



HAL
open science

Barcoding et bioindication : développement du metabarcoding des diatomées pour l'évaluation de la qualité des cours d'eau

Valentin Vasselon

► To cite this version:

Valentin Vasselon. Barcoding et bioindication : développement du metabarcoding des diatomées pour l'évaluation de la qualité des cours d'eau. Biodiversité et Ecologie. Université Grenoble Alpes, 2017. Français. NNT: . tel-01815806

HAL Id: tel-01815806

<https://hal.science/tel-01815806>

Submitted on 14 Jun 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE

Pour obtenir le grade de
**DOCTEUR DE LA COMMUNAUTE UNIVERSITE
GRENOBLE ALPES**

Spécialité : **Biodiversité, Ecologie, Environnement**

Arrêté ministériel : 25 mai 2016

Présentée par

Valentin Vasselon

Thèse dirigée par **Agnès BOUCHEZ, DR, CARTEL**
et codirigée par **Isabelle DOMAIZON, DR, CARTEL**

préparée au sein du **Laboratoire CARTEL (Université Savoie
Mont-Blanc, INRA)**
dans l'**École Doctorale SISEO**

**Barcoding et bioindication : développement
du metabarcoding des diatomées pour
l'évaluation de la qualité des cours d'eau**

Thèse soutenue publiquement le **08 Décembre 2017**,
devant le jury composé de :

M. Yorick REYJOL

Chargé de mission ONEMA-AFB (Rapporteur)

M. Didier PONT

Directeur de recherche IRSTEA (Rapporteur)

M. Wim VYVERMAN

Professeur de l'Université de Ghent (Membre)

M. Marc BUEE

Directeur de recherche INRA (Membre, Président du jury)

M. Frédéric RIMET

Ingénieur d'étude INRA (Membre invité)



« Un problème créé ne peut être résolu en réfléchissant de la même manière qu'il a été créé. »

Albert Einstein

Remerciements

- Cette thèse n'aurait pas pu être réalisée sans l'appui de l'ONEMA-AFB.
- En premier lieu je souhaite remercier mes deux encadrantes de thèse, **Agnès Bouchez** et **Isabelle Domaizon**, pour m'avoir fait confiance tout au long de la thèse. Elles ont su trouver l'équilibre parfait me permettant d'être autonome et créatif (peut-être parfois trop et je salue leur patience au passage !), tout en me guidant et me conseillant vers les choix les plus judicieux. J'ai vraiment pris énormément de plaisir à travailler avec vous.
- Je tiens à remercier **Wim Vyvermann**, **Marc Buée**, **Didier Pont** et **Yorick Reyjol** qui ont très gentiment accepté de faire partie de mon jury de thèse et de juger mon travail.
- Un très grand merci à **Frédéric Rimet** avec qui j'ai vraiment apprécié collaborer et qui, au même titre que mes encadrantes de thèse, à très largement contribué au bon déroulement de cette thèse. S'il est possible de juger la valeur d'un homme à son T-shirt, alors au premier coup d'œil on sait que Fred est vraiment quelqu'un de super !
- Je souhaite remercier **Cécile Chardon** avec qui j'ai beaucoup travaillé au laboratoire de biomol et que je considère plus comme une amie que comme une collègue. J'ai vraiment apprécié nos soirées avec Céline et Luc, et attend avec grand plaisir les prochaines (ne t'inquiètes pas, un jour tu gagneras au tarot).
- Je tiens à remercier la direction du laboratoire, **Bernard Montuelle** et **Jean Guillard**, pour avoir toléré autant de fantaisie (et je ne parle pas juste de la décoration du bureau...).
- Comme il me paraît compliqué de remercier comme il se doit toutes les personnes qui le méritent (collègues, amis et famille) je me limiterai aux personnes avec qui j'ai accompli de nombreux méfaits au sein du laboratoire (pour les autres un très grand merci quand même !). Afin d'éviter de leur attirer des ennuis et ne pas compromettre leur identité, je ne divulgerai que leur prénom : **Bernard, Eric, François, Frédéric (le grand), Frédéric (le roux), Kálmán, Laurent, Lisandrina, Marine, Rosalie, Stéphan, Teofana, Yoann**.
- Je tiens à remercier **mes parents**, qui ont toujours été présents et m'ont soutenu tout au long de ma vie d'étudiant (avec pas mal de moments de stress pour eux). J'espère être digne de mon titre de chevalier de l'ordre de la confrérie du babet !
- Plus qu'un merci, je souhaite décerner la palme de la patience à ma femme **Amélie**. Même si elle ne s'en est rendue pas compte, elle a grandement contribué au bon déroulement de cette thèse. Je sais que je vais regretter de l'avoir mis par écrit, mais bon ... tu es parfaite !
- Enfin, une pensée émue à Maléfica, Blanche-neige et Mylène disparues dans des circonstances tragiques, victimes de la sélection naturelle.

