

D- Training on Galaxy: Metabarcoding

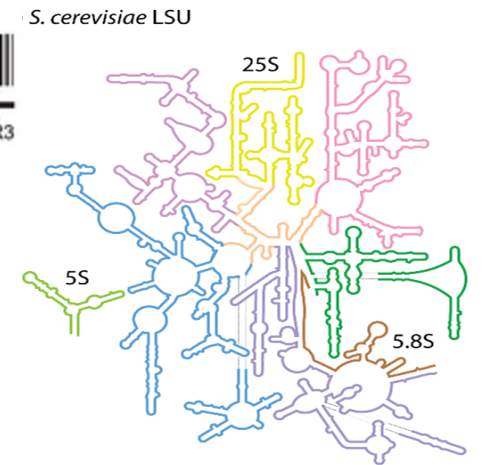
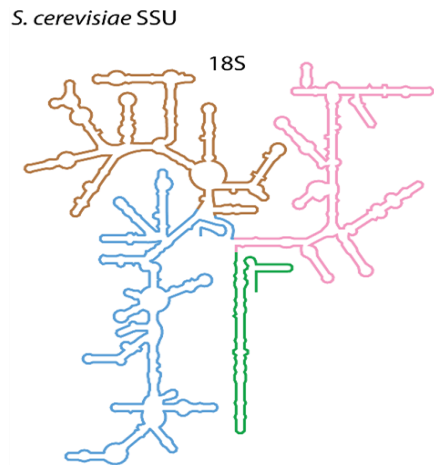
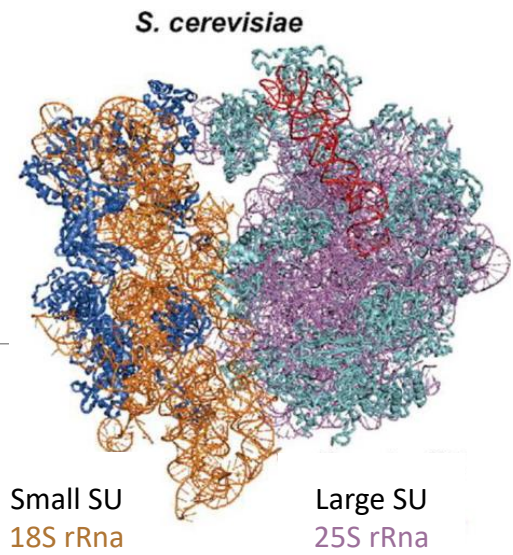
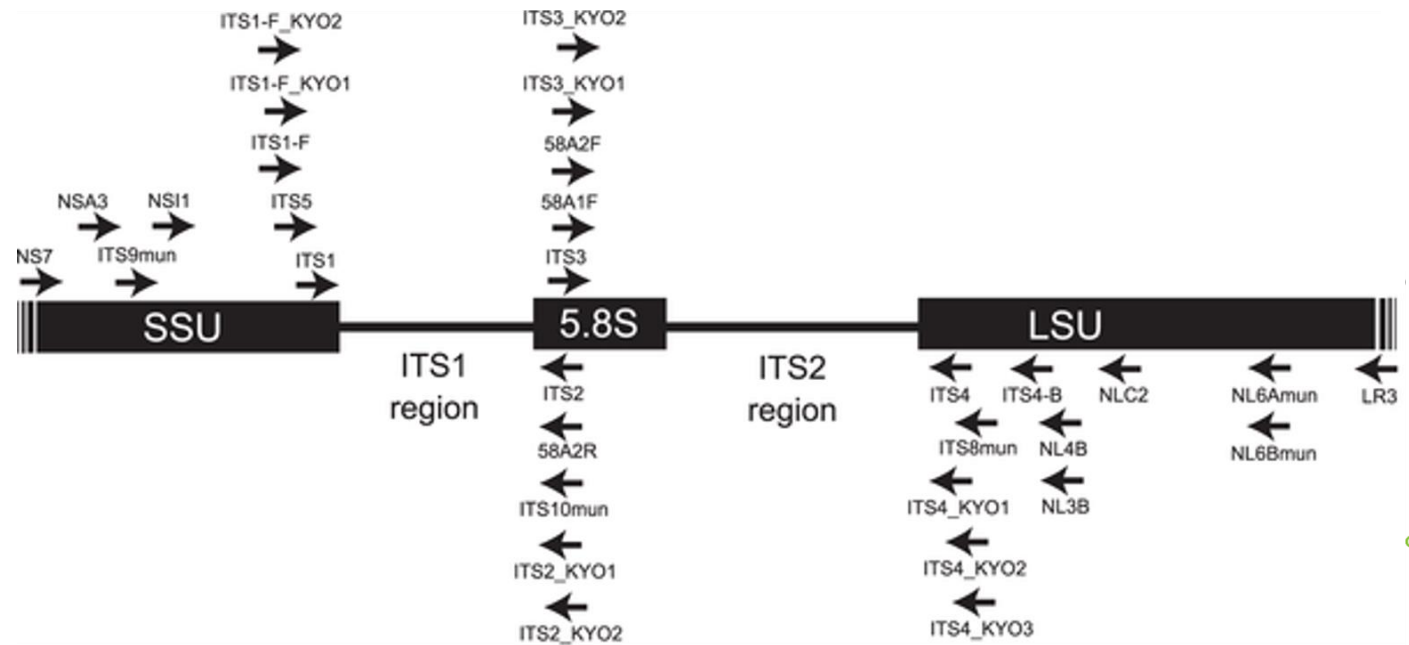
March 2021 - Webinar

FROGS Practice on ITS data

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

What is a ITS ?

Map of nuclear ribosomal RNA genes and their ITS regions.



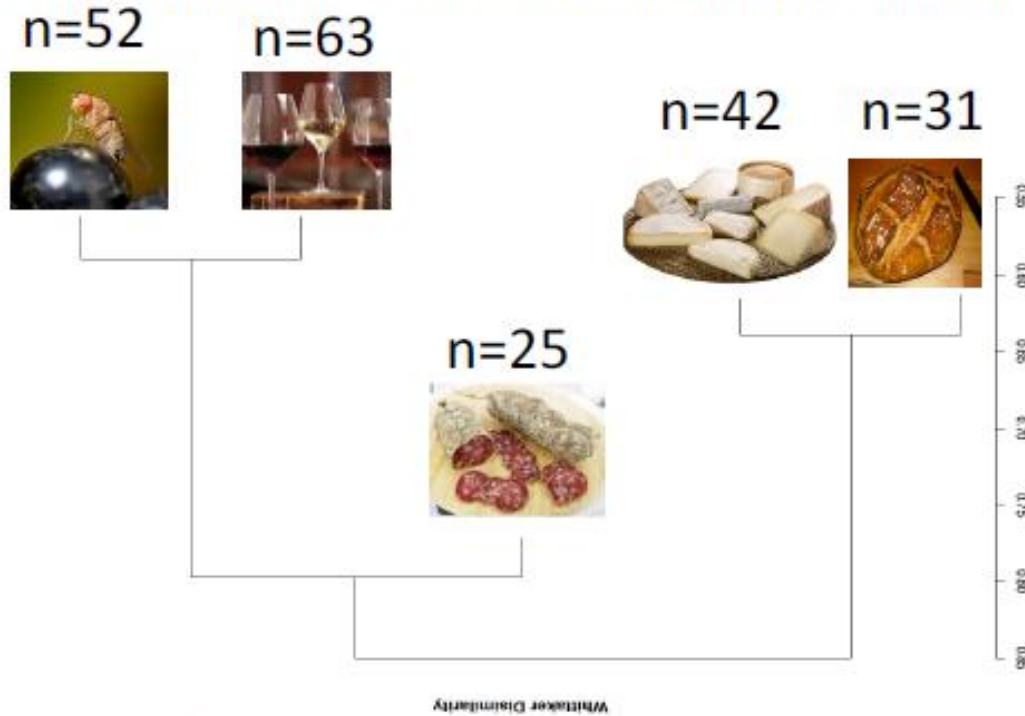
Toju H, Tanabe AS, Yamamoto S, Sato H (2012) High-Coverage ITS Primers for the DNA-Based Identification of Ascomycetes and Basidiomycetes in Environmental Samples. PLOS ONE 7(7): e40863. <https://doi.org/10.1371/journal.pone.0040863>

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

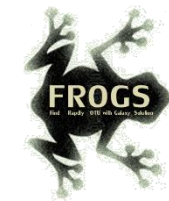


The universal fungal barcode, the Internal Transcribed Spacer (ITS) region, displays considerable size variation amongst yeasts and other micro-eukaryotes.

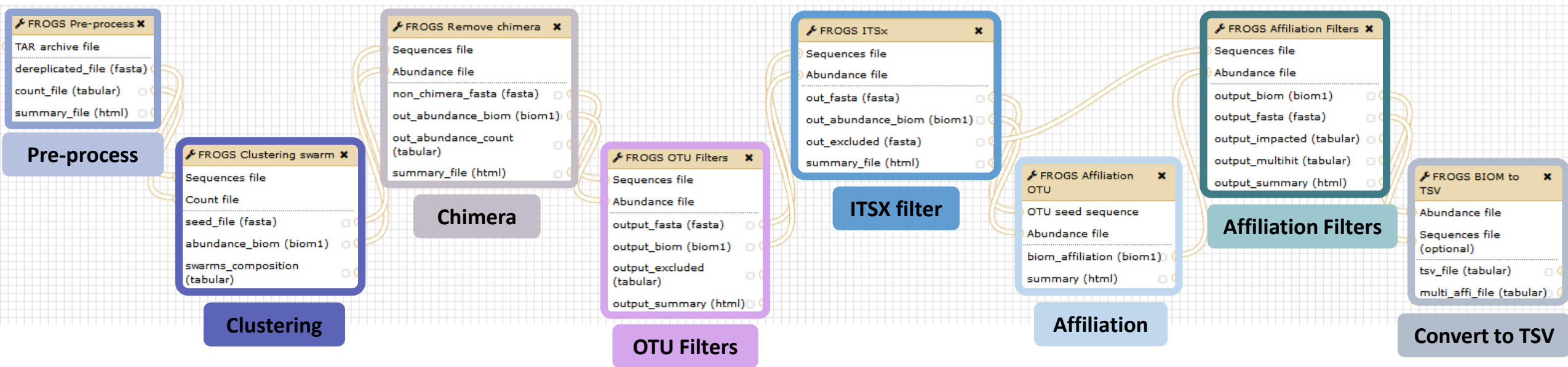
There are also several repeats leading to sequencing errors or termination.

