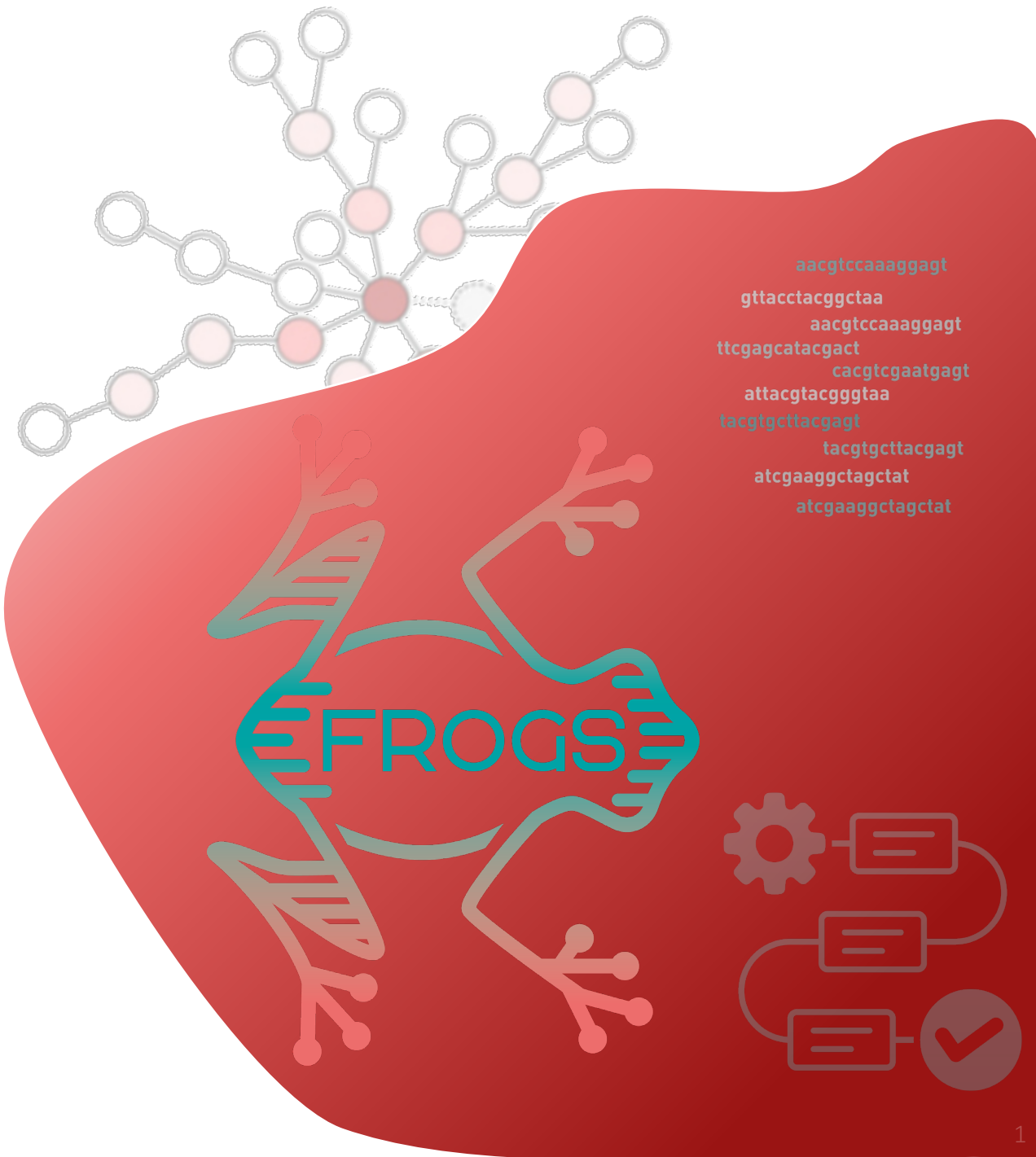


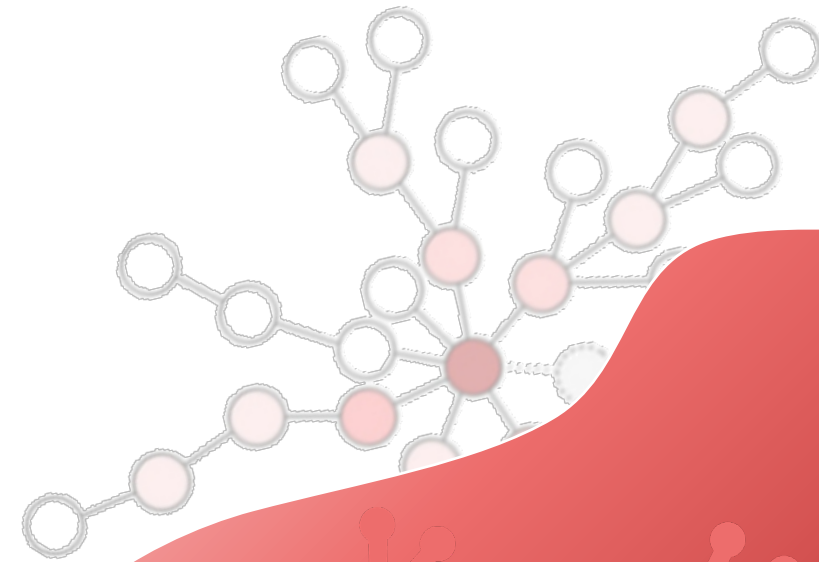
FROGS Core

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
ttcgagcatagact
cacgtcgaatgagt
attacgtacgggtaa
tacgtgcttacgagt
tacgtgcttacgagt
atcgaaggctagctat
atcgaaggctagctat



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>

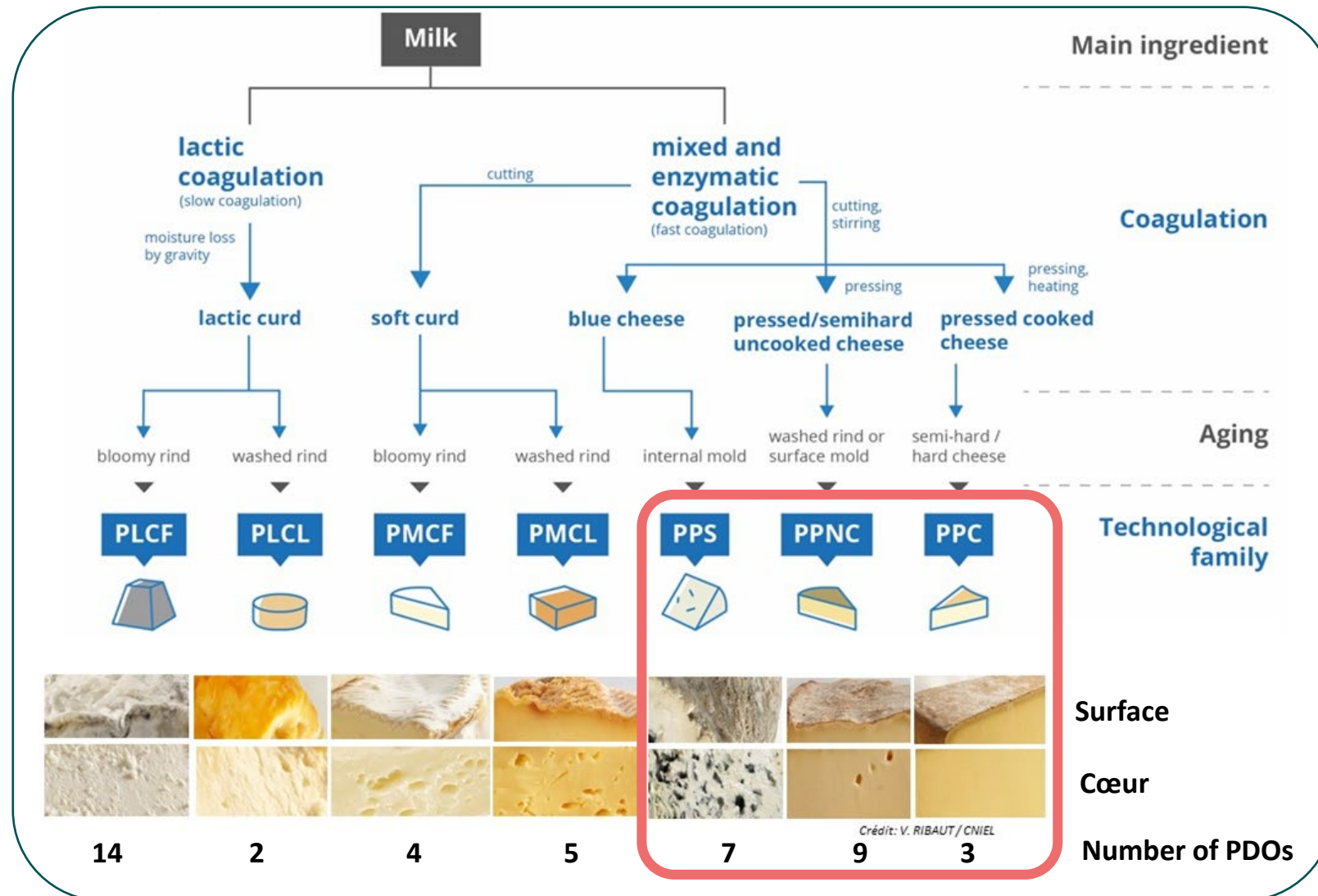
Irlinger, Françoise; Mariadassou, Mahendra; Dugat-Bony, Eric; Rué, Olivier; Neuvéglise, Cécile; Renault, Pierre; Rifa, Etienne; Theil, Sébastien; Loux, Valentin; Cruaud, Corinne; Gavory, Frederick; Barbe, Valérie; Lasbleiz, Ronan; Gaucheron, Frédéric; Spelle, Céline; Delbès, Céline, 2024, "metapdocheese_Fungi_sample_metadata.tab", 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]

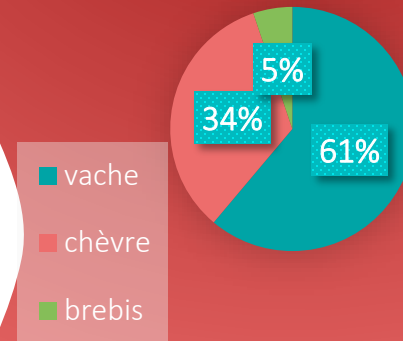
Data:

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

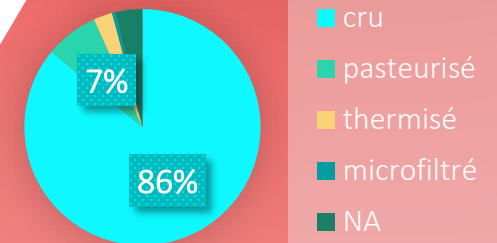
44 PDOs divided into 7 technological families:



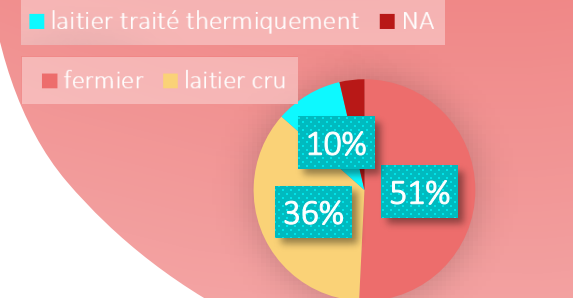
Dairy species



Milk processing



Type of production



From F. Iringer communication

Adapted from Monserrat & Mietton, 2014

72 samples from MetaPDOCheese

- 2 PDOs (34 & 35) → AOP1 & 2
- 2 seasons (winter & summer)
- 2 production units/PDO
- 3 replicats/samples

- Surface only
- ITS & 16S
- Cow milk

PPC
Pâte pressée cuite



- 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

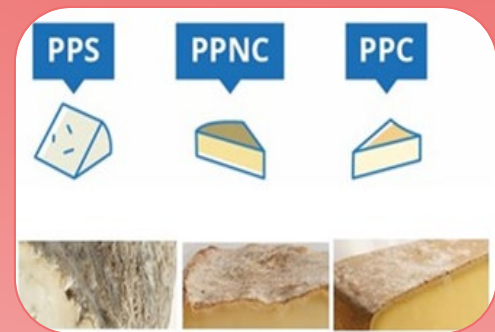
PPNC
Pâte pressée non
cuite



- 2 PDOs (25 & 3) → AOP5 & 6
- 2 seasons (winter & summer)
- 2 production units/PDO
- 3 replicats/samples

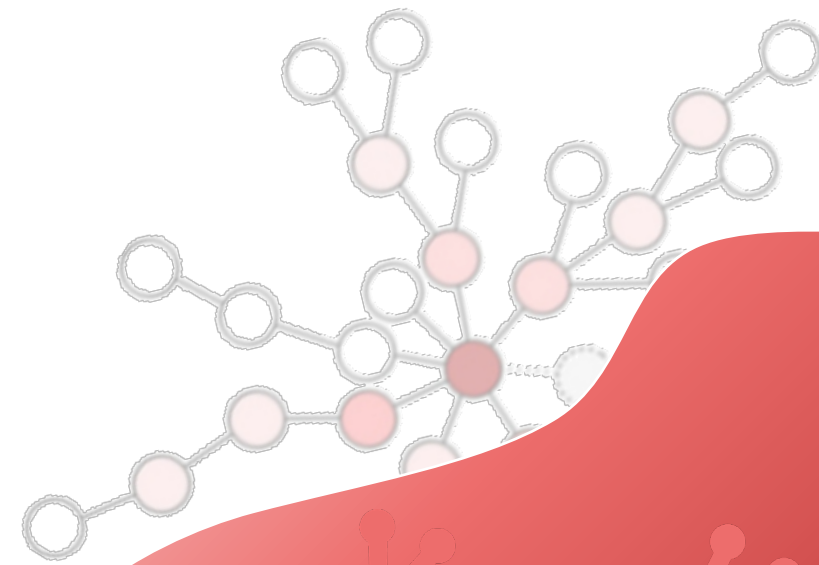
- Surface only
- ITS & 16S
- Cow milk

PPS
Pâte persillée



FROGS Core

Overview



aacgtccaaaggagt
gttacctacggctaa
aacgtccaaaggagt
ttcgagcatagact
cacgtcgaatgagt
attacgtacggtaa
tacgtgcttacgagt
tacgtgcttacgagt
atcgaaggctagctat
atcgaaggctagctat



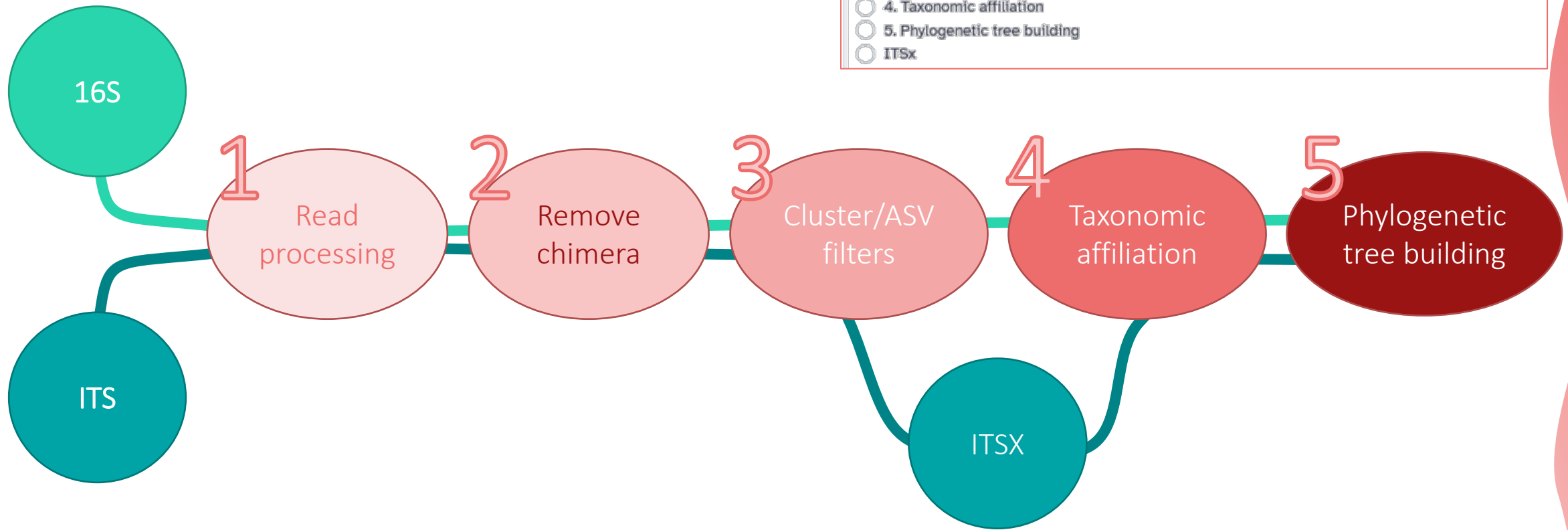
FROGS Core 1 Main tools

 **FROGS Core 1-Main** ASV reconstruction and taxonomic affiliation (Galaxy Version 5.1.0+galaxy0)

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



FROGS Core 2 Companion tools

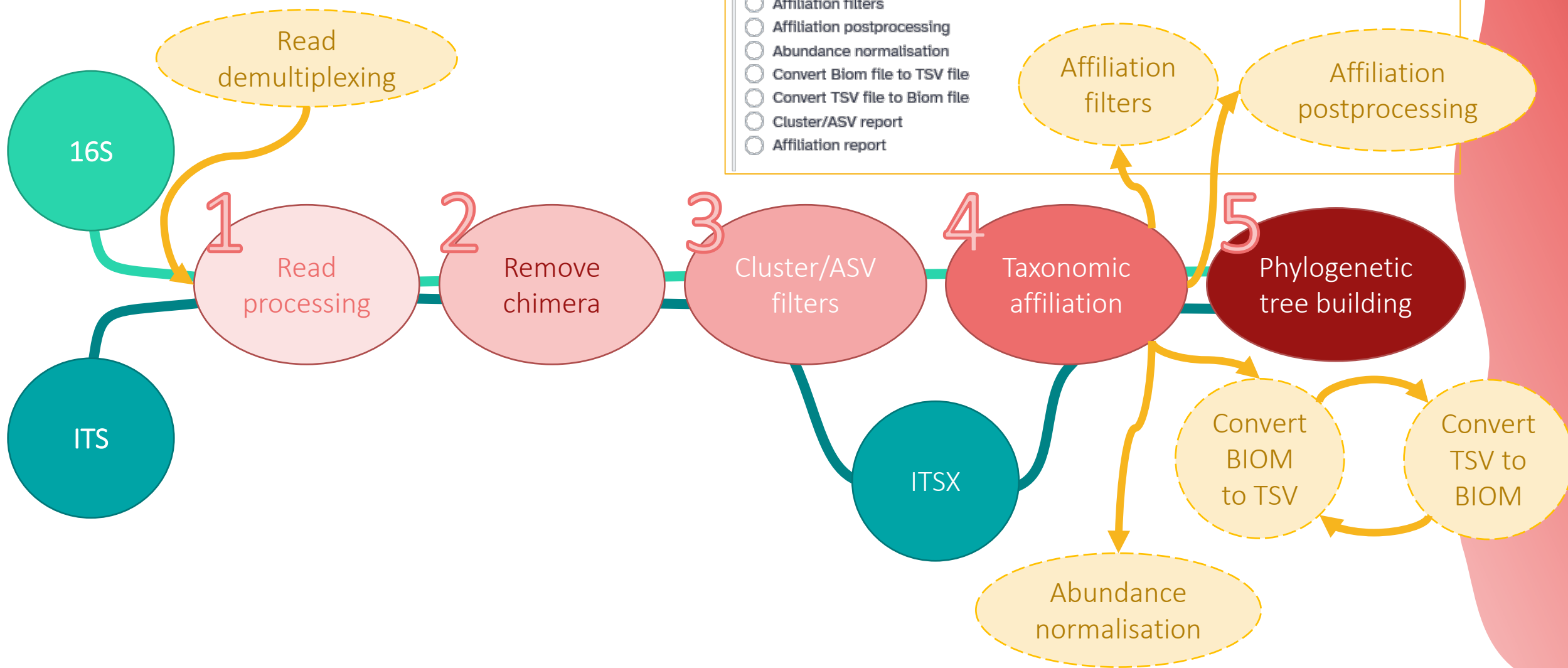
FROGS Core 2-Companion Optional process, converter, and report (Galaxy Version 5.1.0+galaxy0)

Tool Parameters

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

☒ — Please select a tool —

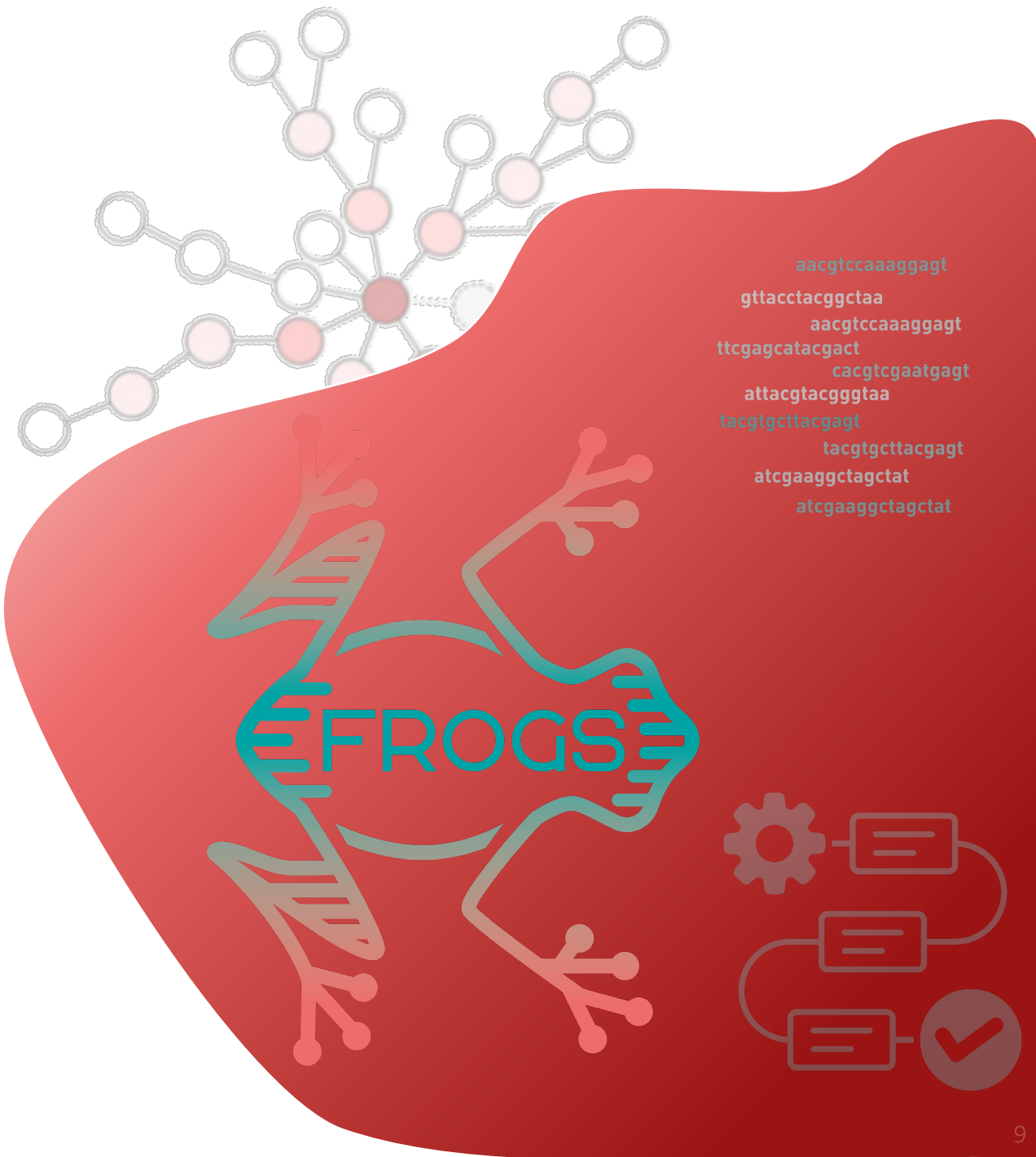
- ☐ Read demultiplexing
- ☐ Affiliation filters
- ☐ Affiliation postprocessing
- ☐ Abundance normalisation
- ☐ Convert Biom file to TSV file
- ☐ Convert TSV file to Biom file
- ☐ Cluster/ASV report
- ☐ Affiliation report



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?

What marker is the target ?

What is the length of target ?

What length of sequencing reads ?

Are the reads paired or not?

What sequencing technology was used ?

What PCR primers were used ?

What is the quality of sequences ?

What is sequencing deep ?

