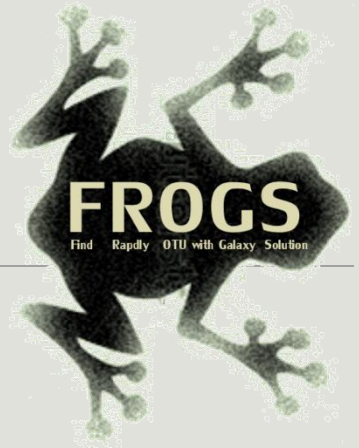
Exercises: Metagenomics

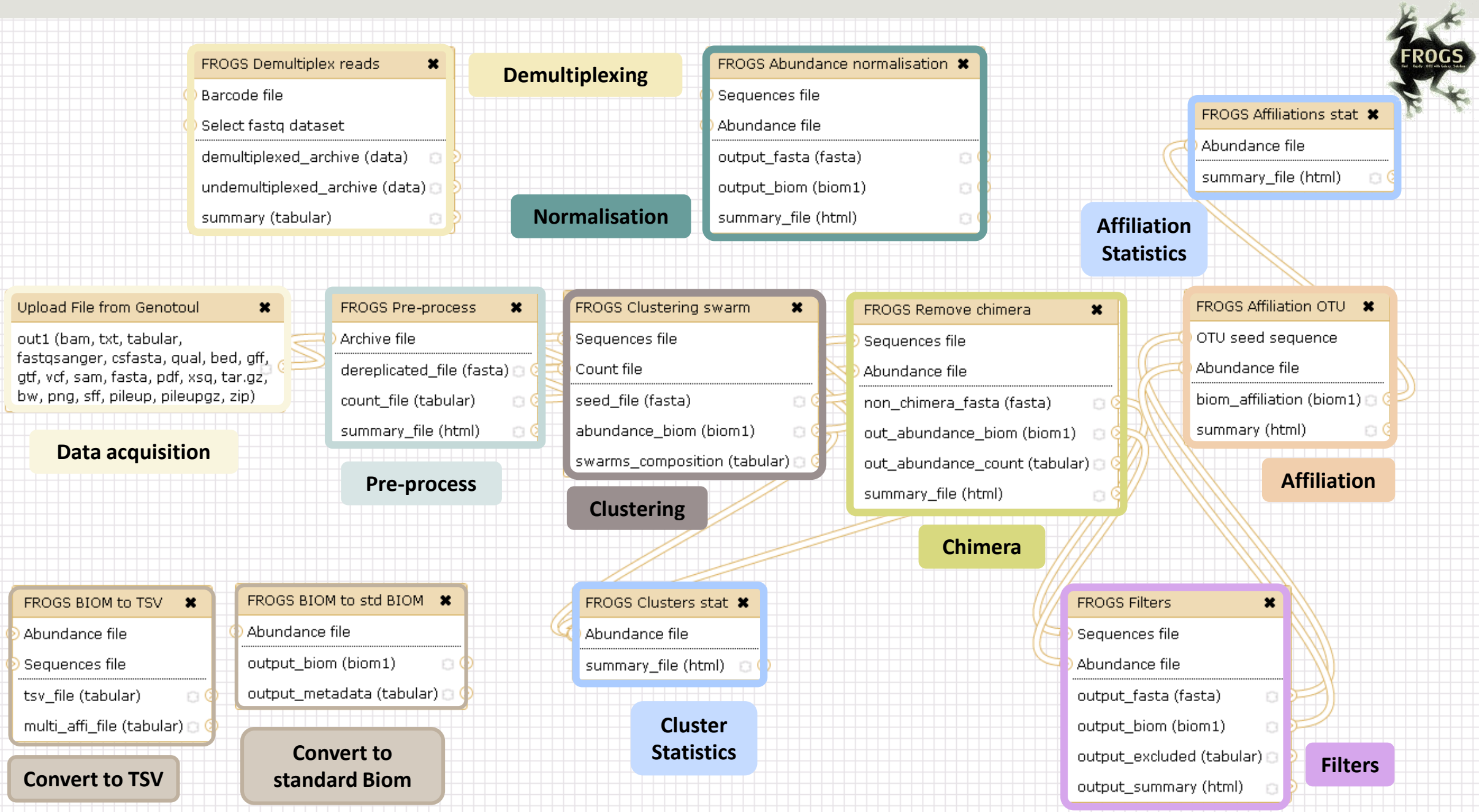
June 2016

Find Rapidly OTU with Galaxy Solution



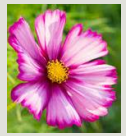
FRÉDÉRIC ESCUDIÉ* and LUCAS AUER*, MARIA BERNARD, LAURENT CAUQUIL, KATIA VIDAL, SARAH MAMAN, MAHENDRA MARIADASSOU, GUILLERMINA HERNANDEZ-RAQUET, GÉRALDINE PASCAL

*THESE AUTHORS HAVE CONTRIBUTED EQUALLY TO THE PRESENT WORK.


FROGS Demultiplex reads
Barcode file
Select fastq dataset
demultiplexed_archive (data)
undemultiplexed_archive (data)
summary (tabular)
Demultiplexing
FROGS Abundance normalisation
Sequences file
Abundance file
output_fasta (fasta)
output_biom (biom1)
summary_file (html)
Normalisation
FROGS Affiliations stat
Abundance file
summary_file (html)
Affiliation Statistics
Upload File from Genotoul
out1 (bam, txt, tabular, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png, sff, pileup, pileupgz, zip)
Data acquisition
FROGS Pre-process
Archive file
dereplicated_file (fasta)
count_file (tabular)
summary_file (html)
Pre-process
FROGS Clustering swarm
Sequences file
Count file
seed_file (fasta)
abundance_biom (biom1)
swarms_composition (tabular)
Clustering
FROGS Remove chimera
Sequences file
Abundance file
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
out_abundance_count (tabular)
summary_file (html)
Chimera
FROGS Affiliation OTU
OTU seed sequence
Abundance file
biom_affiliation (biom1)
summary (html)
Affiliation
FROGS BIOM to TSV
Abundance file
Sequences file
tsv_file (tabular)
multi_affi_file (tabular)
Convert to TSV
FROGS BIOM to std BIOM
Abundance file
output_biom (biom1)
output_metadata (tabular)
Convert to standard Biom
FROGS Clusters stat
Abundance file
summary_file (html)
Cluster Statistics
FROGS Filters
Sequences file
Abundance file
output_fasta (fasta)
output_biom (biom1)
output_excluded (tabular)
output_summary (html)
Filters

Accounts:





- anemone
- arome
- aster
- bleuet
- camelia
- capucine
- chardon
- clematite
- cobee
- coquelicot
- cosmos
- cyclamen
- Password: f1o2r3!

Exercise 1

LAUNCH UPLOAD TOOLS

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/
(datatype fastq or fastq.gz is the same !)

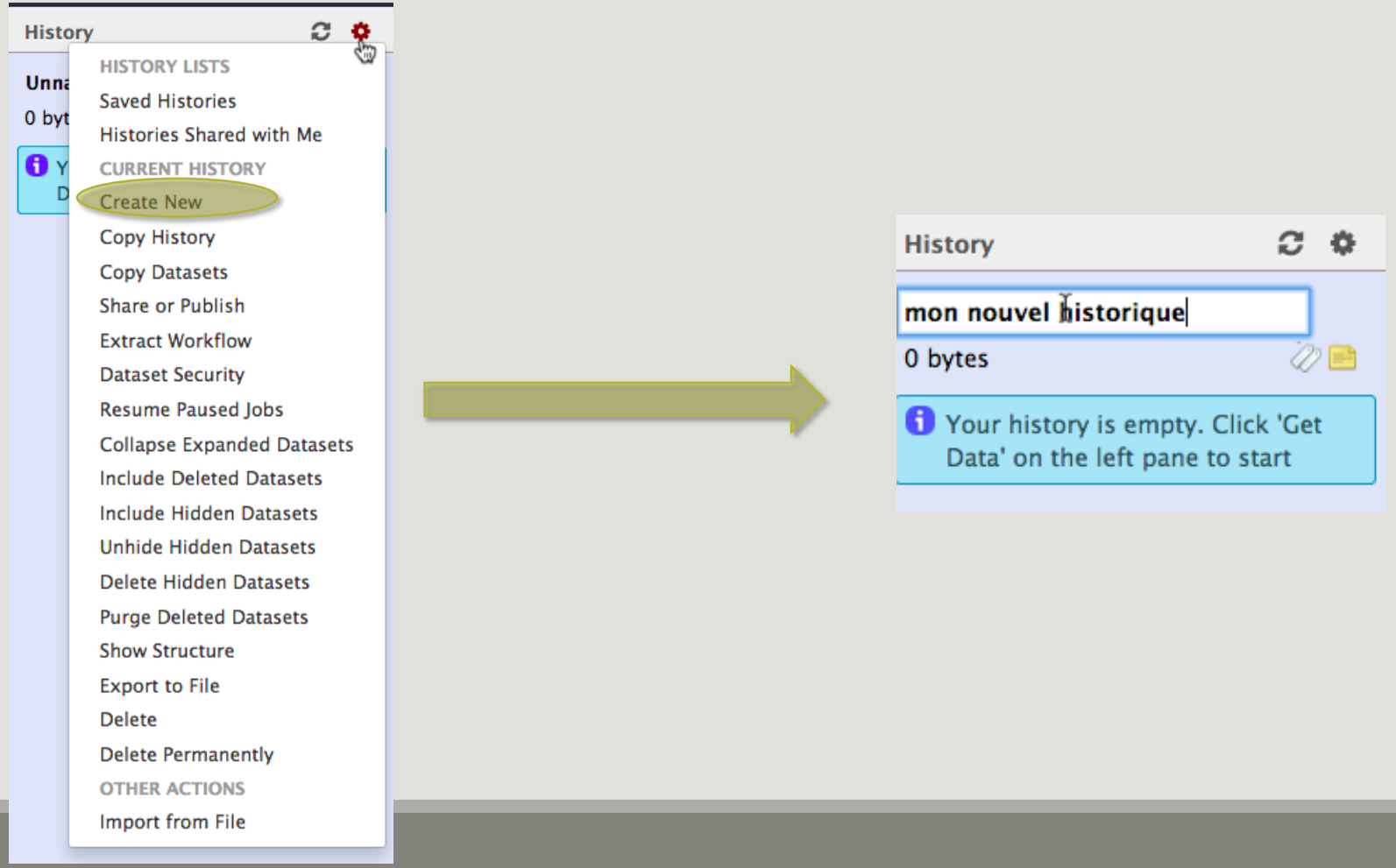
Create the 3rd history **MiSeq R1 R2**

Import files « **sampleA_R1.fastq** » and « **sampleA_R2.fastq** » present in the **Genotoul** folder /work/formation/FROGS/

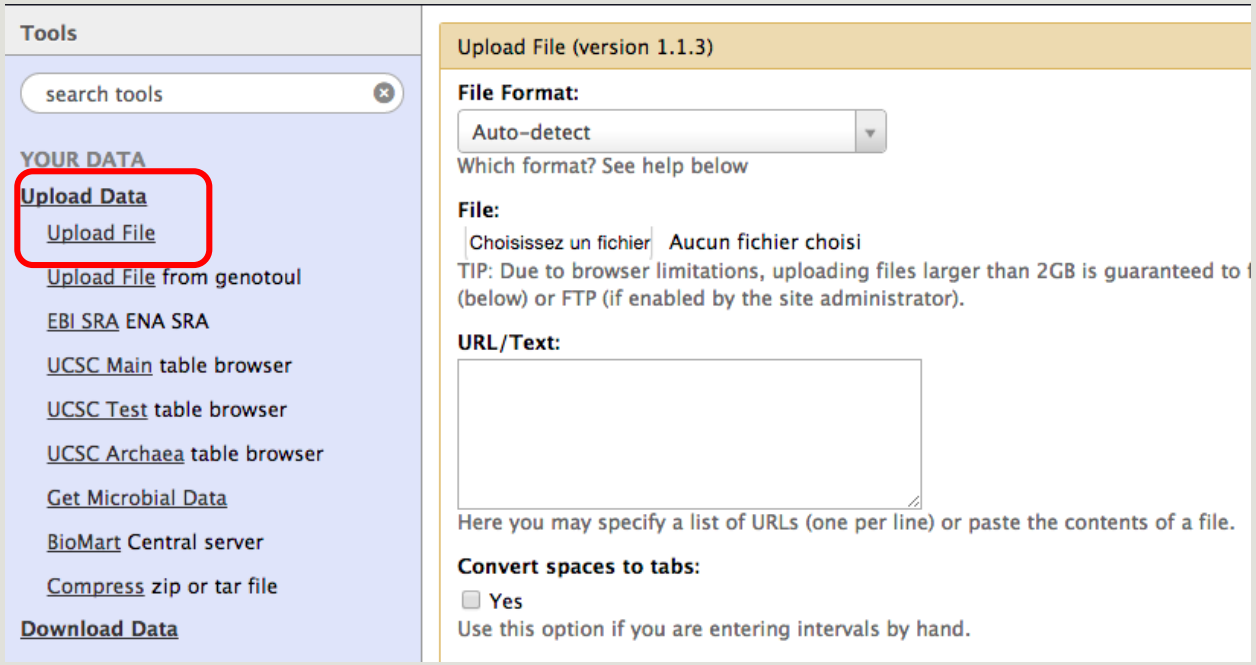
Create the 4th history **MiSeq contiged**

Import archive file « **100spec_90000seq_9samples.tar.gz** » present in the **Genotoul** folder /work/formation/FROGS/

History creation



Upload data: different methods

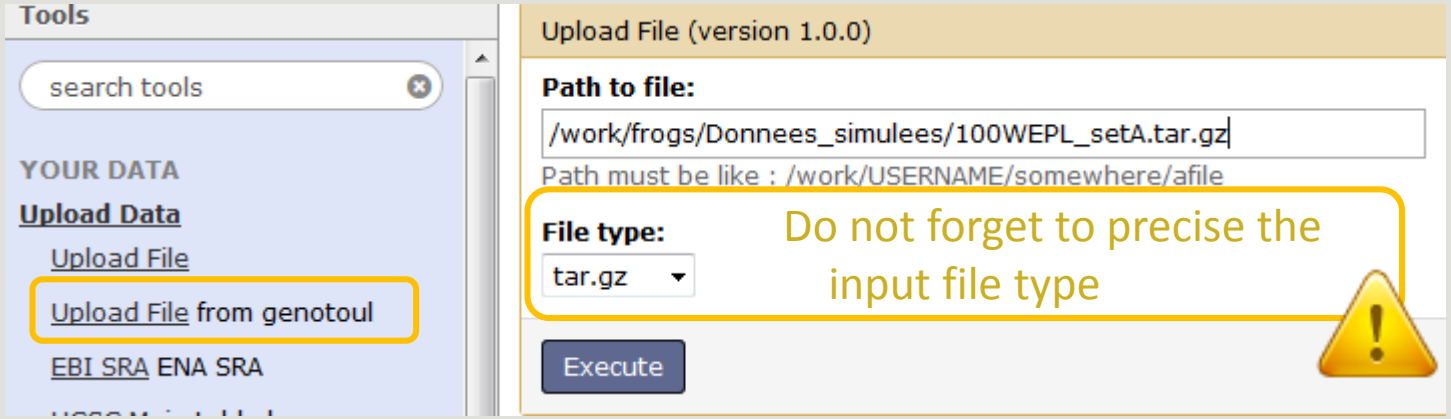


Default method, your files are on **your** computer or accessible on the internet, they are copied on your Galaxy account

You can only upload one local file at a time
→ 10 samples ≥ 10 uploads
You can upload multiple files using URLs but only smaller than 2Go

Each uploaded file will consume your Galaxy's quota!

