



# Training on Galaxy: Metabarcoding

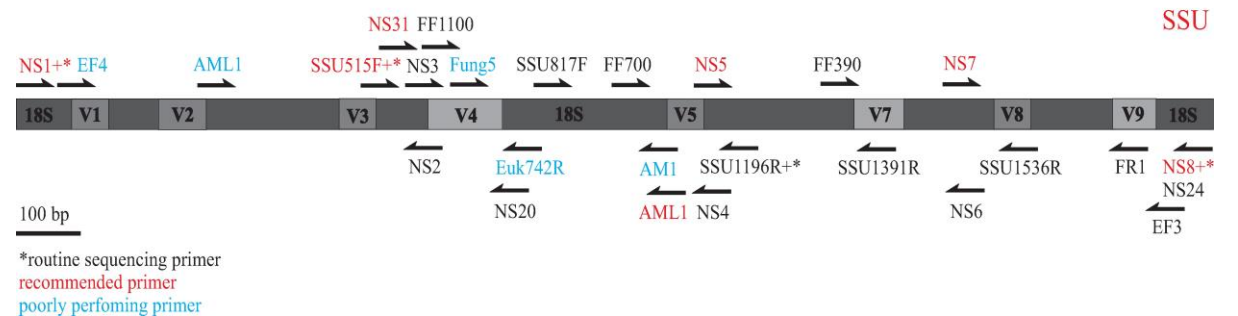
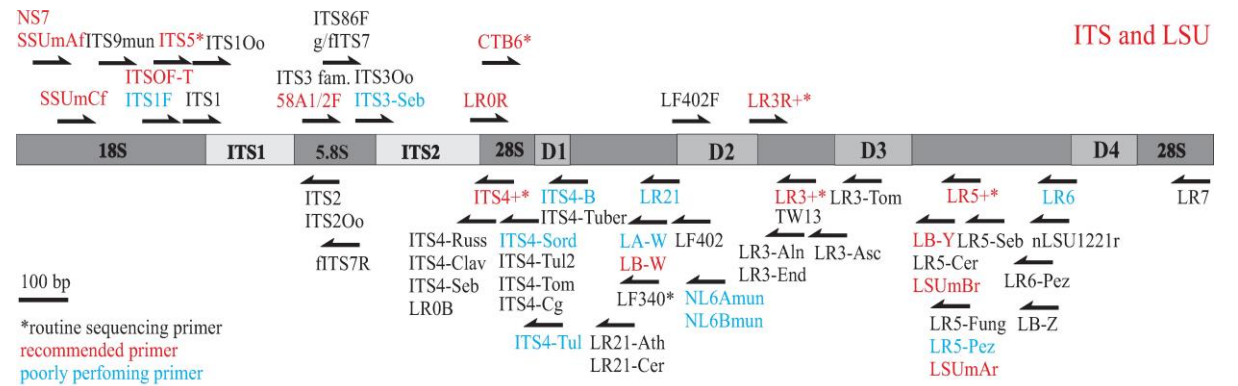
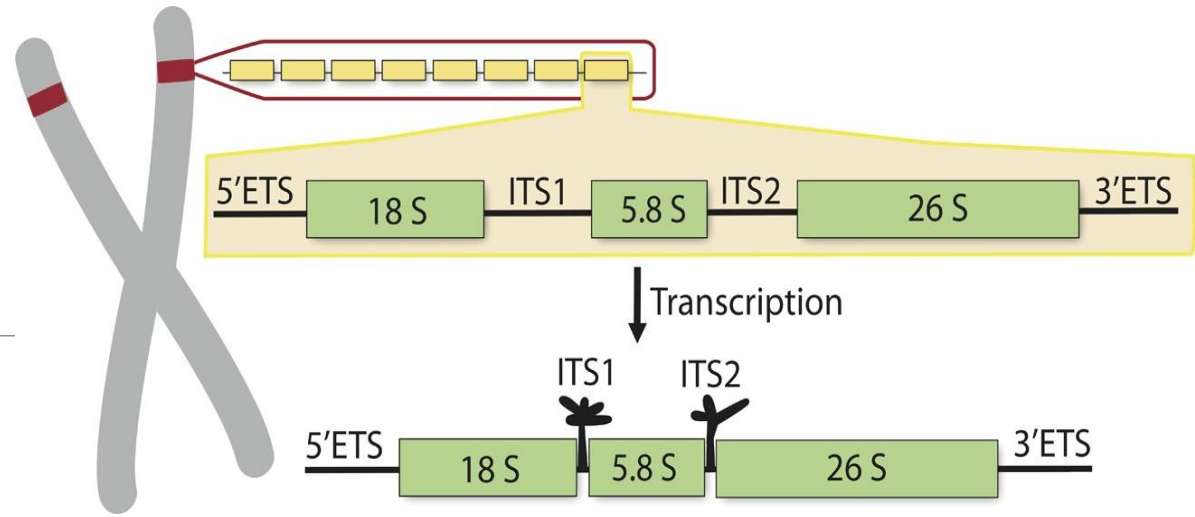
October 2024 - Webinar

## FROGS Practice on ITS data and Workflow creation

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

# What is a ITS ?

ITS: Internal Transcribed Spacer



# 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 at 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 from manipulated organic soil (MOS network)



While in the past forest biomass exports concerned only trunks, these exports recently increased and now concern also the branches and smaller parts that were previously left on the ground (for pellet production).

The [MOS network](#) (18 sites in France) was designed to reveal the long-term effects of intense biomass exports on soil fertility and biodiversity. Different treatment of biomass export are applied with or without supplementation of nutrients.

The aim is to analyse the **impact** of these new forestry practices on **soil microbiota** and **tree health**.

The present dataset concerned one of the site (Champenoux) after 5 years of total Organic Matter removal (OMR treatment : all the organic matter on the ground including leaves was removed), **with** our **without nitrogen supplementation**.

