Training on Galaxy: Metagenomics

June 2016

Find Rapidly OTU with Galaxy Solution

FRÉDÉRIC ESCUDIÉ* and LUCAS AUER*, MARIA BERNARD, LAURENT CAUQUIL, KATIA VIDAL, SARAH MAMAN, MAHENDRA MARIADASSOU, GUILLERMINA HERNANDEZ-RAQUET, GÉRALDINE PASCAL

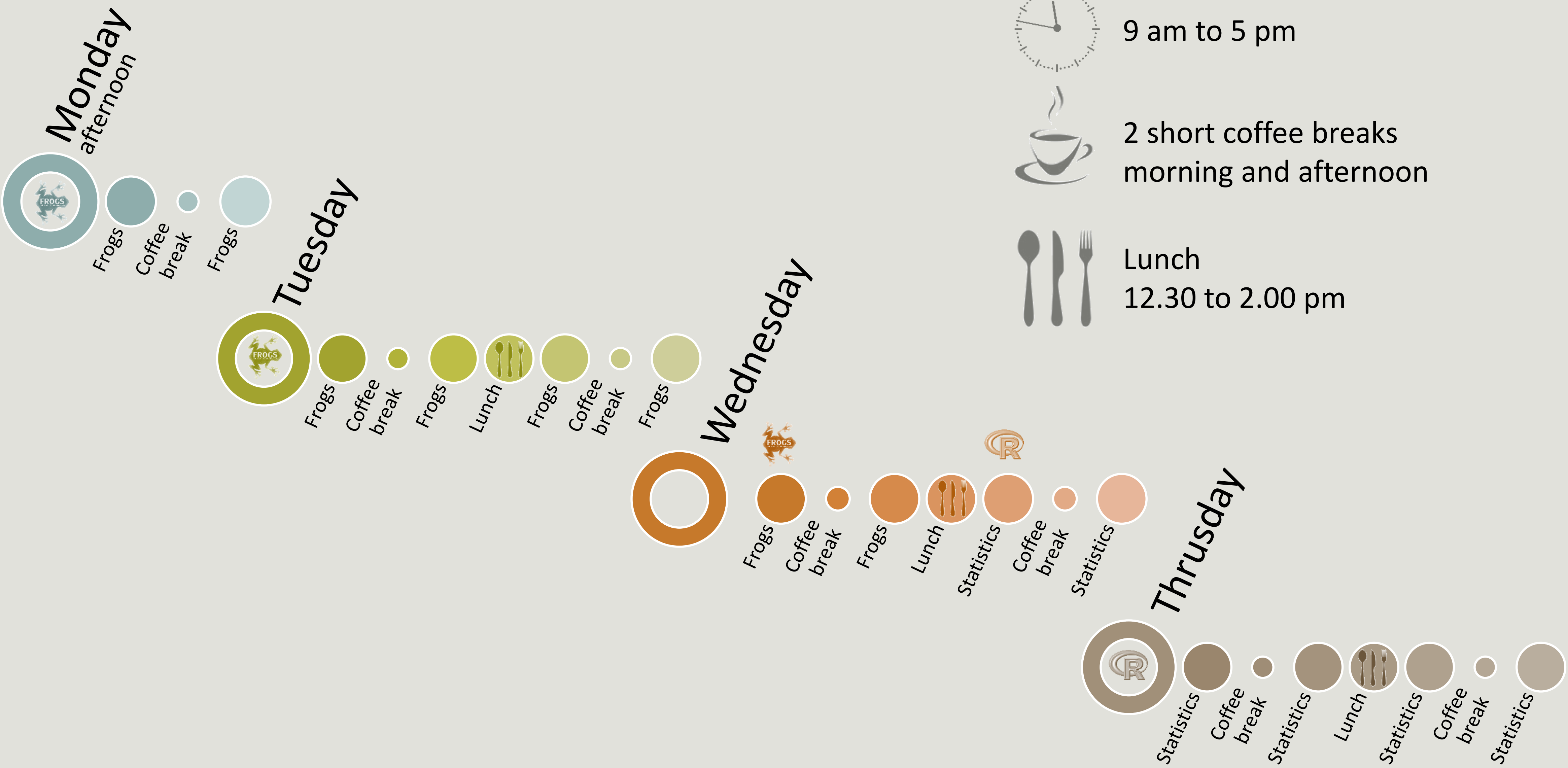
*THESE AUTHORS HAVE CONTRIBUTED EQUALLY TO THE PRESENT WORK.

Feedback:

What are your needs in “metagenomics"?

454 / MiSeq ?

Your background ?



Monday afternoon

Frogs
Coffee break
Frogs

Tuesday

Frogs
Coffee break
Frogs
Lunch
Frogs
Coffee break
Frogs

Wednesday

Frogs
Coffee break
Frogs
Lunch
Statistics
Coffee break
Statistics

Thrusday

Statistics
Coffee break
Statistics
Lunch
Statistics
Coffee break
Statistics

9 am to 5 pm

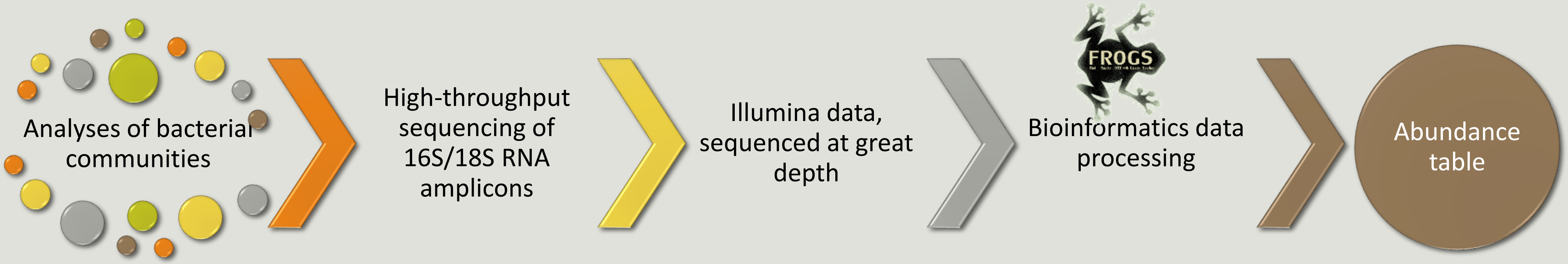
2 short coffee breaks
morning and afternoon

Lunch
12.30 to 2.00 pm

Overview

- Objectives
- Material: data + FROGS
- Data upload into galaxy environment
- Demultiplex tool
- Preprocessing
- Clustering + Cluster Statistics
- Chimera removal
- Filtering
- Affiliation + Affliation Statistics
- Normalization
- Tool descriptions
- Workflow creation
- Download data
- Some figures

Objectives



Analyses of bacterial communities

High-throughput sequencing of 16S/18S RNA amplicons

Illumina data, sequenced at great depth

Bioinformatics data processing

Abundance table

WITH
operational taxonomic units (OTUs) and
their taxonomic affiliation.

Objectives

	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

Objectives

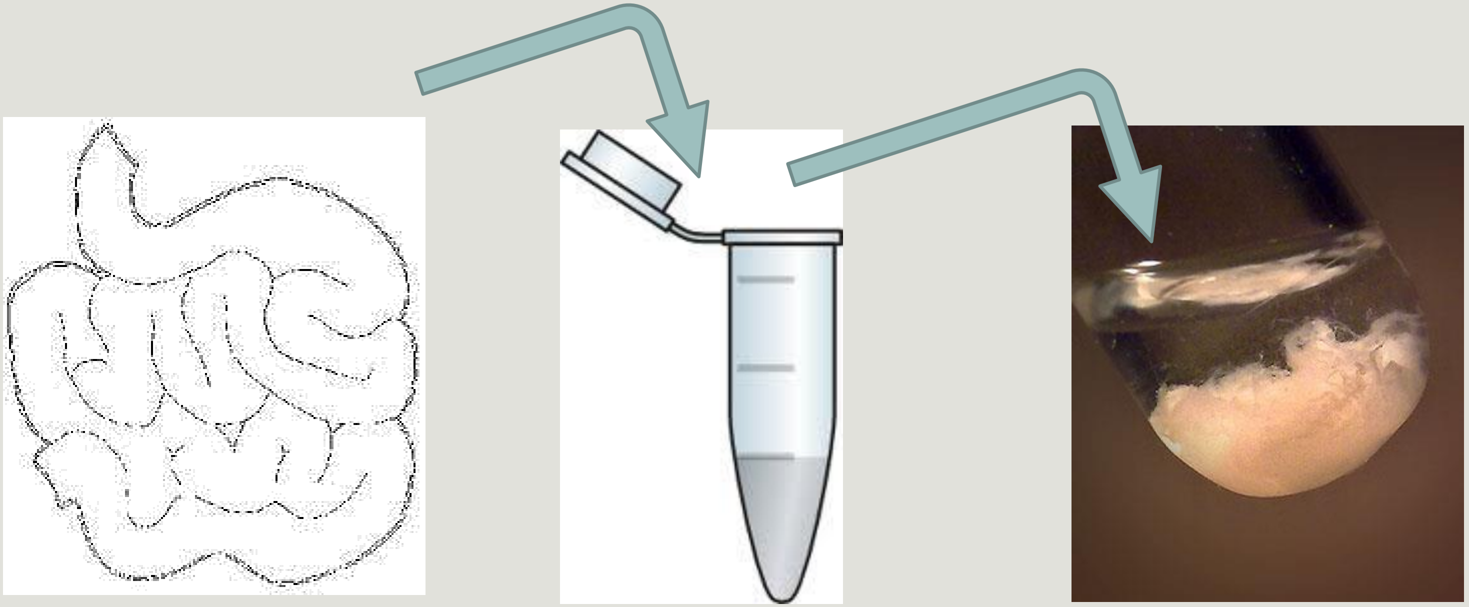
The current processing pipelines struggle to run in a reasonable time.

The most effective solutions are often designed for specialists making access difficult for the whole community.

In this context we developed the pipeline FROGS*: « Find Rapidly OTU with Galaxy Solution ».*

Material

Sample collection and DNA extraction



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

DNA

RNA

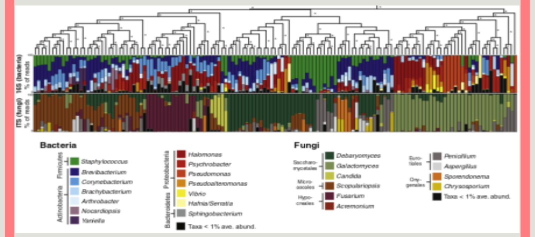
Metagenomics

Metatranscriptomics

Amplicon sequencing

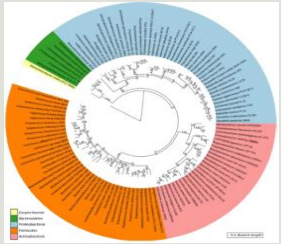
Shotgun sequencing

RNA sequencing



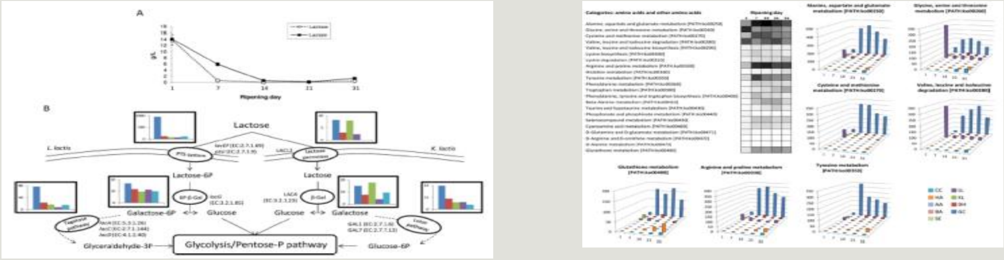
Wolfe *et al.*, 2014

Who is here?



Almeida *et al.*, 2014

What can they do?



Dugat-Bony *et al.*, 2015

What are they doing?

The gene encoding the small subunit of the ribosomal RNA

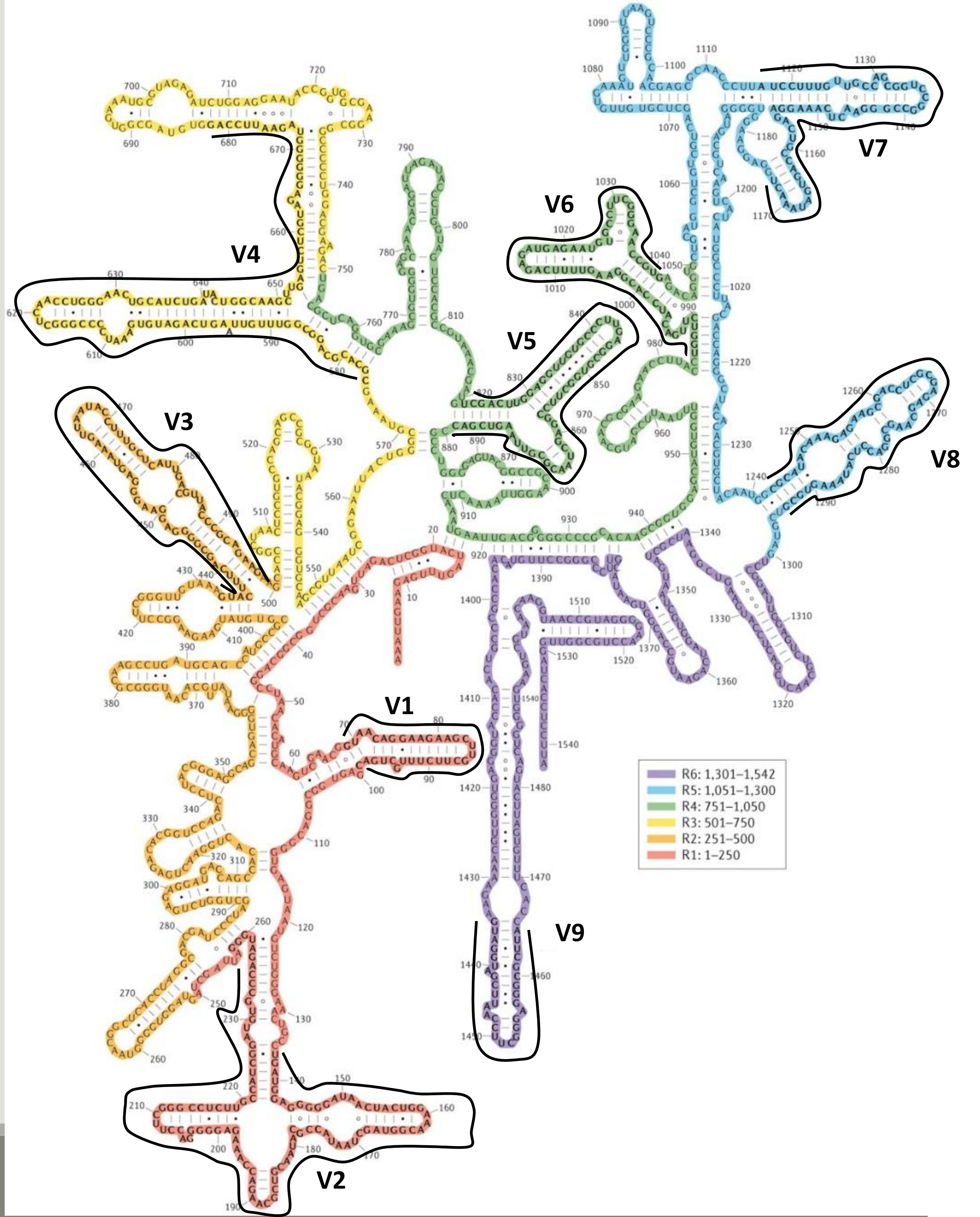
The most widely used gene in **molecular phylogenetic** studies

Ubiquist gene : **16S rDNA** in prokayotes ; **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 2015: >22000 type strains)



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;

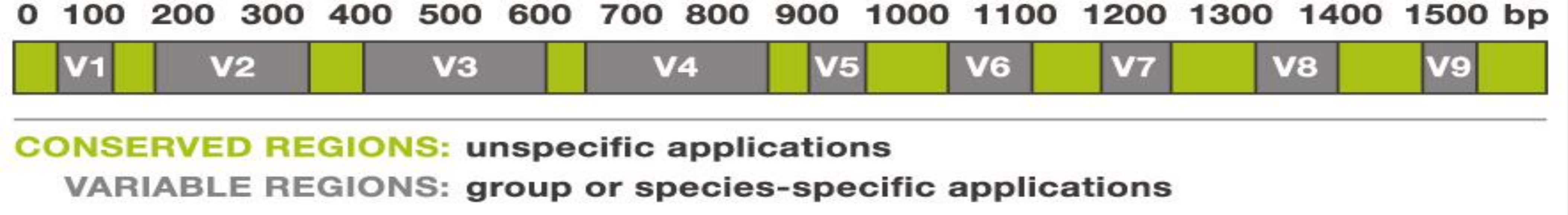
in green, fragment R4 including regions V5 and V6;

in blue, fragment R5 including regions V7 and V8;

and in purple, fragment R6 including region V9.

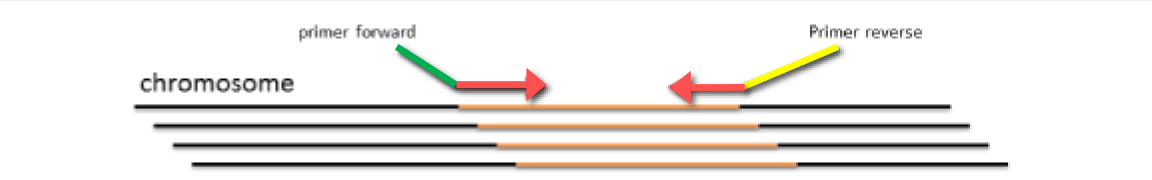
Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences
Pablo Yarza, et al.
Nature Reviews Microbiology 12, 635–645 (2014) doi:10.1038/nrmicro3330

The gene encoding the small subunit of the ribosomal RNA

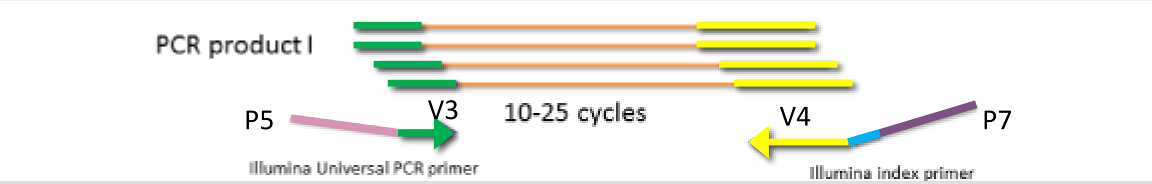

0 100 200 300 400 500 600 700 800 900 1000 1100 1200 1300 1400 1500 bp
V1
V2
V3
V4
V5
V6
V7
V8
V9
CONSERVED REGIONS: unspecific applications
VARIABLE REGIONS: group or species-specific applications

Steps for Illumina sequencing

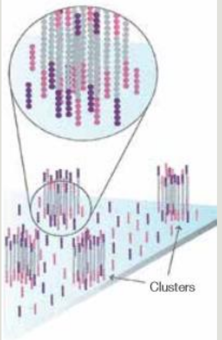
- 1st step : one PCR



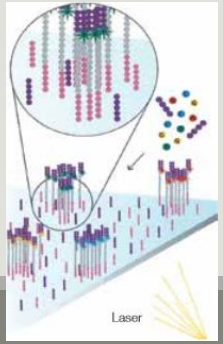
- 2nd step: one PCR



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



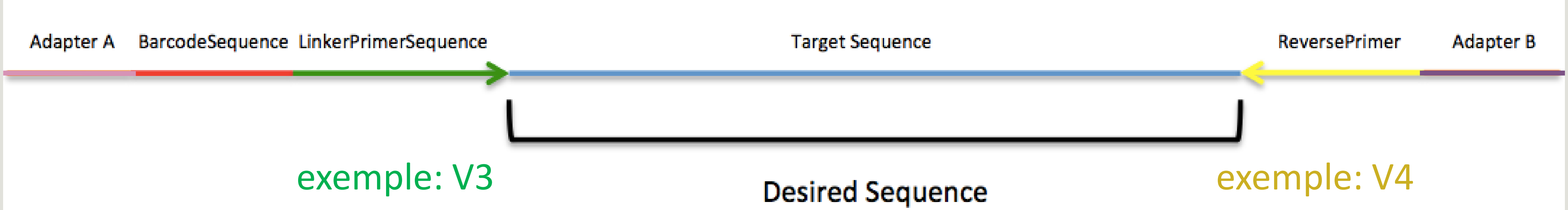
- 4th step: sequencing



Amplification and sequencing

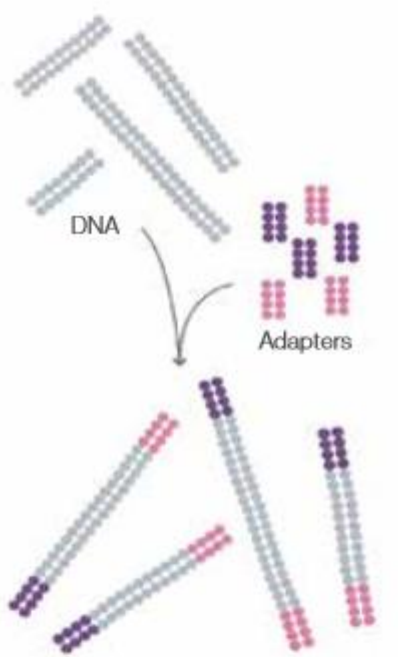
« Universal » primer sets are used for PCR amplification of the phylogenetic biomarker

The primers contain adapters used for the sequencing step and barcodes (= tags = MIDs) to distinguish the samples (multiplexing = sequencing several samples on the same run)



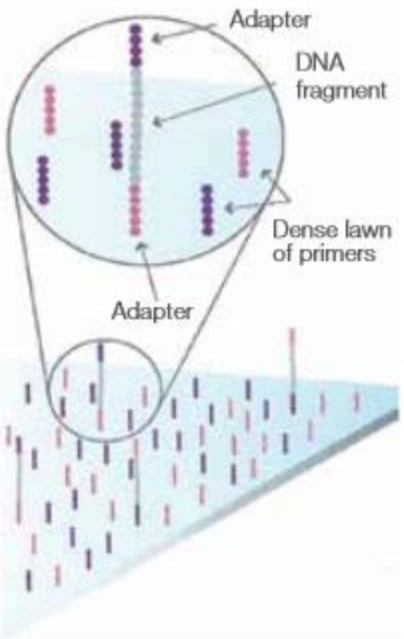
Cluster generation

Prepare Genomic DNA Sample



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

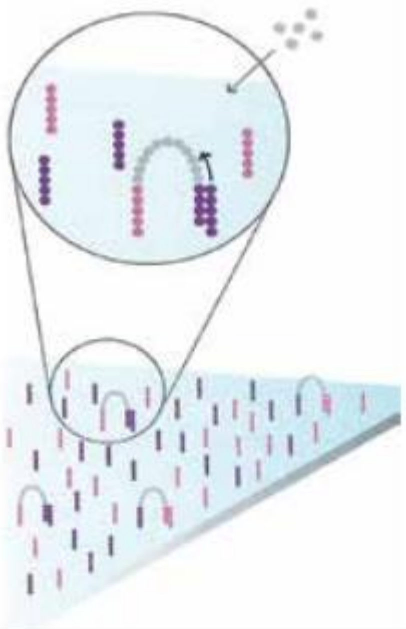
Attach DNA to Surface



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

Attach DNA to surface

Bridge Amplification

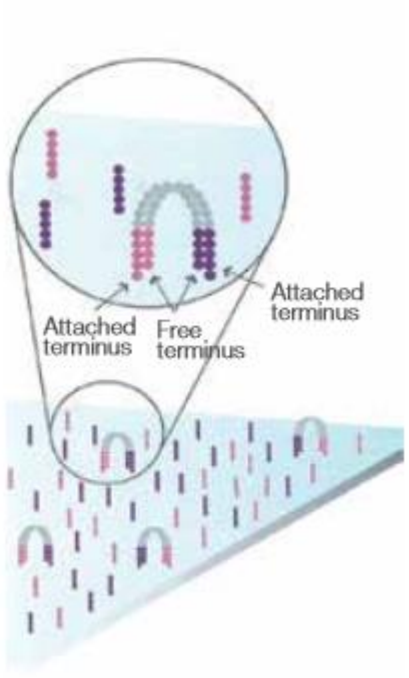


Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

Bridge amplification

Cluster generation

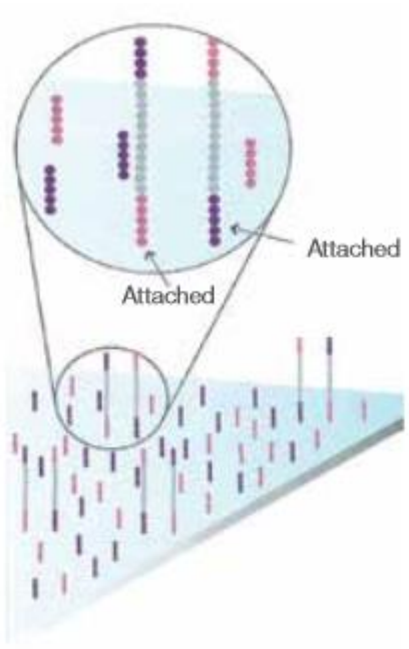
Fragments Become Double Stranded



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

Fragments become double stranded

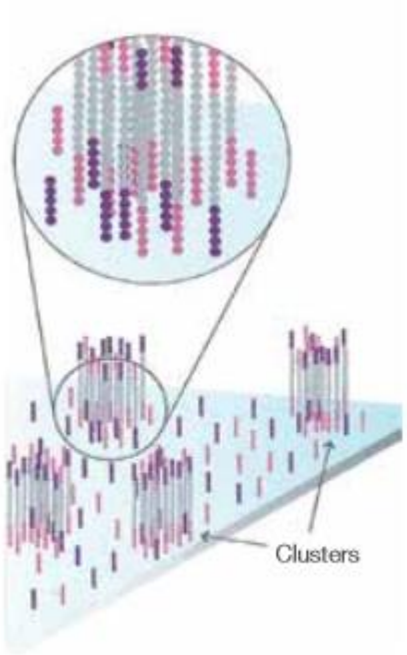
Denature the Double-Stranded Molecules



Denaturation leaves single-stranded templates anchored to the substrate.

Denature the double-stranded molecule

Complete Amplification

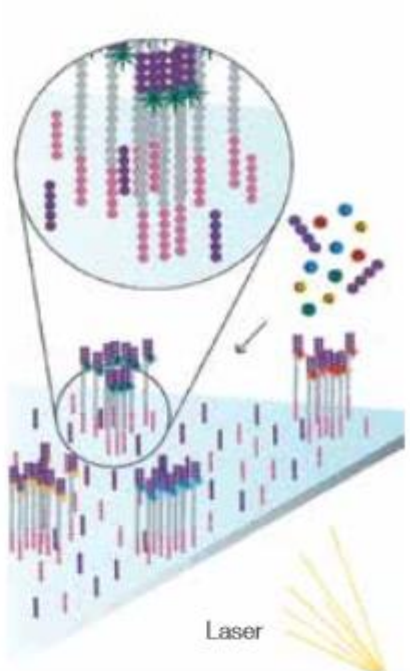


Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

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

Sequencing by synthesis

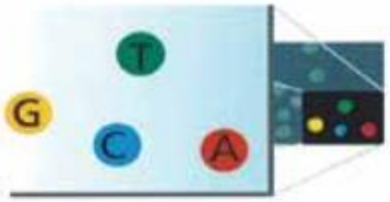
Determine First Base



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

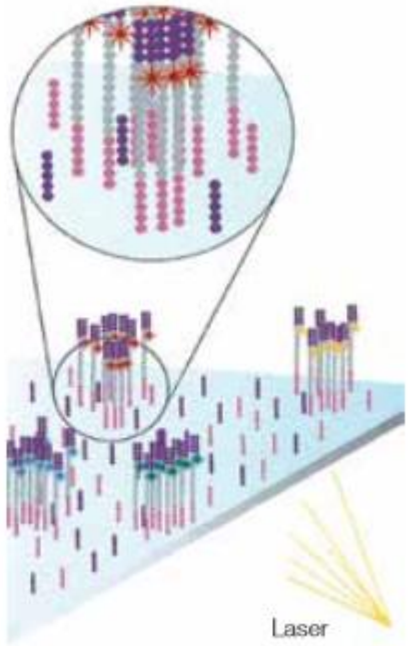
Light signal is more strong in cluster

Image First Base



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

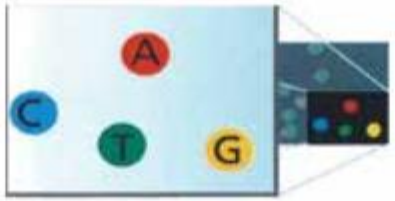
Determine Second Base



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

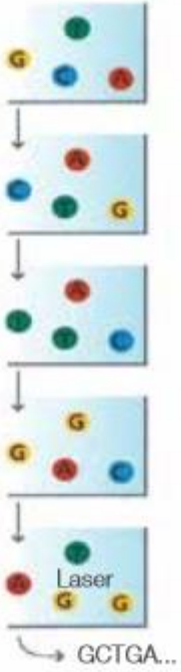
Sequencing by synthesis

Image Second Chemistry Cycle



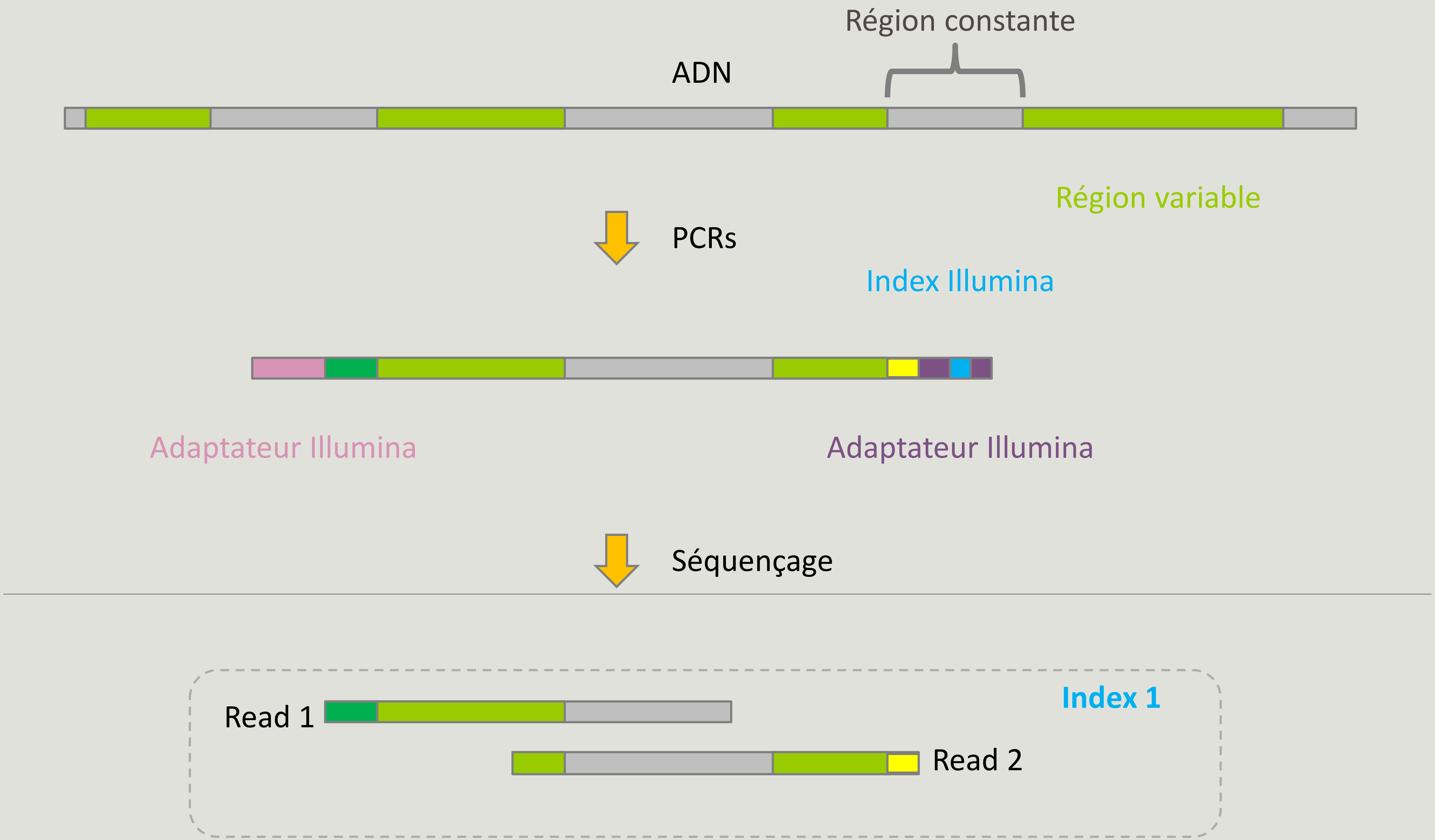
After laser excitation, the image is captured as before, and the identity of the second base is recorded.

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.



Région constante

ADN

Région variable

PCRs

Index Illumina

Adaptateur Illumina

Adaptateur Illumina

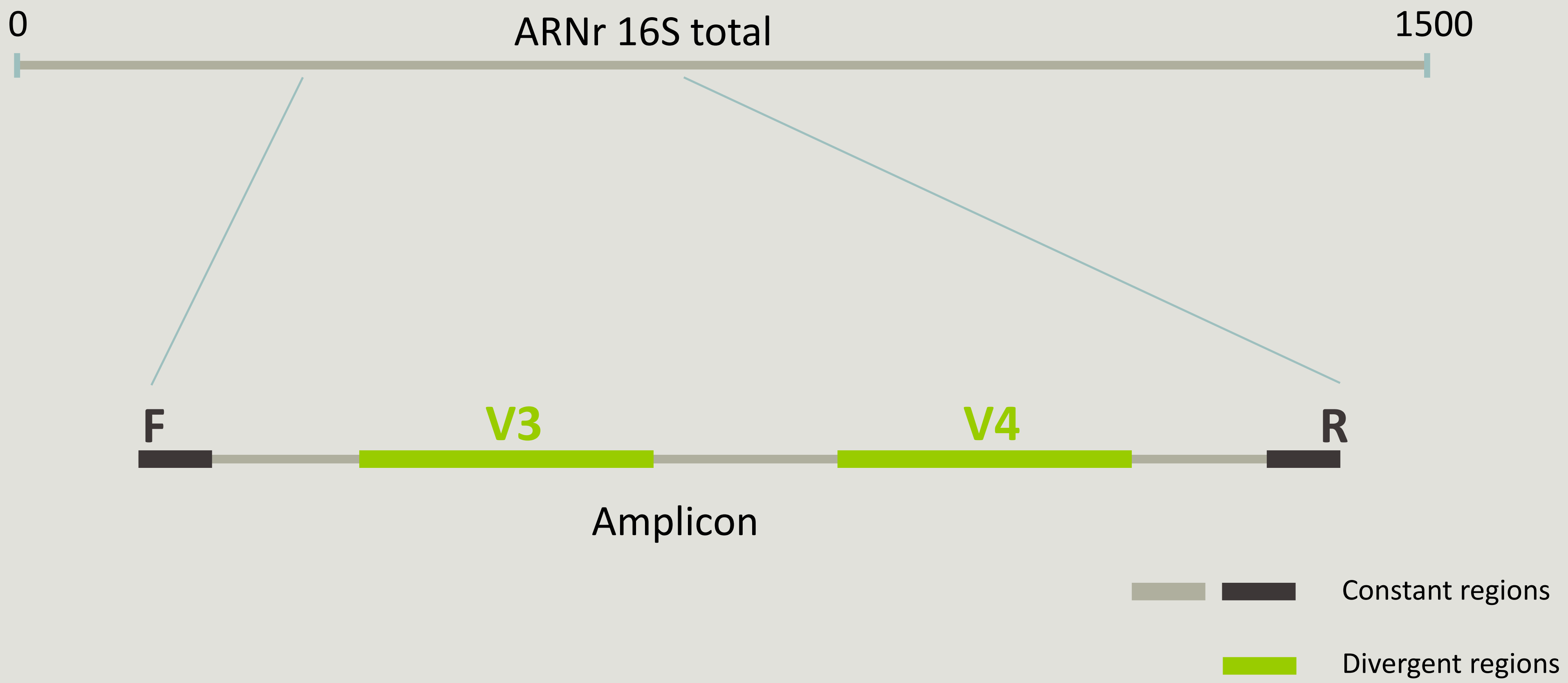
Séquençage

Index 1

Read 1

Read 2

Identification of bacterial populations may be not discriminating



0

ARNr 16S total

1500

F

V3

V4

R

Amplicon

Constant regions

Divergent regions

Amplification and sequencing

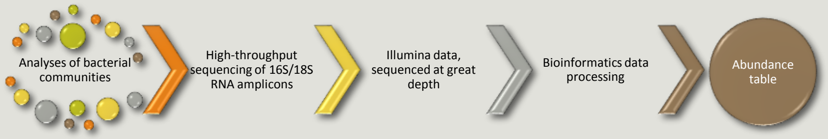
Sequencing is generally perform on **Roche-454** or **Illumina MiSeq** platforms.

Roche-454 generally produce ~ 10 000 reads per sample

MiSeq ~ 30 000 reads per sample

Sequence length is **>650 bp** for pyrosequencing technology (Roche-454) and **2 x 300 bp** for the MiSeq technology in paired-end mode.

Methods


Analyses of bacterial communities
High-throughput sequencing of 16S/18S RNA amplicons
Illumina data, sequenced at great depth
Bioinformatics data processing
Abundance table

Which bioinformatics solutions ?

	Disadvantages
QIIME	Installation problem Command lines
UPARSE	Global clustering command lines
MOTHUR	Not MiSeq data without normalization Global hierarchical clustering Command lines
MG-RAST	No modularity No transparence

QIIME allows analysis of high-throughput community sequencing data
J Gregory Caporaso et al, Nature Methods, 2010; doi:10.1038/nmeth.f.303

Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities.
Schloss, P.D., et al., Appl Environ Microbiol, 2009, doi: 10.1128/AEM.01541-09

UPARSE: Highly accurate OTU sequences from microbial amplicon reads
Edgar, R.C. et al, *Nature Methods*, 2013, dx.doi.org/10.1038/nmeth.2604

The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes
F Meyer et al, BMC Bioinformatics, 2008, doi:10.1186/1471-2105-9-386

FROGS ?

Use platform Galaxy

Set of modules = Tools to analyze your “big” data

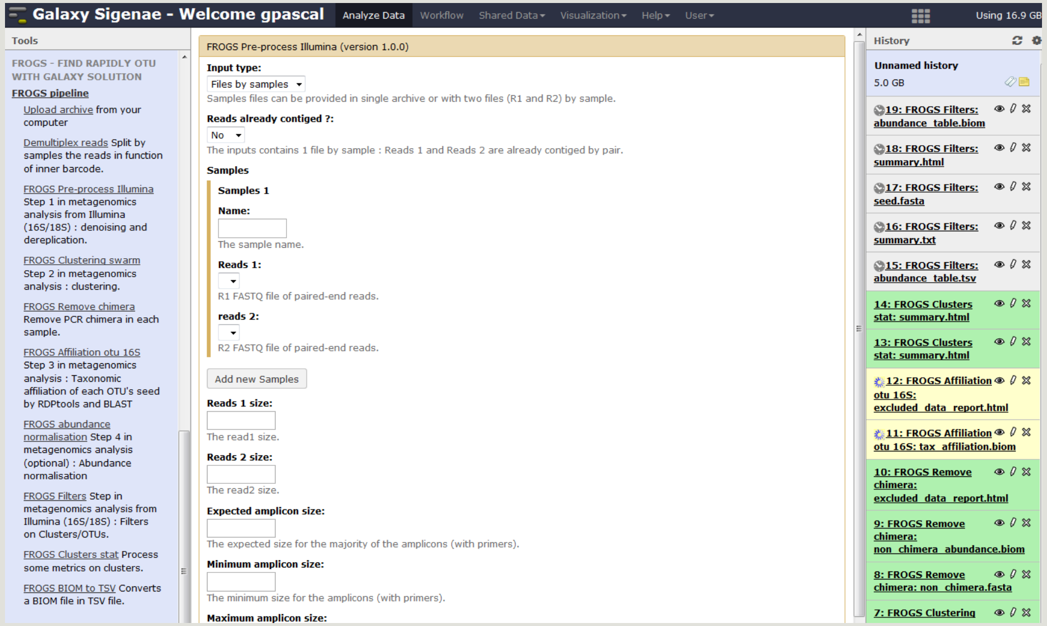
Independent modules

Run on Illumina/454 data 16S, 18S, and 23S

New clustering method

Many graphics for interpretation

User friendly, hiding bioinformatics infrastructure/complexity



FROGS - FIND RAPIDLY OTU WITH GALAXY SOLUTION

FROGS pipeline

Upload archive from your computer

Demultiplex reads Split by samples the reads in function of inner barcode.

FROGS Pre-process Illumina Step 1 in metagenomics analysis from Illumina (16S/18S) : denoising and dereplication.

FROGS Clustering swarm Step 2 in metagenomics analysis : clustering.

FROGS Remove chimera Remove PCR chimera in each sample.

FROGS Affiliation otu 16S Step 3 in metagenomics analysis : Taxonomic affiliation of each OTU's seed by RDPtools and BLAST

FROGS abundance normalisation Step 4 in metagenomics analysis (optional) : Abundance normalisation

FROGS Filters Step in metagenomics analysis from Illumina (16S/18S) : Filters on Clusters/OTUs.

FROGS Clusters stat Process some metrics on clusters.

FROGS BIOM to TSV Converts a BIOM file in TSV file.

FROGS Pre-process Illumina (version 1.0.0)

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

Samples

Samples 1

Name:

The sample name.

Reads 1:

R1 FASTQ file of paired-end reads.

reads 2:

R2 FASTQ file of paired-end reads.

Reads 1 size:

The read1 size.

Reads 2 size:

The read2 size.

Expected amplicon size:

The expected size for the majority of the amplicons (with primers).

Minimum amplicon size:

The minimum size for the amplicons (with primers).

Maximum amplicon size:

Unnamed history

5.0 GB

19: FROGS Filters: abundance_table.biom

18: FROGS Filters: summary.html

17: FROGS Filters: seed.fasta

16: FROGS Filters: summary.txt

15: FROGS Filters: abundance_table.tsv

14: FROGS Clusters stat: summary.html

13: FROGS Clusters stat: summary.html

12: FROGS Affiliation otu 16S: excluded_data_report.html

11: FROGS Affiliation otu 16S: tax_affiliation.biom

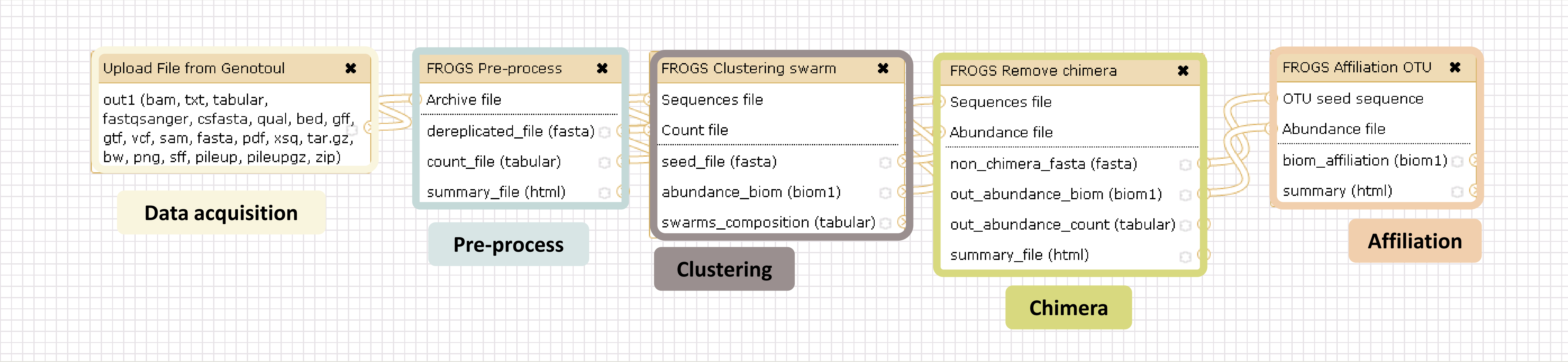
10: FROGS Remove chimera: excluded_data_report.html

9: FROGS Remove chimera: non_chimera_abundance.biom

8: FROGS Remove chimera: non_chimera.fasta

7: FROGS Clustering

FROGS Pipeline



Upload File from Genotoul ✖

out1 (bam, txt, tabular, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png, sff, pileup, pileupgz, zip)

Data acquisition

FROGS Pre-process ✖

Archive file

dereplicated_file (fasta)

count_file (tabular)

summary_file (html)

Pre-process

FROGS Clustering swarm ✖

Sequences file

Count file

seed_file (fasta)

abundance_biom (biom1)

swarms_composition (tabular)

Clustering

FROGS Remove chimera ✖

Sequences file

Abundance file

non_chimera_fasta (fasta)

out_abundance_biom (biom1)

out_abundance_count (tabular)

summary_file (html)

Chimera

FROGS Affiliation OTU ✖

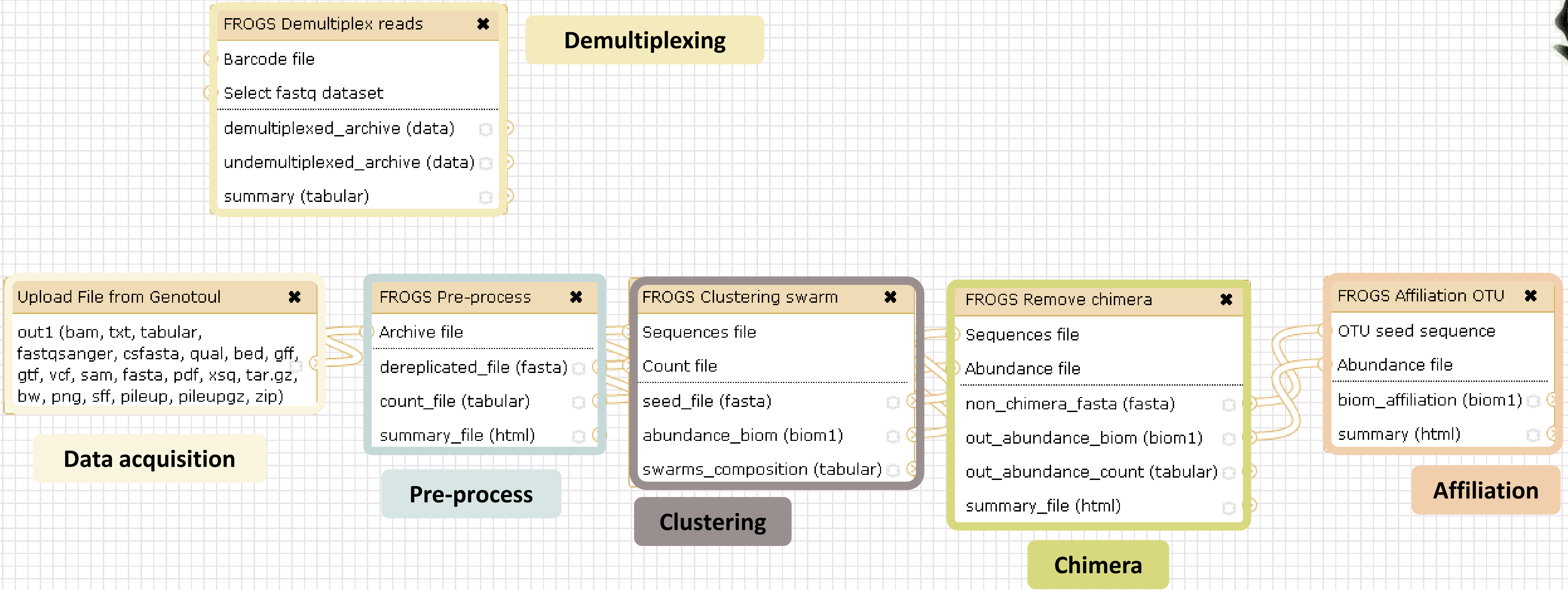
OTU seed sequence

Abundance file

biom_affiliation (biom1)

summary (html)

Affiliation



FROGS Demultiplex reads ✖
Barcode file
Select fastq dataset
demultiplexed_archive (data)
undemultiplexed_archive (data)
summary (tabular)

Demultiplexing

Upload File from Genotoul ✖
out1 (bam, txt, tabular, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png, sff, pileup, pileupgz, zip)

Data acquisition

FROGS Pre-process ✖
Archive file
dereplicated_file (fasta)
count_file (tabular)
summary_file (html)

Pre-process

FROGS Clustering swarm ✖
Sequences file
Count file
seed_file (fasta)
abundance_biom (biom1)
swarms_composition (tabular)

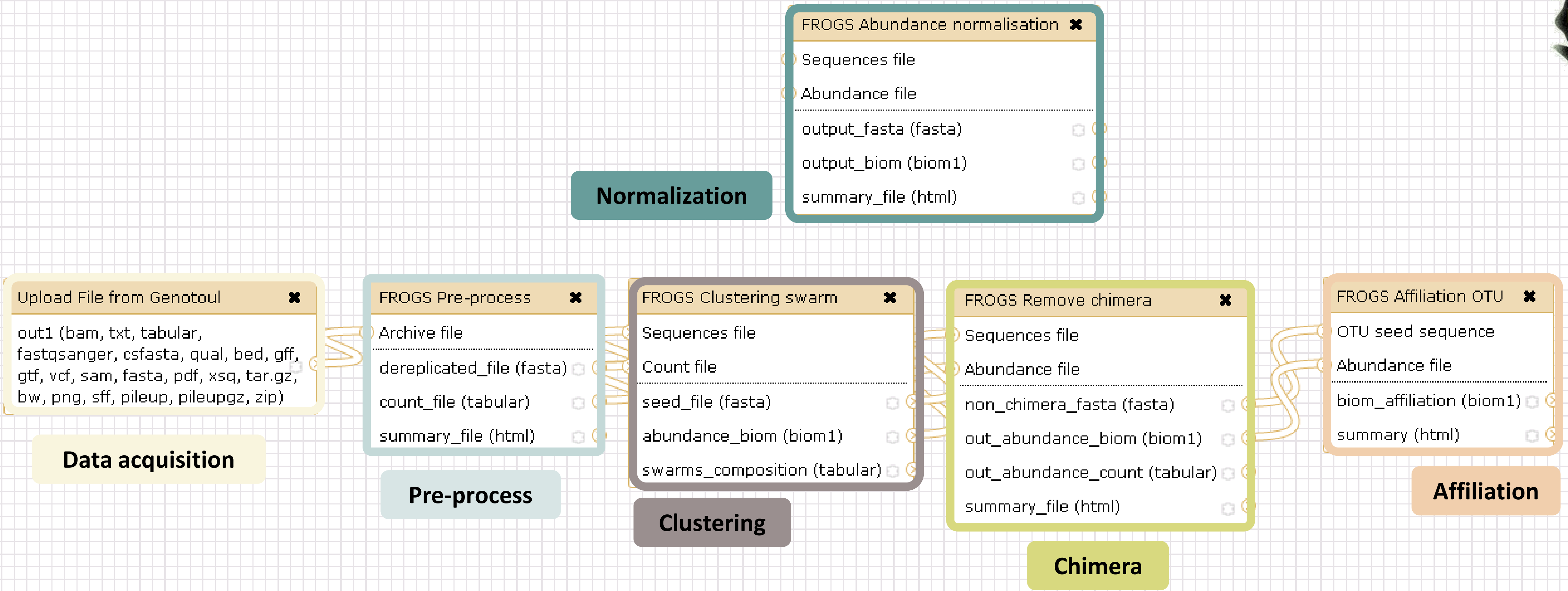
Clustering

FROGS Remove chimera ✖
Sequences file
Abundance file
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
out_abundance_count (tabular)
summary_file (html)

Chimera

FROGS Affiliation OTU ✖
OTU seed sequence
Abundance file
biom_affiliation (biom1)
summary (html)

Affiliation



Normalization

FROGS Abundance normalisation ✖

Sequences file

Abundance file

output_fasta (fasta)

output_biom (biom1)

summary_file (html)

Upload File from Genotoul ✖

out1 (bam, txt, tabular, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png, sff, pileup, pileupgz, zip)

Data acquisition

FROGS Pre-process ✖

Archive file

dereplicated_file (fasta)

count_file (tabular)

summary_file (html)

Pre-process

FROGS Clustering swarm ✖

Sequences file

Count file

seed_file (fasta)

abundance_biom (biom1)

swarms_composition (tabular)

Clustering

FROGS Remove chimera ✖

Sequences file

Abundance file

non_chimera_fasta (fasta)

out_abundance_biom (biom1)

out_abundance_count (tabular)

summary_file (html)

Chimera

FROGS Affiliation OTU ✖

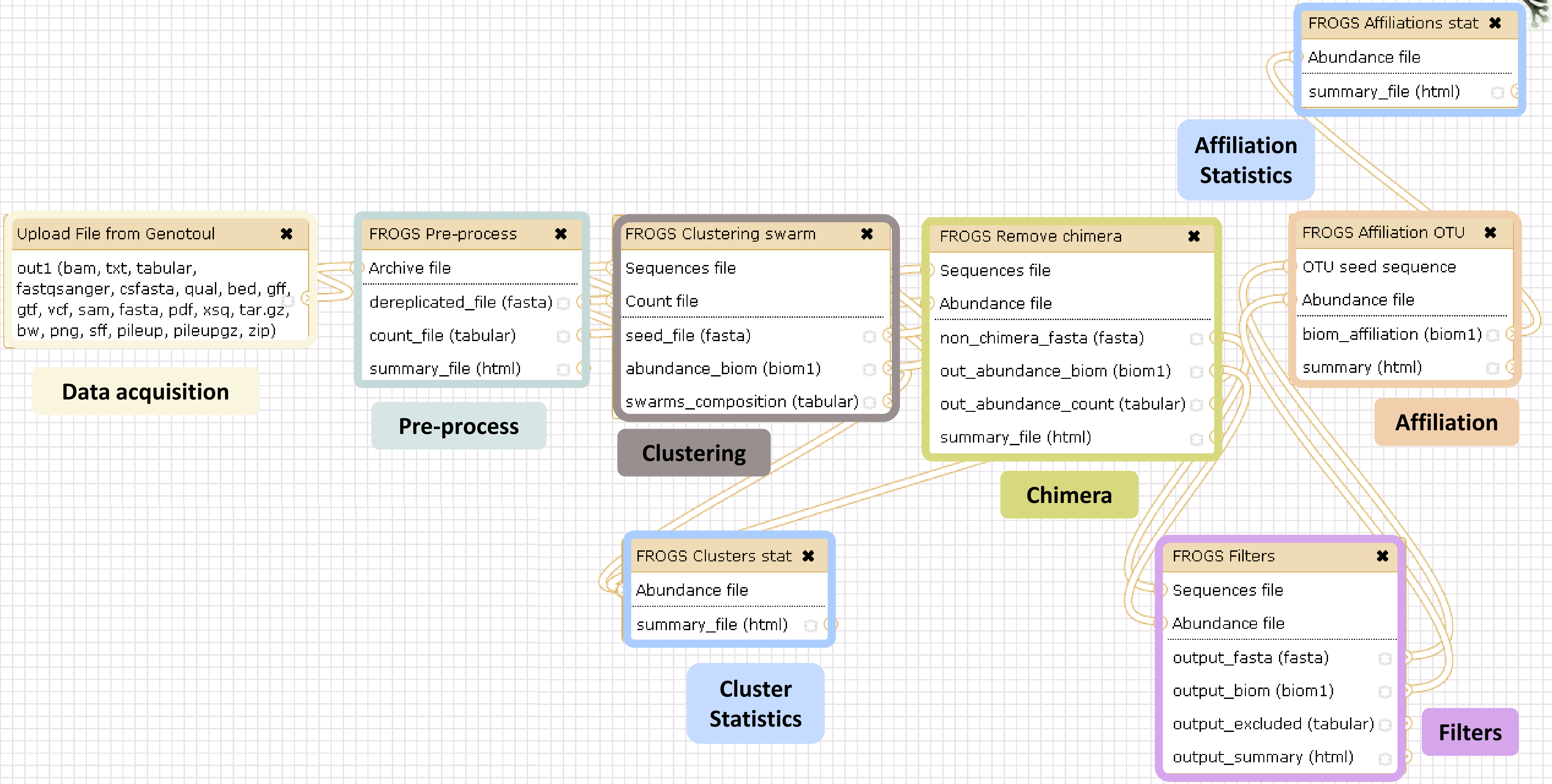
OTU seed sequence

Abundance file

biom_affiliation (biom1)

summary (html)

