Training on Galaxy: Metagenomics

Find Rapidly OTU with Galaxy Solution

FRÉDÉRIC Escudié* and LUCAS Auer*, MARIA Bernard, LAURENT Cauquil, KATIA Vidal, SARAH Maman,
MAHENDRA Mariadassou, GUILLERMINA Hernandez-Raquet, GÉRALDINE Pascal

*THESE AUTHORS HAVE CONTRIBUTED EQUALLY TO THE PRESENT WORK.

Feedback:

What are your needs in “metagenomics"?

454 / MiSeq ?

Your background ?

Overview

First day 2.00 pm to 5.00 pm

- Objectives
- Material: data + FROGS
- Data upload into galaxy environment
- Demultiplex tool
- Preprocess

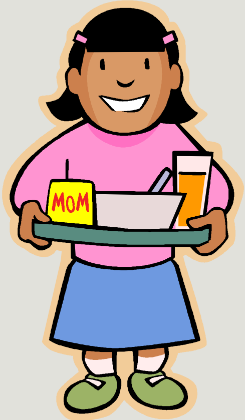
1 short coffee breaks
~3.30 pm

Overview

Second Day: 9.00 am to 5.00 pm

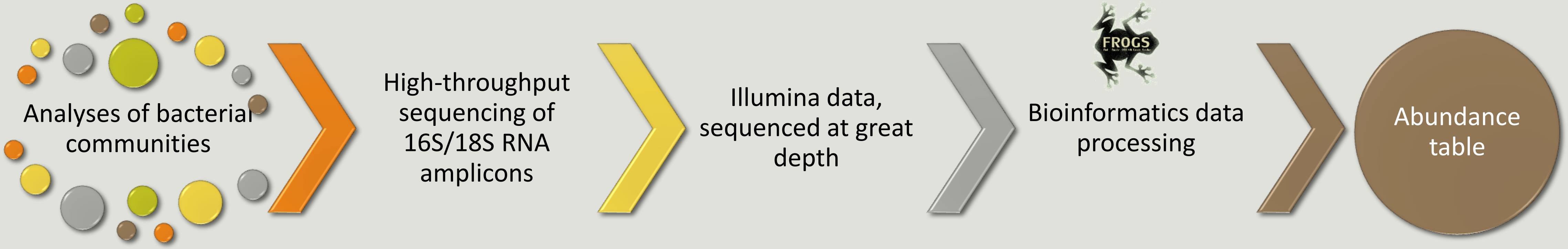
- Clustering + Cluster Statistics
- Removing chimeras
- Filtering
- Affiliation
- Normalization
- Tool Description
- Workflow creation
- Some figures
- Download data

2 short coffee breaks
morning and afternoon



Lunch
12.00 to 1.30 pm

Objectives



Analyses of bacterial communities

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

Illumina data, sequenced at great depth

Bioinformatics data processing

Abundance table

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

The goal:

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

Statistics

Objectives

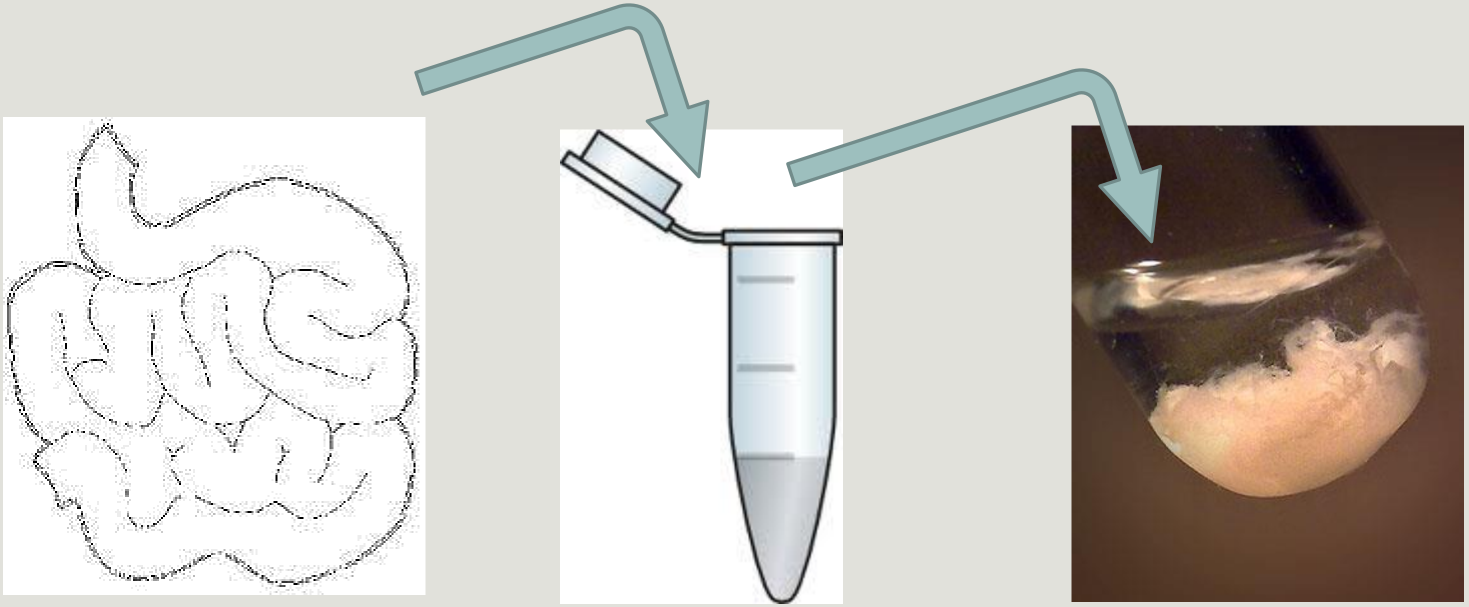
The current processing pipelines struggle to run in a reasonable time.

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

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

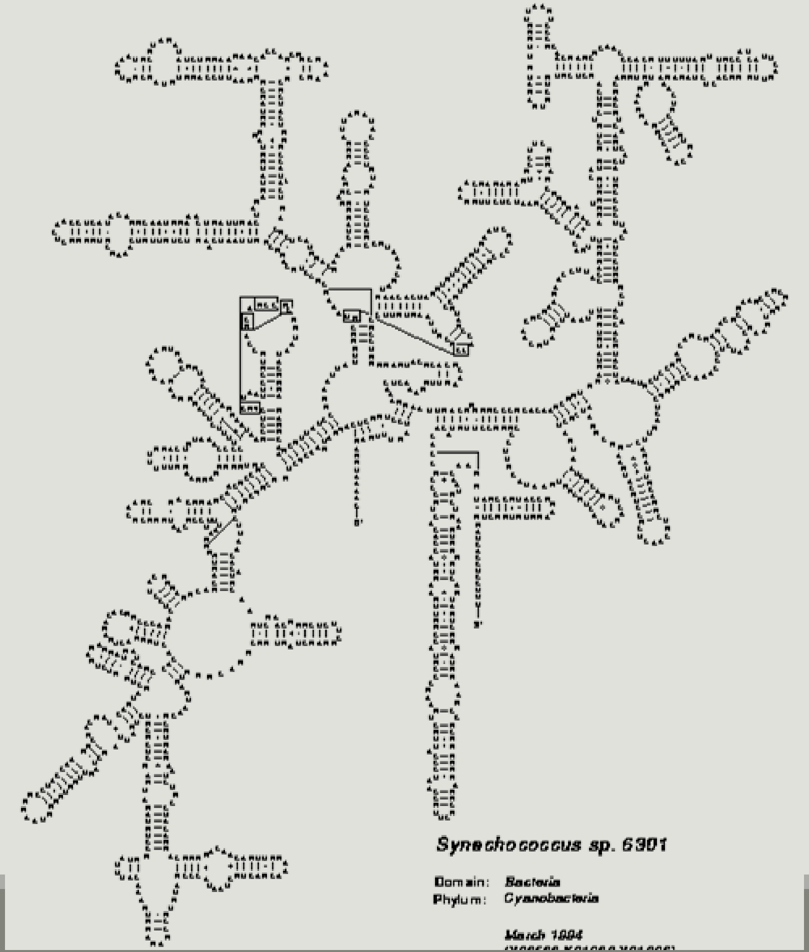
Material

Sample collection and DNA extraction

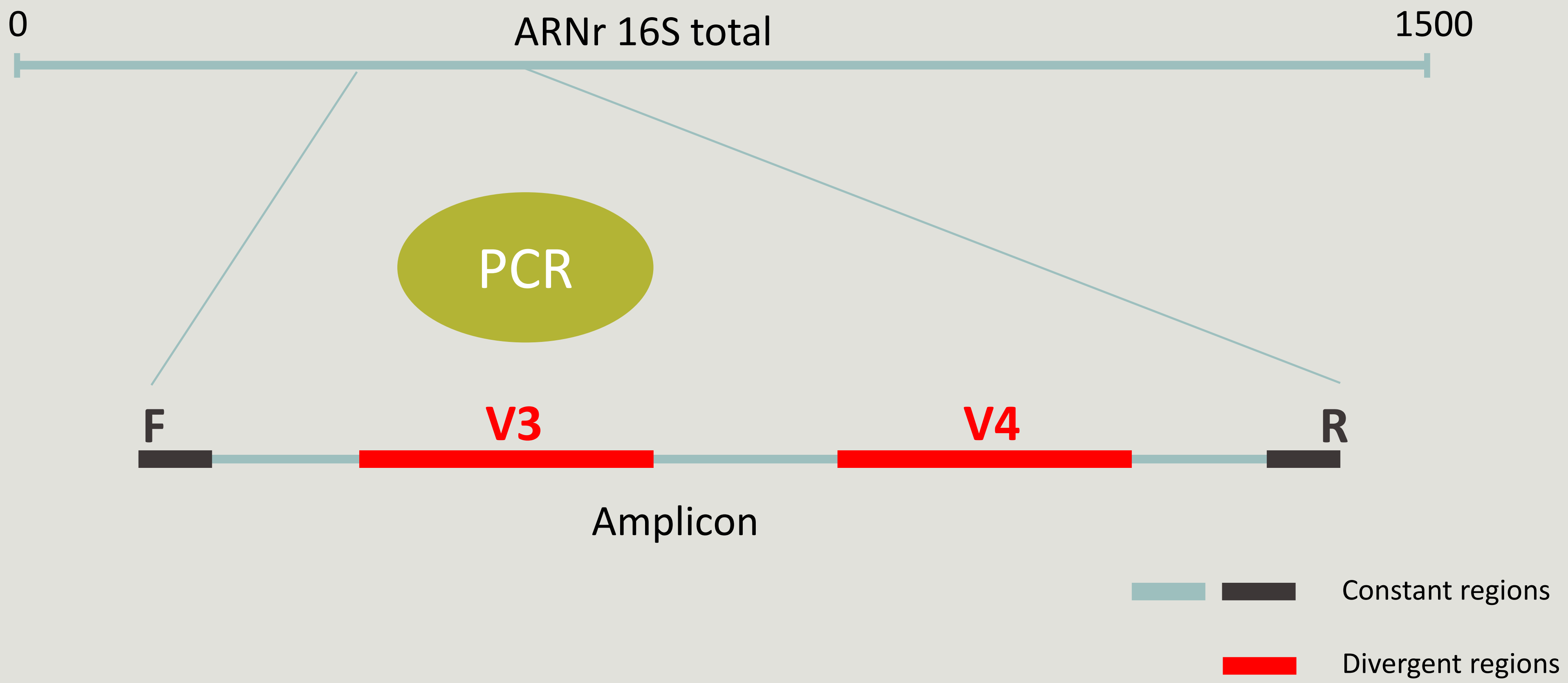


Identification of bacterial populations

Gene encoding the 16S subunit of ribosomal RNA (~ 1500 bp)



Identification of bacterial populations



0

ARNr 16S total

1500

PCR

F

V3

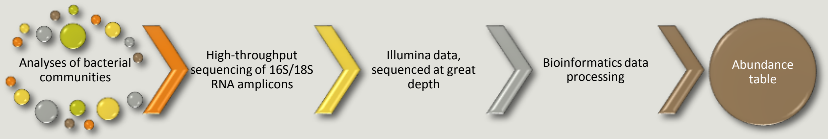
V4

R

Amplicon

Constant regions

Divergent regions



Which bioinformatics solutions ?

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

→ FROGS
Find Rapdly OTU with Galaxy Solution

FROGS ?

Use platform Galaxy

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

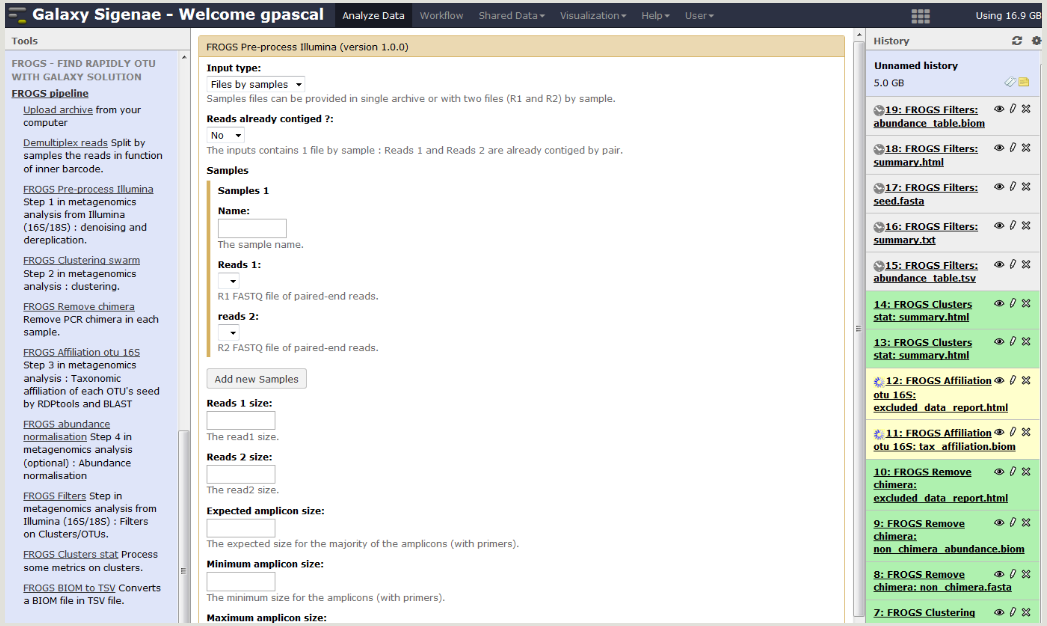
Independent modules

Run on Illumina/454 data 16S and 18S

New clustering method

Many graphics for interpretation

User friendly , hiding bioinformatics infrastructure/complexity



FROGS - FIND RAPIDLY OTU WITH GALAXY SOLUTION

FROGS pipeline

Upload archive from your computer

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

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

FROGS Clustering swarm Step 2 in metagenomics analysis : clustering.

FROGS Remove chimera Remove PCR chimera in each sample.

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

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

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

FROGS Clusters stat Process some metrics on clusters.

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

FROGS Pre-process Illumina (version 1.0.0)

Input type:
Files by samples
Samples files can be provided in single archive or with two files (R1 and R2) by sample.

Reads already contiged ?:
No
The inputs contains 1 file by sample : Reads 1 and Reads 2 are already contiged by pair.

Samples

Samples 1

Name:
The sample name.

Reads 1:
R1 FASTQ file of paired-end reads.

reads 2:
R2 FASTQ file of paired-end reads.

Add new Samples

Reads 1 size:
The read1 size.

Reads 2 size:
The read2 size.

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

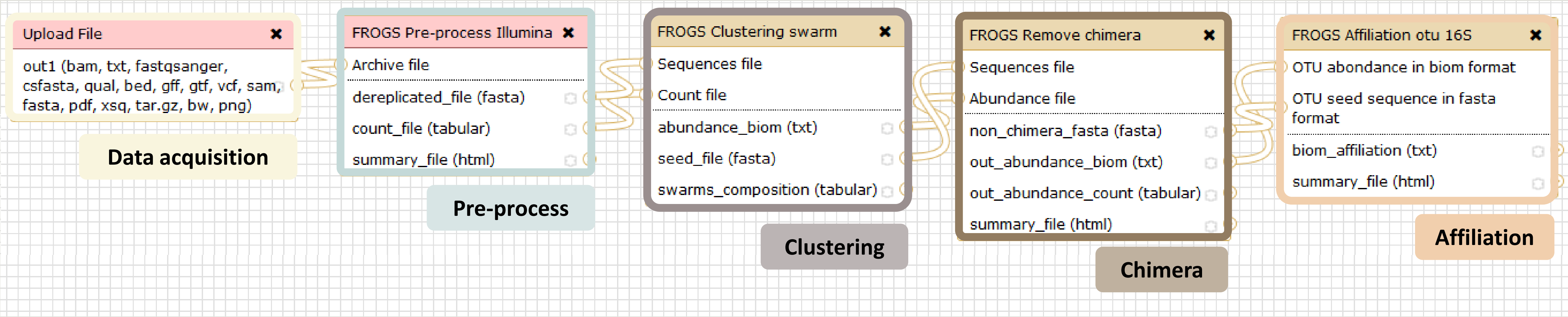
Minimum amplicon size:
The minimum size for the amplicons (with primers).

Maximum amplicon size:

Unnamed history
5.0 GB

19: FROGS Filters: abundance_table.biom
18: FROGS Filters: summary.html
17: FROGS Filters: seed.fasta
16: FROGS Filters: summary.txt
15: FROGS Filters: abundance_table.tsv
14: FROGS Clusters stat: summary.html
13: FROGS Clusters stat: summary.html
12: FROGS Affiliation otu 16S: excluded_data_report.html
11: FROGS Affiliation otu 16S: tax_affiliation.biom
10: FROGS Remove chimera: excluded_data_report.html
9: FROGS Remove chimera: non_chimera_abundance.biom
8: FROGS Remove chimera: non_chimera.fasta
7: FROGS Clustering

FROGS pipeline



Upload File ✖

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

Data acquisition

FROGS Pre-process Illumina ✖

Archive file

dereplicated_file (fasta)

count_file (tabular)

summary_file (html)

Pre-process

FROGS Clustering swarm ✖

Sequences file

Count file

abundance_biom (txt)

seed_file (fasta)

swarms_composition (tabular)

Clustering

FROGS Remove chimera ✖

Sequences file

Abundance file

non_chimera_fasta (fasta)

out_abundance_biom (txt)

out_abundance_count (tabular)

summary_file (html)

Chimera

FROGS Affiliation otu 16S ✖

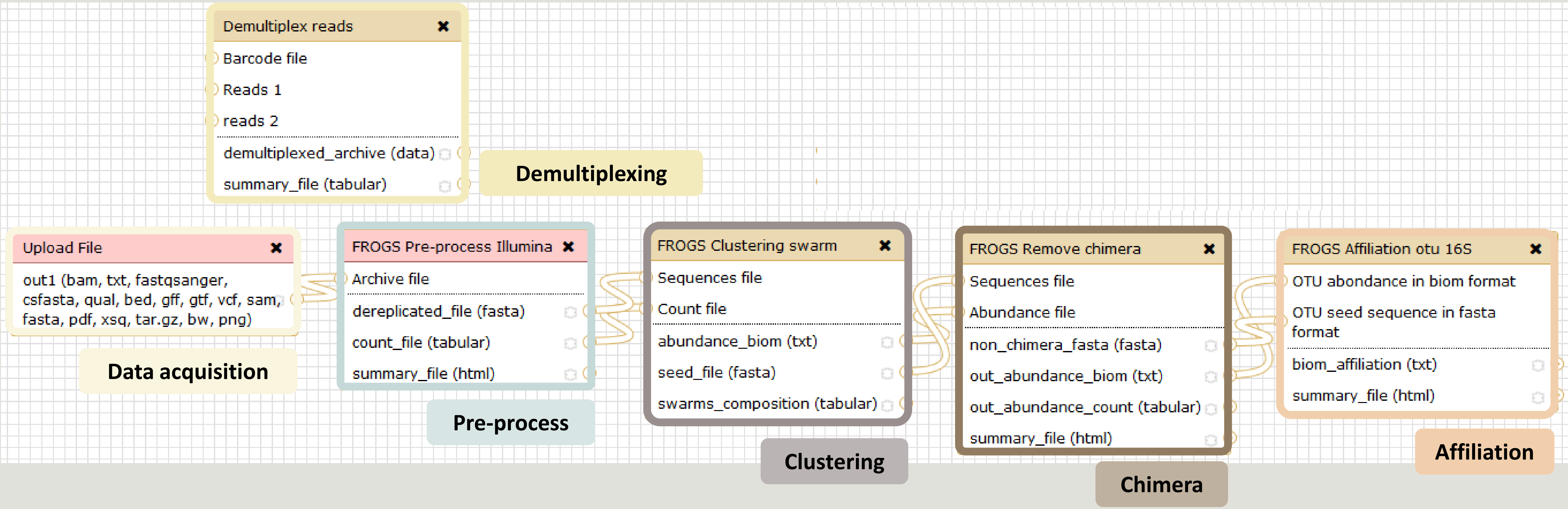
OTU abondance in biom format

OTU seed sequence in fasta format

biom_affiliation (txt)

summary_file (html)

Affiliation



Demultiplex reads ✖

Barcode file

Reads 1

reads 2

demultiplexed_archive (data)

summary_file (tabular)

Demultiplexing

Upload File ✖

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

Data acquisition

FROGS Pre-process Illumina ✖

Archive file

dereplicated_file (fasta)

count_file (tabular)

summary_file (html)

Pre-process

FROGS Clustering swarm ✖

Sequences file

Count file

abundance_biom (txt)

seed_file (fasta)

swarms_composition (tabular)

Clustering

FROGS Remove chimera ✖

Sequences file

Abundance file

non_chimera_fasta (fasta)

out_abundance_biom (txt)

out_abundance_count (tabular)

summary_file (html)

Chimera

FROGS Affiliation otu 16S ✖

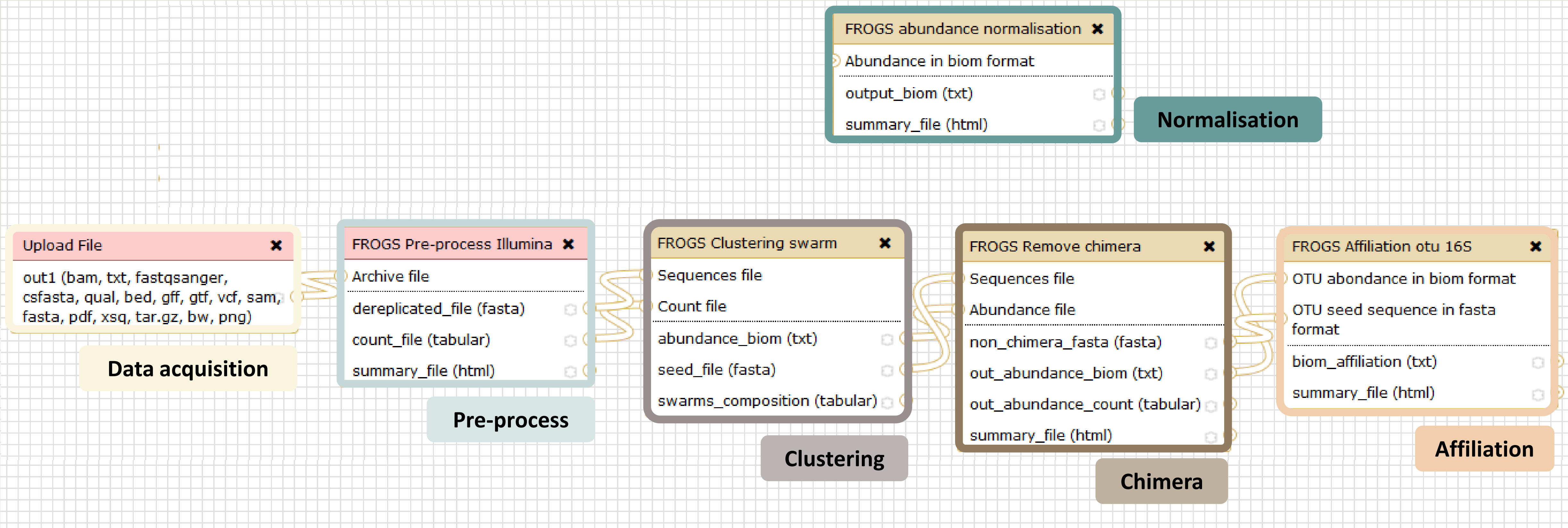
OTU abondance in biom format

OTU seed sequence in fasta format

biom_affiliation (txt)

summary_file (html)

Affiliation



FROGS abundance normalisation ✖

Abundance in biom format

output_biom (txt)

summary_file (html)

Normalisation

Upload File ✖

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

Data acquisition

FROGS Pre-process Illumina ✖

Archive file

dereplicated_file (fasta)

count_file (tabular)

summary_file (html)

Pre-process

FROGS Clustering swarm ✖

Sequences file

Count file

abundance_biom (txt)

seed_file (fasta)

swarms_composition (tabular)

Clustering

FROGS Remove chimera ✖

Sequences file

Abundance file

non_chimera_fasta (fasta)

out_abundance_biom (txt)

out_abundance_count (tabular)

summary_file (html)

Chimera

FROGS Affiliation otu 16S ✖

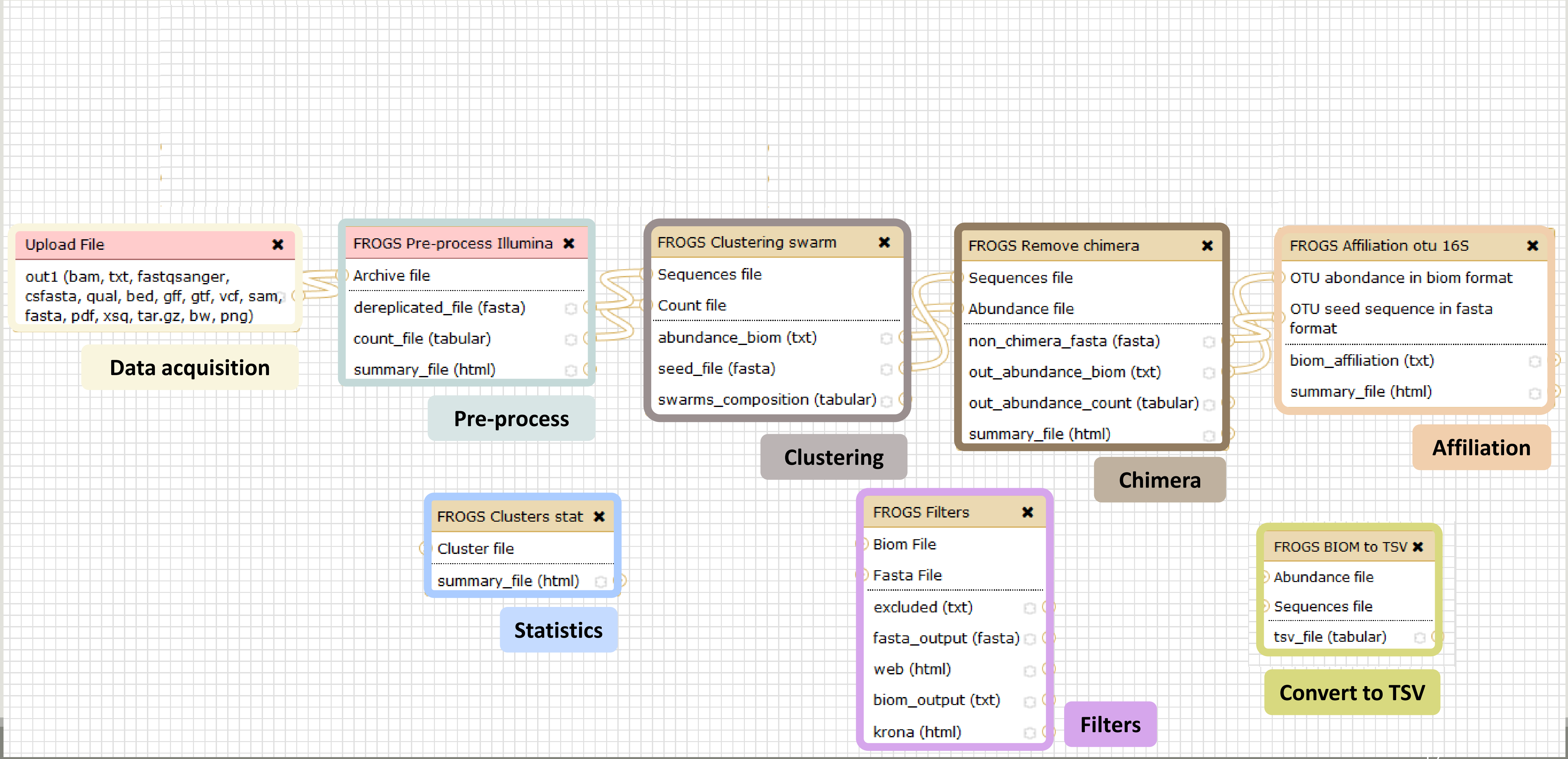
OTU abondance in biom format

OTU seed sequence in fasta format

biom_affiliation (txt)

summary_file (html)

Affiliation



Upload File ✖

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

Data acquisition

FROGS Pre-process Illumina ✖

Archive file

dereplicated_file (fasta)

count_file (tabular)

summary_file (html)

Pre-process

FROGS Clustering swarm ✖

Sequences file

Count file

abundance_biom (txt)

seed_file (fasta)

swarms_composition (tabular)

Clustering

FROGS Remove chimera ✖

Sequences file

Abundance file

non_chimera_fasta (fasta)

out_abundance_biom (txt)

out_abundance_count (tabular)

summary_file (html)

Chimera

FROGS Affiliation otu 16S ✖

OTU abondance in biom format

OTU seed sequence in fasta format

biom_affiliation (txt)

summary_file (html)

Affiliation

FROGS Clusters stat ✖

Cluster file

summary_file (html)

Statistics

FROGS Filters ✖

Biom File

Fasta File

excluded (txt)

fasta_output (fasta)

web (html)

biom_output (txt)

krona (html)

Filters

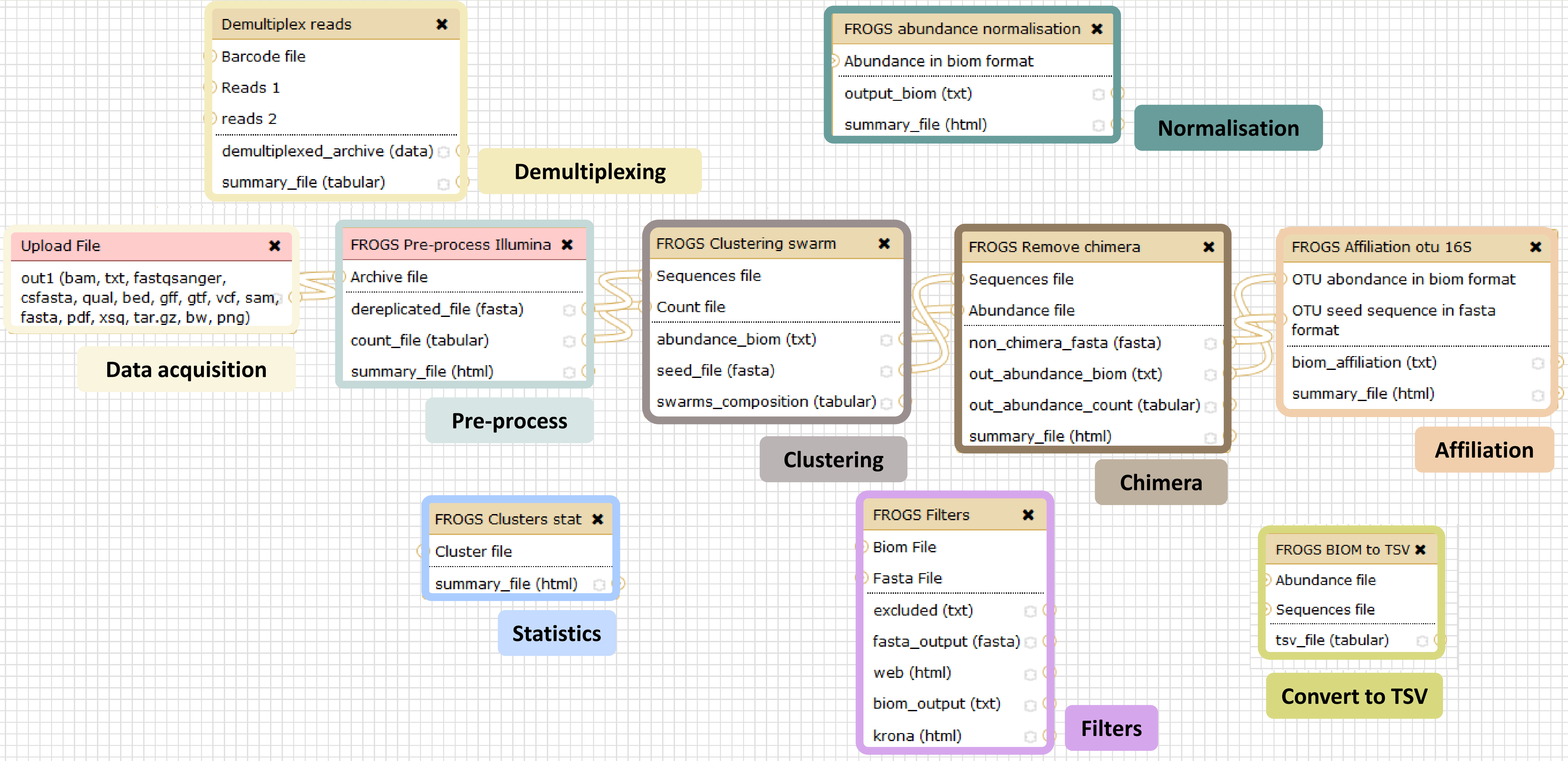
FROGS BIOM to TSV ✖

Abundance file

Sequences file

tsv_file (tabular)

Convert to TSV



Demultiplex reads ✖
Barcode file
Reads 1
reads 2
demultiplexed_archive (data)
summary_file (tabular)

Demultiplexing

FROGS abundance normalisation ✖
Abundance in biom format
output_biom (txt)
summary_file (html)

Normalisation

Upload File ✖
out1 (bam, txt, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png)

Data acquisition

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

Pre-process

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

Clustering

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

Chimera

FROGS Affiliation otu 16S ✖
OTU abondance in biom format
OTU seed sequence in fasta format
biom_affiliation (txt)
summary_file (html)

Affiliation

FROGS Clusters stat ✖
Cluster file
summary_file (html)

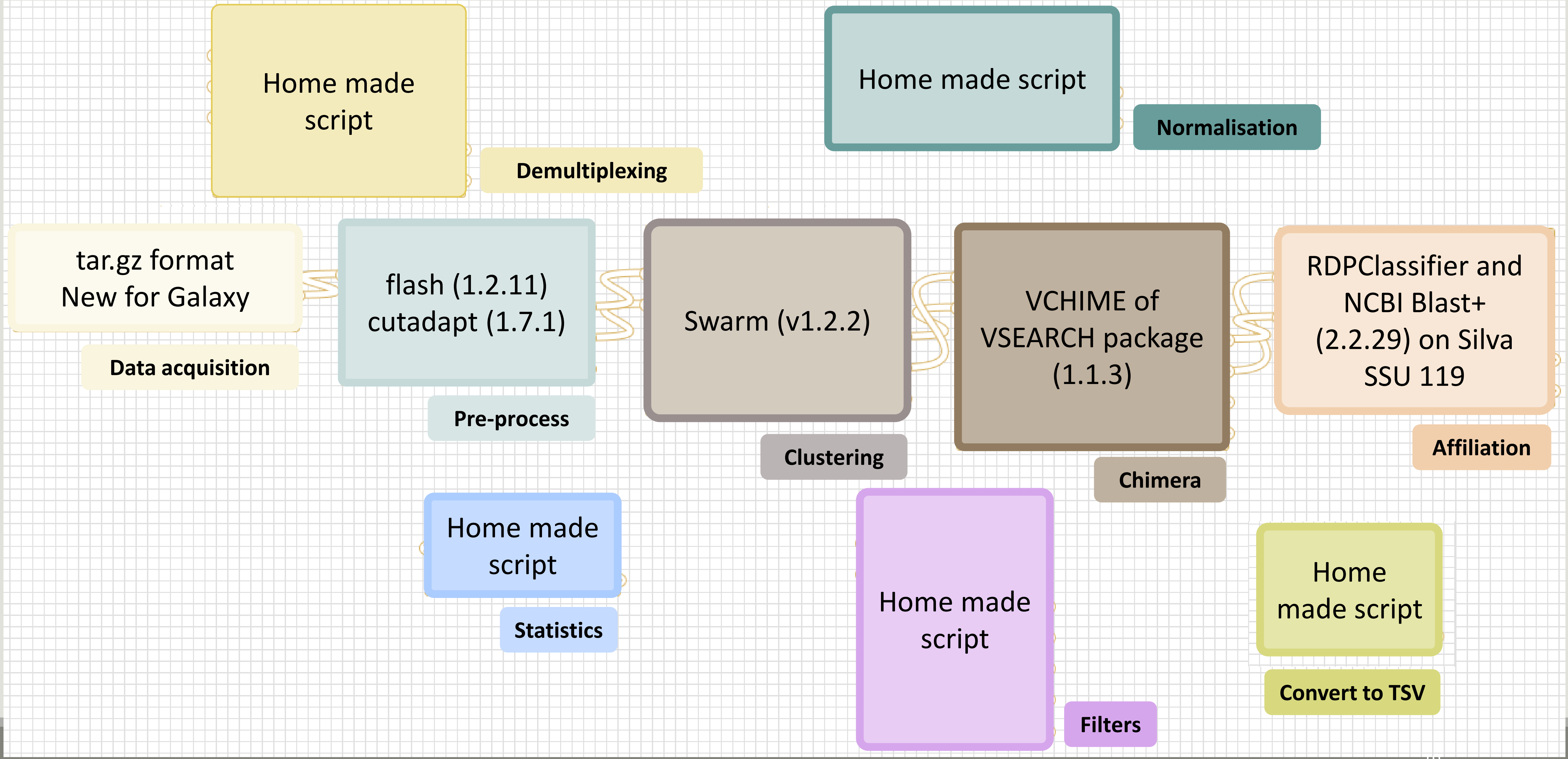
Statistics

FROGS Filters ✖
Biom File
Fasta File
excluded (txt)
fasta_output (fasta)
web (html)
biom_output (txt)
krona (html)

