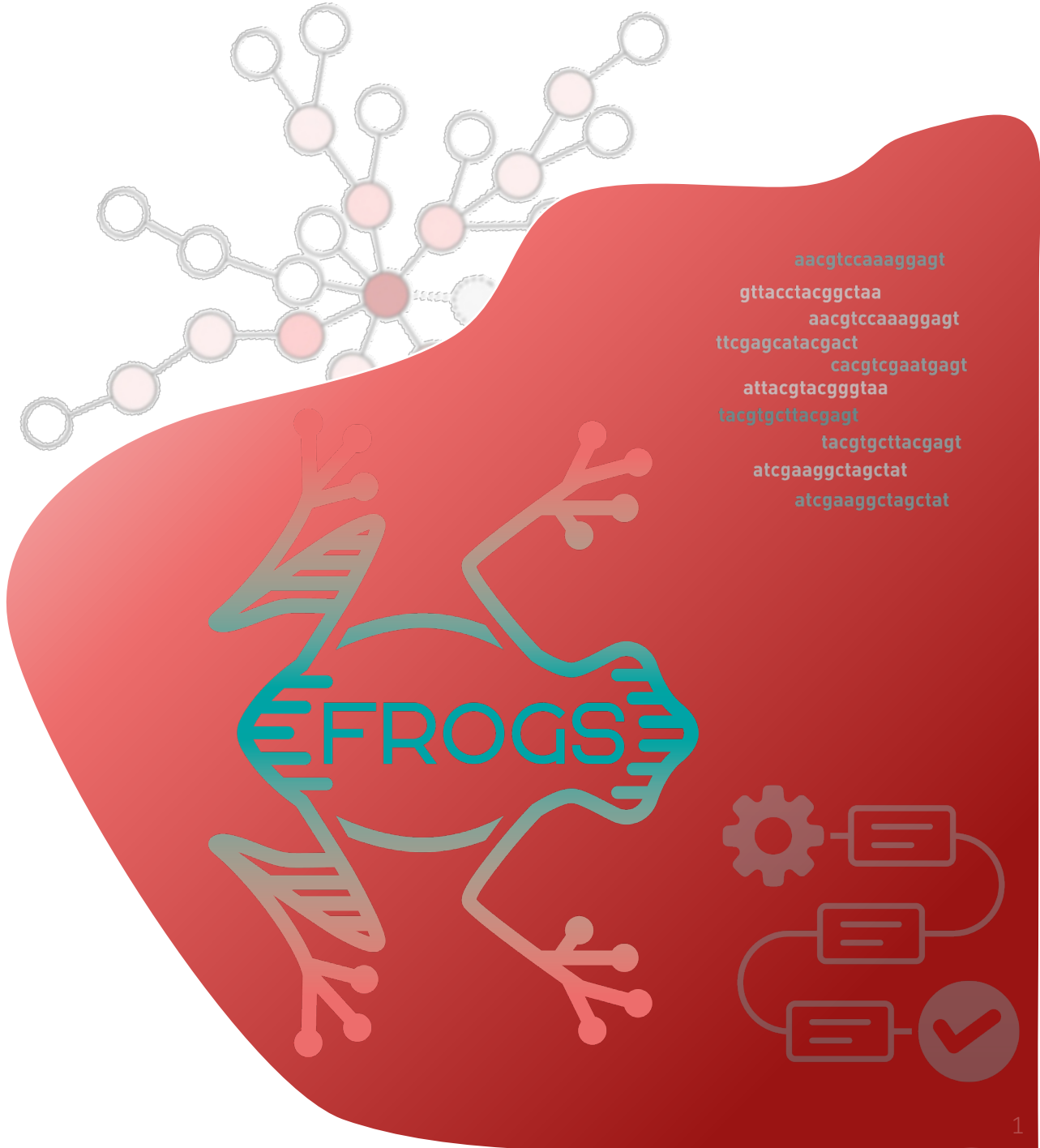
FROGS Core

The essential steps of metabarcoding analysis.

Lucas Auer, Gabryelle Agoutin, Maria Bernard, Géraldine Pascal, Maëlle Pomiès & Olivier Rué



Datasets:
the microbial community in cheese

aacgtccaaaggagt
gttacctacggctaa
aacgtccaaaggagt
ttcgagcatacgact
cacgtcgaatgagt
attacgtacgggtaa
tacgtgcttacgagt
tacgtgcttacgagt
atcgaaggctagctat
atcgaaggctagctat

FROGS

From F. Irlinger communication

MetaPDOcheese

Mapping the microbial diversity of French PDO (*AOP en français*) milks and cheeses and exploring technological determinants

Papers

Irlinger F., Mariadassou M., Dugat-Bony E. et al. (2024). A comprehensive, large-scale analysis of 'terroir' cheese and milk microbiota reveals profiles strongly shaped by both geographical and human factors. ISME Communications, DOI : https://doi.org/10.1093/ismeco/ycae095

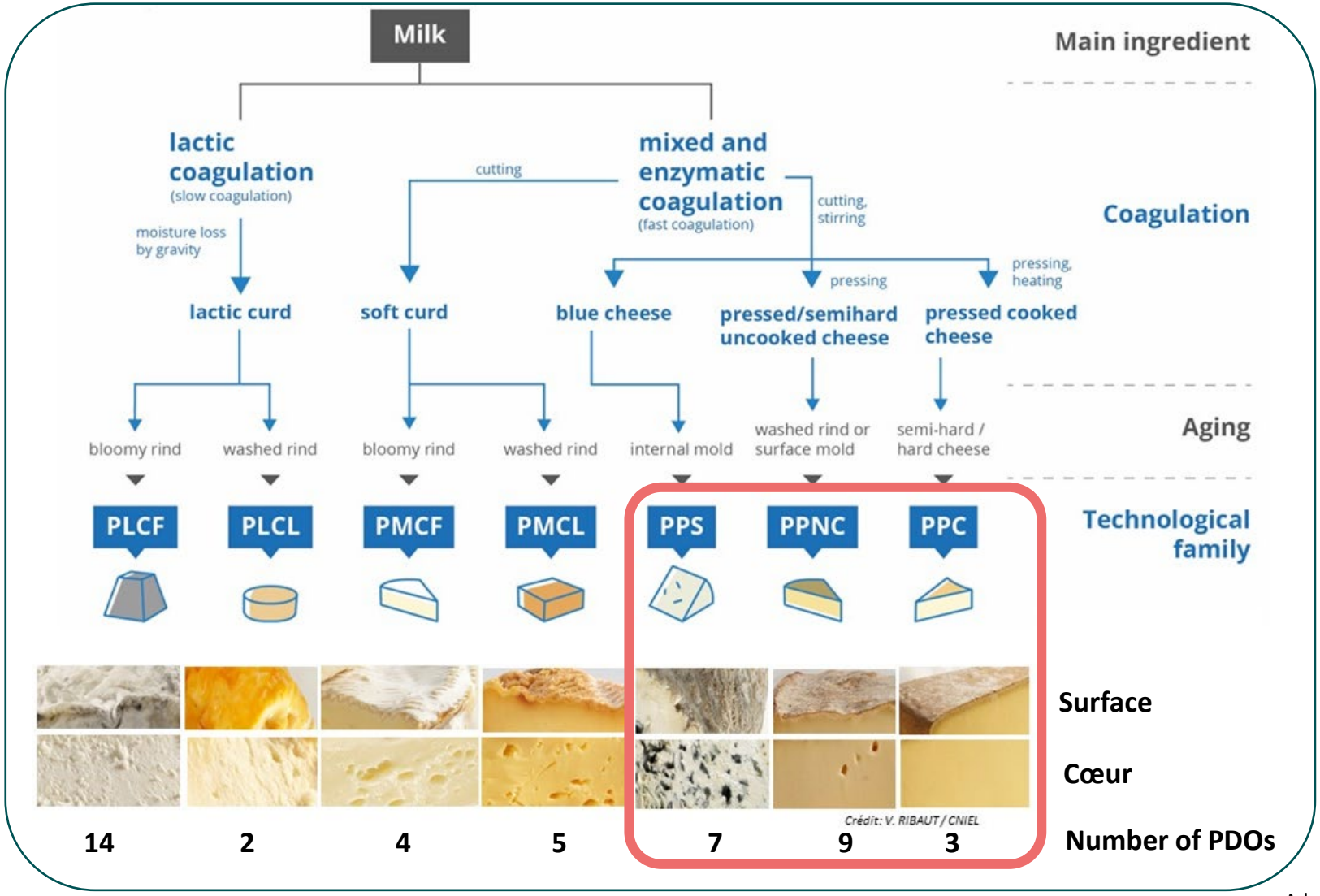
Irlinger, Françoise; Mariadassou, Mahendra; Dugat-Bony, Eric; Rué, Olivier; Neuvéglise, Cécile; Renault, Pierre; Rifa, Etienne; Theil, Sébastien; Loux, Valentin; Cruaud, Corinne; Gavory, Frederick; Barbe, Valérie; Lasbleiz, Ronan; Gaucheron, Frédéric; Spelle, Céline; Delbès, Céline, 2024, "metapdocheese_Fungi_sample_metadata.tab", Metabarcoding data MetaPDOcheese project, https://doi.org/10.57745/E36FNE, Recherche Data Gouv, V1, UNF:6:6wvuMqZyr6R4qIbdfciYcg== [fileUNF]

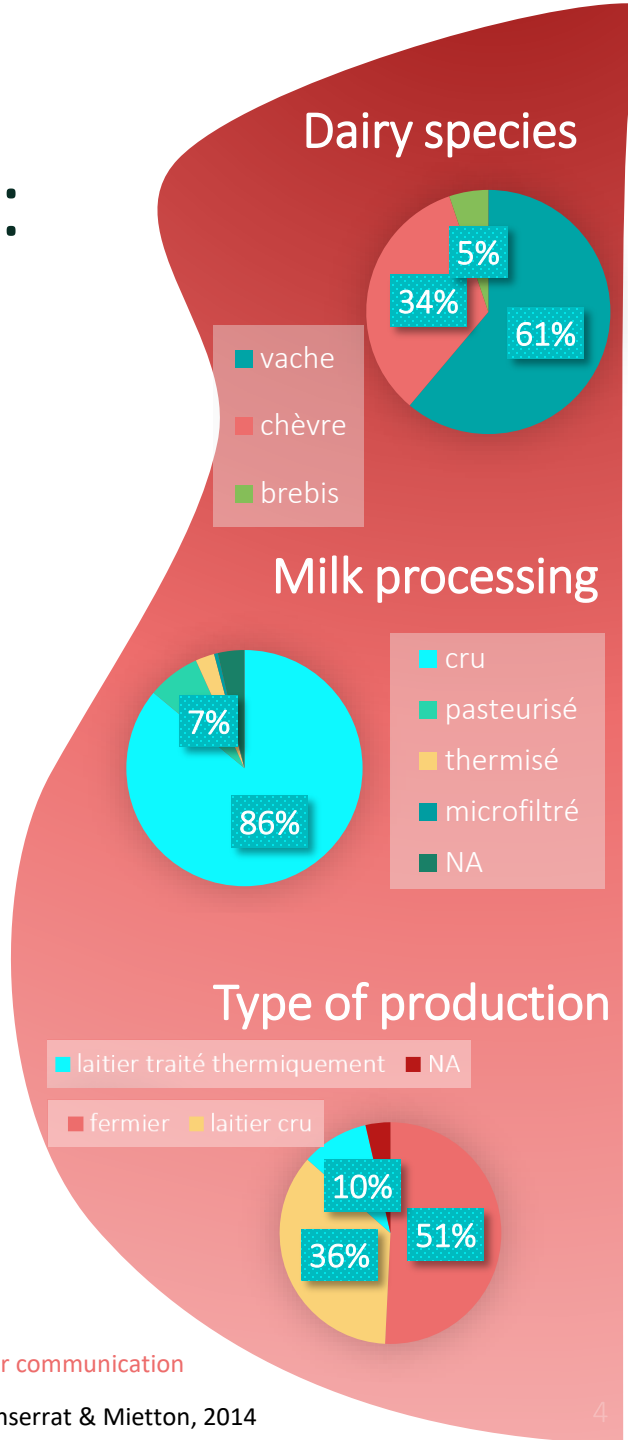
Irlinger, Françoise; Mariadassou, Mahendra; Dugat-Bony, Eric; Rué, Olivier; Neuvéglise, Cécile; Renault, Pierre; Rifa, Etienne; Theil, Sébastien; Loux, Valentin; Cruaud, Corinne; Gavory, Frederick; Barbe, Valérie; Lasbleiz, Ronan; Gaucheron, Frédéric; Spelle, Céline; Delbès, Céline, 2024, "Metabarcoding data MetaPDOcheese project", https://doi.org/10.57745/UCJG6S, Recherche Data Gouv, V1, UNF:6:yT7X4YI5sEi1AUUj44npNg== [fileUNF]

Data:

https://entrepot.recherche.data.gouv.fr/file.xhtml?persistentId=doi:10.57745/E36FNE&version=1.0

44 PDOs divided into 7 technological families:





From F. Irlinger communication

Adapted from Monserrat & Mietton, 2014

72 samples from MetaPDOCheese

- 2 PDOs (25 & 3) → AOP1 & 2
- 2 seasons (winter & summer)
- 2 production units/PDO
- 3 replicats/samples

- Surface only
- ITS & 16S
- Cow milk

PPS
Pâte persillée

- 2 PDOs (28 & 29) → AOP3 & 4
- 2 seasons (winter & summer)
- 2 production units /PDO
- 3 replicats/samples

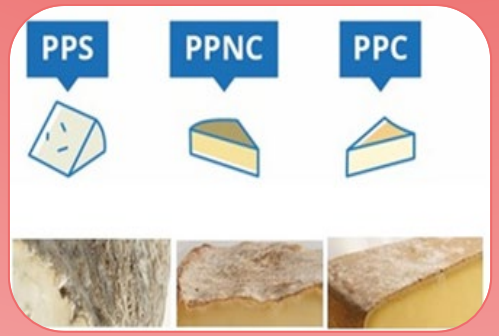
- Surface only
- ITS & 16S
- Cow and goat milks

PPNC
Pâte pressée non cuite

- 2 PDOs (34 & 35) → AOP5 & 6
- 2 seasons (winter & summer)
- 2 production units/PDO
- 3 replicats/samples

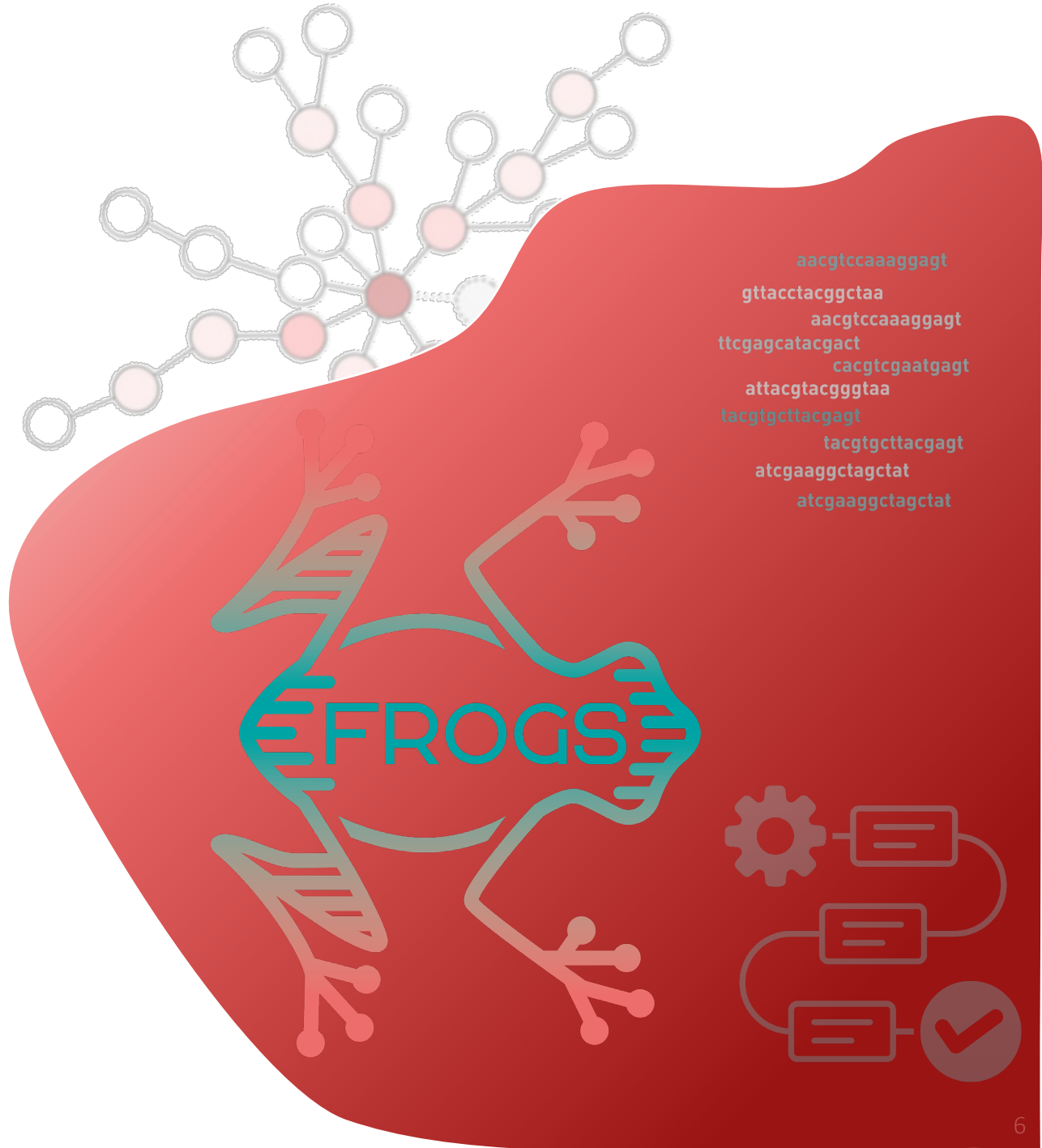
- Surface only
- ITS & 16S
- Cow milk

PPC
Pâte pressée cuite

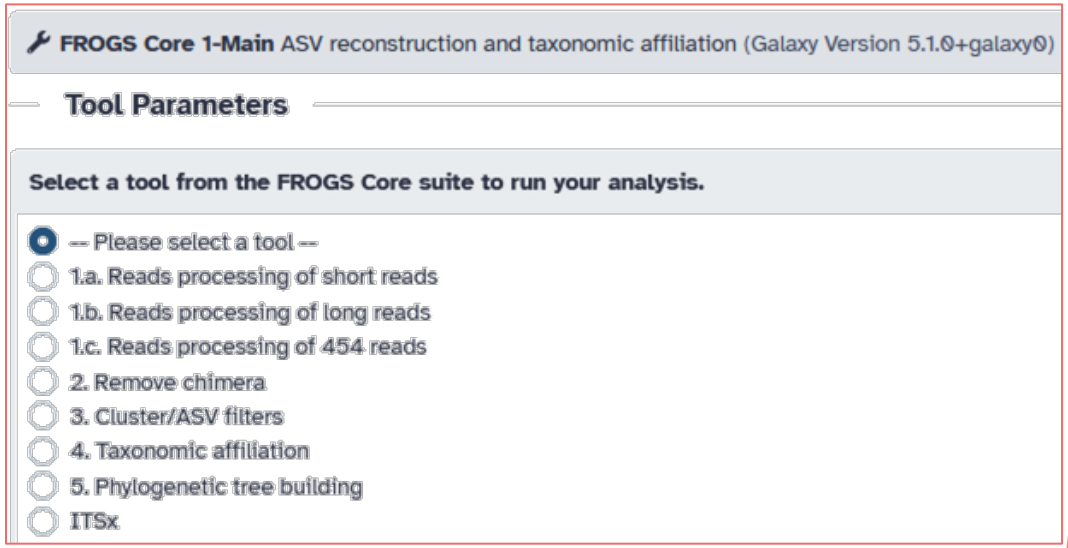


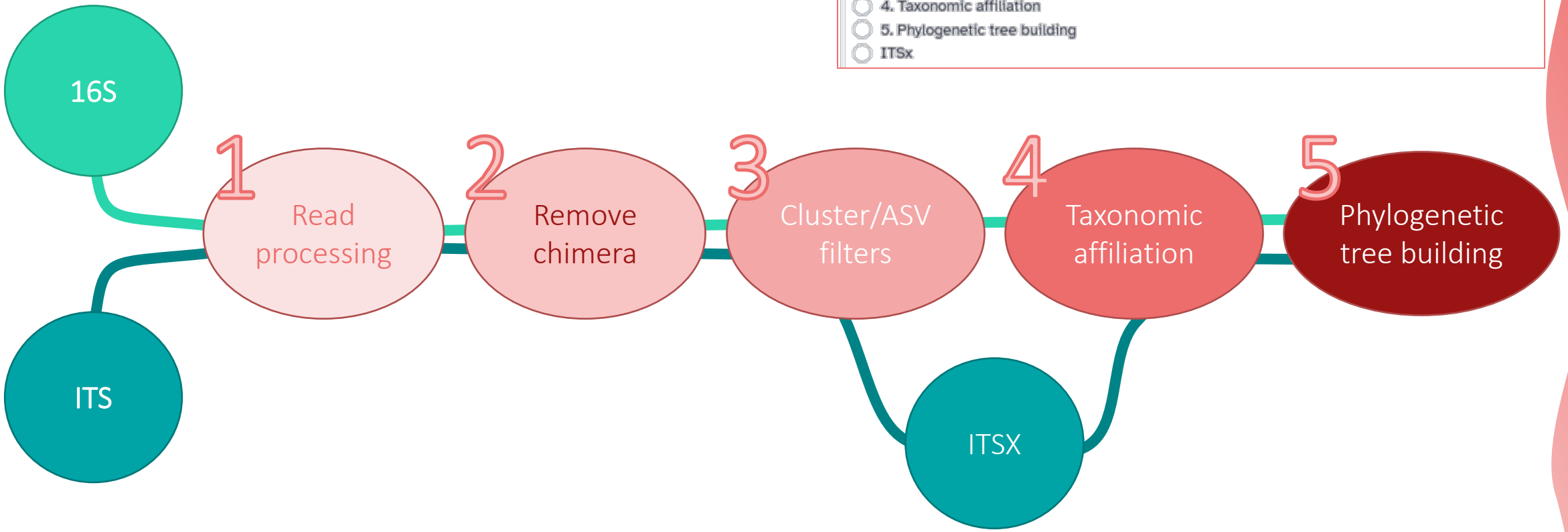
FROGS Core

Overview

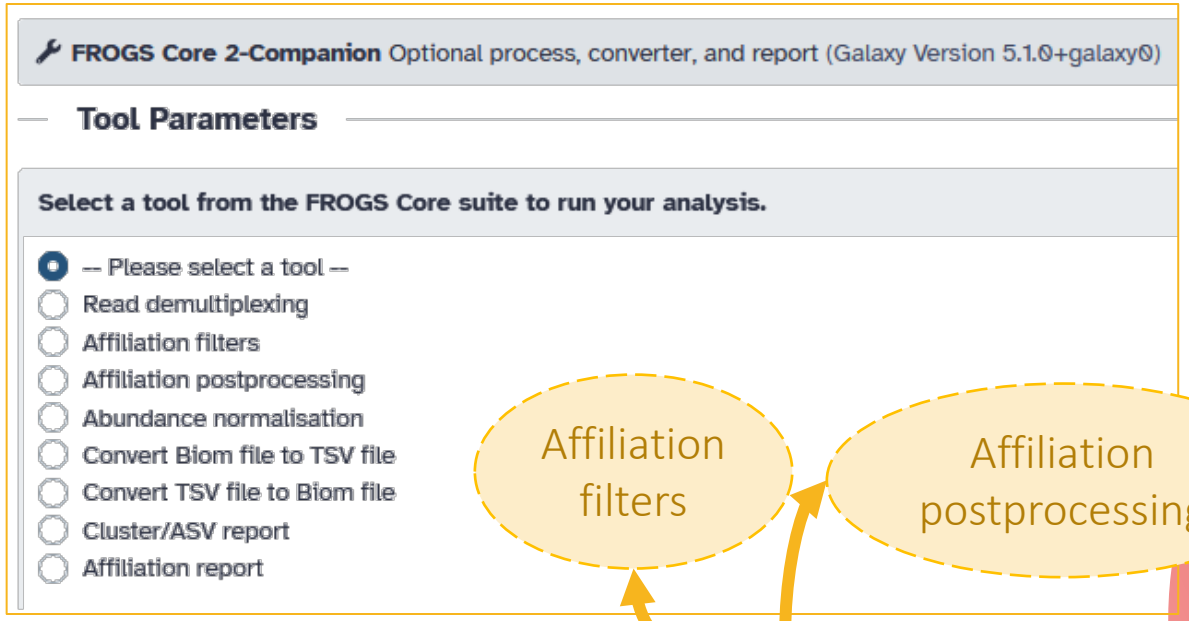


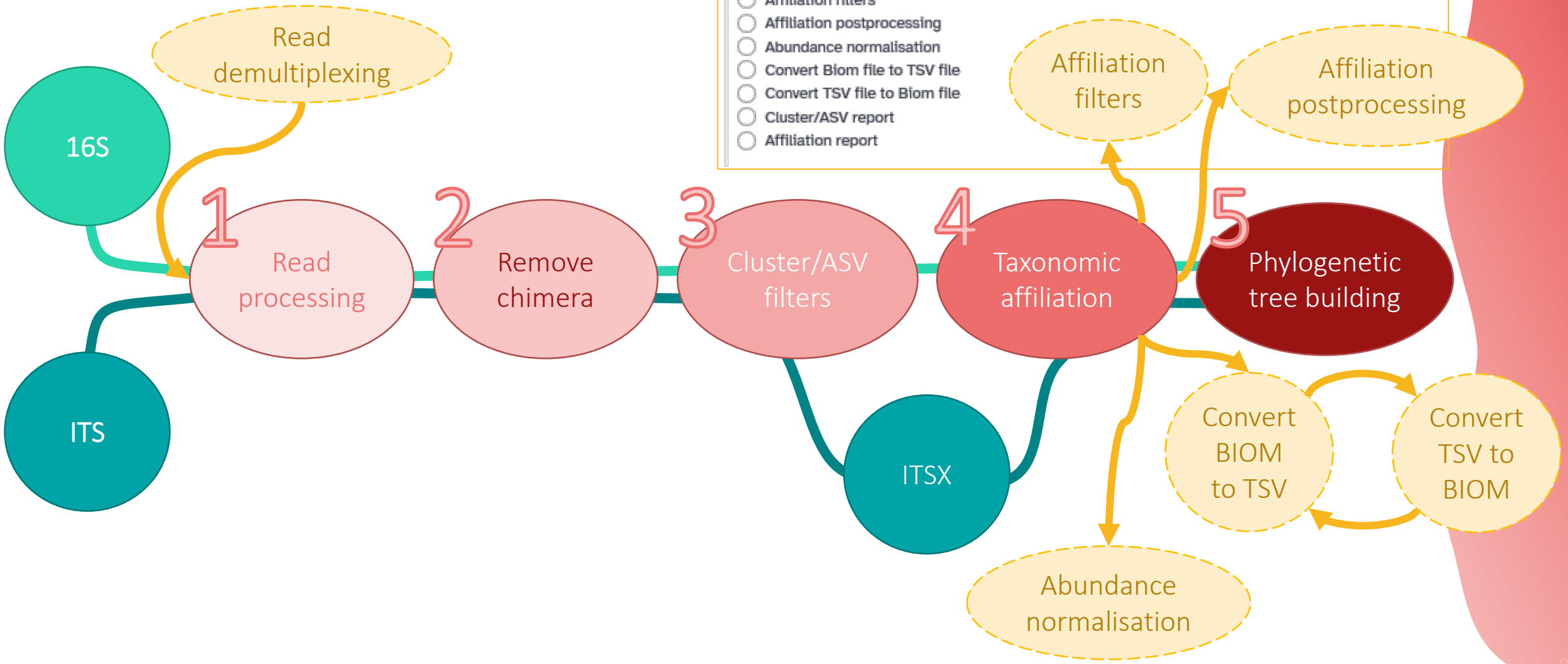
FROGS Core 1 Main tools





FROGS Core 1 Main tools

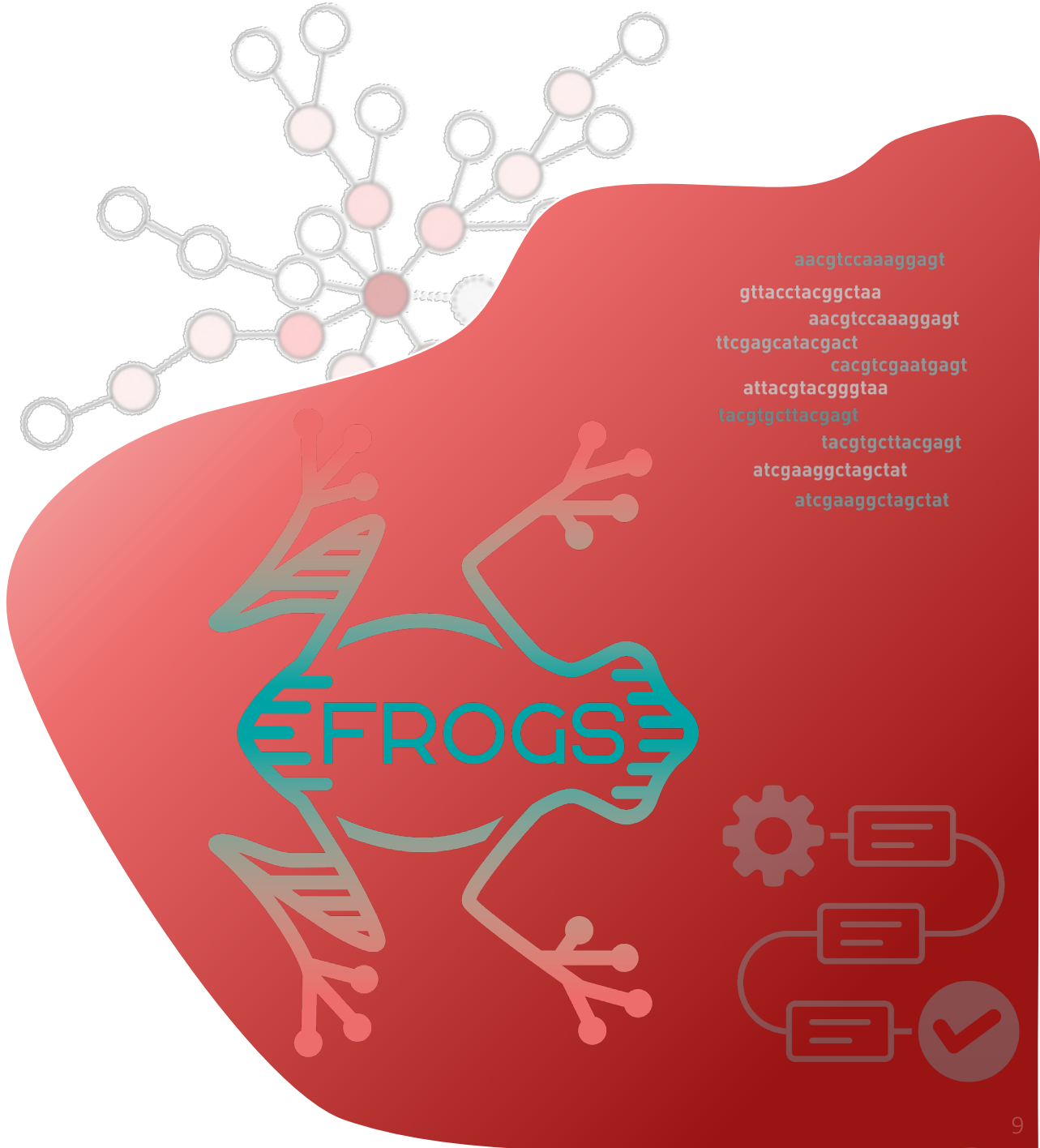




FROGS Core 1

Main tools

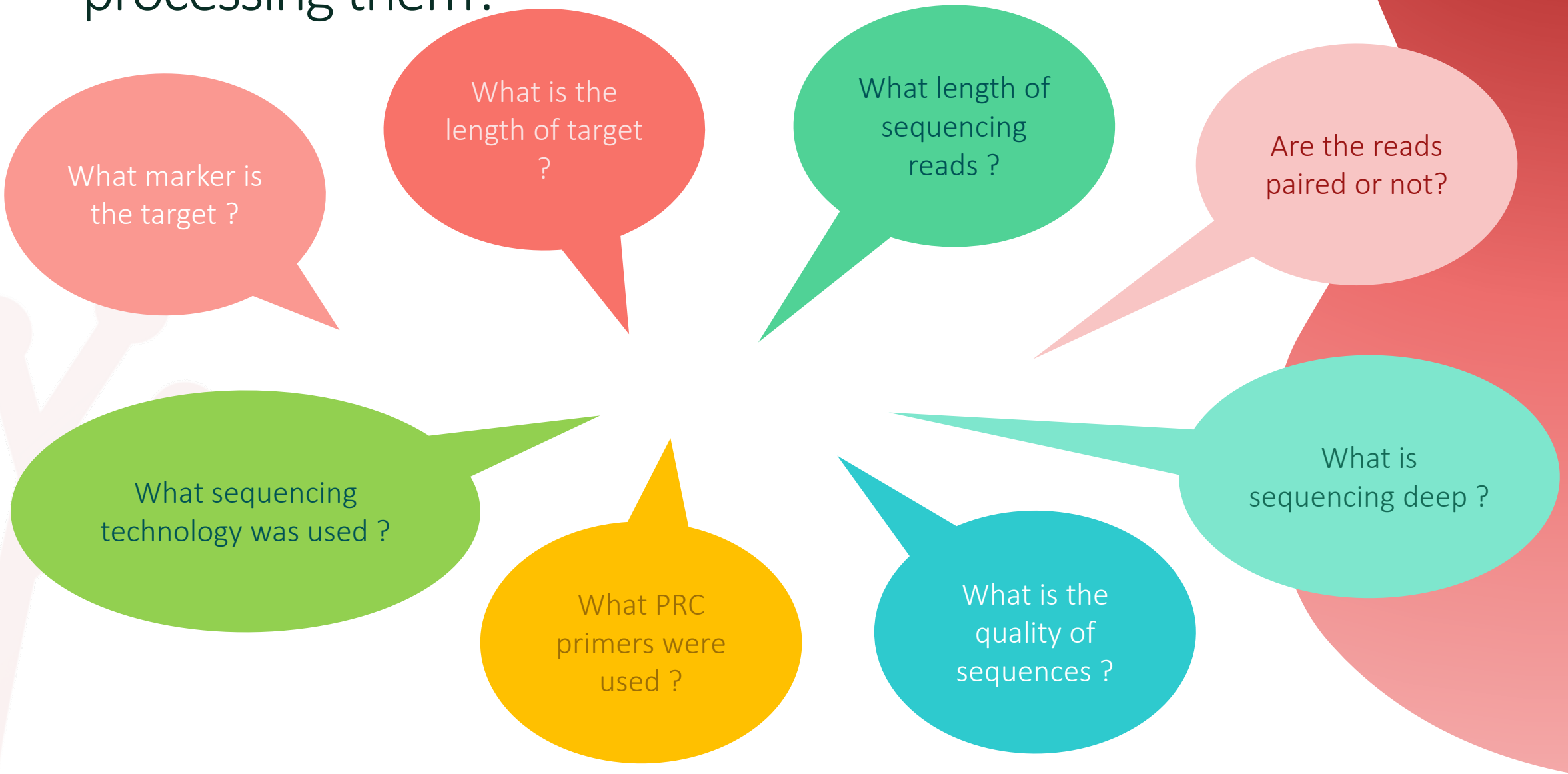
Read processing



What does the tool do?

- Preprocessing
  - Paired-end merging of R1 and R2 reads with vsearch, flash or pear (only in command line)
  - Delete sequences without good primers
  - Finds and removes adapter sequences with cutadapt
  - Delete sequence with not expected lengths
  - Delete sequences with ambiguous bases (N)
  - Dereplication
- Clustering sequences with swarm / or / Denoising with DADA2

What do we need to know about our sequences before processing them?



Features of 16S reads

16S

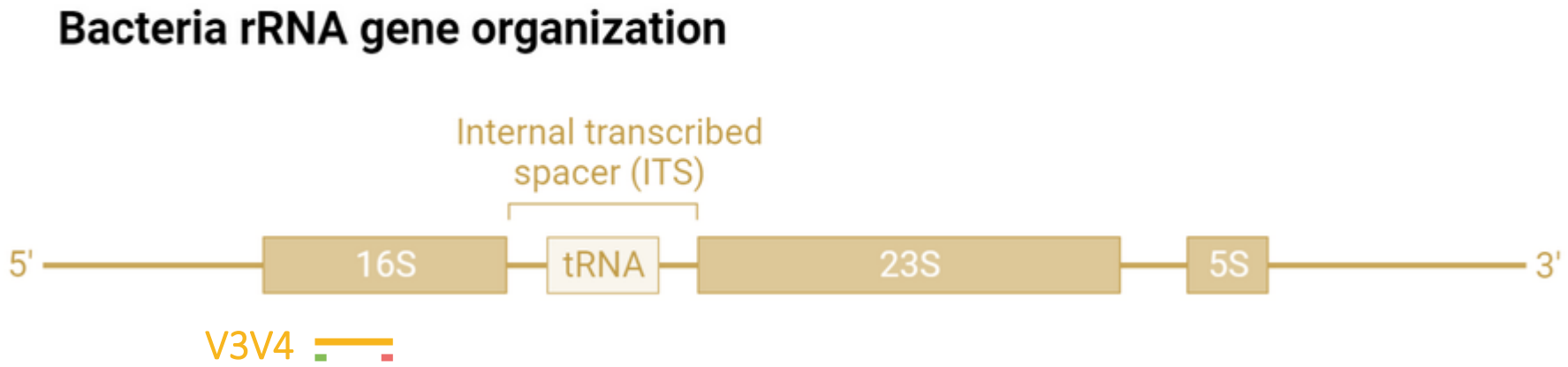
Read lengths: 250

Minimum amplicon length: 250

Maximum amplicon length: 480

5’ PCR primer: 5’ ACGGRAGGCWGCAG 3’

3’ PCR primer: 3’ TACCAGGGTATCTAATCCT 5’



Features of ITS reads

ITS

Read lengths: 250

Minimum amplicon length: 150

Maximum amplicon length: 480

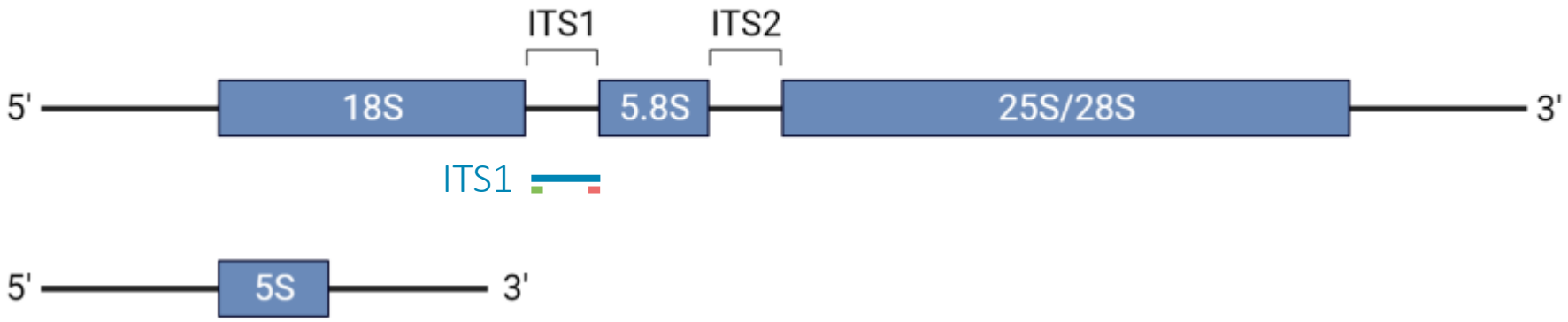
5’ PCR primer: 5’ GCATCGATGAAGAACGCAGC 3’

3’ PCR primer: 3’ TCCTCCGCTTWTTGWTWTGC 5’

In fungi,

- 18S rDNA is relatively non-specific, with little variation between species.
- the transcribed inter-subunit regions (Internal Transcribed Spacers) vary greatly in sequence and size.

Eukaryotes rRNA gene organization



Practice session

Please open the FROGS Core Main 1 tool corresponding to your data and familiarize yourself with the required parameters.

Please, enter all the information you have/understand.

But, no run tool.

Practice session

16S

Select a tool from the FROGS Core suite to run your analysis.

- ○ -- Please select a tool --
- ◉ 1.a. Reads processing of short reads
- ○ 1.b. Reads processing of long reads
- ○ 1.c. Reads processing of 454 reads
- ○ 2. Remove chimera
- ○ 3. Cluster/ASV filters
- ○ 4. Taxonomic affiliation
- ○ 5. Phylogenetic tree building
- ○ ITSx

Paired-end or Single-end reads

- ◉ Paired-end reads
- ○ Single-end reads or Paired reads that have already been merged.

Input

TAR Archive

Sample files can be provided either as a single TAR archive or as se

Archive file (.tar.gz) *

3: MPC_16S.tar.gz

accepted formats

The TAR file containing the short R1 R2 read pairs (.fastq.gz) for eacl
Aviti and IonTorrent. (--input-archive)

R1 read length *

250

Please provide the maximum length of the R1 reads. (--R1-size)

R2 read length *

250

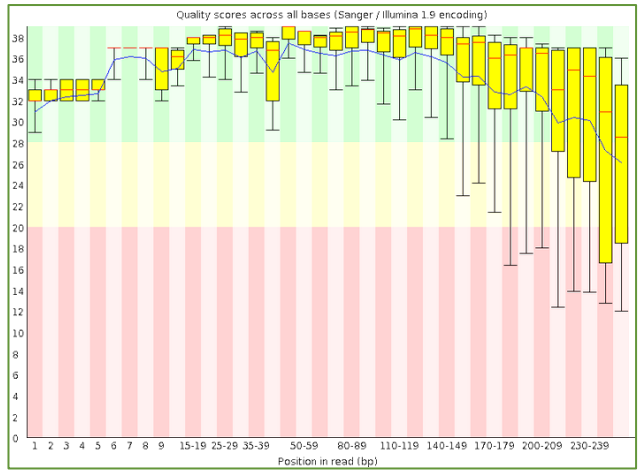
Please provide the maximum length of the R2 reads. (--R2-size)

Mismatch rate (used for R1-R2 merging) *

0,1

If your sequences are of low quality, you can increase this parameter.
But be careful!

Use FASTQC (included in Galaxy Tools) to check the quality of the sequence.



Practice session

16S

Select a tool from the FROGS Core suite to run your analysis.

- ( ) -- Please select a tool --
- (•) 1.a. Reads processing of short reads
- ( ) 1.b. Reads processing of long reads
- ( ) 1.c. Reads processing of 454 reads
- ( ) 2. Remove chimera
- ( ) 3. Cluster/ASV filters
- ( ) 4. Taxonomic affiliation
- ( ) 5. Phylogenetic tree building
- ( ) ITSx

Paired-end or Single-end reads

- (•) Paired-end reads
- ( ) Single-end reads or Paired reads that have already been merged.

Input

TAR Archive

Sample files can be provided either as a single TAR archive or as se

Archive file (.tar.gz) *

3: MPC_16S.tar.gz

accepted formats

The TAR file containing the short R1 R2 read pairs (.fastq.gz) for eacl
Aviti and IonTorrent. (--input-archive)

R1 read length *

250

Please provide the maximum length of the R1 reads. (--R1-size)

R2 read length *

250

Please provide the maximum length of the R2 reads. (--R2-size)

Mismatch rate (used for R1-R2 merging) *

0,1

Paired-end merging tool

- (•) Vsearch
- ( ) Flash

There is a third method, PEAR, but it is only available via the command line.

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

Would you like to keep unmerged reads?

No

No = unmerged reads will be removed; Yes = unmerged reads will be

Minimum amplicon length *

250

The minimum length of the amplicons (including primers). For paired-e

Maximum amplicon length *

480

The maximum length of the amplicons (including primers). For paired-e

Practice session

16S

Select a tool from the FROGS Core suite to run your analysis.

- ( ) -- Please select a tool --
- (•) 1.a. Reads processing of short reads
- ( ) 1.b. Reads processing of long reads
- ( ) 1.c. Reads processing of 454 reads
- ( ) 2. Remove chimera
- ( ) 3. Cluster/ASV filters
- ( ) 4. Taxonomic affiliation
- ( ) 5. Phylogenetic tree building
- ( ) ITSx

Paired-end or Single-end reads

- (•) Paired-end reads
- ( ) Single-end reads or Paired reads that have already been merged.