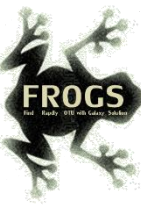


FROGS Pipeline

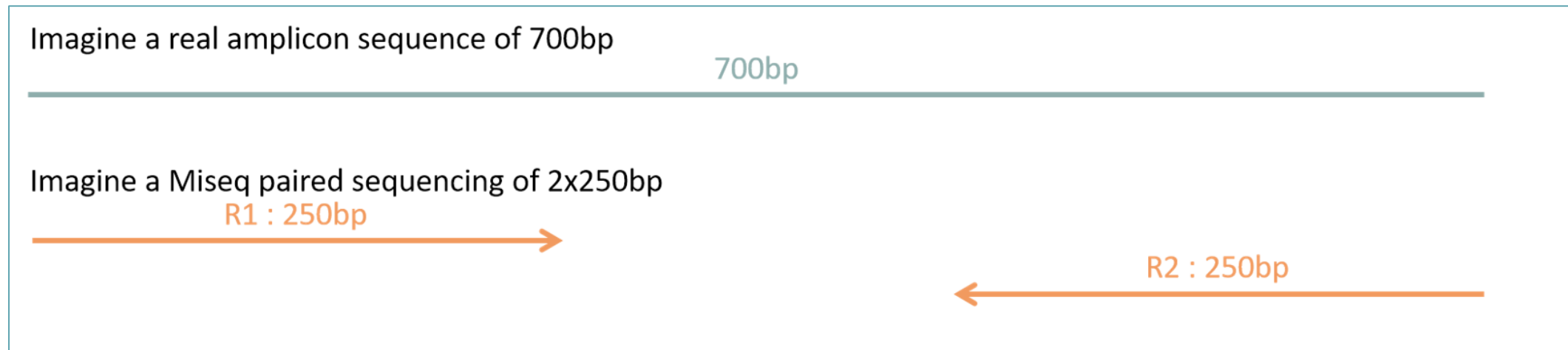


Minimal pipeline for ITS amplicon analyses

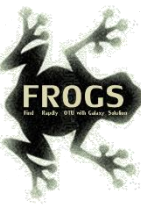




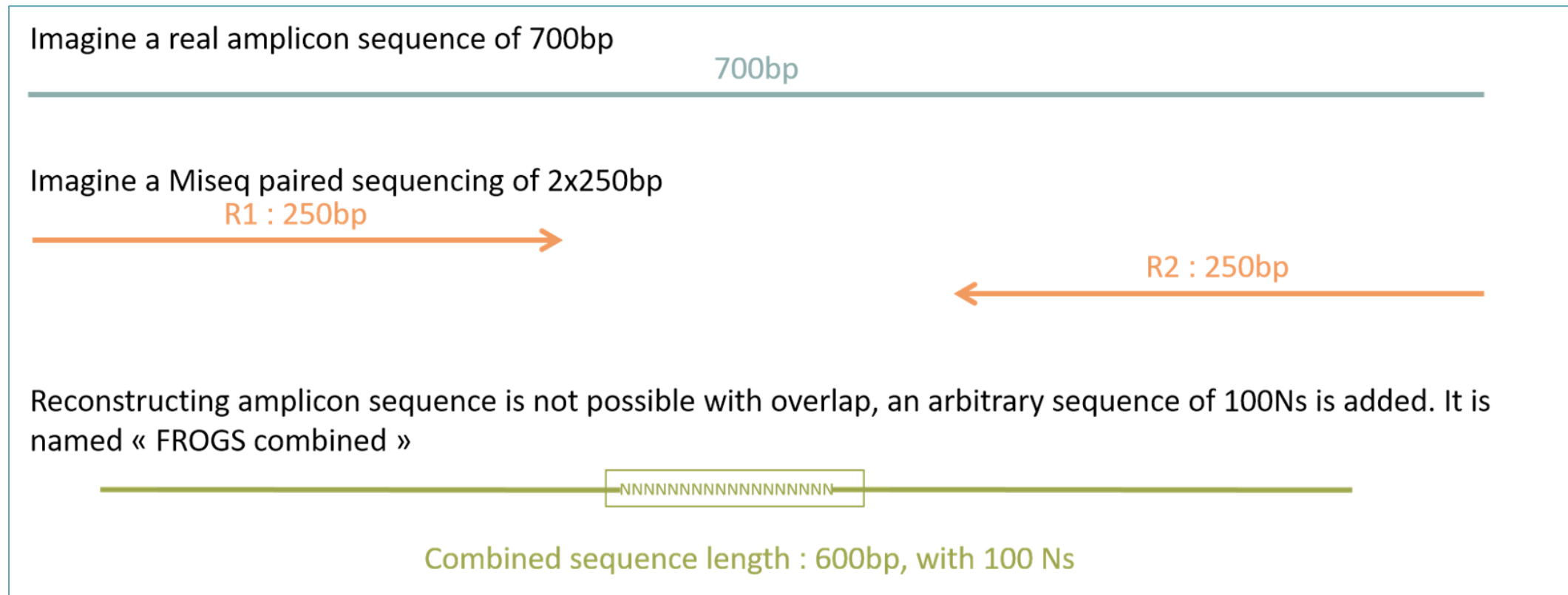
Problematic:
some ITS reads (Miseq sequencing) are non-overlapping
sequences



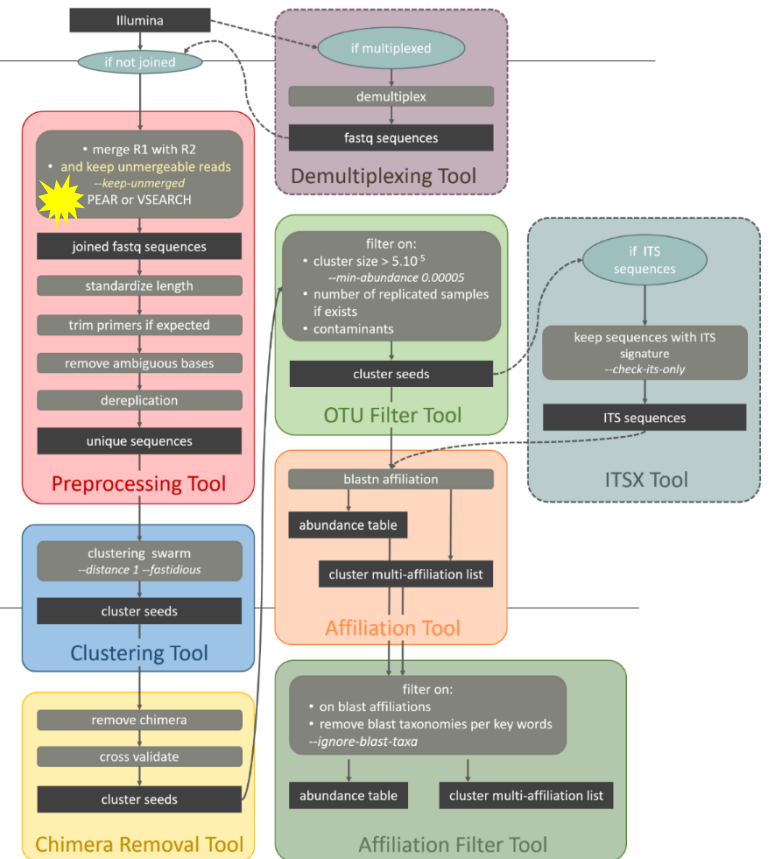
Consequence: during bioinformatics process, these reads are lost and underlying organisms will be never represented in the abundance table.

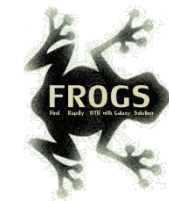


Solution: in preprocess step – creation of “FROGS combined” sequences



Pre-process tool



**FROGS Pre-process** merging, denoising and dereplication. (Galaxy Version r3.0-3.0)

Options

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

5: /work/formation/FROGS/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

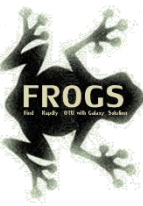
Select the software to merge paired-end reads.

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

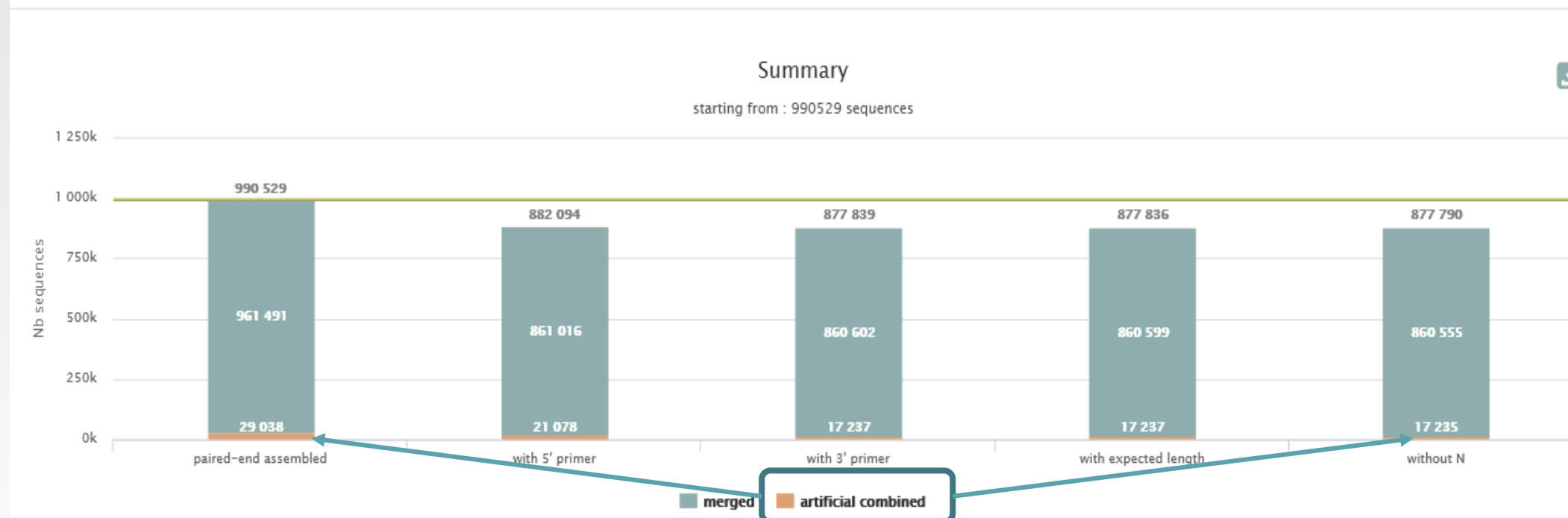
Go to « [ITS](#) » history

Launch the pre-process tool on this data set

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

Explore Preprocess report.html

Preprocess summary





2 tables:

