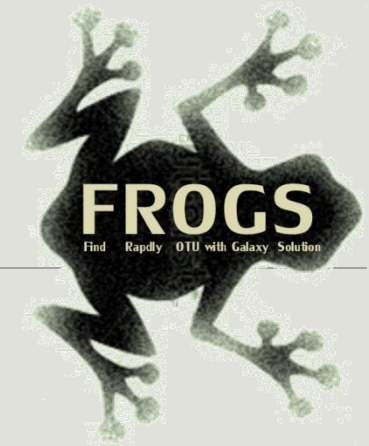


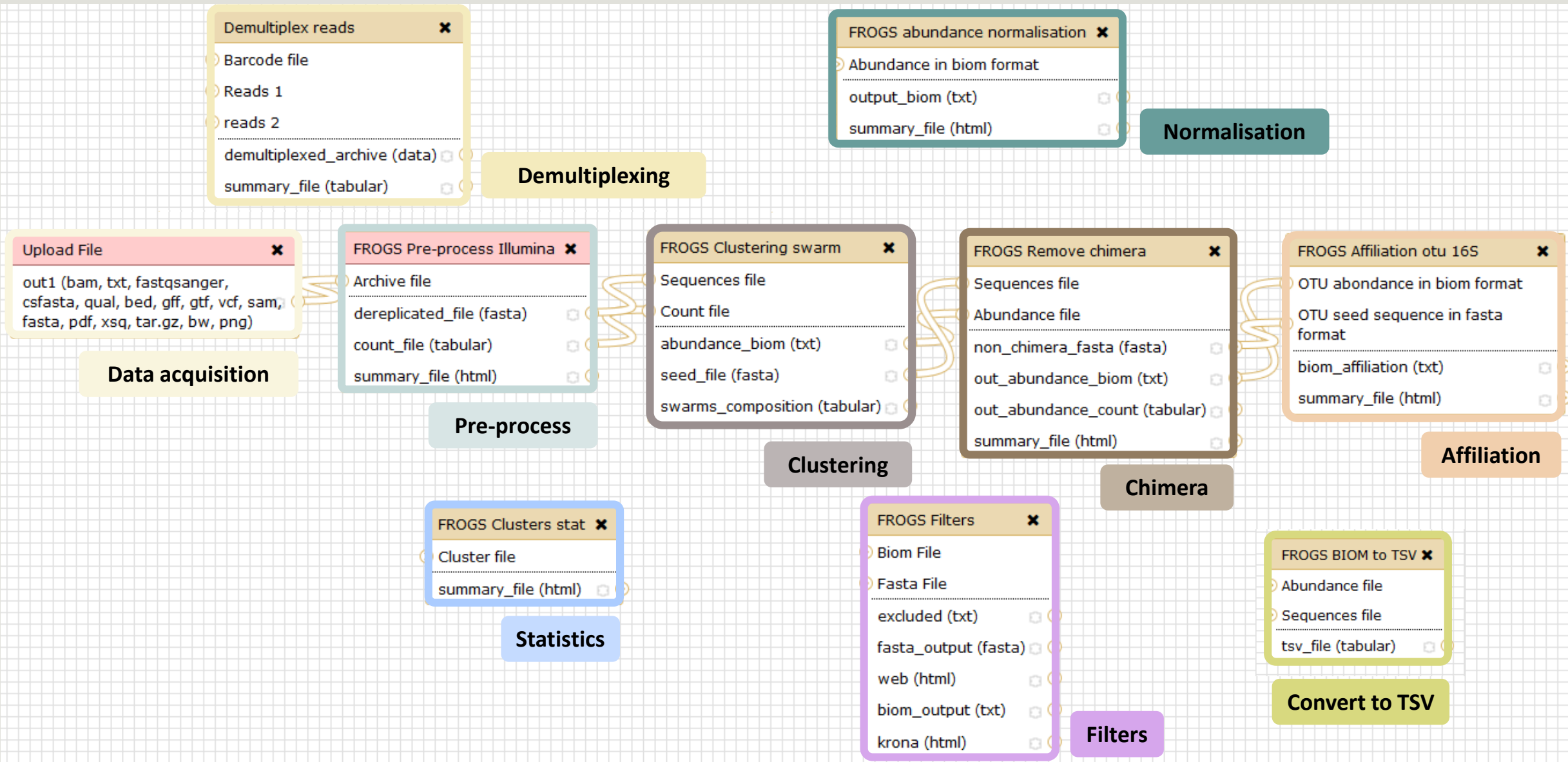
# Exercices: Metagenomics

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.



# Introduction to Galaxy

---

# Galaxy Portal from SIGENAE GENOTOUL

<http://galaxy-workbench.toulouse.inra.fr>

- Tools
- search tools
- YOUR DATA
  - [Upload Data](#)
  - [Download Data](#)
- FILES MANIPULATION
  - [Text Manipulation \(e-learning\)](#)
  - [Filter and Sort](#)
  - [Join, Subtract and Group](#)
  - [Convert Formats](#)
  - [BED Tools](#)
  - [Graph/Display Data](#)
- SEQUENCES MANIPULATION
  - [FASTA manipulation](#)
  - [FASTQ manipulation \(e-learning\)](#)
  - [SAM/BAM manipulation : Picard \(beta\)](#)
  - [SAM/BAM manipulation: SAMtools \(e-learning\)](#)
  - [Fetch Sequences](#)
  - [Sequences Queries](#)
  - [VCF Tools](#)
- SGS MAPPING
  - [BWA - Bowtie \(e-learning\)](#)
  - [BLAT](#)
- SNP / INDEL
  - [NGS: GATK Tools \(beta\)](#)
  - [Indel Analysis](#)
  - [SNP annotation](#)
- TRANSCRIPTOMIC
  - [RNAseq](#)
  - [DEA stats](#)
  - [S-MART](#)
  - [sRNAseq](#)
- CHIP-SEQ
  - [Operate on Genomic Intervals](#)
  - [Nebula](#)
- METAGENOMIC
  - [Clones metagenomiques](#)

## WELCOME TO GALAXY WORKBENCH



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

- Make bioinfo Linux tools accessible to biologists.
- 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](mailto: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 resources thanks to Galaxy instance <http://sigenae-workbench.toulouse.inra.fr>".  
In cases of collaboration, you can directly quote the person who participated to the project : Name, Sigenae group, GenPhySE, INRA Auzeville CS 52627 31326 Castanet Tolosan cedex.

References

X. SIGENAE [<http://www.sigenae.org/>]

### Sigenae e-learning platform

If you need more training about bioinformatic and Galaxy, please connect to [Sigenae e-learning platform](#)

Some of the tools have a direct access to the e-learning platform of sigenae. Those tools will have this  in the help section. Click on this icon to be redirected to the e-learning platform.

[Galaxy](#) is an open, web-based platform for data intensive biomedical research. The [Galaxy team](#) is a part of [BX](#) at [Penn State](#), and the [Biology](#) and [Mathematics and Computer Science](#) departments at [Emory University](#). The [Galaxy Project](#) is supported in part by [NHGRI](#), [NSF](#), [The Huck Institutes of the Life Sciences](#), [The Institute for CyberScience at Penn State](#), and [Emory University](#).

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

## Tools

search tools

## YOUR DATA

[Upload Data](#)[Download Data](#)

## FILES MANIPULATION

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

## SEQUENCES MANIPULATION

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

## SGS MAPPING

[BWA - Bowtie \(e-learning\)](#)[BLAT](#)

## AVAILABLE TOOLS

## WELCOME TO GALAXY WORKBENCH



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

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

**Warnings :****TOOL CONFIGURATION AND EXECUTION**

- 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](mailto:sigenae-support@listes.inra.fr)

If you have some question about Galaxy, please consult your [FAQ](#)

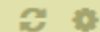
**How to cite Galaxy workbench ?**

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>".

