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

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

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

CANINO Alexis (UMR CARRTEL), LAPLACE-TREYTURE Christophe (UR EABX), BOUCHEZ Agnès (UMR CARRTEL), DOMAIZON Isabelle (UMR CARRTEL), RIMET Frédéric (UMR CARRTEL).

Décembre 2020

SOMMAIRE

I. Contexte 3

II. Rappel des objectifs du projet pour 2020 3

III. Liste des taxons observés en microscopie 4
III.1 Liste des taxons observés en microscopie dans les DROM 4
III.2 Listes de taxons étendues 4

IV. Marqueurs ADN candidats 5
IV.1 Caractéristiques et résumé des principaux marqueurs ADN 5
IV.2 Bibliothèques de séquences utilisées 6
IV.2.1 Silva 6
IV.2.2 PR2 7
IV.2.3 PhytoRef 7
IV.2.4 µgreen-db 7
IV.2.5 NCBI – GenBank 7

V. Homogénéisation des nomenclatures taxonomiques des différentes bases de données 8

VI. Gap-analyses 8
VI.1 Présence des taxons DROM dans PHYTOBS et l'IPLAC 8
VI.2 Présence des taxons DROM dans les bibliothèques de référence ADN 9
VI.3 Présence des taxons PHYTOBS et IPLAC dans les bibliothèques de références ADN 9

VII. Définition des couples de primers pour amplifier des barcodes de chaque marqueur 10
VII.1 Notions clés en metabarcoding 10
VII.2 Méthodologie 10
VII.3 Critères de sélection des primers 11
VII.4 Marqueurs investigués 13
VII.4.1 ARNr 16S 13
VII.4.1.1 Résumé des avantages et inconvénients du marqueur 13
VII.4.1.2 Jeu de données utilisé pour designer les couples d'amorces 13
VII.4.1.3 Couples de primers candidats et critères de sélection 15
VII.4.2 ARNr 18S 18
VII.4.2.1 Résumé des avantages et inconvénients du marqueur 18
VII.4.2.2 Jeu de données utilisé pour designer les couples d'amorces 19
VII.4.2.3 Couples de primers candidats et critères de sélection 21
VII.4.3 ARNr 23S 24
VII.4.3.1 Résumé des avantages et inconvénients du marqueur 24
VII.4.3.2 Jeu de données utilisé pour designer les couples d'amorces 24
VII.4.3.3 Couples de primers candidats et critères de sélection 26
VII.4.4 rbcL 28
VII.4.4.1 Résumé des avantages et inconvénients du marqueur 28
VII.4.4.2 Jeu de données utilisé pour designer les couples d'amorces 28
VII.4.4.3 Couples de primers candidats et critères de sélection 30
VII.4.5 tufA 31
VII.4.5.1 Résumé des avantages et inconvénients du marqueur 31
VII.4.5.2 Jeu de données utilisé pour designer les couples d'amorces 31
VII.4.5.3 Couples de primers candidats et critères de sélection 33

VIII. Protocole d'échantillonnage 36

IX. Discussion et perspectives 36
IX.1 Liste taxonomique 36
IX.2 Choix du barcode 36
IX.3 Stratégie et design de primers 37
IX.4 Les perspectives du projet pour 2021 37

Bibliographie 39

Annexes 41

I. Contexte

Actuellement, les suivis réalisés dans le but d'évaluer l'état écologique des plans d'eau sont basés sur des analyses chimiques et biologiques soumises à des directives (DCE) et des normes Européennes/Françaises. Très peu de méthodes ont été développées spécifiquement pour les plans d'eau des DROM. A défaut, certaines des méthodes développées pour la métropole sont également appliquées pour la surveillance des plans d'eau dans les DROM. Ainsi, pour le maillon biologique phytoplancton, l'indicateur IPLAC (Laplace-Treyture & Feret, 2016) a pu être mis en œuvre pour le plan d'eau de Gaschet à des fins de test (Laplace-Treyture, 2020). Les espèces des communautés phytoplanctoniques de métropole sont utilisées comme bioindicatrices, à défaut d'être en mesure d'utiliser celles propres aux DROM pour l'instant. Les méthodes normalisées demandent une identification et un comptage de ces organismes par microscopie inversée. Ce travail qui demande une expertise taxonomique difficile à acquérir, est long ; de plus la différenciation de certaines espèces peut s'avérer très difficile, même pour les opérateurs les plus expérimentés. Ceci peut entraîner des incertitudes dans l'identification des taxons et donc dans l'évaluation de l'état des plans d'eau. D'autre part, la connaissance des taxons du phytoplancton en milieu tropical est rare, ce qui ajoute un frein à l'utilisation de ce maillon biologique pourtant essentiel pour comprendre le fonctionnement des plans d'eau.

Depuis plusieurs années, le séquençage à haut débit de l'ADN environnemental offre des perspectives intéressantes pour l'évaluation de l'état écologique des milieux aquatiques (Pawlowski et *al*., 2020) et notamment pour l'identification du phytoplancton. Une telle méthodologie, appelée metabarcoding ADN (Pompanon et *al*., 2011), pourrait ainsi venir en appui aux méthodes basées sur la microscopie. L'UMR CARRTEL développe depuis 2010 des méthodes d'identifications moléculaires en se basant sur de petits fragments d'ADN ou 'barcodes' ADN (Hebert et al. 2003) afin de caractériser la diversité des micro-algues aquatiques. En réalisant un séquençage massif d'ADN issus d'échantillons environnementaux, les espèces de microalgues peuvent être identifiées en comparant les résultats du séquençage à une base de référence de barcodes. L'UMR Carrtel a montré pour le bioindicateur diatomées que cette approche de metabarcoding ADN présente plusieurs avantages : elle permet de réduire les coûts et temps d'analyse (ex. Vasselon et al., 2017 ; Rivera et al., 2018). De plus cette méthode peut devenir accessible à des personnes non-spécialistes en taxonomie et elle permet de recenser des espèces difficilement identifiables en microscopie (e.g. picophytoplancton (<3µm) encore mal caractérisé, discrimination d'espèces au sein du genre *Scenedesmus* (Lürling, 1999) ...).

Tous ces avantages constituent des atouts notamment dans les DROM, pour lesquels l'expertise taxonomique est rare et où les plans d'eaux d'eau douce ont une importance capitale.

L'objectif général du projet est de proposer une méthode d'analyse pour le phytoplancton, simple et opérationnelle sur le terrain, ainsi que des outils d'évaluation d'état écologique robustes applicables aux plans d'eau des DROM, basés sur le metabarcoding ADN.

II. Rappel des objectifs du projet pour 2020

1. Liste des taxons phytoplanctoniques identifiés en microscopie dans les DROM :

Une check-list taxonomique doit être établie à partir des identifications microscopiques réalisées dans le cadre des suivis DCE des lacs des DROM. Cette liste s'appuiera sur les données de suivi DCE des lacs des DROM collectées par Angélique Bonnet.

2. *Liste des marqueurs ADN :*

Plusieurs marqueurs ADN sont disponibles dans la littérature pour identifier les taxons phytoplanctoniques (eucaryotes et procaryotes). Nous les listerons et donnerons leurs origine et fonction cellulaire.

3. Gap–analysis des bases de référence :

Sur la base de la check–list établie et de la liste des marqueurs ADN établie, nous évaluerons la complétude des bibliothèques de référence de barcodes ADN.

4. *Fiche résumé pour chaque marqueur* :

Dans ces fiches, les séquences des couples de primers seront données, et une synthèse des avantages et inconvénients de chaque barcode sera donnée.

5. *Protocole d'échantillonnage* adapté au metabarcoding :
Un protocole d'échantillonnage sur le terrain adapté au metabarcoding sera proposé.

III. Liste des taxons observés en microscopie

III.1 Liste des taxons observés en microscopie dans les DROM

Les observations microscopiques du phytoplancton réalisées dans les DROM par différents bureaux d'études ont été rassemblées par Angélique Bonnet (CDD sur projet OFB les années précédentes). Ces données ont été reprises et analysées dans le cadre de ce projet. Pour être utilisables, une curation des noms taxonomiques a été nécessaire (fautes de frappe, fautes de majuscules, espaces en fin de mots, erreurs d'orthographe dans les noms de genres ou espèces).
La liste des taxons observés dans les DROM est basée sur 11 plans d'eau : 5 en Guadeloupe, 4 à Mayotte, 1 en Martinique et 1 en Guyane (Tableau 1).

Tableau 1 : liste des DROM et de leurs plans d'eau associés, dans lesquels ont été réalisées les observations microscopiques.

DROM	Nom du plan d'eau	Données disponibles (années couvertes)
Guadeloupe	Gaschet	2017-2018
	Grand-Etang	2018-2019
	Letaye	2018-2019
	Mitan	2018-2019
	Zombi	2018-2019
Guyane	Retenue de Petit-Saut	2016 à 2018
Martinique	La Manzo	2006, 2007 et 2012 à 2019
Mayotte	Carrières	2006-2007
	Combani	2006 à 2009
	Dzoumougne	2006-2007
	Karihani	2006-2007
Réunion	Gol	Non-réalisées/non-disponibles
	Grand-Etang	Non-réalisées/non-disponibles
	St-Paul	Non-réalisées/non-disponibles

Un premier panorama combine les espèces observées sur l'ensemble des 11 plans d'eau, toutes années disponibles confondues. 670 taxons différents ont été observés. Si on fait abstraction de taxons ne présentant pas d'information taxonomique (ex. « flagellate >10µm »), la liste regroupe un total de 617 taxons différents. L'ensemble des taxons observé s'étend sur 11 phyla, 22 classes, 48 ordres, 92 familles et 199 genres différents.

Une liste de l'ensemble des taxons observées est disponible : *Matériel Supplémentaire 1.*

III.2 Listes de taxons étendues

Les DROM présentent un nombre relativement restreint de plans d'eau qui abritent une diversité phytoplanctonique pouvant être propre à ces derniers. Dans le cadre de ce projet, qui a pour objectif de mettre en place un outil robuste, nous avons cherché des listes de taxons plus complètes afin de ne pas se limiter seulement à certains taxons observés.
Le logiciel PHYTOBS (Laplace-Treyture et *al.*, 2017 https://hydrobio-dce.inrae.fr/phytobs/) est dédié à la bancarisation et au comptage microscopique du phytoplancton des plans d'eau et permet le calcul de l'indice d'évaluation de l'état écologique des plans d'eau - IPLAC. Ce logiciel référence un grand nombre d'espèces (notée 'liste PHYTOBS' dans ce rapport). Cette liste a été utilisée pour avoir une liste taxonomique étendue de la diversité du phytoplancton existant dans les plans d'eau. A ce jour, cette

liste totalise 8832 codes, chacun étant associé à un genre, une espèce et parfois un niveau infraspécifique.
Nous avons également regardé les espèces bio-indicatrices prises en compte dans le calcul de l'indice IPLAC (Laplace-Treyture & Feret, 2016). Cet indice s'appuie actuellement sur 165 espèces bio-indicatrices. Ces dernières ont été regroupées dans une liste appelée 'liste IPLAC' dans ce rapport. L'ensemble des espèces présentes dans la liste IPLAC sont incluses dans la liste PHYTOBS.

IV. Marqueurs ADN candidats

Le phytoplancton *sensu stricto* rassemble les organismes photosynthétiques procaryotes (cyanobactéries) et eucaryotes unicellulaires ou coloniaux vivant en suspension dans l'eau. Nous avons réalisé plusieurs recherches bibliographiques qui ont permis d'établir une sélection des marqueurs les plus souvent utilisés pour réaliser leur identification moléculaire. Leurs caractéristiques et fonctions sont données en **IV.1.** Les bibliothèques de références rassemblant des données taxonomiques pour ces marqueurs et ayant été utilisées dans ce projet sont présentées en **IV.2.**

IV.1 Caractéristiques et résumé des principaux marqueurs ADN

Les marqueurs ADN les plus couramment utilisés pour décrypter la diversité moléculaire des procaryotes et eucaryotes sont des fragments variables des petites sous-unités ribosomales : ARNr 16S (pour les procaryotes et les génomes plastidiaux) et ARNr 18S (pour les eucaryotes). Ces gènes ribosomaux sont universels car ils assurent une fonction essentielle pour tous les organismes vivants : la traduction de l'ARN messager en protéines. Le 18S a une longueur d'environ 1800 pb (paire de bases), et le 16S une longueur d'environ 1500 bp. Ces deux marqueurs ont l'avantage de présenter dans leur séquence des régions bien conservées et d'autres variables : i) les régions bien conservées permettent de constituer des amorces utilisables pour des groupes taxonomiques différents, ii) les régions variables, qui elles permettent de discriminer les espèces entre elles en se basant sur la diversité génétique de cette région d'ADN. Ces marqueurs sont présents en plusieurs copies dans le génome, ils sont donc présents en grande quantité dans les cellules, ce qui facilite leur amplification par PCR (Polymerase Chain Reaction).

Les chloroplastes des eucaryotes photosynthétiques sont des descendants de cyanobactéries, qui ont été intégrés dans les cellules eucaryotes au cours de l'évolution par le phénomène d'endosymbiose (Whatley, 1993). L'ADN plastidial (aussi appelé ADN chloroplastidial) est donc un matériel génétique procaryote 'commun' entre les chloroplastes et les cyanobactéries (ayant bien sûr évolué indépendamment au fil du temps). Il y a donc également dans le génome plastidial des gènes codant pour la synthèse de protéines ayant un rôle dans l'activité photosynthétique. L'utilisation de tels marqueurs pour le métabarcoding du phytoplancton permet de sélectionner uniquement les organismes photosynthétiques, ce qui est un avantage par rapport au 18S et 16S qui eux sont généralistes et couvrent toute la diversité des organismes (photosynthétiques et non photosynthétiques).
Les marqueurs qui ont déjà été utilisés dans des études taxonomiques et de décryptage de la diversité moléculaire d'échantillons naturels sont :

*rbc*L : gène présent sur l'ADN plastidial qui code pour une enzyme essentielle aux producteurs primaires -la RuBPC/O- qui assure la fixation du CO_2.

*tuf*A : gène présent sur l'ADN plastidial qui code pour une fonction essentielle : la synthèse du facteur d'élongation Tu (EF-Tu).

Tout comme les chloroplastes, les mitochondries sont des organites présents dans les cellules eucaryotes à la suite d'une endosymbiose (Gray, 2012). Des gènes de l'ADN mitochondrial (aussi appelé ADNmt) sont souvent utilisés et reconnus comme barcodes, comme par exemple le COI pour le règne animal (Hebert et *al.*, 2003). Or, ce dernier n'est pas optimal pour le barcoding des lignées vertes (Meier et *al.*, 2006).
La figure 1 reprend la localisation cellulaire des différents marqueurs cités ci-dessus.

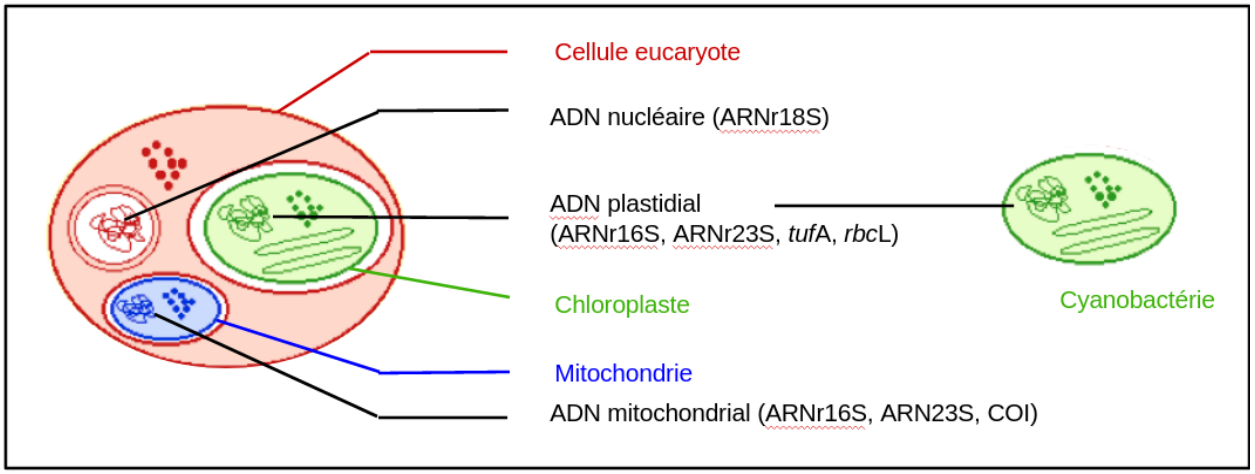

Figure 1 : schéma représentant les localisations des différents organites (et leur matériel génétique associé) au sein d'une cellule eucaryote phytoplanctonique. Les marqueurs investigués ici sont indiqués entre parenthèses.

IV.2 Bibliothèques de séquences utilisées

Pour chaque marqueur cité ci-dessus, nous avons recherché des bibliothèques de séquences publiques rassemblant des séquences de taxons phytoplanctoniques avec leur nom taxonomique. Le tableau 2 donne pour chaque marqueur, son origine et les bases de référence consultées.

Tableau 2 : liste des marqueurs investigués dans cette étude, leur origine et les bases de référence consultées avec les nombres de séquences qu'elles possèdent.