Résumé

Les diatomées sont des algues unicellulaires microscopiques qui sont d'excellents indicateurs de l'état écologique du milieu dans lequel elles se trouvent. Dans le cadre de la Directive Cadre sur l'Eau (DCE), les communautés de diatomées sont utilisées pour évaluer la qualité des cours d'eau. Pour cela, des indices de qualité basés sur la sensibilité des espèces à la pollution sont calculés à partir de la composition et de l'abondance relative des taxa de diatomées. L'identification des espèces est généralement réalisée au microscope, ce qui, en plus d'être complexe, peut être assez long et coûteux lorsqu'il s'agit de traiter de nombreux échantillons.

Une nouvelle méthode développée récemment permet d'identifier les espèces, non plus sur la base de leur variabilité morphologique, mais sur la base de leur variabilité génétique en utilisant de courtes séquences ADN (ou barcodes ADN). Combinée aux technologies de séquençage à haut débit, cette approche moléculaire, appelée metabarcoding, permet d'identifier l'ensemble des espèces présentes au sein d'un échantillon environnemental et de traiter plusieurs centaines d'échantillons en parallèle. Ces avantages font du metabarcoding une alternative à la méthode d'identification morphologique, intéressante dans le cadre de la DCE. Bien que plusieurs études aient montré la capacité de cette approche à identifier correctement les espèces de diatomées retrouvées dans des échantillons environnementaux, le manque de fiabilité dans la quantification des abondances relatives des espèces limite le calcul d'indices fiable et l'utilisation du metabarcoding comme outil pour la bioindication. Les objectifs de ce travail de thèse ont donc été (i) d'identifier et d'optimiser les biais impactant la quantification relative des diatomées en metabarcoding ; (ii) d'appliquer l'approche moléculaire à large échelle, sur des échantillons environnementaux de réseaux de cours d'eau afin de comparer les évaluations de qualité obtenues par les approches morphologique et moléculaire.

Dans un premier temps, nous avons évalué les biais de quantification liés à l'extraction de l'ADN sur des cultures pures de diatomées et des échantillons environnementaux. Bien que le choix de la méthode d'extraction affecte la qualité et la quantité des ADN extraits, ainsi que les abondances relatives de certaines espèces obtenues en metabarcoding, la composition de la communauté ainsi que les notes de qualités ne sont pas significativement affectées. Nous avons donc décidé d'utiliser pour la suite des travaux la méthode GenElute qui produit les plus grandes quantités d'ADN pour un moindre coût. Dans un deuxième temps, grâce à des expériences de qPCR réalisées sur des cultures pures de diatomées, nous avons montré que le nombre de copies du gène *rbcL* (utilisé comme barcode ADN) est proportionnel au biovolume des cellules, ce qui a pour conséquence une surestimation des espèces à gros biovolumes en metabarcoding. A partir de cette corrélation, un facteur de correction a été proposé et appliqué sur les données de

metabarcoding issues de communautés artificielles et d'échantillons environnementaux, permettant d'obtenir des abondances relatives d'espèces comparables à celles obtenues en microscopie et d'améliorer la fiabilité des notes de qualité. Finalement, l'application de l'approche moléculaire à l'échelle des réseaux DCE de surveillance des cours d'eau de Mayotte et de France métropolitaine a permis de montrer que le metabarcoding est une alternative plus rapide et plus économique que l'approche morphologique, tout en permettant une bonne évaluation de la qualité des cours d'eau.

Nos travaux confirment que l'approche moléculaire peut être utilisée pour évaluer la qualité des cours d'eau. Cependant d'autres études devront être réalisées avant d'envisager une application en routine et une implémentation dans la DCE, notamment en termes de standardisation et de normalisation des méthodes utilisées dans l'approche moléculaire.

Abstract

Diatoms are microscopic unicellular algae which are excellent indicators of the ecological status of the environment in which they live. In the Water Framework Directive (WFD), diatom communities are used to assess the quality of rivers. For this purpose, quality indices based on the susceptibility of species to pollution are calculated from the composition and relative abundance of diatom taxa. Species identification is typically carried out under a microscope, which, in addition to being complex, can be time-consuming and costly when many samples are processed.

A newly developed method allows species to be identified, not on the basis of their morphological variability, but on the basis of their genetic variability using short DNA sequences (or DNA barcodes). Combined with high-throughput sequencing technologies, this molecular approach, called metabarcoding, allows to identify all the species present in an environmental sample and to process several hundred samples in parallel. These advantages make metabarcoding an alternative method to those based on morphological identification, which is interesting for the WFD. Although several studies have demonstrated the ability of this approach to correctly identify diatom species found in environmental samples, the lack of reliability in the quantification of the species relative abundances limits the reliability of calculated quality indices and the use of metabarcoding as a tool for biomonitoring. The objectives of this thesis were (i) to identify and optimize the biases impacting the relative quantification of diatoms in metabarcoding; (ii) to apply the molecular approach at a larger scale on environmental samples from river networks in order to compare quality assessments obtained by morphological and molecular approaches.

