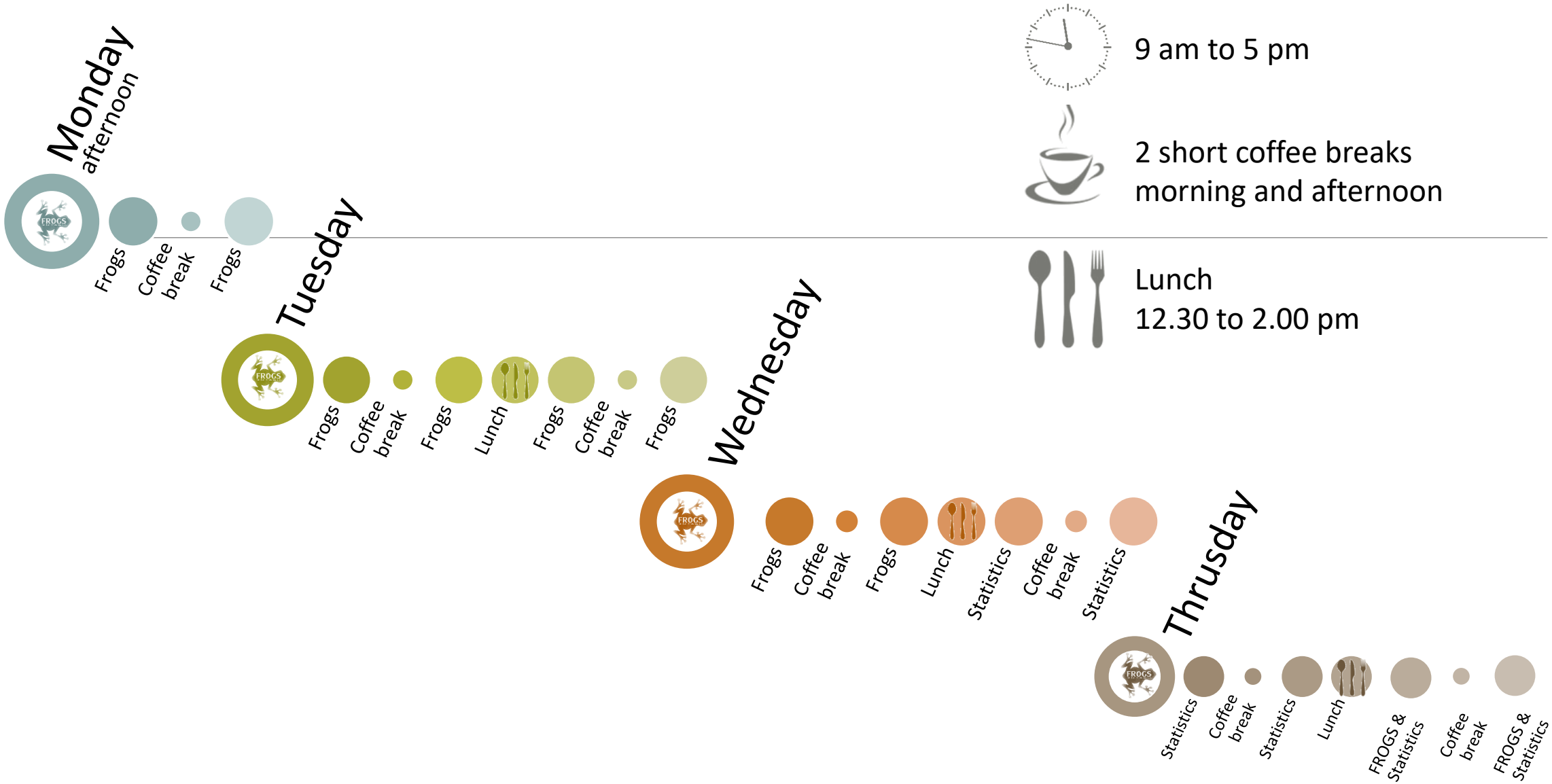
# Training on Galaxy: Metagenomics

March 2019

# Find, Rapidly, OTUs with Galaxy Solution

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

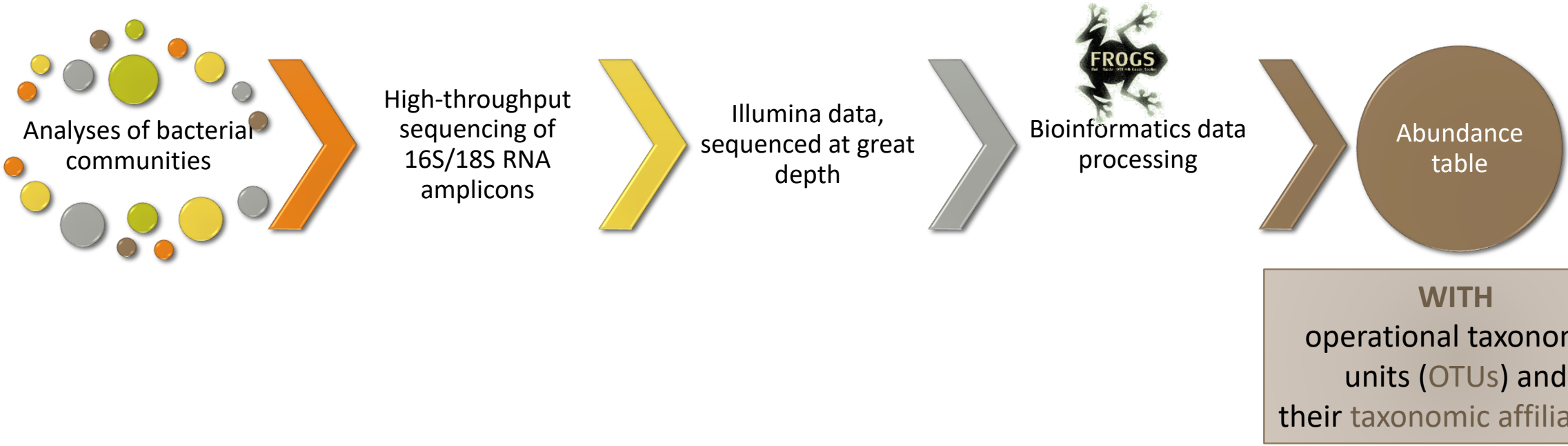
*THESE AUTHORS HAVE CONTRIBUTED EQUALLY TO THE PRESENT WORK.


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
FROGS & Statistics
Coffee break
FROGS & 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
- Demultiplex tool
- Preprocessing
- Clustering + Cluster Statistics
- Chimera removal
- Filtering
- Affiliation + Affiliation Statistics
- Normalization
- Tool descriptions
- Format transformation
- Export your data
- Some figures
- ITS analysis
- Workflow creation

# Objectives


Analyses of bacterial communities
High-throughput sequencing of 16S/18S RNA amplicons
Illumina data, sequenced at great depth
FROGS
Bioinformatics data processing
Abundance table
WITH
operational taxonomic units (OTUs) and their taxonomic affiliation.

# OTUs for ecology

Operational Taxonomy Unit:
a grouping of similar sequences that can be treated as a single « species »

Strengths:

- Conceptually simple
- Mask effect of poor quality data
  - Sequencing error
  - In vitro recombination (chimera)

Weaknesses:

- Limited resolution
- Logically inconsistent definition

# 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 |

# Why FROGS was developed ?

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 ».***

# Who is in the FROGS group?

Maria BERNARD

Olivier RUÉ

Frédéric ESCUDIÉ

## Developers

Lucas AUER

Laurent CAUQUIL

Sylvie COMBES

Guillermina HERNANDEZ-RAQUET

## Biology experts

Sarah MAMAN

## Galaxy support

Mahendra MARIADASSOU

## Statistical expert

Géraldine PASCAL

## Coordinator

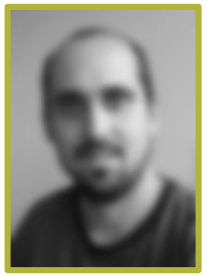
# Who is in the FROGS group?

Maria BERNARD


Olivier RUÉ

Frédéric ESCUDIÉ

## Developers



Lucas AUER

Laurent CAUQUIL

Sylvie COMBES

Guillermina HERNANDEZ-RAQUET

## Biology experts

Sarah MAMAN

## Galaxy support

Mahendra MARIADASSOU

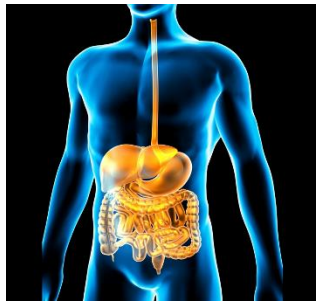
## Statistical expert

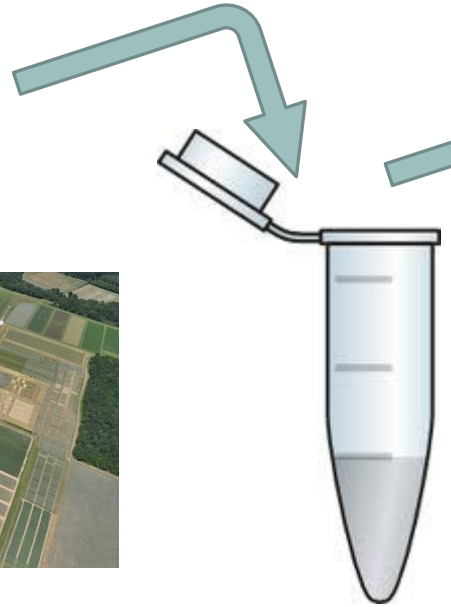
Géraldine PASCAL

## Coordinator

# Material

# Sample collection and DNA extraction





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

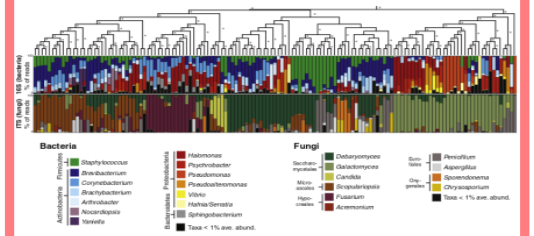
DNA

RNA

**Metagenomics**

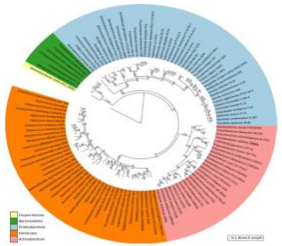
**Metatranscriptomics**

**Amplicon sequencing**



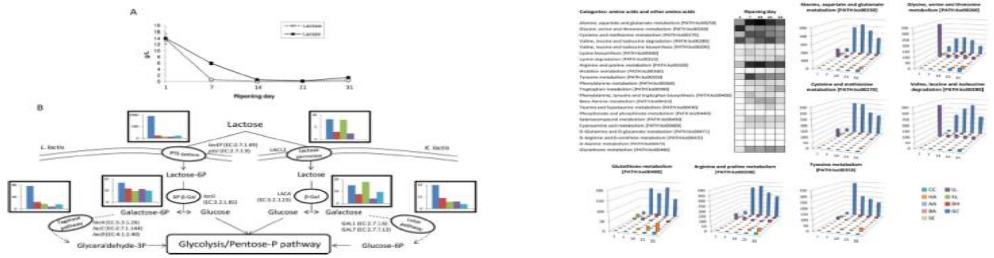

Wolfe *et al.*, 2014

**Shotgun sequencing**



Almeida *et al.*, 2014

**RNA sequencing**




Dugat-Bony *et al.*, 2015

Who is here?

What can they do?

What are they doing?

# The gene encoding the small subunit of the ribosomal RNA

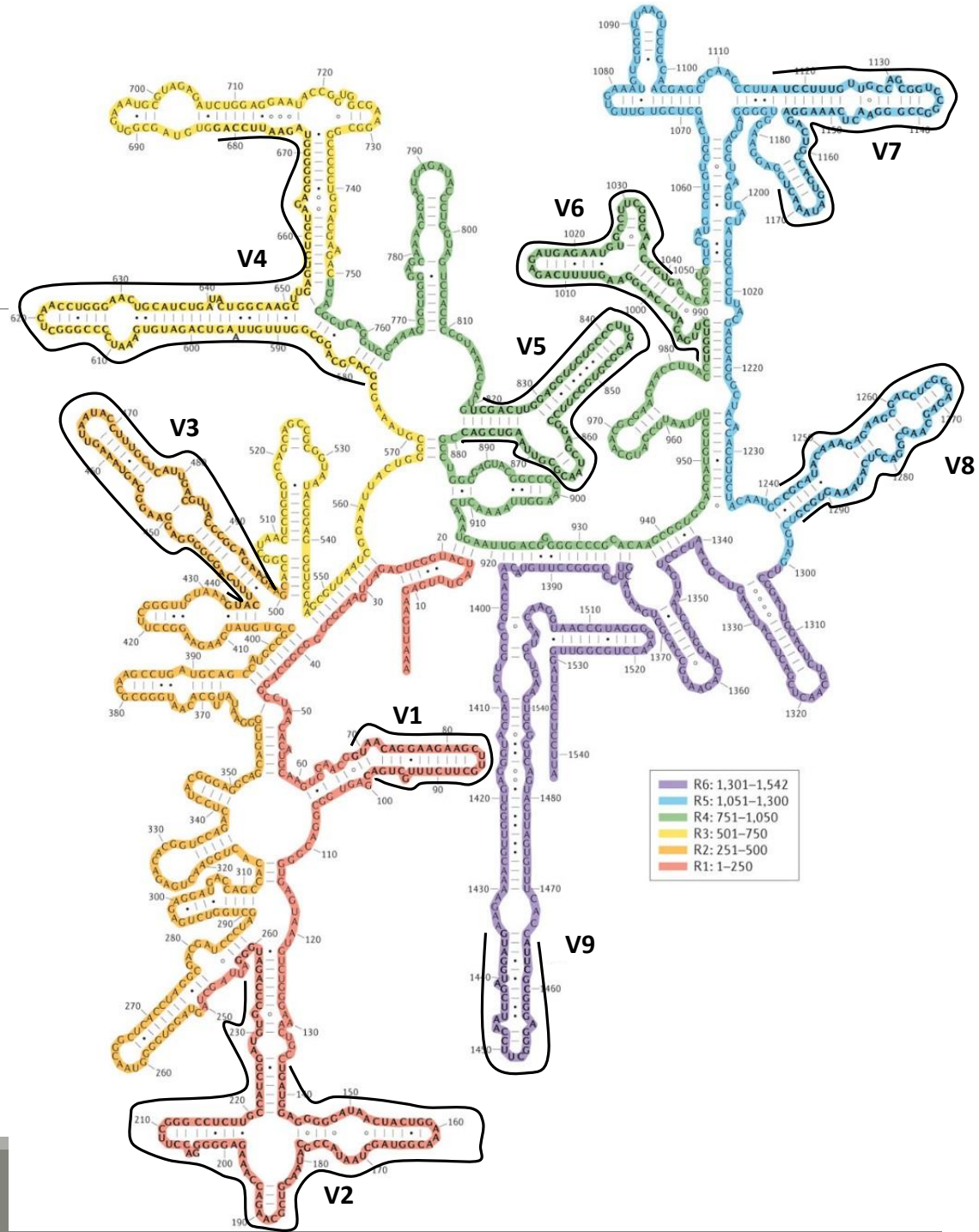
The most widely used gene in **molecular phylogenetic** studies

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)

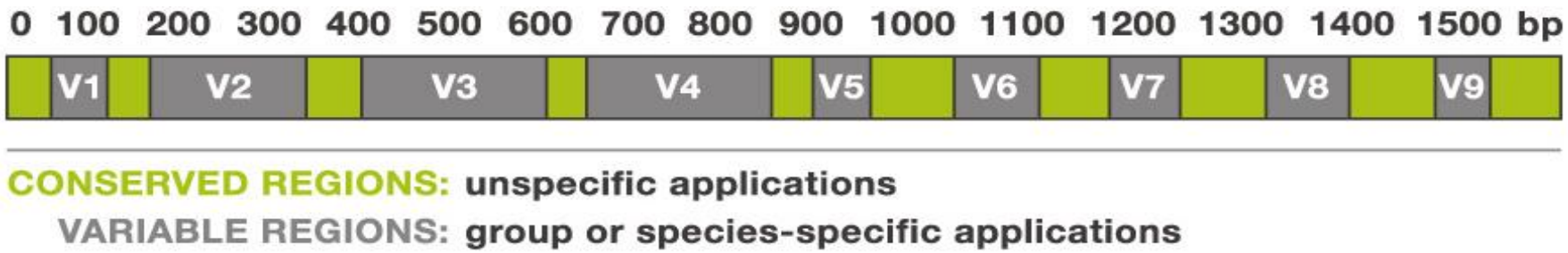



## Secondary structure of the 16S rRNA of *Escherichia coli*

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



# Other targets

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

The genes that have been proposed for this task include those encoding :

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

# Other targets

- gyrB has a higher rate of base substitution than 16S rDNA does, and shows promise for community-profiling applications.
- This gene is essential and ubiquitous in bacteria and
- is sufficiently large in size for use in analysis of microbial communities.
- It is a single-copy housekeeping gene that encodes the subunit B of DNA gyrase, a type II DNA topoisomerase, and therefore plays an essential role in DNA replication.
- Furthermore, the gyrB gene is also present in Eukarya and sometimes in Archaea but it shows enough sequence dissimilarity between the three domains of life to be used selectively for Bacteria.

# Other target

See for gyrB :

Article of Stéphane Chaillou

PLOS ONE

RESEARCH ARTICLE

## Deciphering intra-species bacterial diversity of meat and seafood spoilage microbiota using *gyrB* amplicon sequencing: A comparative analysis with 16S rDNA V3-V4 amplicon sequencing

**Simon Poirier[1], Olivier Rué[2], Raphaëlle Peguilhan[1], Gwendoline Coeuret[1], Monique Zagorec[3], Marie-Christine Champomier-Vergès[1], Valentin Loux[2], Stéphane Chaillou[1]***

**1** MICALIS, INRA, AgroParisTech, Université Paris-Saclay, Jouy-en-Josas, France, **2** MaIAGE, INRA, Université Paris-Saclay, Jouy-en-Josas, France, **3** Secalim, INRA, Oniris, Nantes, France

* stephane.chaillou@inra.fr



OPEN ACCESS

**Citation:** Poirier S, Rué O, Peguilhan R, Coeuret G, Zagorec M, Champomier-Vergès M-C, et al. (2018) Deciphering intra-species bacterial diversity of meat and seafood spoilage microbiota using *gyrB* amplicon sequencing: A comparative analysis with 16S rDNA V3-V4 amplicon sequencing. PLoS ONE 13(9): e0204629. https://doi.org/10.1371/journal.pone.0204629

**Editor:** George-John Nychas, Agricultural University of Athens, GREECE

**Received:** July 6, 2018

**Accepted:** September 11, 2018

**Published:** September 25, 2018



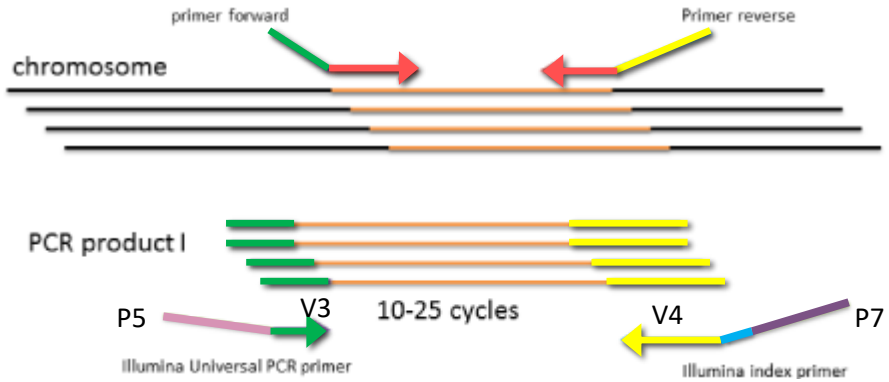
**Data Availability Statement:** Raw read sequences were deposited at the Sequence Read Archive under the accession numbers SAMN09070427 to SAMN09070506. The whole dataset has been uploaded to figshare and is accessible using the following DOI: 10.6084/m9.figshare.7083209. The R script (redlosses_phyloseq_custom.R), which includes all commands performed to create our figures, is available for download at DOI: 10.6084/m9.figshare.7083254.

### Abstract

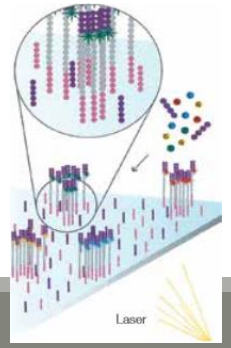
Meat and seafood spoilage ecosystems harbor extensive bacterial genomic diversity that is mainly found within a small number of species but within a large number of strains with different spoilage metabolic potential. To decipher the intraspecies diversity of such microbiota, traditional metagenetic analysis using the 16S rRNA gene is inadequate. We therefore assessed the potential benefit of an alternative genetic marker, *gyrB*, which encodes the subunit B of DNA gyrase, a type II DNA topoisomerase. A comparison between 16S rDNA-based (V3-V4) amplicon sequencing and *gyrB*-based amplicon sequencing was carried out in five types of meat and seafood products, with five mock communities serving as quality controls. Our results revealed that bacterial richness in these mock communities and food samples was estimated with higher accuracy using *gyrB* than using16S rDNA. However, for *Firmicutes* species, 35% of putative *gyrB* reads were actually identified as sequences of a *gyrB* paralog, *parE*, which encodes subunit B of topoisomerase IV; we therefore constructed a reference database of published sequences of both *gyrB* and *pare* for use in all subsequent analyses. Despite this co-amplification, the deviation between relative sequencing quantification and absolute qPCR quantification was comparable to that observed for 16S rDNA for all the tested species. This confirms that *gyrB* can be used successfully alongside 16S rDNA to determine the species composition (richness and evenness) of food microbiota. The major benefit of *gyrB* sequencing is its potential for improving taxonomic assignment and for further investigating OTU richness at the subspecies level, thus allowing more accurate discrimination of samples. Indeed, 80% of the reads of the 16S rDNA dataset were represented by thirteen 16S rDNA-based OTUs that could not be assigned at the species-level. Instead, these same clades corresponded to 44 *gyrB*-based OTUs, which differentiated various lineages down to the subspecies level. The increased ability of *gyrB*-based analyses to track and trace phylogenetically different groups of strains

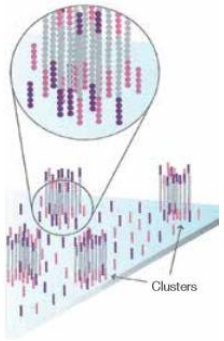
# Steps for Illumina sequencing

- 1st step : one PCR
- 2nd step: one PCR




- 3rd step: on flow cell, the cluster generations
- 4th step: sequencing



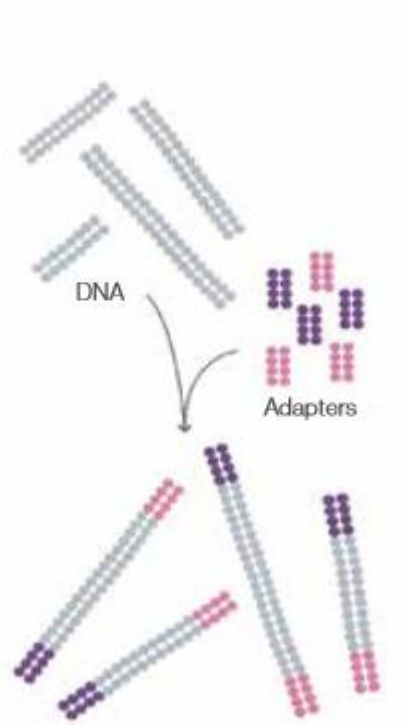
# Amplification and sequencing

« 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)

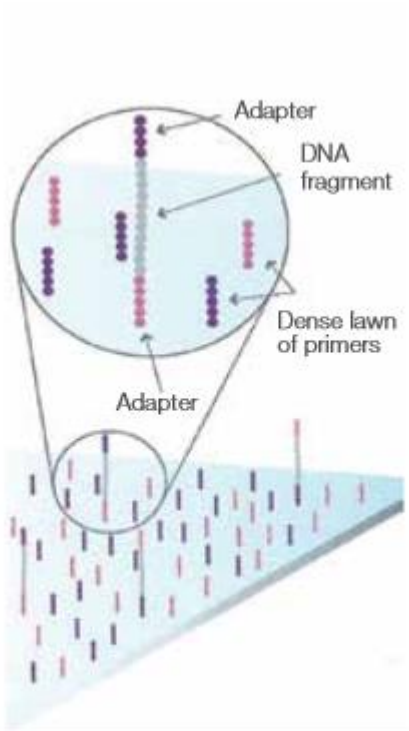
# Cluster generation

## Prepare Genomic DNA Sample




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

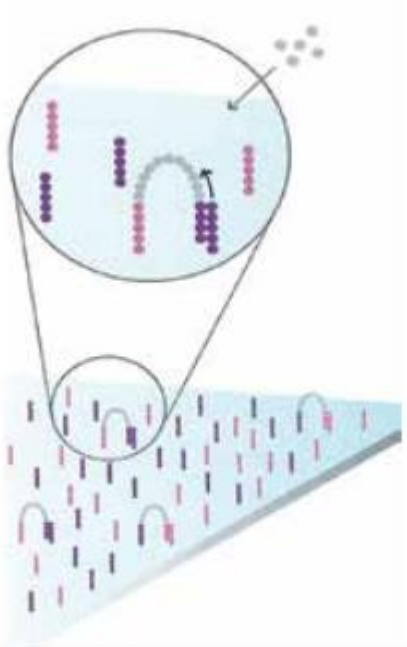
## Attach DNA to Surface




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

Attach DNA to surface

## Bridge Amplification

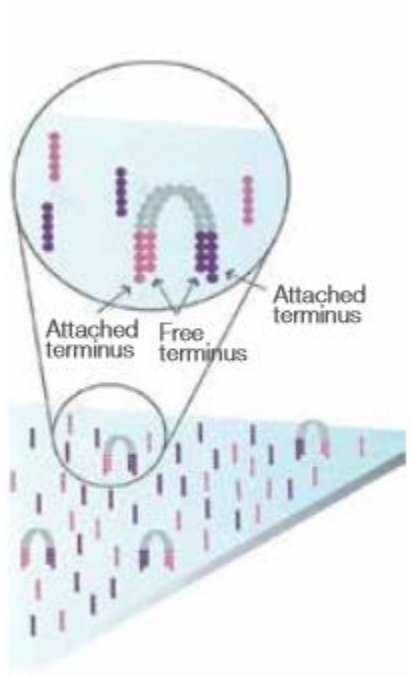


Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

Bridge amplification

# Cluster generation

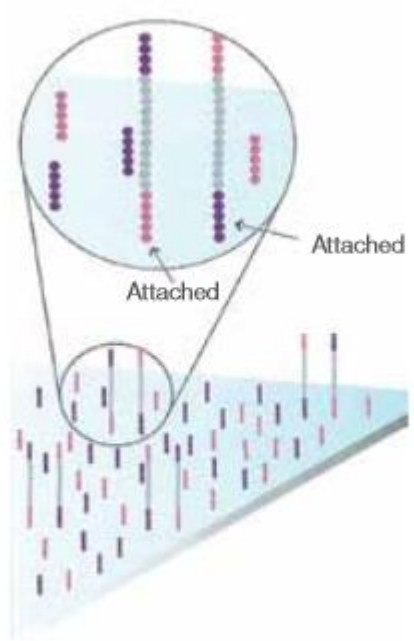
## Fragments Become Double Stranded




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

Fragments become double stranded

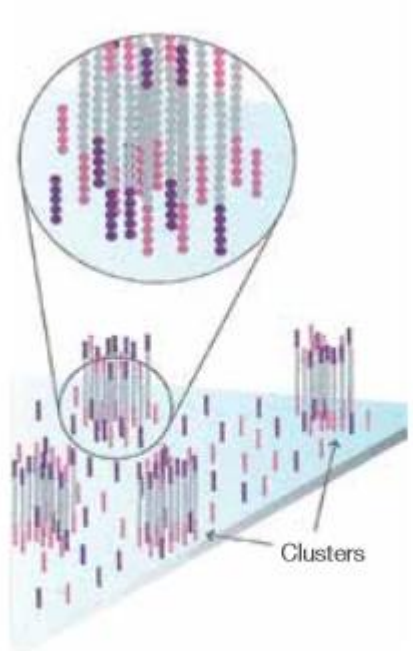
## Denature the Double-Stranded Molecules




Denaturation leaves single-stranded templates anchored to the substrate.

Denature the double-stranded molecule

## Complete Amplification

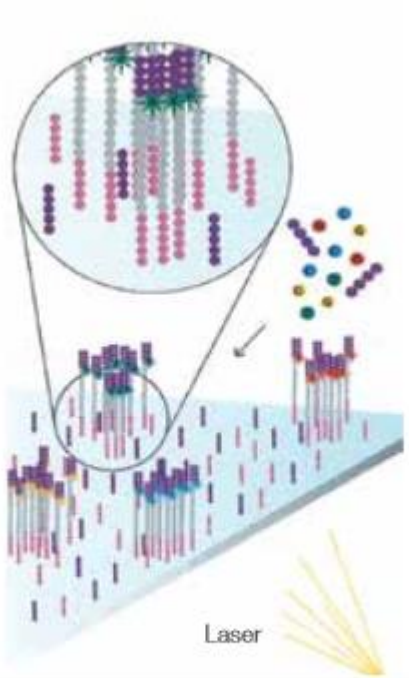



Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

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

# Sequencing by synthesis

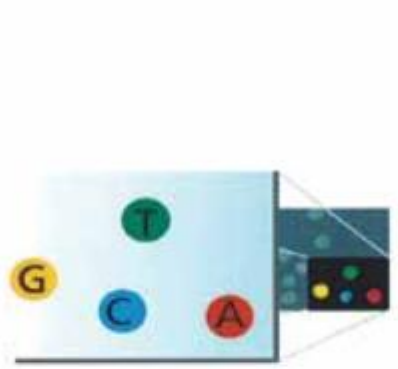
## Determine First Base




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

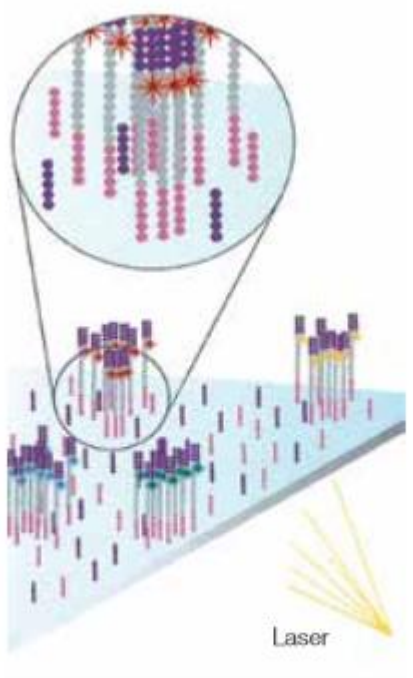
Light signal is more strong in cluster

## Image First Base




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

## Determine Second Base



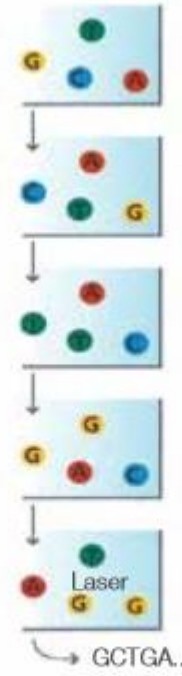

The next cycle repeats the incorporation of four labeled reversible terminators, primers, and DNA polymerase.

# Sequencing by synthesis

## Image Second Chemistry Cycle

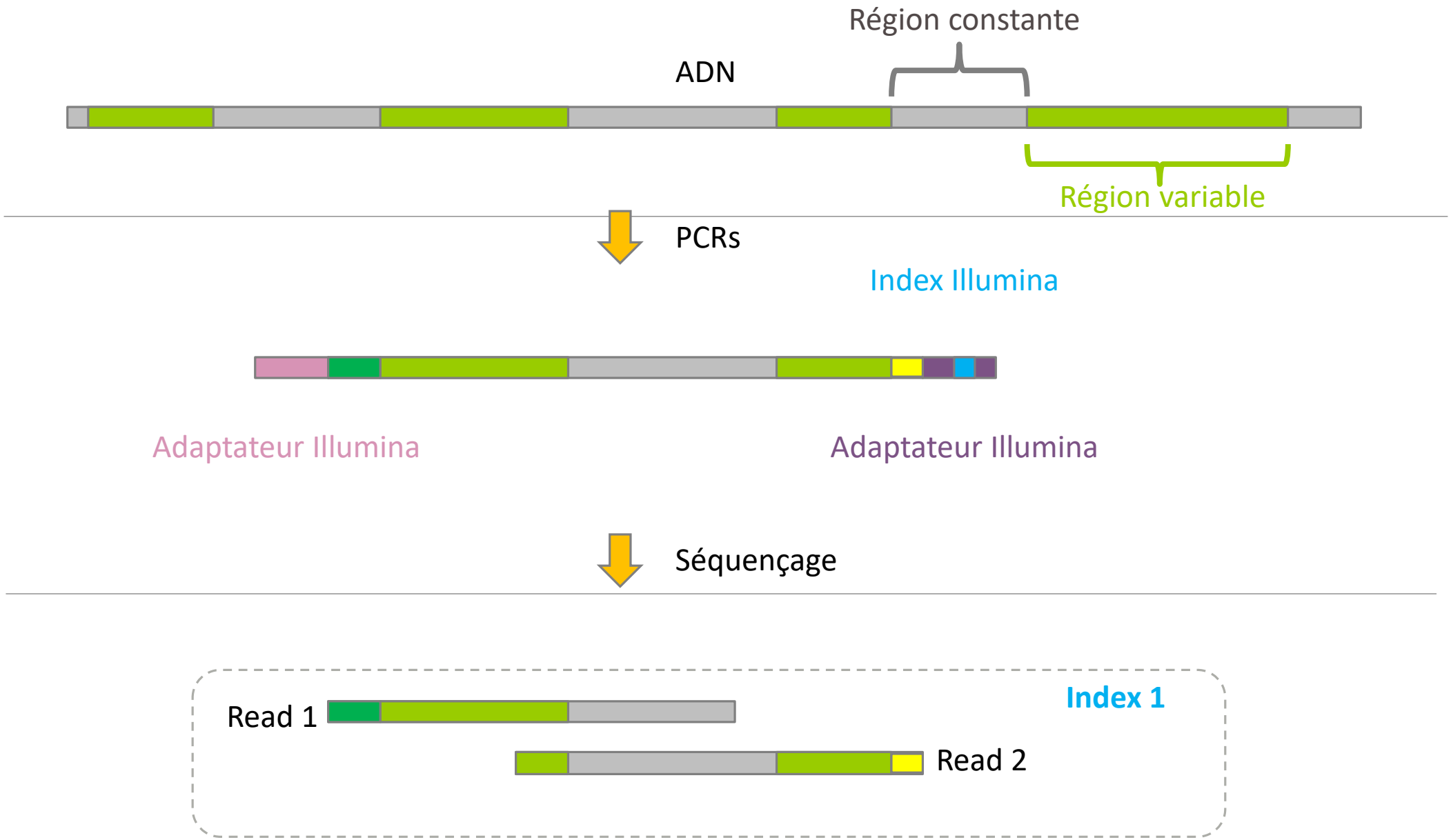

After laser excitation, the image is captured as before, and the identity of the second base is recorded.

## Sequencing Over Multiple Chemistry Cycles

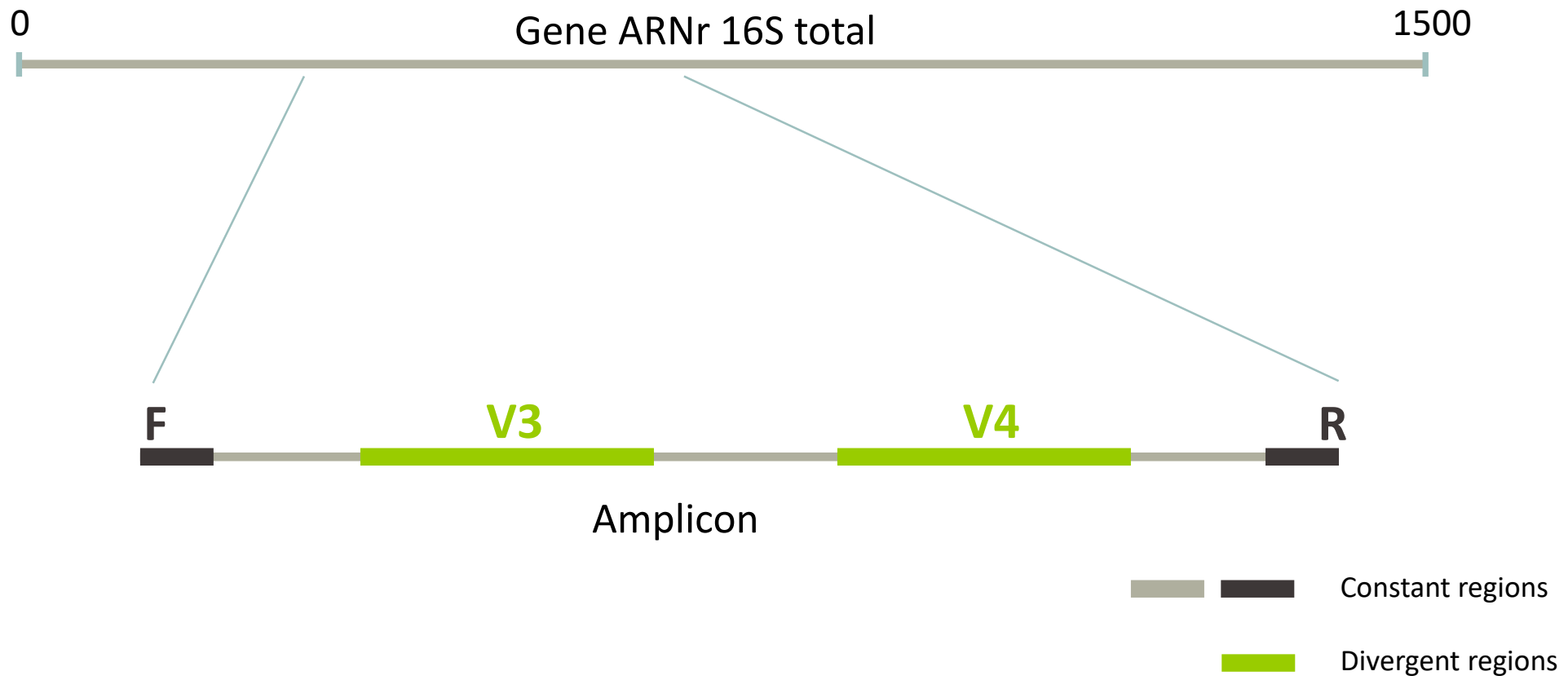



The sequencing cycles are repeated to determine the sequence of bases in a fragment, one base at a time.

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



# Amplification and sequencing

Sequencing is generally perform on **Roche-454** (obsolete now) or **Illumina MiSeq** platforms.

Roche-454 generally produce ~ 10 000 reads per sample

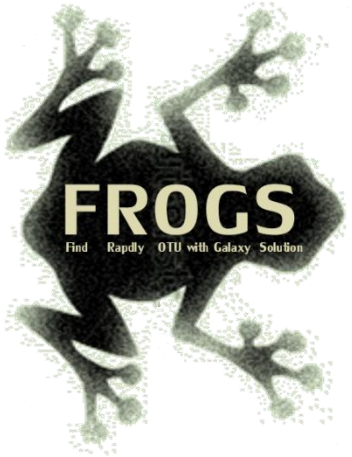
MiSeq ~ 30 000 reads per sample

Sequence length is **>650 bp** for pyrosequencing technology (Roche-454) and **2 x 250 bp or 2 x 300 bp** for the MiSeq technology in paired-end mode.


# Methods


# Which bioinformatics solutions ?

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

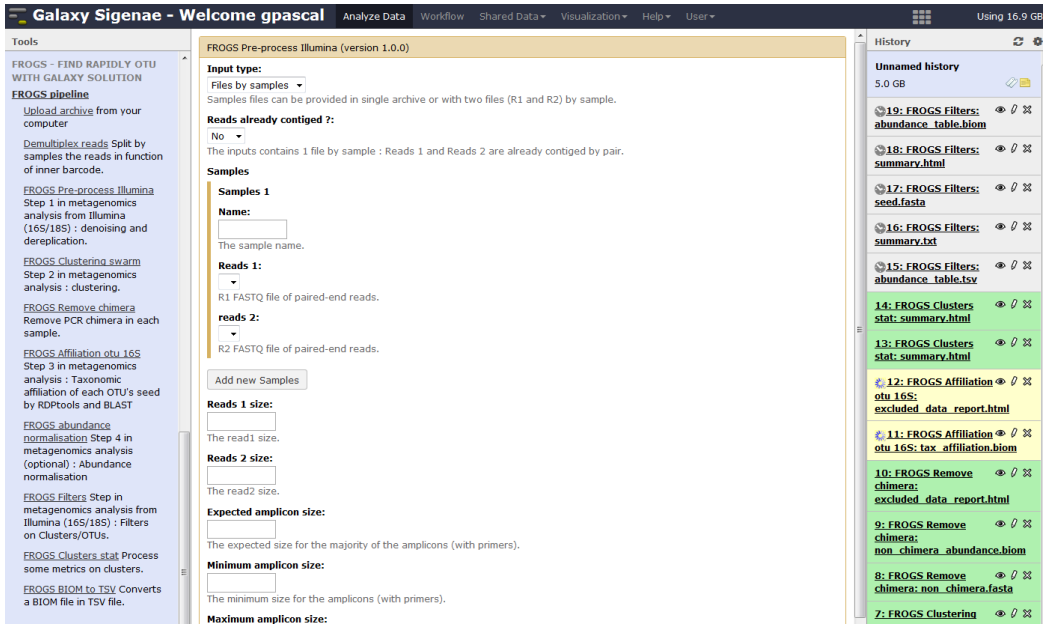
Independent modules

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

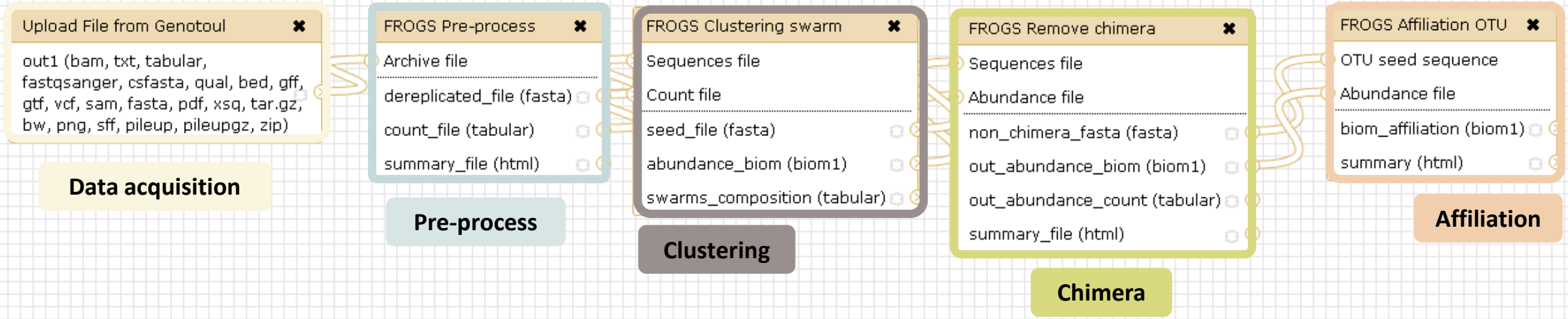
Innovative clustering method

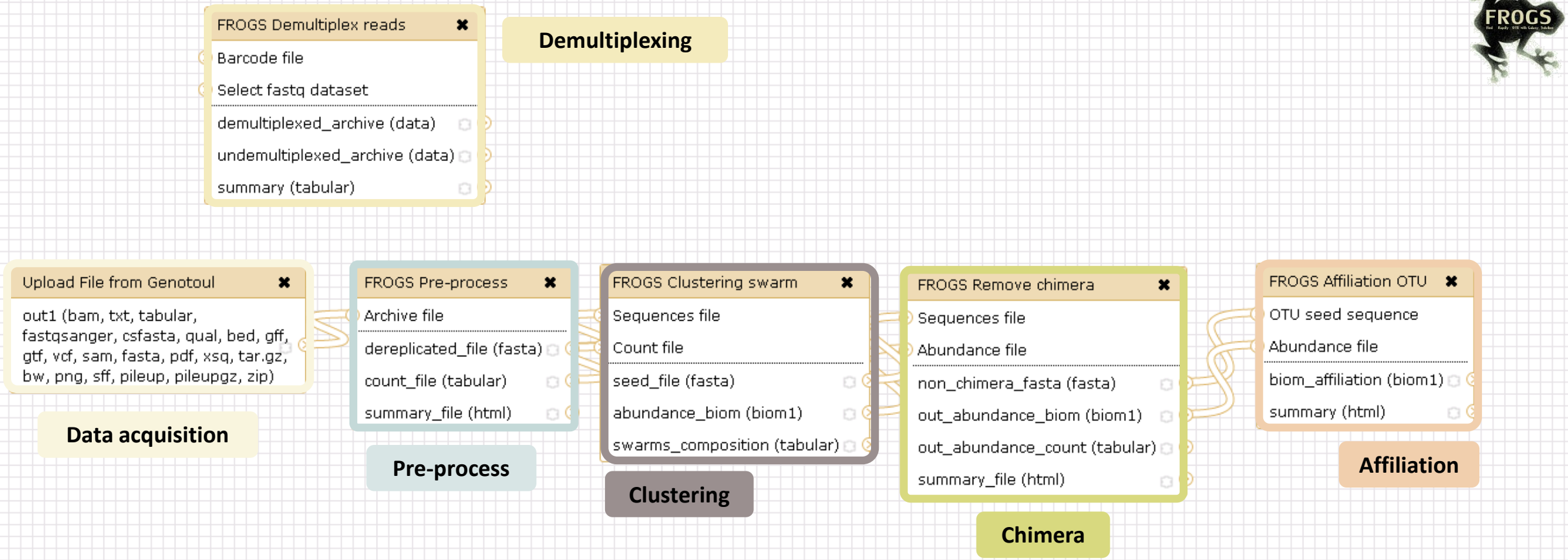
Many graphics for interpretation

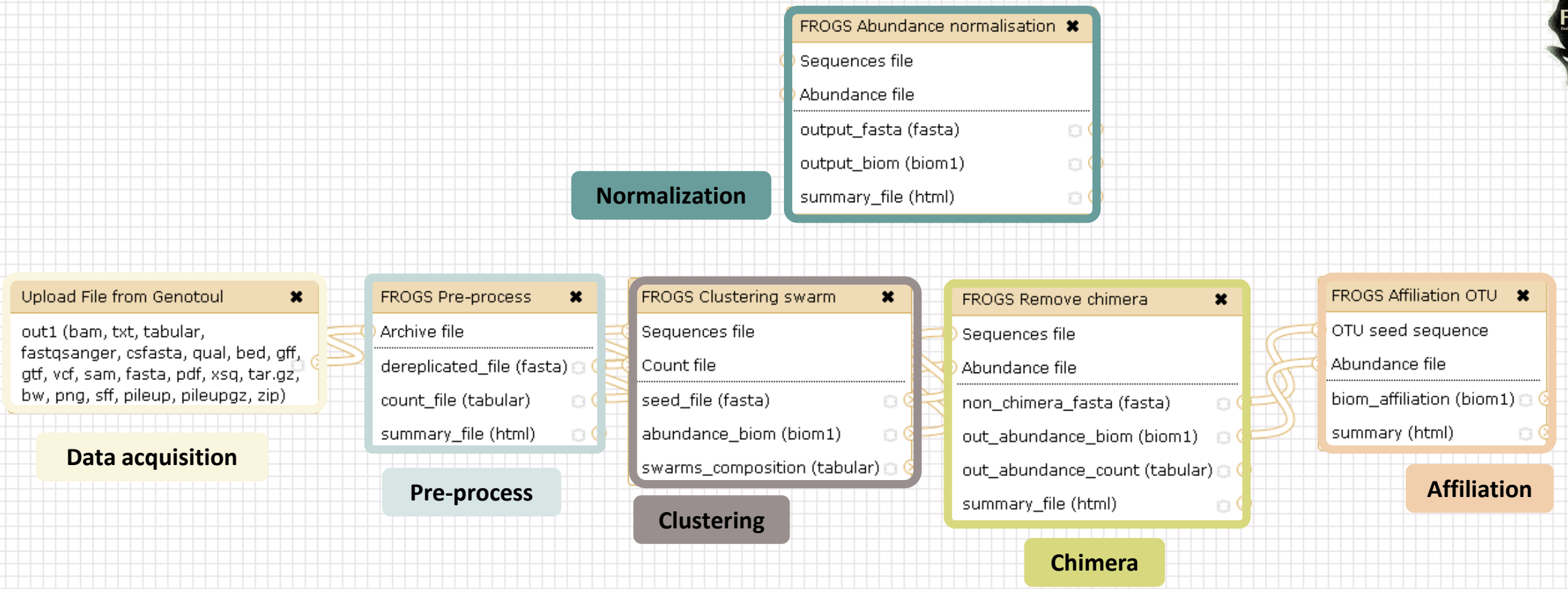
User friendly, hiding bioinformatics infrastructure/complexity

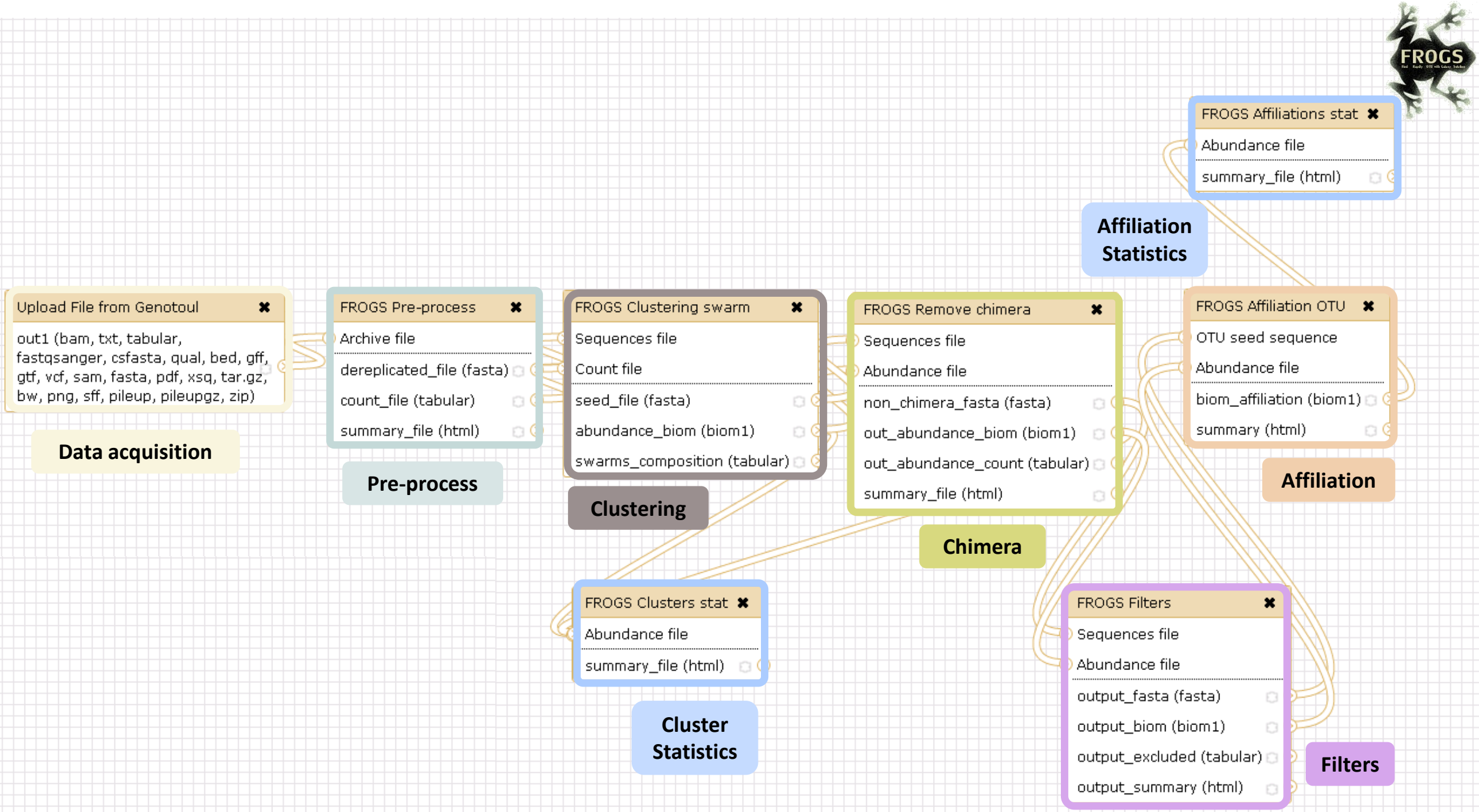


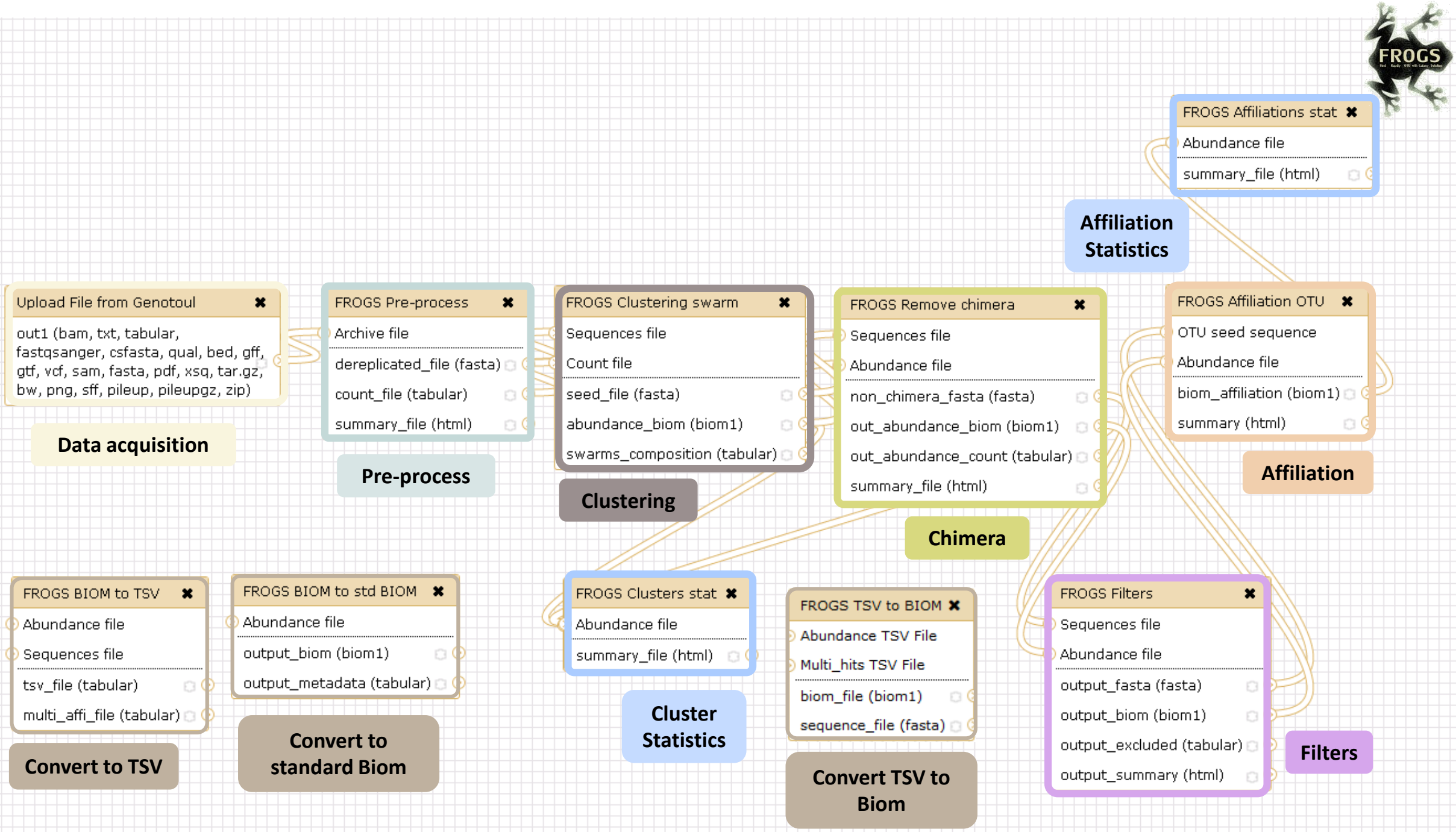
# FROGS Pipeline

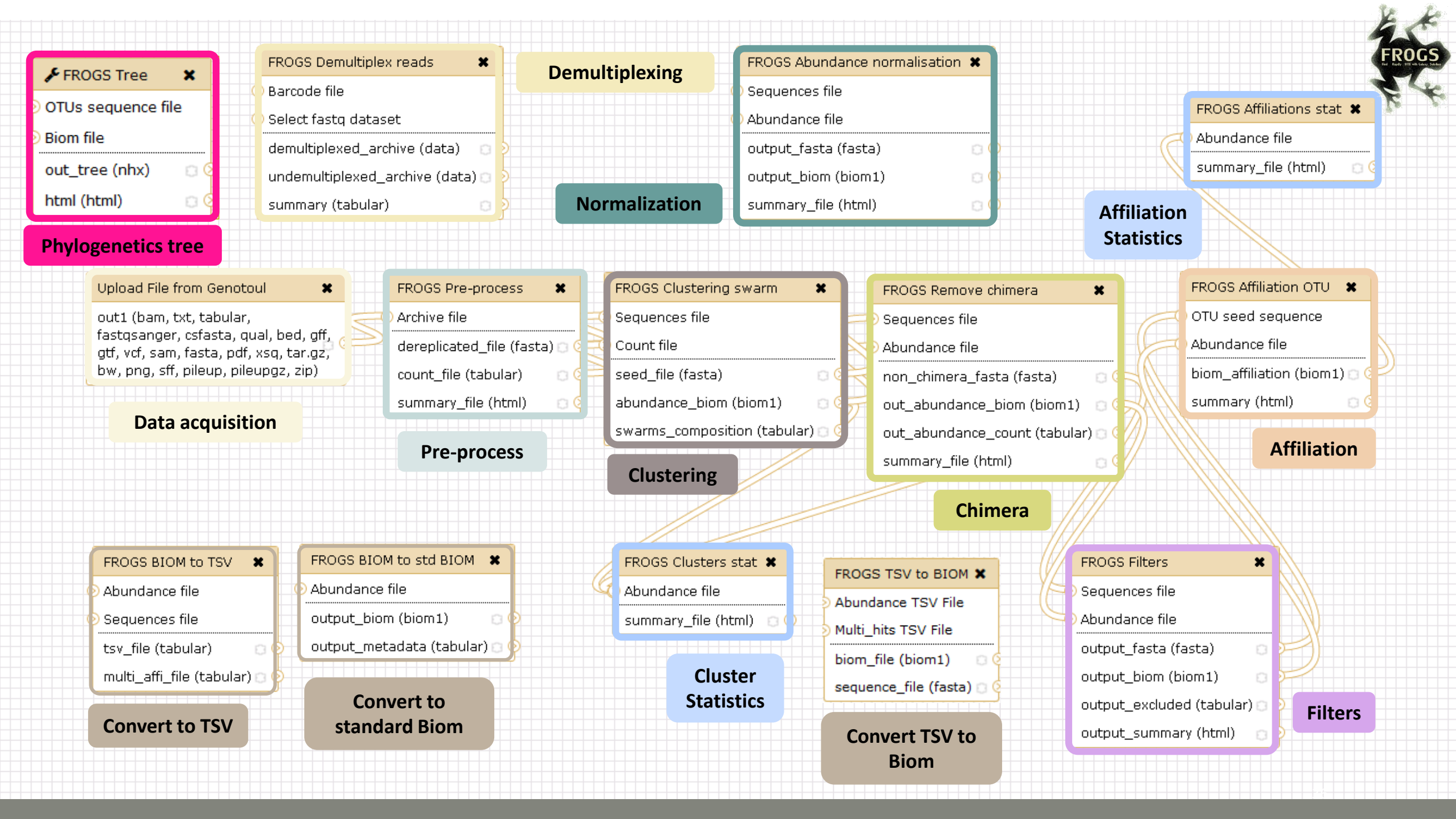



FROGS Demultiplex reads
Barcode file
Select fastq dataset
demultiplexed_archive (data)
undemultiplexed_archive (data)
summary (tabular)
Demultiplexing
Upload File from Genotoul
out1 (bam, txt, tabular, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png, sff, pileup, 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 Abundance normalisation
Sequences file
Abundance file
output_fasta (fasta)
output_biom (biom1)
summary_file (html)
Normalization
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

