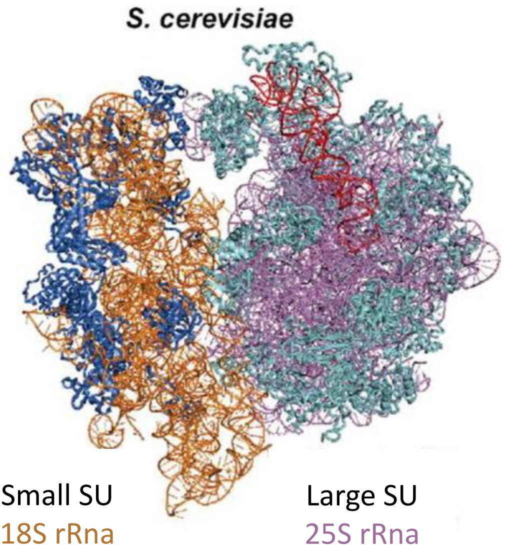
D- Training on Galaxy: Metabarcoding

December 2021 - Webinar

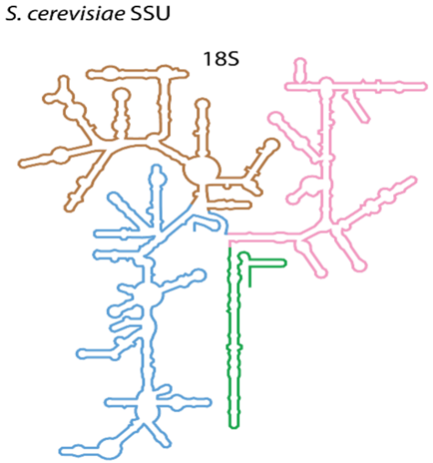
FROGS Practice on ITS data

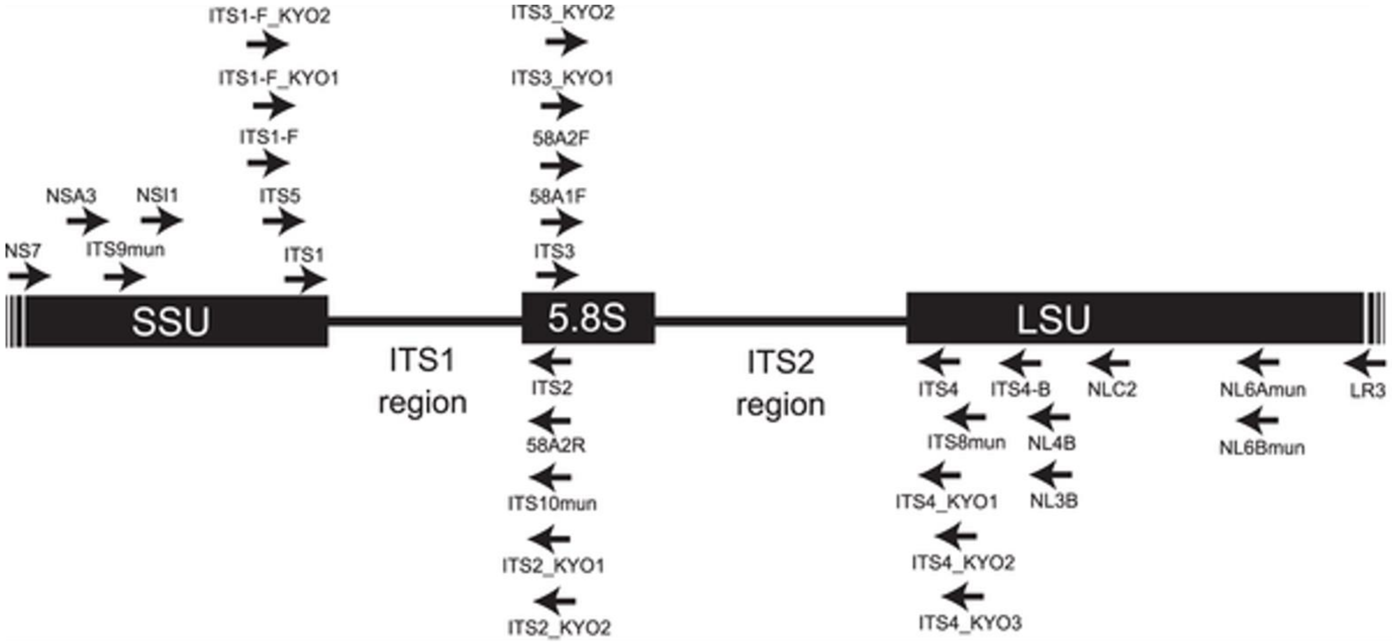
LUCAS AUER, MARIA BERNARD, LAURENT CAUQUIL, MAHENDRA MARIADASSOU, GÉRALDINE PASCAL & OLIVIER RUÉ

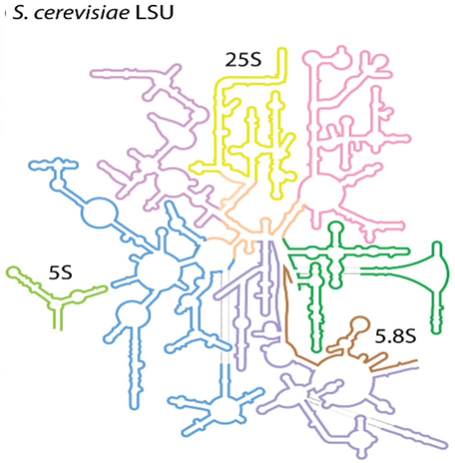
What is a ITS ?



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

What is a ITS ?

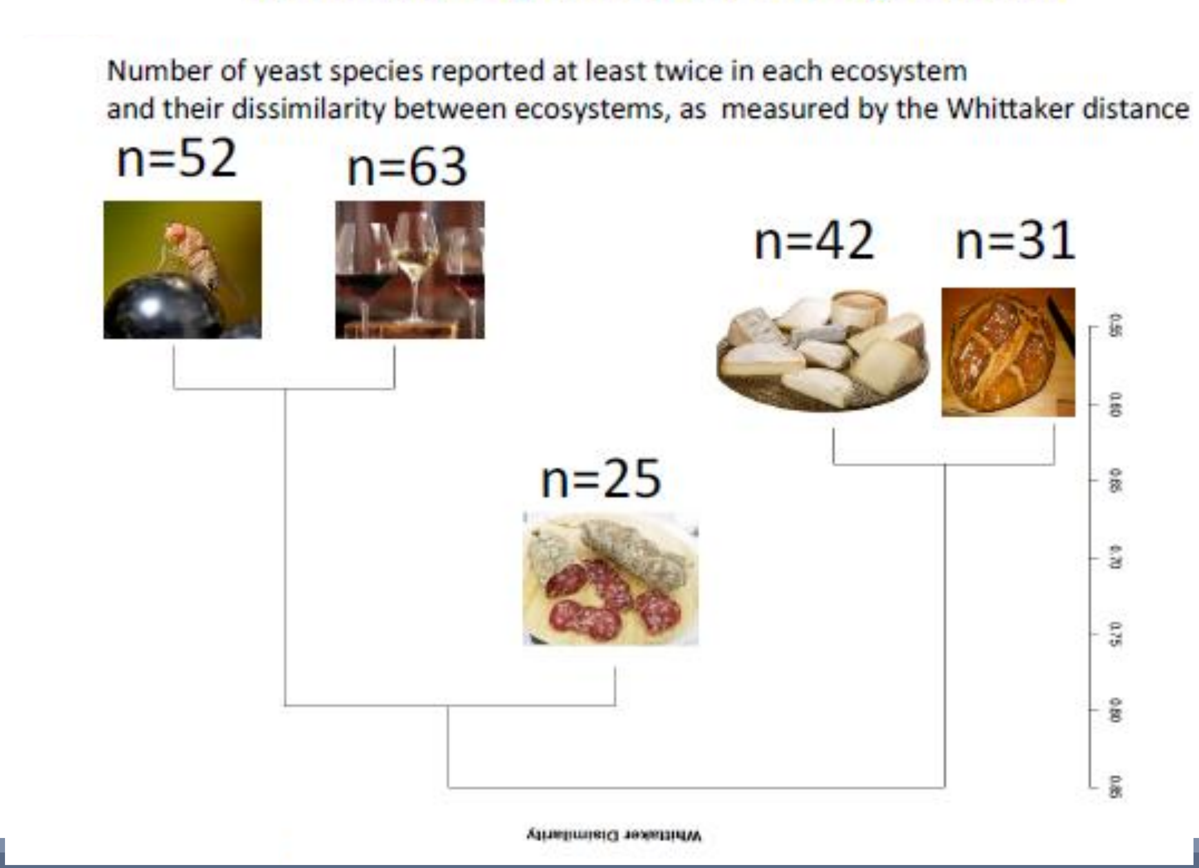
- Size polymorphism of ITS (from 361 to 1475 bases in UNITE 7.1)
- Highly conserved regions of the neighboring of ITS1 and ITS2
- Lack of a generalist and abundant ITS databank (several small specialized databanks)
- Multiple copies[*] (14 to 1400 copies (mean at 113, median et 80))
- Do not target Glomeromycetes/Glomeromycota (→ alternative: 18S)

If your sequencing platform preprocesses your data, it has to keep short and long sequences

[*] https://doi.org/10.1111/mec.14995

ITS data form METABARFOOD project metaprogramme MEM

Yeast catalog in food ecosystem



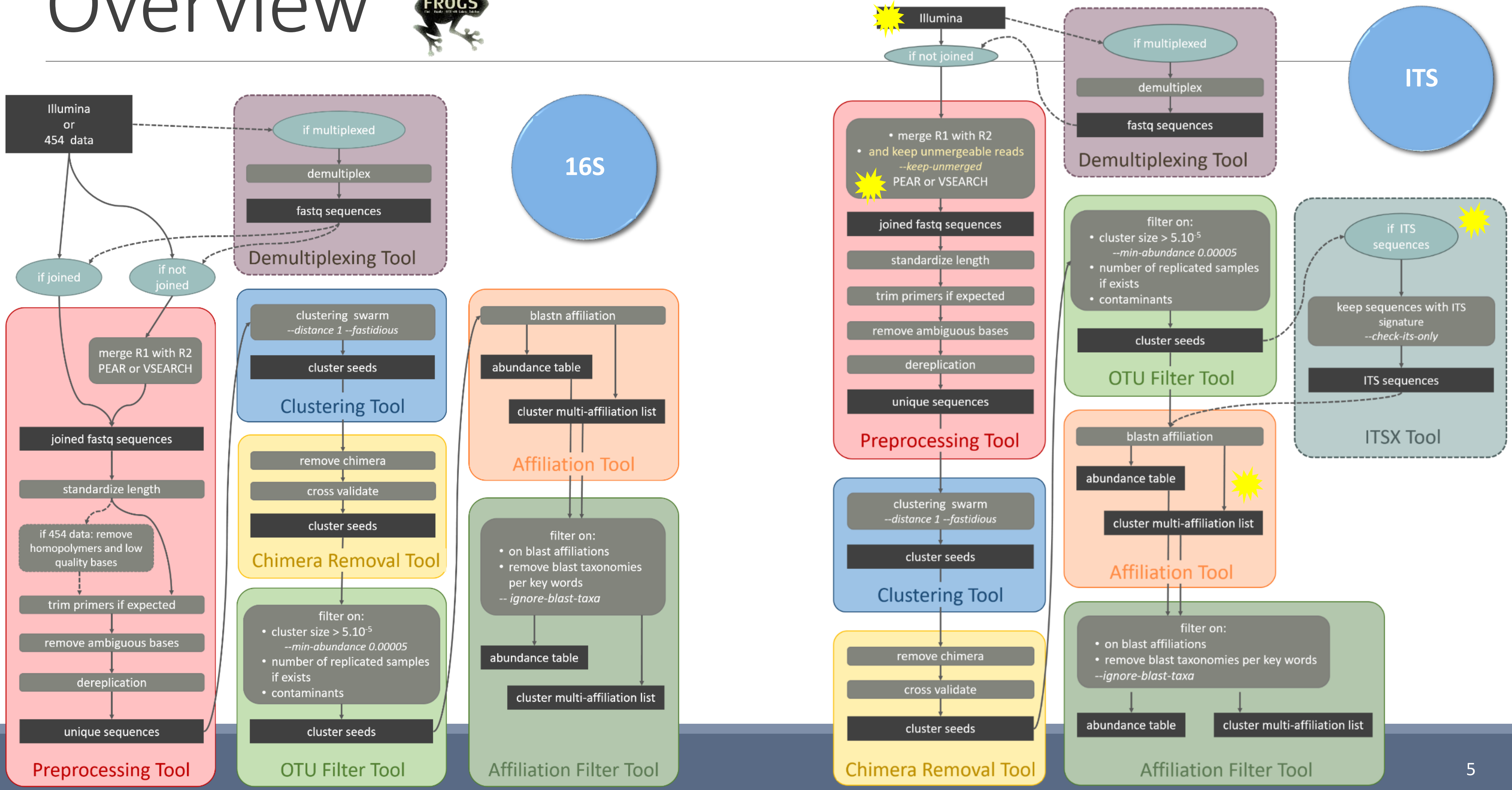
Number of yeast species reported at least twice in each ecosystem
and their dissimilarity between ecosystems, as measured by the Whittaker distance

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.

