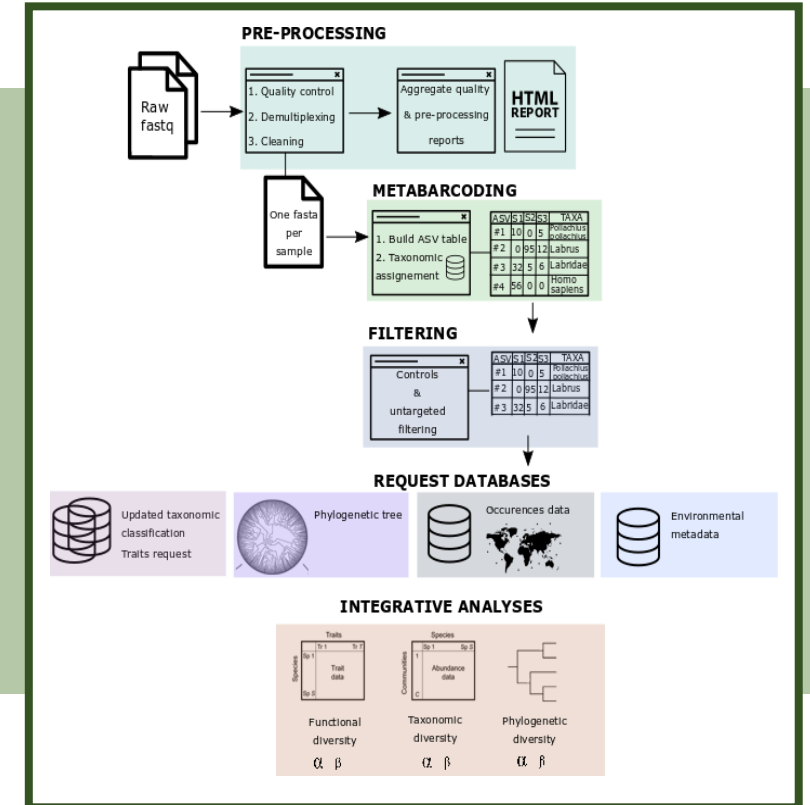
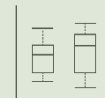


# Faciliter l'intégration des différentes facettes de la biodiversité à partir de données metabarcoding



Anaïs

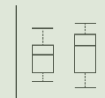
Tuteur : Erwan Corre



## Metabarcoding

- ❑ De plus en plus utilisé en écologie
- ❑ Développement de pipelines de traitement
  - ✓ nombreuses
  - ✓ diverses en fonction des utilisateurs





## Metabarcoding

- ☐ De plus en plus utilisé en écologie
- ☐ Développement de pipelines de traitement
  - ✓ nombreuses
  - ✓ diverses en fonction des utilisateurs



**DADA<sup>2</sup>**  
Amplicon Sequencing. Exactly. Version 1.14



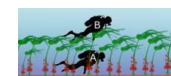
## Questionnement écologique

- ☐ Comprendre l'apport de l'ADN environnemental (ADNe) metabarcoding pour l'analyse de la diversité des poissons côtiers
- ☐ Quelle(s) facette(s) de diversité prendre en compte?

**ADNe** **Comptage en**



**plongée**



**Taxinomique**

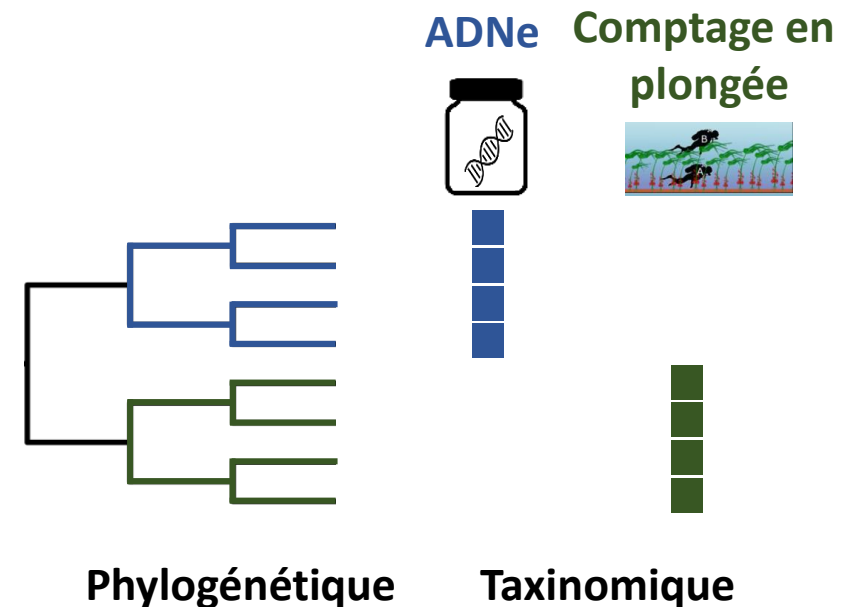
## Metabarcoding

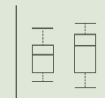
- ☐ De plus en plus utilisé en écologie
- ☐ Développement de pipelines de traitement
  - ✓ nombreuses
  - ✓ diverses en fonction des utilisateurs



## Questionnement écologique

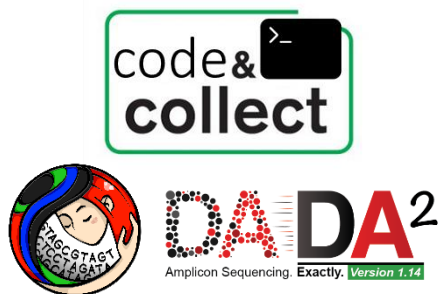
- ☐ Comprendre l'apport de l'ADN environnemental (ADNe) metabarcoding pour l'analyse de la diversité des poissons côtiers
- ☐ Quelle(s) facette(s) de diversité prendre en compte?





## Metabarcoding

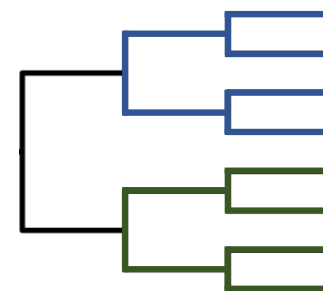
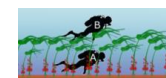
- ☐ De plus en plus utilisé en écologie
- ☐ Développement de pipelines de traitement
  - ✓ nombreuses
  - ✓ diverses en fonction des utilisateurs



## Questionnement écologique

- ☐ Comprendre l'apport de l'ADN environnemental (ADNe) metabarcoding pour l'analyse de la diversité des poissons côtiers
- ☐ Quelle(s) facette(s) de diversité prendre en compte?

ADNe Comptage en plongée



Phylogénétique



Taxinomique



Fonctionnelle



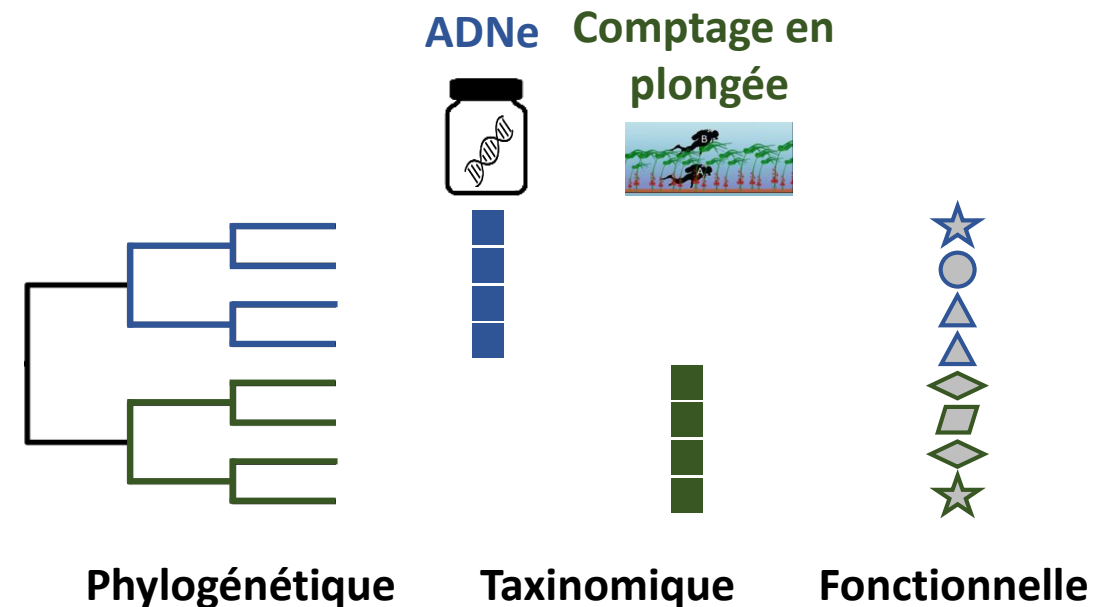
## Metabarcoding

- ☐ De plus en plus utilisé en écologie
- ☐ Développement de pipelines de traitement
  - ✓ nombreuses
  - ✓ diverses en fonction des utilisateurs

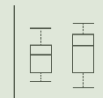


## Questionnement écologique

- ☐ Comprendre l'apport de l'ADN environnemental (ADNe) metabarcoding pour l'analyse de la diversité des poissons côtiers
- ☐ Quelle(s) facette(s) de diversité prendre en compte?



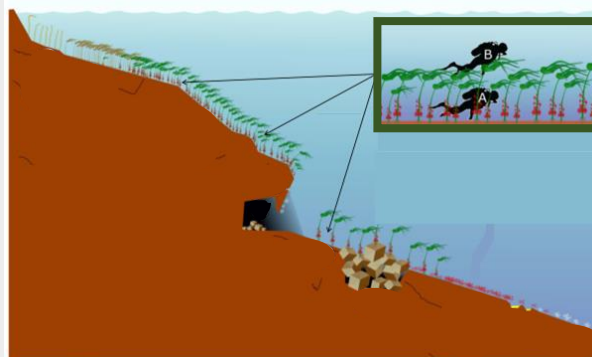
**Intégrer l'analyse des différentes facettes de diversité pour une meilleure compréhension de la biodiversité issue des données de metabarcoding**



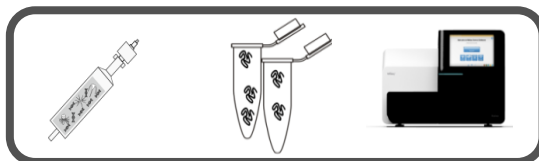
## Poissons côtiers

### ☐ Prélèvement d'eau

- 4 sites au sein d'une baie
- 2 saisons
- 3 méthodes d'échantillonnage



☐ Comparaison avec comptages visuels en plongée

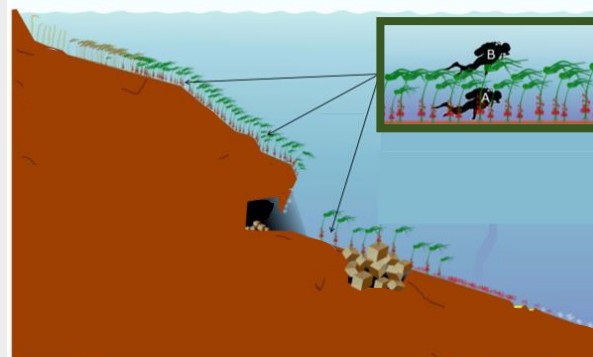




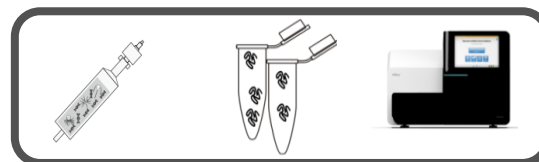
## Poissons côtiers

### ☐ Prélèvement d'eau

- 4 sites au sein d'une baie
- 2 saisons
- 3 méthodes d'échantillonnage



- ☐ Comparaison avec comptages visuels en plongée



## Données du projet

- ☐ Illumina HISEQ et MISEQ

- ☐ Blancs de séquençage & contrôles négatifs

- ☐ Métadonnées

- Échantillonnage : *site, saison, méthode ...*
- Metabarcoding : *librairies, type/run/lane de séquençage ...*

Marker	Lenght (bp)	PCR	Samples	Raw reads
12S	64	12	1536	370
16S	114	4	384	millions





- ☐ En cours de rédaction - Modèle INRAE
- ☐ Réflexion sur l'intégration du projet global FISHDNA
- ☐ Besoin de réunion avec tous les acteurs du projet (MNHN, Station de Roscoff/ABiMS) & prestataires
- ☐ Stockage des données metabarcoding
  - Court terme : ABiMS
  - Long terme : MNHN?



