



**HAL**  
open science

## Metabarcoding, FROGS training - Oct. 2024

Géraldine Pascal, Lucas Auer

► **To cite this version:**

Géraldine Pascal, Lucas Auer. Metabarcoding, FROGS training - Oct. 2024. Master. FROGS metabarcoding training - Oct 2024, [https://web-genobioinfo.toulouse.inrae.fr/formation/15\\_FROGS/current/](https://web-genobioinfo.toulouse.inrae.fr/formation/15_FROGS/current/), France. 2024. <hal-04885591>

**HAL Id: hal-04885591**

**<https://hal.science/hal-04885591v1>**

Submitted on 14 Jan 2025

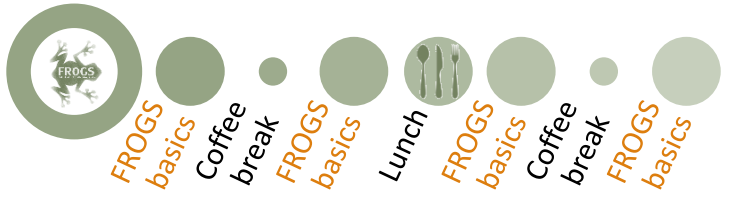
HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY-NC-ND 4.0 - Attribution - Non-commercial use - No Derivative Works - International License

# Monday 1



9 am to 5 pm

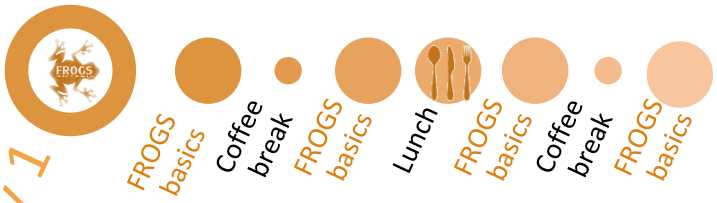


2 short coffee breaks morning and afternoon

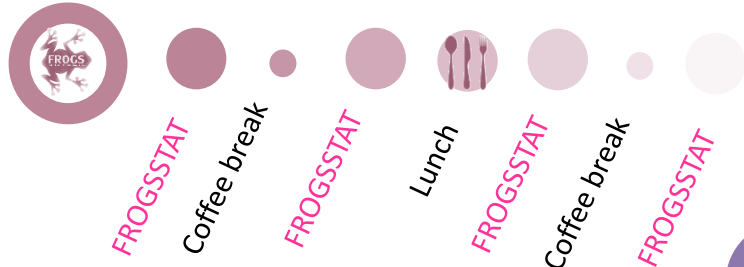


Lunch  
12.30 am to 1.30 pm

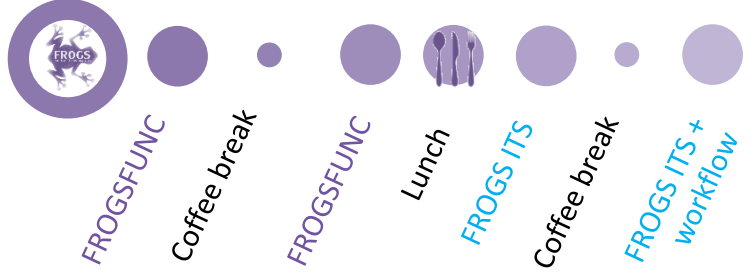
# Tuesday 1



# Monday 2



# Tuesday 2



## Who is in the current FROGS group?

---



Lucas AUER

Biology and statistical  
expert



Maria BERNARD



Olivier RUÉ

Developers



Patrice DÉHAIS

Galaxy  
support

Gabryelle  
AGOUTIN



Databank manager



Géraldine  
PASCAL

Coordinator

# FROGS articles

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

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

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

Sequence analysis

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

Frédéric Escudié<sup>1,†</sup>, Lucas Auer<sup>2,†</sup>, Maria Bernard<sup>3</sup>, Mahendra Mariadassou<sup>4</sup>, Laurent Cauquil<sup>5</sup>, Katia Vidal<sup>6</sup>, Sarah Maman<sup>6</sup>, Guillermina Hernandez-Raquet<sup>6</sup>, Sylvie Combes<sup>7</sup> and Géraldine Pascal<sup>2,\*</sup>

<sup>1</sup>Bioinformatics platform Toulouse Midi-Pyrénées, MIAI, INRA Arceville CS 52627 31228 Castanet Tolosan cedex, France, <sup>2</sup>INRA, UMR 1138, Université de Lorraine, INRA Nancy, 54280, Champenoux, France, <sup>3</sup>SABI, INRA, AgroParisTech, Université Paris Saclay, Jouy-en-Josas, France, <sup>4</sup>MaDiGE, INRA, Université Paris Saclay, 78350, Jouy-en-Josas, France, <sup>5</sup>GenPhySE, Université de Toulouse, INRA, INPT, ENVT, Castanet Tolosan, France and <sup>6</sup>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 16, 2017; revised on December 1, 2017; editorial decision on December 4, 2017; accepted on December 5, 2017

### Abstract

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

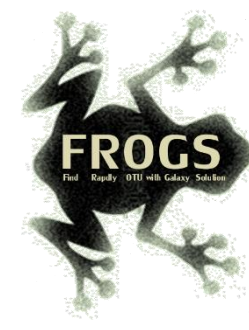
**Availability and implementation:** Source code: geraldinepascal/FROGS.git. A companion web: Contact: geraldine.pascal@inra.fr **Supplementary information:** Supplementary

### 1 Introduction

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

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

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



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

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

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

Maria Bernard<sup>1</sup>, Olivier Rué<sup>1</sup>, Mahendra Mariadassou<sup>2</sup> and Géraldine Pascal<sup>2</sup>

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

\*Maria Bernard and Olivier Rué are joint first authors.

### Abstract

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

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

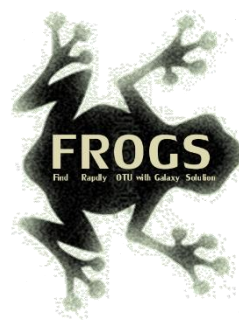
### Introduction

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

for bacteria. The best candidates are internal transcribed spacers (ITS), but these are more difficult to manipulate. The main problem with ITS is size polymorphism, with a size range of 361–1475 bases in ITS1 and 7–11 kb (nucleotide 165 where 95% of the sequences have a length between 1205 and 1556 bases). Most studies describing ITS data analyses process either (i) paired-end reads but filter out non-overlapping, non-mergable reads, thus systematically discarding taxa with longer ITS, or (ii) single-end reads, thus limiting taxonomic resolution and losing the benefit of information contained in longer sequences [4, 5].

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

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



# FROGS'docs

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

All scripts on Github: 

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

Available on :  ANACONDA.ORG

 Galaxy Tool Shed

<https://anaconda.org/bioconda/frogs>

<https://toolshed.g2.bx.psu.edu/view/frogs/frogs/834843ebe569>

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

- FROGS was designed to support multiplexed and demultiplexed sequences.
- FROGS supports ITS, 18S, 23S, 16S and other amplicon reads.
- FROGS supports short or long reads.
- The preprocessing tool is dedicated to paired sequence merging, cleaning and denoising.
- The clustering tool uses Swarm with a novel clustering threshold, but a global clustering threshold for other software too.
- Others removal tool use VSEARCH combined with an innovative chimera removal algorithm.
- A filtering tool allows to remove rare OTUs.
- ATITest tool returns taxonomic affiliation for each OTU using two methods with a unique multi-verification output.
- FROGS offers numerous features for amplicon map (OTU, ...).
- A lot of statistical results and numerous graphical customizations are also produced.
- FROGS is designed for reproducibility thanks to its Galaxy workflow, but it can be adapted with command lines (bash).
- Its tool can be used independently, or as a workflow.
- To install FROGS, visit us on GitHub.

FROGS was tested on many datasets

- The "FROGS ITS Benchmark 16S" and "FROGS ITS Benchmark 18S" show comparisons between FROGS and other popular pipelines.

**Standard Operation Procedure for amplicons**  
(i.e. 18S, 16S, etc., ITS...)

**Citation**

The publications :

Friedrich, David, Lucie Auri, Marie Demaret, Valérie Virelizier, Laurent Chaput, Marie Yvon, Sarah Vahmani, Guillemme Hernandez-Rogues, Soline Cornet, Genevieve Passet (2022), Frogs: Rapidly OTU with Galaxy Solution, *Bioinformatics*, Volume 38, Issue 8, 18 April 2023, Pages 2871-2881

Marie Demaret, Olivier Auri, Valérie Virelizier, and Genevieve Passet, FROGS: a powerful tool to analyze the diversity of fungi with species assignment of their amplicon sequences, *Bioinformatics* 38(21), 10.1093/bioinformatics/btad116

**To test FROGS**

Install FROGS: [conda](#) or [galaxy](#) workflow

FROGS training, documentation and training data: [FROGS data and documentation](#)

Web FROGS workflows with command lines for analysis [23S amplicons](#) (See data but with sequences from JGI and Arabidopsis, [Biology 10.1093/bioinformatics/btad116](#))

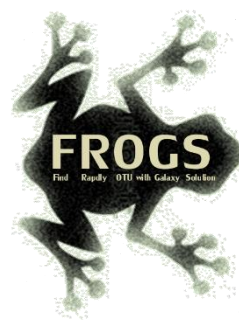
Web FROGS workflows with command lines for analysis [18S amplicons](#) (See data but with sequences from UNITE, [bioRxiv 10.1101/2022.04.14.487116](#))

**License**

GNU GPL3 (default)

**Legal Name**

INRAE  
INSTITUT NATIONAL RECHERCHE POUR L'AGRICULTURE, L'ÉLEVAGE ET LA PÊCHERIE



# To contact




FROGS support


[frogs-support@inrae.fr](mailto:frogs-support@inrae.fr)

Newsletter – subscription request:

[frogs-support@inrae.fr](mailto:frogs-support@inrae.fr)




## June 2023 - FROGS News




- FROGS v4.1.0 is available
  - What has changed since the last version?
  - FROGS produces ASV, Modified tools: More readability of Functional inference; Differential analysis; Affiliation filter
- New documentations for using FROGS v4.1.0
- New databases are available
- You need help to use FROGS, you are looking for training
- Who uses FROGS?

## October 2022 - FROGS News



- FROGS v4.0.1 is available
  - What has changed since the last version?
  - Tools added, Modified tools: Normalisation tool; OTU\_filt
- New documentations for using FROGS v4.0.1
- New databases are available
- You need help to use FROGS, you are looking for training
- Who uses FROGS?

## June 2021 - FROGS News



- FROGS v3.2 is available
- What has changed since the last version?
- New documentations for using FROGS v3.2 on Galaxy
- A redesigned website



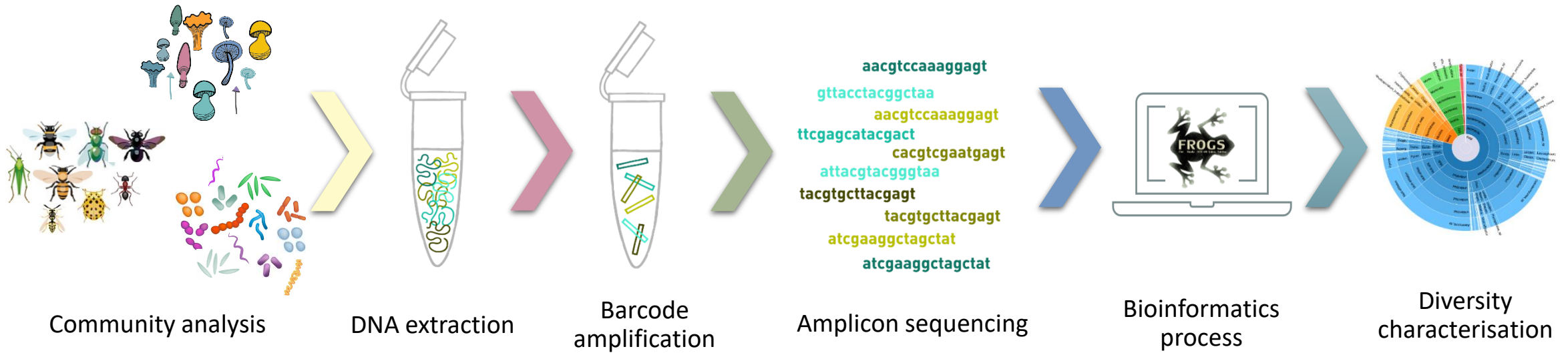
# B- Training on Galaxy: Metabarcoding

October 2024 - Webinar

## FROGS Practice on 16S data

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

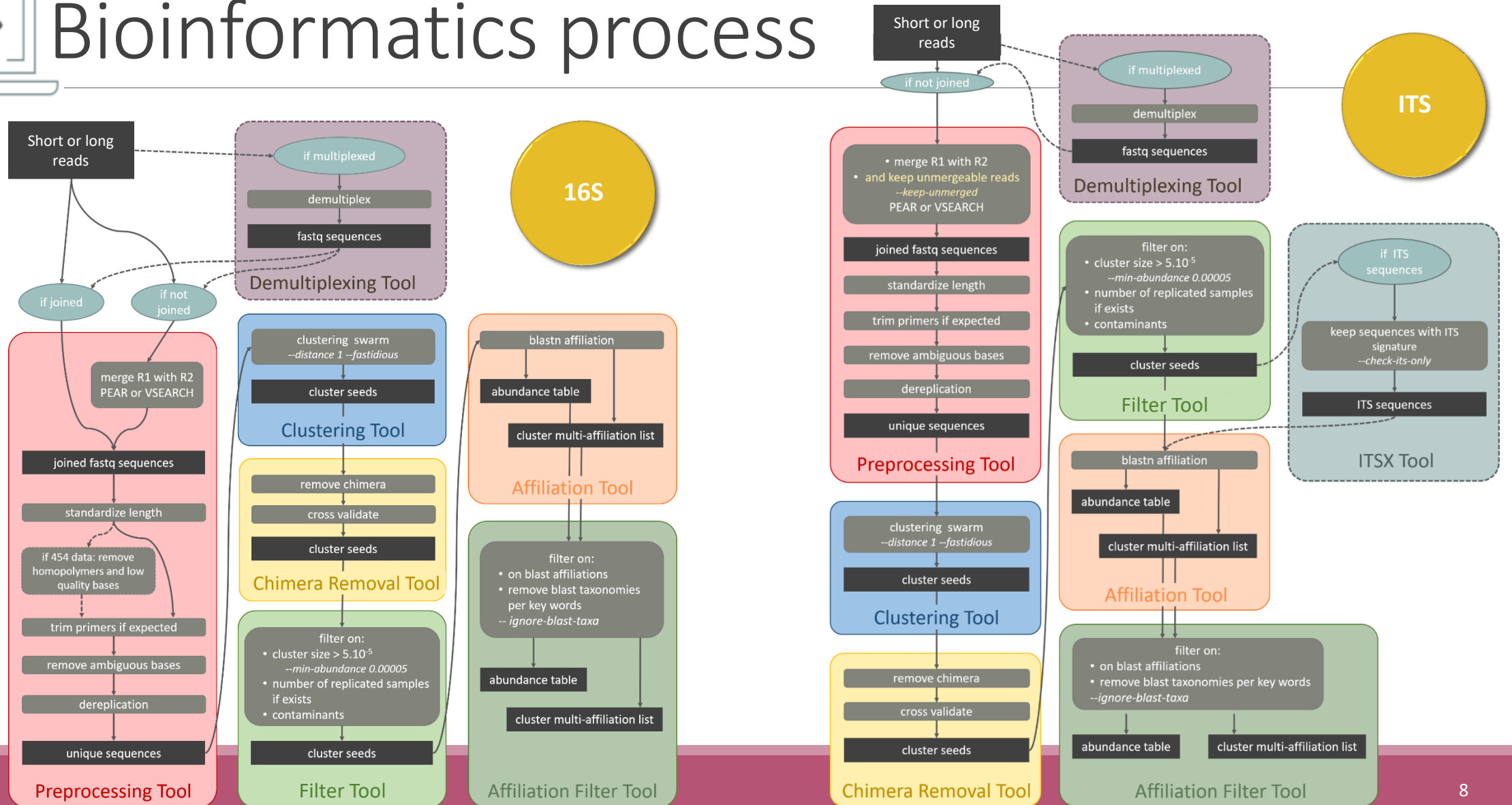
# Objectives



**An abundance table with ASVs and their taxonomic affiliation.**

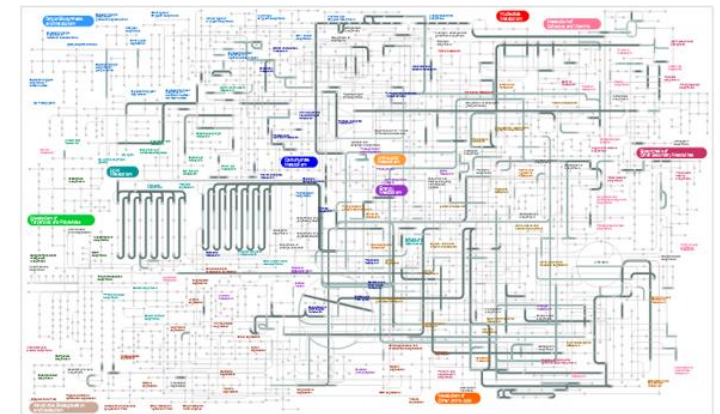
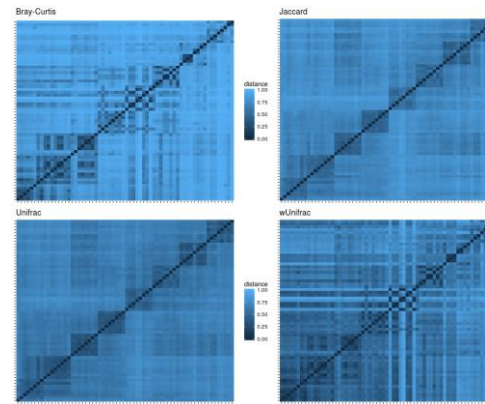
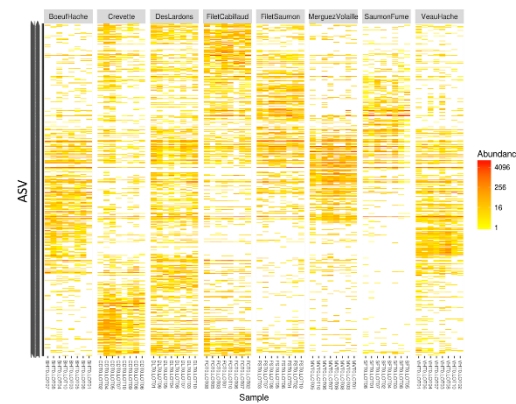
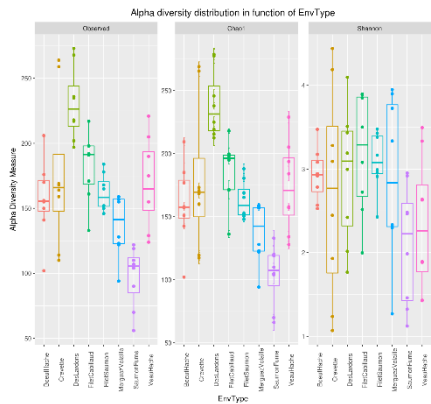


# Bioinformatics process



# Objectives: a count table for statistics analysis

	Affiliation	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
ASV1	Species A	0	100	0	45	75	18645
ASV2	Species B	741	0	456	4421	1255	23
ASV3	Species C	12786	45	3	0	0	0
ASV4	Species D	127	4534	80	456	756	108
ASV5	Species E	8766	7578	56	0	0	200



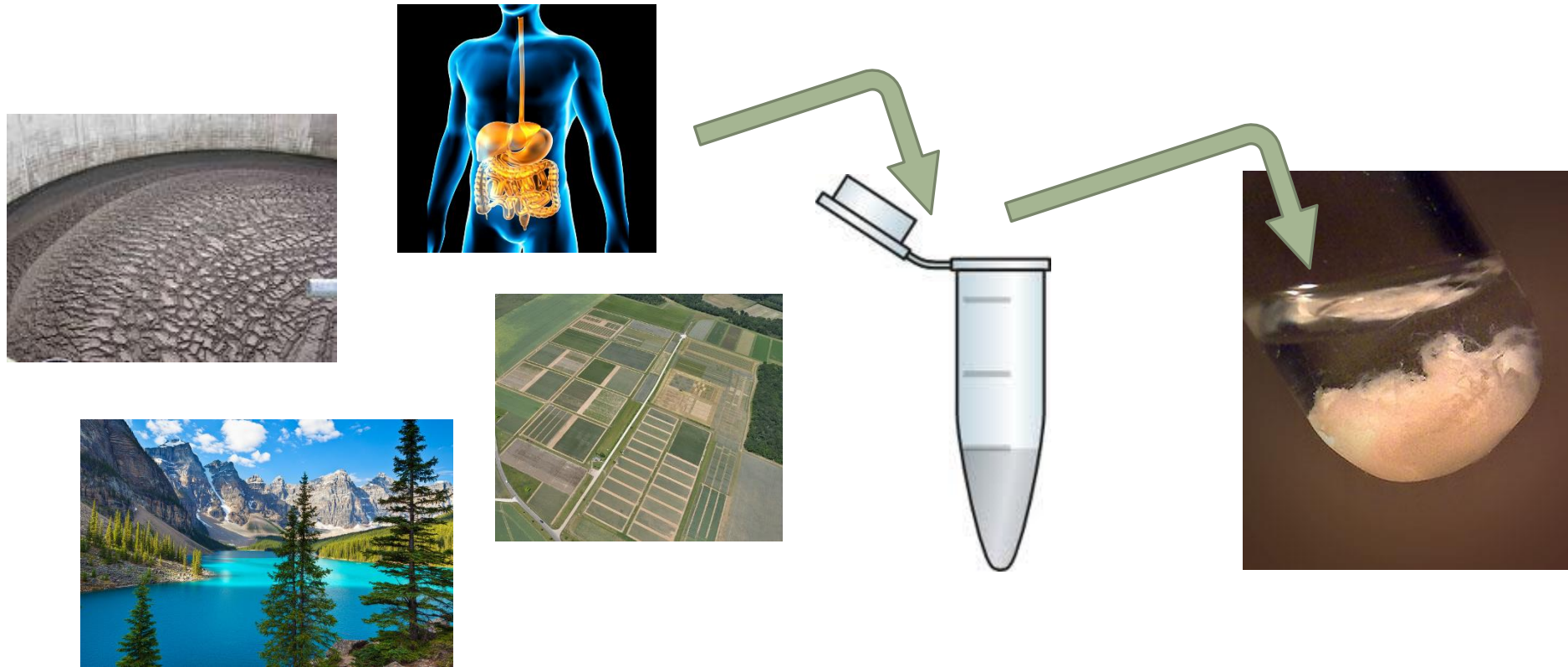
---

# Material

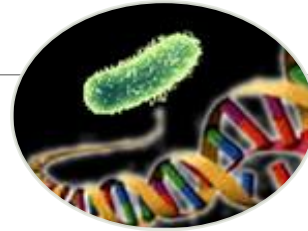
---

# Sample collection and DNA extraction

---



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



DNA

RNA

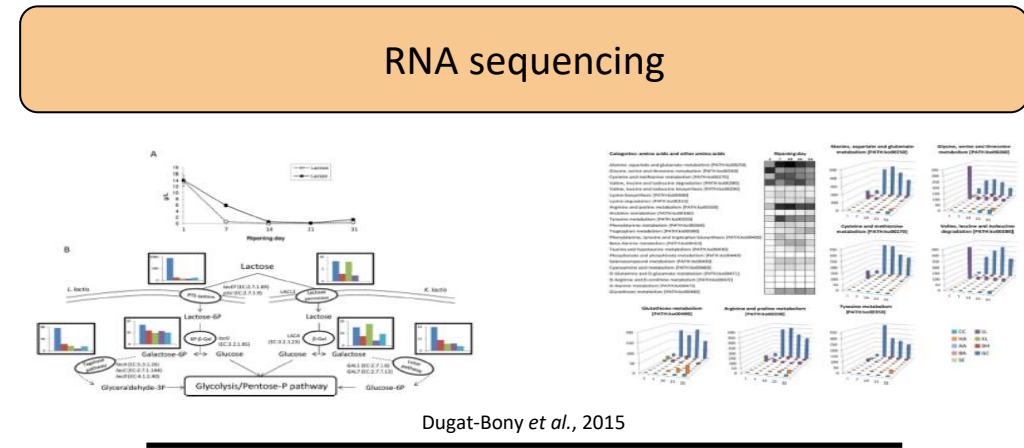
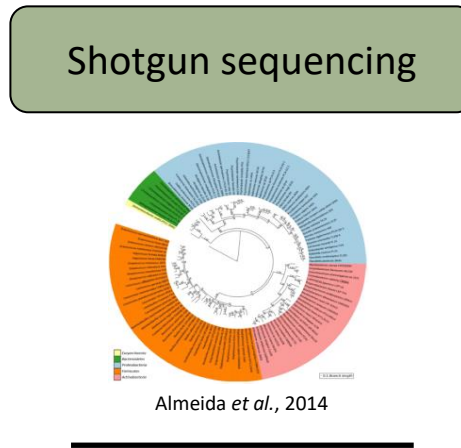
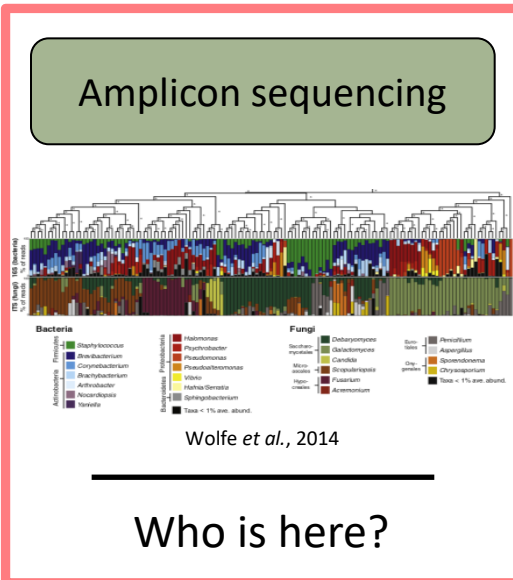
Metagenomics

Metatranscriptomics

Amplicon sequencing

Shotgun sequencing

RNA sequencing



# Story of barcoding

---

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

# Which barcode ?

---

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

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

# The gene encoding the small subunit of the ribosomal RNA

---

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

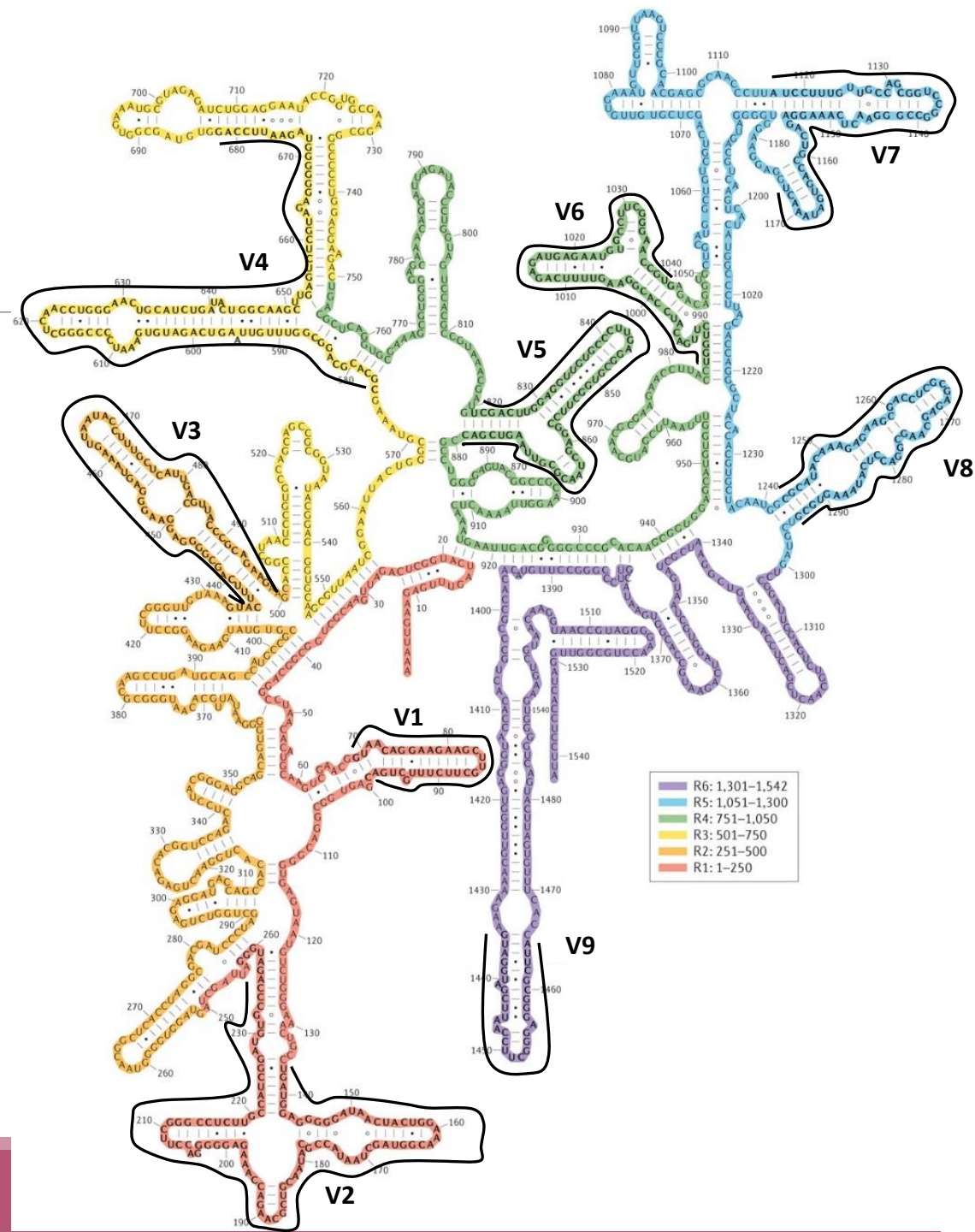
Ubiquist gene : **16S rDNA** in prokaryotes ; **18S rDNA** in eukaryotes

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

Not submitted to lateral gene transfer

Availability of databases facilitating comparison

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

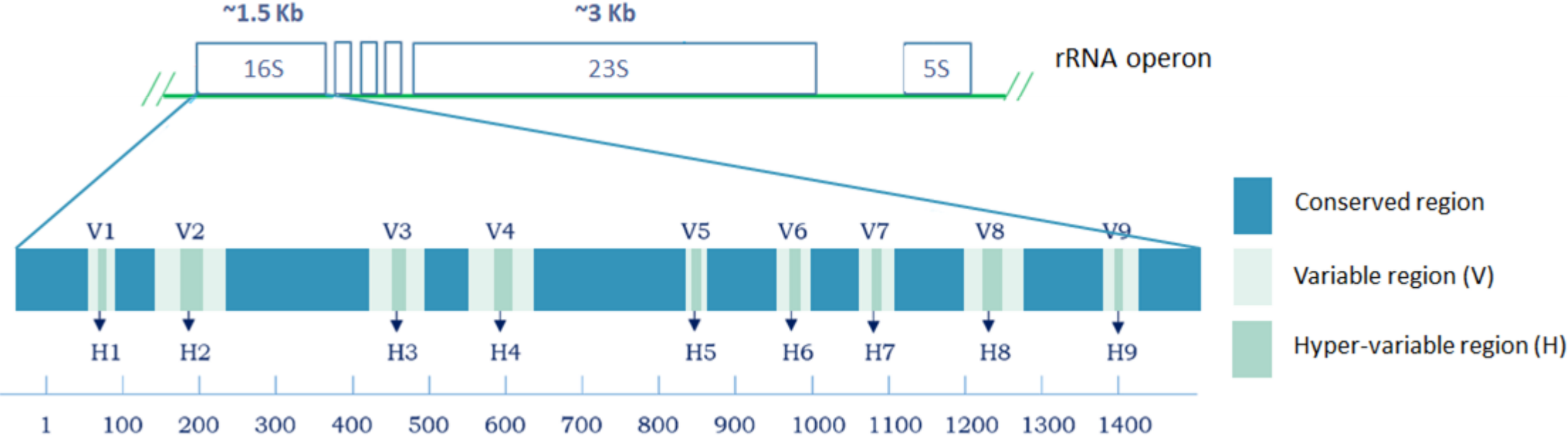


## Secondary structure of the 16S rRNA of *Escherichia coli*

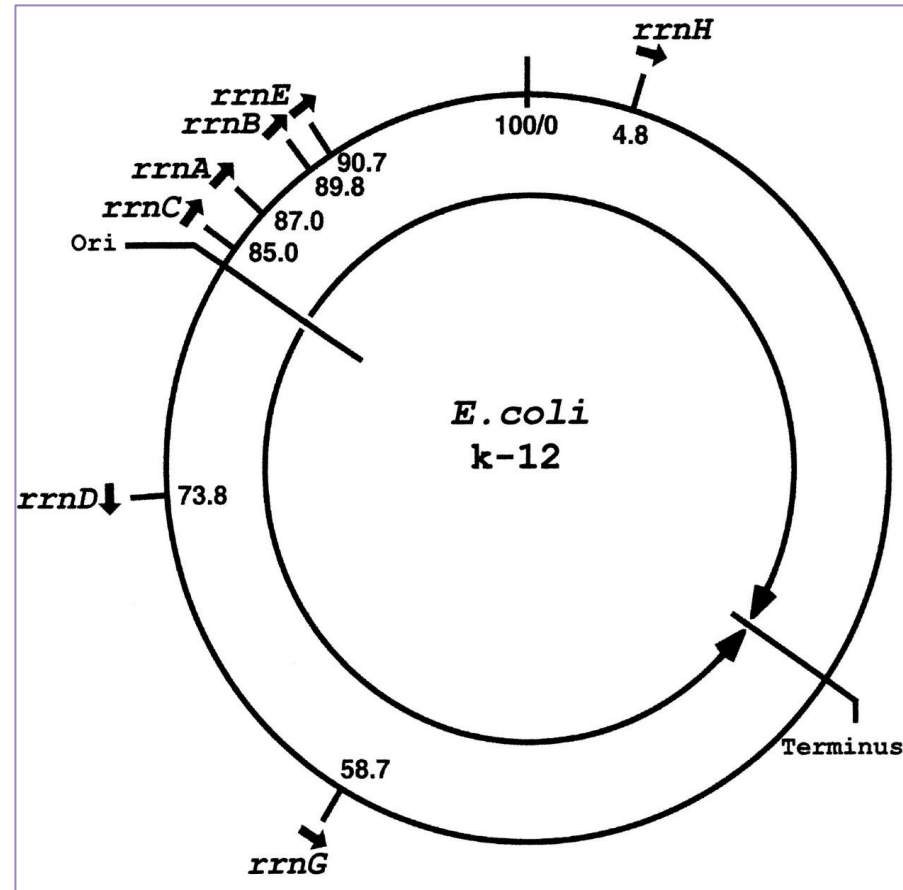
In red, fragment R1 including regions V1 and V2;  
 in orange, fragment R2 including region V3;  
 in yellow, fragment R3 including region V4;  
 in green, fragment R4 including regions V5 and V6;  
 in blue, fragment R5 including regions V7 and V8;  
 and in purple, fragment R6 including region V9.