: Differences between 16S and ITS pipelines

Overview



16S

Illumina or 454 data

Demultiplexing Tool
- if multiplexed
- demultiplex
- fastq sequences

if joined

if not joined

Preprocessing Tool
- merge R1 with R2 PEAR or VSEARCH
- joined fastq sequences
- standardize length
- if 454 data: remove homopolymers and low quality bases
- trim primers if expected
- remove ambiguous bases
- dereplication
- unique sequences

Clustering Tool
- clustering swarm *--distance 1 --fastidious*
- cluster seeds

Chimera Removal Tool
- remove chimera
- cross validate
- cluster seeds

OTU Filter Tool
- filter on:
  - cluster size > 5.10^{-5} *--min-abundance 0.00005*
  - number of replicated samples if exists
  - contaminants
- cluster seeds

Affiliation Tool
- blastn affiliation
- abundance table
- cluster multi-affiliation list

Affiliation Filter Tool
- filter on:
  - on blast affiliations
  - remove blast taxonomies per key words *-- ignore-blast-taxa*
- abundance table
- cluster multi-affiliation list

ITS

Illumina

if not joined

Demultiplexing Tool
- if multiplexed
- demultiplex
- fastq sequences

Preprocessing Tool
- merge R1 with R2
- and keep unmergeable reads *--keep-unmerged* PEAR or VSEARCH
- joined fastq sequences
- standardize length
- trim primers if expected
- remove ambiguous bases
- dereplication
- unique sequences

Clustering Tool
- clustering swarm *--distance 1 --fastidious*
- cluster seeds

Chimera Removal Tool
- remove chimera
- cross validate
- cluster seeds

OTU Filter Tool
- filter on:
  - cluster size > 5.10^{-5} *--min-abundance 0.00005*
  - number of replicated samples if exists
  - contaminants
- cluster seeds

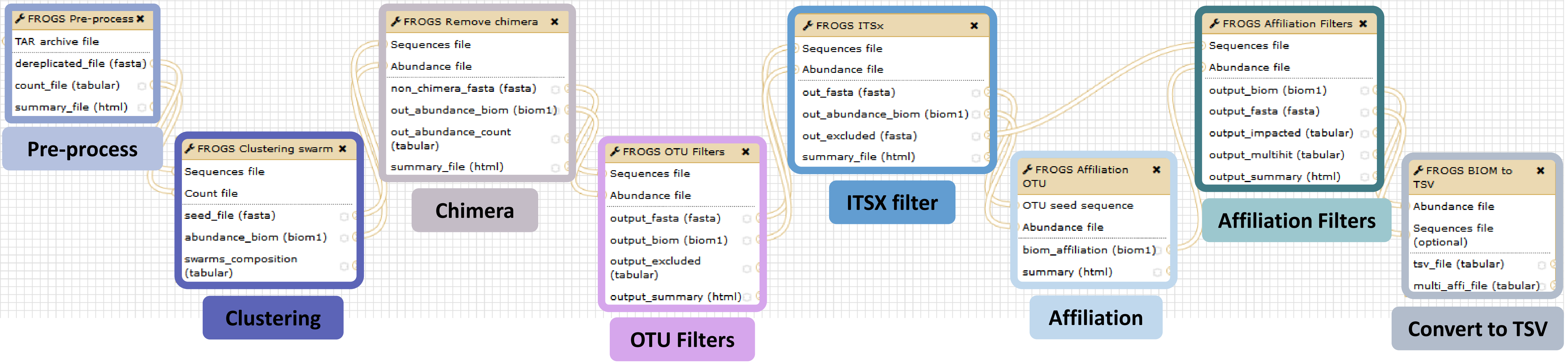
ITSX Tool
- if ITS sequences
- keep sequences with ITS signature *--check-its-only*
- ITS sequences

Affiliation Tool
- blastn affiliation
- abundance table
- cluster multi-affiliation list

Affiliation Filter Tool
- filter on:
  - on blast affiliations
  - remove blast taxonomies per key words *--ignore-blast-taxa*
- abundance table
- cluster multi-affiliation list

FROGS Pipeline

Minimal pipeline for ITS amplicon analyses



FROGS Pre-process
TAR 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 OTU Filters
Sequences file
Abundance file
output_fasta (fasta)
output_biom (biom1)
output_excluded (tabular)
output_summary (html)

OTU Filters

FROGS ITSx
Sequences file
Abundance file
out_fasta (fasta)
out_abundance_biom (biom1)
out_excluded (fasta)
summary_file (html)

ITSX filter

FROGS Affiliation OTU
OTU seed sequence
Abundance file
biom_affiliation (biom1)
summary (html)

Affiliation

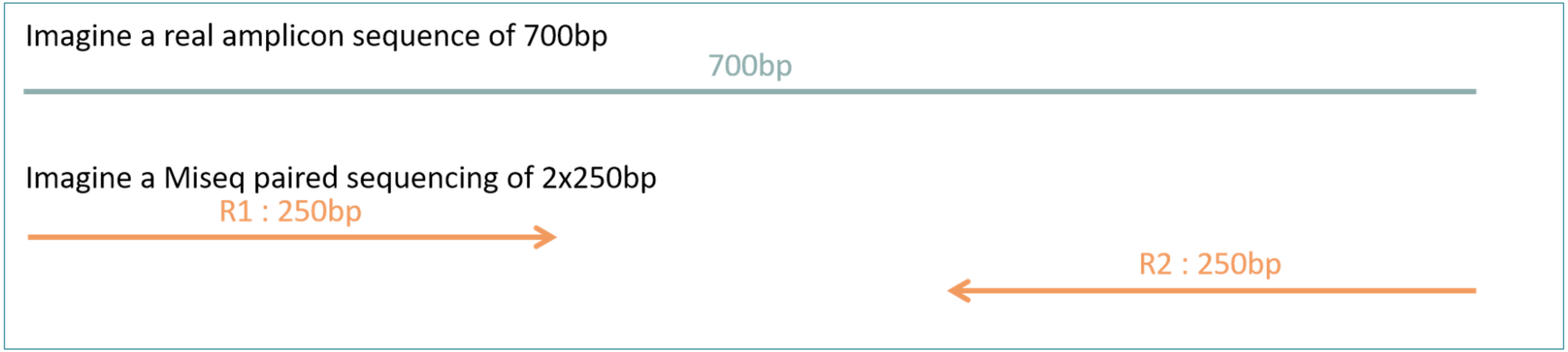
FROGS Affiliation Filters
Sequences file
Abundance file
output_biom (biom1)
output_fasta (fasta)
output_impacted (tabular)
output_multihit (tabular)
output_summary (html)

Affiliation Filters

FROGS BIOM to TSV
Abundance file
Sequences file (optional)
tsv_file (tabular)
multi_affi_file (tabular)

Convert to TSV

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

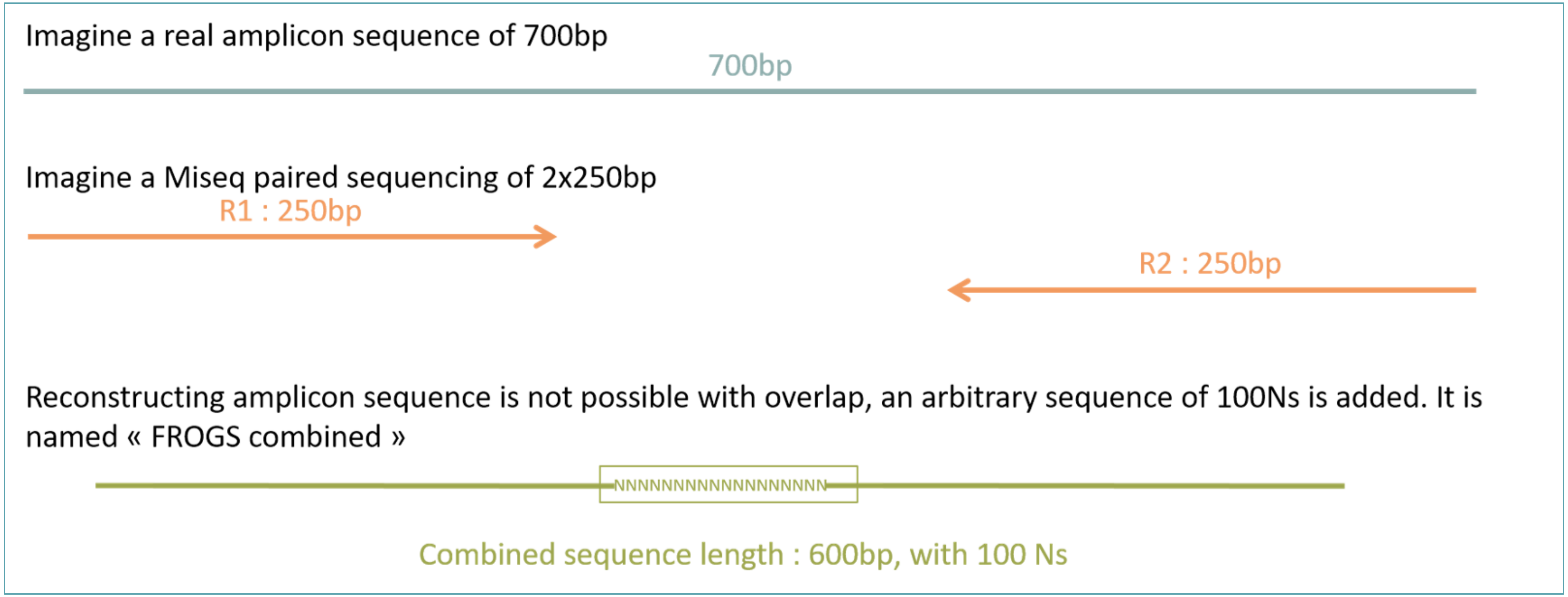


Imagine a real amplicon sequence of 700bp

Imagine a Miseq paired sequencing of 2x250bp

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



Imagine a real amplicon sequence of 700bp

700bp

Imagine a Miseq paired sequencing of 2x250bp

R1 : 250bp

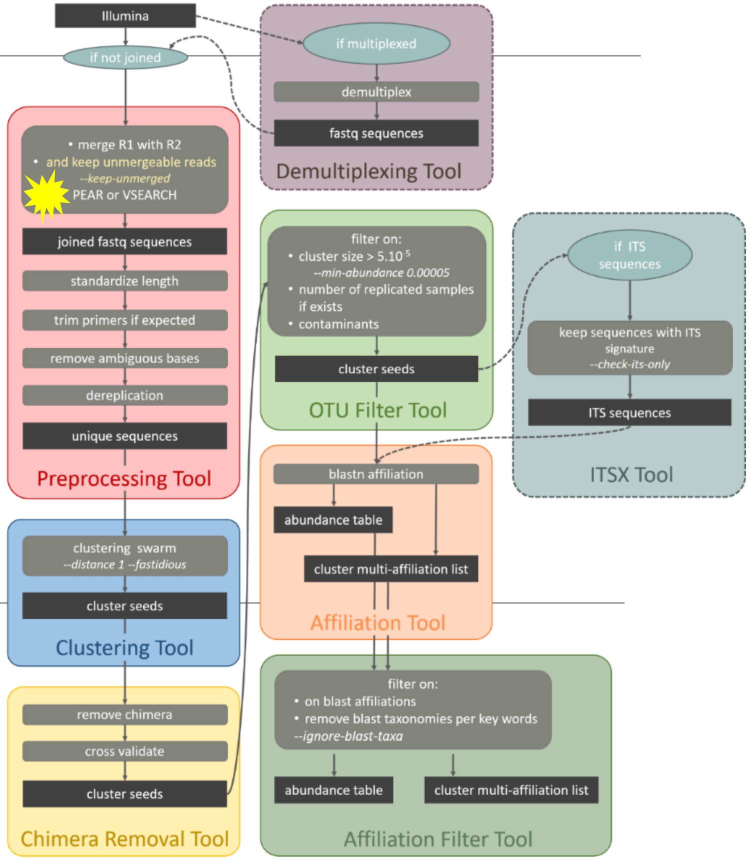
R2 : 250bp

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

NNNNNNNNNNNNNNNNNN

Combined sequence length : 600bp, with 100 Ns

Pre-process tool



Illumina

if not joined

- merge R1 with R2
- and keep unmergeable reads
  --keep-unmerged
  PEAR or VSEARCH

joined fastq sequences

standardize length

trim primers if expected

remove ambiguous bases

dereplication

unique sequences

Preprocessing Tool

if multiplexed

demultiplex

fastq sequences

Demultiplexing Tool

filter on:
- cluster size > 5.10^{-5}
  --min-abundance 0.00005
- number of replicated samples if exists
- contaminants

cluster seeds

OTU Filter Tool

if ITS sequences

keep sequences with ITS signature
--check-its-only

ITS sequences

ITSX Tool

clustering swarm
--distance 1 --fastidious

cluster seeds

Clustering Tool

blastn affiliation

abundance table

cluster multi-affiliation list

Affiliation Tool

remove chimera

cross validate

cluster seeds

Chimera Removal Tool

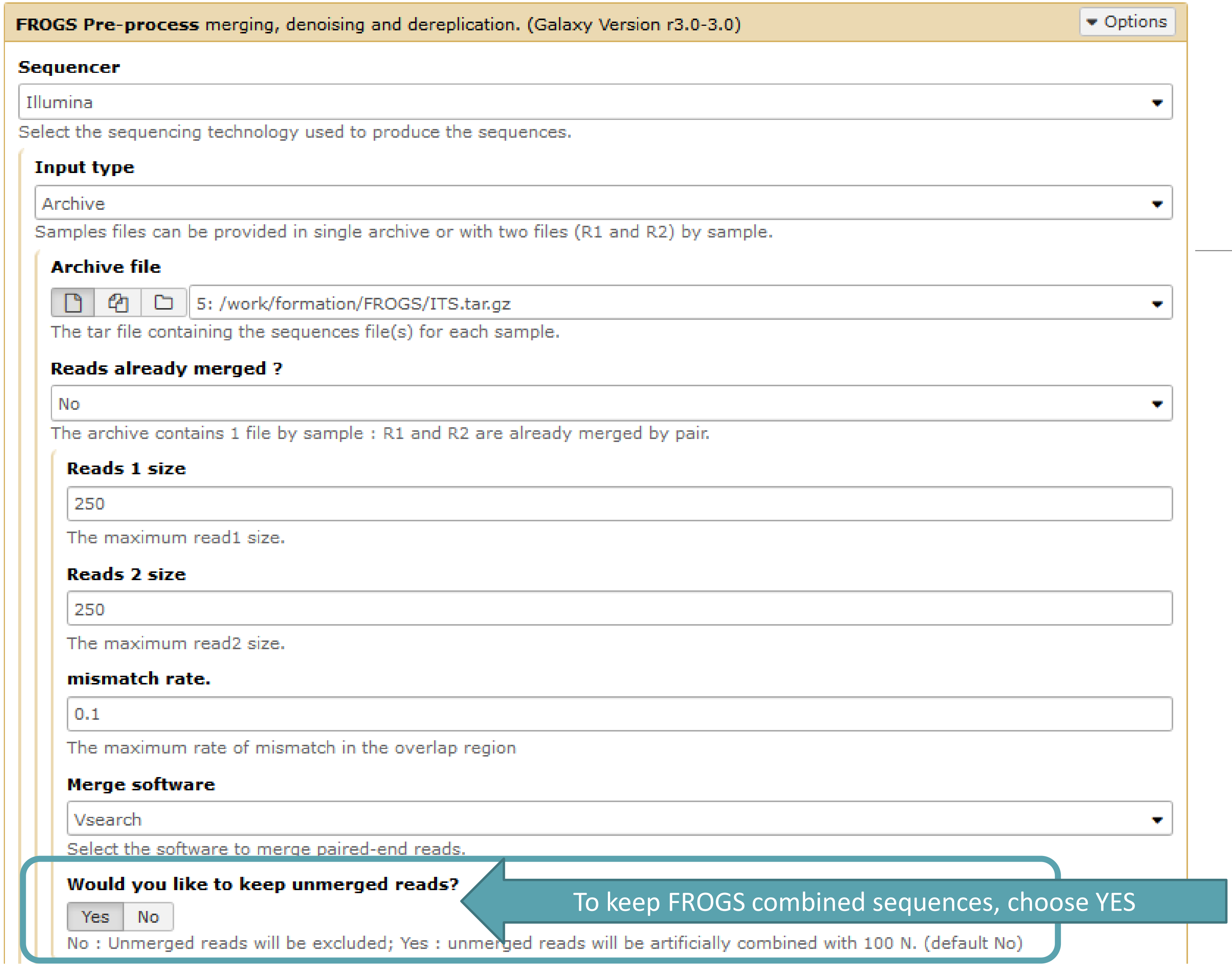
filter on:
- on blast affiliations
- remove blast taxonomies per key words
--ignore-blast-taxa

abundance table

cluster multi-affiliation list

Affiliation Filter Tool

ITS



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

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.