## History



## Unnamed history

0 bytes



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

## DATASETS HISTORY

# Your turn! - 1

---

LAUNCH DEMULTIPLEX READS TOOL

# Upload data

---

# Your turn: exo 1

---



Create the 1st history **multiplexed**

Import files « **multiplex.fastq** » and « **barcode.tabular** » present in the **Genotoul** folder /work/formation/FROGS/



Create the 2nd history **454**

Import file « **454.fastq.gz** » present in the **Genotoul** folder /work/formation/FROGS/



Create the 3rd history **MiSeq R1 R2**

Import files « **sampleA\_R1.fastq** » and « **sampleA\_R2.fastq** » present in the **Genotoul** folder /work/formation/FROGS/

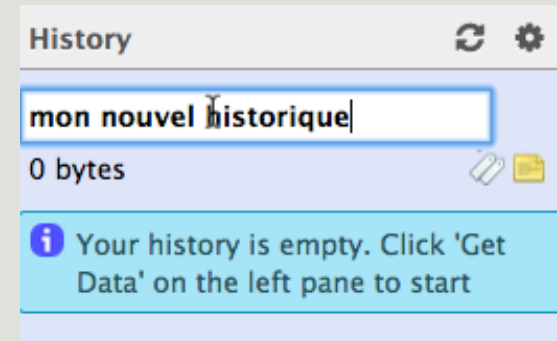
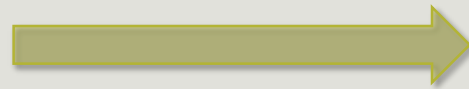
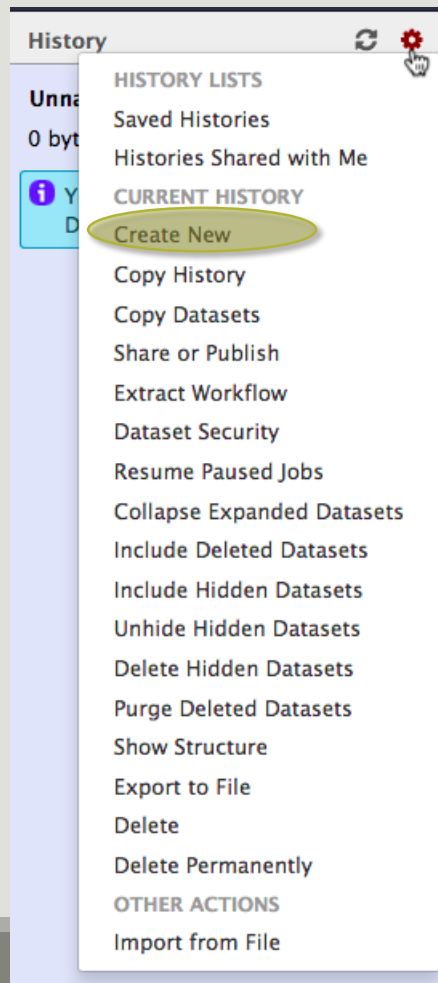


Create the 4th history **MiSeq contiged**

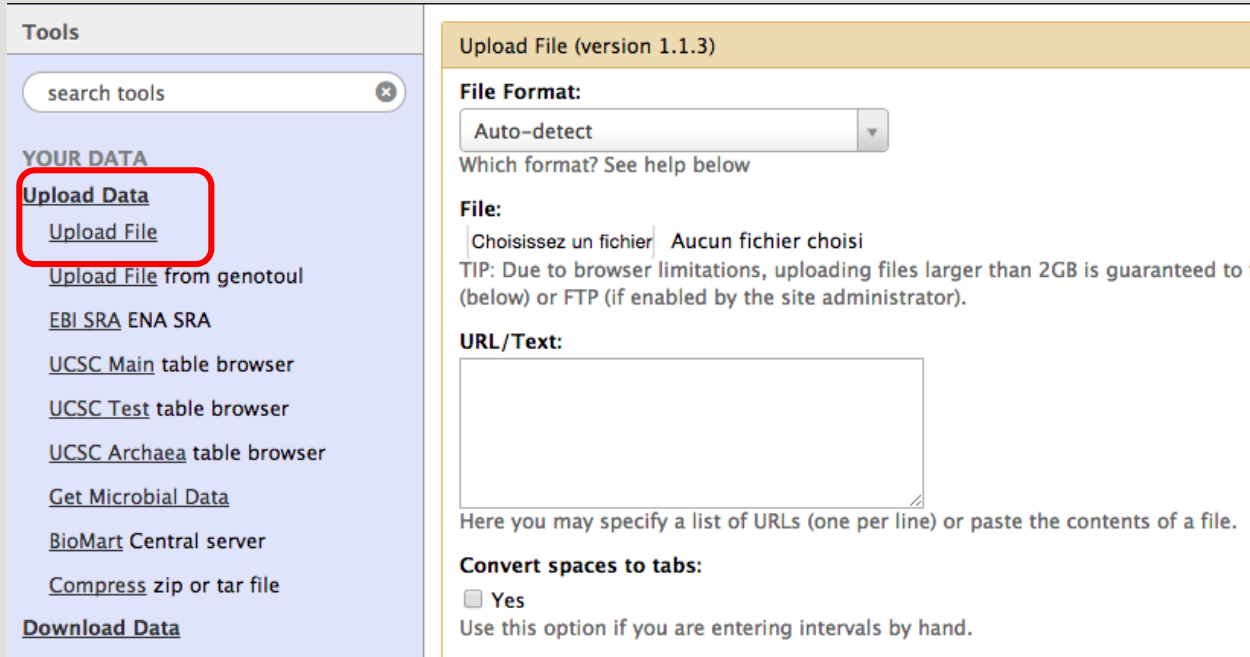
Import archive file « **100spec\_90000seq\_9samples.tar.gz** » present in the **Genotoul** folder /work/formation/FROGS/



# History creation



# Upload data: different methods



The screenshot shows the Galaxy web interface. On the left, under the 'Tools' section, there is a search bar and a list of tools under 'YOUR DATA'. The 'Upload Data' and 'Upload File' links are highlighted with a red box. The main panel shows the 'Upload File (version 1.1.3)' tool configuration. It includes a 'File Format' dropdown menu set to 'Auto-detect', a 'File' section with a file selection button and a tip about browser limitations, a 'URL/Text' text area, and a 'Convert spaces to tabs' checkbox.

Default method, your files are on **your** computer, they are copied on your Galaxy account

You can only upload one file at a time  
→ 10 samples ≥ 10 uploads



Each uploaded file will consume your Galaxy's quota!

# Upload data: different methods

Tools

search tools

YOUR DATA

**Upload Data**

Upload File

**Upload File from genotoul**

EBI SRA ENA SRA

Upload File (version 1.0.0)

**Path to file:**

/work/frogs/Donnees\_simulees/100WEPL\_setA.tar.gz

Path must be like : /work/USERNAME/somewhere/afile

**File type:**

tar.gz

Do not forget to precise the input file type

Execute

Specific SIGENAE GENOTOUL method. It allows you to access to your files in **your work account** on the Genotoul without consuming your Galaxy quota.

And if you have multiple samples ?

See [How to create an archiveTAR.ppt](#)



How to transfer files on /work of Genotoul?

See [How to transfert to genotoul.ppt](#)

# Upload data: different methods

**Tools**

**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](#) Step 1 in metagenomics analysis (16S/18S): denoising and dereplication.

**Upload archive (version 1.0.0)**

**File:**  
 Aucun fichier choisi  
TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method.

**URL:**  
  
Here you may specify the archive URL.

**i What it does**

If you have an **archive** on your **own computer**, you will use this specific FROGS tool to upload your samples archive instead of the default « Upload File » of Galaxy.