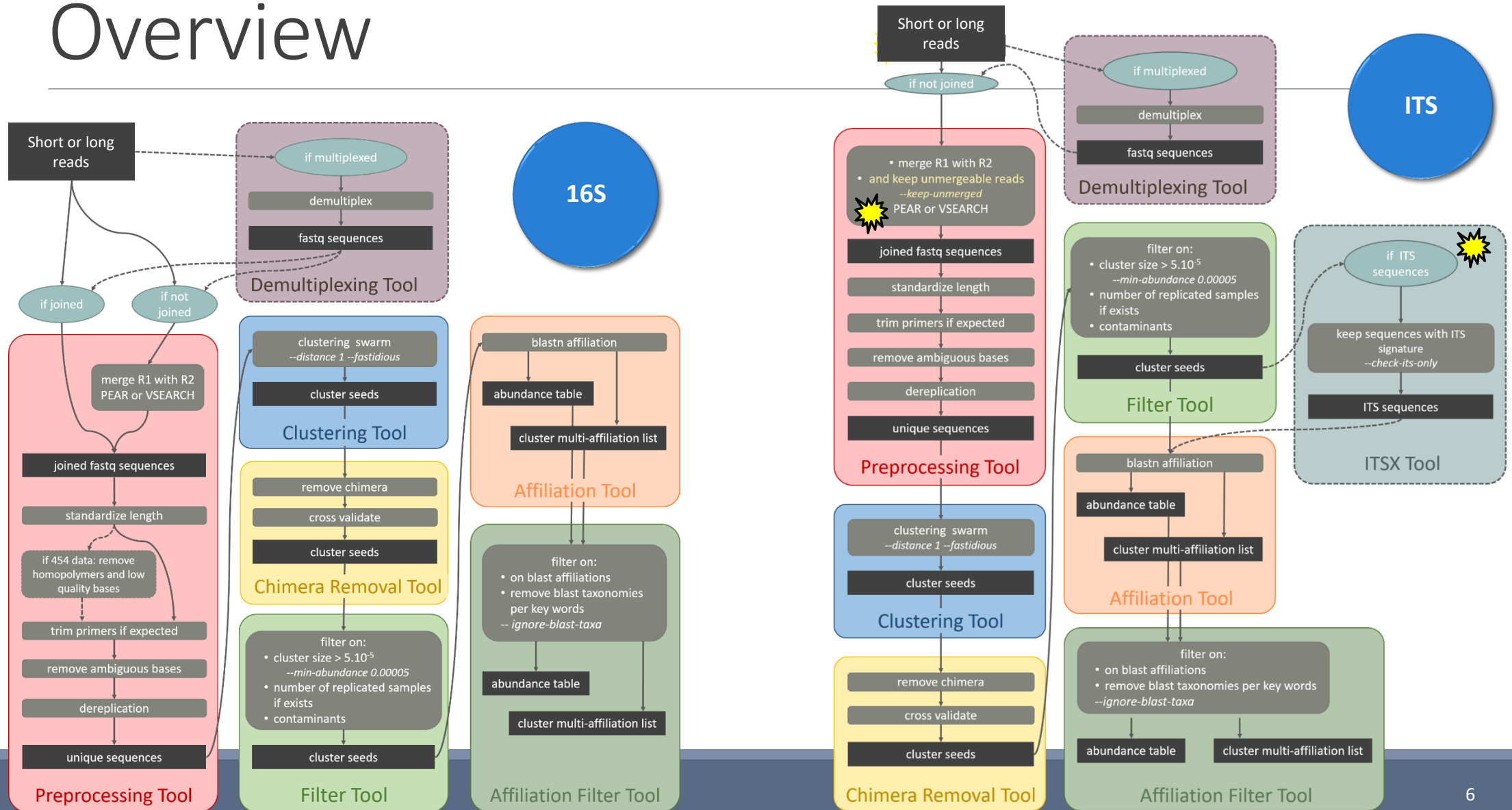
- 5 replicates Control x 2 treatments, 5 replicates OMR x 2 treatments
- DNA is extracted and **ITS1** is sequenced
- 2 x 250 bp Illumina MiSeq
- Primer 5': CTTGGTCATTTAGAGGAAGTAA
- Primer 3': GCATCGATGAAGAACGCAGC



# Metadata for these samples

Samples	kept	Replicas	Incubation	Nitrogen	Forest_management	Quality	Treatment
Ph203	79.76	3	T4	Nitrogen_supplementation	Control	Low degradability	Control_with_N
Ph212	77.64	2	T4	Nitrogen_supplementation	Control	Low degradability	Control_with_N
Ph217	80.26	5	T4	Nitrogen_supplementation	Control	Low degradability	Control_with_N
Ph222	78.65	1	T4	Nitrogen_supplementation	Control	Low degradability	Control_with_N
Ph224	77.18	4	T4	Nitrogen_supplementation	Control	Low degradability	Control_with_N
Ph237	79.68	1	T4	Control	Control	Low degradability	Control
Ph241	78.7	2	T4	Control	Control	Low degradability	Control
Ph243	76.38	4	T4	Control	Control	Low degradability	Control
Ph246	76.37	5	T4	Control	Control	Low degradability	Control
Ph250	77.37	3	T4	Control	Control	Low degradability	Control
Ph407	72.52	3	T4	Nitrogen_supplementation	OMR	Low degradability	OMR_with_N
Ph414	64.98	4	T4	Nitrogen_supplementation	OMR	Low degradability	OMR_with_N
Ph415	78.13	2	T4	Nitrogen_supplementation	OMR	Low degradability	OMR_with_N
Ph417	71.17	1	T4	Nitrogen_supplementation	OMR	Low degradability	OMR_with_N
Ph423	75.2	5	T4	Nitrogen_supplementation	OMR	Low degradability	OMR_with_N
Ph428	73.48	2	T4	Control	OMR	Low degradability	OMR
Ph433	73.21	5	T4	Control	OMR	Low degradability	OMR
Ph434	74.01	3	T4	Control	OMR	Low degradability	OMR
Ph439	74.15	1	T4	Control	OMR	Low degradability	OMR
Ph449	73.77	4	T4	Control	OMR	Low degradability	OMR

# Overview





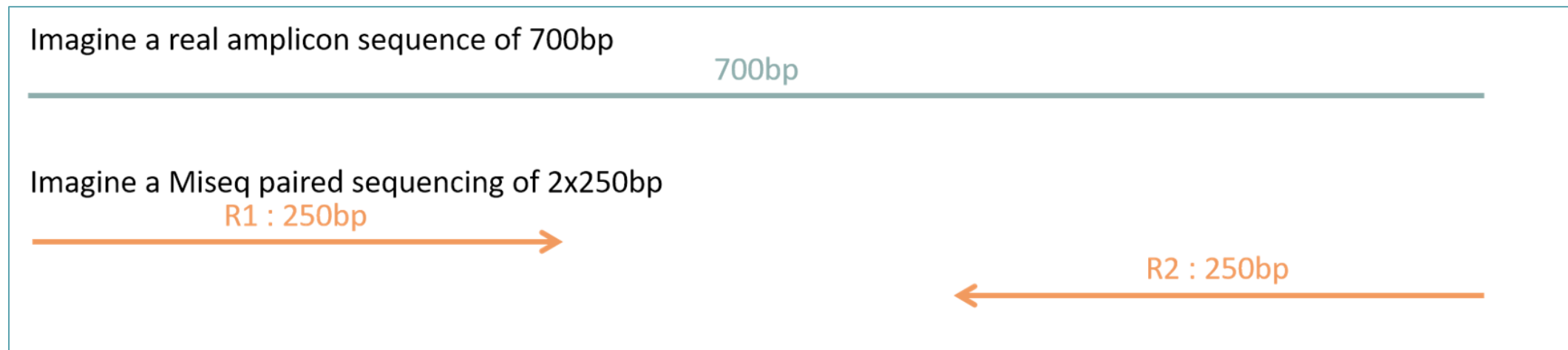
# FROGS Pipeline for ITS





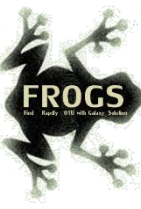
Problematic:  
some ITS reads (Miseq sequencing) are non-overlapping  
sequences

