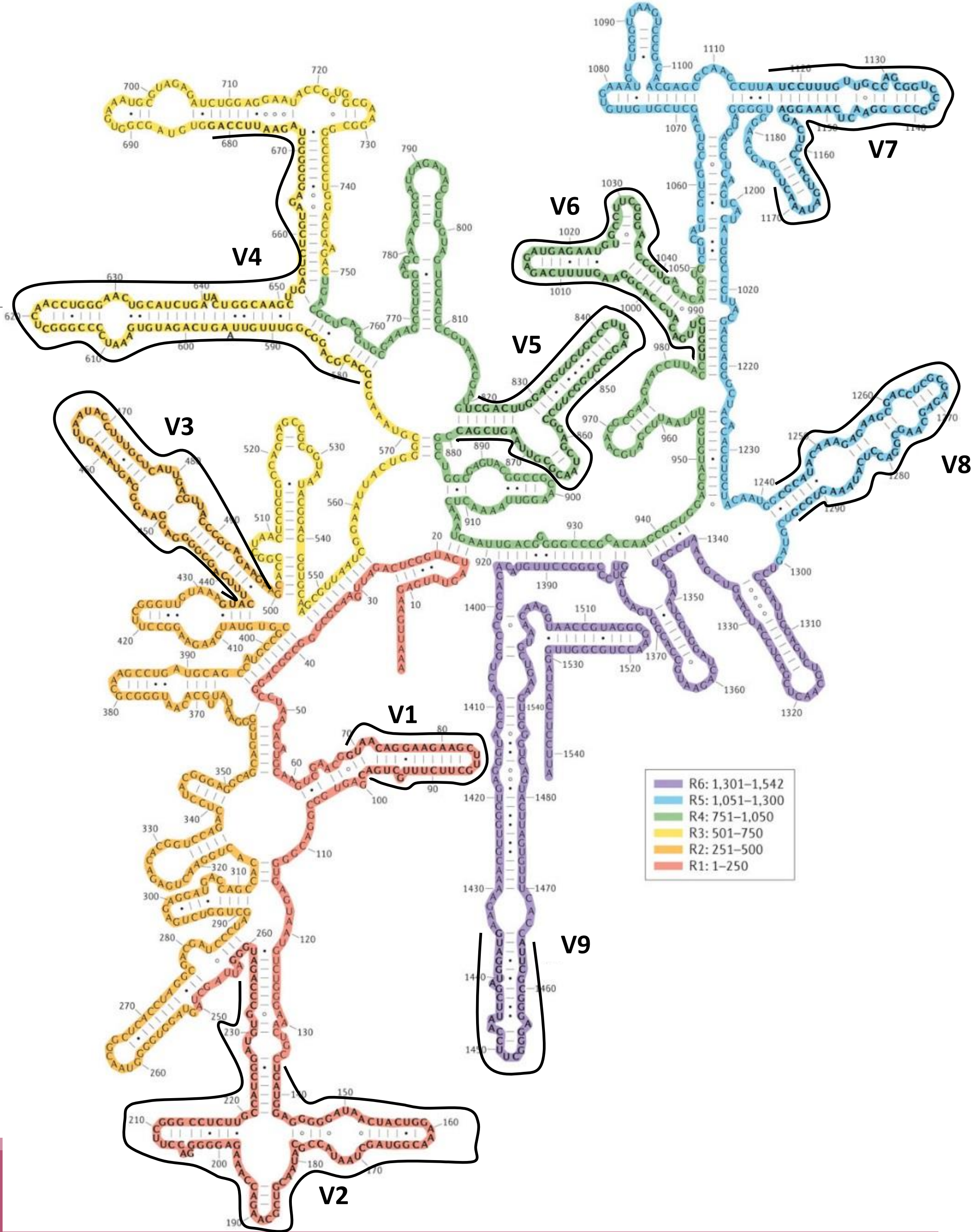
The most widely used gene in **molecular phylogenetic** studies

Ubiquist gene : **16S rDNA** in prokaryotes ; **18S rDNA** in eukaryotes

Gene encoding a ribosomal RNA : non-coding RNA (not translated), part of the small subunit of the ribosome which is responsible for the translation of mRNA in proteins

Not submitted to lateral gene transfer

Availability of databases facilitating comparison
(Silva v138.1 - 2021: available SSU/LSU sequences to over **10,700,000**)

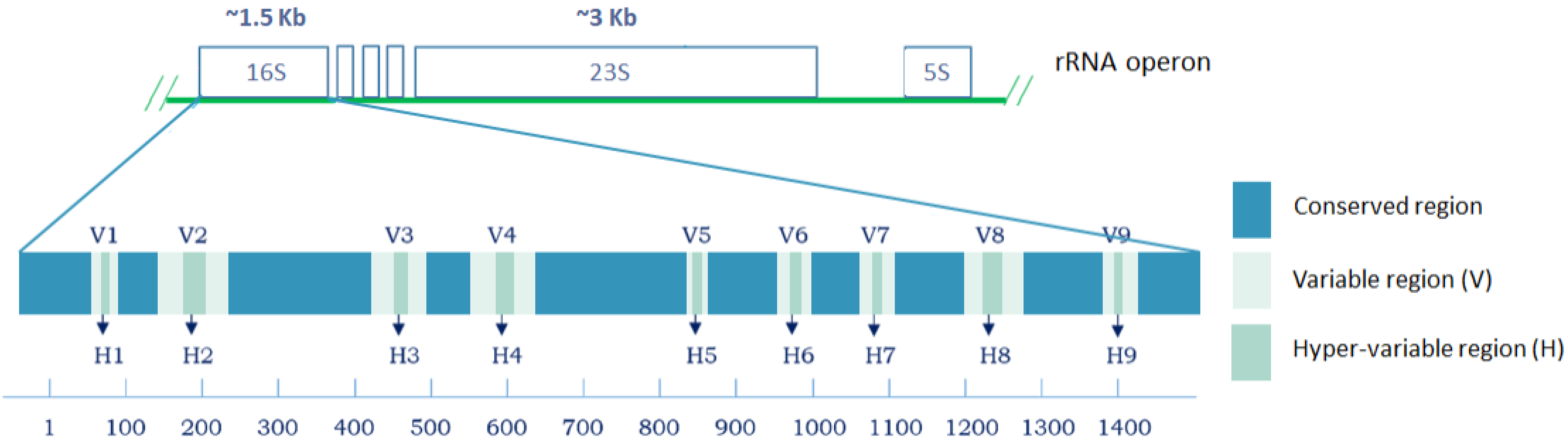


Secondary structure of the 16S rRNA of *Escherichia coli*

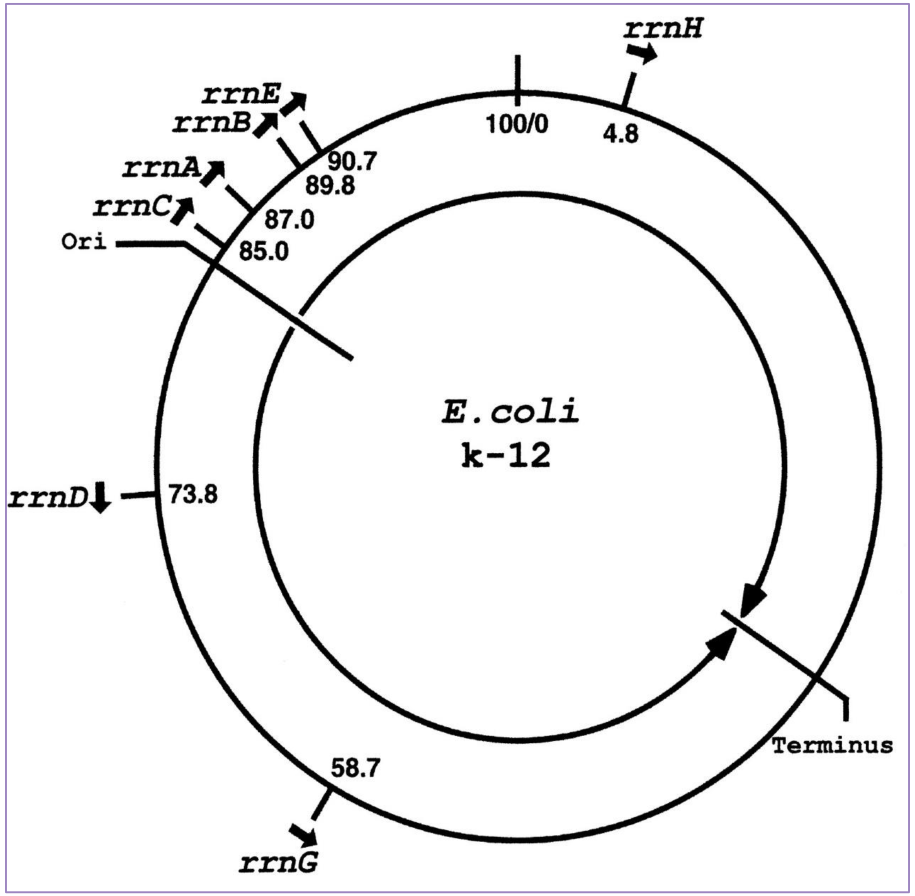
In red, fragment R1 including regions V1 and V2;
in orange, fragment R2 including region V3;
in yellow, fragment R3 including region V4;
in green, fragment R4 including regions V5 and V6;
in blue, fragment R5 including regions V7 and V8;
and in purple, fragment R6 including region V9.

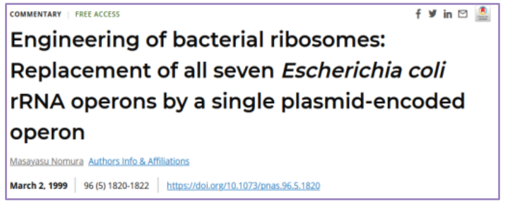
Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences
Pablo Yarza, et al.
Nature Reviews Microbiology 12, 635–645 (2014) doi:10.1038/nrmicro3330

16S rRNA structure



16S rRNA copy number



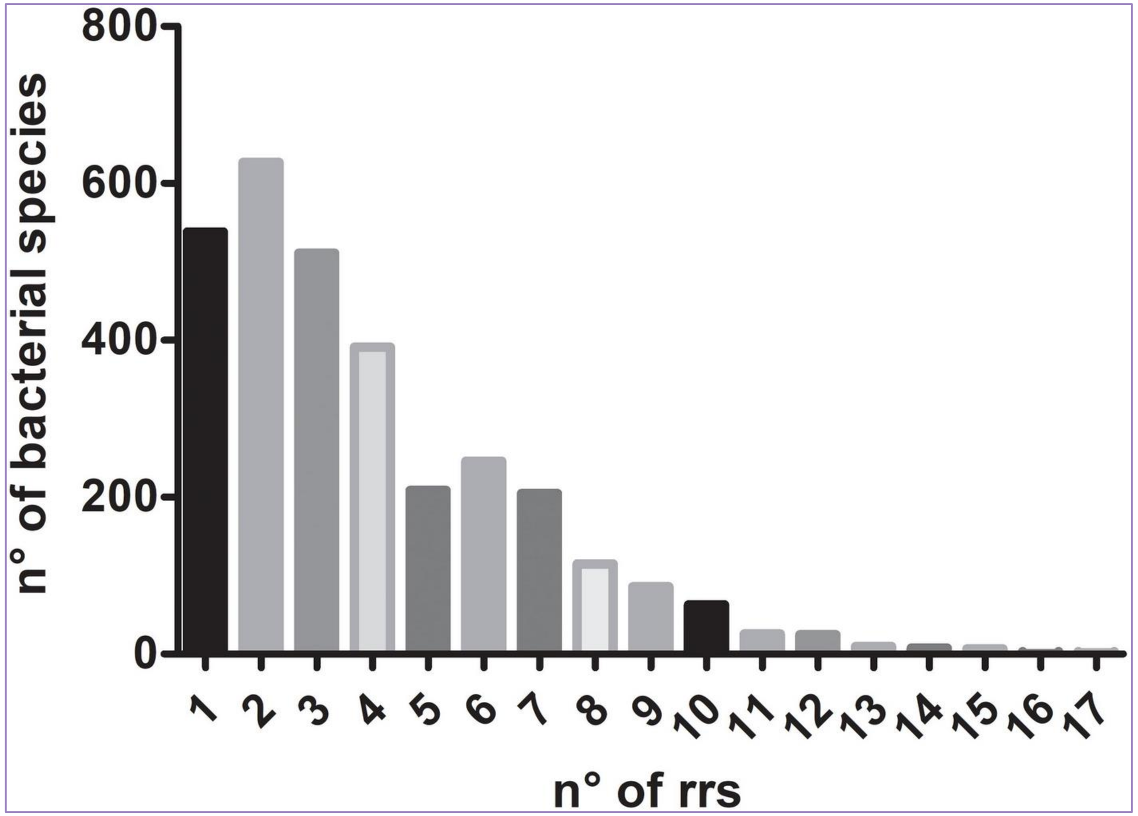

COMMENTARY | FREE ACCESS

Engineering of bacterial ribosomes: Replacement of all seven *Escherichia coli* rRNA operons by a single plasmid-encoded operon

Masayasu Nomura Authors Info & Affiliations

March 2, 1999 | 96 (5) 1820-1822 | https://doi.org/10.1073/pnas.96.5.1820

16S rRNA copy number



MINI REVIEW article

Front. Microbiol., 08 June 2018 | https://doi.org/10.3389/fmicb.2018.01232

Multiple Ribosomal RNA Operons in Bacteria; Their Concerted Evolution and Potential Consequences on the Rate of Evolution of Their 16S rRNA

Romilio T. Espejo[1*] and Nicolás Plaza[1,2]

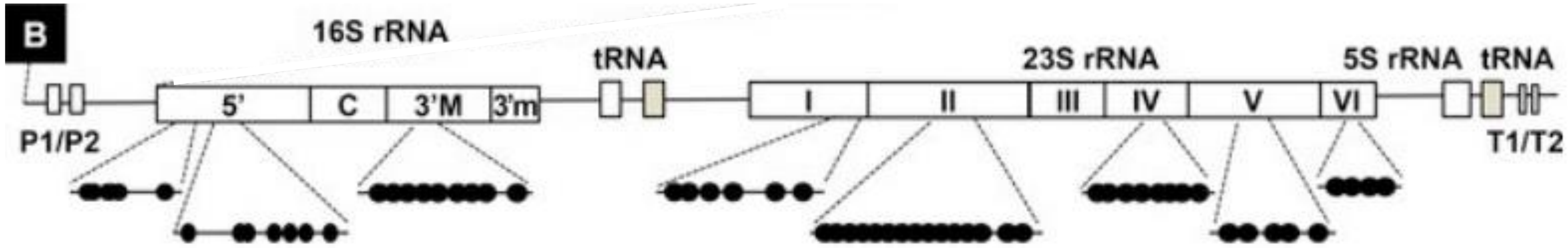
[1]Institute of Nutrition and Food Technology, Universidad de Chile, Santiago, Chile

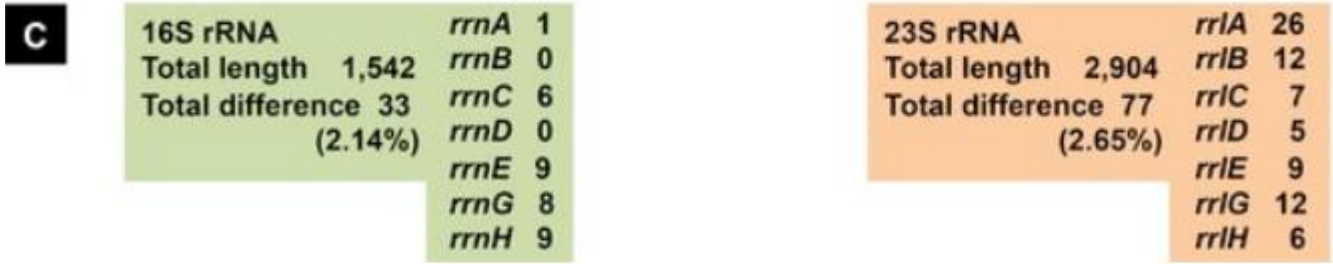
[2]Centro de Investigación Biomédica, Facultad de Ciencias de la Salud, Instituto de Ciencias Biomédicas, Universidad Autónoma de Chile, Santiago, Chile

Median of the number of *16S rRNA* copies in 3,070 bacterial species according to data reported in *rrn*DB database – 2018
https://rrndb.umms.med.umich.edu/search/

2022:
Bacillus megaterium entre 1 à 21 copies selon les souches (médiane à 13)
Photobacterium damselae entre 15 et 21 copie selon les souches (médiane à 17)

16S rRNA copy variation





16S rRNA
Total length 1,542
Total difference 33
(2.14%)

rrnA	1
rrnB	0
rrnC	6
rrnD	0
rrnE	9
rrnG	8
rrnH	9

23S rRNA
Total length 2,904
Total difference 77
(2.65%)

rrlA	26
rrlB	12
rrlC	7
rrlD	5
rrlE	9
rrlG	12
rrlH	6

E. coli

[B] The positions of sequence variation within 16S and 23S rRNA are shown along the gene organization of rrn operons. A total of 33 and 77 differences were identified in 16S rRNA and 23S rRNA, respectively.

[C] The number of bases that are different from the conserved sequence are shown for 16S and 23S rRNA for each rrn operon.

PLOS | ONE

RESEARCH ARTICLE

Strength and Regulation of Seven rRNA Promoters in *Escherichia coli*

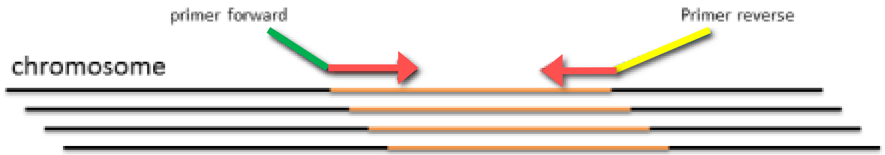
Michihisa Maeda[1], Tomohiro Shimada[2,3], Akira Ishihama[3]*

1 Meiji University, Faculty of Agriculture Chemistry, Kawasaki, Kanagawa 214–8571, Japan, 2 Chemical Resources Laboratory, Tokyo Institute of Technology, Nagatsuda, Yokohama 226–8503, Japan, 3 Research Center for Micro-Nano Technology, Hosei University, Koganei, Tokyo 184–8584, Japan

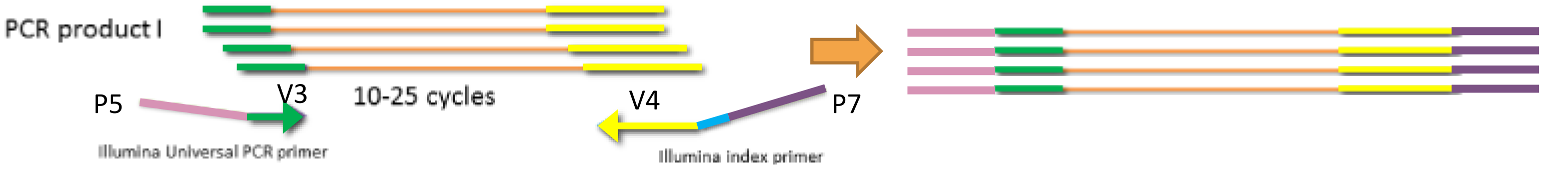
Sequencing produces marker reads

Steps for Illumina sequencing

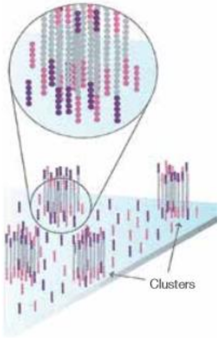
- 1st step : one PCR



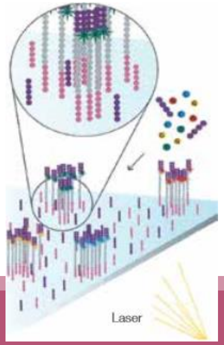
- 2nd step: one PCR



- 3rd step: on flow cell, the cluster generations

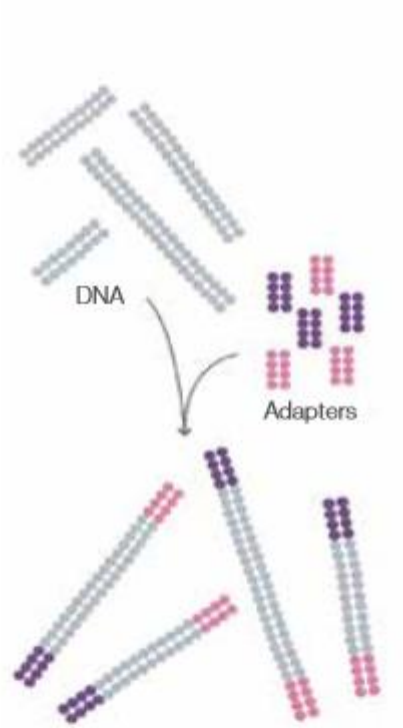


- 4th step: sequencing



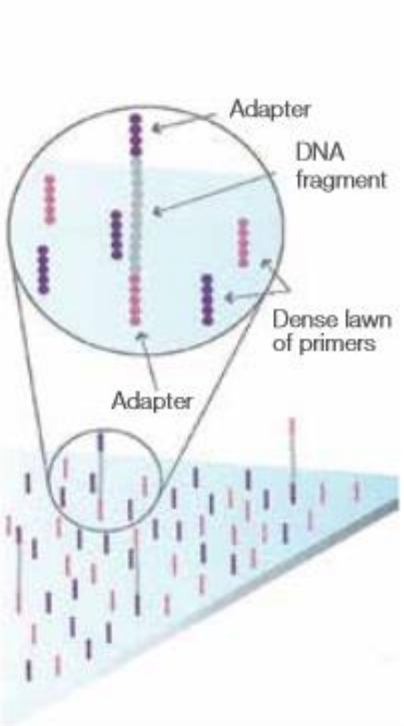
Cluster generation

Prepare Genomic DNA Sample



Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

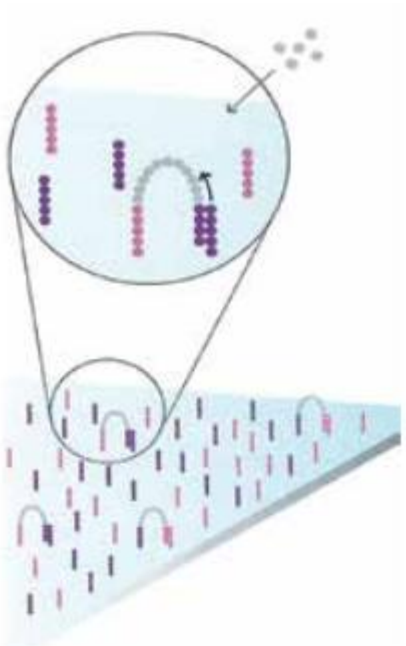
Attach DNA to Surface



Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

Attach DNA to surface

Bridge Amplification

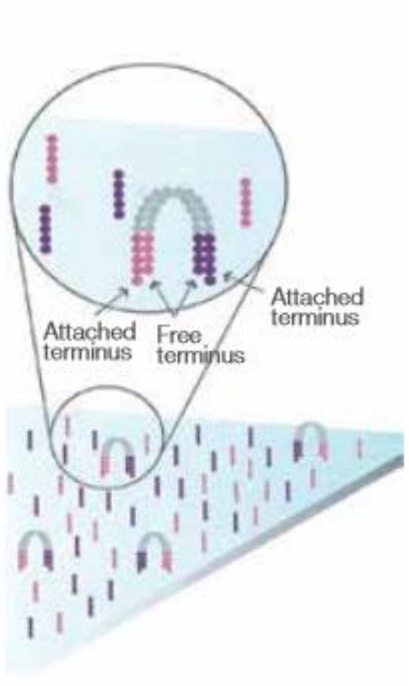


Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

Bridge amplification

Cluster generation

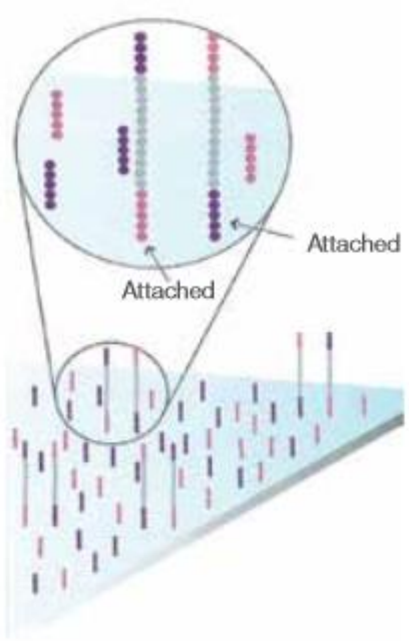
Fragments Become Double Stranded



The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

Fragments become double stranded

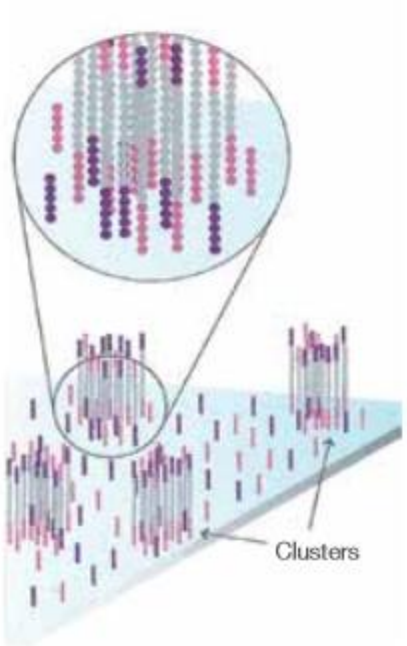
Denature the Double-Stranded Molecules



Denaturation leaves single-stranded templates anchored to the substrate.

Denature the double-stranded molecule

Complete Amplification

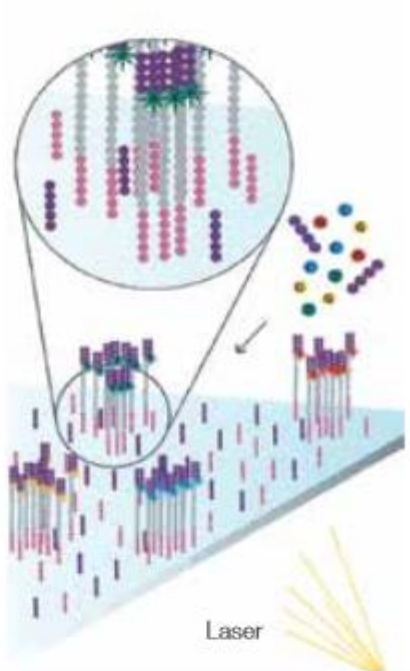


Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

Cycle of new strand synthesis and denaturation to make multiple copies of the same sequence (amplification)
Reverse strands are washed

Sequencing by synthesis

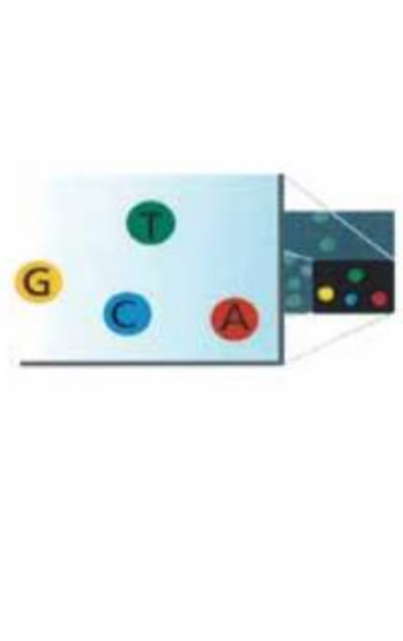
Determine First Base



The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.

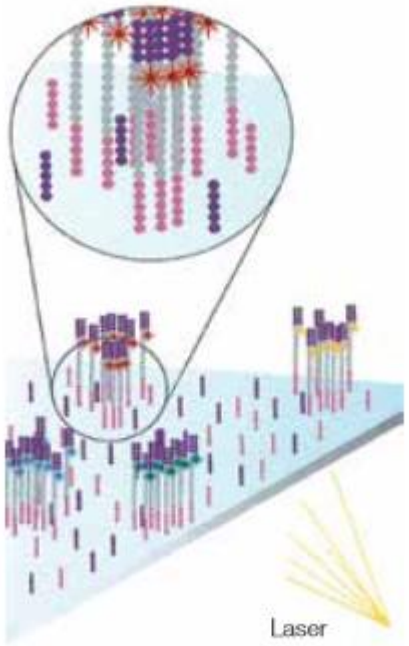
Light signal is more strong in cluster

Image First Base



After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.

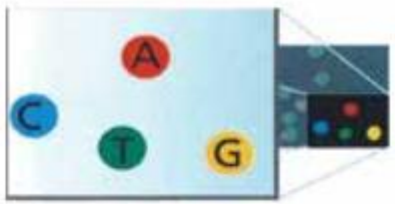
Determine Second Base



The next cycle repeats the incorporation of four labeled reversible terminators, primers, and DNA polymerase.

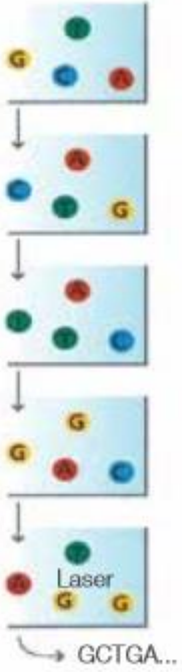
Sequencing by synthesis

Image Second Chemistry Cycle



After laser excitation, the image is captured as before, and the identity of the second base is recorded.

Sequencing Over Multiple Chemistry Cycles

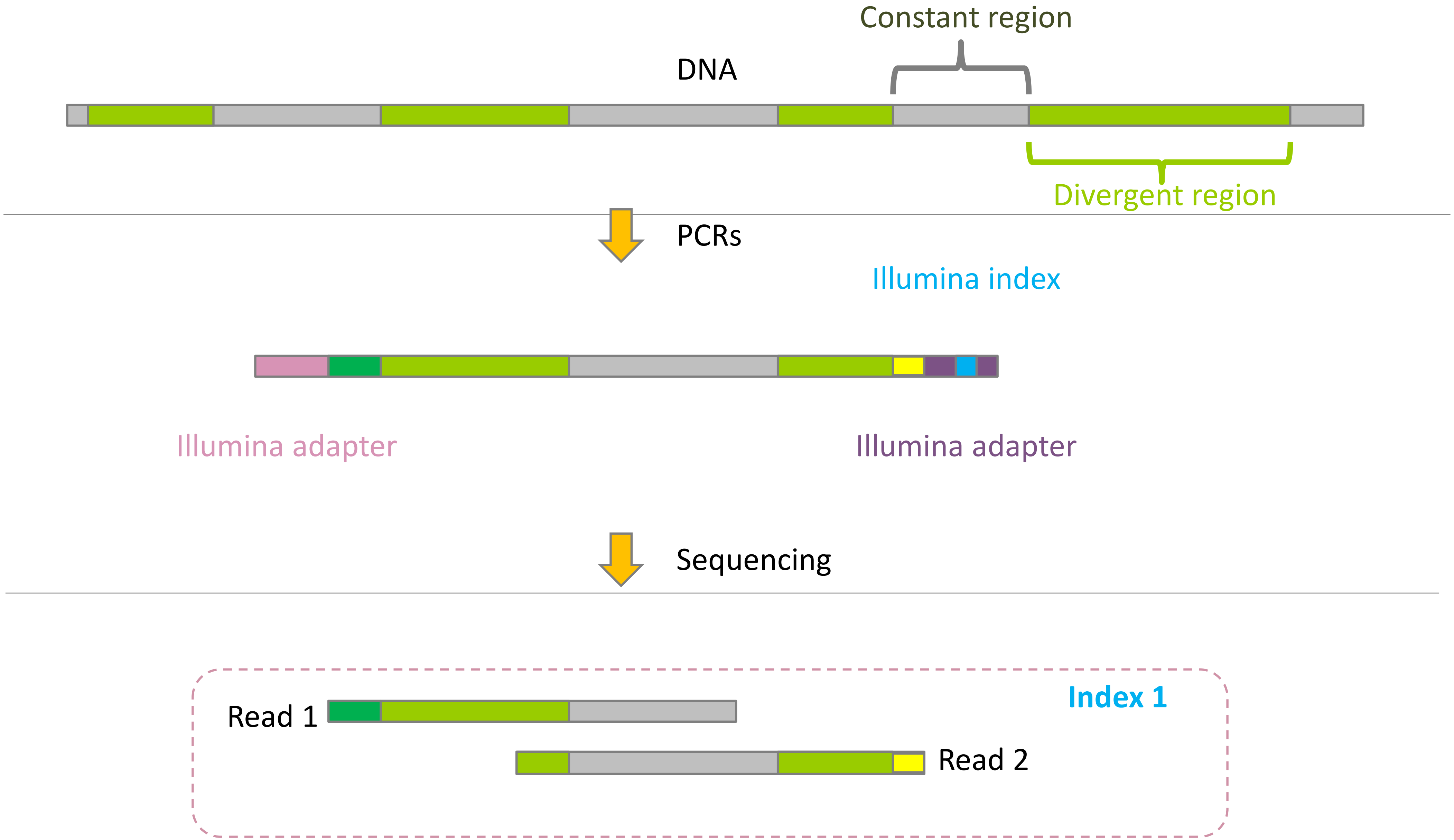


The sequencing cycles are repeated to determine the sequence of bases in a fragment, one base at a time.

Barcode is read, so cluster is identified.
After first sequencing (250 or 300 nt of Reverse strand), fragment form bridges again and Forward strand can be sequenced also.

Illumina sequencing

Illumina video:
https://www.youtube.com/watch?v=fCd6B5HRaZ8


Constant region
DNA
Divergent region
PCRs
Illumina index
Illumina adapter
Illumina adapter
Sequencing
Read 1
Index 1
Read 2

Amplification and sequencing

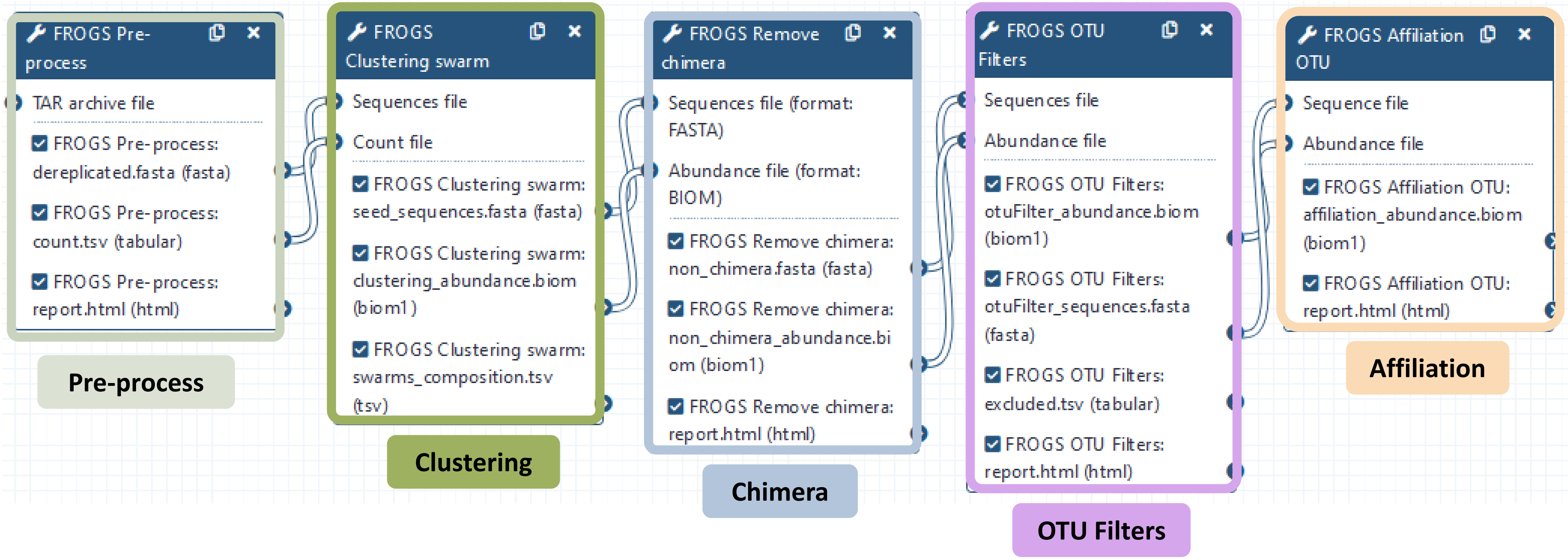
Sequencing is generally perform on Roche-454 (obsolete now) or Illumina MiSeq platforms or Oxford Nanopore Technology or PACBIO platforms.

Read quantity: ~10 000 reads per sample (454), ~30 000 reads per sample (MiSeq), up to several Tera of data (ONT).

Sequence lengths: >650 bp (Roche-454), 2 x 250 bp or 2 x 300 bp (MiSeq), Longest read > 2Mb (ONT or PACBIO)

Methods

Exemple of FROGS Pipeline



FROGS Pre-process

TAR archive file

☑ FROGS Pre-process: dereplicated.fasta (fasta)

☑ FROGS Pre-process: count.tsv (tabular)

☑ FROGS Pre-process: report.html (html)

FROGS Clustering swarm

Sequences file

Count file

☑ FROGS Clustering swarm: seed_sequences.fasta (fasta)

☑ FROGS Clustering swarm: clustering_abundance.biom (biom1)

☑ FROGS Clustering swarm: swarms_composition.tsv (tsv)

FROGS Remove chimera

Sequences file (format: FASTA)

Abundance file (format: BIOM)

☑ FROGS Remove chimera: non_chimera.fasta (fasta)

☑ FROGS Remove chimera: non_chimera_abundance.biom (biom1)

☑ FROGS Remove chimera: report.html (html)

FROGS OTU Filters

Sequences file

Abundance file

☑ FROGS OTU Filters: otuFilter_abundance.biom (biom1)

☑ FROGS OTU Filters: otuFilter_sequences.fasta (fasta)

☑ FROGS OTU Filters: excluded.tsv (tabular)

☑ FROGS OTU Filters: report.html (html)

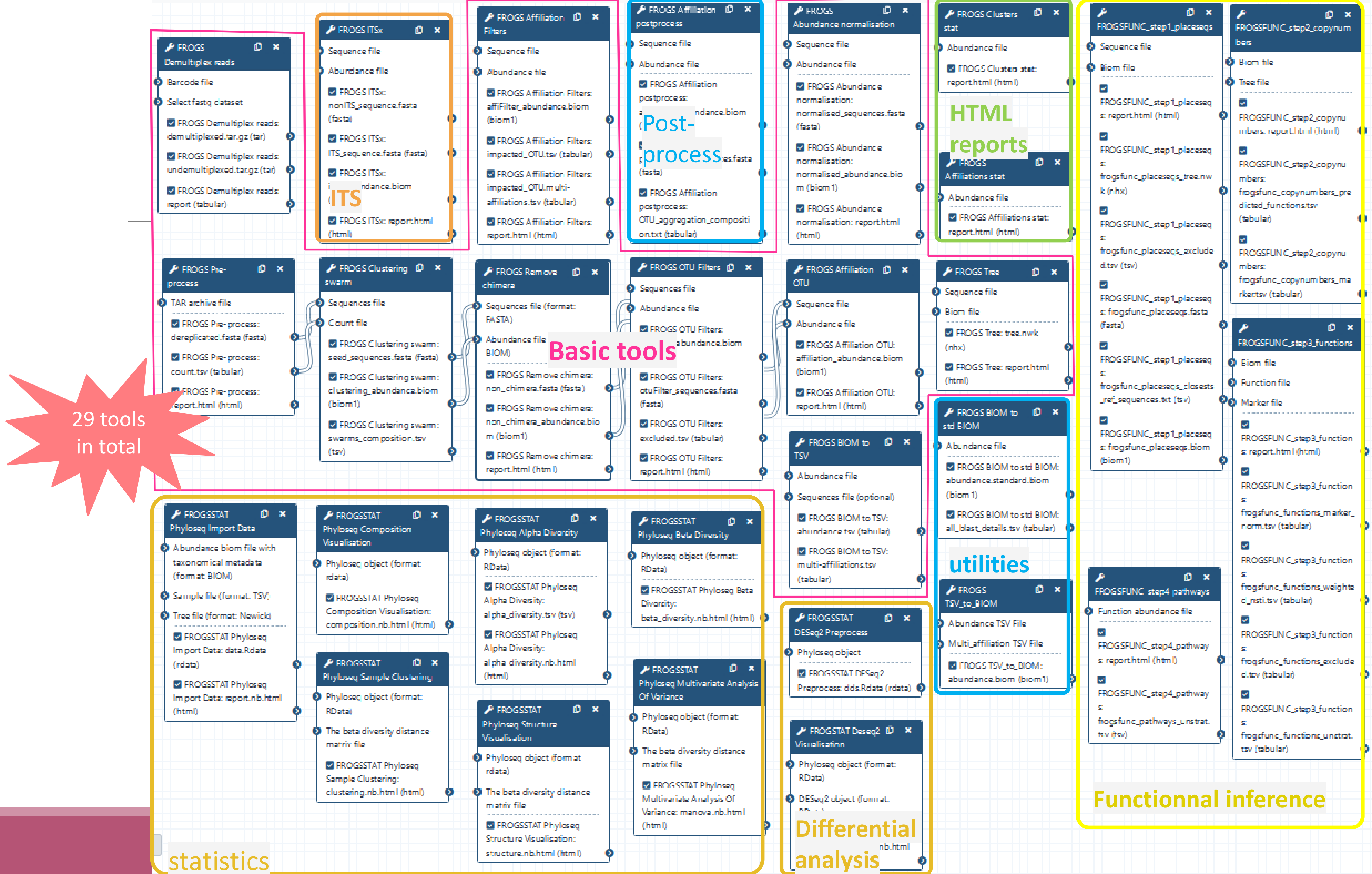
FROGS Affiliation OTU

Sequence file

Abundance file

☑ FROGS Affiliation OTU: affiliation_abundance.biom (biom1)

☑ FROGS Affiliation OTU: report.html (html)


FROGS Demultiplex reads
Barcode file
Select fastq dataset
FROGS Demultiplex reads: demultiplexed.tar.gz (tar)
FROGS Demultiplex reads: undemultiplexed.tar.gz (tar)
FROGS Demultiplex reads: report (tabular)
FROGS ITSx
Sequence file
Abundance file
FROGS ITSx: nonITS_sequence.fasta (fasta)
FROGS ITSx: ITS_sequence.fasta (fasta)
FROGS ITSx: ...ndance.biom
FROGS ITSx: report.html (html)
FROGS Affiliation Filters
Sequence file
Abundance file
FROGS Affiliation Filters: affiFilter_abundance.biom (biom1)
FROGS Affiliation Filters: impacted_OTU.tsv (tabular)
FROGS Affiliation Filters: impacted_OTU.multi-affiliations.tsv (tabular)
FROGS Affiliation Filters: report.html (html)
FROGS Affiliation postprocess
Sequence file
Abundance file
FROGS Affiliation postprocess: ...ndance.biom
...es.fasta (fasta)
FROGS Affiliation postprocess: OTU_aggregation_composition.txt (tabular)
FROGS Abundance normalisation
Sequence file
Abundance file
FROGS Abundance normalisation: normalised_sequences.fasta (fasta)
FROGS Abundance normalisation: normalised_abundance.biom (biom1)
FROGS Abundance normalisation: report.html (html)
FROGS Clusters stat
Abundance file
FROGS Clusters stat: report.html (html)
FROGS Affiliations stat
Abundance file
FROGS Affiliations stat: report.html (html)
FROGSFUNC_step1_placeseqs
Sequence file
Biom file
FROGSFUNC_step1_placeseqs: report.html (html)
FROGSFUNC_step1_placeseqs: frogsfunc_placeseqs_tree.nwk (nhx)
FROGSFUNC_step1_placeseqs: frogsfunc_placeseqs_excluded.tsv (tsv)
FROGSFUNC_step1_placeseqs: frogsfunc_placeseqs.fasta (fasta)
FROGSFUNC_step1_placeseqs: frogsfunc_placeseqs_closests_ref_sequences.txt (tsv)
FROGSFUNC_step1_placeseqs: frogsfunc_placeseqs.biom (biom1)
FROGSFUNC_step2_copynumbers
Biom file
Tree file
FROGSFUNC_step2_copynumbers: report.html (html)
FROGSFUNC_step2_copynumbers: frogsfunc_copynumbers_predicted_functions.tsv (tabular)
FROGSFUNC_step2_copynumbers: frogsfunc_copynumbers_marker.tsv (tabular)
FROGS Pre-process
TAR archive file
FROGS Pre-process: dereplicated.fasta (fasta)
FROGS Pre-process: count.tsv (tabular)
FROGS Pre-process: report.html (html)
FROGS Clustering swarm
Sequences file
Count file
FROGS Clustering swarm: seed_sequences.fasta (fasta)
FROGS Clustering swarm: clustering_abundance.biom (biom1)
FROGS Clustering swarm: swarms_composition.tsv (tsv)
FROGS Remove chimera
Sequences file (format: FASTA)
Abundance file (format: BIOM)
FROGS Remove chimera: non_chimera.fasta (fasta)
FROGS Remove chimera: non_chimera_abundance.biom (biom1)
FROGS Remove chimera: report.html (html)
FROGS OTU Filters
Sequences file
Abundance file
FROGS OTU Filters: ...abundance.biom
FROGS OTU Filters: otuFilter_sequences.fasta (fasta)
FROGS OTU Filters: excluded.tsv (tabular)
FROGS OTU Filters: report.html (html)
FROGS Affiliation OTU
Sequence file
Abundance file
FROGS Affiliation OTU: affiliation_abundance.biom (biom1)
FROGS Affiliation OTU: report.html (html)
FROGS Tree
Sequence file
Biom file
FROGS Tree: tree.nwk (nhx)
FROGS Tree: report.html (html)
FROGSFUNC_step3_functions
Biom file
Function file
Marker file
FROGSFUNC_step3_functions: report.html (html)
FROGSFUNC_step3_functions: frogsfunc_functions_marker_norm.tsv (tabular)
FROGSFUNC_step3_functions: frogsfunc_functions_weighted_nsti.tsv (tabular)
FROGSFUNC_step3_functions: frogsfunc_functions_excluded.tsv (tabular)
FROGSFUNC_step3_functions: frogsfunc_functions_unstrat.tsv (tabular)
FROGS BIOM to TSV
Abundance file
Sequences file (optional)
FROGS BIOM to TSV: abundance.tsv (tabular)
FROGS BIOM to TSV: multi-affiliations.tsv (tabular)
FROGS BIOM to std BIOM
Abundance file
FROGS BIOM to std BIOM: abundance.standard.biom (biom1)
FROGS BIOM to std BIOM: all_blast_details.tsv (tabular)
FROGS TSV_to_BIOM
Abundance TSV File
Multi_affiliation TSV File
FROGS TSV_to_BIOM: abundance.biom (biom1)
FROGSFUNC_step4_pathways
Function abundance file
FROGSFUNC_step4_pathways: report.html (html)
FROGSFUNC_step4_pathways: frogsfunc_pathways_unstrat.tsv (tsv)
FROGSSTAT Phyloseq Import Data
Abundance biom file with taxonomical metadata (format: BIOM)
Sample file (format: TSV)
Tree file (format: Newick)
FROGSSTAT Phyloseq Import Data: data.Rdata (rdata)
FROGSSTAT Phyloseq Import Data: report.nb.html (html)
FROGSSTAT Phyloseq Composition Visualisation
Phyloseq object (format: rdata)
FROGSSTAT Phyloseq Composition Visualisation: composition.nb.html (html)
FROGSSTAT Phyloseq Sample Clustering
Phyloseq object (format: RData)
The beta diversity distance matrix file
FROGSSTAT Phyloseq Sample Clustering: clustering.nb.html (html)
FROGSSTAT Phyloseq Alpha Diversity
Phyloseq object (format: RData)
FROGSSTAT Phyloseq Alpha Diversity: alpha_diversity.tsv (tsv)
FROGSSTAT Phyloseq Alpha Diversity: alpha_diversity.nb.html (html)
FROGSSTAT Phyloseq Structure Visualisation
Phyloseq object (format: rdata)
The beta diversity distance matrix file
FROGSSTAT Phyloseq Structure Visualisation: structure.nb.html (html)
FROGSSTAT Phyloseq Beta Diversity
Phyloseq object (format: RData)
FROGSSTAT Phyloseq Beta Diversity: beta_diversity.nb.html (html)
FROGSSTAT Phyloseq Multivariate Analysis Of Variance
Phyloseq object (format: RData)
The beta diversity distance matrix file
FROGSSTAT Phyloseq Multivariate Analysis Of Variance: manova.nb.html (html)
FROGSSTAT DESeq2 Preprocess
Phyloseq object
FROGSSTAT DESeq2 Preprocess: dds.Rdata (rdata)
FROGSSTAT Deseq2 Visualisation
Phyloseq object (format: RData)
DESeq2 object (format: RData)
...nb.html

ITS

Post-process

HTML reports

Basic tools

29 tools in total

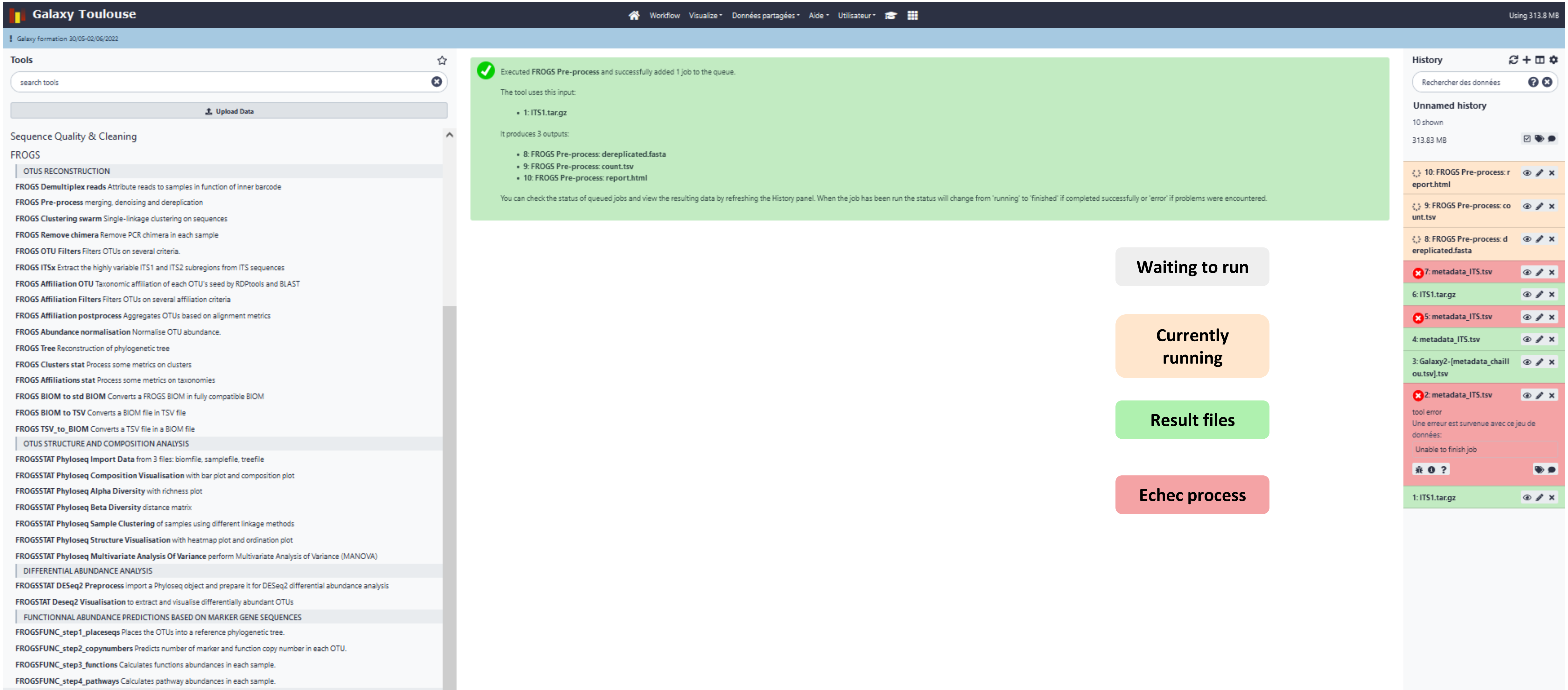
utilities

Functionnal inference

statistics

Differential analysis

FROGS Tools for Bioinfomatics analyses



Sequence Quality & Cleaning

FROGS

OTUS RECONSTRUCTION

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

FROGS Pre-process merging, denoising and dereplication

FROGS Clustering swarm Single-linkage clustering on sequences

FROGS Remove chimera Remove PCR chimera in each sample

FROGS OTU Filters Filters OTUs on several criteria.

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

FROGS Affiliation OTU Taxonomic affiliation of each OTU's seed by RDPtools and BLAST

FROGS Affiliation Filters Filters OTUs on several affiliation criteria

FROGS Affiliation postprocess Aggregates OTUs based on alignment metrics

FROGS Abundance normalisation Normalise OTU abundance.

FROGS Tree Reconstruction of phylogenetic tree

FROGS Clusters stat Process some metrics on clusters

FROGS Affiliations 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

OTUS STRUCTURE AND COMPOSITION ANALYSIS

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)

DIFFERENTIAL ABUNDANCE ANALYSIS

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

FROGSTAT Deseq2 Visualisation to extract and visualise differentially abundant OTUs

FUNCTIONNAL ABUNDANCE PREDICTIONS BASED ON MARKER GENE SEQUENCES

FROGSFUNC_step1_placeseqs Places the OTUs into a reference phylogenetic tree.

FROGSFUNC_step2_copynumbers Predicts number of marker and function copy number in each OTU.

FROGSFUNC_step3_functions Calculates functions abundances in each sample.

FROGSFUNC_step4_pathways Calculates pathway abundances in each sample.

✓ Executed **FROGS Pre-process** and successfully added 1 job to the queue.

The tool uses this input:

- **1: ITS1.tar.gz**

It produces 3 outputs:

- **8: FROGS Pre-process: dereplicated.fasta**
- **9: FROGS Pre-process: count.tsv**
- **10: FROGS Pre-process: report.html**

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

Unnamed history

10 shown

313.83 MB

- 10: FROGS Pre-process: report.html
- 9: FROGS Pre-process: count.tsv
- 8: FROGS Pre-process: dereplicated.fasta
- ☒ 7: metadata_ITS.tsv
- 6: ITS1.tar.gz
- ☒ 5: metadata_ITS.tsv
- 4: metadata_ITS.tsv
- 3: Galaxy2-[metadata_chaillou.tsv].tsv
- ☒ 2: metadata_ITS.tsv

  tool error

  Une erreur est survenue avec ce jeu de données:

  Unable to finish job
- 1: ITS1.tar.gz

Waiting to run

Currently running

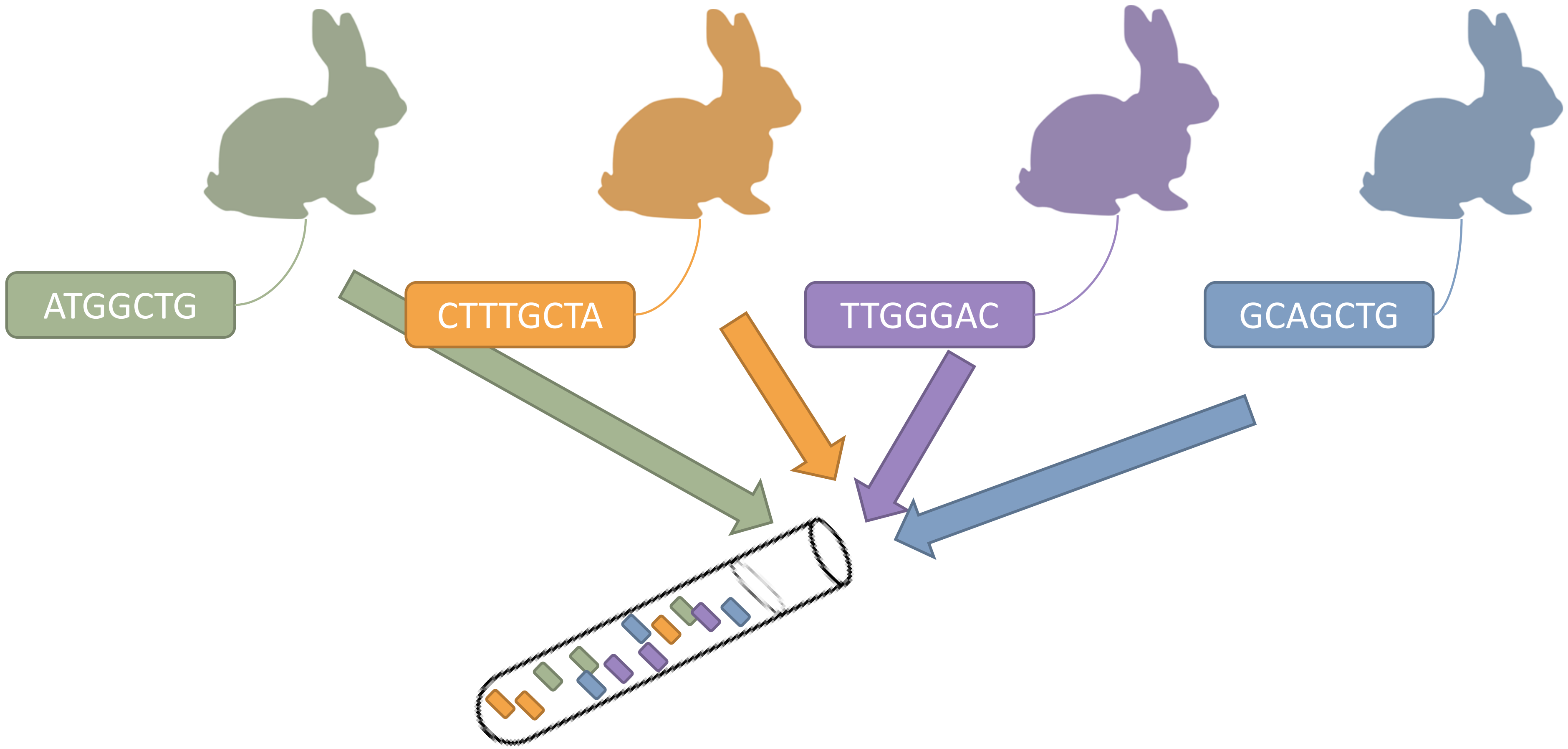
Result files

Echec process

skip

Demultiplexing tool

Barcoding ?



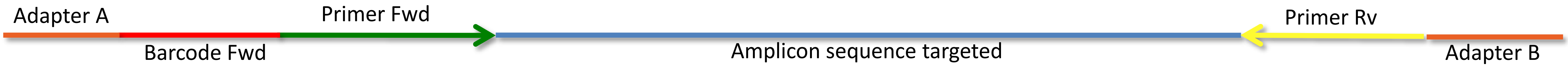
Demultiplexing

Sequence demultiplexing in function of barcode sequences :

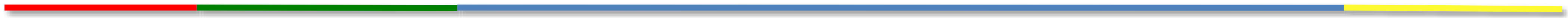
- In forward
- In reverse
- In forward and reverse

Remove unbarcoded or ambiguous sequences

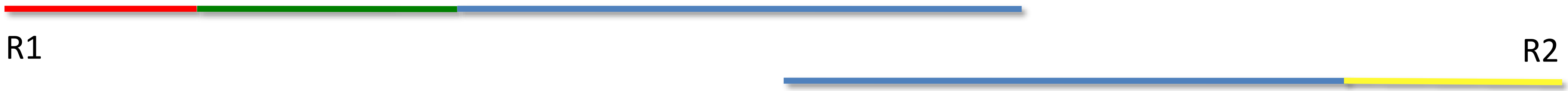
Demultiplexing forward



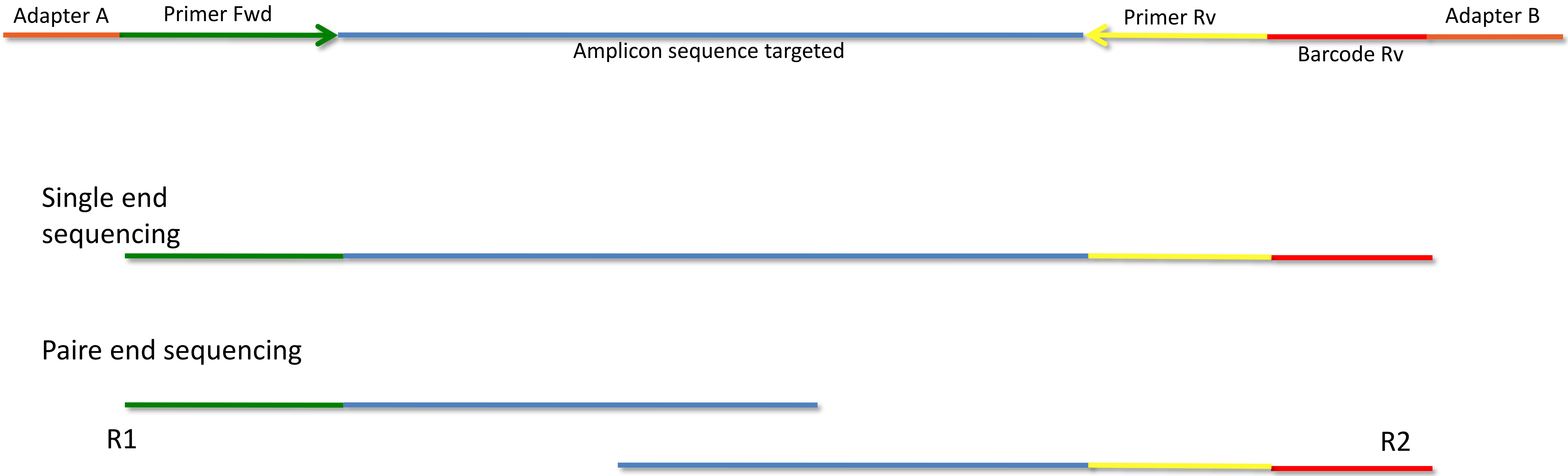
Single-end sequencing



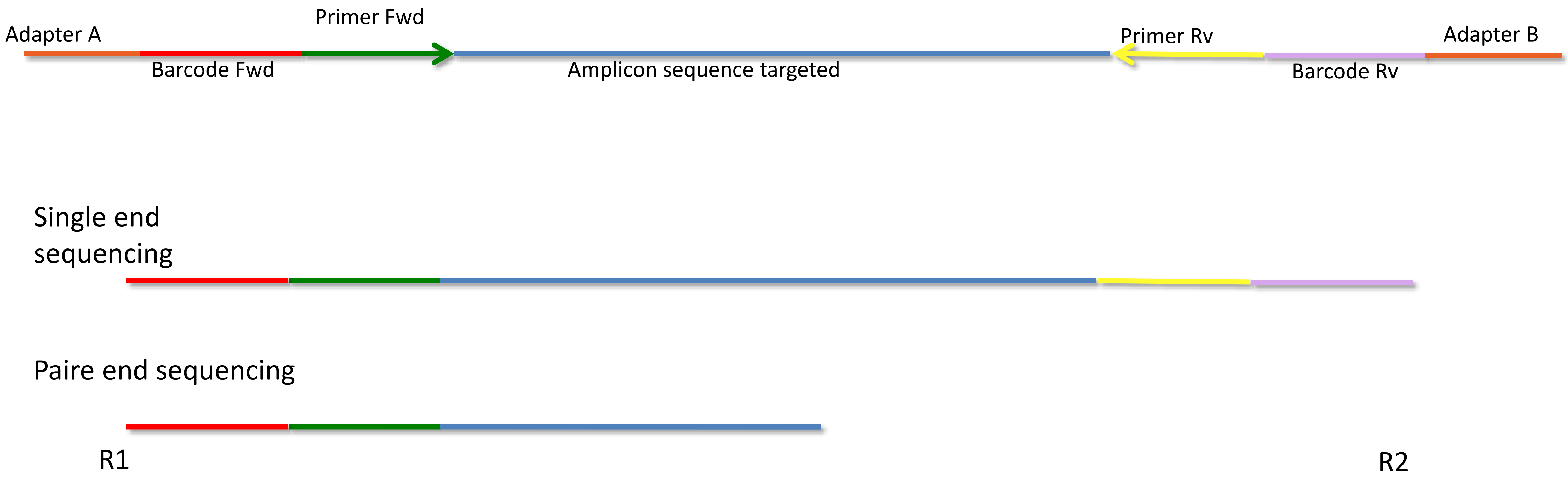
Paired-end sequencing



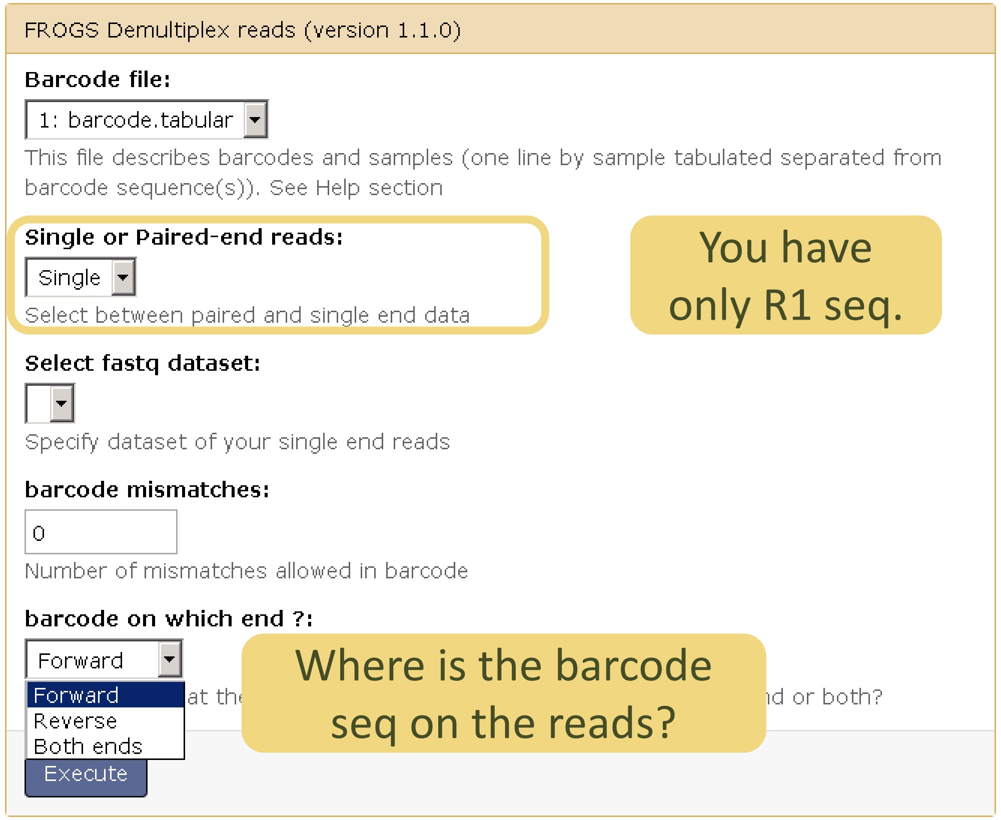
Demultiplexing reverse



Demultiplexing forward and reverse



The tool parameters depend on the input data type



FROGS Demultiplex reads (version 1.1.0)

Barcode file:
1: barcode.tabular
This file describes barcodes and samples (one line by sample tabulated separated from barcode sequence(s)). See Help section

Single or Paired-end reads:
Single
Select between paired and single end data

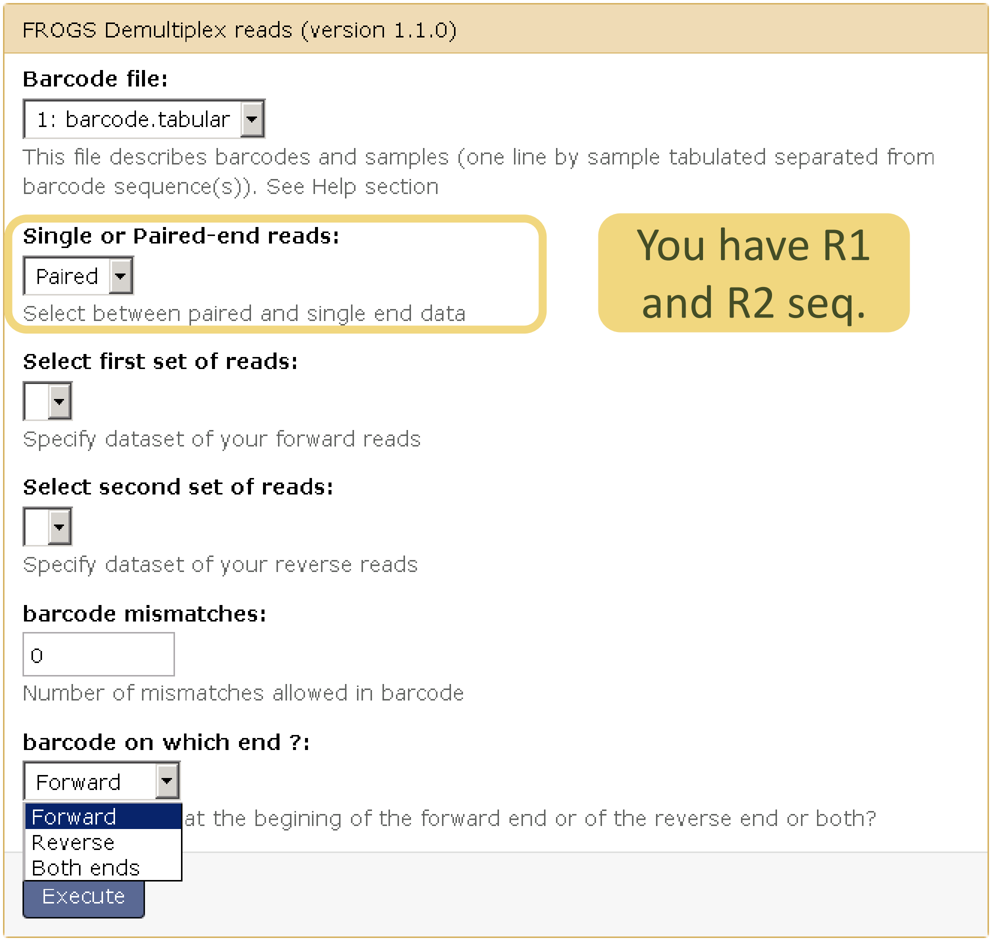
Select fastq dataset:
Specify dataset of your single end reads

barcode mismatches:
0
Number of mismatches allowed in barcode

barcode on which end ?:
Forward
at the … d or both?

