Rapport TP Métagénomique

Valentin Goupille

M2 Bioinformatique

## Introduction

Les lacs méromictiques, caractérisés par des couches d'eau stratifiées qui ne se mélangent jamais, présentent des conditions environnementales particulières. Ces environnements aquatiques, avec des gradients de température et des variations de la concentration en oxygène, influencent de manière significative les communautés microbiennes qui les habitent. L'étude de ces communautés et de leurs fonctions permet de mieux comprendre comment les microbes interagissent avec leur environnement et contribuent à des processus écologiques spécifiques.

Dans cette étude, l'objectif principal est d'examiner la structure et les fonctions des communautés microbiennes présentes dans trois habitats distincts d'un lac méromictique. L'hypothèse sous-jacente est que chaque habitat abrite une communauté microbienne spécifique, adaptée aux conditions particulières de son environnement.

Pour atteindre cet objectif, deux approches métagénomiques sont utilisées : l'approche basée sur les lectures (read-based) et l'approche par assemblage de contigs. Ces méthodes permettent d'identifier les espèces microbiennes présentes dans chaque habitat et d'analyser leurs fonctions métaboliques associées.

## Matériels et méthodes

### 1- Matériel Biologique

Dans ce travail pratique, il est important de noter qu'aucune extraction d'ADN réelle et construction de librairie n'a été effectuée à partir d'échantillons environnementaux pour construire les métagénomes. Au lieu de cela, un jeu de données simplifié a été utilisé. Ce jeu repose sur la création de génomes « jouets » simulés, qui sont des ensembles de séquences génétiques composées de 8 à 10 gènes codant pour des enzymes aux fonctions connues, accompagnés d'un gène 16S rRNA. Ces génomes simulés ont ensuite été utilisés pour générer des métagénomes simplifiés, représentant la diversité microbienne de trois habitats distincts d'un lac méromictique. Cela permet de travailler sur une simulation contrôlée, facilitant ainsi l'exploration de la structure et des fonctions des communautés microbiennes dans ces environnements spécifiques, sans nécessiter d'échantillons réels.

Les habitats étudiés sont les suivants :

- **Mixolimnion (Habitat n°1)** : Il s'agit de la couche supérieure du lac, qui est bien oxygénée et en contact direct avec l'atmosphère. Cette zone reçoit la lumière solaire. Les échantillons simulés A, B et C proviennent de ce milieu.

- **Metalimnion (Habitat n°2)** : Cette couche intermédiaire du lac se situe entre le mixolimnion et le monimolimnion. Elle est caractérisée par une forte variation de température et de concentration en oxygène. Cette zone est généralement peu éclairée et présente des conditions aérobie-anoxiques. Les échantillons simulés D, E et F sont issus de cette zone.

- **Monimolimnion (Habitat n°3)** : Le monimolimnion correspond à la couche inférieure du lac, où les conditions sont anaérobies et où la température est plus stable. Cette zone est souvent pauvre en oxygène. Les échantillons simulés G, H et I proviennent de ce milieu.

## 2- Pipeline Bioinformatique

Le pipeline métagénomique utilisé pour étudier la structure et les fonctions des communautés microbiennes dans un lac méromictique repose sur deux stratégies complémentaires : l'analyse basée sur les reads (séquences brutes) et l'analyse basée sur les contigs (fragments d'ADN assemblés). Ces approches permettent d'obtenir une vue d'ensemble détaillée des communautés microbiennes présentes dans différents habitats du lac.

### I. Analyse basée sur les reads

**Préparation des données** : Les 9 métagénomes sont importés dans l'**environnement *Galaxy*** sous forme de fichiers FASTQ, représentant les séquences brutes obtenus à partir des échantillons simulés (technologie : Illumina MiSeq paired-end ; longueur des reads : 250 pb ; nombre de reads : 8000 par échantillon). Deux fichiers de référence sont également téléchargés : un fichier FASTA des séquences protéiques de référence et un tableau d'annotations fonctionnelles et taxonomiques, organisés pour faciliter l'analyse.

**Alignement des reads avec DIAMOND** : L'outil **DIAMOND**, utilisé en mode BLASTX, permet d'aligner les séquences de reads contre une base de données protéique. Avant cela, un fichier FASTA contenant les séquences protéiques de référence est converti en une base de données DIAMOND à l'aide de l'outil "Diamond makedb". L'alignement identifie les protéines correspondantes et génère des fichiers tabulaires.