---

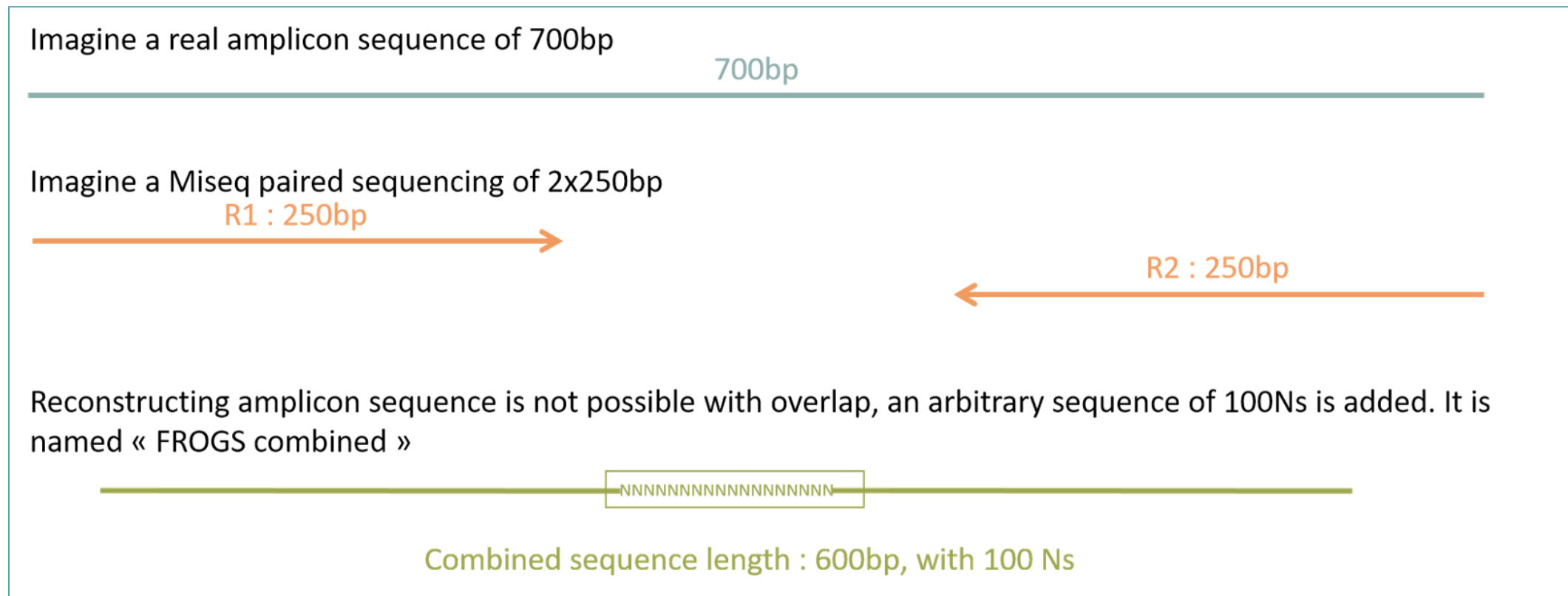



Consequence: during bioinformatics process, these reads are lost and underlying organisms will be never represented in the abundance table.





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





**FROGS\_0 Demultiplex reads** Attribute reads to samples in function of inner barcode  
**FROGS\_1 Pre-process** merging, denoising and dereplication  
**FROGS\_2 Clustering swarm** Single-linkage clustering on sequences  
**FROGS\_Cluster\_Stat** Process some metrics on clusters  
**FROGS\_3 Remove chimera** Remove PCR chimera in each sample  
**FROGS\_4 Cluster filters** Filters clusters on several criteria.  
**FROGS ITSx** Extract the highly variable ITS1 and ITS2 subregions from ITS sequences  
**FROGS\_5 Taxonomic affiliation** Taxonomic affiliation of each ASV's seed by RDPtools and BLAST  
**FROGS\_6 Affiliation\_Stat** Process some metrics on taxonomies

### Basic tools

**FROGSSTAT Phyloseq Import Data** from 3 files: biomfile, samplefile, treefile  
**FROGSSTAT Phyloseq Composition Visualisation** with bar plot and composition plot  
**FROGSSTAT Phyloseq Alpha Diversity** with richness plot  
**FROGSSTAT Phyloseq Beta Diversity** distance matrix  
**FROGSSTAT Phyloseq Sample Clustering** of samples using different linkage methods  
**FROGSSTAT Phyloseq Structure Visualisation** with heatmap plot and ordination plot  
**FROGSSTAT Phyloseq Multivariate Analysis Of Variance** perform Multivariate Analysis of Variance (MANOVA)  
**FROGSSTAT DESeq2 Preprocess** import a Phyloseq object and prepare it for DESeq2 differential abundance analysis a  
**FROGSSTAT DESeq2 Visualisation** extract and visualise differentially abundant ASVs or functions

### Statistics tools

**FROGS Tree** Reconstruction of phylogenetic tree  
**FROGS Affiliation Filters** Filters ASVs on several affiliation criteria  
**FROGS Affiliation postprocess** Aggregates ASVs based on alignment metrics  
**FROGS Abundance normalisation** Normalise ASV abundance.

### Optional basic tools

**FROGS BIOM to std BIOM** Converts a FROGS BIOM in fully compatible BIOM  
**FROGS TSV\_to\_BIOM** Converts a TSV file in a BIOM file 1  
**FROGS BIOM to TSV** Converts a BIOM file in TSV file

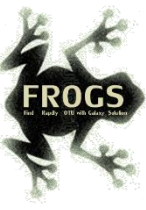
### Utilities tools

**FROGSFUNC\_1\_placeseqs\_and\_copynumbers** Places ASVs into a reference phylogenetic tree.  
**FROGSFUNC\_2\_functions** Calculates functions abundances in each sample.  
**FROGSFUNC\_3\_pathways** Calculates pathway abundances in each sample.

### Functional inference tools



28 tools  
in total

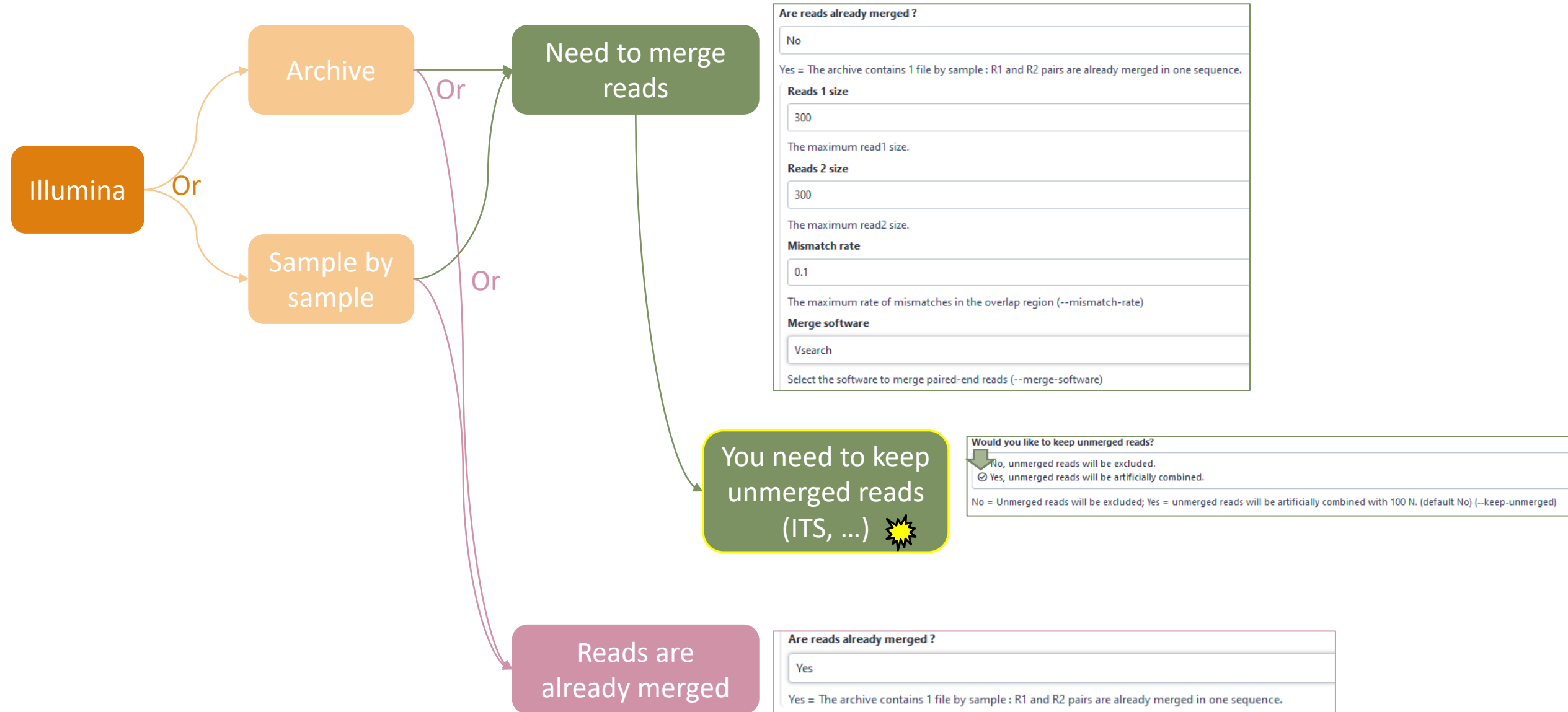


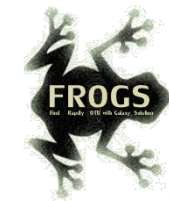
---

# Pre-process tool

---

# For short reads from illumina



**Sequencer**

Illumina

Select the sequencing technology used to produce the sequences.

