

D- Training on Galaxy: Metabarcoding

March 2020 — Nancy

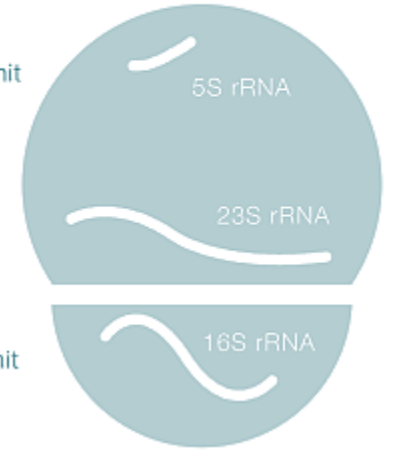
ITS analysis Praticice

MARIA BERNARD, OLIVIER RUÉ, GÉRALDINE PASCAL

What is a ITS ?

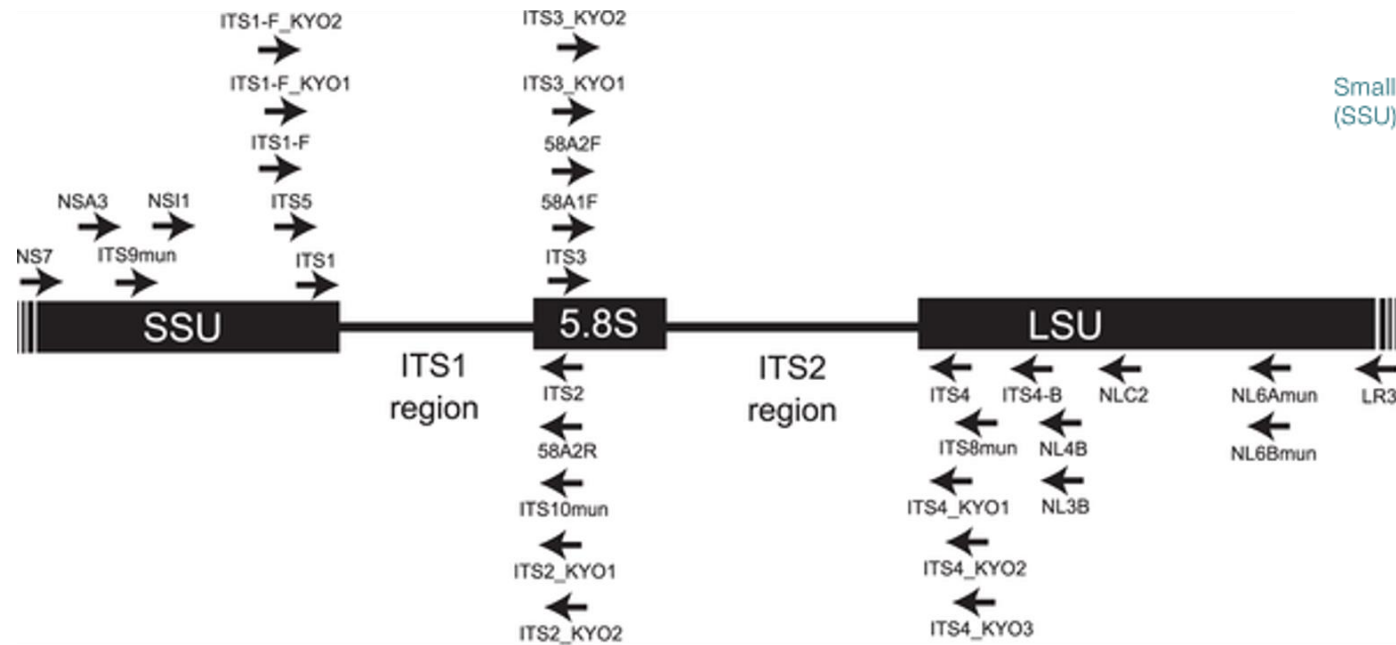
Prokaryotic Ribosome

Large Subunit (LSU)



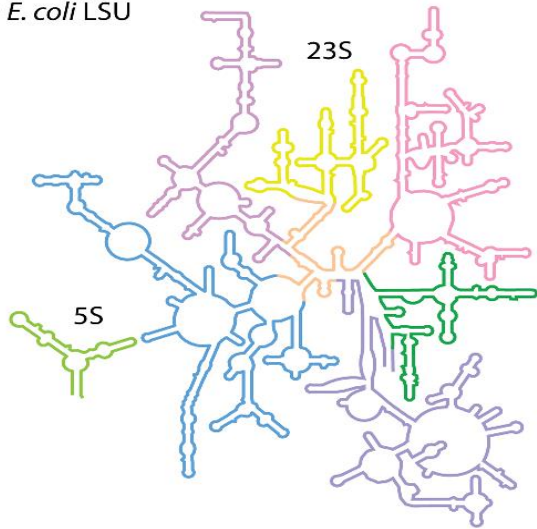
Small Subunit (SSU)

Map of nuclear ribosomal RNA genes and their ITS regions.

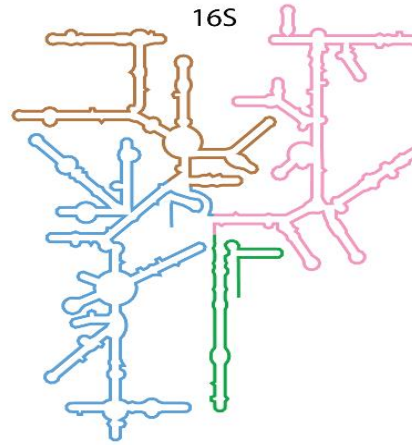


Toju H, Tanabe AS, Yamamoto S, Sato H (2012) High-Coverage ITS Primers for the DNA-Based Identification of Ascomycetes and Basidiomycetes in Environmental Samples. PLOS ONE 7(7): e40863. <https://doi.org/10.1371/journal.pone.0040863>