Upload data: different methods



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

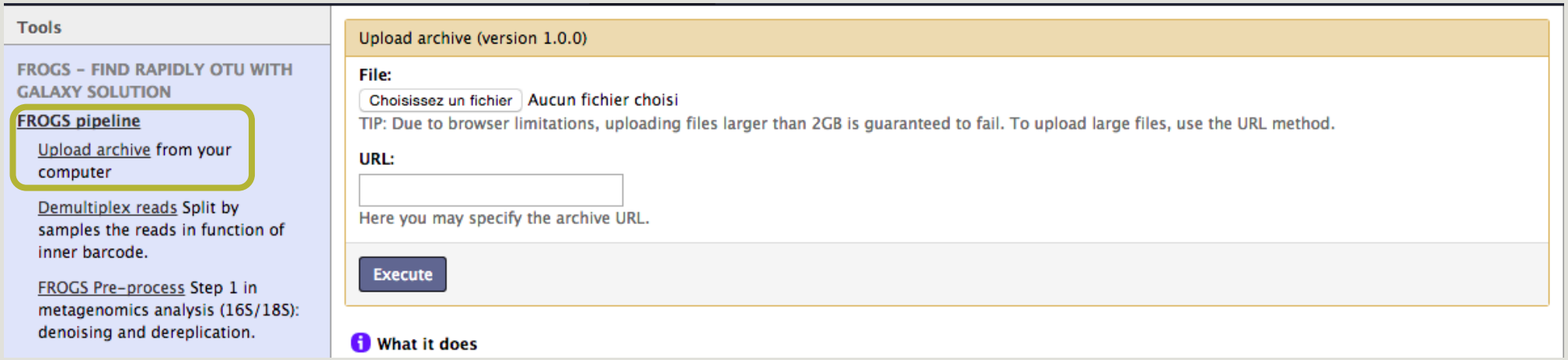
And if you have multiple samples ?

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



If you have an archive on your own computer and smaller than 2Go, you may use this specific FROGS tool to upload your samples archive instead of the default « Upload File » of Galaxy.

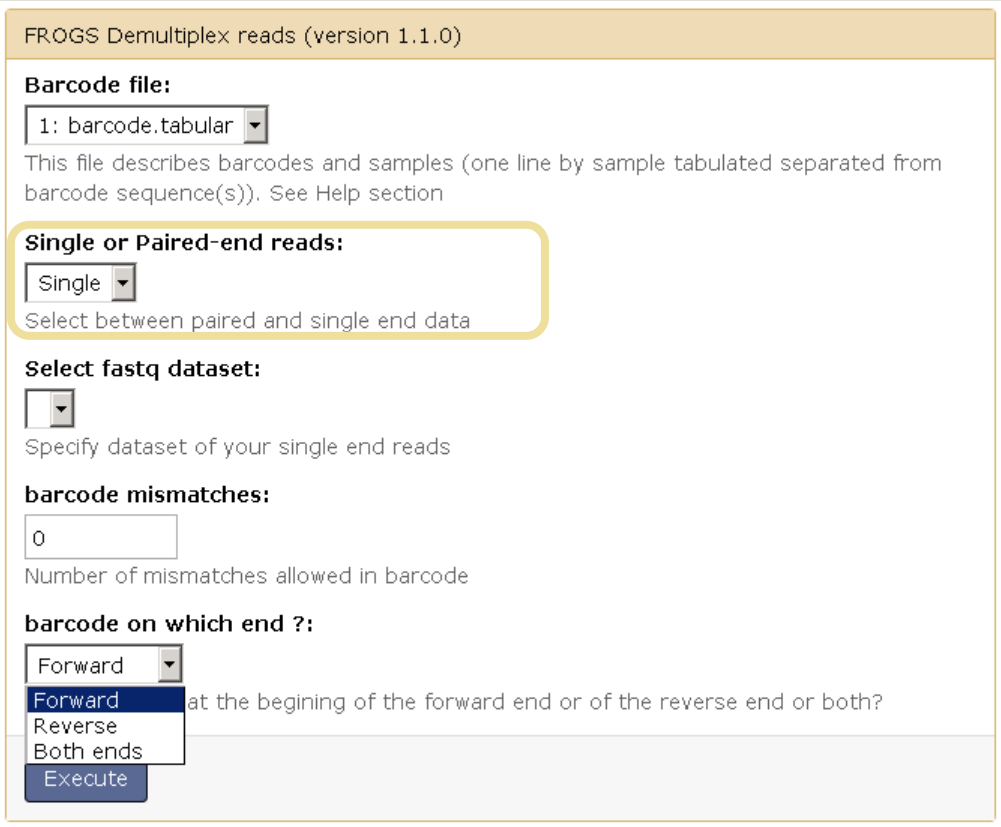
Back main documentation

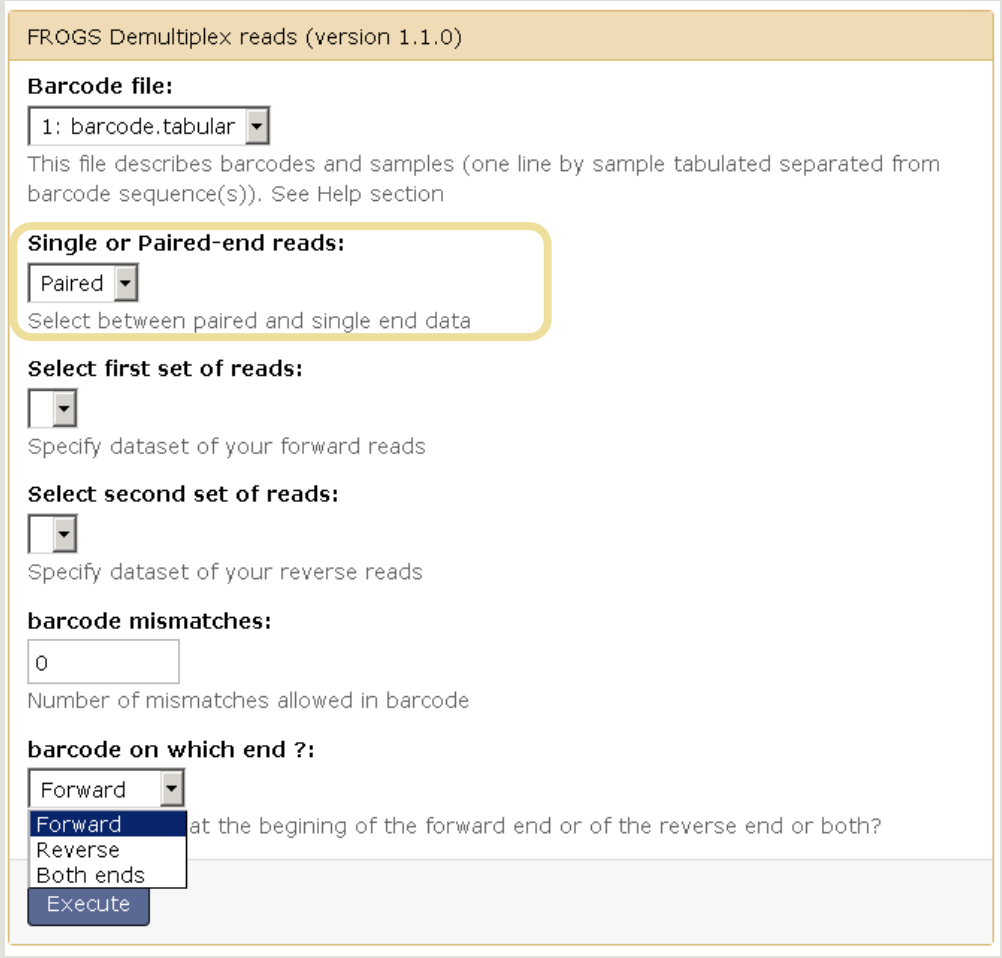
Exercise 2

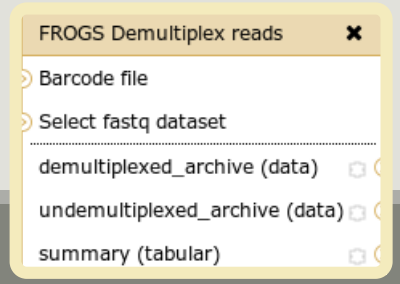
LAUNCH DEMULTIPLEX READS TOOL

multiplexed

The tool parameters depend on the input data type







Demultiplexing

multiplexed

Exercise 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, thay have extracted hundreads of or rats and ticks samples from which they have extracted 16S DNA and sequenced them first time on Roche 454 plateform and in a second time on Illumina Miseq plateform. For this courses, they authorized us to publicly shared some parts of these samples. »

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

multiplexed

Exercise 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 »

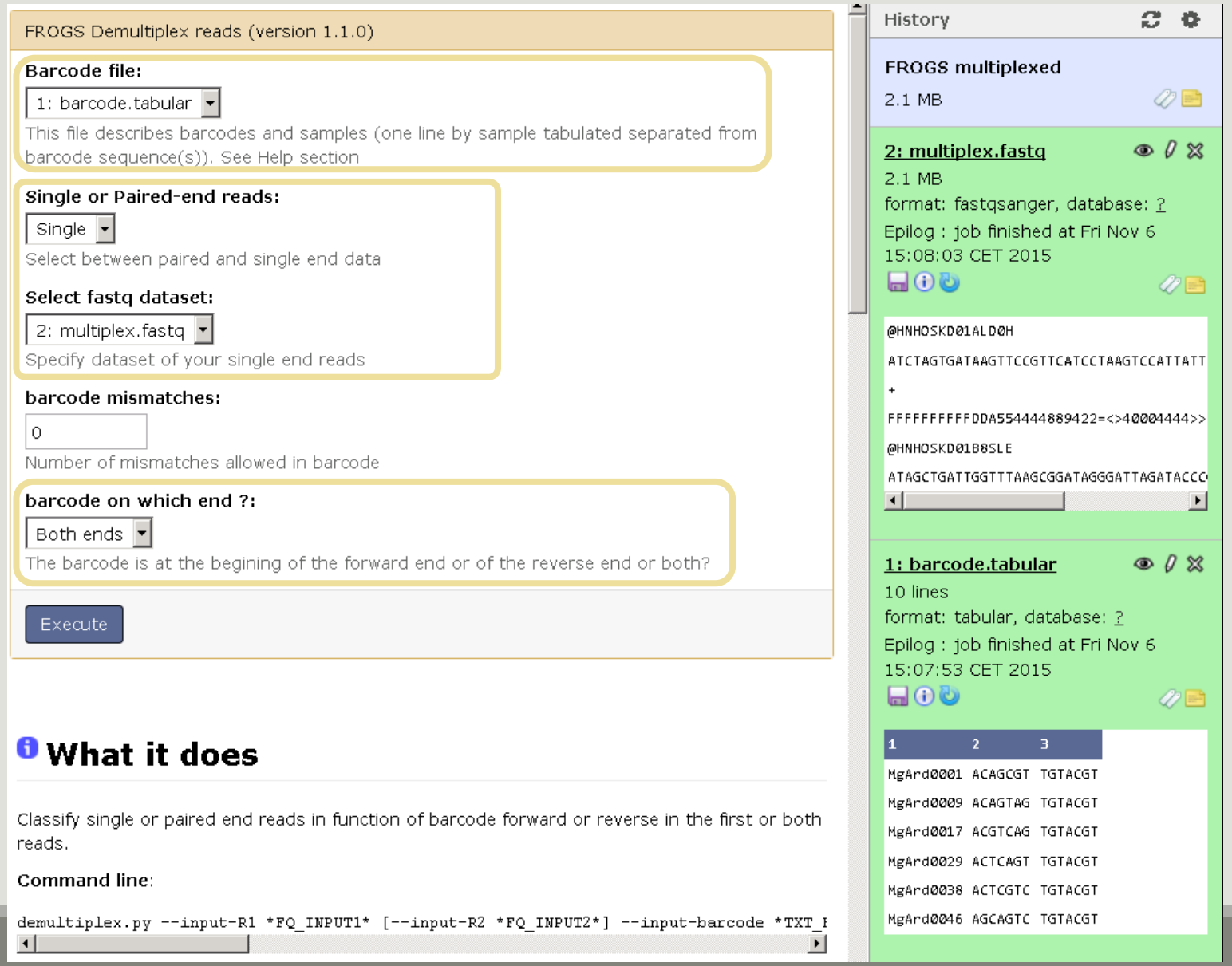
multiplexed

Exercise 2

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

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

multiplexed



multiplexed

For your own data

Advices

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

multiplexed

Results

9: FROGS Demultiplex reads: report

8: FROGS Demultiplex reads: undemultiplexed.tar.gz

7: FROGS 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.