Details on merged sequences

Show 10 entries Search: ↓ CSV

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

Own tag for combined sequences

```
>Cluster_20410 1:N:0:ATATAA
AAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGAT
>Cluster_2881 1:N:0:ATATAA
AAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGAT
>Cluster_10465 1:N:0:ATTACA
AAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGAT
>Cluster_2714_FROGS_combined R1_desc:1:N:0:ATATAA
AAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGAT
>Cluster_6993_FROGS_combined R1_desc:1:N:0:ATATAA
AAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGAT
>Cluster_2580_FROGS_combined R1_desc:1:N:0:ATATAA
AAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGAT
```

Details on artificial combined sequences

Show 10 entries Search: ↓ CSV

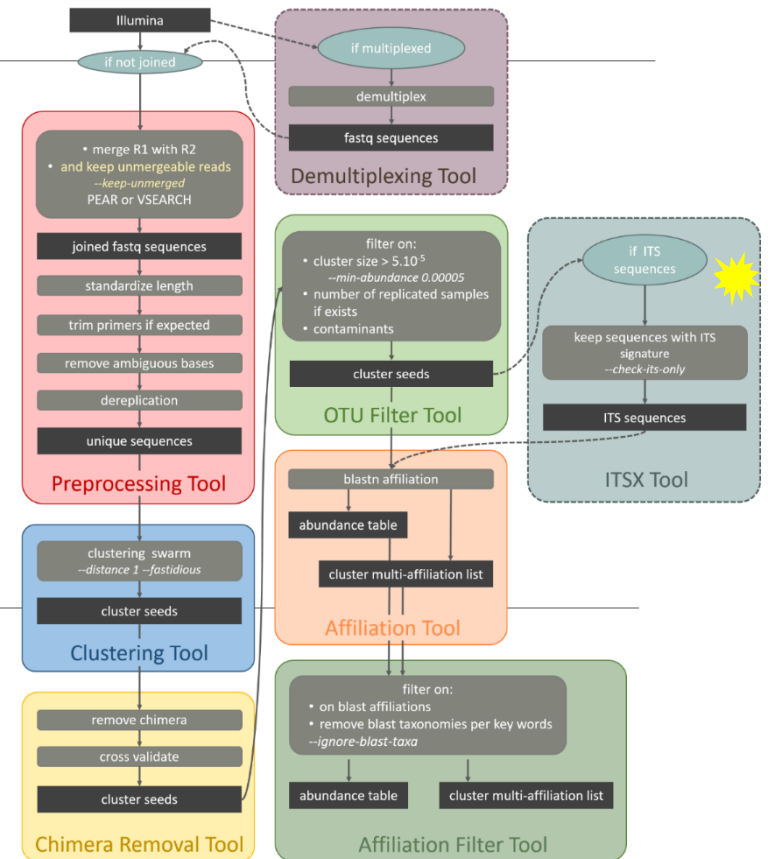
| Samples | % kept | paired-end assembled | with 5' primer | with 3' primer | with expected length | without N |
|----------------|--------|----------------------|----------------|----------------|----------------------|-----------|
| complexe-ADN-1 | 72.45 | 2,163 | 1,797 | 1,567 | 1,567 | 1,567 |
| echantillon1-1 | 57.31 | 1,047 | 745 | 600 | 600 | 600 |
| echantillon1-2 | 63.86 | 1,392 | 1,076 | 890 | 890 | 889 |
| echantillon1-3 | 50.58 | 2,491 | 1,601 | 1,260 | 1,260 | 1,260 |
| echantillon2-1 | 51.30 | 1,421 | 950 | 729 | 729 | 729 |

Filter only on minimum length for « combined ».

Minimum length = $R1 + 100N + R2 - \text{primers sizes}$

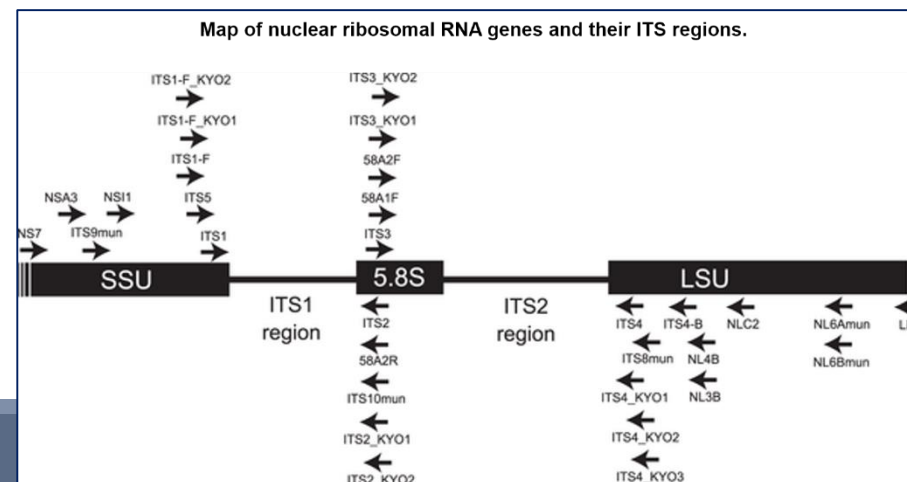
If the primers are very internal to the read, after trimming them, the combined sequence could be smaller than a read.

ITSx tools

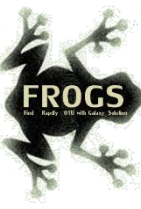


What is the purpose of the ITSx tool?

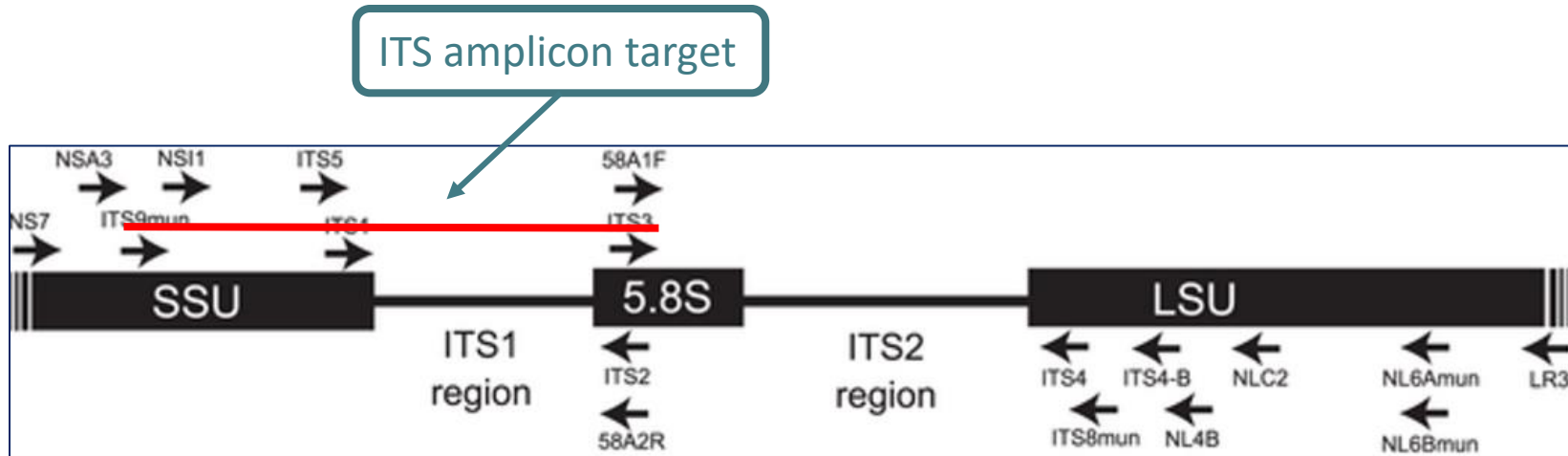
- ITSx is a tool to **filter** sequences.
- ITSx **identifies** and **trims** ITS regions in sequences.
- It **excludes** the highly conserved neighboring sequences **SSU**, **5S** and **LSU** rRNA.
- If the ITS1 or ITS2 region is not detected, the sequence is discarded.
- You can choose to check only 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).





Bengtsson-Palme, J., et al. (2013), Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. *Methods Ecol Evol*, 4: 914-919.
<https://doi.org/10.1111/2041-210X.12073>



What is the purpose of the ITSx tool?