Filters

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

Convert to TSV

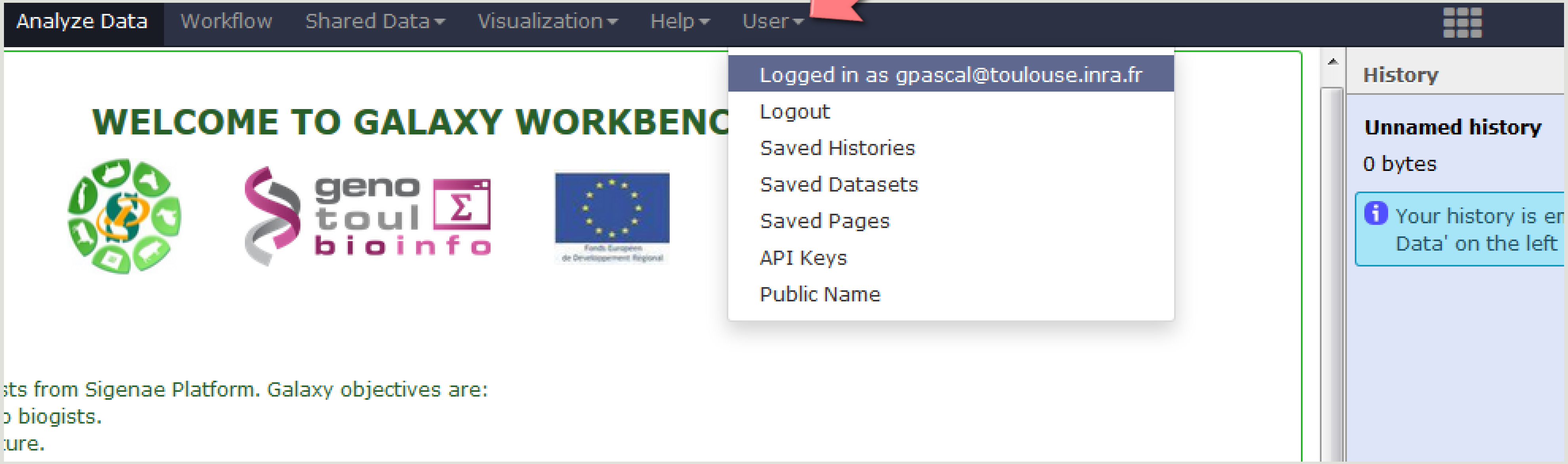


Together go to visit FROGS

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

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

Enter your login and password from GenoToul



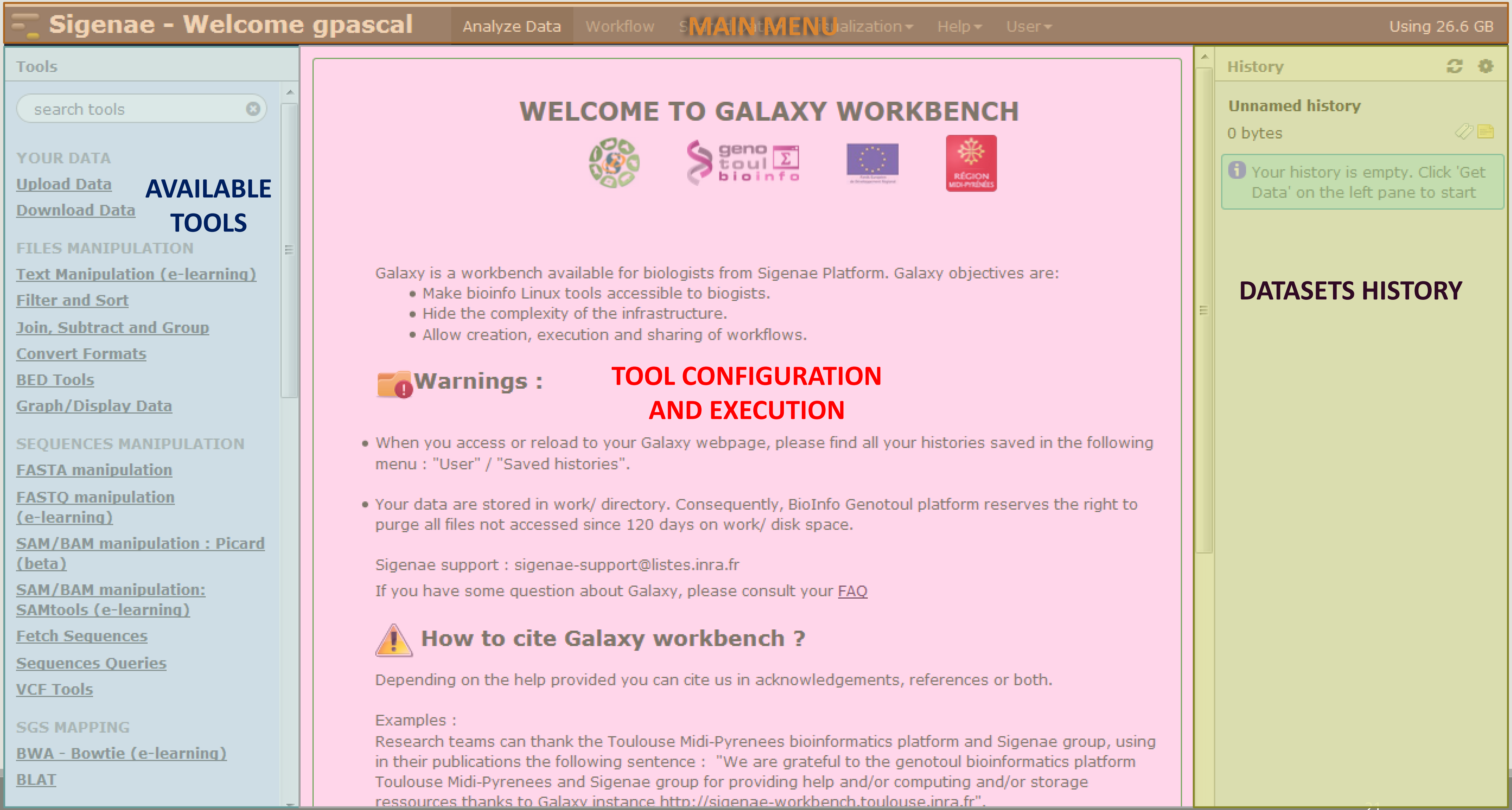
WELCOME TO GALAXY WORKBENC

Unnamed history

0 bytes

Your history is en
Data' on the left

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



MAIN MENU

AVAILABLE TOOLS

TOOL CONFIGURATION AND EXECUTION

DATASETS HISTORY

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

WELCOME TO GALAXY WORKBENCH

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

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

Warnings :

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

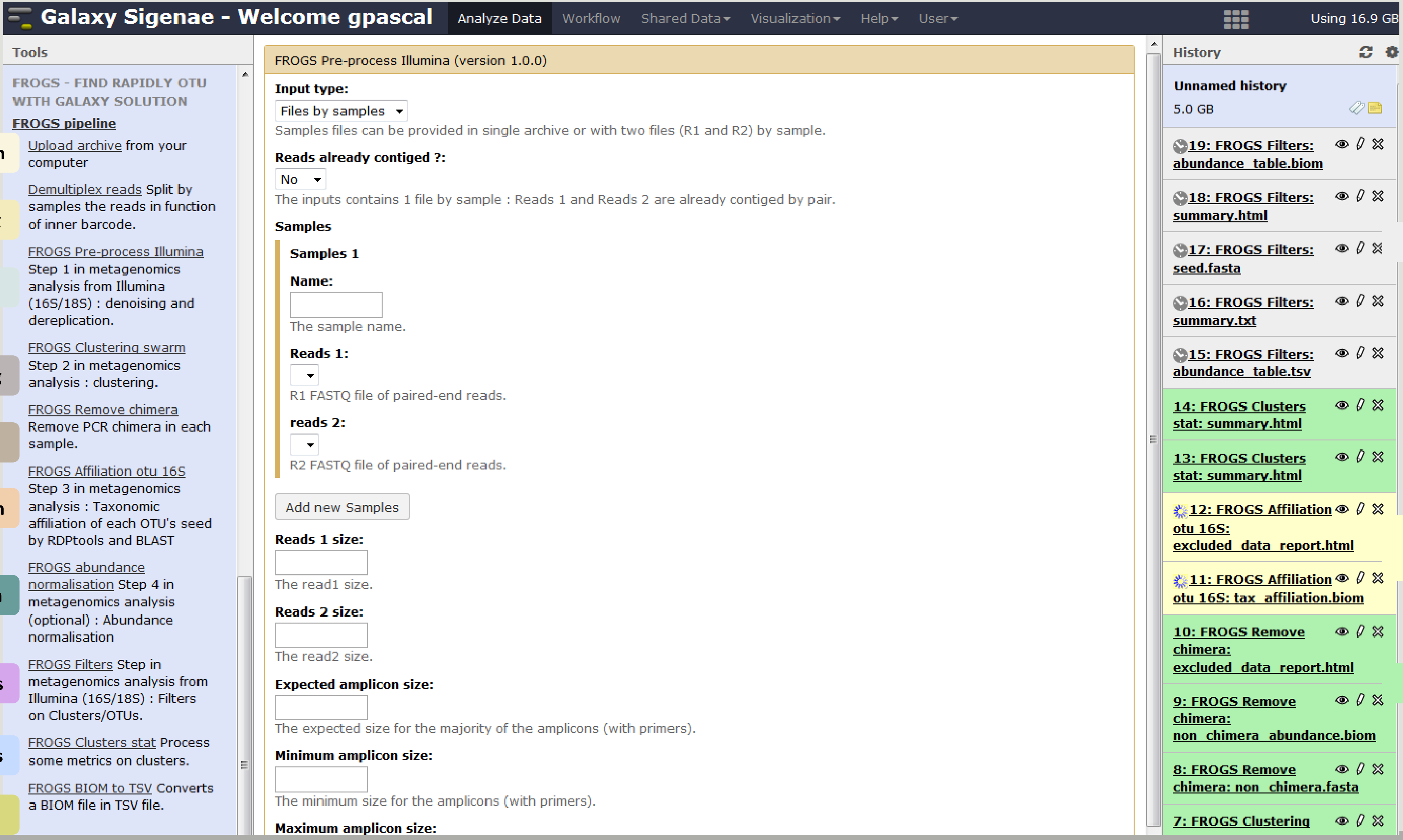
Sigenae support : sigenae-support@listes.inra.fr

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

How to cite Galaxy workbench ?

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

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



Tools

FROGS - FIND RAPIDLY OTU WITH GALAXY SOLUTION

FROGS pipeline

Data acquisition — Upload archive from your computer

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

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

Clustering — FROGS Clustering swarm Step 2 in metagenomics analysis : clustering.

Chimera — FROGS Remove chimera Remove PCR chimera in each sample.

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

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

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

Statistics — FROGS Clusters stat Process some metrics on clusters.

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

FROGS Pre-process Illumina (version 1.0.0)

Input type: Files by samples

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

Reads already contiged ?: No

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

Samples

Samples 1

Name:

The sample name.

Reads 1:

R1 FASTQ file of paired-end reads.

reads 2:

R2 FASTQ file of paired-end reads.

Add new Samples

Reads 1 size:

The read1 size.

Reads 2 size:

The read2 size.

Expected amplicon size:

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

Minimum amplicon size:

The minimum size for the amplicons (with primers).

Maximum amplicon size:

History

Unnamed history

5.0 GB

Waiting to run

- 19: FROGS Filters: abundance_table.biom
- 18: FROGS Filters: summary.html
- 17: FROGS Filters: seed.fasta
- 16: FROGS Filters: summary.txt
- 15: FROGS Filters: abundance_table.tsv

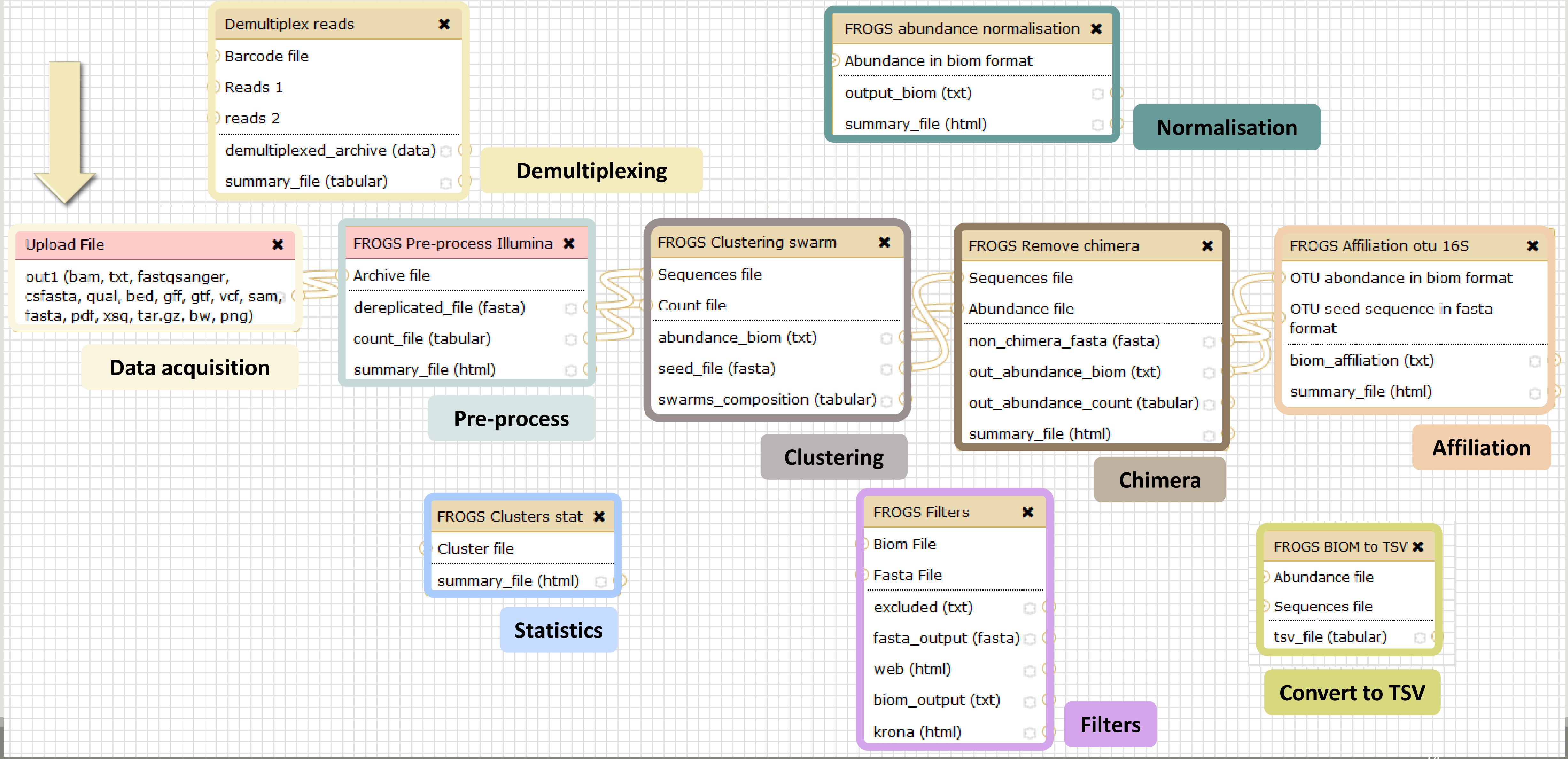
Currently running

- 12: FROGS Affiliation otu 16S: excluded_data_report.html
- 11: FROGS Affiliation otu 16S: tax_affiliation.biom

Result files

- 14: FROGS Clusters stat: summary.html
- 13: FROGS Clusters stat: summary.html
- 10: FROGS Remove chimera: excluded_data_report.html
- 9: FROGS Remove chimera: non_chimera_abundance.biom
- 8: FROGS Remove chimera: non_chimera.fasta
- 7: FROGS Clustering

Upload data



Upload File ✖

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

Data acquisition

Demultiplex reads ✖

Barcode file

Reads 1

reads 2

demultiplexed_archive (data)

summary_file (tabular)

Demultiplexing

FROGS Pre-process Illumina ✖

Archive file

dereplicated_file (fasta)

count_file (tabular)

summary_file (html)

Pre-process

FROGS Clustering swarm ✖

Sequences file

Count file

abundance_biom (txt)

seed_file (fasta)

swarms_composition (tabular)

Clustering

FROGS Remove chimera ✖

Sequences file

Abundance file

non_chimera_fasta (fasta)

out_abundance_biom (txt)

out_abundance_count (tabular)

summary_file (html)

Chimera

FROGS Affiliation otu 16S ✖

OTU abondance in biom format

OTU seed sequence in fasta format

biom_affiliation (txt)

summary_file (html)

Affiliation

FROGS abundance normalisation ✖

Abundance in biom format

output_biom (txt)

summary_file (html)

Normalisation

FROGS Clusters stat ✖

Cluster file

summary_file (html)

Statistics

FROGS Filters ✖

Biom File

Fasta File

excluded (txt)

fasta_output (fasta)

web (html)

biom_output (txt)

krona (html)

Filters

FROGS BIOM to TSV ✖

Abundance file

Sequences file

tsv_file (tabular)

Convert to TSV

What kind of data ?

4 Upload → 4 Histories

Multiplexed data

Pathobiomes
rodents and ticks

multiplex.fastq

barcode.tabular

454 data

Freshwater sediment
metagenome

454.fastq.gz

SRA number
SRR443364

MiSeq
R1 fastq + R2 fastq

Farm animal feces
metagenome

sampleA_R1.fasta

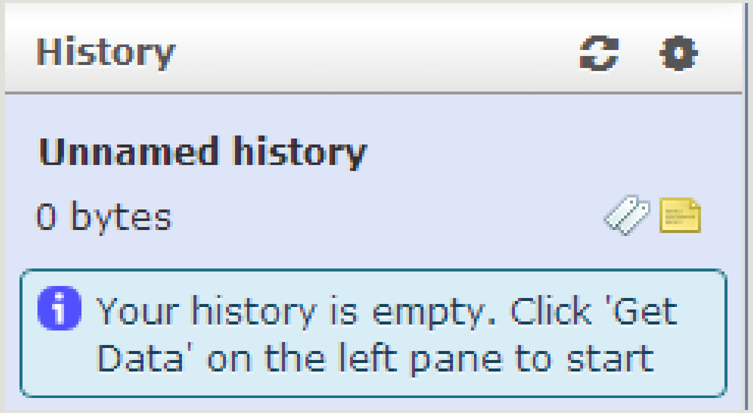
sampleA_R2.fasta

MiSeq contiged fastq
in archive tar.gz

Farm animal feces
metagenome

100spec_90000seq_9samples.tar.gz

1ST CONNEXION


History

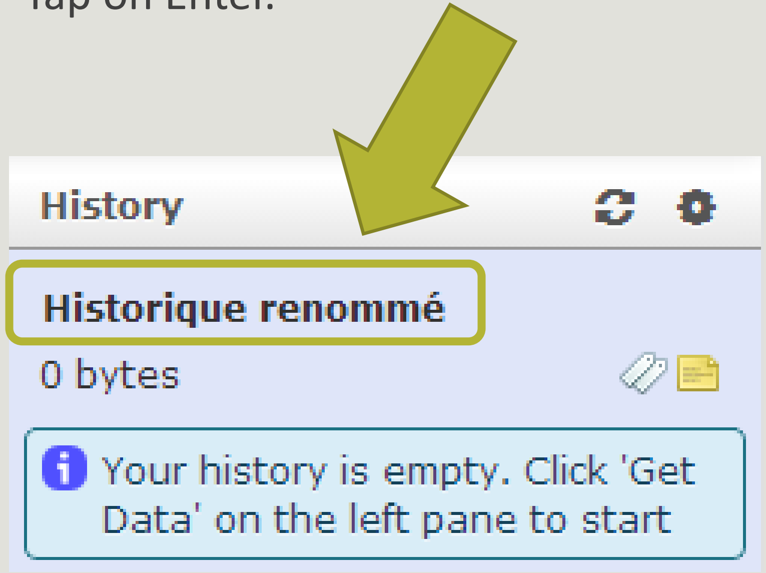
Unnamed history

0 bytes

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

RENAME HISTORY

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


History

Historique renommé

0 bytes

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

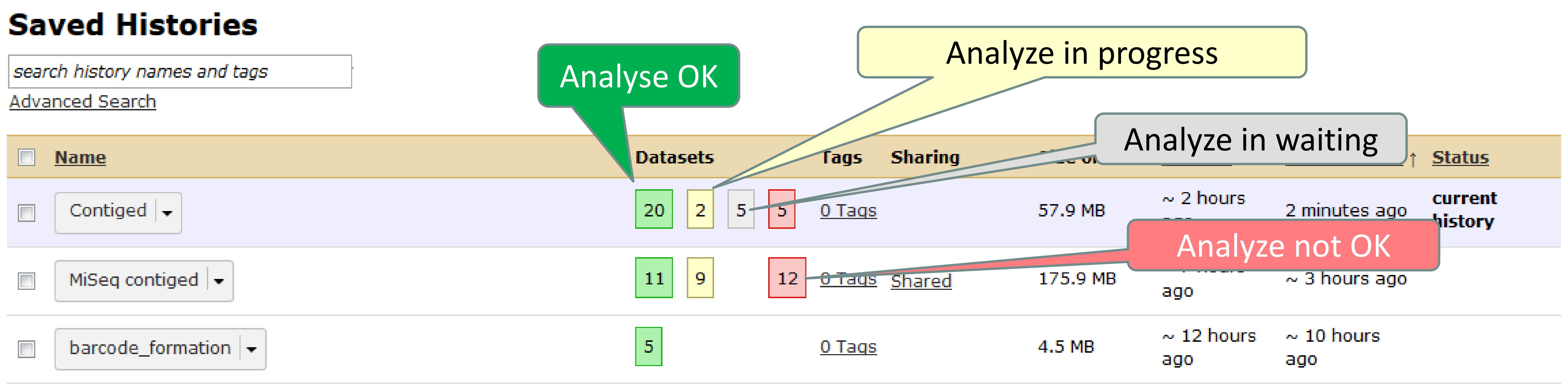
A vous de jouer ! - 1

SEE EXERCISE 1

History gestion

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

Saved Histories



Saved Histories

search history names and tags

Advanced Search

□	Name	Datasets	Tags	Sharing			↑	Status
□	Contiged ▾	20 2 5 5	0 Tags		57.9 MB	~ 2 hours	2 minutes ago	current history
□	MiSeq contiged ▾	11 9 12	0 Tags	Shared	175.9 MB	ago	~ 3 hours ago	
□	barcode_formation ▾	5	0 Tags		4.5 MB	~ 12 hours ago	~ 10 hours ago	

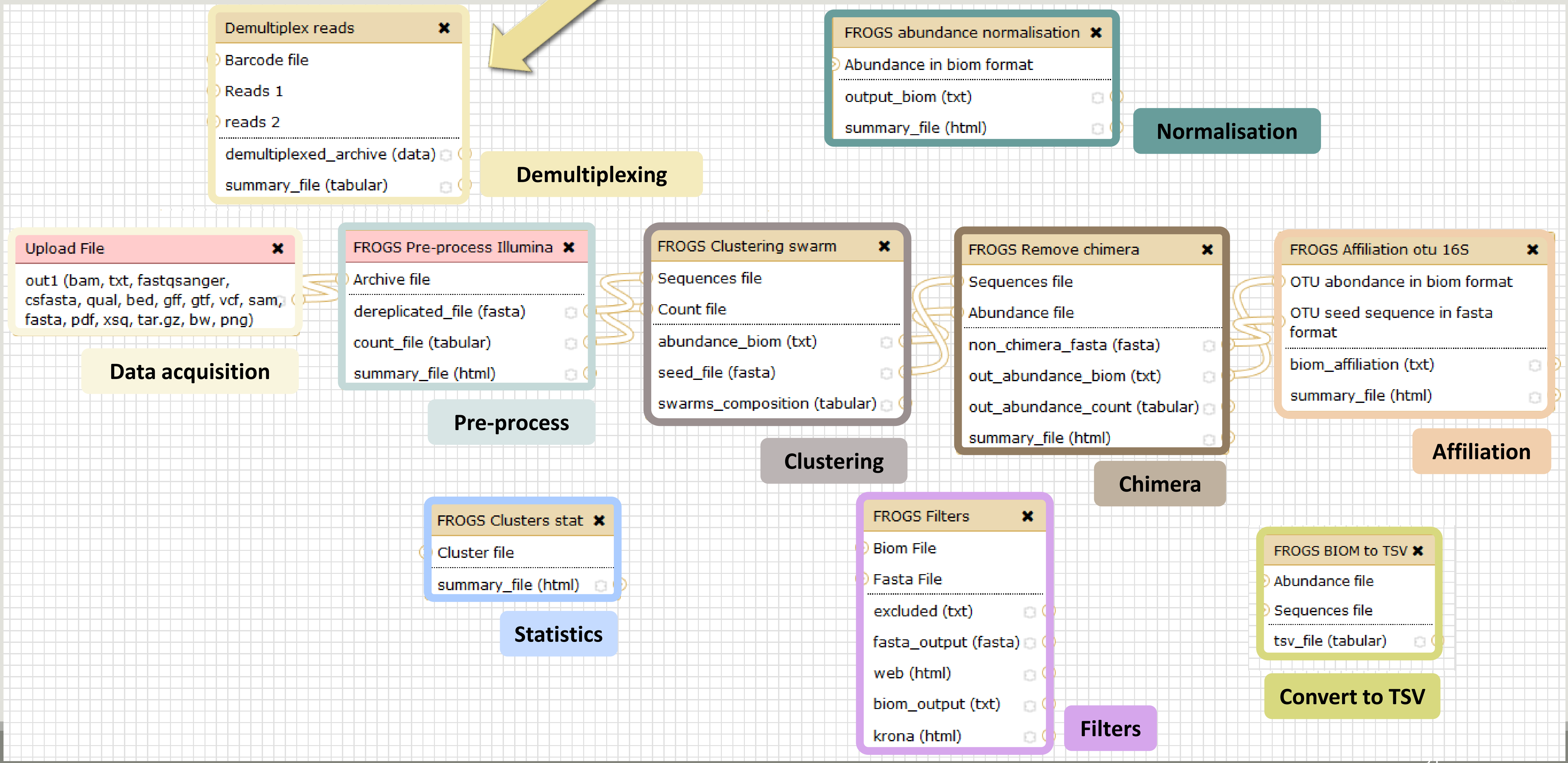
Analyse OK

Analyze in progress

Analyze in waiting

Analyze not OK

Demultiplexing tool



Demultiplex reads ✖
Barcode file
Reads 1
reads 2
demultiplexed_archive (data)
summary_file (tabular)

Demultiplexing

FROGS abundance normalisation ✖
Abundance in biom format
output_biom (txt)
summary_file (html)

Normalisation

Upload File ✖
out1 (bam, txt, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png)

Data acquisition

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

Pre-process

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

Clustering

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

Chimera

FROGS Affiliation otu 16S ✖
OTU abondance in biom format
OTU seed sequence in fasta format
biom_affiliation (txt)
summary_file (html)

Affiliation

FROGS Clusters stat ✖
Cluster file
summary_file (html)

Statistics

FROGS Filters ✖
Biom File
Fasta File
excluded (txt)
fasta_output (fasta)
web (html)
biom_output (txt)
krona (html)

Filters

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

Convert to TSV

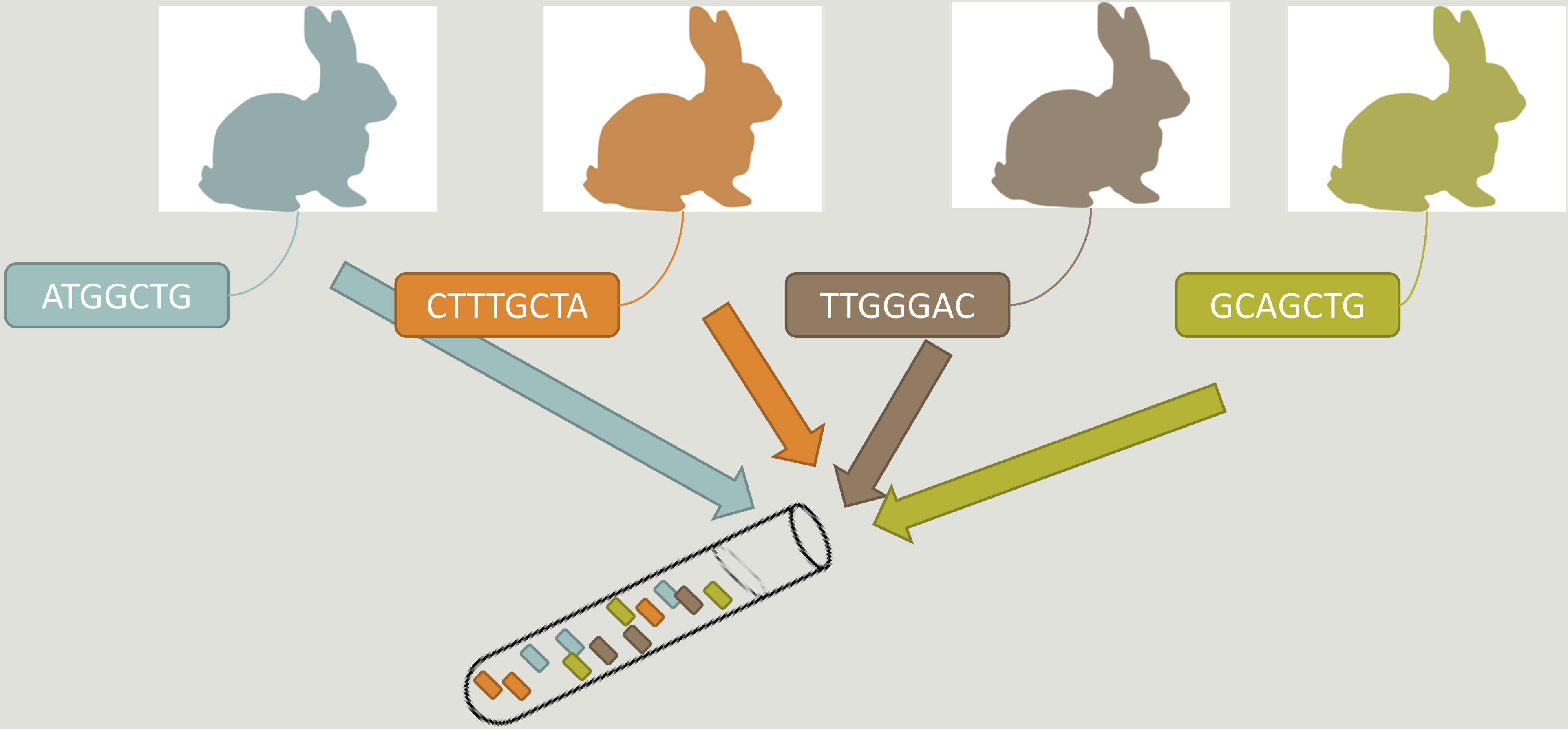
Demultiplexing

Sequence demultiplexing in function of barcode sequences :

- In forward
- In reverse
- In forward and reverse

Remove unbarcoded or ambiguous sequences

Barcoding ?



A vous de jouer ! - 2

GO TO EXERCISE 2

multiplexed

Format: Barcode

BARCODE FILE is expected to be tabulated:

- first column corresponds to the sample name
- second to the sequence barcode used
- optional third is the reverse sequence barcode

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
```

multiplexed

Format : FastQ

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

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