Input

TAR Archive

Sample files can be provided either as a single TAR archive or as se

Archive file (.tar.gz) *

3: MPC_16S.tar.gz

accepted formats

The TAR file containing the short R1 R2 read pairs (.fastq.gz) for eacl Aviti and IonTorrent. (--input-archive)

R1 read length *

250

Please provide the maximum length of the R1 reads. (--R1-size)

R2 read length *

250

Please provide the maximum length of the R2 reads. (--R2-size)

Mismatch rate (used for R1-R2 merging) *

0,1

Paired-end merging tool

- (•) Vsearch
- ( ) Flash

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

Would you like to keep unmerged reads?

- ( ) No

No = unmerged reads will be removed; Yes = unmerged reads will be

Minimum amplicon length *

250

The minimum length of the amplicons (including primers). For paired-e

Maximum amplicon length *

480

The maximum length of the amplicons (including primers). For paired-e

Do the sequences include PCR primers?

- (•) Yes
- ( ) No

Indicate whether the sequences still include PCR primers. Select "Yes" if primers are present, "No" if they have already been removed

5' primer - optional

ACGGRAGGCWGCAG

Enter the 5' primer sequence. Wildcards are allowed. The sequence must be provided in 5' → 3' orientation. (--five-prim-primer)

3' primer - optional

AGGATTAGATACCCTGGTA

Enter the 3' primer sequence. Wildcards are allowed. The sequence must be provided in 5' → 3' orientation. (--three-prim-primer)

Practice session

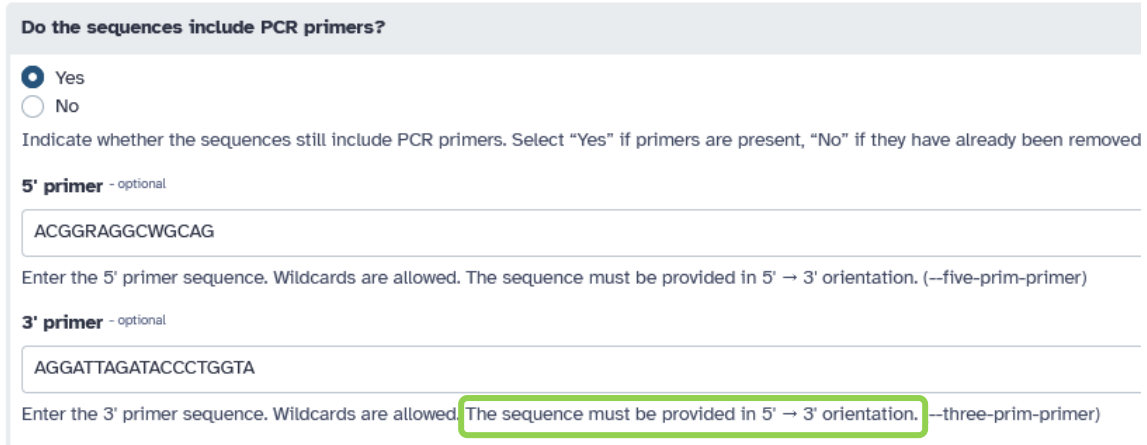
16S

What we knew:
5’ PCR primer: 5’ ACGGRAGGCWGCAG 3’
3’ PCR primer: 3’ TACCAGGGTATCTAATCCT 5’

But reads are separated into 2 sequences R1 and R2, both in 5’ → 3’ direction

You have to reverse transcribed le R2 primer
Use revseq: https://www.bioinformatics.nl/cgi-bin/emboss/revseq

5’ PCR primer: 5’ ACGGRAGGCWGCAG 3’
3’ PCR primer: 5’ AGGATTAGATACCCTGGTA 3’



Practice session

ITS

Select a tool from the FROGS Core suite to run your analysis.

- -- Please select a tool --
- (●) 1.a. Reads processing of short reads
- 1.b. Reads processing of long reads
- 1.c. Reads processing of 454 reads
- 2. Remove chimera
- 3. Cluster/ASV filters
- 4. Taxonomic affiliation
- 5. Phylogenetic tree building
- ITSx

Paired-end or Single-end reads

- (●) Paired-end reads
- Single-end reads or Paired reads that have already been merg

Input

TAR Archive

Sample files can be provided either as a single TAR archive or as

Archive file (.tar.gz) *

1: MPC_ITS.tar.gz

accepted formats

The TAR file containing the short R1 R2 read pairs (.fastq.gz) for e
Aviti and IonTorrent. (--input-archive)

R1 read length *

250

Please provide the maximum length of the R1 reads. (--R1-size)

R2 read length *

250

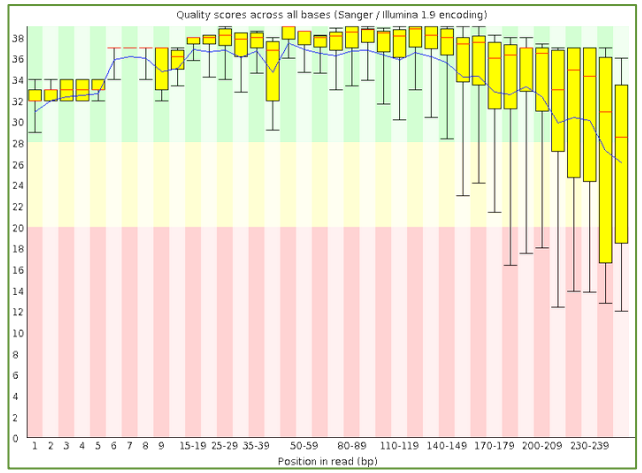
Please provide the maximum length of the R2 reads. (--R2-size)

Mismatch rate (used for R1-R2 merging) *

0,1

If your sequences are of low quality, you can increase this parameter.
But be careful!

Use FASTQC (included in Galaxy Tools) to check the quality of the sequence.



Practice session

ITS

Select a tool from the FROGS Core suite to run your analysis.

- -- Please select a tool --
- (•) 1.a. Reads processing of short reads
- 1.b. Reads processing of long reads
- 1.c. Reads processing of 454 reads
- 2. Remove chimera
- 3. Cluster/ASV filters
- 4. Taxonomic affiliation
- 5. Phylogenetic tree building
- ITSx

Paired-end or Single-end reads

- (•) Paired-end reads
- Single-end reads or Paired reads that have already been merg

Input

TAR Archive

Sample files can be provided either as a single TAR archive or as

Archive file (.tar.gz) *

1: MPC_ITS.tar.gz

accepted formats

The TAR file containing the short R1 R2 read pairs (.fastq.gz) for e
Aviti and IonTorrent. (--input-archive)

R1 read length *

250

Please provide the maximum length of the R1 reads. (--R1-size)

R2 read length *

250

Please provide the maximum length of the R2 reads. (--R2-size)

Mismatch rate (used for R1-R2 merging) *

0,1

Paired-end merging tool

- (•) Vsearch
- Flash

There is a third method, PEAR, but it is only available via the command line.

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

Would you like to keep unmerged reads?

Yes

No = unmerged reads will be removed; Yes = unmerged reads will be artificially combined with 100 N to allow further processing. (keep_unmerged)

Minimum amplicon length *

150

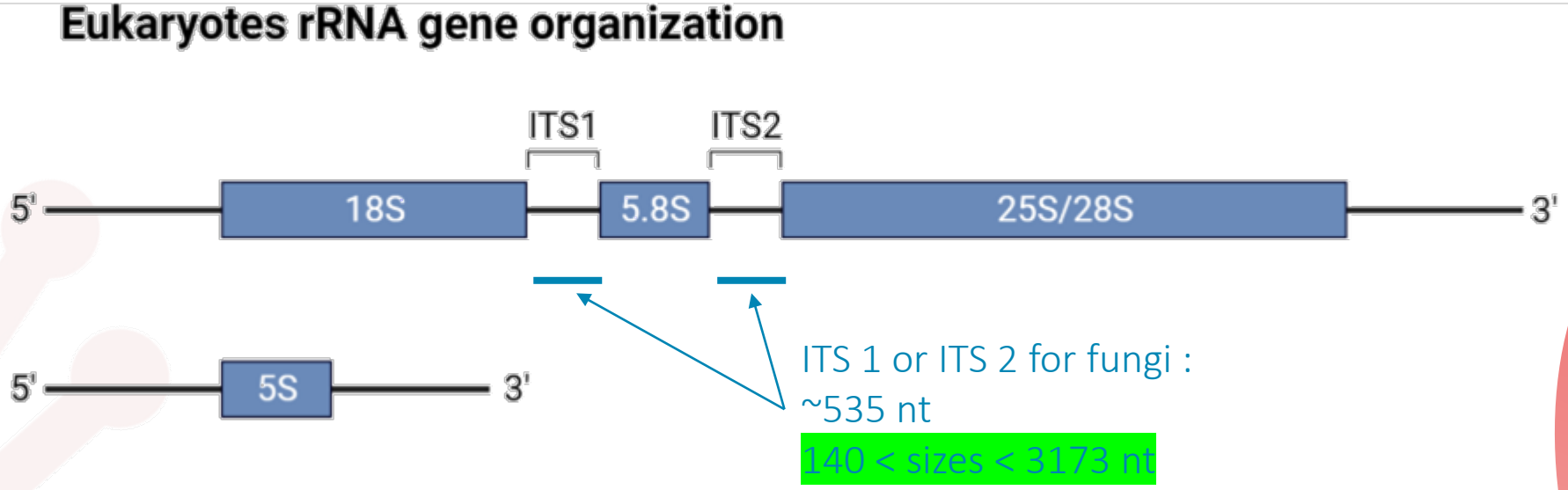
The minimum length of the amplicons (including primers). For paired-end reads, substract 10 bases to account for the minimum overlap between R1 and R2 reads. (--min-amplicon-size)

Maximum amplicon length *

480

The maximum length of the amplicons (including primers). For paired-end reads, substract 10 bases to account for the minimum overlap between R1 and R2 reads. (--max-amplicon-size)

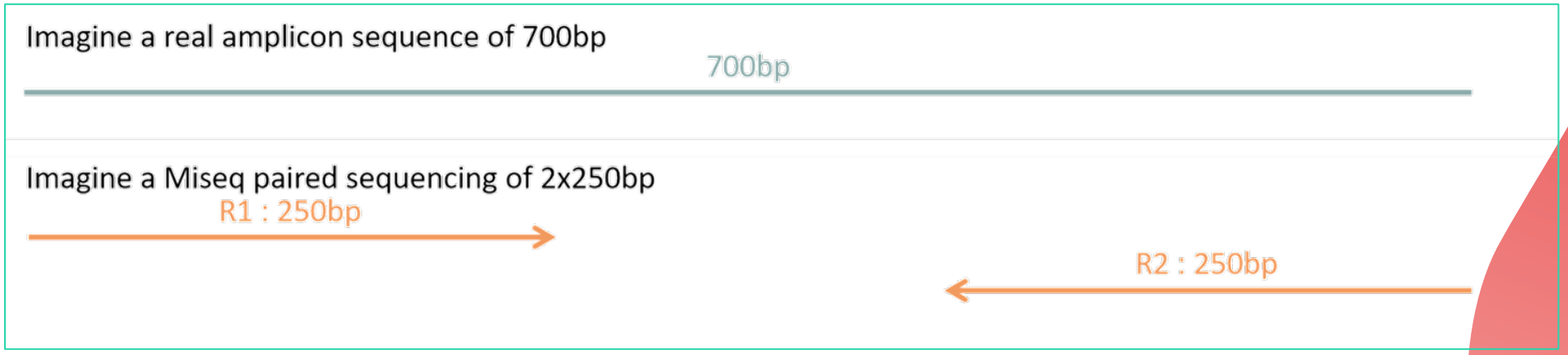
Problematic:
some ITS reads (short-read sequencing) are non-overlapping sequences



ITS

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

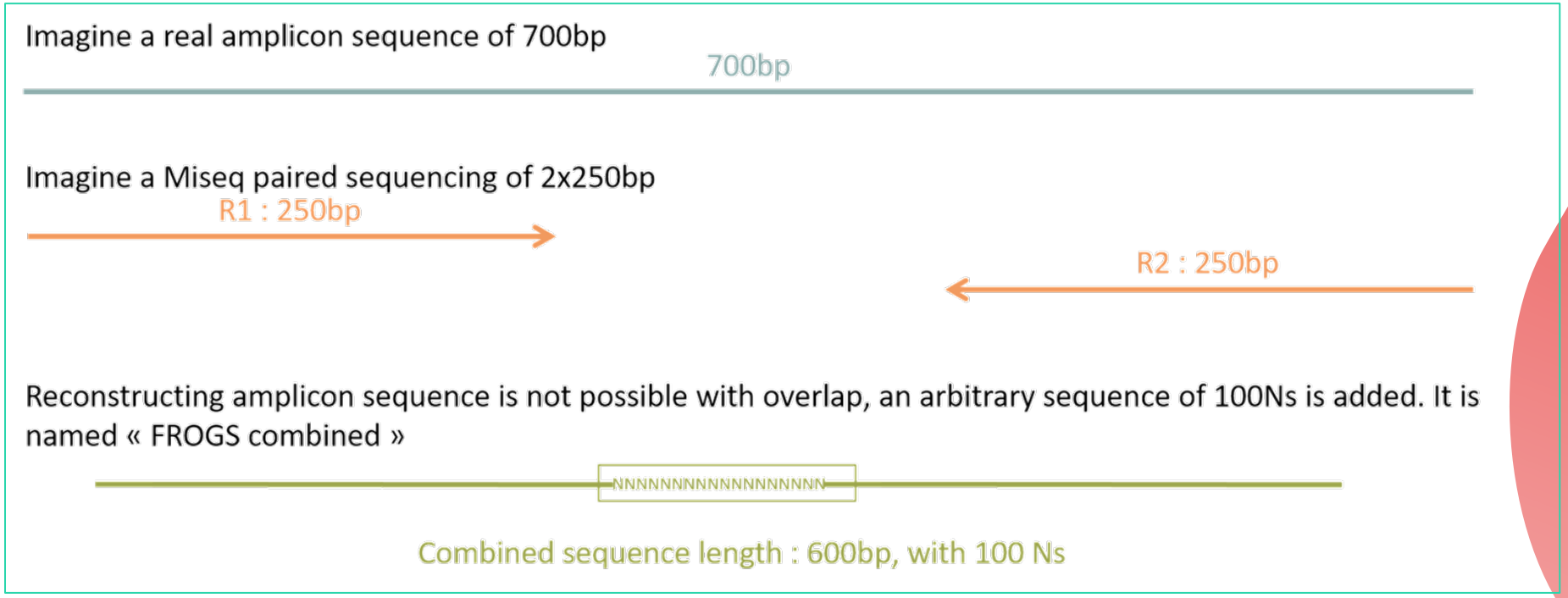
ITS



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

Solution: creation of "FROGS combined" sequences

ITS



Practice session

ITS

Select a tool from the FROGS Core suite to run your analysis.

- -- Please select a tool --
- (•) 1.a. Reads processing of short reads
- 1.b. Reads processing of long reads
- 1.c. Reads processing of 454 reads
- 2. Remove chimera
- 3. Cluster/ASV filters
- 4. Taxonomic affiliation
- 5. Phylogenetic tree building
- ITSx

Paired-end or Single-end reads

- (•) Paired-end reads
- Single-end reads or Paired reads that have already been merg

Input

TAR Archive

Sample files can be provided either as a single TAR archive or as

Archive file (.tar.gz) *

1: MPC_ITS.tar.gz

accepted formats

The TAR file containing the short R1 R2 read pairs (.fastq.gz) for e
Aviti and IonTorrent. (--input-archive)

R1 read length *

250

Please provide the maximum length of the R1 reads. (--R1-size)

R2 read length *

250

Please provide the maximum length of the R2 reads. (--R2-size)

Mismatch rate (used for R1-R2 merging) *

0,1

Paired-end merging tool

- (•) Vsearch
- Flash

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

Would you like to keep unmerged reads?

Yes

No = unmerged reads will be removed; Yes = unmerged reads will be artificially combined with 100 N to allow further processing. (keep_unmerged)

Minimum amplicon length *

150

The minimum length of the amplicons (including primers). For paired-end reads, substract 10 bases to account for the minimum overlap between R1 and R2 reads. (--min-amplicon-size)

Maximum amplicon length *

480

The maximum length of the amplicons (including primers). For paired-end reads, substract 10 bases to account for the minimum overlap between R1 and R2 reads. (--max-amplicon-size)

Do the sequences include PCR primers?

- (•) Yes
- No

Indicate whether the sequences still include PCR primers. Select "Yes" if primers are present, "No" if they have already been removed.

5' primer - optional

GCATCGATGAAGAACGCAGC

Enter the 5' primer sequence. Wildcards are a

3' primer - optional

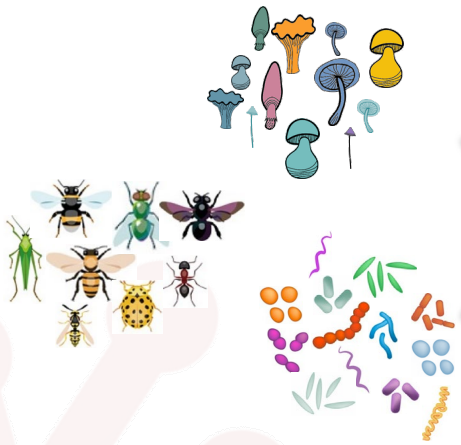
GCAWAWCAAWAAGCGGAGGA

Enter the 3' primer sequence. Wildcards are a

Similarly, you have to reverse transcribed the R2 primer
Use revseq: https://www.bioinformatics.nl/cgi-bin/emboss/revseq

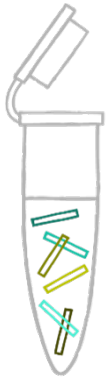
5' PCR primer: 5' GCATCGATGAAGAACGCAGC 3'
3' PCR primer: 5' GCAWAWCAAWAAGCGGAGGA 3'

Why do we need to group sequences together?



Community analysis

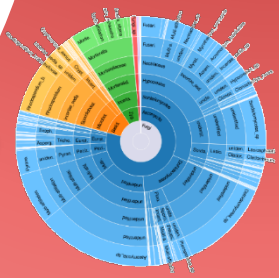
DNA extraction



Barcode amplification

Amplicon sequencing

Bioinformatics process



Diversity characterisation

Biological sampling

Collection methods

Collection methods

PCR biais

Length of reads

Software choice

Natural variability

Primer choice

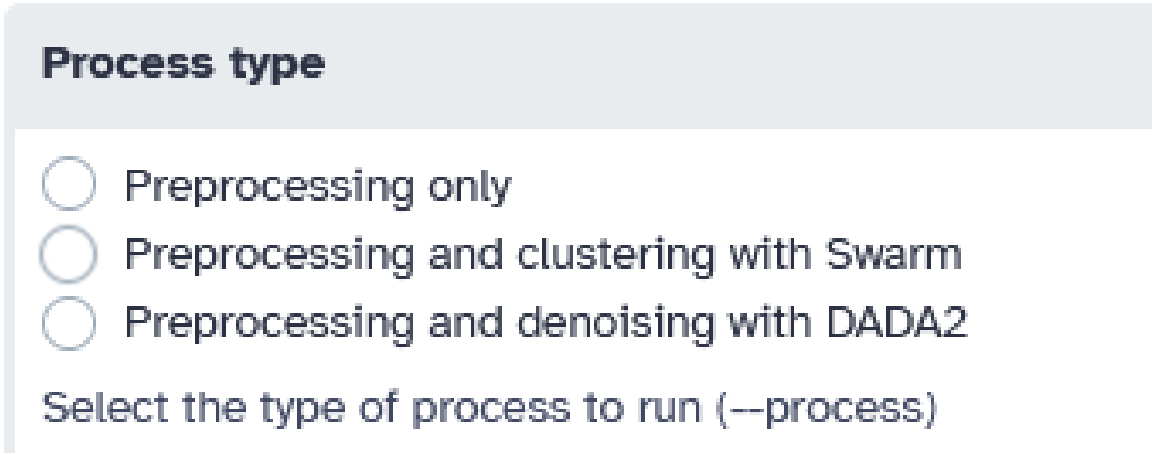
Sequencing protocol

Parameter/filter choice

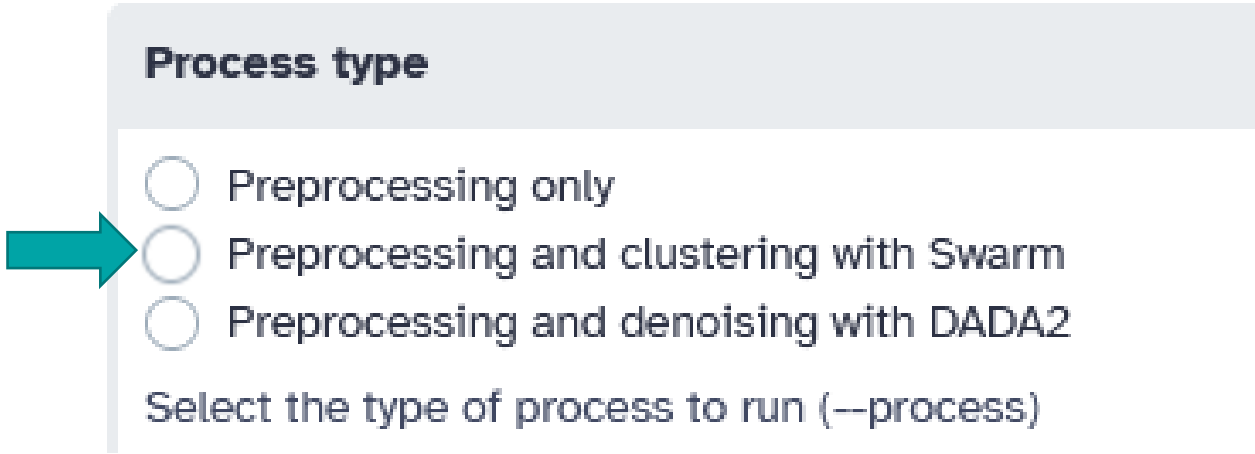
Practice session

Which methods for grouping sequences are available in the settings panel?

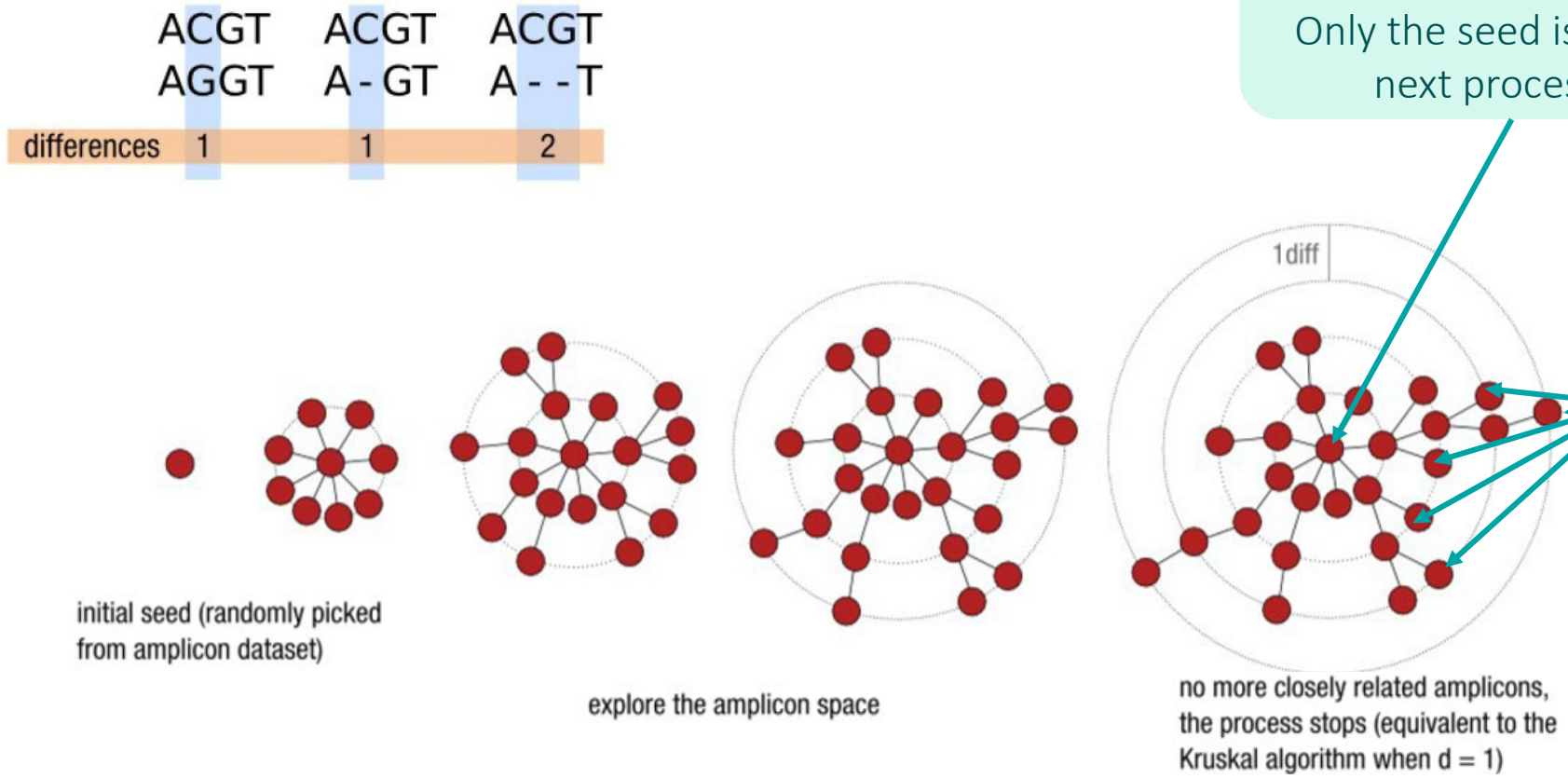
2 methods: Swarm or DADA2



2 methods: Swarm or DADA2



Swarm clustering method



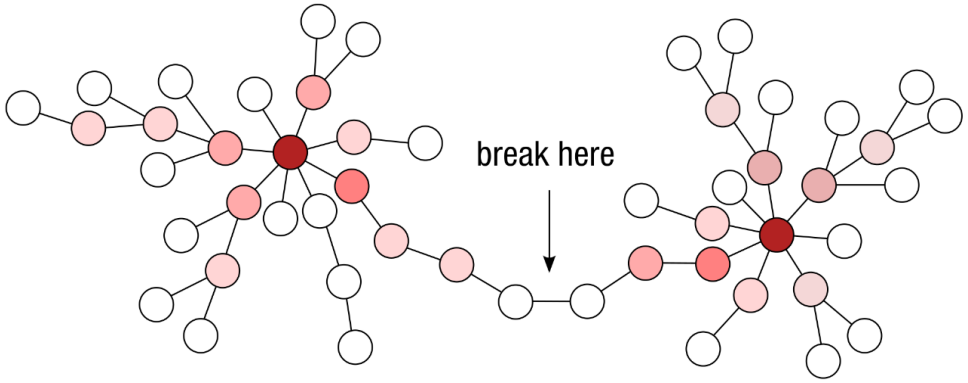
This sequences is the seed of the cluster.
Only the seed is kept for next processes.

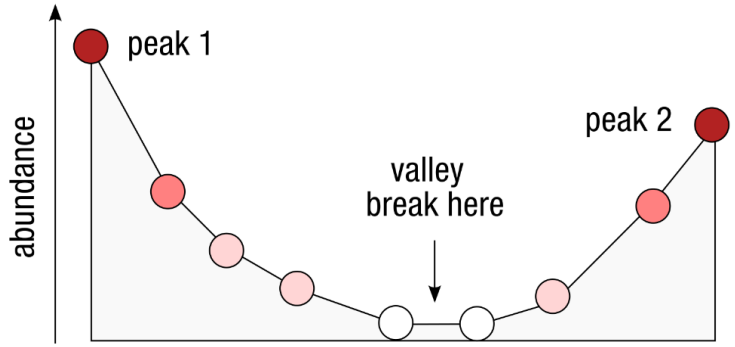
The abundances of each sequence in the cluster are added together.
And the total abundance is given to the seed.

Swarm v2: highly-scalable and high-resolution amplicon clustering. Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. PeerJ. 2015 Dec 10;3:e1420. doi: 10.7717/peerj.1420. eCollection 2015. PMID: 26713226

Swarm clustering method: breaking step

Close clusters can be linked 'by chance' if there is a path of errors between them.





Errors have decreasing abundances, so these situations are easily identified based on the abundance of each sequence (here represented with a red scale).
The breaking step splits merged clusters into natural clusters.

Swarm v2: highly-scalable and high-resolution amplicon clustering. Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. PeerJ. 2015 Dec 10;3:e1420. doi: 10.7717/peerj.1420. eCollection 2015. PMID: 26713226

Swarm options

Process type

- ○ Preprocessing only
- ◉ Preprocessing and clustering with Swarm
- ○ Preprocessing and denoising with DADA2

Select the type of process to run (--process)

Swarm distance threshold *

1

Distance threshold used by Swarm for clustering. (--distance)

Clustering refinement

- ◉ With --distance = 1, refine clusters with Swarm --fastidious option (recommended since FROGS 3.2)
- ○ With --distance > 1, perform a pre-clustering step with FROGS --pre-clustering option
- ○ No clustering refinement

(i) With --distance = 1, use the Swarm --fastidious option to refine clustering (recommended since FROGS 3.2). (ii) With --distance > 1, enable pre-clustering to reduce redundancy before final clustering step. (iii) Select this option to apply neither refinement nor pre-clustering.

Swarm options

Process type

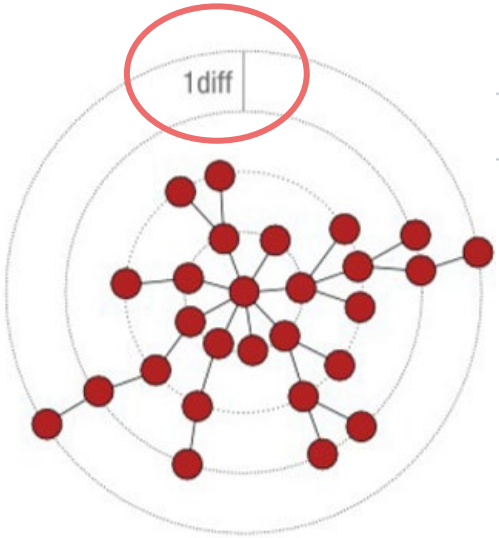
- Preprocessing only
- Preprocessing and clustering with Swarm
- Preprocessing and denoising with DADA2

Select the type of process to run (--process)

Swarm distance threshold *

1

Distance threshold used by Swarm for clustering. (--distance)



Process type

- Preprocessing only
- Preprocessing and clustering with Swarm
- Preprocessing and denoising with DADA2

Select the type of process to run (--process)

Swarm distance threshold *

1

Recommended

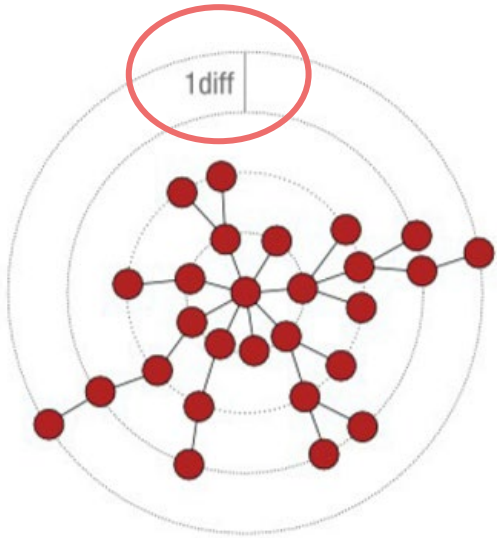
Distance threshold used by Swarm for clustering. (--distance)

Clustering refinement

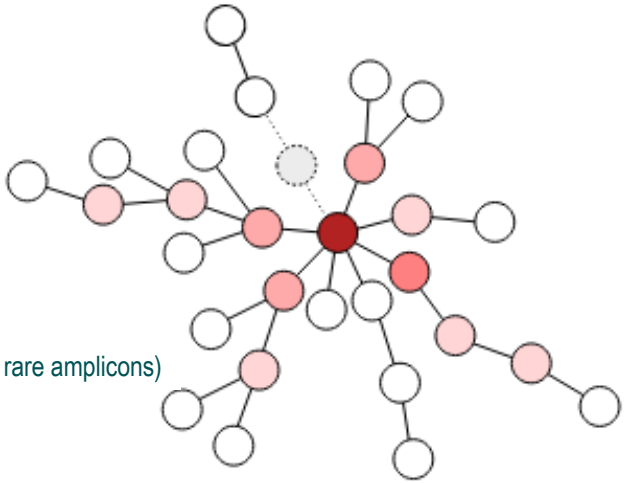
- With --distance = 1, refine clusters with Swarm --fastidious option (recommended since FROGS 3.2)
- With --distance > 1, perform a pre-clustering step with FROGS --pre-clustering option
- No clustering refinement

Recommended

(i) With --distance = 1, use the Swarm --fastidious option to refine clustering (recommended since FROGS 3.2). (ii) With --distance > 1, enable pre-clustering to reduce redundancy before final clustering step. (iii) Select this option to apply neither refinement nor pre-clustering.



longer but more accurate



small cluster (made of 2 rare amplicons)

virtual amplicon

Swarm options

Process type

- Preprocessing only
- Preprocessing and clustering with Swarm
- Preprocessing and denoising with DADA2

Select the type of process to run (--process)

Swarm distance threshold *

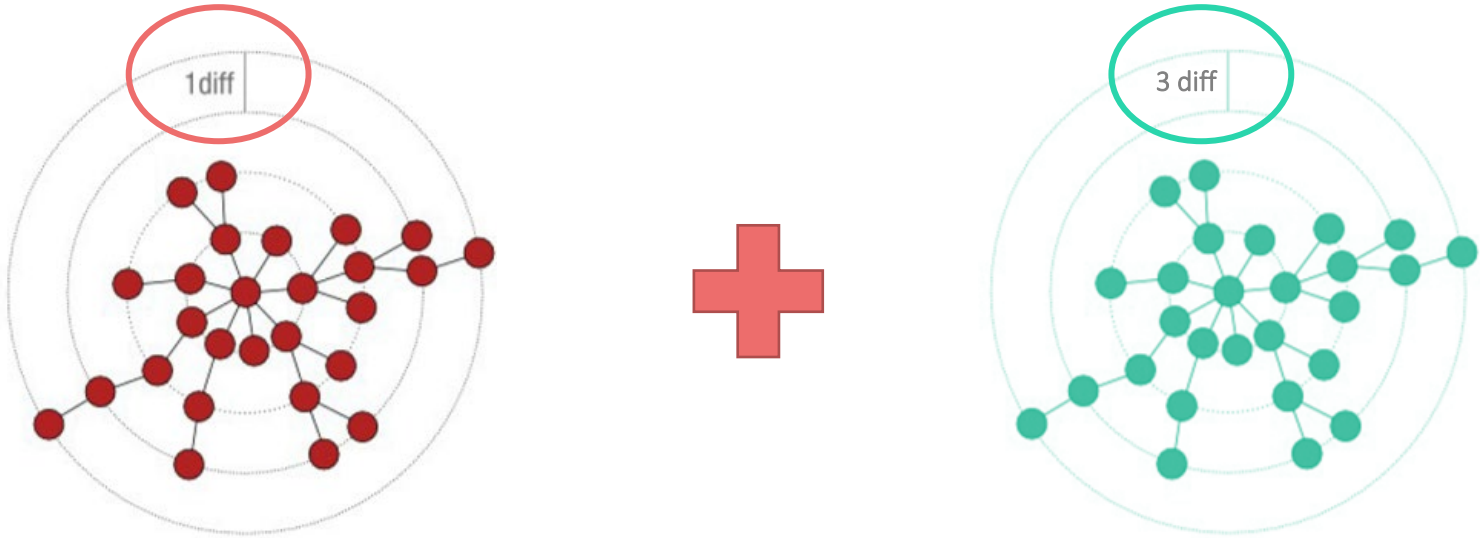
3

Distance threshold used by Swarm for clustering. (--distance)

Clustering refinement

- With --distance = 1, refine clusters with Swarm --fastidious option (recommended since FROGS 3.2)
- With --distance > 1, perform a pre-clustering step with FROGS --pre-clustering option
- No clustering refinement

(i) With --distance = 1, use the Swarm --fastidious option to refine clustering (recommended since FROGS 3.2). (ii) With --distance > 1, enable pre-clustering to reduce redundancy before final clustering step. (iii) Select this option to apply neither refinement nor pre-clustering.



Swarm options

Process type

- ( ) Preprocessing only
- (•) Preprocessing and clustering with Swarm
- ( ) Preprocessing and denoising with DADA2

Select the type of process to run (--process)

Swarm distance threshold *

1

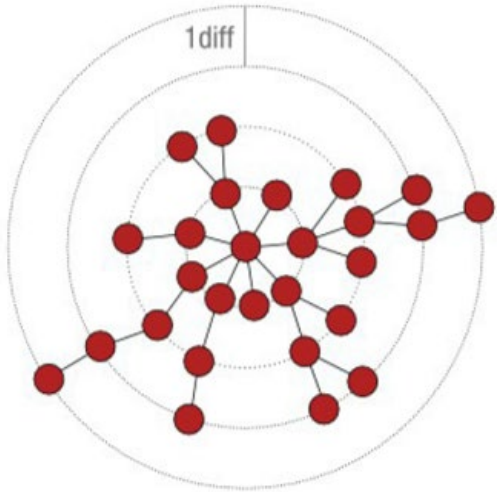
Distance threshold used by Swarm for clustering. (--distance)

Clustering refinement

- ( ) With --distance = 1, refine clusters with Swarm --fastidious option (recommended since FROGS 3.2)
- ( ) With --distance > 1, perform a pre-clustering step with FROGS --pre-clustering option
- (•) No clustering refinement

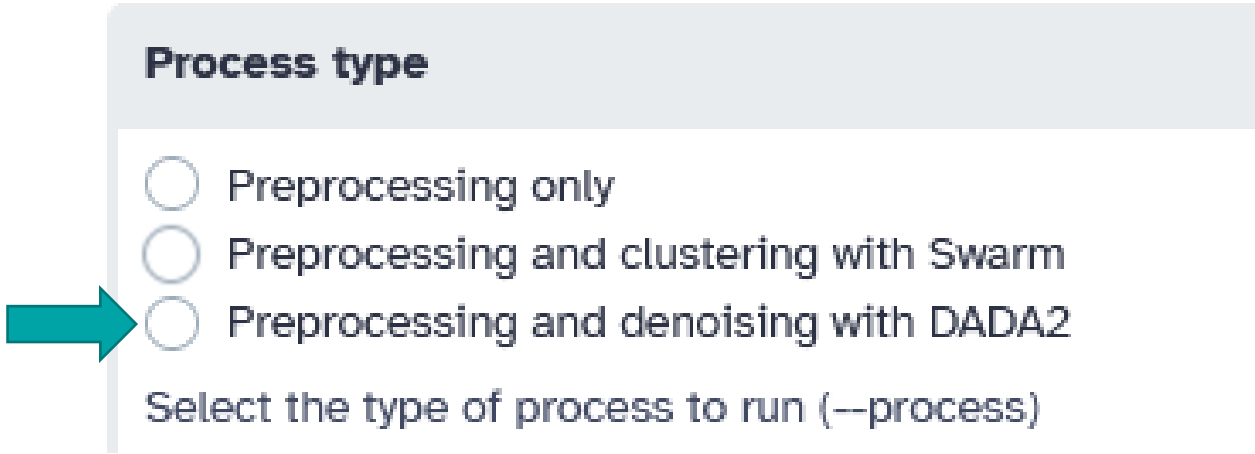
(i) With --distance = 1, use the Swarm --fastidious option to refine clustering (recommended since FROGS 3.2). (ii) With --distance > 1, enable pre-clustering to reduce redundancy before final clustering step. (iii) Select this option to apply neither refinement nor pre-clustering.

A single clustering



Swarm options

2 methods: Swarm or DADA2



DADA2 denoising

It is a complete workflow with a new, original approach to regrouping sequences.

It uses an error model that incorporates read quality information, and estimate the probability that a low-abundance read is an error derived from a more abundant read (incorporating specific base-transitions probabilities, computed on the dataset).

DADA2 relies on two main parameters:

The λ rate at which an amplicon of sequence *i* is produced from sequence *j*, as a function of sequence composition, base transition probabilities and base qualities.

$$\lambda_{ji} = \prod_{l=0}^{L} p(j(l) \rightarrow i(l), q_i(l))$$

According to quality score

Probability of transition between *i* and *j* bases at this position

For each base (from 1 to length L)

The p-value that the abundance of sequence *i* is too abundant to be obtained by sequencing errors of *j* (of abundance n_j).

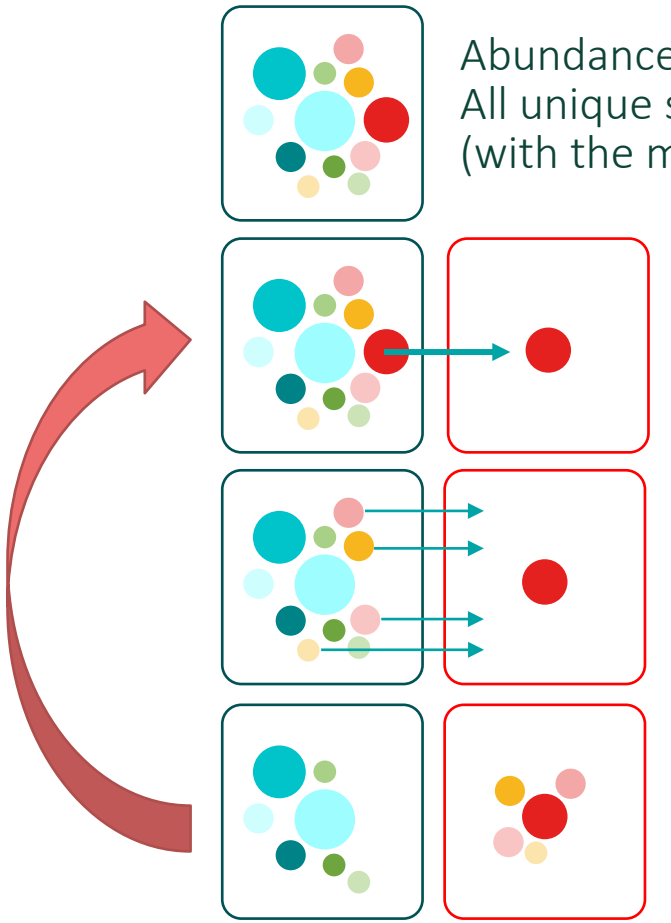
$$p_A(j \rightarrow i) = \frac{1}{1 - \rho_{pois}(n_j \lambda_{ji}, 0)} \sum_{a=a_i}^{\infty} \rho_{pois}(n_j \lambda_{ji}, a)$$

Parent sequence abundance x error rate

For singletons reads $p_A = 1$

A low p_A value indicates that the number of reads of sequence *i* exceeds the number that could be explained by errors introduced during the amplification and sequencing of n_j copies of sequence *j*.

DADA2 denoising - Partitioning



Abundances and consensus quality profile are calculated for each unique sequence.
All unique sequences are placed into a single partition
(with the most abundant being at the centre).

All unique sequences are compared to the centre
(error rates λ and p-value p_A are calculated).

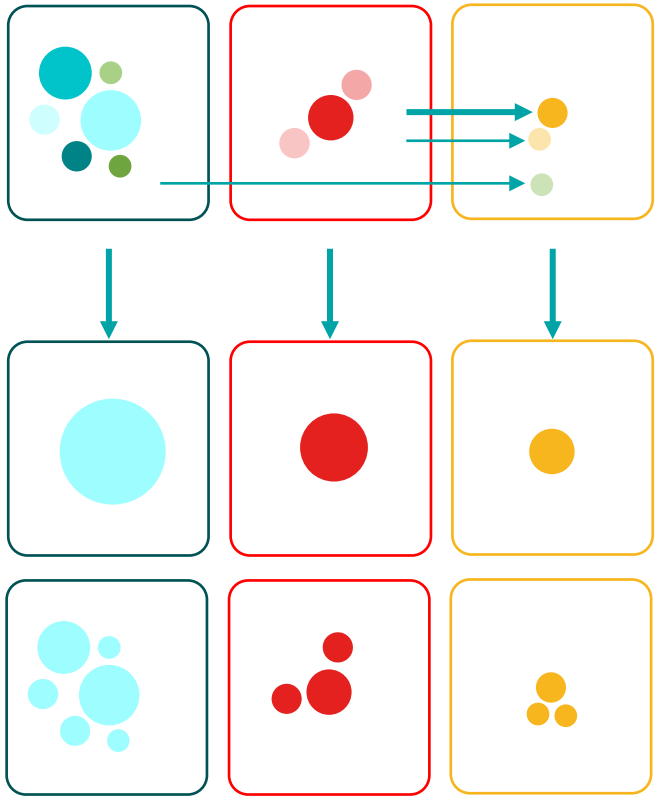
The sequence with the smallest p_A (if smaller than Ω_A) is used to form a new partition.

All unique sequences are compared to this new centre
and attributed to the partition that maximise n x λ.