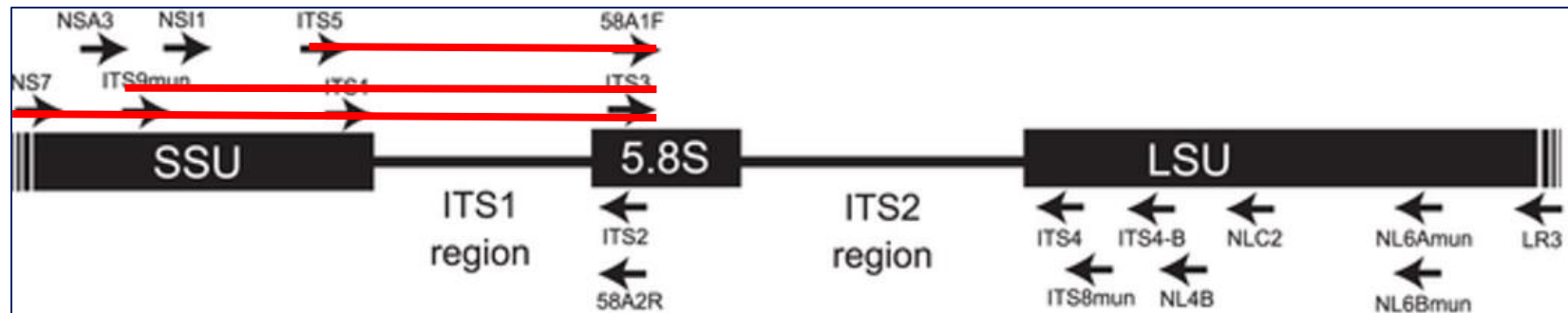


1st case: choose to trim
ITS1 is well detected
SSU part and 5.8S part are trimmed
Result: 

2nd case: choose to check only
ITS1 is well detected
SSU part and 5.8S part are not trimmed
Result: 

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

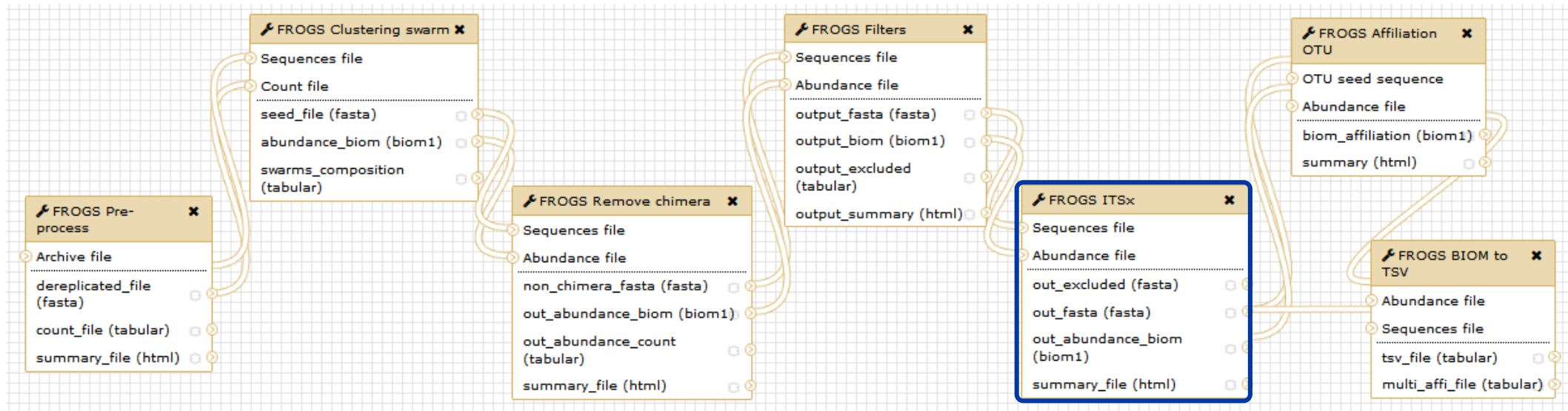
- If not, only ITS1 or ITS2 part will be conserved
- This is interesting to keep only the ITS parts without the flanking sequences in case of comparison of sequenced amplicons with different primers targeting the same region to be amplified.





When should we use ITSx ?

After filtering !



ITSx is a fastidious step



Sequences file

19: FROGS OTU Filters: sequences.fasta

The sequence file to filter (format: fasta).

Abundance file

20: FROGS OTU Filters: abundance.biom

The abundance file to filter (format: BIOM).

ITS region

ITS1

Which fungal ITS region is targeted: either ITS1 or ITS2

Trim conserved sequence (SSU, 5.8S, LSU) ?

Yes No

If Yes, only part of the sequences with ITS signature will be kept, SSU, LSU or 5.8S regions will be trimmed. (default : No)

By default, only part of the sequence with ITS signature is kept (SSU, LSU or 5.8S are trimmed)

Choose pertinent organisms to scan?

Select/Unselect all

- Fungi
- Alveolata
- Bryophyta
- Bacillariophyta
- Amoebozoa
- Euglenozoa
- Chlorophyta
- Rhodophyta
- Phaeophyceae
- Marchantiophyta
- Metazoa
- Oomycota
- Haptophyceae
- Raphidophyceae
- Rhizaria
- Synurophyceae
- Tracheophyta
- Eustigmatophyceae

By default, sequences are considered as FUNGI sequences. Change it, if it is not the case.

Save a lot of time by checking pertinent organism group model to scan.

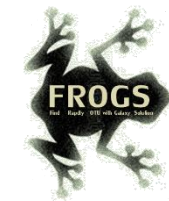
Execute

Careful !

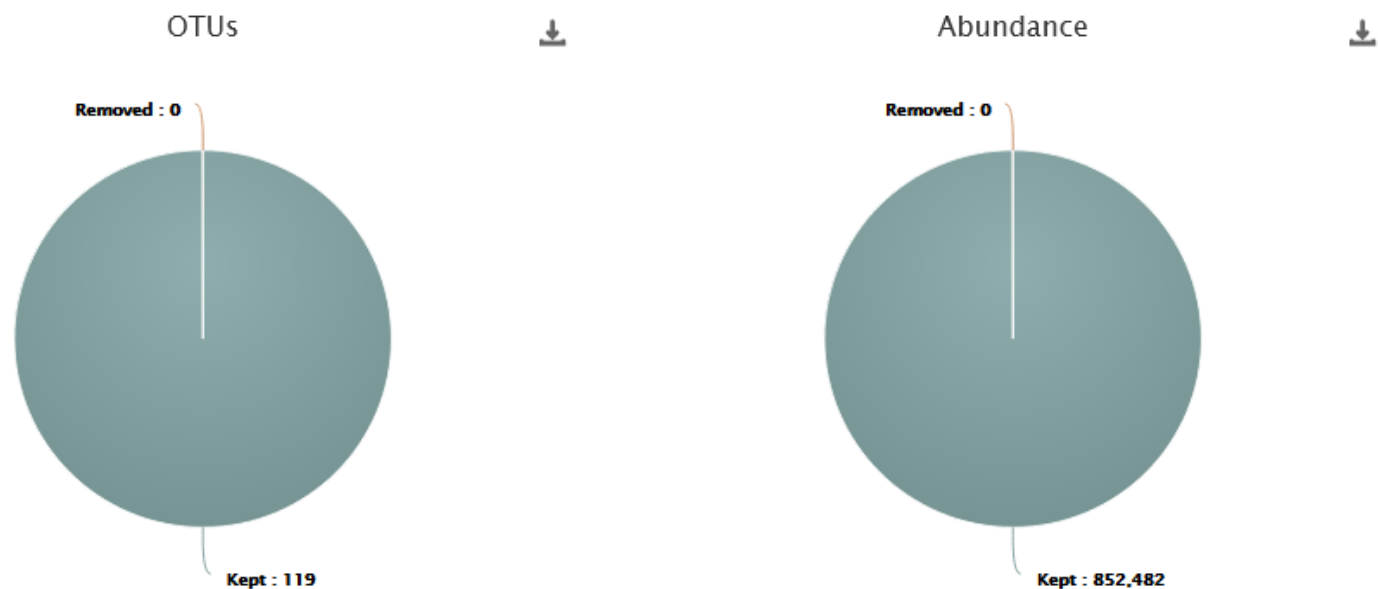


- The ITSx step is time consuming and has to be done on clusters. We advise our users to apply ITSx in 5th step:
 1. Preprocess step,
 2. Clustering step,
 3. Chimera removing step,
 4. Filter on OTUs abundances and replicats step,
 5. ITSx if Fungi ITS amplicons.

- 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 [Download CSV](#)

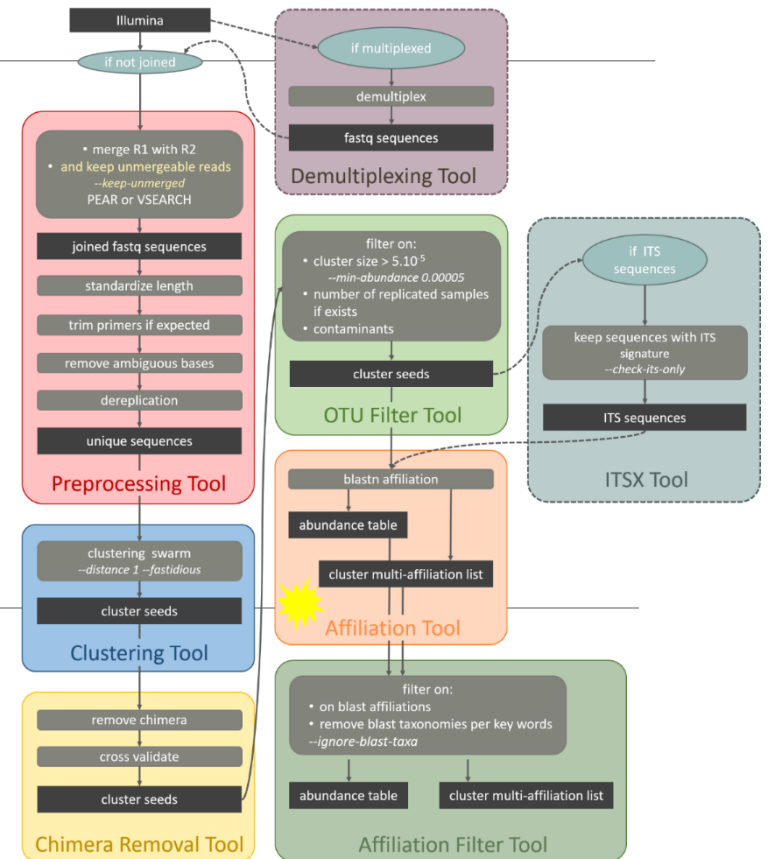
Search:

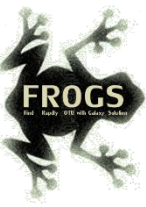
OTUs removed by sample

| Sample name | Initial | Kept | Initial abundance | Kept abundance |
|----------------|---------|------|-------------------|----------------|
| complexe-ADN-1 | 92 | 92 | 47,268 | 47,268 |
| echantillon1-1 | 71 | 71 | 26,783 | 26,783 |
| echantillon1-2 | 72 | 72 | 51,465 | 51,465 |



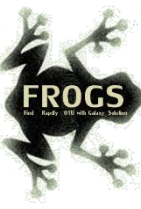
ITS Affiliation





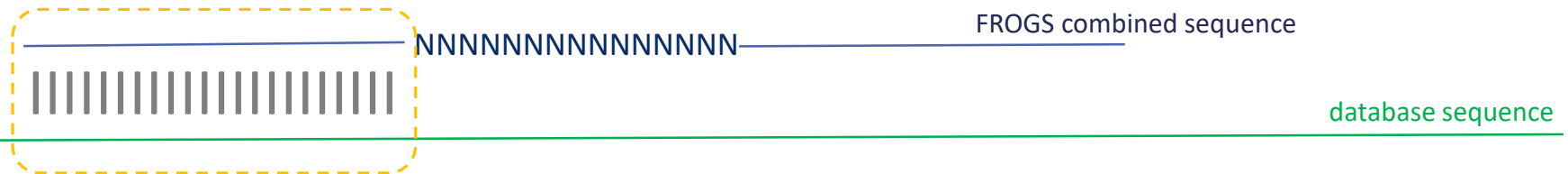
What is special about the affiliation of ITS/FROGS combined sequences?