First, we quantified the bias associated with DNA extraction using pure diatom cultures and environmental samples. Although the choice of the extraction method affects the quality and the quantity of extracted DNA, as the relative abundances of some species identified with metabarcoding, the community composition and water quality index values are not significantly affected. Thus, we decided to use the GenElute method for other works realized during the thesis, as this method produces the largest quantities of DNA at a lower cost. In a second step, using qPCR experiments carried out on pure diatom cultures, we showed that the *rbcL* gene copy number (used as DNA barcode) is correlated to the cell biovolume, which results in an overestimation of large biovolume species with the metabarcoding. On the basis of this correlation, a correction factor was proposed and applied to metabarcoding data obtained from artificial diatom communities and environmental samples, allowing to obtain species relative abundances similar to those obtained with the morphological approach and to improve the reliability of water quality index values. Finally, the application of the molecular approach at the scale of the WFD rivers

monitoring networks of Mayotte and France has enabled us to show that metabarcoding is a faster and cheaper alternative to the morphological approach, while allowing a good water quality assessment of streams.

Our work confirms that the molecular approach can be used to assess the quality of rivers. However, further studies will have to be performed before considering a routine application and its implementation in the WFD, particularly in terms of standardization of the methods used in the molecular approach.

Table des matières

I. Contexte général	1
1. Gestion des écosystèmes d'eau douce	2
1.1. Ecosystèmes d'eau douce	3
1.2. Evolution des politiques de gestion de l'eau	7
1.3. Evaluation de la qualité des cours d'eau	10
2. Les diatomées comme bioindicateurs	13
2.1. Biologie des diatomées	13
2.2. Ecologie générale et réponses aux pressions environnementales.....	18
2.3. Evaluation de la qualité des cours d'eau via les diatomées	23
3. Metabarcoding et bioindication.....	30
3.1. Développement du metabarcoding	31
3.2. Intérêt pour la bioindication	37
4. Metabarcoding des diatomées	41
4.1. Potentiel pour évaluer l'état écologique des cours d'eau	41
4.2. Verrous majeurs liés à la quantification	45
4.3. Implémentation à l'échelle de la DCE	49
5. Objectifs de la thèse.....	51
5.1. Stratégie	51
5.2. Structure de la thèse	55
6. Annexes	56
II. Biais lié à la méthode d'extraction de l'ADN	59
1. Abstract	61
2. Introduction	62
3. Methods	64
3.1. Diatom cultures.....	64
3.2. Environmental community samples	64
3.3. DNA extraction.....	65
3.4. Evaluation of DNA extraction efficiency and DNA quality	66
3.5. Polymerase Chain Reaction (PCR) inhibitor detection (quantitative PCR [qPCR])	67
3.6. Preparation of the library of amplicons and HTS sequencing	68
3.7. Sequence data processing	69
3.8. Statistical analysis on community structure as revealed by HTS.....	70
3.9. Comparison between molecular and morphological taxonomic inventories	70
4. Results	71
4.1. DNA extraction efficiency, quality, and PCR inhibition.....	71
4.2. Comparison of diatom quantification: microscopy vs qPCR.....	72
4.3. Effect of extraction methods on richness, composition, and structure of diatom community	72
4.4. Morphology vs molecular diatom community composition	76
5. Discussion.....	77
5.1. DNA extraction method affects quantity and quality of extracted DNA.....	77
5.2. Diatom community composition is unchanged whatever the extraction method	79
5.3. Accuracy of molecular inventories and downstream quality indices.....	81
6. Conclusion	83
7. Acknowledgements.....	84
8. Author contributions.....	84
9. Supplementary data	85

III. Biais lié à la variation du nombre de copie de gène	93
1. Abstract	95
2. Introduction	95
3. Methods	97
3.1. Evaluation of the quantification bias and development of a quantification correction factor (CF).....	97
3.2. Validation of the quantification CF to mock and environmental HTS data	100
4. Results	102
4.1. Variation of <i>rbcL</i> gene copy number between diatom species	102
4.2. Development of quantification CFs	103
4.3. Application of CFs to mock and environmental HTS data	104
5. Discussion.....	106
5.1. Correlation between <i>rbcL</i> gene copy number and diatom cell biovolume: impacts on HTS quantification.....	107
5.2. Current potential and limits of the quantification CF.....	108
6. Acknowledgments.....	110
7. Data accessibility.....	110
8. Author contributions.....	110
9. Supplementary data.....	111
IV. Application aux cours d'eau de Mayotte.....	121
1. Abstract	123
2. Introduction	124
3. Material and methods.....	126
3.1. Mayotte island monitoring network.....	126
3.2. Diatoms sampling.....	128
3.3. DNA Metabarcoding.....	129
3.4. Morphological analysis	130
3.5. Morphological and molecular SPI	131
3.6. Statistical analysis	131
4. Results	132
4.1. Morphological analysis	132
4.2. HTS analysis.....	132
4.3. Comparison of the diatom communities obtained by molecular or morphological assignment.....	133
4.4. Morphological and molecular SPI calculation.....	134
4.5. Impact of sequencing depth on richness, diversity and the water quality index.....	137
5. Discussion.....	138
5.1. Community structure and ecological quality status inferred by molecular and by morphological approaches are congruent.....	138
5.2. Biases explaining differences between molecular and morphological based diatom indices	140
6. Conclusion and perspectives.....	145
7. Acknowledgments.....	146
8. Author contributions.....	146
9. Supplementary data.....	147
V. Application au réseau de surveillance national	153
1. Résumé.....	155
2. Introduction	156
3. Matériels et méthodes.....	158