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

Back to the main documentation

Exercises 3

LAUNCH THE PRE-PROCESS READS TOOL

454

Exercise 3.1

Go to « 454 » history

Launch the pre-process tool on that data set

→ objective : understand the parameters

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

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

454

FROGS Pre-process (version 1.4.2)

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

Sample name is required

Name:
my_sample
The sample name.

Sequence file:
6: /work/formation/FROGS/454.fastq.gz
FASTQ file of sample.

Add new Samples

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). The orientation is detailed

3' primer:
AGGATTAGATACCCTGGTA
The 3' primer sequence (wildcards are accepted). The orientation is detailed

Execute

Size range of 16S V3-V4:
[380 – 500]

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

454

Exercise 3.1

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

What is the length of your reads before preprocessing ?

Do you understand how enter your primers ?

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

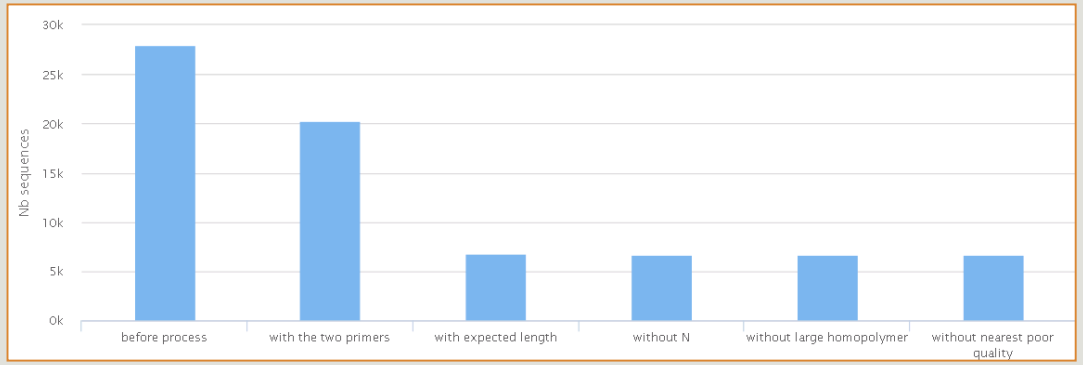
What is the « FROGS Pre-process: count.tsv » file ?

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

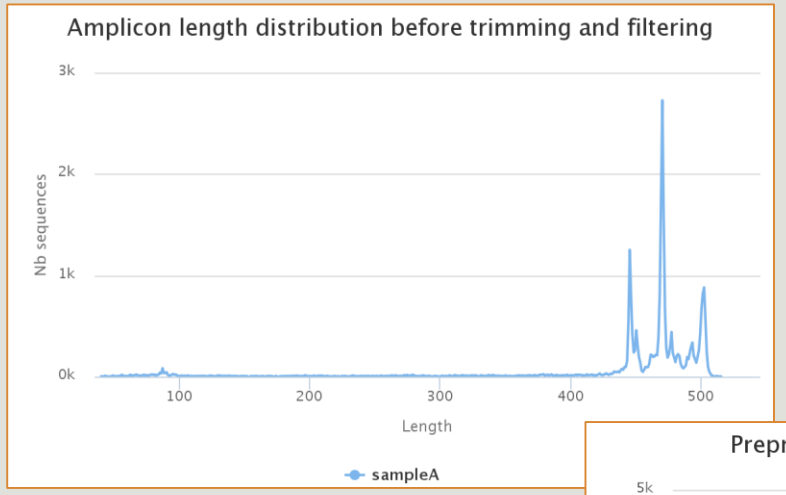
Who loose a lot of sequences ?

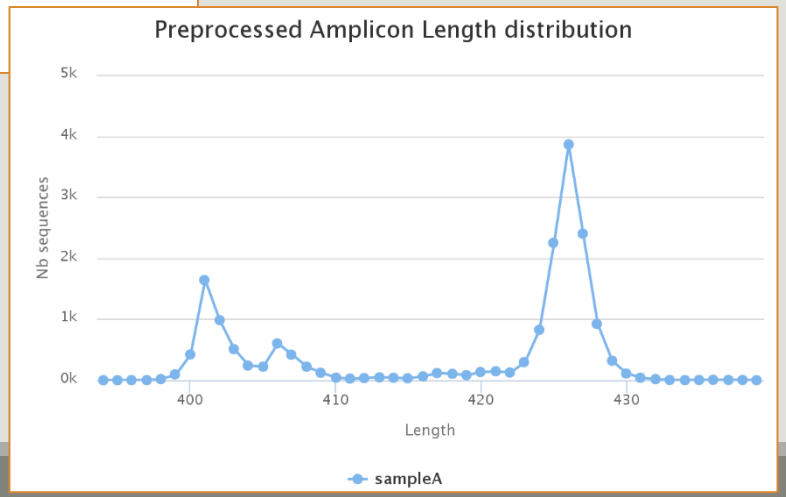
454

Samples	before process	with the two primers	with expected length	without N	without large homopolymer	without nearest poor quality
sample_454	28,009	20,227	6,806	6,677	6,675	6,672



To be kept, sequences must have the 2 primers





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

Back to the main documentation

MiSeq
R1 R2

Exercise 3.2

Go to « MiSeq R1 R2 » history

Launch the pre-process tool on that data set

→ objective: understand flash software

FROGS Pre-process (version 1.4.2)

Sequencer:

Illumina

Select the sequencer family used to produce the sequences.

Input type:

Files by samples

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

Reads already contiged ?:

No

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

Samples

Samples 1

Name:

sampleA

The sample name.

Reads 1:

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

R1 FASTQ file of paired-end reads.

reads 2:

2: /work/formation/FROGS/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'

Flash ?

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

MiSeq
R1 R2

Real 16S sequenced fragment

Expected amplicon size:

410

Maximum amplicon length expected in [illegible]oximately 90% of the amplicons.

Minimum amplicon size:

340

The minimum size for the amplicons.

Size with primers

Maximum amplicon size:

450

The maximum size for the amplicons.

Sequencing protocol:

Illumina standard

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

5' primer:

CCGTCAATTC

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

3' primer:

CCGCNGCTGCT

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

Execute

MiSeq
R1 R2

Exercise 3.2

Interpret « FROGS Pre-process: report.html » file.

Exercise 3.3

Go to« MiSeq contiged » history

Launch the pre-process tool on that data set

→ objective: understand output files

MiSeq contiged

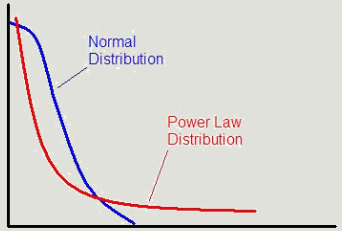
Exercise 3.3

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

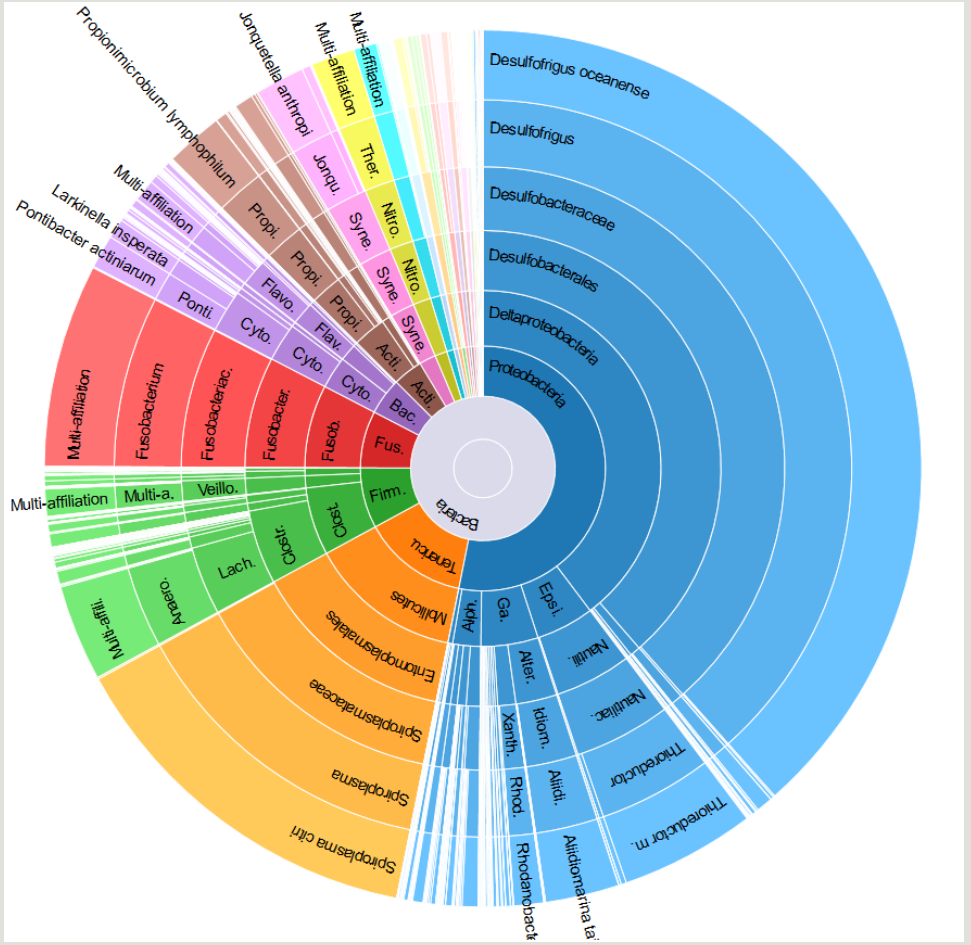
100_10000seq_sampleA1.fastq	100_10000seq_sampleB1.fastq	100_10000seq_sampleC1.fastq
100_10000seq_sampleA2.fastq	100_10000seq_sampleB2.fastq	100_10000seq_sampleC2.fastq
100_10000seq_sampleA3.fastq	100_10000seq_sampleB3.fastq	100_10000seq_sampleC3.fastq

MiSeq contiged

Exercise 3.3



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



MiSeq contiged

Exercise 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 (TACCAGGGTATCTAATCCTA) primers in the V3-V4 region and

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

M
co

FROGS Pre-process (version 1.4.2)

Sequencer:
Illumina
Select the sequencer family used to produce the sequences.

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

Archive file:
1: /work/formation/FROGS/100spec_90000seq_9samples.tar.gz
The tar file containing the sequences file(s) for each sample.

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

Minimum amplicon size:
380
The minimum size for the amplicons.

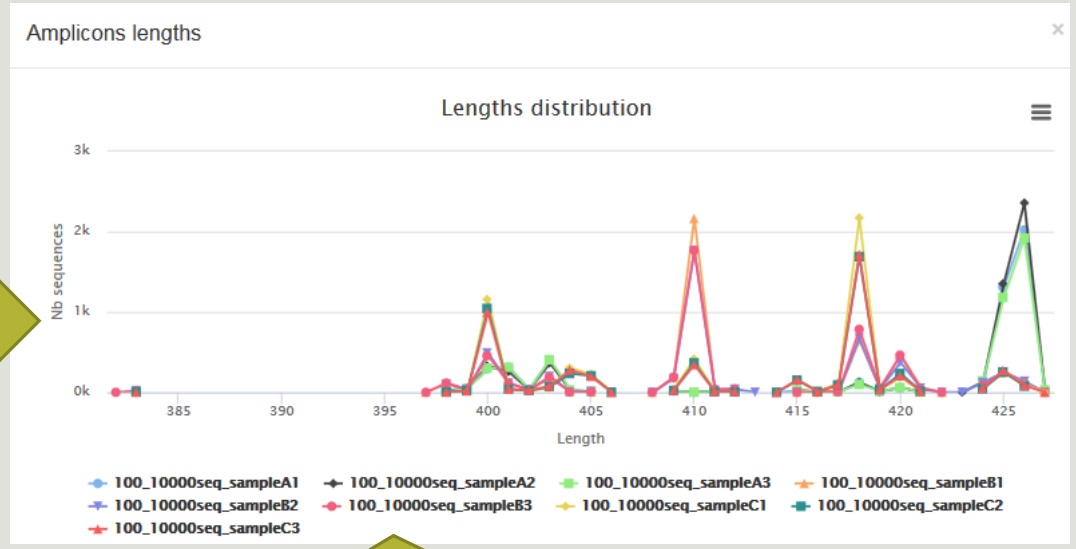
Maximum amplicon size:
500
The maximum size for the amplicons.

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

5' primer:
ACGGGAGGCAGCAG
The 5' primer sequence (wildcards are accepted). The ori

3' primer:
TAGGATTAGATACCCTGGT
The 3' primer sequence (wildcards are accepted). The ori

Execute



Click on legend

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

MiSeq contiged

Back to the main documentation

Exercise 3.3 - Questions

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

Exercise 4

LAUNCH CLUSTERING AND CLUSTERSTAT TOOLS

Clustering tools

MiSeq contiged

Exercise 4

Go to « MiSeq contiged » history

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

→ objectives :

- understand the denoising efficiency
- understand the ClusterStat utility

MiSeq contiged

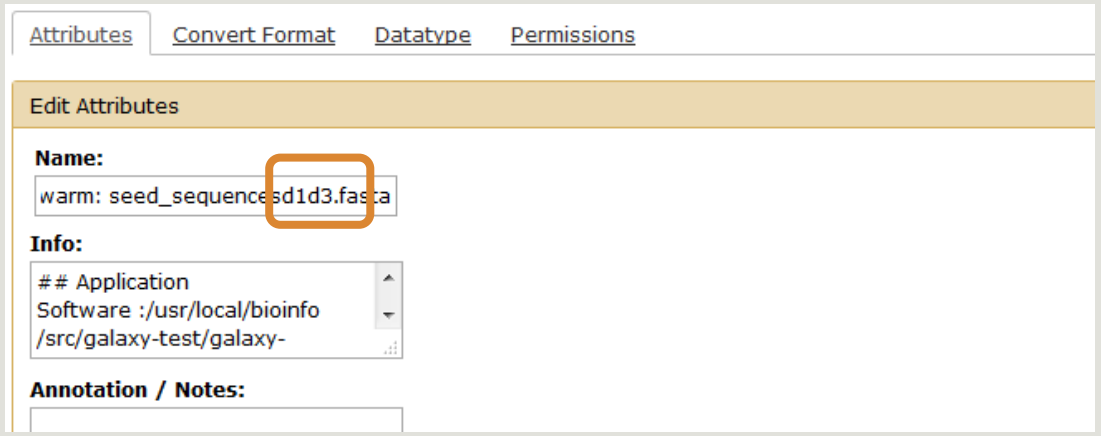
Exercise 4

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

MiSeq contiged

Exercise 4

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



FROGS Clusters stat Process some metrics on clusters.