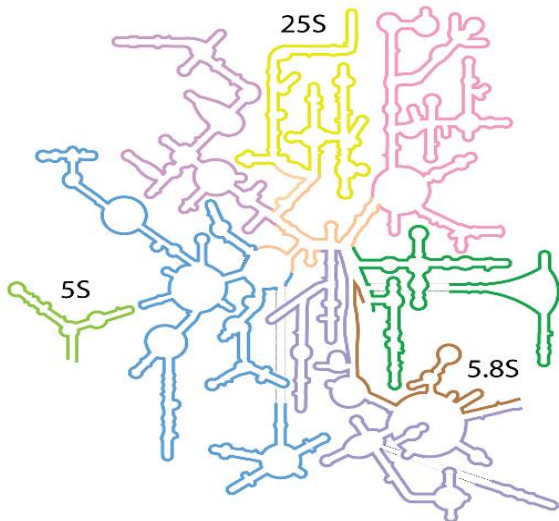
a) *E. coli* LSU



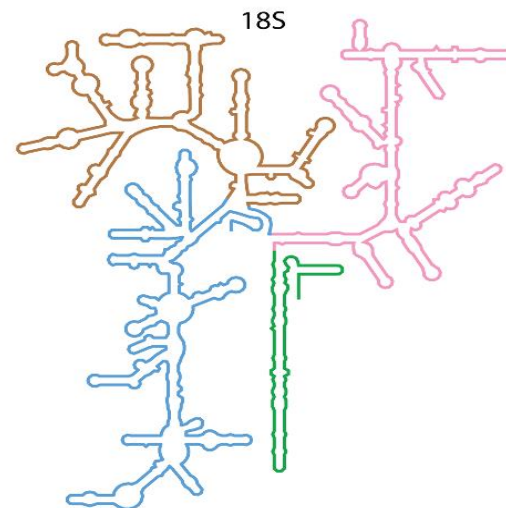
b) *E. coli* SSU



c) *S. cerevisiae* LSU



d) *S. cerevisiae* SSU



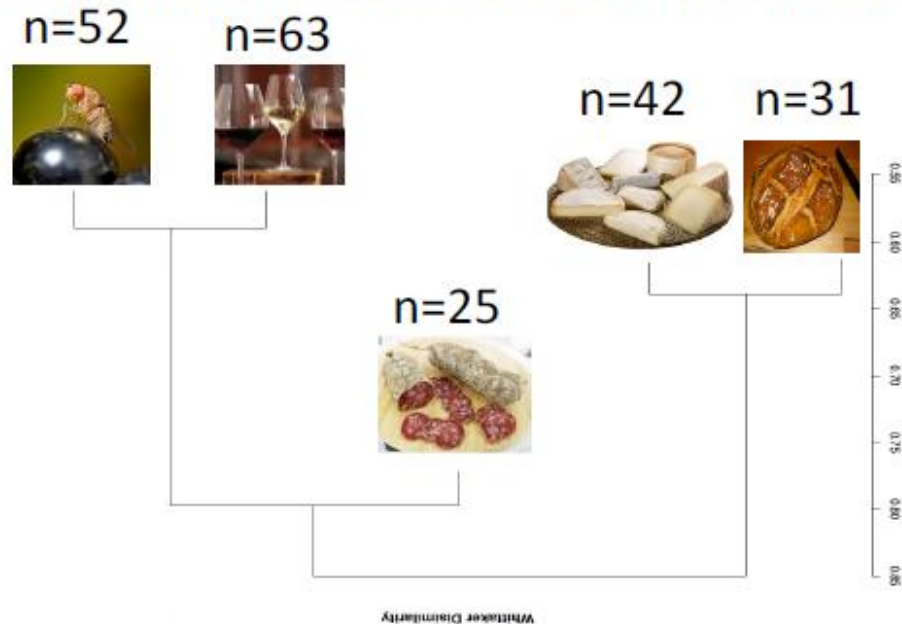
Schematic rRNA 2° structures of a) *E. coli* LSU, b) *E. coli* SSU, c) *S. cerevisiae* LSU, and d) *S. cerevisiae* SSU. These 2° structures are derived from 3D structures, and include non-canonical base pairs.

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

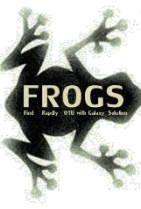
ITS data form METABARFOOD Project metaprogramme MEM

Yeast catalog in food ecosystem

Number of yeast species reported at least twice in each ecosystem and their dissimilarity between ecosystems, as measured by the Whittaker distance



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



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

Imagine a real amplicon sequence of 700bp

700bp

Imagine a Miseq paired sequencing of 2x250bp

R1 : 250bp



R2 : 250bp



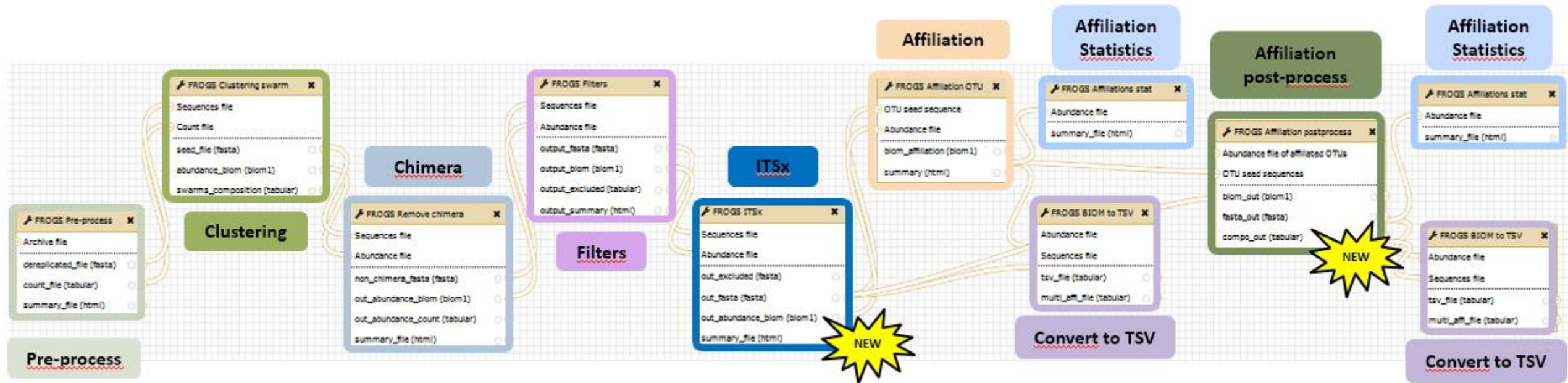
Reconstructing amplicon sequence is not possible with overlap, an arbitrary sequence of 100Ns is added. It is named « FROGS combined »

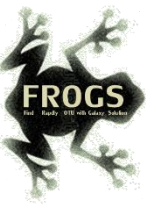


Combined sequence length : 600bp, with 100 Ns

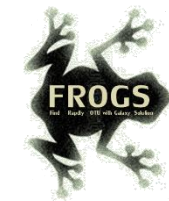


What workflow should we use to analyse ITS ?





Pre-process tool

**Sequencer**


Illumina

Select the sequencing technology used to produce the sequences.

Input type

Archive

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

Archive file 1: /work/frogsfungi/ITS.tar.gz

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

Reads already merged ?

No

The archive contains 1 file by sample : R1 and R2 are already merged by pair.

Reads 1 size

250

The maximum read1 size.

Reads 2 size

250

The maximum read2 size.

mismatch rate.

0.1

The maximum rate of mismatch in the overlap region

Merge software

Vsearch

Select the software to merge paired-end reads.

Would you like to keep unmerged reads? Yes No

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

To keep FROGS combined sequences, choose YES

**Minimum amplicon size**

The minimum size for the amplicons (with primers).

Maximum amplicon size

The maximum size for the amplicons (with primers).

Sequencing protocol

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

5' primer

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

3' primer

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

Primer 5': CTTGGTCATTAGAGGAAGTAA
Primer 3': GCATCGATGAAGAACGCAGC

Exercise

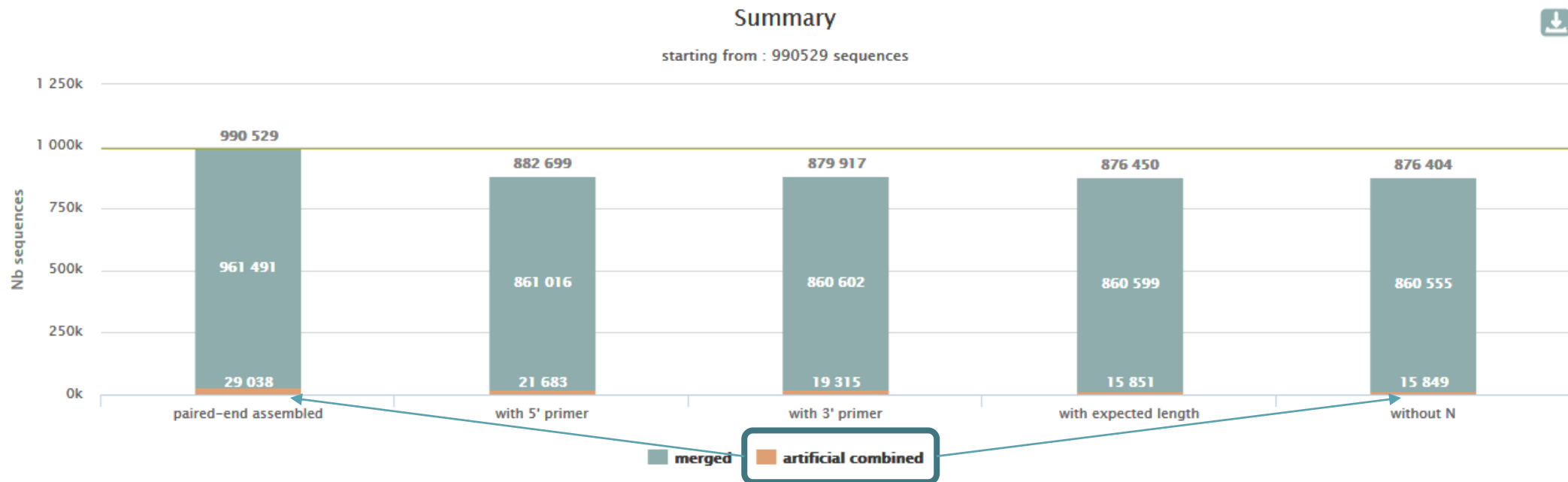
Go to « [ITS](#) » history

Launch the pre-process tool on this data set

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

Explore Preprocess report.html

Preprocess summary



Explore Preprocess report.html

2 tables:

Details on merged sequences

Show entries Search: [CSV](#)