3.1. Sélection des sites	158
3.2. Echantillonnage des biofilms aquatiques	158
3.3. Approche morphologique	160
3.4. Approche moléculaire	160
3.5. Estimation des coûts et temps d'analyse des deux approches	161
3.6. Stratégie de complétion de la base de référence	162
3.7. Evaluation de l'état écologique des cours d'eau	162
4. Résultats préliminaires pour la campagne 2016.....	163
4.1. Mise en place de l'approche moléculaire	163
4.2. Comparaison des coûts et temps d'analyse des deux approches	164
4.3. Résultats du séquençage et complétion de la base de référence	165
4.4. Evaluation de la qualité des cours d'eau	166
5. Remerciements	168
VI. Discussion générale et perspectives.....	169
1. Développement de l'approche moléculaire	172
1.1. Impact des biais techniques : avancées et perspectives	172
1.2. Sources de divergences entre inventaires : deux visions d'une même réalité.....	177
2. Application de l'approche moléculaire pour l'évaluation de la qualité des cours d'eau	179
2.1. Congruence entre approches morphologique et moléculaire	179
2.2. Potentiel d'amélioration et d'évolution des indices diatomiques.....	182
2.3. Autre lecture des données HTS pour évaluer l'état écologique	184
3. Vers une utilisation en routine de l'approche moléculaire ?.....	186
3.1. Fiabilisation des données moléculaires	186
3.2. Intégration dans le paysage actuel de la bioindication et de la DCE	188
Références bibliographiques	190
Contributions scientifiques.....	213
Articles annexes.....	217

Liste des figures

Figure 1 – Impact de l’homme sur les écosystèmes marins.	2
Figure 2 – Cycles naturel et anthropogénique de l’eau à l’échelle globale.	4
Figure 3 – Principales sources de pollutions des eaux douces par les activités humaines.	5
Figure 4 – Répartition des stations des réseaux de contrôle de surveillance (RCS) et de contrôle opérationnel (RCO) pour les eaux de surface.	9
Figure 5 – Critères utilisés pour définir le bon état général des cours d’eau.	10
Figure 6 – Caractéristiques des 5 éléments de qualité biologique.	12
Figure 7 – Ultrastructure d’une cellule de diatomée avec ses principaux composants.	14
Figure 8 – Structure du frustule chez les diatomées centriques (a) et pennées (b).	14
Figure 9 – Multiplication végétative induisant une réduction de la taille des cellules.	16
Figure 10 – Reproduction sexuée et formation de l’auxospore.	17
Figure 11 – Exemples de formes de vie observées chez les diatomées planctoniques (a,b,c,d) et benthiques (e,f,g).	19
Figure 12 – Diversité biologique retrouvée au sein d’un biofilm aquatique de rivière.	22
Figure 13 – Evaluation de la qualité écologique des cours d’eau via l’approche morphologique.	26
Figure 14 – Les différentes étapes du barcoding.	32
Figure 15 – Les différentes étapes du metabarcoding.	33
Figure 16 – Exemples de facteurs influençant les communautés biologiques utilisées en bioindication.	38
Figure 17 – Evaluation de la qualité écologique des cours d’eau via l’approche moléculaire (metabarcoding).	43
Figure 18 – Erreurs et biais potentiellement introduits par la PCR.	48
Figure 19 – Characteristics of the diatom cultures from the Thonon Culture Collection (TCC) (A) and biofilm sampling sites (B).	64
Figure 20 – The main steps of DNA extraction are presented for the 5 methods.	66
Figure 21 – Dissimilarity between molecular inventories obtained with the 5 DNA extraction methods at 8 sampling sites.	73
Figure 22 – Diatom community structure as revealed by molecular inventory for the 5 DNA extraction methods (each performed in duplicates) at 8 sampling sites.	74
Figure 23 – Mean percentage of diatom genera and species detected by microscopy, by molecular inventory, or by both methods for the 8 sampling sites.	76
Figure 24 – Specific pollution-sensitivity index (SPI) values based on morphological inventories (counts obtained from microscopic observations of diatoms valves) and on molecular inventories (reads based on relative abundances estimated by high throughput sequencing [HTS]).	77
Figure 25 – Experimental design applied to the 8 diatom species.	99
Figure 26 – Estimation of the <i>rbcl</i> copy number per diatom cell for the 8 diatom species.	102
Figure 27 – Correlation between the diatom cell biovolume and the <i>rbcl</i> gene copy number per cell after $\log(x+1)$ transformation.	103
Figure 28 – Relative abundances of the 8 diatom species in the 5 DNA mock communities based (A) on mean of HTS DNA reads without (left) and with (right) correcting quantification using the biovolume correction factor and (B) on mean of morphological counts from inverted microscopy.	105
Figure 29 – Dominant taxa (relative abundance > 0.5 %) obtained in HTS Mayotte molecular inventories without (left) and with (right) application of the biovolume correction factor. All samples (n=80) are considered.	105
Figure 30 – Distribution of the differences between the molecular and the morphological SPI (Δ SPI) for all Mayotte samples using original molecular SPI values (left) and new molecular SPI values based on molecular inventories corrected with the biovolume CF (right).	106