You have only R1 seq.

Where is the barcode seq on the reads?



FROGS Demultiplex reads (version 1.1.0)

Barcode file:
1: barcode.tabular
This file describes barcodes and samples (one line by sample tabulated separated from barcode sequence(s)). See Help section

Single or Paired-end reads:
Paired
Select between paired and single end data

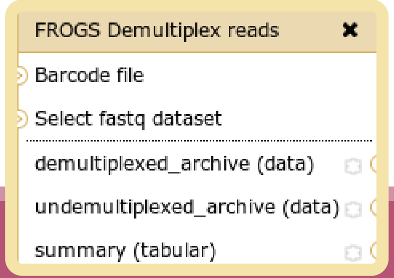
Select first set of reads:
Specify dataset of your forward reads

Select second set of reads:
Specify dataset of your reverse reads

barcode mismatches:
0
Number of mismatches allowed in barcode

barcode on which end ?:
Forward
at the begining of the forward end or of the reverse end or both?

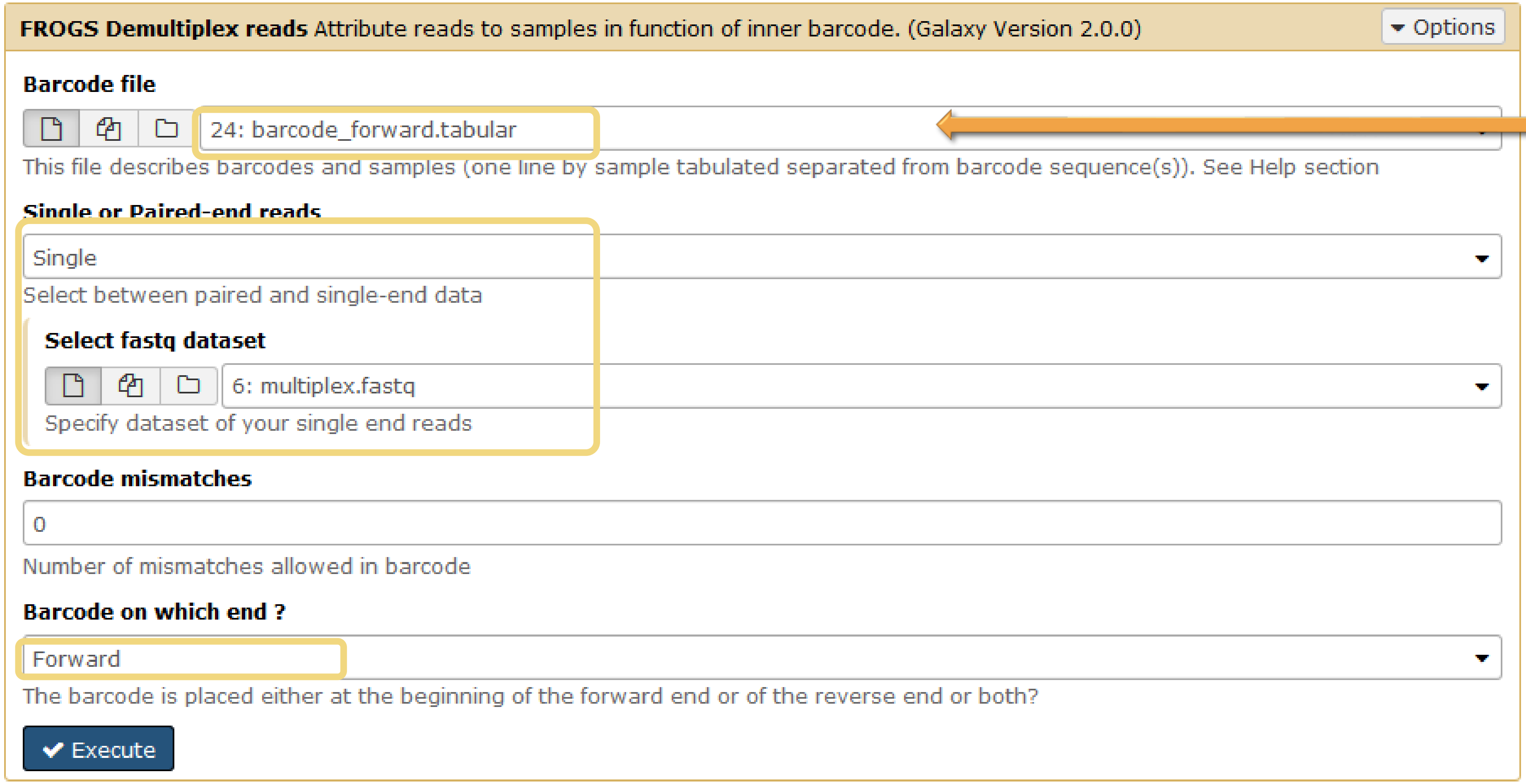
You have R1 and R2 seq.



FROGS Demultiplex reads

- Barcode file
- Select fastq dataset
- demultiplexed_archive (data)
- undemultiplexed_archive (data)
- summary (tabular)

Demultiplexing



FROGS Demultiplex reads Attribute reads to samples in function of inner barcode. (Galaxy Version 2.0.0)

Barcode file

24: barcode_forward.tabular

This file describes barcodes and samples (one line by sample tabulated separated from barcode sequence(s)). See Help section

Single or Paired-end reads

Single

Select between paired and single-end data

Select fastq dataset

6: multiplex.fastq

Specify dataset of your single end reads

Barcode mismatches

0

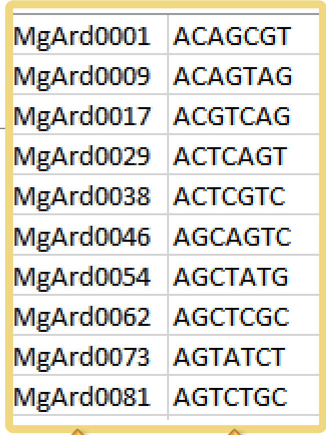
Number of mismatches allowed in barcode

Barcode on which end ?

Forward

The barcode is placed either at the beginning of the forward end or of the reverse end or both?

Input example



MgArd0001	ACAGCGT
MgArd0009	ACAGTAG
MgArd0017	ACGTCAG
MgArd0029	ACTCAGT
MgArd0038	ACTCGTC
MgArd0046	AGCAGTC
MgArd0054	AGCTATG
MgArd0062	AGCTCGC
MgArd0073	AGTATCT
MgArd0081	AGTCTGC

if index is in only at forward: tabular file with 2 columns sample names + barcodes

Advices

For your own data

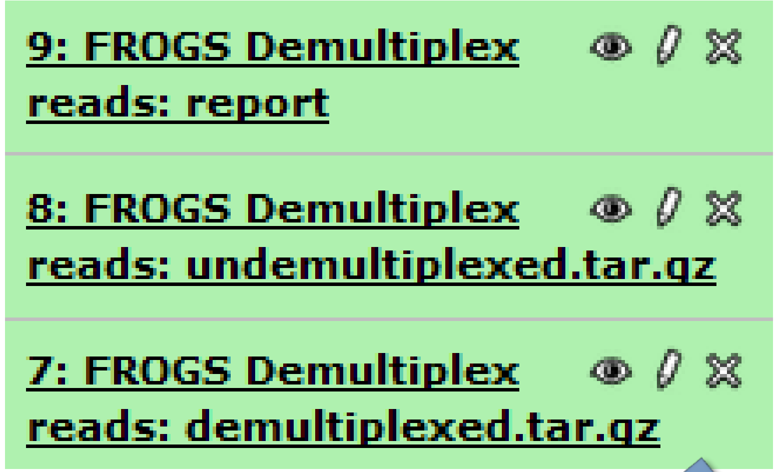
- Do not forget to indicate barcode sequence as they are in the fastq sequence file, especially if you have data multiplexed via the reverse strand.

- For the mismatch threshold, we advised you to let the threshold to 0, and if you are not satisfied by the result, try with 1. The number of mismatch depends on the length of the barcode, but often those sequences are very short so 1 mismatch is already more than the sequencing error rate.

- If you have different barcode lengths, you must demultiplex your data in different times beginning by the longest barcode set and used the "unmatched" or "ambiguous" sequence with smaller barcode and so on.

- If you have Roche 454 sequences in sff format, you must convert them with some program like sff2fastq

Outputs

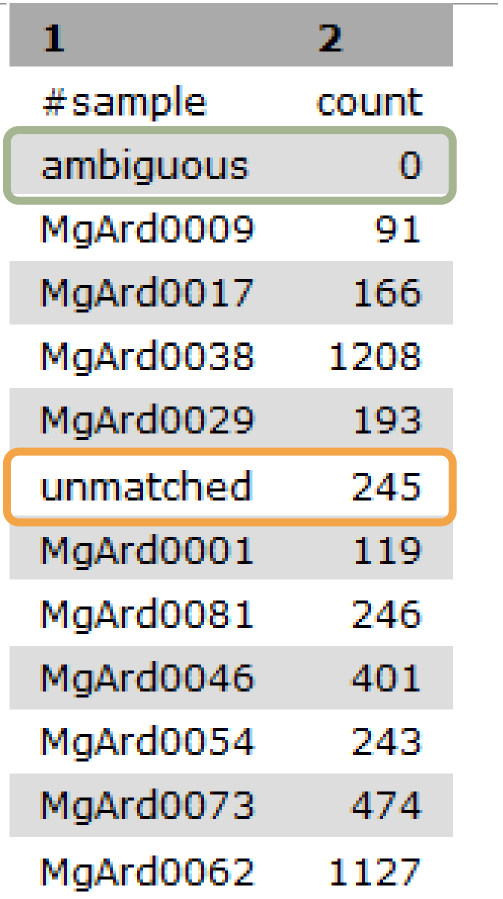


9: FROGS Demultiplex reads: report

8: FROGS Demultiplex reads: undemultiplexed.tar.gz

7: FROGS Demultiplex reads: demultiplexed.tar.gz

A tar archive is created by grouping one (or a pair of) fastq file per sample with the names indicated in the first column of the barcode tabular file.



1	2
#sample	count
ambiguous	0
MgArd0009	91
MgArd0017	166
MgArd0038	1208
MgArd0029	193
unmatched	245
MgArd0001	119
MgArd0081	246
MgArd0046	401
MgArd0054	243
MgArd0073	474
MgArd0062	1127

With barcode mismatches >1 sequence can corresponding to several samples.
Sequences that match at only one sample are affected to this sample but
the others (ambiguous) are not re-affected to a sample.

Sequences without known barcode.
So these sequences are non-affected to a sample.

Format: Barcode

BARCODE FILE is expected to be tabulated:

- first column corresponds to the sample name (unique, without space)
- second to the forward sequence barcode used (None if only reverse barcode)
- optional third is the reverse sequence barcode (optional)

Take care to indicate sequence barcode in the strand of the read, so you may need to reverse complement the reverse barcode sequence. Barcode sequence must have the same length.

Example of barcode file.
The last column is optional, like this, it describes sample multiplexed by both fragment ends.

```
MgArd00001          ACAGCGT     ACGTACA
```

Format : FastQ

FASTQ : Text file describing biological sequence in 4 lines format:

- first line start by "@" correspond to the sequence identifier and optionally the sequence description. "@Sequence_1 description1"
- second line is the sequence itself. "ACAGC"
- third line is a "+" following by the sequence identifier or not depending on the version
- fourth line is the quality sequence, one code per base. The code depends on the version and the sequencer

```
@HNHOSKD01ALD0H
ACAGCGTCAGAGGGTACCAGTCAGCCATGACGTAGCACGTACA
+
CCCFFFFFFHHHHHJJIJJJHHFF@DEDDDDDDD@CDDDDACDD
```

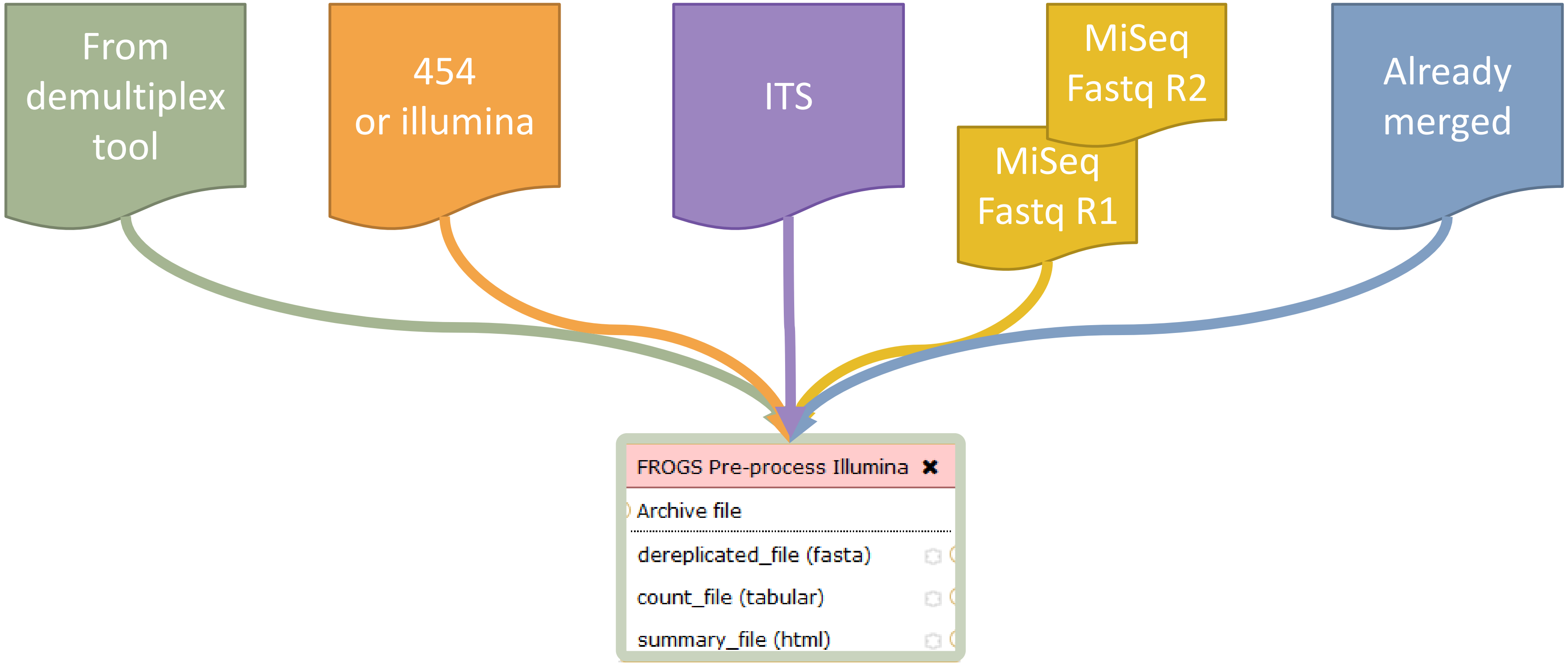
How it works ?

For each sequence or sequence pair the sequence fragment at the beginning (forward multiplexing) of the (first) read or at the end (reverse multiplexing) of the (second) read will be compare to all barcode sequence.

If this fragment is equal (with less or equal mismatch than the threshold) to one (and only one) barcode, the fragment is trimmed and the sequence will be attributed to the corresponding sample.

Finally fastq files (or pair of fastq files) for each sample are included in an archive, and a report describes how many sequence are attributed for each sample.

Pre-process tool



From demultiplex tool

454 or illumina

ITS

MiSeq Fastq R2

MiSeq Fastq R1

Already merged

FROGS Pre-process Illumina ✖

Archive file

dereplicated_file (fasta)

count_file (tabular)

summary_file (html)

Pre-process

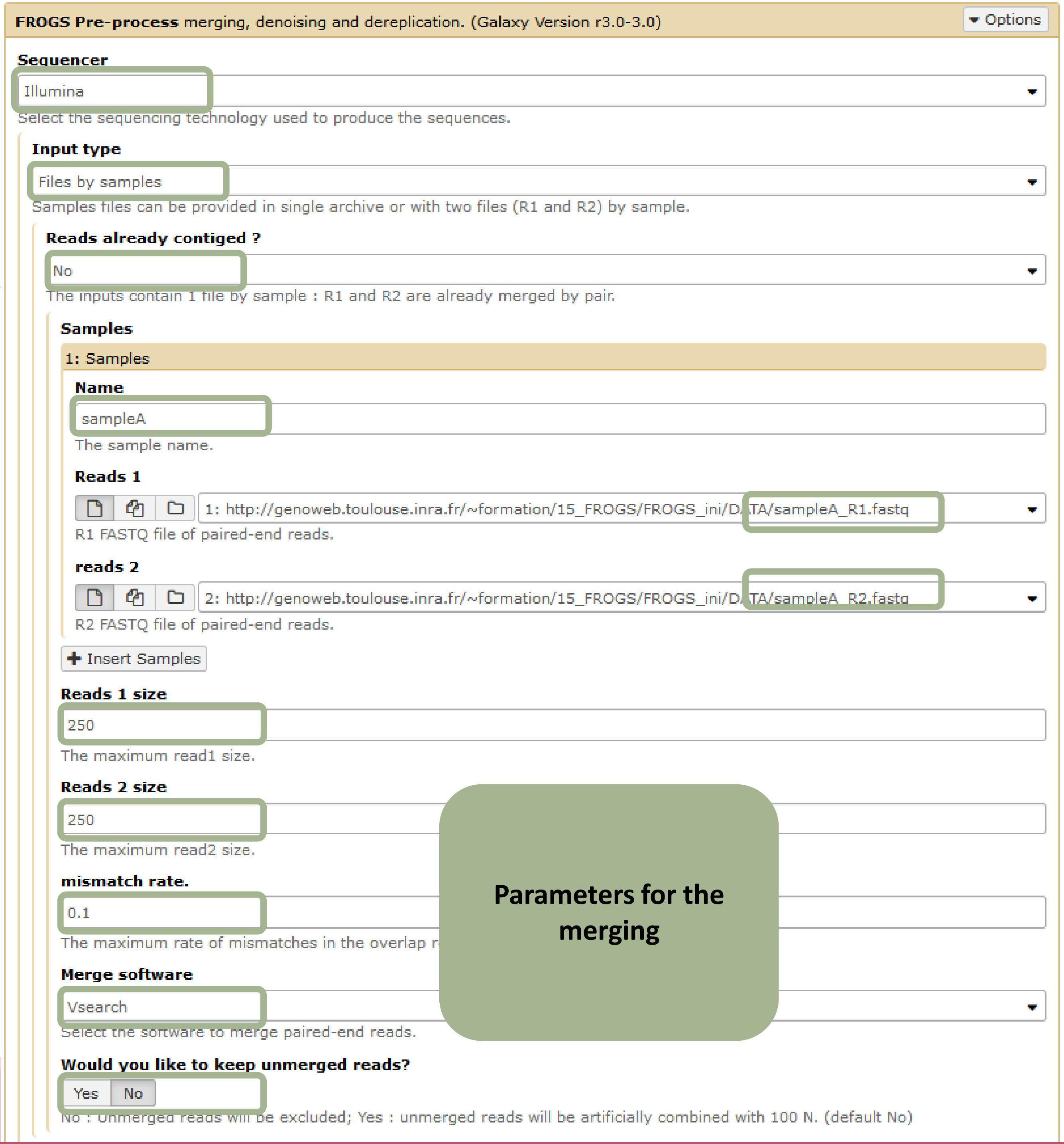
Pre-process

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

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

Example for:

- Illumina MiSeq data
- 1 sample
- Non joined



FROGS Pre-process merging, denoising and dereplication. (Galaxy Version r3.0-3.0)

Sequencer

Illumina

Select the sequencing technology used to produce the sequences.

Input type

Files by samples

Samples files can be provided in single archive or with two files (R1 and R2) by sample.

Reads already contiged ?

No

The inputs contain 1 file by sample : R1 and R2 are already merged by pair.

Samples

1: Samples

Name

sampleA

The sample name.

Reads 1

1: http://genoweb.toulouse.inra.fr/~formation/15_FROGS/FROGS_ini/DATA/sampleA_R1.fastq

R1 FASTQ file of paired-end reads.

reads 2

2: http://genoweb.toulouse.inra.fr/~formation/15_FROGS/FROGS_ini/DATA/sampleA_R2.fastq

R2 FASTQ file of paired-end reads.

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 r

Merge software

Vsearch

Select the software to merge paired-end reads.

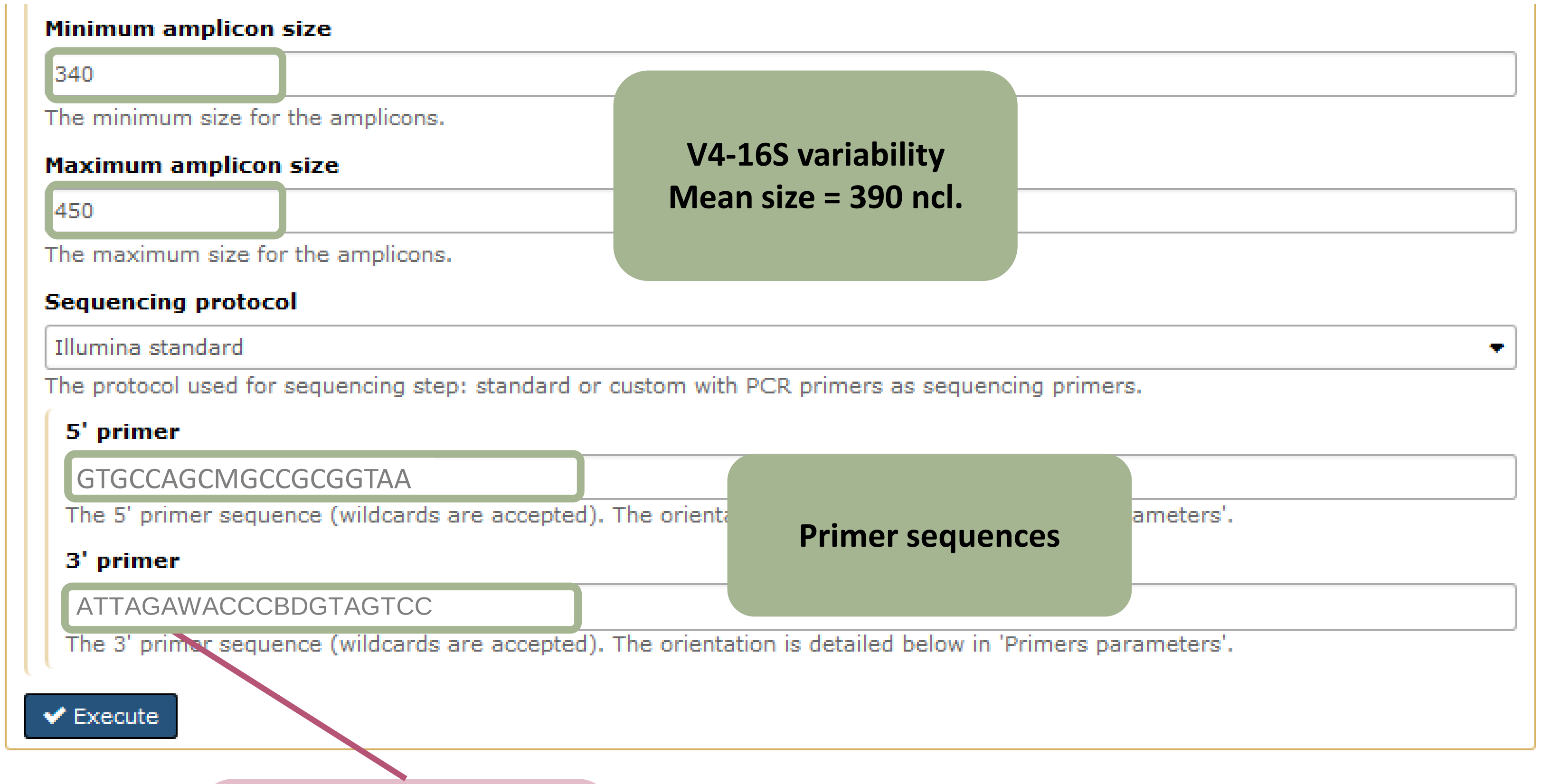
Would you like to keep unmerged reads?

Yes [✓ No]

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

Parameters for the merging

Pre-process example 1



Minimum amplicon size

340

The minimum size for the amplicons.

Maximum amplicon size

450

The maximum size for the amplicons.

Sequencing protocol

Illumina standard

The protocol used for sequencing step: standard or custom with PCR primers as sequencing primers.

5' primer

GTGCCAGCMGCCGCGGTAA

The 5' primer sequence (wildcards are accepted). The orienta ameters'.

3' primer

ATTAGAWACCCBDGTAGTCC

The 3' primer sequence (wildcards are accepted). The orientation is detailed below in 'Primers parameters'.

V4-16S variability
Mean size = 390 ncl.

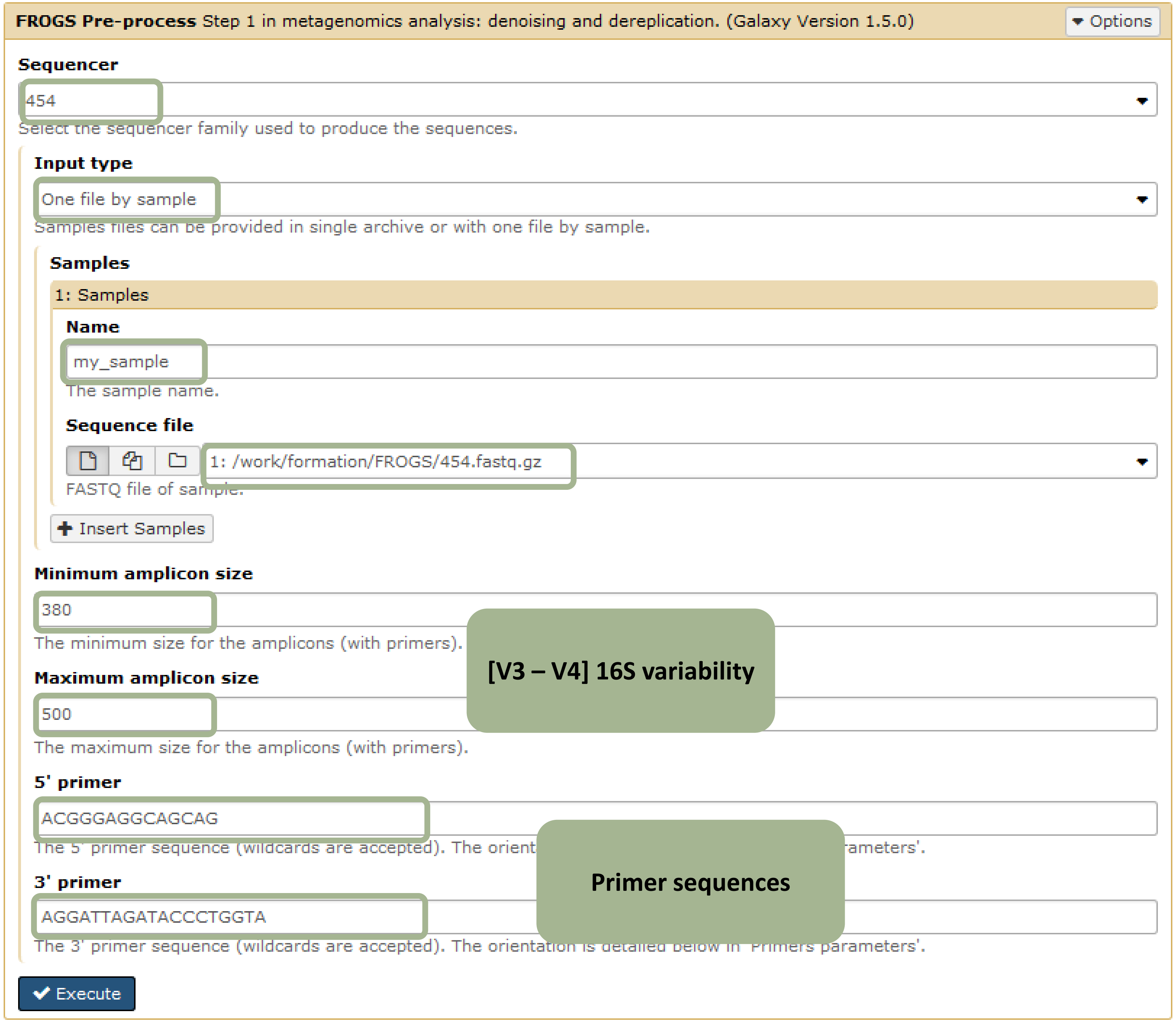
Primer sequences

degenerate primer are accepted (IUPAC code)

Pre-process example 1

Example for:

- Roche 454 data
- 1 sample
- Only one read (454 process)



FROGS Pre-process Step 1 in metagenomics analysis: denoising and dereplication. (Galaxy Version 1.5.0)

Sequencer

454

Select the sequencer family 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

my_sample

The sample name.

Sequence file

1: /work/formation/FROGS/454.fastq.gz

FASTQ file of sample.

Minimum amplicon size

380

The minimum size for the amplicons (with primers).

Maximum amplicon size

500

The maximum size for the amplicons (with primers).

5' primer

ACGGGAGGCAGCAG

The 5' primer sequence (wildcards are accepted). The orient rameters'.

3' primer

AGGATTAGATACCCTGGTA

The 3' primer sequence (wildcards are accepted). The orientation is detailed below in 'Primers parameters'.

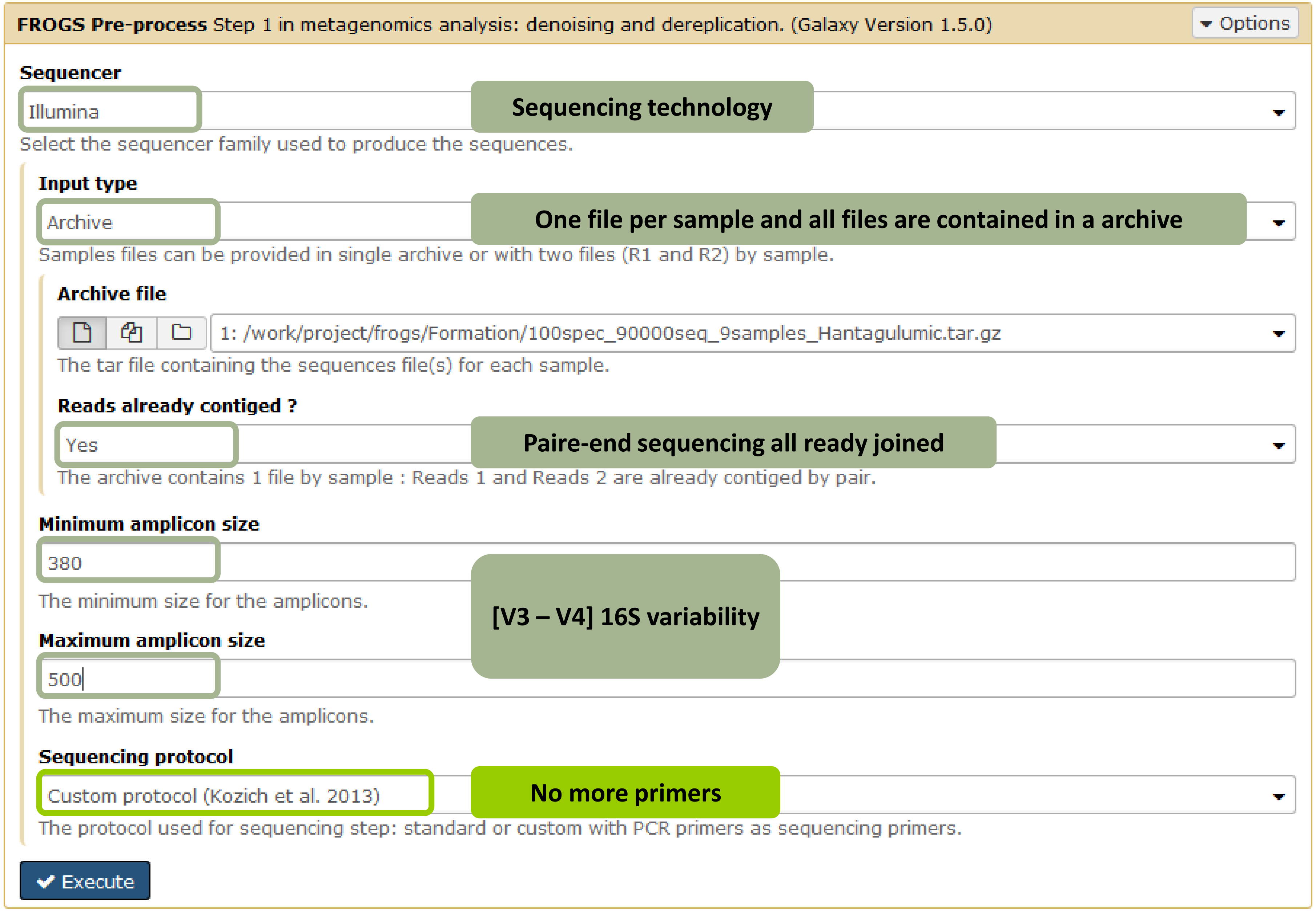
[V3 – V4] 16S variability

Primer sequences

Pre-process example 2

Example for:

- Illumina MiSeq data
- 9 samples in 1 archive
- Joined
- Without sequenced PCR primers (Kozich protocol)



FROGS Pre-process Step 1 in metagenomics analysis: denoising and dereplication. (Galaxy Version 1.5.0)

Sequencer

Illumina

Sequencing technology

Select the sequencer family used to produce the sequences.

Input type

Archive

One file per sample and all files are contained in a archive

Samples files can be provided in single archive or with two files (R1 and R2) by sample.

Archive file

1: /work/project/frogs/Formation/100spec_90000seq_9samples_Hantagulumic.tar.gz

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

Reads already contiged ?

Yes

Paire-end sequencing all ready joined

The archive contains 1 file by sample : Reads 1 and Reads 2 are already contiged by pair.

Minimum amplicon size

380

The minimum size for the amplicons.

[V3 – V4] 16S variability

Maximum amplicon size

500

The maximum size for the amplicons.

Sequencing protocol

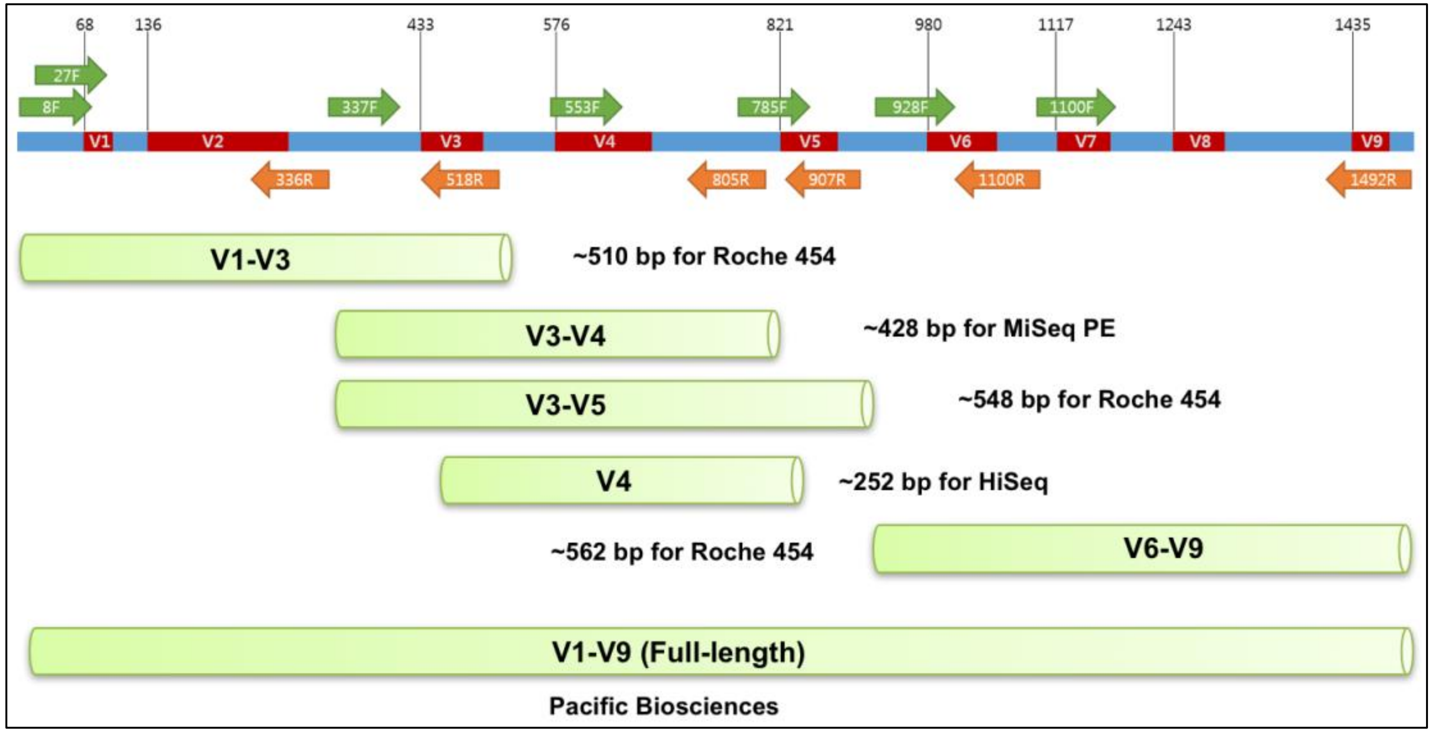
Custom protocol (Kozich et al. 2013)

No more primers

The protocol used for sequencing step: standard or custom with PCR primers as sequencing primers.

Pre-process example 3

Which primers for 16S ?



NGS platforms	16S region	PCR primers	Estimated insert size to read (E. coli)	Sequencing
Illumina MiSeq PE (Pair End)	V3V4	341F & 805R	427 bp	250 bp x 2 or 300 bp x 2
Illumina HiSeq/iSeq100 (Earth Microbiome Project)	V4	515FB & 806RB	250 bp	150 x 2

Name of primer F=forward, R=reverse	Sequence
8F	AGAGTTTGATCCTGGCTCAG
27F	AGAGTTTGATCMTGGCTCAG
336R	ACTGCTGCSYCCCGTAGGAGTCT
337F	GACTCCTACGGGAGGCWGCAG
337F	GACTCCTACGGGAGGCWGCAG
341F	CCTACGGGNGGCWGCAG
515FB	GTGYCAGCMGCCGCGGTAA
518R	GTATTACCGCGGCTGCTGG
533F	GTGCCAGCMGCCGCGGTAA
785F	GGATTAGATACCCTGGTA
805R	GACTACHVGGGTATCTAATCC
806RB	GGACTACNVGGGTWTCTAAT
907R	CCGTCAATTCCTTTRAGTTT
928F	TAAAACTYAAAKGAATTGACGGG
1100F	YAACGAGCGCAACCC
1100R	GGGTTGCGCTCGTTG
1492R	CGGTTACCTTGTTACGACTT

Cf. http://help.ezbiocloud.net/16s-rrna-and-16s-rrna-gene/

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,

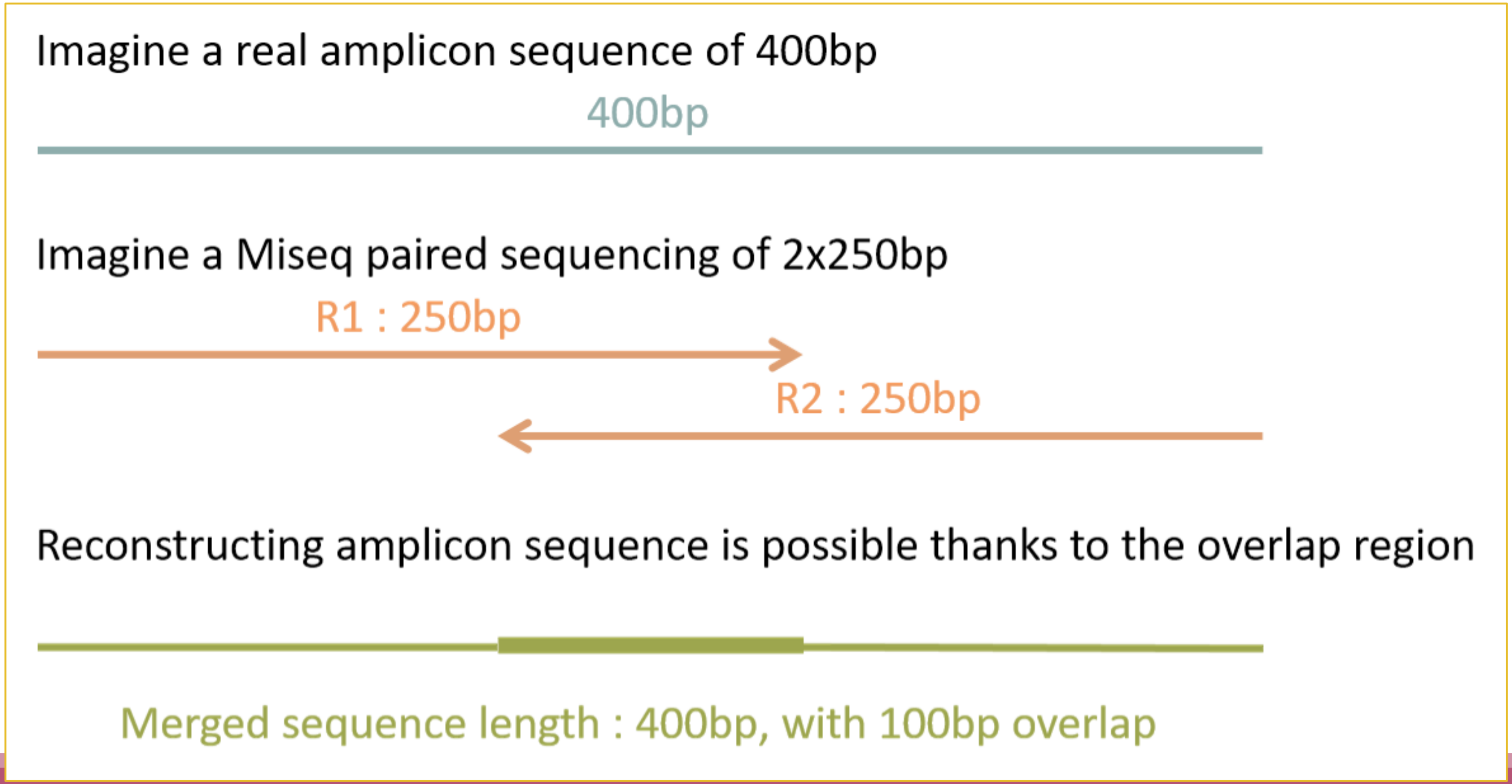
EMBnet Journal, Vol17 no1. doi : 10.14806/ej.17.1.200
Cutadapt removes adapter sequences from high-throughput sequencing reads
Marcel Martin

How work reads merging ?

WITH VSEARCH

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:



Imagine a real amplicon sequence of 400bp

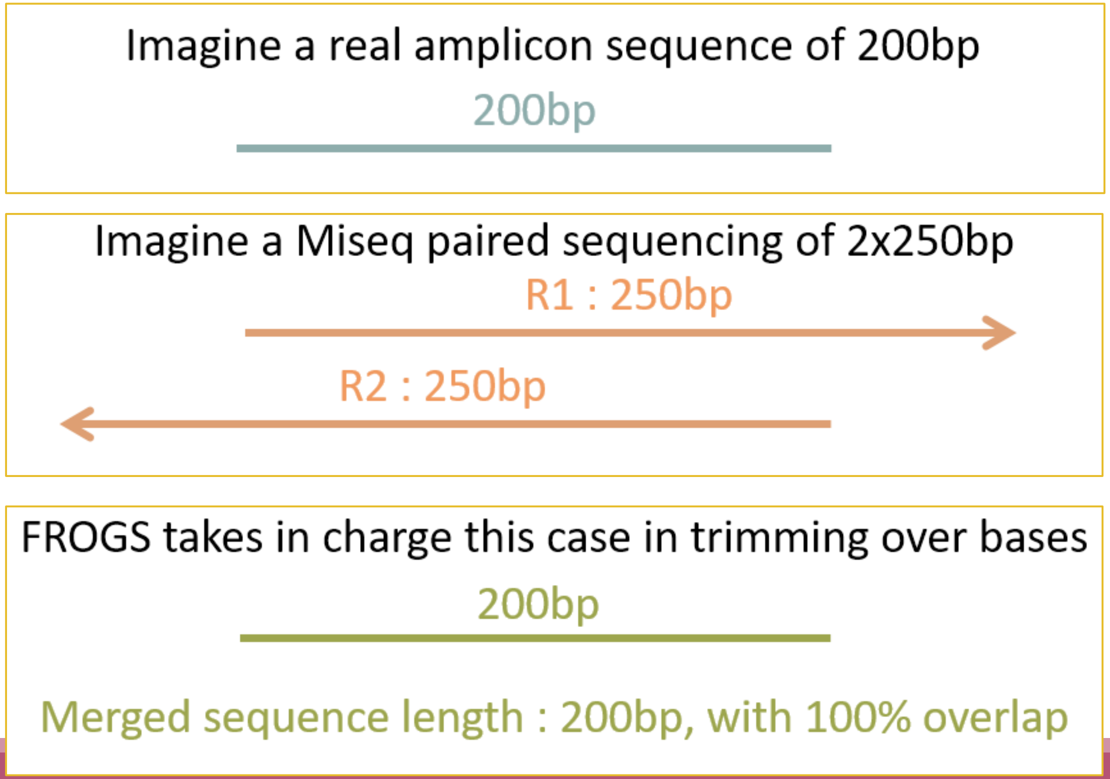
Imagine a Miseq paired sequencing of 2x250bp

Reconstructing amplicon sequence is possible thanks to the overlap region

Merged sequence length : 400bp, with 100bp overlap

The aim of Vsearch is to merge R1 with R2

Case of a sequencing of over-overlapping sequences:



Imagine a real amplicon sequence of 200bp

200bp

Imagine a Miseq paired sequencing of 2x250bp

R1 : 250bp

R2 : 250bp

FROGS takes in charge this case in trimming over bases

200bp

Merged sequence length : 200bp, with 100% overlap

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

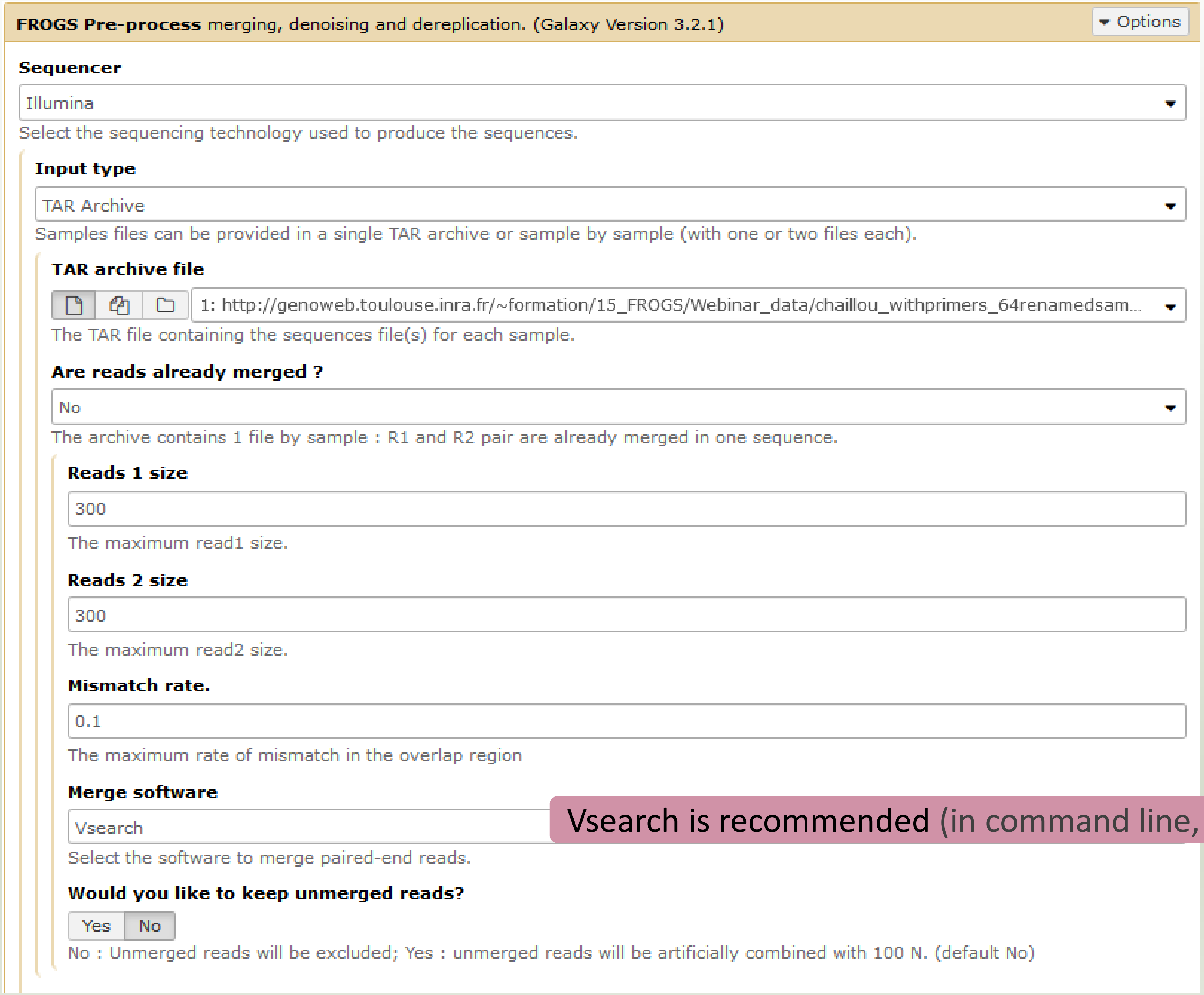
Chaillou, S. et al (2015). Origin and ecological selection of core and food-specific bacterial communities associated with meat and seafood spoilage. ISME J, 9(5):1105-1118.

16S dataset presentation:

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 Pre-process merging, denoising and dereplication. (Galaxy Version 3.2.1)

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: http://genoweb.toulouse.inra.fr/~formation/15_FROGS/Webinar_data/chaillou_withprimers_64renamedsam...

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

Are reads already merged ?

No

The archive contains 1 file by sample : R1 and R2 pair 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 mismatch in the overlap region

Merge software

Vsearch

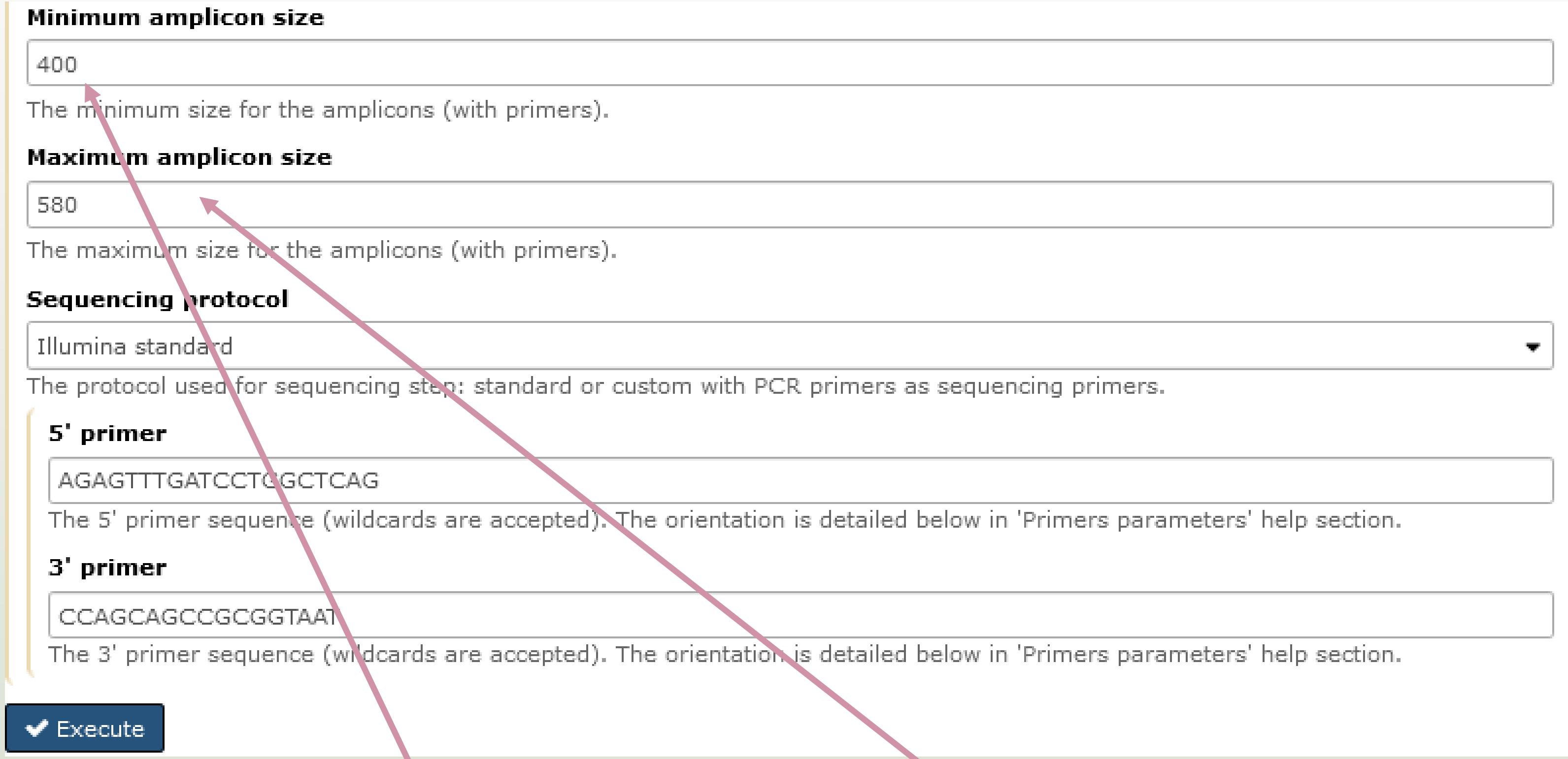
Select the software to merge paired-end reads.

Would you like to keep unmerged reads?

Yes [✓ No]

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

Vsearch is recommended (in command line, prefer pear)



Minimum amplicon size

400

The minimum size for the amplicons (with primers).

Maximum amplicon size

580

The maximum size for the amplicons (with primers).

Sequencing protocol

Illumina standard

The protocol used for sequencing step: standard or custom with PCR primers as sequencing primers.

5' primer

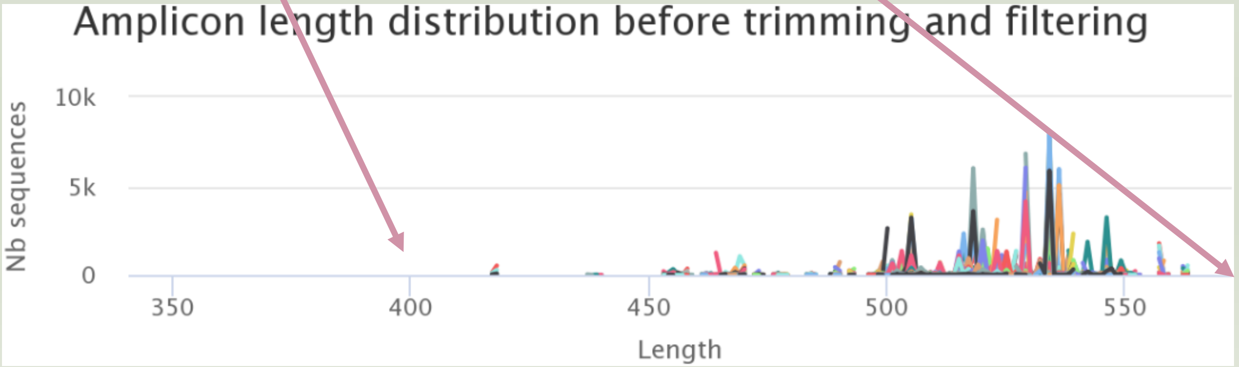
AGAGTTTGATCCTGGCTCAG

The 5' primer sequence (wildcards are accepted). The orientation is detailed below in 'Primers parameters' help section.

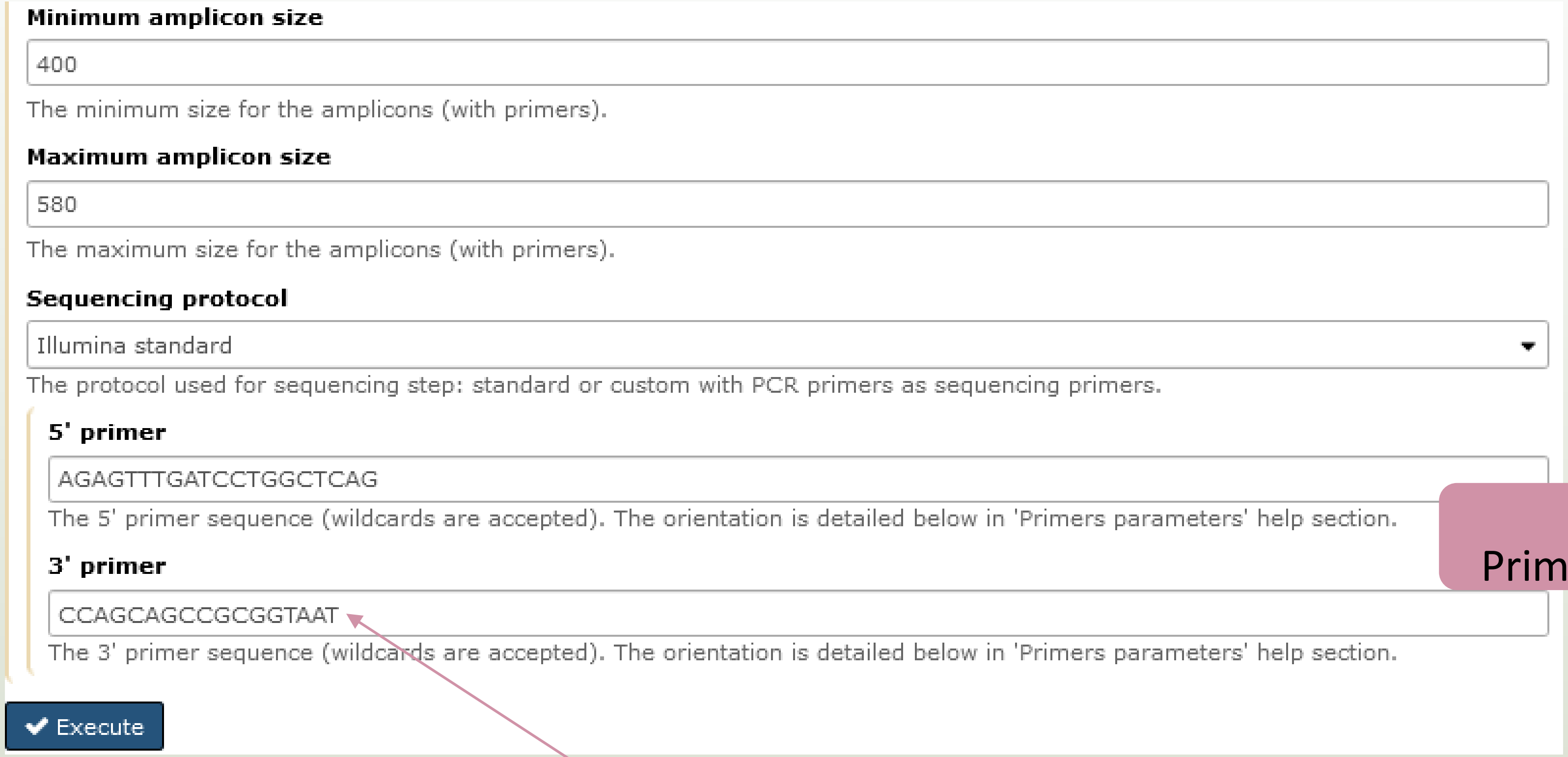
3' primer

CCAGCAGCCGCGGTAAT

The 3' primer sequence (wildcards are accepted). The orientation is detailed below in 'Primers parameters' help section.



Amplicon length distribution before trimming and filtering



Minimum amplicon size

400

The minimum size for the amplicons (with primers).

Maximum amplicon size

580

The maximum size for the amplicons (with primers).

Sequencing protocol

Illumina standard

The protocol used for sequencing step: standard or custom with PCR primers as sequencing primers.

5' primer

AGAGTTTGATCCTGGCTCAG

The 5' primer sequence (wildcards are accepted). The orientation is detailed below in 'Primers parameters' help section.