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

# 16S rRNA structure



# 16S rRNA copy number



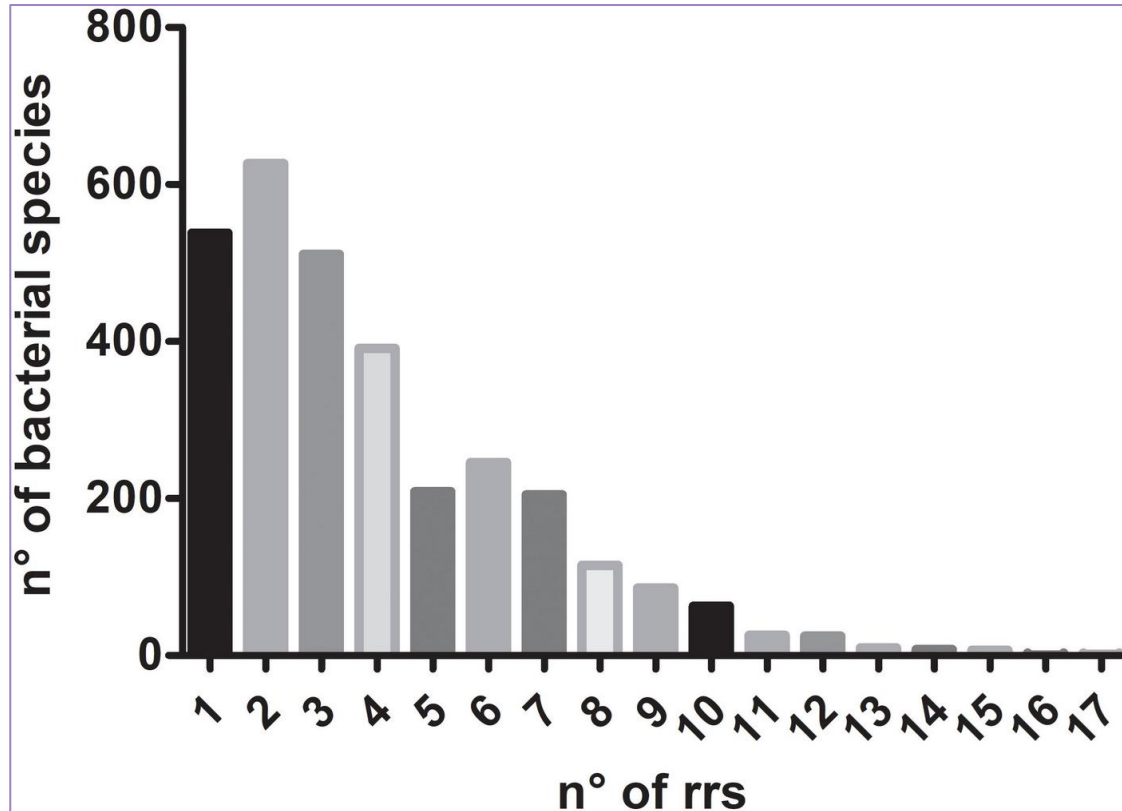
COMMENTARY | FREE ACCESS

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

Masayasu Nomura [Authors Info & Affiliations](#)

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

# 16S rRNA copy number



MINI REVIEW article  
Front. Microbiol., 08 June 2018 | <https://doi.org/10.3389/fmicb.2018.01252>

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

Romilio T. Espejo<sup>1\*</sup> and Nicolás Plaza<sup>2†</sup>

<sup>1</sup>Institute of Nutrition and Food Technology, Universidad de Chile, Santiago, Chile  
<sup>2</sup>Centro de Investigación Biomédica, Facultad de Ciencias de la Salud, Instituto de Ciencias Biomédicas, Universidad Autónoma de Chile, Santiago, Chile

Median of the number of 16S rRNA copies in 3,070 bacterial species according to data reported in *rrnDB* database – 2018

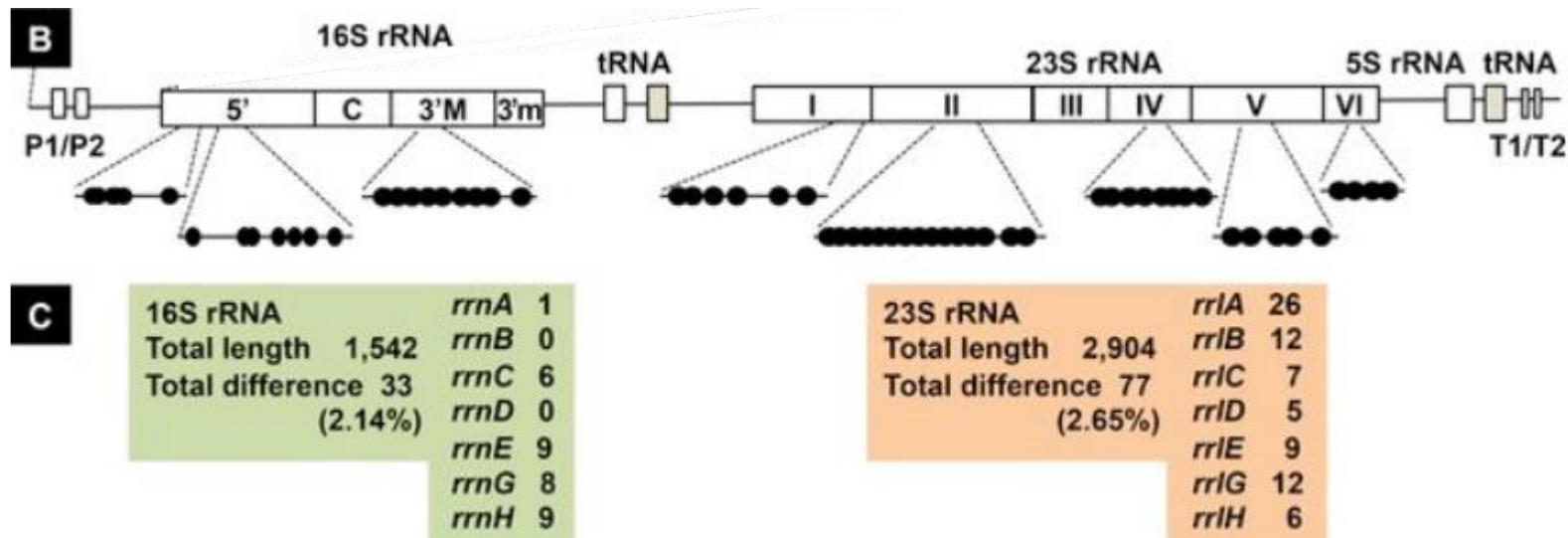
<https://rrndb.umms.med.umich.edu/search/>

2022:

*Bacillus megaterium* entre 1 à 21 copies selon les souches (médiane à 13)

*Photobacterium damsela* entre 15 et 21 copie selon les souches (médiane à 17)

# 16S rRNA copy variation



*E. coli*

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

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

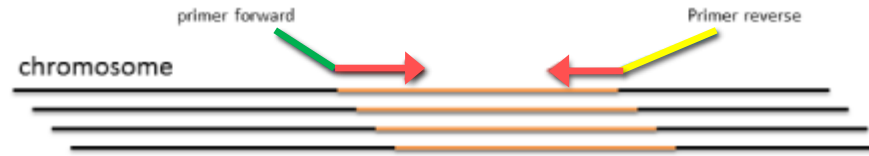
---

Sequencing produces marker  
reads

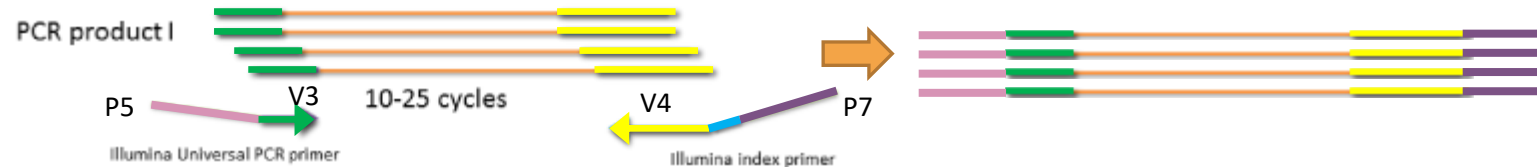
---

# Steps for Illumina sequencing

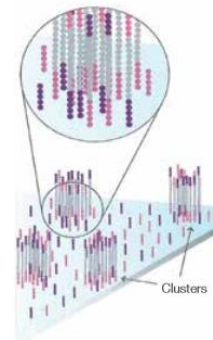
- 1<sup>st</sup> step : one PCR



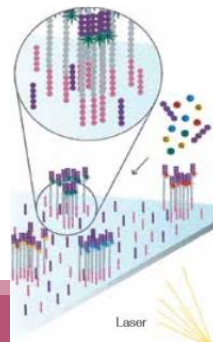
- 2<sup>nd</sup> step: one PCR



- 3<sup>rd</sup> step: on flow cell, the cluster generations

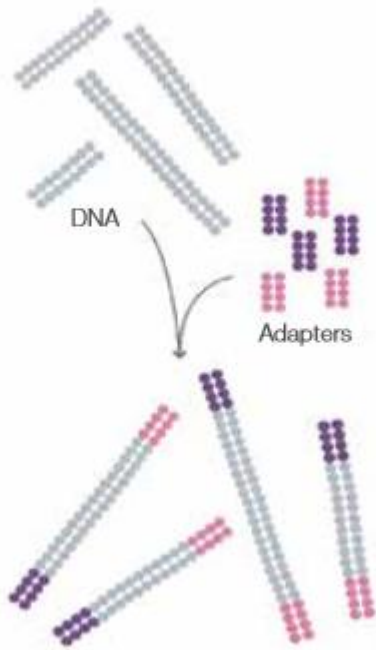


- 4<sup>th</sup> step: sequencing



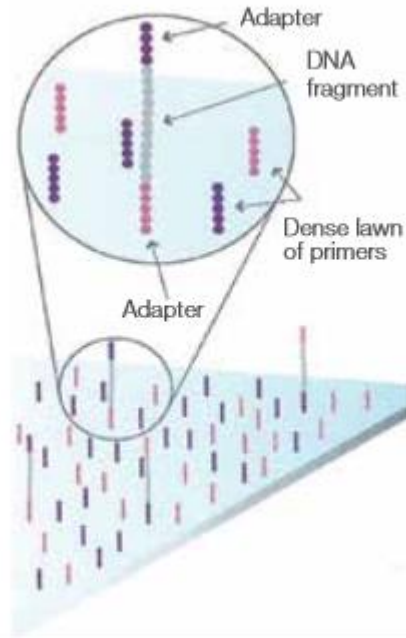
# Cluster generation

## Prepare Genomic DNA Sample



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

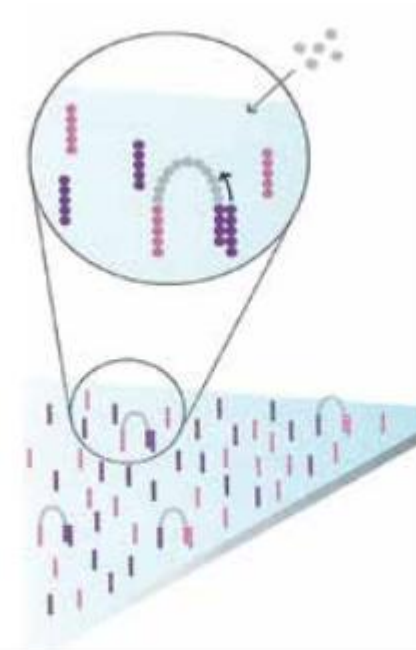
## Attach DNA to Surface



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

Attach DNA to surface

## Bridge Amplification

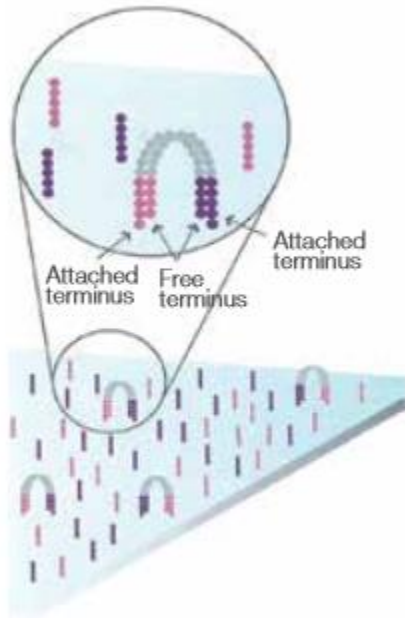


Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

Bridge amplification

# Cluster generation

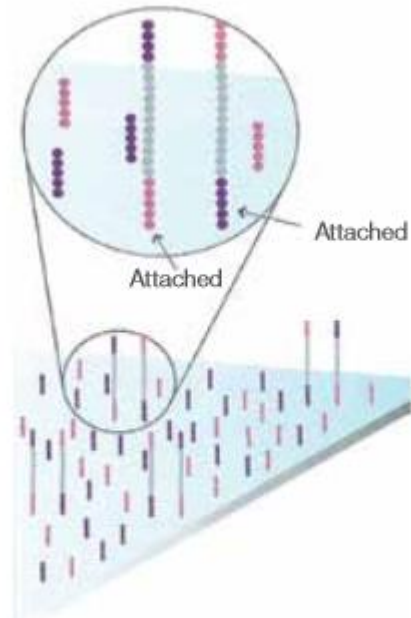
Fragments Become Double Stranded



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

Fragments become double stranded

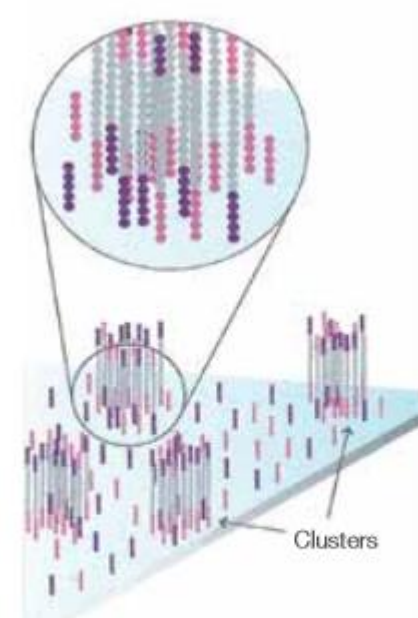
Denature the Double-Stranded Molecules



Denaturation leaves single-stranded templates anchored to the substrate.

Denature the double-stranded molecule

Complete Amplification

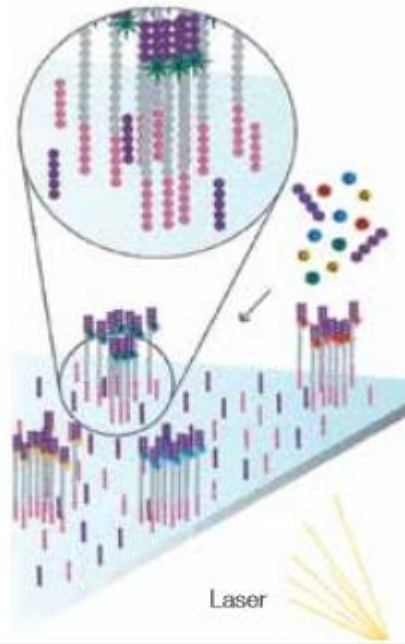


Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

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

# Sequencing by synthesis

## Determine First Base



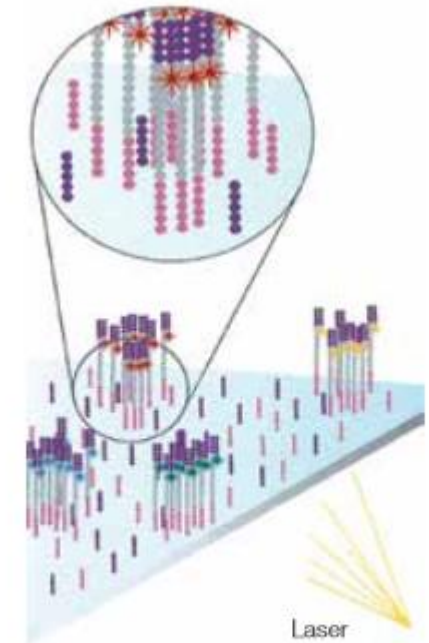
The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.  
Light signal is more strong in cluster

## Image First Base



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

## Determine Second Base



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

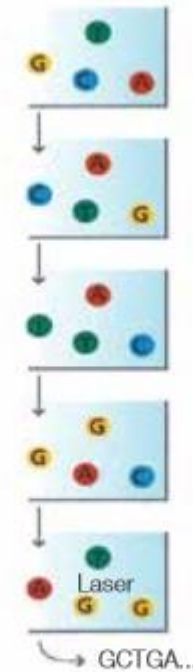
# Sequencing by synthesis

## Image Second Chemistry Cycle



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

## Sequencing Over Multiple Chemistry Cycles

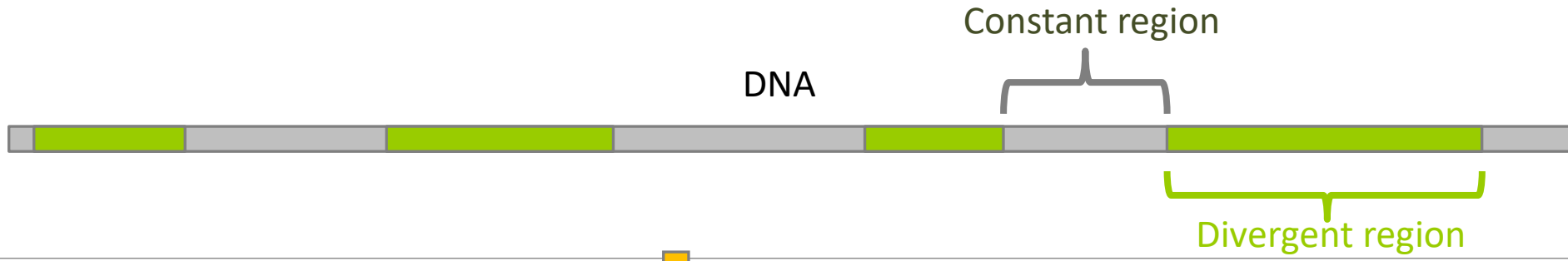


The sequencing cycles are repeated to determine the sequence of bases in a fragment, one base at a time.

Barcode is read, so cluster is identified.

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





↓ PCRs

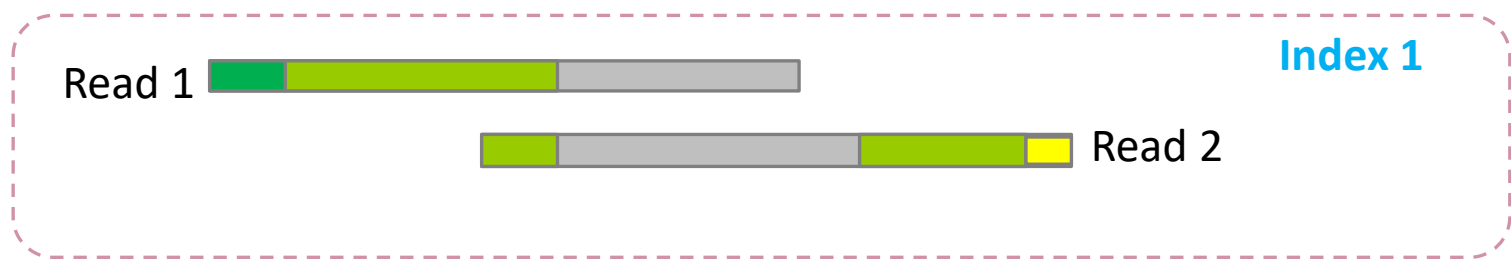
Illumina index



Illumina adapter

Illumina adapter

↓ Sequencing



# Amplification and sequencing

---

Sequencing is generally performed on ~~Roche-454~~ or Illumina MiSeq platforms or Oxford Nanopore Technology or PACBIO platforms.

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

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



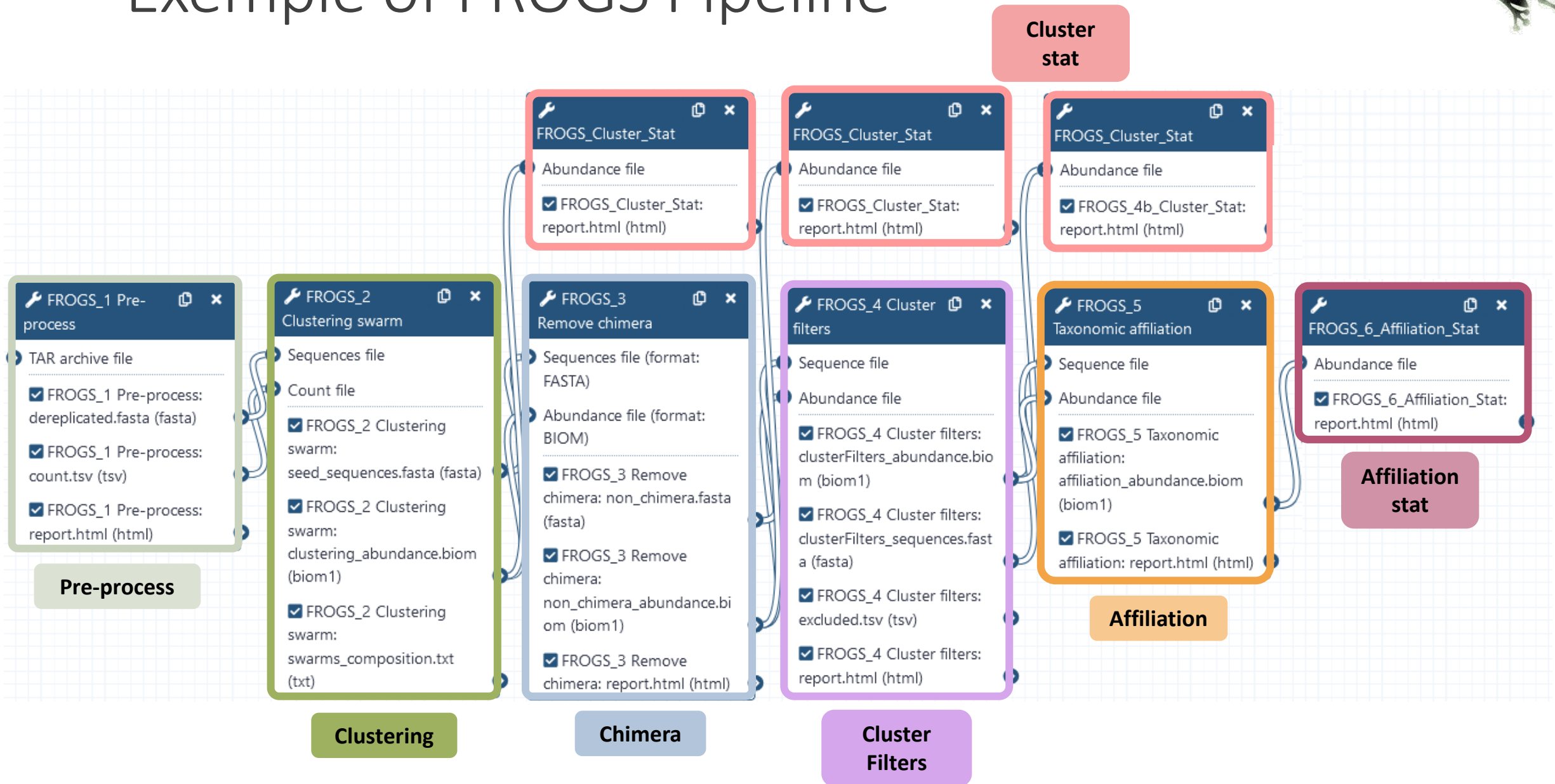
---

# Methods

---



# Example of FROGS Pipeline





**FROGS\_0 Demultiplex reads** Attribute reads to samples in function of inner barcode  
**FROGS\_1 Pre-process** merging, denoising and dereplication  
**FROGS\_2 Clustering swarm** Single-linkage clustering on sequences  
**FROGS\_Cluster\_Stat** Process some metrics on clusters  
**FROGS\_3 Remove chimera** Remove PCR chimera in each sample  
**FROGS\_4 Cluster filters** Filters clusters on several criteria.  
**FROGS ITSx** Extract the highly variable ITS1 and ITS2 subregions from ITS sequences  
**FROGS\_5 Taxonomic affiliation** Taxonomic affiliation of each ASV's seed by RDPtools and BLAST  
**FROGS\_6 Affiliation\_Stat** Process some metrics on taxonomies

### Basic tools

**FROGSSTAT Phyloseq Import Data** from 3 files: biomfile, samplefile, treefile  
**FROGSSTAT Phyloseq Composition Visualisation** with bar plot and composition plot  
**FROGSSTAT Phyloseq Alpha Diversity** with richness plot  
**FROGSSTAT Phyloseq Beta Diversity** distance matrix  
**FROGSSTAT Phyloseq Sample Clustering** of samples using different linkage methods  
**FROGSSTAT Phyloseq Structure Visualisation** with heatmap plot and ordination plot  
**FROGSSTAT Phyloseq Multivariate Analysis Of Variance** perform Multivariate Analysis of Variance (MANOVA)  
**FROGSSTAT DESeq2 Preprocess** import a Phyloseq object and prepare it for DESeq2 differential abundance analysis a  
**FROGSSTAT DESeq2 Visualisation** extract and visualise differentially abundant ASVs or functions

### Statistics tools

**FROGS Tree** Reconstruction of phylogenetic tree  
**FROGS Affiliation Filters** Filters ASVs on several affiliation criteria  
**FROGS Affiliation postprocess** Aggregates ASVs based on alignment metrics  
**FROGS Abundance normalisation** Normalise ASV abundance.

### Optional basic tools

**FROGS BIOM to std BIOM** Converts a FROGS BIOM in fully compatible BIOM  
**FROGS TSV\_to\_BIOM** Converts a TSV file in a BIOM file 1  
**FROGS BIOM to TSV** Converts a BIOM file in TSV file

### Utilities tools

**FROGSFUNC\_1\_placeseqs\_and\_copynumbers** Places ASVs into a reference phylogenetic tree.  
**FROGSFUNC\_2\_functions** Calculates functions abundances in each sample.  
**FROGSFUNC\_3\_pathways** Calculates pathway abundances in each sample.

### Functional inference tools



# FROGS Tools for Bioinformatics analyses

The screenshot displays the Galaxy Toulouse web interface. At the top, the navigation bar includes 'Workflow', 'Visualize', 'Données partagées', 'Aide', 'Utilisateur', and 'Using 313.8 MB'. The left sidebar contains a 'Tools' section with a search bar and an 'Upload Data' button. Below this, a list of tools is categorized under 'Sequence Quality & Cleaning'. The main workspace shows a green notification box indicating a successful execution of 'FROGS Pre-process' with the following details:

- Executed FROGS Pre-process and successfully added 1 job to the queue.
- The tool uses this input:
  - 1: ITS1.tar.gz
- It produces 3 outputs:
  - 8: FROGS Pre-process: dereplicated.fasta
  - 9: FROGS Pre-process: count.tsv
  - 10: FROGS Pre-process: report.html

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

The right sidebar shows the 'History' panel with a search bar and a list of jobs. The jobs are color-coded by status:

- Waiting to run (grey): 10: FROGS Pre-process: report.html, 9: FROGS Pre-process: count.tsv, 8: FROGS Pre-process: dereplicated.fasta
- Currently running (orange): 7: metadata ITS.tsv, 6: ITS1.tar.gz, 5: metadata ITS.tsv
- Result files (green): 4: metadata ITS.tsv, 3: Galaxy2-[metadata\_chaillou.tsv].tsv, 1: ITS1.tar.gz
- Echec process (red): 2: metadata ITS.tsv (with error message: "tool error Une erreur est survenue avec ce jeu de données: Unable to finish job")

Waiting to run

Currently running

Result files

Echec process

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

More name blocks.

FROGS\_  
FROGSSTATS\_  
FROGSFUNC\_

**FROGS\_0 Demultiplex reads** Attribute reads to samples in function of inner barcode

**FROGS\_1 Pre-process** merging, denoising and dereplication

**FROGS\_2 Clustering swarm** Single-linkage clustering on sequences

**FROGS\_Cluster\_Stat** Process some metrics on clusters

**FROGS\_3 Remove chimera** Remove PCR chimera in each sample

**FROGS\_4 Cluster filters** Filters clusters on several criteria.

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

**FROGS\_5 Taxonomic affiliation** Taxonomic affiliation of each ASV's seed by RDPtools and BLAST

**FROGS Affiliation Filters** Filters ASVs on several affiliation criteria

**FROGS Affiliation postprocess** Aggregates ASVs based on alignment metrics

**FROGS Abundance normalisation** Normalise ASV abundance.

**FROGS Tree** Reconstruction of phylogenetic tree

**FROGS\_6\_Affiliation\_Stat** Process some metrics on taxonomies

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

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

**FROGS TSV\_to\_BIOM** Converts a TSV file in a BIOM file 1

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

**FROGSSTAT Phyloseq Composition Visualisation** with bar plot and composition plot

**FROGSSTAT Phyloseq Alpha Diversity** with richness plot

**FROGSSTAT Phyloseq Beta Diversity** distance matrix

**FROGSSTAT Phyloseq Sample Clustering** of samples using different linkage methods

**FROGSSTAT Phyloseq Structure Visualisation** with heatmap plot and ordination plot

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

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

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

**FROGSFUNC\_1\_placeseqs\_and\_copynumbers** Places ASVs into a reference phylogenetic tree.

**FROGSFUNC\_2\_functions** Calculates functions abundances in each sample.

**FROGSFUNC\_3\_pathways** Calculates pathway abundances in each sample.

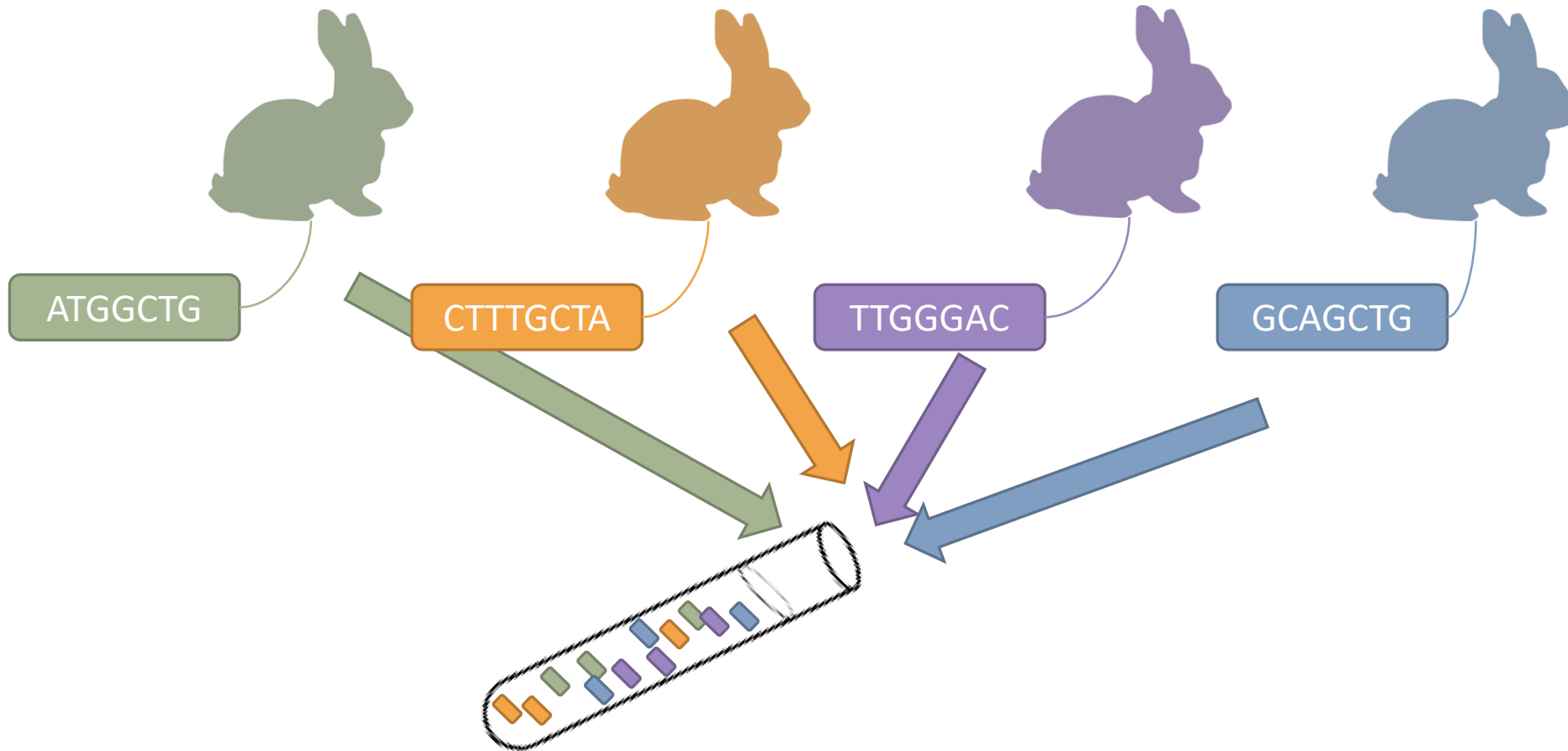


# 0-Demultiplexing tool

---

# Barcoding ?

---



# Demultiplexing

---

Sequence demultiplexing in function of barcode sequences :

- In forward
- In reverse
- In forward and reverse

Remove unbarcoded or ambiguous sequences

# Demultiplexing forward

---



Single-end sequencing

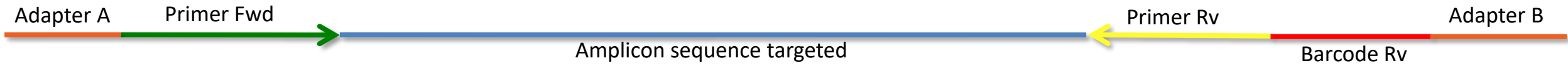


Paired-end sequencing



# Demultiplexing reverse

---



Single end sequencing

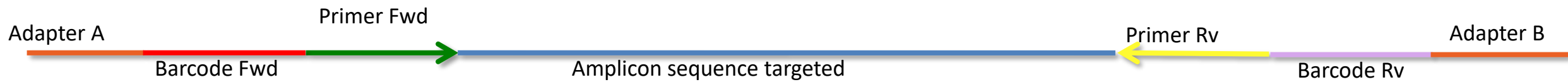


Paire end sequencing



# Demultiplexing forward and reverse

---



Single end sequencing



Paire end sequencing



# The tool parameters depend on the input data type

FROGS Demultiplex reads (version 1.1.0)

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

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

**Select fastq dataset:**  
Specify dataset of your single end reads

**barcode mismatches:**  
0  
Number of mismatches allowed in barcode

**barcode on which end ?:**  
Forward  
Forward at the beginning of the forward end or of the reverse end or both?  
Reverse  
Both ends  
Execute

**Where is the barcode seq on the reads?**

You have only R1 seq.

FROGS Demultiplex reads (version 1.1.0)

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

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

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

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

**barcode mismatches:**  
0  
Number of mismatches allowed in barcode

**barcode on which end ?:**  
Forward  
Forward at the beginning of the forward end or of the reverse end or both?  
Reverse  
Both ends  
Execute

You have R1 and R2 seq.

FROGS Demultiplex reads

- Barcode file
- Select fastq dataset
- demultiplexed\_archive (data)
- undemultiplexed\_archive (data)
- summary (tabular)

Demultiplexing

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

**Barcode file**

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

**Single or Paired-end reads**

Single

Select between paired and single-end data

**Select fastq dataset**

Specify dataset of your single end reads

**Barcode mismatches**

Number of mismatches allowed in barcode

**Barcode on which end ?**

Forward

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

Input example

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

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

# Advices

For your own data

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

# Outputs

7: FROGS\_0 Demultiplex reads: report

6: FROGS\_0 Demultiplex reads: undemultiplexed.tar.gz

5: FROGS\_0 Demultiplex reads: demultiplexed.tar.gz

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

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

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

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

# Format: Barcode

---

BARCODE FILE is expected to be **tabulated**:

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

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

Example of barcode file.