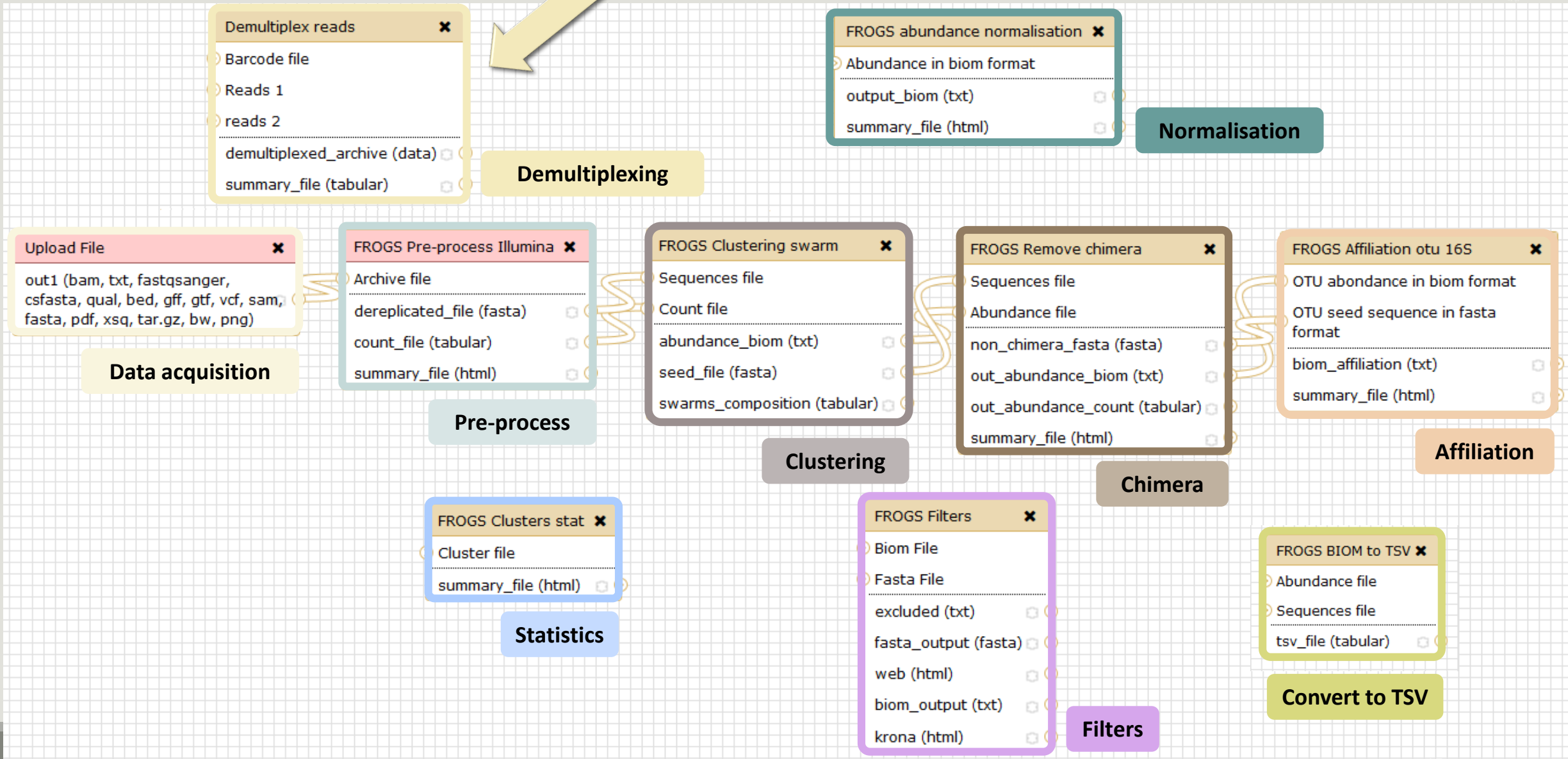
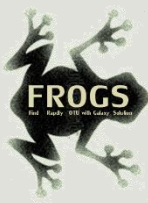
# Your turn! - 2

---

LAUNCH DEMULTIPLEX READS TOOL

# Demultiplexing tool

---



The tool parameters depend on the input data type

### Demultiplex reads (version 1.0.0)

#### Barcode file:

This file describes barcodes and samples (one line by sample). Line format : 'SAMPLE\_NAME BARCODE'.

#### Single or Paired-end reads:

Select between paired and single end data

#### Select fastq dataset:

Specify dataset with single reads

#### barcode mismatches:

Number of mismatches allowed in barcode

#### barcode on which end ?:

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

Execute

### Demultiplex reads



Barcode file

Reads 1

reads 2

demultiplexed\_archive (data)



summary\_file (tabular)



Demultiplexing

### Demultiplex reads (version 1.0.0)

#### Barcode file:

This file describes barcodes and samples (one line by sample). Line format : 'SAMPLE\_NAME BARCODE'.

#### Single or Paired-end reads:

Select between paired and single end data

#### Select first set of reads:

Specify dataset with forward reads

#### Select second set of reads:

Specify dataset with reverse reads

#### barcode mismatches:

Number of mismatches allowed in barcode

#### barcode on which end ?:

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

Execute



# Your turn: exo 2

---

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

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

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

# Your turn: exo 2

---

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

Data are single end reads  
→ only 1 fastq file

Samples are characterized by an association of two barcodes in forward and reverse strands  
→ multiplexing « both ends »

```
2: /work/frogs     
/Formation/multiplex.fastq
```

---

```
1: /work/frogs     
/Formation/barcode.txt
```

# Your turn: exo 2

---

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

1. Play with pictograms



2. Look at the stdout, stderr when available (in the « i » pictogram )

Demultiplex reads (version 1.0.0)

**Barcode file:**  
14: /work/frogs/Formation/barcode.tabular  
This file describes barcodes and samples (one line by sample). Line format : 'SAMPLE\_NAME BARCODE'.

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

**Select fastq dataset:**  
7: /work/frogs/Formation/multiplex.fastq ▾  
Specify dataset with single reads

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

**barcode on which end ?:**  
Both ends ▾  
The barcode is at the beginning of the forward end or of the reverse end or both?

Execute

**History**

**barcode\_formation**  
4.2 MB

**14: /work/frogs /Formation/barcode.txt**  
10 lines  
format: tabular, database: ?  
Epilog : job finished at Tue Jun 16 14:14:52 CEST 2015

1	2	3
MgArd0001	ACAGCGT	TGTACGT
MgArd0009	ACAGTAG	TGTACGT
MgArd0017	ACGTCAG	TGTACGT
MgArd0029	ACTCAGT	TGTACGT
MgArd0038	ACTCGTC	TGTACGT
MgArd0046	AGCAGTC	TGTACGT




**7: /work/frogs /Formation/multiplex.fastq**




# Advices




---

- 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 oftently 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 correspond to several samples. So these sequences are non-affected to a sample.

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

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

# Your turn ! - 3

---

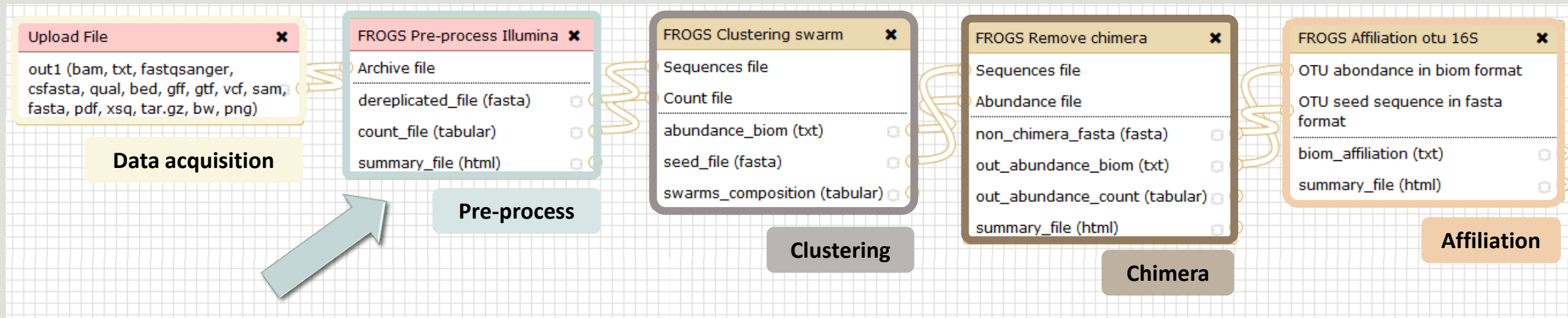
LAUNCH THE PRE-PROCESS READS TOOL

# Pre-process tool

---



# FROGS pipeline



454

# Your turn: exo 3.1

---

Go to « 454 » history

Launch the pre-process tool on that data set

→ objective : understand the parameters

454

FROGS Pre-process (version 1.2.0)

**Sequencer:**

454 ▾

Select the sequencer family used to produce the sequences.

**Input type:**

One file by sample ▾

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

**Samples**

**Samples 1**

**Name:**

sample454

The sample name.

**Sequence file:**

29: /work/frogs/Formation/454.fastq ▾

FASTQ file of sample.

Add new Samples

**Minimum amplicon size:**

340

The minimum size for the amplicons (with primers).

**Maximum amplicon size:**

450

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).