All unique sequences are compared to their centre
(error rates λ and p-value p_A are calculated).

Sequence with the smallest p_A (if smaller than Ω_A) is used to form a new partition.

DADA2 denoising - Partitioning



For singleton reads, $p_A = 1$ so they can never form a new partition and are attributed to the partition maximising n x λ (the least unlikely).
Therefore, there is no partition with a unique sequence of abundance 1.

Partitioning is over, when for all non-singleton sequences $p_A < \Omega_A$,.

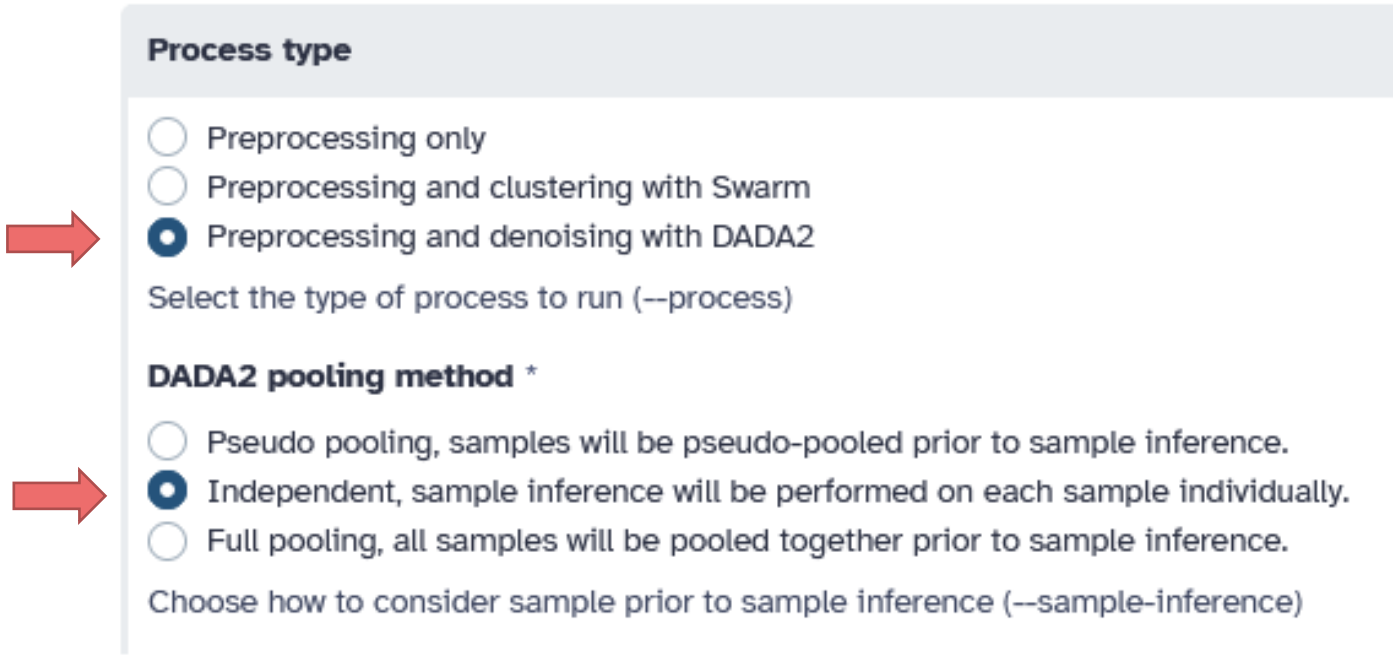
The inferred composition of the sample is the set of central sequences (ASVs), and the corresponding abundances are the sum of their partitions.

or

or each read is denoised and replaced by the central sequence of its partition.

DADA2 initially infers sequences and abundances **of each sample.**
And the abundance tables for each sample are then merged using the ASV sequences as joining key.

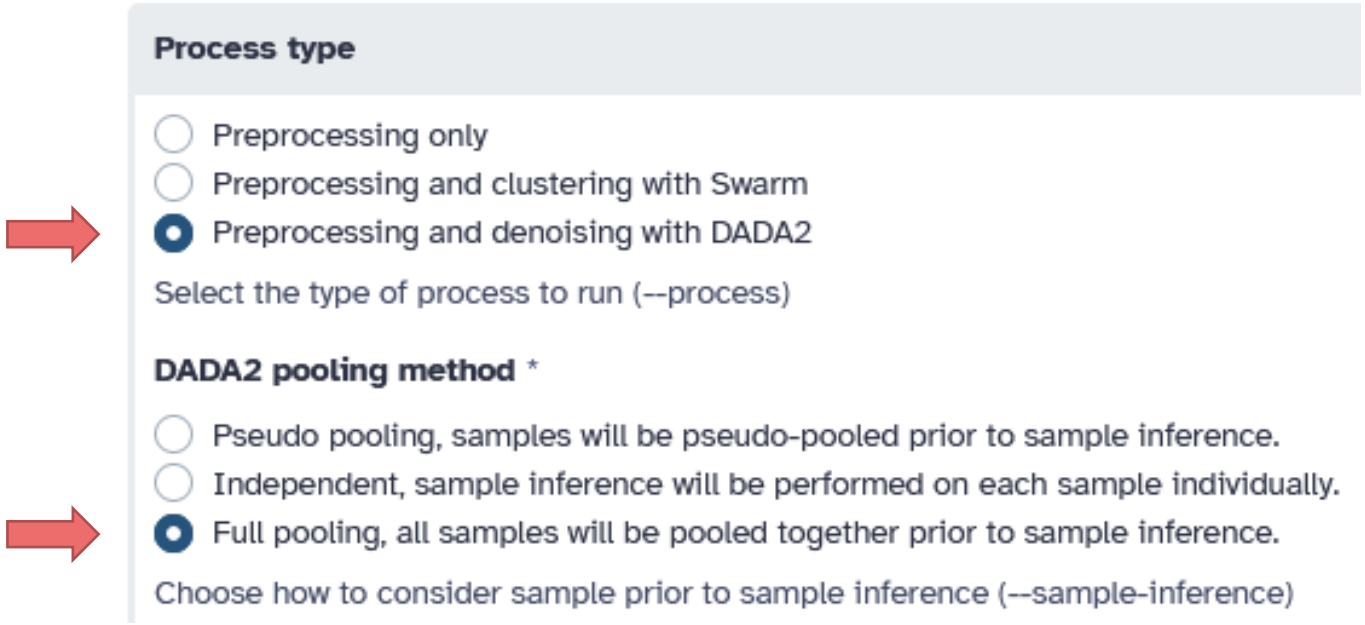
DADA2 options



Initially, independent treatment of samples was the only option. However, this approach has some limitations.

- By construction, it was impossible to observe an abundance of 1 for any ASV.
- Singleton reads are classified in the 'least likely' partition, with potentially significant sequence differences, even though, in some cases, the same sequence exists (and is abundant) in other samples.

DADA2 options

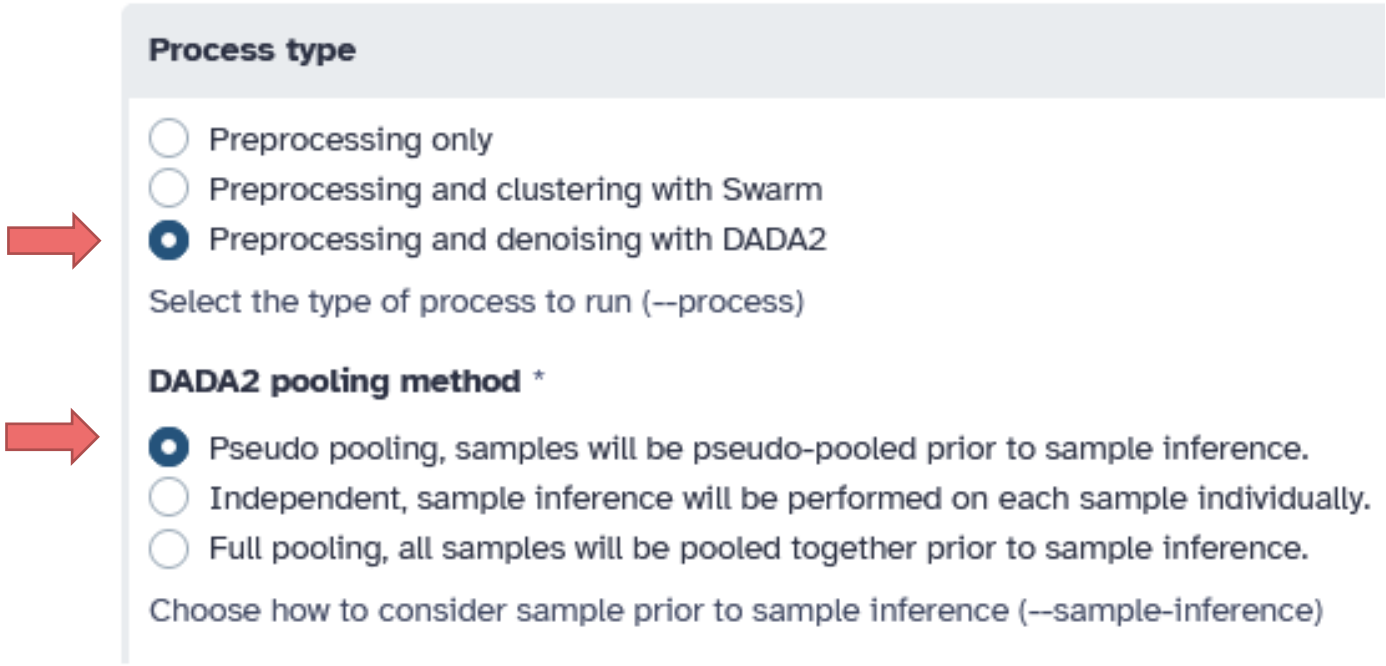


One solution to better reconstruct rare sequences in samples using information from other samples is to pool the reads from all samples (as is done with Swarm).

Partitioning is therefore performed on all samples' reads, and reads that were singletons in one sample are no longer considered as such if they were also present in another sample.

However, full pooling becomes non-linearly scalable with increasing sample numbers, and computation time increases exponentially with large datasets.

DADA2 options



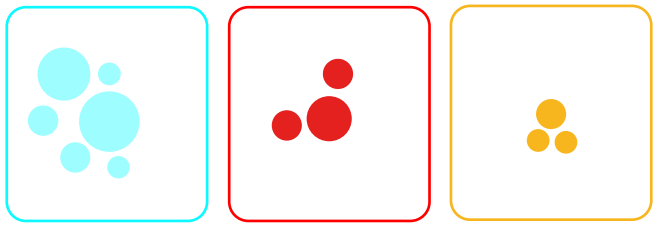
DADA2 offers an alternative 'pseudo-pooling' approach, in which samples are first treated independently before a second treatment using all ASVs identified in the first round as 'priors' (the initial centres of partitions with an abundance of 1).

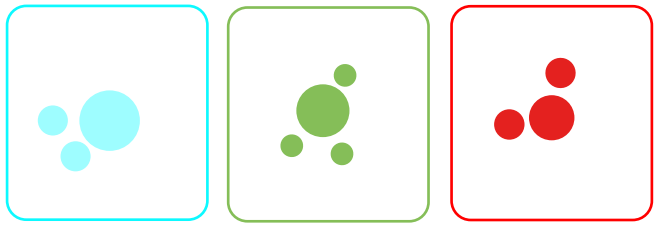
DADA2 pseudo-pooling

Sample A

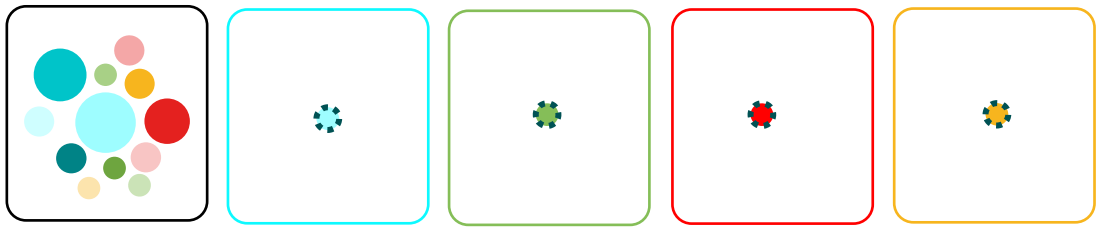
Sample B

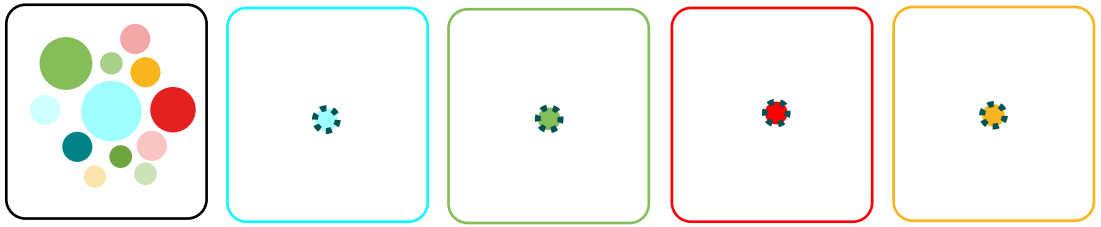
1st round



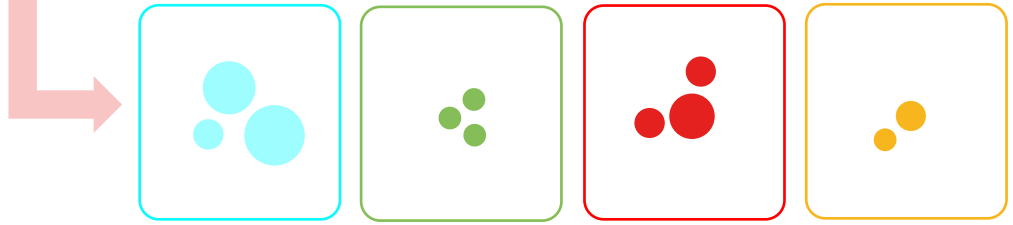


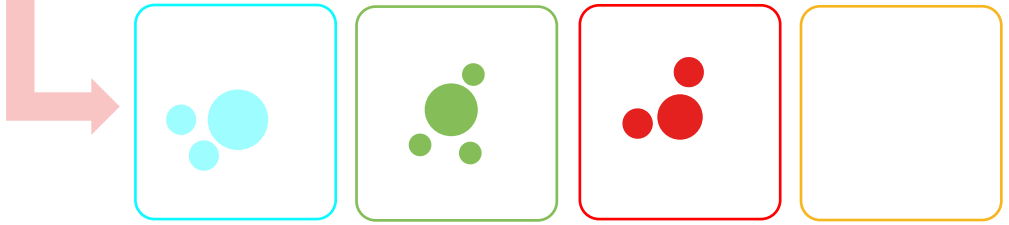
2nd round





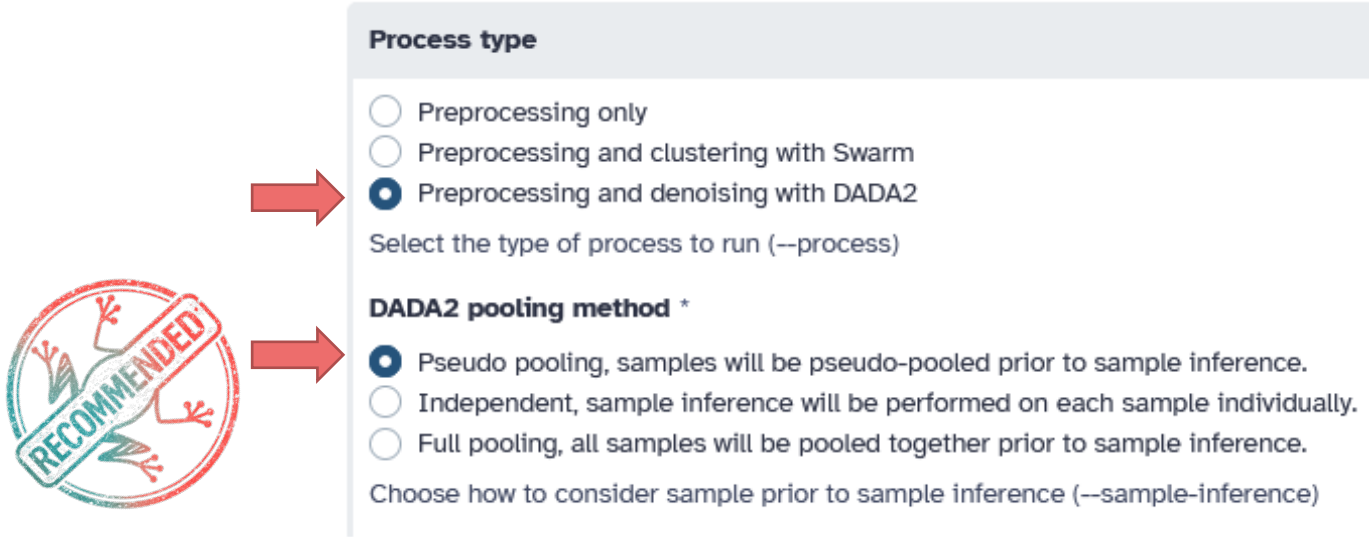
All the unique sequences are compared with the centre of each partition and assigned to the partition that maximises n x λ.

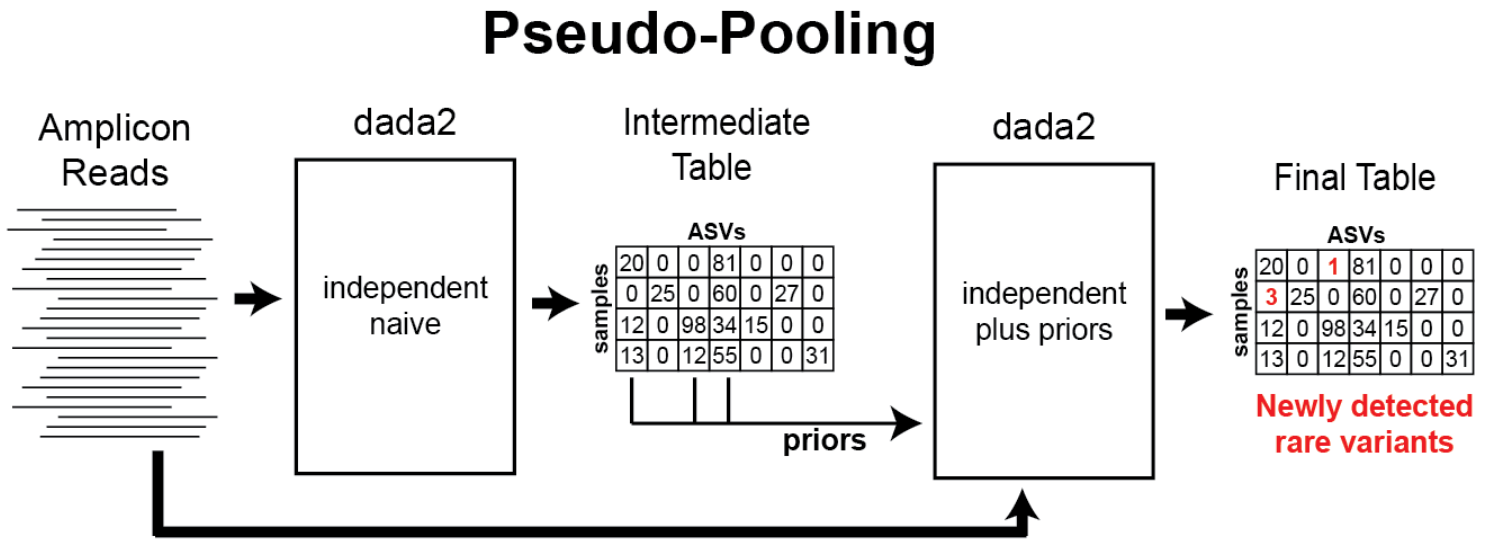




The ● ASV is observable in Sample A thanks to its abundance in Sample B.

DADA2 options





Practice session

Choose swarm or DADA2

For swarm
Swarm distance threshold = 1
With --distance = 1, refine clusters with Swarm --fastidious option

For DADA2
Pseudo pooling, samples will be pseudo-pooled prior to sample inference.

Run the process !

Practice session

What are the outputs ?

FROGS Core Main 1 outputs

5: **FROGS Core 1-Main - processing_short_reads: abundance.biom**

4: **FROGS Core 1-Main - processing_short_reads: sequence.fasta**

3: **FROGS Core 1-Main - processing_short_reads: report.html**

Biological Observation Matrix (BIOM) but not human readable

{"comment": ["FROGS_combined"], "seed_id": "None"}}, {"id": "ID_2650", "metadata": {"comment": [], "seed_id": "None"}}, {"id": "ID_2651", "metadata": {"comment": [], "seed_id": "None"}}, {"id": "ID_2652_FROGS_combined", "metadata": {"comment": ["FROGS_combined"], "seed_id": "None"}}, {"id": "ID_2653_FROGS_combined", "metadata": {"comment": ["FROGS_combined"], "seed_id": "None"}}, {"id": "ID_2654_FROGS_combined", "metadata": {"comment": ["FROGS_combined"], "seed_id": "None"}}, {"id": "ID_2655_FROGS_combined", "metadata": {"comment": ["FROGS_combined"], "seed_id": "None"}}, {"id": "ID_2656", "metadata": {"comment": [], "seed_id": "None"}}, {"id": "ID_2657", "metadata": {"comment": [], "seed_id": "None"}}], "columns": [{"id": "AOP1_PPC_S1", "metadata": null}, {"id": "AOP1_PPC_S2", "metadata": null}, {"id": "AOP1_PPC_S3", "metadata": null}, {"id": "AOP1_PPC_S4", "metadata": null}, {"id": "AOP1_PPC_S5", "metadata": null}, {"id": "AOP1_PPC_S6", "metadata": null}, {"id": "AOP1_PPC_W1", "metadata": null}, {"id": "AOP1_PPC_W2", "metadata": null}, {"id": "AOP1_PPC_W3", "metadata": null}, {"id": "AOP1_PPC_W4", "metadata": null}, {"id": "AOP1_PPC_W5", "metadata": null}, {"id": "AOP1_PPC_W6", "metadata": null}, {"id": "AOP2_PPC_S1", "metadata": null}, {"id": "AOP2_PPC_S2", "metadata": null}, {"id": "AOP2_PPC_S3", "metadata": null}, {"id": "AOP2_PPC_S4", "metadata": null}, {"id": "AOP2_PPC_S5", "metadata": null}, {"id": "AOP2_PPC_S6", "metadata": null}, {"id": "AOP2_PPC_W1", "metadata": null}, {"id": "AOP2_PPC_W2", "metadata": null}, {"id": "AOP2_PPC_W3", "metadata": null}, {"id": "AOP2_PPC_W4", "metadata": null}, {"id": "AOP2_PPC_W5", "metadata": null}, {"id": "AOP2_PPC_W6", "metadata": null}, {"id": "AOP3_PPNC_S1", "metadata": null}, {"id": "AOP3_PPNC_S2", "metadata": null}, {"id": "AOP3_PPNC_S3", "metadata": null}, {"id": "AOP3_PPNC_S4", "metadata": null}, {"id": "AOP3_PPNC_S5", "metadata": null}, {"id": "AOP3_PPNC_S6", "metadata": null}, {"id": "AOP3_PPNC_W1", "metadata": null}, {"id": "AOP3_PPNC_W2", "metadata": null}, {"id": "AOP3_PPNC_W3", "metadata": null}, {"id": "AOP3_PPNC_W4", "metadata": null}, {"id": "AOP3_PPNC_W5", "metadata": null}, {"id": "AOP3_PPNC_W6", "metadata": null}, {"id": "AOP4_PPNC_S1", "metadata": null}, {"id": "AOP4_PPNC_S2", "metadata": null}, {"id": "AOP4_PPNC_S3", "metadata": null}, {"id": "AOP4_PPNC_S4", "metadata": null}, {"id": "AOP4_PPNC_S5", "metadata": null}, {"id": "AOP4_PPNC_S6", "metadata": null}, {"id": "AOP4_PPNC_W1", "metadata": null}, {"id": "AOP4_PPNC_W2", "metadata": null}, {"id": "AOP4_PPNC_W3", "metadata": null}, {"id": "AOP4_PPNC_W4", "metadata": null}, {"id": "AOP4_PPNC_W5", "metadata": null}, {"id": "AOP4_PPNC_W6", "metadata": null}, {"id": "AOP5_PPS_S1", "metadata": null}, {"id": "AOP5_PPS_S2", "metadata": null}, {"id": "AOP5_PPS_S3", "metadata": null}, {"id": "AOP5_PPS_S4", "metadata": null}, {"id": "AOP5_PPS_S5", "metadata": null}, {"id": "AOP5_PPS_S6", "metadata": null}, {"id": "AOP5_PPS_W1", "metadata": null}, {"id": "AOP5_PPS_W2", "metadata": null}, {"id": "AOP5_PPS_W3", "metadata": null}, {"id": "AOP5_PPS_W4", "metadata": null}, {"id": "AOP5_PPS_W5", "metadata": null}, {"id": "AOP5_PPS_W6", "metadata": null}, {"id": "AOP6_PPS_S1", "metadata": null}, {"id": "AOP6_PPS_S2", "metadata": null}, {"id": "AOP6_PPS_S3", "metadata": null}, {"id": "AOP6_PPS_S4", "metadata": null}, {"id": "AOP6_PPS_S5", "metadata": null}, {"id": "AOP6_PPS_S6", "metadata": null}, {"id": "AOP6_PPS_W1", "metadata": null}, {"id": "AOP6_PPS_W2", "metadata": null}, {"id": "AOP6_PPS_W3", "metadata": null}, {"id": "AOP6_PPS_W4", "metadata": null}, {"id": "AOP6_PPS_W5", "metadata": null}, {"id": "AOP6_PPS_W6", "metadata": null}], "matrix_element_type": "int", "data": [[0, 0, 57], [0, 2, 5], [0, 3, 11], [0, 4, 12], [0, 5, 6], [0, 6, 2], [0, 7, 15], [0, 8, 7], [0, 9, 9], [0, 11, 1], [0, 12, 19], [0, 13, 22], [0, 14, 15], [0, 15, 288], [0, 16, 297], [0, 17, 518], [0, 18, 691], [0, 19, 591], [0, 20, 1123], [0, 21, 17], [0, 22, 9], [0, 23, 415], [0, 24, 19093], [0, 25, 45491], [0, 26, 23701], [0, 27, 14317], [0, 28, 15896], [0, 29, 16060], [0, 30, 28722], [0, 31, 19098], [0, 32, 23366], [0, 33, 17283], [0, 34, 18416], [0, 35, 30696], [0, 36, 32325], [0, 37, 29646], [0, 38, 15088], [0, 39, 21002], [0, 40, 21522], [0, 41, 17671], [0, 42, 16361], [0, 43, 20245], [0, 44, 18597], [0, 45, 17355], [0, 46, 17342], [0, 47, 20217], [0, 48, 14], [0, 49, 8], [0, 50, 9], [0, 51, 51], [0, 52, 14], [0, 53, 11], [0, 54, 45], [0, 55, 13], [0, 56, 12], [0, 57, 9], [0, 58, 14], [0, 59, 10], [0, 60, 429], [0, 61, 64], [0, 62, 25], [0, 63, 1029], [0, 64, 176], [0, 65, 634], [0, 66, 3575], [0, 67, 2280], [0, 68, 2027], [0, 69, 203], [0, 70, 330], [0, 71, 315], [1, 0, 3319], [1, 1, 3209], [1, 2, 3118], [1, 3, 8792], [1, 4, 6378], [1, 5, 6801], [1, 6, 5121], [1, 7, 3654], [1, 8, 4721], [1, 9, 3652], [1, 10, 3007], [1, 11, 3493], [1, 12, 8826], [1, 13, 7363], [1, 14, 11760], [1, 15, 5148], [1, 16, 5690], [1, 17, 6361], [1, 18, 2781], [1, 19, 12691], [1, 20, 14291], [1, 21, 48091], [1, 22, 54171], [1, 23, 42551], [1, 24, 1261], [1, 25, 3731], [1, 26, 1161], [1, 27, 31], [1, 28, 21], [1, 29, 21], [1, 30, 38331], [1

FROGS Core Main 1 outputs

5: **FROGS Core 1-Main - processing_short_reads: abundance.biom**

4: **FROGS Core 1-Main - processing_short_reads: sequence.fasta**

3: **FROGS Core 1-Main - processing_short_reads: report.html**

Dereplicated sequences in fasta format

```
>ID_1
GAAACGCGATATTTCTTGTGAATTGCAGAAGTGAATCATCAGTTTTTGAACGCACATTGCACTTTGGGGTATC
>ID_2
GAAATGCGATAAGTAATATGAATTGCAGATTTTCGTGAATCATCGAATCTTTGAACGCACATTGCGCCCTCTG
>ID_3
GAAATGCGATAAGTAATGTGAATTGCAGAATTCAGTGAATCATCGAATCTTTGAACGCACATTGCGCCCGCCA
>ID_4
GAAATGCGATAAGTAATATGAATTGCAGATTTTCGTGAATCATCGAATCTTTGAACGCACATTGCGCCCTCTG
>ID_5
GAAATGCGATAAGTAATATGAATTGCAGATTTTCGTGAATCATCGAATCTTTGAACGCACATTGCGCCCTTTG
>ID_6
GAAATGCGATACGTAATATGAATTGCAGATTTTGTGAATCATCGAATCTTTGAACGCACATTGCGCCCGTGGG
>ID_7
GAAATGCGATAAGTAATGTGAATTGCAGAATTCAGTGAATCATCGAATCTTTGAACGCACATTGCGCCCGGCA
>ID_8
GAAATGCGATAAGTAATGTGAATTGCAGAATTCAGTGAATCATCGAATCTTTGAACGCACATTGCGCCCGGCA
```

FROGS Core Main 1 outputs

6: **FROGS Core 1-Main - processing_short_reads: swarms_composition.txt**

5: **FROGS Core 1-Main - processing_short_reads: abundance.biom**

4: **FROGS Core 1-Main - processing_short_reads: sequence.fasta**

3: **FROGS Core 1-Main - processing_short_reads: report.html**

4th output, for swarm only !

```
H7:1:HTNVFBCX2:1:1101:11476:3603/1_75390 H7:1:HMJYJBCX2:1:2106:11793:25432/1_3575
H7:1:HTNVFBCX2:1:1206:8531:10452/1_2450 H7:1:HMJYJBCX2:1:2113:16449:56604/1_1362 H7:1:HLYVMBCX2:1:2204:20108:18134/1_667
H7:1:HTNVFBCX2:1:2202:10147:30926/1_582 H7:1:HTNVFBCX2:1:2208:6789:40431/1_507 H7:1:HK3WHBCX2:1:2115:17985:38977/1_506
H7:1:HTNVFBCX2:1:1105:6397:51971/1_360 H7:1:HTNVFBCX2:1:1201:19522:70187/1_323 H7:1:HLYVMBCX2:2:2101:14487:7701/1_276
H7:1:HTNVFBCX2:1:1206:8856:80379/1_238 H7:1:HTNVFBCX2:1:2107:13596:99755/1_175 H7:1:HTNVFBCX2:1:2216:9427:70516/1_159
H7:1:HK3WHBCX2:1:1101:20335:76786/1_156 H7:1:HK3WHBCX2:1:2115:20223:7357/1_151 H7:1:HTNVFBCX2:1:2102:12209:85146/1_148
H7:1:HK3WHBCX2:1:1208:17227:76625/1_140 H7:1:HTNVFBCX2:1:1109:11291:68314/1_140 H7:1:HTNVFBCX2:1:2206:9551:63931/1_135
H7:1:HLYVMBCX2:1:2213:4782:91658/1_128 H7:1:HTNVFBCX2:1:2103:19373:84826/1_120 H7:1:HTNVFBCX2:1:1111:10909:23437/1_117
H7:1:HTNVFBCX2:1:1216:7504:24008/1_115 H7:1:HLYVMBCX2:2:1107:20508:16134/1_113 H7:1:HTNVFBCX2:1:1210:15249:91692/1_113
H7:1:HTNVFBCX2:1:1205:7663:94775/1_111 H7:1:HTNVFBCX2:1:2102:8950:79851/1_111 H7:1:HTNVFBCX2:1:2114:9532:14031/1_111
```

FROGS Core Main 1 outputs

6: **FROGS Core 1-Main - processing_short_reads: swarms_composition.txt**

5: **FROGS Core 1-Main - processing_short_reads: abundance.biom**

4: **FROGS Core 1-Main - processing_short_reads: sequence.fasta**

3: **FROGS Core 1-Main - processing_short_reads: report.html**

Data visualization in html report

Reads-processing report *(reads_processing.py, v5.1.0)*

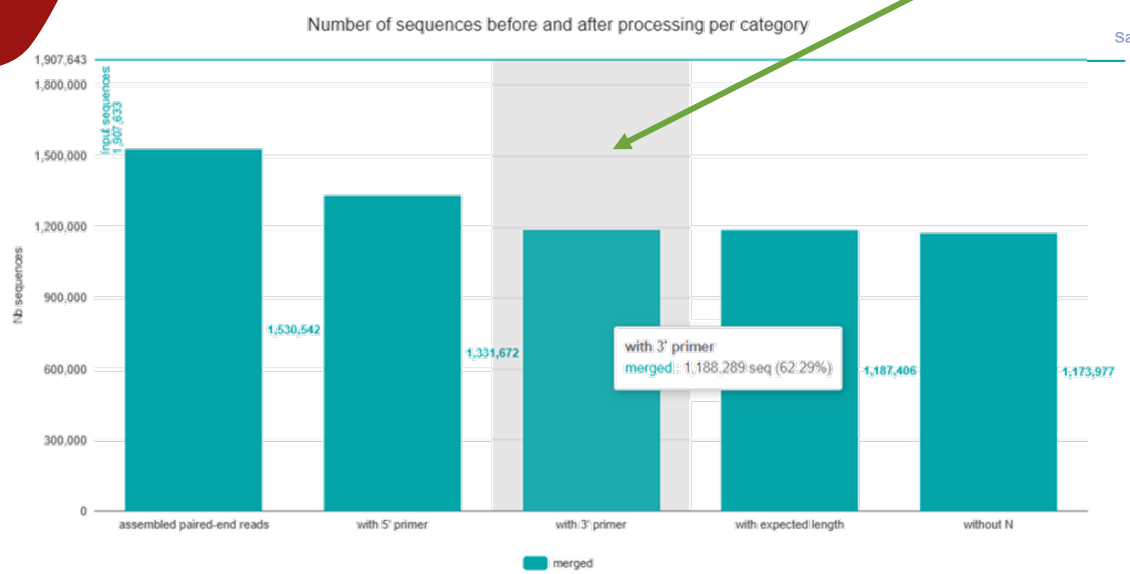
Practice session

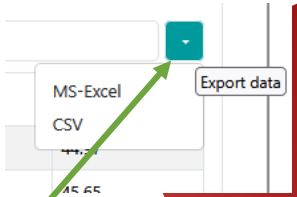
Explore report html: Summary tab

How many sequences are there in the input file ?
How many sequences did not have the 5′ primer ?
How many sequences remain after the data has been preprocessed ?
What is the length of your merged reads before pre-processing ?
Based on amplicon size distributions, what can you tell us about the samples ?

Tips !

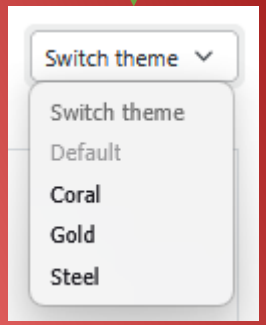
New information appears when you move the mouse over the graphic.





You can download graphics and tables.

You can change the theme color.



Sample name	Number of clusters/ASVs	Number of sha
AOP1_PPC_S1	3,081	655
AOP1_PPC_S2	2,438	566
AOP1_PPC_S3	1,923	362

You can sort the data in the table by clicking on the column headers.

PPNC

ut N	% kept
	53.05

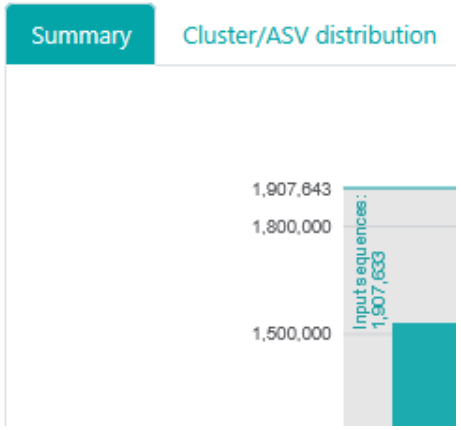
You can filter tables using a keyword.

Practice session

How many sequences are there in the input file ?

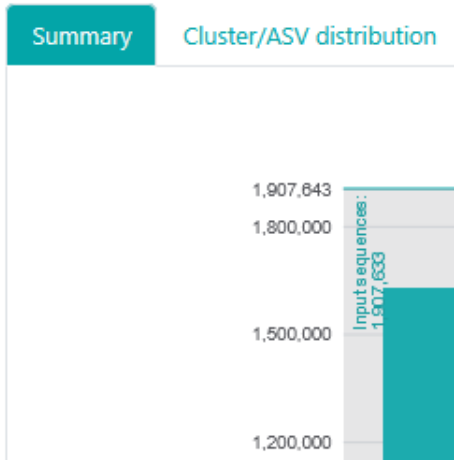
16S

swarm



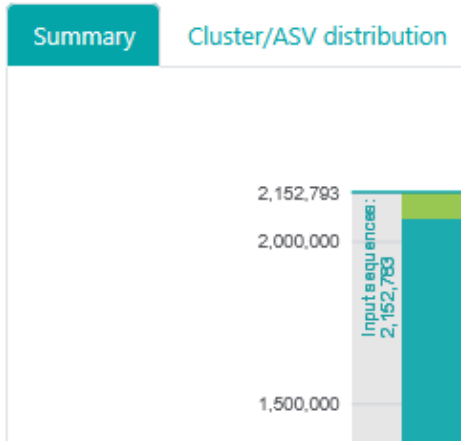
16S

DADA2



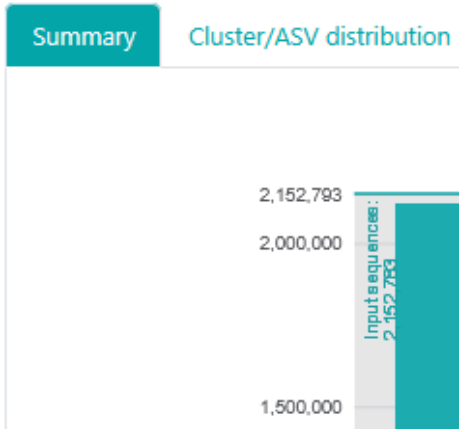
ITS

swarm



ITS

DADA2

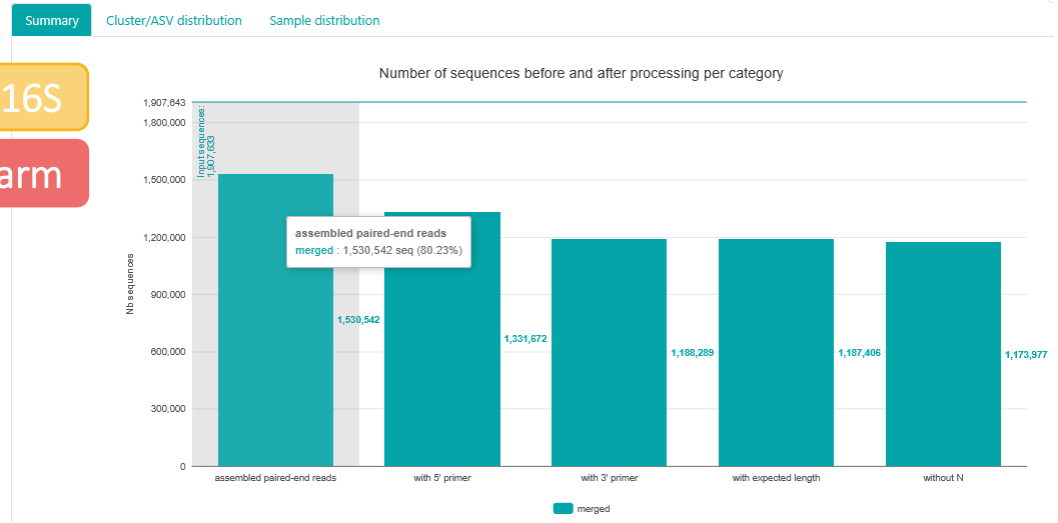


Practice session

How many sequences did not have the 5′ primer ?