**Input type**

TAR Archive

Samples files can be provided in a single TAR archive or sample by sample (with one or two files each).

**TAR archive file** 1: ITS\_fast.tar.gz

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

**Are reads already merged ?**

No

Yes = The archive contains 1 file by sample : R1 and R2 pairs are already merged in one sequence.

**Reads 1 size**

250

The maximum read1 size.

**Reads 2 size**

250

The maximum read2 size.

**Mismatch rate**

0.1

The maximum rate of mismatches in the overlap region (--mismatch-rate)

**Merge software**

Vsearch

Select the software to merge paired-end reads (--merge-software)

**Would you like to keep unmerged reads?**

- No, unmerged reads will be excluded.
- Yes, unmerged reads will be artificially combined.

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



To keep FROGS combined sequences, choose YES



**Minimum amplicon size**

180

The minimum size of the amplicons (with primers) (--min-amplicon-size)

**Maximum amplicon size**

490

The maximum size of the amplicons (with primers) (--max-amplicon-size)

**Do the sequences have PCR primers?** Yes No**5' primer**

CTTGGTCATTTAGAGGAAGTAA

The 5' primer sequence (wildcards are accepted). This primer must be written in 5' to 3' orientation (see details in 'Primers parameters' help section) (--five-prim-primer)

**3' primer**

GCATCGATGAAGAACGCAGC

The 3' primer sequence (wildcards are accepted). This primer must be written in 5' to 3' orientation (see details in 'Primers parameters' help section) (--three-prim-primer)

Primer 5': CTTGGTCATTTAGAGGAAGTAA

Primer 3': GCATCGATGAAGAACGCAGC

# Exercise

---

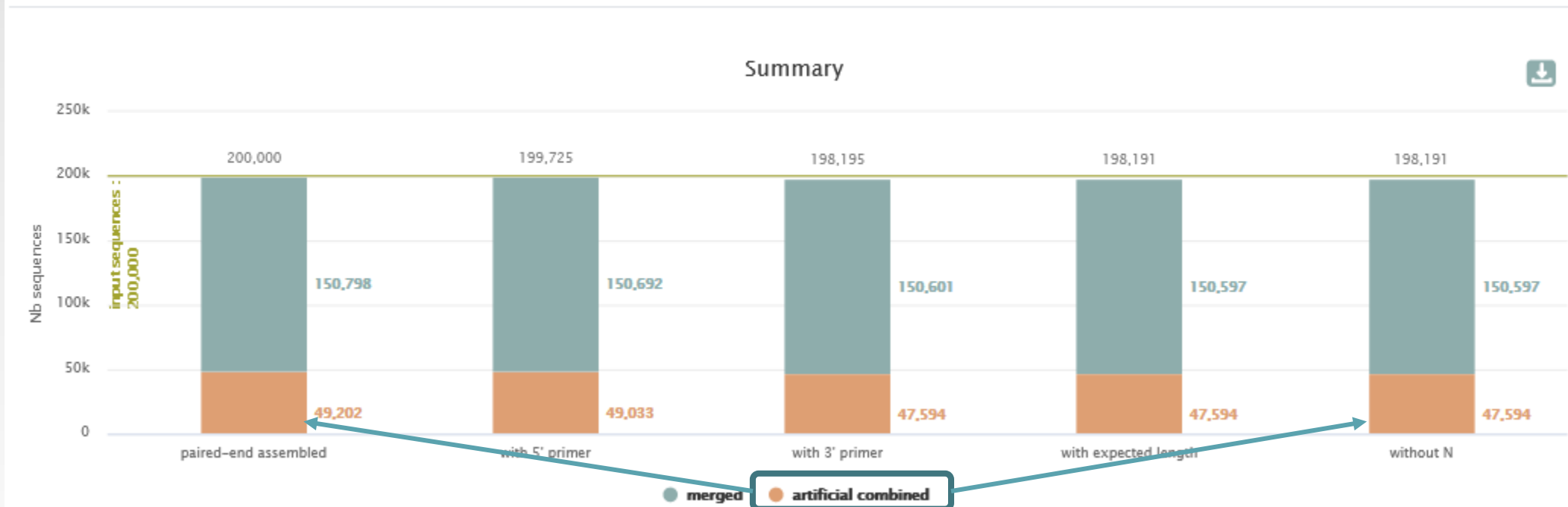
Go to « [ITS](#) » history

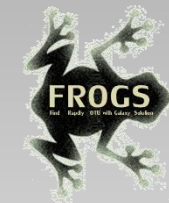
Launch the pre-process tool on this data set

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

# Explore Preprocess report.html

## Preprocess summary





2 tables:

## Details on merged sequences

Show 10 entries

CSV

Search: 

Samples	before process	% kept	paired-end assembled	with 5' primer	with 3' primer	with expected length	without N
Ph203	10,000	79.42	7,954	7,948	7,942	7,942	7,942
Ph212	10,000	78.28	7,837	7,832	7,828	7,828	7,828
Ph217	10,000	80.48	8,061	8,052	8,048	8,048	8,048
Ph222	10,000	78.34	7,839	7,835	7,834	7,834	7,834

### Own tag for combined sequences

```

>M01328:521:000000000-KRPT:1:1103:15714:11240;size=6 1:N:0:238
AAGTCGTAACAAGGTAACCGTAGGTGAACCTGCGGTTGGATCATTAAAAATTTATGAGTTCCGTTGAC
>M01328:521:000000000-KRPT:1:2102:7650:15129;size=1 1:N:0:239
AAGTCGTAACAAGGTAACCGTAGGTGAACCTGCGGTTGGATCATTAAAAATTTATGAGTTCCGTTGAC
>M01328:521:000000000-KRPT:1:1112:8680:15899;size=1 1:N:0:202
AAGTCGTAACAAGGTTATCGTTGCACTAGCTAAGCCCTATTGCAAGCCTTTCCAGCGACTGAAAATAAC
>M01328:521:000000000-KRPT:1:1111:21036:16514_FROGS_combined;size=1
AAGTCGTAACAAGGTTTCGGTAGGTGAACCTGCGGTAAGGATCATTACAAGTTCTGTAGGTCTGTCGCA
>M01328:521:000000000-KRPT:1:1106:19343:17084_FROGS_combined;size=1
AAGTCGTAACAAGGTTTCGGTAGGTGAACCTGCGGTAAGGATCATTACAAGTTCTGTAGGTCTGTCGCA

```

## Details on artificial combined sequences

Show 10 entries

CSV

Search: 

Samples	before process	% kept	paired-end assembled	with 5' primer	with 3' primer	with expected length	without N
Ph203	10,000	19.68	2,046	2,038	1,968	1,968	1,968
Ph212	10,000	20.65	2,163	2,154	2,065	2,065	2,065
Ph217	10,000	18.63	1,939	1,928	1,863	1,863	1,863
Ph222	10,000	20.79	2,161	2,155	2,079	2,079	2,079

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. FROGS rejects these cases.