Features of 16S reads

16S

Read lengths: 250

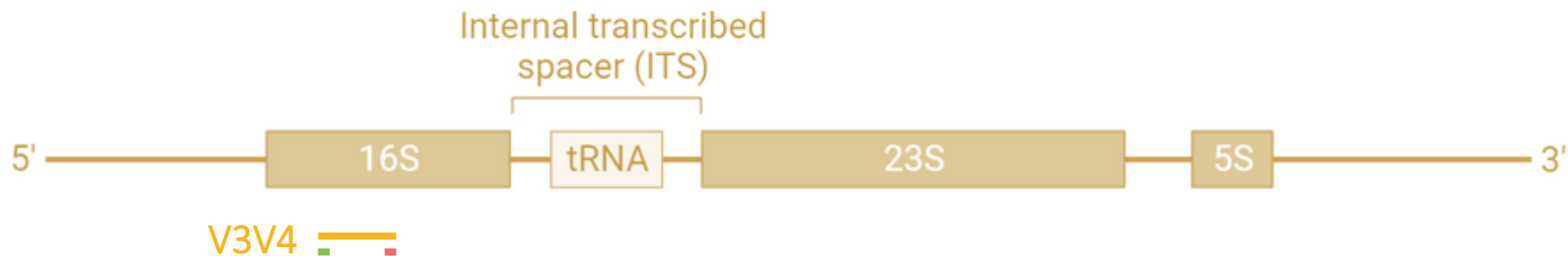
Minimum amplicon length: 250

Maximum amplicon length: 480

5' PCR primer: 5' ACGGRAGGCWGCAG 3'

3' PCR primer: 3' TACCAGGGTATCTAATCCT 5'

Bacteria rRNA gene organization



Features of ITS reads

ITS

Read lengths: 250

Minimum amplicon length: 150

Maximum amplicon length: 480

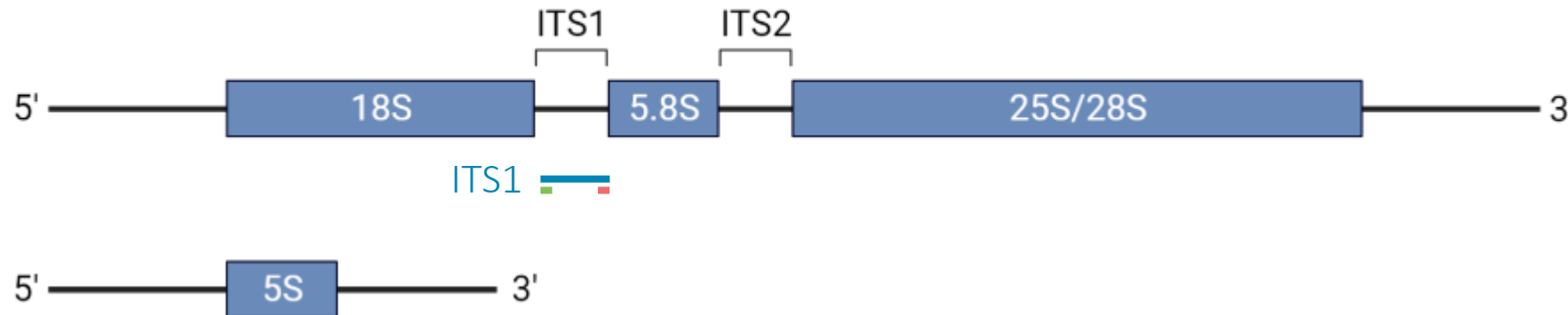
5' PCR primer: 5' GCATCGATGAAGAACGCAGC 3'

3' PCR primer: 3' TCCTCCGCTTWTGWTWTGC 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 each
Aviti and IonTorrent. (--input-archive)

R1 read length *

250

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

R2 read length *

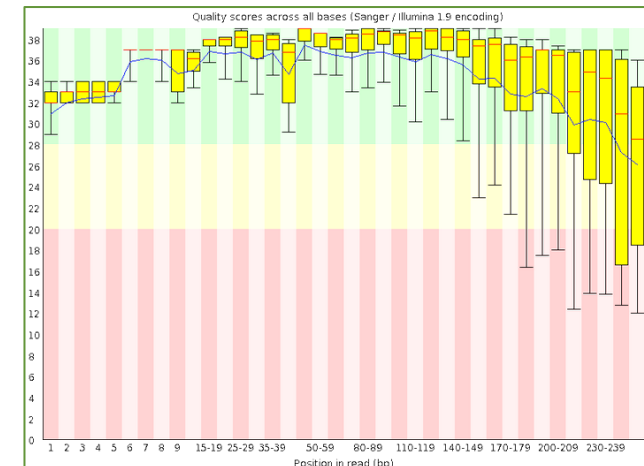
250

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

Mismatch rate (used for R1-R2 merging) *

0,1

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



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

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 each
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

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

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 each 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)

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 the R2 primer

Use revseq: <https://www.bioinformatics.nl/cgi-bin/emboss/revseq>



5' PCR primer: 5' ACGGRAGGCWGCAG 3'

3' PCR primer: 5' AGGATTAGATACCCTGGTA 3'

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

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 merged

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 Illumina and IonTorrent. (--input-archive)

R1 read length *

250

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

R2 read length *

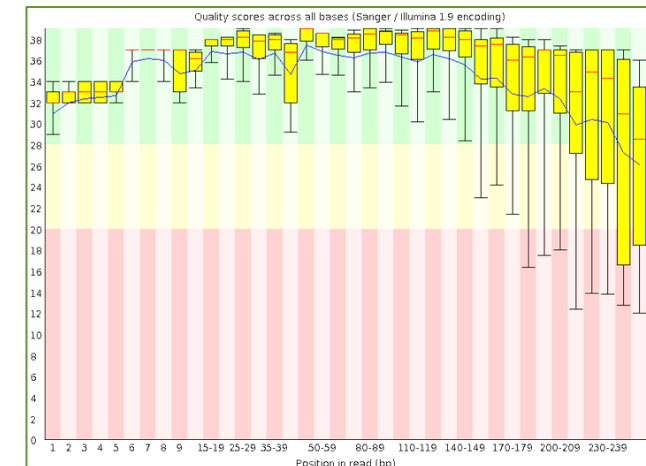
250

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

Mismatch rate (used for R1-R2 merging) *

0,1

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



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

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 merged

Input

TAR Archive

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

Archive file (.tar.gz) *

1: MPC_ITS.tar.gz

accepted formats ▼

The TAR file containing the short R1 R2 read pairs (.fastq.gz) for Illumina, PacBio, 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)

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

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, subtract 10 bases to account for the minimum overlap between R1 and R2 reads. (--min-amplicon-size)

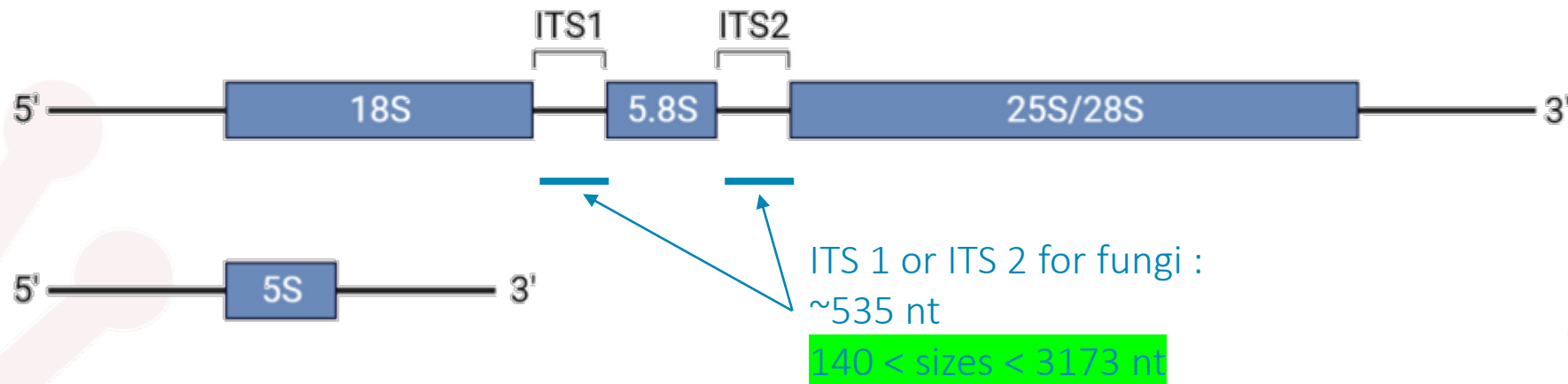
Maximum amplicon length *

480

The maximum length of the amplicons (including primers). For paired-end reads, subtract 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

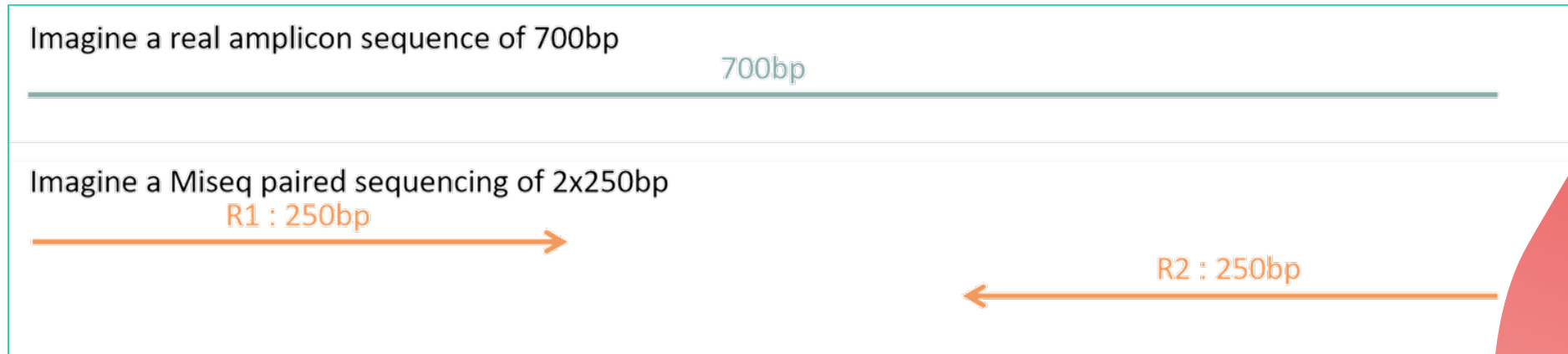
Eukaryotes rRNA gene organization



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

Imagine a real amplicon sequence of 700bp

700bp

Imagine a Miseq paired sequencing of 2x250bp

R1 : 250bp

R2 : 250bp

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

NNNNNNNNNNNNNNNNNNNN

Combined sequence length : 600bp, with 100 Ns

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 merged

Input

TAR Archive

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

Archive file (.tar.gz) *



1: MPC_ITS.tar.gz

accepted formats ▼

The TAR file containing the short R1 R2 read pairs (.fastq.gz) for Illumina, PacBio, Nanopore, MinION, 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, subtract 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, subtract 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 allowed.

3' primer - optional

GCAWAWCAAWAAGCGGAGGA

Enter the 3' primer sequence. Wildcards are allowed.

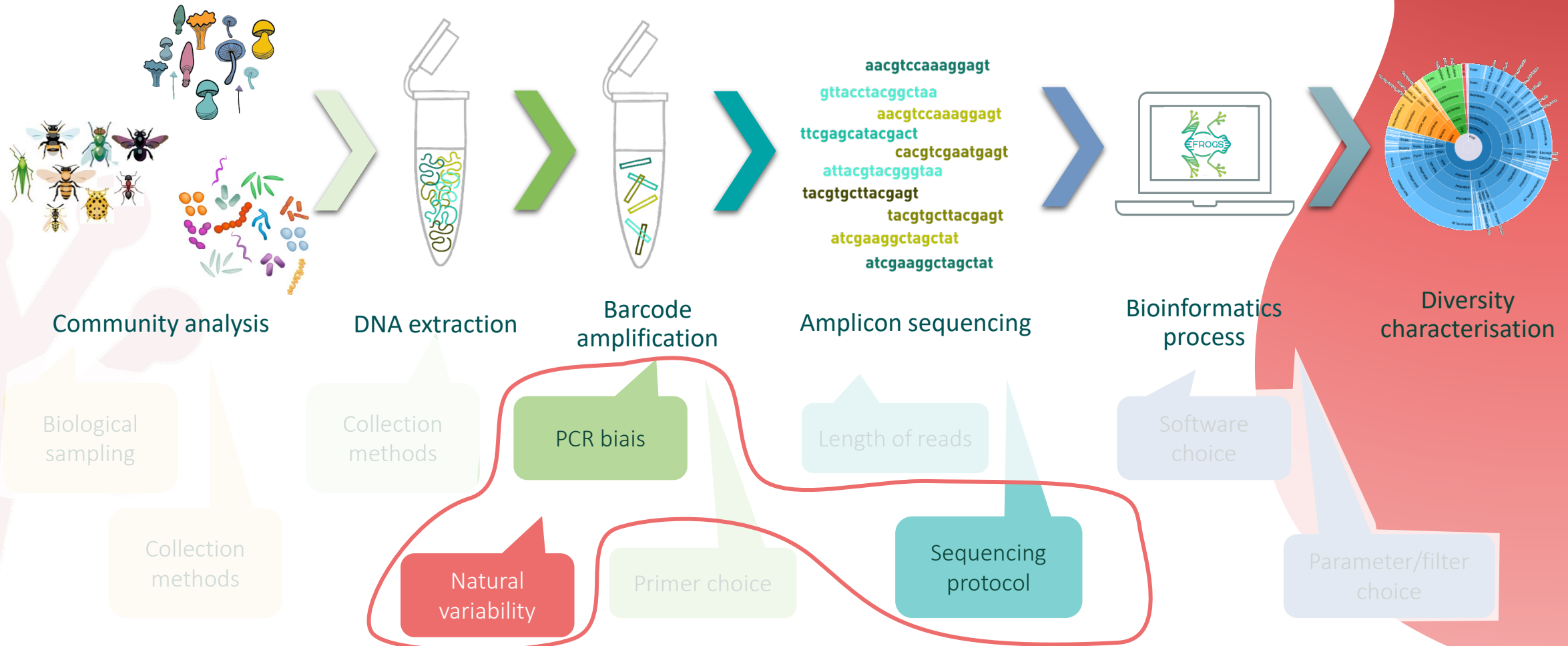
You have to enter the R2 primer in 5'-3'

If necessary, 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?





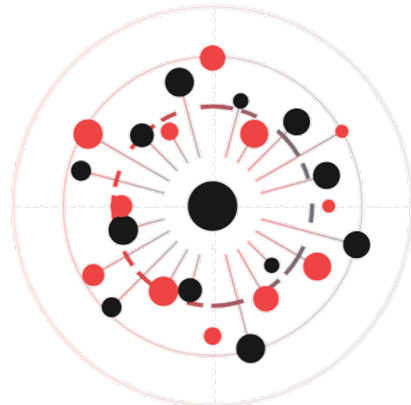
Which methods for grouping sequences are available in the settings panel?

2 methods: Swarm or DADA2

Process type

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

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



SWARM

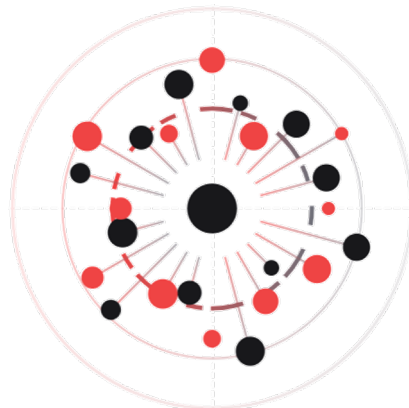


2 methods: Swarm or DADA2

Process type

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

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



SWARM

Skip the detailed
explanations

Swarm clustering method

	ACGT	ACGT	ACGT
	AGGT	A - GT	A - - T
differences	1	1	2

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.

initial seed (randomly picked from amplicon dataset)

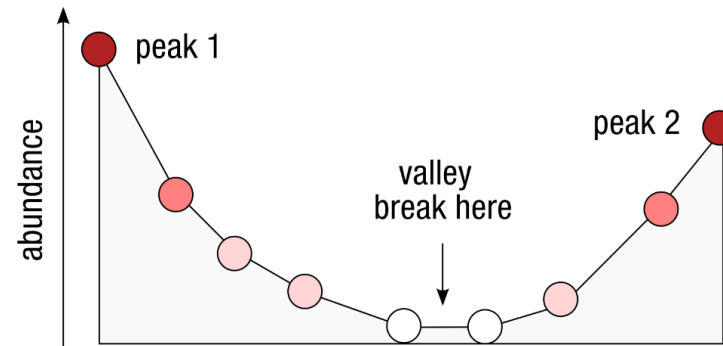
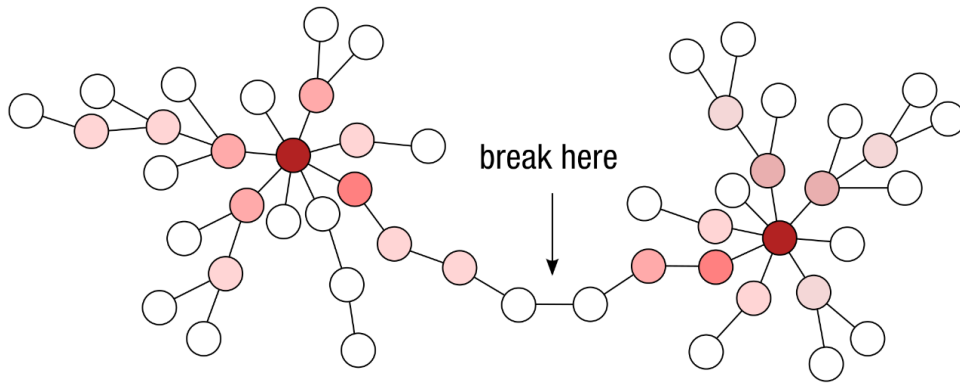
explore the amplicon space

no more closely related amplicons, the process stops (equivalent to the Kruskal algorithm when $d = 1$)



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 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 *



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

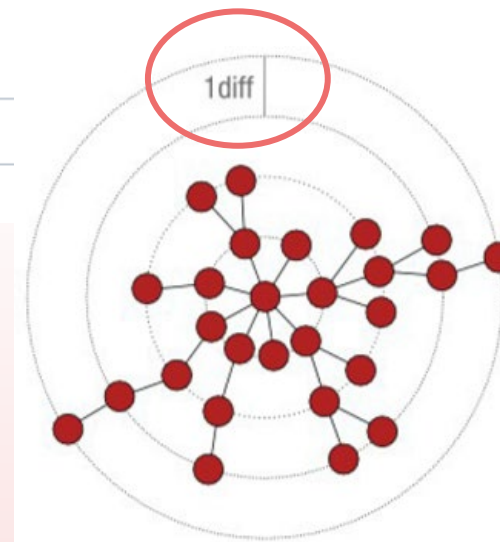
- ☐ Preprocessing only
- ☒ Preprocessing and clustering with Swarm
- ☐ Preprocessing and denoising with DADA2

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

Swarm distance threshold *

1

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



Process type

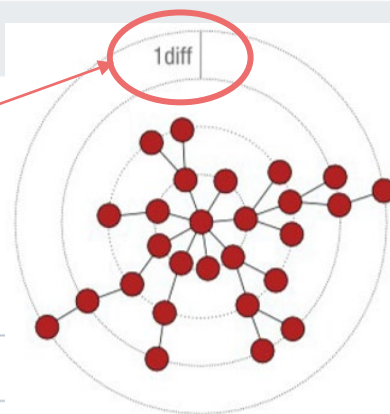
- ☐ Preprocessing only
- ☒ Preprocessing and clustering with Swarm
- ☐ Preprocessing and denoising with DADA2

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

Swarm distance threshold *

1

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

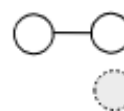


Clustering refinement

- ☒ 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.

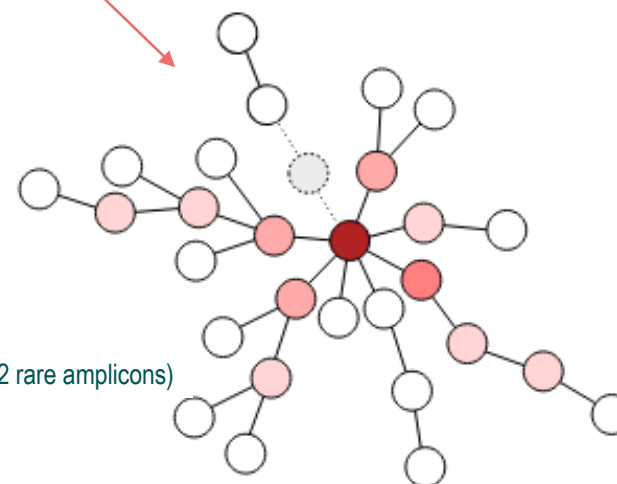
longer but more accurate



small cluster (made of 2 rare amplicons)



virtual amplicon



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 *

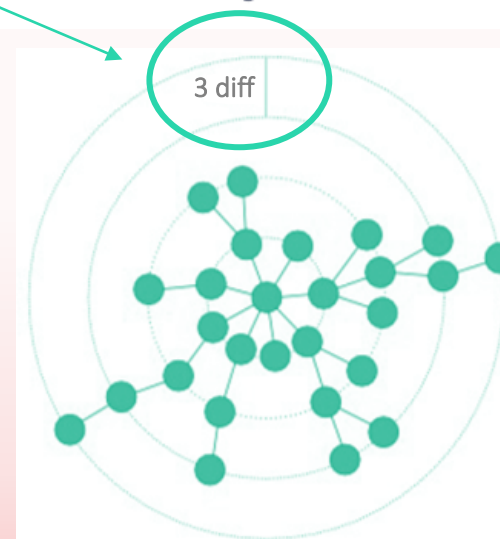
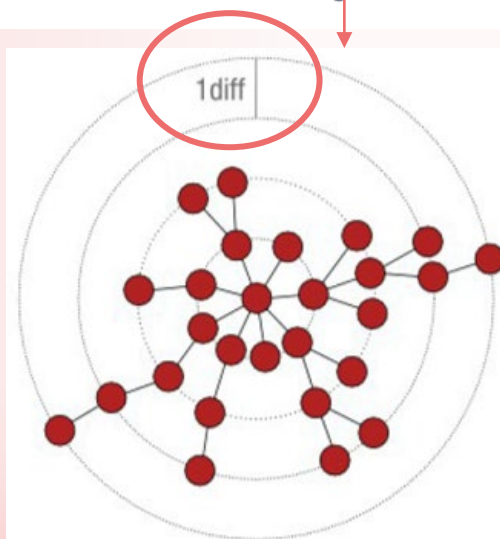
3

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.



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

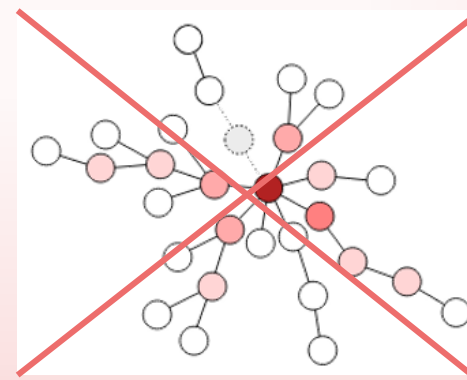
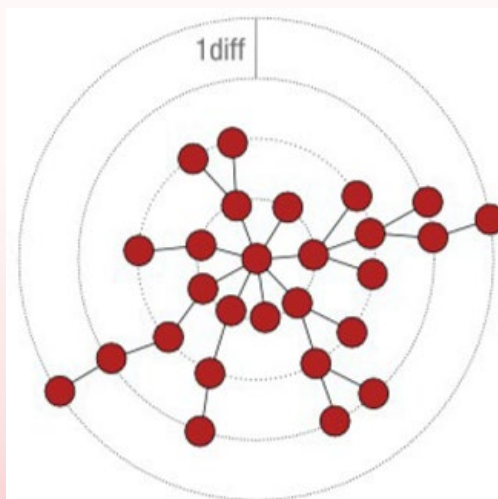
Distance threshold used by Swarm for clustering. (--distance)

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 clustering method: in brief

	ACGT	ACGT	ACGT
	AGGT	A - GT	A - - T
differences	1	1	2

initial seed (randomly picked from amplicon dataset)

explore the amplicon space

no more closely related amplicons, the process stops (equivalent to the Kruskal algorithm when $d = 1$)

This sequence 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.



in brief

Process type

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

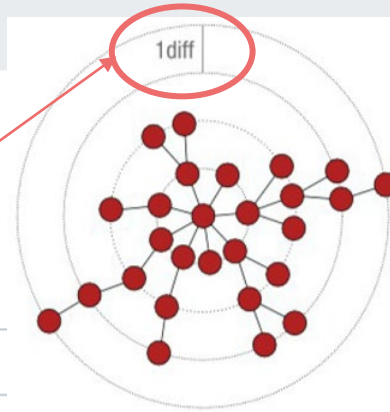
Select the type of process to run (--process)

Swarm distance threshold *

1

Recommended

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



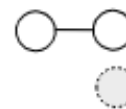
Clustering refinement

- ☒ 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

(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.

Recommended

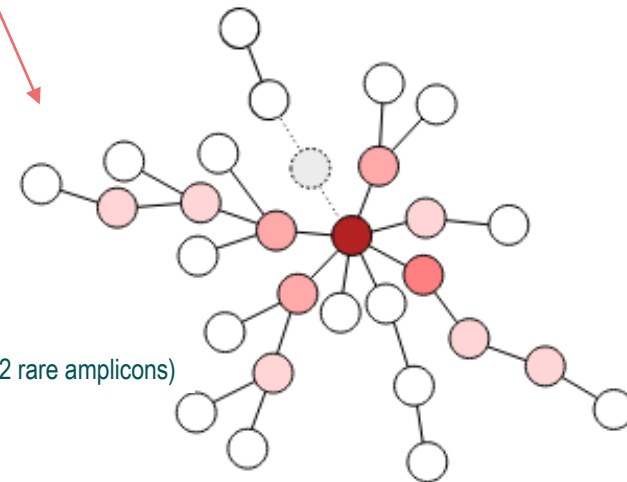
longer but more accurate



small cluster (made of 2 rare amplicons)



virtual amplicon



in brief



U
S
a
g
e

2 methods: Swarm or DADA2

Process type

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

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



DADA2 denoising

- Instead of clustering similar sequences, DADA2 uses a sequencing error model to distinguish true biological sequences from sequences generated by sequencing errors
 - It uses the quality score of each base to estimate the probability that an observed sequence is an erroneous version of a more abundant sequence.
- The result is a set of ASVs (Amplicon Sequence Variants).

Remark:

DADA2 is a complete workflow for analysing amplicon sequencing data. But, in FROGS, we do not use the complete DADA2 workflow. We use only the DADA2 module that generates ASVs.

The complete DADA2 pipeline is not used because, in our benchmarks, the FROGS workflow performs better for our use cases.



Skip the detailed
explanations

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.

$$\lambda_{ji} = \prod_{l=0}^L \underbrace{p(j(l) \rightarrow i(l))}_{\text{Probability of transition between } i \text{ and } j \text{ bases at this position}}, \overbrace{q_i(l)}^{\text{According to quality score}}$$

For each base (from 1 to length L)

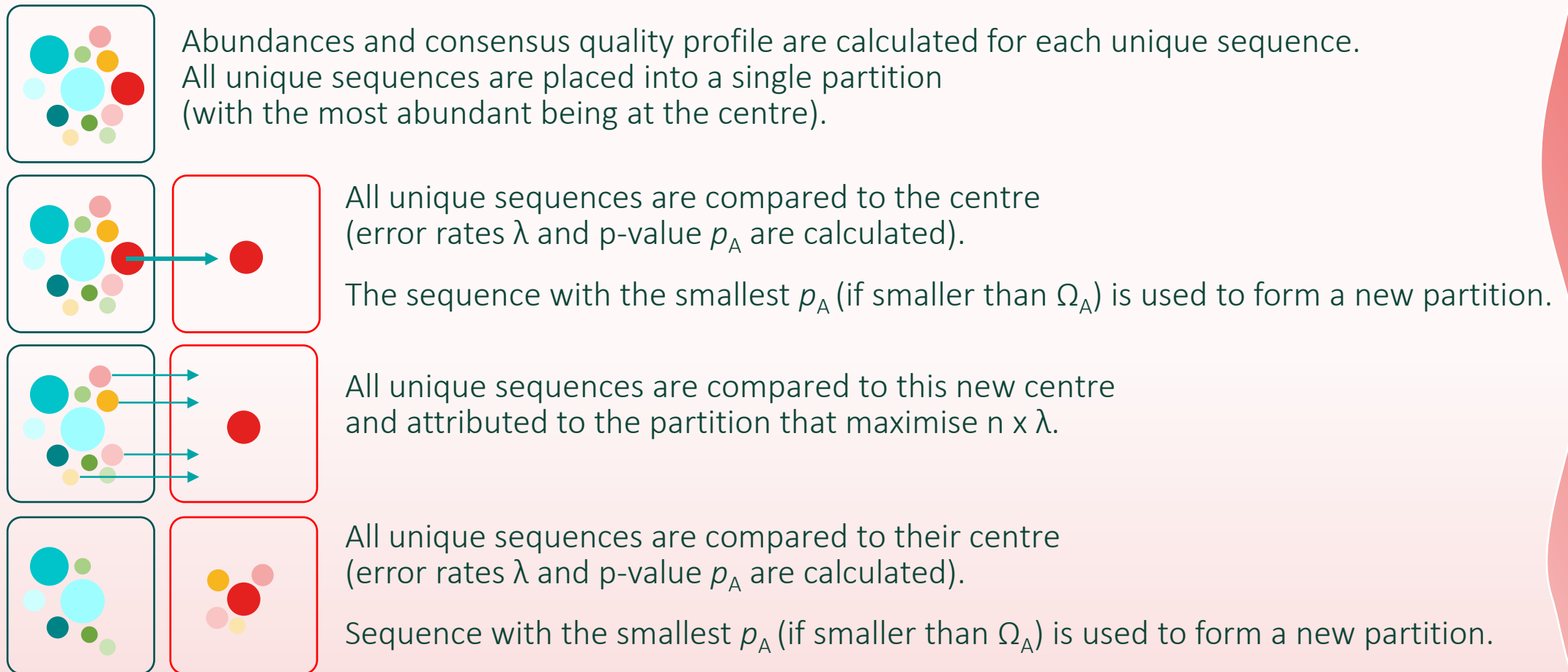
The p-value that the abundance of sequence i is too abundant to be obtained by sequencing errors of j (of abundance n_j).

$$p_A(j \rightarrow i) = \frac{1}{1 - \rho_{\text{pois}}(n_j \lambda_{ji}, 0)} \sum_{a=a_i}^{\infty} \rho_{\text{pois}}(\underbrace{n_j \lambda_{ji}}_{\text{Parent sequence abundance x error rate}}, a)$$

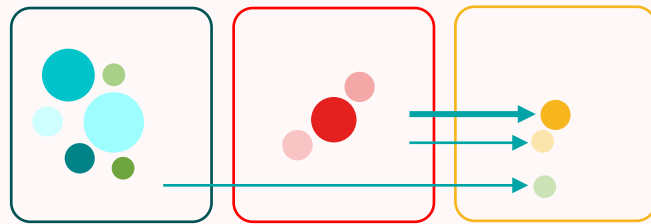
For singletons reads $p_A = 1$

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 .

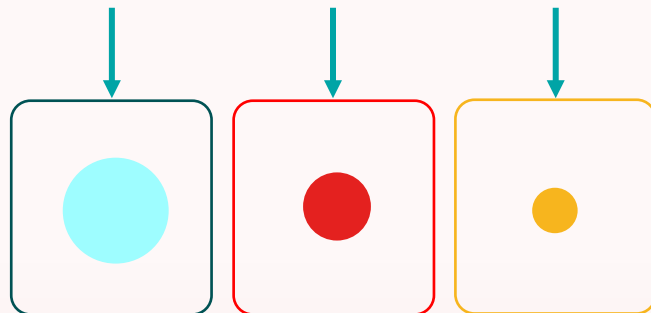
DADA2 denoising - Partitioning



DADA2 denoising - Partitioning



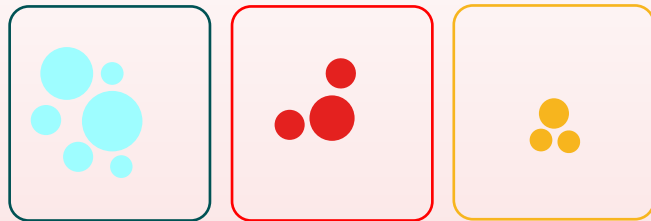
For singleton reads, $p_A = 1$ so they can never form a new partition and are attributed to the partition maximising $n \times \lambda$ (the least unlikely).
Therefore, there is no partition with a unique sequence of abundance 1.