Muséum National d'Histoire Naturelle (MNHN)

## FacetteDivMetaB - Intégration des facettes de diversité à partir de donnée metabarcoding

Renseignements sur le projet   Produits de recherche   Modèle choisi   **Rédiger**   Partager   Télécharger

tout développer | tout réduire

Information concerning the management plan ( 4 questions )

+

Information on the research project ( 10 questions )

+

Brief presentation of project data ( 1 question )

+

Intellectual property rights ( 2 questions )

+

Confidentiality ( 3 questions )

+

Access and sharing of data at the end of the project ( 10 questions )

+

Description and organisation of data ( 6 questions )

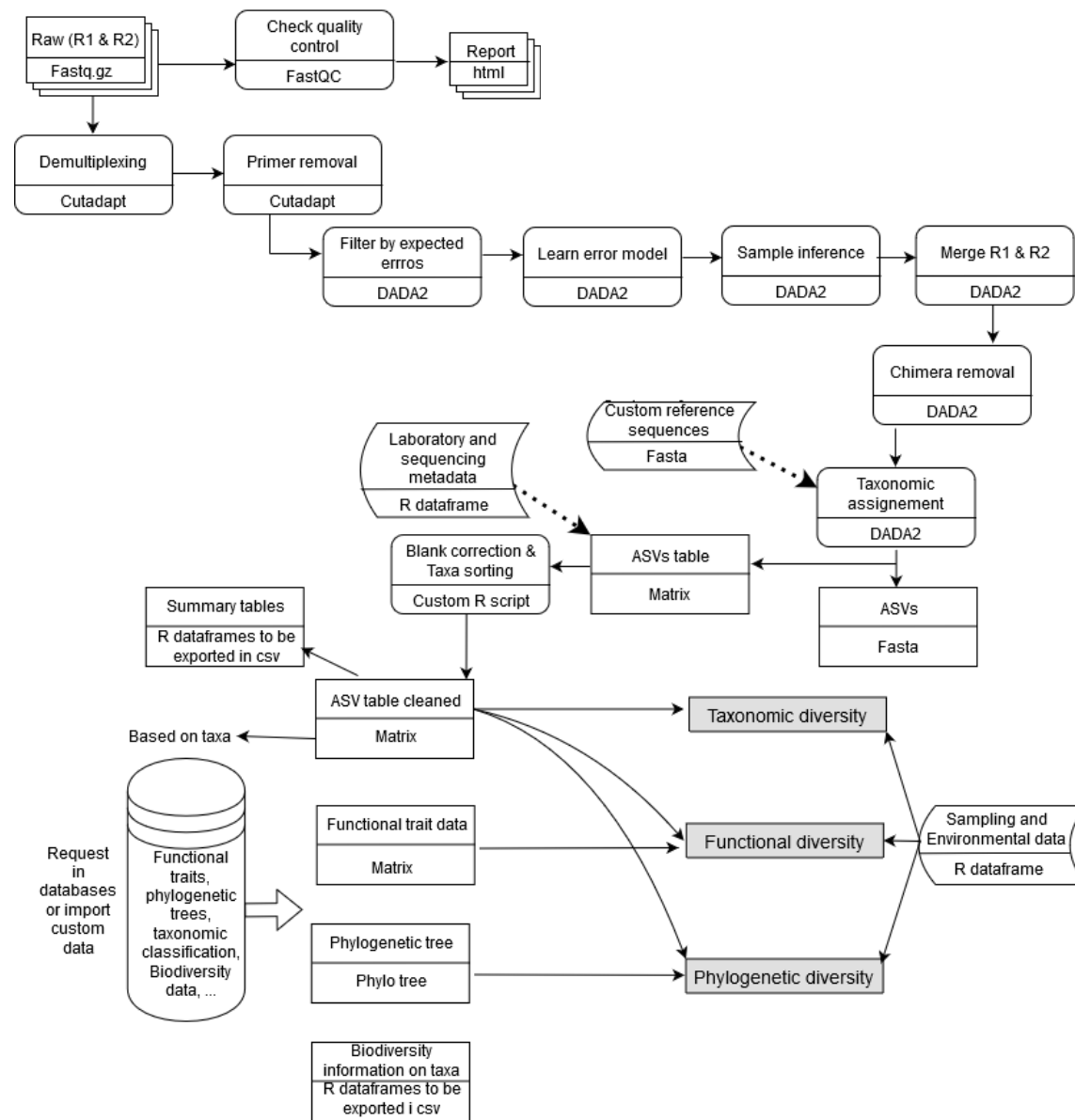
+

Data storage and backup during the project ( 10 questions )

+

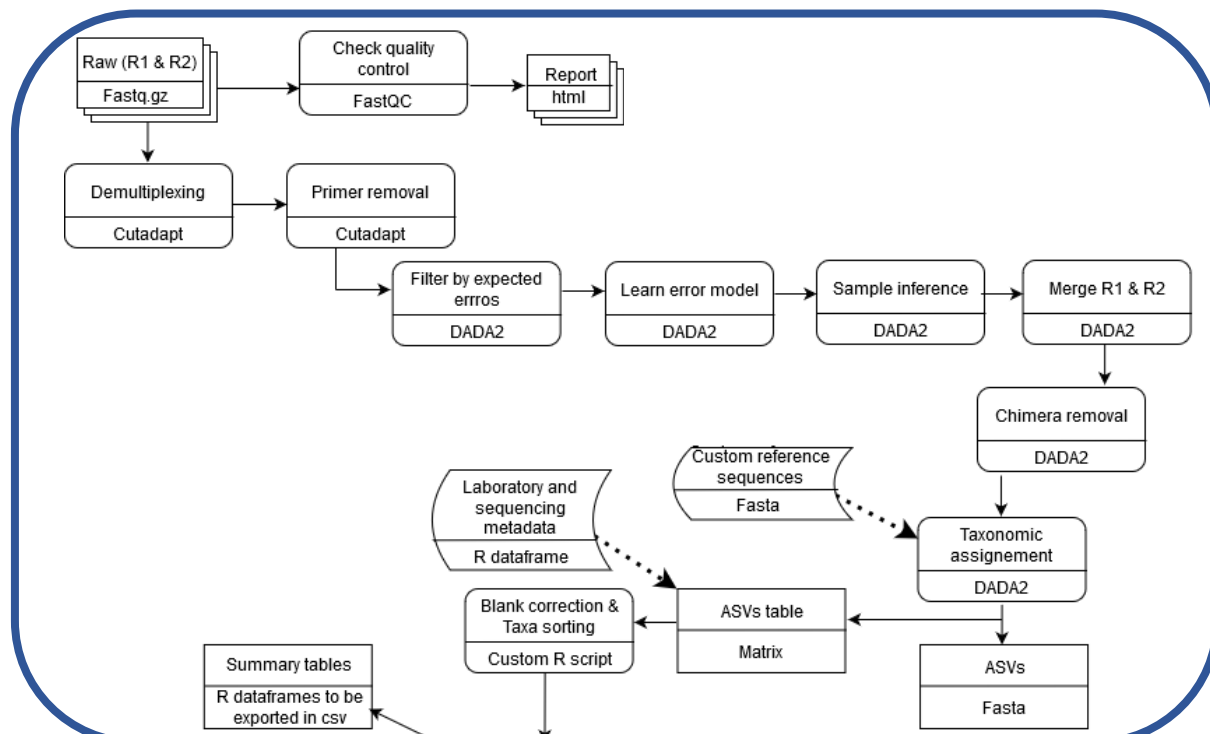
Data archiving and conservation after the end of the project ( 6 questions )

+



## METABARCODING

Fait avant le projet tutoré



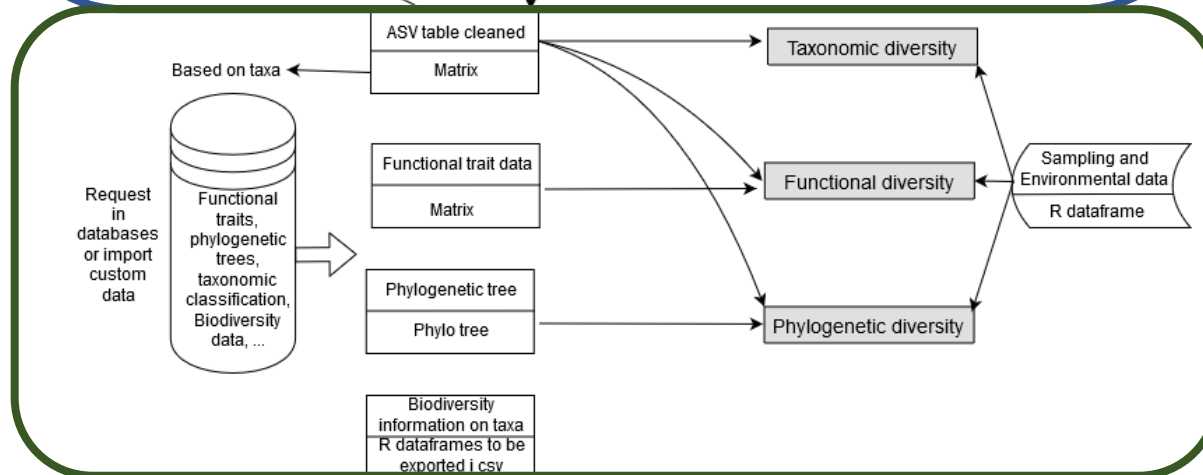
# BASH

# R

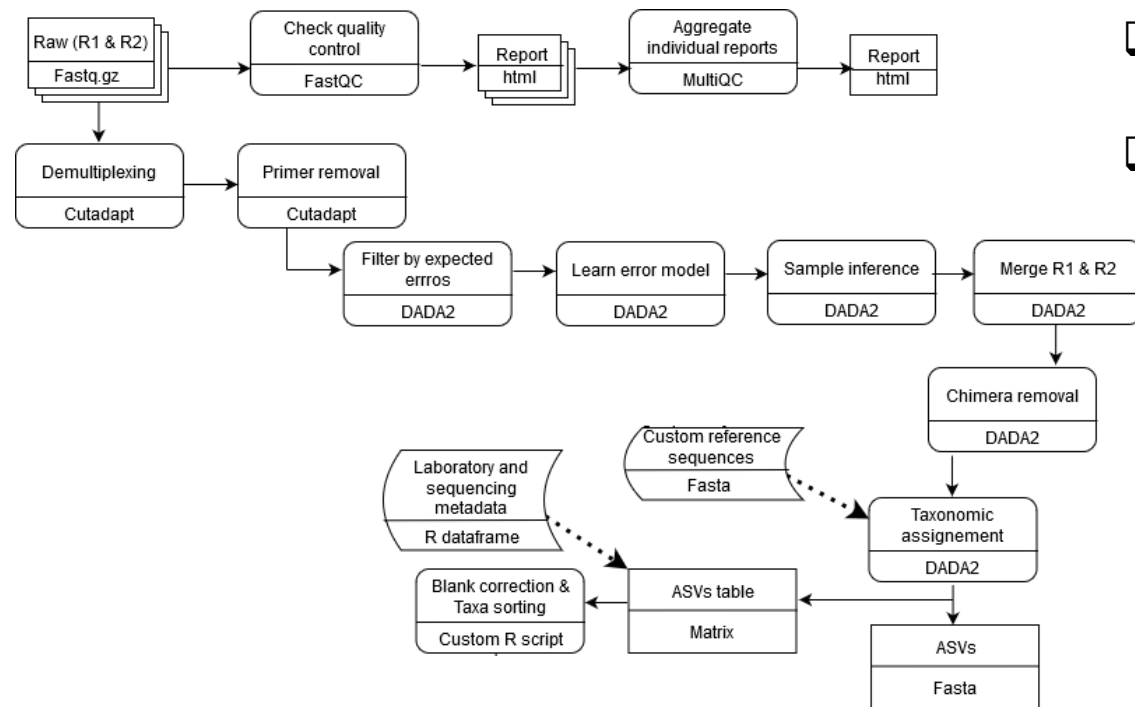
python

## FACETTES DE DIVERSITE

Projet tutoré



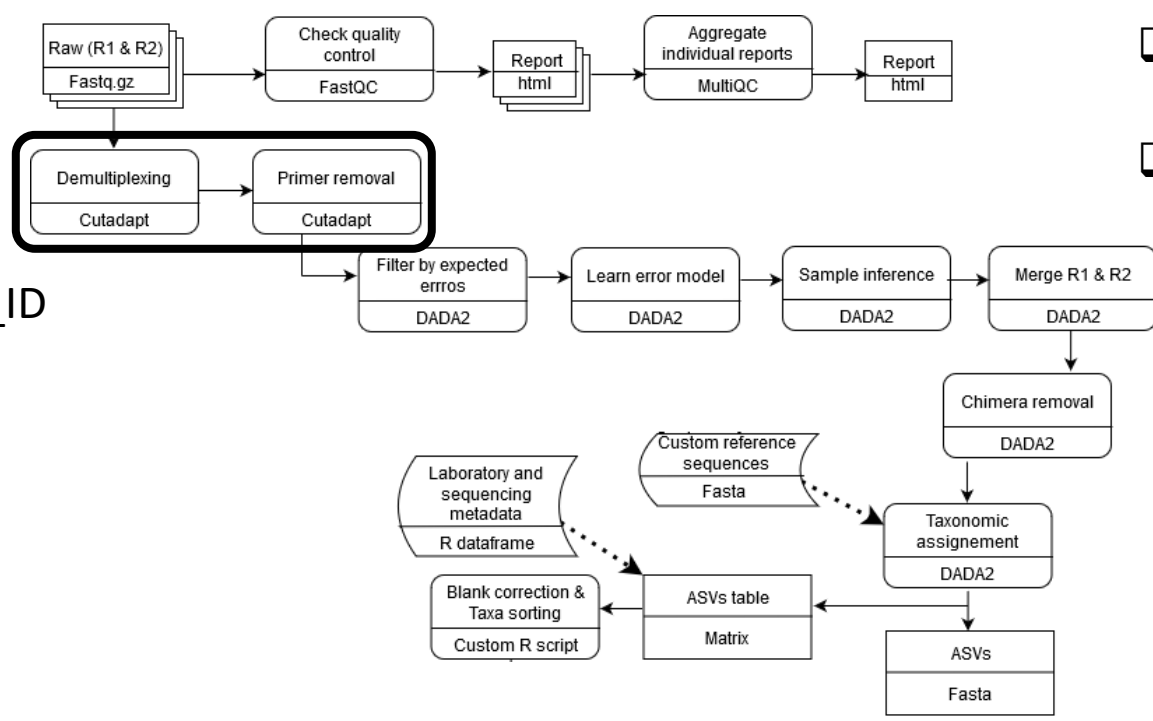
# R



Cluster d'ABiMS

Gestionnaire de ressource QSUB

❑ Multiples librairies : utilisation de boucles & parallélisation avec \$SGE\_TASK\_ID