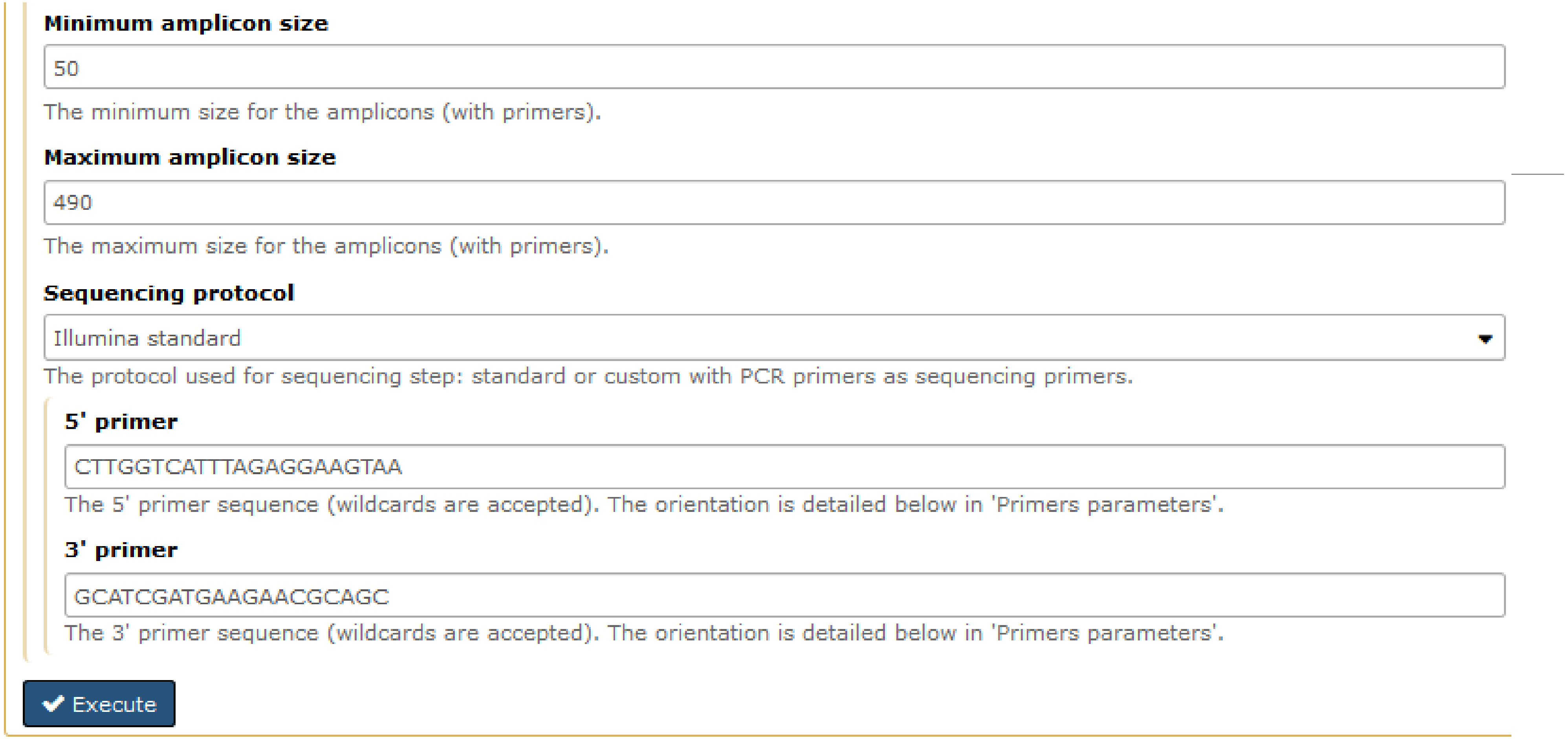
Would you like to keep unmerged reads?

[✓ 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

ITS



Minimum amplicon size

50

The minimum size for the amplicons (with primers).

Maximum amplicon size

490

The maximum size for the amplicons (with primers).

Sequencing protocol

Illumina standard

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

5' primer

CTTGGTCATTTAGAGGAAGTAA

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

3' primer

GCATCGATGAAGAACGCAGC

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

Primer 5’: CTTGGTCATTTAGAGGAAGTAA
Primer 3’: GCATCGATGAAGAACGCAGC

ITS

Exercise

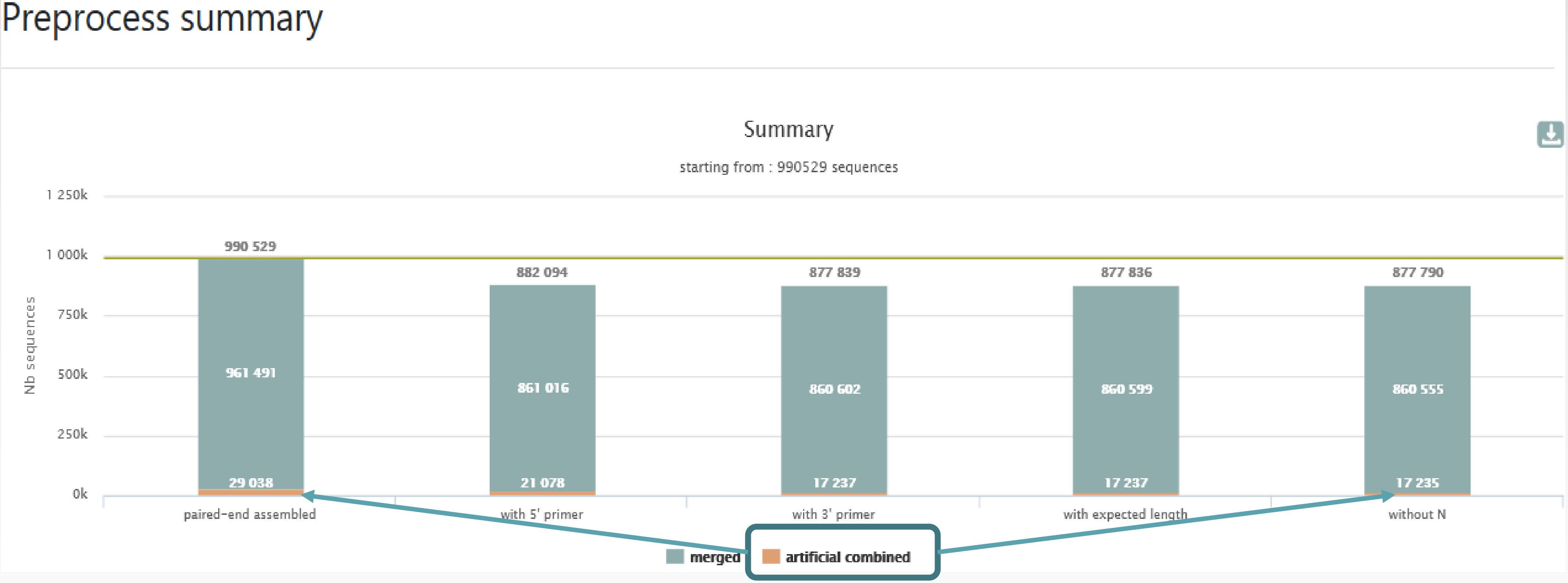
Go to « ITS » history

Launch the pre-process tool on this data set

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

ITS

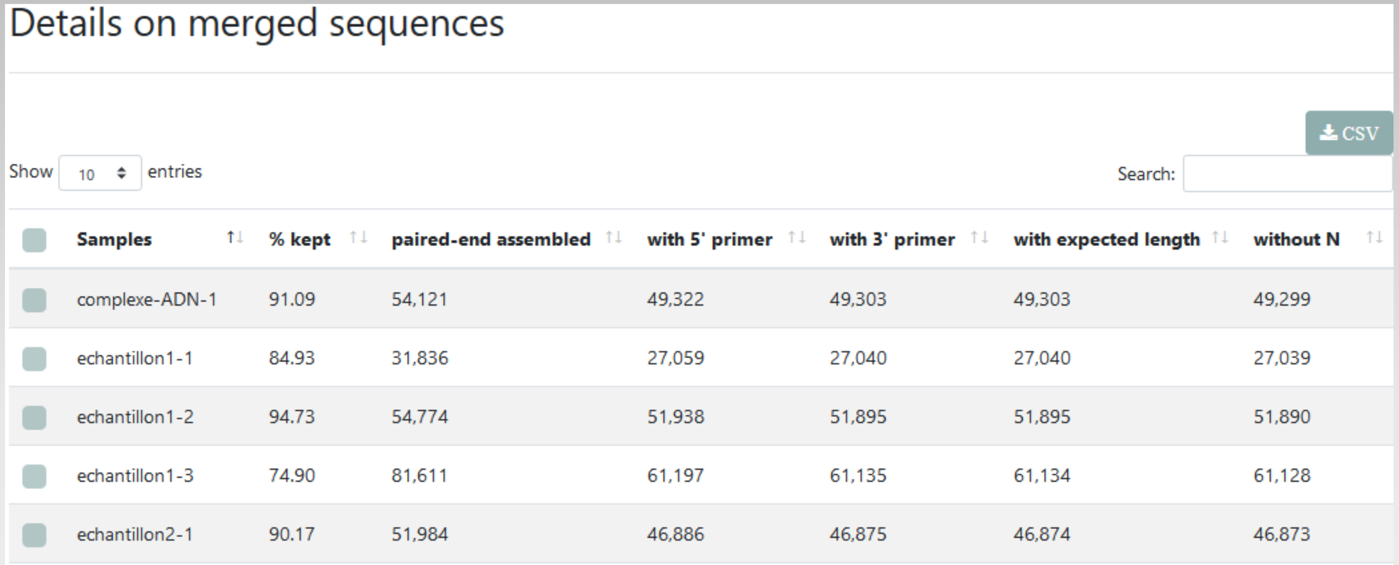
Explore Preprocess report.html



Preprocess summary

ITS

2 tables:

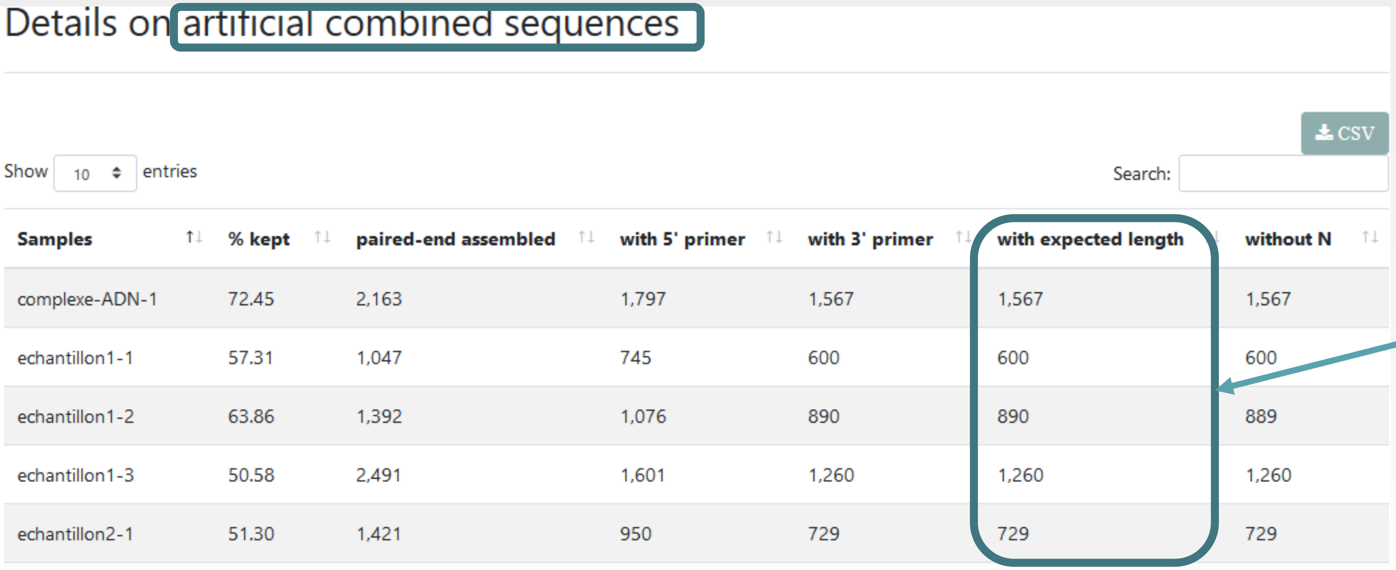


Details on merged sequences

CSV

Show 10 entries   Search:

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



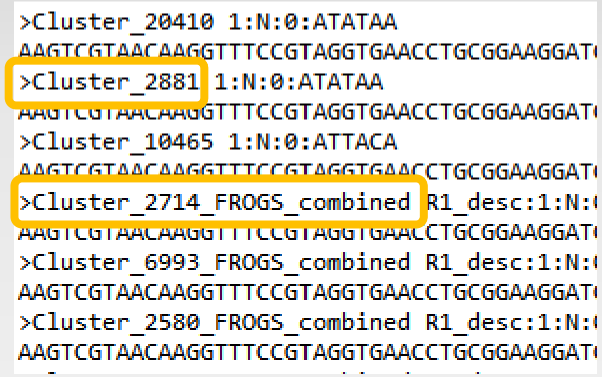
Details on artificial combined sequences

CSV

Show 10 entries   Search:

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

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:
AAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGAT
>Cluster_6993_FROGS_combined R1_desc:1:N:
AAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGAT
>Cluster_2580_FROGS_combined R1_desc:1:N:
AAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGAT
```

Filter only on minimum length for « combined ».

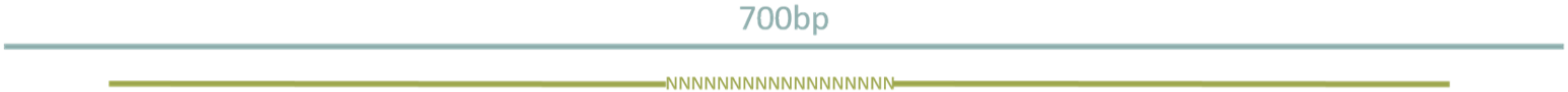
Minimum length =
R1 + 100N + R2 – primers sizes

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

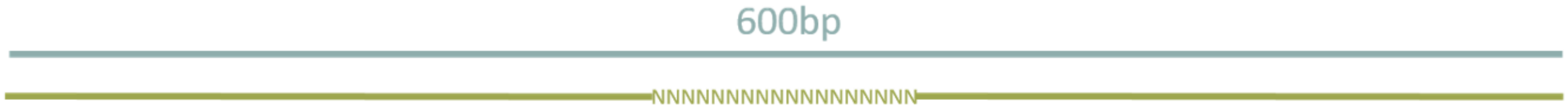
FROGS "combined" sequences are **artificial** and present particular features especially on size.

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

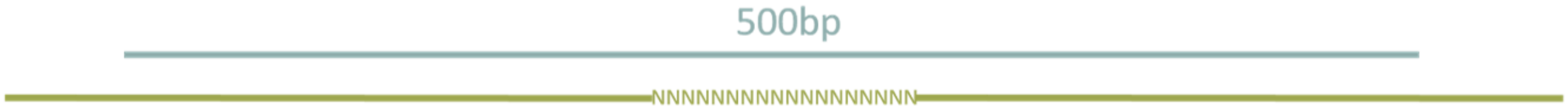
Case 1: real amplicon ≥ 601 bp ➔ “FROGS combined” length is smaller than the reality



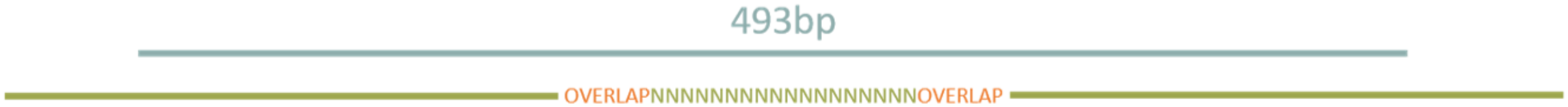
Case 2: real amplicon = 600 bp ➔ “FROGS combined” length is equal to the reality



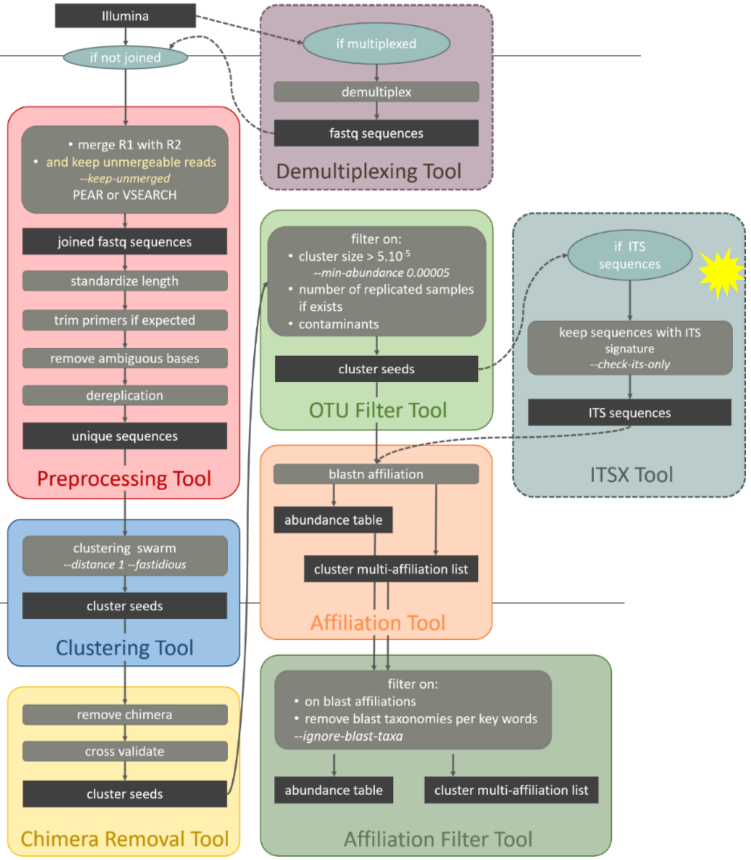
Case 3: real amplicon ≥ 500 and ≤ 599 ➔ “FROGS combined” length is greater than the reality



Case 4 : real amplicon ≥ 491 and ≤ 499 ➔ FROGS combined length is greater than the reality and duplicate small sequences (between 1 and 9 bp flanking the 100 Ns added).



ITSx tools



Illumina

if not joined

- merge R1 with R2
- and keep unmergeable reads
  --keep-unmerged
  PEAR or VSEARCH

joined fastq sequences

standardize length

trim primers if expected

remove ambiguous bases

dereplication

unique sequences

Preprocessing Tool

if multiplexed

demultiplex

fastq sequences

Demultiplexing Tool

clustering swarm
--distance 1 --fastidious

cluster seeds

Clustering Tool

remove chimera

cross validate

cluster seeds

Chimera Removal Tool

filter on:
- cluster size > 5.10^{-5}
  --min-abundance 0.00005
- number of replicated samples if exists
- contaminants

cluster seeds

OTU Filter Tool

if ITS sequences

keep sequences with ITS signature
--check-its-only

ITS sequences

ITSX Tool

blastn affiliation

abundance table

cluster multi-affiliation list

Affiliation Tool

filter on:
- on blast affiliations
- remove blast taxonomies per key words
--ignore-blast-taxa

abundance table

cluster multi-affiliation list

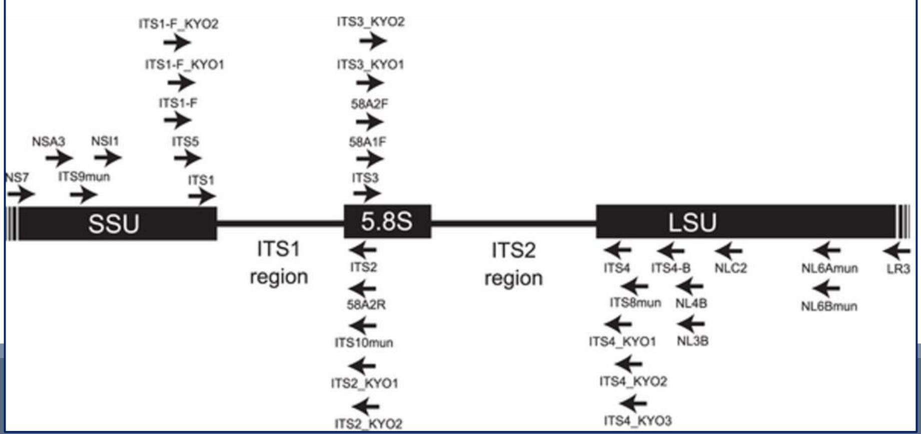
Affiliation Filter Tool

What is the purpose of the ITSx tool?

- ITSx is a tool to **filter** sequences.
- ITSx **identifies** and **trimms** 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).

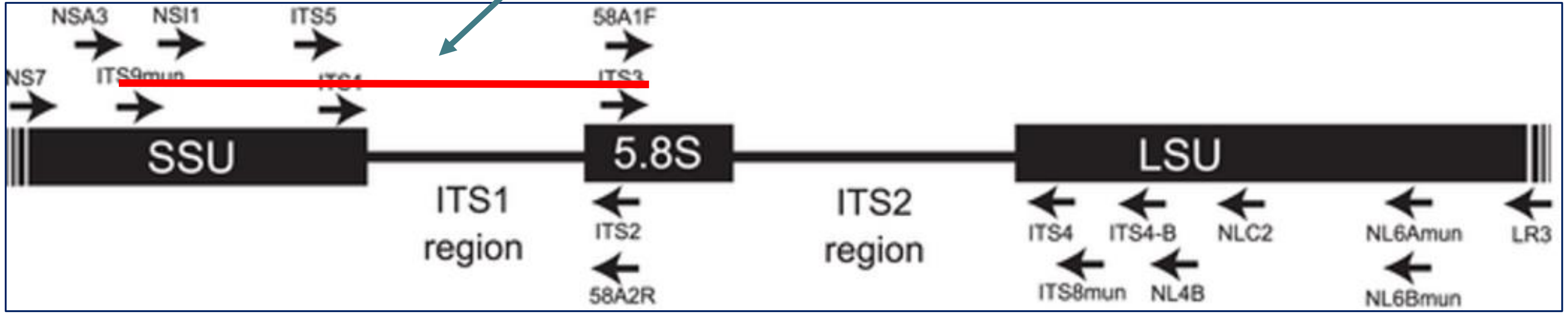
Map of nuclear ribosomal RNA genes and their ITS regions.



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?

ITS amplicon target

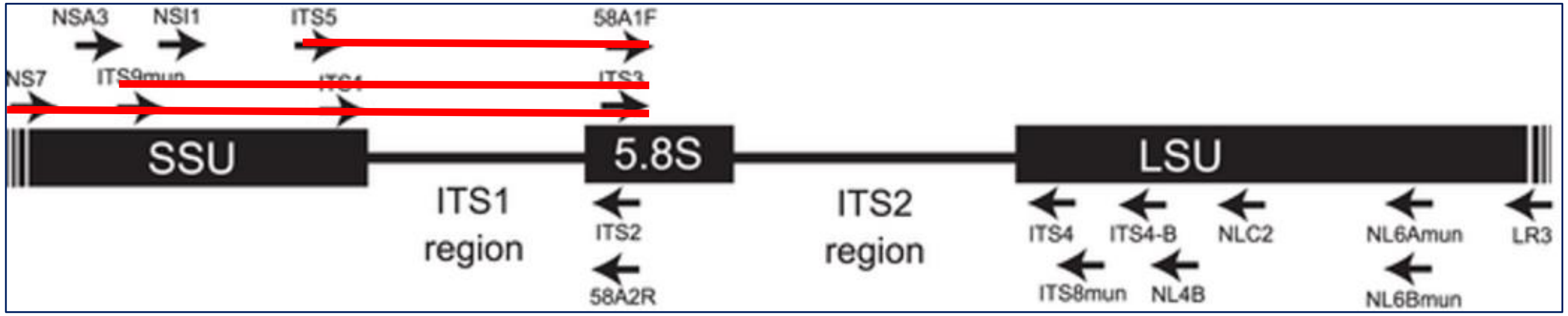


1st case: choose to trim
ITS1 is well detected
SSU part and 5.8S part are trimmed
Result:

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

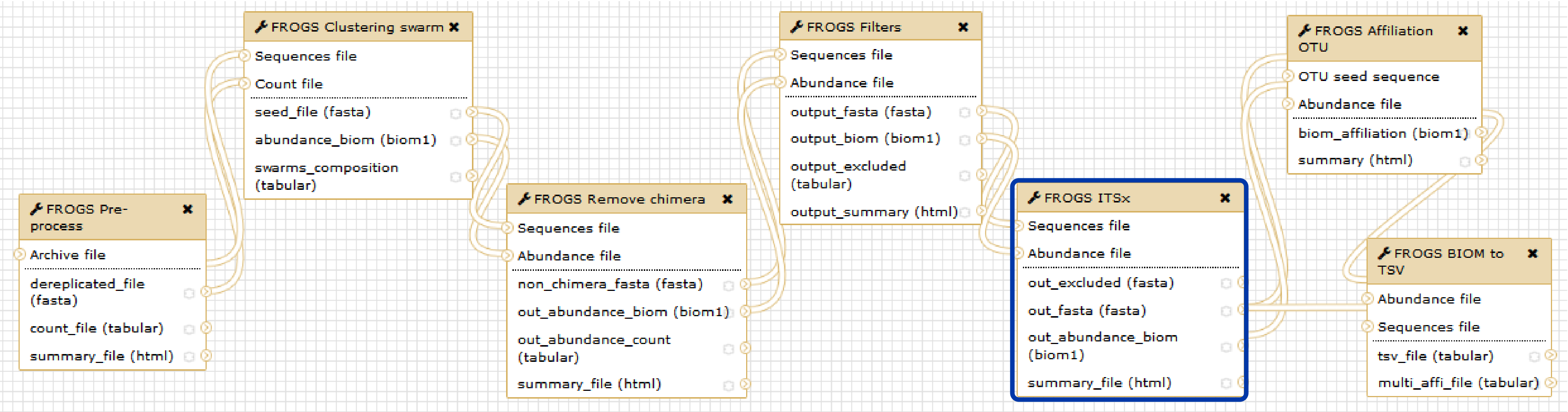
Check only if sequence is detected as ITS? Yes or not?

- 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.
  - using a database with only ITS part

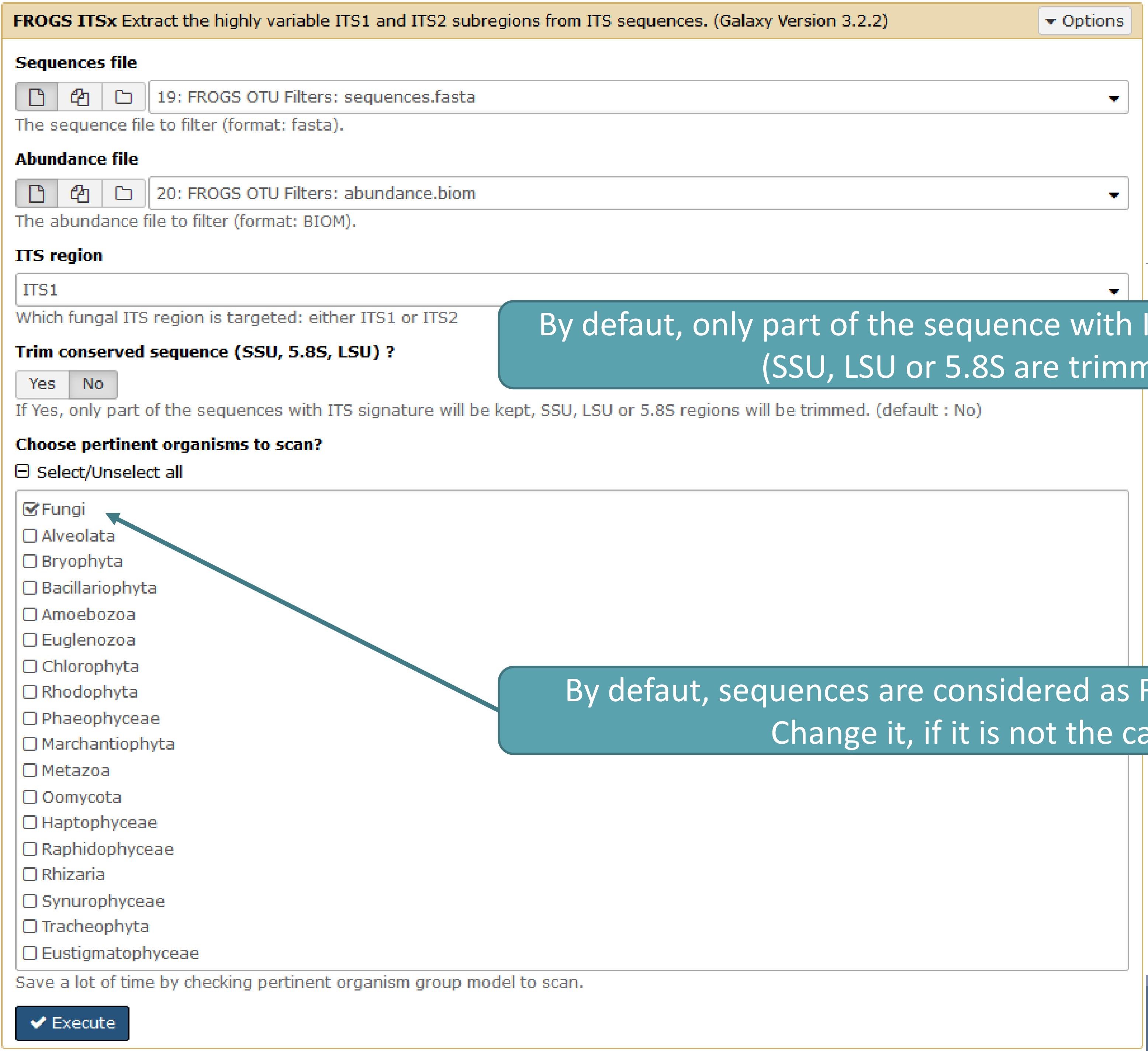


When should we use ITSx ?

After filtering !



ITSx is a fastidious step



FROGS ITSx Extract the highly variable ITS1 and ITS2 subregions from ITS sequences. (Galaxy Version 3.2.2)

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)

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

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