```
@HNHOSKD01ALD0H
ACAGCGTCAGAGGGGTACCAGTCAGCCATGACGTAGCACGTACA
+
CCCFFFFFFHHHHHJJIJJJHHFF@DEDDDDDDD@CDDDDACDD
```

How it works ?

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

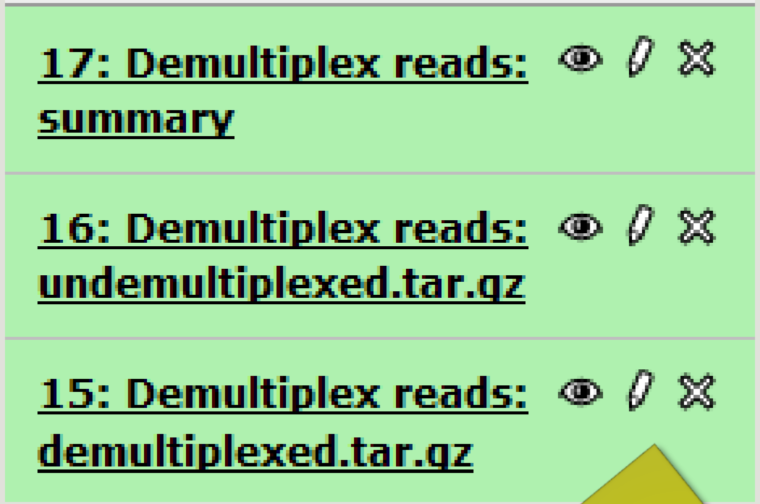
If this fragment is equal (with less or equal mismatch than the threshold) to one (and only one) barcode, the fragment is trimmed and the sequence will be attributed to the corresponding sample.

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

Advice

- Do not forget to indicate barcode sequence as they actually 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 oftenly those sequence are very short so 1 mismatch is already more than the sequencing error rate.
- If you have different barcode length, 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 it with some program like sff2fastq

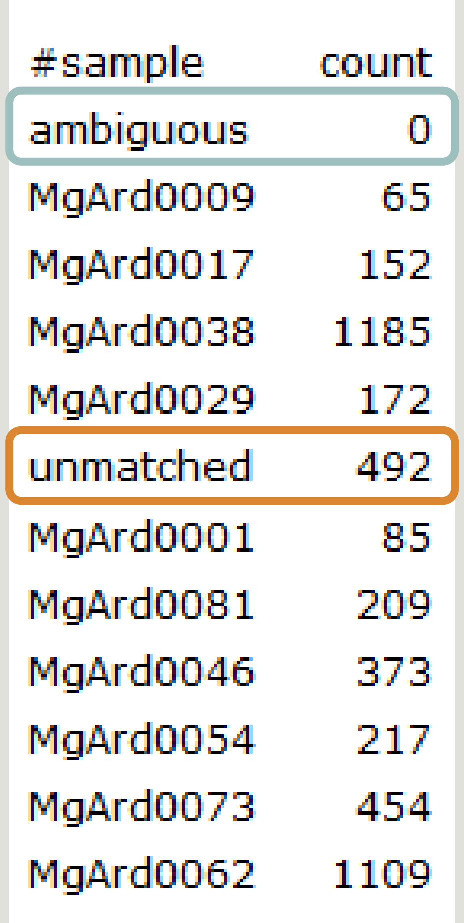
Results



17: Demultiplex reads: summary

16: Demultiplex reads: undemultiplexed.tar.gz

15: Demultiplex reads: demultiplexed.tar.gz



#sample	count
ambiguous	0
MgArd0009	65
MgArd0017	152
MgArd0038	1185
MgArd0029	172
unmatched	492
MgArd0001	85
MgArd0081	209
MgArd0046	373
MgArd0054	217
MgArd0073	454
MgArd0062	1109

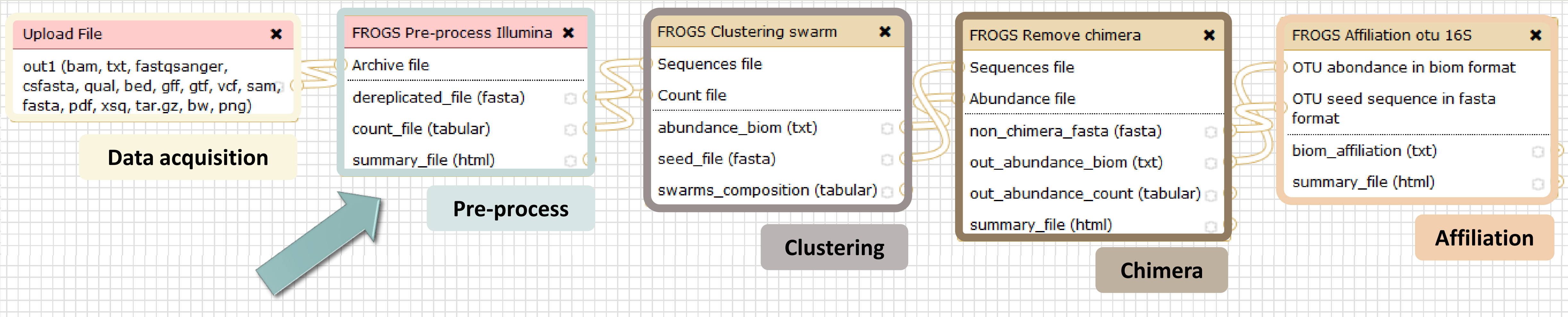
With barcode mismatches >1 sequence can corresponding to several samples.
So these sequences are non-affected to a sample.

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

Create a tar archive by grouping one (pair) fastq file per sample whith names indicate in the first column of the barcode.tsv tabular file

Pre-process tool

FROGS pipeline



Upload File ✖

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

Data acquisition

FROGS Pre-process Illumina ✖

Archive file

dereplicated_file (fasta)

count_file (tabular)

summary_file (html)

Pre-process

FROGS Clustering swarm ✖

Sequences file

Count file

abundance_biom (txt)

seed_file (fasta)

swarms_composition (tabular)

Clustering

FROGS Remove chimera ✖

Sequences file

Abundance file

non_chimera_fasta (fasta)

out_abundance_biom (txt)

out_abundance_count (tabular)

summary_file (html)

Chimera

FROGS Affiliation otu 16S ✖

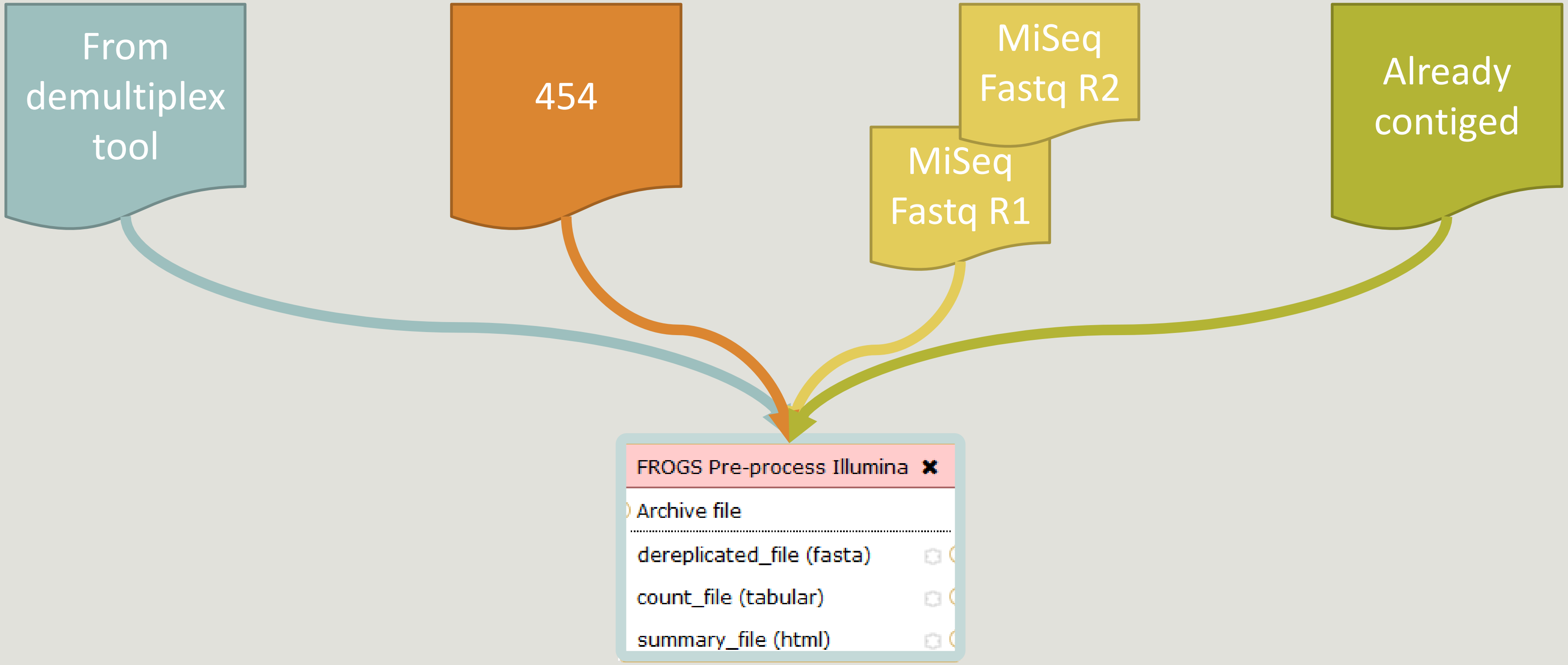
OTU abondance in biom format

OTU seed sequence in fasta format

biom_affiliation (txt)

summary_file (html)

Affiliation



From demultiplex tool

454

MiSeq Fastq R2

MiSeq Fastq R1

Already contiged

FROGS Pre-process Illumina ✖

Archive file

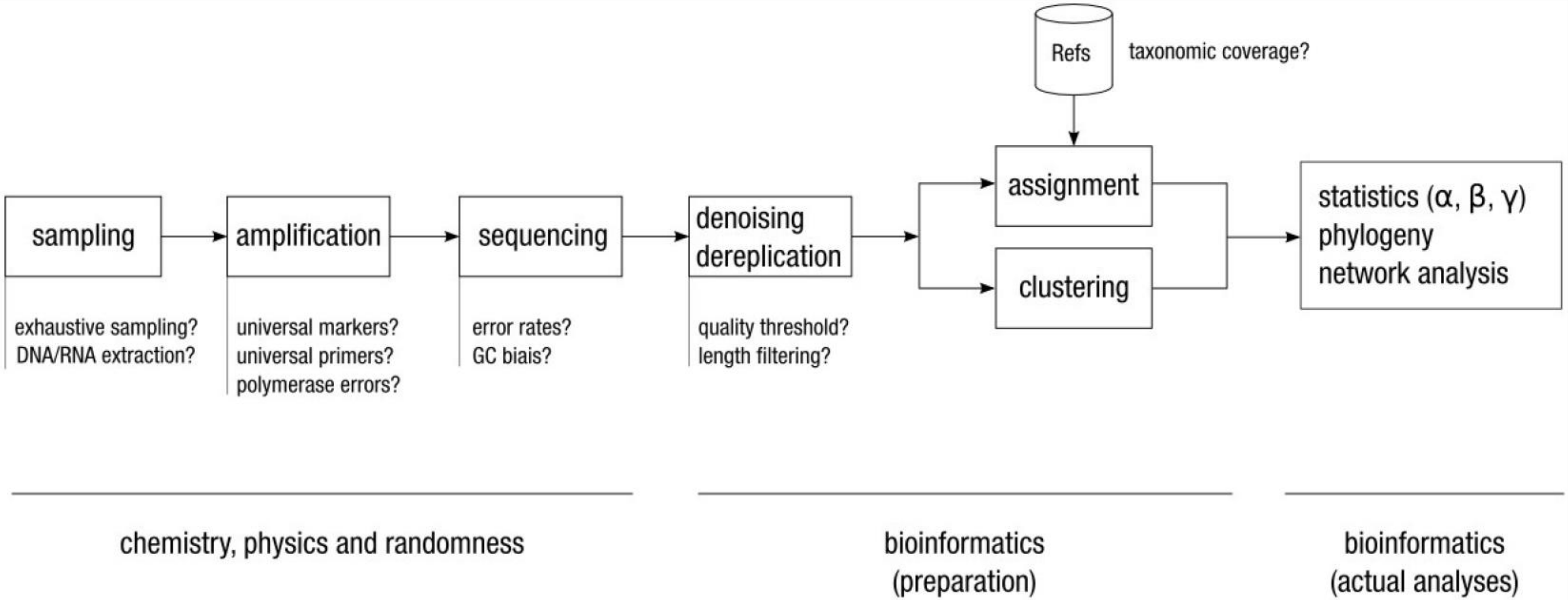
dereplicated_file (fasta)

count_file (tabular)

summary_file (html)

Pre-process

Amplicon-based studies general pipeline



Pre-process

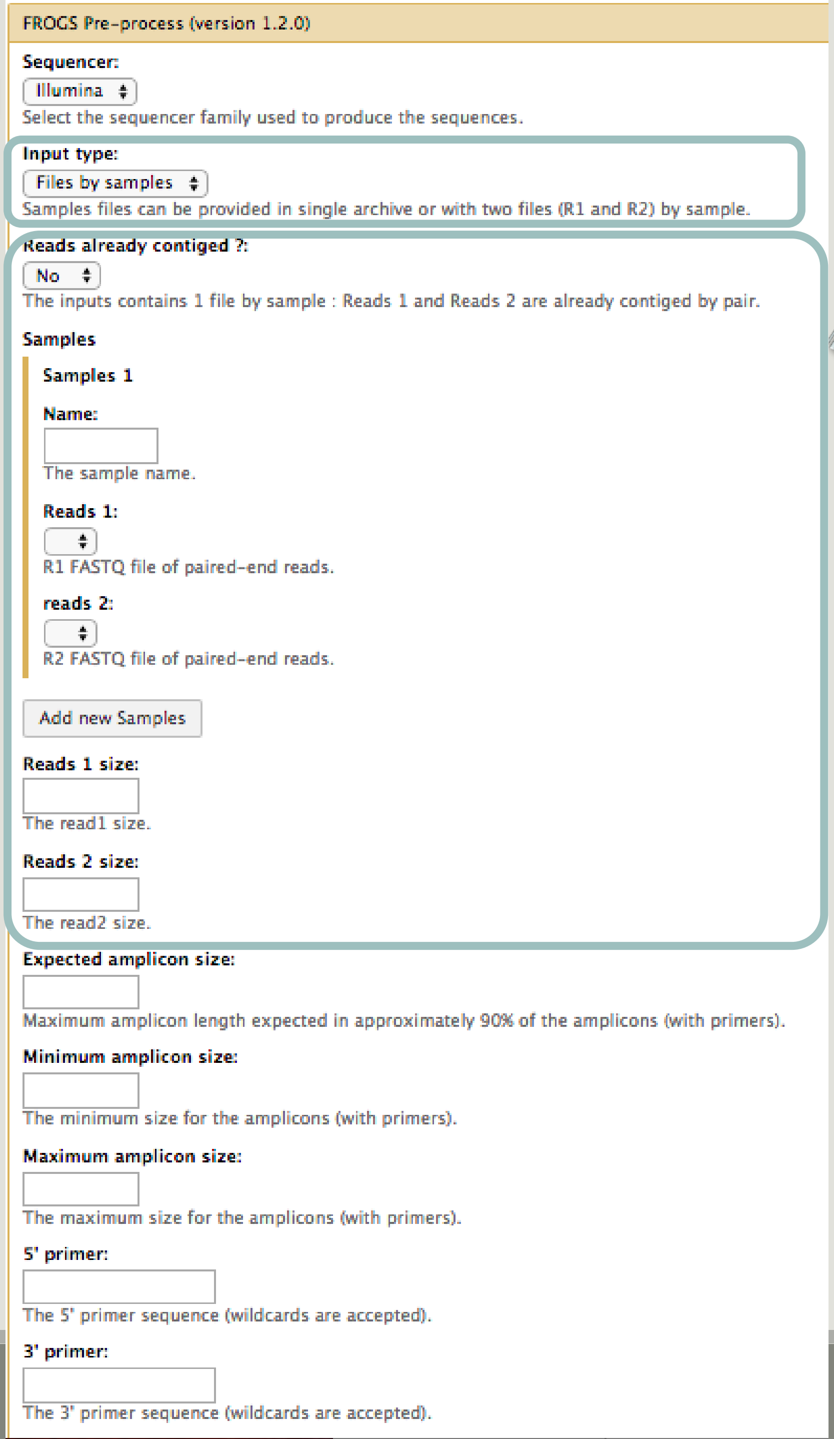
- Delete sequence with not expected lengths
- Delete sequences with ambiguous bases (N)
- Delete sequences do not contain good primers
- Dereplication

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

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

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

Pre-process



FROGS Pre-process (version 1.2.0)

Sequencer:
Illumina

Select the sequencer family used to produce the sequences.

Input type:
Files by samples

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

Reads already contiged ?:
No

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

Samples

Samples 1

Name:

The sample name.

Reads 1:

R1 FASTQ file of paired-end reads.

reads 2:

R2 FASTQ file of paired-end reads.

Add new Samples

Reads 1 size:

The read1 size.

Reads 2 size:

The read2 size.

Expected amplicon size:

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

Minimum amplicon size:

The minimum size for the amplicons (with primers).

Maximum amplicon size:

The maximum size for the amplicons (with primers).

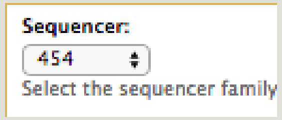
5' primer:

The 5' primer sequence (wildcards are accepted).

3' primer:

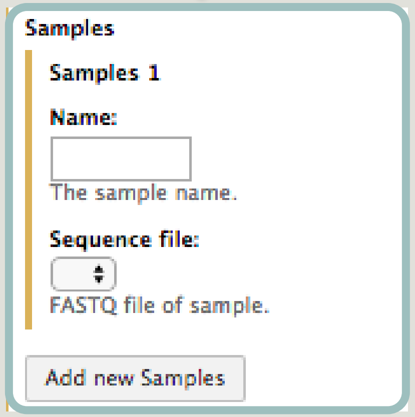
The 3' primer sequence (wildcards are accepted).

OR



Sequencer:
454

Select the sequencer family



Samples

Samples 1

Name:

The sample name.

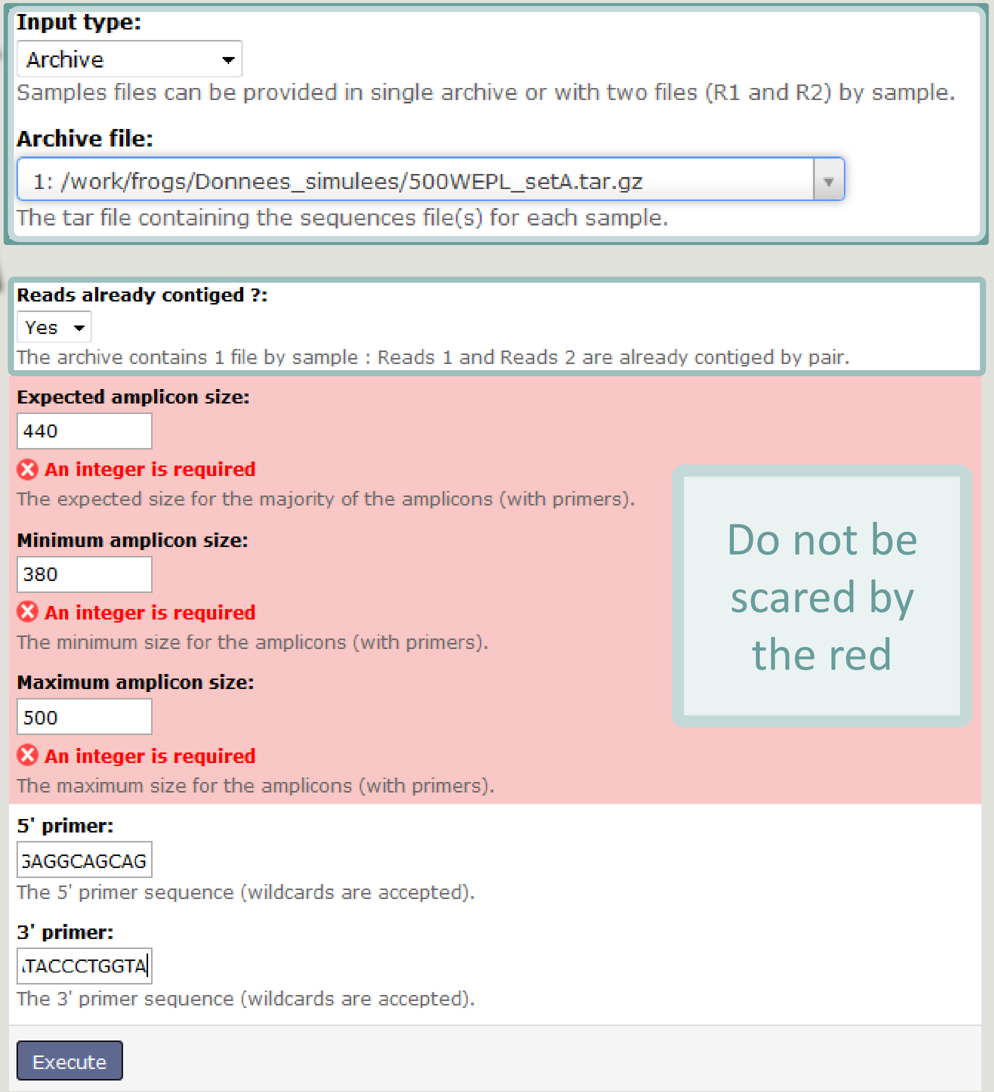
Sequence file:

FASTQ file of sample.

Add new Samples

OR

OR



Input type:
Archive

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

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

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

Reads already contiged ?:
Yes

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

Expected amplicon size:
440

ⓧ An integer is required

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

Minimum amplicon size:
380

ⓧ An integer is required

The minimum size for the amplicons (with primers).

Maximum amplicon size:
500

ⓧ An integer is required

The maximum size for the amplicons (with primers).

5' primer:
3AGGCAGCAG

The 5' primer sequence (wildcards are accepted).

3' primer:
.TACCCTGGTA

The 3' primer sequence (wildcards are accepted).

Execute

Do not be scared by the red

A vous de jouer ! - 3

GO TO EXERCISE 3

MiSeq
R1 R2

Flash, how it work ?

To contig read1 and read2 with FLASh with :

a minimum overlap equal to

[(R1-size + R2-size) - expected-amplicon-size]

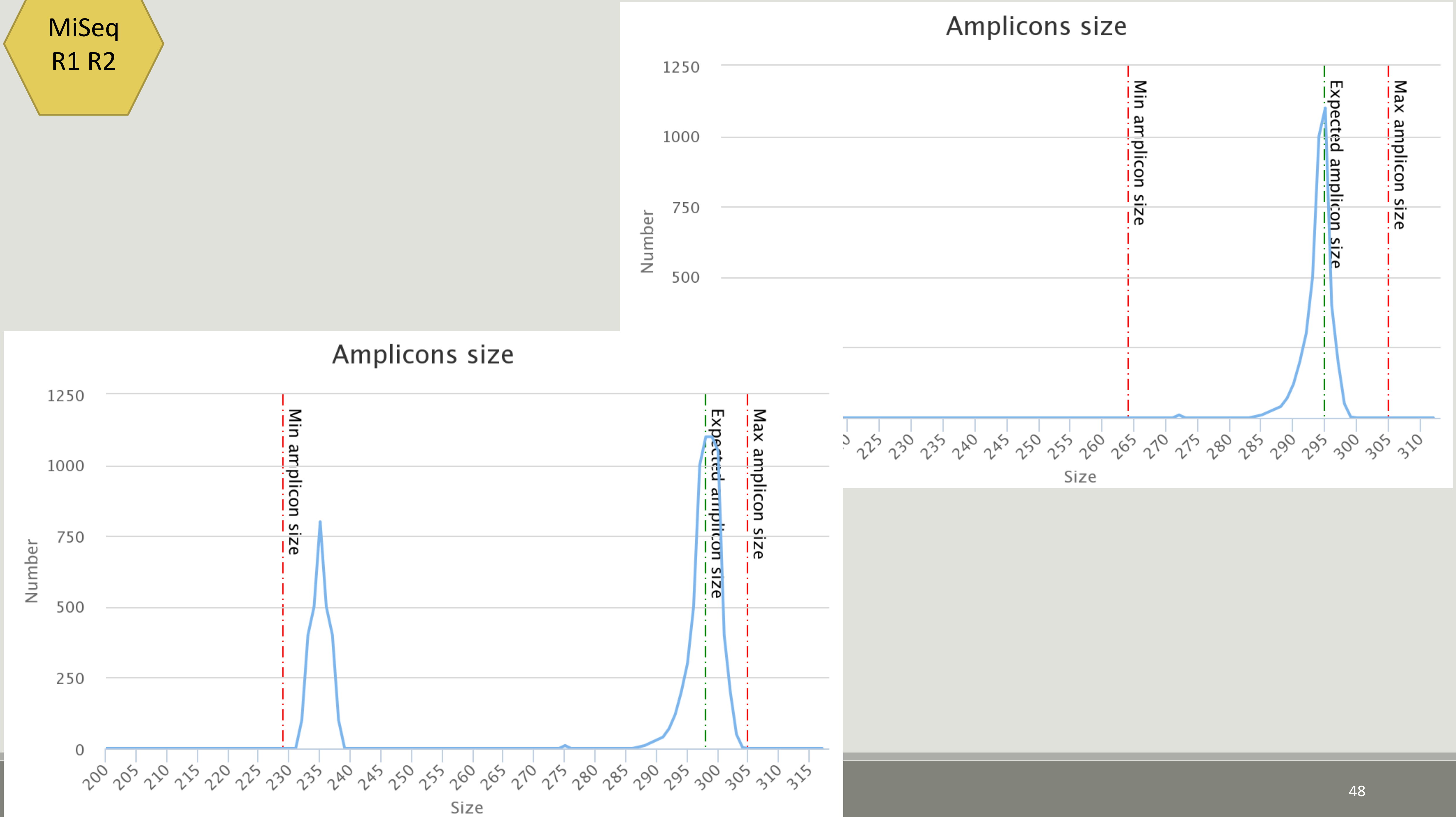
ex: (250+250) - 450 = 50

and a maximum overlap equal to

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

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

MiSeq
R1 R2


Amplicons size
1250
1000
750
500
Number
Min amplicon size
Expected amplicon size
Max amplicon size
225
230
235
240
245
250
255
260
265
270
275
280
285
290
295
300
305
310
Size
Amplicons size
1250
1000
750
500
250
0
Number
Min amplicon size
Expected amplicon size
Max amplicon size
200
205
210
215
220
225
230
235
240
245
250
255
260
265
270
275
280
285
290
295
300
305
310
315
Size

MiSeq
R1 R2

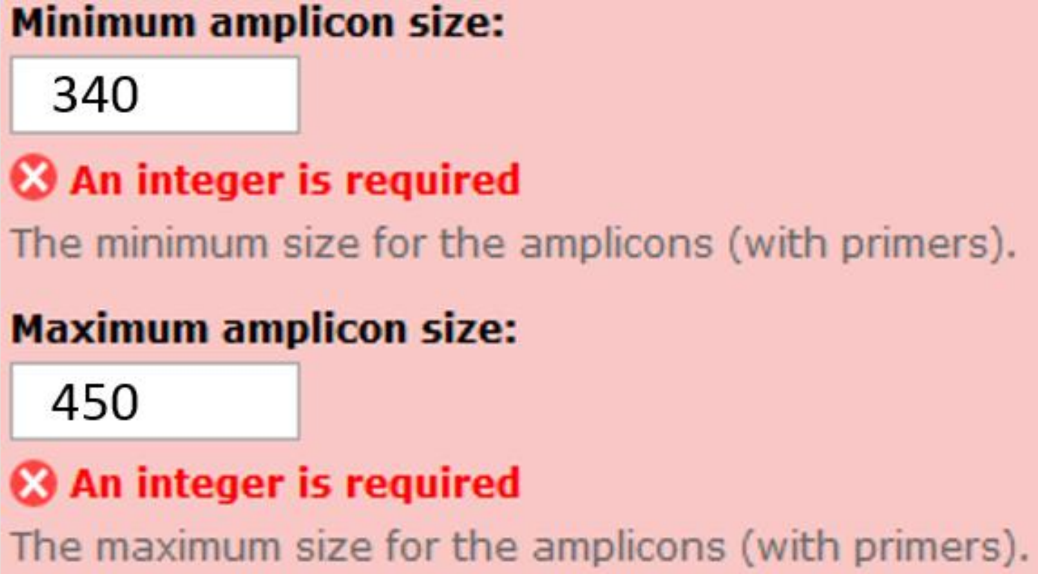
Cleaning, how it work ?

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

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

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

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



Minimum amplicon size:

340

⊗ **An integer is required**

The minimum size for the amplicons (with primers).

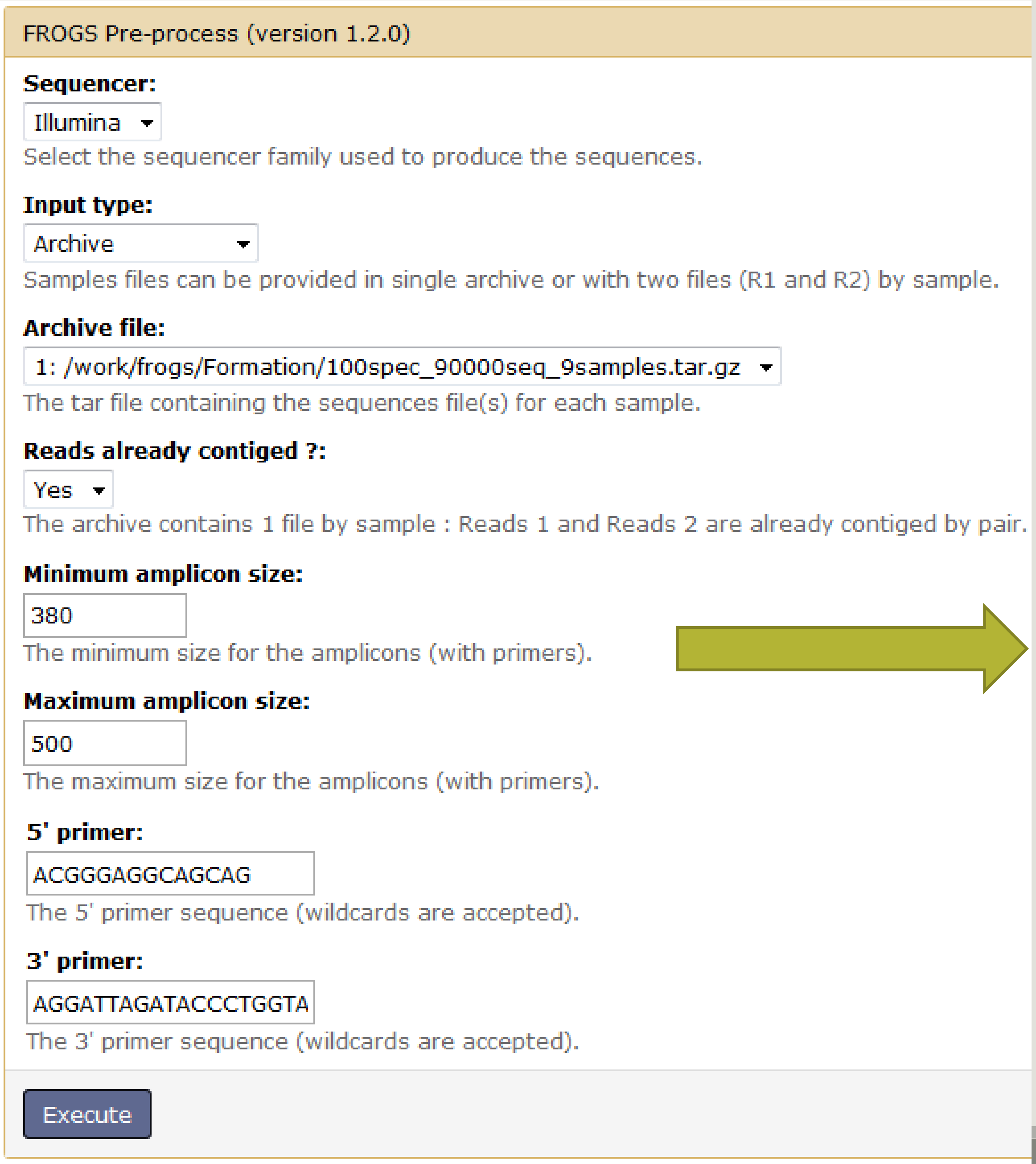
Maximum amplicon size:

450

⊗ **An integer is required**

The maximum size for the amplicons (with primers).

MiSeq contiged



FROGS Pre-process (version 1.2.0)

Sequencer:

Illumina

Select the sequencer family used to produce the sequences.

Input type:

Archive

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

Archive file:

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

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

Reads already contiged ?:

Yes

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

Minimum amplicon size:

380

The minimum size for the amplicons (with primers).

Maximum amplicon size:

500

The maximum size for the amplicons (with primers).

5' primer:

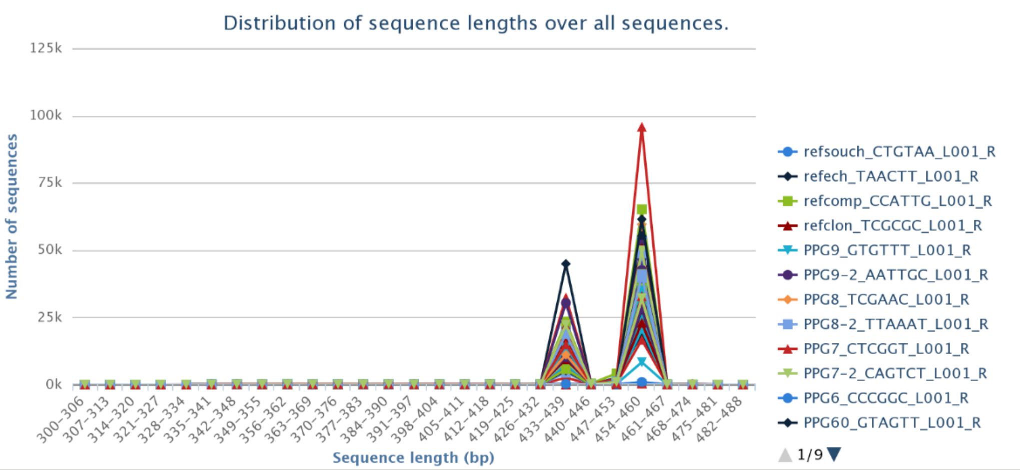
ACGGGAGGCAGCAG

The 5' primer sequence (wildcards are accepted).

3' primer:

AGGATTAGATACCCTGGTA

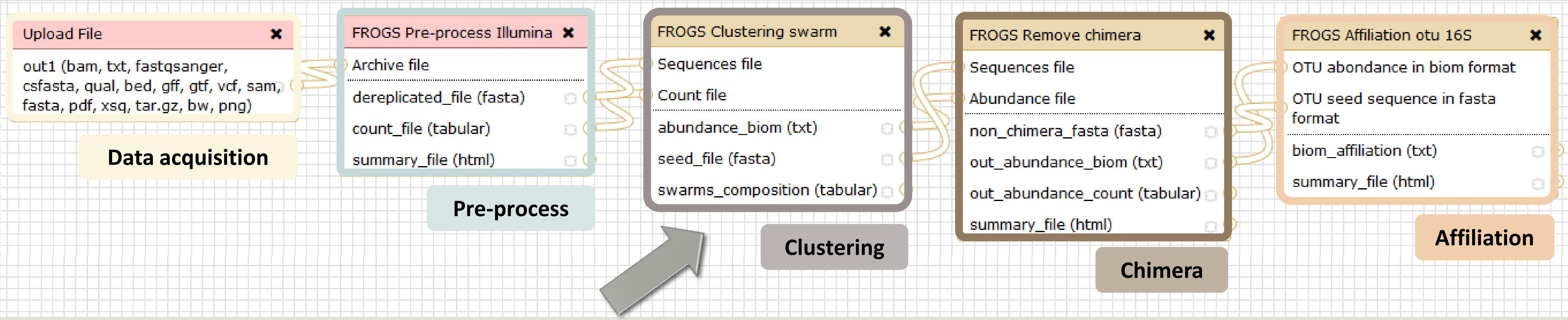
The 3' primer sequence (wildcards are accepted).



