ITS analysis in FROGS ?

Training on Galaxy: Metagenomics
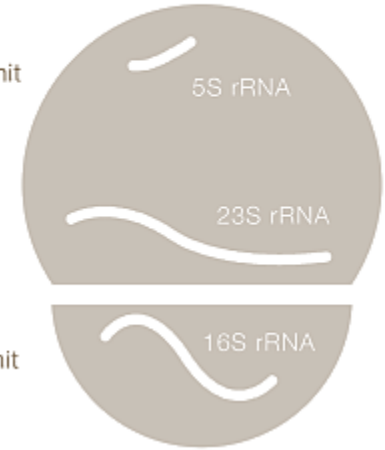
MARCH 2019
MARIA Bernard, OLIVIER Rué, GÉRALDINE Pascal

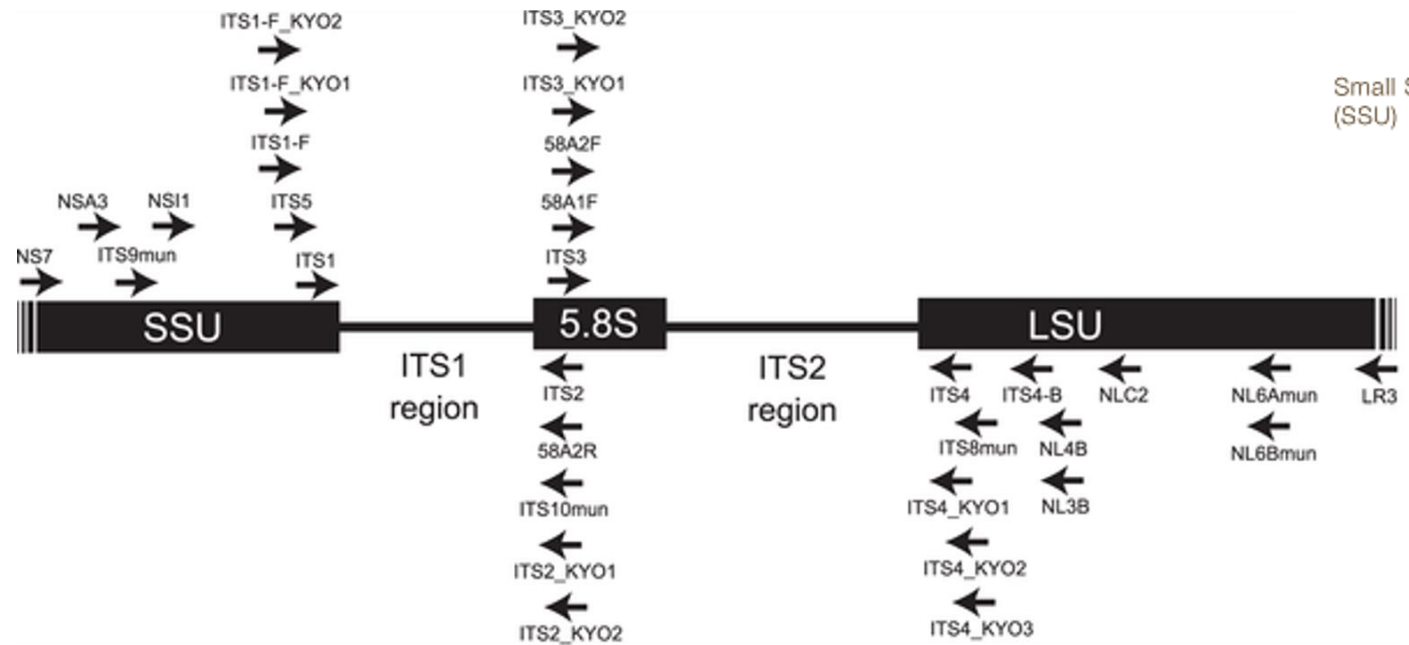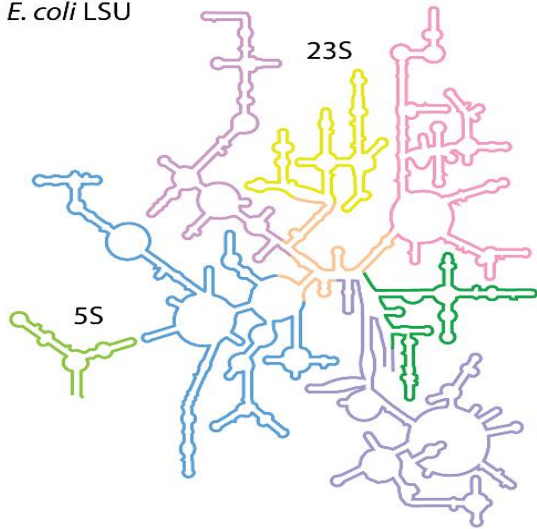# What is a ITS ?