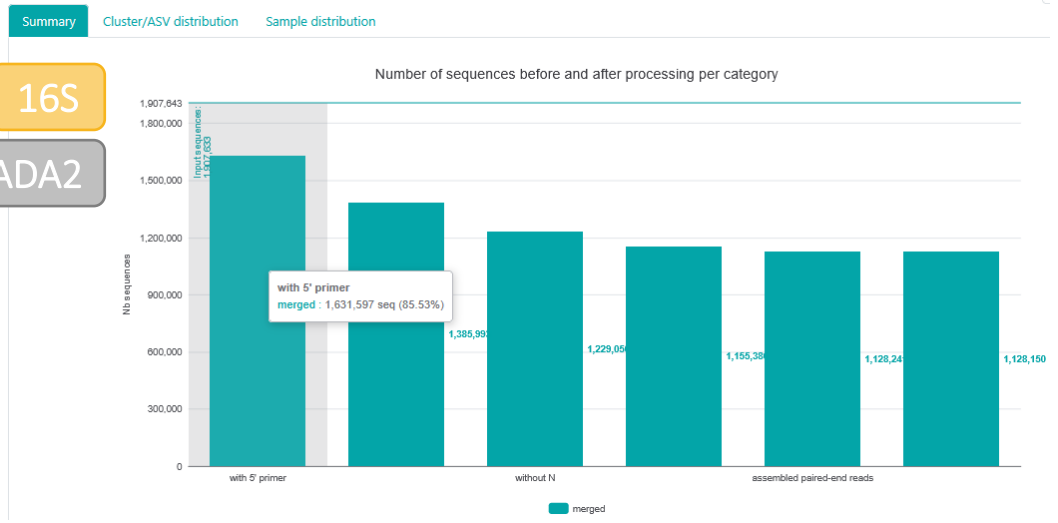
16S

swarm



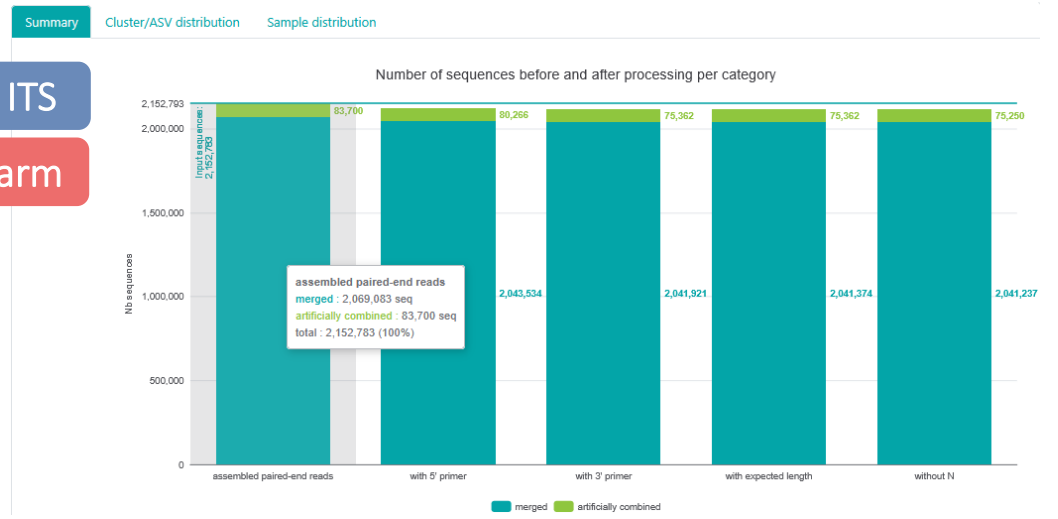
16S

DADA2



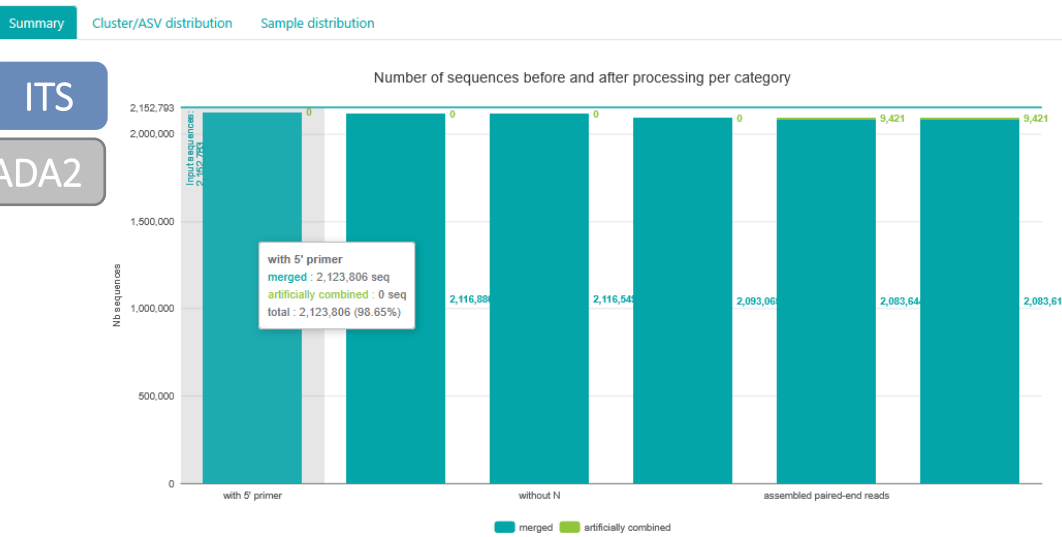
ITS

swarm



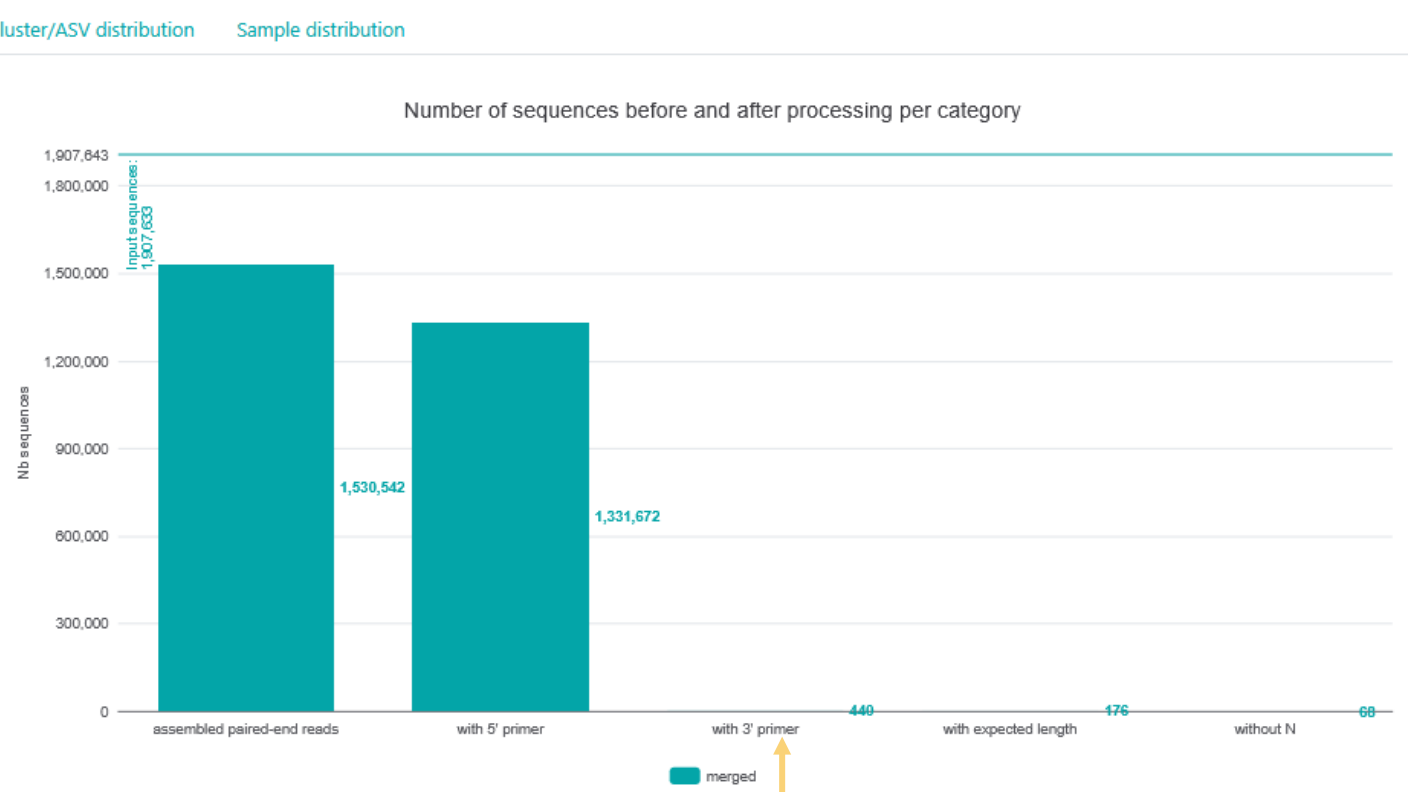
ITS

DADA2



Practice session

What conclusions can you draw from this graph?



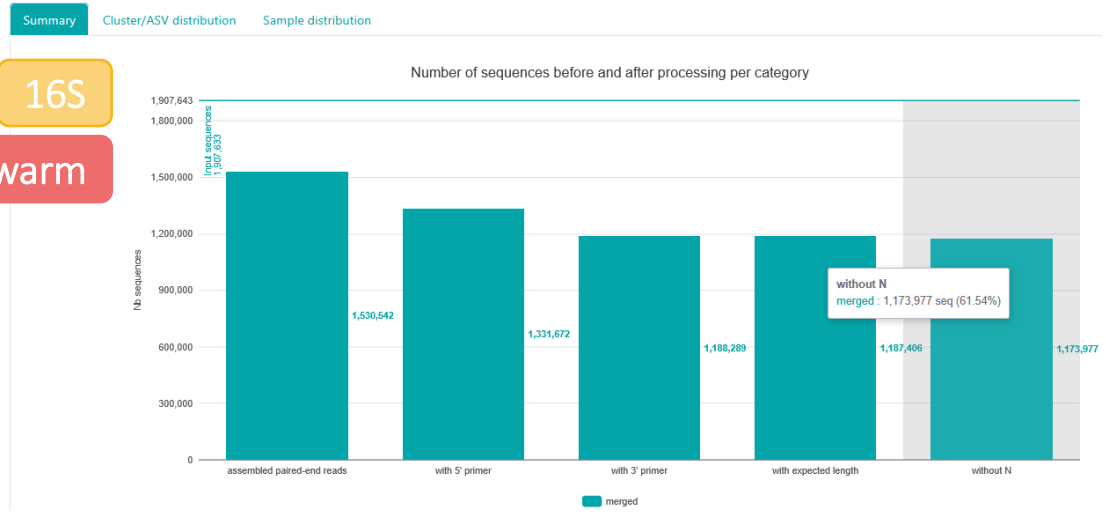
It is likely that the user entered the 3' primer in the wrong direction.

Practice session

How many sequences remain after the data has been preprocessed ?

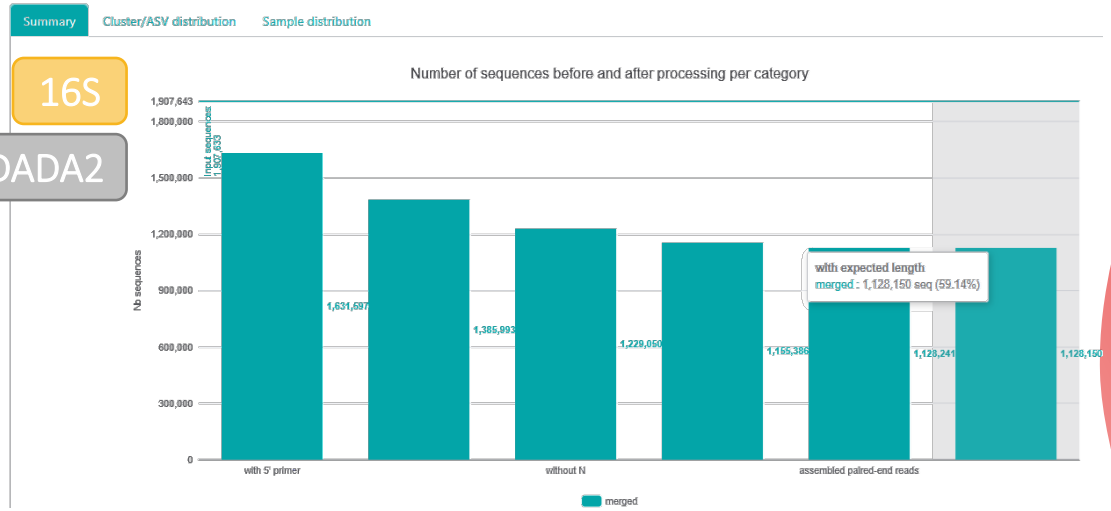
16S

swarm



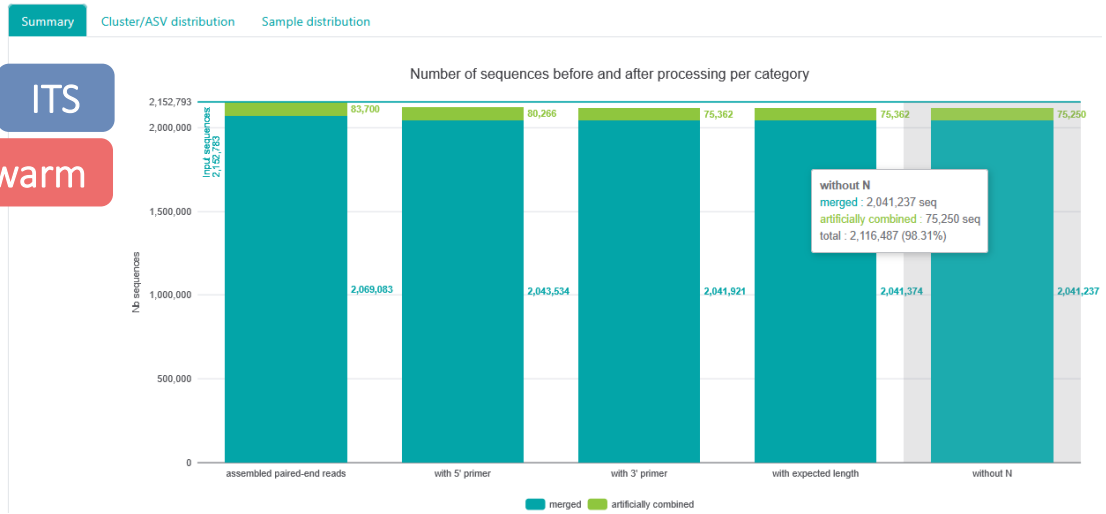
16S

DADA2



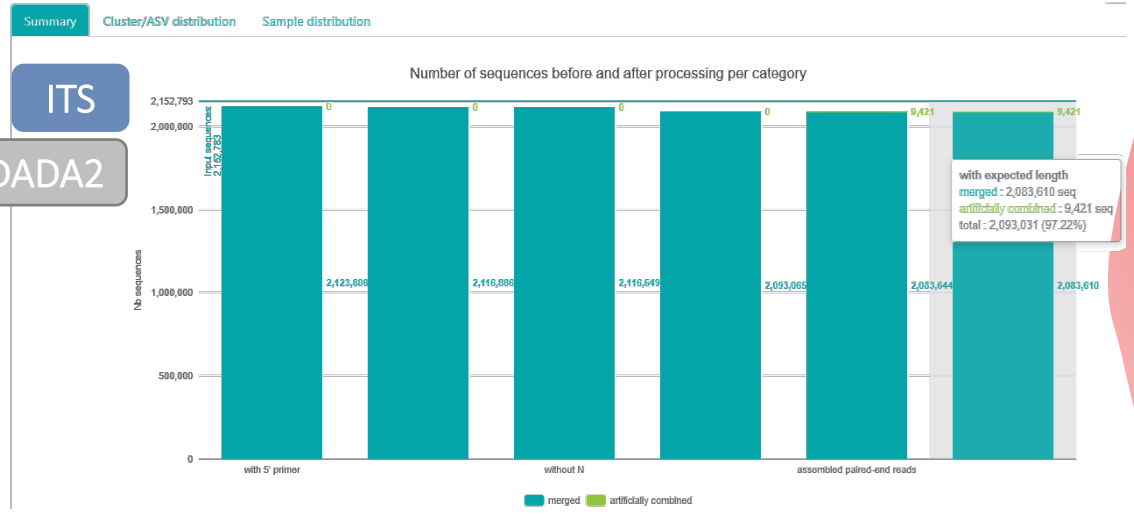
ITS

swarm



ITS

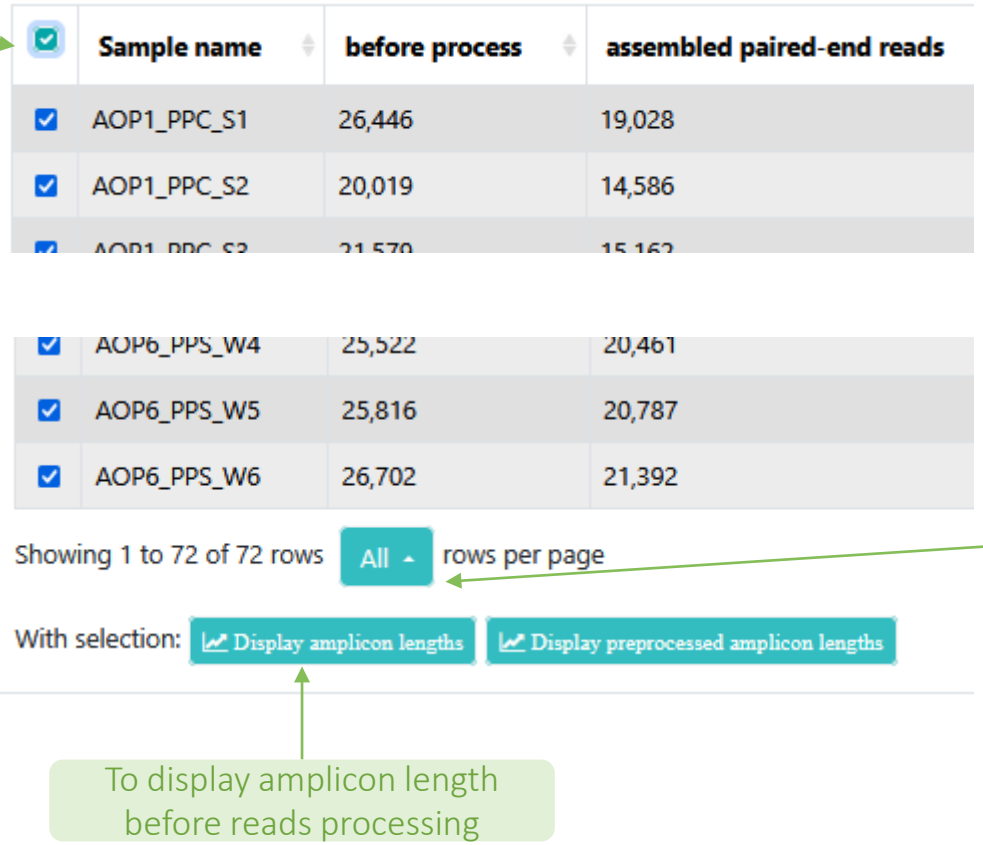
DADA2



Practice session

What is the length of your merged reads before pre-processing ?

To select all samples



To show all samples

To display amplicon length before reads processing

Practice session

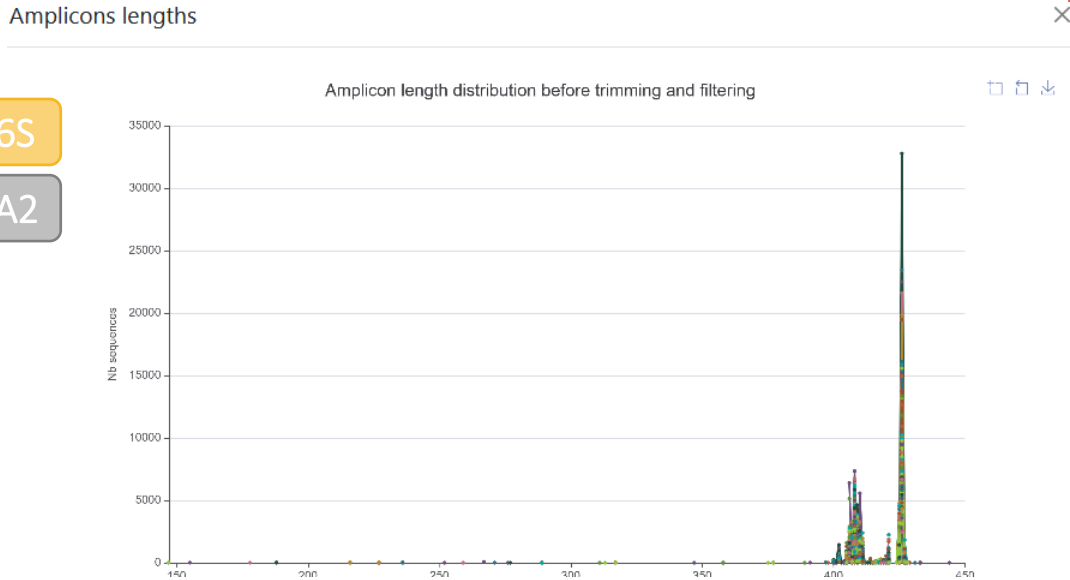
What is the length of your merged reads before pre-processing ?

16S

swarm

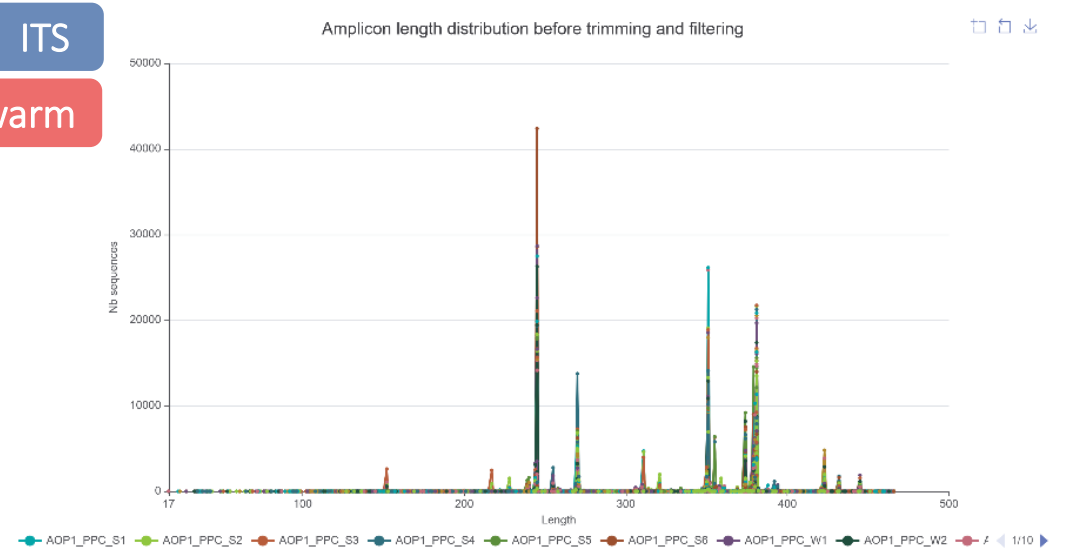
16S

DADA2



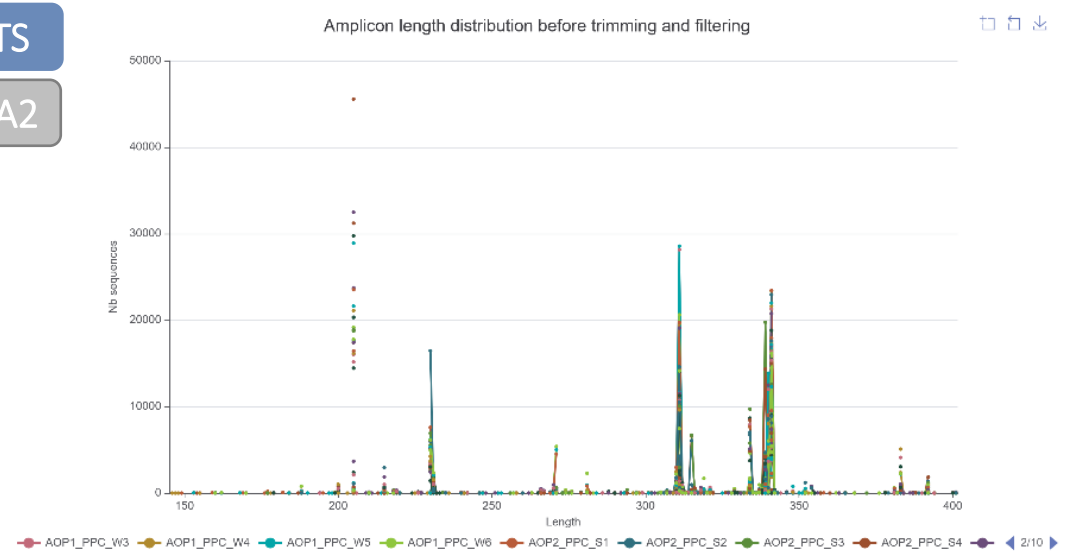
ITS

swarm



ITS

DADA2

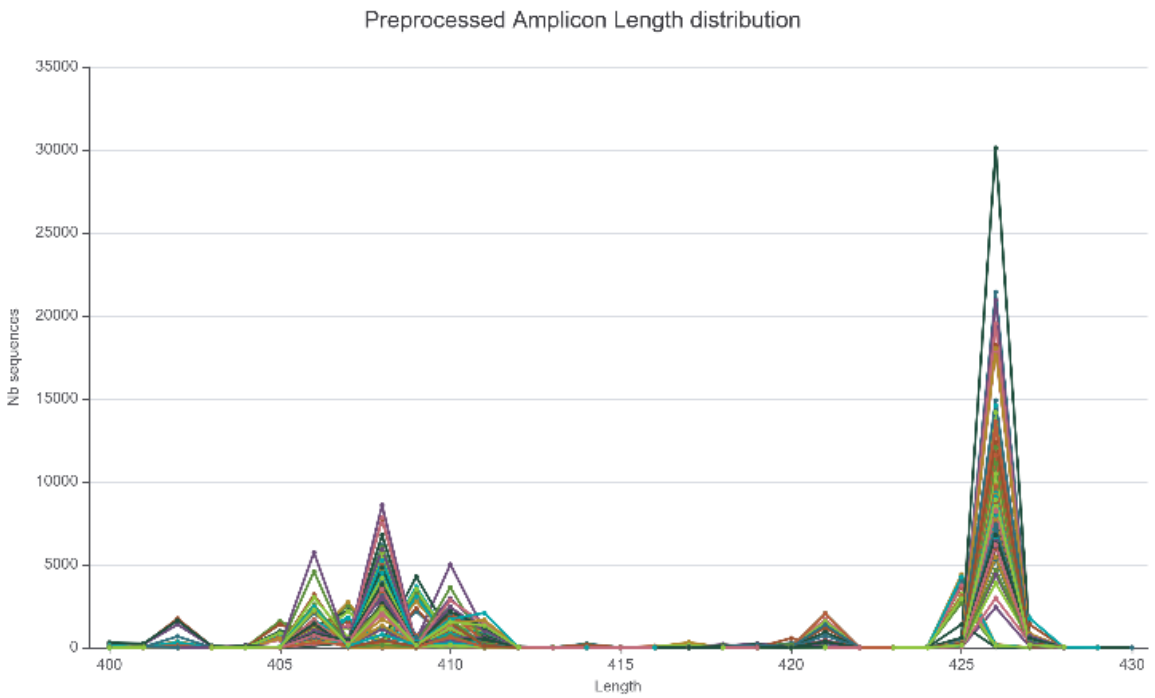


Practice session

Based on amplicon size distributions, what can you tell us about the samples ?

16S

swarm



One pic = one amplicon length ≈ one species

You can zoom in on graphics using the mouse wheel or the selection tool.

AOP6_PPS_W4	25,522	20,461
AOP6_PPS_W5	25,816	20,787
AOP6_PPS_W6	26,702	21,392

Showing 1 to 72 of 72 rows All rows per page

With selection: Display amplicon lengths | Display preprocessed amplicon lengths

To display amplicon length after reads processing

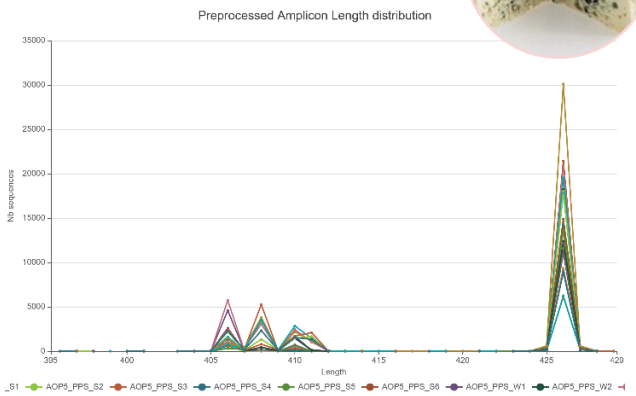
Practice session

Based on amplicon size distributions, what can you tell us about the samples per cheese category ?

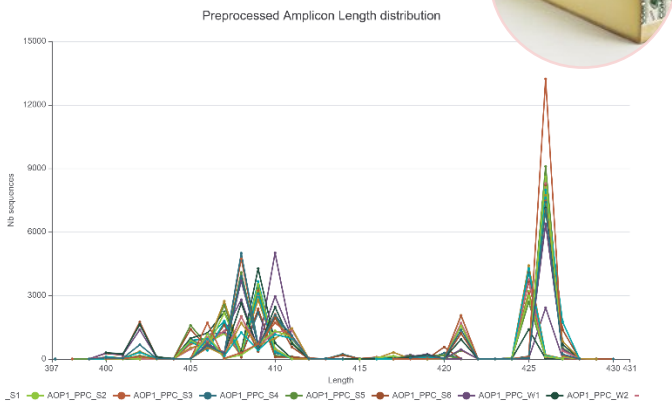
16S

swarm

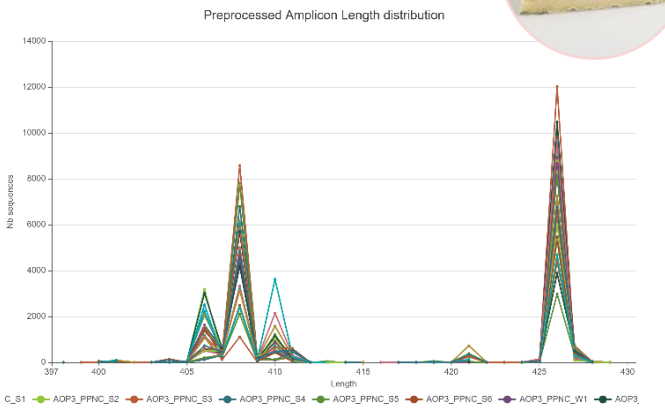
PPS



PPC



PPNC



Practice session

Explore report html: Cluster/ASV distribution tab

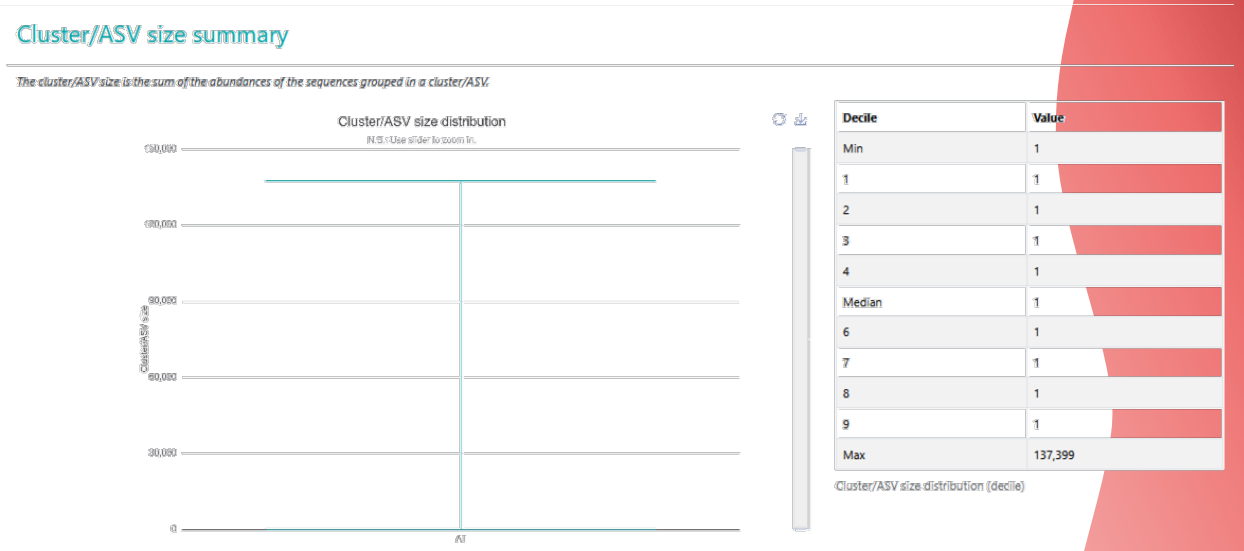
How many clusters do you get ?
Interpret the boxplot: Cluster/ASV size summary
Interpret the table: Cluster/ASV size details
How many single singletons do you find?
What conclusions can we draw from observing the sequence distribution?
How many clusters share “AOP1_PPC_S3” with at least one other sample?
How many clusters could we expect to be shared ?
How many sequences represent the specific clusters of “AOP1_PPC_S3”?
How would you interpret the 'hierarchical clustering'?

Practice session

How many clusters do you get ?

16S swarm	Clusters/ASVs 179,511	Sequences 1,173,977
16S DADA2	Clusters/ASVs 7,489	Sequences 1,128,150
ITS swarm	Clusters/ASVs 112,329	Sequences 2,116,487
ITS DADA2	Clusters/ASVs 2,657	Sequences 2,093,031

Interpret the boxplot: Cluster/ASV size summary



Decile	Value
Min	1
1	1
2	1
3	1
4	1
Median	1
6	1
7	1
8	1
9	1
Max	137,399

Most of clusters are singletons

Practice session

How many single singletons do you find?

16S swarm

Cluster/ASV size	Number of clusters/ASVs	% of all clusters/ASVs
1	171,973	95.80

16S DADA2

Cluster/ASV size	Number of clusters/ASVs	% of all clusters/ASVs
1	2,907	38.82

ITS swarm

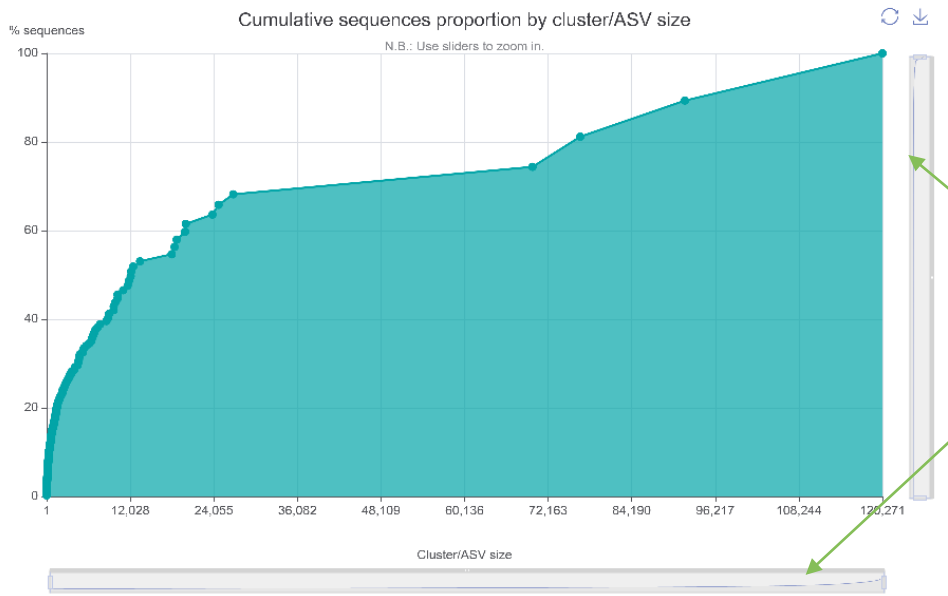
Cluster/ASV size	Number of clusters/ASVs	% of all clusters/ASVs
1	110,033	97.96

ITS DADA2

Cluster/ASV size	Number of clusters/ASVs	% of all clusters/ASVs
1	649	24.43

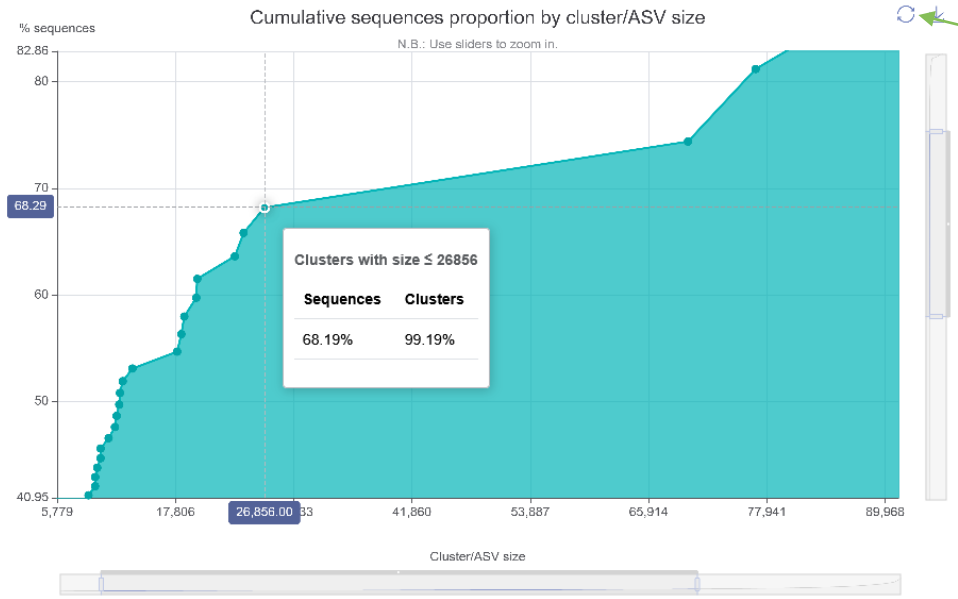
Tips !

Cluster/ASV size distribution



You can zoom in on graphics using the mouse wheel or the selection tool.

Cluster/ASV size distribution



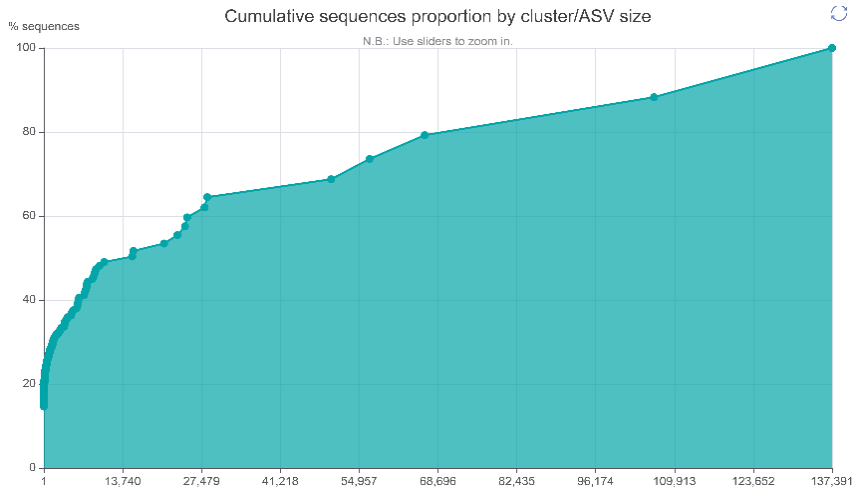
Restore the initial picture.

Practice session

What is the length of your merged reads before pre-processing ?

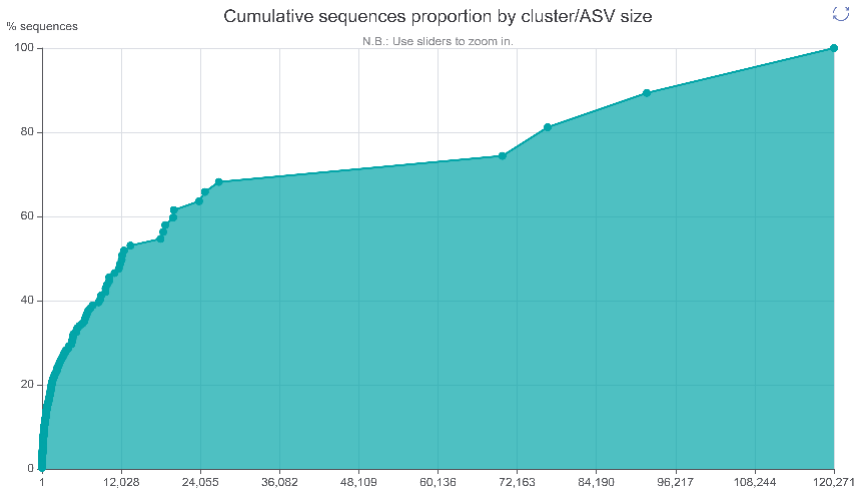
16S

swarm



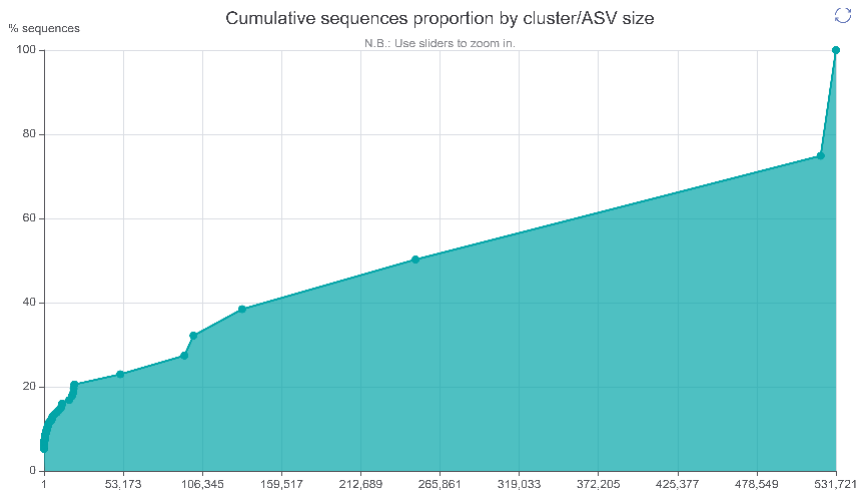
16S

DADA2



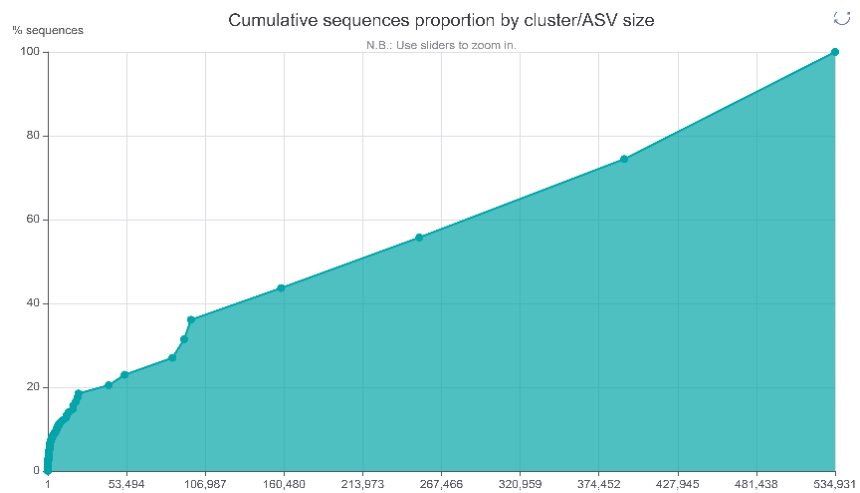
ITS

swarm



ITS

DADA2



Practice session

How many clusters share “AOP1_PPC_S3” with at least one other sample?

ITS
swarm

Sequence count

Sample name	Number of clusters/ASVs	Number of shared clusters/ASVs	Number of own clusters/ASVs
AOP1_PPC_S1	1,774	264	1,510
AOP1_PPC_S2	4,017	467	3,550
AOP1_PPC_S3	3,592	395	3,197
AOP1_PPC_S4	2,105	366	1,739
AOP1_PPC_S5	2,560	333	2,227
AOP1_PPC_S6	2,049	325	1,724

How many clusters could we expect to be shared ?

All, since AOP1_PPC_S1, S2 and S3 are replicates

Practice session

How many clusters share “AOP1_PPC_S3” with at least one other sample?

ITS
DADA2

Sequence count

AOP1_PPC_S

Sample name	Number of clusters/ASVs	Number of shared clusters/ASVs	Number of own clusters/ASVs
AOP1_PPC_S1	195	172	23
AOP1_PPC_S2	410	277	133
AOP1_PPC_S3	288	235	53
AOP1_PPC_S4	305	235	70
AOP1_PPC_S5	363	200	163
AOP1_PPC_S6	316	222	94

Sample information

How many clusters could we expect to be shared ?

All, since AOP1_PPC_S1, S2 and S3 are replicates

Practice session

How many sequences represent the number of own cluster of “AOP1_PPC_S3”?

ITS

swarm

Sample name	Number of own clusters/ASVs	Number of sequences	Number of shared sequences	Number of own sequences
AOP1_PPC_S1	1,510	29,195	27,684	1,511
AOP1_PPC_S2	3,550	33,712	30,135	3,577
AOP1_PPC_S3	3,197	36,260	33,027	3,233
AOP1_PPC_S4	1,739	27,471	25,711	1,760
AOP1_PPC_S5	2,227	33,831	31,521	2,310
AOP1_PPC_S6	1,724	26,955	25,167	1,788

Sample information

Practice session

How many sequences represent the number of own cluster of “AOP1_PPC_S3”?

ITS

DADA2

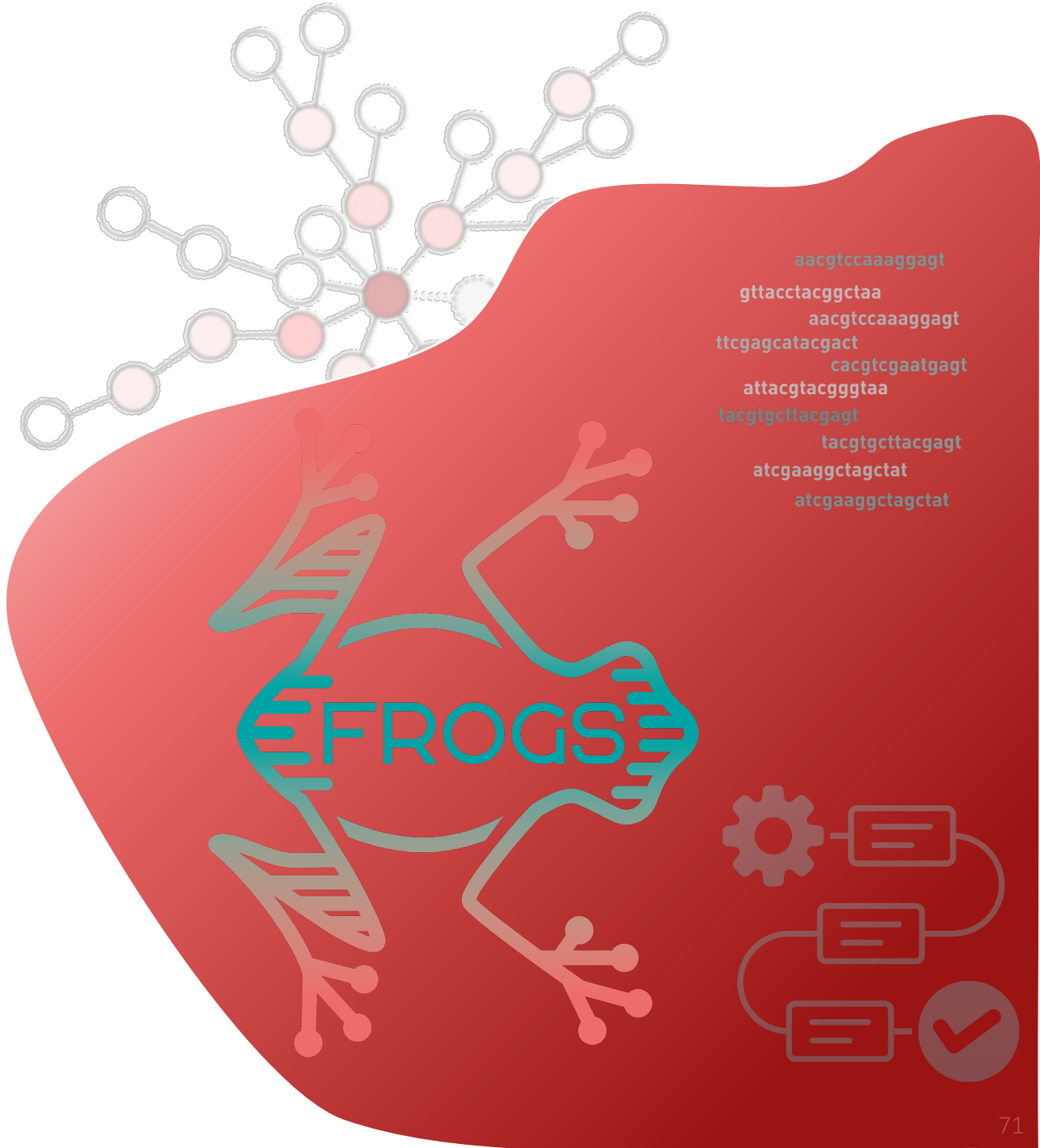
Sample name	Number of own clusters/ASVs	Number of sequences	Number of shared sequences	Number of own sequences
AOP1_PPC_S1	23	28,861	28,785	76
AOP1_PPC_S2	133	32,900	31,555	1,345
AOP1_PPC_S3	53	35,299	34,825	474
AOP1_PPC_S4	70	27,060	26,831	229
AOP1_PPC_S5	163	33,358	31,493	1,865
AOP1_PPC_S6	94	26,540	26,030	510

Sample information

FROGS Core 1

Main tools

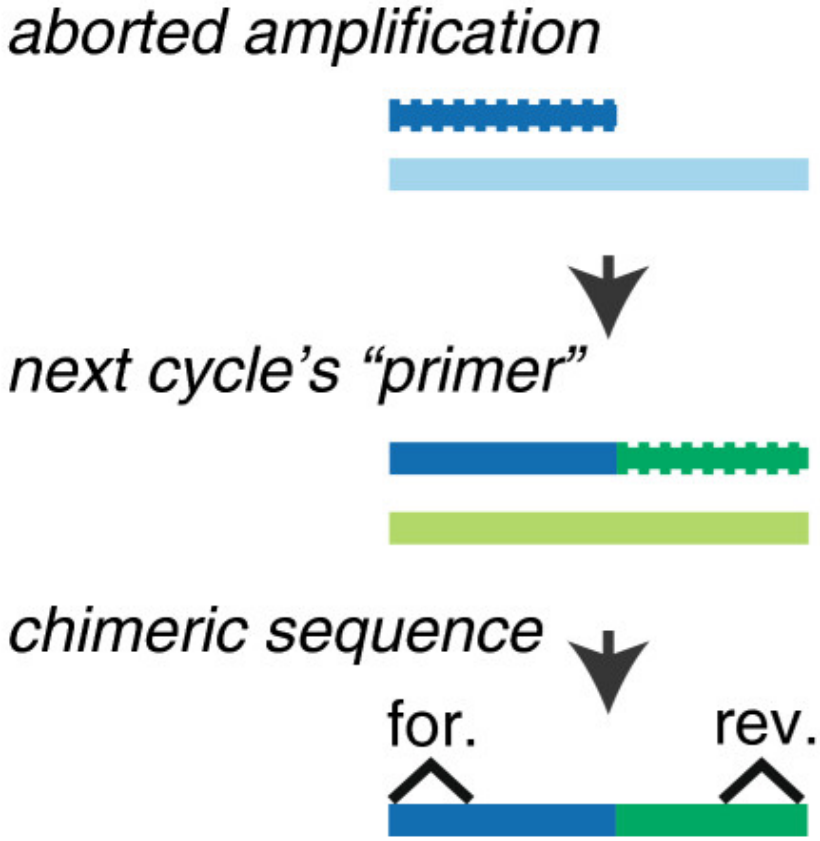
Remove chimera



What is chimera ?