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

Process type

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

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

DADA2 pooling method *

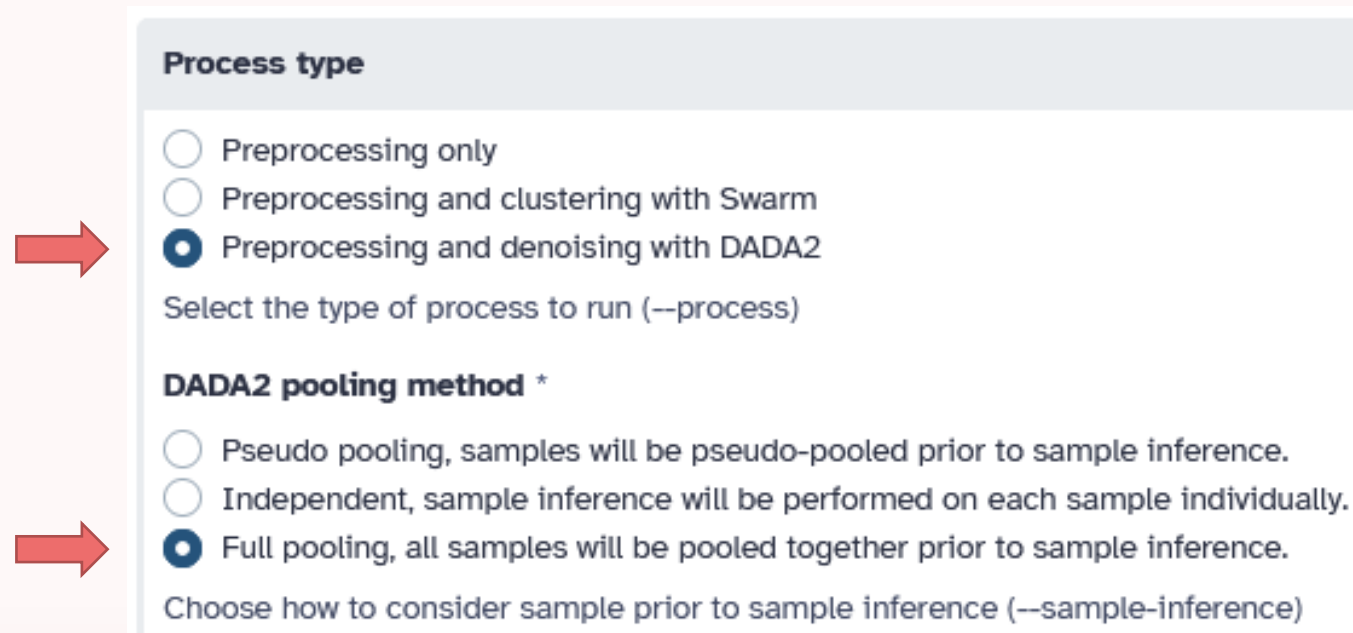
- ☐ Pseudo pooling, samples will be pseudo-pooled prior to sample inference.
- ☒ Independent, sample inference will be performed on each sample individually.
- ☐ Full pooling, all samples will be pooled together prior to sample inference.

Choose how to consider sample prior to sample inference (`--sample-inference`)

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



Process type

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

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

DADA2 pooling method *

- ☐ Pseudo pooling, samples will be pseudo-pooled prior to sample inference.
- ☐ Independent, sample inference will be performed on each sample individually.
- ☒ Full pooling, all samples will be pooled together prior to sample inference.

Choose how to consider sample prior to sample inference (`--sample-inference`)

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

Process type

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

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

DADA2 pooling method *

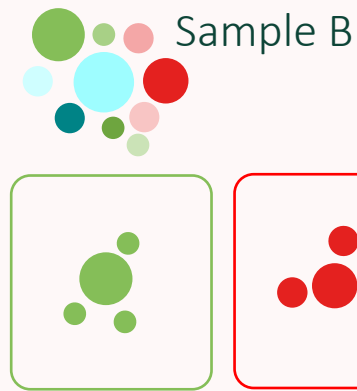
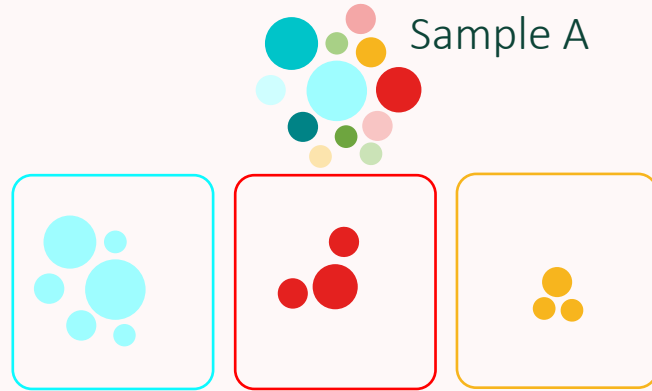
- ☒ Pseudo pooling samples will be pseudo-pooled prior to sample inference.
- ☐ Independent, sample inference will be performed on each sample individually.
- ☐ Full pooling, all samples will be pooled together prior to sample inference.

Choose how to consider sample prior to sample inference (--sample-inference)

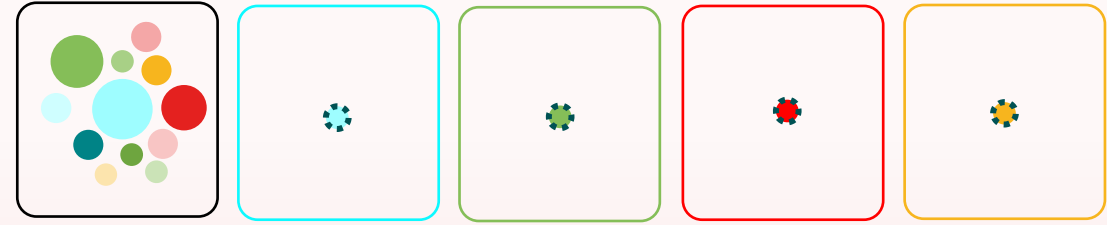
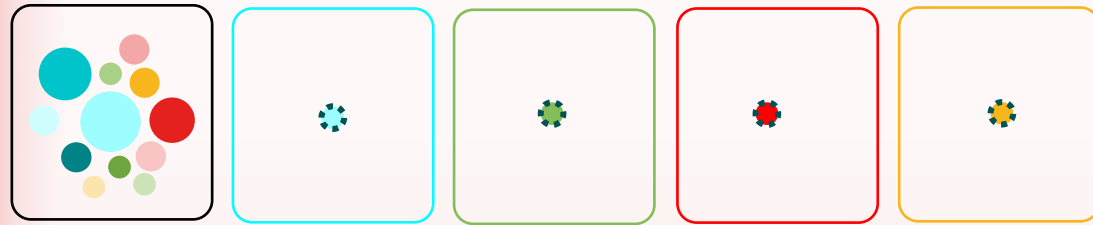
DADA2 offers an alternative 'pseudo-pooling' approach, in which samples are first treated independently before a second treatment using all ASVs identified in the first round as 'priors' (the initial centres of partitions with an abundance of 1).

DADA2 pseudo-pooling

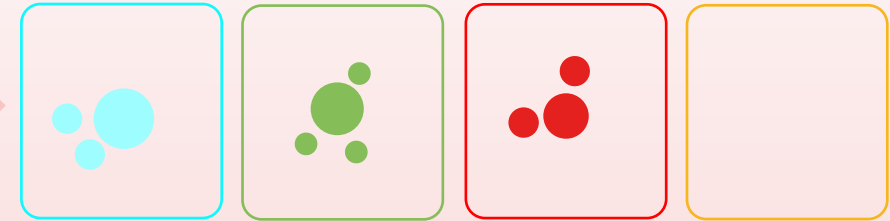
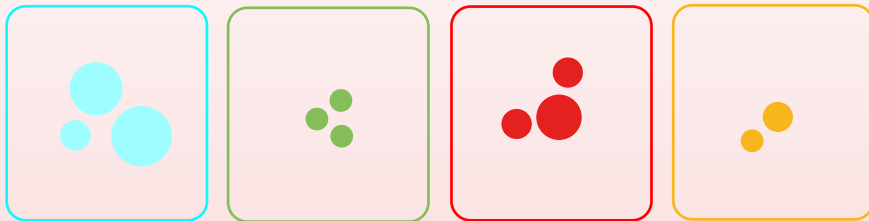
1st round



2nd round



All the unique sequences are compared with the centre of each partition and assigned to the partition that maximises $n \times \lambda$.



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

DADA2 options



Process type

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

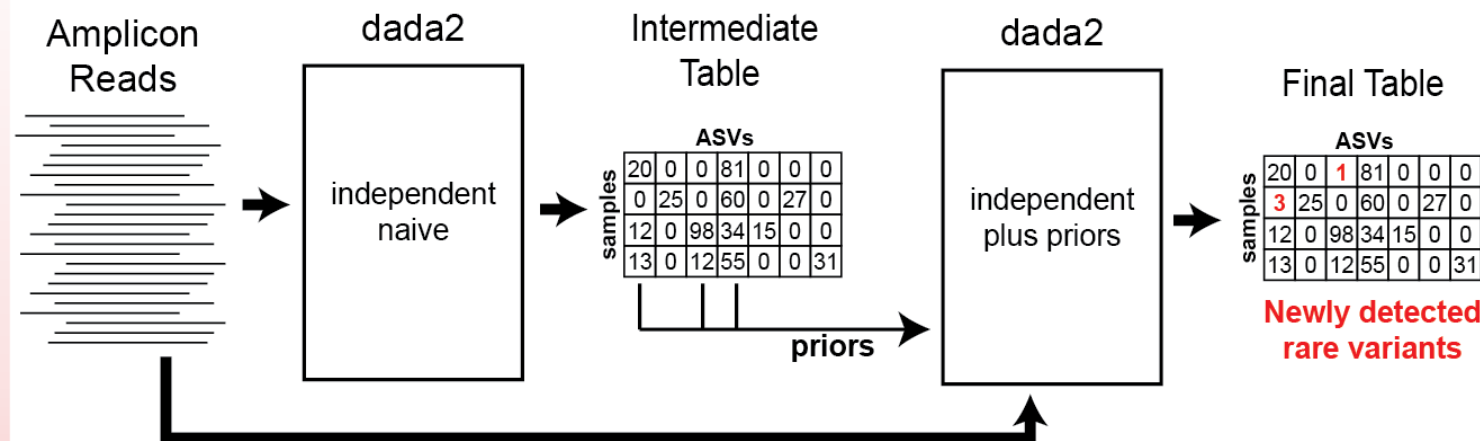
Select the type of process to run (--process)

DADA2 pooling method *

- ☒ Pseudo pooling, samples will be pseudo-pooled prior to sample inference.
- ☐ Independent, sample inference will be performed on each sample individually.
- ☐ Full pooling, all samples will be pooled together prior to sample inference.

Choose how to consider sample prior to sample inference (--sample-inference)

Pseudo-Pooling



DADA2 recommended usage: in brief

Process type

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

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

DADA2 pooling method *

- ☒ Pseudo pooling, samples will be pseudo-pooled prior to sample inference.
- ☐ Independent, sample inference will be performed on each sample individually.
- ☐ Full pooling, all samples will be pooled together prior to sample inference.

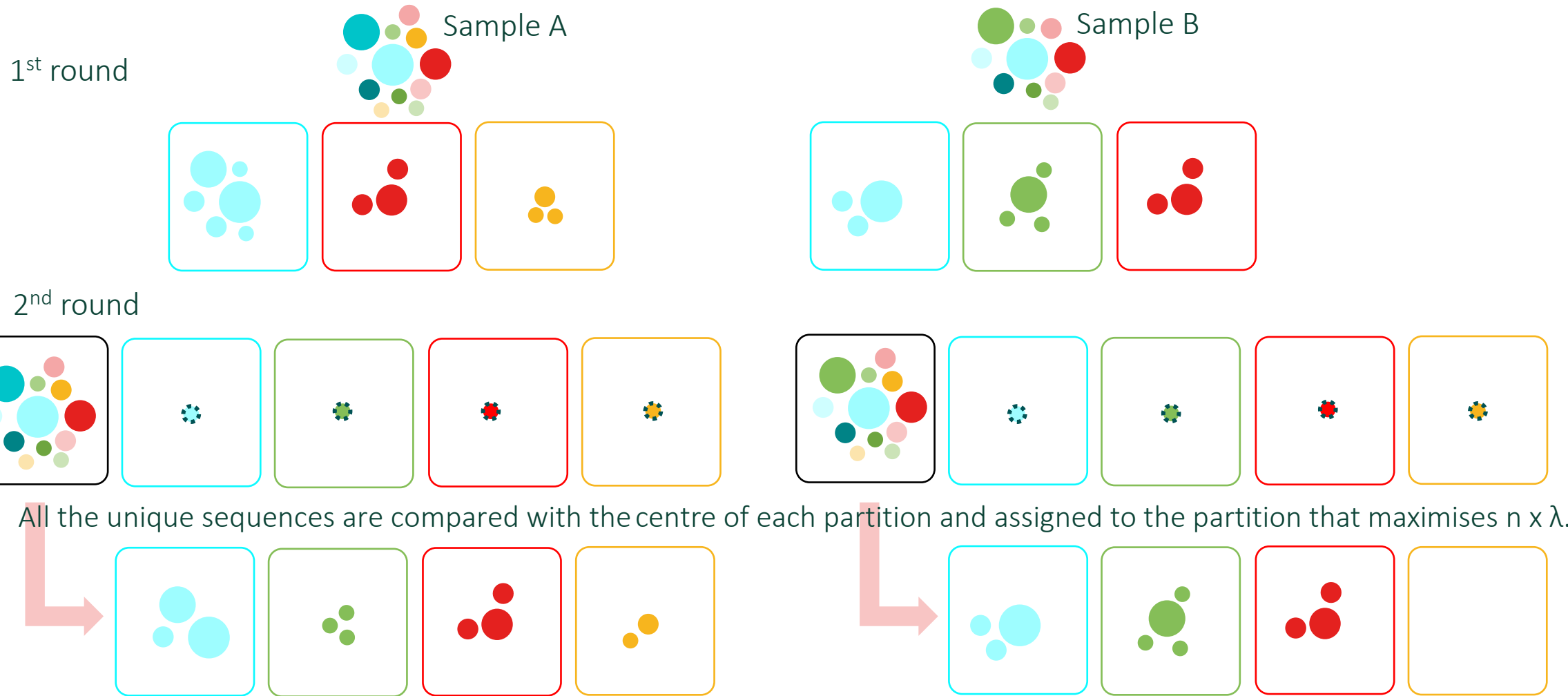
Choose how to consider sample prior to sample inference (--sample-inference)

Recommended

in brief



DADA2 pseudo-pooling



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

in brief





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

Run the process !



What are the outputs ?

FROGS Core Main 1 outputs

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_PPC_S1", "metadata": null}, {"id": "AOP3_PPC_S2", "metadata": null}, {"id": "AOP3_PPC_S3", "metadata": null}, {"id": "AOP3_PPC_S4", "metadata": null}, {"id": "AOP3_PPC_S5", "metadata": null}, {"id": "AOP3_PPC_S6", "metadata": null}, {"id": "AOP3_PPC_W1", "metadata": null}, {"id": "AOP3_PPC_W2", "metadata": null}, {"id": "AOP3_PPC_W3", "metadata": null}, {"id": "AOP3_PPC_W4", "metadata": null}, {"id": "AOP3_PPC_W5", "metadata": null}, {"id": "AOP3_PPC_W6", "metadata": null}, {"id": "AOP4_PPC_S1", "metadata": null}, {"id": "AOP4_PPC_S2", "metadata": null}, {"id": "AOP4_PPC_S3", "metadata": null}, {"id": "AOP4_PPC_S4", "metadata": null}, {"id": "AOP4_PPC_S5", "metadata": null}, {"id": "AOP4_PPC_S6", "metadata": null}, {"id": "AOP4_PPC_W1", "metadata": null}, {"id": "AOP4_PPC_W2", "metadata": null}, {"id": "AOP4_PPC_W3", "metadata": null}, {"id": "AOP4_PPC_W4", "metadata": null}, {"id": "AOP4_PPC_W5", "metadata": null}, {"id": "AOP4_PPC_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, 51], [0, 3, 11], [0, 4, 12], [0, 5, 61], [0, 6, 2], [0, 7, 15], [0, 8, 7], [0, 9, 9], [0, 10, 11], [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, 29], [0, 23, 415], [0, 24, 19993], [0, 25, 45491], [0, 26, 23701], [0, 27, 14317], [0, 28, 15896], [0, 29, 18680], [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, 5890], [1, 17, 6361], [1, 18, 2781], [1, 19, 1360], [1, 20, 1420], [1, 21, 4800], [1, 22, 5417], [1, 23, 4255], [1, 24, 136], [1, 25, 373], [1, 26, 116], [1, 27, 3], [1, 28, 2], [1, 29, 2], [1, 30, 3], [1, 31, 3], [1, 32, 3], [1, 33, 3], [1, 34, 3], [1, 35, 3], [1, 36, 3], [1, 37, 3], [1, 38, 3], [1, 39, 3], [1, 40, 3], [1, 41, 3], [1, 42, 3], [1, 43, 3], [1, 44, 3], [1, 45, 3], [1, 46, 3], [1, 47, 3], [1, 48, 3], [1, 49, 3], [1, 50, 3], [1, 51, 3], [1, 52, 3], [1, 53, 3], [1, 54, 3], [1, 55, 3], [1, 56, 3], [1, 57, 3], [1, 58, 3], [1, 59, 3], [1, 60, 3], [1, 61, 3], [1, 62, 3], [1, 63, 3], [1, 64, 3], [1, 65, 3], [1, 66, 3], [1, 67, 3], [1, 68, 3], [1, 69, 3], [1, 70, 3], [1, 71, 3], [1, 72, 3], [1, 73, 3], [1, 74, 3], [1, 75, 3], [1, 76, 3], [1, 77, 3], [1, 78, 3], [1, 79, 3], [1, 80, 3], [1, 81, 3], [1, 82, 3], [1, 83, 3], [1, 84, 3], [1, 85, 3], [1, 86, 3], [1, 87, 3], [1, 88, 3], [1, 89, 3], [1, 90, 3], [1, 91, 3], [1, 92, 3], [1, 93, 3], [1, 94, 3], [1, 95, 3], [1, 96, 3], [1, 97, 3], [1, 98, 3], [1, 99, 3], [1, 100, 3], [1, 101, 3], [1, 102, 3], [1, 103, 3], [1, 104, 3], [1, 105, 3], [1, 106, 3], [1, 107, 3], [1, 108, 3], [1, 109, 3], [1, 110, 3], [1, 111, 3], [1, 112, 3], [1, 113, 3], [1, 114, 3], [1, 115, 3], [1, 116, 3], [1, 117, 3], [1, 118, 3], [1, 119, 3], [1, 120, 3], [1, 121, 3], [1, 122, 3], [1, 123, 3], [1, 124, 3], [1, 125, 3], [1, 126, 3], [1, 127, 3], [1, 128, 3], [1, 129, 3], [1, 130, 3], [1, 131, 3], [1, 132, 3], [1, 133, 3], [1, 134, 3], [1, 135, 3], [1, 136, 3], [1, 137, 3], [1, 138, 3], [1, 139, 3], [1, 140, 3], [1, 141, 3], [1, 142, 3], [1, 143, 3], [1, 144, 3], [1, 145, 3], [1, 146, 3], [1, 147, 3], [1, 148, 3], [1, 149, 3], [1, 150, 3], [1, 151, 3], [1, 152, 3], [1, 153, 3], [1, 154, 3], [1, 155, 3], [1, 156, 3], [1, 157, 3], [1, 158, 3], [1, 159, 3], [1, 160, 3], [1, 161, 3], [1,
```

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

FROGS Core Main 1 outputs

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
```

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

FROGS Core Main 1 outputs

4th output, for swarm only !