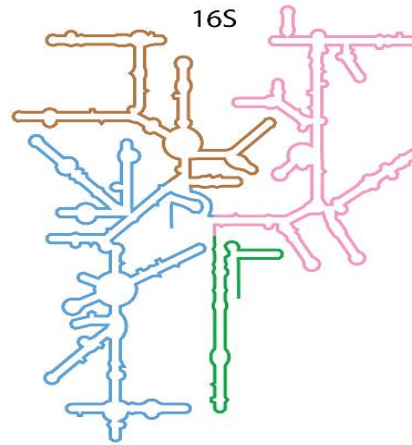**Map of nuclear ribosomal RNA genes and their ITS regions.**

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
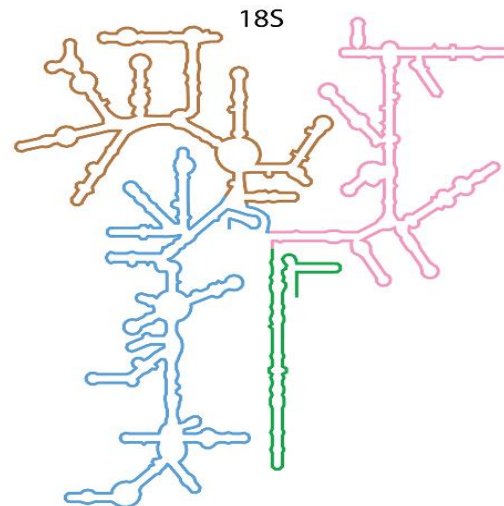
a) *E. coli* LSU

23S

5S

b) *E. coli* SSU

16S

c) *S. cerevisiae* LSU

25S

5S

5.8S

d) *S. cerevisiae* SSU

18S

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

ITS data form METABARFOOD Project metaprogramme MEM
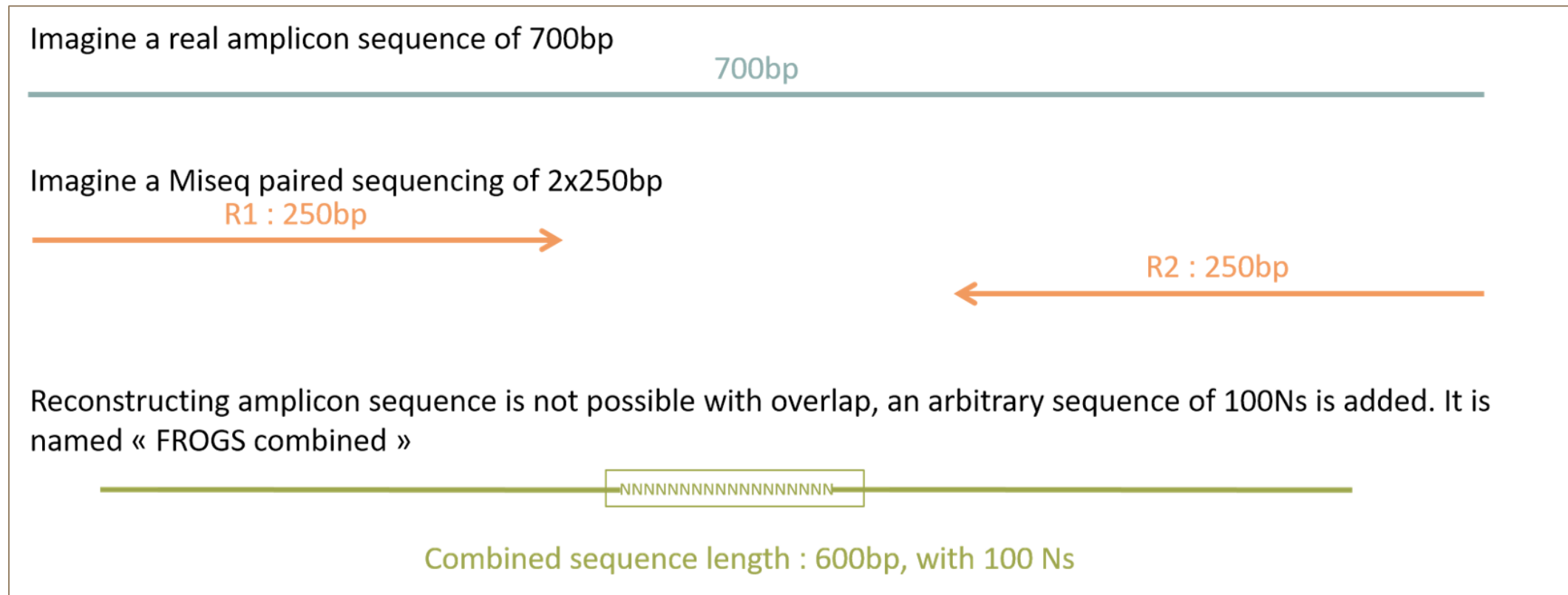