Execute

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

# Your turn: exo 3.1

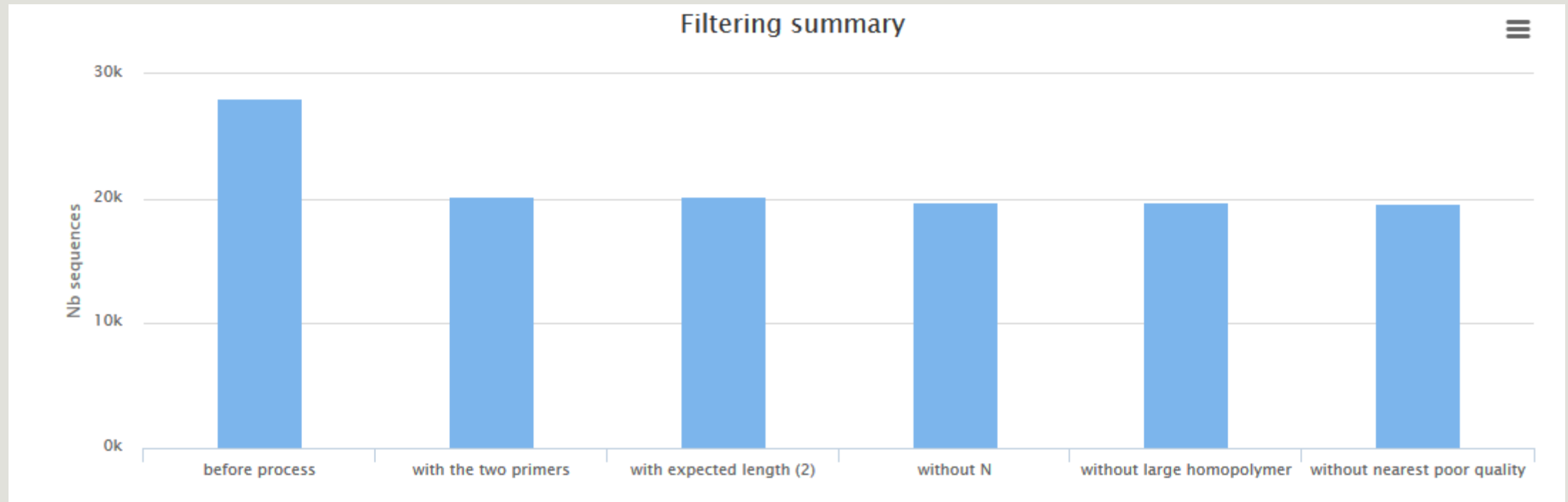
---

What does the « dereplicated.fasta » file ? 

What does the « count.tsv » file ? 

Explore the file « excluded\_data.html » 

Sample	before process	with the two primers	with expected length (2)	without N	without large homopolymer	without nearest poor quality
sample454	28,009	20,227	20,227	19,753	19,746	19,651



To be kept,  
sequences  
must have  
the 2 primers

# Your turn: exo 3.2

---

Go to « [MiSeq R1 R2](#) » history

Launch the pre-process tool on that data set

→ objective: understand flash software

**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:**

sampleA

The sample name.

**Reads 1:**

33: /work/frogs/Formation/sampleA\_R1.fastq

R1 FASTQ file of paired-end reads.

**reads 2:**

34: /work/frogs/Formation/sampleA\_R2.fastq

R2 FASTQ file of paired-end reads.

Add new Samples

**Reads 1 size:**

250

The read1 size.

**Reads 2 size:**

250

The read2 size.

Primers used for this sequencing :  
 Forward: CCGTCAATTC  
 Reverse: CCGCNGCTGCT  
 Lecture 5' → 3'

>ERR619083.M00704

**CCGTCAATTC**ATTGAGTTTCAACCTTGCGGCCGTACTTCCCAGGCGGTACGTT  
 TATCGCGTTAGCTTCGCCAAGCACAGCATCCTGCGCTTAGCCAACGTACATCG  
 TTTAGGGTGTGGACTACCCGGGTATCTAATCCTGTTTCGCTACCCACGCTTTCG  
 AGCCTCAGCGTCAGTGACAGACCAGAGGCCGCTTTCGCCACTGGTGTTCCTC  
 CATATATCTACGCATTTCCACCGCTACACATGGAATTCCACTCTCCCCTTCTGC  
 ACTCAAGTCAGACAGTTTCCAGAGCACTCTATGGTTGAGCCATAGCCTTTTAC  
 TCCAGACTTTCCTGACCGACTGCACTCGCTTTACGCCAATAAATCCGGACAA  
 CGCTTGCCACCTACGTATTA**CCGCNGCTGCT**



**Expected amplicon size:**

410

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

**Minimum amplicon size:**

340

The minimum size for the amplicons (with primers).

**Maximum amplicon size:**

450

The maximum size for the amplicons (with primers).

**5' primer:**

CCGTCAATTC

The 5' primer sequence (wildcards are accepted).

**3' primer:**

CCGCNGCTGCT

The 3' primer sequence (wildcards are accepted).

Execute

# Your turn: exo 3.2

---

Interpret the « excluded\_data.html » file.



# Your turn: exo 3.3

---

Go to« [MiSeq contiged](#) » history

Launch the pre-process tool on that data set

→ objective: understand output files

# Your turn: exo 3.3

---

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

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

## Your turn: exo 3.3

---

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

- (1) filtering out sequences with a N,
- (2) keeping only type species
- (3) with a match for the forward (ACGGRAGGCAGCAG) and reverse (TACCAGGGTATCTAATCCT) primers in the V3-V4 region and
- (4) maximizing the phylogenetic diversity (PD) for a given database size. The PD was computed from the NJ tree distributed with the LTP.”

# MiSeq 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

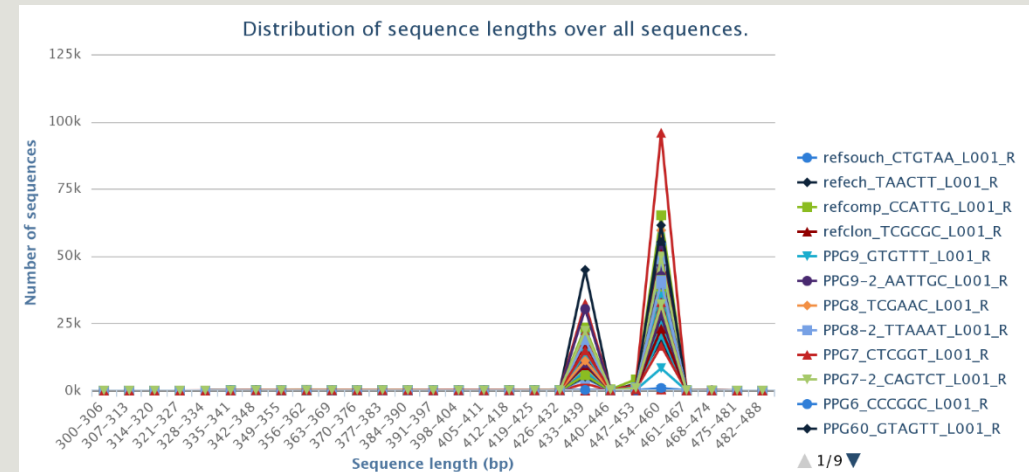
The 5' primer sequence (wildcards are accepted).

### 3' primer:

AGGATTAGATACCCTGGTA

The 3' primer sequence (wildcards are accepted).

Execute



Primers used for this sequencing :  
5' primer: ACGGGAGGCAGCAG  
3' primer: AGGATTAGATACCCTGGTA  
Lecture 5' → 3'

# Your turn: exo 3.3 - Questions

---

1. How many sequences are there in the file ?
2. How many sequences did not have the 5' primer?
3. How many sequences still are after pre-processing the data?
4. How much time did it take to pre-process the data ?