**Analyse fonctionnelle** : Les résultats d'alignement sont associés à des annotations fonctionnelles via l'outil **Join two datasets**, liant chaque alignement à un nom de protéine. À l'aide de l'outil **Group**, le nombre de reads associés à chaque protéine est quantifié pour chaque métagénome, fournissant une estimation de leur abondance. Les fichiers résultants sont structurés en utilisant les outils **Add Header** et **Multi-Join**, avant d'être exportés sous format Excel.

**Normalisation et visualisation** : Un **script R** normalise les données en tenant compte de la longueur des gènes associés. Ces données sont ensuite visualisées sous forme de graphiques (barplots) pour illustrer les fonctions biologiques dominantes dans les différents échantillons et habitats, fournissant une vue d'ensemble des activités fonctionnelles des communautés microbiennes.

**Analyse taxonomique** : L'analyse commence par l'utilisation de l'outil **FASTQ de-interlacer** qui permet de transformer chaque fichier FASTQ interlacé en quatre fichiers distincts pour chaque métagénome : un fichier contenant les reads appariées "forward" (left mate), un fichier pour les reads appariées "reverse" (right mate), et deux fichiers pour les reads non appariées.
L'outil **Kraken** est utilisé sur les reads appariées, pour classer les séquences selon leur taxonomie, en se basant sur une base de données dédiée aux bactéries. Les résultats sont ensuite traités par **Kraken-mpa-report**, générant des profils taxonomiques détaillés. Ces resultats sont copiés dans un tableur xlsx, où les en-têtes des colonnes sont modifiés pour correspondre aux noms des métagénomes. Un script R est utilisé pour créer un barplot à partir de ce tableau, offrant une comparaison de l'abondance relative des différents genres bactériens entre les échantillons et les habitats du lac, afin de visualiser les différences entre les communautés bactériennes des trois habitats.

## II. Analyse basée sur les contigs

**Préparation des données** : Les 9 métagénomes sont importés dans l'**environnement *Galaxy*** sous forme de fichiers FASTQ, ainsi qu'un fichier FASTA qui contient des séquences protéiques de référence utilisées comme priorités pour l'annotation des contigs.

**Assemblage des reads en contigs** : L'outil **SPAdes** est utilisé pour assembler les reads en contigs, permettant de reconstituer des séquences d'ADN plus longues à (contigs) partir des fragments initiaux (reads). Cet assemblage est réalisé sous forme de co-assemblage en utilisant les 9 métagénomes comme données d'entrée.

**Annotation des contigs** : Une fois l'assemblage effectué, les *contigs* sont annotés à l'aide de l'outil ***Prokka***, qui permet d'identifier les gènes codants et d'attribuer des fonctions aux *contigs*. Dans ce cas, l'option "–metagenome" est utilisée pour adapter le logiciel aux spécificités des métagénomes fragmentés. Les résultats de l'annotation incluent des fichiers contenant les séquences traduites des protéines sous format *FAA*, ainsi que des tableaux d'annotations au format *GFF3*. Ces annotations permettent d'identifier des gènes codant pour des ARN ribosomiques, des enzymes spécifiques, et d'autres éléments fonctionnels importants pour la compréhension des processus biologiques dans les communautés microbiennes

**Analyse taxonomique des contigs** : Après l'annotation des contigs, chaque contig est classifié selon son origine taxonomique en utilisant deux approches complémentaires :

1. **Identification des séquences ribosomiques 16S** : Les séquences ribosomiques 16S présentes dans les contigs sont extraites et soumises à un **BLAST sur NCBI**. Cette étape permet d'identifier le genre correspondant à chaque séquence ribosomique.

2. **Affectation de la taxonomie à partir des protéines codées** : La taxonomie d'un contig est également déterminée en examinant certaines protéines codées par ce contig. Pour chaque protéine sélectionnée, l'identifiant RefSeq est recherché dans la base de données **NCBI RefSeq**, afin de confirmer le taxon associé. Cette approche, appliquée à un échantillon restreint de protéines, permet de valider la classification taxonomique du contig, en croisant les informations issues des deux méthodes.

**Couverture des contigs et visualisation** :
La couverture des contigs est une mesure essentielle pour évaluer l'abondance relative des séquences dans les échantillons. Pour cela, les reads des métagénomes sont alignés sur les contigs assemblés à l'aide de l'outil **Bowtie2**. L'alignement génère un fichier au format **SAM/BAM**, qui contient des informations détaillées sur la position des reads sur les contigs.