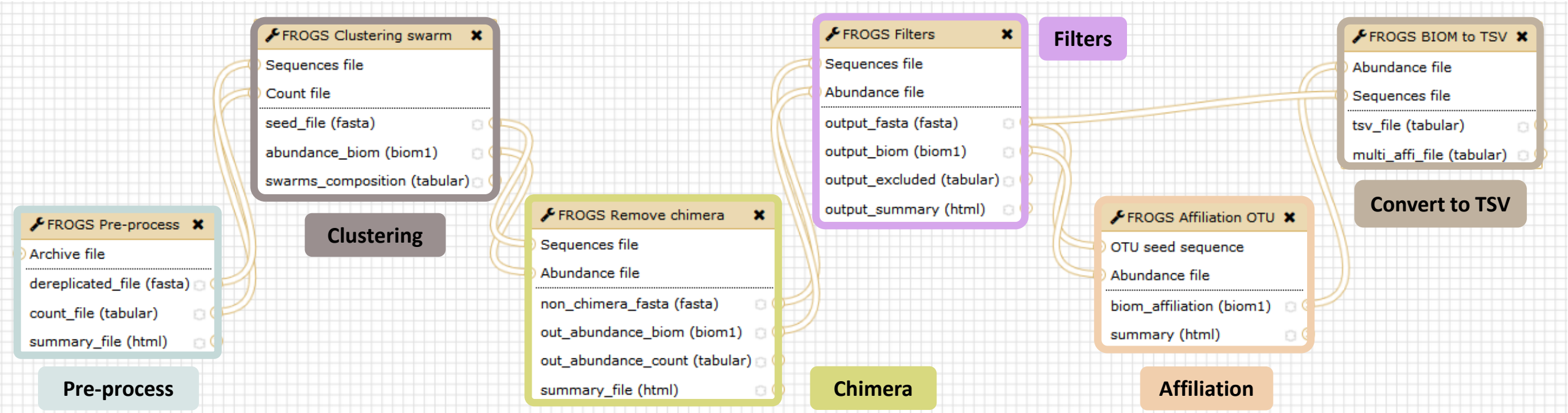

Upload File from Genotoul
out1 (bam, txt, tabular, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png, sff, pileup, pileupgz, zip)
Data acquisition
FROGS Pre-process
Archive file
dereplicated_file (fasta)
count_file (tabular)
summary_file (html)
Pre-process
FROGS Clustering 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 Affiliations stat
Abundance file
summary_file (html)
Affiliation Statistics
FROGS Affiliation OTU
OTU seed sequence
Abundance file
biom_affiliation (biom1)
summary (html)
Affiliation
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


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 Affiliations stat
Abundance file
summary_file (html)
Affiliation Statistics
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


FROGS Tree
OTUs sequence file
Biom file
out_tree (nhx)
html (html)
Phylogenetics tree
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

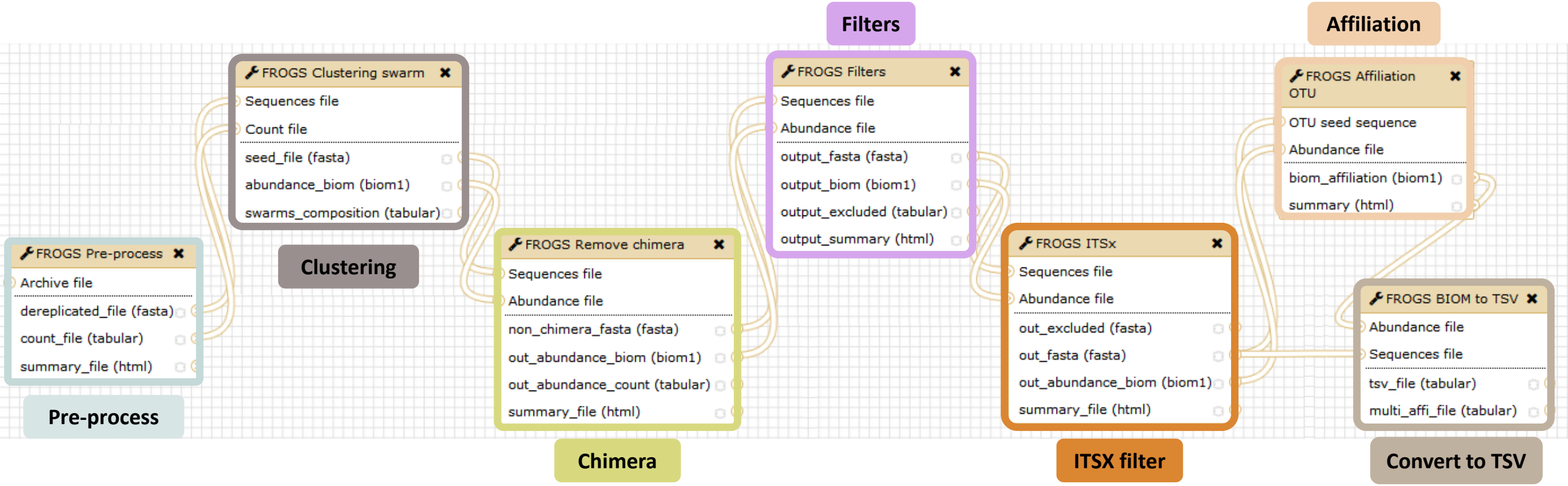
# FROGS Pipeline

Minimal pipeline for bacterial amplicon analyses



# FROGS Pipeline

Minimal pipeline for ITS amplicon analyses



# FROGS Tools for Bioinfomatics analyses

Demultiplexing

Pre-process

Clustering

Chimera

Filters

ITSX

Affiliation

Cluster Stat

Affiliation Stat

Affiliation postprocess

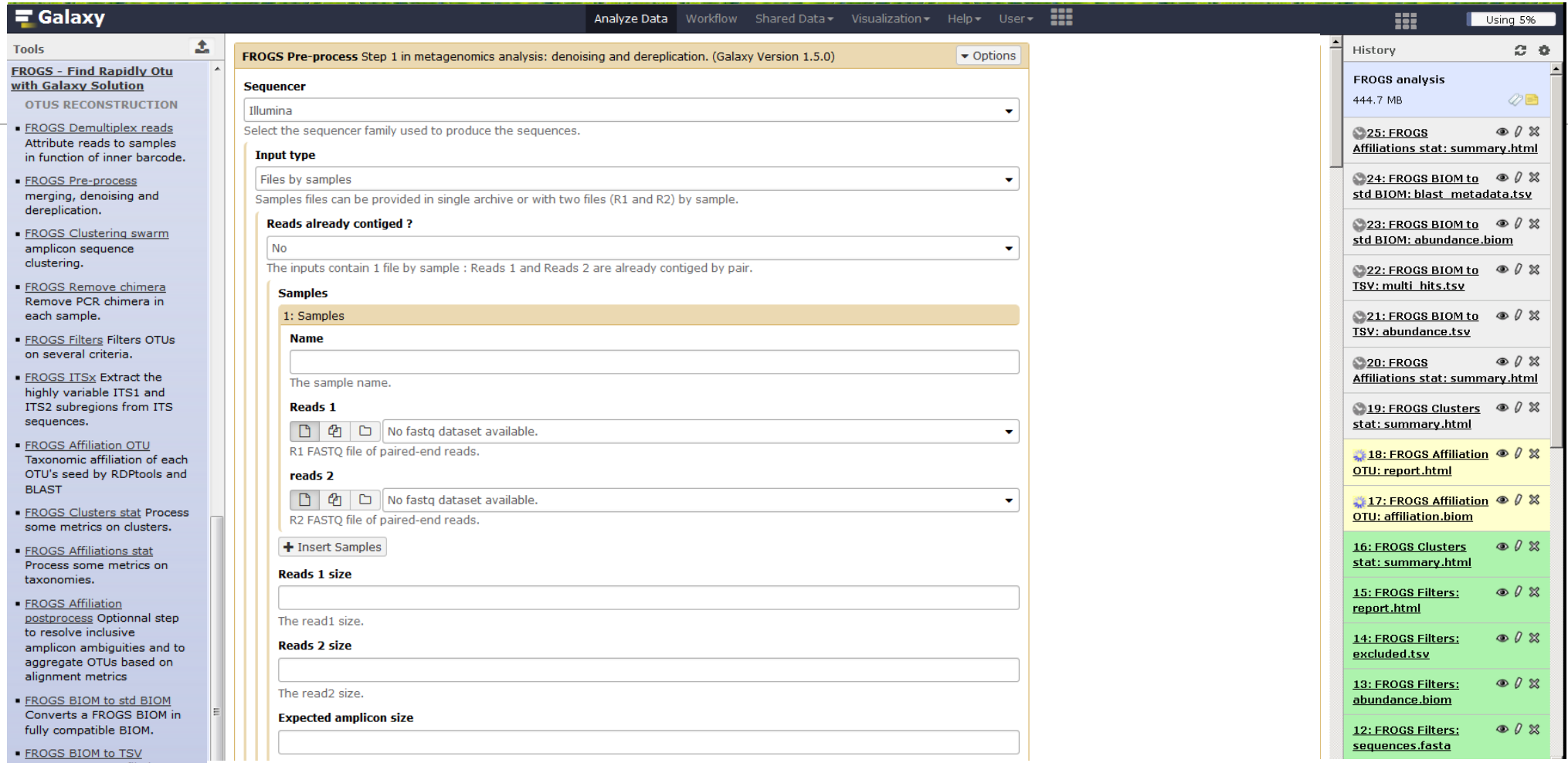
Biom to std Biom

Biom to TSV

TSV to Biom

Normalization

Phylogenetics Tree




Waiting to run

Currently running

Result files

# FROGS Tools for Statistic analyses

Import data

Composition visualisation

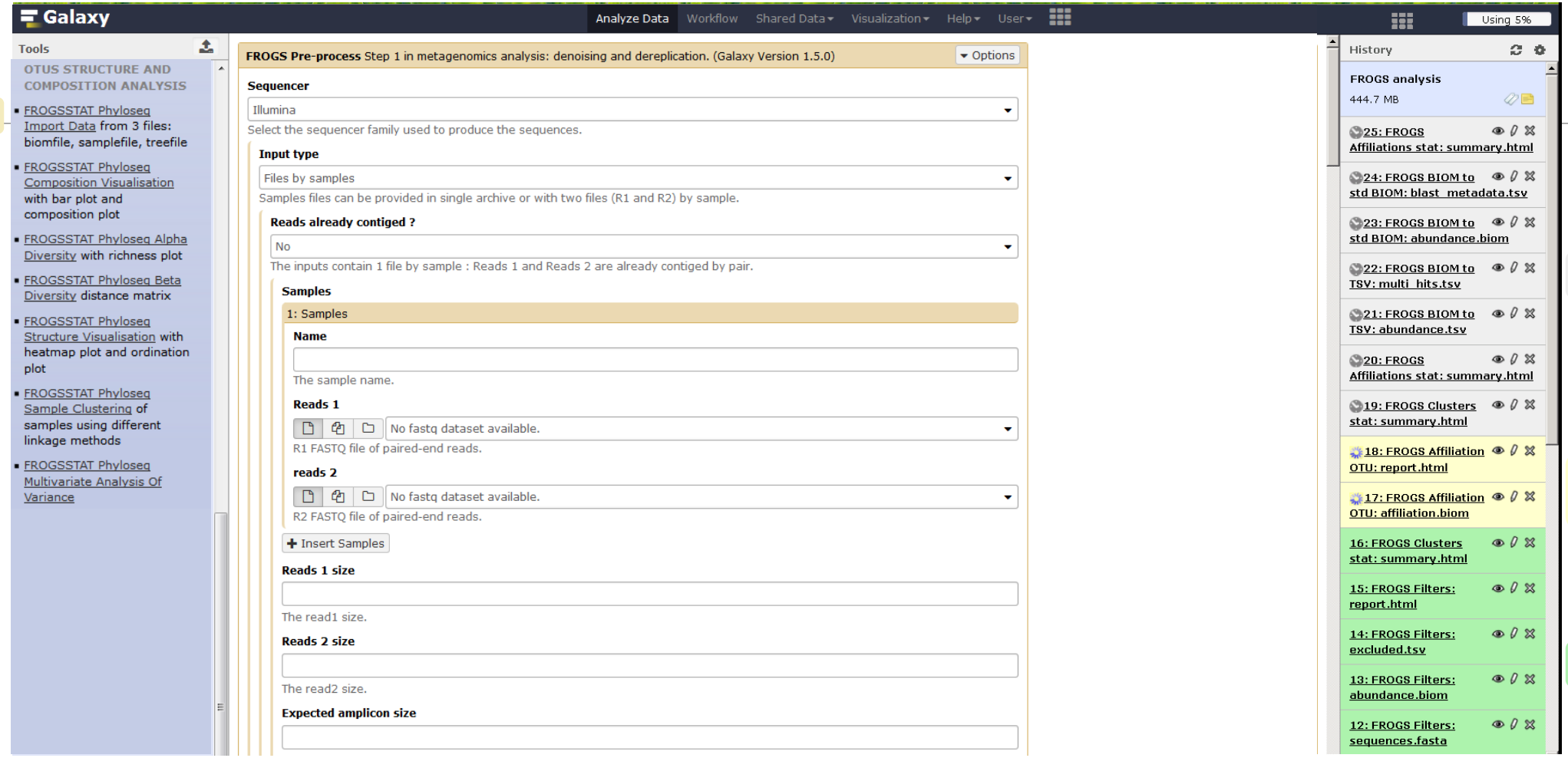
Alpha diversity

Beta diversity

Structure visualisation

Sample clustering

Multivariate analysis of variance




Waiting to run

Currently running

Result files

# What kind of data ?

## 4 Upload → 4 Histories

Multiplexed data

Pathobiomes rodents and ticks

multiplex.fastq

barcode_forward.tabular

ITS data

METABARFOOD project

ITS.tar.gz

**OU**

454 data

Freshwater sediment metagenome

454.fastq.gz

SRA number: SRR443364

MiSeq
R1 fastq + R2 fastq

Farm animal feces metagenome

sampleA_R1.fastq

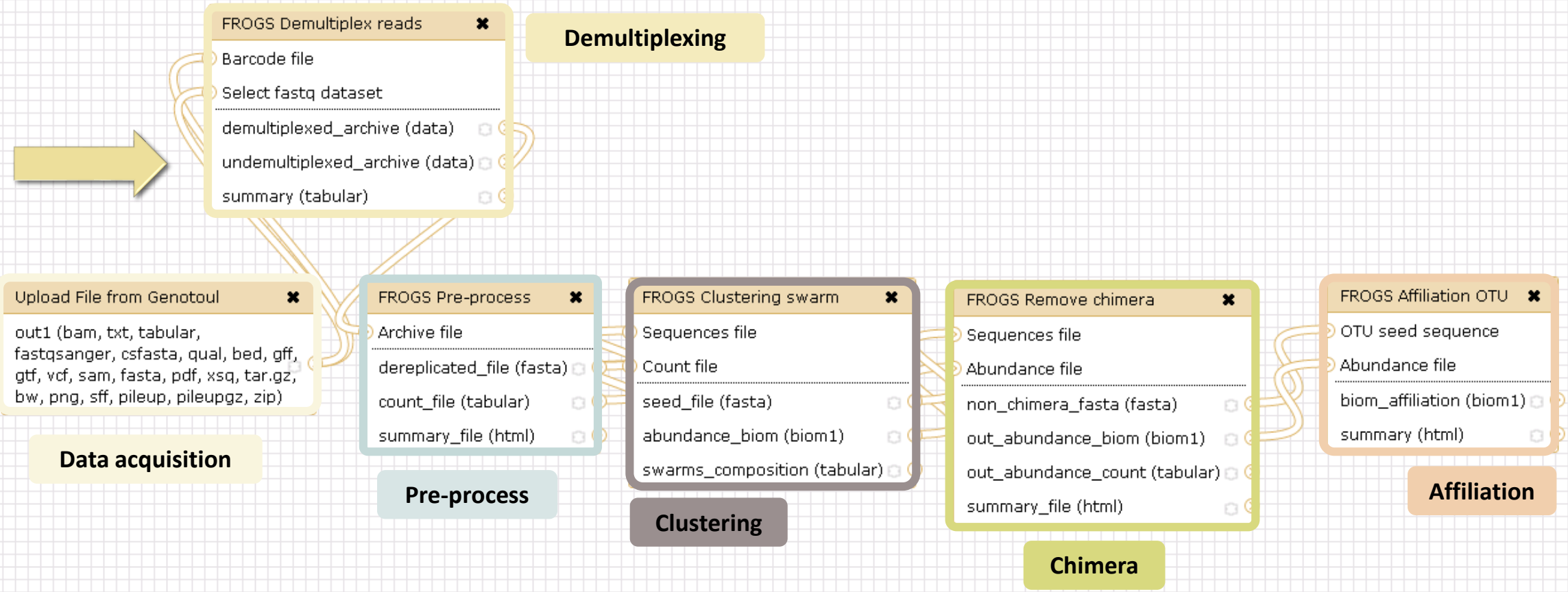
sampleA_R2.fastq

MiSeq merged fastq in archive tar.gz

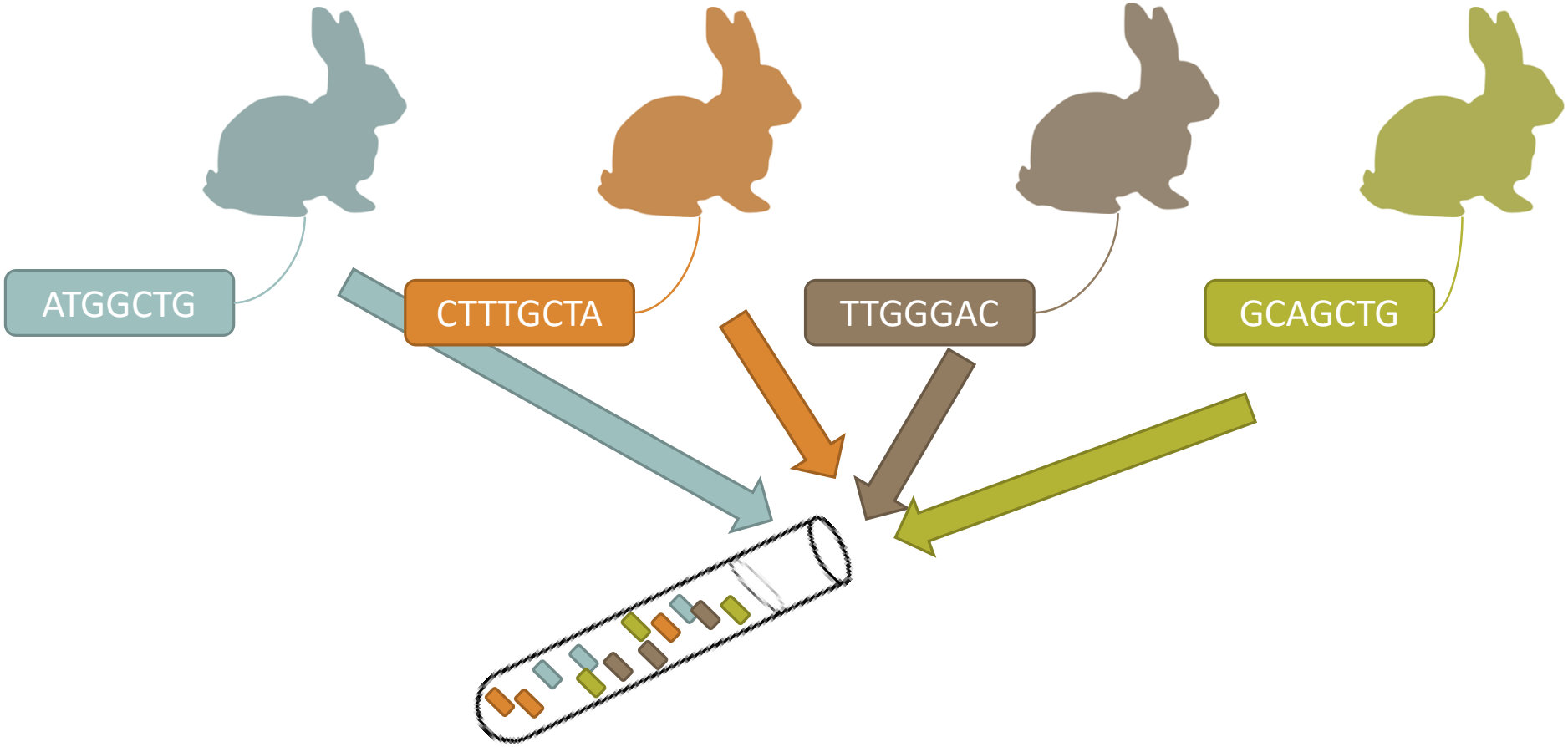
Farm animal feces metagenome

100spec_90000seq_9samples.tar.gz

# 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

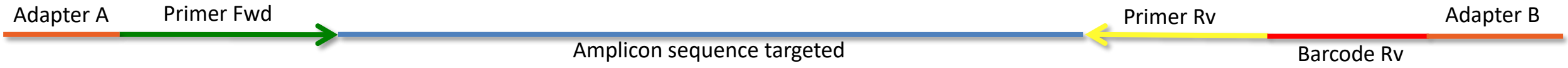
Remove unbarcoded or ambiguous sequences

# Demultiplexing forward


Single-end sequencing

Paired-end sequencing

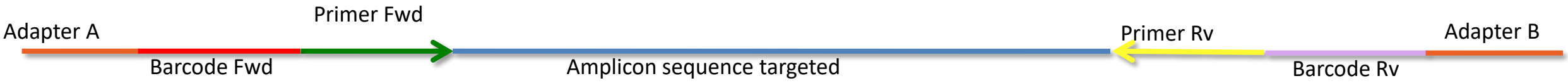
# Demultiplexing reverse




Single end sequencing

Paire end sequencing

# Demultiplexing forward and reverse




Single end sequencing

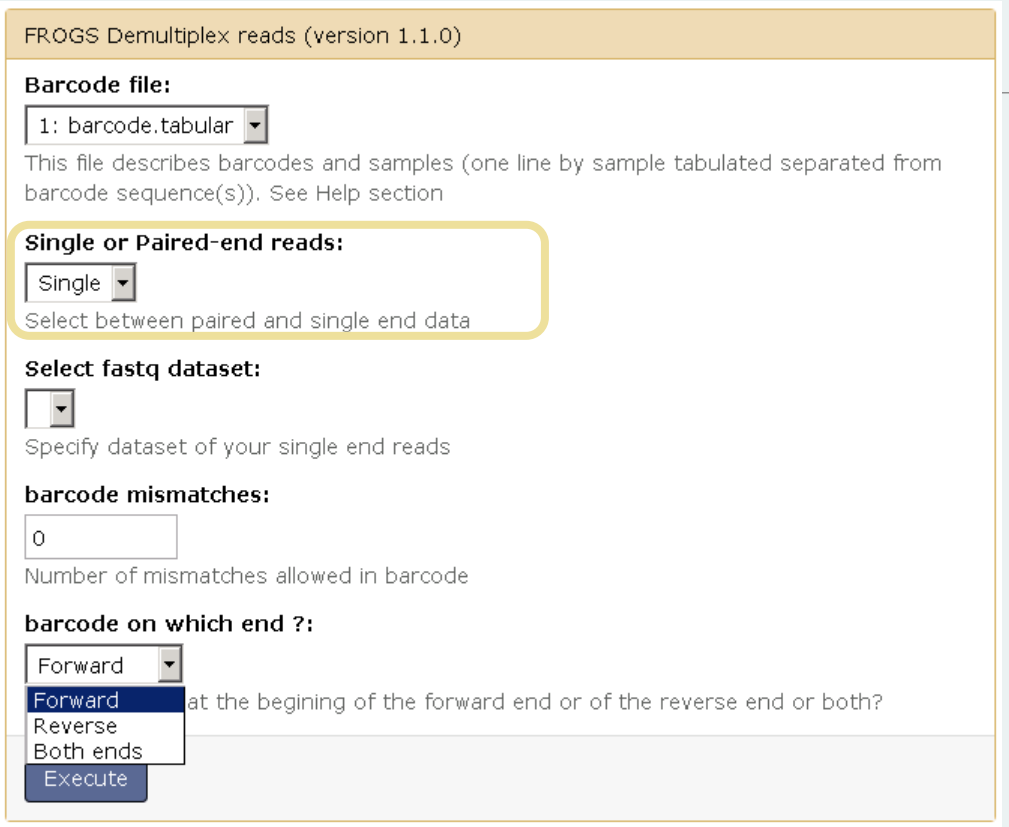
Paire end sequencing

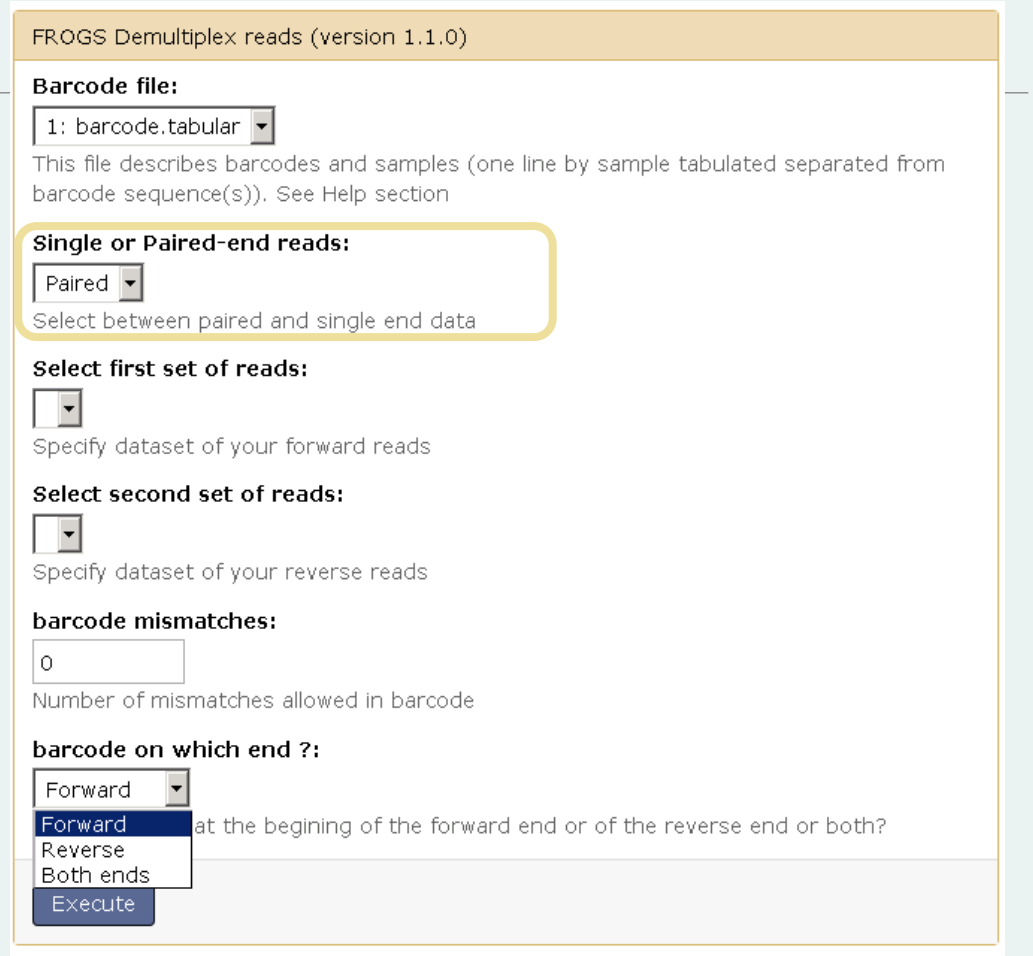
# Your turn! - 1

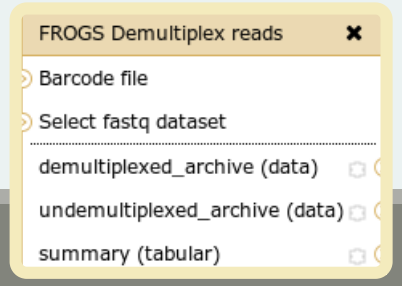
LAUNCH DEMULTIPLEX READS TOOL

Multiplex

The tool parameters depend on the input data type




**Demultiplexing**

# Exercise 1

In **multiplexed** history launch the demultiplex tool:

« The Patho-ID project, rodent and tick‘s pathobioms study, financed by the metaprogram INRA-MEM, studies zoonoses on rats and ticks from multiple places in the world, the co-infection systems and the interactions between pathogens. In this aim, thay have extracted hundreads of or rats and ticks samples from which they have extracted 16S DNA and sequenced them first time on Roche 454 plateform and in a second time on Illumina Miseq plateform. For this courses, they authorized us to publicly shared some parts of these samples. »

Parasites & Vectors (2015) 8:172 DOI 10.1186/s13071-015-0784-7. **Detection of Orientia sp. DNA in rodents from Asia, West Africa and Europe.** Jean François Cosson, Maxime Galan, Emilie Bard, Maria Razzauti, Maria Bernard, Serge Morand, Carine Brouat, Ambroise Dalecky, Khalilou Bâ, Nathalie Charbonnel and Muriel Vayssier-Taussat

# Exercise 1

In **multiplexed** history launch the demultiplex tool:

Data are single end reads
→ only 1 fastq file

Samples are characterized by one barcode in forward strands
→ multiplexing « forward »

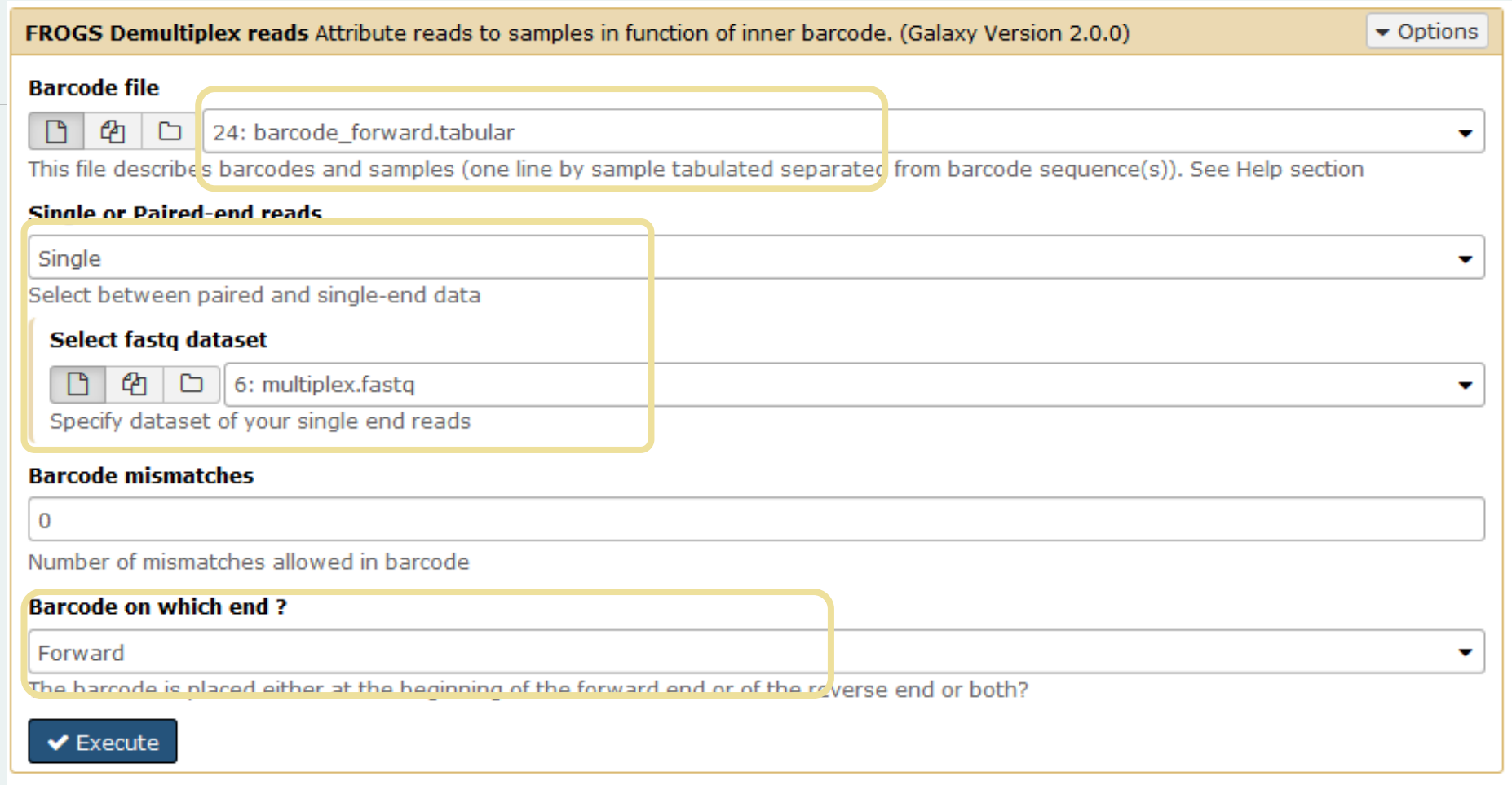
Inputs :

# Exercise 1

Demultiplex tool asks for 2 files: one « fastq » and one « tabular »

1. Play with pictograms 
2. Observe how is built a fastq file. 
3. Look at the stdout, stderr when available (in the  pictogram )

# Multiplex


FROGS Demultiplex reads Attribute reads to samples in function of inner barcode. (Galaxy Version 2.0.0)
Options
Barcode file
24: barcode_forward.tabular
This file describes barcodes and samples (one line by sample tabulated separated from barcode sequence(s)). See Help section
Single or Paired-end reads
Single
Select between paired and single-end data
Select fastq dataset
6: multiplex.fastq
Specify dataset of your single end reads
Barcode mismatches
0
Number of mismatches allowed in barcode
Barcode on which end ?
Forward
The barcode is placed either at the beginning of the forward end or of the reverse end or both?
Execute

# Advices

For your own data

- Do not forget to indicate barcode sequence as they are in the fastq sequence file, especially if you have data multiplexed via the reverse strand.
- For the mismatch threshold, we advised you to let the threshold to 0, and if you are not satisfied by the result, try with 1. The number of mismatch depends on the length of the barcode, but often those sequences are very short so 1 mismatch is already more than the sequencing error rate.
- If you have different barcode lengths, you must demultiplex your data in different times beginning by the longest barcode set and used the "unmatched" or "ambiguous" sequence with smaller barcode and so on.
- If you have Roche 454 sequences in sff format, you must convert them with some program like sff2fastq

# Results


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

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

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

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

# 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
ACAGCGTCAGAGGGGTACCAGTCAGCCATGACGTAGCACGTACA
+
CCCFFFFFFHHHHHJJIJJJHHFF@DEDDDDDDD@CDDDDACDD
```

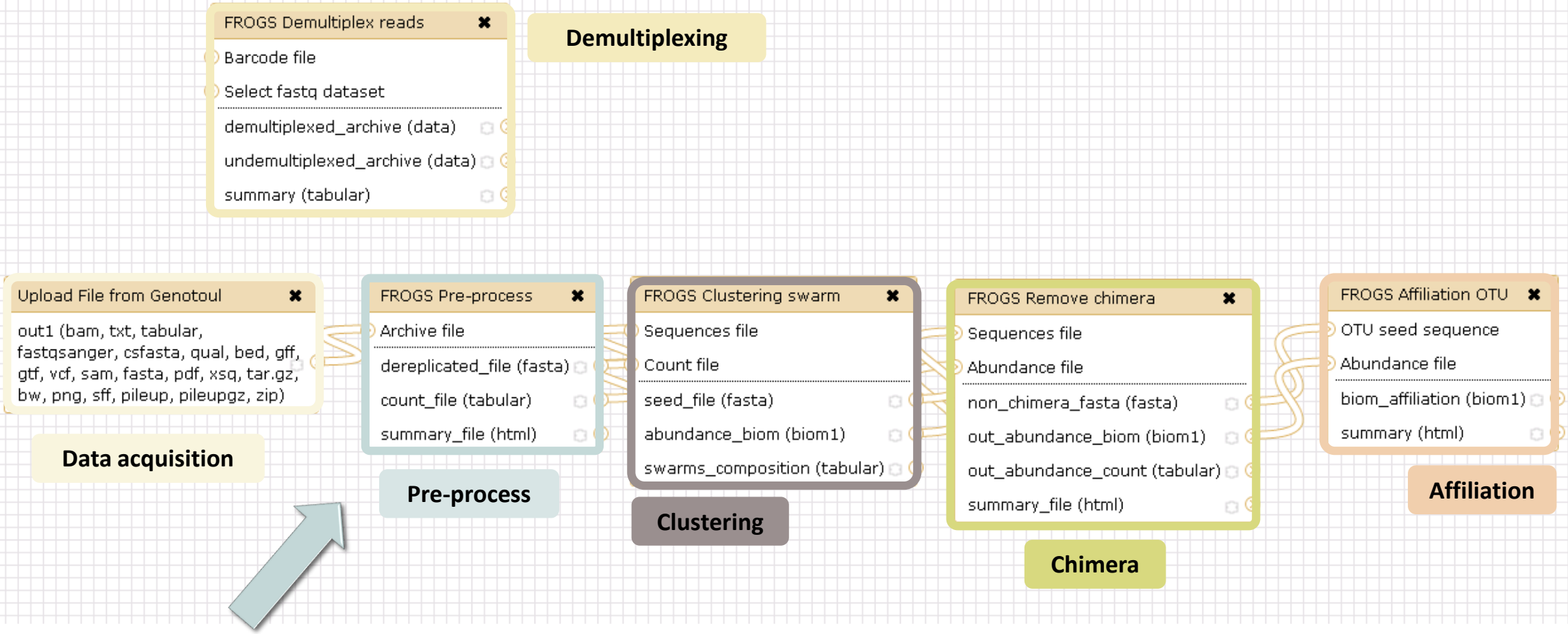
# How it works ?

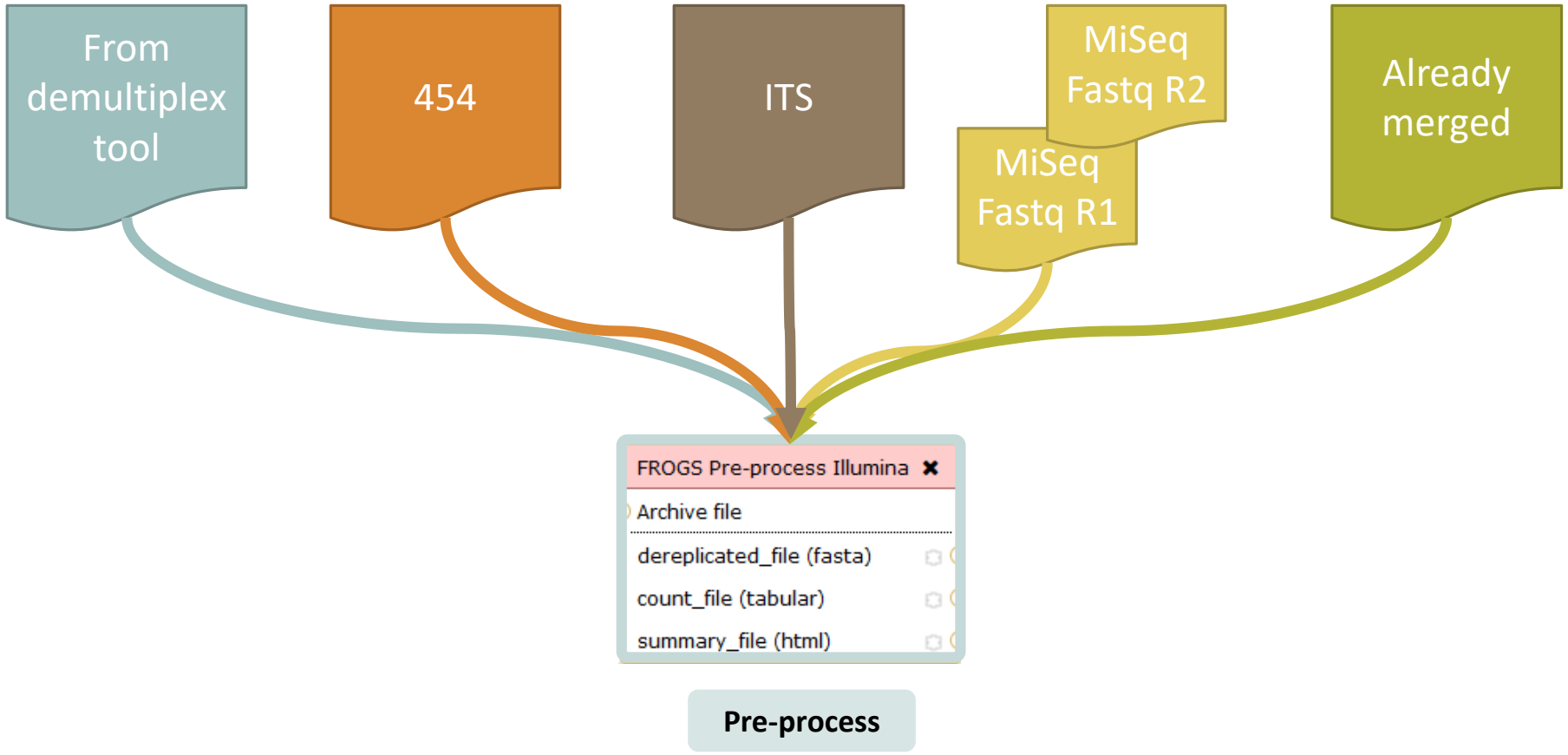
For each sequence or sequence pair the sequence fragment at the beginning (forward multiplexing) of the (first) read or at the end (reverse multiplexing) of the (second) read will be compare to all barcode sequence.

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.

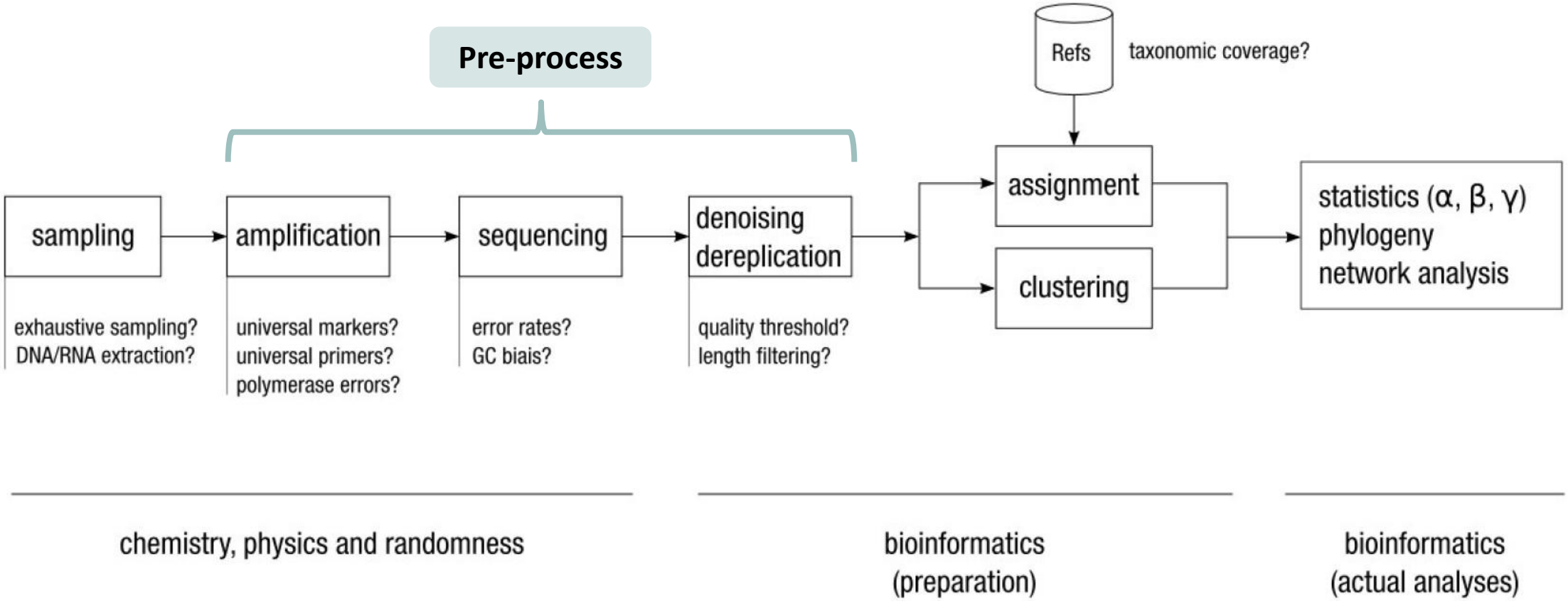
Finally fastq files (or pair of fastq files) for each sample are included in an archive, and a summary describes how many sequence are attributed for each sample.

# Pre-process tool


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


From demultiplex tool
454
ITS
MiSeq Fastq R2
MiSeq Fastq R1
Already merged
FROGS Pre-process Illumina
Archive file
dereplicated_file (fasta)
count_file (tabular)
summary_file (html)
Pre-process

# Amplicon-based studies general pipeline



# Pre-process

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

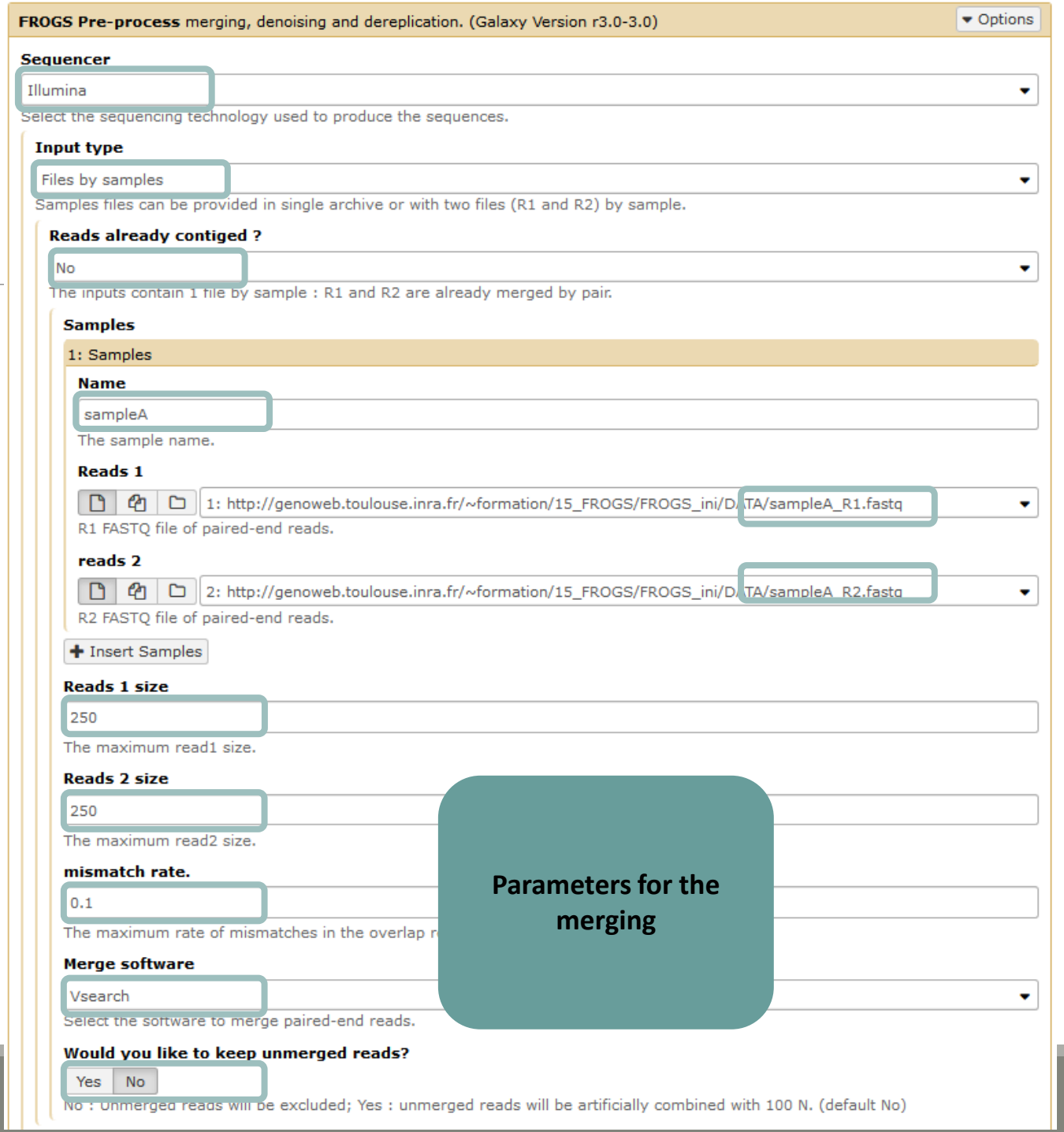
- + removing homopolymers (size = 8 ) for 454 data
- + quality filter for 454 data

**VSEARCH: a versatile open source tool for metagenomics.**
Rognes T, Flouri T, Nichols B, Quince C, Mahé F.
PeerJ. 2016 Oct 18;4:e2584. eCollection 2016.

## Example for:

- Illumina MiSeq data
- 1 sample
- Non joined

**Pre-process example 1**



**Minimum amplicon size**

340

The minimum size for the amplicons.

**Maximum amplicon size**

450

The maximum size for the amplicons.

**[V5] 16S variability**

**Sequencing protocol**

Illumina standard

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

**5' primer**

CCGTCAATTC

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

**Primer sequences**

**3' primer**

CCGCNGCTGCT

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

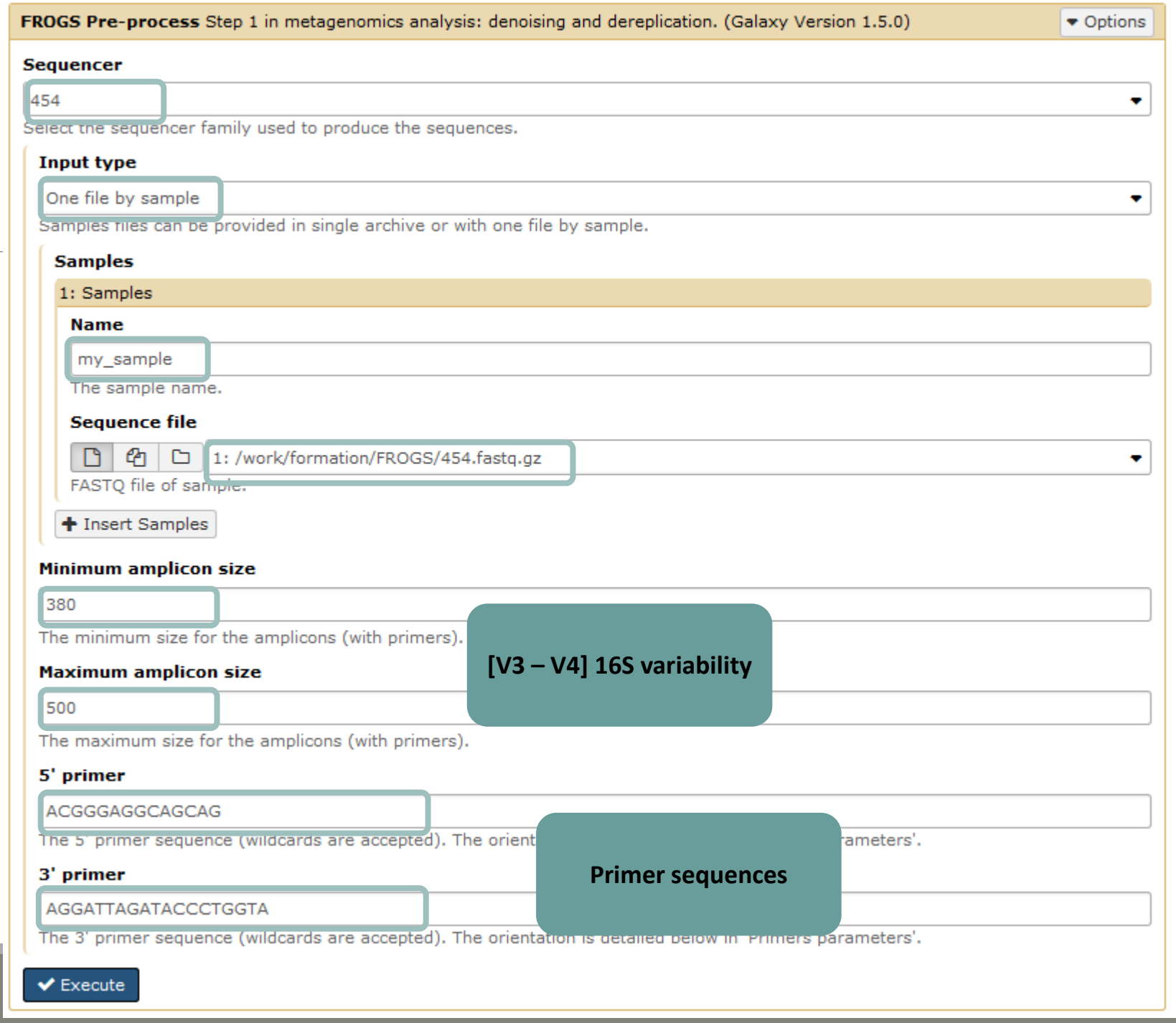
Execute

**Pre-process example 1**

Example for:

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

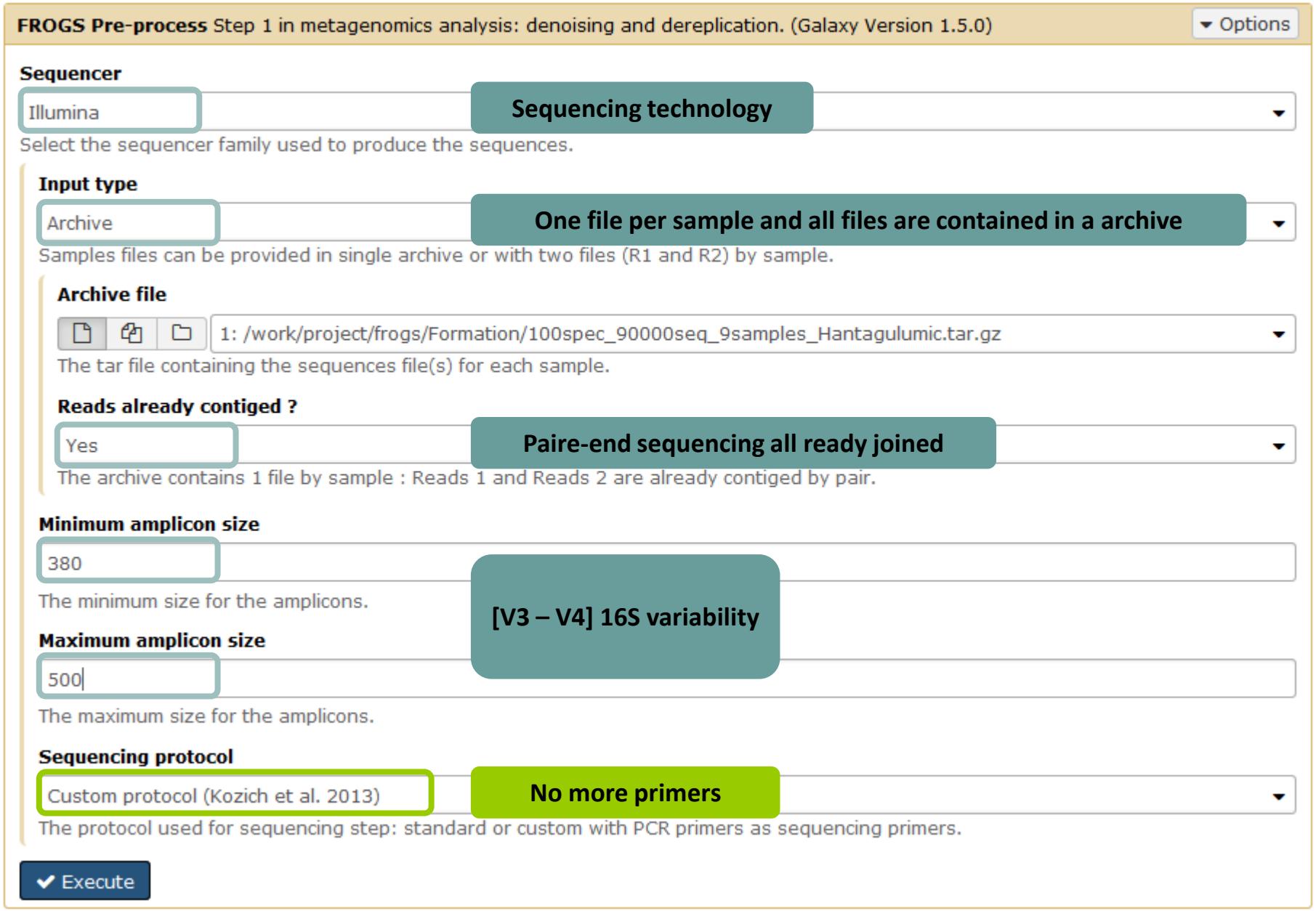
**Pre-process example 2**



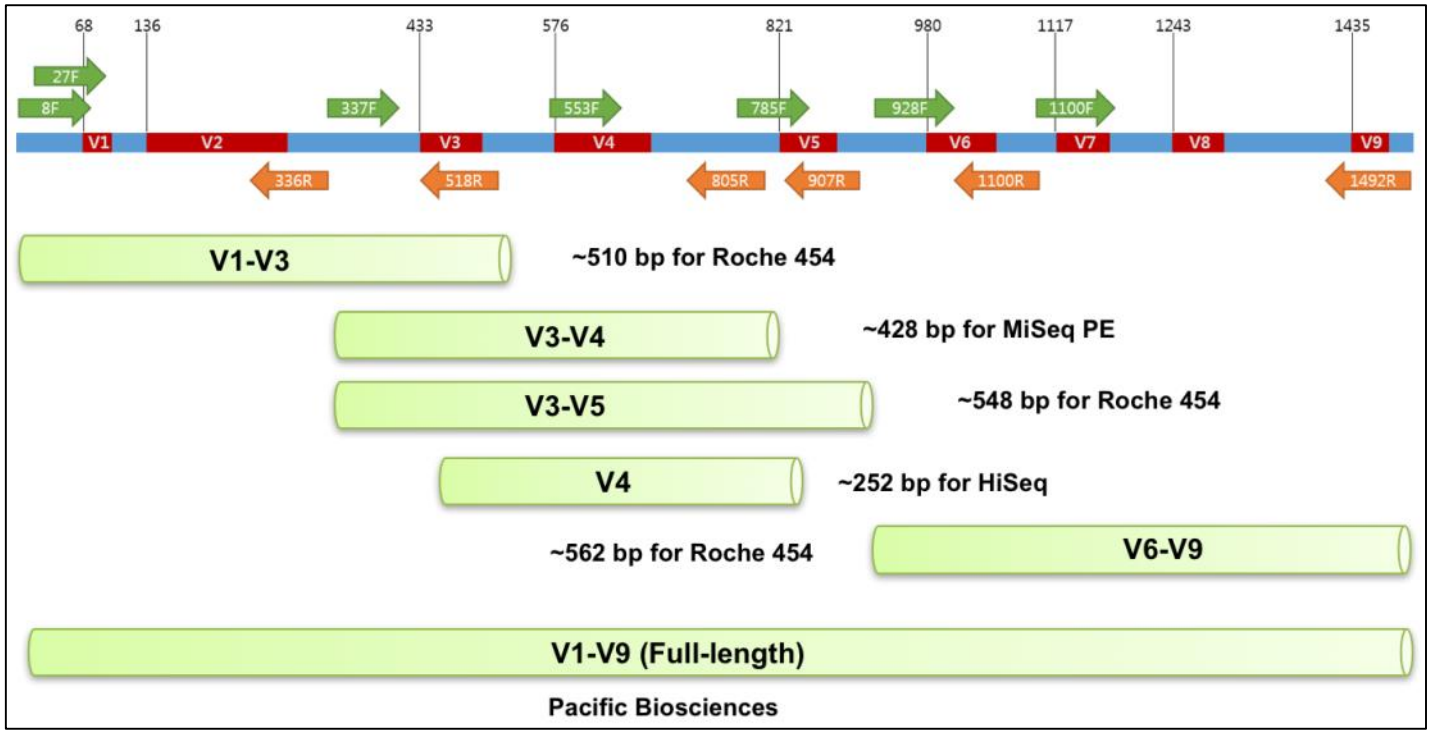
Example for:

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

**Pre-process example 3**



# Which primers for 16S ?




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

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

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

# Your turn! - 2

454

Escape 454 data

# Exercise 2.1

Go to « 454 » history

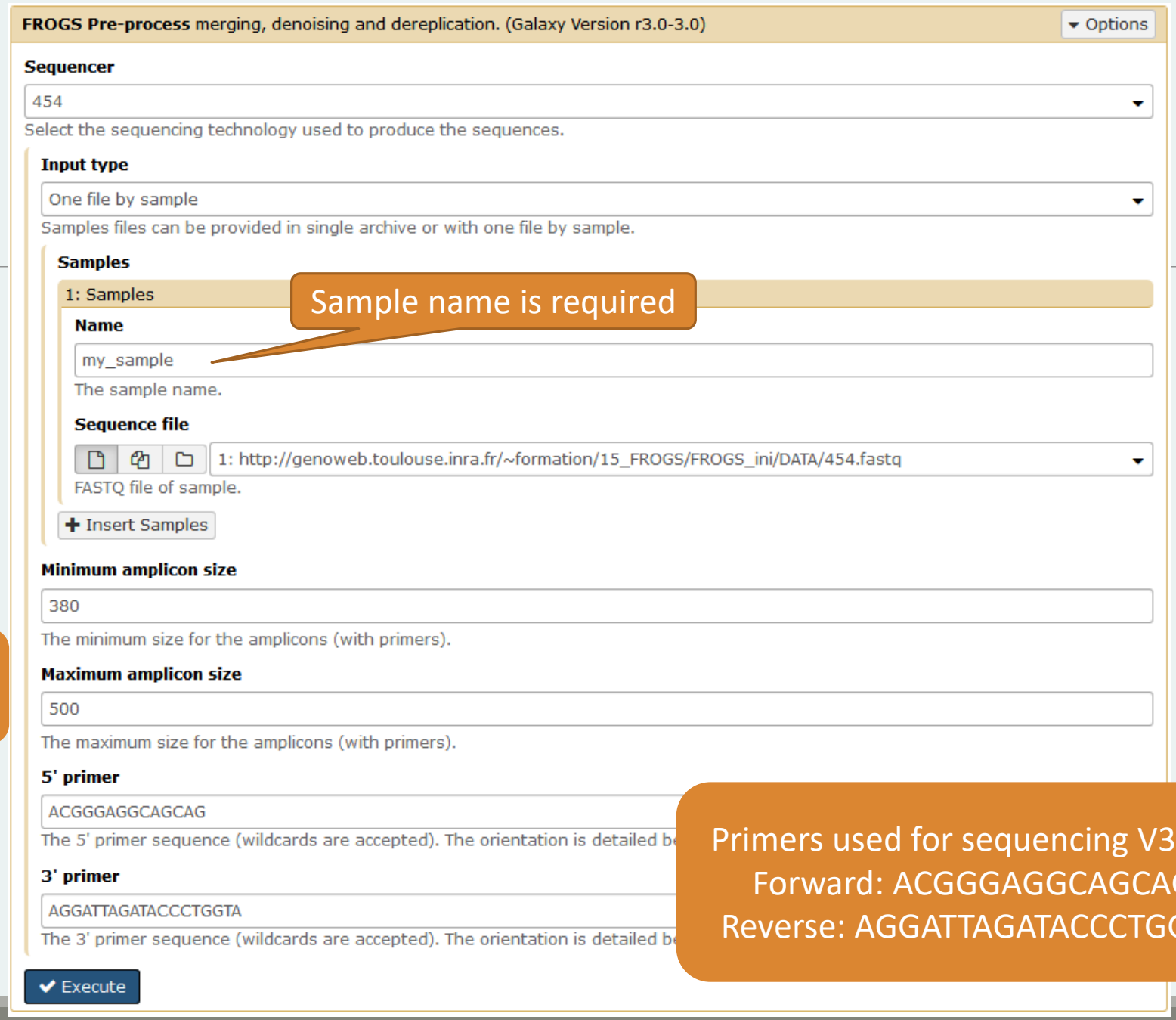
Launch the pre-process tool on that data set

→ objective : understand the parameters

1- Test different parameters for « minimum and maximum amplicon size »

2- Enter these primers: Forward: ACGGGAGGCAGCAG  Reverse: AGGATTAGATACCCTGGTA

454




Size range of 16S V3-V4:
[ 380 – 500 ]

Primers used for sequencing V3-V4:
Forward: ACGGGAGGCAGCAG
Reverse: AGGATTAGATACCCTGGTA

454

# Exercise 2.1

What do you understand about amplicon size, which file can help you ?

What is the length of your reads before preprocessing ?

Do you understand how enter your primers ? 

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

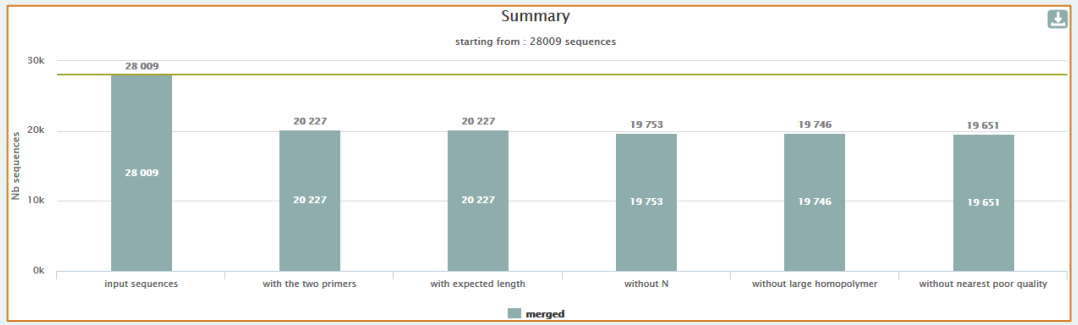
What is the «  FROGS Pre-process: count.tsv » file ? 

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

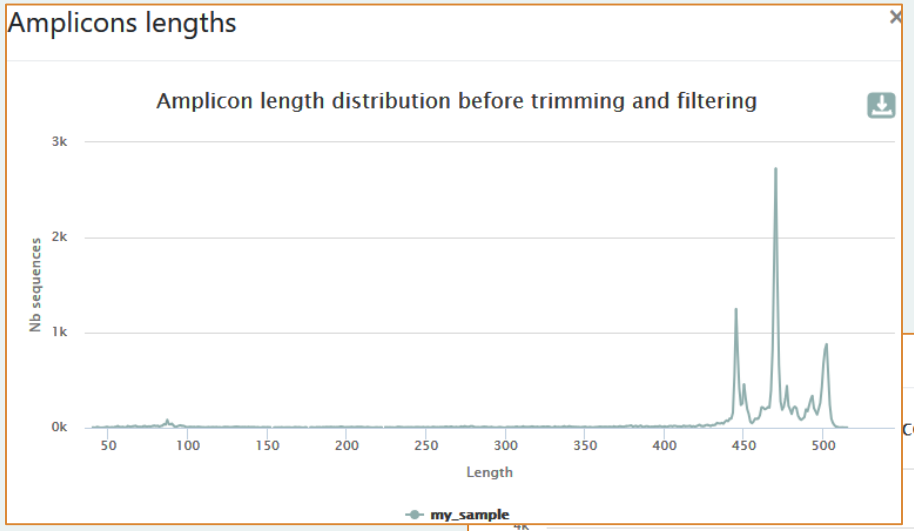
*Who loose a lot of sequences ?*

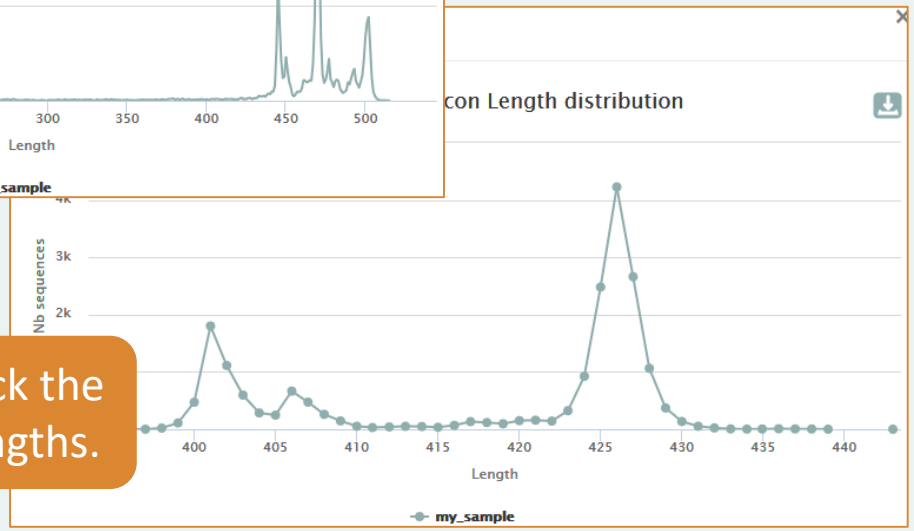
454

| Samples | % kept | input sequences | with the two primers | with expected length | without N | without large homopolymer | without nearest poor quality |
|---|---|---|---|---|---|---|---|
| my_sample | 70.16 | 28,009 | 20,227 | 20,227 | 19,753 | 19,746 | 19,651 |




To be kept, sequences must have the 2 primers



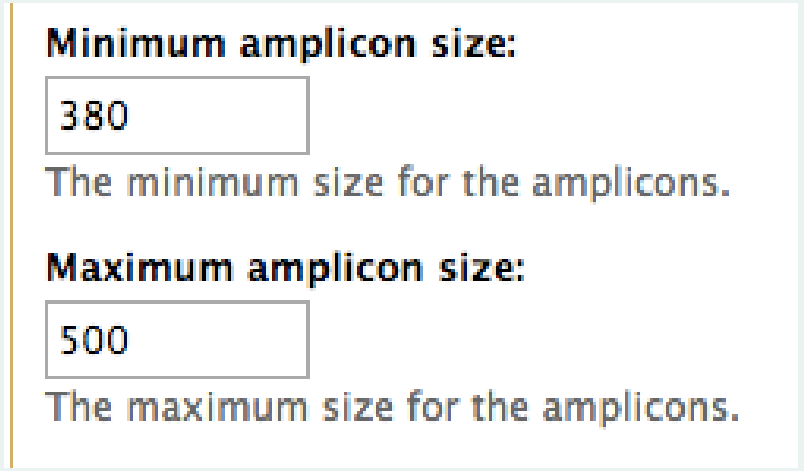




To adjust your filtering, check the distribution of sequence lengths.

454

# Cleaning, how it work ?

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




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

454

# 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.

# Pre-process

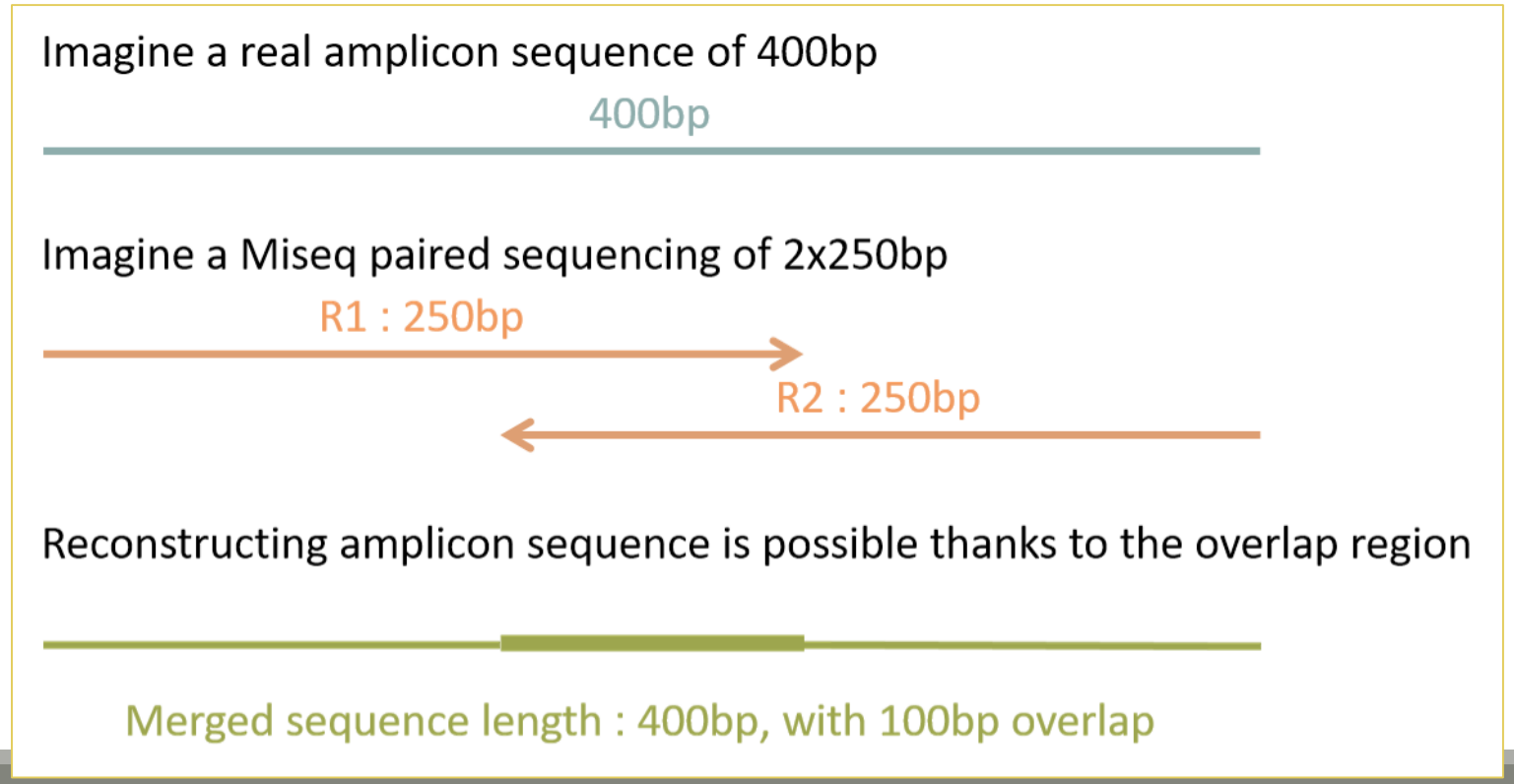
- Delete sequence with not expected lengths
- Delete sequences with ambiguous bases (N)
- Delete sequences do not contain good primers
- Merging of reads ← VSEARCH
- Dereplication

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

**VSEARCH: a versatile open source tool for metagenomics.**
Rognes T, Flouri T, Nichols B, Quince C, Mahé F.
PeerJ. 2016 Oct 18;4:e2584. eCollection 2016.

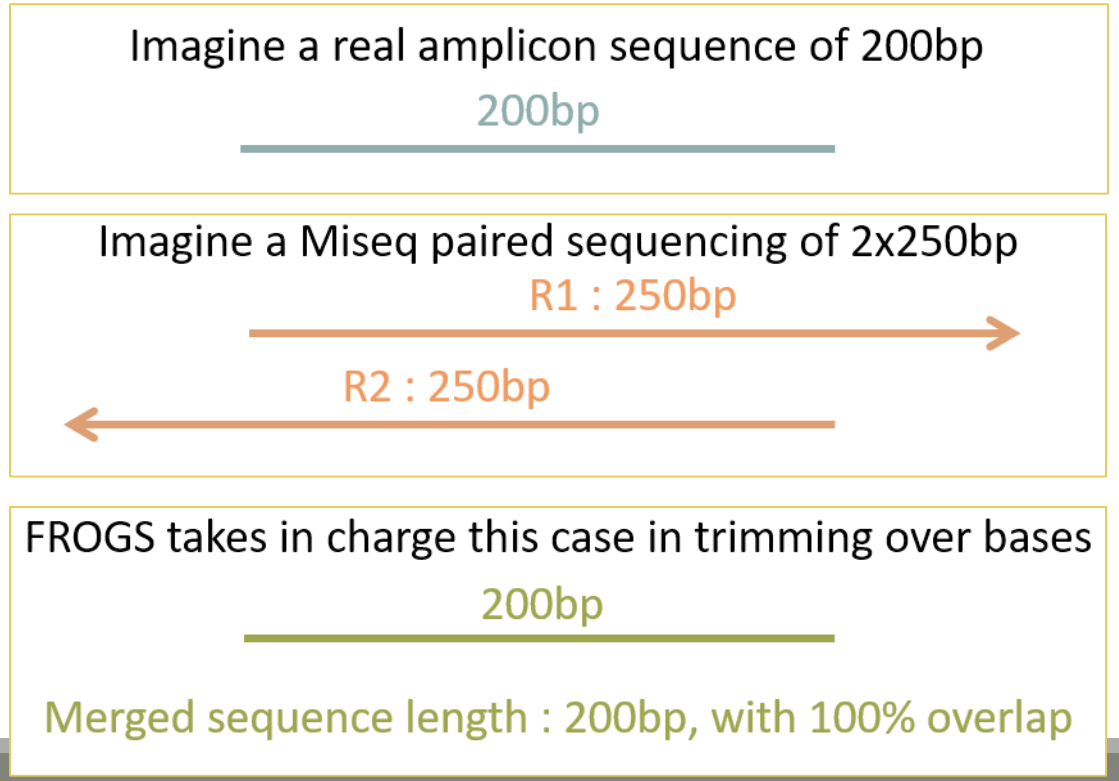
# The aim of Vsearch is to merge R1 with R2

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



# The aim of Vsearch is to merge R1 with R2

Case of a sequencing of over-overlapping sequences:



Miseq R1 R2

# Exercise 2.2

Go to « MiSeq R1 R2 » history

Launch the pre-process tool on that data set

→ objective: understand Vsearch software

**FROGS Pre-process** Step 1 in metagenomics analysis: denoising and dereplication. (Galaxy Ver

**Sequencer**

Illumina

Select the sequencing technology used to produce the sequences.

**Input type**

Files by samples

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

**Reads already contiged ?**

No

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

**Samples**

1: Samples

**Name**

sampleA

The sample name.

Sample name is required

**Reads 1**

59: /work/formation/FROGS/sampleA_R1.fastq

R1 FASTQ file of paired-end reads.

**reads 2**

60: /work/formation/FROGS/sampleA_R2.fastq

R2 FASTQ file of paired-end reads.

+ Insert Samples

**Reads 1 size**

250

The read1 size.

**Reads 2 size**

250

The read2 size.