- 2 alignment tools - blastn+ or needleall - are used to find alignments between each OTU and the database.
- Only the best hits with the same score are reported.
- blastn+ is used for classical **merged read pair**, and blastn+ then needleall are used for **artificially combined sequence**.
- For each alignment, several metrics are computed: %identity, %coverage and alignment length.



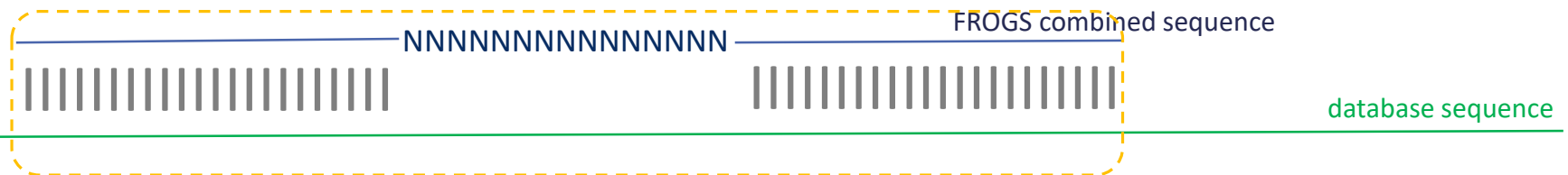
What is special about the affiliation of ITS/FROGS combined sequences?

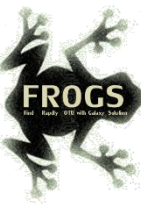
- blastn+ *i.e.* a local aligner, is not usable for "**combined**" sequences



Between combined and the database sequence, alignment is perfect until N stretch with blastn+. Information about the 2nd part of sequence are not explored !

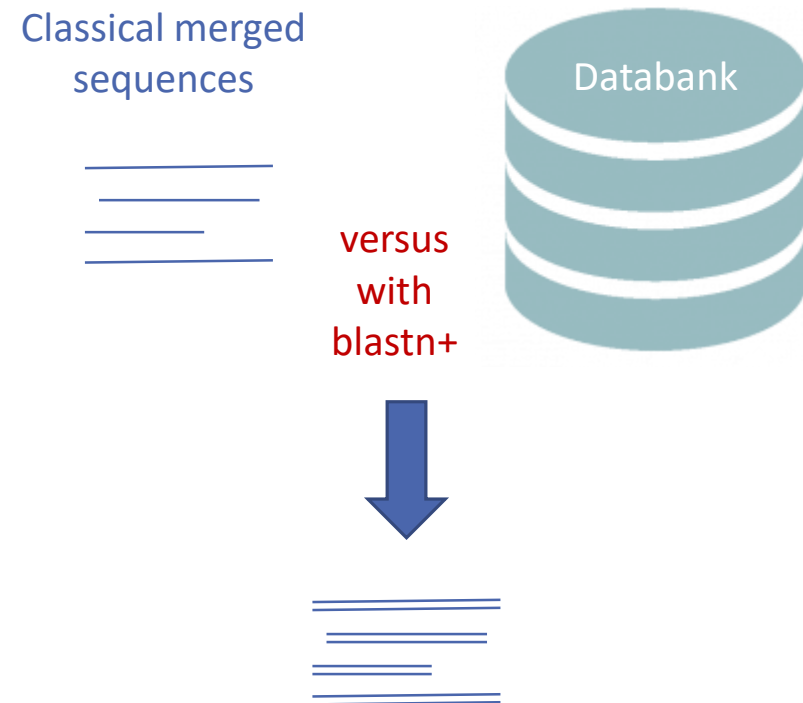
- It is necessary to use a global aligner *i.e.* Needleall (the sequence must be aligned in its entirety), but it is computationally too hard.





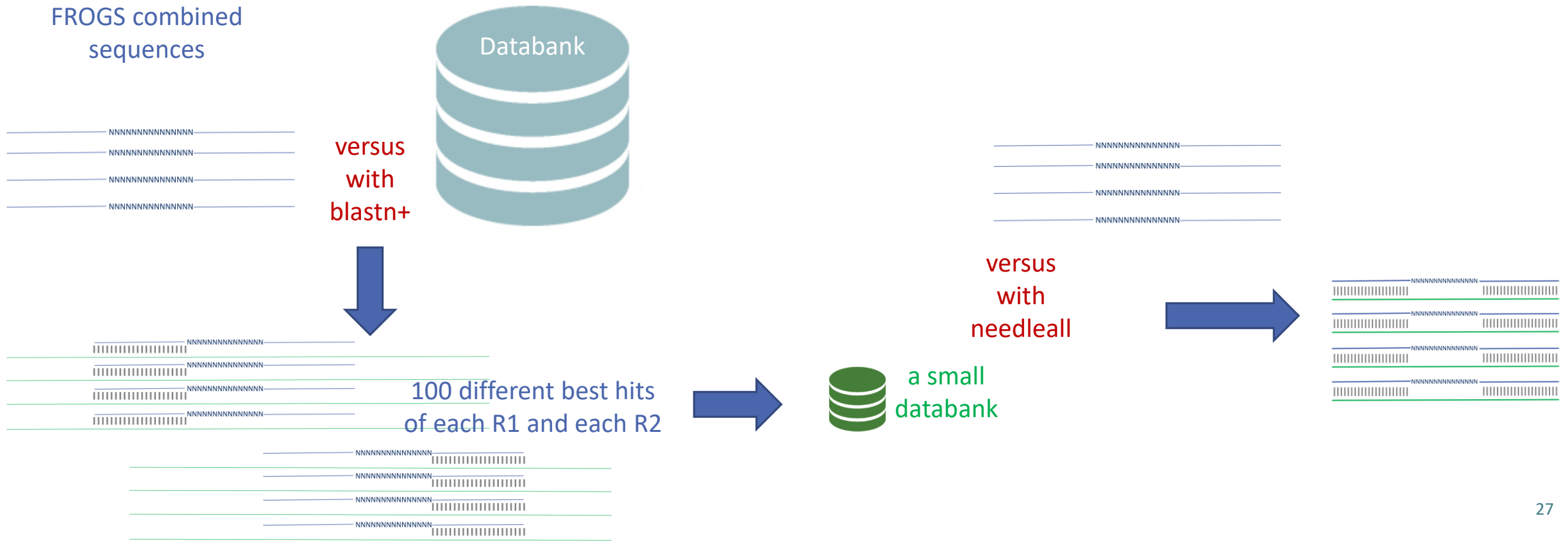
What is special about the affiliation of ITS/FROGS combined sequences?

- Solution: 1st step treat classical merged sequences with blastn+



What is special about the affiliation of ITS/FROGS combined sequences?

- 2nd step for FROGS combined sequences: use blastn+ to create a small databank and align with needleall this small databank versus FROGS combined sequences





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)

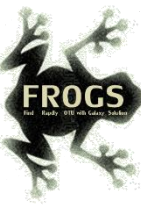


Affiliation is notably made by a local alignment with NCBI Blast+



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 needleall

NNNNNNNNNNNNNNNNNN

Imagine a perfect sequencing without error:

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



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

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

So, we replaced the classical %id by a %id computed on the sequenced bases only.

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

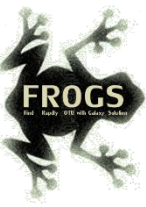
Case 1 : 16S V3V4 → overlapped sequence

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

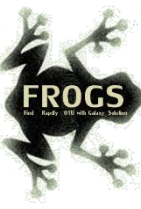
Case 2 : very large ITS1 → “FROGS combined” shorter than the real sequence

% 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



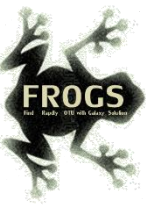
What is the purpose of the *Affiliation post-process* tool ?

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.



FROGS Affiliation postprocess Optionnal step to resolve inclusive amplicon ambiguities and to aggregate OTUs based on alignment metrics (Galaxy Version 3.2.2) Options

Sequences file
23: FROGS ITSx: itsx.fasta
The sequence file to filter (format: fasta).

Abundance file
24: FROGS ITSx: itsx.biom
The abundance file to filter (format: BIOM).

Is this an amplicon hyper variable in length?
 Yes No
Multi-affiliation tag may be resolved by selecting the shortest amplicon reference. For this you need the reference fasta file of your targetted amplicon.

Using reference database
UNITE_8.2 ITS1
Select reference from the list

minimum identity for aggregation
100
OTUs will be aggregated if they share the same taxonomy with at least X% identity.

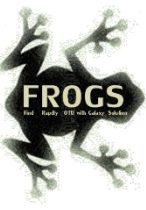
minimum coverage for aggregation
100
OTUs will be aggregated if they share the same taxonomy with at least X% alignment coverage.