Une fois l'alignement réalisé, l'outil **Samtools depth** est utilisé pour calculer la profondeur de couverture (X) de chaque contig dans chaque métagénome, c'est-à-dire le nombre de fois que chaque position du contig est couverte par les reads. Cela permet d'estimer l'abondance relative des contigs dans les échantillons. Ensuite, l'outil **Group** de Galaxy est utilisé pour regrouper les données et obtenir un tableau unique, calculant la couverture moyenne par contig sur l'ensemble des métagénomes.

La couverture moyenne des contigs est ensuite visualisée à l'aide de **R**, ce qui permet d'analyser la répartition de l'abondance des contigs dans les différents métagénomes et habitats du lac.

**Analyse et discussion des résultats :**

1. **Structure des communautés dans les trois habitats à partir des informations taxonomiques et de couverture (Figures 1 et 2)**

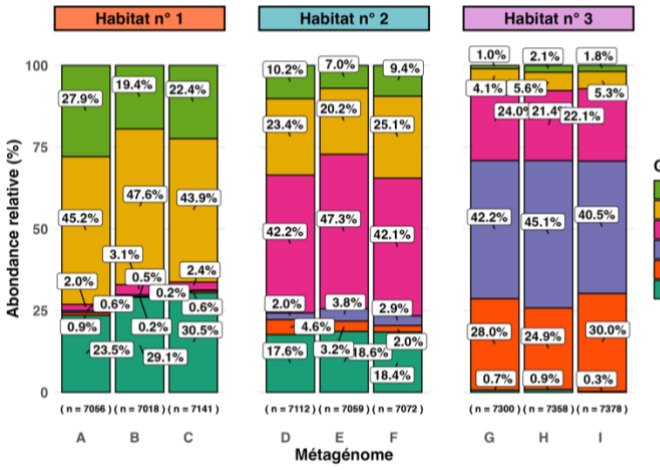


Figure 1 : Abondance relative des genres microbiens dans trois habitats distincts d'un lac méromictique (Mixolimnion, Metalimnion et Monimolimnion). Chaque habitat comprend trois métagénomes (1 : ABC ; 2 : DEF ; 3 : GHI). Les genres microbiens représentés incluent *Synechocystis*, *Nitrosomonas*, *Methylococcus*, *Methanospirillum*, *Desulfovibrio* et *Anabaena*.

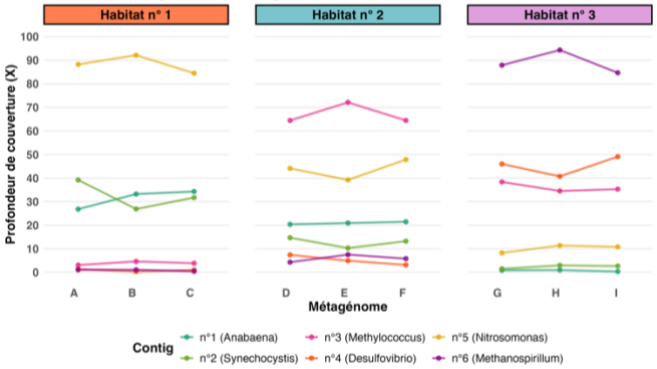


Figure 2 : Profondeur moyenne de couverture (X) des différents contigs co-assemblés à partir des 9 métagenomes dans trois habitats distincts d'un lac méromictique (Mixolimnion, Metalimnion et Monimolimnion). Chaque habitat comprend 3 métagénomes (1: ABC; 2: DEF; 3: GHI). Les six contigs analysés sont associés aux genres bactériens suivants : *Synechocystis*, *Nitrosomonas*, *Methylococcus*, *Methanospirillum*, *Desulfovibrio* et *Anabaena* (données non présentées).

La structure des communautés microbiennes diffère nettement entre les trois habitats du lac méromictique, comme le montrent les données taxonomiques (Figure 1) et de couverture des contigs (Figure 2). Ces différences reflètent l'adaptation des microorganismes aux gradients environnementaux spécifiques de chaque habitat.

**Mixolimnion :**

- **Taxonomie (Figure 1)** : Cette zone est dominée par les genres ***Nitrosomonas*** (45-47 %), suivis par les cyanobactéries ***Synechocystis*** (19-28 %) et ***Anabaena*** (24-31 %). Les autres genres, comme ***Methylococcus***, ***Methanospirillum***, et ***Desulfovibrio***, sont présents en très faibles proportions (< 3 %).