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

```
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:HLVVMBCX2:1:2204:20108:18134/1_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:HLVVMBCX2:2:2101:14487:7701/1_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:HLVVMBCX2:1:2213:4782:91658/1_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:HLVVMBCX2:2:1107:20508:16134/1_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

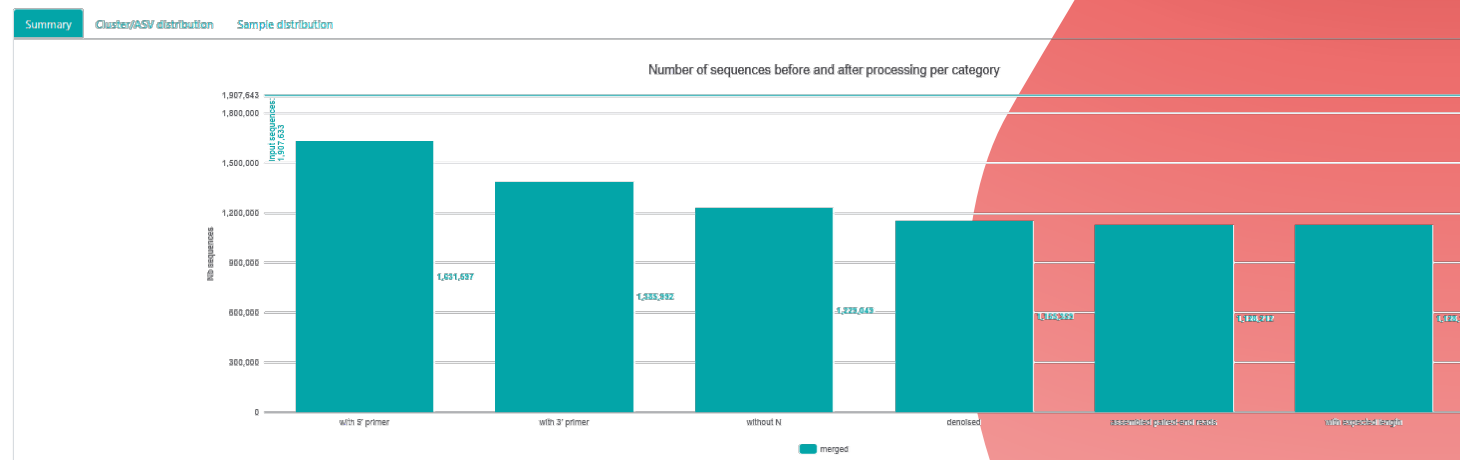
4: FROGS Core 1-Main - processing_short_reads: sequence.fasta

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

Data visualization in html report



Reads-processing report (*reads_processing.py, v5.2.0*)



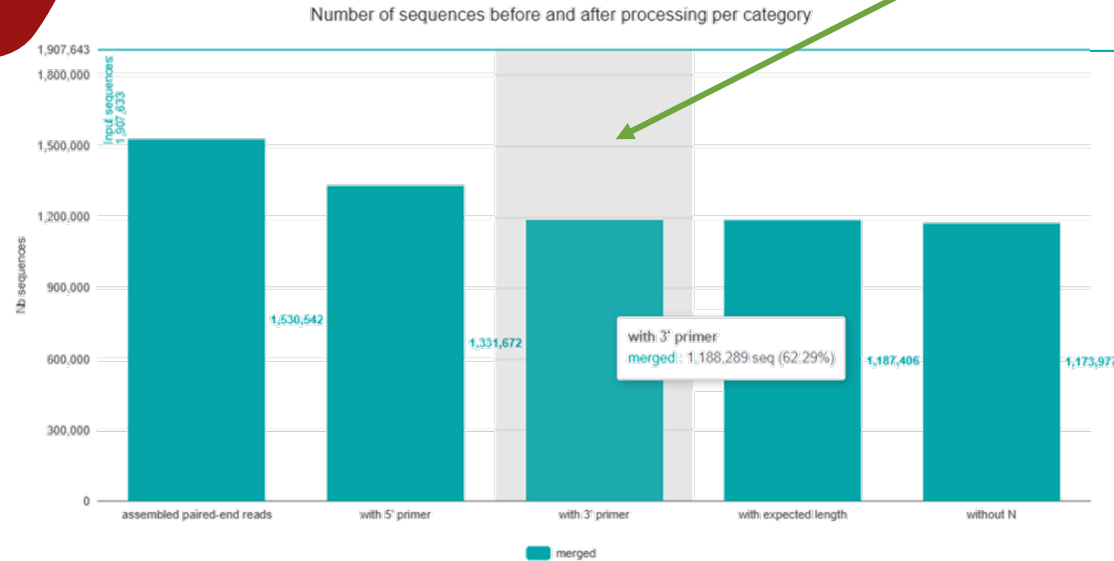
Practice session



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.

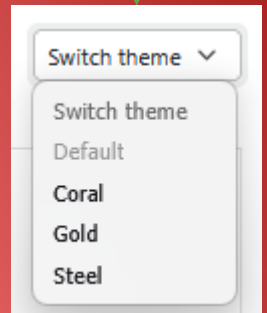
You can change the theme color.

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

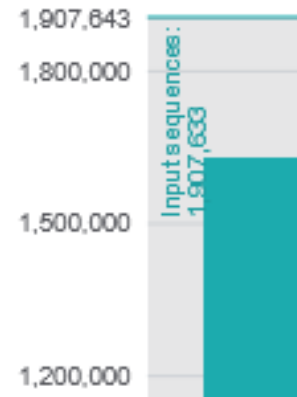
How many sequences are there in the input file ?

16S

DADA2

Summary

Cluster/ASV distribution

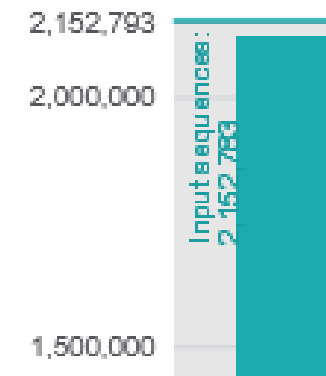


ITS

DADA2

Summary

Cluster/ASV distribution



Practice session

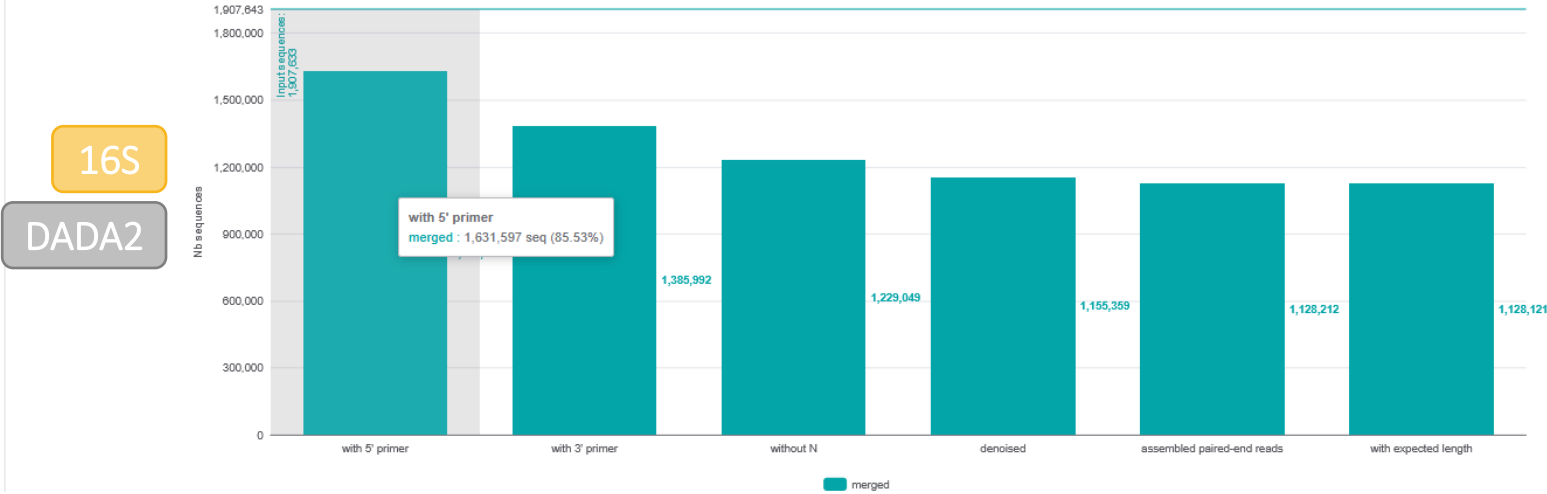
How many sequences did not have the 5' primer ?

Summary

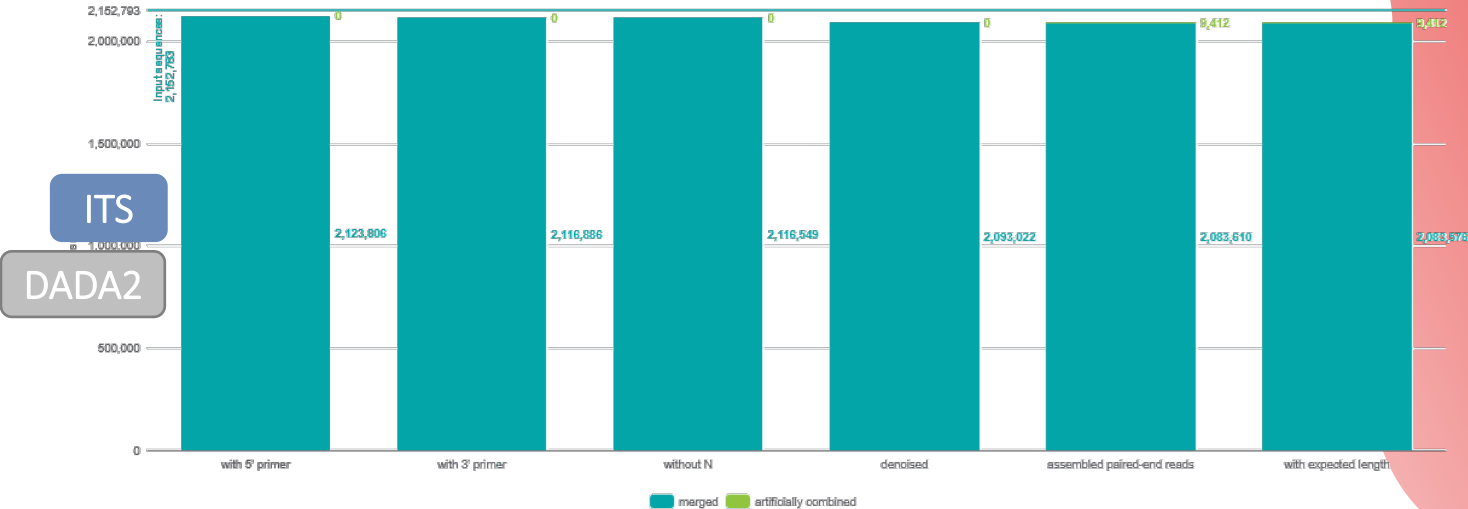
Cluster/ASV distribution

Sample distribution

Number of sequences before and after processing per category

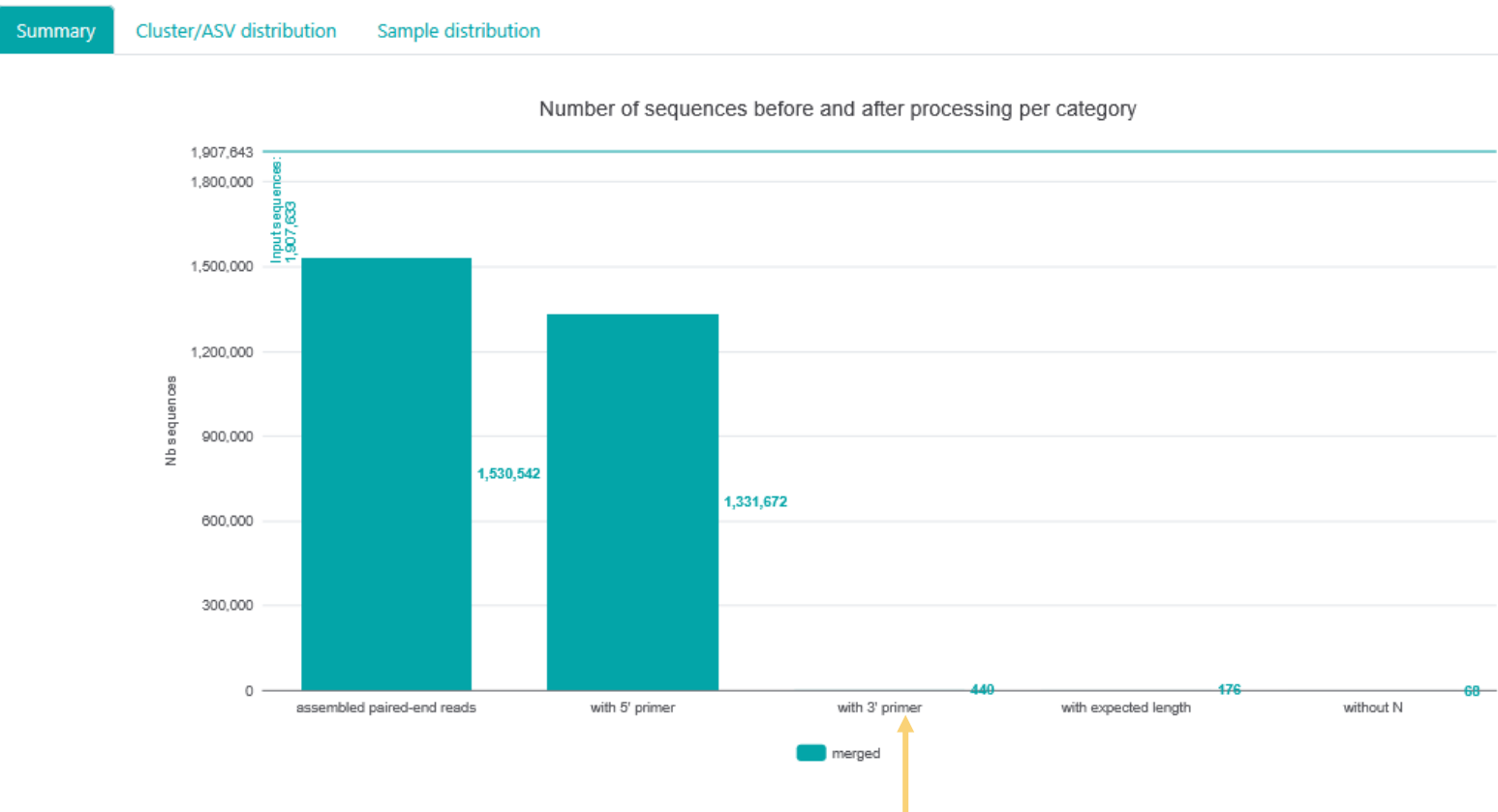


Number of sequences before and after processing per category



Practice session

What conclusions can you draw from this graph?

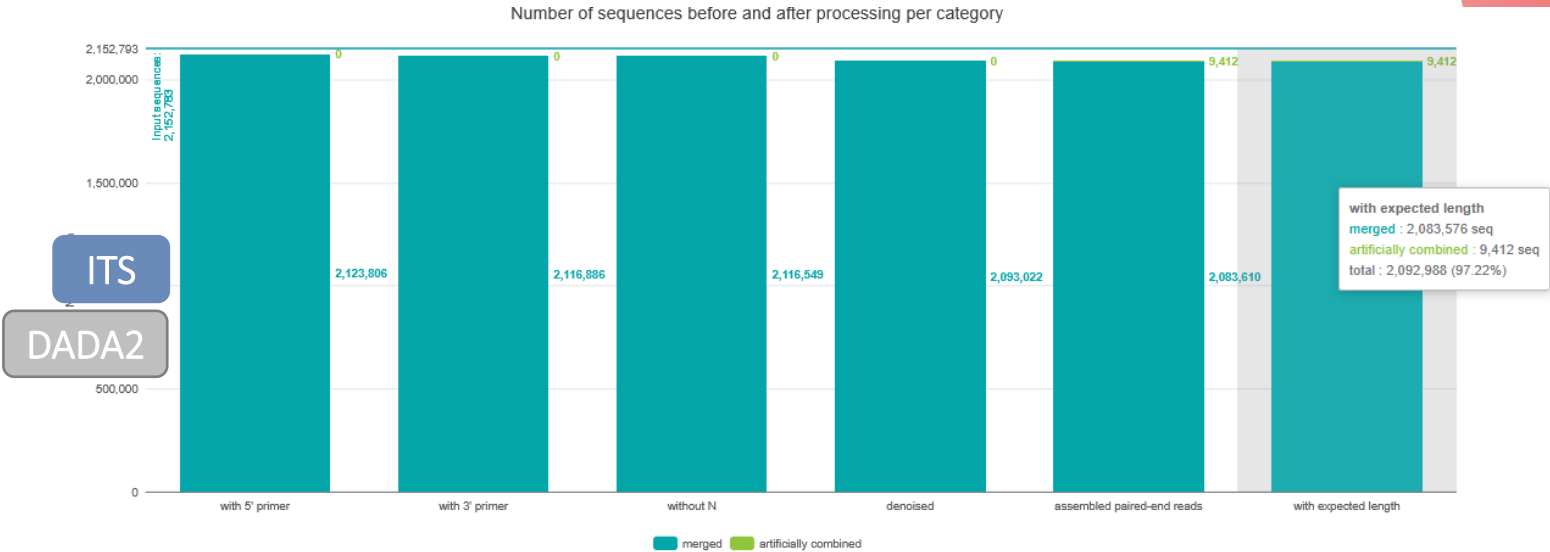
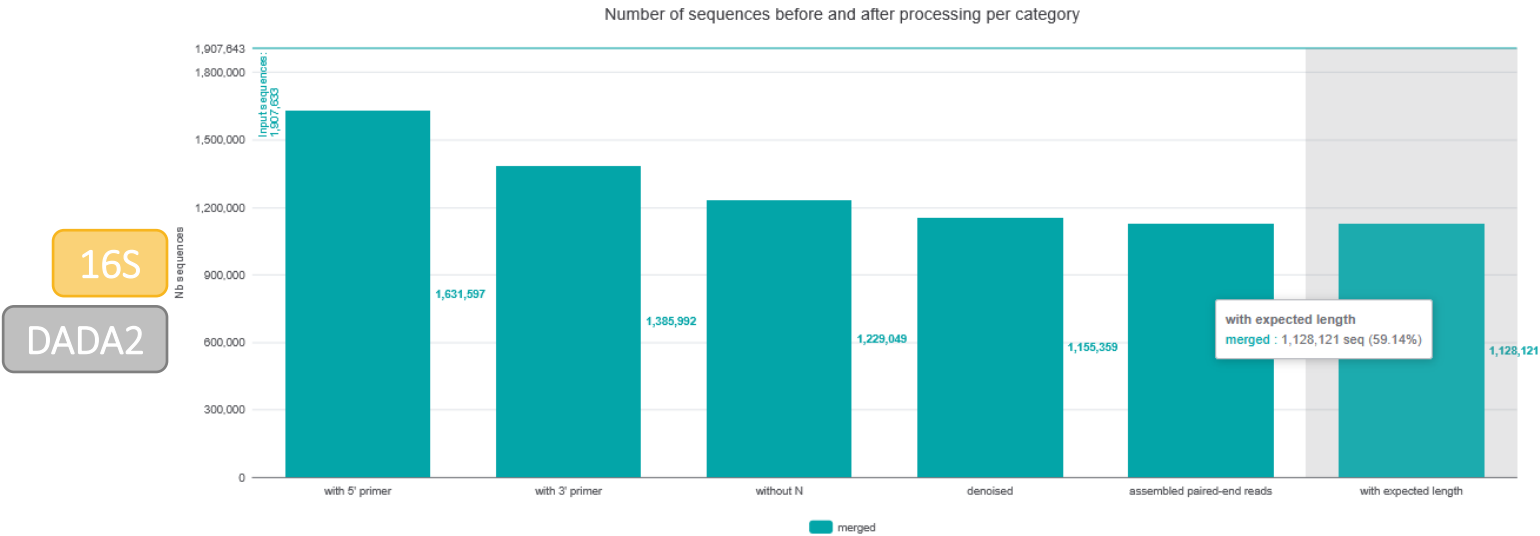


It is likely that the user entered the 3' primer in the wrong direction.

Practice session

How many sequences remain after the data has been preprocessed ?

Summary Cluster/ASV distribution Sample distribution



Practice session

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

To select all samples

☒	Sample name	before process	assembled paired-end reads
☒	AOP1_PPC_S1	26,446	19,028
☒	AOP1_PPC_S2	20,019	14,586
☒	AOP1_PPC_S3	21,570	15,162
☒	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
before reads processing

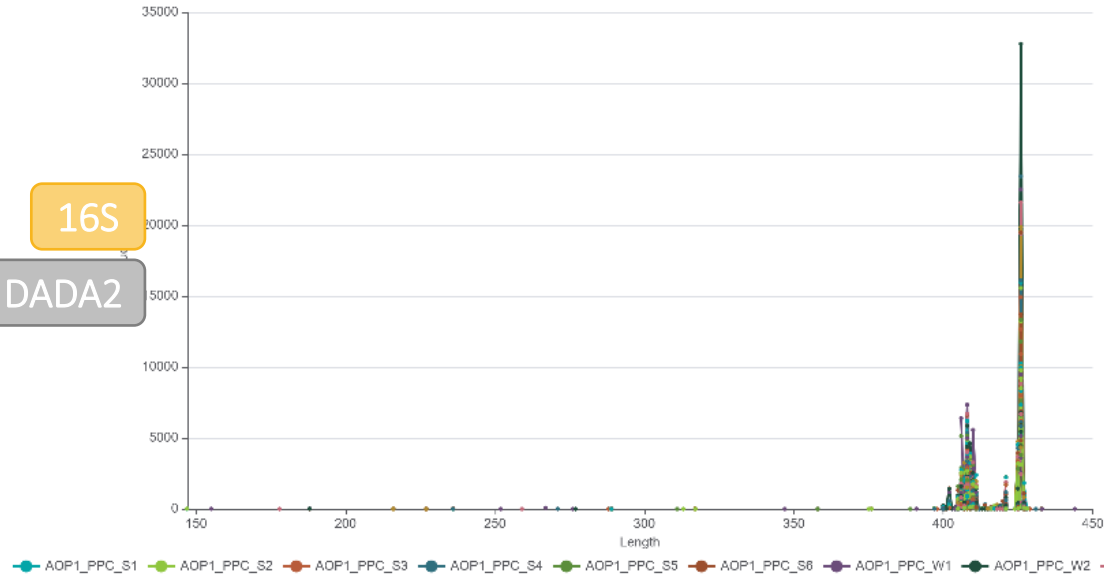
To show all samples

Practice session

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

Amplicons lengths

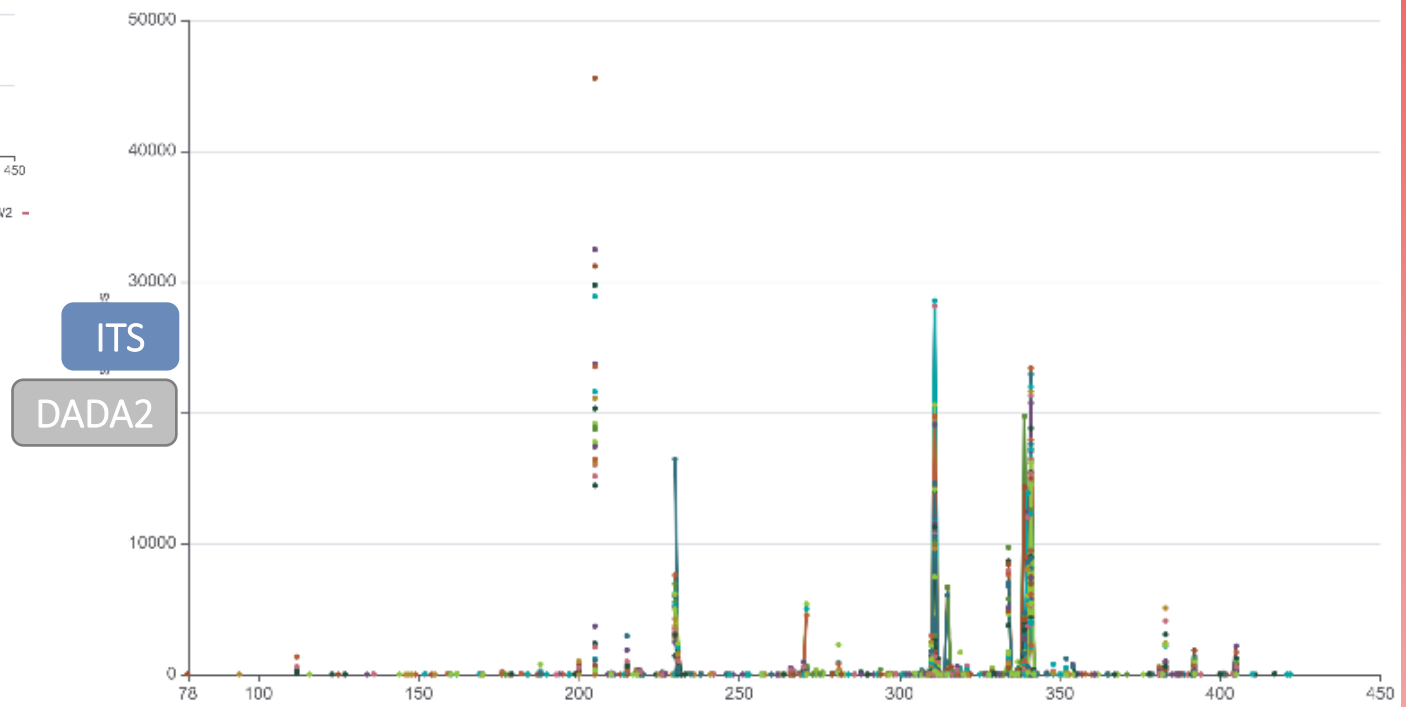
Amplicon length distribution before trimming and filtering



16S

DADA2

Amplicon length distribution before trimming and filtering



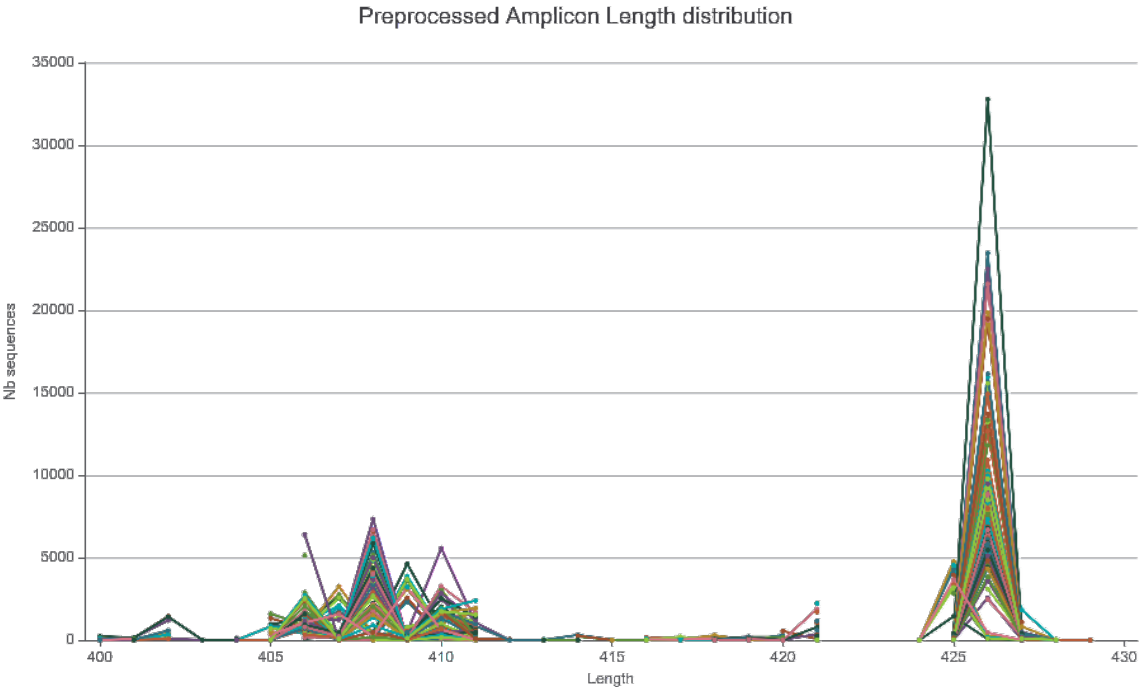
ITS

DADA2

Practice session

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

16S
DADA2



One pic = one amplicon length $\approx$ 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

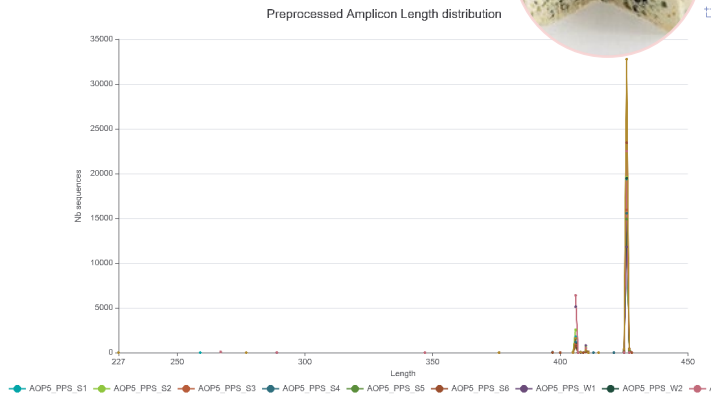
To display amplicon length after reads processing

Practice session

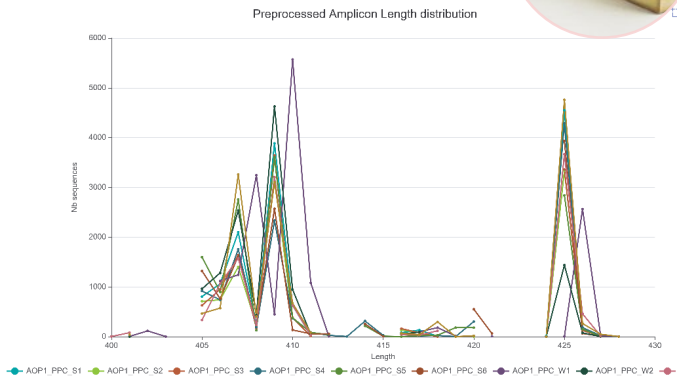
Based on amplicon size distributions, what can you tell us about the samples per cheese category ?

16S
DADA2

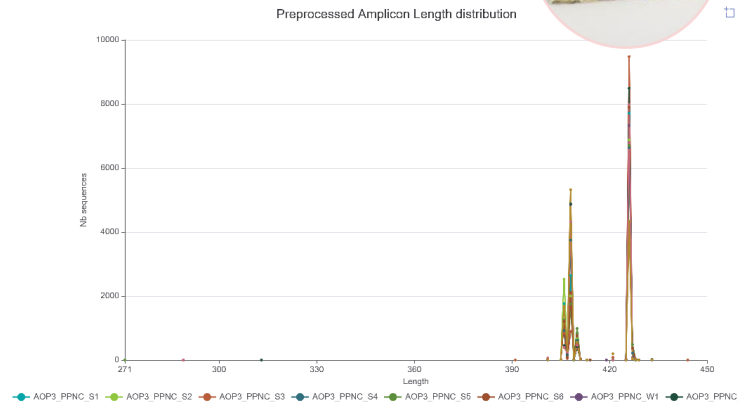
PPS



PPC



PPNC



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

DADA2

Clusters/ASVs

7,480

Sequences

1,128,121

ITS

DADA2

Clusters/ASVs

2,657

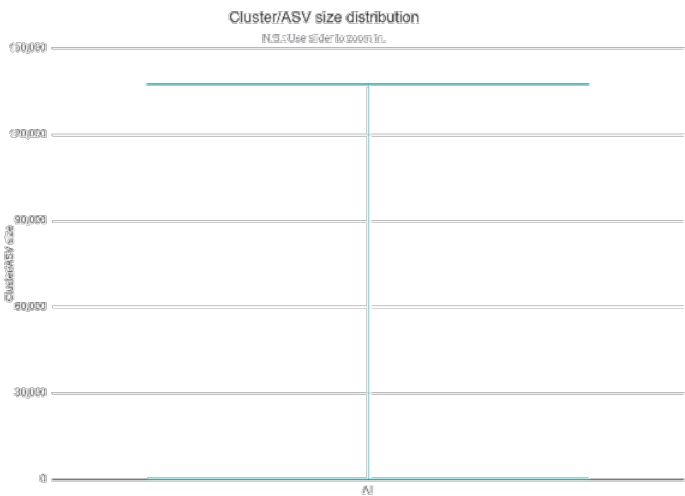
Sequences

2,092,988

Interpret the boxplot: Cluster/ASV size summary

Cluster/ASV size summary

The cluster/ASV size is the sum of the abundances of the sequences grouped in a cluster/ASV.



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

Cluster/ASV size distribution (decile)

Most of clusters are singletons

Practice session

How many single singletons do you find?

16S
DADA2

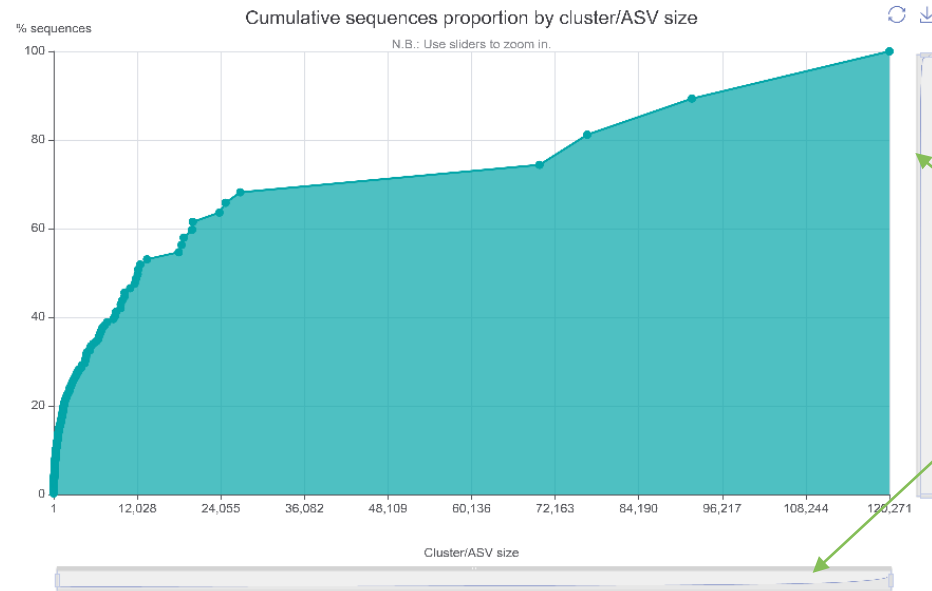
Cluster/ASV size	Number of clusters/ASVs	% of all clusters/ASVs
1	2,907	38.82

ITS
DADA2

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

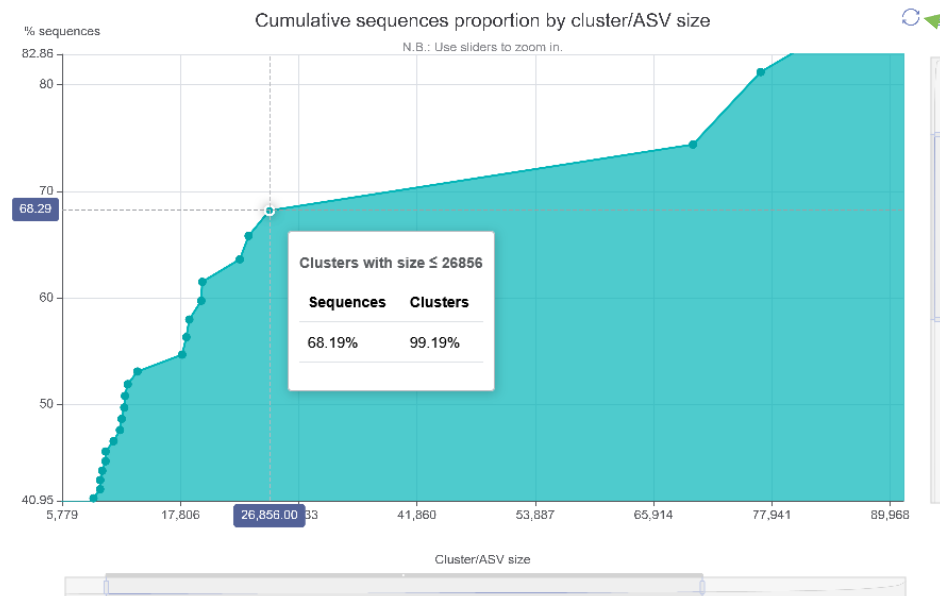
Tips !

Cluster/ASV size distribution



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

Cluster/ASV size distribution



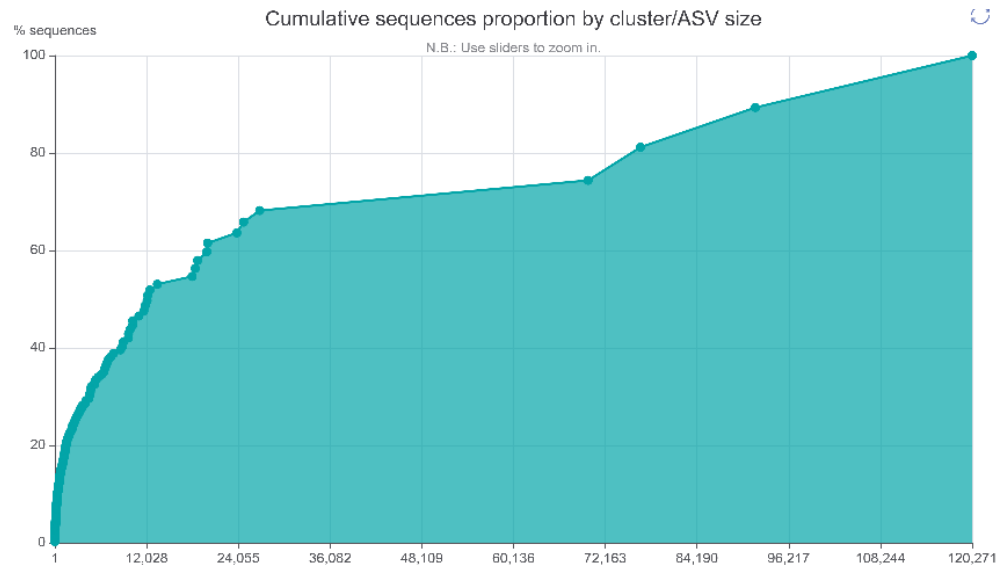
Restore the initial picture.

Practice session

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

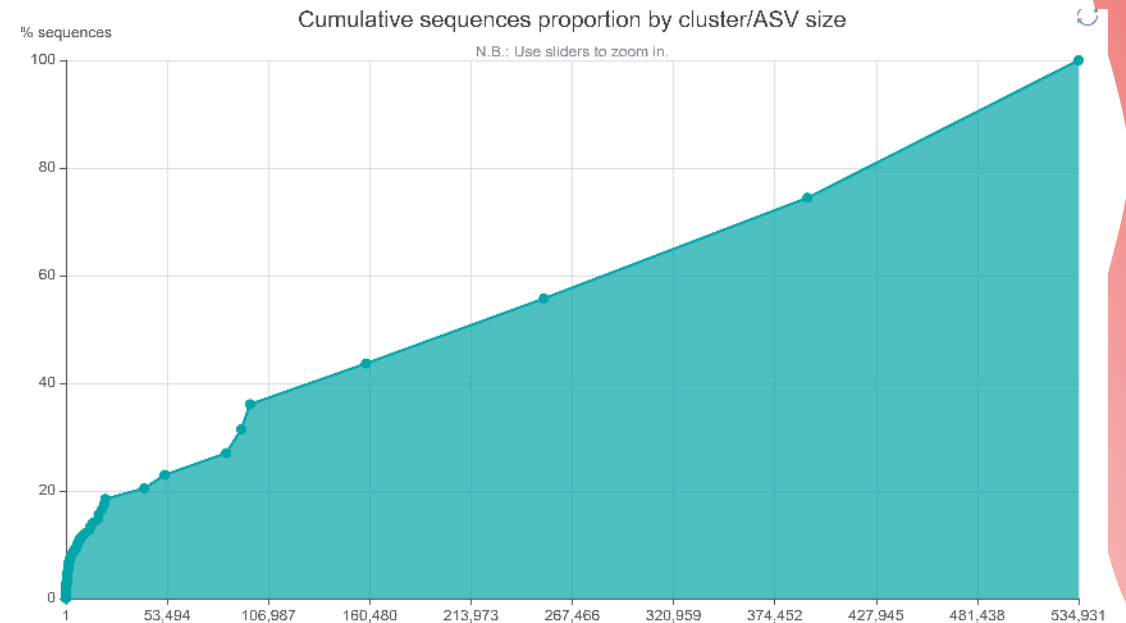
16S

DADA2



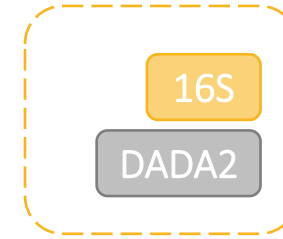
ITS

DADA2



Practice session

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



Sequence count

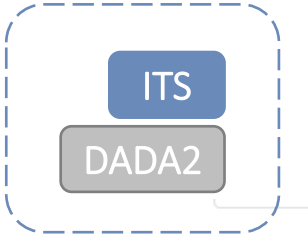
Search 					
Sample name	Number of clusters/ASVs	Number of shared clusters/ASVs	Number of own clusters/ASVs	Number of sequences	Number
AOP1_PPC_S1	343	232	111	13,729	13,169
AOP1_PPC_S2	285	206	79	10,417	10,171
AOP1_PPC_S3	173	151	22	11,169	11,056
AOP1_PPC_S4	346	248	98	11,610	11,303
AOP1_PPC_S5	345	236	109	13,120	12,558
AOP1_PPC_S6	294	225	69	11,645	11,232
AOP1_PPC_W1	370	270	100	15,727	14,759
AOP1_PPC_W2	283	224	59		
AOP1_PPC_W3	283	214	69	11,499	11,277
AOP1_PPC_W4	326	248	78		

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?



Summary Cluster/ASV distribution Sample distribution

Sequence count

AOP1_PPC_S|

