Phytoplancton des plans d'eau d'Outre-Mer : développement d'outils de monitoring basés sur l'ADN - Livrable 2 : "Phytoplancton DOM : première application du protocole de terrain et comparaisons entre microscopie/ADNe "

Alexis Canino, Christophe Laplace-Treyture, Agnes Bouchez, Domaizon Isabelle, Frédéric Rimet

Phytoplancton des plans d'eau d'Outre-Mer : développement d'outils de monitoring basés sur l'ADN.

Livrable 2 : «Phytoplancton DOM : première application du protocole de terrain et comparaisons entre microscopie/ADNe »

CANINO Alexis*[1,3], LAPLACE-TREYTURE Christophe*[2,3],
BOUCHEZ Agnès*[1,3], DOMAIZON Isabelle*[1,3], RIMET Frédéric*[1,3]

*1 INRAE - UMR CARRTEL, F-74500 Thonon-les-Bains, France
*2 INRAE - UR EABX, F-33612 Cestas, France
*3 Pôle R&D ECLA

12 - 2021

Pôle R&D ECLA

Site INRAE d'Aix-en-Provence
3275 route Cézanne – 13100 Le Tholonet

https://professionnels.ofb.fr/fr/pole-ecla-ecosystemes-lacustres

SOMMAIRE

Pôle R&D ECLA

Site INRAE d'Aix-en-Provence
3275 route Cézanne – 13100 Le Tholonet

https://professionnels.ofb.fr/fr/pole-ecla-ecosystemes-lacustres

I. Contexte

I.1 Rappels brefs des travaux réalisés en 2020

Les investigations réalisées en 2020 ont permis d'établir une liste taxonomique des taxa présents dans les plans d'eau d'Outre-Mer. Elles ont également permis de mettre en évidence un besoin d'homogénéisation entre les nomenclatures taxonomiques du phytoplancton utilisées par les bureaux d'étude et dans les différents domaines (*i.e.* biologie moléculaire et microscopie). Une méthode a été mise en place pour pallier ce biais, ce qui était nécessaire pour mener à bien ce projet et constituer des bases de références robustes et comparables entre-elles. Ce point est rediscuté dans ce livrable (*cf.* §V) puisqu'il a abouti à la mise en place d'une web-application d'homogénéisation taxonomique (Canino et *al*. 2021).

Parallèlement à ces travaux, les résultats obtenus l'an passé portaient majoritairement sur l'efficacité *in silico* de différents couples d'amorces. En effet, une stratégie avait été développée pour rechercher des régions conservées, avec la possibilité de modifier des paramètres dans le but de dessiner des couples d'amorces pour le métabarcoding de taxa phylogénétiquement diversifiés (dans notre cas l'ensemble de la communauté phytoplanctonique). Cette stratégie avait été appliquée à différents marqueurs génétiques dans le but de trouver des régions candidates comme barcodes pour le métabarcoding du phytoplancton d'eau douce. L'ensemble des tests in silico conduits a permis de réaliser une pré-selection et d'avoir une vision d'ensemble des différents barcodes potentiels et leurs couples d'amorces associés. La sélection de ceux qui ont pu être testés en laboratoire en 2021 a été réalisée en amont et est détaillée dans la suite de ce livrable (*cf.* §III).

Enfin, un dernier point consistait à mettre en place les protocoles à fournir aux bureaux d'études des départements d'Outre-Mer, en vue des campagnes d'échantillonnage de 2021.

I.2 Principaux objectifs pour 2021

Les principaux objectifs annoncés pour cette année étaient :

- **la mise en place des campagnes d'échantillonnage dans les DROM.**

Cela implique la prise de contact avec les bureaux d'études des différents DROM, leur accord pour mutualiser les échantillonnages demandés avec ceux prévus dans le cadre de la DCE, la préparation et l'envoi du matériel ainsi que la communication du protocole correspondant et *in fine* le rapatriement des échantillons au laboratoire (INRAE UMR Carrtel).

- **la sélection du/des barcode(s) candidat(s) pour les tests *in vitro*.**

Le choix s'appuie sur les travaux réalisés *in silico* sur les différents couples d'amorces pour chacun des marqueurs génétiques investigués.

- **les tests *in vitro* des couples d'amorces sélectionnés.**

Ceci implique d'utiliser des communautés synthétiques sur lesquelles s'appuyer comme échantillons témoins (*i.e*. « mock ») : pour cela des souches de la collection de culture de Thonon (appelée TCC par la suite) ont été utilisées. Une fois la validation de l'amplification du barcode pour ces cultures (et autres échantillons) avec les différents couples d'amorces, un run de séquençage a été réalisé dans le but de déterminer l'efficacité de chacun de ces couples (*cf*. §III).

- **la comparaison des résultats obtenus en microscopie et moléculaire.**

Il s'agit ici de comparer les résultats du run de séquençage des différents échantillons avec ce qui est attendu *a priori* (*i.e*. communautés témoins) et avec ce qui est observé en microscopie (pour un échantillon environnemental). Cela permettra de conclure sur l'efficacité des amorces à produire des inventaires similaires aux observations en microscopie et in fine de définir la stratégie à utiliser en 2022 sur l'ensemble des échantillons des DROM.

En parallèle de ces quatre objectifs, une tâche importante était de nettoyer et optimiser les scripts bio-informatiques d'intérêt (réalisés en 2020) dans le but de les rendre facilement accessibles à toutes personnes en ayant besoin. Les curations des bases de données devaient se poursuivre afin d'être intégrées dans une même application regroupant l'ensemble de ces outils informatiques essentiels à la mise en place de ce projet. Cette application développée sous forme d'une ShinyApp a vocation à poursuivre l'intégration, de manière progressive, des outils développés tout au long du projet afin qu'ils soient disponibles en open-source (*cf.* §V).

II. Mise en place des campagnes d'échantillonnage

Dans le but de récolter des échantillons dans les DROM au cours de cette année, les Offices de l'Eau et Deal, ont été contactés pour avoir la permission de solliciter les bureaux d'études responsables des suivis DCE dans les DOM. Afin de bénéficier également des comptages du phytoplancton réalisés dans le cadre de ces suivis, les prélèvements ont été mutualisés avec les campagnes DCE. Le protocole ainsi que l'ensemble du matériel nécessaire pour les prélèvements de phytoplancton sur le terrain ont été fournis aux bureaux d'étude ayant répondu favorablement à la requête. Les bureaux d'étude ayant été contactés sont les suivants :

- Hydrecolab, s'occupant de la retenue de Petit Saut en Guyane et de la retenue de Gaschet en Guadeloupe
- Oce Consult, échantillonnant plusieurs plans d'eau et rivière à La Réunion dans le cadre de la DCE mais aussi du projet Transphyt.

L'ensemble des sites d'échantillonnage et le nombre d'échantillons associés sont présentés sur le Tableau 1. Les protocoles ont été fournis en format .pdf (un aperçu est disponible en Mat. Supp. 1) et une vidéo explicative a également été réalisée. Ce protocole a pour but d'assurer un prélèvement de phytoplancton dans les plans d'eau en limitant les risques de contamination et est entièrement réalisable sur le terrain en une dizaine de minutes. Il s'appuie sur le protocole utilisé dans le cadre du projet EcoAlpsWater mis en place par Domaizon et *al*. (2019), disponible en ligne. Il utilise des cartouches filtrantes (Sterivex™ avec une porosité de 0.45µm) qui seront rapatriées en fin d'année 2021 et depuis lesquelles l'ADN sera extrait en laboratoire.

Tableau 1 : répartition des différents sites d'échantillonnage prévus pour 2021. Entre parenthèses est indiqué le nombre de stations relatives à chaque site et dans la colonne d'en face, le nombre d'échantillons attendu pour ce site.

<table>
<tr><th colspan="2">La Réunion</th><th>Guyane</th><th colspan="2">Guadeloupe</th><th colspan="2">Métropole suivis OLA* 2020-2021</th></tr>
<tr><td rowspan="2">St Paul (3)</td><td rowspan="2">36</td><td>Dégrad Corrèze (1)</td><td rowspan="10">Gaschet (1)</td><td rowspan="10">6</td><td rowspan="2">Aiguebelette</td><td rowspan="2">8</td></tr>
<tr><td>Patagai (1)</td></tr>
<tr><td>Gol (3)</td><td>36</td><td>Aval Apatou (1)</td><td>Annecy</td><td>16</td></tr>
<tr><td rowspan="2">Rivière
Ste Suzanne</td><td rowspan="2">12</td><td>Angoulême (1)</td><td rowspan="2">Bourget</td><td rowspan="2">19</td></tr>
<tr><td>Athanase (1)</td></tr>
<tr><td rowspan="2">Grand-Etang (1)</td><td rowspan="2">4</td><td>Twenke (1)</td><td rowspan="4">Léman</td><td rowspan="4">21</td></tr>
<tr><td>Papaichton (1)</td></tr>
<tr><td rowspan="2">Bois Rouge (1)</td><td rowspan="2">12</td><td>Langatabiki (1)</td></tr>
<tr><td>Fourmi (1)</td></tr>
<tr><td>TOTAL</td><td>100</td><td>9</td><td></td><td>6</td><td></td><td>64</td></tr>
</table>

*Observatoire des Lacs Alpins

III. Sélection des marqueurs génétiques et couples d'amorces

III.1. Sélection sur les bases *in silico*

A l'issue des résultats produits l'année précédente, il a été décidé de s'orienter vers le choix de deux marqueurs génétiques : l'**ARNr 16S** et l'**ARNr 23S** (*i.e*. respectivement la petite et grande sous-unité ribosomale des procaryotes). A titre de rappel, ces marqueurs, bien que propres à des organismes procaryotes, peuvent également être détectés chez des organismes eucaryotes grâce à la présence d'endobiontes procaryotes (tels que les chloroplastes pour le phytoplancton). C'est donc à l'intérieur du génome des chloroplastes que se trouvent les marqueurs génétiques cibles. Ces derniers contiennent des barcodes d'intérêt pour le métabarcoding du phytoplancton d'eau douce. En effet, les investigations autour du marqueur *rbc*L ont montré que, bien qu'il soit résolutif, ce dernier ne présentait pas de couple d'amorces universelles adapté à toute la diversité du phytoplancton. Il en va de même pour le *tuf*A. Même si des couples d'amorces ont pu être dessinés, le nombre d'espèces possédant des séquences pour ce gène est beaucoup trop faible. Cela diminue également les chances que les amorces trouvées soient adaptées pour le métabarcoding de l'ensemble de la diversité phytoplanctonique.

L'ARNr 18S est le marqueur qui offre le plus grand nombre de séquences associées à des espèces phytoplanctoniques (avec une résolution d'environ 2000 espèces *in silico* pour certains couples d'amorces dont certaines caractéristiques laissent penser que les résultats *in vitro* ne se révéleront pas aussi fructueux). Cependant, ce dernier exclut les cyanobactéries (présentant un rôle clé pour la bioindication des plans d'eau) mais aussi le phylum des Euglenozoa, pour la région ciblée de l'ARNr18S (ici la région variable v4). Cette région a été ciblée car c'est celle qui offrait la possibilité d'avoir des couples de amorces universelles (pour le phytoplancton d'eau douce). C'est aussi celle qui est la plus utilisée dans les études de métabarcoding sur le phytoplancton avec les amorces développées par Stoeck et *al*, 2010 et produisant des amplicons dont la taille est compatible avec les technologies de séquençage les plus courantes : MiSeq (2x300 pb ou 2x250pb). Toutes ces informations ont conduit à se focaliser davantage sur les autres marqueurs ribosomaux.

L'ARNr 23S présente des caractéristiques qui font de lui un bon candidat pour le métabarcoding du phytoplancton d'eau douce : un simple couple d'amorces suffit à cibler l'ensemble de la diversité phytoplanctoniques. Les tests *in silico* réalisés en 2020 ont mis en avant la capacité discriminante d'une région spécifique : le domaine V du 23S appelé aussi UPA (Universal Plastidial Amplicon). Cette région peut être amplifiée par un couple d'amorces (comme évoqué précédemment, mais d'autres couples sont également disponibles pour cette même région) et est déjà bien documentée comme une région d'intérêt pour le barcoding d'organismes photosynthétiques (Presting, 2006 ; Sherwood & Presting, 2007). De plus, des références bibliographiques récentes montrent le bon potentiel de cette région (Djemiel et *al*., 2020 ; Qiao et *al*., 2021 ; Gorzerino 2021) pour des applications similaires. L'intérêt croissant pour ce marqueur est assez encourageant car son principal défaut réside dans la pauvreté des bibliothèques de références ADN, ce qui en soit n'est pas un obstacle puisque ces bibliothèques pourront être complétées à l'avenir. C'est pour ces raisons que le 23S a été sélectionné pour les tests de cette année.

Enfin, l'ARNr 16S, déjà très utilisé pour le phytoplancton sera également utilisé ici. Il permettra de réaliser une comparaison avec le 23S en terme d'efficacité et servira également d'alternative dans le cas où ce dernier s'avérerait inefficace ou problématique lorsque les références sont absentes. En effet, le 16S possède plus de références dans les bibliothèques de barcodes que le 23S. Cependant les investigations *in silico* réalisées en 2020 ont montré une plus faible résolution spécifique des régions du 16S par rapport à l'UPA du 23S.