Yeast catalog in food ecosystems

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

n=52 n=63 n=42 n=31 n=25

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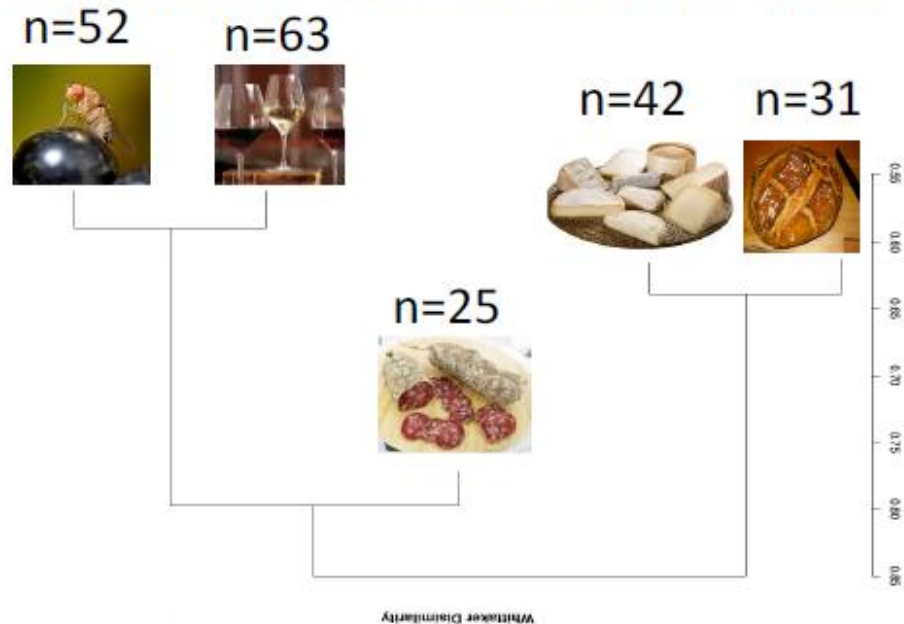
- While metabarcoding is commonly used to describe prokaryotes in the microbiome of many environments, methods for describing micro-eukaryote diversity is lacking and requires better methodology and standardisation.
- One reason is that the universal fungal barcode, the Internal Transcribed Spacer (ITS)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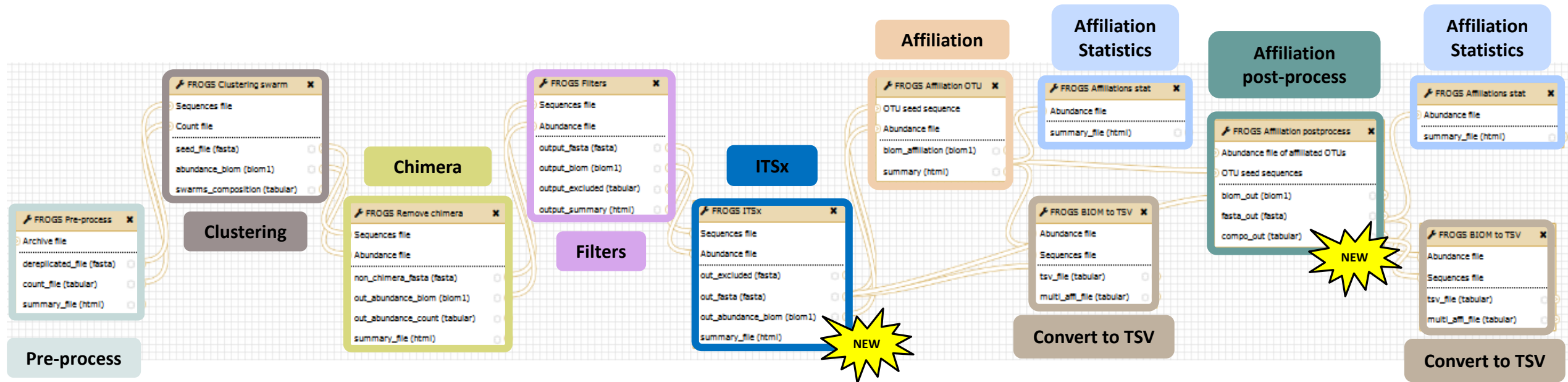- Detecting micro-eukaryotes at the species level, and potentially strain level, is therefore necessary.

Case of ITS1 amplicon MiSeq sequencing, a case of a sequencing of non-overlapping sequences
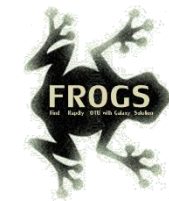
Imagine a real amplicon sequence of 700bp

700bp

Imagine a Miseq paired sequencing of 2x250bp

R1 : 250bp

R2 : 250bp

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

NNNNNNNNNNNNNNNNNN

Combined sequence length : 600bp, with 100 Ns

What workflow should we use to analyse ITS ?

Pre-process tool

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'.
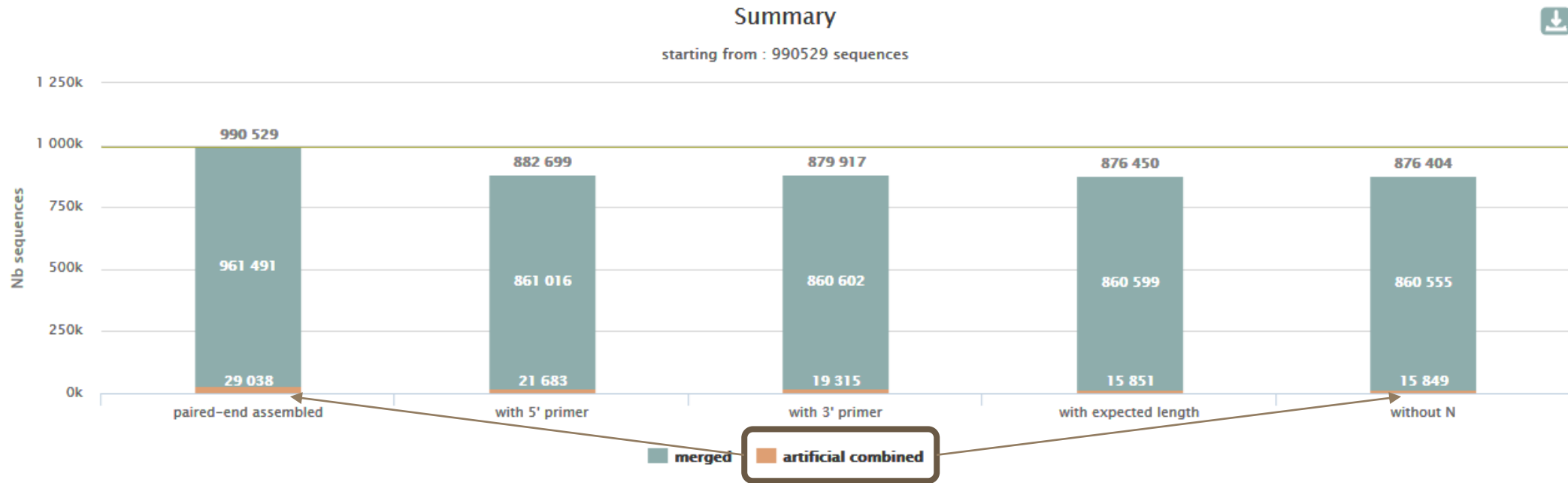
✔ Execute