```
>ERR619083.M00704
CCGTCAATTCATTGAGTTTCAACCTTGCGGCCGTACTTCCCAGGCGGTACGTT
TATCGCGTTAGCTTCGCCAAGCACAGCATCCTGCGCTTAGCCAACGTACATCG
TTTAGGGTGTGGACTACCCGGGTATCTAATCCTGTTCGCTACCCACGCTTTCG
AGCCTCAGCGTCAGTGACAGACCAGAGAGCCGCTTTCGCCACTGGTGTTCCTC
CATATATCTACGCATTTCACCGCTACACATGGAATTCCACTCTCCCCTTCTGC
ACTCAAGTCAGACAGTTTCCAGAGCACTCTATGGTTGAGCCATAGCCTTTTAC
TCCAGACTTTCCTGACCGACTGCACTCGCTTTACGCCCAATAAATCCGGACAA
CGCTTGCCACCTACGTATTACCGCNGCTGCT
```

Miseq R1 R2

Real 16S sequenced fragment

**mismatch rate.**

0.1

The maximum rate of mismatches in the overlap region

**Merge software**

Vsearch

Select the software to merge paired-end reads.

Do not use flash

**Would you like to keep unmerged reads?**

Yes | No

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

**Minimum amplicon size**

340

The minimum size for the amplicons (with primers).

Reads can be overlapped

**Maximum amplicon size**

450

The maximum size for the amplicons (with primers).

**Sequencing protocol**

Illumina standard

The protocol used for sequencin

**5' primer**

CCGTCAATTC

The 5' primer sequence (wildc

**3' primer**

CCGCNGCTGCT

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

Primers used for sequencing V5 region:
Forward: CCGTCAATTC
Reverse: CCGCNGCTGCT
Lecture 5’ → 3’

✔ Execute

Miseq R1
R2

# Exercise 2.2

What do you understand about amplicon size, which file can help you ?

What is the length of your reads before preprocessing ?

Do you understand how enter your primers ? 

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

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

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

*Who loose a lot of sequences ?*

Miseq R1
R2

**Expected amplicon size**

410

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

**mismatch rate.**

0.1

The maximum ra… of mismatches in the overlap region

To increase, if your sequences have low qualities
Use FASTQC to know it!

**Minimum amplicon size**

340

The minimum size for the amplicons.

**Maximum amplicon size**

450

The maximum size for the amplicons.

**Sequencing protocol**

Illumina standard

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

**5' primer**

CCGTCAATTC

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

**3' primer**

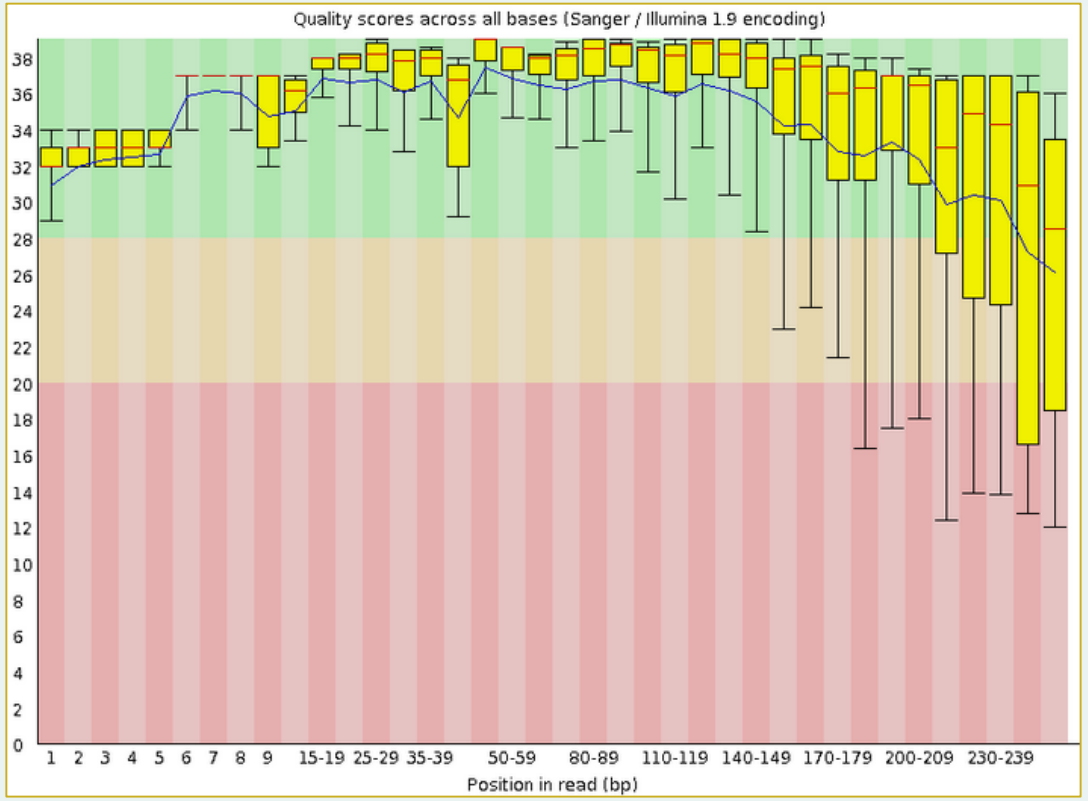
CCGCNGCTGCT

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

✔ Execute

**FastQC: fastq/sam/bam**

FastQC:Read QC reports using FastQC



Escape ITS data

# Exercise 2.3

Go to « ITS » history

Launch the pre-process tool on this data set

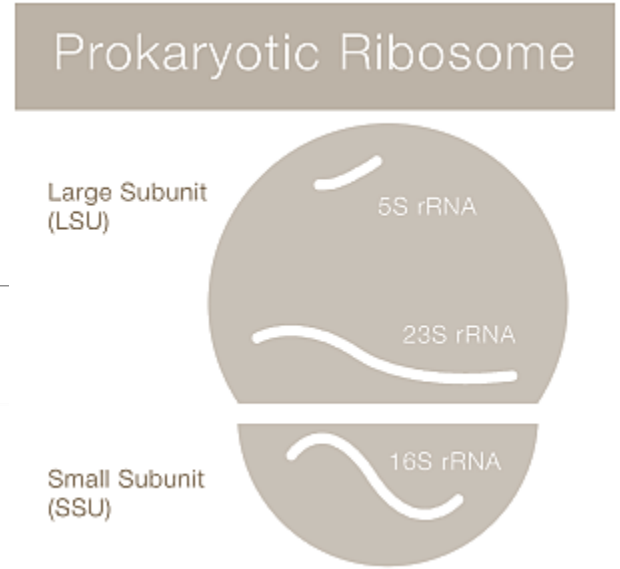
→ objective : understand the « combined sequences »

→ objective : work with non-overlapping reads

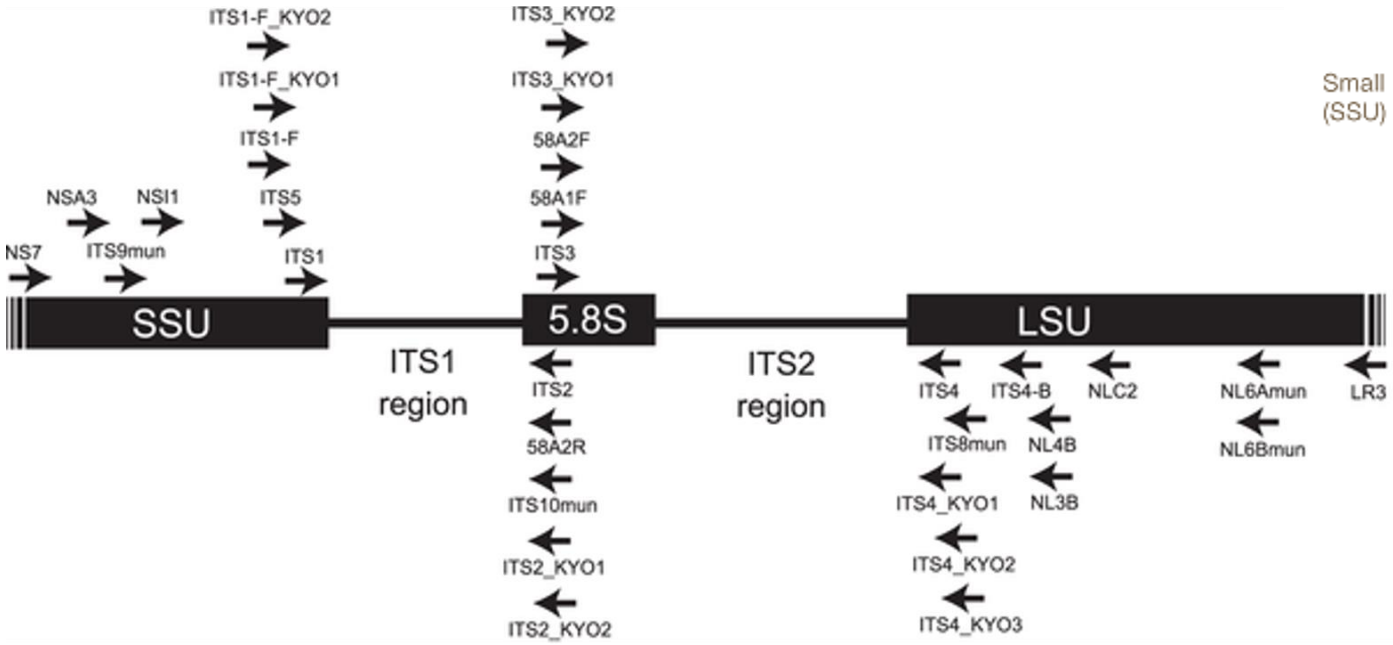
1- Enter these primers:

Forward: CTTGGTCATTTAGAGGAAGTAA  Reverse: GCATCGATGAAGAACGCAGC

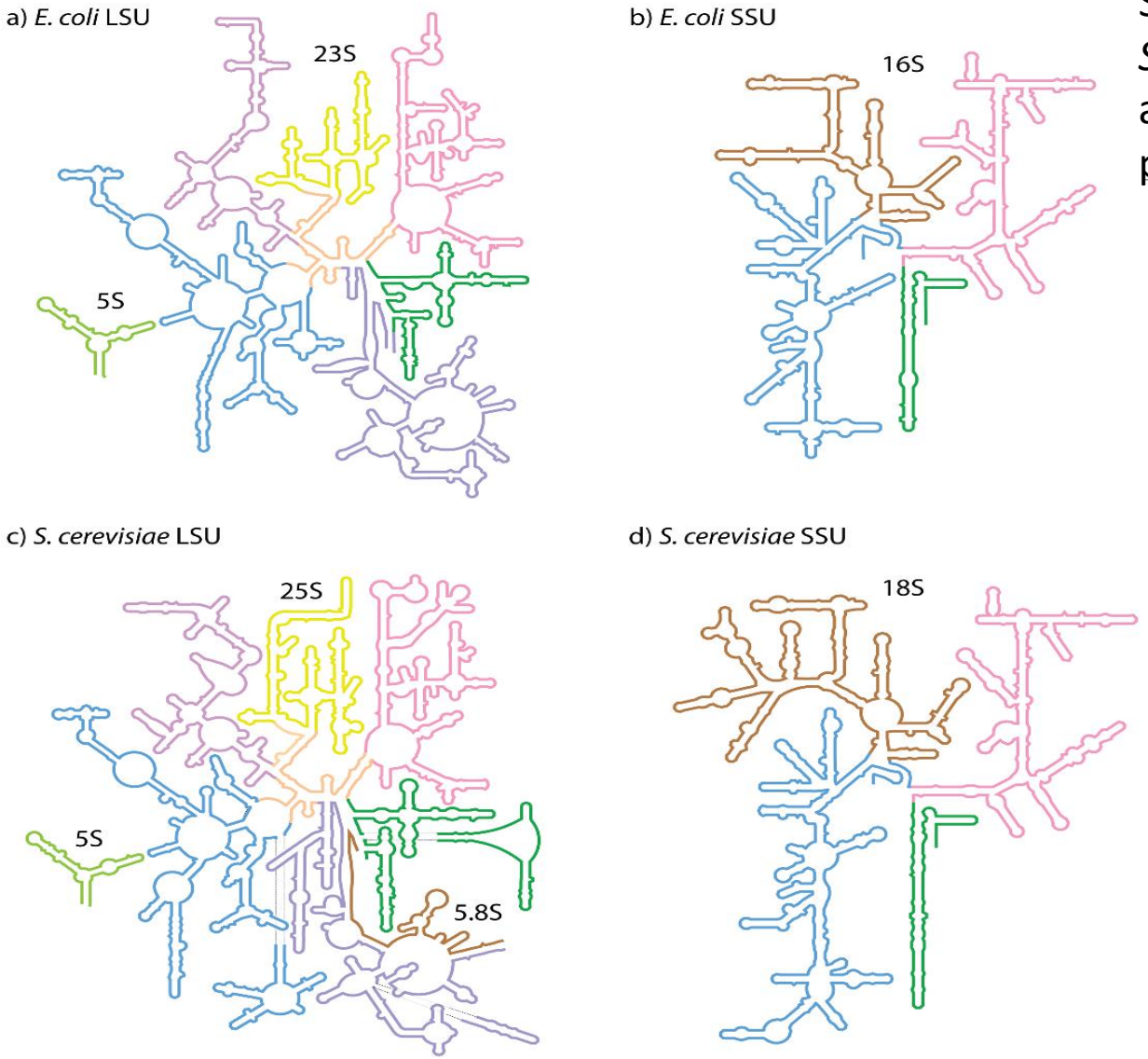
# What is a ITS ?




**Map of nuclear ribosomal RNA genes and their ITS regions.**



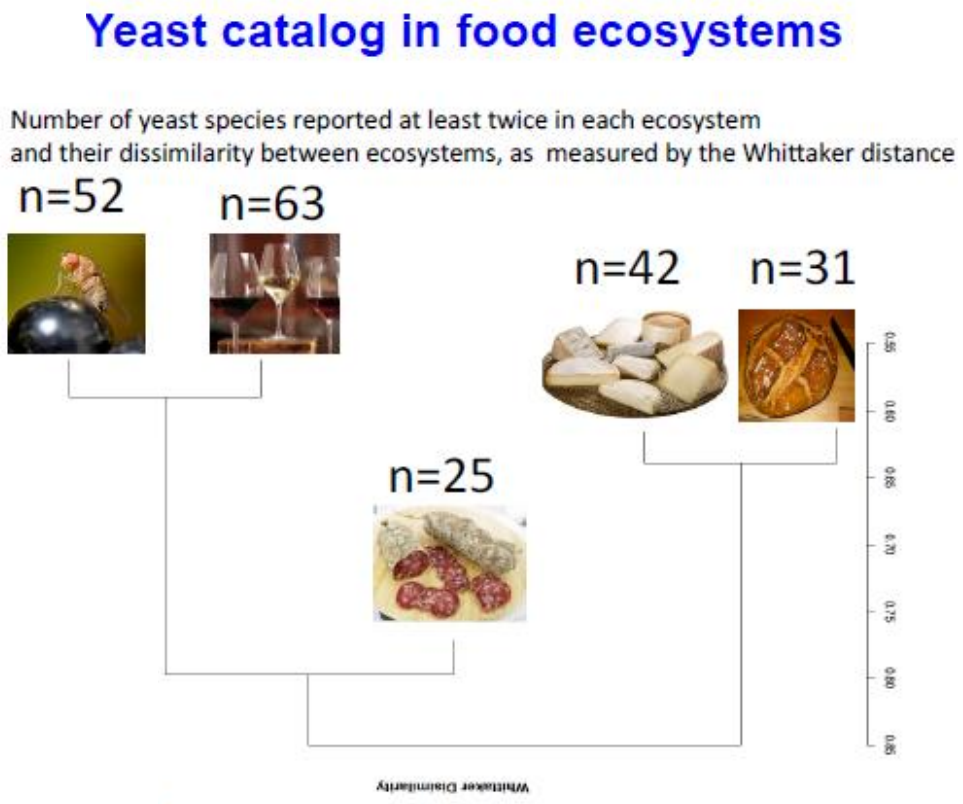

Toju H, Tanabe AS, Yamamoto S, Sato H (2012) High-Coverage ITS Primers for the DNA-Based Identification of Ascomycetes and Basidiomycetes in Environmental Samples. PLOS ONE 7(7): e40863. https://doi.org/10.1371/journal.pone.0040863




Schematic rRNA 2° structures of a) *E. coli* LSU, b) *E. coli* SSU, c) *S. cerevisiae* LSU, and d) *S. cerevisiae* SSU. These 2° structures are derived from 3D structures, and include non-canonical base pairs.

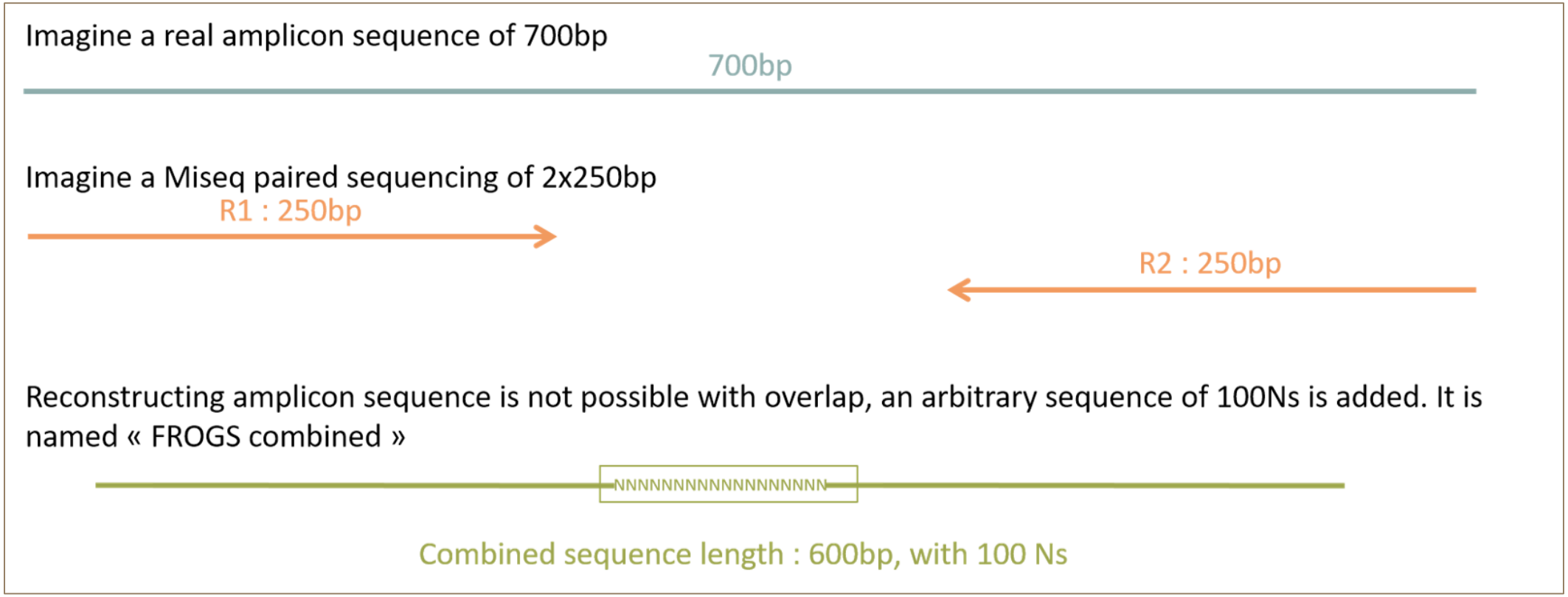
**Secondary Structures of rRNAs from All Three Domains of Life**
Anton S. Petrov , Chad R. Bernier, Burak Gulen, Chris C. Waterbury, Eli Hershkovits, Chiaolong Hsiao, Stephen C. Harvey, Nicholas V. Hud, George E. Fox, Roger M. Wartell, Loren Dean Williams
February 5, 2014 https://doi.org/10.1371/journal.pone.0088222

# ITS data form METABARFOOD Project metaprogramme MEM

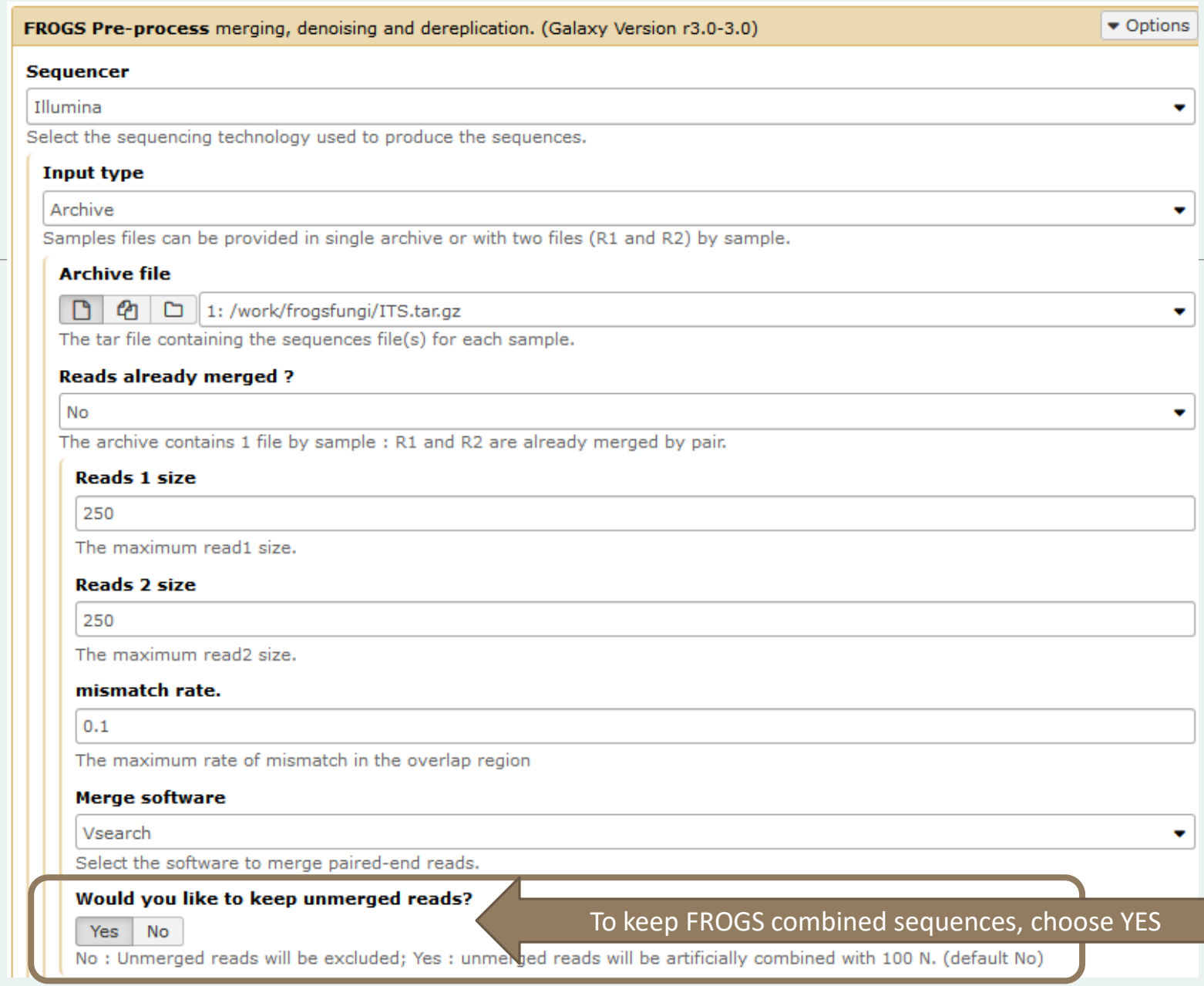



- While metabarcoding is commonly used to describe prokaryotes in the microbiome of many environments, methods for describing micro-eukaryote diversity is lacking and requires better methodology and standardisation.
- One reason is that the universal fungal barcode, the Internal Transcribed Spacer (ITS) region, displays considerable size variation amongst yeasts and other micro-eukaryotes.
- There are also several repeats leading to sequencing errors or termination.
- Additionally, the ITS databases are far from complete, especially for Ascomycota that are commonly found in food.
- Other rDNA barcodes have been used but often do not harbor enough polymorphism to detect taxa to the species level.
- In food, microbiota are usually composed of a reduced number of species compared to wild environments.
- Detecting micro-eukaryotes at the species level, and potentially strain level, is therefore necessary.

# Case of ITS1 amplicon MiSeq sequencing, a case of a sequencing of non-overlapping sequences



ITS


FROGS Pre-process merging, denoising and dereplication. (Galaxy Version r3.0-3.0)
Options
Sequencer
Illumina
Select the sequencing technology 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/frogsfungi/ITS.tar.gz
The tar file containing the sequences file(s) for each sample.
Reads already merged ?
No
The archive contains 1 file by sample : R1 and R2 are already merged by pair.
Reads 1 size
250
The maximum read1 size.
Reads 2 size
250
The maximum read2 size.
mismatch rate.
0.1
The maximum rate of mismatch in the overlap region
Merge software
Vsearch
Select the software to merge paired-end reads.
Would you like to keep unmerged reads?
Yes
No
No : Unmerged reads will be excluded; Yes : unmerged reads will be artificially combined with 100 N. (default No)
To keep FROGS combined sequences, choose YES

ITS

**Minimum amplicon size**

50

The minimum size for the amplicons (with primers).

**Maximum amplicon size**

490

The maximum size for the amplicons (with primers).

**Sequencing protocol**

Illumina standard

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

**5' primer**

CTTGGTCATTTAGAGGAAGTAA

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

**3' primer**

GCATCGATGAAGAACGCAGC

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

✔ Execute

# Exercise 2.3

Go to « ITS » history

Launch the pre-process tool on this data set

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

# Explore Preprocess report.html

ITS

# Explore Preprocess report.html

2 tables:

## Details on merged sequences

CSV

Show 10 entries Search:

| Samples | % kept | paired-end assembled | with 5' primer | with 3' primer | with expected length | without N |
|---|---|---|---|---|---|---|
| complexe-ADN-1 | 91.09 | 54,121 | 49,322 | 49,303 | 49,303 | 49,299 |
| echantillon1-1 | 84.93 | 31,836 | 27,059 | 27,040 | 27,040 | 27,039 |
| echantillon1-2 | 94.73 | 54,774 | 51,938 | 51,895 | 51,895 | 51,890 |
| echantillon1-3 | 74.90 | 81,611 | 61,197 | 61,135 | 61,134 | 61,128 |
| echantillon2-1 | 90.17 | 51,984 | 46,886 | 46,875 | 46,874 | 46,873 |

## Details on artificial combined sequences

CSV

Show 10 entries Search:

| Samples | % kept | paired-end assembled | with 5' primer | with 3' primer | with expected length | without N |
|---|---|---|---|---|---|---|
| complexe-ADN-1 | 68.47 | 2,163 | 1,833 | 1,656 | 1,481 | 1,481 |
| echantillon1-1 | 54.92 | 1,047 | 751 | 620 | 575 | 575 |
| echantillon1-2 | 61.57 | 1,392 | 1,096 | 942 | 858 | 857 |
| echantillon1-3 | 49.54 | 2,491 | 1,617 | 1,334 | 1,234 | 1,234 |
| echantillon2-1 | 44.62 | 1,421 | 996 | 899 | 634 | 634 |

ITS

# Explore Preprocess report.html

2 tables:

## Details on merged sequences

CSV

Show 10 entries Search:

| Samples | % kept | paired-end assembled | with 5' primer | with 3' primer | with expected length | without N |
|---|---|---|---|---|---|---|
| complexe-ADN-1 | 91.09 | 54,121 | 49,322 | 49,303 | 49,303 | 49,299 |
| echantillon1-1 | 84.93 | 31,836 | 27,059 | 27,040 | 27,040 | 27,039 |
| echantillon1-2 | 94.73 | 54,774 | 51,938 | 51,895 | 51,895 | 51,890 |
| echantillon1-3 | 74.90 | 81,611 | 61,197 | 61,135 | 61,134 | 61,128 |
| echantillon2-1 | 90.17 | 51,984 | 46,886 | 46,875 | 46,874 | 46,873 |

## Details on artificial combined sequences

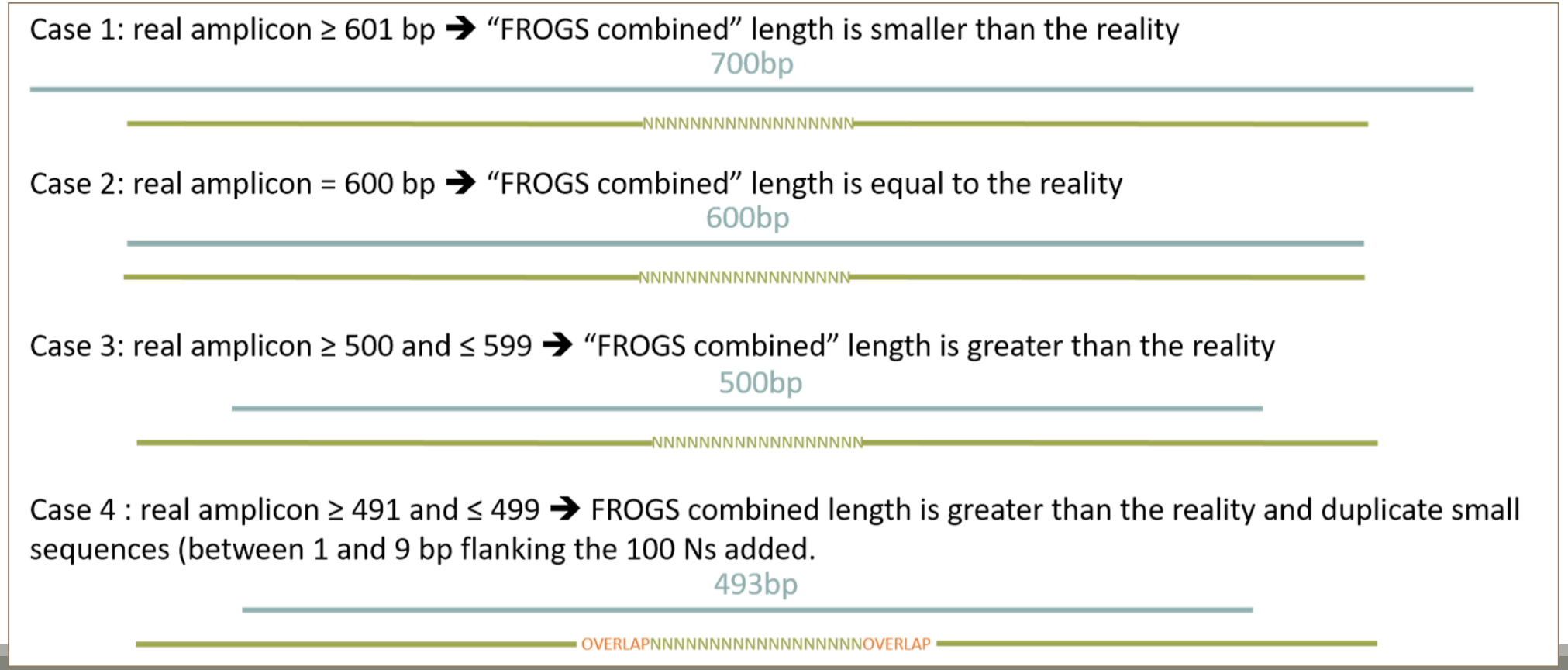
CSV

Show 10 entries Search:

| Samples | % kept | paired-end assembled | with 5' primer | with 3' primer | with expected length | without N |
|---|---|---|---|---|---|---|
| complexe-ADN-1 | 68.47 | 2,163 | 1,833 | 1,656 | 1,481 | 1,481 |
| echantillon1-1 | 54.92 | 1,047 | 751 | 620 | 575 | 575 |
| echantillon1-2 | 61.57 | 1,392 | 1,096 | 942 | 858 | 857 |
| echantillon1-3 | 49.54 | 2,491 | 1,617 | 1,334 | 1,234 | 1,234 |
| echantillon2-1 | 44.62 | 1,421 | 996 | 899 | 634 | 634 |

# FROGS "combined" sequences are artificial and present particular features especially on size.

Imagine a MiSeq sequencing of 2x250pb with reads impossible to overlap. So FROGS "combined" length = 600 bp.



Miseq merged

# Exercise 2.4

Go to« MiSeq merged » history

Launch the pre-process tool on that data set

→ objective: understand output files

Miseq merged

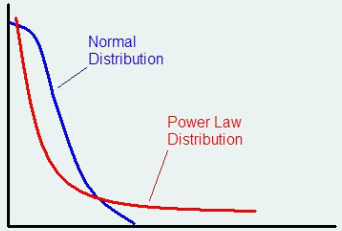
# Exercise 2.4

3 samples are **technically replicated** 3 times : 9 samples of 10 000 sequences each.

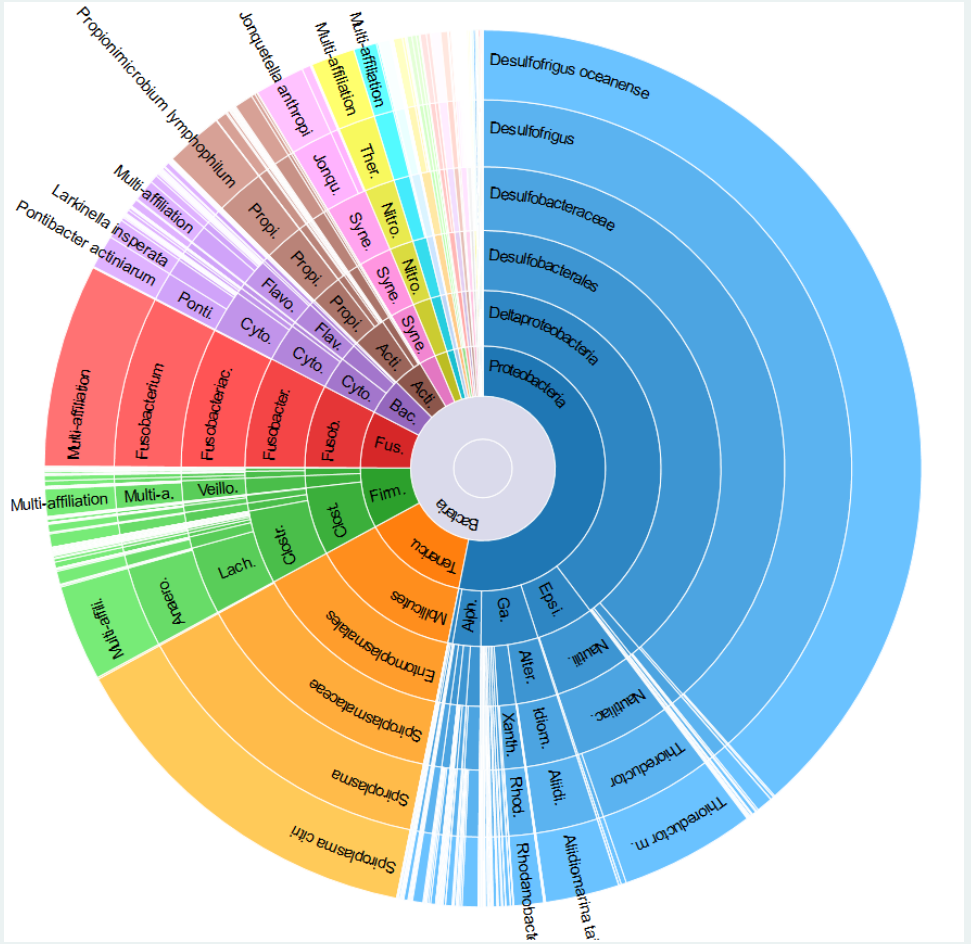
| | | |
|---|---|---|
| 100_10000seq_sampleA1.fastq | 100_10000seq_sampleB1.fastq | 100_10000seq_sampleC1.fastq |
| 100_10000seq_sampleA2.fastq | 100_10000seq_sampleB2.fastq | 100_10000seq_sampleC2.fastq |
| 100_10000seq_sampleA3.fastq | 100_10000seq_sampleB3.fastq | 100_10000seq_sampleC3.fastq |

Miseq merged

# Exercise 2.4




- 100 species, covering all bacterial phyla
- Power Law distribution of the species abundances
- Error rate calibrated with real sequencing runs
- 10% chimeras
- 9 samples of 10 000 sequences each (90 000 sequences)



Miseq merged

# Exercise 2.4

"Grinder (v 0.5.3) (Angly et al., 2012) was used to simulate the PCR amplification of full-length (V3-V4) sequences from reference databases. The reference database of size 100 were generated from the LTP SSU bank (version 115) (Yarza et al., 2008) by

(1) filtering out sequences with a N,

(2) keeping only type species

(3) with a match for the forward (ACGGGAGGCAGCAG) and reverse (TACCAGGGTATCTAATCCTA) primers in the V3-V4 region and

(4) maximizing the phylogenetic diversity (PD) for a given database size. The PD was computed from the NJ tree distributed with the LTP."

Miseq merged

**FROGS Pre-process** Step 1 in metagenomics analysis: denoising and dereplication. (Galaxy Version 2.0.0) ▾ Options

**Sequencer**

Illumina

Select the sequencing technology 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**

2: /work/formation/FROGS/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 : R1 and R2 are already merged 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

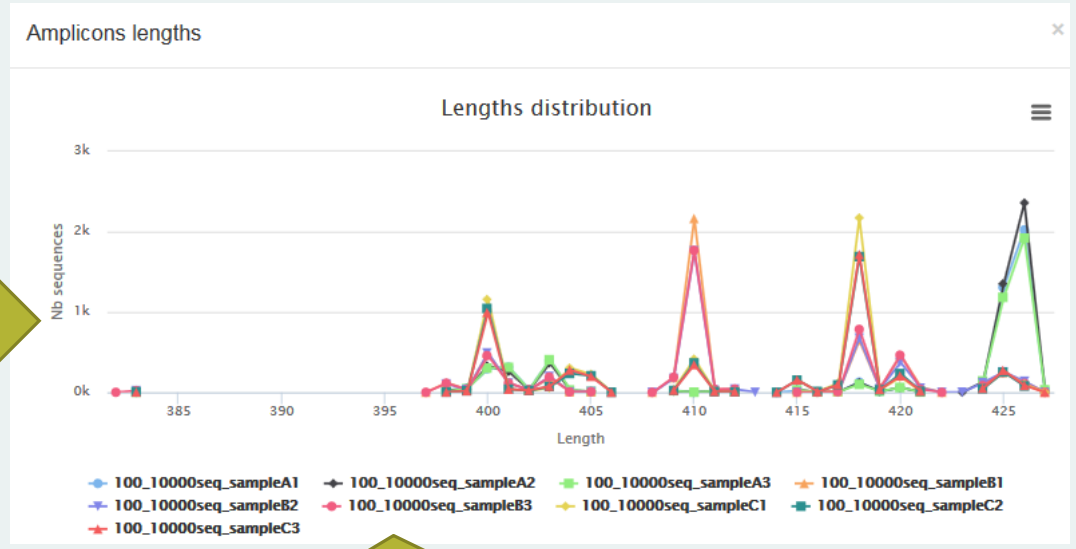
The 5' primer sequence (wildcards are accepted). The orientation is detai

**3' primer**

TAGGATTAGATACCCTGGTA

The 3' primer sequence (wildcards are accepted). The orientation is detai

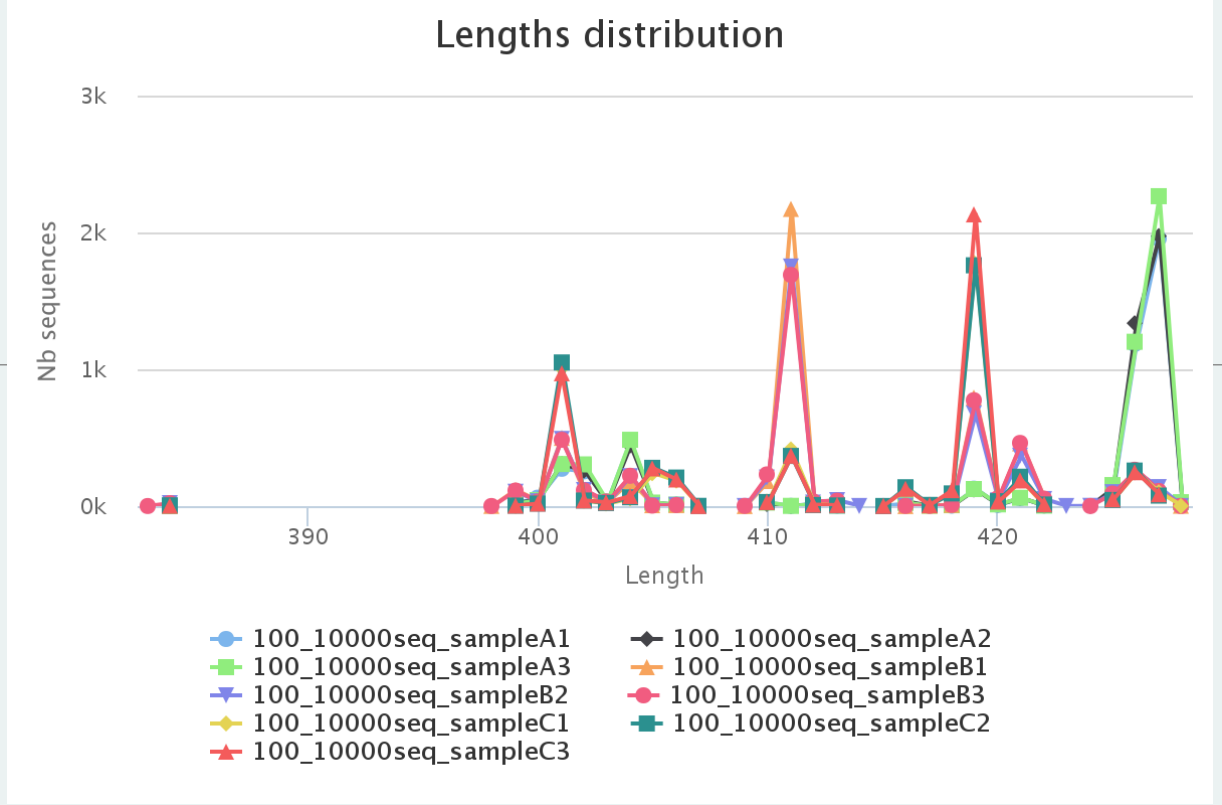
✔ Execute



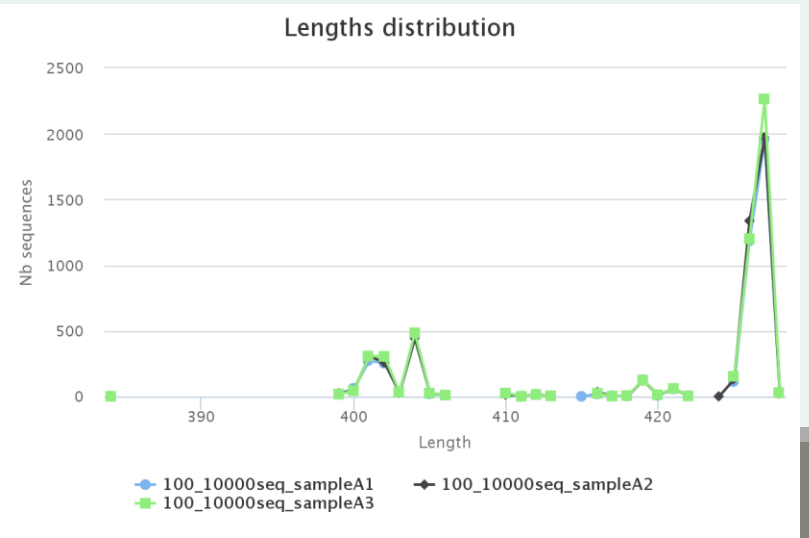

Click on legend

Primers used for this sequencing :
5’ primer: ACGGGAGGCAGCAG
3’ primer: TAGGATTAGATACCCTGGTA
Lecture 5’ → 3’

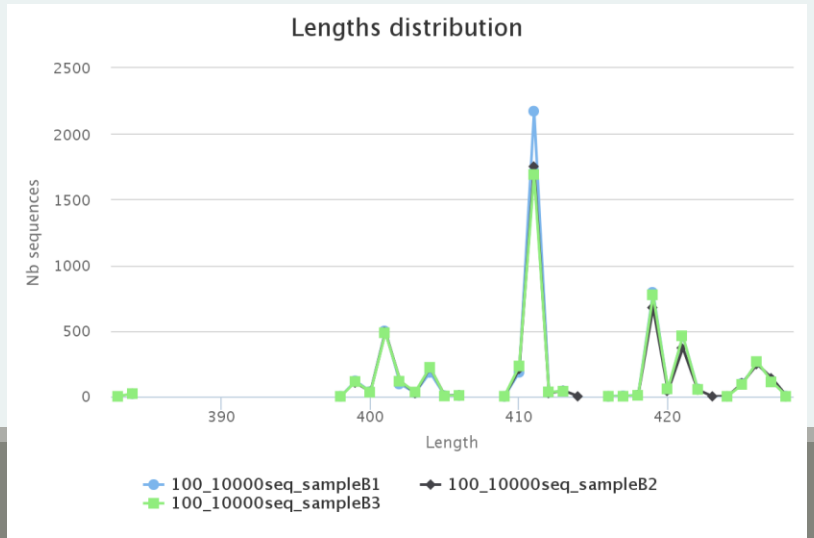
Miseq merged



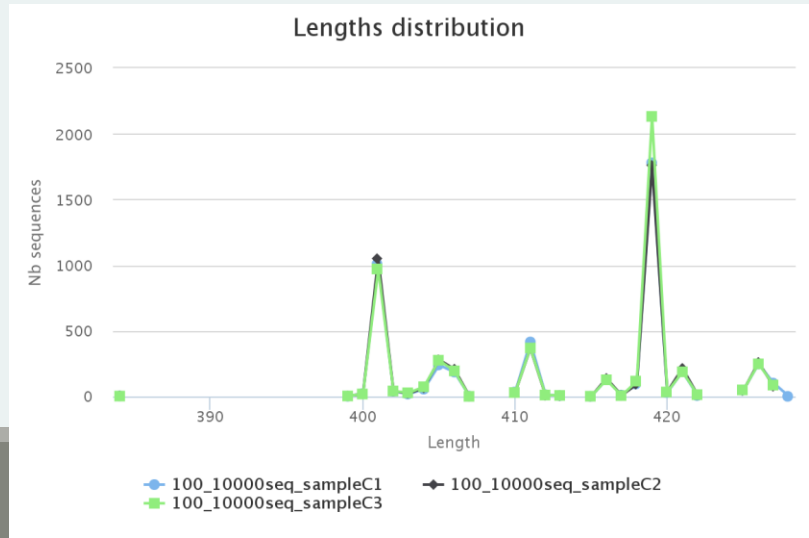

Samples A only




Samples B only




Samples C only



Miseq merged

# Exercise 2.4 - Questions

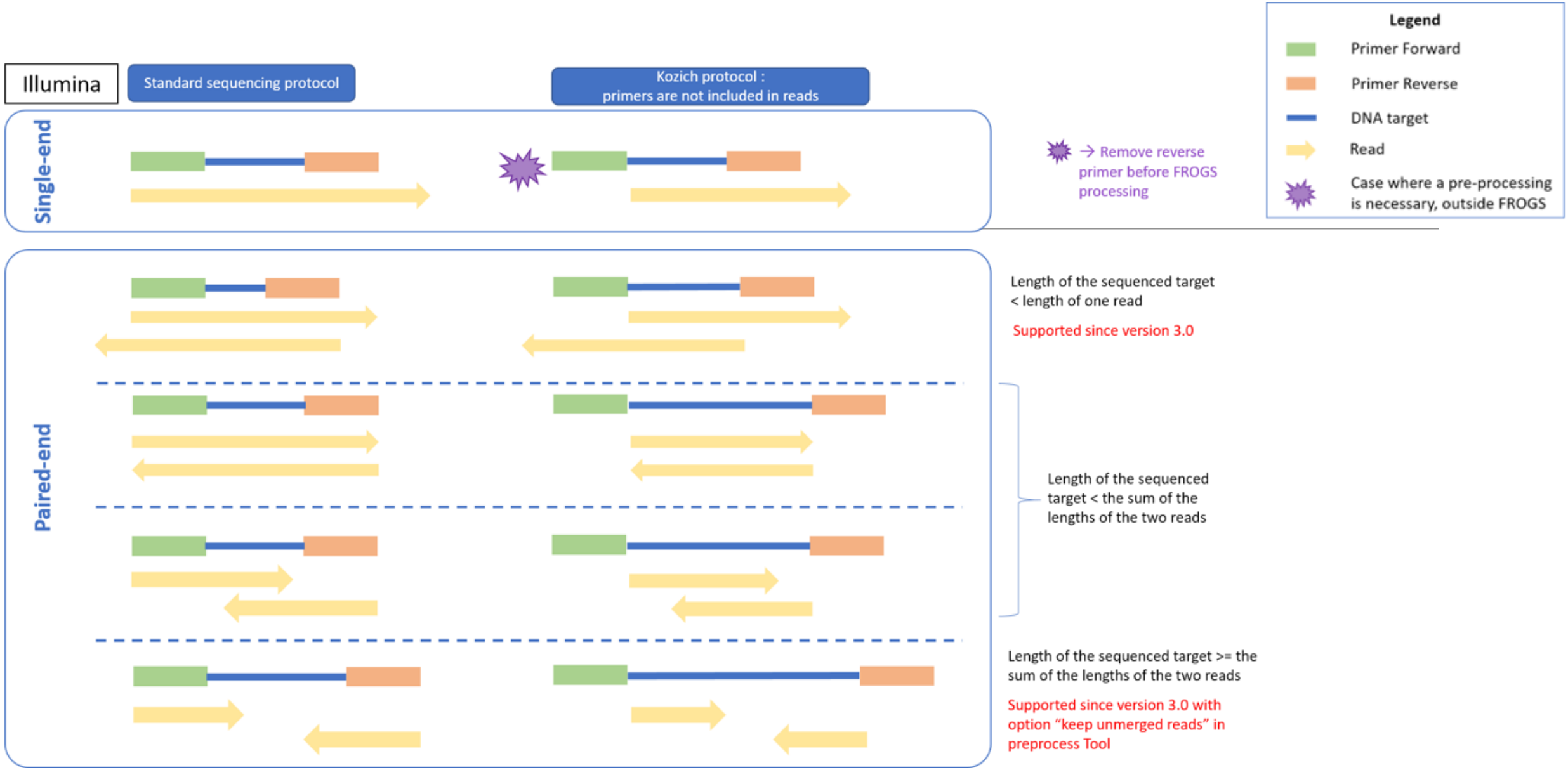
1. How many sequences are there in the input file ?
2. How many sequences did not have the 5’ primer?
3. How many sequences still are after pre-processing the data?
4. How much time did it take to pre-process the data ?
5. What can you tell about the sample based on sequence length distributions ?

# Preprocess tool in bref

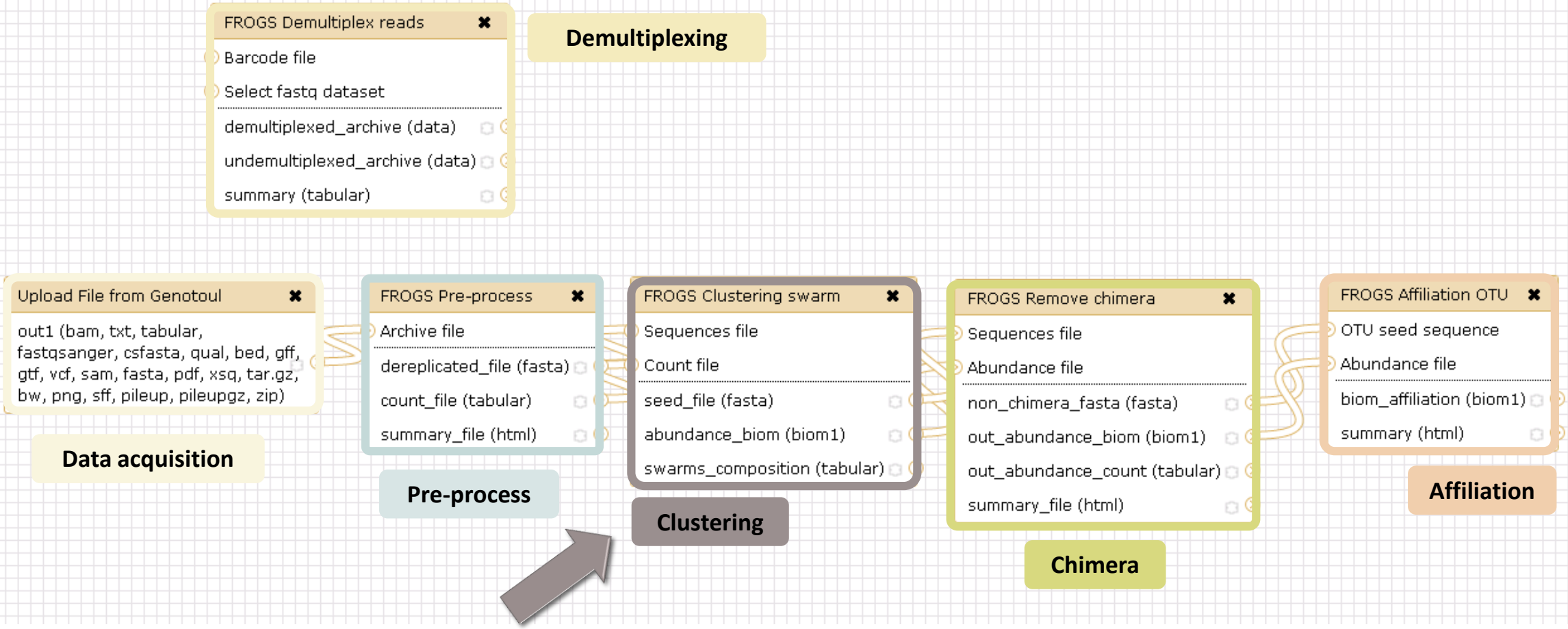
| | Take in charge |
|---|---|
| Illumina | ✓ |
| 454 | ✓ |
| Merged data | ✓ |
| Not merged data | ✓ |
| Without primers | ✓ |
| Only R1 or only R2 | ⃠ |
| Too distant R1 and R2 to be merged | ✓ |
| Over-overlapping R1 R2 | ✓ |

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

# Processed data by FROGS in bref

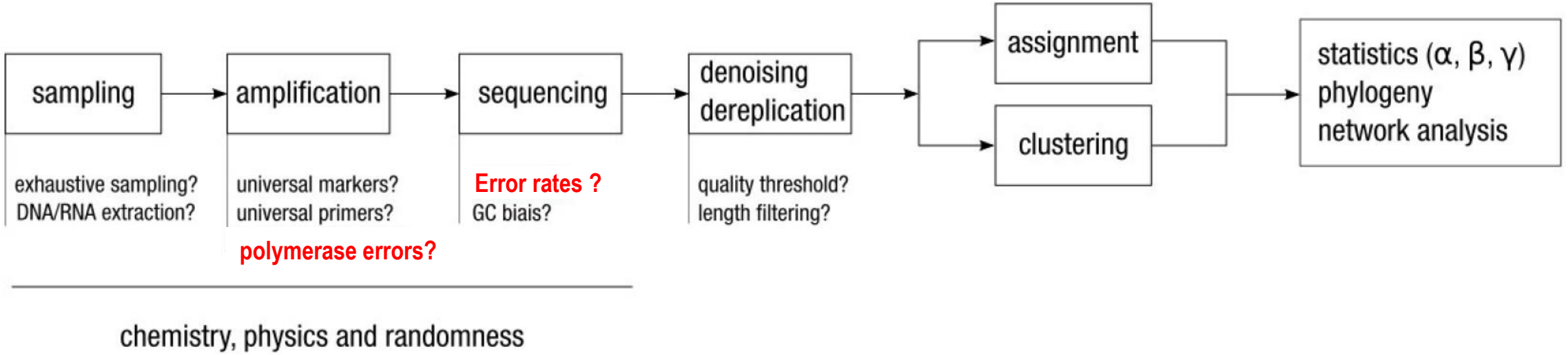

Illumina
Standard sequencing protocol
Kozich protocol :
primers are not included in reads
Single-end
Paired-end
→ Remove reverse primer before FROGS processing
Legend
Primer Forward
Primer Reverse
DNA target
Read
Case where a pre-processing is necessary, outside FROGS
Length of the sequenced target < length of one read
Supported since version 3.0
Length of the sequenced target < the sum of the lengths of the two reads
Length of the sequenced target >= the sum of the lengths of the two reads
Supported since version 3.0 with option “keep unmerged reads” in preprocess Tool

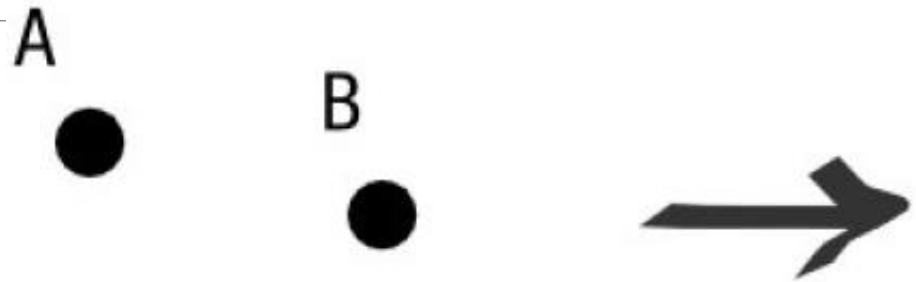
# Clustering tool

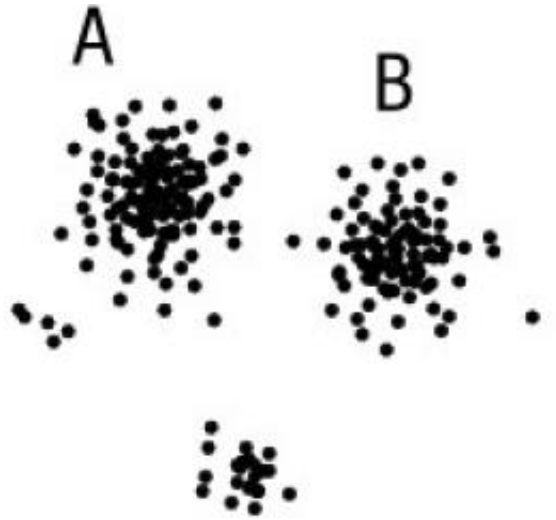

FROGS Demultiplex reads
Barcode file
Select fastq dataset
demultiplexed_archive (data)
undemultiplexed_archive (data)
summary (tabular)
Demultiplexing
Upload File from Genotoul
out1 (bam, txt, tabular, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png, sff, pileup, 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



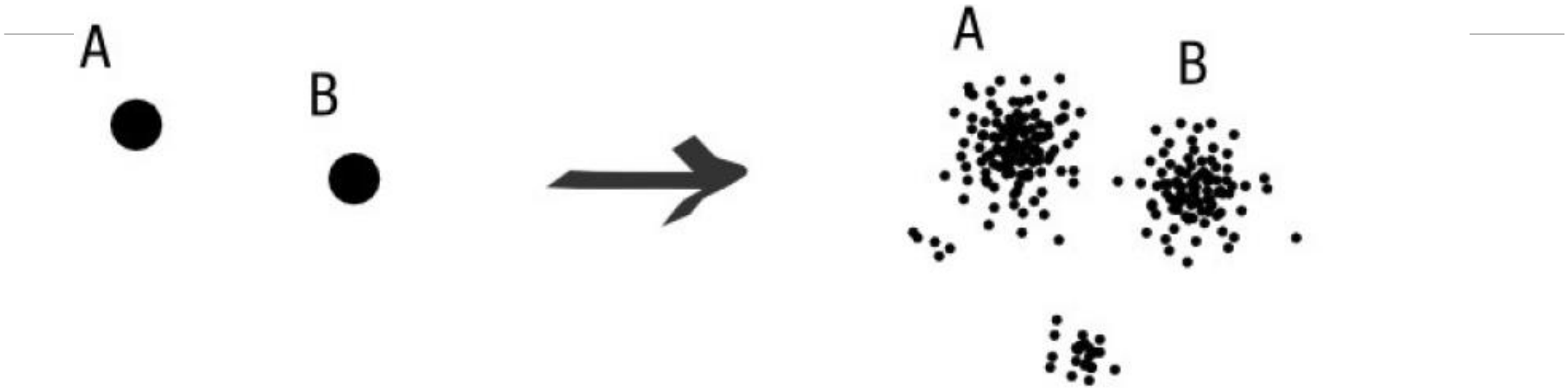



Natural variability ?
Technical noise?
Contaminant?
Chimeras?

Expected

Results

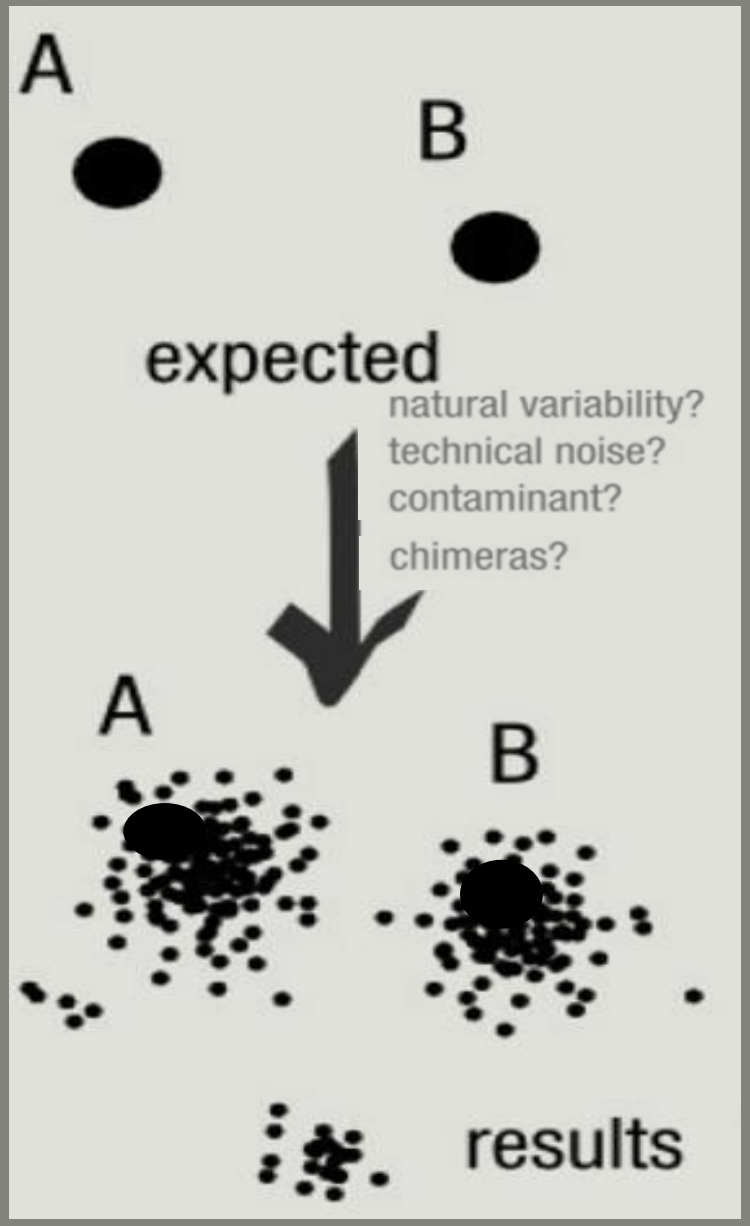



Natural variability ?
Technical noise?
Contaminant?
Chimeras?

Expected

Results

16S variability
*Cf.* RRNDB (ribosomal RNA operons database)
max. 21 copies of 16S in bacteria (*Photobacterium damselae*)
ex. *E. coli* 7 copies



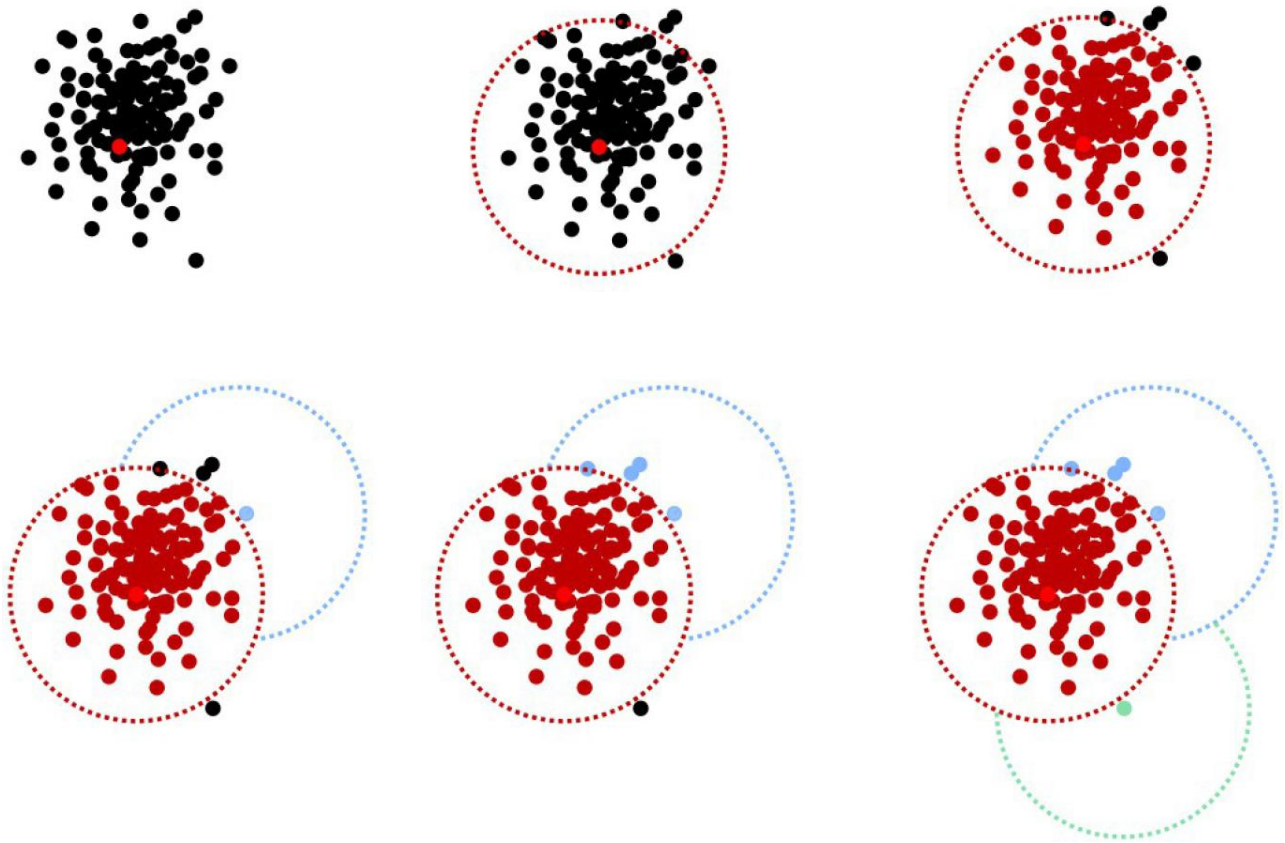

# To have the best accuracy:

## Method: All against all

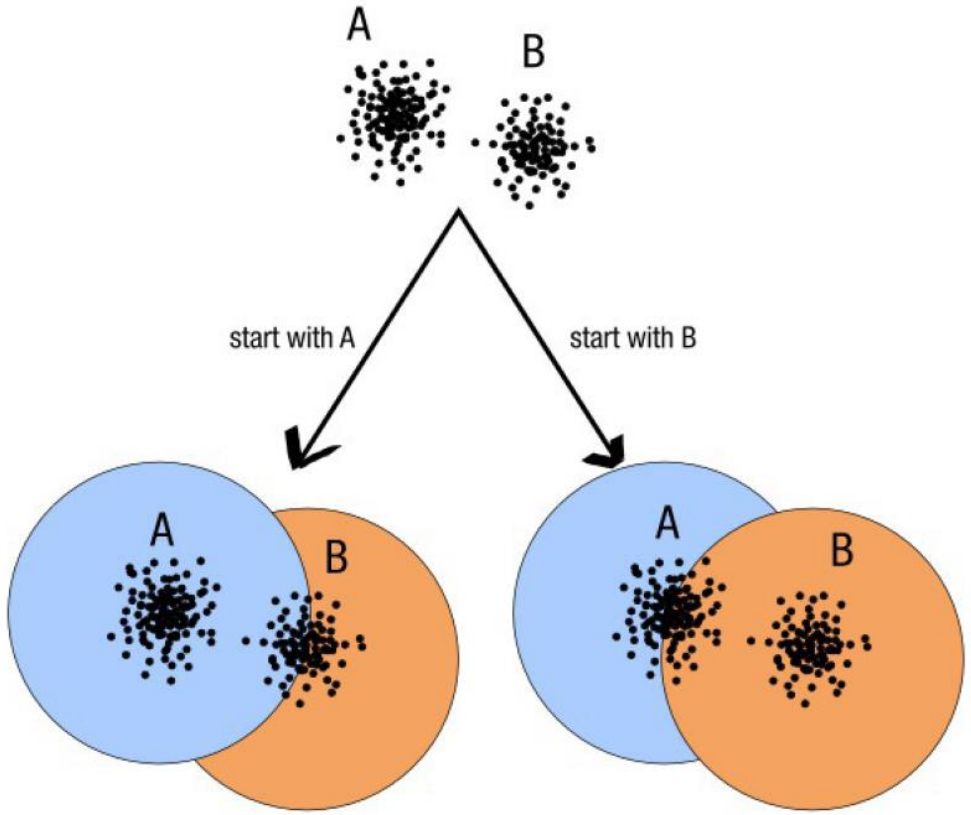
- Very accurate
- Requires a lot of memory and/or time

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

# How traditional clustering works ?

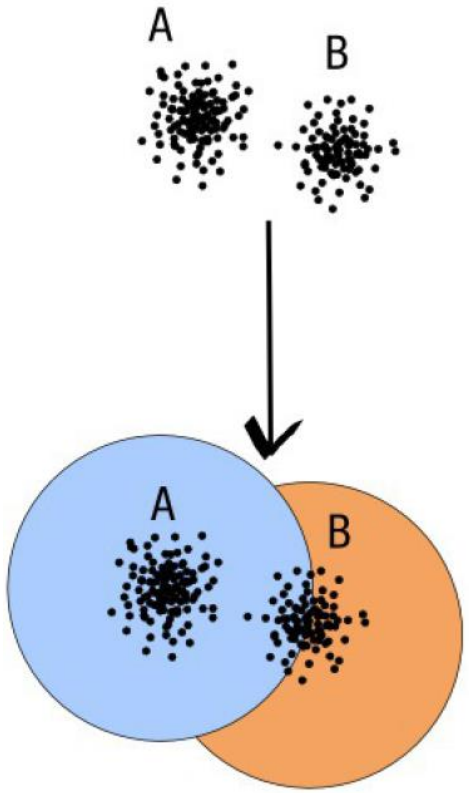


# Input order dependent results

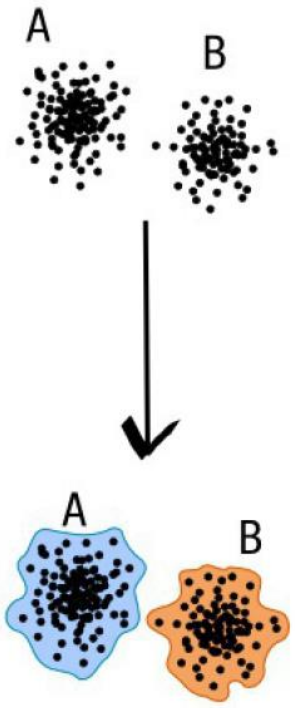



decreasing length,
decreasing abundance,
external references

# Single a priori clustering threshold

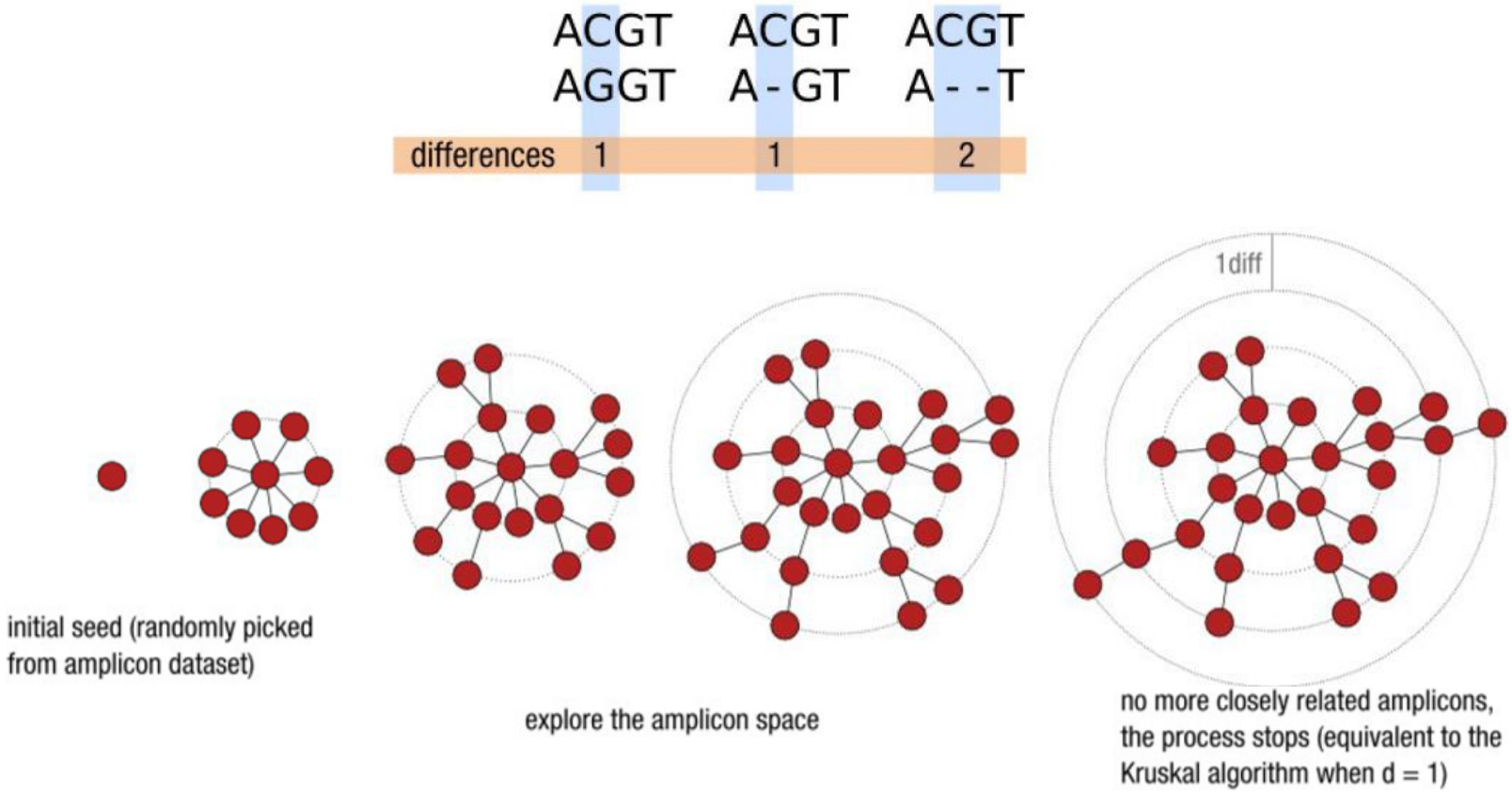



compromise threshold
unadapted threshold

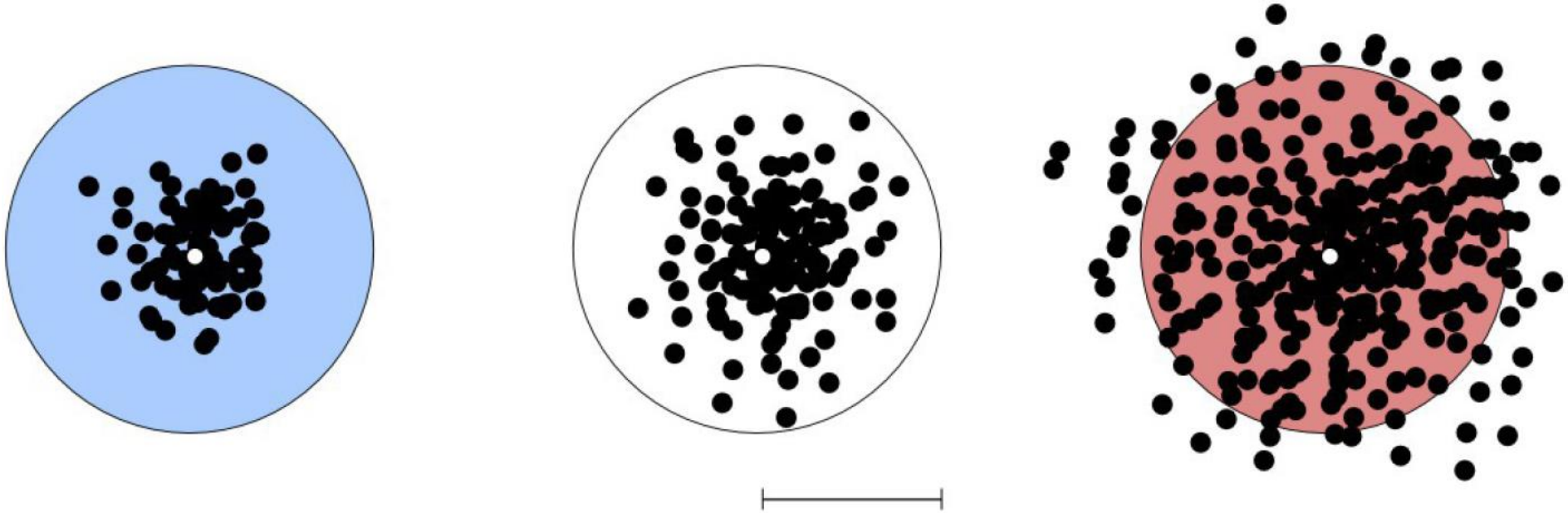



natural limits of clusters

# Swarm clustering method

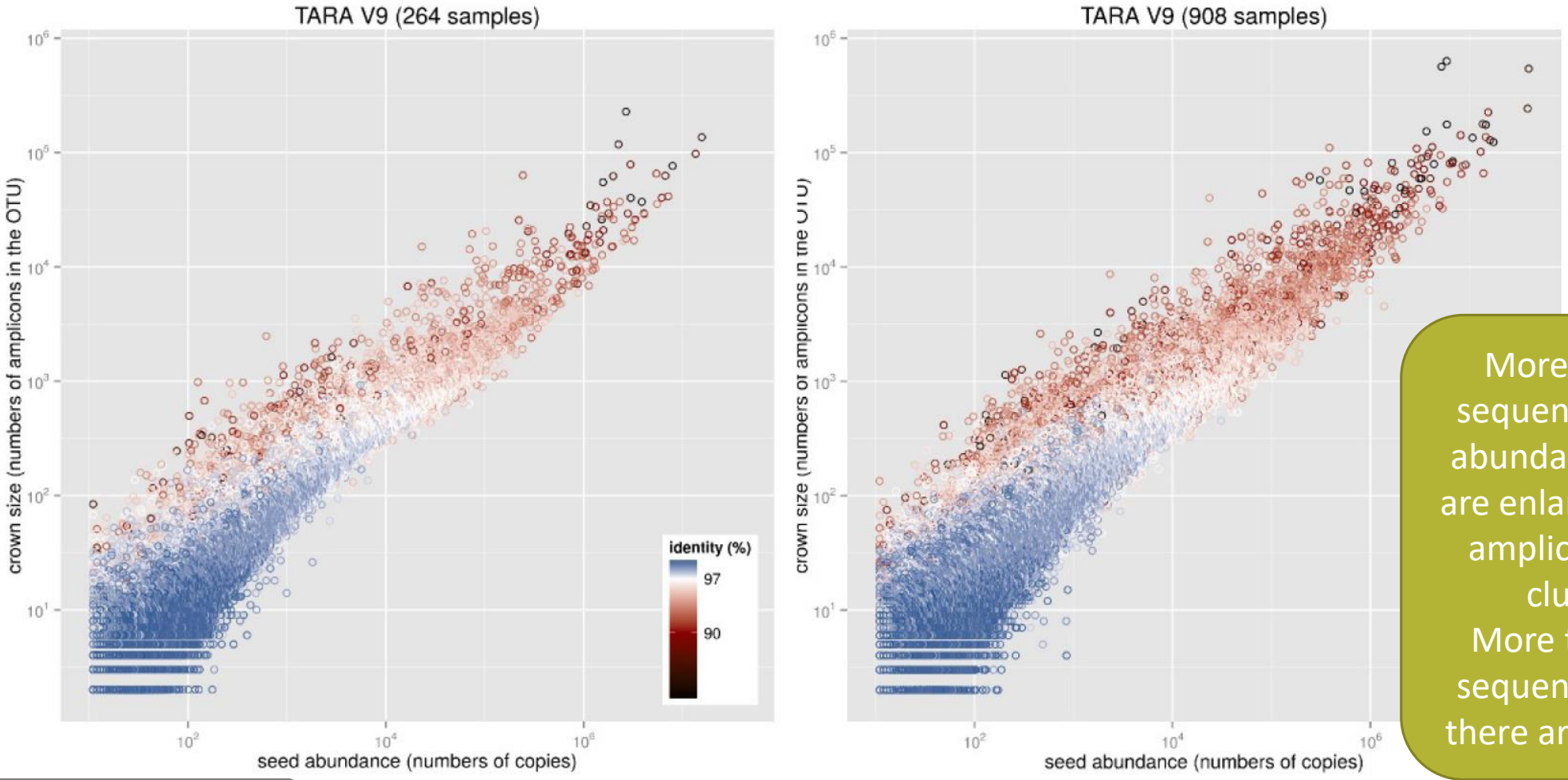


# Comparison Swarm and 3% clusterings




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

# Comparison Swarm and 3% clusterings




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

clusters produced with swarm using d = 1

# SWARM

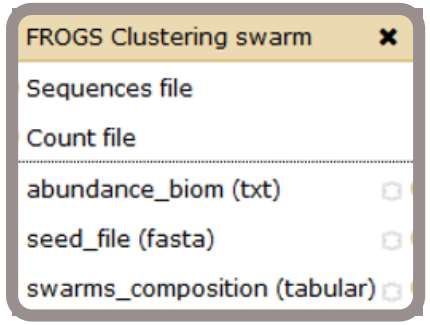
A robust and fast clustering method for amplicon-based studies.

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




**Clustering**

**FROGS Clustering swarm** Step 2 in metagenomics analysis : clustering. (Galaxy Version 2.3.0) — Options

**Sequences file**

2: FROGS Pre-process: dereplicated.fasta

The sequences file (format: fasta).