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

Details on **artificial combined sequences**

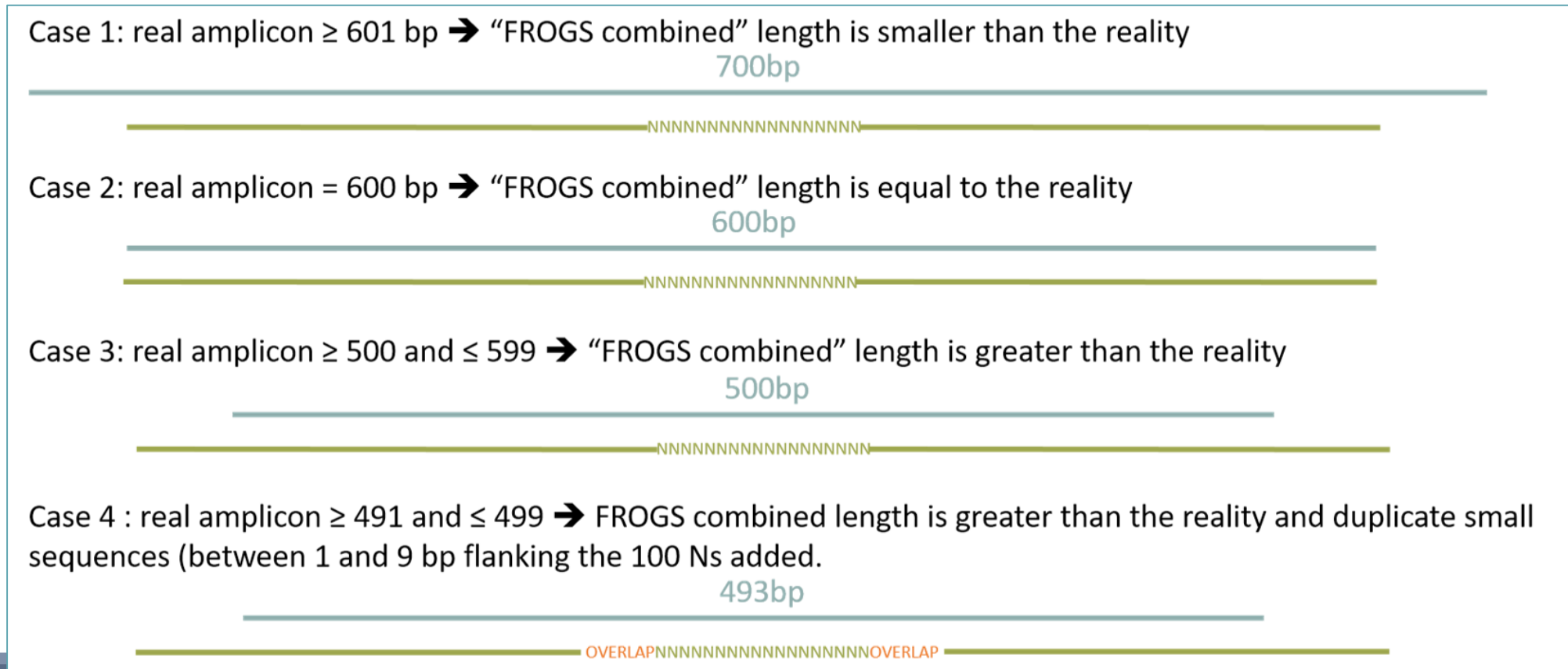
Show entries Search: [CSV](#)

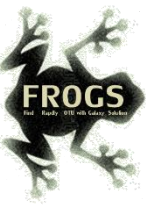
<input type="checkbox"/>	Samples	% kept	paired-end assembled	with 5' primer	with 3' primer	with expected length	without N
<input type="checkbox"/>	complexe-ADN-1	68.47	2,163	1,833	1,656	1,481	1,481
<input type="checkbox"/>	echantillon1-1	54.92	1,047	751	620	575	575
<input type="checkbox"/>	echantillon1-2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

Here, these are the too small artefactual sequences that are filtered. Amplicon_combined length (trimmed primers, with 100 N) < read_1 size.

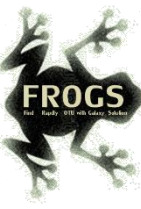
FROGS "combined" sequences are artificial and present particular features especially on size.

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



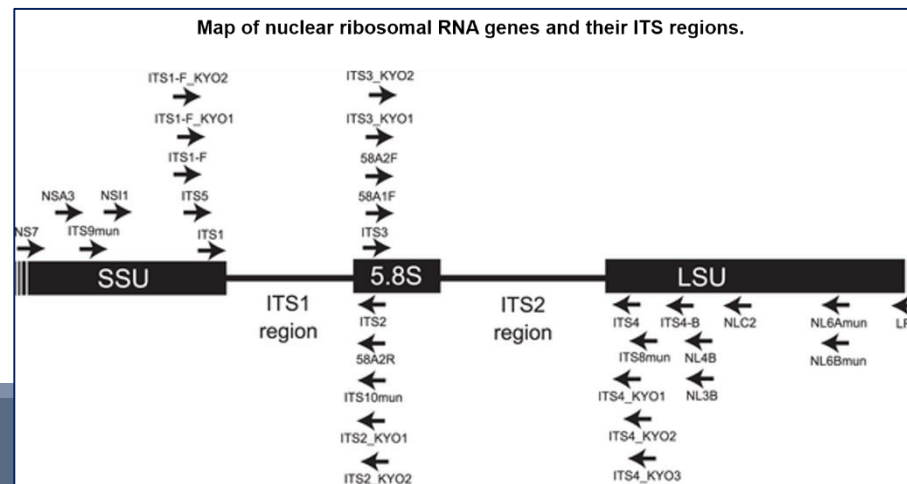


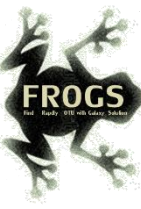
ITSx tools



What is the purpose of the ITSx tool?

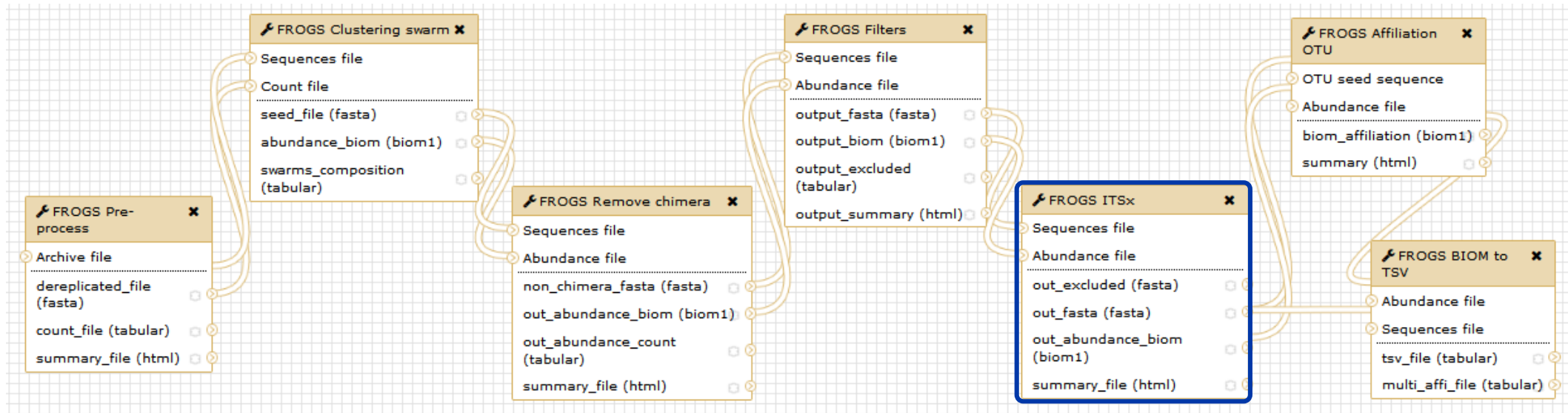
- ITSx is a tool to filter sequences.
- ITSx identifies and trims ITS regions in our sequences.
- It excludes the highly conserved neighboring sequences SSU, 5S and rRNA LSU.
- If the ITS1 or ITS2 region is not detected, the sequence is discarded.
- You can choose to check only if the sequence is detected as an ITS.
- In this case, the sequence is not trimmed, only sequences not detected as ITS are rejected (*e.g.* contaminants).





When should we use ITSx ?

After filtering !





FROGS ITSx Extract the highly variable ITS1 and ITS2 subregions from ITS sequences. (Galaxy Version r3.0-1.0)

Options

Sequences file

13: FROGS Filters: sequences.fasta

The sequence file to filter (format: fasta).

Abundance file

14: FROGS Filters: abundance.biom

The abundance file to filter (format: BIOM).

ITS region

ITS1

Which fungal ITS region is targeted: either ITS1 or ITS2

Check only if sequence detected as ITS ?