Distribution of sequence lengths over all sequences.

Clustering tool

FROGS pipeline



Upload File ✖

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

FROGS Pre-process Illumina ✖

Archive file

dereplicated_file (fasta)

count_file (tabular)

summary_file (html)

FROGS Clustering swarm ✖

Sequences file

Count file

abundance_biom (txt)

seed_file (fasta)

swarms_composition (tabular)

FROGS Remove chimera ✖

Sequences file

Abundance file

non_chimera_fasta (fasta)

out_abundance_biom (txt)

out_abundance_count (tabular)

summary_file (html)

FROGS Affiliation otu 16S ✖

OTU abondance in biom format

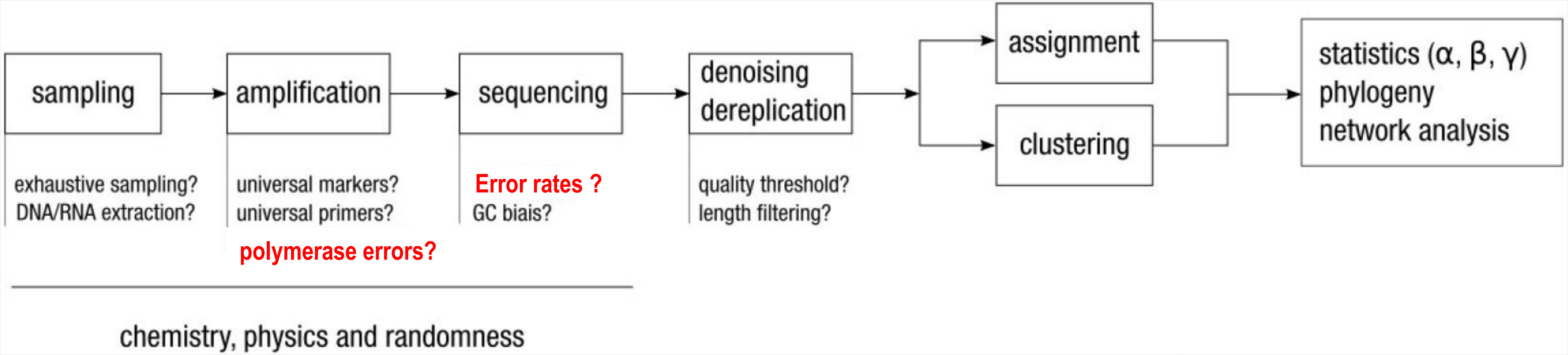
OTU seed sequence in fasta format

biom_affiliation (txt)

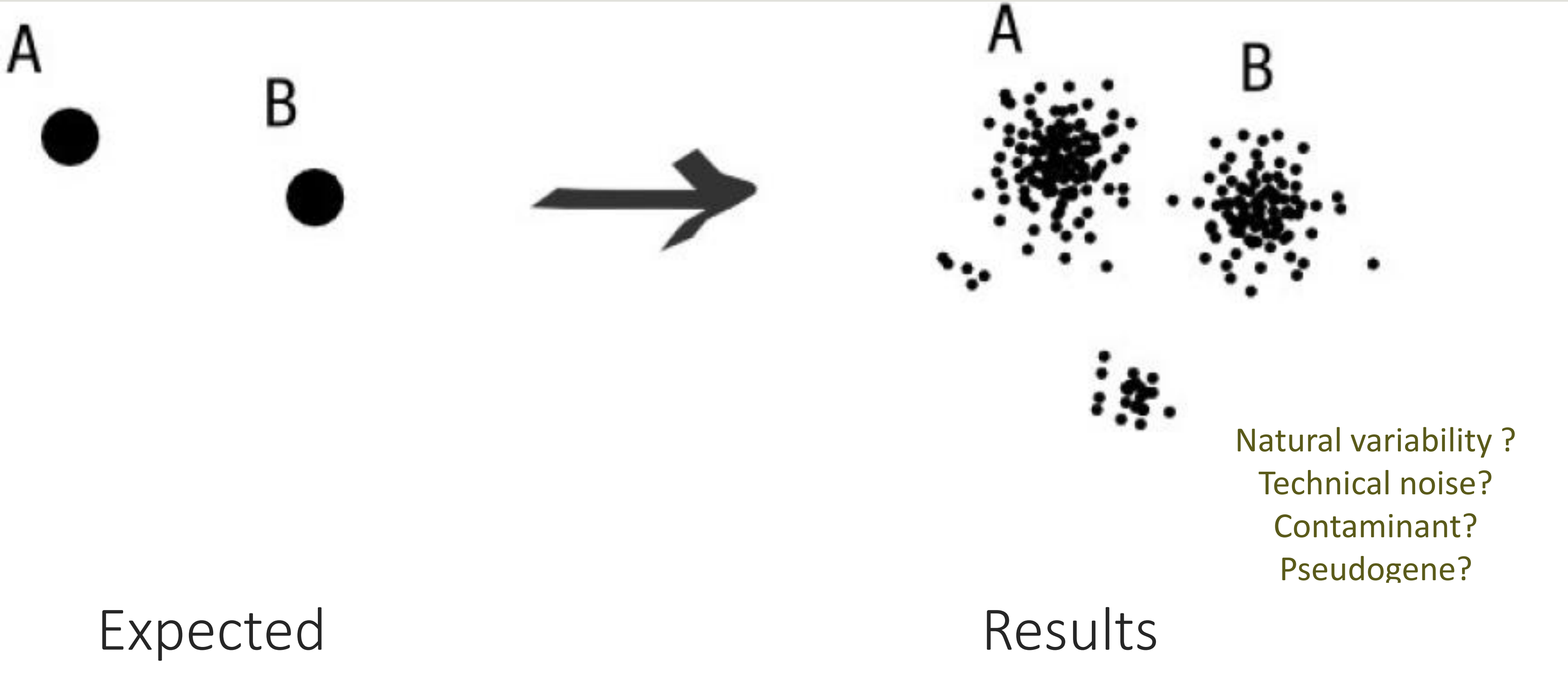
summary_file (html)

Why do we need clustering ?

Amplication and sequencing and are not perfect processes



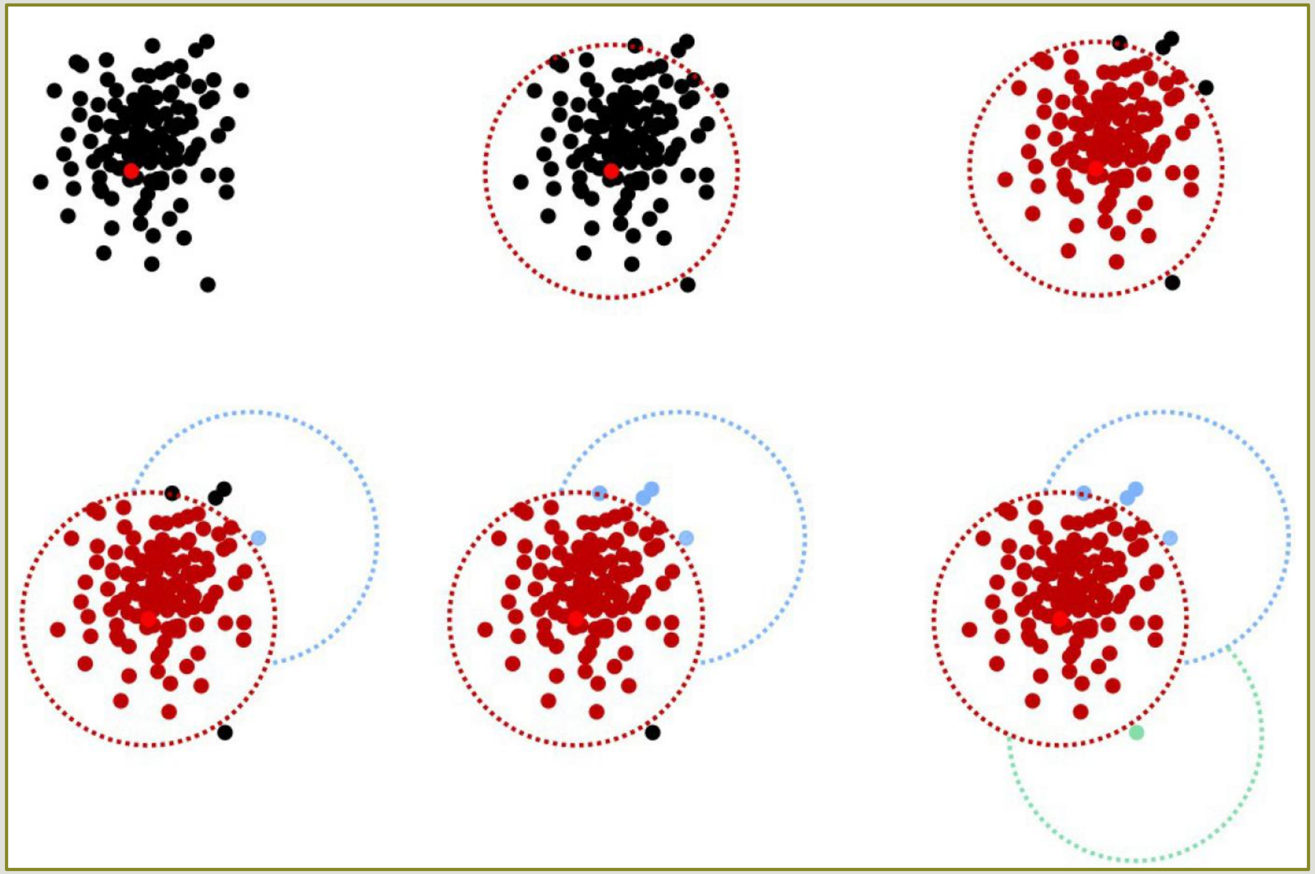
Fréderic Mahé communication



Natural variability ?
Technical noise?
Contaminant?
Pseudogene?

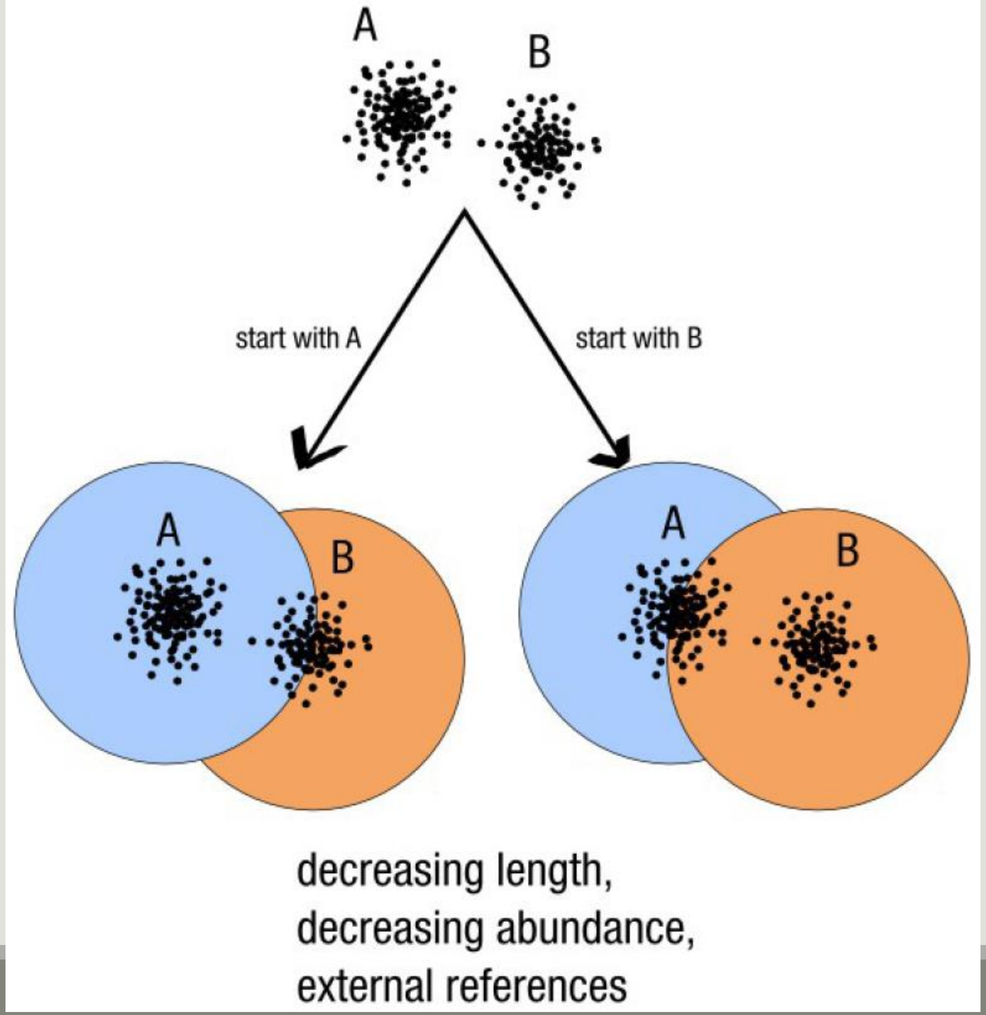
Fréderic Mahé communication

How traditional clustering works ?



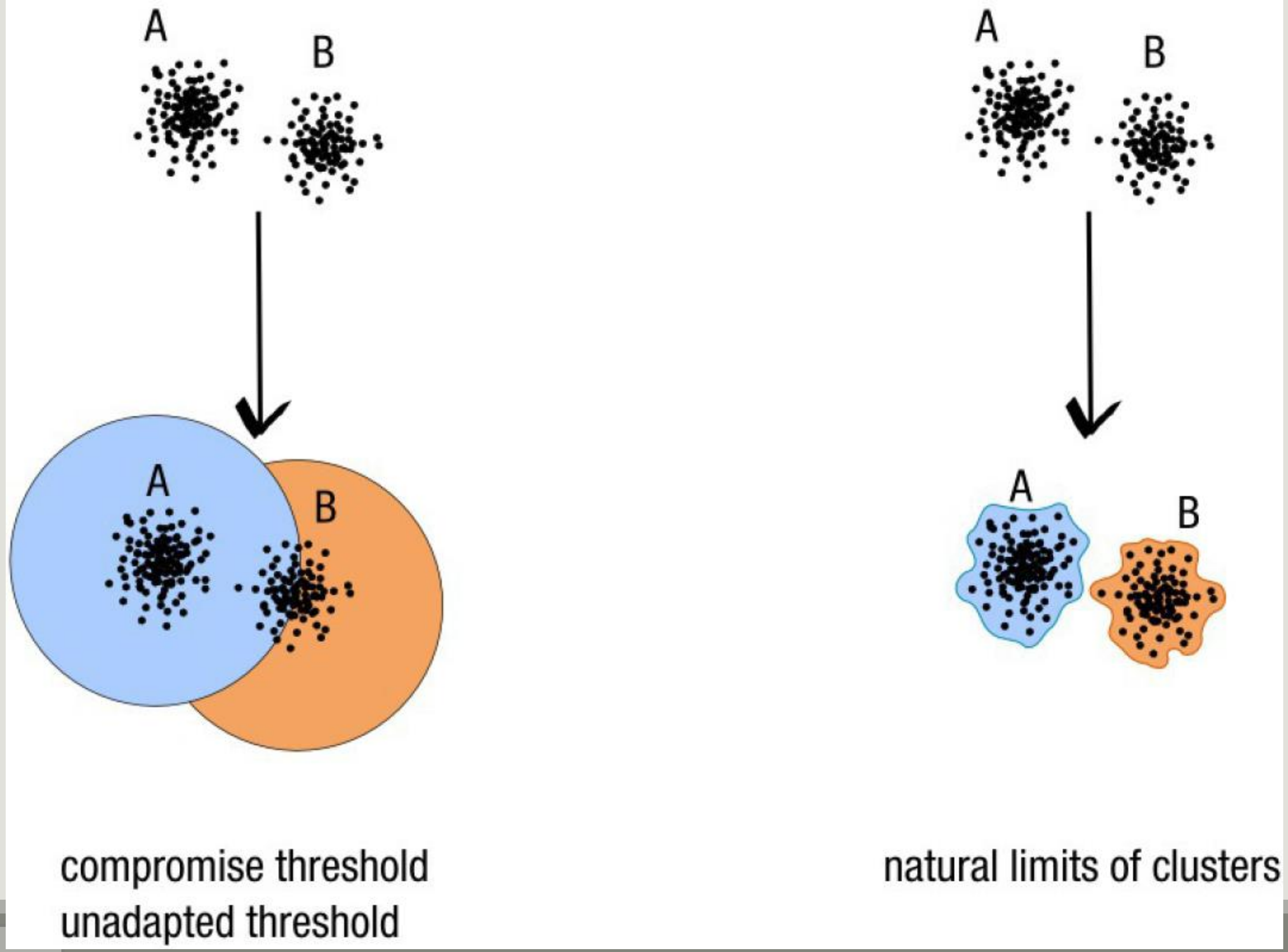
Fréderic Mahé communication

Input order dependent results



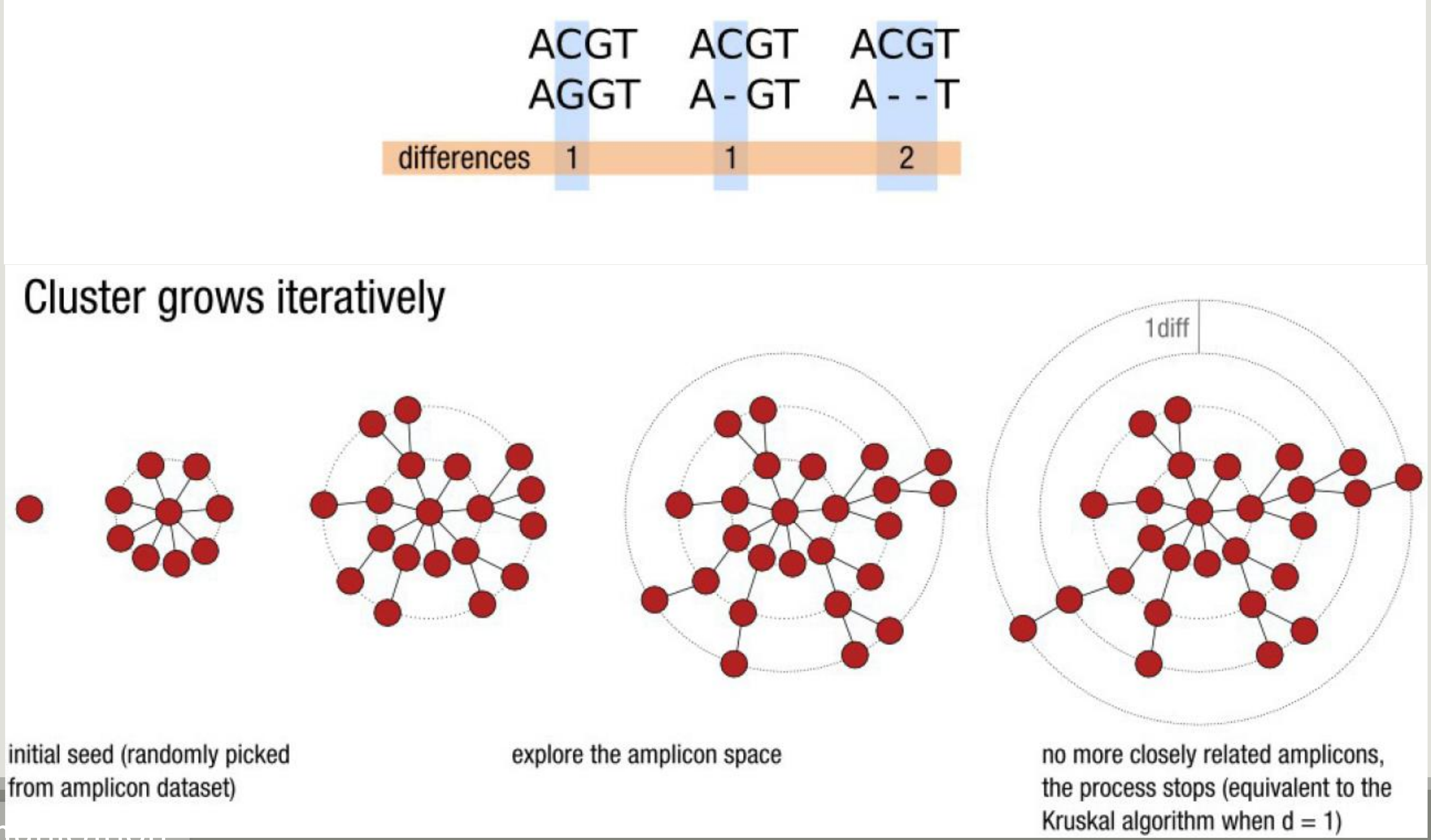
Fréderic Mahé communication

Single a priori clustering threshold



Fréderic Mahé communication

Swarm clustering method



Cluster grows iteratively

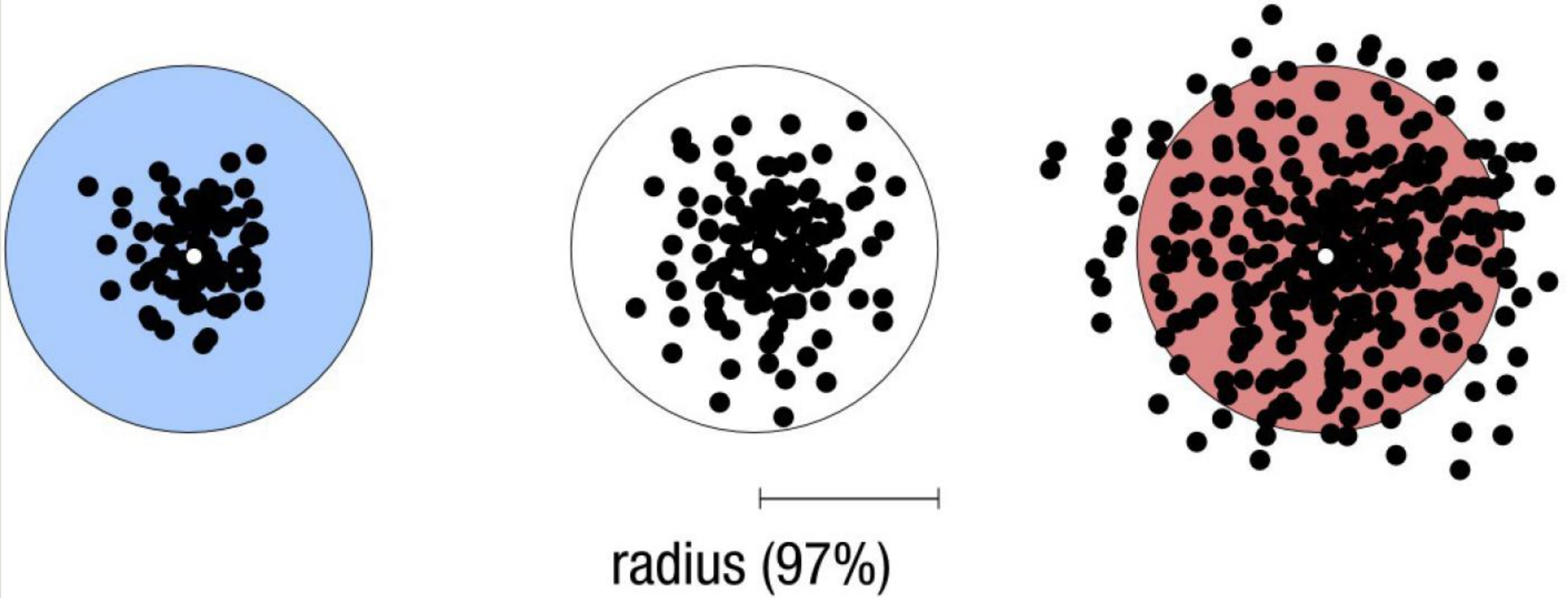
initial seed (randomly picked from amplicon dataset)

explore the amplicon space

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

Fréderic Mahé communication

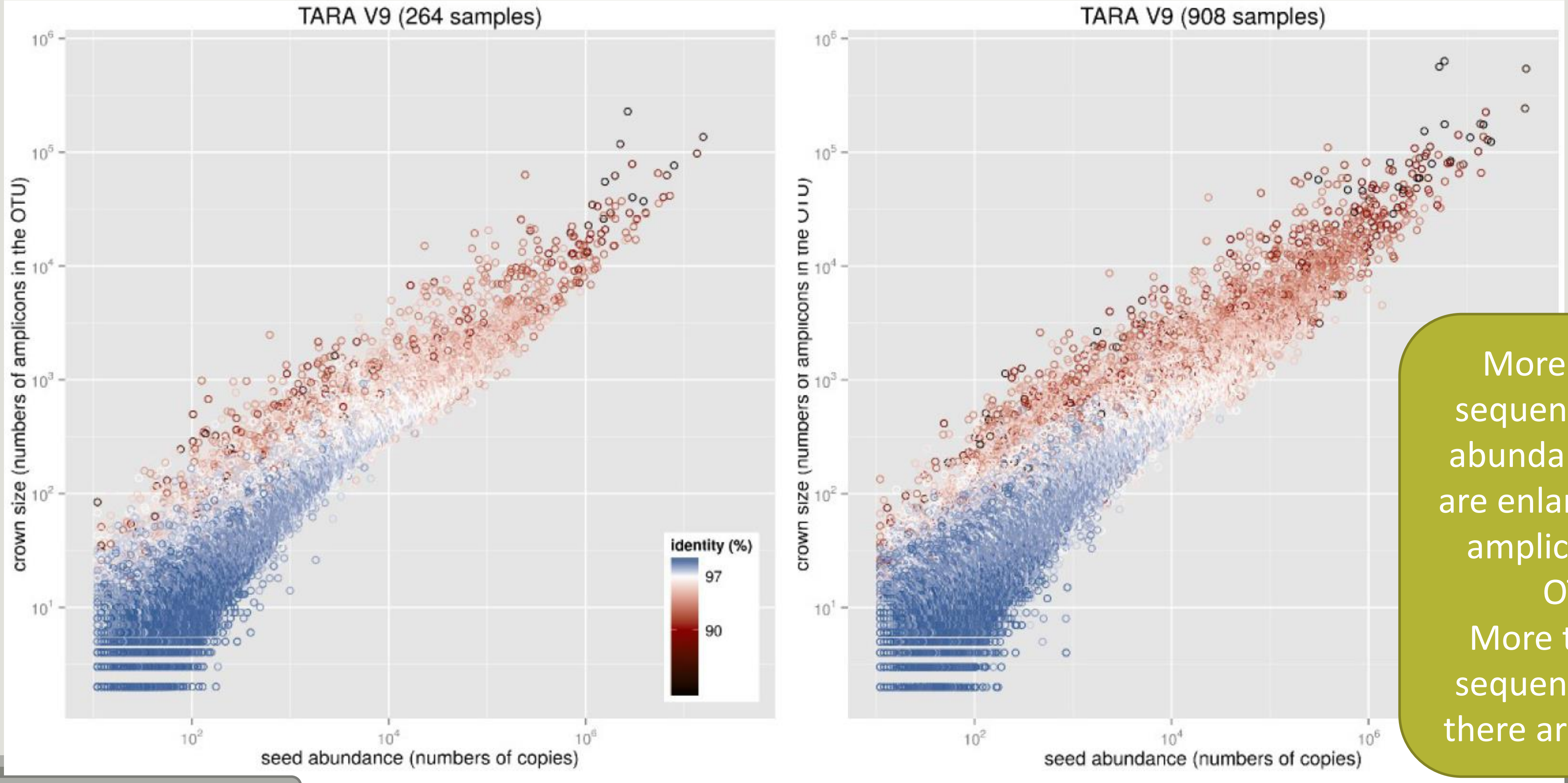
Comparison Swarm and 3% clusterings



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

Fréderic Mahé communication

Comparison Swarm and 3% clusterings



clusters produced with swarm using d = 1

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

Fréderic Mahé communication

SWARM

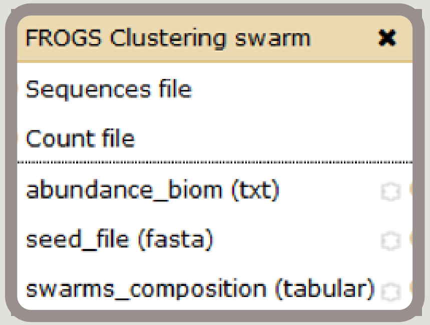
A robust and fast clustering method for amplicon-based studies.

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

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

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

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



FROGS Clustering swarm ✖

Sequences file

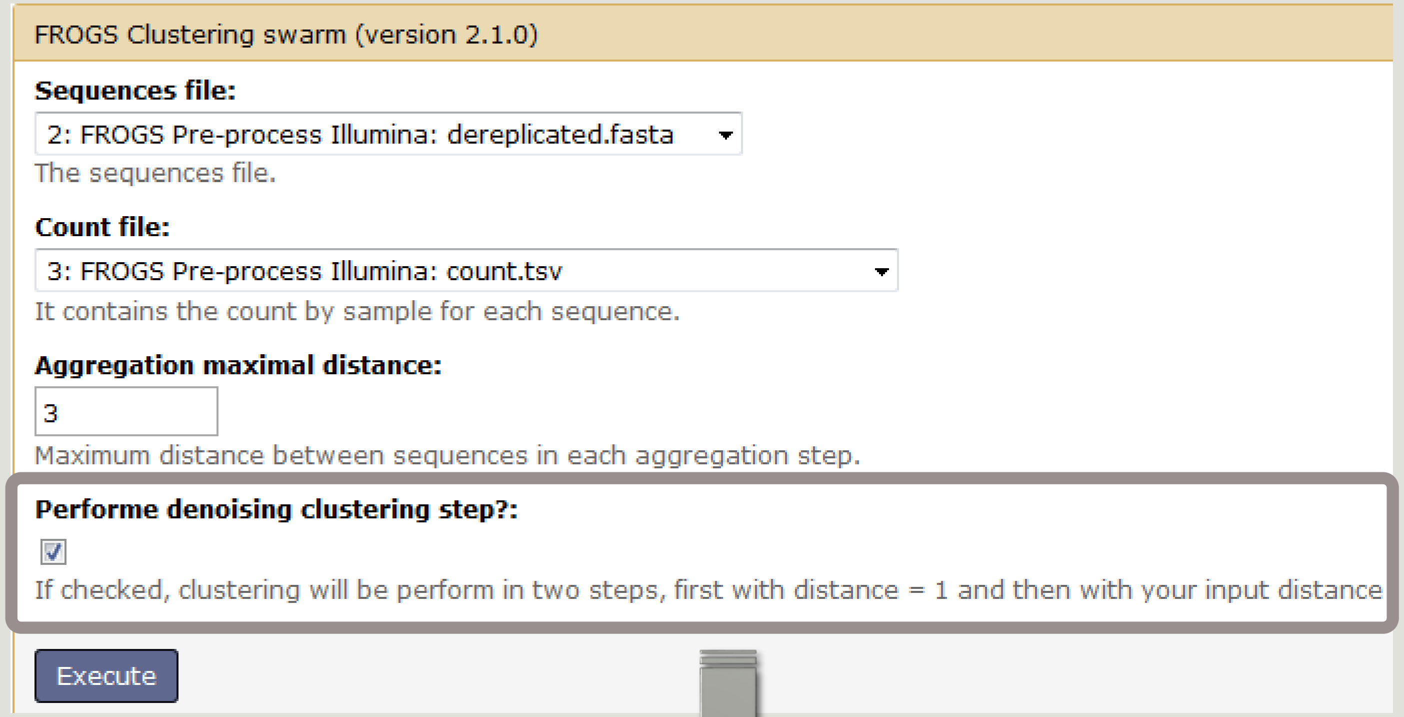
Count file

abundance_biom (txt)

seed_file (fasta)

swarms_composition (tabular)

Clustering



FROGS Clustering swarm (version 2.1.0)

Sequences file:

2: FROGS Pre-process Illumina: dereplicated.fasta

The sequences file.

Count file:

3: FROGS Pre-process Illumina: count.tsv

It contains the count by sample for each sequence.

Aggregation maximal distance:

3

Maximum distance between sequences in each aggregation step.

Performe denoising clustering step?:

☑

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

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

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

Gain time !

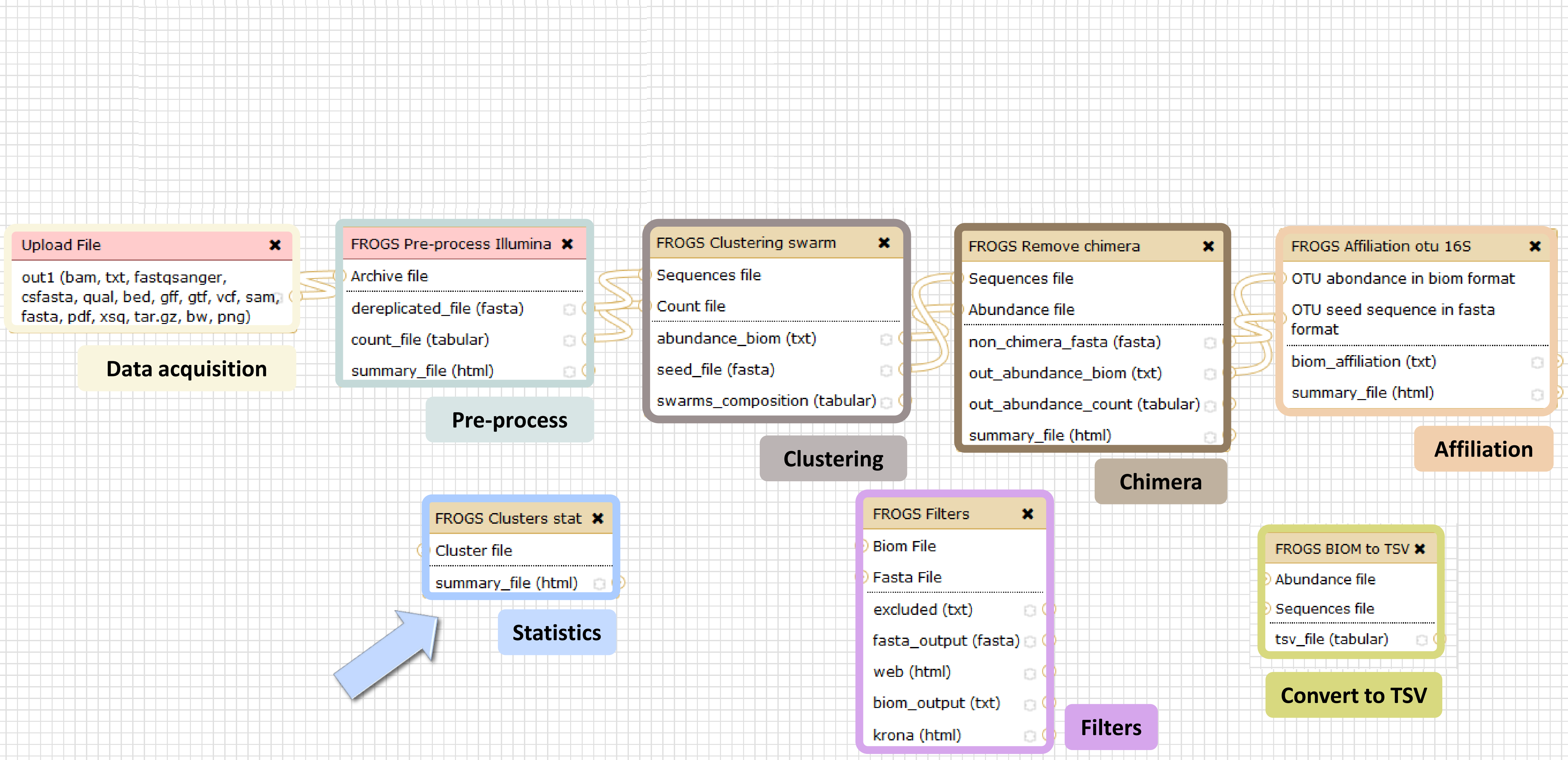
Remove false positives !

A vous de jouer ! - 4

EXERCISE 4

Cluster stat tool

SOME SLIDES TO KEEP EXPLANATIONS IN THE MEMORY



Data acquisition

Pre-process

Clustering

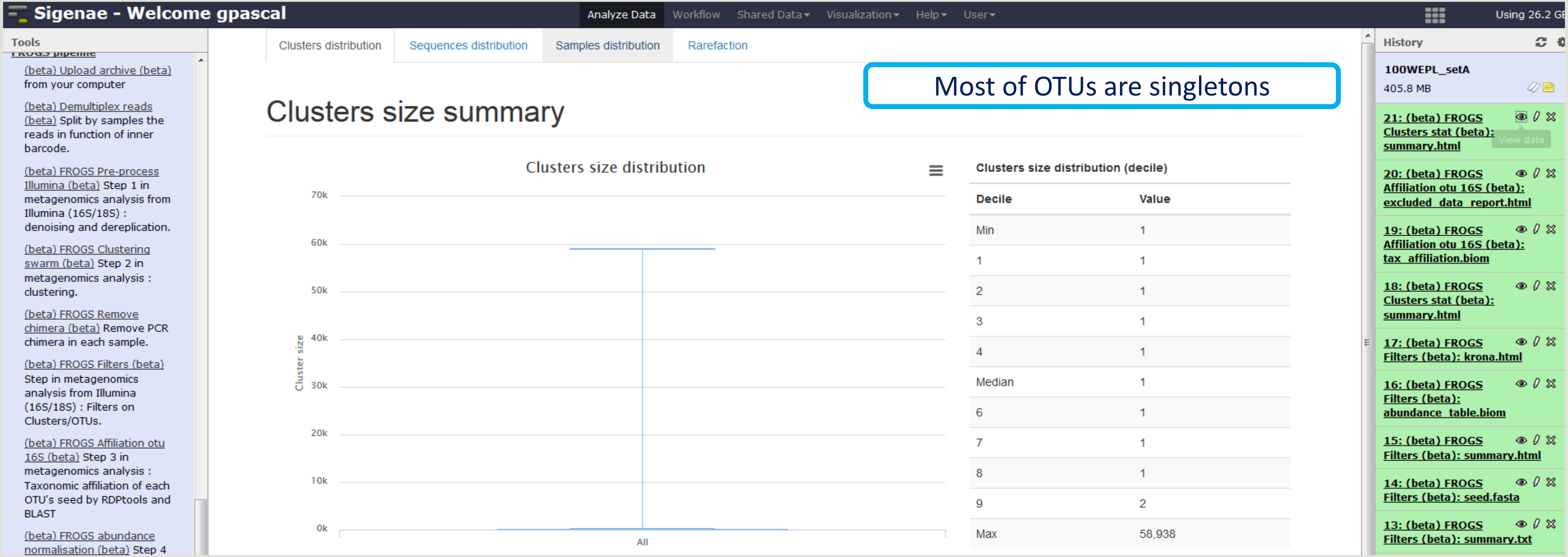
Chimera

Affiliation

Statistics

Filters

Convert to TSV



(beta) Upload archive (beta) from your computer

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

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

(beta) FROGS Clustering swarm (beta) Step 2 in metagenomics analysis : clustering.

(beta) FROGS Remove chimera (beta) Remove PCR chimera in each sample.

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

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

(beta) FROGS abundance normalisation (beta) Step 4

Clusters distribution | Sequences distribution | Samples distribution | Rarefaction

Most of OTUs are singletons

Clusters size summary

Clusters size distribution

Clusters size distribution (decile)

Decile	Value
Min	1
1	1
2	1
3	1
4	1
Median	1
6	1
7	1
8	1
9	2
Max	58,938

100WEPL_setA
405.8 MB

21: (beta) FROGS Clusters stat (beta): summary.html

20: (beta) FROGS Affiliation otu 16S (beta): excluded_data_report.html

19: (beta) FROGS Affiliation otu 16S (beta): tax_affiliation.biom

18: (beta) FROGS Clusters stat (beta): summary.html

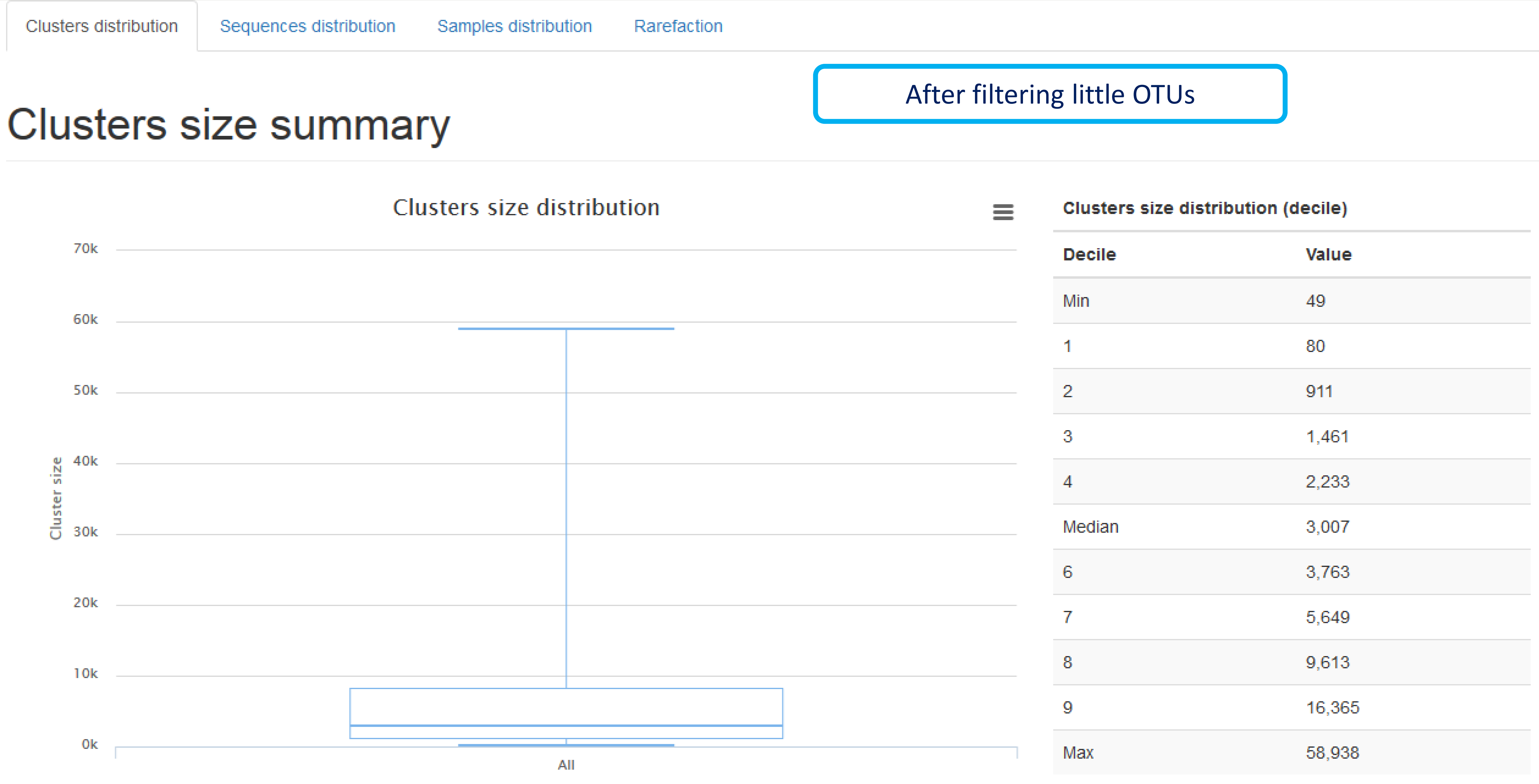
17: (beta) FROGS Filters (beta): krona.html

16: (beta) FROGS Filters (beta): abundance_table.biom

15: (beta) FROGS Filters (beta): summary.html

14: (beta) FROGS Filters (beta): seed.fasta

13: (beta) FROGS Filters (beta): summary.txt

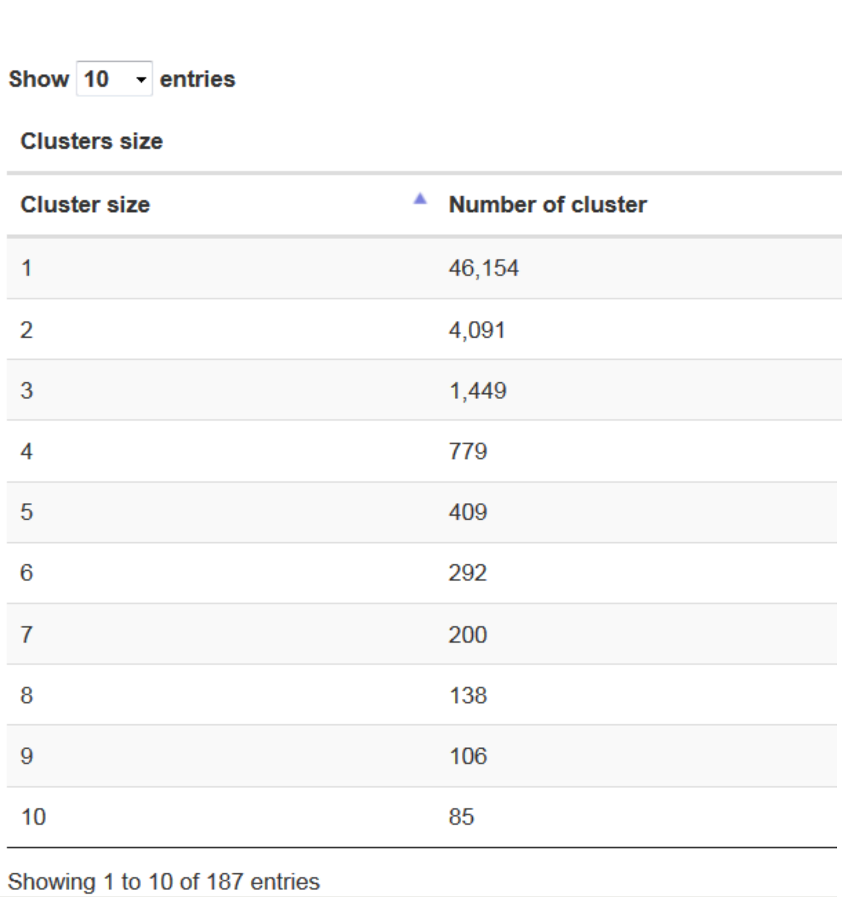


After filtering little OTUs

Clusters size summary

Clusters size distribution (decile)

Decile	Value
Min	49
1	80
2	911
3	1,461
4	2,233
Median	3,007
6	3,763
7	5,649
8	9,613
9	16,365
Max	58,938



Show 10 entries

Search:

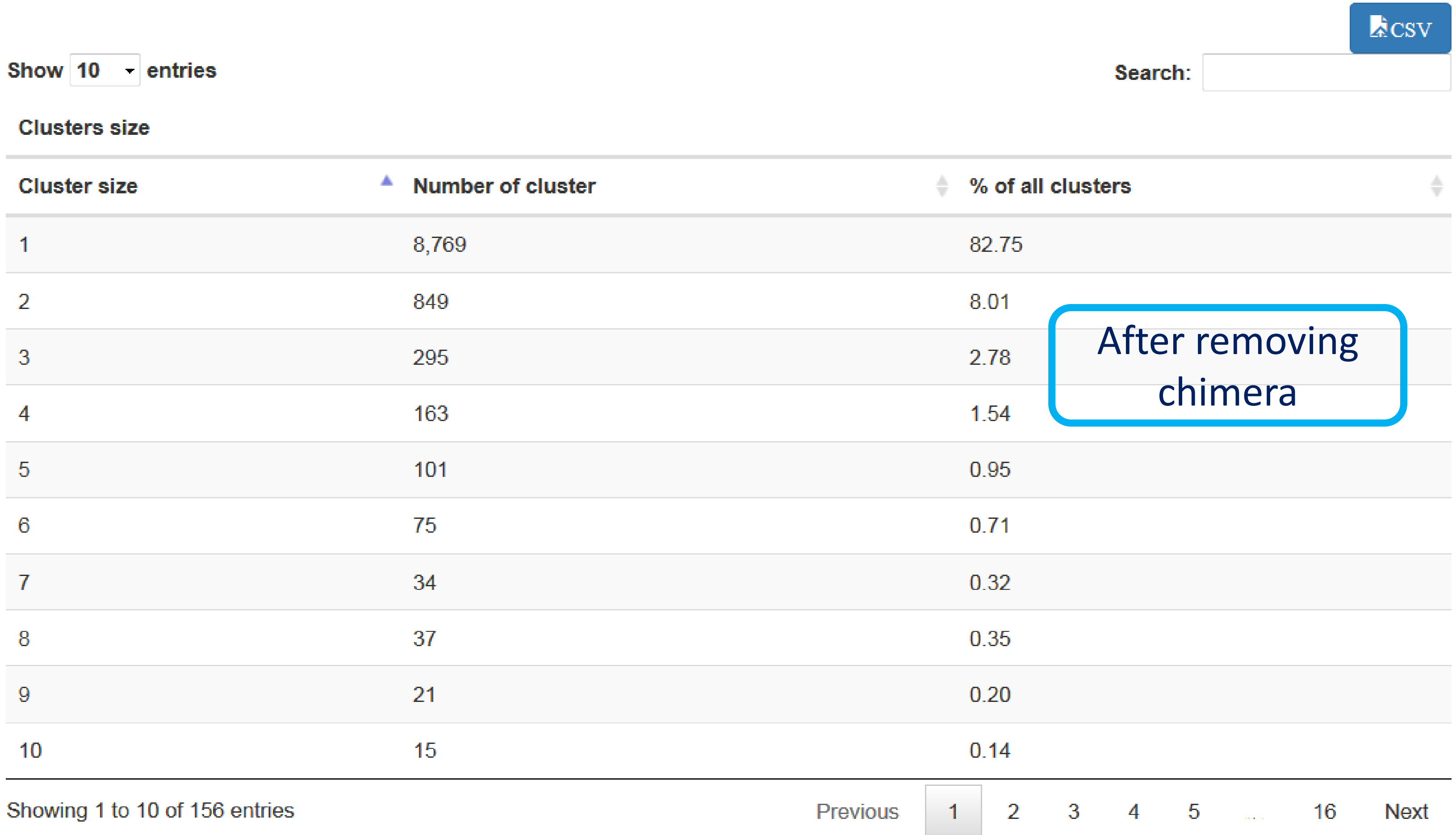
Clusters size

Cluster size	Number of cluster	% of all clusters
1	46,154	84.72
2	4,091	7.51
3	1,449	2.66
4	779	1.43
5	409	
6	292	
7	200	
8	138	
9	106	
10	85	

Showing 1 to 10 of 187 entries

Most of OTUs are singletons

After clustering



Show 10 entries

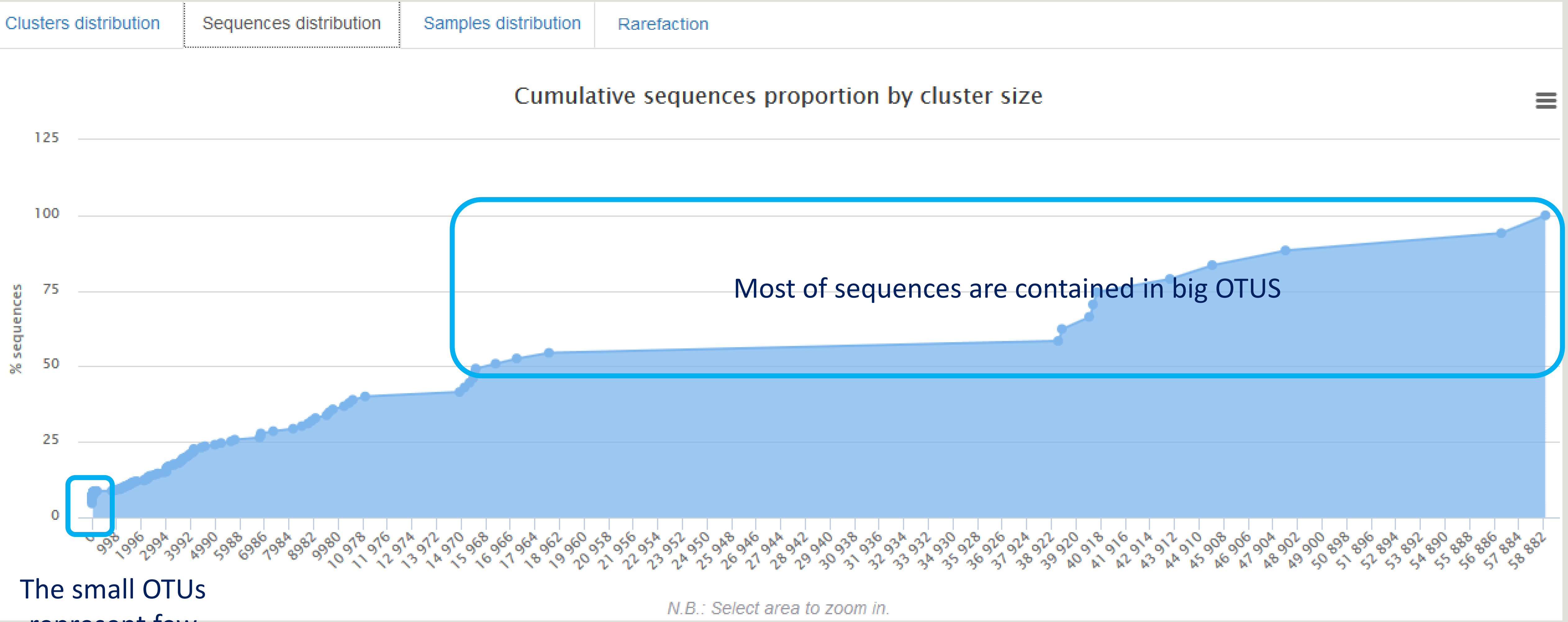
Search:

Clusters size

Cluster size	Number of cluster	% of all clusters
1	8,769	82.75
2	849	8.01
3	295	2.78
4	163	1.54
5	101	0.95
6	75	0.71
7	34	0.32
8	37	0.35
9	21	0.20
10	15	0.14

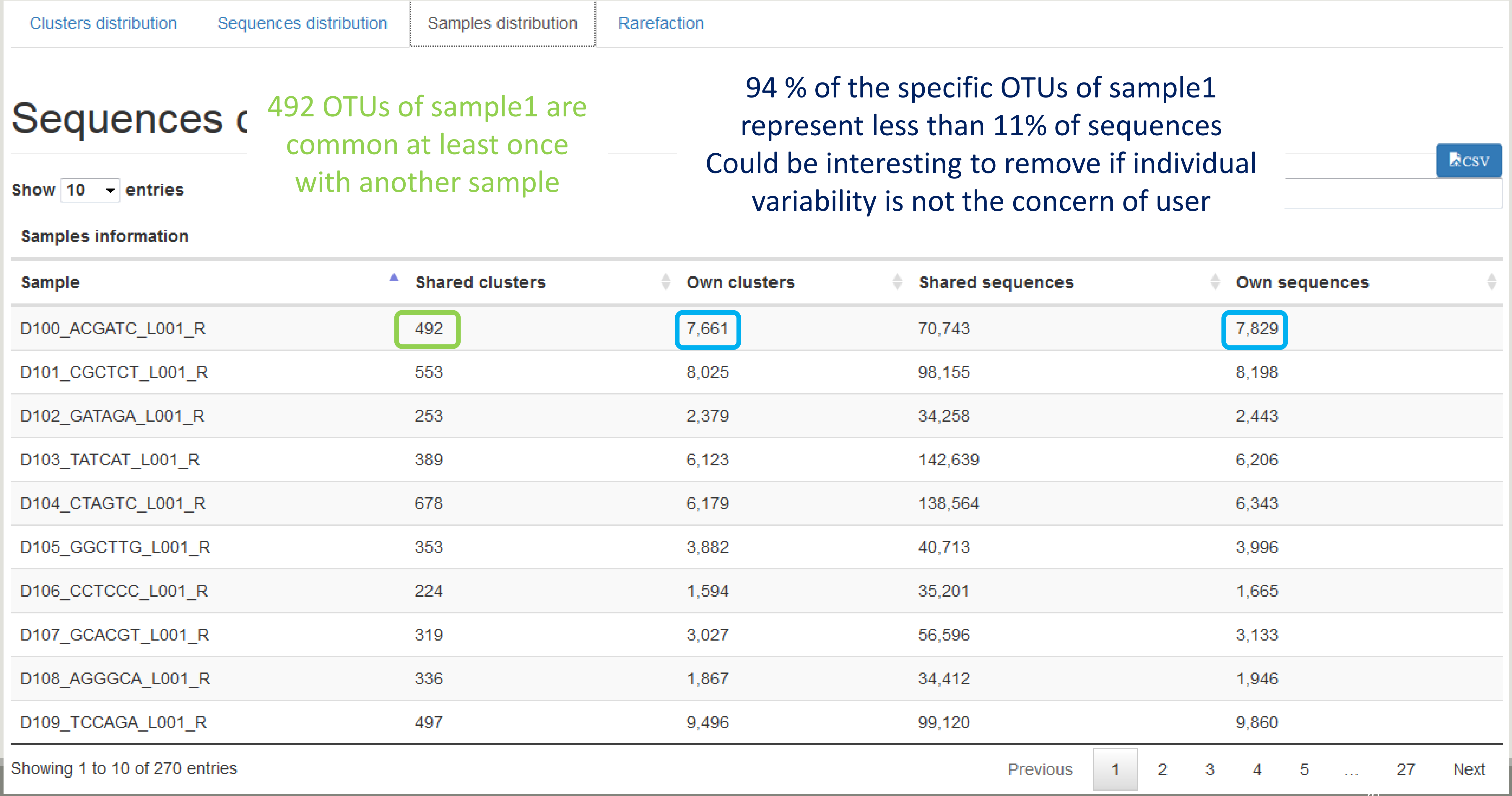
Showing 1 to 10 of 156 entries

After removing chimera



Most of sequences are contained in big OTUS

The small OTUs represent few sequences



Sequences d

492 OTUs of sample1 are common at least once with another sample

94 % of the specific OTUs of sample1 represent less than 11% of sequences
Could be interesting to remove if individual variability is not the concern of user

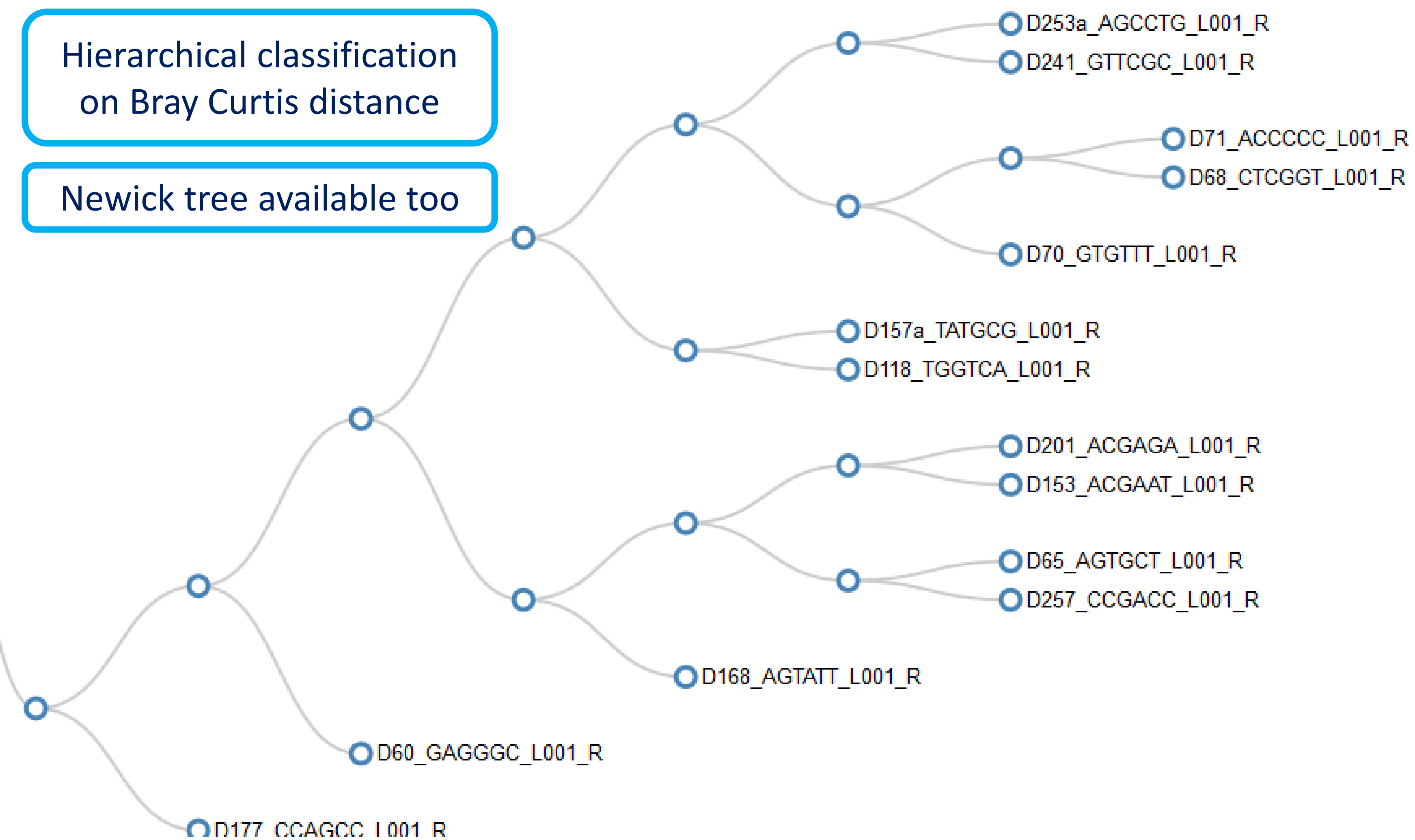
Samples information

Sample	Shared clusters	Own clusters	Shared sequences	Own sequences
D100_ACGATC_L001_R	492	7,661	70,743	7,829
D101_CGCTCT_L001_R	553	8,025	98,155	8,198
D102_GATAGA_L001_R	253	2,379	34,258	2,443
D103_TATCAT_L001_R	389	6,123	142,639	6,206
D104_CTAGTC_L001_R	678	6,179	138,564	6,343
D105_GGCTTG_L001_R	353	3,882	40,713	3,996
D106_CCTCCC_L001_R	224	1,594	35,201	1,665
D107_GCACGT_L001_R	319	3,027	56,596	3,133
D108_AGGGCA_L001_R	336	1,867	34,412	1,946
D109_TCCAGA_L001_R	497	9,496	99,120	9,860

Showing 1 to 10 of 270 entries

Hierarchical classification on Bray Curtis distance

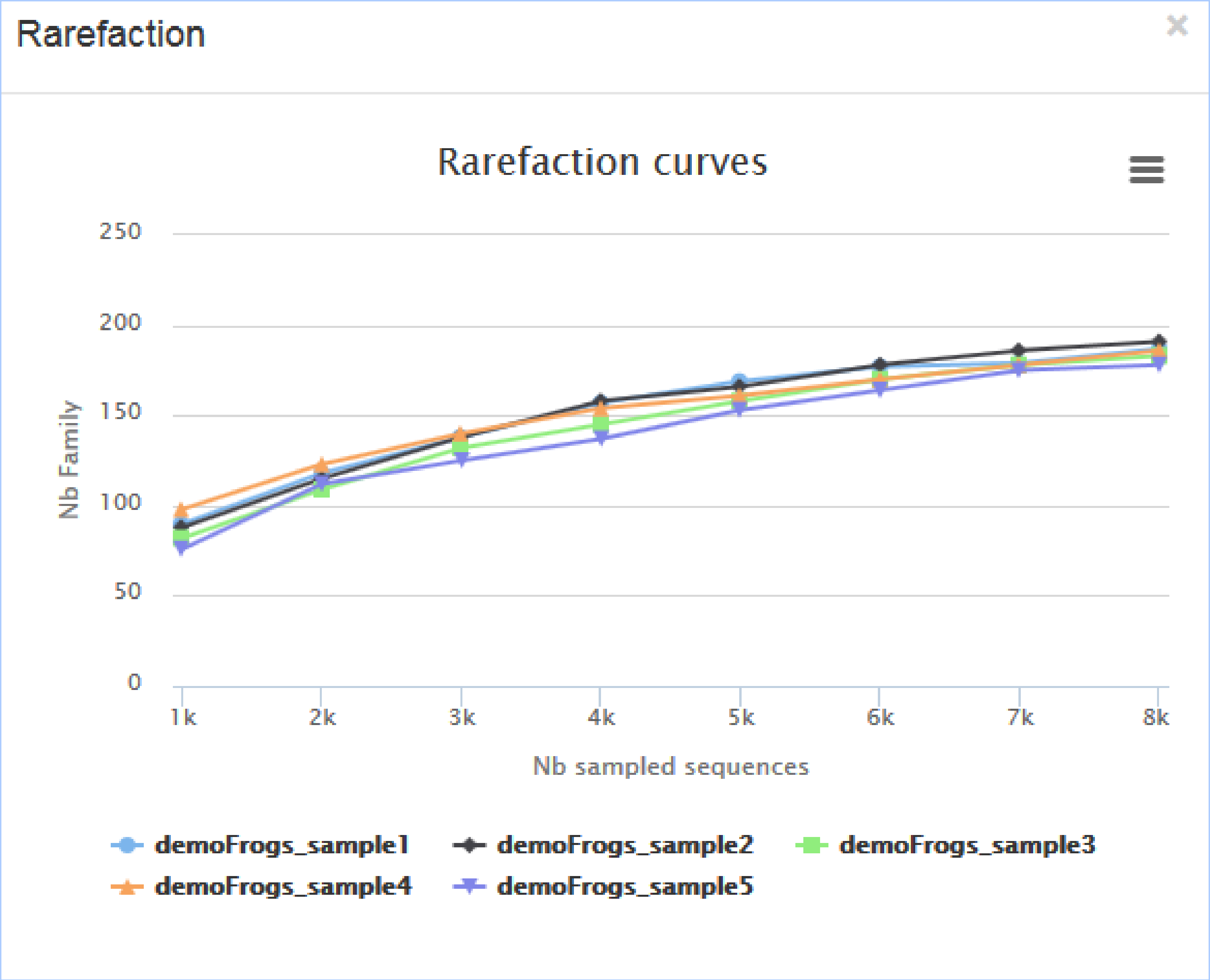
Newick tree available too



Samples distribution tab

Available only after AFFILIATION TOOL

Samples size ~8500 sequences



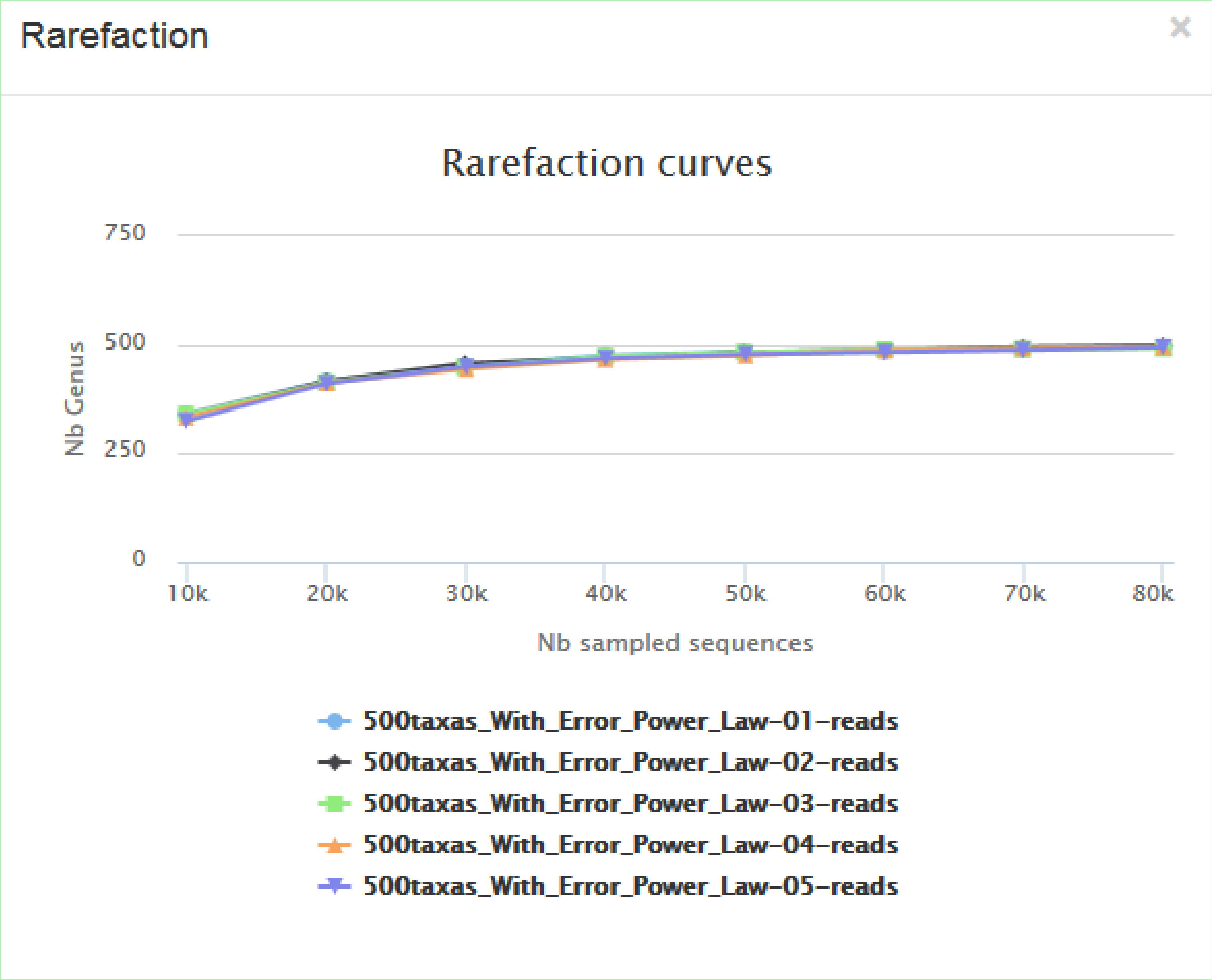
The curve continues to rise

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

Rarefaction tab

Available only after AFFILIATION TOOL

Samples size ~85 000 sequences

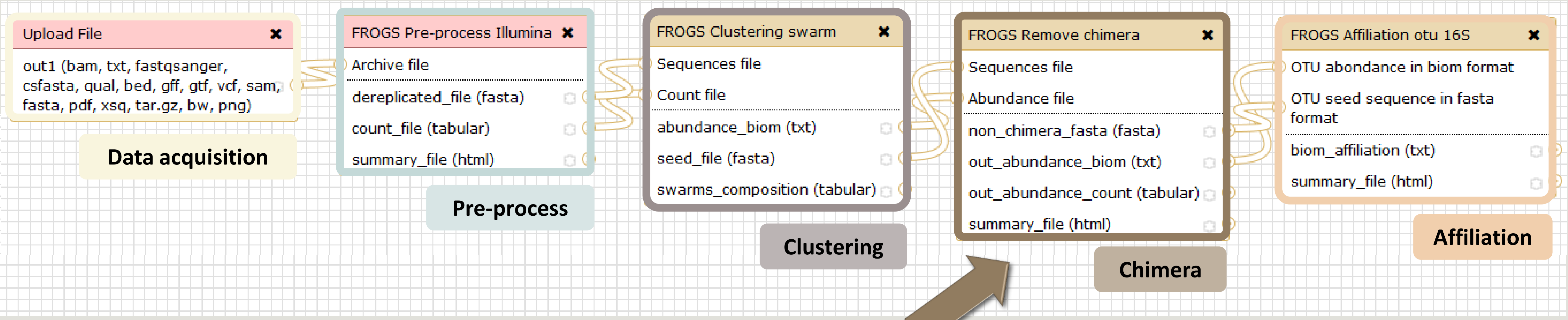


The curve slows to rise with ~50 000 sequences

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

Removing chimera tool

FROGS pipeline



Upload File ✖

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