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

MiSeq contiged

Back to the main documentation

Exercise 4

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

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

Exercise 5

LAUNCH THE REMOVE CHIMERA TOOL

MiSeq contiged

Exercise 5

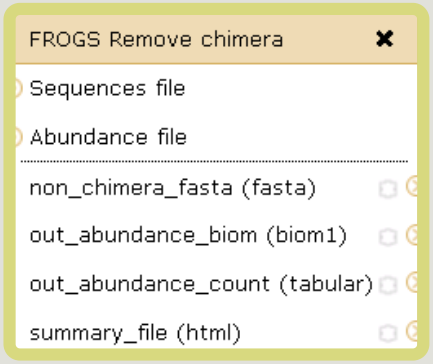
Go to « MiSeq contiged » history

Launch the « FROGS Remove Chimera » tool

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

→ objectives :

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



Chimera

FROGS Remove chimera (version 1.3.0)

Sequences file:

5: FROGS Clustering swarm: d1d3_seed_sequences.fasta

The sequences file (format: fasta).

Abundance type:

BIOM file

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

Abundance file:

6: FROGS Clustering swarm: d1d3_abundance.biom

It contains the count by sample for each sequence.

Execute

MiSeq contiged

Exercise 5

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

MiSeq contiged

Back to the main documentation

Exercise 5

2. Launch « FROGS ClusterStat » tool
   on non_chimera_abundanced1d3.biom
3. Rename output in summary_nonchimera_d1d3.html
4. Compare the HTML files
   a. Of what are mainly composed singleton ? (compare with precedent summary.html)
   b. What are their abundance?
   c. What do you conclude ?

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

Exercise 6

LAUNCH DE LA TOOL FILTERS

MiSeq contiged

Your turn: exo 6

Go to history « MiSeq contiged »

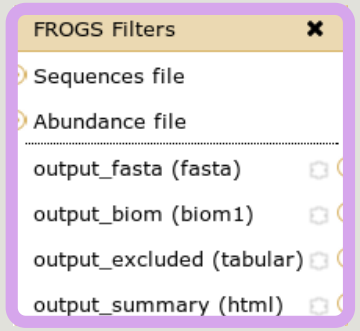
Launch « Filters » tool with non_chimera_abundanced1d3.biom, non_chimerad1d3.fasta

Apply 2 filters :

- proportion/number of sequences threshold to remove an OTU: 0.00005*
- Remove OTUs that are not present at least in XX samples;
  how many samples do you choose? : 3

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

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



Filters

Input

FROGS Filters (version 1.1.0)

Sequences file:

9: FROGS Remove chimera: non_chimera.fasta

The sequence file to filter (format: fasta).

Abundance file:

10: FROGS Remove chimera: non_chimera_abundance.biom

The abundance file to filter (format: BIOM).

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

Apply filters

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

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

3

Fill the field only if you want this treatment.

Proportion/number of sequences threshold to remove an OTU:

0.00005

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

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

Fill the fields only if you want this treatment.

*** THE FILTERS ON RDP:

No filters

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

*** THE FILTERS ON BLAST:

No filters

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

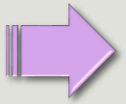
*** THE FILTERS ON CONTAMINATIONS:

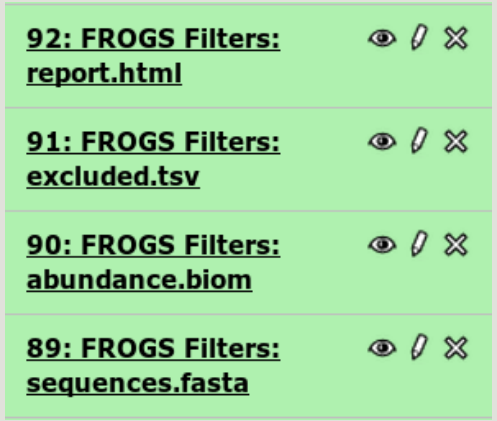
No filters

If you want to filter OTUs on classical contaminations.

Execute

Output





MiSeq contiged

Back to the main documentation

Your turn: exo 6

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

Exercises 7

LAUNCH THE « FROGS AFFILIATION » TOOL

MiSeq contiged

Exercise 7.1

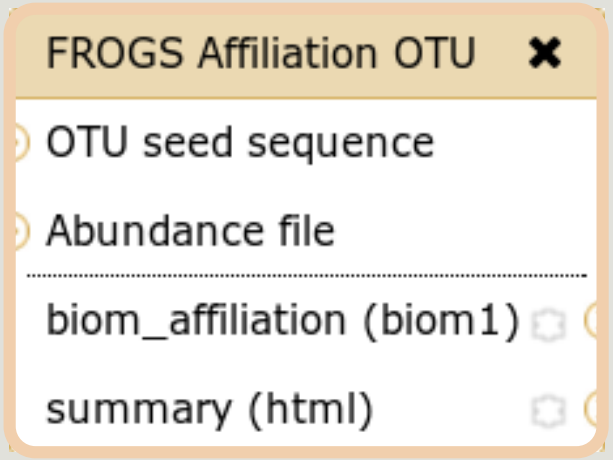
Go to « MiSeq contiged » history

Launch the « FROGS Affiliation » tool with

- SILVA 123 16S database
- FROGS Filters abundance biom and fasta files (after swarm d1d3, remove chimera and filter low abundances)

→ objectives :

- understand abundance tables columns
- understand the RDP and BLAST affiliation complementarity



Affiliation

FROGS Affiliation OTU (version 0.7.0)

Using reference database:

silva123 16S

Select reference from the list

OTU seed sequence:

13: FROGS Filters: sequences.fasta

OTU sequences (format: fasta).

Abundance file:

14: FROGS Filters: abundance.biom

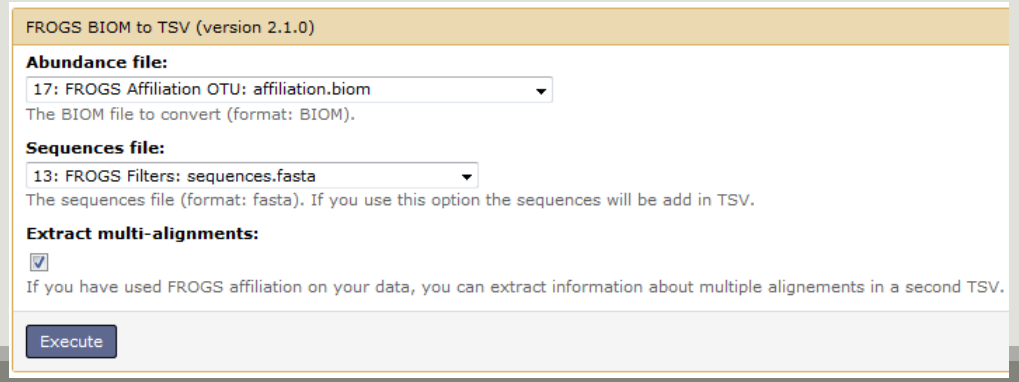
OTU abundances (format: BIOM).

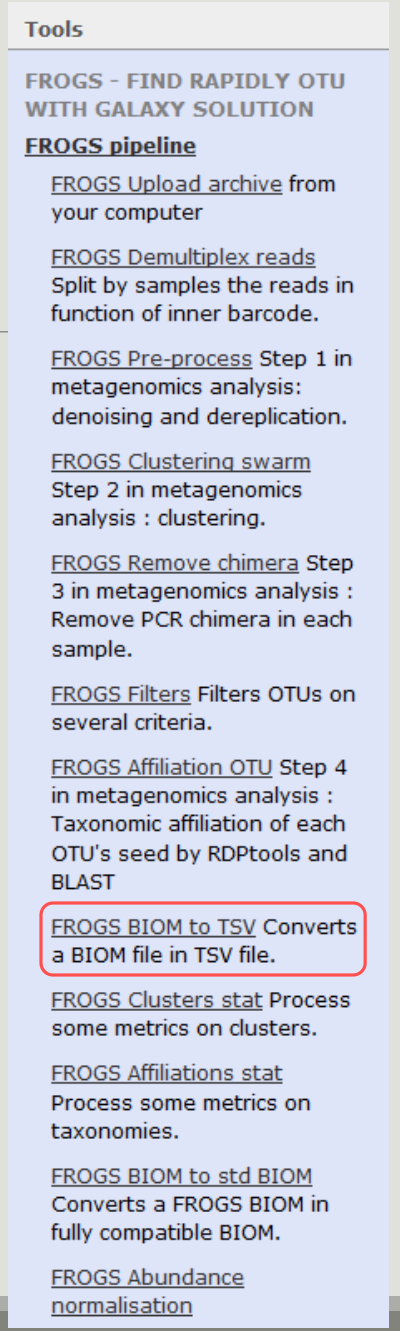
Execute

MiSeq contiged

Exercise 7.1

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





MiSeq contiged

Exercise 7.1

5. Compare RDP and Blast affiliations - Cluster_2388

#rdp_tax_and_bootstrap	blast_subject	blast_evalue	blast_len	blast_perc_query_coverage	blast_perc_identity	blast_taxonomy
Bacteria;(1.0);Planctomycetes;(1.0);Planctomycetacia;(1.0);Planctomycetales;(1.0);Planctomycetaceae;(1.0);Telmatocola;(1.0);Telmatocola sphagniphila;(1.0);	JN880417.1.1422	0.0	360	88.88	99.44	Bacteria;Planctomycetes;Planctomycetacia;Planctomycetales;Planctomycetaceae;Telmatocola;Telmatocola sphagniphila

Blast JN880417.1.1422 vs our OTU

OTU length : 405

Excellent blast but no matches at the beginning of OTU.

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

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

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

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

```
Query  46   CGCGTGCGCGATGAAGGCCTTCGGGTTGTAAAGCGCGAAAGAGGTAATAAAGGGAAACCT  105
            ||||||||||||||||||||||||||||||||||||||||||||||||||| ||||||||||
Sbjct  375  CGCGTGCGCGATGAAGGCCTTCGGGTTGTAAAGCGCGAAAGAGGSAATAAAGGGAAACTT  434

Query  106  GATTGAACCTCAGTAAGCTCGGGCTAAGTTTGTGCCAGCAGCCGCGGTAAGACGAACCGA  165
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  435  GATTGAACCTCAGTAAGCTCGGGCTAAGTTTGTGCCAGCAGCCGCGGTAAGACGAACCGA  494

Query  166  GCGAACGTTGTTCGGAATCACTGGGCATAAAGGGCGCGTAGGCGGGTTTCTAAGTCCGTG  225
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  495  GCGAACGTTGTTCGGAATCACTGGGCATAAAGGGCGCGTAGGCGGGTTTCTAAGTCCGTG  554

Query  226  GTGAAATACTTCAGCTCAACTGGAGAACTGCCTCGGATACTGGGAATCTCGAGTAATGTA  285
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  555  GTGAAATACTTCAGCTCAACTGGAGAACTGCCTCGGATACTGGGAATCTCGAGTAATGTA  614

Query  286  GGGGCACGTGGAACGGCTGGTGGAGCGGTGAAATGCGTTGATATCAGTCGGAACTCCGGT  345
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  615  GGGGCACGTGGAACGGCTGGTGGAGCGGTGAAATGCGTTGATATCAGTCGGAACTCCGGT  674

Query  346  GGCGAAGGCGATGTGCTGGACATTTACTGACGCTGAGGCGCGAAAGCCAGGGGAGCAAAC  405
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  675  GGCGAAGGCGATGTGCTGGACATTTACTGACGCTGAGGCGCGAAAGCCAGGGGAGCAAAC  734
```

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

NCBI Reference Sequence: NR_118328.1

FASTA Graphics

Go to:

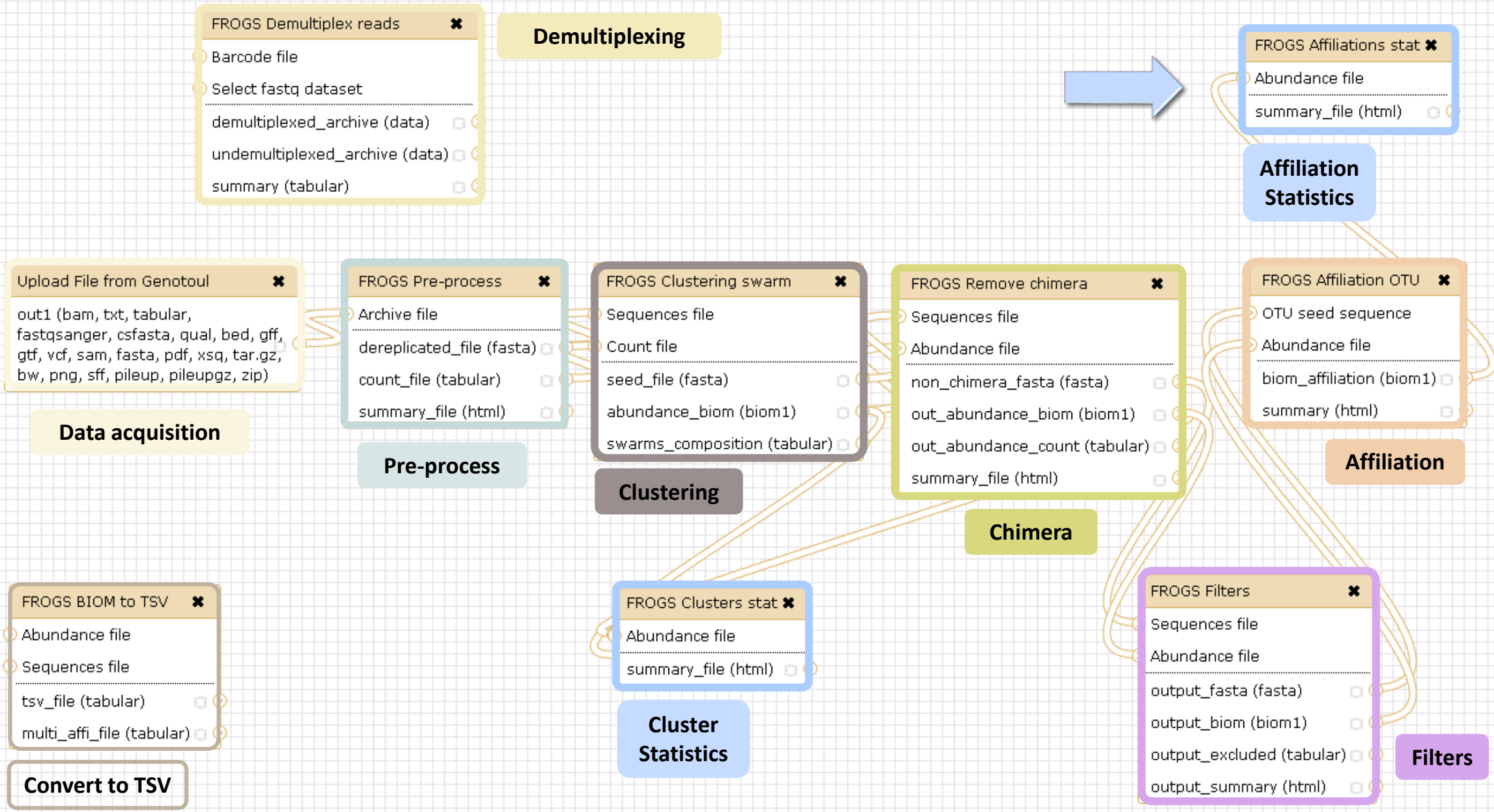
```
LOCUS       NR_118328              1422 bp    rRNA    linear   BCT 03-FEB-2015
DEFINITION  Telmatocola sphagniphila strain SP2 16S ribosomal RNA gene, partial
            sequence.
ACCESSION   NR_118328
VERSION     NR_118328.1  GI:645321338
DBLINK      Project: 33175
            BioProject: PRJNA33175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

What do you think about this case ?

Observe the Cluster_4 line (big abundance!):

#rdp_tax_and_bootstrap	blast_subject	blast_evalue	blast_len	blast_perc_query_coverage	blast_perc_identity	blast_taxonomy
Bacteria;(1.0);Thermotogae;(1.0);Thermotogae;(1.0);Thermotogales;(1.0);Thermotogaceae;(1.0);Petrotoga;(1.0);Petrotoga miotherma;(0.62);	FR733705.1.1499	0.0	419	100.0	100.0	Bacteria;Thermotogae;Thermotogae;Thermotogales;Thermotogaceae;Petrotoga;Petrotoga miotherma

Back to the main documentation


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

MiSeq contiged

Exercise 7.2

Launch the « FROGS Affiliation Stat» tool

- On FROGS blast affiliation
- On FROGS RDP affiliation

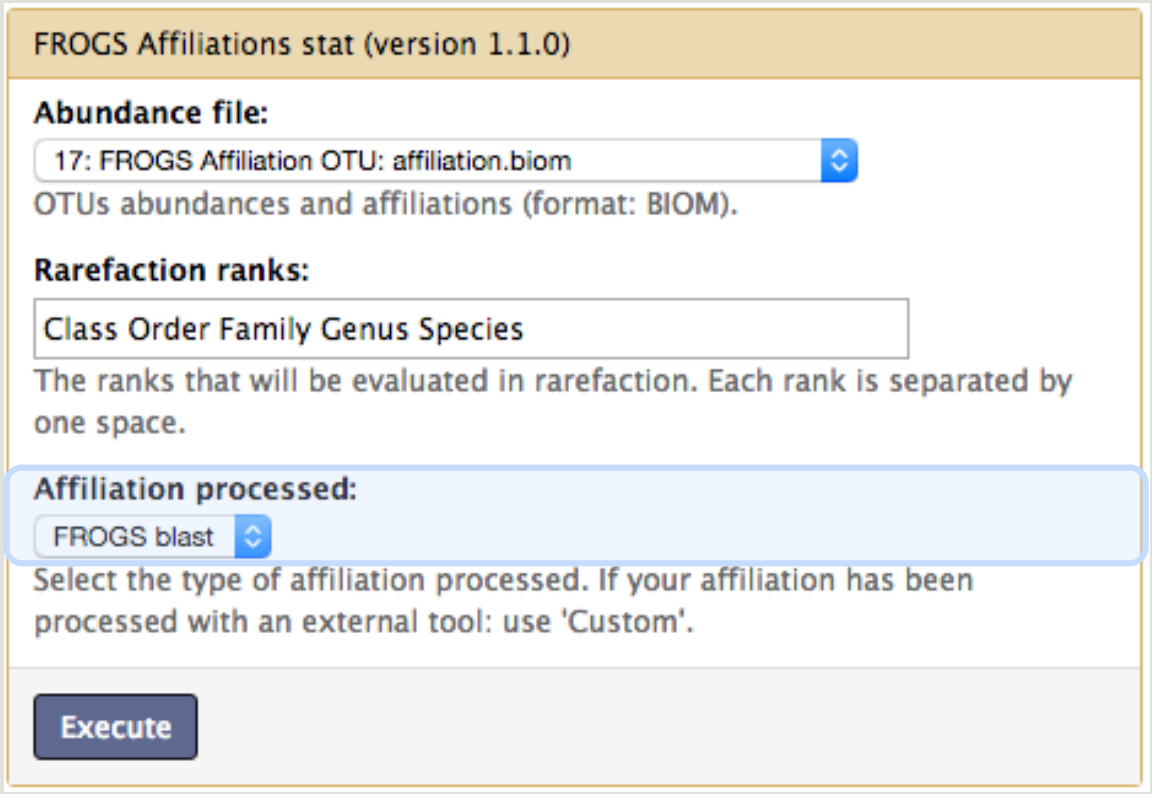
→ objectives :

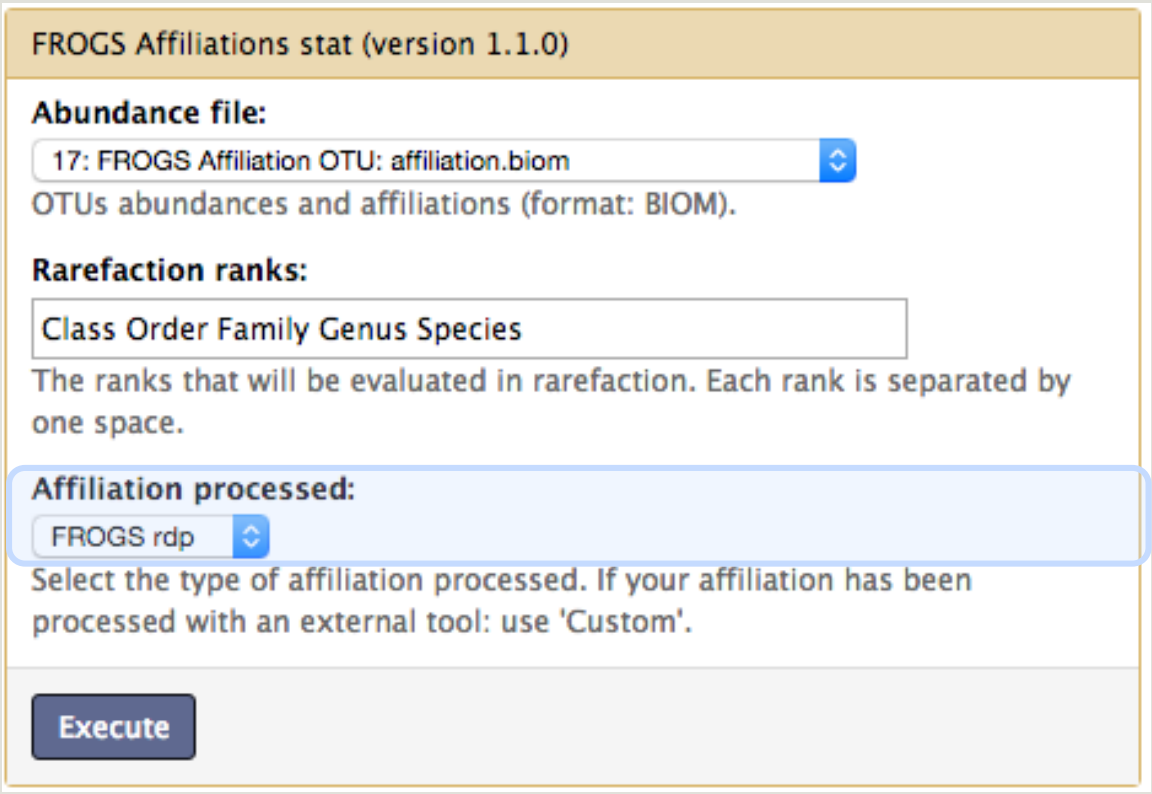
understand rarefaction curve and sunburst

understand the RDP and BLAST affiliation complementarity

MiSeq contiged

Exercise 7.2





MiSeq contiged

Exercise 7.2

1. Explore the Affiliation stat results.
2. What kind of graphs can you generate? What do they mean?
   a) Common to Blast and RDP affiliation results
   b) On Blast results
   c) On RDP results
3. Launch Filters Tool on blasts results (100% id and 100% cov) + biom_to_tsv

Back to the main documentation

Exercise 8

LAUNCH NORMALIZATION TOOL

Exercise 8

Launch Normalization Tool