3' primer

CCAGCAGCCGCGGTAAT

The 3' primer sequence (wildcards are accepted). The orientation is detailed below in 'Primers parameters' help section.

N.B.
Primers in 5’ → 3’ sens

Ex: read R1

```
@63_0 reference=otu_00517 position=1..300
AGAGTTTGATCCTGGCTCAGgatgaacgctagcgggaggcttaacacatgcaagccgagggg
tagaattagcttgctaatttgagaccggcgcacgggtgcgtaacgcgtatgcaacttgccctactgaaaa
ggatagcccagagaaatttggattaatactttataatagactgaatggcatcatttagttttgaaagattt
atcgcagtaggataggcatgcgtaagattagatagttggtgaggtaacggctcaccaagtcgacgatct
ttagggggcctgagagggtgaacccca
```

Ex: read R2

```
@63_0 reference=otu_00517 position=1..300 errors=5%G
ATTACCGCGGCTGCTGGcacggagttagccggtgcttattcttctggtaccttcagctacttacac
gtaagtaggtttatccccagataaaagtagtttacaacccataaggccgtcatcctacacgcgggatggc
tggatcaggcttccacccattgtccaatattcctcactgctgcctcccgtaggagtctggtccgtgtctcag
taccagtgtgggggttcaccctctcaggccccctaaagatcgtcgacttggtgagccgttacctcaccaa
ctatctaatcttacgcatgcct
```

R2 primer must be reverse transcribed

Exercise

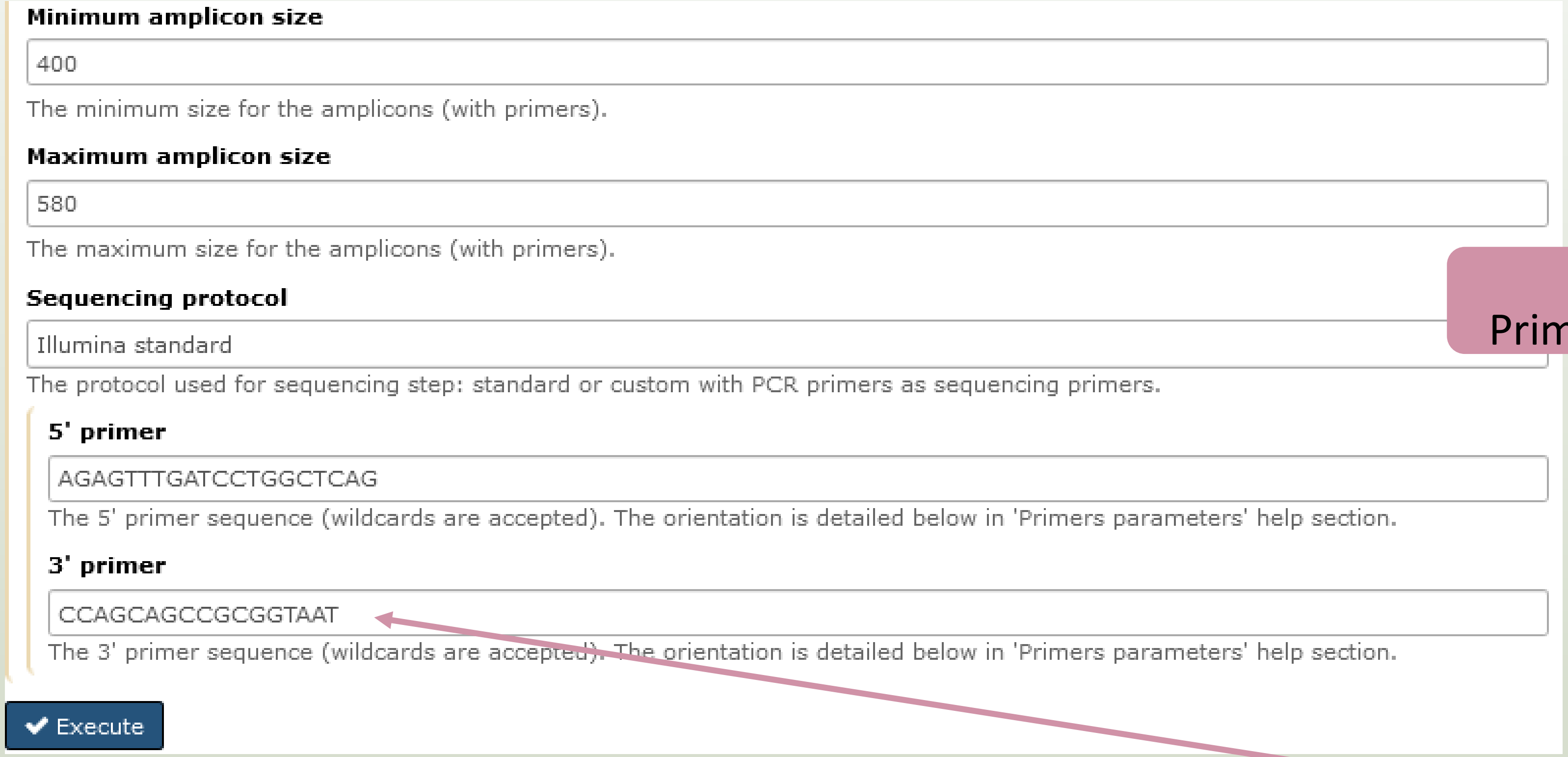
1. Do you understand how enter your primers ?
2. What is the « FROGS Pre-process: dereplicated.fasta » file ? 👁
3. What is the « FROGS Pre-process: count.tsv » file ? 👁
4. Explore the file « FROGS Pre-process: report.html » 👁
5. *Who loose a lot of sequences ?*

Exercise

6. How many sequences are there in the input file ?
7. How many sequences did not have the 5’ primer?
8. How many sequences still are after pre-processing the data?
9. How much time did it take to pre-process the data ?
10. What is the length of your merged reads before preprocessing ?
11. What can you tell about the samples, based on amplicon size distributions ?

Answer 1

Q1: Do you understand how enter your primers ?



Minimum amplicon size

400

The minimum size for the amplicons (with primers).

Maximum amplicon size

580

The maximum size for the amplicons (with primers).

Sequencing protocol

Illumina standard

The protocol used for sequencing step: standard or custom with PCR primers as sequencing primers.

5' primer

AGAGTTTGATCCTGGCTCAG

The 5' primer sequence (wildcards are accepted). The orientation is detailed below in 'Primers parameters' help section.

3' primer

CCAGCAGCCGCGGTAAT

The 3' primer sequence (wildcards are accepted). The orientation is detailed below in 'Primers parameters' help section.

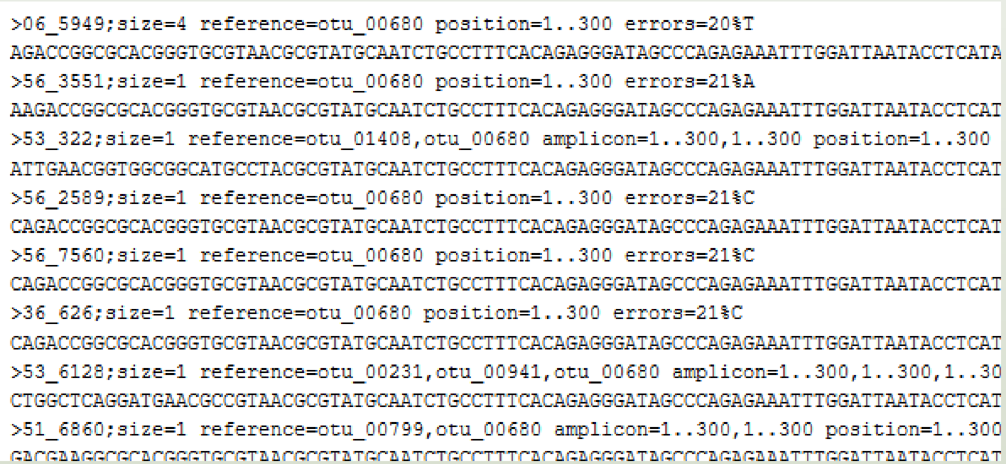
N.B.
Primers in 5’ → 3’ sens

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

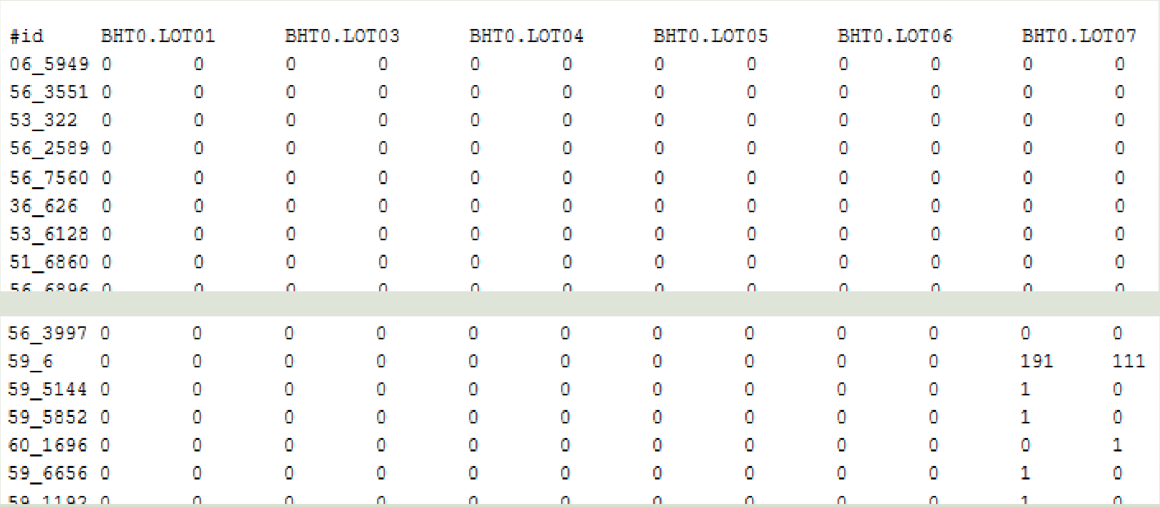
Answer 2 & 3

Q2: What is the « FROGS Pre-process: dereplicated.fasta » file ?

Q3: What is the « FROGS Pre-process: count.tsv » file ?



```
>06_5949;size=4 reference=otu_00680 position=1..300 errors=20%T
AGACCGGCGCACGGGTGCGTAACGCGTATGCAATCTGCCTTTCACAGAGGGATAGCCCAGAGAAATTTGGATTAATACCTCATA
>56_3551;size=1 reference=otu_00680 position=1..300 errors=21%A
AAGACCGGCGCACGGGTGCGTAACGCGTATGCAATCTGCCTTTCACAGAGGGATAGCCCAGAGAAATTTGGATTAATACCTCAT
>53_322;size=1 reference=otu_01408,otu_00680 amplicon=1..300,1..300 position=1..300
ATTGAACGGTGGCGGCATGCCTACGCGTATGCAATCTGCCTTTCACAGAGGGATAGCCCAGAGAAATTTGGATTAATACCTCAT
>56_2589;size=1 reference=otu_00680 position=1..300 errors=21%C
CAGACCGGCGCACGGGTGCGTAACGCGTATGCAATCTGCCTTTCACAGAGGGATAGCCCAGAGAAATTTGGATTAATACCTCAT
>56_7560;size=1 reference=otu_00680 position=1..300 errors=21%C
CAGACCGGCGCACGGGTGCGTAACGCGTATGCAATCTGCCTTTCACAGAGGGATAGCCCAGAGAAATTTGGATTAATACCTCAT
>36_626;size=1 reference=otu_00680 position=1..300 errors=21%C
CAGACCGGCGCACGGGTGCGTAACGCGTATGCAATCTGCCTTTCACAGAGGGATAGCCCAGAGAAATTTGGATTAATACCTCAT
>53_6128;size=1 reference=otu_00231,otu_00941,otu_00680 amplicon=1..300,1..300,1..30
CTGGCTCAGGATGAACGCCGTAACGCGTATGCAATCTGCCTTTCACAGAGGGATAGCCCAGAGAAATTTGGATTAATACCTCAT
>51_6860;size=1 reference=otu_00799,otu_00680 amplicon=1..300,1..300 position=1..300
GACGAAGGCGCACGGGTGCGTAACGCGTATGCAATCTGCCTTTCACAGAGGGATAGCCCAGAGAAATTTGGATTAATACCTCAT
```



```
#id     BHT0.LOT01          BHT0.LOT03          BHT0.LOT04          BHT0.LOT05          BHT0.LOT06          BHT0.LOT07
06_5949 0         0         0         0         0         0         0         0         0         0         0         0
56_3551 0         0         0         0         0         0         0         0         0         0         0         0
53_322  0         0         0         0         0         0         0         0         0         0         0         0
56_2589 0         0         0         0         0         0         0         0         0         0         0         0
56_7560 0         0         0         0         0         0         0         0         0         0         0         0
36_626  0         0         0         0         0         0         0         0         0         0         0         0
53_6128 0         0         0         0         0         0         0         0         0         0         0         0
51_6860 0         0         0         0         0         0         0         0         0         0         0         0
56_6896 0         0         0         0         0         0         0         0         0         0         0         0

56_3997 0         0         0         0         0         0         0         0         0         0         0         0
59_6    0         0         0         0         0         0         0         0         0         0         191       111
59_5144 0         0         0         0         0         0         0         0         0         0         1         0
59_5852 0         0         0         0         0         0         0         0         0         0         1         0
60_1696 0         0         0         0         0         0         0         0         0         0         0         1
59_6656 0         0         0         0         0         0         0         0         0         0         1         0
59_1192 0         0         0         0         0         0         0         0         0         0         1         0
```

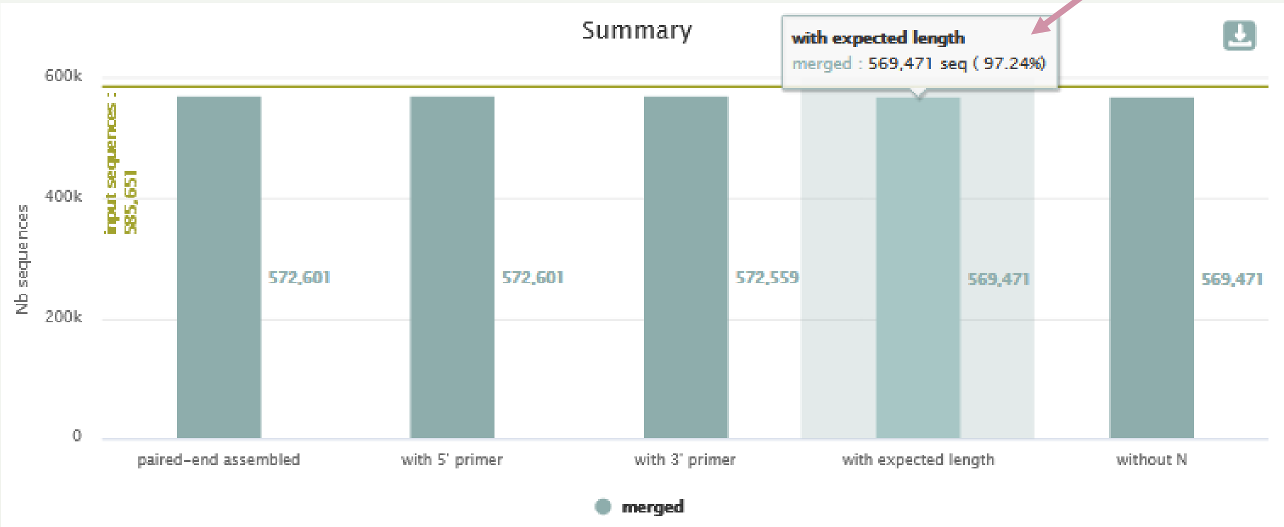
Fasta sequence of all clean and dereplicated sequence *i.e.* only one copy of each sequence is kept

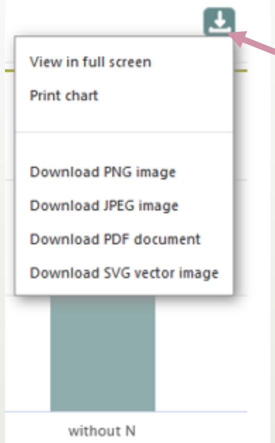
count table for each sequence in each sample

Answer 4

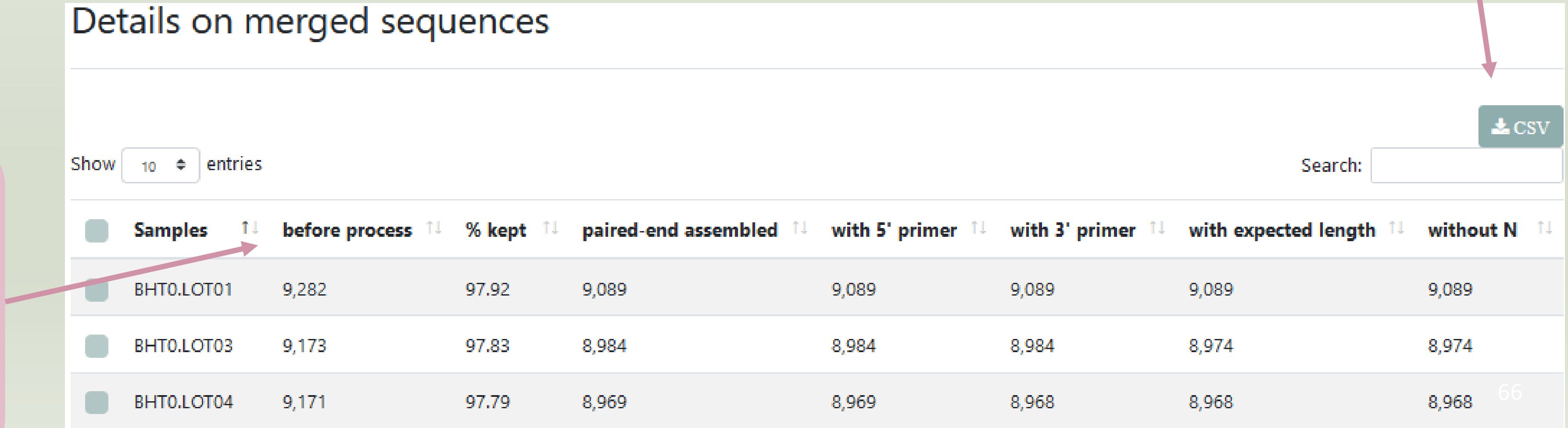
Q4: Explore the file « FROGS Pre-process: report.html »

By moving the mouse over the graphic, new information appears





You can download graphics or table in different formats



Details on merged sequences

CSV

Show 10 entries

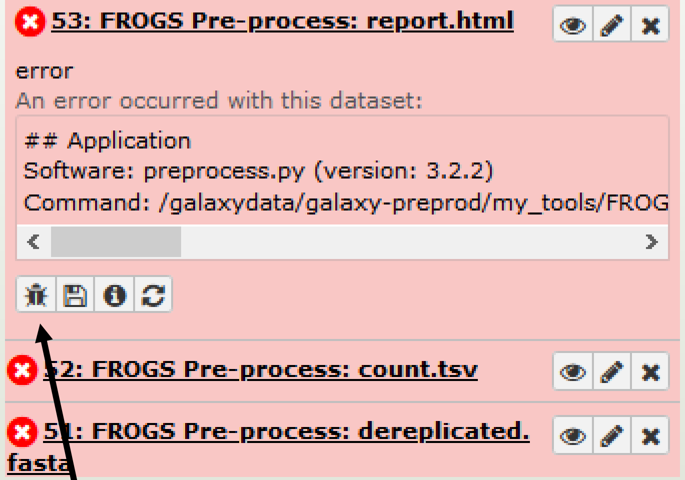
Search:

	Samples	before process	% kept	paired-end assembled	with 5' primer	with 3' primer	with expected length	without N
	BHT0.LOT01	9,282	97.92	9,089	9,089	9,089	9,089	9,089
	BHT0.LOT03	9,173	97.83	8,984	8,984	8,984	8,974	8,974
	BHT0.LOT04	9,171	97.79	8,969	8,969	8,968	8,968	8,968

You can sort data in the table by clicking on the column headers

Answer 5

Q5: Who loose a lot of sequences ?



53: FROGS Pre-process: report.html

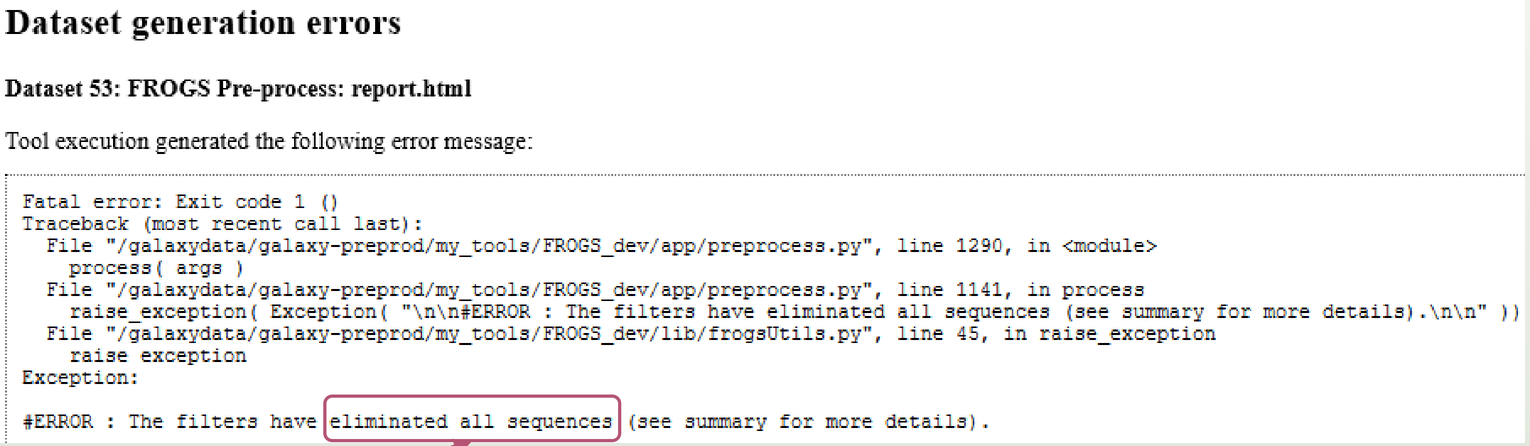
error

An error occurred with this dataset:

```
## Application
Software: preprocess.py (version: 3.2.2)
Command: /galaxydata/galaxy-preprod/my_tools/FROG
```

52: FROGS Pre-process: count.tsv

51: FROGS Pre-process: dereplicated.fasta



Dataset generation errors

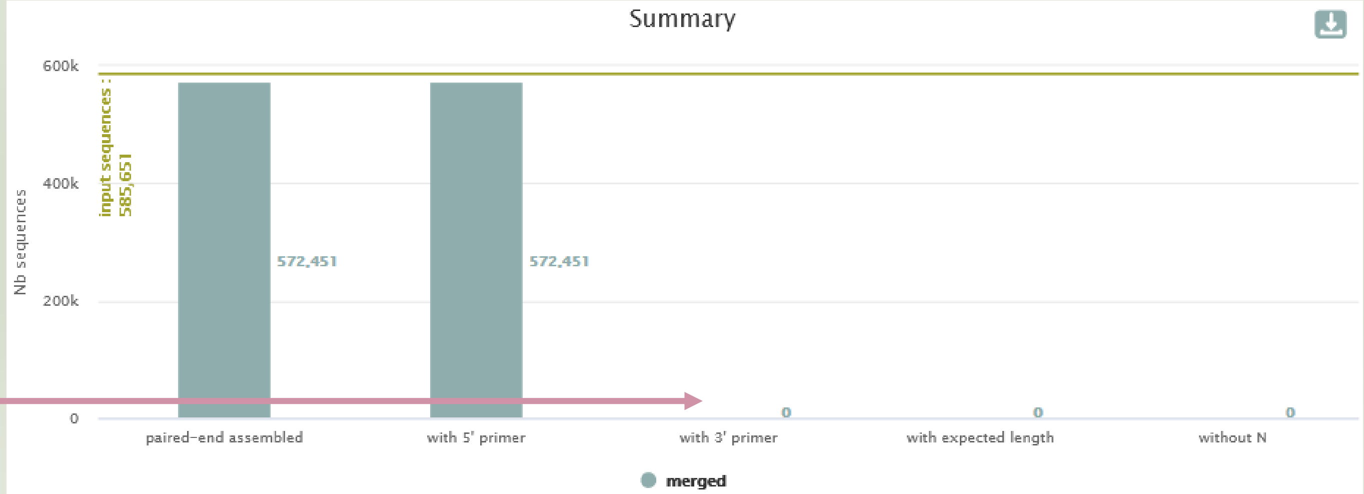
Dataset 53: FROGS Pre-process: report.html

Tool execution generated the following error message:

```
Fatal error: Exit code 1 ()
Traceback (most recent call last):
  File "/galaxydata/galaxy-preprod/my_tools/FROGS_dev/app/preprocess.py", line 1290, in <module>
    process( args )
  File "/galaxydata/galaxy-preprod/my_tools/FROGS_dev/app/preprocess.py", line 1141, in process
    raise_exception( Exception( "\n\n#ERROR : The filters have eliminated all sequences (see summary for more details).\n\n" ))
  File "/galaxydata/galaxy-preprod/my_tools/FROGS_dev/lib/frogsUtils.py", line 45, in raise_exception
    raise exception
Exception:

#ERROR : The filters have eliminated all sequences (see summary for more details).
```

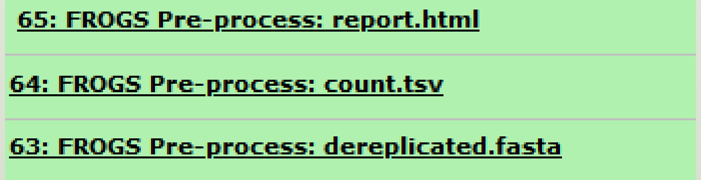
If your outputs are red, click on the bug to read the error message



it is likely that you did not enter the 3' primer in the right direction

Answer 5

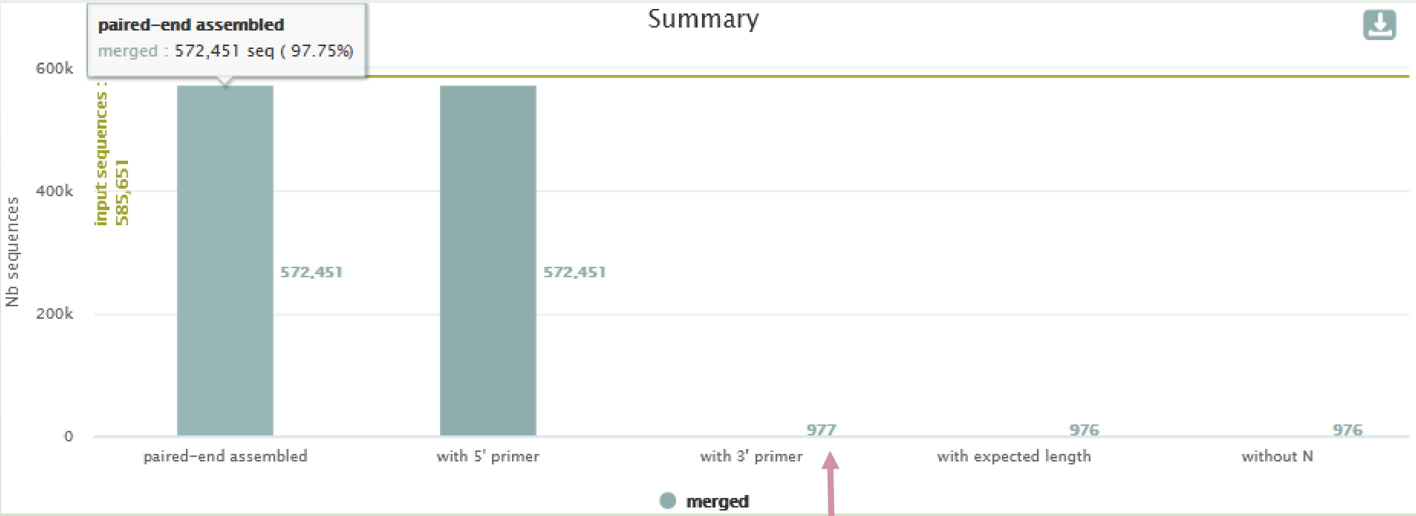
All outputs are green but check the report.html



65: FROGS Pre-process: report.html

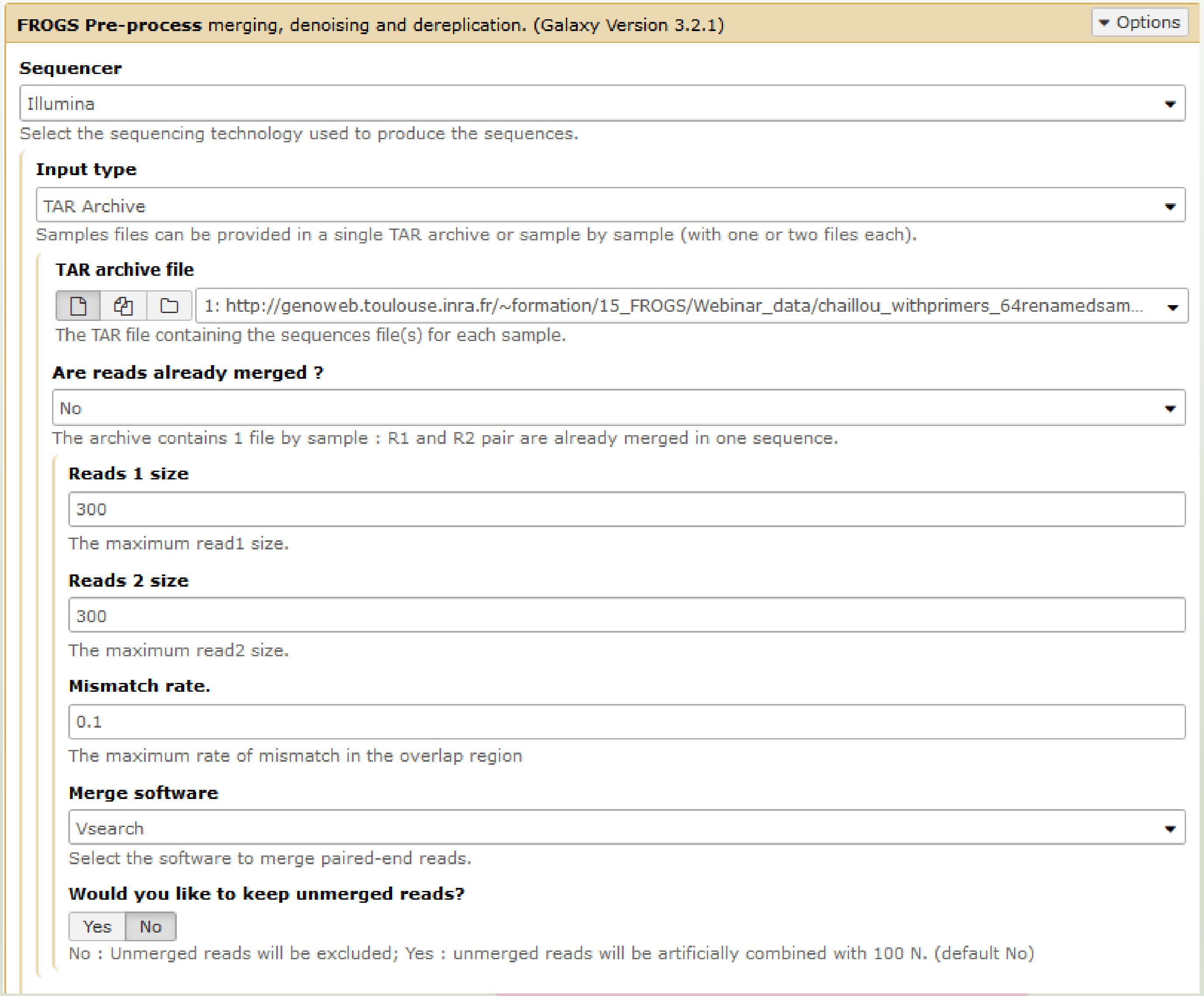
64: FROGS Pre-process: count.tsv

63: FROGS Pre-process: dereplicated.fasta



Error in 3' primer sequence. Primers must be similar with 10% of errors (~1 or 2 bases per primer)

Answer 5



FROGS Pre-process merging, denoising and dereplication. (Galaxy Version 3.2.1)

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: http://genoweb.toulouse.inra.fr/~formation/15_FROGS/Webinar_data/chaillou_withprimers_64renamedsam...

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

Are reads already merged ?

No

The archive contains 1 file by sample : R1 and R2 pair 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 mismatch in the overlap region

Merge software

Vsearch

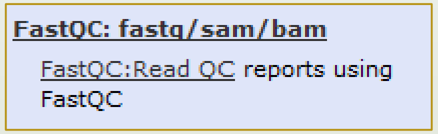
Select the software to merge paired-end reads.

Would you like to keep unmerged reads?

Yes [✓ No]

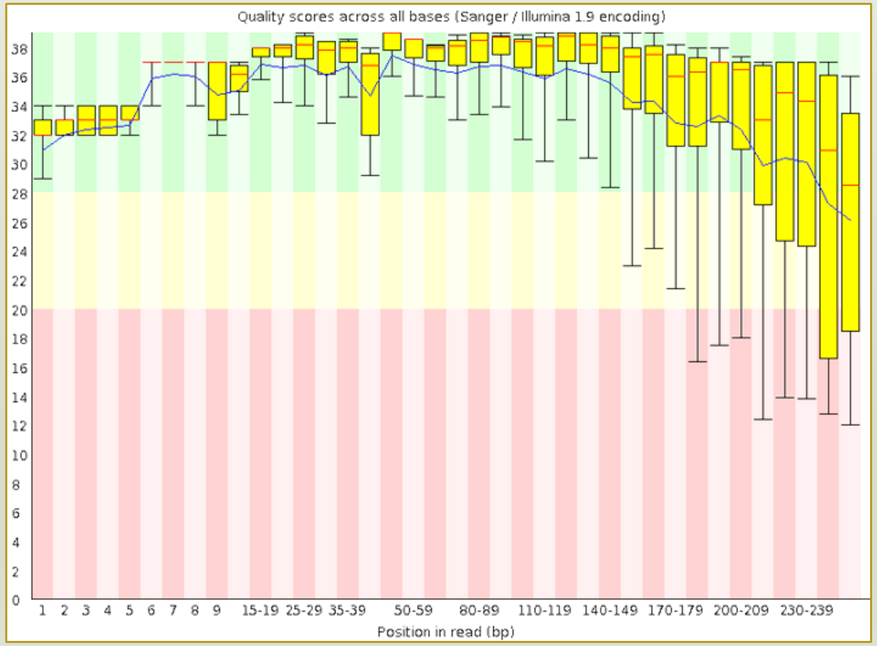
No : Unmerged reads will be excluded; Yes : unmerged reads will be artificially combined with 100 N. (default No)

To check the sequence quality use FASTQC (present in galaxy tools)



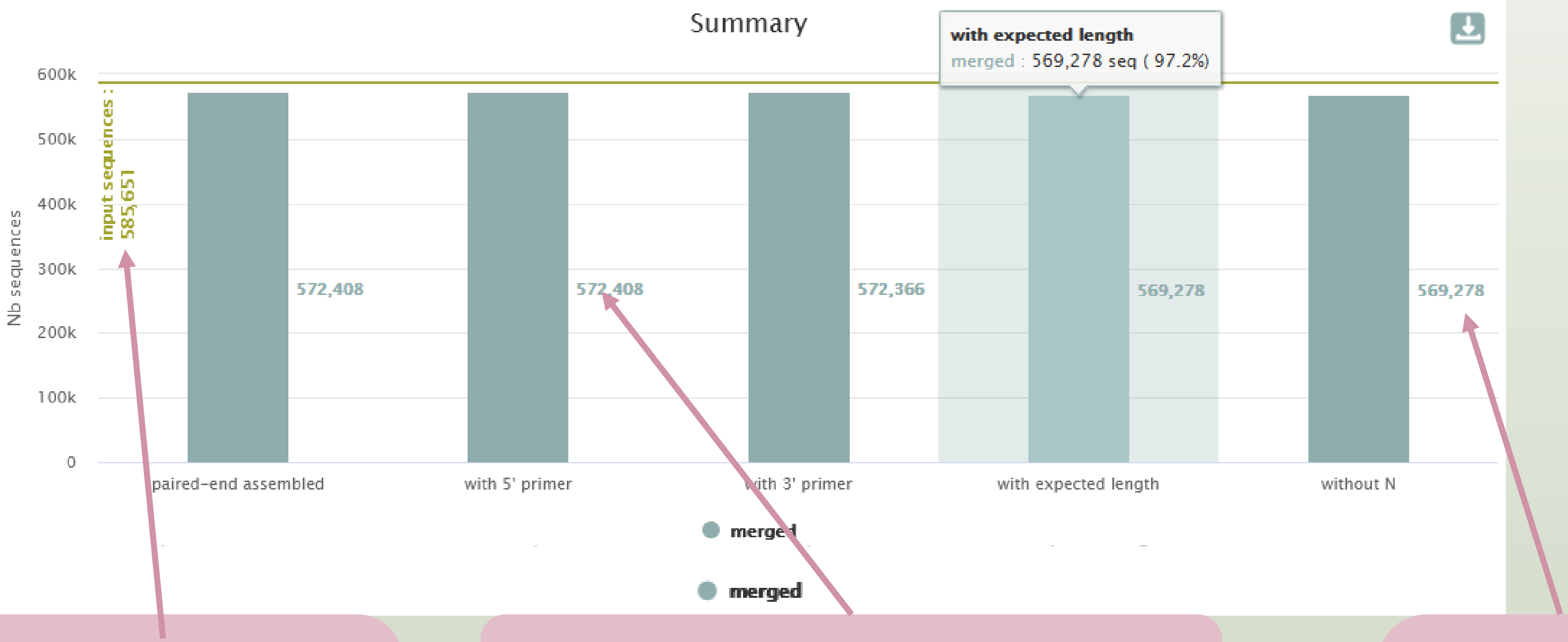
FastQC: fastq/sam/bam

FastQC:Read QC reports using FastQC



Answer 6, 7 & 8

Q6: How many sequences are there in the input file ?
Q7: How many sequences did not have the 5’ primer?
Q8: How many sequences still are after pre-processing the data?



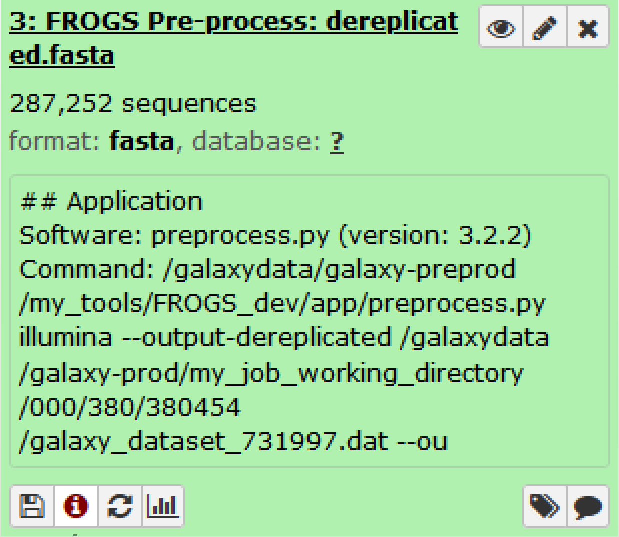
Total number of sequences before preprocessing: 585 651

All sequences have the 5’ primer

569 278 sequences are still after preprocessing

Answer 9

Q9: How much time did it take to pre-process the data ?

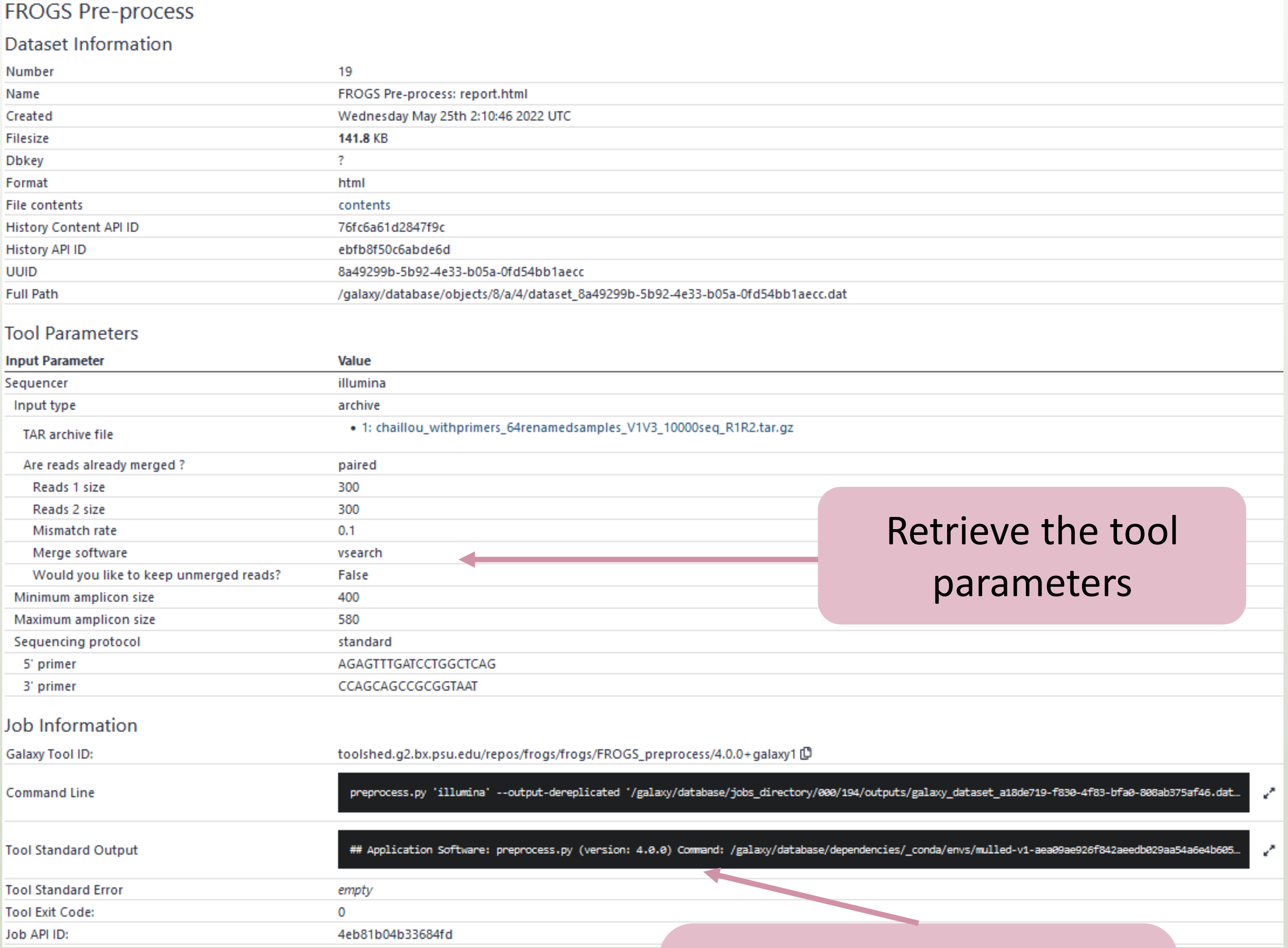


3: FROGS Pre-process: dereplicated.fasta

287,252 sequences
format: **fasta**, database: ?

```
## Application
Software: preprocess.py (version: 3.2.2)
Command: /galaxydata/galaxy-preprod
/my_tools/FROGS_dev/app/preprocess.py
illumina --output-dereplicated /galaxydata
/galaxy-prod/my_job_working_directory
/000/380/380454
/galaxy_dataset_731997.dat --ou
```

Click on « i »



FROGS Pre-process

Dataset Information

Number	19
Name	FROGS Pre-process: report.html
Created	Wednesday May 25th 2:10:46 2022 UTC
Filesize	**141.8** KB
Dbkey	?
Format	html
File contents	contents
History Content API ID	76fc6a61d2847f9c
History API ID	ebfb8f50c6abde6d
UUID	8a49299b-5b92-4e33-b05a-0fd54bb1aecc
Full Path	/galaxy/database/objects/8/a/4/dataset_8a49299b-5b92-4e33-b05a-0fd54bb1aecc.dat

Tool Parameters

Input Parameter	Value
Sequencer	illumina
Input type	archive
TAR archive file	• 1: chaillou_withprimers_64renamedsamples_V1V3_10000seq_R1R2.tar.gz
Are reads already merged ?	paired
Reads 1 size	300
Reads 2 size	300
Mismatch rate	0.1
Merge software	vsearch
Would you like to keep unmerged reads?	False
Minimum amplicon size	400
Maximum amplicon size	580
Sequencing protocol	standard
5' primer	AGAGTTTGATCCTGGCTCAG
3' primer	CCAGCAGCCGCGGTAAT

Retrieve the tool parameters

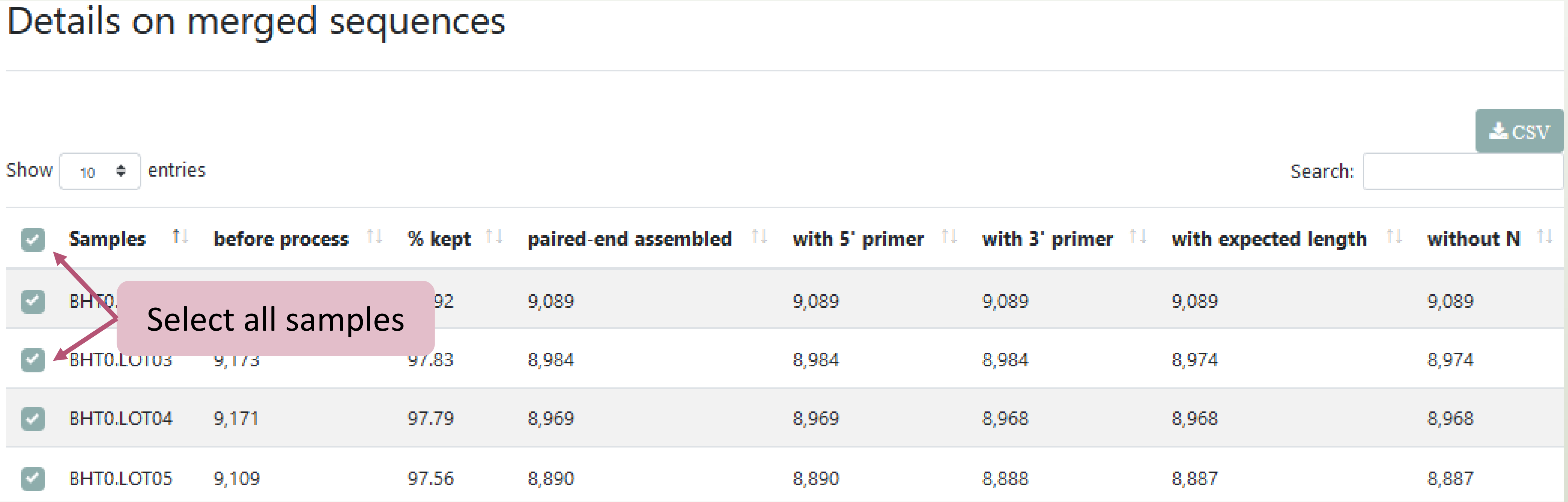
Job Information

Galaxy Tool ID:	toolshed.g2.bx.psu.edu/repos/frogs/frogs/FROGS_preprocess/4.0.0+galaxy1
Command Line	`preprocess.py 'illumina' --output-dereplicated '/galaxy/database/jobs_directory/000/194/outputs/galaxy_dataset_a18de719-f830-4f83-bfa0-808ab375af46.dat…`
Tool Standard Output	`## Application Software: preprocess.py (version: 4.0.0) Command: /galaxy/database/dependencies/_conda/envs/mulled-v1-aea09ae926f842aeedb029aa54a6e4b605…`
Tool Standard Error	*empty*
Tool Exit Code:	0
Job API ID:	4eb81b04b33684fd

Stdout contains FROGS command lines and time execution

Answer 10

Q10: What is the length of your merged reads before preprocessing ?



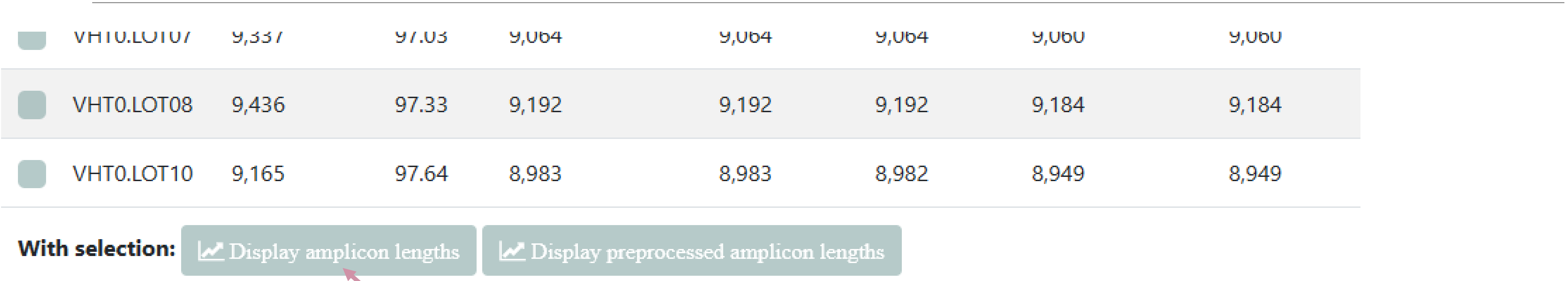
Details on merged sequences

☑	Samples	before process	% kept	paired-end assembled	with 5' primer	with 3' primer	with expected length	without N
☑	BHT0.		92	9,089	9,089	9,089	9,089	9,089
☑	BHT0.LOT03	9,173	97.83	8,984	8,984	8,984	8,974	8,974
☑	BHT0.LOT04	9,171	97.79	8,969	8,969	8,968	8,968	8,968
☑	BHT0.LOT05	9,109	97.56	8,890	8,890	8,888	8,887	8,887

Select all samples

Answer 10

Q10: What is the length of your merged reads before preprocessing ?

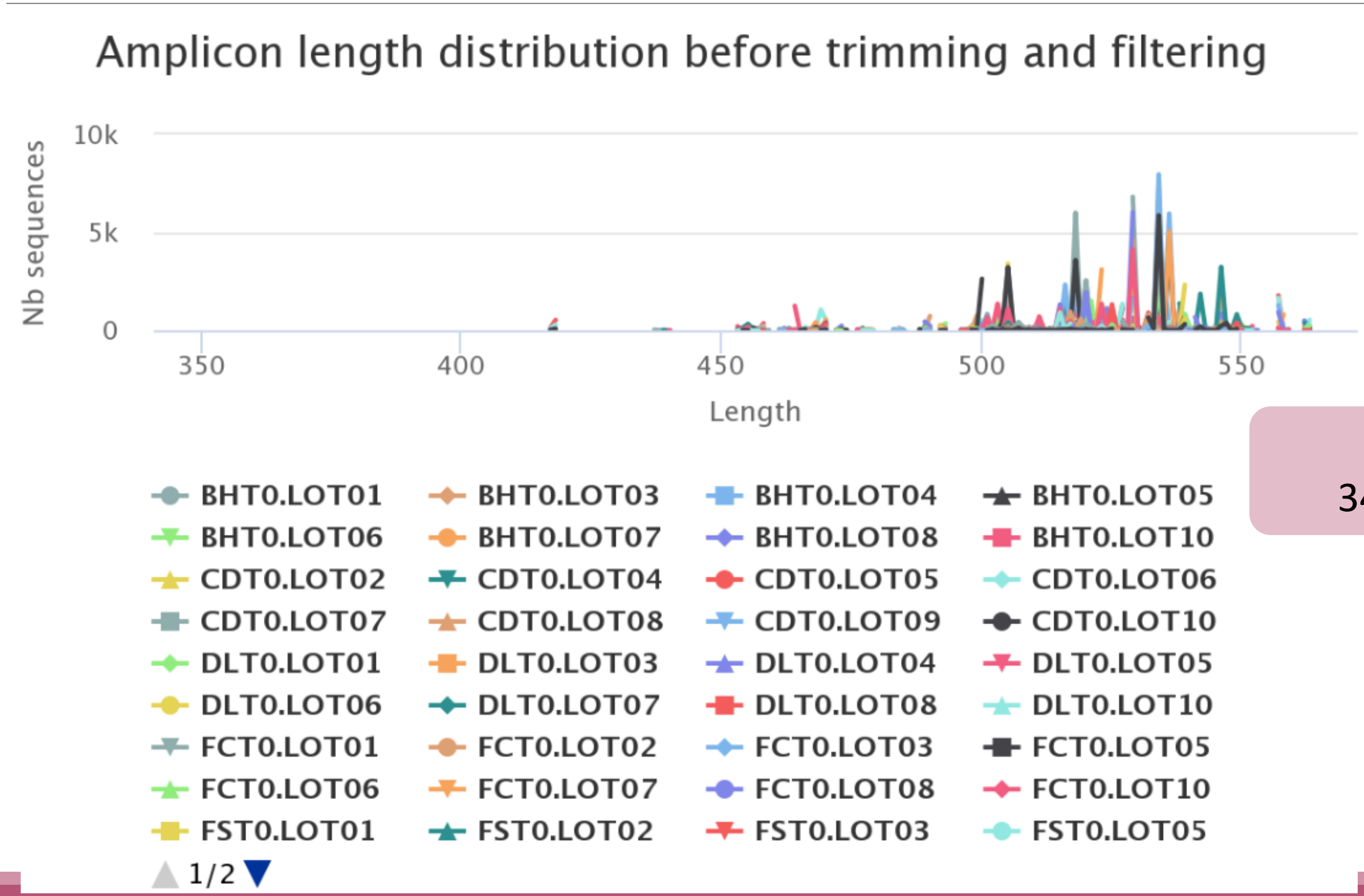


□	VHT0.LOT07	9,337	97.03	9,064	9,064	9,064	9,060	9,060
□	VHT0.LOT08	9,436	97.33	9,192	9,192	9,192	9,184	9,184
□	VHT0.LOT10	9,165	97.64	8,983	8,983	8,982	8,949	8,949

at the bottom of the table

Answer 10

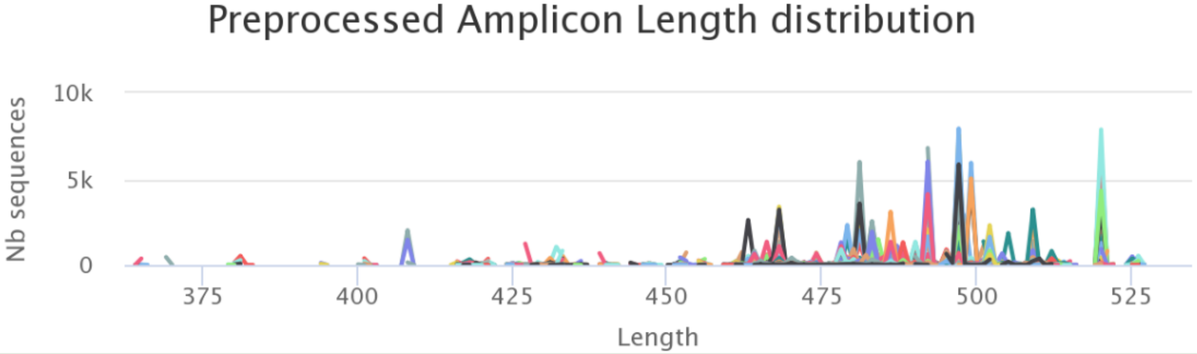
Q10: What is the length of your merged reads before preprocessing ?



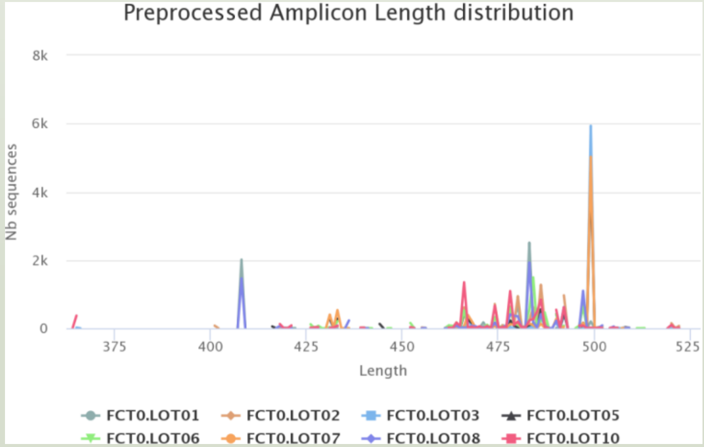
Before preprocessing:
343 < sequence length < 570

Answer 11

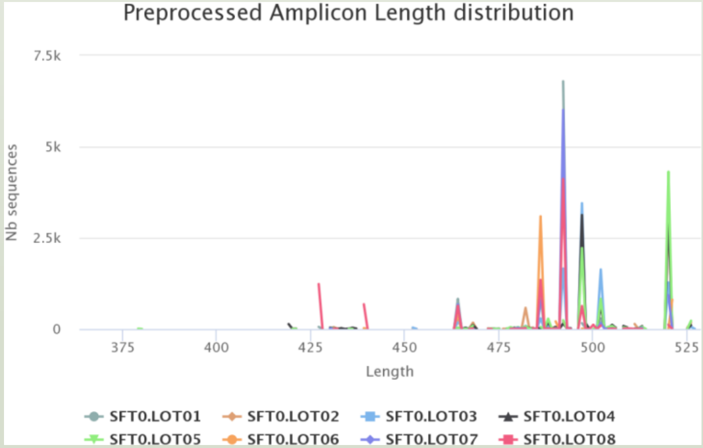
Q11: What can you tell about the samples, based on amplicon size distributions ?



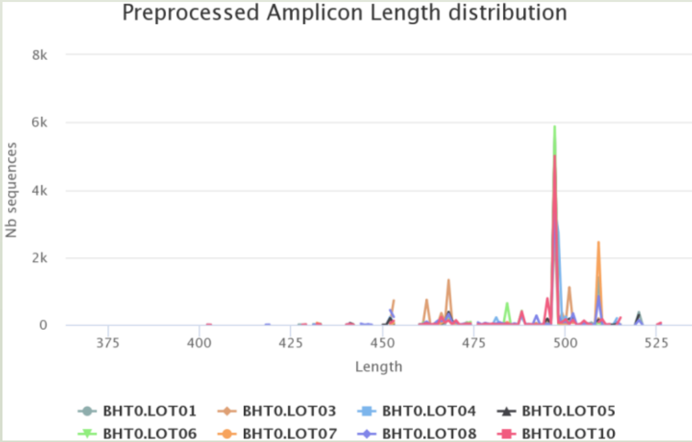
« Filet Cabillaud » samples



« Saumon Fumé » samples



« Bœuf Haché » samples



For each EnvType, we can observe different amplicon sizes. They correspond to different species.
N.B. amplicons with same size can represent different species.

Preprocess tool in brief

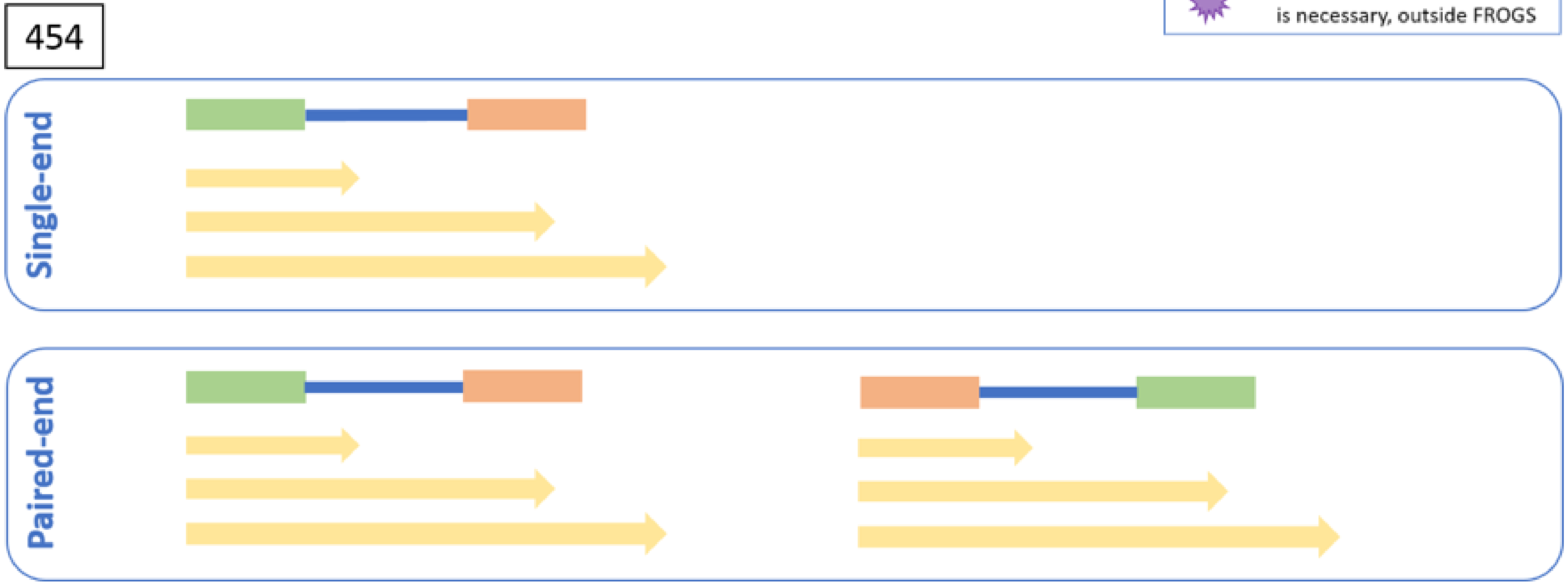
	Take in charge
Illumina	✓
454	✓
Merged data	✓
Not merged data	✓
Without primers	✓
Only R1 or only R2	🚫
Too distant R1 and R2 to be merged	✓
Over-overlapping R1 R2	✓

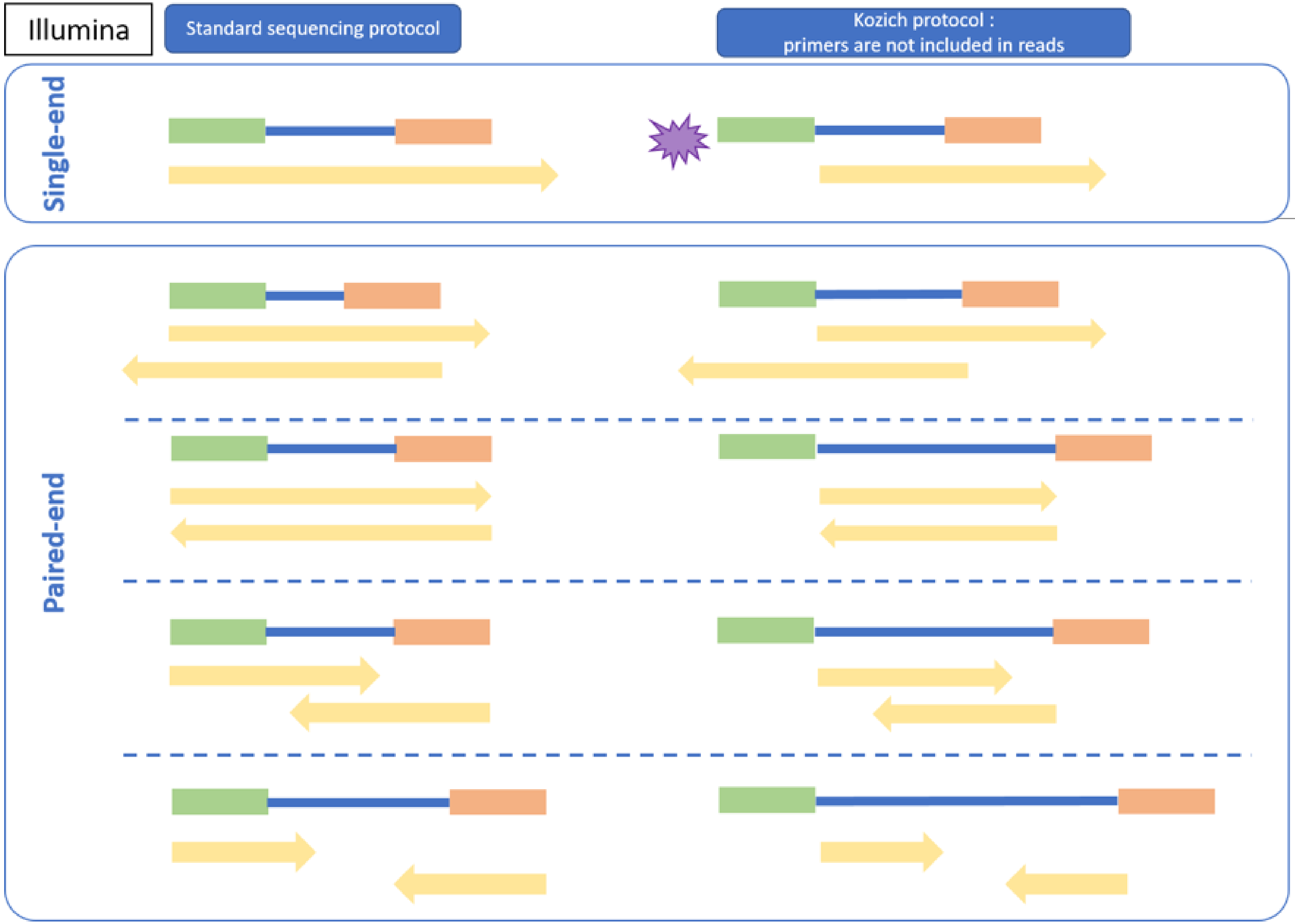
	Take in charge
Archive .tar.gz	✓
Fastq	✓
Fasta	🚫
With only 1 primer	🚫
Multiplexed data	🚫
Demultiplexed data	✓

Processed data by FROGS in brief

Legend

- 🟩 Primer Forward
- 🟧 Primer Reverse
- ▬ DNA target
- ⇨ Read
- ✷ Case where a pre-processing is necessary, outside FROGS




Illumina
Standard sequencing protocol
Kozich protocol :
primers are not included in reads
Single-end
Paired-end

✸ → Remove reverse primer before FROGS processing

Legend

🟩 Primer Forward

🟧 Primer Reverse

▬ DNA target

⇨ Read

✸ Case where a pre-processing is necessary, outside FROGS

Length of the sequenced target < length of one read

Supported since version 3.0

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

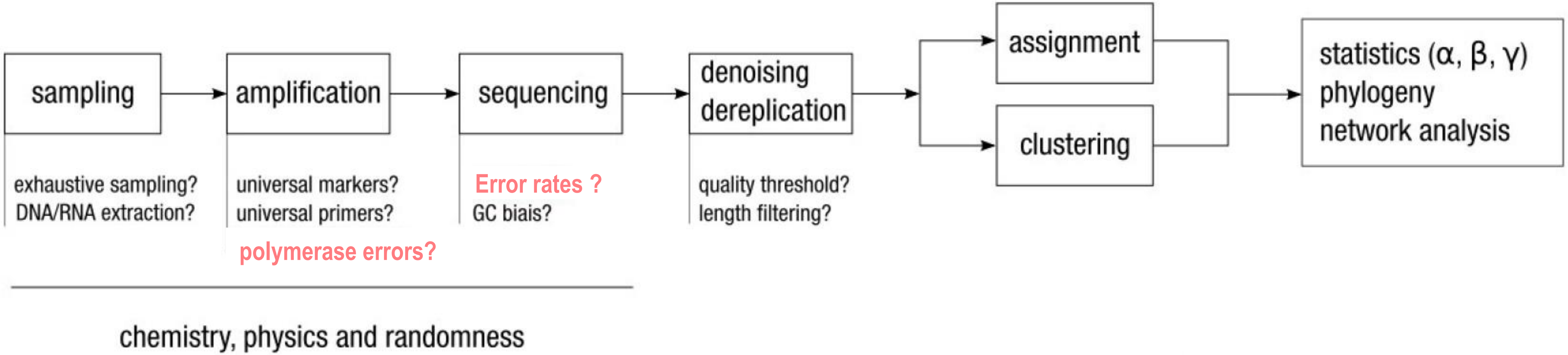
Length of the sequenced target >= the sum of the lengths of the two reads

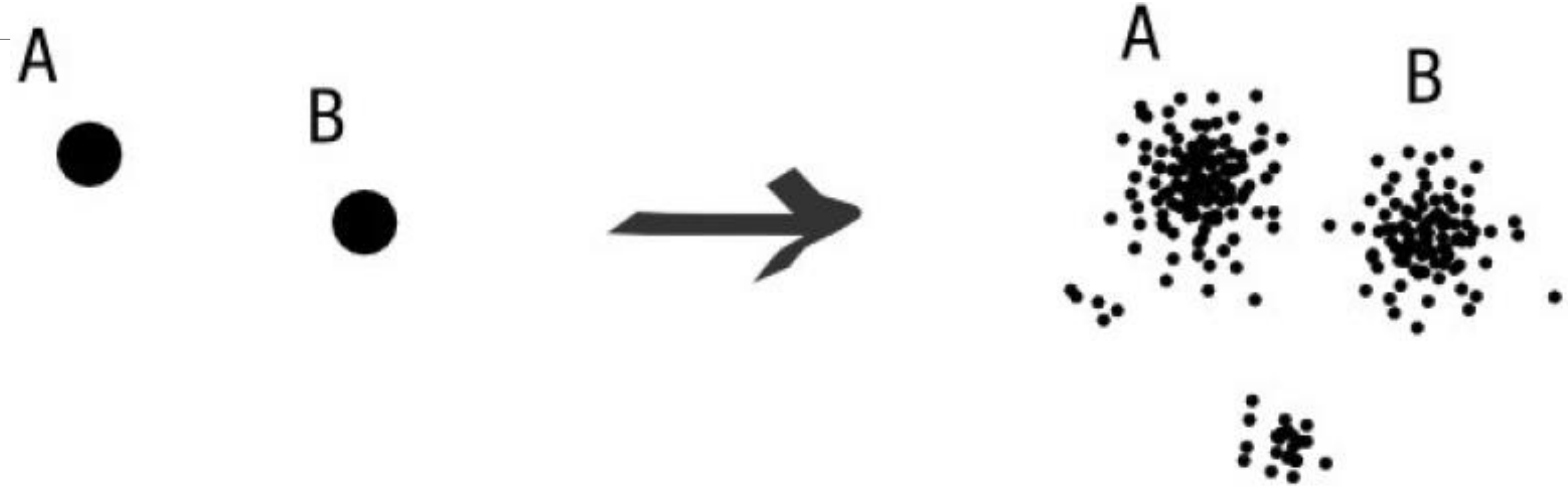
Supported since version 3.0 with option “keep unmerged reads” in preprocess Tool

Clustering tool

Why do we need clustering ?