Exercise 2.3

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

ITS

Explore Preprocess report.html

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

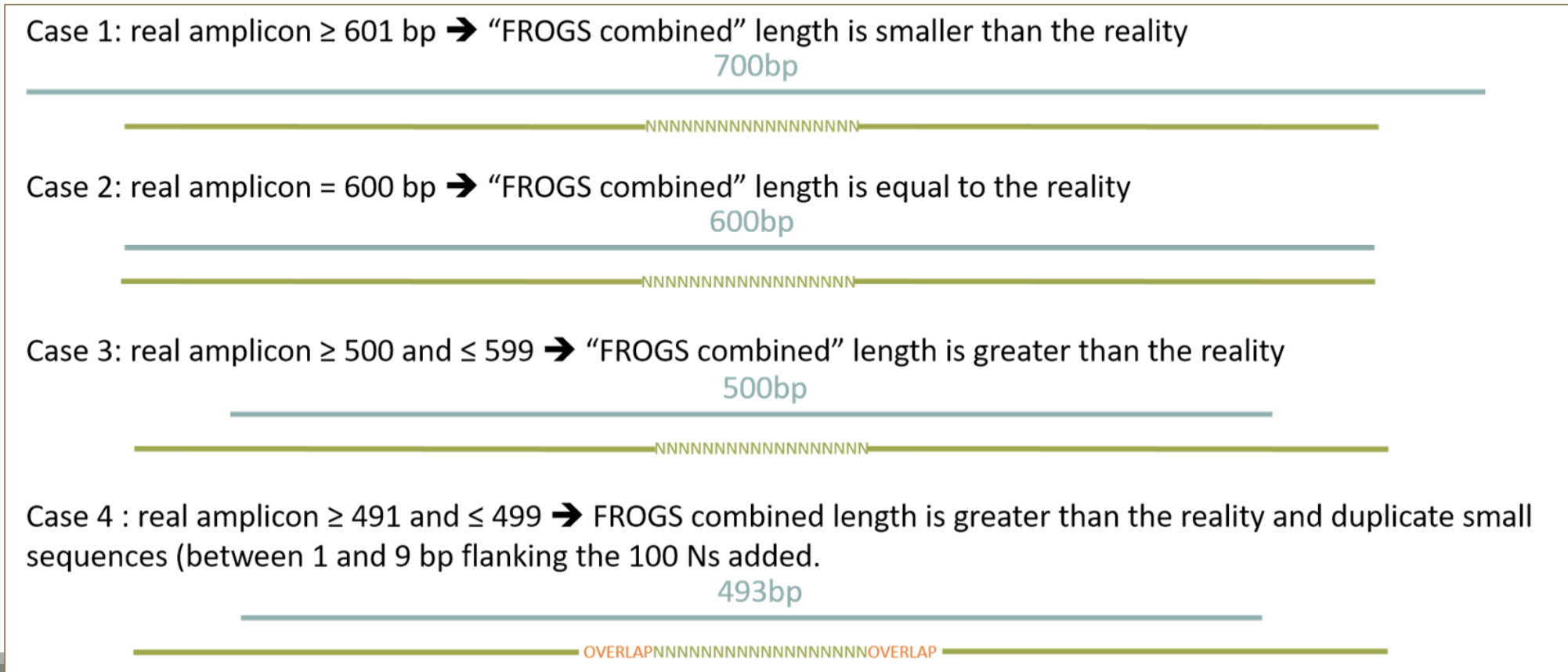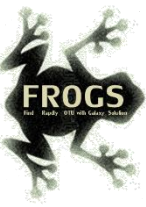Show [10 ⬍] entries   Search: [_____]