1. What is the smallest sequenced samples ?
2. Normalize your data from Affiliation based on that number of sequence
3. Explore the report HTML result.

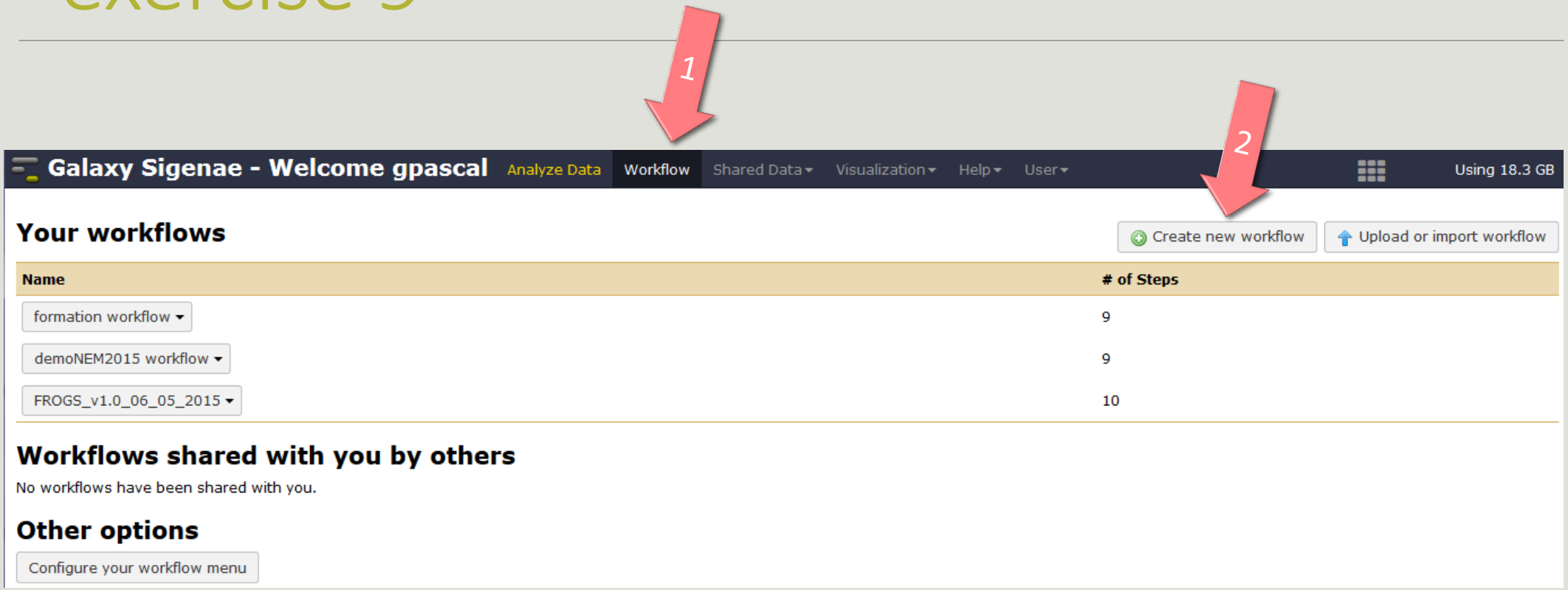
Back to the main documentation

Exercise 9

CREATE YOUR OWN WORKFLOW !

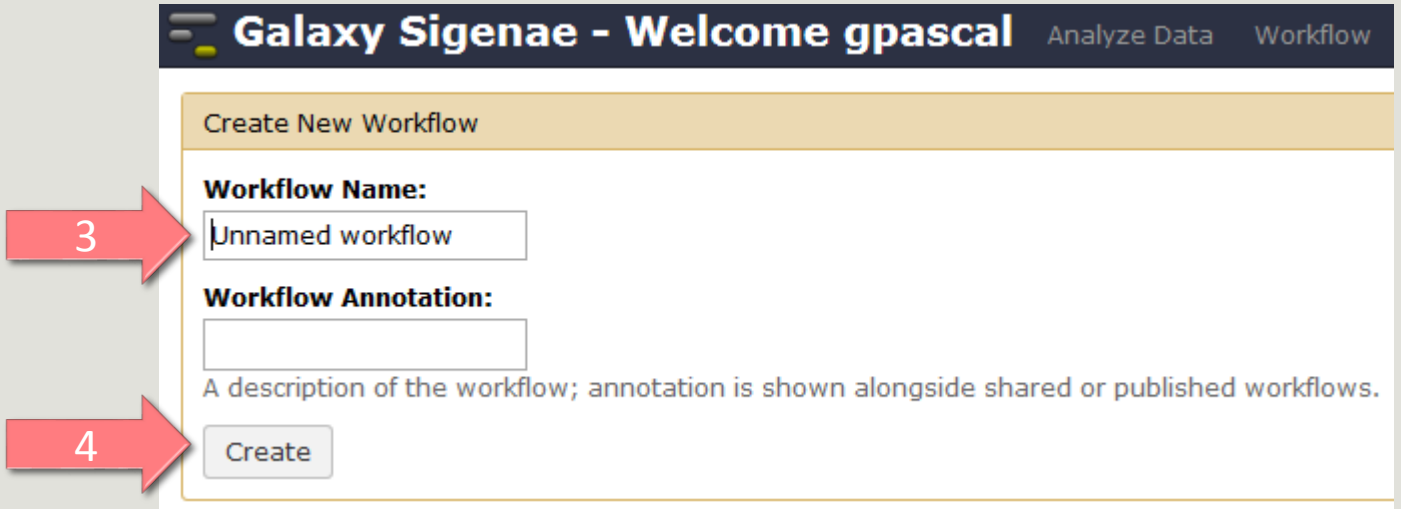
MiSeq contiged

exercise 9



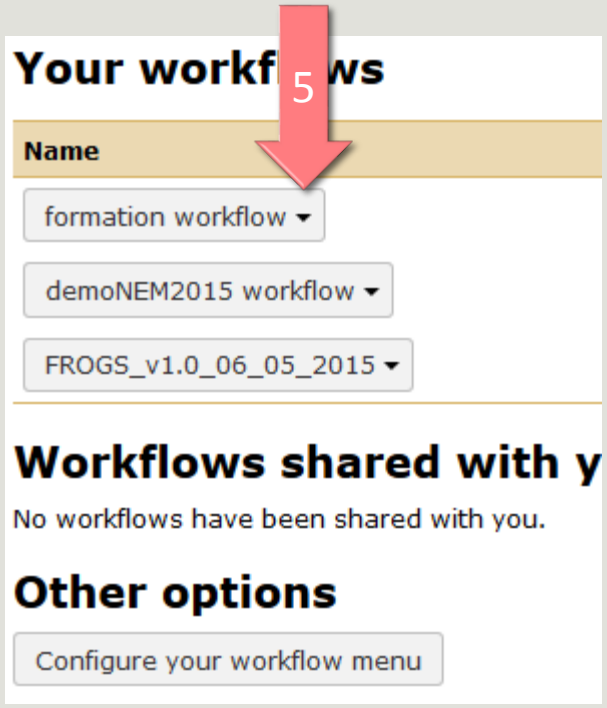
MiSeq contiged

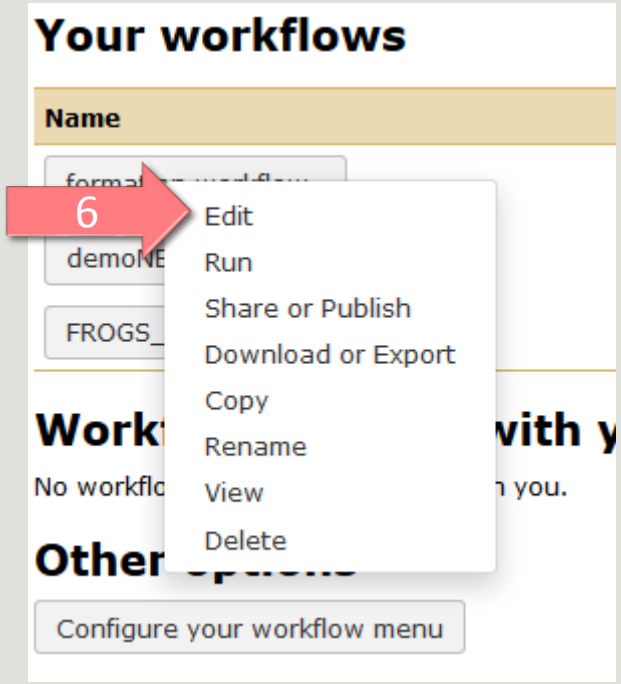
exercise 9

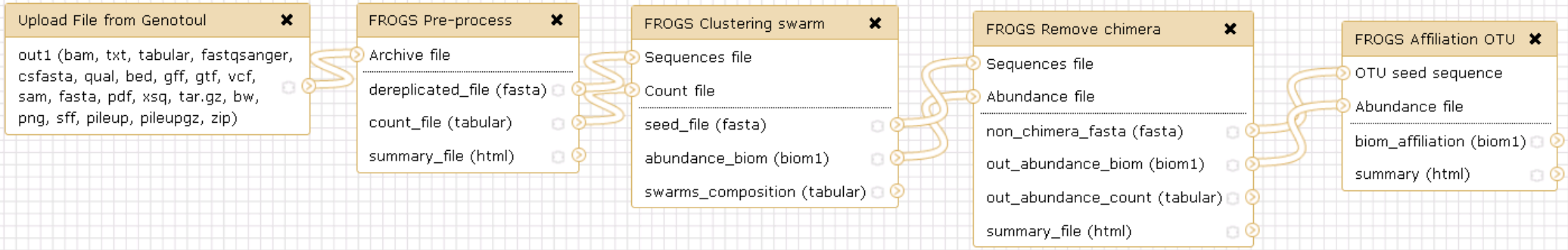


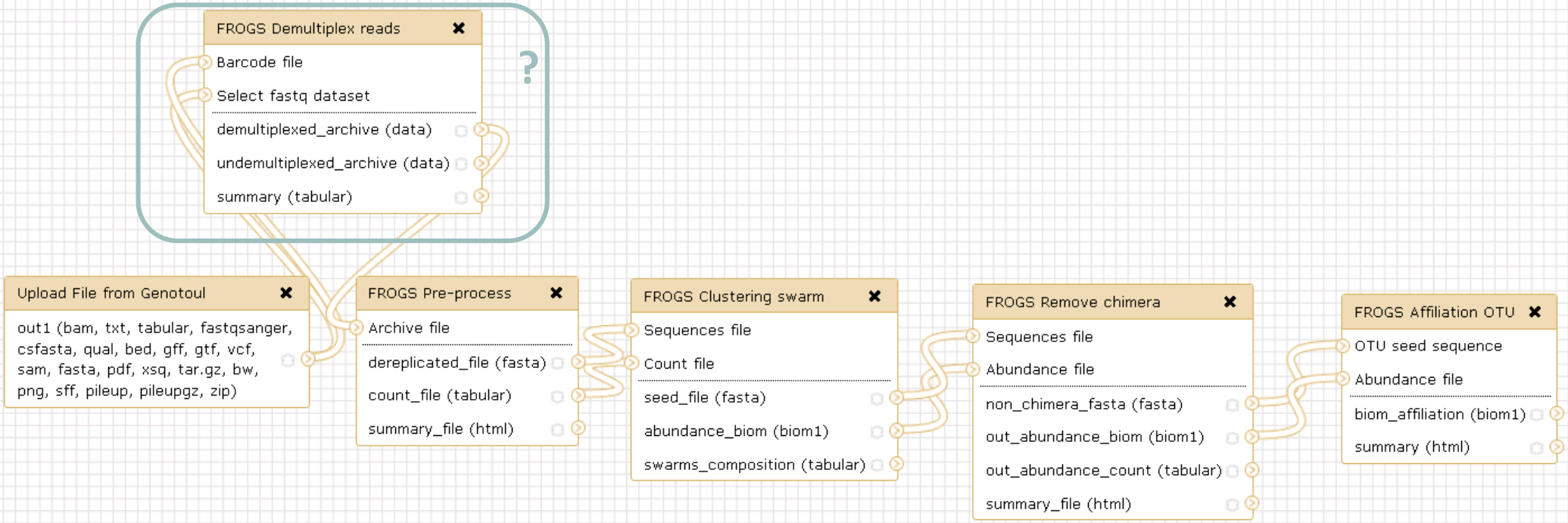
MiSeq contiged

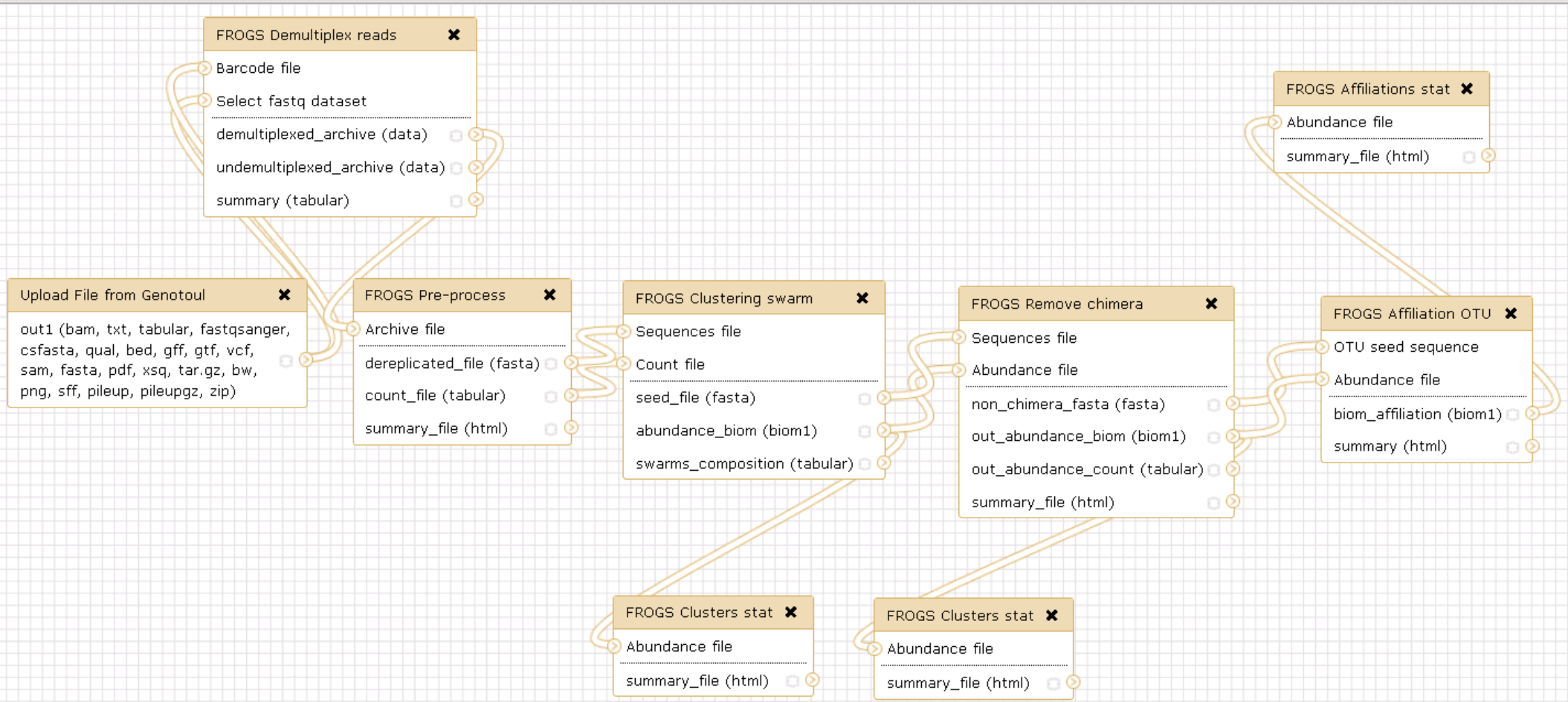
exercise 9

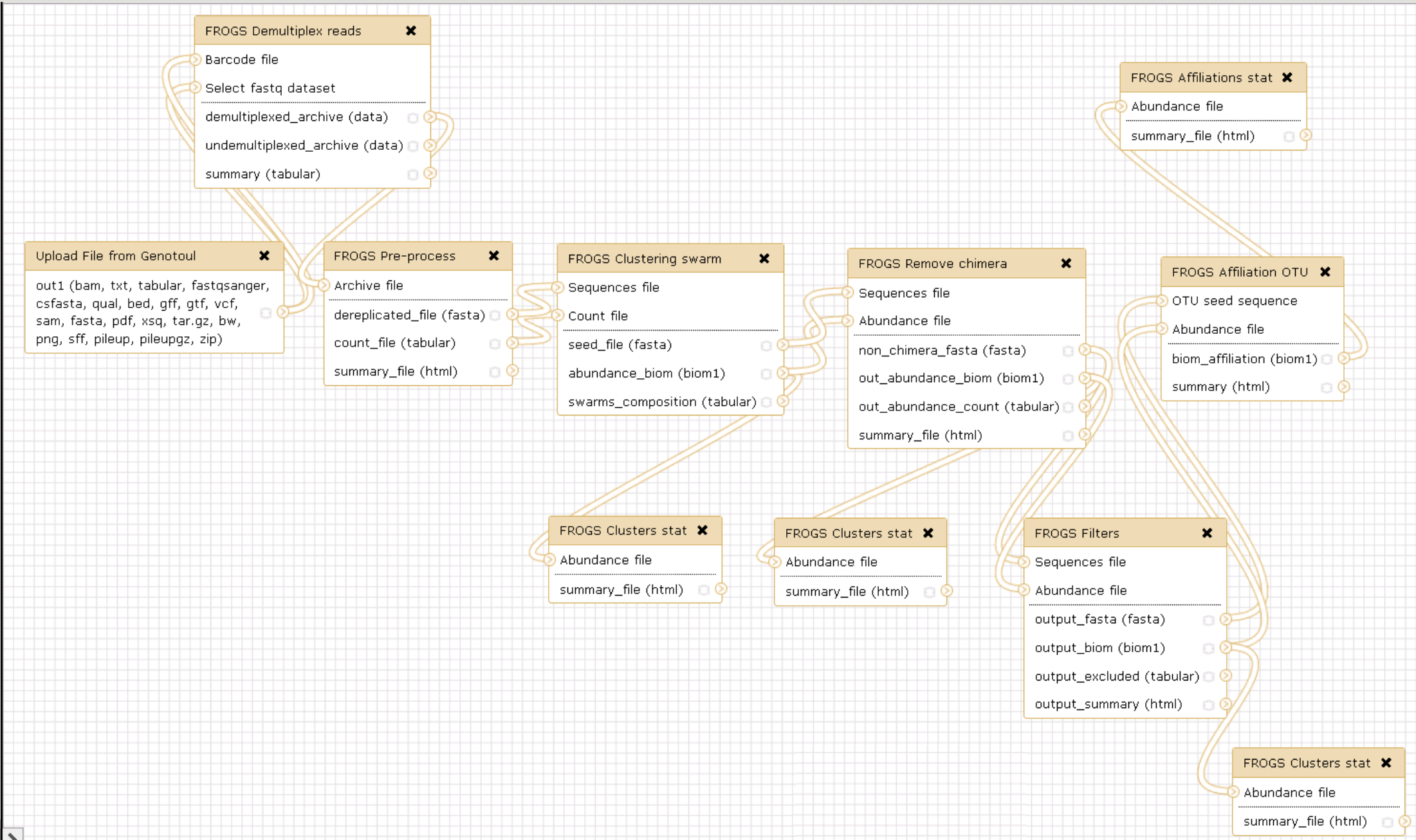


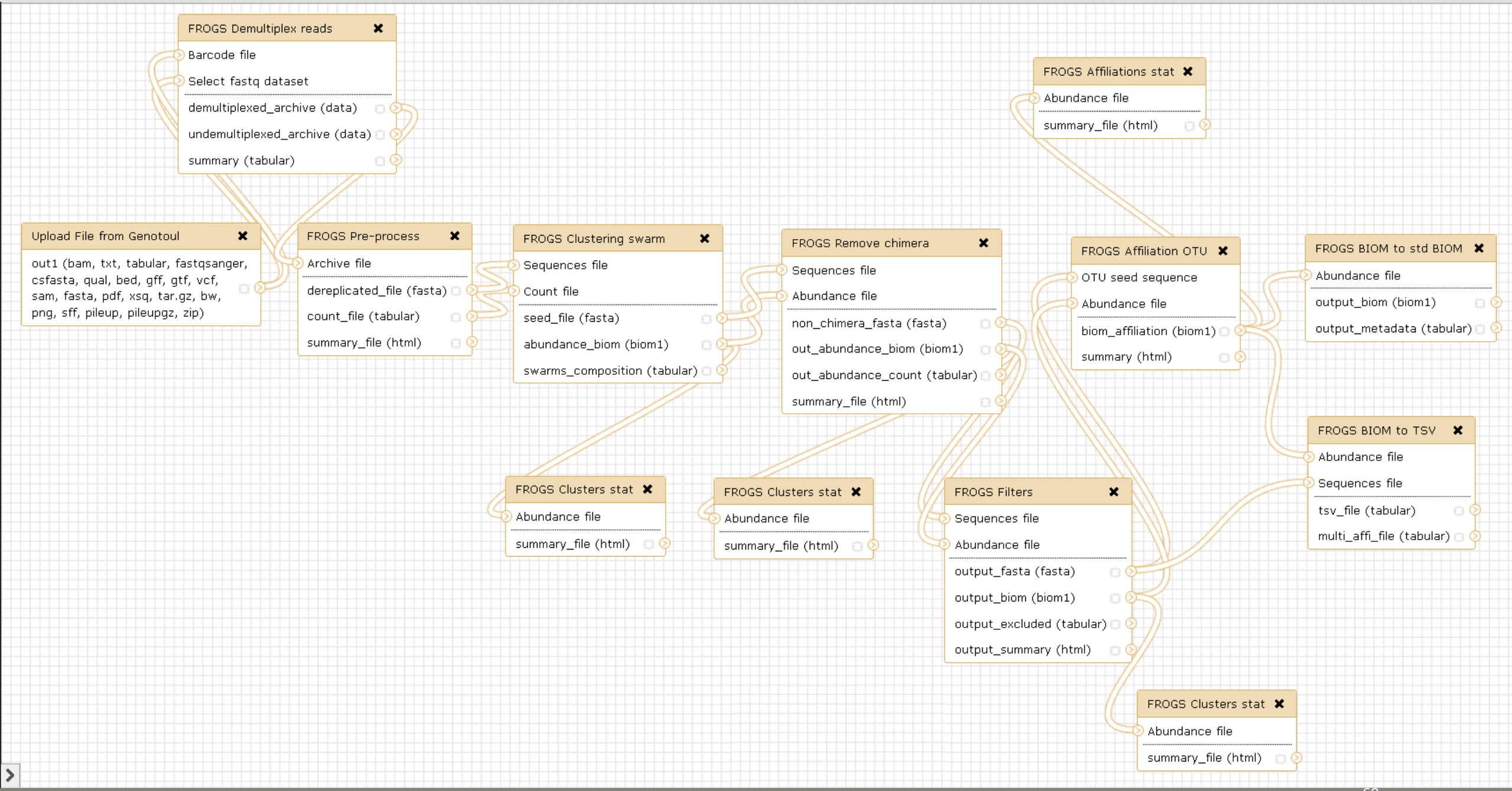


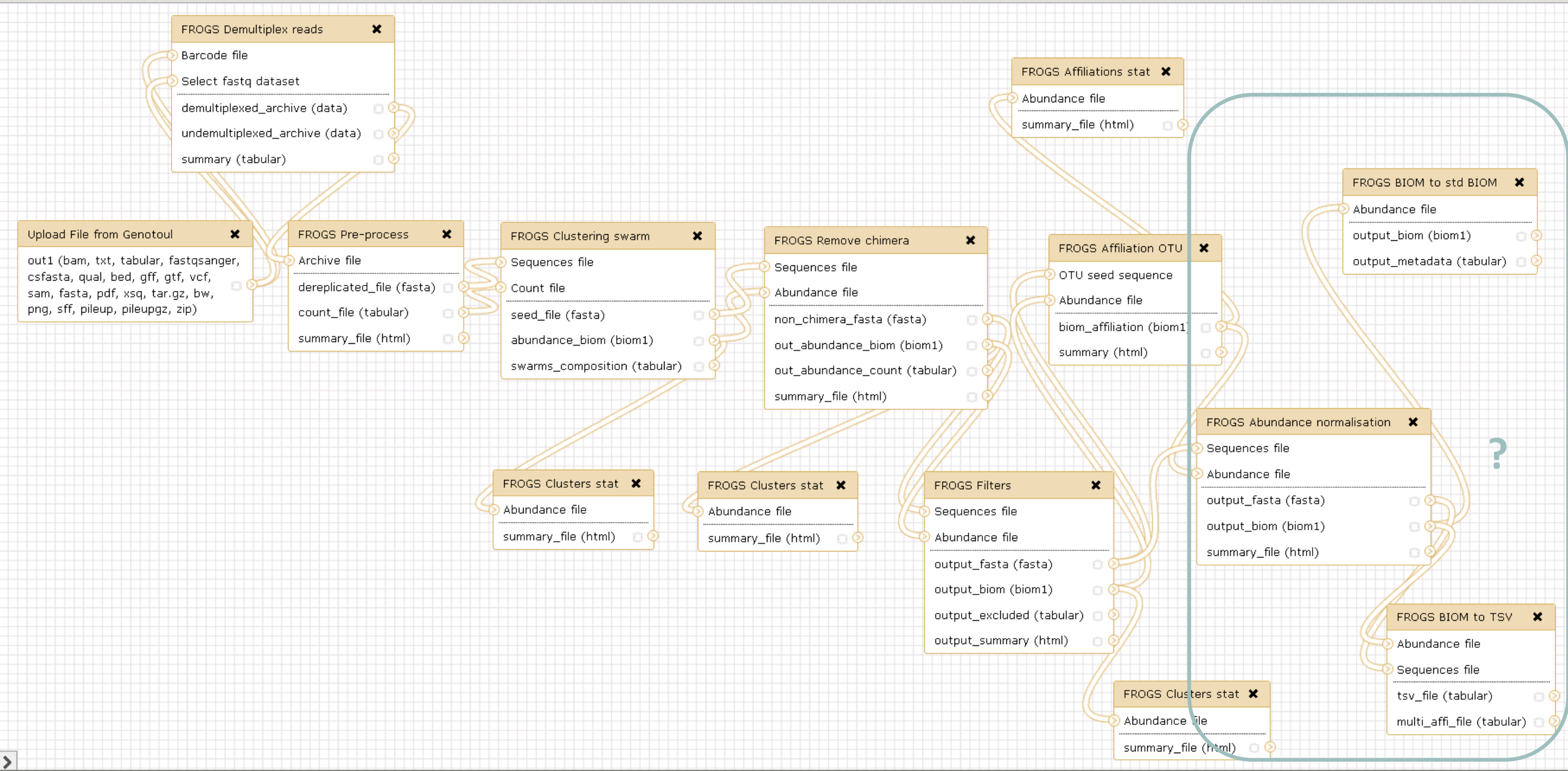

Upload File from Genotoul
out1 (bam, txt, tabular, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png, sff, pileup, pileupgz, zip)
FROGS Pre-process
Archive file
dereplicated_file (fasta)
count_file (tabular)
summary_file (html)
FROGS Clustering swarm
Sequences file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

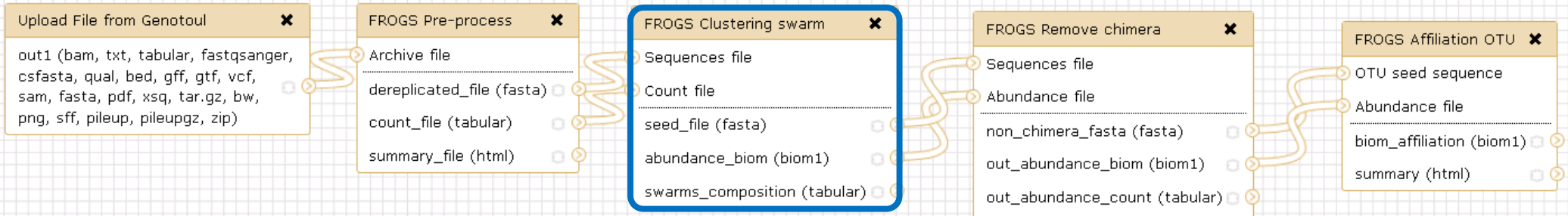

FROGS Demultiplex reads
Barcode file
Select fastq dataset
demultiplexed_archive (data)
undemultiplexed_archive (data)
summary (tabular)
?
Upload File from Genotoul
out1 (bam, txt, tabular, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png, sff, pileup, pileupgz, zip)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


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


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

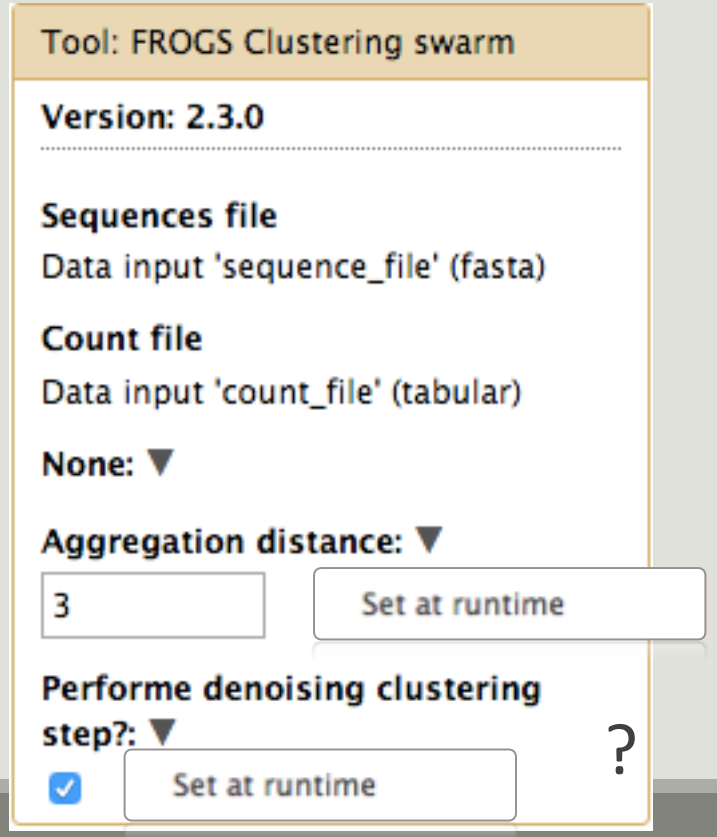

FROGS Demultiplex reads
Barcode file
Select fastq dataset
demultiplexed_archive (data)
undemultiplexed_archive (data)
summary (tabular)
FROGS Affiliations stat
Abundance file
summary_file (html)