Amplication and sequencing and are not perfect processes



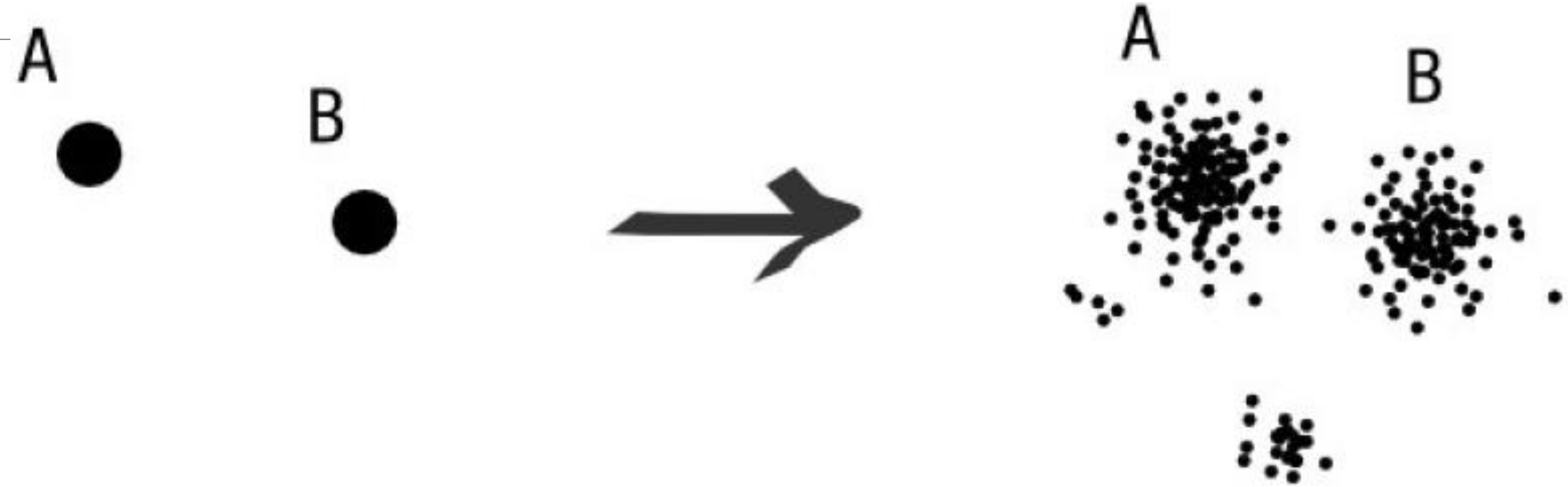


Natural variability ?
Technical noise?
Contaminant?
Chimeras?

Expected

Results

Fréderic Mahé communication



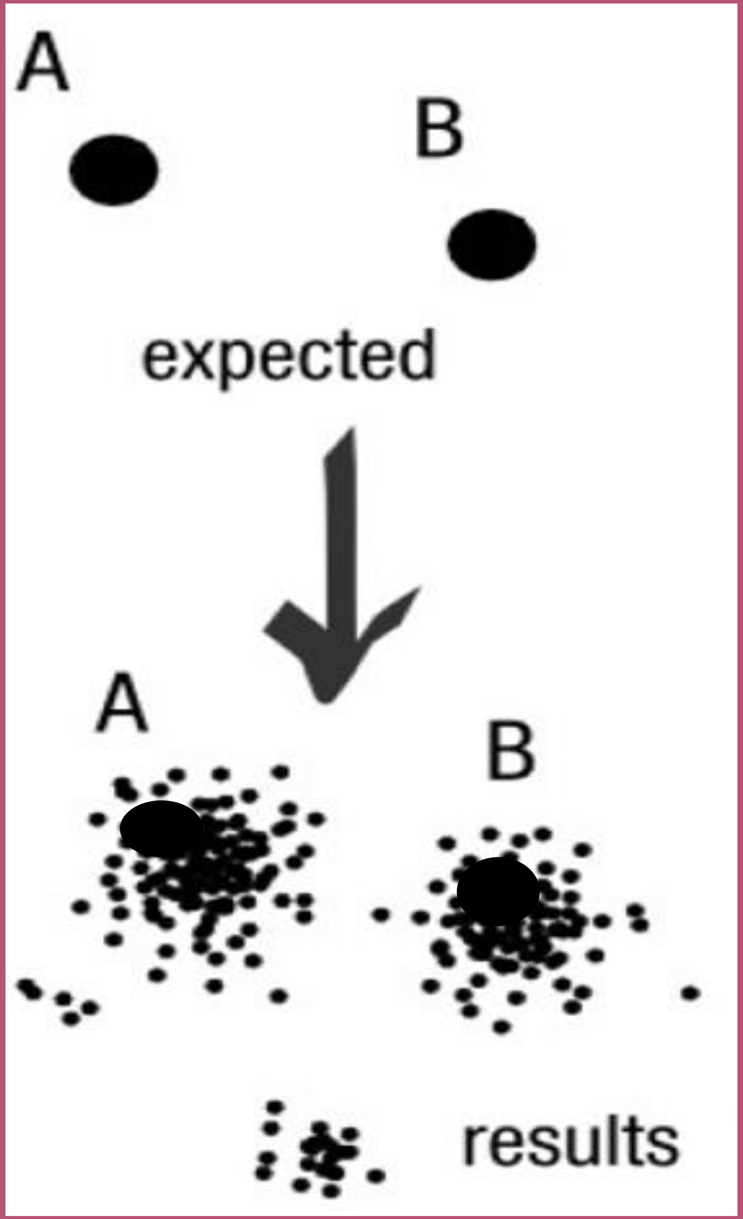
Natural variability ?
Technical noise?
Contaminant?
Chimeras?

Expected

Results

16S variability
Cf. RRNDB (ribosomal RNA operons database)
https://rrndb.umms.med.umich.edu/search/
max. 21 copies of 16S in bacteria (*Photobacterium damselae*)
ex. *E. coli* 7 copies

Fréderic Mahé communication



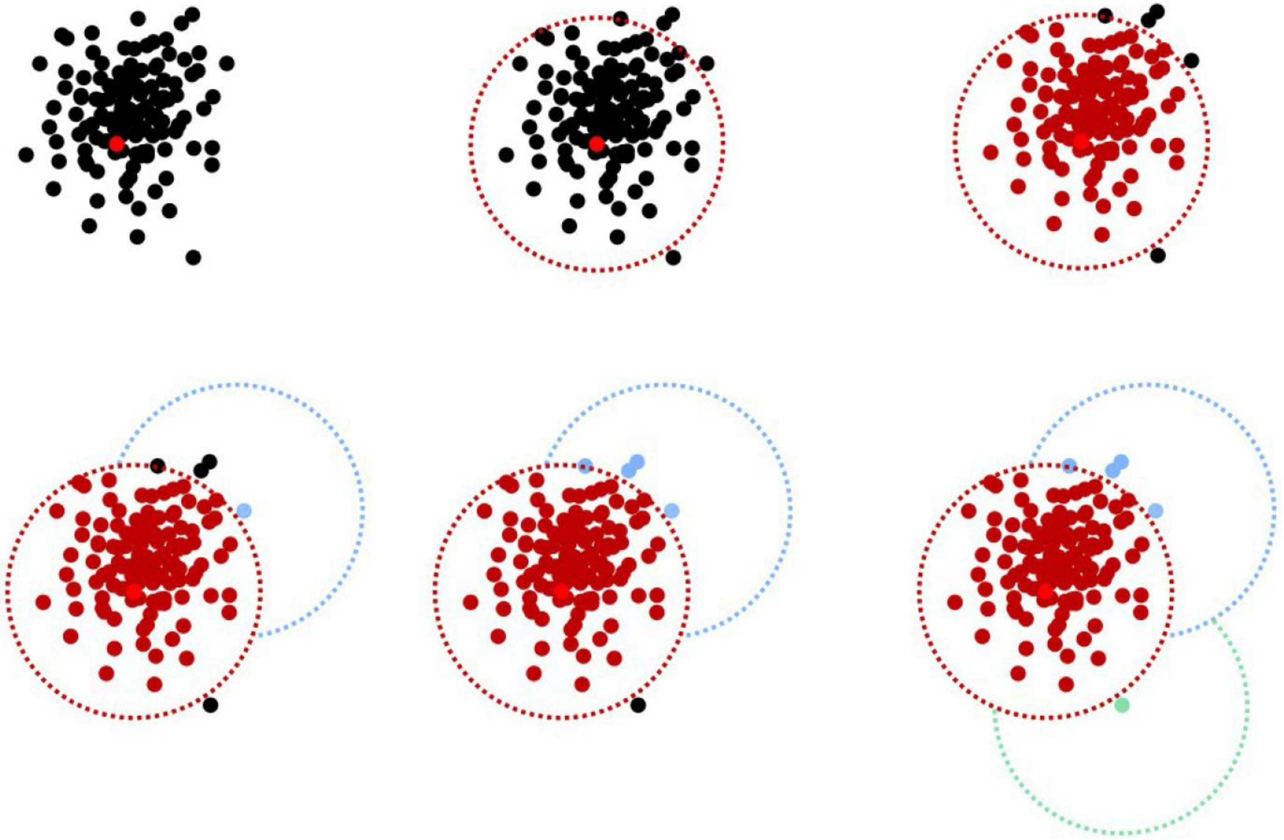
To have the best accuracy:

Method: All against all

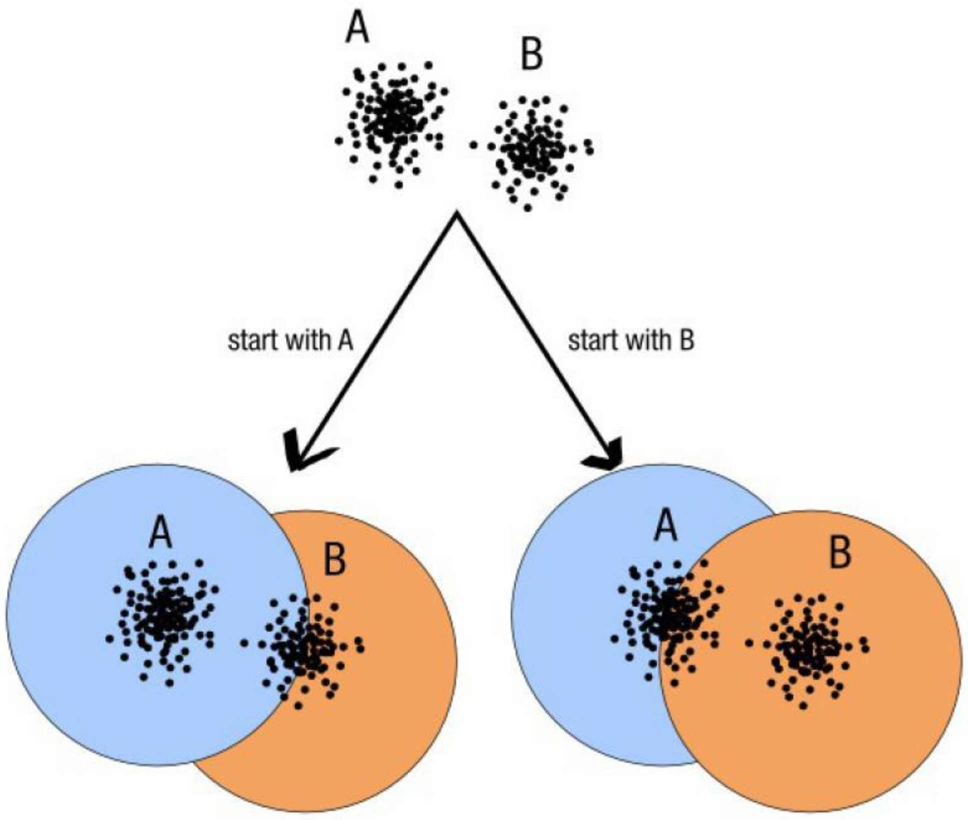
- Very accurate
- Requires a lot of memory and/or time

=> Impossible on very large datasets without strong filtering or sampling

How traditional clustering works ?

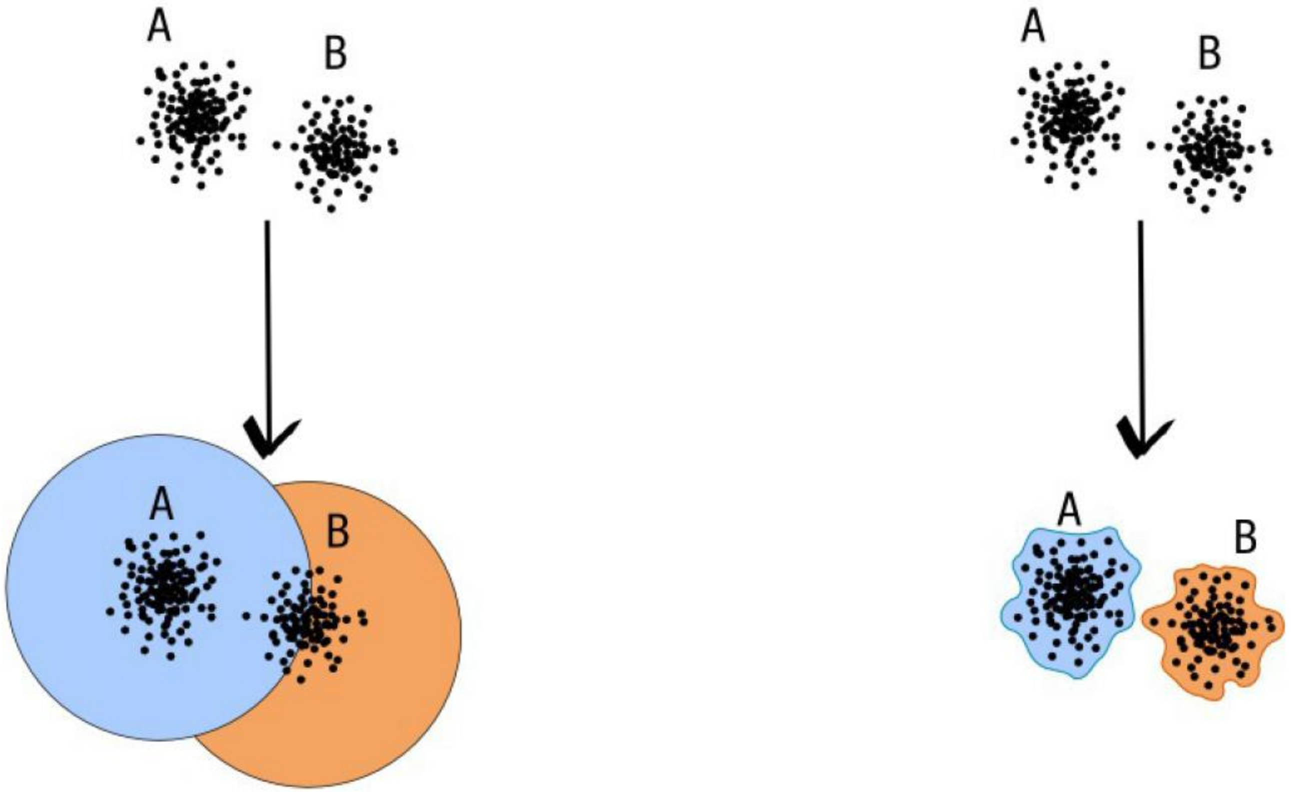


Input order dependent results



decreasing length,
decreasing abundance,
external references

Single a priori clustering threshold

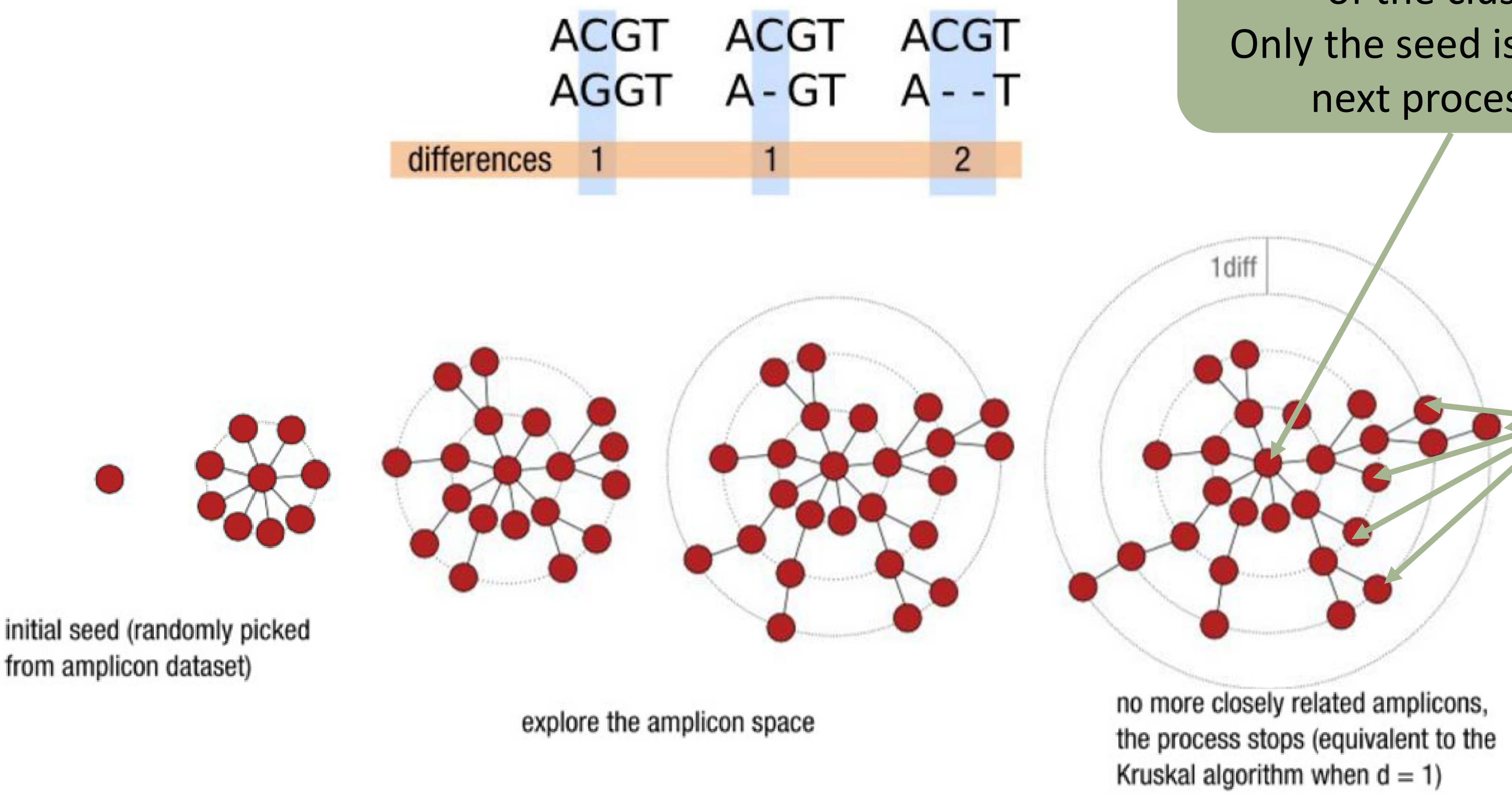


compromise threshold
unadapted threshold

natural limits of clusters

Fréderic Mahé communication

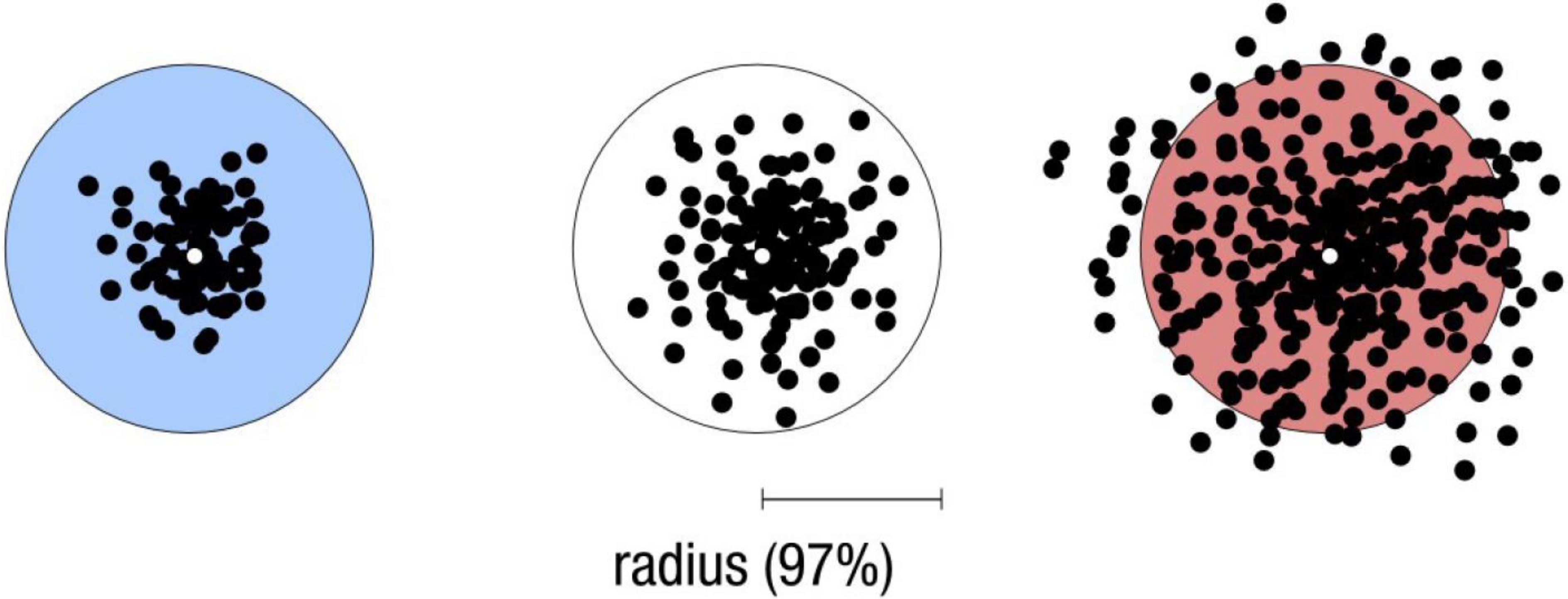
Swarm clustering method


ACGT
AGGT
ACGT
A - GT
ACGT
A - - T
differences
1
1
2
1diff
initial seed (randomly picked from amplicon dataset)
explore the amplicon space
no more closely related amplicons, the process stops (equivalent to the Kruskal algorithm when d = 1)

This sequences is the seed of the cluster.
Only the seed is kept for next processes.

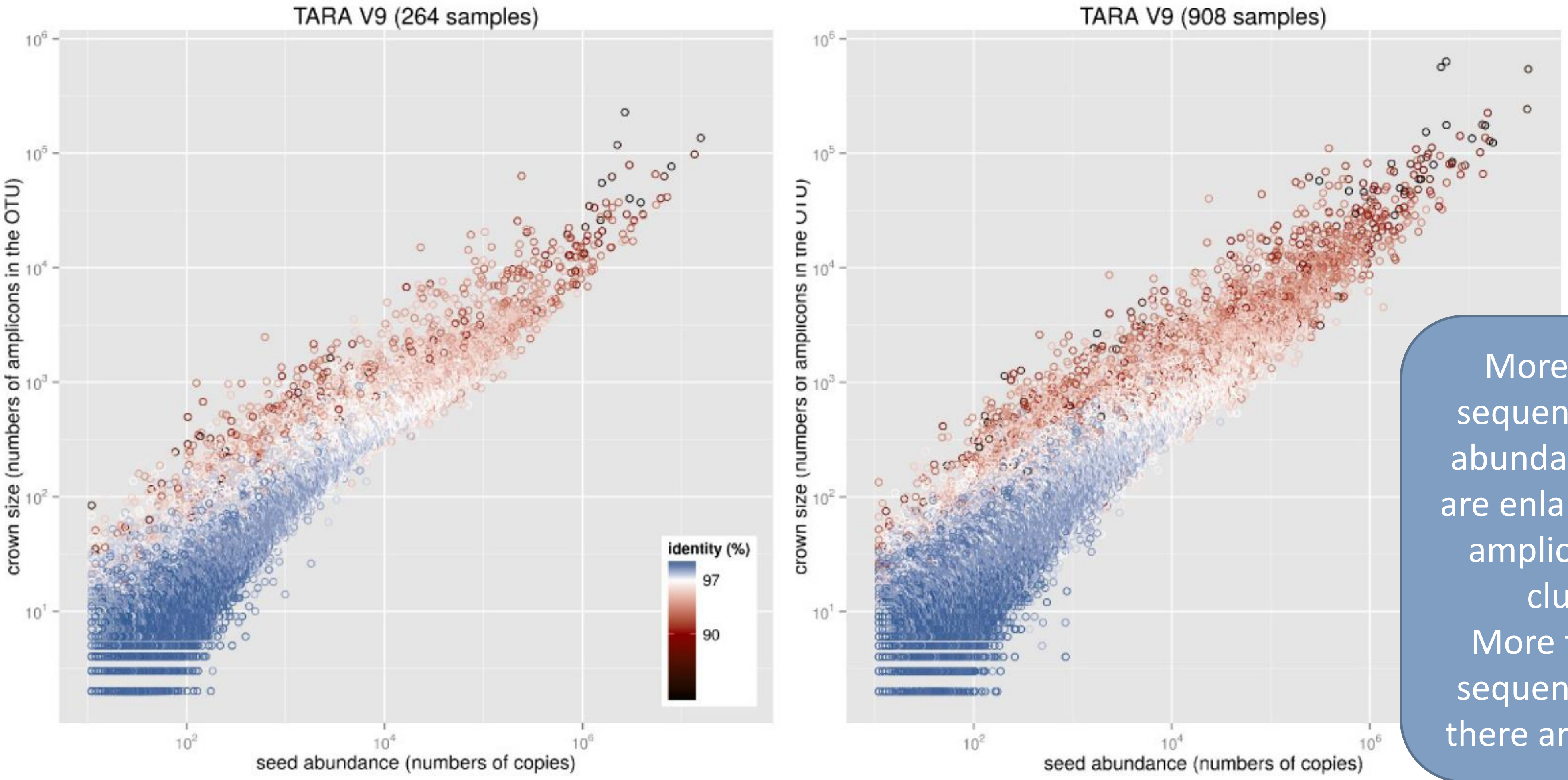
The abundances of each sequence in the cluster are added together.
And the total abundance is given to the seed.

Comparison Swarm and 3% clusterings



Radius expressed as a percentage of identity with the central amplicon (97% is by far the most widely used clustering threshold)

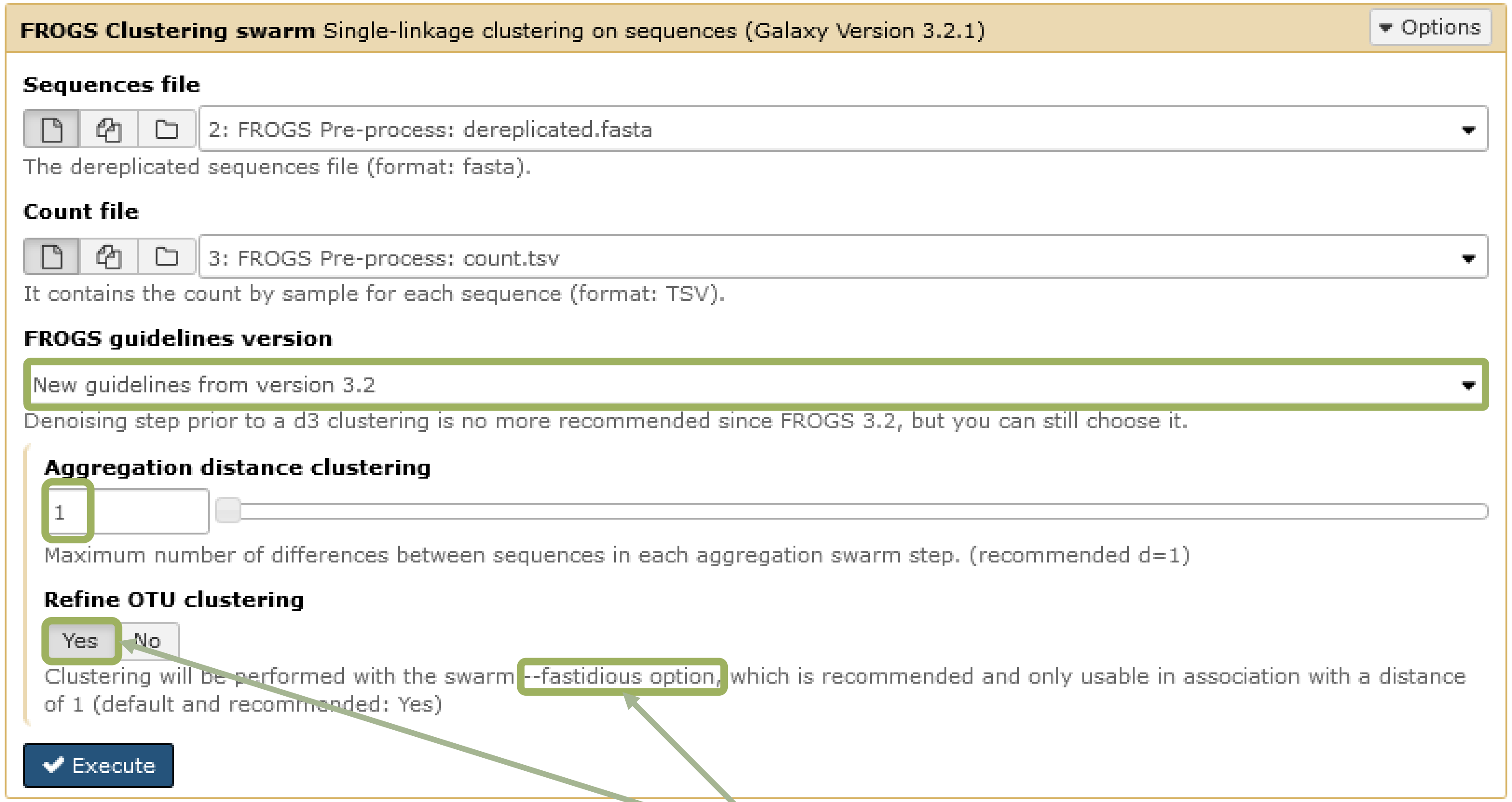
Comparison Swarm and 3% clusterings



clusters produced with swarm using d = 1

More there is sequences, more abundant clusters are enlarged (more amplicon in the cluster).
More there are sequences, more there are artefacts

Fréderic Mahé communication



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

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

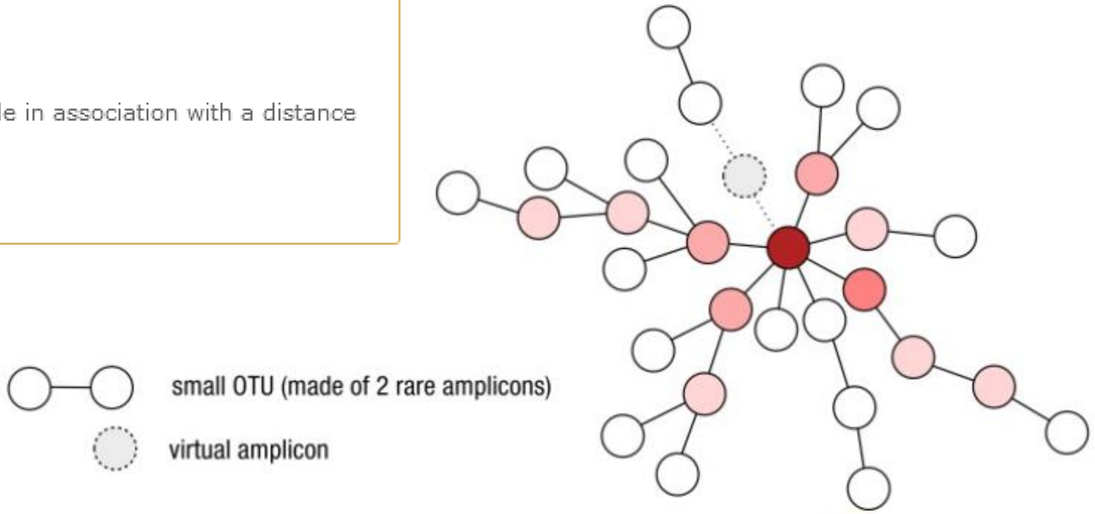
Maximum number of differences between sequences in each aggregation swarm step. (recommended d=1)

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)

longer but more accurate



Mahé, Frédéric et al. "Swarm v2: highly-scalable and high-resolution amplicon clustering." *PeerJ* vol. 3 e1420. 10 Dec. 2015, doi:10.7717/peerj.1420

SWARM

A robust and fast clustering method for amplicon-based studies.

The purpose of **swarm** is to provide a novel clustering algorithm to handle large sets of amplicons.

swarm results are resilient to input-order changes and rely on a small **local** linking threshold *d*, the maximum number of differences between two amplicons.

swarm forms stable high-resolution clusters, with a high yield of biological information.

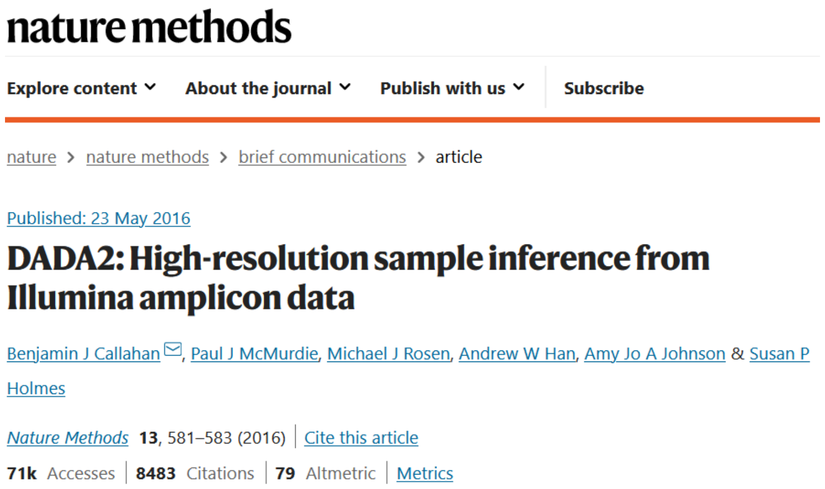
swarm produces ASV-like clusters.

Swarm: robust and fast clustering method for amplicon-based studies.
Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M.
PeerJ. 2014 Sep 25;2:e593. doi: 10.7717/peerj.593. eCollection 2014.
PMID:25276506

« ASV *vs* OTU » debate

One of the popular methods to process amplicon data is the ASV (**A**mplicon **S**equence **V**ariant) analysis which groups sequences according to their abundance and an error model, as proposed by DADA2 or others.

The probability that a sequence is derived from errors in another sequence is estimated by taking into account their abundances and the transition rates between bases (sequencing or PCR errors).



nature methods

nature › nature methods › brief communications › article

Published: 23 May 2016

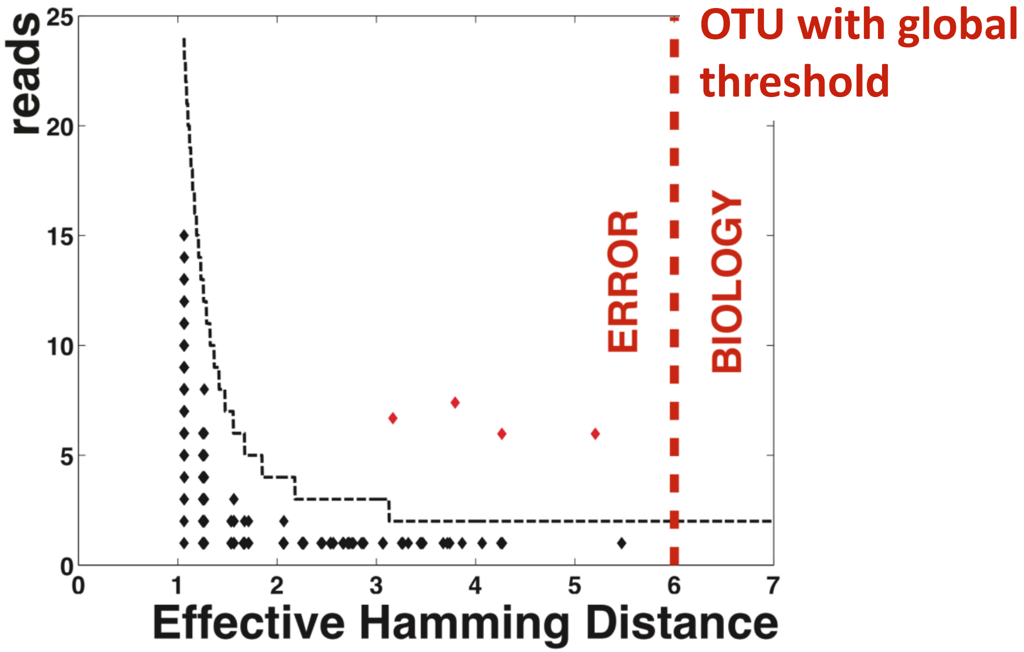
DADA2: High-resolution sample inference from Illumina amplicon data

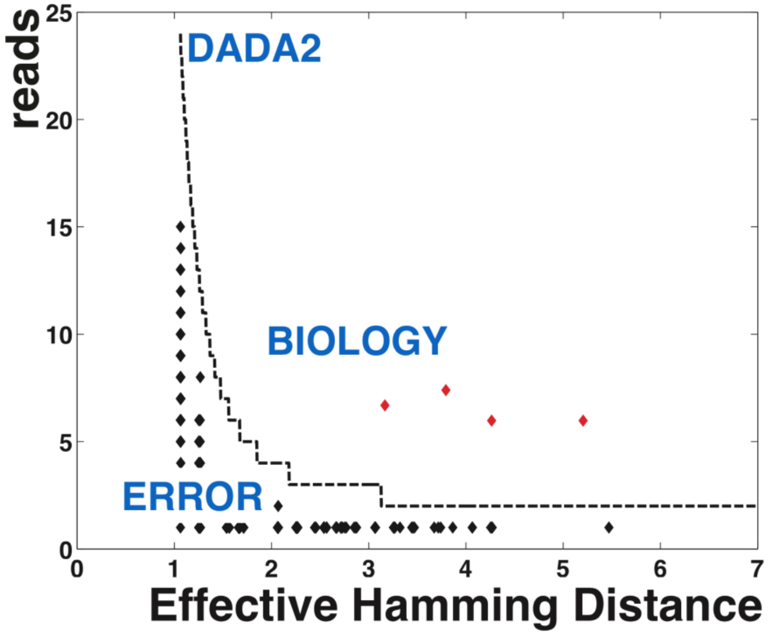
Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson & Susan P Holmes

Nature Methods **13**, 581–583 (2016) | Cite this article

71k Accesses | **8483** Citations | **79** Altmetric | Metrics

« ASV *vs* OTU » debate

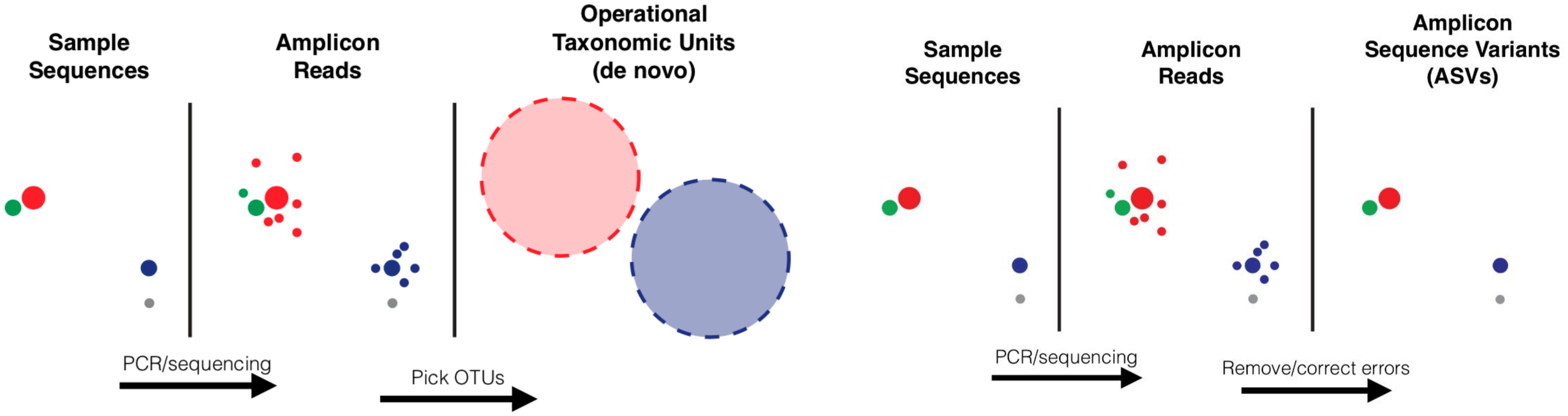




Due to the "ASV vs OTU" debate, the term OTU is currently negatively connoted and creates confusion by suggesting that all methods producing OTUs use a fixed clustering threshold (classically at 97% similarity) and are therefore bad. This is of course not the case: the criticism of fixed threshold methods preceded the use of the term ASV and several previously published tools produce ASV-like clusters, including swarm, the clustering tool used in FROGS.

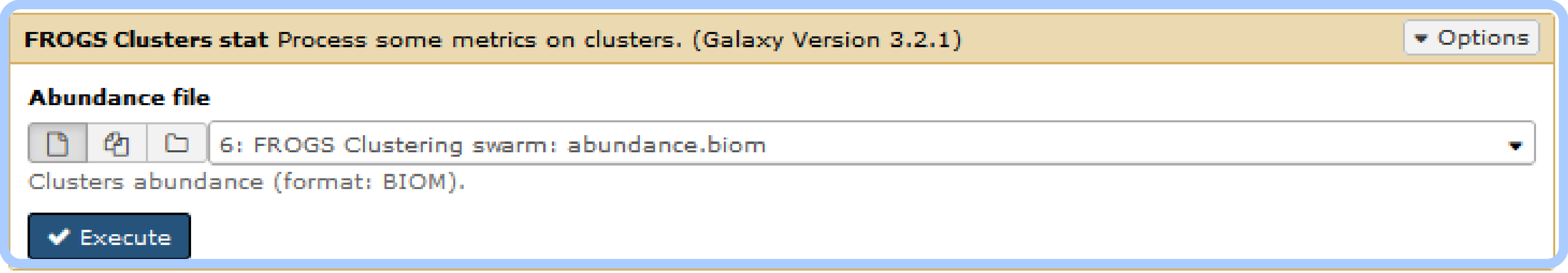
« ASV *vs* OTU » debate

- The ASV vs. OTU debate is actually about fixed-threshold clustering approaches, the criticism of which preceded the ASV term. Many methods, including **swarm**, pre-existed and produce "ASV-like" clusters/OTUs (figure below is from Callahan, author of dada2 and really similar to swarm).



- FROGS OTUs are therefore not concerned by the criticism made of OTUs in comparison to ASVs.

Cluster stat tool



FROGS Clusters stat Process some metrics on clusters. (Galaxy Version 3.2.1)

Abundance file

6: FROGS Clustering swarm: abundance.biom

Clusters abundance (format: BIOM).

Practice:

LAUNCH CLUSTERING AND CLUSTERSTAT TOOLS

Exercise

Go to « 16S » history

Launch the Clustering SWARM tool on that data set with guideline 3.2 *i.e. aggregation distance =1*

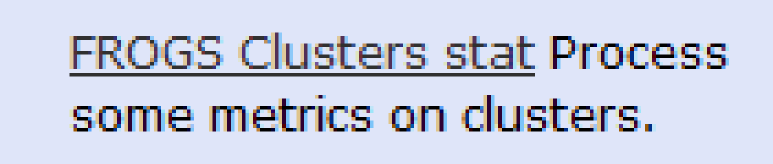
→ objectives :

- understand the outputs from clustering
- understand the ClusterStat utility

Exercise

1. How many clusters do you get ?

Launch FROGS **Cluster Stat tools** on the previous abundance biom file



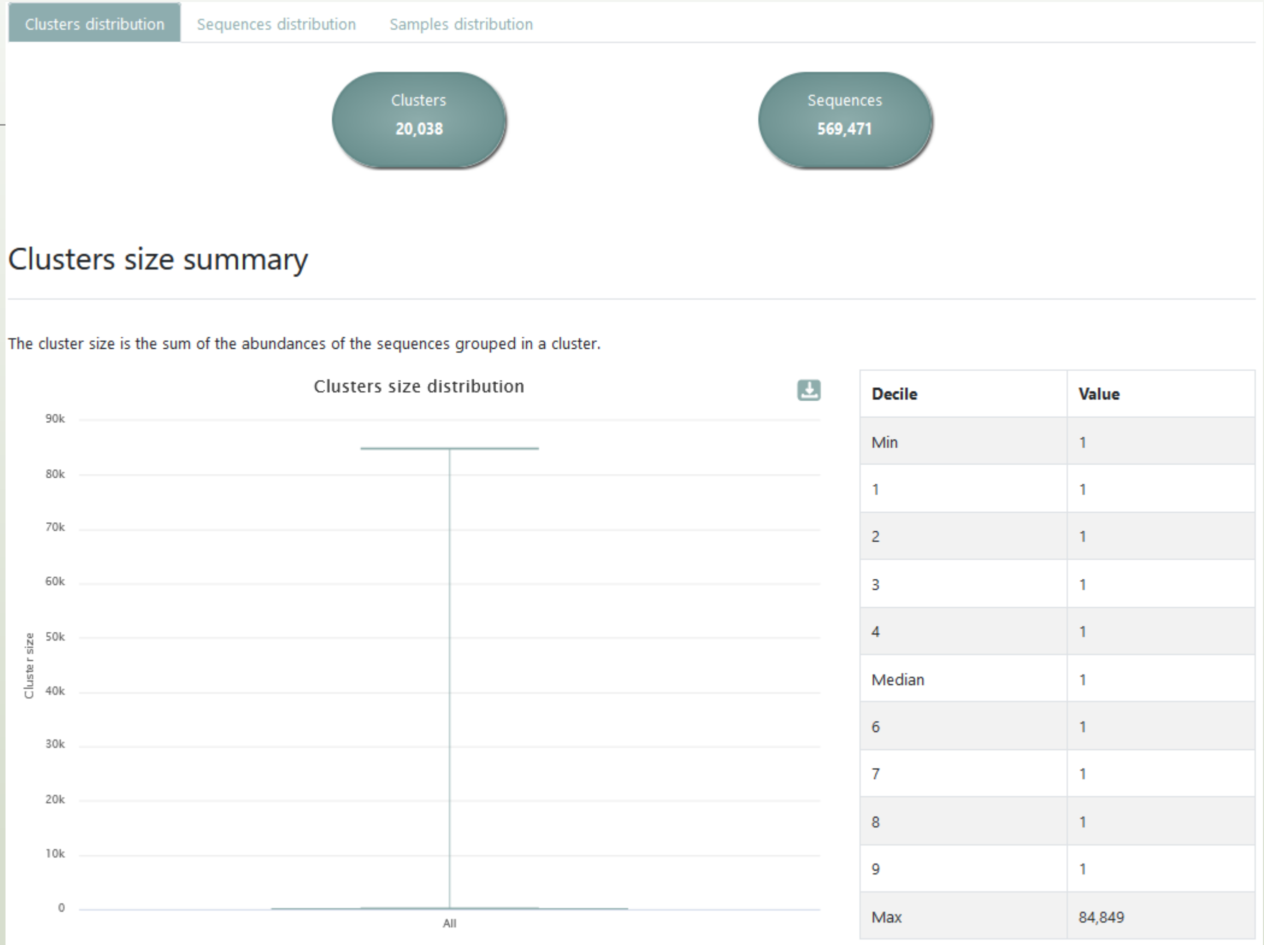
FROGS Clusters stat Process some metrics on clusters.

Exercise

2. Interpret the boxplot: **Clusters size summary**
3. Interpret the table: **Clusters size details - How many single singletons do you find?**
4. What can we say by observing the **sequence distribution**?
5. How many clusters share “BHT0.LOT08” with at least one other sample?
6. How many clusters could we expect to be shared ?
7. How many sequences represent the 106 specific clusters of “CDT0.LOT06”?
8. This represents what proportion of “CDT0.LOT06”?
9. What do you think about it?
10. How do you interpret the « Hierarchical clustering » ?

Answer 1, 2 & 3

Q1: How many clusters do you get ?
Q2: Interpret the boxplot: **Clusters size summary**
Q3: Interpret the table: **Clusters size details - How many single singletons do you find?**



Clusters 20,038

Sequences 569,471

Clusters size summary

The cluster size is the sum of the abundances of the sequences grouped in a cluster.

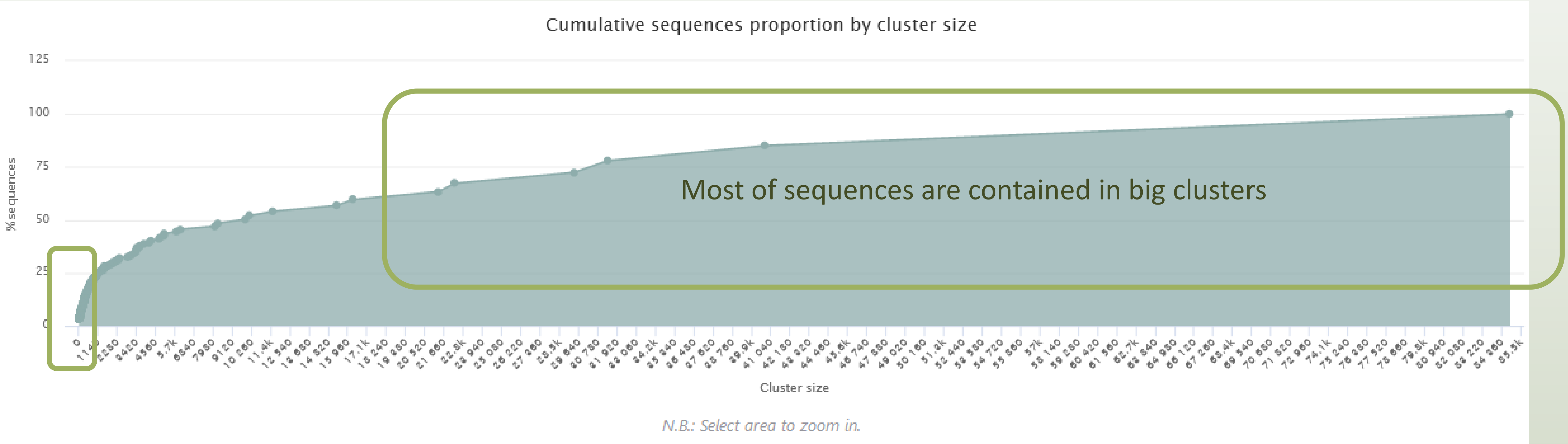
Clusters size distribution

Decile	Value
Min	1
1	1
2	1
3	1
4	1
Median	1
6	1
7	1
8	1
9	1
Max	84,849

Most of OTUs are singletons

Answer 4

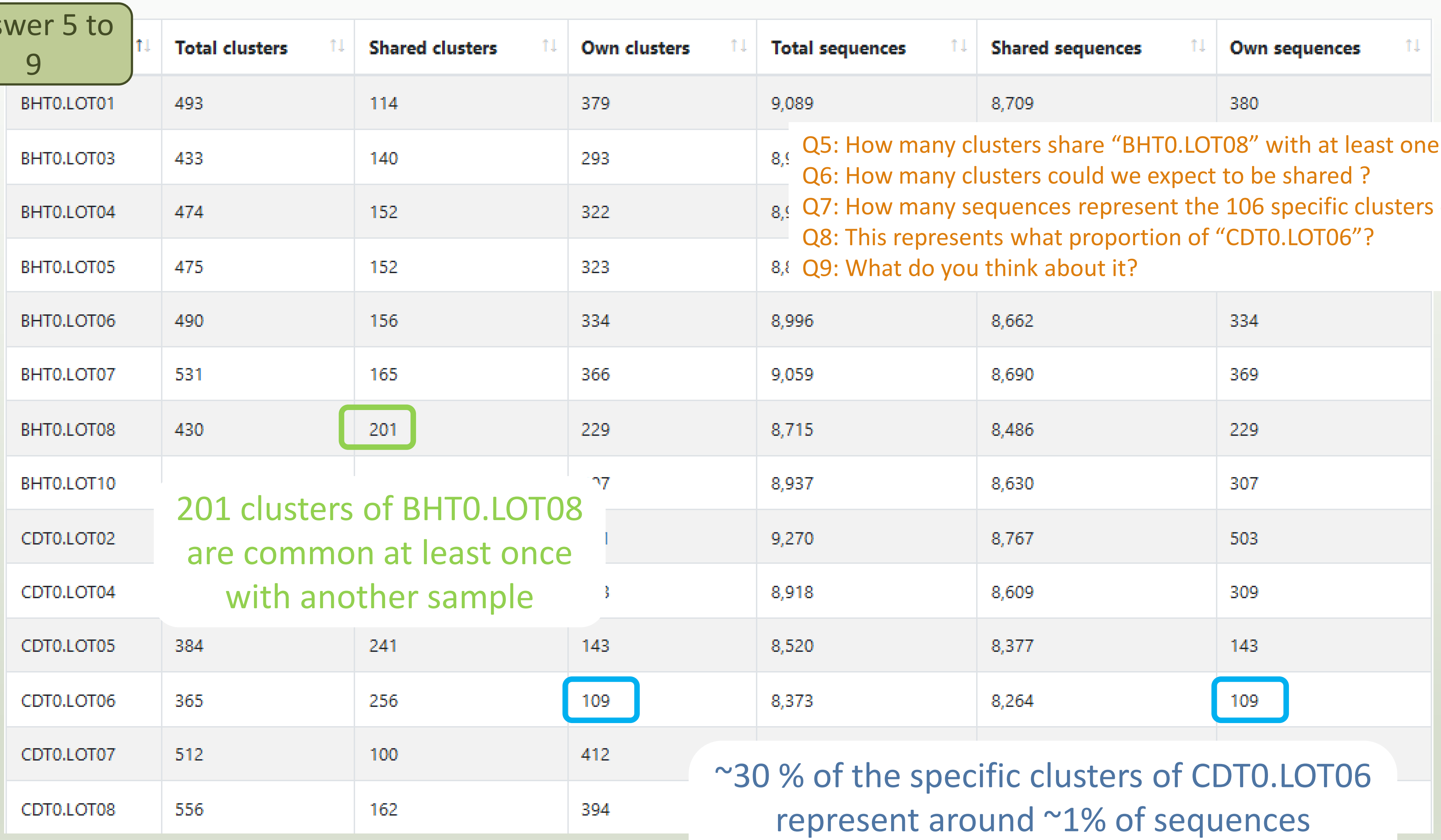
Q4: What can we say by observing the **sequence distribution**?



Most of sequences are contained in big clusters

The small clusters represent few sequences

Answer 5 to 9



	Total clusters	Shared clusters	Own clusters	Total sequences	Shared sequences	Own sequences
BHT0.LOT01	493	114	379	9,089	8,709	380
BHT0.LOT03	433	140	293	8,9		
BHT0.LOT04	474	152	322	8,9		
BHT0.LOT05	475	152	323	8,8		
BHT0.LOT06	490	156	334	8,996	8,662	334
BHT0.LOT07	531	165	366	9,059	8,690	369
BHT0.LOT08	430	201	229	8,715	8,486	229
BHT0.LOT10			07	8,937	8,630	307
CDT0.LOT02			1	9,270	8,767	503
CDT0.LOT04			3	8,918	8,609	309
CDT0.LOT05	384	241	143	8,520	8,377	143
CDT0.LOT06	365	256	109	8,373	8,264	109
CDT0.LOT07	512	100	412			
CDT0.LOT08	556	162	394			

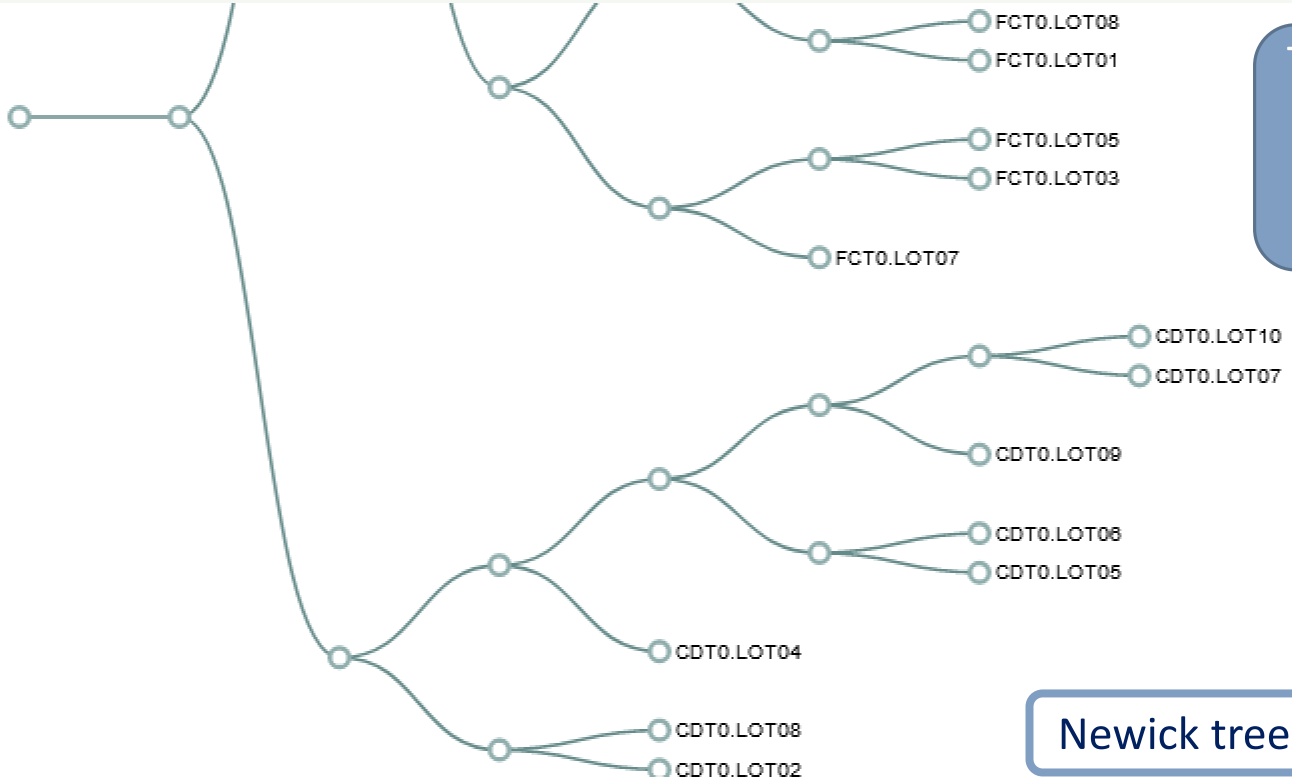
Q5: How many clusters share “BHT0.LOT08” with at least one other sample?
Q6: How many clusters could we expect to be shared ?
Q7: How many sequences represent the 106 specific clusters of “CDT0.LOT06”?
Q8: This represents what proportion of “CDT0.LOT06”?
Q9: What do you think about it?

201 clusters of BHT0.LOT08 are common at least once with another sample

~30 % of the specific clusters of CDT0.LOT06 represent around ~1% of sequences
Could be interesting to remove if individual variability is not the concern of user

Answer 10

Q10: How do you interpret the « Hierarchical clustering » ?



The « Hierachical clustering » is established with a Bray Curtis distance particularly well adapted to abundance table of very heterogenous values (very big and very small figures).