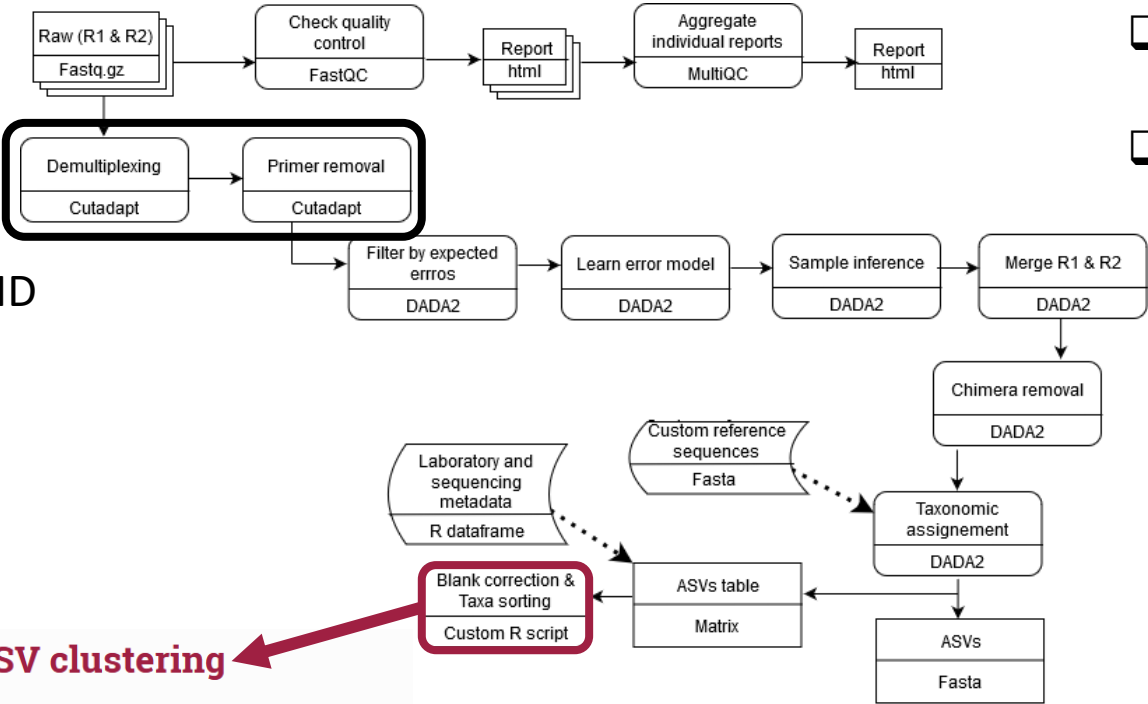


❑ Cluster d’ABiMS

❑ Gestionnaire de ressource QSUB

Multiples librairies :  
 utilisation de boucles &  
 parallélisation avec \$SGE\_TASK\_ID

- Cluster d’ABiMS
- Gestionnaire de ressource QSUB



### Processing post ASV clustering

This script performs:

- Index jumping filtering
- Normalization of librairies for 12S only
- Removing of ASVs assigned to "non-fishes" taxa
- LULU post-clustering filtering
- Make the final taxonomic assignment of ASVs by comparing RDP Classifier, ECOTAG and BLAST
- Clustering ASVs into OTUs with SWARM
- Removing OTUs singletons
- Building final OTU tables datasets

The final OTU tables for 12S and 16s can be used as input for the 03\_Statistical\_analyses.Rmd script.

### Set working directory and path to data

Set the directory where the Data\_processing folder is located

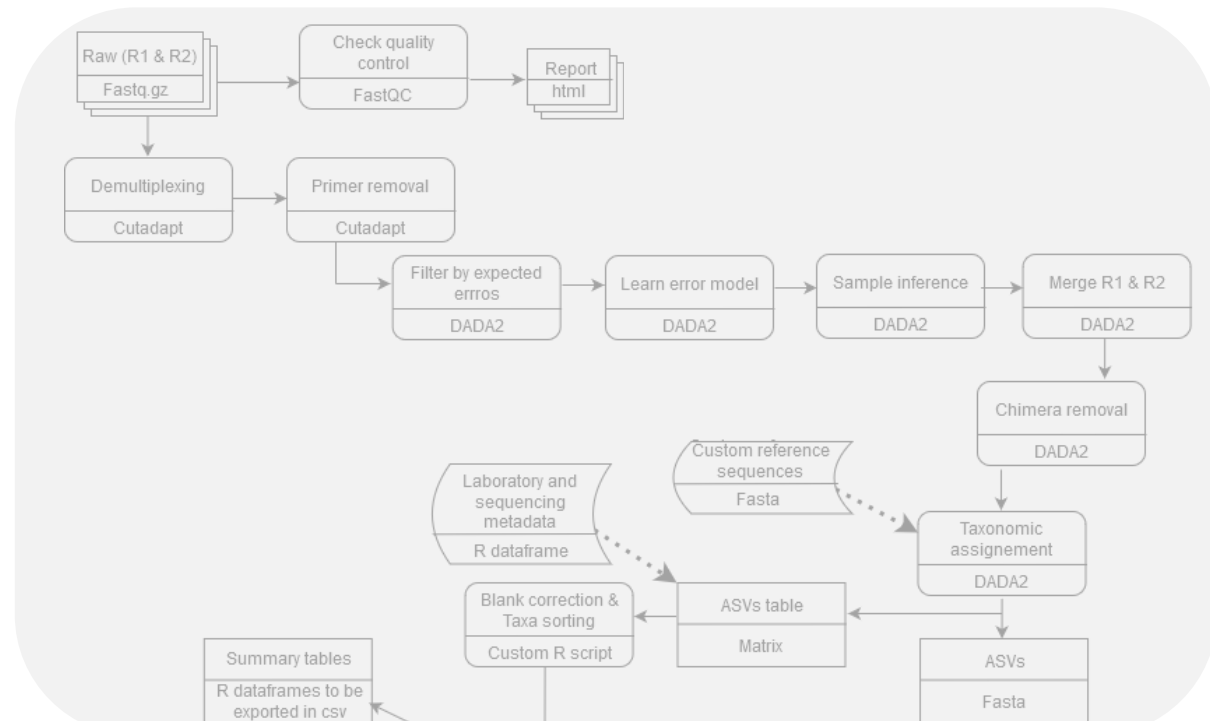


Get the path to required files for this script

Processing post ASV clustering

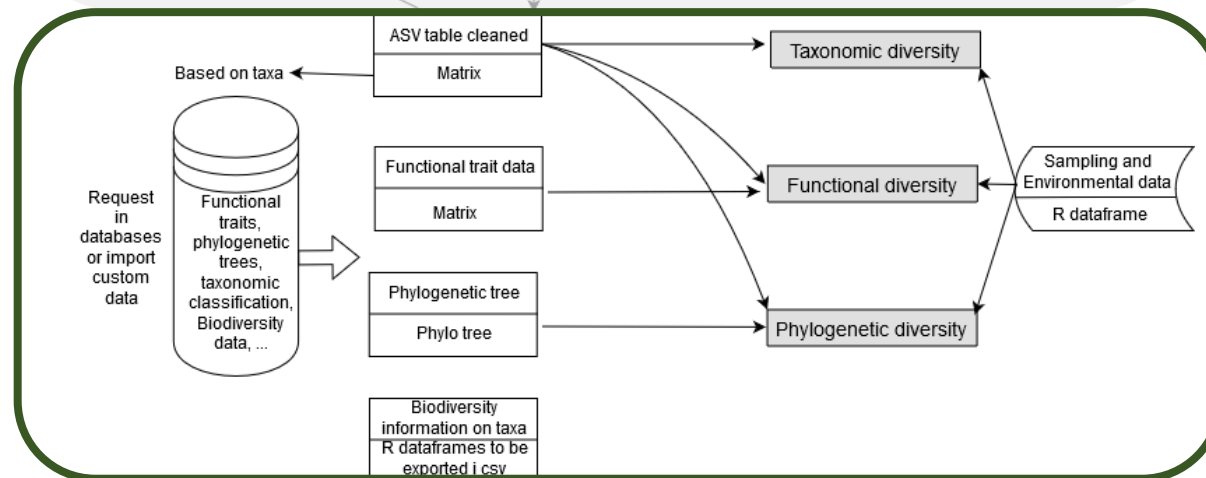
Set working directory and path to data  
 Load packages  
 Import data  
 ASV table formatting  
 Data trimming  
 Final taxonomic assignment  
 Homogenization with WORMS taxonomy  
 ASV to OTU  
 Get the final datasets  
 Sequencing results after each steps  
 Session info

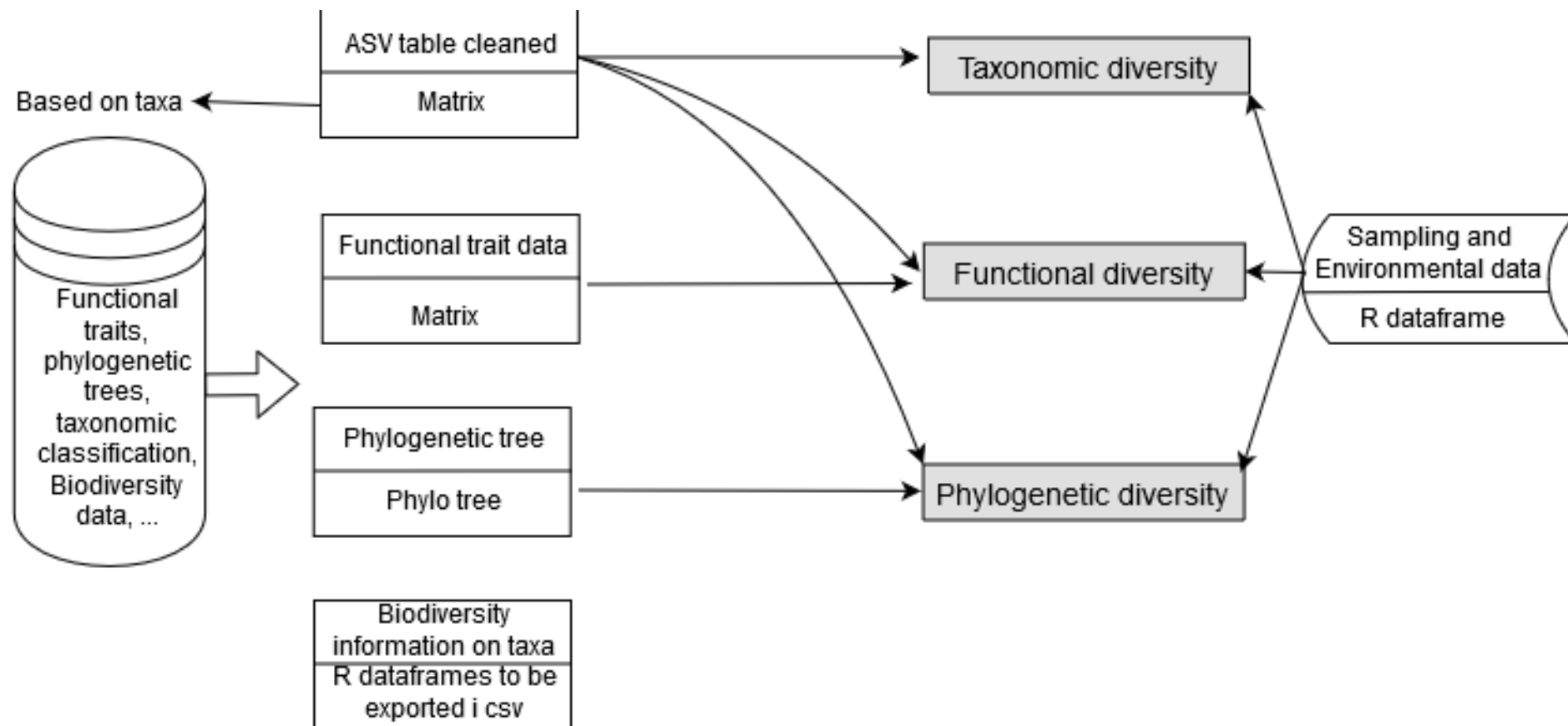
Anaïs Rey



## FACETTES DE DIVERSITE

### Projet tutoré







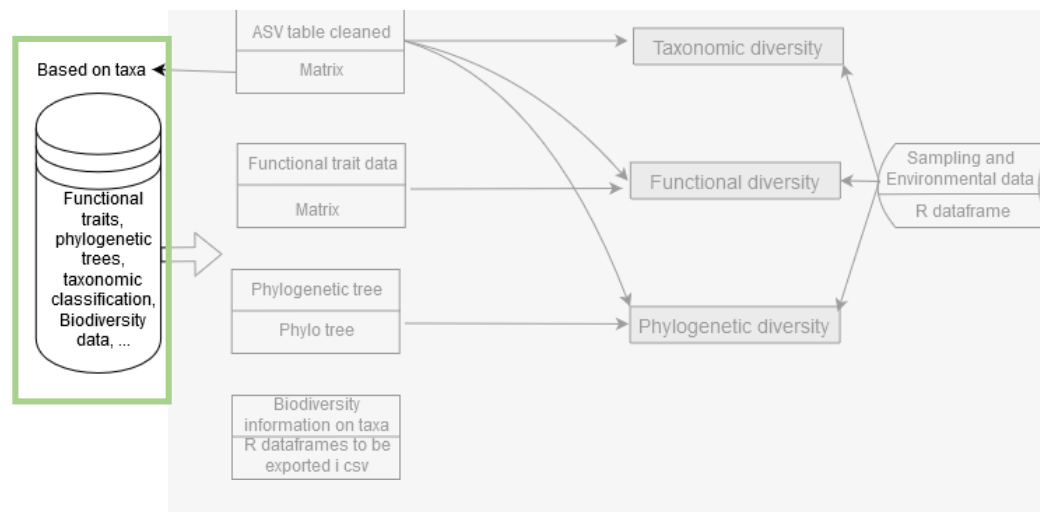
## Module 1 :