Data acquisition

FROGS Pre-process Illumina ✖

Archive file

dereplicated_file (fasta)

count_file (tabular)

summary_file (html)

Pre-process

FROGS Clustering swarm ✖

Sequences file

Count file

abundance_biom (txt)

seed_file (fasta)

swarms_composition (tabular)

Clustering

FROGS Remove chimera ✖

Sequences file

Abundance file

non_chimera_fasta (fasta)

out_abundance_biom (txt)

out_abundance_count (tabular)

summary_file (html)

Chimera

FROGS Affiliation otu 16S ✖

OTU abondance in biom format

OTU seed sequence in fasta format

biom_affiliation (txt)

summary_file (html)

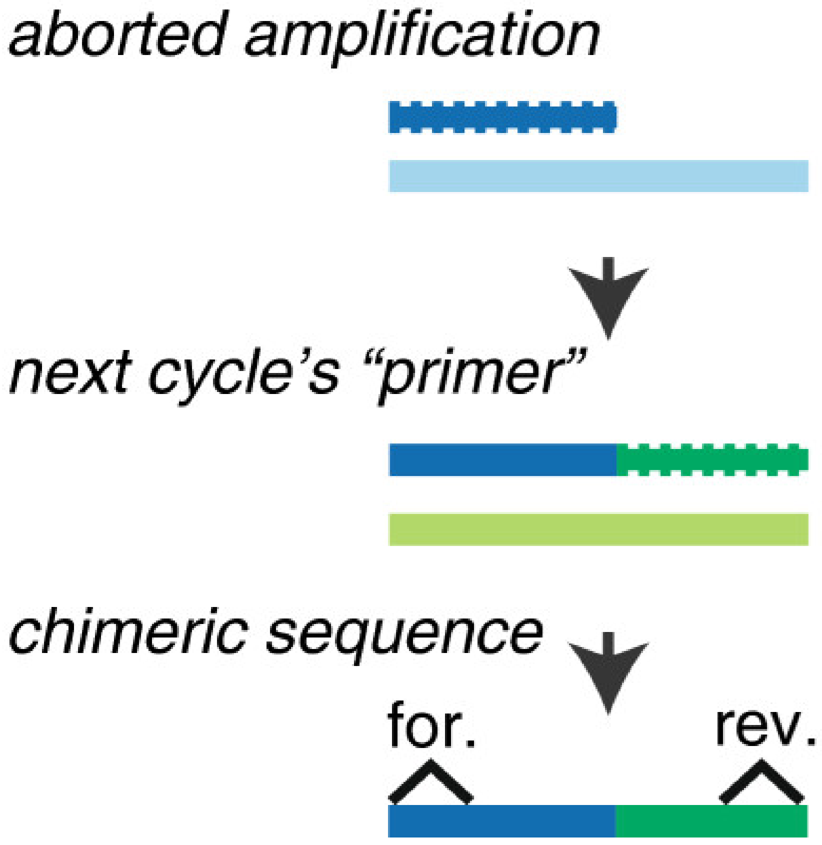
Affiliation

Our advice:
Removing Chimera after
Swarm denoising + Swarm d=3

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.

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



A vous de jouer ! - 5

EXERCISE 5

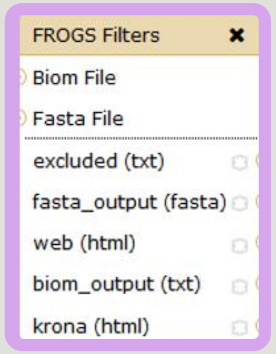
Filters tool

Affiliation runs long time

Advise:
Apply filters between “Remove Chimera” and “Affiliation”.
Remove OTUs with weak abundance and non redundant before affiliation.
You will gain time !

A vous de jouer ! - 6

EXERCISE 6



FROGS Filters ✖

Biom File

Fasta File

excluded (txt)

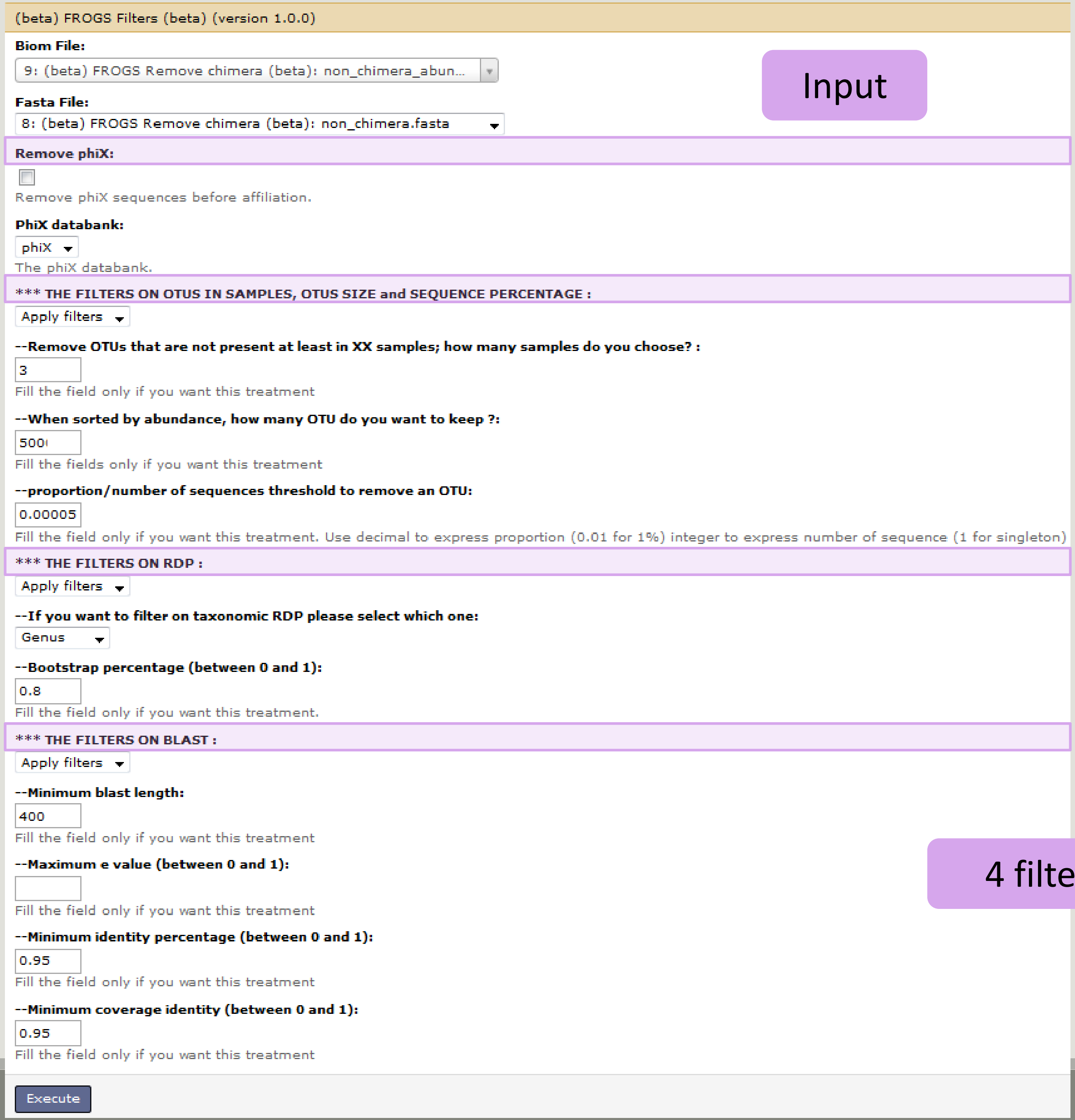
fasta_output (fasta)

web (html)

biom_output (txt)

krona (html)

Filters



(beta) FROGS Filters (beta) (version 1.0.0)

Biom File:

9: (beta) FROGS Remove chimera (beta): non_chimera_abun...

Fasta File:

8: (beta) FROGS Remove chimera (beta): non_chimera.fasta

Remove phiX:

☐

Remove phiX sequences before affiliation.

PhiX databank:

phiX

The phiX databank.

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

Apply filters

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

3

Fill the field only if you want this treatment

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

500(

Fill the fields only if you want this treatment

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

0.00005

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

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

Apply filters

--If you want to filter on taxonomic RDP please select which one:

Genus

--Bootstrap percentage (between 0 and 1):

0.8

Fill the field only if you want this treatment.

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

Apply filters

--Minimum blast length:

400

Fill the field only if you want this treatment

--Maximum e value (between 0 and 1):

Fill the field only if you want this treatment

--Minimum identity percentage (between 0 and 1):

0.95

Fill the field only if you want this treatment

--Minimum coverage identity (between 0 and 1):

0.95

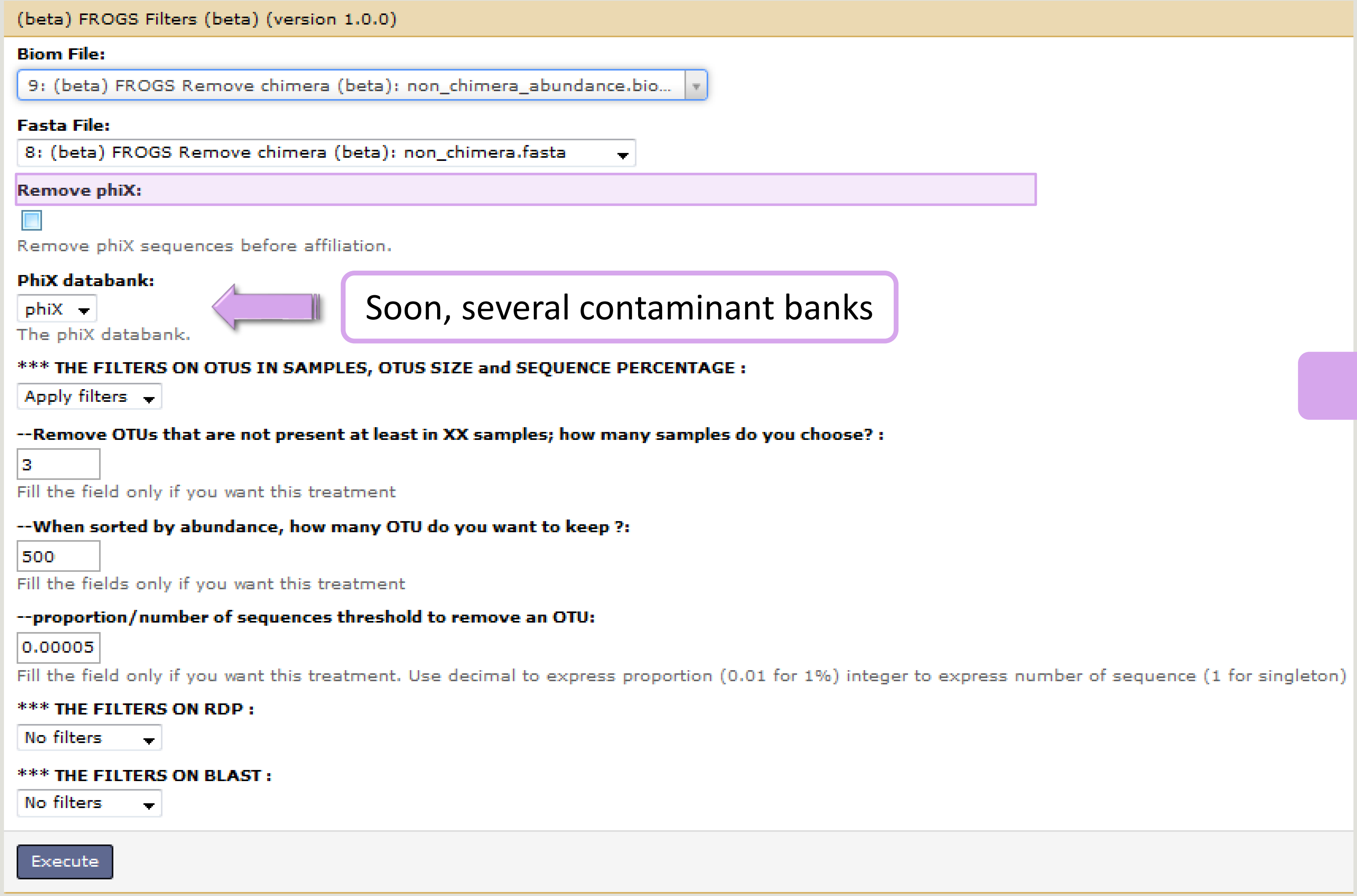
Fill the field only if you want this treatment

Execute

Input

4 filter sections

Input



(beta) FROGS Filters (beta) (version 1.0.0)

Biom File:

9: (beta) FROGS Remove chimera (beta): non_chimera_abundance.bio...

Fasta File:

8: (beta) FROGS Remove chimera (beta): non_chimera.fasta

Remove phiX:

□

Remove phiX sequences before affiliation.

PhiX databank:

phiX

The phiX databank.

Soon, several contaminant banks

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

Apply filters

Filter 1

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

3

Fill the field only if you want this treatment

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

500

Fill the fields only if you want this treatment

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

0.00005

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

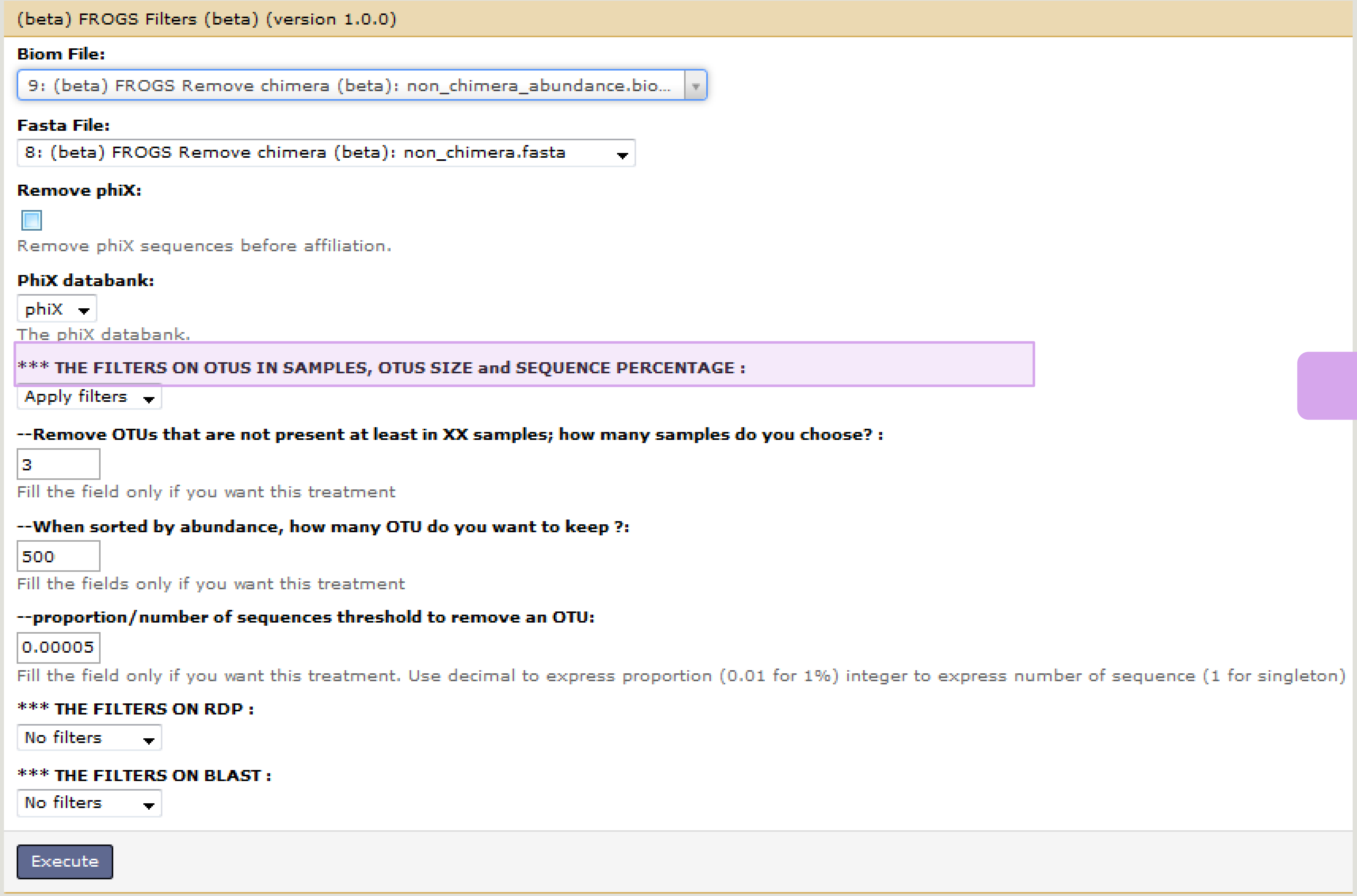
***** THE FILTERS ON RDP :**

No filters

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

No filters

Input



(beta) FROGS Filters (beta) (version 1.0.0)

Biom File:

9: (beta) FROGS Remove chimera (beta): non_chimera_abundance.bio...

Fasta File:

8: (beta) FROGS Remove chimera (beta): non_chimera.fasta

Remove phiX:

□

Remove phiX sequences before affiliation.

PhiX databank:

phiX

The phiX databank.

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

Apply filters

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

3

Fill the field only if you want this treatment

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

500

Fill the fields only if you want this treatment

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

0.00005

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

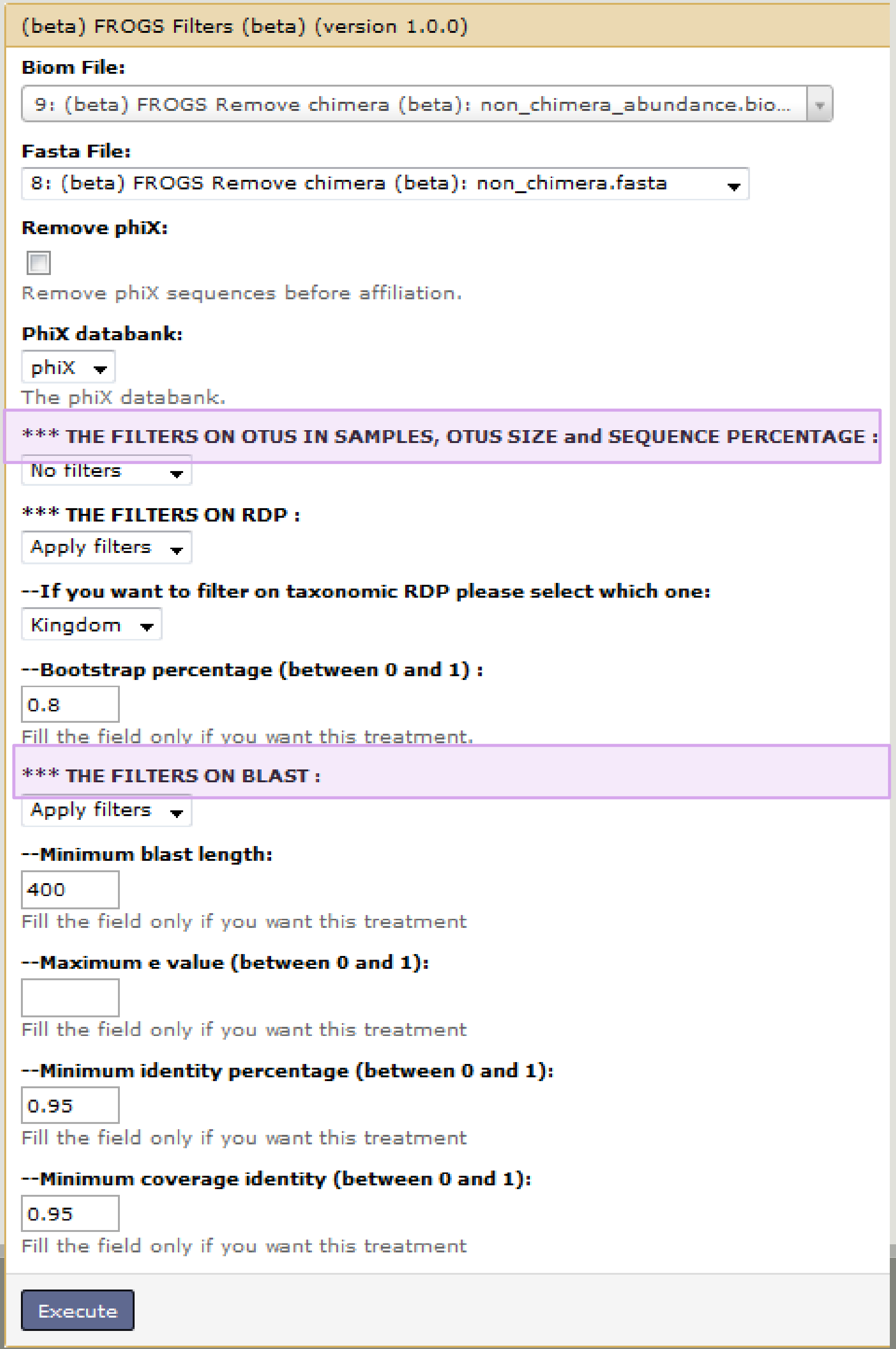
***** THE FILTERS ON RDP :**

No filters

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

No filters

Filter 2



(beta) FROGS Filters (beta) (version 1.0.0)

Biom File:

9: (beta) FROGS Remove chimera (beta): non_chimera_abundance.bio...

Fasta File:

8: (beta) FROGS Remove chimera (beta): non_chimera.fasta

Remove phiX:

□

Remove phiX sequences before affiliation.

PhiX databank:

phiX

The phiX databank.

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

No filters

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

Apply filters

--If you want to filter on taxonomic RDP please select which one:

Kingdom

--Bootstrap percentage (between 0 and 1) :

0.8

Fill the field only if you want this treatment.

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

Apply filters

--Minimum blast length:

400

Fill the field only if you want this treatment

--Maximum e value (between 0 and 1):

Fill the field only if you want this treatment

--Minimum identity percentage (between 0 and 1):

0.95

Fill the field only if you want this treatment

--Minimum coverage identity (between 0 and 1):

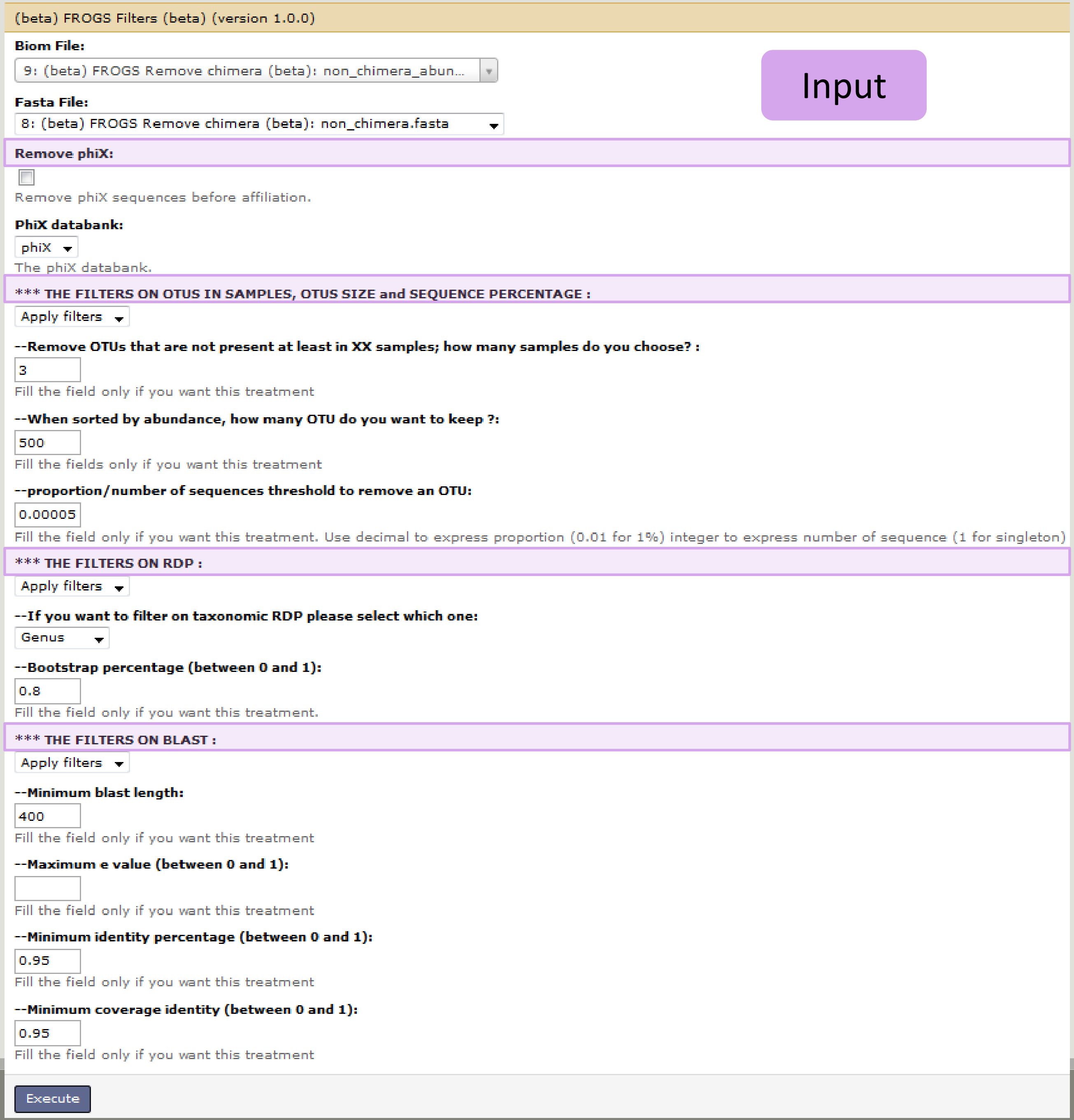
0.95

Fill the field only if you want this treatment

Input

filters 3 & 4

Input



(beta) FROGS Filters (beta) (version 1.0.0)

Biom File:
9: (beta) FROGS Remove chimera (beta): non_chimera_abun...

Fasta File:
8: (beta) FROGS Remove chimera (beta): non_chimera.fasta

Remove phiX:
☐
Remove phiX sequences before affiliation.

PhiX databank:
phiX
The phiX databank.

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

--Remove OTUs that are not present at least in XX samples; how many samples do you choose? :
3
Fill the field only if you want this treatment

--When sorted by abundance, how many OTU do you want to keep ?:
500
Fill the fields only if you want this treatment

--proportion/number of sequences threshold to remove an OTU:
0.00005
Fill the field only if you want this treatment. Use decimal to express proportion (0.01 for 1%) integer to express number of sequence (1 for singleton)

*** THE FILTERS ON RDP :
Apply filters

--If you want to filter on taxonomic RDP please select which one:
Genus

--Bootstrap percentage (between 0 and 1):
0.8
Fill the field only if you want this treatment.

*** THE FILTERS ON BLAST :
Apply filters

--Minimum blast length:
400
Fill the field only if you want this treatment

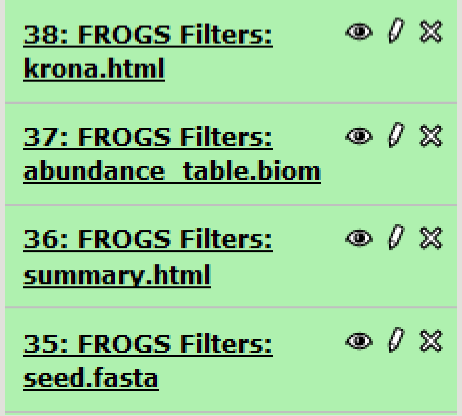
--Maximum e value (between 0 and 1):

Fill the field only if you want this treatment

--Minimum identity percentage (between 0 and 1):
0.95
Fill the field only if you want this treatment

--Minimum coverage identity (between 0 and 1):