**FROGS\_0 Demultiplex reads** Attribute reads to samples in function of inner barcode  
**FROGS\_1 Pre-process** merging, denoising and dereplication  
**FROGS\_2 Clustering swarm** Single-linkage clustering on sequences  
**FROGS\_Cluster\_Stat** Process some metrics on clusters  
**FROGS\_3 Remove chimera** Remove PCR chimera in each sample  
**FROGS\_4 Cluster filters** Filters clusters on several criteria.  
**FROGS ITSx** Extract the highly variable ITS1 and ITS2 subregions from ITS sequences  
**FROGS\_5 Taxonomic affiliation** Taxonomic affiliation of each ASV's seed by RDPtools and BLAST  
**FROGS\_6 Affiliation\_Stat** Process some metrics on taxonomies

### Basic tools

**FROGSSTAT Phyloseq Import Data** from 3 files: biomfile, samplefile, treefile  
**FROGSSTAT Phyloseq Composition Visualisation** with bar plot and composition plot  
**FROGSSTAT Phyloseq Alpha Diversity** with richness plot  
**FROGSSTAT Phyloseq Beta Diversity** distance matrix  
**FROGSSTAT Phyloseq Sample Clustering** of samples using different linkage methods  
**FROGSSTAT Phyloseq Structure Visualisation** with heatmap plot and ordination plot  
**FROGSSTAT Phyloseq Multivariate Analysis Of Variance** perform Multivariate Analysis of Variance (MANOVA)  
**FROGSSTAT DESeq2 Preprocess** import a Phyloseq object and prepare it for DESeq2 differential abundance analysis a  
**FROGSSTAT DESeq2 Visualisation** extract and visualise differentially abundant ASVs or functions

### Statistics tools

**FROGS Tree** Reconstruction of phylogenetic tree  
**FROGS Affiliation Filters** Filters ASVs on several affiliation criteria  
**FROGS Affiliation postprocess** Aggregates ASVs based on alignment metrics  
**FROGS Abundance normalisation** Normalise ASV abundance.

### Optional basic tools

**FROGS BIOM to std BIOM** Converts a FROGS BIOM in fully compatible BIOM  
**FROGS TSV\_to\_BIOM** Converts a TSV file in a BIOM file 1  
**FROGS BIOM to TSV** Converts a BIOM file in TSV file

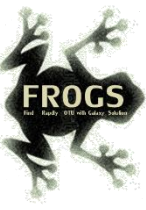
### Utilities tools

**FROGSFUNC\_1\_placeseqs\_and\_copynumbers** Places ASVs into a reference phylogenetic tree.  
**FROGSFUNC\_2\_functions** Calculates functions abundances in each sample.  
**FROGSFUNC\_3\_pathways** Calculates pathway abundances in each sample.

### Functional inference tools

28 tools  
in total





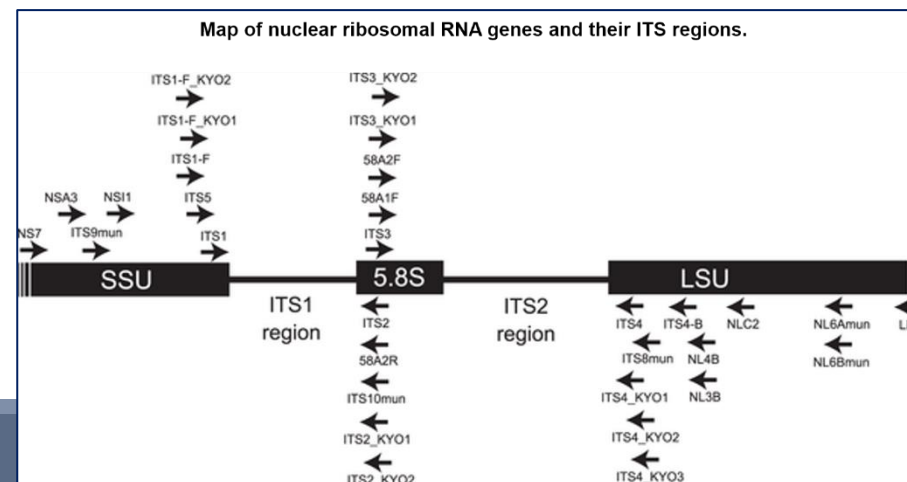
---

# ITSx tools

---

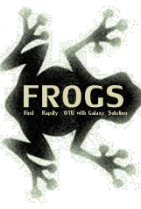
# What is the purpose of the ITSx tool?

- 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).

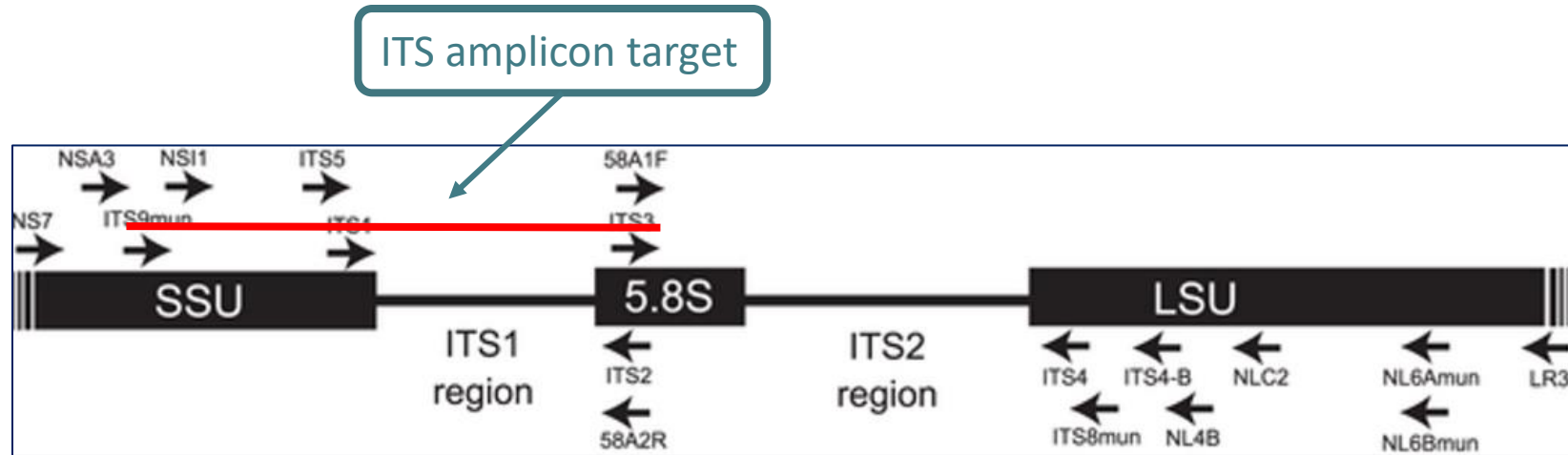



Bengtsson-Palme, J., et al. (2013), Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. *Methods Ecol Evol*, 4: 914-919.