Newick tree available too, can be copied and pasted an tree viewer

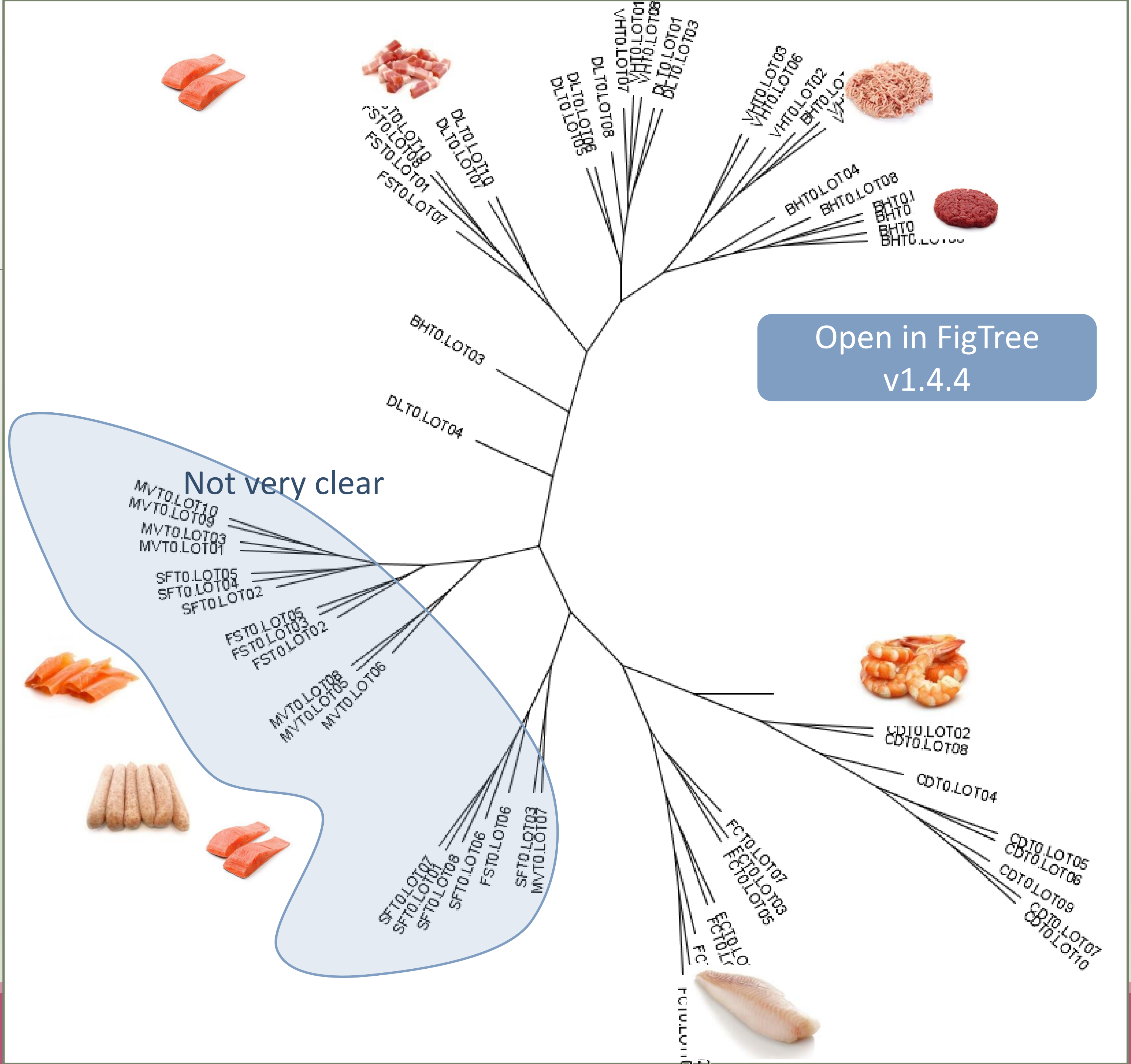
Newick

```
((((CDT0.LOT02,CDT0.LOT08):0.312,(CDT0.LOT04,((CDT0.LOT05,CDT0.LOT06):0.518,(CDT0.LOT09,(CDT0.LOT07,CDT0.LOT10):0.533):0.582):0.757):0.816):0.840,(((FCT0.LOT07,(FCT0.LOT03,FCT0.LOT05):0.257):0.262,
((FCT0.LOT01,FCT0.LOT08):0.352,(FCT0.LOT06,(FCT0.LOT02,FCT0.LOT10):0.427):0.631):0.805):0.832,(((MVT0.LOT07,SFT0.LOT03):0.493,(FST0.LOT06,(SFT0.LOT06,(SFT0.LOT08,
(SFT0.LOT01,SFT0.LOT07):0.132):0.345):0.354):0.570):0.655,(((MVT0.LOT06,(MVT0.LOT05,MVT0.LOT08):0.439):0.511,((FST0.LOT02,(FST0.LOT03,FST0.LOT05):0.147):0.179,((SFT0.LOT02,
(SFT0.LOT04,SFT0.LOT05):0.211):0.227,((MVT0.LOT01,MVT0.LOT03):0.161,(MVT0.LOT09,MVT0.LOT10):0.341):0.466):0.526):0.661):0.681,(DLT0.LOT04,((((DLT0.LOT05,DLT0.LOT06):0.173,(DLT0.LOT08,((VHT0.LOT07,
(VHT0.LOT01,VHT0.LOT08):0.095):0.184,(DLT0.LOT01,DLT0.LOT03):0.231):0.267):0.325):0.411,((BHT0.LOT04,(BHT0.LOT08,((BHT0.LOT01,BHT0.LOT07):0.224,(BHT0.LOT05,BHT0.LOT06):0.231):0.309):0.352):0.462,
((VHT0.LOT03,VHT0.LOT06):0.387,(VHT0.LOT02,(BHT0.LOT10,(VHT0.LOT04,VHT0.LOT10):0.272):0.336):0.401):0.463):0.590):0.711,(BHT0.LOT03,((FST0.LOT07,(FST0.LOT01,
(FST0.LOT08,FST0.LOT10):0.254):0.388):0.408,(DLT0.LOT07,DLT0.LOT10):0.440):0.666):0.734):0.745):0.827):0.856):0.875):0.911):0.938);
```

Answer 10

Q10: How do you interpret the « Hierarchical clustering » ?

N.B.: Hierarchical clustering is not all a phylogenetic tree ! Please consult with caution.



Not very clear

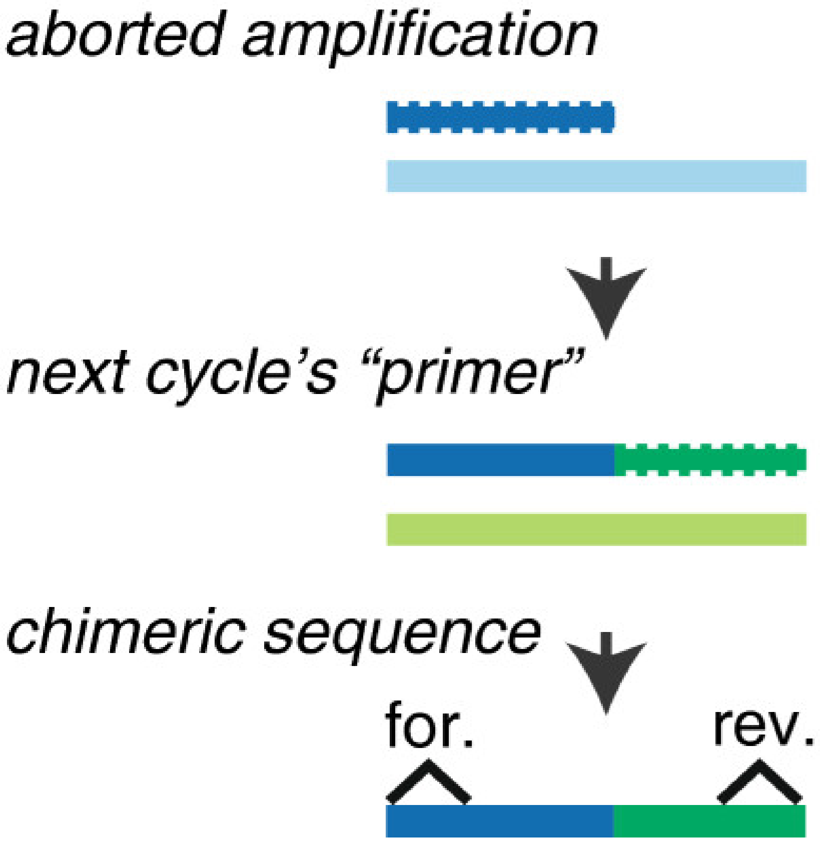
Open in FigTree v1.4.4

Chimera removal tool

What is chimera ?

PCR-generated chimeras are typically created when an aborted amplicon acts as a primer for a heterologous template. Subsequent chimeras are about the same length as the non-chimeric amplicon and contain the forward (for.) and reverse (rev.) primer sequence at each end of the amplicon.

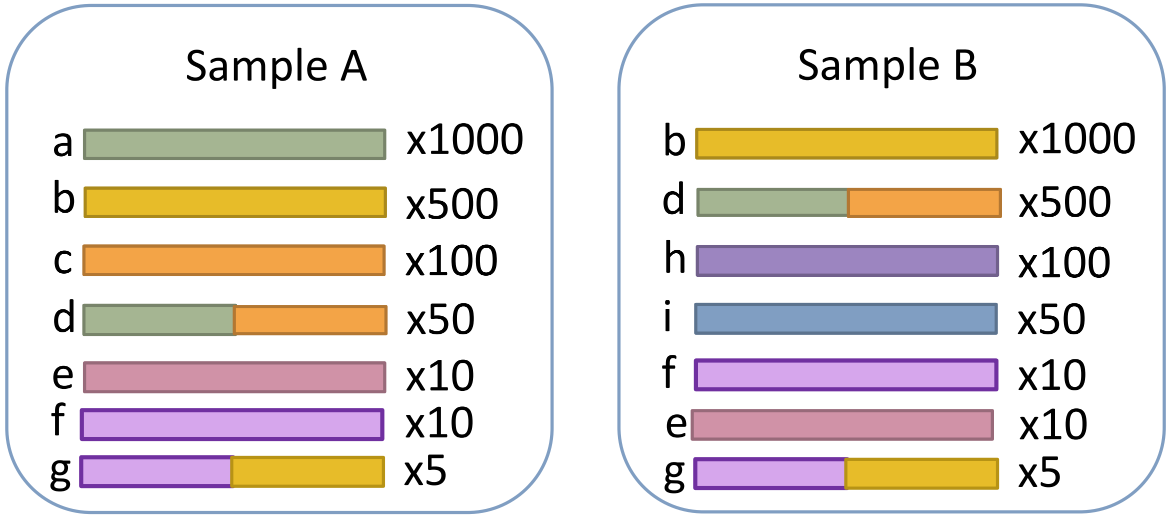
Chimera: from 5 to 45% of reads (Haas 2011 doi: 10.1101/gr.112730.110)



Fichot and Norman *Microbiome* 2013 **1**:10
doi:10.1186/2049-2618-1-10

A smart removal chimera to be accurate

We use a sample cross-validation



“ **d** ” is view as chimera by Vsearch
Its “ parents ” are presents

“ **d** ” is view as normal sequence by Vsearch
Its “ parents ” are absents

⇒ For FROGS “d” is not a chimera
⇒ For FROGS “g” is a chimera, “g” is removed
⇒ FROGS increases the detection specificity

Practice:

LAUNCH THE REMOVE CHIMERA TOOL

Exercise

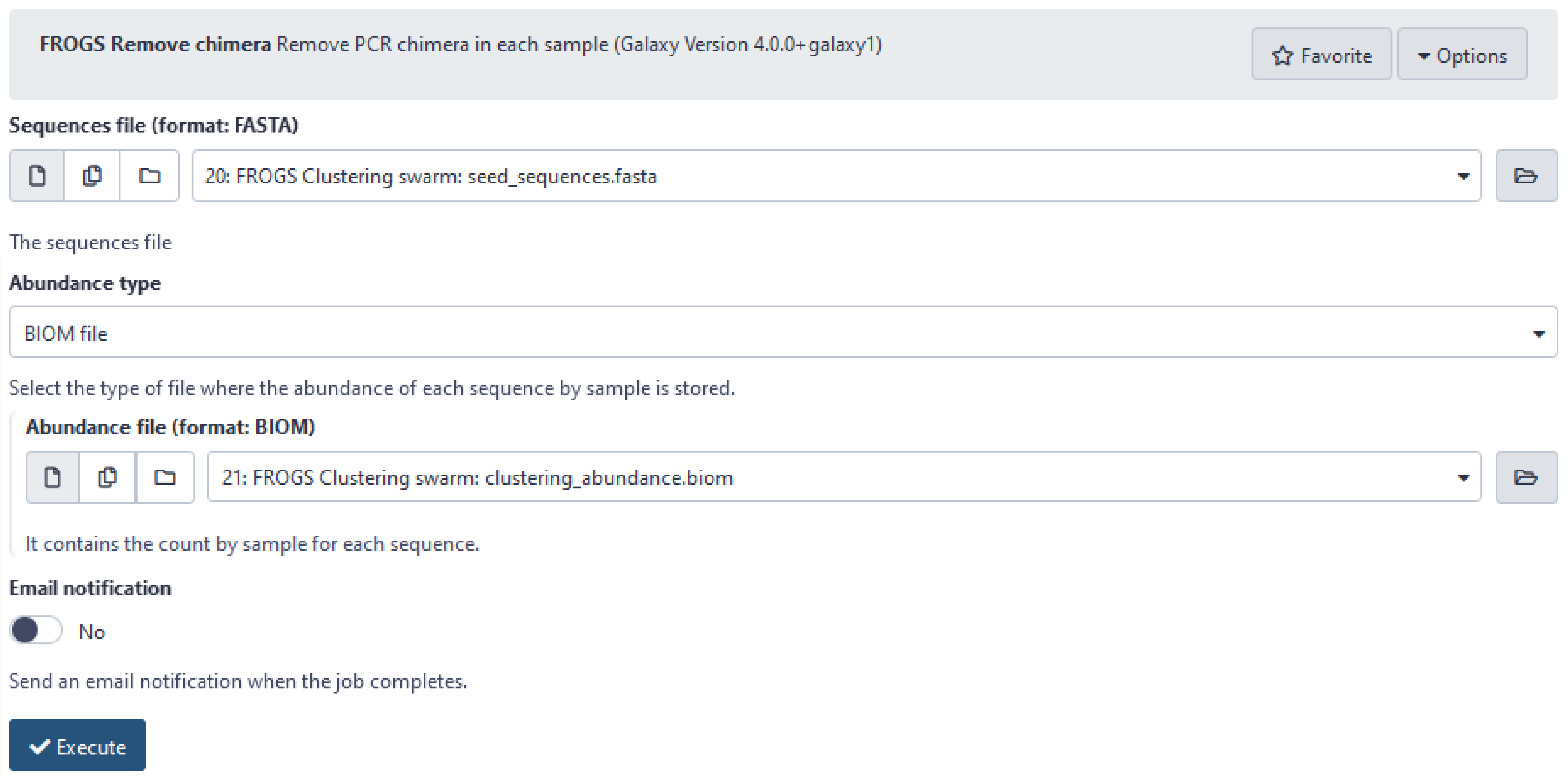
Go to « 16S » history

Launch the « FROGS Remove Chimera » tool

Follow by the « FROGS ClusterStat » tool

→ objectives :

- understand the efficiency of the chimera removal
- make links between small abundant OTUs and chimeras



FROGS Remove chimera Remove PCR chimera in each sample (Galaxy Version 4.0.0+galaxy1)

Sequences file (format: FASTA)

20: FROGS Clustering swarm: seed_sequences.fasta

The sequences file

Abundance type

BIOM file

Select the type of file where the abundance of each sequence by sample is stored.

Abundance file (format: BIOM)

21: FROGS Clustering swarm: clustering_abundance.biom

It contains the count by sample for each sequence.

Email notification

No

Send an email notification when the job completes.

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?

Exercise

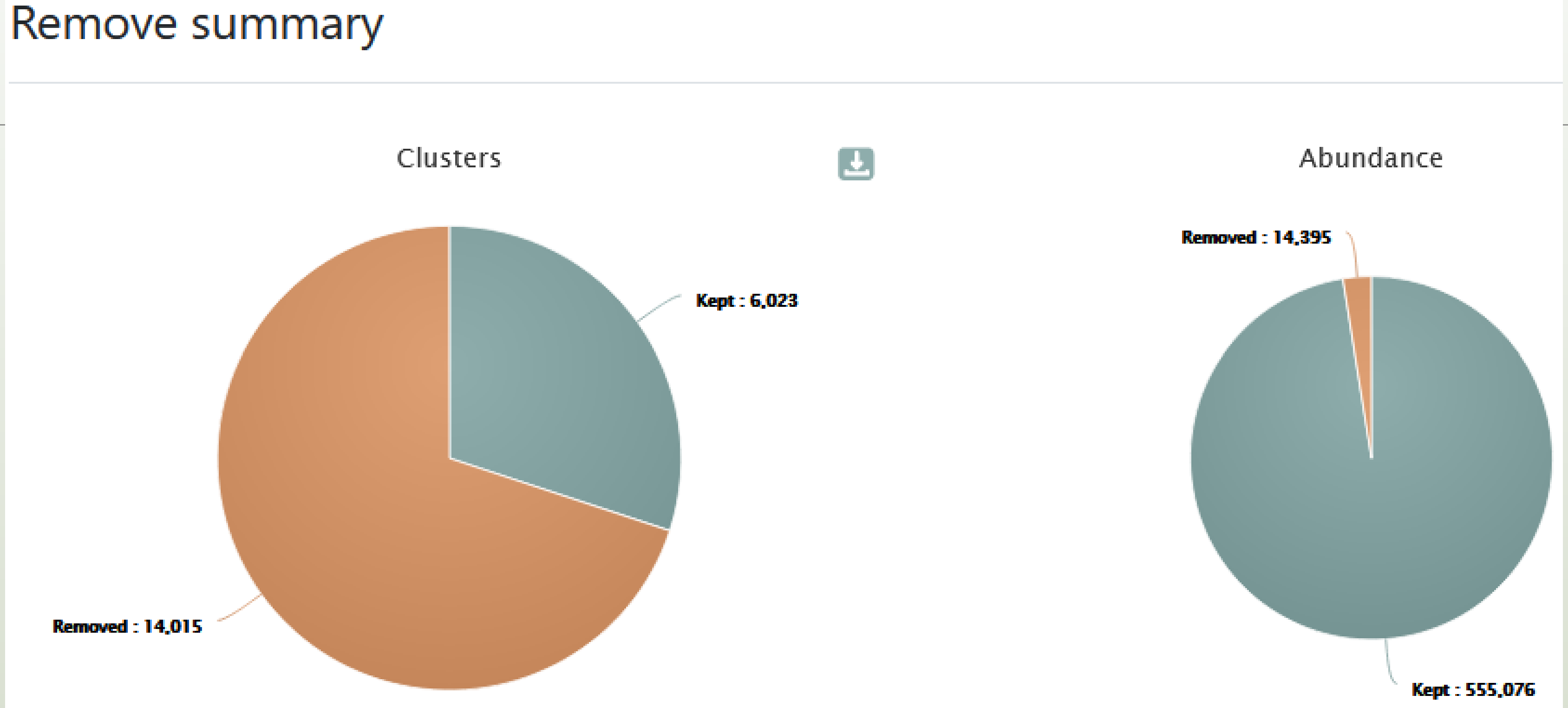
3. Rename html output in Chimera_report.html

Launch « FROGS ClusterStat » tool on non_chimera_abundance.biom

4. Compare the HTML files
   a. Of what are mainly composed singleton ? (compare with previous report.html)
   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 ?


Remove summary

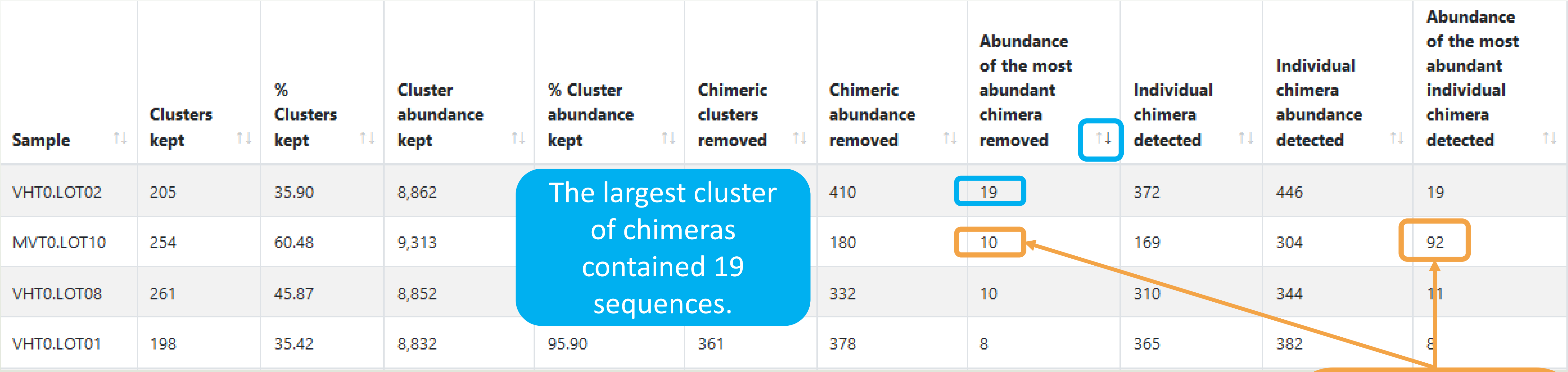
6023 clusters are kept.
The 14015 removed clusters represent ~2.5 % of sequences

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?



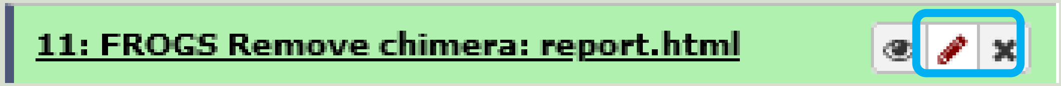
Sample	Clusters kept	% Clusters kept	Cluster abundance kept	% Cluster abundance kept	Chimeric clusters removed	Chimeric abundance removed	Abundance of the most abundant chimera removed	Individual chimera detected	Individual chimera abundance detected	Abundance of the most abundant individual chimera 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.

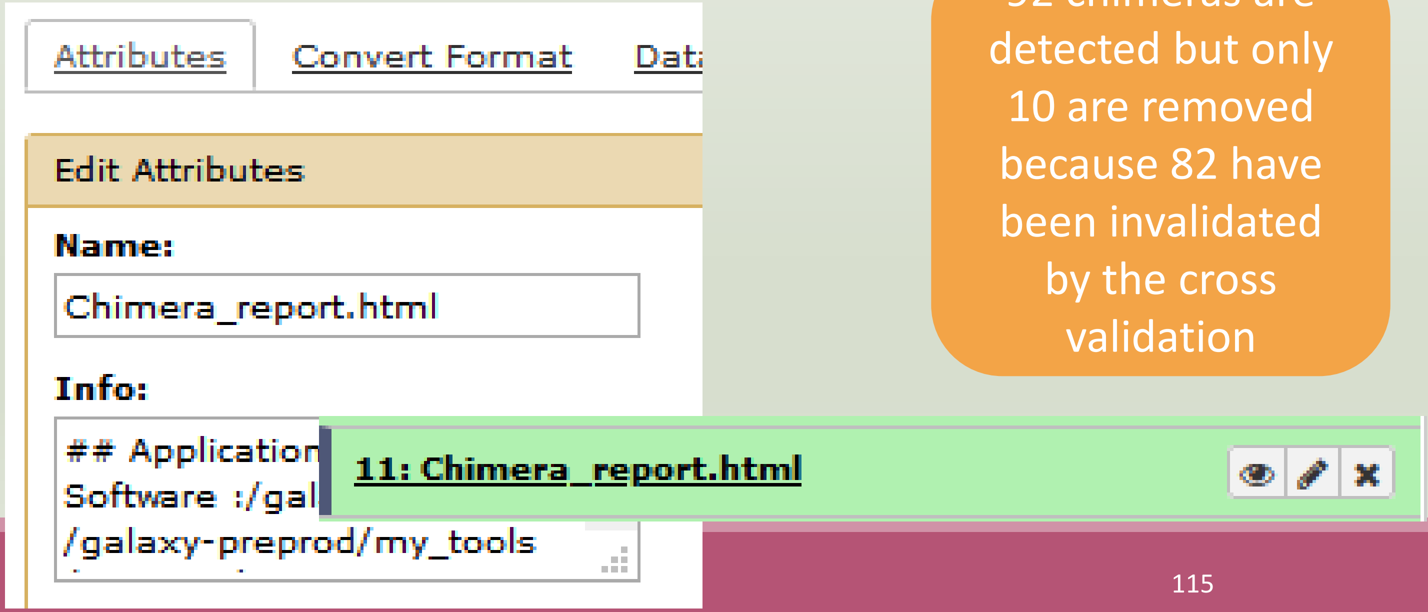
92 chimeras are detected but only 10 are removed because 82 have been invalidated by the cross validation

Q3: Rename html output in Chimera_report.html

Answer 3



11: FROGS Remove chimera: report.html



Edit Attributes

Name:

Chimera_report.html

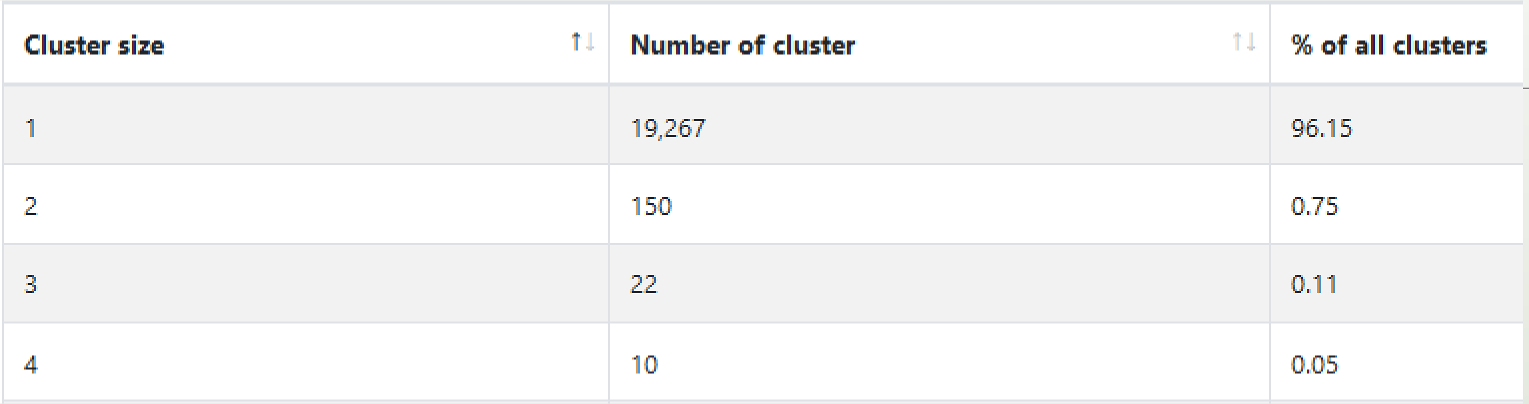
Info:

```
## Application
Software :/gal
/galaxy-preprod/my_tools
```

11: Chimera_report.html

Answer 4

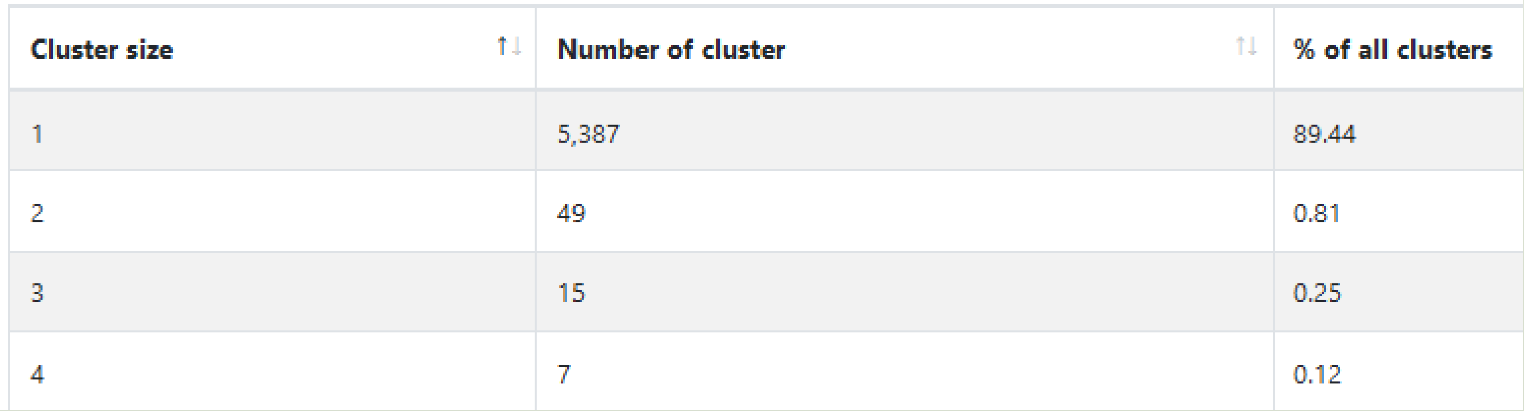
Q4a: Of what are mainly composed singleton ? (compare with previous report.html)
Q4b: What are their abundance?
Q4c: What do you conclude ?



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

OTU Filter tool

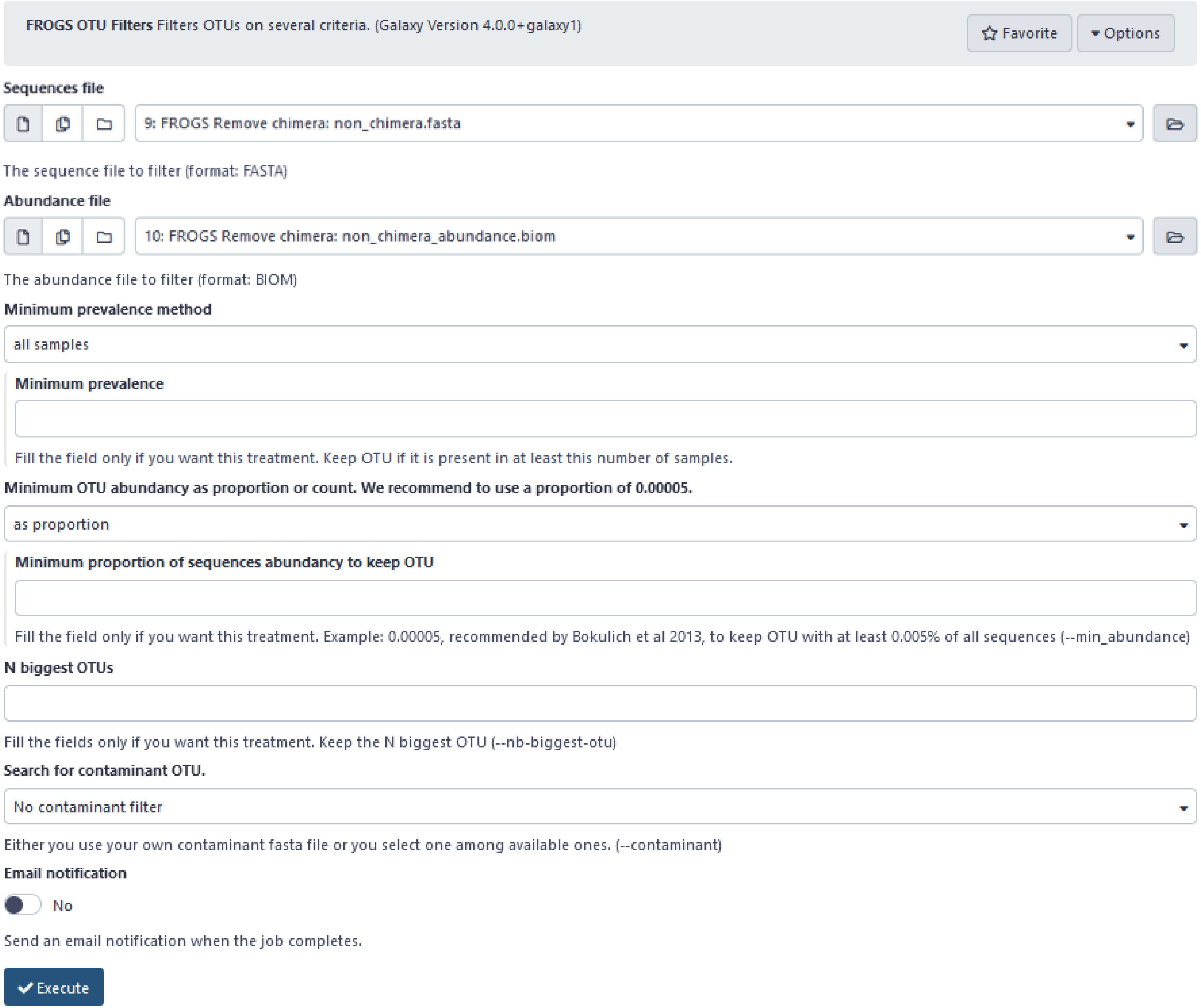
OTU Filter

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

Criteria:

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

One tool, 4 criteria



FROGS OTU Filters Filters OTUs on several criteria. (Galaxy Version 4.0.0+galaxy1)

Sequences file

9: FROGS Remove chimera: non_chimera.fasta

The sequence file to filter (format: FASTA)

Abundance file

10: FROGS Remove chimera: non_chimera_abundance.biom

The abundance file to filter (format: BIOM)

1 **Minimum prevalence method**

all samples

Minimum prevalence

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

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

as proportion

Minimum proportion of sequences abundancy to keep OTU

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

3 **N biggest OTUs**

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

4 **Search for contaminant OTU.**

No contaminant filter

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

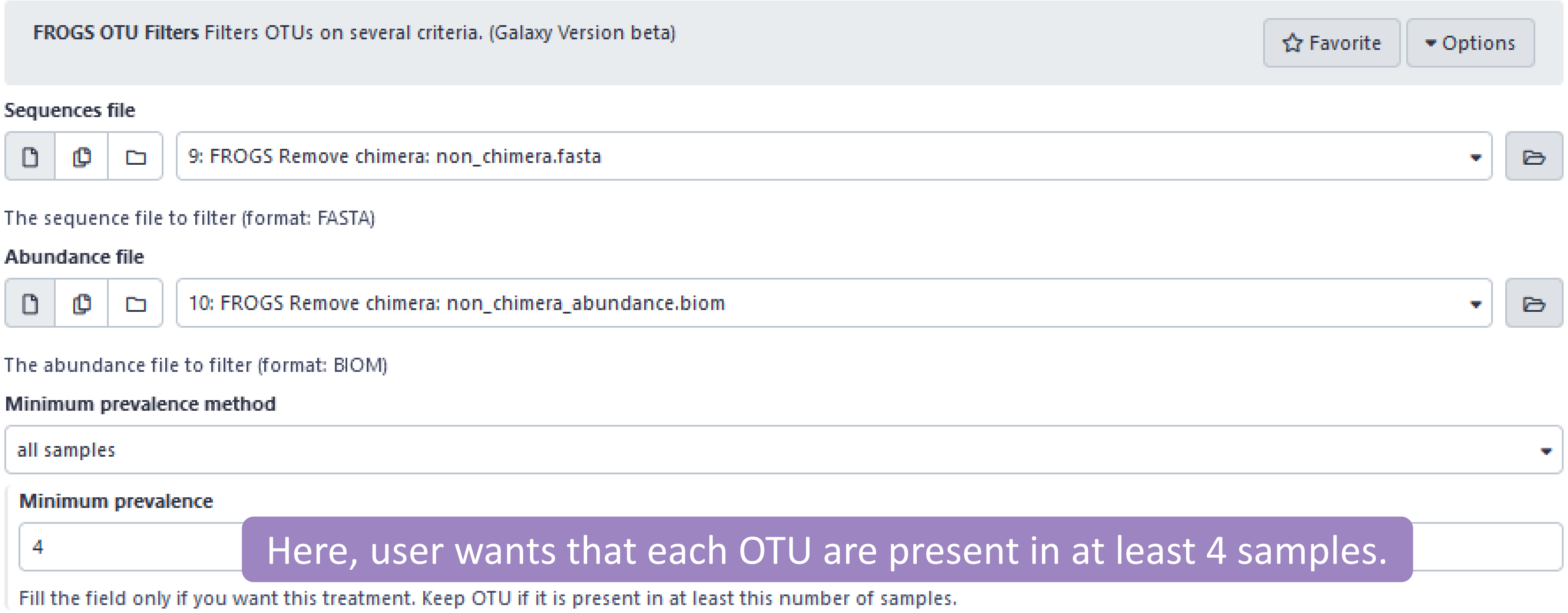
Email notification

No

Send an email notification when the job completes.

1

Prevalence filter – option 1



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

Sequences file

9: FROGS Remove chimera: non_chimera.fasta

The sequence file to filter (format: FASTA)

Abundance file

10: FROGS Remove chimera: non_chimera_abundance.biom

The abundance file to filter (format: BIOM)

Minimum prevalence method

all samples

Minimum prevalence

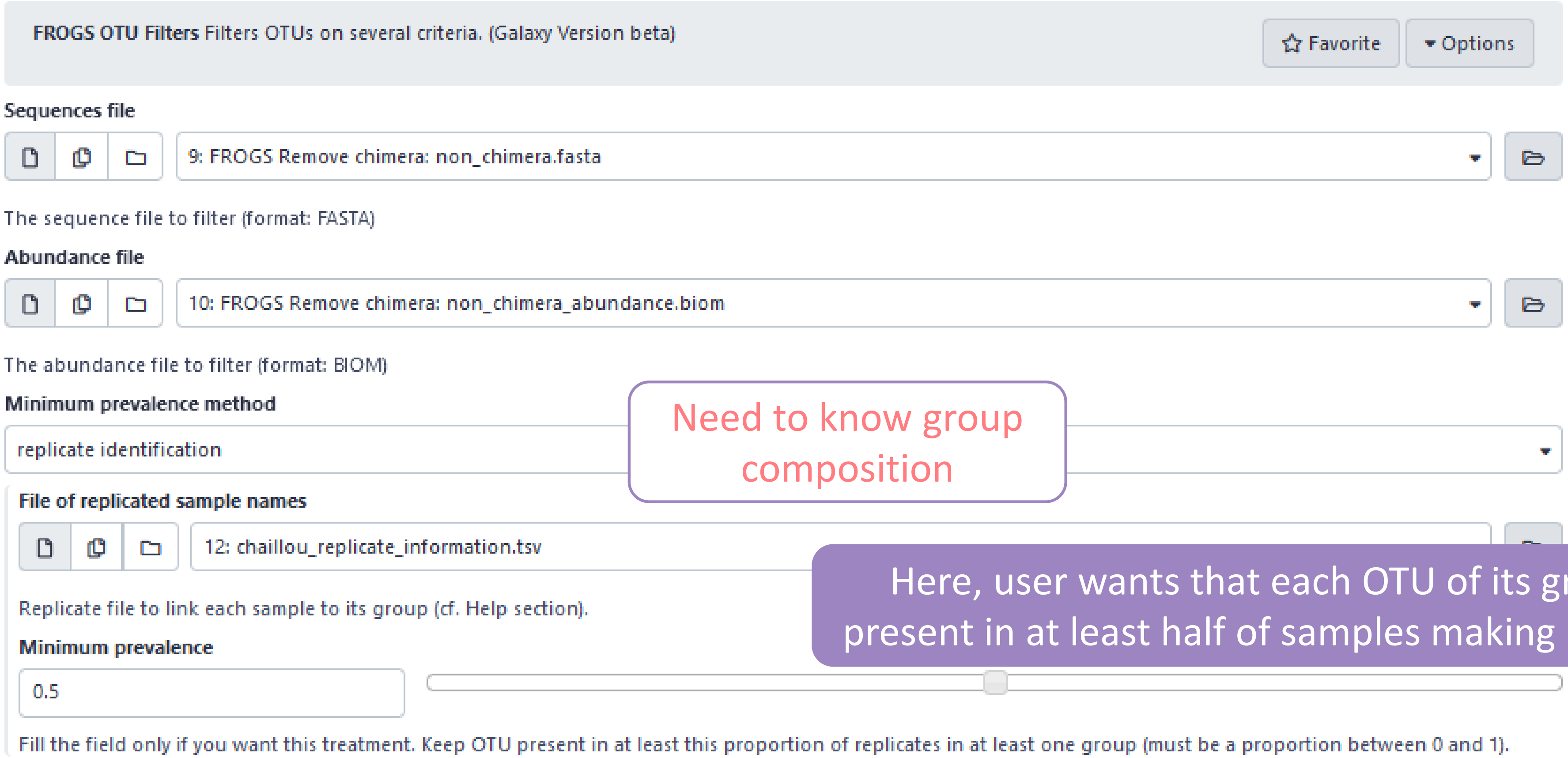
4

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

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

1

Prevalence filter – option 2



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

Sequences file

9: FROGS Remove chimera: non_chimera.fasta

The sequence file to filter (format: FASTA)

Abundance file

10: FROGS Remove chimera: non_chimera_abundance.biom

The abundance file to filter (format: BIOM)

Minimum prevalence method

replicate identification

File of replicated sample names

12: chaillou_replicate_information.tsv

Replicate file to link each sample to its group (cf. Help section).

Minimum prevalence

0.5

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

Need to know group composition

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

1

Prevalence filter – option 2

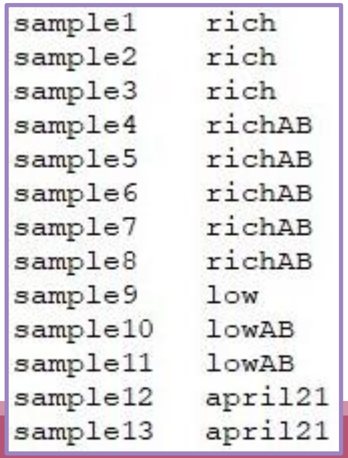
How to build the file of replicated sample names ?

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

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

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

Example:



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

Thanks to get data tool, add it in your history

1

Prevalence filter – option 2

Results:

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

The process will therefore keep the OTUs present in at least

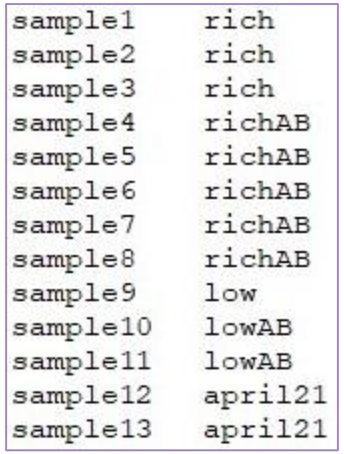
2 "rich" samples

3 "richAB" samples,

1 "lowAB" sample

1 "april21" sample

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

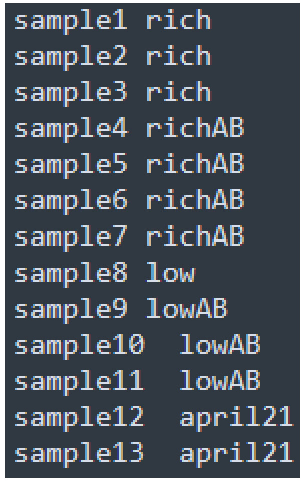


sample1	rich
sample2	rich
sample3	rich
sample4	richAB
sample5	richAB
sample6	richAB
sample7	richAB
sample8	richAB
sample9	low
sample10	lowAB
sample11	lowAB
sample12	april21
sample13	april21

1

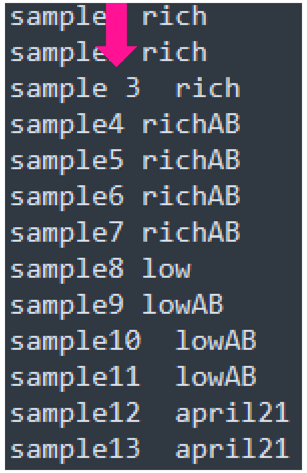
Prevalence filter – option 2

mistakes not to be made:



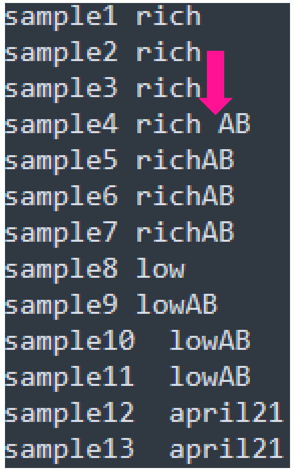
```
sample1 rich
sample2 rich
sample3 rich
sample4 richAB
sample5 richAB
sample6 richAB
sample7 richAB
sample8 low
sample9 lowAB
sample10  lowAB
sample11  lowAB
sample12  april21
sample13  april21
```

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

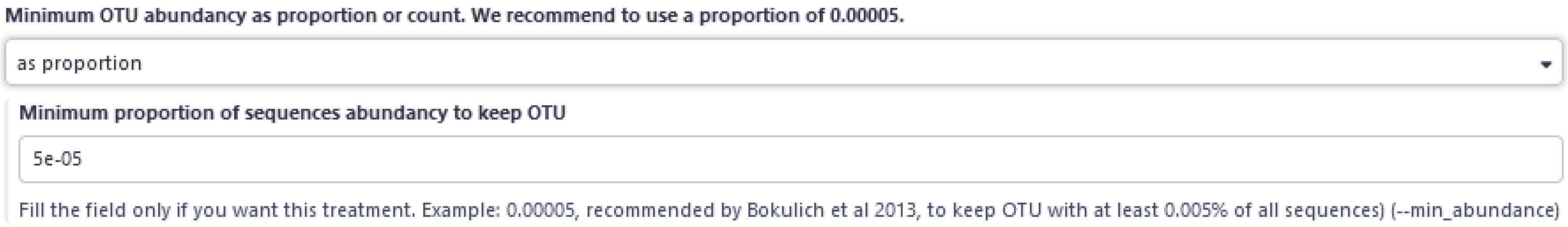


```
sample1 rich
sample2 rich
sample3 rich
sample4 rich AB
sample5 richAB
sample6 richAB
sample7 richAB
sample8 low
sample9 lowAB
sample10  lowAB
sample11  lowAB
sample12  april21
sample13  april21
```

Creates artificially 3 columns

2

OTU size filter



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

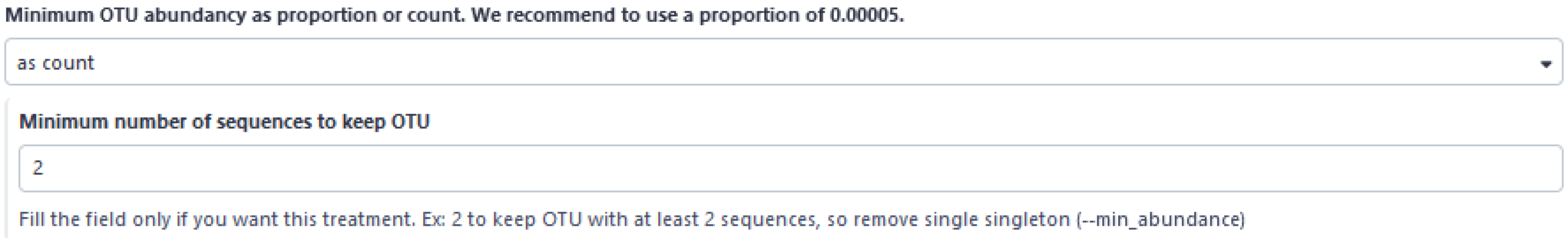
as proportion

Minimum proportion of sequences abundancy to keep OTU

5e-05

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

OR



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

as count

Minimum number of sequences to keep OTU

2

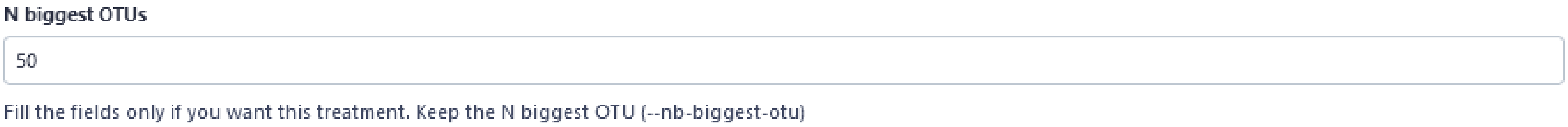
Fill the field only if you want this treatment. Ex: 2 to keep OTU with at least 2 sequences, so remove single singleton (--min_abundance)

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

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

3

Filter : Keep biggest OTU



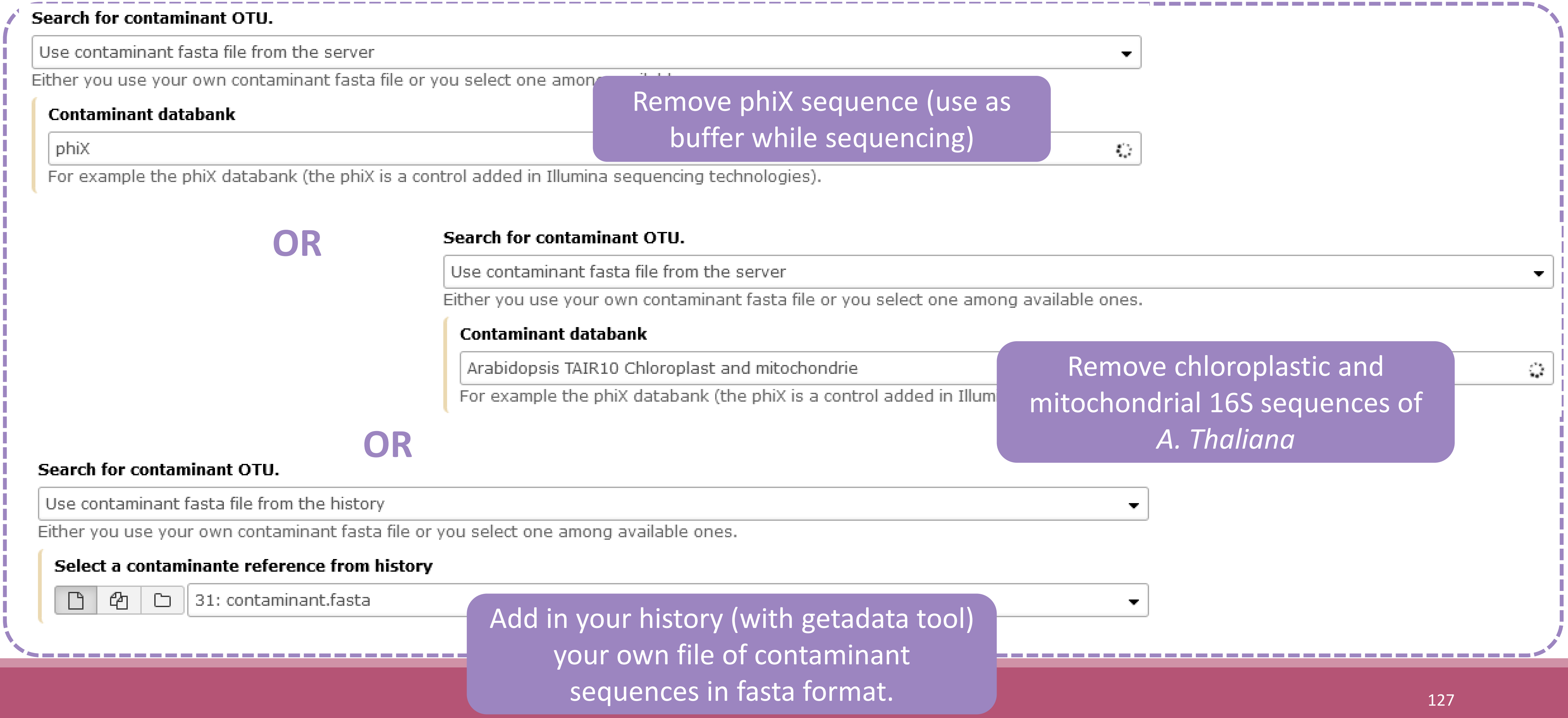
N biggest OTUs

50

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

Here, user wants to keep the 50 biggest OTUs.

4 Contaminant filter



Search for contaminant OTU.

Use contaminant fasta file from the server

Either you use your own contaminant fasta file or you select one amon

Contaminant databank

phiX

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

Remove phiX sequence (use as buffer while sequencing)

OR

Search for contaminant OTU.

Use contaminant fasta file from the server

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

Contaminant databank

Arabidopsis TAIR10 Chloroplast and mitochondrie

For example the phiX databank (the phiX is a control added in Illum

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

OR

Search for contaminant OTU.

Use contaminant fasta file from the history

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

Select a contaminante reference from history

31: contaminant.fasta

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

Practice:

LAUNCH THE OTU FILTER TOOL

Exercice:

Go to history « 16S » history

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

Use 3 criteria to filter OTUs:

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

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

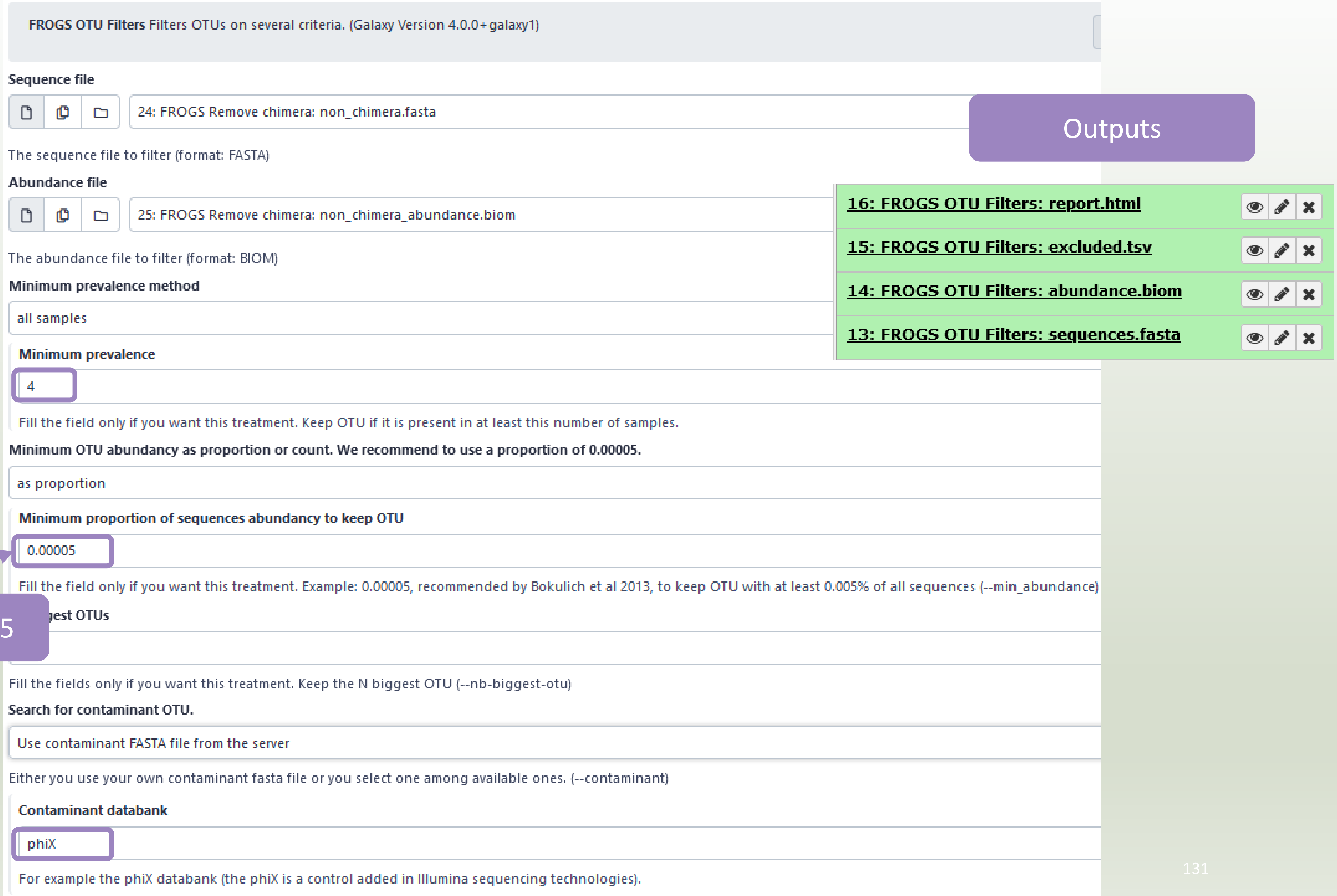
[1] *Nat Methods. 2013 Jan;10(1):57-9. doi: 10.1038/nmeth.2276. Epub 2012 Dec 2.*
Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing.
Bokulich NA1, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, Mills DA, Caporaso JG.

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

Answer 1



FROGS OTU Filters Filters OTUs on several criteria. (Galaxy Version 4.0.0+galaxy1)

Sequence file

24: FROGS Remove chimera: non_chimera.fasta

The sequence file to filter (format: FASTA)

Abundance file

25: FROGS 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 OTU if it is present in at least this number of samples.

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

as proportion

Minimum proportion of sequences abundancy to keep OTU

0.00005

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

gest OTUs

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

Search for contaminant OTU.

Use contaminant FASTA file from the server

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

Contaminant databank

phiX

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

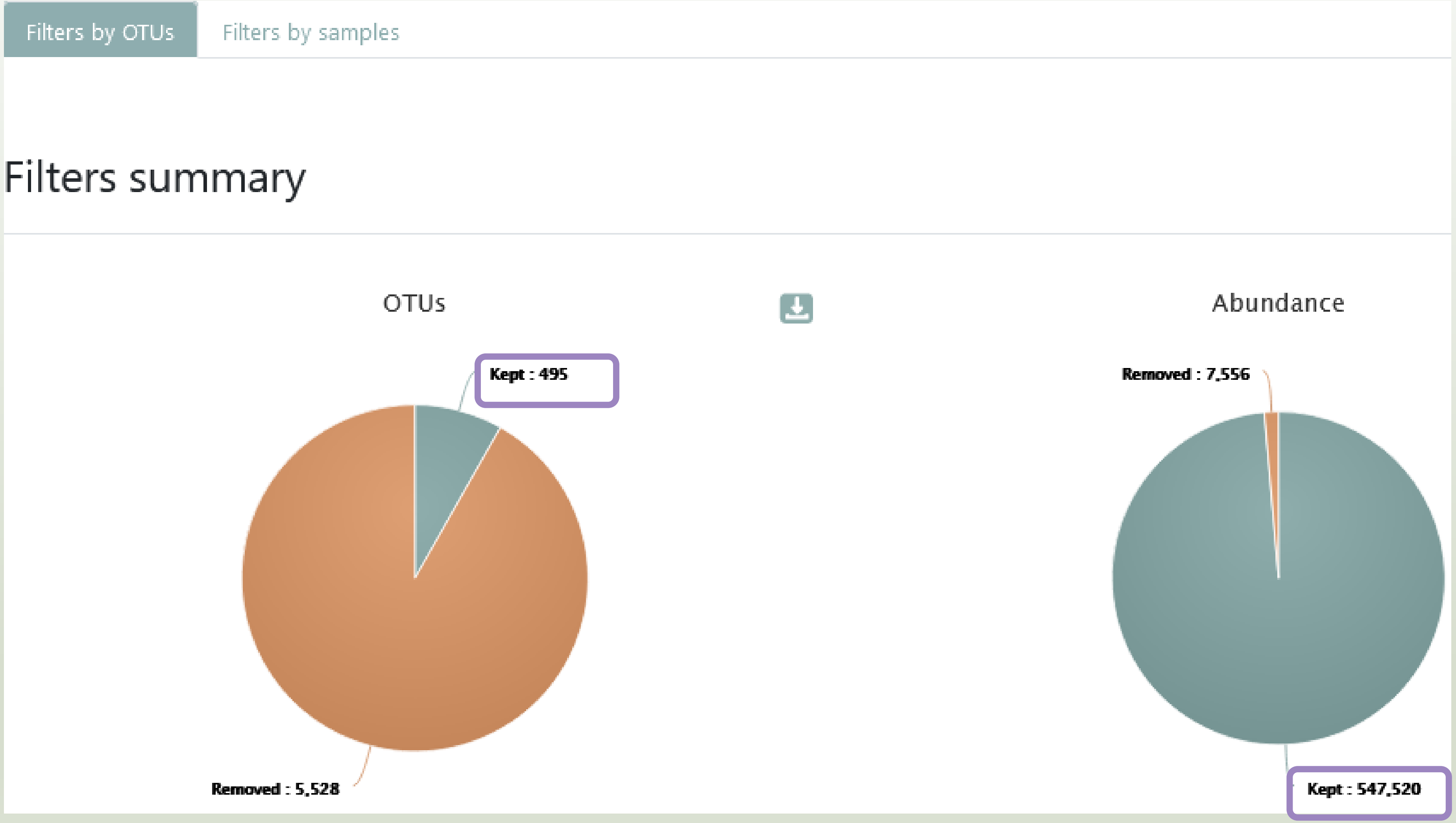
0.005% = 0.00005

Outputs

- 16: FROGS OTU Filters: report.html
- 15: FROGS OTU Filters: excluded.tsv
- 14: FROGS OTU Filters: abundance.biom
- 13: FROGS OTU Filters: sequences.fasta

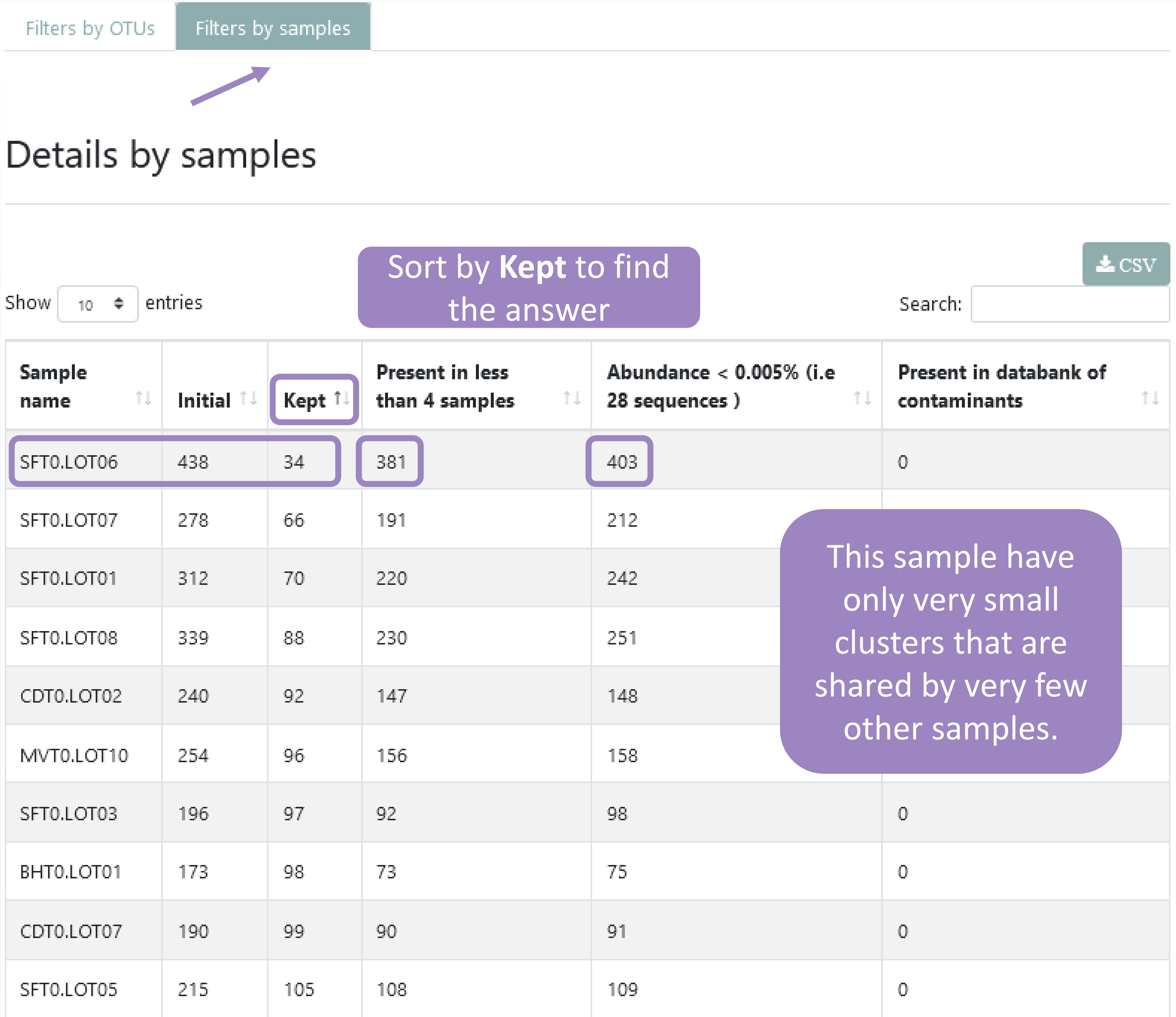
Answer 2

Two tabs to explore



Filters summary

Answer 2



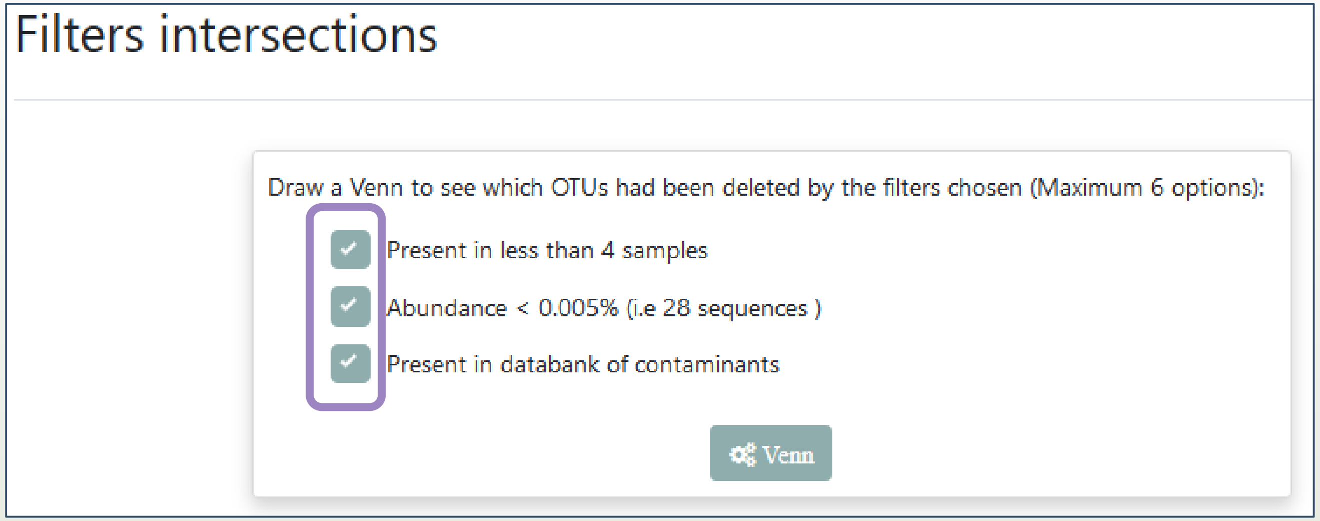
Details by samples

Sort by **Kept** to find the answer

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

This sample have only very small clusters that are shared by very few other samples.