**Count file**

3: FROGS Pre-process: count.tsv

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

**Aggregation distance**

3

Maximum number of differences between sequences in each aggregation step.

**Performe denoising clustering step?**

Yes | No

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

Execute

<u>1st run for denoising:</u>
Swarm with d = 1 -> high clusters definition
linear complexity

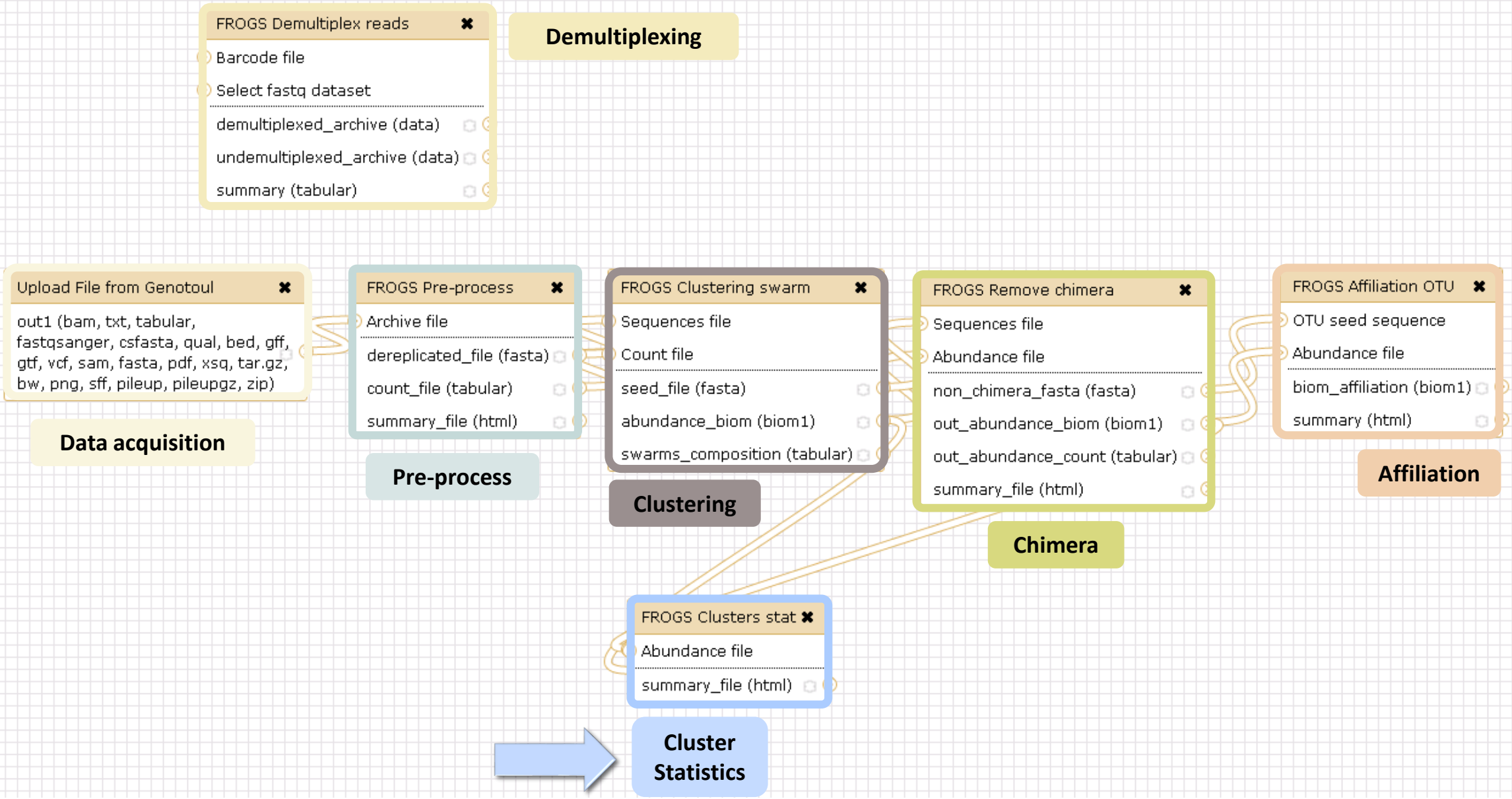
<u>2nd run for clustering:</u>
Swarm with d = 3 on the seeds of first Swarm
quadratic complexity

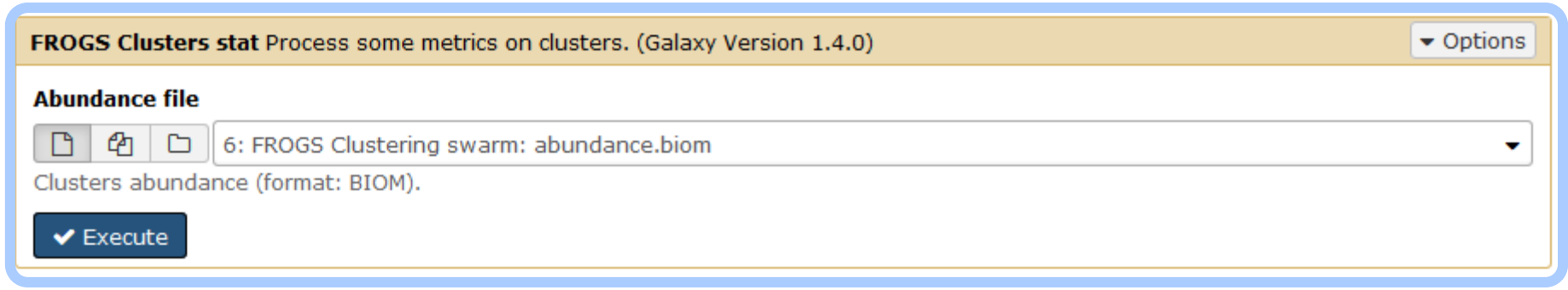
Gain time !

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


FROGS Clusters stat Process some metrics on clusters. (Galaxy Version 1.4.0)
Options
Abundance file
6: FROGS Clustering swarm: abundance.biom
Clusters abundance (format: BIOM).
Execute

# Your Turn! - 3

LAUNCH CLUSTERING AND CLUSTERSTAT TOOLS

Miseq merged

# Exercise 3

Go to « MiSeq merged » history

Launch the Clustering SWARM tool on that data set with aggregation distance = 3 and the denoising

→ objectives :

- understand the denoising efficiency
- understand the ClusterStat utility

Miseq merged

# Exercise 3

1. How much time does it take to finish?
2. How many clusters do you get ?

Miseq merged

# Exercise 3

3. Launch FROGS Cluster Stat tools on the previous abundance biom file

FROGS Clusters stat Process some metrics on clusters.

Miseq merged

# Exercise 3

4. Interpret the boxplot: **Clusters size summary**
5. Interpret the table: **Clusters size details**
6. What can we say by observing the **sequence distribution**?
7. How many clusters share “sampleB3” with at least one other sample?
8. How many clusters could we expect to be shared ?
9. How many sequences represent the 550 specific clusters of “sampleC2”?
10. This represents what proportion of “sampleC2”?
11. What do you think about it?
12. How do you interpret the « Hierarchical clustering » ?

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

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Help | User | Using 29%

Tools

matrix

**FROGS - Find Rapidly Otu with Galaxy Solution**

OTUS RECONSTRUCTION

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

FROGS Pre-process merging, denoising and dereplication.

FROGS Clustering swarm amplicon sequence clustering.

FROGS Remove chimera Remove PCR chimera in each sample.

FROGS Filters Filters OTUs on several criteria.

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

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

FROGS Clusters stat Process some metrics on clusters.

FROGS Affiliations stat Process some metrics on taxonomies.

FROGS Affiliation postprocess Optionnal step to resolve inclusive amplicon ambiguities and to aggregate OTUs based on alignment metrics

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

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

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

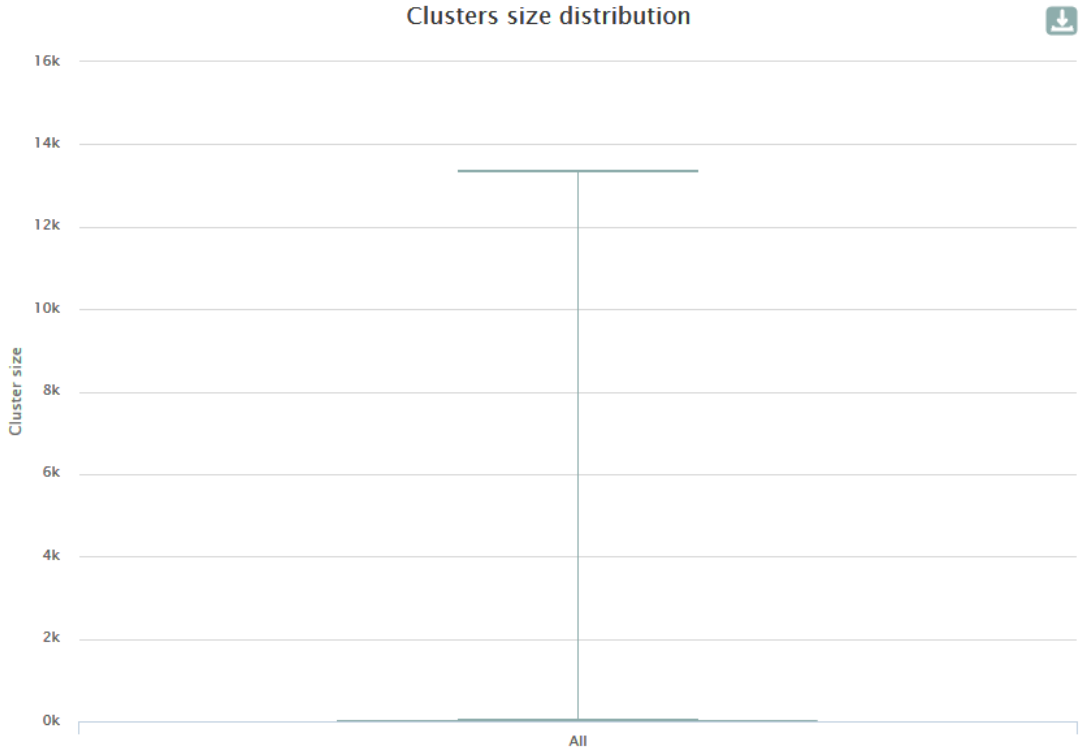
Clusters distribution | Sequences distribution | Samples distribution

Clusters 5,940

Sequences 89,739

Most of clusters are singletons

## Clusters size summary




| 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 |

History

chimera: report.html

13: FROGS Remove chimera: non_chimera_abundance.biom

12: FROGS Remove chimera: non_chimera.fasta

11: FROGS Clusters stat: summary.html

186.7 KB

format: **html**, database: ?

## Application
Software :/galaxydata/galaxy-prod/my_tools/FROGS /app/clusters_stat.py (version : r3.0-3.0)
Command : /galaxydata /galaxy-prod/my_tools/FROGS /app/clusters_stat.py --input-biom /galaxydata/galaxy-prod/my_files/000/330 /dataset_330065.dat --out

HTML file

7: FROGS Clustering swarm: swarms_composition.tsv

6: FROGS Clustering swarm: abundance.biom

5: FROGS Clustering swarm: seed_sequences.fasta

4: FROGS Pre-process: report.html

3: FROGS Pre-process: count.tsv

2: FROGS Pre-process: dereplicated.fasta

1: /work/project /frogs/Formation

# Clusters size details

Most of clusters are singletons

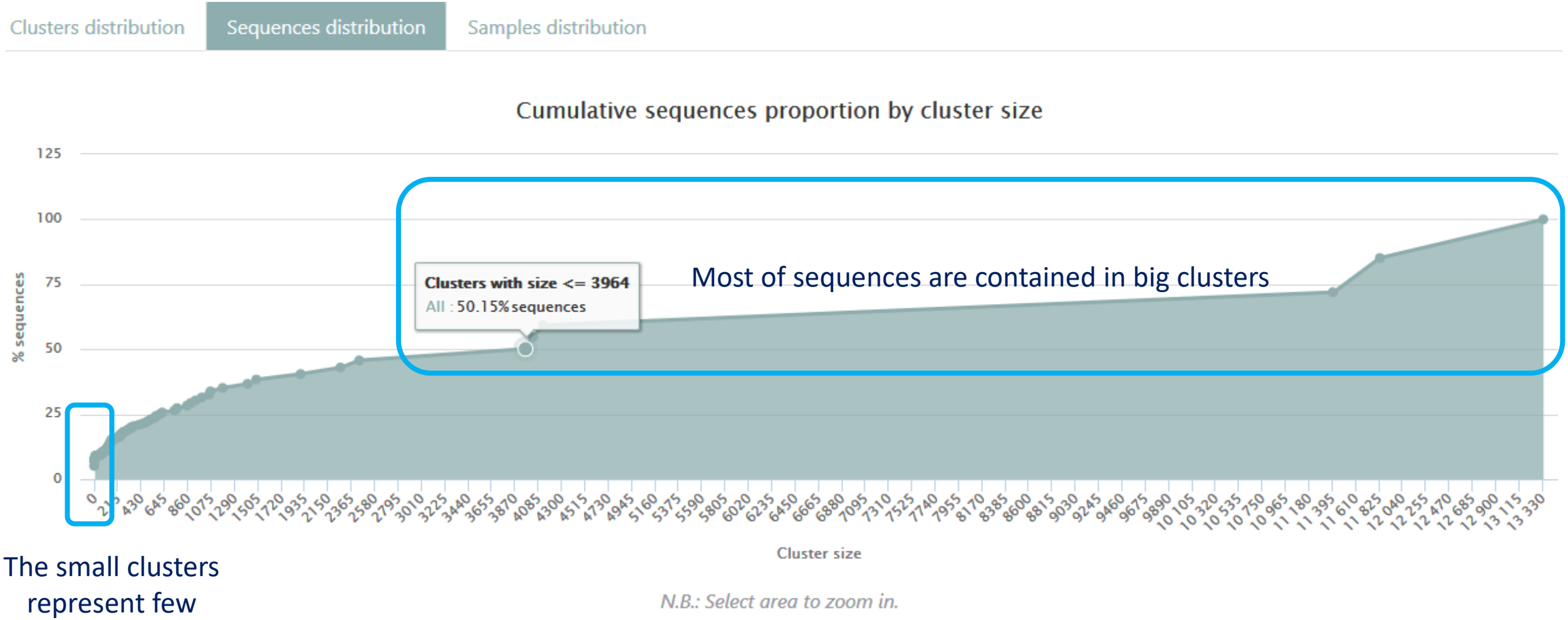
CSV

Show 10 entries

Search:

| Cluster size | Number of cluster | % of all clusters |
|---|---|---|
| 1 | 4,595 | 77.36 |
| 2 | 865 | 14.56 |
| 3 | 154 | 2.59 |
| 4 | 84 | 1.41 |
| 5 | 42 | 0.71 |
| 6 | 29 | 0.49 |
| 7 | 23 | 0.39 |
| 8 | 13 | 0.22 |
| 9 | 6 | 0.10 |
| 10 | 6 | 0.10 |

After clustering




The small clusters represent few sequences

## Sequences count

368 clusters of sampleA1 are common at least once with another sample

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

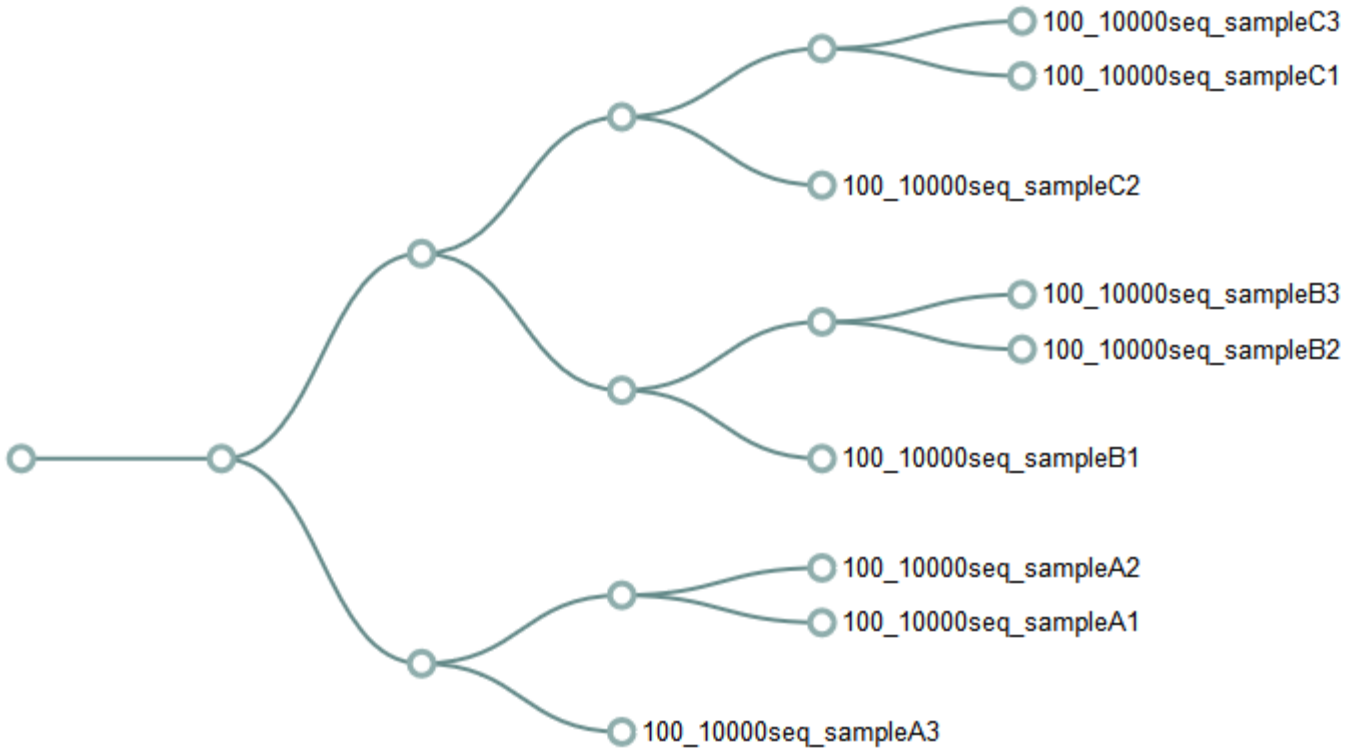
Show 10 entries

| Sample | Total clusters | Shared clusters | Own clusters | Total sequences | Shared sequences | Own sequences |
|---|---|---|---|---|---|---|
| 100_10000seq_sampleA1 | 881 | 368 | 513 | 9,975 | 9,447 | 528 |
| 100_10000seq_sampleA2 | 856 | 366 | 490 | 9,979 | 9,476 | 503 |
| 100_10000seq_sampleA3 | 867 | 384 | 483 | 9,972 | 9,478 | 494 |
| 100_10000seq_sampleB1 | 942 | 394 | 548 | 9,969 | 9,397 | 572 |
| 100_10000seq_sampleB2 | 881 | 373 | 508 | 9,970 | 9,455 | 515 |
| 100_10000seq_sampleB3 | 941 | 379 | 562 | 9,967 | 9,388 | 579 |
| 100_10000seq_sampleC1 | 910 | 371 | 539 | 9,965 | 9,413 | 552 |
| 100_10000seq_sampleC2 | 938 | 388 | 550 | 9,975 | 9,408 | 567 |
| 100_10000seq_sampleC3 | 878 | 362 | 516 | 9,967 | 9,442 | 525 |

Showing 1 to 9 of 9 entries

Previous 1 Next

# Hierarchical clustering




Hierarchical classification on Bray Curtis distance

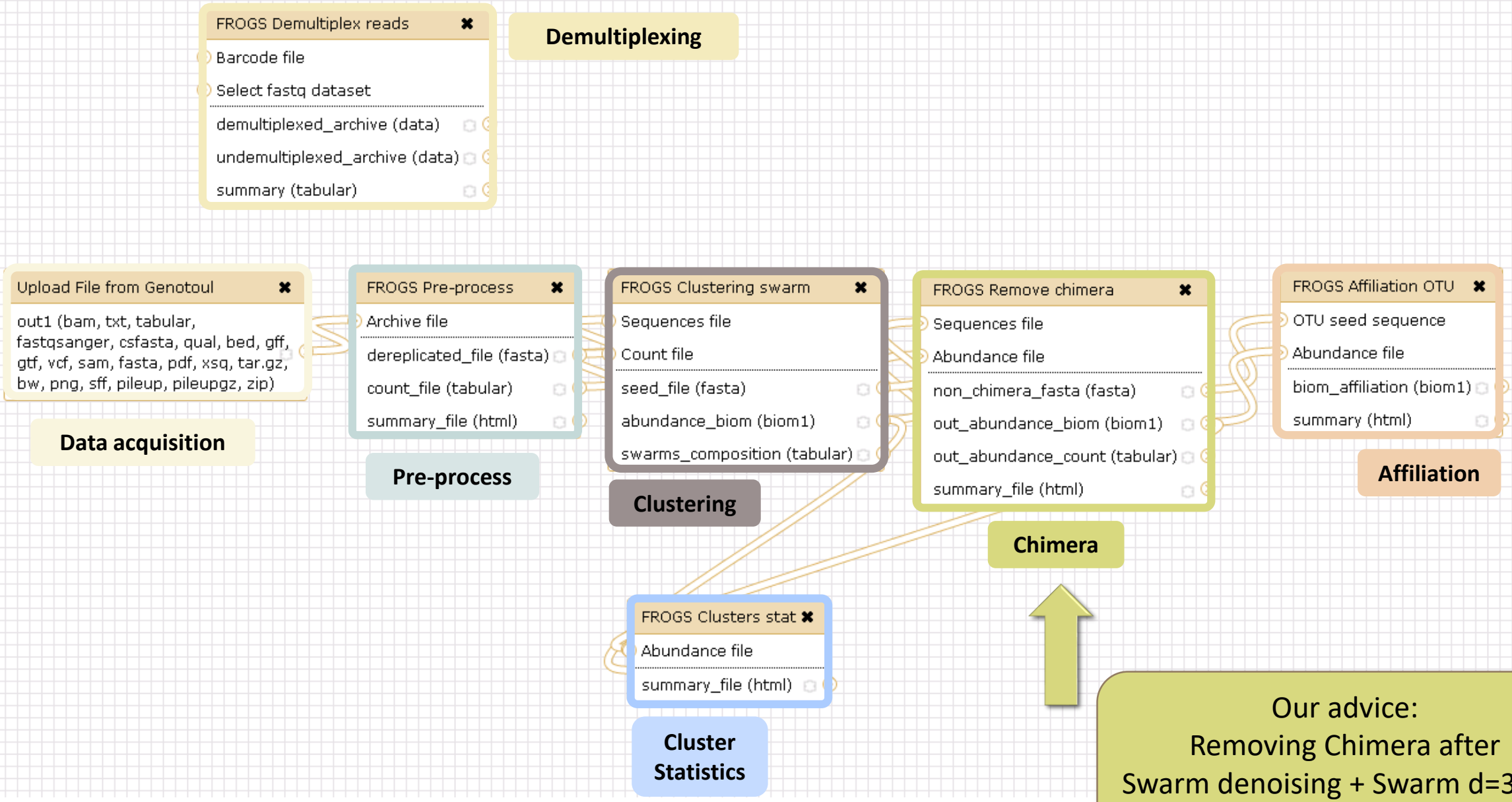
Newick tree available too

Newick

(((100_10000seq_sampleA3,(100_10000seq_sampleA1,100_10000seq_sampleA2):0.096):0.100,((100_10000seq_sampleB1,
(100_10000seq_sampleB2,100_10000seq_sampleB3):0.101):0.102,(100_10000seq_sampleC2,(100_10000seq_sampleC1,100_10000seq_sampleC3):0.098):0.105):0.830):0.883);

Samples distribution tab

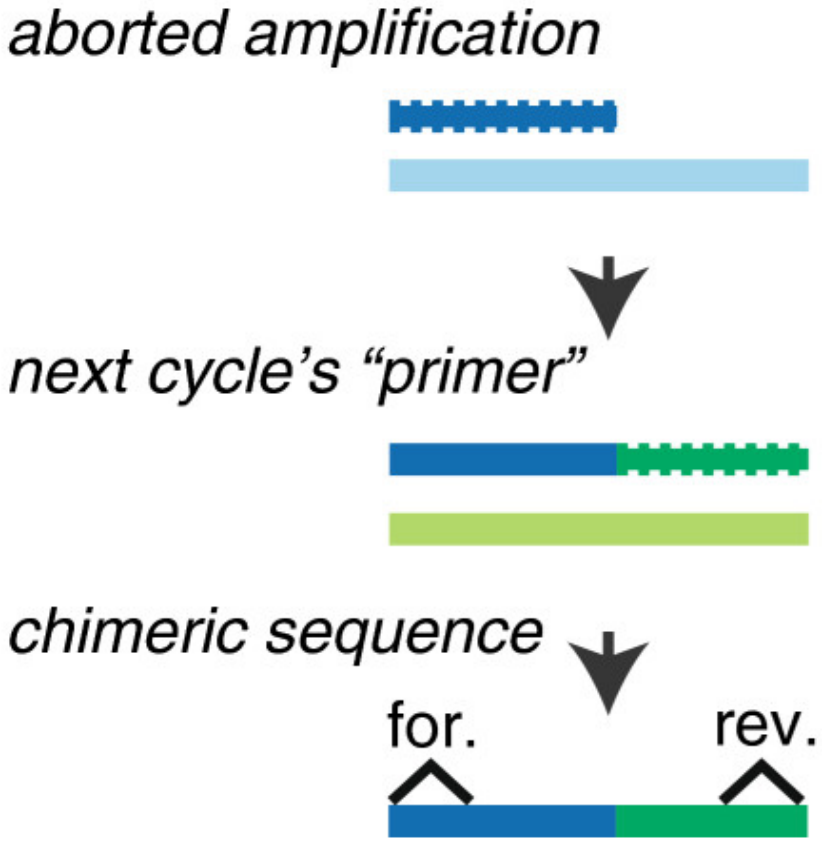
# Chimera removal tool


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.

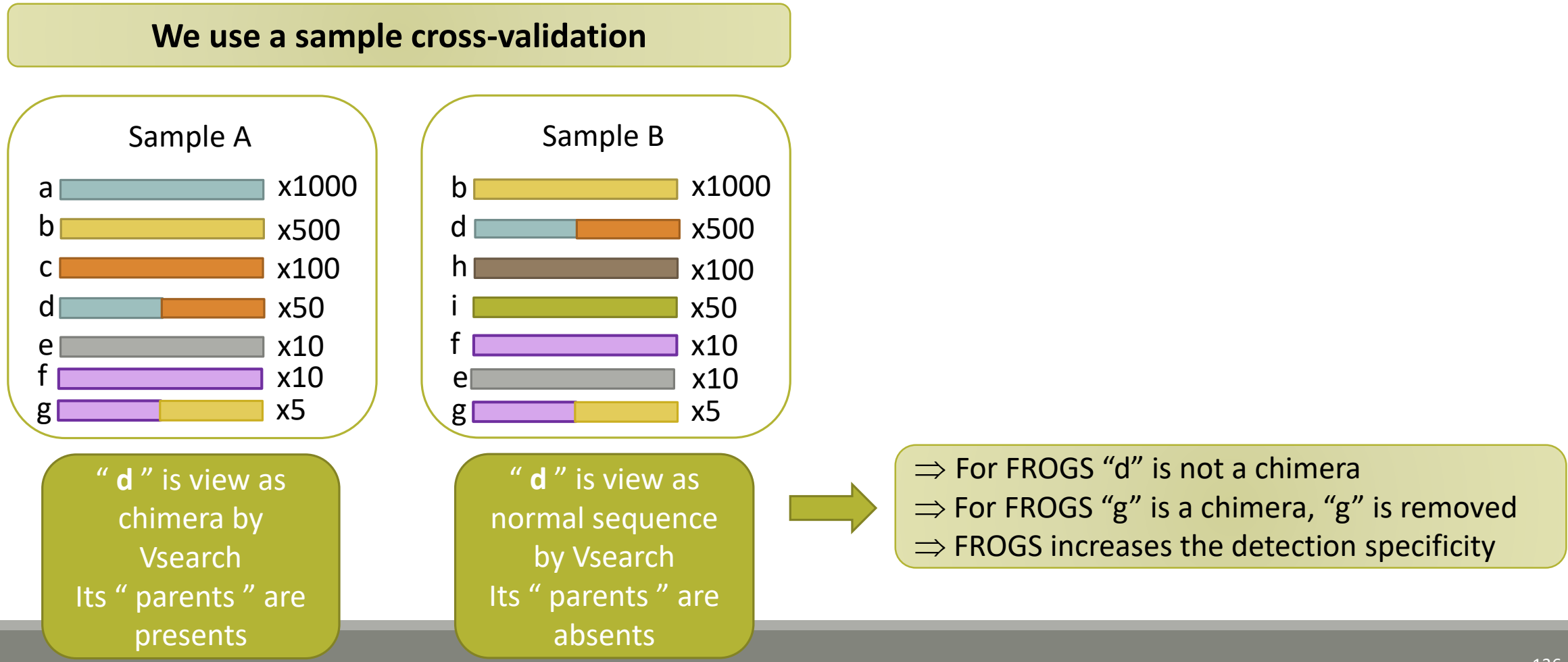
**Chimera: from 5 to 45% of reads** (Schloss 2011)




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**



# Your Turn! - 4

LAUNCH THE REMOVE CHIMERA TOOL

Miseq merged

# Exercise 5

Go to « MiSeq merged » history

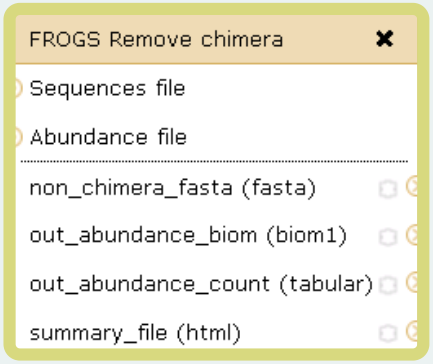
Launch the « FROGS Remove Chimera » tool

Follow by the « FROGS ClusterStat » tool on the swarm d1d3 non chimera abundance biom

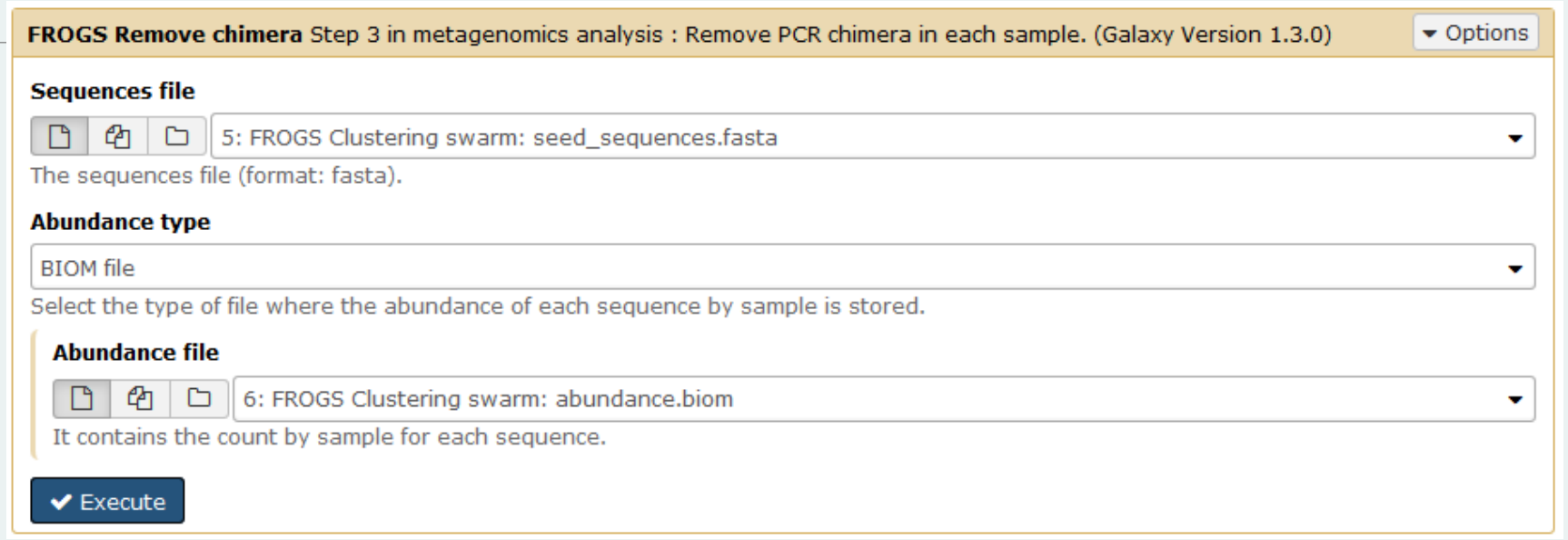
→ objectives :

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

Miseq merged




Chimera



Miseq merged

# Exercise 4

1. Understand the « FROGS remove chimera : report.html»
   a. How many clusters are kept after chimera removal?
   b. How many sequences that represent ? So what abundance?
   c. What do you conclude ?

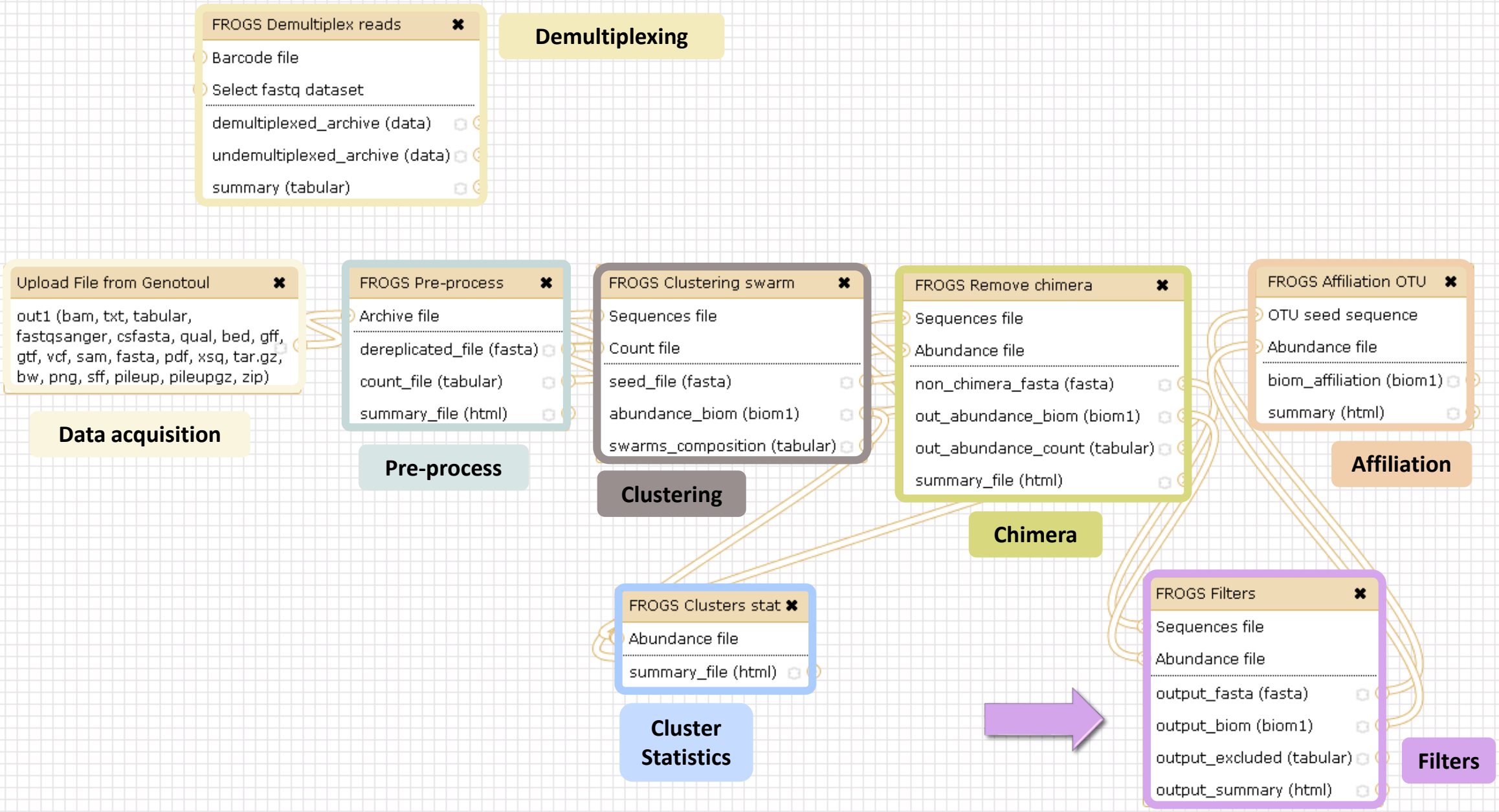
Miseq merged

# Exercise 4

2. Launch « FROGS ClusterStat » tool on non_chimera_abundance.biom
3. Rename output in summary_nonchimera.html
4. Compare the HTML files
   a. Of what are mainly composed singleton ? (compare with previous summary.html)
   b. What are their abundance?
   c. What do you conclude ?

The weakly abundant Clusters are mainly false positives, our data would be much more exact if we remove them

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

Affiliation runs long time

Advise:

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

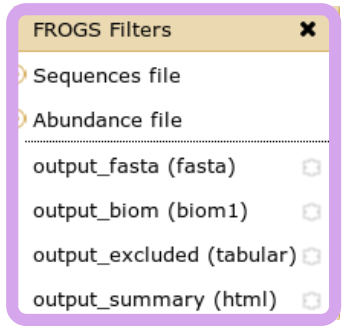
You will gain time !

# Filters

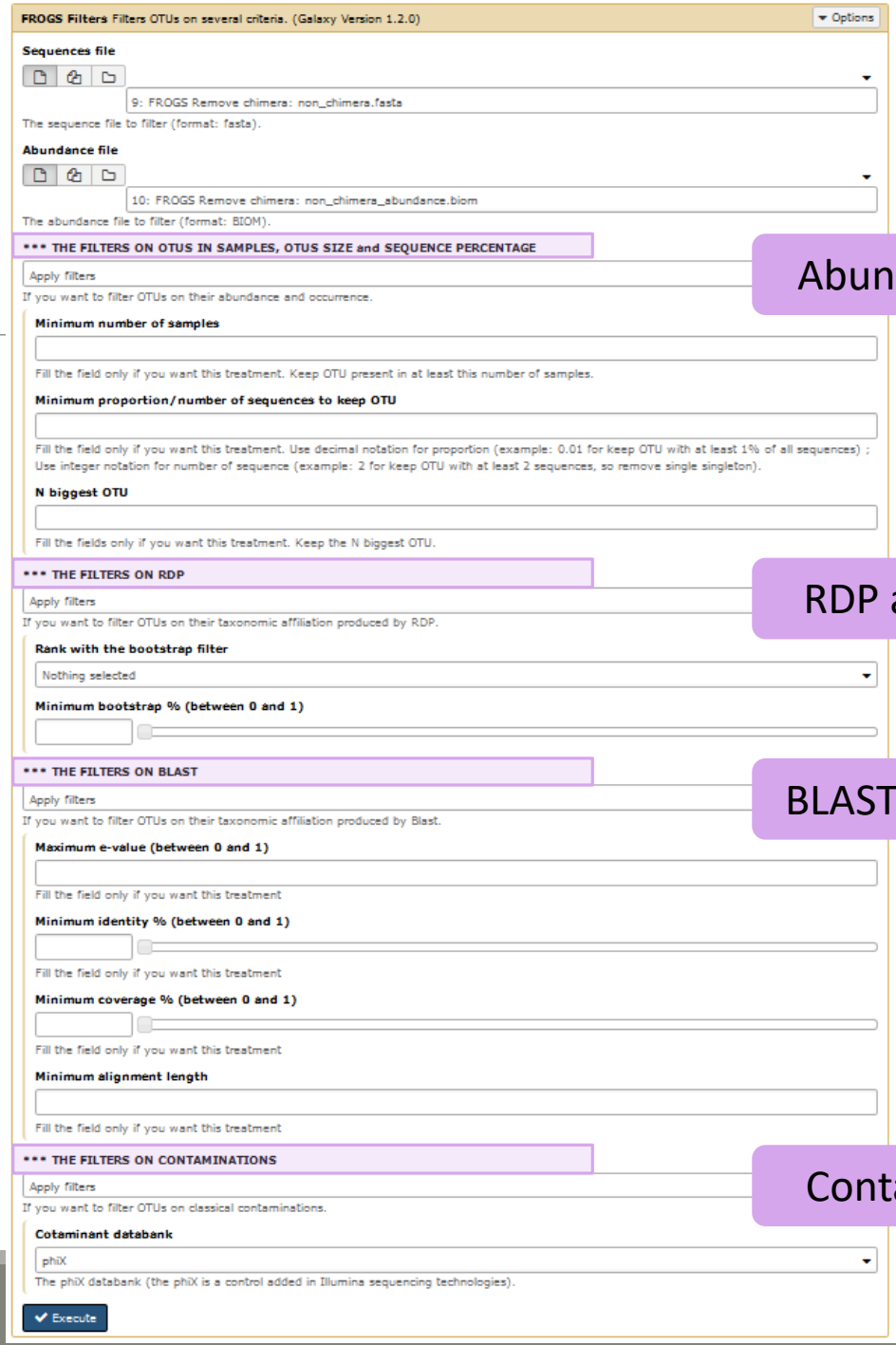
Filters allows to filter the result thanks to different criteria et may be used after different steps of pipeline :

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

**After Affiliation tool**




**Filters**




Abundance filters

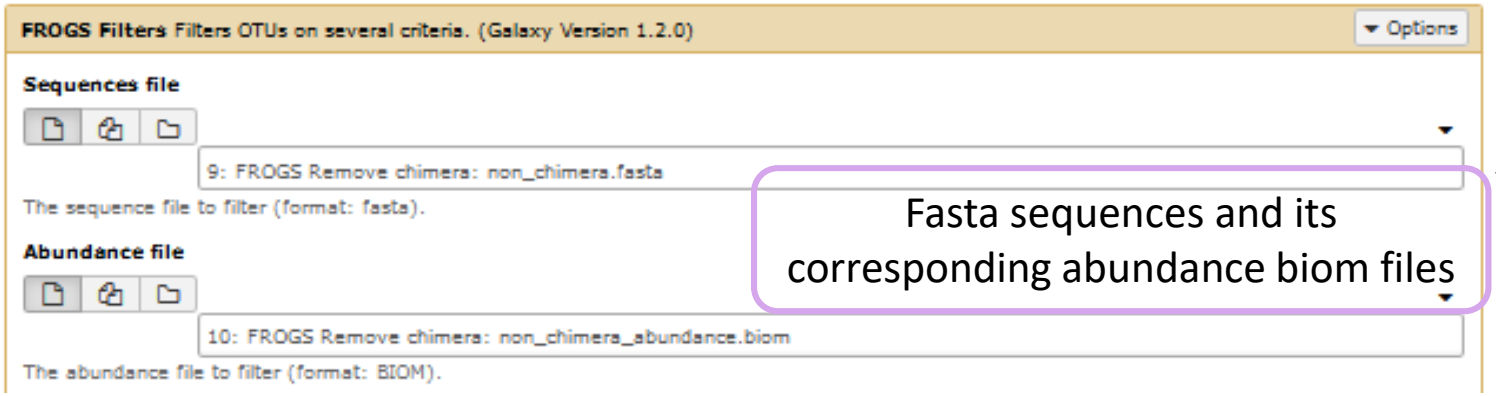
RDP affiliation filters

BLAST affiliation filters

Contamination filter

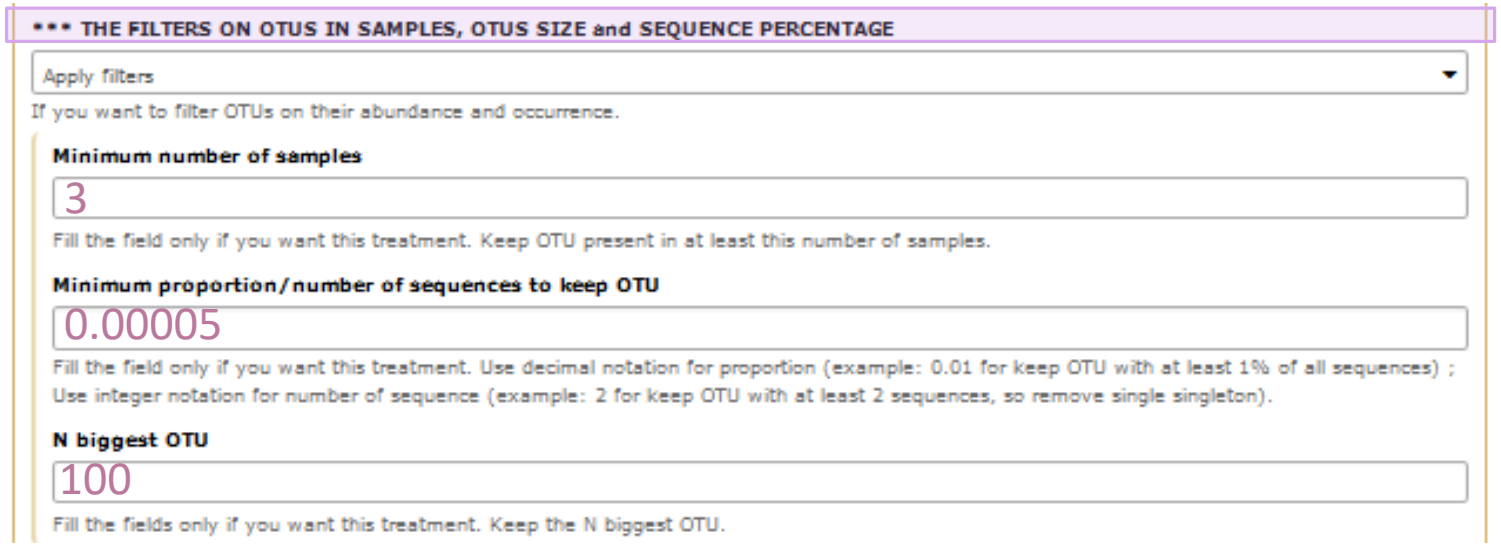
4 filter sections

Input



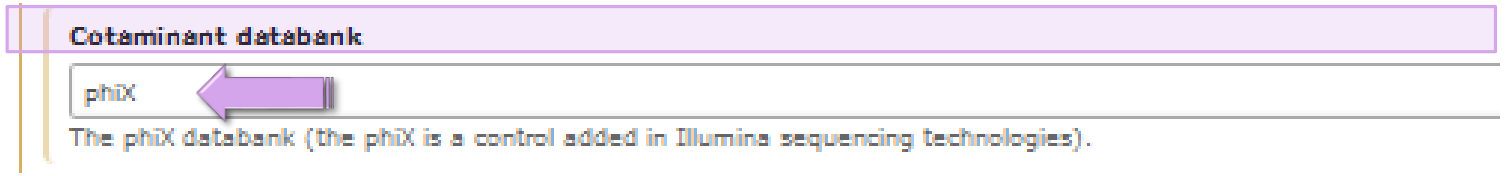

Fasta sequences and its corresponding abundance biom files

Filter 1 : abundance




Filter 2 & 3: affiliation

## Filter 4 : contamination




Soon, several contaminant banks

# Your Turn! - 5

LAUNCH THE « FILTERS » TOOL

Miseq merged

# Exercise 5

Go to history « MiSeq merged »

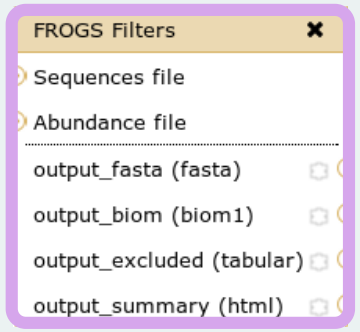
Launch « Filters » tool with non_chimera_abundance.biom, non_chimera.fasta

Apply 2 filters :

- Minimum proportion/number of sequences to keep OTU: 0.00005*
- Minimum number of samples: 3

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

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




Filters

**FROGS Filters** Filters OTUs on several criteria. (Galaxy Version 1.2.0) — Options

**Sequences file**

9: FROGS Remove chimera: non_chimera.fasta

The sequence file to filter (format: fasta).

**Abundance file**

10: FROGS Remove chimera: non_chimera_abundance.biom

The abundance file to filter (format: BIOM).

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

Apply filters

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

**Minimum number of samples**

3

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

**Minimum proportion/number of sequences to keep OTU**

0.00005

Fill the field only if you want this treatment. Use decimal notation for proportion (example: 0.01 for keep OTU with at least 1% of all sequences) ; Use integer notation for number of sequence (example: 2 for keep OTU with at least 2 sequences, so remove single singleton).

**N biggest OTU**

Fill the fields only if you want this treatment. Keep the N biggest OTU.

***** THE FILTERS ON RDP**

No filters

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

***** THE FILTERS ON BLAST**

No filters

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

***** THE FILTERS ON CONTAMINATIONS**

No filters

If you want to filter OTUs on classical contaminations.

Execute

Output

**92: FROGS Filters: report.html**

**91: FROGS Filters: excluded.tsv**

**90: FROGS Filters: abundance.biom**

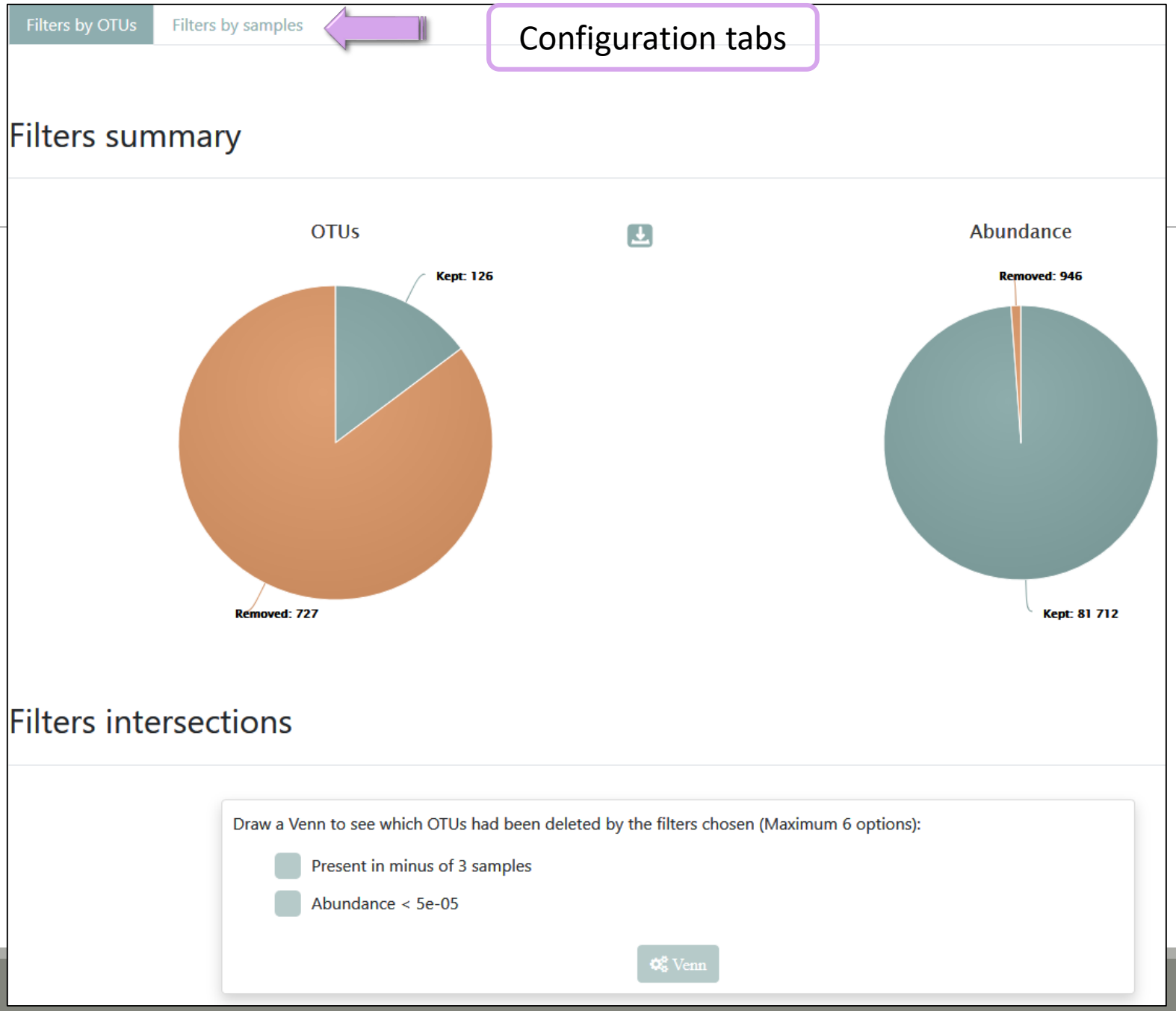
**89: FROGS Filters: sequences.fasta**

If Filters fields are « Apply » so you have to fill at one field. Otherwise, galaxy become red !

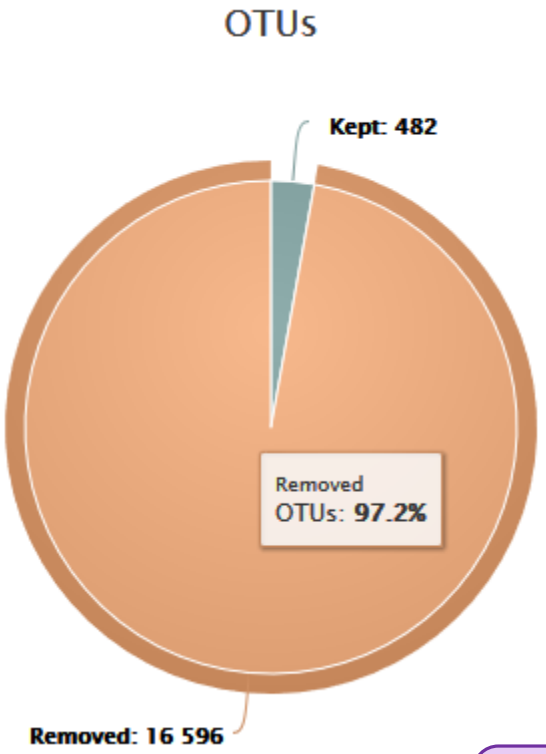
Miseq merged

# Exercise 5

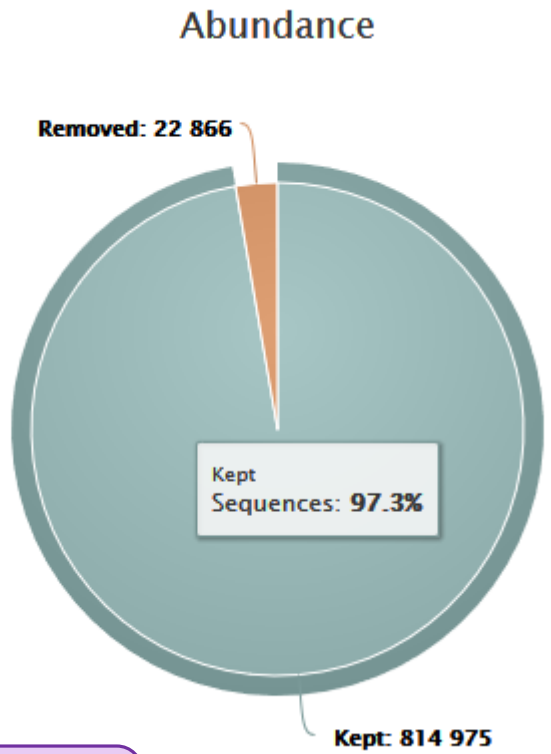
1. What are the output files of “Filters” ?
2. Explore “FROGS Filter : report.html” file.
3. How many OTUs have you removed ?
4. Build the Venn diagram on the two filters.
5. How many OTUs have you removed with each filter “abundance > 0.005% ”, “Remove OTUs that are not present at least in 3 samples”?
6. How many OTUs do they remain ?
7. Is there a sample more impacted than the others ?
8. To characterize these new OTUs, do not forget to launch “FROGS Cluster Stat” tool, and rename the output HTML file.


Filters by OTUs
Filters by samples
Configuration tabs
Filters summary
OTUs
Kept: 126
Removed: 727
Abundance
Removed: 946
Kept: 81 712
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
Venn

# What are the remaining singletons ?

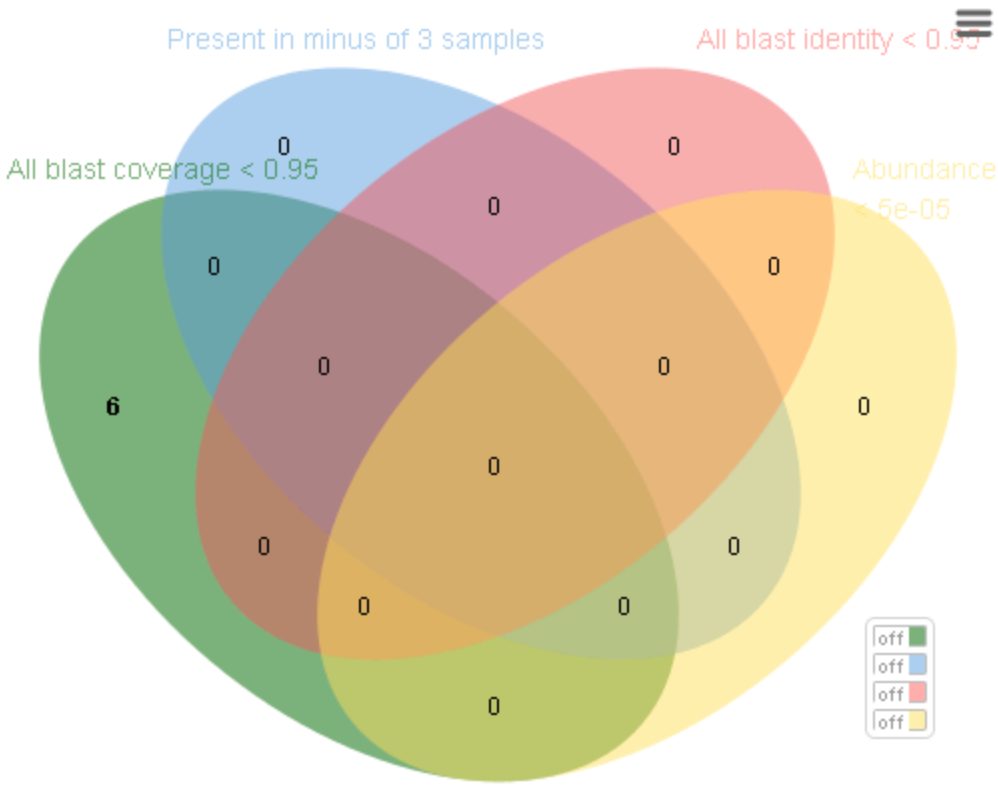



On simulated data, singleton are:
~99,9% are chimera
and
~0,1% are sequences with sequencing errors, non clustered



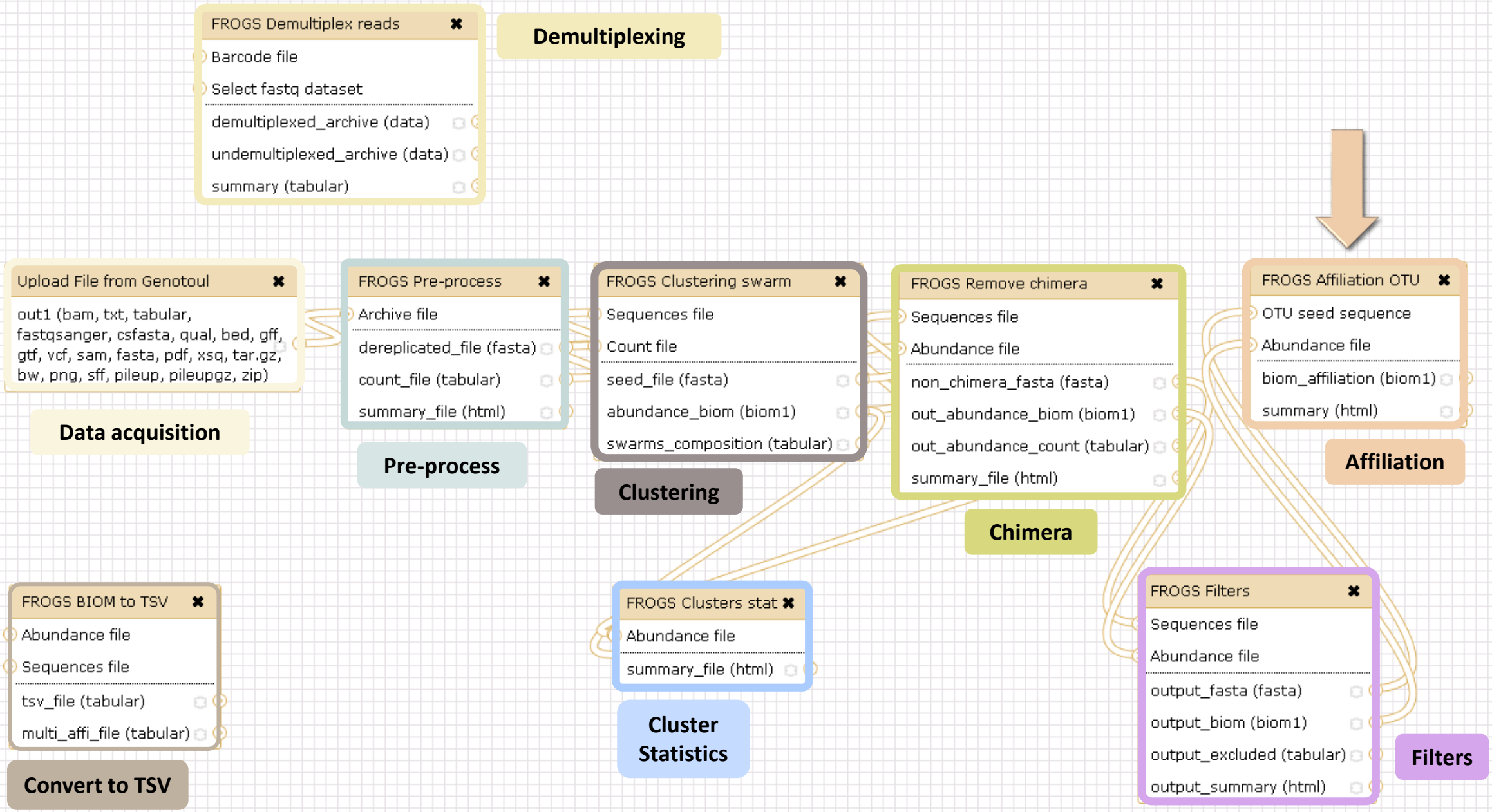

Removing little OTUs (conservation rate =0.005%) and non shared OTU (in less than 2 samples)

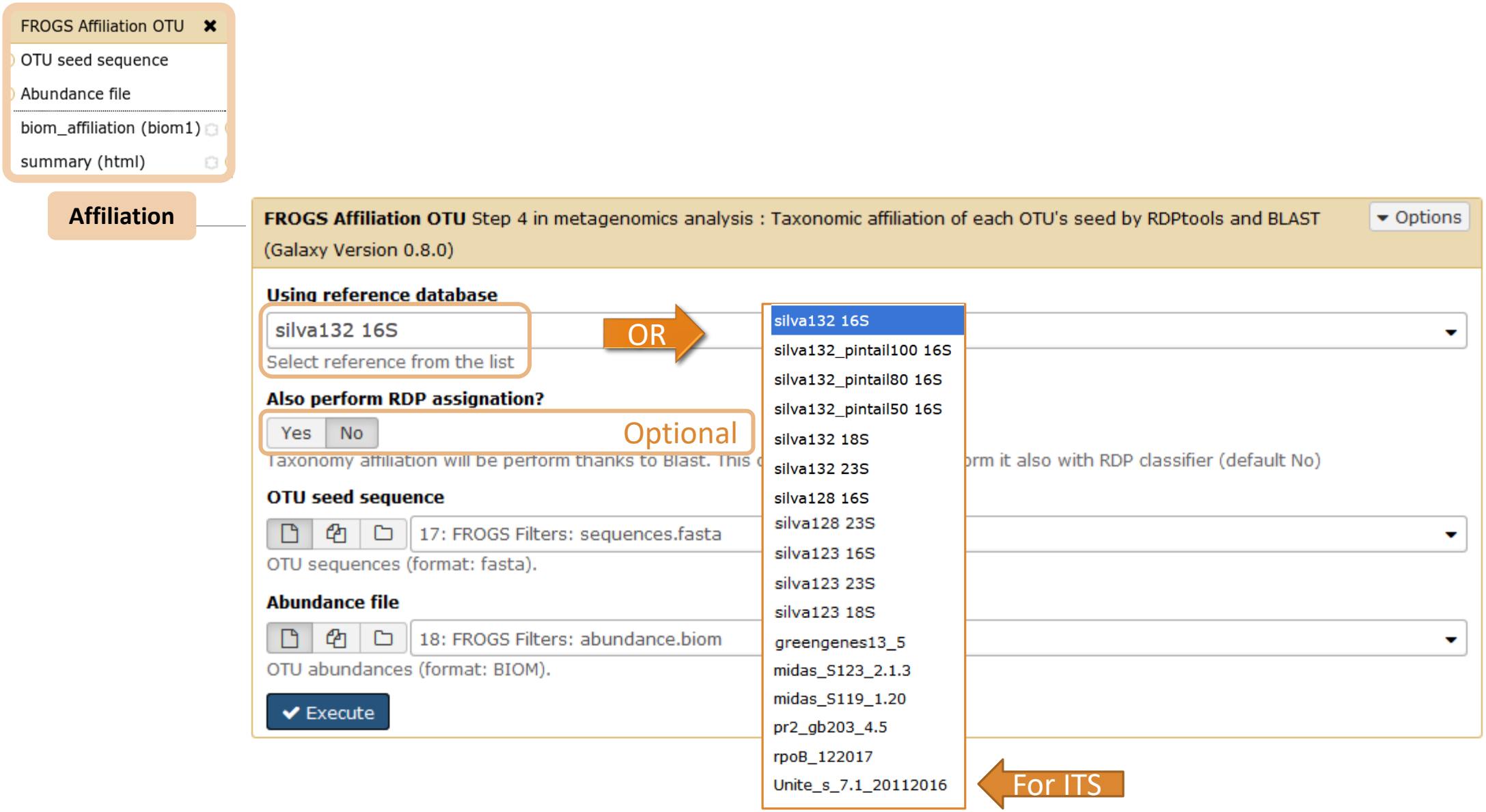
Venn on removed OTUs




Close

# Affiliation 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 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 Affiliation OTU
OTU seed sequence
Abundance file
biom_affiliation (biom1)
summary (html)
Affiliation
FROGS Affiliation OTU Step 4 in metagenomics analysis : Taxonomic affiliation of each OTU's seed by RDPtools and BLAST
(Galaxy Version 0.8.0)
Options
Using reference database
silva132 16S
Select reference from the list
OR
silva132 16S
silva132_pintail100 16S
silva132_pintail80 16S
silva132_pintail50 16S
silva132 18S
silva132 23S
silva128 16S
silva128 23S
silva123 16S
silva123 23S
silva123 18S
greengenes13_5
midas_S123_2.1.3
midas_S119_1.20
pr2_gb203_4.5
rpoB_122017
Unite_s_7.1_20112016
For ITS
Also perform RDP assignation?
Yes
No
Optional
Taxonomy affiliation will be perform thanks to Blast. This
rm it also with RDP classifier (default No)
OTU seed sequence
17: FROGS Filters: sequences.fasta
OTU sequences (format: fasta).
Abundance file
18: FROGS Filters: abundance.biom
OTU abundances (format: BIOM).
Execute

# 1 Cluster = 2 affiliations

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! – 6

LAUNCH THE « FROGS AFFILIATION » TOOL

Miseq merged

# Exercise 6.1

Go to « MiSeq merged » history

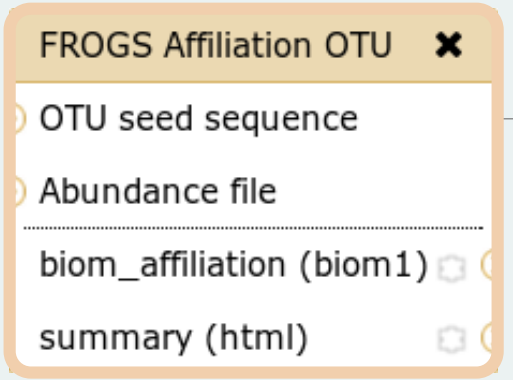
Launch the « FROGS Affiliation » tool with

- SILVA 123 or 128 or 132 16S database
- FROGS Filters abundance biom and fasta files (after swarm d1+d3, remove chimera and filter low abundances)

→ objectives :

- understand abundance tables columns
- understand the BLAST affiliation

Miseq merged




**Affiliation**

**FROGS Affiliation OTU** Step 4 in metagenomics analysis : Taxonomic affiliation of each OTU's seed by RDPtools and BLAST (Galaxy Version 0.8.0)

Options

**Using reference database**

silva123 16S

Select reference from the list

**Also perform RDP assignation?**

Yes | No

Taxonomy affiliation will be perform thanks to Blast. This option allow you to perform it also with RDP classifier (default No)

**OTU seed sequence**

17: FROGS Filters: sequences.fasta

OTU sequences (format: fasta).

**Abundance file**

18: FROGS Filters: abundance.biom

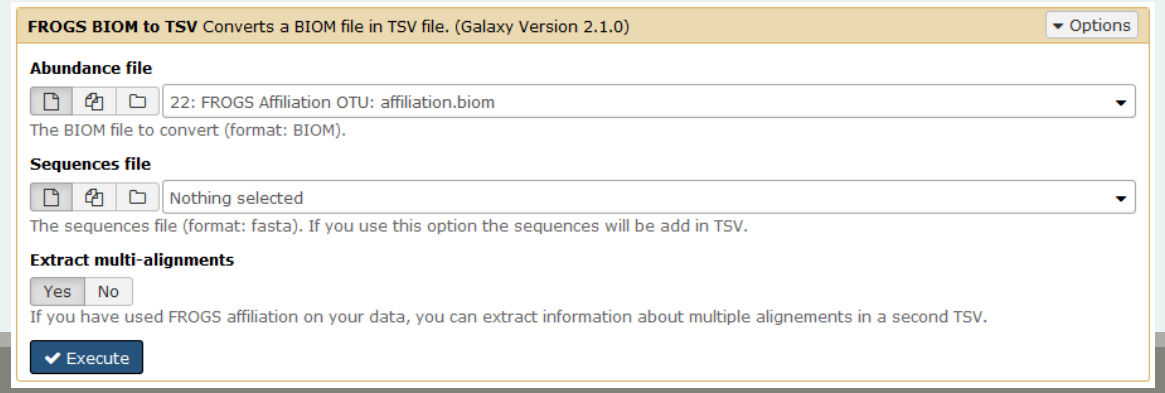
OTU abundances (format: BIOM).

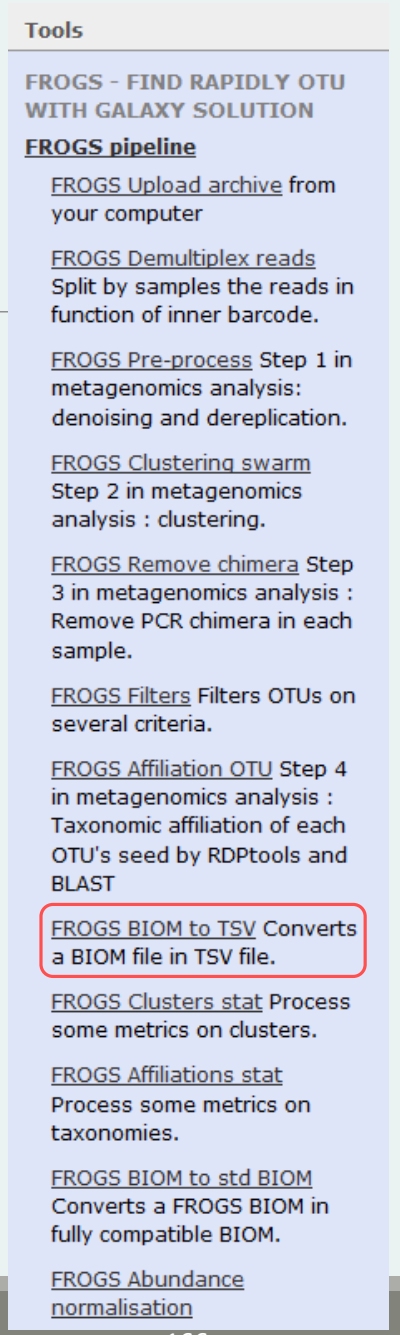
Execute

Miseq merged

# Exercise 6.1

1. What are the « FROGS Affiliation » output files ?
2. How many sequences are affiliated by BLAST ?
3. Click on the « eye » button on the BIOM output file, what do you understand ?
4. Use the Biom_to_TSV tool on this last file and click again on the ”eye” on the new output generated.
   What do the columns ?
   What is the difference if we click on case or not ? What consequence about weight of your file ?



Miseq merged

# Exercise 6.1

5. Understand Blast affiliations - Cluster_2388 (affiliation from silva 123)

| blast_subject | blast_evalue | blast_len | blast_perc_query_coverage | blast_perc_identity | blast_taxonomy |
|---|---|---|---|---|---|
| JN880417.1.1422 | 0.0 | 360 | 88.88 | 99.44 | Bacteria;Planctomycetes;Planctomycetacia;Planctomycetales;Planctomycetaceae;Telmatocola;Telmatocola sphagniphila |

# Blast JN880417.1.1422 vs our OTU

OTU length : 405

Excellent blast but no matches at the beginning of OTU.

Telmatocola sphagniphila strain SP2 16S ribosomal RNA gene, partial sequence

Sequence ID: ref|NR_118328.1 Length: 1422 Number of Matches: 1

Range 1: 375 to 734 GenBank Graphics ▼ Next Match ▲ Previous Match

| Score | Expect | Identities | Gaps | Strand |
|---|---|---|---|---|
| 654 bits(354) | 0.0 | 358/360(99%) | 0/360(0%) | Plus/Plus |