III.2. Spécificité des amorces pour le phytoplancton

Puisque les marqueurs candidats retenus pour ce projet sont tous deux des ARN ribosomaux codant pour des sous-unités ribosomales procaryotes, un critère important à retenir est que les amorces permettent d'exclure les bactéries hétérotrophes. Il est ainsi possible de ne conserver que les autotrophes (*i.e*. cyanobactéries) et les chloroplastes cibles des organismes eucaryotes. Les tests réalisés, PCR *in silico* sur un programme écrit en shell utilisant des commandes de Mothur (Schloss et *al*., 2009) et RDP ProbeMatch (http://rdp.cme.msu.edu/probematch/search.jsp), montrent une très

forte spécificité aux chloroplastes et cyanobactéries pour l'ensemble des couples d'amorces pour la région UPA de l'ARNr 23S. Pour l'ARNr 16S, le couple CYA359F/CYA781R présente cette même spécificité, c'est d'ailleurs ce pour quoi il a été dessiné par Nübel et *al*. (1997). Les autres couples d'amorces donnés dans la littérature pour l'ARNr 16S ne sont pas spécifiques aux taxa phytoplanctoniques, et amplifient beaucoup de bactéries hétérotrophes, ce qui peut être problématique (*e.g*. saturation de la profondeur de séquençage). Leurs avantages sont au niveau de la couverture et de la résolution qu'ils offrent et, *a priori*, des caractéristiques optimales *in vitro*. Un couple d'amorces non-spécifique a, malgré tout, été utilisé pour les tests *in vitro* dans le but de mettre en évidence, concrètement, les potentiels problèmes liés à l'utilisation de celui-ci dans le cadre de ce projet.

III.3. Sélection définitive pour les tests in *vitro*

Sur la base des résultats et critères donnés en §III.1 et §III.2, 5 couples d'amorces différents ont été retenus pour les tests en laboratoire ; ces derniers sont présentés dans le tableau 2. Les couples relatifs à l'**ARNr 16S** qui ont été testés *in vitro* sont :

- '**PhytoF'/'PhytoR'** (couple nommé 16S_PHY / 16PHY dans les analyses) :
bon compromis entre sa résolution et ses caractéristiques optimisant l'amplification *in silico*. Il n'est cependant pas spécifique aux organismes phytoplanctoniques et sera susceptible d'amplifier des bactéries. Ces amorces visent à amplifier les régions variables v5 et v6, avec une taille moyenne des amplicons attendue de 386 pb, amorces incluses ;

- **CYA359F/CYA781R** (couple nommé 16S_CYA / 16CYA dans les analyses) :
déjà utilisé couramment dans d'autres études, il présente l'avantage d'être bien spécifique aux cyanobactéries et chloroplastes. Ses caractéristiques, quant à l'amplification, semblent être correctes ; il a récemment été utilisé pour une étude quasi-similaire (Ivanova et *al.*, 2019). Ces amorces visent à amplifier les régions variables v3 et v4, avec une taille moyenne des amplicons attendue de 425 pb, amorces incluses.

En ce qui concerne l'**ARNr 23S** les couples testés *in vitro* sont :

- **test587F/test587R** (couple nommé 23S_587 / 587 dans les analyses) :
in silico il possède les meilleurs critères parmi les différents couples d'amorces investigués optimisant l'amplification par PCR. Ces amorces visent à amplifier la région variable UPA, avec une taille moyenne des amplicons attendue de 408 pb, amorces incluses ;

- **test108F/test108R** (couple nommé 23S_108 / 108 dans les analyses) :
très similaire au couple précédent, mais celui-ci présente davantage de correspondances exactes avec les séquences cibles. Ces amorces visent à amplifier la région variable UPA, avec une taille moyenne des amplicons attendue de 402 pb, amorces incluses ;

- **p23SrV_f1/p23SrV_r1** (couple nommé 23S_SHE / SHE dans les analyses) :
afin d'avoir un contrôle ayant déjà été utilisé dans la littérature, il présente également des températures d'hybridation plus proches et plus optimales pour la PCR que les amorces de Yoon et *al.,* 2016. Ces amorces amplifient également la région UPA et produisent des amplicons d'une taille moyenne attendue de 408 pb, amorces incluses.

Tableau 2 : récapitulatif des amorces testées in vitro.

Couples d'amorces	Forward	Reverse	Référence
PhytoF/PhytoR [16S v5-v6]	GKAGCGGTGAAATGCGTAGAK	GCTGACGACAGCCATGCA	Ce projet
CYA359F/CYA781R [16S v3-v4]	GGGGAATYTTCCGCAATGGG	GACTACWGGGGTATCTAATCCCWTT	Nübel et *al.*, 1997
test587F/test587R [23S UPA]	GACAGWAAGACCCTATGAAGCT	ATCAGCCTGTTATCCCTAGAG	Ce projet
test108F/test108R [23S UPA]	ACAGWAAGACCCTATGAAGCTT	CCTGTTATCCCTAGAGTAACTT	Ce projet

p23SrV_f1/p23SrV_r1 [23S UPA]	GGACAGAAAGACCCTATGAA	TCAGCCTGTTATCCCTAGAG	Sherwood & Presting, 2007

IV. Premiers tests *in vitro*

IV.1. Déroulement des tests en laboratoire

Les tests *in vitro* des différents couples d'amorces se sont appuyés dans un premier temps sur la validation de l'amplification de différentes souches phytoplanctoniques (*cf.*§IV.2). Ces dernières, phylogénétiquement diversifiées, ont été utilisées dans le but d'évaluer l'efficacité d'amplification pour une large communauté phytoplanctonique à disposition grâce à la collection de microalgues du laboratoire (TCC https://www6.inrae.fr/carrtel-collection/).

Une fois cette étape réalisée avec succès, les couples d'amorces en question ont pu être utilisés pour le séquençage. Les amorces utilisées ont été fusionnées à une queue moléculaire demandée par la plateforme PGTB (Plateforme Génome Transcriptome de Bordeaux) assurant le séquençage. Cette plateforme fournit en effet un protocole à suivre pour l'envoi des échantillons destinés au séquençage sur leur plateforme (protocole disponible en Mat. Supp. 2). La technologie de séquençage choisie pour cette expérience est le MiSeq Reagent Kit v2 Nano (2x250 pb, output : 500Mb), appelé 'NanoMiSeq' dans la suite de ce rapport. Cette technologie a été choisie car adaptée au nombre d'échantillons relativement faible et la disponibilité d'un créneau de séquençage dans des délais courts (en comparaison à un run MiSeq classique). Sa limite réside dans sa capacité en sortie de données : 500Mb (contre 13.2-15Gb pour un séquençage MiSeq 2x300pb classique), qui reste cependant tout à fait suffisant pour le test présenté ici.

Les objectifs de ce séquençage NanoMiSeq sont les suivants :

- évaluer l'efficacité et la spécificité des différents marqueurs amplifiés par différents couples d'amorces pour les différents groupes biologiques constituant le phytoplancton. À l'issue de ce séquençage, les meilleures amorces seront retenues pour la suite du projet.
- Évaluer la réplicabilité des résultats obtenus en moléculaire (à l'aide de triplicats sur des échantillons connus) ;
- réaliser une comparaison des inventaires taxonomiques obtenus en moléculaire à ceux obtenus en microscopie (échantillon environnemental), ou à ceux attendus (échantillons témoins composés d'un mélange de cultures connues) ;
- obtenir un aperçu des potentiels impacts de différentes techniques de filtration (filtre ouvert *vs* filtre fermé Sterivex) et de différentes méthodes d'extractions d'ADN (GenElute *vs* NucleoSpin Soil MACHEREY-NAGEL) peuvent occasionner sur la composition taxonomique d'un même échantillon environnemental ;
- obtenir de nouveaux barcodes de référence à partir de cultures pures qui pourront être intégrées dans la bibliothèque de référence de barcodes Phytool.

IV.2. Préparation des échantillons

Un total de 39 souches de la TCC ont été sélectionnées et remises en culture individuellement dans le but de représenter au mieux les différents clades phytoplanctoniques et de pouvoir tester l'efficacité des différents couples d'amorces. Elles ont été complétées par 2 souches achetées à la CCAP (*Culture Collection of Algae & Protozoa*) pour couvrir au mieux toute la diversité taxonomique présente habituellement dans le phytoplancton. Au total, 41 souches ont ainsi été utilisées et mises en culture le 27.01.2021. Les durées de croissance n'ont pas été les mêmes pour chacune des souches. De ce fait, après quelques semaines de maintien en culture, une première série de prélèvements a été réalisée sur certaines souches qui présentaient une biomasse suffisante (changement de couleur significatif du milieu). Les souches en question ont été soumises à des comptages sur cellules de Malassez afin d'avoir une estimation de la densité cellulaire à un instant t. A partir du moment où la

densité cellulaire a été suffisamment importante, 2 échantillons ont été conservés à -20°C. La même opération a été reconduite sur les souches restantes en leur laissant davantage de temps de croissance.

Parmi les différentes souches cultivées, toutes n'ont pas été amenées à l'étape du séquençage, en effet seules celles répondant à certains critères (voir plus loin) ont été conservées pour réaliser une communauté de souches qui constituera les échantillons témoins. Les autres ont cependant été utiles lors des nombreux tests d'optimisation des réactions de PCR pour lesquelles elles constituaient un matériel essentiel (*cf*. §IV.3). Sur les 39 souches, 10 d'entre elles (voir tableau 3) constituent la communauté témoin qui sera représentée sous 2 types d'échantillon pour le séquençage. Ces dernières ont été sélectionnées sur la base des critères suivants :
(1) l'ensemble des souches devait bien couvrir la diversité taxonomique du phytoplancton ;
(2) les souches devaient être facilement cultivables (parmi les 39 souches, certaines se développaient lentement) ;
(3) l'estimation de la densité cellulaire par observations microscopiques (cellule de Malassez) devait être fiable (certaines souches formaient des agrégats difficiles à comptabiliser) ;
(4) les souches devaient présenter un barcode déjà référencé pour le 16S et 23S ; si ce n'était pas le cas elles devaient présenter une espèce apparentée au même genre présentant un barcode déjà référencé. Ce dernier point permet de faciliter l'assignation taxonomique des séquences obtenues à l'issue du séquençage.

Tableau 3 : présentation des souches de la TCC utilisées pour constituer les échantillons témoins. La concentration d'ADN estimée au Qubit ainsi que la concentration cellulaire estimée grâce aux observations sur cellules de Malassez sont données. Les volumes utilisés pour réaliser les mélanges sont également donnés à titre informatif. Dans le tableau, la mention **MA** indique le mélange d'ADN des souches et **MS** le mélange des souches, comme explicité dans le paragraphe suivant.

Souche et sa référence TCC	**[ADN]$_{Qubit}$** (ng.µL^{-1})	**Quantité mélangée pour MA** (µL)	**[cellules] estimée** (#cell.µL^{-1})	**Quantité mélangée pour MS** (µL)
Asterionella formosa TCC362	4.17	1.2	478	209
Botryococcus braunii TCC57	1.82	2.74	103.5	966
Chlamydomonas reinhardtii TCC234-2	19.7	1*	4366.7	23
Cosmarium regnellii TCC56	25.4	1*	4550	22
Cyclotella meneghiniana TCC640	19.3	1*	152.5	656
Dolichospermum flosaquae TCC79	22.2	1*	7140	14
Mougeotia sp. TCC814	9.13	1*	27	3704
Stichococcus bacillaris TCC145-1	7.6	1*	5062.5	20
Tetradesmus obliquus TCC142-1	3.98	1.26	10275	9.7
Xanthonema montanum TCC165	20.2	1*	852	117

* 1µL d'une dilution préalablement réalisée avec une concentration en ADN ~ 5ng.µL^{-1}

Les différents types d'échantillons utilisés pour le run de séquençage peuvent être distingués en quatre catégories principales, explicitées dans les paragraphes qui suivent et sur la figure 1.

1. Mélange d'ADN des souches témoins (nommée **MA** ci-après)

Les premiers extraits récupérés des souches ont ensuite été exploités comme suit : des extractions individuelles d'ADN de chaque souche ont été réalisées avec le protocole GenElute

(Mat.Supp. 3) à partir des culots cellulaires de chacune des 10 souches (Tableau 3). Les ADN extraits des souches ont été dosés au NanoDrop et au Qubit (les résultats du Qubit, plus fiables, ont été retenus). Ces ADN ont ensuite été mélangés ensemble à des quantités variables en fonction de la concentration d'ADN estimée, afin que chaque souche soit représentée à une concentration d'ADN similaire (~ 5ng.µL-1). Cependant, il faut garder à l'esprit que l'ADN quantifié constitue l'ADN total présent dans l'échantillon. Il ne correspond pas uniquement à celui des communautés phytoplanctoniques mais inclut également celui des autres microorganismes présents dans les cultures, dans des proportions variables et difficiles à estimer : les cultures ne sont pas axéniques et peuvent présenter des bactéries par exemple. Le mélange ainsi constitué sera appelé « MA » (pour Mélange ADN) par la suite. Pour chacun des couples d'amorces, il sera amplifié en triplicats et envoyé au séquençage.