Execute

Yes, we have combined sequences

same database used for affiliation

Here, we wanted to aggregate OTUs only if they are identical



Exemple

After Preprocessing + Clustering + OTU Filter + **ISTX** :

>Cluster_3

GTGATTGCCTTTATAGGCTTATAACTATATCCACTTACACCTGTGAACTGTTCTACTACTTGACGCAAGTCGAGTATTTTTACAAACAATGTGTAATGAACGTCGTTTTATTATAACAAAATA

>Cluster_54

GTGATTGCCTTTATAGGCTTATAACTATATCCACTTACACCTGTGAACTGTTCTACTACTTGACGCAAGTCGAGTATTTTTACAAACAATGTGTAATGAACGTCGTTTTATTATAACAAAATA

>Cluster_414_FROGS_combined

GTGATTGCCTTTATAGGCTTATAACTATATCCACTTACACCTGTGAACTGTTCTACTACTTGACGCAAGTCGAGTATTTTTACAAACAATGTGTAATGAACGTCGTTTTATTATAACAAAATA

These 3 sequences have become **strictly identical** !²



Example

After Preprocessing + Clustering + OTU Filter + ISTX + **Affiliation Post-process** :

Cluster_3, Cluster_54 and Cluster_414_FROGS_combined are **aggregated** in a same OTU

FROGS Affiliation postprocess: aggregation_composition.txt

Cluster_1 Cluster_244 Cluster_448_FROGS_combined Cluster_471_FROGS_combined

Cluster_2 Cluster_320 Cluster_357 Cluster_435 Cluster_468 Cluster_312 Cluster_364 Cluster_477 Cluster_466 Cluster_480

Cluster_3 Cluster_54 Cluster_414_FROGS_combined

Cluster_4 Cluster_15 Cluster_27 Cluster_42 Cluster_67 Cluster_77 Cluster_137 Cluster_209 Cluster_422

Cluster_5 Cluster_5171

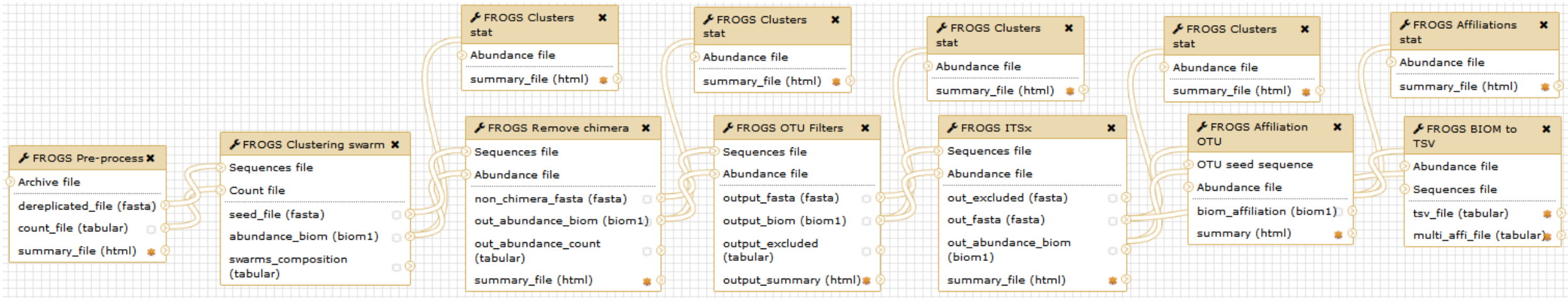
Cluster_6 Cluster_53

Cluster_9 Cluster_71

Cluster_7

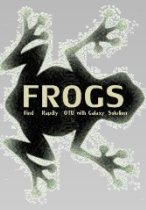


Workflow creation



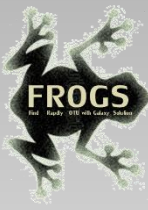
Workflows are useful for routine analyses

A workflow links FROGS steps together and when it is launched, all the steps run automatically.

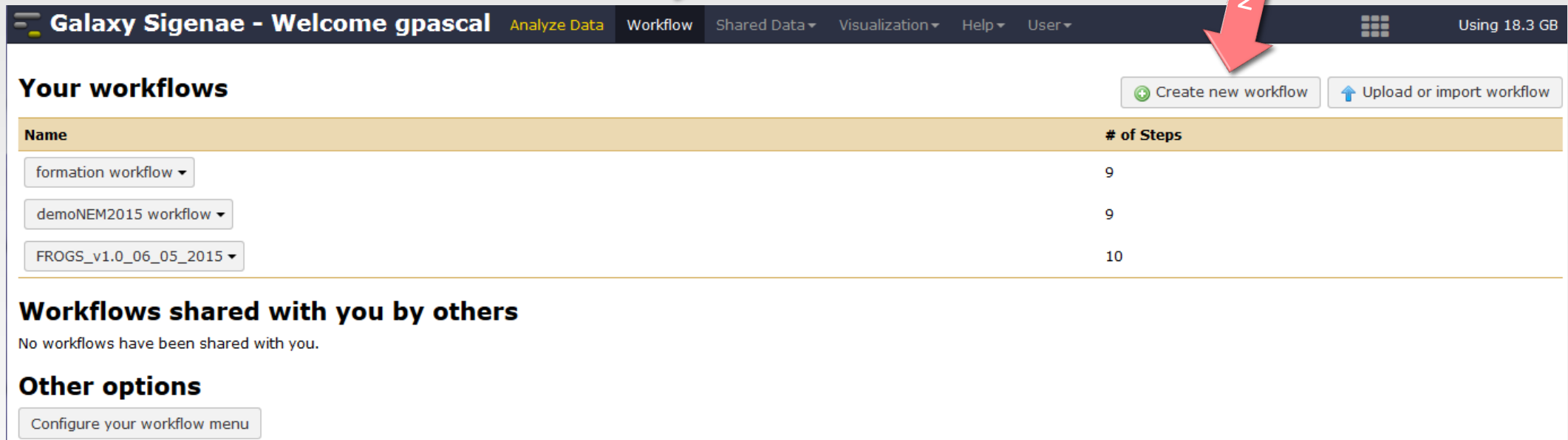


Practice

CREATE YOUR OWN WORKFLOW !



Exercise



The screenshot shows the Galaxy Sigenae web interface. A dark navigation bar at the top contains the text 'Galaxy Sigenae - Welcome gpascal' followed by menu items: 'Analyze Data', 'Workflow', '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'. A red arrow labeled '1' points to the 'Workflow' menu item. Below the navigation bar, the main content area is titled 'Your workflows'. To the right of this title are two buttons: 'Create new workflow' (with a plus icon) and 'Upload or import workflow' (with an upload icon). A red arrow labeled '2' points to the 'Create new workflow' button. 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 another 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 |

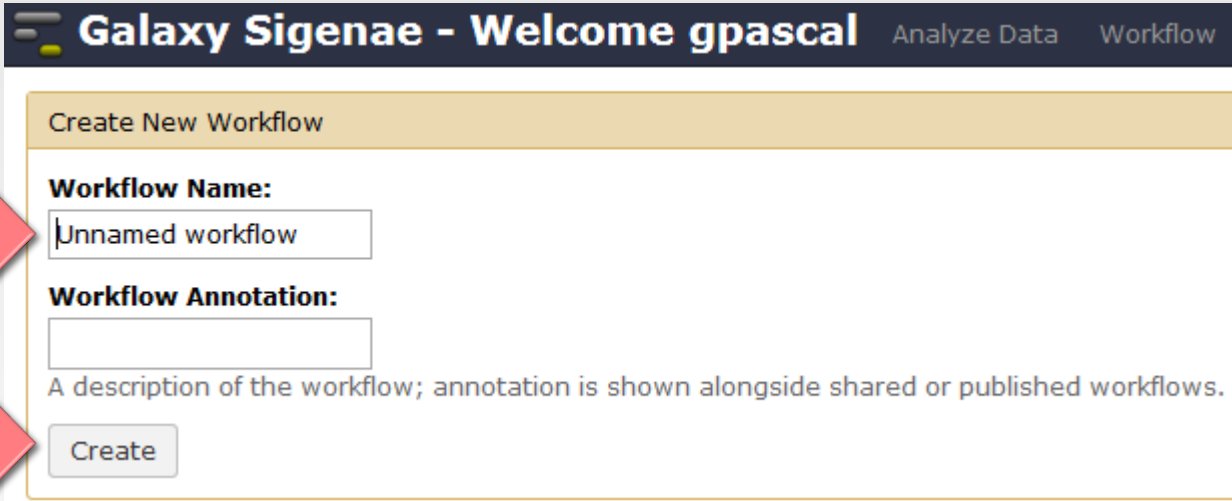
Workflows shared with you by others

No workflows have been shared with you.

Other options

[Configure your workflow menu](#)

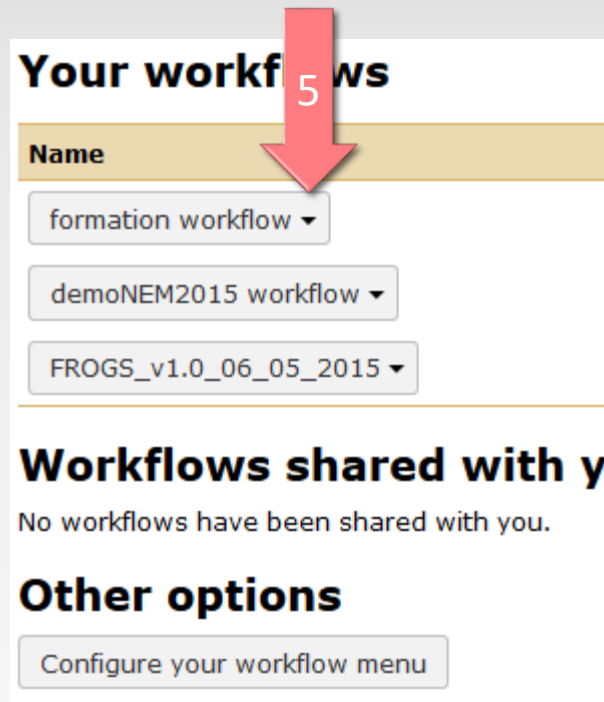
Exercise



The screenshot shows the 'Create New Workflow' interface in Galaxy Sigeneae. The header bar includes the text 'Galaxy Sigeneae - Welcome gpascal' and navigation links for 'Analyze Data' and 'Workflow'. The form contains the following elements:

- Create New Workflow** (Section Header)
- Workflow Name:** A text input field containing 'Unnamed workflow'. A red arrow with the number '3' points to this field.
- Workflow Annotation:** An empty text input field.
- A descriptive text: 'A description of the workflow; annotation is shown alongside shared or published workflows.'
- Create** (Submit Button). A red arrow with the number '4' points to this button.

Exercise

This screenshot shows the 'Your workflows' section of a web interface. A red arrow with the number '5' points to the first workflow name, 'formation workflow', in a list. Below the list are sections for 'Workflows shared with you' (with the text 'No workflows have been shared with you.') and 'Other options' (with a button 'Configure your workflow menu').