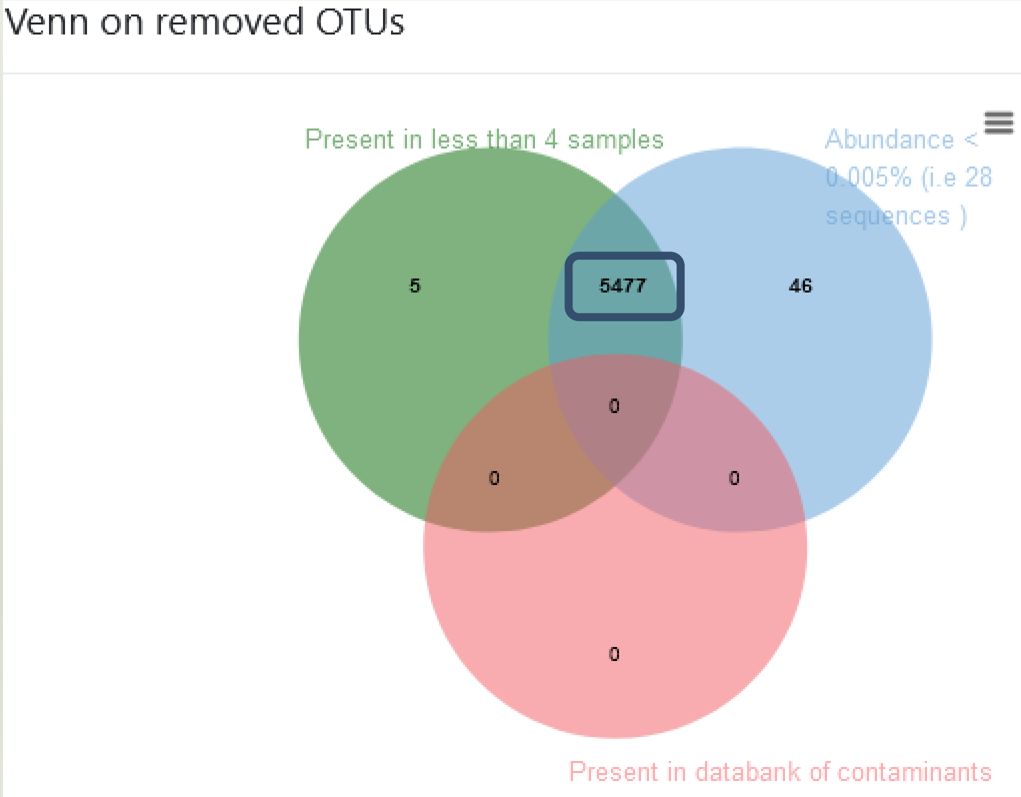
Answer 3



Filters intersections

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

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

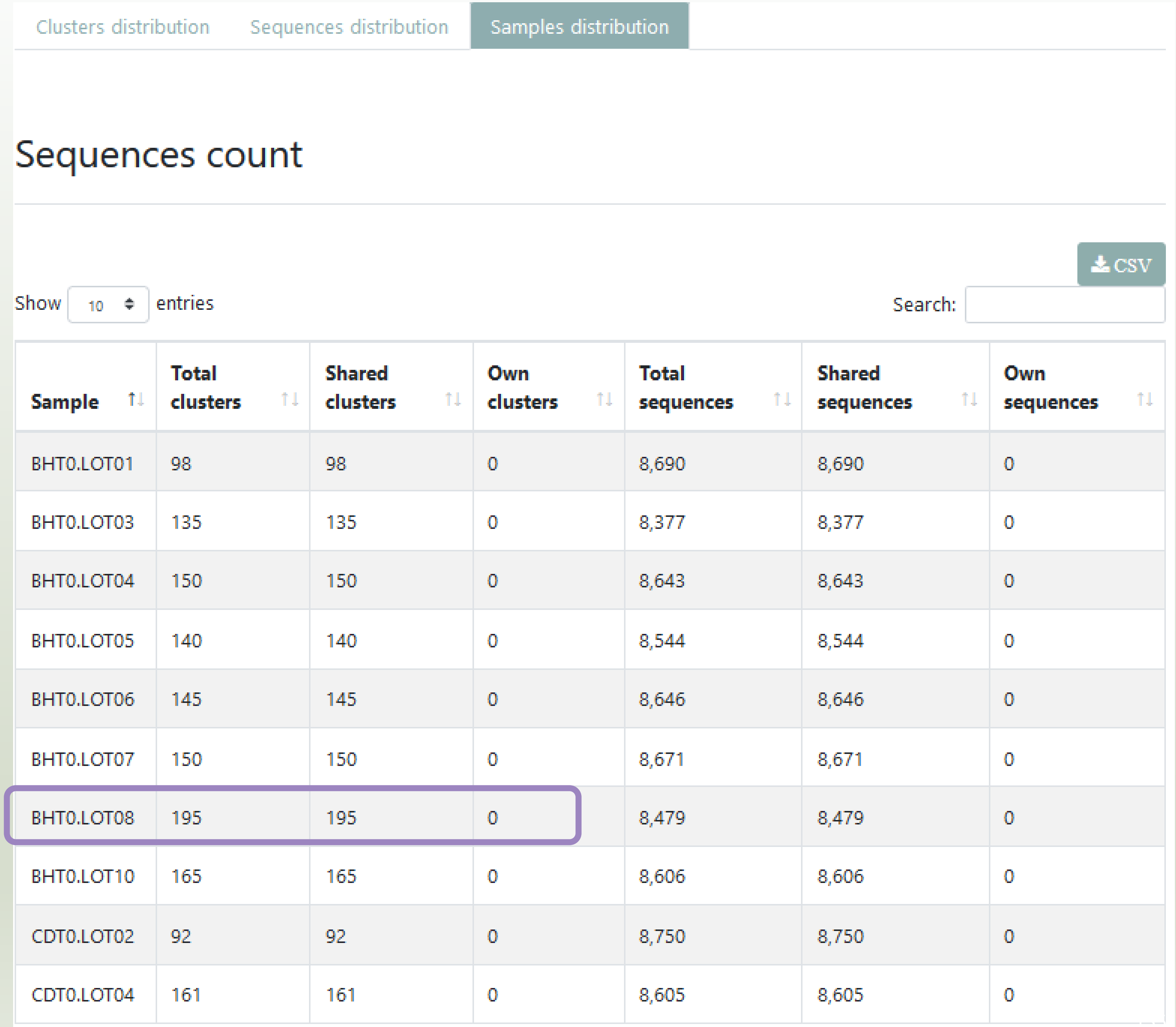


Venn on removed OTUs

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

Answer 4

report.html of **ClusterStat tool**



Sequences count

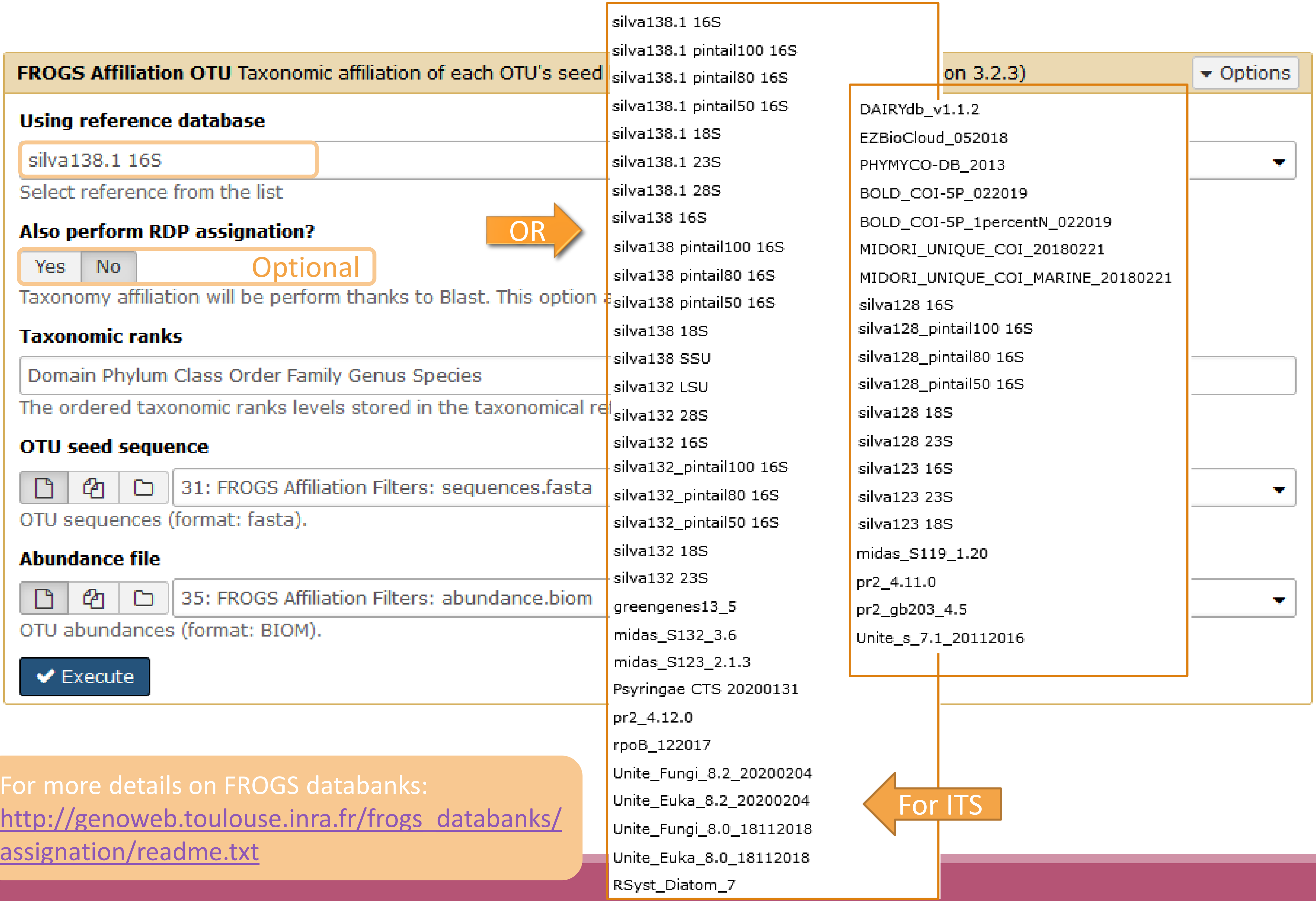
Show 10 entries

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

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

Affiliation tool



FROGS Affiliation OTU Taxonomic affiliation of each OTU's seed … on 3.2.3)

Using reference database

silva138.1 16S

Select reference from the list

Also perform RDP assignation?

Yes No

Optional

Taxonomy affiliation will be perform thanks to Blast. This option a

Taxonomic ranks

Domain Phylum Class Order Family Genus Species

The ordered taxonomic ranks levels stored in the taxonomical ref

OTU seed sequence

31: FROGS Affiliation Filters: sequences.fasta

OTU sequences (format: fasta).

Abundance file

35: FROGS Affiliation Filters: abundance.biom

OTU abundances (format: BIOM).

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

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

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

1 Cluster = 2 affiliations

RDPClassifier*: one affiliation with bootstrap, on each taxonomic subdivision.

Bacteria;(1.0);Actinobacteriota;(1.0);Actinobacteria;(1.0);Propionibacteriales;(1.0);Propionibacteriaceae;(1.0);Cutibacterium;(1.0);Cutibacterium acnes;(0.57);

NCBI Blastn+** : one affiliation with identity %, coverage %, e-value, alignment length and a special tag “**Multi-affiliation**”.

Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Multi-affiliation

Identity: 100% and Coverage: 100%

* Appl. Environ. Microbiol. August 2007 vol. 73 no. 16 5261-5267. doi : 10.1128/AEM.00062-07
Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy.
Qiong Wang, George M.Garrity, James M. Tiedje and James R. Cole

** BMC Bioinformatics 2009, 10:421. doi:10.1186/1471-2105-10-421
BLAST+: architecture and applications
Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos,Kevin Bealer and Thomas L Madden

Affiliation Strategy of FROGS

Blastn+ with “**Multi-affiliation**” management

Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus

Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus saprophyticus

Strictly identical (V1-V3 amplification) on 499 nucleotides

Which one to choose?

Affiliation Strategy of FROGS

Blastn+ with “**Multi-affiliation**” management

Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus xylosus

Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;Staphylococcus saprophyticus

Strictly identical (V1-V3 amplification) on 499 nucleotides

Bacteria;Firmicutes;Bacilli;Staphylococcales;Staphylococcaceae;Staphylococcus;**Multi-affiliation**

We cannot choose without preconceived ideas.

Practice:

LAUNCH THE FROGS AFFILIATION TOOL

Exercice:

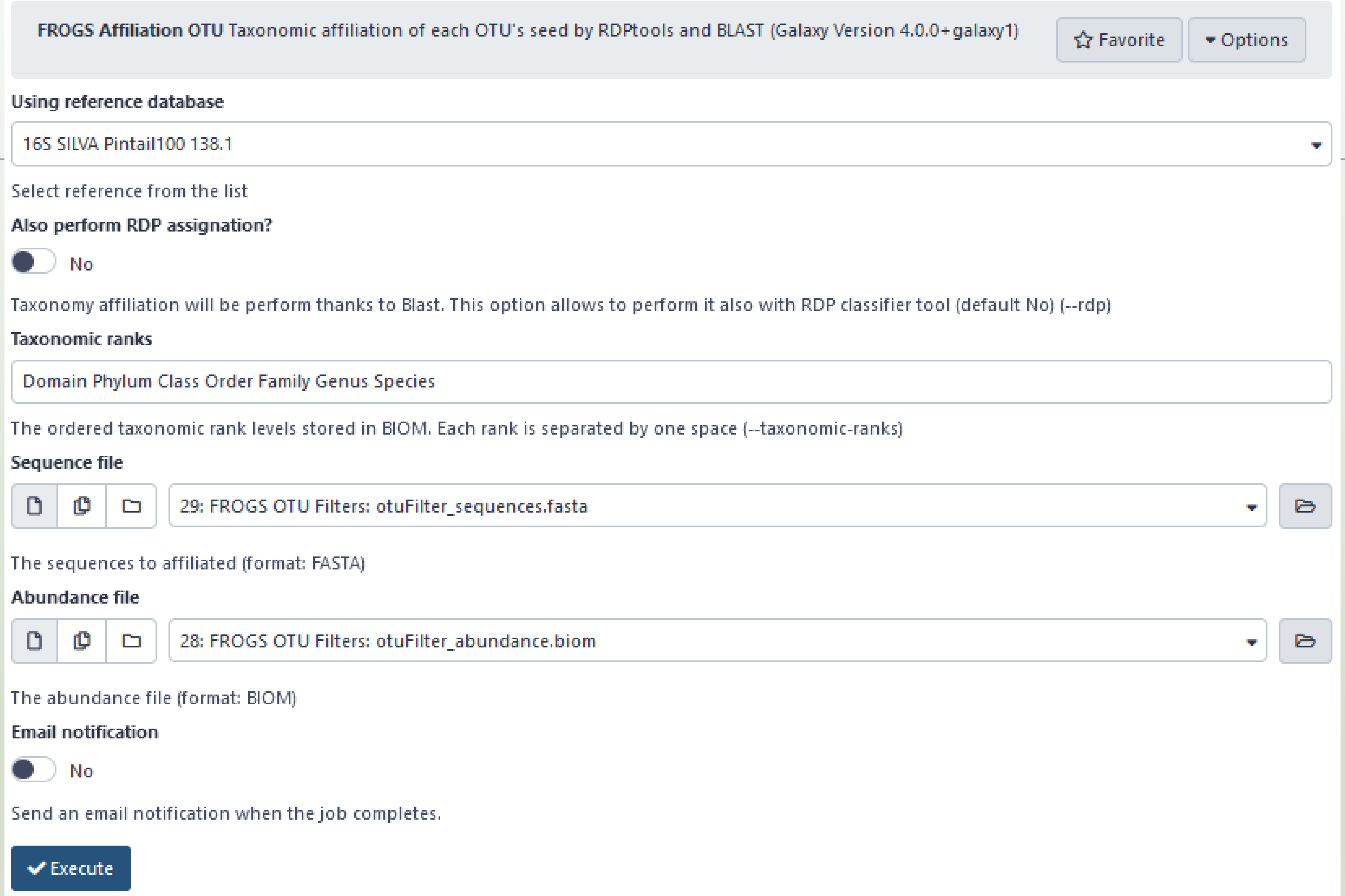
Go to history « 16S » history

Launch the « FROGS Affiliation » tool with

- SILVA 138.1 16S database pintail 100

→ objectives :

  - understand abundance tables columns
  - understand the BLAST affiliation



FROGS Affiliation OTU Taxonomic affiliation of each OTU's seed by RDPtools and BLAST (Galaxy Version 4.0.0+galaxy1)

Using reference database

16S SILVA Pintail100 138.1

Select reference from the list

Also perform RDP assignation?

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

29: FROGS OTU Filters: otuFilter_sequences.fasta

The sequences to affiliated (format: FASTA)

Abundance file

28: FROGS OTU Filters: otuFilter_abundance.biom

The abundance file (format: BIOM)

Email notification

No

Send an email notification when the job completes.

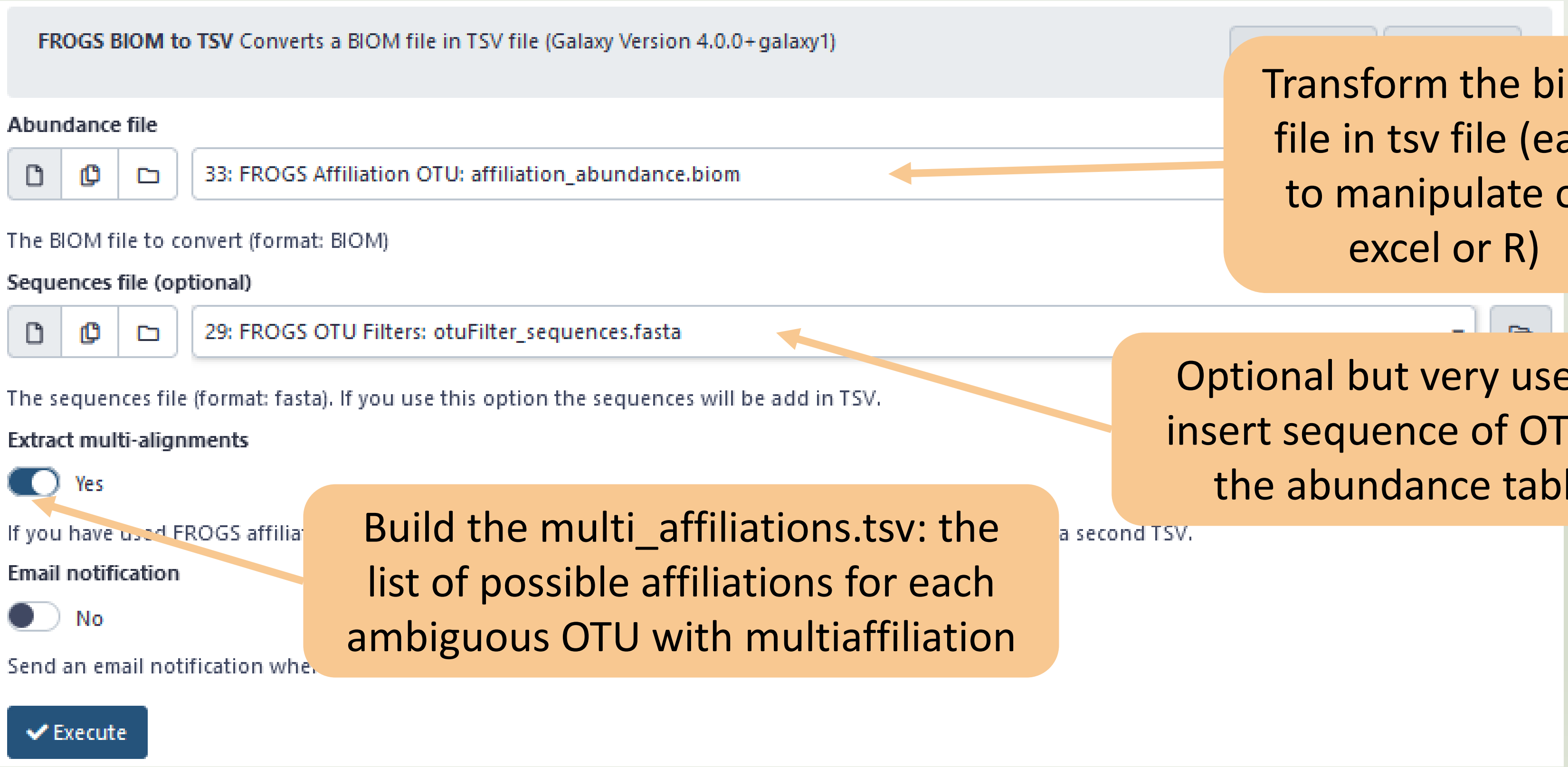
Exercise

1. What are the « **FROGS Affiliation tool** » output files ?
2. How many sequences are affiliated by BLAST ?
3. How many OTU have a “multiaffiliation” at Order ranks ?
4. Click on the « eye » button on the BIOM output file, what do you understand ?

👁

Exercise

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



FROGS BIOM to TSV Converts a BIOM file in TSV file (Galaxy Version 4.0.0+galaxy1)

Abundance file

33: FROGS Affiliation OTU: affiliation_abundance.biom

The BIOM file to convert (format: BIOM)

Sequences file (optional)

29: FROGS OTU Filters: otuFilter_sequences.fasta

The sequences file (format: fasta). If you use this option the sequences will be add in TSV.

Extract multi-alignments

● Yes

If you have used FROGS affilia a second TSV.

Email notification

○ No

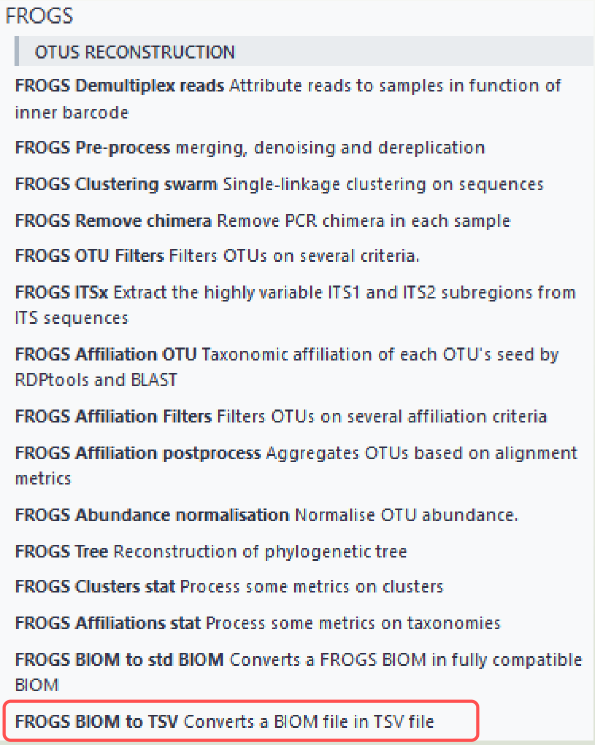
Send an email notification whe

✓ Execute

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

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

Build the multi_affiliations.tsv: the list of possible affiliations for each ambiguous OTU with multiaffiliation



FROGS

OTUS RECONSTRUCTION

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

FROGS Pre-process merging, denoising and dereplication

FROGS Clustering swarm Single-linkage clustering on sequences

FROGS Remove chimera Remove PCR chimera in each sample

FROGS OTU Filters Filters OTUs on several criteria.

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

FROGS Affiliation OTU Taxonomic affiliation of each OTU's seed by RDPtools and BLAST

FROGS Affiliation Filters Filters OTUs on several affiliation criteria

FROGS Affiliation postprocess Aggregates OTUs based on alignment metrics

FROGS Abundance normalisation Normalise OTU abundance.

FROGS Tree Reconstruction of phylogenetic tree

FROGS Clusters stat Process some metrics on clusters

FROGS Affiliations 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

Exercise

👁

5. Click again on the ”eye” on the new output generated.

Or open it in your favorite spreadsheet (Excel, google sheet, Calc…) !

Now, what do you think about the file format? What does it contain?

Exercise

6. Observe and describe

- In FROGS BIOM to TSV: abundance_silva.tsv, the different columns of cluster 3

a. how would you qualify the alignment between the OTU3 seed and the sequences of the silva database?

b. What does it mean e-value = 0 ?

c. What is the header of column that shows the sequence of OTU seed ?

d. How many sequences have OTU3 in total ?

e. How many sequences have OTU3 in MVT0.LOT10 ? What is the sample where OTU3 is absent ?

Exercise

7. Observe and describe

- In FROGS BIOM to TSV: multi_affiliations.tsv, identifies the lines corresponding to cluster3

a. Why cluster3 has a multiaffiliation for species ?

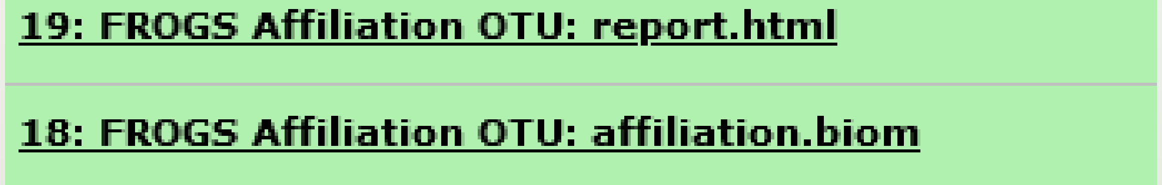
b. Why “Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Latilactobacillus;Lactobacillus sakei” is present 74 times ?

Q1: What are the « **FROGS Affiliation tool** » output files ?

Q2: How many sequences are affiliated by BLAST ?

Exercise

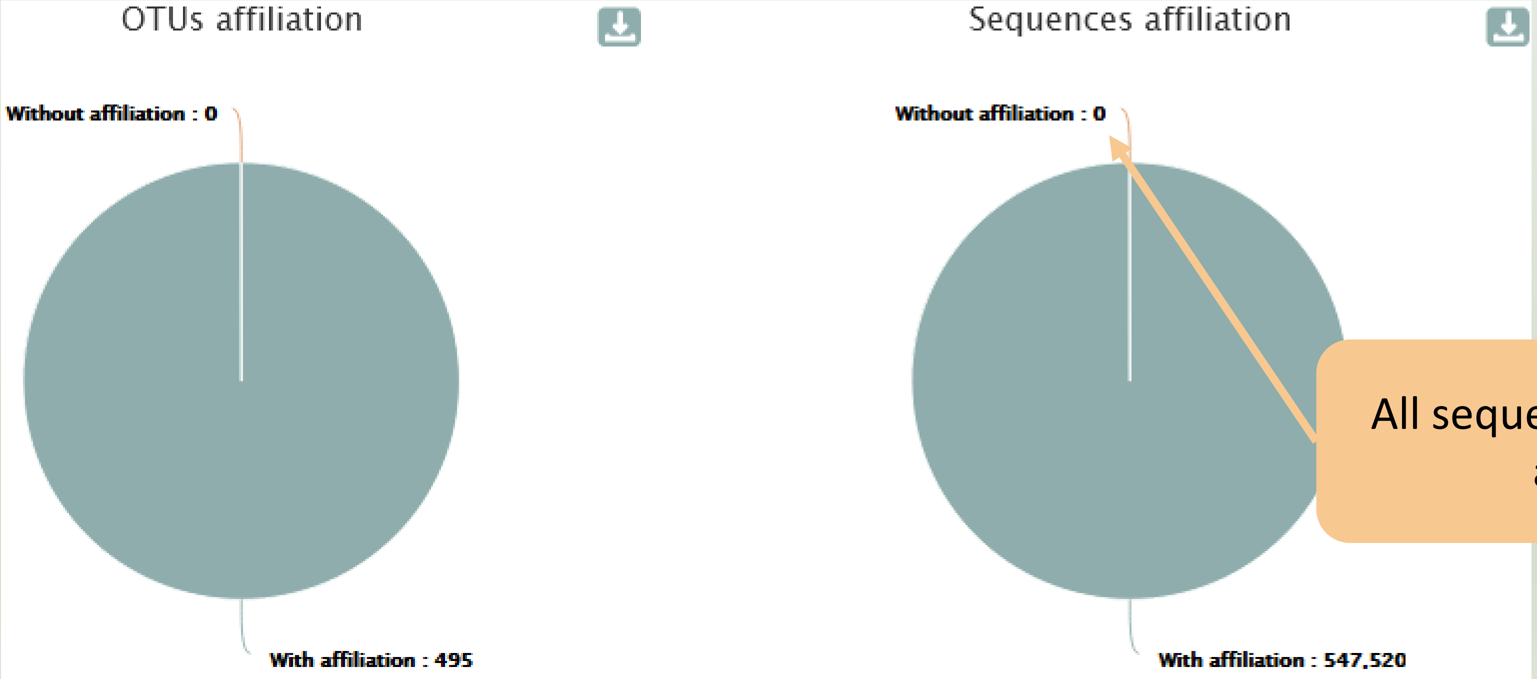
Answer 1



19: FROGS Affiliation OTU: report.html

18: FROGS Affiliation OTU: affiliation.biom

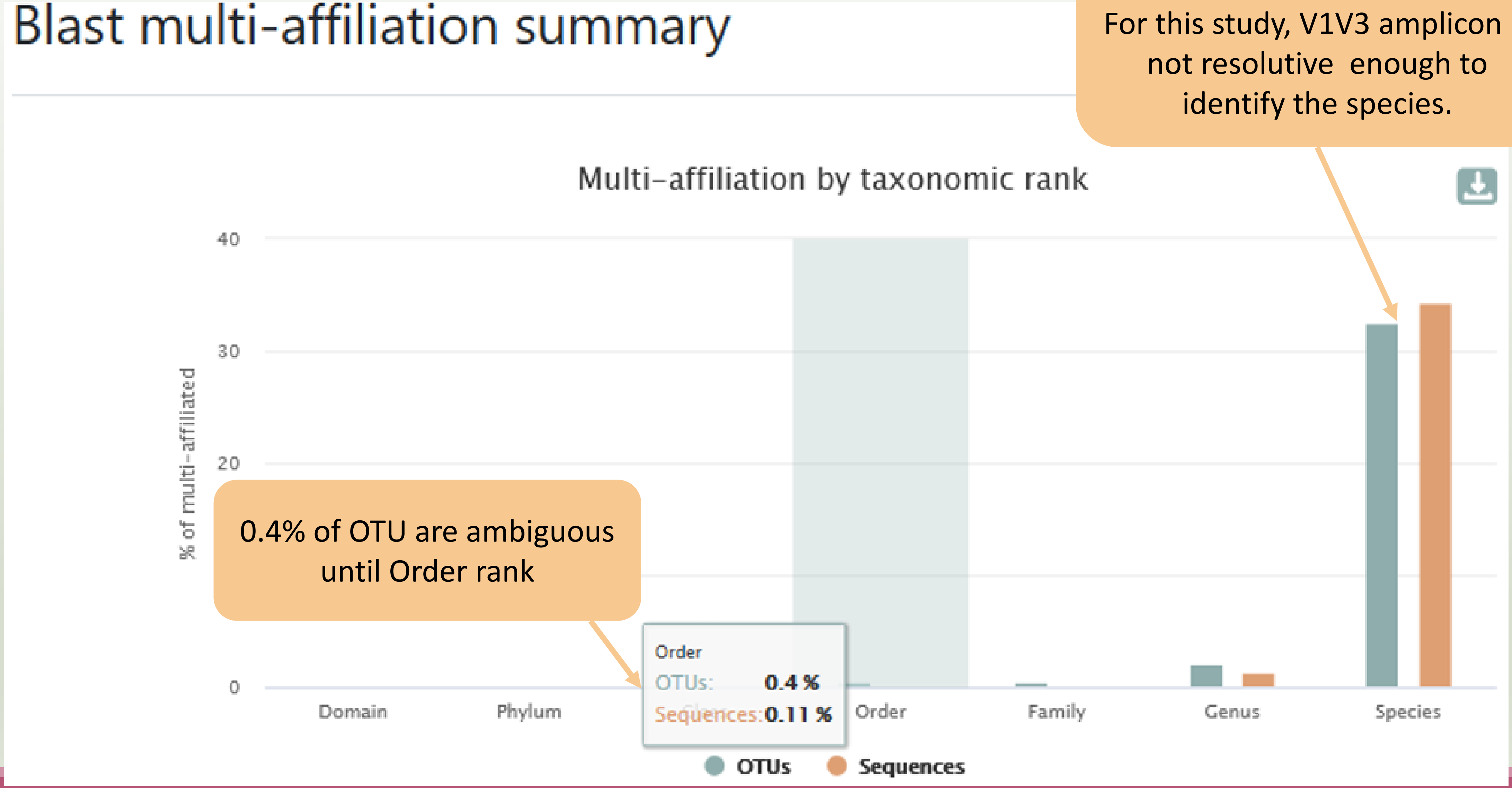
Answer 2



All sequences have a blast affiliation

Answer 3

Q3: How many OTU have a “multiaffiliation” at Order ranks ?



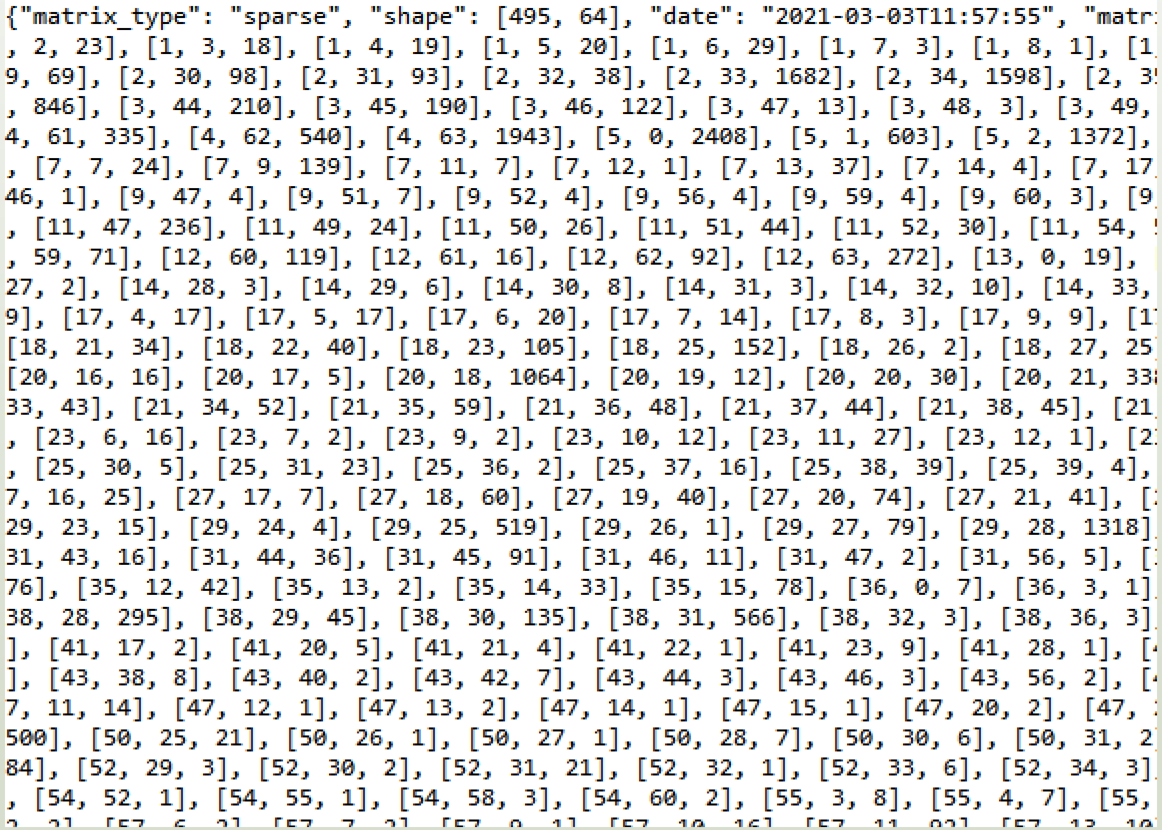
Blast multi-affiliation summary

Most of OTUs are ambiguous at species rank.
For this study, V1V3 amplicon is not resolutive enough to identify the species.

0.4% of OTU are ambiguous until Order rank

Answer 4

Q4: Click on the « eye » button on the BIOM output file, what do you understand ?



```
{"matrix_type": "sparse", "shape": [495, 64], "date": "2021-03-03T11:57:55", "matr
, 2, 23], [1, 3, 18], [1, 4, 19], [1, 5, 20], [1, 6, 29], [1, 7, 3], [1, 8, 1], [1
9, 69], [2, 30, 98], [2, 31, 93], [2, 32, 38], [2, 33, 1682], [2, 34, 1598], [2, 3
, 846], [3, 44, 210], [3, 45, 190], [3, 46, 122], [3, 47, 13], [3, 48, 3], [3, 49,
4, 61, 335], [4, 62, 540], [4, 63, 1943], [5, 0, 2408], [5, 1, 603], [5, 2, 1372],
, [7, 7, 24], [7, 9, 139], [7, 11, 7], [7, 12, 1], [7, 13, 37], [7, 14, 4], [7, 17
46, 1], [9, 47, 4], [9, 51, 7], [9, 52, 4], [9, 56, 4], [9, 59, 4], [9, 60, 3], [9
, [11, 47, 236], [11, 49, 24], [11, 50, 26], [11, 51, 44], [11, 52, 30], [11, 54,
, 59, 71], [12, 60, 119], [12, 61, 16], [12, 62, 92], [12, 63, 272], [13, 0, 19],
27, 2], [14, 28, 3], [14, 29, 6], [14, 30, 8], [14, 31, 3], [14, 32, 10], [14, 33,
9], [17, 4, 17], [17, 5, 17], [17, 6, 20], [17, 7, 14], [17, 8, 3], [17, 9, 9], [1
[18, 21, 34], [18, 22, 40], [18, 23, 105], [18, 25, 152], [18, 26, 2], [18, 27, 25
[20, 16, 16], [20, 17, 5], [20, 18, 1064], [20, 19, 12], [20, 20, 30], [20, 21, 33
33, 43], [21, 34, 52], [21, 35, 59], [21, 36, 48], [21, 37, 44], [21, 38, 45], [21
, [23, 6, 16], [23, 7, 2], [23, 9, 2], [23, 10, 12], [23, 11, 27], [23, 12, 1], [2
, [25, 30, 5], [25, 31, 23], [25, 36, 2], [25, 37, 16], [25, 38, 39], [25, 39, 4],
7, 16, 25], [27, 17, 7], [27, 18, 60], [27, 19, 40], [27, 20, 74], [27, 21, 41], [
29, 23, 15], [29, 24, 4], [29, 25, 519], [29, 26, 1], [29, 27, 79], [29, 28, 1318]
31, 43, 16], [31, 44, 36], [31, 45, 91], [31, 46, 11], [31, 47, 2], [31, 56, 5], [
76], [35, 12, 42], [35, 13, 2], [35, 14, 33], [35, 15, 78], [36, 0, 7], [36, 3, 1]
38, 28, 295], [38, 29, 45], [38, 30, 135], [38, 31, 566], [38, 32, 3], [38, 36, 3]
], [41, 17, 2], [41, 20, 5], [41, 21, 4], [41, 22, 1], [41, 23, 9], [41, 28, 1], [
], [43, 38, 8], [43, 40, 2], [43, 42, 7], [43, 44, 3], [43, 46, 3], [43, 56, 2], [
7, 11, 14], [47, 12, 1], [47, 13, 2], [47, 14, 1], [47, 15, 1], [47, 20, 2], [47,
500], [50, 25, 21], [50, 26, 1], [50, 27, 1], [50, 28, 7], [50, 30, 6], [50, 31, 2
84], [52, 29, 3], [52, 30, 2], [52, 31, 21], [52, 32, 1], [52, 33, 6], [52, 34, 3]
, [54, 52, 1], [54, 55, 1], [54, 58, 3], [54, 60, 2], [55, 3, 8], [55, 4, 7], [55,
2, 2], [57, 6, 2], [57, 7, 2], [57, 9, 1], [57, 10, 16], [57, 11, 92], [57, 13, 10
```

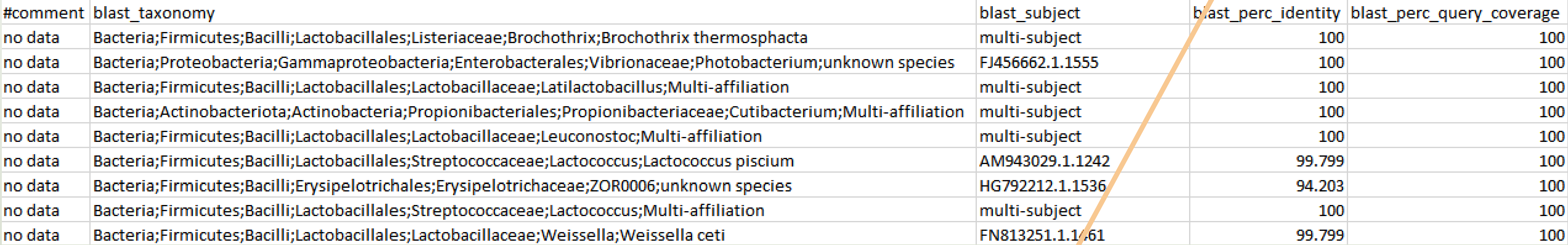
The biom file is not a human readable format. It is only very useful for bioinformaticians.
To read the abundance table you have to transform the BIOM file in TSV file thanks to **BIOM_to_TSV tool**.

Answer 5

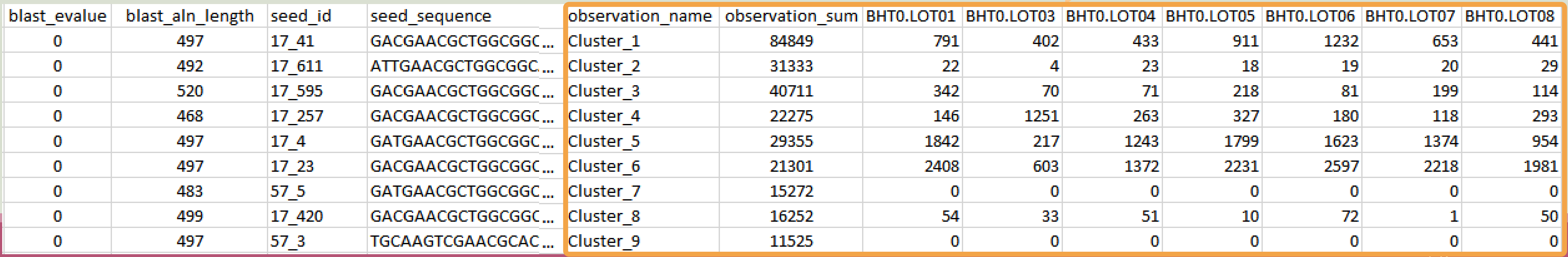
Q5: what do you think about the TSV file format? What does it contain?

The TSV format: tabular separated Value.
Universal format, ideal for different spreadsheets.

This file contain the abundance table and information about affiliation of OTUs.



#comment	blast_taxonomy	blast_subject	blast_perc_identity	blast_perc_query_coverage
no data	Bacteria;Firmicutes;Bacilli;Lactobacillales;Listeriaceae;Brochothrix;Brochothrix thermosphacta	multi-subject	100	100
no data	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales;Vibrionaceae;Photobacterium;unknown species	FJ456662.1.1555	100	100
no data	Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Latilactobacillus;Multi-affiliation	multi-subject	100	100
no data	Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Multi-affiliation	multi-subject	100	100
no data	Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Leuconostoc;Multi-affiliation	multi-subject	100	100
no data	Bacteria;Firmicutes;Bacilli;Lactobacillales;Streptococcaceae;Lactococcus;Lactococcus piscium	AM943029.1.1242	99.799	100
no data	Bacteria;Firmicutes;Bacilli;Erysipelotrichales;Erysipelotrichaceae;ZOR0006;unknown species	HG792212.1.1536	94.203	100
no data	Bacteria;Firmicutes;Bacilli;Lactobacillales;Streptococcaceae;Lactococcus;Multi-affiliation	multi-subject	100	100
no data	Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Weissella;Weissella ceti	FN813251.1.1461	99.799	100



blast_evalue	blast_aln_length	seed_id	seed_sequence	observation_name	observation_sum	BHT0.LOT01	BHT0.LOT03	BHT0.LOT04	BHT0.LOT05	BHT0.LOT06	BHT0.LOT07	BHT0.LOT08
0	497	17_41	GACGAACGCTGGCGGC...	Cluster_1	84849	791	402	433	911	1232	653	441
0	492	17_611	ATTGAACGCTGGCGGC...	Cluster_2	31333	22	4	23	18	19	20	29
0	520	17_595	GACGAACGCTGGCGGC...	Cluster_3	40711	342	70	71	218	81	199	114
0	468	17_257	GACGAACGCTGGCGGC...	Cluster_4	22275	146	1251	263	327	180	118	293
0	497	17_4	GATGAACGCTGGCGGC...	Cluster_5	29355	1842	217	1243	1799	1623	1374	954
0	497	17_23	GACGAACGCTGGCGGC...	Cluster_6	21301	2408	603	1372	2231	2597	2218	1981
0	483	57_5	GATGAACGCTGGCGGC...	Cluster_7	15272	0	0	0	0	0	0	0
0	499	17_420	GACGAACGCTGGCGGC...	Cluster_8	16252	54	33	51	10	72	1	50
0	497	57_3	TGCAAGTCGAACGCAC...	Cluster_9	11525	0	0	0	0	0	0	0

Answer 6

a. how would you qualify the alignment between the OTU3 seed and the sequences of the silva database?

Alignment is perfect ! 100% indentity and 100% coverage between OTU3 seed and the 520 nucleotides of sequence from silva database

b. What does it mean e-value = 0 ?

The expect value is a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size. The lower the e-value, or the closer it is to zero, the more "significant" the match is.

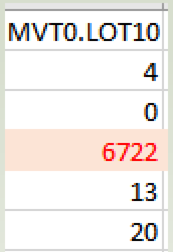
c. What is the header of column that shows the sequence of OTU seed ?

Seed_sequence

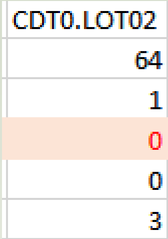
d. How many sequences have OTU3 in total ?

40711 found in column “ observation_sum”

e. How many sequences have OTU3 in MVT0.LOT10 ? What is the sample where OTU3 is absent ?



MVT0.LOT10
4
0
6722
13
20



CDT0.LOT02
64
1
0
0
3

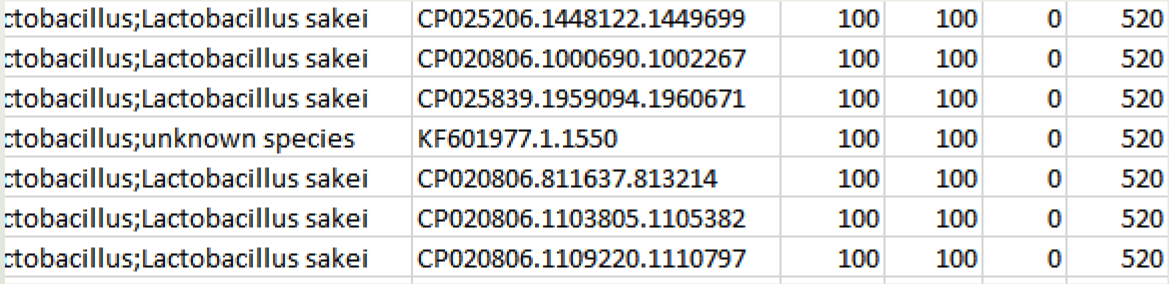
We can remark that OTU3 is particularly present in MV samples and rare in CD samples

Answer 7

a. Why cluster3 has a multiaffiliation for species ?

In multi-affiliations.tsv file, for cluster_3, we observe that 75 affiliations are possible for this OTU at species rank.

All strictly equivalent 100% identity and 100% coverage with 75 different sequences of silva database.



ctobacillus;Lactobacillus sakei	CP025206.1448122.1449699	100	100	0	520
ctobacillus;Lactobacillus sakei	CP020806.1000690.1002267	100	100	0	520
ctobacillus;Lactobacillus sakei	CP025839.1959094.1960671	100	100	0	520
ctobacillus;unknown species	KF601977.1.1550	100	100	0	520
ctobacillus;Lactobacillus sakei	CP020806.811637.813214	100	100	0	520
ctobacillus;Lactobacillus sakei	CP020806.1103805.1105382	100	100	0	520
ctobacillus;Lactobacillus sakei	CP020806.1109220.1110797	100	100	0	520

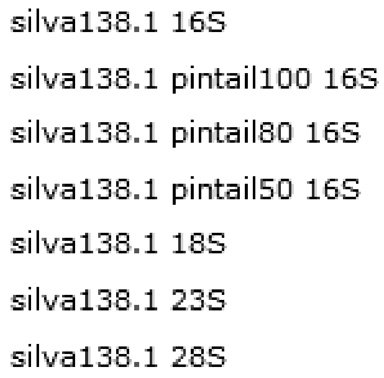
b. Why “Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Latilactobacillus;Lactobacillus sakei” is present 74 times ?

Because these are 74 different strains of *L. sakei*. They have blast ID different.

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

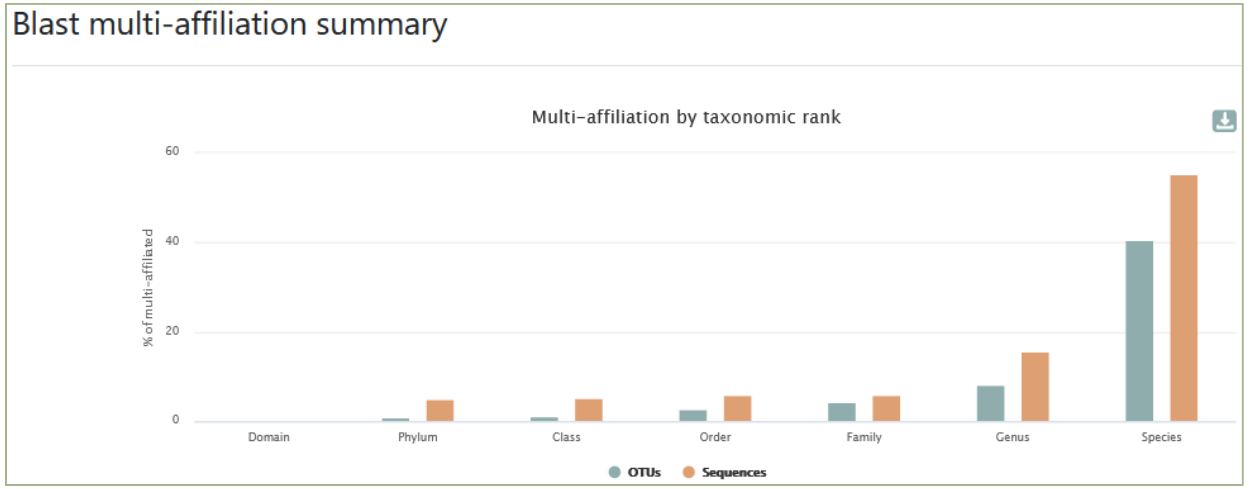
silva138.1 23S

silva138.1 28S

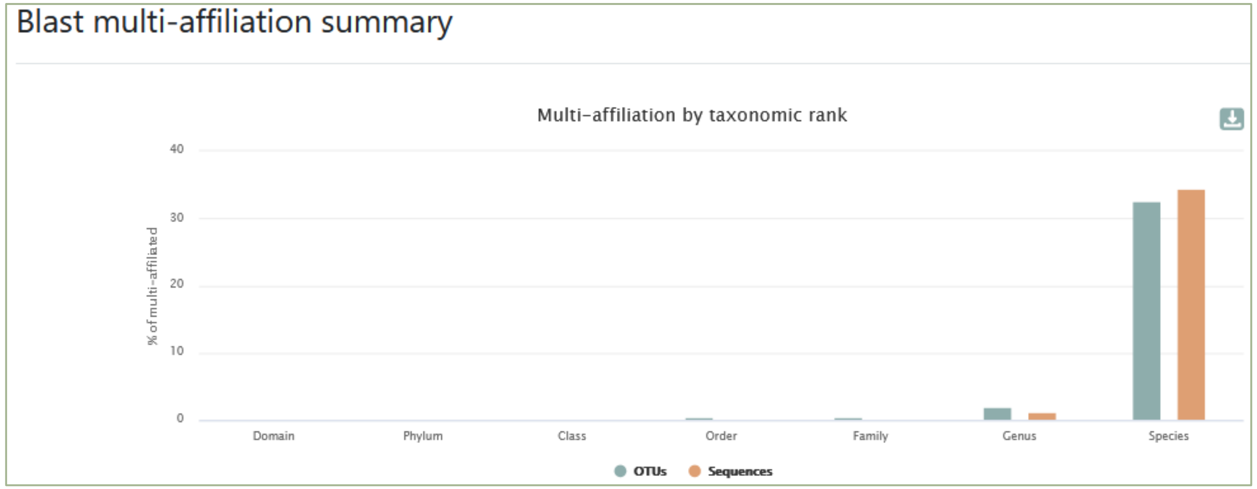
Only for 16S !

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

Silva pintail or not pintail ?



Blast multi-affiliation summary



Blast multi-affiliation summary

Exemple between silva 138.1 and silva 138.1 pintail 100

130 identical blast best hits on **SILVA 138.1 pintail 100** databank

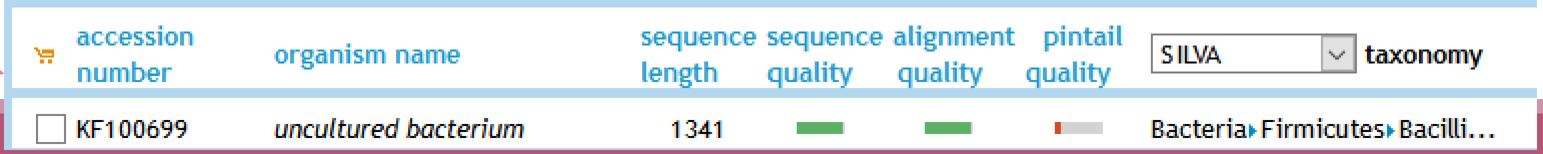
Cluster_4	Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes
Cluster_4	Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes 6609
Cluster_4	Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium 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 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	CP013922.15879[illegible]				
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

2 choices for cluster 64

How choose the good affiliation ?

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

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

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

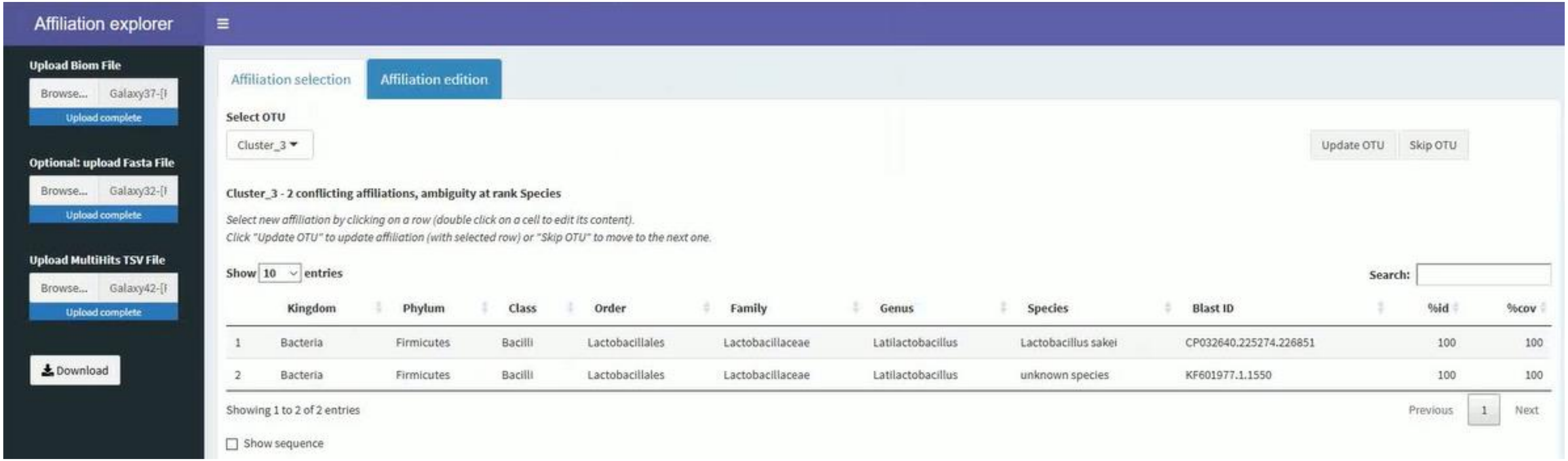
and

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

Maybe, for this cluster, S. xylosus is better

Affiliation explorer

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



Cluster_3 - 2 conflicting affiliations, ambiguity at rank Species

Select new affiliation by clicking on a row (double click on a cell to edit its content).
Click "Update OTU" to update affiliation (with selected row) or "Skip OTU" to move to the next one.

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

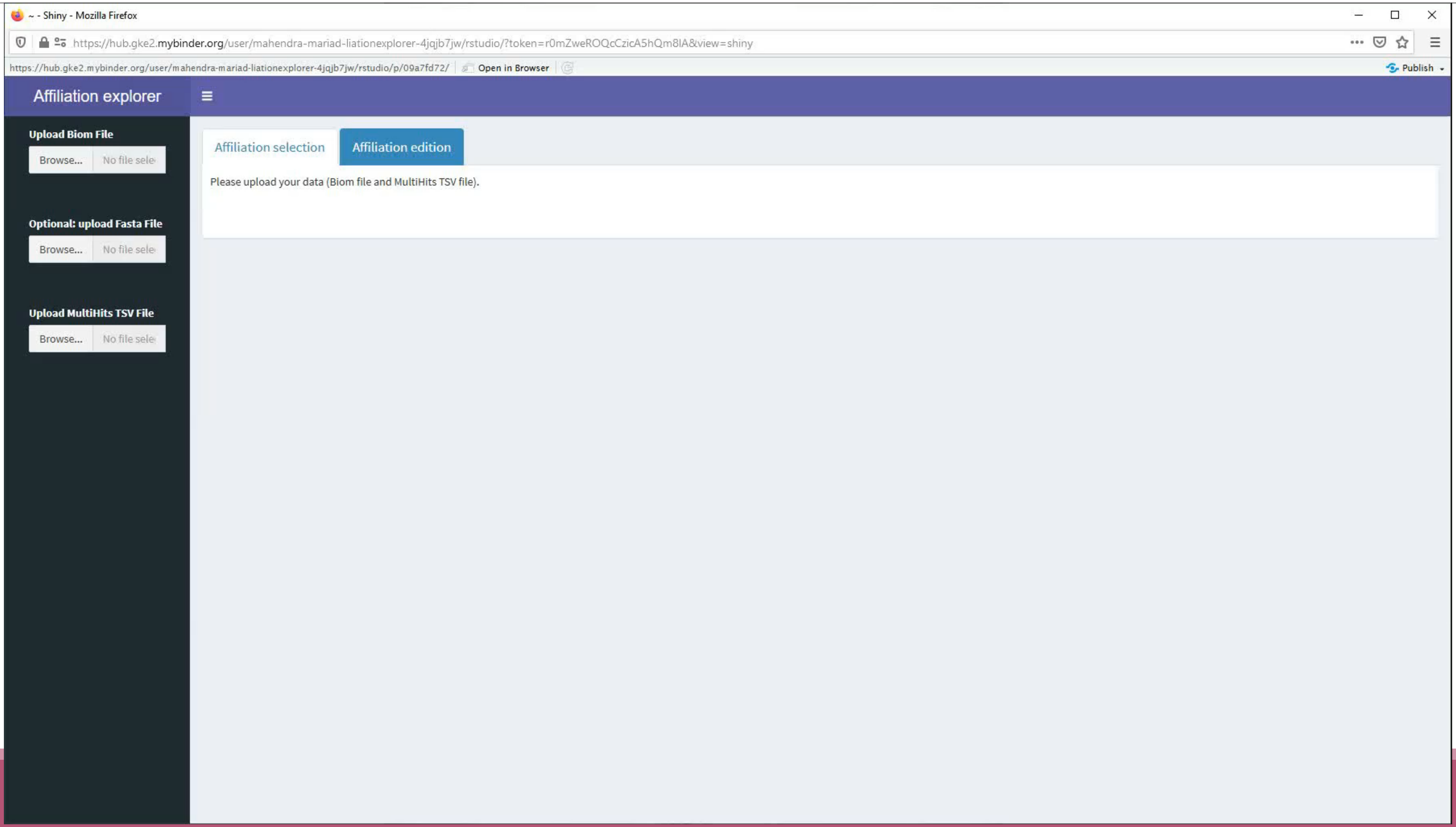
Showing 1 to 2 of 2 entries

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



~ - Shiny - Mozilla Firefox

https://hub.gke2.mybinder.org/user/mahendra-mariad-liationexplorer-4jqjb7jw/rstudio/?token=r0mZweROQcCzicA5hQm8lA&view=shiny

https://hub.gke2.mybinder.org/user/mahendra-mariad-liationexplorer-4jqjb7jw/rstudio/p/09a7fd72/

Please upload your data (Biom file and MultiHits TSV file).

Divergence on the composition of microbial communities at the different taxonomic ranks

Affiliations and abundances of FROGS OTUs are they reliable ?

With the first versions of FROGS where multi-affiliation did not yet exist.