Marqueur ADN	Origine	Base(s) de référence publiques consultées	Nombre de séquences	
			total	spécifiques phytoplancton*
16S	Chloroplastes, Cyanobactéries	Silva 138	2225272	6881
		pr2 - 16S	6010	1174
		PhytoRef	9190	1336
18S	Nucléaire, Eucaryotes	Silva 138	2225272	9147
		pr2 - 18S	177934	9061
23S	Chloroplastes, Cyanobactéries	Silva 138	227331	1330
		µgreen-db	2326	1432
*rbc*L	Chloroplastes, Cyanobactéries	NCBI-GenBank	187709	11391
*tuf*A	Chloroplastes, Cyanobactéries	NCBI-GenBank	5453	1974

*après homogénéisation de la taxonomie avec AlgaeBase (*cf.* **partie 5.**) et comparaison avec PHYTOBS pour conserver uniquement les séquences appartenant à ces taxons.

IV.2.1 Silva

SILVA (Quast et *al.*, 2013) est une bibliothèque de séquences qui fournit des ensembles de données complets, contrôlés d'un point de vue de leur qualité et régulièrement mis à jour pour des séquences d'ARN ribosomal (ARNr) des petites (16S/18S, SSU) et grandes sous-unités (23S/28S, LSU) pour les trois domaines de la vie (bactéries, archées et eucaryotes). Cette base est disponible sur https://www.arb-silva.de/.

IV.2.2 PR2

PR2 (Guillou et *al.*, 2013) est une bibliothèque qui rassemble uniquement des séquences codantes pour la petite sous-unité ribosomal (ARNr 18S / SSU) présente uniquement chez les eucaryotes). Les données de cette base sont contrôlées par des experts pour les différents groupes taxonomiques eucaryotes. Cette dernière est disponible sur https://pr2-database.org/. PR2 propose également des séquences 16S (uniquement plastidiales ou appartenant à des cyanobactéries) issues de PhytoRef.

IV.2.3 PhytoRef

PhytoRef (Decelle et *al.*, 2015) est une bibliothèque regroupant uniquement des séquences codantes pour la petite sous-unité ribosomale des procaryotes (ARNr 16S / SSU). Elle regroupe les séquences appartenant aux cyanobactéries et aux chloroplastes (présents donc également chez les organismes phytoplanctoniques eucaryotes). Les séquences de cette bibliothèque sont également régulièrement contrôlées par des experts et des taxonomistes spécialisés dans le phytoplancton, et sont disponibles à l'adresse suivante http://phytoref.sb-roscoff.fr/ .

IV.2.4 µgreen-db

µgreen-db est une base de référence récemment créée par des experts (Djemiel et *al.*, 2019) qui regroupe uniquement des séquences codantes pour la grande sous-unité ribosomale des procaryotes (cyanobactéries mais aussi chloroplastes présents chez les eucaryotes) : ARNr 23S / LSU. Un intérêt particulier est porté sur le domaine V de celui-ci, aussi connu sous le nom de UPA (*Universal Plastid Amplicon*). La première version de cette bibliothèque est disponible en ligne http://microgreen-23sdatabase.ea.inra.fr/ .

IV.2.5 NCBI – GenBank

NCBI (National Center for Biotechnology Information) est une plateforme en ligne qui héberge plusieurs bases de données, centrées sur la biologie moléculaire. Parmi celles-ci, on y retrouve des bibliothèques de séquences ADN de toutes sortes, c'est le cas de GenBank (Benson et *al.*, 2005). De nombreuses séquences sont référencées puisque tout chercheur peut y en ajouter, cependant ces dernières ne font pas l'objet de validation, ce qui peut poser des problèmes pour des usages ultérieurs. NCBI propose également des bibliothèques de séquences vérifiées (*e.g.* RefSeq). Toutes ces données sont disponibles sur la plateforme NCBI https://www.ncbi.nlm.nih.gov/ .

V. Homogénéisation des nomenclatures taxonomiques des différentes bases de données

Afin de comparer les différentes bases de données entre elles (listes floristiques des différents plans d'eau, listes de taxon des bibliothèques de séquences de référence), il a fallu rendre leurs nomenclatures taxonomiques homogènes. En effet, ces nomenclatures taxonomiques sont retrouvées sous de nombreuses et diverses formes selon les bases de données (sans compter les erreurs d'orthographe des noms taxonomiques…). Ces corrections et cette homogénéisation des nomenclatures a été un important travail préalable nécessaire pour pouvoir réaliser les Gap-analyses. La nomenclature taxonomique d'AlgaeBase (Guiry & Guiry, 2020) a été utilisée comme référence pour homogénéiser les différentes bases de données, et ainsi pouvoir les comparer : AlgaeBase est une base de données taxonomiques qui est alimentée par une centaines d'experts internationaux, c'est la meilleure référence en termes de taxonomie algale et elle est consultable en ligne gratuitement (https://www.algaebase.org/).
Le tableau 3 reprends les différentes bases de données qui ont été homogénéisées taxonomiquement.

Tableau 3 : résumé des différentes bases ayant été homogénéisées avec la taxonomie d'AlgaeBase. La source des différentes bases est donnée.

Bases de données	Intérêt(s)	Source	Lien vers fichier correspondant
PHYTOBS	Représente la diversité globale du phytoplancton dans les plans d'eau	https://hydrobio-dce.inrae.fr/phytobs/	Mat.Supp. 2
IPLAC	Regroupe les espèces ayant une valeur bio-indicatrice dans l'indice IPLAC	http://seee.eaufrance.fr/	Mat.Supp. 3
DROM	Liste les espèces phytoplanctoniques observées dans les plans d'eau des DROM	Créée à partir des observations des bureaux d'études (rassemblées par A. Bonnet)	Mat.Supp. 1
AlgaeBase	Base de données de référence mondiale pour la taxonomique des algues. Elle regroupe tous les taxons et décrit leur classification taxonomique	https://www.algaebase.org/ 08.2020	Mat.Supp. 4
Listes ADN	Rechercher les identités moléculaires des espèces phytoplanctoniques présentes dans les DROM et celles sur la liste PHYTOBS de façon plus générale.	La liste des différentes sources est donnée en 4.2.	Mat.Supp. 5 (dossier)

VI. Gap-analyses

Les « gap-analyses » réalisées ci-après ont pour objectif de :
(VI.1) vérifier la présence des espèces observées dans les DROM au sein de PHYTOBS et voir celles étant considérées comme bio-indicatrices (dans l'IPLAC) ;
(VI.2) vérifier la couverture de la liste DROM pour les différents barcodes ADN ;
(VI.3) vérifier la couverture des listes PHYTOBS et IPLAC pour les différents barcode ADN.

VI.1 Présence des taxons DROM dans PHYTOBS et l'IPLAC

La présence de l'ensemble de ces espèces dans la liste PhytObs a été vérifiée grâce à un script R. Le tableau 4 donne pour chaque niveau taxonomique le nombre de taxons identifiés dans les DROM qui sont pris en compte dans PHYTOBS et dans l'IPLAC.
Les taxons sans valeurs taxonomiques (*e.g.* « Dino A » ; « Dino B ») ont été retirés, ce qui a permis d'arriver à un total de 617 taxon différents. Cependant, comme l'explicite le tableau 4, de nombreux taxons (311) restent non-identifiés au niveau spécifique (*e.g.* « *sp.* » ; « *nd.* »). Cependant, 306 taxons sont associés à une espèce définie et 302 d'entre-elles sont retrouvées dans PHYTOBS. En ce qui concerne les taxons avec une valeur bio-indicatrice (liste IPLAC) seulement 55% des espèces de cette liste sont couvertes par les observations dans les DROM. En effet, parmi les 165 taxons présents sur

la liste IPLAC, 92 d'entre eux ont été observés dans les DROM. Etant donné que la liste IPLAC met en évidence des espèces bio-indicatrices, il n'a pas été jugé nécessaire de décrire la couverture des rangs taxonomiques supérieures.

Tableau 4 : taxons observés dans les DROM et pris en compte dans PHYTOBS et dans l'IPLAC. (Les taxons absents dans PHYTOBS pour les rangs taxonomiques de la classe jusqu'au genre sont des taxons retrouvés en milieux marins ou saumâtres).

	Nombre de taxons total	Règne	Embran-chement	Classe	Ordre	Famille	Genre	Espèce
DROM	617	5	11	22	48	92	199	306/617
PHYTOBS	8832	5	11	21	47	91	195/199	302/306
IPLAC	165	-	-	-	-	-	-	92/165

VI.2 Présence des taxons DROM dans les bibliothèques de référence ADN

La vérification de la présence des espèces identifiées dans les DROM dans les différentes bases de références pour les différents marqueurs ADN a été réalisée à l'aide d'un script R. Le tableau 5 résume les résultats de ces gap-analyses. Comme le montre celui-ci, les taxons observés dans les DROM sont peu représentés dans les bases de données ADN. Pour l'ensemble des marqueurs, les bibliothèques de référence sont incomplètes surtout si le niveau spécifique est considéré.

Tableau 5 : couverture (en %) des taxons observés dans les DROM par les différentes bases de référence ADN investiguées.

	16S			18S		23S		*rbc*L	*tuf*A
	Silva	PhytoRef	pr2	Silva	pr2	Silva	µgreen-db	NCBI	NCBI
Genres (%)	47,24	30,15	25,13	66,33	66,83	35,68	33,17	63,82	31,66
Especes (%)	4,54	5,19	4,38	17,02	21,72	3,08	3,89	11,83	4,05

VI.3 Présence des taxons PHYTOBS et IPLAC dans les bibliothèques de références ADN

La présence des espèces de PHYTOBS et de l'IPLAC dans les différentes bibliothèques de séquences et pour les différents marqueurs ADN a également été réalisée avec un autre script R.

Tableau 6 : pourcentages des taxons de PHYTOBS et de l'IPLAC dont des séquences pour les différents marqueurs sont enregistrées dans les bibliothèques ADN de référence.

		16S			18S		23S		*rbc*L	*tuf*A
(%)		Silva	Phyto Ref	pr2	Silva	pr2	Silva	µgreen-db	NCBI	NCBI
PHYTOBS	Genres	33,07	23,14	14,33	50,45	47,63	22,12	20,54	50,23	19,86
PHYTOBS	Espèces	2,58	2,67	1,55	9,85	9,75	1,42	2,25	7,47	1,44
IPLAC	Espèces	13,33	10,91	7,88	44,85	47,88	7,88	10,30	34,55	11,52

Remarque : indiquer les genres de la liste IPLAC représentés par les différents marqueurs ADN n'est pas particulièrement utile ; en effet, si une espèce appartient à un genre présent dans l'IPLAC mais que l'espèce associée au genre en question n'apparait pas dans la liste, alors cela peut prêter à confusion. Cette information n'a donc pas été retenue pour ce tableau.

VII. Définition des couples de primers pour amplifier des barcodes de chaque marqueur

VII.1 Notions clés en metabarcoding

Ce qui est généralement qualifié de 'barcode' en metabarcoding correspond à une portion de séquence d'ADN dont le code oligonucleotique (génétique) permet de discriminer les espèces les unes des autres. Le barcode peut provenir de différentes séquences d'ADN (*e.g.* gènes, portions de gènes…) selon le contexte de l'étude (*i.e.* groupes d'espèces que l'on veut cibler et discriminer). Il doit posséder une variabilité des nucléotides suffisante pour différencier les espèces entre elles, on parle de résolution. Enfin, sa taille est courte, généralement quelques dizaines ou centaines de paires de bases (longueur en paires de bases –abrégé *pb* dans la suite du rapport-) doit être compatible avec la technologie de séquençage utilisée.
Pour obtenir ces barcodes, il faut identifier des zones dites primers (ou amorces) positionnées aux 2 extrémités de la région souhaitée (*i.e.* barcode) ; ces primers permettent une hybridation spécifique sur la zone d'intérêt pour réaliser la réaction de PCR (Polymerase Chain Reaction). Le primer se fixant en amont de la zone à amplifier est appelé forward, celui en aval est appelé reverse. Les zones où les primers s'hybrident vont constituer des zones de fixation de la Taq polymérase (une enzyme capable de synthétiser le brin d'ADN complémentaire d'un autre brin matrice). A la fin de la réaction de PCR, une grande quantité de barcodes est synthétisée, on parle également d'amplicons pour désigner ces derniers.

Pour ce projet, le barcode idéal devra être présent chez des organismes appartenant à des groupes taxonomiques diversifiés : l'ensemble du phytoplancton (dont des eucaryotes et des procaryotes). Un barcode présent sur un gène chloroplastidial par exemple est un bon candidat, puisqu'il limiterait le champ des possibilités aux organismes photosynthétiques, comme le phytoplancton. Il devra être capable d'avoir une résolution suffisante pour discriminer des espèces : c'est un point important notamment en bio-indication / bio-monitoring. Enfin, étant donné que la technologie de séquençage utilisé dans ce projet sera le MiSeq 2x250pb (Illumina), la taille du barcode ne devra pas excéder environ 450 pb (primers inclus).

L'objectif est de faire ressortir pour chacun des marqueurs les avantages et inconvénients liés à leur utilisation pour du metabarcoding du phytoplancton. A cela, une liste de primers utilisés dans la littérature sera ajoutée, ainsi que des primers qui auront été « designés » (ou dessinés) dans le cadre de ce projet. La résolution des régions qu'ils permettent d'amplifier sera évaluée et constituera un argument important dans le choix du/des primer(s) à utiliser.

VII.2 Méthodologie

Pour chaque marqueur (18S, 16S, 23S, *rbc*L, *tuf*A), des couples de primers candidats ont été définis de deux manières différentes :

1. Sur base bibliographique : une revue bibliographique a été réalisée pour chaque marqueur afin d'identifier les primers les plus souvent utilisés en metabarcoding
2. Sur base bioinformatique : à l'aide des bibliothèques de référence évoquées précédemment (Silva, pr2 *etc.*) nous avons recherché s'il était possible de définir de nouveaux primers. Pour cela la méthodologie suivante a été utilisée pour chaque marqueur :

a. Les séquences du marqueur considéré (ex. 16S) présentes dans les différentes bibliothèques (ex. Silva, pr2, PhytoRef) ont été combinées dans un jeu de données commun.
b. Les redondances dans le jeu de données commun ont été traitées de manière à ne conserver qu'une seule séquence associée à une espèce. Ainsi il n'y avait pas d'espèces surreprésentées, car cela pourrait entraîner des biais dans le design de primers qui serait alors plus spécifiques pour ces espèces. De la même manière, lorsqu'un groupe taxonomique entier était surreprésenté par rapport aux autres, l'analyse du jeu de données était optimisée en prenant en compte cette considération afin que les primers « designés » ne soient pas trop influencés par celui-ci. En effet, il est important que les primers amplifient de manière homogène toute la richesse en microalgues potentiellement présentes dans un échantillon de phytoplancton.

c. L'alignement des séquences récupérées est une étape cruciale pour rechercher les zones conservées qui contiennent de potentiels primers. Dans le cadre de ce projet, les séquences sont nombreuses et diversifiées dans l'arbre du vivant (l'ensemble du phytoplancton…). L'alignement multiple d'un tel jeu de données est chronophage et le résultat ne serait pas exploitable pour la recherche de séquences conservées : l'algorithme insère de nombreux 'gaps' (*i.e.* des espaces vides) pour les positions qu'il n'arrive pas à aligner localement.

La stratégie choisie dans cette étude est donc différente, de manière à pallier ce problème. Les résultats de celle-ci permettront de voir son efficacité pour le design de primers généralistes adaptés au metabarcoding d'un groupe taxonomique cible.

La stratégie se décompose comme suit :

- Partitionnement aléatoire du jeu de séquences en plusieurs sous-jeux de données (ex. jeu de données initial de 1000 séquences partitionné en 100 sous-jeux de 10 séquences).
- Alignements multiples de chacun des sous-jeux de séquences (les alignements sont ainsi rapides, et plus fiables). Les alignements sont réalisés grâce à Muscle (Edgar, 2004), avec les paramètres par défaut.
- Recherche des zones conservées et motifs similaires dans tous les alignements réalisés, les plus redondants sont conservés et soumis à différents critères (voir § suivant) qui font d'eux des primers optimaux. Ces opérations sont réalisées grâce à des scripts Python et au module BioPython (Cock et *al.*, 2009).
- Les primers candidats récupérés sont testés par PCR *in silico* grâce à des commandes de Mothur (Schloss et *al.*, 2009) sur les jeux de séquences correspondants (*cf.* suite).

L'intégralité de ce pipeline bio-informatique, qui regroupe différents langages informatiques, pour designer des primers a été intégré dans Rstudio dans le but d'être facilement reproductible.