Affiliation



Data acquisition

Pre-process

Clustering

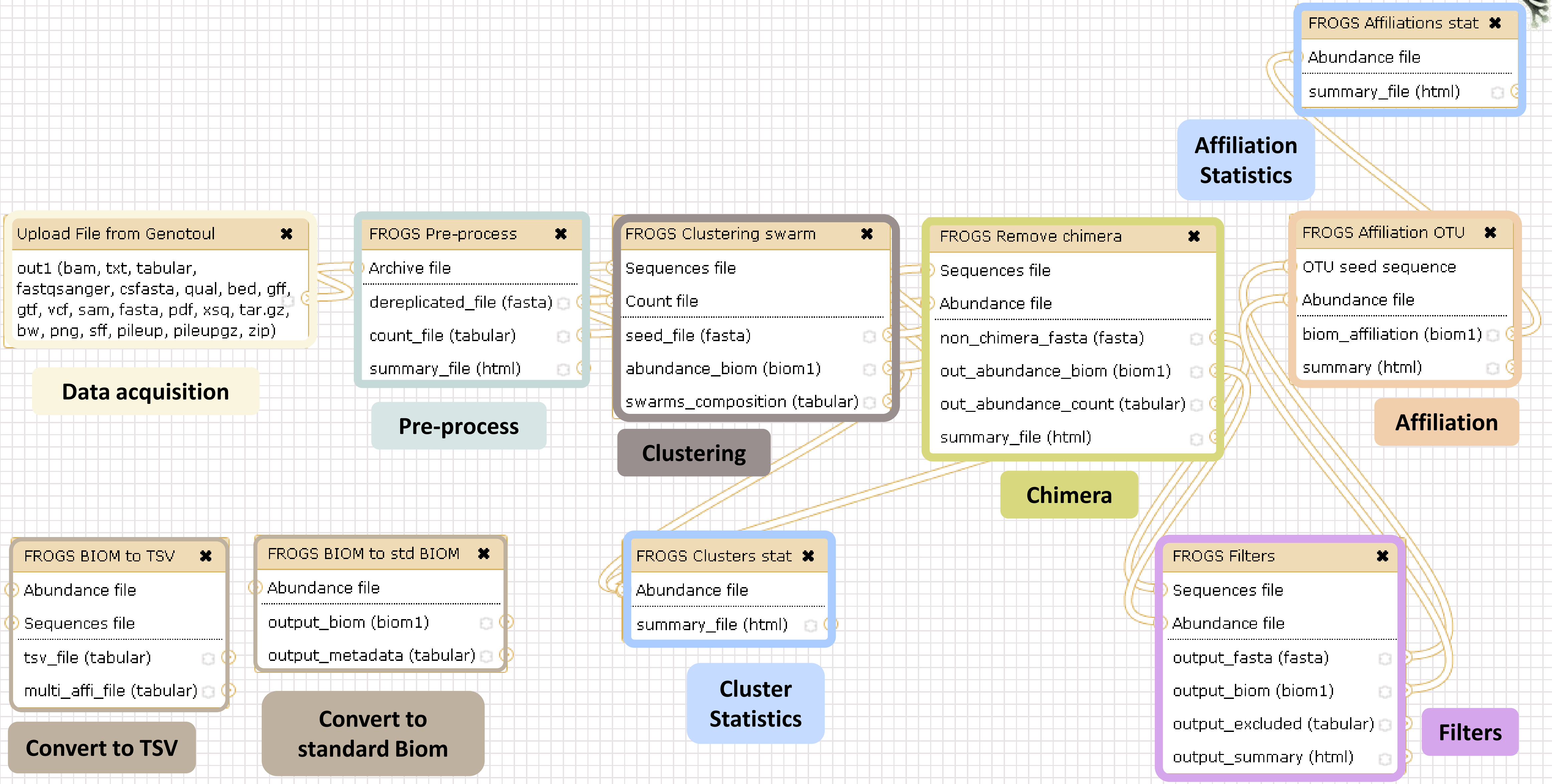
Chimera

Affiliation Statistics

Affiliation

Cluster Statistics

Filters



Upload File from Genotoul ✖
out1 (bam, txt, tabular, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png, sff, pileup, pileupgz, zip)

Data acquisition

FROGS Pre-process ✖
Archive file
dereplicated_file (fasta)
count_file (tabular)
summary_file (html)

Pre-process

FROGS Clustering swarm ✖
Sequences file
Count file
seed_file (fasta)
abundance_biom (biom1)
swarms_composition (tabular)

Clustering

FROGS Remove chimera ✖
Sequences file
Abundance file
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
out_abundance_count (tabular)
summary_file (html)

Chimera

FROGS Affiliation OTU ✖
OTU seed sequence
Abundance file
biom_affiliation (biom1)
summary (html)

Affiliation

FROGS Affiliations stat ✖
Abundance file
summary_file (html)

Affiliation Statistics

FROGS BIOM to TSV ✖
Abundance file
Sequences file
tsv_file (tabular)
multi_affi_file (tabular)

Convert to TSV

FROGS BIOM to std BIOM ✖
Abundance file
output_biom (biom1)
output_metadata (tabular)

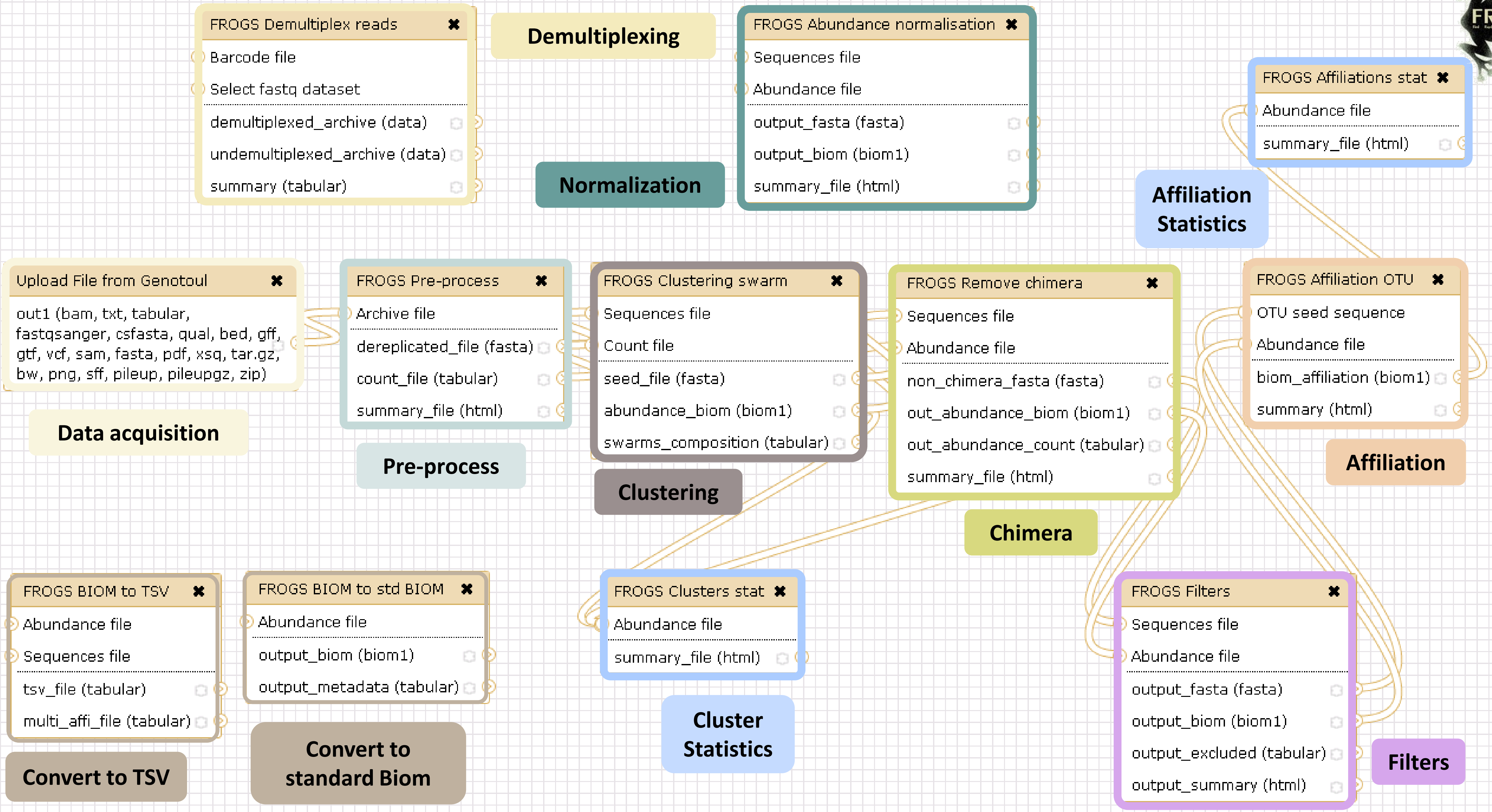
Convert to standard Biom

FROGS Clusters stat ✖
Abundance file
summary_file (html)

Cluster Statistics

FROGS Filters ✖
Sequences file
Abundance file
output_fasta (fasta)
output_biom (biom1)
output_excluded (tabular)
output_summary (html)

Filters



FROGS Demultiplex reads ✖
Barcode file
Select fastq dataset
demultiplexed_archive (data)
undemultiplexed_archive (data)
summary (tabular)

Demultiplexing

Normalization

FROGS Abundance normalisation ✖
Sequences file
Abundance file
output_fasta (fasta)
output_biom (biom1)
summary_file (html)

FROGS Affiliations stat ✖
Abundance file
summary_file (html)

Affiliation Statistics

Upload File from Genotoul ✖
out1 (bam, txt, tabular, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png, sff, pileup, pileupgz, zip)

Data acquisition

FROGS Pre-process ✖
Archive file
dereplicated_file (fasta)
count_file (tabular)
summary_file (html)

Pre-process

FROGS Clustering swarm ✖
Sequences file
Count file
seed_file (fasta)
abundance_biom (biom1)
swarms_composition (tabular)

Clustering

FROGS Remove chimera ✖
Sequences file
Abundance file
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
out_abundance_count (tabular)
summary_file (html)

Chimera

FROGS Affiliation OTU ✖
OTU seed sequence
Abundance file
biom_affiliation (biom1)
summary (html)

Affiliation

FROGS BIOM to TSV ✖
Abundance file
Sequences file
tsv_file (tabular)
multi_affi_file (tabular)

Convert to TSV

FROGS BIOM to std BIOM ✖
Abundance file
output_biom (biom1)
output_metadata (tabular)

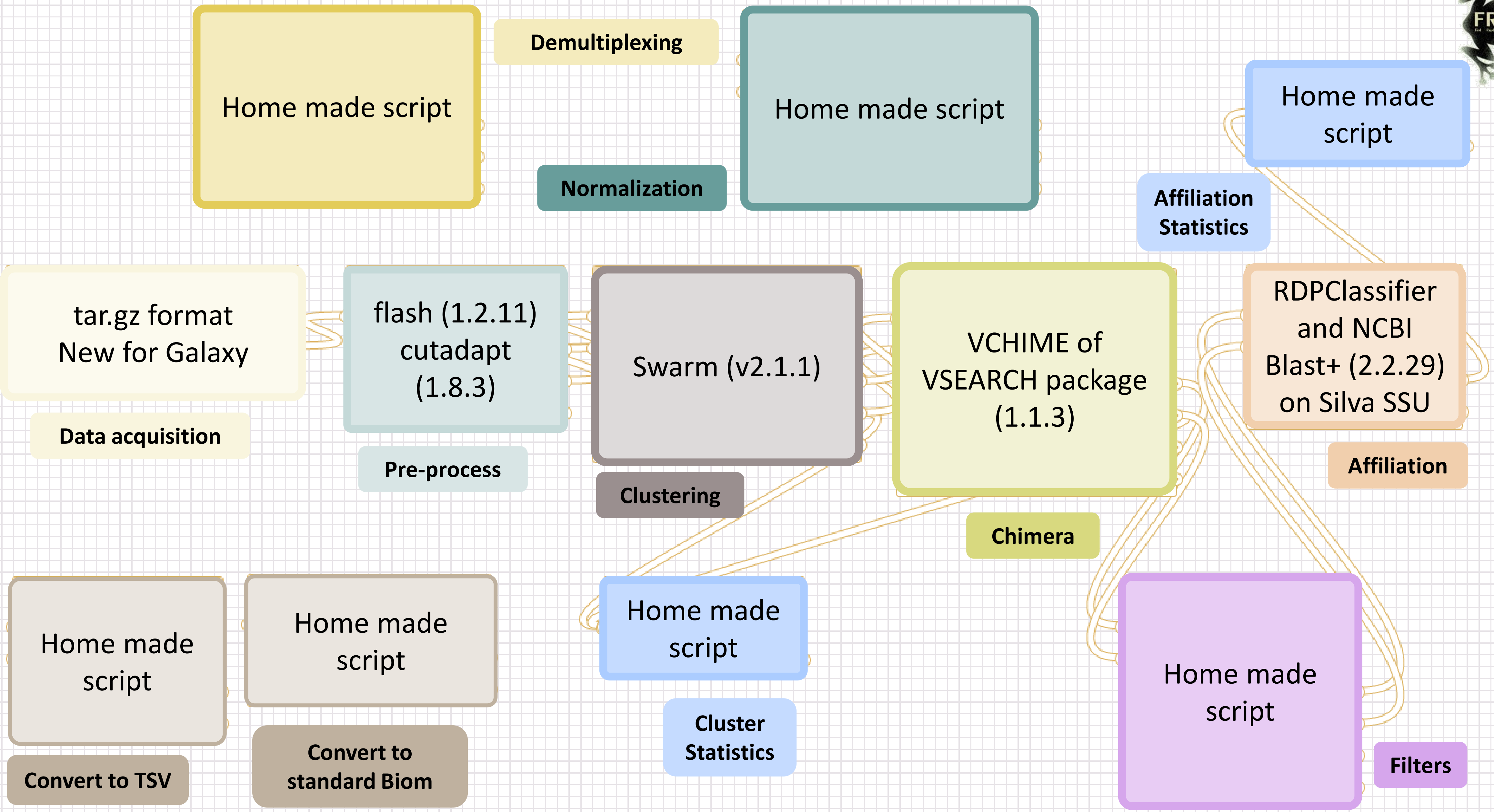
Convert to standard Biom

FROGS Clusters stat ✖
Abundance file
summary_file (html)

Cluster Statistics

FROGS Filters ✖
Sequences file
Abundance file
output_fasta (fasta)
output_biom (biom1)
output_excluded (tabular)
output_summary (html)

Filters



Home made script

Demultiplexing

Normalization

Home made script

Home made script

Affiliation Statistics

tar.gz format
New for Galaxy

Data acquisition

flash (1.2.11)
cutadapt
(1.8.3)

Pre-process

Swarm (v2.1.1)

Clustering

VCHIME of
VSEARCH package
(1.1.3)

Chimera

RDPClassifier
and NCBI
Blast+ (2.2.29)
on Silva SSU

Affiliation

Home made script

Convert to TSV

Home made script

Convert to standard Biom

Home made script

Cluster Statistics

Home made script

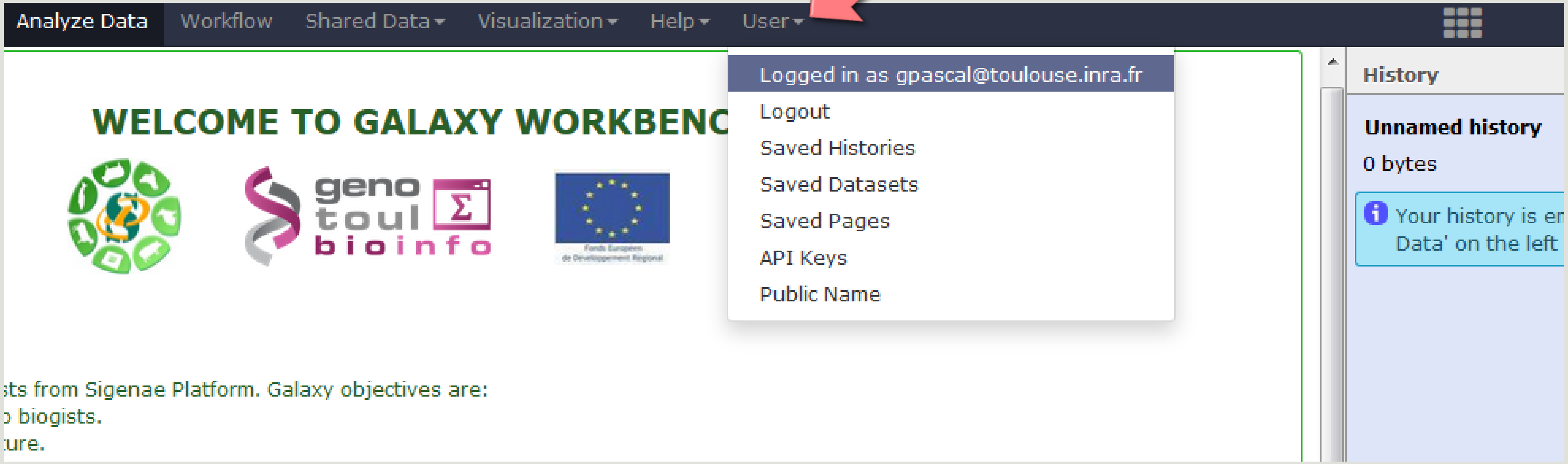
Filters

Together go to visit FROGS

In your internet browser (Firefox, chrome, Internet explorer) :

http://sigenae-workbench.toulouse.inra.fr/

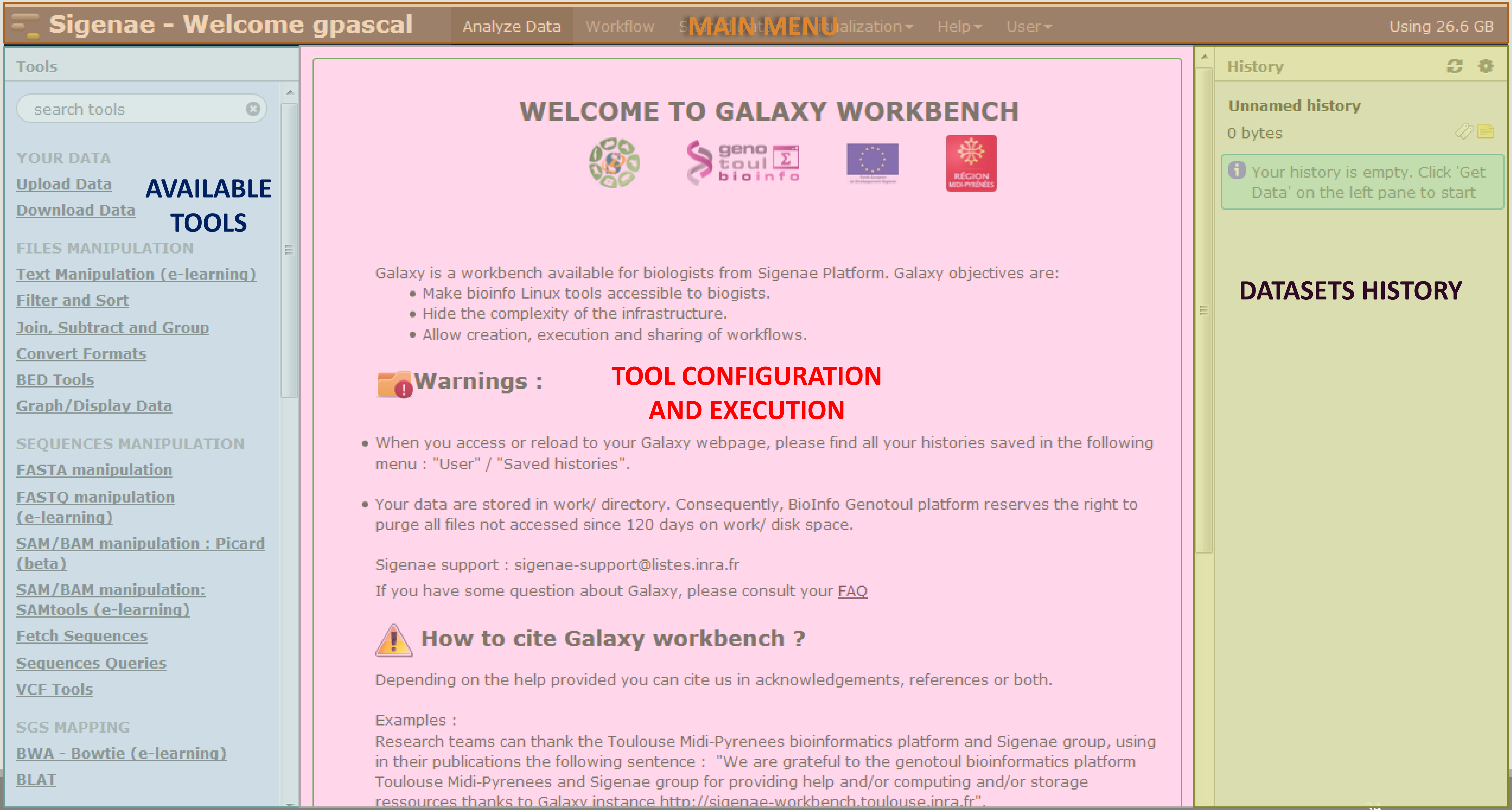
Enter your login and password from GenoToul



WELCOME TO GALAXY WORKBENC

Your history is er Data' on the left

ıts from Sigenae Platform. Galaxy objectives are:
ɔ biogists.
ture.



YOUR DATA

- Upload Data
- Download Data

FILES MANIPULATION

- Text Manipulation (e-learning)
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- BED Tools
- Graph/Display Data

SEQUENCES MANIPULATION

- FASTA manipulation
- FASTQ manipulation (e-learning)
- SAM/BAM manipulation : Picard (beta)
- SAM/BAM manipulation: SAMtools (e-learning)
- Fetch Sequences
- Sequences Queries
- VCF Tools

SGS MAPPING

- BWA - Bowtie (e-learning)
- BLAT

WELCOME TO GALAXY WORKBENCH

Galaxy is a workbench available for biologists from Sigenae Platform. Galaxy objectives are:

- Make bioinfo Linux tools accessible to biogists.
- Hide the complexity of the infrastructure.
- Allow creation, execution and sharing of workflows.

Warnings :

- When you access or reload to your Galaxy webpage, please find all your histories saved in the following menu : "User" / "Saved histories".
- Your data are stored in work/ directory. Consequently, BioInfo Genotoul platform reserves the right to purge all files not accessed since 120 days on work/ disk space.

Sigenae support : sigenae-support@listes.inra.fr

If you have some question about Galaxy, please consult your FAQ

How to cite Galaxy workbench ?

Depending on the help provided you can cite us in acknowledgements, references or both.

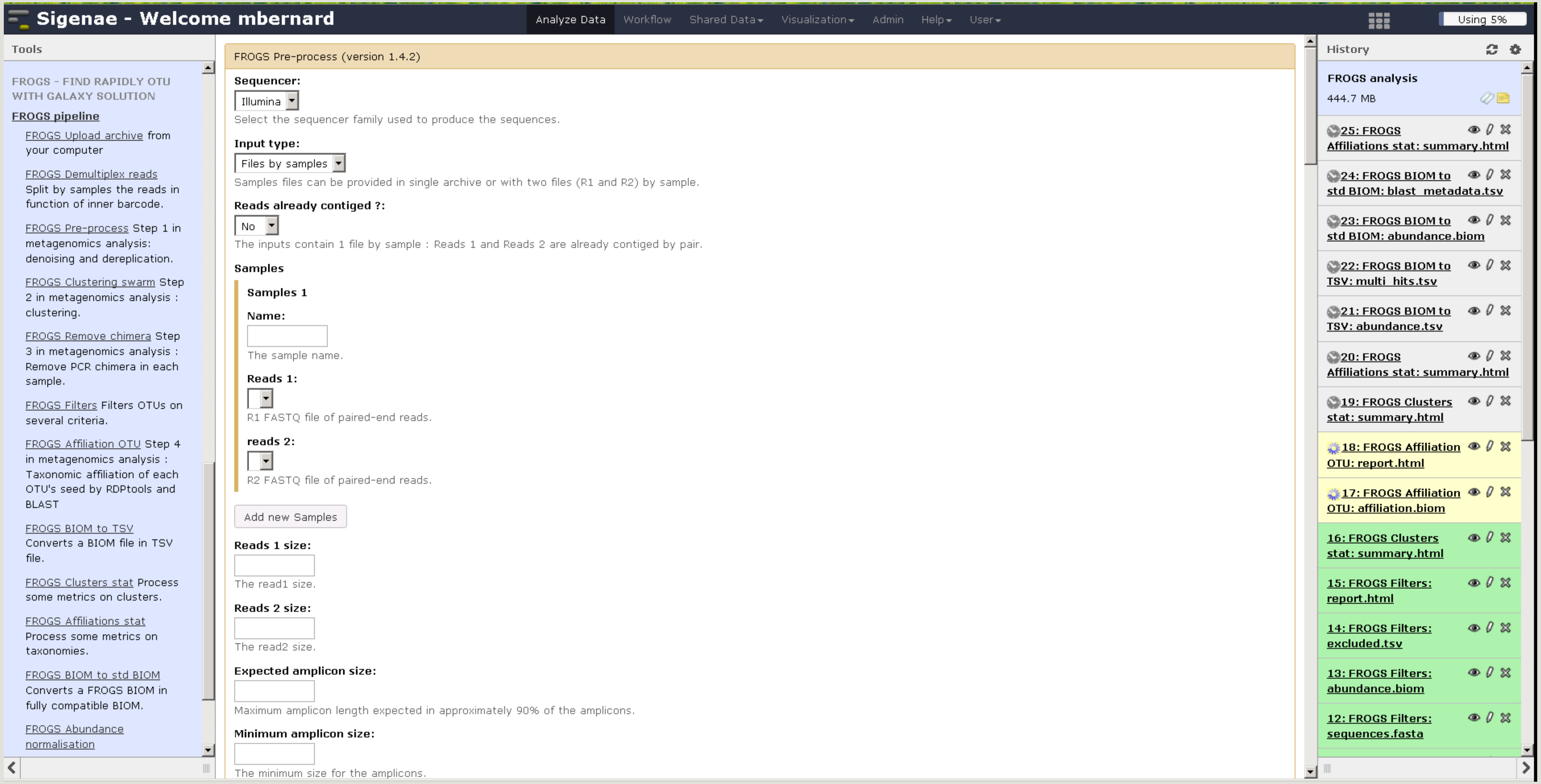
Examples :

Research teams can thank the Toulouse Midi-Pyrenees bioinformatics platform and Sigenae group, using in their publications the following sentence : "We are grateful to the genotoul bioinformatics platform Toulouse Midi-Pyrenees and Sigenae group for providing help and/or computing and/or storage ressources thanks to Galaxy instance http://sigenae-workbench.toulouse.inra.fr".

Unnamed history

0 bytes

Your history is empty. Click 'Get Data' on the left pane to start



Tools

FROGS - FIND RAPIDLY OTU WITH GALAXY SOLUTION

FROGS pipeline

- FROGS Upload archive from your computer
- FROGS Demultiplex reads Split by samples the reads in function of inner barcode.
- FROGS Pre-process Step 1 in metagenomics analysis: denoising and dereplication.
- FROGS Clustering swarm Step 2 in metagenomics analysis : clustering.
- FROGS Remove chimera Step 3 in metagenomics analysis : Remove PCR chimera in each sample.
- FROGS Filters Filters OTUs on several criteria.
- FROGS Affiliation OTU Step 4 in metagenomics analysis : Taxonomic affiliation of each OTU's seed by RDPtools and BLAST
- FROGS BIOM to TSV Converts a BIOM file in TSV file.
- 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 Abundance normalisation

FROGS Pre-process (version 1.4.2)

Sequencer: Illumina

Select the sequencer family 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 : Reads 1 and Reads 2 are already contiged by pair.

Samples

Samples 1

Name:

The sample name.

Reads 1:

R1 FASTQ file of paired-end reads.

reads 2:

R2 FASTQ file of paired-end reads.

Add new Samples

Reads 1 size:

The read1 size.

Reads 2 size:

The read2 size.

Expected amplicon size:

Maximum amplicon length expected in approximately 90% of the amplicons.

Minimum amplicon size:

The minimum size for the amplicons.

History

FROGS analysis

444.7 MB

- 25: FROGS Affiliations stat: summary.html
- 24: FROGS BIOM to std BIOM: blast_metadata.tsv
- 23: FROGS BIOM to std BIOM: abundance.biom
- 22: FROGS BIOM to TSV: multi_hits.tsv
- 21: FROGS BIOM to TSV: abundance.tsv
- 20: FROGS Affiliations stat: summary.html
- 19: FROGS Clusters stat: summary.html
- 18: FROGS Affiliation OTU: report.html
- 17: FROGS Affiliation OTU: affiliation.biom
- 16: FROGS Clusters stat: summary.html
- 15: FROGS Filters: report.html
- 14: FROGS Filters: excluded.tsv
- 13: FROGS Filters: abundance.biom
- 12: FROGS Filters: sequences.fasta

Data acquisition

Demultiplexing

Pre-process

Clustering

Chimera

Filters

Affiliation

Biom to TSV

Cluster Stat

Affiliation Stat

Biom to std Biom

Normalization

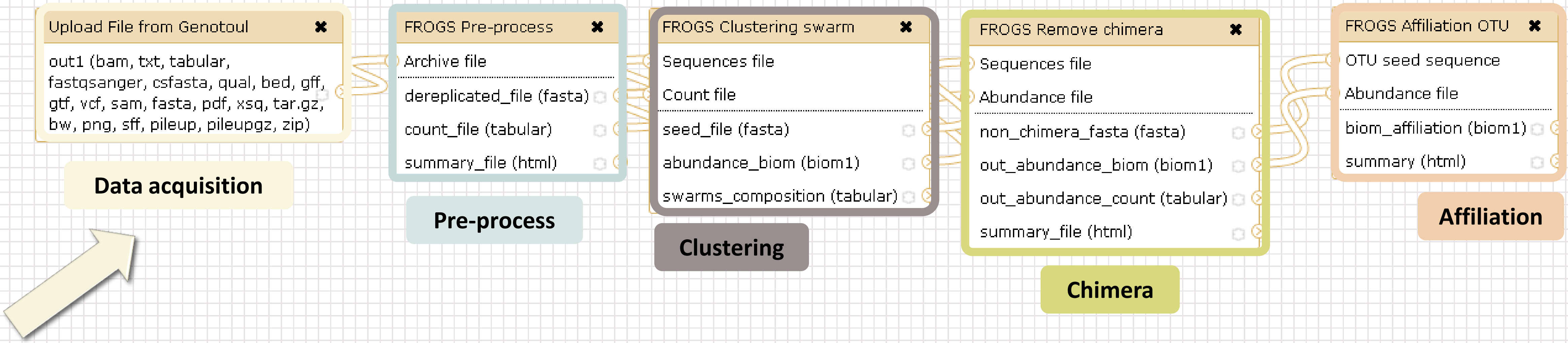
Waiting to run

Currently running

Result files

Upload data

Go to demultiplexing tool



Upload File from Genotoul ✖

out1 (bam, txt, tabular, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png, sff, pileup, pileupgz, zip)

Data acquisition

FROGS Pre-process ✖

Archive file

dereplicated_file (fasta)

count_file (tabular)

summary_file (html)

Pre-process

FROGS Clustering swarm ✖

Sequences file

Count file

seed_file (fasta)

abundance_biom (biom1)

swarms_composition (tabular)

Clustering

FROGS Remove chimera ✖

Sequences file

Abundance file

non_chimera_fasta (fasta)

out_abundance_biom (biom1)

out_abundance_count (tabular)

summary_file (html)

Chimera

FROGS Affiliation OTU ✖

OTU seed sequence

Abundance file

biom_affiliation (biom1)

summary (html)

Affiliation

What kind of data ?

4 Upload → 4 Histories

Multiplexed data

Pathobiomes
rodents and ticks

multiplex.fastq

barcode.tabular

454 data

Freshwater sediment
metagenome

454.fastq.gz

SRA number

- SRR443364

MiSeq
R1 fastq + R2 fastq

Farm animal feces
metagenome

sampleA_R1.fastq

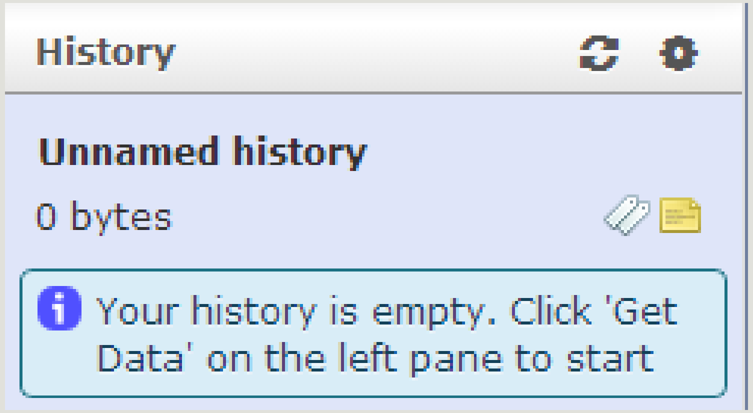
sampleA_R2.fastq

MiSeq contiged fastq
in archive tar.gz

Farm animal feces
metagenome

100spec_90000seq_9samples.tar.gz

1ST CONNEXION



History

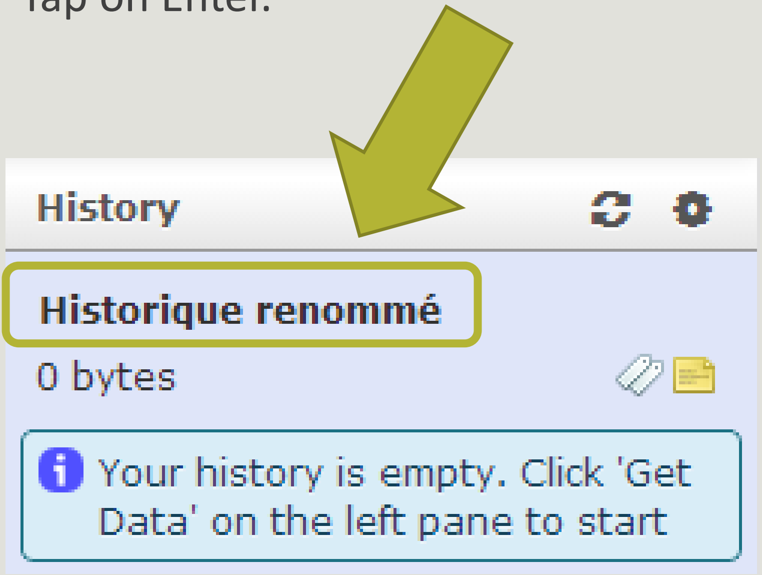
Unnamed history

0 bytes

Your history is empty. Click 'Get Data' on the left pane to start

RENAME HISTORY

- click on Unnamed history,
- Write your new name,
- Tap on Enter.



History

Historique renommé

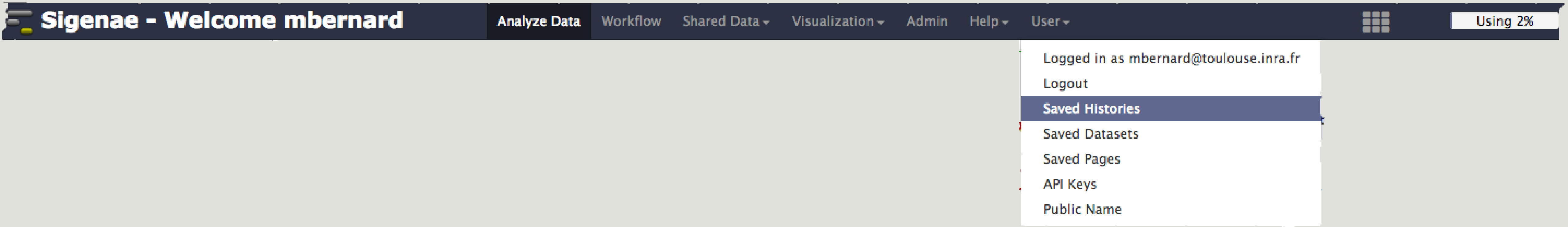
0 bytes

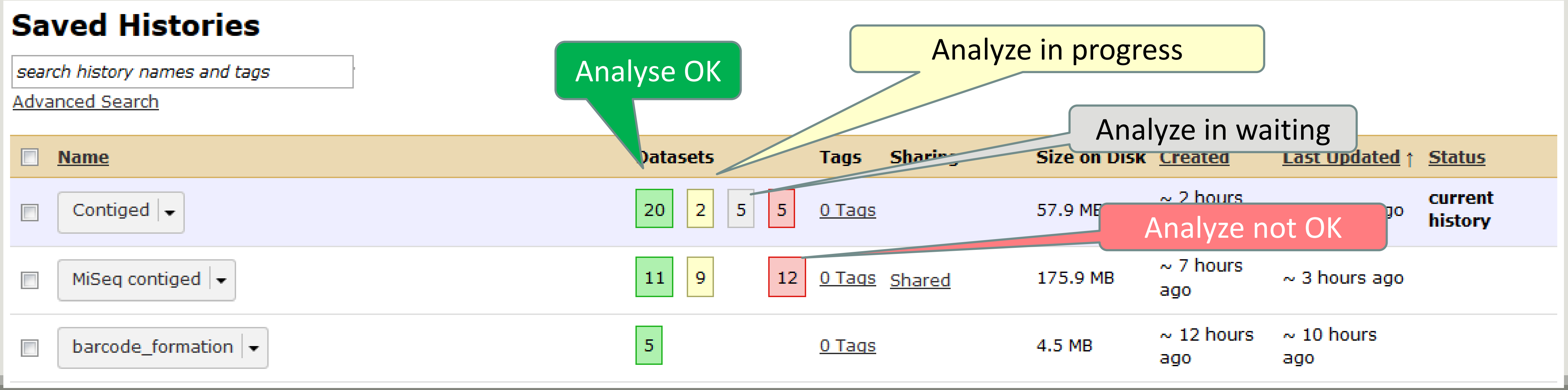
Your history is empty. Click 'Get Data' on the left pane to start

History gestion

- Keep all steps of your analysis.
- Share your analyzes.
- At each run of a tool, a new dataset is created. The data are not overwritten.
- Repeat, as many times as necessary, an analysis.
- All your logs are automatically saved.
- Your published histories are accessible to all users connected to Galaxy (Shared Data / Published Histories).
- Shared histories are accessible only to a specific user (History / Option / Histories Shared With Me).
- To share or publish a history: User / Saved histories / Click the history name / Share or Publish

Saved Histories





Saved Histories

search history names and tags

Advanced Search

☐	Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated ↑	Status
☐	Contiged ▾	20 2 5 5	0 Tags		57.9 MB	~ 2 hours	go	current history
☐	MiSeq contiged ▾	11 9 12	0 Tags	Shared	175.9 MB	~ 7 hours ago	~ 3 hours ago	
☐	barcode_formation ▾	5	0 Tags		4.5 MB	~ 12 hours ago	~ 10 hours ago	

Analyse OK

Analyze in progress

Analyze in waiting

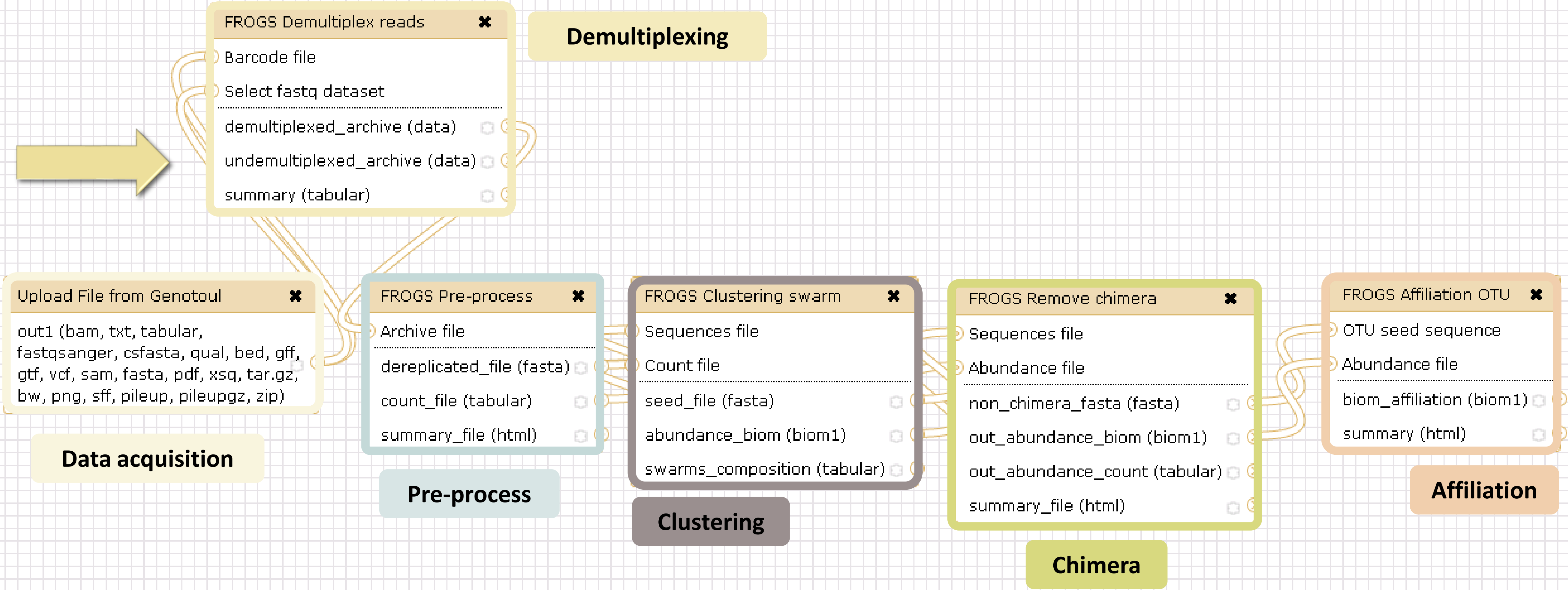
Analyze not OK

Your turn! - 1

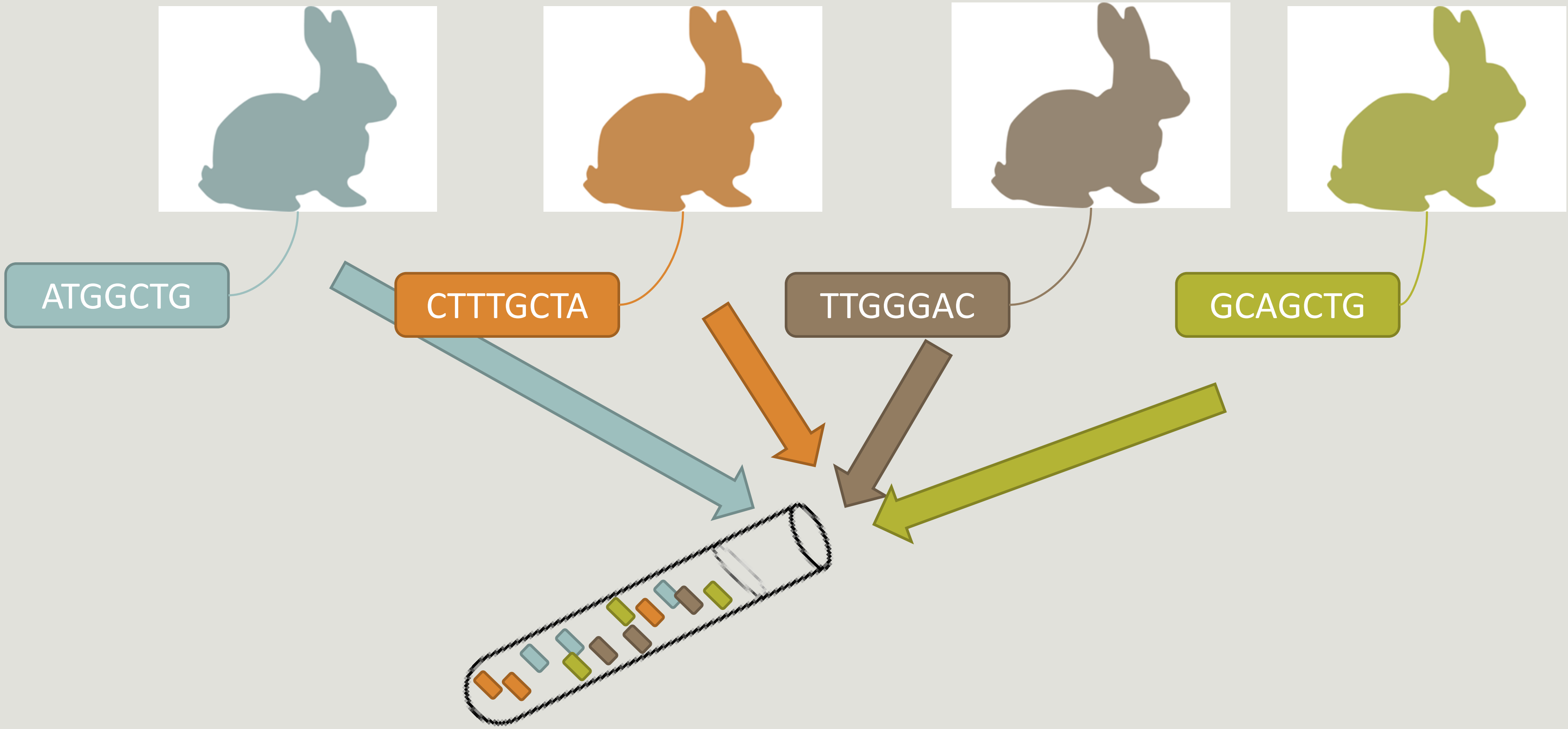
Go to practice

SEE EXERCISE 1

Demultiplexing tool


FROGS Demultiplex reads
Barcode file
Select fastq dataset
demultiplexed_archive (data)
undemultiplexed_archive (data)
summary (tabular)
Demultiplexing
Upload File from Genotoul
out1 (bam, txt, tabular, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png, sff, pileup, pileupgz, zip)
Data acquisition
FROGS Pre-process
Archive file
dereplicated_file (fasta)
count_file (tabular)
summary_file (html)
Pre-process
FROGS Clustering swarm
Sequences file
Count file
seed_file (fasta)
abundance_biom (biom1)
swarms_composition (tabular)
Clustering
FROGS Remove chimera
Sequences file
Abundance file
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
out_abundance_count (tabular)
summary_file (html)
Chimera
FROGS Affiliation OTU
OTU seed sequence
Abundance file
biom_affiliation (biom1)
summary (html)
Affiliation

Barcoding ?



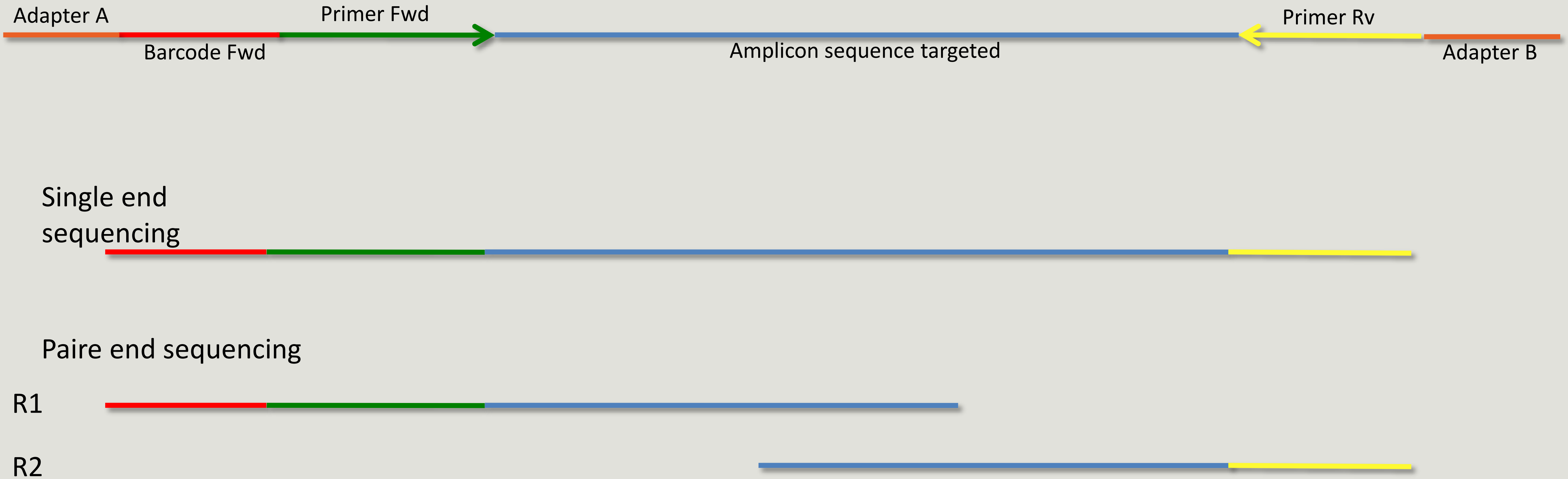
Demultiplexing

Sequence demultiplexing in function of barcode sequences :