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

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.
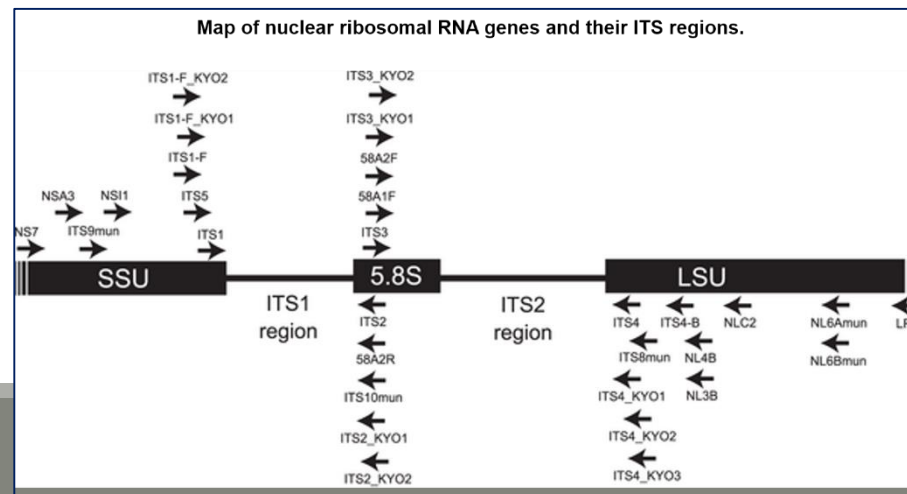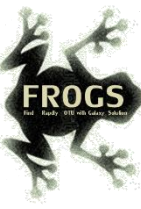

Case 1: real amplicon ≥ 601 bp ➜ "FROGS combined" length is smaller than the reality
700bp
NNNNNNNNNNNNNNNNNN

Case 2: real amplicon = 600 bp ➜ "FROGS combined" length is equal to the reality
600bp
NNNNNNNNNNNNNNNNNN

Case 3: real amplicon ≥ 500 and ≤ 599 ➜ "FROGS combined" length is greater than the reality
500bp
NNNNNNNNNNNNNNNNNN

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.
493bp
OVERLAPNNNNNNNNNNNNNNNNNNNNNNOVERLAP

ITSx tools

What is the purpose of the ITSx tool?

- ITSx is a tool to filter sequences.

- ITSx identifies and trimms ITS regions in our sequences.

- It excludes the highly conserved neighbouring sequences SSU, 5S and ARNr LSU.

- If the ITS1 or ITS2 region is not detected, the sequence is discarded.

- You can choose to check only if the sequence is detected as an ITS.

- In this case, the sequence is not trimmed, only sequences not detected as ITS are rejected (*e.g.* contaminants).
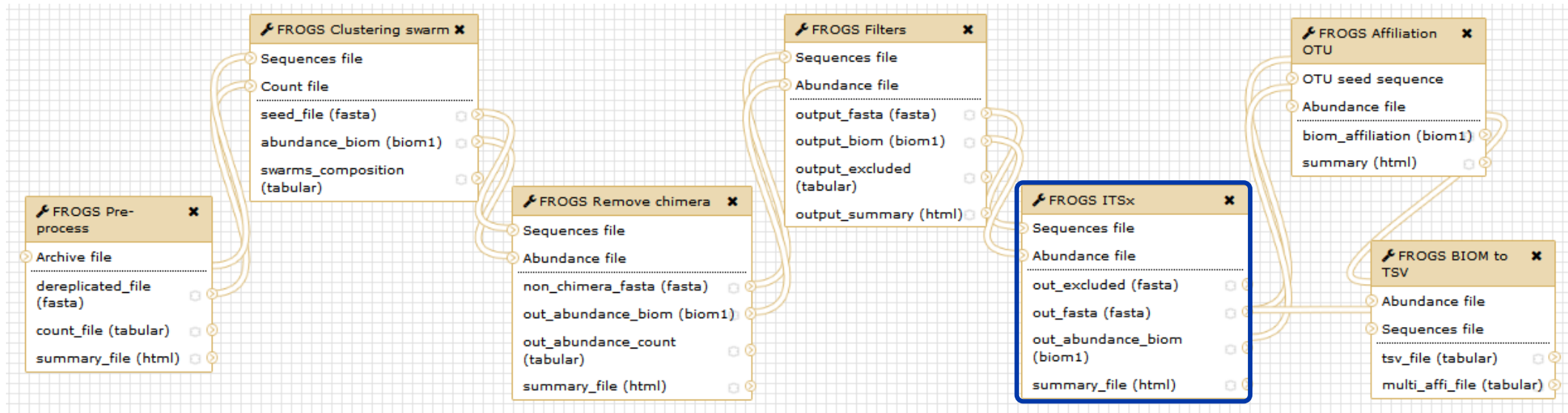
Map of nuclear ribosomal RNA genes and their ITS regions.

When should we use ITSx ?

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

- It is interesting to keep only the ITS parts without the flanking sequences in case one would like to compare sequenced amplicons with different primers targeting the same region to be amplified.

- You can choose this option on configuration panel of ITSx Tool.

- Reply "No" to question "Check only if sequence is detected as ITS?".
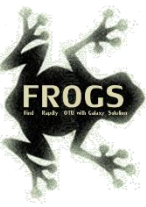
- In opposite, if "Yes" is chosen, sequences with ITS signature will be kept without trimming SSU, LSU or 5.8S regions.

Carreful !

- The ITSx step is time consuming and has to be done on clusters. We advise our users to apply ITSx in 5th step:

1. Preprocess step,

2. Clustering step,

3. Chimera removing step,

4. Filter on OTUs abundances and replicats step,

5. ITSx if Fungi ITS amplicons.

- Carreful, ITSx is currently usable for the detection of fungi ITS <u>neither</u> plants <u>nor</u> other eukaryotes.