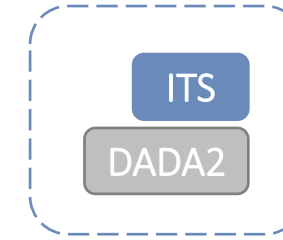
Sample name	Number of clusters/ASVs	Number of shared clusters/ASVs	Number of own clusters/ASVs	Number of sequences
AOP1_PPC_S1	195	174	21	28,861
AOP1_PPC_S2	410	277	133	32,900
AOP1_PPC_S3	286	233	53	35,299
AOP1_PPC_S4	305	235	70	27,060
AOP1_PPC_S5	363	200	163	
AOP1_PPC_S6	316	222	94	26,540

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”?



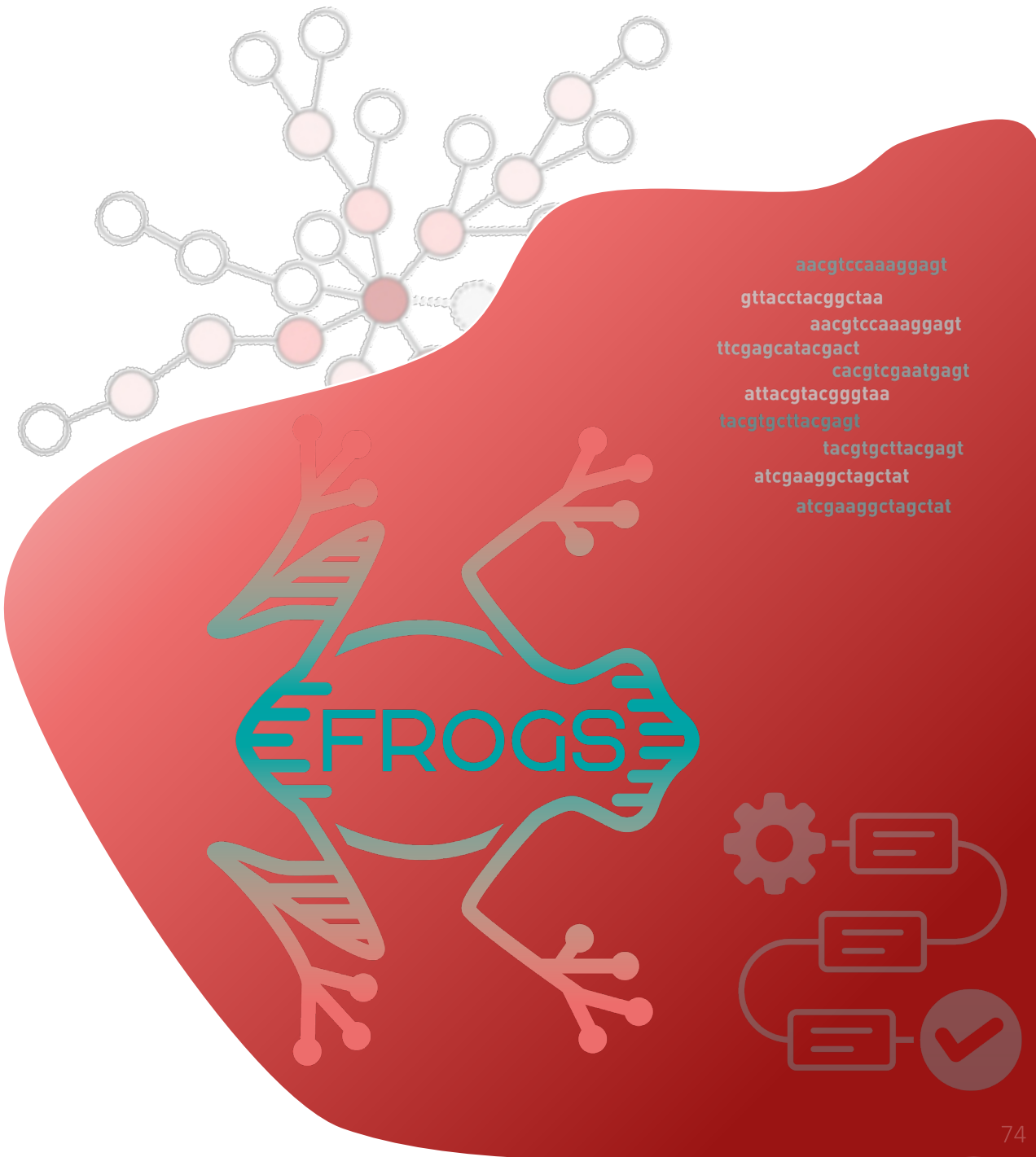
Sample name	Number of own clusters/ASVs	Number of sequences	Number of shared sequences	Number of own sequences
AOP1_PPC_S1	21	28,861	28,787	74
AOP1_PPC_S2	133	32,900	31,555	1,345
AOP1_PPC_S3	53	35,299	34,825	474
AOP1_PPC_S4	70	27,060	26,831	229
AOP1_PPC_S5	163	33,343	31,487	1,856
AOP1_PPC_S6	94	26,540	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)

aborted amplification



next cycle's "primer"



chimeric sequence

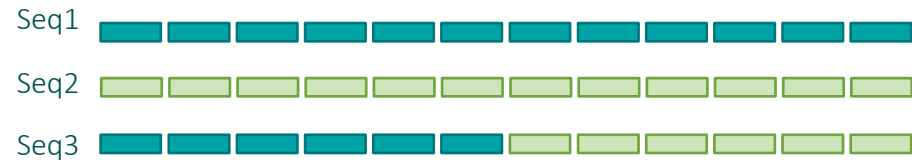


Chimera detection

1 Sample sequences.

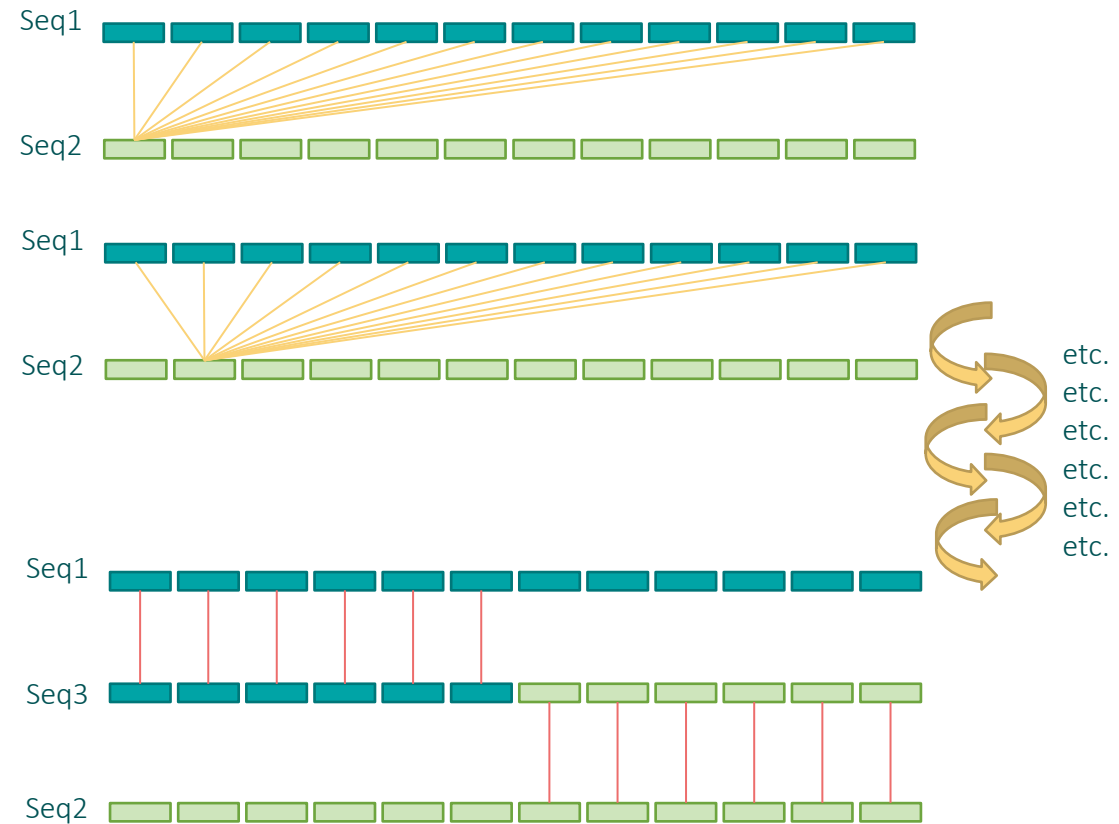


2 Dividing into fragments.



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.

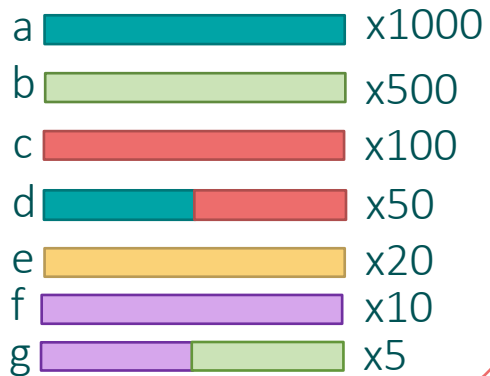


The best hits of Seq3's alignment are found in both Seq1 and Seq2.
Seq3 is considered as a chimera.

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

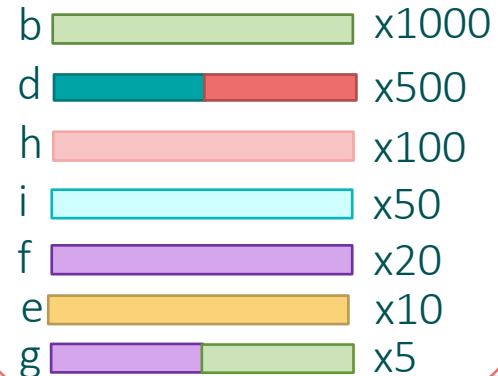
Cross-validation to delete false-positive

Sample A

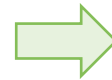


“d” is view as chimera by Vsearch
Its “parents” are presents

Sample B



“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 ?

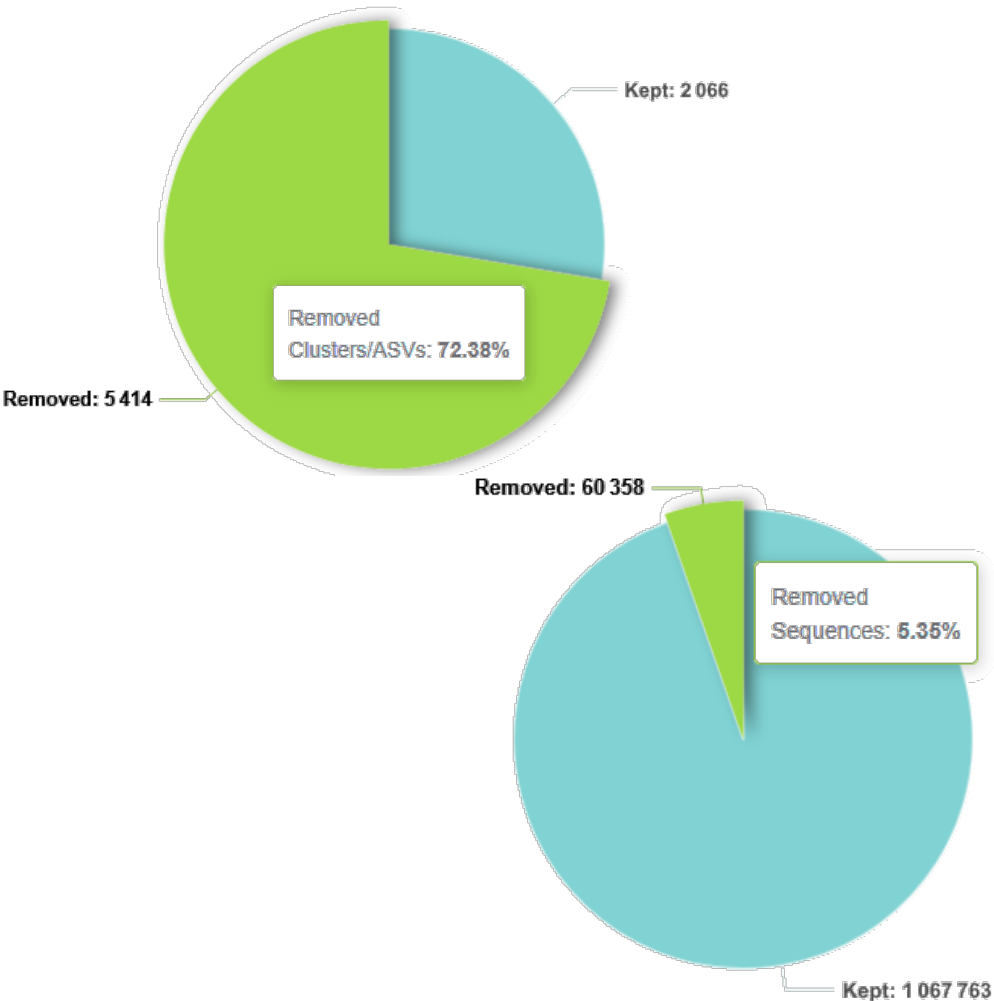
What is your conclusion ?

Practice session

Removed clusters are low abundance clusters.

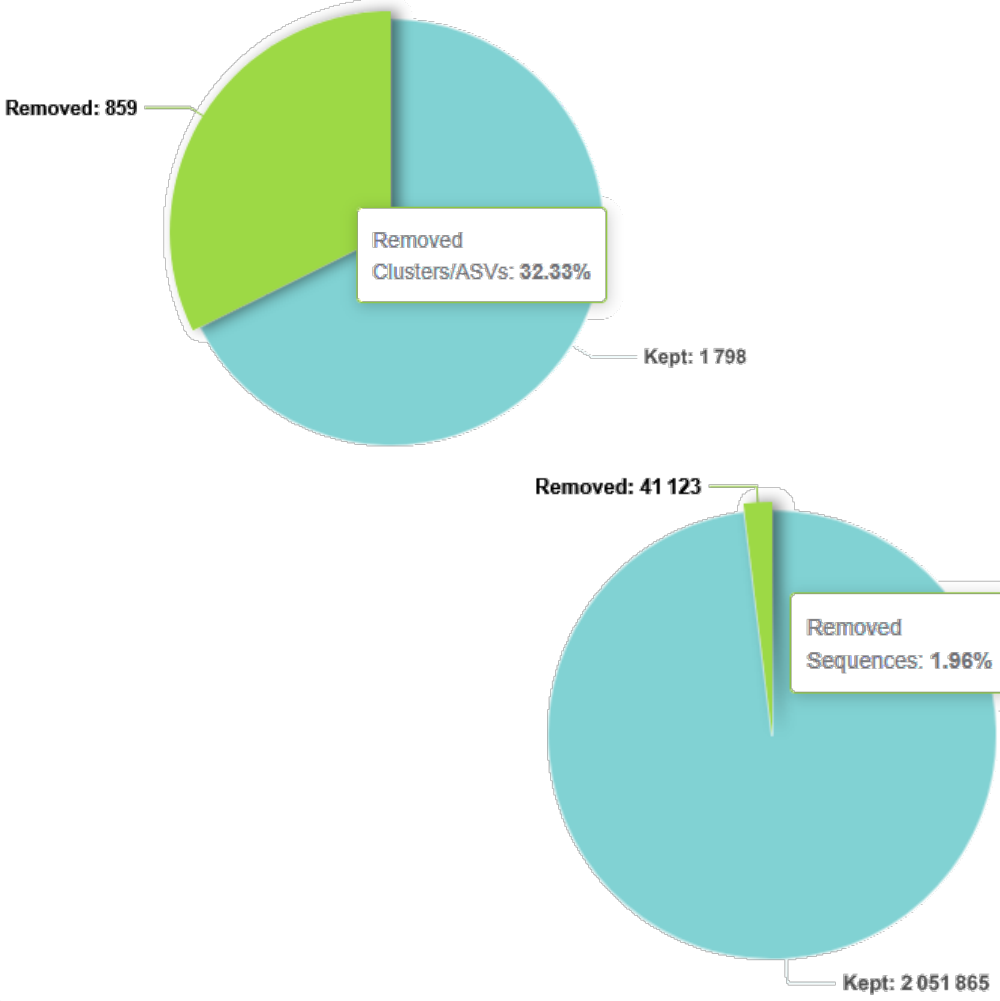
16S

DADA2



ITS

DADA2



impact of chimera removing per sample

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

Number of chimeric removed clusters ▲	Abundance of chimeric removed clusters ◆	Abundance of the most abundant removed chimera ◆
53	141	10
72	273	13
85	637	41

Practice session



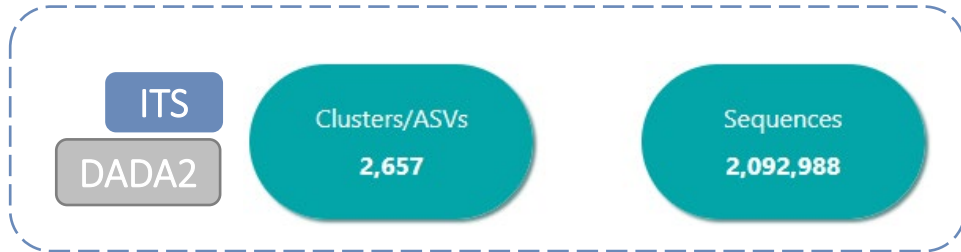
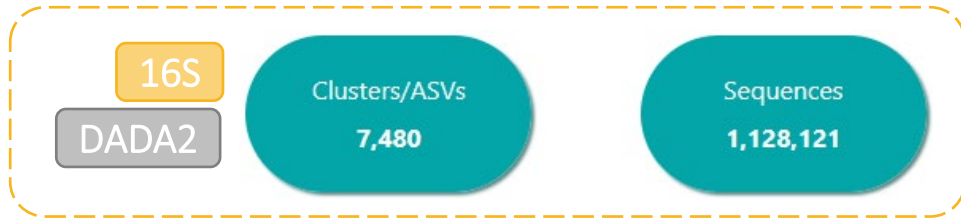
How many clusters do you get ?

Please, focus on singletons

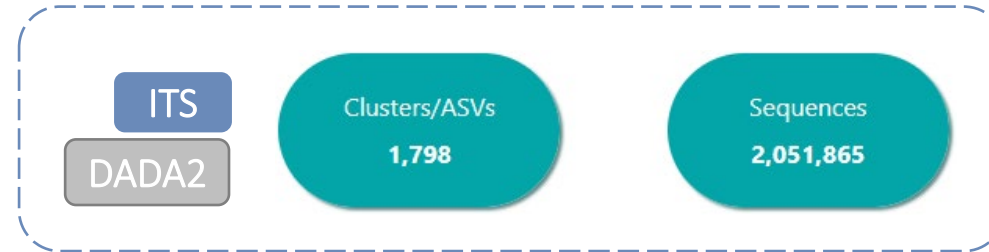
Practice session

How many clusters do you get ?

Before



Now



Practice session

Focus on singleton

16S

DADA2

Before

Cluster/ASV size	Number of clusters/ASVs	% of all clusters/ASVs
1	2,912	38.93

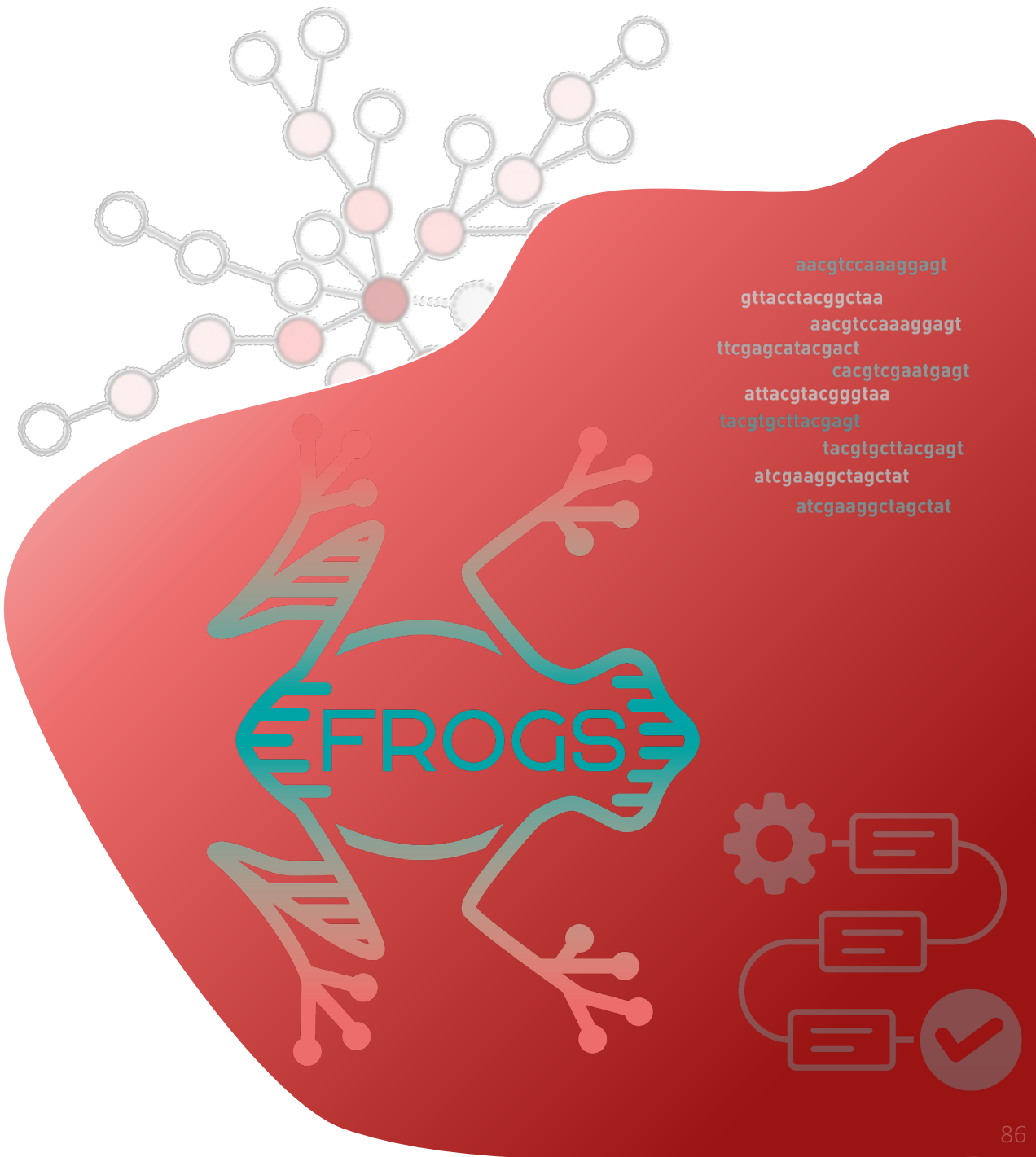
Now

Cluster/ASV size	Number of clusters/ASVs	% of all clusters/ASVs
1	452	21.88

FROGS Core 1

Main tools

Cluster/ASV filter



aacgtccaaaggagt
gttacctacggctaa
aacgtccaaaggagt
ttcgagcatagact
cacgtcgaatgagt
attacgtacgggtaa
tacgtgcttacgagt
tacgtgcttacgagt
atcgaaggctagctat
atcgaaggctagctat



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 abundance

Filter on the most abundant

Filter on contaminant

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

Prevalence across all samples

Prevalence by sample groups (metadata file required)

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`)

The user wants each ASV to be present in at least 3 samples.

Filter on prevalence

Filter on prevalence

Prevalence across all samples

Prevalence by sample groups (metadata file required)

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

Prevalence by sample groups (metadata file required)

Sample replicates group (.tsv) - optional



44: MPC_replicate_metadata.tsv

accepted formats ▼

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

Minimum prevalence (%) - optional

0.5

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)

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

Filter on prevalence

Prevalence across all samples

Prevalence by sample groups (metadata file required)

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

Filter on prevalence

Prevalence across all samples

Prevalence by sample groups (metadata file required)

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

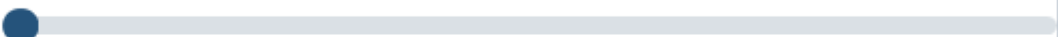
Filter on abundance ^

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





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`)

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

Filter on abundance

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 number of sequences to keep ASVs/clusters (integer > 1) - optional

100

Leave empty to skip this abundance filtering. For example, 2 keeps ASVs/cluster with at least two sequences (i.e. removes single singleton) (--min-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

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`)

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 prevalence

Filter on abundance

Filter on the most abundant

Filter on contaminant



Filter on contaminant

Filter on contaminant

Search for contaminant ASVs/clusters

Use contaminant FASTA file from FROGS

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

Contaminant reference database *

Select Value

phiX *

Arabidopsis TAIR10 Chloroplast and mitochondria

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>