Figure 31 – Graphical abstract.	124
Figure 32 – Graphical abstract Location of Mayotte island (France) and the 45 river sites of the three monitoring networks: Reference sites network (REF - white), Regular WFD monitoring network (RCS – grey) and Polluted sites network (POLL – black).	127
Figure 33 – Main environmental pressure gradients in rivers of Mayotte: dissolved organic carbon (DOC), total organic carbon (TOC), turbidity and suspended solids (SS).	128
Figure 34 – Venn diagrams comparing the diatom inventories assigned at family, genus and species levels either by the molecular (right circles) or by the morphological (left circles) approach (80 river samples).	133
Figure 35 – Distribution of the values of the diatom Specific Pollution Index (SPI) based on molecular (left) and morphological (right) inventories for all 80 samples within the 3 monitoring networks (POLL, RCS, REF).	134
Figure 36 – Correlation between the diatom Specific Pollution Index (SPI) based on molecular (y axis) and morphological (x axis) inventories for all 80 samples.	135
Figure 37 – Three-dimensional NMDS plots of Bray-Curtis dissimilarity based on OTU composition of all the 80 samples.	135
Figure 38 – Correlation between the Δ SPI (difference between the molecular SPI and the morphological SPI values) and the proportion in molecular inventories of (A) Eunotia taxa and (B) unclassified reads at Genus level, for all samples.	136
Figure 39 – Impact of the sequencing depth on molecular SPI values (left), OTUs richness (Chao estimator, middle), OTUs diversity (Shannon index, right) evaluated on all samples by performing random subsampling of DNA reads (subsampling decreasing from 5710 to 50 reads per sample).	137
Figure 40 – Localisation des 461 sites utilisés dans cette étude.	159
Figure 41 – Estimations des coûts et temps relatifs à la mise en place des deux approches d'identification des diatomées.	165
Figure 42 – Corrélation entre valeurs d'indices IBD moléculaire et IBD morphologique.	166
Figure 43 – Corrélation entre valeurs d'indices IPS moléculaire et IPS morphologique.	167
Figure 44 – Approches et techniques utilisées au cours de ces travaux de thèse.	172
Figure 45 – Evolution au cours du temps du nombre d'articles incluant les termes « metabarcoding » OU « environmental barcoding ».	173
Figure 46 – Abondances relatives des taxons de diatomées sensibles (cercle blanc) et tolérants (cercle noir) aux nutriments en fonction du ratio de qualité écologique (EQR) obtenus pour des cours d'eau.	181
Figure 47 – Différentes stratégies d'évaluations de l'état écologique des cours d'eau à partir des données moléculaires obtenues en metabarcoding.	185

Liste des tables

<i>Table 1 – Mean (SD, n = 72) DNA extraction efficiency and 260/280 DNA ratios obtained for the 5 extraction methods with the pure culture and environmental samples.</i>	<i>71</i>
<i>Table 2 – Mean (SD, n = 72) estimation of the inhibition level for DNA extracted from the environmental samples.</i>	<i>72</i>
<i>Table 3 – Results of the similarity percentages (SIMPER) analysis performed to identify the operational taxonomic units (OTUs) contributing to >1% of the dissimilarity between diatom communities obtained from the SA-Gen (SA) and MN-Soil extraction methods (MN).</i>	<i>75</i>
<i>Table 4 – Characteristics of the 8 diatom species selected in the Thonon Culture Collection (TCC) and used in this study.</i>	<i>98</i>
<i>Table 5 – CF calculated for the 8 diatom species using their respective cell biovolume (Table 4) and the linear equation between the rbcL copy number and the cell biovolume (Figure 27).....</i>	<i>104</i>

I. Contexte général

1. Gestion des écosystèmes d'eau douce

Au cours de son histoire, l'homme a sans cesse cherché à étudier et à comprendre son environnement. Dans un premier temps en décrivant ce qui l'entoure (*e.g.* la faune, la flore, les phénomènes naturels), il tente par la suite de déterminer quel impact l'environnement peut avoir sur lui et sur son mode de vie. Ainsi on retrouve dès l'antiquité des traités, comme « des airs, des eaux et des lieux » attribué à Hippocrate en 400 avant JC (Daremberg 1855), qui expliquent comment les vents, la qualité de l'air ou encore la qualité de l'eau peuvent affecter la santé humaine. Il faudra attendre les 17^{ème} et 18^{ème} siècles, avec l'expansion coloniale de l'Europe dans les tropiques, pour voir apparaître les premiers écrits où l'homme prend réellement conscience de son impact sur l'environnement (Grove & Damodaran 2006). A l'époque, les activités humaines pratiquées dans les îles, en lien avec la production du sucre, avaient accéléré les phénomènes de déforestation et d'érosion des sols notamment aux Bermudes (Tryon 1684) et à la Barbade (Grove 1995). Les colons réalisent alors qu'il est nécessaire de protéger les ressources environnementales afin de pérenniser leurs activités commerciales.