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

MgArd00001	ACAGCGT	ACGTACA
------------	---------	---------

# Format : FastQ

---

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

- first line start by "@" correspond to the sequence identifier and optionally the sequence description. "@Sequence\_1 description1"
- second line is the sequence itself. "ACAGC"
- third line is a "+" following by the sequence identifier or not depending on the version
- fourth line is the quality sequence, one code per base. The code depends on the version and the sequencer

```
@HNHOSKD01ALD0H  
ACAGCGTCAGAGGGGTACCAGTCAGCCATGACGTAGCACGTACA  
+  
CCCFHHHHHHJJJJHHFF@DEDDDDDDDD@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 compared to all barcode sequences.

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

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

---

# 1-Preprocess tool

---



# What does the Pre-process tool do?

---

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

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

Bioinformatics (2011) 27 (21):2957-2963. doi:10.1093/bioinformatics/btr507  
**FLASH: fast length adjustment of short reads to improve genome assemblies**  
TanjaMagoc, Steven L. Salzberg

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

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

---

Examples of different preprocess panels for your future personal uses.

---

# A – for short reads from illumina

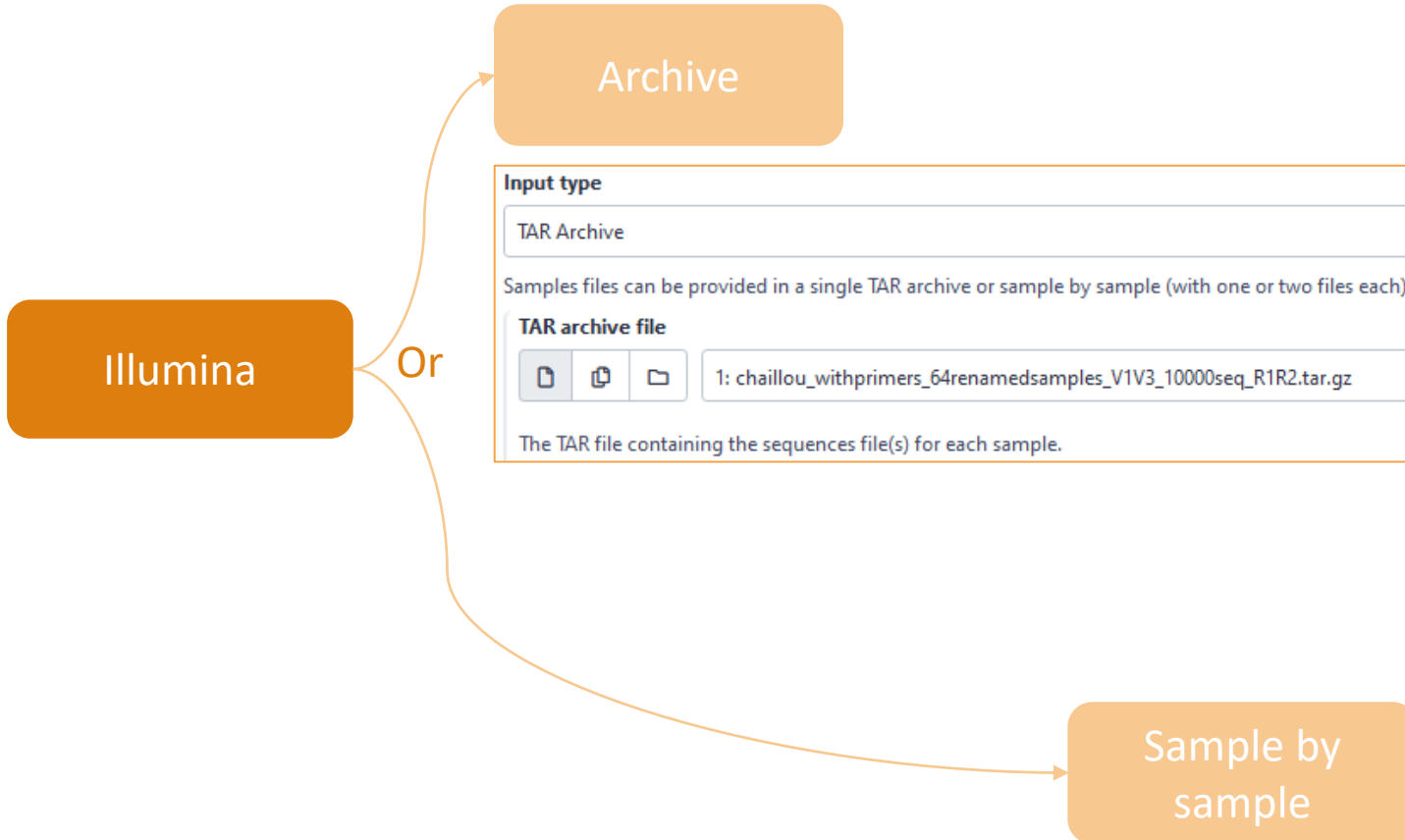
illumina

Sequencer

illumina

Select the sequencing technology used to produce the sequences.

# A – for short reads from illumina



**Input type**

TAR Archive

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

**TAR archive file**

1: chaillou\_withprimers\_64renamedsamples\_V1V3\_10000seq\_R1R2.tar.gz

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

**Input type**

Files by samples

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

**Are reads already merged ?**

No

Yes = The inputs contain 1 file by sample : R1 and R2 pairq are already merged in one sequence.

**Samples**

1: Samples

**Name**

sampleA

The sample name.

**Reads 1**

252: sampleA\_R1.fastq

R1 FASTQ file of paired-end reads.

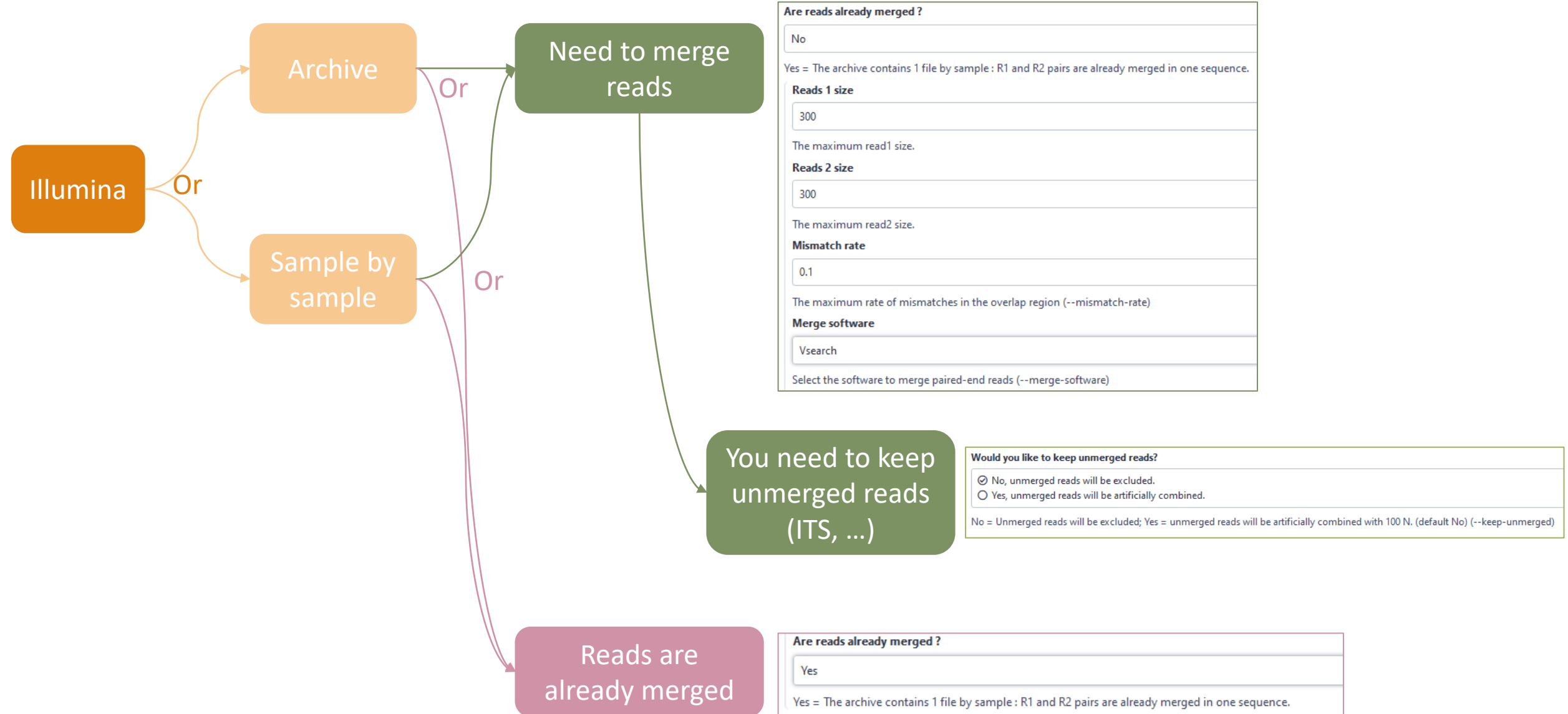
**Reads 2**

251: sampleA\_R2.fastq

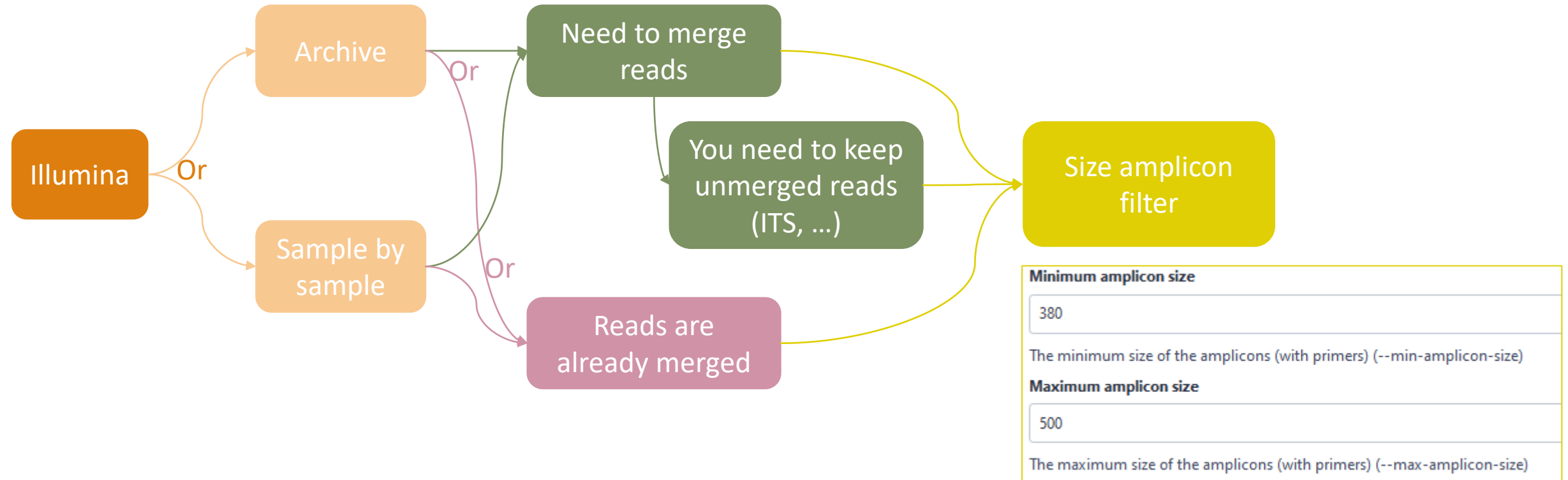
R2 FASTQ file of paired-end reads.

+ Insert Samples

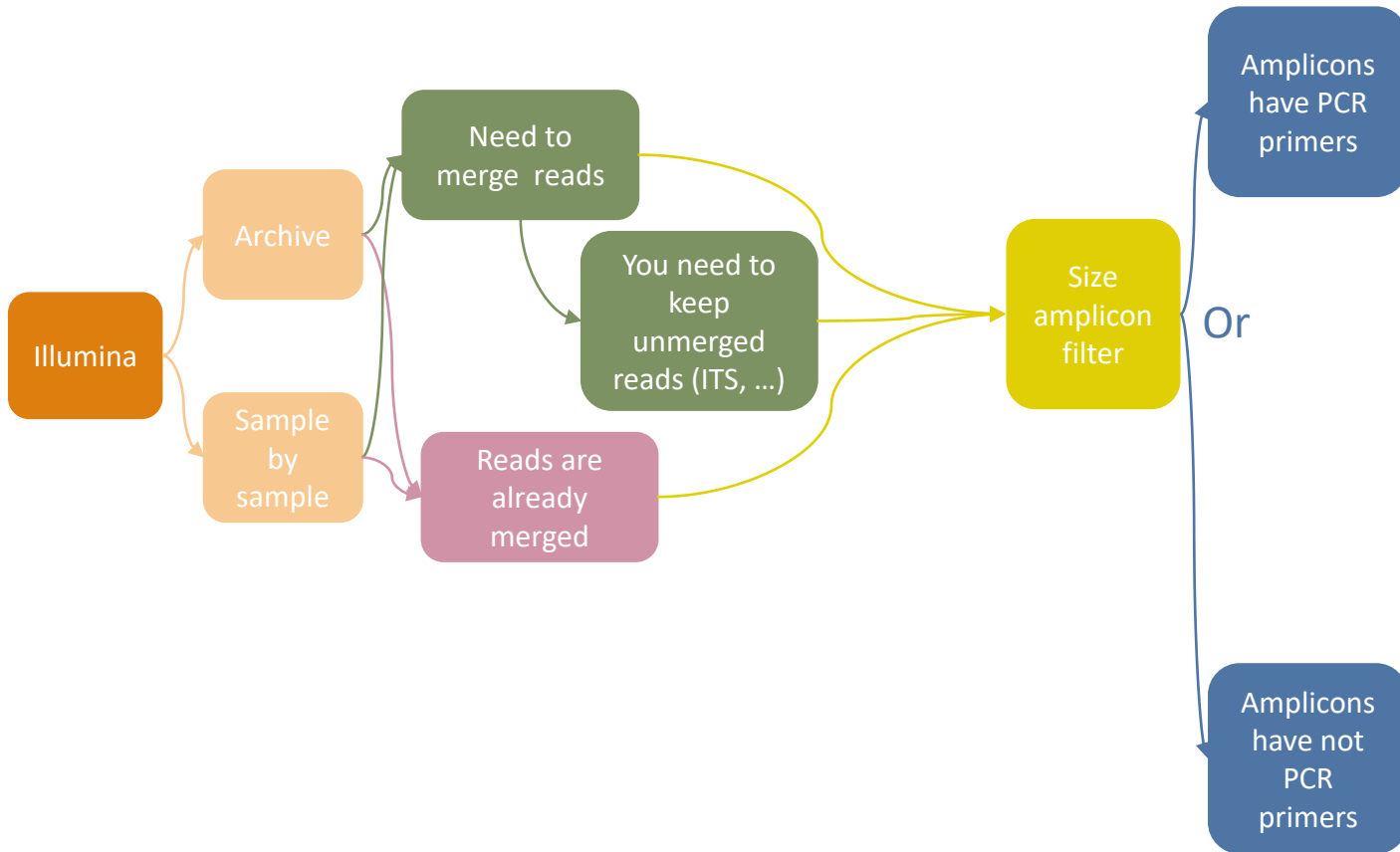
# A – for short reads from illumina



# A – for short reads from illumina



# A – for short reads from illumina



**Do the sequences have PCR primers?**

Yes  
 No

**5' primer**

AGAGTTTGATCCTGGCTCAG

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

**3' primer**

CCAGCAGCCGCGTAAT

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

degenerate primer are accepted (IUPAC code)

**Do the sequences have PCR primers?**

Yes  
 No

# B – for long reads from Pacbio or ONT

Longreads

**Sequencer**

Longreads (PACBIO, ONT)

Select the sequencing technology used to produce the sequences.

# B – for long reads from Pacbio or ONT

Longreads

Or

Archive

**Input type**  
TAR Archive

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

**TAR archive file**

The TAR file containing the sequences file for each sample.

**Sequencer**

Longreads (PACBIO, ONT)

Select the sequencing technology used to produce the sequences.

Sample by sample

**Input type**  
One file by sample

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

**Samples**

1: Samples

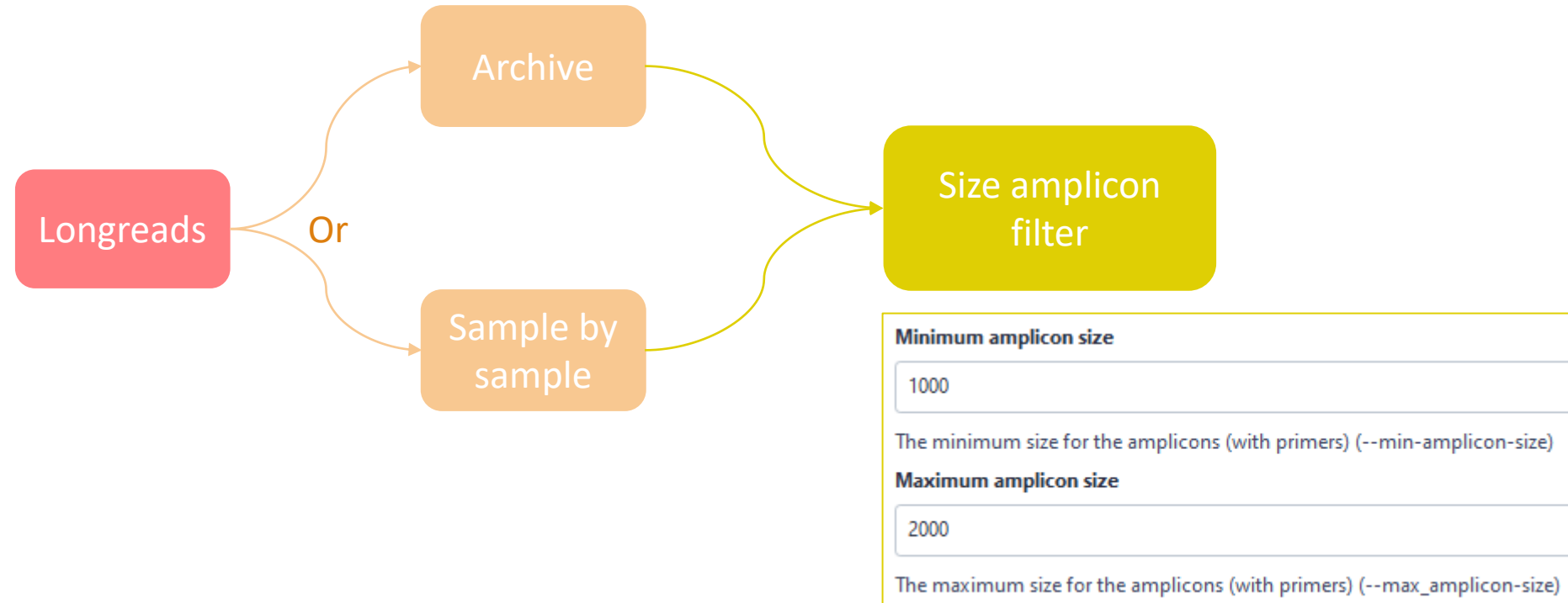
**Name**  
Mockbact

The sample name.

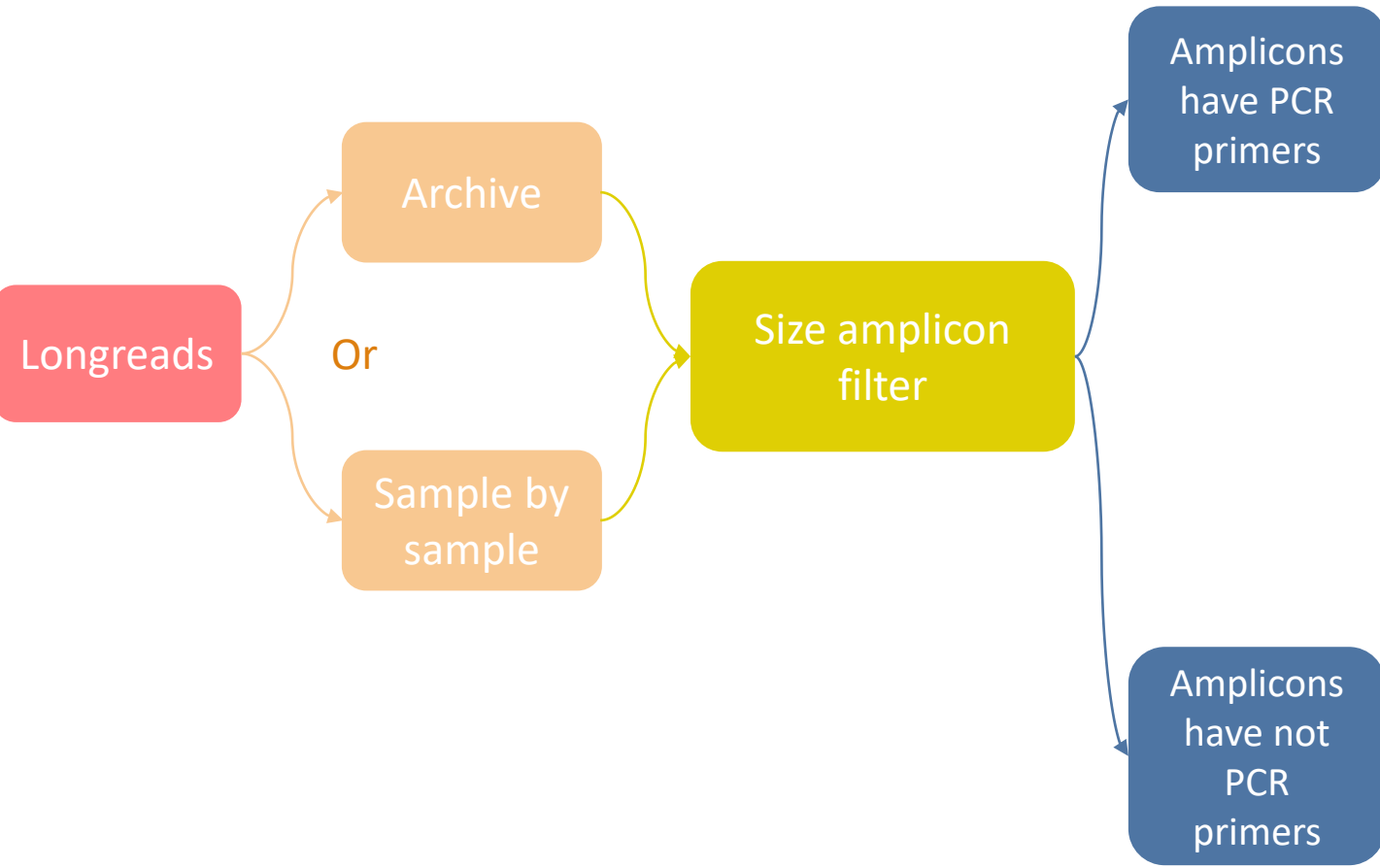
**Sequence file**

FASTQ file of sample.

# B – for long reads from Pacbio or ONT



# B – for long reads from Pacbio or ONT



**Do the sequences have PCR primers?**

Yes  
 No

**5' primer**

AGRGTTYGATYMTGGCTCAG

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

**3' primer**

AAGTCGTAACAAGGTARCY

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

**Do the sequences have PCR primers?**

Yes  
 No

# C – for short reads from 454

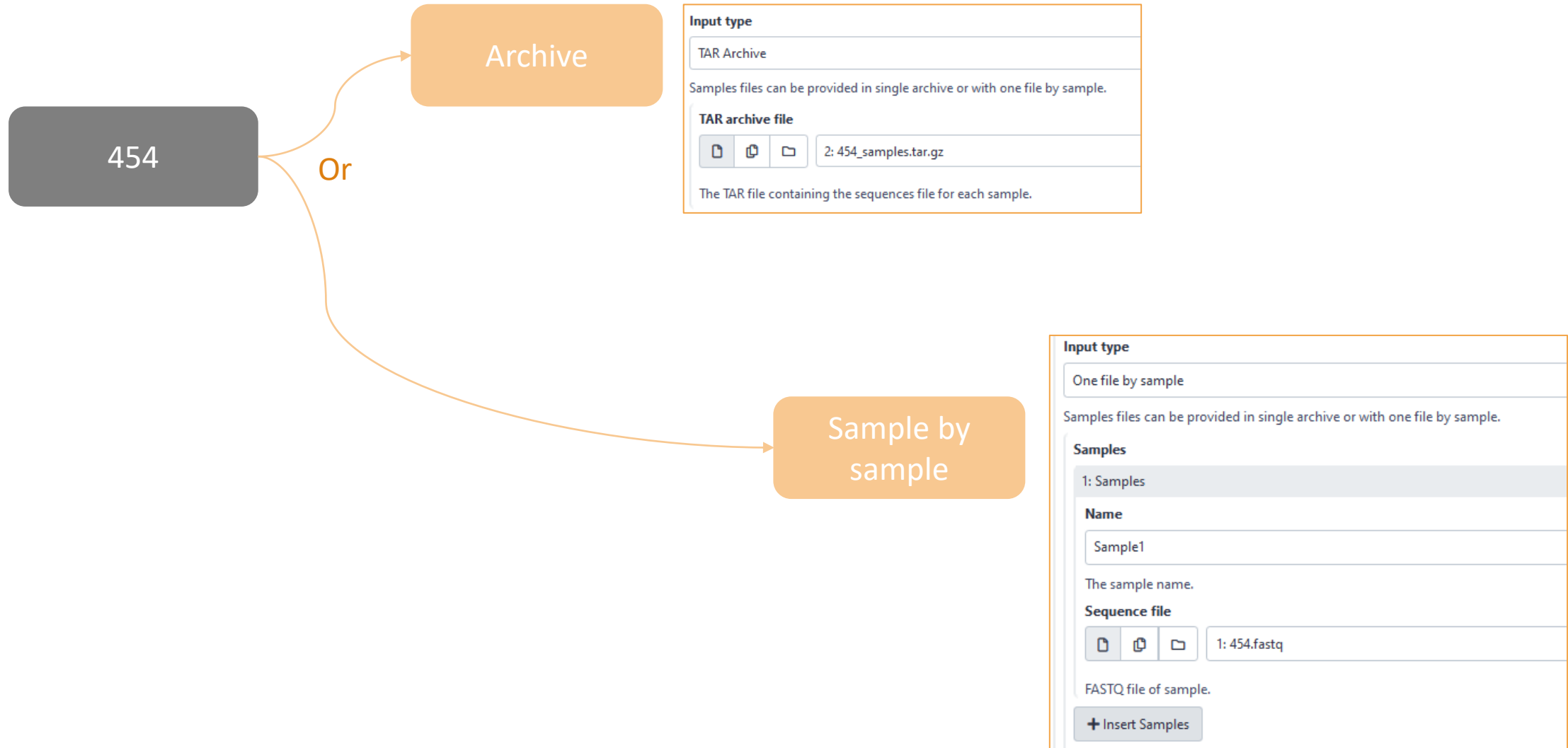
454

**Sequencer**

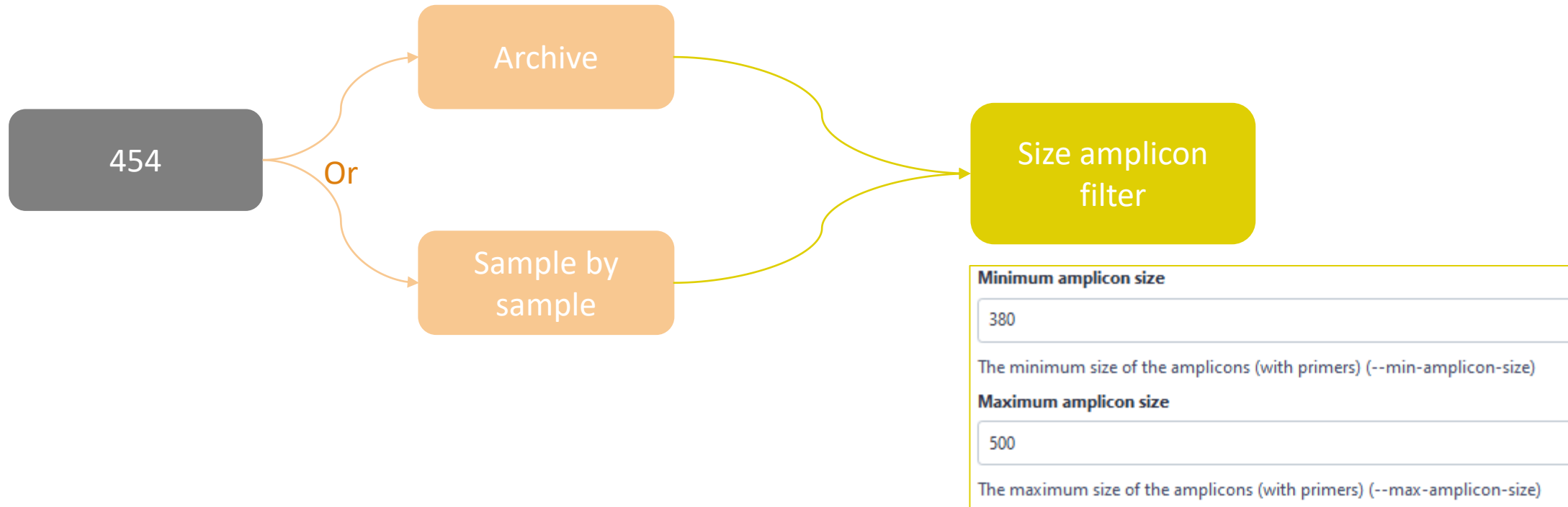
454

Select the sequencing technology used to produce the sequences.

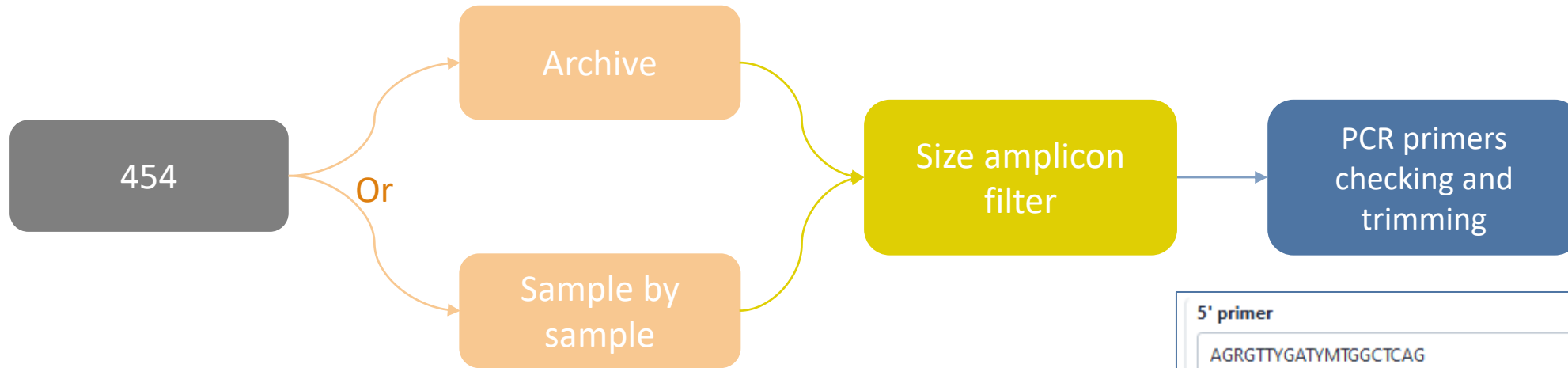
# C – for short reads from 454



# C – for short reads from 454



# C – for short reads from 454



## 5' primer

```
AGRGTTYGATYMTGGCTCAG
```

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

## 3' primer

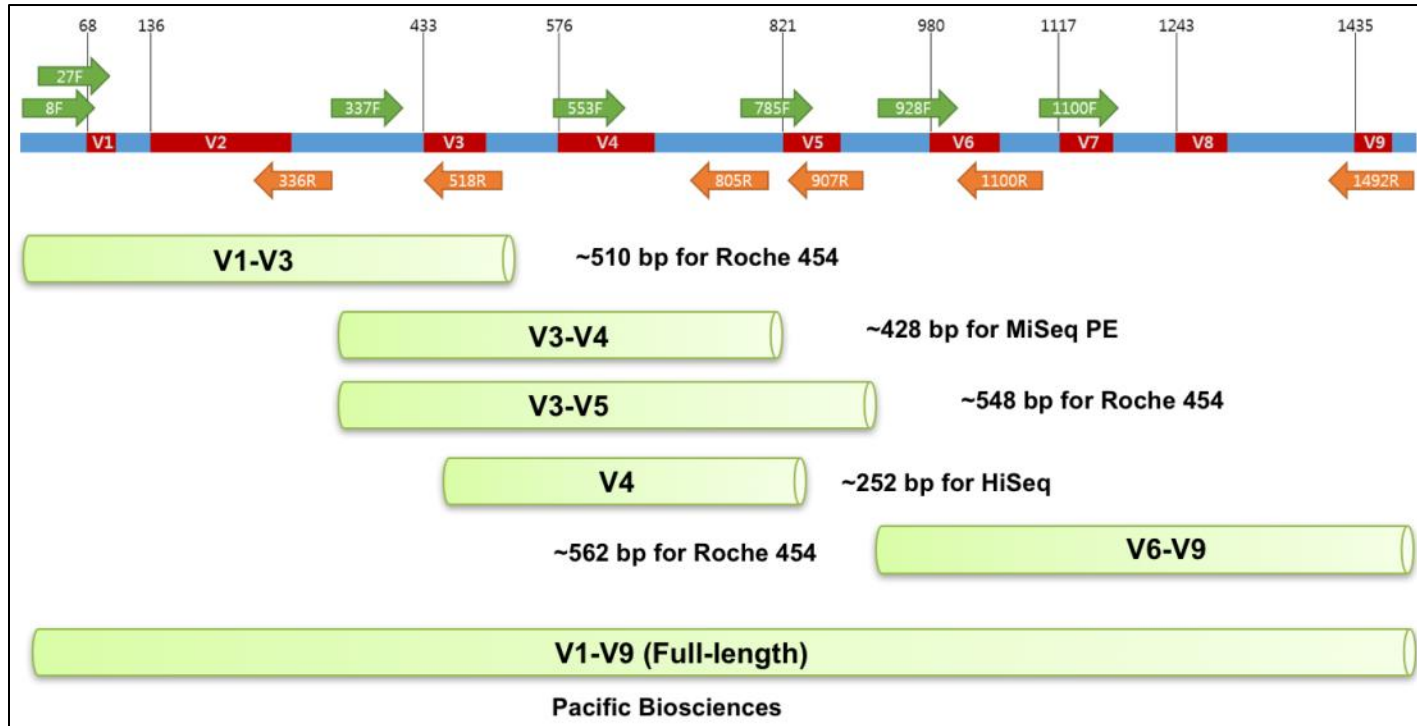
```
AAGTCGTAACAAGGTARCY
```

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

---

Which primers for 16S ?

---



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	TAAACTYAAAKGAATTGACGGG
1100F	YAACGAGCGCAACCC
1100R	GGGTTGCGCTCGTTG
1492R	CGGTTACCTTGTTACGACTT

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

---

# How work reads merging ?

---

WITH VSEARCH

# The aim of Vsearch is to merge R1 with R2

---

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

Imagine a real amplicon sequence of 400bp

400bp



Imagine a Miseq paired sequencing of 2x250bp

R1 : 250bp



R2 : 250bp



Reconstructing amplicon sequence is possible thanks to the overlap region



Merged sequence length : 400bp, with 100bp overlap

# The aim of Vsearch is to merge R1 with R2

---

Case of a sequencing of over-overlapping sequences:

Imagine a real amplicon sequence of 200bp

200bp



Imagine a Miseq paired sequencing of 2x250bp

R1 : 250bp

R2 : 250bp



FROGS takes in charge this case in trimming over bases

200bp



Merged sequence length : 200bp, with 100% overlap

---

# Practice:

---

# Exercise

---

Go to « [16S](#) » history

Launch the pre-process tool on that data set

→ objective: understand Vsearch software

# 16S dataset presentation:

---

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

Comparison of meat and seafood bacterial communities.

8 environment types (EnvType) :

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



# 16S dataset presentation:

---



From Chaillou paper, we produced simulated data:

- 64 samples of 16S amplicons
- R1 and R2 overlapping reads of 300 bases.
- 8 replicates per condition
- with errors among the linear curve  $2.54e-1$   $2.79e-1$

- with 10% chimeras
- Primers for V1-V3:
  - 5' AGAGTTTGATCCTGGCTCAG 3'
  - 5' CCAGCAGCCGCGGTAAT 3'

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

**Sequencer**

Illumina

Select the sequencing technology used to produce the sequences.

**Input type**

TAR Archive

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

**TAR archive file**

1: chaillou\_withprimers\_64renamedsamples\_V1V3\_10000seq\_R1R2.tar.gz

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

**Are reads already merged ?**

No

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

**Reads 1 size**

300

The maximum read1 size.

**Reads 2 size**

300

The maximum read2 size.

**Mismatch rate**

0.1

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

**Merge software**

Vsearch

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

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

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

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

Vsearch is recommended (in command line, prefer pear)

**Minimum amplicon size**

400

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

**Maximum amplicon size**

580

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

**Do the sequences have PCR primers?**

Yes  
 No

**5' primer**

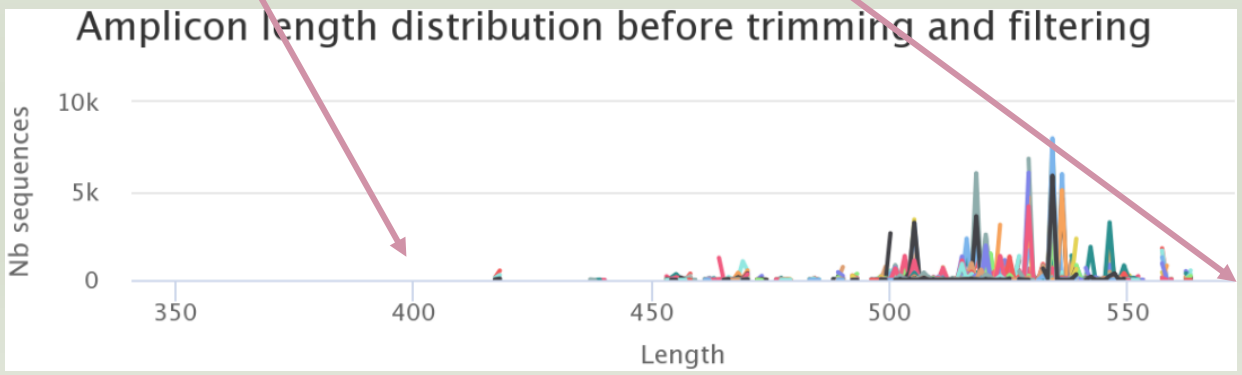
AGAGTTTGATCCTGGCTCAG

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

**3' primer**

CCAGCAGCCGCGTAAT

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



**Minimum amplicon size**

400

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

**Maximum amplicon size**

580

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

**Do the sequences have PCR primers?**

Yes  
 No

**5' primer**

AGAGTTTGATCCTGGCTCAG

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

**3' primer**

CCAGCAGCCGCGGTAAT

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

Primer R1: AGAGTTTGATCCTGGCTCAG  
reverse transcribed Primer R2 : CCAGCAGCCGCGGTAAT

Ex: read R1

@63\_0 reference=ASV\_00517 position=1..300

AGAGTTTGATCCTGGCTCAGgatgaacgctagcgggaggcttaacacatgcaagccgagggg  
tagaattagcttgctaattgagaccggcgacgggtgctgaacgcgatgcaacttgcctactgaaa  
ggatagcccagagaaattggattaatactttataatagactgaatggcatcatttagttttaaagatt  
atcgcagtaggataggcatgctgaagattagatagttggtagagtaacggctcaccaagtcgacgatct  
ttagggggcctgagagggtgaaccccca

Ex: read R2

@63\_0 reference=ASV\_00517 position=1..300 errors=5%G




ATTACCGCGGCTGCTGGcacggagtagccggtgcttattcttctgttacctcagctacttacac  
gtaagtaggtttatccccagataaaaagtagtttacaaccataaggccgtcatcctacacgcgggatggc  
tggatcaggcttccaccattgtccaatattcctcactgctgcctcccgtaggagtctggctcgtgtctcag  
taccagtgtgggggttcaccctctcaggccccctaaagatcgtcgacttggtagccgttacctcacca  
ctatctaattctacgcatgcct



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

# Exercise

---

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

# Exercise

---

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

## Q1: Do you understand how enter your primers ?

**Minimum amplicon size**

The minimum size for the amplicons (with primers).

**Maximum amplicon size**

The maximum size for the amplicons (with primers).

**Sequencing protocol**

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

**5' primer**

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

**3' primer**

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

✓ Execute

N.B.  
Primers in 5' → 3' sens



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

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

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