```
Query  46   CGCGTGCGCGATGAAGGCCTTCGGGTTGTAAAGCGCGAAAGAGGTAATAAAGGGAAACCT  105
Sbjct  375  CGCGTGCGCGATGAAGGCCTTCGGGTTGTAAAGCGCGAAAGAGGSAATAAAGGGAAACTT  434

Query  106  GATTGAACCTCAGTAAGCTCGGGCTAAGTTTGTGCCAGCAGCCGCGGTAAGACGAACCGA  165
Sbjct  435  GATTGAACCTCAGTAAGCTCGGGCTAAGTTTGTGCCAGCAGCCGCGGTAAGACGAACCGA  494

Query  166  GCGAACGTTGTTCGGAATCACTGGGCATAAAGGGCGCGTAGGCGGGTTTCTAAGTCCGTG  225
Sbjct  495  GCGAACGTTGTTCGGAATCACTGGGCATAAAGGGCGCGTAGGCGGGTTTCTAAGTCCGTG  554

Query  226  GTGAAATACTTCAGCTCAACTGGAGAACTGCCTCGGATACTGGGAATCTCGAGTAATGTA  285
Sbjct  555  GTGAAATACTTCAGCTCAACTGGAGAACTGCCTCGGATACTGGGAATCTCGAGTAATGTA  614

Query  286  GGGGCACGTGGAACGGCTGGTGGAGCGGTGAAATGCGTTGATATCAGTCGGAACTCCGGT  345
Sbjct  615  GGGGCACGTGGAACGGCTGGTGGAGCGGTGAAATGCGTTGATATCAGTCGGAACTCCGGT  674

Query  346  GGCGAAGGCGATGTGCTGGACATTTACTGACGCTGAGGCGCGAAAGCCAGGGGAGCAAAC  405
Sbjct  675  GGCGAAGGCGATGTGCTGGACATTTACTGACGCTGAGGCGCGAAAGCCAGGGGAGCAAAC  734
```

**Telmatocola sphagniphila strain SP2 16S ribosomal RNA gene, partial sequence**

NCBI Reference Sequence: NR_118328.1

FASTA Graphics

Go to:

```
LOCUS       NR_118328              1422 bp    rRNA    linear   BCT 03-FEB-2015
DEFINITION  Telmatocola sphagniphila strain SP2 16S ribosomal RNA gene, partial
            sequence.
ACCESSION   NR_118328
VERSION     NR_118328.1  GI:645321338
DBLINK      Project: 33175
            BioProject: PRJNA33175
KEYWORDS    RefSeq.
SOURCE      Telmatocola sphagniphila
  ORGANISM  Telmatocola sphagniphila
            Bacteria; Planctomycetes; Planctomycetia; Planctomycetales;
            Planctomycetaceae.
REFERENCE   1  (bases 1 to 1422)
  AUTHORS   Kulichevskaya,I.S., Serkebaeva,Y.M., Kim,Y., Rijpstra,W.I.,
            Damste,J.S., Liesack,W. and Dedysh,S.N.
  TITLE     Telmatocola sphagniphila gen. nov., sp. nov., a novel dendriform
            planctomycete from northern wetlands
  JOURNAL   Front Microbiol 3, 146 (2012)
   PUBMED   22529844
  REMARK    Publication Status: Online-Only
REFERENCE   2  (bases 1 to 1422)
  CONSRTM   NCBI RefSeq Targeted Loci Project
  TITLE     Direct Submission
  JOURNAL   Submitted (28-APR-2014) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
REFERENCE   3  (bases 1 to 1422)
  AUTHORS   Dedysh,S.N.
  TITLE     Direct Submission
  JOURNAL   Submitted (20-OCT-2011) Winogradsky Institute of Microbiology RAS,
            Prospect 60-Letya Octyabrya 7/2, Moscow 117312, Russia
COMMENT     REVIEWED REFSEQ: This record has b[illegible] NCBI staff. The
            reference sequence is identical to JN880417:1-1422.
```

# Blast columns

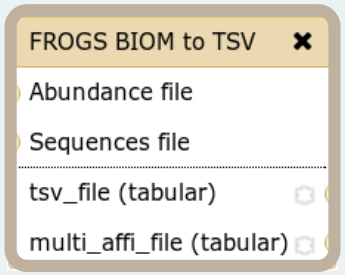
OTU_2 seed has a best BLAST hit with the reference sequence AJ496032.1.1410

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;Actinobacteria;Actinobacteria;Bifidobacteriales;Bifidobacteriaceae;Metascardovia;Multi-affiliation | multi-subject | 100.0 | 100.0 | 0.0 | 411 |
| Bacteria;Fibrobacteres;Fibrobacteria;Fibrobacterales;Fibrobacteraceae;Fibrobacter;Fibrobacter succinogenes | AJ496032.1.1410 | 100.0 | 100.0 | 0.0 | 419 |
| Bacteria;Firmicutes;Bacilli;Bacillales;Staphylococcaceae;Nosocomiicoccus;Nosocomiicoccus ampullae | EU240886.1.1502 | 100.0 | 100.0 | 0.0 | 427 |
| Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Moraxellaceae;Psychrobacter;Psychrobacter immobilis | U39399.1.1477 | 100.0 | 100.0 | 0.0 | 426 |
| Bacteria;Thermotogae;Thermotogae;Thermotogales;Thermotogaceae;Petrotoga;Petrotoga miotherma | FR733705.1.1499 | 100.0 | 100.0 | 0.0 | 419 |
| Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Phyllobacteriaceae;Pseudahrensia;Pseudahrensia aquimaris | GU575117.1.1441 | 100.0 | 100.0 | 0.0 | 401 |
| Bacteria;Bacteroidetes;Cytophagia;Cytophagales;Cytophagaceae;Persicitalea;Persicitalea jodogahamensis | multi-subject | 100.0 | 100.0 | 0.0 | 421 |
| Bacteria;Proteobacteria;Deltaproteobacteria;Bdellovibrionales;Bdellovibrionaceae;Bdellovibrio;Multi-affiliation | multi-subject | 100.0 | 100.0 | 0.0 | 404 |

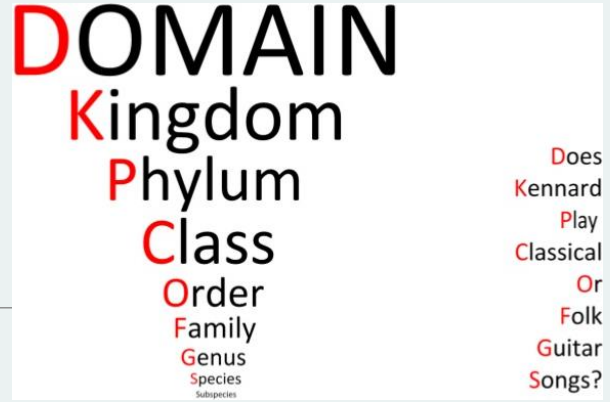
Evaluation variables of BLAST

**Convert to TSV**



# Focus on “Multi-”

*(affiliation from silva 123)*




Observe line of Cluster 1 inside abundance.tsv and multi_hit.tsv files, what do you conclude ?

| #blast_taxonomy | blast_subject | blast_perc_identity | blast_perc_query_coverage | blast_evalue | blast_aln_length |
|---|---|---|---|---|---|
| Bacteria;Actinobacteria;Actinobacteria;Bifidobacteriales;Bifidobacteriaceae;Metascardovia;Multi-affiliation | multi-subject | 100.0 | 100.0 | 0.0 | 411 |
| Bacteria;Fibrobacteres;Fibrobacteria;Fibrobacterales;Fibrobacteraceae;Fibrobacter;Fibrobacter succinogenes | AJ496032.1.1410 | 100.0 | 100.0 | 0.0 | 419 |
| Bacteria;Firmicutes;Bacilli;Bacillales;Staphylococcaceae;Nosocomiicoccus;Nosocomiicoccus ampullae | EU240886.1.1502 | 100.0 | 100.0 | 0.0 | 427 |
| Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Moraxellaceae;Psychrobacter;Psychrobacter immobilis | U39399.1.1477 | 100.0 | 100.0 | 0.0 | 426 |
| Bacteria;Thermotogae;Thermotogae;Thermotogales;Thermotogaceae;Petrotoga;Petrotoga miotherma | FR733705.1.1499 | 100.0 | 100.0 | 0.0 | 419 |
| Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Phyllobacteriaceae;Pseudahrensia;Pseudahrensia aquimaris | GU575117.1.1441 | 100.0 | 100.0 | 0.0 | 401 |
| Bacteria;Bacteroidetes;Cytophagia;Cytophagales;Cytophagaceae;Persicitalea;Persicitalea jodogahamensis | multi-subject | 100.0 | 100.0 | 0.0 | 421 |
| Bacteria;Proteobacteria;Deltaproteobacteria;Bdellovibrionales;Bdellovibrionaceae;Bdellovibrio;Multi-affiliation | multi-subject | 100.0 | 100.0 | 0.0 | 404 |

Cluster_1 has 5 identical blast hits, with different taxonomies as the species level

# Focus on “Multi-”

*(affiliation from silva 123)*

Observe line of Cluster 11 inside abundance.tsv and multi_hit.tsv files, what do you conclude ?

| | | | |
|---|---|---|---|
| Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Hyphomonadaceae;Henriciella;Henriciella marina | multi-subject | 100.0 | 100.0 |

Cluster_11 has 2 identical blast hits, with identical species but with different strains (strains are not written in our data)

# Focus on “Multi-”

*(affiliation from silva 123)*

Observe line of Cluster 43 inside abundance.tsv and multi_hit.tsv files, what do you conclude ?

| | | | | |
|---|---|---|---|---|
| Bacteria;Firmicutes;Negativicutes;Selenomonadales;Veillonellaceae;Multi-affiliation;Multi-affiliation | | multi-subject | 99.3 | 100.0 |

| | | |
|---|---|---|
| Cluster_43 | Bacteria;Firmicutes;Negativicutes;Selenomonadales;Veillonellaceae;Selenomonas 3;unknown species | JQ447821.1.1420 |
| Cluster_43 | Bacteria;Firmicutes;Negativicutes;Selenomonadales;Veillonellaceae;Centipeda;Centipeda periodontii | AJ010963.1.1494 |

Cluster_43 has 2 identical blast hits, with different taxonomies at the genus level

# Back on Blast parameters

| #blast_taxonomy | blast_subject | blast_perc_identity | blast_perc_query_coverage | blast_evalue | blast_aln_length |
|---|---|---|---|---|---|
| Bacteria;Actinobacteria;Actinobacteria;Bifidobacteriales;Bifidobacteriaceae;Metascardovia;Multi-affiliation | multi-subject | 100.0 | 100.0 | 0.0 | 411 |
| Bacteria;Fibrobacteres;Fibrobacteria;Fibrobacterales;Fibrobacteraceae;Fibrobacter;Fibrobacter succinogenes | AJ496032.1.1410 | 100.0 | 100.0 | 0.0 | 419 |
| Bacteria;Firmicutes;Bacilli;Bacillales;Staphylococcaceae;Nosocomiicoccus;Nosocomiicoccus ampullae | EU240886.1.1502 | 100.0 | 100.0 | 0.0 | 427 |
| Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Moraxellaceae;Psychrobacter;Psychrobacter immobilis | U39399.1.1477 | 100.0 | 100.0 | 0.0 | 426 |
| Bacteria;Thermotogae;Thermotogae;Thermotogales;Thermotogaceae;Petrotoga;Petrotoga miotherma | FR733705.1.1499 | 100.0 | 100.0 | 0.0 | 419 |
| Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Phyllobacteriaceae;Pseudahrensia;Pseudahrensia aquimaris | GU575117.1.1441 | 100.0 | 100.0 | 0.0 | 401 |
| Bacteria;Bacteroidetes;Cytophagia;Cytophagales;Cytophagaceae;Persicitalea;Persicitalea jodogahamensis | multi-subject | 100.0 | 100.0 | 0.0 | 421 |
| Bacteria;Proteobacteria;Deltaproteobacteria;Bdellovibrionales;Bdellovibrionaceae;Bdellovibrio;Multi-affiliation | multi-subject | 100.0 | 100.0 | 0.0 | 404 |

Evaluation variables of BLAST

# 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    TGGGGAATATTGCACAATGGGGGGAACCCTGATGCAGCGACGCCGCGTGCGGGATGACGG  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  331  TGGGGAATATTGCACAATGGGGGGAACCCTGATGCAGCGACGCCGCGTGCGGGATGACGG  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    TGGGGAATATTGCACAATGGGGGGAACCCTGATGCAGCGACGCCGCGTGCGGGATGACGG  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  331  TGGGGAATATTGCACAATGGGGGGAACCCTGATGCAGCGACGCCGCGTGCGGGATGACGG  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
```