VII.3 Critères de sélection des primers

Une fois des couples de primers sélectionnés sur base bibliographique et bioinformatique pour les différents marqueurs, certaines de leurs caractéristiques ont été mises en avant afin de les comparer et de conclure sur leur efficacité. Parmi celles-ci, il faudra prendre en considération :

- **La taille moyenne de l'amplicon** : celle-ci doit être compatible avec la technologie MiSeq (Illumina) 2x250pb, comme évoqué précédemment (*cf.* VII.1).
- **La température hybridation** : chaque primer possède une température d'hybridation qui lui est propre (à laquelle il va pouvoir s'hybrider avec la séquence d'ADN complémentaire). Les températures recommandées pour une réaction de PCR optimale sont dans une gamme entre 52 et 58°C. La différence des températures entre le primer forward et le reverse doit être la plus faible possible (et il est couramment indiqué qu'elle ne doit pas excéder 5°C) pour garantir une amplification optimale. Cette température d'hybridation se calcule via différents algorithmes, qui se basent principalement sur le contenu en nucléotides G et C mais donnent des résultats assez différents. Pour la sélection des couples de primers en bioinformatique, l'algorithme d'un package R (TmCalculator) a été utilisé dans un premier temps afin d'avoir une pré-sélection automatique de ce critère (applicable rapidement à de nombreux primers). Par la suite, l'algorithme implémenté dans l'outil OligoAnalyzer (Owczarzy et *al.*, 2008 ; www.idtdna.com) a été utilisé pour revérifier ce critère pour les meilleurs couples de primers candidats retenus. Les résultats de cet algorithme, différents de celui utilisé dans R, ont été retenus et seront présentés ici.
- **Potentiels problèmes liés à l'hybridation ou à la configuration de la molécule** : c'est également l'outil en ligne OligoAnalyzer qui permet d'estimer si les primers sont susceptibles de poser des problèmes lors de la réaction de PCR. Par exemple, ces derniers peuvent, à une température donnée, avoir une structure tridimensionnelle repliée sur elle-même. Cette conformation, qualifiée de 'hairpin' empêche toute hybridation si elle est présente à une température similaire ou supérieure à la température d'hybridation recommandée. La présence de ce phénomène sera symbolisée par 'Hairpin' (F ou R selon l'amorce concernée) dans les tableaux descriptifs suivants. Les primers peuvent aussi s'hybrider avec eux-mêmes, on parle d'homo-dimères ce phénomène peut réduire les chances que les primers s'hybrident avec l'ADN cible. Il sera symbolisé par 'Selfdimer'. De la même manière les primers forward et reverse peuvent également s'hybrider entre eux, on parle alors d'hétéro-dimères, symbolisés par la suite par 'Heterodimer'.
- **L'efficacité d'amplification *in silico* :** contrôlée grâce à des commandes sur Mothur simulant

des PCR (se basant uniquement sur la composition en nucléotide des séquences). Les paramètres suivant ont été recensés afin d'avoir une meilleure idée de l'efficacité des couples de primers :

- Nombre d'amplicons obtenus avec les primers sans aucun mismatch (*i.e.* chaque nucléotide composant le primers s'hybride parfaitement avec son complémentaire sur l'ADN cible).
- Nombre d'amplicons obtenus avec un maximum de 2 mismatches. Si quelques bases ne sont pas exactement similaires à celle de l'ADN cible, l'hybridation peut quand même avoir lieu en pratique (selon les conditions de températures, la composition du mélange réactionnel *etc.*). Un seuil de 2 mismatches maximum est défini car il peut être représentatif de PCR en conditions réelles. Ce seuil a été choisi pour définir le nombre de séquences à conserver par cette étape. La taille de l'amplicon devait également être dans un intervalle de ± 20 % de la taille médiane des amplicons obtenus pour s'assurer que la séquence amplifiée correspondait au barcode attendu.
- La résolution spécifique des amplicons obtenus. Cette dernière s'obtient en divisant le nombre de séquences uniques récupérées des amplicons par le nombre d'amplicons (avec 2 mismatches maximum autorisés). Elle permet de savoir combien d'espèces sont discriminés parmi les amplicons.

Dans la partie suivante, cette méthodologie est mise en œuvre pour chaque marqueur (18S, 16S, *etc.*). Pour chaque marqueur, un résumé des résultats et d'abord donné (avantages/inconvénients du marqueur et de ses couples de primers), puis le détail des bibliothèques de référence utilisées, des couples de primers et des critères de sélection sont donnés.

VII.4 Marqueurs investigués

VII.4.1 ARNr 16S

Gène codant pour la petite sous-unité ribosomale chez les procaryotes. Le 16S est également présent dans les plastes (chloroplastes) des eucaryotes. Il possède des régions variables et d'autres conservées, qui sont bien documentées (*cf*. figure 2). Or, l'ARNr 16S est souvent utilisé en écologie microbienne et les régions les plus résolutives ne seront pas forcément les mêmes pour le phytoplancton que pour caractériser la diversité microbienne totale. Pour vérifier laquelle des régions semble être la plus résolutive pour les taxons d'intérêt, une analyse de la résolution de chacune de ces dernière a été réalisée sur le jeu de donnée qui a également servi au design de primers (cf. § VII.4.1.2) et est disponible en annexe (*cf.* Annexe 1).

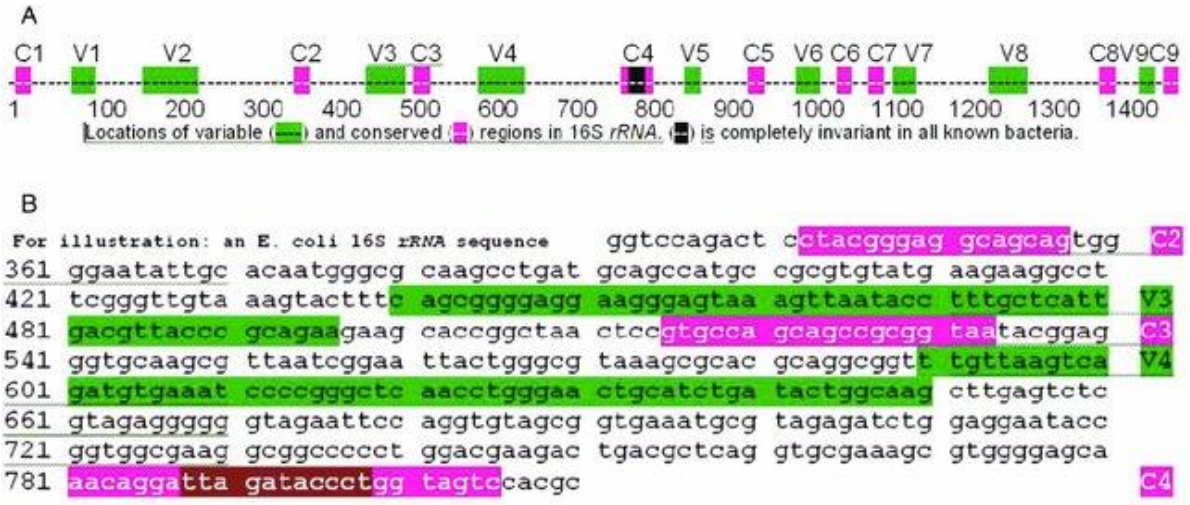

Figure 2 : représentation des régions variables et conservées sur l'ARNr 16S pour un organisme de référence : *E. coli* (de Ram et *al.*, 2011).

VII.4.1.1 Résumé des avantages et inconvénients du marqueur

Avantages :

+ Cible l'ensemble du phytoplancton : cyanobactéries (procaryotes) et eucaryotes via le génome des chloroplastes ;
+ Bases de références bien documentées ;
+ Certains couples d'amorces adaptés au phytoplancton sont renseignés dans la littérature ;
+ Localisation des régions variables et conservées connus (*cf.* Figure 2).

Inconvénients :

- Amplification de bactéries hétérotrophes (nombreuses et diversifiées dans les milieux ciblés) si les amorces ne sont pas assez spécifiques : cela peut conduire à une saturation de la profondeur de séquençage et potentiellement à une incapacité à caractériser efficacement la diversité du phytoplancton ;
- La résolution taxonomique de ce marqueur peut être faible pour certains taxons.

VII.4.1.2 Jeu de données utilisé pour « designer » les couples d'amorces

Les séquences présentes dans les bases Silva v.138 (Quast et *al.*, 2013), pr2 (Guillou et *al.*, 2013) et PhytoRef (Decelle et *al.*, 2015) ont été combinées dans un jeu de données commun, totalisant ainsi 2120 séquences. Comme indiqué précédemment (*cf.* 2b § VII.2), pour chaque espèce une seule séquence a été conservée parmi toutes celles redondantes. Le nombre de séquences sélectionnées pour chaque phylum pour le design des primers est donné dans la figure 3. La diversité taxonomique des séquences représentants chaque phylum a également été mise en évidence en calculant un indice

de la diversité taxonomique (Warwick & Clarke, 1995).

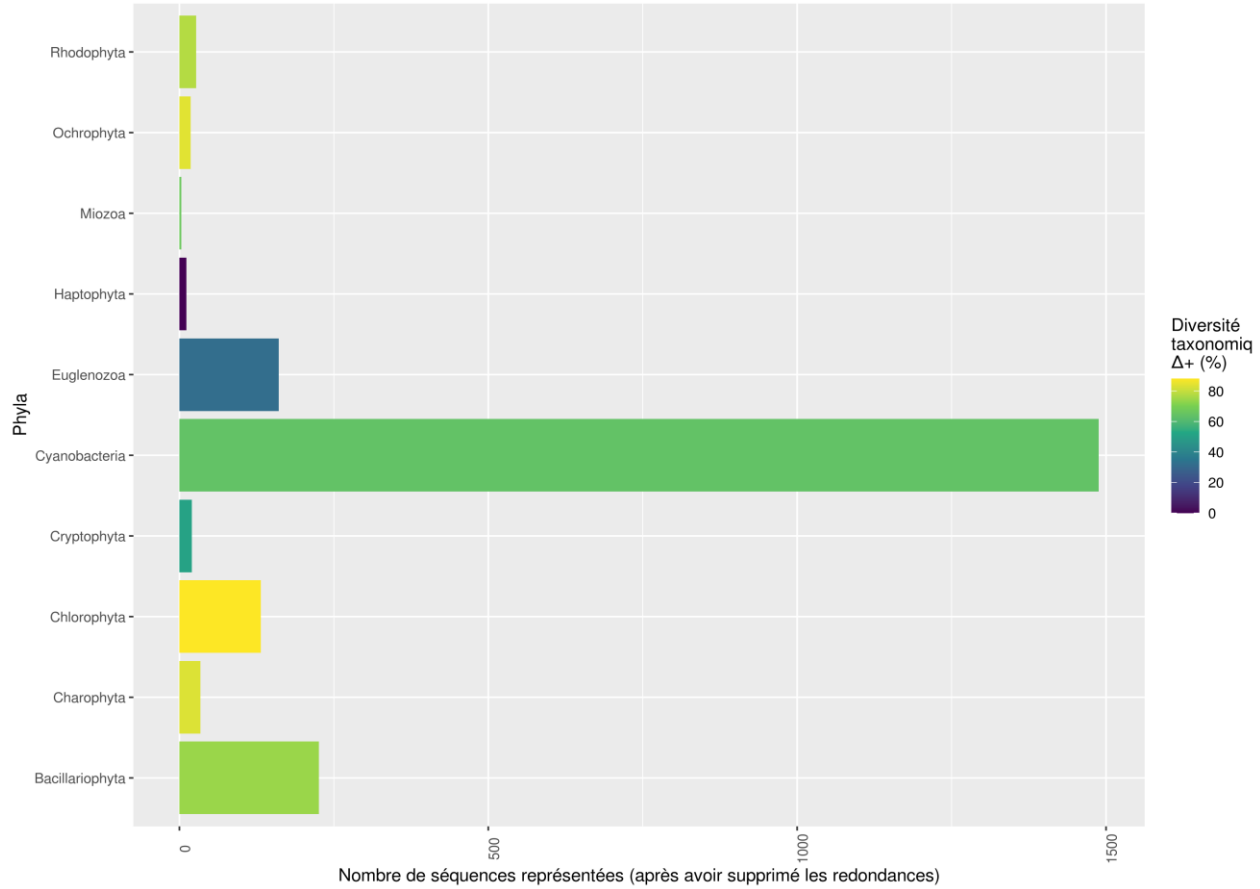

Figure 3 : Pour chaque phylum, le nombre de séquences 16S sélectionnées pour le design de primers est donnée. Ces séquences d'ARNr 16S sont issues d'une fusion entre les bases de données Silva (138), pr2 et PhytoRef. Si une espèce présentait plusieurs séquences, ou si des séquences présentaient des codes d'accession identiques, alors dans les deux cas, une seule des séquences a été conservée. La couleur des histogrammes représente la diversité taxonomique (Warwick & Clarke, 1995) au sein du phylum (0 : toutes les espèces appartiennent au même groupe taxonomique, 100 : chaque espèce appartient à un groupe taxonomique différent).

On observe dans la figure 3 qu'il y a davantage de séquences de Cyanobactéries (1488 en retirant les redondances pour une même espèce) que de séquences de microalgues eucaryotes (646). Cette différence a été prise en compte pour la recherche d'un couple de primers capable d'amplifier de manière efficace à la fois les Cyanobactéries et les microalgues eucaryotes. Pour cela, l'analyse a été réalisée à nouveau en séparant ces 2 lignées en 2 jeux de données différents. A partir des résultats obtenus, on a cherché des zones conservées communes sur lesquels un design de primers était possible.

VII.4.1.3 Couples de primers candidats et critères de sélection

Tableau 7: présentation des couples de primers récupérés de la littérature ou designés dans cette étude et leurs caractéristiques. Différentes informations sont fournies (pour plus de détails voir § VII.3.). Pour rappel, les températures d'hybridation et les potentiels problèmes *in vitro* sont donnés par OligoAnalyzer (idt-dna). Les informations relatives aux séquences procaryotes (*i.e.* cyanobactéries) sont représentées en bleu, celles des eucaryotes en vert et l'ensemble en noir.

Couple	Région variable	Taille moyenne amplicons	Tm* en °C (F/R)	Potentiels problèmes *in vitro**	Nombre séquences matchant exactement avec le couple	Nombre de séquences amplifiées et de longueur acceptable en autorisant 2 mismatches maximum	Nombre de séquences uniques récupérées parmi les amplicons (2mm)	Résolution spécifique	Etudes
515F/926R	V4-V5	413	63,6 / 48,9	SelfDimer F HeteroDimer	1992/2120 563/646 1442/1488	1989/2120 587/646 1415/1488	1201/1989 476/587 726/1415	60,38 % 81,09 % 51,3 %	Parada et *al.,* 2016 Watanabe et *al*., 2001 recommandé par Santoferrara 2019
341F/805R	V3-V4	443	57,5 / 51,3	SelfDimer F SelfDimer R	1793/2120 376/646 1427/1488	1845/2120 446/646 1414/1488	1156/1845 363/446 795/1414	62,65 % 81,39 % 56,22 %	Recommandé par Klindworth et *al.*, 2013
CYA359F/CYA781R	V3-V4	425	57,9 / 55,5	HeteroDimer	1611/2120 210/646 1408/1488	1756/2120 360/646 1407/1488	1070/1756 283/360 787/1407	60,93 % 78,61 % 55,93 %	Nübel et *al*., 1997 récemment utilisé par Ivanova et *al.*, 2019
PLA491F/805R	V4	315	52 / 51,3	Hairpin F Hairpin R	711/2120 155/646 558/1488	1677/2120 452/646 1236/1488	899/1677 345/452 555/1236	53,6 % 76,32 % 44,9 %	Füller et *al*., 2006 déjà utilisé à l'UMR CARRTEL

CYA781R/E1115R	V5-V6	336	55,5 / 56,1	HeteroDimer Hairpin R	1587/2120 311/646 1284/1488	1677/2120 400/646 1287/1488	881/1677 268/400 613/1287	52,53 % 67 % 47,63 %	Reysenbach & Pace, 1995
E939F/OXY1313R	V6-V7	417	65,1 / 60,6	Selfdimer F	30/2120 4/646 26/1488	1657/2120 391/646 1278/1488	1043/1657 311/391 733/1278	62,94 % 79,53% 57,35%	Rudi et *al*., 1997 West et *al*., 2001
PhytF/PhytR	V5-V6	386	56,6 / 58,6		1870/2120 458/646 1423/1488	1946/2120 538/646 1420/1488	1165/1946 421/538 745/1420	59,86 % 78,25 % 52,46 %	Cette étude
Test1F/test1R	V5-V6	385	54,8 / 59,8	HairpinR à 54°C	1871/2120 459/646 1423/1488	2030/2120 562/646 1481/1488	1243/2030 443/562 801/1481	61,23 % 78,82 % 54,08 %	Cette étude
Test2F/test2R	V5-V6	385	55,2 / 59,8	HairpinR à 54°C	1806/2120 453/646 1364/1488	2033/2120 562/646 1484/1488	1246/2033 443/562 804/1484	61,28 % 78,82 % 54,17 %	Cette étude
Test3F/test3R	V5-V6	385	55,3 / 59,8	HairpinR à 54°C	1806/2120 453/646 1364/1488	2032/2120 562/646 1484/1488	1245/2032 443/562 803/1483	61,26 % 78,82 % 54,14 %	Cette étude
Test4F/test4R	V5-V6	385	55,8 / 59,8	HairpinR à 54°C	1798/2120 445/646 1364/1488	2027/2120 557/646 1484/1488	1241/2027 439/557 803/1483	61,22 % 78,81 % 54,14 %	Cette étude
Test5F/test5R	V5-V6	384	53,8 / 59,8	HairpinR à 54°C	1807/2120 454/646 1364/1488	2030/2120 562/646 1481/1488	1243/2030 443/562 801/1481	61,23 % 78,82 % 54,08 %	Cette étude

Ce que met en évidence ce tableau, c'est que certains couples de primers utilisés par le passé (et encore actuellement) présentent des caractéristiques non-optimales pour une amplification efficace. C'est notamment le cas du couple 515F/926R dont les températures d'hybridation présentent une différence largement supérieure à celle préconisée pour garantir de bons résultats en PCR. Le tableau présente également d'autres informations importantes comme le nombre d'amplicons produits *in silico* et la résolution spécifique de ces derniers. Ces informations vont être cruciales pour le choix des couples de primers à tester. En combinant ces dernières avec celles de la figure 4, nous pouvons voir le spectre de la diversité des amplicons au travers des différents phyla pour chacun des marqueurs évoqués dans le tableau précédent. Cela permet de vérifier si le choix d'un couple de primers va être plus ou moins favorable à l'amplification d'une lignée ou d'une autre par exemple et si la diversité amplifiée au sein de cette lignée est important ou ne concerne que quelques représentants

de celle-ci.