Yes No

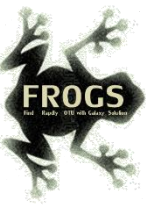
If Yes, sequences with ITS signature will be kept without trimming SSU, LSU or 5.8S regions.

Execute



Check only if sequence is detected as ITS? Yes or not?

- It is interesting to keep only the ITS parts without the flanking sequences in case one would like to compare sequenced amplicons with different primers targeting the same region to be amplified.
- You can choose this option on configuration panel of ITSx Tool.
- Reply "No" to question "Check only if sequence is detected as ITS?".
- In opposite, if "Yes" is chosen, sequences with ITS signature will be kept without trimming SSU, LSU or 5.8S regions.



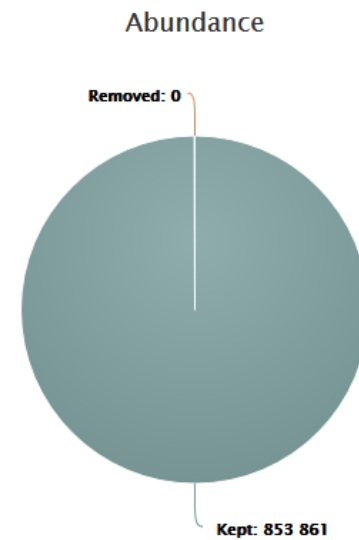
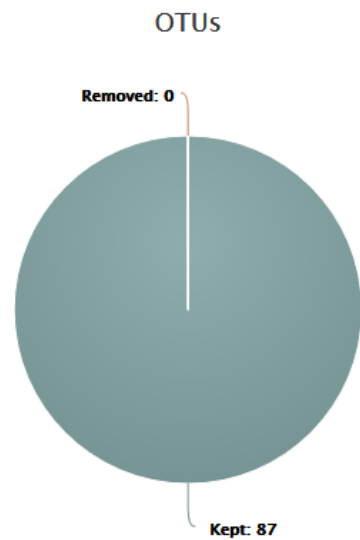
Carreful !

- The ITSx step is time consuming and has to be done on clusters. We advise our users to apply ITSx in 5th step:
 1. Preprocess step,
 2. Clustering step,
 3. Chimera removing step,
 4. Filter on OTUs abundances and replicats step,
 5. ITSx if Fungi ITS amplicons.

- Careful, ITSx is currently usable for the detection of fungi ITS neither plants nor other eukaryotes.



Filters (ITSx) summary



Filters (ITSx) by samples

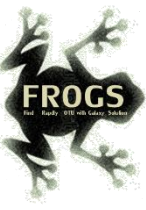
Show entries

 CSV

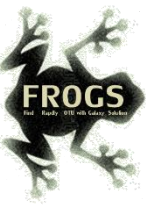
Search:

OTUs removed by sample

Sample name	↑↓ Initial	↑↓ Kept	↑↓ Initial abundance	↑↓ Kept abundance	↑↓
complexe-ADN-1	65	65	47,980	47,980	
echantillon1-1	63	63	26,797	26,797	
echantillon1-2	64	64	51,499	51,499	
echantillon1-3	66	66	68,500	68,500	

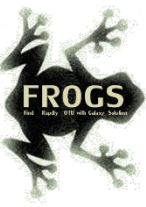


ITS Affiliation



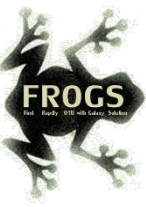
What is special about the affiliation of ITS (with combined sequences more broadly)?

- blastn+ or needlall is used to find alignment between each OTU and the database.
- Only the bests hits with the same score are reported.
- blastn+ is used for merged read pair, and needall is used for **artificially combined sequence**.
- For each alignment returned, several metrics are computed: identity percentage, coverage percentage, and alignment length.
- If "combined" sequences are stayed presents in OTUs, blastn+ is not usable as for classical merged sequences.

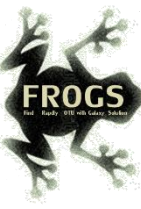


What is special about the affiliation of ITS (with combined sequences more broadly)?

- So, sequences are affiliated in 3 steps: Alignment of classical "merged" sequences with blastn+ *versus* chosen database (*e.g.* UNITE),
- Alignment of "combined" sequences with blastn+ *versus* chosen database, best hits are collected and a very small new databank (at most 200 references per blast hit) is created composed exclusively of "subject" sequences from these best hits,
- Alignment of "combined" sequences with needlall (global alignment: very time consuming) *versus* these small new databank.



Careful, with "combined" sequences, we introduced some modification on identity percentage



Case 1: a sequencing of overlapping sequences i.e. 16S V3-V4 amplicon MiSeq sequencing

Imagine a real amplicon sequence of 400bp



Reconstructing amplicon sequence is a merged sequence (length : 400bp, with 100bp overlap)

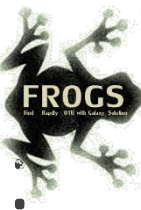


Affiliation is notably made by a local alignment with NCBI Blast+



Imagine a perfect sequencing without error:

classical %id = number of matches / alignment length = 400 matches / 400 positions = 100% identity



Case2: a sequencing of non-overlapping sequences: case of ITS1 amplicon MiSeq sequencing

Imagine a real amplicon sequence of 700bp

700bp

Reconstructing a FROGS combined sequence (length : 600bp, with 100Ns)

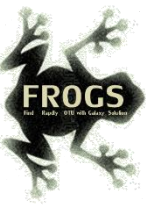
NNNNNNNNNNNNNNNNNN

Affiliation could not be made by a local alignment but with a global alignment with Emboss needle

NNNNNNNNNNNNNNNNNN

Imagine a perfect sequencing without error:

classical %id = number of matches / alignment length = (250+250 matches) / 700 positions = 71%



Conclusion on identity percentage for ITS

Filtering on %id will systematically removed “FROGS combined” OTUs.

So, we proposed to replace the classical %id by a %id computed on the sequenced bases only.

% sequenced bases identity = number of matches / (seed length – artificial added N)

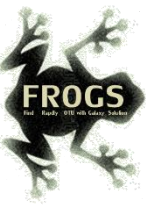
Case 1 : 16S V3V4 → overlapped sequence

% sequenced bases identity = 400 matches / 400 bp = **100 %**

Case 2 : very large ITS1 → “FROGS combined” shorter than the real sequence

% sequenced bases identity = (250 + 250) / (600 - 100) = **100%**

This calculation allows the 100% identity score to be returned on FROGS "combined" shorter or longer than reality in case of perfect sequencing. And returns a lower percentage of identity in the case of repeated small overlaps kept in the FROGS "combined".



Affiliation Post-process



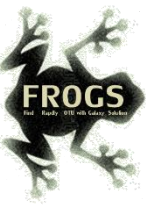
What is the purpose of the *Affiliation post-process* tool ?

This tool allows grouping OTUs together in accordance with the %id and %cov chosen by the user and according to the following criteria:

1. They must have the same affiliation

Or

2. If they have "multi-affiliation" tag in FROGS taxonomy, they must have in common in their list of possible affiliations at least one identical affiliation.



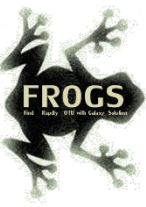
What is the purpose of the *Affiliation post-process* tool ?

In consequence:

The different affiliations involved in multi-affiliation are merged.

The abundances are added together.

It is the most abundant OTU seed that is kept.



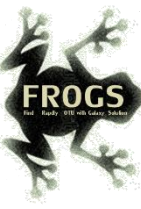
What is the purpose of the *Affiliation post-process* tool ?

In case of ITS amplicon analyses,

you may have ambiguities due to inclusive ITS sequence coming from different species.