2. Mélange de cellules des Souches témoins (nommé **MS** ci-après)

Les seconds extraits récupérés des souches ont été mélangés ensemble à des quantités variables en fonction des densités cellulaires précédemment estimées. Le but étant, de réaliser un mélange homogène des 10 souches témoins avec un apport estimé de 100 000 cellules de chacune d'elles dans le mélange. Ce mélange a été complété jusqu'à 200mL avec de l'eau stérile (et filtrée à 0.22µm), la première moitié de celui-ci (*i.e.* 100mL) a été filtrée sur un filtre 0.45µm et l'autre sur une capsule filtrante (Sterivex™) 0.45µm. La même extraction d'ADN a été réalisée à partir de ces 2 supports suivant le protocole NucleoSpin Soil (Mat. Supp. 4) et la version de celui-ci adapté pour les Sterivex™ (Domaizon et *al.*, 2019). Les extraits d'ADN récupérés par ces deux conditions constituent les échantillons appelés « MS » (pour Mélange Souches) dans la suite de ce rapport avec « MSF » pour « Mélange Souche Filtre » et « MSS » pour « Mélange Souche Sterivex ».

3. Echantillon environnemental (nommé **ENV** ci-après)

Un échantillon environnemental de 1 litre a été prélevé en surface du lac Léman, en milieu côtier, proche du port de l'UMR Carrtel, le 13.04.2021. Il a été divisé en 4 parts égales (*i.e.* 250mL). La première a été conservée au Lugol pour réaliser une observation microscopique. Cette dernière a été réalisée par Frédéric Rimet, selon les protocoles standardisés (Utermöhl, 1958 ; CEN, 2006) basées sur le comptage de 400 individus. Les 3 autres parts ont été filtrées par des techniques différentes (2 filtres *vs* 1 Stérivex) combinées avec des méthodes d'extraction d'ADN également différentes (1 GenElute *vs* 2 NucleoSpin Soil).

4. Séquençage de souches pour la complétion de la bibliothèque de barcode

L'ensemble des échantillons (MS, MA, ENV) présentés ci-dessus totalisait 40 puits au final sur la plaque à envoyer à la plateforme de séquençage. Or, la plateforme demandait un nombre multiple de 24 pour un run de type NanoMiSeq (*e.g.* 24, 48, 72 …) moins 1 puit qui doit être laissé libre pour leur contrôle interne (donc 23, 47, 71 … échantillons disponibles). Il a été choisi d'envoyer 47 échantillons. Il a donc été possible de rajouter 7 échantillons supplémentaires pour compléter la plaque. Ces derniers ont été constitués de mélanges de cultures de la TCC dont les barcodes n'étaient pas renseignés dans les bases de références. Plusieurs souches ont été mélangées au sein de chacun des échantillons avec pour critère qu'elles soient suffisamment éloignées phylogénétiquement entre-elles de manière à ce qu'il soit évident de leur réattribuer *a posteriori* leur barcode. Grâce à cette technique, il a été possible d'obtenir leur barcode ADN, et de compléter les bibliothèques de barcodes de référence.

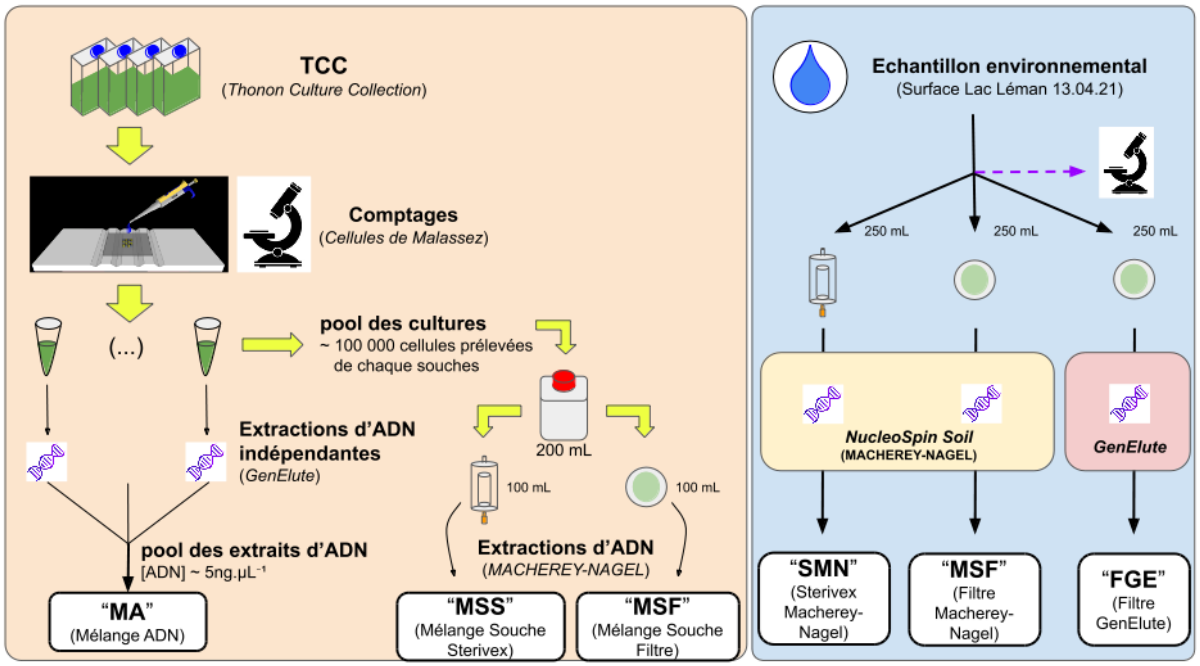

Figure 1 : les différents types d'échantillons destinés au séquençage. Ces derniers seront chacun amplifiés par les 5 différents couples d'amorces à tester. D'un côté les échantillons synthétiques (en orange) composés de souches issues de la TCC avec différents traitements, de l'autre un échantillon environnemental (en bleu) présentant également plusieurs modalités de traitements.

IV.3. Optimisation des réactions de PCR et préparation des échantillons pour le séquençage

Les recommandations fournies par la plateforme PGTB (Mat. Supp. 2) ont été utilisées pour le choix des réactifs et des programmes de PCR. Les couples d'amorces ont été commandés sur GATC (https://eurofinsgenomics.eu) accompagnés du couple d'oligonucléotides (queues moléculaires fournies par PGTB) fusionnées en 5' de chacune des amorces forward et reverse. Le Mix 2X KAPA HiFi HotStart (Roche) a été utilisé comme principal mélange réactionnel pour la PCR. Ce dernier contient tous les constituants nécessaires au déroulement de la réaction de PCR (dont une enzyme haute fidélité) excepté bien sûr les amorces et l'ADN à amplifier. Le thermocycleur GeneExplorer (Bioer Technology) a été utilisé tout le long de l'expérience. Le programme PCR utilisé correspond à celui préconisé par la plateforme PGTB (*cf.* Tableau 4).

Tableau 4 : présentation du programme PCR préconisé par PGTB sur lequel s'appuient les différents tests d'optimisation.

Température (°C)	Durée	# cycles	commentaire
95	3'	-	activation/dénaturation initiale
95	30''	30 à 35	dénaturation
55-65	30''		hybridation
72	30''		élongation
72	5'	-	élongation finale
10	∞	-	conservation

Afin de déterminer une plage de température optimale pour l'amplification de l'ADN des cibles de l'ensemble des couples d'amorces (queues moléculaires incluses), une première PCR à gradient de température a été réalisée. Pour cette première PCR à gradient de température, ce dernier s'étendait de 54 à 62°C pour la température d'hybridation des amorces et comportait 30 cycles d'amplifications. Les échantillons à amplifier utilisés étaient des mélanges d'extraits d'ADN de souches TCC, un témoin positif bactérien (pour tester la spécificité) et un témoin négatif. Le mélange réactionnel (de 25µL) utilisé était le suivant :

- 2X KAPA HiFi HotStart RM : 12.5µL ;
- amorce forward (1µM) : 5µL ;
- amorce reverse (1µM) : 5µL ;
- ADN cible (~ 5ng.µL^{-1}) : 2.5µL.

Cette première PCR a donné des premiers résultats encourageants : les amplifications ont toutes fonctionné avec les amorces couplées aux queues moléculaires PGTB. Une température de 58°C a été retenue pour la suite, température médiane de la plage fournie mais aussi à laquelle les amplifications ont bien fonctionné (*i.e.* bandes intenses observées sur le gel d'électrophorèse) pour tous les couples d'amorces. (photos des gels d'électrophorèse en Mat. Supp. 5.1). En revanche, les résultats ont aussi mis en évidence que tous les couples d'amorces (excepté le couple CYA359F/CYA781R pour le 16S) ont amplifié le témoin bactérien, et n'étaient donc pas spécifiques au phytoplancton. L'amplification du témoin était cependant moins marquée pour les amorces ciblant le 23S que le 16S ('16S_PHY').

Il est possible d'utiliser des programmes de PCR alternatifs, pour améliorer la spécificité de l'amplification pour les taxa cibles (ici les microalgues). Pour cela, une approche par TD-PCR (Touch Down-PCR) a alors été testée afin de cibler plus spécifiquement les organismes phytoplanctoniques et s'affranchir des bactéries. Ce programme PCR a consisté à réaliser les 10 premiers cycles de PCR (*i.e.* dénaturation, hybridation et élongation) à une température d'hybridation initiale plus élevée (ici 68°C), décroissant de 1°C à chaque cycle (pour arriver ici à une température d'hybridation finale de 58°C). Les 20 cycles supplémentaires sont ensuite réalisés à cette température finale. Une autre série de tests de TD-PCR a été réalisée avec des températures d'hybridation s'étendant de 65 à 55°C. Les résultats de ces TD-PCR montrent qu'il est ainsi possible de s'affranchir de l'amplification des bactéries hétérotrophes. Cela a été observé pour les couples d'amorces visant le 23S, mais a été sans succès pour le couple '16S_PHY' pour le 16S. Ces résultats ont été confirmés *a posteriori* après l'analyse des résultats de séquençage pour certains échantillons (*cf.* §VI.2.). La TD-PCR dont les températures d'hybridation initiales s'étendaient de 65 à 55°C semblent montrer une meilleure efficacité d'amplification pour davantage de taxa (*cf.* gels d'électrophorèse en Mat. Supp. 5.2). Cependant, il a été observé que certaines souches étaient peu ou pas amplifiées par cette approche, alors qu'elles l'étaient via une approche PCR classique. Cela peut s'expliquer par exemple lorsque les couples d'amorces présentent quelques nucléotides de différences avec les régions cibles de l'ADN à amplifier, l'approche TD-PCR peut alors s'avérer un peu trop spécifique pour certaines de ces souches. Afin de s'assurer que l'amplification des différentes souches ne soit pas altérée par ce phénomène, il a donc été décidé de conserver un programme PCR classique avec 30 cycles d'amplification à 58°C, puisque ce programme a permis l'amplification de la majorité des souches. Cela permet de ne pas prendre le risque de réduire la diversité amplifiée.

Enfin un dernier problème a reposé sur la présence de smears (des trainées laissées par une bande lors de l'électrophorèse), dont il était préférable de s'affranchir pour les envois des échantillons au séquençage. Ces derniers peuvent être causés par une amplification trop importante pouvant être liée à un nombre de cycles trop élevé et/ou une concentration en ADN initiale trop importante. Le nombre de cycles a été fixé au nombre minimal recommandé par PGTB (30), la quantité d'ADN a donc été diminuée : pour une concentration toujours égale à 5 ng.µL-1 une quantité de 0.66µL a été déposée à la place de 2.5µL pour les échantillons synthétiques. En ce qui concerne les amplifications de l'échantillon environnemental, où les smears étaient moins présents, 2µL ont été déposés (au lieu de 2.5µL). L'ajustement de ces paramètres sera un point à discuter et améliorer lors des futures analyses, qui se révélera certainement être un facteur dépendant de la nature des échantillons. Concernant les souches dont l'ADN a été amplifié individuellement en vue de séquencer leur barcodes, l'approche par TD-PCR a été favorisée ; lorsqu'elle ne donnait pas lieu à des amplifications, la PCR a alors été sélectionnée.

Une fois la plaque (SuperPlate PCR Plate AB2400 – THERMOFISCHER AB2400) complétée avec les 47 échantillons, elle a été scellée avec un film adhésif (Adhesive PCR Seal 4TiTUDE 4Ti-0500), placée dans un colis avec de la glace et envoyée à la plateforme pour séquençage le 19.04.2021. Le plan complet de la plaque et les gels d'électrophorèse associés à chaque échantillon sont disponibles en Mat. Supp. 5.3.

V. Outils bio-informatiques et adaptation des bases de références

V.1. Création d'outils bio-informatiques opensource

Certains scripts écrits l'année précédente ont été optimisés et réécrits dans le but de créer des applications accessibles à toutes les personnes travaillant sur les mêmes thématiques. Nous espérons que la mise à disposition de ces outils favorisera les collaborations permettant d'améliorer la robustesse et la fiabilité des comparaisons de résultats obtenus par des approches en microscopie ou en moléculaire. L'interface ShinyApp, proposée par Rstudio (package *shiny*), a été choisie pour réaliser ces applications. En effet, en plus de disposer d'un accès à un serveur R gratuit pour héberger les applications sur le web, le codage reste un langage R (utilisé en majeure partie pour les scripts rédigés l'an dernier) dans lequel il est possible d'intégrer d'autres langages (*e.g.* bash, Python ...).