Figure 4 : couverture des différents phyla du phytoplancton par les couples d'amorces du Tableau 7 pour le marqueur 16S. L'efficacité d'amplification est exprimée (en %) pour chaque phylum et les couleurs explicitent la diversité taxonomique qui est amplifiée (en %) au sein de ce phylum.

VII.4.2 ARNr 18S

Gène codant pour la petite sous-unité ribosomale chez les eucaryotes, utilisé de manière conventionnelle dans les études phylogénétiques (Chapman et *al.*, 1998). Parmi les différentes régions variables (*cf.* figure 5) qu'il présente, V2, V4 et V9 du 18S sont les plus adaptées pour recenser la biodiversité des eucaryotes (Hadziavdic et *al.*, 2014) et communément utilisées pour les organismes phototrophes eucaryotes (Bradley et *al*., 2016). La région V4 est la plus classiquement utilisée en metabarcoding car c'est celle qui est la plus variable sur le 18S (ex. Alverson et *al.*, 2006) avec de nombreux primers documentés et qu'elle a été déclarée comme « pre-barcode » pour l'étude des protistes (Pawlowski et *al.*, 2012).

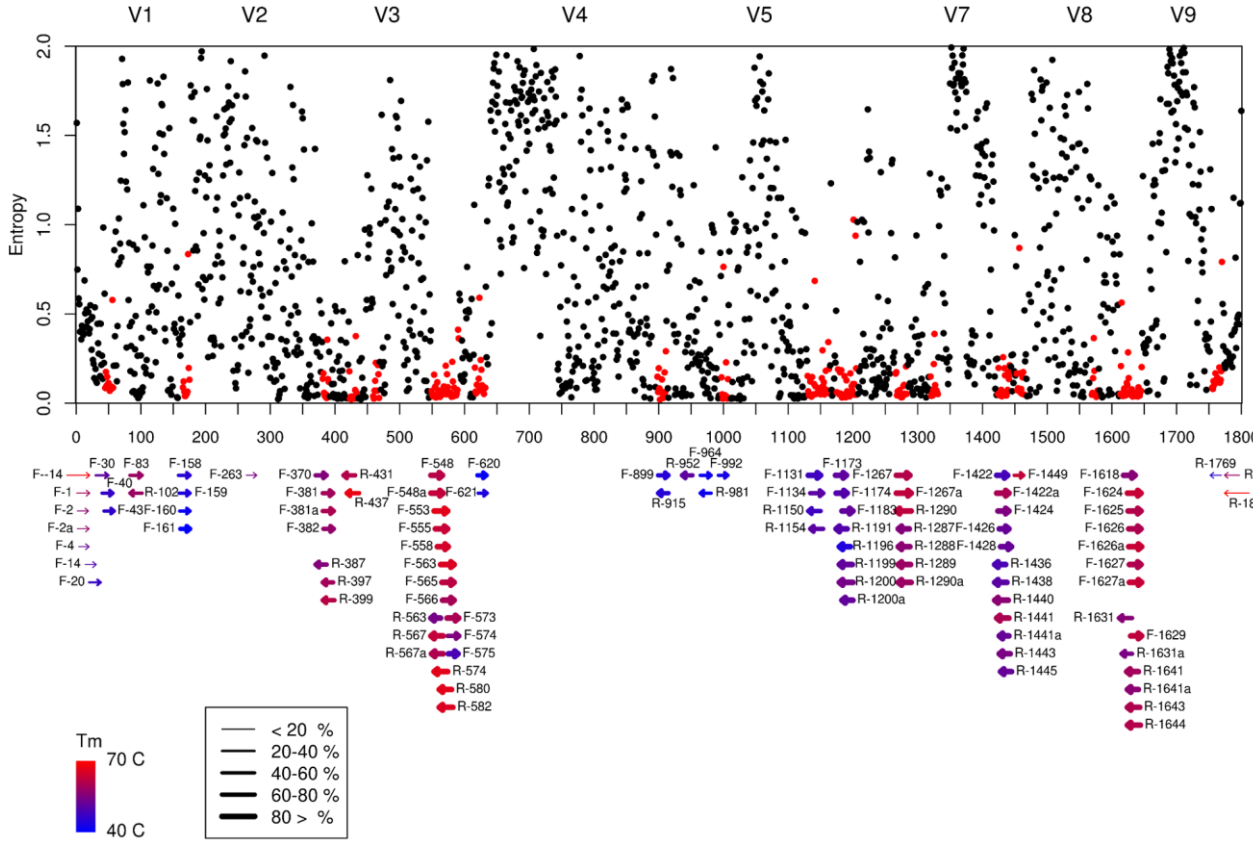

Figure 5 : représentation des zones variables et conservées du 18S et localisation de différents primers. La variabilité est représentée par les valeurs d'entropie de Shannon calculées pour chaque position de séquences eucaryotes alignées selon celle d'un organisme modèle : S. cerevisiae (Figure de Hadziavdic et *al.*, 2014)

VII.4.2.1 Résumé des avantages et inconvénients du marqueur

Avantages :

+ Assure la plus grande couverture du phytoplancton eucaryote pour les listes DROM et PHYTOBS (*cf.* Gap-analyses) ;
+ Bases de références bien documentées ;
+ Certains couples d'amorces adaptés au phytoplancton déjà renseignés dans la littérature ;
+ Présence de régions variables et conservées (*cf.* Figure 5).

Inconvénients :

- Ne cible pas les procaryotes, et donc nécessité d'être combiné à un autre marqueur pour couvrir également les cyanobactéries ;

- Ne cible pas seulement les micro-algues mais aussi une large diversité e microbes hétérotrophes (protistes) qui peut nécessiter une profondeur de séquençage accrue pour bien caractériser la diversité des micro-algues.
- la région V4 du 18S varie en longueur entre les taxons : cause des problèmes d'alignements et ainsi des difficultés de mise en place d'un seuil de divergence entre les séquences (Zhang et *al*., 2018 + réf. associées, et ici avec les Euglenozoa – *cf.* suite -) ;
- Marqueur assez conservé (même dans les régions variables) qui ne discrimine pas certaines espèces (qui peuvent l'être sur des critères morphologiques) (Krietniz & Bock 2012 + réf. associées).

VII.4.2.2 Jeu de données utilisé pour « designer » les couples d'amorces

Les séquences présentes sur les bases de données Silva v.138 (Quast et *al.*, 2013) et pr2 (Guillou et *al*., 2013) ont été combinées dans un jeu de données commun totalisant ainsi 3040 séquences. Comme indiqué précédemment (*cf.* 2b § VII.2), pour chaque espèce une seule séquence a été conservée parmi toutes celles redondantes. Le nombre de séquences sélectionnées pour chaque phylum pour le design des primers est donné dans la figure 6. La diversité taxonomique des séquences représentants chaque phylum a également été mise en évidence en calculant un indice de la diversité taxonomique (Warwick & Clarke, 1995). Cette figure nous permet également de constater que les phyla qui sont le plus représentés en moléculaire (*i.e.* qui présentent le plus grand nombre de séquences documentées) sont ceux des Chlorophyta et Bacillariophyta. D'autres, comme les Haptophyta ou les Bigyra restent encore très peu documentés pour ce marqueur.

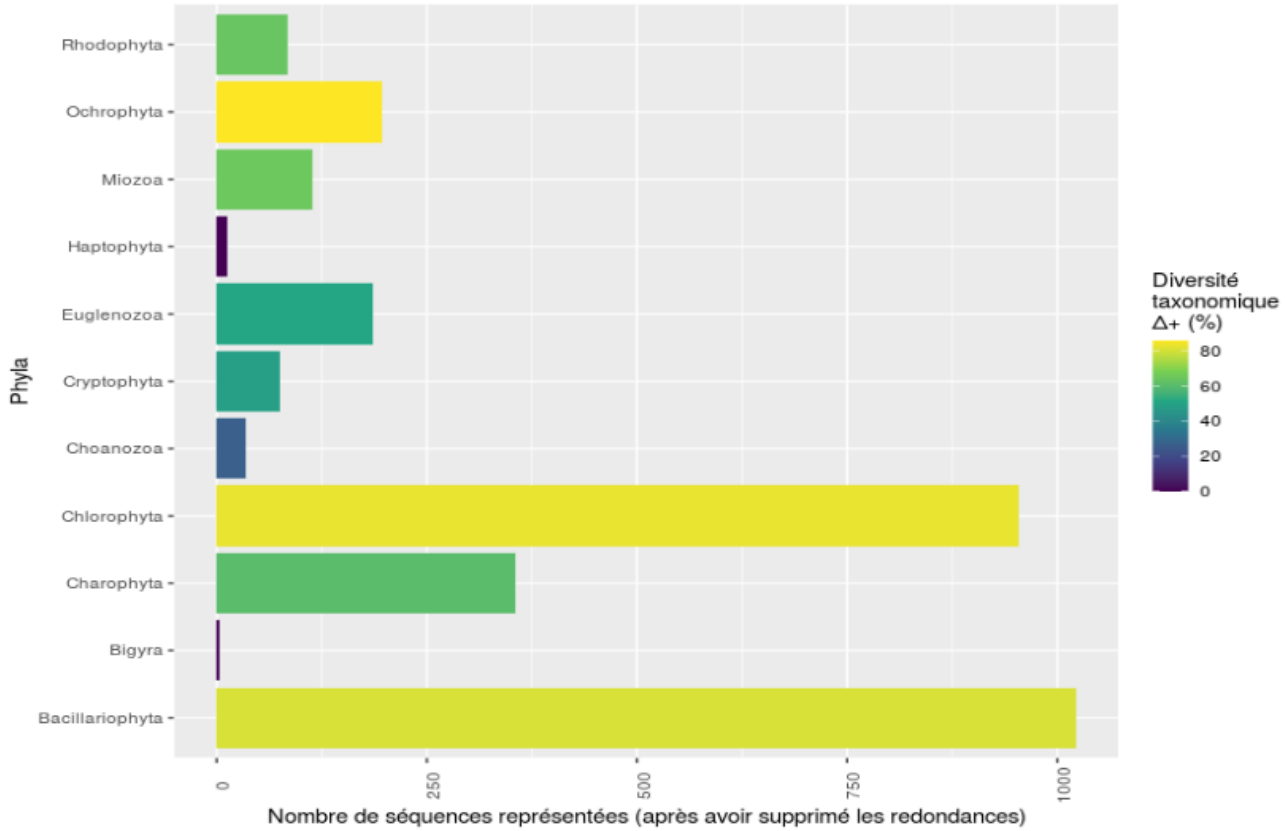

Figure 6 : Pour chaque phylum, le nombre de séquences 18S sélectionnées pour le design de primers est donnée. Ces séquences d'ARNr 18S sont issues d'une fusion entre les bases de données Silva (138), pr2. Si une espèce présentait plusieurs séquences, ou si des séquences présentaient des codes d'accession identiques, alors dans les deux cas, une seule séquence a été conservée. La couleur des histogrammes représente la diversité taxonomique (Warwick & Clarke, 1995) au sein du phylum (0 : toutes les espèces appartiennent au même groupe taxonomique, 100 : chaque espèce appartient à un groupe taxonomique différent).

Etant donné que la région V4 est la plus résolutive, et l'une des plus adaptée au séquençage MiSeq Illumina, les analyses de primers se concentreront sur celles-ci. Même si d'autres régions ont été investiguées, c'est sur la région V4 que les meilleurs résultats ont été obtenus et que des couples de primers ont pu être trouvés. Cependant, il faut noter que la région V4 exclut totalement le phylum des Euglenozoa : en effet ces derniers présentent un intron d'environ 250 pb portant la taille de la séquence amplifiée à environ 680 pb (ce qui est trop grand pour le séquençage MiSeq que l'on souhaite réaliser).

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

VII.4.2.3 Couples de primers candidats et critères de sélection

Tableau 8: présentation des couples de primers récupérés de la littérature ou designés dans cette étude et leurs caractéristiques. Différentes informations sont fournies (pour plus de détail voir § VII.3.). Pour rappel, les températures d'hybridation et les potentiels problèmes *in vitro* sont donnés par OligoAnalyzer (idt-dna).

Couple	Région variable	Taille moyenne amplicons	Tm* en °C (F/R)	Potentiels problèmes *in vitro**	Nombre séquences matchant exactement avec le couple	Nombre de séquences amplifiées et de longueur acceptable en autorisant 2 mismatches maximum	Nombre de séquences uniques récupérées parmi les amplicons (2mm)	Résolution spécifique	Etudes
TAREuk454FWD1/TAREukR	V4	417	60,1 / 45,9	SelfDimer F	2547/3040	2731/3040	2170/2731	79,45 %	Stoeck et *al*., 2010
D514for18S/D978rev18S	V4	437	52,2 / 46,7		1053/3040	2193/3040	1726/2193	78,70 %	Zimmerman et *al*., 2011
D512for18S/Euk690R	V4	337	52,2 / 51,4		2337/3040	2648/3040	2061/2648	77,83 %	Zimmerman et *al*., 2011 + Elwood et *al*., 1985
Test101F/test101R	V4	433	52,1 / 49,1	SelfDimer R limite	2443/3040	2599/3040	2057/2599	79,14 %	Cette étude
Test101F/Euk690R	V4	342	52,1 / 51,4		2352/3040	2584/3040	2017/2584	78,05 %	Cette étude
Rhodo101F/Euk690RhodoTm	V4	344	55,6 / 50,1	SelfDimerF	2521/3040	2678/3040	2089/2678	78,00 %	Cette étude

Ce que met en évidence ce tableau, c'est que certains couples de primers utilisés par le passé (et encore actuellement) présentent des caractéristiques non-optimales pour une amplification efficace. C'est notamment le cas du couple TAREuk454FWD/TAREukR dont les températures d'hybridation présentent une différence largement supérieure à celle préconisée pour garantir de bons résultats en PCR ainsi qu'un risque élevée d'hybridation des amorces forward entre elles. Le tableau présente également d'autres informations importantes comme le nombre d'amplicons produits *in silico* et la résolution spécifique de ces derniers. Ces informations vont être cruciales pour le choix des couples de primers à tester. En combinant ces dernières avec celles de la figure 7, nous pouvons voir le spectre de la diversité des amplicons au travers des différents phyla pour chacun des marqueurs évoqués dans le tableau précédent. Cela permet de vérifier si le choix d'un couple de primers va être plus ou moins favorable à l'amplification d'une lignée ou d'une autre par exemple et si la diversité amplifiée au sein de

cette llignée est important ou ne concerne que quelques représentants de celle-ci. Comme le montre ces résultats, il est difficile de trouver un couple de primers avec de bons critères pour la réaction de PCR qui cible le phylum des Rhodophyta. Une autre remarque importante concerne l'absence du phylum des Euglenozoa, pas représenté ici. En effet, une insertion d'environ 250 pb a été trouvée pour ce phylum pour la région V4 du 18S. La conséquence est qu'avec les couples d'amorces utilisés, les amplicons produits pour ce phylum ont une taille qui dépasse les 600 bases et ne permet donc pas le séquençage de ces amplicons avec la technologie MiSeq (Illumina) 2x250pb.

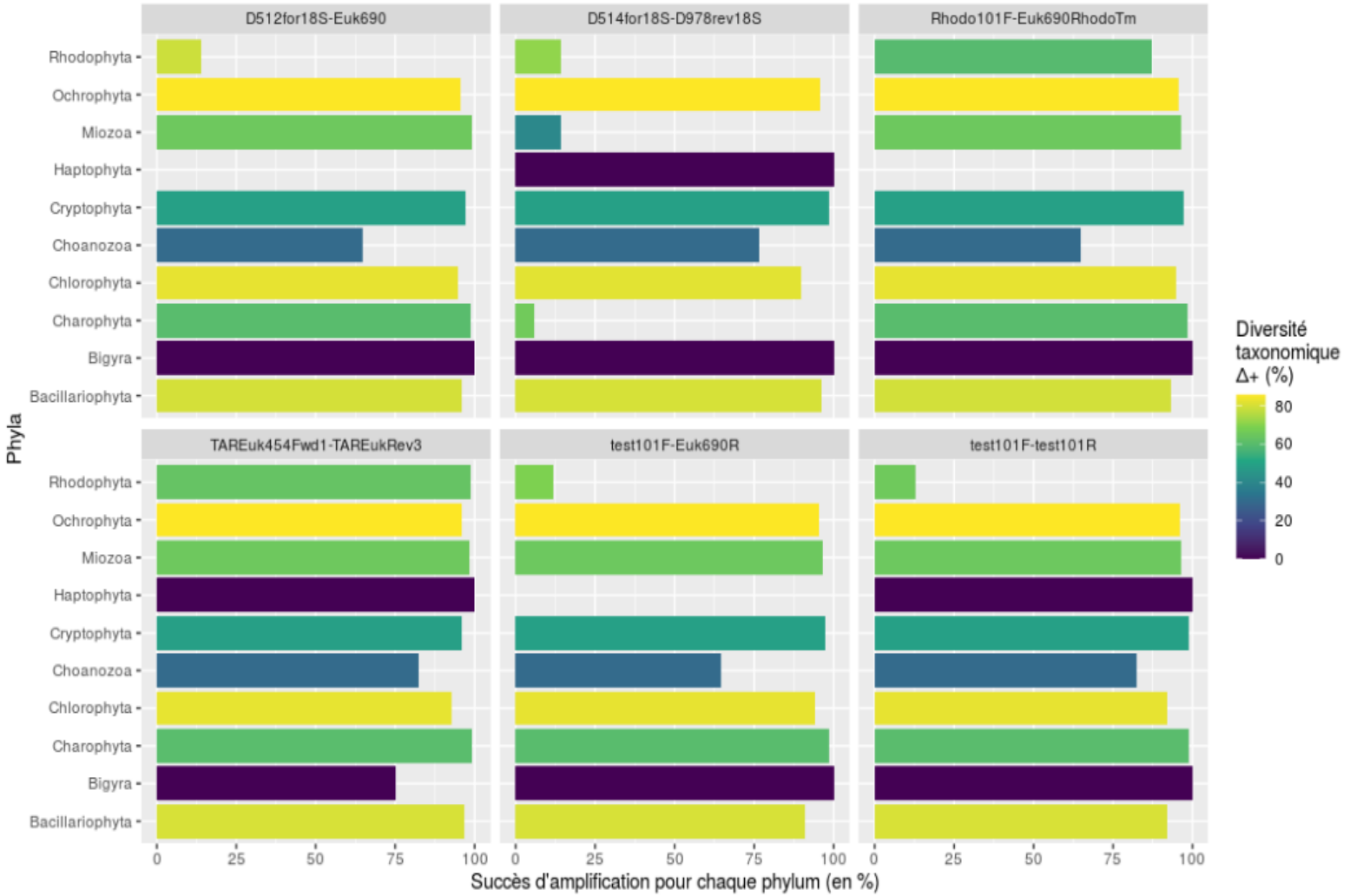

Figure 7 : couverture des différents phyla du phytoplancton par les couples d'amorces du Tableau 8 pour le marqueur 18S. L'efficacité d'amplification est exprimée (en %) pour chaque phylum et les couleurs explicitent la diversité taxonomique qui est amplifiée (en %) au sein de ce phylum.