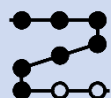
### Récolte des données écologiques/biologiques liées aux assignations taxonomiques

- Actualisation de la classification taxinomique



- Collecte des traits

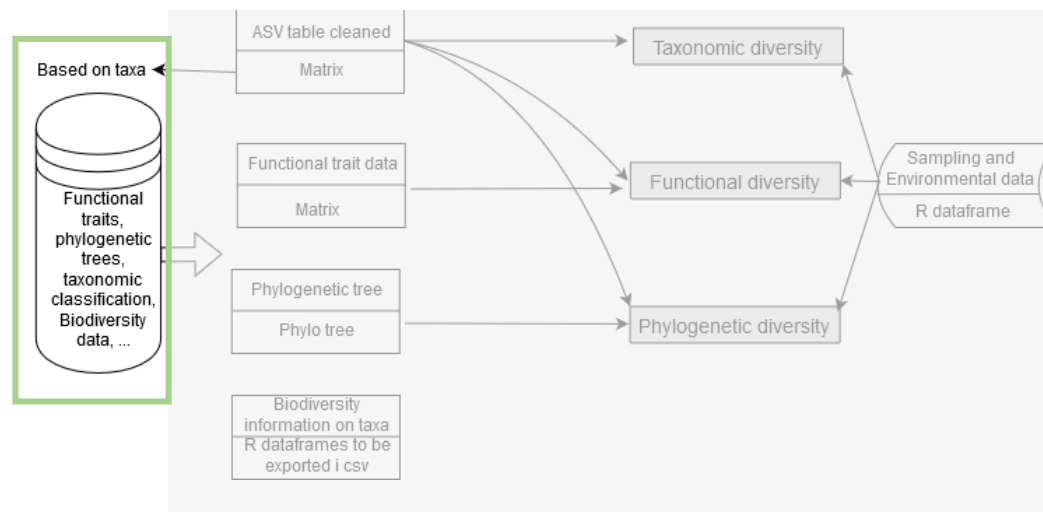


**Module 1 :****Récolte des données  
écologiques/biologiques liées aux  
assignations taxonomiques**

- Actualisation de la classification taxinomique



- Collecte des traits

**Module 2 :****Intégration des données  
d'occurrence des espèces**

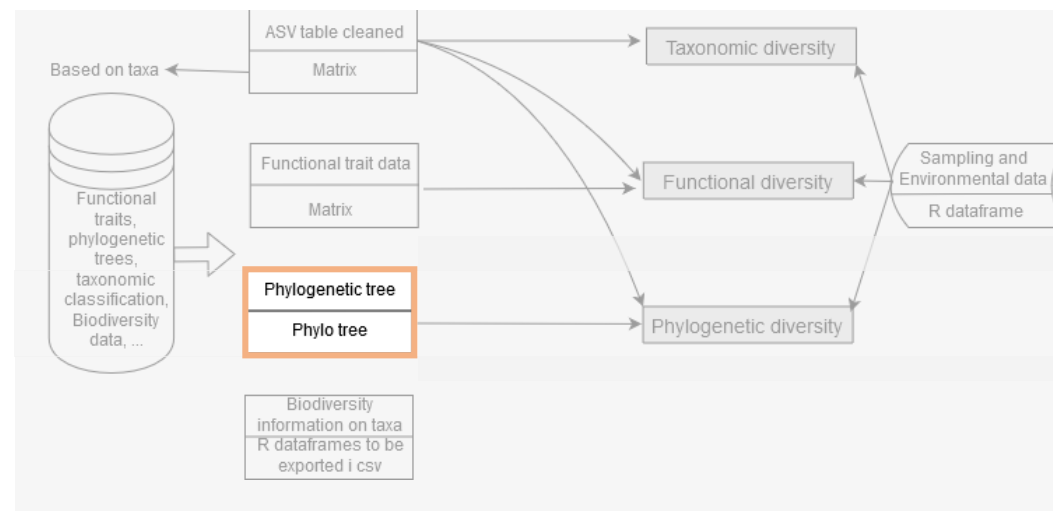
- Récolte et visualise les occurrences des espèces



- Identification de potentielles incorrectes assignations

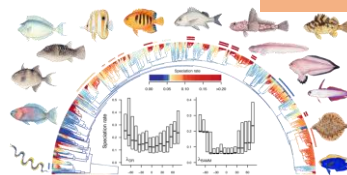


## Module 1 : Récolte des données écologiques/biologiques liées aux assignations taxonomiques

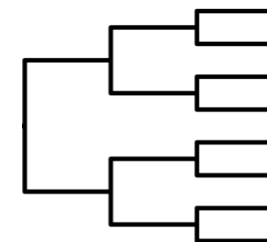


## Module 2 : Intégration des données d'occurrence des espèces

## Module 3 : Intégration des données phylogénétiques



- « Subset » arbres existant sur taxa identifiés
- Placement phylogénétique des OTUs





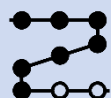
Introduction



Données



PGD



Flowchart



Outils



Résultats



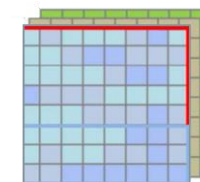
Retours



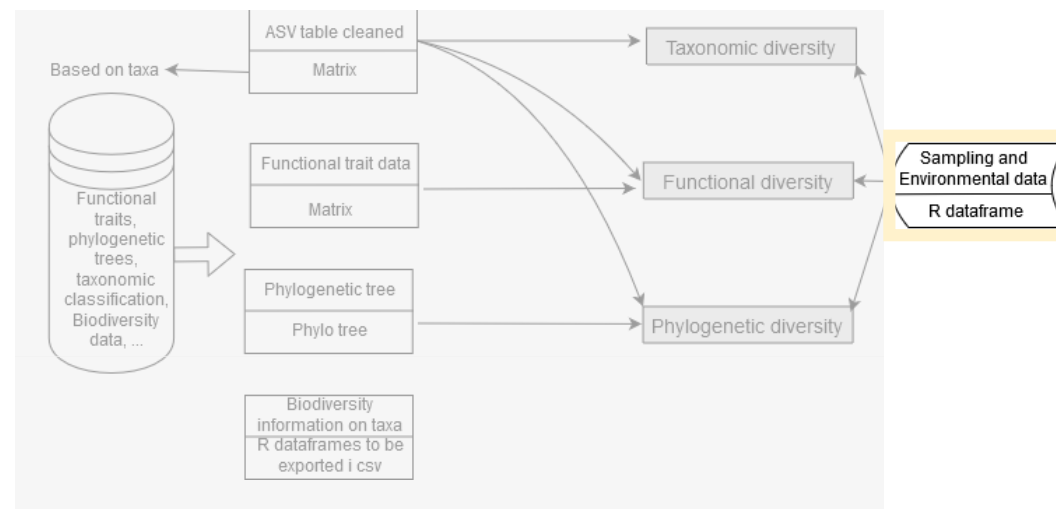
Perspectives

**Module 1 :**  
**Récolte des données**  
**écologiques/biologiques liées aux**  
**assignations taxonomiques**

**Module 4 :**  
**Récolte des métadonnées**  
**environnementales**



**Module 2 :**  
**Intégration des données**  
**d'occurrence des espèces**



**Module 3 :**  
**Intégration des données**  
**phylogénétiques**



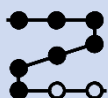
Introduction



Données



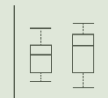
PGD



Flowchart



Outils



Résultats



Retours

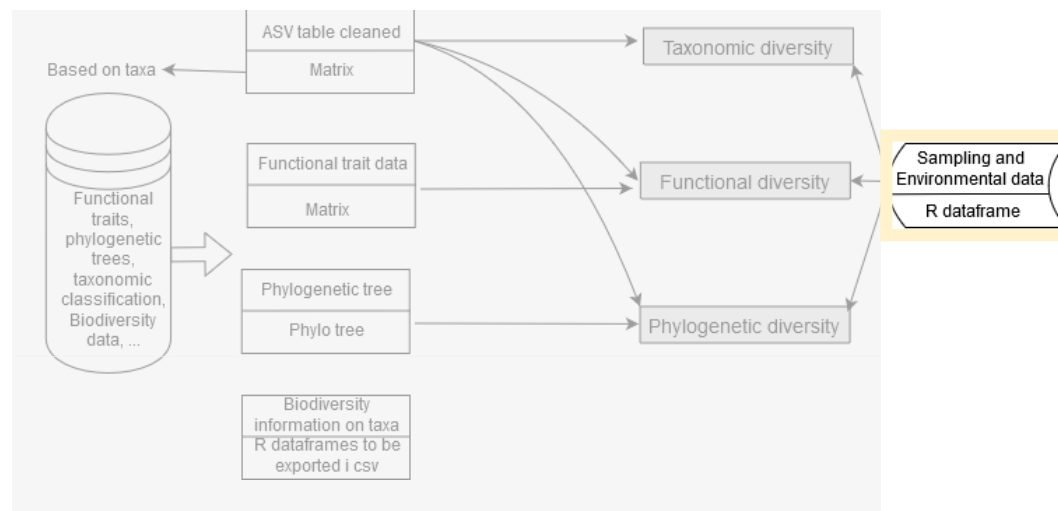


Perspectives

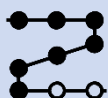
**Module 1 :**  
**Récolte des données**  
**écologiques/biologiques liées aux**  
**assignations taxonomiques**

**Module 4 :**  
**~~Récolte des métadonnées~~**  
**environnementales**

**Module 2 :**  
**Intégration des données**  
**d'occurrence des espèces**

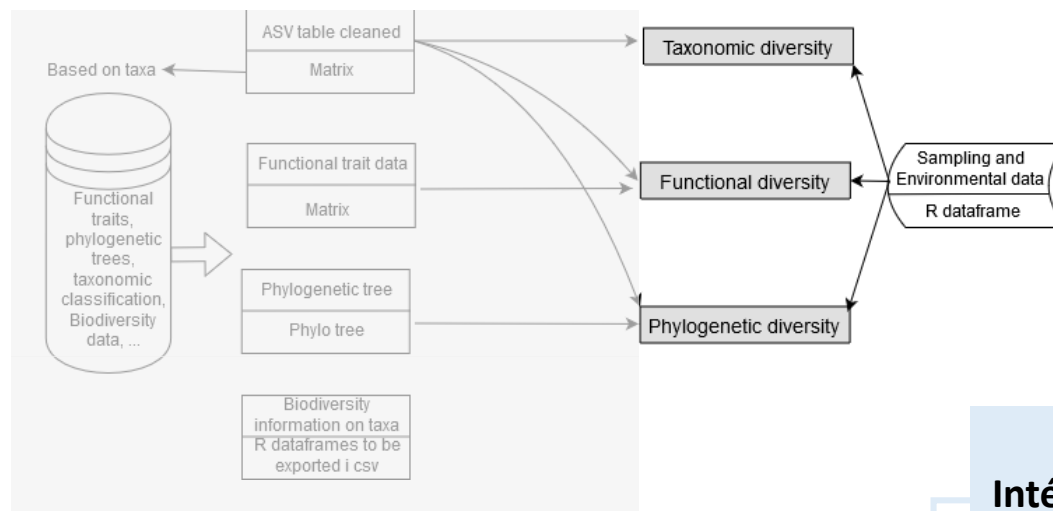


**Module 3 :**  
**Intégration des données**  
**phylogénétiques**



**Module 1 :**  
**Récolte des données**  
**écologiques/biologiques liées aux**  
**assignations taxonomiques**

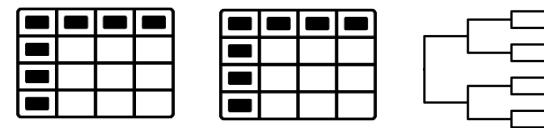
**Module 4 :**  
**Récolte des métadonnées**  
**environnementales**



**Module 2 :**  
**Intégration des données**  
**d'occurrence des espèces**

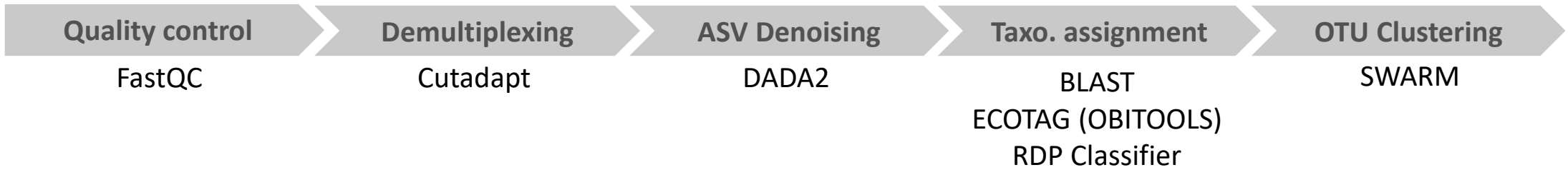
**Module 3 :**  
**Intégration des données**  
**phylogénétiques**