```
>06_5949;size=4 reference=otu_00680 position=1..300 errors=20%T
AGACCGGCGCACGGGTGCGTAACGCGTATGCAATCTGCCTTTCACAGAGGGATAGCCCAGAGAAAATTTGGATTAATACCTCATA
>56_3551;size=1 reference=otu_00680 position=1..300 errors=21%A
AAGACCGGCGCACGGGTGCGTAACGCGTATGCAATCTGCCTTTCACAGAGGGATAGCCCAGAGAAAATTTGGATTAATACCTCATA
>53_322;size=1 reference=otu_01408,otu_00680 amplicon=1..300,1..300 position=1..300
ATTGAACGGTGGCGGCATGCCTACGCGTATGCAATCTGCCTTTCACAGAGGGATAGCCCAGAGAAAATTTGGATTAATACCTCATA
>56_2589;size=1 reference=otu_00680 position=1..300 errors=21%C
CAGACCGGCGCACGGGTGCGTAACGCGTATGCAATCTGCCTTTCACAGAGGGATAGCCCAGAGAAAATTTGGATTAATACCTCATA
>56_7560;size=1 reference=otu_00680 position=1..300 errors=21%C
CAGACCGGCGCACGGGTGCGTAACGCGTATGCAATCTGCCTTTCACAGAGGGATAGCCCAGAGAAAATTTGGATTAATACCTCATA
>36_626;size=1 reference=otu_00680 position=1..300 errors=21%C
CAGACCGGCGCACGGGTGCGTAACGCGTATGCAATCTGCCTTTCACAGAGGGATAGCCCAGAGAAAATTTGGATTAATACCTCATA
>53_6128;size=1 reference=otu_00231,otu_00941,otu_00680 amplicon=1..300,1..300,1..30
CTGGCTCAGGATGAACGCGTAACGCGTATGCAATCTGCCTTTCACAGAGGGATAGCCCAGAGAAAATTTGGATTAATACCTCATA
>51_6860;size=1 reference=otu_00799,otu_00680 amplicon=1..300,1..300 position=1..300
GACGAAAGGCGCACGGGTGCGTAACGCGTATGCAATCTGCCTTTCACAGAGGGATAGCCCAGAGAAAATTTGGATTAATACCTCATA
```

#id	BHT0.LOT01	BHT0.LOT03	BHT0.LOT04	BHT0.LOT05	BHT0.LOT06	BHT0.LOT07
06_5949	0	0	0	0	0	0
56_3551	0	0	0	0	0	0
53_322	0	0	0	0	0	0
56_2589	0	0	0	0	0	0
56_7560	0	0	0	0	0	0
36_626	0	0	0	0	0	0
53_6128	0	0	0	0	0	0
51_6860	0	0	0	0	0	0
56_6896	0	0	0	0	0	0
56_3997	0	0	0	0	0	0
59_6	0	0	0	0	0	191
59_5144	0	0	0	0	0	1
59_5852	0	0	0	0	0	1
60_1696	0	0	0	0	0	0
59_6656	0	0	0	0	0	1
59_1182	0	0	0	0	0	1

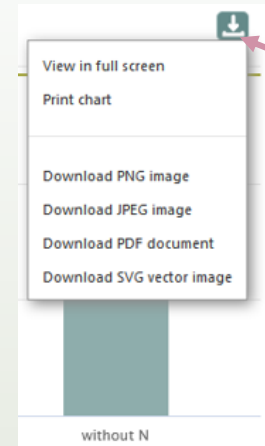
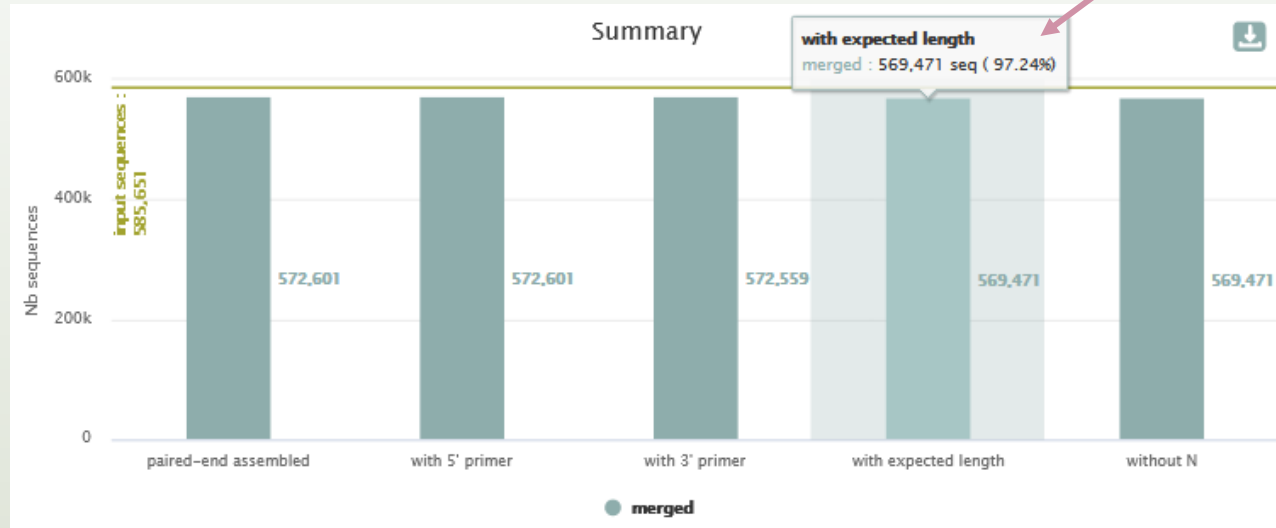
Fasta sequence of all clean and dereplicated sequence *i.e.* only one copy of each sequence is kept

count table for each sequence in each sample

## Answer 4

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

By moving the mouse over the graphic, new information appears



You can download graphics or table in different formats

### Details on merged sequences

Show 10 entries

Search:

CSV

Samples	before process	% kept	paired-end assembled	with 5' primer	with 3' primer	with expected length	without N
BHT0.LOT01	9,282	97.92	9,089	9,089	9,089	9,089	9,089
BHT0.LOT03	9,173	97.83	8,984	8,984	8,984	8,974	8,974
BHT0.LOT04	9,171	97.79	8,969	8,969	8,968	8,968	8,968 <sup>81</sup>

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

Q5: Who loose a lot of sequences ?

**53: FROGS Pre-process: report.html** [eye] [edit] [close]

error  
An error occurred with this dataset:

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

[back] [forward]

[bug] [download] [info] [refresh]

**52: FROGS Pre-process: count.tsv** [eye] [edit] [close]

**51: FROGS Pre-process: dereplicated.fasta** [eye] [edit] [close]

**Dataset generation errors**

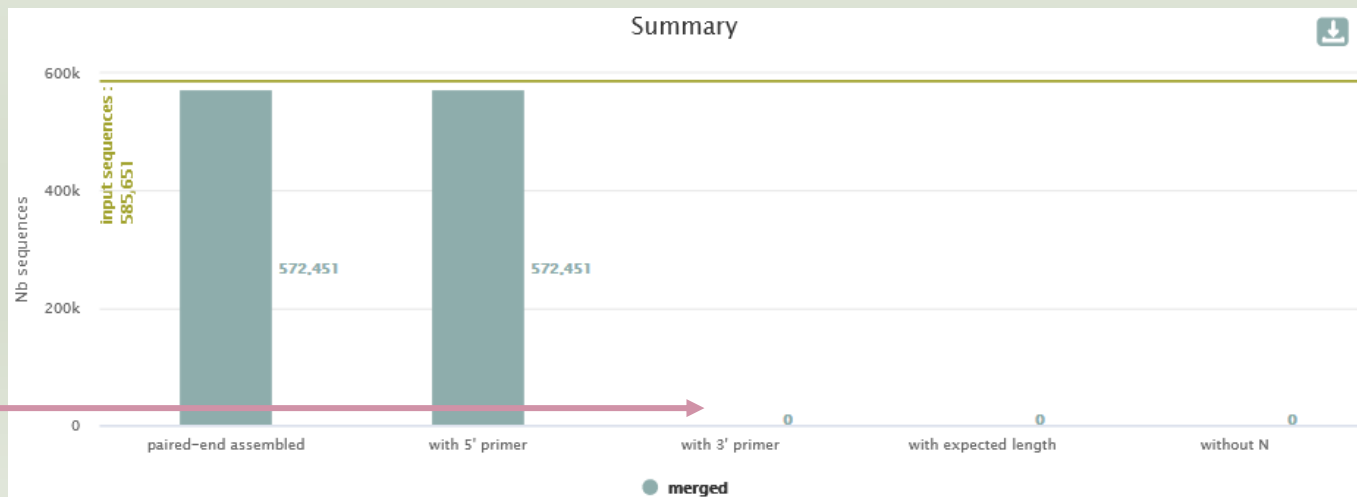
**Dataset 53: FROGS Pre-process: report.html**

Tool execution generated the following error message:

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

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

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



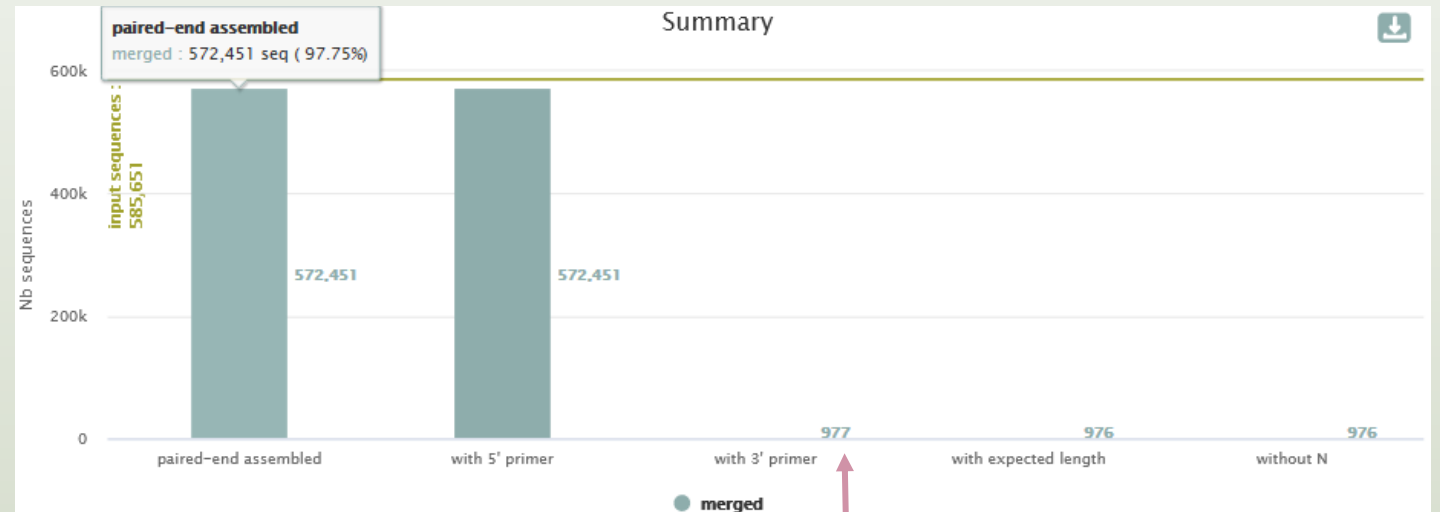
## Answer 5

All outputs are green  
but check the  
report.html

5: FROGS\_1 Pre-process: report.html

4: FROGS\_1 Pre-process: count.tsv

3: FROGS\_1 Pre-process: dereplicated.fasta



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

Sequencer

ILLUMINA

Select the sequencing technology used to produce the sequences.

Input type

TAR Archive

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

TAR archive file

1: chaillou\_withprimers\_64renamedsamples\_V1V3\_10000seq\_R1R2.tar.gz

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

Are reads already merged ?

No

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

Reads 1 size

300

The maximum read1 size.

Reads 2 size

300

The maximum read2 size.

Mismatch rate

0.1

The maximum rate of mismatches in the overlap

Merge software

Vsearch

Select the software to merge paired-end reads

Would you like to keep unmerged reads?

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

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

if your sequences have low qualities, you can increase this parameter  
But careful !

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

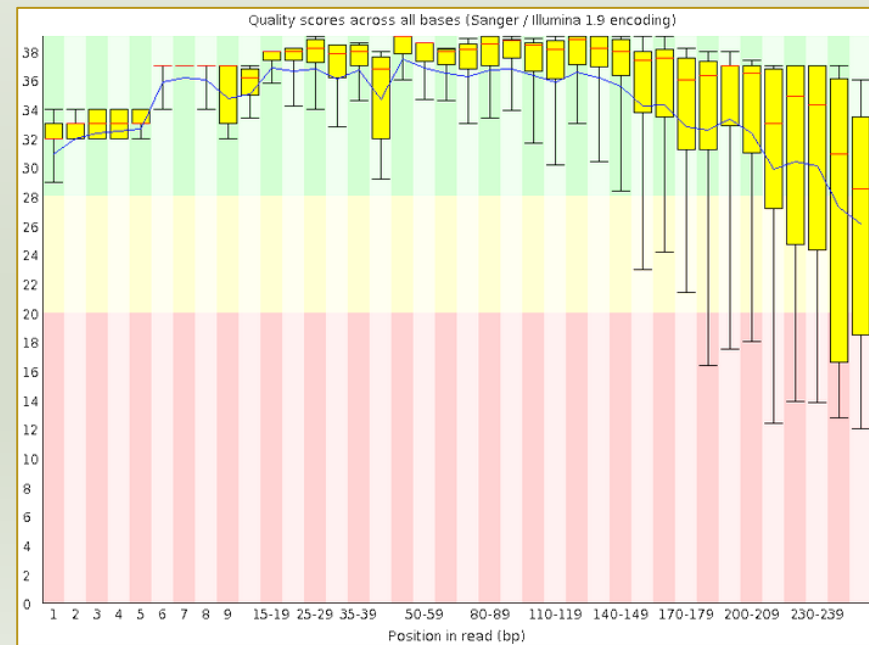
**Tools** ☆

fastqc ✕

⬆ Upload Data

👁 Show Sections

**FastQC Read Quality reports**

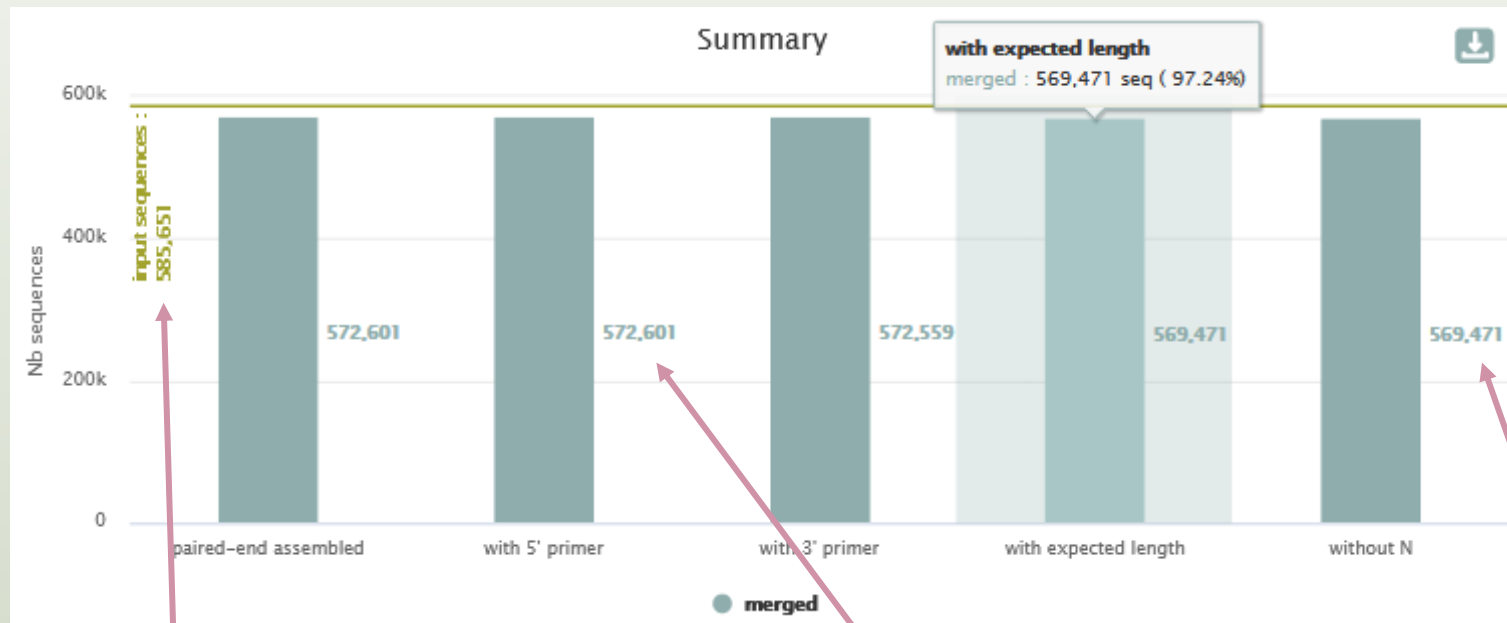


Answer 6, 7 & 8

Q6: How many sequences are there in the input file ?

Q7: How many sequences did not have the 5' primer?

Q8: How many sequences still are after pre-processing the data?



Total number of sequences before preprocessing: 585 651

All sequences have the 5' primer

569 471 sequences are still after preprocessing

## Answer 9

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

### 3: FROGS\_1 Pre-process: dereplicated.fasta

287,403 sequences

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

## Application

Software: preprocess.py (version: 4.1.0)