Your workflows

Name

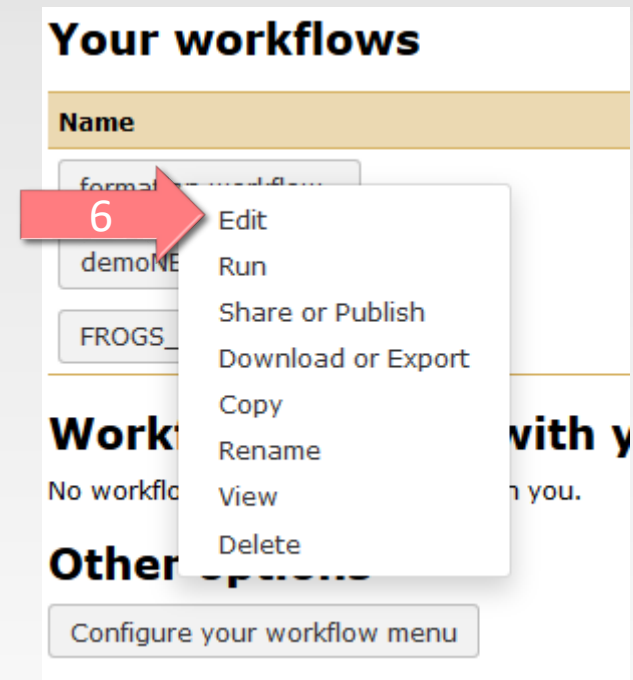
- formation workflow ▾
- demoNEM2015 workflow ▾
- FROGS_v1.0_06_05_2015 ▾

Workflows shared with you

No workflows have been shared with you.

Other options

[Configure your workflow menu](#)

This screenshot shows the 'Your workflows' section with a context menu open over the first workflow, 'formation workflow'. A red arrow with the number '6' points to the menu. The menu options are: Edit, Run, Share or Publish, Download or Export, Copy, Rename, View, and Delete. The rest of the page content is partially visible behind the menu.

Your workflows

Name

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

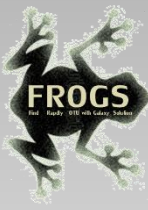
Workflows shared with you

No workflows have been shared with you.

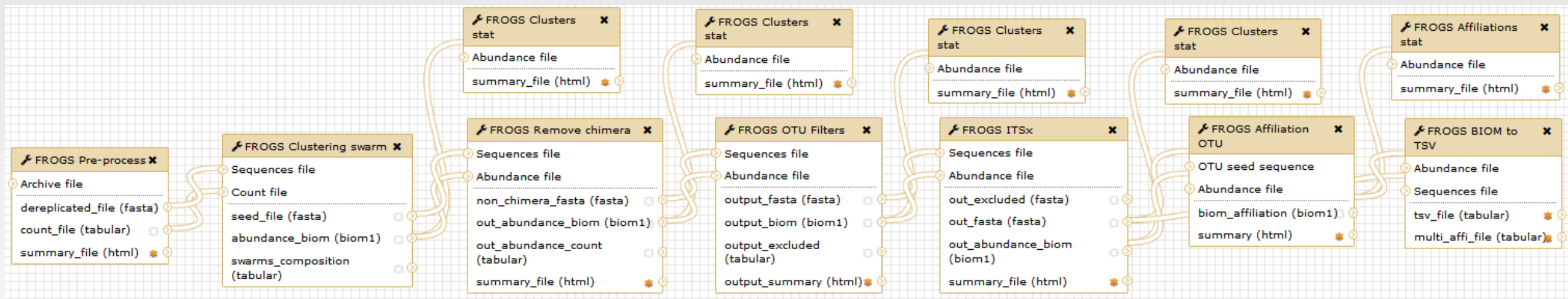
Other options

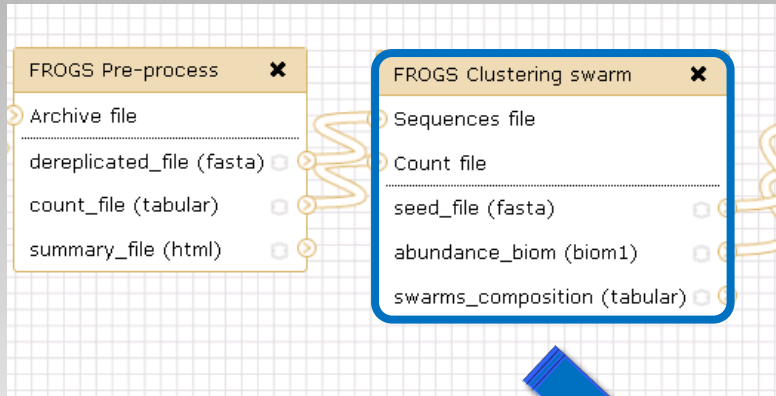
[Configure your workflow menu](#)

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



Solution of exercise:





For each tool, think to:
1. Fixe parameter ?

?

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

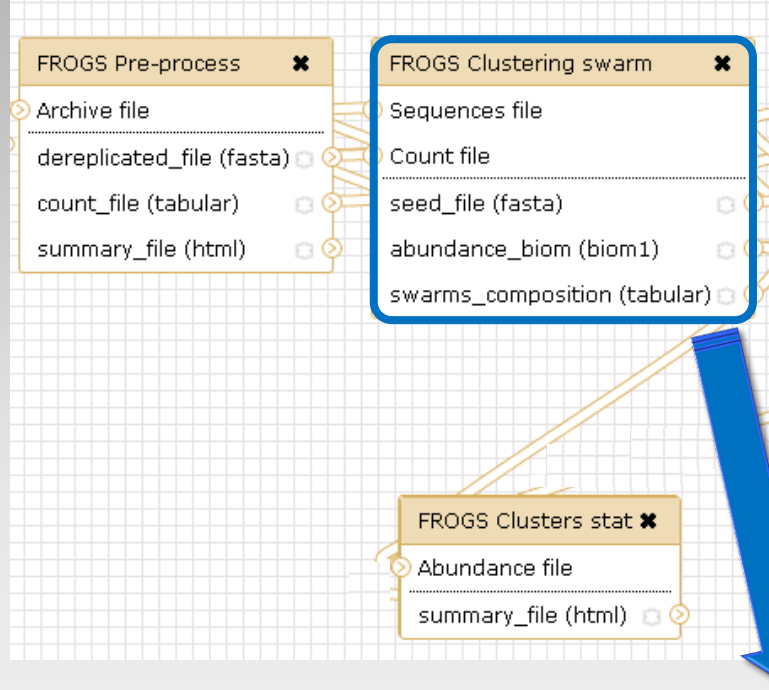
Sequences file
Data input 'sequence_file' (fasta)
The sequences file (format: fasta).

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



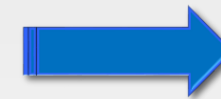
For each tool, think to:

1. Fixe parameter ?
2. 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

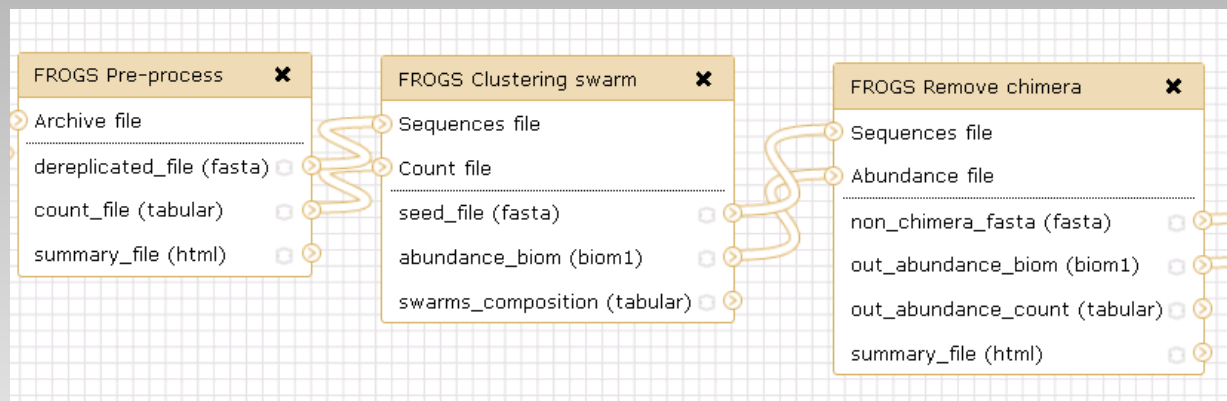
This action will rename the output dataset. Click [here](#) for more information. Valid inputs are: **sequence_file**, **count_file**.

Change datatype

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

Tags

This action will set tags for the dataset.



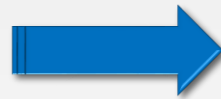
For each tool, think to:

1. Fixe parameter ?
2. Rename output files
3. Hide intermediate files

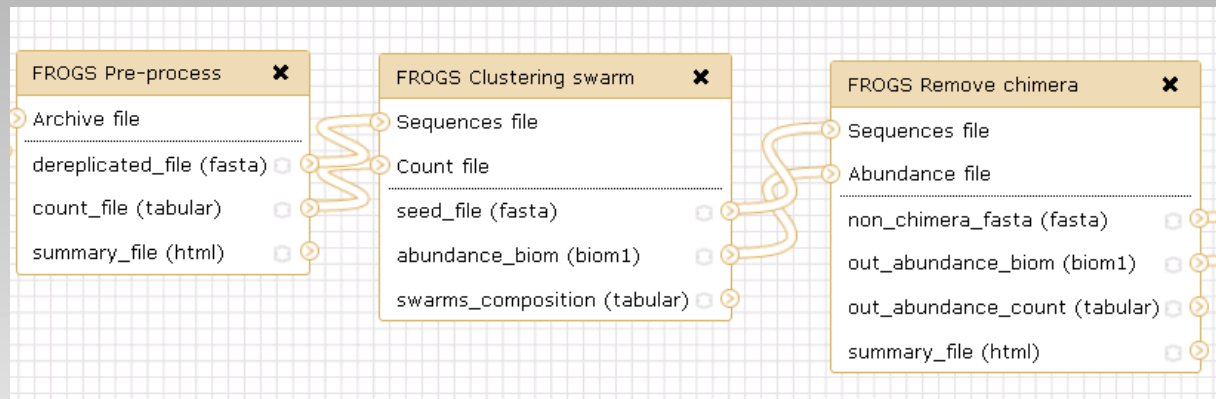
FROGS Remove chimera

- Sequences file
- Abundance file

- non_chimera_fasta (fasta)
- out_abundance_biom (biom1)
- out_abundance_count (tabular)
- summary_file (html)