**Module 5 :**  
**Intégration des différentes facettes**  
**de diversité**



- Préparation des 3 jeux de données
- Analyses des diversité  $\alpha$  et  $\beta$  multi-facettes

☐ **Outils bioinformatiques – partie metabarcoding**



❑ Outils bioinformatiques – partie metabarcoding

❑ **Packages R et fonctions – partie projet tutoré**

### Manipulation de données

dplyr – tibble – tidyr – rlang

### Visualisation

ggplot – ggVennDiagram –  
RColorBrewer

### Données d'occurrence

robis – sp – ggmap

### Facettes de diversité

FD – hillr – adiv

### Taxonomie - Traits

taxize – rfishbase – worrms

### Phylogénie

fishtree – ape – picante

→ **Intégration des différentes fonctions des différents packages au sein de fonctions**





❑ Outils bioinformatiques – partie metabarcoding

❑ Packages R et fonctions – partie projet tutoré



## Taxonomie

taxize – worms

## Manipulation de données

dplyr – tibble – tidyr

### Module "Data Harvest"

#### Taxonomic classification

**Description:** Give the accepted taxonomic classification under [WORMS](#) and if not found under [GBIF](#)

There are two steps:

- **Resolve taxonomic name:** for instance, get the correct spelling when fuzzy spelling, remove non-scientific name information (spp., subspecies x, ...). This is done by comparing with WORMS and GBIF database : First it looks if the taxa is present in WORMS and if not, it goes to GBIF
- **Retrieve the taxonomic classification:** It retrieves from the package [worms](#) the accepted taxonomy (scientific accepted name and associated classification) of each query by putting `marine_only=FALSE` to increase the number of species for which the taxonomy is found. When the taxon is not found in WORMS, a second search is done in GBIF (useful for instance when non-marine/aquatic taxa are present)

**Users' action:** Execute the blow chunk and move to the following chunk

```
Entrée [9]: taxo.table.class <- get_taxo_class(taxo.table.in=taxo.table)
```

```
== 1 queries =====
```

```
Retrieving data for taxon 'Spariformes'
```

```
Not found. Consider checking the spelling or alternate classification
```

```
x Not Found: Spariformes  
== Results =====
```

```
* Total: 1  
* Found: 0  
* Not Found: 1  
[1] "All taxa from the otu table were searched for taxonomic classification"
```

☐ Outils bioinformatiques – partie metabarcoding

☐ Packages R et fonctions – partie projet tutoré

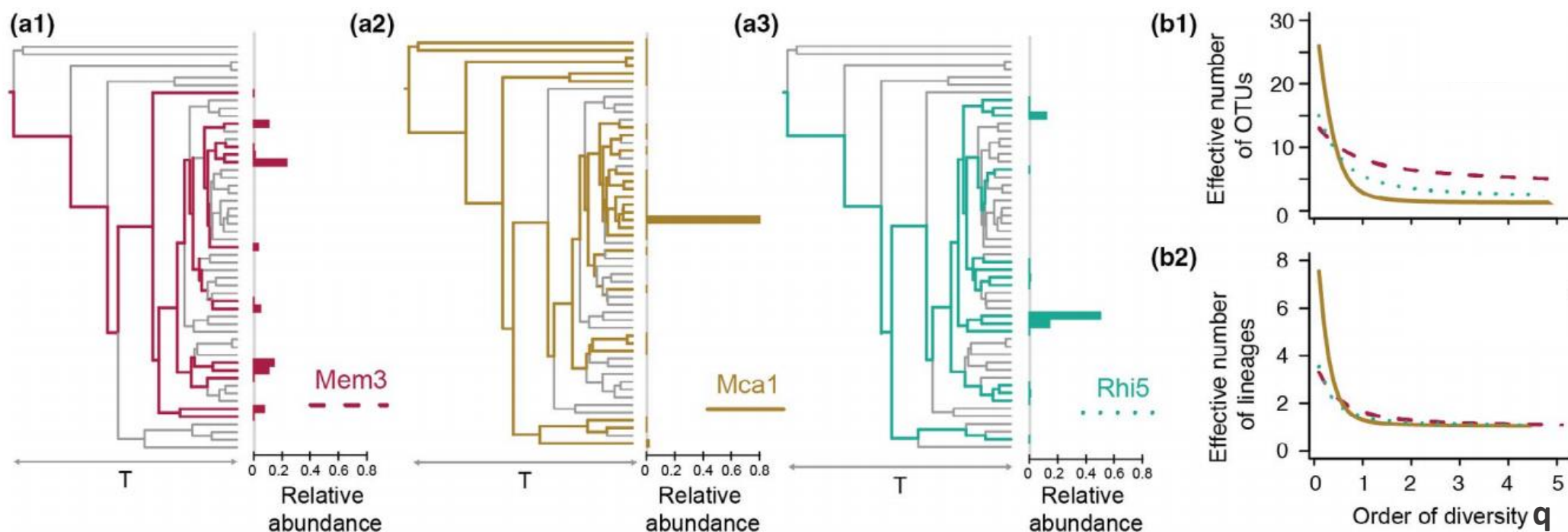
☐ **Outils statistiques – partie projet tutoré**

### Nombres de Hill $^qD$

Unifient multiples indices de diversité en écologie

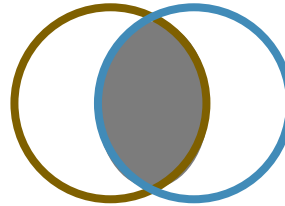
Intègrent diversité  $\alpha$  et  $\beta$  taxinomique, phylogénétique et fonctionnelle

Basés sur l'utilisation du paramètre  $q$  – plus  $q$  est élevé plus de poids est donné aux espèces abondantes



**Module 2 :**  
**Intégration des données  
d'occurrence des espèces**

**Occurrences des taxa  
identifiés dans le jeu de  
données de l'utilisateur**



**« Redflag »**

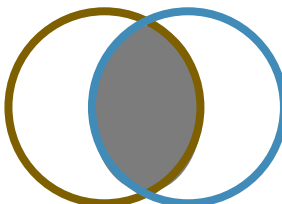
**Occurrence d'une espèce**

**Occurrence d'un genre**



## Module 2 : Intégration des données d'occurrence des espèces

Occurrences des taxa  
identifiés dans le jeu de  
données de l'utilisateur



### « Redflag »

Identifie taxa à  
vérifier

Occurrence d'une espèce

Occurrence d'un genre

```
> spList <- taxo.table.class %>% filter(rank=="Species") %>% pull(valid_name)
>
> id.redflag.obis(spList,
+               startdate_chosen = NULL, # enter a year
+               type_zone="area",
+               coord_zone="40024") # North East Atlantic : 40024

  species                explanation
1 Chromogobius britoi    REDFLAG: not found in obis
2 Barbatula barbatula    REDFLAG: not found in obis
3 Phoxinus phoxinus      REDFLAG: not found in the area
4 Rutilus rutilus        REDFLAG: not found in the area
```



Introduction



Données



PGD



Flowchart



Outils



Résultats



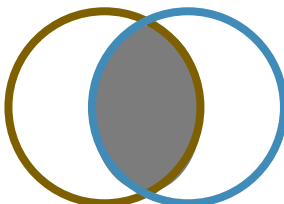
Retours



Perspectives

## Module 2 : Intégration des données d'occurrence des espèces

Occurrences des taxa  
identifiés dans le jeu de  
données de l'utilisateur

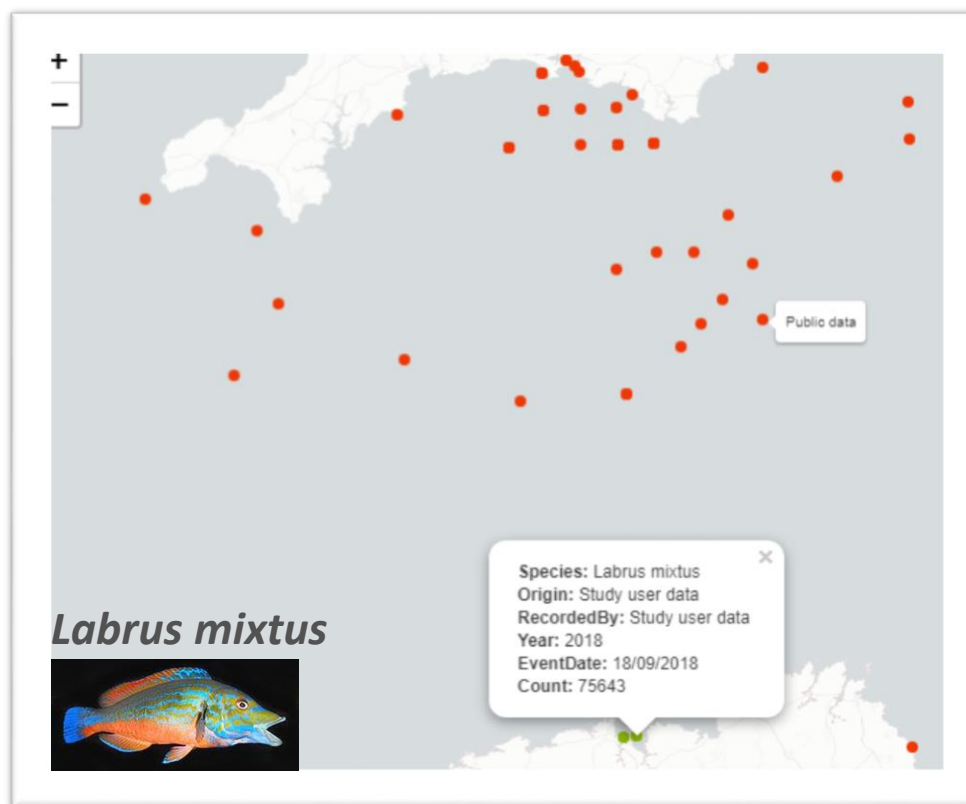


« Redflag »

**Occurrence d'une espèce**

Récupère données d'occurrence

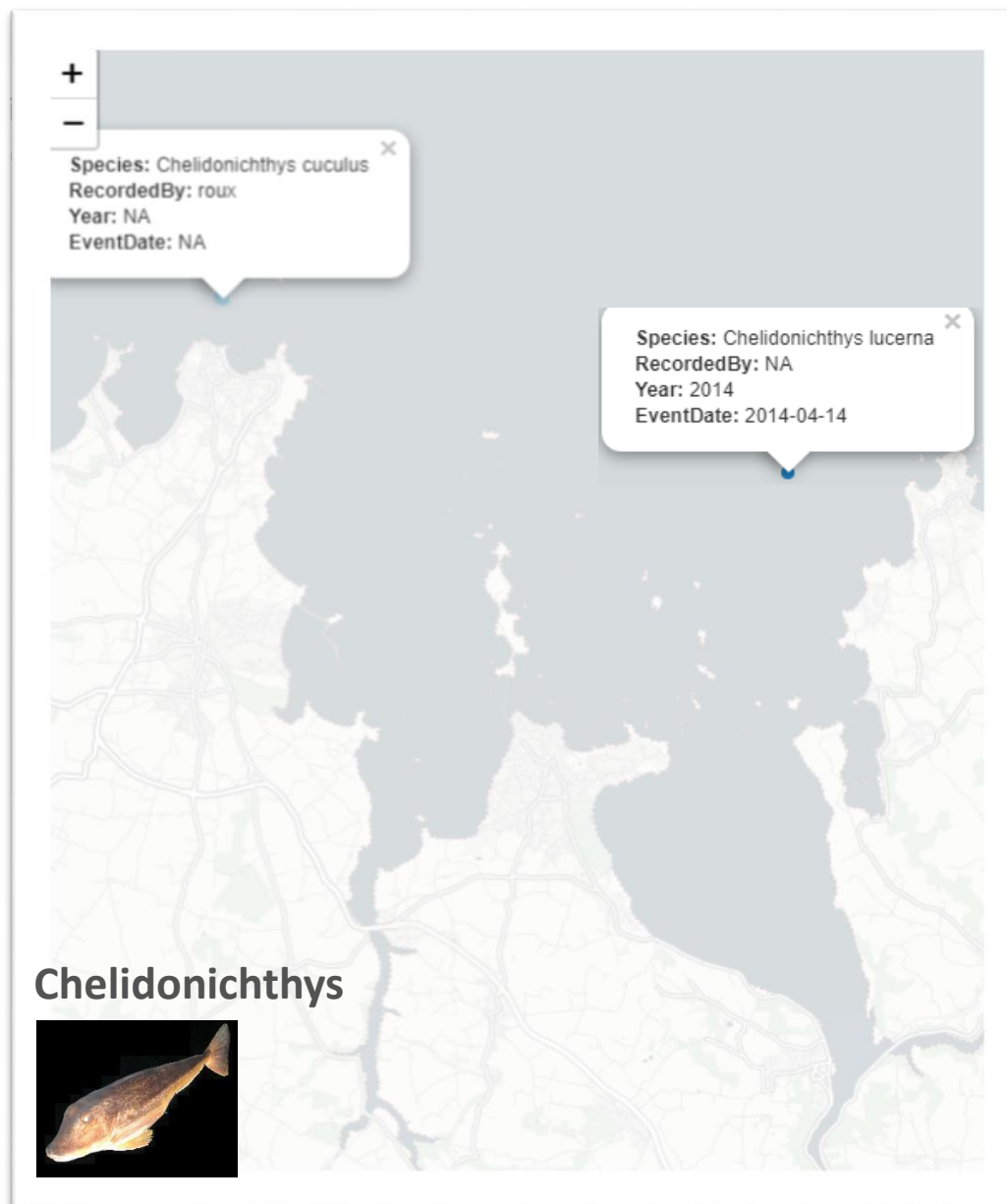
**Occurrence d'un genre**





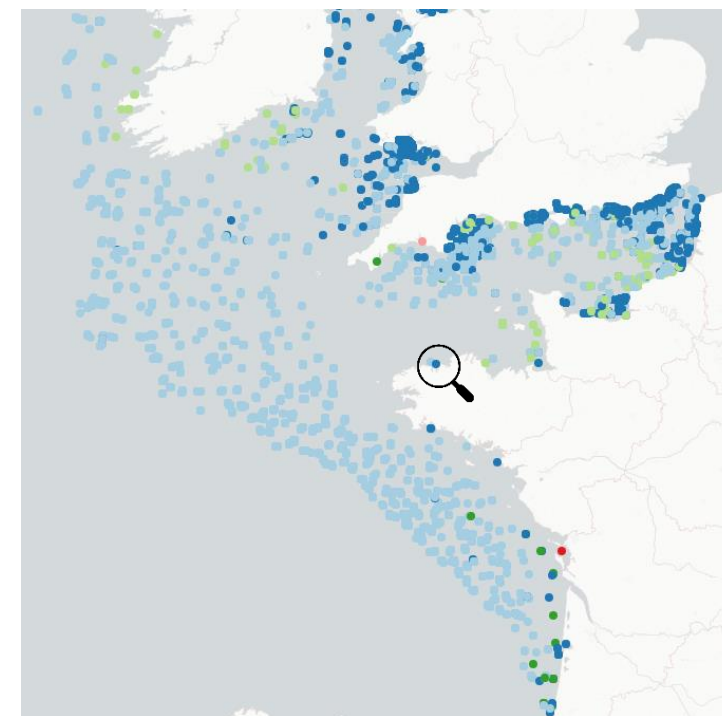
## Module 2 : Intégration des données d'occurrence des espèces