Taxonomic ranks	**Average divergence of the affiliations of the 10 samples (%)** **500setA**	**Average divergence of the affiliations of the 10 samples (%)** **100setA**
Kingdom	0.00	0.00
Phylum	0.46	0.41
Class	0.64	0.50
Order	0.94	0.68
Familly	1.18	0.78
Genus	1.76	1.30
Species	23.87	34.80

Affiliation was chosen with arbitrary criterion among all strictly equivalent affiliation

solution

Report on abundance table, the multiple identical affiliations

Only one best hit

Taxonomic ranks	Average divergence of the affiliations of the 10 samples (%) 500setA	Average divergence of the affiliations of the 10 samples (%) 100setA
Kingdom	0.00	0.00
Phylum	0.46	0.41
Class	0.64	0.50
Order	0.94	0.68
Familly	1.18	0.78
Genus	1.76	1.30
Species	23.87	34.80

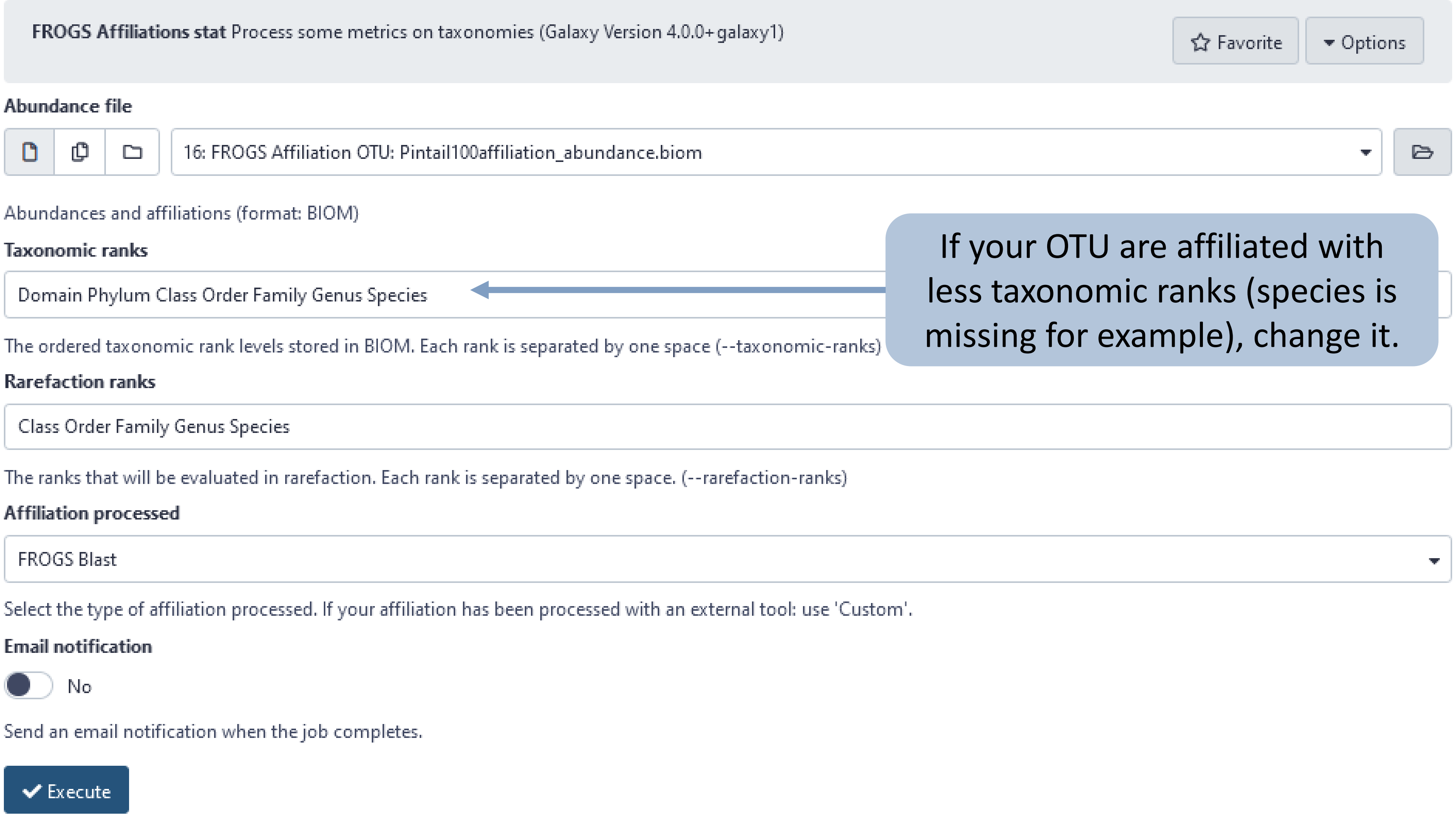
Multiple best hit

Taxonomic ranks	Median divergence of the affiliations of the 10 samples (%) 500setA	Median divergence of the affiliations of the 10 samples (%) 100setA
Kingdom	0.00	0.00
Phylum	0.46	0.41
Class	0.64	0.50
Order	0.93	0.68
Familly	1.17	0.78
Genus	1.60	1.00
Species	6.63	5.75

With the FROGS guideline
OTU filter on abundance < 0.005%

Taxonomic ranks	Median divergence of the affiliations of the 10 samples (%) 500setA filter: 0.005% - 505 OTUs	Median divergence of the affiliations of the 10 samples (%) 100setA filter: 0.005% - 100 OTUs
Kingdom	0.00	0.00
Phylum	0.38	0.38
Class	0.57	0.48
Order	0.81	0.64
Familly	1.08	0.74
Genus	1.43	0.76
Species	1.53	0.78

Affiliation Stat



FROGS Affiliations stat Process some metrics on taxonomies (Galaxy Version 4.0.0+galaxy1)

Abundance file

16: FROGS Affiliation OTU: Pintail100affiliation_abundance.biom

Abundances and affiliations (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)

If your OTU are affiliated with less taxonomic ranks (species is missing for example), change it.

Rarefaction ranks

Class Order Family Genus Species

The ranks that will be evaluated in rarefaction. Each rank is separated by one space. (--rarefaction-ranks)

Affiliation processed

FROGS Blast

Select the type of affiliation processed. If your affiliation has been processed with an external tool: use 'Custom'.

Email notification

○ No

Send an email notification when the job completes.

Practice:

LAUNCH THE FROGS AFFILIATION STAT TOOL

Exercice:

Go to history « 16S » history

Launch the « FROGS Affiliation Stat » tool on last affiliation_abundance.biom

→ objectives :

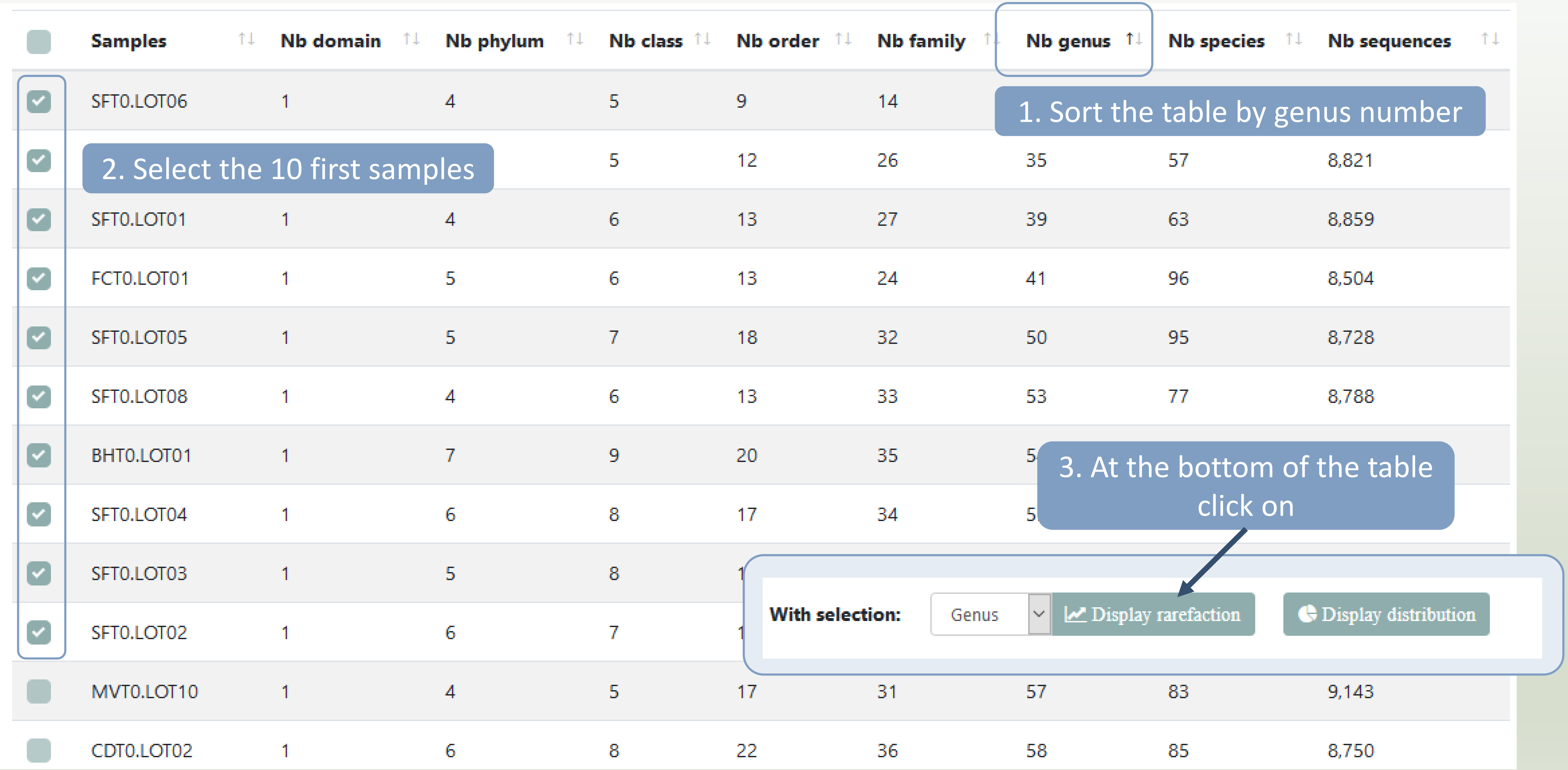
understand rarefaction curves and the diversity diagram

Exercice:

1. Build the **rarefaction** curve on genus rank with the 10 samples that contain the least number of different genus.
2. SFT0.LOT06 and MVT0.LOT10 have they been sequenced deeply enough?
3. Build the **distribution** on FC samples *i.e.* “Filet de Cabillaud”
4. How many sequences are some *Brochothrix thermosphacta* ?
5. On the total of sequences, what is the proportion affiliated to the Firmicutes?
6. Among Firmicutes, how many are Bacilli ?
7. But what is the proportion of Firmicutes in the total of sequence of all sample ?
8. How many OTUs are align perfectly with a database sequence ?

Answer 1

Q1: Build the **rarefaction** curve on genus rank with the 10 samples that contain the least number of different genus.



1. Sort the table by genus number

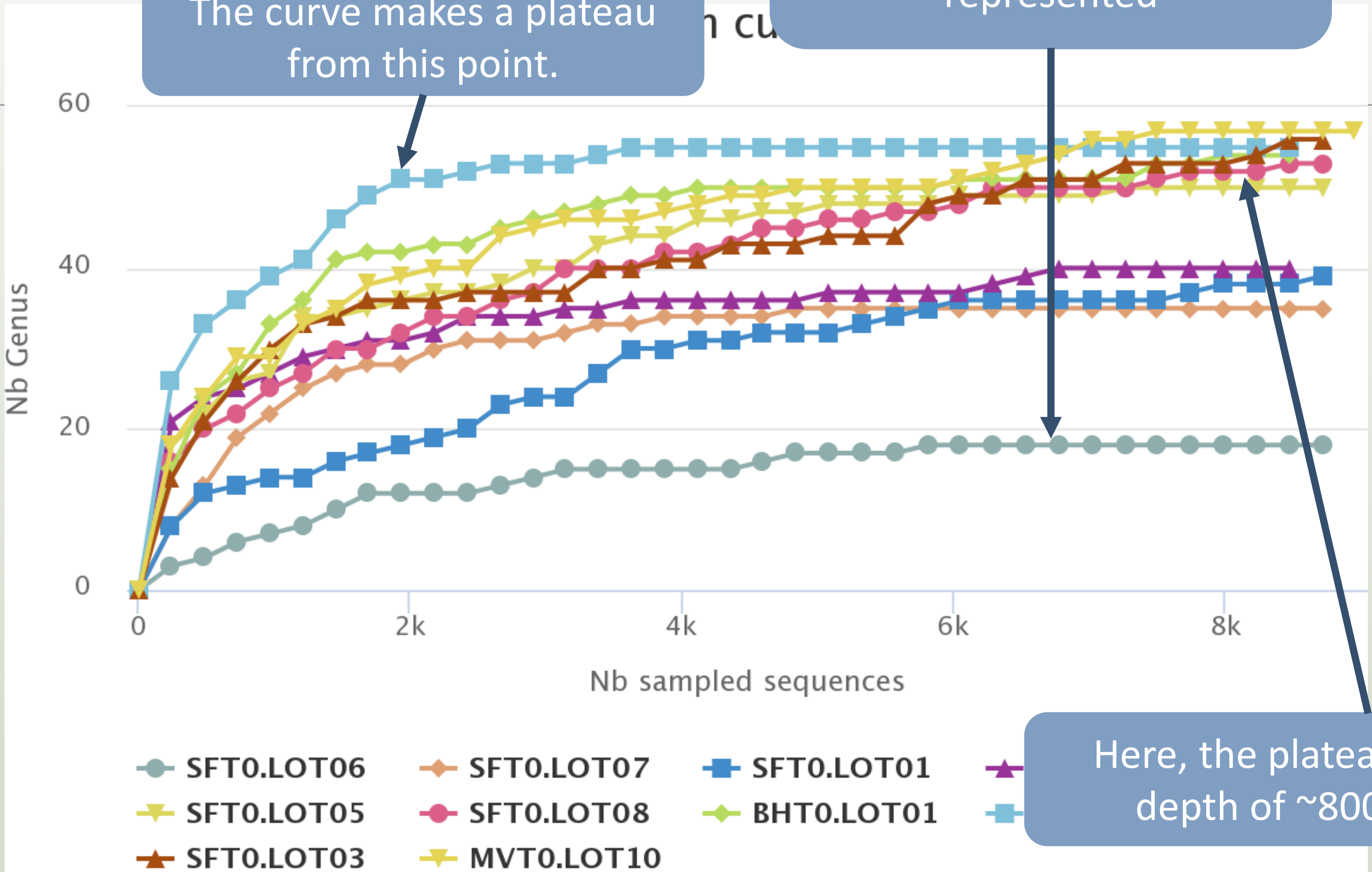
2. Select the 10 first samples

3. At the bottom of the table click on

☐	Samples ↑↓	Nb domain ↑↓	Nb phylum ↑↓	Nb class ↑↓	Nb order ↑↓	Nb family ↑↓	Nb genus ↑↓	Nb species ↑↓	Nb sequences ↑↓
☑	SFT0.LOT06	1	4	5	9	14			
☑				5	12	26	35	57	8,821
☑	SFT0.LOT01	1	4	6	13	27	39	63	8,859
☑	FCT0.LOT01	1	5	6	13	24	41	96	8,504
☑	SFT0.LOT05	1	5	7	18	32	50	95	8,728
☑	SFT0.LOT08	1	4	6	13	33	53	77	8,788
☑	BHT0.LOT01	1	7	9	20	35	5		
☑	SFT0.LOT04	1	6	8	17	34	5		
☑	SFT0.LOT03	1	5	8	1				
☑	SFT0.LOT02	1	6	7	1				
☐	MVT0.LOT10	1	4	5	17	31	57	83	9,143
☐	CDT0.LOT02	1	6	8	22	36	58	85	8,750

Answer 2

Q2: SFT0.LOT06 and MVT0.LOT10 have they been sequenced deeply enough?



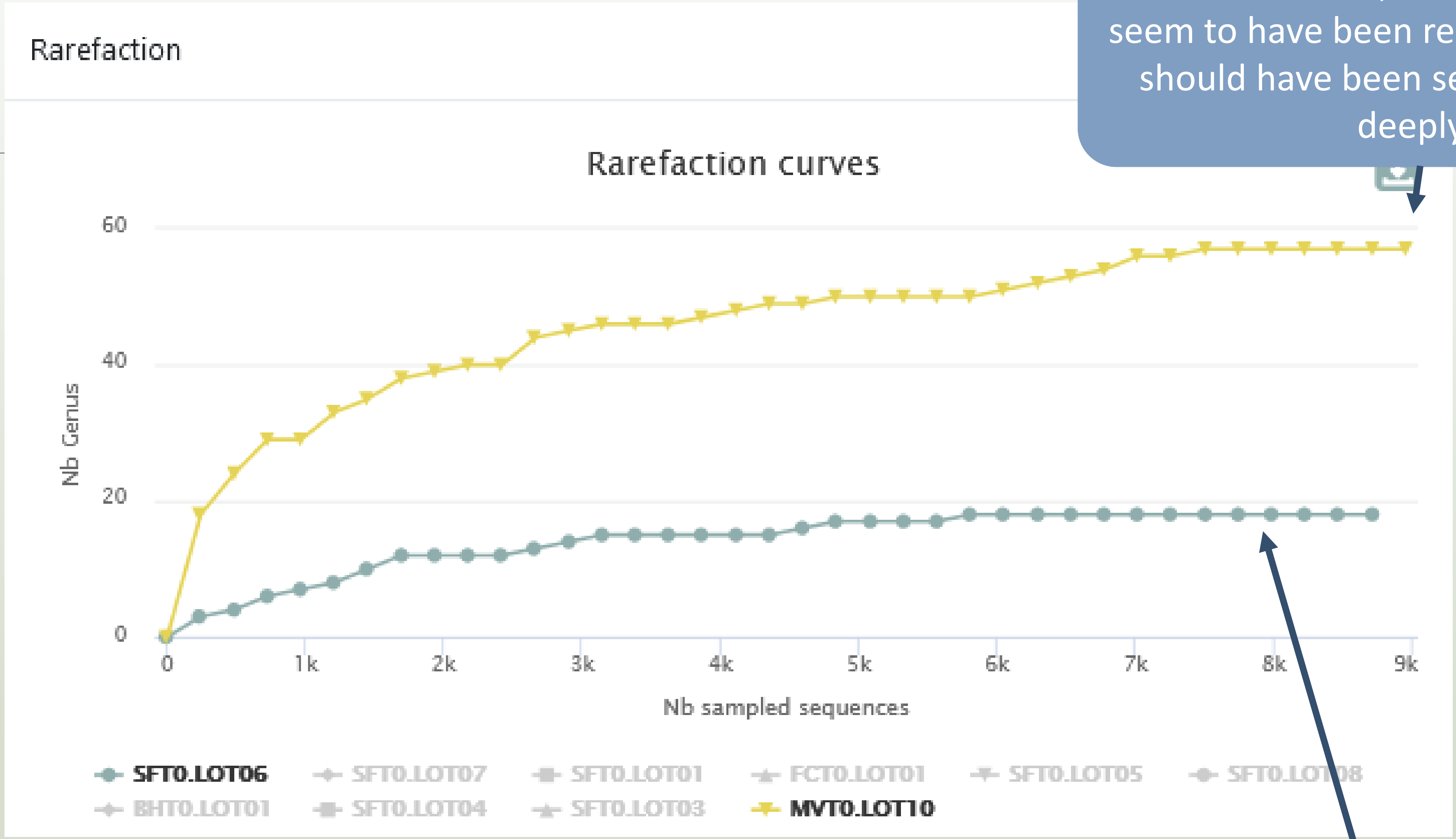
The curve makes a plateau from this point.

For this sample, with ~6500 sequences, all genus are represented

Here, the plateau comes with a depth of ~8000 sequences

Answer 2

Q2: SFT0.LOT06 and MVT0.LOT10 have they been sequenced deeply enough?

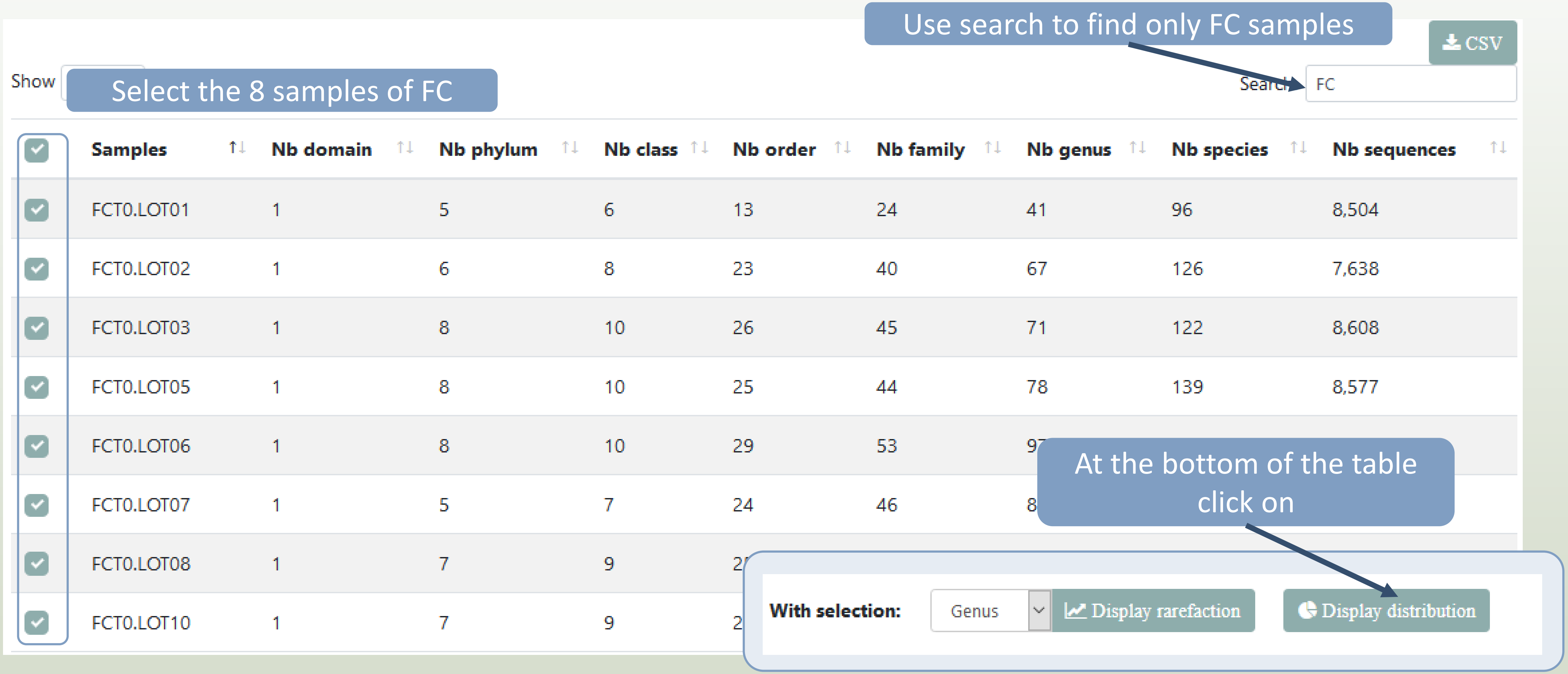


For MVTO.LOT10, the plateau does no seem to have been reached. Perhaps should have been sequenced more deeply?

With ~8000 sequences, all genus for this species are represented

Answer 3

Q3: Build the **distribution** on FC samples *i.e.* “Filet de Cabillaud”



Use search to find only FC samples

Select the 8 samples of FC

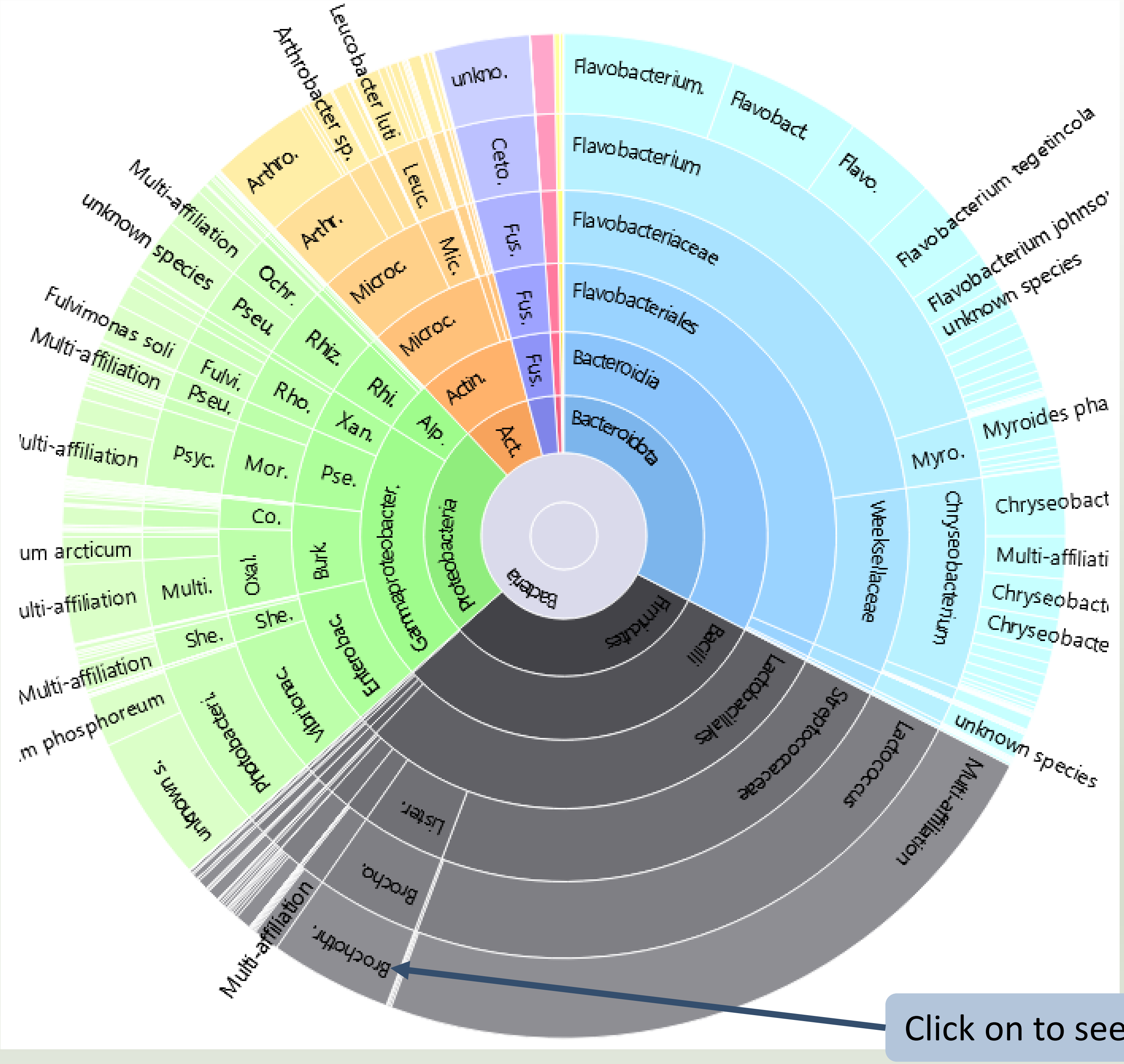
At the bottom of the table click on

☑	Samples	Nb domain	Nb phylum	Nb class	Nb order	Nb family	Nb genus	Nb species	Nb sequences
☑	FCT0.LOT01	1	5	6	13	24	41	96	8,504
☑	FCT0.LOT02	1	6	8	23	40	67	126	7,638
☑	FCT0.LOT03	1	8	10	26	45	71	122	8,608
☑	FCT0.LOT05	1	8	10	25	44	78	139	8,577
☑	FCT0.LOT06	1	8	10	29	53	97		
☑	FCT0.LOT07	1	5	7	24	46	8		
☑	FCT0.LOT08	1	7	9	2				
☑	FCT0.LOT10	1	7	9	2				

With selection: Genus

Answer 3 4 & 5

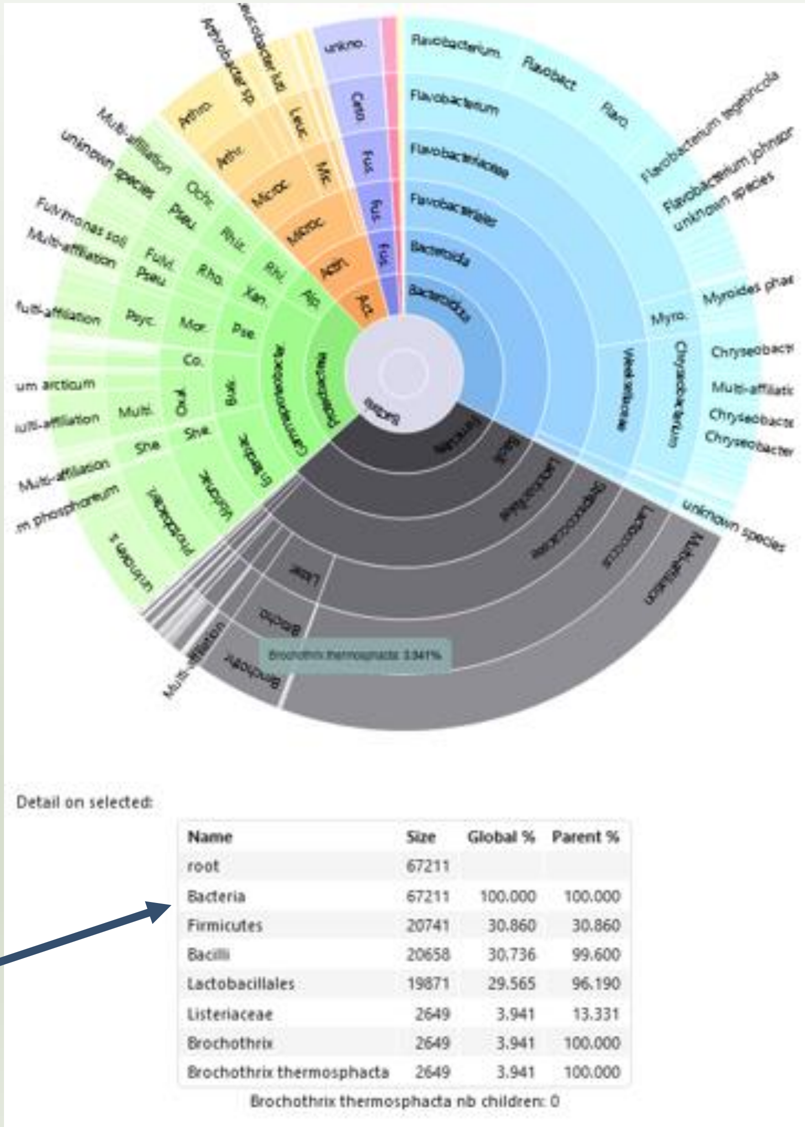
Q3: Build the **distribution** on FC samples *i.e.* “Filet de Cabillaud”



Click on to see *Brochothrix thermosphacta*

Answer 3, 4, 5 & 6

Q4: How many sequences are some *Brochothrix thermosphacta* ?
Q5: On the total of sequences, what is the proportion affiliated to the Firmicutes?
Q6: Among Firmicutes, how many are Bacilli ?

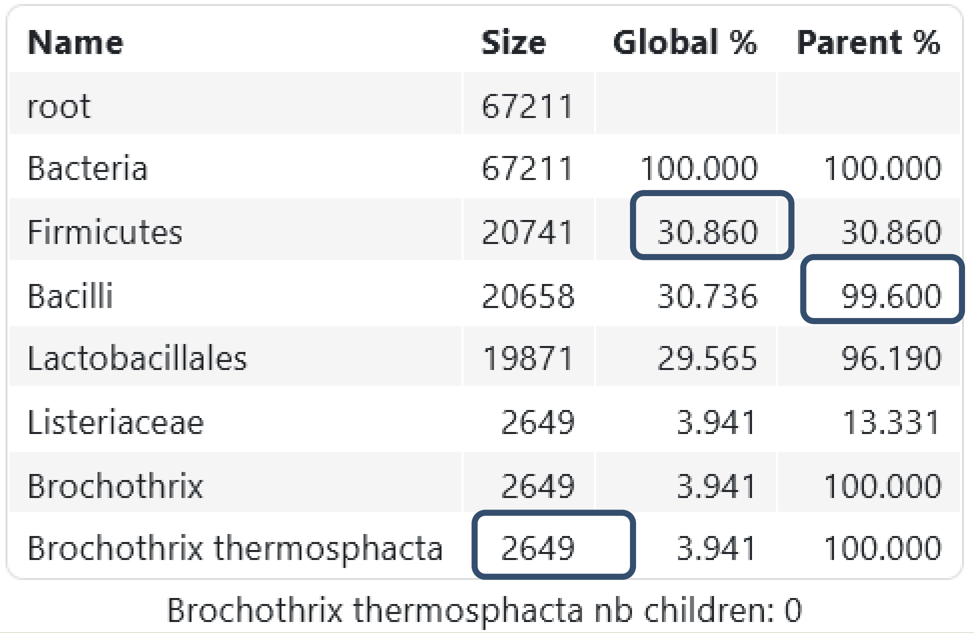


Detail on selected:

Name	Size	Global %	Parent %
root	67211		
Bacteria	67211	100.000	100.000
Firmicutes	20741	30.860	30.860
Bacilli	20658	30.736	99.600
Lactobacillales	19871	29.565	96.190
Listeriaceae	2649	3.941	13.331
Brochothrix	2649	3.941	100.000
Brochothrix thermosphacta	2649	3.941	100.000

Brochothrix thermosphacta nb children: 0

A table appears



Name	Size	Global %	Parent %
root	67211		
Bacteria	67211	100.000	100.000
Firmicutes	20741	30.860	30.860
Bacilli	20658	30.736	99.600
Lactobacillales	19871	29.565	96.190
Listeriaceae	2649	3.941	13.331
Brochothrix	2649	3.941	100.000
Brochothrix thermosphacta	2649	3.941	100.000

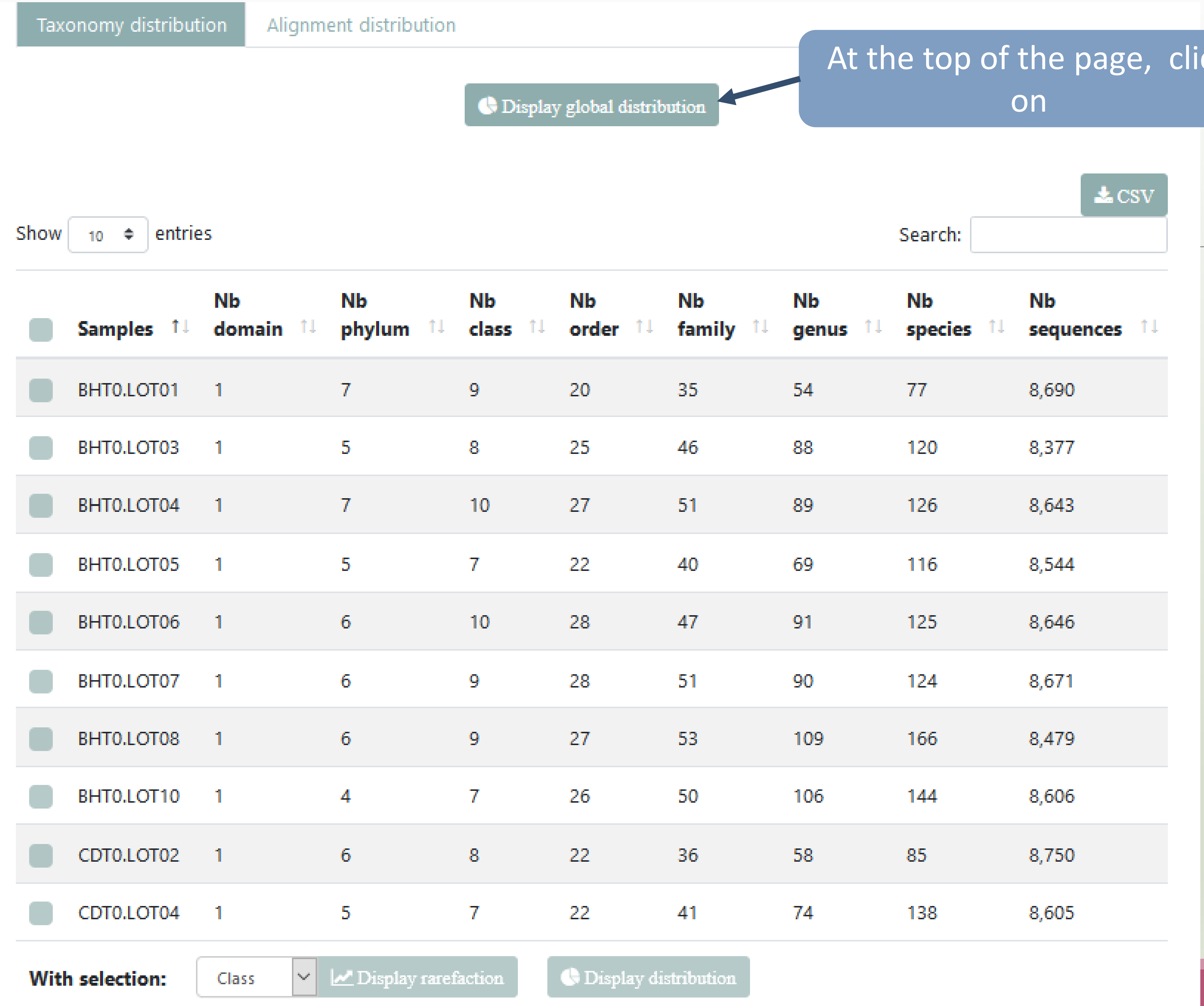
Brochothrix thermosphacta nb children: 0

- 2649 sequences are some *Brochothrix thermosphacta*
- Firmicutes represent ~30% of total of sequences of these samples
- 99.6% of Firmicutes are Bacilli

Answer 7

Q7: But what is the proportion of Firmicutes in the total of sequence of all sample ?

At the top of the page, click on



Show 10 entries

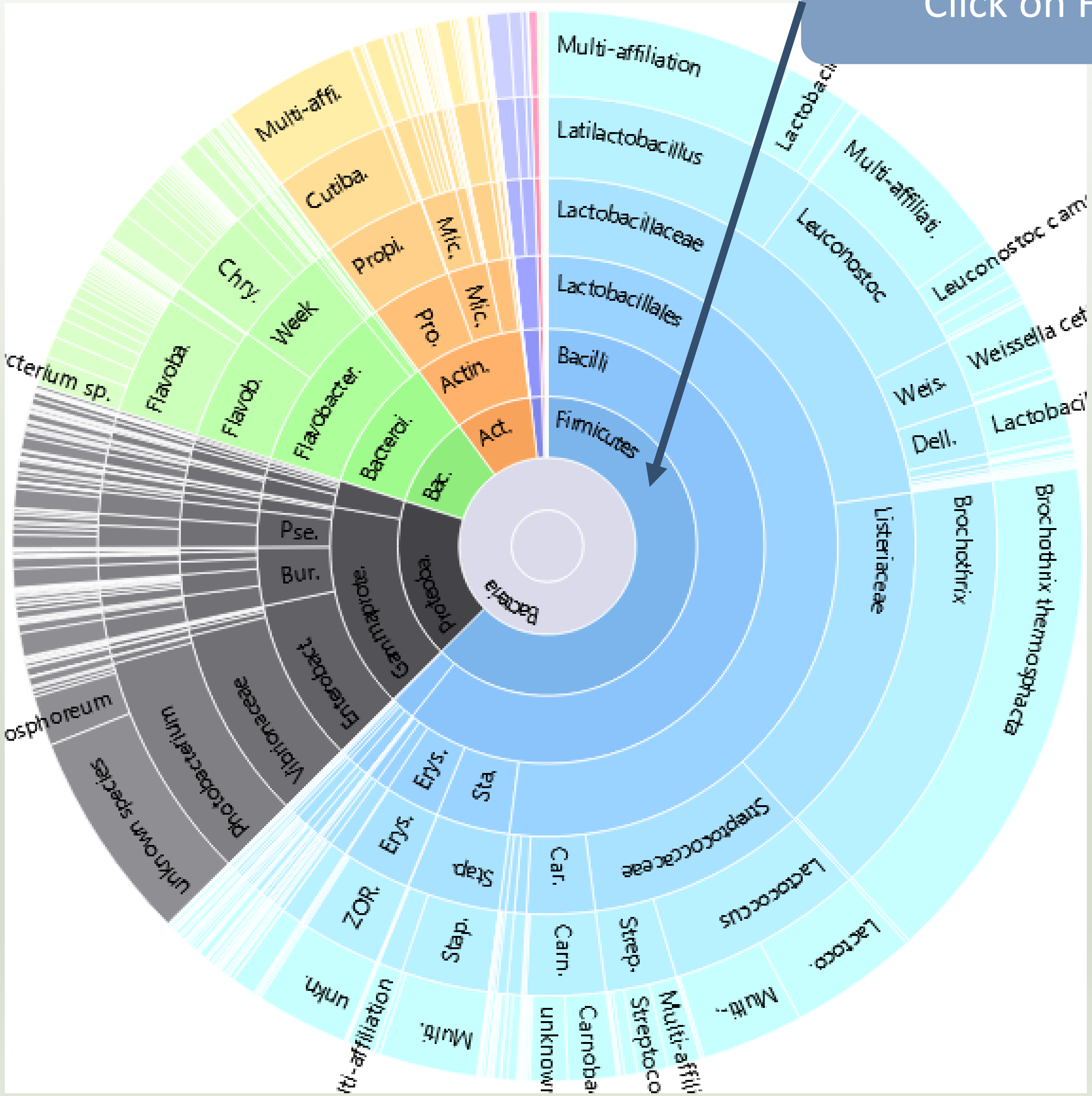
Search:

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

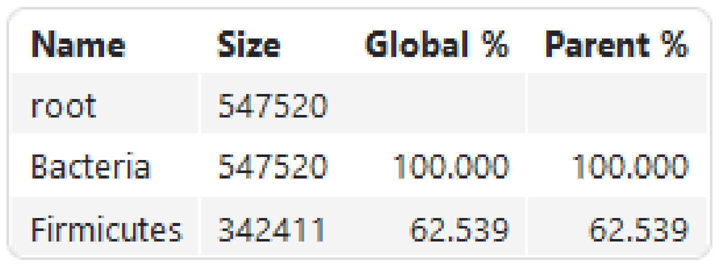
With selection: Class

Answer 7

Q7: But what is the proportion of Firmicutes in the total of sequence of all sample ?



Click on Firmicutes

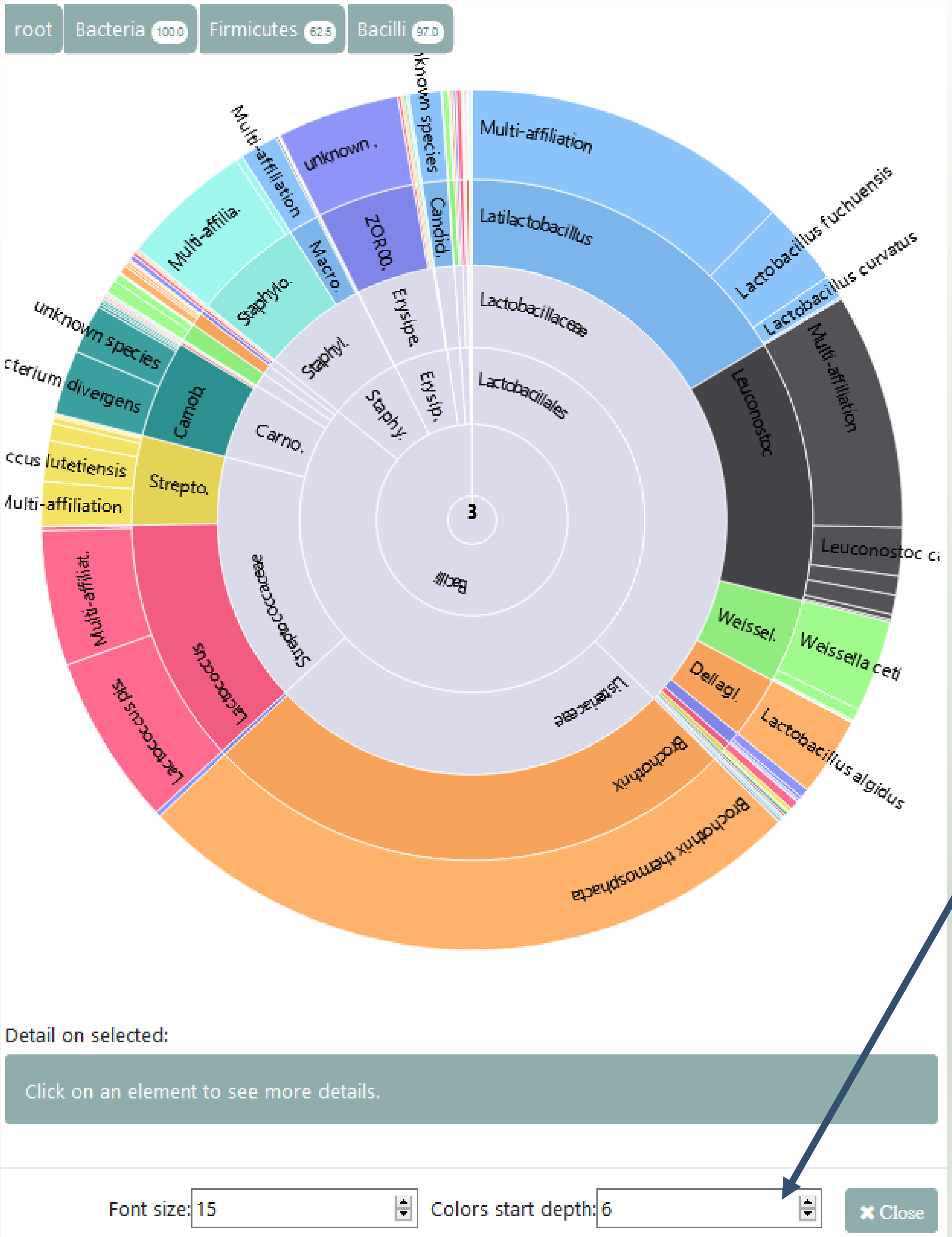


Name	Size	Global %	Parent %
root	547520		
Bacteria	547520	100.000	100.000
Firmicutes	342411	62.539	62.539

Firmicutes represent 62% of Bacteria

Answer 7

Q7: But what is the proportion of Firmicutes in the total of sequence of all sample ?



Detail on selected:

Click on an element to see more details.

Font size: 15

Colors start depth: 6

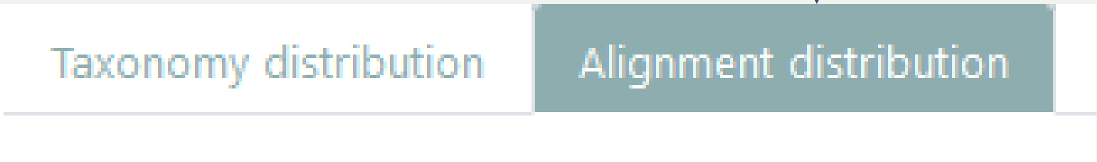
✖ Close

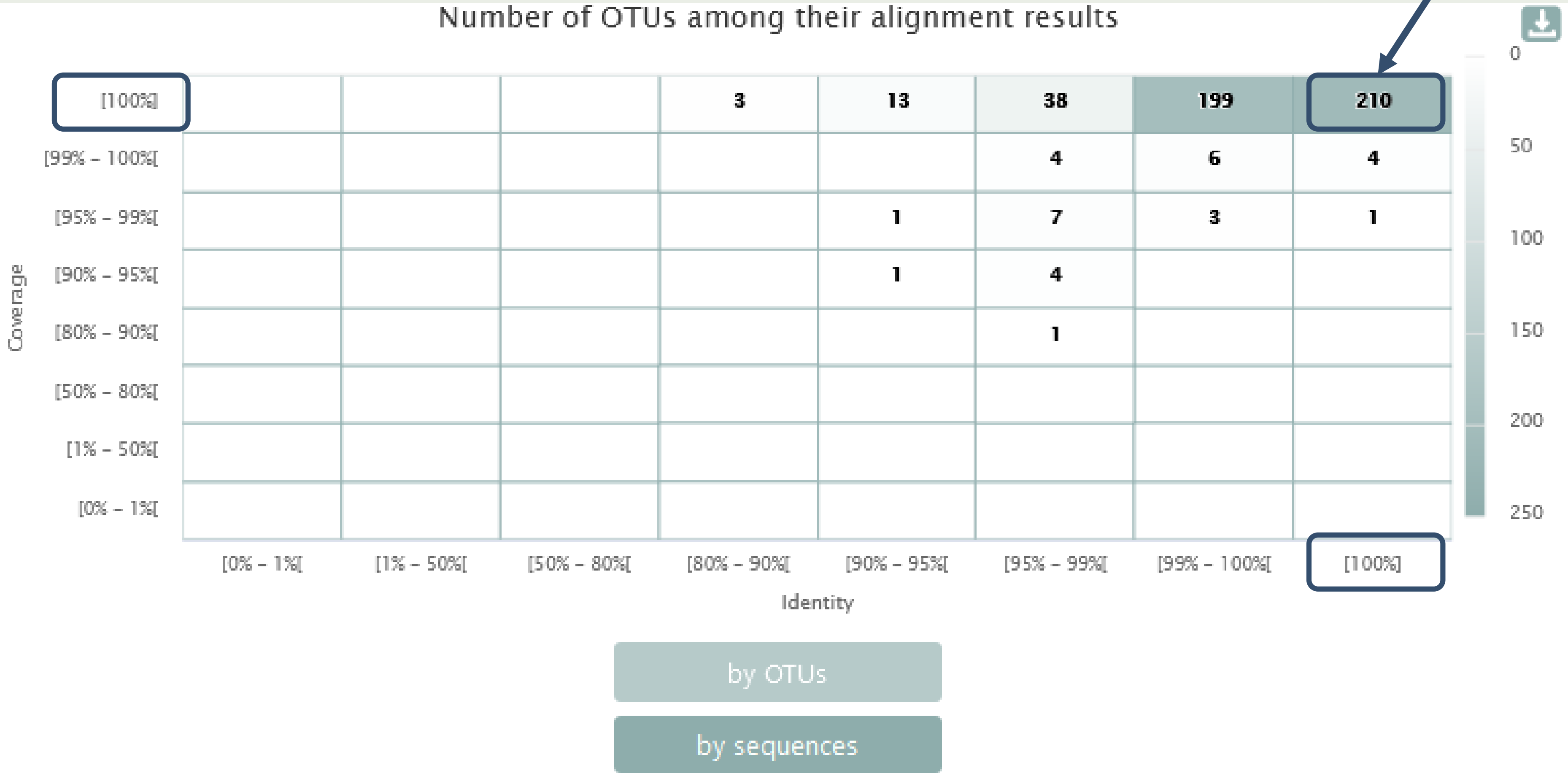
To focus on Firmicutes, **double click on**. After you can apply color among rank depth.

Answer 8

Q8: How many OTUs are align perfectly with a database sequence ?

At the top of the page, click on this tab



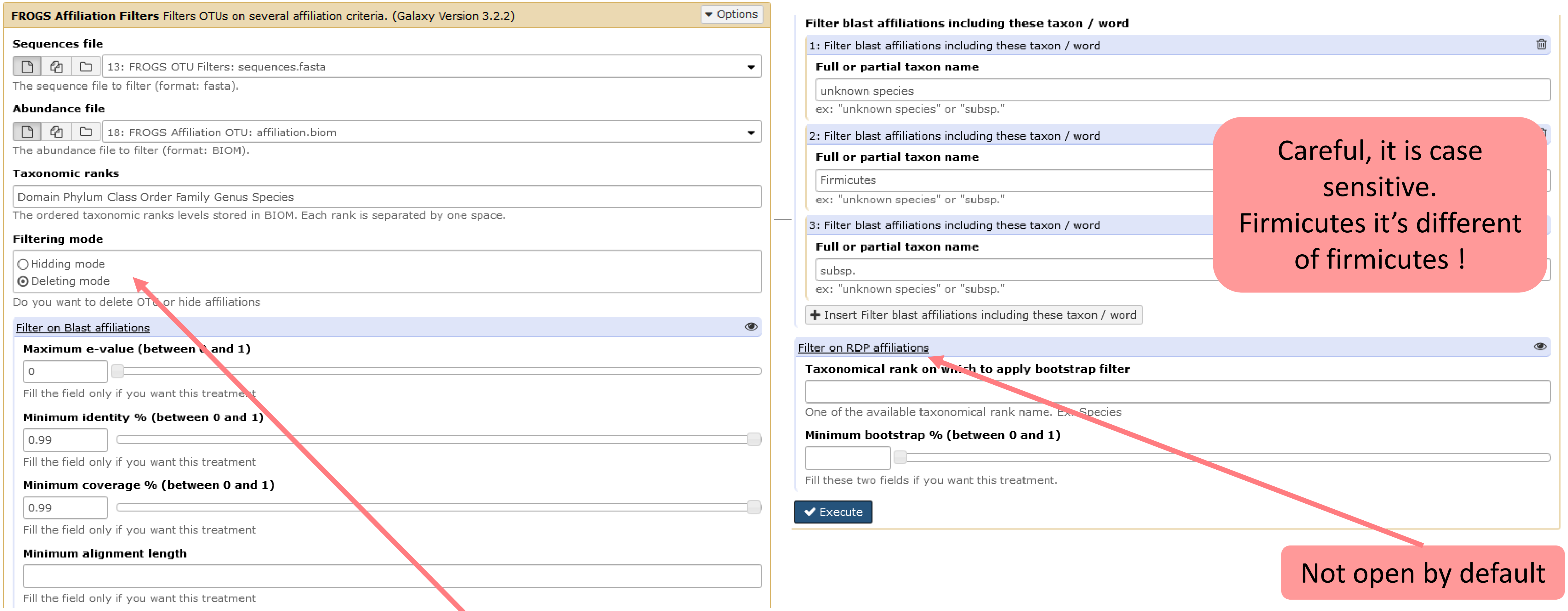


Number of OTUs among their alignment results

Coverage \ Identity	[0% – 1%[	[1% – 50%[	[50% – 80%[	[80% – 90%[	[90% – 95%[	[95% – 99%[	[99% – 100%[	[100%]
[100%]				3	13	38	199	210
[99% – 100%[						4	6	4
[95% – 99%[					1	7	3	1
[90% – 95%[					1	4		
[80% – 90%[						1		
[50% – 80%[								
[1% – 50%[								
[0% – 1%[								

210 sequences are aligned with 100% identity and 100% coverage with a sequence of silva.

Filters on affiliations



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

Sequences file

13: FROGS OTU Filters: sequences.fasta

The sequence file to filter (format: fasta).

Abundance file

18: FROGS Affiliation OTU: affiliation.biom

The abundance file to filter (format: BIOM).

Taxonomic ranks

Domain Phylum Class Order Family Genus Species

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

Filtering mode

○ Hidding mode
● Deleting mode

Do you want to delete OTU or hide affiliations

Filter on Blast affiliations

Maximum e-value (between 0 and 1)

0

Fill the field only if you want this treatment

Minimum identity % (between 0 and 1)

0.99

Fill the field only if you want this treatment

Minimum coverage % (between 0 and 1)

0.99

Fill the field only if you want this treatment

Minimum alignment length

Fill the field only if you want this treatment

Filter blast affiliations including these taxon / word

1: Filter blast affiliations including these taxon / word

Full or partial taxon name

unknown species

ex: "unknown species" or "subsp."

2: Filter blast affiliations including these taxon / word

Full or partial taxon name

Firmicutes

ex: "unknown species" or "subsp."

3: Filter blast affiliations including these taxon / word

Full or partial taxon name

subsp.

ex: "unknown species" or "subsp."

Options
+ Insert Filter blast affiliations including these taxon / word
Execute

Filter on RDP affiliations

Taxonomical rank on which to apply bootstrap filter

One of the available taxonomical rank name. Ex: Species

Minimum bootstrap % (between 0 and 1)

Fill these two fields if you want this treatment.

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

Not open by default

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

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

Practice:

LAUNCH THE FROGS AFFILIATION FILTER TOOL

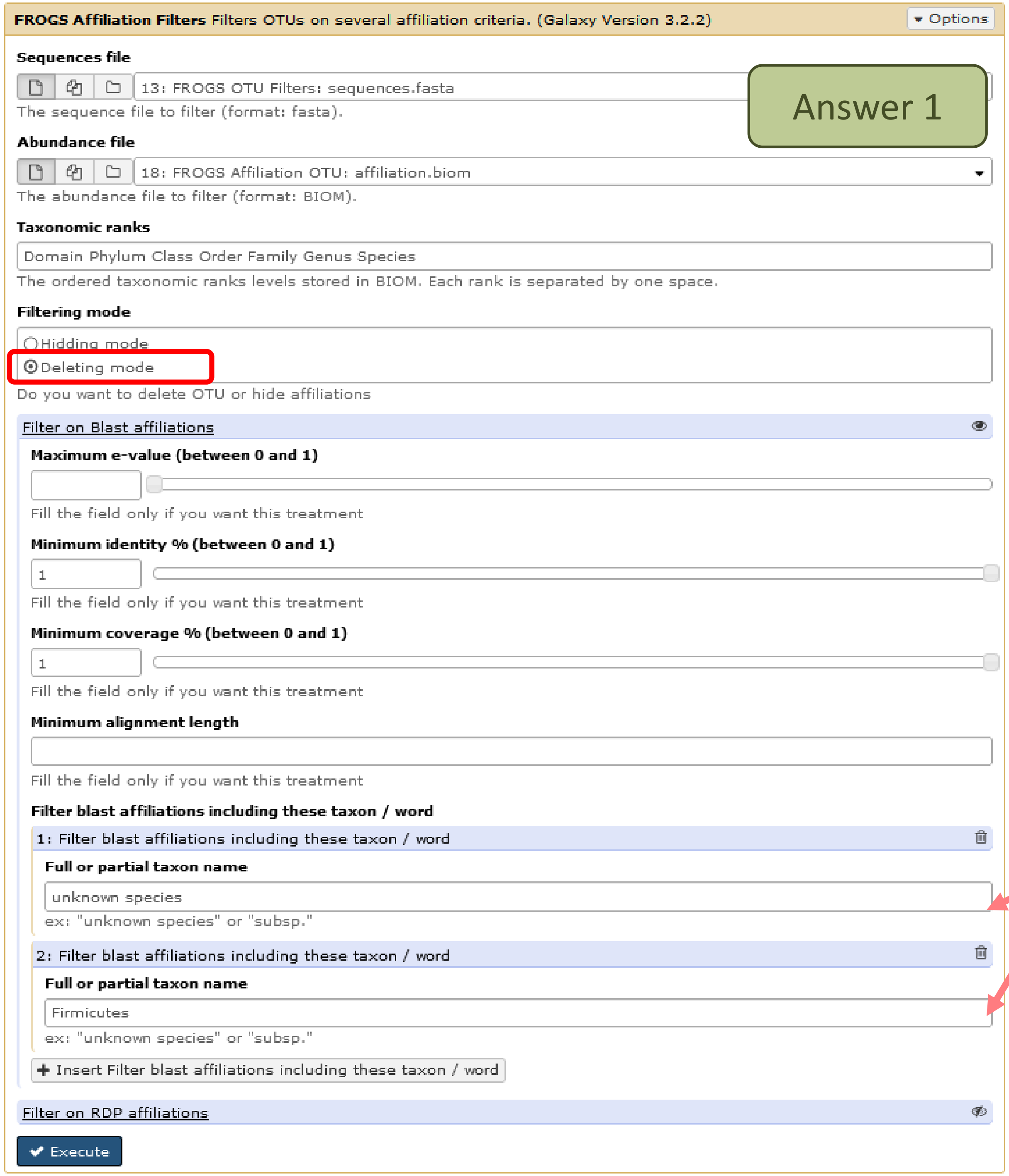
Exercice:

1. Apply filters to keep only sequences with perfect alignment with Silva sequences and affilliations without « unknown species » and « Firmicutes » terms. (deleting mode)

2. Apply filters to hide OTU affiliations that have not a perfect alignment with Silva sequences and the affilliations without « unknown species » and « Firmicutes » terms.

3. In deleting mode:
   - How many OTUs remain?
   - Among OTUs with multiaffiliation, How many were impacted/modified ?

4. In hidding mode:
   - What outputs change between deleted mode and hidding mode ?



Answer 1

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

Sequences file

13: FROGS OTU Filters: sequences.fasta

The sequence file to filter (format: fasta).

Abundance file

18: FROGS Affiliation OTU: affiliation.biom

The abundance file to filter (format: BIOM).

Taxonomic ranks

Domain Phylum Class Order Family Genus Species

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

Filtering mode

○ Hidding mode
● Deleting mode

Do you want to delete OTU or hide affiliations

Filter on Blast affiliations

Maximum e-value (between 0 and 1)

Fill the field only if you want this treatment

Minimum identity % (between 0 and 1)

1

Fill the field only if you want this treatment

Minimum coverage % (between 0 and 1)

1

Fill the field only if you want this treatment

Minimum alignment length

Fill the field only if you want this treatment

Filter blast affiliations including these taxon / word

1: Filter blast affiliations including these taxon / word

Full or partial taxon name

unknown species

ex: "unknown species" or "subsp."

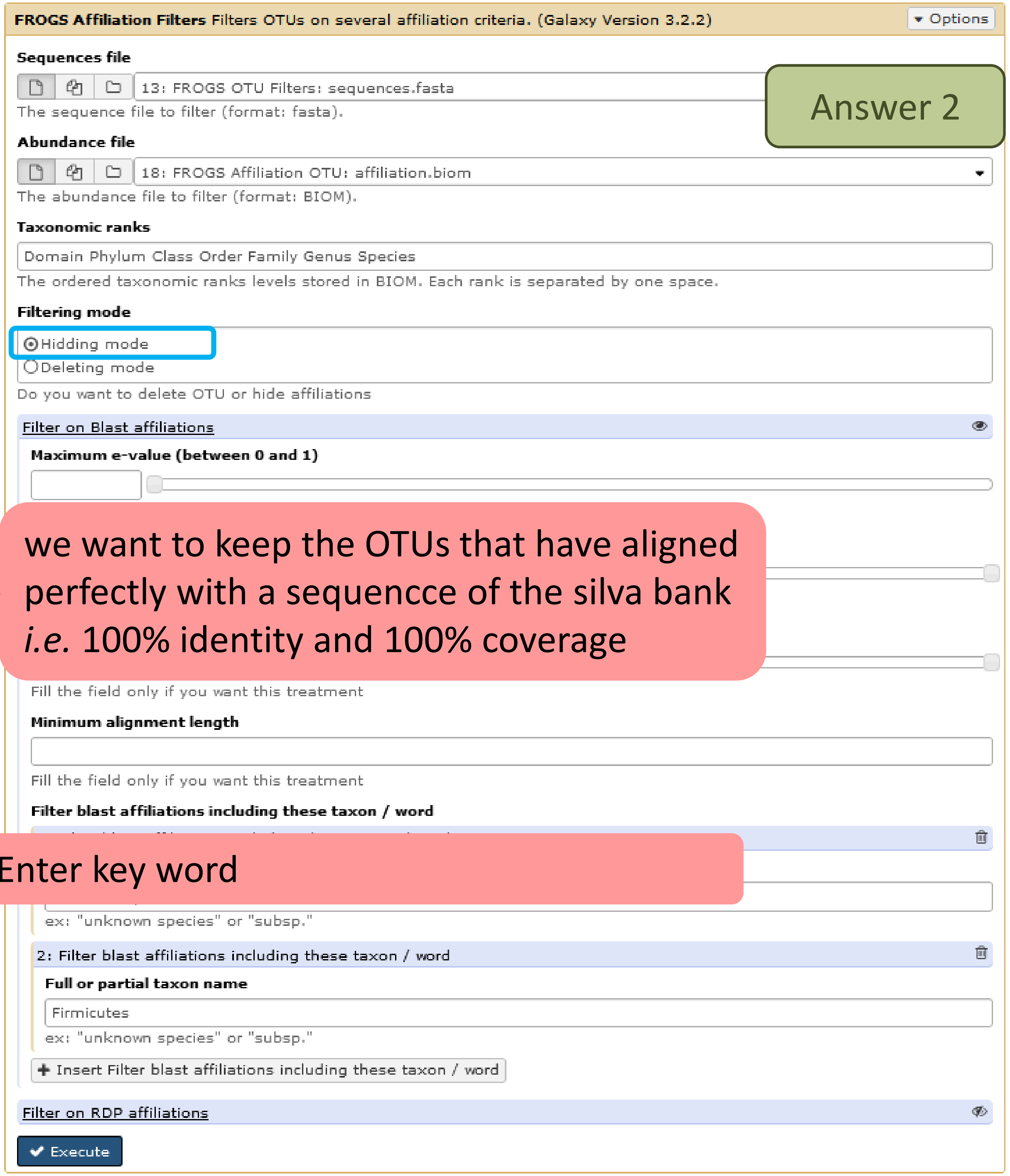
2: Filter blast affiliations including these taxon / word

Full or partial taxon name

Firmicutes

ex: "unknown species" or "subsp."

✚ Insert Filter blast affiliations including these taxon / word
Filter on RDP affiliations
✔ Execute



Answer 2

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

Sequences file

13: FROGS OTU Filters: sequences.fasta

The sequence file to filter (format: fasta).

Abundance file

18: FROGS Affiliation OTU: affiliation.biom

The abundance file to filter (format: BIOM).

Taxonomic ranks

Domain Phylum Class Order Family Genus Species

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

Filtering mode

● Hidding mode
○ Deleting mode

Do you want to delete OTU or hide affiliations

Filter on Blast affiliations

Maximum e-value (between 0 and 1)

Fill the field only if you want this treatment

Minimum alignment length

Fill the field only if you want this treatment

Filter blast affiliations including these taxon / word

ex: "unknown species" or "subsp."

2: Filter blast affiliations including these taxon / word

Full or partial taxon name

Firmicutes

ex: "unknown species" or "subsp."