- In forward
- In reverse
- In forward and reverse

Remove unbarcoded or ambiguous sequences

Demultiplexing forward



Adapter A

Primer Fwd

Primer Rv

Barcode Fwd

Amplicon sequence targeted

Adapter B

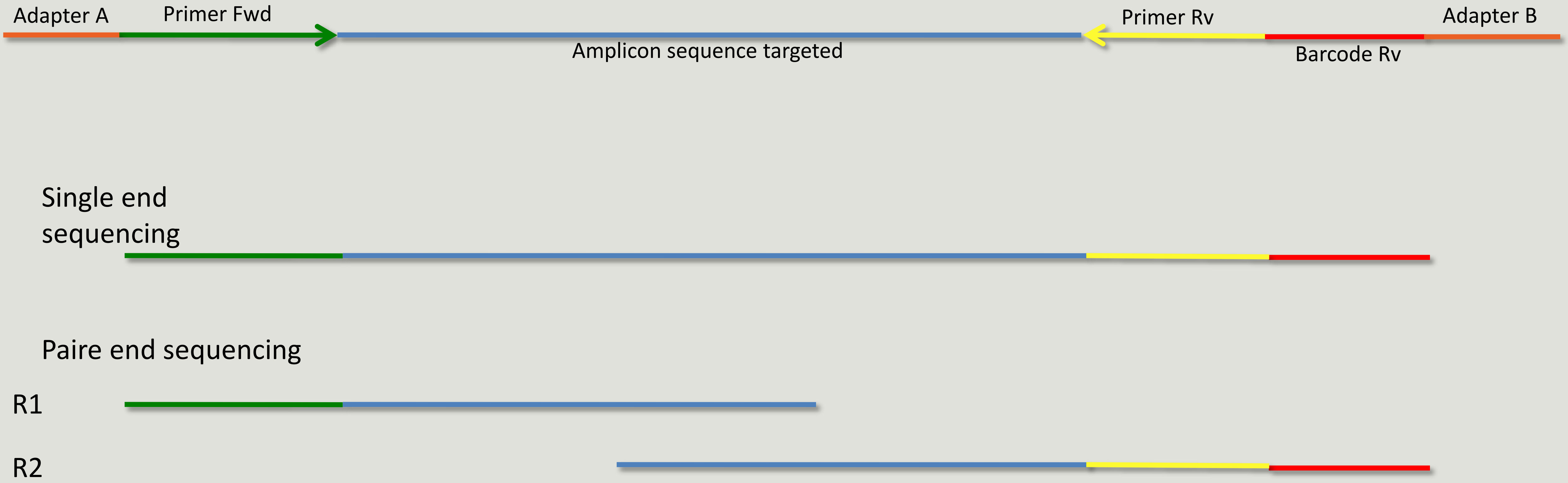
Single end sequencing

Paire end sequencing

R1

R2

Demultiplexing reverse



Adapter A

Primer Fwd

Amplicon sequence targeted

Primer Rv

Barcode Rv

Adapter B

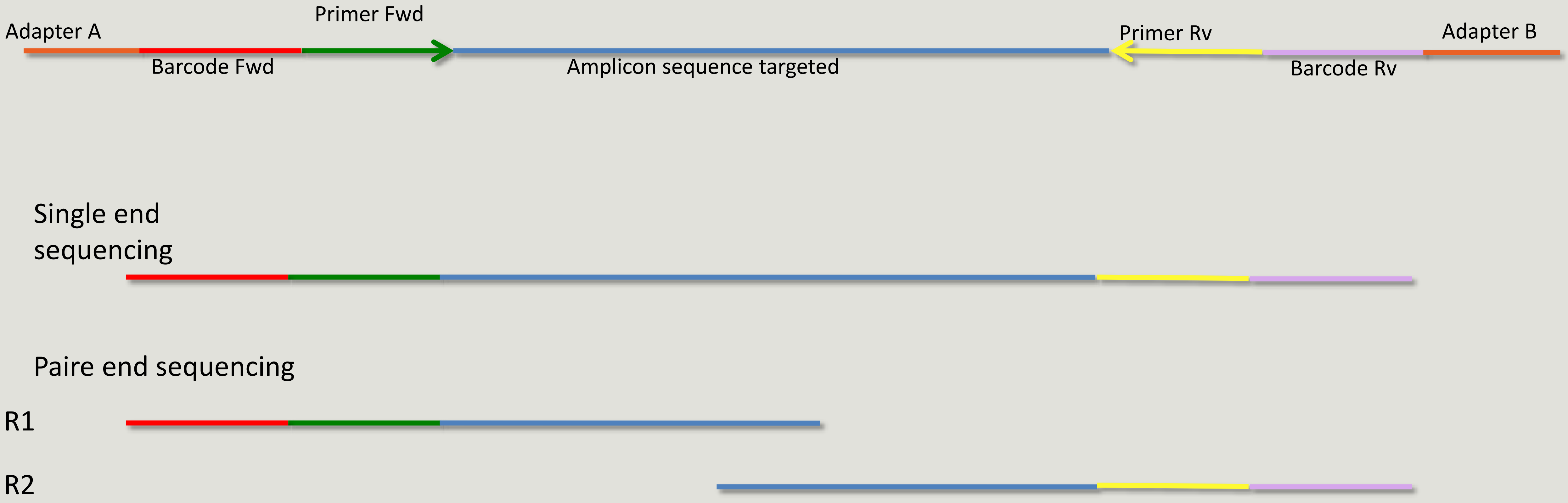
Single end sequencing

Paire end sequencing

R1

R2

Demultiplexing forward and reverse



Primer Fwd

Adapter A

Barcode Fwd

Amplicon sequence targeted

Primer Rv

Barcode Rv

Adapter B

Single end sequencing

Paire end sequencing

R1

R2

Your turn! - 2

Go to practice

GO TO EXERCISE 2

multiplexed

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
```

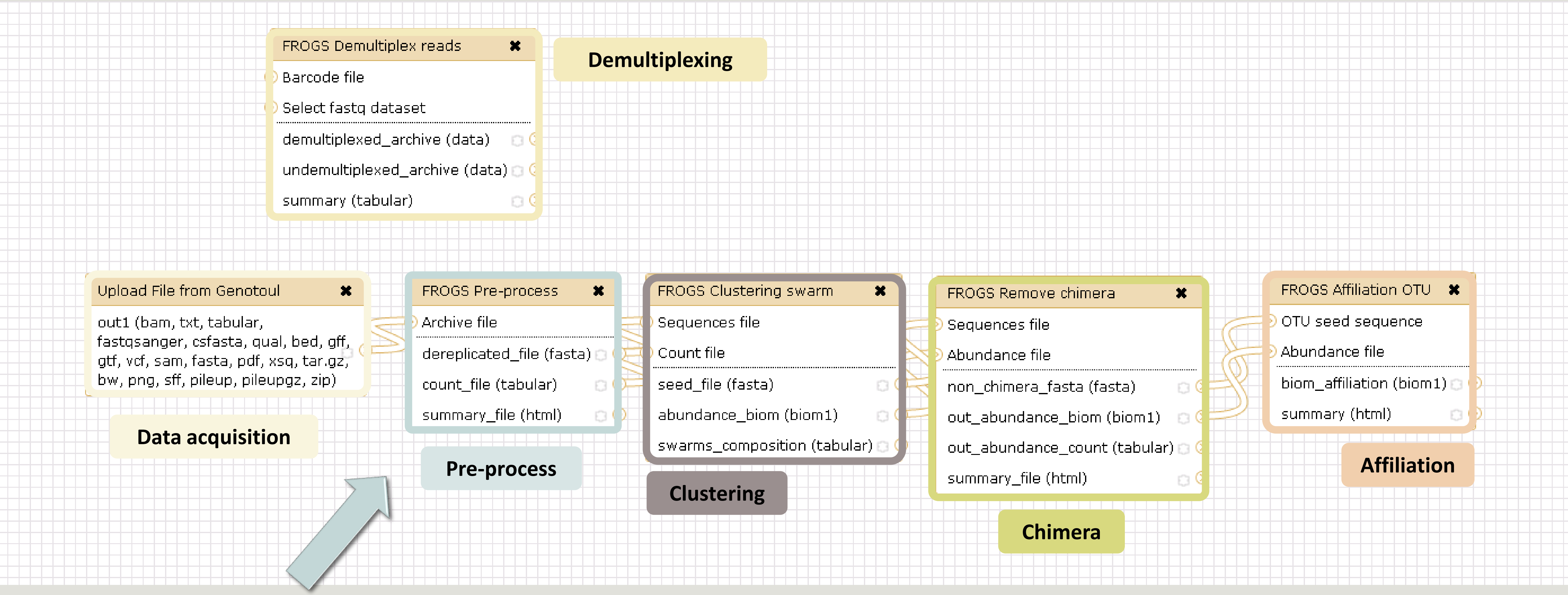
How it works ?

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 summary describes how many sequence are attributed for each sample.

Pre-process tool



Demultiplexing

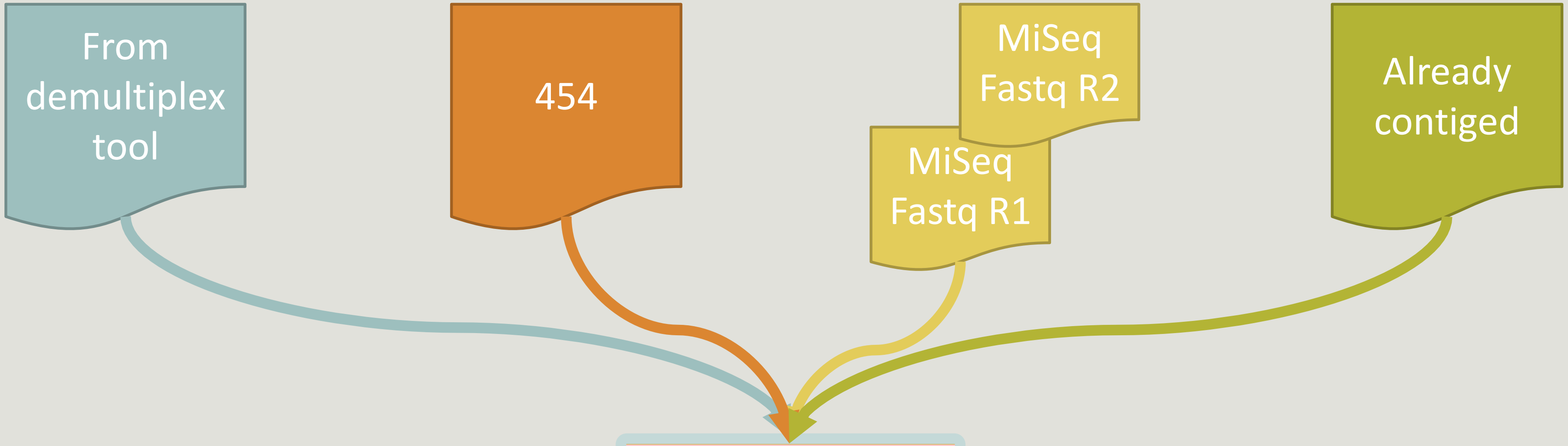
Data acquisition

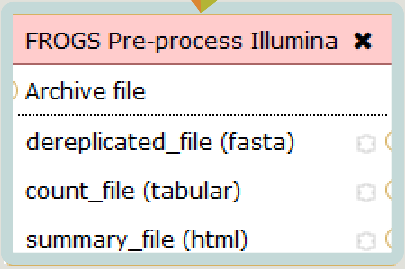
Pre-process

Clustering

Chimera

Affiliation




FROGS Pre-process Illumina ✖

Archive file

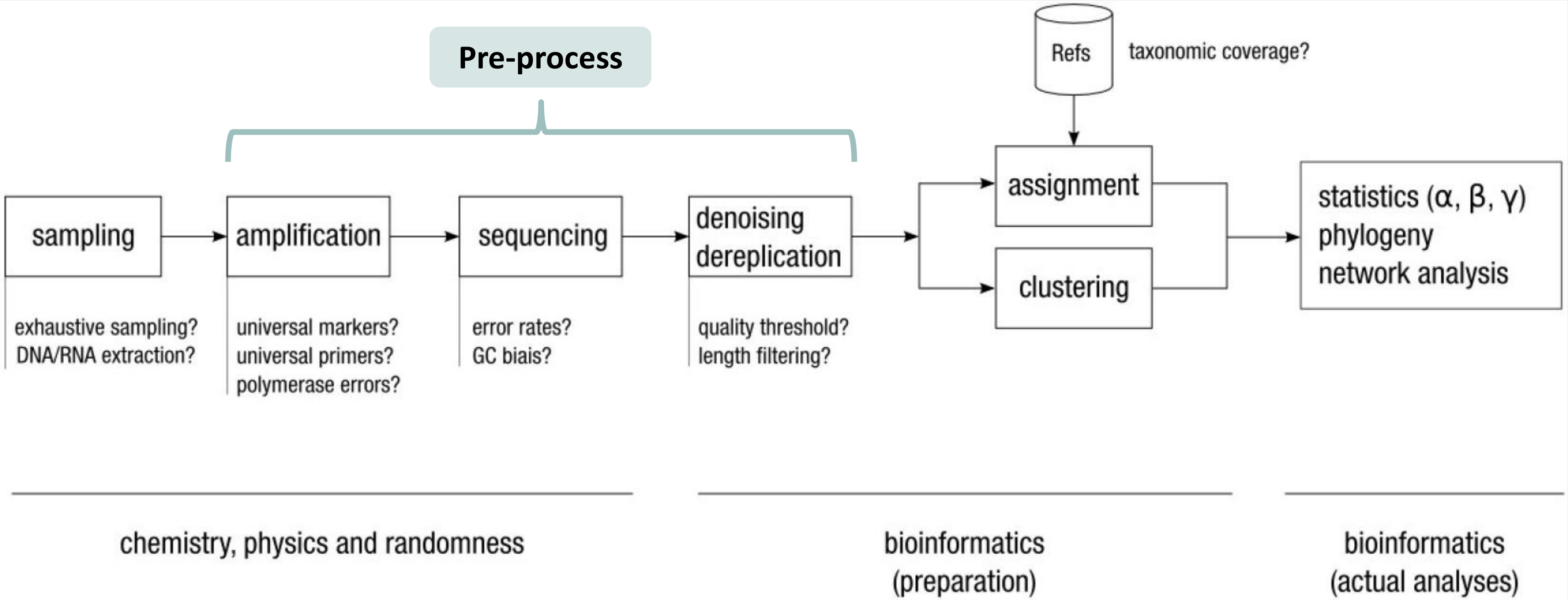
dereplicated_file (fasta)

count_file (tabular)

summary_file (html)

Pre-process

Amplicon-based studies general pipeline



Pre-process

sampling → amplification → sequencing → denoising dereplication → assignment / clustering → statistics (α, β, γ) phylogeny network analysis

Refs → assignment

taxonomic coverage?

exhaustive sampling?
DNA/RNA extraction?

universal markers?
universal primers?
polymerase errors?

error rates?
GC biais?

quality threshold?
length filtering?

chemistry, physics and randomness

bioinformatics (preparation)

bioinformatics (actual analyses)

Pre-process

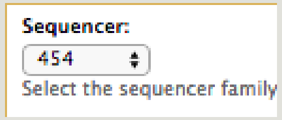
- Delete sequence with not expected lengths
- Delete sequences with ambiguous bases (N)
- Delete sequences do not contain good primers
- Dereplication

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

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

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

Pre-process

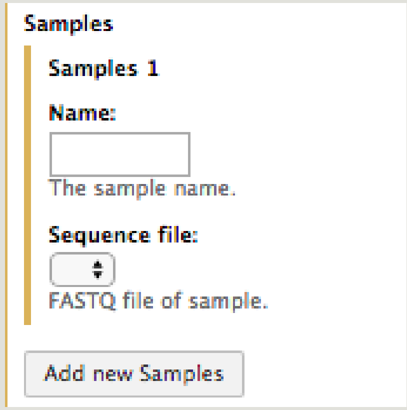


Sequencer:

454

Select the sequencer family

OR



Samples

Samples 1

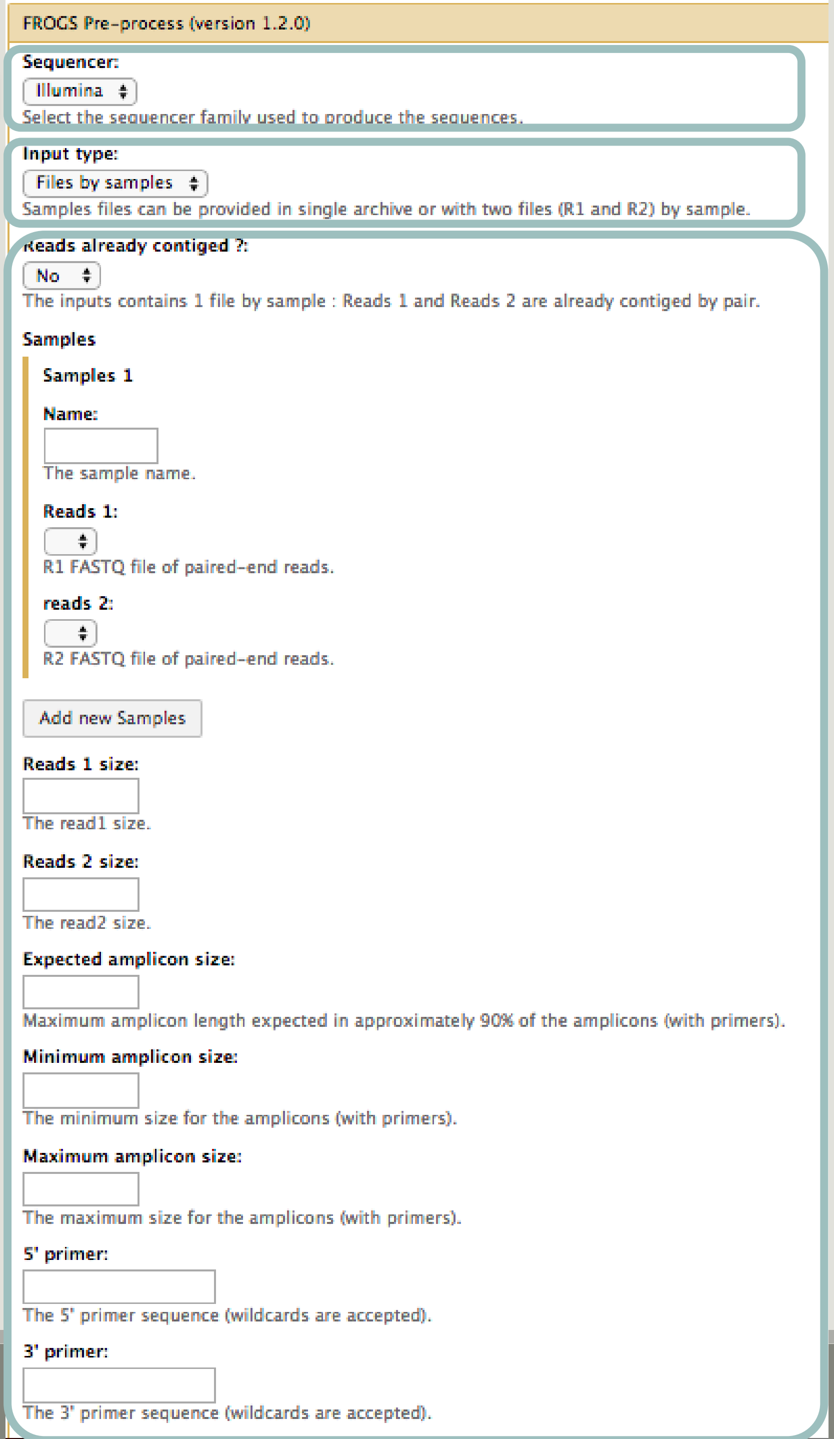
Name:

The sample name.

Sequence file:

FASTQ file of sample.

Add new Samples



FROGS Pre-process (version 1.2.0)

Sequencer:

Illumina

Select the sequencer family 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 contains 1 file by sample : Reads 1 and Reads 2 are already contiged by pair.

Samples

Samples 1

Name:

The sample name.

Reads 1:

R1 FASTQ file of paired-end reads.

reads 2:

R2 FASTQ file of paired-end reads.

Add new Samples

Reads 1 size:

The read1 size.

Reads 2 size:

The read2 size.

Expected amplicon size:

Maximum amplicon length expected in approximately 90% of the amplicons (with primers).

Minimum amplicon size:

The minimum size for the amplicons (with primers).

Maximum amplicon size:

The maximum size for the amplicons (with primers).

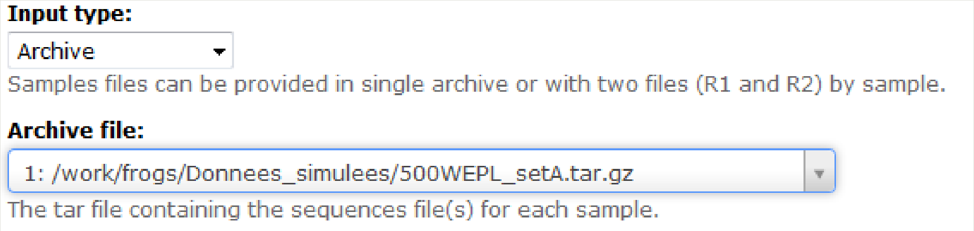
5' primer:

The 5' primer sequence (wildcards are accepted).

3' primer:

The 3' primer sequence (wildcards are accepted).

OR



Input type:

Archive

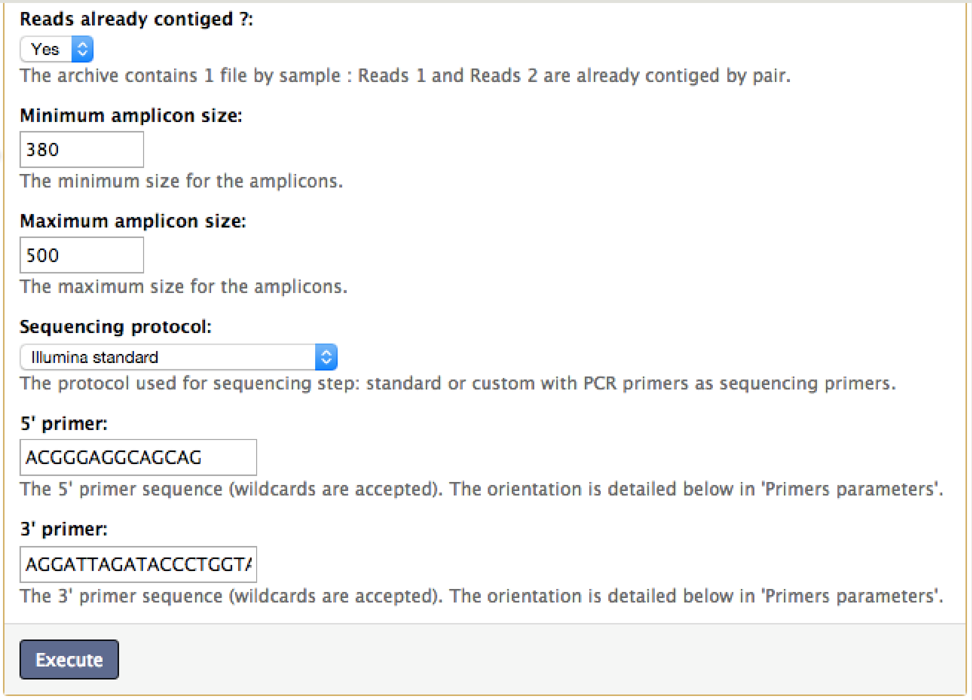
Samples files can be provided in single archive or with two files (R1 and R2) by sample.

Archive file:

1: /work/frogs/Donnees_simulees/500WEPL_setA.tar.gz

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

OR



Reads already contiged ?:

Yes

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.

Maximum amplicon size:

500

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:

ACGGGAGGCAGCAG

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

3' primer:

AGGATTAGATACCCTGGTA

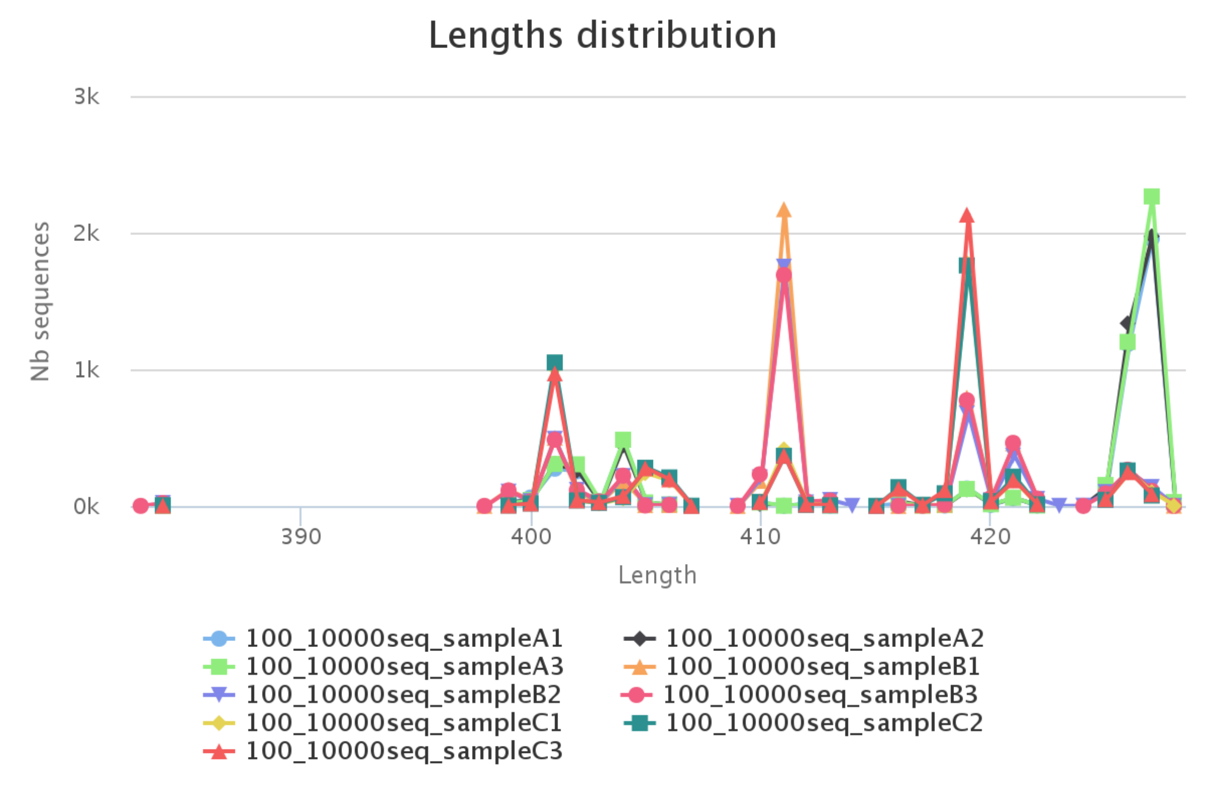
The 3' primer sequence (wildcards are accepted). The orientation is detailed below in 'Primers parameters'.

Your turn! - 3

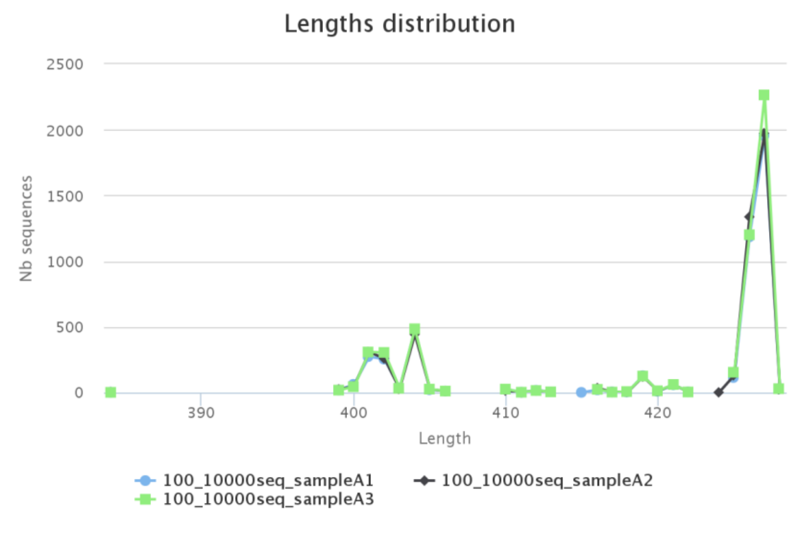
Go to practice

GO TO EXERCISES 3

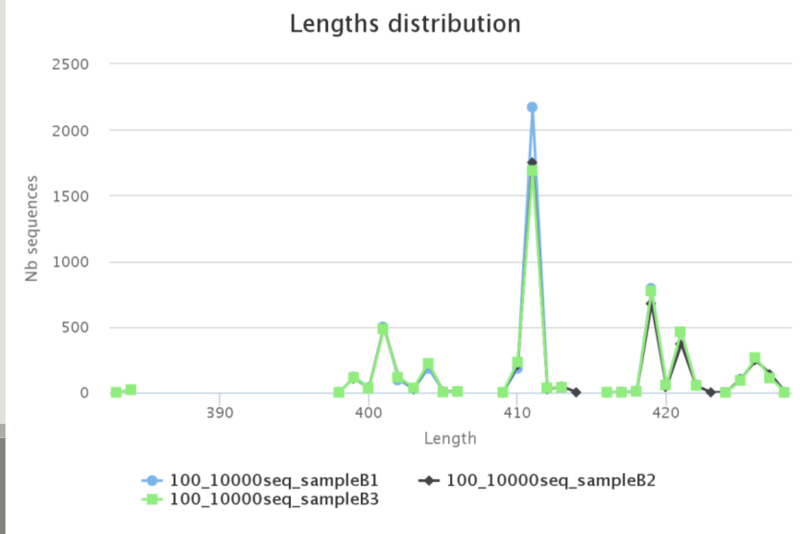
454



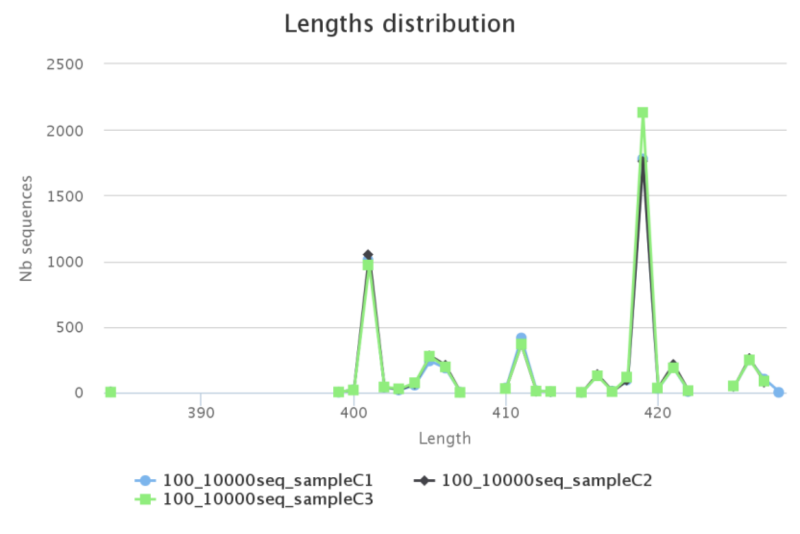
Samples A only



Samples B only



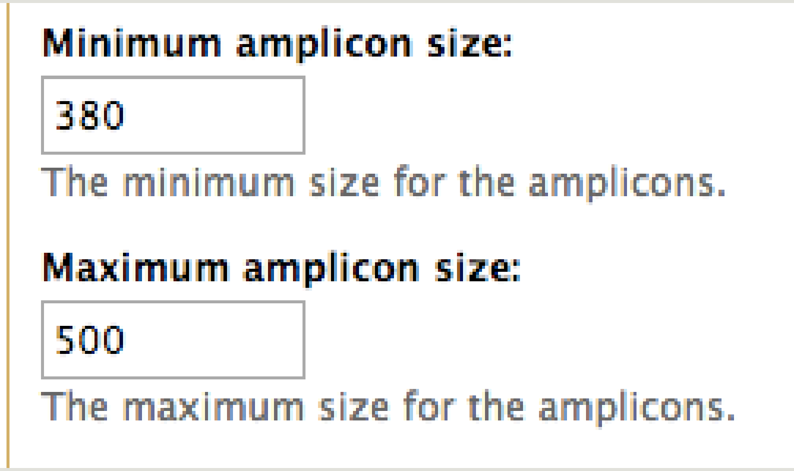
Samples C only



454

Cleaning, how it work ?

Filter contig sequence on its length which must be between min-amplicon-size and max-amplicon-size



Minimum amplicon size:

380

The minimum size for the amplicons.

Maximum amplicon size:

500

The maximum size for the amplicons.

use cutadapt to search and trim primers sequences with less than 10% differences

454

Go to practice

Cleaning, how it work ?

dereplicate sequences and return one uniq fasta file for all sample and a count table to indicate sequence abundances among sample.

In the HTML report file, you will find for each filter the number of sequences passing it, and a table that details these filters for each sample.

MiSeq
R1 R2

Flash, how it works ?

To contig read1 and read2 with FLASh with :

a minimum overlap equals to

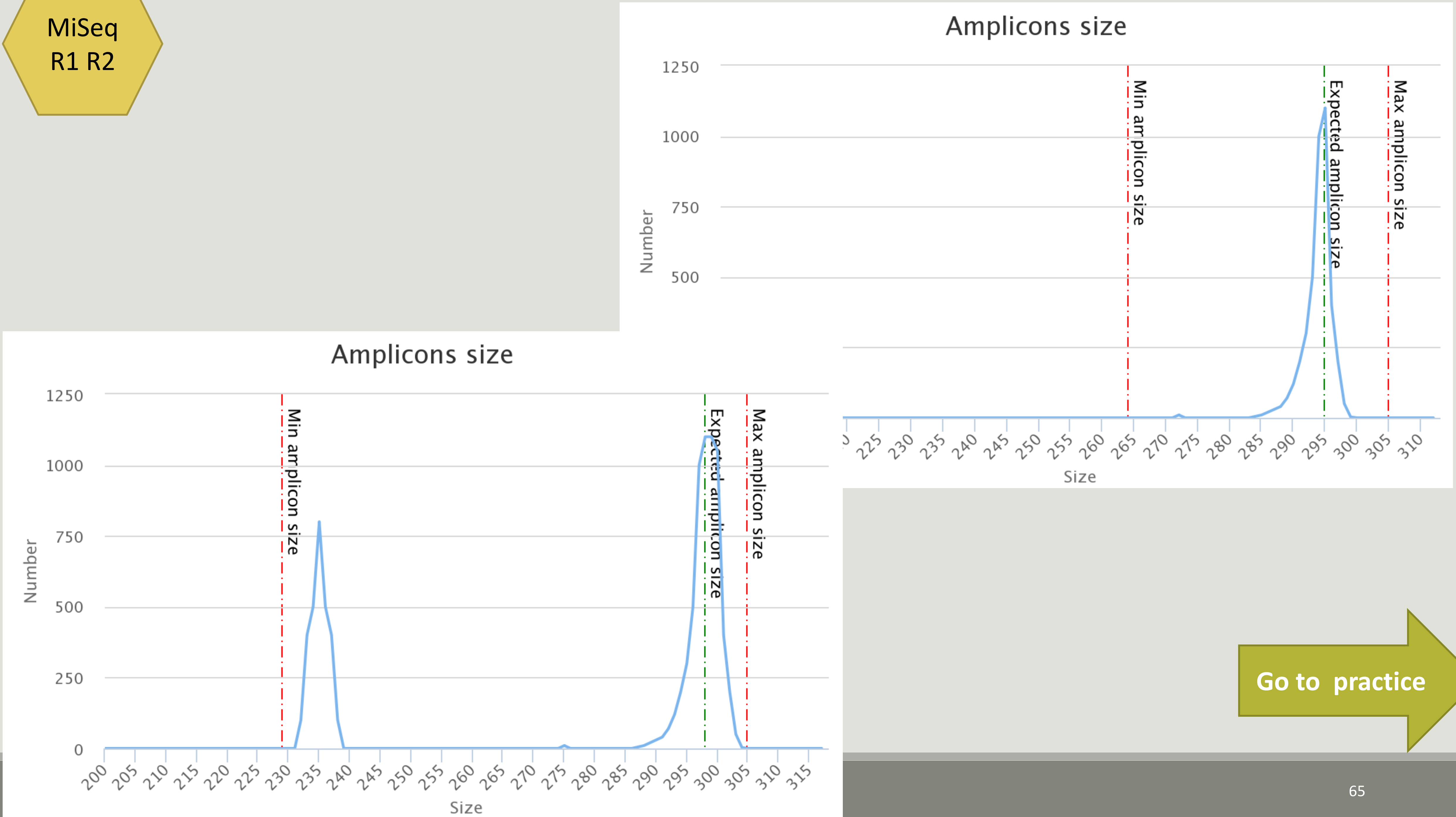
[(R1-size + R2-size) - expected-amplicon-size] ex: (250+250) - 450 = 50

and a maximum overlap equal to

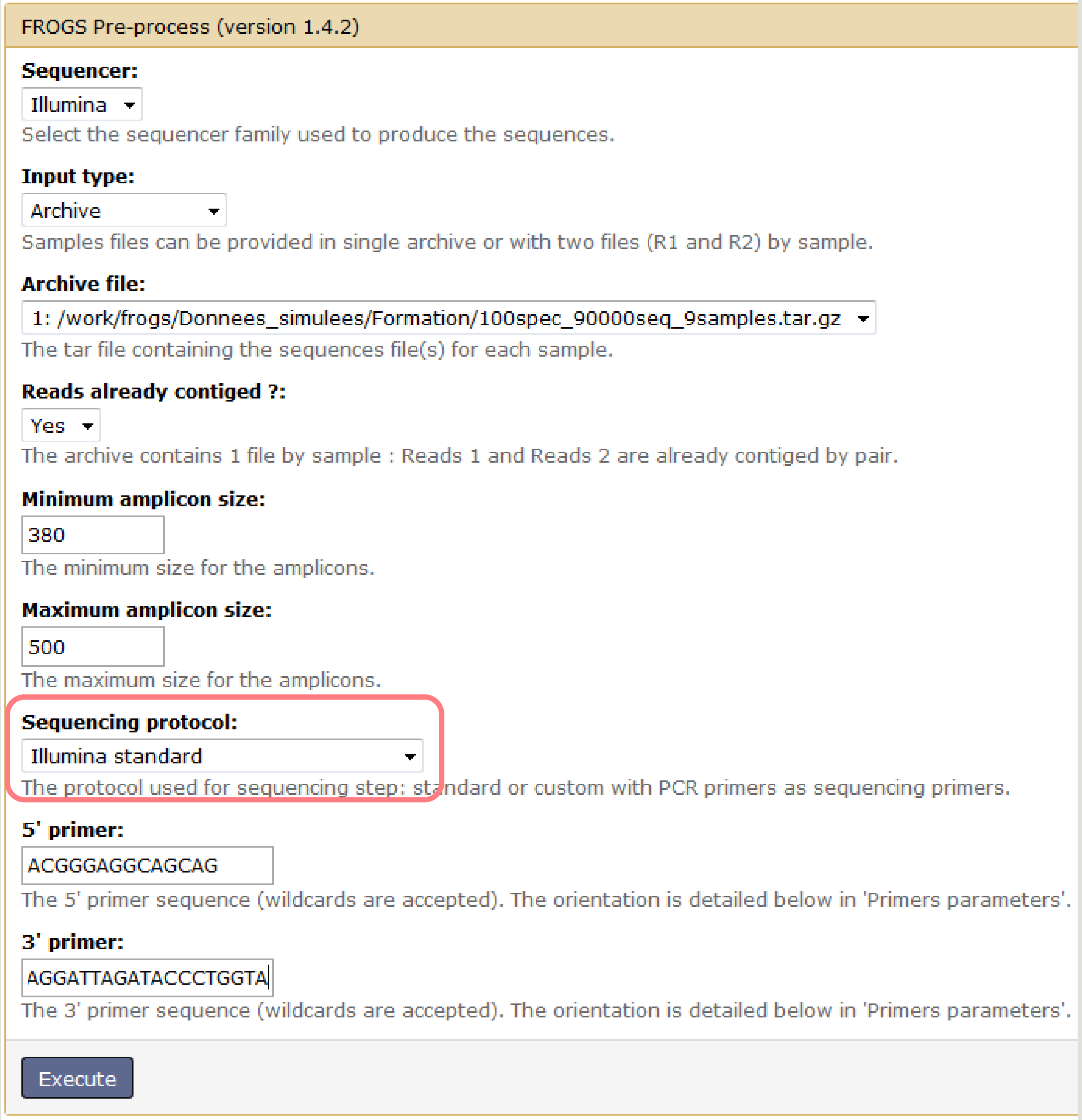
[expected-amplicon-size] with a maximum of 10% mismatch among this overlap

90% of the amplicon are smaller than [expected-amplicon-size]

MiSeq
R1 R2



Go to practice



FROGS Pre-process (version 1.4.2)

Sequencer:

Illumina

Select the sequencer family used to produce the sequences.

Input type:

Archive

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

Archive file:

1: /work/frogs/Donnees_simulees/Formation/100spec_90000seq_9samples.tar.gz

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

Reads already contiged ?:

Yes

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.

Maximum amplicon size:

500

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:

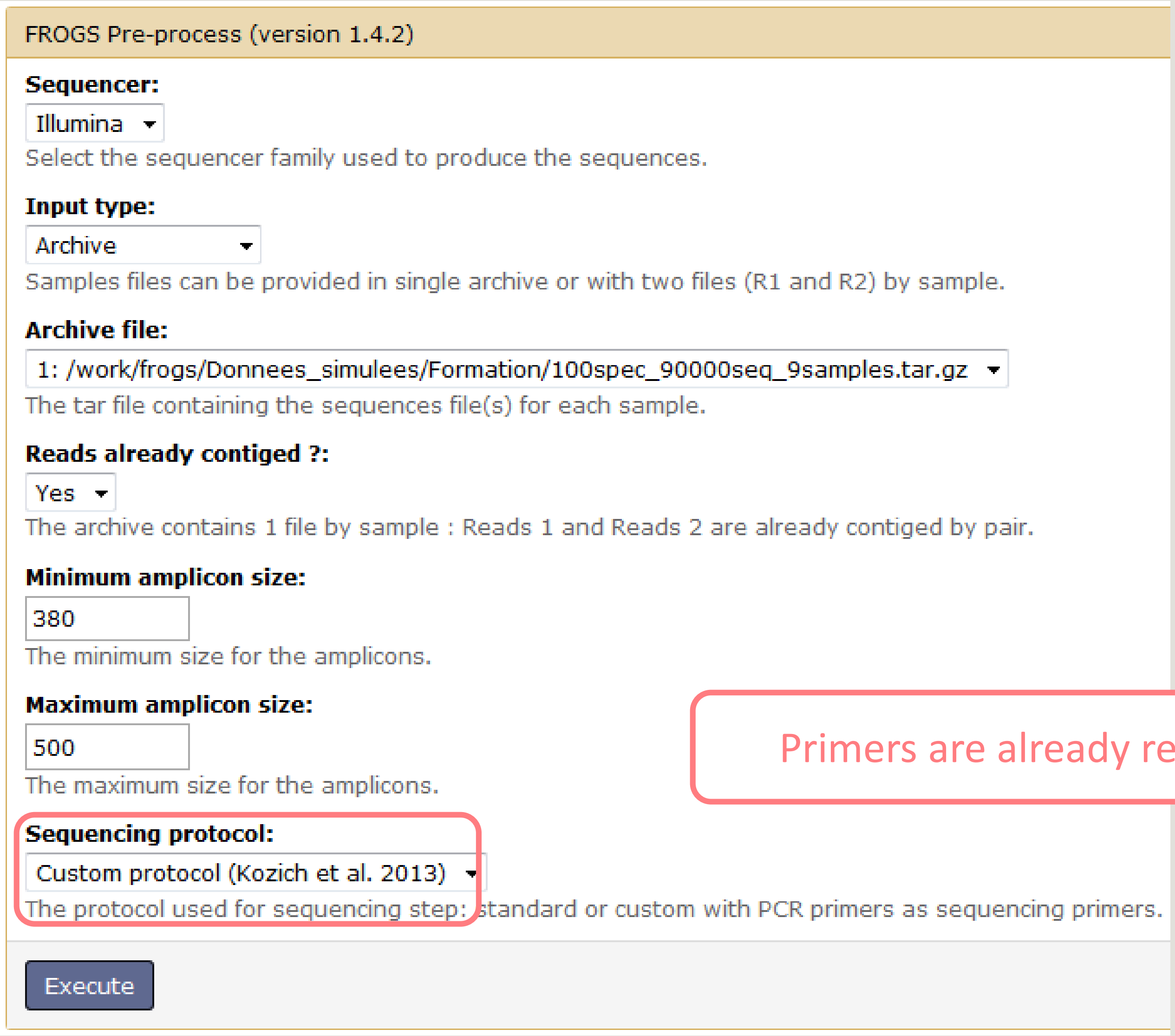
ACGGGAGGCAGCAG

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

3' primer:

AGGATTAGATACCCTGGTA

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



FROGS Pre-process (version 1.4.2)

Sequencer:

Illumina

Select the sequencer family used to produce the sequences.

Input type:

Archive

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

Archive file:

1: /work/frogs/Donnees_simulees/Formation/100spec_90000seq_9samples.tar.gz

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

Reads already contiged ?:

Yes

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.

Maximum amplicon size:

500

The maximum size for the amplicons.

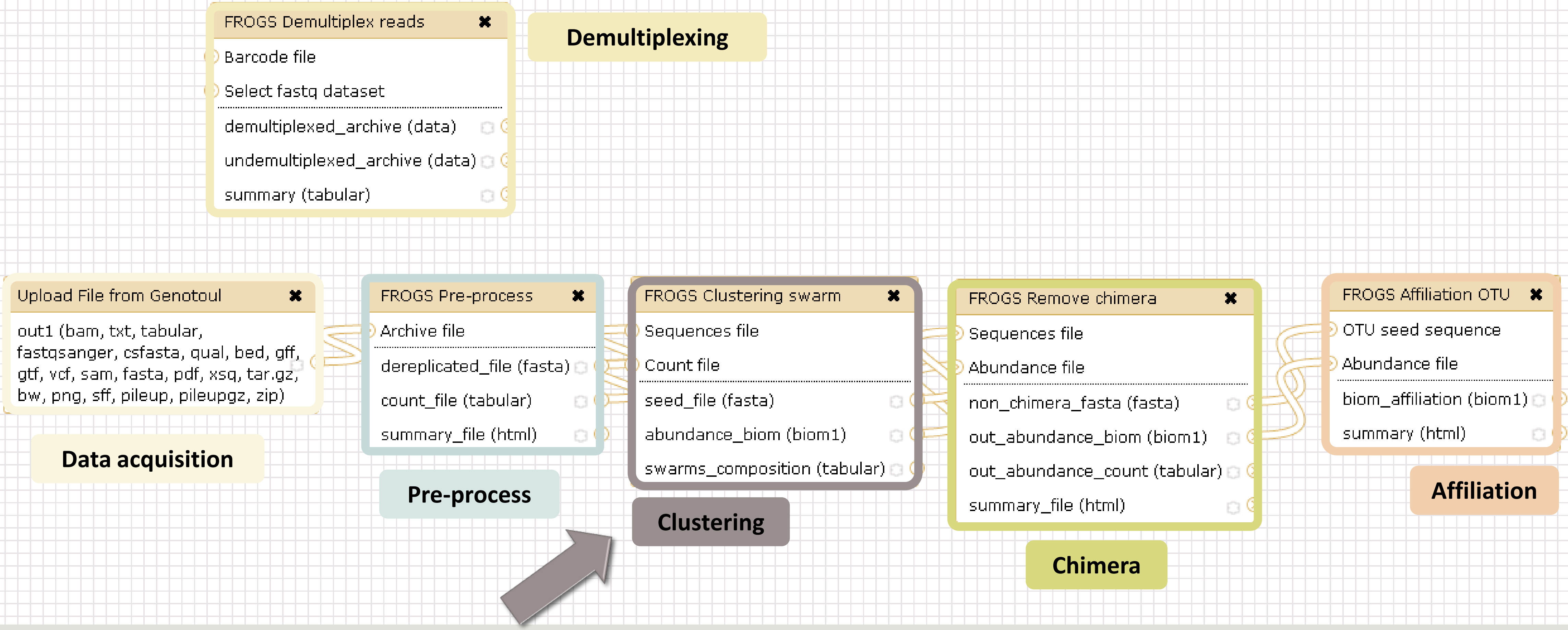
Sequencing protocol:

Custom protocol (Kozich et al. 2013)

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

Primers are already removed

Clustering tool



FROGS Demultiplex reads ✖

Barcode file

Select fastq dataset

demultiplexed_archive (data)

undemultiplexed_archive (data)

summary (tabular)

Demultiplexing

Upload File from Genotoul ✖

out1 (bam, txt, tabular, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png, sff, pileup, pileupgz, zip)

Data acquisition

FROGS Pre-process ✖

Archive file

dereplicated_file (fasta)

count_file (tabular)

summary_file (html)

Pre-process

FROGS Clustering swarm ✖

Sequences file

Count file

seed_file (fasta)

abundance_biom (biom1)

swarms_composition (tabular)

Clustering

FROGS Remove chimera ✖

Sequences file

Abundance file

non_chimera_fasta (fasta)

out_abundance_biom (biom1)

out_abundance_count (tabular)

summary_file (html)

Chimera

FROGS Affiliation OTU ✖

OTU seed sequence

Abundance file

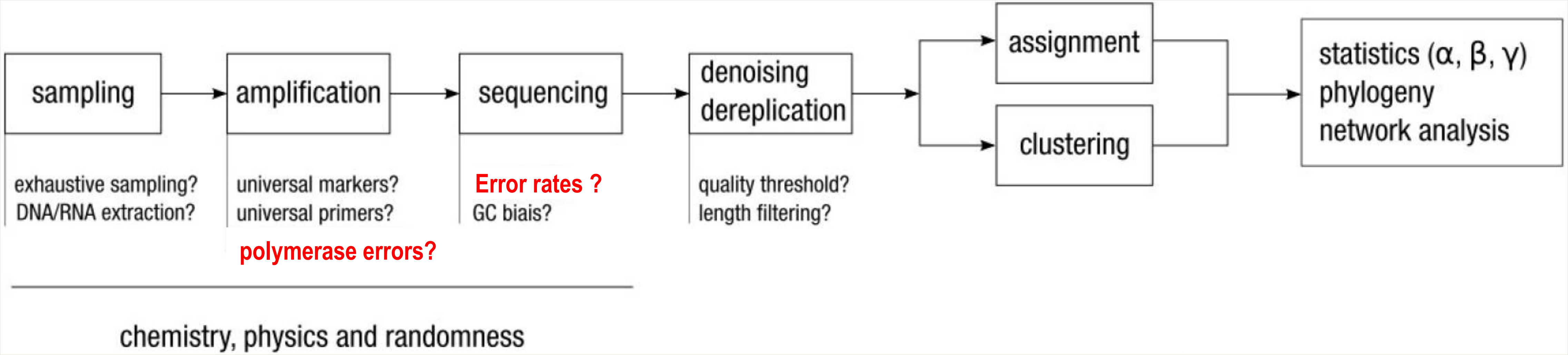
biom_affiliation (biom1)

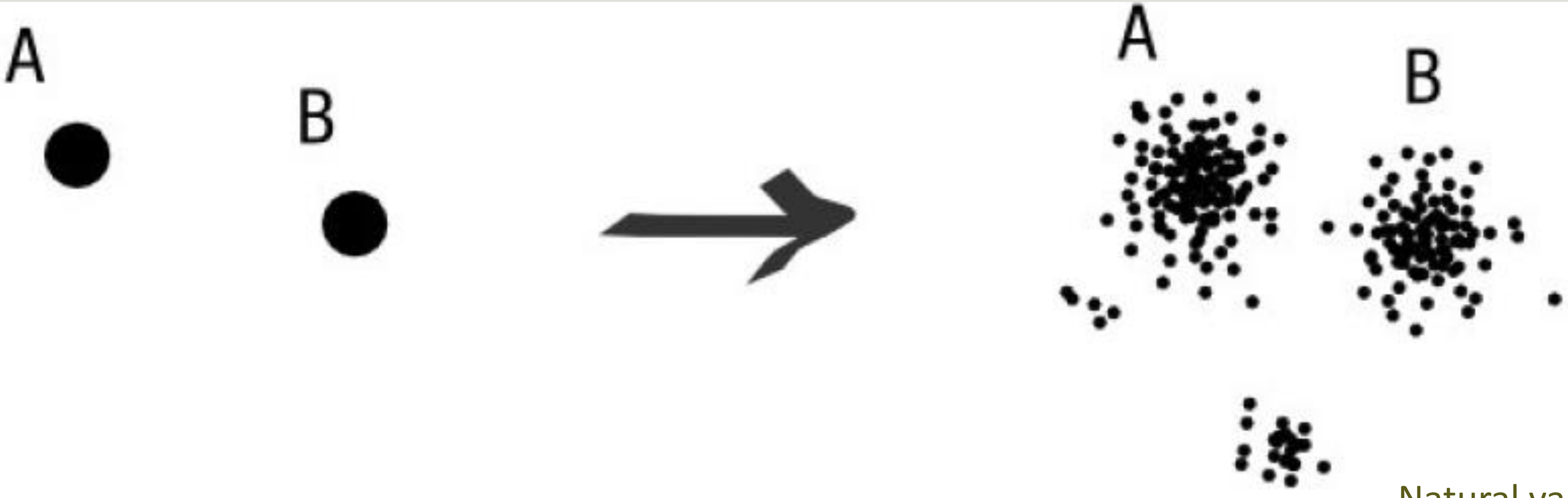
summary (html)

Affiliation

Why do we need clustering ?

Amplication and sequencing and are not perfect processes



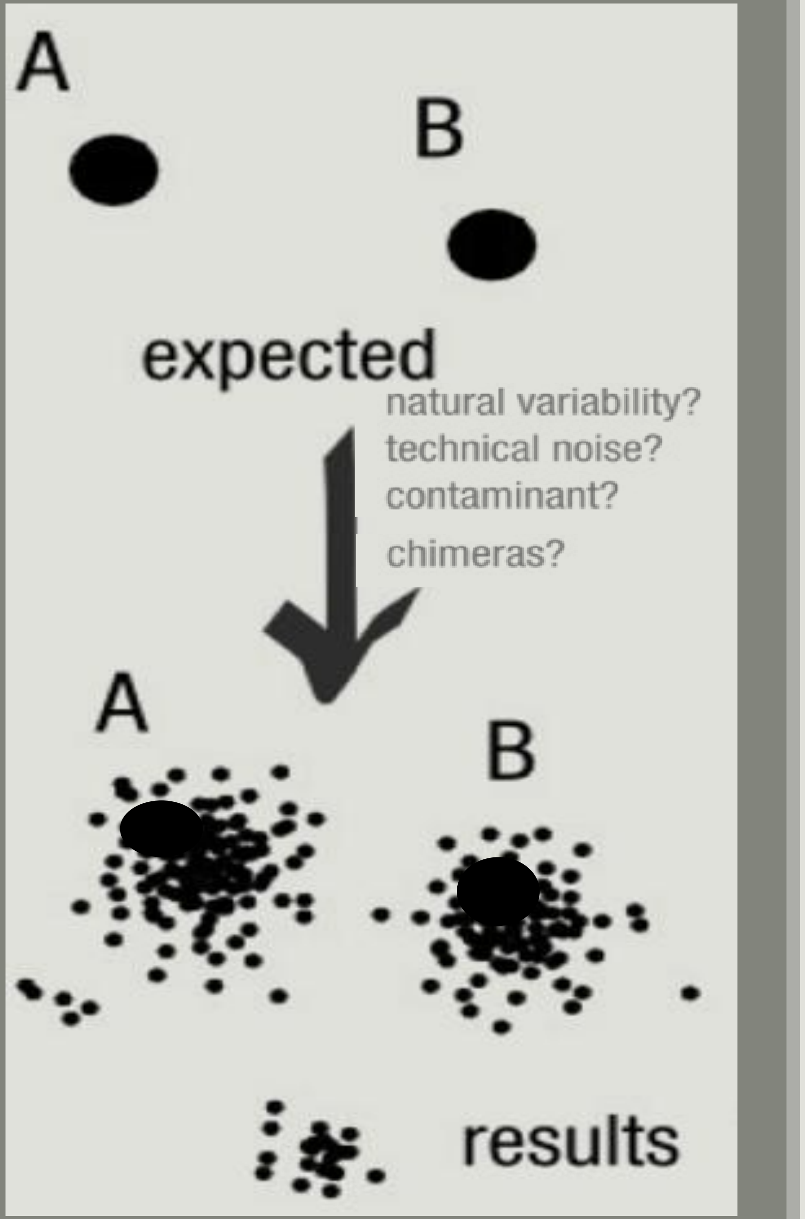


Expected

Results

Natural variability ?
Technical noise?
Contaminant?
Pseudogene?
Chimeras?

Fréderic Mahé communication



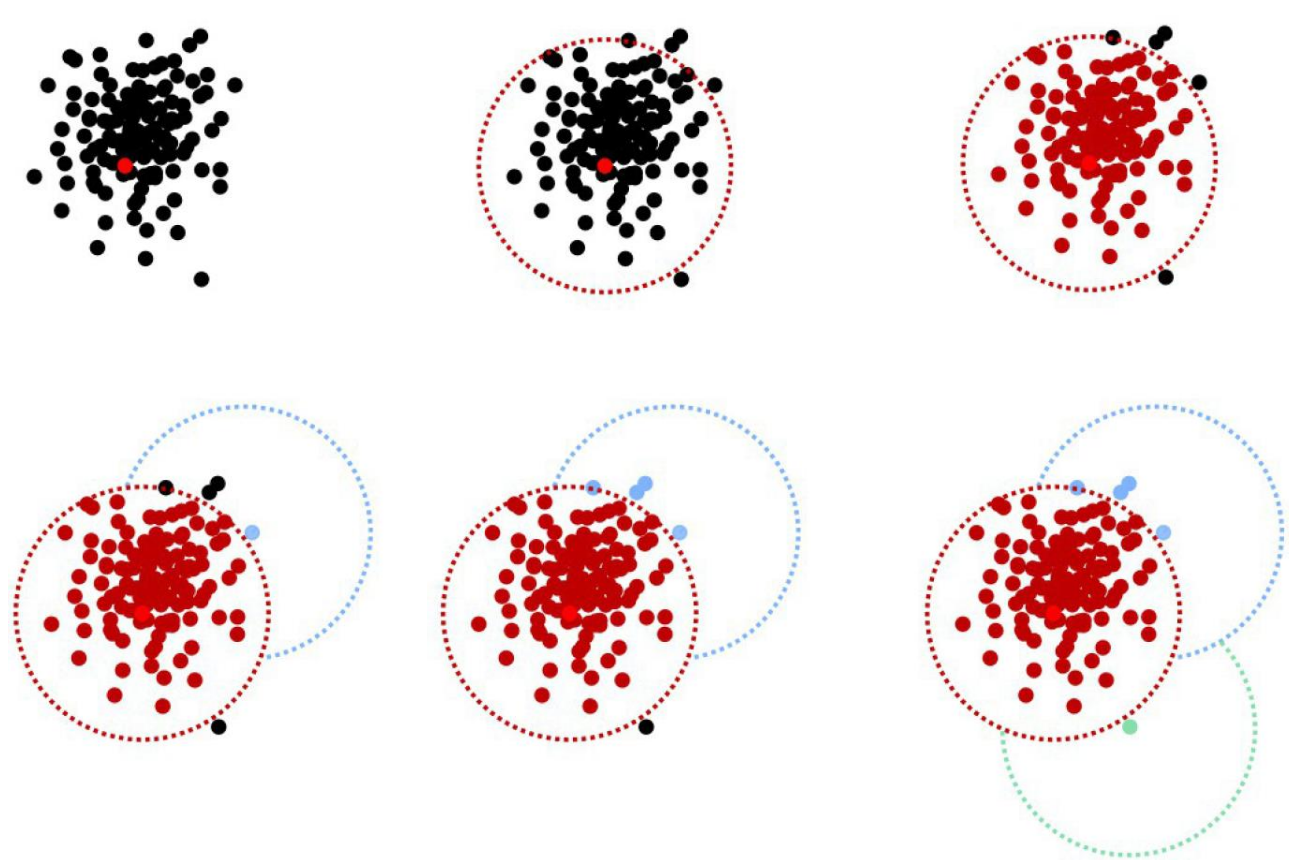
To have the best accuracy:

Method: All against all

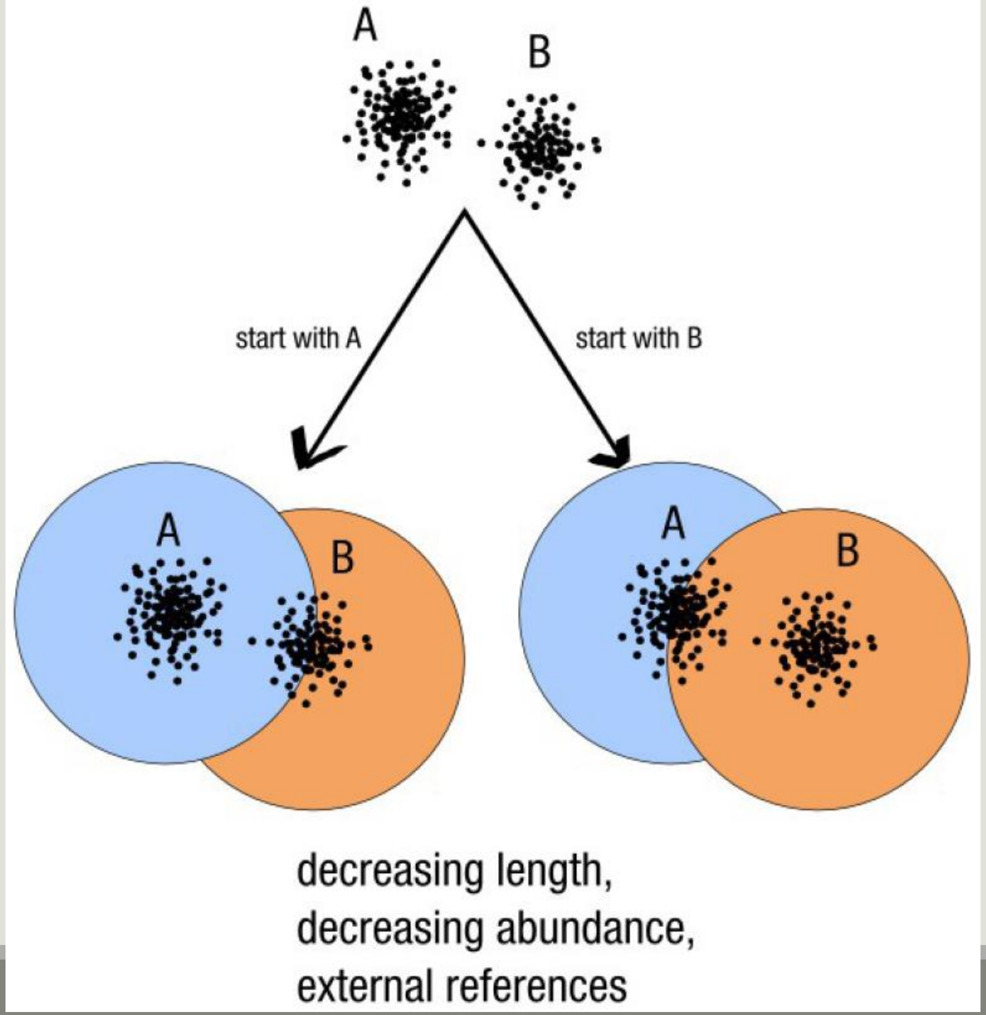
- Very accurate
- Requires a lot of memory and/or time

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

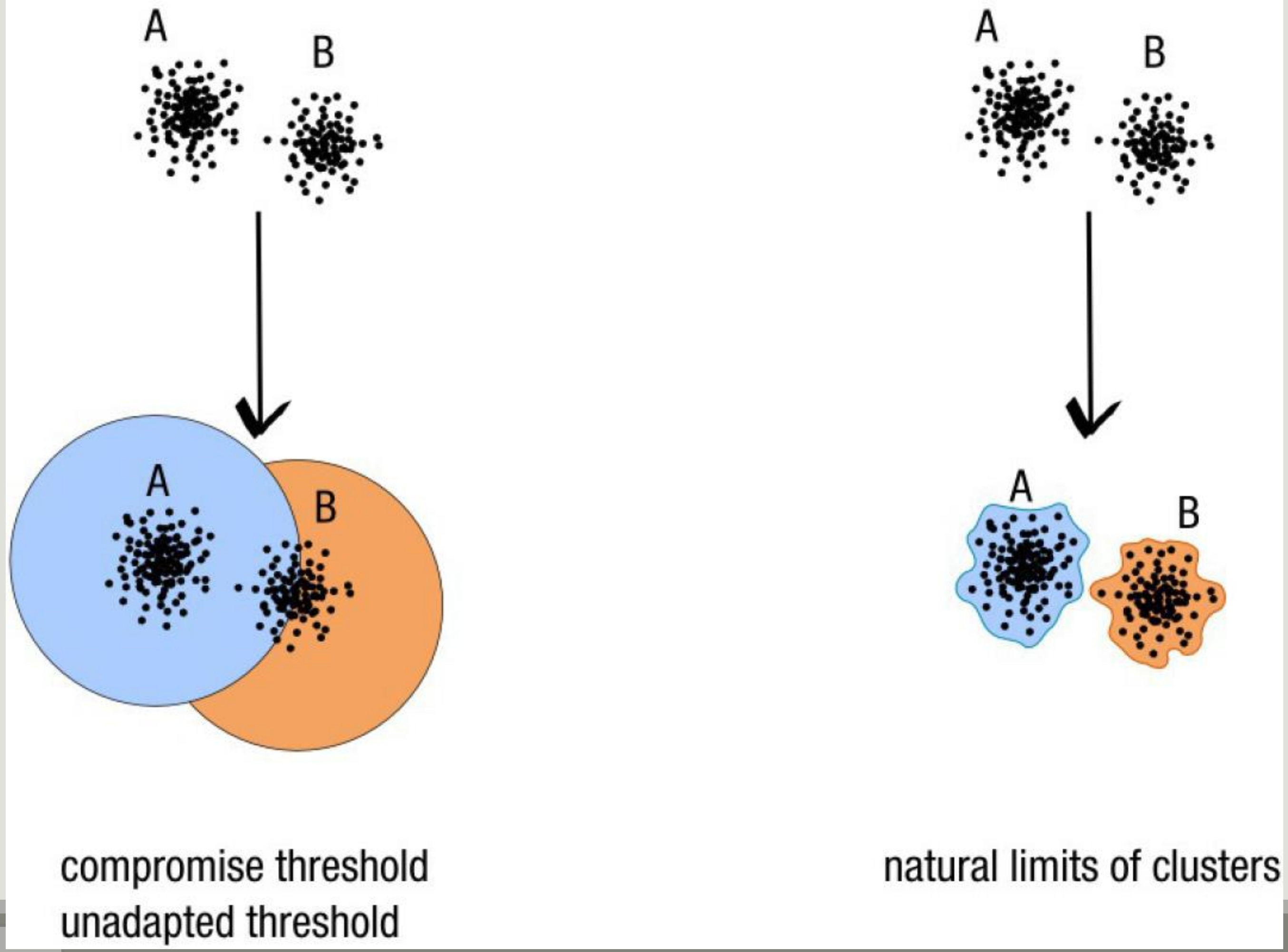
How traditional clustering works ?



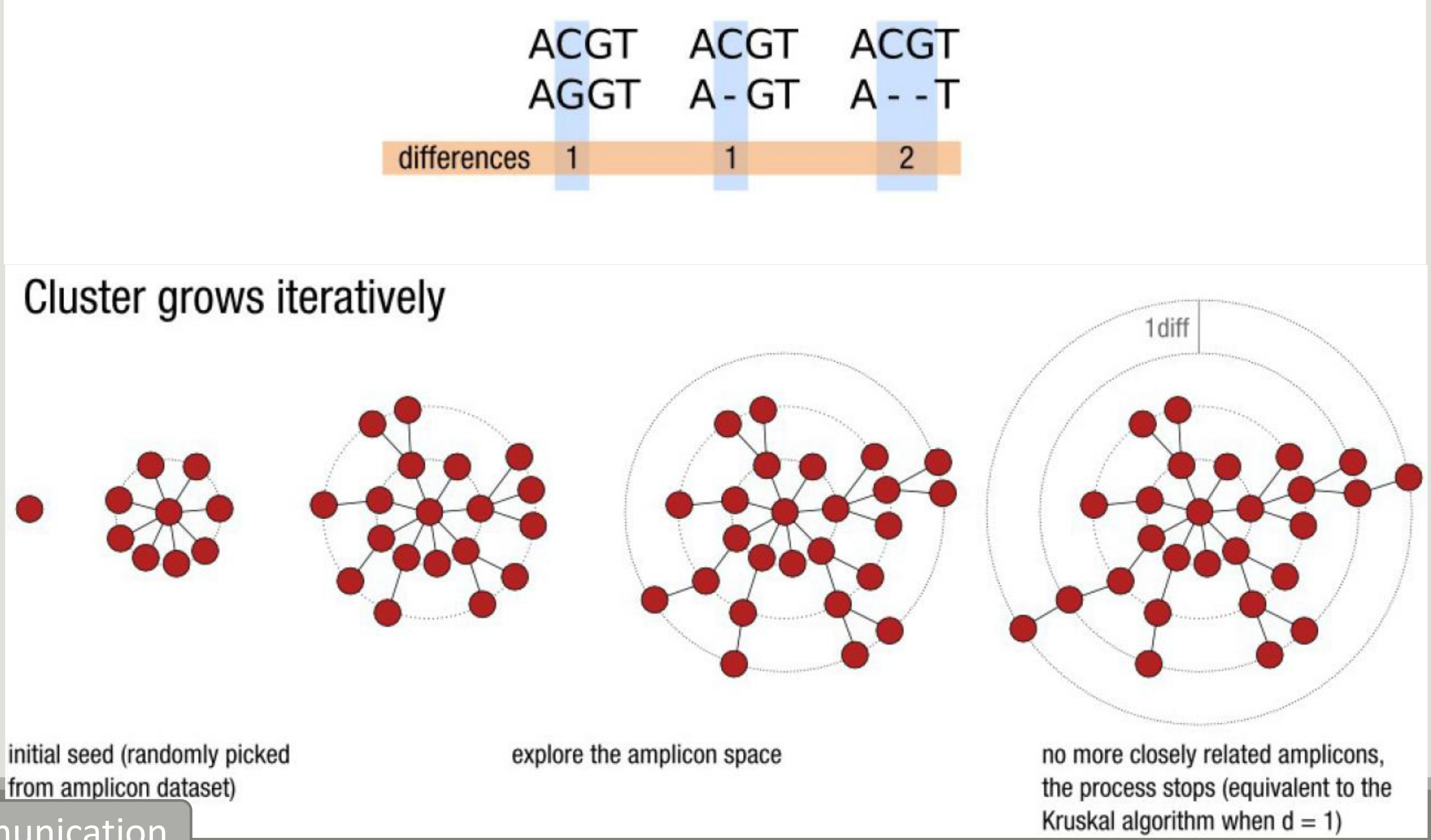
Input order dependent results



Single a priori clustering threshold



Swarm clustering method



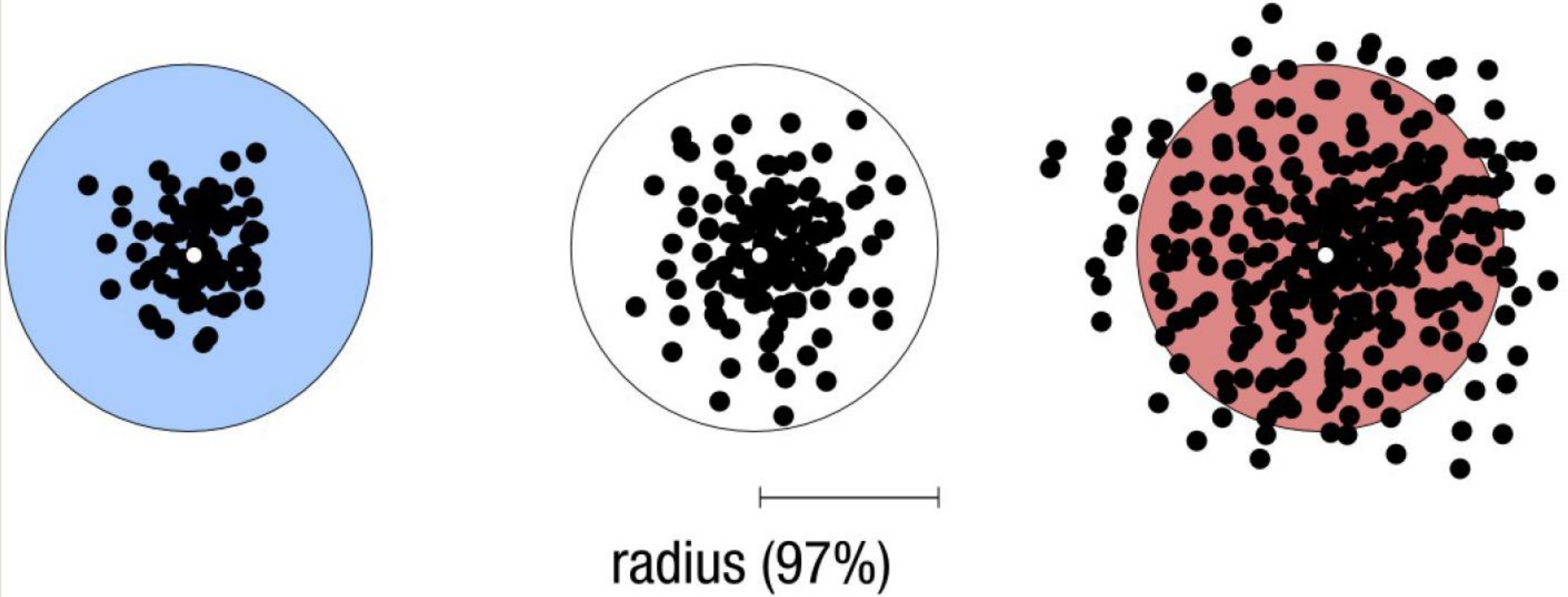
initial seed (randomly picked from amplicon dataset)

explore the amplicon space

no more closely related amplicons, the process stops (equivalent to the Kruskal algorithm when d = 1)

Fréderic Mahé communication

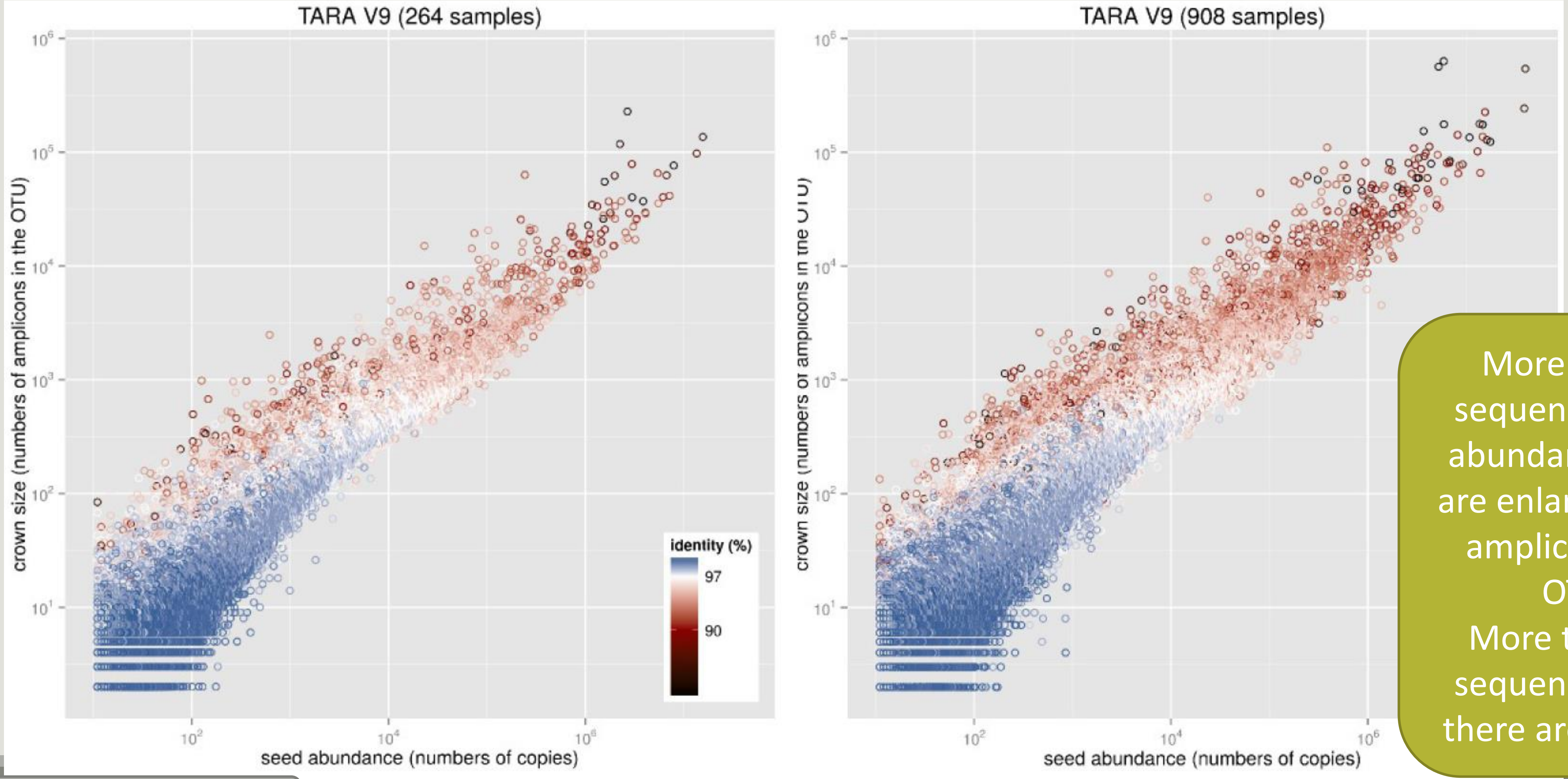
Comparison Swarm and 3% clusterings



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

Fréderic Mahé communication

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 OTU).
More there are sequences, more there are artefacts

Fréderic Mahé communication

SWARM

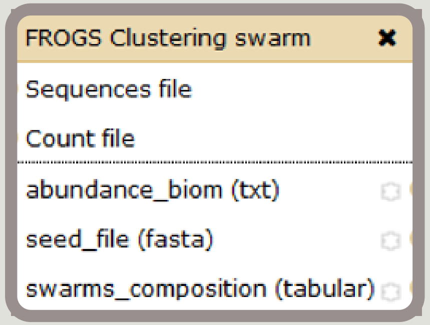
A robust and fast clustering method for amplicon-based studies.

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: 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



FROGS Clustering swarm ✖

Sequences file

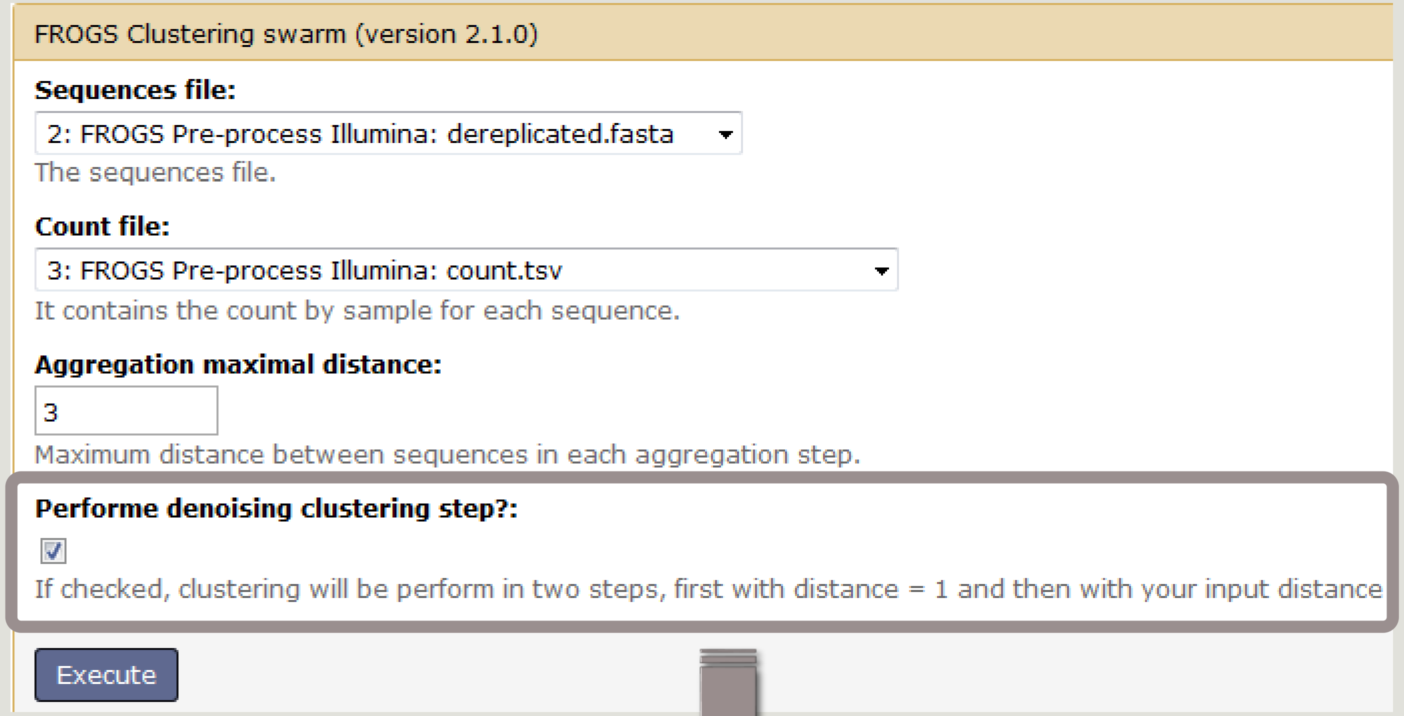
Count file

abundance_biom (txt)

seed_file (fasta)

swarms_composition (tabular)

Clustering



FROGS Clustering swarm (version 2.1.0)

Sequences file:

2: FROGS Pre-process Illumina: dereplicated.fasta ▾

The sequences file.

Count file:

3: FROGS Pre-process Illumina: count.tsv ▾

It contains the count by sample for each sequence.

Aggregation maximal distance:

3

Maximum distance between sequences in each aggregation step.

Performe denoising clustering step?:

☑

If checked, clustering will be perform in two steps, first with distance = 1 and then with your input distance

<u>1st run for denoising:</u>

Swarm with d = 1 -> high OTUs definition
linear complexity

<u>2nd run for clustering:</u>

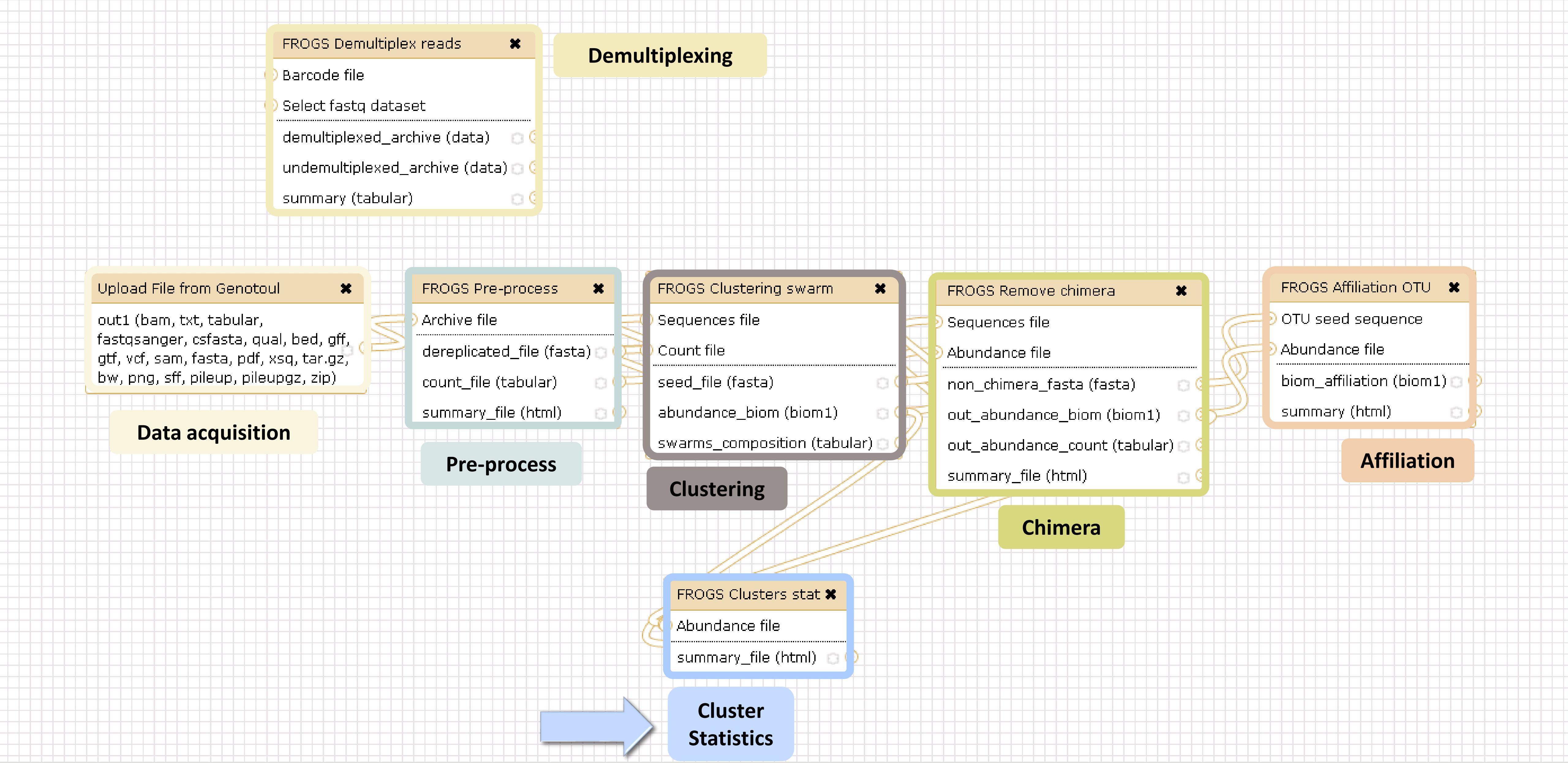
Swarm with d = 3 on the seeds of first Swarm
quadratic complexity

Gain time !

Remove false positives !

PeerJ PrePrints 2:e386v1. 2014. doi: 10.7287/peerj.preprints.386v1
Swarm: robust and fast clustering method for amplicon-based studies.
Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M.

Cluster stat tool



FROGS Demultiplex reads ✖

Barcode file

Select fastq dataset

demultiplexed_archive (data)

undemultiplexed_archive (data)

summary (tabular)

Demultiplexing

Upload File from Genotoul ✖

out1 (bam, txt, tabular, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png, sff, pileup, pileupgz, zip)

Data acquisition

FROGS Pre-process ✖

Archive file

dereplicated_file (fasta)

count_file (tabular)

summary_file (html)

Pre-process

FROGS Clustering swarm ✖

Sequences file

Count file

seed_file (fasta)

abundance_biom (biom1)

swarms_composition (tabular)

Clustering

FROGS Remove chimera ✖

Sequences file

Abundance file

non_chimera_fasta (fasta)

out_abundance_biom (biom1)

out_abundance_count (tabular)

summary_file (html)

Chimera

FROGS Affiliation OTU ✖

OTU seed sequence

Abundance file

biom_affiliation (biom1)

summary (html)

Affiliation

FROGS Clusters stat ✖

Abundance file

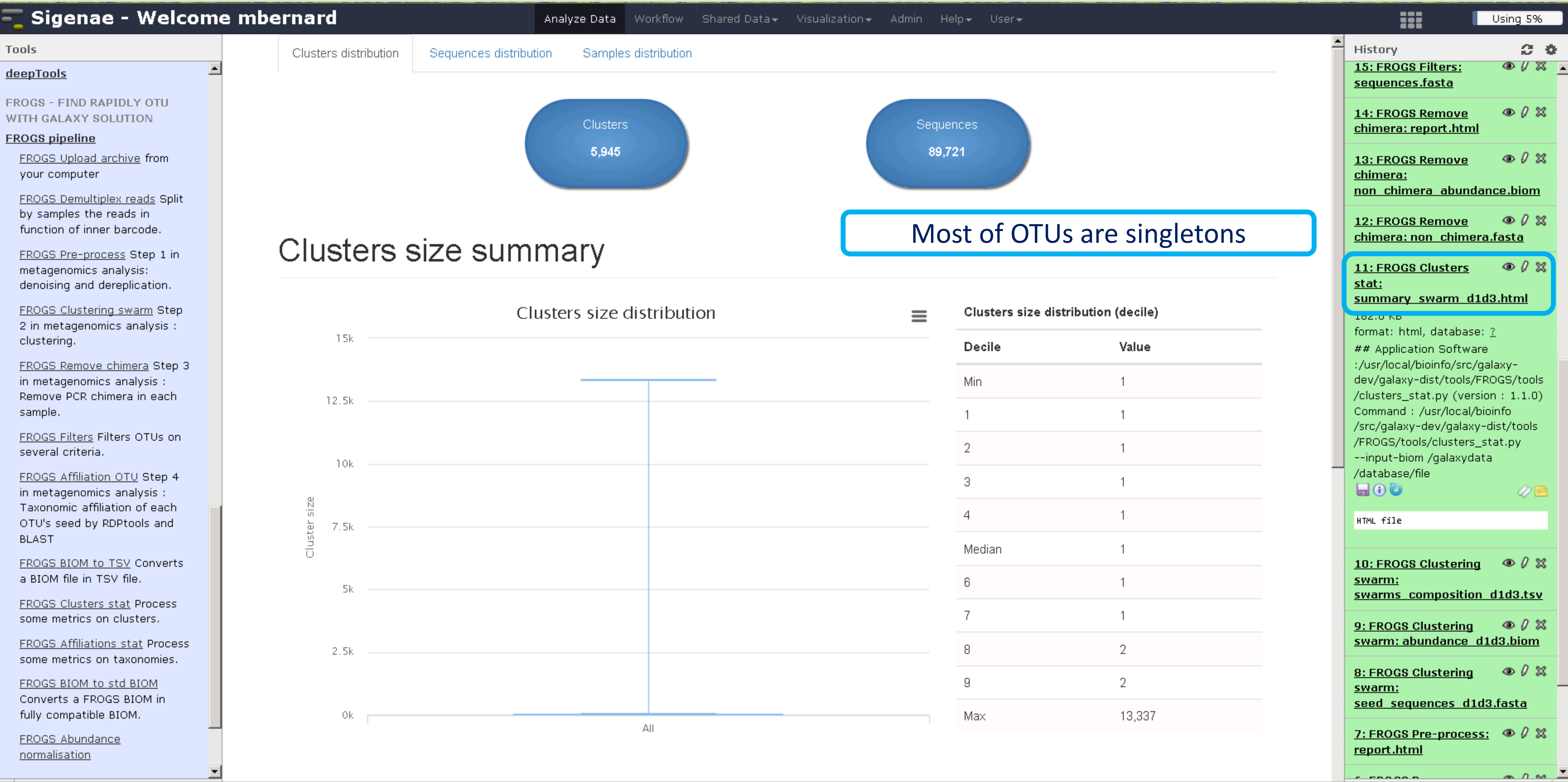
summary_file (html)

Cluster Statistics

Your Turn! - 4

EXERCISE 4

Go to practice



deepTools

FROGS - FIND RAPIDLY OTU WITH GALAXY SOLUTION

FROGS pipeline

FROGS Upload archive from your computer

FROGS Demultiplex reads Split by samples the reads in function of inner barcode.

FROGS Pre-process Step 1 in metagenomics analysis: denoising and dereplication.

FROGS Clustering swarm Step 2 in metagenomics analysis : clustering.

FROGS Remove chimera Step 3 in metagenomics analysis : Remove PCR chimera in each sample.

FROGS Filters Filters OTUs on several criteria.

FROGS Affiliation OTU Step 4 in metagenomics analysis : Taxonomic affiliation of each OTU's seed by RDPtools and BLAST

FROGS BIOM to TSV Converts a BIOM file in TSV file.

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 Abundance normalisation