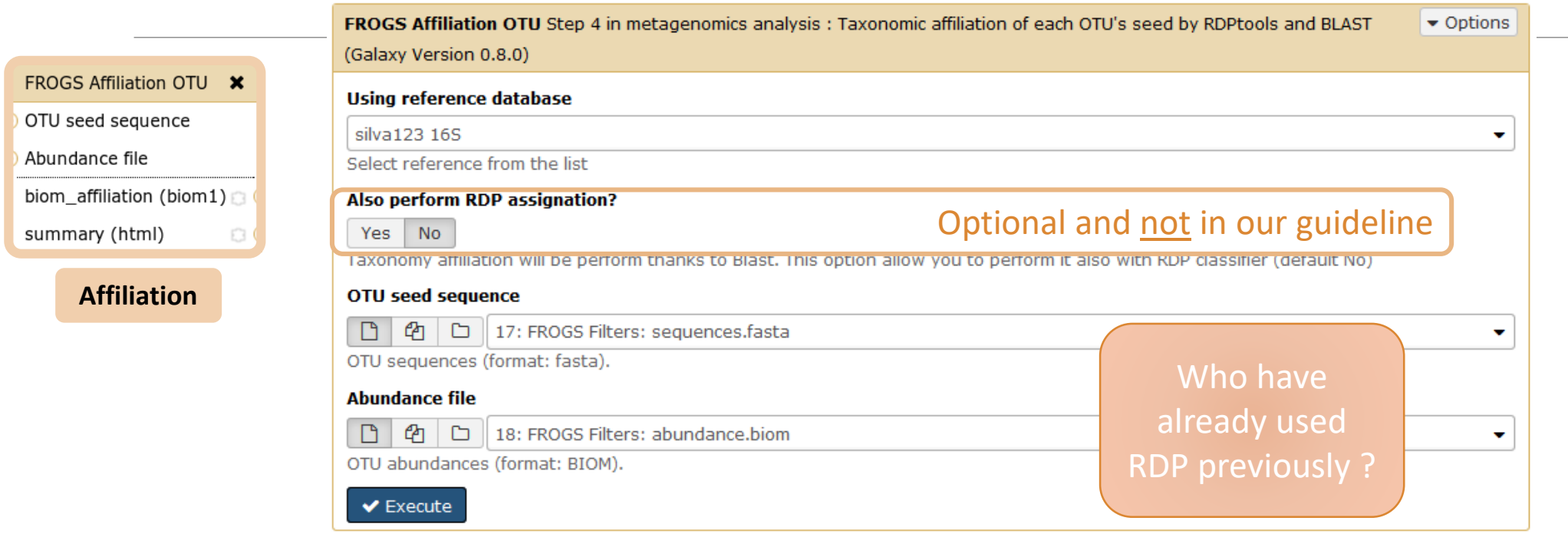
Query length = 411
100% coverage

# Blast variables : blast-length

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

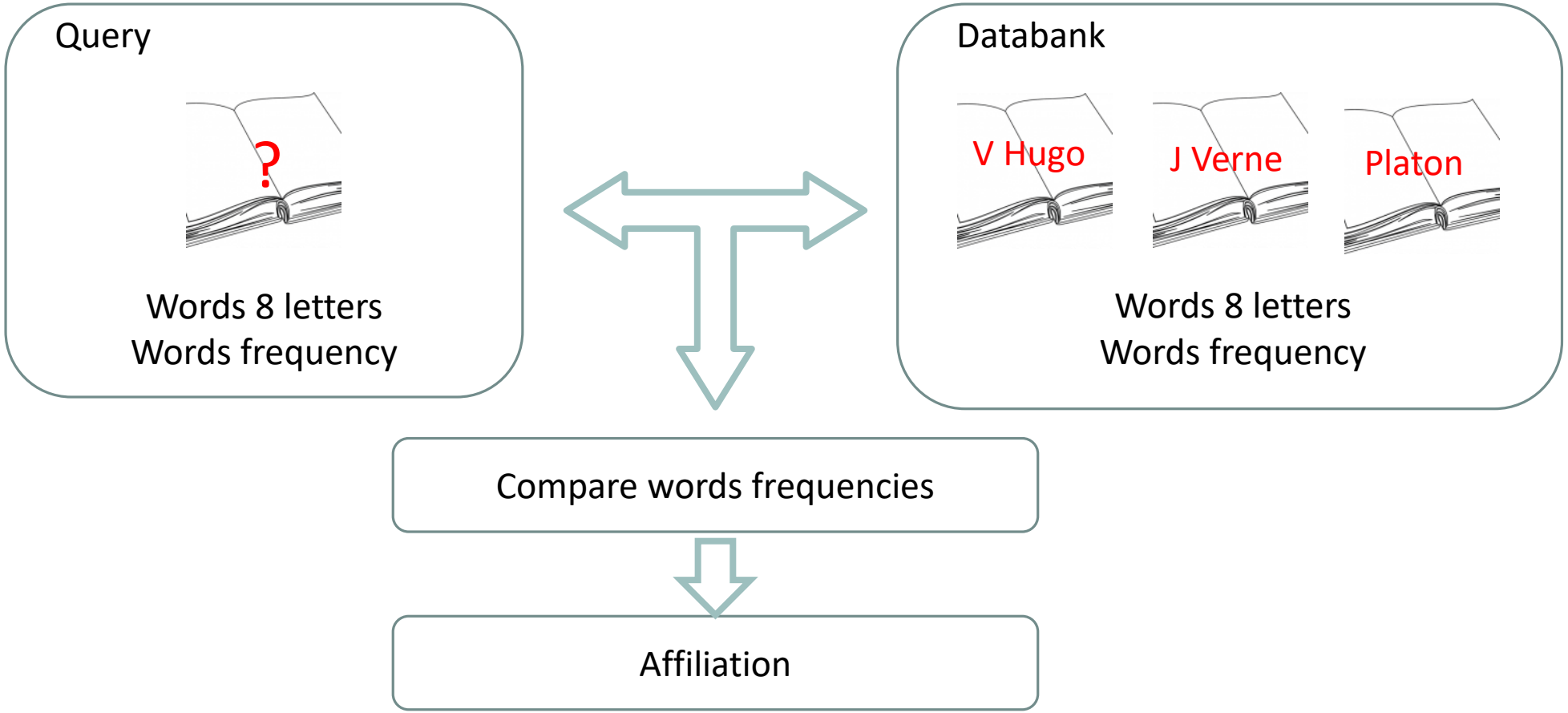
| | Coverage % | Identity % | Length alignment |
|---|---|---|---|
| OTU1 | 100 | 98 | 400 |
| OTU2 | 100 | 98 | 500 |

More mismatches/gaps

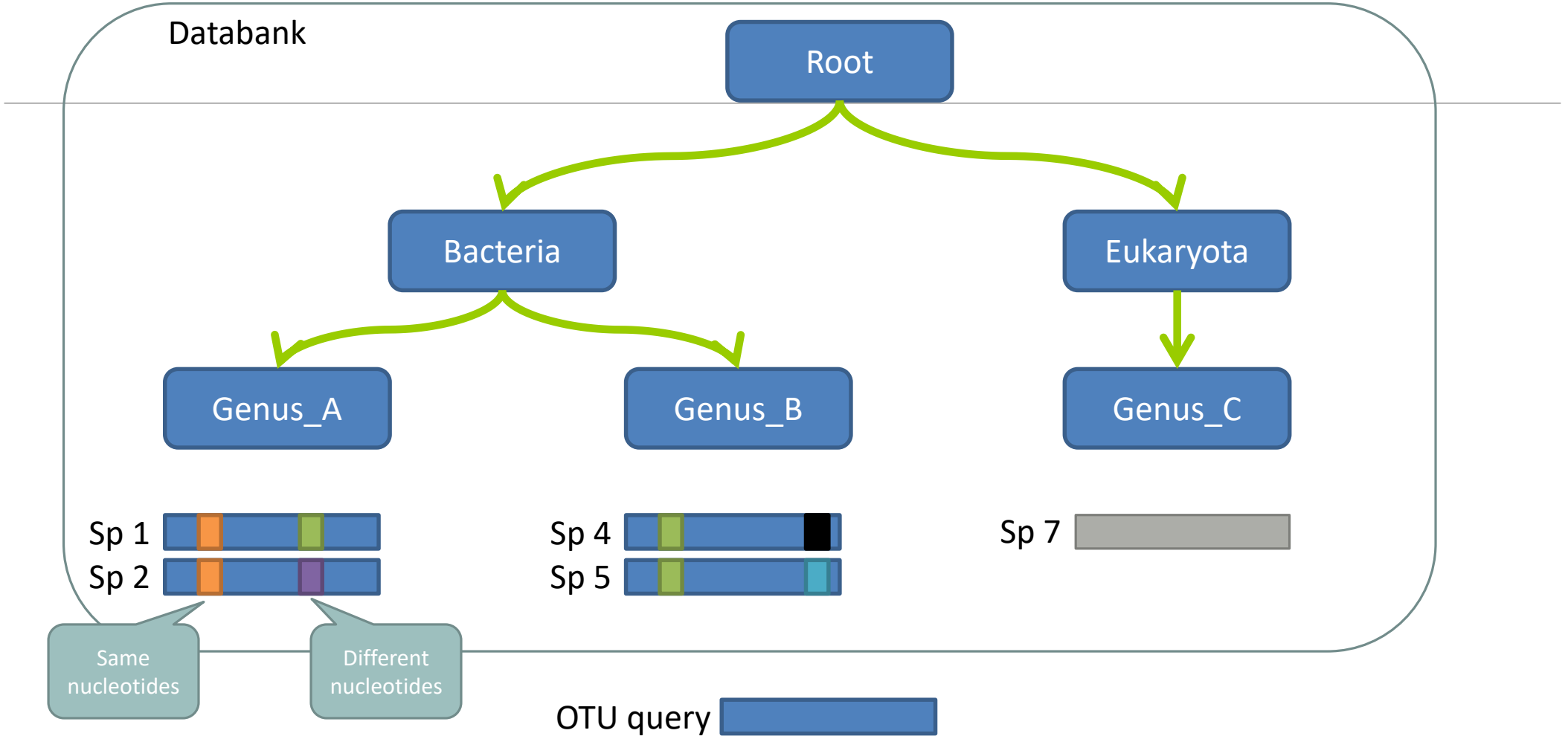



**Escape RDP explanation**

# How works RDP ?

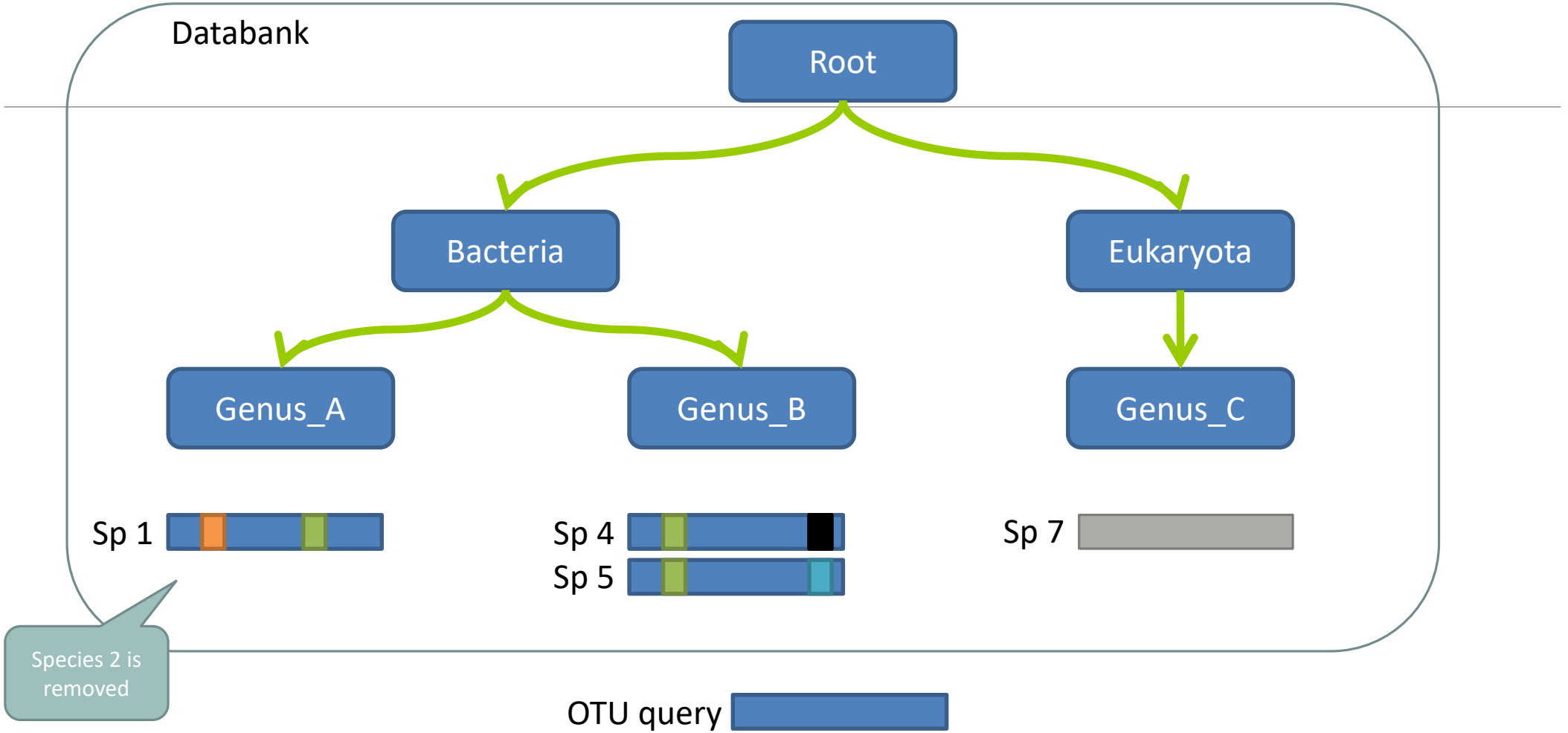


# How works RDP ?



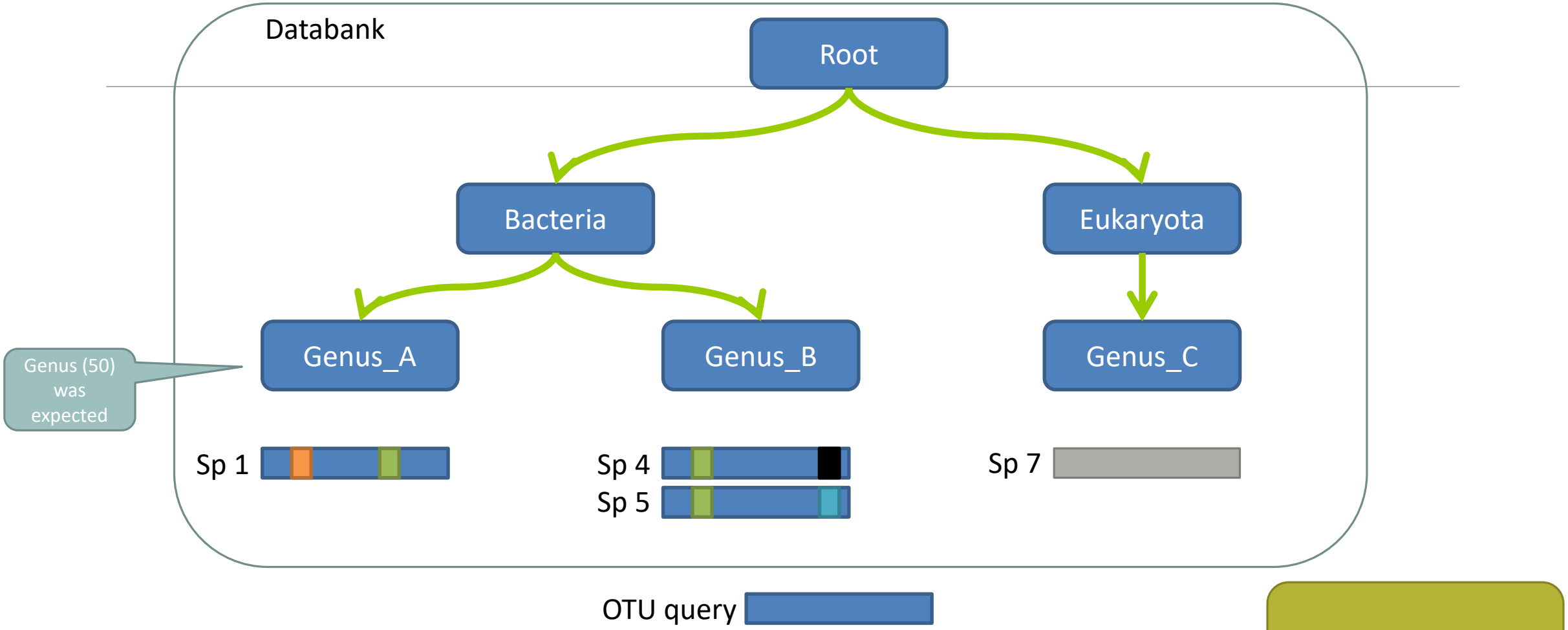

**Result:**
Bacteria(100) ; Genus_A(50) ; Sp1(25)

# The dysfunctions of RDP ?




**Result:**
?

# The dysfunctions of RDP n°1 ?

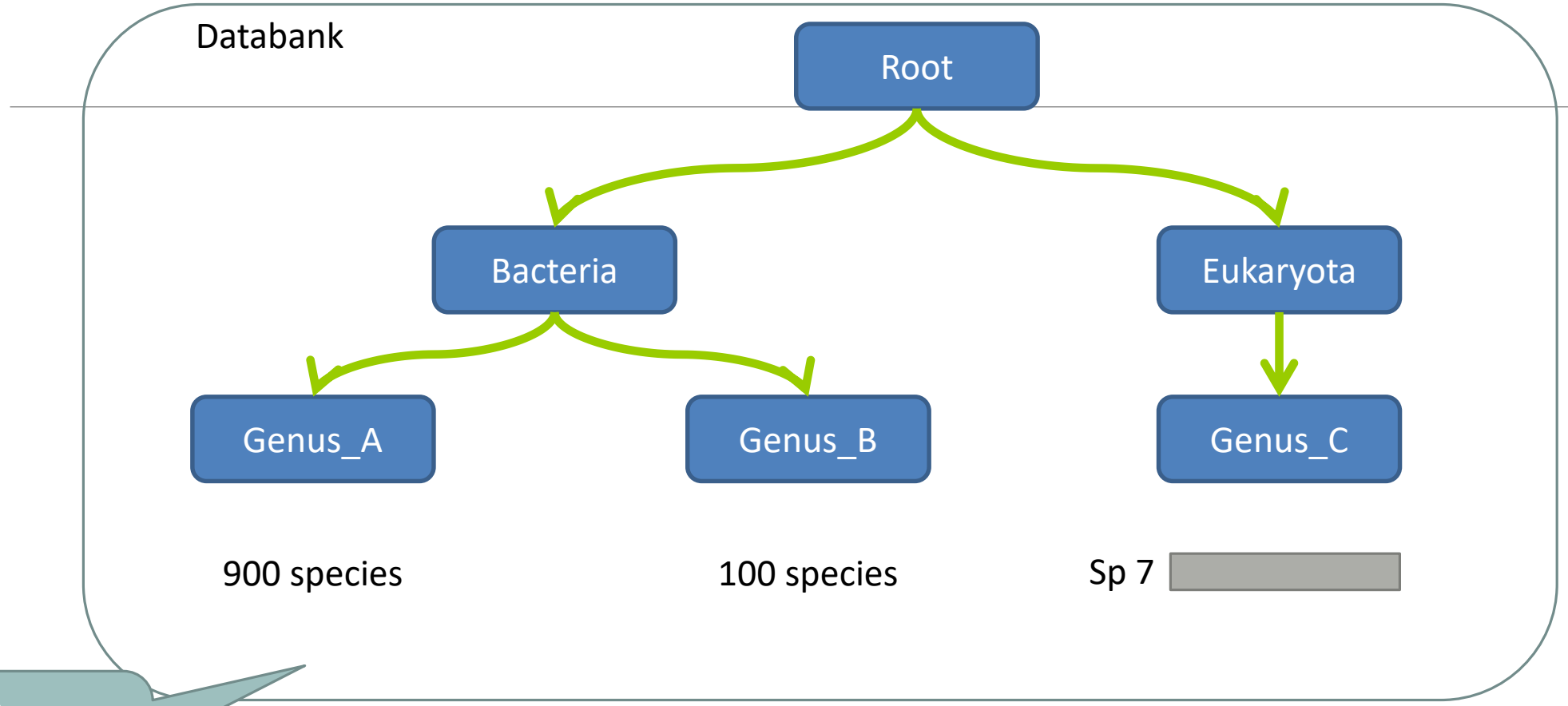



Order dependent

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

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

# The dysfunctions of RDP n°2 ?



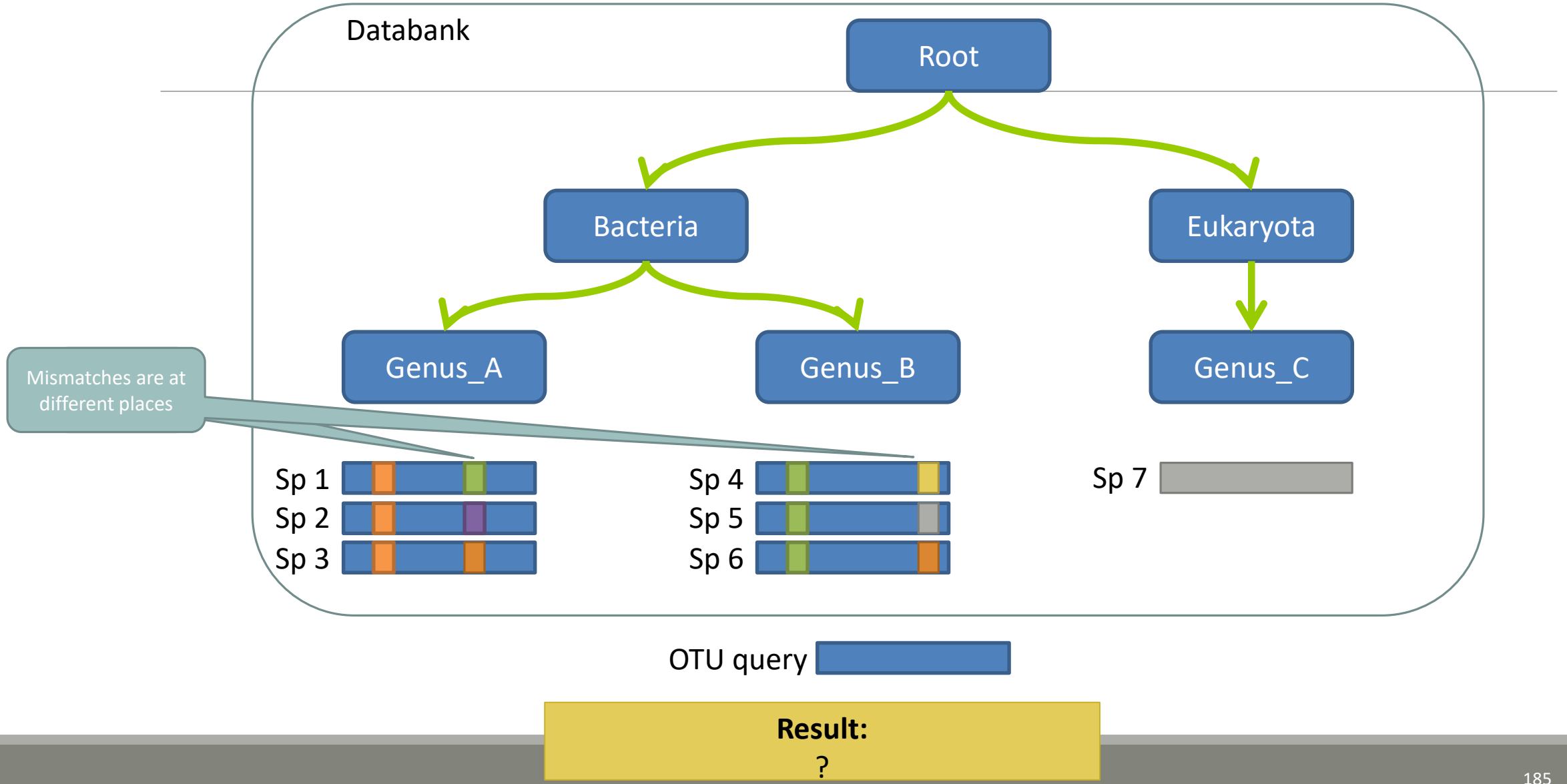

Many species in one genus and little in the other: So, RDP can give very different results

Influenced by heterogeneity in last ranks

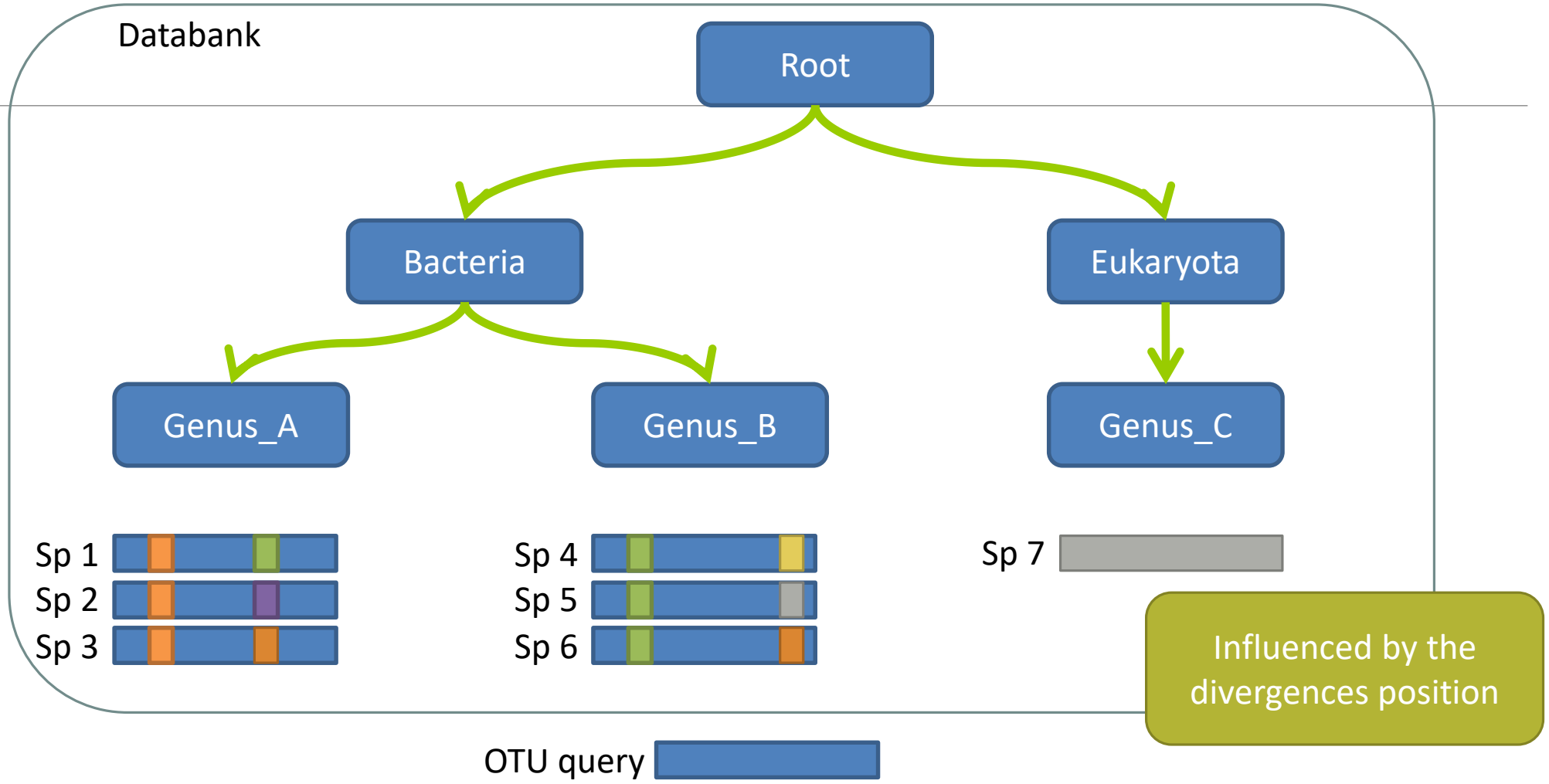
**Result:**

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

# The dysfunctions of RDP n°3 ?



# The dysfunctions of RDP n°3 ?




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.

# 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 |

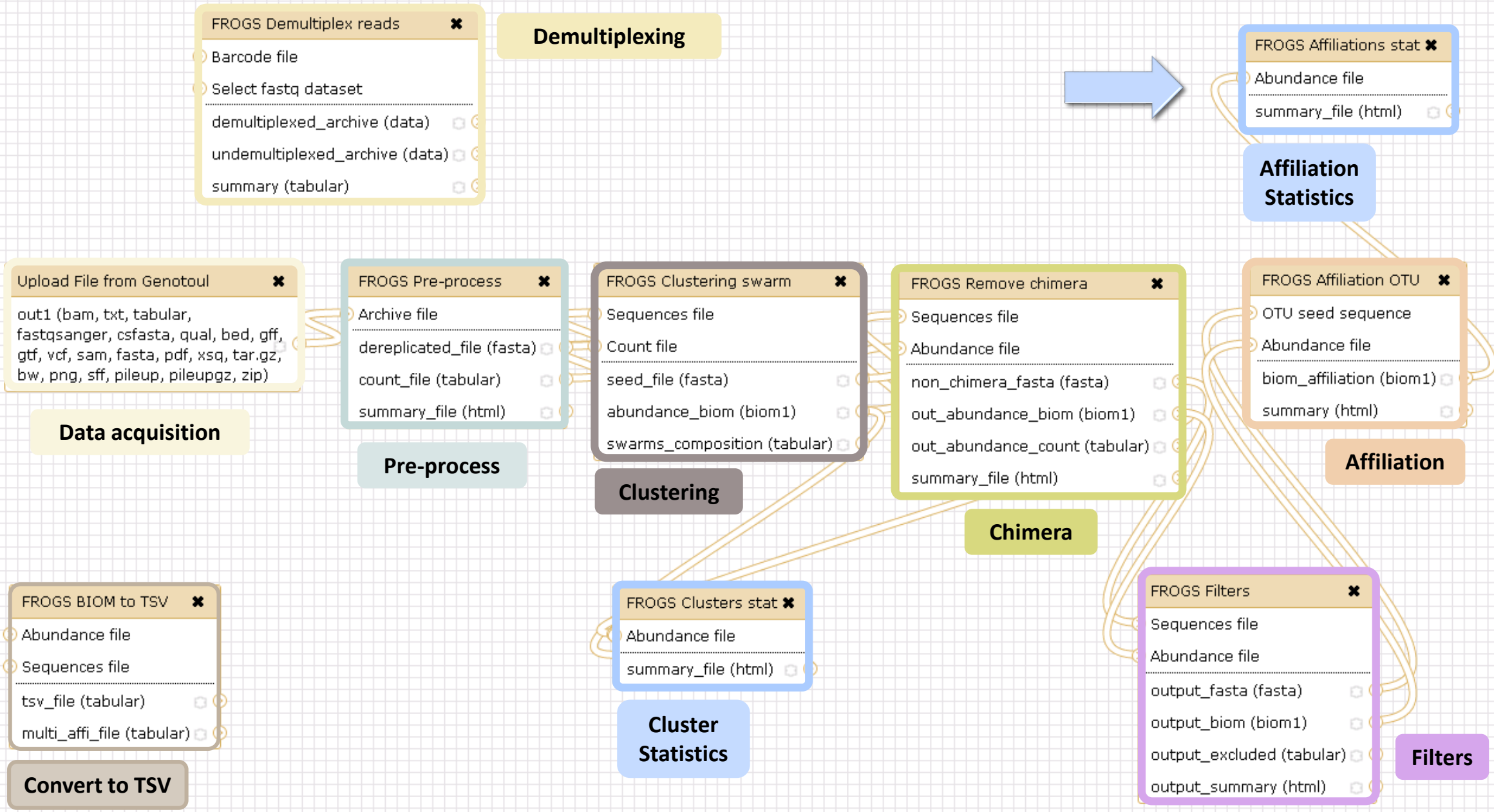
With the FROGS guideline

| 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

# 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** Process some metrics on taxonomies. (Galaxy Version 1.1.0) ▾ Options

**Abundance file**

22: 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'.

✔ Execute

OR

**FROGS Affiliations stat** Process some metrics on taxonomies. (Galaxy Version 1.1.0) ▾ Options

**Abundance file**

22: 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'.

✔ Execute

Taxonomy distribution | Alignment distribution

OR

Taxonomy distribution | Bootstrap distribution

**FROGS Affiliations stat** Process some metrics on taxonomies. (Galaxy Version 1.1.0) ▾ Options

**Abundance file**

22: 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**

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**

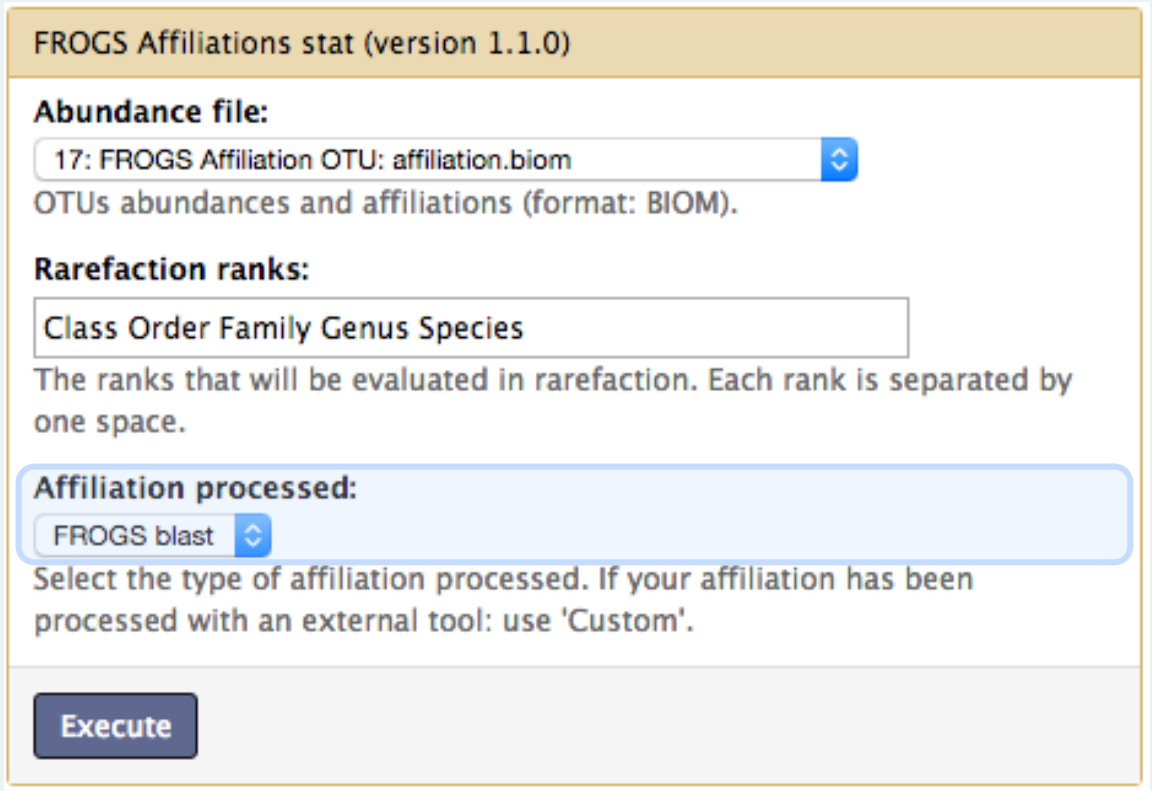
The metadata tag used in BIOM file to store the alignment identity.

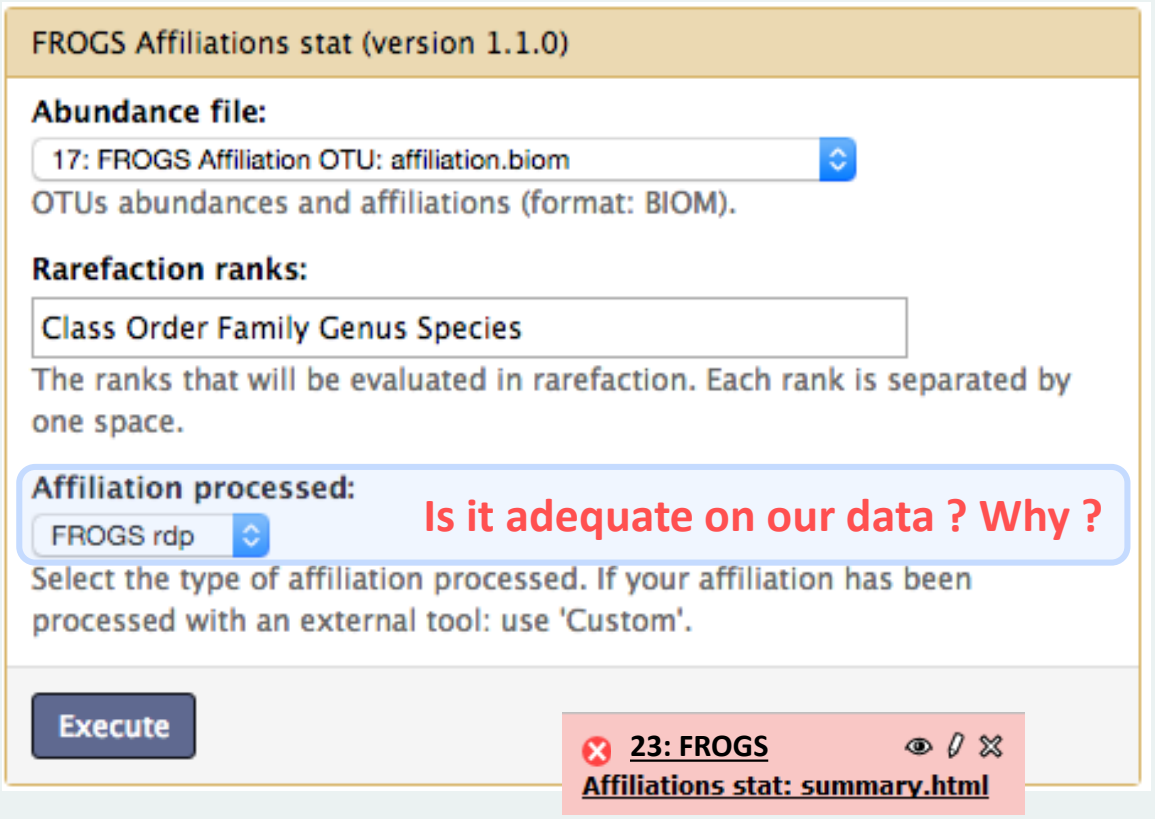
**Coverage tag**

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

✔ Execute

# Exercise 6.2


FROGS Affiliations stat (version 1.1.0)
Abundance file:
17: 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'.
Execute

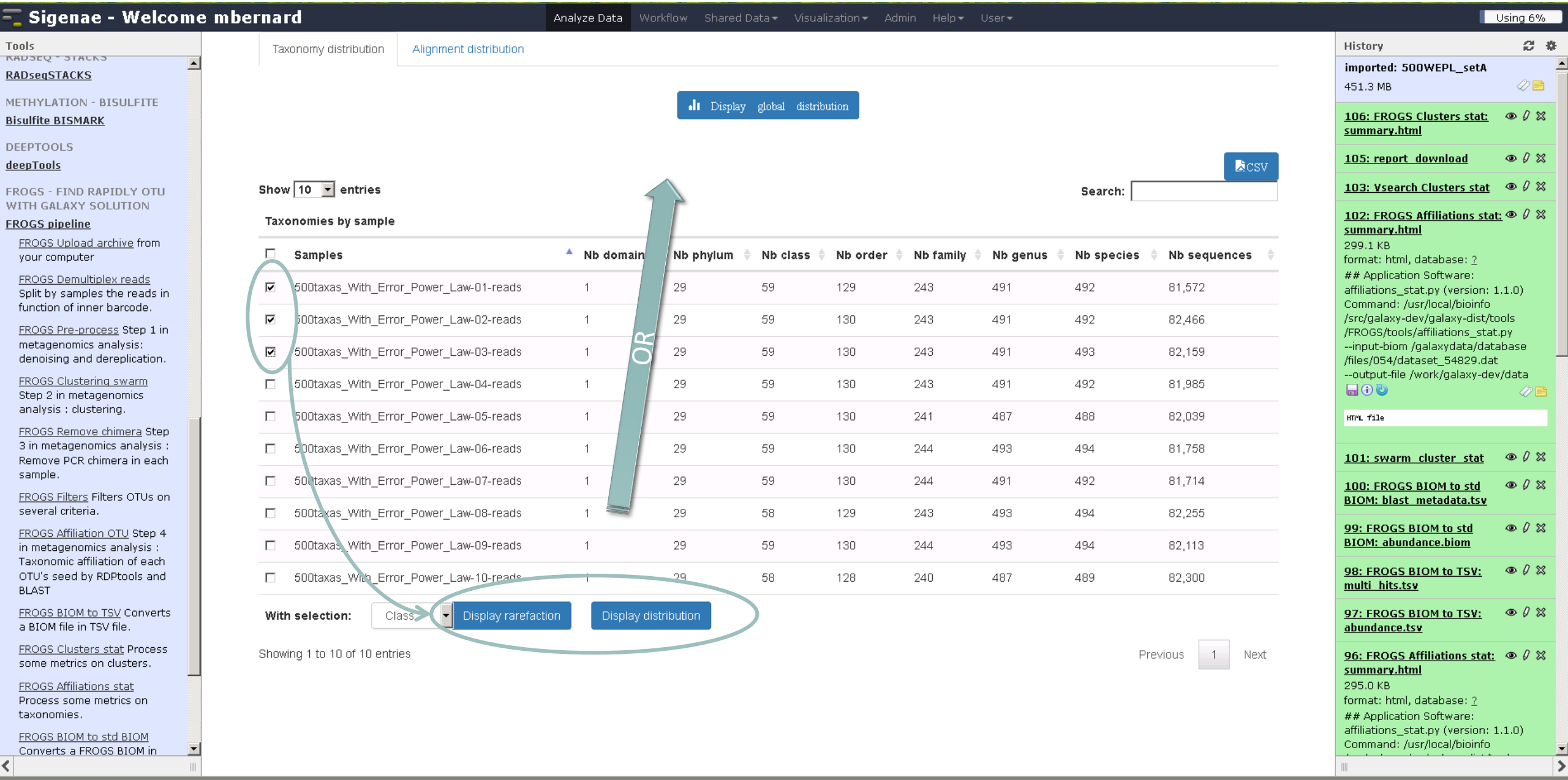

FROGS Affiliations stat (version 1.1.0)
Abundance file:
17: FROGS Affiliation OTU: affiliation.biom
OTUs abundances and affiliations (format: BIOM).
Rarefaction ranks:
Class Order Family Genus Species
The ranks that will be evaluated in rarefaction. Each rank is separated by one space.
Affiliation processed:
FROGS rdp
Is it adequate on our data ? Why ?
Select the type of affiliation processed. If your affiliation has been processed with an external tool: use 'Custom'.
Execute
23: FROGS
Affiliations stat: summary.html

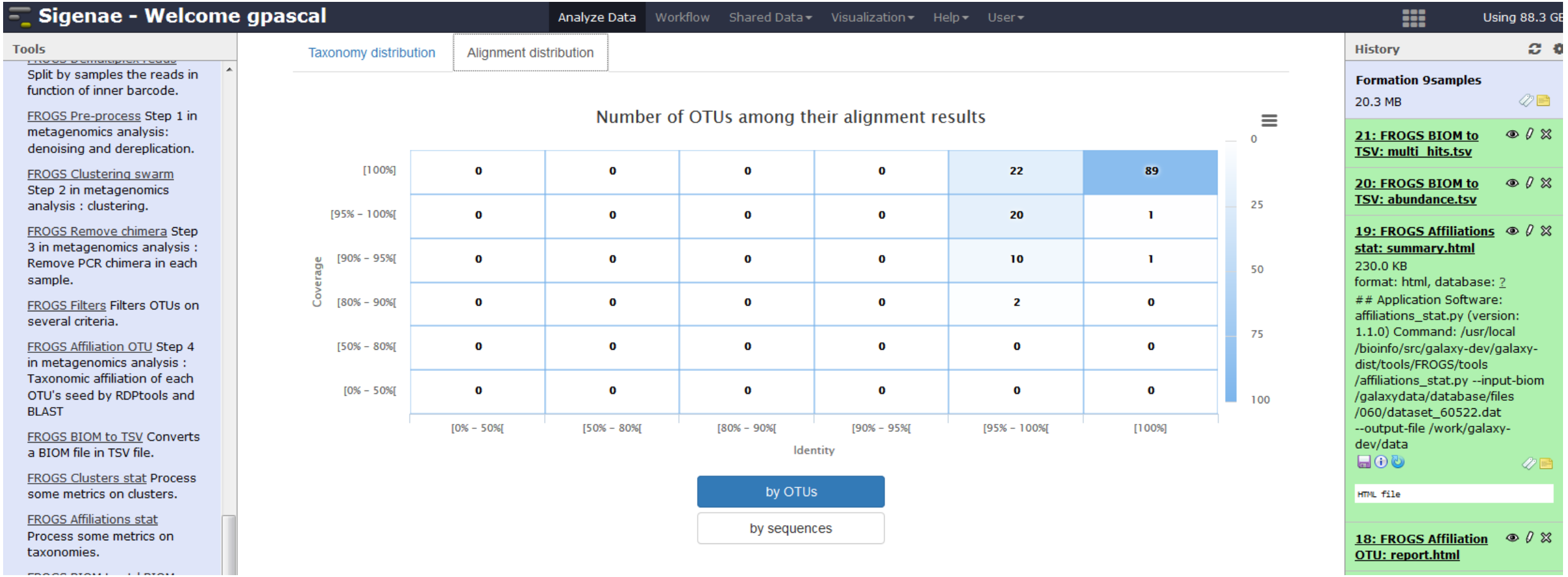
# Exercise 6.2

→ objectives :

understand rarefaction curve and sunburst

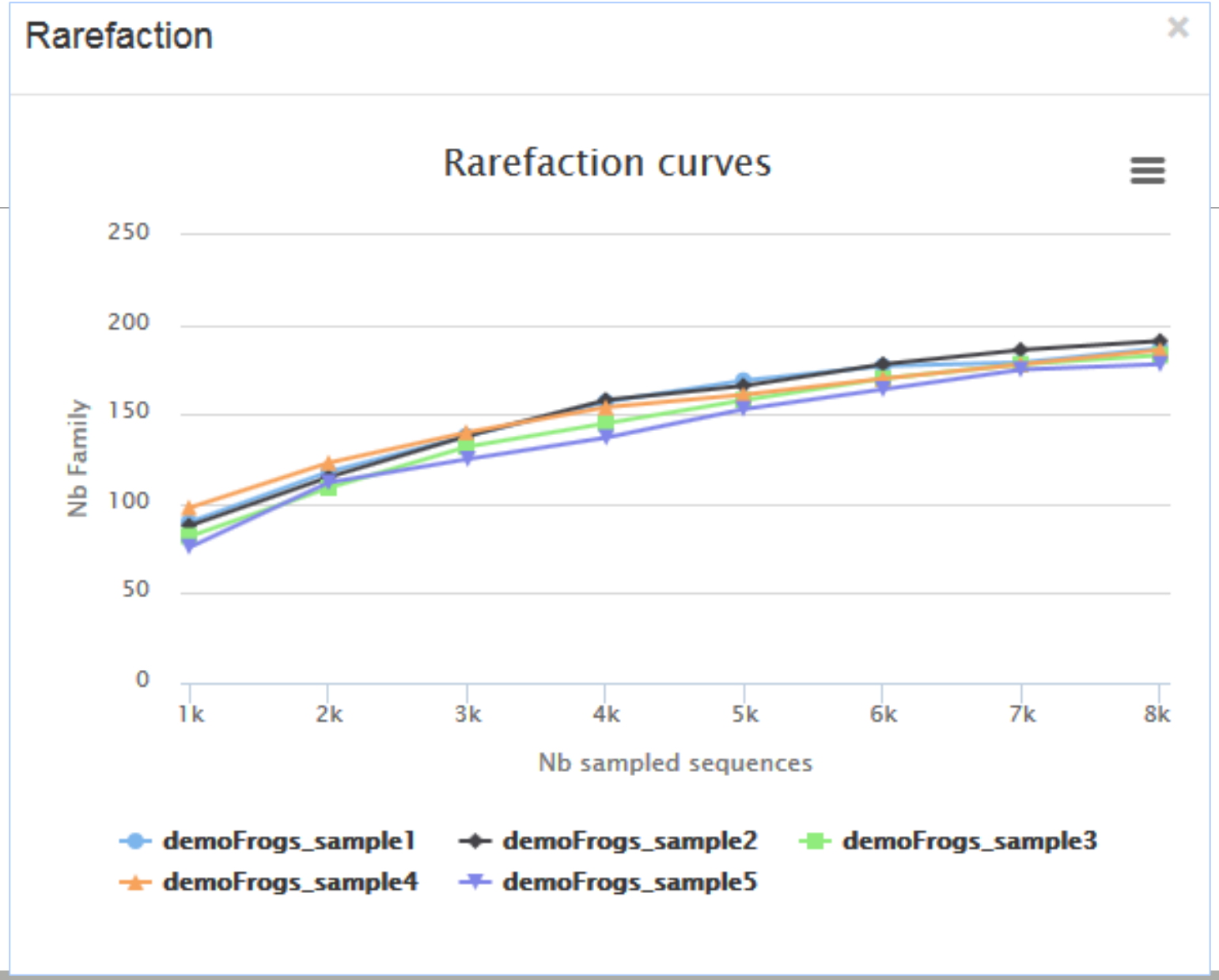
1. Explore the Affiliation stat results on FROGS blast affiliation.
2. What kind of graphs can you generate? What do they mean?


Sigenae - Welcome mbernard
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 6%
Tools
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
OR
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


Sigenae - Welcome gpascal
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 88.3 GB
Tools
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.
Taxonomy distribution
Alignment distribution
Number of OTUs among their alignment results
Coverage
[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
[0% – 50%[ [50% – 80%[ [80% – 90%[ [90% – 95%[ [95% – 100%[ [100%]
Identity
0 25 50 75 100
by OTUs
by sequences
History
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



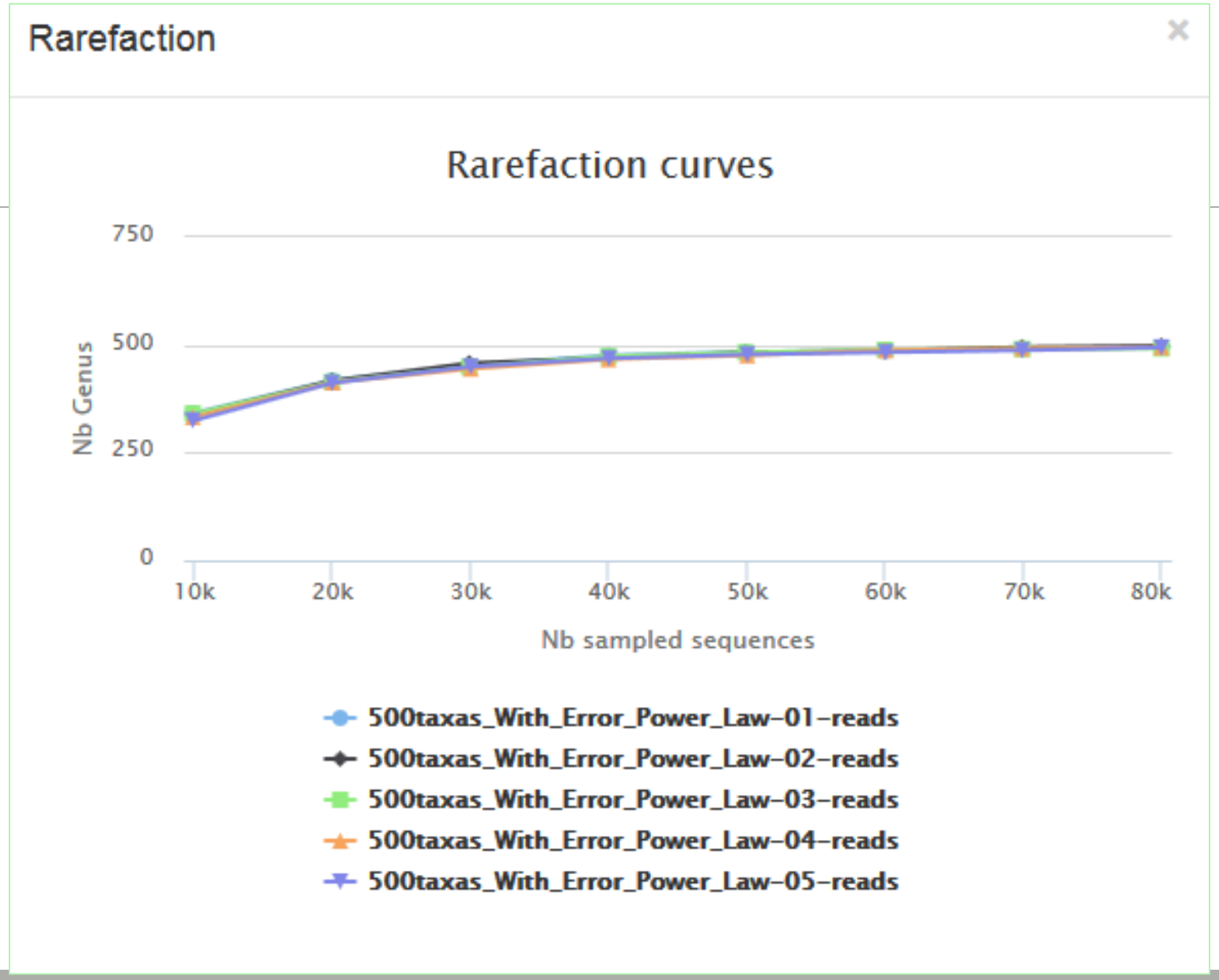

The curve continues to rise

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

Rarefaction tab

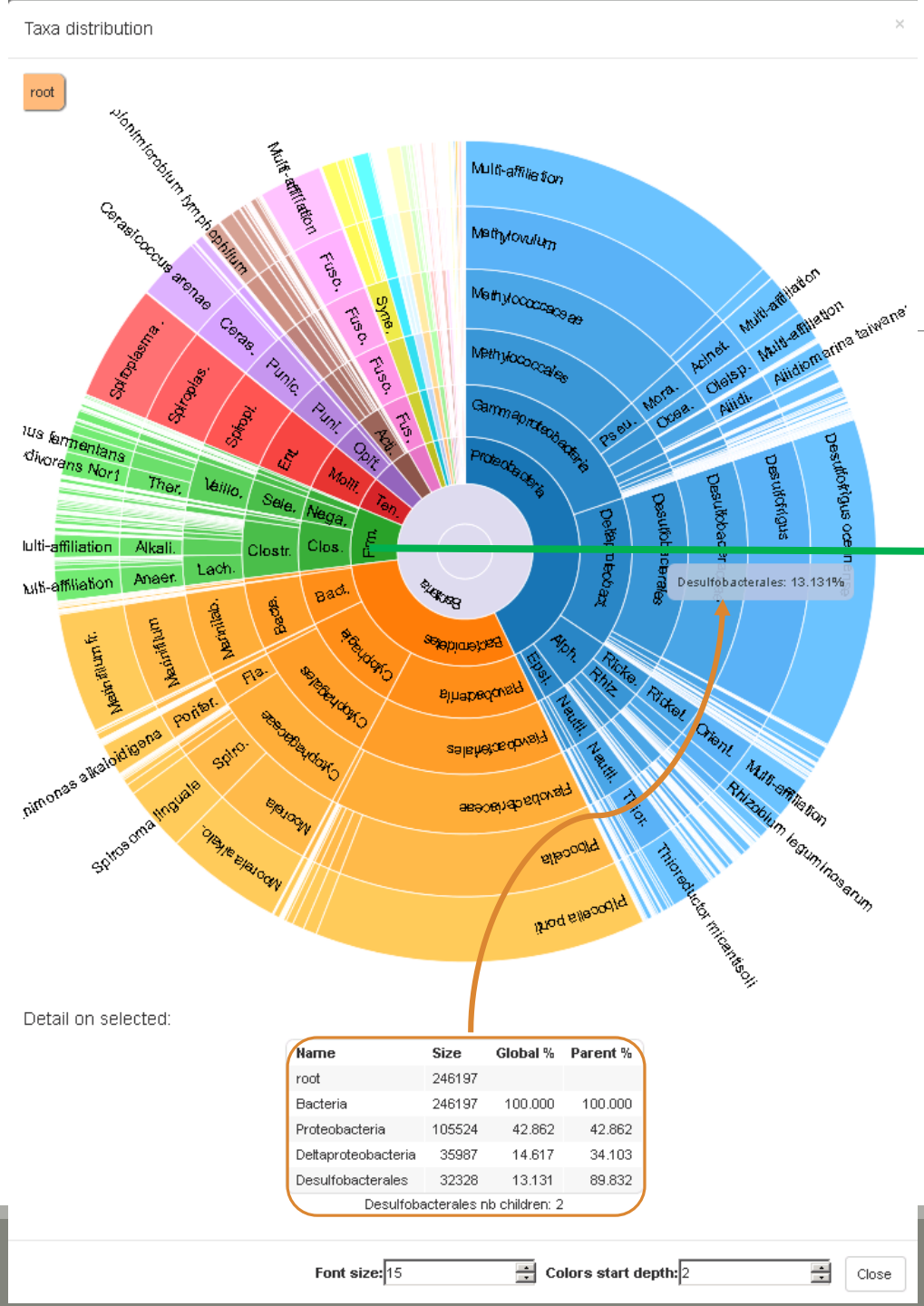
Available only after AFFILIATION TOOL

Samples size ~85 000 sequences

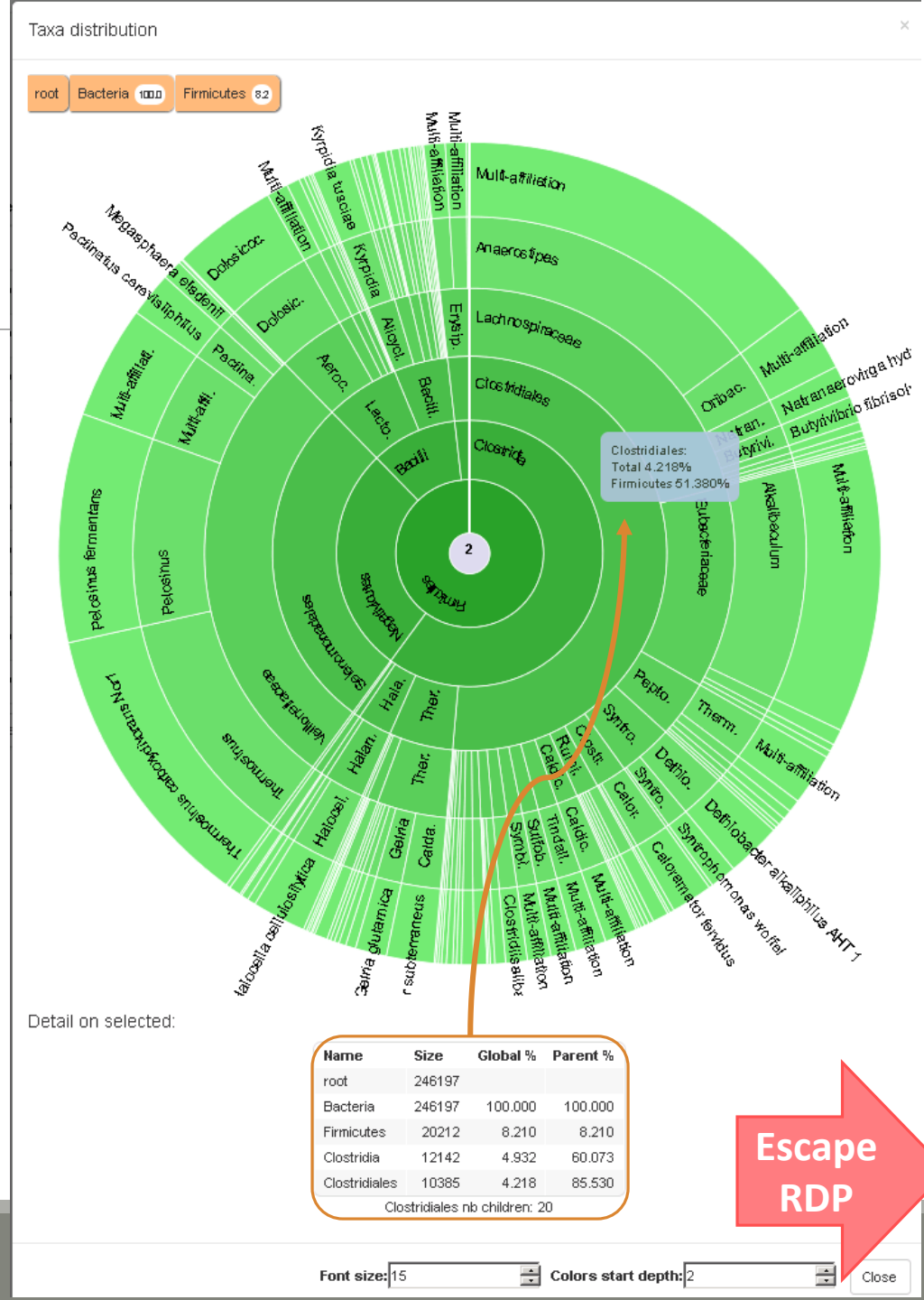



The curve slows to rise with ~50 000 sequences

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

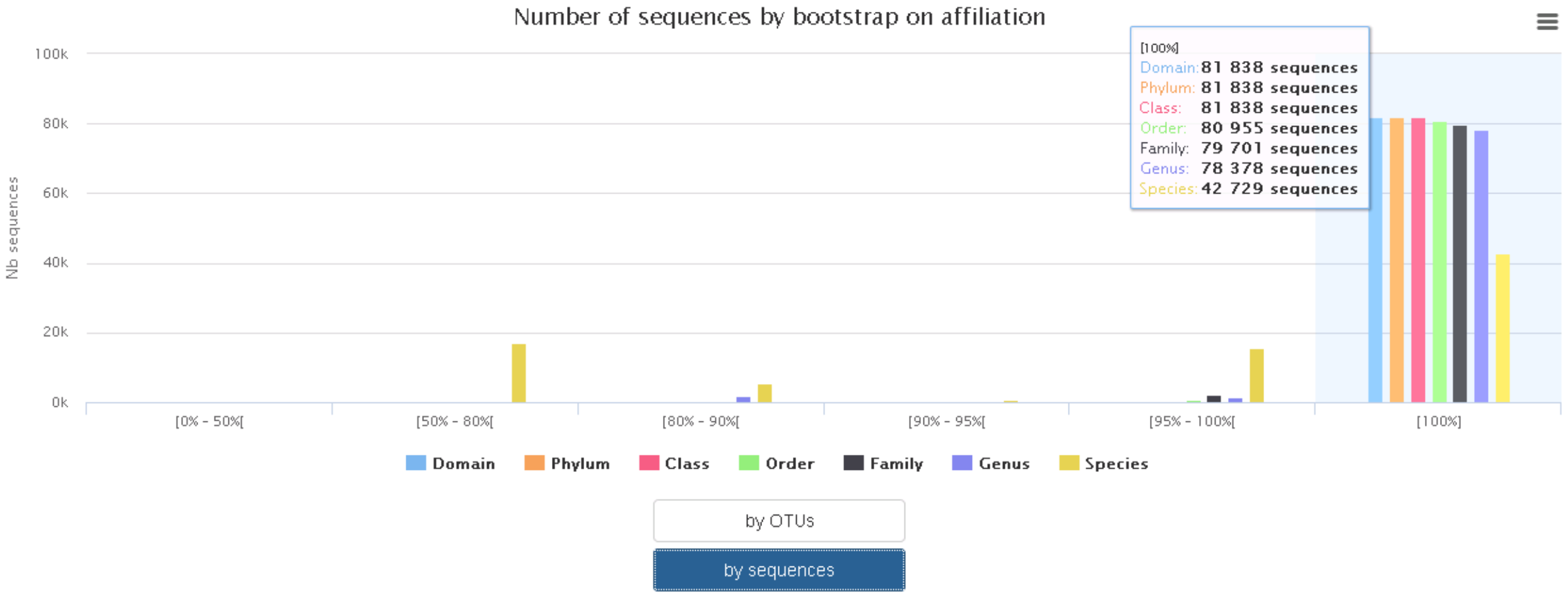



Zoom in on firmicutes




Escape RDP

Taxonomy distribution | Bootstrap distribution

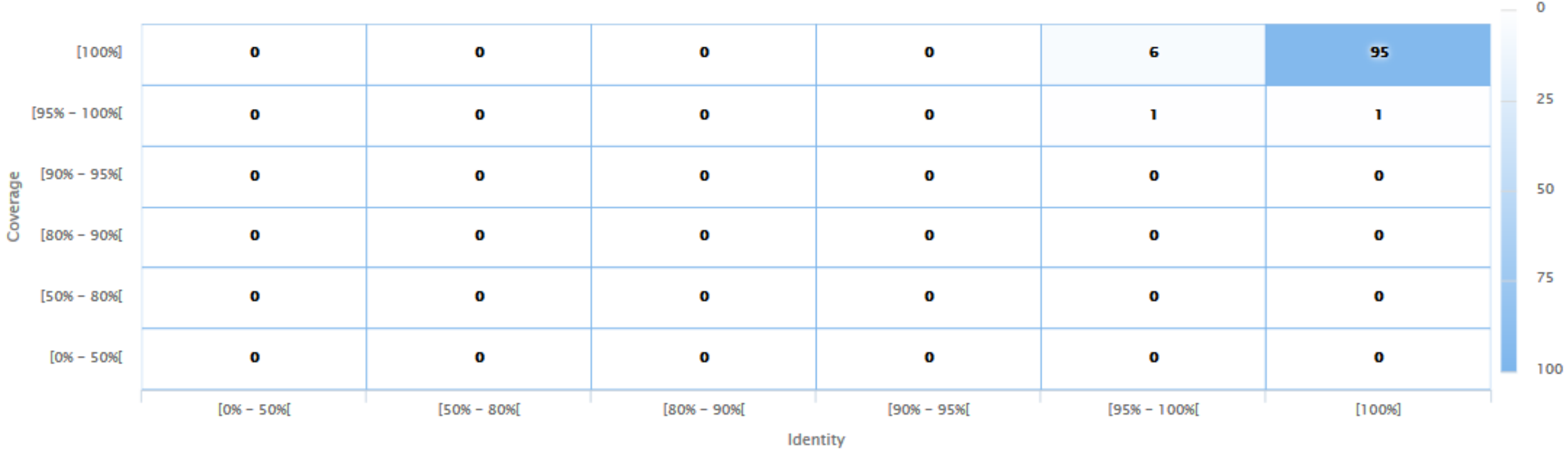



by OTUs

by sequences

Taxonomy distribution | Alignment distribution

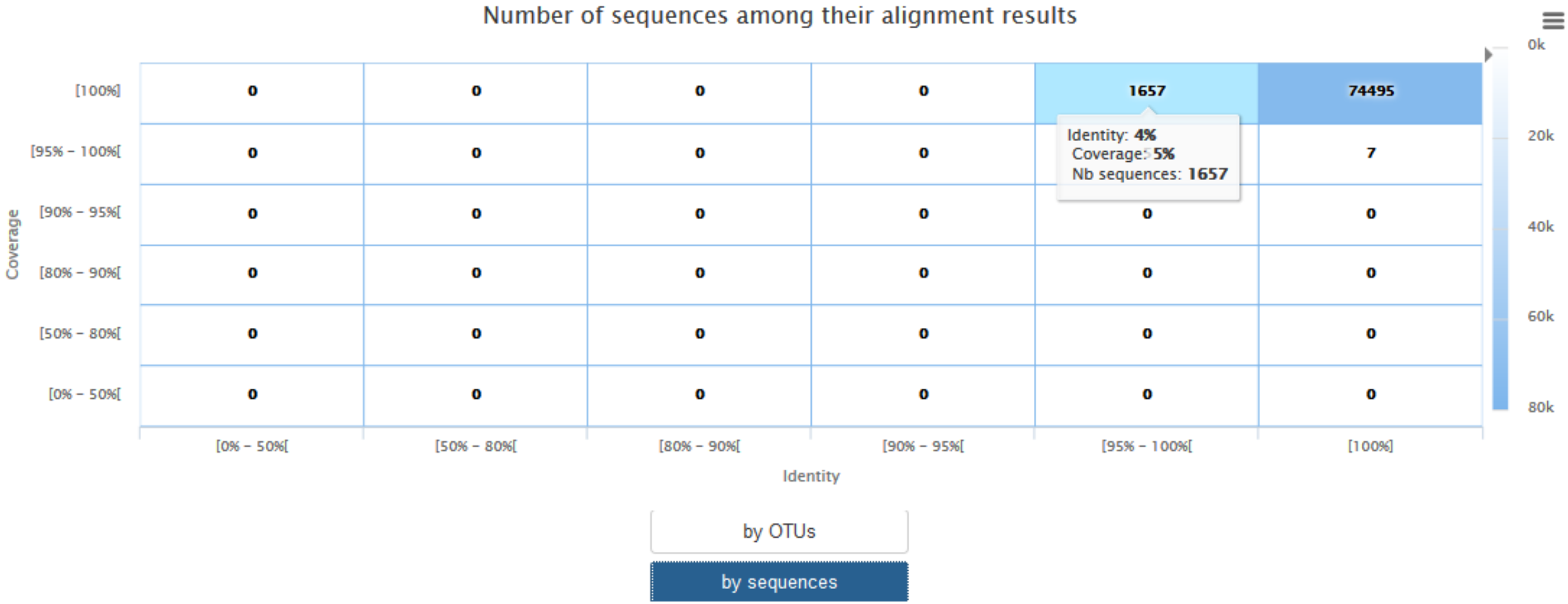
## Number of OTUs among their alignment results