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

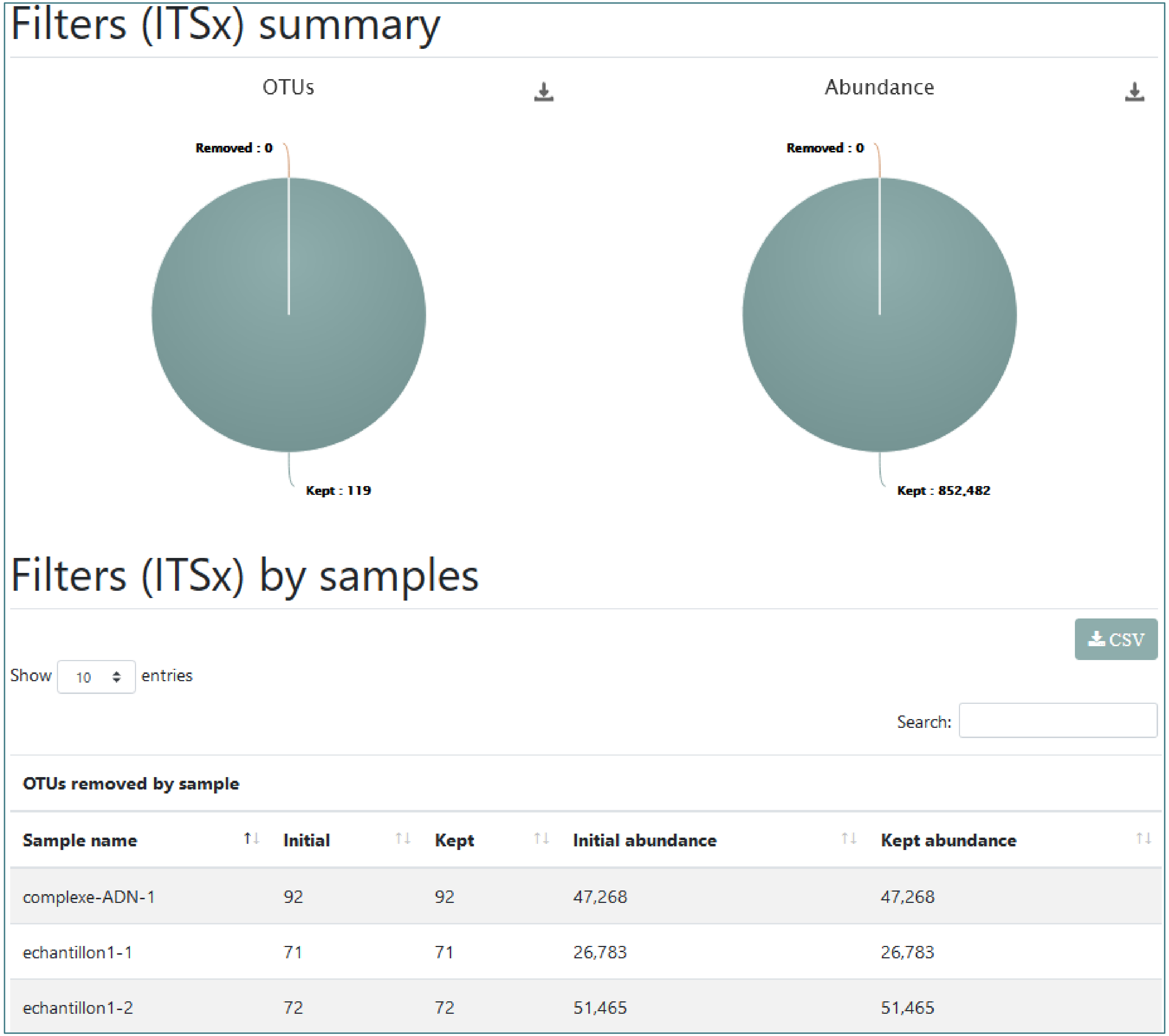
By defaut, sequences are considered as FUNGI sequences. Change it, if it is not the case.

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.

Report.html, ITSX output



Filters (ITSx) summary

Filters (ITSx) by samples

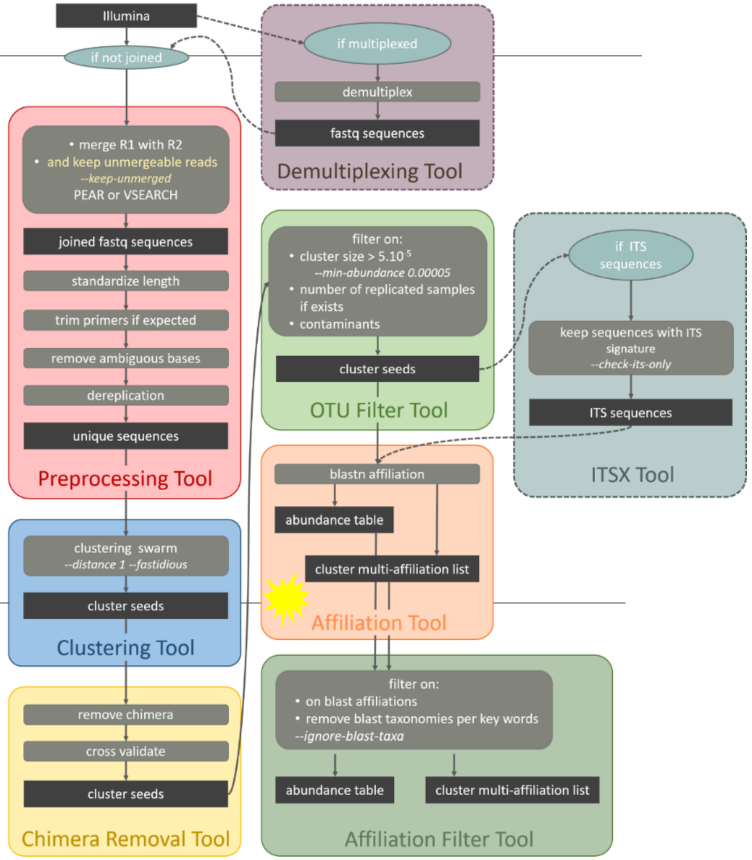
Show 10 entries

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



Illumina

if not joined

if multiplexed

demultiplex

fastq sequences

Demultiplexing Tool

- merge R1 with R2
- and keep unmergeable reads
--keep-unmerged
PEAR or VSEARCH

joined fastq sequences

standardize length

trim primers if expected

remove ambiguous bases

dereplication

unique sequences

Preprocessing Tool

filter on:
- cluster size > 5.10^{-5}
--min-abundance 0.00005
- number of replicated samples if exists
- contaminants

cluster seeds

OTU Filter Tool

if ITS sequences

keep sequences with ITS signature
--check-its-only

ITS sequences

ITSX Tool

blastn affiliation

abundance table

cluster multi-affiliation list

Affiliation Tool

clustering swarm
--distance 1 --fastidious

cluster seeds

Clustering Tool

remove chimera

cross validate

cluster seeds

Chimera Removal Tool

filter on:
- on blast affiliations
- remove blast taxonomies per key words
--ignore-blast-taxa

abundance table

cluster multi-affiliation list

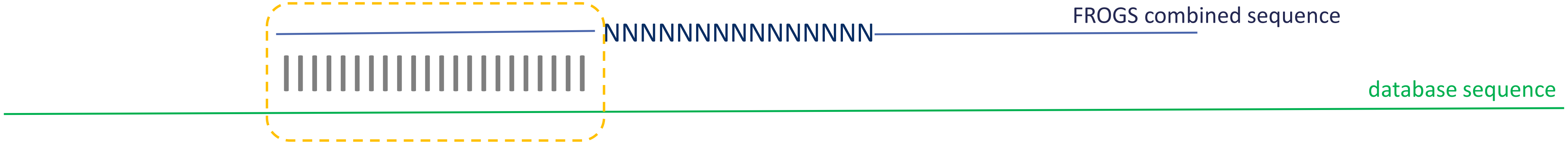
Affiliation Filter Tool

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 bests 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.

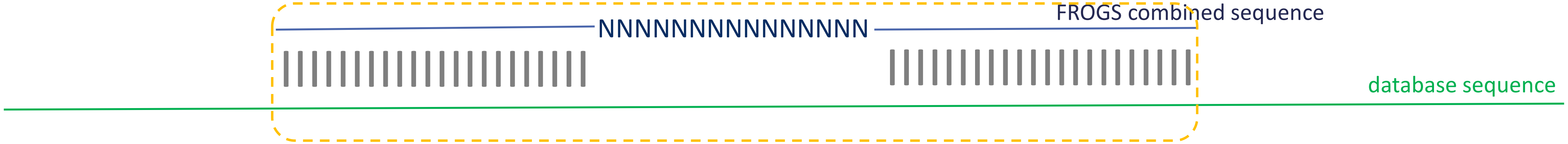
What is special about the affiliation of ITS/FROGS combined sequences?

- 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 !

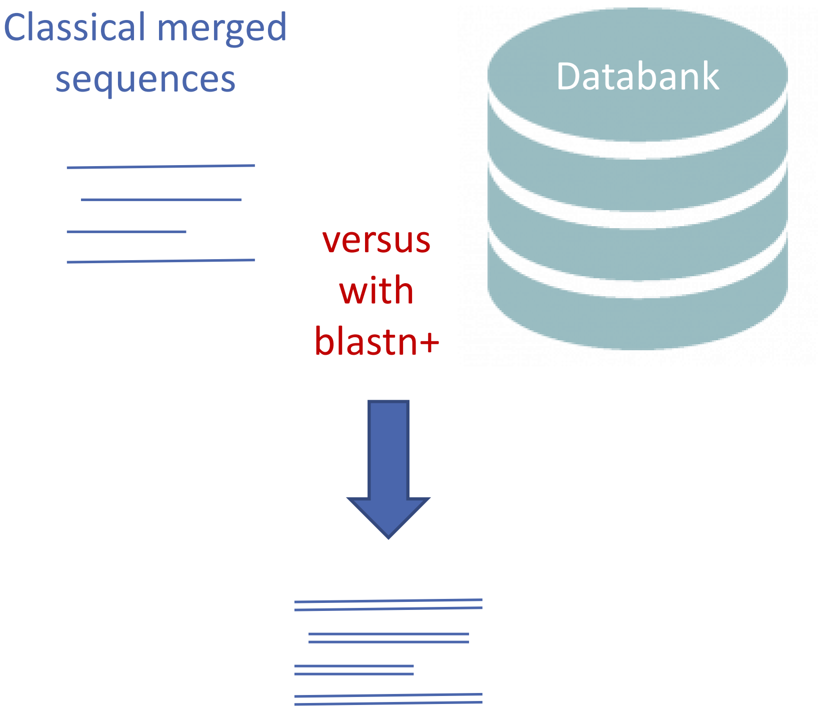
- It is necessary to use a global aligner *i.e.* Needleall (the sequence must be aligned in its entirety), but it is computationally too hard.



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

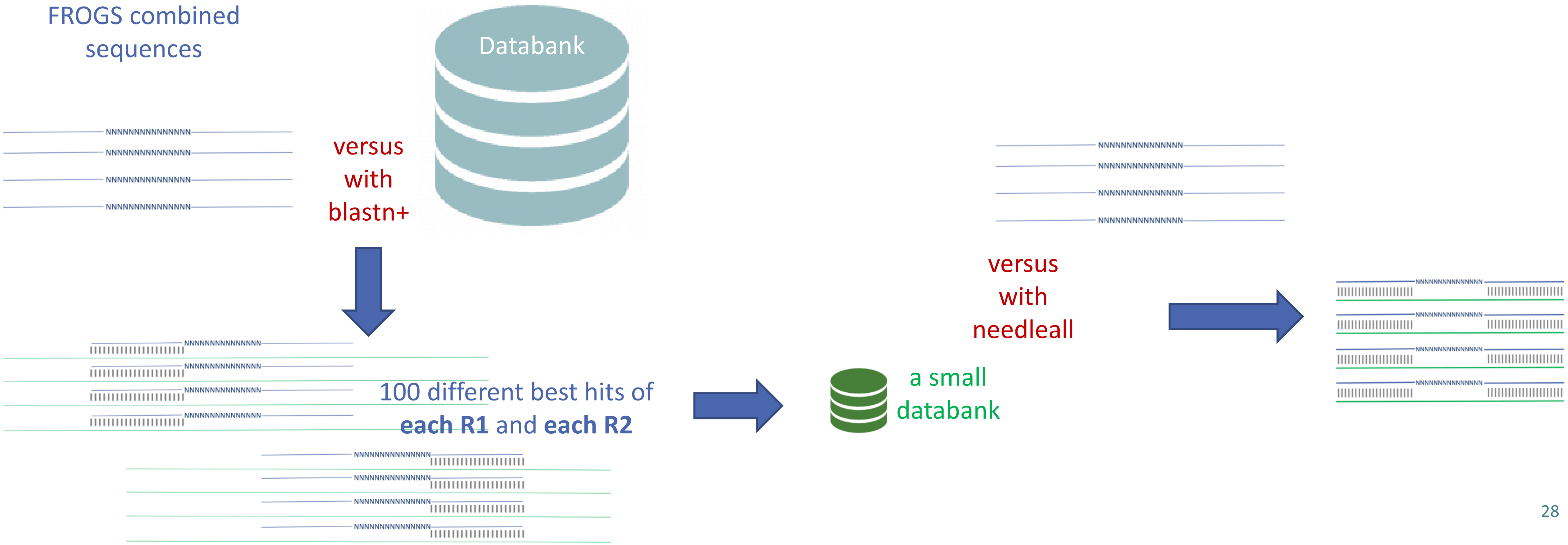
Solution:

- 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



FROGS combined sequences

Databank

versus with blastn+

100 different best hits of **each R1** and **each R2**

a small databank

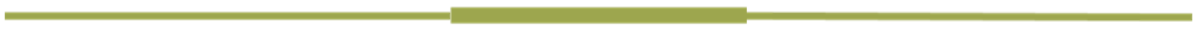
versus with needleall

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

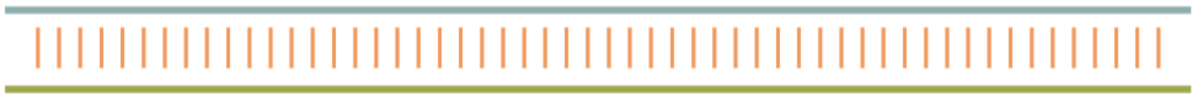
Case 1: a sequencing of overlapping sequences *i.e.* 16S V3-V4 amplicon MiSeq sequencing

Imagine a real amplicon sequence of 400bp

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



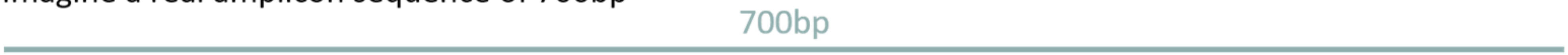
Affiliation is notably made by a local alignment with NCBI Blast+



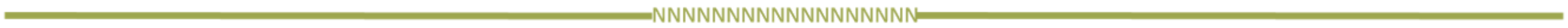
Imagine a perfect sequencing without error:
classical %id = number of matches / alignment length = 400 matches / 400 positions = 100% identity

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

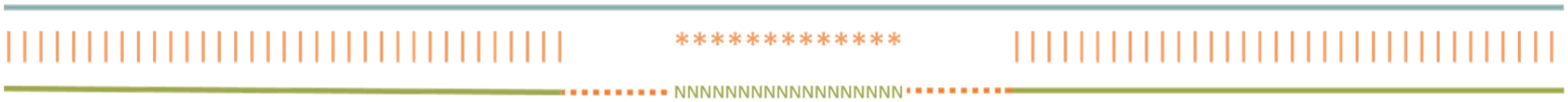
Imagine a real amplicon sequence of 700bp



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



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



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.

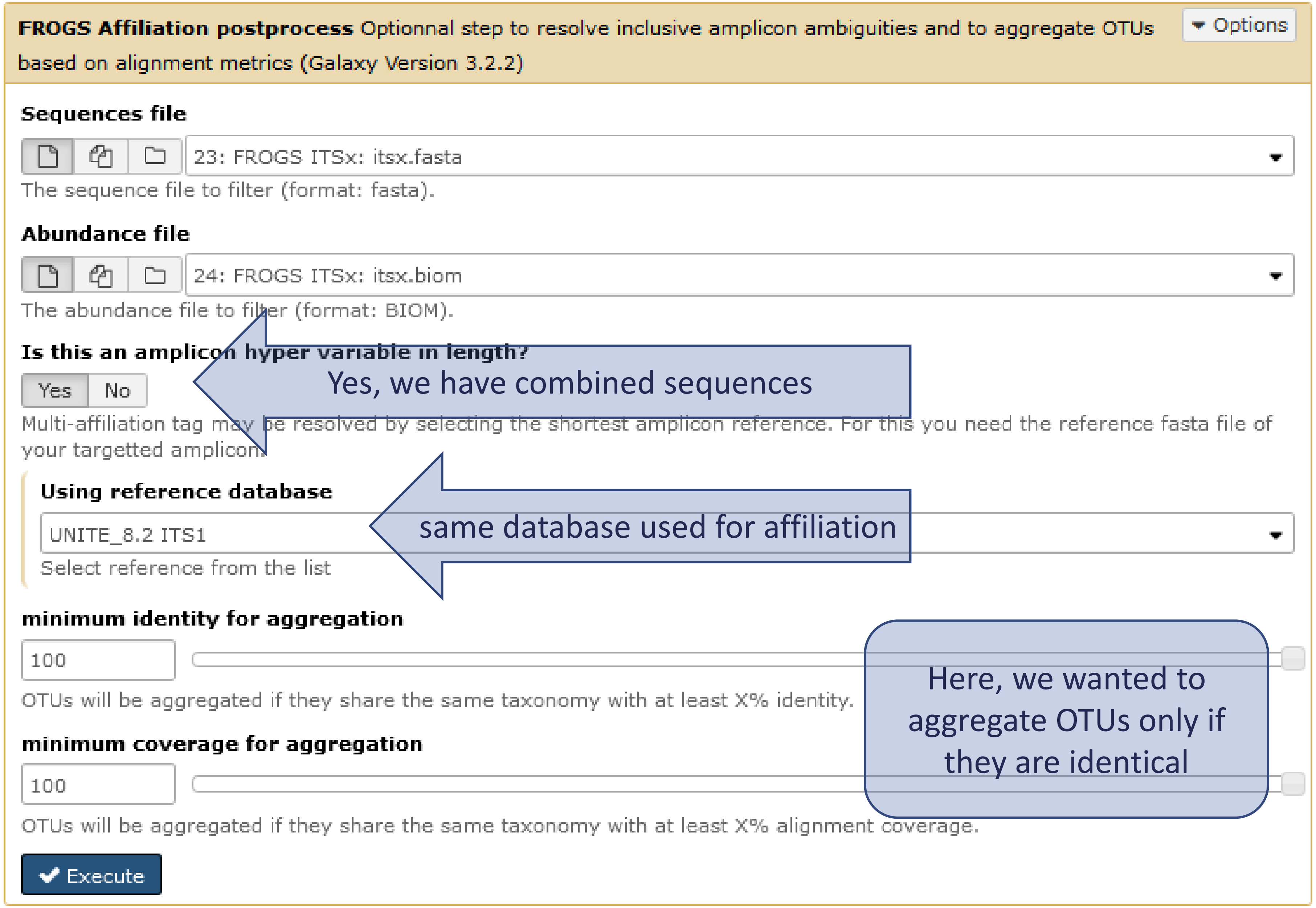
What is the purpose of the *Affiliation post-process* tool ?

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)

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.

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:

```
>Cluster_3
AAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTAGTGATTGCCTTTATAGGCTTATAACTATATCCACTTACACCTGTGAACTGTTCTACTACTTGACGCAAGTCGAGT
ATTTTTACAAACAATGTGTAATGAACGTCGTTTTATTATAACAAAATAAAACTTTCAACAACGGATCTCTTGGCTCTC

>Cluster_54
AAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTAGTGATTGCCTTTATAGGCTTATAACTATATCCACTTACACCTGTGAACTGTTCTACTACTTGACGCAAGTCGAGT
ATTTTTACAAACAATGTGTAATGAACGTCGTTTTATTATAACAAAATAAAACTTTCAACAACGGATCTCTTGGTTCCG

>Cluster_414_FROGS_combined
AAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTAGTGATTGCCTTTATAGGCTTATAACTATATCCACTTACACCTGTGAACTGTTCTACTACTTGACGCAAGTCGAGT
ATTTTTACAAACAATGTGTAATGAACGTCGTTTTATTATAACAAAATAAAACTTTCAACAACGGATCTCTTGGCTCTCGCATCGATGAAGAACGCAGCAGATCGGAANNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTCCGATCTCTTGGT
CATTTAGAGGAAGTAAAAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCGGAAGGATCATTAGTGATTGCCTTTATAGGCTTATAACTATATCCACTTACACCTGTGAACTGTTCTACTAC
TTGACGCAAGTCGAGTATTTTTACAAACAATGTGTAATGAACGTCGTTTTATTATAACAAAATAAAACTTTCAACAACGGATCTCTTGGCTCTC
```

Exemple

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

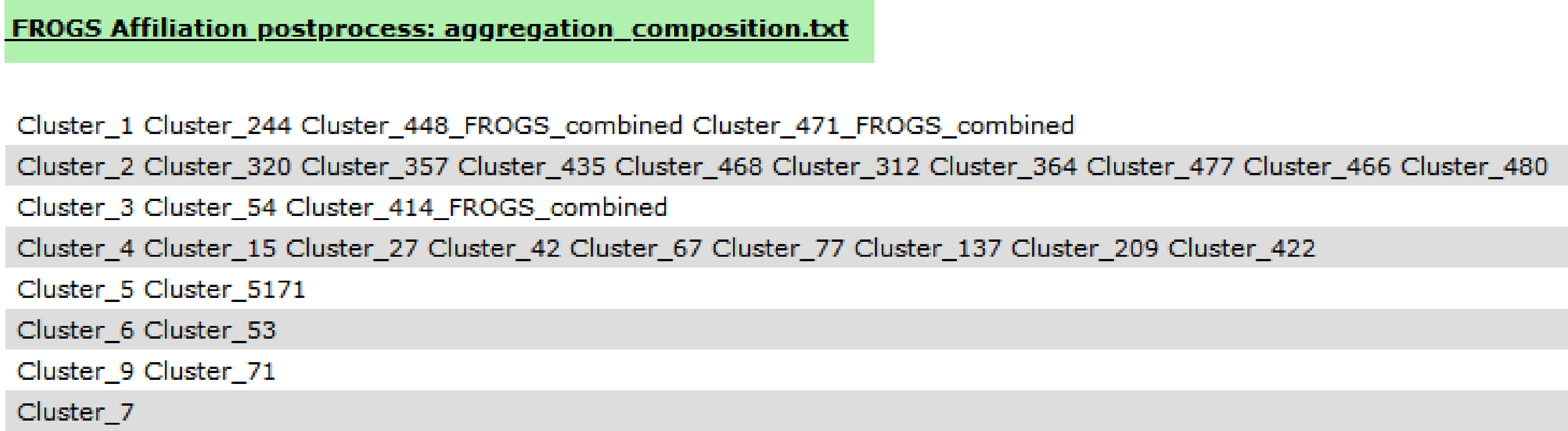
```
>Cluster_3
GTGATTGCCTTTATAGGCTTATAACTATATCCACTTACACCTGTGAACTGTTCTACTACTTGACGCAAGTCGAGTATTTTTACAAACAATGTGTAATGAACGTCGTTTTATTATAACAAAATA

>Cluster_54
GTGATTGCCTTTATAGGCTTATAACTATATCCACTTACACCTGTGAACTGTTCTACTACTTGACGCAAGTCGAGTATTTTTACAAACAATGTGTAATGAACGTCGTTTTATTATAACAAAATA