VII.4.3 ARNr 23S

Gène codant pour la grande sous-unité ribosomale chez les procaryotes. Une région de ce gène, appartenant au domaine V, appelée 'UPA' (Universal Plastid Amplicon) est utilisée pour les organismes que l'on souhaite cibler (phytoplancton eucaryote et cyanobactéries) (Sherwood & Presting 2007).

VII.4.3.1 Résumé des avantages et inconvénients du marqueur

Avantages :

+ Cible l'ensemble du phytoplancton : cyanobactéries (procaryotes) et eucaryotes via le génome des chloroplastes ;
+ Apporte une meilleure résolution phylogénétique que les petites sous-unités ribosomales (Gutell et *al*., 1994 ; Pei et *al.*, 2009) ;
+ Couvre mieux la diversité des communautés des cyanobactéries que le 16S ou *tuf*A (Marcelino & Verbruggen, 2016) ;
+ Universalité du marqueur : avec une simple paire de primers, il est possible d'amplifier l'ensemble des clades associés au phytoplancton (Sherwood & Presting, 2007 ; Saunders & Kucera, 2010) ;
+ Marqueur d'intérêt grandissant, notamment 'en interne' (INRAE) (INRAE Bordeaux Djemiel et *al*., 2020 ; INRAE Dijon : Marie-Agnès Coutellec et *al*., *in press* (séquences 23S culture TCC) et bases de références en développement.

Inconvénients :

- Bases de données encore 'pauvres' en séquences ;
- Des résultats variables pour quelques groupes taxonomiques pour lesquels la résolution n'est pas suffisante pour distinguer certaines espèces (Saunders & Kucera, 2010).
- Amplification de bactéries hétérotrophes (nombreuses et diversifiées dans les milieux ciblés) si les amorces ne sont pas assez spécifiques : cela peut conduire à une saturation de la profondeur de séquençage et potentiellement à une incapacité à caractériser efficacement la diversité du phytoplancton ;

VII.4.3.2 Jeu de données utilisé pour « designer » les couples d'amorces

Les séquences présentes sur les bases de données Silva v.138 (Quast et *al.*, 2013) et µgreen-db (Djemiel et *al.,* 2020) ont été combinées dans un jeu de données commun totalisant ainsi 866 séquences. Comme indiqué précédemment (*cf.* 2b § VII.2), pour chaque espèce une seule séquence a été conservée parmi toutes celles redondantes. Le nombre de séquences sélectionnées pour chaque phylum pour le design des primers est donné dans la figure 8. La diversité taxonomique des séquences représentants chaque phylum a également été mise en évidence en calculant un indice de la diversité taxonomique (Warwick & Clarke, 1995). Cette figure nous permet également de constater que les séquences des cyanobactéries sont les plus représentées pour ce marqueur. D'autres phyla, comme les Haptophyta ou les Miozoa restent encore très peu représentés par ce marqueur. Nous pouvons également constater que de manière globale, les séquences de ce marqueur pour le phytoplancton sont moins nombreuses que pour les autres séquences ribosomales évoquées précédemment (*i.e.* 16S et 18S).

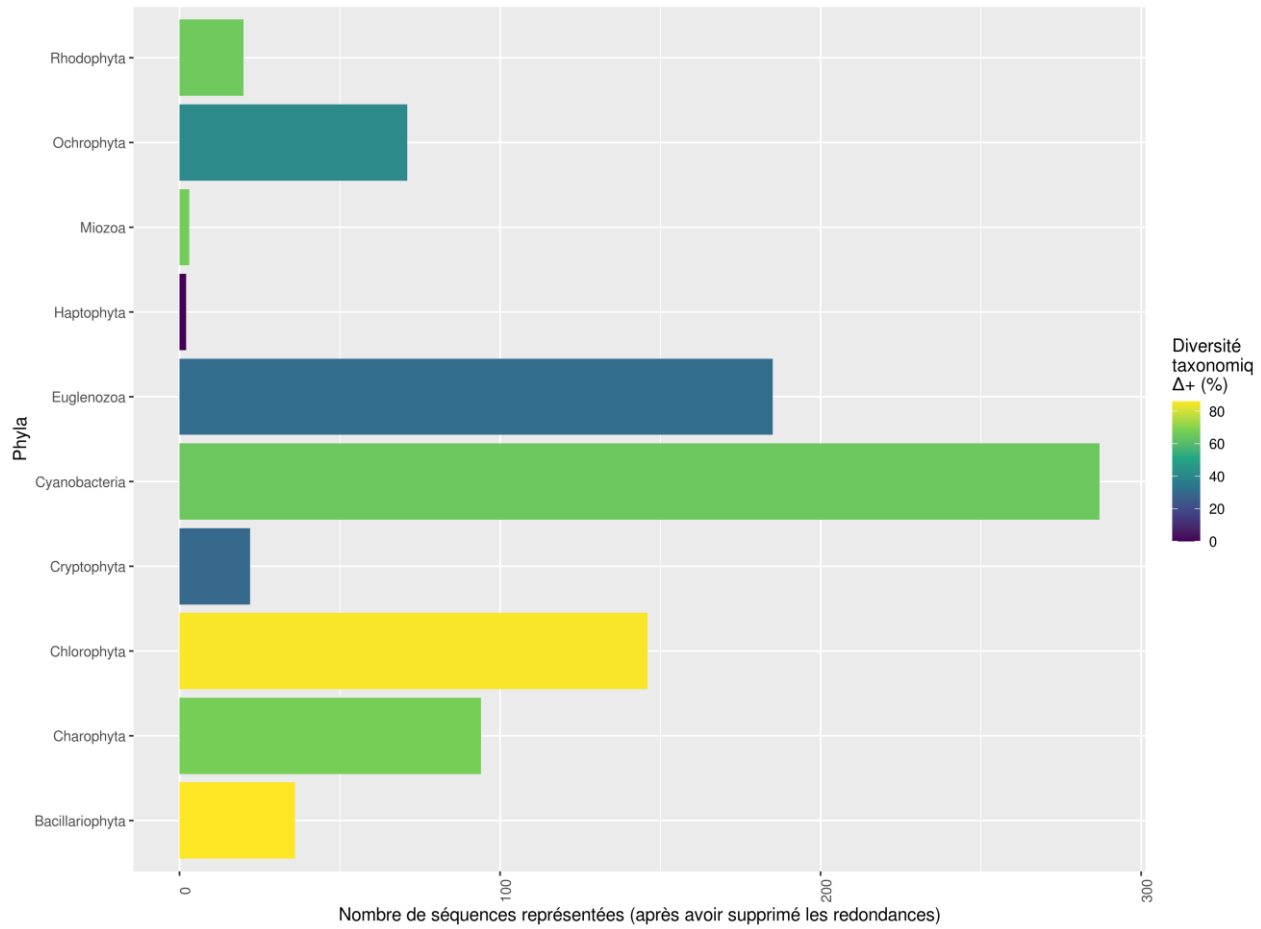

Figure 8 : pour chaque phylum, le nombre de séquences 23S sélectionnées pour le design de primers est donnée. Ces séquences d'ARNr 23S sont issues d'une fusion entre les bases de données Silva (138), µgreen-db. Si une espèce présentait plusieurs séquences, ou si des séquences présentaient des codes d'accession identiques, alors dans les deux cas, une seule séquence a été conservée. La couleur des histogrammes représente la diversité taxonomique (Warwick & Clarke, 1995) au sein du phylum (0 : toutes les espèces appartiennent au même groupe taxonomique, 100 : chaque espèce appartient à un groupe taxonomique différent).

VII.4.3.3 Couples de primers candidats et critères de sélection

Tableau 9: présentation des couples de primers récupérés de la littérature ou designés dans cette étude et leur caractéristiques. Différentes informations sont fournies (pour plus de détail voir § VII.3.). Pour rappel, les températures d'hybridation et les potentiels problèmes *in vitro* sont donnés par OligoAnalyzer (idt-dna).

Couple	Région variable	Taille moyenne amplicons	Tm* en °C (F/R)	Potentiels problèmes *in vitro**	Nombre séquences matchant exactement avec le couple	Nombre de séquences amplifiées et de longueur acceptable en autorisant 2 mismatches maximum	Nombre de séquences uniques récupérées parmi les amplicons (2mm)	Résolution spécifique	Etudes
p23SrV_f1/p23SrV_r1	UPA	408	51,3 / 53		651/866	732/866	628/732	85,79 %	Sherwood & presting (2007)
A23SrV_F1/A23SrV_R1	UPA	411	48,3 / 47,8		652/866	740/866	639/740	86,35%	Yoon et *al.*, 2016
A23SrV_F2/A23SrV_R2	UPA	405	52.3 / 50,3		654/866	729/866	625/729	85,73 %	Yoon et *al.*, 2016
test253F/test253R	UPA	408	53,9 / 54,8		709/866	745/866	643/745	86,30 %	Cette étude
test587F/test587R	UPA	408	53,9 / 53,2		709/866	745/866	643/745	86,30 %	Cette étude
test108F/test108R	UPA	402	53,5 / 50,9		731/866	751/866	646/751	86,01 %	Cette étude

Ce que met principalement en évidence ce tableau, c'est l'efficacité d'amplification de l'ensemble des couples de primers présentés ici ainsi que la bonne résolution fournie par ces amplicons. L'universalité de ces primers est un avantage majeur qui garantit une bonne amplification à la fois chez les organismes procaryotes et eucaryotes. De plus, comme le montre la figure 9, les différents phyla du phytoplancton sont plutôt bien représentés. Le seul biais que nous pourrions évoquer ici est le fait que le nombre de séquences soit faible et donc pas assez représentatif de la diversité naturelle. Il serait intéressant de savoir si ces mêmes couples de primers seraient capables d'amplifier davantage de séquences 23S, qui ne sont pas encore répertoriées dans des bibliothèques de références pour s'assurer de l'universalité des primers à plus grande échelle.

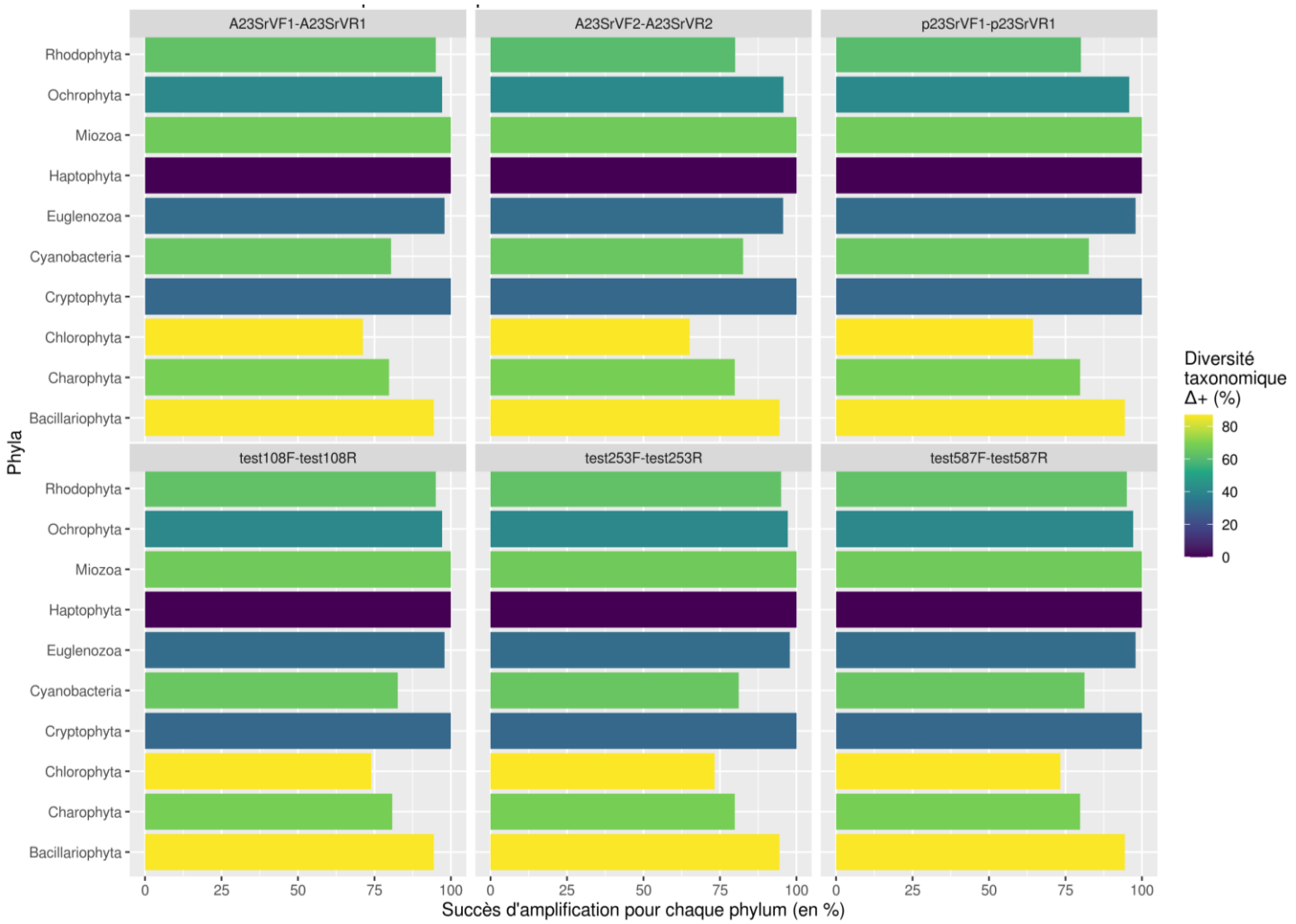

Figure 9 : couverture des différents phyla représentant le phytoplancton par les potentiels couples d'amorces proposés. L'efficacité d'amplification est exprimée (en %) pour chaque phylum et les couleurs explicitent la diversité taxonomique qui est amplifiée (en %) au sein de ce phylum.

*VII.4.4 rbc*L

Gène codant pour l'enzyme ribulose-bisphosphate-carboxylase/oxygenase (RuBPC/O). Cette dernière assure la fixation du CO_2 par les producteurs primaires et est donc essentielle au fonctionnement de tout organisme photosynthétique.

VII.4.4.1 Résumé des avantages et inconvénients du marqueur

Avantages :

+ Marqueur plastidial qui couvre à la fois les eucaryotes et procaryotes photosynthétiques ;
+ Des études montrent qu'il offre une bonne résolution intra-lignées, il est même un des marqueurs recommandé pour le barcoding des plantes (CBOL Plant Working Groupe, 2009) ;
+ Efficacité démontrée pour les diatomées avec Diat.barcode (Rimet et *al*., 2019) ;
+ Nombreuses séquences disponibles sur GenBank.

Inconvénients :

- Pas de bases de données existantes pour l'ensemble des classes algales, il faut donc les récupérer sur GenBank ;
- Les primers sont spécifiques à certaines lignées, pas de primers universels efficaces pour l'ensemble du phytoplancton, difficulté d'en développer pour du metabarcoding.

VII.4.4.2 Jeu de données utilisé pour « designer » les couples d'amorces

Les séquences *rbc*L présentes sur GenBank ont été récupérées grâce à un script écrit sur Python. Une fois ces nombreuses séquences traitées et associées à celle des espèces présentes sur PHYTOBS, le jeu de données a été traité comme indiqué précédemment (*cf.* 2b § VII.2). Pour chaque espèce, une seule séquence a donc été conservée parmi toutes celles redondantes. Le jeu de données totalisait alors 4605 séquences. Le nombre de séquences sélectionnées pour chaque phylum pour le design des primers est donné dans la figure 10. La diversité taxonomique des séquences représentant chaque phylum a également été mise en évidence en calculant un indice de la diversité taxonomique (Warwick & Clarke, 1995). Cette figure nous permet également de constater que les séquences des cyanobactéries sont les plus représentées pour ce marqueur. Les phyla des Bacillariophyta et des Chlorophyta sont également bien représentés, en revanche, d'autres comme les Cryptophyta, Haptophyta ou encore les Miozoa restent encore très peu représentés par ce marqueur.