<https://doi.org/10.1111/2041-210X.12073>



# What is the purpose of the ITSx tool?

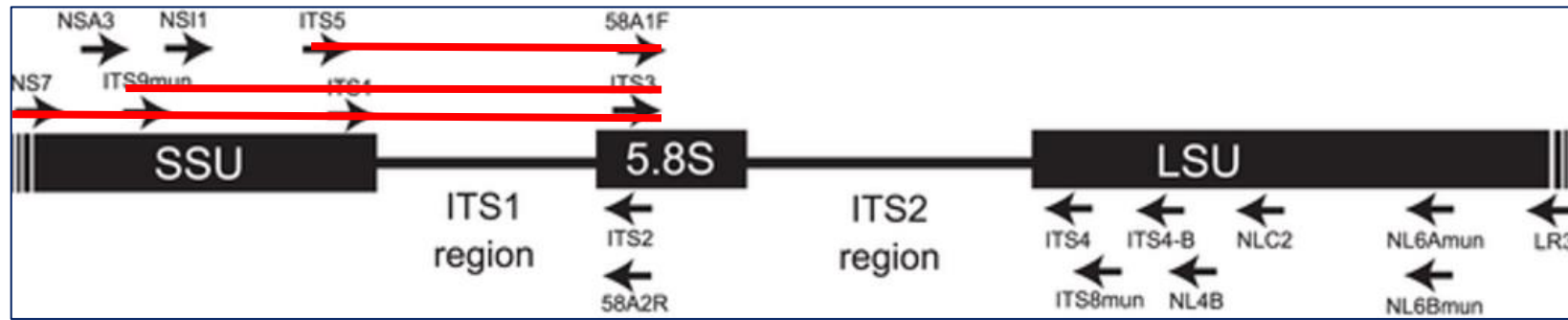


1<sup>st</sup> case: choose to trim  
ITS1 is well detected  
SSU part and 5.8S part are trimmed  
Result: 

2<sup>nd</sup> 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

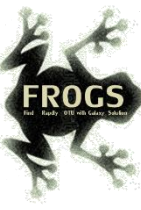




# When should we use ITSx ?







Sequence file

16: FROGS\_4 Cluster filters: clusterFilters\_sequences.fasta

The sequence file to filter (format: FASTA).

Abundance file

15: FROGS\_4 Cluster filters: clusterFilters\_abundance.biom

The abundance file to filter (format: BIOM)

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

No, keep conserved regions  
 Yes, trim conserved regions

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

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 (--organism-groups)

Email notification

No

Send an email notification when the job completes.

By default, the ITSs are kept in their entirety.

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

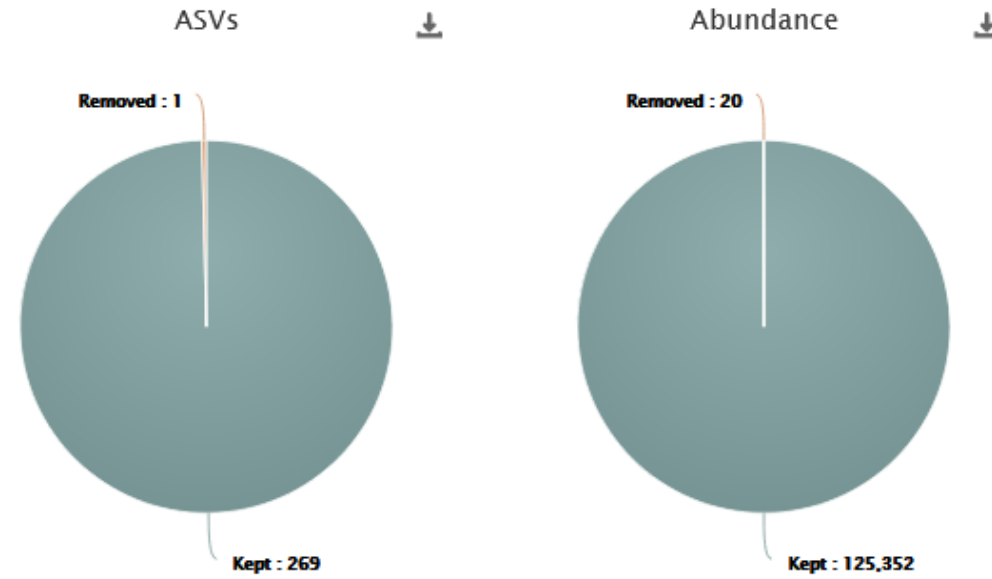
# Careful !



- The ITSx step is time consuming and has to be done on minimum of clusters.
  1. Preprocess step,
  2. Clustering step,
  3. Chimera removing step,
  4. Filter on ASVs abundances and replicates step,
  5. ITSx



## Filters (ITSx) summary



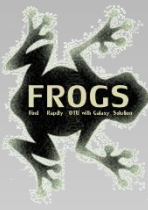
## Filters (ITSx) by samples

Show  entries  CSV

Search:

### ASVs removed by sample

Sample name	Initial	Kept	Initial abundance	Kept abundance
Ph203	105	105	7,065	7,065
Ph212	65	65	7,474	7,474
Ph217	89	89	5,990	5,990



**FROGS\_0 Demultiplex reads** Attribute reads to samples in function of inner barcode  
**FROGS\_1 Pre-process** merging, denoising and dereplication  
**FROGS\_2 Clustering swarm** Single-linkage clustering on sequences  
**FROGS\_Cluster\_Stat** Process some metrics on clusters  
**FROGS\_3 Remove chimera** Remove PCR chimera in each sample  
**FROGS\_4 Cluster filters** Filters clusters on several criteria.  
**FROGS ITSx** Extract the highly variable ITS1 and ITS2 subregions from ITS sequences  
**FROGS\_5 Taxonomic affiliation** Taxonomic affiliation of each ASV's seed by RDPtools and BLAST  
**FROGS\_6 Affiliation\_Stat** Process some metrics on taxonomies

### Basic tools

**FROGSSTAT Phyloseq Import Data** from 3 files: biomfile, samplefile, treefile  
**FROGSSTAT Phyloseq Composition Visualisation** with bar plot and composition plot  
**FROGSSTAT Phyloseq Alpha Diversity** with richness plot  
**FROGSSTAT Phyloseq Beta Diversity** distance matrix  
**FROGSSTAT Phyloseq Sample Clustering** of samples using different linkage methods  
**FROGSSTAT Phyloseq Structure Visualisation** with heatmap plot and ordination plot  
**FROGSSTAT Phyloseq Multivariate Analysis Of Variance** perform Multivariate Analysis of Variance (MANOVA)  
**FROGSSTAT DESeq2 Preprocess** import a Phyloseq object and prepare it for DESeq2 differential abundance analysis a  
**FROGSSTAT DESeq2 Visualisation** extract and visualise differentially abundant ASVs or functions

### Statistics tools

**FROGS Tree** Reconstruction of phylogenetic tree  
**FROGS Affiliation Filters** Filters ASVs on several affiliation criteria  
**FROGS Affiliation postprocess** Aggregates ASVs based on alignment metrics  
**FROGS Abundance normalisation** Normalise ASV abundance.

### Optional basic tools

Not specific for ITS  
but often useful

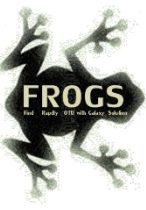
**FROGS BIOM to TSV** Converts a BIOM file in TSV file  
**FROGS TSV to BIOM** Converts a TSV file in a BIOM file 1

### Utilities tools

**FROGSFUNC\_1\_placeseqs\_and\_copynumbers** Places ASVs into a reference phylogenetic tree.  
**FROGSFUNC\_2\_functions** Calculates functions abundances in each sample.  
**FROGSFUNC\_3\_pathways** Calculates pathway abundances in each sample.

### Functional inference tools

28 tools  
in total



---

# Affiliation Post-process

---



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

---

This tool allows **grouping ASVs 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 ?

---

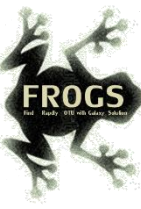
## In consequence:

The different affiliations involved in multi-affiliation are merged.

The abundances are added together.

It is the most abundant ASV seed that is kept.





Sequence file

21: FROGS ITSx: ITS\_sequence.fasta

The sequence file to filter (format: FASTA).

Abundance file

25: FROGS\_5 Taxonomic affiliation: affiliation\_abundance.biom

The abundance file to filter (format: BIOM)

Is this an amplicon hyper variable in length?

- No
- Yes

Yes, we have combined sequences

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

Using reference database

UNITE 8.2 ITS1

same database used for taxonomic affiliation

Select reference from the list (--reference)

Minimum identity for aggregation

99

ASVs will be aggregated if they share the same taxonomy with at least X% identity (--identity)

Minimum coverage for aggregation

99

ASVs will be aggregated if they share the same taxonomy with at least X% alignment coverage (--coverage)

Here, we wanted to aggregate ASVs only if they are very closed

Email notification

No

Send an email notification when the job completes.