```
Command: /galaxydata/galaxy2021/galaxy02/galaxy/database/dependencies/_conda/envs/mulled-v1-916ab06f682ad01c3fd3dce3cb781eab380d0e2ea46de6f24ab32102bee0/bin/preprocess.py illumina --output-derep
```

Click on « i »

#### Dataset Information

Number	7
Name	FROGS_1 Pre-process: dereplicated.fasta
Created	Friday Sep 29th 15:59:54 2023 GMT+2
Filesize	152.5 MB
Dbkey	?
Format	fasta
File contents	contents
History Content API ID	46ff8dc5d79db92
History API ID	3e1619f828ae7909
UUID	be0a0235-6c54-4b64-85de-31c76e9cd875
Originally Created From a File Named	dereplicated.fasta

#### Tool Parameters

Input Parameter	Value
Sequencer	illumina
Input type	archive
TAR archive file	2 : chaillou_withprimers_64renamedsamples_V1V3_10000seq_R1R2.tar.gz
Are reads already merged ?	paired
Reads 1 size	300
Reads 2 size	300
Mismatch rate	0.1
Merge software	vsearch
Would you like to keep unmerged reads?	No, unmerged reads will be excluded.
Minimum amplicon size	400
Maximum amplicon size	500
Do the sequences have PCR primers?	true
5' primer	AGAGTTTGATCCTGGCTCAG
3' primer	CCAGCAGCCGCGGTAAT

Retrieve the tool parameters

#### Job Outputs

Tool Outputs	Dataset
FROGS_1 Pre-process: report.html	9 : FROGS_1 Pre-process: report.html
FROGS_1 Pre-process: count.tsv	8 : FROGS_1 Pre-process: count.tsv
FROGS_1 Pre-process: dereplicated.fasta	7 : FROGS_1 Pre-process: dereplicated.fasta

Stdout contains FROGS command lines and time execution

#### Job Information

Galaxy Tool ID:	toolshed.g2.bx.psu.edu/repos/frogs/frogs/FROGS_preprocess/4.1.0+galaxy3
Command Line	empty
Tool Standard Output	## Application Software: preprocess.py (version: 4.1.0) Command: /shared/ibfstor1/galaxy/utable-data/dependencies/_conda/envs/mulled-v1-916ab0...
Tool Standard Error	empty
Tool Exit Code:	0
Job API ID:	ecd2f76b136f78ca

Answer 10

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

## Details on merged sequences

Show  entries

 CSV

Search:

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

Select all samples

Answer 10

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

---

<input type="checkbox"/>	VHT0.LOT07	9,337	97.03	9,064	9,064	9,064	9,060	9,060
<input checked="" type="checkbox"/>	VHT0.LOT08	9,436	97.33	9,192	9,192	9,192	9,184	9,184
<input type="checkbox"/>	VHT0.LOT10	9,165	97.64	8,983	8,983	8,982	8,949	8,949

---

With selection:

Display amplicon lengths

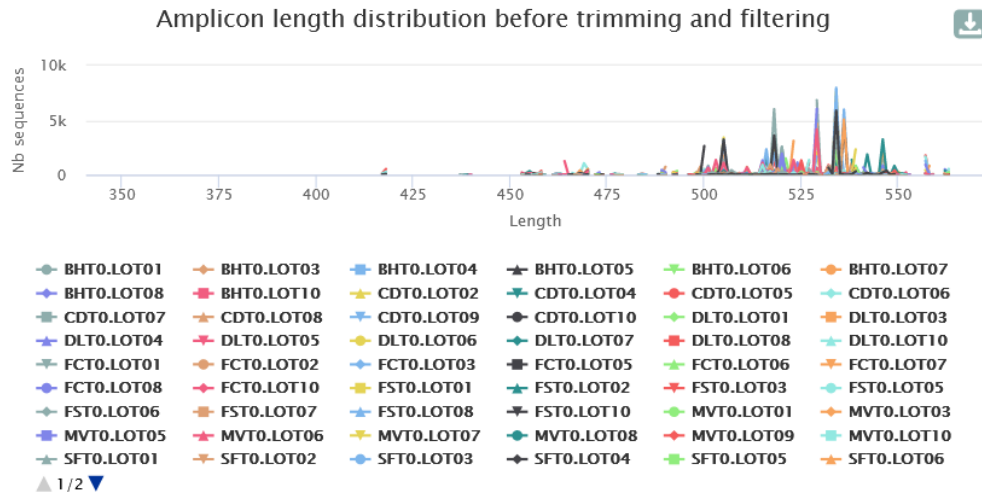
Display preprocessed amplicon lengths

at the bottom of the table

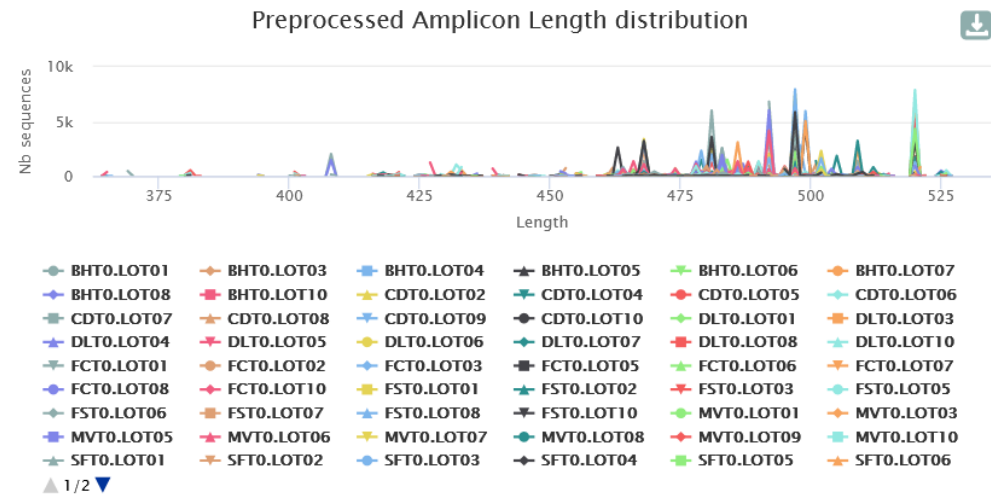
# Answer 10

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

Amplicons lengths



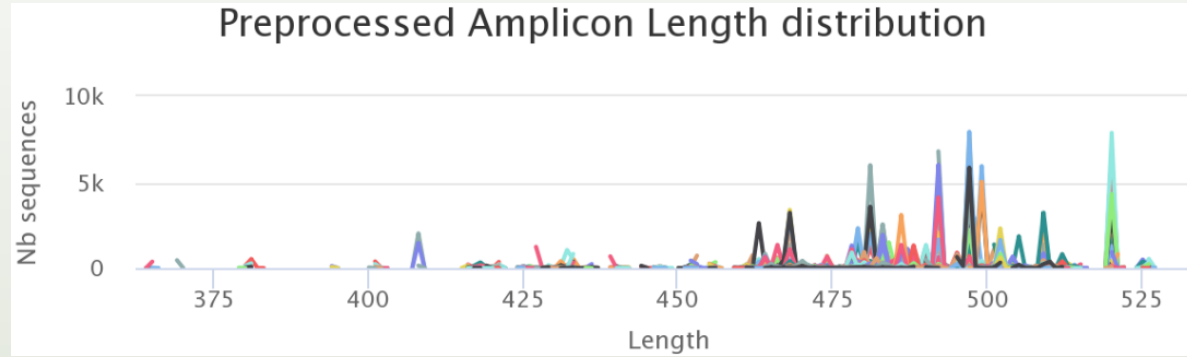
Amplicons lengths



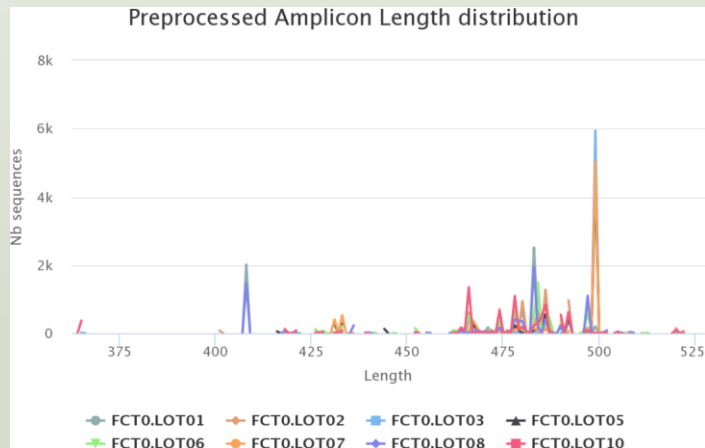
Before the preprocessing,  
 $400 < \text{sequence length} < 555$

After the preprocessing, the  
sequences were shortened

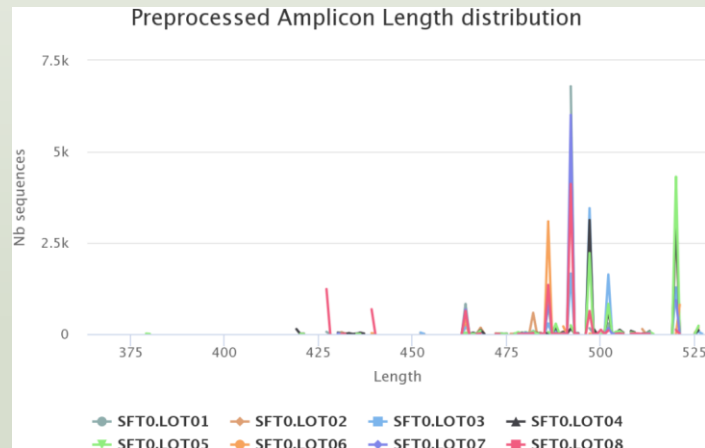
Q11: What can you tell about the samples, based on amplicon size distributions ?



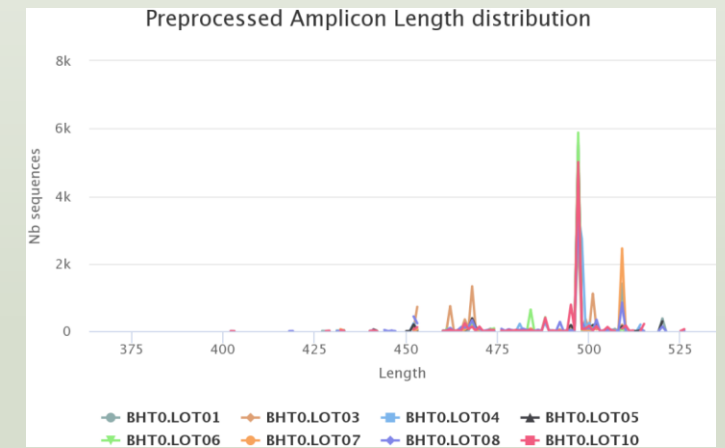
« Filet Cabillaud » samples



« Saumon Fumé » samples



« Bœuf Haché » samples



For each EnvType, we can observe different amplicon sizes. They correspond to different species.

*N.B.* amplicons with same size can represent different species.

---

# 2-Clustering tool

---

# Why do we need clustering ?

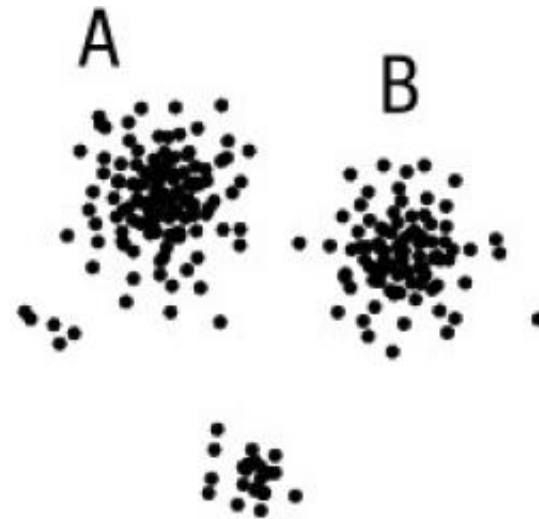
---

Amplification and sequencing are not perfect processes

- Polymerase error during PCR?
- Sequencing errors ?
- Natural variations ?



Expected

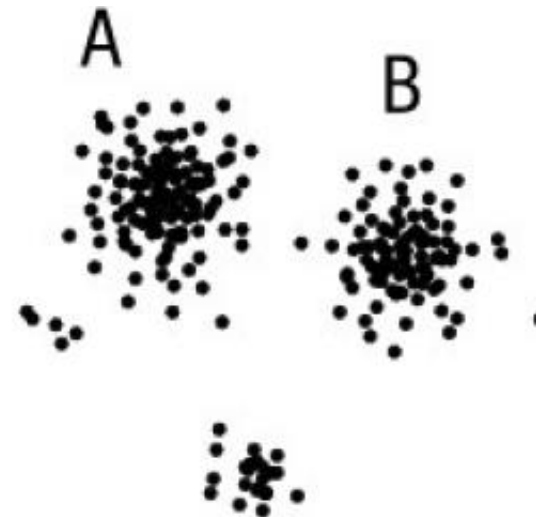


Results

Natural variability?  
Technical noise?  
Contaminant?  
Chimeras?



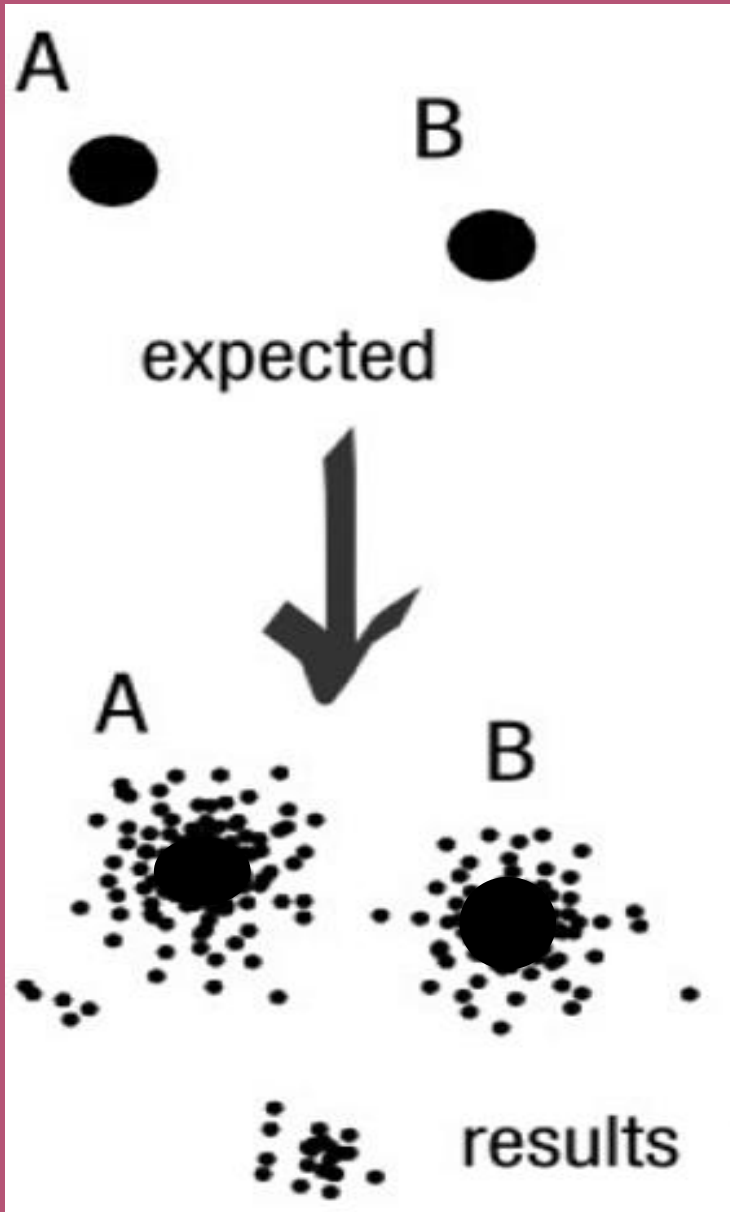
Expected



Results

Natural variability ?  
Technical noise?  
Contaminant?  
Chimeras?

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



# To have the best accuracy:

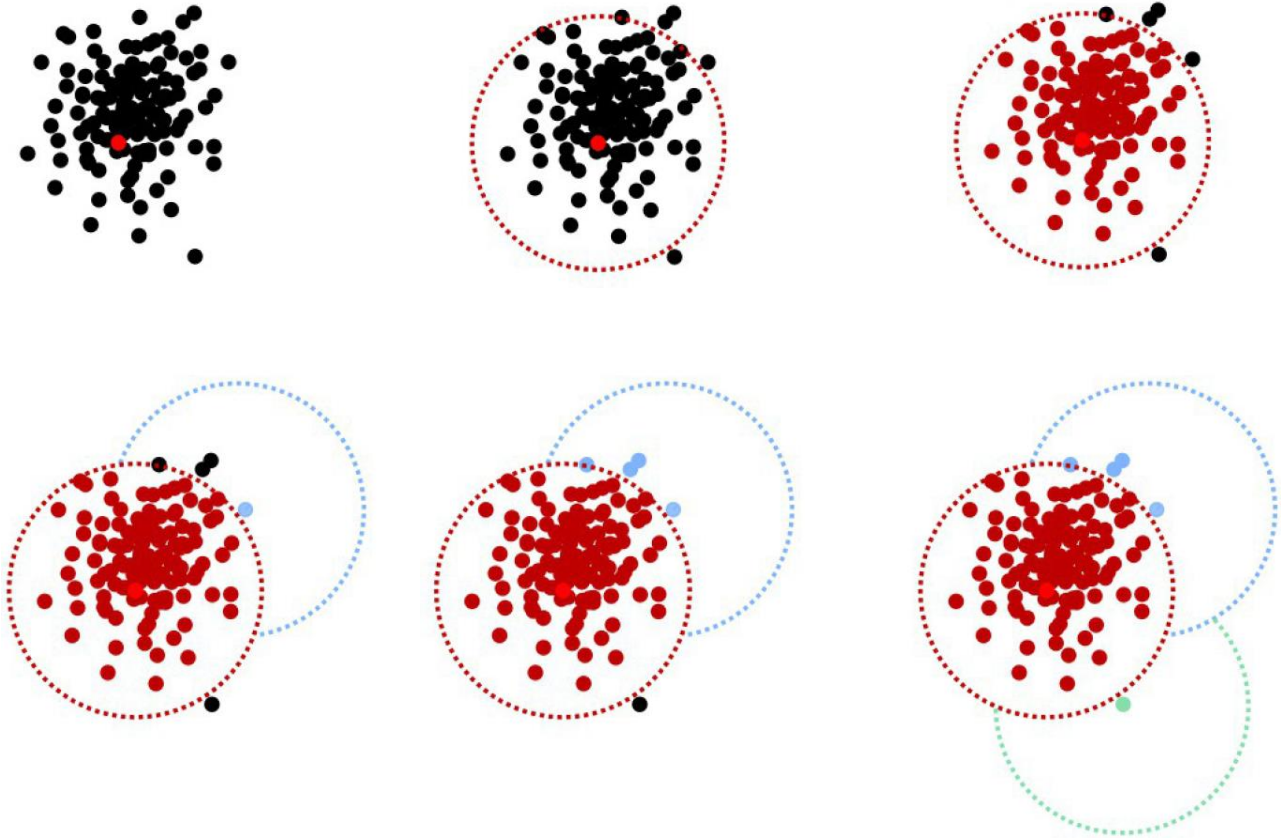
Method: All against all

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

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

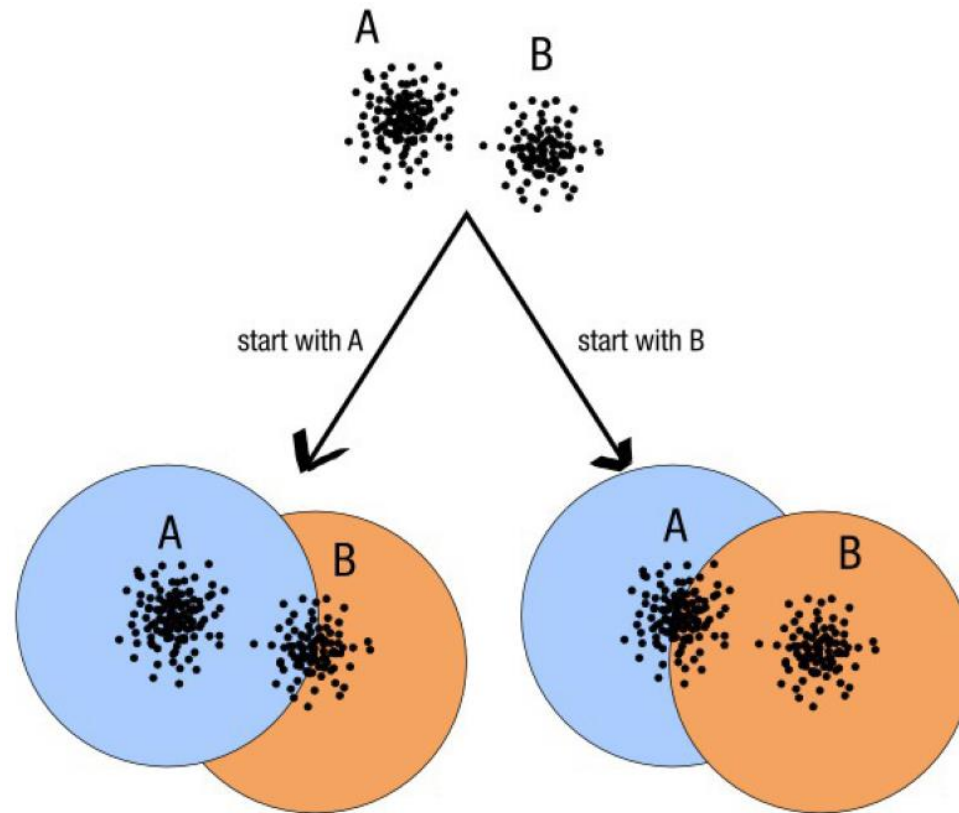
# How traditional clustering works ?

---



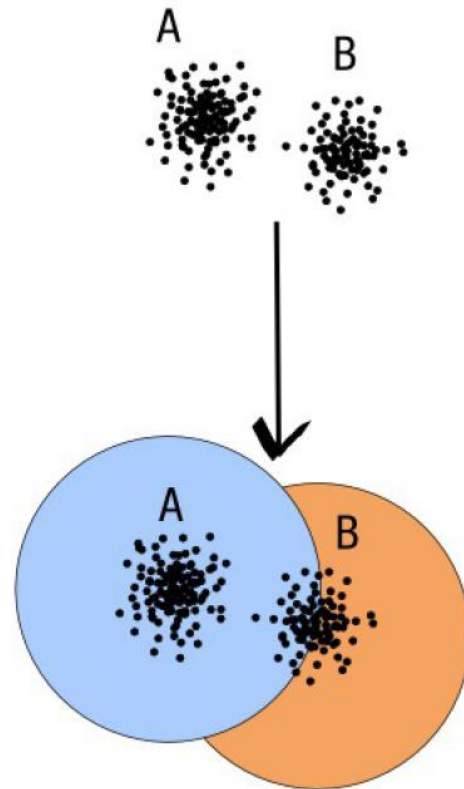
# Input order dependent results

---

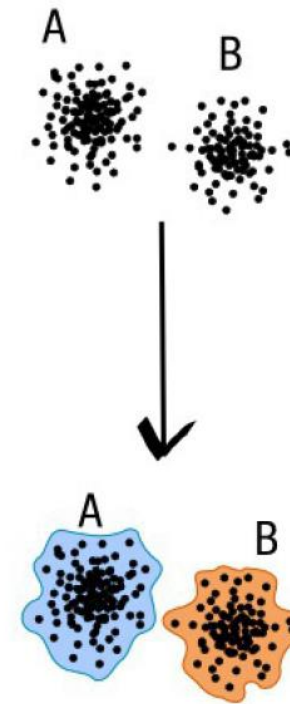


decreasing length,  
decreasing abundance,  
external references

# Single a priori clustering threshold

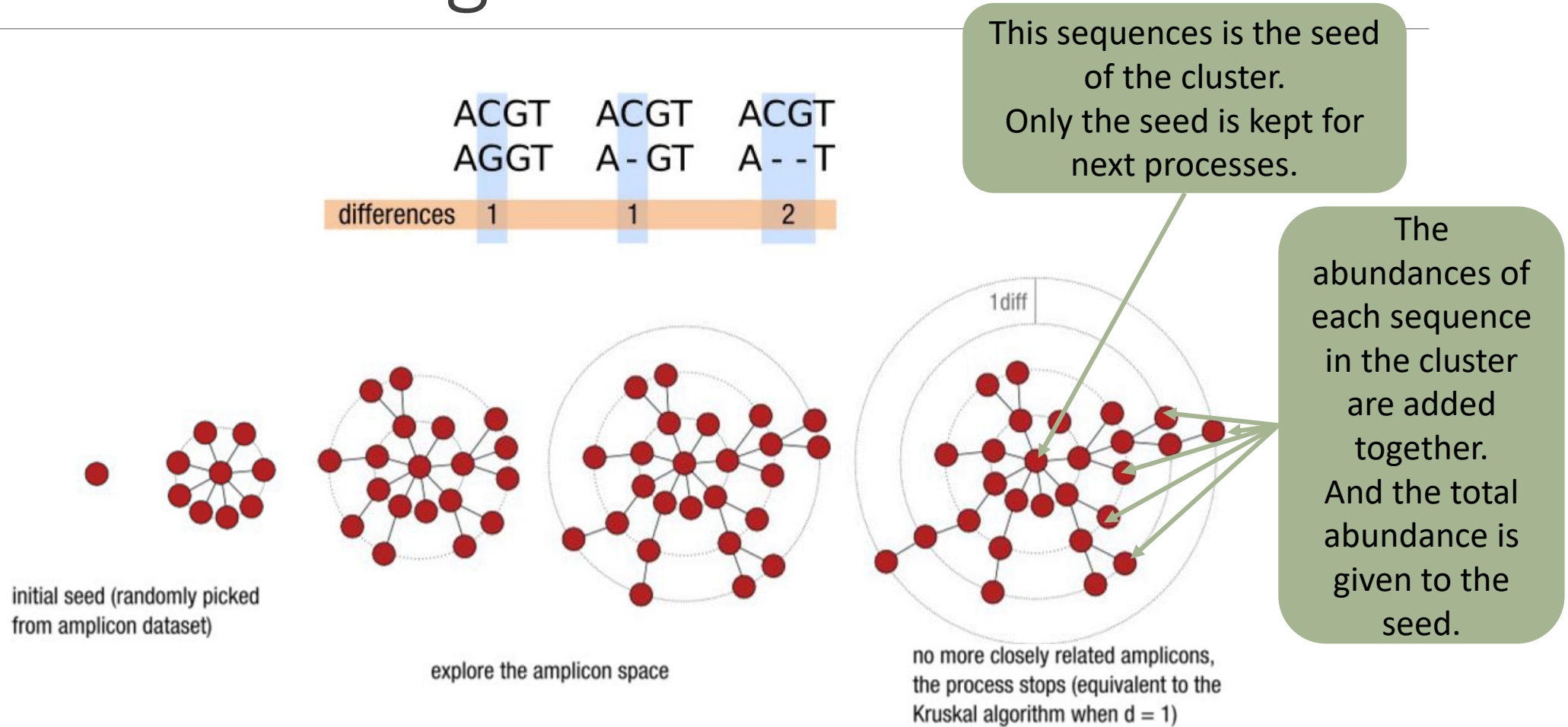


compromise threshold  
unadapted threshold



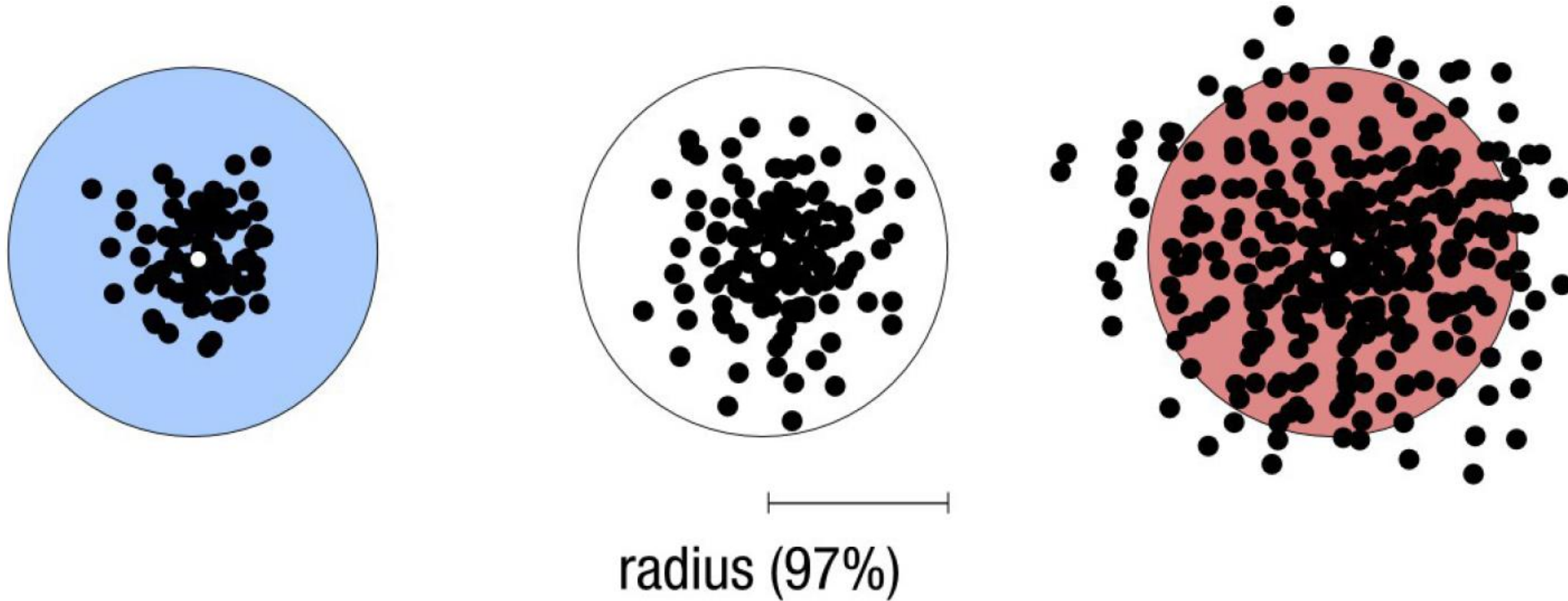
natural limits of clusters

# Swarm clustering method



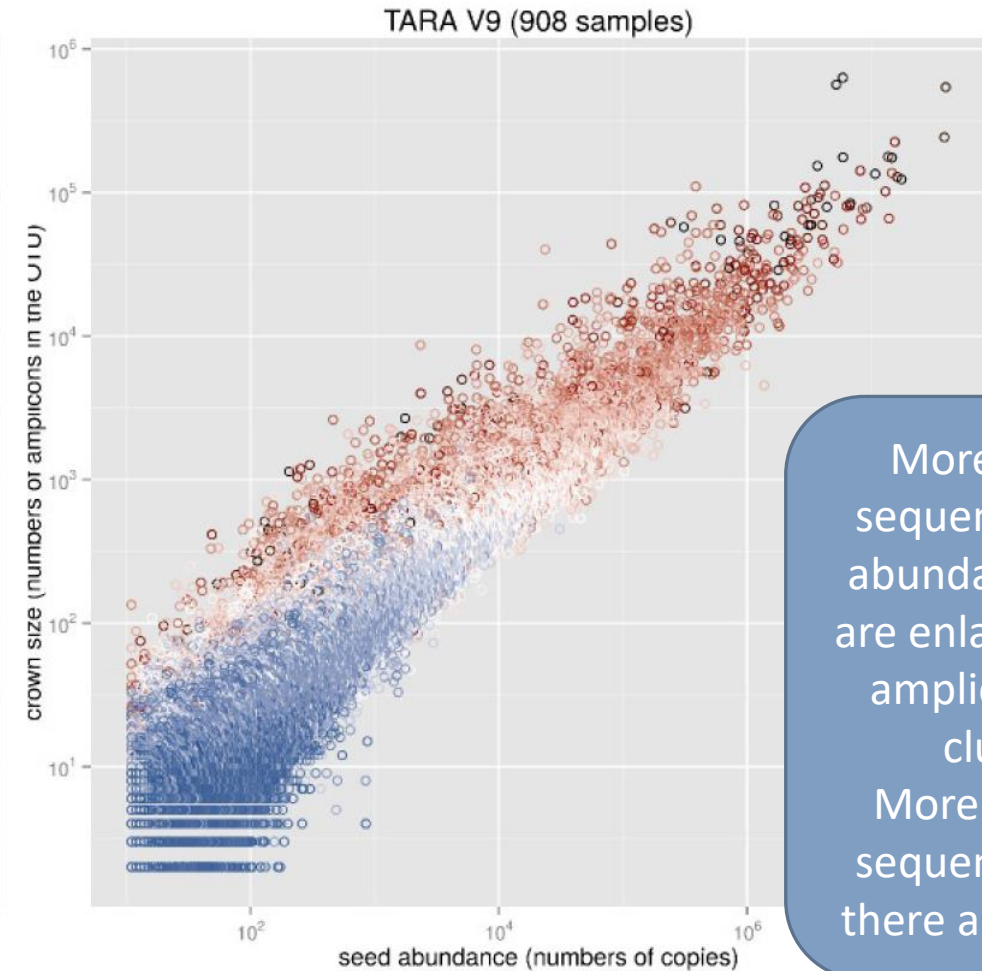
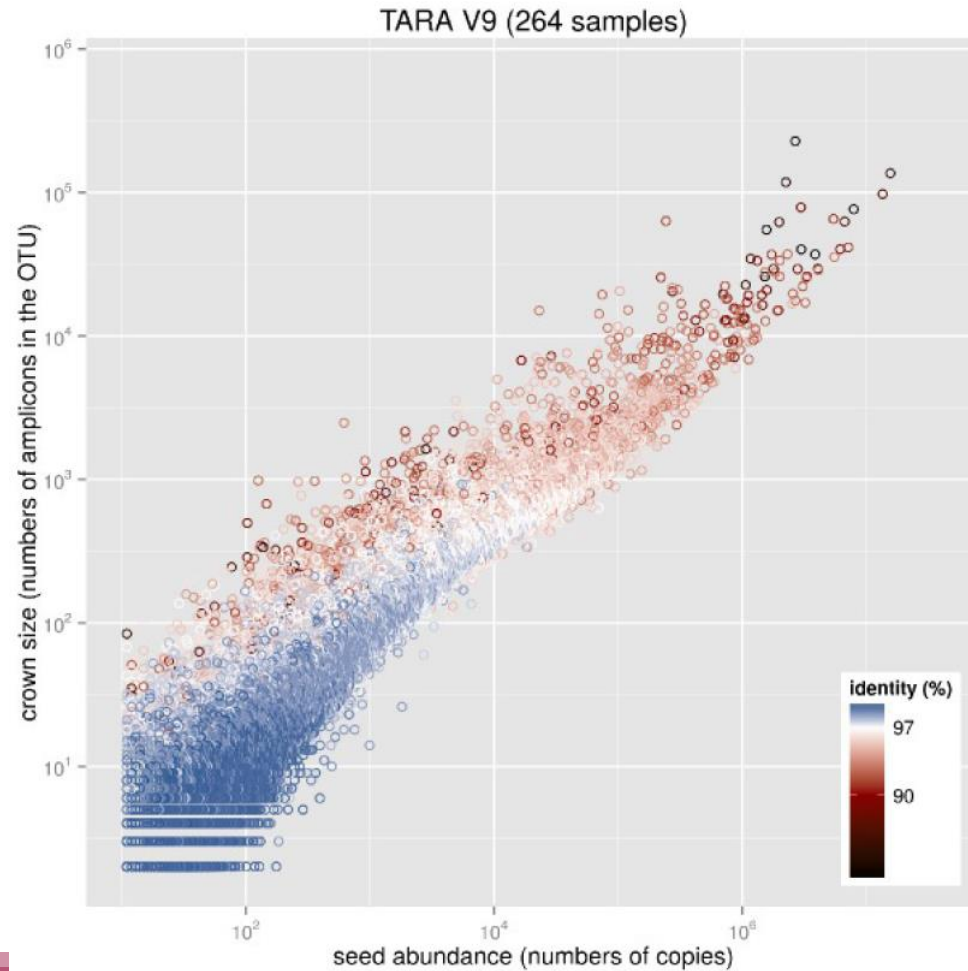
# Comparison Swarm and 3% clusterings

---



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

**FROGS\_2 Clustering swarm** Single-linkage clustering on sequences  
(Galaxy Version 4.1.0+galaxy1)

☆ Favorite    🔄 Versions    ▼ Options

**Sequences file**

3: FROGS\_1 Pre-process: dereplicated.fasta

The dereplicated sequences file (format: FASTA)

**Count file**

4: FROGS\_1 Pre-process: count.tsv

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

**FROGS guidelines version**

New guidelines from version 3.2

The denoising step before a d3 clustering is no longer recommended since FROGS 3.2, but you can still choose it.

**Aggregation distance clustering**

1

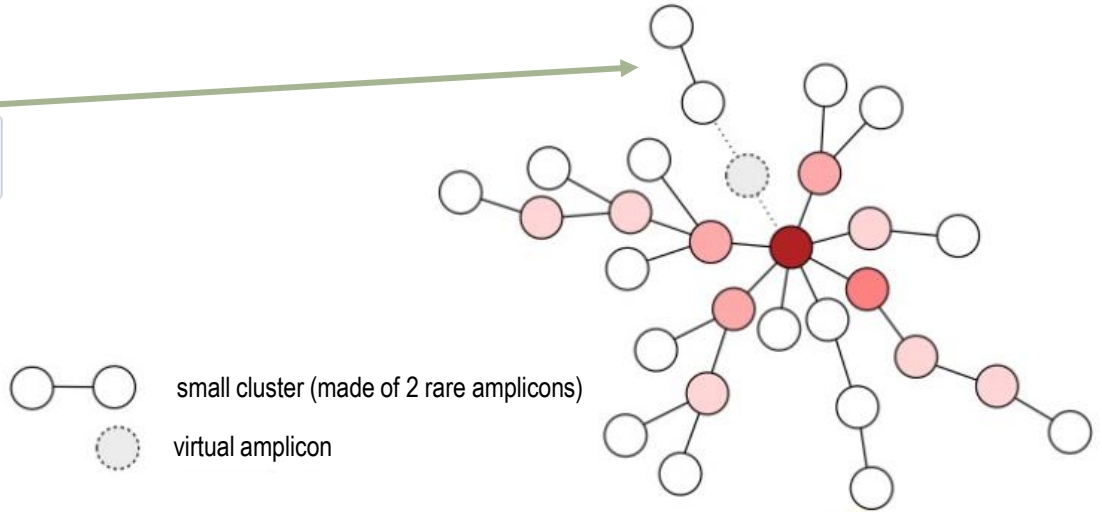
Maximum number of differences between sequences in each aggregation Swarm step. (recommended d=1) (--distance)

**Refine clustering**

Yes, refine clustering with --fastidious swarm option  
 No, perform clustering without refinement

Clustering will be performed with the Swarm --fastidious option. It is recommended and only usable in association with a distance of 1 (default and recommended: Yes) (--fastidious)

longer but more accurate



---

# Cluster stat tool

---

A RECURRENT TOOL

---

**FROGS\_Cluster\_Stat** Process some metrics on clusters (Galaxy Version 4.1.0+galaxy1)

**Abundance file**



7: FROGS\_2 Clustering swarm: clustering\_abundance.biom

Clusters abundance (format: BIOM)

---

# Practice:

---

LAUNCH CLUSTERING AND CLUSTERSTAT TOOLS

# Exercise

---

Go to « 16S » history

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

→ objectives :

- understand the outputs from clustering
- understand the ClusterStat utility

# Exercise

---

1. How many clusters do you get ?

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

FROGS Clusters stat Process  
some metrics on clusters.

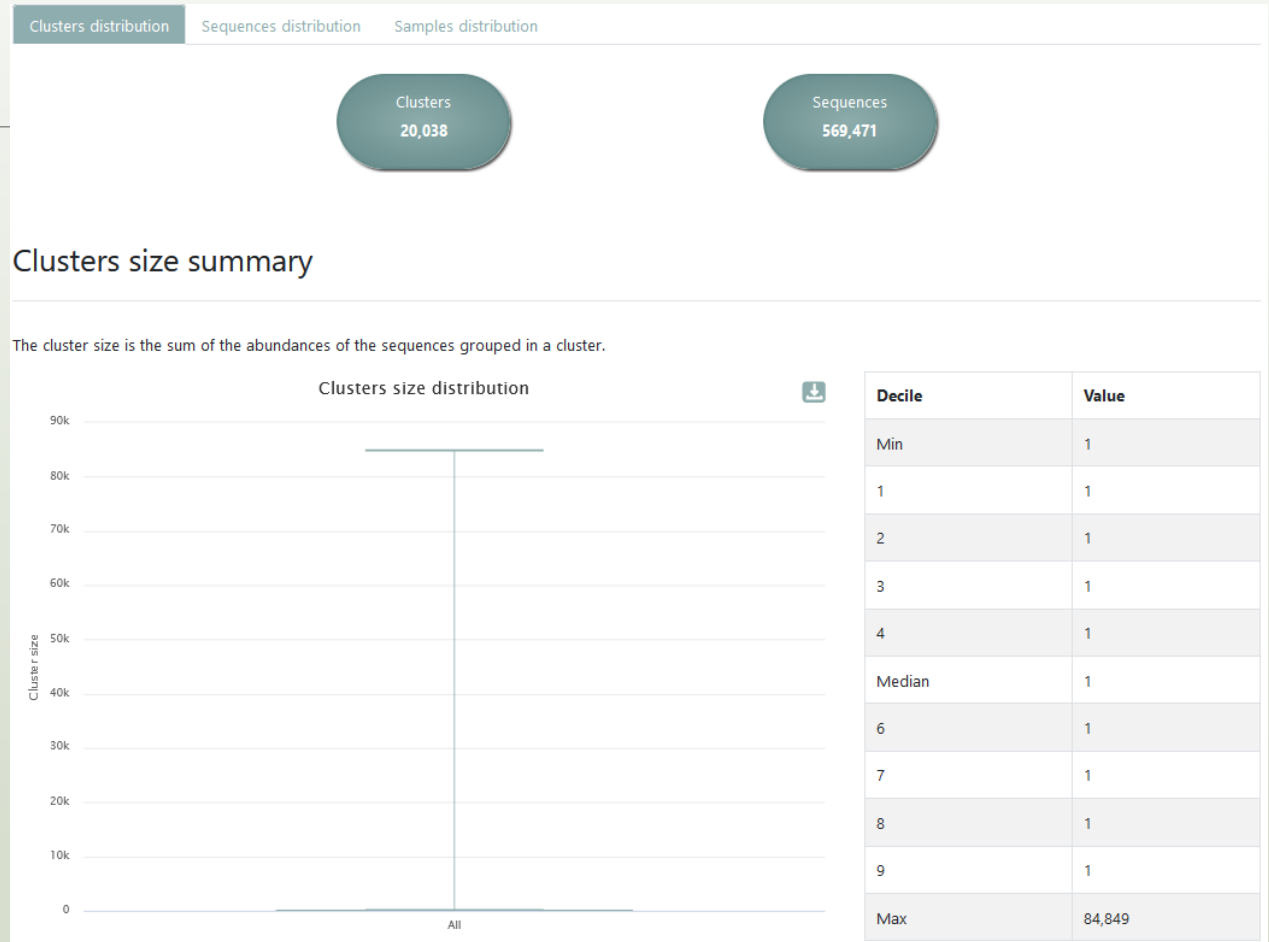
# Exercise

---

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

Answer 1, 2  
& 3

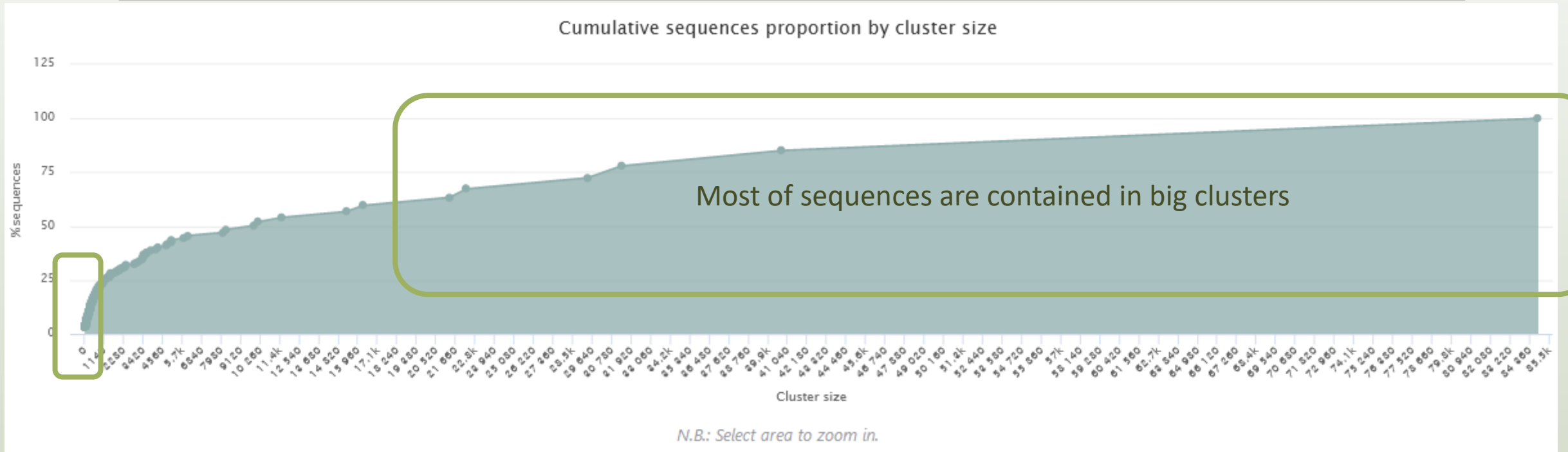
- Q1: How many clusters do you get ?  
Q2: Interpret the boxplot: **Clusters size summary**  
Q3: Interpret the table: **Clusters size details** -  
**How many single singletons do you find?**



Most of clusters are singletons

Answer 4

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



The small clusters represent few sequences

Answer 5 to 9

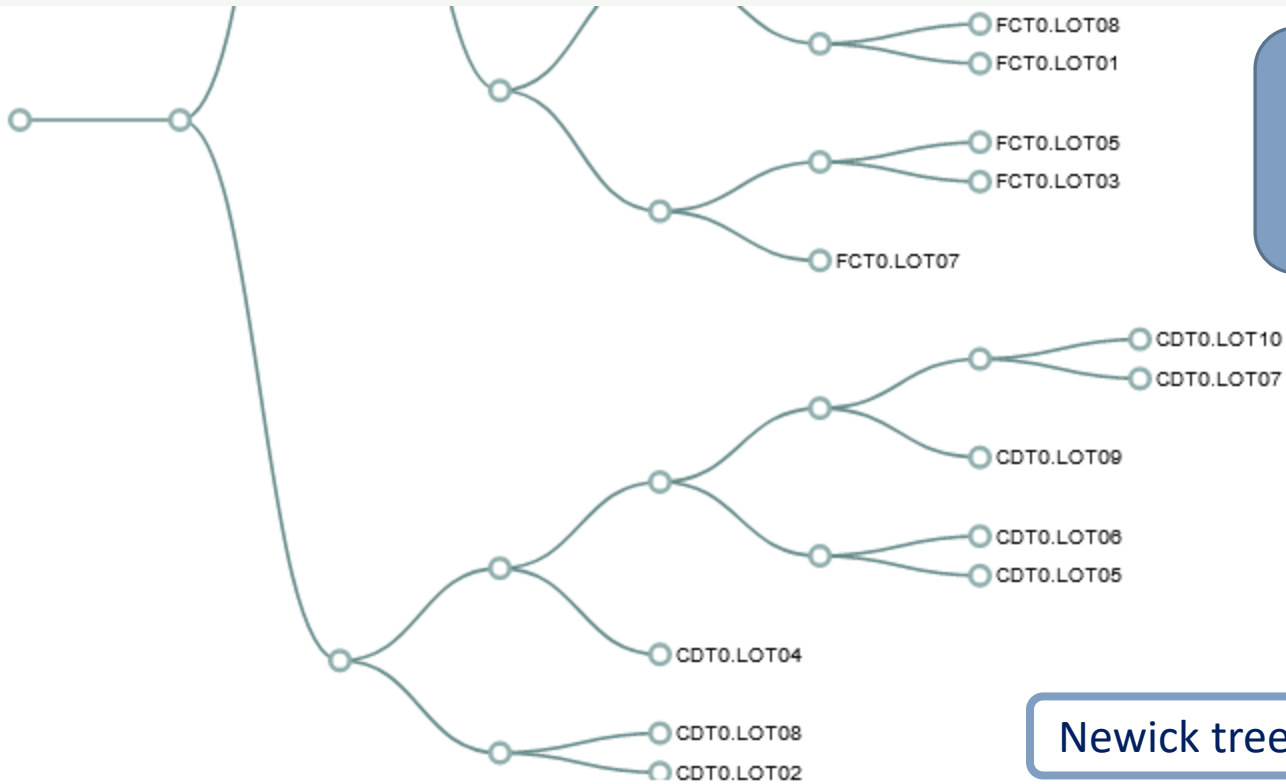
	Total clusters	Shared clusters	Own clusters	Total sequences	Shared sequences	Own sequences
BHT0.LOT01	493	114	379	9,089	8,709	380
BHT0.LOT03	433	140	293	8,937	8,630	307
BHT0.LOT04	474	152	322	9,270	8,767	503
BHT0.LOT05	475	152	323	8,918	8,609	309
BHT0.LOT06	490	156	334	8,520	8,377	143
BHT0.LOT07	531	165	366	8,373	8,264	109
BHT0.LOT08	430	201	229	8,715	8,486	229
BHT0.LOT10	477	157	320	8,937	8,630	307
CDT0.LOT02	477	157	320	9,270	8,767	503
CDT0.LOT04	477	157	320	8,918	8,609	309
CDT0.LOT05	384	241	143	8,520	8,377	143
CDT0.LOT06	365	256	109	8,373	8,264	109
CDT0.LOT07	512	100	412	8,715	8,486	229
CDT0.LOT08	556	162	394	8,937	8,630	307

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

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

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

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



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

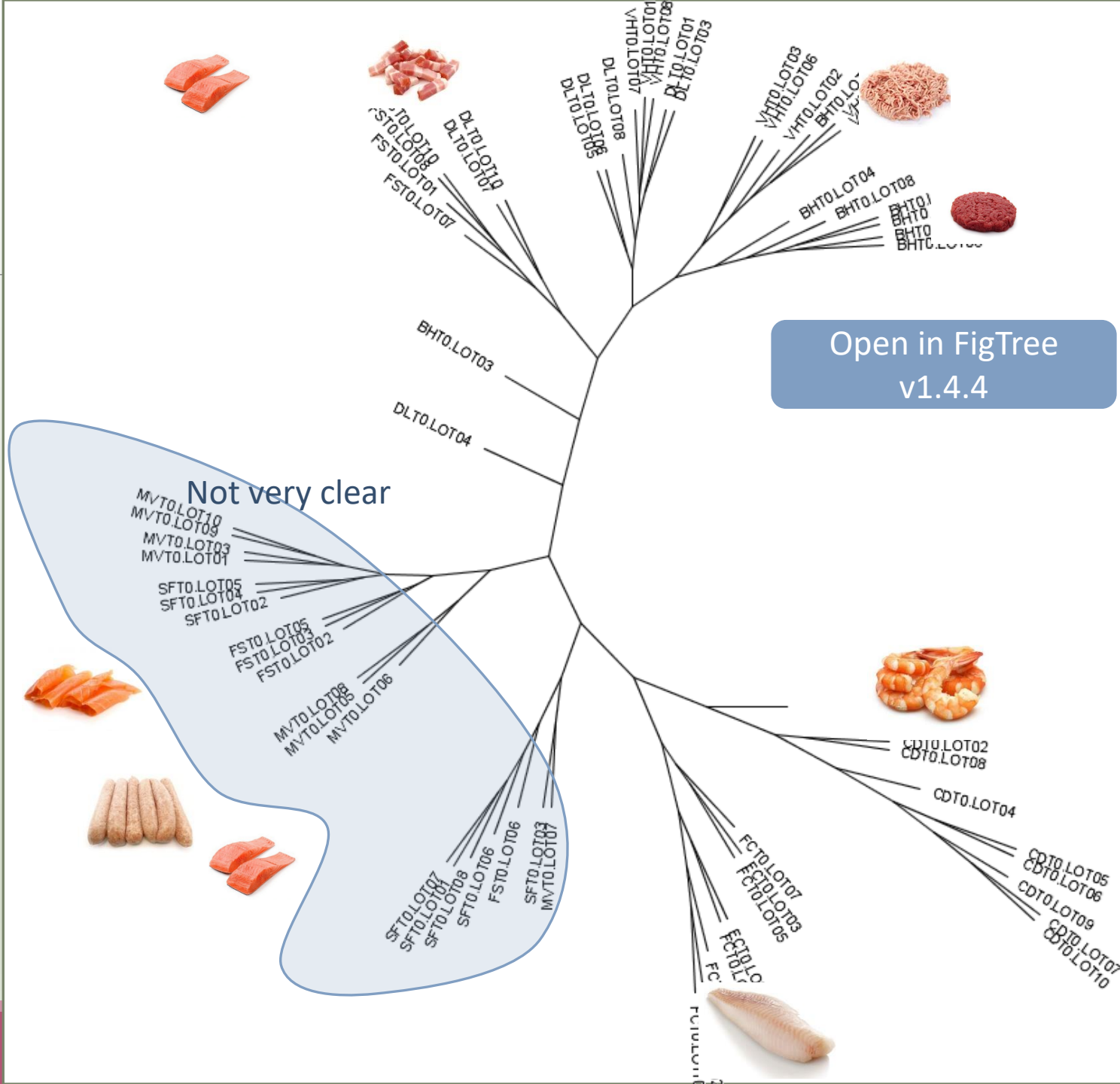
Newick tree available too, can be copied and pasted an tree viewer

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

Answer 10

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

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



---

# 3-Chimera removal tool

---

# What is chimera ?

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

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

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

*aborted amplification*



*next cycle's "primer"*



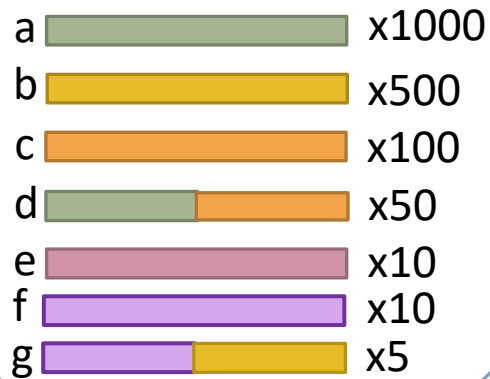
*chimeric sequence*



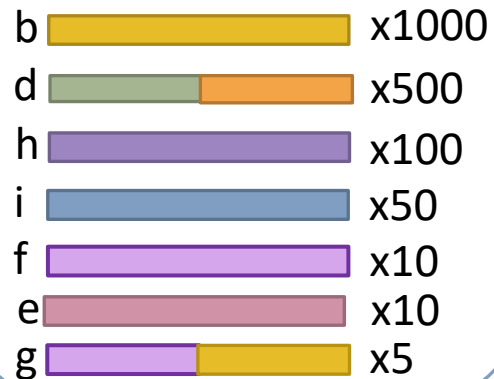
# A smart removal chimera to be accurate

We use a sample cross-validation

Sample A



Sample B



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

“d” is view as normal sequence by Vsearch  
because it have not “parents”.

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

---

# Practice:

---

LAUNCH THE REMOVE CHIMERA TOOL

# Exercise

---

Go to « 16S » history

Launch the « FROGS\_3 Remove Chimera » tool

Follow by the « FROGS ClusterStat » tool

→ objectives :

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

**FROGS\_3 Remove chimera** Remove PCR chimera in each sample (Galaxy Version 4.1.0+galaxy1)

**Sequences file (format: FASTA)**



6: FROGS\_2 Clustering swarm: seed\_sequences.fasta

The sequences file

**Abundance type**

BIOM file

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

**Abundance file (format: BIOM)**



7: FROGS\_2 Clustering swarm: clustering\_abundance.biom

It contains the count by sample for each sequence.

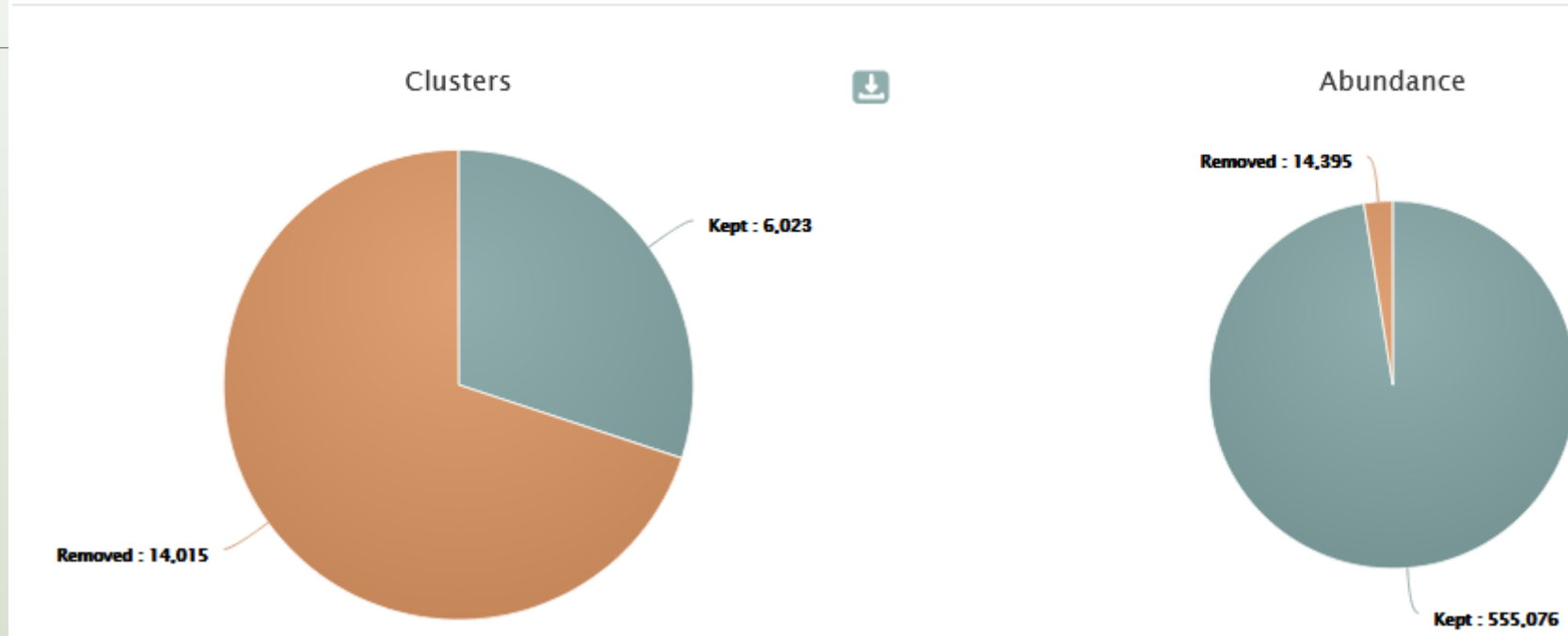
# Exercise

---

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

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

## Remove summary



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

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

Removed clusters are low  
 abundance clusters.

Answer 2

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

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

The largest cluster of chimeras contained 19 sequences.

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

### Answer 3

Q3a: Of what are mainly composed singleton ? (compare with previous report.html)

Q3b: What are their abundance?

Q3c: What do you conclude ?

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

Cluster\_Stat report  
after clustering

Most small clusters  
are composed of  
chimeras

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

Cluster\_Stat report after  
chimera removing

---

# 4- Cluster Filter tool

---

# 4- Cluster Filter

---

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

**Criteria:**

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

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

**Biggest Cluster :** Only the X biggest are conserved.

**Contaminant:** If cluster sequence matches with phiX, or the whole chromosomes chloroplastic and mitochondrial of *A. Thaliana* or your own contaminant sequence.

## One tool, 4 criteria

## Sequence file

The sequence file to filter (format: FASTA)

## Abundance file

The abundance file to filter (format: BIOM)

## Minimum prevalence method

## Minimum prevalence

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

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

## Minimum proportion of sequences abundance to keep cluster

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

## N biggest clusters

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

## Search for contaminant clusters.

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

1

# Prevalence filter – option 1

**FROGS\_4 Cluster filters** Filters clusters on several criteria. (Galaxy Version 4.1.0+galaxy1) ☆ Favorite ▼ Options

**Sequences file**

📁

The sequence file to filter (format: FASTA)

**Abundance file**

📁

The abundance file to filter (format: BIOM)

**Minimum prevalence method**

**Minimum prevalence**

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

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

1

# Prevalence filter – option 2

**FROGS\_4 Cluster filters** Filters clusters on several criteria. (Galaxy Version 4.1.0+galaxy1) Favorite Options

**Sequences file**  
9: FROGS Remove chimera: non\_chimera.fasta  
The sequence file to filter (format: FASTA)

**Abundance file**  
10: FROGS Remove chimera: non\_chimera\_abundance.biom  
The abundance file to filter (format: BIOM)

**Minimum prevalence method**  
replicate identification Need to know group composition

**File of replicated sample names**  
12: chaillou\_replicate\_information.tsv  
Replicate file to link each sample to its group (cf. Help section).

**Minimum prevalence**  
0.5  
Fill the field only if you want this treatment. Keep OTU present in at least this proportion of replicates in at least one group (must be a proportion between 0 and 1).

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

1

# Prevalence filter – option 2

## How to build the file of replicated sample names ?

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

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

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

Example:

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

Thanks to get data tool,  
add it in your history

# 1 Prevalence filter – option 2

## Results:

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

The process will therefore keep the clusters present in at least

2 "rich" samples

3 "richAB" samples,

1 "lowAB" sample

1 "april21" sample

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

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

1

# Prevalence filter – option 2

mistakes not to be made:

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

valid

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

Creates artificially 3 columns

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

Creates artificially 3 columns

2

# Cluster size filter

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

as proportion

Minimum proportion of sequences abundance to keep cluster

5e-05

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

OR

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

as count

Minimum number of sequences to keep cluster

2

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

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

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

3

# Filter : Keep biggest cluster

---

**N biggest clusters**

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

Here, user wants to keep the 50 biggest clusters.



---

# Practice:

---

LAUNCH THE CLUSTER FILTER TOOL

# Exercise:

---

Go to history « 16S » history

Launch « cluster Filter » tool with non\_chimera\_abundance.biom, non\_chimera.fasta

Use 3 criteria to filter clusters:

- cluster must be present at least in 4 samples
- Each cluster must represented a minimum of 0.005 % = 0.00005 <sup>(1)</sup> of the totality of the sequences
- cluster of phiX <sup>(2)</sup> must be removed

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

<sup>(1)</sup> *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.

<sup>(2)</sup> <https://www.illumina.com/products/by-type/sequencing-kits/cluster-gen-sequencing-reagents/phix-control-v3.html>

# Exercise:

---

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

Answer 1

FROGS\_4 Cluster filters Filters clusters on several criteria. (Galaxy Version 4.1.0+galaxy1)

☆ Favorite

▼ Options

### Sequence file

10: FROGS\_3 Remove chimera: non\_chimera.fasta

The sequence file to filter (format: FASTA)