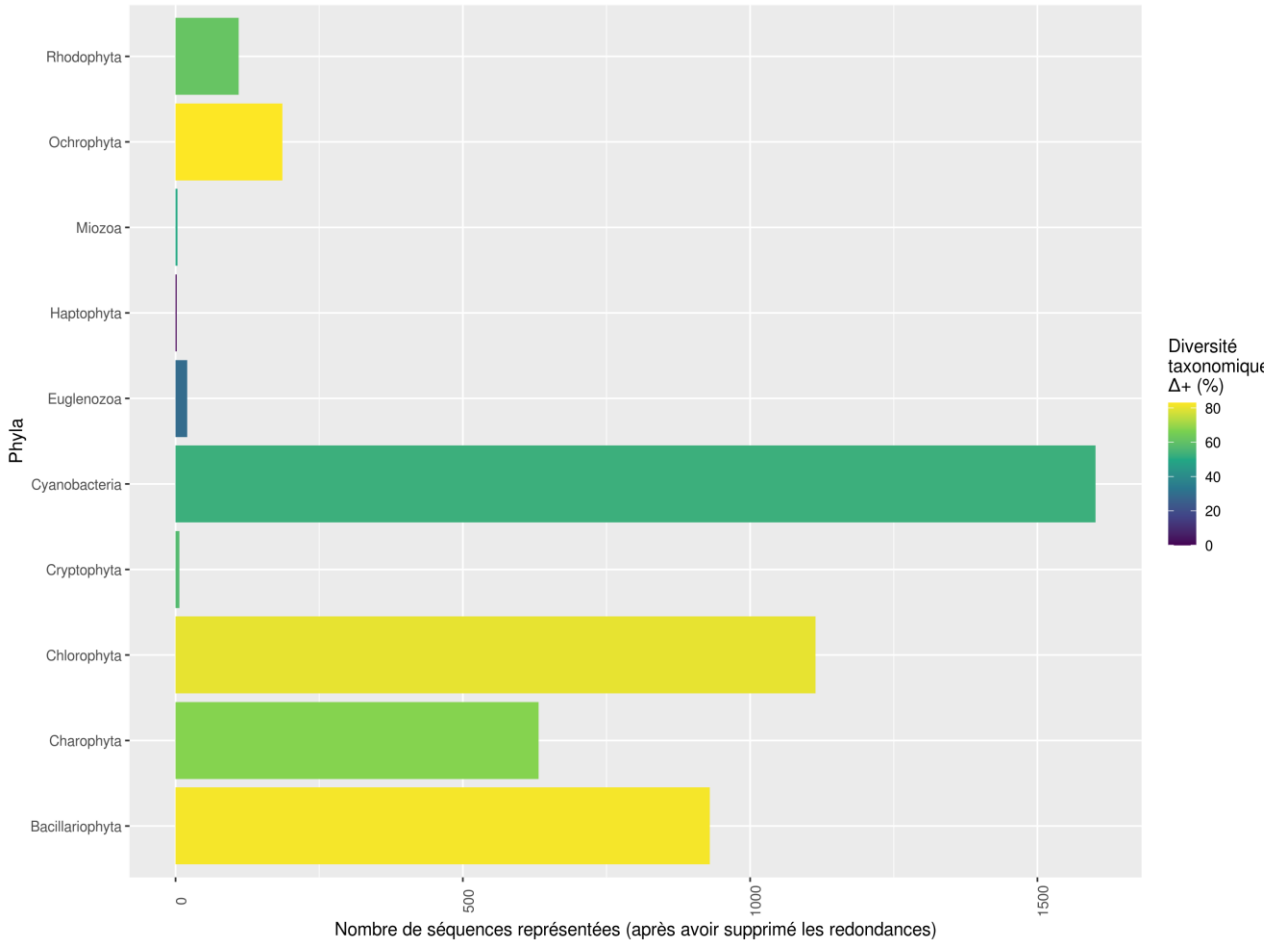

Figure 10 : pour chaque phylum, le nombre de séquences *rbc*L sélectionnées pour le design de primers est donnée. Ces séquences *rbc*L ont été initialement récupérées de GenBank grâce à un script Python. Si une espèce présentait plusieurs séquences, ou si des séquences présentaient des codes d'accession identiques, alors dans les deux cas, une seule séquence a été conservée. La couleur des histogrammes représente la diversité taxonomique (Warwick & Clarke, 1995) au sein du phylum (0 : toutes les espèces appartiennent au même groupe taxonomique, 100 : chaque espèce appartient à un groupe taxonomique différent).

VII.4.4.3 Couples de primers candidats et critères de sélection

Tableau 10: présentation des couples de primers récupérés de la littérature ou designés dans cette étude et leurs caractéristiques. Différentes informations sont fournies (pour plus de détail voir § VII.3.). Pour rappel, les températures d'hybridation et les potentiels problèmes *in vitro* sont donnés par OligoAnalyzer (idt-dna) ; étant donné qu'ici les primers ne répondaient pas au critère de taille attendue pour les amplicons ces paramètres n'ont pas été décrits.

Couple	Région variable	Taille moyenne amplicons	Tm* en °C (F/R)	Potentiels problèmes *in vitro**	Nombre séquences matchant exactement avec le couple	Nombre de séquences amplifiées et de longueur acceptable en autorisant 2 mismatches maximum	Nombre de séquences uniques récupérées parmi les amplicons (2mm)	Résolution spécifique	Etudes
P169/P328	Forme 1D	554	-	-	0/4605	699/4605	652/699	93.27 %	Paul et *al.*, 1990
21-mer P3 / 18-mer P6	Forme 1A/1B	623	-	-	273/4605	1189/4605	962/1189	80.90 %	

Aucun autre couple de primers adapté au barcoding n'a été trouvé dans la littérature. Cela s'explique par la difficulté de trouver des régions conservées au sein du gène *rbc*L qui a déjà été décrit pour accueillir des introns, notamment chez des macroalgues marines (Hanyuda et *al.*, 2000), et cela a l'air d'être également confirmé pour les microalgues d'eau douce. Même au sein d'un seul groupe taxonomique comme les diatomées, il est nécessaire de designer différents primers pour couvrir la diversité du groupe (Vasselon et *al.*, 2017). Malgré les nombreux avantages qu'offrent l'utilisation du marqueur *rbc*L pour le barcoding du phytoplancton, celui-ci ne peut être retenu à cause de la non-universalité des primers. L'investigation des primers pour ce marqueur s'arrête donc là.

VII.4.5 tufA

Utilisé pour la première fois en phylogénie par Iwabe et *al*., 1989 puis par Delwiche et *al.,* 1995 pour démontrer l'origine cyanobactérienne des chloroplastes présents chez les organismes eucaryotes. Ce gène est relativement bien conservé au fil de l'évolution : il code pour la synthèse d'une protéine essentielle, le facteur d'élongation Tu (EF-Tu).

VII.4.5.1 Résumé des avantages et inconvénients du marqueur

Avantages :

+ Couvre à la fois les eucaryotes et les procaryotes ;
+ Offre une bonne résolution intra-lignées pour certaines espèces phytoplanctoniques d'eau douce difficile à différencier morphologiquement (Vieira et *al.,* 2016 ; Zou et *al*., 2016) ;
+ Présente moins d'introns que d'autres marqueurs (*e.g. rbc*L) et le rend donc plus adapté au metabarcoding ;
+ Ce marqueur présenterait un taux d'évolution moléculaire intermédiaire (Saez et *al*., 2008) permettant ainsi une meilleure résolution intra-lignées (Sauvage et *al.*, 2016).

Inconvénients :

- Peu de séquences en ligne (pas de bases de données existante, il faut les récupérer sur GenBank) ;
- Très utilisé pour les Ulvophyceae (beaucoup de séquences pour cette classe mais peu pour le phytoplancton en général) ;
- Primers développés dans la littérature pas forcément adaptés, difficulté d'en développer de nouveaux 'fiables' étant donné le peu de données représentant nos cibles.

VII.4.5.2 Jeu de données utilisé pour designer les couples d'amorces

Les séquences *tuf*A présentes sur GenBank ont été récupérées grâce à un script écrit sur Python. Une fois ces nombreuses séquences traitées et associées à celle des espèces présentes sur PHYTOBS, le jeu de données a été traité comme indiqué précédemment (*cf.* 2b § VII.2). Pour chaque espèce, une seule séquence a donc été conservée parmi toutes celles redondantes. Le jeu de données totalisait alors 430 séquences. Le nombre de séquences sélectionnées pour chaque phylum pour le design des primers est donné dans la figure 11. La diversité taxonomique des séquences représentants chaque phylum a également été mise en évidence en calculant un indice de la diversité taxonomique (Warwick & Clarke, 1995). Cette figure nous permet également de constater que les séquences du phylum des Chlorophyta sont les plus représentées pour ce marqueur. Relativement peu de séquences appartenant à des taxons phytoplanctoniques sont retrouvés pour le marqueur *tuf*A, en effet ce dernier a davantage été utilisé dans la littérature pour les plantes terrestres et les macrophytes.

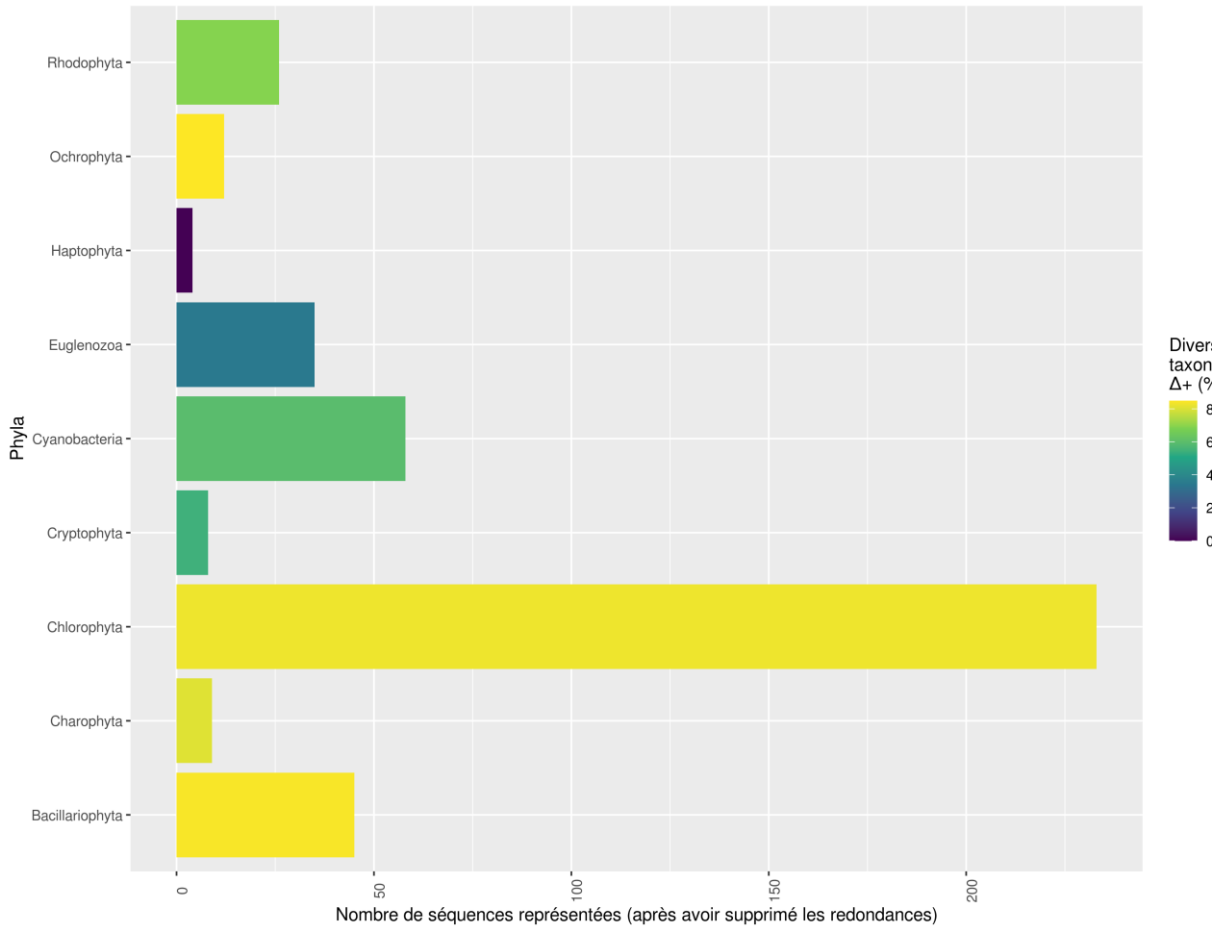

Figure 11 : pour chaque phylum, le nombre de séquences *tuf*A sélectionnées pour le design de primers est donnée. Ces séquences *tuf*A ont été initalement récupérées de GenBank grâce à un script Python. Si une espèce présentait plusieurs séquences, ou si des séquences présentaient des codes d'accession identiques, alors dans les deux cas, une seule séquence a été conservée. La couleur des histogrammes représente la diversité taxonomique (Warwick & Clarke, 1995) au sein du phylum (0 : toutes les espèces appartiennent au même groupe taxonomique, 100 : chaque espèce appartient à un groupe taxonomique différent).

VII.4.5.3 Couples de primers candidats et critères de sélection

Tableau 11 : présentation des couples de primers récupérés de la littérature ou designés dans cette étude et leurs caractéristiques. Différentes informations sont fournies (pour plus de détail voir § 7.3. 4). Pour rappel, les températures d'hybridation et les potentiels problèmes *in vitro* sont donnés par OligoAnalyzer (idt-dna) ; lorsque les primers ne répondaient pas au critère de taille attendue pour les amplicons, ces paramètres n'ont pas été décrits.

Couple	Région variable	Taille moyenne amplicons	Tm* en °C (F/R)	Potentiels problèmes *in vitro**	Nombre séquences matchant exactement avec le couple	Nombre de séquences amplifiées et de longueur acceptable en autorisant 2 mismatches maximum	Nombre de séquences uniques récupérées parmi les amplicons (2mm)	Résolution spécifique	Etudes
TufAF/TufAR	-	987			1/430	60/430	60/60	100 %	Fama et *al.* (2002)
TufA_SF/TufA_SR	-	622			35/430	147/430	140/147	95,23 %	Zou et *al.* (2016)
Tubryo_F/Tubryo_R	-	596			4/430	117/430	113/117	95,76 %	Sauvage et *al*. (2016)
envtufA_F/envtufA_R	-	463	55.3 / 54.8	SelfDimer F & R HeteroDimer	56/430	245/430	220/245	89,79 %	Sauvage et *al*. (2016)
Test129F/test129R	-	339	49.8 / 49.8		110/430	308/430	285/308	92,53 %	Cette étude
Test130F/test130R	-	340	49.8 / 51.3		110/430	308/430	285/308	92,53 %	Cette étude
Test126F/test126R	-	340	49.8 /49.8		112/430	306/430	283/306	92,48 %	Cette étude

Ce que met principalement en évidence ce tableau, c'est le manque d'efficacité d'amplification de l'ensemble des couples de primers récupérés ici de la littérature. Lorsque ceux-ci présentent une efficacité un peu meilleure (*e.g.* envtufA_F/envtufA_R) alors les critères d'amplifications peuvent poser problème, ou alors ces primers ne sont pas très fiables, notamment le couple évoqué en question, qui présente énormément d'ambiguités (*cf.* Annexes, tableau des séquences des primers). Les couples de primers trouvés dans cette étude semblent être plus efficaces pour le phytoplancton, même si l'amplification reste un peu limitée.

En revanche, nous pouvons voir la haute résolution spécifique qu'apporte ces amplicons (autour de 92 % pour les couples de primers designés ici). Notons également que le nombre de séquences pour ce marqueur reste assez limité. La figure 12 quant à elle nous confirme la faible couverture des couples de primers de la littérature pour le phytoplancton, et nous montre que la diversité des phyla couverts par les couples d'amorces designés ici est assez hétérogène.