Le développement de deux applications était souhaité : une première permettant l'homogénéisation de listes taxonomiques de communautés phytoplanctoniques (d'eau douce) et une seconde visant à dessiner des couples d'amorces adaptés au métabarcoding. Seule la première a concrètement aboutit en 2021. La seconde s'est avérée trop gourmande en ressource sur l'interface web ; une stratégie pour alléger les temps de calcul a déjà été testée, cependant par manque de temps, il n'a pas été possible de finaliser son développement, elle ne sera donc pas intégrée dans la suite de ce rapport. La première application évoquée est, quant à elle, centrale pour le déroulement du projet mais aussi pour la communauté scientifique travaillant sur le phytoplancton d'eau douce. Cette dernière est d'ores et déjà disponible en ligne (https://caninuzzo.shinyapps.io/phytool_v1/) en version beta (Canino et *al.*, 2021). Comme évoqué précédemment, l'objectif majeur de cette application est de pouvoir mettre à jour automatiquement l'ensemble des rangs taxonomiques du phytoplancton d'eau douce référencé en lien avec le logiciel Phytobs (Laplace-Treyture et *al.*, 2017) dont la taxonomie est à jour et basée sur une référence pour le phytoplancton : AlgaeBase (Guiry & Guiry, 2020). Phytool propose également des outils pour reformater des données ADN (fichiers FASTA), naviguer au sein de la liste des taxa disponibles afin de vérifier l'évolution de leurs noms et la présence de barcodes associés. Des bibliothèques de barcodes pour différents marqueurs (à ce jour : ARNr 16S ; ARNr 18S et ARNr 23S) y sont également présentes et sont détaillées dans le paragraphe suivant.

V.2. Curation et adaptation des bases de référence

Une nouvelle stratégie de curation de bases de données publiques de référence (*cf.* tableau 5) a été utilisée afin de collecter un maximum de barcodes disponibles pour le phytoplancton d'eau douce. Ce travail a été réalisé sur différentes bases de références, dans le but de produire des bibliothèques de barcodes ADN d'intérêt pour le projet (ARNr 16S ; ARNr 23S) mais aussi pour l'ARNr 18S qui est un marqueur habituellement utilisé pour décrypter la diversité taxonomique eucaryotique, notamment du phytoplancton.

Tableau 5 : aperçu des marqueurs génétiques utilisés, et mis en ligne dans les bibliothèques de barcodes Phytool et leurs origines.

Marqueurs génétiques	**Bibliothèques de séquences utilisées**	**Références**
Petite sous-unité ribosomale procaryotes **ARNr 16S**	Silva_138.1	Quast et *al.*, 2013
	PR2	Guillou et *al.*, 2013
	PhytoRef	del Campo et *al.*, 2018

Petite sous-unité ribosomale eucaryotes **ARNr 18S**	Silva_138.1	Quast et *al.*, 2013
Grande sous-unité ribosomale procaryotes **ARNr 23S**	Silva_138.1	Quast et *al.*, 2013
	µgreen-db	Djemiel et *al.*, 2020

La stratégie utilisée et schématisée sur la figure 2, est la suivante : après sélection d'une base de référence en ligne, celle-ci est reformatée pour pouvoir être homogénéisée taxonomiquement sur l'application Phytool en ne conservant que les séquences des taxa appartenant au phytoplancton d'eau douce. Une fois cette étape réalisée, un pipeline sur R cherche les séquences redondantes pour une même espèce, pour n'en conserver qu'une seule, et détecte également les espèces redondantes pour ne garder qu'une seule séquence les représentant. La séquence retenue est la plus longue et avec le minimum d'ambiguïtés possible. Si une même séquence est détectée pour des taxa différents, alors elle est exclue du jeu de données et mise de côté dans un fichier, sa curation sera faite manuellement et sera réalisée dans un second temps. Une fois ce procédé réalisé pour chaque base de référence individuellement, les séquences curées obtenues sont rassemblées par marqueurs. Le même pipeline est reconduit à nouveau sur ce jeu de données et les séquences sans conflits obtenues à l'issue de celui-ci constituent les bibliothèques de références fournies dans l'application Phytool dans sa version actuelle.

Il est prévu d'implémenter les bases de références régulièrement au cours du projet, notamment pour les marqueurs ARNr16S et ARNr23S. L'enrichissement des bases de références se fera grâce à l'implémentation de nouveaux barcodes obtenus par des séquençages réalisés durant ce projet, mais également en consultant les bases publiques (*e.g.* nouvelles données de séquençage Sanger disponibles sur GenBank) et les nouvelles publications.

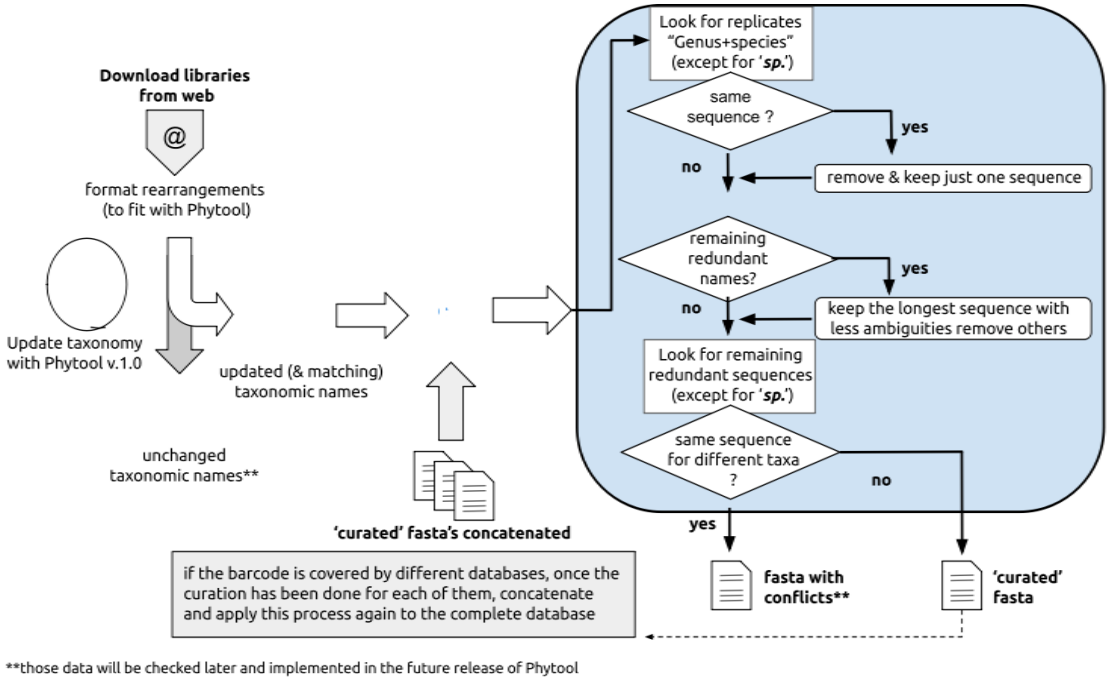

Figure 2 : pipeline schématisé de la curation des bases de références en ligne en vue de leur implémentation sur Phytool. Figure de Canino et *al.*, 2021.

VI. Analyses bio-informatiques des résultats de séquençage

VI.1. Pipeline bio-informatique utilisé

Le traitement des données de séquençage a été réalisé en intégralité sur Rstudio v.1.4.1106 à partir d'un pipeline combinant le programme cutadapt v3.5 (Martin, 2011), les commandes du package DADA2 v1.20.0 (Callahan et *al.*, 2016) et également le programme Mothur (Schloss et *al.*, 2009). Les bases de références utilisées sont celles constituées pour ce projet et accessibles en ligne sur la première version Beta (appelée aussi v.1.0) de l'application Phytool. La figure 3 présente les principales étapes du pipeline bio-informatique utilisé depuis l'importation des fichiers brutes obtenus de la plateforme de séquençage (.fastq) jusqu'aux assignations taxonomiques des ASV (*i.e.* Amplicon Sequence Variant). En complément, les valeurs attribuées aux paramètres nécessaires à certaines commandes sont décrites dans les lignes suivantes.

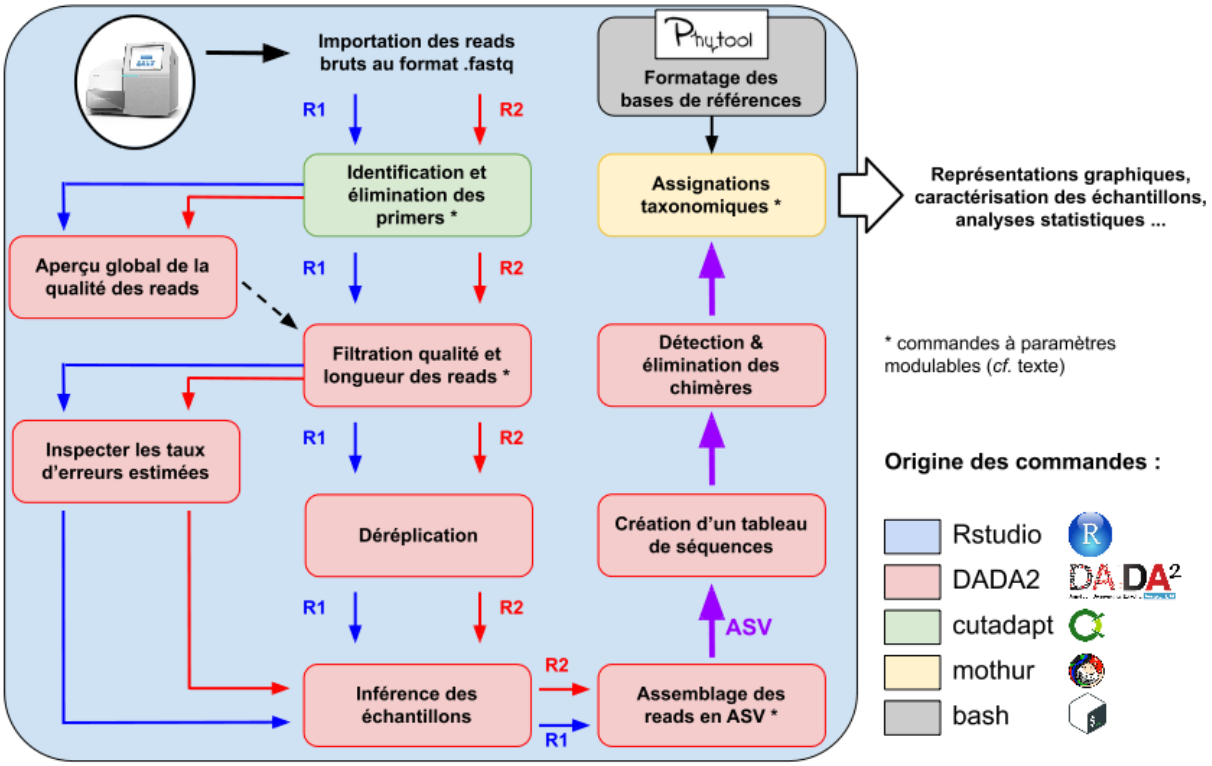

Figure 3 : schéma général du pipeline bio-informatique utilisé pour le traitement des données de séquençage.

Voici les étapes pour lesquelles différentes valeurs de paramètres ont été testés :

Identification et élimination des amorces :

La commande cutadapt a été utilisée, avec un maximum d'ambiguïté de 0 et l'instruction de rejeter les reads pour lesquels les amorces requises n'ont pas été détectés.

Filtre en fonction de la qualité et de la longueur des reads :

Au vu de la qualité des reads, qui semblait se dégrader au-delà de 230pb (sur les 250pb initiales), il a été choisi de tronquer les reads à 220pb. Etant donné qu'une fois les amorces retirées, les reads mesuraient environ 230 pb, seulement 10 nucléotides finaux ont été en réalité retirés par cette étape. Ce choix a été fait afin de conserver davantage de paires de bases chevauchantes pour l'étape d'assemblage des reads ; plus d'explications à ce sujet sont disponibles à la section suivante. En plus de ce paramètre, le maximum d'ambiguïté accepté était fixé à 0, le maximum d'erreurs attendues à 2 et le Q score minimal à 2.

Assemblage des reads en ASV :

Cette étape est très complémentaire et dépendante de la précédente (*i.e.* la filtration des reads). L'assemblage va dépendre de 2 paramètres principaux : la longueur minimale du chevauchement des reads R1 et R2 (en paires de bases, voir la figure 4 qui décrit le cas du barcode avec la plus grande taille et donc pour lequel le chevauchement est le plus limitant) et le nombre maximal d'erreurs autorisées sur les bases chevauchantes. La longueur minimale chevauchante a été définie à 50 paires de bases, ce qui est bien plus contraignant que celle définie par défaut sur DADA2 (12 paires de bases). Les reads ont été donc assemblés en ASV s'ils présentaient un minimum de 50 nucléotides chevauchants identiques car aucune erreur n'était autorisée. Ce choix a été fait dans le but de réduire le nombre de potentiel ASV chimères. Dans le cas où la zone chevauchante était conservée, 12 nucléotides similaires restent potentiellement un seuil où des reads appartenant à des espèces proches peuvent être assemblés entre eux. Des investigations ont permis de confirmer cette hypothèse, même si, *a posteriori* dans le pipeline, l'étape d'élimination des chimères s'est avérée efficace pour retirer ces dernières.