# Your turn ! - 4

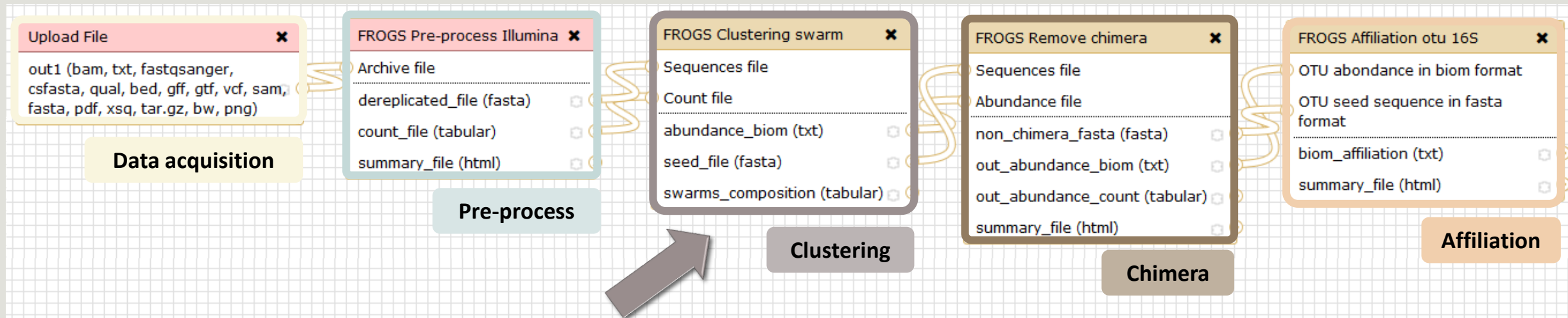
---

LAUNCH CLUSTERING AND CLUSTERSTAT TOOLS

# Clustering tool

---

# FROGS pipeline





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

Execute



1st run for denoising:

Swarm with  $d = 1$  → high OTUs definition  
linear complexity

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

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

Gain time !

Remove false positives !

# Your turn: exo 4

---

Go to « [MiSeq contiged](#) » history

Launch the Clustering SWARM tool on that data set

→ objectives :

- understand the denoising efficiency

- understand the ClusterStat utility

# Your turn: exo 4

---

1. Launch FROGS Clustering with  $d = 3$  and with denoising option checked
  - a. How much time does it take to finish?
  - b. How many clusters do you get ?

# Your turn: exo 4

3. Edit the biom and fasta output dataset by adding **d1d3**



Attributes Convert Format Datatype Permissions

Edit Attributes

**Name:**  
warm: seed\_sequencesd1d3.fasta

**Info:**  
## Application  
Software :/usr/local/bioinfo  
/src/galaxy-test/galaxy-

**Annotation / Notes:**  
Add an annotation or notes to a dataset; annotations are available when a history is viewed.

**Database/Build:**  
unspecified (?)

Save

Auto-detect  
This will inspect the dataset and attempt to correct the above column values if they are not accurate.

# Your turn: exo 4

2. Use the FROGS ClusterStat tool
3. Interpret the boxplot: **Clusters size summary**
4. Interpret the table: **Clusters size details**
5. What can we say by observing the sequence distribution?
6. How many clusters share “sampleB3” with at least one other sample?
7. How many clusters could we expect to be shared ?
8. How many sequences represent the 668 specific clusters of “sampleC2”?
9. This represents what proportion of “sampleC2”?
10. What do you think about it?
11. How do you interpret the « Hierarchical clustering » ?

FROGS Clusters stat Process  
some metrics on clusters.

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).

# Your turn ! - 5

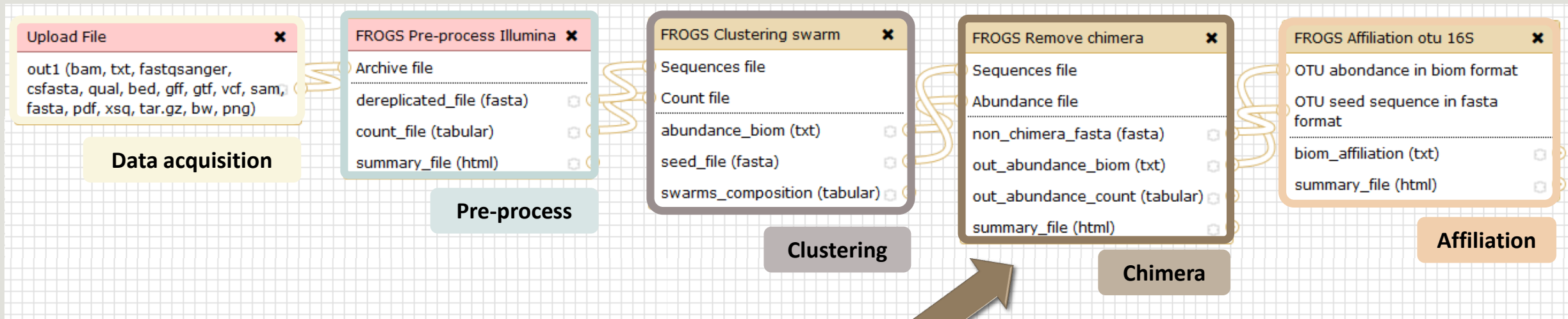
---

LAUNCH THE REMOVE CHIMERA TOOL

# Removing chimera tool

---

# FROGS pipeline



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



FROGS Remove chimera ✕

Sequences file

Abundance file

---

non\_chimera\_fasta (fasta) 🗑️

out\_abundance\_biom (txt) 🗑️

out\_abundance\_count (tabular) 🗑️

summary\_file (html) 🗑️

Chimera

FROGS Remove chimera (version 1.0.0)

**Sequences file:**

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

The sequences file.

**Abundance type:**

BIOM file ▾

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

**Abundance file:**

5: FROGS Clustering swarm: abundance.biom ▾

It contains the count by sample for each sequence.

Execute

# Your turn: exo 5

---

Go to « [MiSeq contiged](#) » history

Launch the « FROGS Remove Chimera » tool

Follow by the « FROGS ClusterStat » tool on the d1d3 biom and fasta

→ objectives :

- understand the efficiency of the chimera removal

- make links between small abundant OTUs and chimeras

# Your turn: exo 5

---

1. Understand the « `excluded_data_reportd1d3.html` »
  - a. How many clusters are kept after chimera removal?
  - b. How many sequences that represent ? So what abundance?
  - c. What do you conclude ?

# Your turn: exo 5

---

2. Launch « FROGS ClusterStat » tool  
on non\_chimera\_abundanced1d3.biom
3. Rename outputs in summarynonchimerad1d3.html
4. Compare the HTML files
  - a. Of what are mainly composed singleton are weakly abundant OTUs ?
  - b. What sequence abundances are they representing ?
  - c. What do you conclude ?