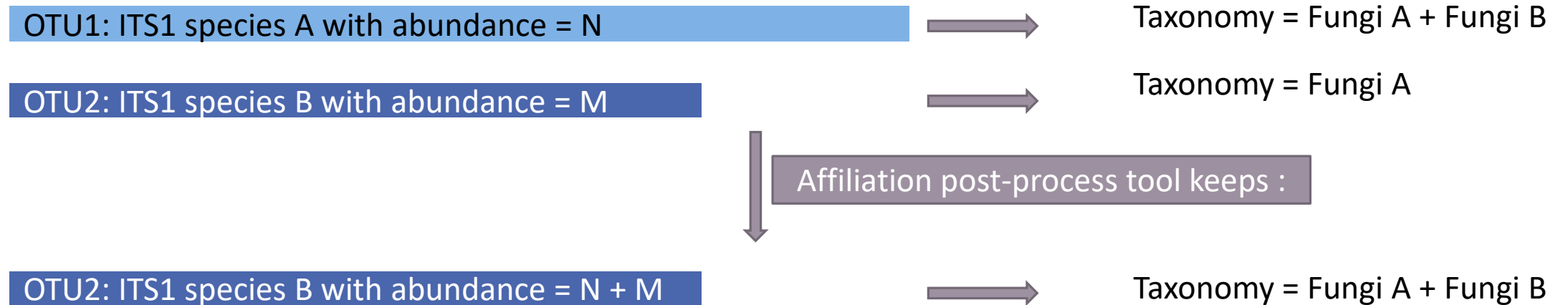
The tool will keep affiliation of the shortest sequence in case of multi-affiliation tag.

This "Affiliation post-process" tool helps to resolve ambiguities due to potentially inclusive sequences such as ITS.



What is the purpose of the *Affiliation post-process* tool ?

ITS1 blue is completely included (with 100% identity) in ITS1 yellow





FROGS Affiliation postprocess Optional step to resolve inclusive amplicon ambiguities and to aggregate OTUs based on alignment metrics (Galaxy Version r3.0-1.0)

Options

Abundance file of affiliated OTUs

23: FROGS ITSx: itsx.biom

Abundances of affiliated OTUs (format: BIOM).

OTU seed sequences

22: FROGS ITSx: itsx.fasta

OTU sequences (format: fasta).

Is this an hyper variable in length amplicon ?

Yes No

Multi-affiliation tag may be resolved by selecting the shortest amplicon reference. For this you need the reference fasta file of your kind of amplicon.

Using reference database

UNITE_7.1_ITS1

Which ITS 1 or 2 do you want to analyze?

UNITE_7.1_ITS1

UNITE_7.1_ITS2

OTUs will be aggregated if they share the same taxonomy with at least X% identity.

minimum coverage for aggregation

99

OTUs will be aggregated if they share the same taxonomy with at least X% alignment coverage.

Execute

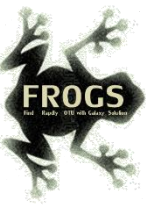


Post-affiliation Tool - output

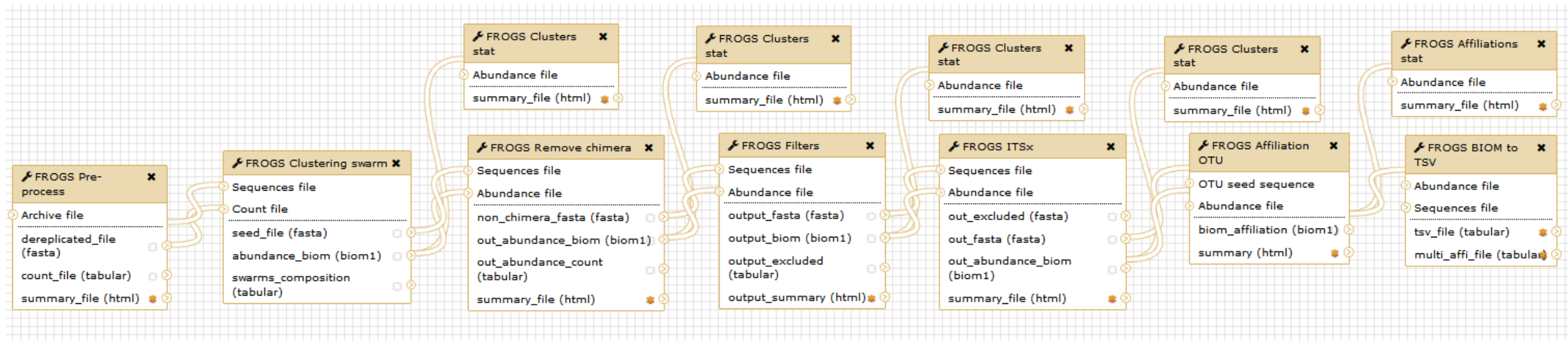
```
Cluster_1 Cluster_781 Cluster_922 Cluster_930 Cluster_3573 Cluster_1298 Cluster_798 Cluster_738 Cluster_918
Cluster_2 Cluster_313 Cluster_469 Cluster_445 Cluster_105 Cluster_912 Cluster_471 Cluster_1152 Cluster_1145
Cluster_3 Cluster_599 Cluster_114
Cluster_4 Cluster_109
Cluster_5 Cluster_140 Cluster_3850
Cluster_6 Cluster_195 Cluster_905 Cluster_388 Cluster_275
Cluster_7
Cluster_8
Cluster_9
Cluster_10
Cluster_11
Cluster_12
Cluster_13
Cluster_14
Cluster_15
Cluster_16
Cluster_17
Cluster_18
Cluster_20
Cluster_19
Cluster_21
Cluster_22
Cluster_23
Cluster_25
Cluster_24
Cluster_26
```

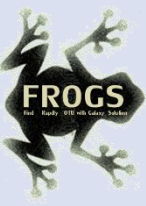
Cluster 1 encapsulate also clusters 781, 922, 930, 3573, 1298, 798 and 918

```
>Cluster_1 reference=AB241105 amplicon=1..444 position=1..444 errors=440%A
TGGGGAATATTGCACAATGGGGGAAACCCTGATGCAGCGACGCCGCGTGCGGGATGACGGCCCTTCGGGTGTAAACCGCTTTTAAATGGGAGCAAGCAGTTTTA
>Cluster_2 reference=AJ496032 amplicon=1..452 position=1..452 errors=440%G
TAGGGGAATATTGCACAATGGGGGAAACCCTGATGCAGCAACGCCACGTGTGGGAAGAAGCATTTCGGTGTGTAAACCACTGTCATGAGGGAATAAGGCCCGCCT
>Cluster_3 reference=EU240886 amplicon=1..460 position=1..460
TAGGGGAATCTTCCGCAATGGGGGAAAGCCTGACGGAGCAACGCCGCGTGTGCGAAGAAGGCTTTCGGATCGTAAAGCACTGTTGTTAAGGAAGAACGACAGTAA
>Cluster_4 reference=U39399 amplicon=1..459 position=1..459
TGGGGAATATTGGACAATGGGGGAAACCCTGATCCAGCCATGCCGCGTGTGTGAAGAAGGCCCTTTGGTGTAAAGCACTTTAAGCAGTGAAGAAGACTCCGTG
>Cluster_5 reference=FR733705 amplicon=1..452 position=1..452 errors=436%C
TGGGGAATATTGGACAATGGGGGAAAGCCTGATCCAGCGACGCCGCGTGAAGGAAGAAGTCTTTCGGGATGTAAACTTCTGAACTAATCGAATAAGAGGGTAGI
>Cluster_6 reference=GU575117 amplicon=1..434 position=1..434
TGGGGAATATTGGACAATGGGGCGAAGCCTGATCCAGCCATGCCGCGTGTGTGATGAAGGCCCTAGGGTGTAAAGCACTTTCAACGGTGAAGATAATGACGGI
>Cluster_7 reference=AB272165 amplicon=1..454 position=1..454 errors=441%A
TAGGGGAATATTGGCAATGGGTGAGAGCCTGACCCAGCCATGCCGCGTGCCTGCGGACGAAGGCCCTCAGGGTGTAAACCGCTTTTCCGGGGAAGAAGAGGTTCC
>Cluster_8 reference=AJ292759 amplicon=1..437 position=1..437
TAGGGGAATATTGCACAATGGAGGAACTCTGATGCAGCGACGCCGCGTGAAGTGTGAAGGCCCTTCGGGTCGTAAAGCTCTGTGCGAGGGGAATAACACAATGAA
>Cluster_9 reference=CP000027 amplicon=1..435 position=1..435
CAAGGAATCTTGGGCAATGGGGGAAAGCCTGACCCAGCAACGCCGCGTGAAGGCTTTCGGGTGTAAACCTCTTTTCACAGGGAAGAATAATGACGG
>Cluster_10 reference=JN880417 amplicon=1..438 position=1..438 errors=11%A
TCGAGGATCTTCGTCATGGGGGAAAGCCTGAACGAGCGATTAGCCGCGTGCCTGATGAAGGCCCTTCGGGTGTAAAGCGGAAAGAGGTAATAAAGGGAAACT
>Cluster_11 reference=EF660760 amplicon=1..434 position=1..434 errors=424%C
TGGGGAATCTTGCACAATGGGGGAAAGCCTGATGCAGCCATGCCGCGTGAATGTGAAGGCCCTTAGGGTGTAAATCTTTCCGCCAGGGATGATAATGACAGI
>Cluster_12 reference=AB594446 amplicon=1..443 position=1..443 errors=438%G
TGGGGAATATTGCACAATGGGGCGAAGCCTGATGCAGCGACGCCGCGTGGGGGATGACGGCCCTTCGGGTGTAAACTCCTTTCCGCATTGACGAAGCCTTTTTG
>Cluster_13 reference=U93332 amplicon=1..439 position=1..439 errors=426%C
```



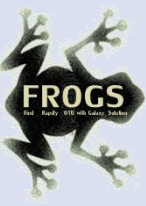
Workflow creation





Your Turn! – 10

CREATE YOUR OWN WORKFLOW !