Most of OTUs are singletons

Clusters size summary

Clusters size distribution (decile)

Decile	Value
Min	1
1	1
2	1
3	1
4	1
Median	1
6	1
7	1
8	2
9	2
Max	13,337

15: FROGS Filters: sequences.fasta

14: FROGS Remove chimera: report.html

13: FROGS Remove chimera: non_chimera_abundance.biom

12: FROGS Remove chimera: non_chimera.fasta

11: FROGS Clusters stat: summary_swarm_d1d3.html

format: html, database: ?

Application Software :/usr/local/bioinfo/src/galaxy-dev/galaxy-dist/tools/FROGS/tools/clusters_stat.py (version : 1.1.0) Command : /usr/local/bioinfo/src/galaxy-dev/galaxy-dist/tools/FROGS/tools/clusters_stat.py --input-biom /galaxydata/database/file

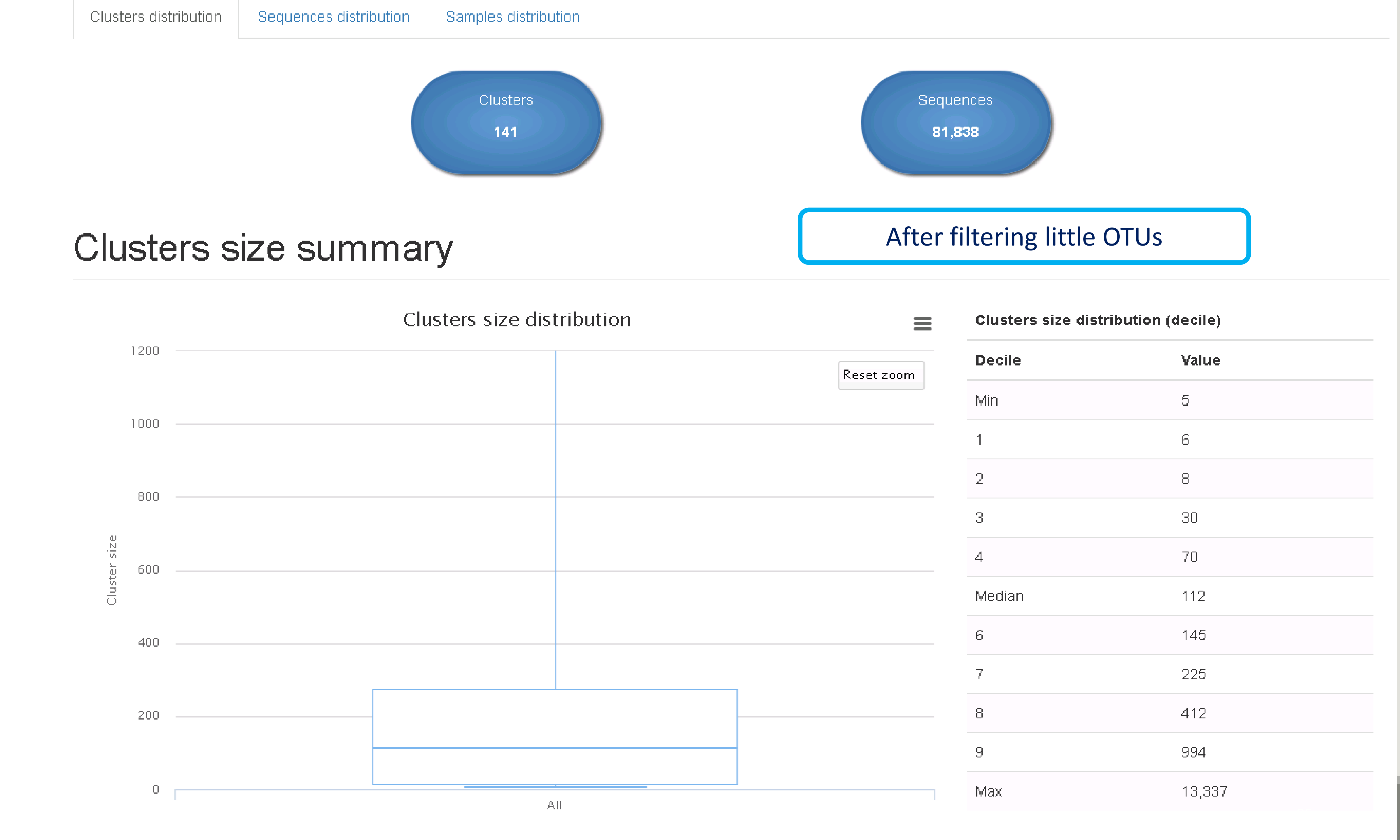
HTML file

10: FROGS Clustering swarm: swarms_composition_d1d3.tsv

9: FROGS Clustering swarm: abundance_d1d3.biom

8: FROGS Clustering swarm: seed_sequences_d1d3.fasta

7: FROGS Pre-process: report.html



Clusters 141

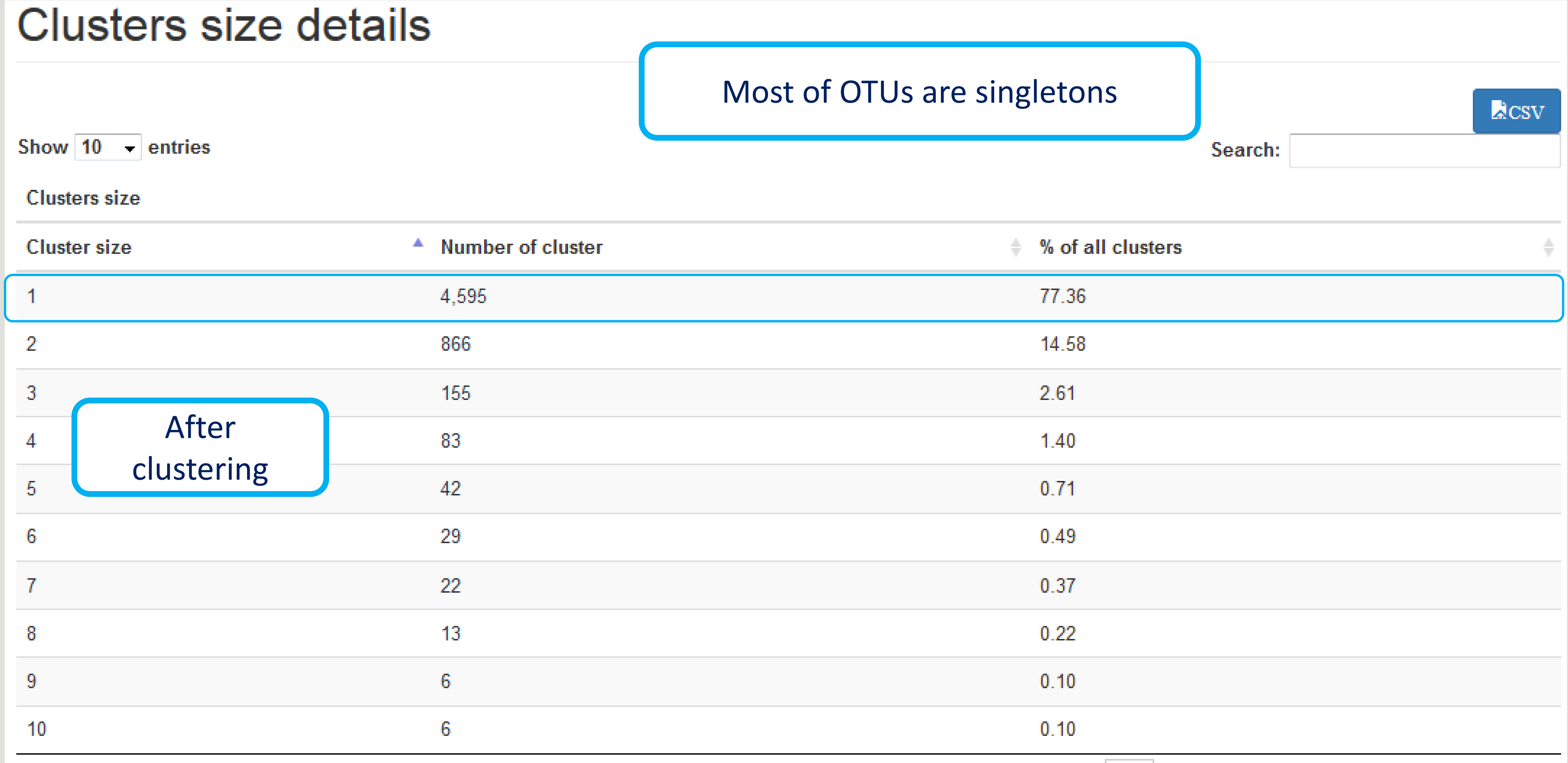
Sequences 81,838

After filtering little OTUs

Clusters size summary

Clusters size distribution (decile)

Decile	Value
Min	5
1	6
2	8
3	30
4	70
Median	112
6	145
7	225
8	412
9	994
Max	13,337



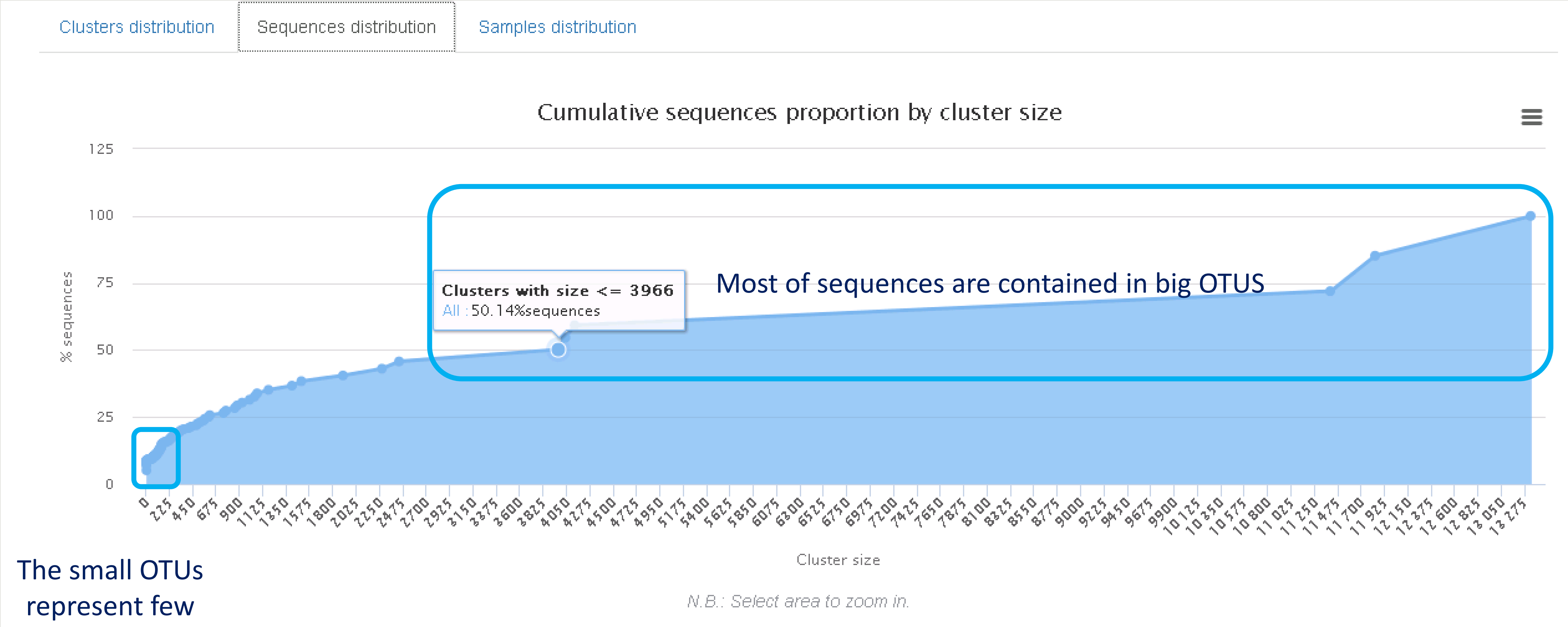
Clusters size details

Most of OTUs are singletons

Clusters size

Cluster size	Number of cluster	% of all clusters
1	4,595	77.36
2	866	14.58
3	155	2.61
4	83	1.40
5	42	0.71
6	29	0.49
7	22	0.37
8	13	0.22
9	6	0.10
10	6	0.10

After clustering



Clusters distribution | Sequences distribution | Samples distribution

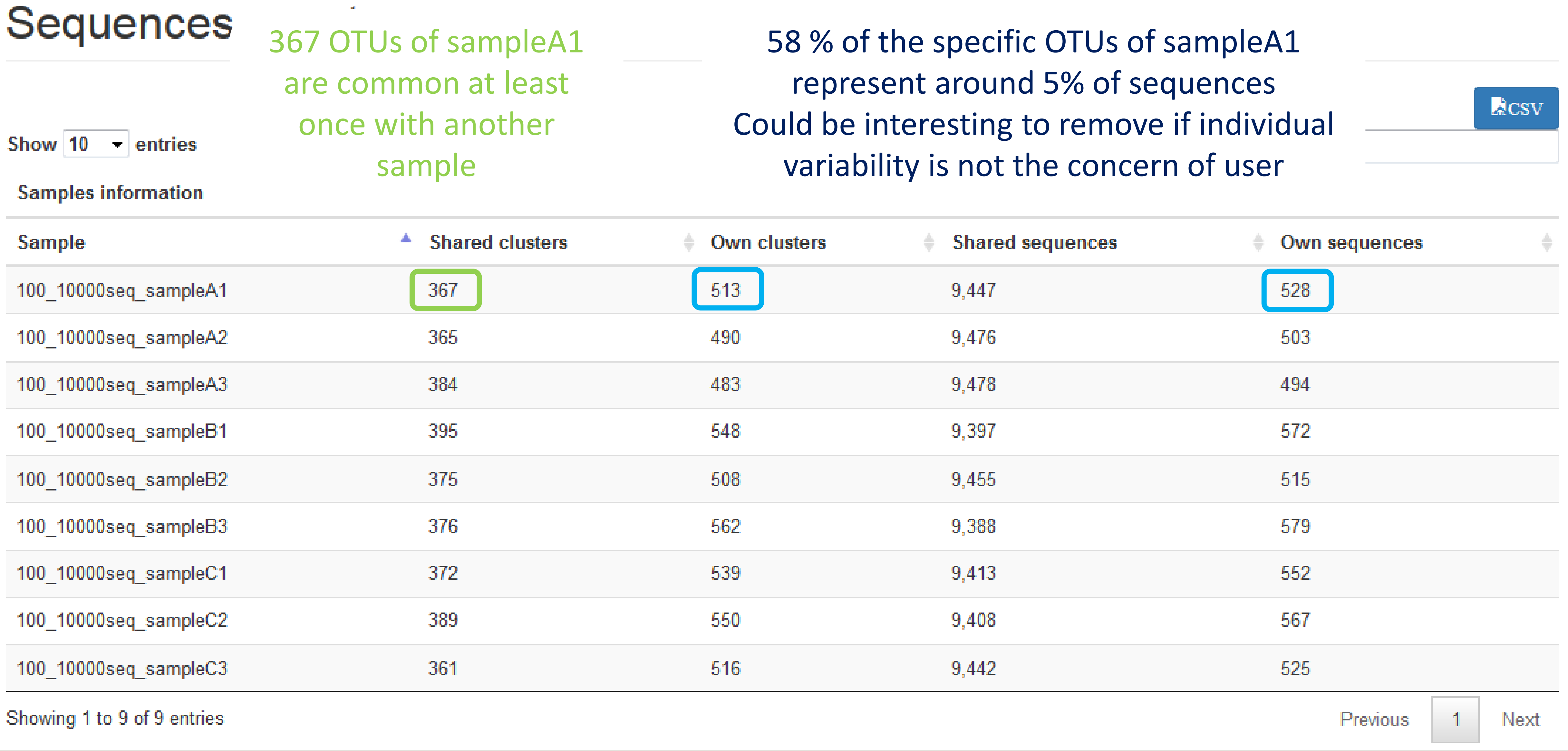
Cumulative sequences proportion by cluster size

Clusters with size <= 3966
All : 50.14%sequences

Most of sequences are contained in big OTUS

N.B.: Select area to zoom in.

The small OTUs represent few sequences



Sequences

367 OTUs of sampleA1 are common at least once with another sample

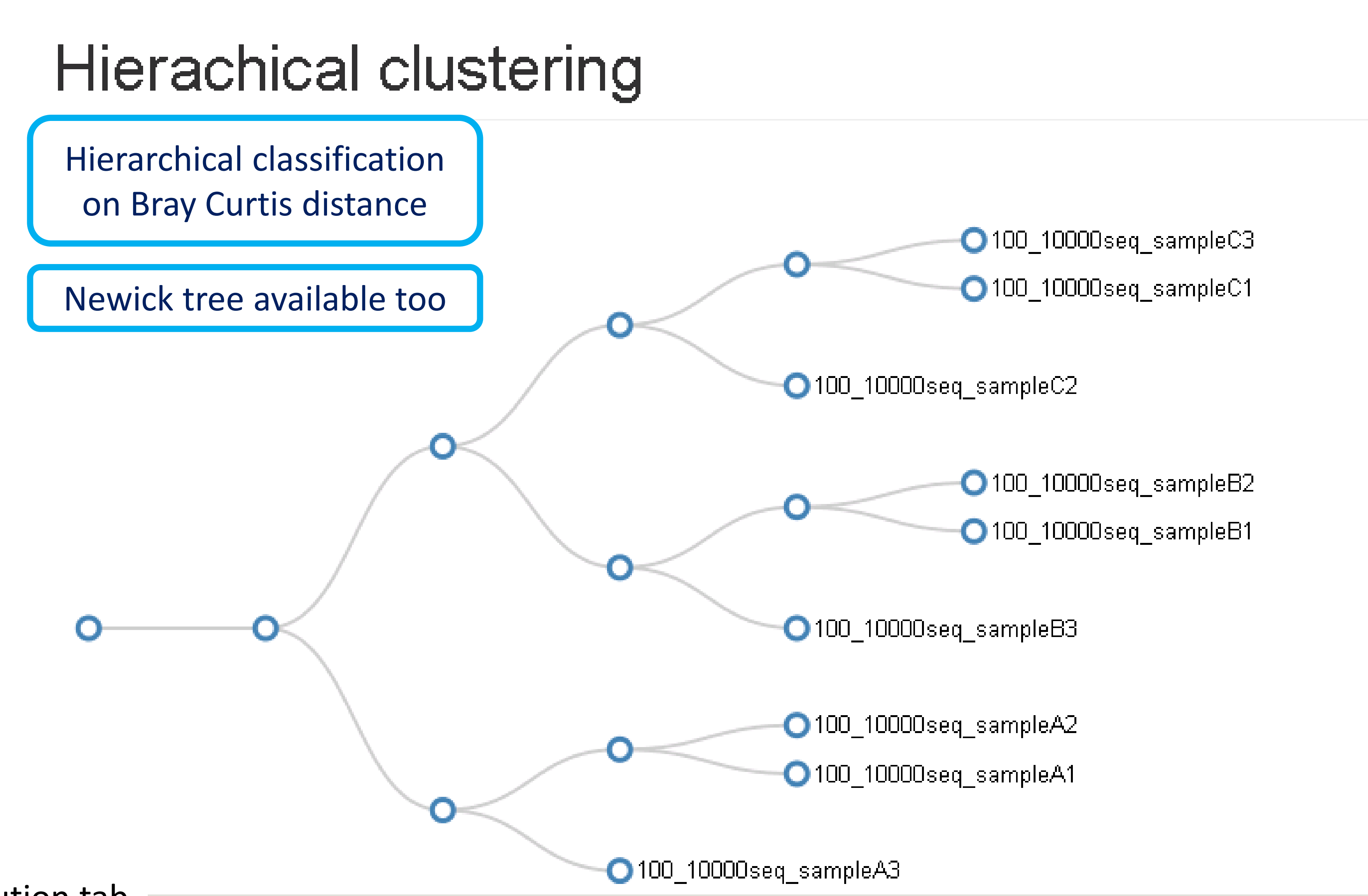
58 % of the specific OTUs of sampleA1 represent around 5% of sequences
Could be interesting to remove if individual variability is not the concern of user

Show 10 entries

Samples information

Sample	Shared clusters	Own clusters	Shared sequences	Own sequences
100_10000seq_sampleA1	367	513	9,447	528
100_10000seq_sampleA2	365	490	9,476	503
100_10000seq_sampleA3	384	483	9,478	494
100_10000seq_sampleB1	395	548	9,397	572
100_10000seq_sampleB2	375	508	9,455	515
100_10000seq_sampleB3	376	562	9,388	579
100_10000seq_sampleC1	372	539	9,413	552
100_10000seq_sampleC2	389	550	9,408	567
100_10000seq_sampleC3	361	516	9,442	525

Showing 1 to 9 of 9 entries



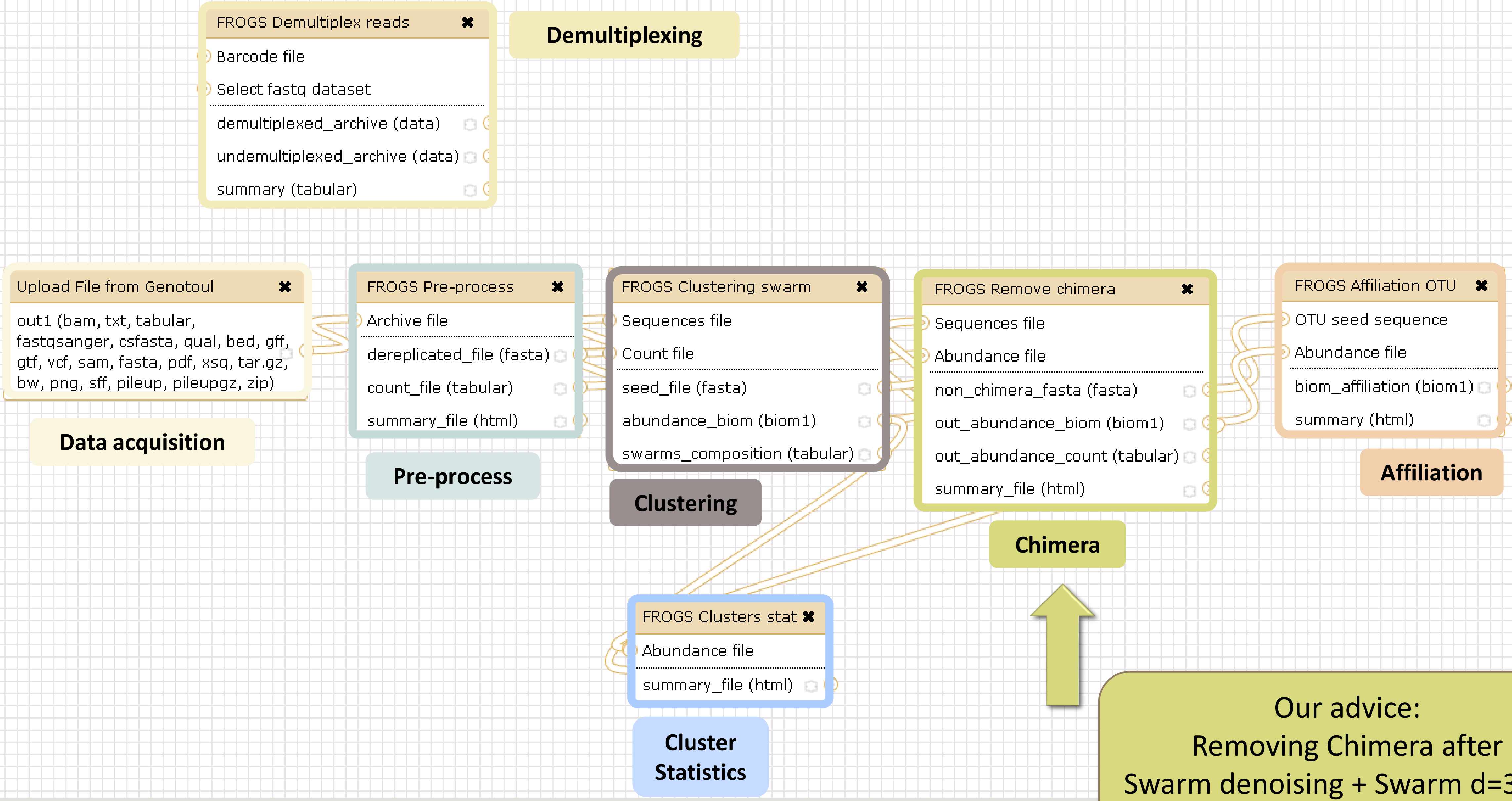
Hierachical clustering

Hierarchical classification on Bray Curtis distance

Newick tree available too

Samples distribution tab

Chimera removal tool



FROGS Demultiplex reads ✖
Barcode file
Select fastq dataset
demultiplexed_archive (data)
undemultiplexed_archive (data)
summary (tabular)

Demultiplexing

Upload File from Genotoul ✖
out1 (bam, txt, tabular, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png, sff, pileup, pileupgz, zip)

Data acquisition

FROGS Pre-process ✖
Archive file
dereplicated_file (fasta)
count_file (tabular)
summary_file (html)

Pre-process

FROGS Clustering swarm ✖
Sequences file
Count file
seed_file (fasta)
abundance_biom (biom1)
swarms_composition (tabular)

Clustering

FROGS Remove chimera ✖
Sequences file
Abundance file
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
out_abundance_count (tabular)
summary_file (html)

Chimera

FROGS Affiliation OTU ✖
OTU seed sequence
Abundance file
biom_affiliation (biom1)
summary (html)

Affiliation

FROGS Clusters stat ✖
Abundance file
summary_file (html)

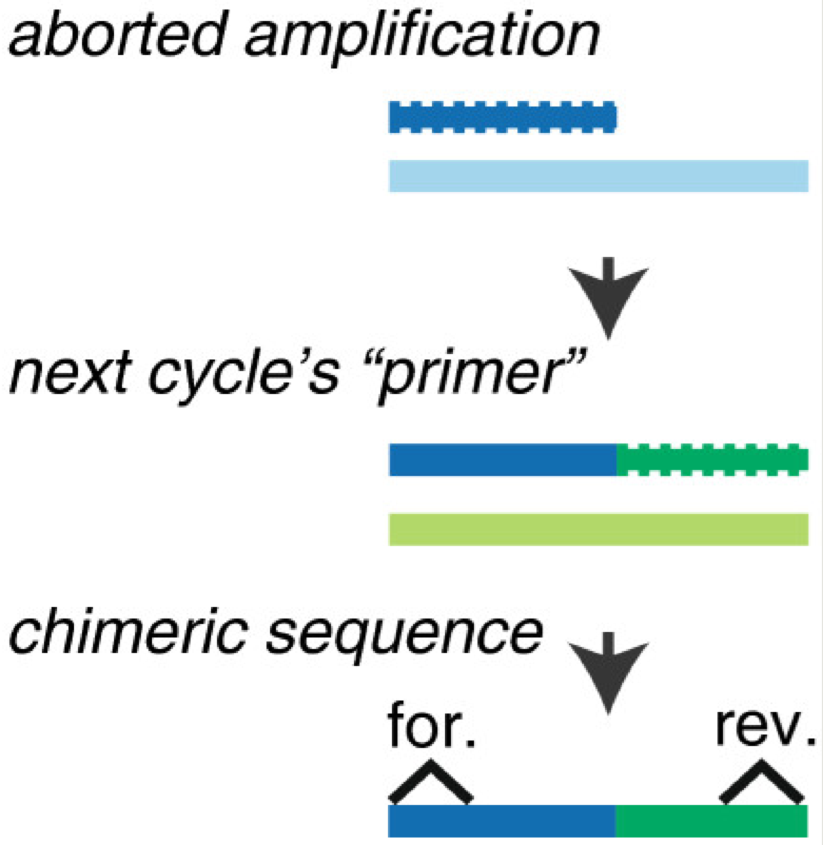
Cluster Statistics

Our advice:
Removing Chimera after
Swarm denoising + Swarm d=3, for
saving time without sensitivity loss

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.

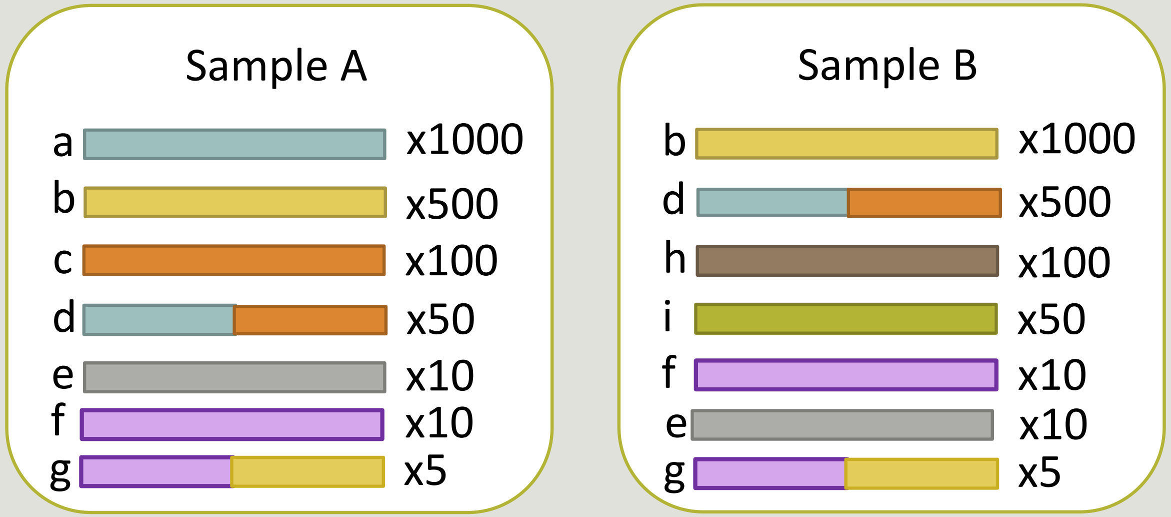
Chimera: from 5 to 45% of reads (Schloss 2011)



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



Sample A

a x1000
b x500
c x100
d x50
e x10
f x10
g x5

Sample B

b x1000
d x500
h x100
i x50
f x10
e x10
g x5

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

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

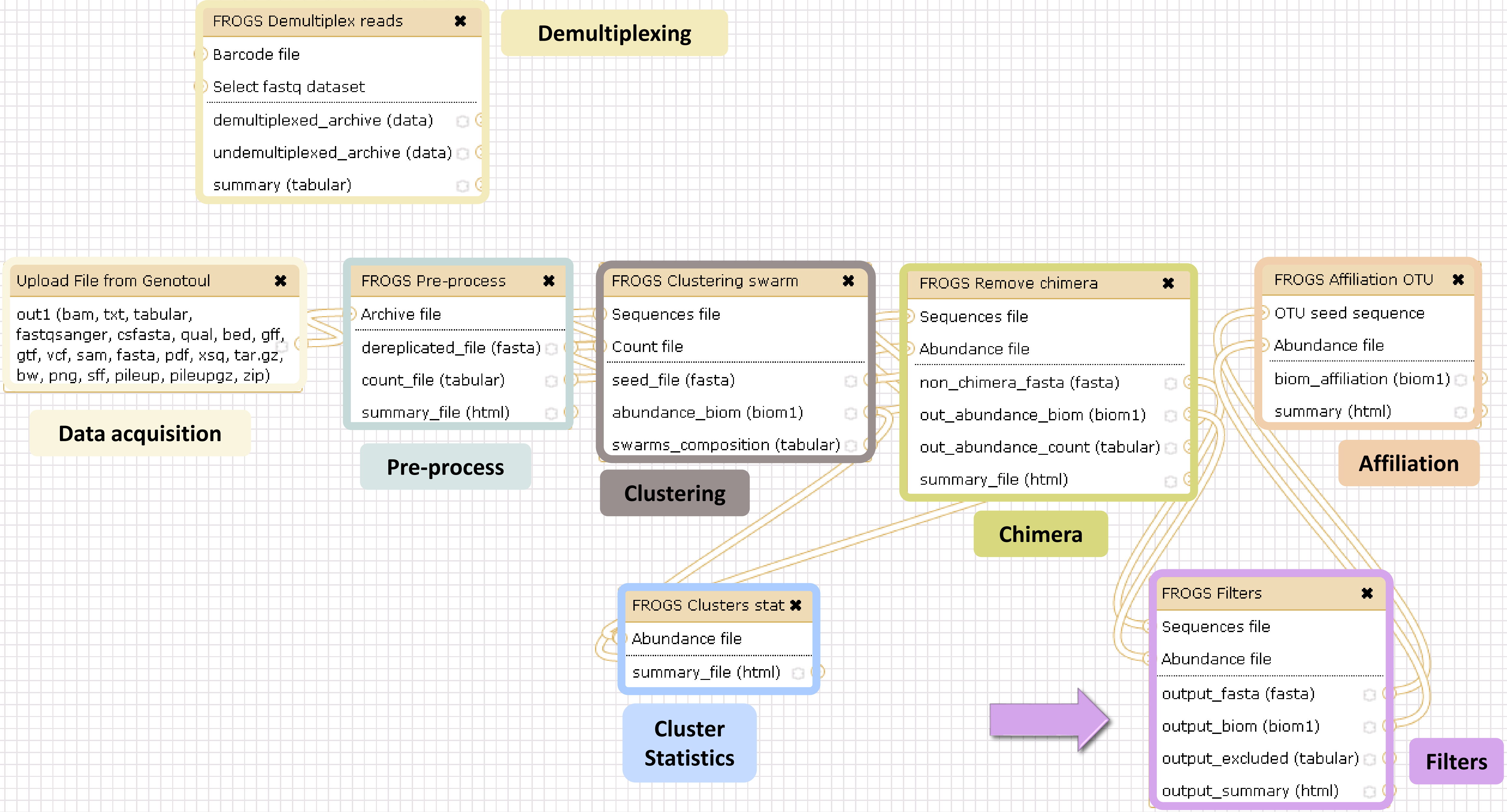
⇒ For FROGS “d” is not a chimera
⇒ For FROGS “g” is a chimera, “g” is removed
⇒ FROGS increases the detection specificity

Your Turn! - 5

Go to practice

EXERCISE 5

Filters tool



FROGS Demultiplex reads ✖

Barcode file

Select fastq dataset

demultiplexed_archive (data)

undemultiplexed_archive (data)

summary (tabular)

Upload File from Genotoul ✖

out1 (bam, txt, tabular, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png, sff, pileup, pileupgz, zip)

FROGS Pre-process ✖

Archive file

dereplicated_file (fasta)

count_file (tabular)

summary_file (html)

FROGS Clustering swarm ✖

Sequences file

Count file

seed_file (fasta)

abundance_biom (biom1)

swarms_composition (tabular)

FROGS Remove chimera ✖

Sequences file

Abundance file

non_chimera_fasta (fasta)

out_abundance_biom (biom1)

out_abundance_count (tabular)

summary_file (html)

FROGS Affiliation OTU ✖

OTU seed sequence

Abundance file

biom_affiliation (biom1)

summary (html)

FROGS Clusters stat ✖

Abundance file

summary_file (html)

FROGS Filters ✖

Sequences file

Abundance file

output_fasta (fasta)

output_biom (biom1)

output_excluded (tabular)

output_summary (html)

Affiliation runs long time

Advise:

Apply filters between “Chimera Removal ” and “Affiliation”.
Remove OTUs with weak abundance and non redundant before affiliation.

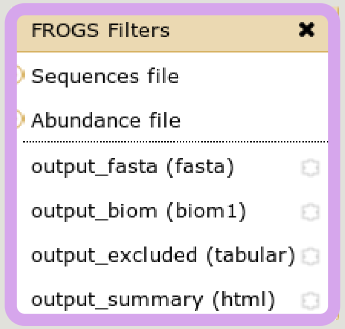
You will gain time !

Filters

Filters allows to filter the result thanks to different criteria et may be used after different steps of pipeline :

- On the abundance
- On RDP affiliation
- On Blast affiliation
- On phix contaminant

After Affiliation tool



FROGS Filters

Sequences file

Abundance file

output_fasta (fasta)

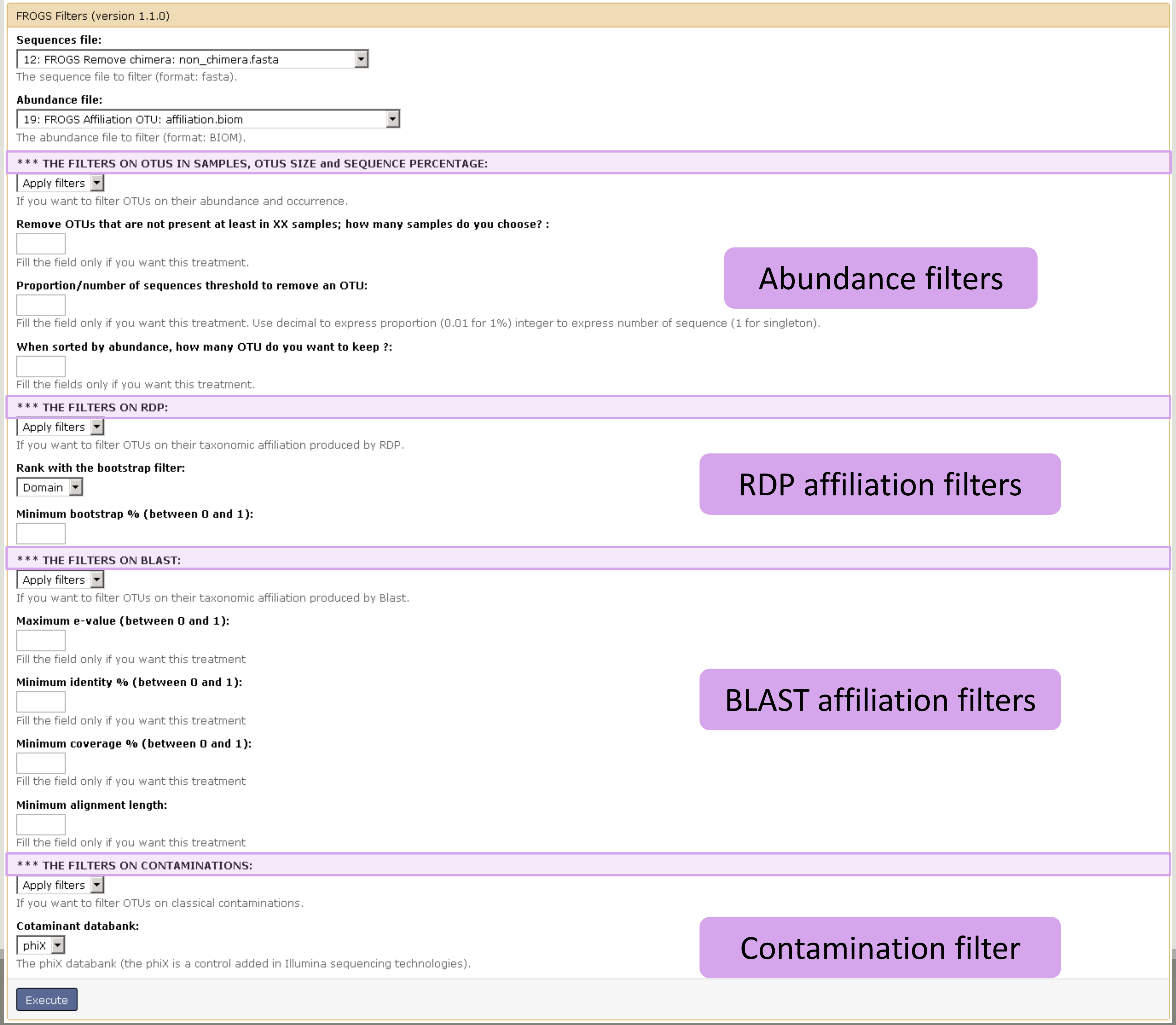
output_biom (biom1)

output_excluded (tabular)

output_summary (html)

Filters

4 filter sections



FROGS Filters (version 1.1.0)

Sequences file:

12: FROGS Remove chimera: non_chimera.fasta

The sequence file to filter (format: fasta).

Abundance file:

19: FROGS Affiliation OTU: affiliation.biom

The abundance file to filter (format: BIOM).

*** THE FILTERS ON OTUS IN SAMPLES, OTUS SIZE and SEQUENCE PERCENTAGE:

Apply filters

If you want to filter OTUs on their abundance and occurrence.

Remove OTUs that are not present at least in XX samples; how many samples do you choose? :

Fill the field only if you want this treatment.

Proportion/number of sequences threshold to remove an OTU:

Fill the field only if you want this treatment. Use decimal to express proportion (0.01 for 1%) integer to express number of sequence (1 for singleton).

When sorted by abundance, how many OTU do you want to keep ?:

Fill the fields only if you want this treatment.

Abundance filters

*** THE FILTERS ON RDP:

Apply filters

If you want to filter OTUs on their taxonomic affiliation produced by RDP.

Rank with the bootstrap filter:

Domain

Minimum bootstrap % (between 0 and 1):

RDP affiliation filters

*** THE FILTERS ON BLAST:

Apply filters

If you want to filter OTUs on their taxonomic affiliation produced by Blast.

Maximum e-value (between 0 and 1):

Fill the field only if you want this treatment

Minimum identity % (between 0 and 1):

Fill the field only if you want this treatment

Minimum coverage % (between 0 and 1):

Fill the field only if you want this treatment

Minimum alignment length:

Fill the field only if you want this treatment

BLAST affiliation filters

*** THE FILTERS ON CONTAMINATIONS:

Apply filters

If you want to filter OTUs on classical contaminations.

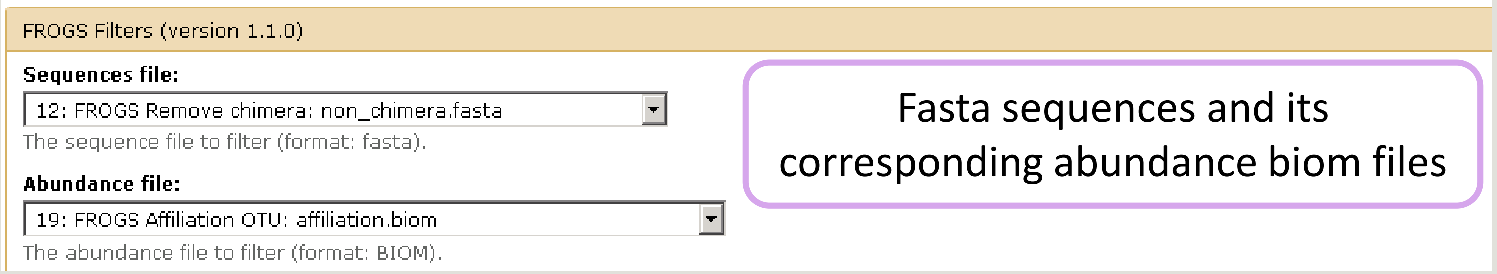
Cotaminant databank:

phiX

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

Contamination filter

Input



FROGS Filters (version 1.1.0)

Sequences file:

12: FROGS Remove chimera: non_chimera.fasta

The sequence file to filter (format: fasta).

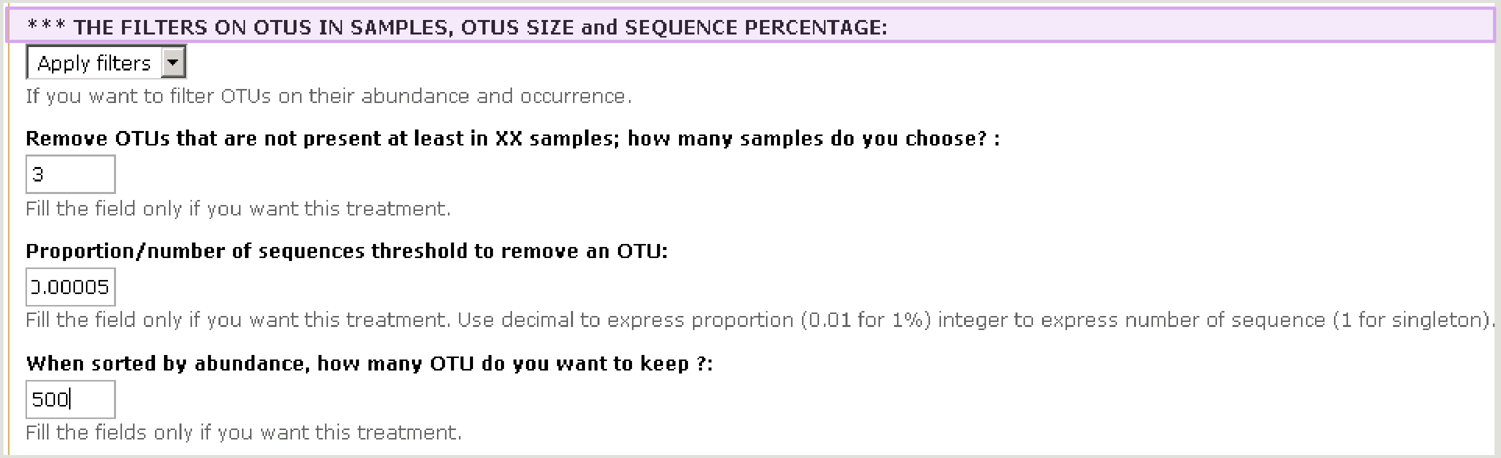
Abundance file:

19: FROGS Affiliation OTU: affiliation.biom

The abundance file to filter (format: BIOM).

Fasta sequences and its corresponding abundance biom files

Filter 1 : abundance



***** THE FILTERS ON OTUS IN SAMPLES, OTUS SIZE and SEQUENCE PERCENTAGE:**

Apply filters

If you want to filter OTUs on their abundance and occurrence.

Remove OTUs that are not present at least in XX samples; how many samples do you choose? :

3

Fill the field only if you want this treatment.

Proportion/number of sequences threshold to remove an OTU:

0.00005

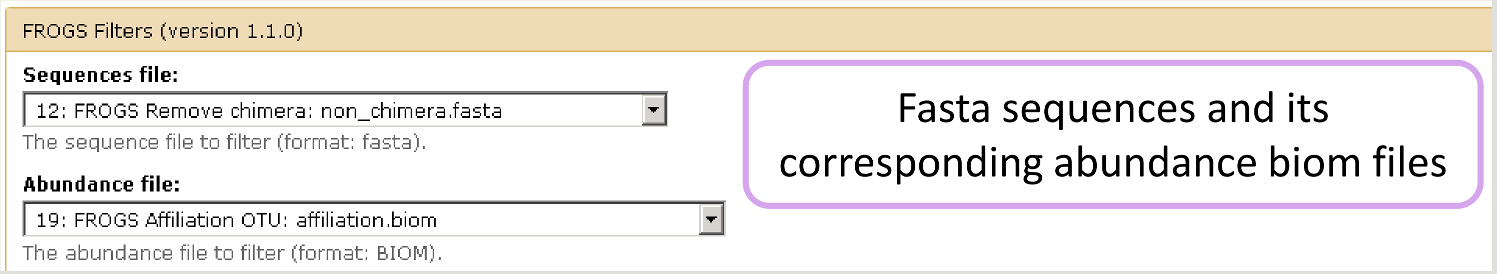
Fill the field only if you want this treatment. Use decimal to express proportion (0.01 for 1%) integer to express number of sequence (1 for singleton).

When sorted by abundance, how many OTU do you want to keep ?:

500

Fill the fields only if you want this treatment.

Input



FROGS Filters (version 1.1.0)

Sequences file:

12: FROGS Remove chimera: non_chimera.fasta

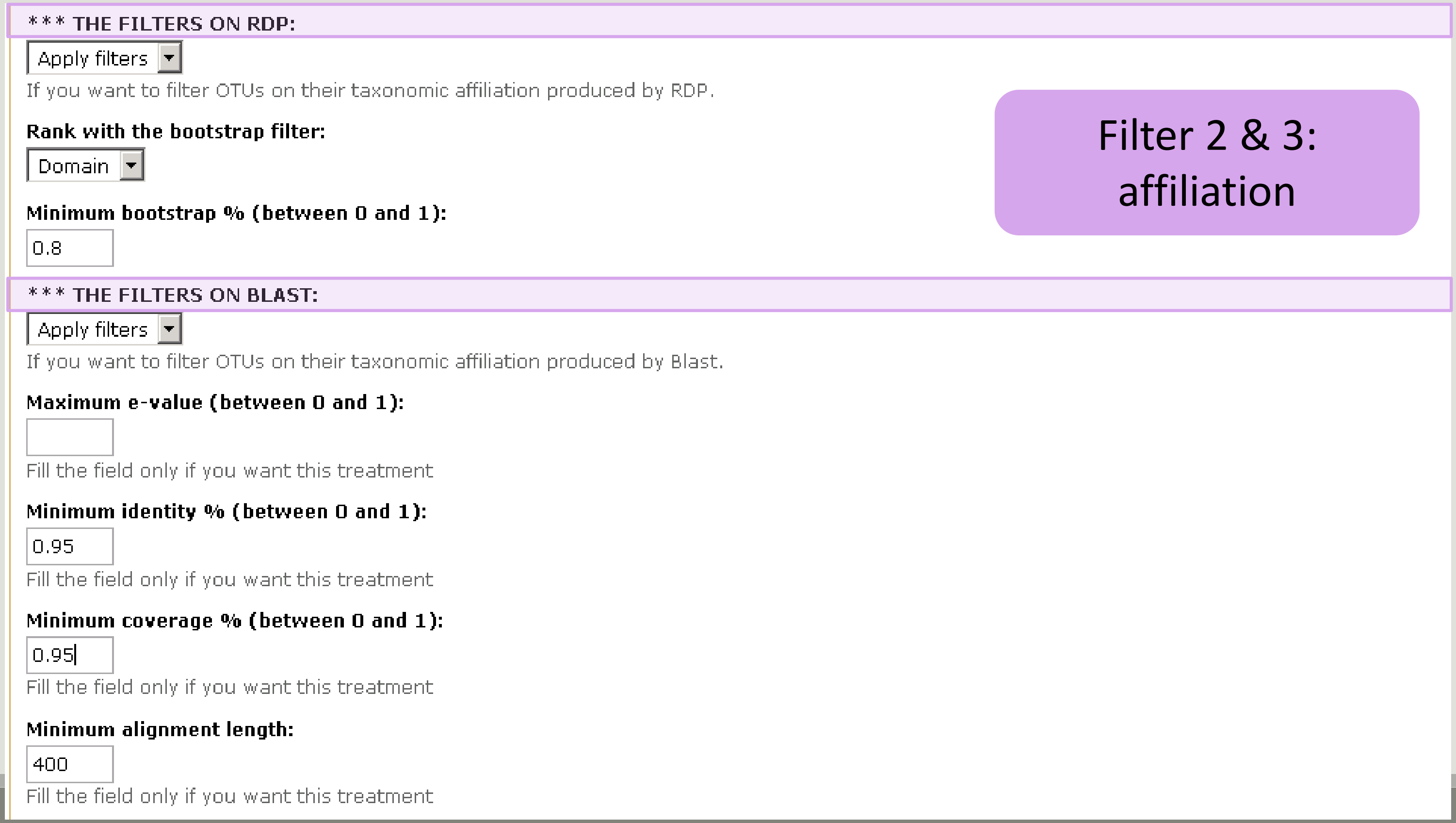
The sequence file to filter (format: fasta).

Abundance file:

19: FROGS Affiliation OTU: affiliation.biom

The abundance file to filter (format: BIOM).

Fasta sequences and its corresponding abundance biom files



*** THE FILTERS ON RDP:

Apply filters

If you want to filter OTUs on their taxonomic affiliation produced by RDP.

Rank with the bootstrap filter:

Domain

Minimum bootstrap % (between 0 and 1):

0.8

Filter 2 & 3: affiliation

*** THE FILTERS ON BLAST:

Apply filters

If you want to filter OTUs on their taxonomic affiliation produced by Blast.

Maximum e-value (between 0 and 1):

Fill the field only if you want this treatment

Minimum identity % (between 0 and 1):

0.95

Fill the field only if you want this treatment

Minimum coverage % (between 0 and 1):

0.95

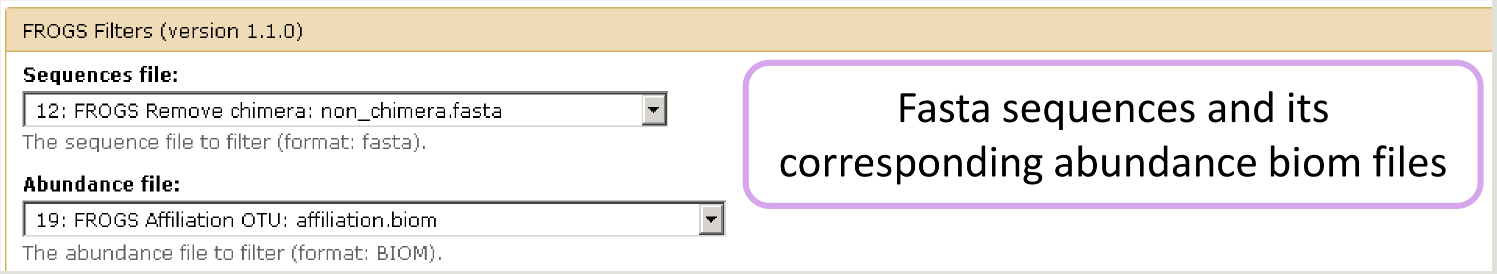
Fill the field only if you want this treatment

Minimum alignment length:

400

Fill the field only if you want this treatment

Input



FROGS Filters (version 1.1.0)

Sequences file:

12: FROGS Remove chimera: non_chimera.fasta

The sequence file to filter (format: fasta).

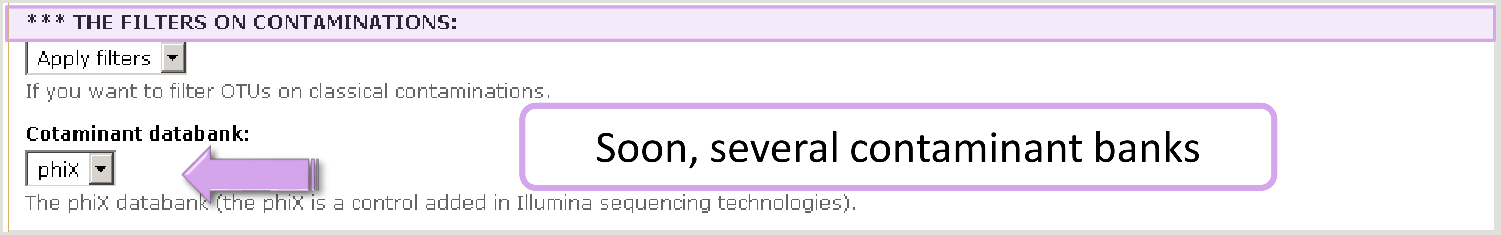
Abundance file:

19: FROGS Affiliation OTU: affiliation.biom

The abundance file to filter (format: BIOM).

Fasta sequences and its corresponding abundance biom files

Filter 4 : contamination



***** THE FILTERS ON CONTAMINATIONS:**

Apply filters

If you want to filter OTUs on classical contaminations.

Cotaminant databank:

phiX

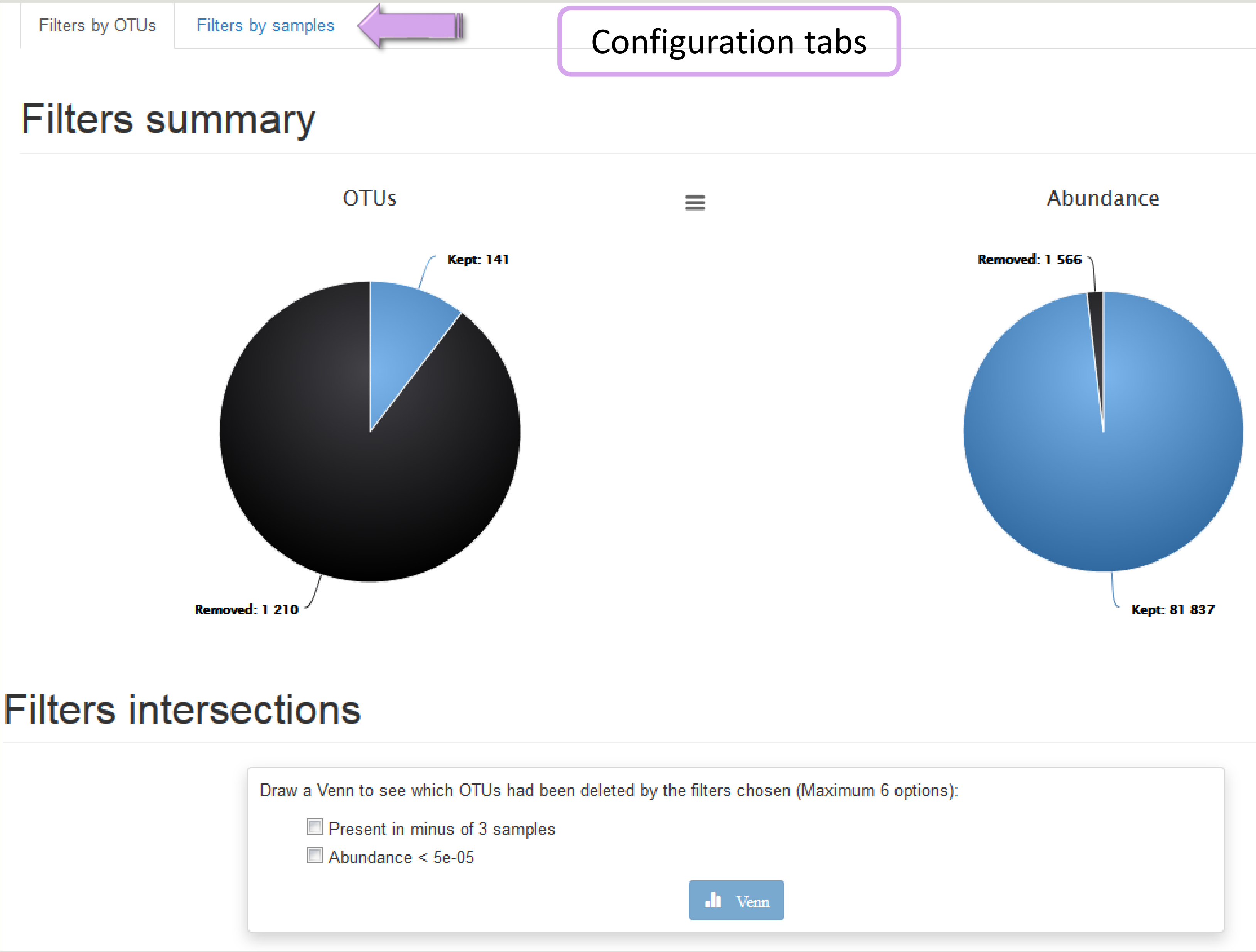
The phiX databank (the phiX is a control added in Illumina sequencing technologies).

Soon, several contaminant banks

Your Turn! - 6

Go to practice

EXERCISE 6



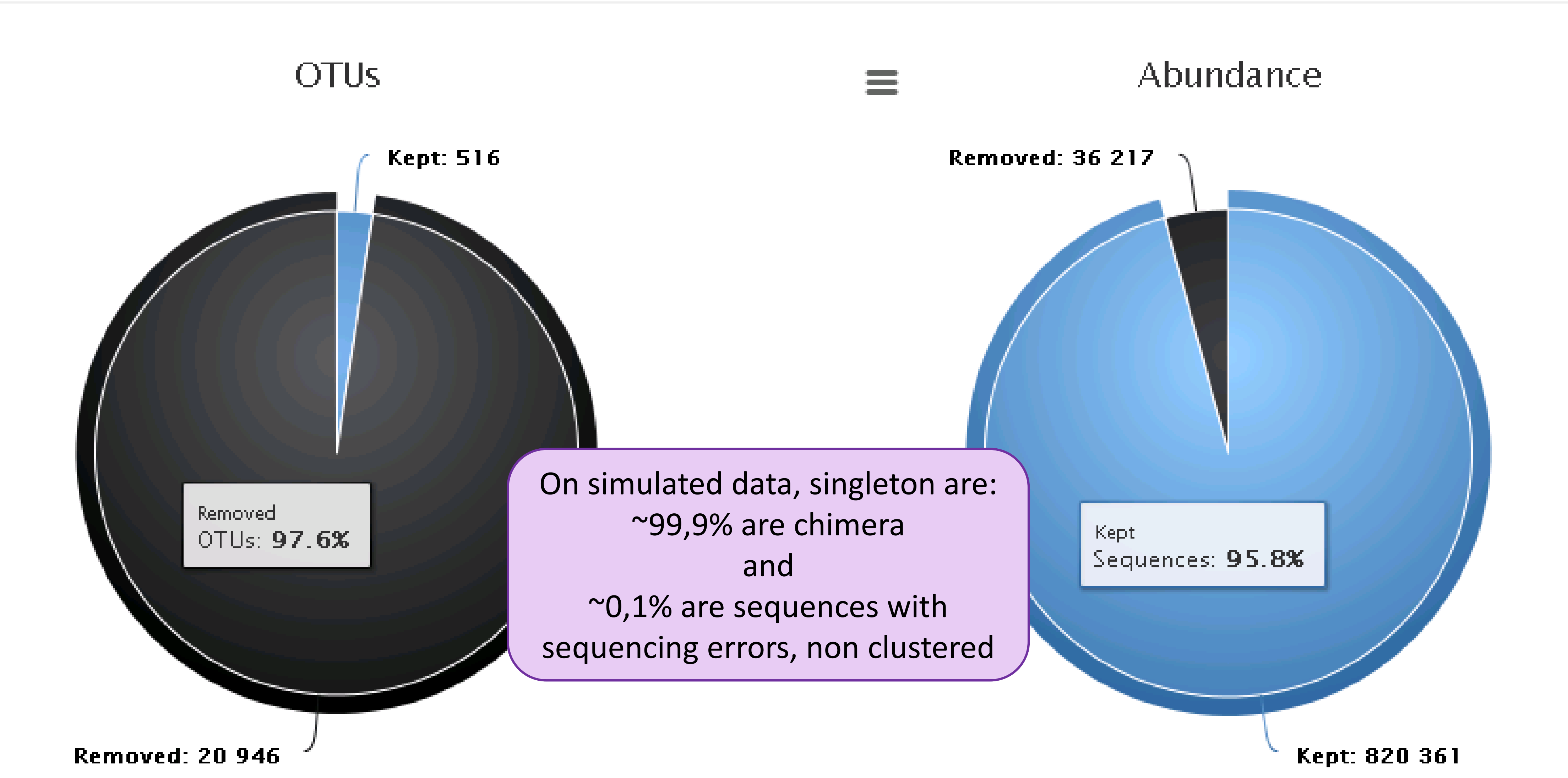
Configuration tabs

Filters summary

Filters intersections

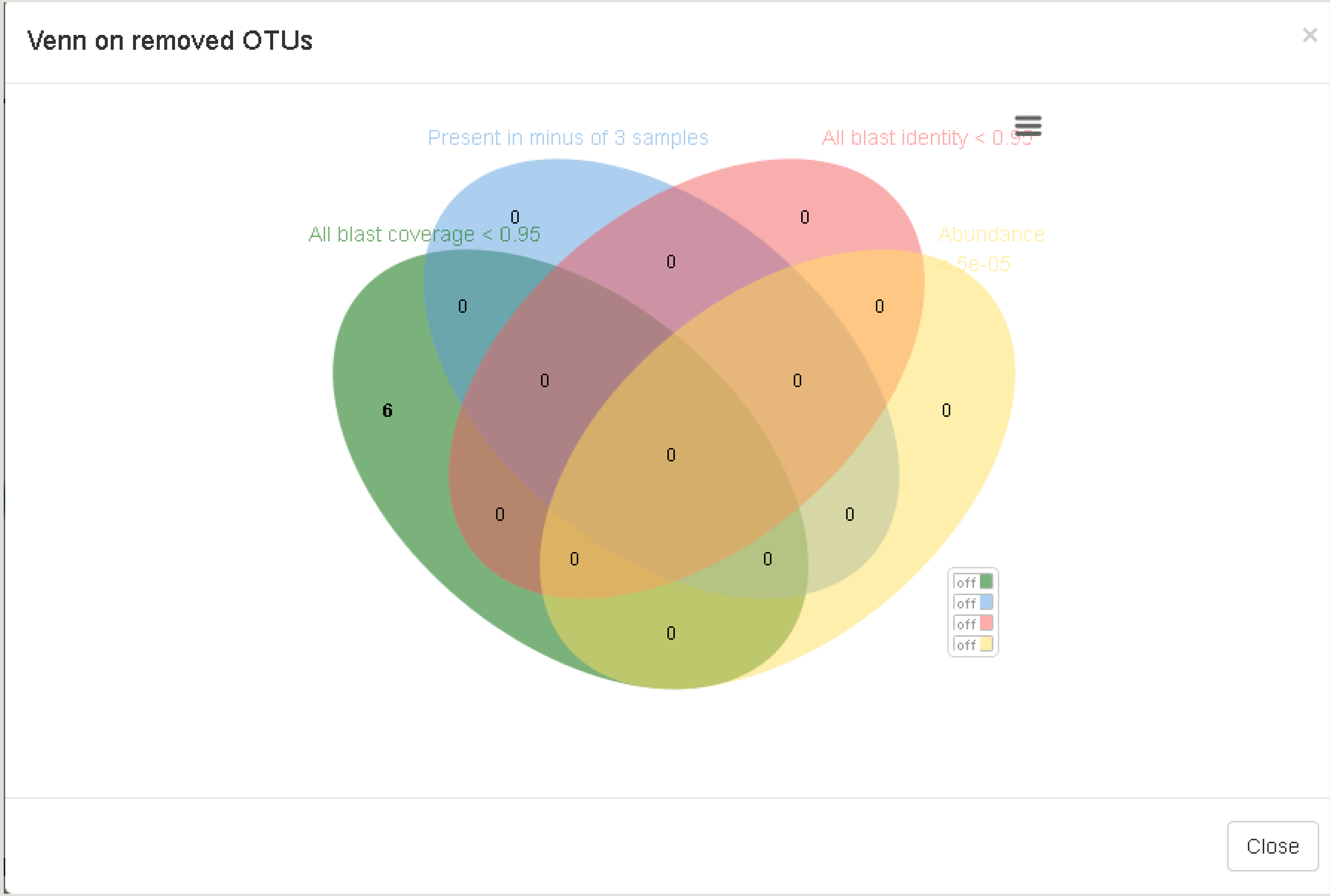
Draw a Venn to see which OTUs had been deleted by the filters chosen (Maximum 6 options):

- □ Present in minus of 3 samples
- □ Abundance < 5e-05



On simulated data, singleton are:
~99,9% are chimera
and
~0,1% are sequences with sequencing errors, non clustered

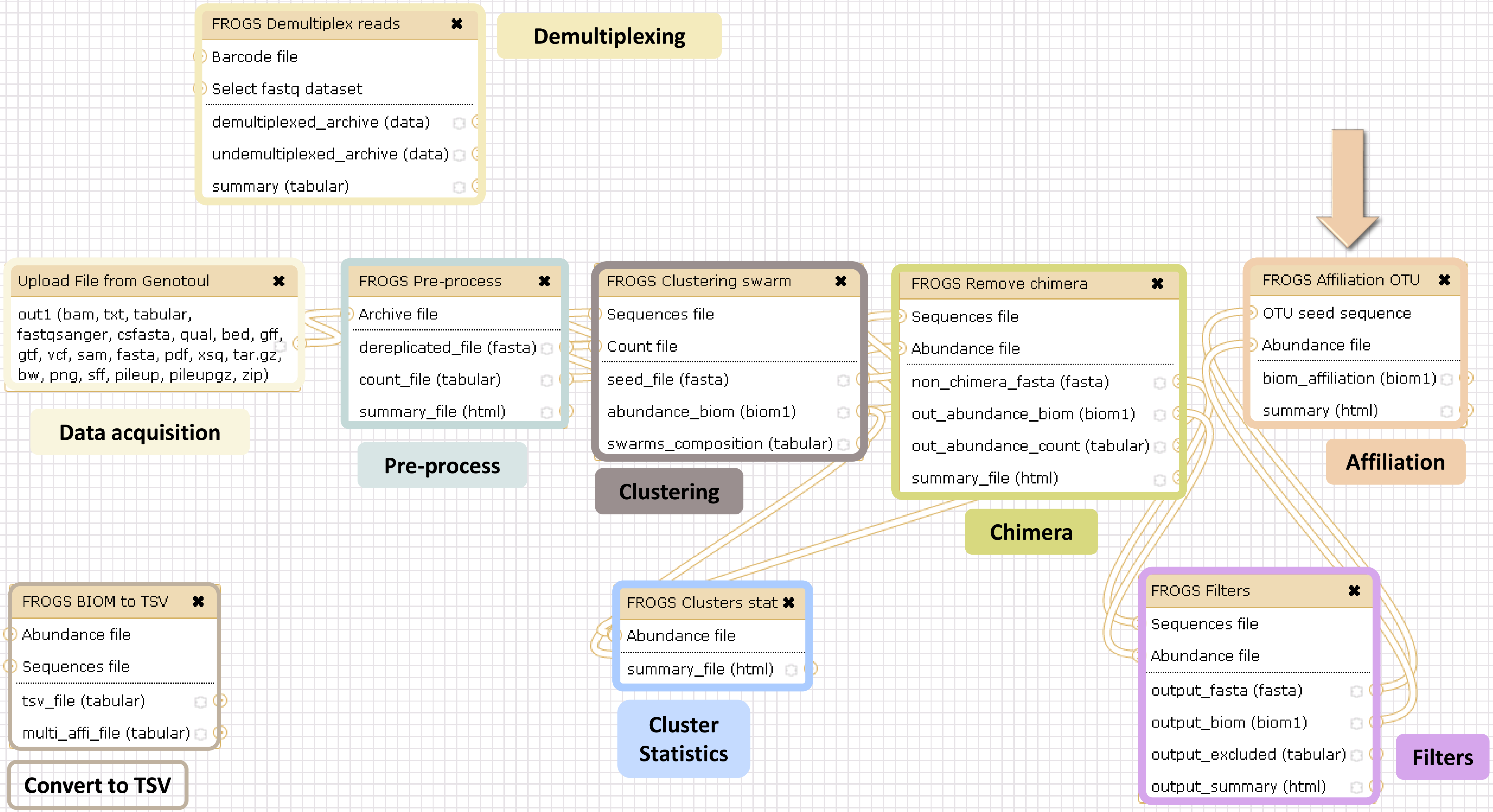
Removing little OTUs (conservation rate =0.005%) and non shared OTU (in less than 2 samples)



Venn on removed OTUs

×
Present in minus of 3 samples
All blast identity < 0.95
All blast coverage < 0.95
Abundance
< 5e-05
0
0
0
0
0
6
0
0
0
0
0
0
0
0
0
off 🟩
off 🟦
off 🟥
off 🟨
Close

Affiliation tool



Demultiplexing

Data acquisition

Pre-process

Clustering

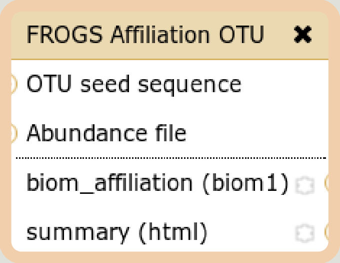
Chimera

Affiliation

Convert to TSV

Cluster Statistics

Filters



FROGS Affiliation OTU ✖

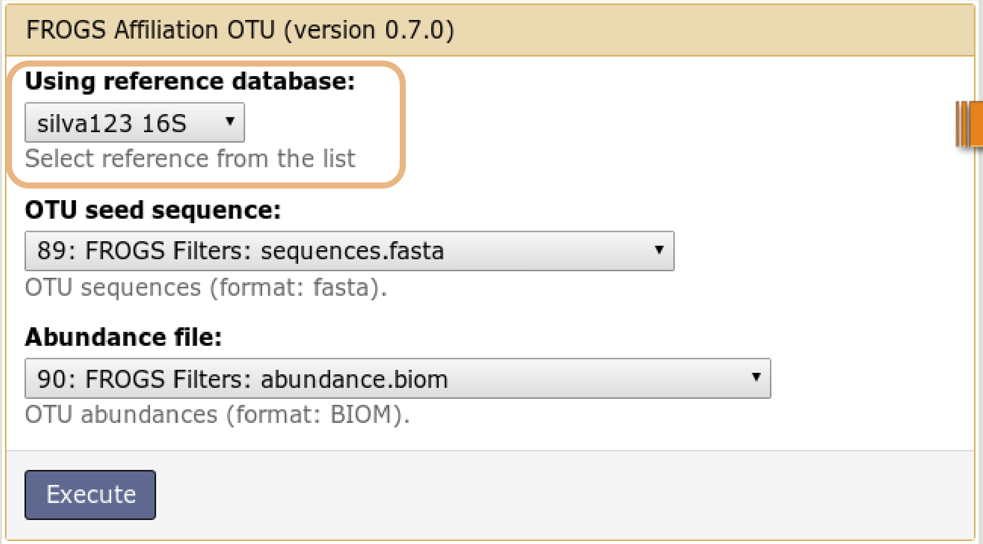
OTU seed sequence

Abundance file

biom_affiliation (biom1)

summary (html)

Affiliation



FROGS Affiliation OTU (version 0.7.0)

Using reference database:

silva123 16S ▾

Select reference from the list

OTU seed sequence:

89: FROGS Filters: sequences.fasta ▾

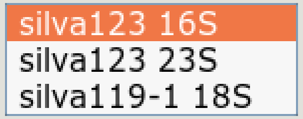
OTU sequences (format: fasta).

Abundance file:

90: FROGS Filters: abundance.biom ▾

OTU abundances (format: BIOM).

OR



silva123 16S
silva123 23S
silva119-1 18S

1 Cluster = 2 affiliations

Double Affiliation vs SILVA 123 (for 16S, 18S or 23S), SILVA 119 (for 18S) or Greengenes **with :**

1. RDPClassifier* (Ribosomal Database Project): one affiliation with bootstrap, on each taxonomic subdivision.

Bacteria(100);Firmicutes(100);Clostridia(100);Clostridiales(100);Lachnospiraceae(100);Pseudobutyrivibrio(80); Pseudobutyrivibrio xylanivorans (80)

2. NCBI Blastn+** : all identical Best Hits with identity %, coverage %, e-value, alignment length and a special tag "**Multi-affiliation".**

Bacteria;Firmicutes;Clostridia;Clostridiales;Lachnospiraceae;Pseudobutyrivibrio;Pseudobutyrivibrio ruminis; Pseudobutyrivibrio xylanivorans

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

V3 – V4	Bacteria|Firmicutes|Clostridia|Clostridiales|Lachnospiraceae|Pseudobutyrivibrio|16S unknown species
V3 – V4	Bacteria|Firmicutes|Clostridia|Clostridiales|Lachnospiraceae|Pseudobutyrivibrio|16S Butyrivibrio fibrisolvens
V3 – V4	Bacteria|Firmicutes|Clostridia|Clostridiales|Lachnospiraceae|Pseudobutyrivibrio|16S rumen bacterium 8|9293-9
V3 – V4	Bacteria|Firmicutes|Clostridia|Clostridiales|Lachnospiraceae|Pseudobutyrivibrio|16S Pseudobutyrivibrio xylanivorans
V3 – V4	Bacteria|Firmicutes|Clostridia|Clostridiales|Lachnospiraceae|Pseudobutyrivibrio|16S Pseudobutyrivibrio ruminis

5 identical blast best hits on SILVA 123 databank

Affiliation Strategy of FROGS

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

V3 – V4	Bacteria|Firmicutes|Clostridia|Clostridiales|Lachnospiraceae|Pseudobutyrivibrio|16S unknown species
V3 – V4	Bacteria|Firmicutes|Clostridia|Clostridiales|Lachnospiraceae|Pseudobutyrivibrio|16S Butyrivibrio fibrisolvens
V3 – V4	Bacteria|Firmicutes|Clostridia|Clostridiales|Lachnospiraceae|Pseudobutyrivibrio|16S rumen bacterium 8|9293-9
V3 – V4	Bacteria|Firmicutes|Clostridia|Clostridiales|Lachnospiraceae|Pseudobutyrivibrio|16S Pseudobutyrivibrio xylanivorans
V3 – V4	Bacteria|Firmicutes|Clostridia|Clostridiales|Lachnospiraceae|Pseudobutyrivibrio|16S Pseudobutyrivibrio ruminis

FROGS Affiliation: Bacteria|Firmicutes|Clostridia|Clostridiales|Lachnospiraceae|Pseudobutyrivibrio|**Multi-affiliation**

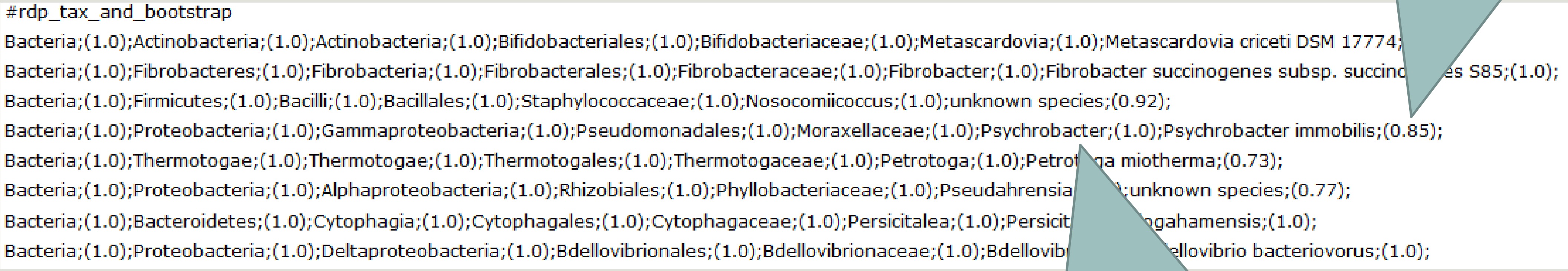
Your Turn! – 7

Go to practice

EXERCISE 7

1st column - RDP

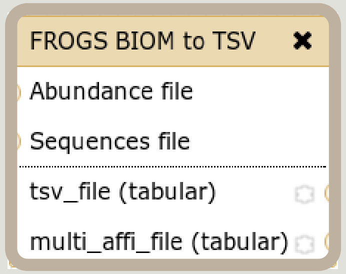
85% of RDP iterations have affiliated the sequence to the species « Psychrobacter immobilis »



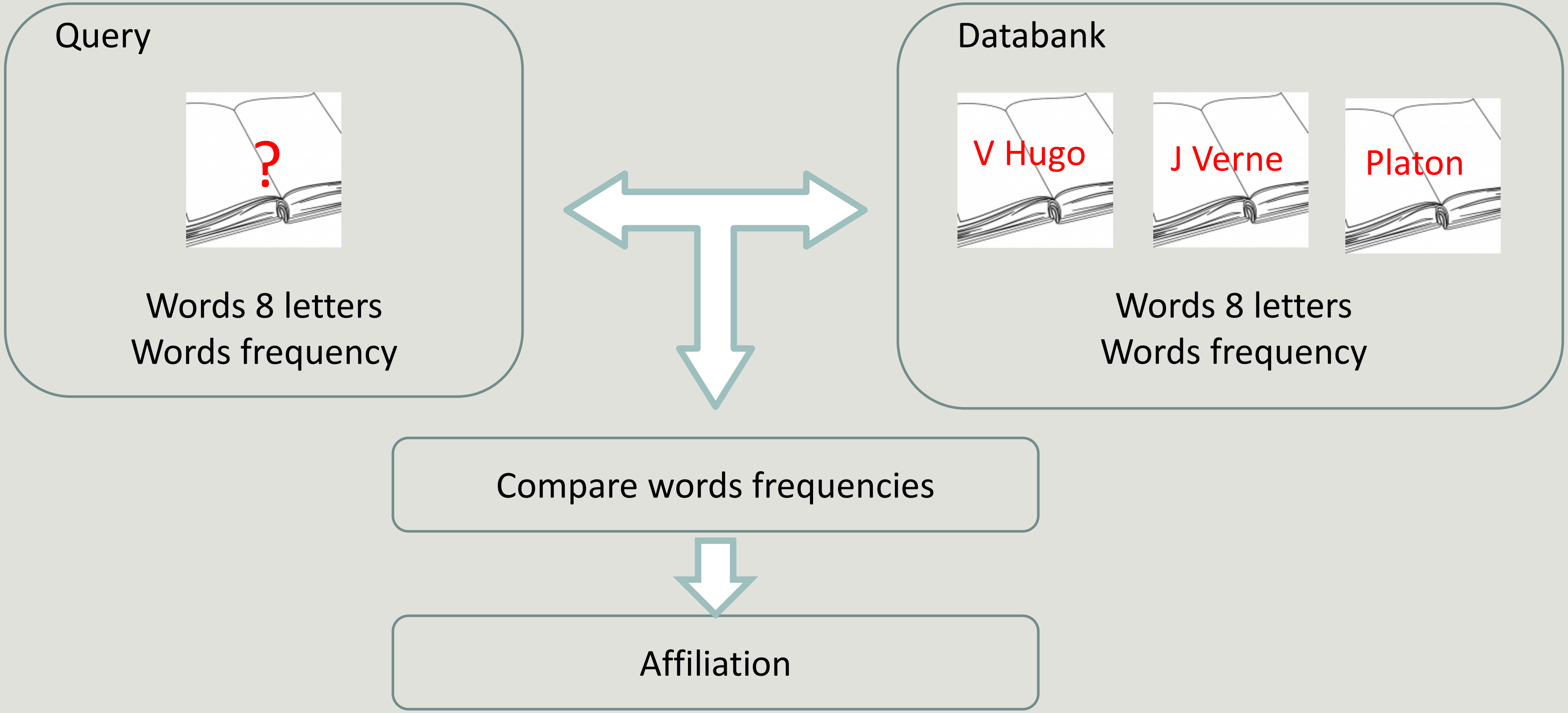
```
#rdp_tax_and_bootstrap
Bacteria;(1.0);Actinobacteria;(1.0);Actinobacteria;(1.0);Bifidobacteriales;(1.0);Bifidobacteriaceae;(1.0);Metascardovia;(1.0);Metascardovia criceti DSM 17774;
Bacteria;(1.0);Fibrobacteres;(1.0);Fibrobacteria;(1.0);Fibrobacterales;(1.0);Fibrobacteraceae;(1.0);Fibrobacter;(1.0);Fibrobacter succinogenes subsp. succin  es S85;(1.0);
Bacteria;(1.0);Firmicutes;(1.0);Bacilli;(1.0);Bacillales;(1.0);Staphylococcaceae;(1.0);Nosocomiicoccus;(1.0);unknown species;(0.92);
Bacteria;(1.0);Proteobacteria;(1.0);Gammaproteobacteria;(1.0);Pseudomonadales;(1.0);Moraxellaceae;(1.0);Psychrobacter;(1.0);Psychrobacter immobilis;(0.85);
Bacteria;(1.0);Thermotogae;(1.0);Thermotogae;(1.0);Thermotogales;(1.0);Thermotogaceae;(1.0);Petrotoga;(1.0);Petro  ga miotherma;(0.73);
Bacteria;(1.0);Proteobacteria;(1.0);Alphaproteobacteria;(1.0);Rhizobiales;(1.0);Phyllobacteriaceae;(1.0);Pseudahrensia  );unknown species;(0.77);
Bacteria;(1.0);Bacteroidetes;(1.0);Cytophagia;(1.0);Cytophagales;(1.0);Cytophagaceae;(1.0);Persicitalea;(1.0);Persicit  ngahamensis;(1.0);
Bacteria;(1.0);Proteobacteria;(1.0);Deltaproteobacteria;(1.0);Bdellovibrionales;(1.0);Bdellovibrionaceae;(1.0);Bdellovib  ellovibrio bacteriovorus;(1.0);
```

100% of RDP iterations have affiliated the sequence to the genus « Psychrobacter ». Bootstrap values are between 0 and 1

Convert to TSV



How works RDP ?



Query

Words 8 letters
Words frequency

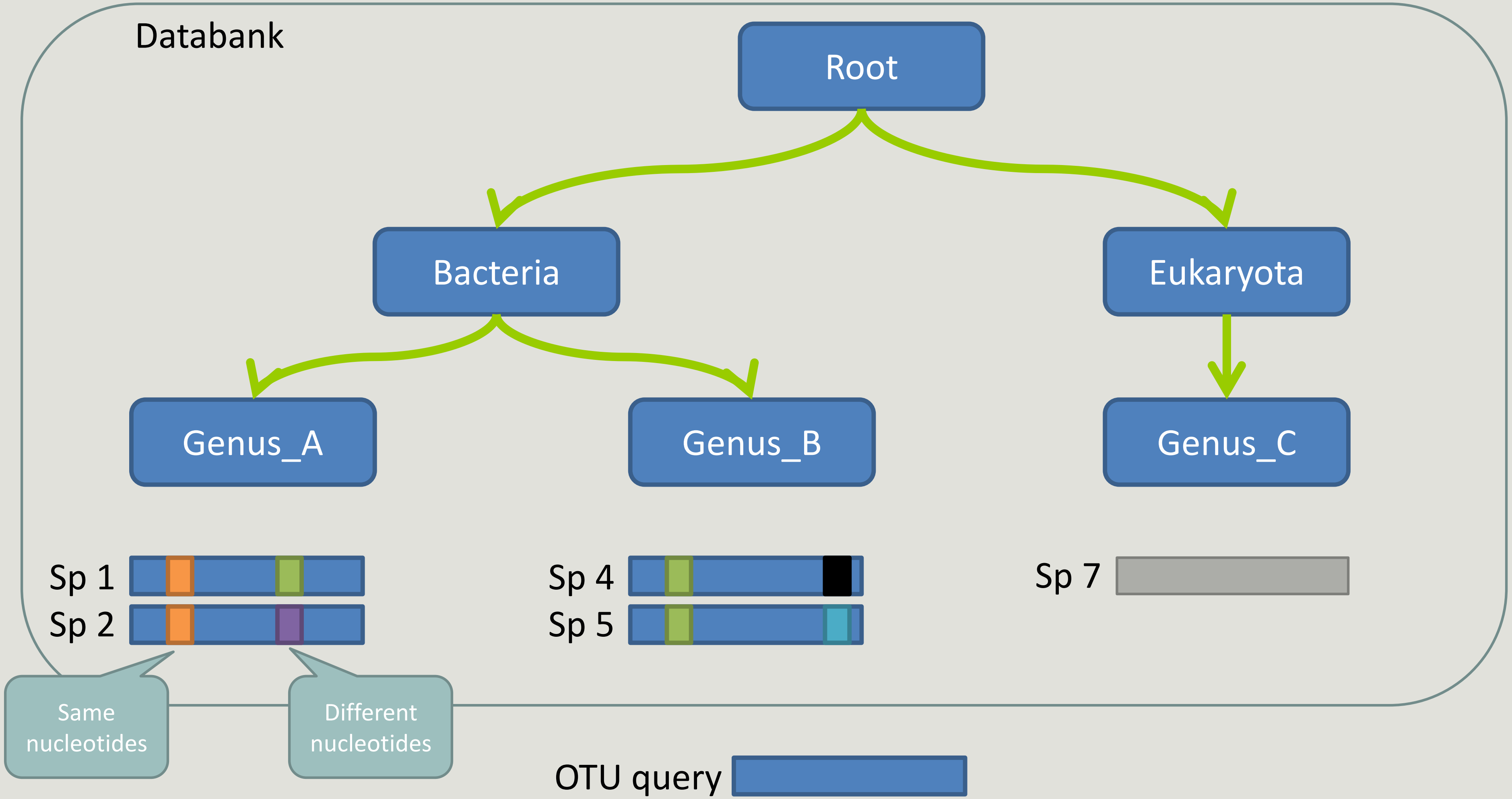
Databank

Words 8 letters
Words frequency

Compare words frequencies

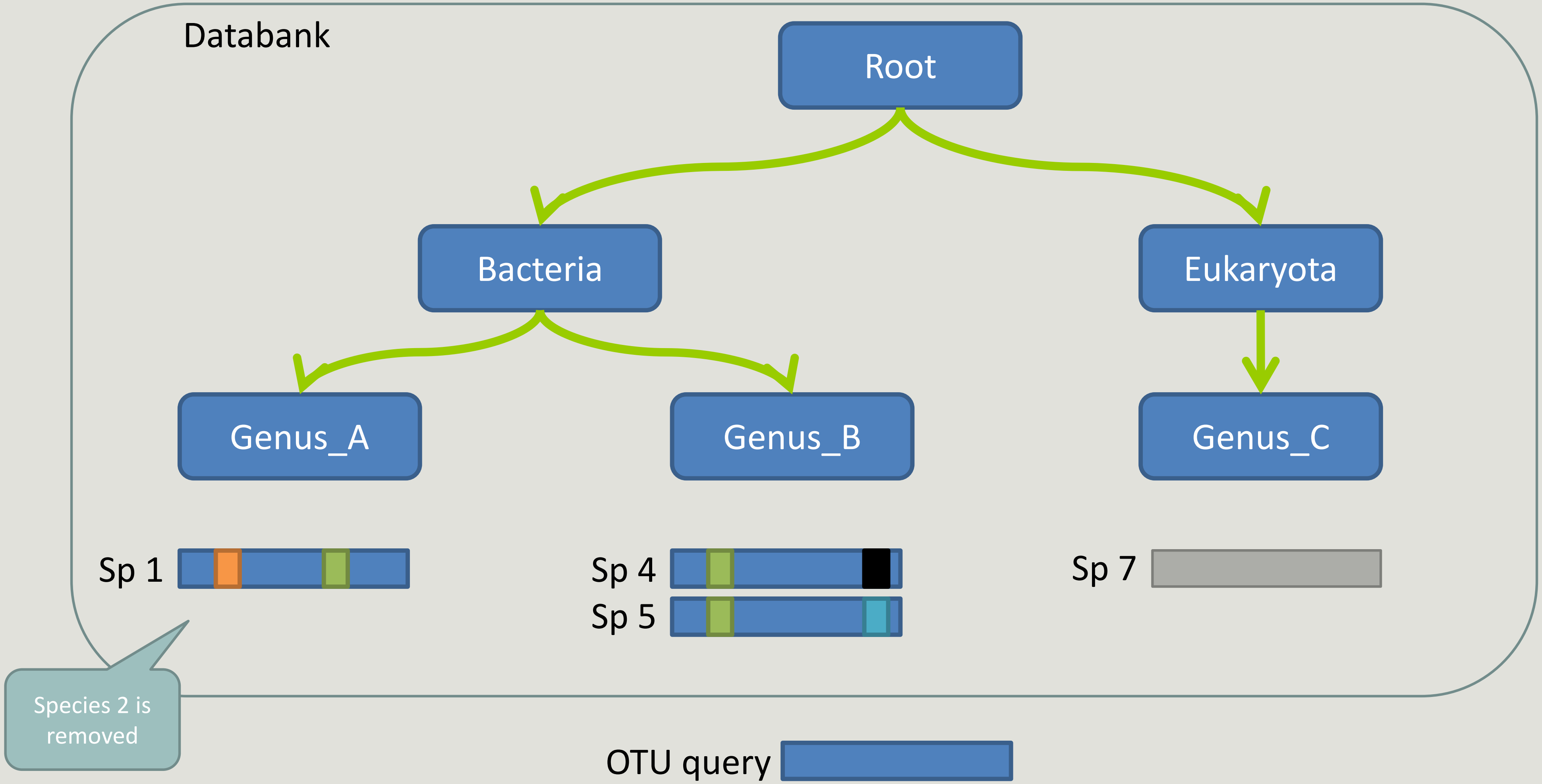
Affiliation

How works RDP ?



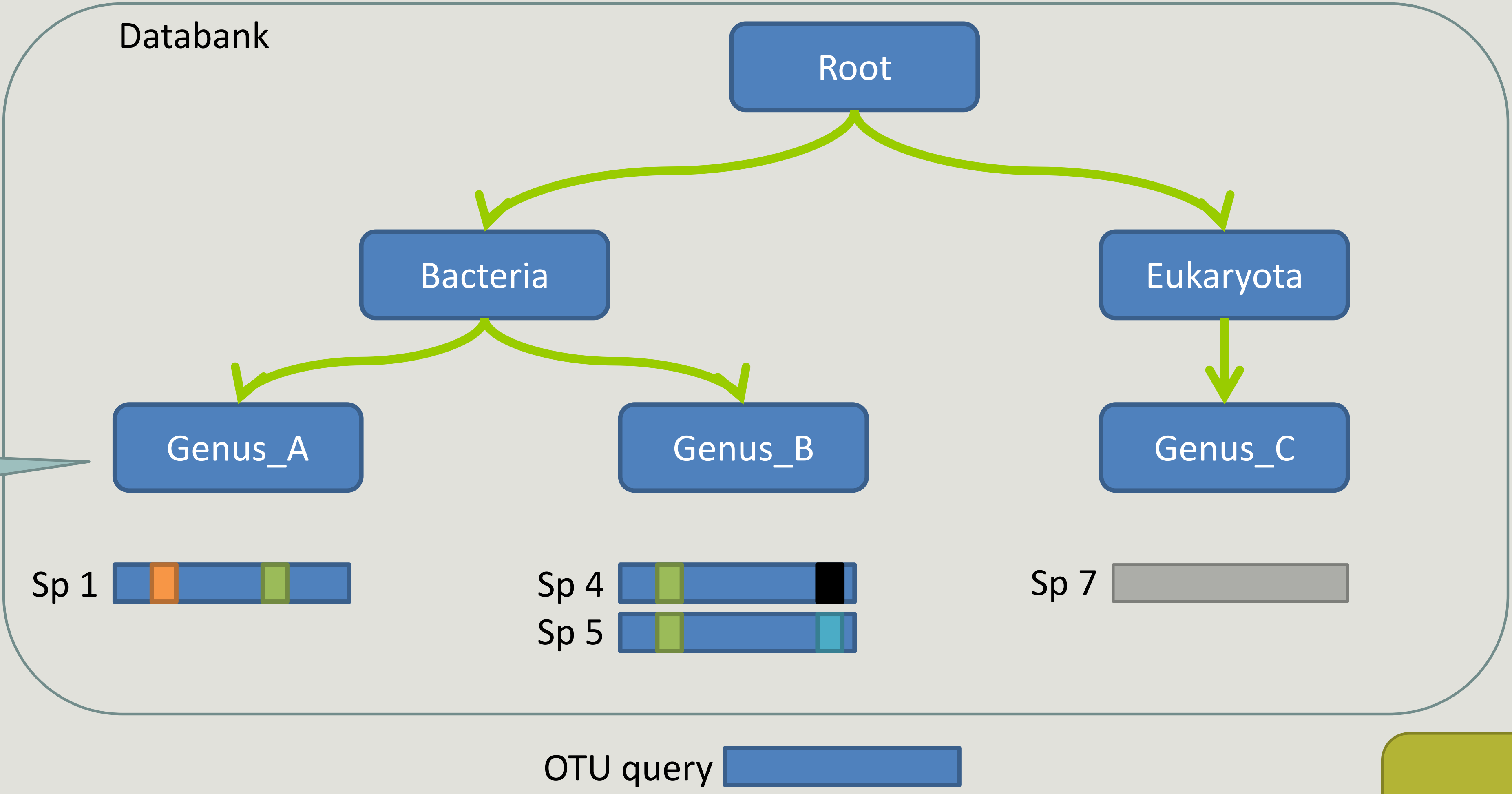
Result:
Bacteria(100) ; Genus_A(50) ; Sp1(25)

The dysfunctions of RDP ?



Result:
?

The dysfunctions of RDP n°1 ?



Genus (50) was expected

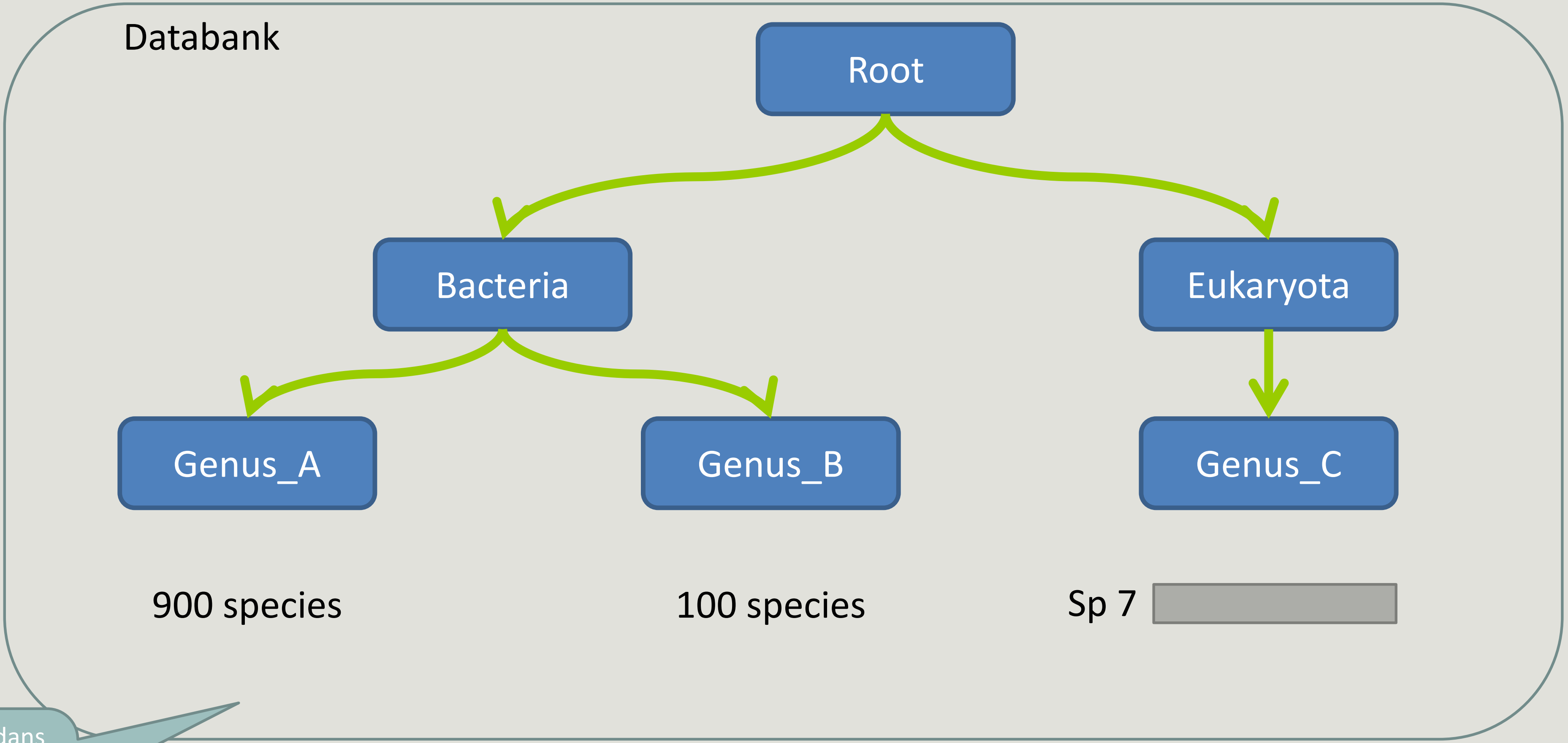
Genus (50) was expected
If seq order was inversed in banks, so we got the second chance

Order dependent

Result:

Bacteria(100); Genus_A(33); sp1(33) OR Bacteria(100); Genus_B(66); sp5(33)

The dysfunctions of RDP n°2 ?



Beaucoup d’espèces dans un genre et peu dans l’autre, alors RDP peut donner des résultats très différents

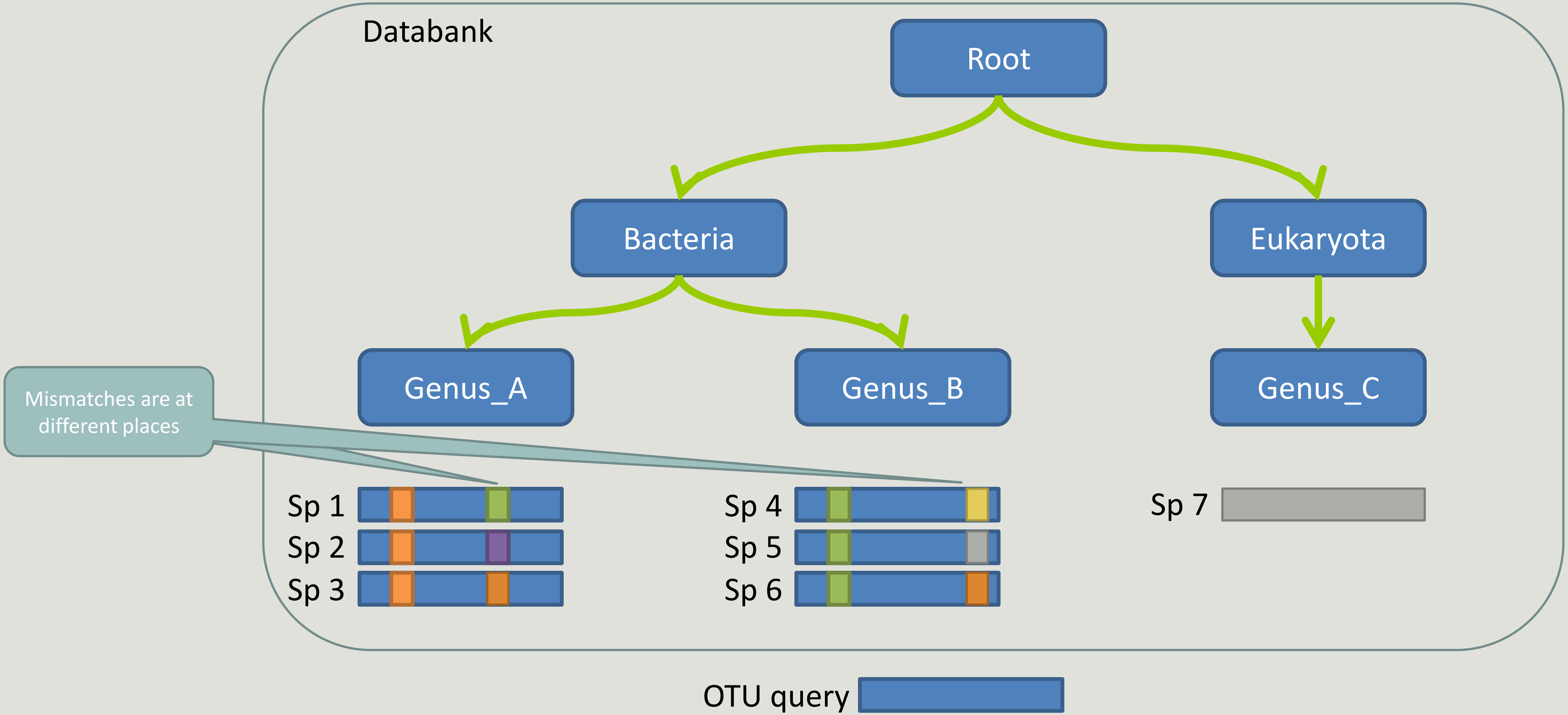
OTU query

Influenced by heterogeneity in last ranks

Result:

Bacteria(100); Genus_A(90); spX(0.1) OR Bacteria(100); Genus_B(10); spX(0.1)

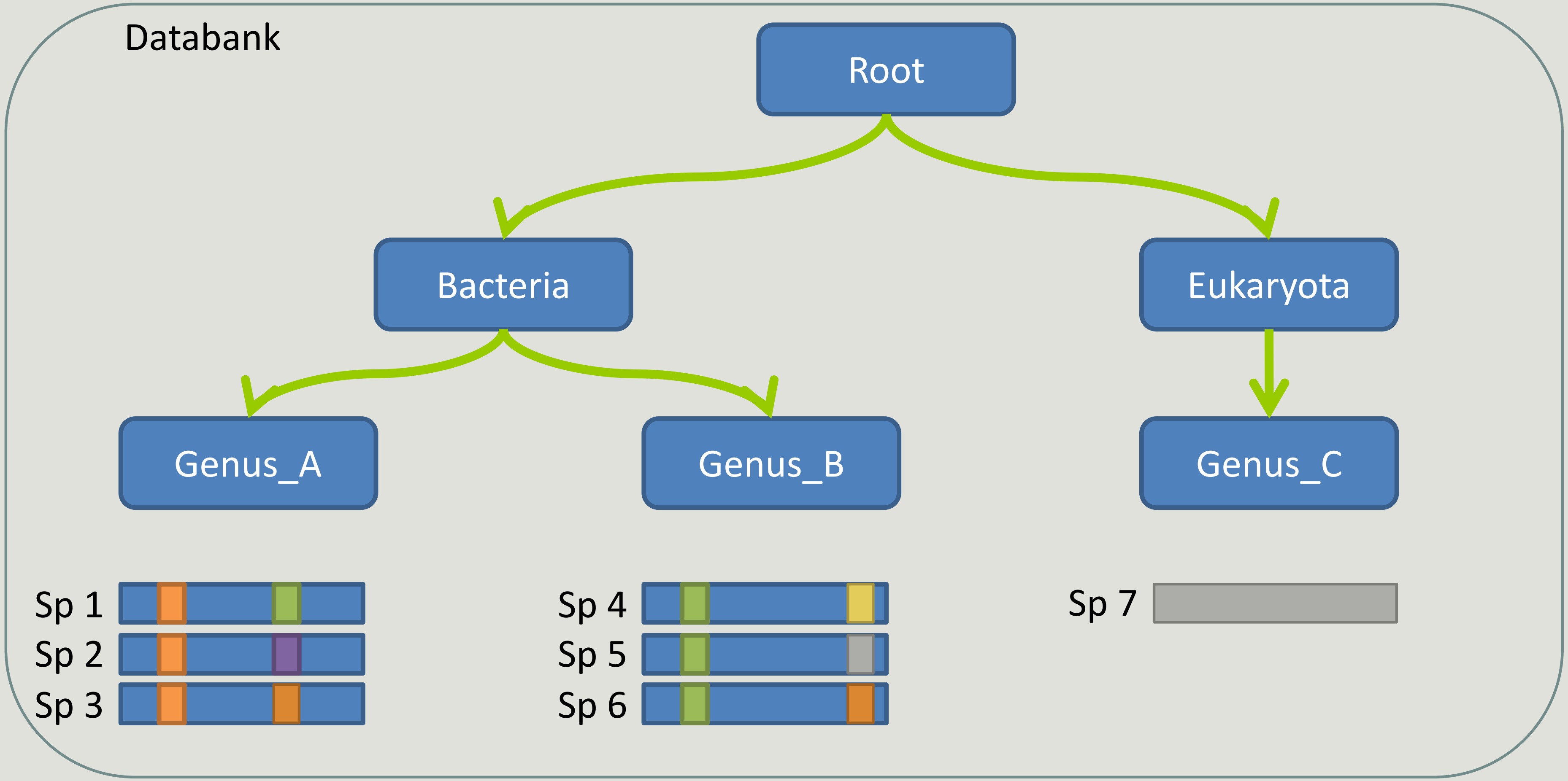
The dysfunctions of RDP n°3 ?

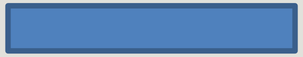


Result:
?

The dysfunctions of RDP n°3 ?

Go to practice





Si le mismatch se fait sur un mot très "significatif" dans le profil de k-mers, RDP ne tombera que rarement sur l'espèce lors du bootstrap. Avec une même distance d'édition (2 mismatchs) on peut donc avoir une grande différence de bootstrap pour peu que le mot affecté soit important dans le profil.

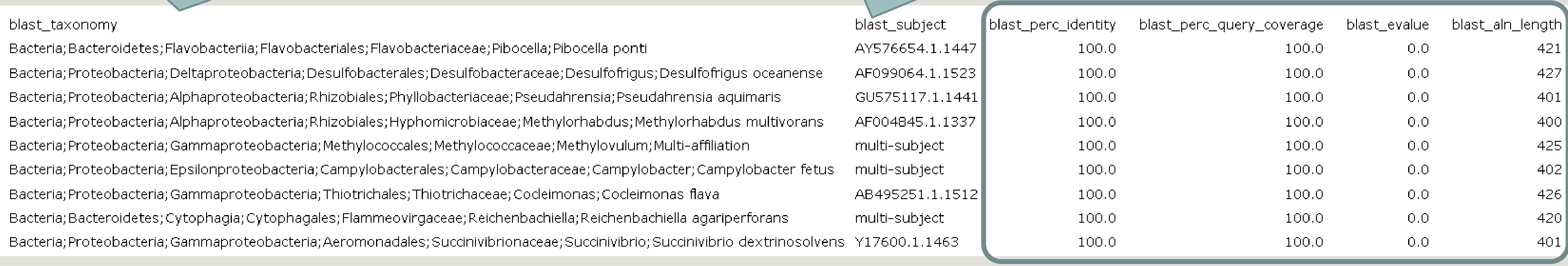
Result:
Bacteria(100); Genus_A(50); sp1(20)

Influenced by the divergences position

2nd to 7th columns – Blast

OTU_1 seed has a best BLAST hit with the reference sequence AQXR01000005.3811.5326

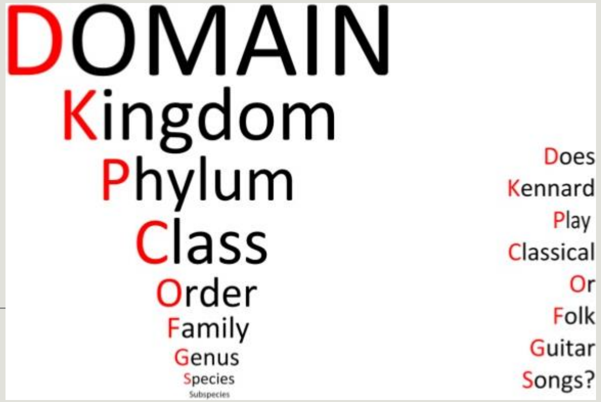
The reference sequence taxonomic affiliation is this one.

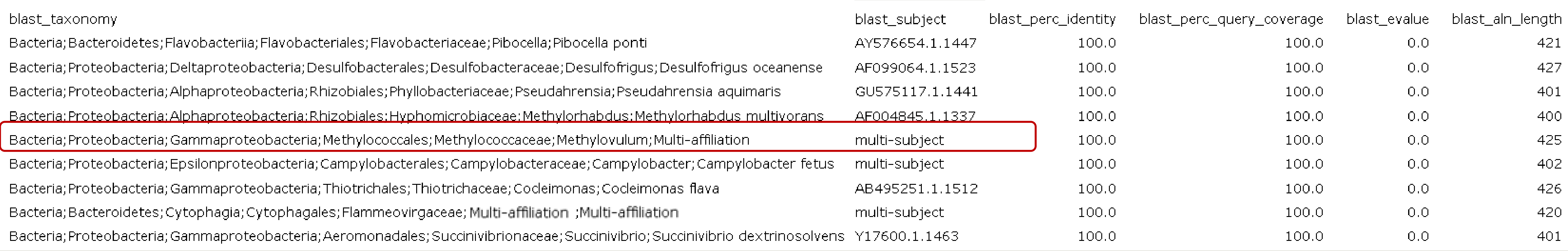


blast_taxonomy	blast_subject	blast_perc_identity	blast_perc_query_coverage	blast_evalue	blast_aln_length
Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;Flavobacteriaceae;Pibocella;Pibocella ponti	AY576654.1.1447	100.0	100.0	0.0	421
Bacteria;Proteobacteria;Deltaproteobacteria;Desulfobacterales;Desulfobacteraceae;Desulfofrigus;Desulfofrigus oceanense	AF099064.1.1523	100.0	100.0	0.0	427
Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Phyllobacteriaceae;Pseudahrensia;Pseudahrensia aquimaris	GU575117.1.1441	100.0	100.0	0.0	401
Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Hyphomicrobiaceae;Methylorhabdus;Methylorhabdus multivorans	AF004845.1.1337	100.0	100.0	0.0	400
Bacteria;Proteobacteria;Gammaproteobacteria;Methylococcales;Methylococcaceae;Methylovulum;Multi-affiliation	multi-subject	100.0	100.0	0.0	425
Bacteria;Proteobacteria;Epsilonproteobacteria;Campylobacterales;Campylobacteraceae;Campylobacter;Campylobacter fetus	multi-subject	100.0	100.0	0.0	402
Bacteria;Proteobacteria;Gammaproteobacteria;Thiotrichales;Thiotrichaceae;Cocleimonas;Cocleimonas flava	AB495251.1.1512	100.0	100.0	0.0	426
Bacteria;Bacteroidetes;Cytophagia;Cytophagales;Flammeovirgaceae;Reichenbachiella;Reichenbachiella agariperforans	multi-subject	100.0	100.0	0.0	420
Bacteria;Proteobacteria;Gammaproteobacteria;Aeromonadales;Succinivibrionaceae;Succinivibrio;Succinivibrio dextrinosolvens	Y17600.1.1463	100.0	100.0	0.0	401

Evaluation variables of BLAST

2nd to 7th columns – Blast

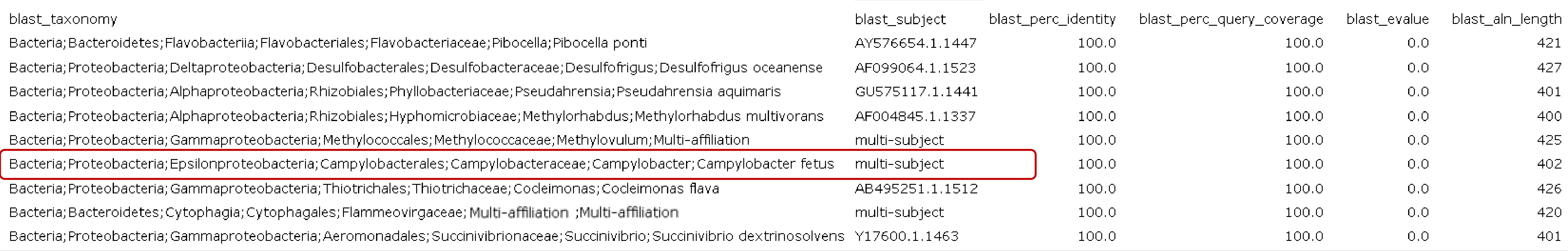




blast_taxonomy	blast_subject	blast_perc_identity	blast_perc_query_coverage	blast_evalue	blast_aln_length
Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;Flavobacteriaceae;Pibocella;Pibocella ponti	AY576654.1.1447	100.0	100.0	0.0	421
Bacteria;Proteobacteria;Deltaproteobacteria;Desulfobacterales;Desulfobacteraceae;Desulfofrigus;Desulfofrigus oceanense	AF099064.1.1523	100.0	100.0	0.0	427
Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Phyllobacteriaceae;Pseudahrensia;Pseudahrensia aquimaris	GU575117.1.1441	100.0	100.0	0.0	401
Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Hyphomicrobiaceae;Methylorhabdus;Methylorhabdus multivorans	AF004845.1.1337	100.0	100.0	0.0	400
Bacteria;Proteobacteria;Gammaproteobacteria;Methylococcales;Methylococcaceae;Methylovulum;Multi-affiliation	multi-subject	100.0	100.0	0.0	425
Bacteria;Proteobacteria;Epsilonproteobacteria;Campylobacterales;Campylobacteraceae;Campylobacter;Campylobacter fetus	multi-subject	100.0	100.0	0.0	402
Bacteria;Proteobacteria;Gammaproteobacteria;Thiotrichales;Thiotrichaceae;Cocleimonas;Cocleimonas flava	AB495251.1.1512	100.0	100.0	0.0	426
Bacteria;Bacteroidetes;Cytophagia;Cytophagales;Flammeovirgaceae;Multi-affiliation ;Multi-affiliation	multi-subject	100.0	100.0	0.0	420
Bacteria;Proteobacteria;Gammaproteobacteria;Aeromonadales;Succinivibrionaceae;Succinivibrio;Succinivibrio dextrinosolvens	Y17600.1.1463	100.0	100.0	0.0	401

Cluster_5 has 4 identical blast hits, with different taxonomies as the species level

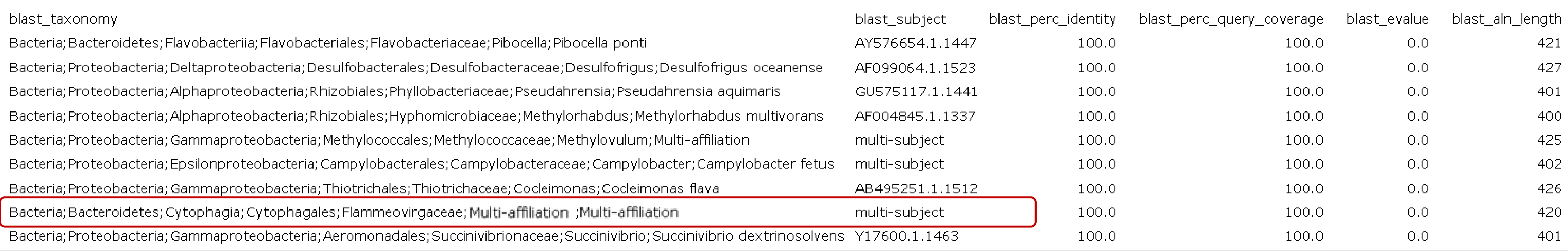
2nd to 7th columns – Blast



blast_taxonomy	blast_subject	blast_perc_identity	blast_perc_query_coverage	blast_evalue	blast_aln_length
Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;Flavobacteriaceae;Pibocella;Pibocella ponti	AY576654.1.1447	100.0	100.0	0.0	421
Bacteria;Proteobacteria;Deltaproteobacteria;Desulfobacterales;Desulfobacteraceae;Desulfofrigus;Desulfofrigus oceanense	AF099064.1.1523	100.0	100.0	0.0	427
Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Phyllobacteriaceae;Pseudahrensia;Pseudahrensia aquimaris	GU575117.1.1441	100.0	100.0	0.0	401
Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Hyphomicrobiaceae;Methylorhabdus;Methylorhabdus multivorans	AF004845.1.1337	100.0	100.0	0.0	400
Bacteria;Proteobacteria;Gammaproteobacteria;Methylococcales;Methylococcaceae;Methylovulum;Multi-affiliation	multi-subject	100.0	100.0	0.0	425
Bacteria;Proteobacteria;Epsilonproteobacteria;Campylobacterales;Campylobacteraceae;Campylobacter;Campylobacter fetus	multi-subject	100.0	100.0	0.0	402
Bacteria;Proteobacteria;Gammaproteobacteria;Thiotrichales;Thiotrichaceae;Cocleimonas;Cocleimonas flava	AB495251.1.1512	100.0	100.0	0.0	426
Bacteria;Bacteroidetes;Cytophagia;Cytophagales;Flammeovirgaceae;Multi-affiliation ;Multi-affiliation	multi-subject	100.0	100.0	0.0	420
Bacteria;Proteobacteria;Gammaproteobacteria;Aeromonadales;Succinivibrionaceae;Succinivibrio;Succinivibrio dextrinosolvens	Y17600.1.1463	100.0	100.0	0.0	401

Cluster_6 has 38 identical blast hits, with different taxonomies as the species level

2nd to 7th columns – Blast



blast_taxonomy	blast_subject	blast_perc_identity	blast_perc_query_coverage	blast_evalue	blast_aln_length
Bacteria;Bacteroidetes;Flavobacteriia;Flavobacteriales;Flavobacteriaceae;Pibocella;Pibocella ponti	AY576654.1.1447	100.0	100.0	0.0	421
Bacteria;Proteobacteria;Deltaproteobacteria;Desulfobacterales;Desulfobacteraceae;Desulfofrigus;Desulfofrigus oceanense	AF099064.1.1523	100.0	100.0	0.0	427
Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Phyllobacteriaceae;Pseudahrensia;Pseudahrensia aquimaris	GU575117.1.1441	100.0	100.0	0.0	401
Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Hyphomicrobiaceae;Methylorhabdus;Methylorhabdus multivorans	AF004845.1.1337	100.0	100.0	0.0	400
Bacteria;Proteobacteria;Gammaproteobacteria;Methylococcales;Methylococcaceae;Methylovulum;Multi-affiliation	multi-subject	100.0	100.0	0.0	425
Bacteria;Proteobacteria;Epsilonproteobacteria;Campylobacterales;Campylobacteraceae;Campylobacter;Campylobacter fetus	multi-subject	100.0	100.0	0.0	402
Bacteria;Proteobacteria;Gammaproteobacteria;Thiotrichales;Thiotrichaceae;Cocleimonas;Cocleimonas flava	AB495251.1.1512	100.0	100.0	0.0	426
Bacteria;Bacteroidetes;Cytophagia;Cytophagales;Flammeovirgaceae; Multi-affiliation ;Multi-affiliation	multi-subject	100.0	100.0	0.0	420
Bacteria;Proteobacteria;Gammaproteobacteria;Aeromonadales;Succinivibrionaceae;Succinivibrio;Succinivibrio dextrinosolvens	Y17600.1.1463	100.0	100.0	0.0	401

Cluster_8 has 2 identical blast hits, with different taxonomies as the genus level

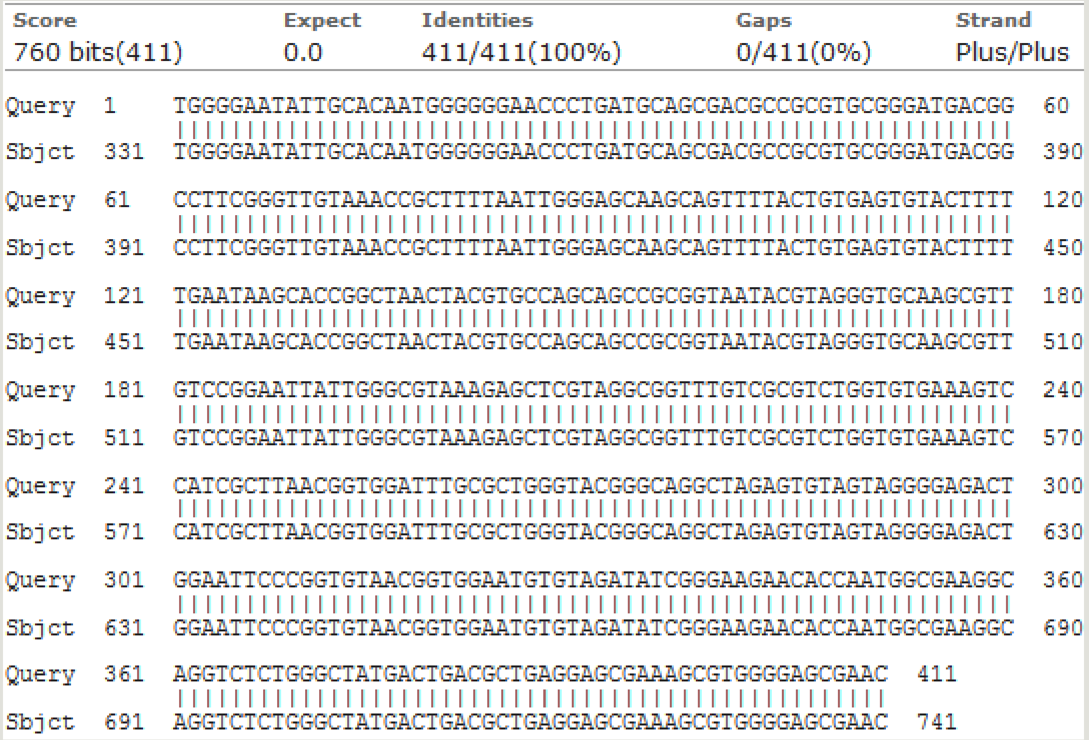
Blast variables : e-value

The Expect value (E) 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.

Blast variables : blast_perc_identity

Identity percentage between the Query (OTU) and the subject in the alignment (length subject = 1455 bases)

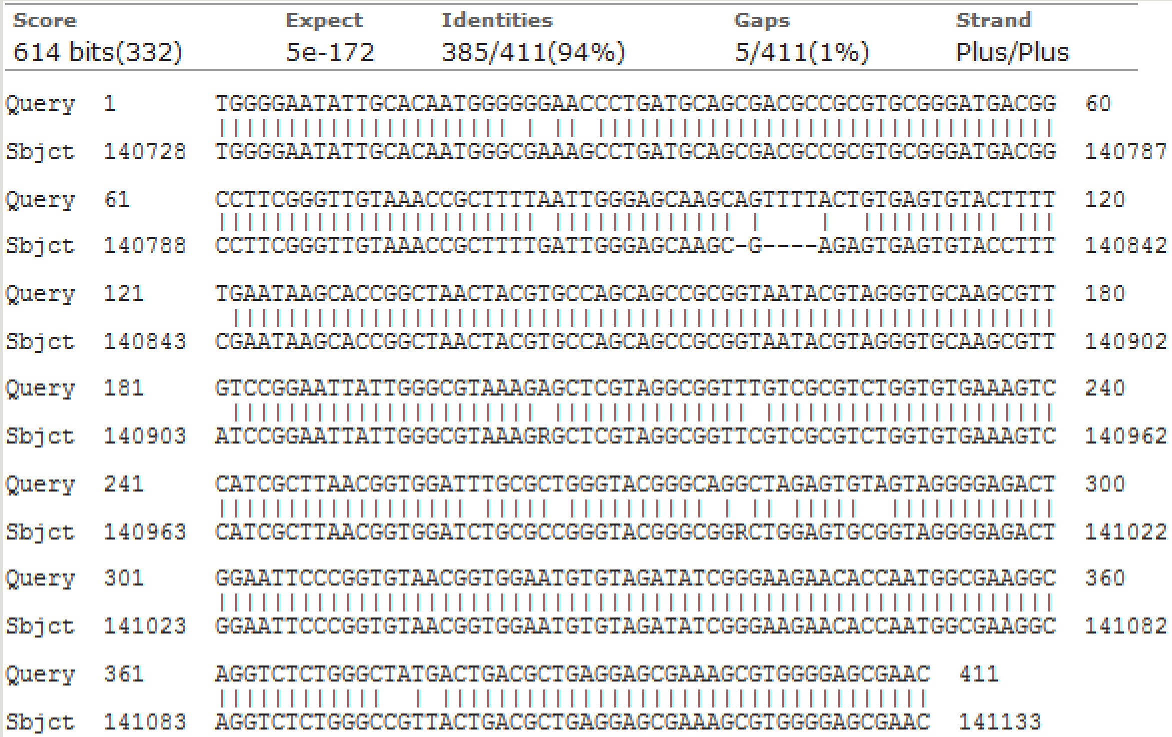