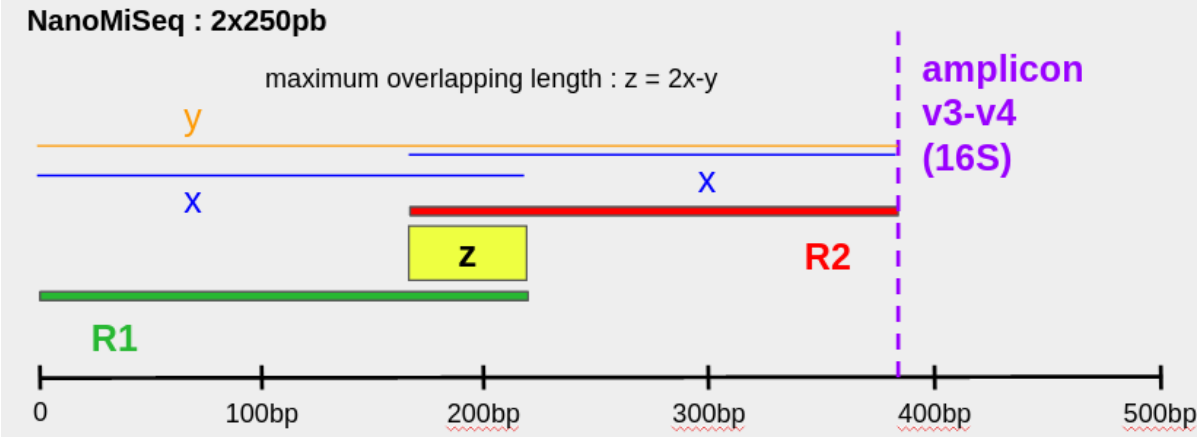

Figure 4 : Mise en évidence de l'étape d'assemblage des reads et présentation de la longueur maximale de chevauchement possible pour le barcode présentant la plus grande taille.

Assignations taxonomiques :

Enfin l'assignation taxonomique a d'abord été réalisée sur DADA2. L'écriture d'une fonction annexe a permis de mettre en évidence le fait que les résultats des assignations taxonomiques sur DADA2 peuvent se montrer différents pour des conditions initiales identiques. Une part d'aléatoire subsiste et il n'est pas possible de contrôler le nombre d'itérations lors de la classification taxonomique des séquences sur DADA2. Pour pallier ce problème et améliorer la robustesse des résultats d'assignation, la commande équivalente sur Mothur a été utilisée (faisant appel au même algorithme de classification : Wang et *al.*, 2007), celle-ci ayant l'avantage de laisser le choix du nombre d'itérations. Les valeurs de bootstrap supérieures à 80 (défaut) ont été retenues et 10'000 itérations ont été réalisées.

Dans le but de justifier ce choix, deux tests ont été réalisés en réassignant la base de référence Phytool v1 sur elle-même avec la fonction assignTaxonomy de DADA2 et avec classify.seqs de Mothur. Le 1er test a été réalisé en utilisant la totalité des barcodes présents dans Phytool v.1, le 2ème test a été réalisé en utilisant uniquement les barcodes résolutifs (c'est-à-dire les barcodes strictement différents entre taxa[1]). Les résultats sont présentés dans le tableau 6 et montrent que la commande de Mothur offre une meilleure capacité d'assignation pour l'ensemble des barcodes. En revanche, il n'y a pas de différences significatives si le processus est appliqué sur les barcodes résolutifs uniquement, outre le fait que l'approche Mothur permette d'avoir une mesure plus robuste de l'appartenance de la séquence aux rangs taxonomiques supérieurs lorsque l'espèce n'est pas détectée.

Les bases de références utilisées pour les assignations taxonomiques sont issues d'un découpage ciblant uniquement le barcode d'intérêt pour le couple d'amorces concerné. Un script écrit en bash, et intégré au pipeline sur R, a permis cela. Il se base sur une commande Mothur et permet de fournir les

[1] En effet, les séquences entières qui sont différentes lorsqu'elles ne sont pas tronquées aux longueurs relatives des barcodes, peuvent devenir identiques alors qu'elles appartiennent à des taxa différents. Les barcodes ne sont alors plus résolutifs (*i.e.* discriminants). Pour le 2ème test, ces barcodes non-résolutifs ont alors été exclus.

fichiers requis par la commande d'assignation taxonomique de Mothur, cette fonctionnalité s'avérant très utile et efficace, sera implémentée prochainement dans Phytool pour être en libre accès et permettre de générer des bases de références 'sur mesure' adaptées aux pipelines de métabarcoding.

Tableau 6 : efficacité de la réassignation taxonomique des barcodes de la base Phytool sur eux-mêmes. Les réassignations sont exprimées en nombre de barcodes correctement réassignés (jusqu'à l'espèce) / le nombre total de barcodes, avec les proportions correspondantes en pourcentage.

	DADA2 (assignTaxonomy, BS=80)		Mothur (classify.seqs, cutoff=80, iter=10000)	
	Barcodes complets	Barcodes résolutifs	Barcodes complets	Barcodes résolutifs
16S_v3v4 'CYA'	1087/3262 33.32 %	1317/2033 64.78 %	2069/3262 63.43 %	1351/2033 66.45 %
UPA '108F/R'	513/835 61.44 %	605/681 88.84 %	688/835 82.4 %	601/681 88.25 %
UPA '587F/R'	528/851 62.04 %	612/697 87.8 %	700/851 82.26 %	614/697 88.09 %
UPA 'Sherwood'	510/836 61.00 %	601/682 88.12 %	683/836 81.7 %	598/682 87.68 %

VI.2. Analyse des échantillons synthétiques

Echantillons témoins « MA » (Mélanges d'ADN de souches)

En premier lieu, les échantillons synthétiques « MA », réalisés en triplicats, ont été traités afin de tester l'efficacité *in vitro* des différents couples d'amorces. Sur l'ensemble des ASV récupérés au final pour chaque échantillon, seuls ceux, dont l'abondance relative (pour l'échantillon donné) était supérieure à 0.05 %, ont été conservés dans le but d'éliminer les ASV rares. Cette valeur a été choisie car elle permettait de réduire le bruit et d'accéder à l'ensemble des taxa phytoplanctoniques détectés. Dans le cadre des échantillons témoins, étant donné que les organismes cibles étaient connus *a priori*, différents paramètres ont été testés dans le but d'optimiser les assignations taxonomiques des ASV (*cf.* paragraphe précédent sur les assignations taxonomiques).

La figure 5 présente les résultats de l'analyse du séquençage pour ces échantillons, et pour l'ensemble des couples d'amorces utilisés. Les résultats des triplicats ont été regroupés ensemble dans cette analyse, et ne seront pas présentés individuellement dans ce livrable, mais dans l'ensemble, les réplicats présentaient peu de variabilité entre eux (*cf.* figure en Mat .Supp. 6). Cette figure permet d'observer clairement le problème de spécificité du couple d'amorce '16S_PHY' : une importante proportion de bactéries hétérotrophes a été amplifiée et cela biaise considérablement le signal des organismes phytoplanctoniques, car les souches témoins ne sont pas bien représentées. Pour l'ensemble des autres couples d'amorces, la majorité des souches a été détectée, même si la plupart du temps, l'assignation des ASV jusqu'à l'espèce correspondante n'est pas précise et doit être déduite avec la connaissance des souches témoins (ceci est dû à l'incomplétude de la bibliothèque de référence). Les abondances relatives sont variables et éloignées de ce qui est attendu en théorie, cela se justifie dans un premier temps par le fait que les dosages de l'ADN ont été réalisés sur l'ADN total de chacun des échantillons, incluant donc l'ADN des bactéries hétérotrophes présentes dans les échantillons. Cette présence d'ADN est d'ailleurs justifiée par son importante abondance relative pour les échantillons amplifiés avec les amorces '16S_PHY', non-spécifiques du phytoplancton. La quantité d'ADN relative au phytoplancton n'était donc pas connue et estimée précisément dans ces échantillons et cela fait ressortir ce besoin pour la constitution de futures témoins de contrôles plus robustes pour l'année 2022 (*cf.* §VII.1). Le second facteur expliquant ce phénomène est lié aux nombres de copies variables des gènes ribosomaux combinés à des biais d'extraction d'ADN et d'amplification. D'autres biais existent et sont bien repris dans l'article de McLaren et *al.* (2019) ; ils

peuvent intervenir lors des étapes de bio-informatique ou du séquençage, mais sont moins importants que ceux intervenant lors de l'extraction ou de l'amplification des acides nucléiques.

En dépit de cela, l'ensemble des couples d'amorces ciblant la région UPA présentent des résultats comparables, une bonne spécificité pour le phytoplancton et une meilleure capacité à détecter les souches témoins que le couples d'amorces amplifiant le barcode v3/v4 de l'ARNr 16S. Parmi les taxa constituants les échantillons témoins, *Chlamydomonas reinhardtii* n'a pas été détecté par les barcodes du 16S, et la détection de *Tetradesmus obliquus* s'est avérée très faible pour le barcode v3-v4 du 16S et l'UPA amplifié par le couple d'amorces « 108F/108R ».

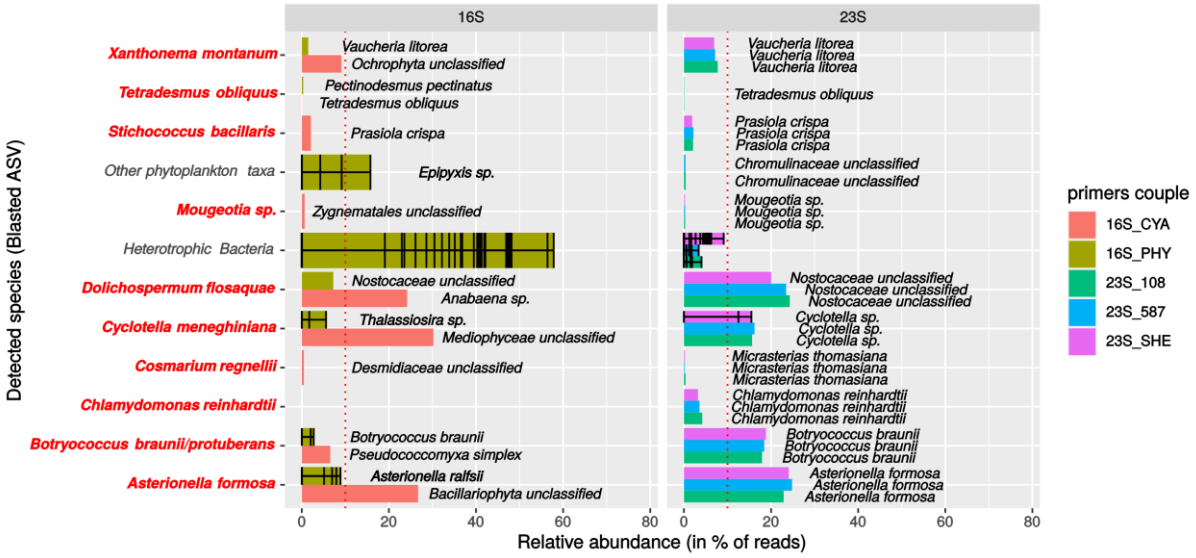

Figure 5 : histogrammes représentant les abondances relatives des ASV issus des différents couples d'amorces utilisés pour les échantillons témoins 'MA'. Lorsque plusieurs ASV composent un même taxon, des barres noires sont visibles sur l'histogramme, indiquant les proportions correspondantes à ces ASV. Les taxa témoins sont représentés en rouge, et l'abondance théorique est représentée en pointillées sur les graphes ; les taxa auxquels les ASV ont été originellement assignés sont écrit en noir. Les ASV mal assignés ou non rattachés à des organismes phytoplanctoniques ont été respectivement placés dans « Other phytoplankton taxa » ou « Heterotrophic Bacteria » .

Echantillons témoins « MS » (Mélange de cellules des Souches)

Les résultats de l'analyse des échantillons synthétiques « MS » correspondent au mélange de souches filtrées sur filtre ouvert ou Sterivex sont présentés sur la figure 6. Cette analyse confirme les principales tendances mises en évidence par les échantillons « MA ». Une autre tendance semble être mise en avant par ces résultats : certains taxa dont l'abondance relative est faible (*e.g. Stichococcus bacillaris*) semblent mieux détectés par l'approche filtre ouvert que l'approche Stérivex. Les taxa les plus abondants (*e.g. Cyclotella meneghiniana*) semblent plus représentés avec les Stérivex. Ces résultats sont cependant peu représentatifs, car non réalisés en réplicats.