>Cluster_414_FROGS_combined
GTGATTGCCTTTATAGGCTTATAACTATATCCACTTACACCTGTGAACTGTTCTACTACTTGACGCAAGTCGAGTATTTTTACAAACAATGTGTAATGAACGTCGTTTTATTATAACAAAATA
```

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

Exemple

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

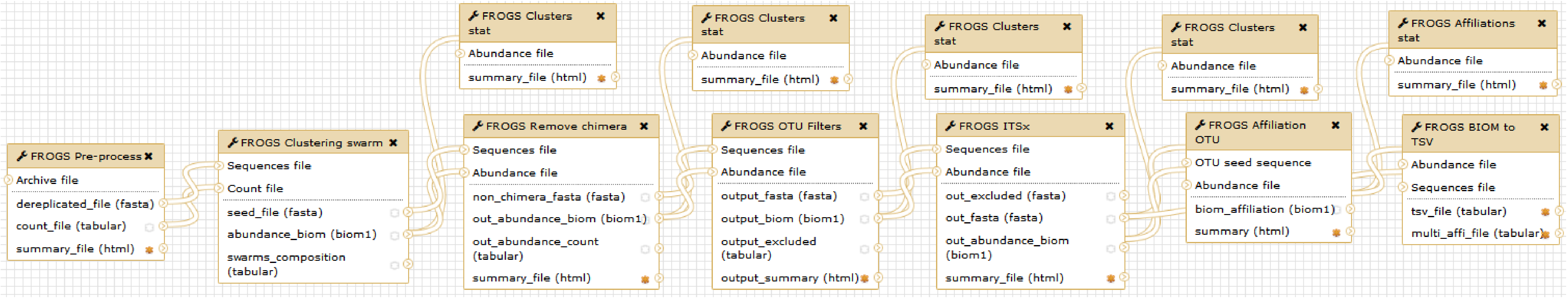
Cluster_3, Cluster_54 and Cluster_414_FROGS_combined are **aggregated** in a same OTU



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



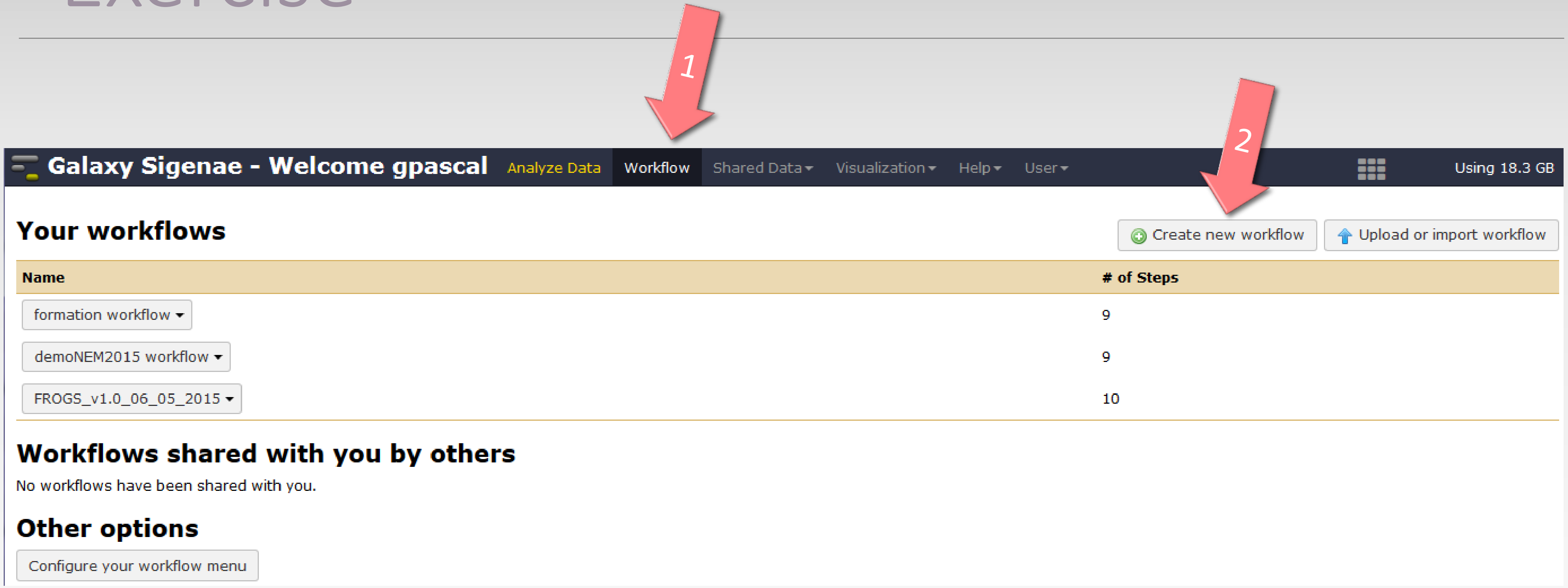
Workflow are useful for routine analyses

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

Practice

CREATE YOUR OWN WORKFLOW !

Exercise



Your workflows

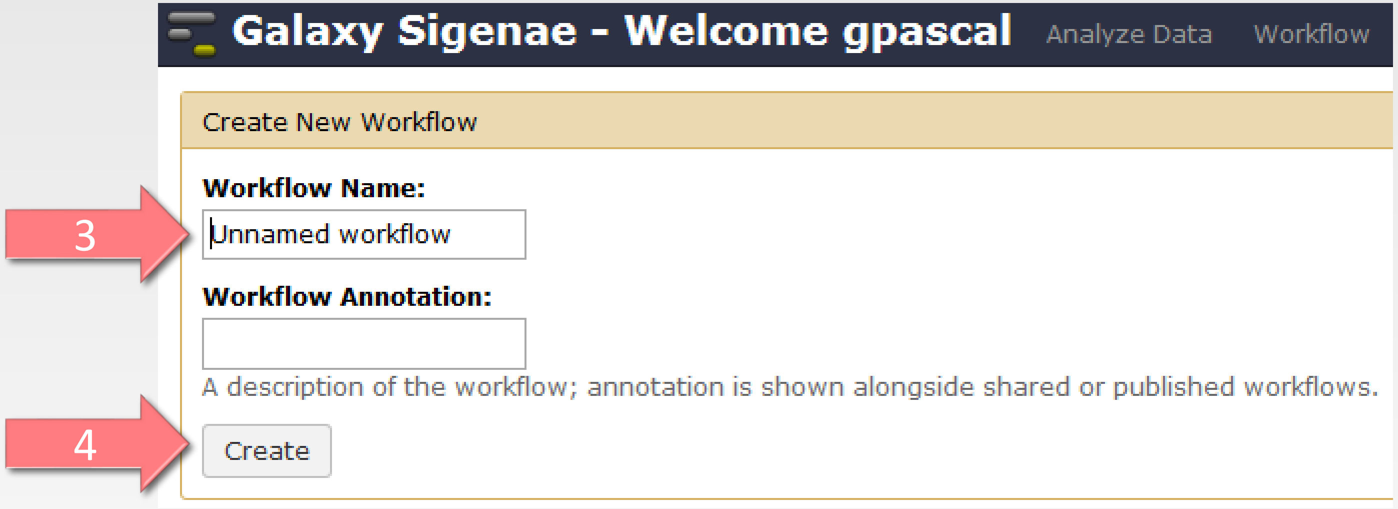
Name	# of Steps
formation workflow	9
demoNEM2015 workflow	9
FROGS_v1.0_06_05_2015	10

Workflows shared with you by others

No workflows have been shared with you.

Other options

Exercise



Galaxy Sigenae - Welcome gpascal

Create New Workflow

3

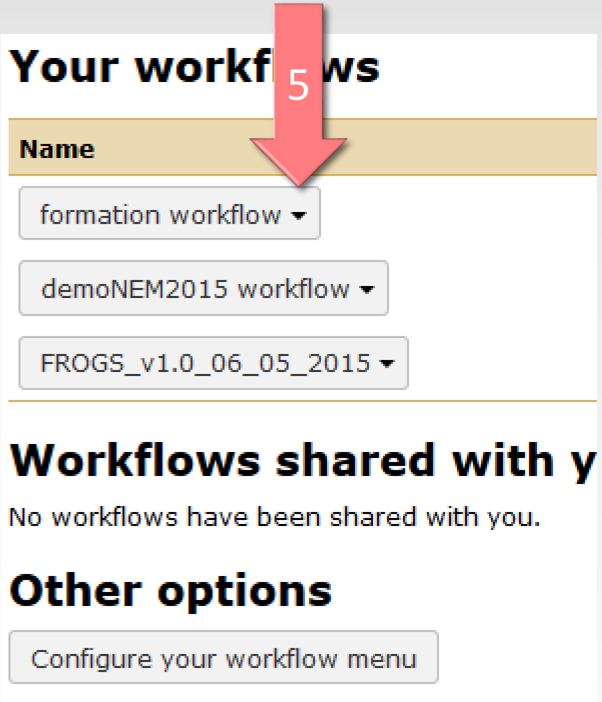
Workflow Name: Unnamed workflow

Workflow Annotation:

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

4

Exercise



5

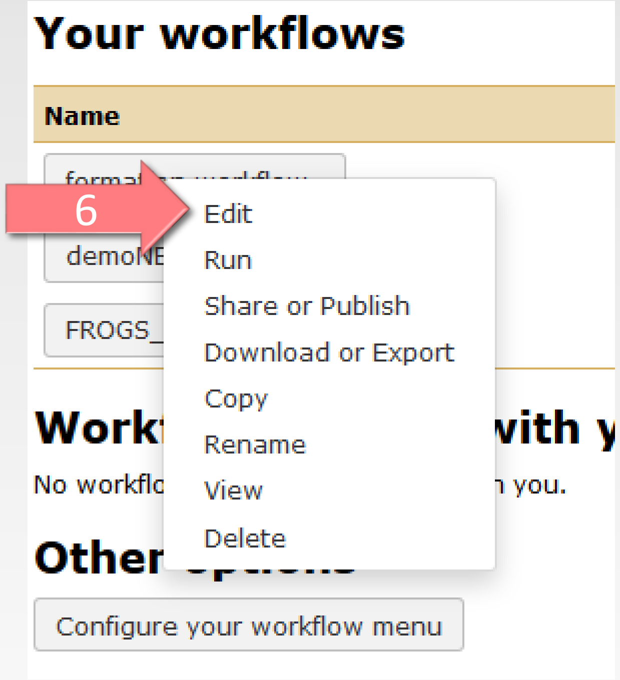
Your workflows

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

Workflows shared with y

No workflows have been shared with you.

Other options

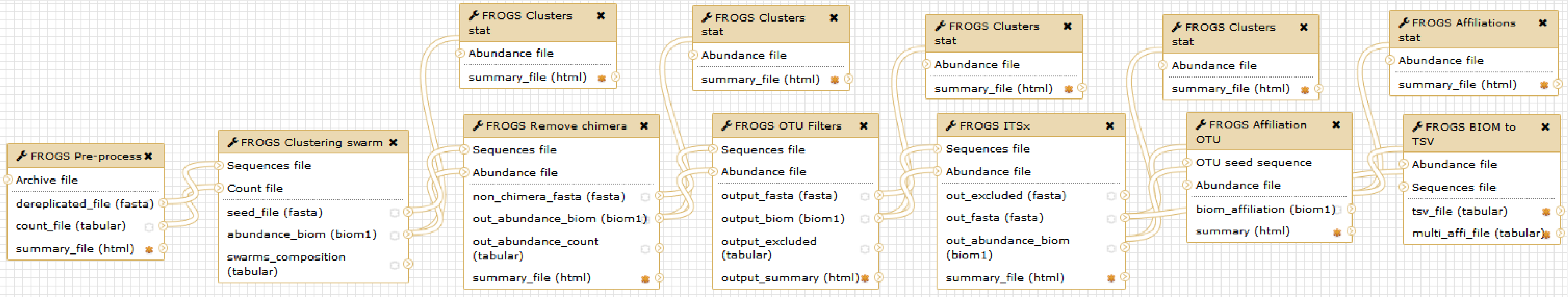


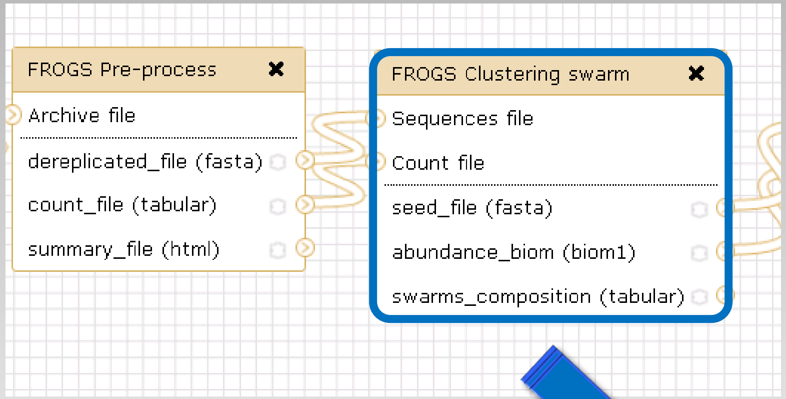
6

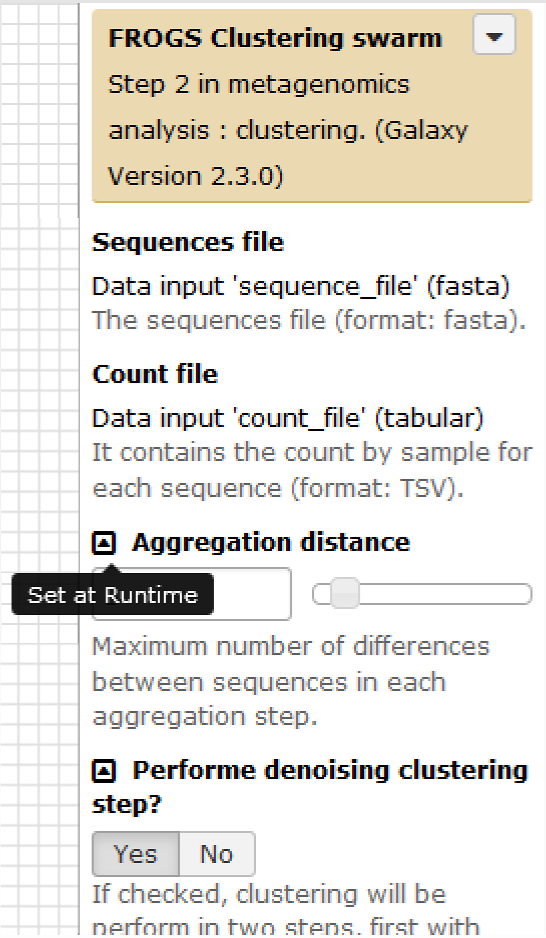
Your workflows

Name

Solution of exercise:







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?

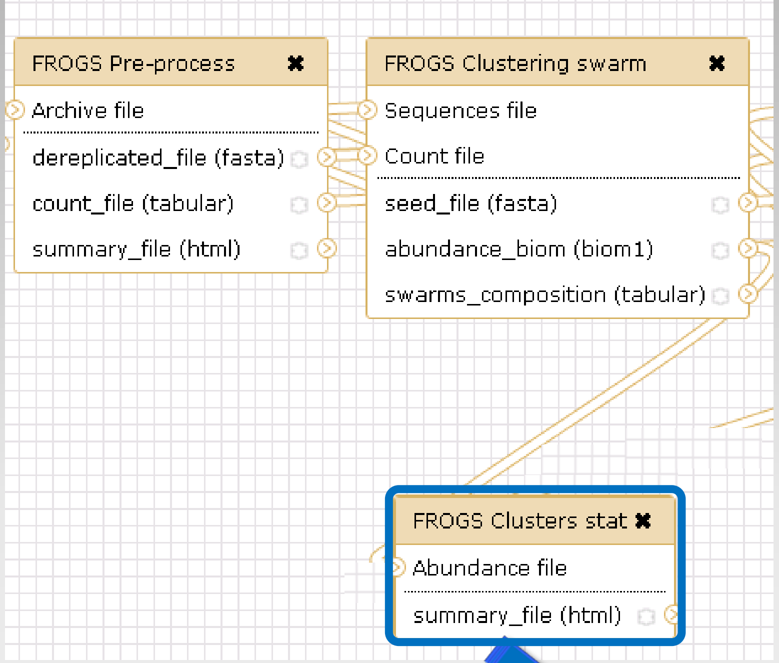
Yes No

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

For each tool, think to:

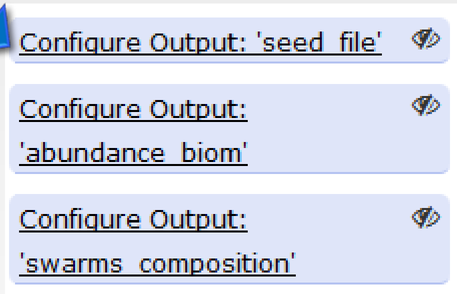
1. Fixe parameter ?

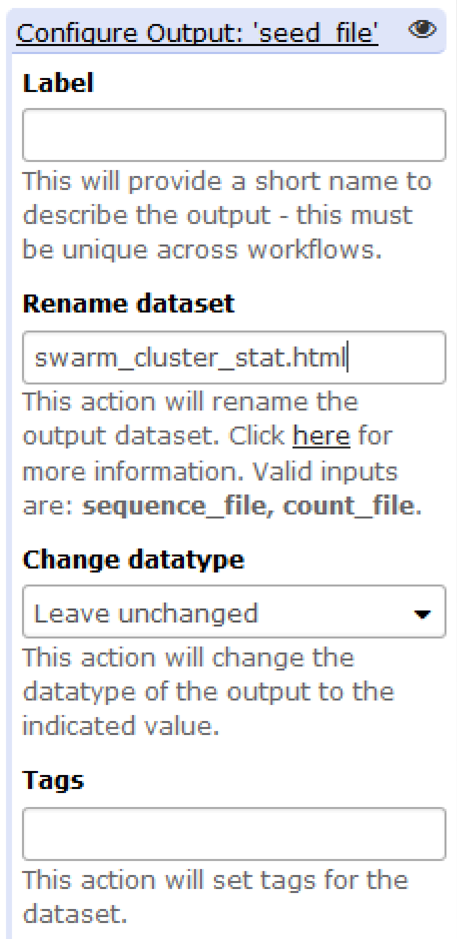
?



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

swarm_cluster_stat.html

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

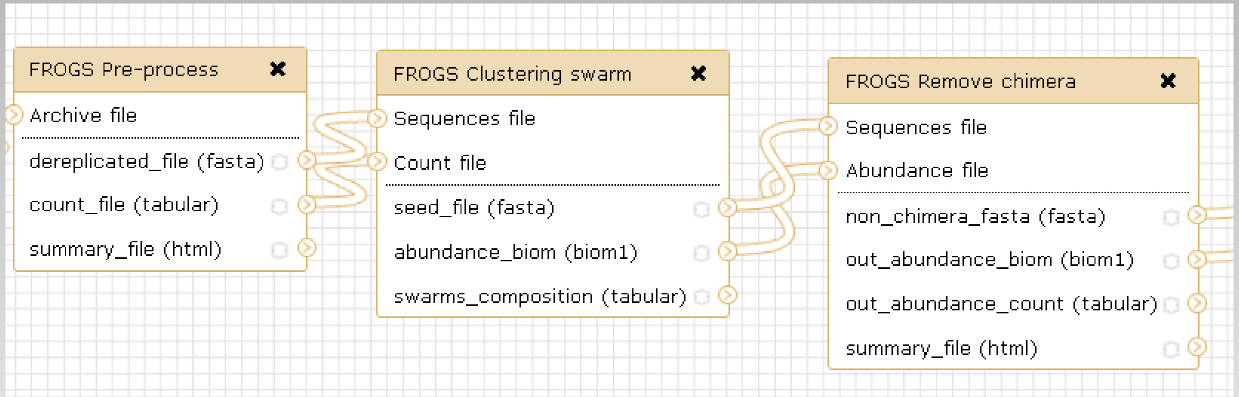
Change datatype

Leave unchanged

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

Tags

This action will set tags for the dataset.



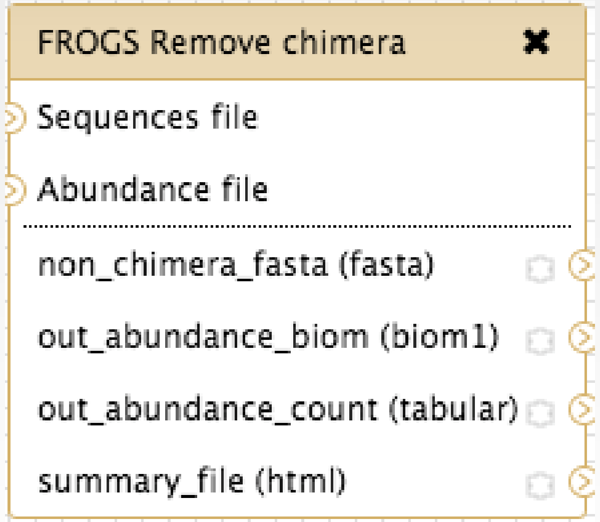
FROGS Pre-process
Archive file
dereplicated_file (fasta)
count_file (tabular)
summary_file (html)

FROGS Clustering swarm
Sequences file
Count file
seed_file (fasta)
abundance_biom (biom1)
swarms_composition (tabular)

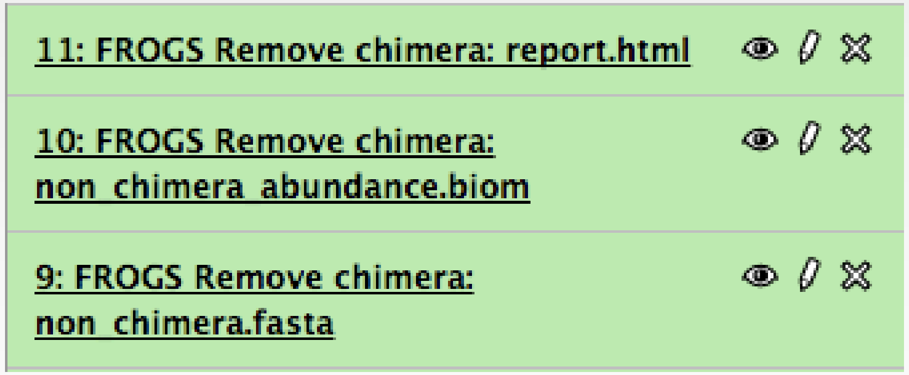
FROGS Remove chimera
Sequences file