PCR-generated chimeras are typically created when an aborted amplicon acts as a primer for a heterologous template. Subsequent chimeras are about the same length as the non-chimeric amplicon and contain the forward (for.) and reverse (rev.) primer sequence at each end of the amplicon.

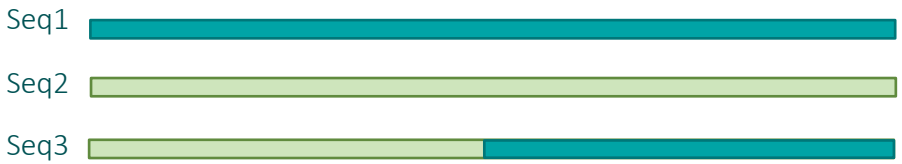
Chimera: from 5 to 45% of reads
(Haas 2011 doi: 10.1101/gr.112730.110)



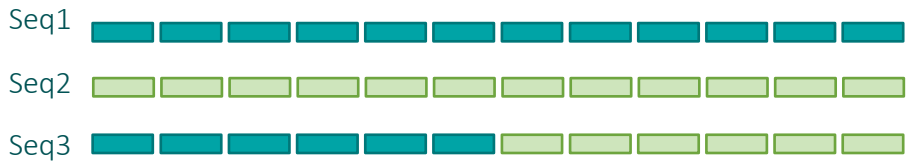
Fichot and Norman *Microbiome* 2013 **1**:10 doi:10.1186/2049-2618-1-10

Chimera detection

1 Sample sequences.

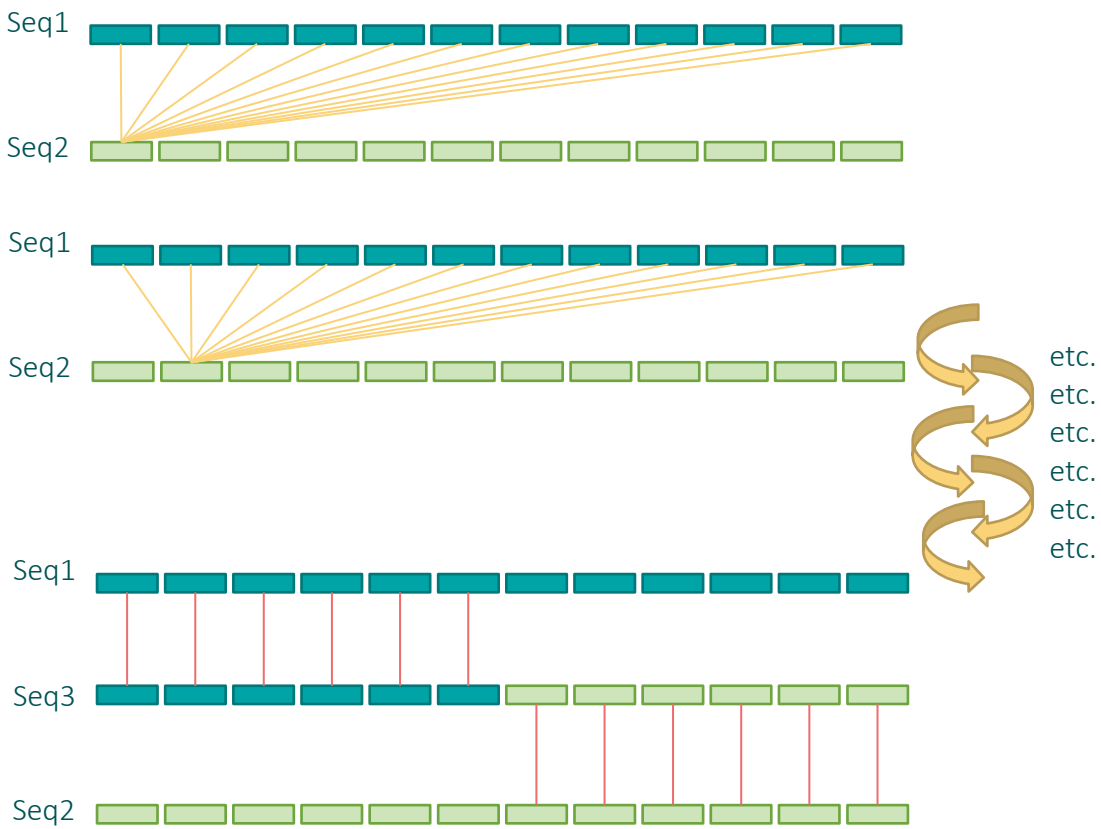


2 Dividing into fragments.



Rule: chimeras are by-products of PCR, so they occur less frequently than the parent sequences.

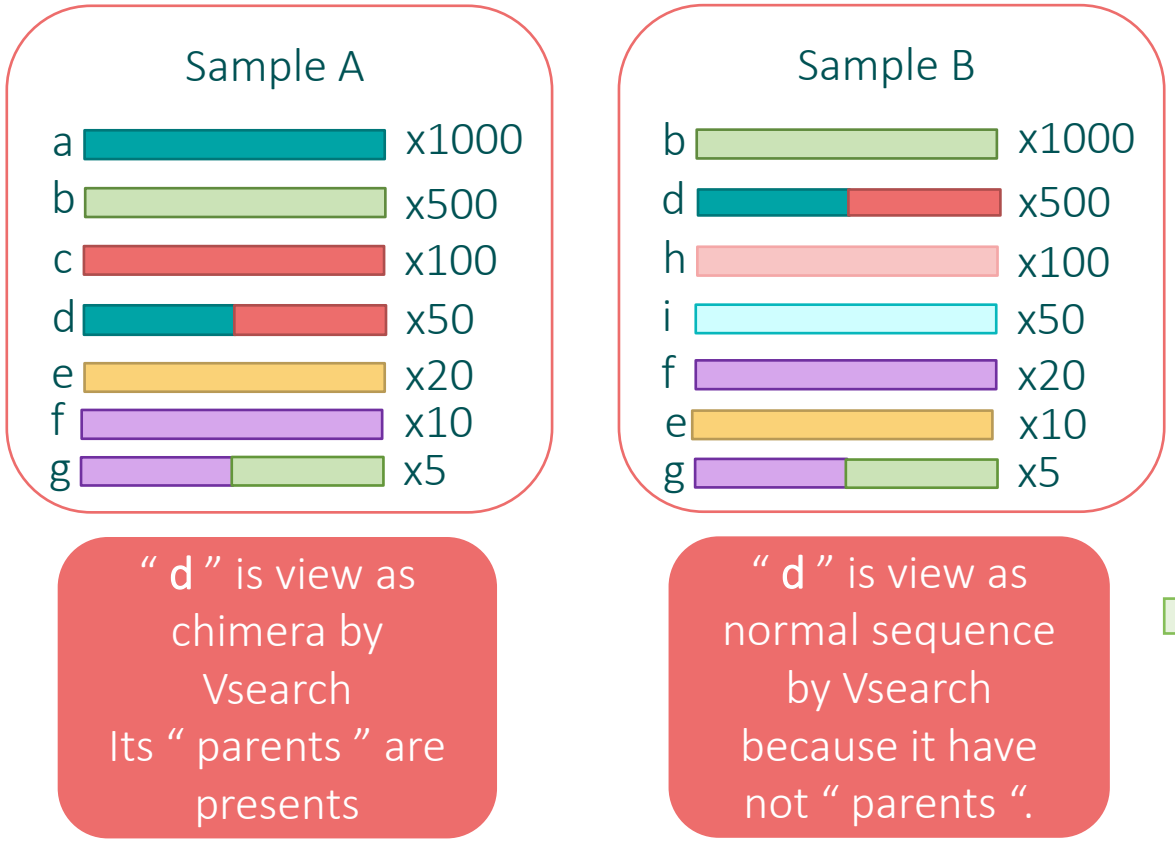
3 Fragment alignments of fragments of SeqX *vs* fragments of SeqY.



The best hits of Seq3's alignment are found in both Seq1 and Seq2.
Seq3 is considered as a chimera.

A method adapted to both short-reads and long-reads.

Cross-validation to delete false-positive



“ **d** ” is view as chimera by Vsearch
Its “ parents ” are presents

“ **d** ” is view as normal sequence by Vsearch because it have not “ parents “.

- For FROGS “d” is not a chimera
- For FROGS “g” is a chimera, “g” is removed
- FROGS increases the detection specificity

With sample-by-sample analysis, and by classifying a sequence as a true chimera only if it is identified as such in every sample in which it is present, the risk of a false positive is nearly zero.

Practice session

Please open the FROGS Core Main 2 tool and familiarize yourself with the required parameters.

Please, enter all the information you have/understand.

Run the process !

Practice session

Select a tool from the FROGS Core suite to run your analysis.

- ( ) -- Please select a tool --
- ( ) 1.a. Reads processing of short reads
- ( ) 1.b. Reads processing of long reads
- ( ) 1.c. Reads processing of 454 reads
- (•) 2. Remove chimera
- ( ) 3. Cluster/ASV filters
- ( ) 4. Taxonomic affiliation
- ( ) 5. Phylogenetic tree building
- ( ) ITSx

Sequence file (.fasta) *

4: FROGS Core 1-Main - processing_short_reads: sequence.fasta

accepted formats

The sequence file to filter. (--input-fasta)

Abundance file (.biom) *

5: FROGS Core 1-Main - processing_short_reads: abundance.biom

accepted formats

The abundance file to filter. (--input-biom)

The data originally comes from long-read sequencing *

- ( ) Yes
- (•) No

For long reads sequencing, chimera will be detected using chimeras_denovo algorithm of vsearch (instead of uchime_denovo for short reads sequencing). (--long-read)

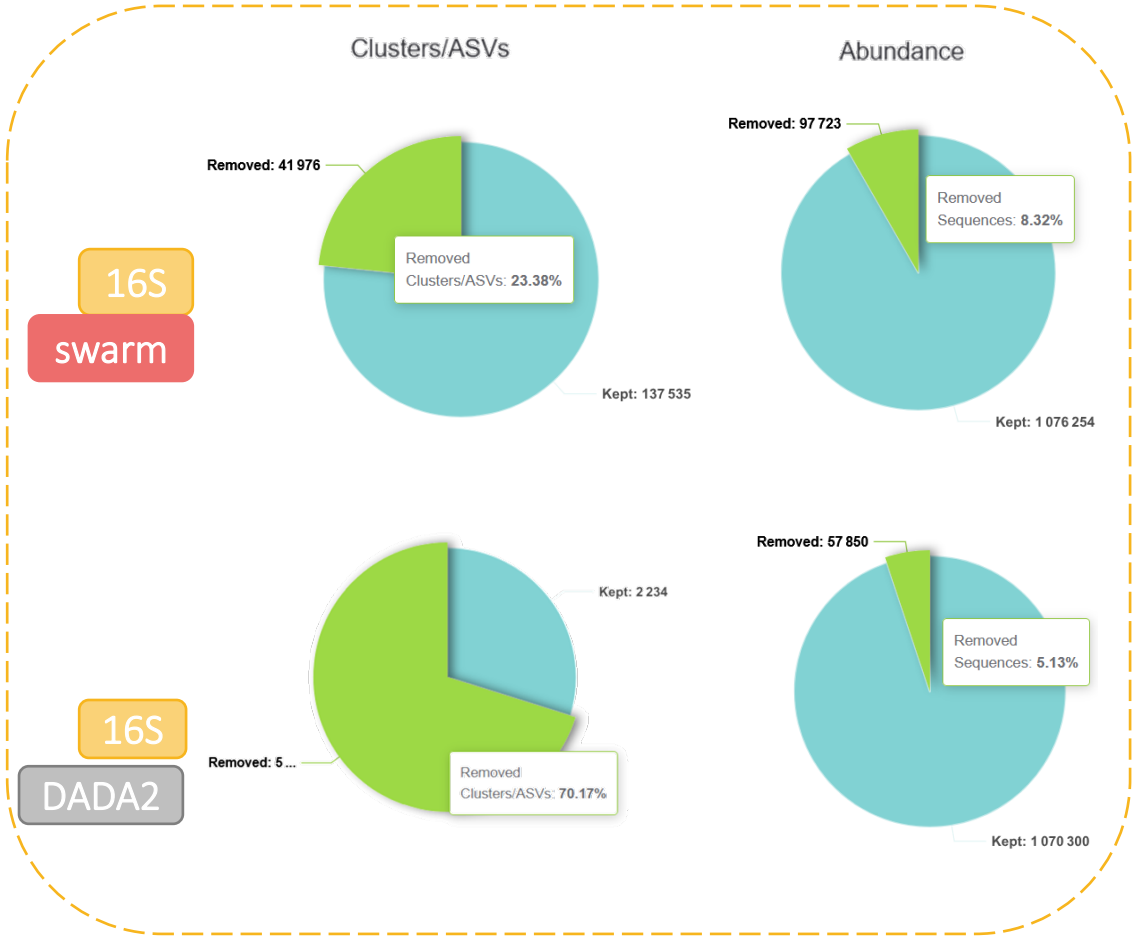
Practice session

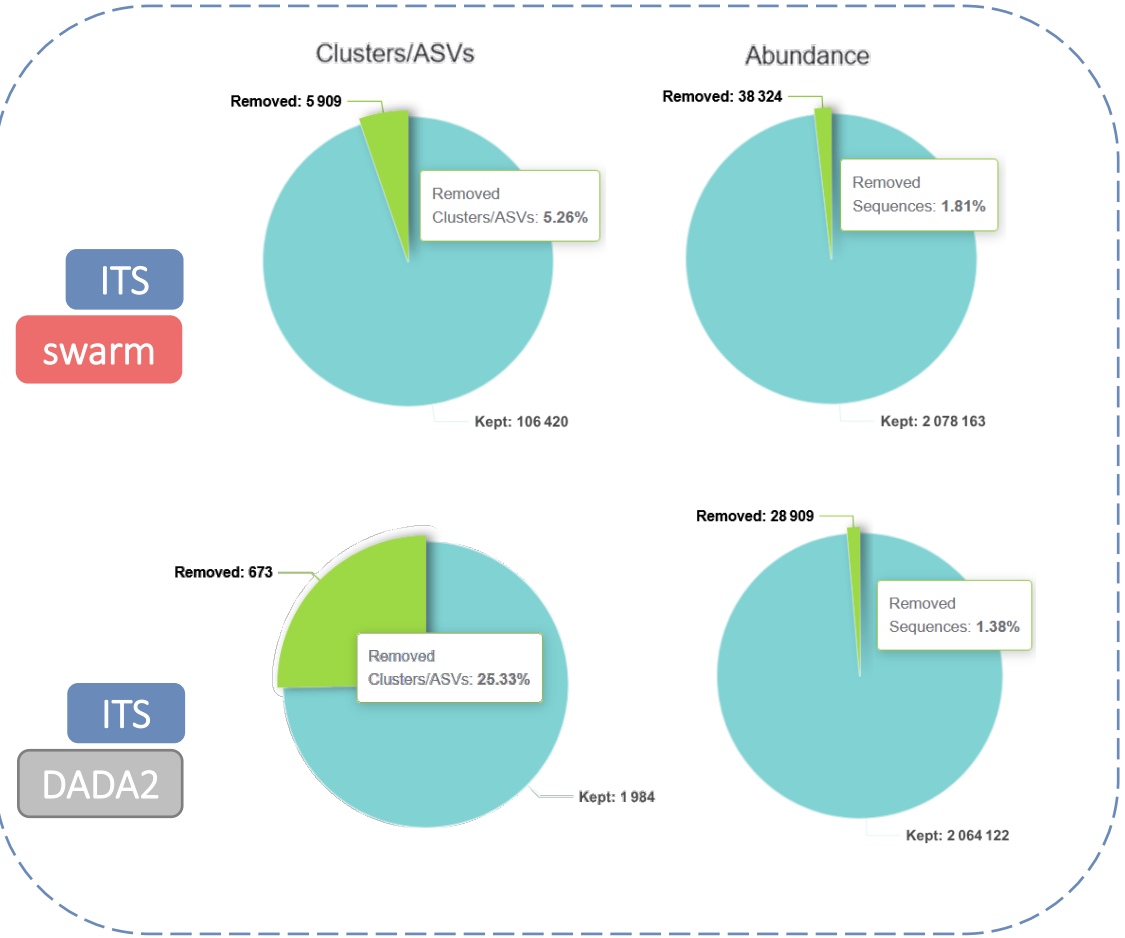
How many clusters remain after the removal of chimeras ?

How many sequences represents this ?

What is your conclusion ?

Practice session





Removed clusters are low abundance clusters.

impact of cross-validation

Chimera detection by sample

Chimeras are first detected independently in each sample. Only clusters/ASVs detected as chimeric in every sample (cross-validated chimeras) are removed.

Search

Chimeric clusters removed	Chimeric abundance removed	Abundance of the most abundant chimera removed	Individual chimera detected
41	535	179	66
96	1766	317	136
53	1291	243	93

66 chimeras are detected but only 41 are removed because 25 have been invalidated by the cross validation

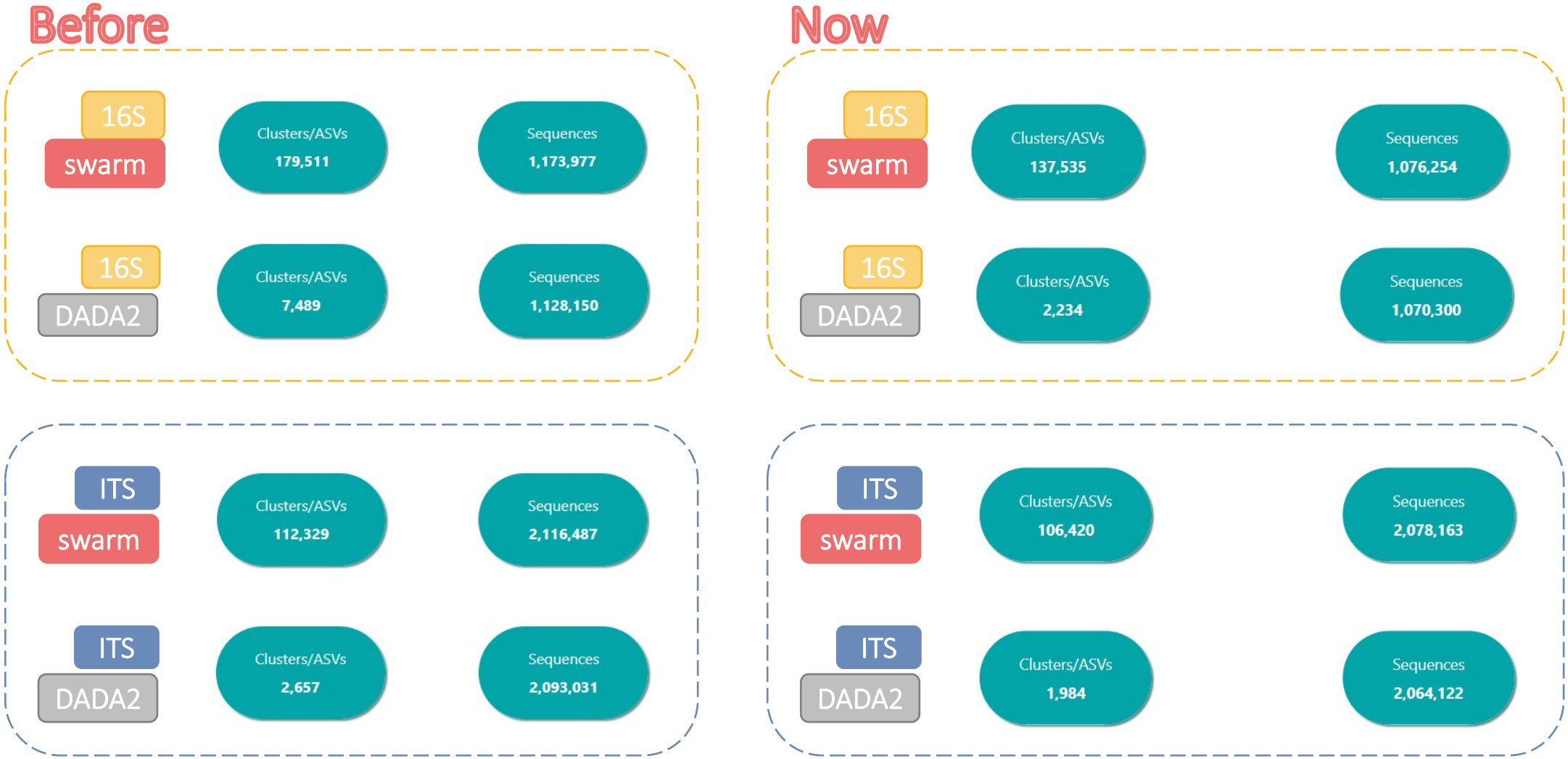
Practice session

How many clusters do you get ?

Please, focus on singletons

Practice session

How many clusters do you get ?



Practice session

Focus on singleton

16S

swarm

Before

Cluster/ASV size	Number of clusters/ASVs	% of all clusters/ASVs
1	171,973	95.80

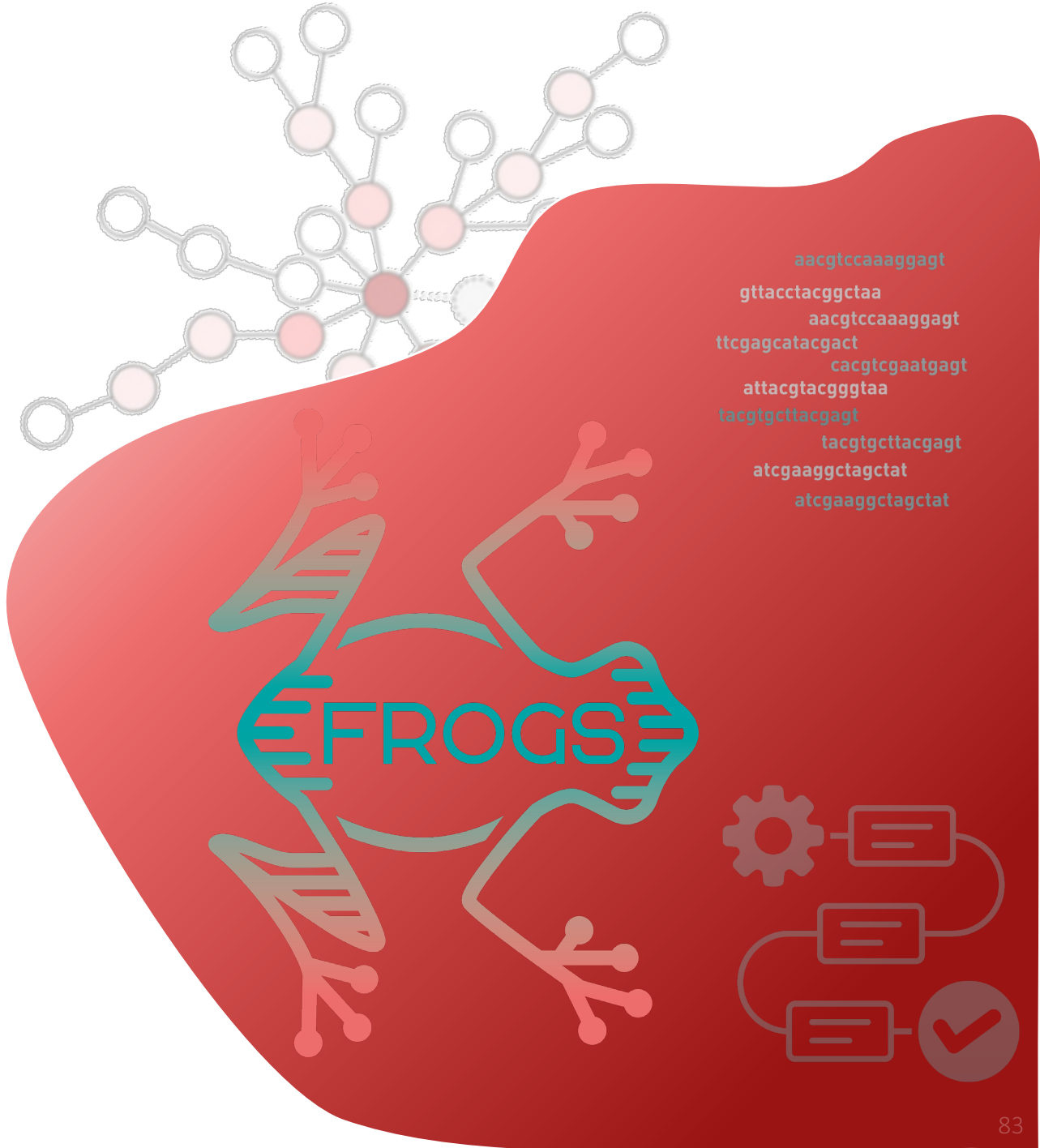
Now

Cluster/ASV size	Number of clusters/ASVs	% of all clusters/ASVs
1	134,825	98.03

FROGS Core 1

Main tools

Cluster/ASV filter



Objectives

Goal: This tool deletes clusters based on the conditions entered by the user.

If a cluster responds to at least one criterion, the cluster is deleted.

Criteria:

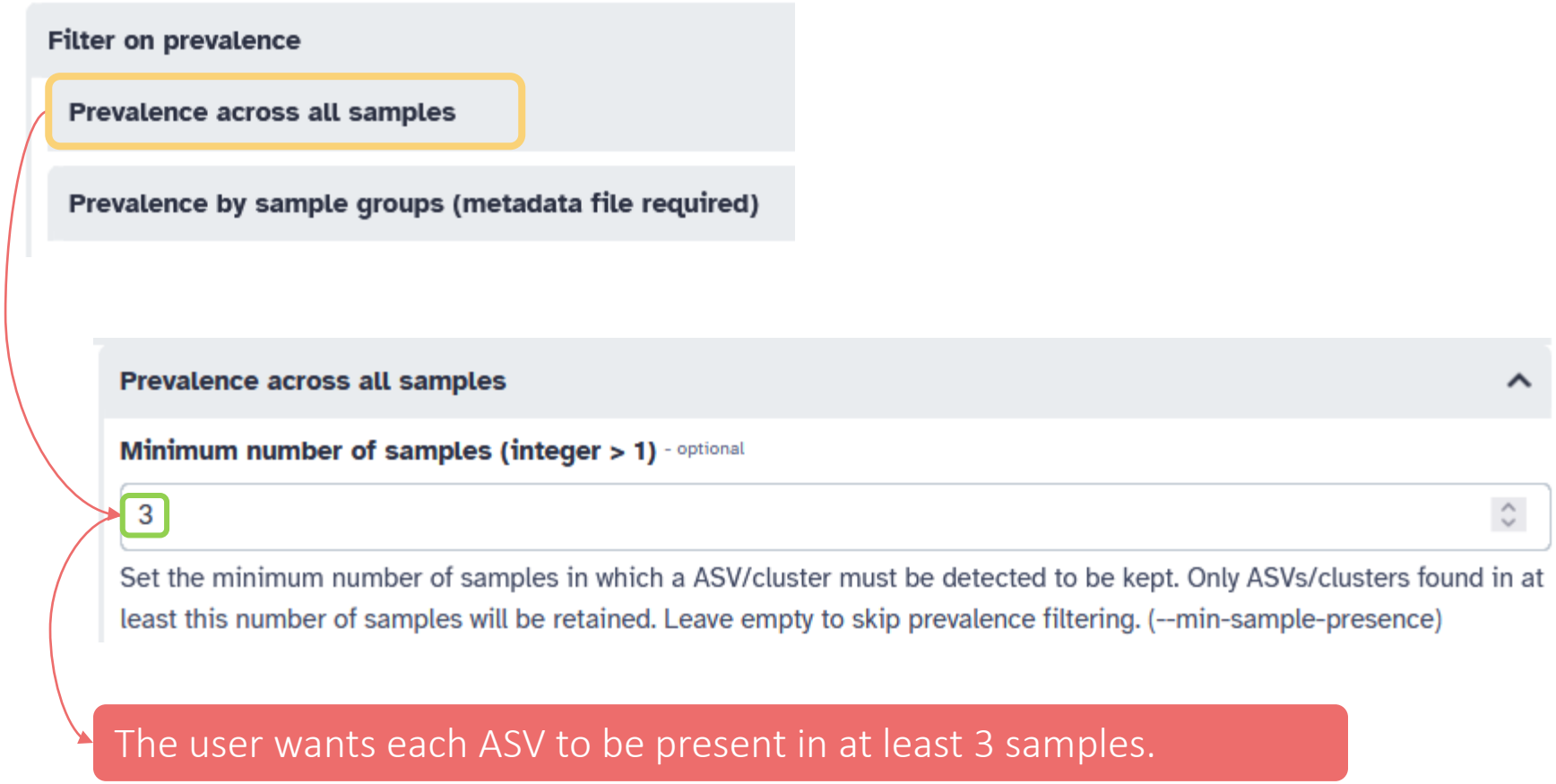
Objectives

Goal: This tool deletes clusters based on the conditions entered by the user.

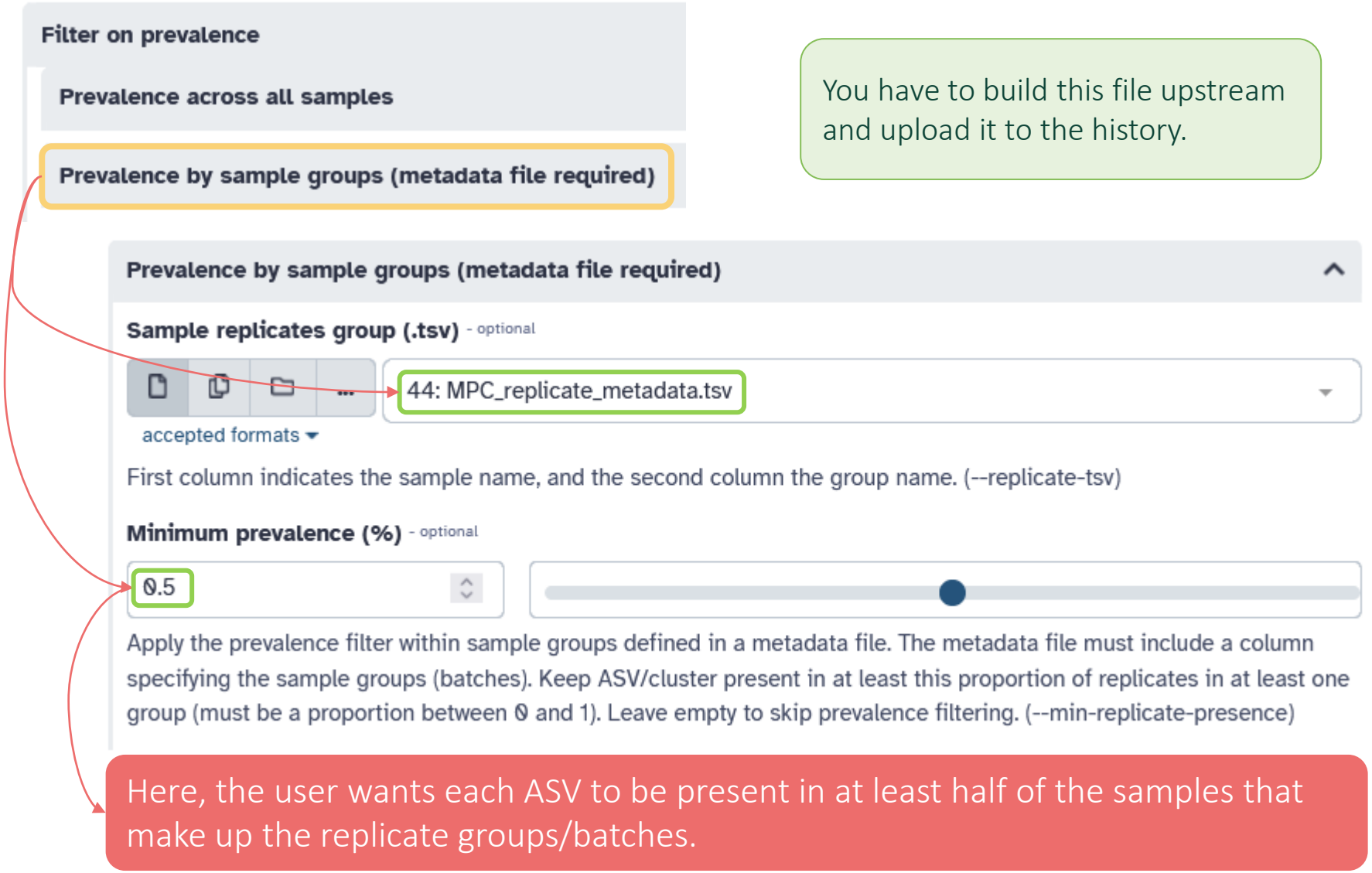
If a cluster responds to at least one criterion, the cluster is deleted.

Criteria:

Filter on prevalence



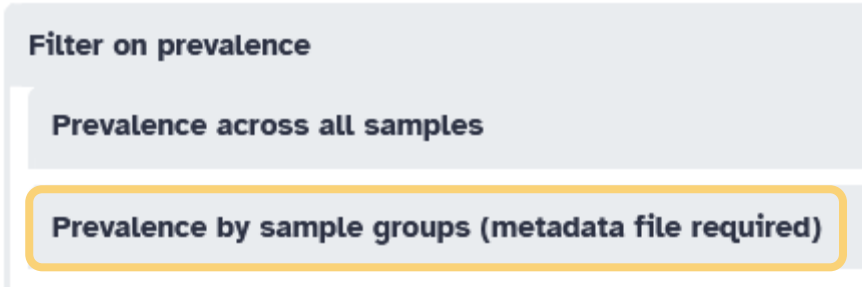
Filter on prevalence



You have to build this file upstream and upload it to the history.

Here, the user wants each ASV to be present in at least half of the samples that make up the replicate groups/batches.

Filter on prevalence: the replicate metadata file



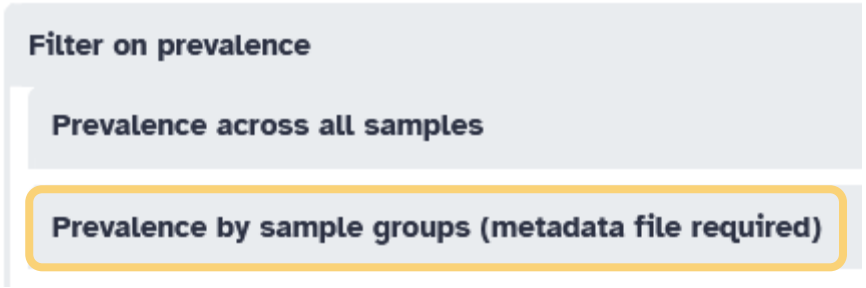
You have to build this file upstream and upload it to the history.

How should the file containing the replicated sample names be built?

The file must consist of two columns only, separated by a tab.
The first column contains the exact names of the samples (i.e. those contained in the biom file).
The second contains the name of the group to which they belong.
Please note that group names must not contain accents, spaces, or special characters.
There are no headers on the columns!

AOP1_PPC_S1	AOP1_S
AOP1_PPC_S2	AOP1_S
AOP1_PPC_S3	AOP1_S
AOP1_PPC_S4	AOP1_S
AOP1_PPC_S5	AOP1_S
AOP1_PPC_S6	AOP1_S
AOP1_PPC_W1	AOP1_W
AOP1_PPC_W2	AOP1_W
AOP1_PPC_W3	AOP1_W
AOP1_PPC_W4	AOP1_W
AOP1_PPC_W5	AOP1_W
AOP1_PPC_W6	AOP1_W
AOP2_PPC_S1	AOP2_S
AOP2_PPC_S2	AOP2_S

Filter on prevalence: the replicate metadata file



You have to build this file upstream and upload it to the history.

How the filter works

If we want to retain ASVs present in at least 50% of samples within a group, we set the threshold to 0.5.
The process will therefore retain the ASVs present in at least:

- two "rich" samples
- three "richAB" samples and
- one "lowAB" sample
- one “april21" sample
- and all the ASVs in sample9, as this is the only sample representing the 'low' condition.

sample1	rich
sample2	rich
sample3	rich
sample4	richAB
sample5	richAB
sample6	richAB
sample7	richAB
sample8	richAB
sample9	low
sample10	lowAB
sample11	lowAB
sample12	april21
sample13	april21

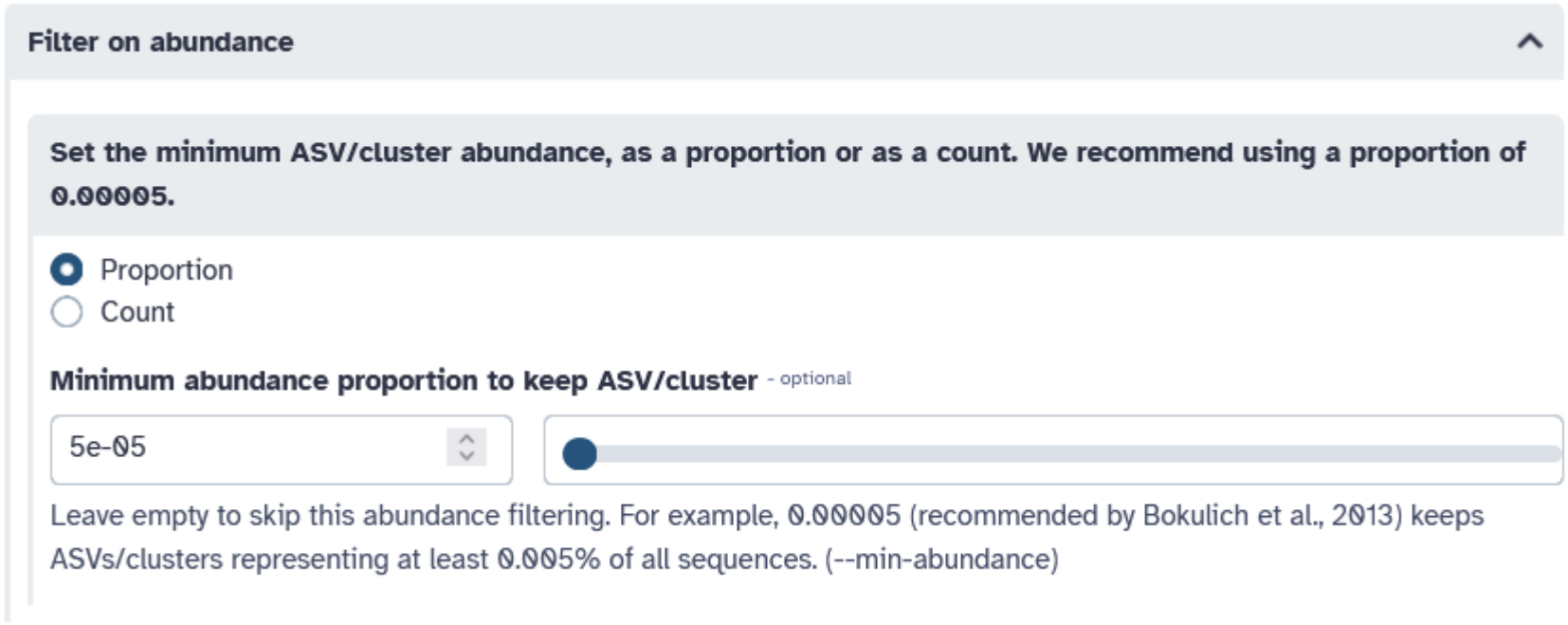
Objectives

Goal: This tool deletes clusters based on the conditions entered by the user.

If a cluster responds to at least one criterion, the cluster is deleted.

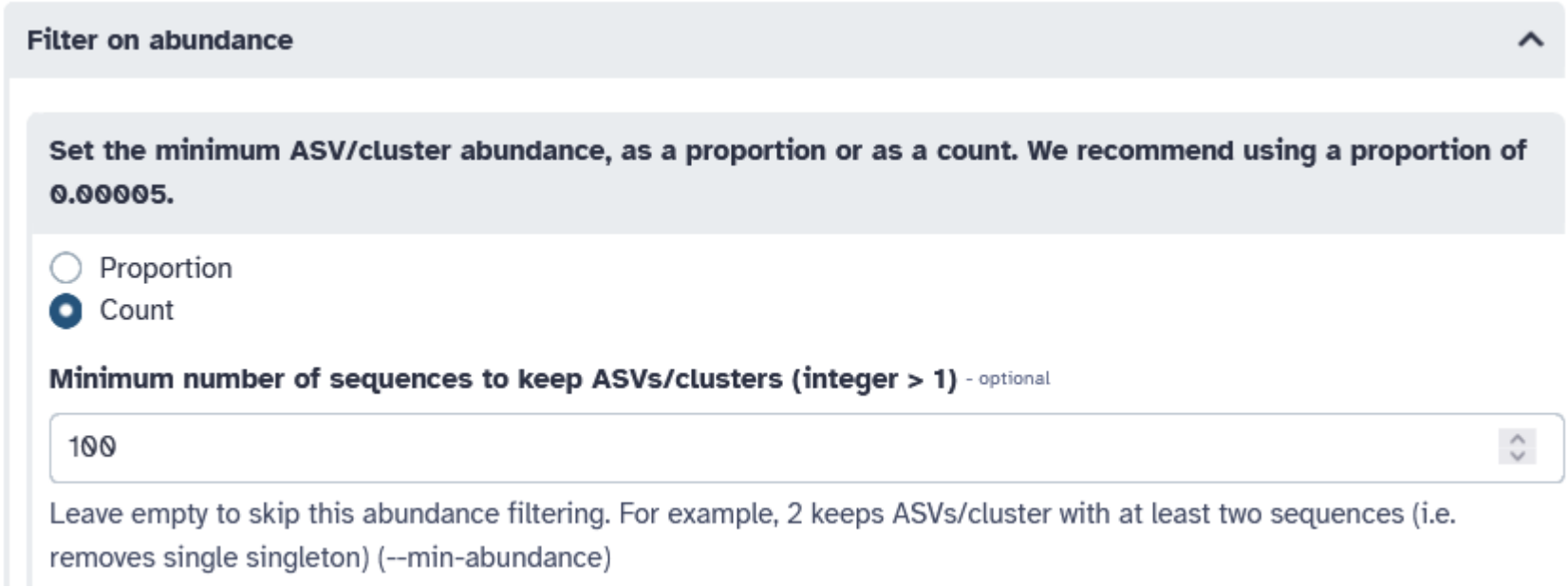
Criteria:

Filter on abundance



The user wants each ASV to represent at least 0.005% (*i.e.* 0.00005) of the total number of sequences.