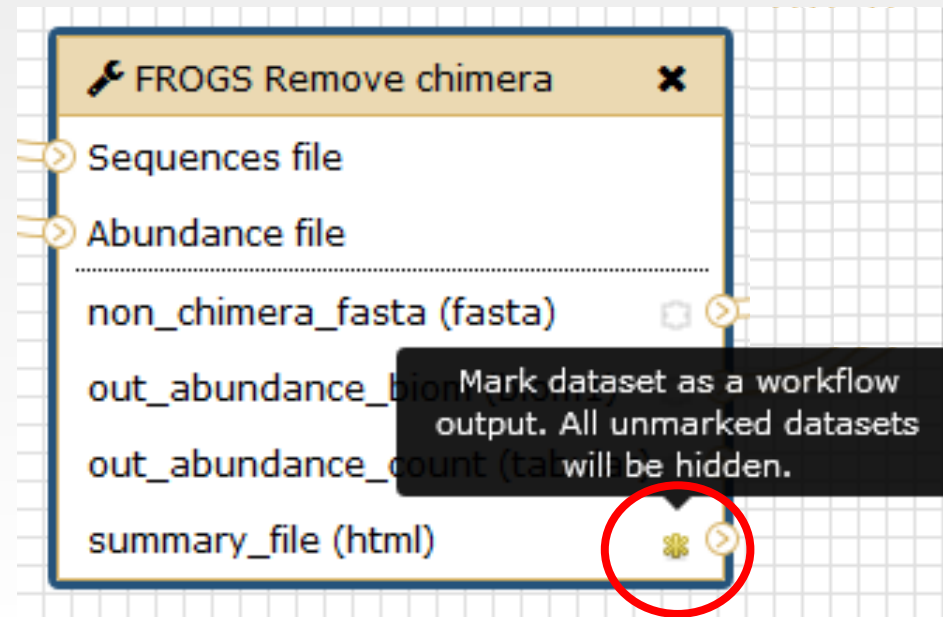


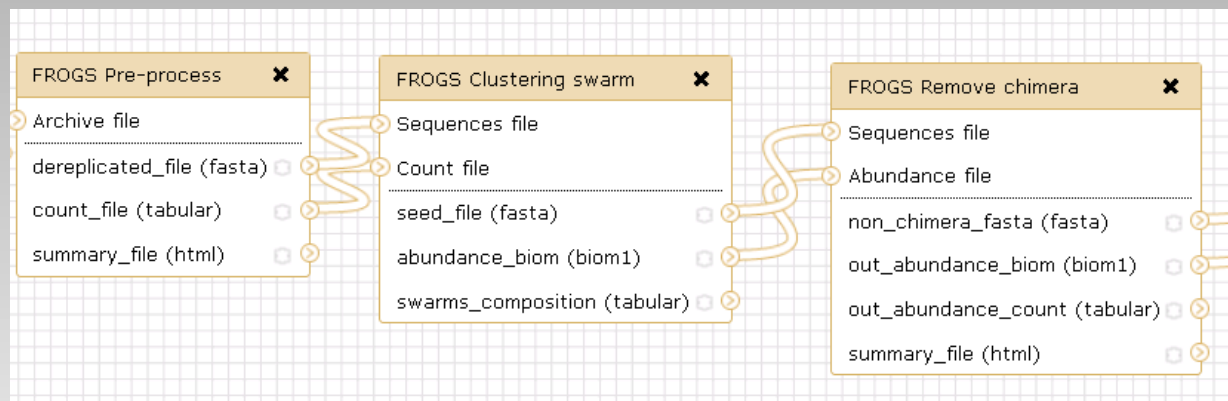
- 11: FROGS Remove chimera: report.html**
- 10: FROGS Remove chimera: non chimera abundance.biom**
- 9: FROGS Remove chimera: non chimera.fasta**



For each tool, think to:

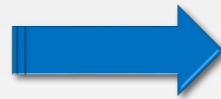
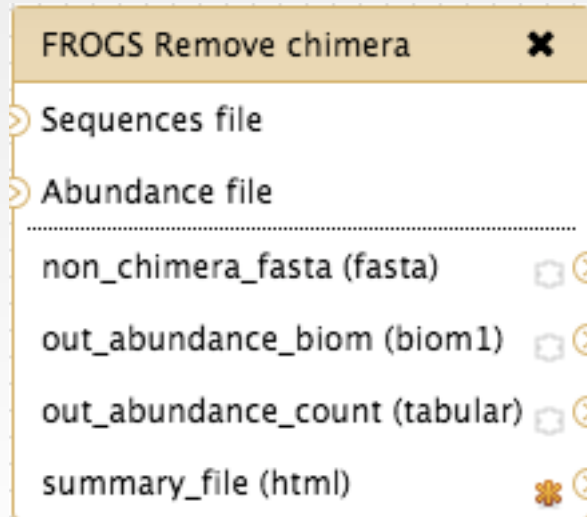
1. Fixe parameter ?
2. Rename output files
3. Hide intermediate files








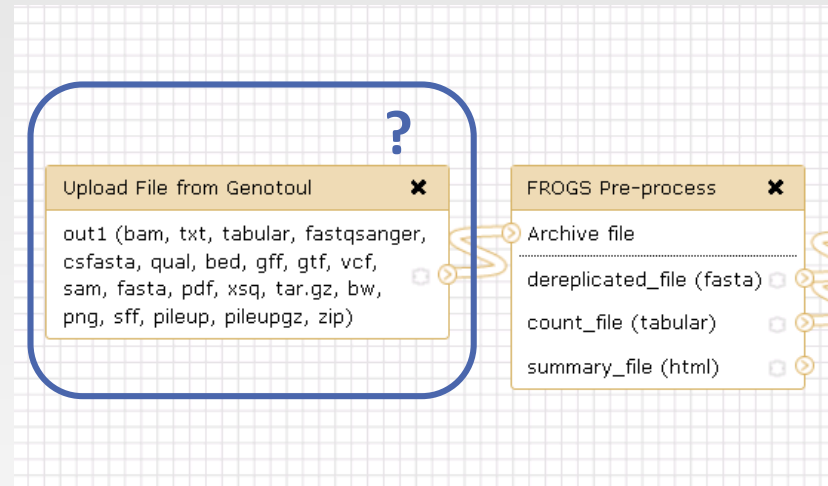
For each tool, think to:

1. Fixe parameter ?
2. Rename output files
3. Hide intermediate files



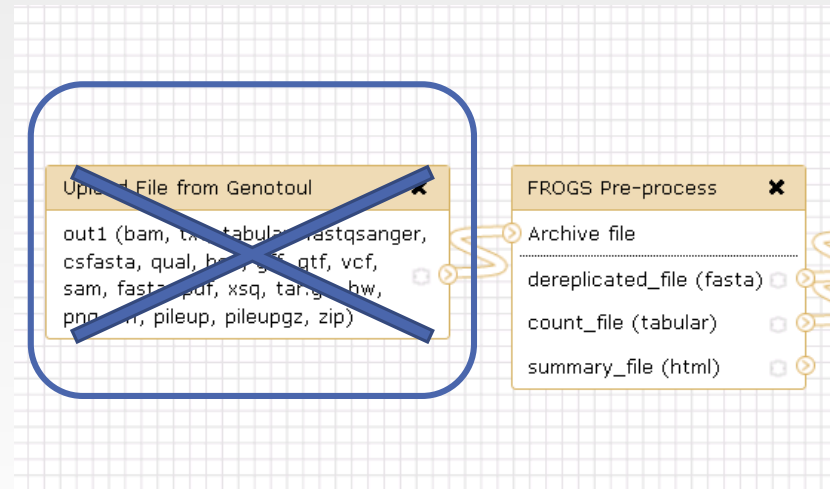
11: FROGS Remove chimera: report.html   

Could you integrate « upload file » in the workflow ?

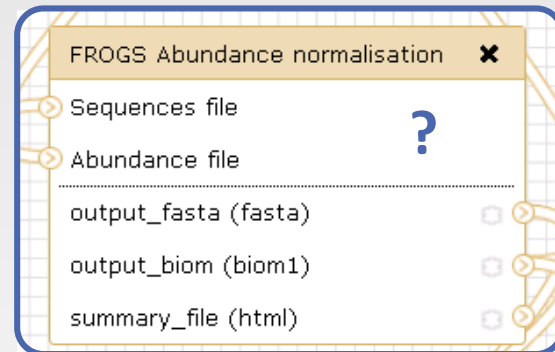


Could you integrate « upload file » in the workflow ?

Upload file cannot be automatized because the workflow, at each run, will be processed with different input data

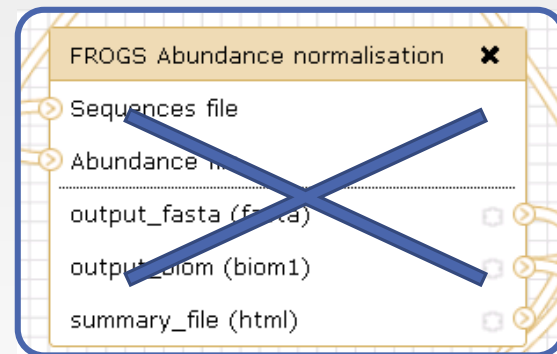


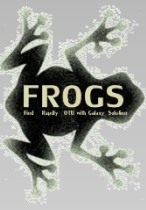
Could you integrate « Normalisation tool » in the workflow ?



Could you integrate « Normalisation tool » in the workflow ?

You need to know by which number you will normalize data and this maximal number is known during the process, you need to enter in a clusterStat_report.html after OTU filter step.





Exercise

When your workflow is built

1. Run your own workflow with ITS data with :

http://genoweb.toulouse.inra.fr/~formation/15_FROGS/Webinar_data/ITS1.tar.gz

2. Import metadata for statistics analyses

http://genoweb.toulouse.inra.fr/~formation/15_FROGS/Webinar_data/metadata ITS.tsv

3. Run FROGS_stat tools