### Abundance file

11: FROGS\_3 Remove chimera: non\_chimera\_abundance.biom

The abundance file to filter (format: BIOM)

### Minimum prevalence method

all samples

### Minimum prevalence

4

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

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

as proportion

### Minimum proportion of sequences abundance to keep cluster

0.00005

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

### Keep N biggest clusters

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

### Search for contaminant clusters.

Use contaminant FASTA file from the server

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

### Contaminant databank

phiX

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

Outputs

17: FROGS\_4 Cluster filters: report.html

16: FROGS\_4 Cluster filters: excluded.tsv

15: FROGS\_4 Cluster filters: clusterFilters\_sequences.fasta

14: FROGS\_4 Cluster filters: clusterFilters\_abundance.biom

0.005% = 0.00005

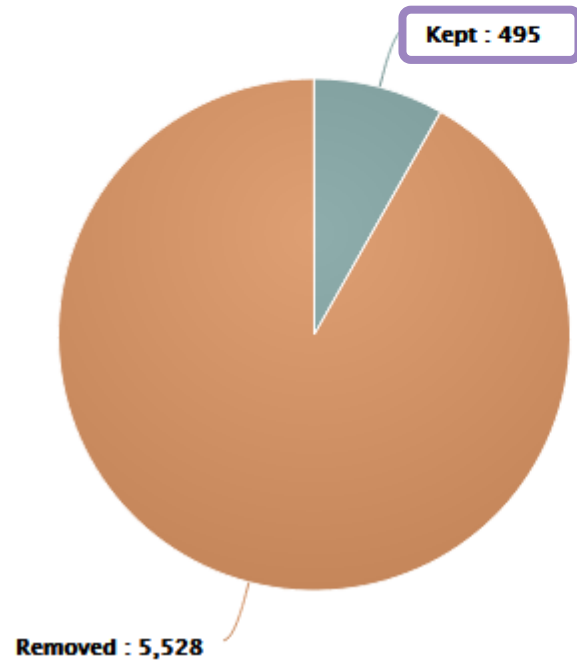
Two tabs to explore

Filters by ASVs

Filters by samples

## Filters summary

ASVs

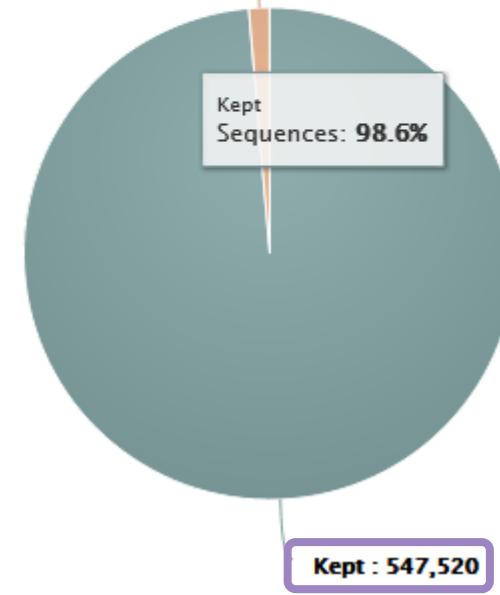


Abundance

Removed : 7,556

Kept Sequences: 98.6%

Kept : 547,520



## Answer 2

Filters by ASVs

Filters by samples



### Details by samples

Show 10 entries

You can sort the table by header

Search:

CSV

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

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

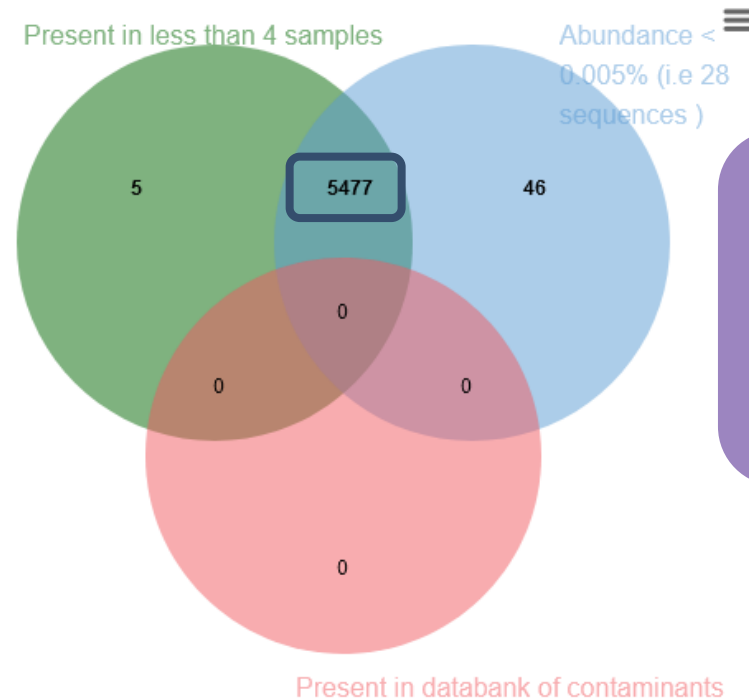
## Filters intersections

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

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

Venn

### Venn on removed ASVs



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

## Answer 4

report.html of ClusterStat tool

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

Clusters distribution Sequences distribution Samples distribution

### Sequences count

Show  entries Search:  [CSV](#)

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

# Overview

---

1. Preprocessing
2. Clustering without fixed-threshold
3. Remove chimera
4. Cluster filters
  - ASV - Amplicon Sequence Variant

---

OTU → ASV

---

# A long-standing discussion

---

The ASV vs OTU debate launched by the arrival of dada2 is not so new and had been bothering us for several months/years.

In fact, the debate largely preceded the term "ASV", and is precisely what made us opt for Swarm in FROGS (just under 10 years ago).

To quote the author of swarm:

“The traditional term "OTU" is negatively charged nowadays. The ASV vs OTU debate is creating confusion in the community and some users now think that all methods producing "OTUs" use a fixed clustering threshold (i.e. 97%-similarity) and are inherently bad. Of course, this is not the case and there are several methods published before the ASV term was coined that produce ASV-like clusters, swarm included.” To avoid that confusion, swarm's manual now only uses the generic term "cluster".

<https://github.com/torognes/swarm/commit/0bb491f9bf646c22a5363c27dc31a6d4b2ad335d> “

# A question of vocabulary

---

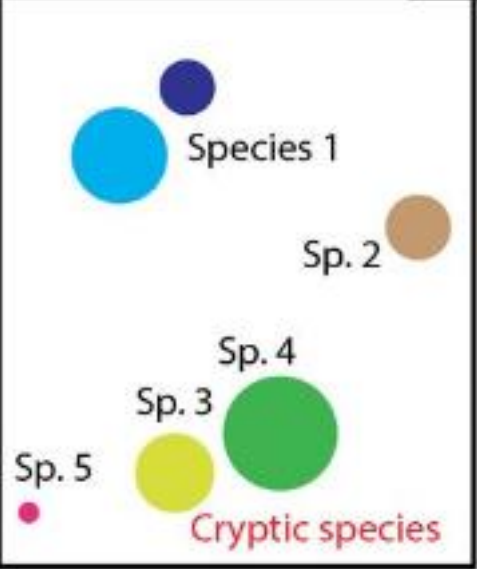
A few years ago, the **semantic problem was the opposite**, and any method that didn't produce OTUs was questioned or even disqualified.

At the start of FROGS, we therefore chose to call our clusters "OTUs" at the end of the analysis (once the filters had been applied), but it's only a question of vocabulary, and the **clusters produced by FROGS/swarm are very close to ASV in their construction**.

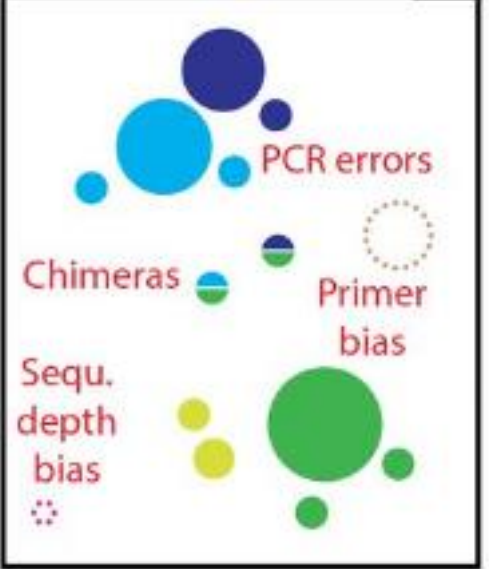
In any case, they look much more like ASVs than "fixed threshold" OTUs. The best thing would have been to use a new term, but **Frédéric Mahé** didn't make that choice at the time introducing **a new term** could have led to **confusion**.

Since version 4.1.0 of FROGS, we have changed our vocabulary and all OTU terms have been changed to **cluster** or **ASV** in FROGS tools and outputs.

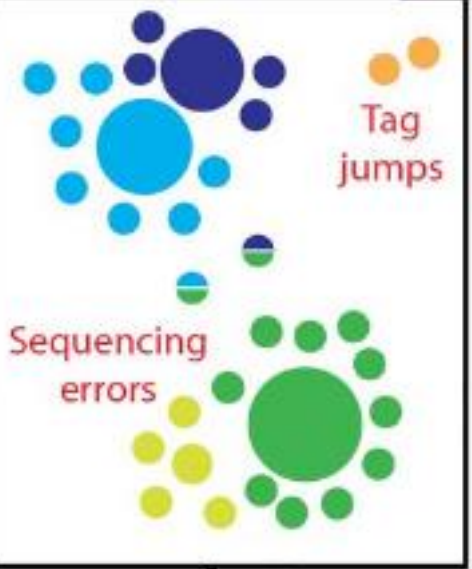
A) Bulk sample PCR



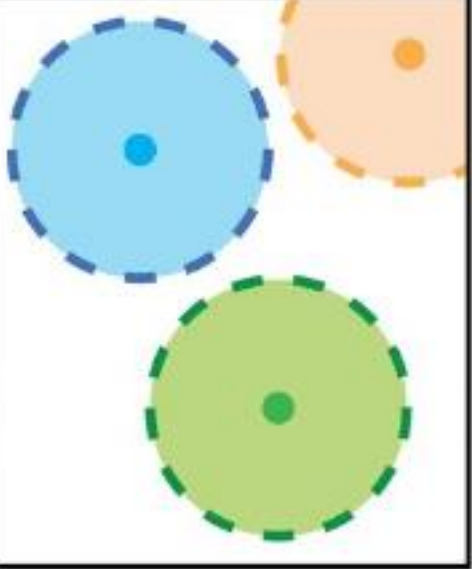
B) Amplicons HTS



C) Sequences Bio inf.



D) OTUs 97%

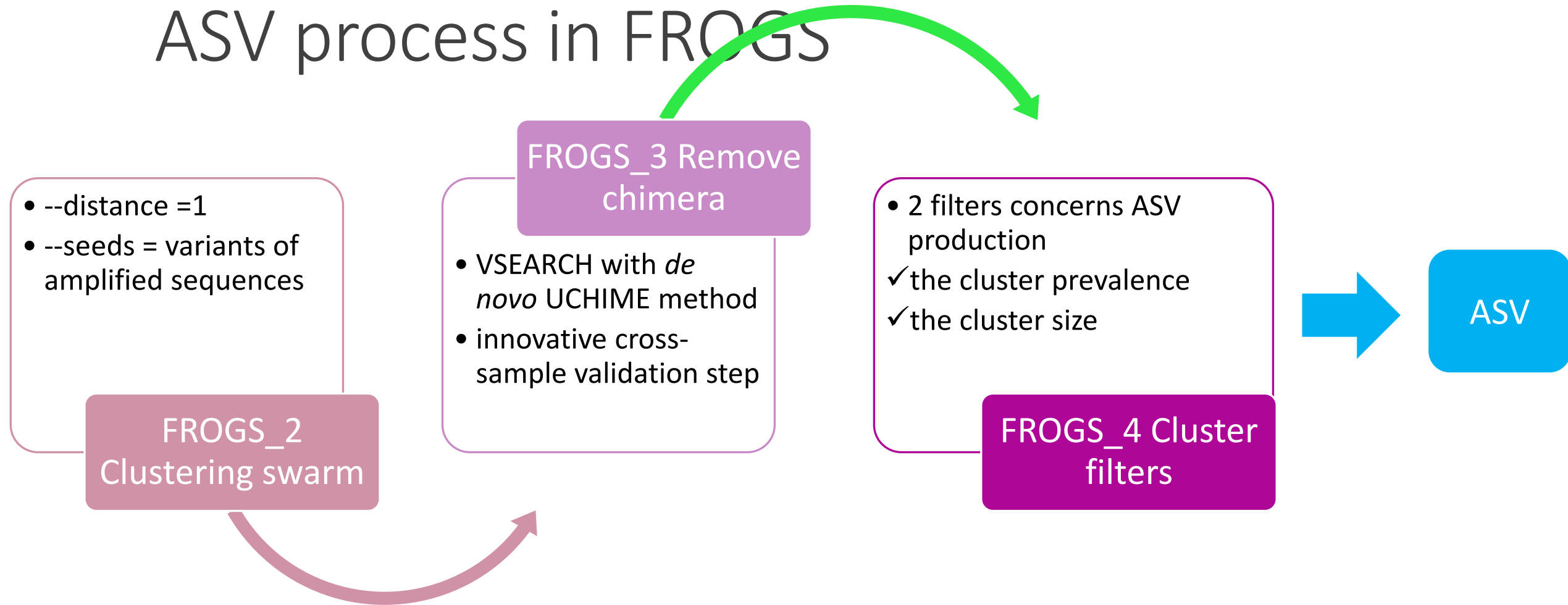


ASV by dada2 or FROGS

Denoising



# ASV process in FROGS



Swarm --seeds produces:  
variants of amplified sequences.

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

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

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

<sup>1</sup>Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia

<sup>2</sup>Department of Marine Microbiology and Biogeochemistry, NIOZ Royal Netherlands Institute for Sea Research, Texel, Netherlands

<sup>3</sup>Unit of Computational Biology, Research and Innovation Centre, Fondazione Edmund Mach, Italy

<sup>4</sup>INRAE, AgroParisTech, GABI, Université Paris-Saclay, Jouy-en-Josas, France

<sup>5</sup>INRAE, SIGENAE, Jouy-en-Josas, France

<sup>6</sup>Aquatic Ecosystem Research, University of Duisburg-Essen, Essen, Germany

<sup>7</sup>Department of Population Health and Pathobiology, College of Veterinary Medicine and Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, USA

<sup>8</sup>Center for Applied Microbiome Science, Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, Arizona, USA

<sup>9</sup>Vermont Biomedical Research Network, University of Vermont, Burlington, Vermont, USA

<sup>10</sup>Agroécologie, INRAE, Institut Agro, Univ. Bourgogne Franche-Comté, Dijon, France

<sup>11</sup>Department of Forest Mycology and Plant Pathology, Swedish University of Agricultural Sciences, Uppsala, Sweden

<sup>12</sup>Zachary Gold, NOAA Pacific Marine Environmental Laboratory, Seattle, Washington, USA

<sup>13</sup>UK Centre for Ecology & Hydrology, Oxfordshire, UK

<sup>14</sup>School of Biological Sciences, University of Reading, Reading, UK

<sup>15</sup>Department of Integrative Biology and Centre for Biodiversity Genomics, University of Guelph, Guelph, Ontario, Canada

<sup>16</sup>Gut Microbes & Health, Quadram Institute Bioscience, Norfolk, UK

<sup>17</sup>Earlham Institute, Norwich Research Park, Norfolk, UK

<sup>18</sup>Institut de Biologie Intégrative et des Systèmes, Université Laval, Québec, Québec, Canada

<sup>19</sup>Center for Forest Mycology Research, Northern Research Station, US Forest Service, Madison, Wisconsin, USA

<sup>20</sup>GenPhySE, Université de Toulouse, INRAE, ENVT, Castanet Tolosan, France

<sup>21</sup>Quantitative Biology Center (QBC), University of Tübingen, Tübingen, Germany

<sup>22</sup>Laboratory of Environmental Microbiology, Institute of Microbiology of the Czech Academy of Sciences, Praha, Czech Republic

<sup>23</sup>KU Leuven, Department of Microbiology, Immunology and Transplantation, Rega Institute for Medical Research, Laboratory of Molecular Bacteriology, Leuven, Belgium

<sup>24</sup>Department of Biological and Environmental Science, University of Jyväskylä, Jyväskylä, Finland

### Correspondence

Sten Anslan, Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia.  
Email: [sten.anslan@ut.ee](mailto:sten.anslan@ut.ee)

### Present address

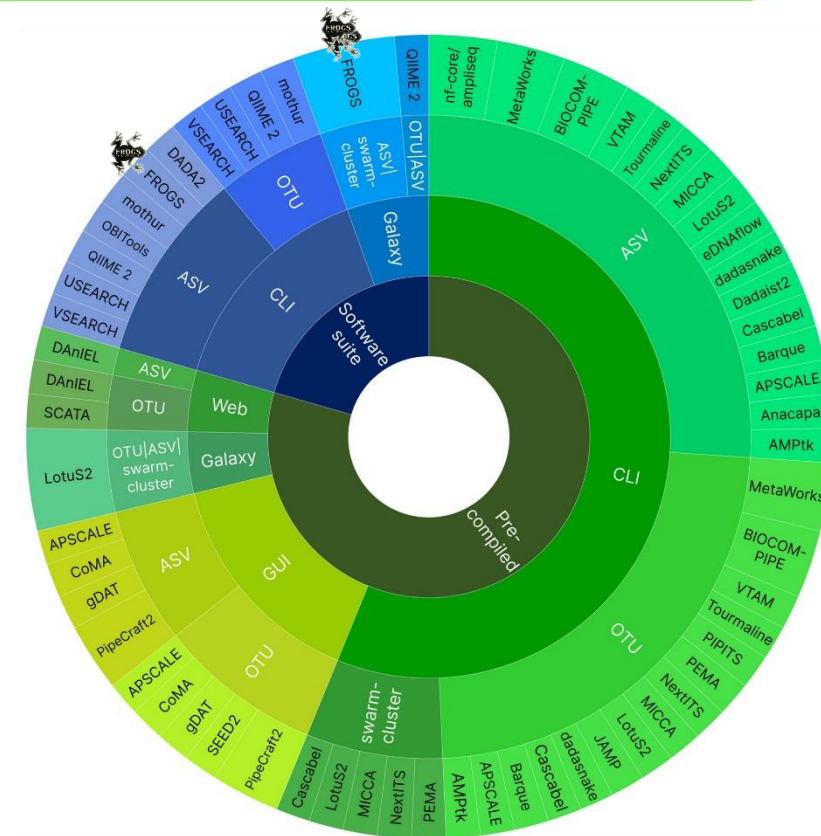
Jonathan M. Palmer, Genencor Technology Center, IFF, Palo Alto, California, USA

### Single-end data

AMPTk	Anacapa	BIOCOM-PIPE	Cascabel	CoMA	DADA2	dadasanke	
eDNAflow	FROGS	gDAT	JAMP	LotuS2	MetaWorks	MICCA	mothur
NextTITS	nf-core/ampliseq	OBITools	PipeCraft2	QIIME 2	SCATA		
SEED2	USEARCH	VSEARCH	Tourmaline	VTAM			

APSCALE  
Barque  
Dadaist2  
DAnIEL  
PIPITS  
PEMA

### Paired-end data



### Linux

#### macOS

eDNAflow	AMPTk	Barque	BIOCOM-PIPE	Anacapa	APSCALE	CoMA
dadasanke	Cascabel	Dadaist2	JAMP	DADA2	gDAT	MICCA
MetaWorks	OBITools	PIPITS	VSEARCH	mothur	nf-core/ampliseq	
NextTITS				PEMA	PipeCraft2	USEARCH
				Tourmaline	VTAM	

SCATA

SEED2

### Windows

### Web-based (including Galaxy)

---

# Affiliation tool

---

Using reference database

16S SILVA 138.1

Select reference from the list

Also perform RDP assignation? **Optional**

Yes

No

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

Taxonomic ranks

Domain Phylum Class Order Family Genus Species

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

Sequence file

15: FROGS\_4 Cluster filters: clusterFilters\_sequences.fasta

The sequences to affiliated (format: FASTA)

Abundance file

14: FROGS\_4 Cluster filters: clusterFilters\_abundance.biom



For more details on FROGS databanks:

[https://web-genobioinfo.toulouse.inrae.fr/frogs\\_databanks/assignation/readme.txt](https://web-genobioinfo.toulouse.inrae.fr/frogs_databanks/assignation/readme.txt)

# Available databases in FROGS

[https://web-genobioinfo.toulouse.inrae.fr/frogs\\_databanks/assignation/readme.txt](https://web-genobioinfo.toulouse.inrae.fr/frogs_databanks/assignation/readme.txt)

For exemples:

## ITS

ITS1 extract  
ITS UNITE Eukaryote 8.2  
ITS UNITE Fungi 8.2  
ITS UNITE 7.1  
ITS UNITE Eukaryote 8.0  
ITS UNITE Fungi 8.3

## 16S

16S SILVA Pintail100 138.1  
16S SILVA Pintail50 138.1  
16S SILVA Pintail80 138.1  
16S SILVA 138.1  
16S MIDAS S132\_3.6  
16S EZBioCloud 52018  
16S DAIRYdb V1.1.2  
16S Greengenes 13.5  
16S MIDAS S138.1\_v4.8.1  
16S DAIRYdb v2.0 20210401V2.0\_20210401  
16S REFseq Bacteria 20230726  
16S REFseq Archaea 20230726  
16S-ITS-23S GTDB 08-RS214

## coi

COI MIDORI LONGEST SP GB242  
COI MIDORI MARINE 20180221  
COI MIDORI 20180221  
COI BOLD 1percentN 22019  
COI BOLD 22019  
COI BOLD 052022  
COI MIDORI UNIQ SP GB249  
COI MIDORI LONGEST SP GB249

NCBI

complete operon

# 1 Cluster = 2 affiliations

---

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

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

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

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

Identity: 100% and Coverage: 100%

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

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

# Affiliation Strategy of FROGS

---

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

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

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

Strictly identical (V1-V3 amplification) on 499 nucleotides

Which one to choose?

# Affiliation Strategy of FROGS

---

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

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

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

Strictly identical (V1-V3 amplification) on 499 nucleotides



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

We cannot choose without preconceived ideas.

---

# Practice:

---

LAUNCH THE FROGS\_5 TAXONOMIC AFFILIATION TOOL

# Exercice:

---

Go to history « 16S » history

Launch the « FROGS\_5 taxonomic affiliation » tool with

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

→ objectives :

- understand abundance tables columns
- understand the BLAST affiliation

**FROGS\_5 Taxonomic affiliation** Taxonomic affiliation of each ASV's seed by RDPtools and BLAST (Galaxy Version 4.1.0+galaxy1)

☆ Favorite

▼ Options

#### Using reference database

16S SILVA 138.1\_pintail100 ▼

Select reference from the list

#### Also perform RDP assignment?

- Yes  
 No




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

#### Taxonomic ranks

Domain Phylum Class Order Family Genus Species


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

#### Sequence file

   15: FROGS\_4 Cluster filters: clusterFilters\_sequences.fasta ▼ 

The sequences to affiliated (format: FASTA)

#### Abundance file

   14: FROGS\_4 Cluster filters: clusterFilters\_abundance.biom ▼ 

The abundance file (format: BIOM)


# Exercise

---

1. What are the « **FROGS\_5 taxonomic affiliation tool** » output files ?
2. How many sequences are affiliated by BLAST ?
3. How many ASV have a “multiaffiliation” at Order ranks ?
4. Click on the « eye » button on the BIOM output file, what do you understand ?



# Exercise

Use the **Biom\_to\_TSV** tool on this last file and click again on the "eye"  on the new output generated.

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

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

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

- FROGS\_0 Demultiplex reads Attribute reads to samples in function
- FROGS\_1 Pre-process merging, denoising and dereplication
- FROGS\_2 Clustering swarm Single-linkage clustering on sequences
- FROGS\_Cluster\_Stat Process some metrics on clusters
- FROGS\_3 Remove chimera Remove PCR chimera in each sample
- FROGS\_4 Cluster filters Filters clusters on several criteria.
- FROGS ITSx Extract the highly variable ITS1 and ITS2 subregions from reads
- FROGS\_5 Taxonomic affiliation Taxonomic affiliation of each ASV
- FROGS\_6\_Affiliation\_Stat Process some metrics on taxonomies
- FROGS Tree Reconstruction of phylogenetic tree
- FROGS Affiliation Filters Filters ASVs on several affiliation criteria
- FROGS Affiliation postprocess Aggregates ASVs based on alignments
- FROGS Abundance normalisation Normalise ASV abundance.
- FROGSFUNC\_1\_placeseqs\_and\_copynumbers Places ASVs into a reference
- FROGSFUNC\_2\_functions Calculates functions abundances in each ASV
- FROGSFUNC\_3\_pathways Calculates pathway abundances in each ASV
- FROGS BIOM to std BIOM Converts a FROGS BIOM in fully compatible BIOM
- FROGS TSV\_to\_BIOM Converts a TSV file in a BIOM file 1
- FROGS BIOM to TSV Converts a BIOM file in TSV file**
- FROGSSTAT Phyloseq Import Data from 3 files: biomfile, samplefile, taxfile
- FROGSSTAT Phyloseq Composition Visualisation with bar plot and heatmap
- FROGSSTAT Phyloseq Alpha Diversity with richness plot
- FROGSSTAT Phyloseq Beta Diversity distance matrix
- FROGSSTAT Phyloseq Sample Clustering of samples using different methods
- FROGSSTAT Phyloseq Structure Visualisation with heatmap plot and dendrogram
- FROGSSTAT Phyloseq Multivariate Analysis Of Variance perform MDS
- FROGSSTAT DESeq2 Preprocess import a Phyloseq object and preprocess
- FROGSSTAT DESeq2 Visualisation to extract and visualise differentially expressed

# Exercise

---

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



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

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

# Exercise

---

## 6. Observe and describe

- In FROGS BIOM to TSV: abundance\_silva.tsv, the different columns of cluster 3
  - a. how would you qualify the alignment between the ASV3 seed and the sequences of the silva database?
  - b. What does it mean e-value = 0 ?
  - c. What is the header of column that shows the sequence of ASV seed ?
  - d. How many sequences have ASV3 in total ?
  - e. How many sequences have ASV3 in MVT0.LOT10 ? What is the sample where ASV3 is absent ?

# Exercise

---

## 7. Observe and describe

- In FROGS BIOM to TSV: multi\_affiliations.tsv, identifies the lines corresponding to cluster3
  - a. Why cluster3 has a multi-affiliation for species ?
  - b. Why “Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Lactobacillus;Lactobacillus sakei” is present 74 times ?

Q1: What are the « **FROGS\_5 taxonomic affiliation tool** » output files ?

Q2: How many sequences are affiliated by BLAST ?

# Exercise

Answer 1

21: FROGS\_5 Taxonomic affiliation: report.html

20: FROGS\_5 Taxonomic affiliation: affiliation\_abundance.biom

Answer 2



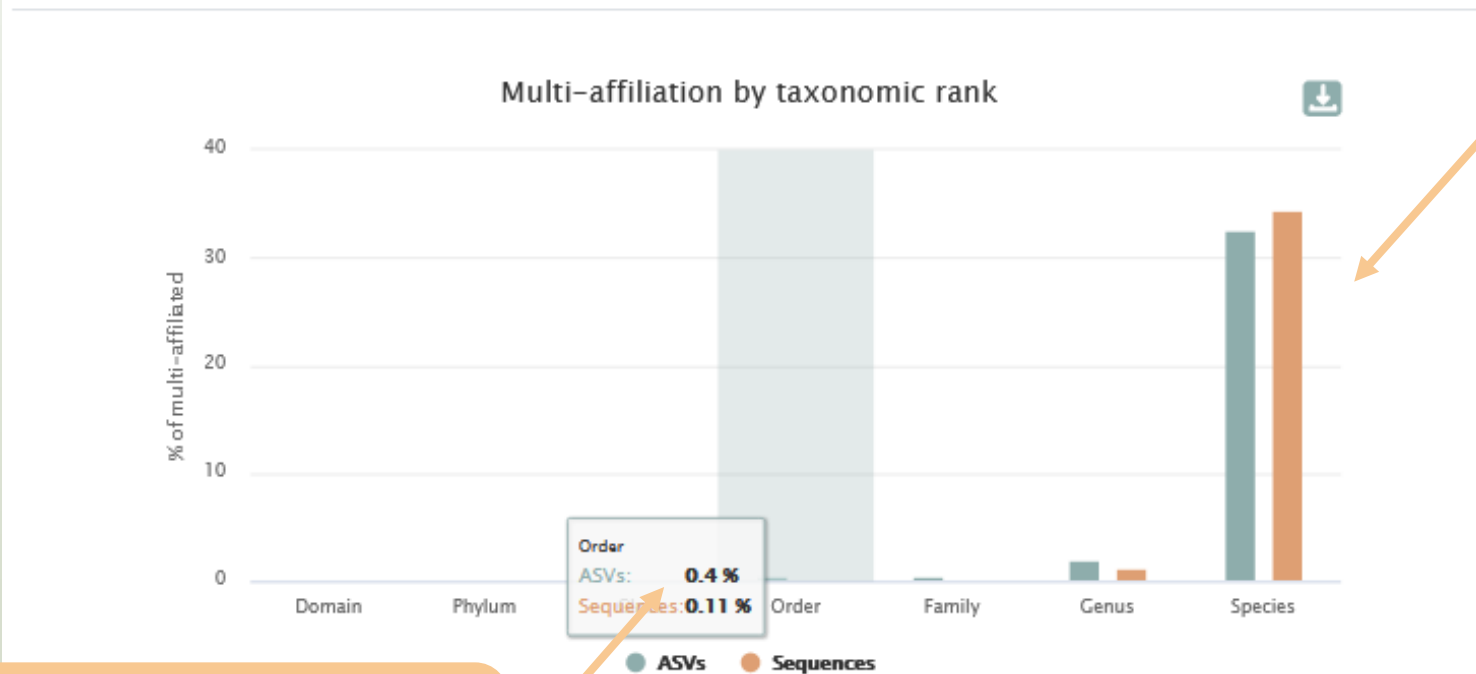
All sequences have a blast affiliation

Answer 3

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

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

### Blast multi-affiliation summary



2.83% of ASV are ambiguous until Order rank

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