Filter on abundance



The user wants each ASV to have at least 100 sequences.

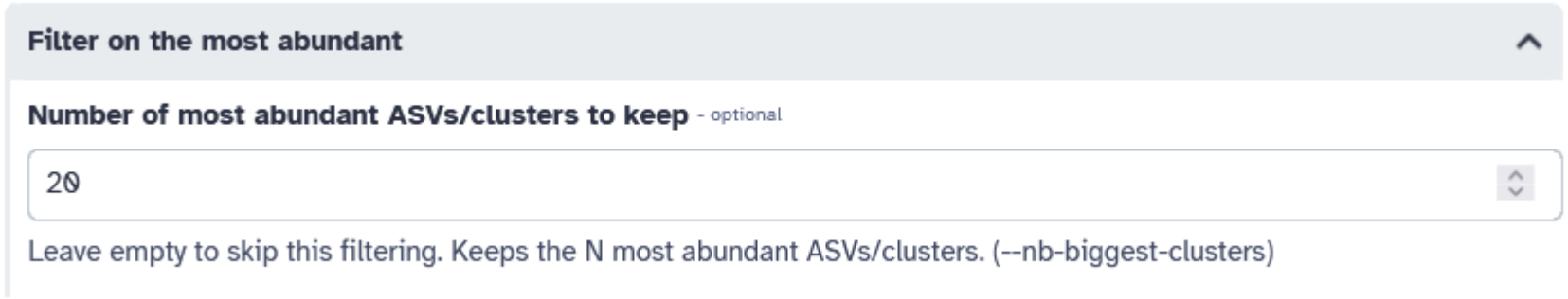
Objectives

Goal: This tool deletes clusters based on the conditions entered by the user.

If a cluster responds to at least one criterion, the cluster is deleted.

Criteria:

Filter on the most abundance



The user only wants to keep the 20 most abundant ASVs.

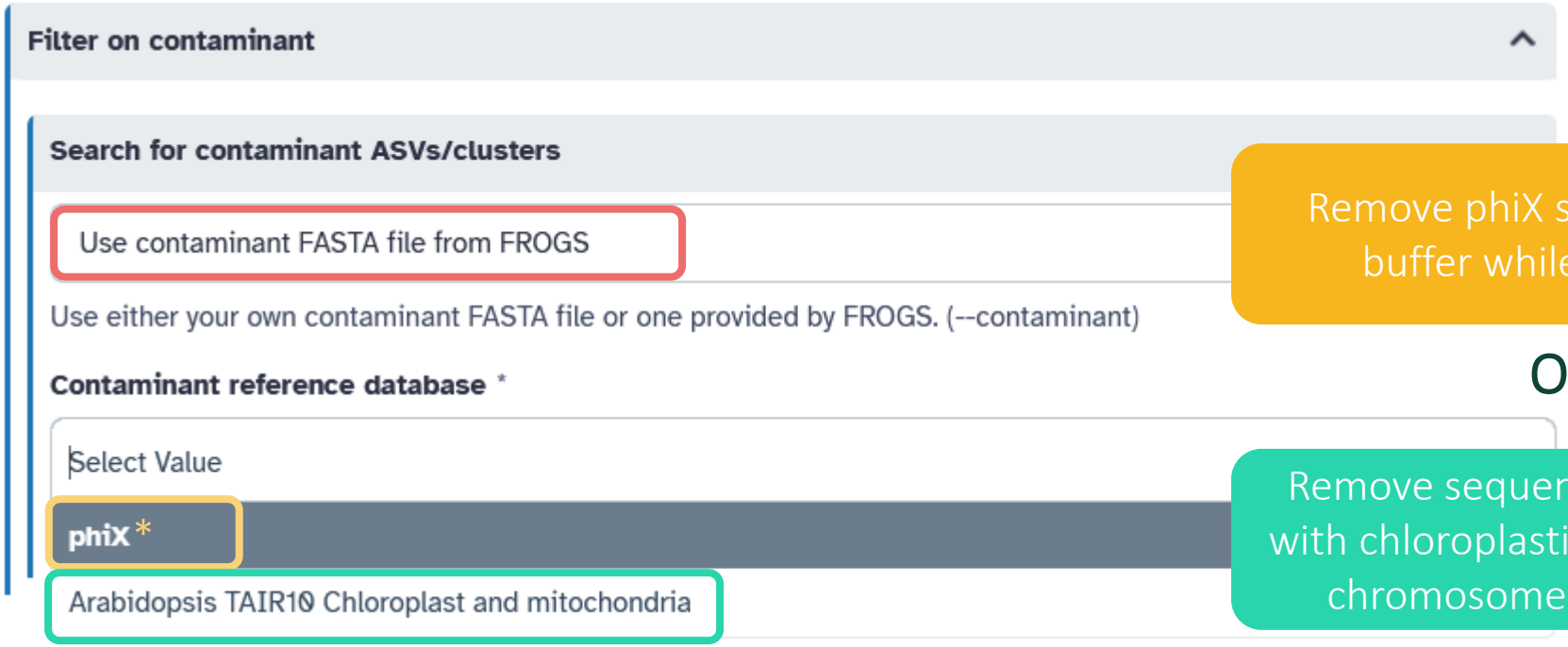
Objectives

Goal: This tool deletes clusters based on the conditions entered by the user.

If a cluster responds to at least one criterion, the cluster is deleted.

Criteria:

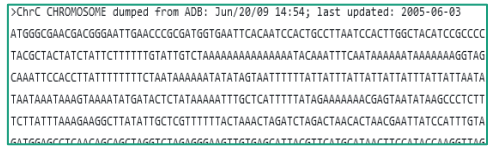
Filter on contaminant

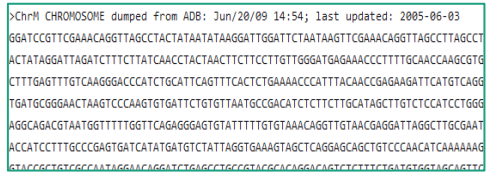


Remove phiX sequence (use as buffer while sequencing)

OR

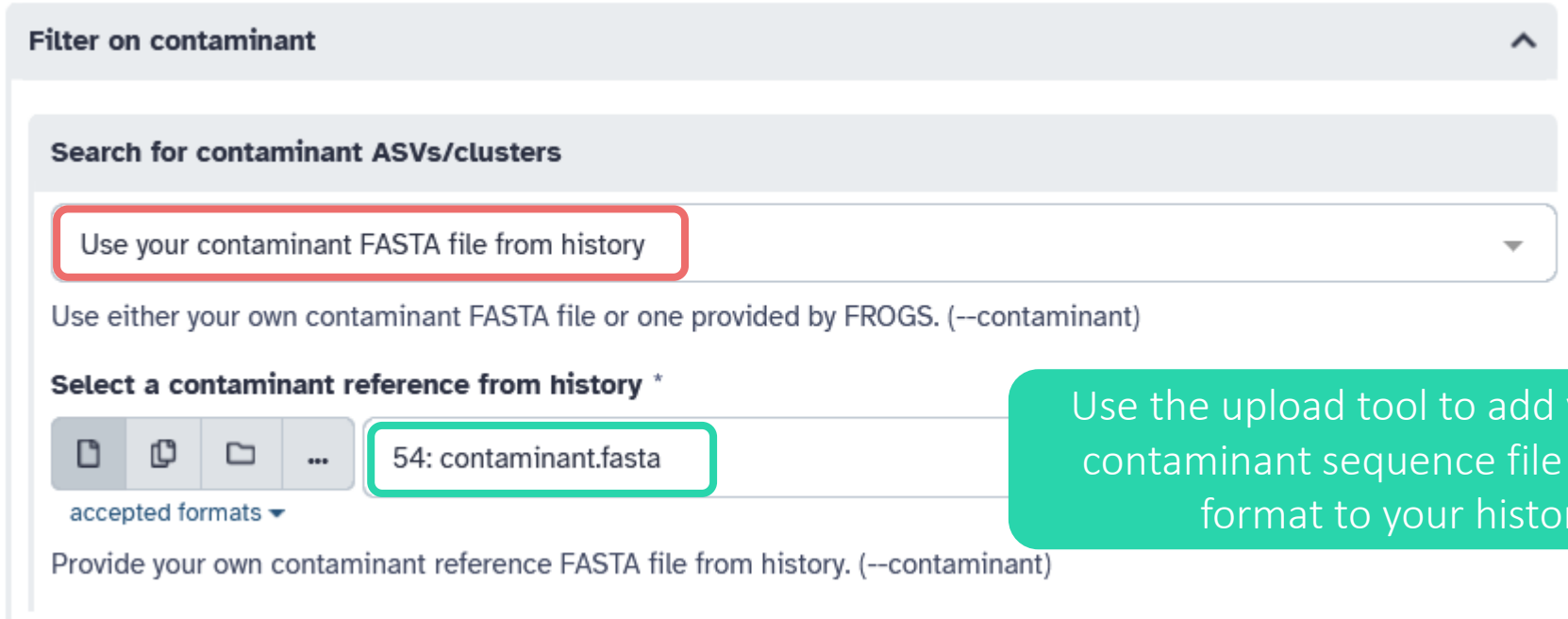
Remove sequences that matches with chloroplastic or mitochondrial chromosomes of *A. Thaliana*





(*) https://www.illumina.com/products/by-type/sequencing-kits/cluster-gen-sequencing-reagents/phix-control-v3.html

Filter on contaminant



Use the upload tool to add your own contaminant sequence file in FASTA format to your history.

Any cluster sequences that match a contaminant sequence will be removed.

Practice session

Please open the FROGS Core Main 3 tool and familiarize yourself with the required parameters.

Please apply the following filters:

- The ASV must be present in at least 4 samples.
- Each ASV must represent a minimum of 0.005% (i.e. 0.00005) of the total number of sequences.

Run the process !

(*) *Nat Methods. 2013 Jan;10(1):57-9. doi: 10.1038/nmeth.2276. Epub 2012 Dec 2.*
Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing.
Bokulich NA1, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, Mills DA, Caporaso JG.

Practice session

Tool Parameters

Select a tool from the FROGS Core suite to run your analysis.

- -- Please select a tool --
- 1.a. Reads processing of short reads
- 1.b. Reads processing of long reads
- 1.c. Reads processing of 454 reads
- 3. Cluster/ASV filters
- 5. Phylogenetic tree building
- ITSx

Sequence file (.fasta) *

9: FROGS Core 1-Main - remove_chimer : sequence.fasta

accepted formats

The sequence file to filter. (--input-fasta)

Abundance file (.biom) *

10: FROGS Core 1-Main - remove_chimera abundance.biom

accepted formats

The abundance file to filter. (--input-biom)

Filter on prevalence

Prevalence across all samples

Minimum number of samples (integer > 1) - optional

3

Set the minimum number of samples in which a ASV/cluster must be detected to be kept. Only ASVs/clusters found in at least this number of samples will be retained. Leave empty to skip prevalence filtering. (--min-sample-presence)

Prevalence by sample groups (metadata file required)

Sample replicates group (.tsv) - optional

Nothing selected

accepted formats

First column indicates the sample name, and the second column the group name. (--replicate-tsv)

Minimum prevalence (%) - optional

Apply the prevalence filter within sample groups defined in a metadata file. The metadata file must include a column specifying the sample groups (batches). Keep ASV/cluster present in at least this proportion of replicates in at least one group (must be a proportion between 0 and 1). Leave empty to skip prevalence filtering. (--min-replicate-presence)

Filter on abundance

Set the minimum ASV/cluster abundance, as a proportion or as a count. We recommend using a proportion of 0.00005.

- Proportion
- Count

Minimum abundance proportion to keep ASV/cluster - optional

5e-05

Leave empty to skip this abundance filtering. For example, 0.00005 (recommended by Bokulich et al., 2013) keeps ASVs/clusters representing at least 0.005% of all sequences. (--min-abundance)

Filter on the most abundant

Number of most abundant ASVs/clusters to keep - optional

Leave empty to skip this filtering. Keeps the N most abundant ASVs/clusters. (--nb-biggest-clusters)

Filter on contaminant

Search for contaminant ASVs/clusters

No contaminant filtering

Use either your own contaminant FASTA file or one provided by FROGS. (--contaminant)

Practice session

- What are the output files of Cluster/ASV Filter tool ?

Explore “FROGS Filter : report.html” file.

- How many clusters have you removed ?
- How many ASVs remain ?
- How are ASVs distributed across the different samples?

Create a Venn diagram using the two filters.

- How many clusters did you remove with each filter?

- How many own ASV remains in AOP1_PPC_S3?

Practice session

What are the output files of 3. Cluster/ASV Filter tool ?

14: **FROGS Core 1-Main - cluster_filters: excluded.tsv**

13: **FROGS Core 1-Main - cluster_filters: abundance.biom**

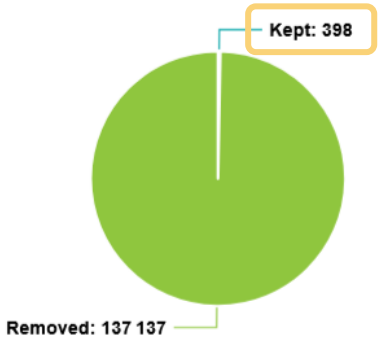
12: **FROGS Core 1-Main - cluster_filters: sequence.fasta**

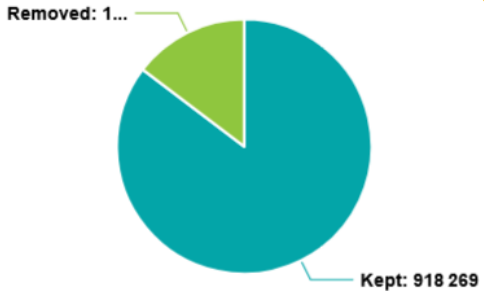
11: **FROGS Core 1-Main - cluster_filters: report.html**

Practice session

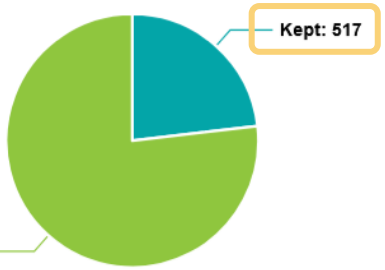
How many clusters have you removed ?
How many ASVs remain ?

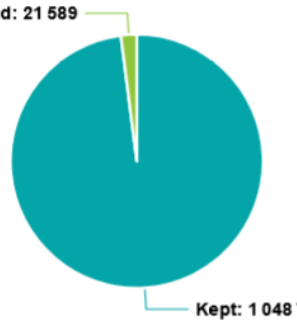
16S swarm



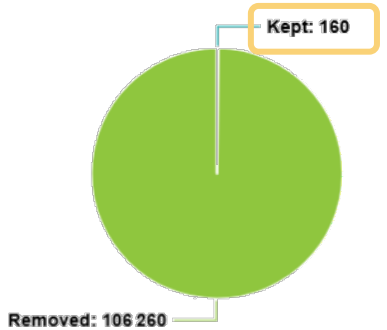


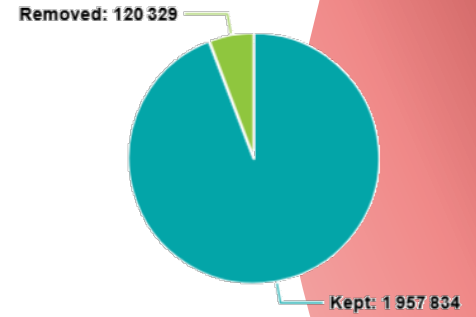
16S DADA2



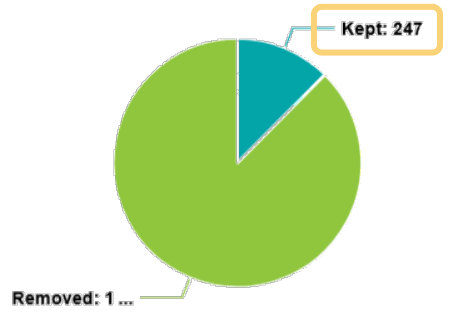


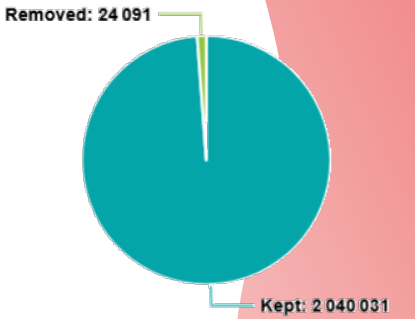
ITS swarm





ITS DADA2

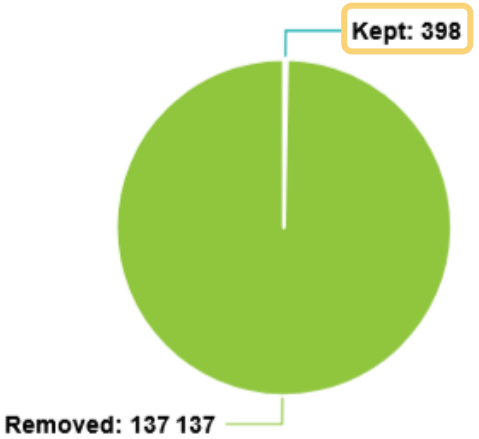




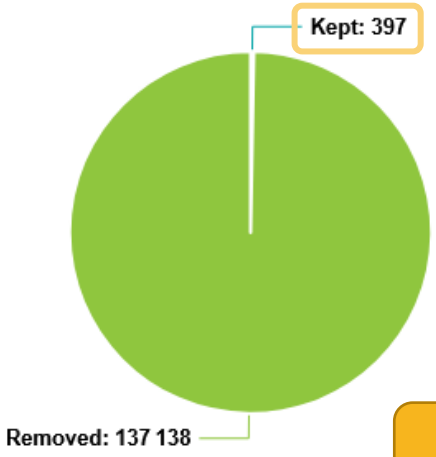
Practice session

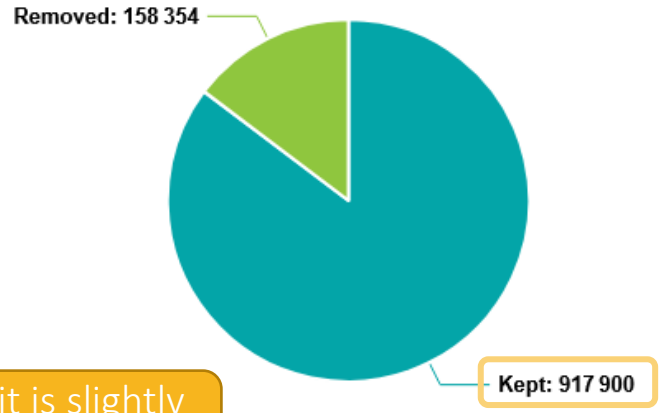
Comparison with a filter applied to batches of samples (replicates)

16S
swarm



16S
swarm





As expected, it is slightly more discriminant.

Tool Parameters

Input Parameter	Value
Select a tool from the FROGS Core suite to run your analysis.	cluster_filters
Sequence file (.fasta)	9: **FROGS Core 1-Main - remove_chimera: sequence.fasta**
Abundance file (.biom)	10: **FROGS Core 1-Main - remove_chimera: abundance.biom**
prevalence_section	
global_prevalence	
Minimum number of samples (integer > 1)	3
replicate_prevalence	
Sample replicates group (.tsv)	
Minimum prevalence (%)	Not available.
abund_section	
Set the minimum ASV/cluster abundance, as a proportion or as a count. We recommend using a proportion of 0.00005.	proportion
Minimum abundance proportion to	0.00005

Tool Parameters

Input Parameter	Value
Select a tool from the FROGS Core suite to run your analysis.	cluster_filters
Sequence file (.fasta)	9: **FROGS Core 1-Main - remove_chimera: sequence.fasta**
Abundance file (.biom)	10: **FROGS Core 1-Main - remove_chimera: abundance.biom**
prevalence_section	
global_prevalence	
Minimum number of samples (integer > 1)	Not available.
replicate_prevalence	
Sample replicates group (.tsv)	45: **MPC_replicate_metadata.tsv**
Minimum prevalence (%)	0.5
abund_section	
Set the minimum ASV/cluster abundance, as a proportion or as a count. We recommend using a proportion of 0.00005.	proportion
Minimum abundance proportion to	0.00005

Practice session

How are ASVs distributed across the different samples?

16S swarm

Sample name	Total number of ASVs	Number of shared ASVs	Number of own ASVs
AOP1_PPC_S1	124	124	0
AOP1_PPC_S2	118	118	0
AOP1_PPC_S3	111	111	0
AOP1_PPC_S4	137	137	0
AOP1_PPC_S5	132	132	0
AOP1_PPC_S6	125	125	0

ASVs **398**

Sequences **918,269**

16S DADA2

Sample name	Total number of ASVs	Number of shared ASVs	Number of own ASVs
AOP1_PPC_S1	87	87	0
AOP1_PPC_S2	85	85	0
AOP1_PPC_S3	75	75	0
AOP1_PPC_S4	104	104	0
AOP1_PPC_S5	101	101	0
AOP1_PPC_S6	89	89	0

ASVs **517**

Sequences **1,048,711**

Here, we observe both a different number of ASVs and a different distribution across the samples.
In this case, DADA2 generates more ASVs in total, but there are fewer per sample.

Practice session

How are ASVs distributed across the different samples?

ITS

swarm

Sample name	Total number of ASVs	Number of shared ASVs	Number of own ASVs
AOP1_PPC_S1	69	69	0
AOP1_PPC_S2	69	69	0
AOP1_PPC_S3	66	66	0
AOP1_PPC_S4	72	72	0
AOP1_PPC_S5	70	70	0
AOP1_PPC_S6	64	64	0

ASVs
160

Sequences
1,957,834

ITS

DADA2

Sample name	Total number of ASVs	Number of shared ASVs	Number of own ASVs
AOP1_PPC_S1	78	78	0
AOP1_PPC_S2	87	87	0
AOP1_PPC_S3	84	84	0
AOP1_PPC_S4	81	81	0
AOP1_PPC_S5	81	81	0
AOP1_PPC_S6	76	76	0

ASVs
247

Sequences
2,040,031

Practice session

How many ASVs do you get ?

Before

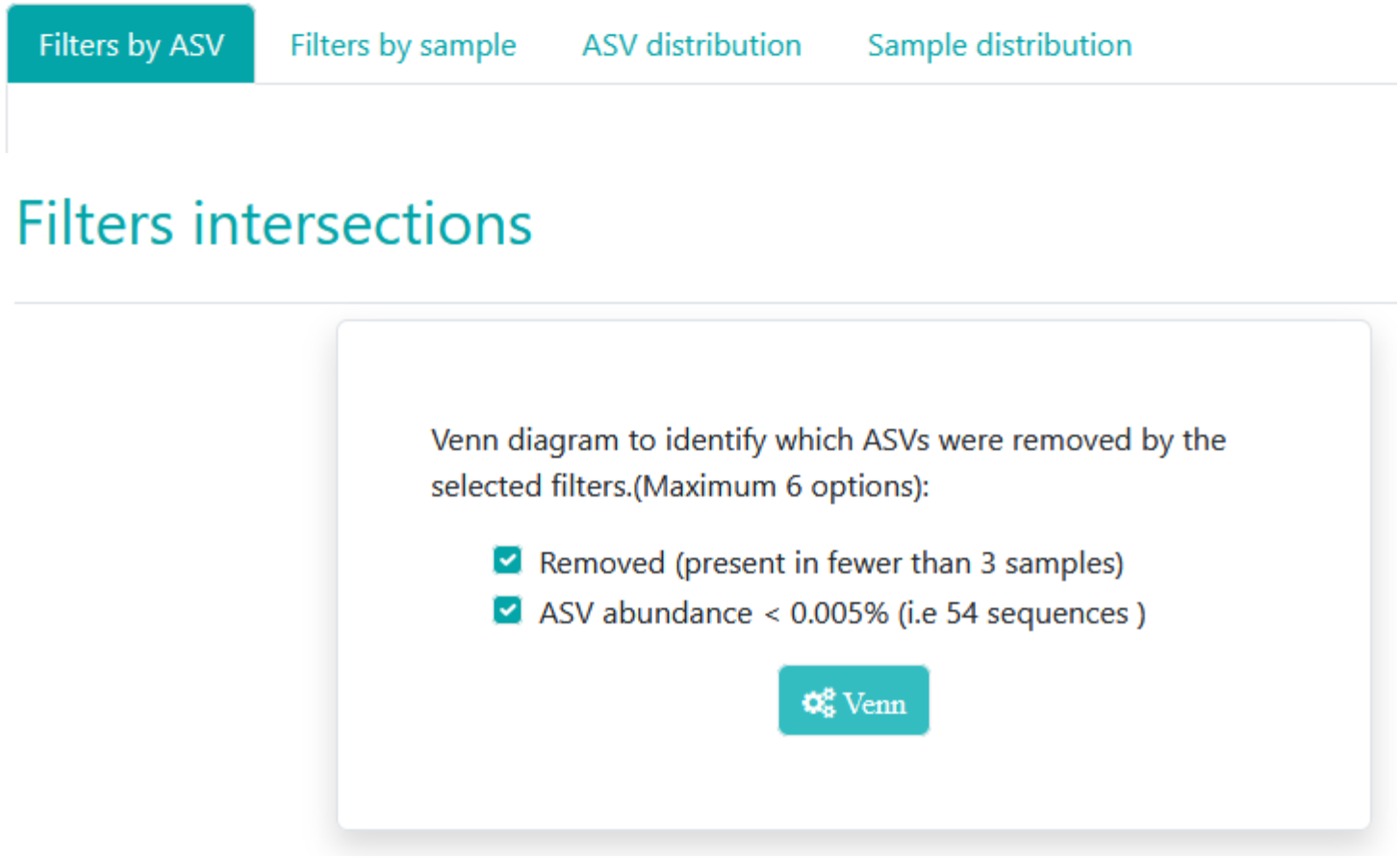
		Clusters/ASVs	Sequences
16S	swarm	137,535	1,076,254
16S	DADA2	2,234	1,070,300
ITS	swarm	106,420	2,078,163
ITS	DADA2	1,984	2,064,122

Now

		ASVs	Sequences
16S	swarm	398	918,269
16S	DADA2	517	1,048,711
ITS	swarm	160	1,957,834
ITS	DADA2	247	2,040,031

Practice session

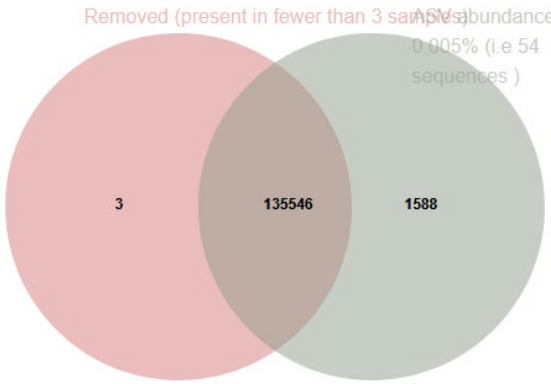
How many clusters did you remove with each filter?



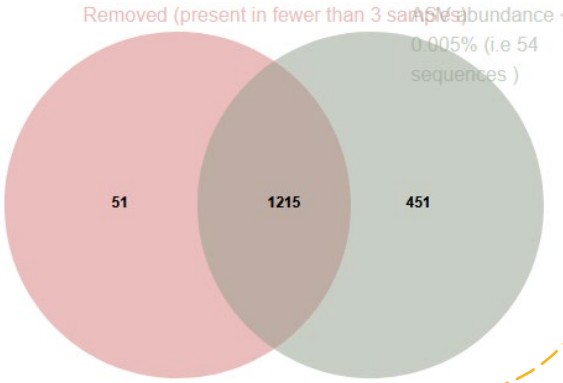
Practice session

How many clusters did you remove with each filter?

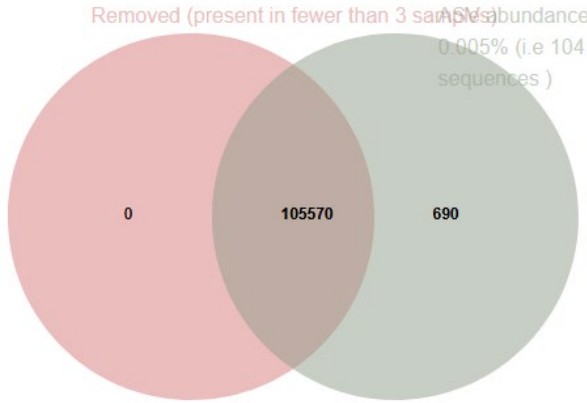
16S
swarm



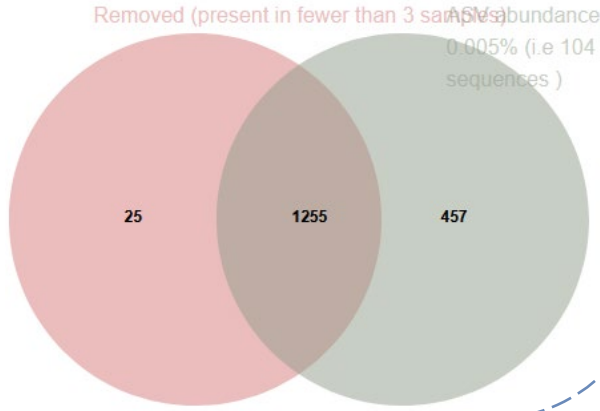
16S
DADA2



ITS
swarm



ITS
DADA2



Practice session

How many own ASV remains in AOP1_PPC_S3 ?

Before

16S

swarm

Sample name	Number of clusters/ASVs	Number of shared clusters/ASVs	Number of own clusters/ASVs
AOP1_PPC_S1	1,999	295	1,704
AOP1_PPC_S2	1,599	261	1,338
AOP1_PPC_S3	1,534	209	1,325
AOP1_PPC_S4	1,868	304	1,564
AOP1_PPC_S5	1,910	301	1,609
AOP1_PPC_S6	1,829	281	1,548

Now

16S

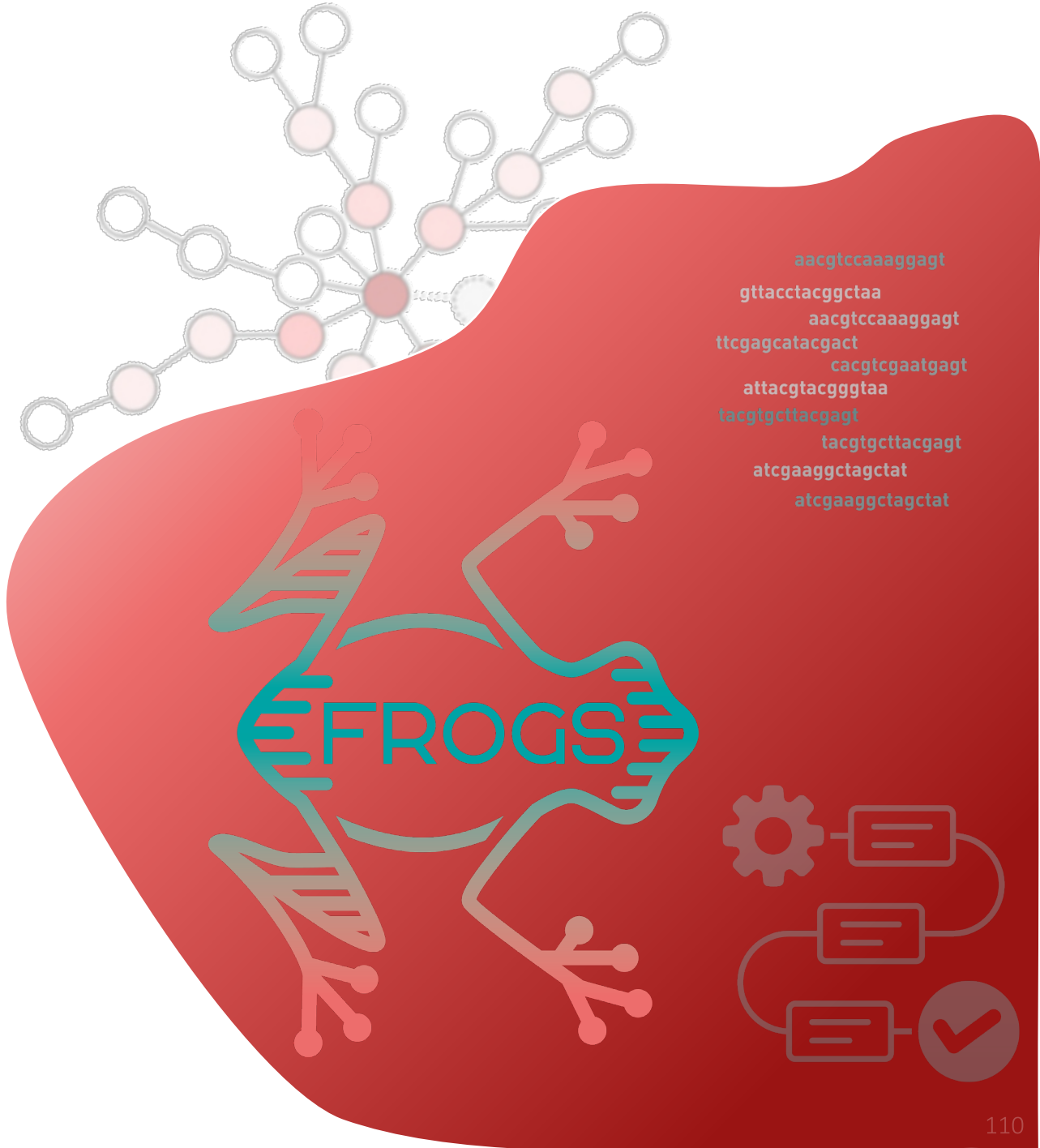
swarm

Sample name	Total number of ASVs	Number of shared ASVs	Number of own ASVs
AOP1_PPC_S1	124	124	0
AOP1_PPC_S2	118	118	0
AOP1_PPC_S3	111	111	0
AOP1_PPC_S4	137	137	0
AOP1_PPC_S5	132	132	0
AOP1_PPC_S6	125	125	0

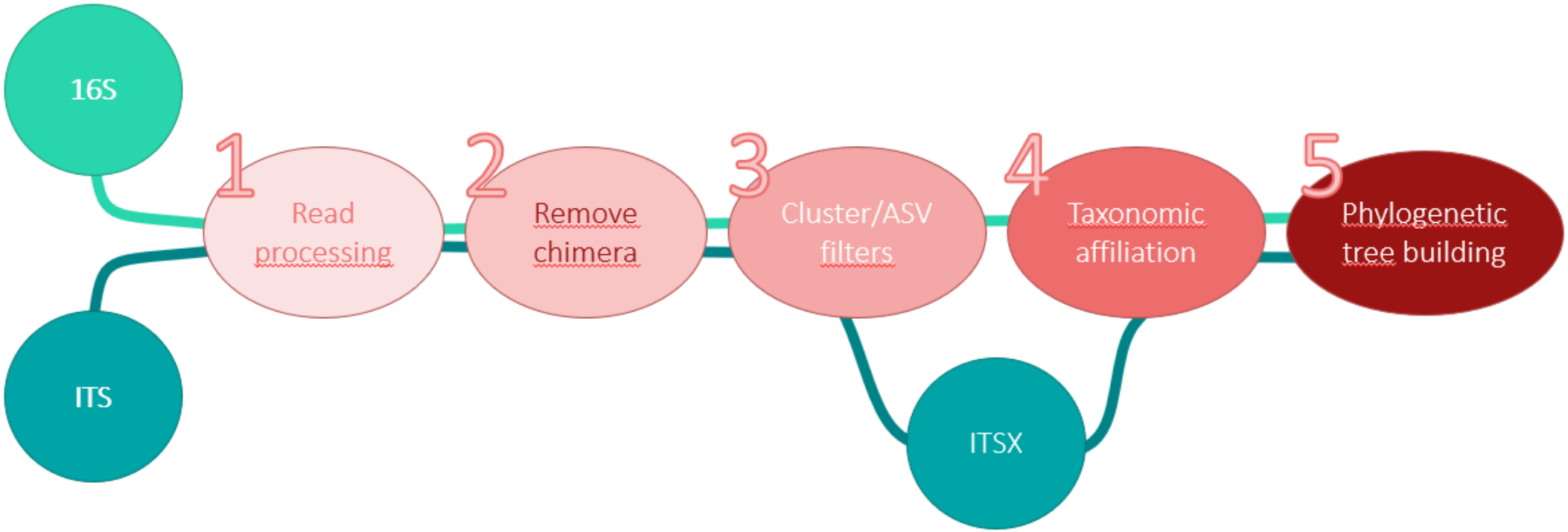
FROGS Core 1

Main tools

ITSX filter



ITS data has an additional filter.



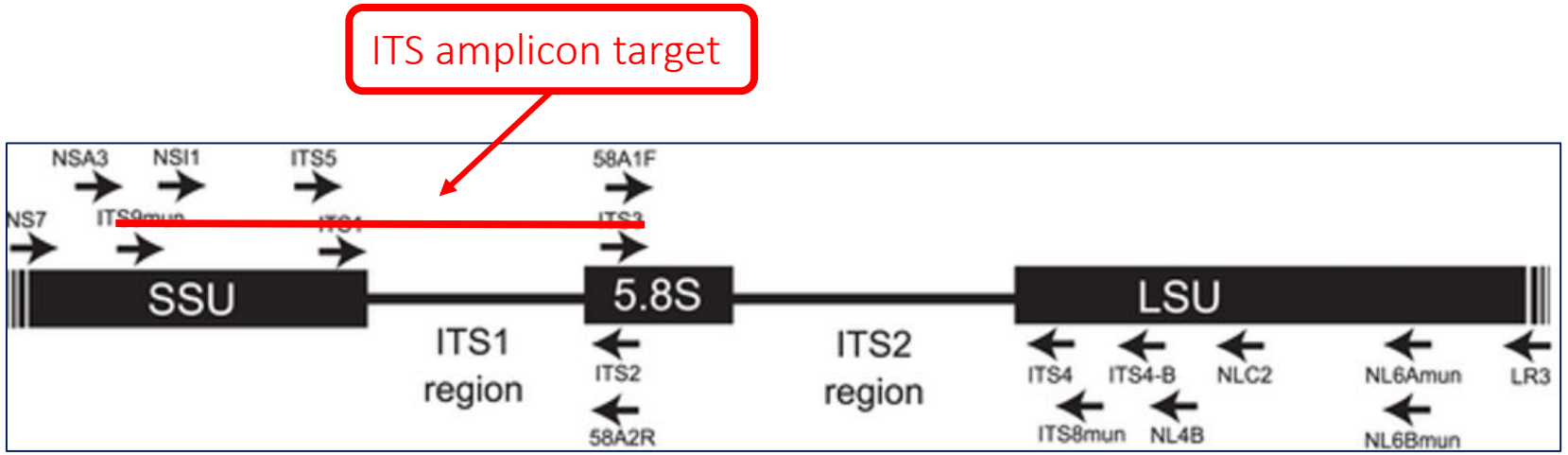
What is the purpose of the ITSx tool?

- ITSx is a tool to filter sequences.
- ITSx identifies and/or trimms ITS regions in sequences.

- Identification: If the ITS1 or ITS2 region is not detected, the sequence is discarded.
- Trimming: It **excludes** the highly conserved neighboring sequences **SSU**, **5S** and **LSU** rRNA.

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

How ITSx tool works ?



1st case: identification

no trimming - keep conserved regions

ITS1 is well detected

SSU part and 5.8S part are not trimmed

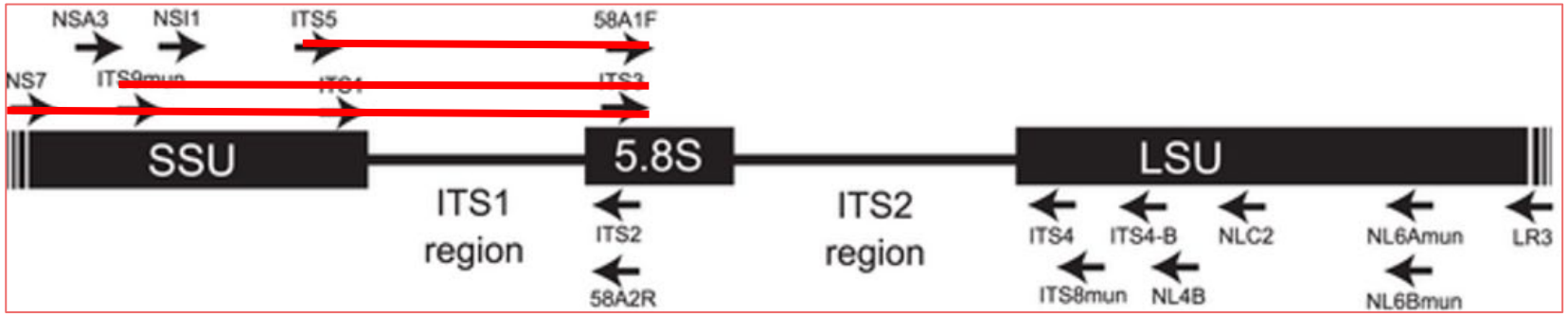
2nd case: identification & trimming of conserved regions

ITS1 is well detected

and SSU part and 5.8S part are trimmed

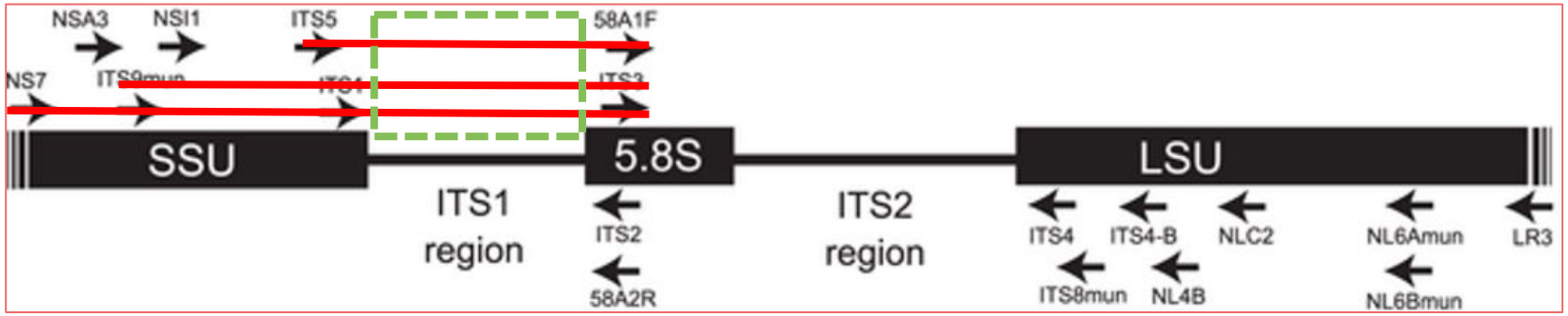
Check only if sequence is identified as ITS? Yes or not?

- If not, only ITS1 or ITS2 part will be conserved
- It is interesting to consider keeping only the ITS parts, without the flanking sequences, in case of:
  - A comparison of sequenced amplicons with different primers that target the same region to be amplified.



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

identification	identification & trimming
Keep more information	A comparison of sequenced amplicons with different primers that target the same region to be amplified
	Using a database with only the ITS part.



Practice session

ITS

Please open the FROGS Core Main ITSX tool and familiarize yourself with the required parameters.

Please apply the following filters:

- No, keep conserved regions
- ITS1
- Fungi

Run the process !

Practice session

Tool Parameters

Select a tool from the FROGS Core suite to run your analysis.

