Proposed pipeline

- Collect marine? nifH amplicon sequences from NCBI and ENA
- Run DADA2 with the latest nifH DB of Molly
- Generate fasta file, taxonomic and count table
- Upload into R package Ampvis
- Identify and extract unidentified sequences (e.g., sequence only identified at phylum level)
- Diamond blast and add hits to the database (remember to note down which was added). Manually, add the taxonomic annotation (i.e. kingdom, phylum, order and so on). Not sure have this can be done automatically.

Alternatively

- Collect marine? nifH amplicon sequences from NCBI and ENA
- Run DADA2 with the latest nifH DB of Molly
- Generate fasta file, taxonomic and count table
- Upload into R package Ampvis
- Identify and extract unidentified sequences (e.g., sequence only identified at phylum level)
- Run through nifhMAP pipeline (already a automated pipeline)
- Add hits manually to Molly nifH DB and rerun amplicon.

What Mathéo did

Pour l'analyse de données de métabarcoding que j'ai réalisé :

- J'ai utilisé la base de données "nifH_dada2_all_v1.1.0.fasta" de Molly Moynihan (je ne manquerai pas de la citer dans l'article concerné par cette banque de donnée!).
- J'ai ensuite réalisé l'analyse de métabarcoding via dada2. En sortie, j'avais beaucoup d'ASV qui n'étaient pas affiliés très précisément ou carrément pas affiliés du tout.
- J'ai donc blasté chaque ASV n'ayant pas une classification suffisamment précise pour notre étude, contre la banque de données NR, en excluant les "Uncultured/environmental sample sequences" et en utilisant l'option "Highly similar sequences (megablast)".
- J'ai ensuite mis au format fasta + au format de la banque de donnée de Molly, les séquences obtenues après blast (paramètres très stringeants, notamment au niveau du % d'identité).
- J'ai intégré ces séquences dans la banque de données initiale "nifH_dada2_all_v1.1.0.fasta".
- J'ai de nouveau réalisé l'analyse de métabarcoding sur mes données avec la base de données de Molly enrichie.

Pre-step before loop

1) Filter unclassified seqs out from molly
    a) This can be done as a separate step using “filter function in R” or perhaps automatically

Ideas to integrate into automatically loop

1) Use molly create_fastq.R to write fasta file with accession number from **“pre-step”**
    a) Know we have a fasta file with all unclassified seqs
2) Run NCBI blast on each entry (accession nr.)
    a) https://blast.ncbi.nlm.nih.gov/doc/blast-help/downloadblastdata.html#blast-executables
    b) Determine a threshold => 90%
3) From each entries, extract subject accession and subject alignment to write a new fasta file
4) Save the new fasta with the updated entries
5) Add the updated files to molly DB version (the fasta file)
6) Loop finish

Post-step

1) Test the updated version on amplicon data set (the once we have)
    a) Check if there is fewer or more nifH annotated