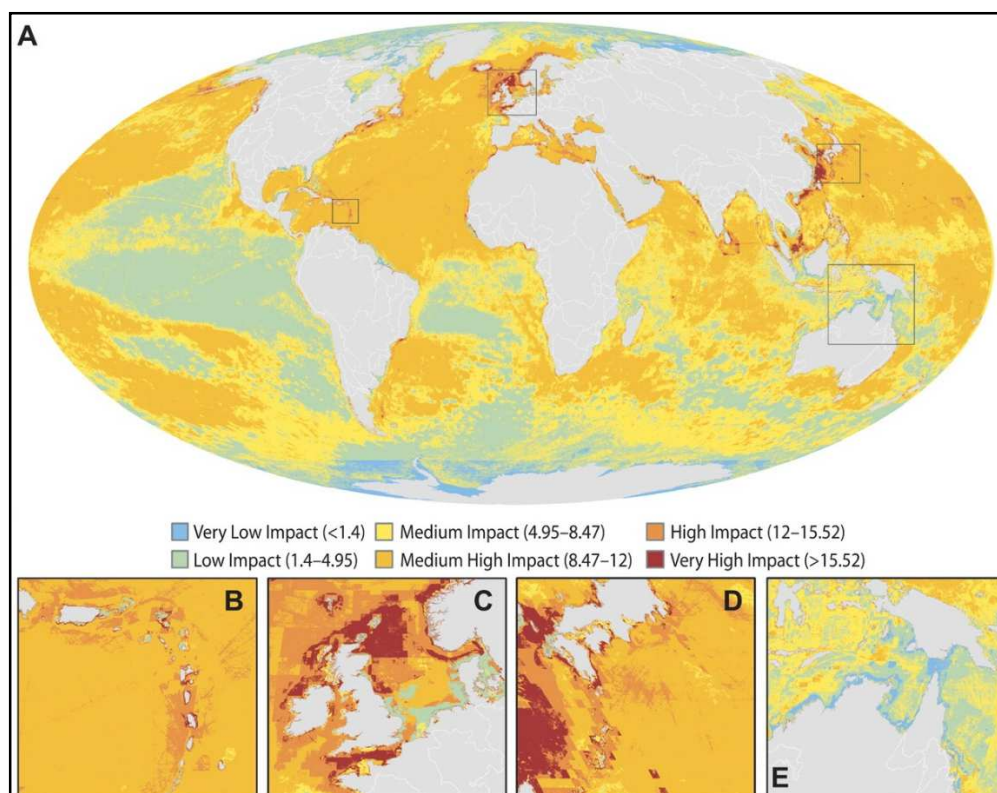


Figure 1 – Impact de l'homme sur les écosystèmes marins.
(source : Halpern *et al.* 2008)

Depuis, les activités humaines se sont développées et diversifiées, évoluant en parallèle des avancées technologiques et répondant aux exigences d'une société en perpétuelle croissance. L'urbanisation, les transports, l'agriculture et l'élevage, l'accès aux ressources, la production/consommation de biens et d'énergies n'ont eu de cesse de modifier et d'impacter l'environnement, si bien qu'actuellement aucun écosystème terrestre ou aquatique ne peut être considéré comme vierge d'activité anthropique (Vitousek *et al.* 1997; Halpern *et al.* 2008). Bien que les écosystèmes marins soient les plus étudiés et apparaissent comme très fortement impactés par les activités humaines (**Figure 1**), les écosystèmes d'eau douce sont les plus menacés et nécessitent donc une attention particulière (Abell 2002; Dudgeon *et al.* 2006; Collen *et al.* 2014).