Abundance file
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
out_abundance_count (tabular)
summary_file (html)

For each tool, think to:

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



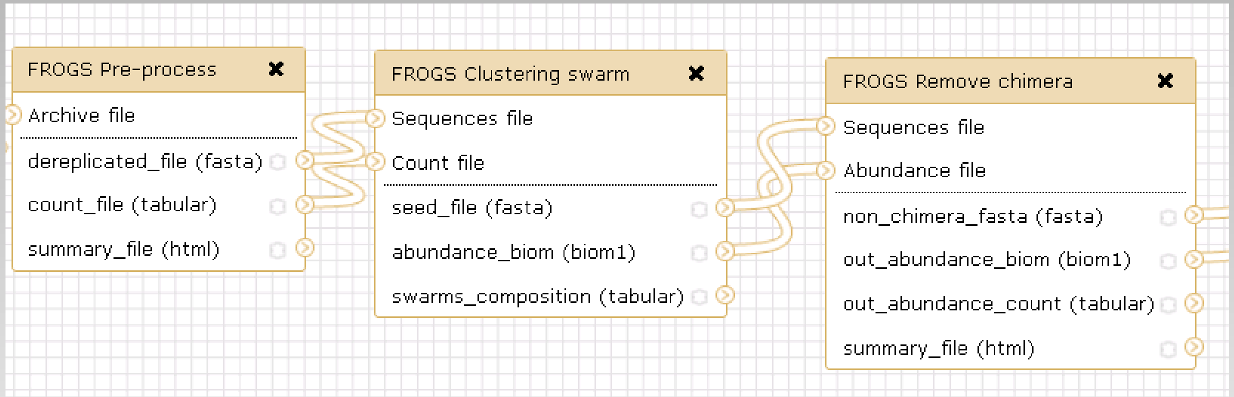
FROGS Remove chimera
Sequences file
Abundance file
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
out_abundance_count (tabular)
summary_file (html)



11: FROGS Remove chimera: report.html

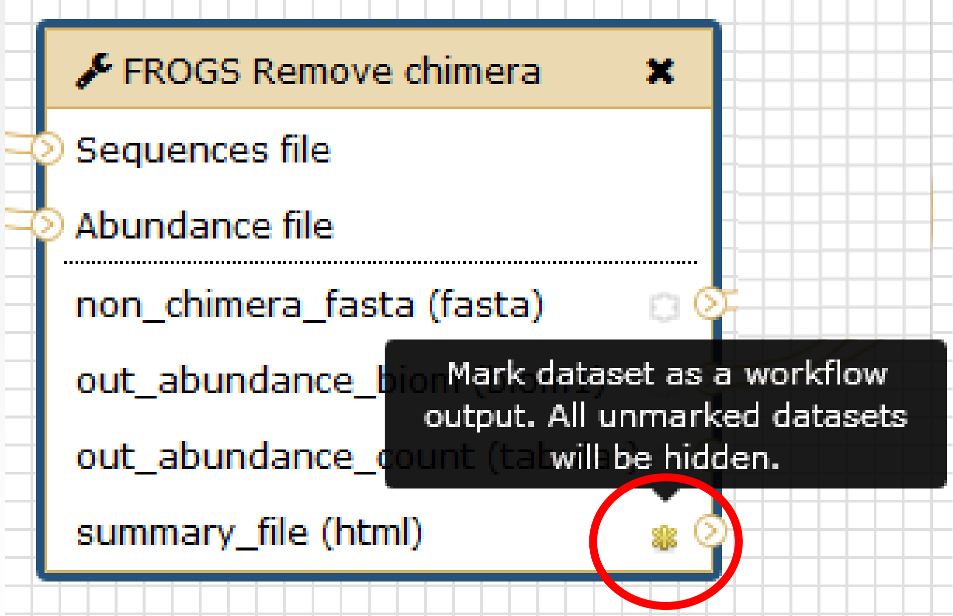
10: FROGS Remove chimera: non_chimera_abundance.biom

9: FROGS Remove chimera: non_chimera.fasta

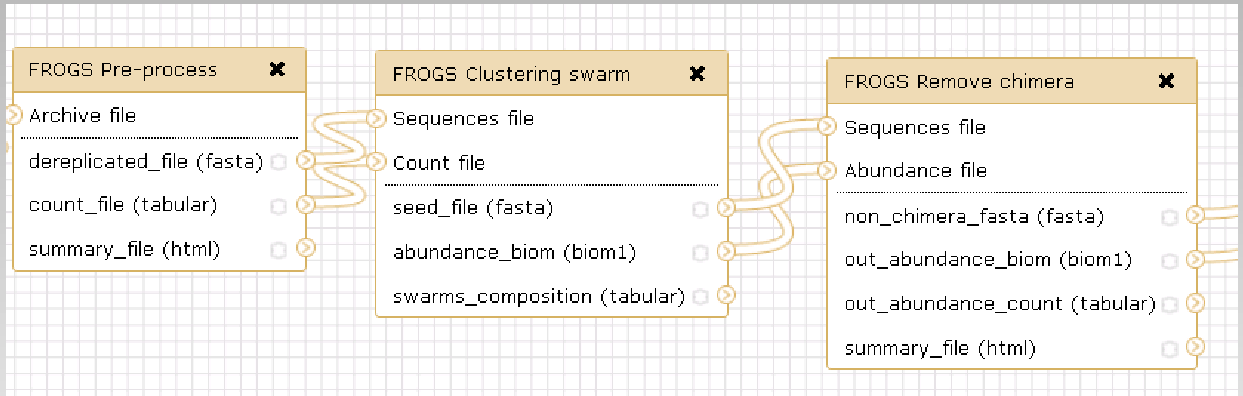


For each tool, think to:

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

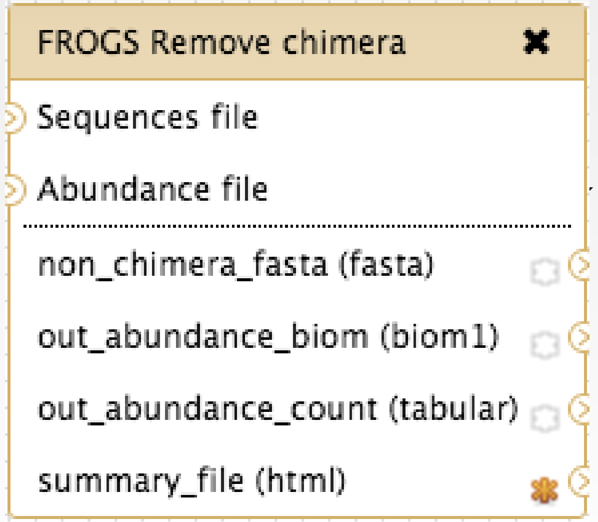

FROGS Remove chimera
Sequences file
Abundance file
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
out_abundance_count (tabular)
summary_file (html)

Mark dataset as a workflow output. All unmarked datasets will be hidden.

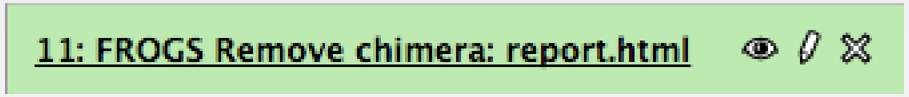


For each tool, think to:

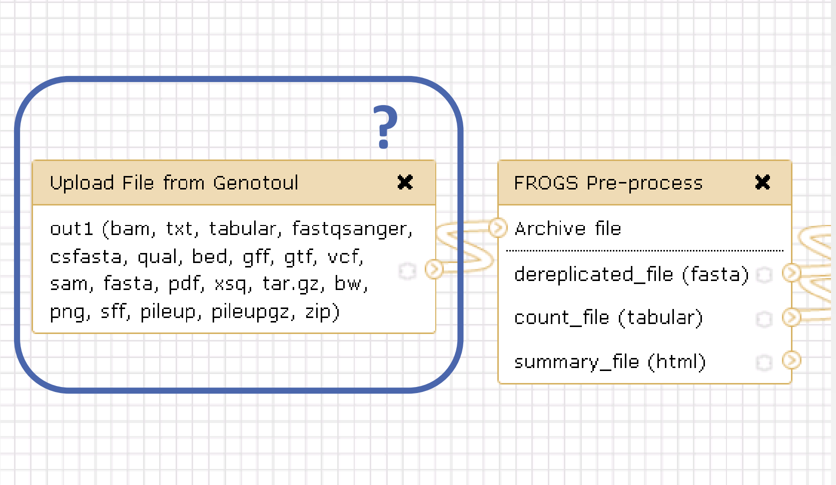
1. Fixe parameter ?
2. Rename output files
3. Hide intermediate files



FROGS Remove chimera
Sequences file
Abundance file
non_chimera_fasta (fasta)
out_abundance_biom (biom1)
out_abundance_count (tabular)
summary_file (html)



Could you integrate « upload file » in the workflow ?



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

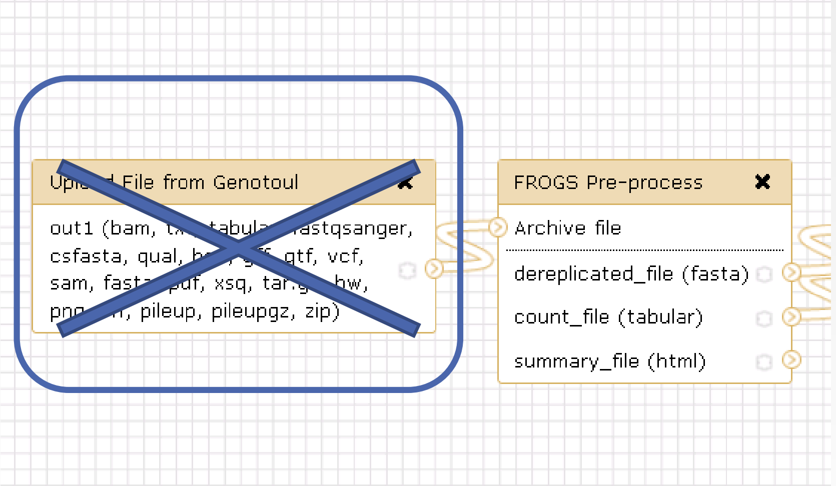
dereplicated_file (fasta)

count_file (tabular)

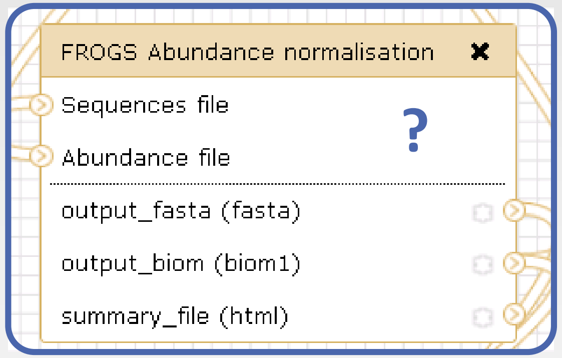
summary_file (html)

Could you integrate « upload file » in the workflow ?

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


Upl... File from Genotoul
out1 (bam, t... tabula... astqsanger, csfasta, qual, b... g...f, gtf, vcf, sam, fasta, ...pdf, xsq, tar.g... bw, png, ...n, pileup, pileupgz, zip)
FROGS Pre-process
Archive file
dereplicated_file (fasta)
count_file (tabular)
summary_file (html)

Could you integrate « Normalisation tool » in the workflow ?



FROGS Abundance normalisation ✖

Sequences file

Abundance file

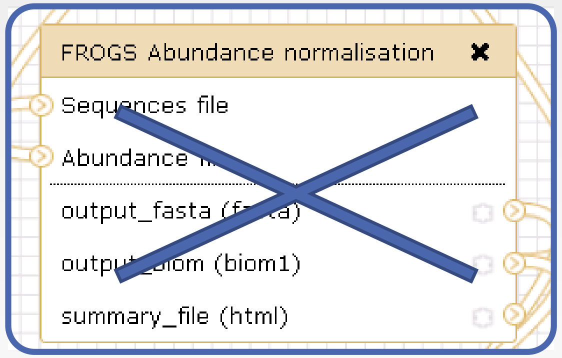
output_fasta (fasta)

output_biom (biom1)

summary_file (html)

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.



FROGS Abundance normalisation

Sequences file

Abundance fi

output_fasta (fasta)

output_biom (biom1)

summary_file (html)

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