Hormis cela, les Sterivex et filtres ouverts donnent des résultats comparables. Il est possible de voir que, pour le barcode v3-v4 du 16S, un des taxa n'est pas détecté : *Chlamydomonas reinhardtii,* il est détecté pour les barcodes du 23S. Inversement *Tetradesmus obliquus* ne l'est pas avec le 23S, mais est (faiblement) détecté avec le 16S. *Chlamydomonas reinhardtii* n'est pas non plus détecté avec les barcodes du 16S pour les échantillons MA, ce qui suggère que ce taxon n'a pas été amplifié, même si, *in silico,* l'amplification a fonctionné avec de la variabilité observée sur les amorces. Cela rappelle bien qu'en pratique et en théorie les résultats peuvent être très variables et que les seuils de non correspondance des nucléotides (mismatches) autorisés pour les expériences *in silico* ne fournissent qu'une estimation de la réalité. Le manque de détection ou avec une abondance relative très faible de certains taxa (*i.e. Tetradesmus obliquus ; Stichococcus bacillaris ; Cosmarium regnellii*) pour les différents barcodes et différentes conditions peut s'expliquer par une première hypothèse ici : la quantité plus faible d'ADN pour ces organismes, liée à leurs plus petites tailles. Cela explique d'ailleurs pourquoi de plus gros organismes, comme les diatomées, sont retrouvées avec

une plus grande abondance ; ce phénomène suit une tendance logique, déjà bien connue (Vasselon et *al.*, 2017).

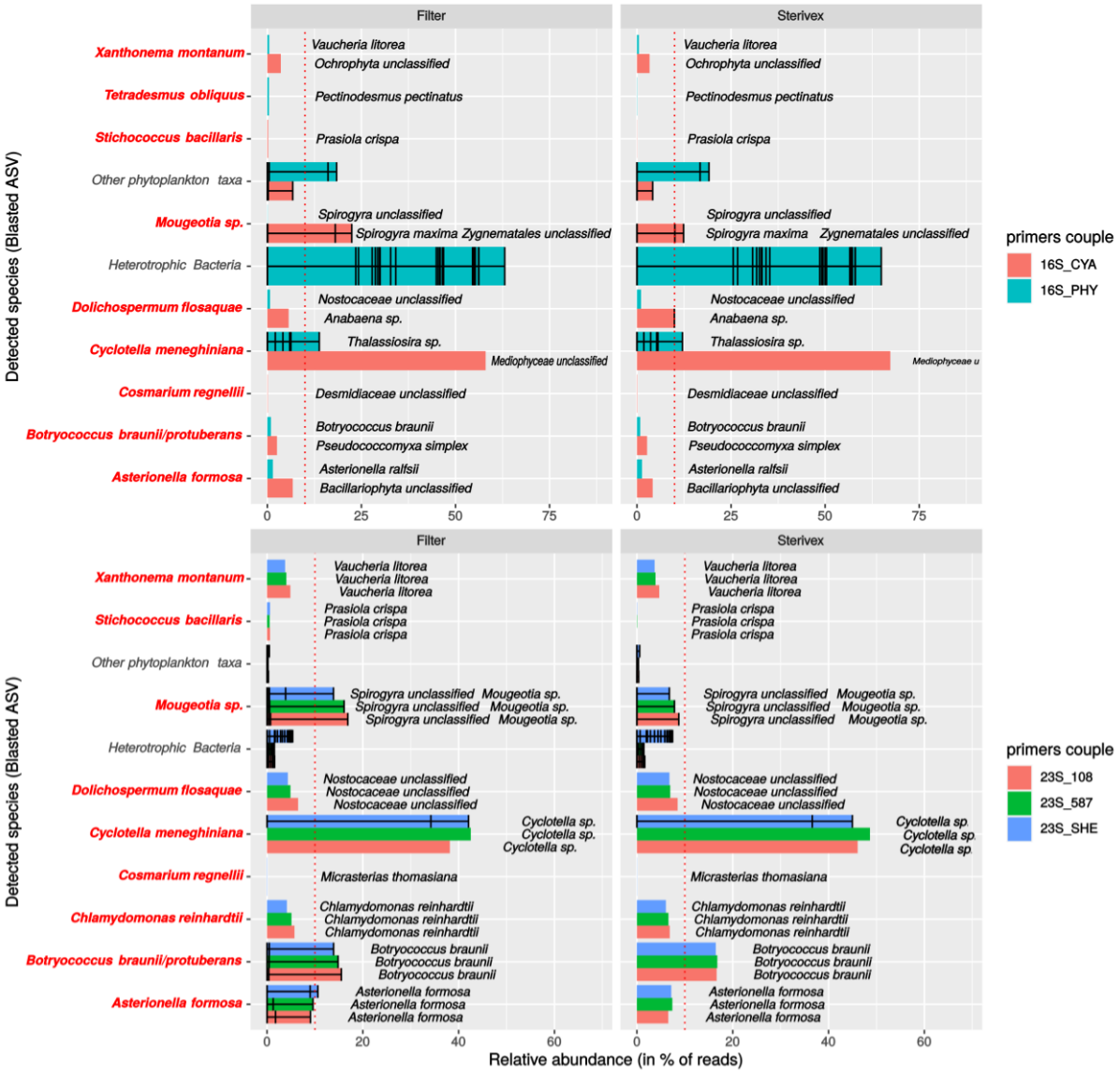

Figure 6 : histogrammes représentant les abondances relatives des ASV issus des différents couples d'amorces utilisés pour les échantillons 'MS'. Lorsque plusieurs ASV composent un même taxon, des barres noires sont visibles sur l'histogramme, indiquant les proportions correspondantes à ces ASV. Les taxa témoins sont représentés en rouge, et l'abondance théorique est représentée en pointillés sur les graphes ; les taxa auxquels les ASV ont été originellement assignés sont écrits en noir. Les ASV mal assignés ou non rattachés à des organismes phytoplanctoniques ont été respectivement placés dans « Other phytoplankton taxa » ou « Heterotrophic Bacteria ».

Nouveaux barcodes

Les 7 échantillons supplémentaires ont permis d'enrichir les bibliothèques de 18 et 23 barcodes supplémentaires pour la région v3-v4 de l'ARNr 16S et l'UPA de l'ARNr23S, respectivement. Les taxa associés à ces derniers (et leur barcode respectif) sont présentés en Mat. Supp.7 et seront ajoutés aux bibliothèques de barcodes Phytool. Certaines séquences (seulement 4) n'ont pas pu être assignées correctement, ou du moins avec certitude, leur barcode ne sont donc toujours pas disponibles. Cela peut s'expliquer par des amplifications peu efficaces ou des confusions dans les souches référencées. L'analyse de ces échantillons a cependant permis de confirmer l'efficacité de

l'approche par TD-PCR pour augmenter la spécificité des amorces aux organismes phytoplanctoniques : les séquences non-assignées, associées à des amplifications de bactéries hétérotrophes étaient en effet très réduites voire inexistantes pour les échantillons où les souches été amplifiées en TD-PCR (pour le barcode UPA).

VI.3. Comparaison microscopie/ADN

Les résultats de l'observation microscopique réalisée sur l'échantillon environnemental et ceux obtenus avec le séquençage de ce dernier (sous les différentes modalités de traitements appliquées) ont été comparés et sont présentés sur la figure 7. Dans le but d'avoir un aperçu global simplifié, une classification ascendante hiérarchique a été réalisée. Celle-ci s'appuie sur une matrice de distance (distance utilisée : Brays-Curtis) se basant sur les abondances relatives des différentes familles. Les histogrammes quant à eux ne représentent que les 10 familles les plus abondantes pour chacune des modalités de traitement de l'échantillon (ce choix a été fait pour améliorer la lisibilité de la figure et diminuer sa complexité).

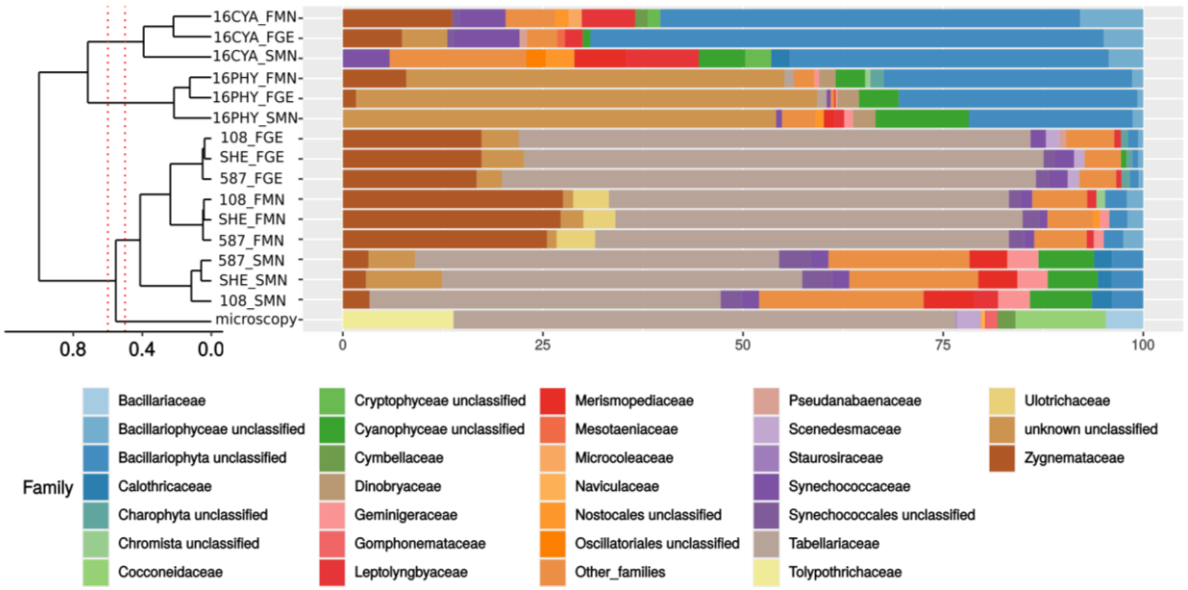

Figure 7 : classification ascendante hiérarchique basée sur les différentes familles détectées pour les différentes modalités de traitement de l'échantillon environnemental amplifié avec les différents couples d'amorces. Les lignes en pointillées rouges symbolisent les valeurs optimales locales d'un critère de clusterisation : le pseudo-F de Calinski-Harabasz Les histogrammes correspondants à chacun des échantillons représentant les 10 familles les plus abondantes. Les autres familles sont regroupées dans le champs 'Other_families' ; le champs 'unknown unclassified' représente les bactéries hétérotrophes. Les couples d'amorces CYA359F/CYA781R, PhytoF/PhytoR, 23S_108F/23S_108R, 23S_587F/23S_587R et 23S Sherwood sont représentés respectivement par «16CYA», «16PHY», «108», «587» et «SHE» sur la figure. La condition Filtre avec le protocole Macherey-Nagel est symbolysée par «FMN» ; Filtre avec GenElute par «FGE» et Sterivex avec Macherey Nagel «SMN».

Cette figure permet de mettre en avant que les échantillons amplifiés avec les amorces associées à la région UPA sont plus représentatifs des taxa observés en microscopie. Cela s'explique notamment par la meilleure résolution de ce barcode pour les principaux organismes observés : les diatomées et notamment la famille des Tabellariaceae, dont certaines des espèces observées en microscopie ont pû être également identifiées par l'approche moléculaire avec l'UPA. Encore une fois, il est possible d'observer la mauvaise spécificité des couples d'amorces 16S_PHY pour le phytoplancton : la famille nommée 'unknown unclassified' correspondant aux bactéries hétérotrophes. Le fait que les échantillons associés aux 2 barcodes de l'ARNr 16S soient similaires entre eux et distants de l'échantillon observé en microscopie s'explique en grande partie par l'abondance de la famille 'Bacillariophyta unclassified', signifiant que l'algorithme n'a pas permis d'assigner précisément les ASV associés aux diatomées correspondantes. Il est déjà possible d'émettre l'hypothèse que ces

barcodes ne soient pas suffisamment résolutifs pour les diatomées, et celle-ci se vérifie déjà et pourra l'être plus explicitement grâce à l'algorithme à venir (*cf.* §VII.2). De nombreuses différences sont tout de même notables sur l'échantillon en microscopie et les données moléculaires, même pour les barcodes UPA. Malgré de grandes tendances qui ressortent de manière quasi-similaire (*i.e.* diatomées), d'autres familles ne sont pas retrouvées avec les approches moléculaires et vice-versa avec la microscopie. La présence d'ADN libres peut en partie expliquer cela, mais également de petits organismes qui n'auraient pas été comptabilisés lors du comptage. Enfin certains organismes visibles ont probablement vu leur signal génétique réduit par l'ensemble de l'ADN présent dans l'échantillon, ils peuvent également ne pas avoir été assignés jusqu'à la famille (c'est le cas pour une espèce du genre *Tolypothrix*, cyanobactérie hétérocystée, dont l'assignation s'est arrêtée à l'ordre : *Nostocales*) ou encore ne pas avoir été amplifiés. L'incomplétude des bibliothèques de références joue un rôle dans ce phénomène.