0.95
Fill the field only if you want this treatment

Output

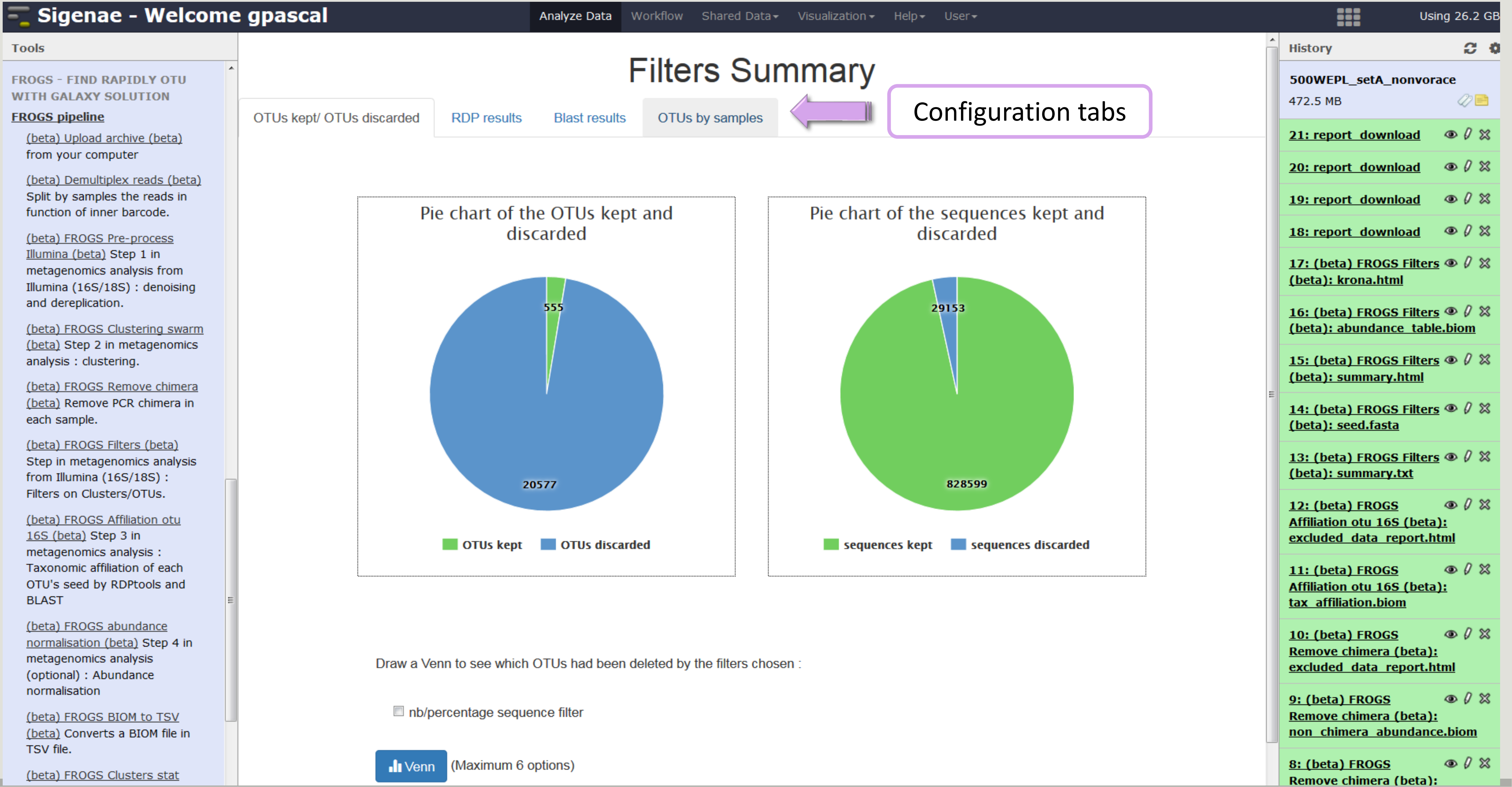


38: FROGS Filters: krona.html

37: FROGS Filters: abundance_table.biom

36: FROGS Filters: summary.html

35: FROGS Filters: seed.fasta



FROGS - FIND RAPIDLY OTU WITH GALAXY SOLUTION

FROGS pipeline

(beta) Upload archive (beta) from your computer

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

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

(beta) FROGS Clustering swarm (beta) Step 2 in metagenomics analysis : clustering.

(beta) FROGS Remove chimera (beta) Remove PCR chimera in each sample.

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

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

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

(beta) FROGS BIOM to TSV (beta) Converts a BIOM file in TSV file.

(beta) FROGS Clusters stat

Filters Summary

Configuration tabs

Pie chart of the OTUs kept and discarded

Pie chart of the sequences kept and discarded

Draw a Venn to see which OTUs had been deleted by the filters chosen :

☐ nb/percentage sequence filter

(Maximum 6 options)

500WEPL_setA_nonvorace

472.5 MB

21: report_download

20: report_download

19: report_download

18: report_download

17: (beta) FROGS Filters (beta): krona.html

16: (beta) FROGS Filters (beta): abundance_table.biom

15: (beta) FROGS Filters (beta): summary.html

14: (beta) FROGS Filters (beta): seed.fasta

13: (beta) FROGS Filters (beta): summary.txt

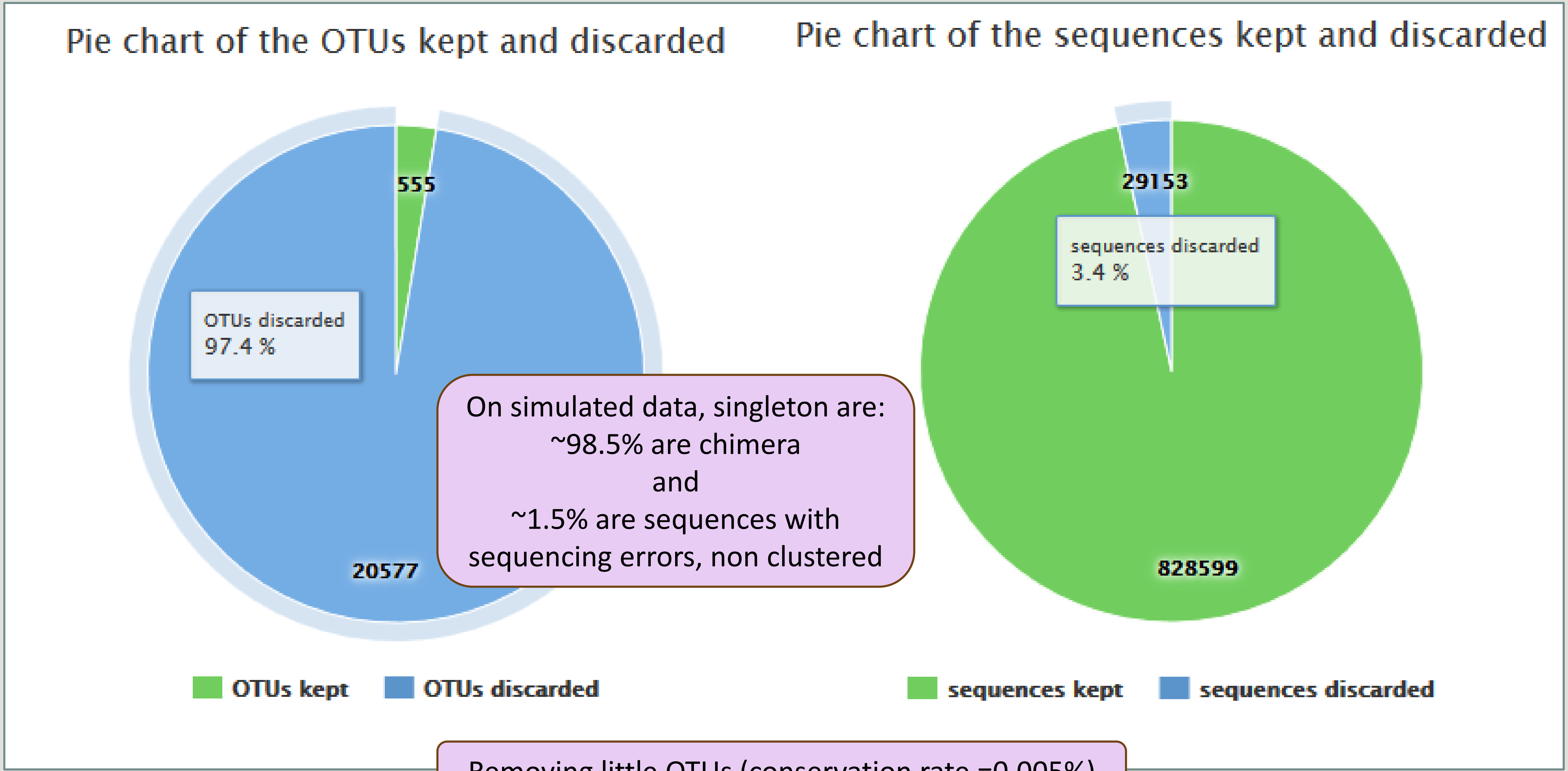
12: (beta) FROGS Affiliation otu 16S (beta): excluded_data_report.html

11: (beta) FROGS Affiliation otu 16S (beta): tax_affiliation.biom

10: (beta) FROGS Remove chimera (beta): excluded_data_report.html

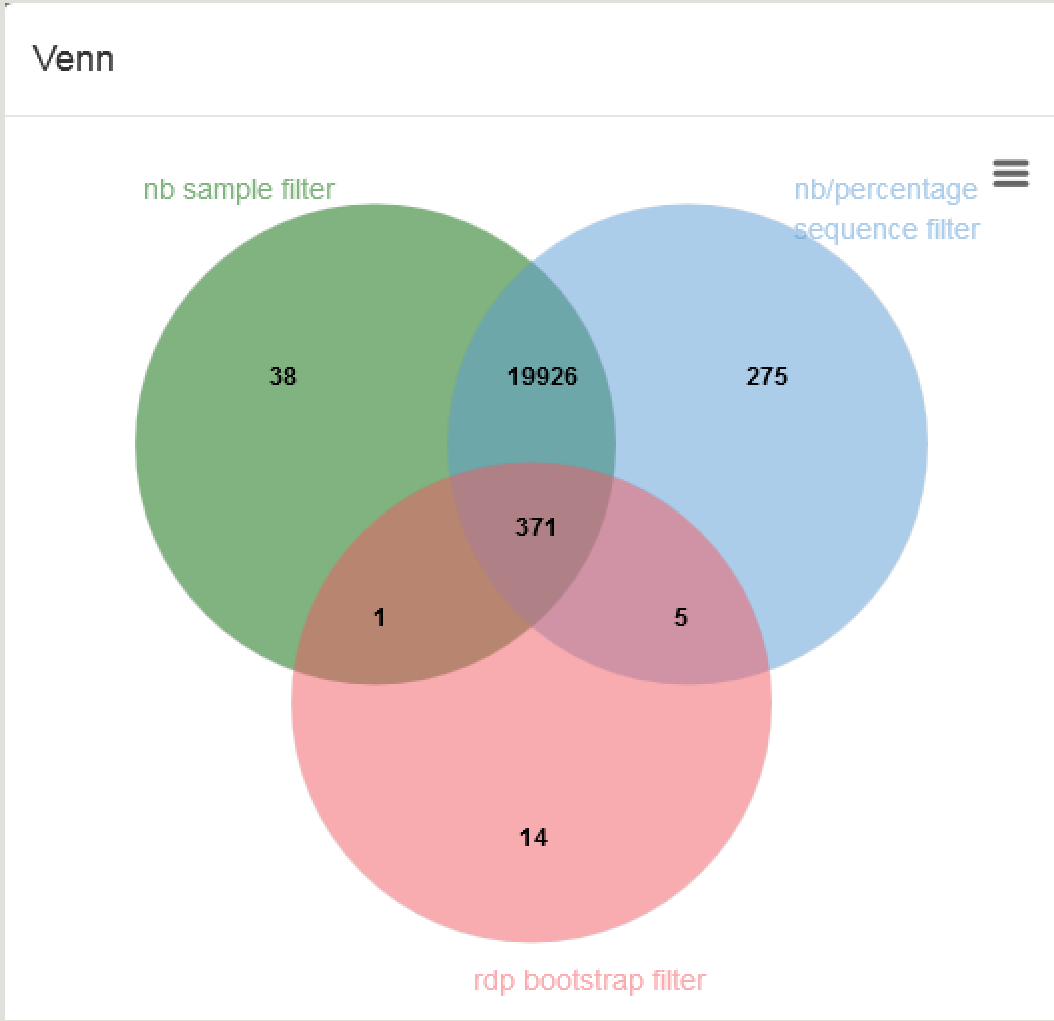
9: (beta) FROGS Remove chimera (beta): non_chimera_abundance.biom

8: (beta) FROGS Remove chimera (beta):



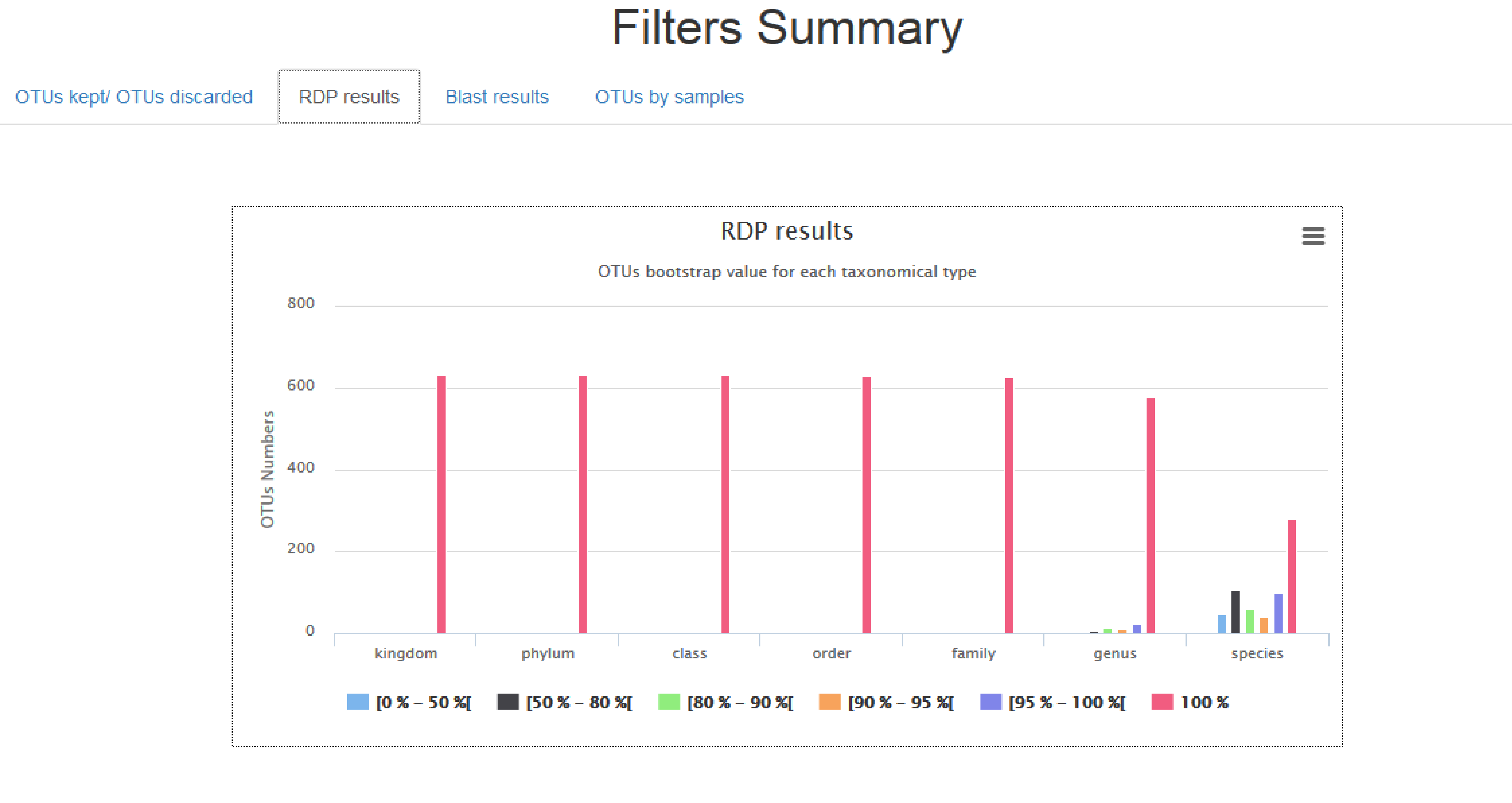
On simulated data, singleton are:
~98.5% are chimera
and
~1.5% are sequences with
sequencing errors, non clustered

Removing little OTUs (conservation rate =0.005%)



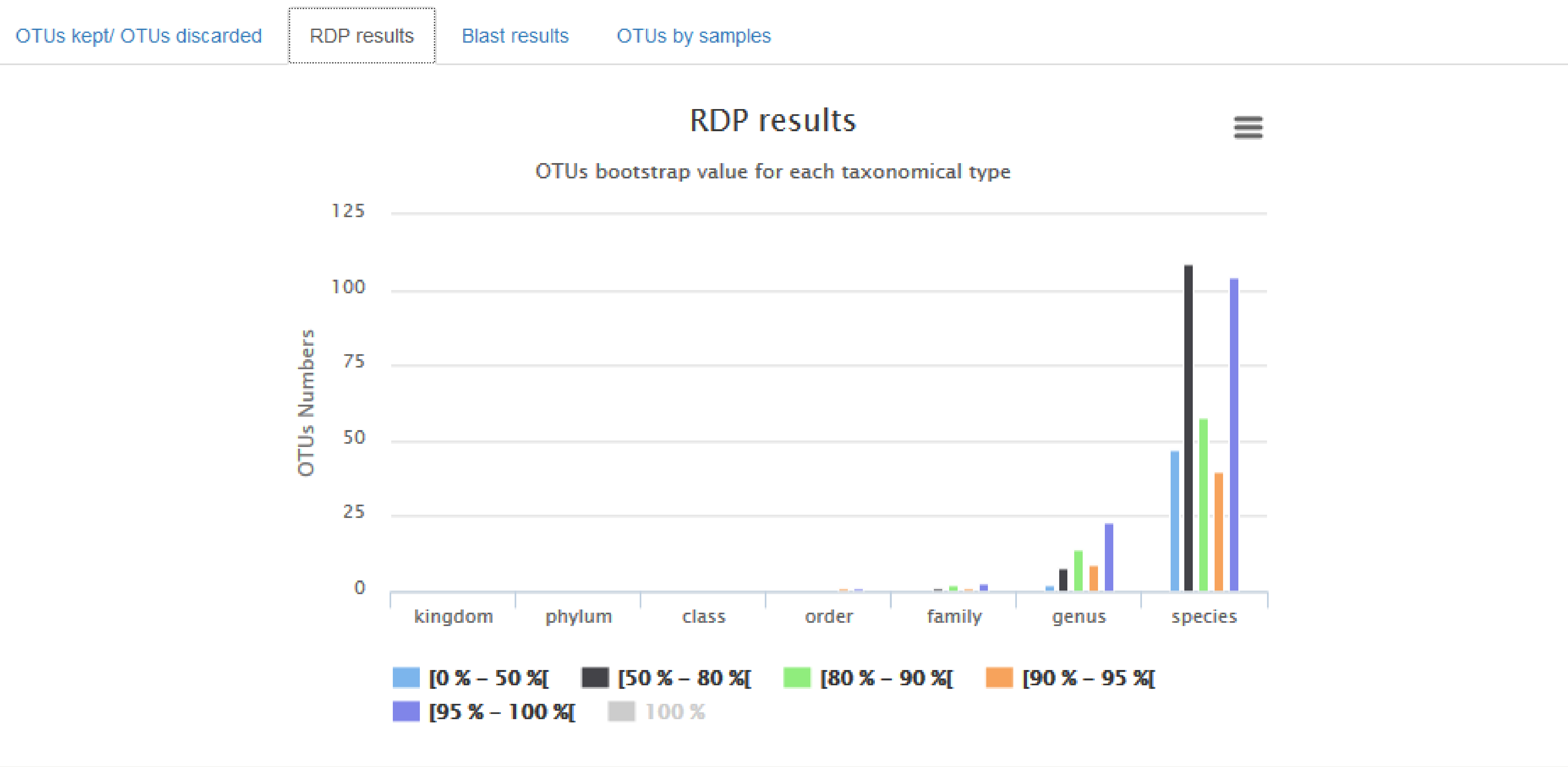
Venn

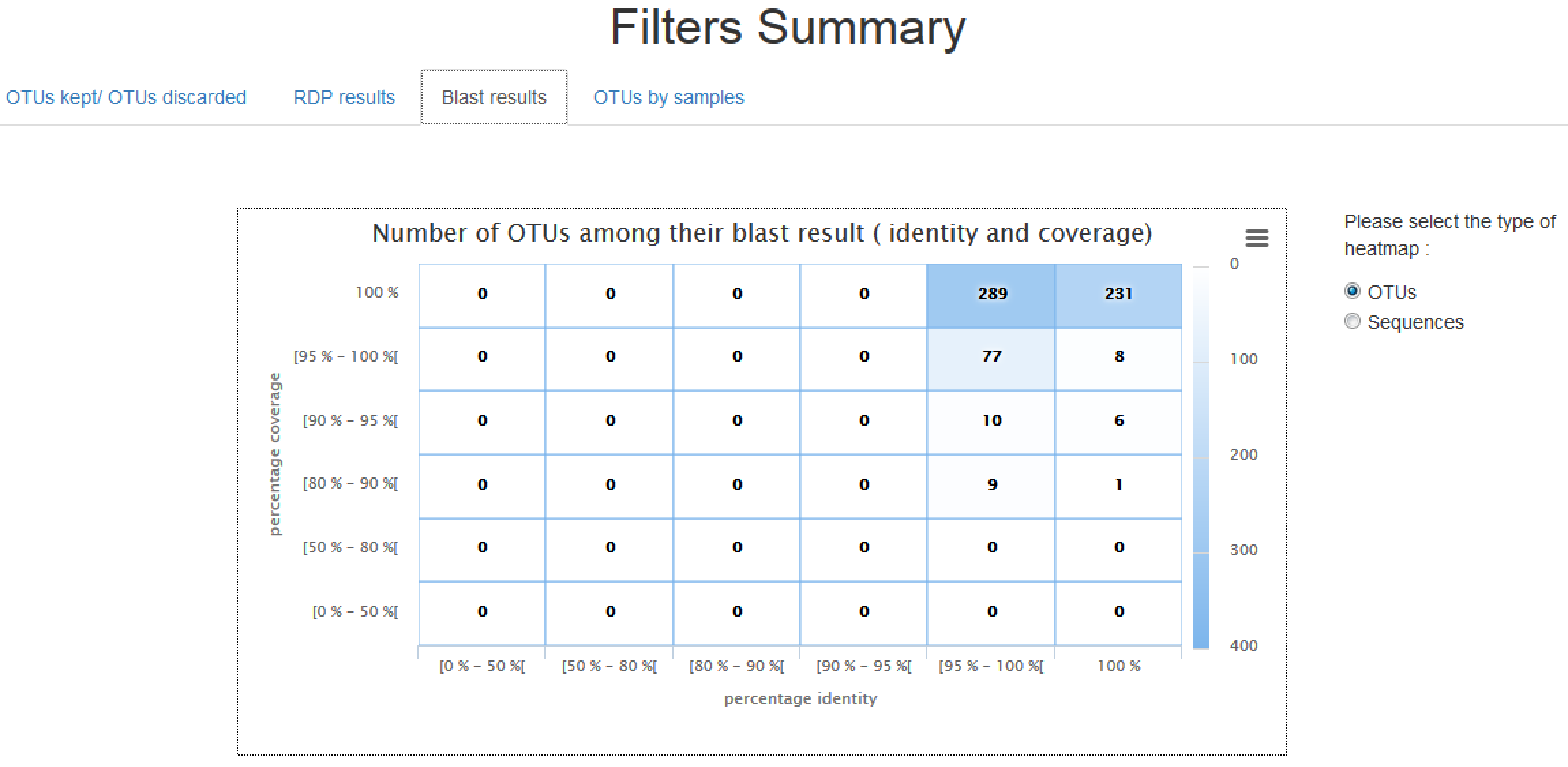
nb sample filter
nb/percentage sequence filter
rdp bootstrap filter
38
19926
275
371
1
5
14



Filters Summary

Filters Summary





Filters Summary

OTUs kept/ OTUs discarded | RDP results | Blast results | OTUs by samples

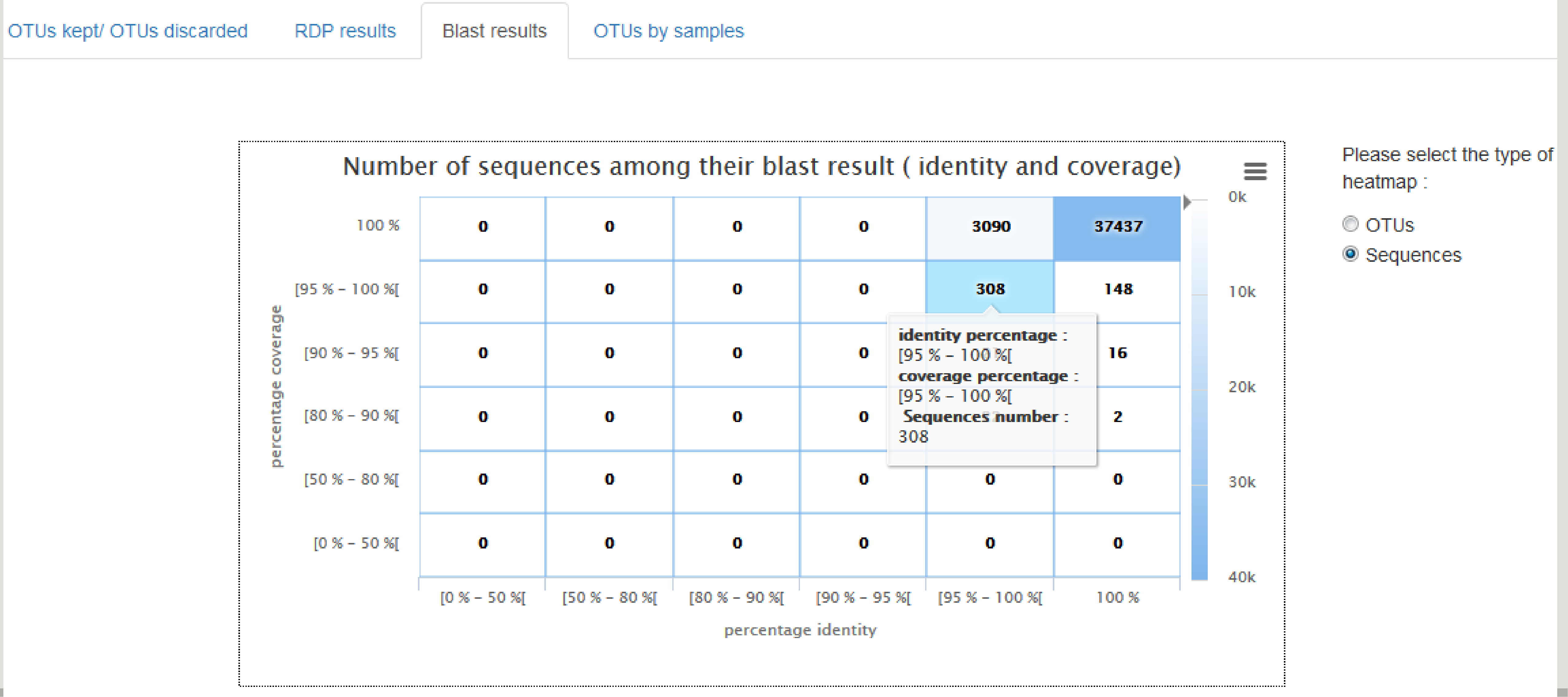
Number of OTUs among their blast result (identity and coverage)

percentage coverage \ percentage identity	[0 % – 50 %[	[50 % – 80 %[	[80 % – 90 %[	[90 % – 95 %[	[95 % – 100 %[	100 %
100 %	0	0	0	0	289	231
[95 % – 100 %[	0	0	0	0	77	8
[90 % – 95 %[	0	0	0	0	10	6
[80 % – 90 %[	0	0	0	0	9	1
[50 % – 80 %[	0	0	0	0	0	0
[0 % – 50 %[	0	0	0	0	0	0

Please select the type of heatmap :

● OTUs
○ Sequences

Filters Summary



Number of sequences among their blast result (identity and coverage)

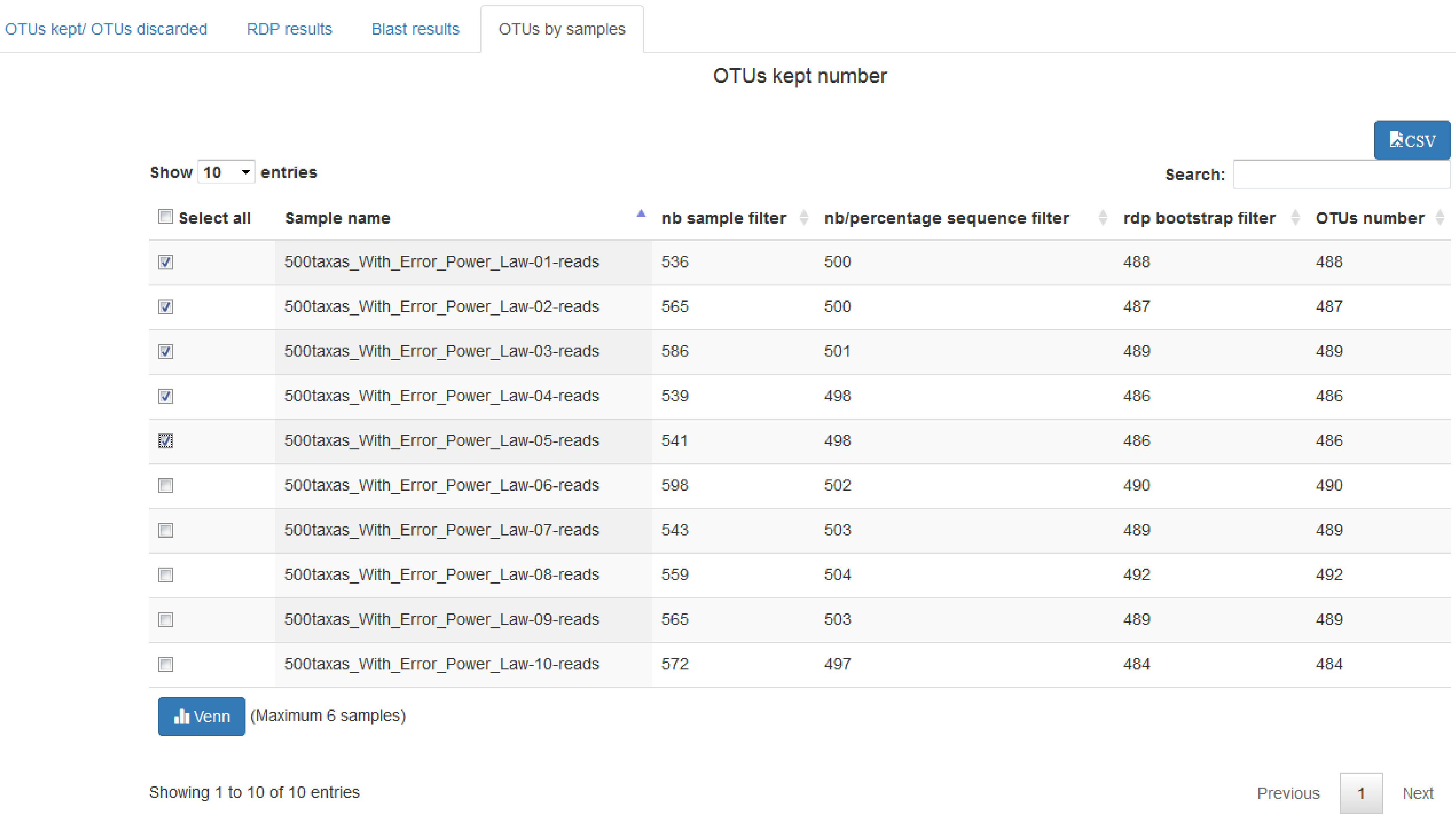
percentage coverage \ percentage identity	[0 % – 50 %[	[50 % – 80 %[	[80 % – 90 %[	[90 % – 95 %[	[95 % – 100 %[	100 %
100 %	0	0	0	0	3090	37437
[95 % – 100 %[	0	0	0	0	308	148
[90 % – 95 %[	0	0	0	0	[illegible]	16
[80 % – 90 %[	0	0	0	0	[illegible]	2
[50 % – 80 %[	0	0	0	0	0	0
[0 % – 50 %[	0	0	0	0	0	0

identity percentage : [95 % – 100 %[
coverage percentage : [95 % – 100 %[
Sequences number : 308

Please select the type of heatmap :

○ OTUs
● Sequences

Filters Summary

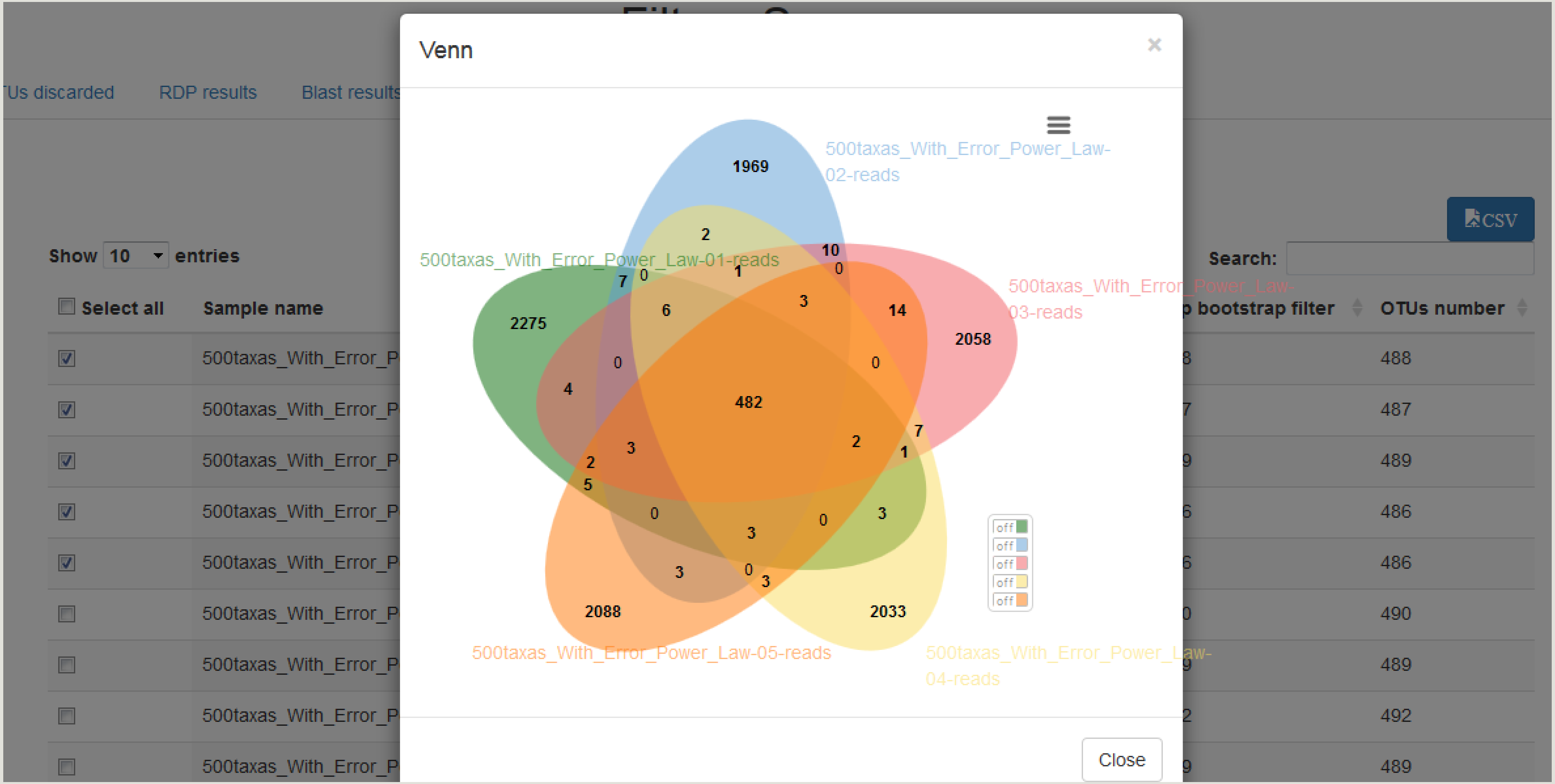


OTUs kept number

☐ Select all	Sample name	nb sample filter	nb/percentage sequence filter	rdp bootstrap filter	OTUs number
☑	500taxas_With_Error_Power_Law-01-reads	536	500	488	488
☑	500taxas_With_Error_Power_Law-02-reads	565	500	487	487
☑	500taxas_With_Error_Power_Law-03-reads	586	501	489	489
☑	500taxas_With_Error_Power_Law-04-reads	539	498	486	486
☑	500taxas_With_Error_Power_Law-05-reads	541	498	486	486
☐	500taxas_With_Error_Power_Law-06-reads	598	502	490	490
☐	500taxas_With_Error_Power_Law-07-reads	543	503	489	489
☐	500taxas_With_Error_Power_Law-08-reads	559	504	492	492
☐	500taxas_With_Error_Power_Law-09-reads	565	503	489	489
☐	500taxas_With_Error_Power_Law-10-reads	572	497	484	484

(Maximum 6 samples)

Showing 1 to 10 of 10 entries



Us discarded RDP results Blast results

Venn

×

Show 10 entries

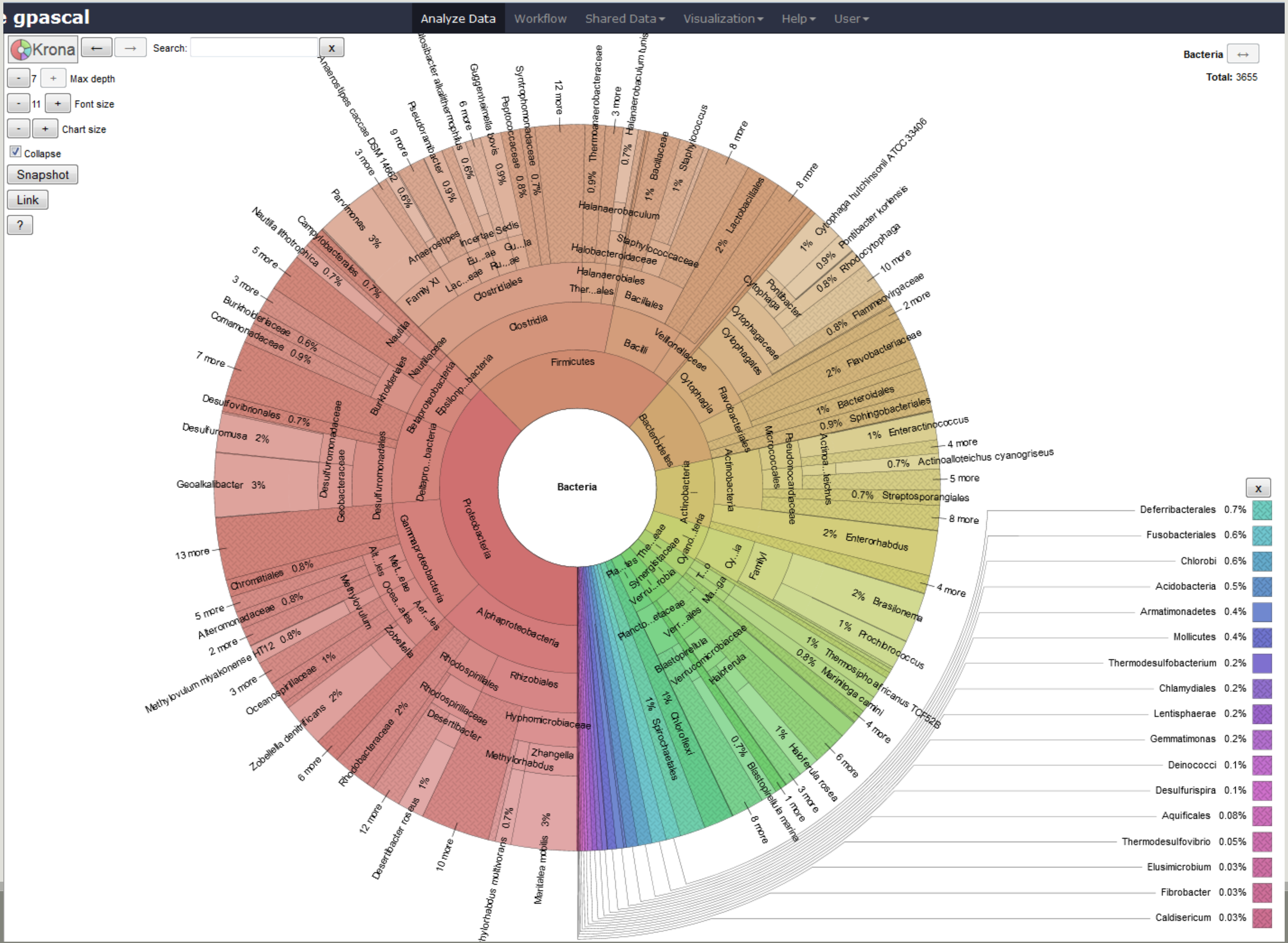
Search:

CSV

☐ Select all	Sample name	p bootstrap filter	OTUs number
☑	500taxas_With_Error_P	8	488
☑	500taxas_With_Error_P	7	487
☑	500taxas_With_Error_P	9	489
☑	500taxas_With_Error_P	6	486
☑	500taxas_With_Error_P	6	486
☐	500taxas_With_Error_P	0	490
☐	500taxas_With_Error_P	9	489
☐	500taxas_With_Error_P	2	492
☐	500taxas_With_Error_P	9	489

500taxas_With_Error_Power_Law-02-reads
500taxas_With_Error_Power_Law-01-reads
500taxas_With_Error_Power_Law-03-reads
500taxas_With_Error_Power_Law-05-reads
500taxas_With_Error_Power_Law-04-reads
1969
2
10
7
0
1
0
3
6
14
2275
2058
0
0
4
482
7
3
2
1
2
5
0
3
0
3
3
0
3
2088
2033
off 🟩
off 🟦
off 🟥
off 🟨
off 🟧

Close



☐ 7 ☐ Max depth

☐ 11 ☐ Font size

☐ ☐ Chart size

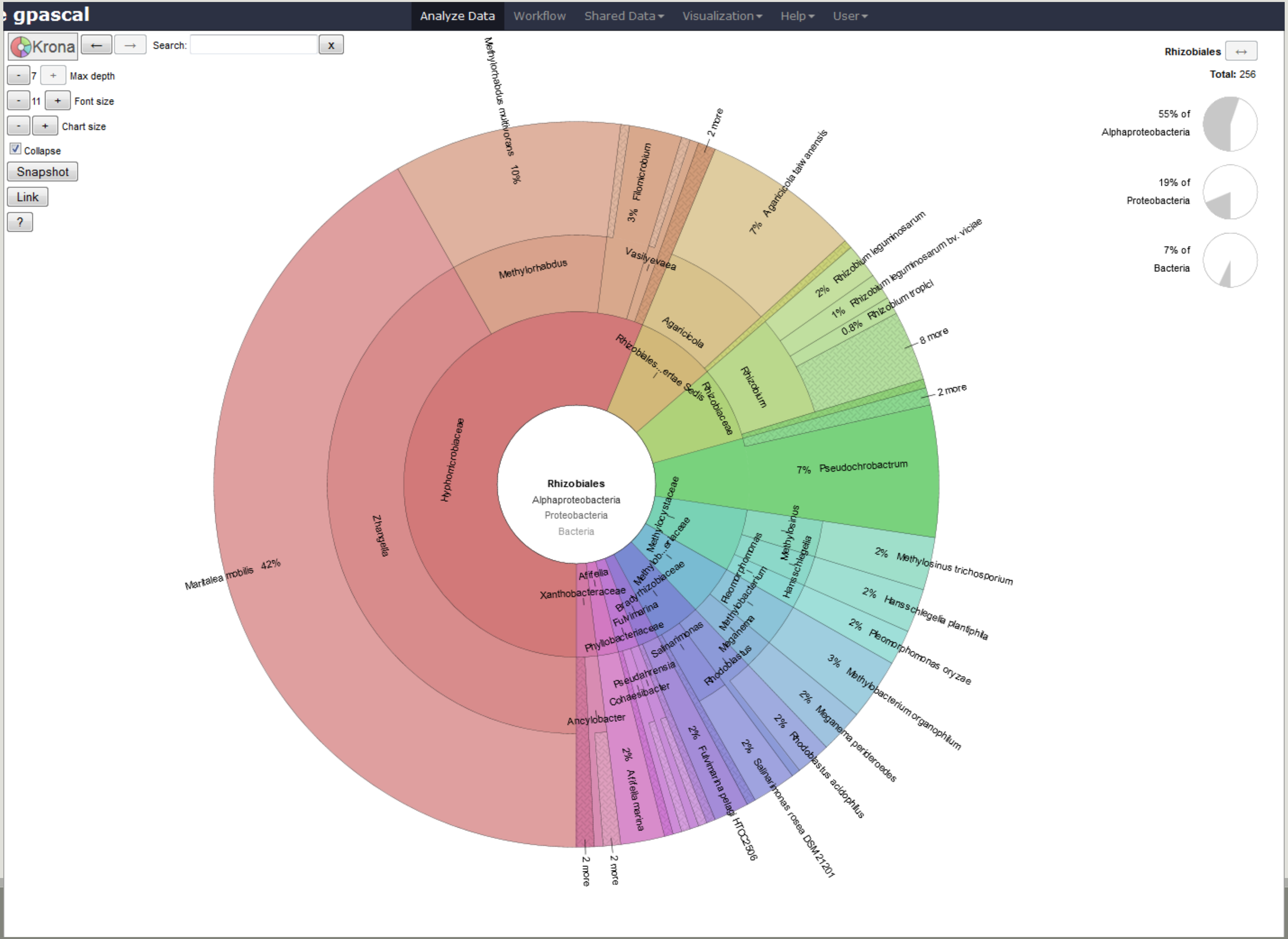
☑ Collapse

Bacteria ↔

Total: 3655

x

Deferribacterales	0.7%	▨
Fusobacteriales	0.6%	▨
Chlorobi	0.6%	▨
Acidobacteria	0.5%	▨
Armatimonadetes	0.4%	▨
Mollicutes	0.4%	▨
Thermodesulfobacterium	0.2%	▨
Chlamydiales	0.2%	▨
Lentisphaerae	0.2%	▨
Gemmatimonas	0.2%	▨
Deinococci	0.1%	▨
Desulfurispira	0.1%	▨
Aquificales	0.08%	▨
Thermodesulfovibrio	0.05%	▨
Elusimicrobium	0.03%	▨
Fibrobacter	0.03%	▨
Caldisericum	0.03%	▨



- 7 + Max depth

- 11 + Font size

☑ Collapse

Rhizobiales ↔

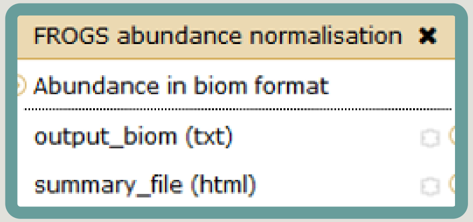
Total: 256

55% of Alphaproteobacteria

19% of Proteobacteria

7% of Bacteria

Normalisation



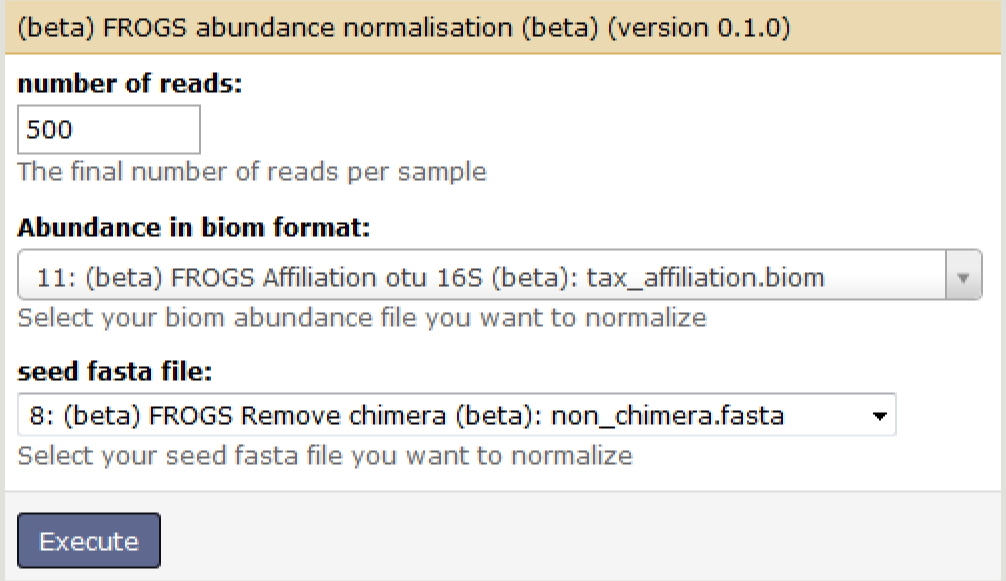
FROGS abundance normalisation ✖

Abundance in biom format

output_biom (txt)

summary_file (html)

Normalisation



(beta) FROGS abundance normalisation (beta) (version 0.1.0)

number of reads:

500

The final number of reads per sample

Abundance in biom format:

11: (beta) FROGS Affiliation otu 16S (beta): tax_affiliation.biom

Select your biom abundance file you want to normalize

seed fasta file:

8: (beta) FROGS Remove chimera (beta): non_chimera.fasta

Select your seed fasta file you want to normalize

Affiliation tool

1 Cluster = 2 affiliations

2 methods used:

RDP classifier (Ribosomal Database Project)

NCBI Blast+ vs. SILVA 119 (16S or 18S)

RDP classifier: bootstrap on each taxonomic subdivision

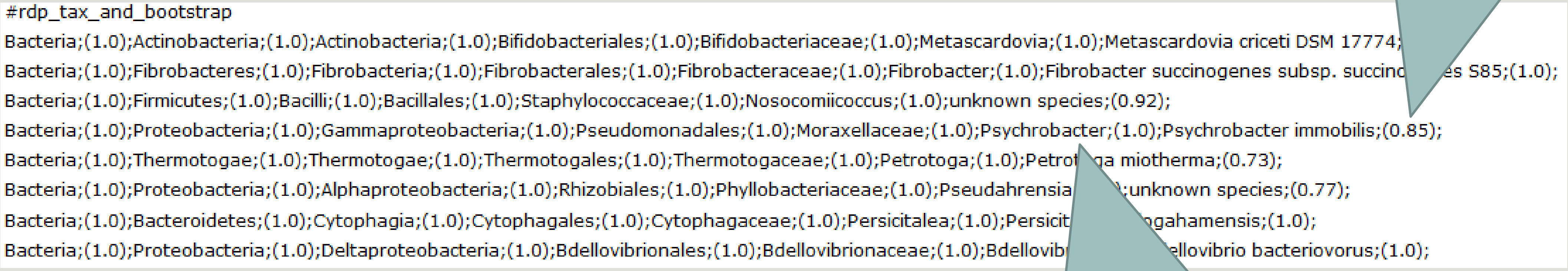
Blast: identity %, coverage %, e-value, alignment length

A vous de jouer ! – 7

EXERCISE 7

1st column - RDP

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



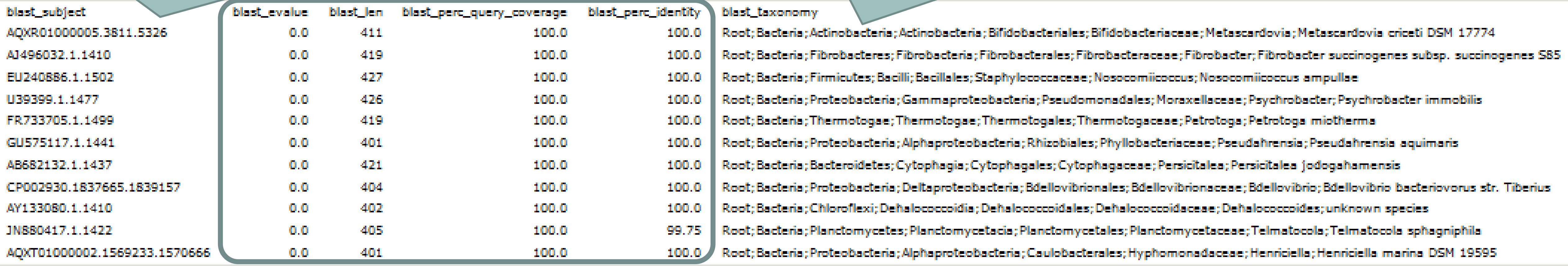
```
#rdp_tax_and_bootstrap
Bacteria;(1.0);Actinobacteria;(1.0);Actinobacteria;(1.0);Bifidobacteriales;(1.0);Bifidobacteriaceae;(1.0);Metascardovia;(1.0);Metascardovia criceti DSM 17774;
Bacteria;(1.0);Fibrobacteres;(1.0);Fibrobacteria;(1.0);Fibrobacterales;(1.0);Fibrobacteraceae;(1.0);Fibrobacter;(1.0);Fibrobacter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