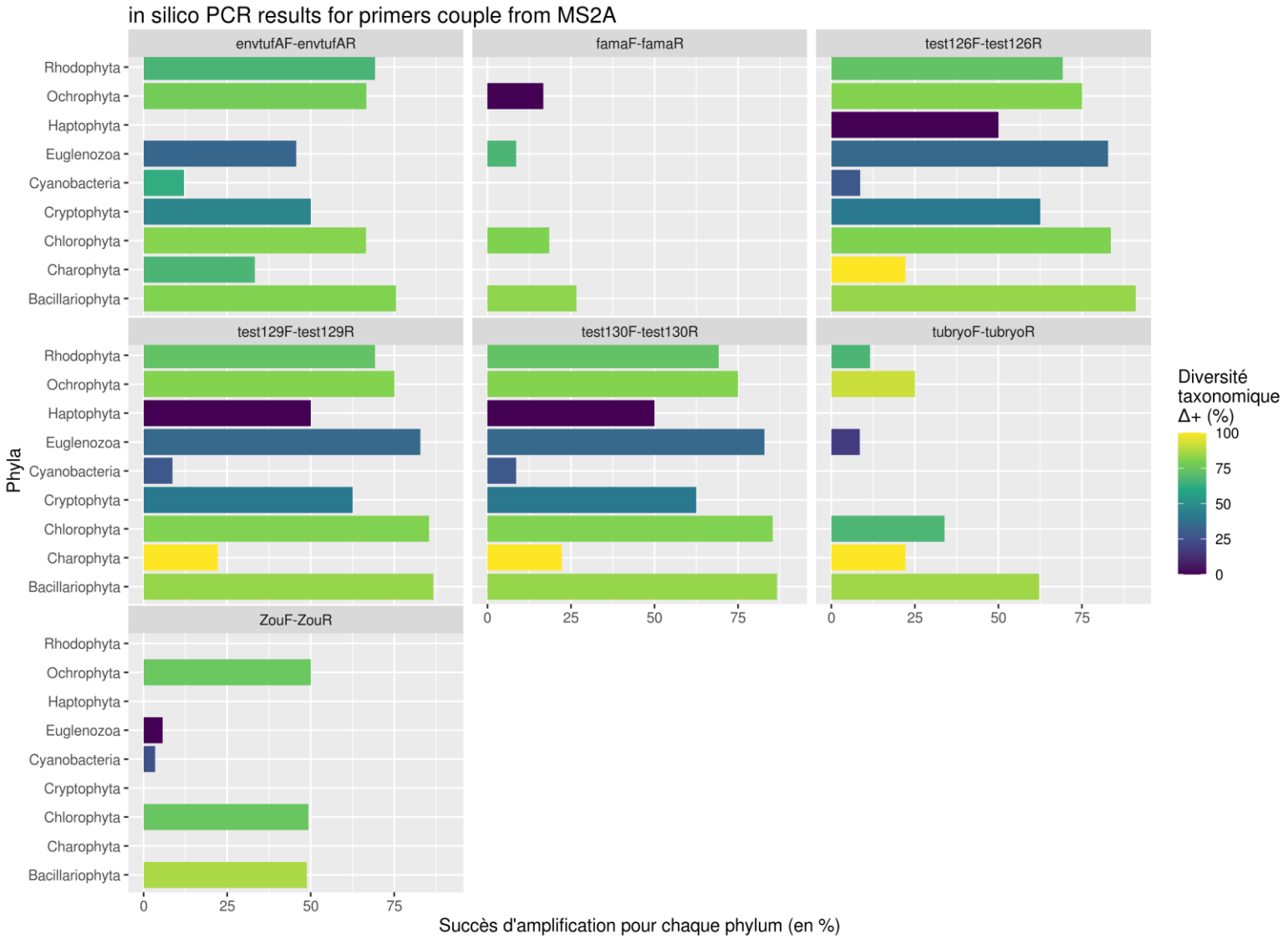

Figure 12 : couverture des différents phyla représentant le phytoplancton par les potentiels couples d'amorces proposés. L'efficacité d'amplification est exprimée (en %) pour chaque phylum et les couleurs explicitent la diversité taxonomique qui est amplifiée (en %) au sein de ce phylum.

VIII. Protocole d'échantillonnage

Dans le cadre de ce projet, il est nécessaire de collecter de nouveaux échantillons de phytoplancton des plans d'eau des DROM sur lesquels l'ADN pourra être extrait. Pour faciliter cette collecte, un protocole d'échantillonnage pour l'ADNe du phytoplancton a été défini et sera fourni aux différents laboratoires préleveurs dans les DROM dès l'année prochaine. Ces derniers pourront alors nous faire des retours sur l'utilisation de celui-ci afin de l'optimiser si nécessaire. Ce protocole se base sur celui élaboré et utilisé dans le cadre du projet EcoAlpsWater (https://www.protocols.io/view/lake-plankton-sample-collection-from-the-field-for-xn6fmhe Domaizon et *al.*, 2019). L'ensemble du matériel nécessaire pour réaliser les campagnes de prélèvements dans les plans d'eau des DROM sera envoyé aux différents opérateurs de terrain (bureaux d'études / DREAL / Office de l'eau...), accompagné du protocole défini. Une version en format vidéo de démonstration de la mise en œuvre sera également fournie. Le protocole en question est disponible en matériel supplémentaire (**Mat.Supp.6**).

L'utilisation de Stérivex sera préconisée pour réaliser les filtrations de l'eau directement sur le terrain. Celle-ci sera réalisée grâce à des seringues et la conservation des échantillons sera possible directement dans les Stérivex grâce à l'utilisation d'un tampon de préservation. Cela permettra de conserver les échantillons jusqu'à la fin de l'année, lorsque les opérateurs de terrain devront alors nous les renvoyer. Nous prendrons contact dès le début de l'an prochain avec ces derniers afin de leur transmettre toutes ces informations plus en détails.

IX. Discussion et perspectives

IX.1 Liste taxonomique

La liste taxonomique réalisée sur la base des observations microscopiques du phytoplancton dans les plans d'eau des DROM met en avant un résultat positif : une proportion considérable de ces taxons sont présents dans la liste des espèces bio-indicatrices contenues dans la méthode IPLAC. Cependant, ce sont les investigations futures qui nous permettront de conclure sur l'efficacité de cet indice pour les plans d'eau des DROM ou sur le besoin d'adapter ce dernier pour ces milieux ou d'en définir un nouveau. De plus, il est nécessaire de rappeler que les observations microscopiques réalisées dans les DROM sont susceptibles de souffrir de certains biais observateurs (références différentes pour les identifications microscopiques, personnes différentes impliquées dans les comptages, expertises différentes *etc.*). Bien qu'elles fournissent un aperçu général de la diversité du phytoplancton dans les DROM, se focaliser uniquement sur celles-ci dans le cadre de notre projet serait une erreur. C'est pourquoi la base PHYTOBS a été utilisée, afin de couvrir l'ensemble du phytoplancton d'eau douce (présent à ce jour sur la base). L'avantage est que la méthode développée dans le cadre de ce projet sera plus robuste et pourra également être appliquée à d'autres milieux lacustres. Les gap-analyses indiquent un manque de représentativité du phytoplancton d'eau douce (PHYTOBS) dans les bibliothèques ADN. Ce projet, ainsi que tous les autres travaillant sur le sujet, permettront d'augmenter le nombre de séquences disponibles pour le phytoplancton au fil du temps en confrontant les observations microscopiques aux outils moléculaires en cours de développement.

IX.2 Choix du barcode

Etant donné le contexte dans lequel s'inscrit ce projet, l'idéal serait de trouver un barcode couvrant à la fois toute la diversité du phytoplancton (eucaryote et procaryote), offrant une bonne résolution et pouvant s'obtenir avec un seul couple de primers afin de faciliter l'utilisation de la méthode. Les différents barcodes investigués dans le début de ce projet montrent que le marqueur idéal tel que décrit précédemment n'existe pas et qu'il va falloir faire des compromis pour trouver celui qui sera le plus adapté pour répondre aux objectifs du projet. Peser les avantages et inconvénients de chacun n'est pas suffisant pour permettre de décider de l'efficacité du barcode, il faut également prendre en compte l'existence de primers pertinents pour celui-ci.

Par exemple, les investigations réalisées présentent le *rbc*L comme un marqueur idéal : il est présent à la fois chez les eucaryotes et les procaryotes. De plus, il ne va cibler que des organismes photosynthétiques et pour finir beaucoup de séquences de celui-ci sont associées à des taxons de la base PHYTOBS. Cependant, il n'est pas possible de trouver pour ce marqueur des couples de primers universels pour le phytoplancton (mais également pour certaines grandes lignées de celui-ci).

Quel que soit le marqueur choisi, les primers jouent donc un rôle des plus primordial et le choix du couple de primer à utiliser va être déterminant dans la suite de l'étude. C'est pour cette raison que beaucoup de temps a été consacré au design de primers au cours de la première phase du projet. Au

vu des premiers résultats des PCR *in silico* réalisées, le choix se porterait sur l'ARNr 23S comme barcode pour la suite du projet. En effet, en plus d'être résolutif, la région 'UPA' constitue un barcode pour l'ensemble du phytoplancton amplifiable par un même couple d'amorce. Enfin, il existe différents couples d'amorces (énoncés précédemment dans le tableau consacré aux primers pour le 23S) dans la littérature, mais aussi « designés » ici, qui présentent de bonnes caractéristiques pour l'amplification *in vitro*. Le seul inconvénient de l'utilisation du 23S réside dans les bases de données encore peu complètes, mais celles-ci tendent à s'enrichir notamment, avec des auteurs travaillant sur des problématiques similaires et dont le choix s'est également porté sur le 23S pour le barcoding des microalgues (Djemiel et *al*., 2020 ; Gorzerino et *al*., *en élaboration*).

Enfin une autre stratégie consisterait à utiliser en parallèle plusieurs marqueurs : en effet cette approche améliorerait nettement la résolution mais serait plus difficile à mettre en œuvre. Ce projet cherche à développer une méthode facile à utiliser, et tentera si possible d'éviter l'utilisation de plusieurs barcodes, toutefois certains auteurs la recommandent (Pawlowski et *al.*, 2012).

IX.3 Stratégie et design de primers

L'investigation réalisée autour des primers a été menée, ici, avec une nouvelle stratégie adaptée à la recherche de zones conservées au travers d'un nombre important de séquences diversifiées d'un point de vue phylogénétique. Cette stratégie s'est avérée efficace car elle a pu mettre en évidence des couples de primers dont les caractéristiques sont considérées comme optimales pour les réactions de PCR.

Les caractéristiques d'un couple de primers qui garantissent de bonnes amplifications par PCR sont :

- Taille de l'amorce entre 18 et 22 bases ;
- Températures d'hybridation comprises entre 52 et 58°C ;
- Pas plus de 5°C de différence entre les températures d'hybridation entre Forward et Reverse ;
- Pas plus de 3 'G' ou 'C' sur les 5 dernières bases de l'amorce ;
- Prêter attention aux différentes conformations spatiales de l'amorce en fonction de la température ainsi qu'à la possibilité de formation de dimères (*cf.* OligoAnalyzer).

Ces informations sont en général bien reconnues en biologie moléculaire, on les retrouve détaillées sur des sites comme par exemple http://www.premierbiosoft.com/tech_notes/PCR_Primer_Design.html .

Les nouveaux couples de primers designés lors de cette étude méritent donc d'être considérés et testés *in vitro* au même titre que ceux qui sont récurrents dans la littérature et donc souvent ré-utilisés malgré certaines limites. Une attention particulière devrait être portée sur l'utilisation des primers - dans le cas où l'on utilise des couples d'amorces designés dans la littérature par le passé : les bibliothèques ADN évoluent vite et certains primers peuvent vite devenir obsolètes.

IX.4 Les perspectives du projet pour 2021

Les travaux réalisés en 2020 ont permis de développer des outils bio-informatiques qui ont été utilisés dans ce rapport et pourront faire l'objet de partages avec les personnes intéressées. Ils ont rendu possible l'homogénéisation des bases de références selon une même nomenclature taxonomique pour le phytoplancton, à savoir, celle d'AlgaeBase (Guiry & Guiry, 2020). Cela a permis la création de bases de données moléculaires adaptées à l'étude du phytoplancton d'eau douce. Une méthode pour la recherche de primers a également été développée. Tout cela a permis de poser les bases du projet pour les années suivantes.

Pour l'année 2021, les principaux objectifs seront alors de :

1. **Prendre contact avec les services locaux de chaque DROM**

   Nous prendrons contact avec les services responsables (office de l'eau, direction de l'environnement...) de chaque DROM afin de demander les autorisations de former à distance les bureaux d'études chargés des prélèvements annuels afin qu'ils réalisent les prélèvements ADN. Le but est de collecter des échantillons de phytoplancton des plans d'eau des DROM permettant de produire des inventaires phytoplanctoniques par ADN et d'évaluer leur capacité à produire des notes indicielles de qualité écologique des plans d'eau dans les DROM. Nous mutualiserons ces prélèvements spécifiques avec ceux réalisés annuellement dans le cadre de la DCE, afin de bénéficier des inventaires en microscopie et de toutes les données environnementales associées. Cela inclut :

- La formation à distance les bureaux d'études aux prélèvements ADNe :

  Cette formation se fera par le biais de vidéos, protocoles écrits ainsi que des discussions en visioconférence. Le matériel nécessaire pour ces prélèvements sera préparé et envoyé à chacun des bureaux d'études, les coûts seront pris en charge par INRAE dans le cadre de ce projet.
- Le prélèvement des échantillons suivant les campagnes DCE :

  Les échantillons seront prélevés lors des campagnes DCE et serons transmis au laboratoire INRAE pour être traités en 2022.

2. **Sélectionner le/les marqueur(s) candidat(s) pour la méthode**

   Un ou plusieurs marqueurs seront sélectionnés pour être mis en application ; la décision s'appuiera sur les travaux réalisés en 2020 en considérant les avantages et inconvénients de chaque marqueur. Le design des primers associés à ces marqueurs sera consolidé *in-silico*. Ensuite, ces primers seront testés *in-vitro* par PCR et/ou qPCR. La collection de cultures d'algues (TCC) pourra être mobilisée.

3. **Tester des sets de primers et séquençage haut-débit**

   Sur la base des résultats de PCR et/ou qPCR, les meilleurs primers seront sélectionnés pour être testés sur différents types d'échantillons :
   - Des échantillons 'synthétiques' (mélanges contrôlés de cultures pures),
   - Des échantillons déjà disponibles à l'UMR Carrtel (venant de Mayotte),
   - Des échantillons prélevés dans le cadre des suivis réalisés par les responsables du projet (ex : grands lacs alpins). Ces échantillons bénéficient d'analyses microscopiques précises et documentées de longue date, qui permettront de réaliser de bonnes comparaisons entre ADNe et microscopie.

     Après extraction de l'ADN de ces échantillons, les marqueurs d'intérêt seront amplifiés par PCR avec les différentes paires de primers retenues. Les amplicons obtenus feront ensuite l'objet d'un séquençage haut-débit. Les données obtenues en séquençage permettront d'établir des inventaires moléculaires.
4. **Comparer les résultats obtenus par les approches ADNe et microscopie**

   Les premières comparaisons entre inventaires en microscopie et en moléculaire seront réalisées : nous nous concentrerons sur les compositions taxonomiques et les diversités. Nous comparerons également les valeurs d'indices de bio-indications (*e.g.* IPLAC) obtenus par les deux méthodes. Au vu des résultats, un premier bilan sera dressé pour définir la stratégie future à déployer l'année suivante pour les échantillons collectés en 2021 dans les DROM.
5. **Recevoir les échantillons prélevés dans les DROM au cours de l'année 2021**

Les échantillons prélevés dans les DROM dans le cadre des suivis DCE seront rapatriés à l'UMR Carrtel en fin d'année 2021 pour extraire l'ADNe et les séquencer en 2022. Nous collecterons également les retours d'expérience sur l'applicabilité de ces protocoles en routine. Les frais de d'acheminement seront pris en charge par le projet.

Bibliographie

Alverson AJ, Cannone JJ, Gutell RR, Theriot EC (2006) The evolution of elongate shape in diatoms. Journal of Phycology42, 655-668.

Benson, Dennis A., et al. "GenBank." *Nucleic acids research* 33.suppl_1 (2005): D34-D38.

Bradley, I. M., Pinto, A. J. & Guest, J. S. Design and Evaluation of Illumina MiSeq-Compatible, 18S rRNA Gene-Specific Primers for Improved Characterization of Mixed Phototrophic Communities. *Applied and Environmental Microbiology* 82, 5878–5891, (2016).

CBOL Plant Working Group, and Daniel H. Janzen. "A DNA barcode for land plants." *Proceedings of the National Academy of Sciences of the United States of America* (2009): 12794-12797.

Chapman, R. L., M. A. Buchheim, C. F. Delwiche, T. Friedl, V. A. R. Huss, K. G. Karol, L. A. Lewis, J. Manhart, R. M. McCourt, J. L. Olsen & D. A. Waters, 1998. Molecular systematics of the green algae. In Soltis, D. E., P. S. Soltis & J. J. Doyle (eds), Molecular Systematics of Plants II. DNA Sequencing. Kluwer Academic Publishers, Boston, Dordrecht, London: 508–540.

Cock PA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B and de Hoon MJL (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25, 1422-1423

Decelle, J., Romac, S., Stern, R. F., Bendif, E. M., Zingone, A., Audic, S., Guiry, M. D., Guillou, L., Tessier, D., Le Gall, F., Gourvil, P., Dos Santos, A. L., Probert, I., Vaulot, D., de Vargas, C. and Christen, R. (2015), PhytoREF: a reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes with curated taxonomy. Mol Ecol Resour, 15: 1435-1445. doi:10.1111/1755-0998.12401

Delwiche CF, Kuhsel M, Palmer JD. Phylogenetic analysis of *tuf*A sequences indicates a cyanobacterial origin of all plastids. Mol Phylog Evol. 1995;4:110–28.

Djemiel, C., Plassard, D., Terrat, S. *et al.* µgreen-db: a reference database for the 23S rRNA gene of eukaryotic plastids and cyanobacteria. *Sci Rep* 10, 5915 (2020).

Isabelle Domaizon, Rainer Kurmayer, Camilla Capelli, Cécile Chardon, Peter Hufnagl, Marine Vautier, Nico Salmaso 2019. Lake plankton sample collection from the field for downstream molecular analysis. protocols.io https://dx.doi.org/10.17504/protocols.io.xn6fmhe

Edgar, R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5,** 113 (2004).