Exercise 10

1. Create your own workflow with ITS data



Exercise 10

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](#)

Exercise 10

A screenshot of the Galaxy Sigeneae web interface showing the 'Create New Workflow' form. The browser title bar at the top reads 'Galaxy Sigeneae - Welcome gpascal' with navigation links for 'Analyze Data' and 'Workflow'. The form has a light brown header with the text 'Create New Workflow'. Below this, there are two input fields: 'Workflow Name:' containing 'Unnamed workflow' and 'Workflow Annotation:' which is empty. A descriptive sentence follows: 'A description of the workflow; annotation is shown alongside shared or published workflows.' At the bottom of the form is a 'Create' button. Two red arrows with white numbers point to the form: arrow '3' points to the 'Workflow Name' input field, and arrow '4' points to the 'Create' button.

Galaxy Sigeneae - Welcome gpascal Analyze Data Workflow

Create New Workflow

Workflow Name:
Unnamed workflow

Workflow Annotation:

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

Create

Exercise 10

Your workflows

5

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](#)

Your workflows

Name
formation workflow ▾
demoNEM2015 workflow ▾
FROGS_v1.0_06_05_2015 ▾

6

- Edit
- Run
- Share or Publish
- Download or Export
- Copy
- Rename
- View
- Delete

Workflows shared with you

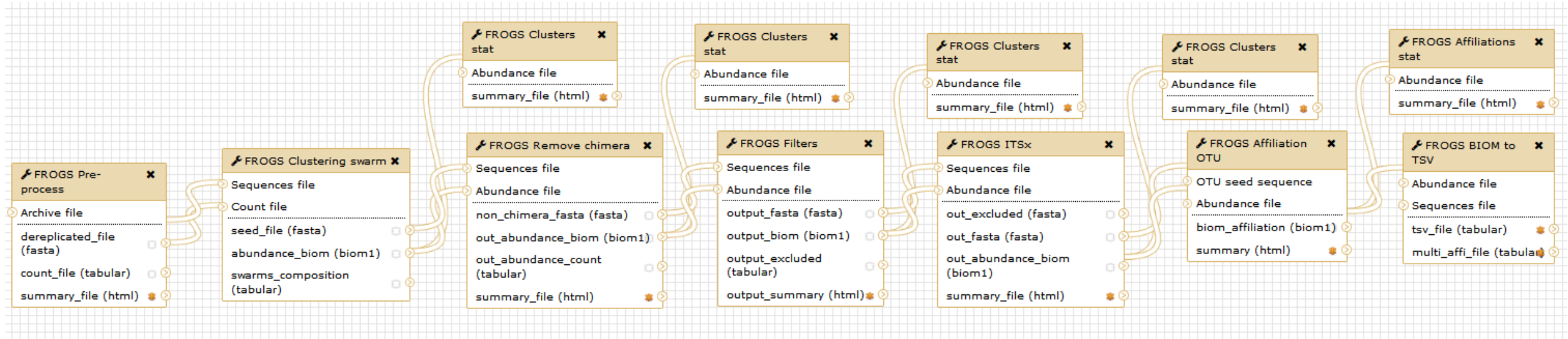
No workflows have been shared with you.

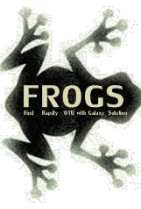
Other options

[Configure your workflow menu](#)

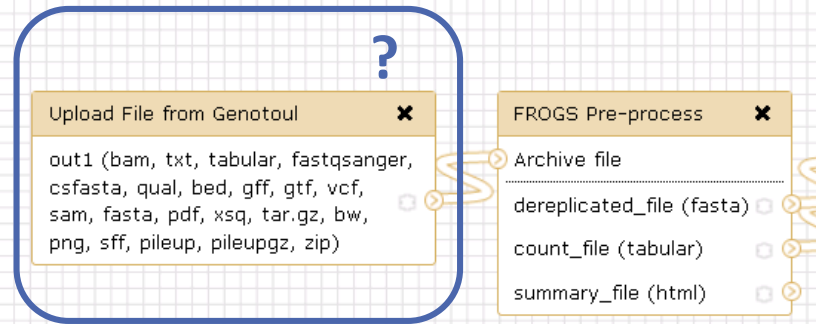


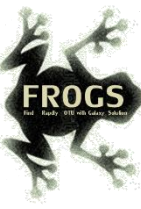
Solution of exercise:



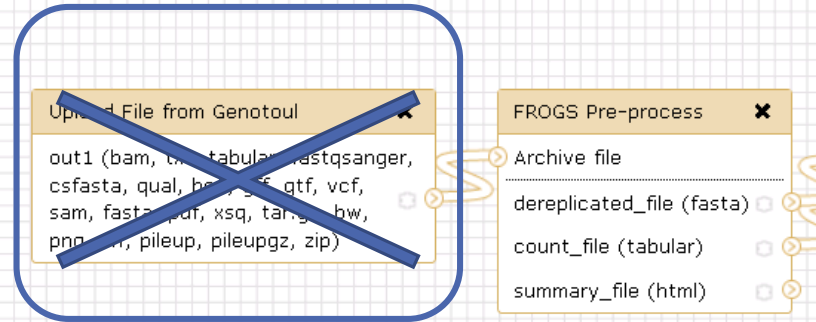


What about « upload file » ?