« Redflag »



## Occurrence d'un genre

Identifie espèces  
potentiellement présentes



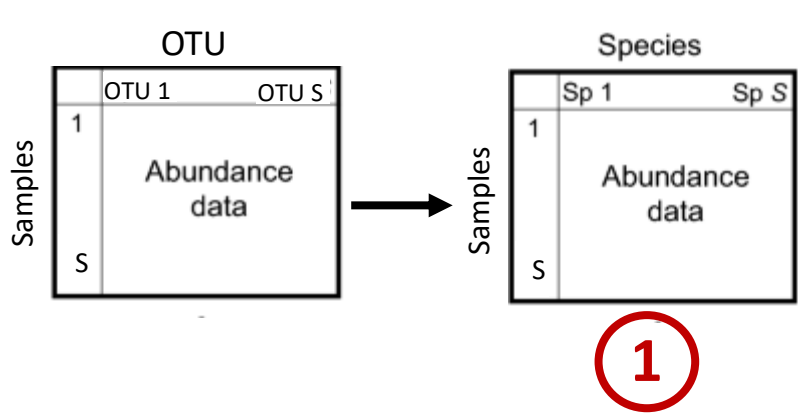
## Module 5 : Intégration des différentes facettes de diversité

- Préparation des **3 jeux de données**

OTU	
	OTU 1 OTU S
1	Abundance data
S	

Module 5 :  
Intégration des différentes facettes  
de diversité

- Préparation des 3 jeux de données

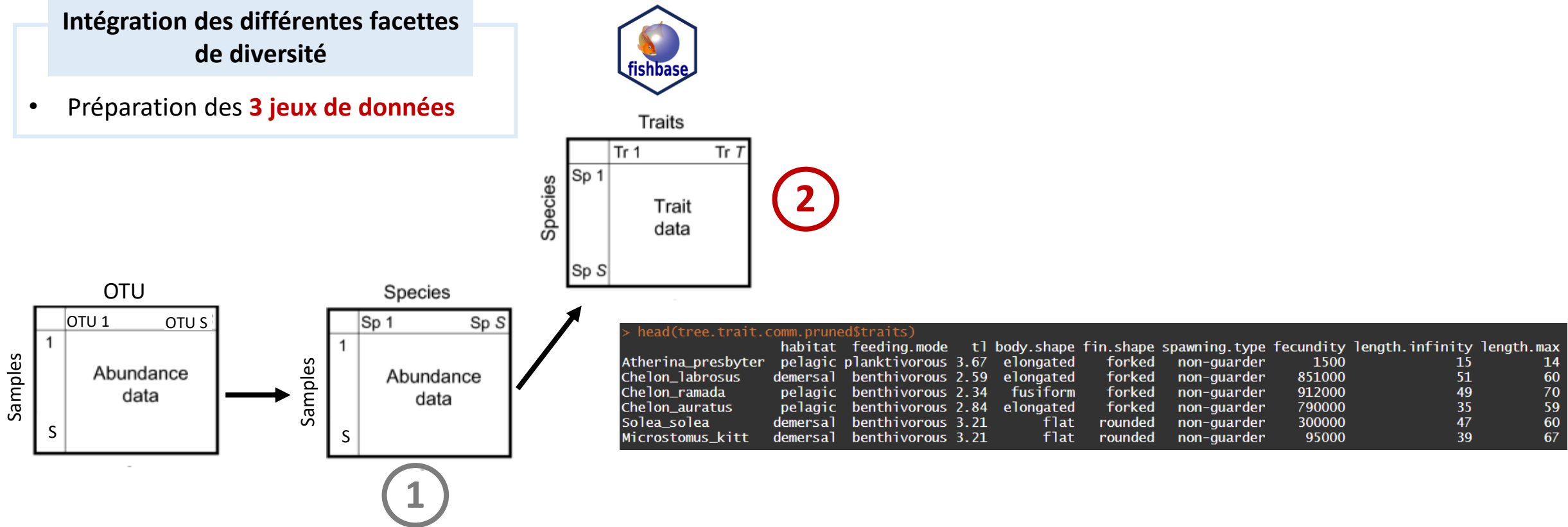


```
> head(tree.trait.comm.pruned$comm)
      Atherina_presbyter Chelon_labrosus Chelon_ramada Chelon_auratus Solea_solea Microstomus_kitt
AST.JUNE.F.R1           0              0              0              0              0              0
AST.JUNE.F.R2           0              0              0              0              0              0
AST.JUNE.F.R3           0              0              0              0              0              0
AST.JUNE.S.R1        41368              0              0          18397              0              0
AST.JUNE.S.R2        38654              0              0              0              0              0
AST.JUNE.S.R3        28090              0              0              0              0              0
      Zeugopterus_punctatus Scophthalmus_maximus Gobius_niger Gobius_paganellus Pomatoschistus_minutus
AST.JUNE.F.R1              0              0              0              0              0
AST.JUNE.F.R2              0              0              0              4861              0
AST.JUNE.F.R3              0              0              0              0              0
AST.JUNE.S.R1              0              0              0              0              0
AST.JUNE.S.R2              0              0              0              0              0
AST.JUNE.S.R3              0              0              0          9879              0
```



Module 5 :  
Intégration des différentes facettes  
de diversité

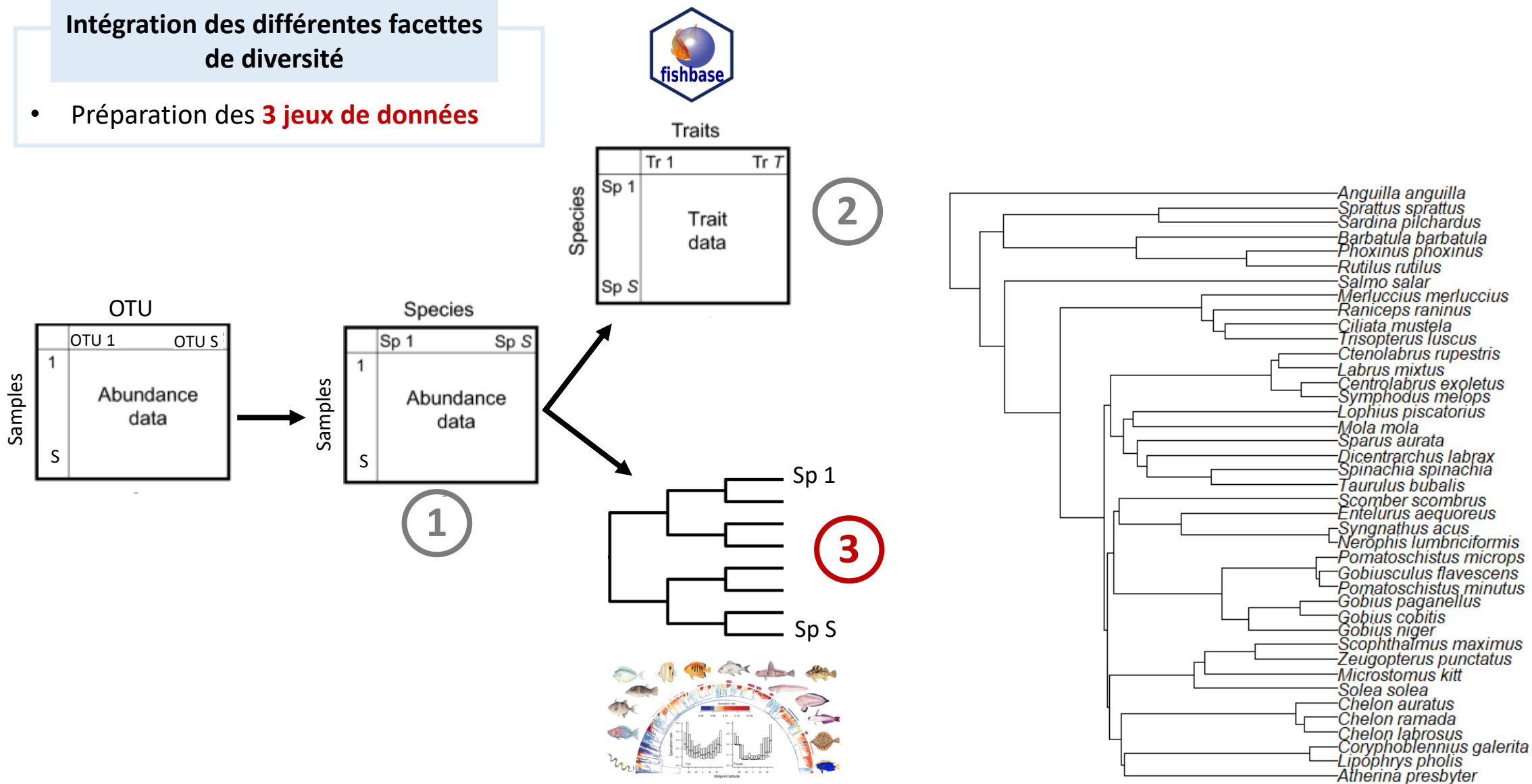
- Préparation des 3 jeux de données



```
> head(tree.trait.comm.pruned$traits)
      habitat feeding.mode  t1 body.shape fin.shape spawning.type fecundity length.infinity length.max
Atherina_presbyter  pelagic planktivorous 3.67 elongated   forked   non-guarder    1500           15          14
Chelon_labrosus     demersal benthivorous 2.59 elongated   forked   non-guarder   851000          51          60
Chelon_ramada       pelagic benthivorous 2.34 fusiform    forked   non-guarder   912000          49          70
Chelon_auratus      pelagic benthivorous 2.84 elongated   forked   non-guarder   790000          35          59
Solea_solea         demersal benthivorous 3.21 flat         rounded   non-guarder   300000          47          60
Microstomus_kitt    demersal benthivorous 3.21 flat         rounded   non-guarder   95000           39          67
```

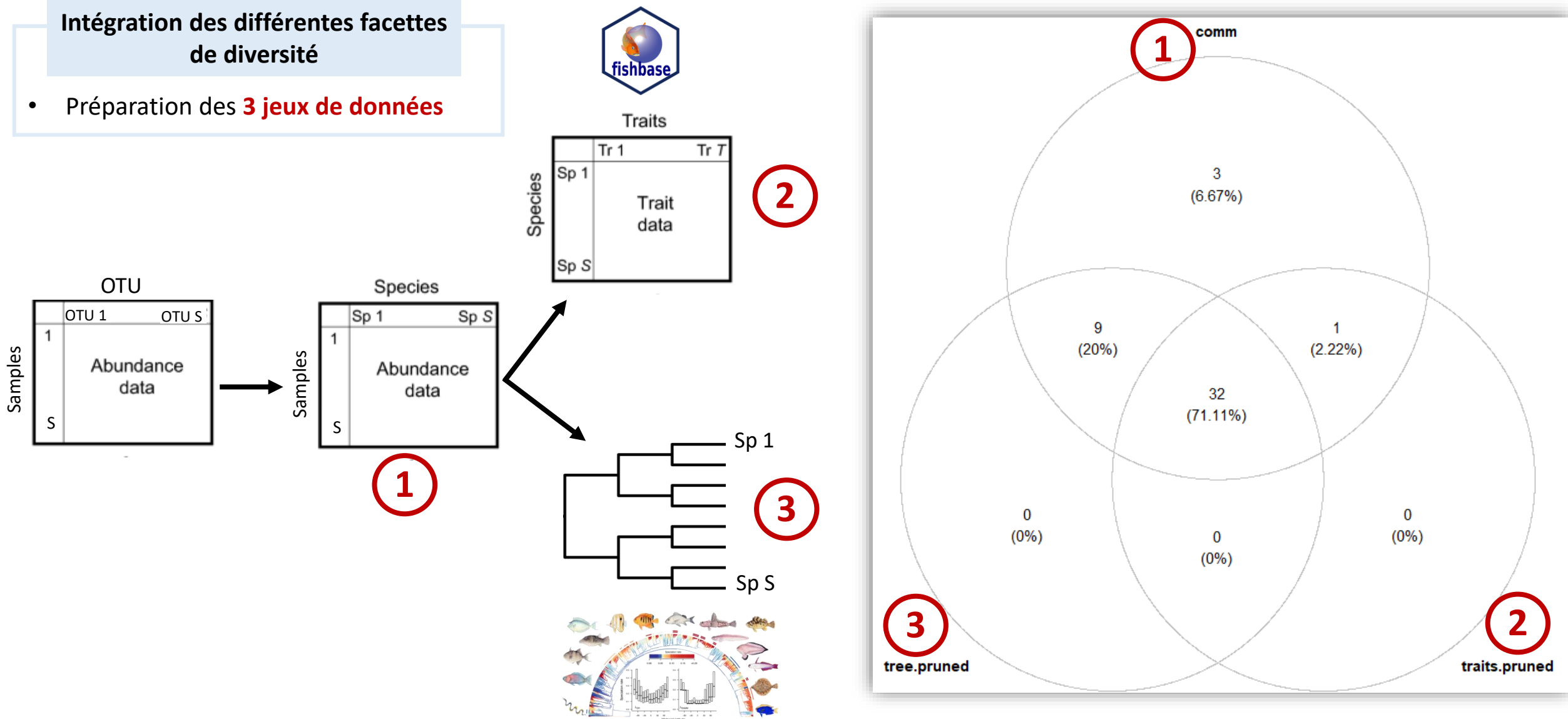
## Module 5 : Intégration des différentes facettes de diversité