```
Score           Expect   Identities        Gaps         Strand
760 bits(411)   0.0      411/411(100%)     0/411(0%)    Plus/Plus

Query  1    TGGGGAATATTGCACAATGGGGGAACCCTGATGCAGCGACGCCGCGTGCGGGATGACGG  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  331  TGGGGAATATTGCACAATGGGGGAACCCTGATGCAGCGACGCCGCGTGCGGGATGACGG  390

Query  61   CCTTCGGGTTGTAAACCGCTTTTAATTGGGAGCAAGCAGTTTTACTGTGAGTGTACTTTT  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  391  CCTTCGGGTTGTAAACCGCTTTTAATTGGGAGCAAGCAGTTTTACTGTGAGTGTACTTTT  450

Query  121  TGAATAAGCACCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGTGCAAGCGTT  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  451  TGAATAAGCACCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGTGCAAGCGTT  510

Query  181  GTCCGGAATTATTGGGCGTAAAGAGCTCGTAGGCGGTTTGTCGCGTCTGGTGTGAAAGTC  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  511  GTCCGGAATTATTGGGCGTAAAGAGCTCGTAGGCGGTTTGTCGCGTCTGGTGTGAAAGTC  570

Query  241  CATCGCTTAACGGTGGATTTGCGCTGGGTACGGGCAGGCTAGAGTGTAGTAGGGGAGACT  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  571  CATCGCTTAACGGTGGATTTGCGCTGGGTACGGGCAGGCTAGAGTGTAGTAGGGGAGACT  630

Query  301  GGAATTCCCGGTGTAACGGTGGAATGTGTAGATATCGGGAAGAACACCAATGGCGAAGGC  360
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  631  GGAATTCCCGGTGTAACGGTGGAATGTGTAGATATCGGGAAGAACACCAATGGCGAAGGC  690

Query  361  AGGTCTCTGGGCTATGACTGACGCTGAGGAGCGAAAGCGTGGGGAGCGAAC  411
            |||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  691  AGGTCTCTGGGCTATGACTGACGCTGAGGAGCGAAAGCGTGGGGAGCGAAC  741
```

Query length = 411
Alignment length = 411
0 mismatch
-> 100% identity

Blast variables : blast_perc_identity

Identity percentage between the Query (OTU) and the subject in the alignment
(length subject = 1455 bases)



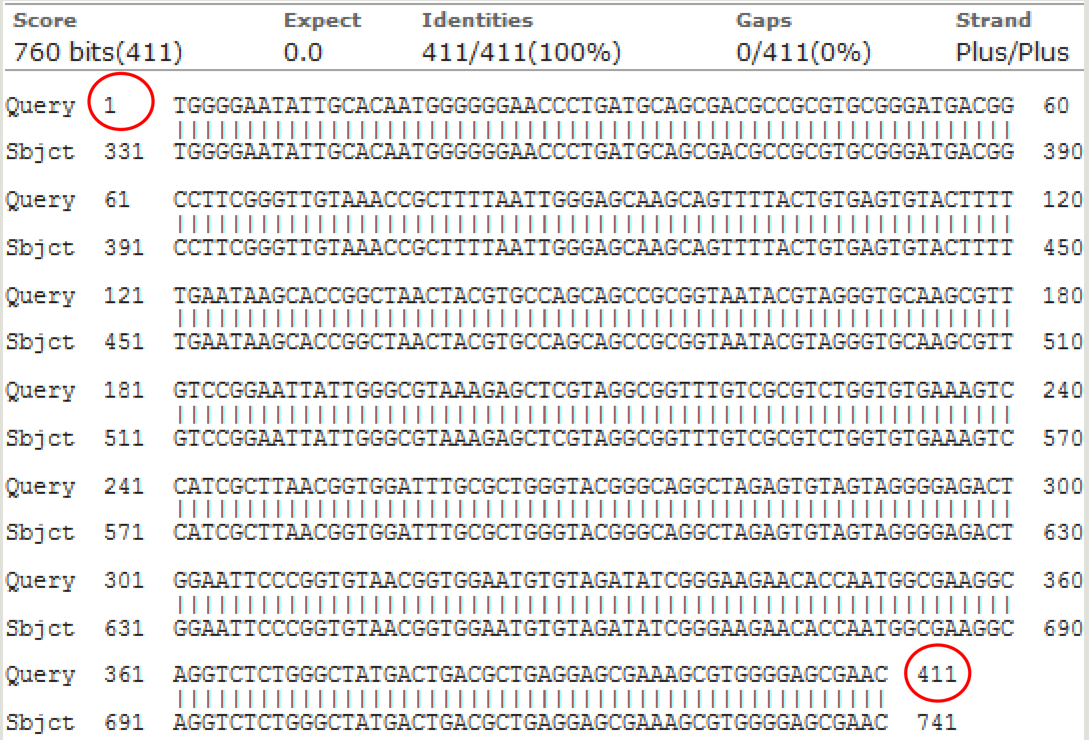
Score	Expect	Identities	Gaps	Strand
614 bits(332)	5e-172	385/411(94%)	5/411(1%)	Plus/Plus

```
Query  1       TGGGGAATATTGCACAATGGGGGGAACCCTGATGCAGCGACGCCGCGTGCGGGATGACGG  60
               ||||||||||||||||||||| | || |||||||||||||||||||||||||||||||||
Sbjct  140728  TGGGGAATATTGCACAATGGGCGAAAGCCTGATGCAGCGACGCCGCGTGCGGGATGACGG  140787

Query  61      CCTTCGGGTTGTAAACCGCTTTTAATTGGGAGCAAGCAGTTTTACTGTGAGTGTACTTTT  120
               ||||||||||||||||||||||| ||||||||||||| |    |  |||||||||| |||
Sbjct  140788  CCTTCGGGTTGTAAACCGCTTTTGATTGGGAGCAAGC-G----AGAGTGAGTGTACCTTT  140842

Query  121     TGAATAAGCACCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGTGCAAGCGTT  180
                |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  140843  CGAATAAGCACCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGTGCAAGCGTT  140902

Query  181     GTCCGGAATTATTGGGCGTAAAGAGCTCGTAGGCGGTTTGTCGCGTCTGGTGTGAAAGTC  240
                |||||||||||||||||||||| |||||||||||||| |||||||||||||||||||||
Sbjct  140903  ATCCGGAATTATTGGGCGTAAAGRGCTCGTAGGCGGTTCGTCGCGTCTGGTGTGAAAGTC  140962

Query  241     CATCGCTTAACGGTGGATTTGCGCTGGGTACGGGCAGGCTAGAGTGTAGTAGGGGAGACT  300
               |||||||||||||||||| ||||| |||||||||| | || |||||  ||||||||||||
Sbjct  140963  CATCGCTTAACGGTGGATCTGCGCCGGGTACGGGCGGRCTGGAGTGCGGTAGGGGAGACT  141022

Query  301     GGAATTCCCGGTGTAACGGTGGAATGTGTAGATATCGGGAAGAACACCAATGGCGAAGGC  360
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  141023  GGAATTCCCGGTGTAACGGTGGAATGTGTAGATATCGGGAAGAACACCAATGGCGAAGGC  141082

Query  361     AGGTCTCTGGGCTATGACTGACGCTGAGGAGCGAAAGCGTGGGGAGCGAAC  411
               ||||||||||||  | |||||||||||||||||||||||||||||||||||
Sbjct  141083  AGGTCTCTGGGCCGTTACTGACGCTGAGGAGCGAAAGCGTGGGGAGCGAAC  141133
```

Query length = 411
Alignment length = 411
26 mismatches (gaps included)
-> 94% identity

Blast variables : blast_perc_query_coverage

Coverage percentage of alignment on query (OTU)



Score	Expect	Identities	Gaps	Strand
760 bits(411)	0.0	411/411(100%)	0/411(0%)	Plus/Plus

```
Query  1    TGGGGAATATTGCACAATGGGGGAACCCTGATGCAGCGACGCCGCGTGCGGGATGACGG  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  331  TGGGGAATATTGCACAATGGGGGAACCCTGATGCAGCGACGCCGCGTGCGGGATGACGG  390

Query  61   CCTTCGGGTTGTAAACCGCTTTTAATTGGGAGCAAGCAGTTTTACTGTGAGTGTACTTTT  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  391  CCTTCGGGTTGTAAACCGCTTTTAATTGGGAGCAAGCAGTTTTACTGTGAGTGTACTTTT  450

Query  121  TGAATAAGCACCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGTGCAAGCGTT  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  451  TGAATAAGCACCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGTGCAAGCGTT  510

Query  181  GTCCGGAATTATTGGGCGTAAAGAGCTCGTAGGCGGTTTGTCGCGTCTGGTGTGAAAGTC  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  511  GTCCGGAATTATTGGGCGTAAAGAGCTCGTAGGCGGTTTGTCGCGTCTGGTGTGAAAGTC  570

Query  241  CATCGCTTAACGGTGGATTTGCGCTGGGTACGGGCAGGCTAGAGTGTAGTAGGGGAGACT  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  571  CATCGCTTAACGGTGGATTTGCGCTGGGTACGGGCAGGCTAGAGTGTAGTAGGGGAGACT  630

Query  301  GGAATTCCCGGTGTAACGGTGGAATGTGTAGATATCGGGAAGAACACCAATGGCGAAGGC  360
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  631  GGAATTCCCGGTGTAACGGTGGAATGTGTAGATATCGGGAAGAACACCAATGGCGAAGGC  690