Filters (ITSx) summary

OTUs

Removed: 0

Kept: 87

Abundance

Removed: 0

Kept: 853 861
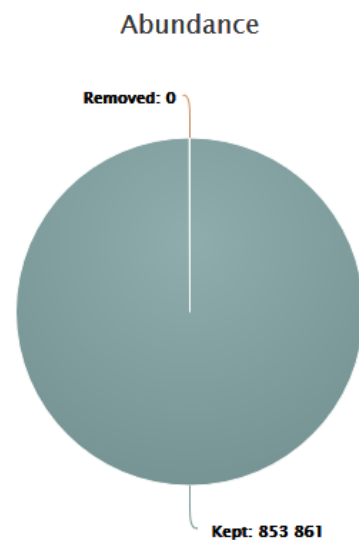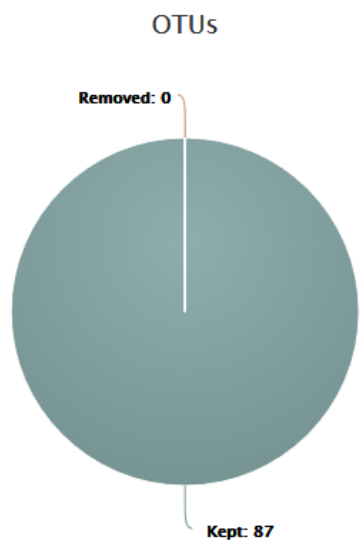
Filters (ITSx) by samples

⬇ CSV

Show [10 ‡] entries
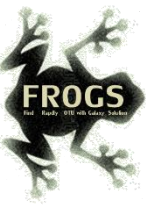
Search: [   ]

OTUs removed by sample

Sample name ↑↓	Initial ↑↓	Kept ↑↓	Initial abundance ↑↓	Kept abundance ↑↓
complexe-ADN-1	65	65	47,980	47,980
echantillon1-1	63	63	26,797	26,797
echantillon1-2	64	64	51,499	51,499
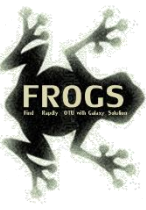
ITS Affiliation

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

- blastn+ or needlall is used to find alignment between each OTU and the database.

- Only the bests hits with the same score are reported.

- blastn+ is used for merged read pair, and needall is used for **artificially combined sequence**.

- For each alignment returned, several metrics are computed: identity percentage, coverage percentage, and alignment length.

- If "combined" sequences are stayed presents in OTUs, blastn+ is not usable as for classical merged sequences.

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

- So, sequences are affiliated in 3 steps: Alignment of classical "merged" sequences with blastn+ *versus* chosen database (*e.g.* UNITE),

- Alignment of "combined" sequences with blastn+ *versus* chosen database, best hits are collected and a very small new databank (at most 200 references per blast hit) is created composed exclusively of "subject" sequences from these best hits,

- Alignment of "combined" sequences with needlall (global alignment: very time consuming) *versus* these small new databank.
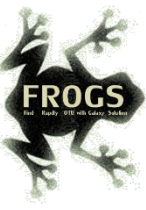
Careful, with "combined" sequences, we introduced some modification on identity percentage
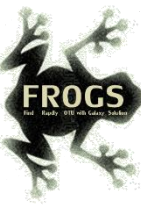
Case 1: a sequencing of overlapping sequences i.e. 16S V3-V4 amplicon MiSeq sequencing

Imagine a real amplicon sequence of 400bp

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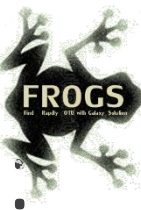
Case2: a sequencing of non-overlapping sequences: case of ITS1 amplicon MiSeq sequencing

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%

Conclusion on identity percentage for ITS

Filtering on %id will systematically removed "FROGS combined" OTUs.

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

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

Case 1 : 16S V3V4 ➜ overlapped sequence

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

Case 2 : very large ITS1 ➜ "FROGS combined" shorter than the real sequence

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

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

Affiliation Post-process

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

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

1. They must have the same affiliation

Or

2. If they have "multi-affiliation" tag in FROGS taxonomy, they must have in common in their list of possible affiliations at least one identical affiliation.
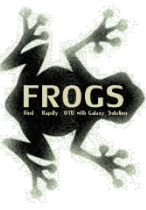
What is the purpose of the *Affiliation post-process* tool ?

In consequence:

The different affiliations involved in multi-affiliation are merged.

The abundances are added together.

It is the most abundant OTU seed that is kept.

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

In case of ITS amplicon analyses,

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

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

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

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

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

OTU1: ITS1 species A with abundance = N → Taxonomy = Fungi A + Fungi B

OTU2: ITS1 species B with abundance = M → Taxonomy = Fungi A

Affiliation post-process tool keeps :