- ( ) -- Please select a tool --
- ( ) 1.a. Reads processing of short reads
- ( ) 1.b. Reads processing of long reads
- ( ) 1.c. Reads processing of 454 reads
- ( ) 2. Remove chimera
- ( ) 3. Cluster/ASV filters
- ( ) 4. Taxonomic affiliation
- ( ) 5. Phylogenetic tree building
- (•) ITSx

Sequence file (.fasta) *

17: FROGS Core 1-Main - cluster_filters: sequence.fasta

accepted formats

The sequence file to filter. (--input-fasta)

Abundance file (.biom) *

18: FROGS Core 1-Main - cluster_filters: abundance.biom

accepted formats

The abundance file to filter. (--input-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: - optional

- [−] Select / Deselect all
- [x] Fungi
- [ ] Alveolata
- [ ] Bryophyta
- [ ] Bacillariophyta
- [ ] Amoebozoa
- [ ] Euglenozoa
- [ ] Chlorophyta
- [ ] Rhodophyta
- [ ] Phaeophyceae
- [ ] Marchantiophyta
- [ ] Metazoa
- [ ] Oomycota
- [ ] Haptophyceae
- [ ] Raphidophyceae
- [ ] Rhizaria
- [ ] Synurophyceae
- [ ] Tracheophyta
- [ ] Eustigmatophyceae

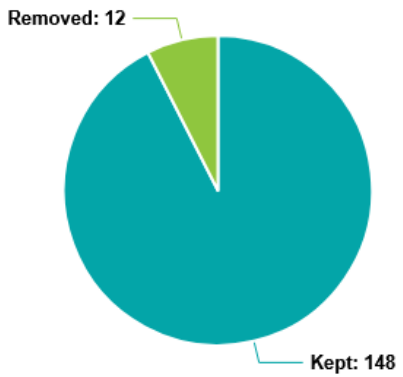
Save a lot of time by checking pertinent organism group model to scan (--organism-groups)

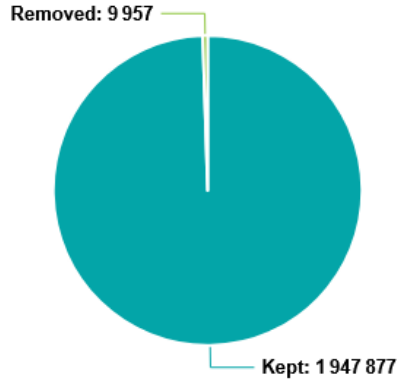
Practice session

ITS

ITS

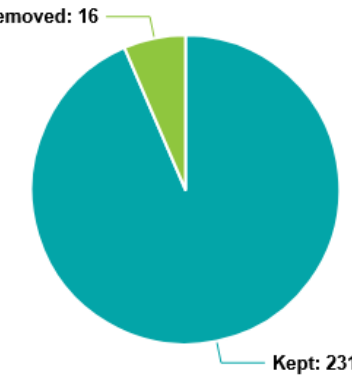
swarm

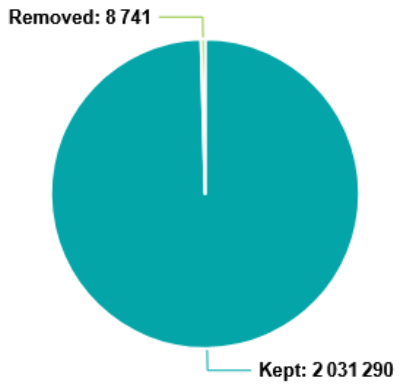




ITS

DADA2

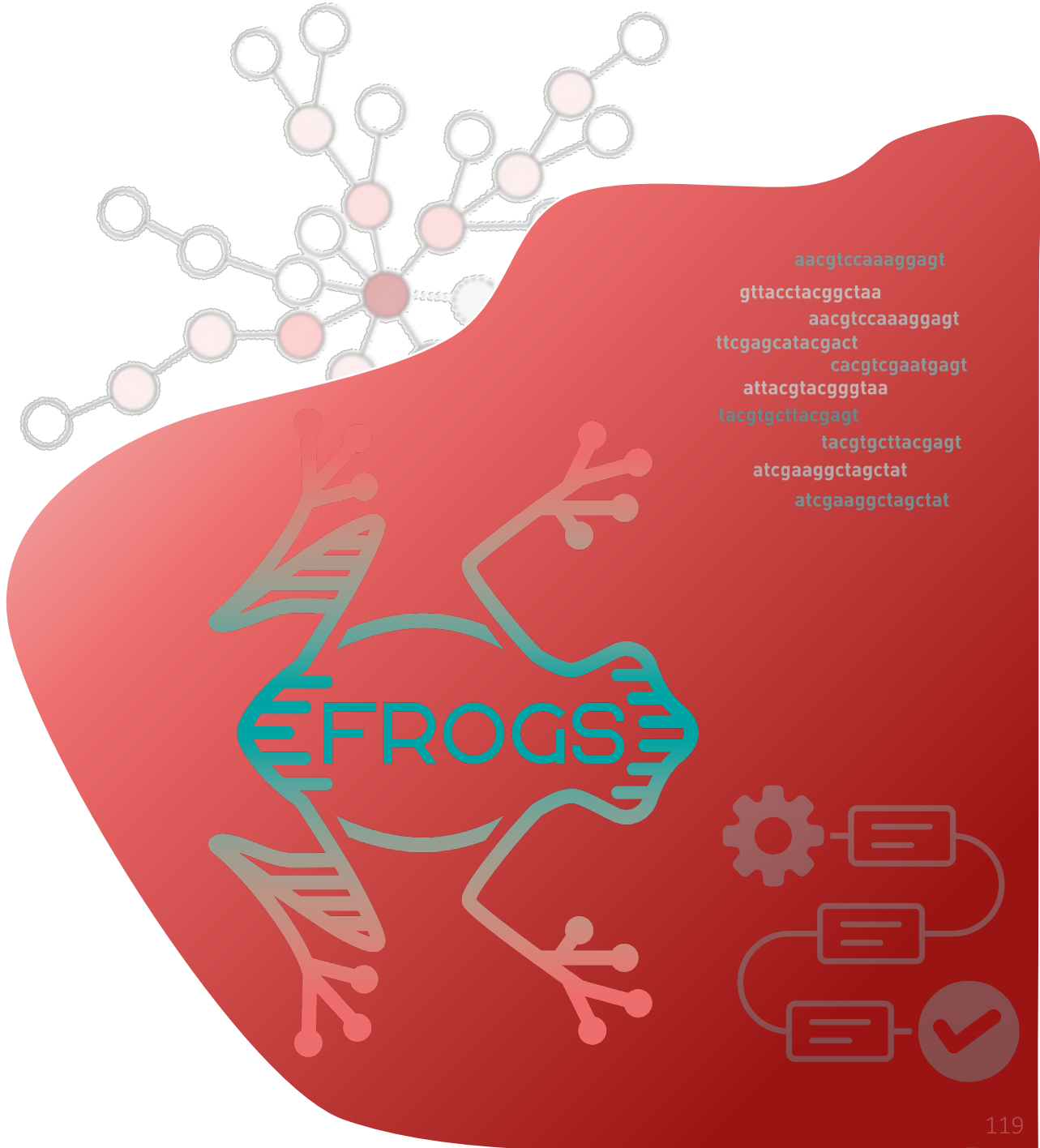




FROGS Core 1

Main tools

Taxonomic affiliation



What does this tool do?

Taxonomic affiliation tool assigns a taxonomic identification to ASVs.

Taxonomic affiliations are determined by comparing ASVs to sequences contained within databases.

Which databases should you choose?

→ LEAP is a companion website

Learn and Evaluate Affiliation databanks on an online Platform

1 Cluster = 2 affiliations

RDPClassifier*: one affiliation with bootstrap, on each taxonomic subdivision.

Bacteria;(1.0);Actinobacteriota;(1.0);Actinobacteria;(1.0);Propionibacteriales;(1.0);Propionibacteriaceae;(1.0);Cutibacterium;(1.0);Cutibacterium acnes;(0.57);

NCBI Blastn+** : one affiliation with identity %, query coverage %, subject coverage %, e-value, alignment length and a special tag "**Multi-affiliation**".

k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Corynebacteriales;f__Corynebacteriaceae;g__Corynebacterium;s__Corynebacterium_nuruki

blast_subject	blast_perc_identity	blast_perc_query_coverage	blast_perc_subject_coverage	blast_evalue	blast_aln_length
Corynebacterium_nuruki_HM165487_TS	99.02	100.0	28.274428274428274	0.0	408

* Appl. Environ. Microbiol. August 2007 vol. 73 no. 16 5261-5267. doi : 10.1128/AEM.00062-07
Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy.
Qiong Wang, George M.Garrity, James M. Tiedje and James R. Cole

** BMC Bioinformatics 2009, 10:421. doi:10.1186/1471-2105-10-421
BLAST+: architecture and applications
Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos,Kevin Bealer and Thomas L Madden

Affiliation Strategy of FROGS

Blastn+ with “**Multi-affiliation**” management

k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__Streptococcus_thermophilus
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__Streptococcus_salivarius

Strictly identical (V3V4 amplification) on 426 nucleotides

Which one to choose?

Affiliation Strategy of FROGS

Blastn+ with "**Multi-affiliation**" management

k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__Streptococcus_thermophilus
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__Streptococcus_salivarius

Strictly identical (V3V4 amplification) on 426 nucleotides.

k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;**Multi-affiliation**

We cannot make a choice without having preconceived ideas.

Practice session

Please visit the LEAP website https://leap.frogs.toulouse.inrae.fr/

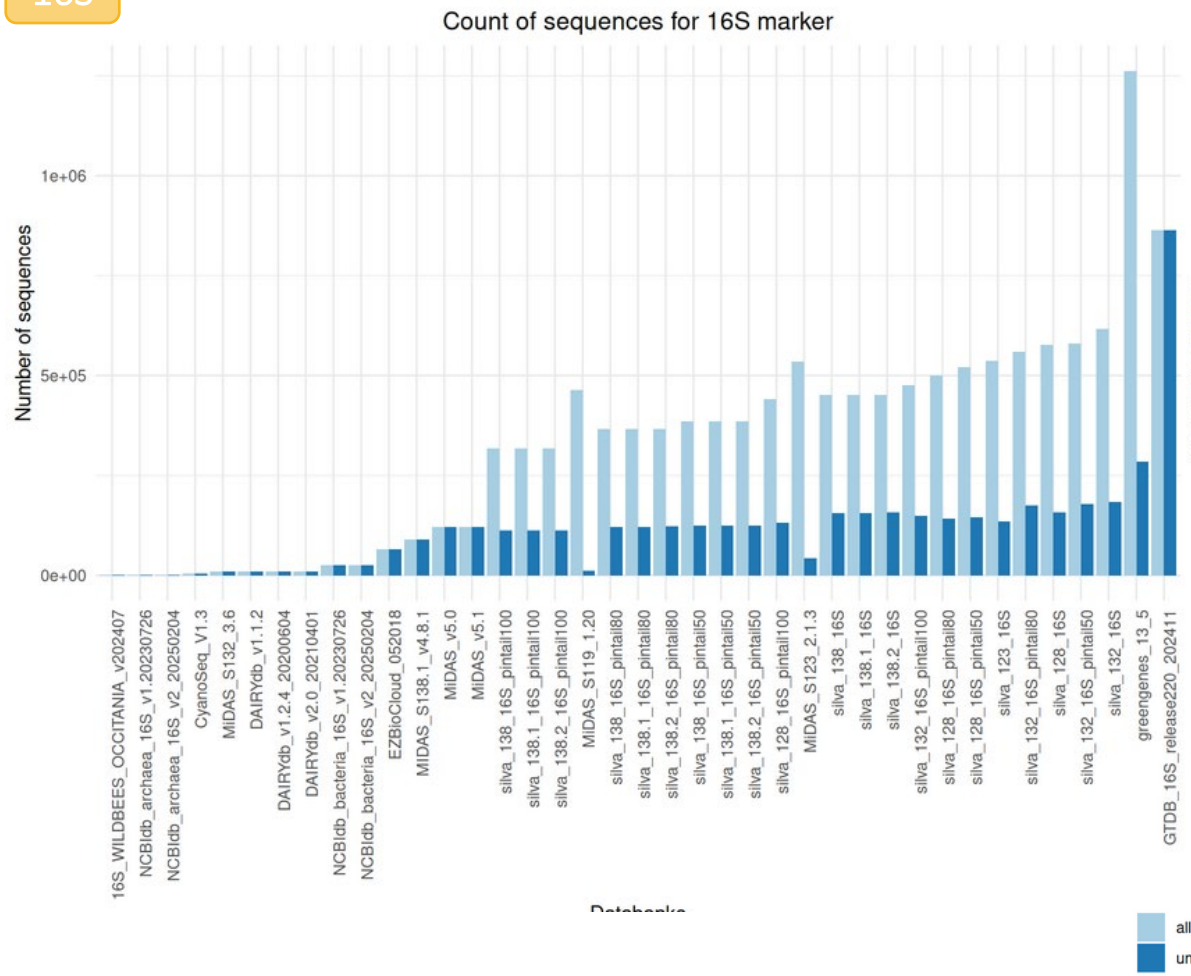
Which banks specialise in 16S sequencing ?
Which banks specialise in the ITS marker?

Focus on 16S SILVA and 16S SILVA pintail 100, 80 and 50

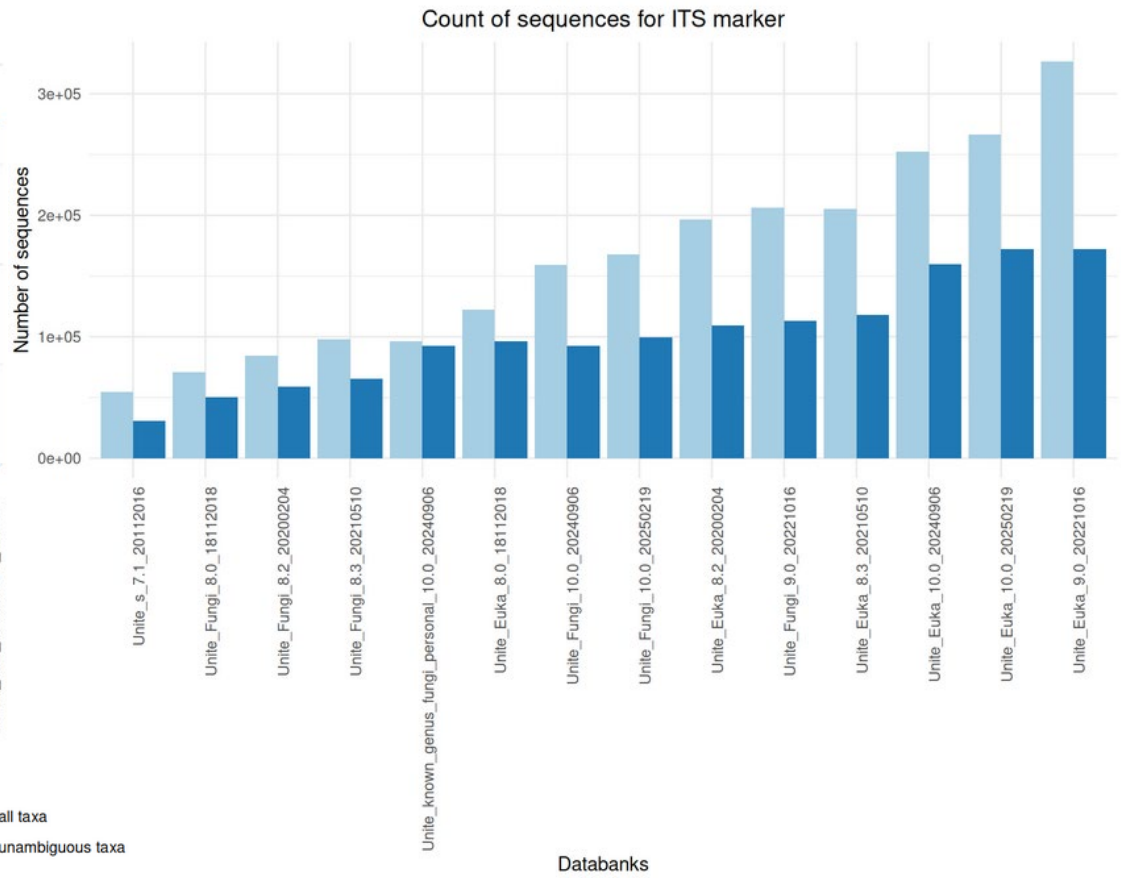
Practice session

Which banks specialise in 16S sequencing ?
Which banks specialise in the ITS marker?

16S



ITS



Practice session

Focus on 16S SILVA and 16S SILVA pintail 100, 80 and 50

Pintail* represents the probability that the rRNA sequence contains anomalies or is a chimera.

Pintail 100 means that the probability for being anomalous or chimeric is low.

4 ranks of available databases in FROGS: 50 pintail, 80 pintail or 100 pintail or no pintail filter.

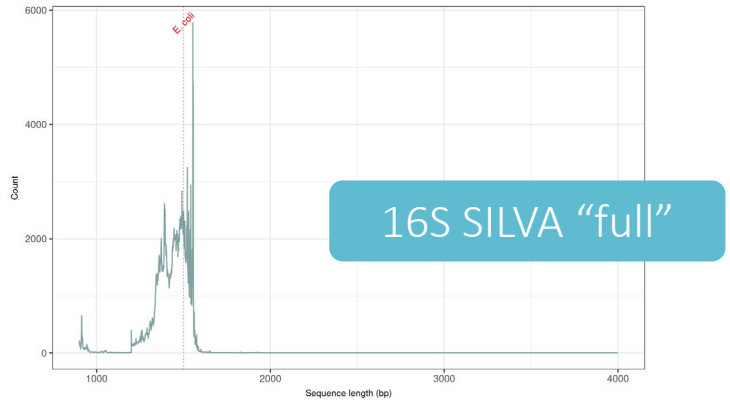
Name	Date	Marker	Sequence number
16S_SILVA_138.2	Oct 1, 2024	16S	451555
16S_SILVA_Pintail100_138.2	Oct 1, 2024	16S	317241
16S_SILVA_Pintail50_138.2	Oct 1, 2024	16S	385166
16S_SILVA_Pintail80_138.2	Oct 1, 2024	16S	366036

Only for 16S !

* http://aem.asm.org/content/71/12/7724.abstract

Practice session

Focus on 16S SILVA and 16S SILVA pintail 100, 80 and 50



16S SILVA "full"

Length statistics

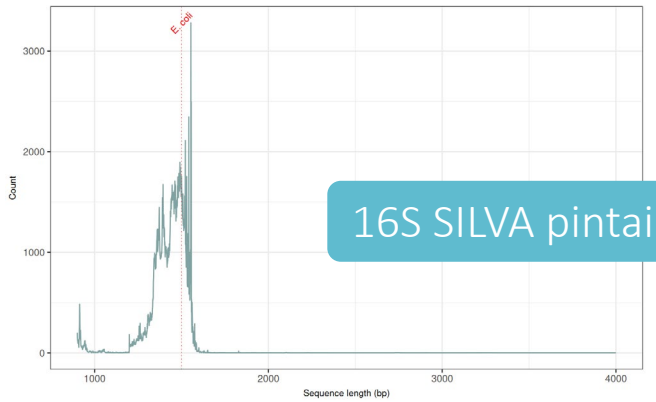
Min	Q1	Mean	Median	Q3	Max
900	1224	1612	1549	1887	4000

Unique taxa

Kingdom	Phylum	Class	Order	Family	Genus	Species
2	112	283	728	1331	5128	43532

Detail of ambiguous key-words

keyword	count
metagenome	8017
unknown	283779
unspecified	0
unidentified	461
unclassified	1
incertae	1696
NA	4
None	0



16S SILVA pintail 100

Length statistics

Min	Q1	Mean	Median	Q3	Max
900	1183	1531	1464	1764	4000

Unique taxa

Kingdom	Phylum	Class	Order	Family	Genus	Species
2	112	280	717	1301	4968	36441

Detail of ambiguous key-words

keyword	count
metagenome	5521
unknown	197060
unspecified	0
unidentified	385
unclassified	0
incertae	1206
NA	4
None	0

Practice session

Focus on 16S SILVA and 16S SILVA pintail 100, 80 and 50

SILVA pintail 100, 80 and 50 comparison

SILVA pintail 100 : clean but small, possibly species missing

SILVA: All the deposited sequences, including artefactual sequences and those with poor annotations.

Your choice of SILVA databank depends on your data, experiments, and habits.

Practice session

Please open the FROGS Core Main 4 tool and familiarize yourself with the required parameters.

Please apply the following filters:

- For 16S marker : ASVs will be affiliated with the 16S_DAIRYdb_V1.1.2 databank.
- For ITS marker: ASVs will be affiliated with the ITS_UNITE_Fungi_10.0 databank.

Run the process !

Practice session

16S

Tool Parameters

Select a tool from the FROGS Core suite to run your analysis.

- ( ) -- Please select a tool --
- ( ) 1.a. Reads processing of short reads
- ( ) 1.b. Reads processing of long reads
- ( ) 1.c. Reads processing of 454 reads
- ( ) 2. Remove chimera
- ( ) 3. Cluster/ASV filters
- (•) 4. Taxonomic affiliation
- ( ) 5. Phylogenetic tree building
- ( ) ITSx

Sequence file (.fasta) *

12: FROGS Core 1-Main - cluster_filters: sequence.fasta

accepted formats

The sequence file to be taxonomically affiliated. (--input-fasta)

Abundance file (.biom) *

13: FROGS Core 1-Main - cluster_filters: abundance.biom

accepted formats

The abundance file linked to the sequence to be taxonomically affiliated. (--input-biom)

Choose the reference database *

16S DAIRYdb V1.1.2

Select the reference database from the list.

Run RDP taxonomic affiliation in addition to BLAST *

- ( ) Yes
- (•) No

Taxonomic affiliation is performed using BLAST. Enabling this option also runs taxonomic affiliation with the RDP Classifier (default: No). (--rdp)

Taxonomic ranks *

Domain Phylum Class Order Family Genus Species

Specify the ordered taxonomic rank levels used in the reference database. Separate each rank by a single space. (--taxonomic-ranks)

Practice session

ITS

Tool Parameters

Select a tool from the FROGS Core suite to run your analysis.

- ( ) -- Please select a tool --
- ( ) 1.a. Reads processing of short reads
- ( ) 1.b. Reads processing of long reads
- ( ) 1.c. Reads processing of 454 reads
- ( ) 2. Remove chimera
- ( ) 3. Cluster/ASV filters
- (•) 4. Taxonomic affiliation
- ( ) 5. Phylogenetic tree building
- ( ) ITSx

Sequence file (.fasta) *

14: FROGS Core 1-Main - itsx: sequence.fasta

accepted formats

The sequence file to be taxonomically affiliated. (--input-fasta)

Abundance file (.biom) *

15: FROGS Core 1-Main - itsx: abundance.biom

accepted formats

The abundance file linked to the sequence to be taxonomically affiliated. (--input-biom)

Choose the reference database *

ITS UNITE Fungi 10.0 20250219

Select the reference database from the list.

Run RDP taxonomic affiliation in addition to BLAST *

- ( ) Yes
- (•) No

Taxonomic affiliation is performed using BLAST. Enabling this option also runs taxonomic affiliation with the RDP Classifier (default: No). (--rdp)

Taxonomic ranks *

Domain Phylum Class Order Family Genus Species

Specify the ordered taxonomic rank levels used in the reference database. Separate each rank by a single space. (--taxonomic-ranks)

Practice session

What are the outputs ?

How many ASVs are affiliated by BLAST ?

How many ASV have a “multiaffiliation” at Genus ranks ?

Practice session

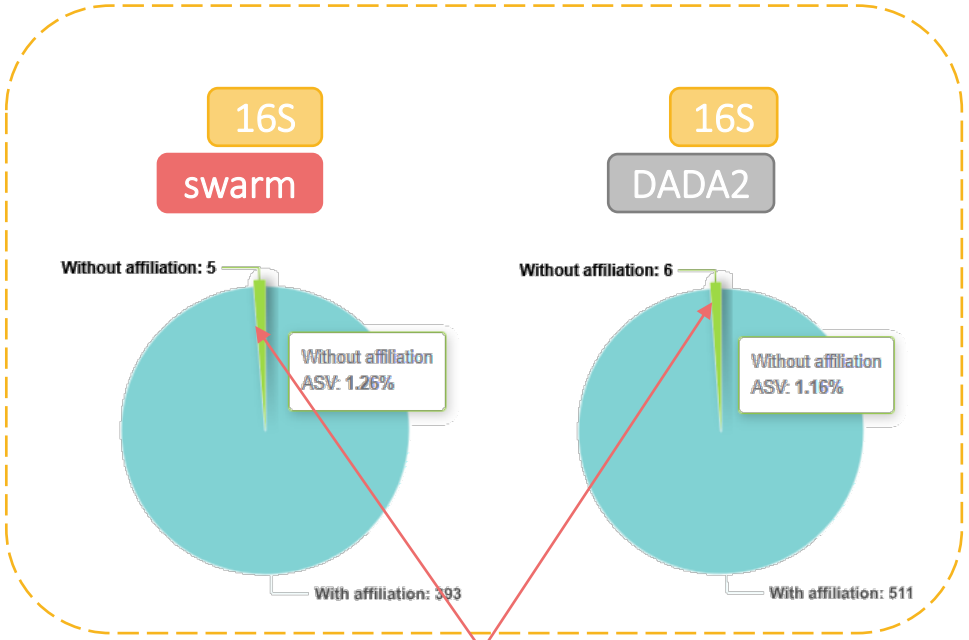
What are the outputs ?

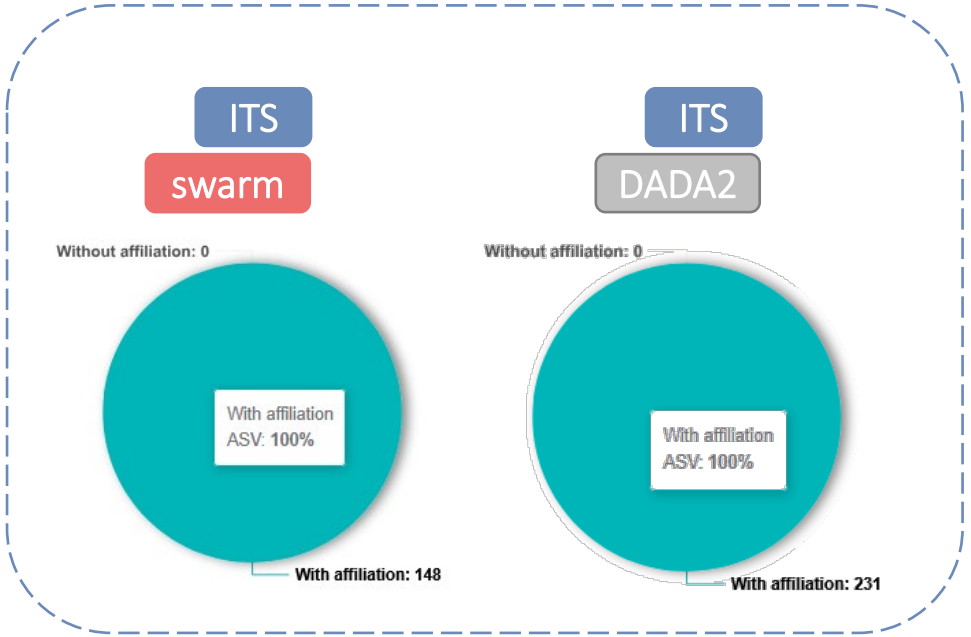
39: **FROGS Core 1-Main - taxonomic_affiliation: abundance.biom**

38: **FROGS Core 1-Main - taxonomic_affiliation: report.html**

No human readable !

How many ASVs are affiliated by BLAST ?





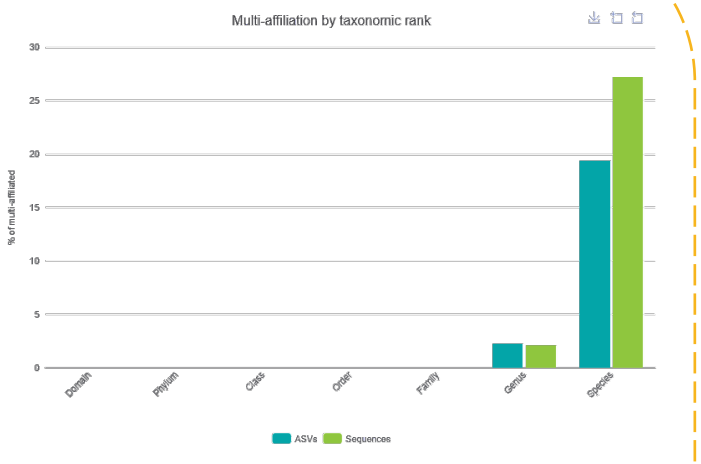
These ASVs are too distant with subject in databank and are not retained.

Practice session

How many ASV have a “multiaffiliation” at Genus ranks ?

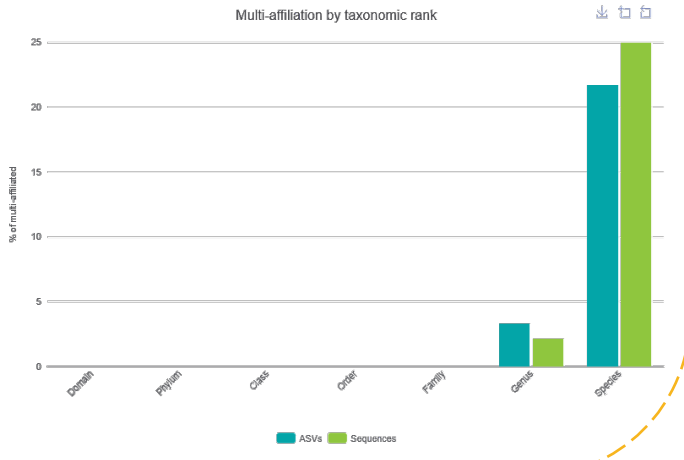
16S swarm

Blast multi-affiliation summary



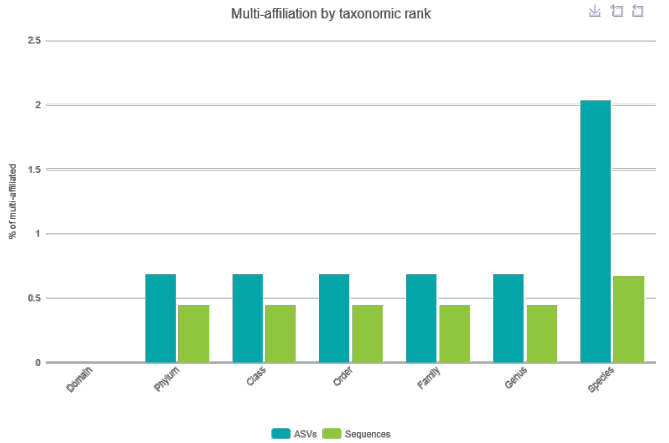
16S DADA2

Blast multi-affiliation summary



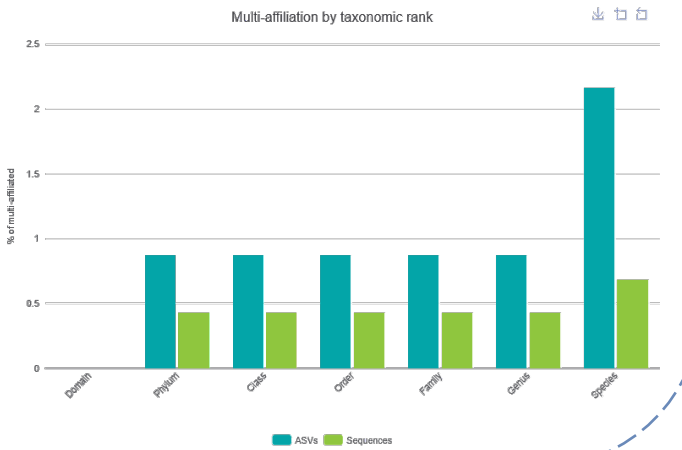
ITS swarm

Blast multi-affiliation summary



ITS DADA2

Blast multi-affiliation summary



Practice session

Which tool do you need to use to read the abundance table in TSV format?

Run it !

Practice session

Which tool do you need to use to read the abundance table in TSV format?

FROGS Core 2-Companion Optional process, converter, and report (Galaxy Version 5.1.0+galaxy0)

Tool Parameters

Select a tool from the FROGS Core suite to run your analysis.

- -- Please select a tool --
- Read demultiplexing
- Affiliation filters
- Affiliation postprocessing
- Abundance normalisation
- Convert Biom file to TSV file
- Convert TSV file to Biom file
- Cluster/ASV report
- Affiliation report

16S

Abundance file (.biom) *

39: FROGS Core 1-Main - taxonomic_affiliation: abundance.biom

accepted formats

The abundance file to convert. (--input-biom)

Sequence file (.fasta) - optional

12: FROGS Core 1-Main - cluster_filters: sequence.fasta

accepted formats

The sequences file. If you use this option the sequences will be added to the TSV. (--input-fasta)

Extract multi-affiliation *

- Yes
- No

This option will extract information about multiple blast affiliation in a second TSV file.

FROGS Core 2-Companion Optional process, converter, and report (Galaxy Version 5.1.0+galaxy0)

Tool Parameters

Select a tool from the FROGS Core suite to run your analysis.

- -- Please select a tool --
- Read demultiplexing
- Affiliation filters
- Affiliation postprocessing
- Abundance normalisation
- Convert Biom file to TSV file
- Convert TSV file to Biom file
- Cluster/ASV report
- Affiliation report

ITS

Abundance file (.biom) *

37: FROGS Core 1-Main - taxonomic_affiliation: abundance.biom

accepted formats

The abundance file to convert. (--input-biom)

Sequence file (.fasta) - optional

21: FROGS Core 1-Main - itsx: sequence.fasta

accepted formats

The sequences file. If you use this option the sequences will be added to the TSV. (--input-fasta)

Extract multi-affiliation *

- Yes
- No

This option will extract information about multiple blast affiliation in a second TSV file.

Outputs :

45: **FROGS Core 2-Companion - biom_to_tsv: multi-affiliations.tsv**

44: **FROGS Core 2-Companion - biom_to_tsv: abundance.tsv**

Practice session

Download these 2 TSV files from “Convert biom file to TSV file”.

Open them in Excel.

Which ASVs are dominant?

What are their affiliations ?

How is the ASV ID1 (for 16S - focus on AOP1 samples) and ASV ID2 (for ITS – focus on AOP6 samples) distributed across different samples?

For the 16S only, how can ASVs without affiliation be identified?

Practice session

Which ASVs are dominant?
What are their affiliations ?

16S swarm

blast_taxonomy	blast_subject	blast_perc_identity	blast_perc_query_coverage	blast_perc_subject_coverage	blast_evalue	blast_aln_length	seed_id	seed_sequ	observation_name	observation_sum
k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Corynebacteriales;f__Corynebacteriaceae;g__Corynebacterium;s__Corynebacterium_Species	Corynebacterium_Species_AJ222817	100	100	28.39248434	0	408	H7:1:HTNVF	TGGGGAA	ID_1	13739
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;Multi-affiliation	multi-subject	100	100	multi-coverage	0	426	H7:1:HTNVF	TAGGGAA	ID_2	10636
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Lactococcus;s__Lactococcus_lactis	Lactococcus_lactis_AB681295_TS	100	100	28.94021739	0	426	H7:1:HTNVF	TAGGGAA	ID_3	6639

16S DADA2

blast_taxonomy	blast_subject	blast_perc_identity	blast_perc_query_coverage	blast_perc_subject_coverage	blast_evalue	blast_aln_length	seed_id	seed_sequ	observation_name	observation_sum
k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Corynebacteriales;f__Corynebacteriaceae;g__Corynebacterium;s__Corynebacterium_Species	Corynebacterium_Species_AJ222	100	100	28.39248434	0	408	None	TGGGGAA	ID_1	120273
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;Multi-affiliation	multi-subject	100	100	multi-coverage	0	426	None	TAGGGAA	ID_2	91846
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Lactococcus;s__Lactococcus_lactis	Lactococcus_lactis_AB681295_TS	100	100	28.94021739	0	426	None	TAGGGAA	ID_3	76802

How is the ASV ID1 distributed across different samples?

16S swarm

AOP1_PPC_S1	AOP1_PPC_S2	AOP1_PPC_S3	AOP1_PPC_S4	AOP1_PPC_S5	AOP1_PPC_S6	AOP1_PPC_W1	AOP1_PPC_W2	AOP1_PPC_W3	AOP1_PPC_W4	AOP1_PPC_W5	AOP1_PPC_W6	AOP2_PPC_S1	AOP2_PPC_S2	AOP2_PPC_S3	AOP2_PPC_S4	AOP2_PPC_S5	AOP2_PPC_S6
986	720	838	1072	1775	1003	947	654	552	809	450	615	665	549	514	29	32	24

16S DADA2

AOP1_PPC_S1	AOP1_PPC_S2	AOP1_PPC_S3	AOP1_PPC_S4	AOP1_PPC_S5	AOP1_PPC_S6	AOP1_PPC_W1	AOP1_PPC_W2	AOP1_PPC_W3	AOP1_PPC_W4	AOP1_PPC_W5	AOP1_PPC_W6	AOP2_PPC_S1	AOP2_PPC_S2	AOP2_PPC_S3	AOP2_PPC_S4	AOP2_PPC_S5	AOP2_PPC_S6
0	0	0	0	0	0	1222	0	0	0	0	0	547	472	450	25	25	27

how would do you explain it ?

Practice session

How would do you explain it ?

16S swarm

blast_taxonomy	blast_subject	blast_perc_identity	blast_perc_query_coverage	blast_perc_subject_coverage	blast_evalue	blast_aln_length	seed_id	seed_sequ	observation_name
k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Corynebacteriales;f__Corynebacteriaceae;g__Corynebacterium;s__Corynebacterium_Species	Corynebacterium_Species_AJ222817	100	100	28.39248434	0	408	H7:1:HTNVI	TGGGGAA	ID_1

observation_sum	AOP1_PPC_S1	AOP1_PPC_S2	AOP1_PPC_S3	AOP1_PPC_S4	AOP1_PPC_S5	AOP1_PPC_S6	AOP1_PPC_W1	AOP1_PPC_W2	AOP1_PPC_W3	AOP1_PPC_W4	AOP1_PPC_W5	AOP1_PPC_W6	AOP2_PPC_S1	AOP2_PPC_S2	AOP2_PPC_S3	AOP2_PPC_S4	AOP2_PPC_S5	AOP2_PPC_S6	AOP2_PPC_W1	AOP2_PPC_W2	AOP2_PPC_W3	AOP2_PPC_W4
137399	986	720	838	1072	1775	1003	947	654	552	809	450	615	665	549	514	29	32	24	347	286	218	

16S DADA2

blast_taxonomy	blast_subject	blast_perc_identity	blast_perc_query_coverage	blast_perc_subject_coverage	blast_evalue	blast_aln_length	seed_id	seed_sequ	observation_name
k__Bacteria;p__Actinobacteri;c__Actinobacteria;o__Corynebacteriales;f__Corynebacteriaceae;g__Corynebacterium;s__Corynebacterium_Species	Corynebacterium_Species_AJ222	100	100	28.39248434	0	408	None	TGGGGAA	ID_1
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;Multi-affiliation	multi-subject	100	100	multi-coverage	0	426	None	TAGGGAA	ID_2
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Lactococcus;s__Lactococcus_lactis	Lactococcus_lactis_AB681295_T	100	100	28.94021739	0	426	None	TAGGGAA	ID_3
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae;g__Lactobacillus;s__Lactobacillus_delbrueckii	Lactobacillus_delbrueckii_AB007	100	100	28.10026385	0	426	None	TAGGGAA	ID_4
k__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Staphylococcaceae;g__Staphylococcus;s__Staphylococcus_equorum	Staphylococcus_equorum_AB009	100	100	28.51405622	0	426	None	TAGGGAA	ID_5
k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Oceanospirillales;f__Halomonadaceae;g__Halomonas;s__Halomonas_variabilis	Halomonas_variabilis_HG795418	100	100	28.06324111	0	426	None	TGGGGAA	ID_6
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Leuconostocaceae;g__Leuconostoc;s__Leuconostoc_mesenteroides	Leuconostoc_mesenteroides_ABC	100	100	29.31865107	0	426	None	TAGGGAA	ID_7
k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Micrococcales;f__Brevibacteriaceae;g__Brevibacterium;Multi-affiliation	multi-subject	multi-identity	100	multi-coverage	0	410	None	TGGGGAA	ID_8
k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Corynebacteriales;f__Corynebacteriaceae;g__Corynebacterium;s__Corynebacterium_casei	Corynebacterium_casei_HQ20286	100	100	29.02542373	0	411	None	TGGGGAA	ID_9
k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Micrococcales;f__Brevibacteriaceae;g__Brevibacterium;s__Brevibacterium_aurantiacum	Brevibacterium_aurantiacum_X76	99.268	100	27.81546811	0	410	None	TGGGGAA	ID_10
k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Moraxellaceae;g__Psychrobacter;s__Psychrobacter_pacificensi	Psychrobacter_pacificensis_AB01	97.887	100	27.91612058	0	426	None	TGGGGAA	ID_11
k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Corynebacteriales;f__Corynebacteriaceae;g__Corynebacterium;s__Corynebacterium_nuruki	Corynebacterium_nuruki_HM1654	99.02	100	28.27442827	0	408	None	TGGGGAA	ID_12
k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;Multi-affiliation	multi-subject	99.765	100	multi-coverage	0	426	None	TAGGGAA	ID_13
k__Bacteria;p__Actinobacteri;c__Actinobacteria;o__Corynebacteriales;f__Corynebacteriaceae;g__Corynebacterium;s__Corynebacterium_Species	Corynebacterium_Species_AJ222	99.755	100	28.32289492	0	408	None	TGGGGAA	ID_14

observation_sum	AOP1_PPC_S1	AOP1_PPC_S2	AOP1_PPC_S3	AOP1_PPC_S4	AOP1_PPC_S5	AOP1_PPC_S6	AOP1_PPC_W1	AOP1_PPC_W2	AOP1_PPC_W3	AOP1_PPC_W4	AOP1_PPC_W5	AOP1_PPC_W6	AOP2_PPC_S1	AOP2_PPC_S2	AOP2_PPC_S3	AOP2_PPC_S4	AOP2_PPC_S5	AOP2_PPC_S6	AOP2_PPC_W1	AOP2_PPC_W2	AOP2_PPC_W3	AOP2_PPC_W4
120273	0	0	0	0	0	0	1222	0	0	0	0	0	547	472	450	25	25	27	413	356	283	
91846	0	0	0	0	0	0	568	0	0	0	0	0	1075	666	1718	1045	682	916	2192	1108	927	63
76802	0	0	18	0	17	0	34	0	0	0	13	0	0	5	0	0	0	0	947	887	388	
69919	0	0	0	0	0	0	0	0	0	0	0	0	45	43	127	0	0	0	0	0	0	
26856	0	0	0	0	0	0	230	0	0	0	0	0	254	255	305	384	483	439	855	694	706	49
24740	0	10	0	0	0	0	280	0	0	0	0	0	2346	2692	2579	932	445	611	1314	1330	1824	383
23849	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	71	80	21	
20022	0	0	0	0	0	0	1242	0	0	0	0	0	373	283	321	387	500	377	410	311	397	55
19924	0	0	0	0	0	0	833	0	0	0	0	0	669	713	532	405	670	503	1459	1569	1061	115
18707	0	0	0	0	0	0	1430	0	0	0	0	0	498	376	474	411	547	447	390	268	362	33
18401	0	3	0	0	0	0	0	0	0	0	0	7	278	219	313	1719	1471	1616	739	1419	953	38
17987	0	0	0	0	0	0	1030	0	0	0	0	0	2686	2833	2217	1045	2024	1378	873	860	505	
13438	1184	870	814	1900	1290	1455	0	182	1245	2030	1411	1057	0	0	0	0	0	0	0	0	0	
12447	1306	957	1134	1456	2339	1356	0	828	719	1001	570	781	0	0	0	0	0	0	0	0	0	

Dada2 has identified two ASVs that belong to the same species. Perhaps they are different strains …? These two ASVs are distributed differently across the various samples. Biostatistical analysis will allow us to explore these interesting data more closely.

Practice session

Which ASVs are dominant?
What are their affiliations ?

ITS swarm

blast_taxonomy	blast_subject	blast_perc_identity	blast_perc_query_coverage	blast_perc_subject_coverage	blast_evalue	blast_aln_length	seed_id	seed_sequ	observation_name	observation_sum
k__Fungi;p__Ascomycota;c__Saccharomycetes;o__Saccharomycetales;f__Dipodascaceae;g__Geotrichum;s__Geotrichum_candidum	LC032055_SH1393006.10FU_	100	90.73170732	66.66666667	2.86E-94	186	V1:1:HGK71	GAAACGC(	ID_1	521467
k__Fungi;p__Ascomycota;c__Saccharomycetes;o__Saccharomycetales;f__Debaryomycetaceae;g__Debaryomyces;s__Debaryomyces_prosopidis	JN942657_SH1242810.10FU_	99.707	100	51.43288084	1.59E-178	341	V1:1:HGK71	GAAATGC(	ID_2	531723
k__Fungi;p__Ascomycota;c__Sordariomycetes;o__Hypocreales;f__Nectriaceae;g__Bisifusarium;s__Bisifusarium_domesticum	JQ434585_SH1298109.10FU_	100	87.45980707	57.62711864	7.03E-142	272	V1:1:HGK71	GAAATGC(	ID_3	249502

ITS DADA2

blast_taxonomy	blast_subject	blast_perc_identity	blast_perc_query_coverage	blast_perc_subject_coverage	blast_evalue	blast_aln_length	seed_id	seed_sequ	observation_name	observation_sum
k__Fungi;p__Ascomycota;c__Saccharomycetes;o__Saccharomycetales;f__Dipodascaceae;g__Geotrichum;s__Geotrichum_candidum	LC032055_SH1393006.10FU_reps	100	90.73170732	66.66666667	2.86E-94	186	None	GAAACGC(	ID_1	534937
k__Fungi;p__Ascomycota;c__Saccharomycetes;o__Saccharomycetales;f__Debaryomycetaceae;g__Debaryomyces;s__Debaryomyces_prosopidis	JN942657_SH1242810.10FU_refs	99.707	100	51.43288084	1.59E-178	341	None	GAAATGC(	ID_2	391681
k__Fungi;p__Ascomycota;c__Sordariomycetes;o__Hypocreales;f__Nectriaceae;g__Bisifusarium;s__Bisifusarium_domesticum	JQ434585_SH1298109.10FU_refs	100	87.45980707	57.62711864	7.03E-142	272	None	GAAATGC(	ID_3	252465

How is the ASV ID2 distributed across AOP6 samples ?