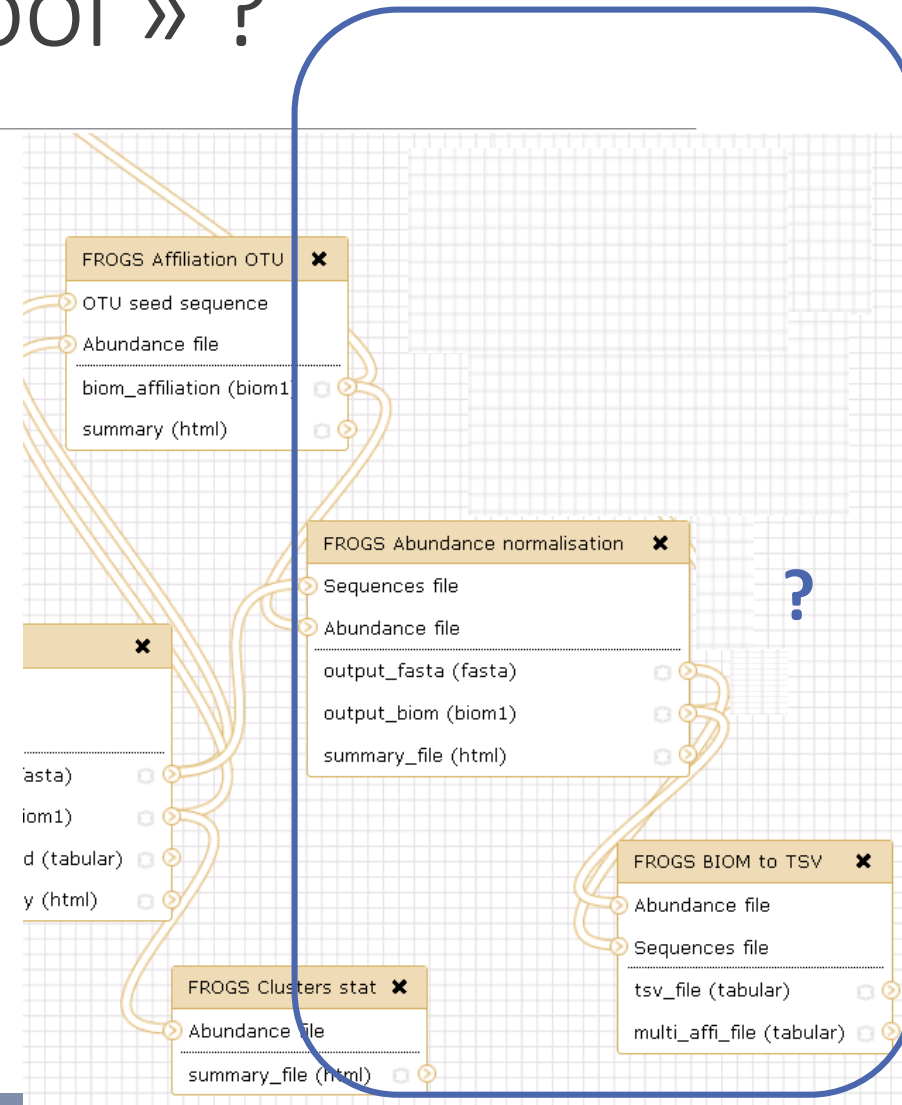




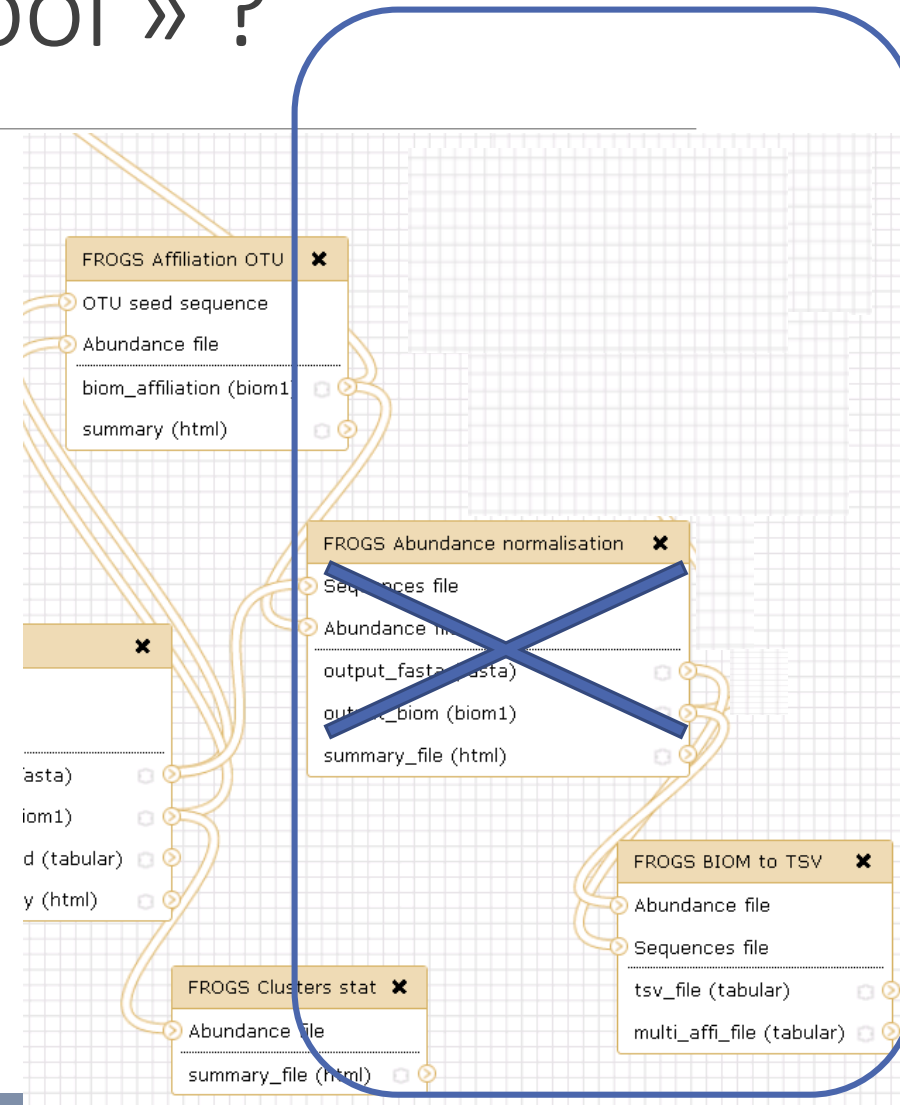
What about « upload file » ?

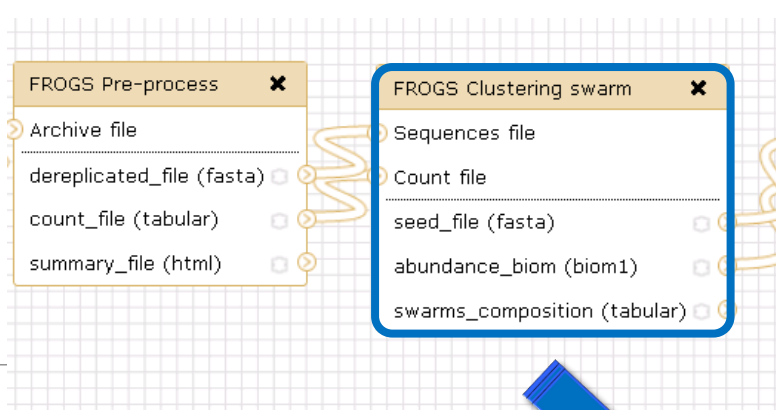


What about « Normalisation Tool » ?



What about « Normalisation Tool » ?





For each tool, think to:

- Fixe parameter ?

?

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

Sequences file
 Data input 'sequence_file' (fasta)
 The sequences file (format: fasta).

Count file
 Data input 'count_file' (tabular)
 It contains the count by sample for each sequence (format: TSV).

Aggregation distance
 Set at Runtime

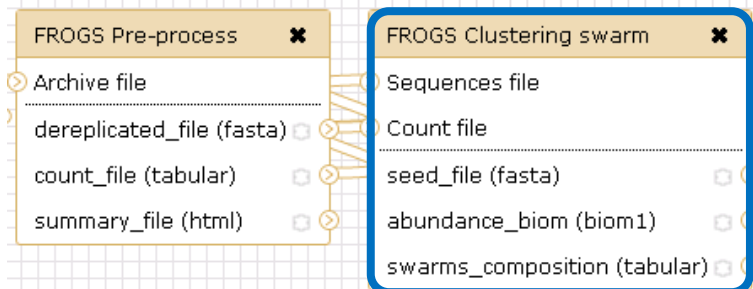
Maximum number of differences between sequences in each aggregation step.

Performe denoising clustering step?

If checked, clustering will be perform in two steps. first with

For each tool, think to:

- Fixe parameter ?
- Automatically rename output files



- Configure Output: 'seed file'
- Configure Output: 'abundance biom'
- Configure Output: 'swarms composition'



Configure Output: 'seed file'

Label

This will provide a short name to describe the output - this must be unique across workflows.

Rename dataset

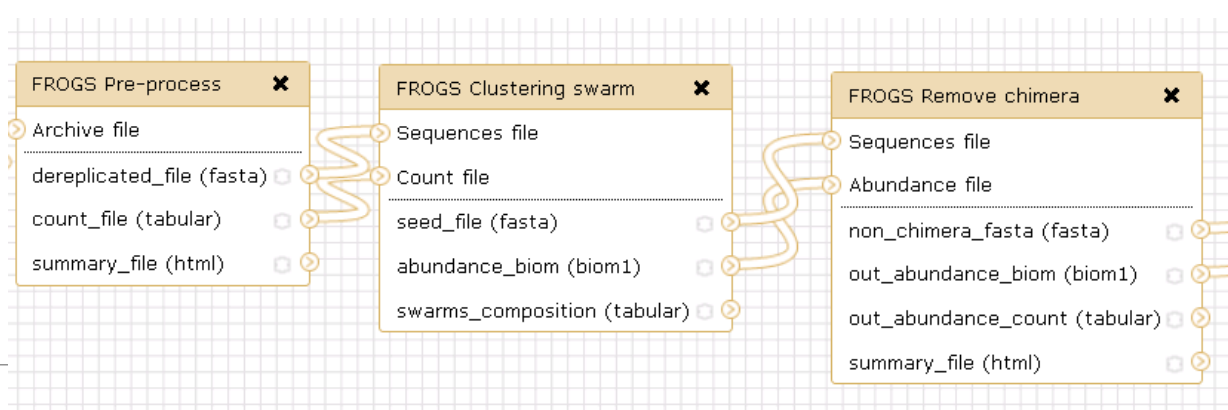
This action will rename the output dataset. Click [here](#) for more information. Valid inputs are: **sequence_file**, **count_file**.

Change datatype

This action will change the datatype of the output to the indicated value.