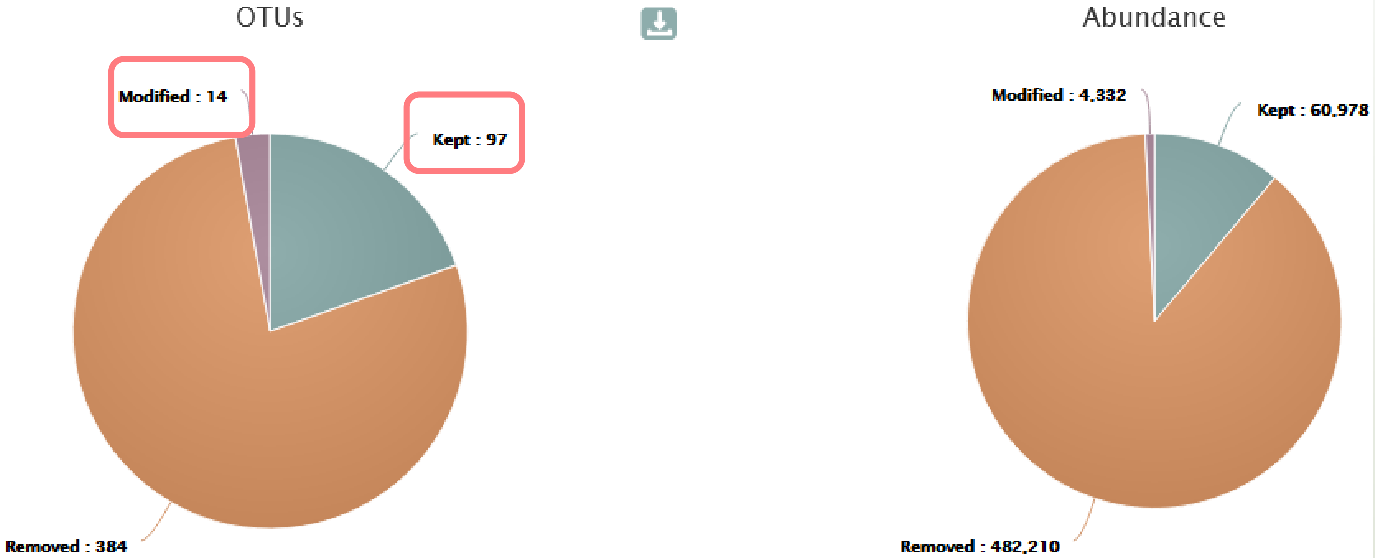
✚ Insert Filter blast affiliations including these taxon / word
Filter on RDP affiliations
✔ Execute

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

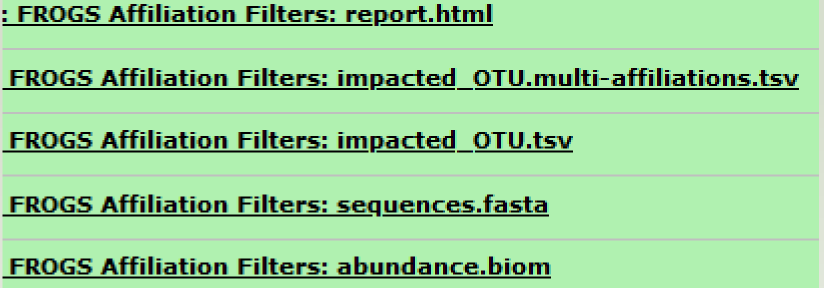
Enter key word

Answer 3

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



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



: FROGS Affiliation Filters: report.html

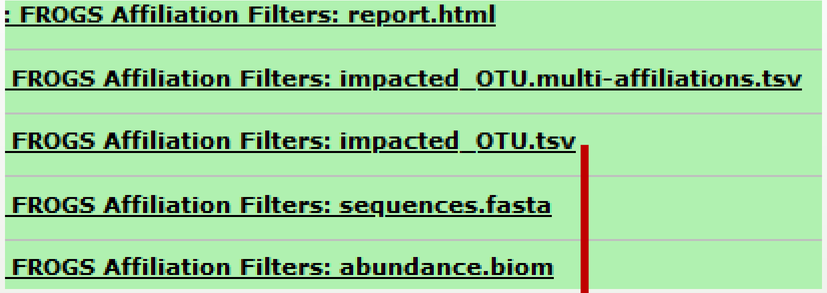
FROGS Affiliation Filters: impacted_OTU.multi-affiliations.tsv

FROGS Affiliation Filters: impacted_OTU.tsv

FROGS Affiliation Filters: sequences.fasta

FROGS Affiliation Filters: abundance.biom

Answer 3



: FROGS Affiliation Filters: report.html

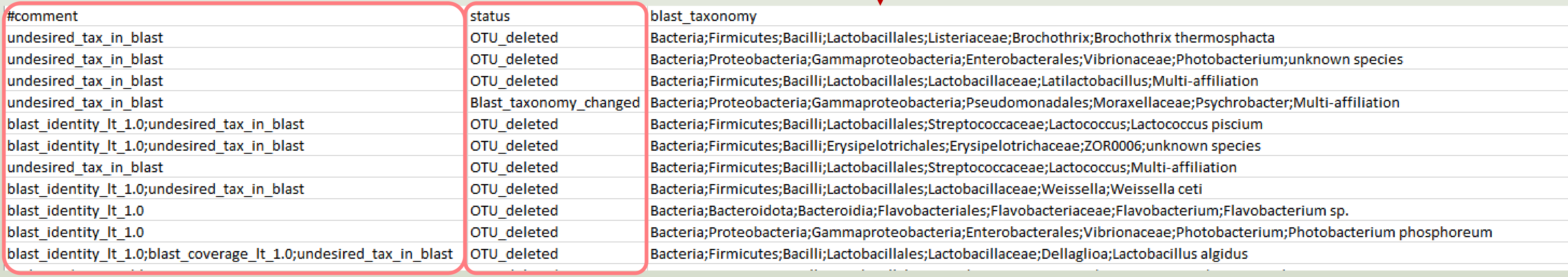
FROGS Affiliation Filters: impacted_OTU.multi-affiliations.tsv

FROGS Affiliation Filters: impacted_OTU.tsv

FROGS Affiliation Filters: sequences.fasta

FROGS Affiliation Filters: abundance.biom

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



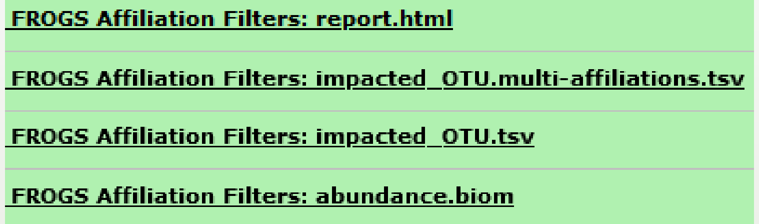
#comment	status	blast_taxonomy
undesired_tax_in_blast	OTU_deleted	Bacteria;Firmicutes;Bacilli;Lactobacillales;Listeriaceae;Brochothrix;Brochothrix thermosphacta
undesired_tax_in_blast	OTU_deleted	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales;Vibrionaceae;Photobacterium;unknown species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

In impacted_OTU.tsv

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

Answer 4

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


FROGS Affiliation Filters: report.html

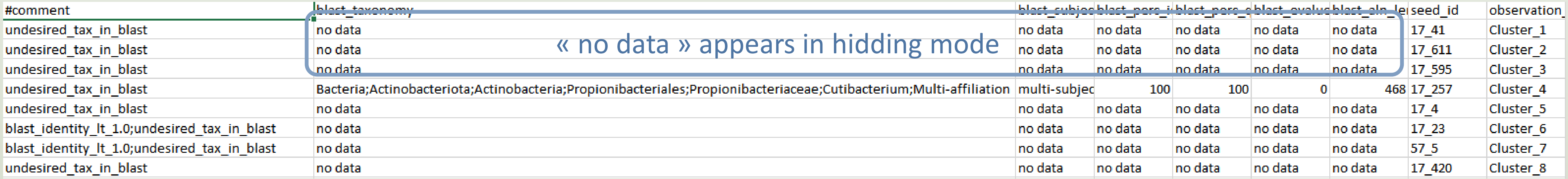
FROGS Affiliation Filters: impacted_OTU.multi-affiliations.tsv

FROGS Affiliation Filters: impacted_OTU.tsv

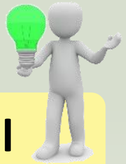
FROGS Affiliation Filters: abundance.biom

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

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


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

« no data » appears in hidding mode



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

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

Exercise 8

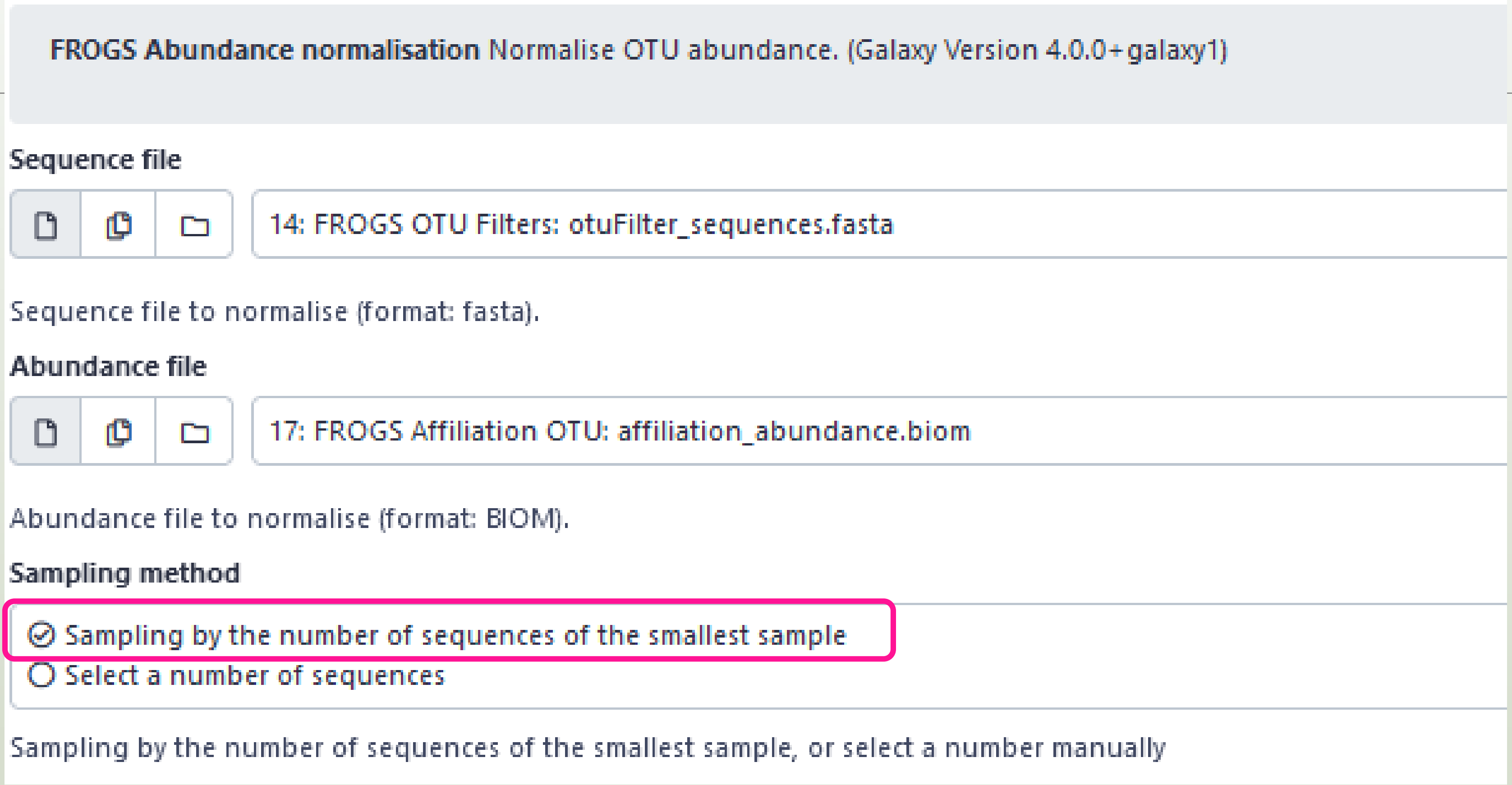
Which values are interesting to test?

Exercise 8

1. Normalize your data from Affiliation based on the smallest samples
2. Normalize your data on 2000 sequences or less
3. Normalize your data on 8000 sequences
4. What differences with or without

Answer 2

Q1: Normalize your data from Affiliation based on this number of sequence



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

Sequence file

14: FROGS OTU Filters: otuFilter_sequences.fasta

Sequence file to normalise (format: fasta).

Abundance file

17: FROGS Affiliation OTU: affiliation_abundance.biom

Abundance file to normalise (format: BIOM).

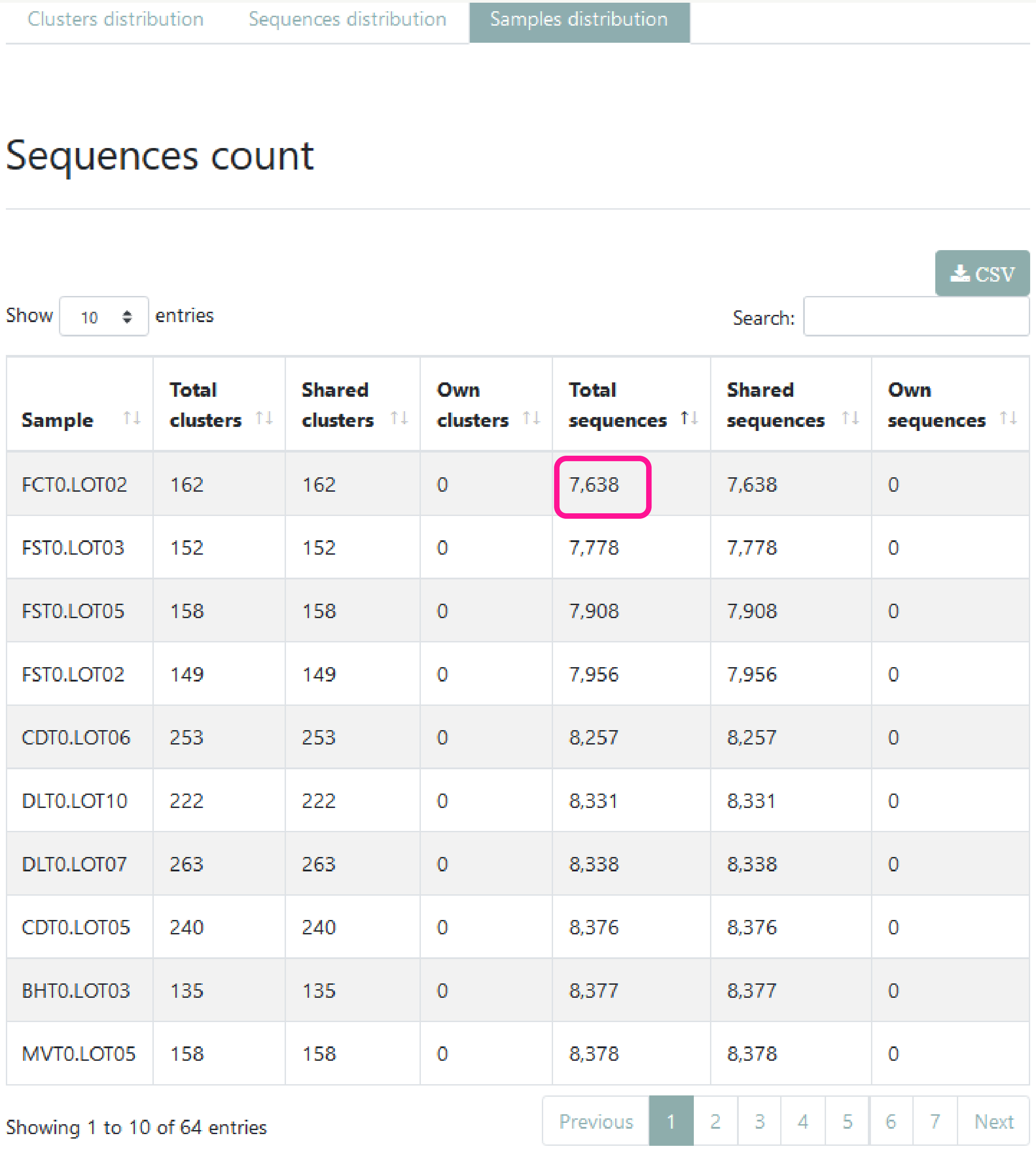
Sampling method

● Sampling by the number of sequences of the smallest sample
○ Select a number of sequences

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

Answer 1

The smallest sequenced samples



Sequences count

Sample	Total clusters	Shared clusters	Own clusters	Total sequences	Shared sequences	Own sequences
FCT0.LOT02	162	162	0	7,638	7,638	0
FST0.LOT03	152	152	0	7,778	7,778	0
FST0.LOT05	158	158	0	7,908	7,908	0
FST0.LOT02	149	149	0	7,956	7,956	0
CDT0.LOT06	253	253	0	8,257	8,257	0
DLT0.LOT10	222	222	0	8,331	8,331	0
DLT0.LOT07	263	263	0	8,338	8,338	0
CDT0.LOT05	240	240	0	8,376	8,376	0
BHT0.LOT03	135	135	0	8,377	8,377	0
MVT0.LOT05	158	158	0	8,378	8,378	0

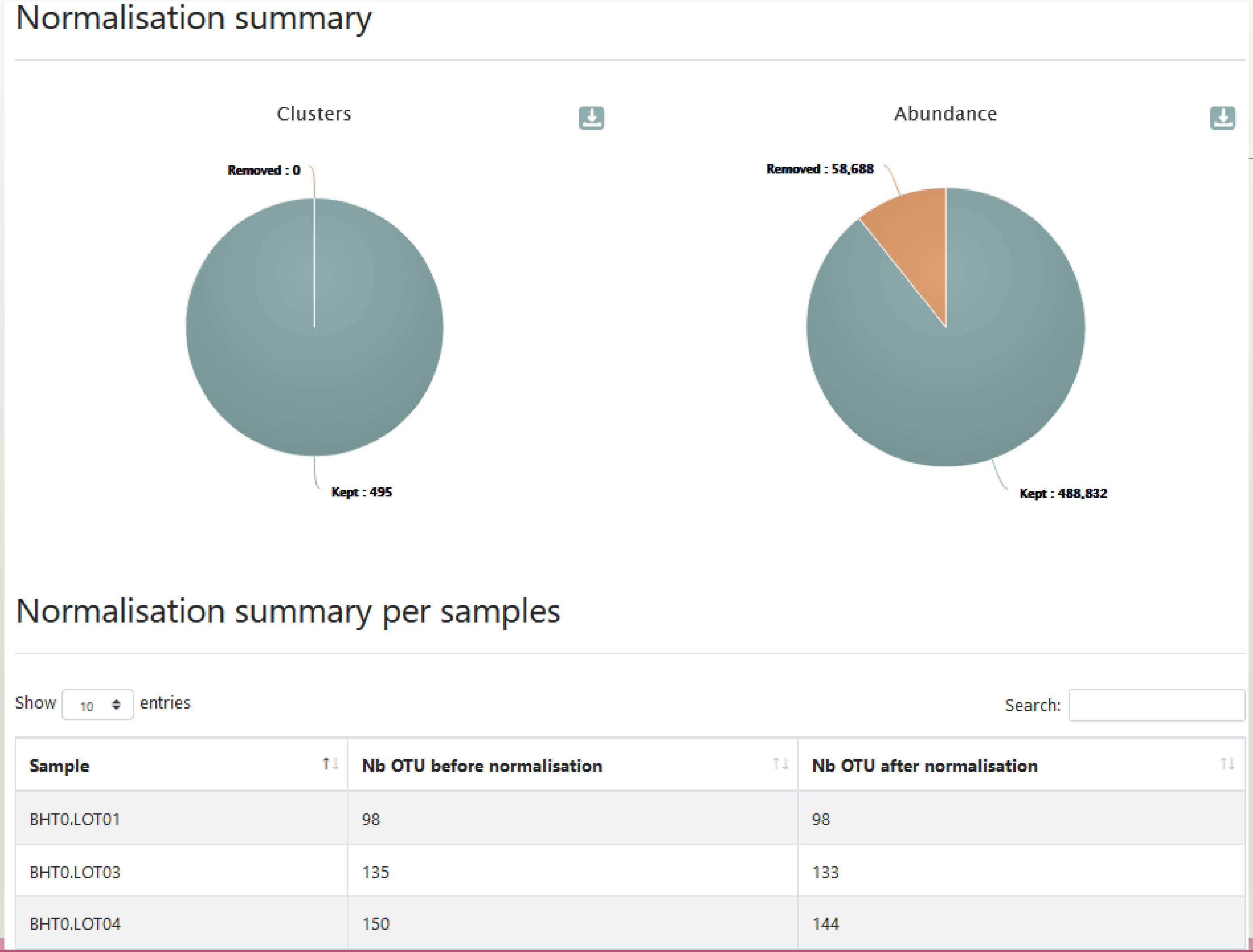
Showing 1 to 10 of 64 entries

Thanks to Clusterstat output, you can know what is the size of the smallest sample.
Sort by **Total sequences**
i.e. 7638 sequences

7638 is the maximal size that you can ask for normalizing the sample sizes.

Answer 1

Q1: Normalize your data from Affiliation based on this number of sequence



Normalisation summary

Normalisation summary per samples

Show 10 entries

Search:

Sample	Nb OTU before normalisation	Nb OTU after normalisation
BHT0.LOT01	98	98
BHT0.LOT03	135	133
BHT0.LOT04	150	144

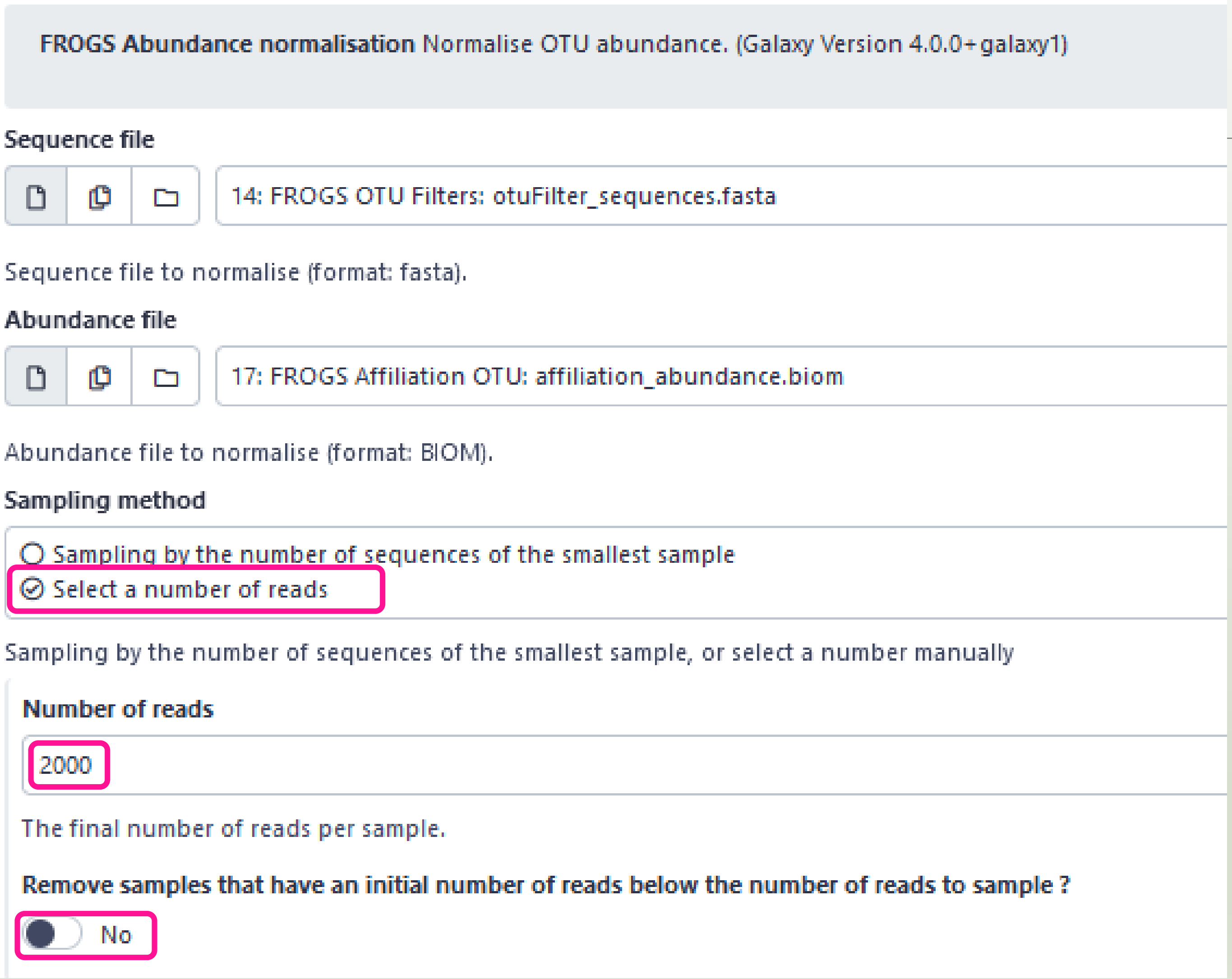
Auto-selection of the minimal number of OTUs *i.e.* 7638 sequences

495 OTUs
488832 sequences

The minimum impact of OTU number per sample

Answer 2

Q2: Normalize your data on 2000 sequences or less



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

Sequence file

14: FROGS OTU Filters: otuFilter_sequences.fasta

Sequence file to normalise (format: fasta).

Abundance file

17: FROGS Affiliation OTU: affiliation_abundance.biom

Abundance file to normalise (format: BIOM).

Sampling method

○ Sampling by the number of sequences of the smallest sample

☑ Select a number of reads

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

Number of reads

2000

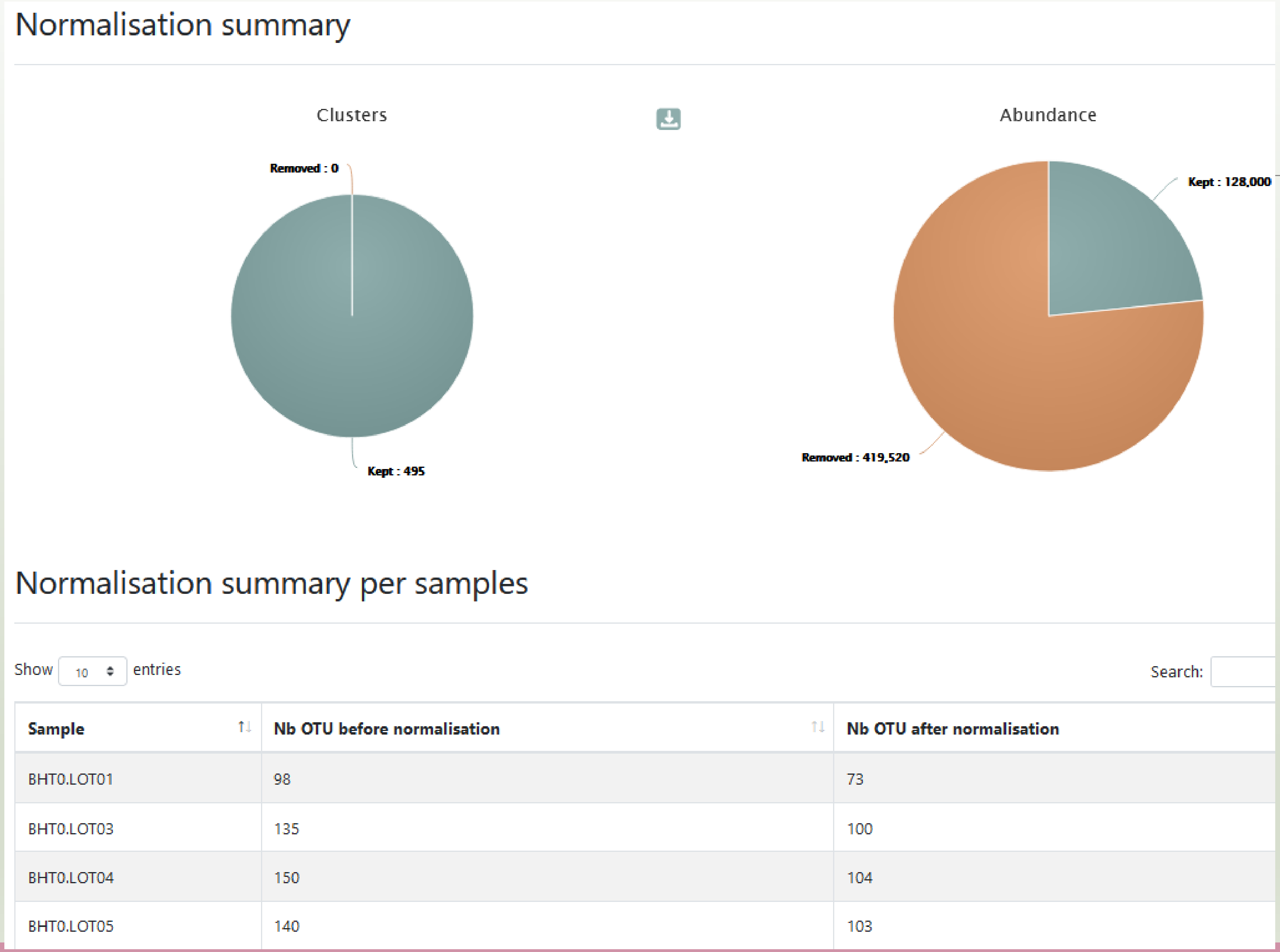
The final number of reads per sample.

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

No

Answer 2

Q2: Normalize your data on 2000 sequences or less


Normalisation summary

Normalisation summary per samples

Show 10 entries

Search:

Sample	Nb OTU before normalisation	Nb OTU after normalisation
BHT0.LOT01	98	73
BHT0.LOT03	135	100
BHT0.LOT04	150	104
BHT0.LOT05	140	103

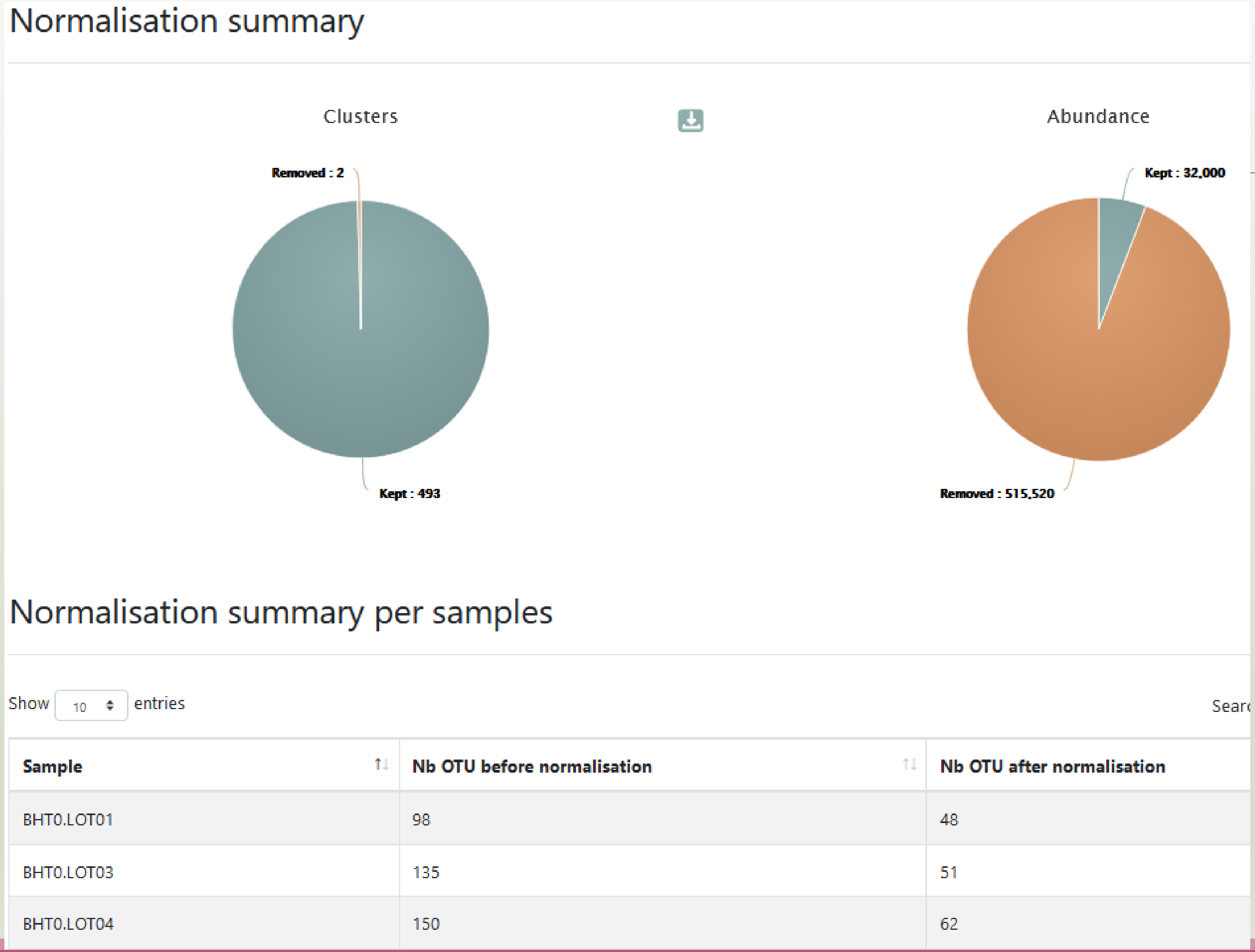
Normalization at 2000 sequences

495 OTUs
128000 sequences

Big impact of OTU number per sample

Answer 2

Q2: Normalize your data on 2000 sequences or less



Normalisation summary

Normalisation summary per samples

Show 10 entries

Searc

Sample	Nb OTU before normalisation	Nb OTU after normalisation
BHT0.LOT01	98	48
BHT0.LOT03	135	51
BHT0.LOT04	150	62

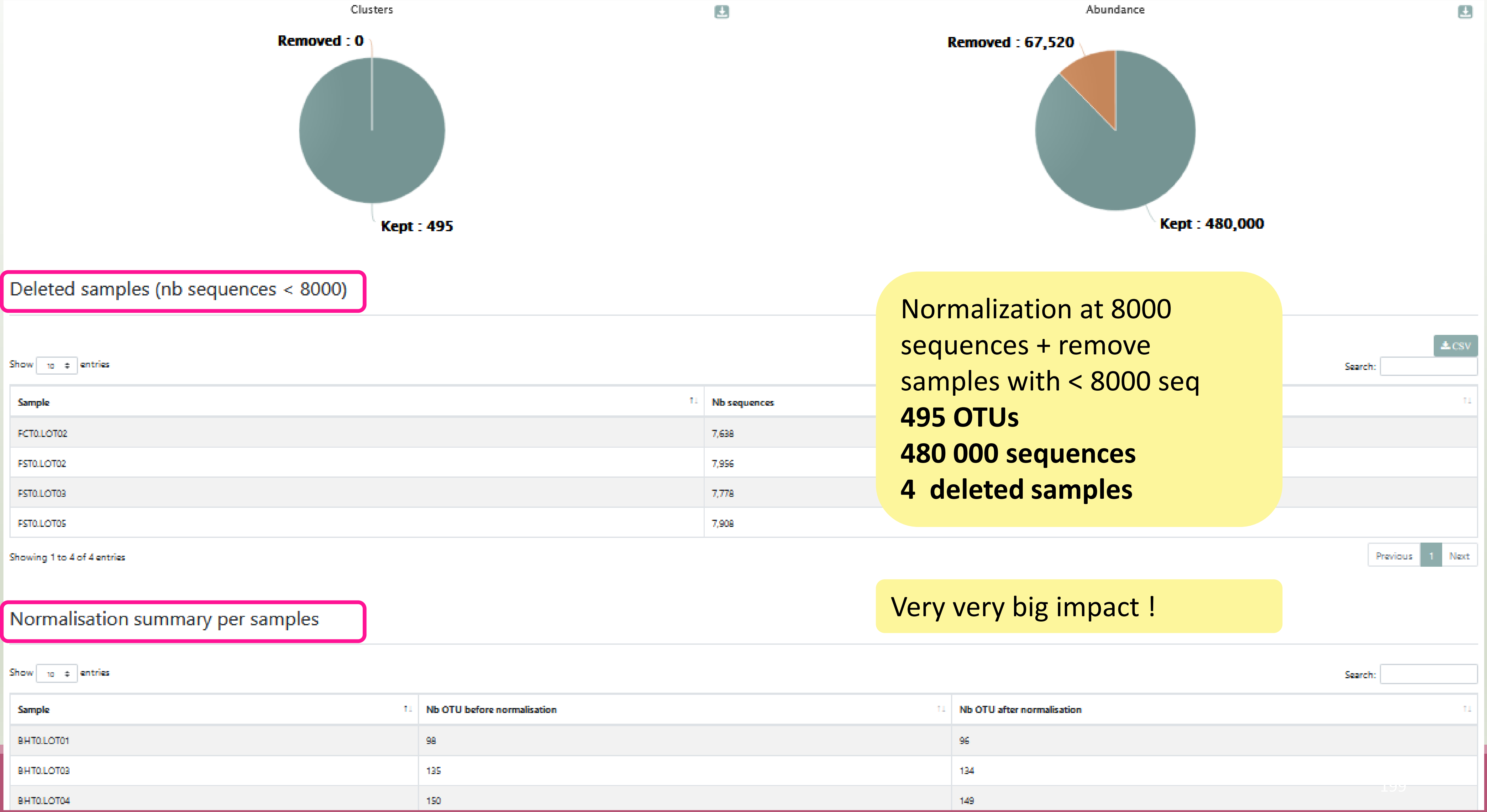
Normalization at 500 sequences

493 OTUs
32000 sequences

Very big impact of OTU number per sample

Answer 3

Q3: Normalize your data on 8000 sequences – with option “removing sample”



Deleted samples (nb sequences < 8000)

CSV

Show 10 entries

Search:

Sample	Nb sequences
FCT0.LOT02	7,638
FST0.LOT02	7,956
FST0.LOT03	7,778
FST0.LOT05	7,908

Showing 1 to 4 of 4 entries

Previous 1 Next

Normalisation summary per samples

Show 10 entries

Search:

Sample	Nb OTU before normalisation	Nb OTU after normalisation
BHT0.LOT01	98	96
BHT0.LOT03	135	134
BHT0.LOT04	150	149

Normalization at 8000 sequences + remove samples with < 8000 seq
495 OTUs
480 000 sequences
4 deleted samples

Very very big impact !

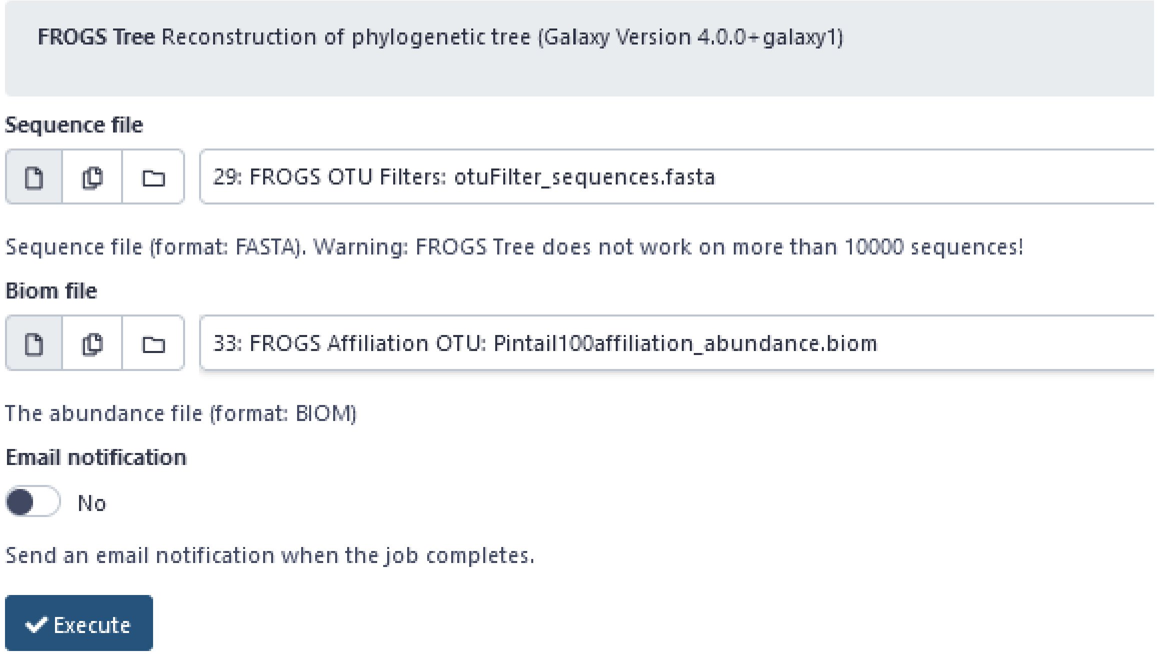
FROGS Tree

CREATE A PHYLOGENETICS TREE OF OTUS

FROGS Tree

This tool builds a phylogenetic tree thanks to affiliations of OTUs contained in the BIOM file

It uses MAFFT for the multiple alignment and FastTree for the phylogenetic tree.



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

Sequence file

29: FROGS OTU Filters: otuFilter_sequences.fasta

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

Biom file

33: FROGS Affiliation OTU: Pintail100affiliation_abundance.biom

The abundance file (format: BIOM)

Email notification

● No

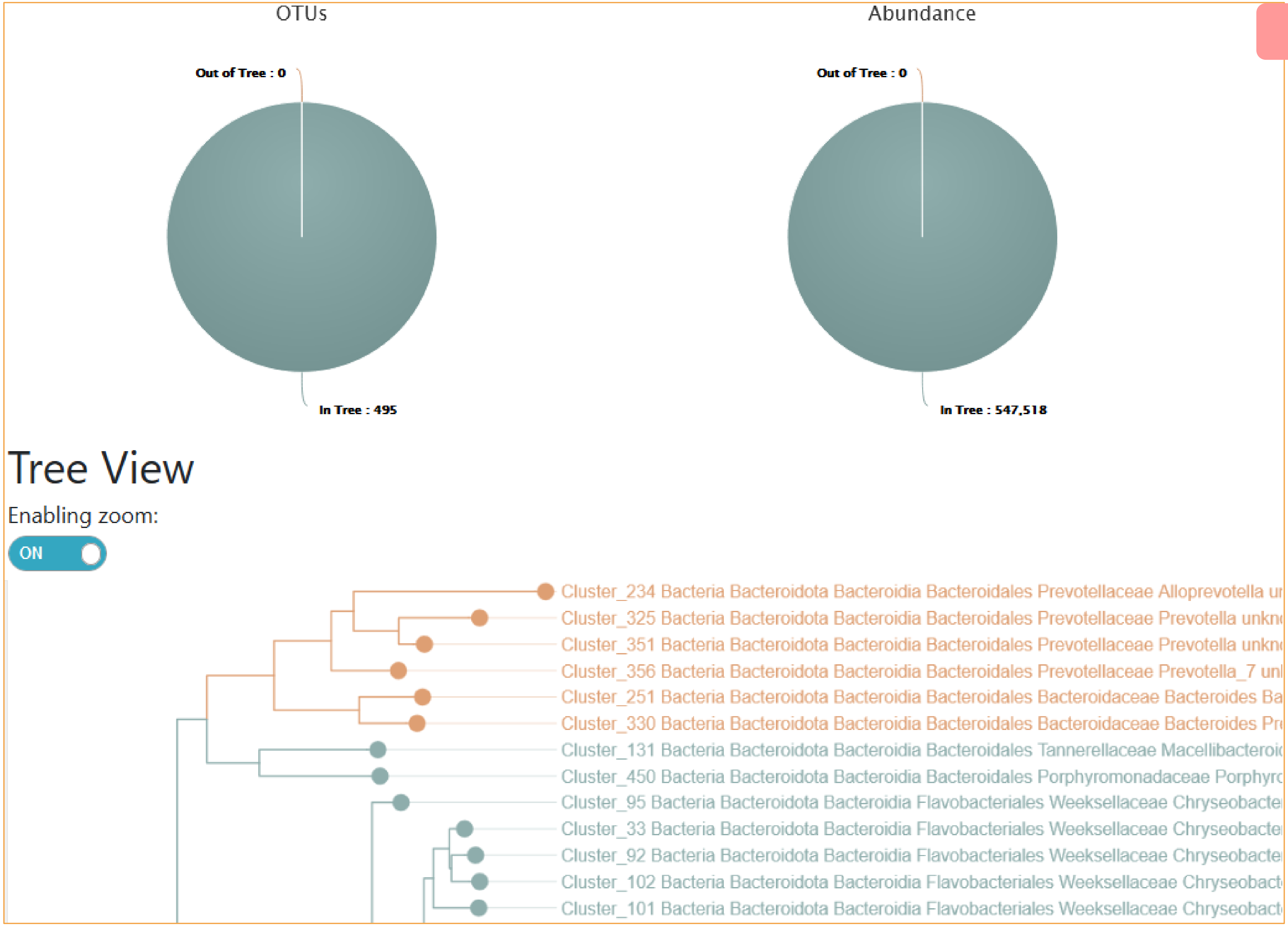
Send an email notification when the job completes.

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

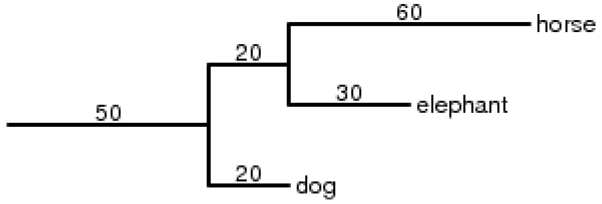
Enabling zoom:

ON

Cluster_234 Bacteria Bacteroidota Bacteroidia Bacteroidales Prevotellaceae Alloprevotella ur
Cluster_325 Bacteria Bacteroidota Bacteroidia Bacteroidales Prevotellaceae Prevotella unkn
Cluster_351 Bacteria Bacteroidota Bacteroidia Bacteroidales Prevotellaceae Prevotella unkn
Cluster_356 Bacteria Bacteroidota Bacteroidia Bacteroidales Prevotellaceae Prevotella_7 unl
Cluster_251 Bacteria Bacteroidota Bacteroidia Bacteroidales Bacteroidaceae Bacteroides Ba
Cluster_330 Bacteria Bacteroidota Bacteroidia Bacteroidales Bacteroidaceae Bacteroides Pr
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

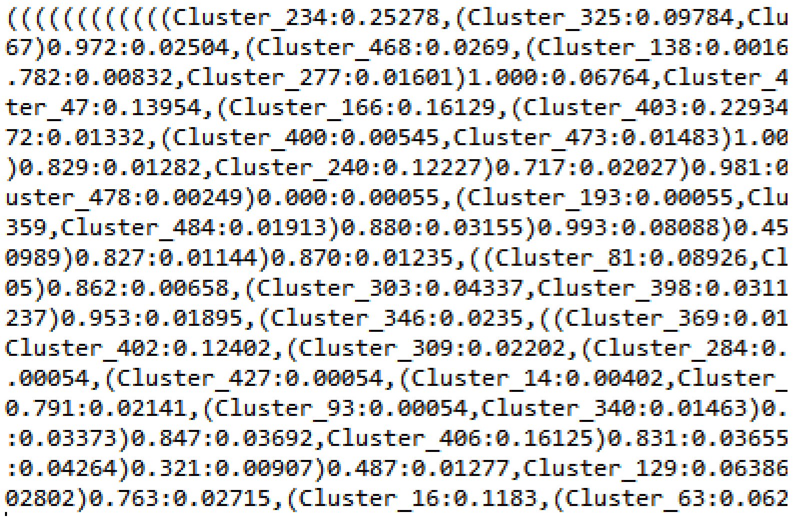
2nd output

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



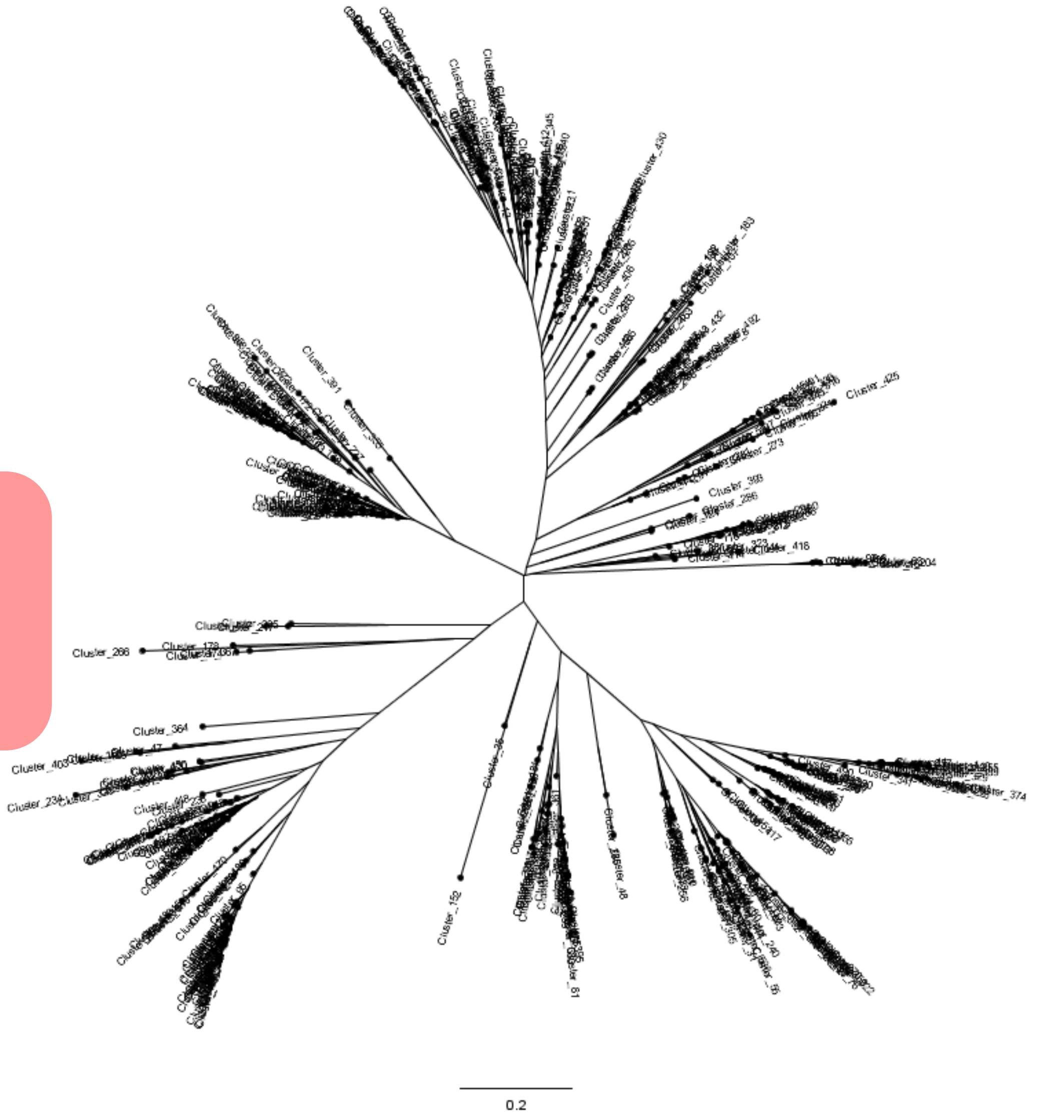
→ (dog:20, (elephant:30, horse:60):20):50

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
```

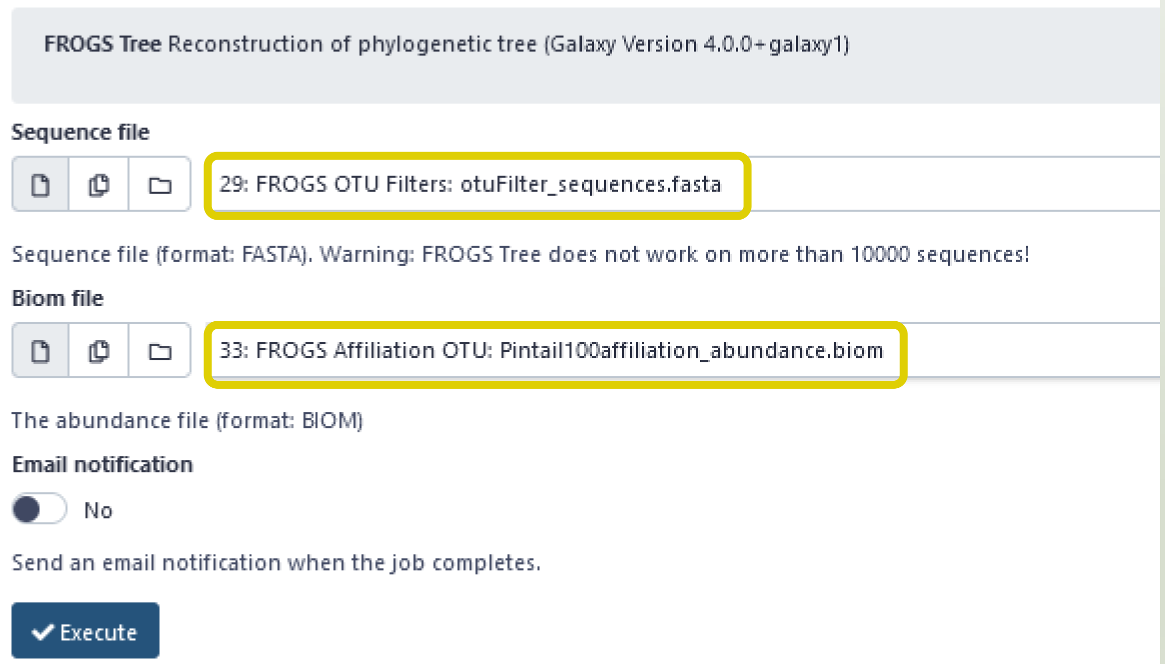
Exemple of visualization in FigTree from nhx file



Practice:

Exercice:

1. Create the phylogenetic tree that will be used for statistical analyses.



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

Sequence file

29: FROGS OTU Filters: otuFilter_sequences.fasta

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

Biom file

33: FROGS Affiliation OTU: Pintail100affiliation_abundance.biom

The abundance file (format: BIOM)

Email notification

● No

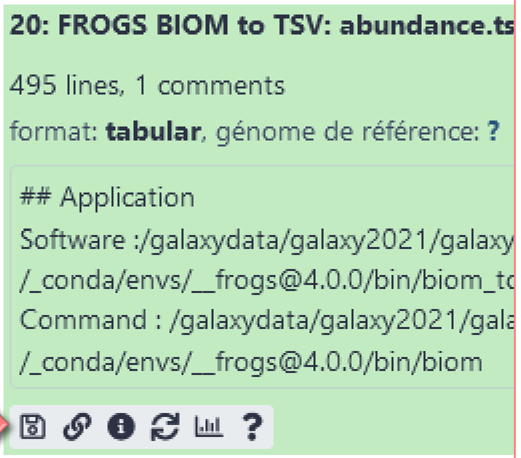
Send an email notification when the job completes.

For tutorial, we ask you to create a phylogentic tree on affiliation.biom **before** "**affiliation filter**" process. Otherwise on your own data, create the phylogenetic tree on cleaned affiliation.biom

Download your data

In order to share resources as well as possible, files that have not been accessed for more than 120 days are regularly purged. The backup of data generated using
of Galaxy is <u>your responsibility</u>.

You have 2 backup possibilities:
1. Save your datasets one by one using the "floppy disk" icon.



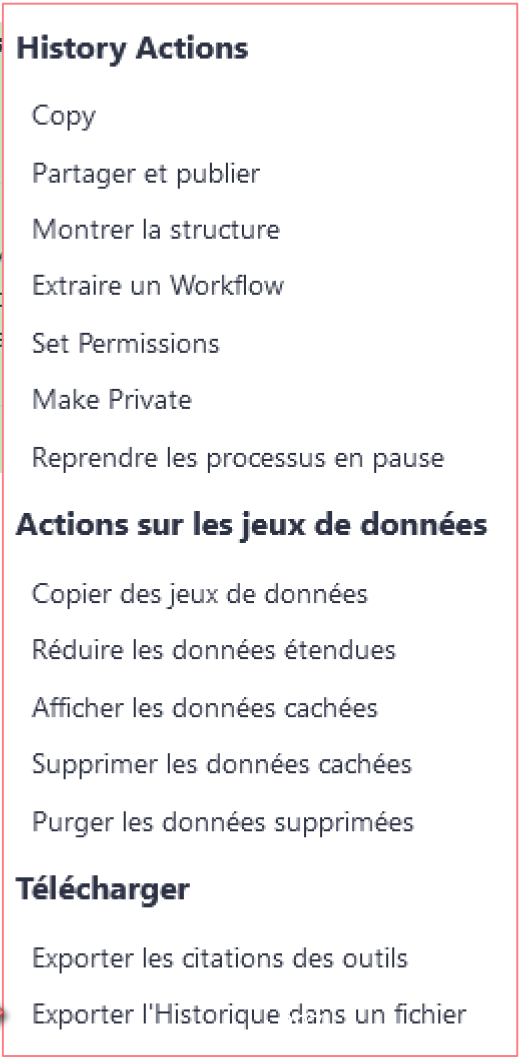
20: FROGS BIOM to TSV: abundance.ts

495 lines, 1 comments

format: **tabular**, génome de référence: **?**

```
## Application
Software :/galaxydata/galaxy2021/galaxy
/_conda/envs/__frogs@4.0.0/bin/biom_tc
Command : /galaxydata/galaxy2021/gala
/_conda/envs/__frogs@4.0.0/bin/biom
```

2. Or export each history.
To export a history, from the "History" menu, click on the wheel, then "Export History to File":



History Actions

Copy

Partager et publier

Montrer la structure

Extraire un Workflow

Set Permissions

Make Private

Reprendre les processus en pause

Actions sur les jeux de données

Copier des jeux de données

Réduire les données étendues

Afficher les données cachées

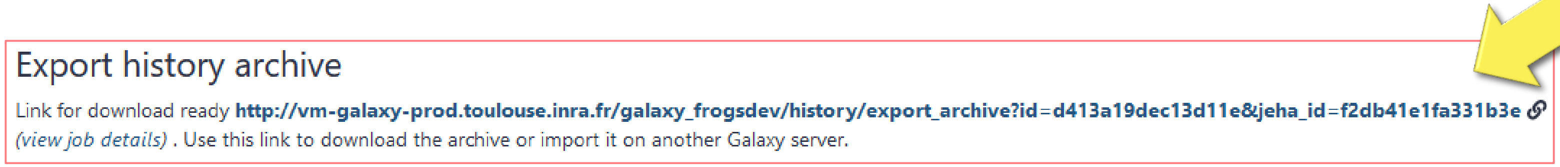
Supprimer les données cachées

Purger les données supprimées

Télécharger

Exporter les citations des outils

Exporter l'Historique dans un fichier



Export history archive

Link for download ready **http://vm-galaxy-prod.toulouse.inra.fr/galaxy_frogsdev/history/export_archive?id=d413a19dec13d11e&jeha_id=f2db41e1fa331b3e**
(view job details) . Use this link to download the archive or import it on another Galaxy server.

How to cite FROGS

Frédéric Escudié, Lucas Auer, Maria Bernard, Mahendra Mariadassou, Laurent Cauquil, Katia Vidal, Sarah Maman, Guillermina Hernandez-Raquet, Sylvie Combes, Géraldine Pascal.

"**FROGS**: Find, Rapidly, OTUs with Galaxy Solution." ***Bioinformatics,*** , Volume 34, Issue 8, 15 April 2018, Pages 1287–1294

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

Bioinformatics, 2017, 1–8
doi: 10.1093/bioinformatics/btx791
Advance Access Publication Date: 7 December 2017
Original Paper

OXFORD

Sequence analysis

FROGS: Find, Rapidly, OTUs with Galaxy Solution

Frédéric Escudié[1,†], Lucas Auer[2,†], Maria Bernard[3], Mahendra Mariadassou[4], Laurent Cauquil[5], Katia Vidal[5], Sarah Maman[5], Guillermina Hernandez-Raquet[6], Sylvie Combes[5] and Géraldine Pascal[5,*]

[1]Bioinformatics platform Toulouse Midi-Pyrenees, MIAT, INRA Auzeville CS 52627 31326 Castanet Tolosan cedex, France, [2]INRA, UMR 1136, Université de Lorraine, INRA-Nancy, 54280, Champenoux, France, [3]GABI, INRA, AgroParisTech, Université Paris-Saclay, Jouy-en-Josas, France, [4]MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France, [5]GenPhySE, Université de Toulouse, INRA, INPT, ENVT, Castanet Tolosan, France and [6]Laboratoire d'ingénierie des Systèmes Biologiques et des Procédés-LISBP, Université de Toulouse, INSA, INRA, CNRS, Toulouse, France

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Bonnie Berger

Received on May 10, 2017; revised on December 1, 2017; editorial decision on December 4, 2017; accepted on December 5, 2017

Abstract

Motivation: Metagenomics leads to major advances in microbial ecology and biologists need user friendly tools to analyze their data on their own.
Results: This Galaxy-supported pipeline, called FROGS, is designed to analyze large sets of amplicon sequences and produce abundance tables of Operational Taxonomic Units (OTUs) and their taxonomic affiliation. The clustering uses Swarm. The chimera removal uses VSEARCH, combined
with original cross-sample validation. The
affiliation output to highlight databases confl
ous graphical illustrations are produced along
for the detection and quantification of OTUs
robust and highly sensitive. It compares fa
QIIME.
Availability and implementation: Source cod
geraldinepascal/FROGS.git. A companion web
Contact: geraldine.pascal@inra.fr
Supplementary information: Supplementary

1 Introduction

The expansion of high-throughput sequencing of rRN
has opened new horizons for the study of microbial
By making it possible to study all micro-organisms f
environment without the need to cultivate them, metag
led to major advances in many fields of microbial ecolo
study of the impact of microbiota on human and animal

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

OXFORD

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

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

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

Maria Bernard †, Olivier Rué †, Mahendra Mariadassou and Géraldine Pascal

Corresponding author: Geraldine Pascal, GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France. Tel.: +33 (0)5 61 28 51 05; E-mail: geraldine.pascal@inrae.fr
†Maria Bernard and Olivier Rué are joint first authors.

Abstract

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

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

Introduction

Using amplicon sequencing to describe the microbial composition of an environment is a time saving and cost-effective strategy and can be used even for very large-scale surveys [1]. Most studies currently focus on the bacterial fraction of microbial communities but the fungal fraction is equally important, as fungi are ubiquitous and provide several ecosystem services [2]. Unfortunately, studying the fungal fraction using metabarcoding has its own challenges. Indeed, in fungi, there is no equivalent of the 16S rRNA gene, which is widely used and highly suitable for bacteria. The best candidates are internal transcribed spacers (ITS), but these are more difficult to manipulate. The main problem with ITS is size polymorphism, with a size range of 361–1475 bases in UNITE 7.1 [3] (unlike 16S where 95% of the sequences have a length between 1205 and 1556 bases). Most studies describing ITS data analyses process either (i) paired-end reads but filter out non-overlapping, non-mergeable reads, thus systematically discarding taxa with longer ITS, or (ii) single-end reads, thus limiting taxonomic resolution and losing the benefit of information contained in longer sequences [4, 5].

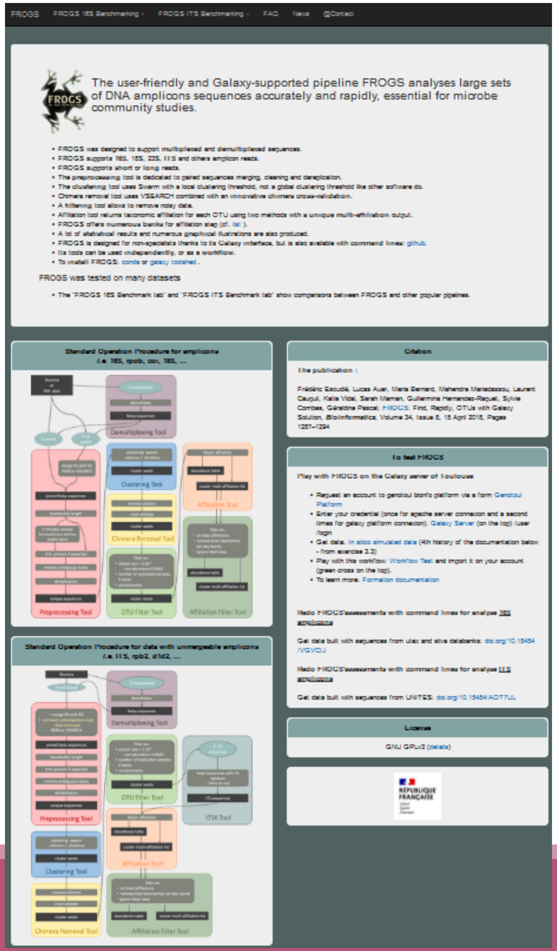
Maria Bernard is a bioinformatics engineer. She is a member of a platform team conducting NGS sequence analysis and designing software. She specializes in workflow development in particular for metabarcoding analysis.
Olivier Rué is a bioinformatics engineer. He is in charge of data analysis at the Migale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

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

1

FROGS’docs

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



FROGS FROGS 16S Benchmarking - FROGS ITS Benchmarking - FAQ News @Contact

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

- FROGS was designed to support multiplexed and demultiplexed sequences.
- FROGS supports 16S, 18S, 23S, ITS and others amplicon reads.
- FROGS supports short or long reads.
- The preprocessing tool is dedicated to paired sequences merging, cleaning and dereplication.
- The clustering tool uses Swarm with a local clustering threshold, not a global clustering threshold like other software do.
- Chimera removal tool uses VSEARCH combined with an innovative chimera cross-validation.
- A filtering tool allows to remove noisy data.
- Affiliation tool returns taxonomic affiliation for each OTU using two methods with a unique multi-affiliation output.
- FROGS offers numerous banks for affiliation step (cf. list).
- A lot of statistical results and numerous graphical illustrations are also produced.
- FROGS is designed for non-specialists thanks to its Galaxy interface, but is also available with command lines: github.
- Its tools can be used independently, or as a workflow.
- To install FROGS: conda or galaxy toolshed .

FROGS was tested on many datasets

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

Standard Operation Procedure for amplicons
i.e. 16S, rpob, cox, 18S, ...

Standard Operation Procedure for data with unmergeable amplicons
i.e. ITS, rpb2, d1d2, ...

Citation

The publication :

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

To test FROGS

Play with FROGS on the Galaxy server of Toulouse

- Request an account to genotoul bioinfo platform via a form Genotoul Platform
- Enter your credential (once for apache server connexion and a second times for galaxy platform connexion). Galaxy Server (on the top) /user /login
- Get data: In silico simulated data (4th history of the documentation below - from exercise 3.2)
- Play with this workflow: Workflow Test and import it on your account (green cross on the top).
- To learn more: Formation documentation

Redo FROGS'assessments with command lines for analyses 16S amplicons

Get data built with sequences from utax and silva databanks: doi.org/10.15454/VGVOIJ

Redo FROGS'assessments with command lines for analyses ITS amplicons

Get data built with sequences from UNITES: doi.org/10.15454/AOTTUL

License

GNU GPLv3 (details)

RÉPUBLIQUE FRANÇAISE

All scripts on Github:

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

To contact

FROGS support:

frogs-support@inrae.fr

Newsletter – subscription request:

frogs-support@inrae.fr

Play list FROGS:

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