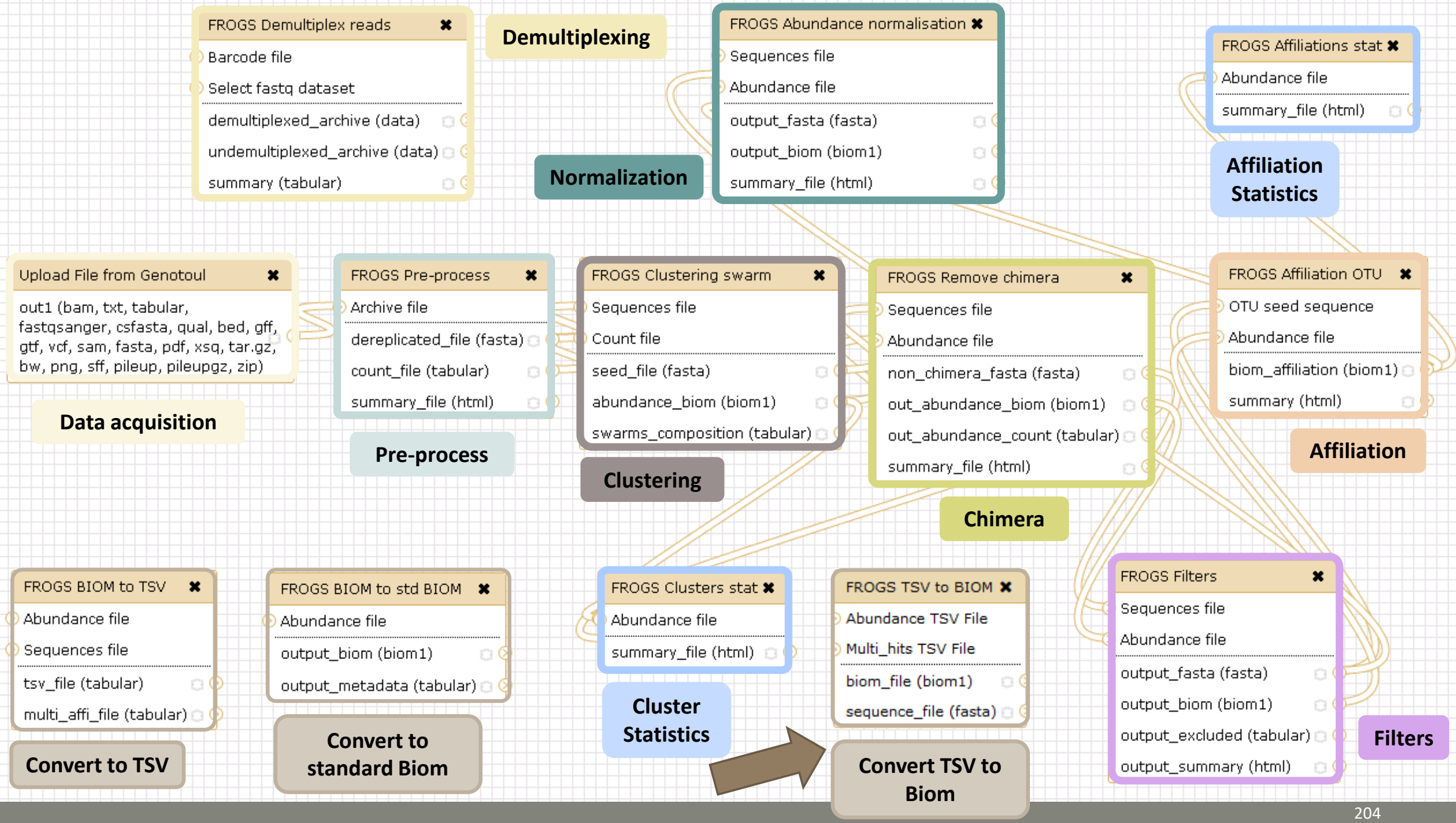

by OTUs

by sequences

Taxonomy distribution | Alignment distribution



# 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

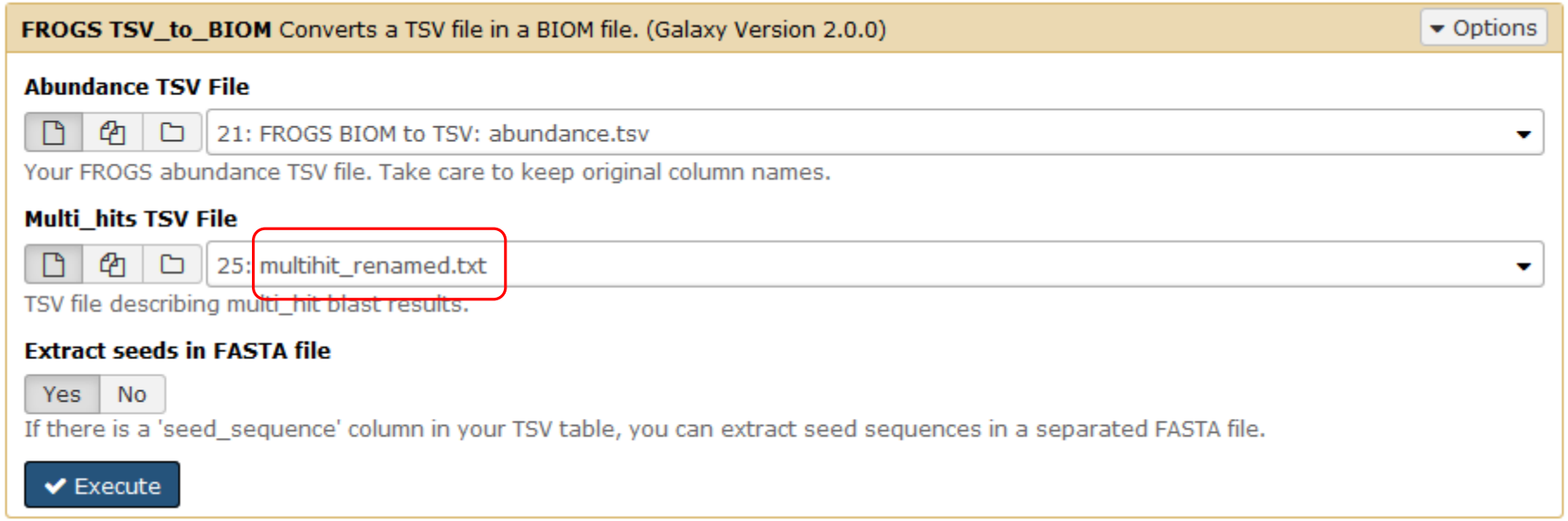
After modifying your abundance TSV file you can again:

- 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 multi_hit, 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 multi_hit TSV file.

# TSV to BIOM


FROGS TSV_to_BIOM Converts a TSV file in a BIOM file. (Galaxy Version 2.0.0)
Options
Abundance TSV File
21: FROGS BIOM to TSV: abundance.tsv
Your FROGS abundance TSV file. Take care to keep original column names.
Multi_hits TSV File
25: multihit_renamed.txt
TSV file describing multi_hit blast results.
Extract seeds in FASTA file
Yes
No
If there is a 'seed_sequence' column in your TSV table, you can extract seed sequences in a separated FASTA file.
Execute

# Your Turn! – 7

PLAY WITH TSV_TO_BIOM

# Exercise 7

→ objectives : Play with multi-affiliation and TSV_to_BIOM

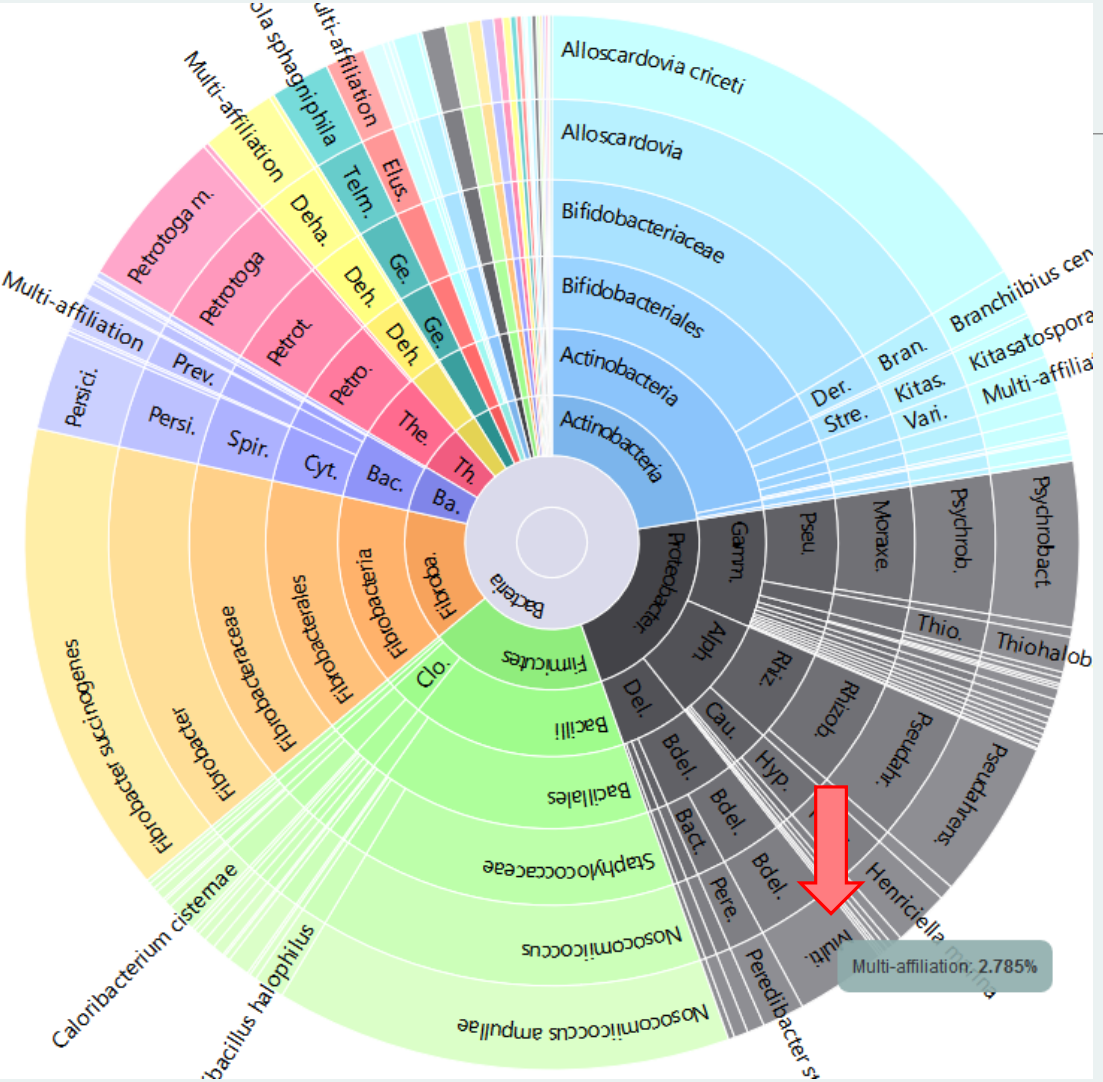
1. Observe in Multi_hit.tsv and abundance.tsv cluster_8 annotation

| #blast_taxonomy | blast_subject | observation_name | observation_sum |
|---|---|---|---|
| Bacteria;Actinobacteria;Actinobacteria;Bifidobacteriales;Bifidobacteriaceae;Metascardovia;Multi-affiliation | multi-subject | Cluster_1 | 13337 |
| Bacteria;Fibrobacteres;Fibrobacteria;Fibrobacterales;Fibrobacteraceae;Fibrobacter;Fibrobacter succinogenes | AJ496032.1.1410 | Cluster_2 | 11830 |
| Bacteria;Firmicutes;Bacilli;Bacillales;Staphylococcaceae;Nosocomiicoccus;Nosocomiicoccus ampullae | EU240886.1.1502 | Cluster_3 | 11405 |
| Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Moraxellaceae;Psychrobacter;Psychrobacter immobilis | U39399.1.1477 | Cluster_4 | 4125 |
| Bacteria;Thermotogae;Thermotogae;Thermotogales;Thermotogaceae;Petrotoga;Petrotoga miotherma | FR733705.1.1499 | Cluster_5 | 4034 |
| Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Phyllobacteriaceae;Pseudahrensia;Pseudahrensia aquimaris | GU575117.1.1441 | Cluster_6 | 3966 |
| Bacteria;Bacteroidetes;Cytophagia;Cytophagales;Cytophagaceae;Persicitalea;Persicitalea jodogahamensis | multi-subject | Cluster_7 | 2433 |
| Bacteria;Proteobacteria;Deltaproteobacteria;Bdellovibrionales;Bdellovibrionaceae;Bdellovibrio;Multi-affiliation | multi-subject | Cluster_8 | 2268 |

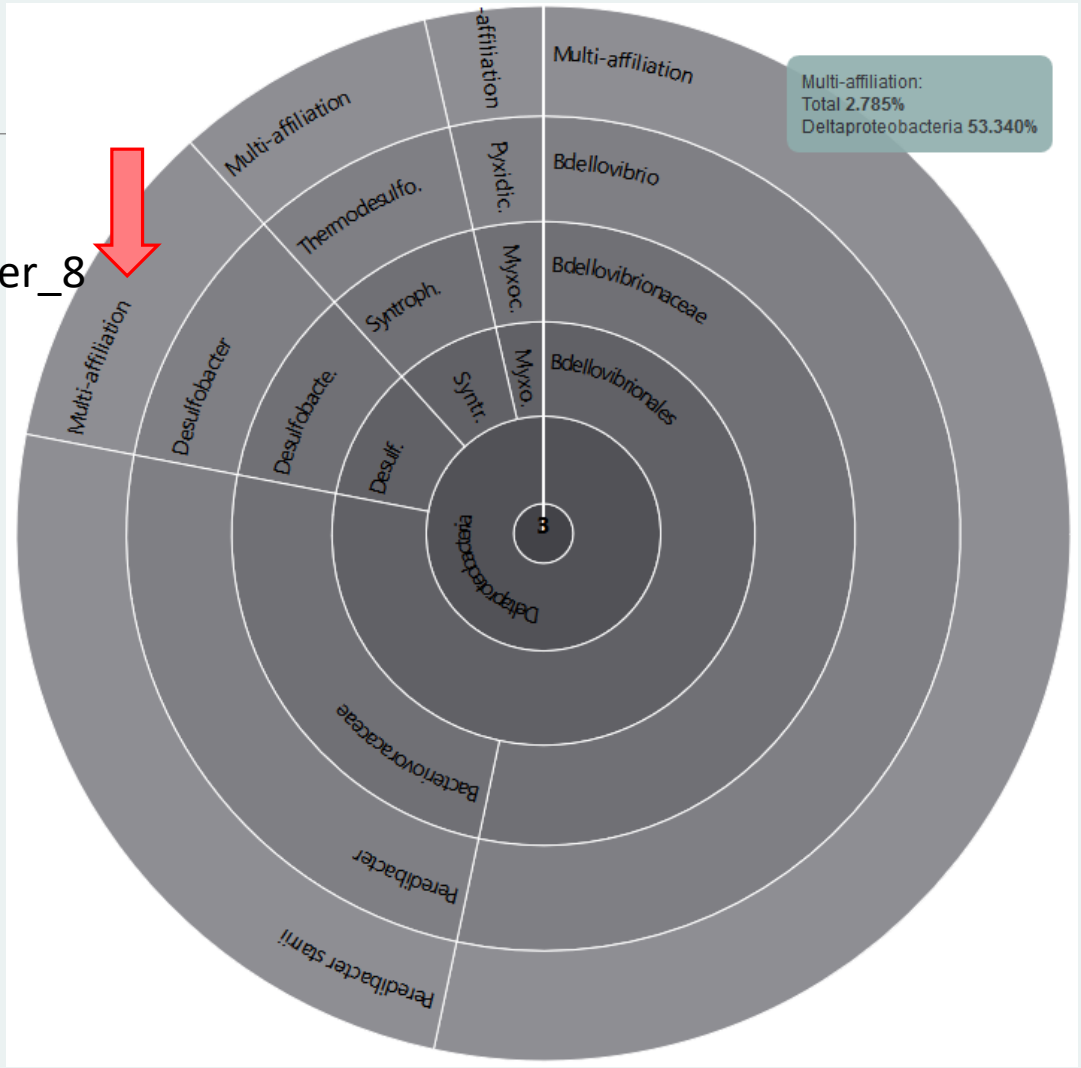
| | | | |
|---|---|---|---|
| Cluster_8 | Bacteria;Proteobacteria;Deltaproteobacteria;Bdellovibrionales;Bdellovibrionaceae;Bdellovibrio | Bdellovibrio bacteriovorus | CP007656.1036900.1038415 |
| Cluster_8 | Bacteria;Proteobacteria;Deltaproteobacteria;Bdellovibrionales;Bdellovibrionaceae;Bdellovibrio | Bdellovibrio bacteriovorus str. Tiberius | CP002930.1837665.1839157 |
| Cluster_8 | Bacteria;Proteobacteria;Deltaproteobacteria;Bdellovibrionales;Bdellovibrionaceae;Bdellovibrio | Bdellovibrio bacteriovorus str. Tiberius | CP002930.842397.843889 |
| Cluster_8 | Bacteria;Proteobacteria;Deltaproteobacteria;Bdellovibrionales;Bdellovibrionaceae;Bdellovibrio | Bdellovibrio bacteriovorus | AJ292760.1.1334 |
| Cluster_8 | Bacteria;Proteobacteria;Deltaproteobacteria;Bdellovibrionales;Bdellovibrionaceae;Bdellovibrio | Bdellovibrio bacteriovorus | |
| Cluster_8 | Bacteria;Proteobacteria;Deltaproteobacteria;Bdellovibrionales;Bdellovibrionaceae;Bdellovibrio | Bdellovibrio bacteriovorus | |
| Cluster_8 | Bacteria;Proteobacteria;Deltaproteobacteria;Bdellovibrionales;Bdellovibrionaceae;Bdellovibrio | Bdellovibrio bacteriovorus | AF084850.1.1436 |
| Cluster_8 | Bacteria;Proteobacteria;Deltaproteobacteria;Bdellovibrionales;Bdellovibrionaceae;Bdellovibrio | Bdellovibrio bacteriovorus HD100 | BX842648.123565.125058 |
| Cluster_8 | Bacteria;Proteobacteria;Deltaproteobacteria;Bdellovibrionales;Bdellovibrionaceae;Bdellovibrio | Bdellovibrio bacteriovorus HD100 | BX842650.295616.297109 |

→ Bdellovibrio bacteriovorus

2. Observe le diversity diagramm




Cluster_8



# Exercise 7

3. How to change affiliation of cluster 8 ????

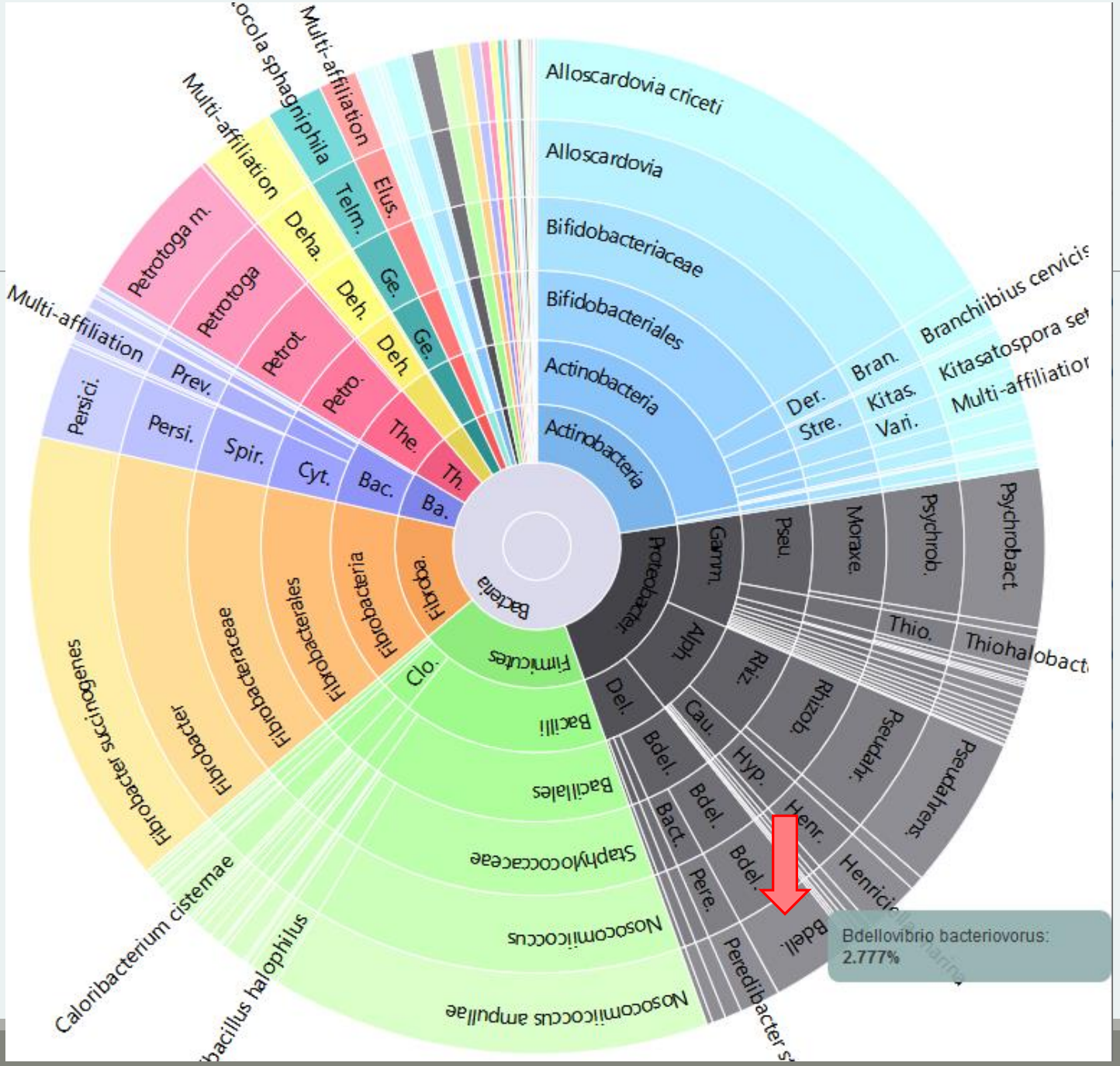
# Exercise 7

4. Modify multi_hit.tsv under excel for example and keep only :

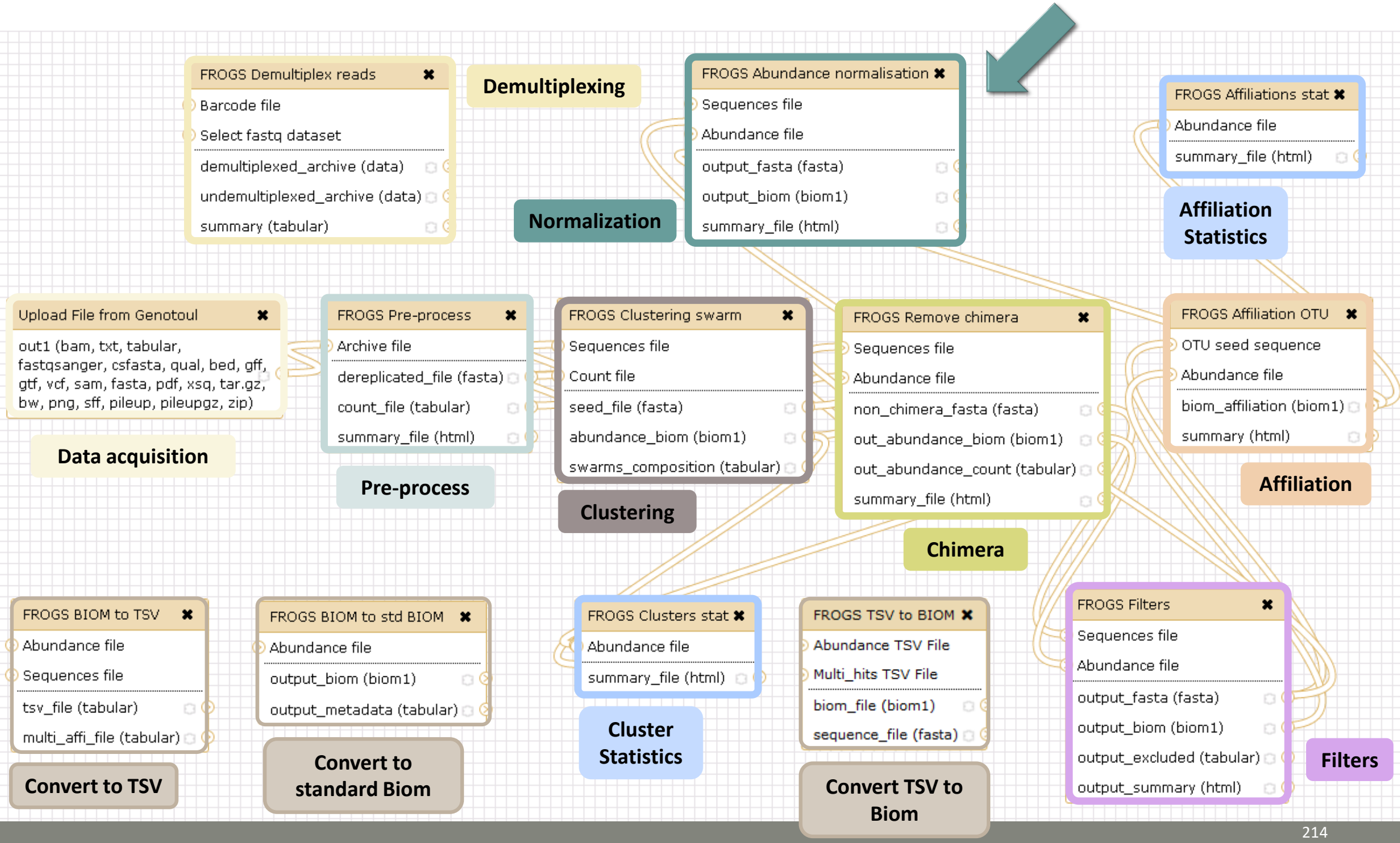
| | | |
|---|---|---|
| Cluster_8 | Bacteria;Proteobacteria;Deltaproteobacteria;Bdellovibrionales;Bdellovibrionaceae;Bdellovibrio;Bdellovibrio bacteriovorus | CP007656.1036900.1038415 |

5. Save in multihit_cluster8_modified.tsv
6. Upload the new multihit file.
7. Create a new biom with a TSV_to_BIOM tool
8. Launch again the affilation_stat tool on this new biom
9. Observe the diversity diagram

# Exercise 7


Alloscardovia criceti
Alloscardovia
Bifidobacteriaceae
Bifidobacteriales
Actinobacteria
Actinobacteria
Bacteria
Proteobacter.
Gamm.
Alph.
Del.
Cau.
Bdel.
Bact.
Pere.
Bdell.
Pseu.
Morax.
Psychrob.
Psychrobact.
Thio.
Thiohalobact.
Rhiz.
Rhizob.
Pseudahr.
Pseudahrens.
Hyp.
Henr.
Henric.
Peredibacter s.
Bdellovibrio bacteriovorus:
2.777%
Firmicutes
Bacilli
Bacillales
Staphylococcaceae
Nosocomiicoccus
Nosocomiicoccus ampullae
Clo.
Caloribacterium cisternae
Bacillus halophilus
Fibrobacter succinogenes
Fibrobacter
Fibrobacteraceae
Fibrobacterales
Fibrobacteria
Fibro.
Persici.
Persi.
Spir.
Cyt.
Bac.
Ba.
Prev.
Multi-affiliation
Petrotoga m.
Petrotoga
Petrot.
Petro.
The.
Th.
Deh.
Deh.
Deha.
Multi-affiliation
Ge.
Ge.
Telm.
Elus.
Multi-affiliation
ocola sphagniphila
Branchiibius cervicis
Kitasatospora se
Multi-affiliation
Der.
Stre.
Bran.
Kitas.
Vari.

# 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 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

# 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

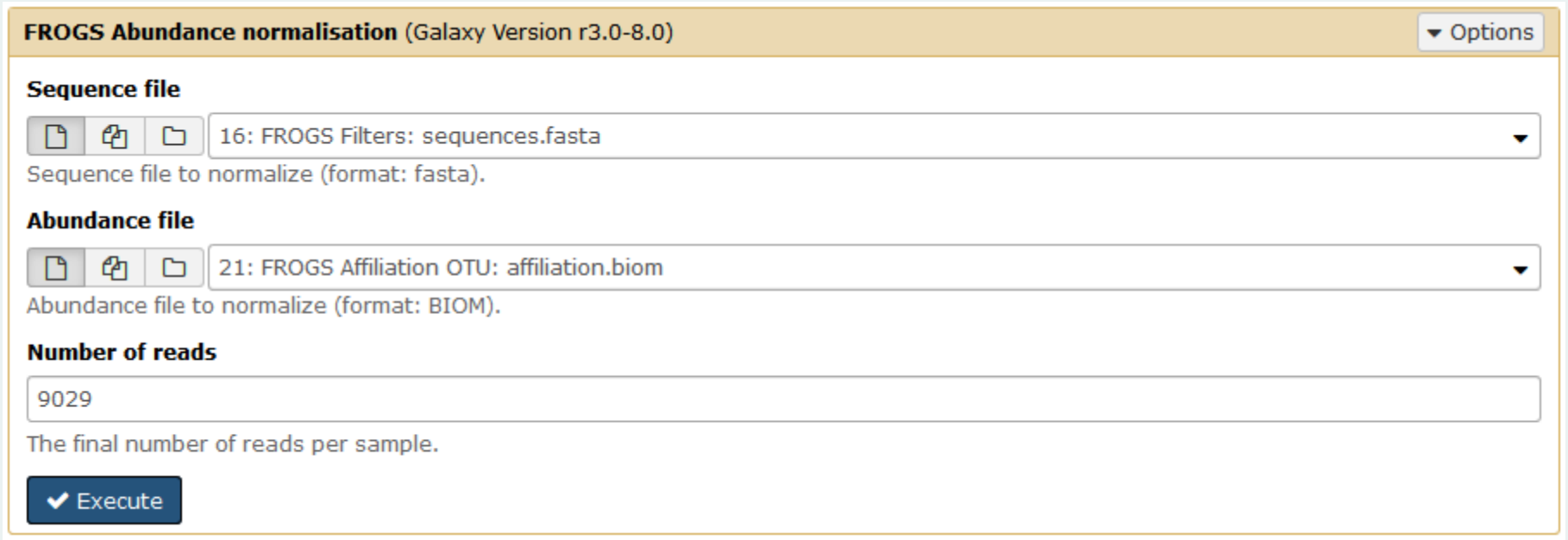
# Your Turn! – 8

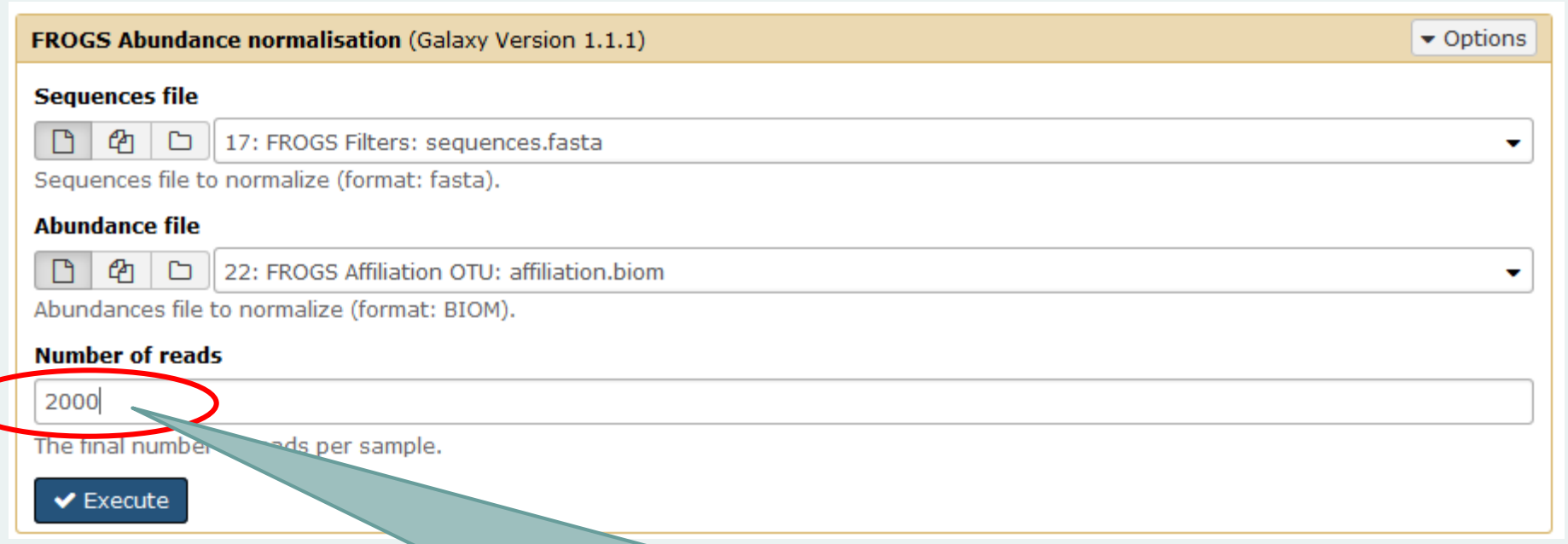
LAUNCH NORMALIZATION TOOL

# Exercise 8

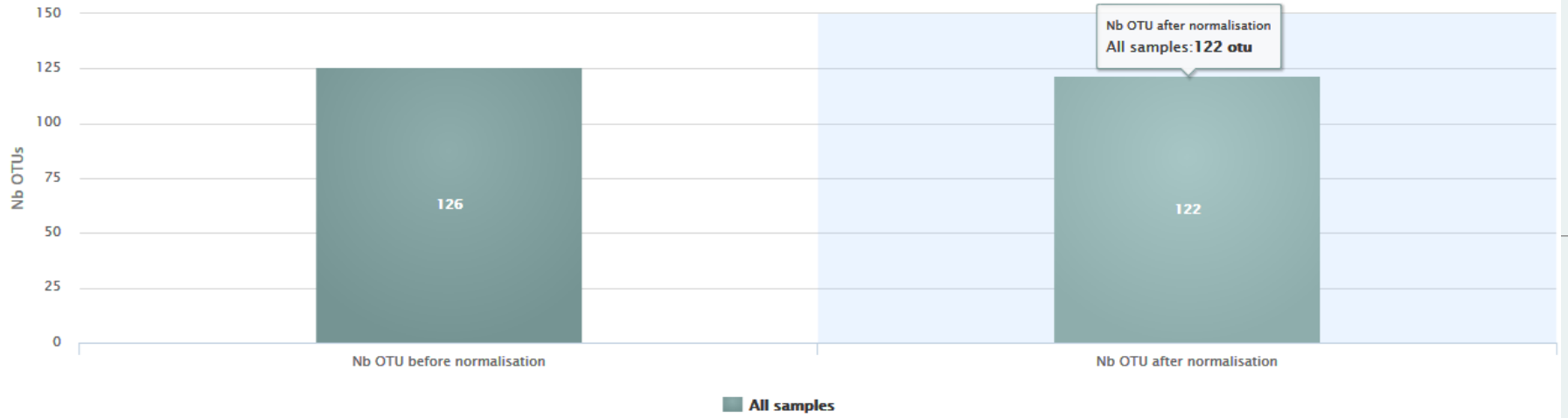
Launch Normalization Tool

1. What is the smallest sequenced samples ?
2. Normalize your data from Affiliation based on this number of sequence
3. Explore the report HTML result.
4. Try other threshold and explore the report HTML result
   What do you remark ?


FROGS Abundance normalisation (Galaxy Version r3.0-8.0)
Options
Sequence file
16: FROGS Filters: sequences.fasta
Sequence file to normalize (format: fasta).
Abundance file
21: FROGS Affiliation OTU: affiliation.biom
Abundance file to normalize (format: BIOM).
Number of reads
9029
The final number of reads per sample.
Execute




Or, this number can be chosen according to the rarefaction curve. For example, we can choose the smallest number of sequences that still retain all the genus.




CSV

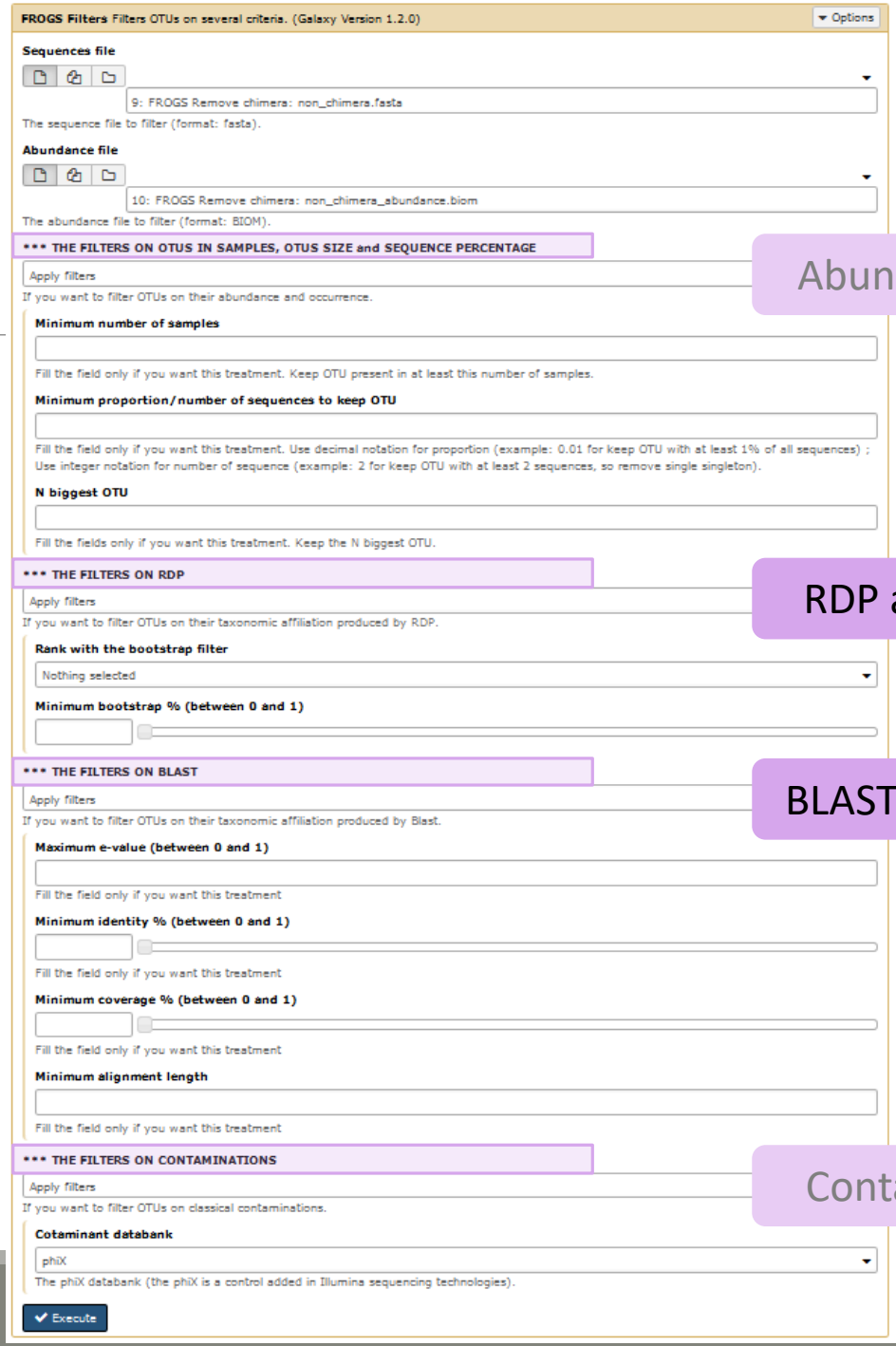
Show 10 entries

Search:

| Sample | Nb OTU before normalisation | Nb OTU after normalisation |
|---|---|---|
| 100_10000seq_sampleA1 | 126 | 122 |
| 100_10000seq_sampleA2 | 126 | 122 |
| 100_10000seq_sampleA3 | 126 | 122 |
| 100_10000seq_sampleB1 | 126 | 122 |
| 100_10000seq_sampleB2 | 126 | 122 |
| 100_10000seq_sampleB3 | 126 | 122 |
| 100_10000seq_sampleC1 | 126 | 122 |
| 100_10000seq_sampleC2 | 126 | 122 |
| 100_10000seq_sampleC3 | 126 | 122 |

# Filters on affiliations

Do not forget, with filter tool we can filter the data based on their affiliation




Abundance filters

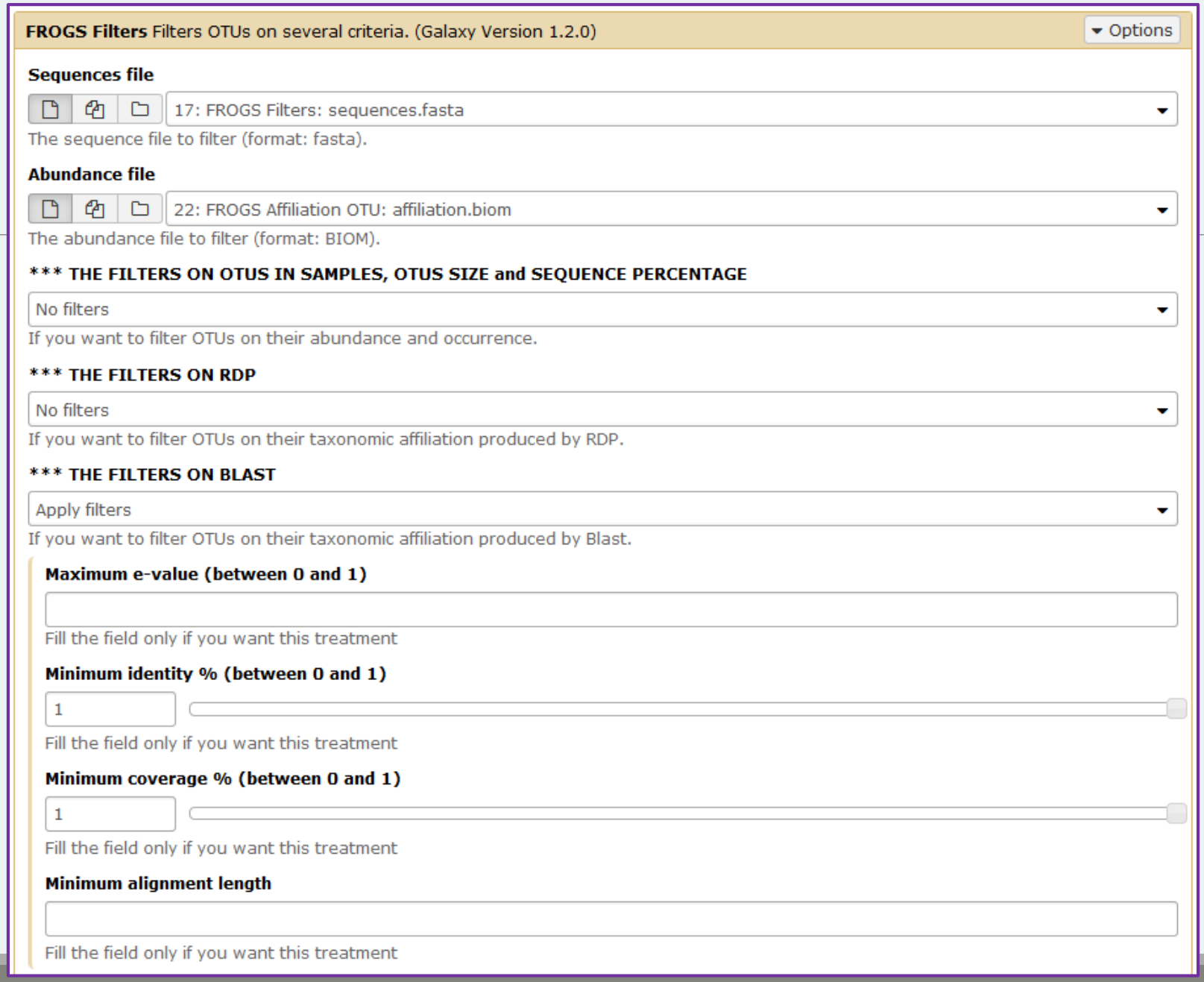
RDP affiliation filters

BLAST affiliation filters

Contamination filter

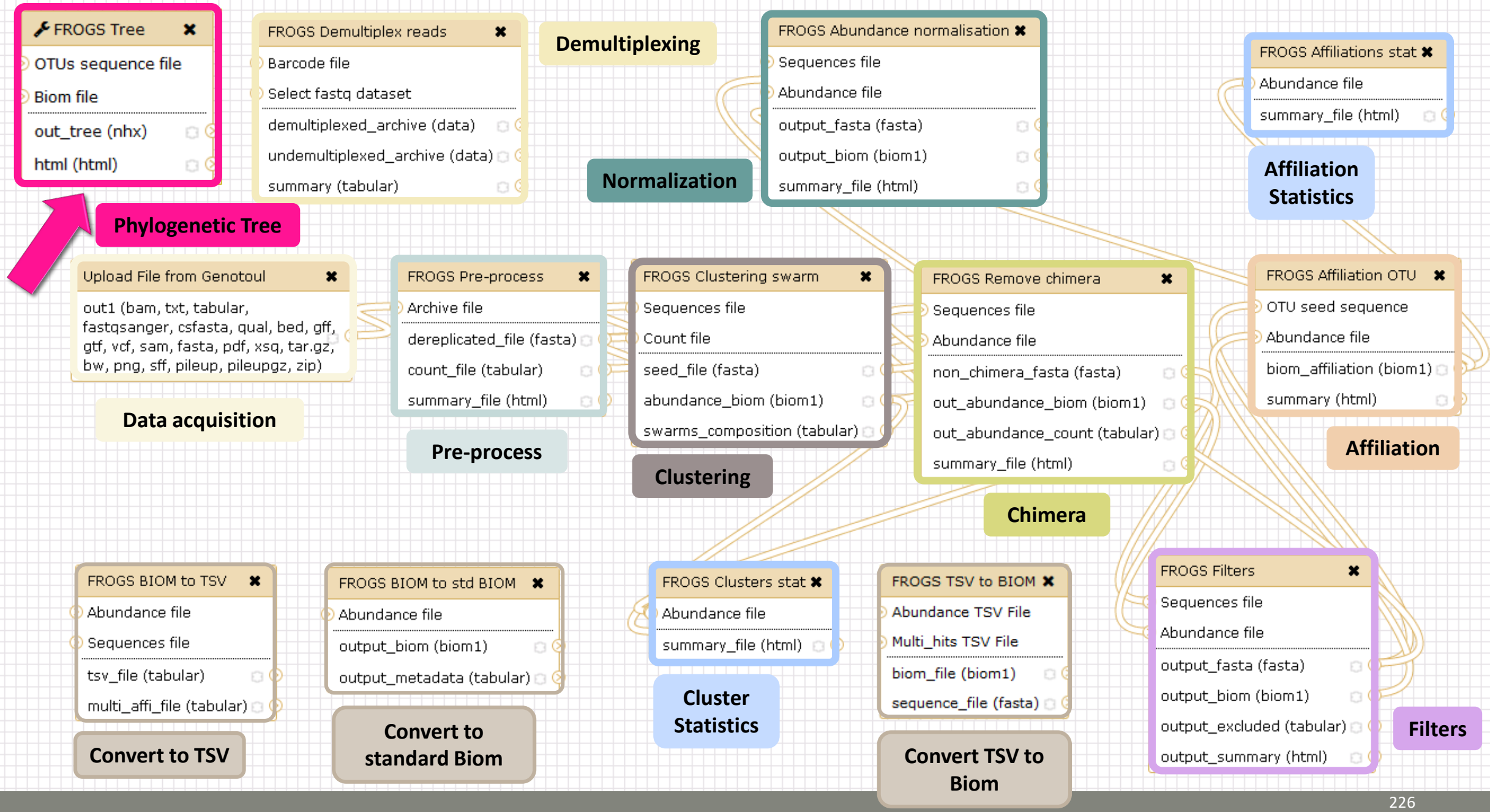
# Exercise 9

1. Apply filters to keep only data with perfect alignment.
2. How many clusters have you keep ?

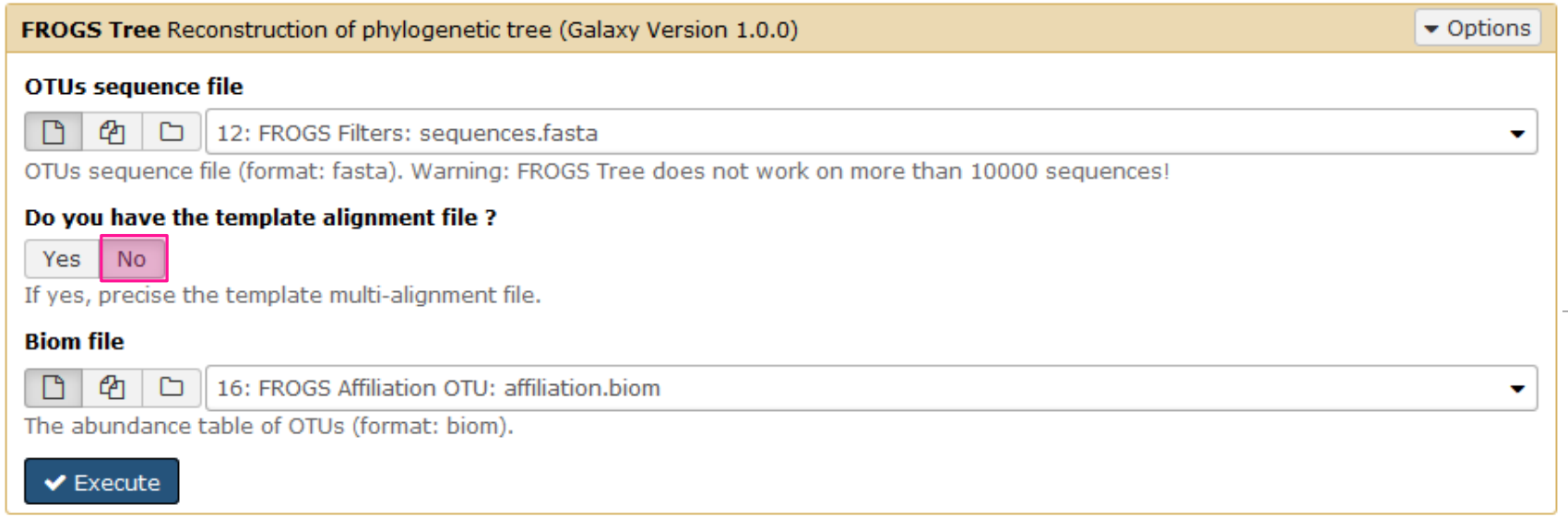

FROGS Filters Filters OTUs on several criteria. (Galaxy Version 1.2.0)
Options
Sequences file
17: FROGS Filters: sequences.fasta
The sequence file to filter (format: fasta).
Abundance file
22: FROGS Affiliation OTU: affiliation.biom
The abundance file to filter (format: BIOM).
*** THE FILTERS ON OTUS IN SAMPLES, OTUS SIZE and SEQUENCE PERCENTAGE
No filters
If you want to filter OTUs on their abundance and occurrence.
*** THE FILTERS ON RDP
No filters
If you want to filter OTUs on their taxonomic affiliation produced by RDP.
*** 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)
1
Fill the field only if you want this treatment
Minimum coverage % (between 0 and 1)
1
Fill the field only if you want this treatment
Minimum alignment length
Fill the field only if you want this treatment