```
{ "id": null, "format": "Biological Observation Matrix 1.0.0", "format_url": "http://biom-format.org", "type": "OTU table"
"2023-03-28T11:27:32", "rows": [{"id": "Cluster_1", "metadata": {"comment": [], "seed_id": "17_41", "blast_affiliations":
["Bacteria", "Firmicutes", "Bacilli", "Lactobacillales", "Listeriaceae", "Brochothrix", "unknown species"], "value": "0.
"perc_query_coverage": 100.0}, {"subject": "CP023643.1319711.1321267", "taxonomy": ["Bacteria", "Firmicutes", "Bacilli",
"Brochothrix", "Brochothrix thermosphacta"], "value": "0.0", "aln_length": 497, "perc_identity": 100.0, "perc_query_cove
"CP023483.1387851.1389407", "taxonomy": ["Bacteria", "Firmicutes", "Bacilli", "Lactobacillales", "Listeriaceae", "Brochot
"0.0", "aln_length": 497, "perc_identity": 100.0, "perc_query_coverage": 100.0}, {"subject": "CP023643.1330505.1332061",
"Bacilli", "Lactobacillales", "Listeriaceae", "Brochothrix", "Brochothrix thermosphacta"], "value": "0.0", "aln_length":
"perc_query_coverage": 100.0}, {"subject": "CP023483.1398643.1400199", "taxonomy": ["Bacteria", "Firmicutes", "Bacilli",
"Brochothrix", "Brochothrix thermosphacta"], "value": "0.0", "aln_length": 497, "perc_identity": 100.0, "perc_query_cove
"CP023643.1325108.1326664", "taxonomy": ["Bacteria", "Firmicutes", "Bacilli", "Lactobacillales", "Listeriaceae", "Brochot
"0.0", "aln_length": 497, "perc_identity": 100.0, "perc_query_coverage": 100.0}, {"subject": "CP023643.1248577.1250133",
"Bacilli", "Lactobacillales", "Listeriaceae", "Brochothrix", "Brochothrix thermosphacta"], "value": "0.0", "aln_length":
"perc_query_coverage": 100.0}, {"subject": "CP023483.1393248.1394804", "taxonomy": ["Bacteria", "Firmicutes", "Bacilli",
"Brochothrix", "Brochothrix thermosphacta"], "value": "0.0", "aln_length": 497, "perc_identity": 100.0, "perc_query_cove
"CP023483.1316717.1318273", "taxonomy": ["Bacteria", "Firmicutes", "Bacilli", "Lactobacillales", "Listeriaceae", "Brochot
"0.0", "aln_length": 497, "perc_identity": 100.0, "perc_query_coverage": 100.0}, {"subject": "CP023643.722570.724126", "t
```

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

Answer 5

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

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

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

#comment	blast_taxonomy	blast_subject	blast_perc_identity	blast_perc_query_coverage
no data	Bacteria;Firmicutes;Bacilli;Lactobacillales;Listeriaceae;Brochothrix;Brochothrix thermosphacta	multi-subject	100	100
no data	Bacteria;Proteobacteria;Gammaproteobacteria;Enterobacterales;Vibrionaceae;Photobacterium;unknown species	FJ456662.1.1555	100	100
no data	Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Latilactobacillus;Multi-affiliation	multi-subject	100	100
no data	Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Multi-affiliation	multi-subject	100	100
no data	Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Leuconostoc;Multi-affiliation	multi-subject	100	100
no data	Bacteria;Firmicutes;Bacilli;Lactobacillales;Streptococcaceae;Lactococcus;Lactococcus piscium	AM943029.1.1242	99.799	100
no data	Bacteria;Firmicutes;Bacilli;Erysipelotrichales;Erysipelotrichaceae;ZOR0006;unknown species	HG792212.1.1536	94.203	100
no data	Bacteria;Firmicutes;Bacilli;Lactobacillales;Streptococcaceae;Lactococcus;Multi-affiliation	multi-subject	100	100
no data	Bacteria;Firmicutes;Bacilli;Lactobacillales;Lactobacillaceae;Weissella;Weissella ceti	FN813251.1.1761	99.799	100

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

## Answer 6

- a. how would you qualify the alignment between the ASV3 (cluster\_3) seed and the sequences of the silva database?

Alignment is perfect ! 100% identity and 100% coverage between ASV3 (cluster 3) seed and the 520 nucleotides of sequence from silva database

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

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

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

Seed\_sequence

- d. How many sequences have ASV3 (cluster\_3) in total ?

40711 found in column " observation\_sum"

- e. How many sequences have ASV3 (cluster\_3) in MVT0.LOT10 ? What is the sample where ASV3 (cluster\_3) is absent ?

MVT0.LOT10
4
0
6722
13
20

CDT0.LOT02
64
1
0
0
3

We can remark that ASV3 is particularly present in MV samples and rare in CD samples

## Answer 7

- a. Why ASV3 (cluster\_3) has a multiaffiliation for species ?

In multi-affiliations.tsv file, for cluster\_3, we observe that 75 affiliations are possible for this ASV at species rank.

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

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

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

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

# Silva pintail or not pintail ?

---

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

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

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



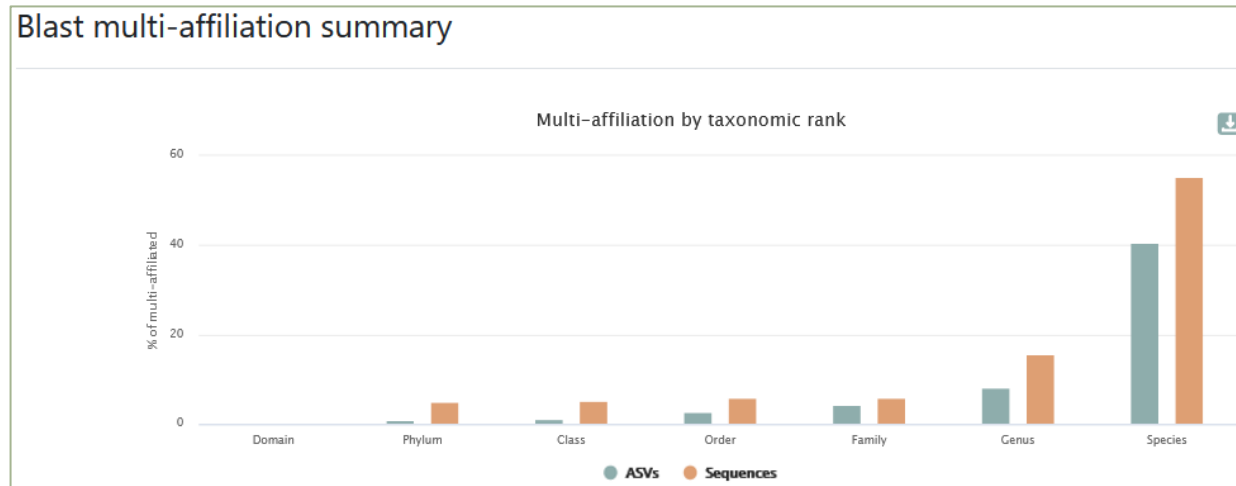
Only for 16S !

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

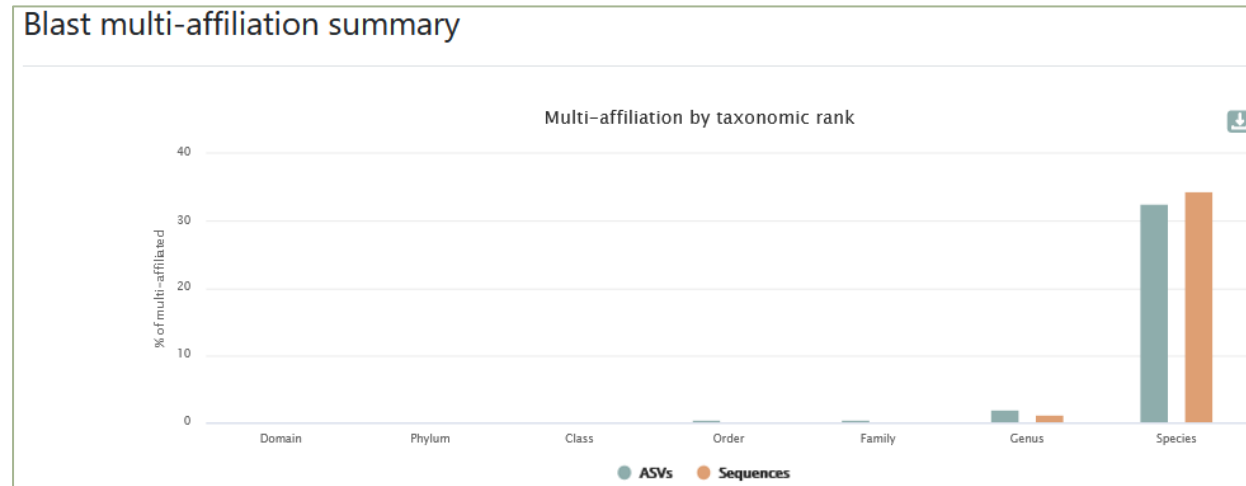
# Silva pintail or not pintail ?

---

Blast multi-affiliation summary



Blast multi-affiliation summary



# Exemple between silva 138.1 and silva 138.1 pintail 100

---

130 identical blast best hits on SILVA 138.1 pintail 100 databank

Cluster\_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes  
Cluster\_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes 6609  
Cluster\_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes C1  
Cluster\_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes KPA171202  
Cluster\_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes TypeIA2 Pacn17  
Cluster\_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes TypeIA2 Pacn31  
Cluster\_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes TypeIA2 Pacn33

# Exemple between silva 138.1 and silva 138.1 pintail 100

267 identical blast best hits on SILVA 138.1 full databank

- ? Cluster\_4 Bacteria;Actinobacteriota;Actinobacteria;Corynebacteriales;Corynebacteriaceae;Corynebacterium;unknown species
- ? Cluster\_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Aureobasidium melanogenum
- Cluster\_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes
- Cluster\_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes 266
- Cluster\_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes 6609
- Cluster\_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes C1
- Cluster\_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes hdn-1
- Cluster\_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes HL096PA1
- Cluster\_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes KPA171202
- Cluster\_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes SK137
- Cluster\_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;unknown species
- Cluster\_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes TypeA2 P.acn17
- Cluster\_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes TypeA2 P.acn31
- Cluster\_4 Bacteria;Actinobacteriota;Actinobacteria;Propionibacteriales;Propionibacteriaceae;Cutibacterium;Cutibacterium acnes TypeA2 P.acn33
- ? Cluster\_4 Bacteria;Firmicutes;Bacilli;Lactobacillales;Carnobacteriaceae;Dolosigranulum;unknown species

Induces a multi-affiliation up to phylum rank

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

# How choose the good affiliation ?

---

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

2 choices for cluster 64

# How choose the good affiliation ?

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

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

Ex:

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

and

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

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

# Affiliation explorer

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

The screenshot shows the Affiliation Explorer web application. On the left, there are three upload sections: 'Upload Biom File' (Galaxy37-[f]), 'Optional: upload Fasta File' (Galaxy32-[f]), and 'Upload MultiHits TSV File' (Galaxy42-[f]). Each has a 'Browse...' button and an 'Upload complete' button. A 'Download' button is at the bottom left. The main area has two tabs: 'Affiliation selection' and 'Affiliation edition'. Under 'Affiliation selection', there is a 'Select OTU' dropdown menu set to 'Cluster\_3', and 'Update OTU' and 'Skip OTU' buttons. Below this, a message states: 'Cluster\_3 - 2 conflicting affiliations, ambiguity at rank Species'. Instructions follow: 'Select new affiliation by clicking on a row (double click on a cell to edit its content). Click "Update OTU" to update affiliation (with selected row) or "Skip OTU" to move to the next one.' There is a 'Show 10 entries' dropdown and a search box. A table displays two entries with columns: Kingdom, Phylum, Class, Order, Family, Genus, Species, Blast ID, %id, and %cov. The first entry is Lactobacillus sakei and the second is unknown species. At the bottom, it says 'Showing 1 to 2 of 2 entries' and has 'Previous', '1', and 'Next' navigation buttons. A 'Show sequence' checkbox is at the bottom left.

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

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

# Affiliation explorer

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

Demo  
video

The screenshot shows a web browser window displaying the 'Affiliation explorer' application. The browser's address bar shows the URL: <https://hub.gke2.mybinder.org/user/mahendra-mariad-liationexplorer-4jqjb7jw/rstudio/?token=r0mZweROqcCzicA5hQm8IA&view=shiny>. The application interface has a dark blue header with the title 'Affiliation explorer' and a hamburger menu icon. On the left side, there is a dark sidebar with three sections: 'Upload Biom File', 'Optional: upload Fasta File', and 'Upload MultiHits TSV File'. Each section contains a 'Browse...' button and a 'No file sele...' button. The main content area has a light blue background and features two tabs: 'Affiliation selection' (active) and 'Affiliation edition'. Below the tabs, a message reads: 'Please upload your data (Biom file and MultiHits TSV file)'. The browser window also shows standard navigation and window control buttons.

---

# 6- Affiliation Stat

---

FROGS\_6\_Affiliation\_Stat Process some metrics on taxonomies (Galaxy Version 4.1.0+galaxy1)

☆ Favorite

▼ Options

**Abundance file**

20: FROGS\_5 Taxonomic affiliation: affiliation\_abundance.biom

Abundances and affiliations (format: BIOM)

**Taxonomic ranks**

Domain Phylum Class Order Family Genus Species

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

**Rarefaction ranks**

Class Order Family Genus Species

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

**Affiliation processed**

FROGS Blast

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

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

---

# Practice:

---

LAUNCH THE FROGS\_6 AFFILIATION STAT TOOL

# Exercice:

---

Go to history « 16S » history

Launch the « FROGS\_6 Affiliation Stat » tool on last affiliation\_abundance.biom

→ objectives :

understand rarefaction curves and the diversity diagram

# Exercise:

---

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

Answer 1

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

<input type="checkbox"/>	Samples	Nb domain	Nb phylum	Nb class	Nb order	Nb family	Nb genus	Nb species	Nb sequences
<input checked="" type="checkbox"/>	SFT0.LOT06	1	4	5	9	14	35	57	8,821
<input checked="" type="checkbox"/>	SFT0.LOT01	1	4	6	13	27	39	63	8,859
<input checked="" type="checkbox"/>	FCT0.LOT01	1	5	6	13	24	41	96	8,504
<input checked="" type="checkbox"/>	SFT0.LOT05	1	5	7	18	32	50	95	8,728
<input checked="" type="checkbox"/>	SFT0.LOT08	1	4	6	13	33	53	77	8,788
<input checked="" type="checkbox"/>	BHT0.LOT01	1	7	9	20	35	55	83	8,750
<input checked="" type="checkbox"/>	SFT0.LOT04	1	6	8	17	34	55	83	8,750
<input checked="" type="checkbox"/>	SFT0.LOT03	1	5	8	17	34	55	83	8,750
<input checked="" type="checkbox"/>	SFT0.LOT02	1	6	7	17	34	55	83	8,750
<input type="checkbox"/>	MVT0.LOT10	1	4	5	17	31	57	83	9,143
<input type="checkbox"/>	CDT0.LOT02	1	6	8	22	36	58	85	8,750

1. Sort the table by genus number

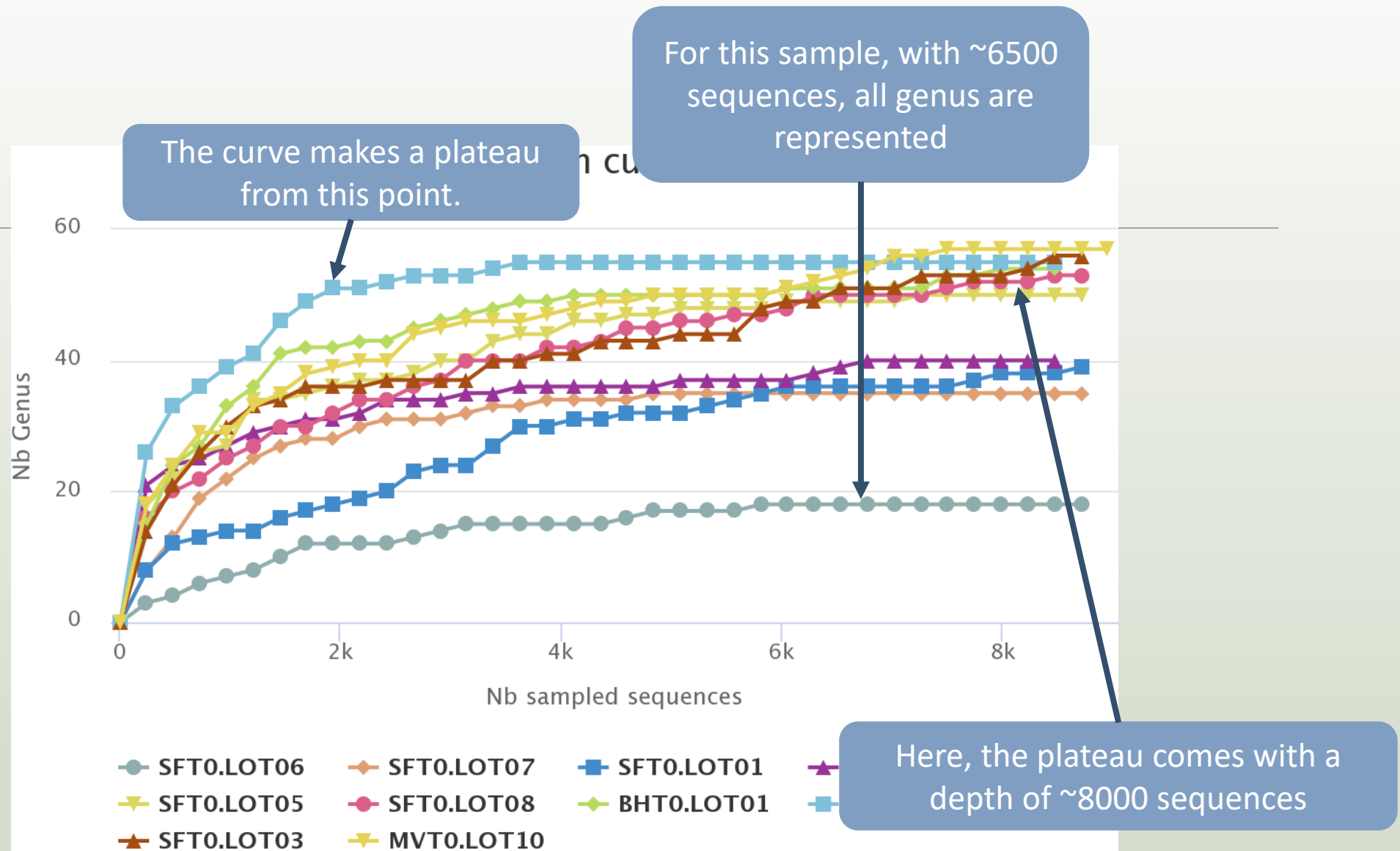
2. Select the 10 first samples

3. At the bottom of the table click on

With selection: Genus

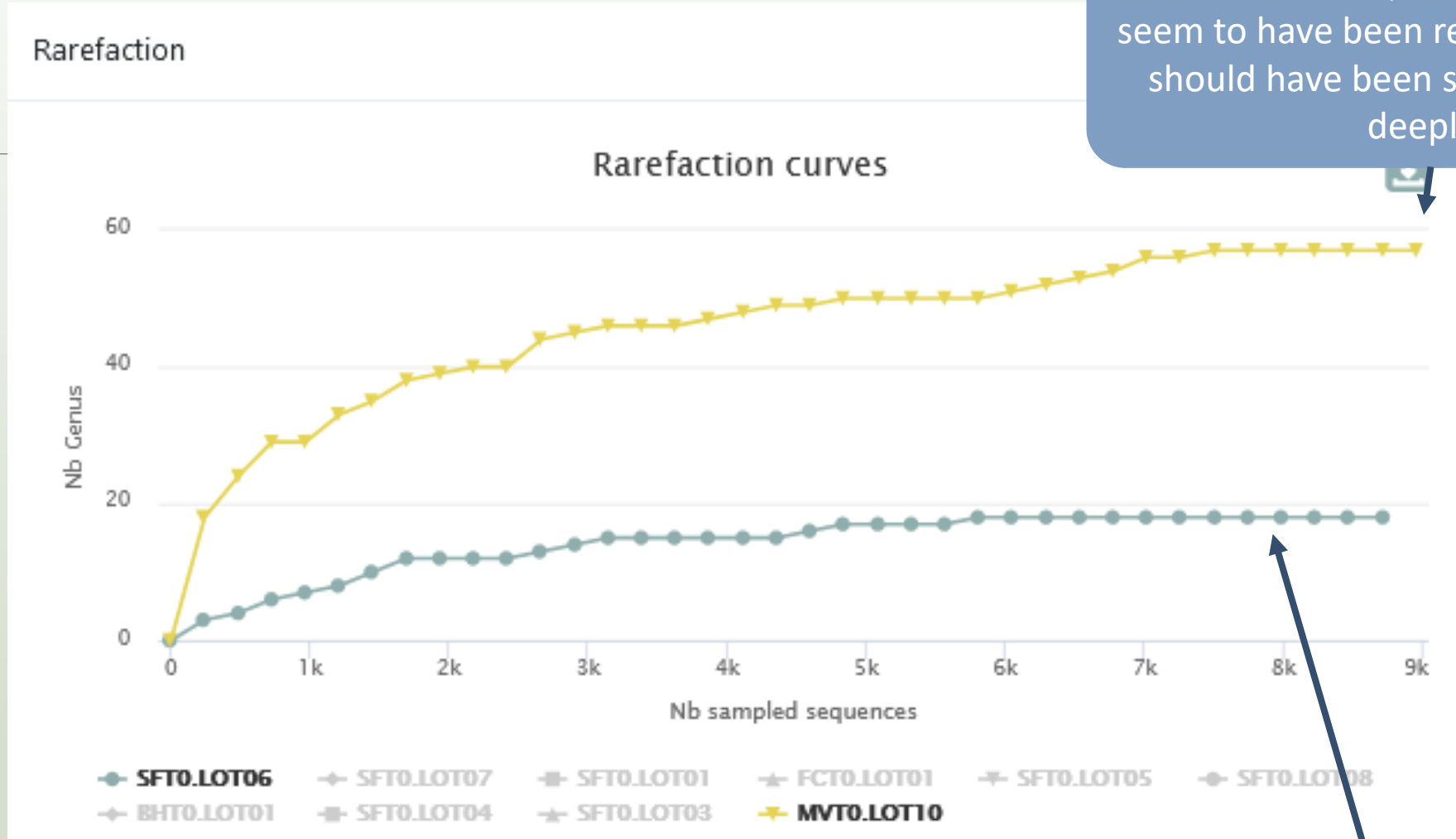
Answer 2

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



Answer 2

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



For MVT0.LOT10, the plateau does not seem to have been reached. Perhaps it should have been sequenced more deeply?

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

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

Use search to find only FC samples

CSV

Show

Select the 8 samples of FC

Search

FC

<input checked="" type="checkbox"/>	Samples	Nb domain	Nb phylum	Nb class	Nb order	Nb family	Nb genus	Nb species	Nb sequences
<input checked="" type="checkbox"/>	FCT0.LOT01	1	5	6	13	24	41	96	8,504
<input checked="" type="checkbox"/>	FCT0.LOT02	1	6	8	23	40	67	126	7,638
<input checked="" type="checkbox"/>	FCT0.LOT03	1	8	10	26	45	71	122	8,608
<input checked="" type="checkbox"/>	FCT0.LOT05	1	8	10	25	44	78	139	8,577
<input checked="" type="checkbox"/>	FCT0.LOT06	1	8	10	29	53	97	151	8,577
<input checked="" type="checkbox"/>	FCT0.LOT07	1	5	7	24	46	80	126	8,577
<input checked="" type="checkbox"/>	FCT0.LOT08	1	7	9	27	50	97	151	8,577
<input checked="" type="checkbox"/>	FCT0.LOT10	1	7	9	27	50	97	151	8,577

At the bottom of the table click on

With selection:

Genus

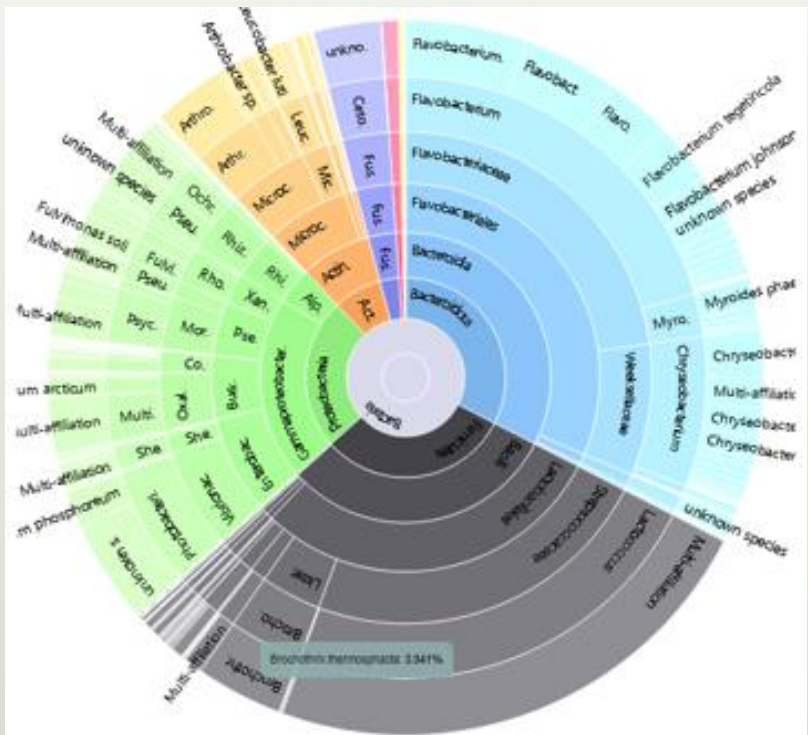
Display rarefaction

Display distribution



Answer 3,  
4, 5 & 6

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



Detail on selected:

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

Brochothrix thermosphacta nb children: 0

A table appears

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

Brochothrix thermosphacta nb children: 0

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

Answer 7

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

Taxonomy distribution Alignment distribution

At the top of the page, click on [Display global distribution](#)

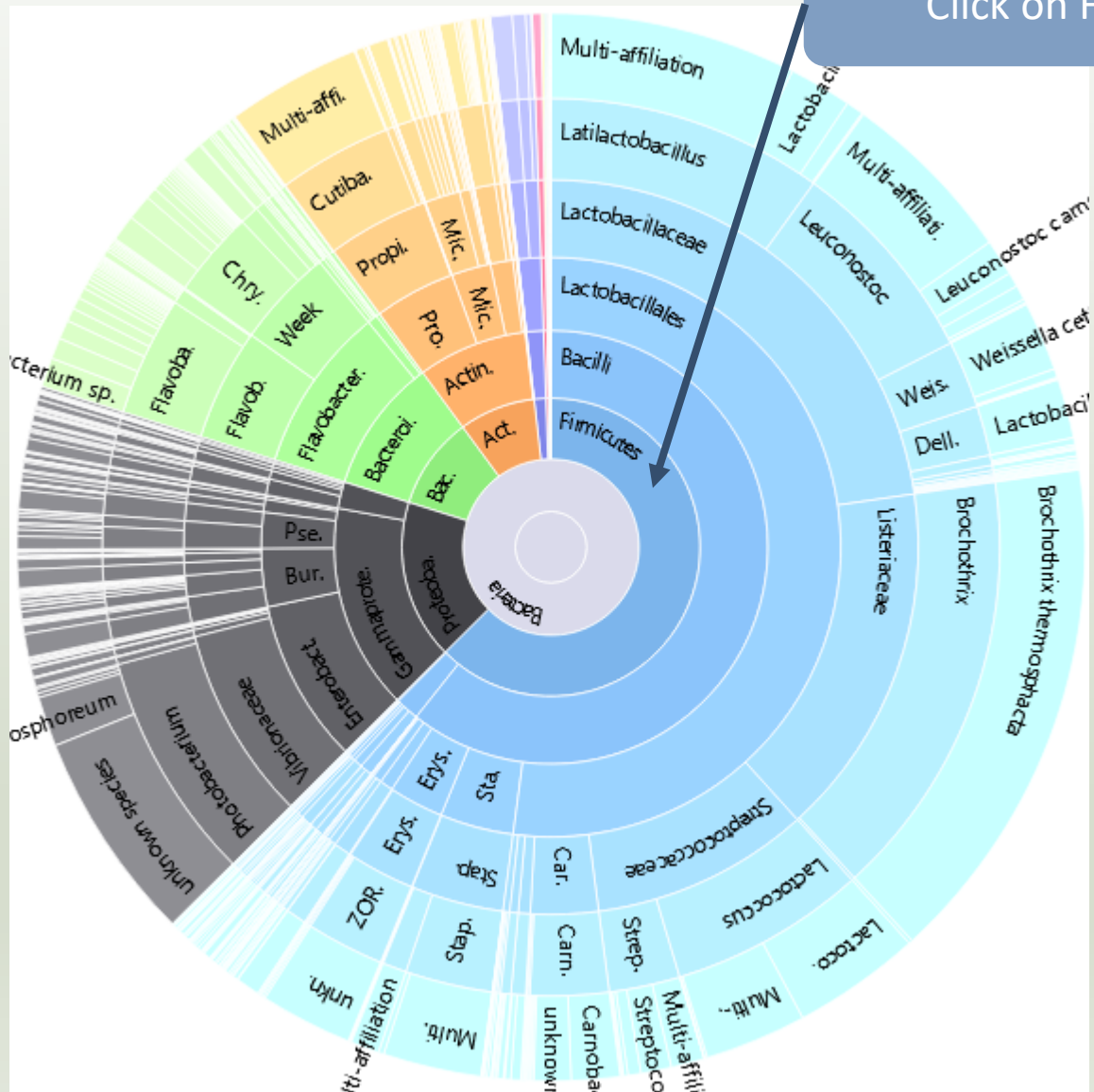
Show  entries  Search:

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

With selection:

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

Click on Firmicutes

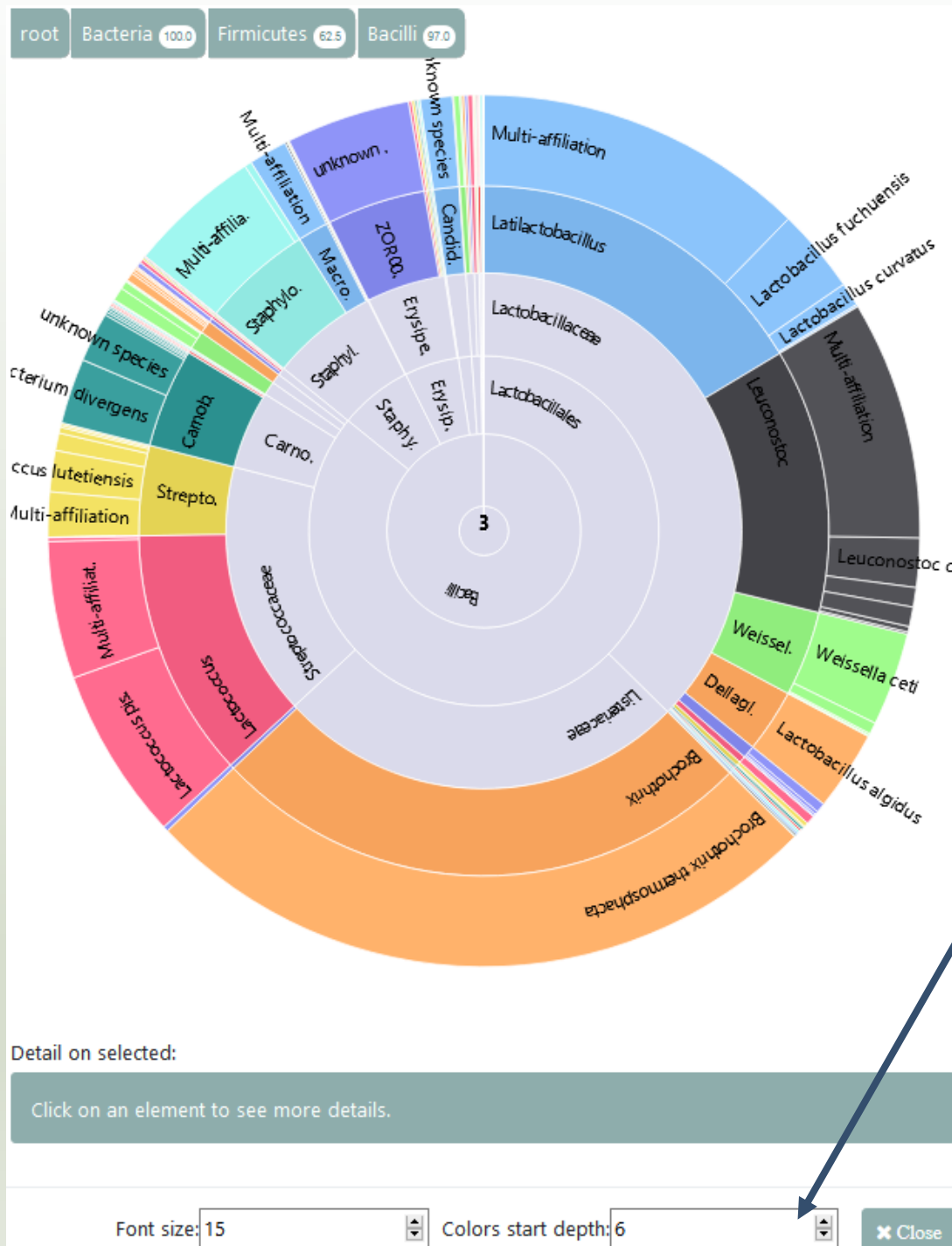


Name	Size	Global %	Parent %
root	547520		
Bacteria	547520	100.000	100.000
Firmicutes	342411	62.539	62.539

Firmicutes represent 62% of Bacteria

Answer 7

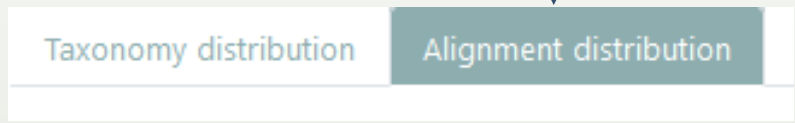
Q7: But what is the proportion of Firmicutes in the total of sequence of all sample ?



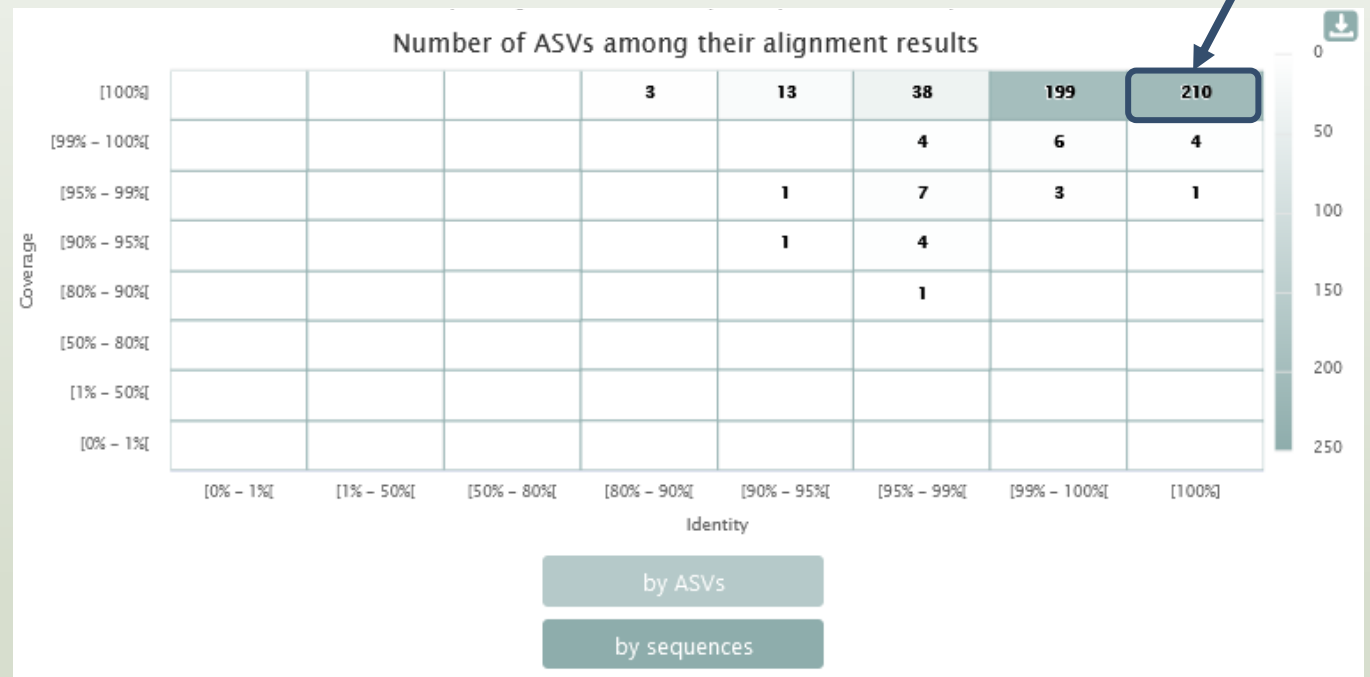
To focus on Firmicutes, double click on. After you can apply color among rank depth.

Q8: How many ASVs align perfectly with a database sequence ?

At the top of the page, click on this tab



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



---

# 7- Filters on affiliations

---

Sequence file

📄 📄 📄 15: FROGS\_4 Cluster filters: clusterFilters\_sequences.fasta 📄

The sequence file to filter (format: FASTA)

Abundance file

📄 📄 📄 25: FROGS\_5 Taxonomic affiliation: affiliation\_abundance.biom 📄

The abundance file to filter (format: BIOM)

Taxonomic ranks

Domain Phylum Class Order Family Genus Species

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

Filtering mode

Hidding mode  
 Deleting mode

Do you want to delete ASV or hide affiliations?

Filter on Blast affiliations

**Maximum e-value**

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

**Minimum identity**

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

**Minimum coverage**

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

**Minimum alignment length**

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

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

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

**Sequence file**

📄 📄 📁 15: FROGS\_4 Cluster filters: clusterFilters\_sequences.fasta 📄

The sequence file to filter (format: FASTA)

**Abundance file**

📄 📄 📁 25: FROGS\_5 Taxonomic affiliation: affiliation\_abundance.biom 📄

The abundance file to filter (format: BIOM)

**Taxonomic ranks**

Domain Phylum Class Order Family Genus Species

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

**Filtering mode**

Hidding mode  
 Deleting mode

Do you want to delete ASV or hide affiliations?

Filter on Blast affiliations 👁

**Maximum e-value**

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

**Minimum identity**

99

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

**Minimum coverage**

99

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

**Minimum alignment length**

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

Possibility to filter affiliations according to blast metrics

**Keyword filters of blast affiliation**

No filter  
 Ignore taxa  
 Keep taxa

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

**Remove blast affiliations including these taxon / word**

1: Remove blast affiliations including these taxon / word

Full or partial taxon name

unknown species

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

2: Remove blast affiliations including these taxon / word

Full or partial taxon name

Firmicutes

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

+ Insert Remove blast affiliations including these taxon / word

---

Filter on RDP affiliations 🔇

Possibility to filter for keeping or for ignore ASV according keywords

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

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

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

Possibility to filter on RDP taxonomic affiliation

Not open by default

---

# Practice:

---

LAUNCH THE FROGS AFFILIATION FILTER TOOL

# Exercise:

---

## 1. Mask

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

## 2. Explore the report.html

- How many ASVs remain?
- How are impacted affiliation?

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

**Sequence file**

111: FROGS\_4 Cluster filters: clusterFilters\_sequences.fasta

The sequence file to filter (format: FASTA)

**Abundance file**

115: FROGS\_5 Taxonomic affiliation: affiliation\_abundance.biom

The abundance file to filter (format: BIOM)

**Taxonomic ranks**

Domain Phylum Class Order Family Genus Species

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

**Filtering mode**

Hidding mode  Deleting mode

Do you want to delete ASV or hide affiliations?

**Filter on Blast affiliations**

**Maximum e-value**

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

**Minimum identity**

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

**Minimum coverage**

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

**Minimum alignment length**

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

**Keyword filters of blast affiliation**

No filter  Ignore taxa  Keep taxa

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

**Remove blast affiliations including these taxon / word**

1: Remove blast affiliations including these taxon / word

**Full or partial taxon name**

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

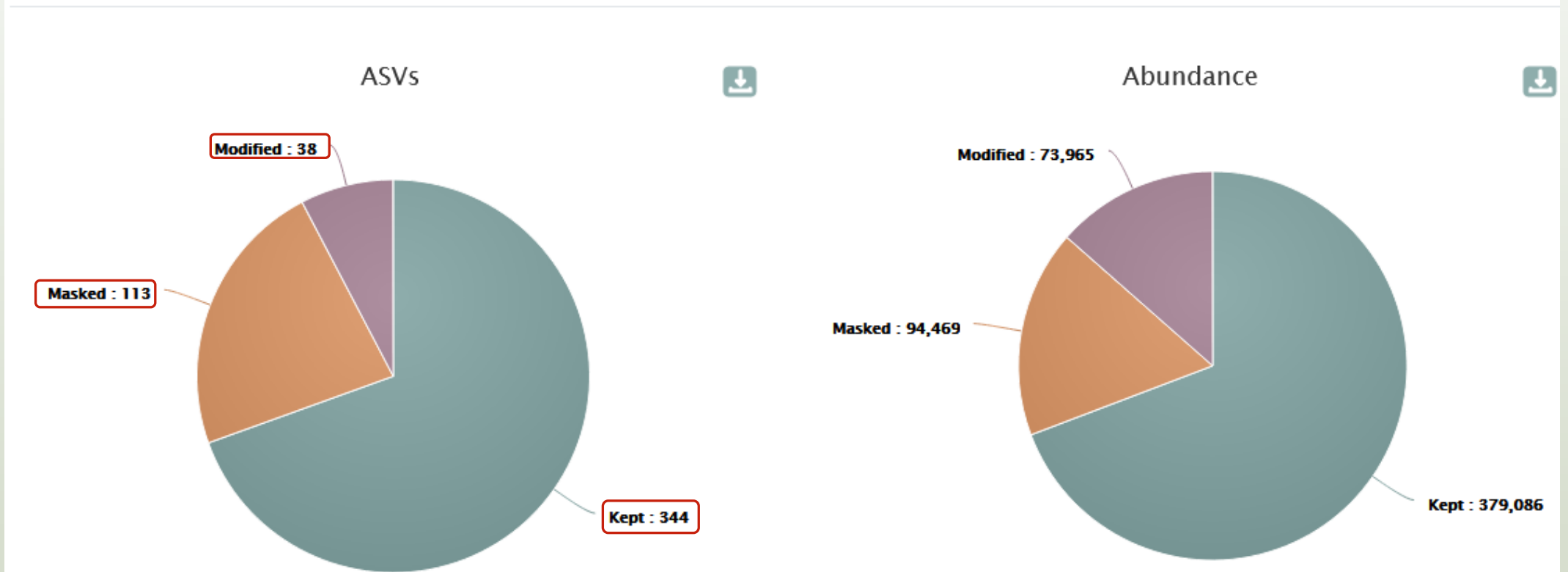
**Filter on RDP affiliations**

**Email notification**

No

Send an email notification when the job completes.

## Filters summary



- 344 ASV are kept without modification
- 38 ASV are kept with modification (see **impacted\_clusters.multi-affiliation.tsv**)
- It's remain 382 ASVs !

## 42: FROGS Affiliation Filters: impacted\_clusters.multi-affiliations.tsv

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

Exemple: Cluster\_3 is an impacted clusters because

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



To see the content, think to transform the BIOM to TSV file with **BIOM\_to\_TSV** tool

## 41: FROGS Affiliation Filters: impacted\_clusters.tsv

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

In impacted\_cluster.tsv

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

## Hidding mode

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

## Deleting mode

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

### Remark

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

---

# Abundance normalization

---

# Abundance normalization

---

Conserve a predefined number of sequence per sample:

- update Biom abundance file
- update seed fasta file

May be used when :

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

# Exercise 8

---



Which values are interesting to test?

# Exercise 8

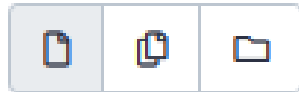
---

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

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

### FROGS Abundance normalisation Normalise ASV abundance. (Galaxy Version 4.1.0+ galaxy1)

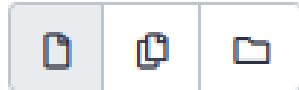
#### Sequence file



15: FROGS\_4 Cluster filters: clusterFilters\_sequences.fasta

Sequence file to normalise (format: fasta). (--input-fasta)

#### Abundance file



19: FROGS\_5 Taxonomic affiliation: affiliation\_abundance.biom

Abundance file to normalise (format: BIOM). (--input-biom)

#### Sampling method

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

Sampling by the number of sequences of the smallest sample, or select a number manually (--sampling-by-min)

## Answer 1

## The smallest sequenced samples

Clusters distribution Sequences distribution **Samples distribution**

### Sequences count

Show  entries  [Download CSV](#)

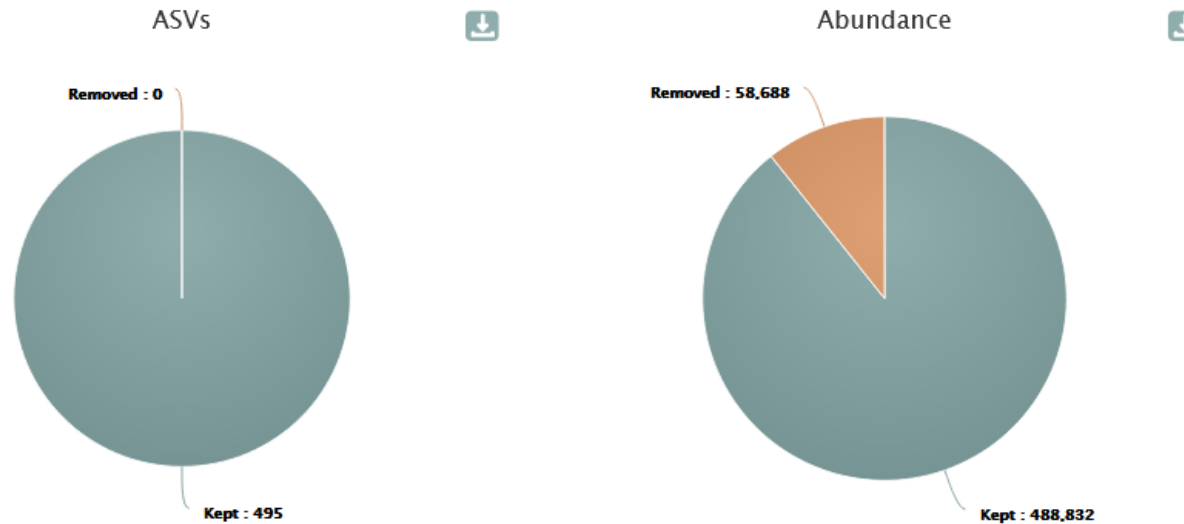
Sample	Total clusters	Shared clusters	Own clusters	Total sequences	Shared sequences	Own sequences
FCT0.LOT02	162	162	0	7,638	7,638	0
FST0.LOT03	152	152	0	7,778	7,778	0
FST0.LOT05	158	158	0	7,908	7,908	0
FST0.LOT02	149	149	0	7,956	7,956	0
CDT0.LOT06	253	253	0	8,257	8,257	0
DLT0.LOT10	222	222	0	8,331	8,331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

Showing 1 to 10 of 64 entries [Previous](#) [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [Next](#)

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

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

### Normalisation summary



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

**495 ASVs**  
**488832 sequences**

### Normalisation summary per samples

Show  entries

Sample	Nb ASV before normalisation	Nb ASV after normalisation
BHT0.LOT01	98	96
BHT0.LOT03	135	135
BHT0.LOT04	150	148

The minimum impact of ASV number per sample

## Q2: Normalize your data on 2000 sequences or less

**FROGS Abundance normalisation** Normalise ASV abundance. (Galaxy Version 4.1.0+galaxy1) ☆ Favorite ▼ Options

**Sequence file**

Sequence file to normalise (format: fasta). (--input-fasta)

**Abundance file**

Abundance file to normalise (format: BIOM). (--input-biom)

**Sampling method**

Sampling by the number of sequences of the smallest sample

Select a number of sequences

Sampling by the number of sequences of the smallest sample, or select a number manually (--sampling-by-min)

**Number of reads**

The final number of reads per sample. (--num-reads)

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

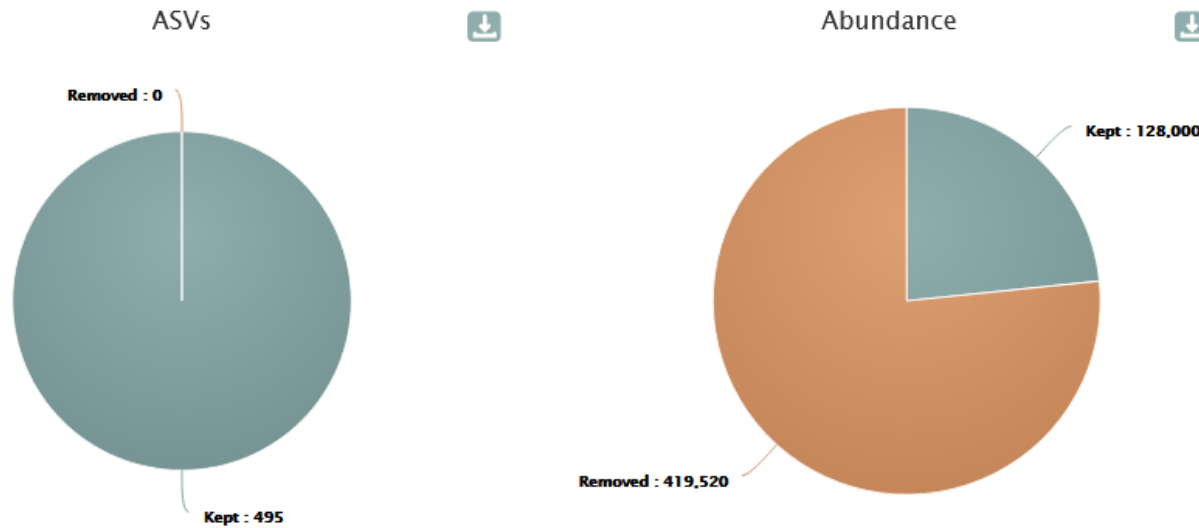
No, subsampling threshold need to at most equal to the smallest sample

Yes, subsampling threshold may be greater than the smallest sample

(--delete-samples)

If “yes”, the smallest samples can be removed !

### Normalisation summary



Normalization at **2000** sequences

**495 ASVs**  
**128000 sequences**

### Normalisation summary per samples

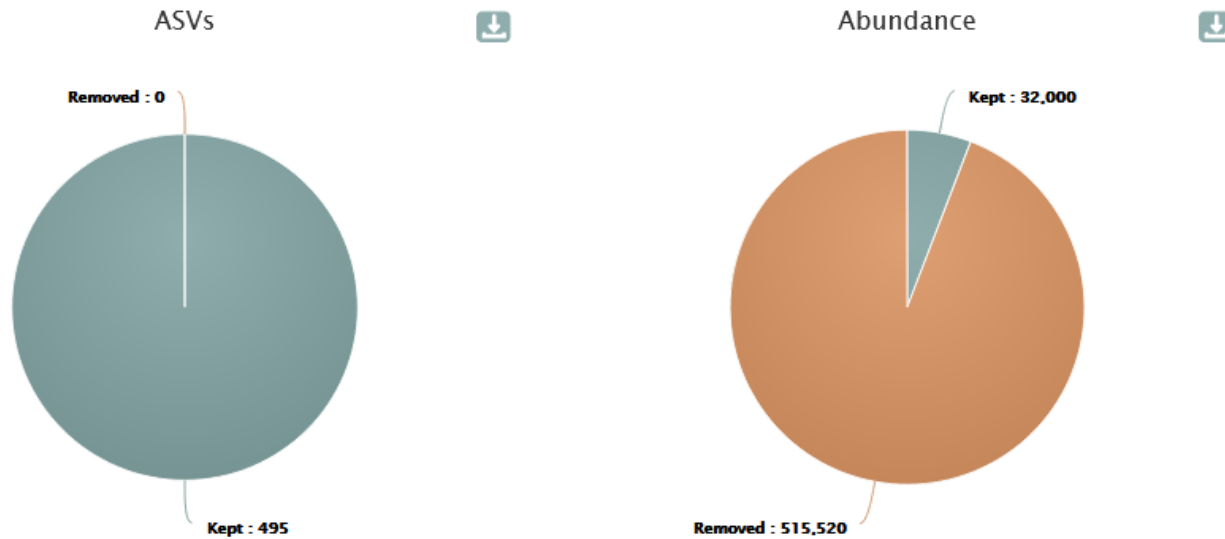
Show 10 entries

Search:

Sample	Nb ASV before normalisation	Nb ASV after normalisation
BHT0.LOT01	98	73
BHT0.LOT03	135	102
BHT0.LOT04	150	105

Big impact of ASV number per sample

## Normalisation summary



Normalization at **500**  
sequences

**495 ASVs**  
**32000 sequences**

## Normalisation summary per samples

Show  entries

Search:

Sample	Nb ASV before normalisation	Nb ASV after normalisation
BHT0.LOT01	98	41
BHT0.LOT03	135	58
BHT0.LOT04	150	57

Very big impact of ASV  
number per sample

**FROGS Abundance normalisation** Normalise ASV abundance. (Galaxy Version 4.1.0+galaxy1) ☆ Favorite ▼ Options

**Sequence file**

15: FROGS\_4 Cluster filters: clusterFilters\_sequences.fasta

Sequence file to normalise (format: fasta). (--input-fasta)

**Abundance file**

19: FROGS\_5 Taxonomic affiliation: affiliation\_abundance.biom

Abundance file to normalise (format: BIOM). (--input-biom)

**Sampling method**

Sampling by the number of sequences of the smallest sample

Select a number of sequences

Sampling by the number of sequences of the smallest sample, or select a number manually (--sampling-by-min)

**Number of reads**

8000

The final number of reads per sample. (--num-reads)

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

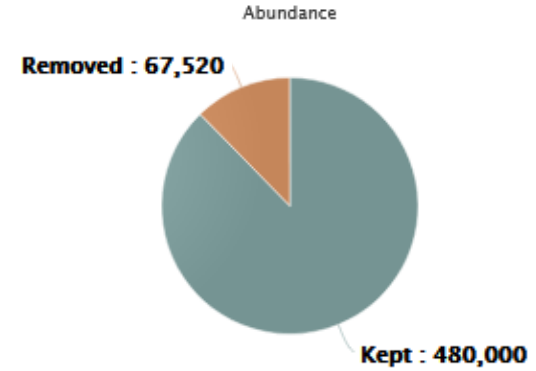
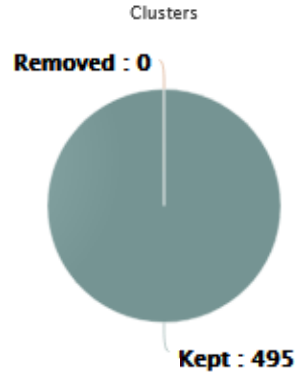
No, subsampling threshold need to at most equal to the smallest sample

Yes, subsampling threshold may be greater than the smallest sample

(--delete-samples)

Answer 3

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



Deleted samples (nb sequences < 8000)

Show 10 entries

Sample	T1	Nb sequences
FCT0.LOT02		7,638
FST0.LOT02		7,956
FST0.LOT03		7,778
FST0.LOT05		7,908

Showing 1 to 4 of 4 entries

Search:

CSV

Previous 1 Next

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

Normalisation summary per samples

Show 10 entries

Sample	T1	Nb OTU before normalisation	T1	Nb OTU after normalisation	T1
BHT0.LOT01		98		96	
BHT0.LOT03		135		134	
BHT0.LOT04		150		149	

Search:

Very very big impact !

---

# FROGS Tree

---

CREATE A PHYLOGENETICS TREE OF ASVS

# FROGS Tree

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

FROGS Tree Reconstruction of phylogenetic tree (Galaxy Version 4.1.0+galaxy1) ☆ Favorite 🔄 Versions ▼ Options

**Sequence file**

15: FROGS\_4 Cluster filters: clusterFilters\_sequences.fasta

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

**Biom file**

25: FROGS\_5 Taxonomic affiliation: affiliation\_abundance.biom

The abundance file (format: BIOM)

**Email notification**

No

Send an email notification when the job completes.

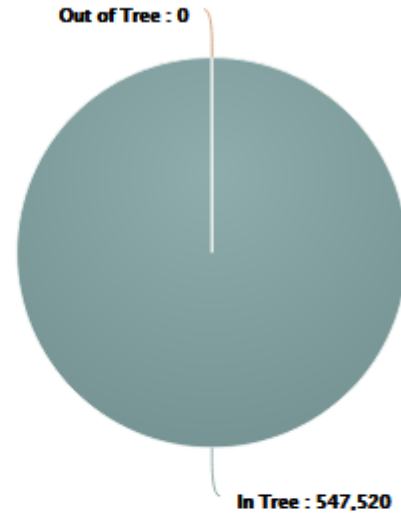
2 outputs:

- FROGS Tree: report.html**
- FROGS Tree: tree.nwk**

ASVs



Abundance

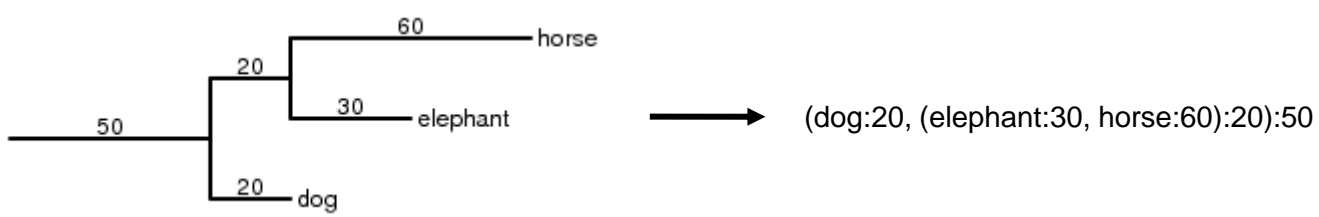


## Tree View

Enabling zoom:



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



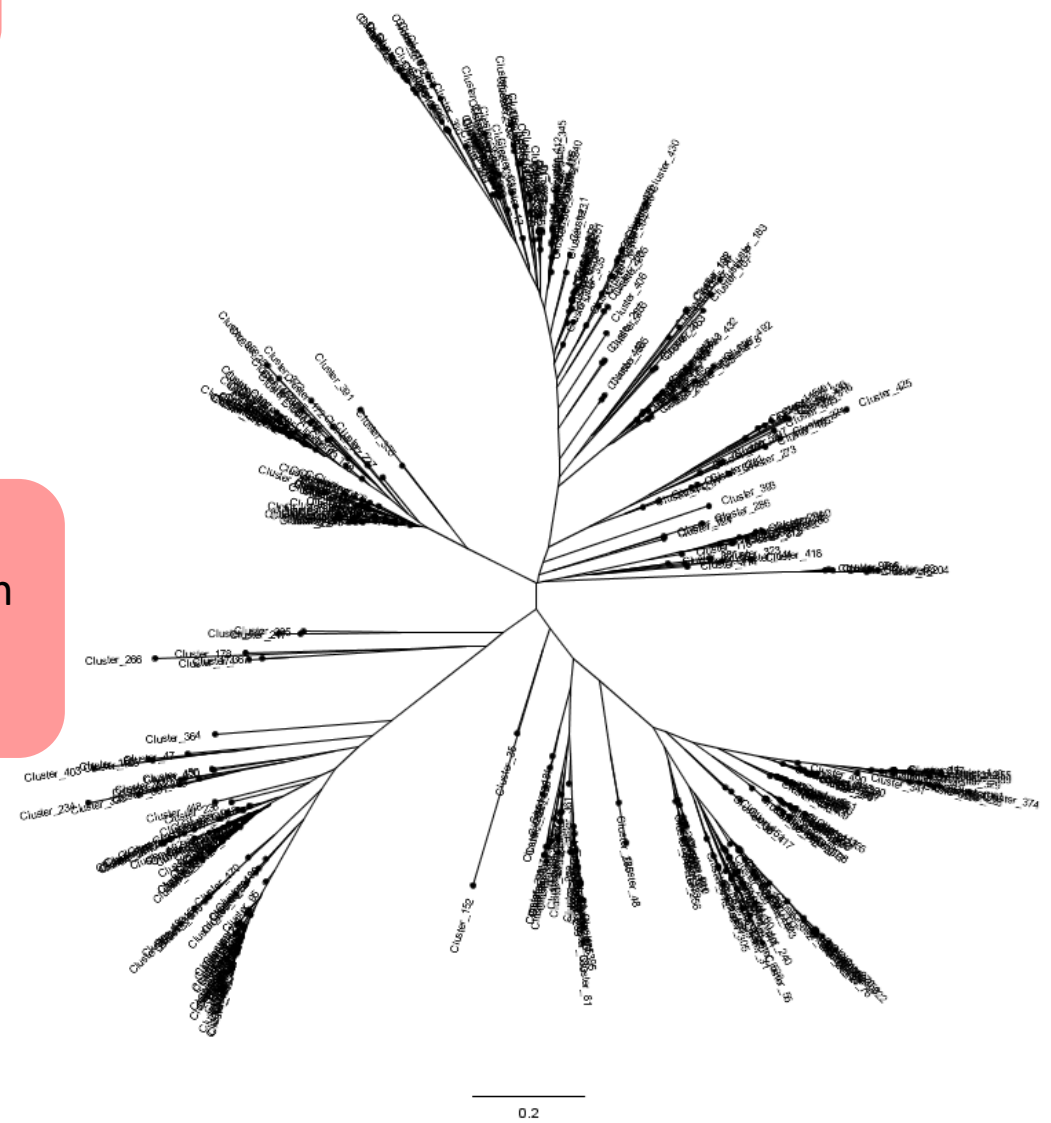
Our tree in nhx (= nwk) format