```
>ChrC CHROMOSOME dumped from ADB: Jun/20/09 14:54; last updated: 2005-06-03
ATGGGCGAACGACGGGAATTGAACCGCGATGGTGAATTCACAATCCACTGCCTTAATCCACTTGGCTACATCCGCCCC
TACGCTACTATCTATCTTTTTGATTGTCTAAAAAATACAAATTCATAAAAAAAGGTAG
CAAATCCACCTATTTTTCTAATAAAAAATATAGTAATTTTTATTTATTTATTTATTTATTTATTAATA
TAATAAATAAGTAAATATGATACTCTATAAAATTTGCTCATTTTATAGAAAAAACGAGTAATATAAGCCCTCTT
TCTTATTTAAGAAGGCTTATATTGCTGTTTTTACTAACTAGATCTAGACTAACACTAACGAATATCCATTGTGTA
GATGGAGCTCAACAGAGCTAGGCTAGAGGAGCTTGTGAGCATTAGGCTCATGCTAATCTTCATAGCAGGTAG
```

```
>ChrH CHROMOSOME dumped from ADB: Jun/20/09 14:54; last updated: 2005-06-03
GGATCCGTTTGAACAGGTTAGCCTACTATAATAAGGATTGGATTCTAATAAGTTTGAACAGGTTAGCCTTAGCCT
ACTATAGGATTAGATCTTTCTATCAACCTACTAATCTTCTTCTTGTGGGATGAGAACCTTTTGAACCAAGCGTG
CTTTGAGTTTGTCAAGGACCATCTGCATTCAGTTTCACTCTGAAACCCATTTACAACGAGAGATTATGTCAGG
TGATGCGGGAACATAAGTCCAGGTGATCTGTGTTAATGCGGACATCTCTTCTGATAGCTTGTCTCCATCCTGCG
AGGACAGACTAATGGTTTTGGTTCAAGGGAGTGATTTTTGTGTAACAGGTTGTAACGAGGATTAGGCTTGCAGAT
ACCATCCTTTGCCGAGTGATCATATGATGTCATTAGGTGAAAGTAGCTCAGGAGCAGCTGTCCCAACATCAAAAAAG
GTAGGCTGTGCGCATAGGAACAGGATCTGAGGCTGCGGATGCAAGAGGAGATCTTTCTGATGCTGAGGATG
```

Filter on contaminant

Filter on contaminant ^

Search for contaminant ASVs/clusters

Use your contaminant FASTA file from history

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

Select a contaminant reference from history *

...

54: contaminant.fasta

accepted formats ▼

Provide your own contaminant reference FASTA file from history. (--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 JJ, 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

☐ 2. Remove chimeras

☒ 3. Cluster/ASV filters

☐ 4. Taxonomic annotation

☐ 5. Phylogenetic tree building

☐ ITSx

Sequence file (.fasta) *



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

accepted formats ▼

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

Abundance file (.biom) *



10: FROGS Core 1-Main - remove_chimeras: 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)



- 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**

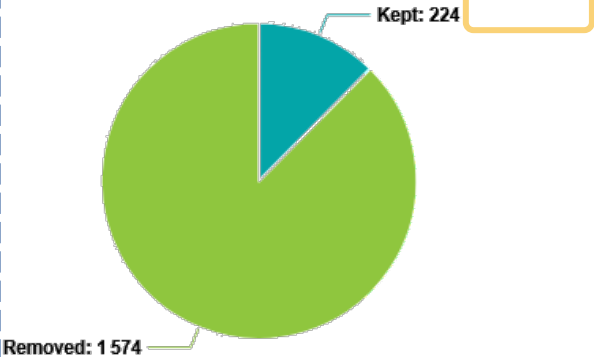
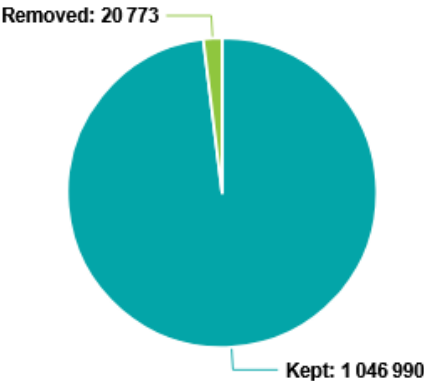
Practice session

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



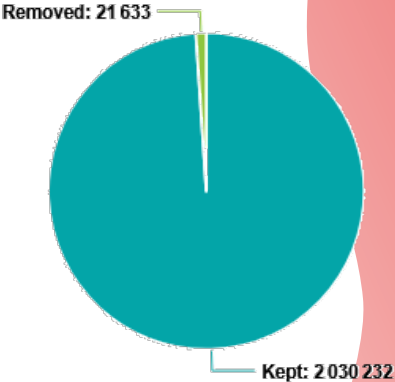
16S

DADA2



ITS

DADA2



Practice session

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



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)	
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

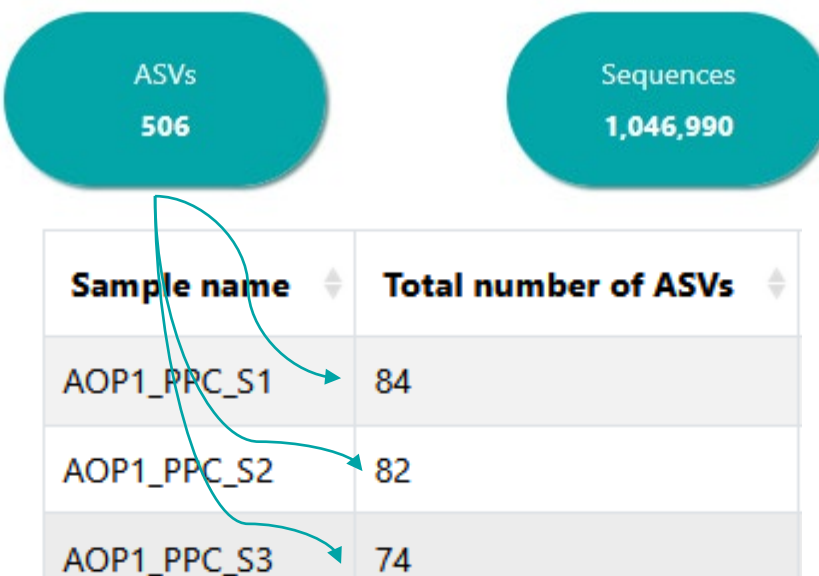
DADA2

ASVs

506

Sequences

1,046,990



Sample name	Total number of ASVs
AOP1_PPC_S1	84
AOP1_PPC_S2	82
AOP1_PPC_S3	74
AOP1_PPC_S4	96
AOP1_PPC_S5	96
AOP1_PPC_S6	84

Practice session

How are ASVs distributed across the different samples?

ITS

DADA2

ASVs
224

Sequences
2,030,232

Sample name	Total number of ASVs	Number of shared ASVs	Number of own ASVs
AOP1_PPC_S1	67	67	0
AOP1_PPC_S2	73	73	0
AOP1_PPC_S3	71	71	0
AOP1_PPC_S4	74	74	0
AOP1_PPC_S5	75	75	0
AOP1_PPC_S6	70	70	0

Practice session

How many ASVs do you get ?

Before



Now



How many clusters did you remove with each filter?

Filters by ASV

Filters by sample

ASV distribution

Sample distribution

Filters intersections

Venn diagram to identify which ASVs were removed by the selected filters.(Maximum 6 options):

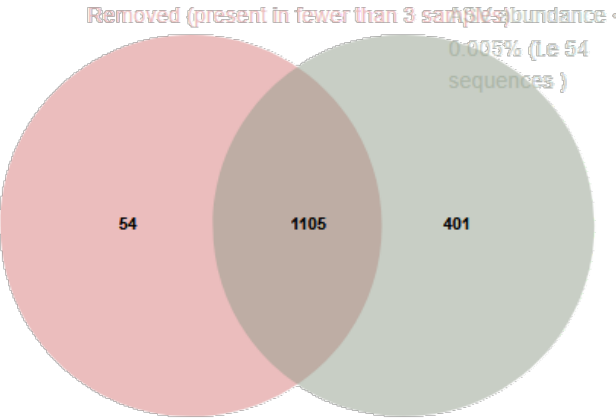
- ☒ Removed (present in fewer than 3 samples)
- ☒ ASV abundance < 0.005% (i.e 54 sequences)

 Venn

Practice session

How many clusters did you remove with each filter?

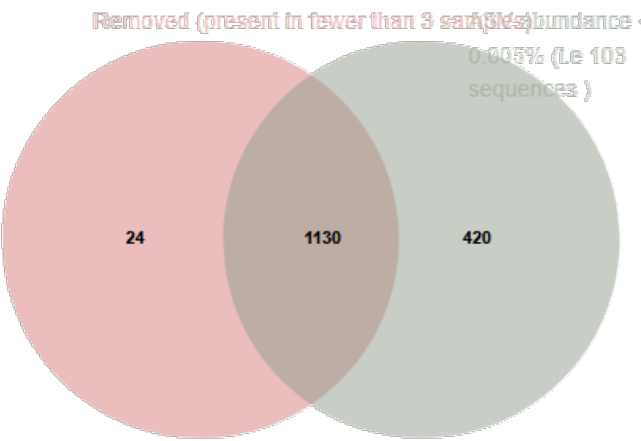
Venn on removed ASVs



16S

DADA2

Venn on removed ASVs



ITS

DADA2

Practice session

Summary Cluster/ASV distribution Sample distribution

How many own ASV remains in AOP1_PPC_S3 ?

Before

16S

DADA2

Sequence count

Search

Sample name	Number of clusters	Number of shared clusters	Number of own clusters
AOP1_PPC_S1	138	120	18
AOP1_PPC_S2	132	119	13
AOP1_PPC_S3	101	92	9
AOP1_PPC_S4	167	145	22
AOP1_PPC_S5	173	144	29
AOP1_PPC_S6	145	127	18

Now

16S

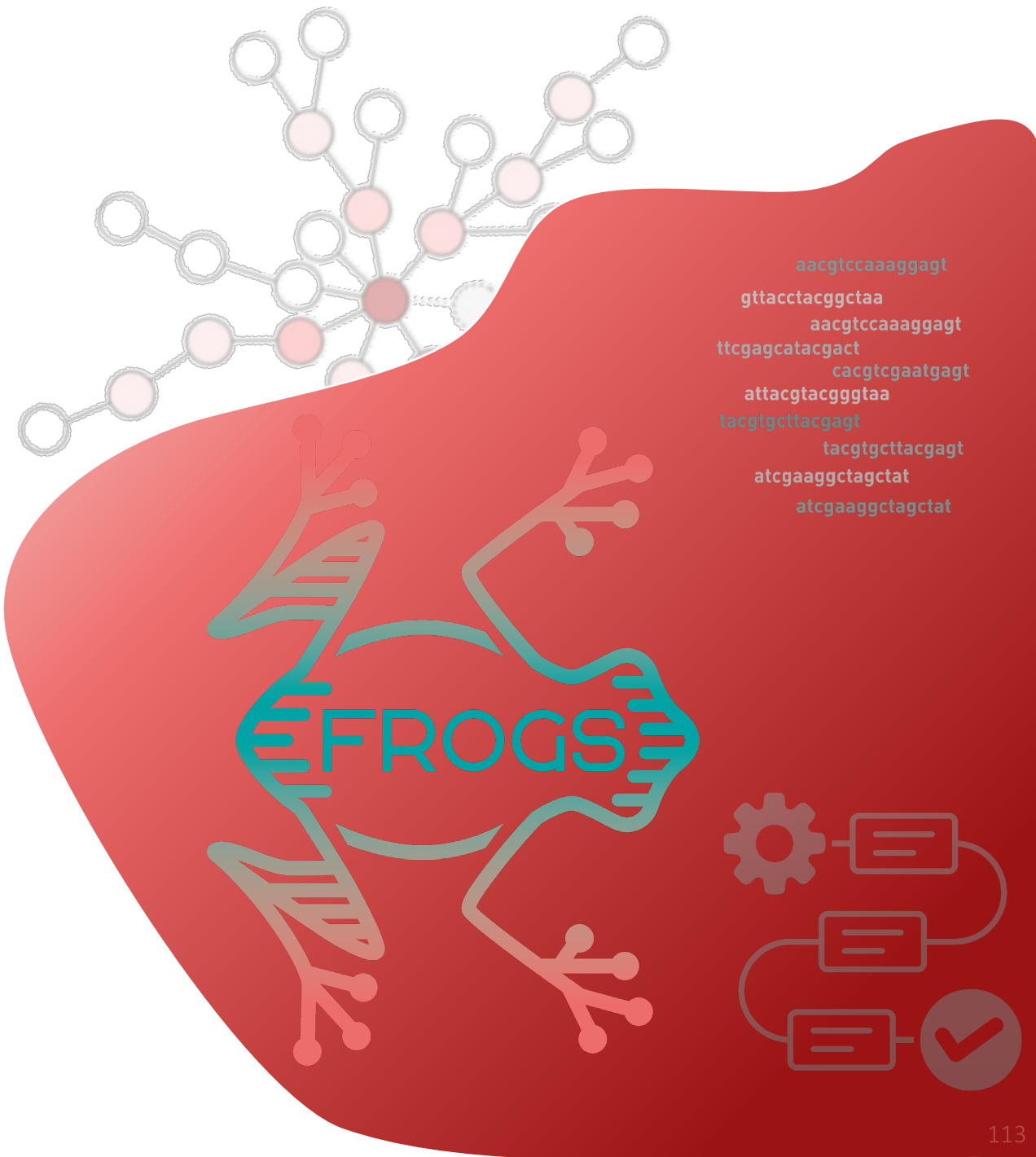
DADA2

Sample name	Total number of ASVs	Number of shared ASVs	Number of own ASVs
AOP1_PPC_S1	84	84	0
AOP1_PPC_S2	82	82	0
AOP1_PPC_S3	74	74	0
AOP1_PPC_S4	96	96	0
AOP1_PPC_S5	96	96	0
AOP1_PPC_S6	84	84	0

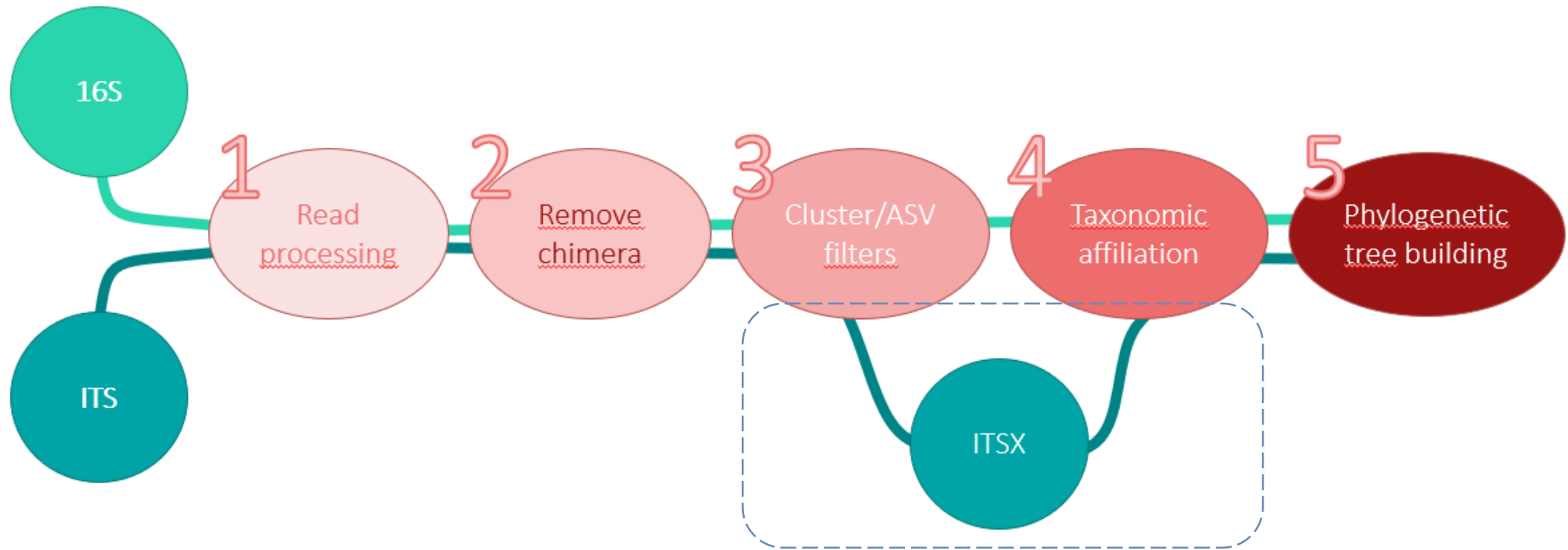
FROGS Core 1

Main tools

ITSX filter



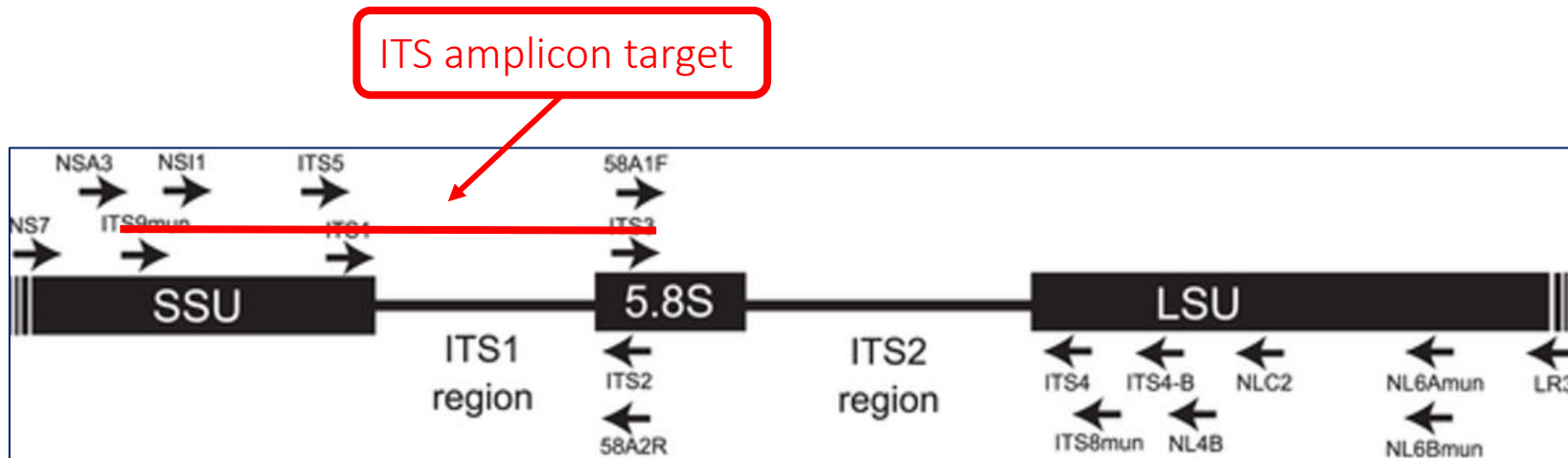
ITS data has an additional filter.



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.

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

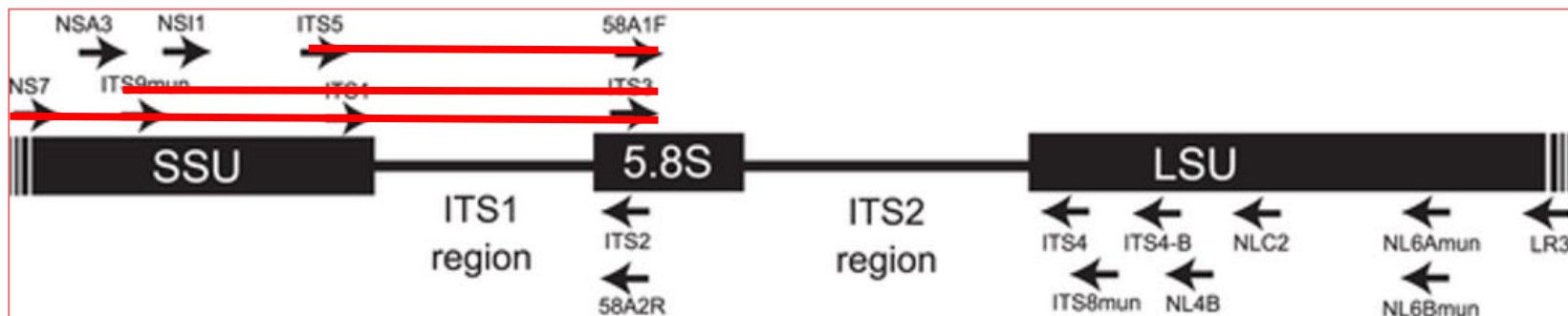
2nd case: identification & trimming of conserved regions

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

- ☒ 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)

ITS

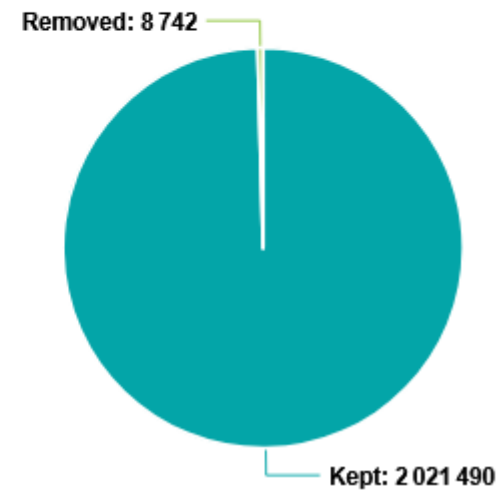
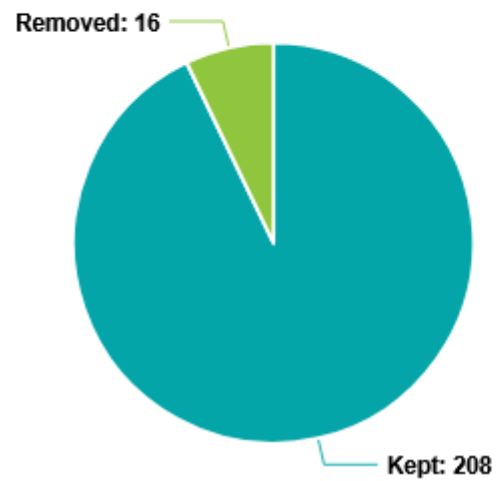
ITS

DADA2

ASVs



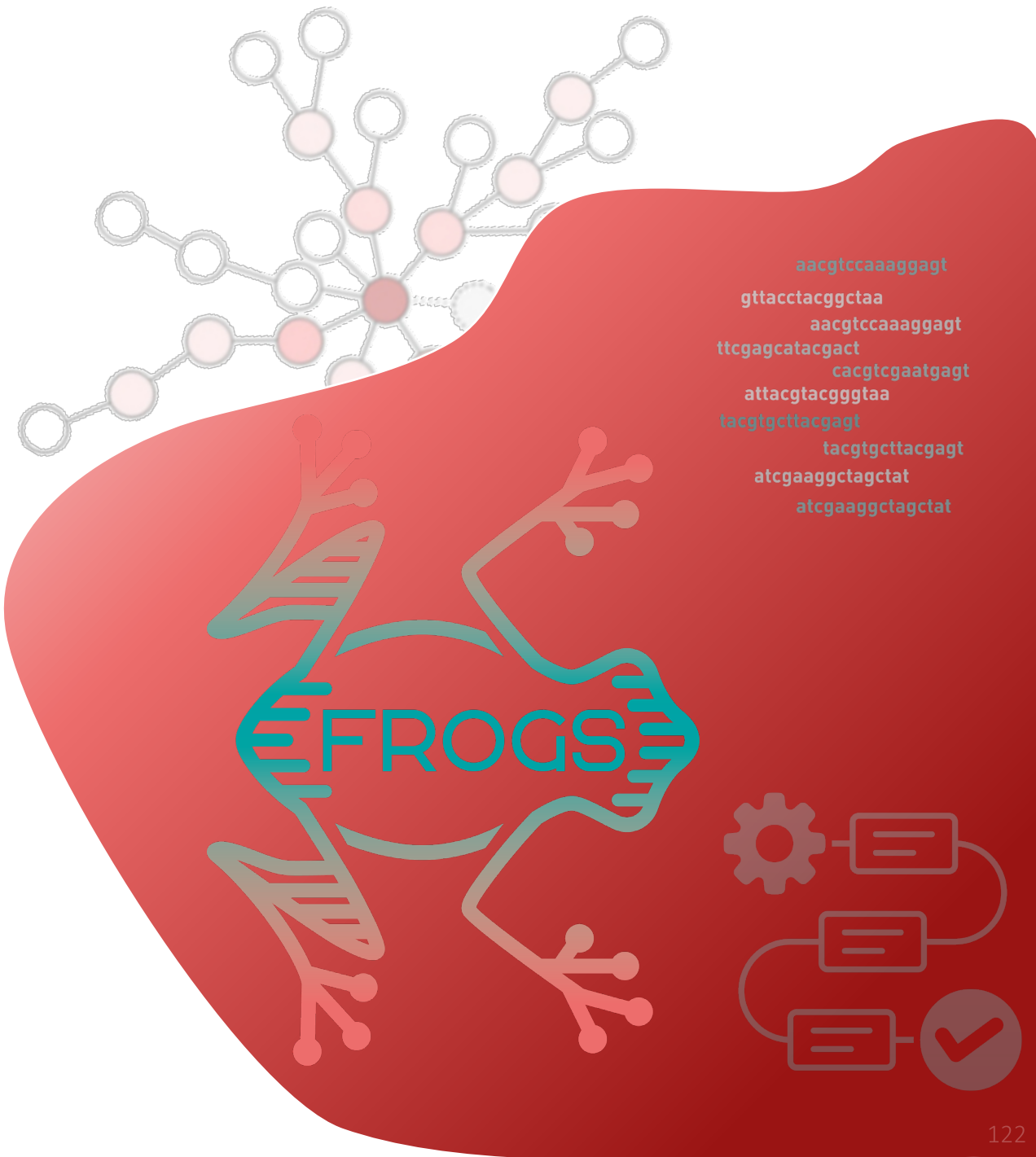
Abundance



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_nuru

blast_subject	blast_perc_identity	blast_perc_query_coverage	blast_perc_subject_coverage	blast_evalue	blast_aln_length
Corynebacterium_nuru	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?

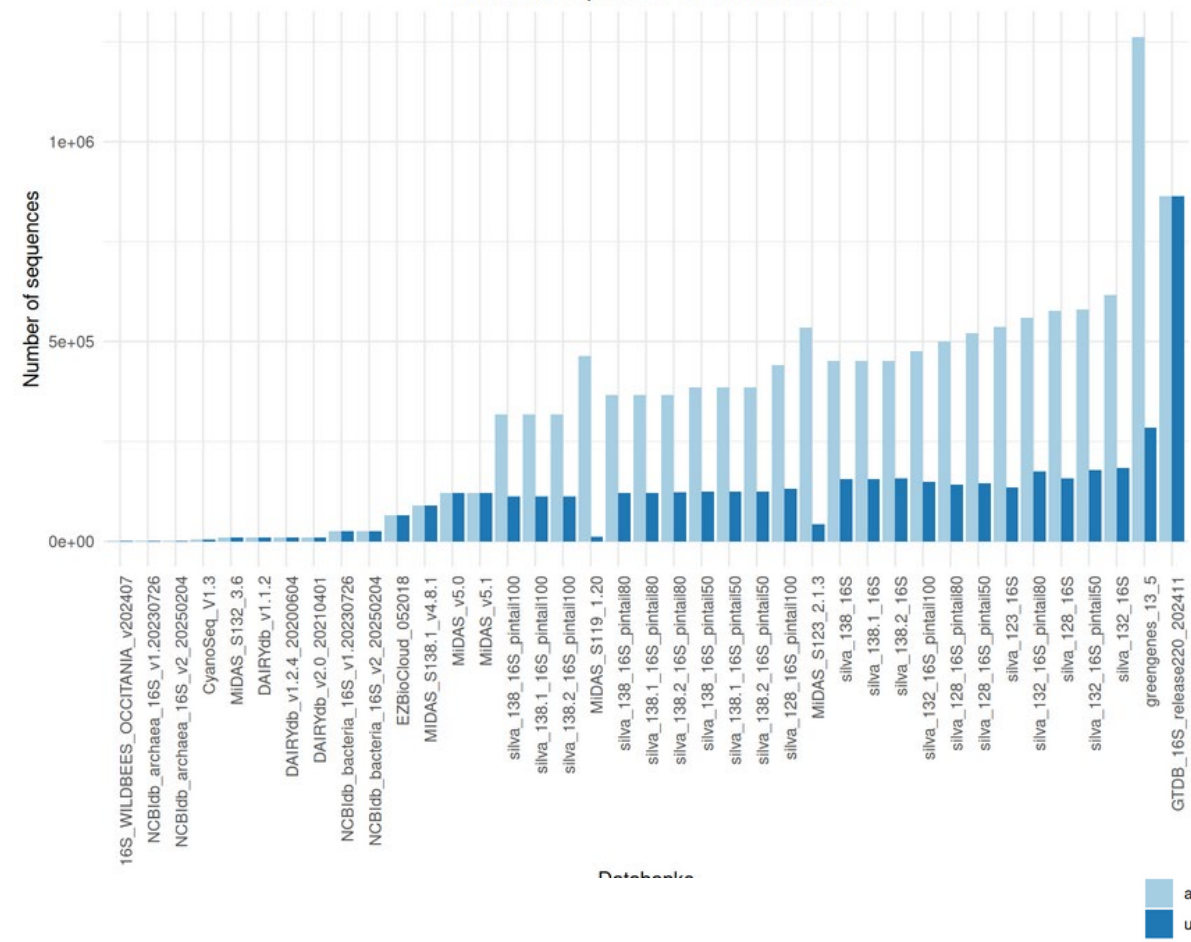
Focus on 16S SILVA and 16S SILVA pintail 100, 80 and 50

Practice session

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

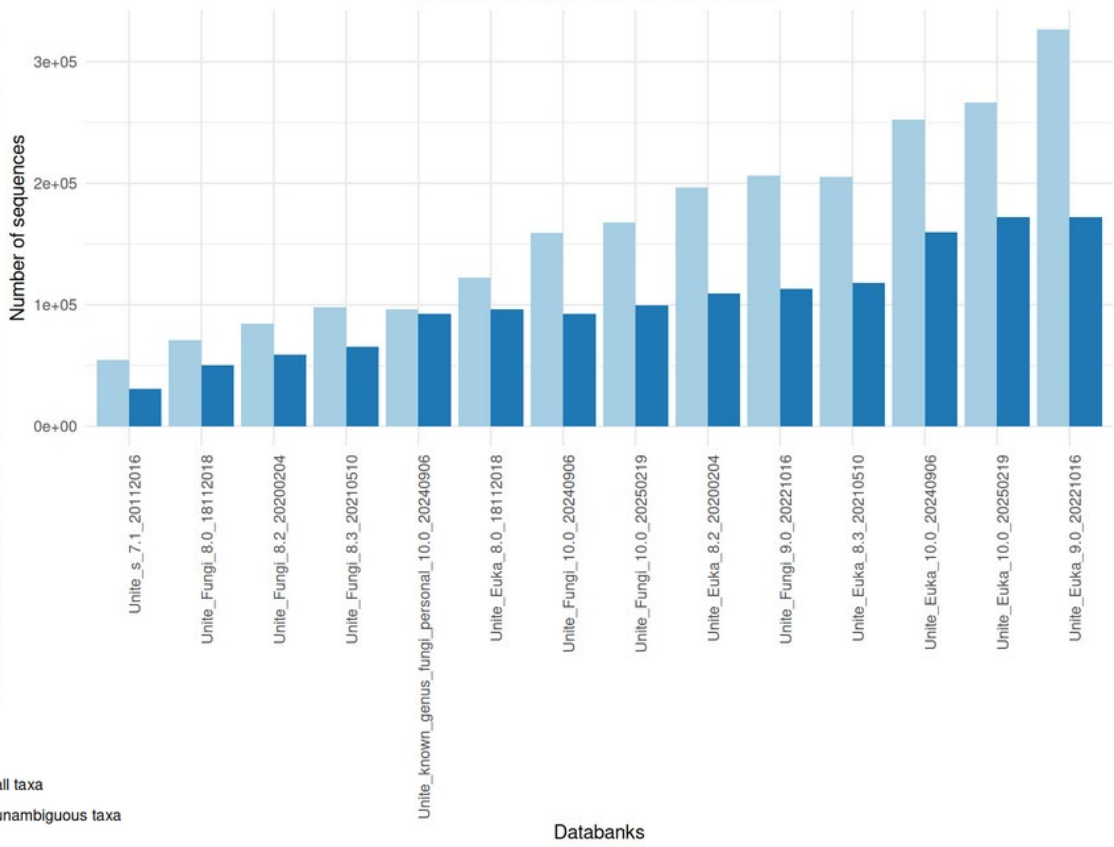
16S

Count of sequences for 16S marker



ITS

Count of sequences for ITS marker



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

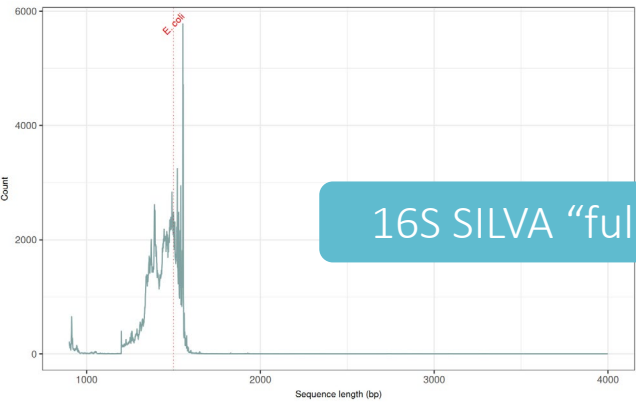


Only for 16S !

* <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

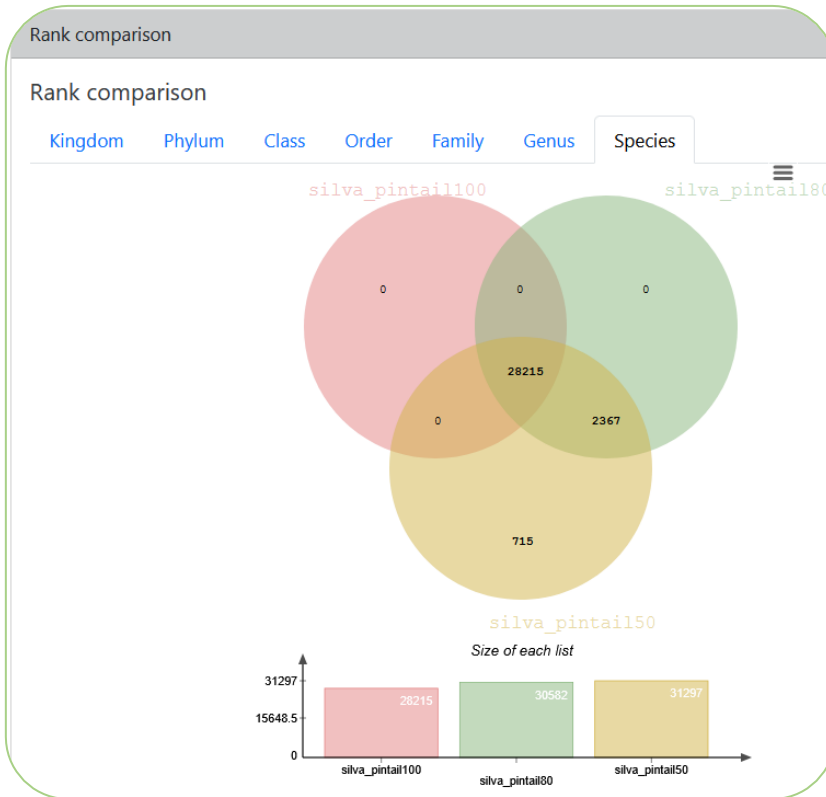
Detail of ambiguous key-words

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 DAIRYdb_v3.0.0_202304 databank.
- For ITS marker: ASVs will be affiliated with the ITS_UNITE_Fungi_10.0_20250219 databank.

Run the process !

Tool Parameters

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

- ☐ -- Please select a tool --
- ☐ 1.a. Short read processing
- ☐ 1.b. Long read processing
- ☐ 1.c. 454 read processing
- ☐ 2. Remove chimera
- ☐ 3. Cluster/ASV filters
- ☒ 4. Taxonomic affiliation
- ☐ 5. Phylogenetic tree building
- ☐ ITSx

Sequence file (.fasta) *



11: (unavailable) FROGS Core 1-Main cluster_filters: sequence.fasta

accepted formats ▼

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

Abundance file (.biom) *



12: (unavailable) 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 *

DAIRYdb v3.0.0_202304

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

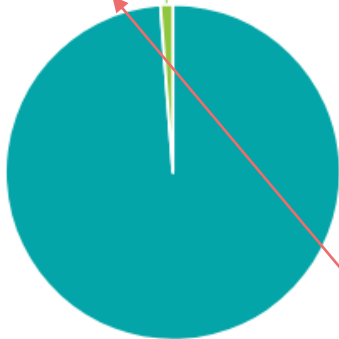
No human readable !

How many ASVs are affiliated by BLAST ?

16S

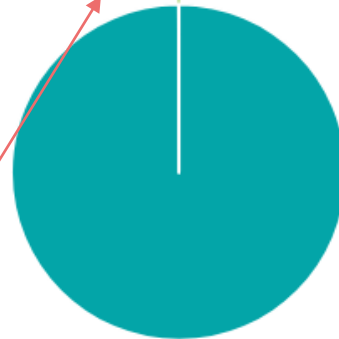
DADA2

Without affiliation: 6



With affiliation: 500

Without affiliation: 514



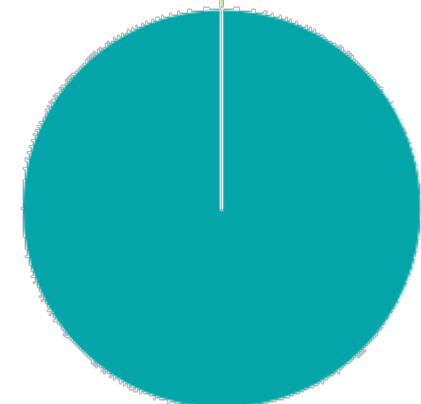
With affiliation: 1 046 476

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

ITS

DADA2

Without affiliation: 0



With affiliation: 208

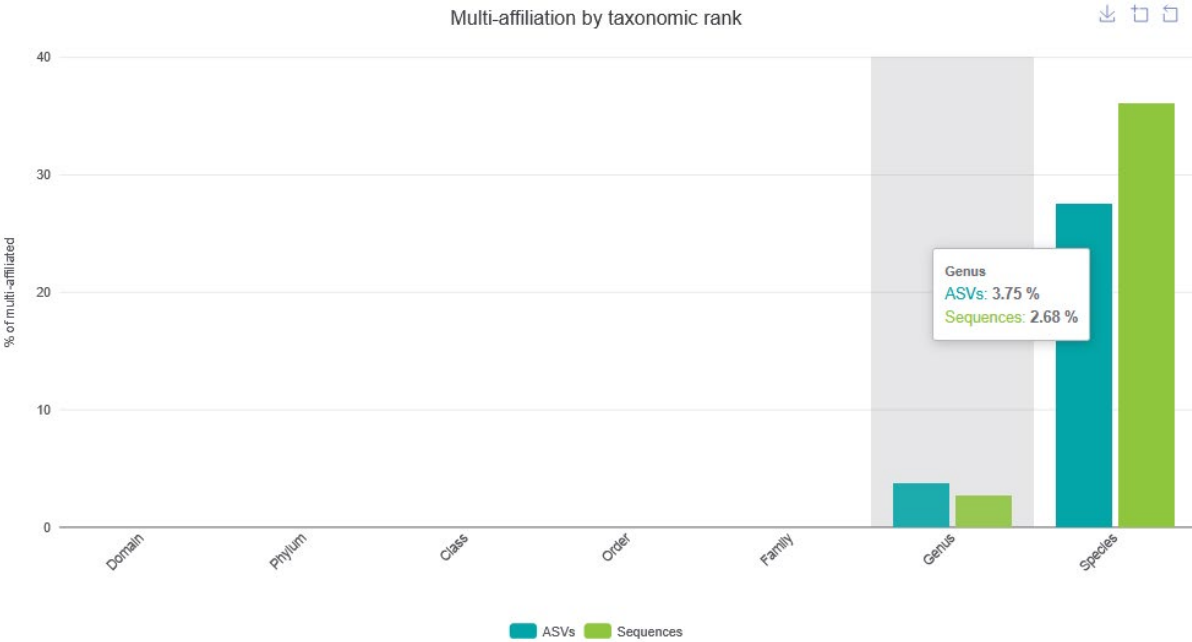
Practice session

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

16S

DADA2

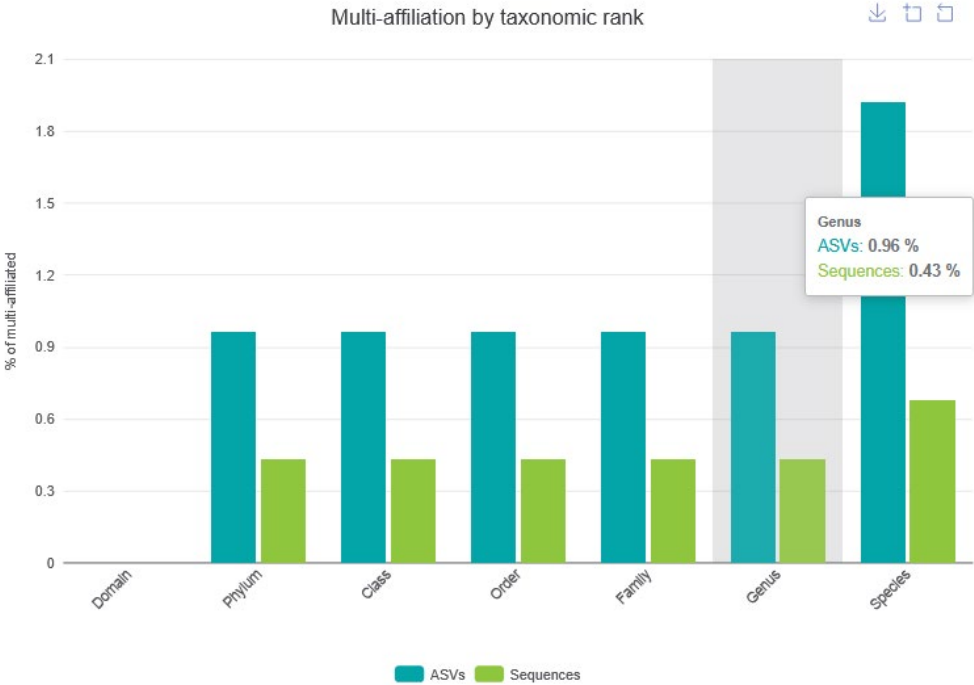
Blast multi-affiliation summary



ITS

DADA2

Blast multi-affiliation summary





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

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

Sequence file (.fasta) - optional

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

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

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

Sequence file (.fasta) - optional

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

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 ID10** (for 16S - focus on AOP1 samples) and **ASV ID7** (for ITS – focus on AOP1 and AOP2 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 DADA2

blast_taxonomy	blast_subject	blast_perc_identity	blast_perc_query_coverage	blast_perc_subject_coverage	blast_evalue	blast_aln_length	seed_id	seed_sequence	observation_name	observation_sum
Bacteria;Actinomycetota;Actinomycetes;Mycobacteriales;Corynebacteriaceae;Corynebacterium;Multi-affiliation	multi-subject	100	100	multi-coverage	0	408	None	TGGGGAATATTGC	ID_1	120201
Bacteria;Bacillota;Bacilli;Lactobacillales;Streptococcaceae;Streptococcus;Streptococcus_thermophilus	multi-subject	100	100	multi-coverage	0	426	None	TAGGGAATCTTCG	ID_2	91844
Bacteria;Bacillota;Bacilli;Lactobacillales;Streptococcaceae;Lactococcus;Multi-affiliation	multi-subject	100	100	multi-coverage	0	426	None	TAGGGAATCTTCG	ID_3	76806
Bacteria;Bacillota;Bacilli;Lactobacillales;Lactobacillaceae;Lactobacillus;Lactobacillus_delbrueckii	multi-subject	100	100	multi-coverage	0	426	None	TAGGGAATCTTCCA	ID_4	69921
Bacteria;Bacillota;Bacilli;Caryophanales;Staphylococcaceae;Staphylococcus;Staphylococcus_equorum	Staphylococcus	100	100	multi-coverage	0	426	None	TAGGGAATCTTCCC	ID_5	26989
Bacteria;Pseudomonadota;Gammaproteobacteria;Oceanospirillales;Halomonadaceae;Halomonas;Halomonas_v	Halomonas_va	100	100	28.51405622	0	426	None	TGGGGAATATTGG	ID_6	24741