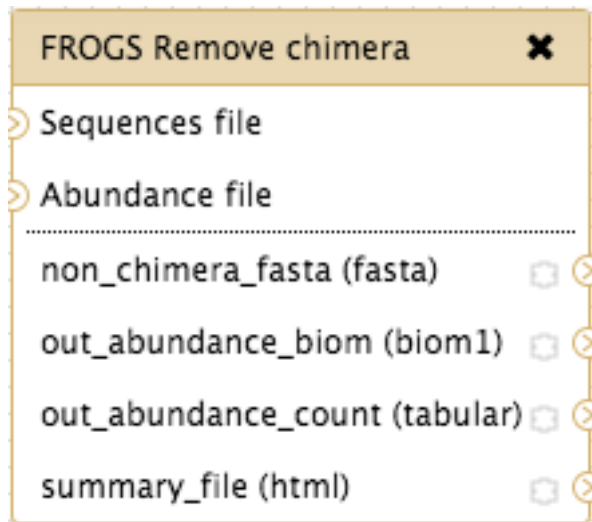
Tags

This action will set tags for the dataset.

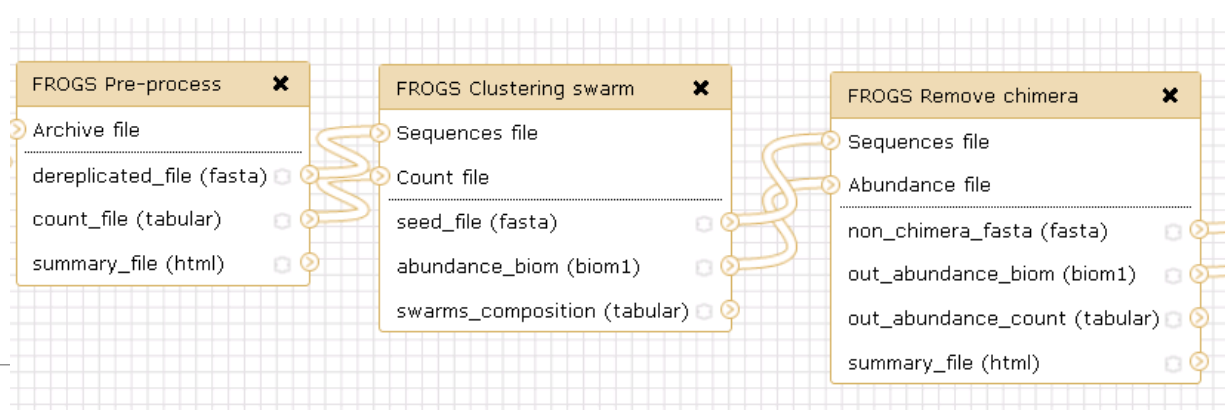


For each tool, think to:

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

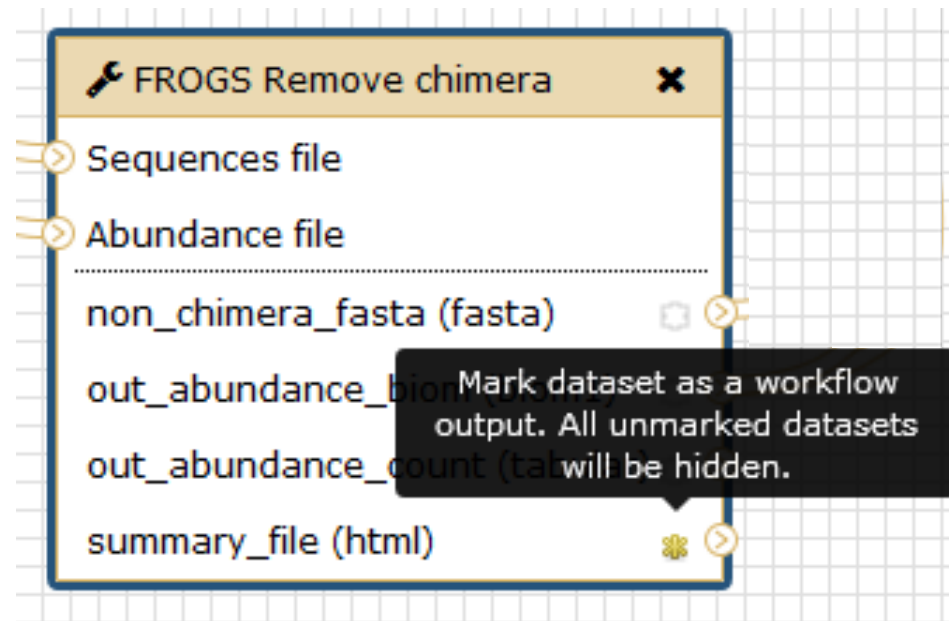


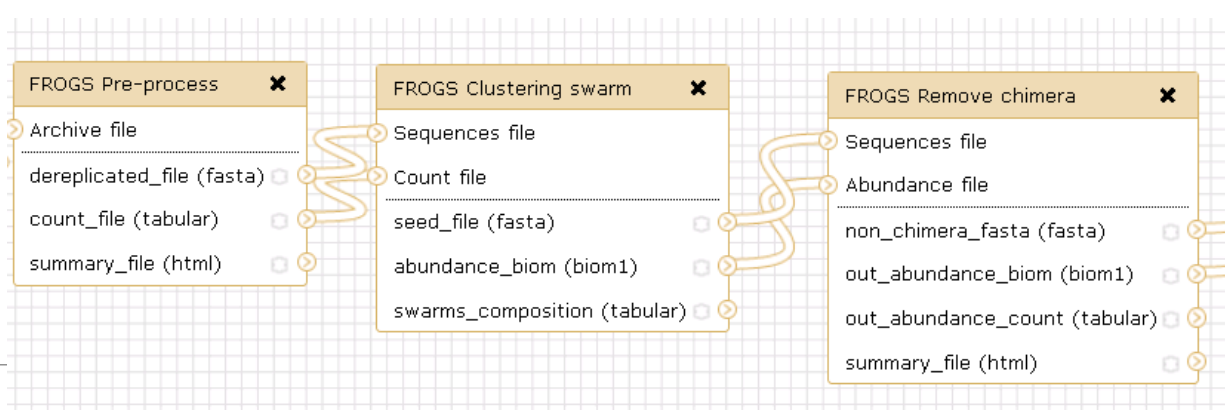
<u>11: FROGS Remove chimera: report.html</u>	👁️ ✎ ✕
<u>10: FROGS Remove chimera: non chimera abundance.biom</u>	👁️ ✎ ✕
<u>9: FROGS Remove chimera: non chimera.fasta</u>	👁️ ✎ ✕



For each tool, think to:

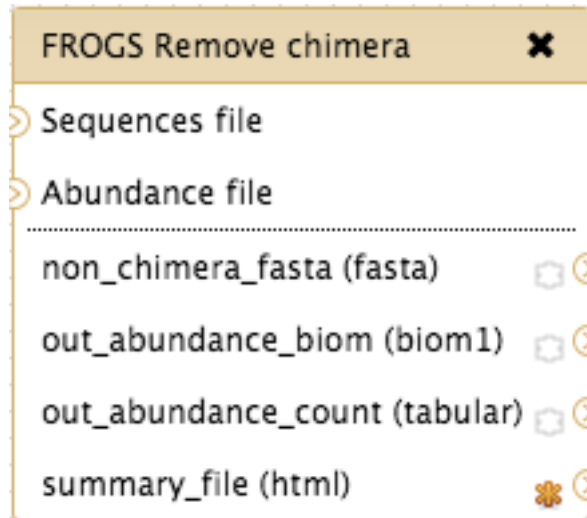
- Fixe parameter ?
- Automatically rename output files
- Hide intermediate files ?






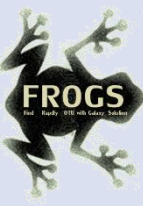


For each tool, think to:

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

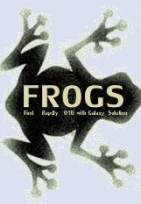


11: FROGS Remove chimera: report.html   



Your Turn! – 11

PLAY WITH YOUR OWN WORKFLOW !



Exercise 11

1. Run your own workflow with ITS data with :

http://genoweb.toulouse.inra.fr/~formation/15_FROGS/15-July2019/ITS.tar.gz

2. Import metadata for statistics analyses

http://genoweb.toulouse.inra.fr/~formation/15_FROGS/15-July2019/meta_data ITS.tsv

3. Run FROGS_stat tools