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

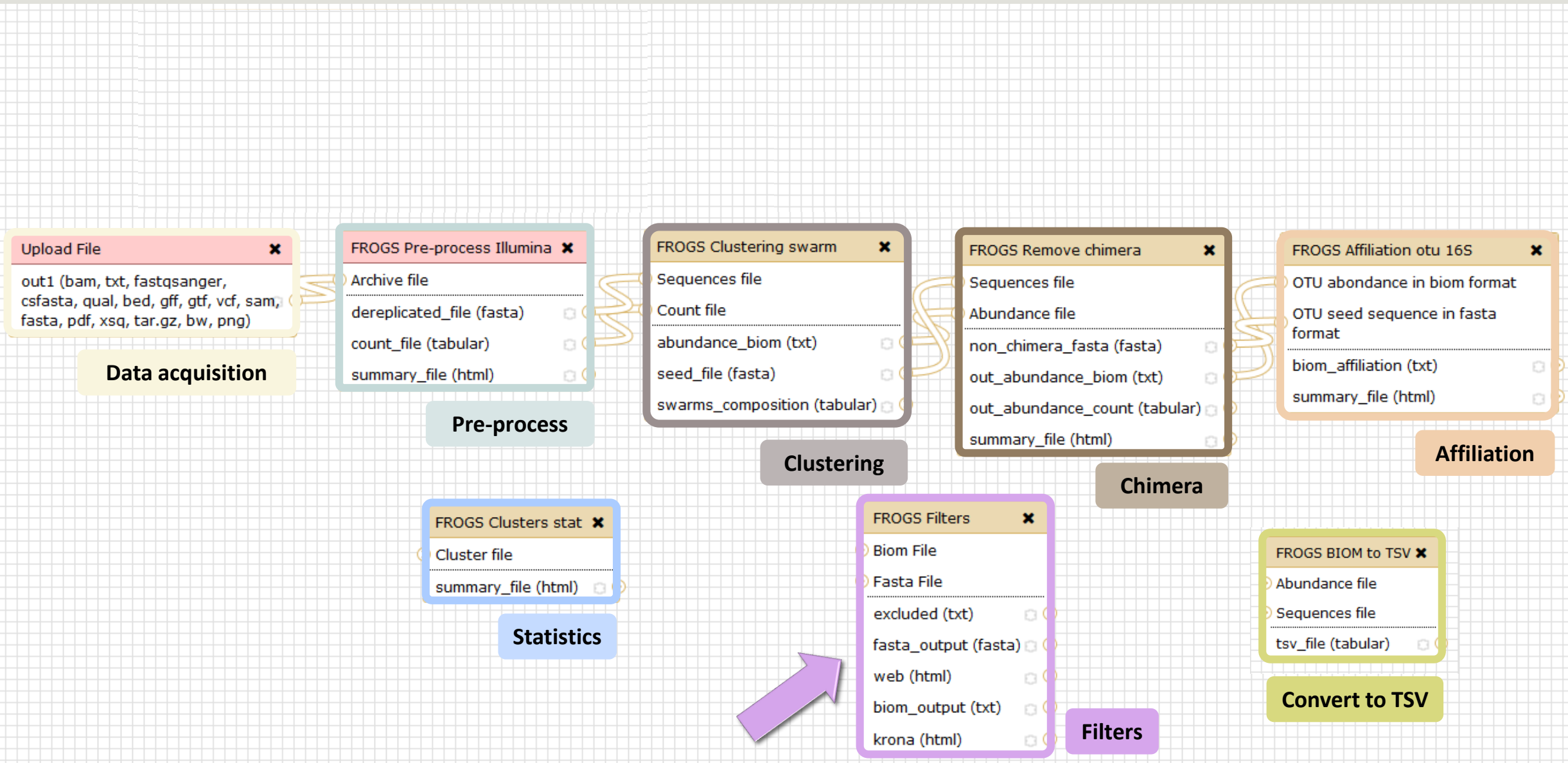
# Your turn ! - 6

---

LAUNCH DE LA TOOL FILTERS

# Filters tool

---



# Filters

---

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

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

After Affiliation tool



**FROGS Filters** ✕

- Biom File
- Fasta File
- excluded (txt)
- fasta\_output (fasta)
- web (html)
- biom\_output (txt)
- krona (html)

Filters

4 filter sections

FROGS Filters (version 1.0.0)

**Biom File:**

**Fasta File:**

**Remove phiX:**  
  
 Remove phiX sequences before affiliation.

**PhiX databank:**  
  
 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? :  
  
 Fill the field only if you want this treatment

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

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

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

Apply filters

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

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

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

Apply filters

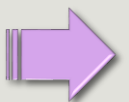
--Minimum blast length:  
  
 Fill the field only if you want this treatment

--Maximum e value (between 0 and 1):  
  
 Fill the field only if you want this treatment













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

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

Input



Output

- 38: FROGS Filters:** [krona.html](#)   
- 37: FROGS Filters:** [abundance table.biom](#)   
- 36: FROGS Filters:** [summary.html](#)   
- 35: FROGS Filters:** [seed.fasta](#)   

# Your turn: exo 6

---

- I. Go to history« **MiSeq contiged** »
- II. Launch « Filters » tool with non\_chimera\_abundanced1d3.biom, non\_chimerad1d3.fasta
- III. Apply 2 filters **--proportion/number of sequences threshold to remove an OTU: 0.00005\***  
and **--Remove OTUs that are not present at least in XX samples; how many samples do you choose? : 3**

→ objective : play with filters, understand their impact on false-positives OTUs

\*Nat Methods. 2013 Jan;10(1):57-9. doi: 10.1038/nmeth.2276. Epub 2012 Dec 2.

**Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing.**

Bokulich NA<sup>1</sup>, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, Mills DA, Caporaso JG.

# Your turn: exo 6

---

1. What are the output files of “Filters” ?
2. Explore summary.html file.
3. How many OTUs have you removed with the filter “0.00005 ” ?
4. How many OTUs have you removed with the filter “Remove OTUs that are not present at least in 3 samples”?
5. How many sequences represent these for each of the filters?
6. How many OTUs do they remain ?
7. Build the Venn diagram on the two filters.
8. What you says krona.html ?

# Your turn ! - 7

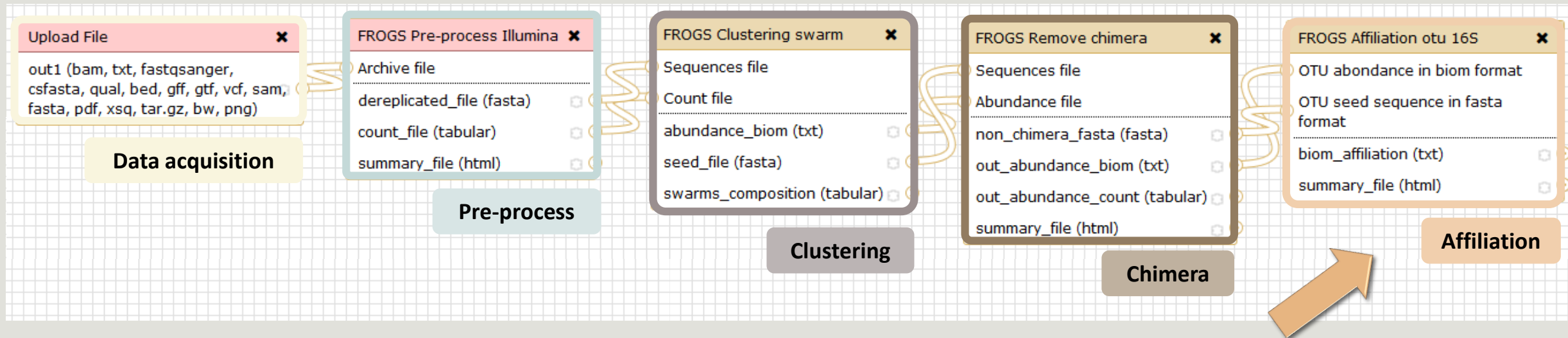
---

LAUNCH THE « FROGS AFFILIATION » TOOL

# Affiliation tool

---

# FROGS pipeline



# Affiliation

---

2 methods used on one reference database, here SILVA 119 (16S or 18S):

- RDP Classifier (Ribosomal Database Project) \*
- NCBI Blast+ \*\*

RDP Classifier affiliation characteristics:

- Bootstrap value for each taxonomic subdivision

NCBI Blast+ affiliation characteristics:

- identity %
- coverage %
- e-value
- alignment length

\* 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

FROGS Affiliation otu 16S ✕

- OTU abundance in biom format
- OTU seed sequence in fasta format

---

- biom\_affiliation (txt) ⊞
- summary\_file (html) ⊞

**Affiliation**

FROGS Affiliation OTU (version 0.4.0)

**Using reference database:**  
 ▼  
Select reference from the list

**OTU abundance in biom format:**  
 ▼  
Select your biom abundance file which contain the OTU you want to affiliate

**OTU seed sequence in fasta format:**  
 ▼  
Select your OTU's seed fasta file



# Your turn: exo 7

---

Go to « **MiSeq contiged** » history

Launch the « FROGS Affiliation » tool with

- silva\_119-1\_prokaryotes
- abundance\_tabled1d3.biom
- seed1d3.fasta


→ objectives :

understand abundance tables columns

understand the RDP and BLAST affiliation complementarity

# Your turn: exo 7

---

1. What are the « FROGS Affiliation » output files ?
2. How many sequences are affiliated by BLAST ?
3. Click on the « eye » button on the BIOM output file, what do you understand ? 
4. Use the Biom\_to\_TSV tool on this last file and click again on the "eye" on the new output generated, on what correspond the columns ?

# Your turn: exo 7

## 5. Compare RDP and Blast affiliations

#rdp_tax_and_bootstrap	blast_subject	blast_evalue	blast_len	blast_perc_query_coverage	blast_perc_identity	blast_taxonomy
Bacteria;(1.0);Fibrobacteres;(1.0);Fibrobacteria;(1.0);Fibrobacterales;(1.0);Fibrobacteraceae;(1.0);Fibrobacter;(1.0);Fibrobacter succinogenes subsp. succinogenes S85;(1.0);	<a href="#">JX218783.1.1459</a>	0.0	360	89.78	99.72	Root;Bacteria;Fibrobacteres;Fibrobacteria;Fibrobacterales;Fibrobacteraceae;Fibrobacter;unknown species

# Blast JX218783.1.1459 vs our OTU

OTU length : 401

Excellent blast but no matches at the end of OTU. Chimera ?