ITS swarm

AOP6_PPS_S1	AOP6_PPS_S2	AOP6_PPS_S3	AOP6_PPS_S4	AOP6_PPS_S5	AOP6_PPS_S6	AOP6_PPS_W1	AOP6_PPS_W2	AOP6_PPS_W3	AOP6_PPS_W4	AOP6_PPS_W5	AOP6_PPS_W6
423	62	25	1021	178	614	3496	2227	1965	194	321	309
11376	13459	16705	17114	14528	13998	19674	14466	14571	16313	16138	15150

ITS DADA2

AOP6_PPS_S1	AOP6_PPS_S2	AOP6_PPS_S3	AOP6_PPS_S4	AOP6_PPS_S5	AOP6_PPS_S6	AOP6_PPS_W1	AOP6_PPS_W2	AOP6_PPS_W3	AOP6_PPS_W4	AOP6_PPS_W5	AOP6_PPS_W6
429	64	25	1029	176	634	3575	2280	2027	203	330	315
240	512	336	7831	5948	6669	130	122	164	4876	5471	5346

how would do you explain it ?

Practice session

How would do you explain it ?

ITS swarm

blast_taxonomy
k__Fungi;p__Ascomycota;c__Saccharomycetes;o__Saccharomycetales;f__Dipodascaceae;g__Geotrichum;s__Geotrichum_candidum
k__Fungi;p__Ascomycota;c__Saccharomycetes;o__Saccharomycetales;f__Debaryomycetaceae;g__Debaryomyces;s__Debaryomyces_prosopidis

AOP6_PPS_S1	AOP6_PPS_S2	AOP6_PPS_S3	AOP6_PPS_S4	AOP6_PPS_S5	AOP6_PPS_S6	AOP6_PPS_W1	AOP6_PPS_W2	AOP6_PPS_W3	AOP6_PPS_W4	AOP6_PPS_W5	AOP6_PPS_W6
423	62	25	1021	178	614	3496	2227	1965	194	321	309
11376	13459	16705	17114	14528	13998	19674	14466	14571	16313	16138	15150

ITS DADA2

blast_taxonomy
k__Fungi;p__Ascomycota;c__Saccharomycetes;o__Saccharomycetales;f__Dipodascaceae;g__Geotrichum;s__Geotrichum_candidum
k__Fungi;p__Ascomycota;c__Saccharomycetes;o__Saccharomycetales;f__Debaryomycetaceae;g__Debaryomyces;s__Debaryomyces_prosopidis
k__Fungi;p__Ascomycota;c__Sordariomycetes;o__Hypocreales;f__Nectriaceae;g__Bisifusarium;s__Bisifusarium_domesticum
k__Fungi;p__Ascomycota;c__Saccharomycetes;o__Saccharomycetales;f__Debaryomycetaceae;g__Debaryomyces;s__Debaryomyces_prosopidis

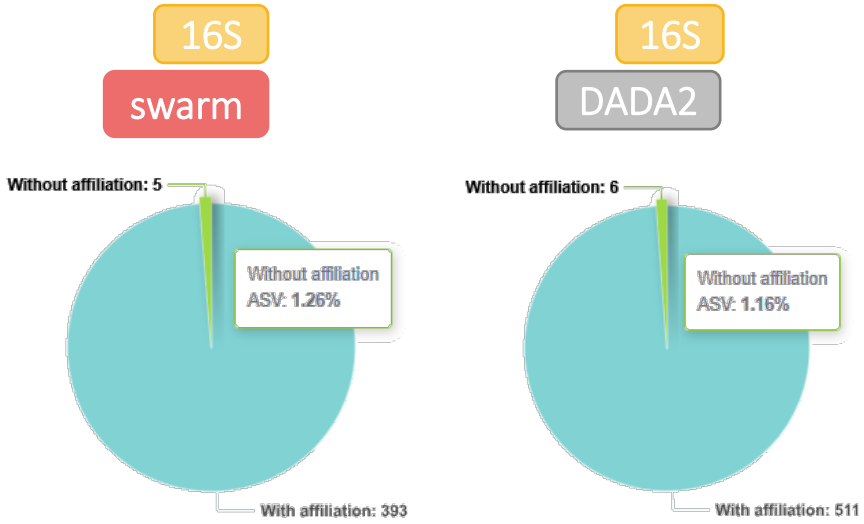
AOP6_PPS_S1	AOP6_PPS_S2	AOP6_PPS_S3	AOP6_PPS_S4	AOP6_PPS_S5	AOP6_PPS_S6	AOP6_PPS_W1	AOP6_PPS_W2	AOP6_PPS_W3	AOP6_PPS_W4	AOP6_PPS_W5	AOP6_PPS_W6
429	64	25	1029	176	634	3575	2280	2027	203	330	315
240	512	336	7831	5948	6669	130	122	164	4876	5471	5346
0	0	0	0	0	0	0	0	0	0	0	0
11629	13584	17312	8797	9232	7971	20361	15043	15134	12148	11158	10305

Dada2 has identified two ASVs that belong to the same species. Perhaps they are different strains ...? These two ASVs are distributed differently across the various samples. Biostatistical analysis will allow us to explore these interesting data more closely.

Practice session

For the 16S only, how can ASVs without affiliation be identified?

16S



no data		no data		no data		no data								no data		no data					
H7:1:HLYVI	GCAATGC	ID_159	165	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	12	18	23

'No data' replaces all affiliation metadata.
No affiliation has been found for this ASV, but it is widely distributed across the various samples.
In this dataset, the “without affiliation” entries relate only to a very small minority of ASVs.

Practice session

What is the distinctive feature of the ASV ID2 (for 16S) and ASV ID19 or ASV ID15 (for ITS swarm or DADA2 respectively) affiliations ?

Explore the respective 'multi-affiliation.tsv' files to understand why this special tag is used.

Practice session

What is the distinctive feature of the ASV ID2 (for 16S) and ASV ID19 or ASV ID15 (for ITS swarm or DADA2 respectively) affiliations ?

'multi-affiliation.tsv' files analysis

16S

#observation_name	blast_taxonomy	blast_subject	blast_perc_identity	blast_perc_query_coverage	blast_perc_subject_coverage	blast_evalue	blast_aln_leng
ID_2	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__Streptococcus_thermophil	Streptococcus_thermophilus_KC429781	100	100	30.84721217	0	42
ID_2	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__Streptococcus_salivarius	Streptococcus_salivarius_AY188354_TS	100	100	27.68031189	0	42

ITS

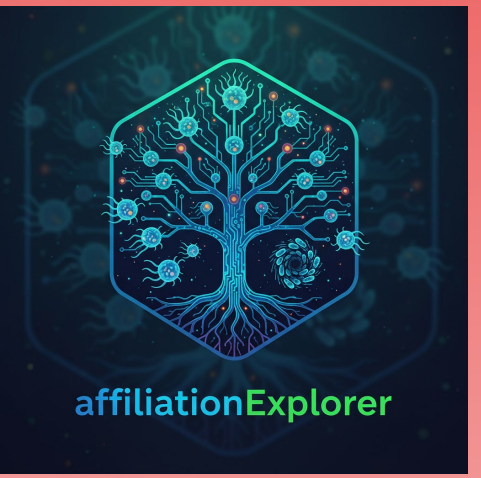
#observation_name		blast_taxonomy	blast_subject	blast_perc_identity	blast_perc_query_covera	blast_perc_subject_coverag	blast_evalue	blast_aln_length
ID_15	ID_19	k__Fungi;p__Ascomycota;c__Saccharomycetes;o__Saccharomycetales;f__Metschnikowiaceae;g__Clavispora;s__Clavispora_lusitaniae	KP674552_SH1056863.10FU_	100	81.86046512	59.25925926	1.09E-88	17
ID_15	ID_19	k__Fungi;p__Ascomycota;c__Saccharomycetes;o__Saccharomycetales;f__Metschnikowiaceae;g__Clavispora;s__Clavispora_lusitaniae	MT534186_SH1056848.10FU_r	100	81.86046512	60.06825939	1.09E-88	17
ID_15	ID_19	k__Fungi;p__Ascomycota;c__Saccharomycetes;o__Saccharomycetales;f__Metschnikowiaceae;g__Clavispora;s__Clavispora_lusitaniae	KP674537_SH1056845.10FU_	100	81.86046512	60.2739726	1.09E-88	17
ID_15	ID_19	k__Fungi;p__Ascomycota;c__Saccharomycetes;o__Saccharomycetales;f__Metschnikowiaceae;g__Clavispora;s__Clavispora_lusitaniae	AY321464_SH1056834.10FU_r	100	81.86046512	59.25925926	1.09E-88	17
ID_15	ID_19	k__Fungi;p__Ascomycota;c__Saccharomycetes;o__Saccharomycetales;f__Metschnikowiaceae;g__Clavispora;s__Clavispora_lusitaniae	KP131851_SH0005010.10FU_re	100	81.86046512	54.15384615	1.09E-88	17
ID_15	ID_19	k__Fungi;p__Ascomycota;c__Saccharomycetes;o__Saccharomycetales;f__Metschnikowiaceae;g__Clavispora;s__Clavispora_lusitaniae	KP764960_SH0005006.10FU_	100	81.86046512	60.48109966	1.09E-88	17
ID_15	ID_19	k__Fungi;p__Ascomycota;c__Saccharomycetes;o__Saccharomycetales;f__Metschnikowiaceae;g__Clavispora;s__Clavispora_lusitaniae	KP764965_SH0005003.10FU_	100	81.86046512	59.25925926	1.09E-88	17
ID_15	ID_19	k__Fungi;p__Ascomycota;c__Saccharomycetes;o__Saccharomycetales;f__Saccharomycetales_fam_Incertae_sedis;g__Saccharomycetales_gen_Incertae_sedis;s__Saccharomycetales_sp	KY073532_SH1056831.10FU_r	100	81.86046512	65.67164179	1.09E-88	17
ID_15	ID_19	k__Fungi;p__Fungi_phy_Incertae_sedis;c__Fungi_cls_Incertae_sedis;o__Fungi_ord_Incertae_sedis;f__Fungi_fam_Incertae_sedis;g__Fungi_gen_Incertae_sedis;s__Fungi_sp	HG937182_SH0206531.10FU_r	100	81.86046512	46.19422572	1.09E-88	17

To easily modify your abundance files, you should use the ‘affiliationExplorer' tool.

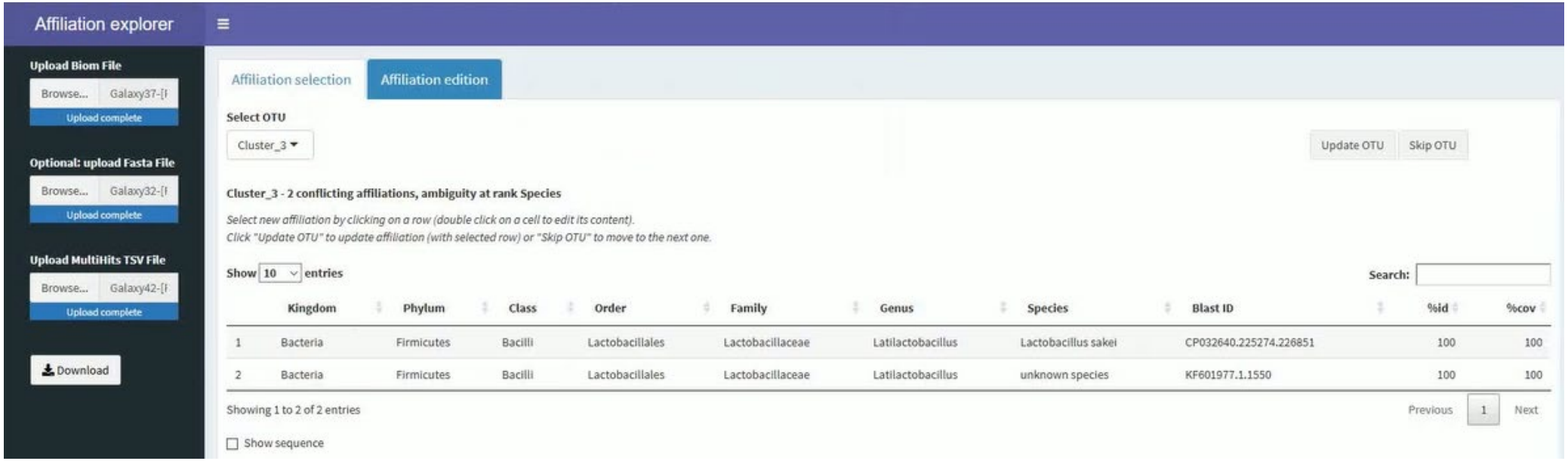
affliationExplorer is a companion tool of FROGS.

- Shiny app for exploration and disambiguation of FROGS multi-affiliation.
- The goal of affiliationExplorer is to provide a user-friendly graphical interface to choose between different conflicting affiliations.
- The app will then sort the multi-affiliated taxa from most abundant to least abundant and help you pick one (or none) of the conflicting affiliations.
- You can also edit the affiliation manually.
  The documentation: https://forge.inrae.fr/migale/affiliationexplorer

https://shiny.migale.inrae.fr/app/affiliationexplorer



affiliationExplorer



A very user-friendly tool, developed by Mahendra Mariadassou and his collaborators (Maiage unit - INRAE Jouy-en-Josas). It allows to modify very simply the affiliations of an abundance table from FROGS.

affiliationExplorer

Demo video

Practice session

Explore other tabs of taxonomic_affiliation_report.html file.

Explore Taxonomy Distribution tab.

- Build the global distribution.
- Build the distribution for AOP1 samples.
- Build the rarefaction curves at species rank.

Explore Blast alignment metrics Distribution tab.

How would you interpret these matrices?

Practice session

Explore Taxonomy Distribution tab. Build the global distribution.

Summary | Taxonomy Distribution | Blast Alignment metrics Distribution

Display global distribution

Search

Sample name	Number of Domain	Number of Phylum	Number of Class	Number of Order	Numb
AOP1_PPC_S1	1	4	6	12	22
AOP1_PPC_S2	1	4	5	11	22
AOP1_PPC_S3	1	4	6	12	20
AOP1_PPC_S4	1	4	6	14	27
AOP1_PPC_S5	1	4	6	15	28
AOP1_PPC_S6	1	4	5	13	25
AOP1_PPC_W1	1	4	8	14	25
AOP1_PPC_W2	1	4	6	12	22
AOP1_PPC_W3	1	4	5	11	22
AOP1_PPC_W4	1	4	7	15	26

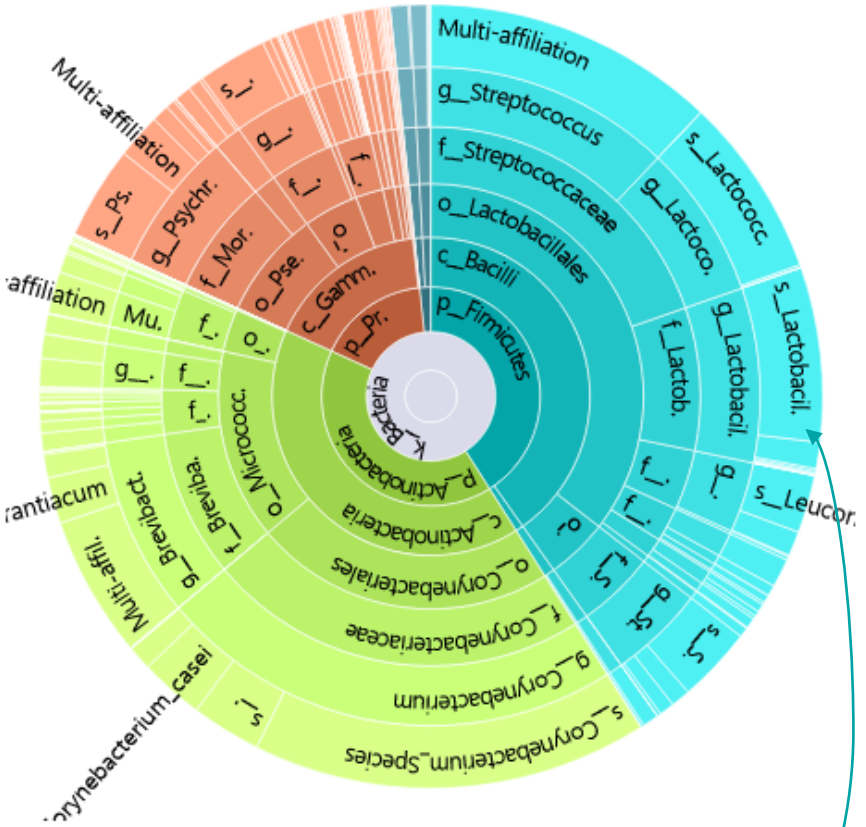
Showing 1 to 10 of 72 rows 10 rows per page

‹ 1 2 3 4 5 ... 8 ›

With selection:

Display distribution

Display rarefaction Domain



Detail on selected:

Name	Size	Global %	Parent %
root	917846		
k__Bacteria	917846	100.000	100.000
p__Firmicutes	376387	41.008	41.008
c__Bacilli	368919	40.194	98.016
o__Lactobacillales	325683	35.483	88.280
f__Lactobacillaceae	78681	8.572	24.159
g__Lactobacillus	78681	8.572	100.000
s__Lactobacillus_delbrueckii	66117	7.203	84.032

s__Lactobacillus_delbrueckii nb children: 0

Font size: 22 Colors start depth: 2 Close

Practice session

Explore Taxonomy Distribution tab. Build the distribution for AOP1 samples.

Summary | **Taxonomy Distribution** | Blast Alignment metrics Distribution

Display global distribution

Search

	Sample name	Number of Domain	Number of Phylum	Number of Class	Number of Order	Numb
☐						
☑	AOP1_PPC_S1	1	4	6	12	22
☑	AOP1_PPC_S2	1	4	5	11	22
☑	AOP1_PPC_S3	1	4	6	12	20
☑	AOP1_PPC_S4	1	4	6	14	27
☑	AOP1_PPC_S5	1	4	6	15	28
☑	AOP1_PPC_S6	1	4	5	13	25
☐	AOP1_PPC_W1	1	4	8	14	25
☐	AOP1_PPC_W2	1	4	6	12	22
☐	AOP1_PPC_W3	1	4	5	11	22
☐	AOP1_PPC_W4	1	4	7	15	26

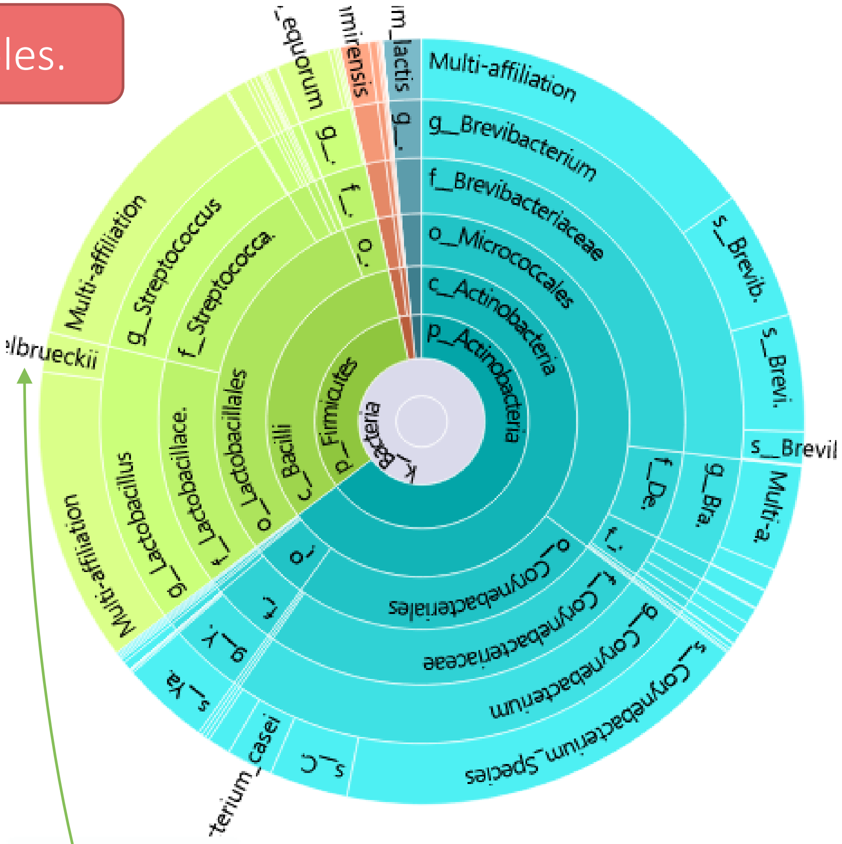
Showing 1 to 10 of 72 rows 10 rows per page

‹ 1 2 3 4 5 ... 8 ›

With selection:

Display distribution

Display rarefaction Species



Detail on selected:

Name	Size	Global %	Parent %
root	48353		
k__Bacteria	48353	100.000	100.000
p__Firmicutes	15414	31.878	31.878
c__Bacilli	15413	31.876	99.994
o__Lactobacillales	14100	29.161	91.481
f__Lactobacillaceae	6811	14.086	48.305
g__Lactobacillus	6811	14.086	100.000
s__Lactobacillus_delbrueckii	804	1.663	11.804

s__Lactobacillus_delbrueckii nb children: 0

Font size: 22 Colors start depth: 2 Close

Practice session

Explore Taxonomy Distribution tab. Build the rarefaction curves at species rank.

Summary | Taxonomy Distribution | Blast Alignment metrics Distribution

Display global distribution

Search

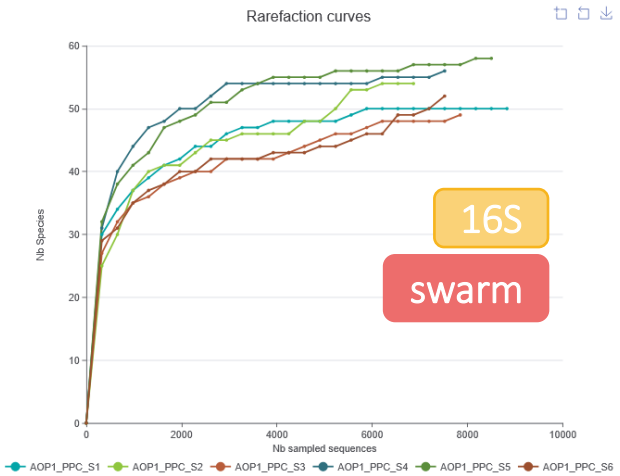
Sample name	Number of Domain	Number of Phylum	Number of Class	Number of Order	Numb
AOP1_PPC_S1	1	4	6	12	22
AOP1_PPC_S2	1	4	5	11	22
AOP1_PPC_S3	1	4	6	12	20
AOP1_PPC_S4	1	4	6	14	27
AOP1_PPC_S5	1	4	6	15	28
AOP1_PPC_S6	1	4	5	13	25
AOP1_PPC_W1	1	4	8	14	25
AOP1_PPC_W2	1	4	6	12	22
AOP1_PPC_W3	1	4	5	11	22
AOP1_PPC_W4	1	4	7	15	26

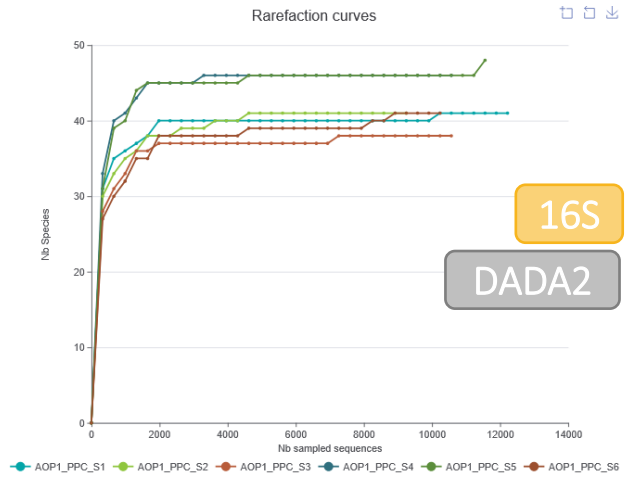
Showing 1 to 10 of 72 rows 10 rows per page ‹ 1 2 3 4 5 ... 8 ›

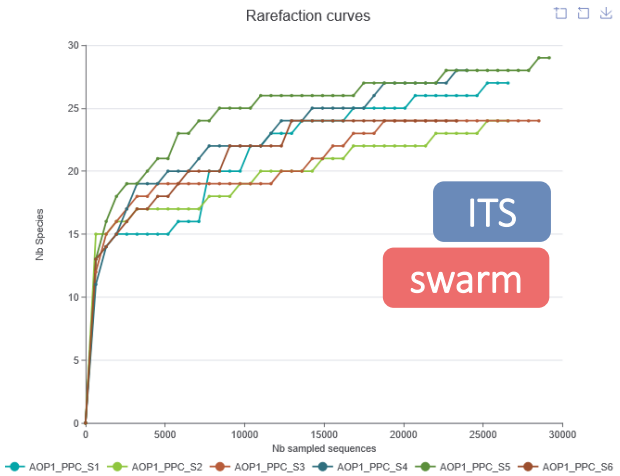
With selection:

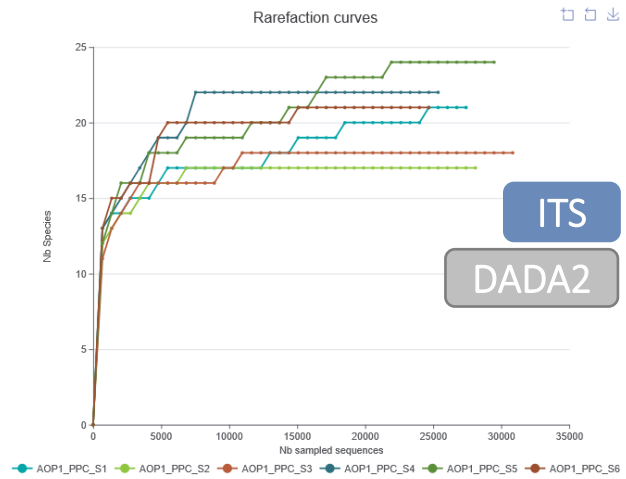
Display distribution

Display rarefaction Species







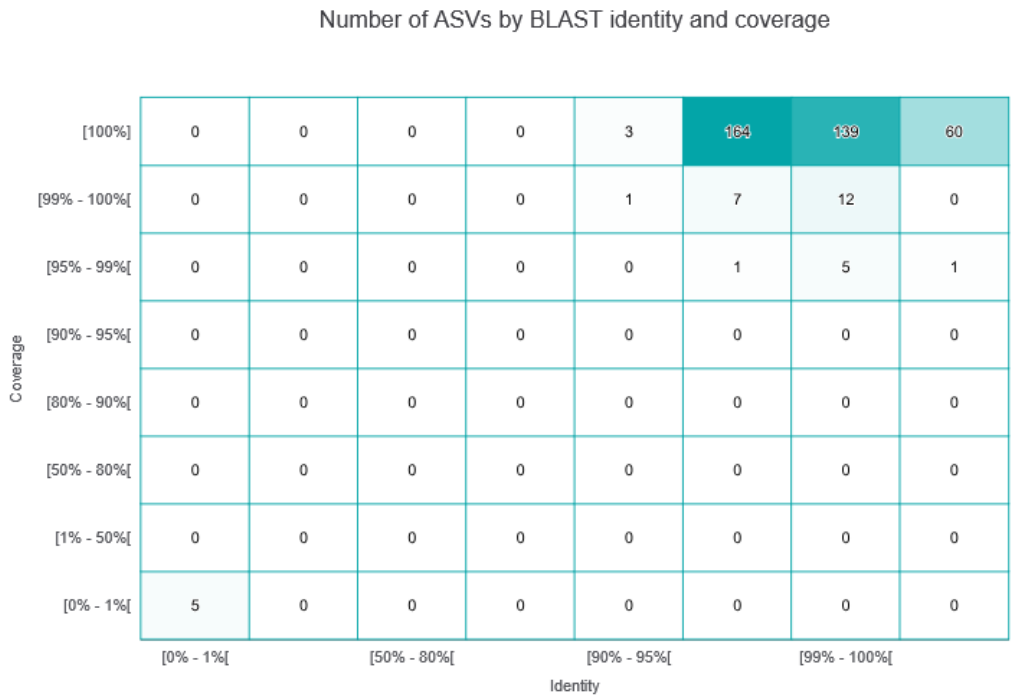


Practice session

Explore Blast alignment metrics Distribution tab.

16S

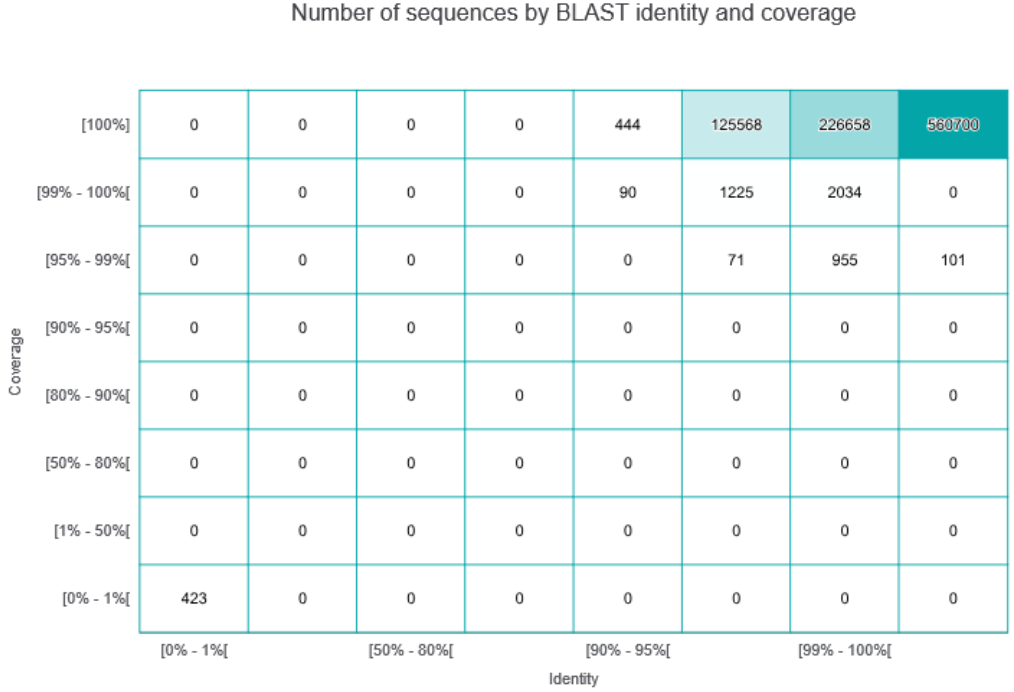
swarm



By ASV | By sequence

16S

swarm



By ASV | By sequence

The results obtained using 16S DADA2 data are equivalent.

We observe that, much of the data can be found in the DAIRYDB database.

Practice session

Explore Blast alignment metrics Distribution tab.

ITS

DADA2

Number of ASVs by BLAST identity and coverage

Coverage \ Identity	[0% - 1%[	[50% - 80%[	[90% - 95%[	[99% - 100%[
[100%]	0	1	4	1	6	35	18	4
[99% - 100%[	0	0	0	0	0	0	0	0
[95% - 99%[	0	0	0	0	2	13	0	0
[90% - 95%[	0	0	0	0	0	19	12	11
[80% - 90%[	0	0	0	0	6	41	25	24
[50% - 80%[	0	0	0	0	0	4	1	0
[1% - 50%[	0	0	0	0	2	2	0	0
[0% - 1%[	0	0	0	0	0	0	0	0

By ASV | By sequence

ITS

DADA2

Number of sequences by BLAST identity and coverage

Coverage \ Identity	[0% - 1%[	[50% - 80%[	[90% - 95%[	[99% - 100%[
[100%]	0	578	1123	141	1064	26617	456696	196834
[99% - 100%[	0	0	0	0	0	0	0	0
[95% - 99%[	0	0	0	0	244	107892	0	0
[90% - 95%[	0	0	0	0	0	9656	23145	582890
[80% - 90%[	0	0	0	0	2833	47129	105204	463385
[50% - 80%[	0	0	0	0	0	3213	124	0
[1% - 50%[	0	0	0	0	2090	432	0	0
[0% - 1%[	0	0	0	0	0	0	0	0

By ASV | By sequence

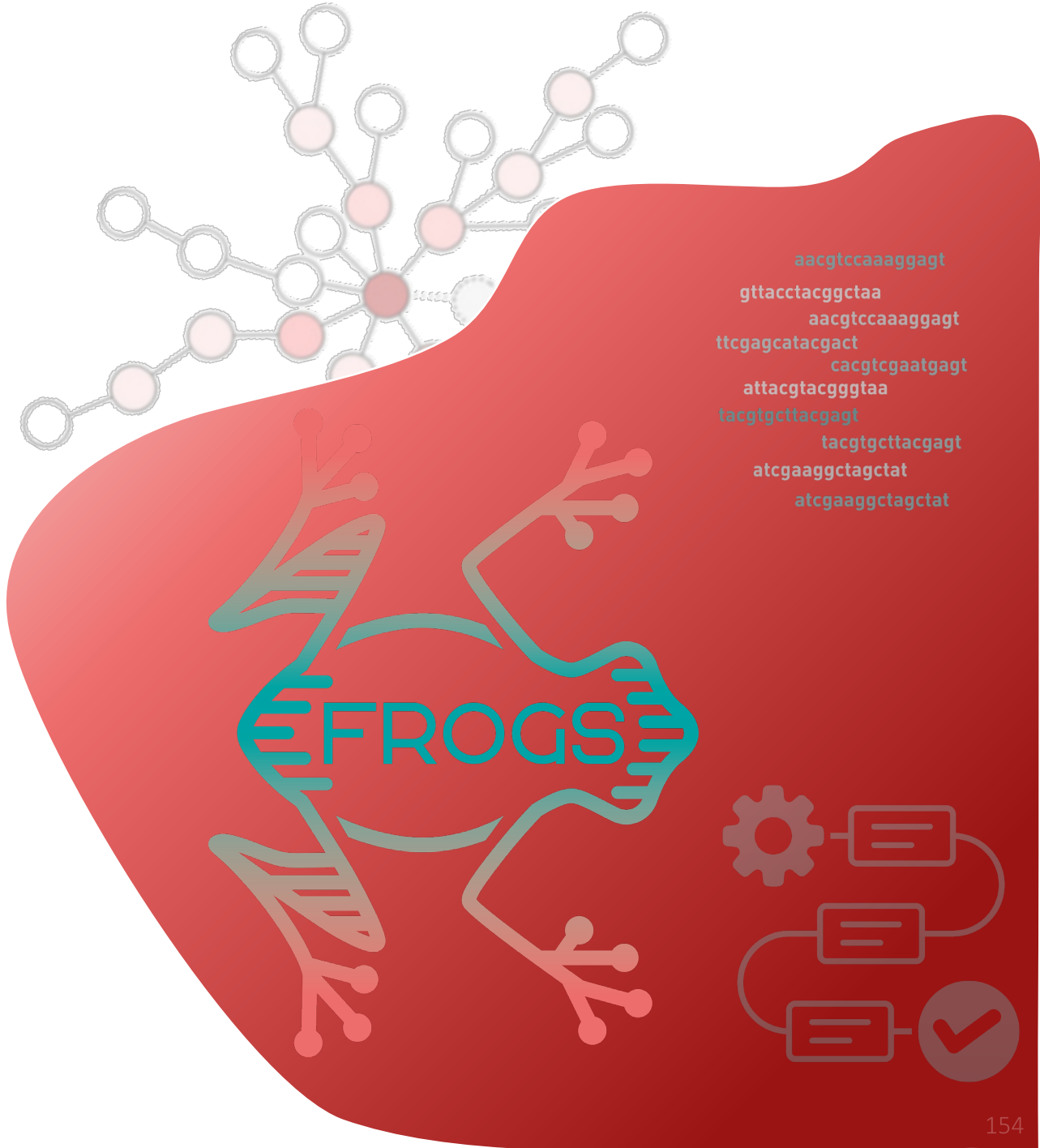
The results obtained using ITS swarm data are equivalent.

We observe that, much of the data cannot be found in the UNITE database.

FROGS Core 1

Main tools

Phylogenetic tree building



What does this tool do?

This tool generates a phylogenetic tree using the sequences of the ASVs contained within the fasta file.

It uses MAFFT for multiple sequence alignment and FastTree for phylogenetic reconstruction.

The BIOM file serves as a metadata resource for annotating each branch of the reconstructed tree with the affiliation of each ASV.

Practice session

Please open the FROGS Core Main 5 tool and familiarize yourself with the required parameters.

Run the process !

Explore the output files.

Practice session

Tool Parameters

Select a tool from the FROGS Core suite to run your analysis.

- -- Please select a tool --
- 1.a. Reads processing of short reads
- 1.b. Reads processing of long reads
- 1.c. Reads processing of 454 reads
- 2. Remove chimera
- 3. Cluster/ASV filters
- 4. Taxonomic affiliation
- (•) 5. Phylogenetic tree building
- ITSx

16S

Sequence file (.fasta) *

12: FROGS Core 1-Main - cluster_filters: sequence.fasta

accepted formats

FASTA file containing the representative sequences to be used for phylogenetic tree construction. (--input-fasta)

Abundance file (.biom) *

39: FROGS Core 1-Main - taxonomic_affiliation: abundance.biom

accepted formats

BIOM file containing the abundance table associated with the input sequences. (--input-biom)

Tool Parameters

Select a tool from the FROGS Core suite to run your analysis.

- -- Please select a tool --
- 1.a. Reads processing of short reads
- 1.b. Reads processing of long reads
- 1.c. Reads processing of 454 reads
- 2. Remove chimera
- 3. Cluster/ASV filters
- 4. Taxonomic affiliation
- (•) 5. Phylogenetic tree building
- ITSx

ITS

Sequence file (.fasta) *

21: FROGS Core 1-Main - itsx: sequence.fasta

accepted formats

FASTA file containing the representative sequences to be used for phylogenetic tree construction. (--input-fasta)

Abundance file (.biom) *

37: FROGS Core 1-Main - taxonomic_affiliation: abundance.biom

accepted formats

BIOM file containing the abundance table associated with the input sequences. (--input-biom)

Outputs :

47: **FROGS Core 1-Main - tree: tree.nwk**

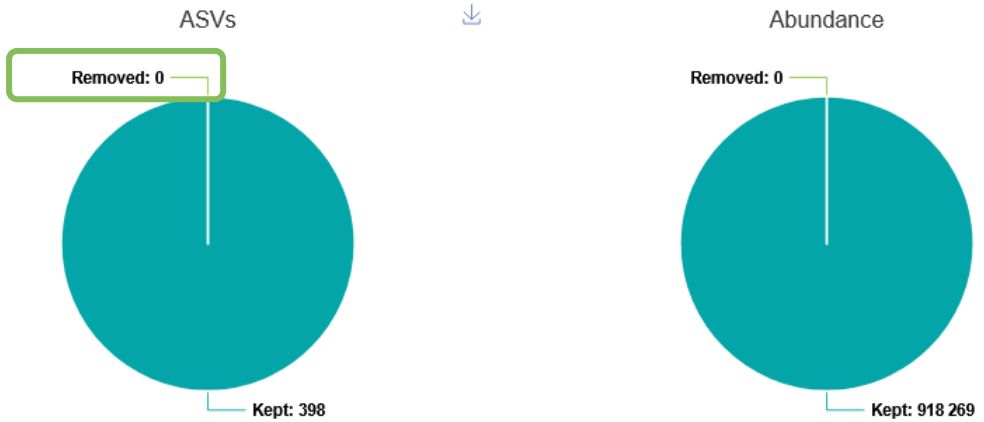
46: **FROGS Core 1-Main - tree: report.html**

Practice session

Explore the report.html file.

16S

ITS



It is important to check that all ASVs have been included in the phylogenetic tree; otherwise, the tree does not represent the abundance file entirely.

You can use the mouse wheel to zoom in on the tree and see it in more detail.

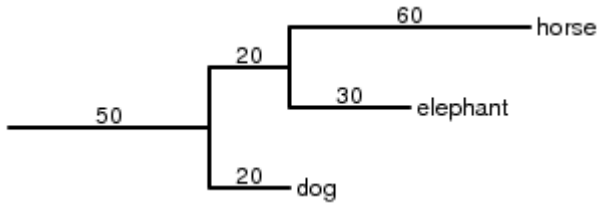
Practice session

Explore the tree.nhx file.

16S

ITS

This is a phylogenetic tree in Newick format, where each node is represented by brackets. This universal format can be used with all tree viewers.

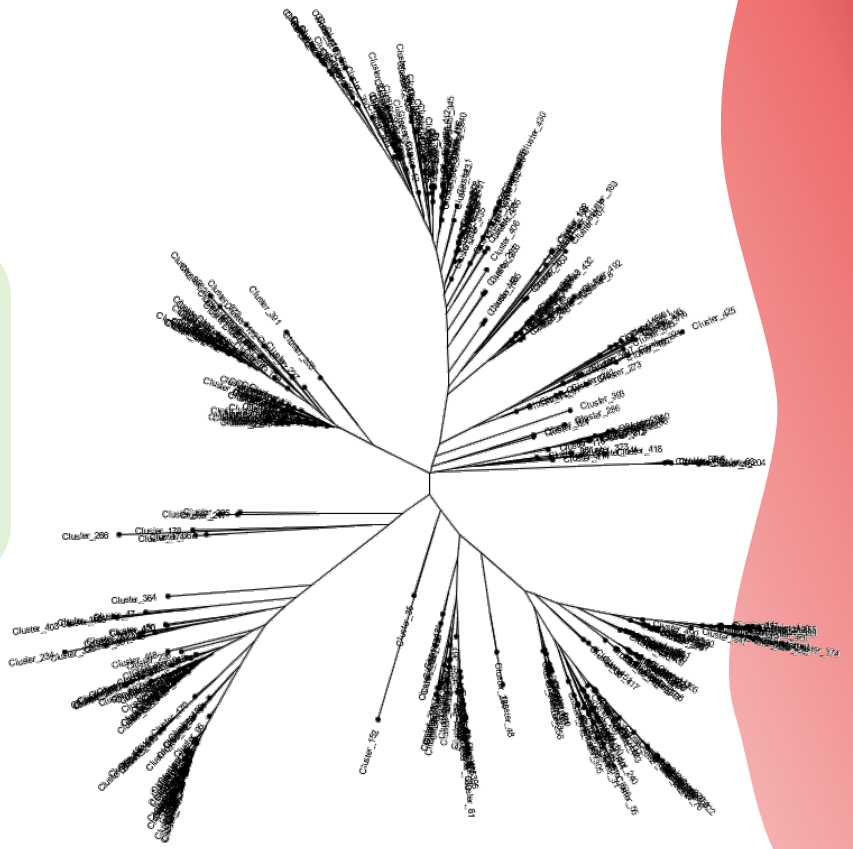


(dog:20, (elephant:30, horse:60):20):50

Our tree is in NHX (= nwk) format.

```
(((((ID_299:0.351858451,(((ID_48:0.002219021,(ID_37:5e-09,ID_
(ID_82:5e-09,ID_264:0.004941303)1.000:0.060755678)0.749:0.011
(ID_332:0.004694167,ID_307:5e-09)0.827:0.002338801)0.978:0.07
(ID_12:5e-09,ID_20:0.00912482)0.998:0.02188453)1.000:0.369577
((ID_58:0.002592757,ID_300:0.011147743)0.997:0.061295548,
(ID_125:0.029079135,ID_225:0.045514945)0.936:0.034654793)0.78
(ID_159:0.070427204,((ID_236:5e-09,(ID_239:0.002226577,(ID_17
(ID_170:0.004760105,ID_153:0.019412678)0.000:5e-09)0.766:0.00
((ID_150:0.02177976,ID_297:0.025118579)0.718:0.003146269,(ID_
(ID_97:5e-09,ID_109:0.005402298)0.922:0.005433275)0.000:5e-09
(((ID_30:5e-09,
((ID_46:5e-09,ID_193:0.016595317)0.926:0.013849197,ID_108:0.0
(((ID_301:0.002683217,ID_215:0.005377519)0.000:5e-09,ID_158:5
:5e-09,ID_111:0.002698725)0.849:0.002754821,(ID_116:5e-09,
```

This is an example of a visualization created using FigTree software from an NHX file.



https://github.com/rambaut/figtree/releases