Query  361  AGGTCTCTGGGCTATGACTGACGCTGAGGAGCGAAAGCGTGGGGAGCGAAC  411
            |||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  691  AGGTCTCTGGGCTATGACTGACGCTGAGGAGCGAAAGCGTGGGGAGCGAAC  741
```

Query length = 411
100% coverage

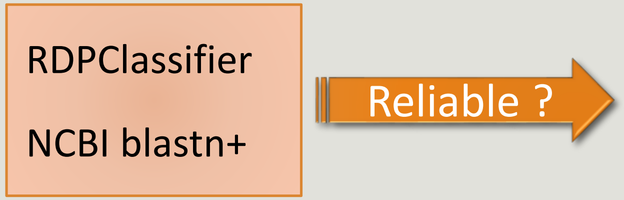
Blast variables : blast-length

Length of alignment between the OTUs = “Query” and “subject” sequence of database

	Coverage %	Identity %	Length alignment
OTU1	100	98	400
OTU2	100	98	500

More mismatches/gaps

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

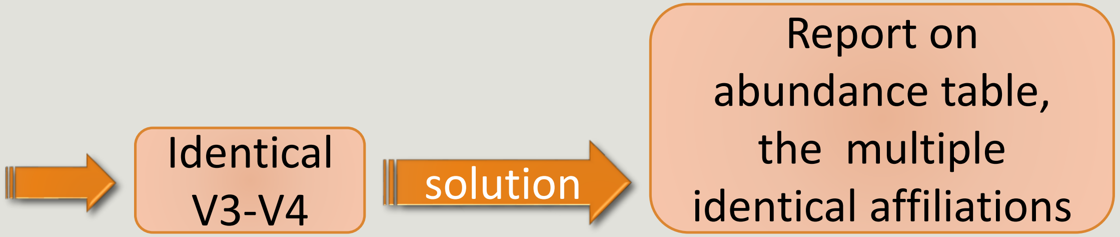


RDPClassifier

NCBI blastn+

Reliable ?

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



Identical V3-V4

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

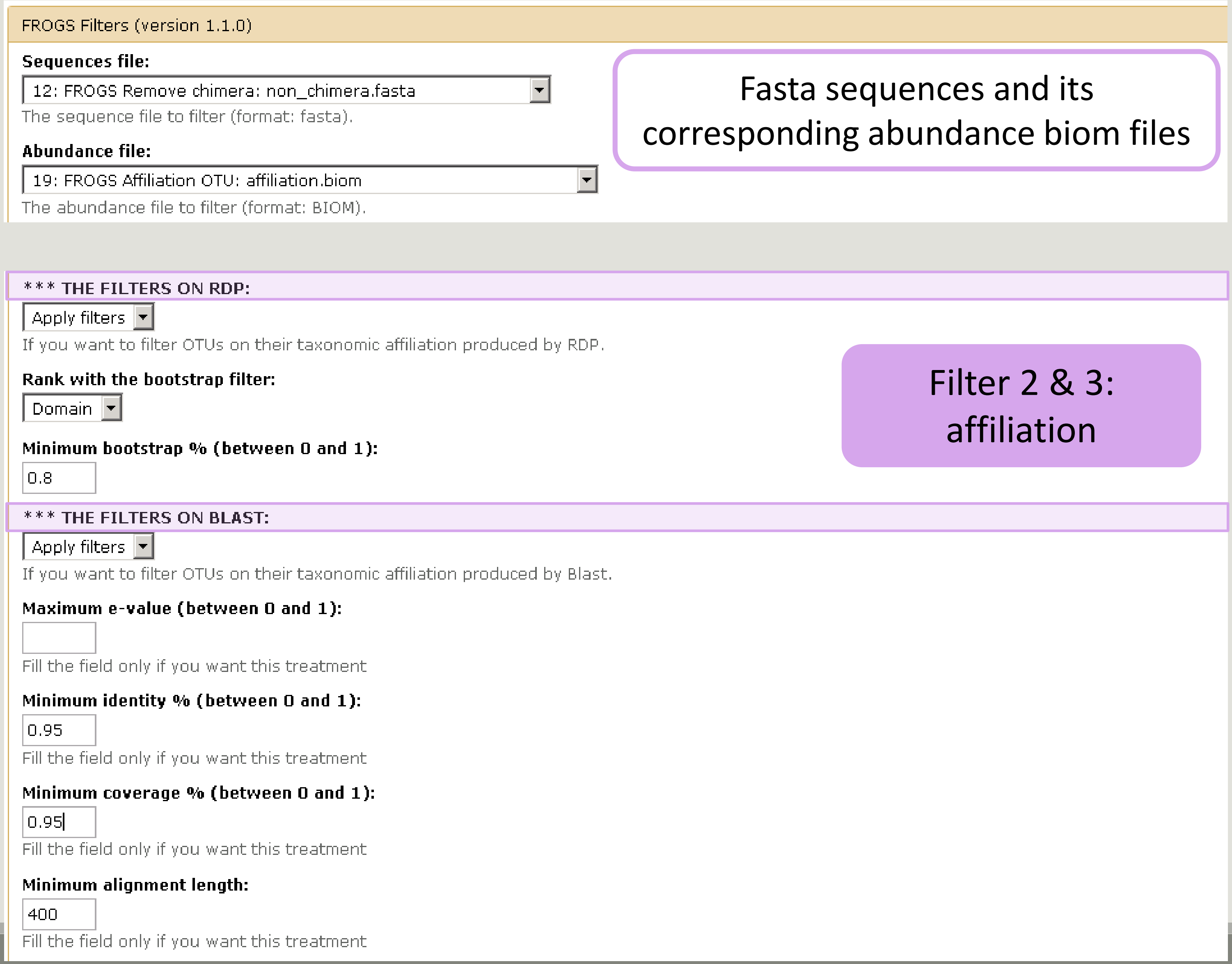
Taxonomic ranks	Median divergence of the affiliations of the 10 samples (%) 500setA filter: 0.005% - 505 OTUs	Median divergence of the affiliations of the 10 samples (%) 100setA filter: 0.005% - 100 OTUs
Kingdom	0.00	0.00
Phylum	0.38	0.38
Class	0.57	0.48
Order	0.81	0.64
Familly	1.08	0.74
Genus	1.43	0.76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

Careful: Multi hit blast table is non exhaustive !

- Chimera (multiple affiliation)
- V3V4 included in others
- Missed primers on some 16S during database building

Do not forget, with filter tool we can filter the data

Input



FROGS Filters (version 1.1.0)

Sequences file:

12: FROGS Remove chimera: non_chimera.fasta

The sequence file to filter (format: fasta).

Abundance file:

19: FROGS Affiliation OTU: affiliation.biom

The abundance file to filter (format: BIOM).

Fasta sequences and its corresponding abundance biom files

***** THE FILTERS ON RDP:**

Apply filters

If you want to filter OTUs on their taxonomic affiliation produced by RDP.

Rank with the bootstrap filter:

Domain

Minimum bootstrap % (between 0 and 1):

0.8

Filter 2 & 3: affiliation

***** THE FILTERS ON BLAST:**

Apply filters

If you want to filter OTUs on their taxonomic affiliation produced by Blast.

Maximum e-value (between 0 and 1):

Fill the field only if you want this treatment

Minimum identity % (between 0 and 1):

0.95

Fill the field only if you want this treatment

Minimum coverage % (between 0 and 1):

0.95

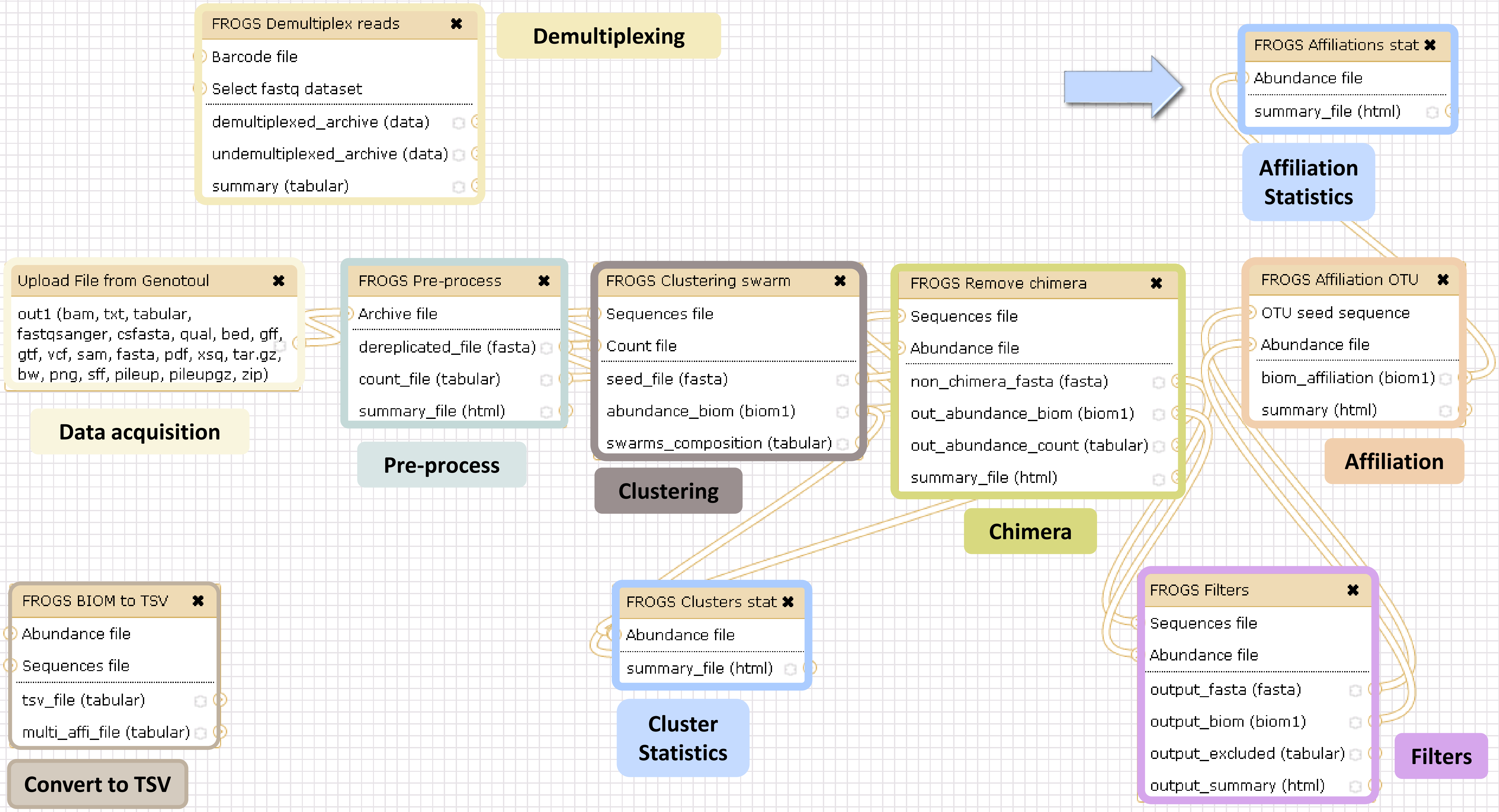
Fill the field only if you want this treatment

Minimum alignment length:

400

Fill the field only if you want this treatment

Affiliation Stat



FROGS Demultiplex reads ✖
Barcode file
Select fastq dataset
demultiplexed_archive (data)
undemultiplexed_archive (data)
summary (tabular)

Demultiplexing

FROGS Affiliations stat ✖
Abundance file
summary_file (html)

Affiliation Statistics

Upload File from Genotoul ✖
out1 (bam, txt, tabular, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png, sff, pileup, pileupgz, zip)

Data acquisition

FROGS Pre-process ✖
Archive file
dereplicated_file (fasta)
count_file (tabular)
summary_file (html)

Pre-process

FROGS Clustering swarm ✖
Sequences file
Count file
seed_file (fasta)
abundance_biom (biom1)
swarms_composition (tabular)

Clustering

FROGS Remove chimera ✖
Sequences file
Abundance file
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
out_abundance_count (tabular)
summary_file (html)

Chimera

FROGS Affiliation OTU ✖
OTU seed sequence
Abundance file
biom_affiliation (biom1)
summary (html)

Affiliation

FROGS BIOM to TSV ✖
Abundance file
Sequences file
tsv_file (tabular)
multi_affi_file (tabular)

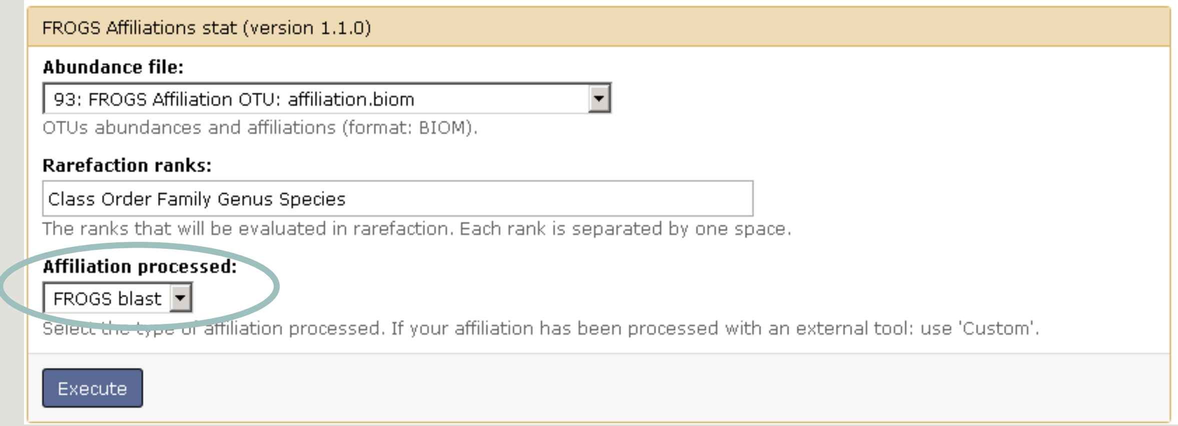
Convert to TSV

FROGS Clusters stat ✖
Abundance file
summary_file (html)

Cluster Statistics

FROGS Filters ✖
Sequences file
Abundance file
output_fasta (fasta)
output_biom (biom1)
output_excluded (tabular)
output_summary (html)

Filters



FROGS Affiliations stat (version 1.1.0)

Abundance file:

93: FROGS Affiliation OTU: affiliation.biom

OTUs abundances and affiliations (format: BIOM).

Rarefaction ranks:

Class Order Family Genus Species

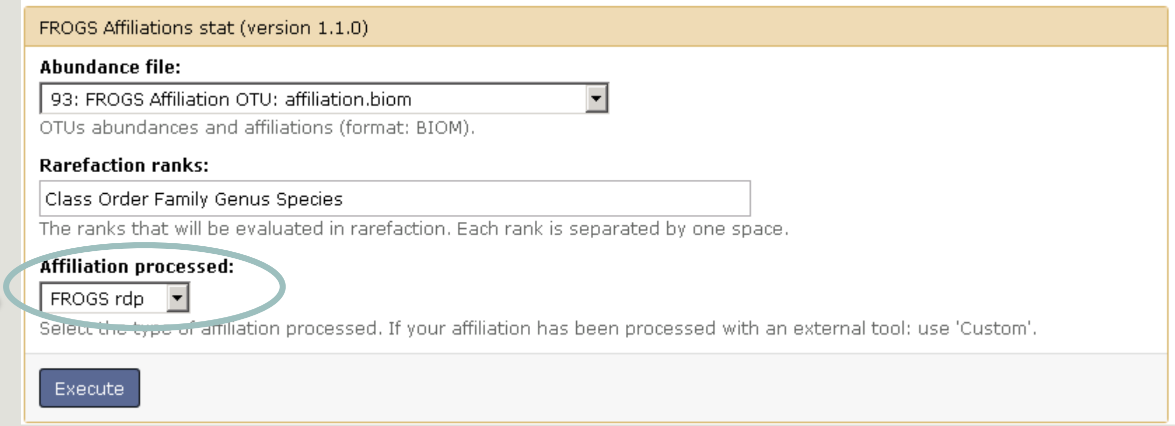
The ranks that will be evaluated in rarefaction. Each rank is separated by one space.

Affiliation processed:

FROGS blast

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

OR



FROGS Affiliations stat (version 1.1.0)

Abundance file:

93: FROGS Affiliation OTU: affiliation.biom

OTUs abundances and affiliations (format: BIOM).

Rarefaction ranks:

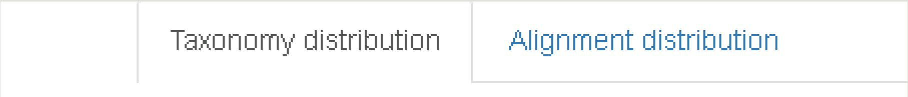
Class Order Family Genus Species

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

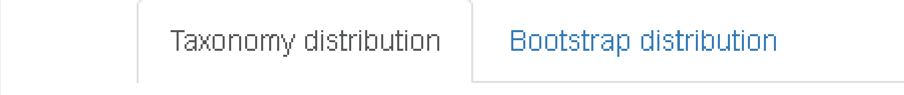
Affiliation processed:

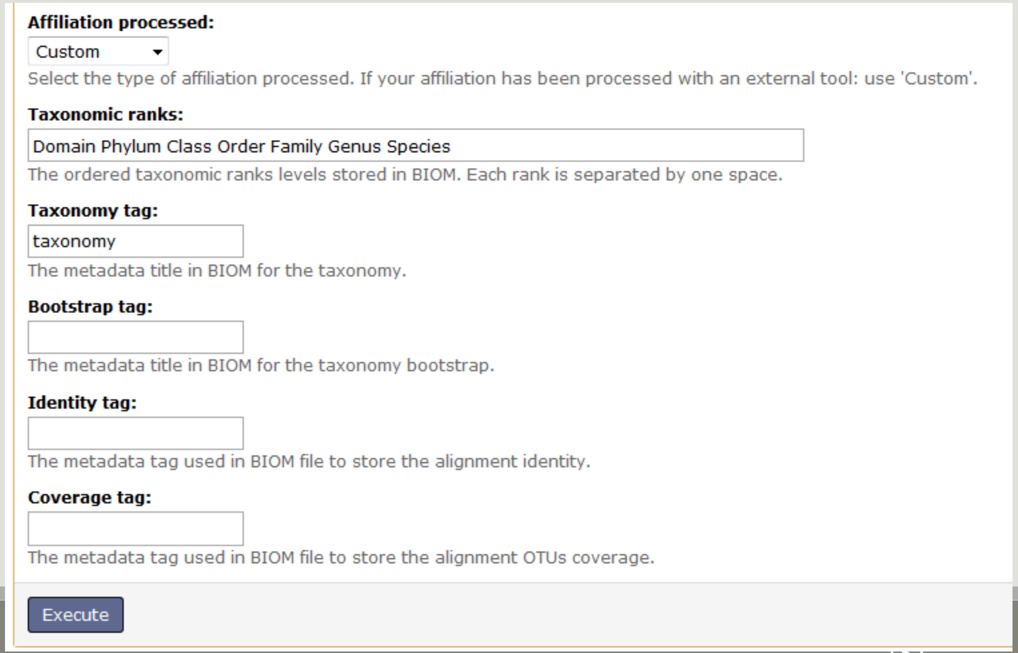
FROGS rdp

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



OR





Affiliation processed:

Custom

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

Taxonomic ranks:

Domain Phylum Class Order Family Genus Species

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

Taxonomy tag:

taxonomy

The metadata title in BIOM for the taxonomy.

Bootstrap tag:

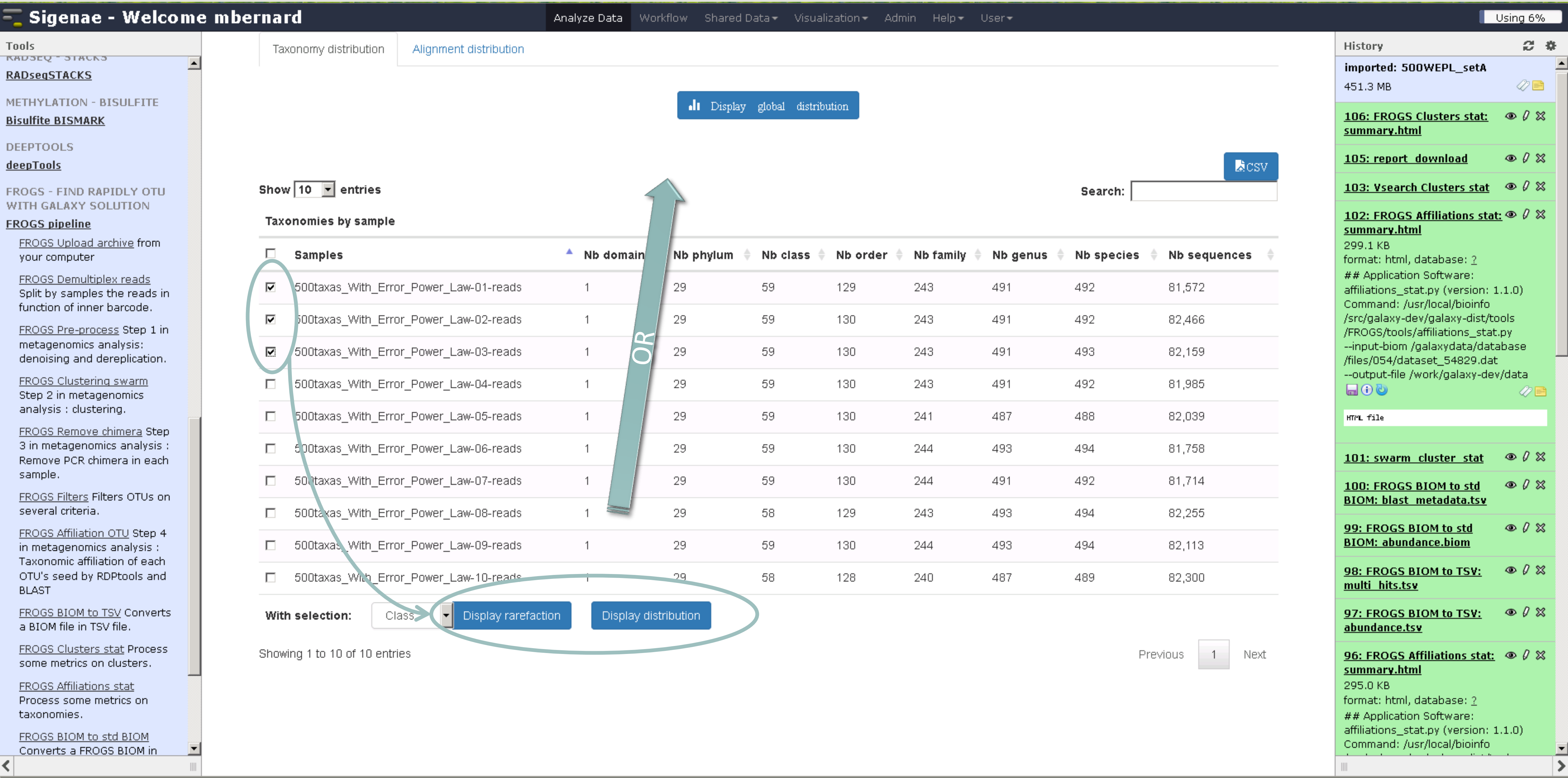
The metadata title in BIOM for the taxonomy bootstrap.

Identity tag:

The metadata tag used in BIOM file to store the alignment identity.

Coverage tag:

The metadata tag used in BIOM file to store the alignment OTUs coverage.



Tools

RADSEQ - STACKS

RADseqSTACKS

METHYLATION - BISULFITE

Bisulfite BISMARK

DEEPTOOLS

deepTools

FROGS - FIND RAPIDLY OTU WITH GALAXY SOLUTION

FROGS pipeline

FROGS Upload archive from your computer

FROGS Demultiplex reads Split by samples the reads in function of inner barcode.

FROGS Pre-process Step 1 in metagenomics analysis: denoising and dereplication.

FROGS Clustering swarm Step 2 in metagenomics analysis : clustering.

FROGS Remove chimera Step 3 in metagenomics analysis : Remove PCR chimera in each sample.

FROGS Filters Filters OTUs on several criteria.

FROGS Affiliation OTU Step 4 in metagenomics analysis : Taxonomic affiliation of each OTU's seed by RDPtools and BLAST

FROGS BIOM to TSV Converts a BIOM file in TSV file.

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

Taxonomy distribution | Alignment distribution

Display global distribution

CSV

Show 10 entries

Search:

Taxonomies by sample

□	Samples	Nb domain	Nb phylum	Nb class	Nb order	Nb family	Nb genus	Nb species	Nb sequences
☑	500taxas_With_Error_Power_Law-01-reads	1	29	59	129	243	491	492	81,572
☑	500taxas_With_Error_Power_Law-02-reads	1	29	59	130	243	491	492	82,466
☑	500taxas_With_Error_Power_Law-03-reads	1	29	59	130	243	491	493	82,159
□	500taxas_With_Error_Power_Law-04-reads	1	29	59	130	243	491	492	81,985
□	500taxas_With_Error_Power_Law-05-reads	1	29	59	130	241	487	488	82,039
□	500taxas_With_Error_Power_Law-06-reads	1	29	59	130	244	493	494	81,758
□	500taxas_With_Error_Power_Law-07-reads	1	29	59	130	244	491	492	81,714
□	500taxas_With_Error_Power_Law-08-reads	1	29	58	129	243	493	494	82,255
□	500taxas_With_Error_Power_Law-09-reads	1	29	59	130	244	493	494	82,113
□	500taxas_With_Error_Power_Law-10-reads	1	29	58	128	240	487	489	82,300

With selection: Class | Display rarefaction | Display distribution

Showing 1 to 10 of 10 entries

Previous 1 Next

History

imported: 500WEPL_setA

451.3 MB

106: FROGS Clusters stat: summary.html

105: report_download

103: Vsearch Clusters stat

102: FROGS Affiliations stat: summary.html

299.1 KB

format: html, database: ?

Application Software:

affiliations_stat.py (version: 1.1.0)

Command: /usr/local/bioinfo/src/galaxy-dev/galaxy-dist/tools/FROGS/tools/affiliations_stat.py --input-biom /galaxydata/database/files/054/dataset_54829.dat --output-file /work/galaxy-dev/data

HTML file

101: swarm_cluster_stat

100: FROGS BIOM to std BIOM: blast_metadata.tsv

99: FROGS BIOM to std BIOM: abundance.biom

98: FROGS BIOM to TSV: multi_hits.tsv

97: FROGS BIOM to TSV: abundance.tsv

96: FROGS Affiliations stat: summary.html

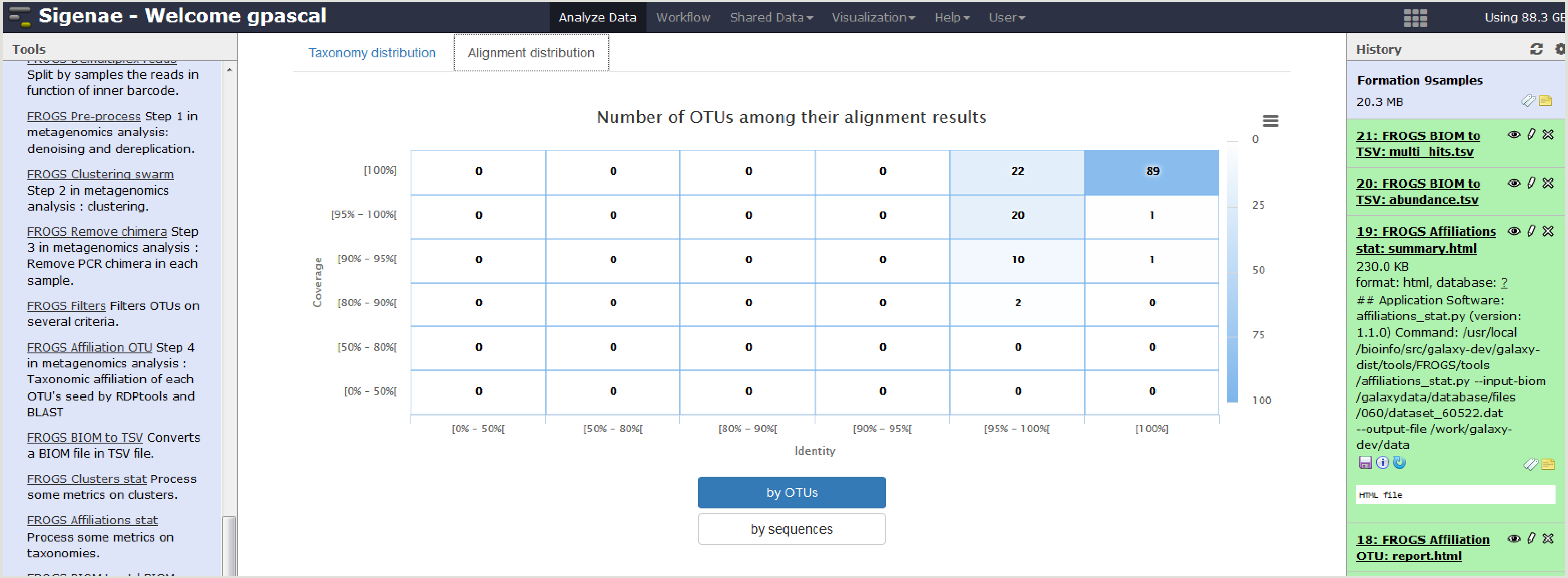
295.0 KB

format: html, database: ?

Application Software:

affiliations_stat.py (version: 1.1.0)

Command: /usr/local/bioinfo



Split by samples the reads in function of inner barcode.

FROGS Pre-process Step 1 in metagenomics analysis: denoising and dereplication.

FROGS Clustering swarm Step 2 in metagenomics analysis : clustering.

FROGS Remove chimera Step 3 in metagenomics analysis : Remove PCR chimera in each sample.

FROGS Filters Filters OTUs on several criteria.

FROGS Affiliation OTU Step 4 in metagenomics analysis : Taxonomic affiliation of each OTU's seed by RDPtools and BLAST

FROGS BIOM to TSV Converts a BIOM file in TSV file.

FROGS Clusters stat Process some metrics on clusters.

FROGS Affiliations stat Process some metrics on taxonomies.

Number of OTUs among their alignment results

Coverage \ Identity	[0% – 50%[	[50% – 80%[	[80% – 90%[	[90% – 95%[	[95% – 100%[	[100%]
[100%]	0	0	0	0	22	89
[95% – 100%[	0	0	0	0	20	1
[90% – 95%[	0	0	0	0	10	1
[80% – 90%[	0	0	0	0	2	0
[50% – 80%[	0	0	0	0	0	0
[0% – 50%[	0	0	0	0	0	0

Formation 9samples
20.3 MB

21: FROGS BIOM to TSV: multi_hits.tsv

20: FROGS BIOM to TSV: abundance.tsv

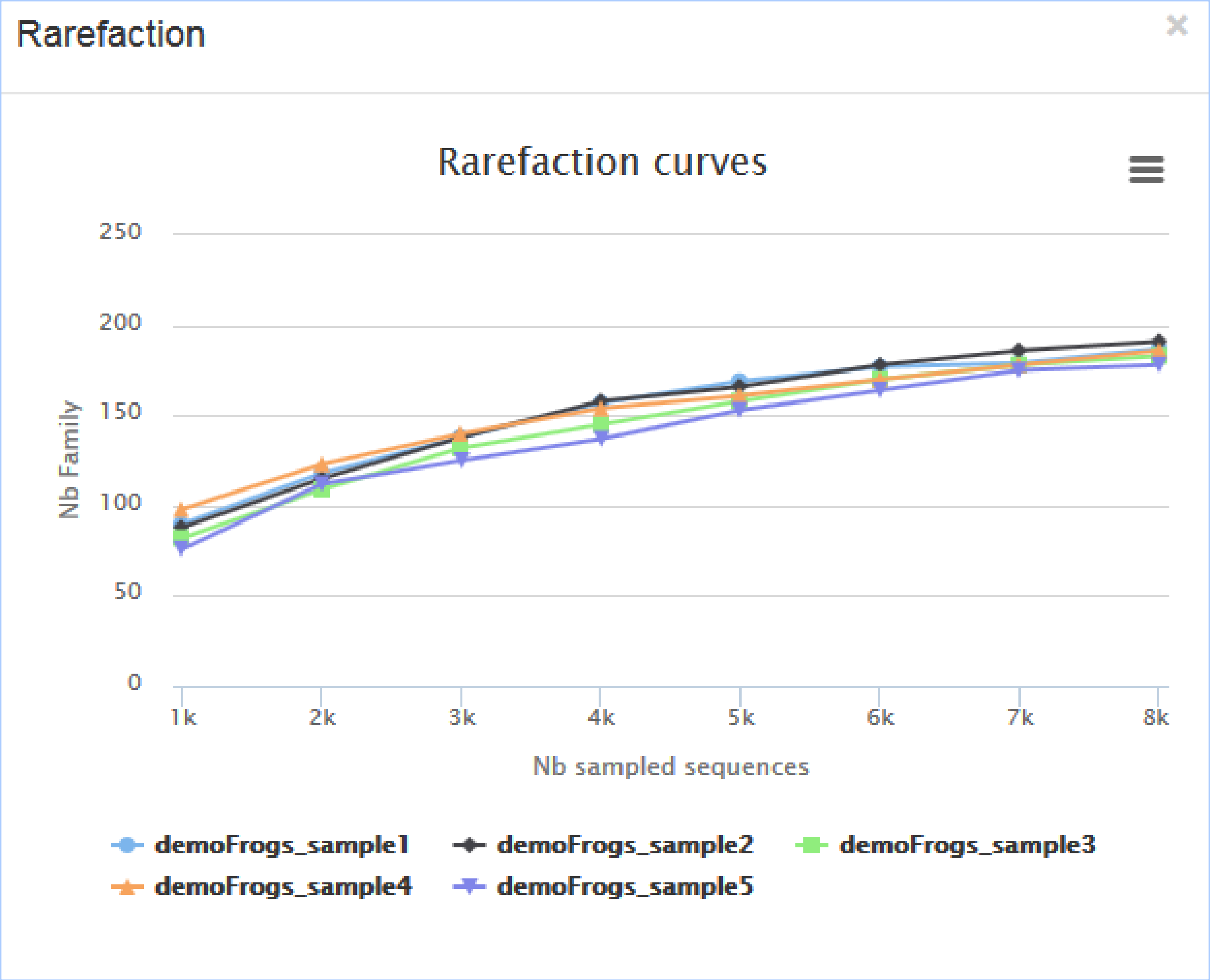
19: FROGS Affiliations stat: summary.html
230.0 KB
format: html, database: ?
Application Software: affiliations_stat.py (version: 1.1.0) Command: /usr/local /bioinfo/src/galaxy-dev/galaxy-dist/tools/FROGS/tools /affiliations_stat.py --input-biom /galaxydata/database/files /060/dataset_60522.dat --output-file /work/galaxy-dev/data

HTML file

18: FROGS Affiliation OTU: report.html

Available only after AFFILIATION TOOL

Samples size ~8500 sequences



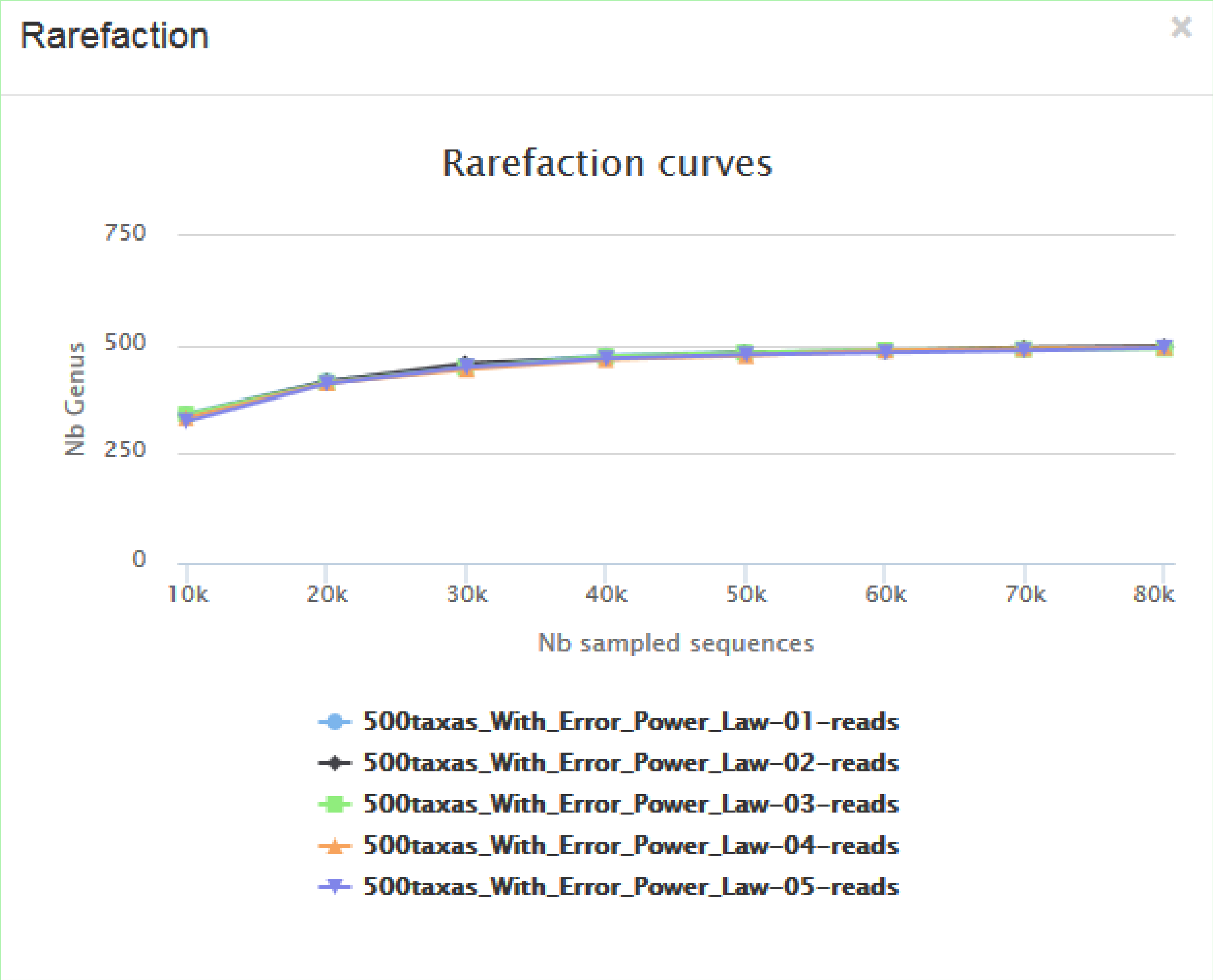
The curve continues to rise

The number of sequences per sample is not large enough to cover all of the bacterial families

Rarefaction tab

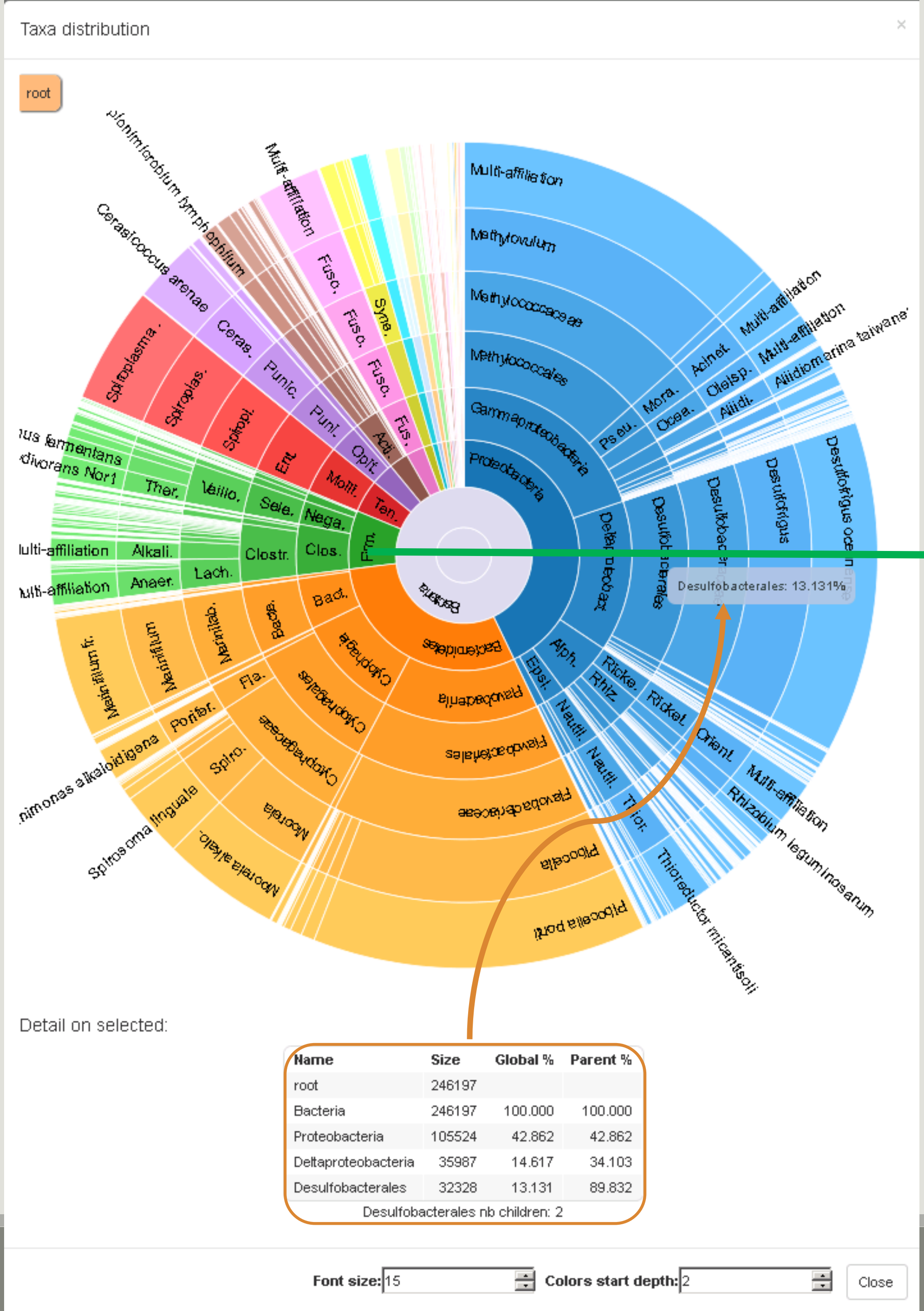
Available only after AFFILIATION TOOL

Samples size ~85 000 sequences



The curve slows to rise with ~50 000 sequences

With 60 000 sequences, we catch almost all genus of bacteria



Taxa distribution

root

Desulfobacterales: 13.131%

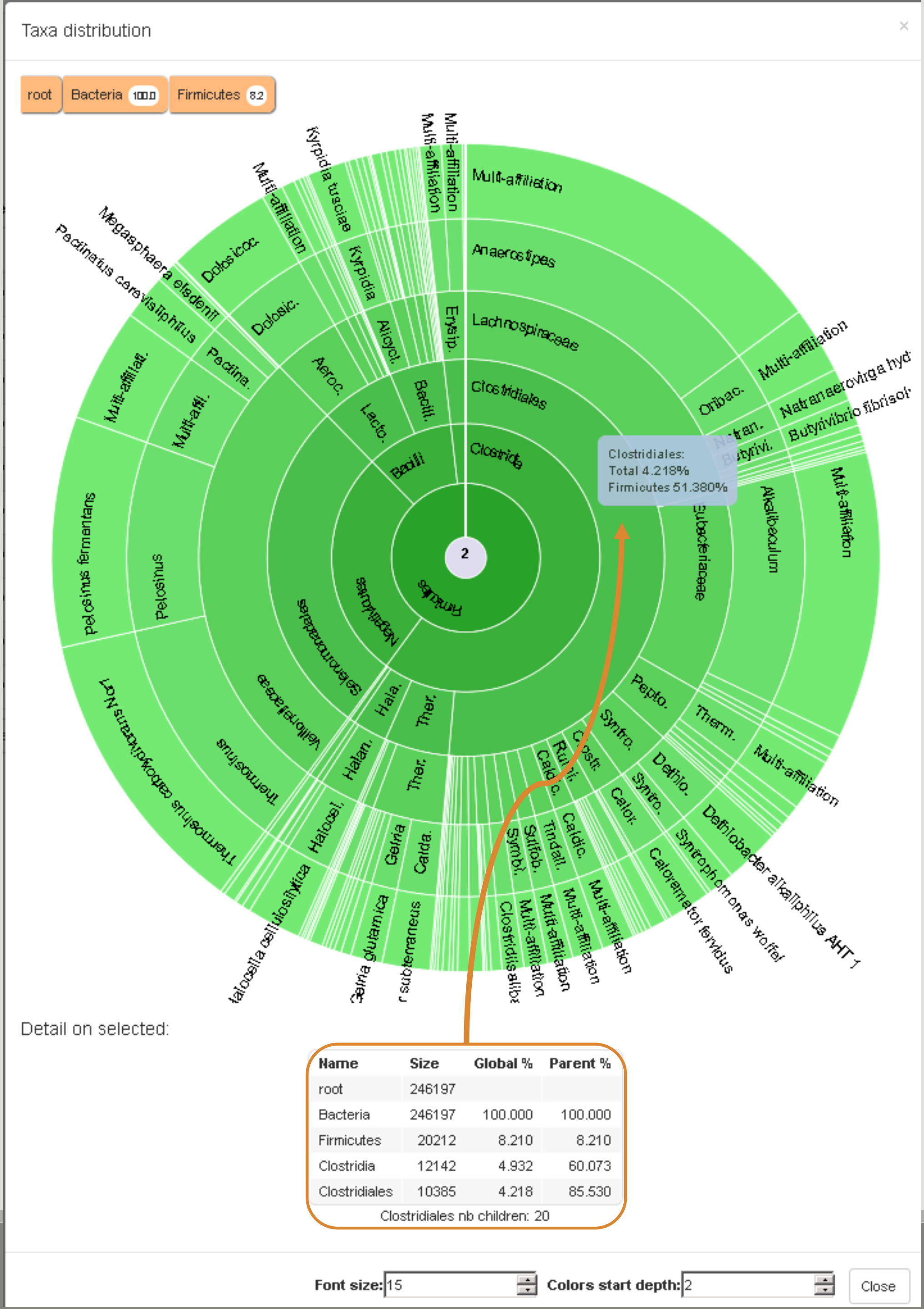
Detail on selected:

Name	Size	Global %	Parent %
root	246197		
Bacteria	246197	100.000	100.000
Proteobacteria	105524	42.862	42.862
Deltaproteobacteria	35987	14.617	34.103
Desulfobacterales	32328	13.131	89.832

Desulfobacterales nb children: 2

Font size: 15 Colors start depth: 2 Close

Zoom in on firmicutes



Taxa distribution

root Bacteria 100.0 Firmicutes 8.2

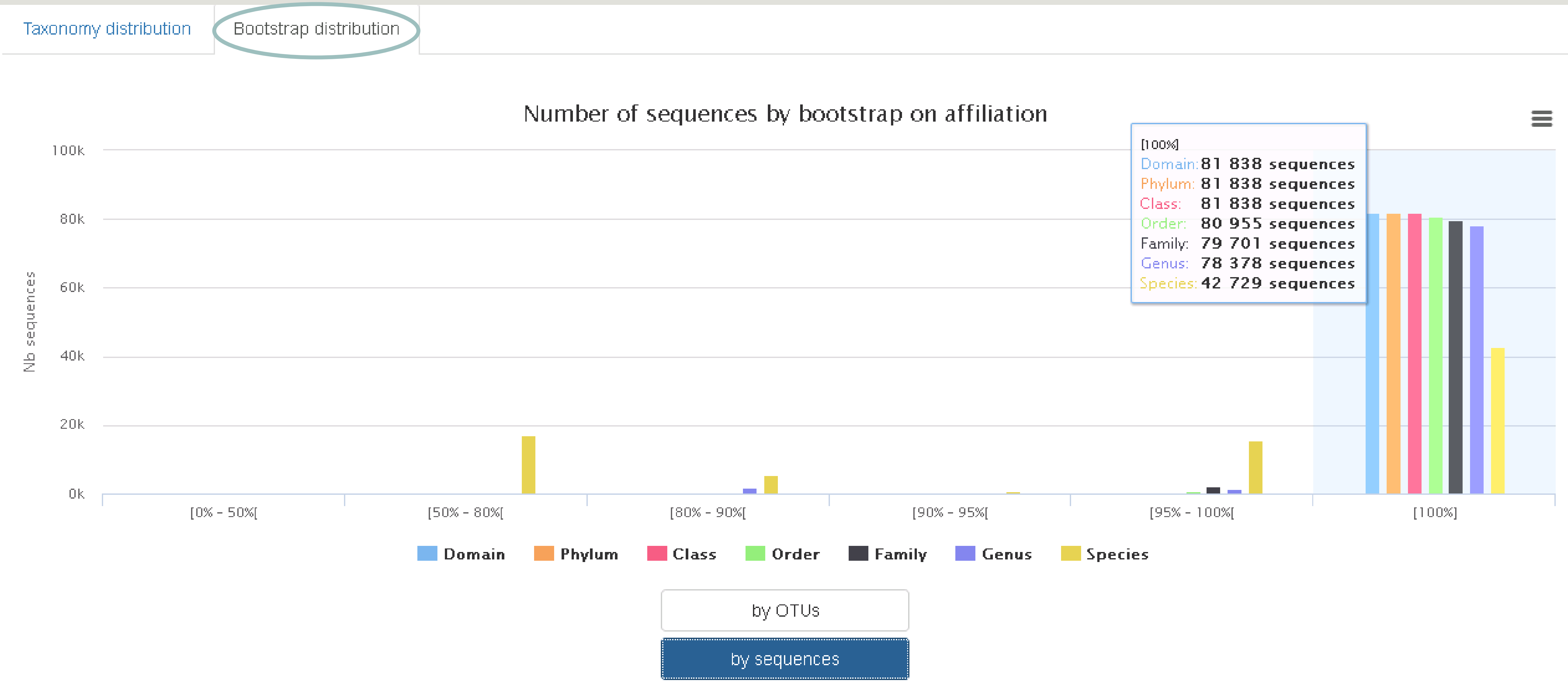
Clostridiales:
Total 4.218%
Firmicutes 51.380%

Detail on selected:

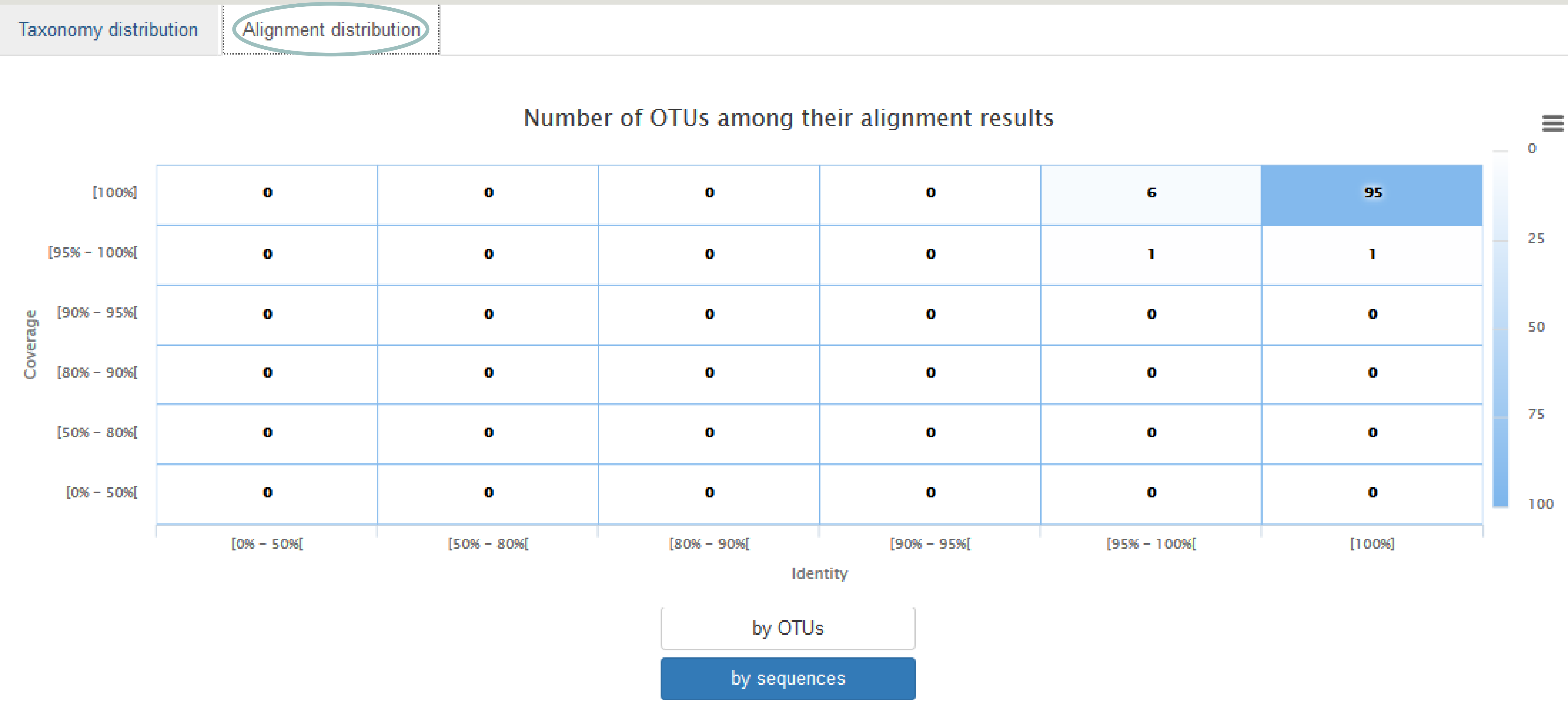
Name	Size	Global %	Parent %
root	246197		
Bacteria	246197	100.000	100.000
Firmicutes	20212	8.210	8.210
Clostridia	12142	4.932	60.073
Clostridiales	10385	4.218	85.530

Clostridiales nb children: 20

Font size: 15 Colors start depth: 2 Close

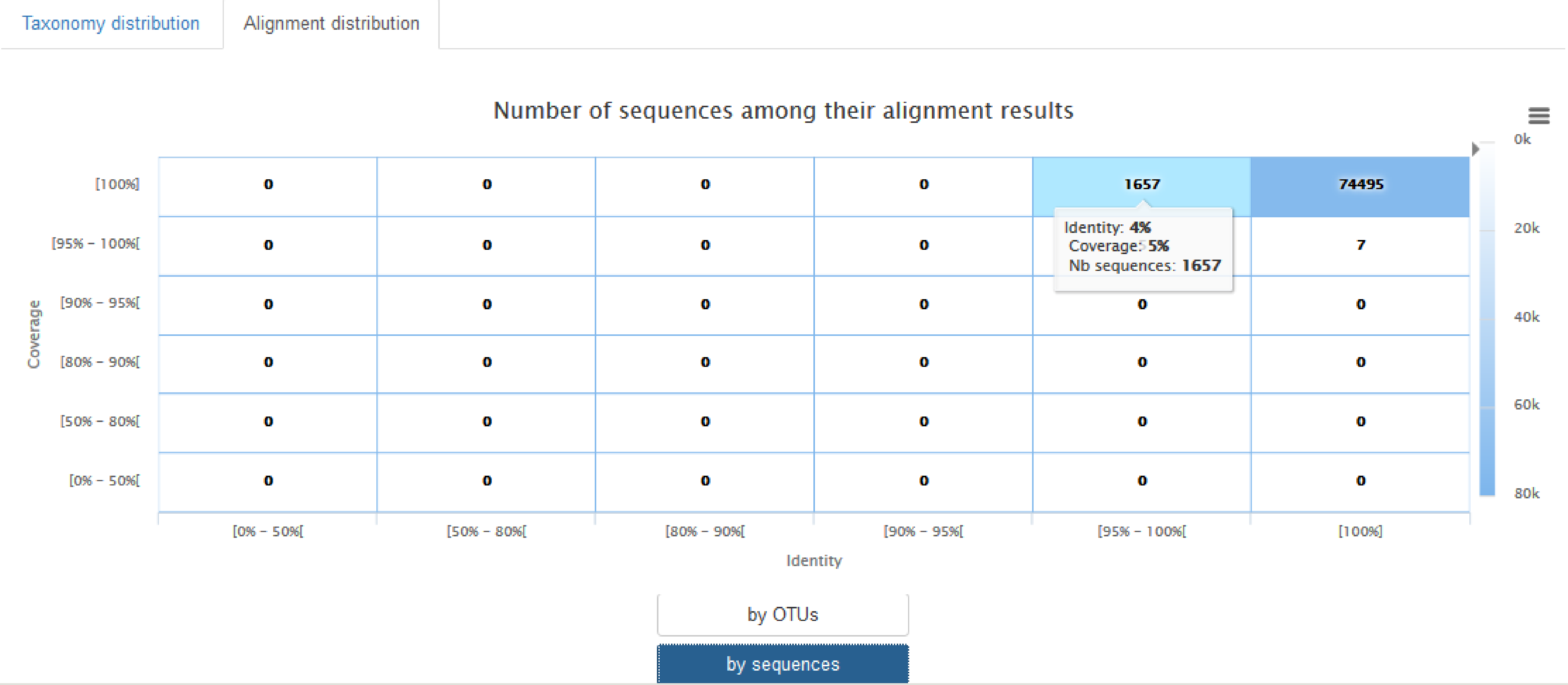


[100%]
Domain: 81 838 sequences
Phylum: 81 838 sequences
Class: 81 838 sequences
Order: 80 955 sequences
Family: 79 701 sequences
Genus: 78 378 sequences
Species: 42 729 sequences



Number of OTUs among their alignment results

Coverage \ Identity	[0% - 50%[	[50% - 80%[	[80% - 90%[	[90% - 95%[	[95% - 100%[	[100%]
[100%]	0	0	0	0	6	95
[95% - 100%[	0	0	0	0	1	1
[90% - 95%[	0	0	0	0	0	0
[80% - 90%[	0	0	0	0	0	0
[50% - 80%[	0	0	0	0	0	0
[0% - 50%[	0	0	0	0	0	0

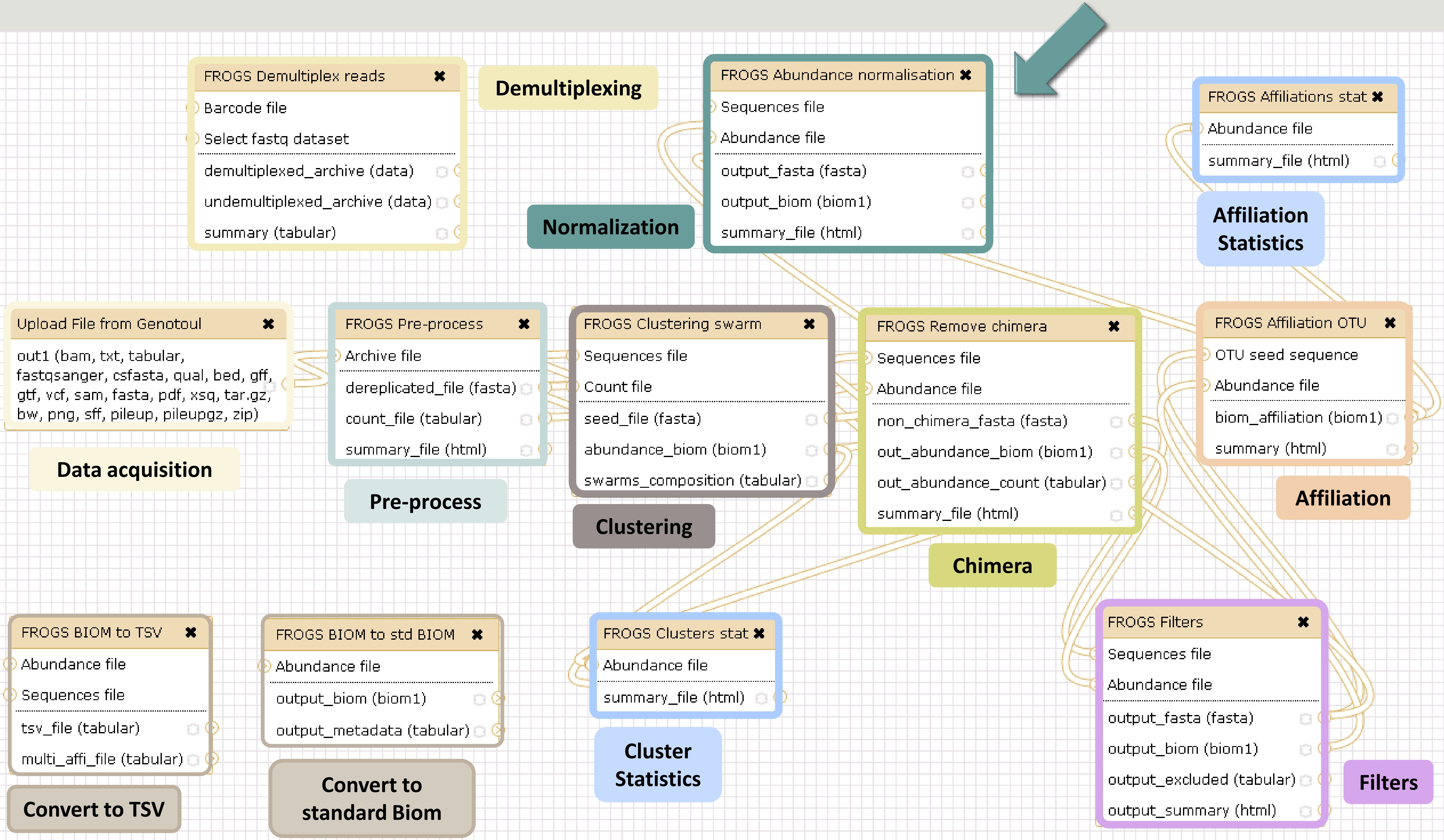


Number of sequences among their alignment results

Coverage \ Identity	[0% - 50%[	[50% - 80%[	[80% - 90%[	[90% - 95%[	[95% - 100%[	[100%]
[100%]	0	0	0	0	1657	74495
[95% - 100%[	0	0	0	0		7
[90% - 95%[	0	0	0	0	0	0
[80% - 90%[	0	0	0	0	0	0
[50% - 80%[	0	0	0	0	0	0
[0% - 50%[	0	0	0	0	0	0

Identity: **4%**
Coverage: **5%**
Nb sequences: **1657**

Normalization



FROGS Demultiplex reads ✖
Barcode file
Select fastq dataset
demultiplexed_archive (data)
undemultiplexed_archive (data)
summary (tabular)

Demultiplexing

FROGS Abundance normalisation ✖
Sequences file
Abundance file
output_fasta (fasta)
output_biom (biom1)
summary_file (html)

Normalization

FROGS Affiliations stat ✖
Abundance file
summary_file (html)

Affiliation Statistics

Upload File from Genotoul ✖
out1 (bam, txt, tabular, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png, sff, pileup, pileupgz, zip)

Data acquisition

FROGS Pre-process ✖
Archive file
dereplicated_file (fasta)
count_file (tabular)
summary_file (html)

Pre-process

FROGS Clustering swarm ✖
Sequences file
Count file
seed_file (fasta)
abundance_biom (biom1)
swarms_composition (tabular)

Clustering

FROGS Remove chimera ✖
Sequences file
Abundance file
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
out_abundance_count (tabular)
summary_file (html)

Chimera

FROGS Affiliation OTU ✖
OTU seed sequence
Abundance file
biom_affiliation (biom1)