1.1. Ecosystèmes d'eau douce

1.1.1. Ressources en eau douce

L'eau douce ne représente que 0,8 % de la surface totale du globe et seulement 3 % de l'eau sur Terre (Pimentel *et al.* 1997), ce qui en fait une ressource précieuse. Précieuse pour la biodiversité car, malgré sa faible proportion sur terre, les écosystèmes d'eau douce abritent 9,5 % des espèces animales et 1 % des plantes vasculaires décrites sur terre (Balian *et al.* 2008). Mais surtout précieuse car c'est une ressource vitale pour tous les organismes, dont l'homme, et qui offre de nombreux services. L'homme a ainsi pu bénéficier des ressources en eau douce pour ses besoins comme la consommation pour usages domestiques et industriels, l'irrigation, la pêche ou encore les loisirs, si bien que les activités humaines font parties intégrantes des flux d'eau à l'échelle globale (**Figure 2**). Bien que l'homme n'utilise que 10 % de la ressource en eau douce disponible chaque année (Oki & Kanae 2006), il n'en reste pas moins que 80 % de la population mondiale est exposée à un risque en lien avec la sécurité de l'eau (Vörösmarty *et al.* 2010). Le risque majeur est principalement lié au manque d'eau, avec 2/3 de la population mondiale affectée par de sévères pénuries au moins 1 mois par an, principalement en Chine et en Inde (Vörösmarty *et al.* 2000). Si les pays d'Europe semblent moins affectés par les pénuries d'eau, ils sont confrontés aux risques liés à une diminution de la qualité des eaux en lien avec les activités humaines, qui se traduit notamment par une perte de biodiversité dans les milieux aquatiques (Dodds *et al.* 2013).

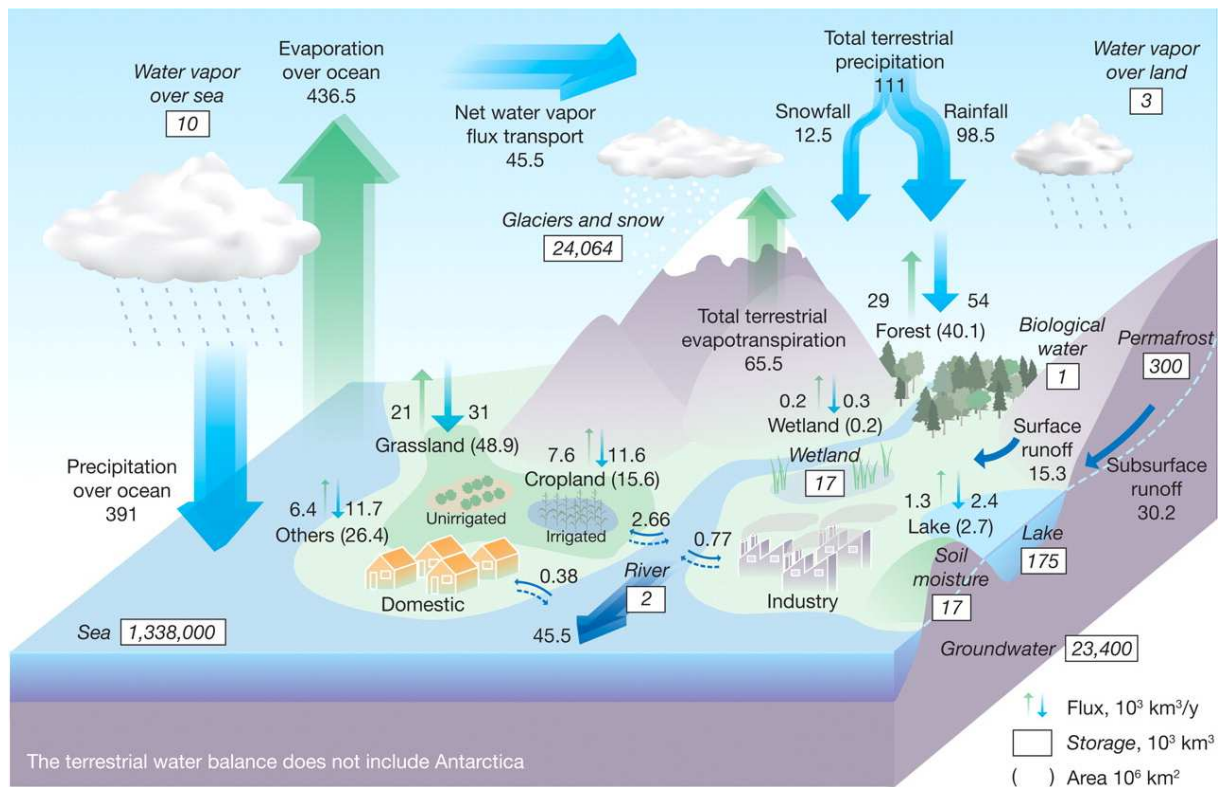


Figure 2 – Cycles naturel et anthropogénique de l'eau à l'échelle globale.
(source : Oki & Kanae 2006)

1.1.2. Sources de dégradation des milieux aquatiques

La plupart des activités anthropiques présentent un risque pour les milieux aquatiques et peuvent entraîner la dégradation des ressources en eau douce ainsi qu'une perte de biodiversité (Vörösmarty *et al.* 2010). On peut ainsi regrouper ces activités en fonction de leur impact sur l'environnement :