Bacteria;Bacillota;Bacilli;Lactobacillales;Leuconostocaceae;Leuconostoc;Leuconostoc_mesenteroides	multi-subject	100	100	multi-coverage	0	426	None	TAGGGAATCTTCCA	ID_7	23851

How is the ASV ID10 distributed across different samples?

16S DADA2

observation_name	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
ID_10	18687	0	0	0	0	0	0	1430	0	0	0	0	0	505	378	475	416	547	446	388



how would do you explain it ?

Practice session

How would do you explain it ?

16S

DADA2

blast_taxonomy	blast_subject	blast_perc_identity	blast_perc_query_coverage
Bacteria;Actinomycetota;Actinomycetes;Mycobacteriales;Corynebacteriaceae;Corynebacterium;Multi-affiliation	multi-subject	100	100
Bacteria;Bacillota;Bacilli;Lactobacillales;Streptococcaceae;Streptococcus;Streptococcus_thermophilus	multi-subject	100	100
Bacteria;Bacillota;Bacilli;Lactobacillales;Streptococcaceae;Lactococcus;Multi-affiliation	multi-subject	100	100
Bacteria;Bacillota;Bacilli;Lactobacillales;Lactobacillaceae;Lactobacillus;Lactobacillus_delbrueckii	multi-subject	100	100
Bacteria;Bacillota;Bacilli;Caryophanales;Staphylococcaceae;Staphylococcus;Staphylococcus_equorum	Staphylococcus	100	100
Bacteria;Pseudomonadota;Gammaproteobacteria;Oceanospirillales;Halomonadaceae;Halomonas;Halomonas_variabilis	Halomonas_va	100	100
Bacteria;Bacillota;Bacilli;Lactobacillales;Leuconostocaceae;Leuconostoc;Leuconostoc_mesenteroides	multi-subject	100	100
Bacteria;Actinomycetota;Actinomycetes;Micrococcales;Brevibacteriaceae;Brevibacterium;Multi-affiliation	multi-subject	multi-identity	100
Bacteria;Actinomycetota;Actinomycetes;Mycobacteriales;Corynebacteriaceae;Corynebacterium;Corynebacterium_casei	Corynebacteriu	100	100
Bacteria;Actinomycetota;Actinomycetes;Micrococcales;Brevibacteriaceae;Brevibacterium;Brevibacterium_aurantiacum	Brevibacterium	99.268	100
Bacteria;Pseudomonadota;Gammaproteobacteria;Pseudomonadales;Moraxellaceae;Psychrobacter;Psychrobacter_pacificensis	Psychrobacter_	97.887	100
Bacteria;Actinomycetota;Actinomycetes;Mycobacteriales;Corynebacteriaceae;Corynebacterium;Corynebacterium_nuruki	Corynebacteriu	99.02	100
Bacteria;Bacillota;Bacilli;Lactobacillales;Streptococcaceae;Streptococcus;Streptococcus_thermophilus	multi-subject	99.765	100
Bacteria;Actinomycetota;Actinomycetes;Mycobacteriales;Corynebacteriaceae;Corynebacterium;Multi-affiliation	multi-subject	99.755	100
Bacteria;Bacillota;Bacilli;Lactobacillales;Leuconostocaceae;Leuconostoc;Leuconostoc_pseudomesenteroides	Leuconostoc_p	99.765	100
Bacteria;Bacillota;Bacilli;Lactobacillales;Carnobacteriaceae;Alkalibacterium;Multi-affiliation	multi-subject	100	100
Bacteria;Actinomycetota;Actinomycetes;Micrococcales;Micrococccaceae;Multi-affiliation;Multi-affiliation	multi-subject	100	100
Bacteria;Pseudomonadota;Gammaproteobacteria;Pseudomonadales;Moraxellaceae;Psychrobacter;Psychrobacter_pacificensis	Psychrobacter_	97.887	100
Bacteria;Bacillota;Bacilli;Lactobacillales;Streptococcaceae;Streptococcus;Streptococcus_thermophilus	multi-subject	99.765	100
Bacteria;Pseudomonadota;Gammaproteobacteria;Pseudomonadales;Moraxellaceae;Psychrobacter;Multi-affiliation	multi-subject	97.887	100
Bacteria;Bacillota;Bacilli;Lactobacillales;Lactobacillaceae;Multi-affiliation;Multi-affiliation	multi-subject	99.296	100
Bacteria;Actinomycetota;Actinomycetes;Micrococcales;Promicromonosporaceae;Cellulosimicrobium;Multi-affiliation	multi-subject	100	100
Bacteria;Pseudomonadota;Gammaproteobacteria;Pseudomonadales;Moraxellaceae;Psychrobacter;Multi-affiliation	multi-subject	98.122	100
Bacteria;Actinomycetota;Actinomycetes;Micrococcales;Dermabacteraceae;Brachybacterium;Brachybacterium_xerosis	Brachybacteriu	100	100
Bacteria;Actinomycetota;Actinomycetes;Micrococcales;Brevibacteriaceae;Brevibacterium;Brevibacterium_aurantiacum	Brevibacterium	99.024	100

Dada2 has identified at least 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.

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
0	0	0	0	0	0	1430	0	0	0	0	0	505	378	475
0	3	0	0	0	0	0	0	0	0	0	7	278	219	313
0	0	0	0	0	0	1030	0	0	0	0	0	2686	2833	2217
1184	870	814	1900	1290	1455	0	182	1245	2030	1410	1057	0	0	0
1299	957	1134	1456	2339	1356	0	828	719	1001	570	781	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	223	0	0	0	0	0	186	197	172
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2208	1686	1349	938	616	948	0	138	1111	445	493	261	0	0	0
0	0	12	0	0	0	0	0	0	0	22	9	22	20	33
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
48	0	6	0	0	76	161	0	2	2	0	0	1	0	0
819	598	776	528	860	514	0	1092	882	929	915	1097	0	0	0

Practice session

Which ASVs are dominant?
What are their affiliations ?

ITS

DADA2

blast_taxonomy	blast_subject	blast_perc_identity	blast_perc_query_coverage	blast_perc_subject_coverage	blast_evalue	blastaln_length	seed_id	seed_sequence	observation_name
k__Fungi;p__Ascomycota;c__Saccharomycetes;o__Saccharomycetales;f__Dipodascaceae;g__Geotrichum;s__Geotrichum_candidum	LC032055_SH13	100	90.73170732	66.66666667	2.86E-94	186	None	GAAACGCGATA	ID_1
k__Fungi;p__Ascomycota;c__Saccharomycetes;o__Saccharomycetales;f__Debaryomycetaceae;g__Debaryomyces;s__Debaryomyces_prosopidis	JN942657_SH12	99.707	100	51.43288084	1.59E-178	341	None	GAAATGCGATA	ID_2
k__Fungi;p__Ascomycota;c__Sordariomycetes;o__Hypocreales;f__Nectriaceae;g__Bisifusarium;s__Bisifusarium_domesticum	JQ434585_SH12	100	87.45980707	57.62711864	7.03E-142	272	None	GAAATGCGATA	ID_3
k__Fungi;p__Ascomycota;c__Saccharomycetes;o__Saccharomycetales;f__Debaryomycetaceae;g__Debaryomyces;s__Debaryomyces_prosopidis	JN942657_SH12	100	100	51.43288084	3.42E-180	341	None	GAAATGCGATA	ID_4
k__Fungi;p__Ascomycota;c__Saccharomycetes;o__Saccharomycetales;f__Debaryomycetaceae;g__Kurtzmaniella;s__Kurtzmaniella_sp	AY821840_SH12	96.951	97.30538922	55.08474576	3.46E-155	328	None	GAAATGCGATA	ID_5
k__Fungi;p__Ascomycota;c__Saccharomycetes;o__Saccharomycetales;f__Saccharomycetales_fam_Incertae_sedis;g__Diutina;s__Diutina_catenuolata	KY103376_SH13	100	83.47826087	61.14649682	1.50E-97	192	None	GAAATGCGATA	ID_6
k__Fungi;p__Ascomycota;c__Sordariomycetes;o__Microascales;f__Microascaceae;g__Scopulariopsis;s__Scopulariopsis_flava	LM652493_SH1	99.668	88.52941176	70.3271028	9.81E-156	301	None	GAAATGCGATA	ID_7

How is the ASV ID7 distributed across AOP6 samples ?

ITS

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	AOP2_PPC_W1	AOP2_PPC_W2	AOP2_PPC_W3	AOP2_PPC_W4
57	0	5	11	12	6	2	15	7	9	0	1	19	22	15	288	297	518	691	591	1123	17
3319	3209	3118	8792	6378	6801	5128	3654	4721	3652	3007	3493	8841	7364	11760	5147	5690	6357	2781	1269	1429	4809
5012	4144	2411	2614	1395	639	15781	6167	3610	7881	9135	12904	9374	9989	9444	10901	10961	10667	25812	18241	26754	17450
269	1048	848	0	0	0	263	280	198	661	608	722	0	0	0	52	98	210	120	61	37	21
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0
5	0	0	3	21	0	0	1	0	1	0	0	0	0	0	3	3	2	0	0	0	0
3723	3950	8058	7489	0	0	11991	0	3144	4651	5382	5780	0	0	0	0	0	0	10700	7222	12571	0



how would do you explain it ?

Practice session

How would do you explain it ?

ITS

DADA2

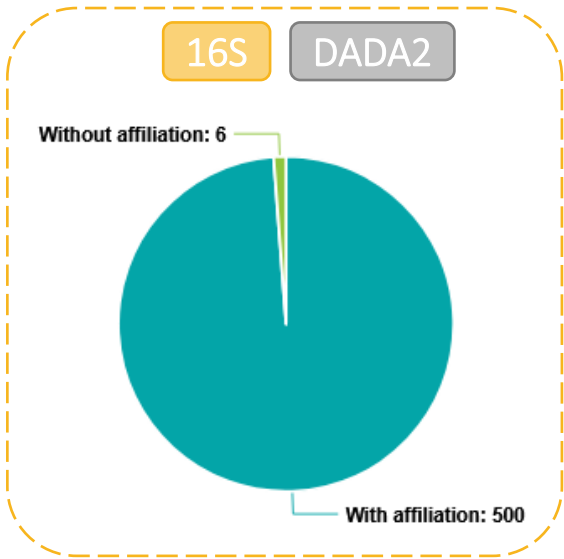
```
blast_taxonomy
k_Fungi;p_Ascomycota;c_Saccharomycetes;o_Saccharomycetales;f_Dipodascaceae;g_Geotrichum;s_Geotrichum_candidum
k_Fungi;p_Ascomycota;c_Saccharomycetes;o_Saccharomycetales;f_Debaryomycetaceae;g_Debaryomyces;s_Debaryomyces_prosopidis
k_Fungi;p_Ascomycota;c_Sordariomycetes;o_Hypocreales;f_Nectriaceae;g_Bisifusarium;s_Bisifusarium_domesticum
k_Fungi;p_Ascomycota;c_Saccharomycetes;o_Saccharomycetales;f_Debaryomycetaceae;g_Debaryomyces;s_Debaryomyces_prosopidis
k_Fungi;p_Ascomycota;c_Saccharomycetes;o_Saccharomycetales;f_Debaryomycetaceae;g_Kurtzmaniella;s_Kurtzmaniella_sp
k_Fungi;p_Ascomycota;c_Saccharomycetes;o_Saccharomycetales;f_Saccharomycetales_fam_Incertae_sedis;g_Diutina;s_Diutina_catnulata
k_Fungi;p_Ascomycota;c_Sordariomycetes;o_Microascales;f_Microascaceae;g_Scopulariopsis;s_Scopulariopsis_flava
k_Fungi;p_Ascomycota;c_Sordariomycetes;o_Microascales;f_Microascaceae;g_Scopulariopsis;s_Scopulariopsis_flava
```

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.

observation_name	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
ID_1	534925	57	0	5	11	12	6	2	15	7	9	0	1	19	22	15	288	297	518
ID_2	391654	3319	3209	3118	8792	6378	6801	5128	3654	4721	3652	3007	3493	8841	7364	11760	5147	5690	6357
ID_3	252335	5012	4144	2411	2614	1395	639	15781	6167	3610	7881	9135	12904	9374	9989	9444	10901	10961	10667
ID_4	158482	269	1048	848	0	0	0	263	280	198	661	608	722	0	0	0	52	98	210
ID_5	97286	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
ID_6	92604	5	0	0	3	21	0	0	1	0	1	0	0	0	0	0	3	3	2
ID_7	84661	3723	3950	8058	7489	0	0	11991	0	3144	4651	5382	5780	0	0	0	0	0	0
ID_8	52207	0	0	0	0	10011	11612	0	3522	0	0	0	0	3873	2805	4178	4019	2525	3145

Practice session

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



16S DADA2

Bacteria;Bacillota;Bacilli;Lactobacillales;Leuconostocaceae;Leuconostoc;Leuconostoc_mesenteroides	multi-subject	98.592	100	multi-coverage	0	426	None	TAGGGAATCTTCC	ID_459
Bacteria;Pseudomonadota;Gammaproteobacteria;Alteromonadales;Celerinatantimonadaceae;Celerinatantimonas;Celerinat	Celerinatantim	96.479	100	28.53315472	0	426	None	TGGGGAATATTGC	ID_460