Uncultured rumen bacterium clone MXMP-H11 16S ribosomal RNA gene, partial sequence  
Sequence ID: [gbJX218783.1](#) Length: 1459 Number of Matches: 1

Range 1: 334 to 693 [GenBank](#) [Graphics](#) ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
660 bits(357)	0.0	359/360(99%)	0/360(0%)	Plus/Plus
Query 1	TAGGGAATATTGCACAATGGGGGAAACCCTGATGCAGCAACGCCACGTGTGGGAAGAAAC	60		
Sbjct 334	TAGGGAATATTGCACAATGGGGGAAACCCTGATGCAGCAACGCCACGTGTGGGAAGAAGC	393		
Query 61	ATTCGGTGTGTAACCCTGTCATGAGGGAATAAGGCCCGCCTTCGGGCGGGATTGAAT	120		
Sbjct 394	ATTCGGTGTGTAACCCTGTCATGAGGGAATAAGGCCCGCCTTCGGGCGGGATTGAAT	453		
Query 121	GTACCTTGAGAGGAAGCACCGGCAAACCTTCGTGCCAGCAGCCGCGGTAATACGAGGGGTG	180		
Sbjct 454	GTACCTTGAGAGGAAGCACCGGCAAACCTTCGTGCCAGCAGCCGCGGTAATACGAGGGGTG	513		
Query 181	CAAGCGTTGTTTCGGAATTACTGGGCGTAAAGGGAGCGTAGGCGGAGATTCAAGCGGATTG	240		
Sbjct 514	CAAGCGTTGTTTCGGAATTACTGGGCGTAAAGGGAGCGTAGGCGGAGATTCAAGCGGATTG	573		
Query 241	TACAATCCCGGGGCCAACCCCGGCTCTGCAGTCCGAACTGGATCTCTTGGATAGTTCAG	300		
Sbjct 574	TACAATCCCGGGGCCAACCCCGGCTCTGCAGTCCGAACTGGATCTCTTGGATAGTTCAG	633		
Query 301	GGCAGGCGGAATTCCTGGTGTAGCGGTGGAATGCGTAGAGATCAGGAAGAACACCGATG	360		
Sbjct 634	GGCAGGCGGAATTCCTGGTGTAGCGGTGGAATGCGTAGAGATCAGGAAGAACACCGATG	693		

# What do you think about this case ?

---

#rdp_tax_and_bootstrap	blast_subject	blast_evalue	blast_len	blast_perc_query_coverage	blast_perc_identity	blast_taxonomy
Bacteria;(1.0);Proteobacteria;(1.0);Alphaproteobacteria;(1.0);Caulobacterales;(1.0);Hyphomonadaceae;(1.0);Henriciella;(1.0);Henriciella marina;(0.18);	AQXT01000002.1569233.1570666	0.0	401	100.0	100.0	Root;Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Hyphomonadaceae;Henriciella;Henriciella marina DSM 19595

# Your turn ! - 8

---

LAUNCH NORMALIZATION TOOL

# You turn : exo 8

---

1. Normalize your data from Clustering

# Your turn ! - 9

---

CREATE YOUR OWN WORKFLOW !

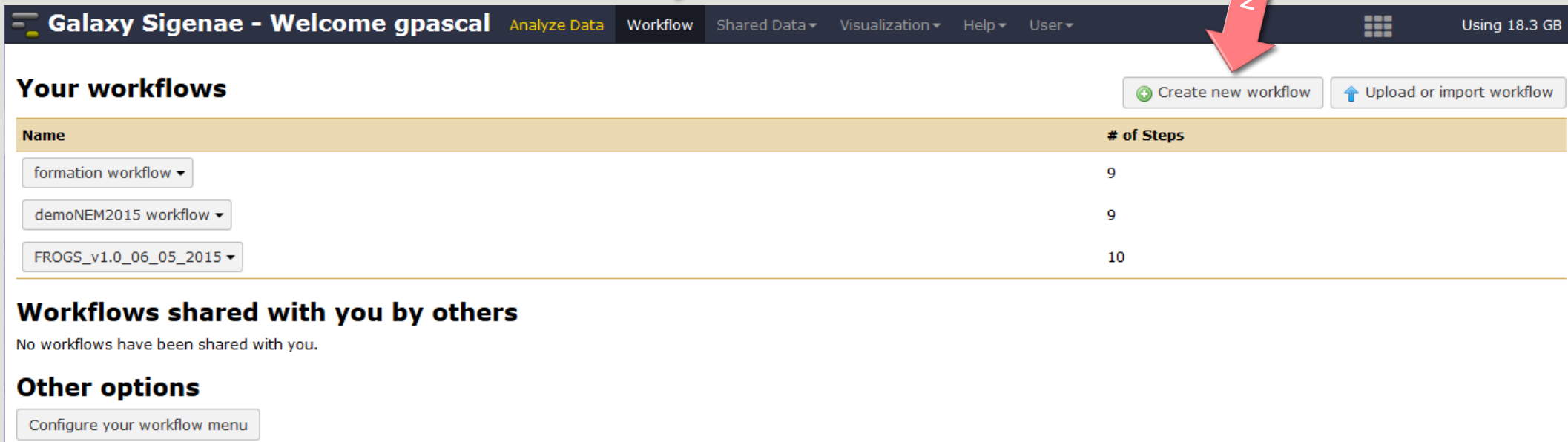


# Workflow creation

---

MiSeq  
contiged

# Your turn: exo 9



The screenshot shows the Galaxy Sigenae interface. The top navigation bar includes the following items: Galaxy Sigenae - Welcome gpascal, Analyze Data, Workflow (highlighted with a red arrow labeled '1'), Shared Data, Visualization, Help, and User. On the right side of the navigation bar, there is a grid icon and the text 'Using 18.3 GB'. Below the navigation bar, the 'Your workflows' section is visible. It contains two buttons: 'Create new workflow' (with a plus icon) and 'Upload or import workflow' (with an upload icon), the latter of which is highlighted with a red arrow labeled '2'. Below these buttons is a table with two columns: 'Name' and '# of Steps'. The table lists three workflows: 'formation workflow' (9 steps), 'demoNEM2015 workflow' (9 steps), and 'FROGS\_v1.0\_06\_05\_2015' (10 steps). Below the table, there is a section titled 'Workflows shared with you by others' with the text 'No workflows have been shared with you.' and a section titled 'Other options' with a button 'Configure your workflow menu'.

**Galaxy Sigenae - Welcome gpascal** Analyze Data **Workflow** Shared Data Visualization Help User Using 18.3 GB

### Your workflows

[+ Create new workflow](#) [↑ Upload or import workflow](#)

Name	# of Steps
formation workflow	9
demoNEM2015 workflow	9
FROGS_v1.0_06_05_2015	10

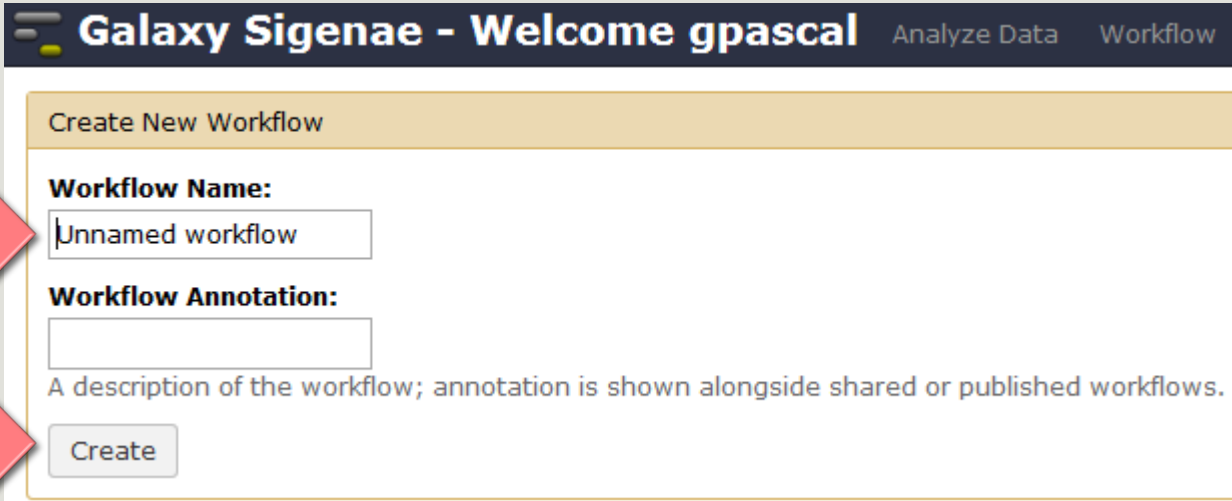
### Workflows shared with you by others

No workflows have been shared with you.