Une autre tendance intéressante provient de la clusterisation de l'échantillon associé à l'UPA : il est possible d'observer que ces derniers sont regroupés non pas par couple d'amorces utilisés, comme il serait attendu *a priori,* mais plutôt par méthode de filtration et d'extraction. Ce résultat a été analysé de plus près pour comprendre mieux le phénomène et il s'avère que les échantillons filtrés sur Sterivex (extraction NucleoSpin Soil) présentent une diversité plus importante que ceux extraits sur filtres ouverts. Ces résultats peuvent s'expliquer par la plus grande surface de filtre disponible sur un Sterivex *versus* un filtre ouvert classique (Ø 47mm). Ces résultats contradictoires devront être confirmés à l'avenir avec d'avantage d'échantillons (témoins et environnementaux). Les courbes de la figure 8 présentent le nombre total de reads en fonction des abondances relatives, rangées par ordre décroissant, des ASV assignés au rang des familles. Ce genre d'approche permet de mettre en évidence la proportion de reads qui est représentée lorsque seuls les 10 ASV les plus nombreux sont retenus par exemple. Cela permet également de voir visuellement l'abondance et la diversité liée aux ASV plus rares dans les échantillons. Dans le cas présent, dans les modalités de traitement de l'échantillon avec filtres ouverts, les familles principales (car ici les ASV retenus sont ceux assignés à la famille) présentent des abondances relatives qui semblent plus élevées qu'avec Sterivex. La filtration sur Sterivex semble permettre de capter plus de diversité : au-delà de ces 10 familles les plus abondantes, il y a davantage de familles associées à des abondances relatives plus importantes qu'avec l'approche filtre ouvert.

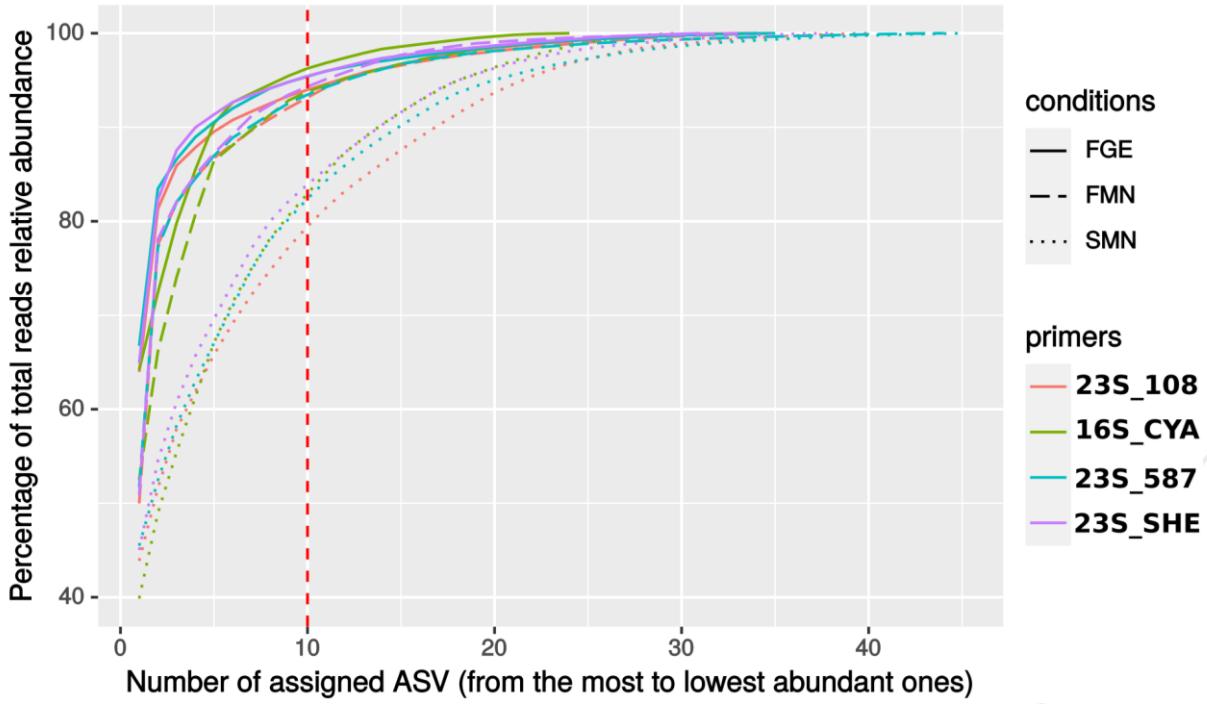

Figure 8 : courbe rang-fréquence basée sur le nombre d'ASV assignés au niveau de la famille. Le graphique présente le pourcentage de reads conservé en fonction du nombre de familles retenues. L'analyse a été réalisée pour l'ensemble des couples d'amorces, excepté '16S_PHY', et sur l'échantillon environnemental avec l'ensemble des conditions expérimentales appliquées. Une sélection des 10 familles les plus abondantes, condition exposée sur la figure 7, est représentée en

pointillées rouges. La condition Filtre avec le protocole Macherey-Nagel est symbolisée par «FMN» ; Filtre avec GenElute par «FGE» et Sterivex avec Macherey Nagel «SMN».

VI.4. Conclusion sur l'efficacité des différents couples d'amorces

Afin d'avoir une vue d'ensemble quant à l'efficacité des différents couples d'amorces testés, différents scores ont été attribués pour des paramètres qui semblaient importants à investiguer. Même si certains d'entre eux ne s'avèrent pas déterminants pour le choix d'un couple d'amorces, ils permettent néanmoins d'avoir une première idée de leur 'comportement' en laboratoire. Ces indices sont les suivants :

- **Spécificité :** la capacité des couples d'amorces à capter les organismes appartenant au phytoplancton au détriment des bactéries hétérotrophes. Etant donné que pour les extractions réalisées avec GenElute et NucleoSpin Soil, des différences ont été observées, les spécificités ont été séparées suivant ces catégories.

$$\frac{\text{number of reads } \in \text{ freshwater phytoplankton}}{\text{total number of reads}}$$

- **Reproductibilité :** sur la base du triplicat de PCR réalisé pour les échantillons « MA », il s'agit de tester la reproductibilité du résultat, même si celui-ci est soumis à une part d'aléa indétectable.

$$\frac{1}{\sum[\text{var}(\text{mock samples relative abundance})] \times 100}$$

- **Amplification totale :** il s'agit là de mesurer l'efficacité d'amplification des couples d'amorces en se basant sur le nombre total d'ASV correspondants à du phytoplancton d'eau douce (*i.e.* les souches témoins). Cette analyse s'est concentrée uniquement sur les échantillons témoins.

- **Résolution spécifique :** basée sur les résultats des investigations *in silico* réalisées en 2020, pour rappel ces dernières se basaient sur les résultats de PCR *in silico* et le nombre de barcodes résolutifs associés. Permet de rappeler pourquoi certains couples ont été choisis.

Ces paramètres ont été calculés uniquement en se basant sur les échantillons témoins, étant donné que leur composition était déjà connue et les résultats sont présentés sur la figure 9.

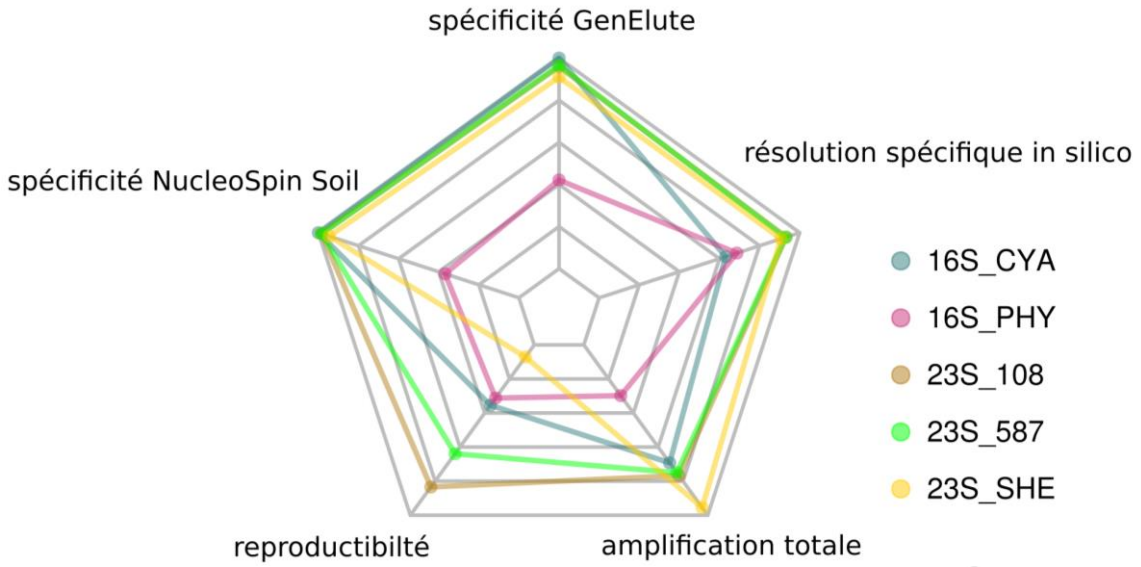

Figure 9 : estimation de l'efficacité des différents couples d'amorces à amplifier des barcodes avec une résolution optimale pour la détection des organismes phytoplanctoniques. Différents paramètres ont été investigués pour établir des scores qualifiant l'efficacité des couples d'amorces testés.

Comme cela a déjà été mis en évidence dans les paragraphes précédents, le couple d'amorces 'phytoF'/'phytoR' souffre d'un manque important de spécificité qui impacte lourdement la détection des organismes phytoplanctoniques. Le radarchart présente le couple d'amorces 16S_CYA comme le plus l'un des plus spécifique au phytoplancton (même si d'autres présentent également une spécificité comparable). En revanche, il rappelle aussi que la résolution du barcode associé est discutable, et cela a été montré lors de l'analyse de l'échantillon environnemental. Concernant les couples d'amorces relatifs à la région UPA, ils présentent des tendances très similaires, et cela s'est également vérifié lors des tests en laboratoire. Le couple d'amorces 23S_SHE semblent montrer un taux d'amplification nettement supérieur aux autres, en effet les nombres de reads étaient bien plus importants, et ce sont d'ailleurs les variations de ceux-ci qui expliquent son faible score en reproductibilité. Il ne faut donc pas attacher trop d'importance à ce critère de reproductibilité. L'efficacité d'amplification (« amplification totale » sur la figure 9) mérite plus d'attention, mais il faut rappeler qu'ici, elle se base seulement sur 10 souches phytoplanctoniques. Même si le nombre de reads est plus élevé, les échantillons amplifiés avec le couple 23S_SHE n'ont pas présenté plus de diversité que les autres pour le même barcode. De plus, les tests *in silico* ont montré que les couples 23S_108 et 23S_587 sont plus spécifiques au phytoplancton (meilleure spécificité confirmée en laboratoire), présentant moins de mismatchs (mauvaises correspondances) entre les amorces et les séquences testées et produisant également des barcodes plus discriminants. Le choix n'est donc pas si évident et la discussion reste encore ouverte.

VII. Discussion et perspectives

VII.1. Laboratoire

1. Préparation de échantillons témoins

La préparation d'échantillons témoins est un atout robuste, en particulier pour un projet comme celui-ci, qu'il faut absolument conserver et développer. Disposer d'une culture de microalgues est une opportunité d'avoir de nombreuses souches témoins. Cependant, certaines des souches utilisées n'ont pas été détectées par les analyses moléculaires et cela s'avère problématique pour des échantillons témoins. Comme déjà formulé précédemment, l'hypothèse la plus probable est que lors de l'extraction d'ADN des microalgues, une importante (et variable) proportion d'ADN appartenant à des bactéries hétérotrophes (présentes dans les cultures) ait été extraite. La quantification de l'ADN propre aux microalgues n'a donc pas été réalisée et c'est cette proportion qui est intéressante et qu'il faudrait pouvoir estimer afin de constituer des échantillons témoins plus efficaces pour la suite du projet. Diverses techniques seront discutées dans le but de réaliser cet objectif (*e.g.* tri cellulaire au cytomètre en flux lorsque possible, quantification par qPCR de fragments propres au phytoplancton, quantifications avec des sondes spécifiques au phytoplancton et rapport avec quantité d'ADN totale *etc.*). Une autre solution serait de commander des séquences synthétiques d'organismes types, mais cette solution s'avère très coûteuse.

2. Investigation des techniques d'amplification

En se focalisant sur les barcodes les plus prometteurs, l'approche utilisée ici (PCR classique) pour leur amplification a donné de bons résultats et sera sûrement conservée. Même si certains résultats ont montré que l'approche par TD-PCR garantissait une meilleure spécificité, il y a un risque non-négligeable de passer à côté d'une certaine part de diversité phytoplanctonique. L'approche par PCR classique s'est révélée suffisamment spécifique sur les échantillons synthétiques et environnementaux (très peu de bactéries hétérotrophes amplifiées pour les barcodes UPA de l'ARNr 23S, voir aucune pour le barcode v3-v4 de l'ARNr 16S). De futurs tests pourront être envisagés pour clarifier davantage ce point.