- Préparation des **3 jeux de données**



## Module 5 : Intégration des différentes facettes de diversité

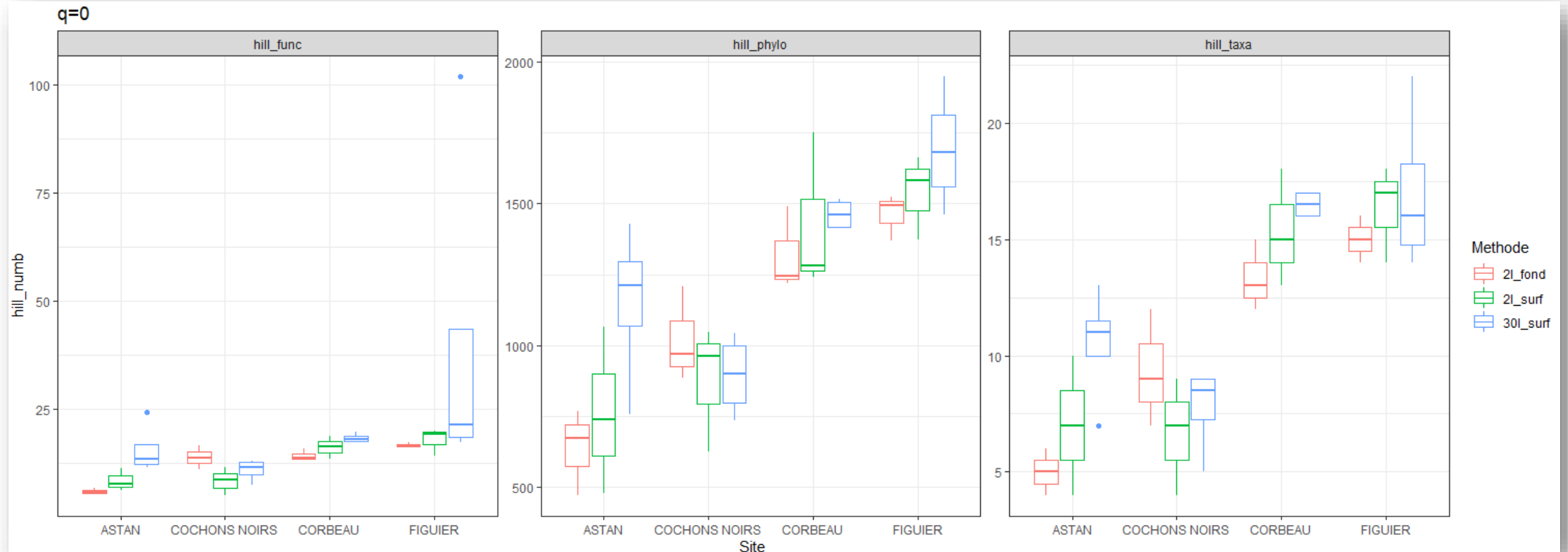
- Préparation des **3 jeux de données**



## Module 5 : Intégration des différentes facettes de diversité

- Analyses des diversité  $\alpha$  multi-facettes

## Nombres de Hill – Diversité $\alpha$





- ✓ **Projet en total continuité avec mon poste actuel** → connaissance des données, cluster, plateforme
- ✓ **Echanges réguliers avec les différents membres de la plateforme** → bases de données biologiques (format Darwin Core), organisation du workflow, reproductibilité



## Difficultés

### ---- Diversité des tâches

fonctions R reproductibles, appréhender analyses statistiques, accès aux bases de données environnementales

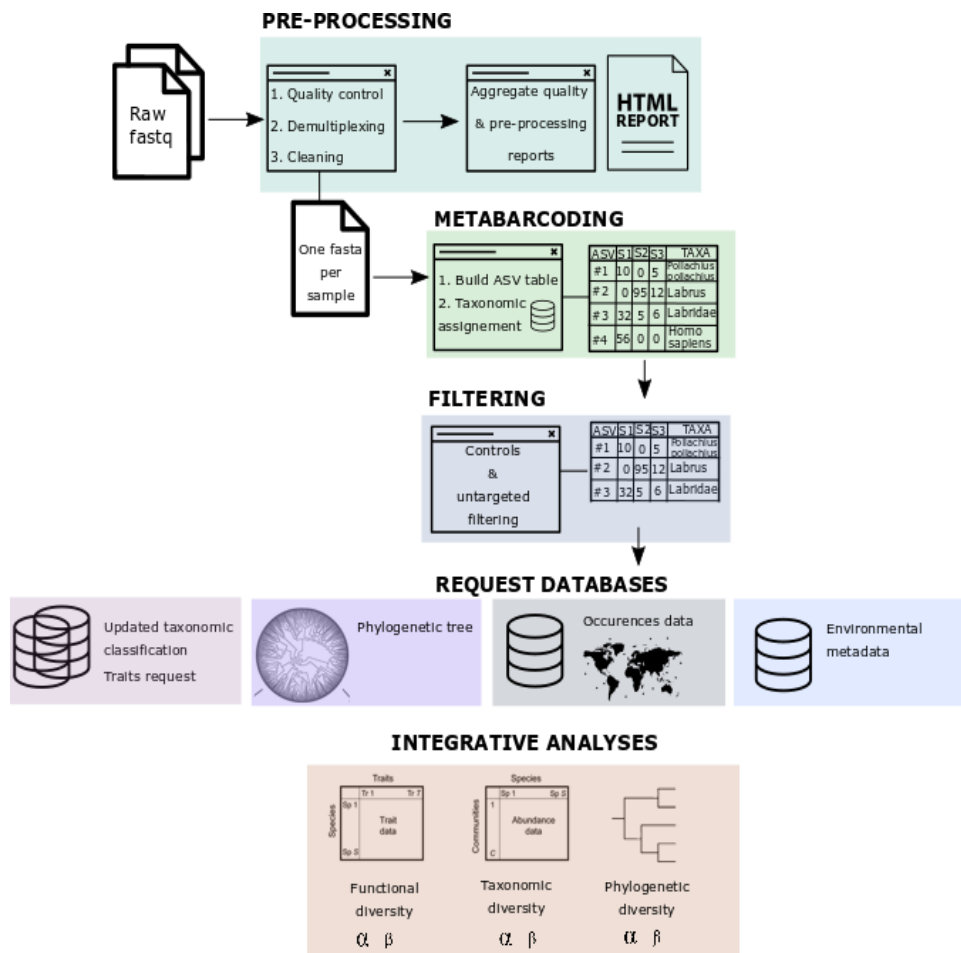
### ---- Gérer le temps

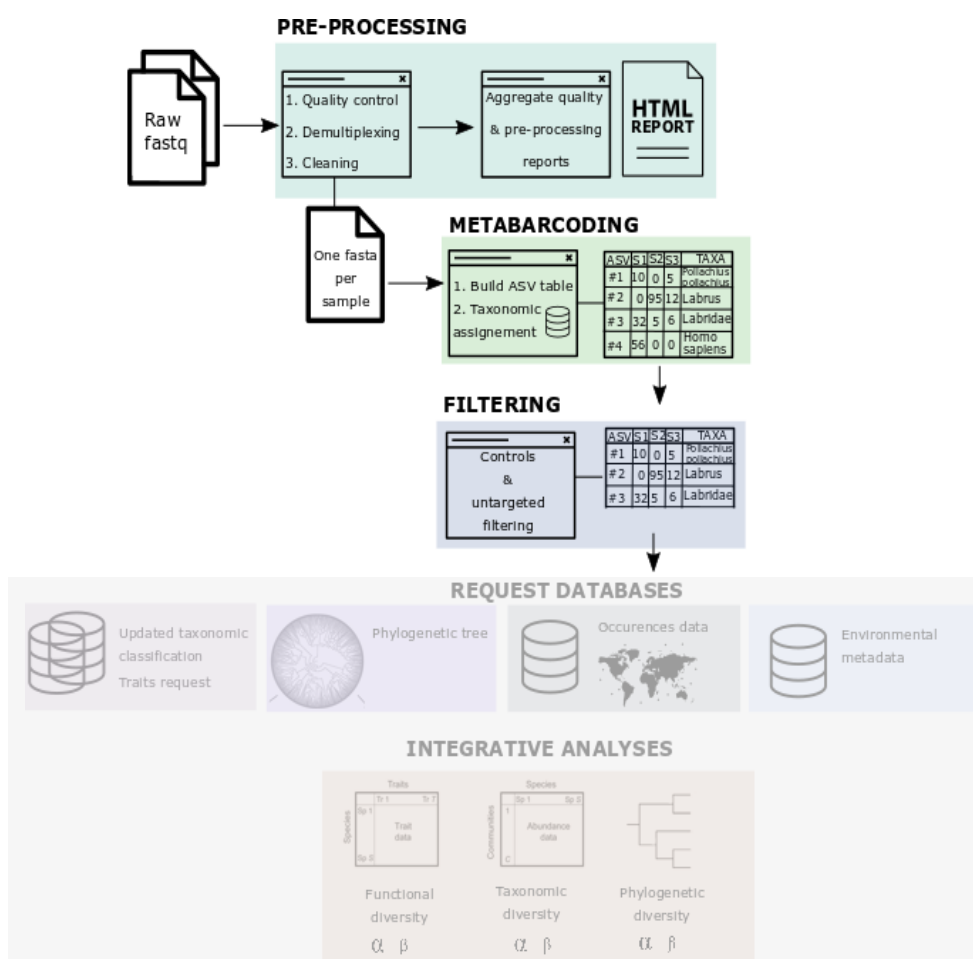
Gain de temps sur le module phylogénétique → utilisation d'arbre phylogénétique déjà existant au lieu de le créer par placement phylogénétique

### ---- Choix de la forme du workflow

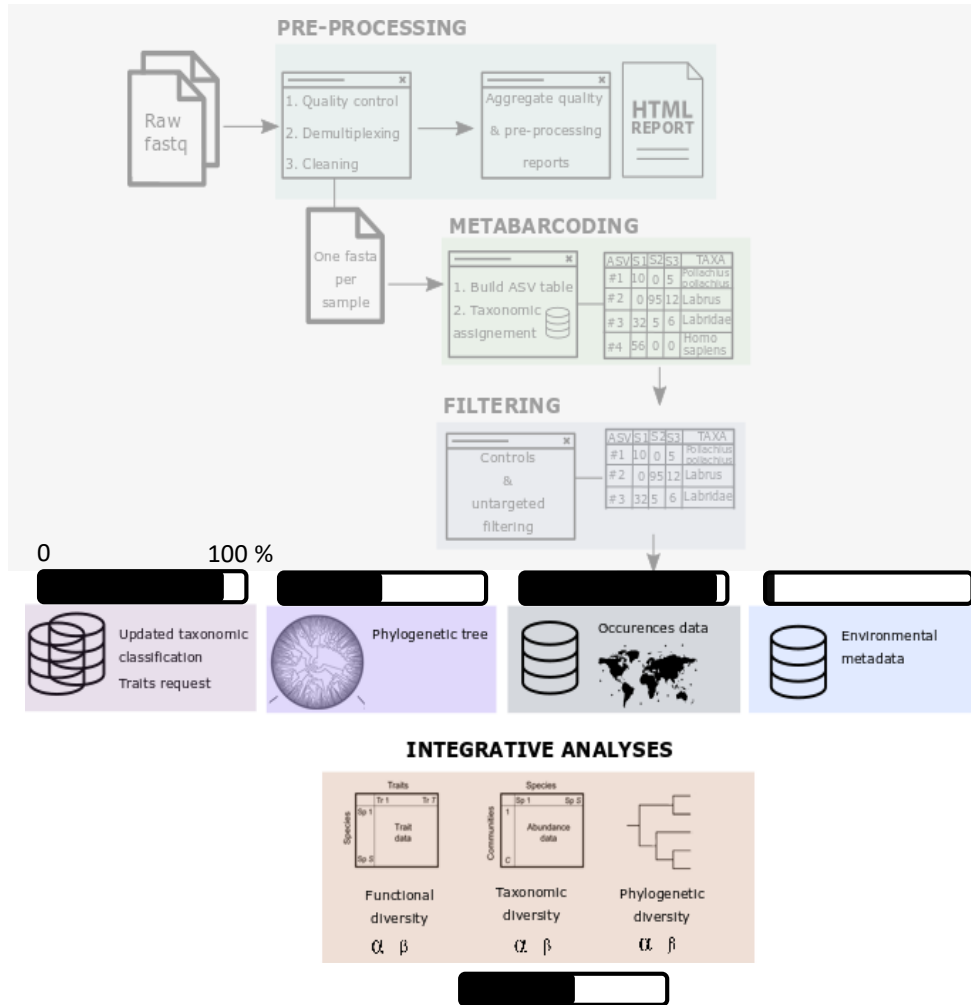


→ compromis **Jupyterlab** *(pour l'instant)*





- Finaliser les scripts metabarcoding :
- QSUB → SLURM
- Analyses reproductibles pour les publications