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

2nd to 7th columns – Blast

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

The reference sequence taxonomic affiliation is this one.



blast_subject	blast_evalue	blast_len	blast_perc_query_coverage	blast_perc_identity	blast_taxonomy
AQXR01000005.3811.5326	0.0	411	100.0	100.0	Root;Bacteria;Actinobacteria;Actinobacteria;Bifidobacteriales;Bifidobacteriaceae;Metascardovia;Metascardovia criceti DSM 17774
AJ496032.1.1410	0.0	419	100.0	100.0	Root;Bacteria;Fibrobacteres;Fibrobacteria;Fibrobacterales;Fibrobacteraceae;Fibrobacter;Fibrobacter succinogenes subsp. succinogenes S85
EU240886.1.1502	0.0	427	100.0	100.0	Root;Bacteria;Firmicutes;Bacilli;Bacillales;Staphylococcaceae;Nosocomiicoccus;Nosocomiicoccus ampullae
U39399.1.1477	0.0	426	100.0	100.0	Root;Bacteria;Proteobacteria;Gammaproteobacteria;Pseudomonadales;Moraxellaceae;Psychrobacter;Psychrobacter immobilis
FR733705.1.1499	0.0	419	100.0	100.0	Root;Bacteria;Thermotogae;Thermotogae;Thermotogales;Thermotogaceae;Petrotoga;Petrotoga miotherma
GU575117.1.1441	0.0	401	100.0	100.0	Root;Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Phyllobacteriaceae;Pseudahrensia;Pseudahrensia aquimaris
AB682132.1.1437	0.0	421	100.0	100.0	Root;Bacteria;Bacteroidetes;Cytophagia;Cytophagales;Cytophagaceae;Persicitalea;Persicitalea jodogahamensis
CP002930.1837665.1839157	0.0	404	100.0	100.0	Root;Bacteria;Proteobacteria;Deltaproteobacteria;Bdellovibrionales;Bdellovibrionaceae;Bdellovibrio;Bdellovibrio bacteriovorus str. Tiberius
AY133080.1.1410	0.0	402	100.0	100.0	Root;Bacteria;Chloroflexi;Dehalococcoidia;Dehalococcoidales;Dehalococcoidaceae;Dehalococcoides;unknown species
JN880417.1.1422	0.0	405	100.0	99.75	Root;Bacteria;Planctomycetes;Planctomycetacia;Planctomycetales;Planctomycetaceae;Telmatocola;Telmatocola sphagniphila
AQXT01000002.1569233.1570666	0.0	401	100.0	100.0	Root;Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Hyphomonadaceae;Henriciella;Henriciella marina DSM 19595

Evaluation variables of BLAST

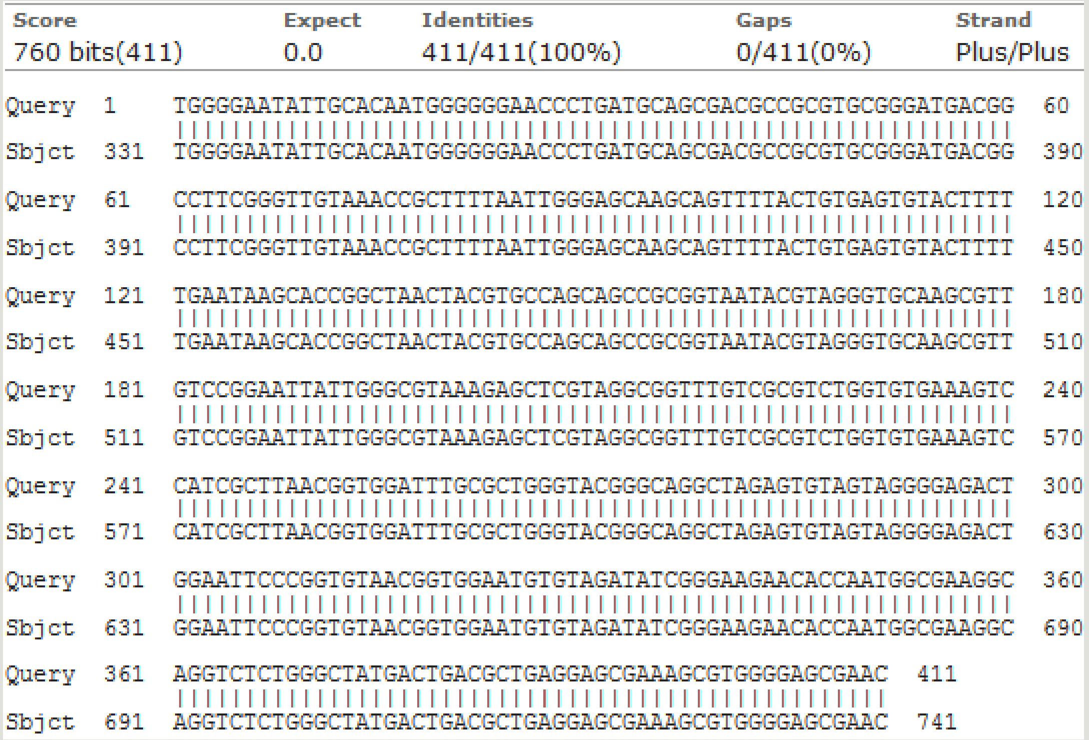
Blast variables : e-value

The Expect value (E) is a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size.

The lower the E-value, or the closer it is to zero, the more "significant" the match is.