```

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

```

Exemple of visualization in FigTree from nhx file



---

# Practice:

---

# Exercise:

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

FROGS Tree Reconstruction of phylogenetic tree (Galaxy Version 4.1.0+galaxy1) Favorite Versions Options

**Sequence file**

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

**Biom file**

The abundance file (format: BIOM)

**Email notification**

No

Send an email notification when the job completes.

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

---

# Download your data

---

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

You have 2 backup possibilities:

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 arrow, then "Export History to File":

The screenshot shows a Galaxy history window titled "20: FROGS BIOM to TSV: abundance.tsv" with 495 lines and 1 comment. The format is tabular. A red arrow labeled "1" points to the floppy disk icon in the bottom toolbar. A second red arrow labeled "2" points to the dropdown arrow in the top right corner of the history window. A third red arrow labeled "2" points to the "Exporter l'Historique dans un fichier" option in the history menu.

The screenshot shows the "Export history archive" dialog box. It has a "to a link" button and a "to a remote file" button. The main text provides a link for download ready: <https://metabarcoding.usegalaxy.fr/api/histories/3e1619f028ae7909/exports/63d11d232c253fb8> (view job details). A note at the bottom states: "History archives are removed at regular intervals. For permanent storage download the archive, export to a remote file or import the archive on another Galaxy server." A red arrow labeled "2" points to the link.