no data	no data	no data	no data	no data	no data	no data	None	GCAATGCATGGAG	ID_462
Bacteria;Actinomycetota;Actinomycetes;Micrococcales;Brevibacteriaceae;Brevibacterium;Brevibacterium_jeikeali	Brevibacterium	99.268	100	26.78454486	0	410	None	TGGGGAATATTGC	ID_464

'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.



What is the distinctive feature of the **ASV ID1** (for 16S) and **ASV ID19** (for ITS) 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 ID1 (for 16S) and ASV ID19 (for ITS) affiliations ?

'multi-affiliation.tsv' files analysis

16S

DADA2

#observation_name	blast_taxonomy	blast_subject
ID_1	Bacteria;Actinomycetota;Actinomycetes;Mycobacteriales;Corynebacteriaceae;Corynebacterium;Corynebacterium_variabale	Corynebacterium_variabale_FAUH01000036_TS
ID_1	Bacteria;Actinomycetota;Actinomycetes;Mycobacteriales;Corynebacteriaceae;Corynebacterium;Corynebacterium_Species	Corynebacterium_Species_AJ222817

ITS

DADA2

#observation_name	blast_taxonomy
ID_19	k_Fungi;p_Ascomycota;c_Saccharomycetes;o_Saccharomycetales;f_Metschnikowiaceae;g_Clavispora;s_Clavispora_lusitaniae
ID_19	k_Fungi;p_Ascomycota;c_Saccharomycetes;o_Saccharomycetales;f_Metschnikowiaceae;g_Clavispora;s_Clavispora_lusitaniae
ID_19	k_Fungi;p_Ascomycota;c_Saccharomycetes;o_Saccharomycetales;f_Metschnikowiaceae;g_Clavispora;s_Clavispora_lusitaniae
ID_19	k_Fungi;p_Ascomycota;c_Saccharomycetes;o_Saccharomycetales;f_Metschnikowiaceae;g_Clavispora;s_Clavispora_lusitaniae
ID_19	k_Fungi;p_Ascomycota;c_Saccharomycetes;o_Saccharomycetales;f_Metschnikowiaceae;g_Clavispora;s_Clavispora_lusitaniae
ID_19	k_Fungi;p_Ascomycota;c_Saccharomycetes;o_Saccharomycetales;f_Metschnikowiaceae;g_Clavispora;s_Clavispora_lusitaniae
ID_19	k_Fungi;p_Ascomycota;c_Saccharomycetes;o_Saccharomycetales;f_Metschnikowiaceae;g_Clavispora;s_Clavispora_lusitaniae
ID_19	k_Fungi;p_Ascomycota;c_Saccharomycetes;o_Saccharomycetales;f_Saccharomycetales_fam_Incertae_sedis;g_Saccharomycetales_gen_Incertae_sedis;s_Saccharomycetales_sp
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

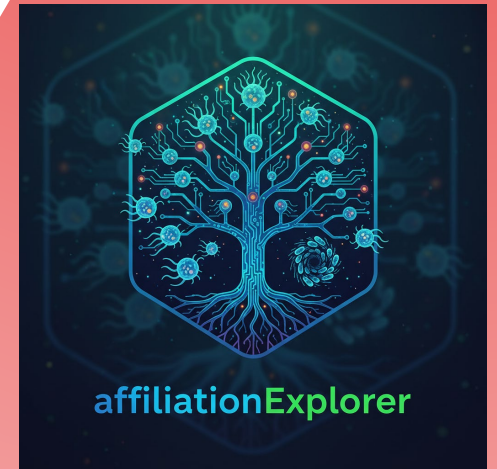
To easily modify your abundance files, you should use the 'affiliationExplorer' tool.

affiliationExplorer is a companion tool of FROGS.

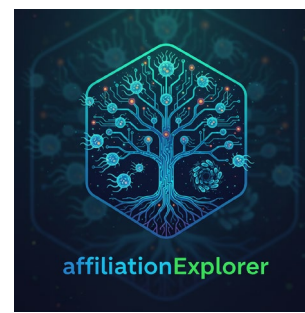
- Shiny app for exploration and disambiguation of FROGS multi-affiliation.
- The goal of affiliationExplorer is to provide a user-friendly graphical interface to choose between different conflicting affiliations.
- The app will then sort the multi-affiliated taxa from most abundant to least abundant 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



Affiliation explorer

Upload Biom File
Browse... Galaxy37-[f]
Upload complete

Optional: upload Fasta File
Browse... Galaxy32-[f]
Upload complete

Upload MultiHits TSV File
Browse... Galaxy42-[f]
Upload complete

Download

Affiliation selection Affiliation edition

Select OTU
Cluster_3 ▼

Update OTU Skip OTU

Cluster_3 - 2 conflicting affiliations, ambiguity at rank Species
Select new affiliation by clicking on a row (double click on a cell to edit its content).
Click "Update OTU" to update affiliation (with selected row) or "Skip OTU" to move to the next one.

Show 10 ▼ entries

	Kingdom	Phylum	Class	Order	Family	Genus	Species	Blast ID	%id	%cov
1	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Latilactobacillus	Lactobacillus sakei	CP032640.225274.226851	100	100
2	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Latilactobacillus	unknown species	KF601977.1.1550	100	100

Showing 1 to 2 of 2 entries

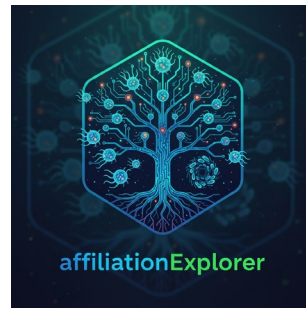
Search:

Previous 1 Next

☐ Show sequence

A very user-friendly tool, developed by Mahendra Mariadassou and his collaborators (Maïage 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	Number of Family	Number of Genus
☐	AOP1_PPC_S1	1	3	4	9	19	30
☐	AOP1_PPC_S2	1	4	5	11	20	32
☐	AOP1_PPC_S3	1	3	4	9	18	28
☐	AOP1_PPC_S4	1	4	5	9	21	32
☐	AOP1_PPC_S5	1	4	6	11	24	35
☐	AOP1_PPC_S6	1	4	5	9	21	29
☐	AOP1_PPC_W1	1	4	7	11	18	30
☐	AOP1_PPC_W2	1	4	6	10	19	30
☐	AOP1_PPC_W3	1	3	4	9	18	26
☐	AOP1_PPC_W4	1	4	6	12	22	31

Showing 1 to 10 of 72 rows 10 rows per page

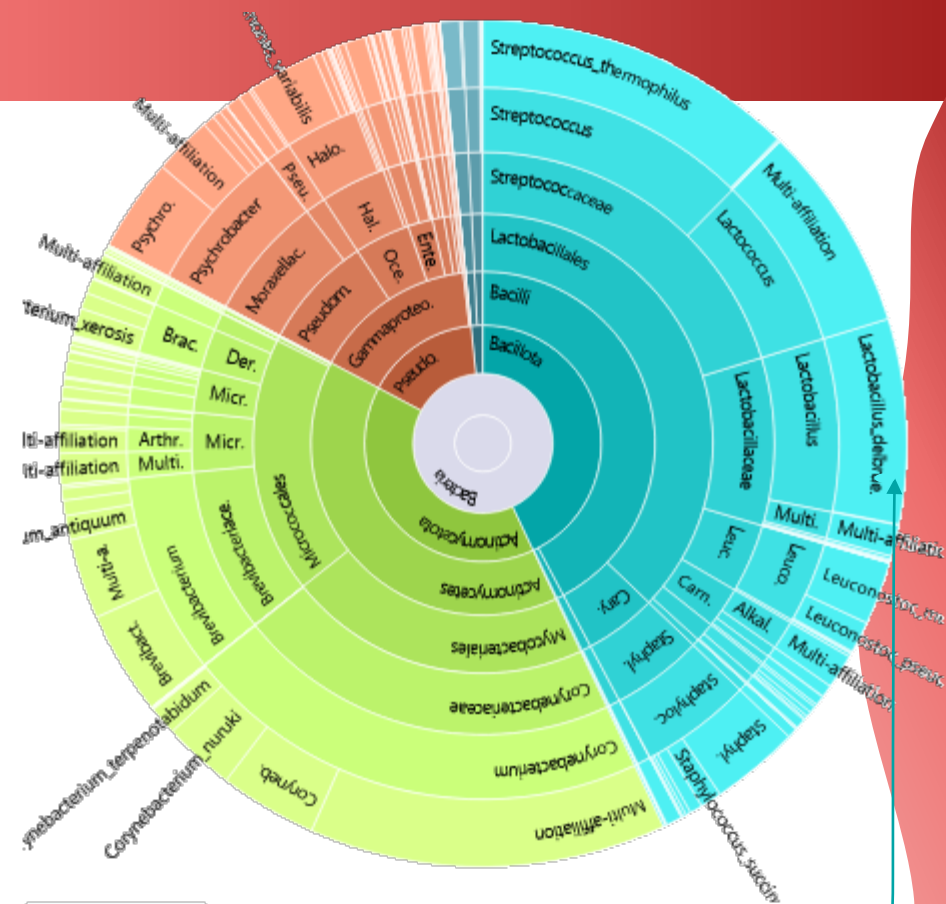
< 1 2 3 4 5 ... 8 >

With selection:

Display distribution

Display rarefaction

Genus



Detail on selected:

Name	Size	Global %	Parent %
root	1046476		
Bacteria	1046476	100.000	100.000
Bacillota	449934	42.995	42.995
Bacilli	442022	42.239	98.242
Lactobacillales	388957	37.168	87.995
Lactobacillaceae	97436	9.311	25.051
Lactobacillus	82185	7.854	84.348
Lactobacillus_delbrueckii	81739	7.811	99.457

Lactobacillus_delbrueckii nb children: 0

Font size: 15

Colors start depth: 2

Close

Practice session

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

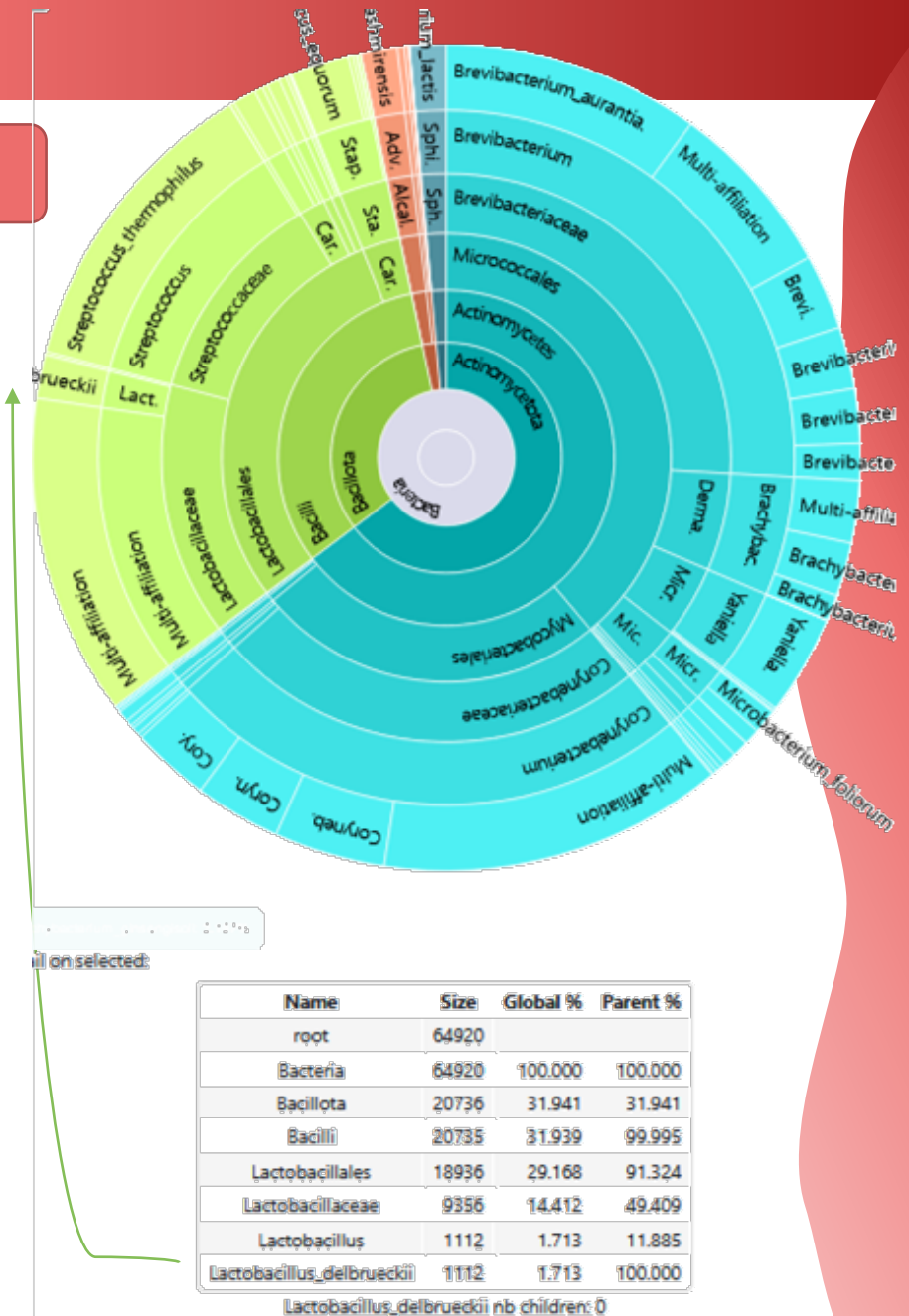
Summary Taxonomy Distribution Blast Alignment metrics Distribution

☐	Sample name	Number of Domain	Number of Phylum	Nur
☒	AOP1_PPC_S1	1	3	4
☒	AOP1_PPC_S2	1	4	5
☒	AOP1_PPC_S3	1	3	4
☒	AOP1_PPC_S4	1	4	5
☒	AOP1_PPC_S5	1	4	6
☒	AOP1_PPC_S6	1	4	5
☐	AOP1_PPC_W1	1	4	7
☐	AOP1_PPC_W2	1	4	6
☐	AOP1_PPC_W3	1	3	4
☐	AOP1_PPC_W4	1	4	6

Showing 1 to 10 of 72 rows 10 rows per page

With selection:

Genus



Practice session

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

Summary Taxonomy Distribution Blast Alignment metrics Distribution

Display global distribution

☐	Sample name	Number of Domain	Number of Phylum	Number of Class	Number of Order	Number of Species
☒	AOP1_PPC_S1	1	2	5	8	14
☒	AOP1_PPC_S2	1	2	5	7	11
☒	AOP1_PPC_S3	1	2	5	8	14
☒	AOP1_PPC_S4	1	1	4	6	12
☒	AOP1_PPC_S5	1	3	5	7	13
☒	AOP1_PPC_S6	1	1	3	5	12
☐	AOP1_PPC_W1	1	2	5	8	14
☐	AOP1_PPC_W2	1	2	5	7	12
☐	AOP1_PPC_W3	1	3	6	8	14
☐	AOP1_PPC_W4	1	2	5	7	12

Showing 1 to 10 of 72 rows 10 rows per page

With selection:

Display distribution

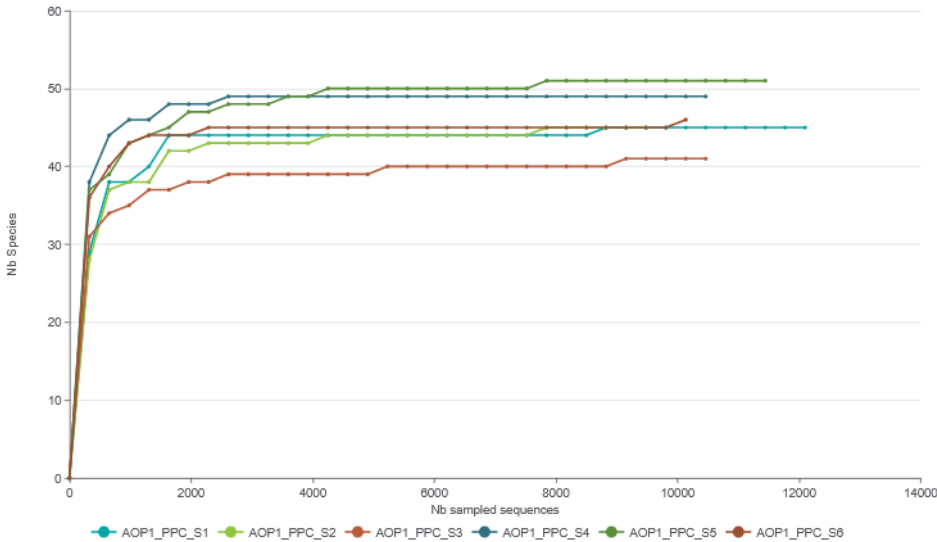
Display rarefaction

Species

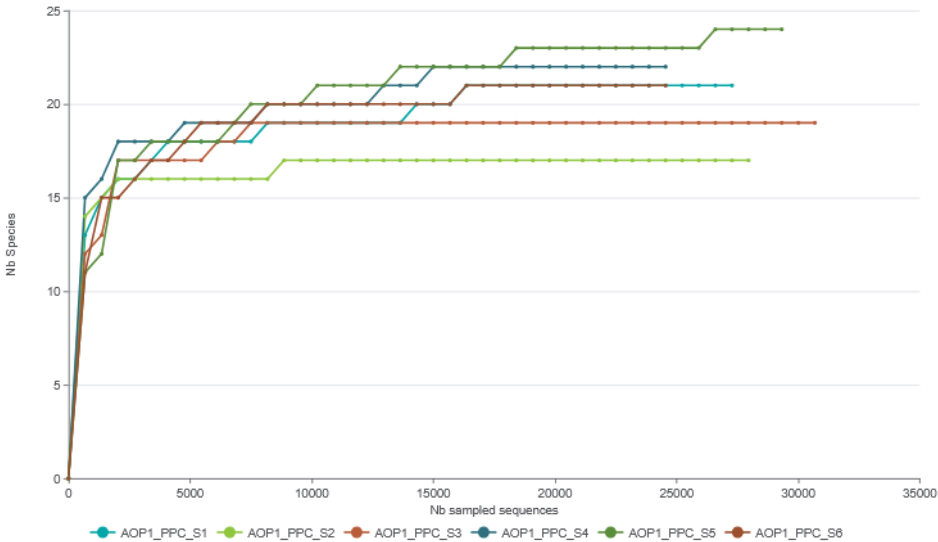


16S
DADA2

Rarefaction curves



Rarefaction curves



ITS
DADA2

Practice session

Explore Blast alignment metrics Distribution tab.

16S

DADA2

Summary Taxonomy Distribution Blast Alignment metrics Distribution

Number of ASVs by BLAST identity and coverage

Coverage	[100%]	[99% - 100%[	[95% - 99%[	[90% - 95%[	[80% - 90%[	[50% - 80%[	[1% - 50%[	[0% - 1%[
	0	0	0	0	2	153	248	80
	0	0	0	0	1	7	3	0
	0	0	0	0	0	1	4	1
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
Identity	[0% - 1%[	[1% - 50%[	[50% - 80%[	[80% - 90%[	[90% - 95%[	[95% - 99%[	[99% - 100%[	[100%]
	6	0	0	0	0	0	0	0

By ASV By sequence

Number of sequences by BLAST identity and coverage

Coverage	[100%]	[99% - 100%[	[95% - 99%[	[90% - 95%[	[80% - 90%[	[50% - 80%[	[1% - 50%[	[0% - 1%[
	0	0	0	0	145	138393	316853	588074
	0	0	0	0	126	1098	718	0
	0	0	0	0	0	105	872	152
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0
Identity	[0% - 1%[	[1% - 50%[	[50% - 80%[	[80% - 90%[	[90% - 95%[	[95% - 99%[	[99% - 100%[	[100%]
	514	0	0	0	0	0	0	0

By ASV By sequence

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

Practice session

Explore Blast alignment metrics Distribution tab.

ITS
DADA2

Summary Taxonomy Distribution Blast Alignment metrics Distribution

Number of ASVs by BLAST identity and coverage

Coverage	[100%]	[90% - 100%]	[95% - 99%]	[90% - 95%]	[80% - 90%]	[50% - 80%]	[1% - 50%]	[0% - 1%]
[100%]	0	1	4	1	6	29	18	4
[90% - 100%]	0	0	0	0	0	0	0	0
[95% - 99%]	0	0	0	0	2	11	0	0
[90% - 95%]	0	0	0	0	0	17	12	11
[80% - 90%]	0	0	0	0	5	27	26	26
[50% - 80%]	0	0	0	0	0	5	1	0
[1% - 50%]	0	0	0	0	2	1	0	0
[0% - 1%]	0	0	0	0	0	0	0	0
Identity								
[0% - 1%] [1% - 50%] [50% - 80%] [80% - 90%] [90% - 95%] [95% - 99%] [99% - 100%] [100%]								

By ASV By sequence

Number of sequences by BLAST identity and coverage

Coverage	[100%]	[90% - 100%]	[95% - 99%]	[90% - 95%]	[80% - 90%]	[50% - 80%]	[1% - 50%]	[0% - 1%]
[100%]	0	578	1123	141	1064	25393	456694	196811
[90% - 100%]	0	0	0	0	0	0	0	0
[95% - 99%]	0	0	0	0	243	107494	0	0
[90% - 95%]	0	0	0	0	0	8263	23155	922878
[80% - 90%]	0	0	0	0	2611	40275	105308	463450
[50% - 80%]	0	0	0	0	0	3467	124	0
[1% - 50%]	0	0	0	0	2090	328	0	0
[0% - 1%]	0	0	0	0	0	0	0	0
Identity								
[0% - 1%] [1% - 50%] [50% - 80%] [80% - 90%] [90% - 95%] [95% - 99%] [99% - 100%] [100%]								

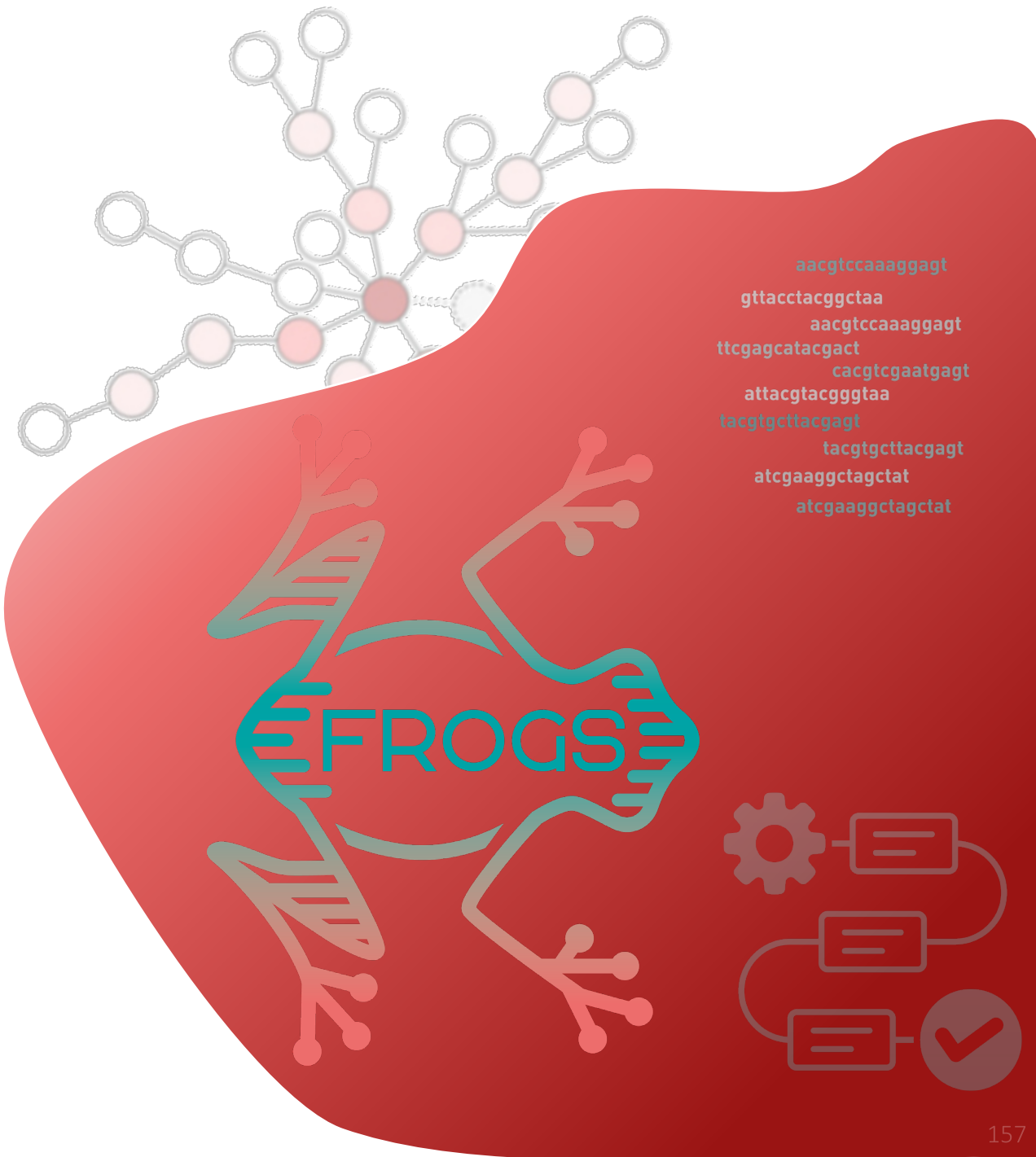
By ASV By sequence

We have found that a large proportion of the data does not appear **exactly** as it is 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

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

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

Outputs :

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

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

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

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

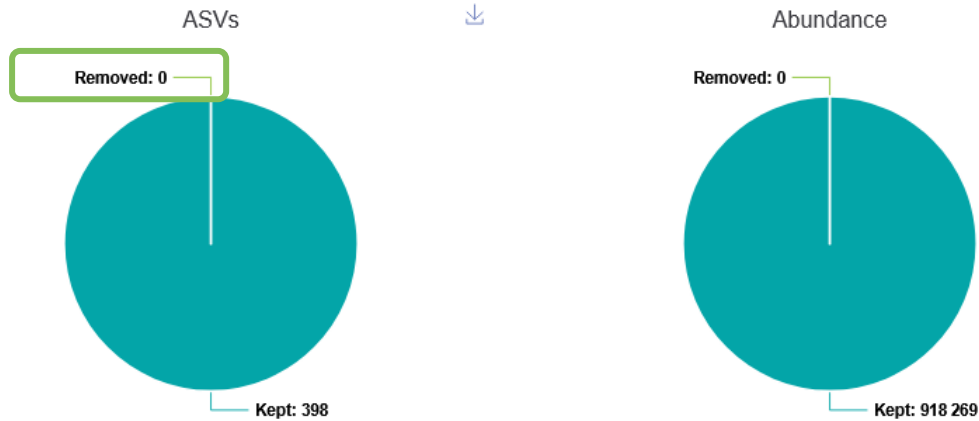
46: FROGS Core 1-Main - tree: report.html

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.

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.

16S

ITS



Our tree is in NHX (= nwk) format.

```
(((((ID_299:0.351858451,(((ID_48:0.002219021,(ID_37:5e-09,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_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.