OTU2: ITS1 species B with abundance = N + M → Taxonomy = Fungi A + Fungi B

Post-affiliation Tool - output

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

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

```
>Cluster_1 reference=AB241105 amplicon=1..444 position=1..444 errors=440%A
TGGGGAATATTGCACAATGGGGGGAACCCTGATGCAGCGACGCCGCGTGCGGGATGACGGCCTTCGGGTTGTAAACCGCTTTTAATTGGGAGCAAGCAGTTTTA
>Cluster_2 reference=AJ496032 amplicon=1..452 position=1..452 errors=440%G
TAGGGAATATTGCACAATGGGGGAAACCCTGATGCAGCAACGCCACGTGTGGGAAGAAGCATTTCGGTGTGTAAACCACTGTCATGAGGGAATAAGGCCCGCCT
>Cluster_3 reference=EU240886 amplicon=1..460 position=1..460
TAGGGAATCTTCCGCAATGGGCGAAAGCCTGACGGAGCAACGCCGCGTGTGCGAAGAAGGTCTTCGGATCGTAAAGCACTGTTGTTAAGGAAGAACGACAGTAA
>Cluster_4 reference=U39399 amplicon=1..459 position=1..459
TGGGGAATATTGGACAATGGGGGAAACCCTGATCCAGCCATGCCGCGTGTGTGAAGAAGGCCTTTTGGTTGTAAAGCACTTTAAGCAGTGAAGAAGACTCCGTG
>Cluster_5 reference=FR733705 amplicon=1..452 position=1..452 errors=436%C
TGGGGAATATTGGACAATGGGCGAAAGCCTGATCCAGCGACGCCGCGTGAAGGAAGAAGTCCTTCGGGATGTAAACTTCTGAACTAATCGAATAAGAGGGTAGT
>Cluster_6 reference=GU575117 amplicon=1..434 position=1..434
TGGGGAATATTGGACAATGGGCGCAAGCCTGATCCAGCCATGCCGCGTGTGTGATGAAGGCCCTAGGGTTGTAAAGCACTTTCAACGGTGAAGATAATGACGGT
>Cluster_7 reference=AB272165 amplicon=1..454 position=1..454 errors=441%A
TAGGGAATATTGGGCAATGGGTGAGAGCCTGACCCAGCCATGCCGCGTGCCGGACGAAGGCCCTCAGGGTCGTAAACGGCTTTTGCCGGGGAAGAAGAGGTTCC
>Cluster_8 reference=AJ292759 amplicon=1..437 position=1..437
TAGGGAATATTGCACAATGGAGGAAACTCTGATGCAGCGACGCCGCGTGAGTGATGAAGGCCTTCGGGTCGTAAAGCTCTGTCGCAGGGGAATAACACAATGAA
>Cluster_9 reference=CP000027 amplicon=1..435 position=1..435
CAAGGAATCTTGGGCAATGGGCGAAAGCCTGACCCAGCAACGCCGCGTGAGGGATGAAGGCTTTCGGGTTGTAAACCTCTTTTCACAGGGAAGAATAATGACGG
>Cluster_10 reference=JN880417 amplicon=1..438 position=1..438 errors=11%A
TCGAGGATCTTCGTCAATGGGCGAAAGCCTGAACGAGCGATTAGCCGCGTGCGCGATGAAGGCCTTCGGGTTGTAAAGCGCGAAAGAGGTAATAAAGGGAAACT
>Cluster_11 reference=EF660760 amplicon=1..434 position=1..434 errors=424%C
TGGGGAATCTTGCACAATGGGCGAAAGCCTGATGCAGCCATGCCGCGTGAATGATGAAGGCCTTAGGGTTGTAAAATTCTTTCGCCAGGGATGATAATGACAGT
>Cluster_12 reference=AB594446 amplicon=1..443 position=1..443 errors=438%G
TGGGGAATATTGCACAATGGGCGCAAGCCTGATGCAGCGACGCCGCGTGGGGGATGACGGCCTTCGGGTTGTAAACTCCTTTCGCCATTGACGAAGCCTTTTG
>Cluster_13 reference=U93332 amplicon=1..439 position=1..439 errors=426%C
```
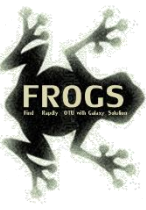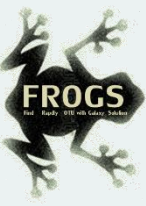
Workflow creation
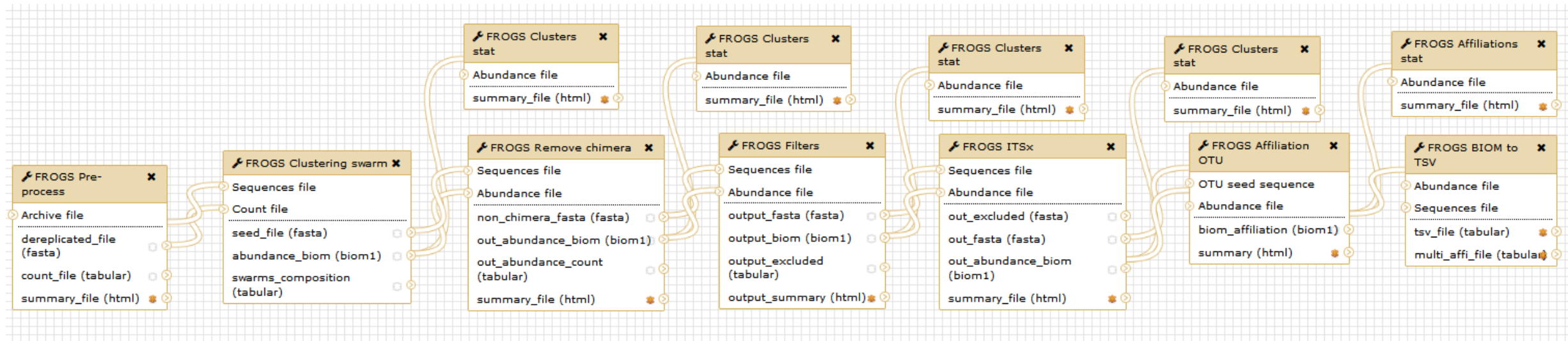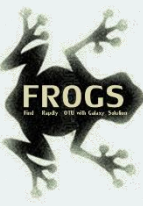
Your Turn! – 10

CREATE YOUR OWN WORKFLOW !