- **Couverture des contigs (Figure 2)** : Le contig 5 (correspondant à ***Nitrosomonas***) a la couverture la plus élevée (environ 85x), suivi des contigs 1 (***Anabaena***) et 2 (***Synechocystis***) avec des couvertures autour de 30x. Les autres contigs (3, 4 et 6) ont des contributions marginales (< 10x).

- **Structure globale** : La communauté est dominée par des microorganismes autotrophes, principalement des cyanobactéries et des bactéries nitrifiantes comme ***Nitrosomonas***, adaptés aux conditions bien éclairées et oxygénées.

**Metalimnion :**

- **Taxonomie (Figure 1)** : Le genre dominant est ***Methylococcus*** (42-47 %), suivi de **Nitrosomonas** (20-25 %), et des cyanobactéries ***Anabaena***, et ***Synechocystis*** (18 % chacun). Les genres ***Methanospirillum*** et ***Desulfovibrio*** sont plus abondants que dans le Mixolimnion mais restent minoritaires (2-4 %).

- **Couverture des contigs (Figure 2)** : Les contigs 3 (***Methylococcus***) et 5 (***Nitrosomonas***) sont les plus abondants avec des couvertures respectives de 70x et 45x. Les cyanobactéries (***Anabaena*** et ***Synechocystis***) ont des contributions secondaires (20x et 12x), et les contigs 4 et 6 ont des couvertures marginales (< 10x).

- **Structure globale** : La communauté se transforme en une zone intermédiaire où les bactéries méthanotrophes comme ***Methylococcus*** et les bactéries nitrifiantes coexistent avec une faible contribution des cyanobactéries. Cela reflète une transition vers des conditions micro-oxiques.

**Monimolimnion :**

- **Taxonomie (Figure 1)** : Les genres dominants sont ***Methanospirillum*** (40-45 %) et ***Desulfovibrio*** (25-30 %), suivis de ***Methylococcus*** (21-24 %). Les genres comme **Nitrosomonas**, ***Synechocystis***, et ***Anabaena*** deviennent rares (< 6 %).

- **Couverture des contigs (Figure 2)** : Le contig 6 (***Methanospirillum***) a la couverture la plus élevée (85x), suivi des contigs 4 (***Desulfovibrio***, 45x) et 3 (***Methylococcus***, 35x). Les contigs associés aux cyanobactéries et à ***Nitrosomonas*** ont des contributions négligeables (< 10x).

- **Structure globale** : La communauté est dominée par des microorganismes anaérobies spécialisés dans la méthanogénèse (***Methanospirillum***) et la réduction du soufre (***Desulfovibrio***). Ces résultats reflètent des conditions anoxiques et riches en composés réduits.

**2. Métabolisme des microorganismes à partir de l'annotation des gènes codant des protéines (Figure 3)**

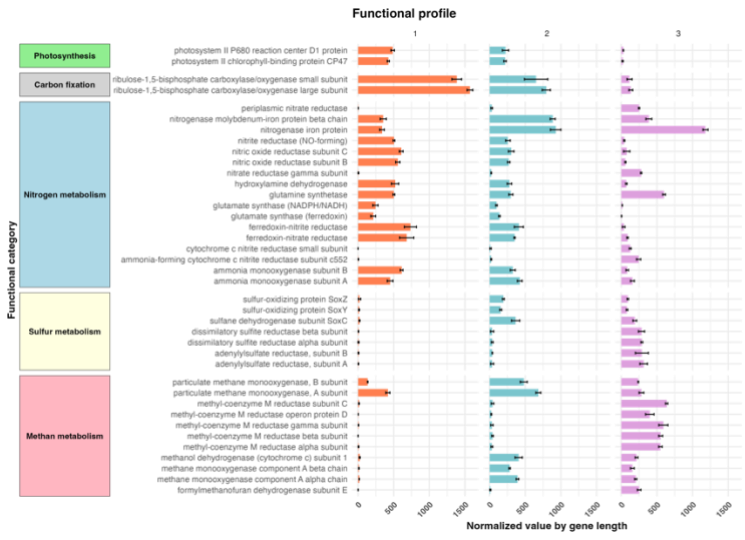


Figure 3 : Profil fonctionnel des communautés microbiennes dans trois habitats d'un lac méromictique (1: Mixolimnion, 2: Metalimnion et 3: Monimolimnion), en fonction des voies métaboliques dominantes (photosynthèse, fixation du carbone, métabolismes de l'azote, du soufre et du méthane) et des abondances des protéines associées, normalisées par la longueur des gènes.