VII.2. Robustesse des barcodes utilisés

Dans le but d'avoir un aperçu de la résolution des différents barcodes utilisés, des analyses en coordonnées principales ont été réalisées en se basant sur la distance existante entre les barcodes. En d'autres termes, cette distance s'apparente à la dissimilarité des barcodes entre eux, pouvant être interprétée comme la distanciation phylogénétique entre les taxa représentés par ces barcodes. Ce travail d'investigation vise à être amélioré par la suite afin de donner lieu au développement de critères statistiques qui pourront améliorer l'assignation taxonomique de barcodes parfois difficile ou ambiguë. La distance utilisée pour représenter les dissimilarités entre barcodes se base sur le modèle d'évolution phylogénétique « K80 » (Kimura, 1980) dans un premier temps, mais d'autres modèles vont être testés par la suite. La figure 10 présente la similarité des barcodes pour la région v3-v4 de l'ARNr 16S -obtenus avec le couple d'amorces 16S_CYA (CYA359F/CYA781R)- et pour la région UPA de l'ARNr 23S (les résultats présentés ont été obtenus avec le couple d'amorces designé par Sherwood, mais les résultats obtenus avec les autres couples d'amorces sont similaires). Cette figure met en évidence une capacité discriminante (au niveau du phylum) plus importante pour la région UPA que pour les régions v3-v4. Cette tendance se vérifie statistiquement en comparant les rapports de l'inertie inter-phyla avec l'inertie intra-phylum basés sur les distances euclidiennes entre les points, où les phyla représentent les différentes catégories investiguées. Ce type d'analyse pourrait être appliqué sur des rangs taxonomiques plus fins par la suite, pour mettre en avant statistiquement les capacités discriminatoires de différents barcodes, mais un programme informatique est nécessaire afin d'optimiser ce processus. Il serait intéressant de développer cette approche pour les barcodes des taxa qui se sont révélés difficiles à réassigner sur eux-mêmes. Une représentation graphique a également été réalisée avec le package R *plotly*, qui permet d'avoir une sélection et représentation interactive des différents barcodes et les taxa auxquels ils sont associés. Cette approche pourrait être développée dans le futur et intégrée à l'application Phytool afin de permettre une meilleure représentation de la capacité discriminante des barcodes. De plus, elle

s'avère efficace pour détecter visuellement d'éventuelles erreurs de référencements ou d'identifications taxonomiques.

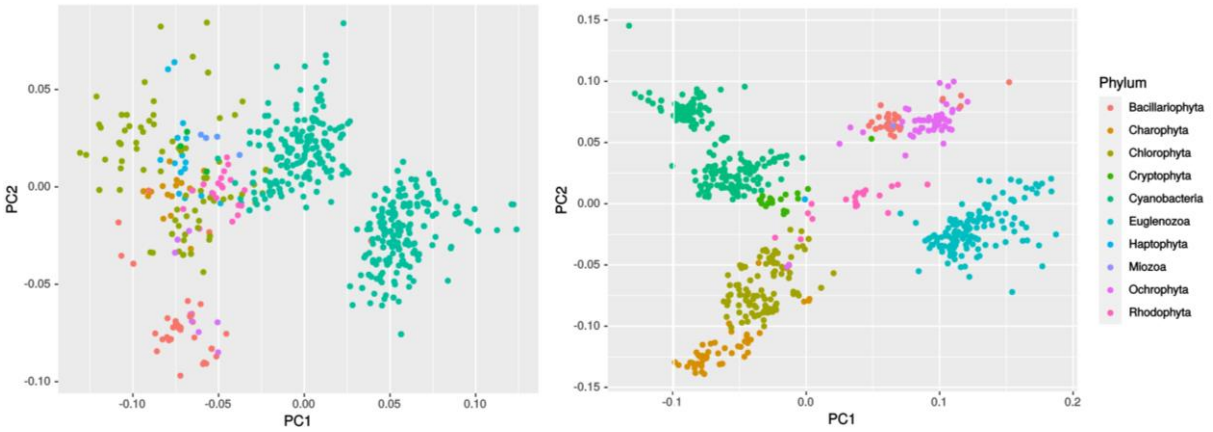

Figure 10 : représentations graphiques des analyses en coordonnées principales appliquées sur les barcodes de la région v3-v4 de l'ARNr 16S (à gauche) et de la région UPA de l'ARNr 23S (à droite). Les phyla associés aux barcodes sont représentés par des couleurs différentes.

Sur la base de classifications (supervisées et non-supervisées), il est possible d'envisager de créer des arbres de décisions construits sur l'attribution d'un rang taxonomique à un nouveau barcode. Des tests statistiques permettraient de montrer si les barcodes associées à différents taxa d'un même clade sont significativement différents, pouvant alors renseigner de manière plus précise sur la capacité du barcode à discriminer des groupes taxonomiques.

VII.3. Perspectives futures

Au vu des résultats obtenus cette année, les barcodes UPA de l'ARNr 23S semblent les plus prometteurs. Le « talon d'Achille » de ce barcode est la pauvreté de ces bibliothèques de références. Des efforts de complétion pourront être réalisés :

(1) en séquençant les souches disponibles à la TCC, qui seront remises en culture ;
(2) en séquençant les souches disponibles dans d'autres collections, (ex. MNHN) ;
(3) en réalisant de nouveaux isolements à l'UMR-CARRTEL à partir de prélèvements réalisés dans l'environnement. Il serait intéressant par exemple de réaliser des isolements à partir de prélèvements d'eau des DROM. Cela pourrait être envisagé dès l'an prochain si la demande d'envoi d'échantillons lugolés en provenance des DROM est acceptée.
(4) En parallèle à cela, des efforts de curation des bibliothèques en ligne (*i.e.* NCBI) seront réalisés.

L'application en ligne 'Phytool' développée cette année sera enrichie de nouvelles fonctionnalités, utiles pour le bon déroulement de ce projet mais qui seront également utiles à toute une communauté d'utilisateurs. Les avis de ces derniers seront d'ailleurs précieux pour améliorer l'application. Parmi les nouvelles fonctionnalités envisagées figurent :

-
- Le reformatage des bibliothèques de barcodes pour pouvoir être compatibles avec l'assignation taxonomique de Mothur.

- Des estimations de la résolution de barcodes, ciblées par des couples d'amorces entrés par l'utilisateur, seront aussi implémentées.

- Même s'il n'a pas d'intérêt particulier dans ce projet, l'information sur le barcode *rbc*L pour les taxa référencés dans Phytobs sera ajoutée, sur demande d'utilisateurs. Un lien vers la base de données diat.barcode (Rimet et *al.*, 2019) sera également fournie et certaines fonctionnalités pourront être appliquées à cette bibliothèque.

Enfin, de manière plus générale les objectifs établis pour l'année 2022 sont repris ci-dessous.

-

- Récupération des prélèvements réalisés en 2020 et 2021 (DOM, Lacs Alpins) ainsi que des comptages microscopiques réalisés sur ces prélèvements auprès des différentes structures. Le rappatriement de ces échantillons a été initié dès la fin de l'année 2021 et se poursuivra début 2022 ;

- Traitement de ces échantillons : extractions d'ADN, PCR et séquençages en MiSeq, auprès de la même plateforme (PGTB) en suivant les mêmes protocoles que pour le NanoMiSeq réalisé cette année. Deux runs sont prévus, un pour le 16S (couple d'amorces 16S_CYA) et un autre pour le 23S (barcode UPA encore à sélectionner avant la fin du premier trimestre 2022).

- Tests de différentes stratégies bio-informatiques, puis comparaisons avec les comptages en microscopie. Ceci est, en grande partie, basé sur le travail qui a déjà été fait cette année, mais cette fois, l'application concernera un nombre beaucoup plus important d'échantillons.

- Lancement d'une nouvelle campagne d'échantillonnage pour 2022 : plans d'eau des DROMs (suivis DCE, autres potentielles campagnes).

- Tenter d'initier un réseau de barcoders phytoplancton. Du côté de l'UMR Carrtel, un run NanoMiSeq sera d'ailleurs mis à disposition pour enrichir le barcode UPA (23S).

Bibliographie

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016) DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods 13: 581–583. https://doi.org/10.1038/nmeth.3869

Canino A, Bouchez A, Laplace-Treyture C, Domaizon I, Rimet F (2021) Phytool, a ShinyApp to homogenise taxonomy of freshwater microalgae from DNA barcodes and microscopic observations. Metabarcoding and Metagenomics 5: 199. https://doi.org/10.3897/mbmg.5.74096

CEN (2006) Water quality - Guidance standard on the enumeration of phytoplankton using inverted microscopy (Utermohl technique). European Committee for Standardisation EN 15204: 1–42.

Del Campo J, Kolisko M, Boscaro V, Santoferrara LF, Nenarokov S, Massana R, Guillou L, Simpson A, Berney C, de Vargas C (2018) EukRef: phylogenetic curation of ribosomal RNA to enhance understanding of eukaryotic diversity and distribution. PLoS biology 16: e2005849.

Djemiel C, Plassard D, Terrat S, Crouzet O, Sauze J, Mondy S, Nowak V, Wingate L, Ogée J, Maron P-A (2020) µgreen-db: a reference database for the 23S rRNA gene of eukaryotic plastids and cyanobacteria. Scientific reports 10: 1–11.

Domaizon I, Kurmayer R, Capelli C, Chardon C, Hufnagl P, Vautier M, Salmaso N (2019) Lake plankton sample collection from the field for downstream molecular analysis protocol metadata. protocols.io. Available from: https://www.protocols.io/view/lake-plankton-sample-collection-from-the-field-for-xn6fmhe/metadata (January 12, 2022).

European commission (2000) Directive 2000/60/EC of the European Parliament and of the Council of 23rd October 2000 establishing a framework for Community action in the field of water policy. Official Journal of the European Communities 327: 1–72.

Gorzerino C. (2021) Projet Algo-PAM-Barcode : Couplage PAM-Barcoding pour l'étude des communautés microalgales des milieux d'eau douce. 1er Meeting ADN-O : 15- 16 nov 2021 -Lyon Villeurbanne, livre des résumés, p 6.

Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, Boutte C, Burgaud G, De Vargas C, Decelle J (2012) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. Nucleic acids research 41: D597–D604.

Guiry MD, Guiry GM (2012) World-wide electronic publication, National University of Ireland, Galway. AlgaeBase.[En línea]. Disponible en: http://www. algaebase. org. Fecha de consulta 10.

Ivanova NV, Watson LC, Comte J, Bessonov K, Abrahamyan A, Davis TW, Bullerjahn GS, Watson SB (2019) Rapid assessment of phytoplankton assemblages using Next Generation Sequencing – Barcode of Life database: a widely applicable toolkit to monitor biodiversity and harmful algal blooms (HABs). 2019.12.11.873034pp. https://doi.org/10.1101/2019.12.11.873034

Laplace-Treyture C, Hadoux E, Plaire M, Esmieu P, Dubertrand A, Crampe F (2017) PHYTOBS v3. 0: Phytoplankton counting tool in laboratory and calculation of IPLAC. Version 3.0. Java application.

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17: 10–12. https://doi.org/10.14806/ej.17.1.200

McLaren MR, Willis AD, Callahan BJ (2019) Consistent and correctable bias in metagenomic sequencing experiments. Turnbaugh P, Garrett WS, Turnbaugh P, Quince C, Gibbons S (Eds). eLife 8: e46923. https://doi.org/10.7554/eLife.46923

Nübel U, Garcia-Pichel F, Muyzer G (1997) PCR primers to amplify 16S rRNA genes from cyanobacteria. Applied and Environmental Microbiology. Available from: https://journals.asm.org/doi/abs/10.1128/aem.63.8.3327-3332.1997 (January 12, 2022).

Presting, G. G. Identification of conserved regions in the plastid genome: implications for DNA barcoding and biological function. Available from: https://cdnsciencepub.com/doi/abs/10.1139/B06-117 (January 12, 2022).

QIAO L, REN C, SUN X, SONG K, LI T, MU X (2021) SELECTIVE FEEDING OF TWO BIVALVE SPECIES ON THE PHYTOPLANKTON COMMUNITY IN AN AQUACULTURE POND REVEALED BY HIGH-THROUGHPUT SEQUENCING. APPLIED ECOLOGY AND ENVIRONMENTAL RESEARCH 19: 4477–4491.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2012) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic acids research 41: D590–D596.

Rimet F, Gusev E, Kahlert M, Kelly MG, Kulikovskiy M, Maltsev Y, Mann DG, Pfannkuchen M, Trobajo R, Vasselon V (2019) Diat. barcode, an open-access curated barcode library for diatoms. Scientific Reports 9: 1–12.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and environmental microbiology 75: 7537–7541.

Sherwood AR, Presting GG (2007) Universal primers amplify a 23S rDNA plastid marker in eukaryotic algae and cyanobacteria 1. Journal of phycology 43: 605–608.

Stoeck T, Bass D, Nebel M, Christen R, Jones MD, BREINER H-W, Richards TA (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Molecular ecology 19: 21–31.

Utermohl H (1958) Zur Vervollkommung der quantitativen phytoplankton-methodik. Mitt Int. Ver Limnol. 9: 38.

Vasselon V, Bouchez A, Rimet F, Jacquet S, Trobajo R, Corniquel M, Tapolczai K, Domaizon I (2018) Avoiding quantification bias in metabarcoding: Application of a cell biovolume correction factor in diatom molecular biomonitoring. Methods in Ecology and Evolution 9: 1060–1069.

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and environmental microbiology 73: 5261–5267.

Yoon T-H, Kang H-E, Kang C-K, Lee SH, Ahn D-H, Park H, Kim H-W (2016) Development of a cost-effective metabarcoding strategy for analysis of the marine phytoplankton community. PeerJ 4: e2115.

Matériels supplémentaires

L’ensemble des matériels supplémentaires évoqués dans ce rapport a été regroupé sur la plateforme Data INRAE à l’adresse suivante : https://data.inrae.fr/dataset.xhtml?persistentId=doi:10.15454/DI74PS .