✓ Execute

FROGS Affiliation postprocess: OTU\_aggregation\_composition.txt  
FROGS Affiliation postprocess: sequences.fasta  
FROGS Affiliation postprocess: affiliation\_abundance.biom

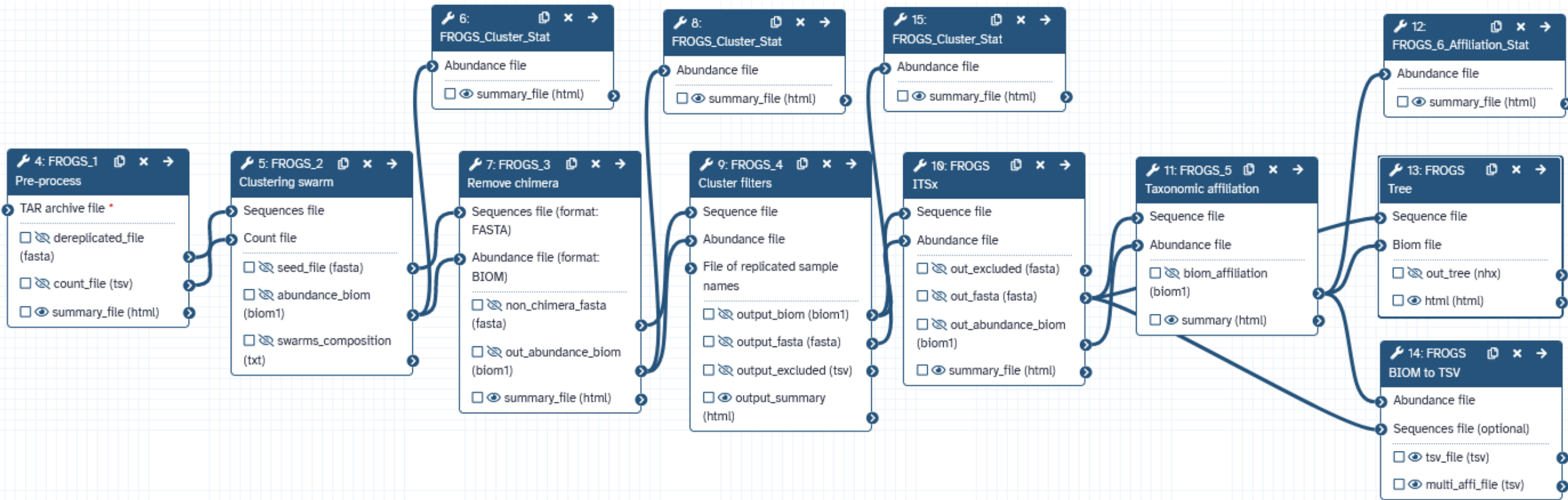
Cluster\_1  
Cluster\_2  
Cluster\_8  
Cluster\_3  
Cluster\_5  
Cluster\_4  
Cluster\_6  
Cluster\_7  
Cluster\_9  
Cluster\_13  
Cluster\_10  
Cluster\_11  
Cluster\_16  
Cluster\_17  
Cluster\_14  
Cluster\_12  
Cluster\_15  
Cluster\_22  
Cluster\_18  
Cluster\_23  
Cluster\_25  
Cluster\_19  
Cluster\_21  
Cluster\_26  
Cluster\_29  
Cluster\_34  
Cluster\_35  
Cluster\_28  
Cluster\_31  
Cluster\_32  
Cluster\_42  
Cluster\_33  
Cluster\_75\_FROGS\_combined Cluster\_121\_FROGS\_combined Cluster\_137\_FROGS\_combined Cluster\_144\_FROGS\_combi



---

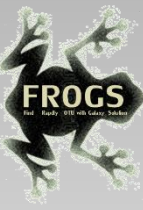
# Workflow creation

---



Workflows are useful for routine analyses

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

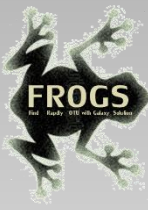


---

# Practice

---

CREATE YOUR OWN WORKFLOW !



# Exercise



New Galaxy server, needed tools/databanks are added on demand

Tools

search tools

Upload Data

BASIC TOOLS

- Monitoring
- Get Data
- Send Data
- Collection Operations
- Lift-Over
- Text Manipulation
- Convert Formats
- Filter and Sort

Search Workflows

+ Create Import

Name	Tags	Updated	Sharing	Bookmarked
16S		a few seconds ago		<input type="checkbox"/>
ITS		2 minutes ago		<input type="checkbox"/>

History

Rechercher des données

ITS

21 shown

252.24 MB

- 21: FROGS Affiliation O TU: report.html
- 20: FROGS Affiliation O TU: affiliation\_abundance.biom
- 19: FROGS ITSx: report.html
- 18: FROGS ITSx: itsx\_abundance.biom

# Exercise

---

Create Workflow

**Name**

3

ITS\_formation

**Annotation**

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

✕ Cancel







✓ Create

4











# Exercise

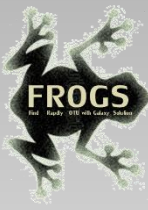
Search Workflows + Create Import

Name	Tags	Updated	Sharing	Bookmarked	
▼ ITS_formation		a few seconds ago		<input type="checkbox"/>	
▼ 16S		3 minutes ago		<input type="checkbox"/>	
▼ ITS		5 minutes ago		<input type="checkbox"/>	