`

La figure 3 illustre le profil fonctionnel des communautés microbiennes dans trois habitats distincts d'un lac méromictique (mixolimnion, metalimnion et monimolimnion), en lien avec les principaux processus métaboliques. Ces résultats reflètent les adaptations des microorganismes aux gradients environnementaux spécifiques de chaque habitat, en cohérence avec les données taxonomiques (Figure 1) et de couverture des contigs (Figure 2).

Dans le **mixolimnion**, les conditions éclairées et oxygénées favorisent des processus tels que la photosynthèse, fortement représentée par l'abondance des protéines des photosystèmes (P680 et protéines liant les chlorophylles), en accord avec la dominance des cyanobactéries (*Synechocystis* et *Anabaena*). La fixation du carbone, catalysée par l'enzyme RubisCO, est également prédominante dans cet habitat autotrophe. De plus, la forte présence d'enzymes impliquées dans la nitrification (ammoniac monooxygénase) reflète l'activité de bactéries nitrifiantes telles que *Nitrosomonas*. En revanche, les métabolismes du soufre et du méthane sont peu représentés, ce qui correspond à la faible abondance de genres tels que *Desulfovibrio* et *Methanospirillum*.

Le **metalimnion** constitue une zone de transition micro-oxique où les profils métaboliques sont plus diversifiés. La photosynthèse et la fixation du carbone, bien que présentes, diminuent en importance en raison de la baisse des cyanobactéries. Les enzymes liées à la nitrification et à la dénitrification montrent une activité mixte, reflétant la coexistence de Nitrosomonas et d'autres microorganismes adaptés à des conditions moins oxygénées. Par ailleurs, la méthanotrophie devient un processus clé, comme en témoigne l'abondance accrue d'enzymes telles que la méthane monooxygénase associée à *Methylococcus*.

Dans le **monimolimnion**, les conditions anoxiques favorisent les processus métaboliques anaérobies. La photosynthèse et la fixation du carbone sont quasi absentes, reflétant la rareté des cyanobactéries. Le métabolisme du soufre joue également un rôle central, comme le montre la forte abondance de la sulfite réductase, caractéristique de *Desulfovibrio*. Enfin, la méthanogenèse est le processus prédominant, avec une activité marquée de la méthyl-coenzyme M réductase, confirmant la dominance de *Methanospirillum* dans cet habitat.

Ces profils fonctionnels mettent en évidence une stratification métabolique claire dans le lac méromictique, directement liée aux gradients environnementaux spécifiques des trois habitats. Ils révèlent l'adaptation des microorganismes aux variations de lumière, d'oxygène et de disponibilité des substrats énergétiques le long de la colonne d'eau.

## Conclusion

Cette étude sur les communautés microbiennes des trois habitats distincts d'un lac méromictique a permis de mettre en évidence la diversité et les spécificités fonctionnelles des communautés microbiennes en réponse aux gradients environnementaux propres à chaque habitat. Les résultats obtenus à partir des analyses taxonomiques et de couverture des contigs confirment l'adaptation des micro-organismes à leurs environnements respectifs : les cyanobactéries et les bactéries nitrifiantes prédominent dans le mixolimnion oxygéné et lumineux, tandis que le metalimnion, plus complexe, est dominé par des bactéries méthanotrophes et nitrifiantes. En revanche, le monimolimnion, avec ses conditions anoxiques, abrite une communauté microbienne adaptée aux processus méthanogènes et de réduction du soufre.

L'analyse fonctionnelle a montré que les voies métaboliques dominantes étaient en cohérence avec la structure taxonomique observée. La photosynthèse, la fixation du carbone et les métabolismes de l'azote, du soufre et du méthane ont joué un rôle central dans la survie et l'activité des micro-organismes. Ces résultats soulignent l'importance de l'environnement dans la structuration des communautés microbiennes et leur adaptation aux conditions physico-chimiques locales.

En conclusion, cette étude démontre l'utilité des approches métagénomiques pour explorer la diversité microbienne dans des écosystèmes complexes, et met en lumière les rôles écologiques spécifiques des micro-organismes dans chaque habitat d'un lac méromictique. Ces connaissances peuvent servir de base pour des études futures sur les interactions microbiennes et leur contribution aux cycles biogéochimiques dans des environnements aquatiques stratifiés.