Upload File from Genotoul
out1 (bam, txt, tabular, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png, sff, pileup, pileupgz, zip)
FROGS Pre-process
Archive file
dereplicated_file (fasta)
count_file (tabular)
summary_file (html)
FROGS Clustering swarm
Sequences file
Count file
seed_file (fasta)
abundance_biom (biom1)
swarms_composition (tabular)
FROGS Remove chimera
Sequences file
Abundance file
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
out_abundance_count (tabular)
summary_file (html)
FROGS Affiliation OTU
OTU seed sequence
Abundance file
biom_affiliation (biom1)
summary (html)
FROGS BIOM to std BIOM
Abundance file
output_biom (biom1)
output_metadata (tabular)
FROGS BIOM to TSV
Abundance file
Sequences file
tsv_file (tabular)
multi_affi_file (tabular)
FROGS Clusters stat
Abundance file
summary_file (html)
FROGS Clusters stat
Abundance file
summary_file (html)
FROGS Filters
Sequences file
Abundance file
output_fasta (fasta)
output_biom (biom1)
output_excluded (tabular)
output_summary (html)
FROGS Clusters stat
Abundance file
summary_file (html)

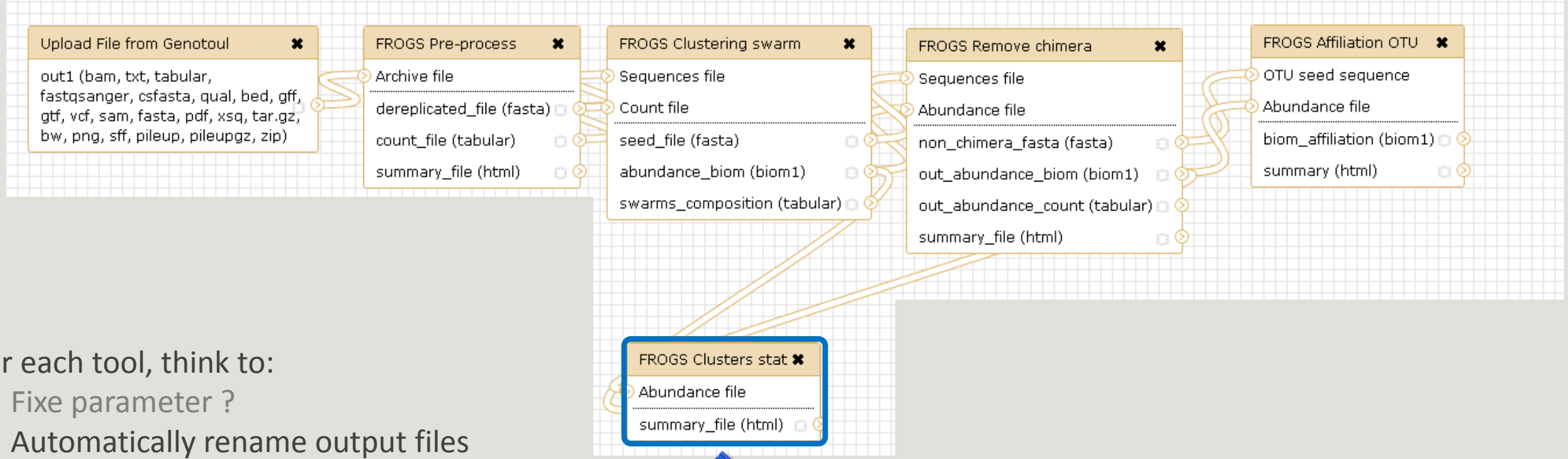

FROGS Demultiplex reads
Barcode file
Select fastq dataset
demultiplexed_archive (data)
undemultiplexed_archive (data)
summary (tabular)
Upload File from Genotoul
out1 (bam, txt, tabular, fastqsanger, csfasta, qual, bed, gff, gtf, vcf, sam, fasta, pdf, xsq, tar.gz, bw, png, sff, pileup, pileupgz, zip)
FROGS Pre-process
Archive file
dereplicated_file (fasta)
count_file (tabular)
summary_file (html)
FROGS Clustering swarm
Sequences file
Count file
seed_file (fasta)
abundance_biom (biom1)
swarms_composition (tabular)
FROGS Remove chimera
Sequences file
Abundance file
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