Blast variables : blast_perc_identity

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

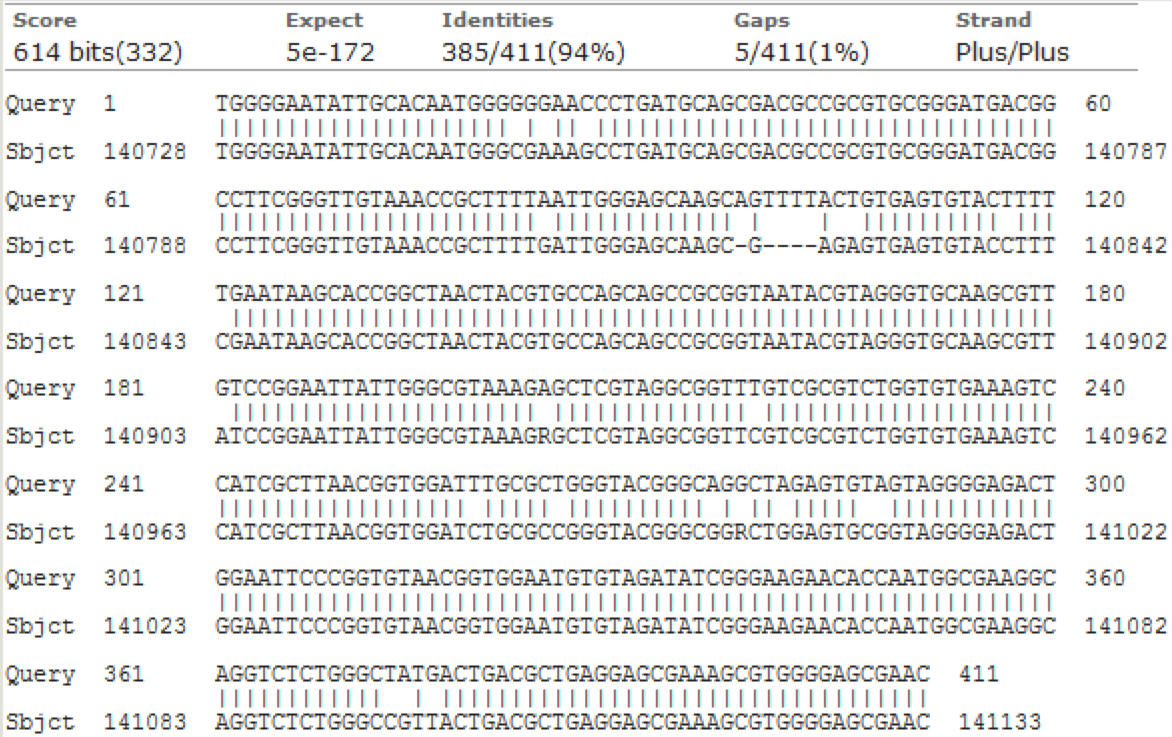


```
Score            Expect   Identities        Gaps         Strand
760 bits(411)    0.0      411/411(100%)     0/411(0%)    Plus/Plus

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

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

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

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

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

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

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

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

Blast variables : blast_perc_identity

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

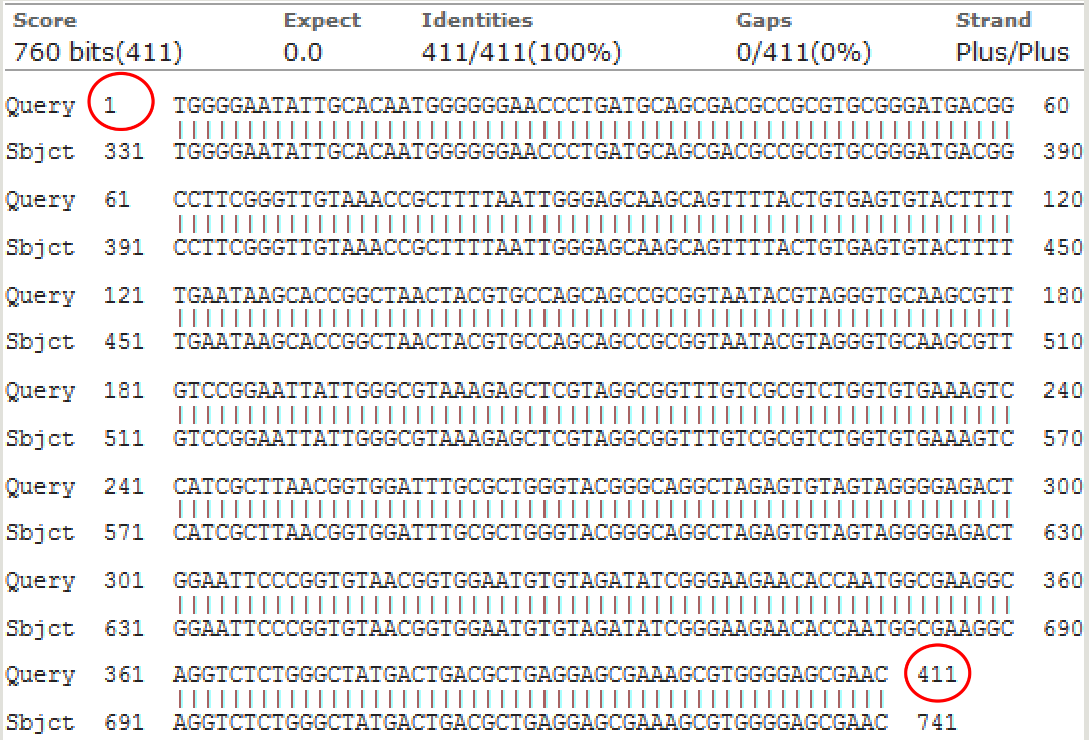


```
Score            Expect    Identities       Gaps         Strand
614 bits(332)    5e-172    385/411(94%)     5/411(1%)    Plus/Plus

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

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

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

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

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

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

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

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

Blast variables : blast_perc_query_coverage

Coverage percentage of alignment on query (OTU)



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

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

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

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

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

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

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

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

Query length = 411
100% coverage

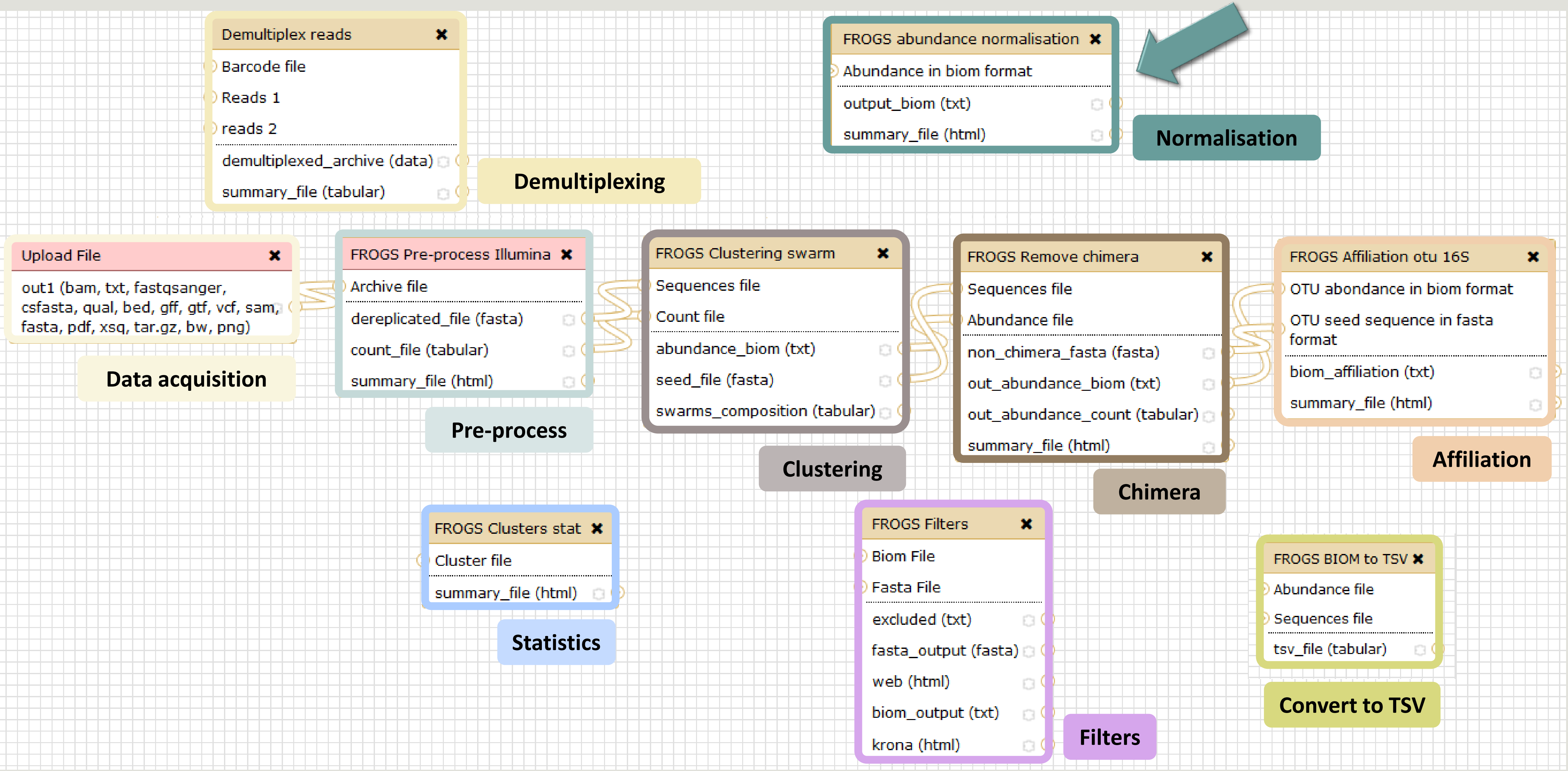
Blast variables : blast-length

Length of alignment between the OUT = “Query” and “subject” sequence of database (SILVA 119)

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

More mismatches/gaps

Normalisation



Demultiplex reads ✖
- Barcode file
- Reads 1
- reads 2
- demultiplexed_archive (data)
- summary_file (tabular)

Demultiplexing

FROGS abundance normalisation ✖
- Abundance in biom format
- output_biom (txt)
- summary_file (html)

Normalisation

Upload File ✖
- out1 (bam, txt, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png)

Data acquisition

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

Pre-process

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

Clustering

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

Chimera

FROGS Affiliation otu 16S ✖
- OTU abondance in biom format
- OTU seed sequence in fasta format
- biom_affiliation (txt)
- summary_file (html)

Affiliation

FROGS Clusters stat ✖
- Cluster file
- summary_file (html)

Statistics

FROGS Filters ✖
- Biom File
- Fasta File
- excluded (txt)
- fasta_output (fasta)
- web (html)
- biom_output (txt)
- krona (html)

Filters

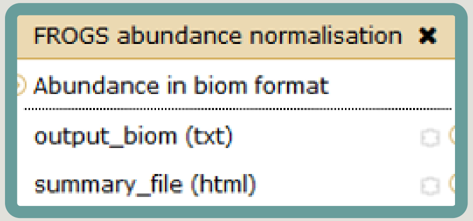
FROGS BIOM to TSV ✖
- Abundance file
- Sequences file
- tsv_file (tabular)

Convert to TSV

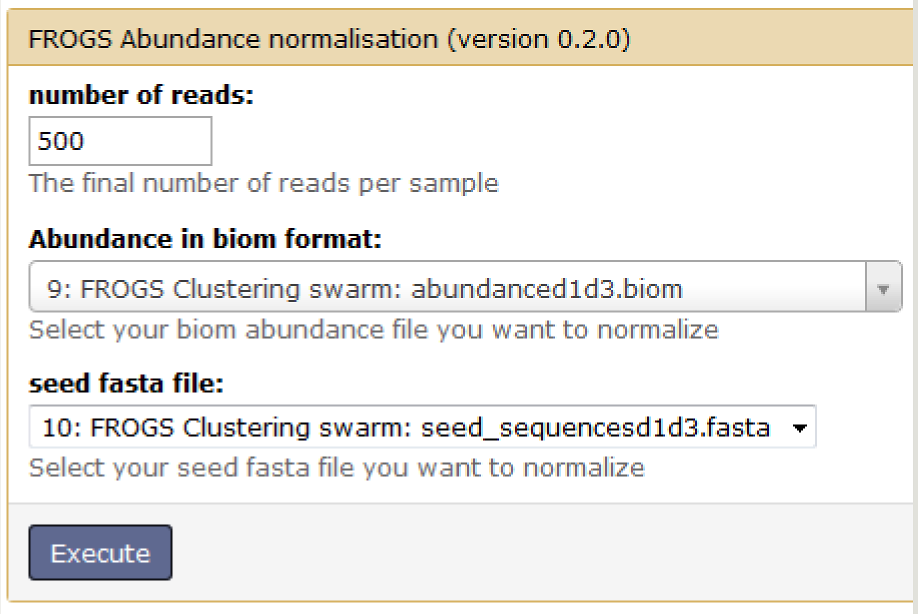
Normalisation

Conserve a predefined number of sequence per sample:

- update Biom abundance file
- update seed fasta file



Normalisation



FROGS Abundance normalisation (version 0.2.0)

number of reads:

500

The final number of reads per sample

Abundance in biom format:

9: FROGS Clustering swarm: abundanced1d3.biom

Select your biom abundance file you want to normalize

seed fasta file:

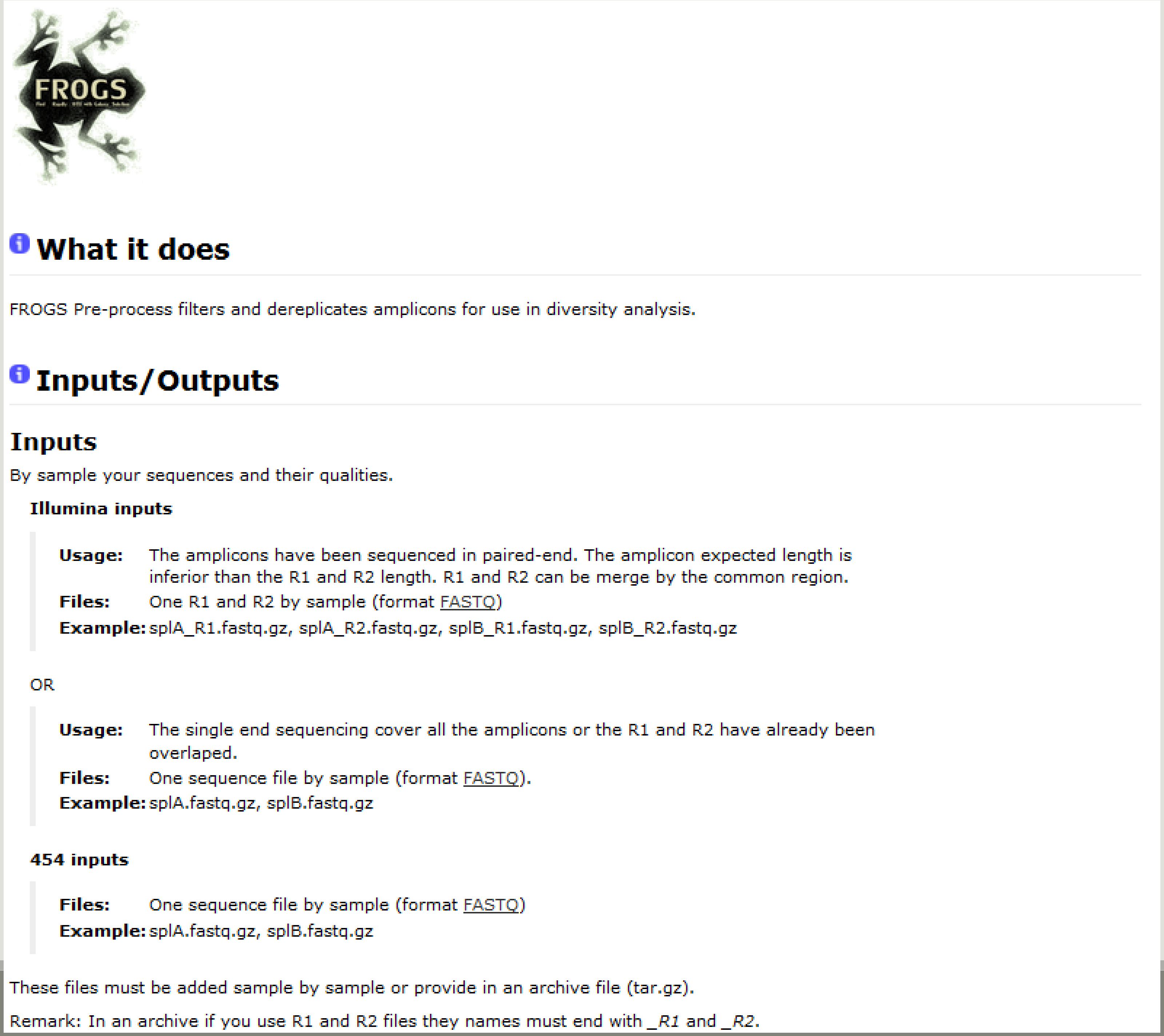
10: FROGS Clustering swarm: seed_sequencesd1d3.fasta

Select your seed fasta file you want to normalize

A vous de jouer ! – 8

EXERCISE 8

Tool descriptions



What it does

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

Inputs/Outputs

Inputs

By sample your sequences and their qualities.

Illumina inputs

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

OR

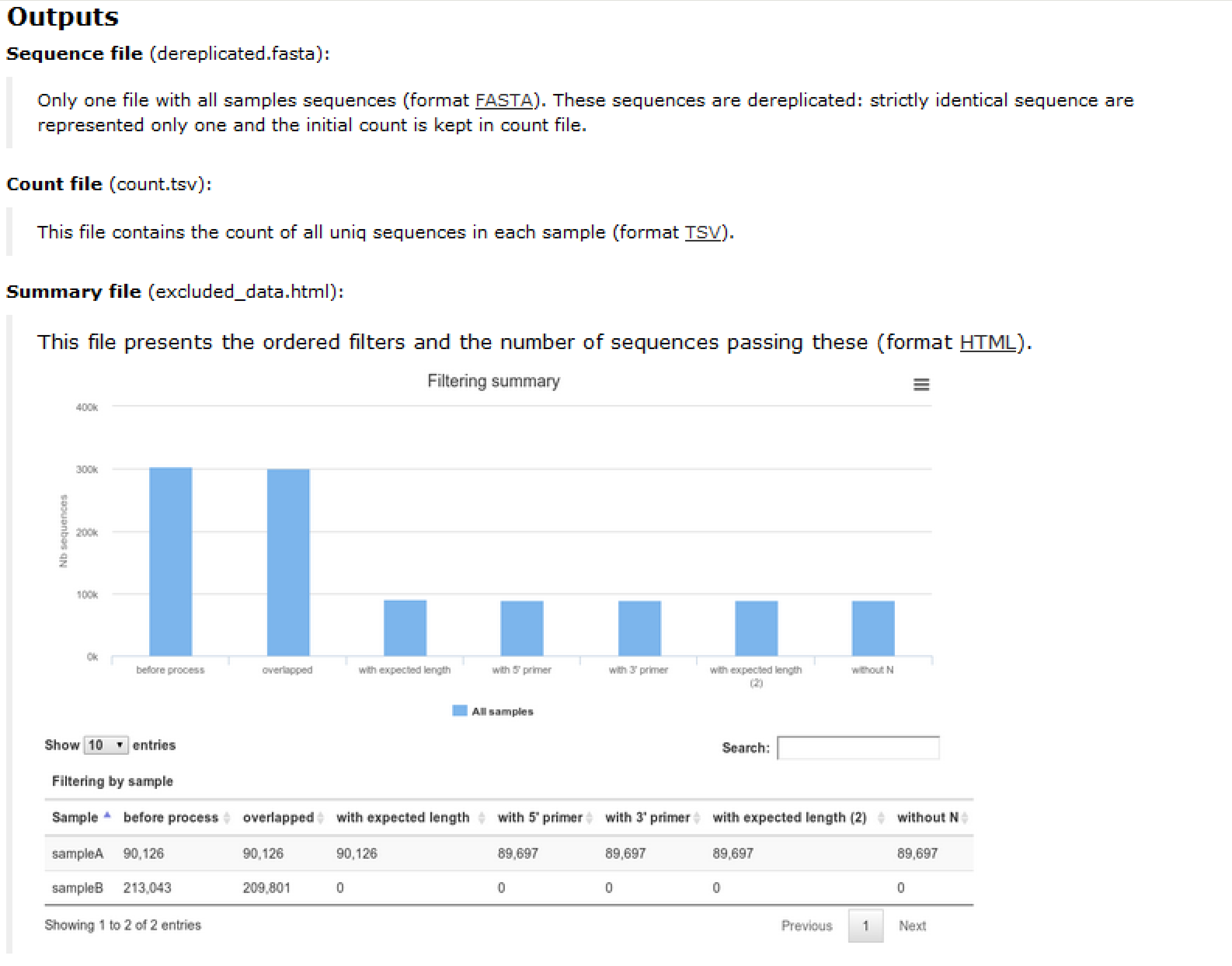
> **Usage:** The single end sequencing cover all the amplicons or the R1 and R2 have already been overlaped.
>
> **Files:** One sequence file by sample (format FASTQ).
>
> **Example:** splA.fastq.gz, splB.fastq.gz

454 inputs

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

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

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



Outputs

Sequence file (dereplicated.fasta):

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

Count file (count.tsv):

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

Summary file (excluded_data.html):

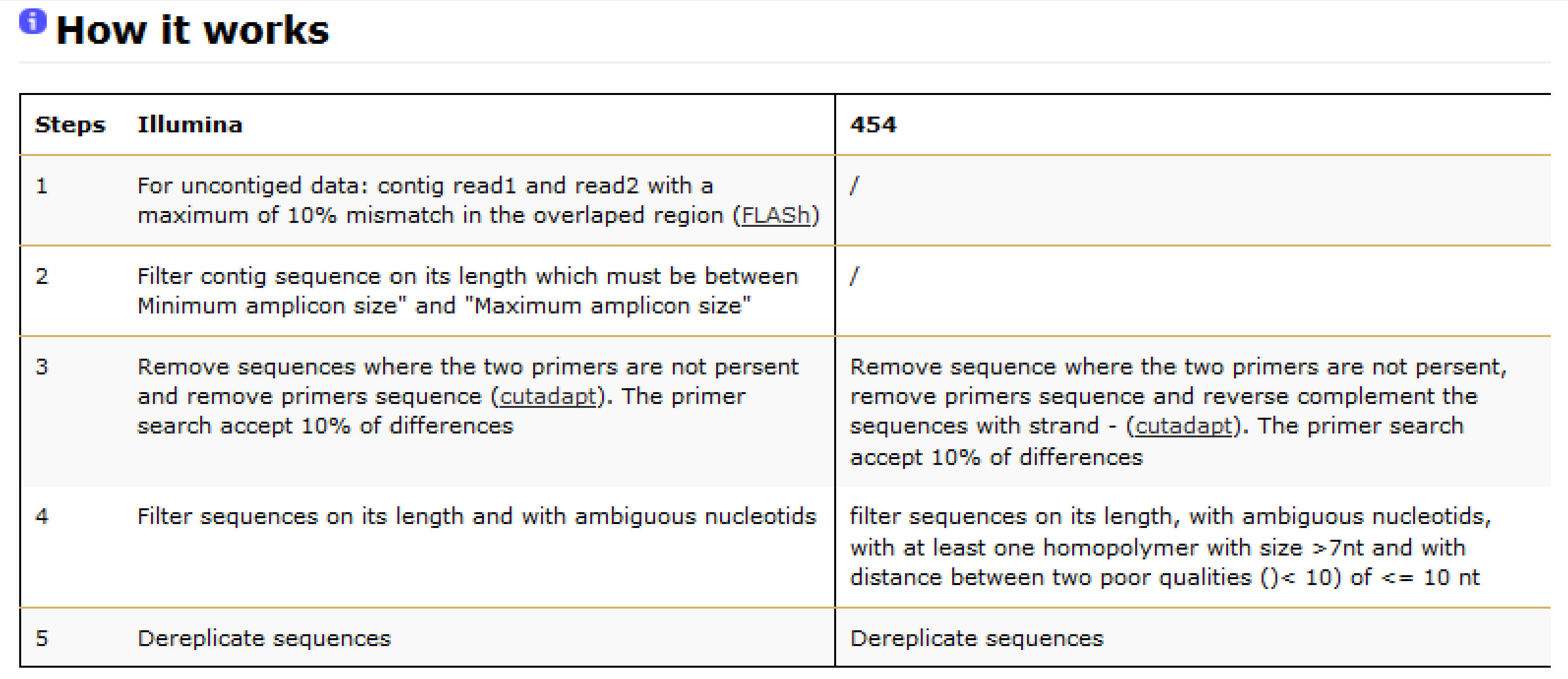
> This file presents the ordered filters and the number of sequences passing these (format HTML).

Filtering summary

Filtering by sample

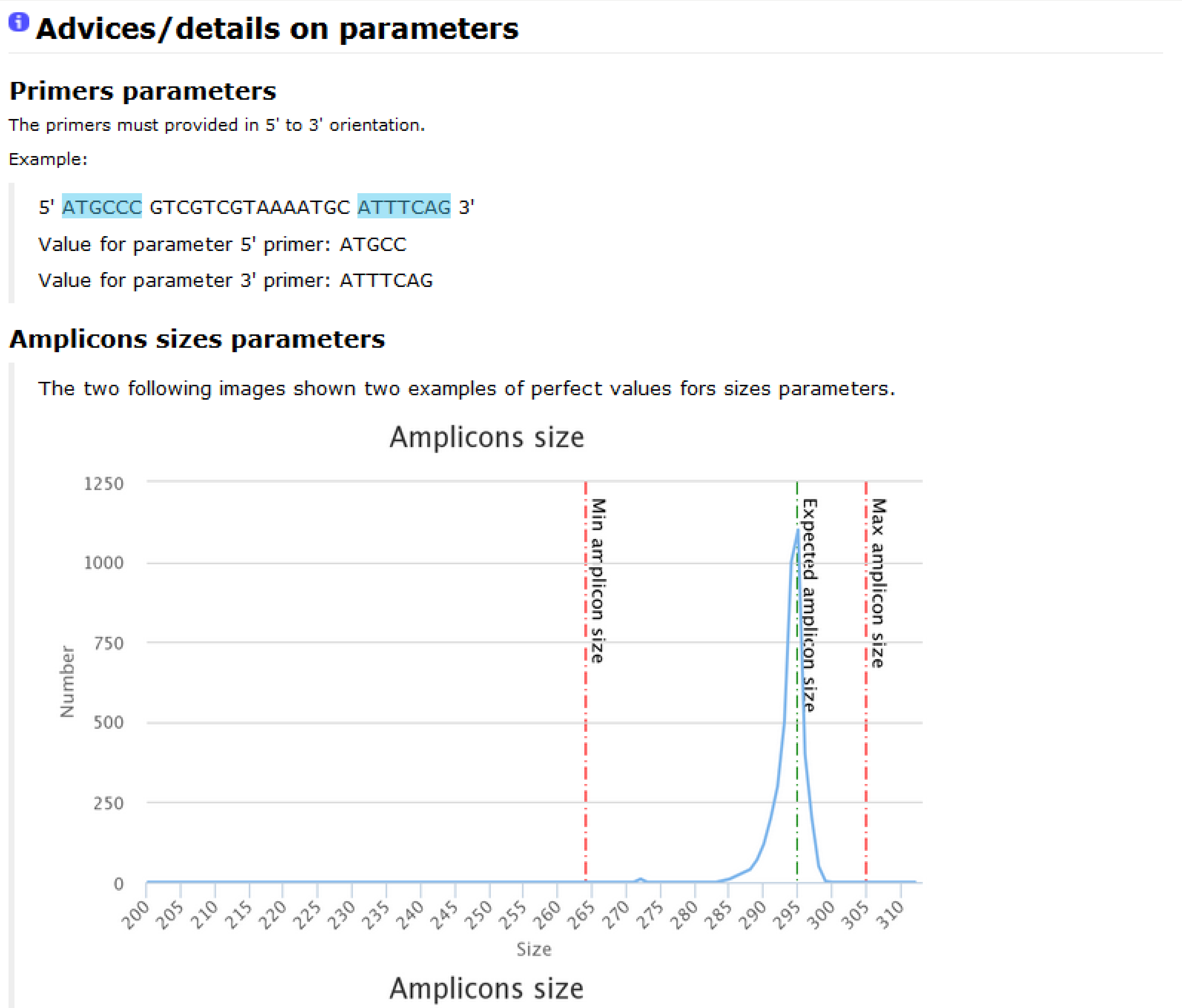
Sample ▲	before process	overlapped	with expected length	with 5' primer	with 3' primer	with expected length (2)	without N
sampleA	90,126	90,126	90,126	89,697	89,697	89,697	89,697
sampleB	213,043	209,801	0	0	0	0	0

Showing 1 to 2 of 2 entries



How it works

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



Advices/details on parameters

Primers parameters

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

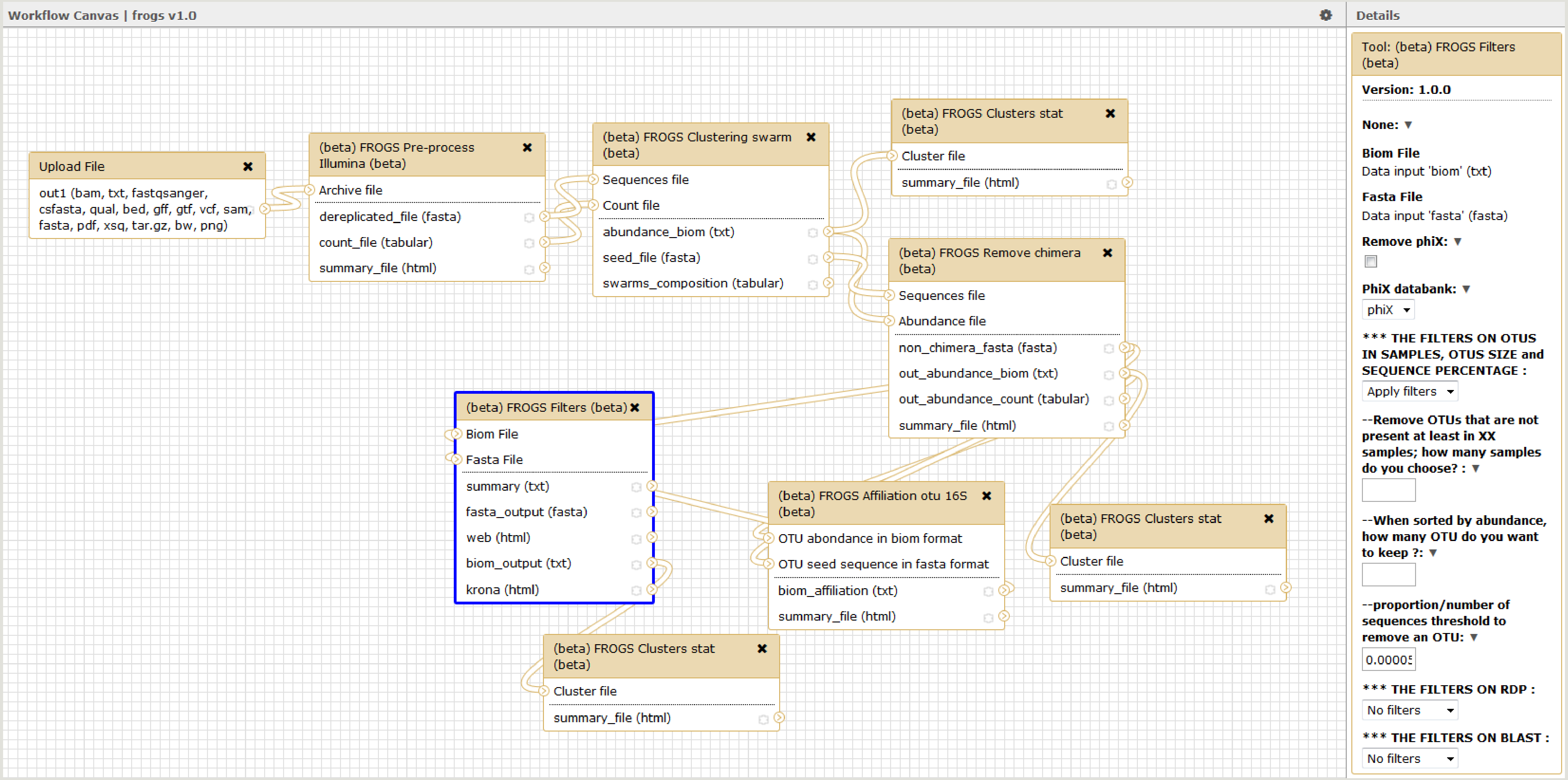
Example:

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

Amplicons sizes parameters

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

Workflow creation



Workflow Canvas | frogs v1.0

Upload File ✖
out1 (bam, txt, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png)

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

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

(beta) FROGS Clusters stat (beta) ✖
Cluster file
summary_file (html)

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

(beta) FROGS Filters (beta) ✖
Biom File
Fasta File
summary (txt)
fasta_output (fasta)
web (html)
biom_output (txt)
krona (html)

(beta) FROGS Affiliation otu 16S (beta) ✖
OTU abondance in biom format
OTU seed sequence in fasta format
biom_affiliation (txt)
summary_file (html)

(beta) FROGS Clusters stat (beta) ✖
Cluster file
summary_file (html)

(beta) FROGS Clusters stat (beta) ✖
Cluster file
summary_file (html)

Details

Tool: (beta) FROGS Filters (beta)

Version: 1.0.0

None: ▼

Biom File
Data input 'biom' (txt)

Fasta File
Data input 'fasta' (fasta)

Remove phiX: ▼
☐

PhiX databank: ▼
phiX ▾

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

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

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

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

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

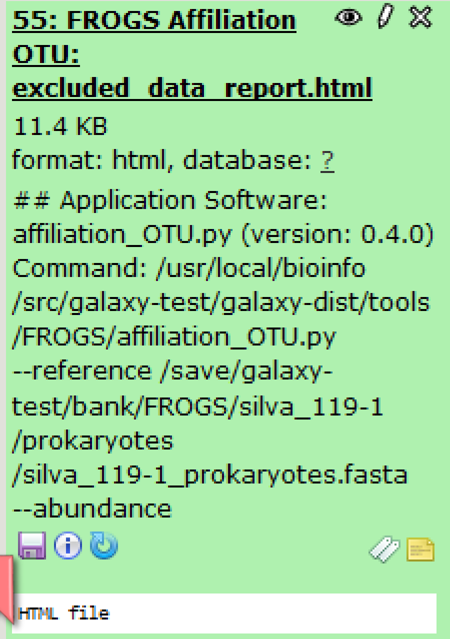
***** THE FILTERS ON BLAST :**
No filters ▾

A vous de jouer ! – 9

EXERCISE 9

Download your data

You have to download one per one your files



55: FROGS Affiliation OTU: excluded_data_report.html

11.4 KB

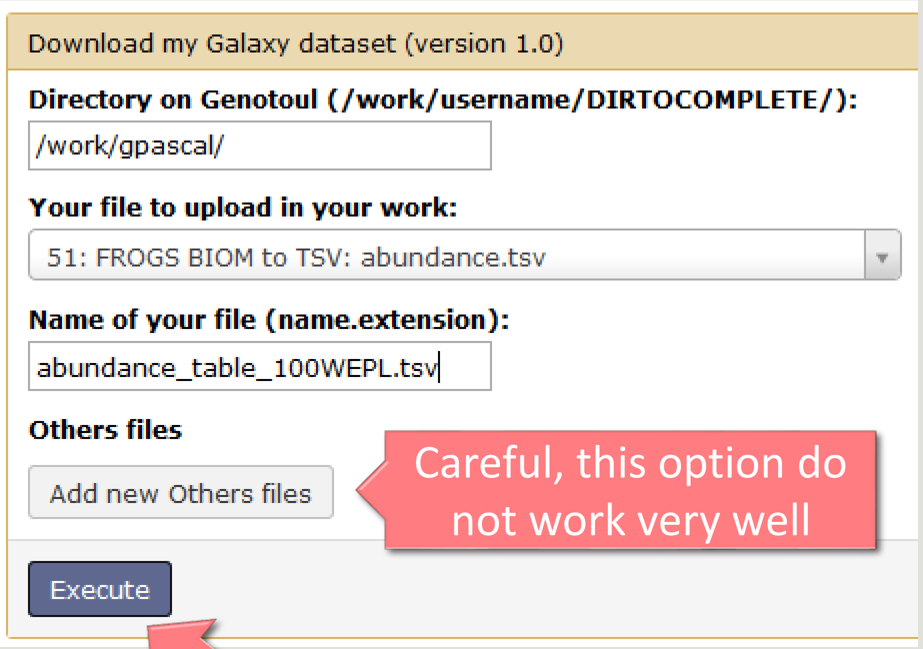
format: html, database: ?

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

HTML file

OR

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



Download my Galaxy dataset (version 1.0)

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

/work/gpascal/

Your file to upload in your work:

51: FROGS BIOM to TSV: abundance.tsv

Name of your file (name.extension):

abundance_table_100WEPL.tsv

Others files

Careful, this option do not work very well

Conclusions

Why Use FROGS ?

User-friendly

Fast

454 data and Illumina data
→ sequencing methods change but same tool

→easier for comparisons

Clustering without global threshold and independent of sequence order

Filters tool

Cluster Stat tool

How to cite FROGS

In waiting for the publication:

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

To contact

FROGS:

geraldine.pascal@toulouse.inra.fr

Or

maria.bernard@jouy.inra.fr

Galaxy:

sigenae-support@listes.inra.fr

Next training sessions

December 2, 3 and 4th 2015 (with a Galaxy day)

Galaxy e-learning (user account)

And soon FROGS e-learning