Fuller, Nicholas J., et al. "Analysis of photosynthetic picoeukaryote diversity at open ocean sites in the Arabian Sea using a PCR biased towards marine algal plastids." *Aquatic Microbial Ecology* 43.1 (2006): 79-93.

Gray, Michael W. "Mitochondrial evolution." *Cold Spring Harbor perspectives in biology* 4.9 (2012): a011403.

Guiry, M.D. & Guiry, G.M. 2020. *AlgaeBase*. World-wide electronic publication, National University of Ireland, Galway. https://www.algaebase.org; searched on 05 November 2020.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C. et al. 2013. The Protist Ribosomal Reference database (PR^2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids Res. 41:D597–604.

Gutell, R. R., Larsen, N. & Woese, C. R. Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiological Reviews* 58, 10–26, (1994)

Hadziavdic, Kenan, et al. "Characterization of the 18S rRNA gene for designing universal eukaryote specific primers." *PloS one* 9.2 (2014): e87624.

Hadoux, E., et al. "PHYTOBS v2. 3: Outil de comptage du phytoplancton en laboratoire et de calcul de l'IPLAC. Version 2.3. Application JAVA." (2015).

Hanyuda, Takeaki, Shogo Arai, and Kunihiko Ueda. "Variability in the rbcL introns of Caulerpalean algae (Chlorophyta, Ulvophyceae)." *Journal of Plant Research* 113.4 (2000): 403.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society of London. Series B: Biological Sciences 270: 313–321.

Hebert, Paul DN, Sujeevan Ratnasingham, and Jeremy R. De Waard. "Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species." *Proceedings of the Royal Society of London. Series B: Biological Sciences* 270.suppl_1 (2003): S96-S99.

Hugerth, L. W., Muller, E. E., Hu, Y. O., Lebrun, L. A., Roume, H., Lundin, D., ... & Andersson, A. F. (2014). Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. *PloS one*, *9*(4), e95567.

Iwabe N, Kuma KI, Hasegawa M, Osawa S, Miyata T. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci USA. 1989;86:9355–9.

Krienitz, Lothar, and Christina Bock. "Present state of the systematics of planktonic coccoid green algae of inland waters." *Hydrobiologia* 698.1 (2012): 295-326.

Laplace-Treyture, Christophe, and Thibaut Feret. "Performance of the Phytoplankton Index for Lakes (IPLAC): A multimetric phytoplankton index to assess the ecological status of water bodies in France." *Ecological Indicators* 69 (2016): 686-698.

Laplace-Treyture, C. 2020. Surveillance de la communauté phytoplanctonique du plan d'eau de Gaschet et applicabilité de l'IPLAC– Années 2017-2019.

Lürling, Miquel. *The smell of water: grazer-induced colony formation in Scenedesmus*. 1999.

Marcelino, V. R. & Verbruggen, H. Multi-marker metabarcoding of coral skeletons reveals a rich microbiome and diverse evolutionary origins of endolithic algae. *Scientific Reports* 6, 1–9 (2016).

Meier, Rudolf, et al. "DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success." *Systematic biology* 55.5 (2006): 715-728.

Needham, David M., and Jed A. Fuhrman. "Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom." *Nature microbiology* 1.4 (2016): 1-7.

Nübel, Ulrich, Ferran Garcia-Pichel, and Gerard Muyzer. "PCR primers to amplify 16S rRNA genes from cyanobacteria." *Applied and environmental microbiology* 63.8 (1997): 3327-3332.

Owczarzy, Richard, et al. "IDT SciTools: a suite for analysis and design of nucleic acid oligomers." *Nucleic acids research* 36.suppl_2 (2008): W163-W169.

Paul, J. H., L. Cazares, and J. Thurmond. "Amplification of the rbcL gene from dissolved and particulate DNA from aquatic

environments." *Applied and Environmental Microbiology* 56.6 (1990): 1963-1966.
Pawlowski J, Audic S, Adl S, et al. (2012) CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms. PLoS Biol10, e1001419.
Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A., Borja, A., Bouchez, A., Cordier, T., Domaizon, I., Feio, M.J., Filipe, A.F., Fornaroli, R., Graf, W., Herder, J., van der Hoorn, B., Iwan Jones, J., Sagova-Mareckova, M., Moritz, C., Barquín, J., Piggott, J.J., Pinna, M., Rimet, F., Rinkevich, B., Sousa-Santos, C., Specchia, V., Trobajo, R., Vasselon, V., Vitecek, S., Zimmerman, J., Weigand, A., Leese, F., Kahlert, M., 2018. The future of biotic indices in the ecogenomic era: Integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. Sci. Total Environ. 637–638, 1295–1310.
Pei, A. *et al*. Diversity of 23S rRNA Genes within Individual Prokaryotic Genomes. *PLoS One* 4, e5437,(2009).
Pompanon F, Coissac E, Taberlet P (2011) Metabarcoding a new way to analyze biodiversity. Biofutur: 30–32.
Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.
Ram, Jeffrey L., et al. "Strategy for microbiome analysis using 16S rRNA gene sequence analysis on the Illumina sequencing platform." *Systems biology in reproductive medicine* 57.3 (2011): 162-170.
Reysenbach, A.-L., Pace, N.R., 1995. In: Robb, F.T., Place, A.R. (Eds.), Archaea: A Laboratory Manual—Thermophiles. Cold Spring Harbour Laboratory Press, New York, pp. 101 -107.
Rimet, Frédéric, et al. "Diat. barcode, an open-access curated barcode library for diatoms." *Scientific reports* 9.1 (2019): 1-12.
Rivera, S.F., Vasselon, V., Jacquet, S. *et al.* Metabarcoding of lake benthic diatoms: from structure assemblages to ecological assessment. *Hydrobiologia* 807, 37–51 (2018).
Rudi, K., Skulberg, O.M., Larsen, F., Jacoksen, K.S., 1997. Strain classification of oxyphotobacteria in clone cultures on the basis of 16S rRNA sequences from variable regions V6, V7 and V8.Appl. Environ. Microbiol. 63, 2593-2599.
Saez, Alberto G., Alejandro Zaldivar-Riverón, and Linda K. Medlin. "Molecular systematics of the Pleurochrysidaceae, a family of coastal coccolithophores (Haptophyta)." *Journal of plankton research* 30.5 (2008): 559-566.
Saunders, Gary W., and Hana Kucera. "An evaluation of rbcL, tufA, UPA, LSU and ITS as DNA barcode markers for the marine green macroalgae." *Cryptogamie, Algologie* 31.4 (2010): 487-528.
Sauvage, T., Schmidt, W.E., Suda, S. *et al.* A metabarcoding framework for facilitated survey of endolithic phototrophs with *tuf*A. *BMC Ecol* 16, 8 (2016).
Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., … others. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, *75*(23), 7537–7541.
Sherwood, A. R. & Presting, G. G. Universal primers amplify a 23S rDNA plastid marker in eukaryotic algae and cyanobacteria. *Journal of Phycology* 43, 605–608, (2007).
Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M.D., Breiner, H.W., Richards, T.A., 2010.Multiple marker parallel tag environmental DNA sequencing reveals a highly com-plex eukaryotic community in marine anoxic water. Mol. Ecol. 19 (Suppl. 1):21–31.
Vasselon, Valentin, et al. "Assessing ecological status with diatoms DNA metabarcoding: scaling-up on a WFD monitoring network (Mayotte island, France)." *Ecological Indicators* 82 (2017): 1-12.
Vieira, Helena Henriques, et al. "tufA gene as molecular marker for freshwater Chlorophyceae." *Algae* 31.2 (2016): 155-165.
Wang, Yong, and Pei-Yuan Qian. "Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies." *PloS one* 4.10 (2009): e7401.
Wang, Y., Tian, R. M., Gao, Z. M., Bougouffa, S., & Qian, P. Y. (2014). Optimal eukaryotic 18S and universal 16S/18S ribosomal RNA primers and their application in a study of symbiosis. *PloS one*, *9*(3), e90053.
Warwick, R. M., and K. R. Clarke. "New\'biodiversity\'measures reveal a decrease in taxonomic distinctness with increasing stress." *Marine ecology progress series* 129 (1995): 301-305.
Watanabe, K., Kodama, Y., Harayama, S., 2001. Design and evaluation of PCR primers to amplify 16S ribosomal DNA fragments used for community fingerprinting. J. Microbiol. Methods44, 253-262.
Whatley JM (1993) The Endosymbiotic Origin of Chloroplasts. In: Jeon KW, Jarvik J (Eds), International Review of Cytology. Academic Press, 259–299.
Zhang, Guang K., et al. "Metabarcoding using multiplexed markers increases species detection in complex zooplankton communities." *Evolutionary Applications* 11.10 (2018): 1901-1914.
Zimmermann, J., Jahn, R., & Gemeinholzer, B. (2011). Barcoding diatoms: evaluation of the V4 subregion on the 18S rRNA gene, including new primers and protocols. *Organisms Diversity & Evolution*, *11*(3), 173.
Zou, Shanmei, et al. "How DNA barcoding can be more effective in microalgae identification: a case of cryptic diversity revelation in Scenedesmus (Chlorophyceae)." *Scientific reports* 6 (2016): 36822.

Annexes

16S résolution des régions variables

Tableau A1 : Représentation brève des résolutions de certains couples de primers sur les séquences 16S récupérées (2120 dont 1480 cyanobactéries) pour le phytoplancton d'eau douce.

16S	Approx pos	Fwd	Rvr	Taille amplicon*	matrix	Amplicons	Amplicons 2	Resolution
V1-V2		27F	CYA359F	340		336	282	83,92 %
V3		CYA359F	PLA491F	130		1701	607	35,68 %
V3-V4		341F	805R	443		1845	1156	62,65 %
V3-V4		CYA359F	CYA781Rd	425		1756	1070	60,93 %
V3-V4		CYA359F	805R	425		1794	1104	61,53 %
V4		PLA491F	CYA781Rd	315		1611	841	52,2 %
V4		PLA491F	805R	315		1677	899	53,6 %
V5		CYA781Rd	E939R	160		1871	502	26,83 %
V5-V6		CYA781Rd	E1115R	336		1677	881	52,53 %
V6		E939R	E1115R	199		1799	857	47.63 %
V6-V7		E939R	OXY1313R	417		1675	1043	62,94 %
V7		E1115R	OXY1313R	234		1662	829	49,87 %
V7-V8		E1115R	U1406/15	308		1809	994	54,94 %
V9		U1406/15	U1541R	153		585	349	59,65 %
V7-V8-V9		E1115R	U1541R	443		569	430	75,57 %

***primers inclus**

Remarque : attention ce ne sont pas des couples de primers valides, simplement ils délimitent les extrémités des régions simplement pour vérifier leur résolution de manière plus large (donc avec des primers 'valides *in silico/in vitro*' on aurait des résolutions moins élevées).

Séquences de l'ensemble des primers évoqués dans ce rapport

Tableau A2 : Séquences et références pour l'ensemble des primers évoqués dans le rapport.
Repères : en bleu les primers issus de la littérature, en vert ceux trouvés dans cette étude.

16S		Ref
27F	AGAGTTTGATCMTGGCTCAG	West et *al.*, 2001
341F	CCTAYGGGRBGCASCAG	Klindworth et *al.*, 2013
CYA359F	GGGGAATYTTCCGCAATGGG	Nübel et *al.*, 1997
PLA491F	GAGGAATAAGCATCGGCTAA	Fuller et al., 2006
CYA781R(a)	GACTACTGGGGTATCTAATCCCATT	Nübel et *al.*, 1997
CYA781(b)	GACTACAGGGGTATCTAATCCCTTT	Nübel et *al.*, 1997
CYA781Rd	GACTACWGGGGTATCTAATCCCWTT	adapté
805R	GACTACHVGGGTATCTAATCC	Füller et *al.*, 2006
E939R	GAATTGACGGGGGCCCGCACAAG	Rudi et *al.*, 1997
E1115R	AGGGTTGCGCTCGTTG	Reysenbach & Pace, 1995
OXY1313R	CTTCACGTAGGCGAGTTGCAGC	West et *al.*, 2001
U1406/15	TGTACACACCGCCCGTCA	
U1541R	AAGGAGGTGATCCANCCRCA	
515F	GTGCCAGCMGCCGCGGTAA	Parada et *al.*, 2016
926R	CCGYCAATTYMTTTRAGTTT	Parada et *al.*, 2016
PhytoF	GKAGCGGTGAAATGCGTAGAK	ici
PhytoR	GCTGACGACAGCCATGCA	Ici
Test1F	AGCGGTGAAATGCGTAGAK	Ici
Test1R	AGCTGACGACAGCCATGCA	Ici
Test2F	AGCGGTGAAATGCGTAGAKA	Ici
Test2R	AGCTGACGACAGCCATGCA	Ici
Test3F	AGCGGTGAAATGCGTAGAKAT	Ici
Test3R	AGCTGACGACAGCCATGCA	Ici
Test4F	AGCGGTGAAATGCGTAGAKATY	ici
Test4R	AGCTGACGACAGCCATGCA	Ici
Test5F	GCGGTGAAATGCGTAGAKA	Ici
Test5R	AGCTGACGACAGCCATGCA	ici
18S		
TAREuk454FWD1	CCAGCASCYGCGGTAATTCC	Stoeck et *al.*, 2010
TAREukRev3	ACTTTCGTTCTTGATYRA	Stoeck et *al.*, 2010
D514for18S	TCCAGCTCCAATAGCGTA	Zimmerman et *al.*, 2011
D978rev18S	GACTACGATGGTATCTAATC	Zimmerman et *al.*, 2011
D512for18S	ATTCCAGCTCCAATAGCG	Zimmerman et *al.*, 2011
Euk690R	ATCCAAGAATTTCACCTCTGA	Elwood et *al.*, 1985
Test101F	CGGTAATTCCAGCTCCAA	ici
Test101R	GGTATCTRATCRTCTTCGAK	Ici
Rhodo101F	CGCGGTAATTCCAGCTCY	Ici
Euk690RhodoTm	ATCCAAGAATTTCACCTCTRA	Ici
23S		
P23SrV_f1	GGACAGAAAGACCCTATGAA	Sherwood & Presting, 2007
P23SrV_r1	TCAGCCTGTTATCCCTAGAG	Sherwood & Presting, 2007
A23SrV_F1	GGACARAAAGACCCTATG	Yoon et *al.*, 2016
A23SrV_R1	AGATCAGCCTGTTATCC	Yoon et *al.*, 2016
A23SrV_F2	CARAAAGACCCTATGMAGCT	Yoon et *al.*, 2016
A23SrV_R2	TCAGCCTGTTATCCCTAG	Yoon et *al.*, 2016
Test253F	GACAGWAAGACCCTATGAAGCT	ici
Test253R	ATCAGCCTGTTATCCCTAGAGT	Ici
Test587F	GACAGWAAGACCCTATGAAGCT	Ici
Test587R	ATCAGCCTGTTATCCCTAGAG	Ici
Test108F	ACAGWAAGACCCTATGAAGCTT	Ici
Test108R	CCTGTTATCCCTAGAGTAACTT	ici
***rbc*L**		
P169	GATGATGARAAYATTAACCT	Paul et *al.*, 1990

P328	ATTTGDCCACAGTGDATACCA	Paul et *al.*, 1990
21-merP3	TCIGCIAARAACTAYGGTCG	Paul et *al.*, 1990
18-merP6	GGCATRTGCCAIACRTGRAT	Paul et *al.*, 1990
***tuf*A**		
tufAF	TGAAACAGAAMAWCGTCATTATGC	Fama et *al.*, 2002
tufAR	CCTTCNCGAATMGCRAAWCGC	Fama et *al.*, 2002
tufA_SF	TGGATGGTGCWATTYTWG	Zou et *al.*, 2016
tufA_SR	GGTTTTGCWAAAACCATWCCACG	Zou et *al.*, 2016
Tubryo_F	GCAGATGGTCCAATGCCWCAAAC	Sauvage et *al.*, 2016
Tubryo_R	CCWGGTTTAGCTAAAACCATNCC	Sauvage et *al.*, 2016
envtufA_F	TGGGTDGAHAADATTTWYNMNYTRATGR	Sauvage et *al.*, 2016
envtufA_R	TNACATCHGTWGTWCKNACATARAAYTG	Sauvage et *al.*, 2016
Test129F	CWAAACAAGTWGGWGTWCCW	ici
Test129R	TWCCWCGWCCWGTAATWGA	Ici
Test130F	CWAAACAAGTWGGWGTWCCW	Ici
Test130R	GTWCCWCGWCCWGTAATWGA	Ici
Test126F	CWAAACAAGTWGGWGTWCCW	Ici
Test126R	GTWCCWCGWCCWGTAATWG	ici

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

Matériels Supplémentaires

- Mat.Supp.1 : Liste DOM
- Mat.Supp.2 : Liste PHYTOBS homogénéisée AlgaeBase
- Mat.Supp.3 : Liste IPLAC
- Mat.Supp.4 : Taxonomie brute AlgaeBase
- Mat.Supp.5 : Listes ADN phytoplancton (.zip)
- Mat.Supp.6 : Protocole échantillonnage

L'ensemble des matériels supplémentaires sont retrouvés en ligne *via* le lien suivant :

CANINO Alexis; LAPLACE-TREYTURE Christophe; BOUCHEZ Agnès; DOMAIZON Isabelle; RIMET Frédéric, 2021, "Données de réplication pour : PhytoDOM - Matériels Supplémentaires Rapport OFB 2020", https://doi.org/10.15454/W9JGOA, Portail Data INRAE, V1

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