## Finaliser les scripts metabarcoding :

- QSUB → SLURM avec nextflow ou snakemake
- Analyses reproductibles pour les publications

## Continuer le workflow :

- Choix des outils statistiques intégrant diversité phylogénétique, fonctionnelle et taxinomique → mixKERNEL
- Effort dans la documentation
- Version test sera publiée sur github
- Contrôle des versions : git+github
- FAIRification:

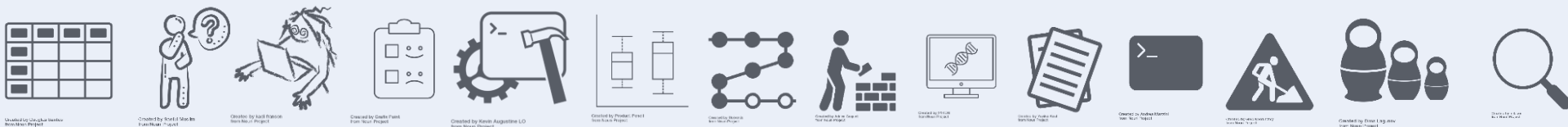
? Créer un environnement CONDA → gère dépendances et versions des packages











# Merci de votre attention !

Un grand merci à Erwan, Fred, Pierre, Mark, Gildas, Romain, ABiMS et tout le DUBII !!

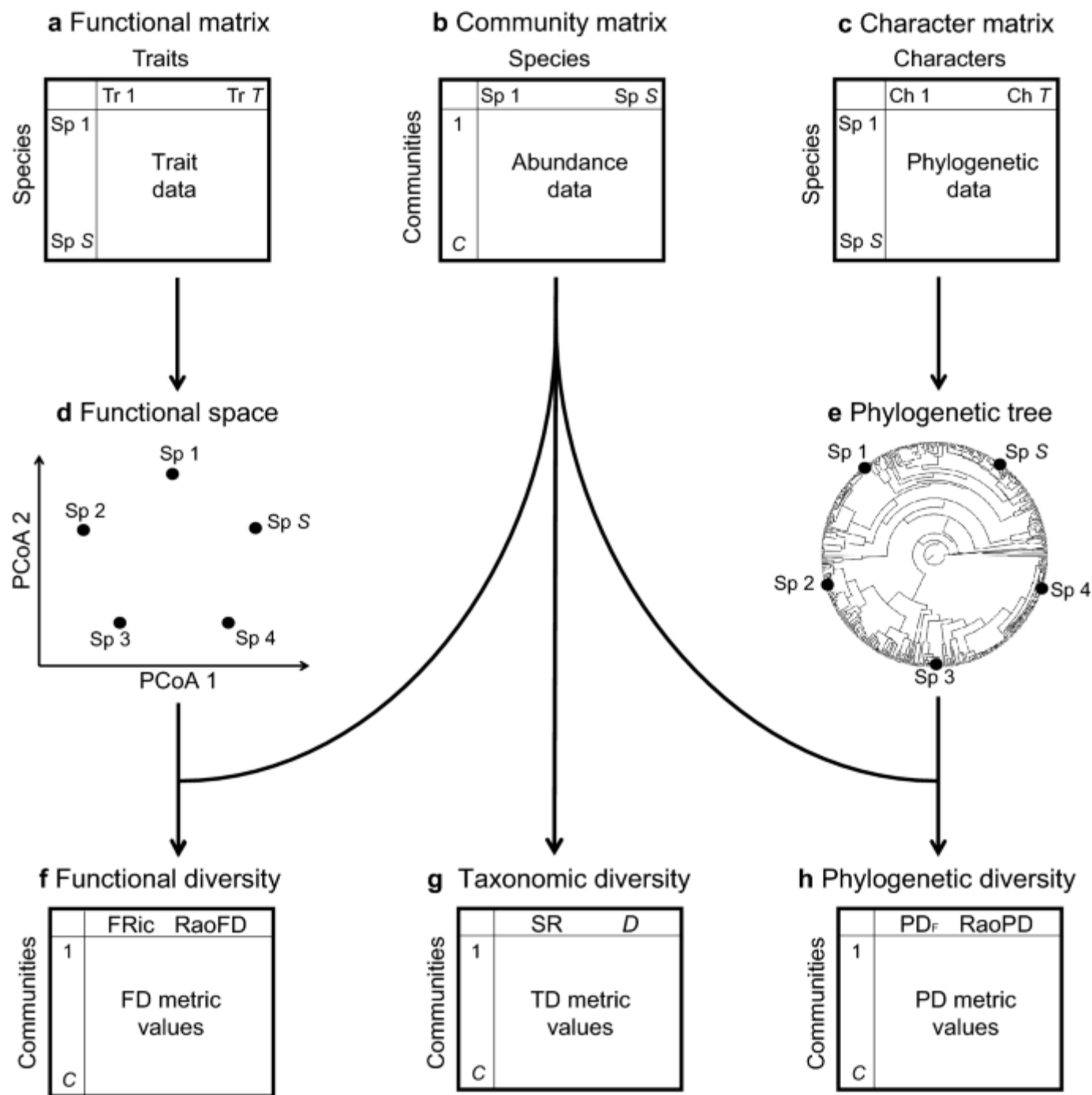
Source logo: <https://thenounproject.com/>

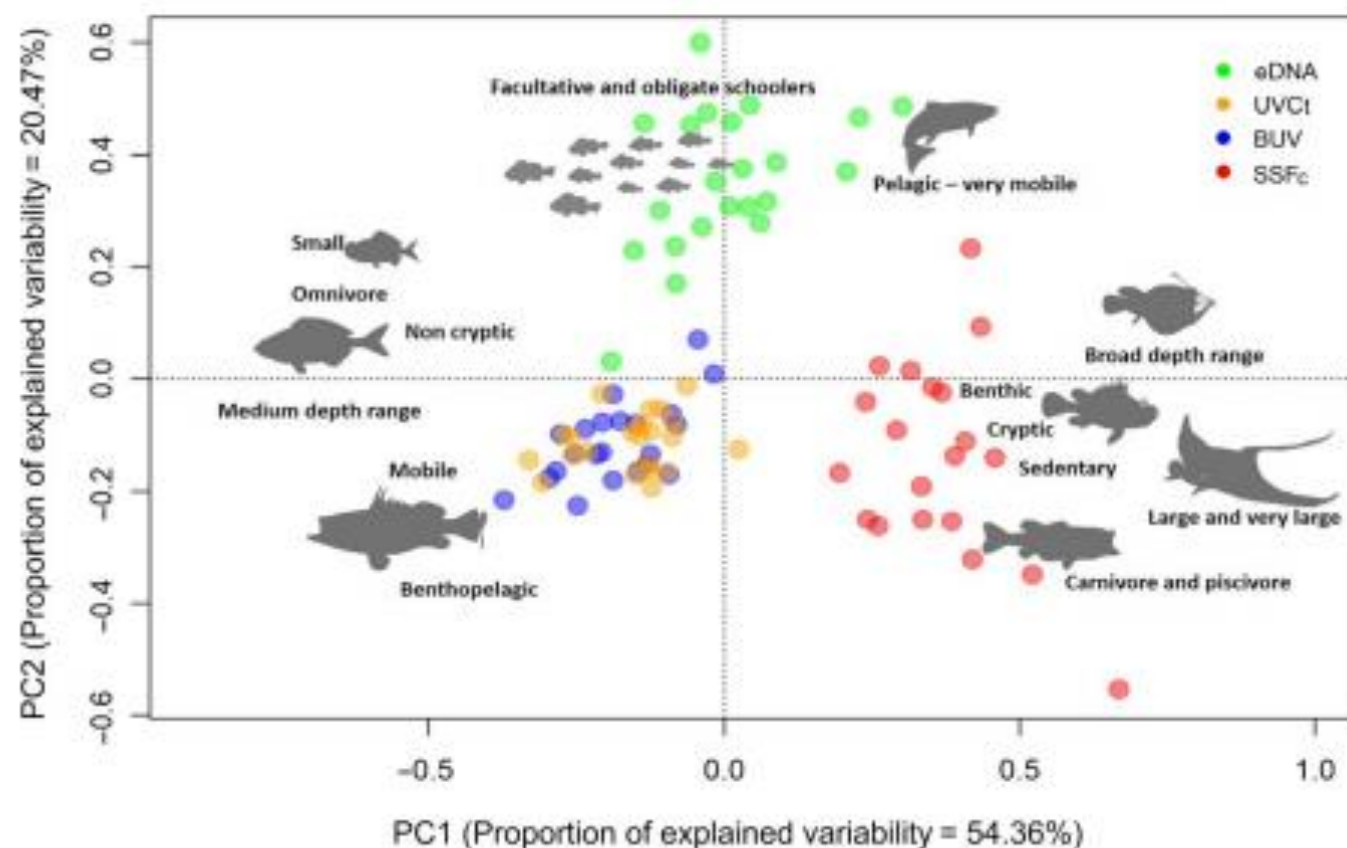


 <b>Introduction</b>	 Données	 PGD	 Flowchart	 Outils	 Résultats	 Retours	 Perspectives
1min	1min	1min	3min	3min	4min	1min	1min

**Table 1** A unified framework for quantifying species diversity, phylogenetic diversity, and functional diversity on the basis of Hill numbers (by treating each entity as a “species” with its relative abundance as indicated in the table)

Type of diversity	Element in collection $C$	Attribute value	Entity	Relative abundance of an entity	Attribute diversity (effective number of entities or total attribute value)	Generalized Hill numbers (effective number of species or lineages)
Attribute diversity (unified framework)	Element $u$ in collection $C$ , $u \in C$	$v_u$	An entity is an element with a unit of attribute value	$a_u / \hat{V}$ , where $a_u =$ weight or abundance, $\hat{V} = \sum_{u \in C} v_u a_u$	${}^q AD(\hat{V}) = \left[ \sum_{u \in C} v_u \times \left( \frac{a_u}{\hat{V}} \right)^q \right]^{1/(1-q)}$	${}^q D(\hat{V}) = \left[ \frac{{}^q AD(\hat{V})}{\hat{V}} \right]^{1/\lambda}$ $\lambda = 1$ or $\lambda = 2$ (see below)
Species diversity	Species $i$ , $C = \{1, \dots, S\}$	Unity for species $i$	Taxonomic entity (species)	$p_i$ , species relative abundance ( $\hat{V} = 1$ )	Species diversity (Hill numbers, effective number of species) ${}^q D = \left[ \sum_{i=1}^S 1 \times \left( \frac{p_i}{\sum_{k=1}^S p_k} \right)^q \right]^{1/(1-q)}$	Hill numbers ${}^q D$
Phylogenetic diversity	Branch segment $i$ , $C = \{1, \dots, B\}$	Branch length $L_i$ for branch $i$	Phylogenetic entity (branch of unit-length)	$a_i / \hat{T}$ , where $a_i =$ branch abundance, $\hat{T} = \sum_{j=1}^B L_j a_j$	Phylogenetic diversity (effective total branch-length) ${}^q PD(\hat{T}) = \left[ \sum_{i=1}^B L_i \times \left( \frac{a_i}{\hat{T}} \right)^q \right]^{1/(1-q)}$	Phylogenetic Hill numbers ${}^q \hat{D}(\hat{T}) = \frac{{}^q PD(\hat{T})}{\hat{T}}$
Functional diversity	Species-pair $(i, j)$ , $C = \{(i, j); i, j = 1, \dots, S\}$	Distance $d_{ij}$ for species-pair $(i, j)$	Functional entity (species-pair of unit-distance)	$p_i p_j / Q$ , where $Q = \sum_{i,j=1}^S d_{ij} p_i p_j$	Functional diversity (effective sum of species pairwise distances) ${}^q FD(Q) = \left[ \sum_{i,j=1}^S d_{ij} \times \left( \frac{p_i p_j}{Q} \right)^q \right]^{1/(1-q)}$	Functional Hill numbers ${}^q D(Q) = \left( \frac{{}^q FD(Q)}{Q} \right)^{1/2}$





**FIGURE 6** Principal component analysis (PCA) of the functional trait proportions of fish assemblages identified by the eDNA, UVCt, BUV and SSFc techniques. The first four dimensions of the PCA cumulatively explained 88.24% of the projected inertia in the distribution of fish species traits, 74.82% of which was explained by the first two axes. Each point refers to samples collected in MPAs and their flanking unprotected locations (i.e., a total of 22 locations). Correlations with main fish traits (represented by different fish shapes) are also superimposed. The original PCA graph is provided in Figure S5. Fish shapes are modified free of rights images. Sources: flyclipart.com, cleanpng.com, www.shareicon.net, netclipart.com, publicdomainvectors.org