### Other options

[Configure your workflow menu](#)

# Your turn: exo 9



Galaxy Sigenae - Welcome gpascal Analyze Data Workflow

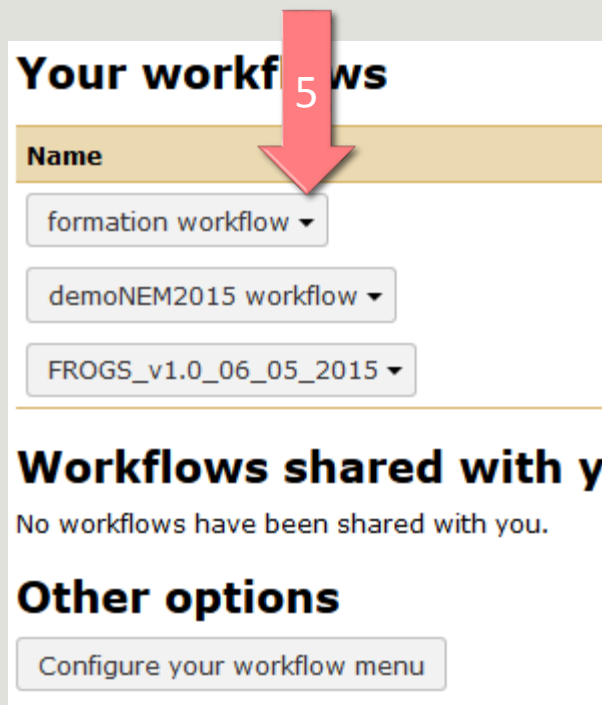
Create New Workflow

**Workflow Name:**

**Workflow Annotation:**

A description of the workflow; annotation is shown alongside shared or published workflows.

# Your turn: exo 9



**Your workflows**

**Name**

formation workflow ▾

demoNEM2015 workflow ▾

FROGS\_v1.0\_06\_05\_2015 ▾

---

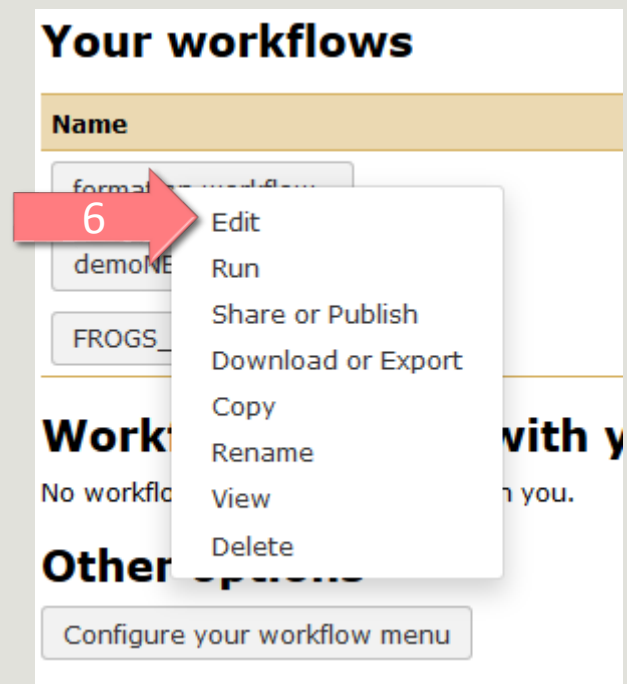
**Workflows shared with you**

No workflows have been shared with you.

**Other options**

Configure your workflow menu

A red arrow with the number 5 points to the first workflow name 'formation workflow'.



**Your workflows**

**Name**

formation workflow ▾

demoNEM2015 workflow ▾

FROGS\_v1.0\_06\_05\_2015 ▾

---

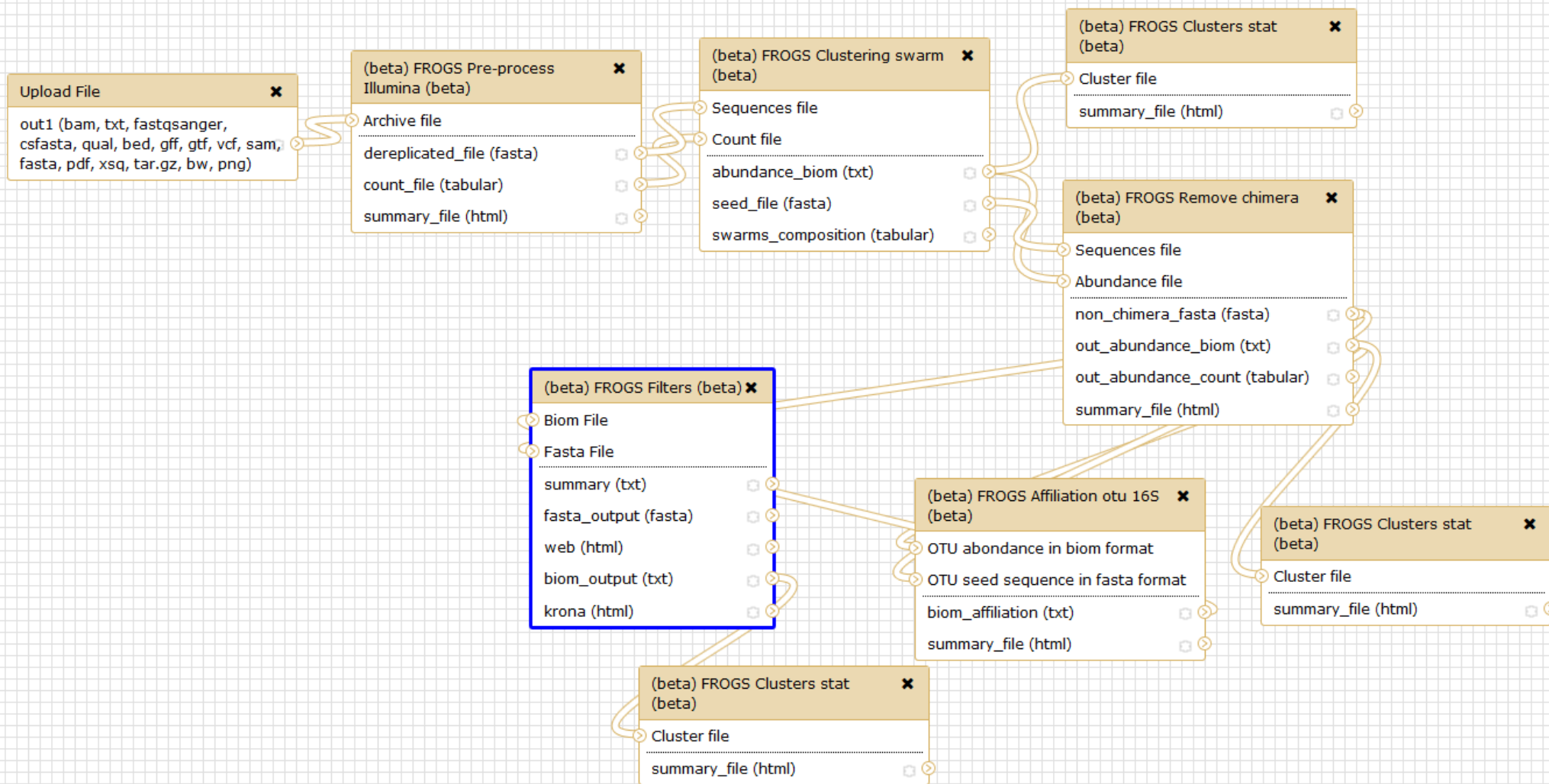
**Workflows shared with you**

No workflows have been shared with you.

**Other options**

Configure your workflow menu

A red arrow with the number 6 points to the first workflow name 'formation workflow', which has a context menu open over it. The menu items are: Edit, Run, Share or Publish, Download or Export, Copy, Rename, View, and Delete.



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: ▾

0.0000 ▾

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

No filters ▾

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

No filters ▾