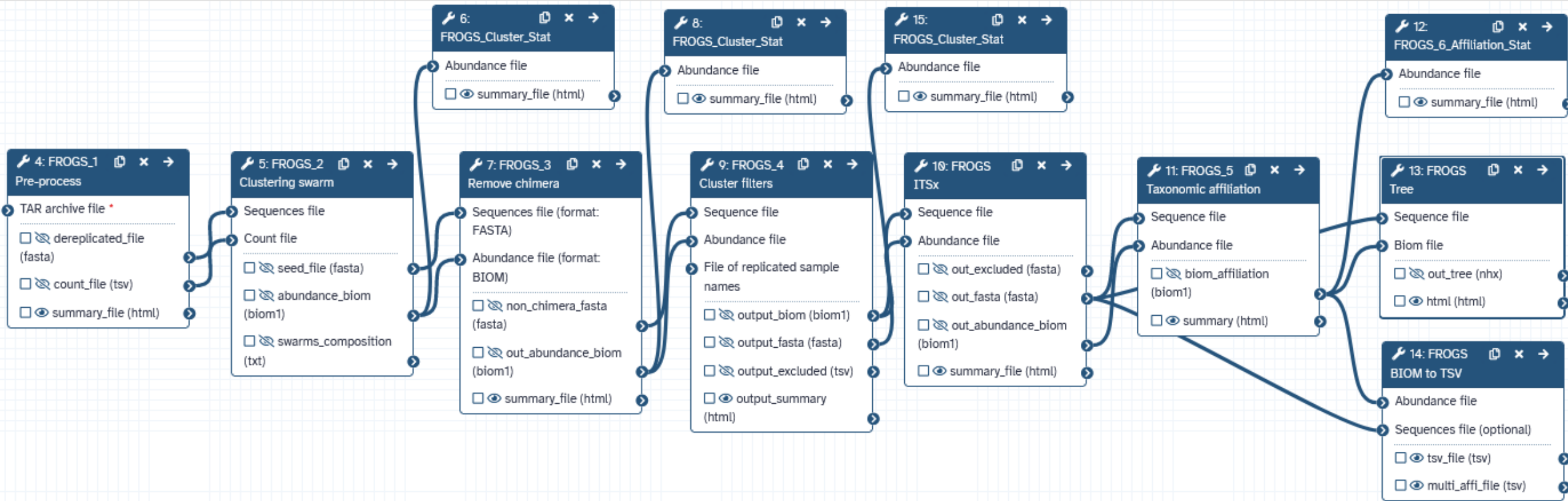
Name Tags

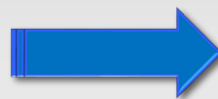
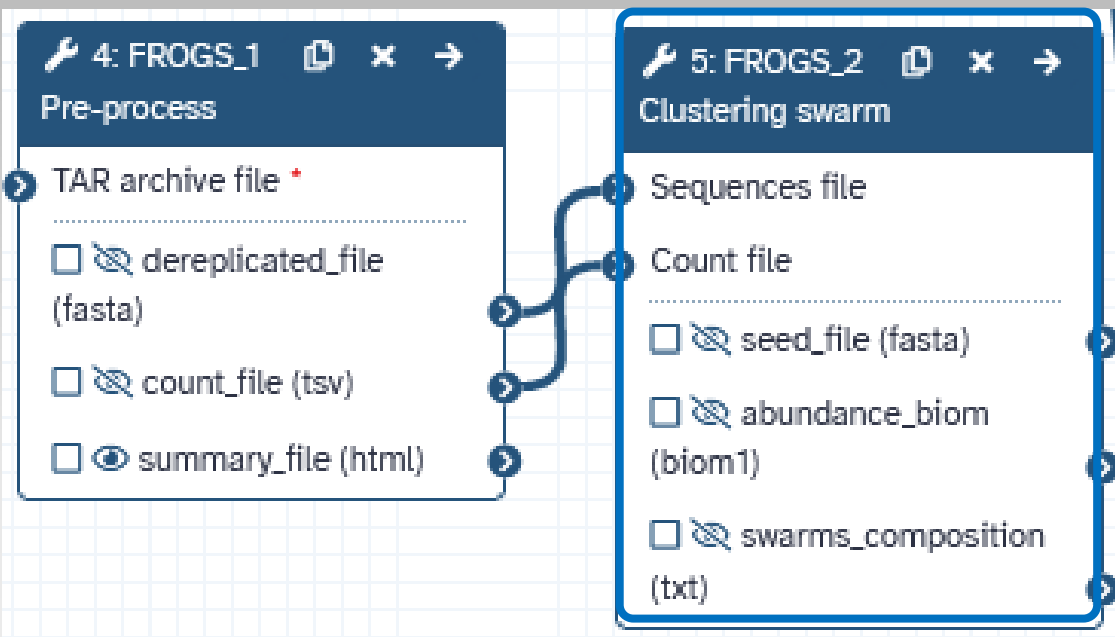
▼ ITS\_formation 

-  Edit
-  Copy
-  Download
-  Rename
-  Share
-  View
-  Delete



# Solution of exercise:





FROGS\_2 Clustering swarm Single-linkage clustering on sequences (Galaxy Version 4.1.0+galaxy1)

**Label**

Add a step label.

**Step Annotation**

Add an annotation or notes to this step. Annotations are available when a workflow is viewed.

**Sequences file**  
Data input 'sequence\_file' (fasta)  
The dereplicated sequences file (format: FASTA)

**Count file**  
Data input 'count\_file' (tabular or tsv)  
It contains the count by sample for each sequence (format: TSV)

**FROGS guidelines version**  
New guidelines from version 3.2

The denoising step before a d3 clustering is no longer recommended since FROGS 3.2, but you can still choose it.

**↔ Aggregation distance clustering**  
1

Maximum number of differences between sequences in each aggregation Swarm step. (recommended d=1) (--distance)

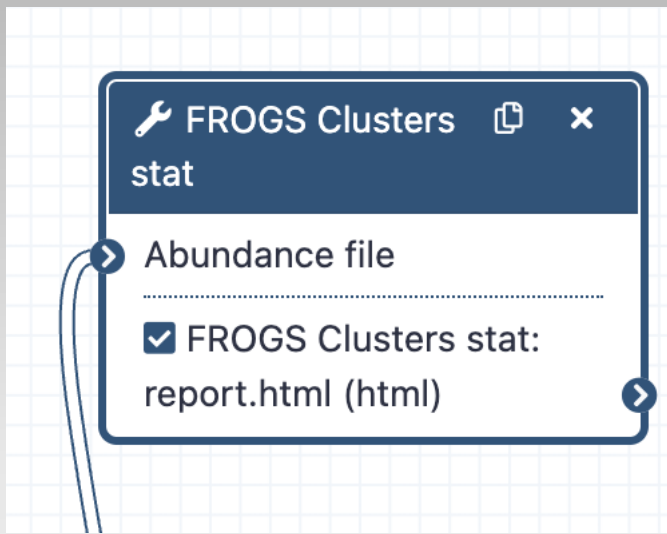
**↔ Refine clustering**

Yes, refine clustering with --fastidious swarm option  
 No, perform clustering without refinement

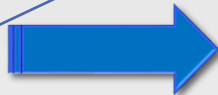
Clustering will be performed with the Swarm --fastidious option. It is recommended and only usable in association with a distance of 1 (default and recommended: Yes) (--fastidious)

For each tool, think to:

1. Set parameters



Configure Output: 'FROGS Clusters stat: report.html'



Configure Output: 'FROGS Clusters stat: report.html'

**Label**

**Don't use 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 input variables are:

- **biom** (Abundance file)

**Change datatype**

Leave unchanged

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

**Add Tags**

This action will set tags for the dataset.

**Remove Tags**

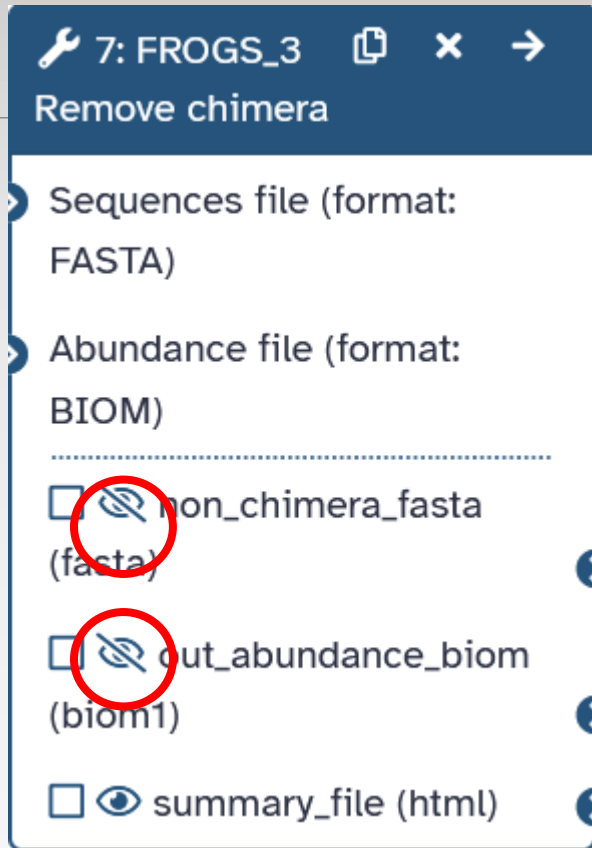
This action will remove tags for the dataset.

Assign columns

For each tool, think to:

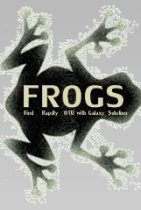
1. Set parameters
2. Rename output files





For each tool, think to:

1. Set parameters
2. Rename output files
3. Hide intermediate files to simplify your history



# Exercise

---

When your workflow is built

1. Run your own workflow with ITS data with :

[ITS fast.tar.gz](#)

[ITS fast metadata.tsv](#)

[ITS fast replicates.tsv](#)

2. Run FROGSSTAT tools