Exercise 10

1. Create your own workflow with ITS data

Exercise 10

Exercise 10

Exercise 10

Solution of exercise:

What about « upload file » ?

What about « Normalisation Tool » ?

For each tool, think to:
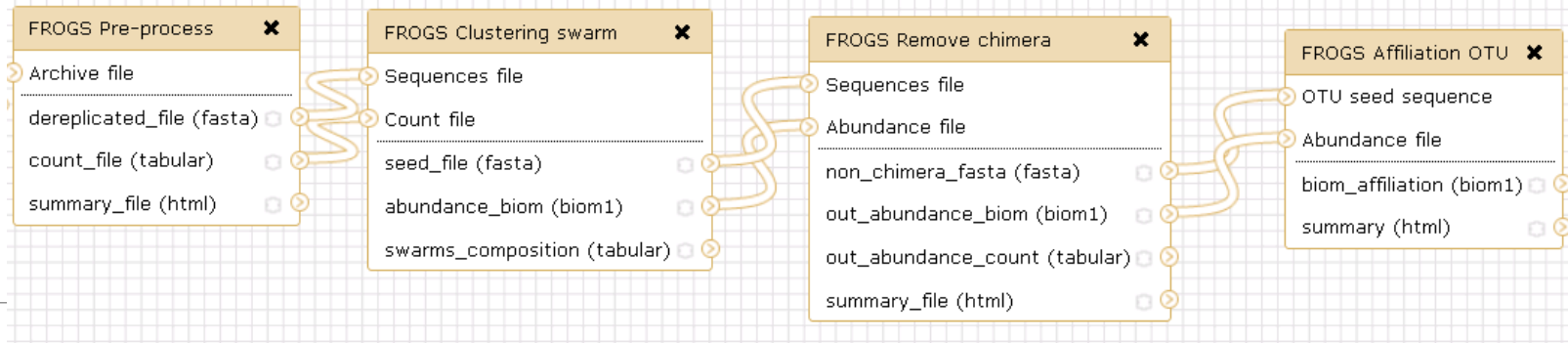• Fixe parameter ?

?

For each tool, think to:
- Fixe parameter ?
- Automatically rename output files

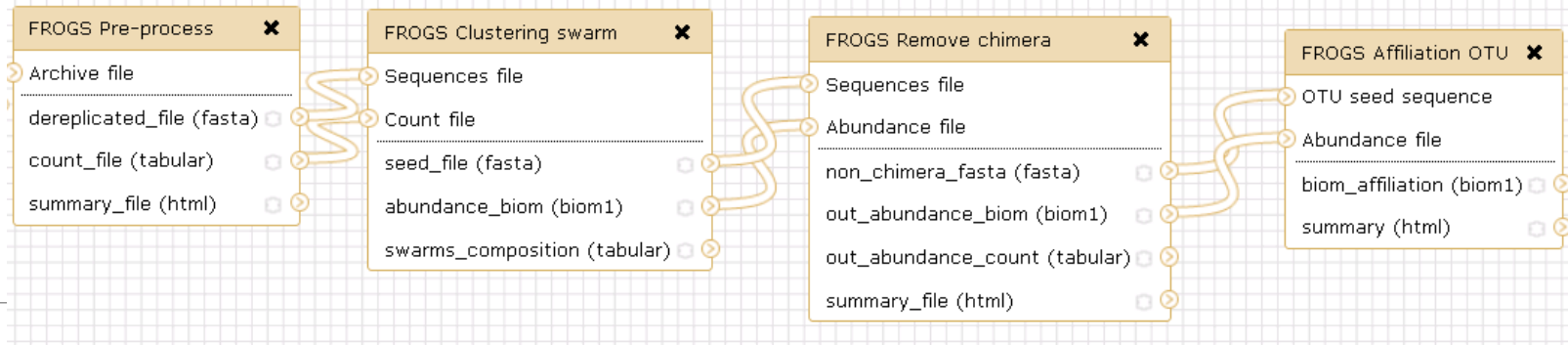For each tool, think to:
- Fixe parameter ?
- Automatically rename output files
- Hide intermediate files ?

For each tool, think to:
- Fixe parameter ?
- Automatically rename output files
- Hide intermediate files ?

For each tool, think to:
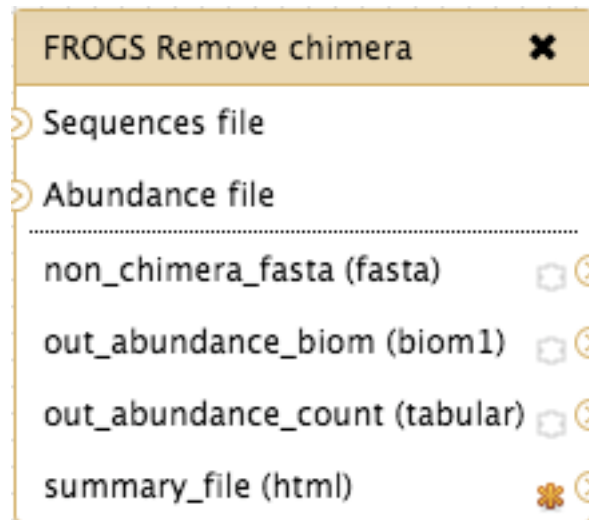- Fixe parameter ?
- Automatically rename output files
- Hide intermediate files ?