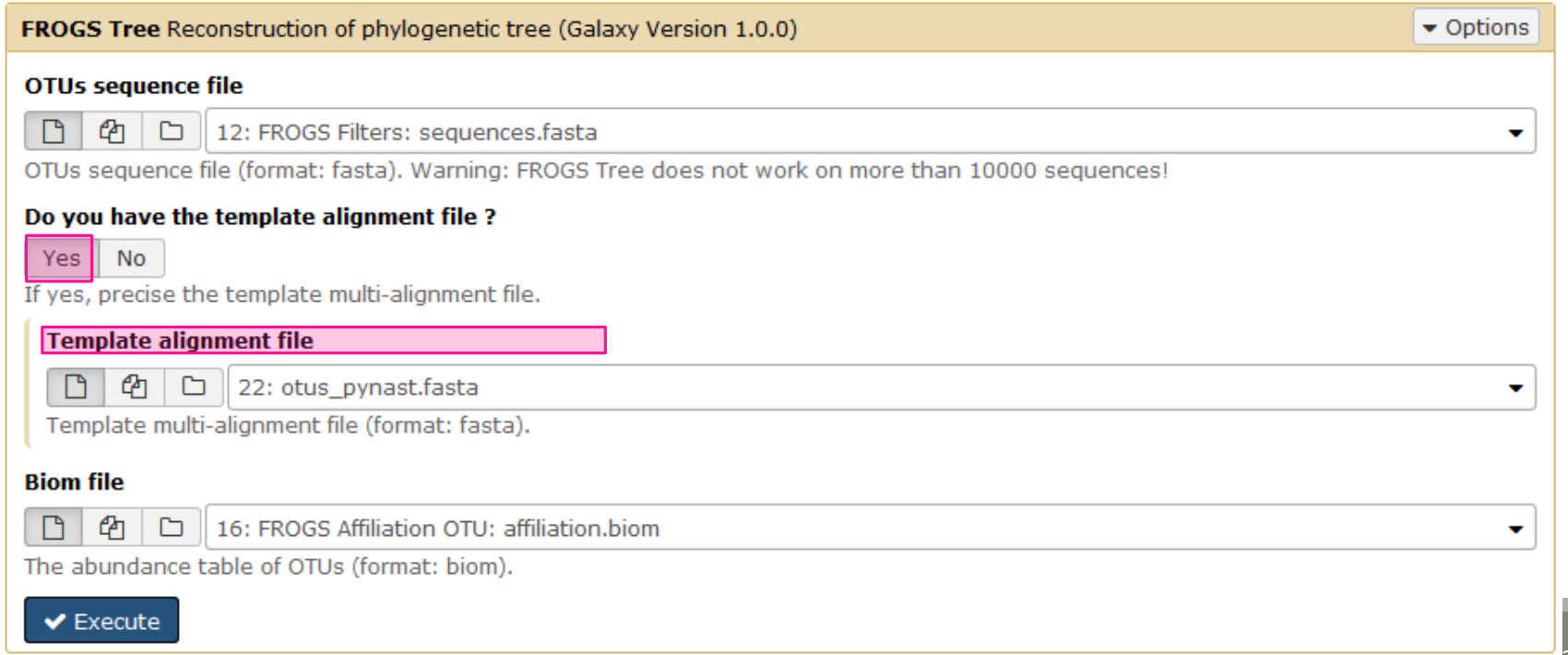
# FROGS Tree

CREATE A PHYLOGENETICS TREE OF OTUS


FROGS Tree
OTUs sequence file
Biom file
out_tree (nhx)
html (html)
Phylogenetic Tree
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

2 choices to do your phylogenetics tree



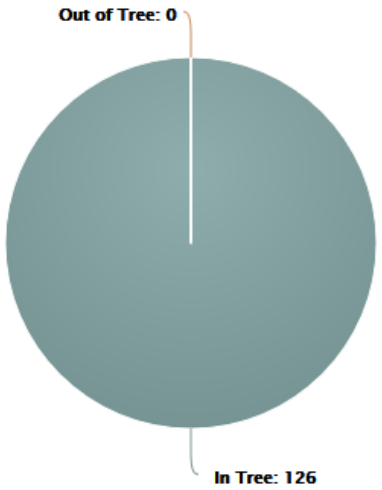



# Exercise 9

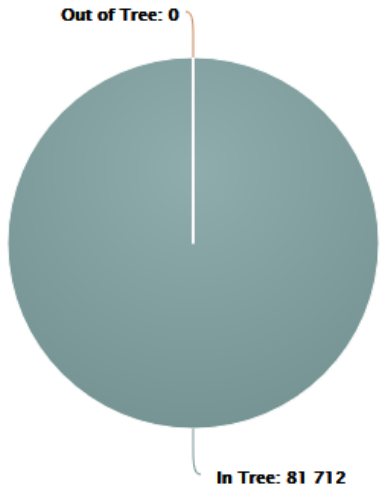
1. Create a tree with the filtered OTUs without template
2. Explore the HTML file
3. Look tree.nwk

# Summary

OTUs



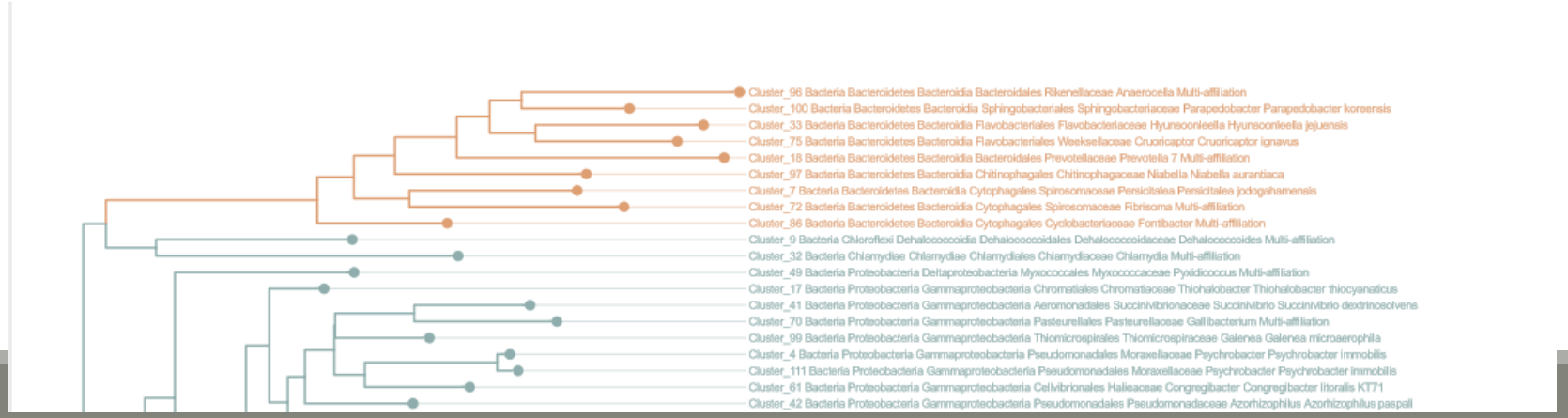

Abundance




# Tree View

Enabling zoom:

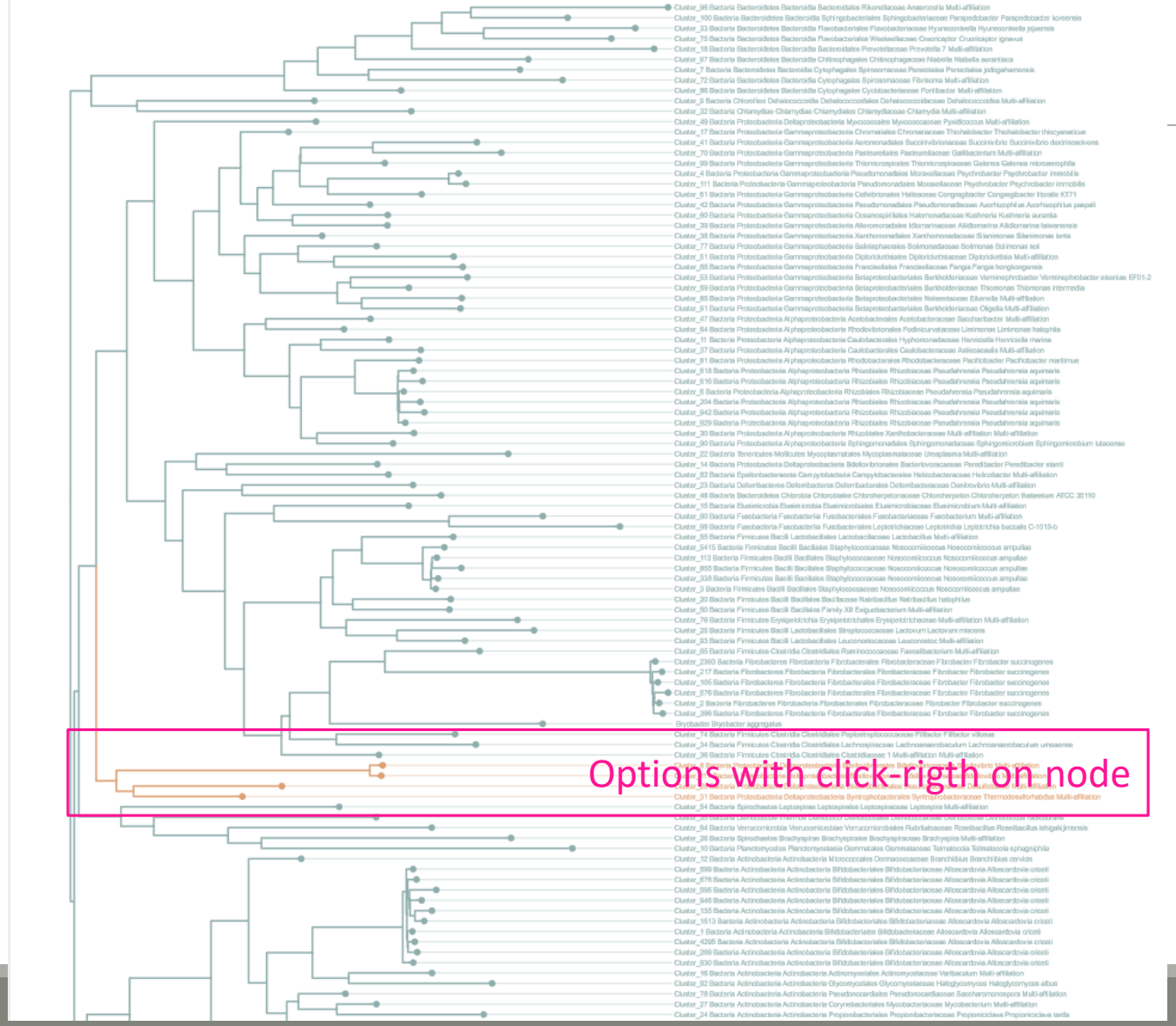
ON



# Tree View

Enabling zoom:

ON



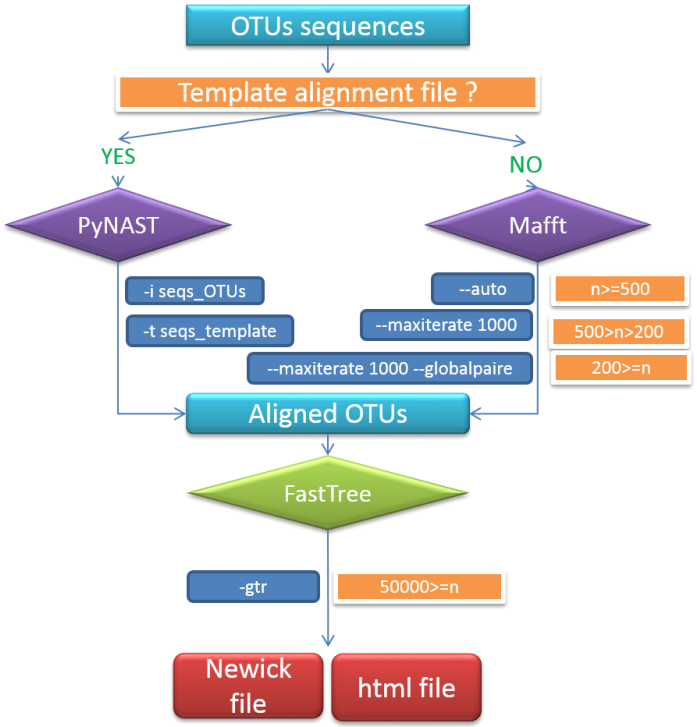

Tree.nwk:

((Cluster_8 Bacteria Proteobacteria Deltaproteobacteria Bdellovibrionales Bdellovibrionaceae Bdellovibrio Multi-affiliation:0.00879,Cluster_117 Bacteria Proteobacteria Deltaproteobacteria Bdellovibrionales Bdellovibrionaceae Bdellovibrio Multi-affiliation:0.00744):0.25827,(Cluster_28 Bacteria Proteobacteria Deltaproteobacteria Desulfobacterales Desulfobacteraceae Desulfobacter Multi-affiliation:0.14675,Cluster_31 Bacteria Proteobacteria Deltaproteobacteria Syntrophobacterales Syntrophobacteraceae Thermodesulforhabdus Multi-affiliation:0.10644):0.01759):0.02059;

# How works FROGS TREE ?

Pynast needs alignment template to go fast

But if your species is not similar at 75% with a sequence in the template, your species will be not in the tree !




To find templates:

Based on 16S GreenGenes databank
https://github.com/biocore/qiime-default-reference/blob/master/qiime_default_reference/gg_13_8_otus/rep_set_aligned/85_otus.pynast.fasta.gz

Based on 16S SILVA databank
https://www.arb-silva.de/fileadmin/silva_databases/qiime/Silva_128_release.tgz

# Tool descriptions

# Example of Preprocess tool HELP


## What it does

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

## Inputs/Outputs

### Inputs

Sample files added one after another or provide in an archive file (tar.gz).

**Illumina inputs**

> **Usage:** For samples sequenced in paired-end. The amplicon length must be inferior to the length of the R1 plus R2 length. R1 and R2 are merged 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:** For samples sequenced in single-ends or when R1 and R2 reads are already merged.
> **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

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 represent

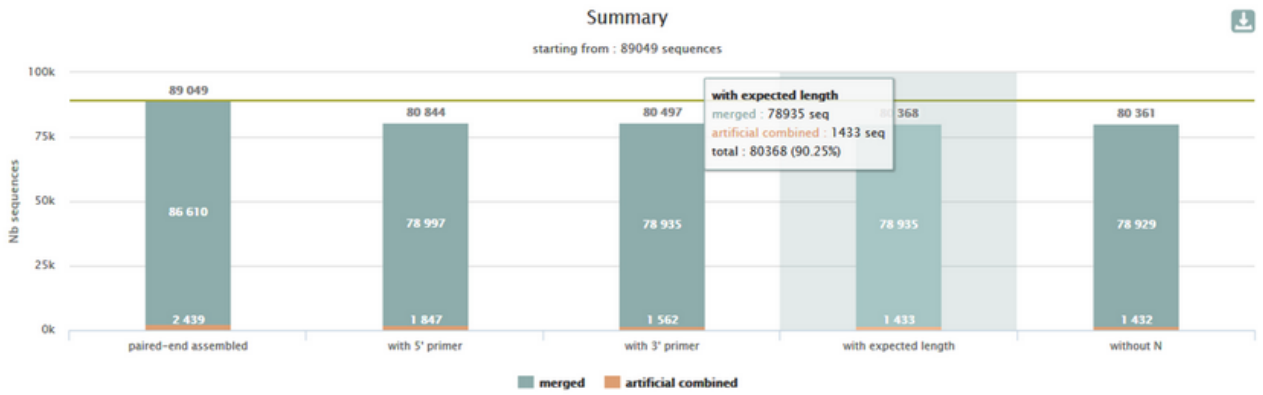
**Count file** (count.tsv):

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

**Summary file** (report.html):

> This file reports the number of remaining sequences after each filter (format HTML).

### Preprocess summary




### Details on merged sequences

CSV

Show 10 entries Search:

| Samples | % kept | paired-end assembled | with 5' primer | with 3' primer | with expected length | without N |
|---|---|---|---|---|---|---|
| echantillon1-1 | 84.93 | 31,836 | 27,059 | 27,040 | 27,040 | 27,039 |
| echantillon1-2 | 94.73 | 54,774 | 51,938 | 51,895 | 51,895 | 51,890 |

**With selection:** Display amplicon lengths | Display preprocessed amplicon lengths

Showing 1 to 2 of 2 entries Previous Next

## How it works

| Steps | Illumina | 454 |
|---|---|---|
| 1 | For un-merged data: merges R1 and R2 with a maximum of M% mismatch in the overlaped region(VSEARCH or FLASH or optionnaly PEAR). Resulting un-merged reads may optionnaly be artificially combined by adding 100 N between the reads | / |
| 2 | If sequencing protocol is the illumina standard protocol : Removes sequences where the two primers are not present and then remove primers in the remaining sequence (cutadapt). The primer search accepts 10% of differences | Removes sequences where the two primers are not present, removes primers sequence and reverse complement the sequences on strand - (cutadapt). The primer search accepts 10% of differences |
| 3 | Filters sequences with ambiguous nucleotides and for merged sequences filters on their length which must be range between 'Minimum amplicon size - primer length' and 'Maximum amplicon size - primer length' | the tool removes sequences with at least one homopolymer with more than seven nucleotides and with a distance of less than or equal to 10 nucleo-tides between two poor quality positions, i.e. with a Phred quality score lesser than 10 |
| 4 | Dereplicates sequences | Dereplicates sequences |

# Advices/details on parameters

## Primers parameters

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

Example:

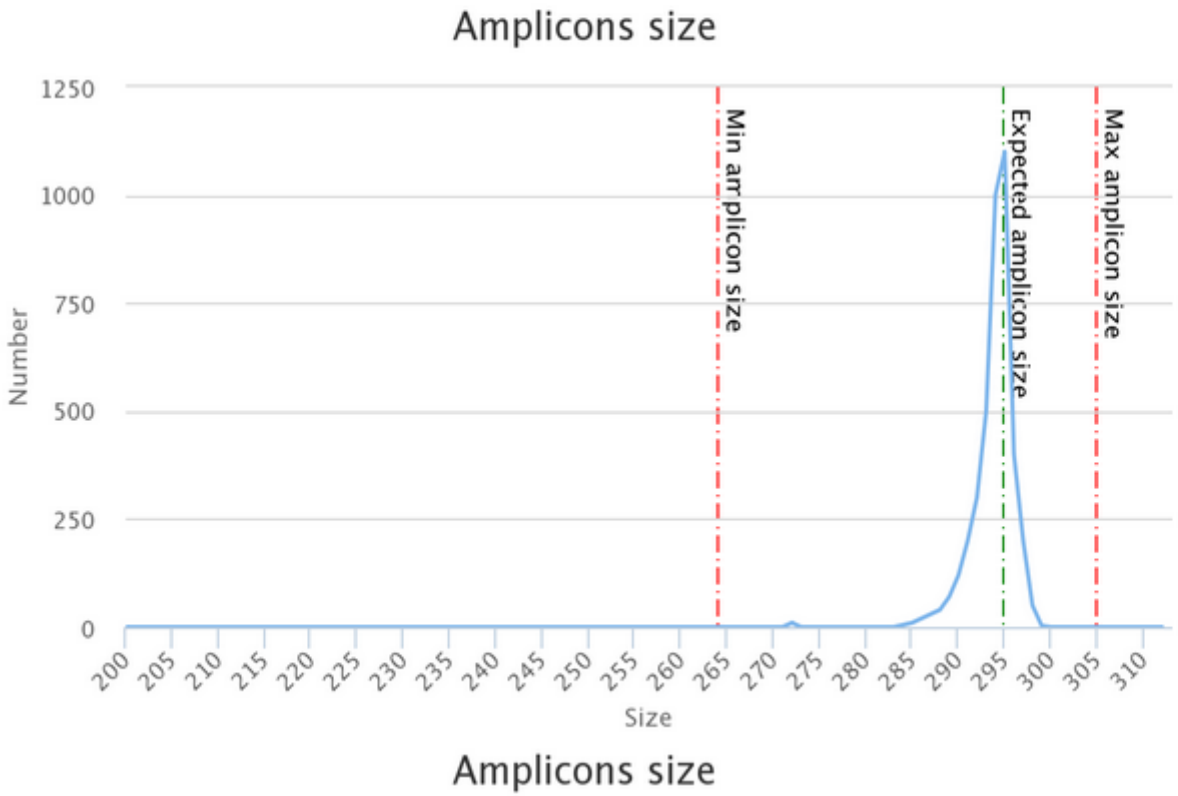
5' ATGCCC GTCGTCGTAAAATGC ATTTCAG 3'

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.



## Advices/details on parameters

### What is the difference between overlapped sequences and combined sequences?

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

Imagine a real amplicon sequence of 400bp


Imagine a Miseq paired sequencing of 2x250bp


Reconstructing amplicon sequence is possible thanks to the overlap region

Merged sequence length : 400bp, with 100bp overlap

**Case of a sequencing of non-overlapping sequences: case of ITS1 amplicon MiSeq sequencing**

Imagine a real amplicon sequence of 700bp


Imagine a Miseq paired sequencing of 2x250bp


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


Combined sequence length : 600bp, with 100 Ns

**"FROGS combined" warning points**

Reads pair are not merged because:

> the real amplicon length is greater than de number of base sequences (500 bp for MiSeq 2x250bp)
> the overlapped region is smaller than 10 (fixed parameter in FROGS).

Thus, "FROGS combined" sequences are artificial and present particular features especially on size. Imagine a MiSeq sequencing of 2x25
sequences length will be 600 bp.

Case 1: real amplicon > 601 bp → "FROGS combined" length is smaller than the reality

**Contact**

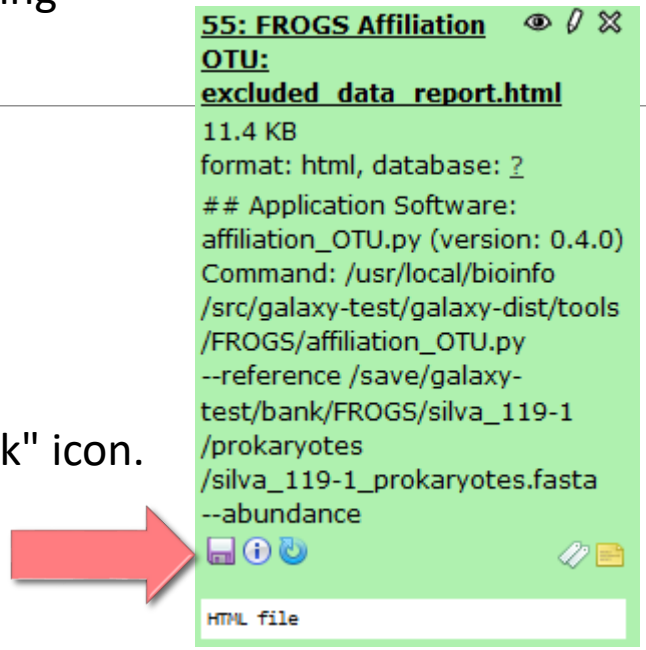
Contacts: frogs@inra.fr

Repository: https://github.com/geraldinepascal/FROGS website: http://frogs.toulouse.inra.fr/

Please cite the **FROGS article**: *Escudie F., et al. Bioinformatics, 2018. FROGS: Find, Rapidly, OTUs with Galaxy Solution.*

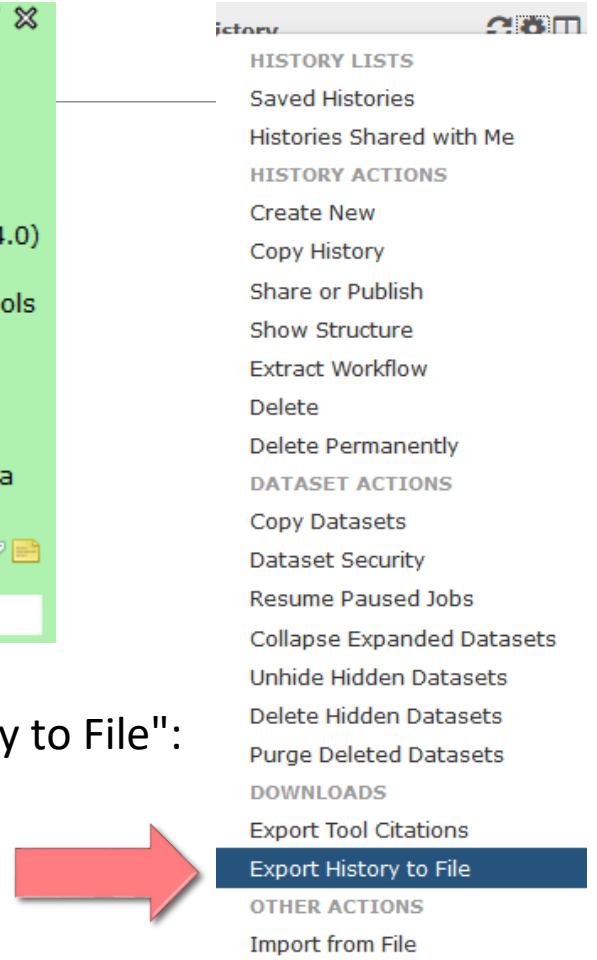
# Download your data

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




You have the opportunity:
1/ Save your datasets one by one using the "floppy disk" icon.

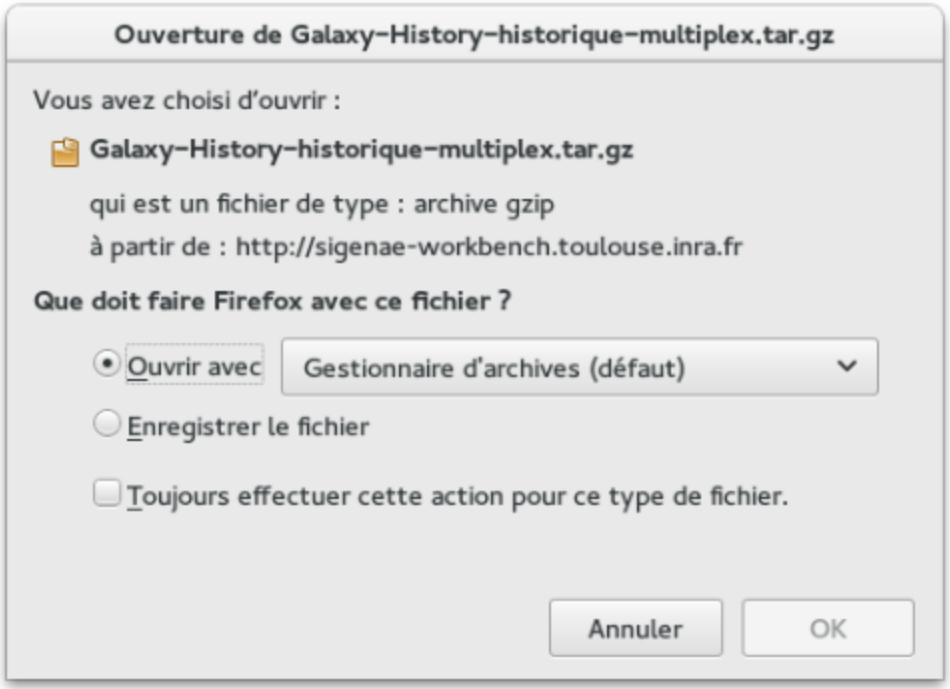
2/ Or export each history.
To export a history, from the "History" menu, click on the wheel, then "Export History to File":



To retrieve your history, click on the http link that appears automatically:

---

It is then possible to record the data :




This directory contains :


1. in the "datasets" directory: Your Galaxy files.
2. in the files "-attrs.txt" : Metadata about your datasets, your jobs and your history.

# FROGS BIOM to Standard BIOM

# FROGS biom to standard Biom

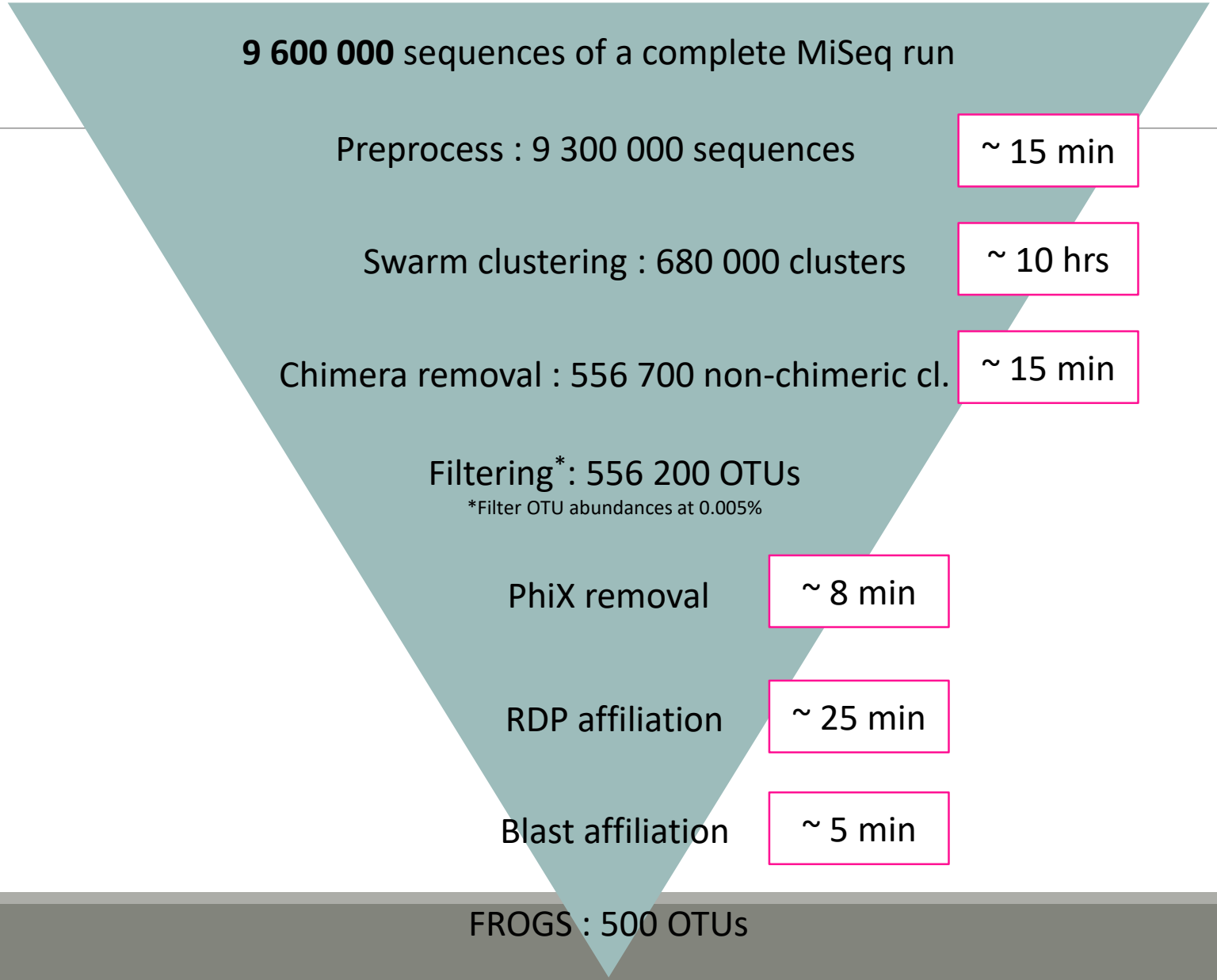
This step is required to run R

# 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 |

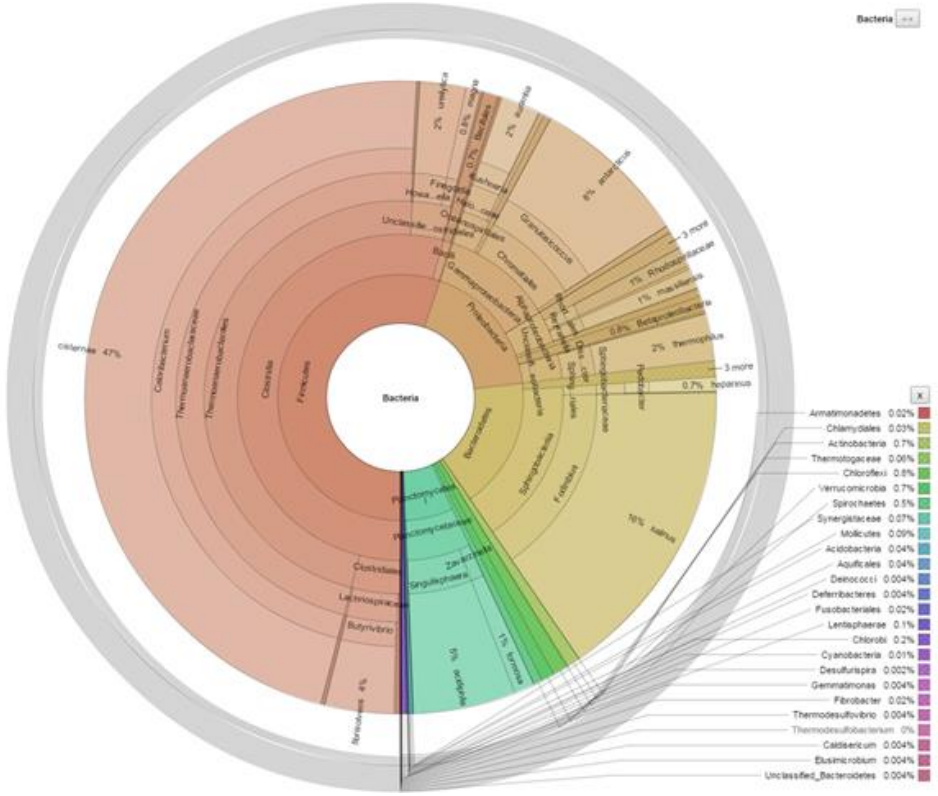
# Speed on real datasets with filter




**Escape statistics on assessments**

# 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**

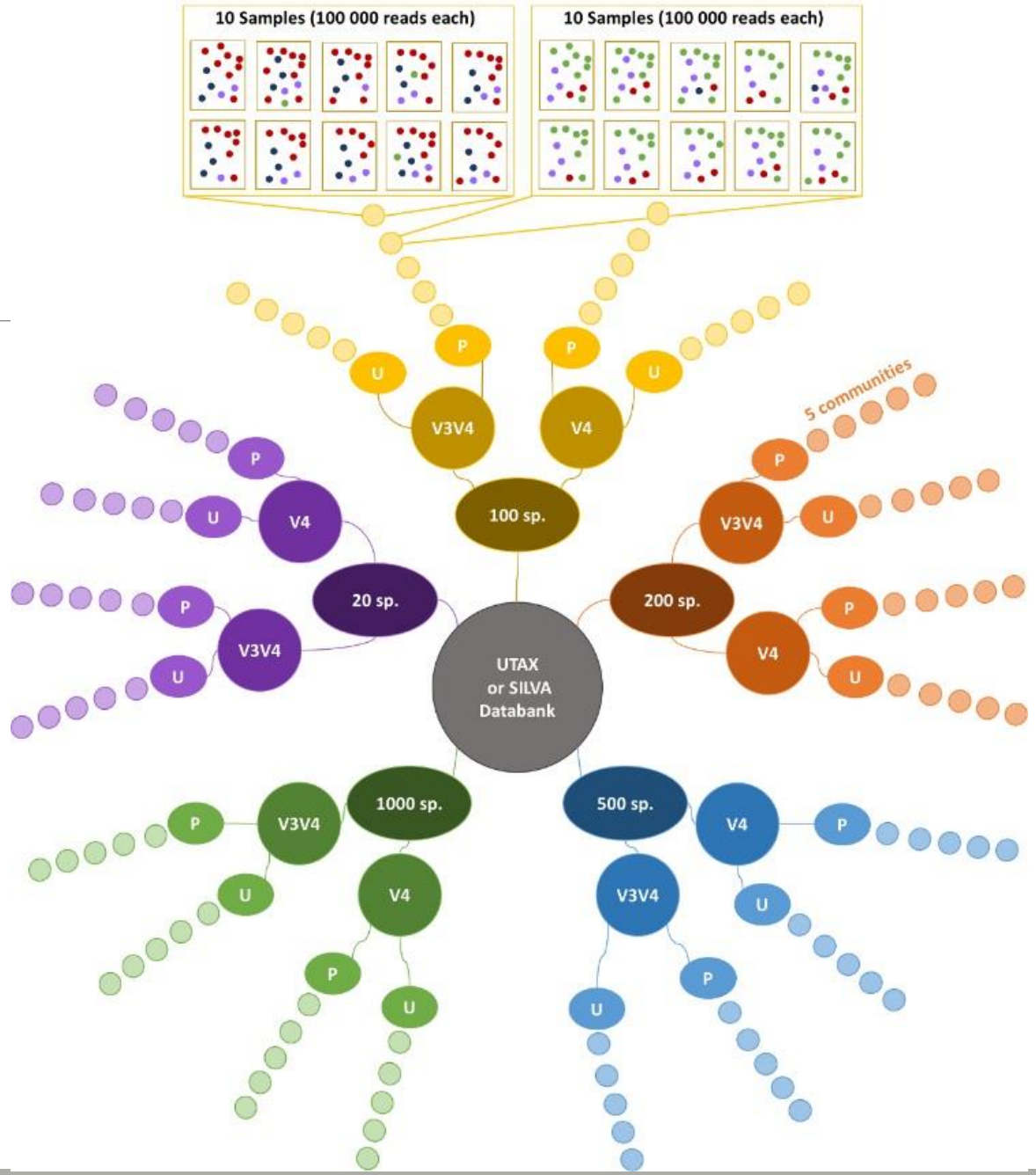
↓

**filters : 0.005%** **505 OTUs**

# FROGS' Accuracy

- $1.10^{+8}$ synthetic sequences were treated with FROGS, UPARSE and MOTHUR, QIIME, with their guidelines, to compare their performances
- 20, 100, 200, 500 or 1000 different species
- power law or a uniform distribution
- 5 to 20% of chimera

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



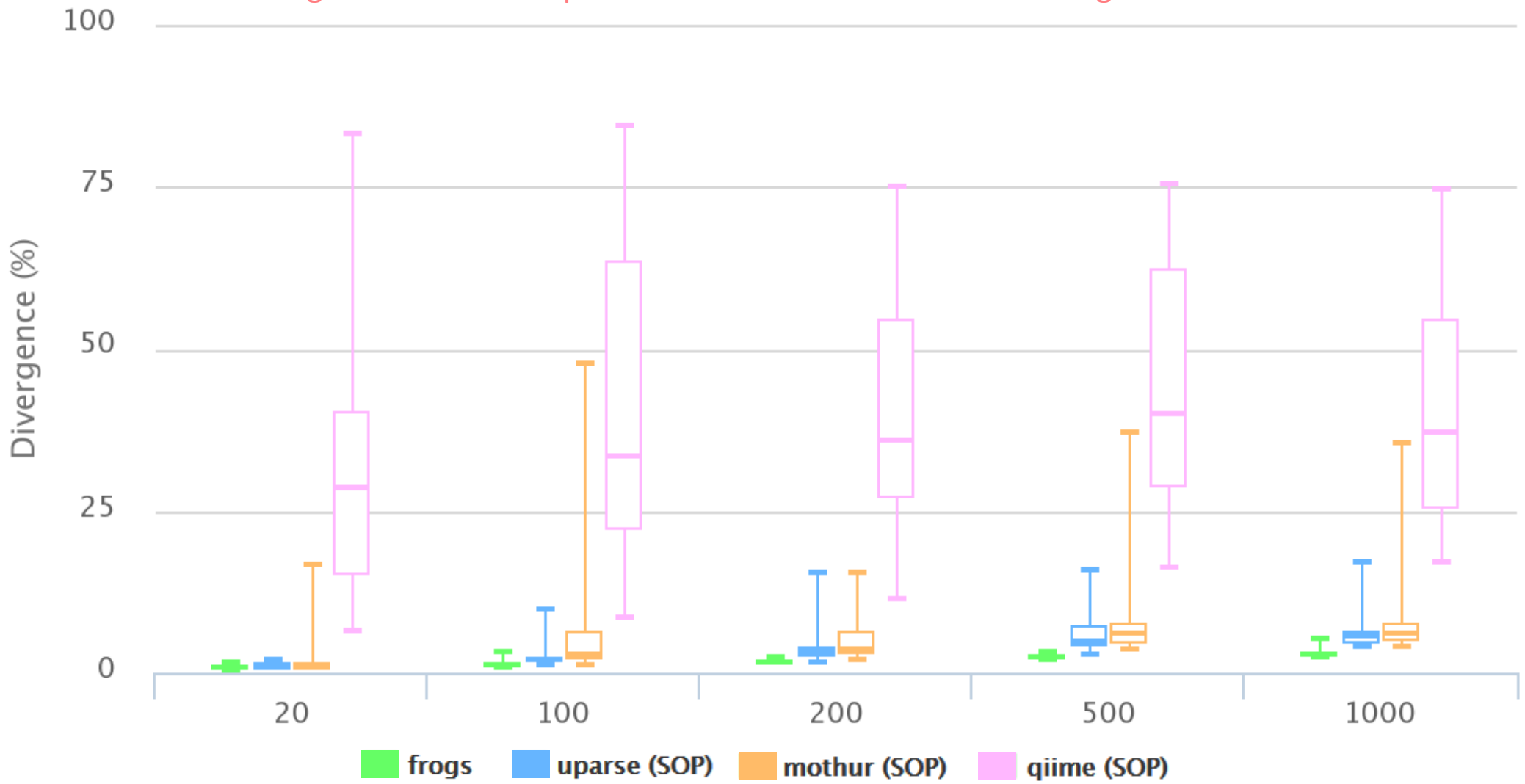
# FROGS' Accuracy

The four metrics used to compare results of FROGS, UPARSE, QIIME and MOTHUR are :

V3V4 Power Law

# Affiliations divergence

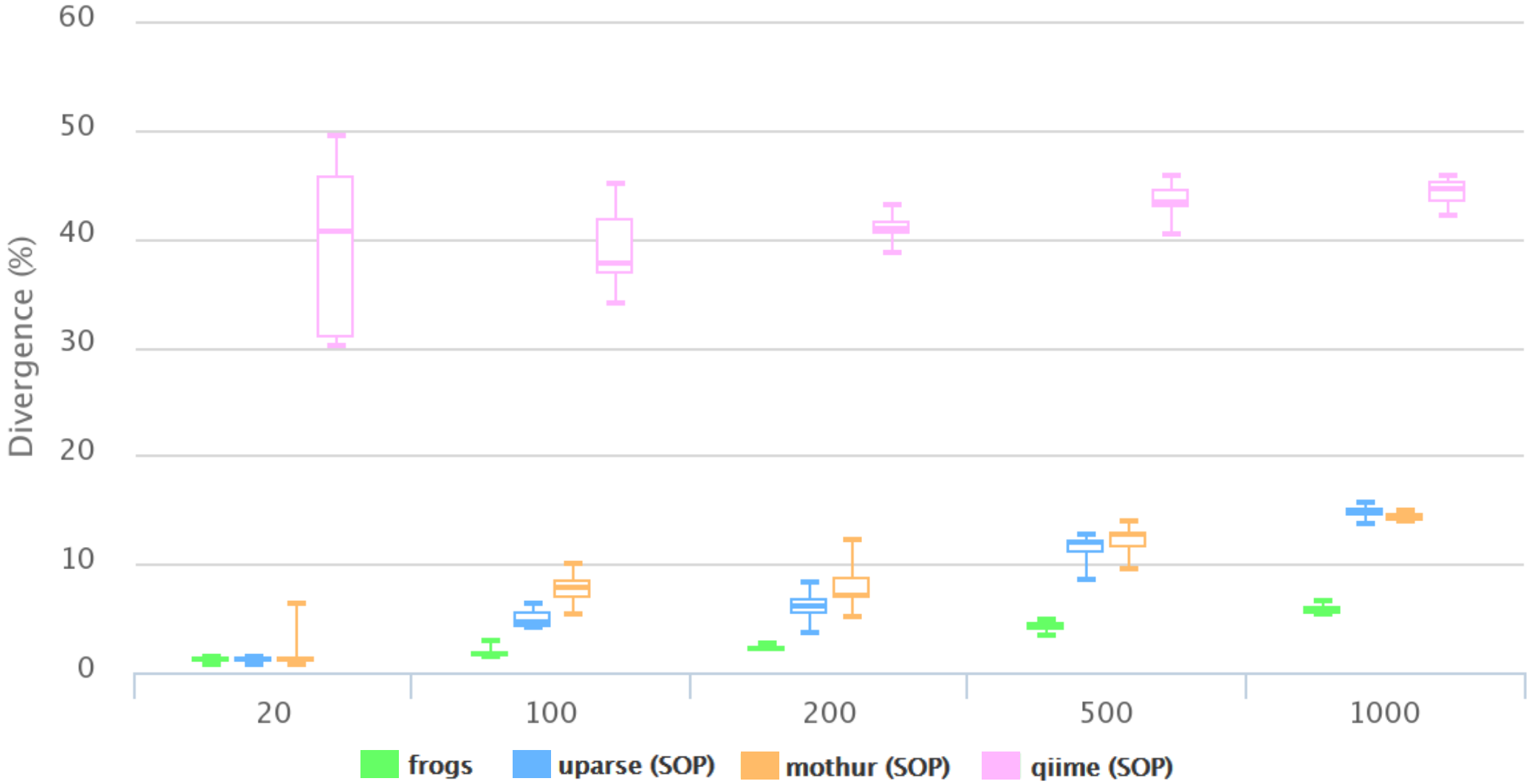
Divergence on the composition of microbial communities at genus rank



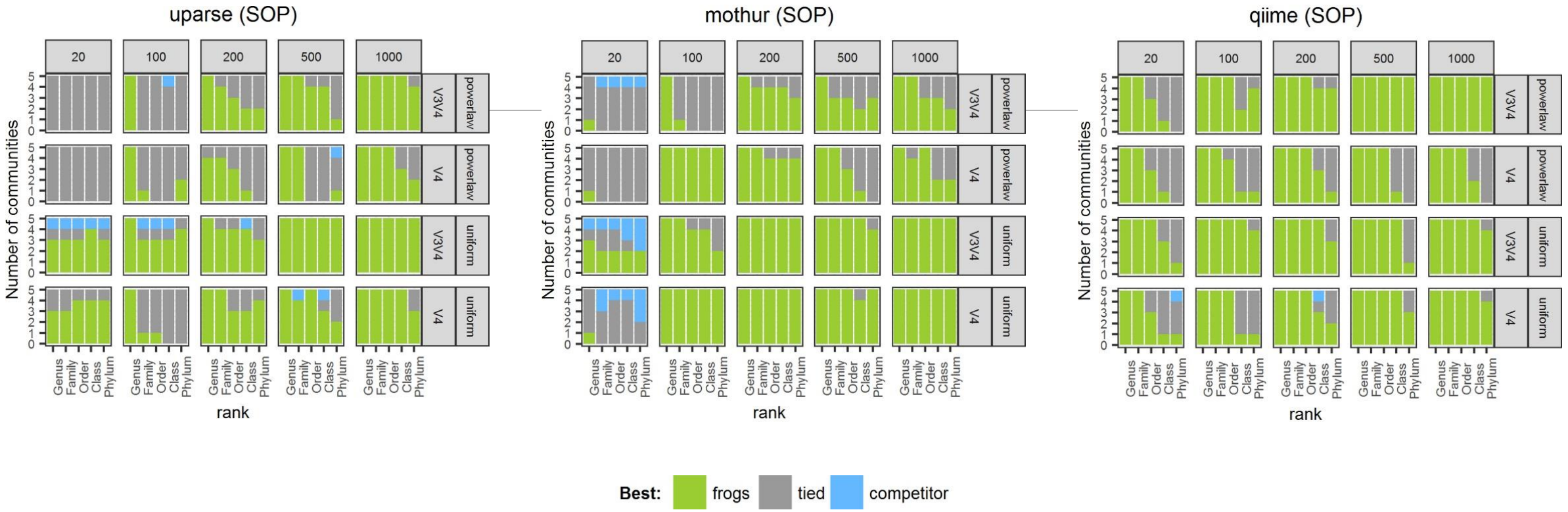
V3V4 Uniform

# Affiliations divergence

Divergence on the composition of microbial communities at genus rank



The results of non-parametric paired tests (signed rank test) of Affiliation divergence on simulated data from UTAX

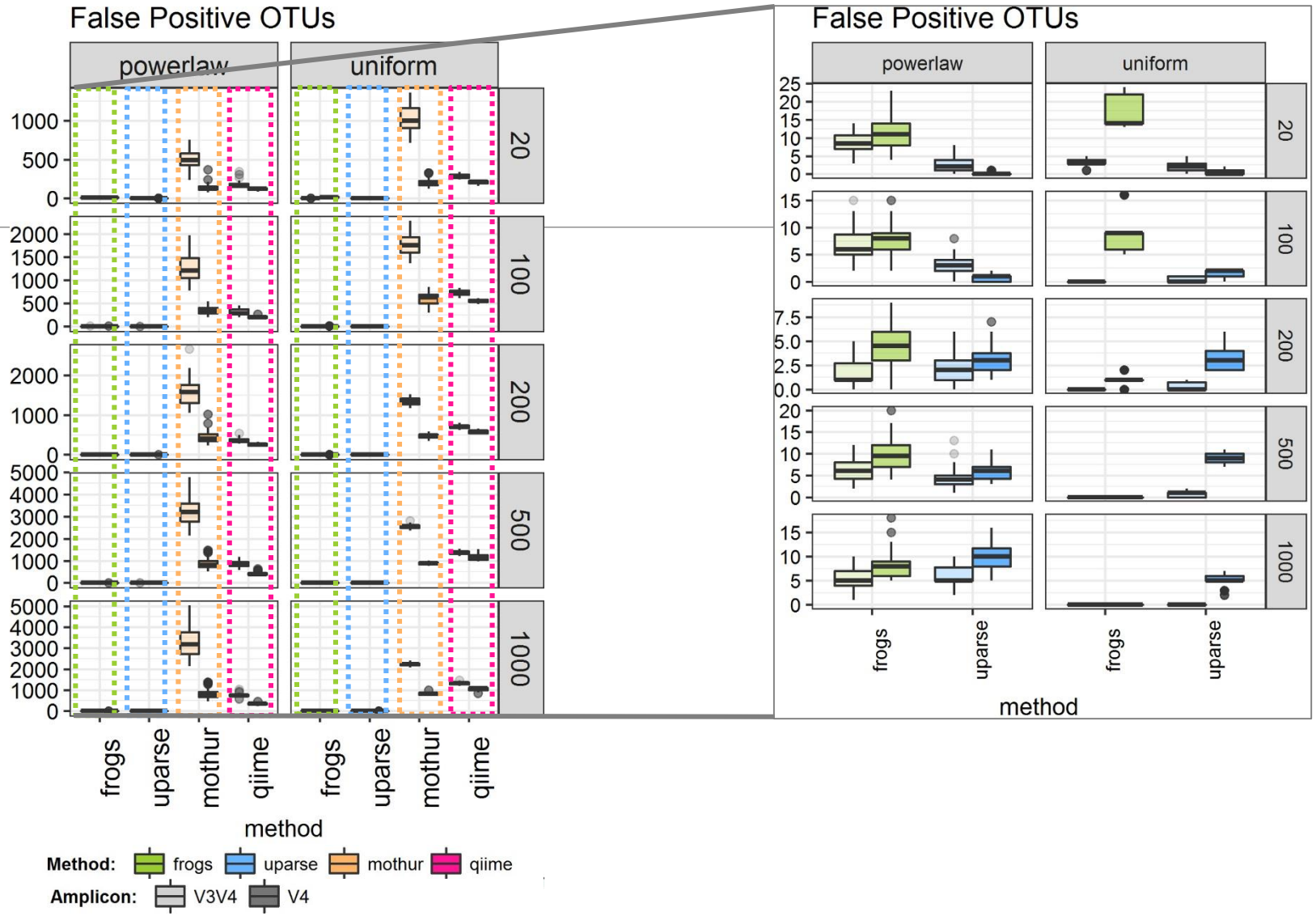



FROGS performed as well as or better than UPARSE and mothur in most settings. The infrequent condition in which FROGS performed worse than UPARSE and mothur was for small community sizes (20 species), except at genus level. It performed better than QIIME in all settings.

Huge number of FP inferred by mothur (up to 20 times more than the expected community size).

a few more FPs under power law abundance distributions and a few less under uniform abundance distributions (except for size < 100 species)

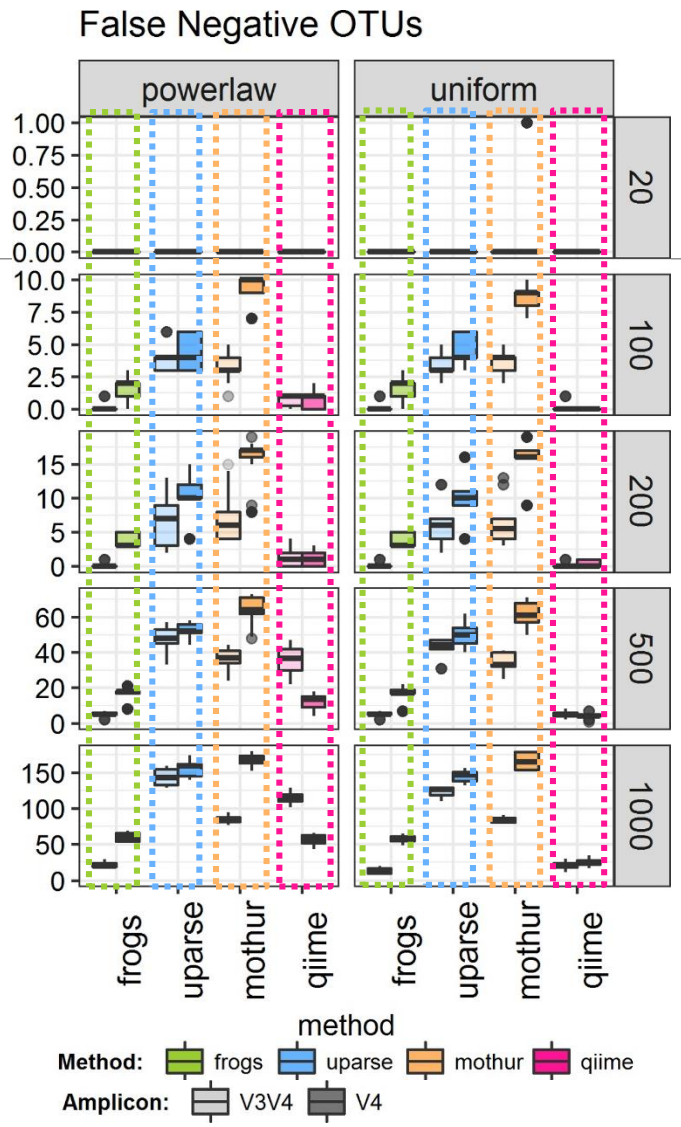
FROGS performed better than QIIME in all settings



FROGS truly outperformed mothur in terms of FN taxa

FROGS always produced fewer FNs than UPARSE.

FROGS sometimes produced more FNs than QIIME, especially on the V4 region.



# Conclusions on assessments

FROGS performed much better than mothur in all settings

FROGS is less conservative than UPARSE for small size communities and better (for both FPs and FNs) for large size communities

FROGS is more conservative than QIIME on the V4 region and better (for both FPs and FNs) on V3V4 regions.

FROGS maintained both the number of FP and FN OTUs low, especially in complex communities.

→ cross-validation of chimeras, only used in FROGS, which avoids confusing real OTUs with chimeras.

→ 3 step strategy (clustering by Swarm + chimera removal with cross-validation + filtering) = a low FP rate and the high probability of detecting a species that is really present in the dataset *i.e.* a high recall rate.

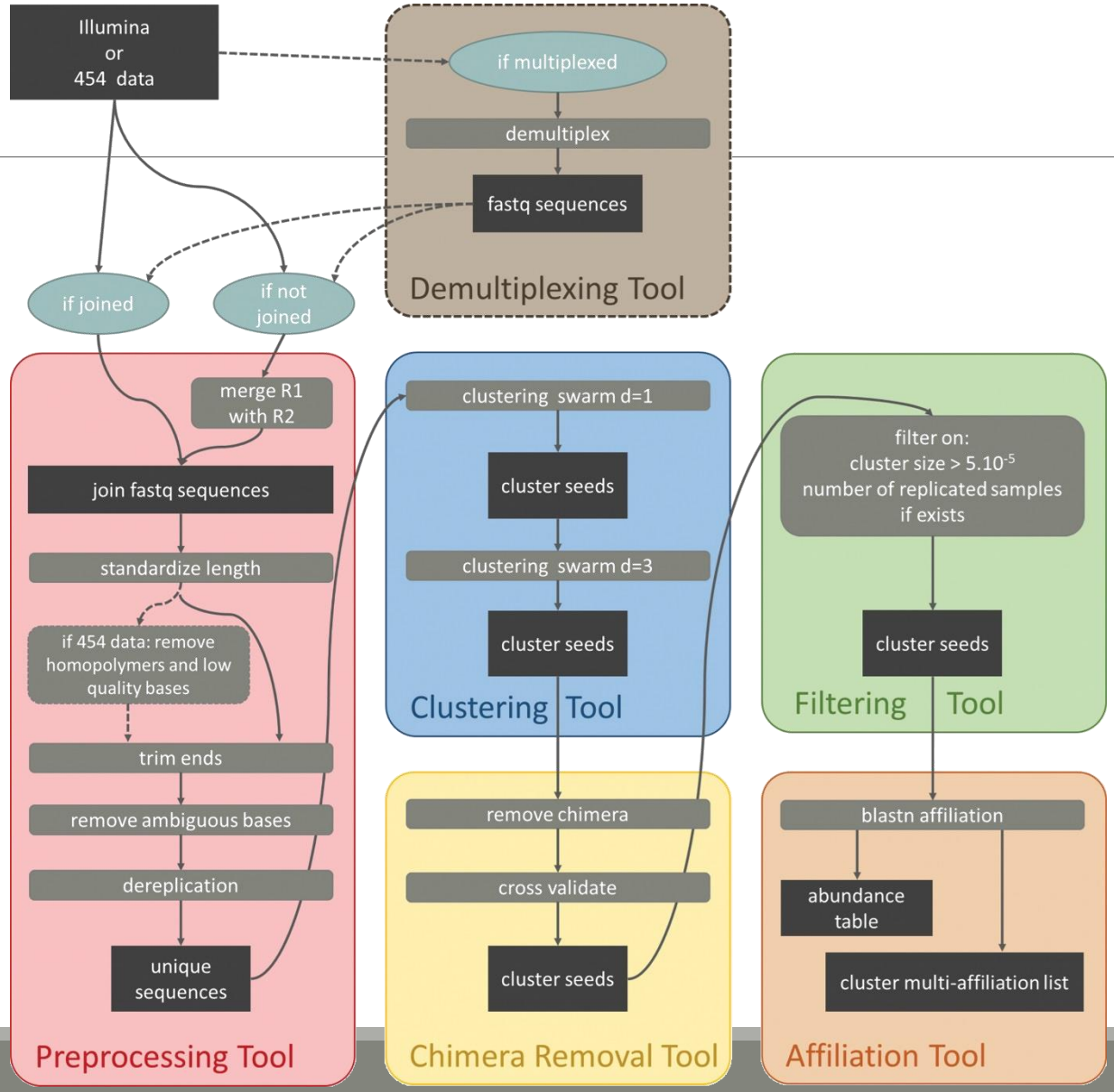
→ unlike QIIME or mothur, FROGS never produced Supernumerary OTUs, which further validates the FROGS OTU picking strategy.

# Conclusions

# Why Use FROGS ?

- User-friendly
- Fast
- 454 data and Illumina data
- Clustering without global threshold and independent of sequence order
- Innovative chimera removal method (Vsearch + cross-validation)
- Filters tool
- Multi-affiliation with 2 taxonomy affiliation procedures
- Cluster Stat and Affiliation Stat tools
- Able to analyse ITS
- A lot of graphics
- Independant tools
- Few False Positives and few False Negatives

# Our recommended guideline for mergeable reads:



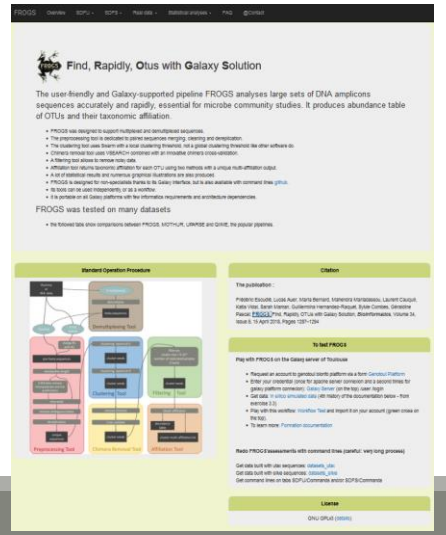
# How to cite FROGS

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

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

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

**Website: http://frogs.toulouse.inra.fr**





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

Sequence analysis

**FROGS: Find, Rapidly, OTUs with Galaxy Solution**

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

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

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

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

**Abstract**

**Motivation:** Metagenomics leads to major advances in microbial ecology and biologists need user friendly tools to analyze their data on their own.
**Results:** This Galaxy-supported pipeline, called FROGS, is designed to analyze large sets of amplicon sequences and produce abundance tables of Operational Taxonomic Units (OTUs) and their taxonomic affiliation. The clustering uses Swarm. The chimera removal uses VSEARCH, combined with original cross-sample validation. The taxonomic affiliation returns an innovative multi-affiliation output to highlight databases conflicts and uncertainties. Statistical results and numerous graphical illustrations are produced along the way to monitor the pipeline. FROGS was tested for the detection and quantification of OTUs on real and *in silico* datasets and proved to be rapid, robust and highly sensitive. It compares favorably with the widespread mothur, UPARSE and QIIME.
**Availability and implementation:** Source code and instructions for installation: https://github.com/geraldinepascal/FROGS.git. A companion website: http://frogs.toulouse.inra.fr.
**Contact:** geraldine.pascal@inra.fr
**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

**1 Introduction**

The expansion of high-throughput sequencing of rRNA amplicons has opened new horizons for the study of microbial communities. By making it possible to study all micro-organisms from a given environment without the need to cultivate them, metagenomics has led to major advances in many fields of microbial ecology, from the study of the impact of microbiota on human and animal pathologies (Ioss *et al.*, 2011; Hooper *et al.*, 2012; Jowel *et al.*, 2016) to the study of biodiversity in environmental ecosystems and the search for biomarkers of pollution (Andres and Bertin, 2016; de Vargas *et al.*, 2015). Determining the composition of a microbial ecosystem, at low cost and great depth, is still largely based on the amplification and sequencing of biodiversity marker genes, also called amplicons, such as rRNA genes and ITS. The clustering of sequences into

# To contact

FROGS:

frogs@inra.fr

Galaxy:

support.sigenae@inra.fr

Newsletter – subscription request:

frogs@inra.fr

# Play list FROGS:

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