out_abundance_count (tabular)
summary_file (html)
FROGS Affiliations stat
Abundance file
summary_file (html)
FROGS Affiliation OTU
OTU seed sequence
Abundance file
biom_affiliation (biom1)
summary (html)
FROGS BIOM to std BIOM
Abundance file
output_biom (biom1)
output_metadata (tabular)
FROGS Clusters stat
Abundance file
summary_file (html)
FROGS Clusters stat
Abundance file
summary_file (html)
FROGS Filters
Sequences file
Abundance file
output_fasta (fasta)
output_biom (biom1)
output_excluded (tabular)
output_summary (html)
FROGS Abundance normalisation
Sequences file
Abundance file
output_fasta (fasta)
output_biom (biom1)
summary_file (html)
?
FROGS BIOM to TSV
Abundance file
Sequences file
tsv_file (tabular)
multi_affi_file (tabular)
FROGS Clusters stat
Abundance file
summary_file (html)



For each tool, think to:

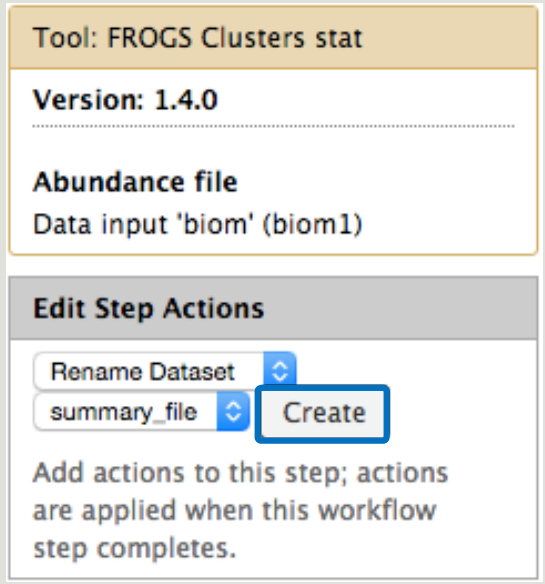
- Fixe parameter ?

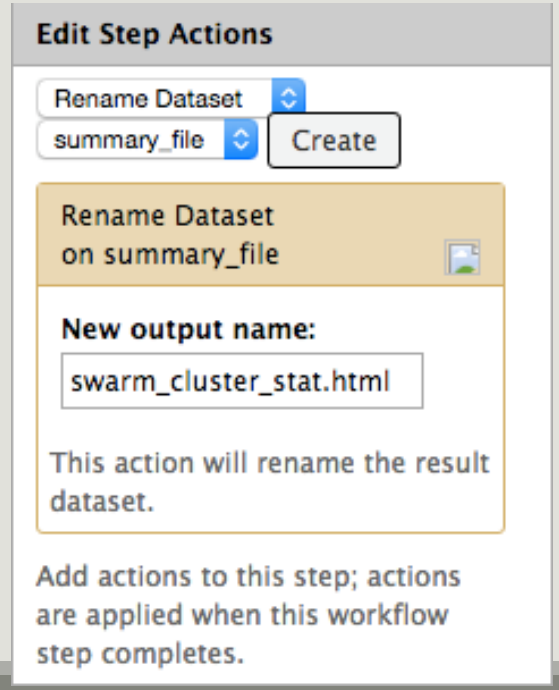


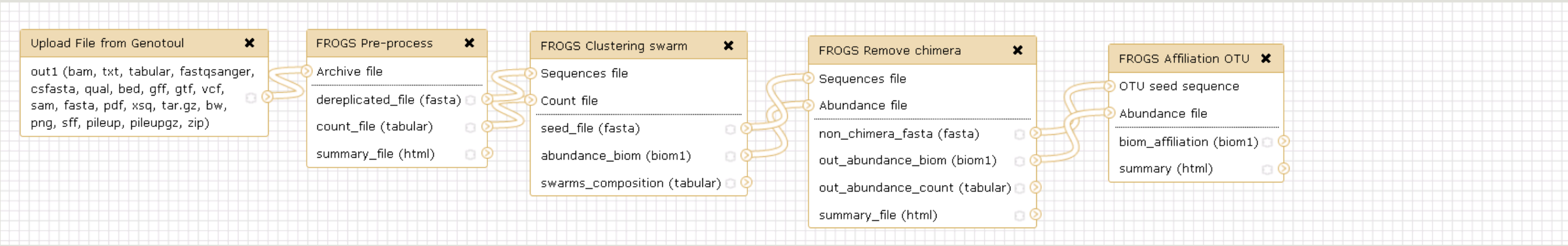


For each tool, think to:

- Fixe parameter ?
- Automatically rename output files

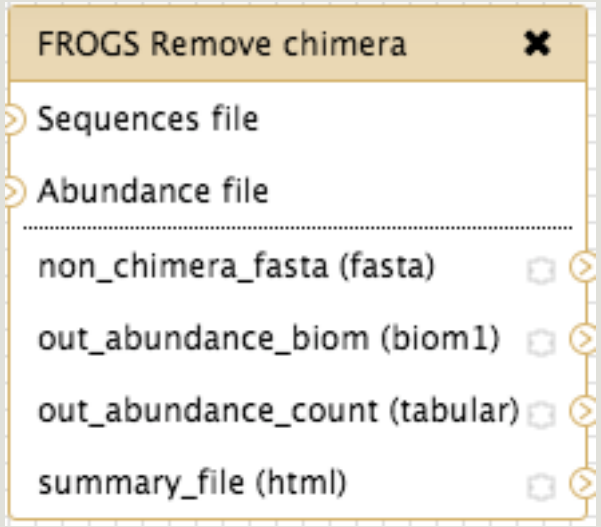


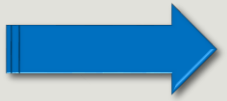


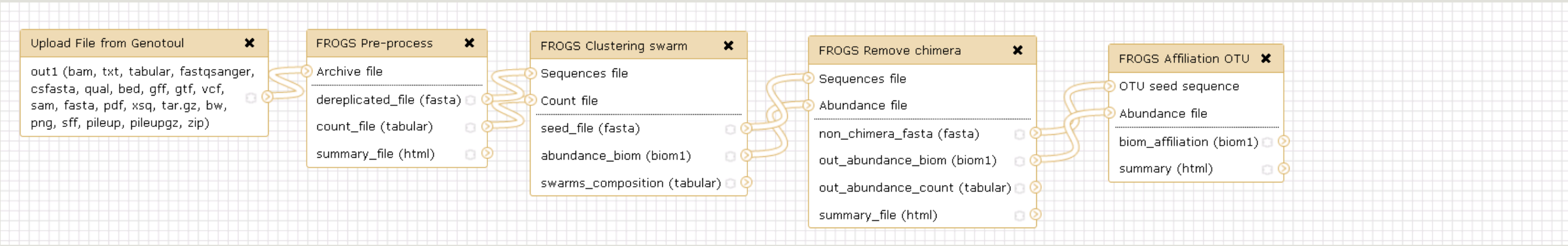


For each tool, think to:

- Fixe parameter ?
- Automatically rename output files
- Hide intermediate files ?

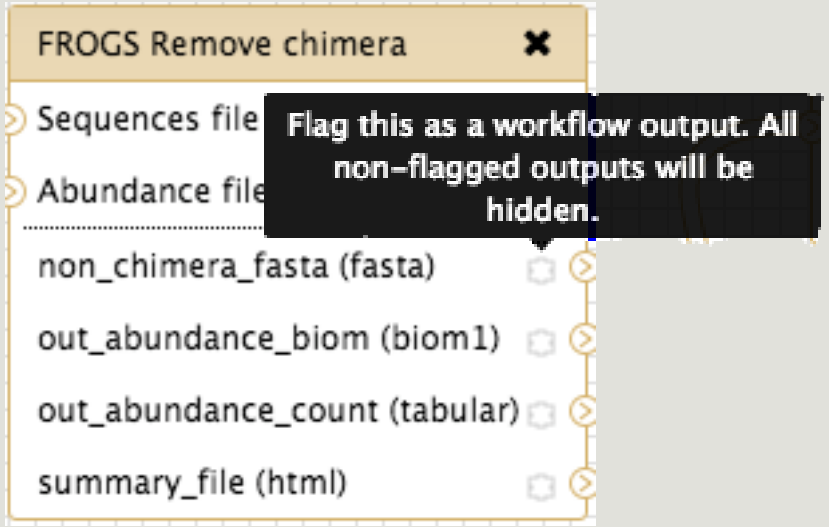


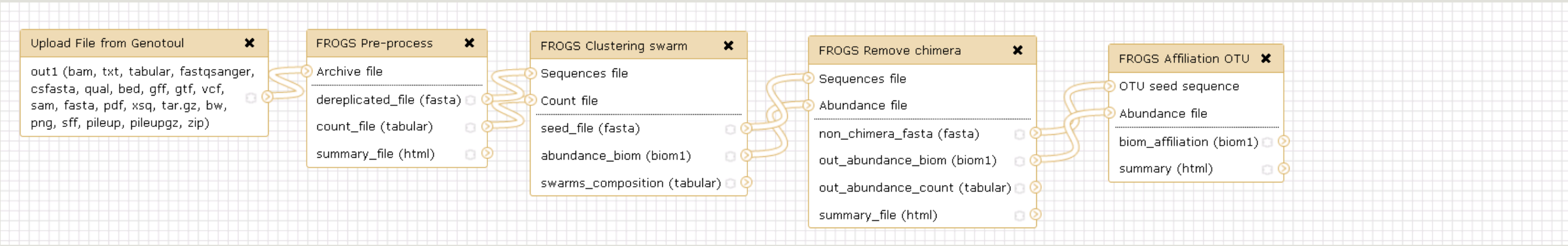




For each tool, think to:

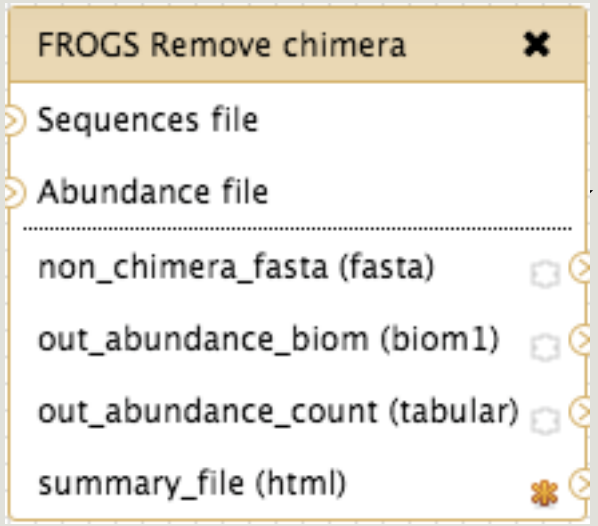
- Fixe parameter ?
- Automatically rename output files
- Hide intermediate files ?

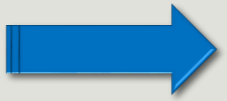


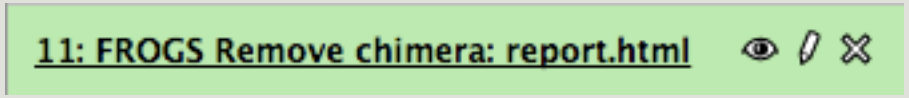


For each tool, think to:

- Fixe parameter ?
- Automatically rename output files
- Hide intermediate files ?







Back to the main documentation