summary (html)

Affiliation

FROGS BIOM to TSV ✖
Abundance file
Sequences file
tsv_file (tabular)
multi_affi_file (tabular)

Convert to TSV

FROGS BIOM to std BIOM ✖
Abundance file
output_biom (biom1)
output_metadata (tabular)

Convert to standard Biom

FROGS Clusters stat ✖
Abundance file
summary_file (html)

Cluster Statistics

FROGS Filters ✖
Sequences file
Abundance file
output_fasta (fasta)
output_biom (biom1)
output_excluded (tabular)
output_summary (html)

Filters

Normalization

Conserve a predefined number of sequence per sample:

- update Biom abundance file
- update seed fasta file

May be used when :

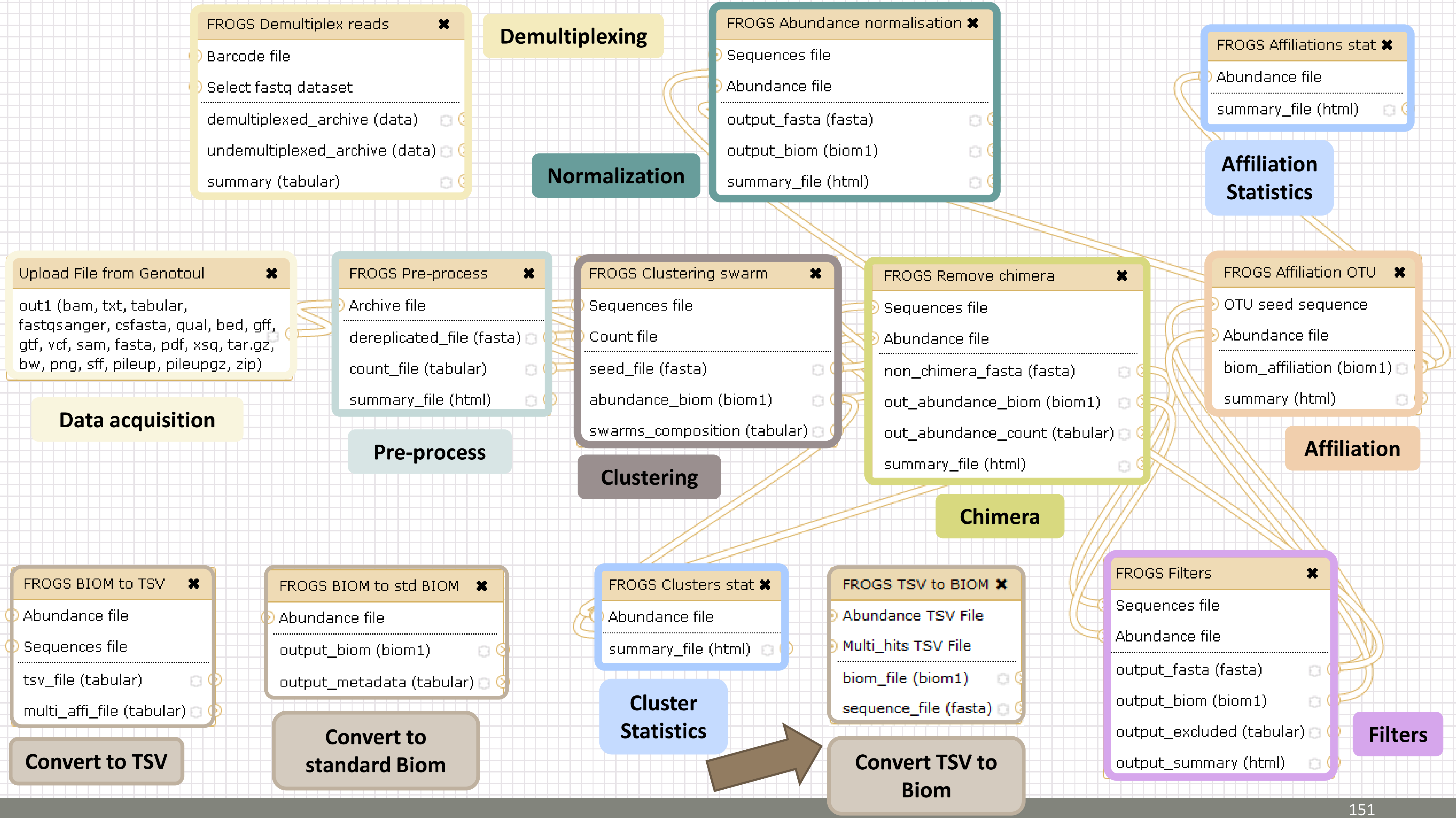
- Low sequencing sample
- Required for some statistical methods to compare the samples in pairs

Your Turn! – 8

Go to practice

EXERCISE 8

TSV to BIOM



FROGS Demultiplex reads ✖
Barcode file
Select fastq dataset
demultiplexed_archive (data)
undemultiplexed_archive (data)
summary (tabular)

Demultiplexing

FROGS Abundance normalisation ✖
Sequences file
Abundance file
output_fasta (fasta)
output_biom (biom1)
summary_file (html)

Normalization

FROGS Affiliations stat ✖
Abundance file
summary_file (html)

Affiliation Statistics

Upload File from Genotoul ✖
out1 (bam, txt, tabular, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png, sff, pileup, pileupgz, zip)

Data acquisition

FROGS Pre-process ✖
Archive file
dereplicated_file (fasta)
count_file (tabular)
summary_file (html)

Pre-process

FROGS Clustering swarm ✖
Sequences file
Count file
seed_file (fasta)
abundance_biom (biom1)
swarms_composition (tabular)

Clustering

FROGS Remove chimera ✖
Sequences file
Abundance file
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
out_abundance_count (tabular)
summary_file (html)

Chimera

FROGS Affiliation OTU ✖
OTU seed sequence
Abundance file
biom_affiliation (biom1)
summary (html)

Affiliation

FROGS BIOM to TSV ✖
Abundance file
Sequences file
tsv_file (tabular)
multi_affi_file (tabular)

Convert to TSV

FROGS BIOM to std BIOM ✖
Abundance file
output_biom (biom1)
output_metadata (tabular)

Convert to standard Biom

FROGS Clusters stat ✖
Abundance file
summary_file (html)

Cluster Statistics

FROGS TSV to BIOM ✖
Abundance TSV File
Multi_hits TSV File
biom_file (biom1)
sequence_file (fasta)

Convert TSV to Biom

FROGS Filters ✖
Sequences file
Abundance file
output_fasta (fasta)
output_biom (biom1)
output_excluded (tabular)
output_summary (html)

Filters

TSV to BIOM

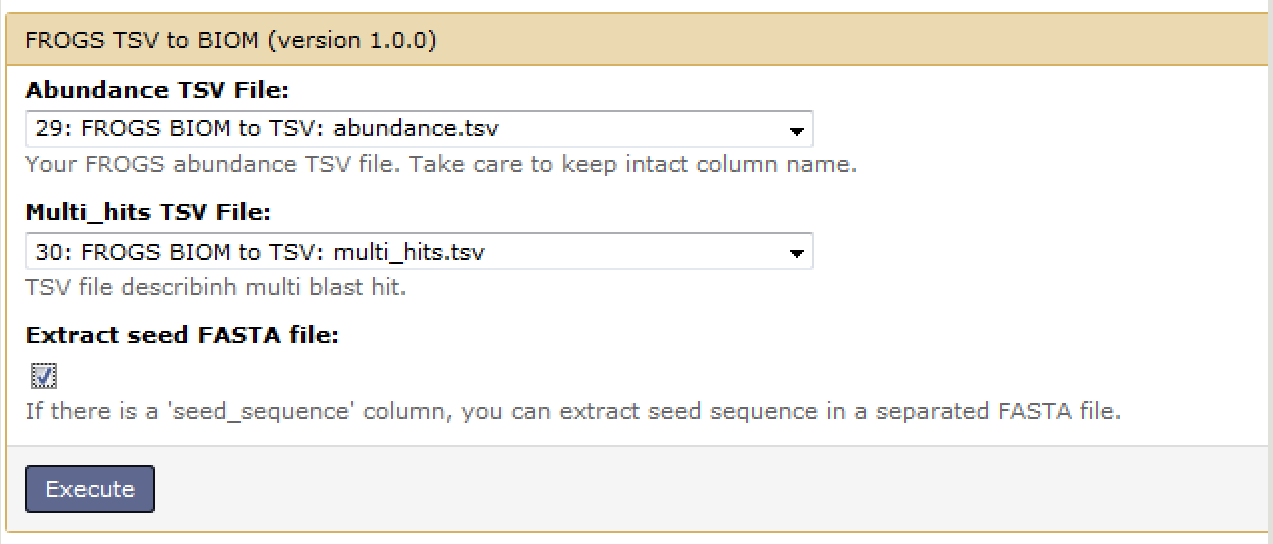
After modifying your abundance TSV file you can again:

- generate rarefaction curve
- sunburst

Careful :

- do not modify column name
- do not remove column
- take care to choose a taxonomy available in your multi_hit TSV file
- if deleting line from multihit, take care to not remove a complete cluster without removing all "multi tags" in you abundance TSV file.
- if you want to rename a taxon level (ex : genus "Ruminiclostridium 5;" to genus "Ruminiclostridium;"), do not forget to modify also your mult_hit TSV file.

TSV to BIOM



FROGS TSV to BIOM (version 1.0.0)

Abundance TSV File:

29: FROGS BIOM to TSV: abundance.tsv

Your FROGS abundance TSV file. Take care to keep intact column name.

Multi_hits TSV File:

30: FROGS BIOM to TSV: multi_hits.tsv

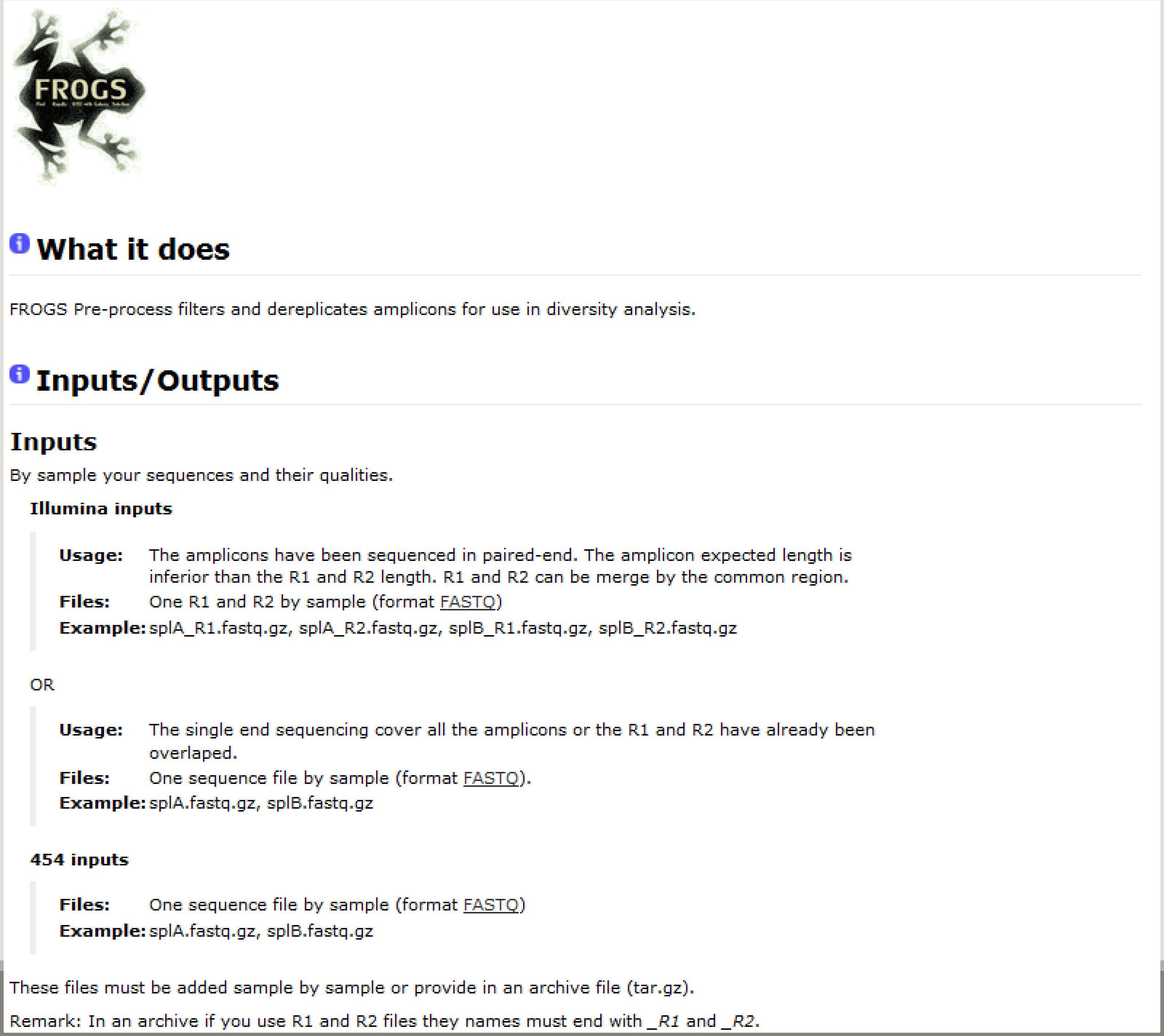
TSV file describinh multi blast hit.

Extract seed FASTA file:

☑

If there is a 'seed_sequence' column, you can extract seed sequence in a separated FASTA file.

Tool descriptions



What it does

FROGS Pre-process filters and dereplicates amplicons for use in diversity analysis.

Inputs/Outputs

Inputs

By sample your sequences and their qualities.

Illumina inputs

> **Usage:** The amplicons have been sequenced in paired-end. The amplicon expected length is inferior than the R1 and R2 length. R1 and R2 can be merge by the common region.
>
> **Files:** One R1 and R2 by sample (format FASTQ)
>
> **Example:** splA_R1.fastq.gz, splA_R2.fastq.gz, splB_R1.fastq.gz, splB_R2.fastq.gz

OR

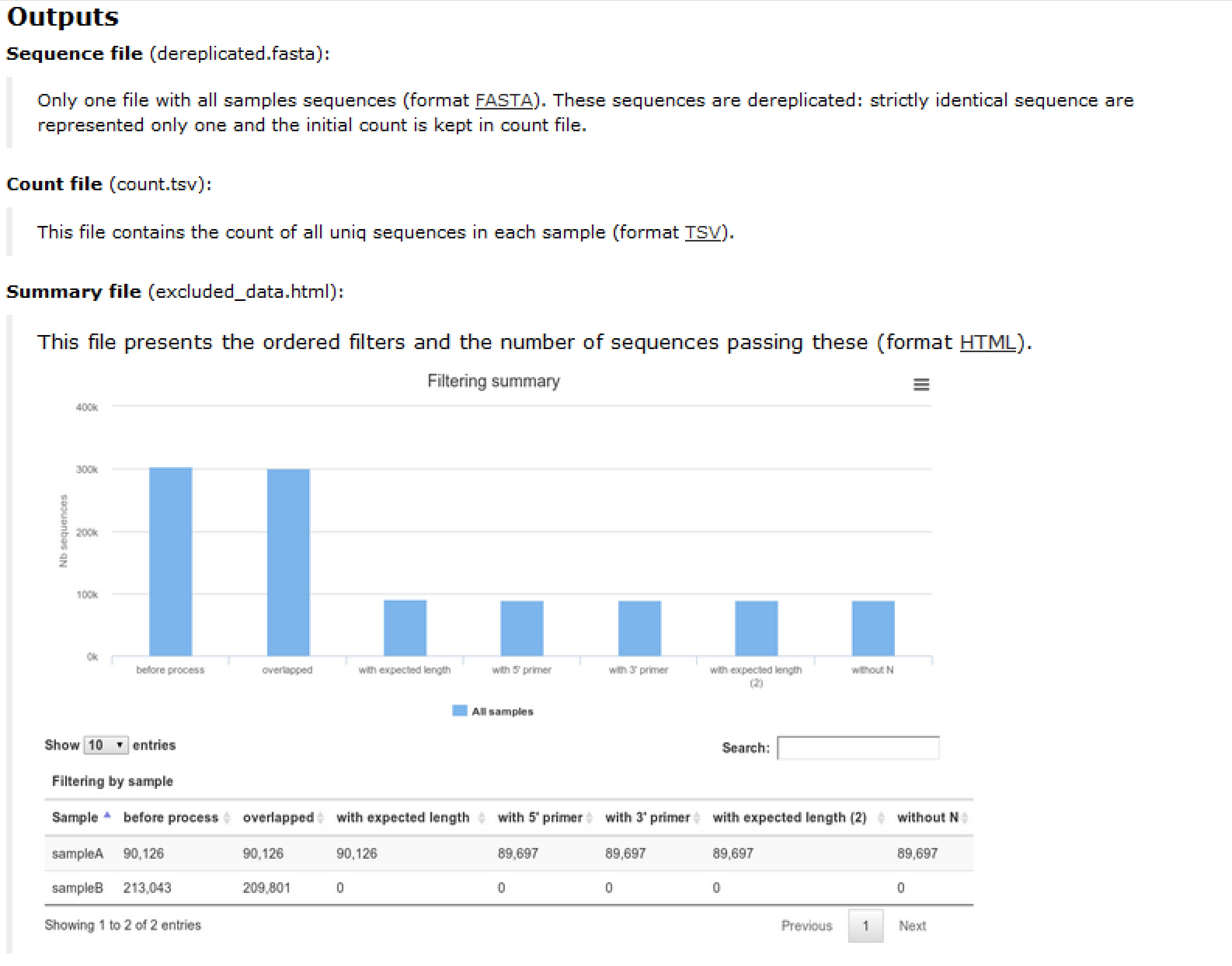
> **Usage:** The single end sequencing cover all the amplicons or the R1 and R2 have already been overlaped.
>
> **Files:** One sequence file by sample (format FASTQ).
>
> **Example:** splA.fastq.gz, splB.fastq.gz

454 inputs

> **Files:** One sequence file by sample (format FASTQ)
>
> **Example:** splA.fastq.gz, splB.fastq.gz

These files must be added sample by sample or provide in an archive file (tar.gz).

Remark: In an archive if you use R1 and R2 files they names must end with _*R1* and _*R2*.



Outputs

Sequence file (dereplicated.fasta):

> Only one file with all samples sequences (format FASTA). These sequences are dereplicated: strictly identical sequence are represented only one and the initial count is kept in count file.

Count file (count.tsv):

> This file contains the count of all uniq sequences in each sample (format TSV).

Summary file (excluded_data.html):

> This file presents the ordered filters and the number of sequences passing these (format HTML).

Show 10 entries

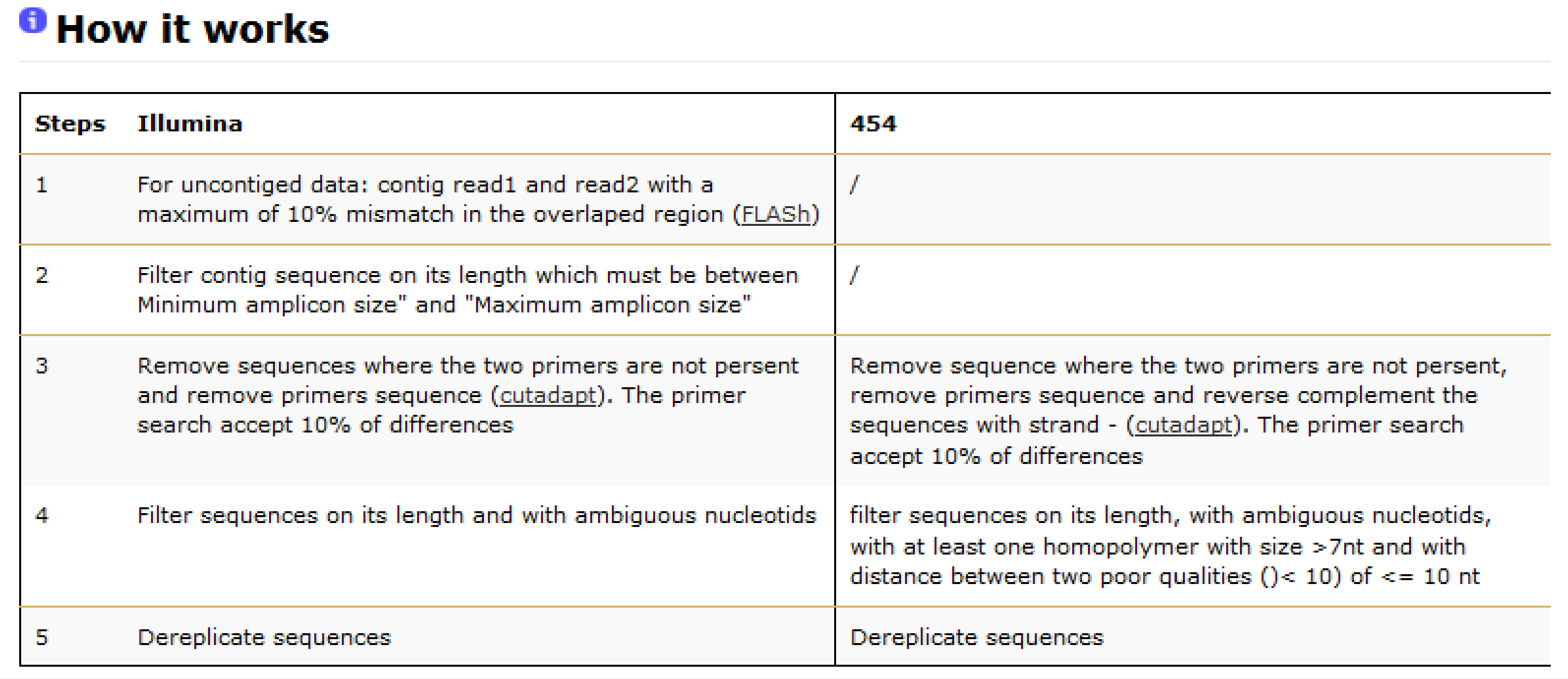
Search:

Filtering by sample

Sample	before process	overlapped	with expected length	with 5' primer	with 3' primer	with expected length (2)	without N
sampleA	90,126	90,126	90,126	89,697	89,697	89,697	89,697
sampleB	213,043	209,801	0	0	0	0	0

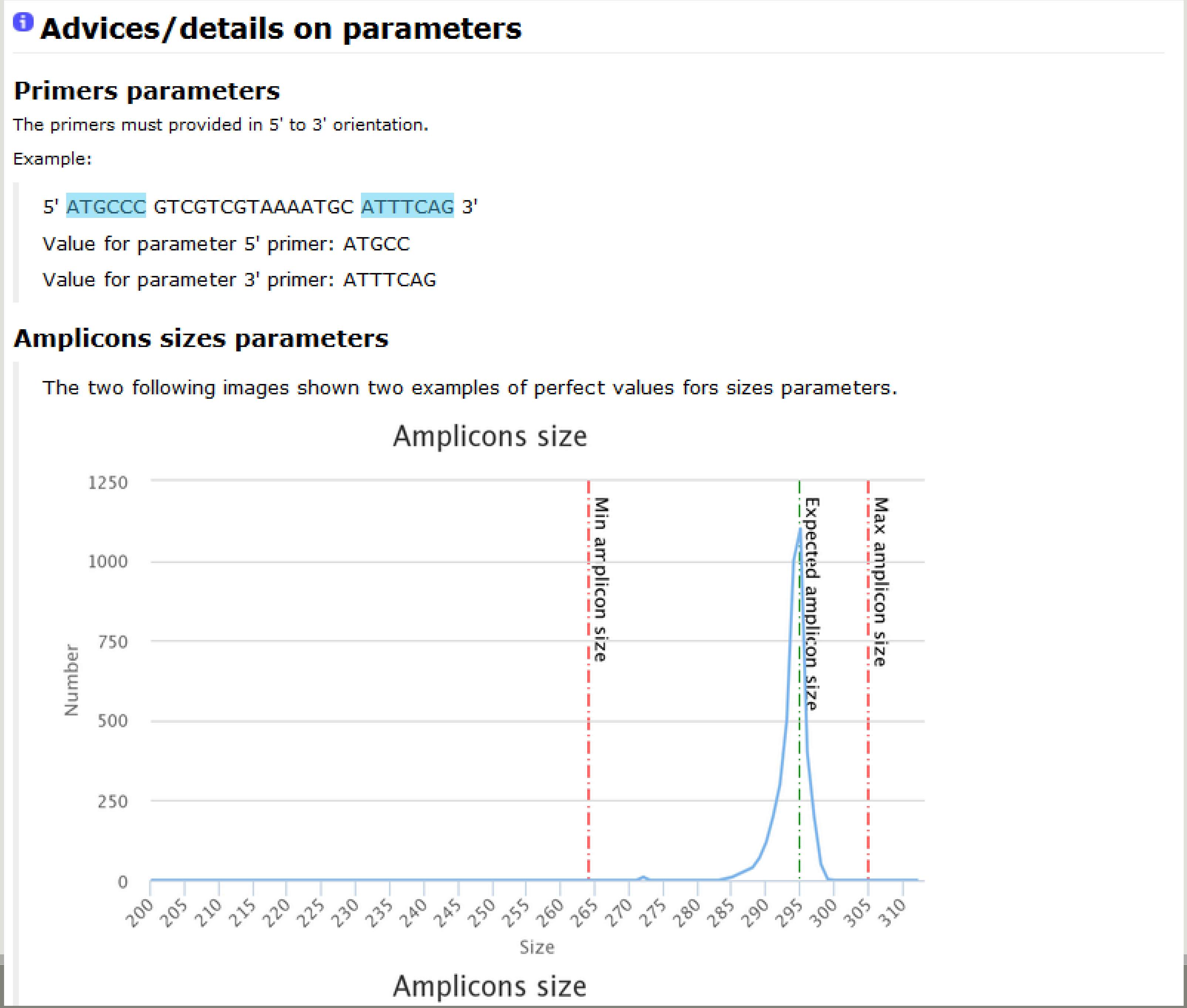
Showing 1 to 2 of 2 entries

Previous 1 Next



How it works

Steps	Illumina	454
1	For uncontiged data: contig read1 and read2 with a maximum of 10% mismatch in the overlaped region (FLASh)	/
2	Filter contig sequence on its length which must be between Minimum amplicon size" and "Maximum amplicon size"	/
3	Remove sequences where the two primers are not persent and remove primers sequence (cutadapt). The primer search accept 10% of differences	Remove sequence where the two primers are not persent, remove primers sequence and reverse complement the sequences with strand - (cutadapt). The primer search accept 10% of differences
4	Filter sequences on its length and with ambiguous nucleotids	filter sequences on its length, with ambiguous nucleotids, with at least one homopolymer with size >7nt and with distance between two poor qualities ()< 10) of <= 10 nt
5	Dereplicate sequences	Dereplicate sequences



Advices/details on parameters

Primers parameters

The primers must provided in 5' to 3' orientation.

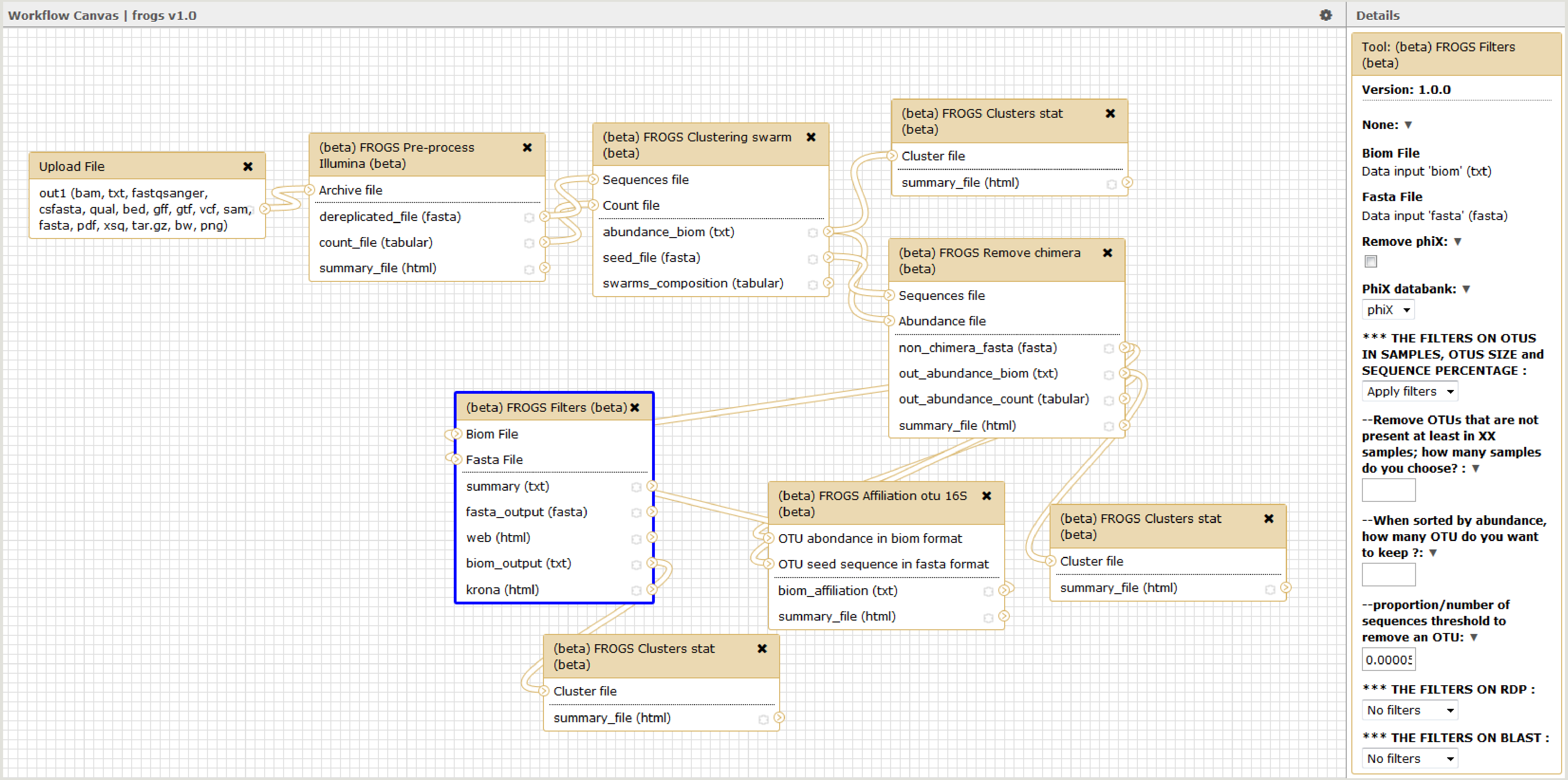
Example:

> 5' ATGCCC GTCGTCGTAAAATGC ATTTCAG 3'
>
> Value for parameter 5' primer: ATGCC
>
> Value for parameter 3' primer: ATTTCAG

Amplicons sizes parameters

> The two following images shown two examples of perfect values fors sizes parameters.

Workflow creation


Workflow Canvas | frogs v1.0
Upload File
out1 (bam, txt, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png)
(beta) FROGS Pre-process Illumina (beta)
Archive file
dereplicated_file (fasta)
count_file (tabular)
summary_file (html)
(beta) FROGS Clustering swarm (beta)
Sequences file
Count file
abundance_biom (txt)
seed_file (fasta)
swarms_composition (tabular)
(beta) FROGS Clusters stat (beta)
Cluster file
summary_file (html)
(beta) FROGS Remove chimera (beta)
Sequences file
Abundance file
non_chimera_fasta (fasta)
out_abundance_biom (txt)
out_abundance_count (tabular)
summary_file (html)
(beta) FROGS Filters (beta)
Biom File
Fasta File
summary (txt)
fasta_output (fasta)
web (html)
biom_output (txt)
krona (html)
(beta) FROGS Affiliation otu 16S (beta)
OTU abondance in biom format
OTU seed sequence in fasta format
biom_affiliation (txt)
summary_file (html)
(beta) FROGS Clusters stat (beta)
Cluster file
summary_file (html)
(beta) FROGS Clusters stat (beta)
Cluster file
summary_file (html)
Details

Tool: (beta) FROGS Filters (beta)

Version: 1.0.0

None: ▼

Biom File
Data input 'biom' (txt)

Fasta File
Data input 'fasta' (fasta)

Remove phiX: ▼
□

PhiX databank: ▼
phiX ▾

***** THE FILTERS ON OTUS IN SAMPLES, OTUS SIZE and SEQUENCE PERCENTAGE :**
Apply filters ▾

--Remove OTUs that are not present at least in XX samples; how many samples do you choose? : ▼

--When sorted by abundance, how many OTU do you want to keep ?: ▼

--proportion/number of sequences threshold to remove an OTU: ▼
0.00005

***** THE FILTERS ON RDP :**
No filters ▾

***** THE FILTERS ON BLAST :**
No filters ▾

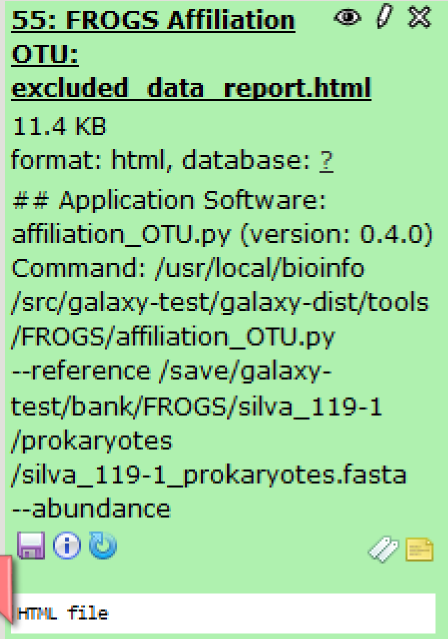
Your Turn! – 9

Go to practice

EXERCISE 9

Download your data

You have to download one per one your files



55: FROGS Affiliation OTU: excluded_data_report.html

11.4 KB

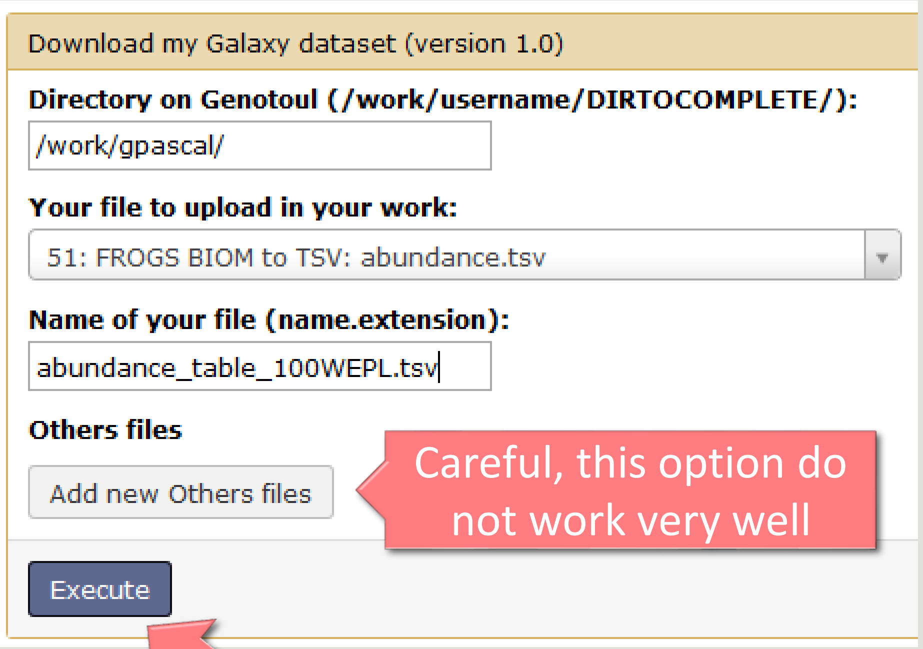
format: html, database: ?

Application Software: affiliation_OTU.py (version: 0.4.0) Command: /usr/local/bioinfo /src/galaxy-test/galaxy-dist/tools /FROGS/affiliation_OTU.py --reference /save/galaxy-test/bank/FROGS/silva_119-1 /prokaryotes /silva_119-1_prokaryotes.fasta --abundance

HTML file

OR

This tool will save your datasets in your work on genotoul (/work/username/dataset-archive-XXX.tar.gz). Then, you could work on these files in your work on Genotoul.



Download my Galaxy dataset (version 1.0)

Directory on Genotoul (/work/username/DIRTOCOMPLETE/):

/work/gpascal/

Your file to upload in your work:

51: FROGS BIOM to TSV: abundance.tsv

Name of your file (name.extension):

abundance_table_100WEPL.tsv

Others files

Add new Others files

Careful, this option do not work very well

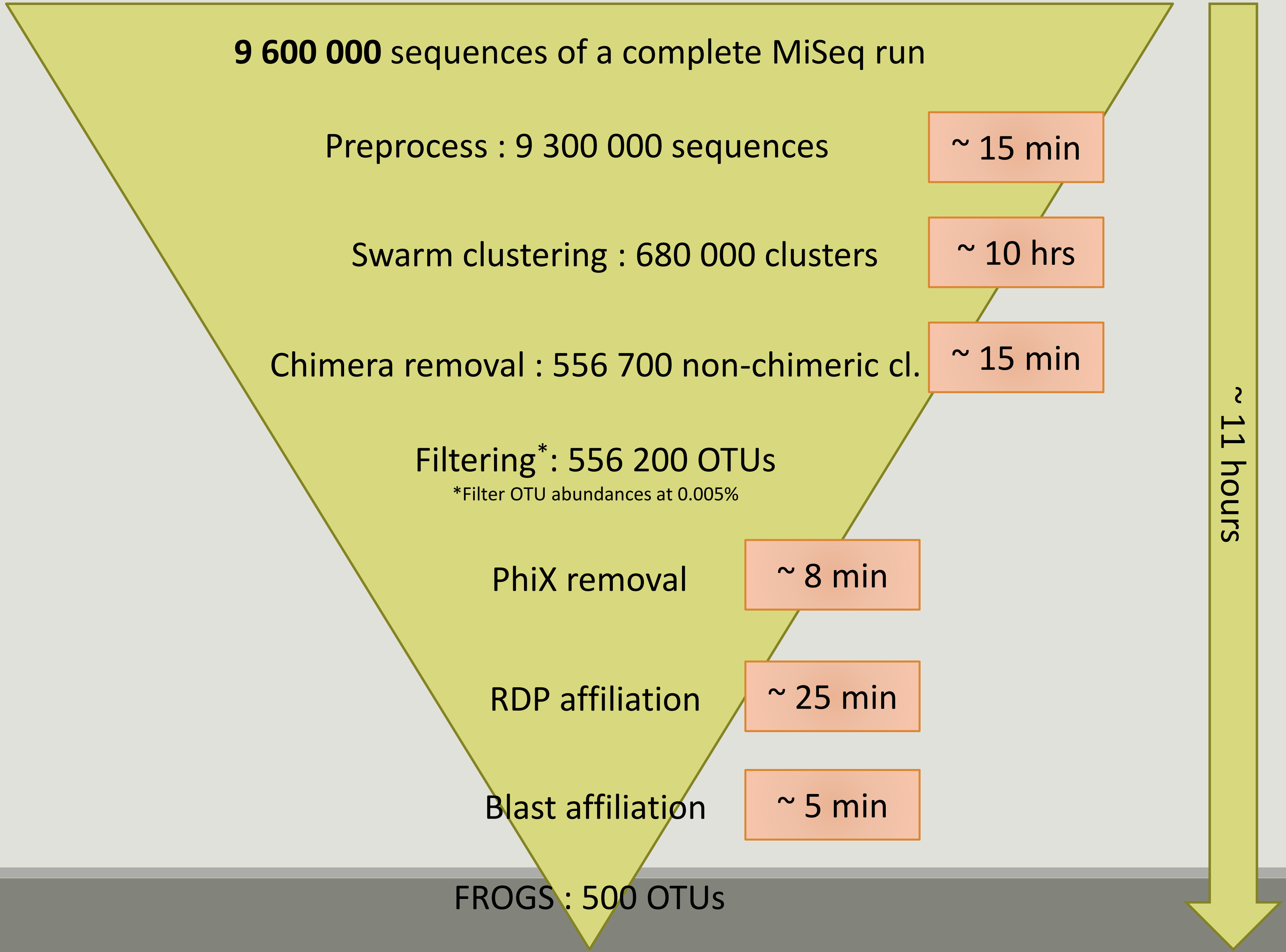
Execute

Some figures

Some figures - Fast

NB SEQ	TIME with complete pipeline without Filters
50 000	40 min
400 000	4 hrs
3 500 000	2 days
10 000 000	5 days

Speed on real datasets



9 600 000 sequences of a complete MiSeq run

Preprocess : 9 300 000 sequences ~ 15 min

Swarm clustering : 680 000 clusters ~ 10 hrs

Chimera removal : 556 700 non-chimeric cl. ~ 15 min

Filtering*: 556 200 OTUs

*Filter OTU abundances at 0.005%

PhiX removal ~ 8 min

RDP affiliation ~ 25 min

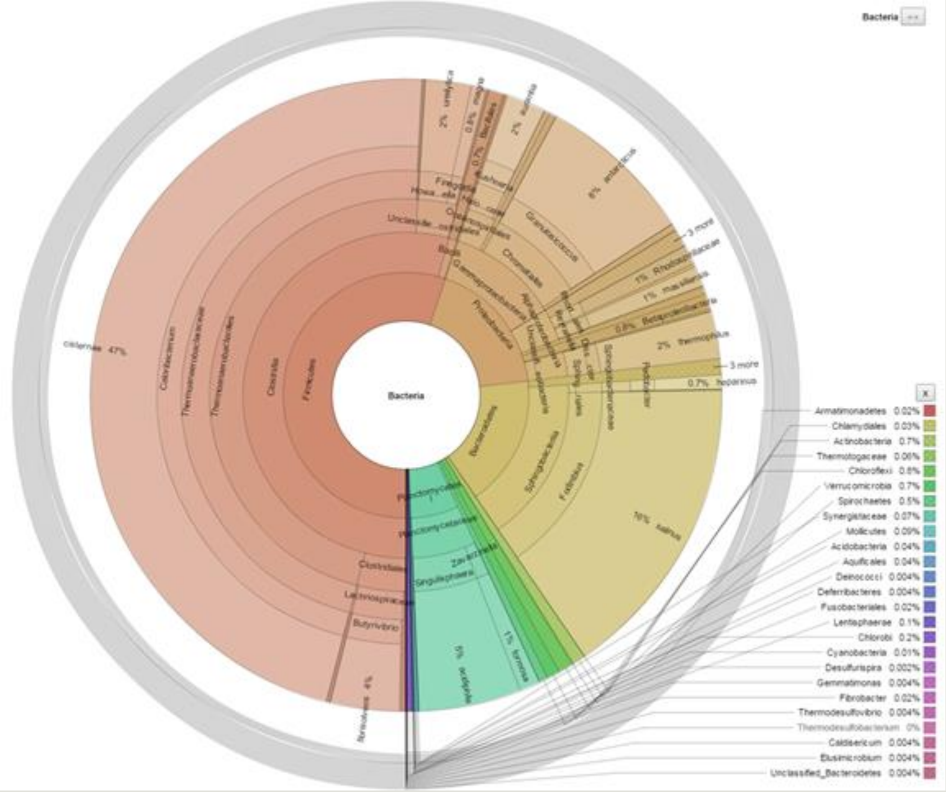
Blast affiliation ~ 5 min

FROGS : 500 OTUs

~ 11 hours

Simulated datasets, for testing FROGS' Accuracy

- 500 species, covering all bacterial phyla
- Power Law distribution of the species abundances
- Error rate calibrated with real sequencing runs
- 20% chimeras
- 10 samples of 100 000 sequences each (1M sequences)



Simulated dataset : 1M sequences

↓

SWARM : 109 000 clusters

↓

VSEARCH: 21 000 clusters

↓

filters : 0.005% **505 OTUs**

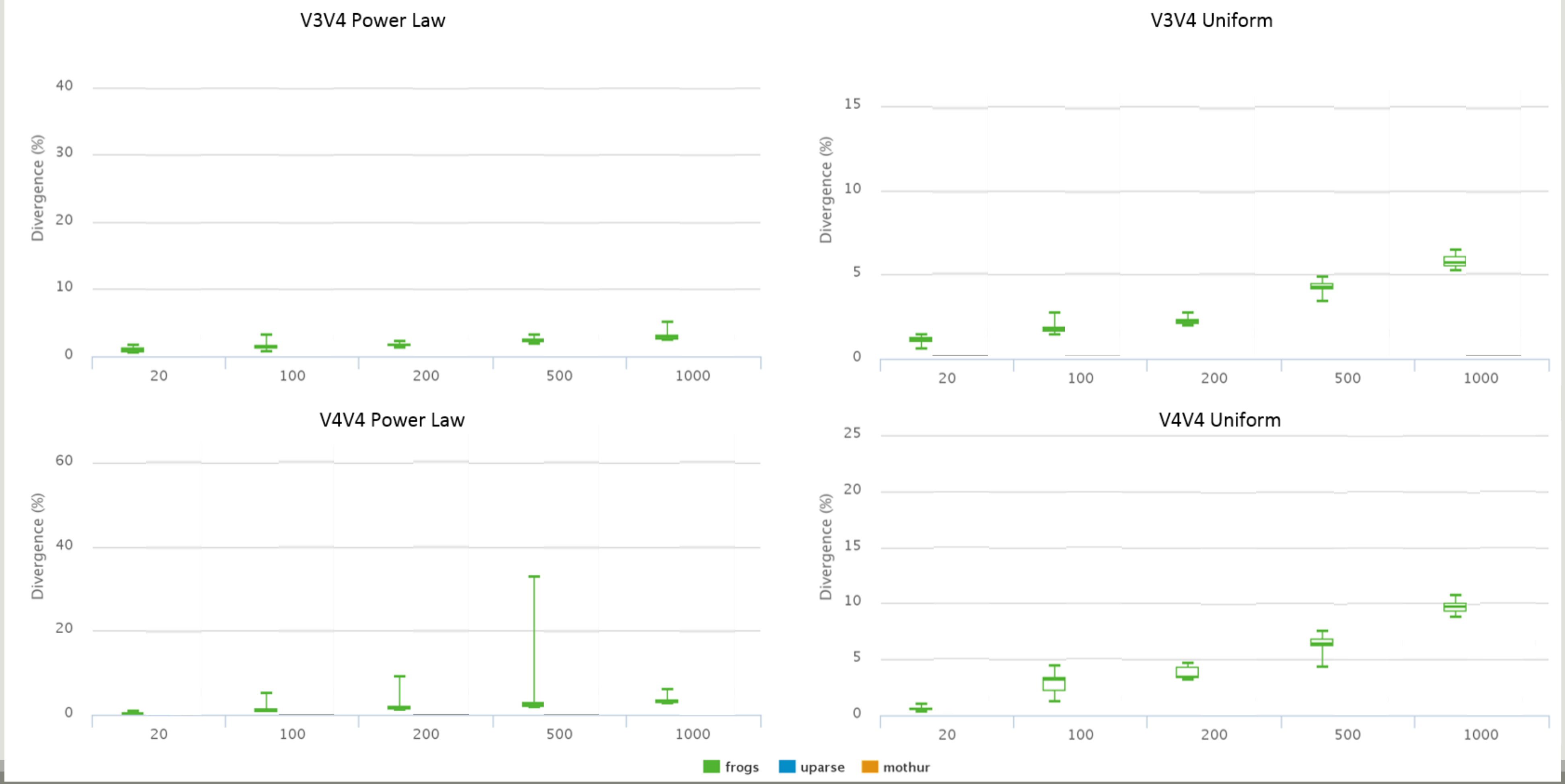
FROGS’ Accuracy

- 10 artificial samples of 100 000 sequences
- 25 sets of species
- 20, 100, 200, 500 or 1000 different species
- power law or a uniform distribution
- 5 to 20% of chimera
- 1.10^{+11} sequences were treated with FROGS, UPARSE and MOTHUR, with their guidelines, to compare their performances

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

FROGS' Accuracy

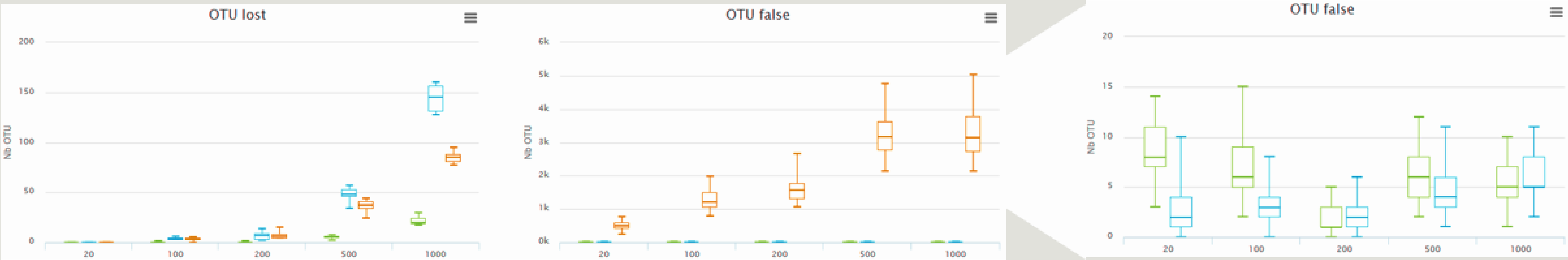
→ divergence at "genus" rank



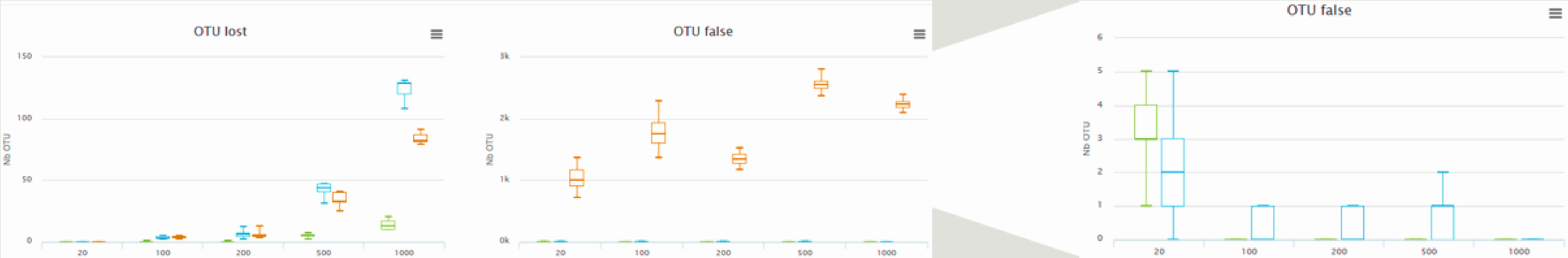
FROGS' Accuracy

→ Lost & False OTU

V3V4 Power Law



V3V4 Uniform

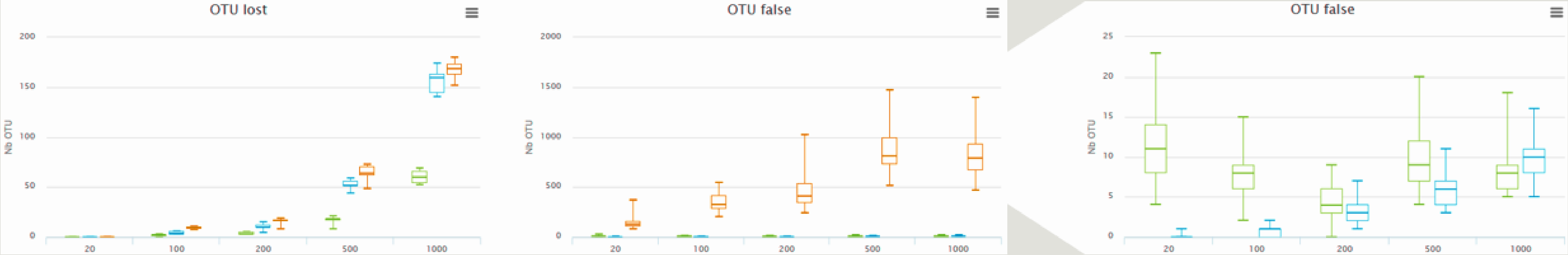


🟩 frogs 🟦 uparse 🟧 mothur

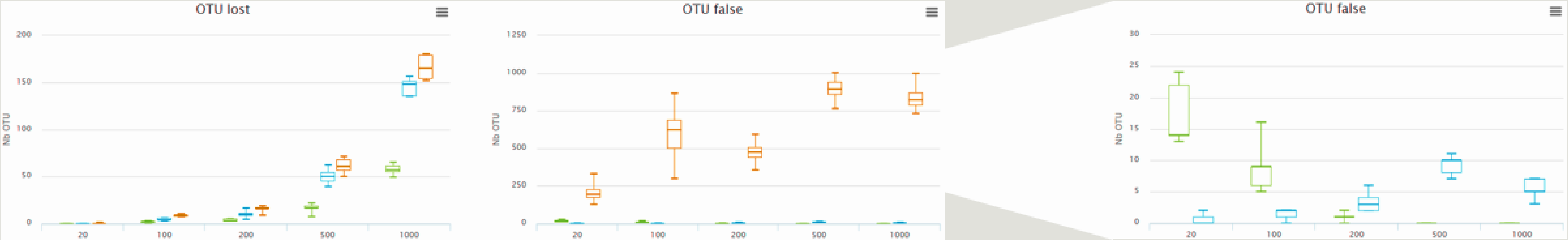
FROGS’ Accuracy

→ Lost & False OTU

V4V4 Power Law



V4V4 Uniform



🟩 frogs 🟦 uparse 🟧 mothur

Conclusions

Why Use FROGS ?

- User-friendly
- Fast
- 454 data and Illumina data
  - sequencing methods change but same tool
  - easier for comparisons
- Clustering without global threshold and independent of sequence order
- New chimera removal method (Vsearch + cross-validation)
- Filters tool
- Multi-affiliation with 2 taxonomy affiliation procedures
- Cluster Stat and Affiliation Stat tools
- A lot of graphics
- Independant tools

How to cite FROGS

In waiting for the publication:

Pipeline FROGS on http://sigenae-workbench.toulouse.inra.fr/

Poster FROGS: Escudie F., Auer L., Bernard M., Cauquil L., Vidal K., Maman S., Mariadassou M., Hernadez-Raquet G., Pascal G., 2015. FROGS: Find Rapidly OTU with Galaxy Solution. In: Environmental Genomics 2015, Montpellier, France, http://bioinfo.genotoul.fr/fileadmin/user_upload/FROGS_2015_GE_Montpellier_poster.pdf

To contact

FROGS:

frogs@toulouse.inra.fr

Galaxy:

sigenae-support@listes.inra.fr

Newsletter – demande d’abonnement:

mailto:sympa@listes.inra.fr?subject=sub%20frogs-newsletter

frogs-newsletter-request@listes.inra.fr

Next training sessions

20th to 23th June 2016 (complete) and 10th or 13th October 2016

4 days : 1 Galaxy day
2 FROGS days
1 Statistics phyloseq day (under R)

Galaxy e-learning (user account)

And soon FROGS e-learning

If we have time

- Play with TSV to BIOM.
- Change clustering option ad compare.
- Make a phylogenetic tree from sequences.fasta built with Filter Tool.

  → use the document about phylogeny.fr