- **Altération physique ou hydromorphologique :** Afin de répondre aux différents besoins en eau liés à la production d'énergie, l'industrie, l'agriculture/l'élevage, la consommation humaine ou encore pour se protéger contre les crues, de nombreux aménagements et constructions ont été réalisés, principalement sur les cours d'eau. Les barrages sont les ouvrages les plus abondants, avec environ 40 000 barrages de plus de 15m de haut ainsi que 800 000 plus petits, ils permettent de retenir $10\,000 \text{ km}^3$ d'eau soit environ 5 fois le volume d'eau contenu dans les rivières du monde (Dudgeon *et al.* 2006; Zarfl *et al.* 2015). De tels ouvrages sont une source de dégradation des habitats en amont (*e.g.* inondations, apparition de nouvelles zones ripariennes), en aval (*e.g.* variation des débits, de la profondeur et de la largeur de la rivière) et sont responsables de la fragmentation physique des cours d'eau (Kikyo *et al.* 1999; Nilsson 2005). Ceci affecte directement les communautés aquatiques et peut se traduire par une perte directe de diversité, l'apparition d'espèces mieux adaptées aux nouvelles conditions (dont des

espèces exotiques envahissantes) ou encore une diminution de la capacité de reproduction de certaines espèces (Carlisle *et al.* 2011; Lehner *et al.* 2011).

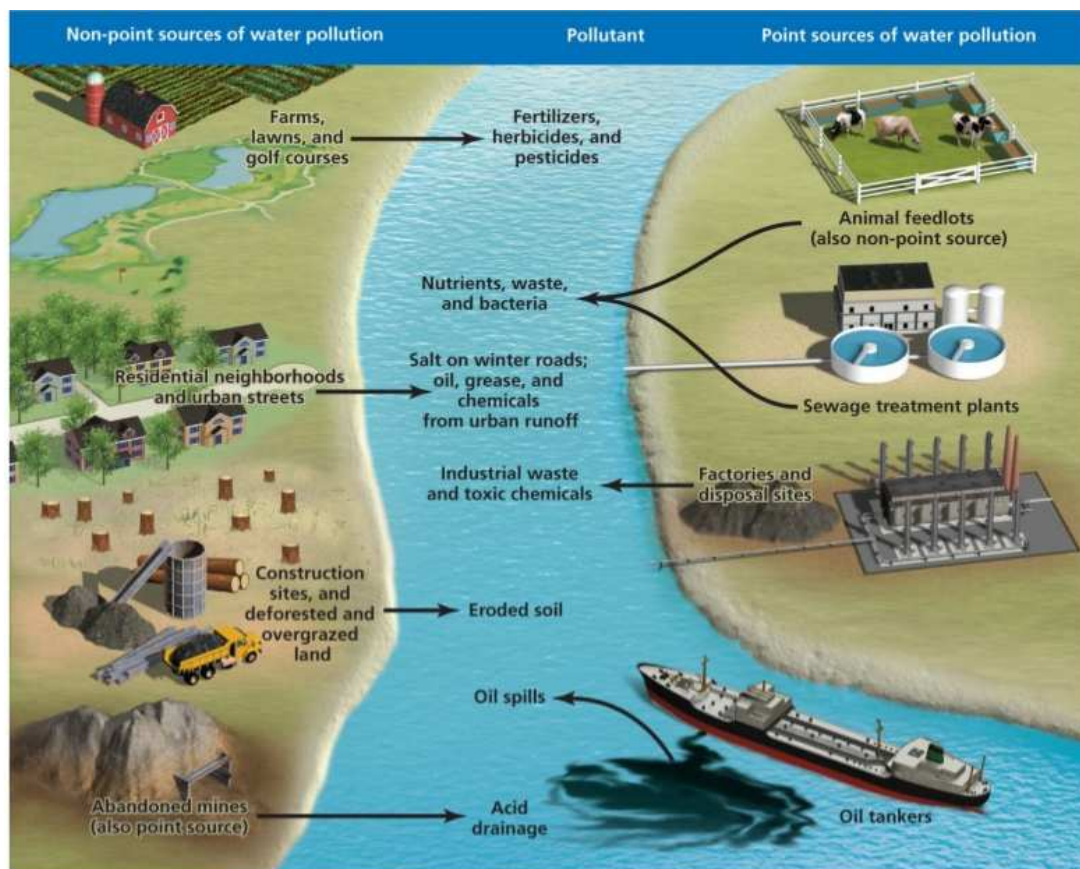


Figure 3 – Principales sources de pollutions des eaux douces par les activités humaines.
(source : Withgott and Brennan 2007)

- **Pollution de l'eau** : Les activités humaines liées à l'industrie, l'agriculture et aux usages domestiques sont les principales sources de pollution des milieux aquatiques (**Figure 3**). Ces pollutions peuvent se regrouper en différentes catégories incluant les pollutions par les composés organiques (*e.g.* pesticides, détergents, solvants) et inorganiques (*e.g.* métaux lourds), les nutriments, les matières en suspension, l'apport d'organismes pathogènes (bactéries, virus) ou encore des perturbations physiques du milieu (Voulvoulis & Georges 2016). La grande diversité des stress apportés par ces pollutions chimiques et physiques aboutit à des effets très variés sur les écosystèmes aquatiques. Certains composés comme les pesticides ont une toxicité directe et peuvent bloquer la photosynthèse ou inhiber le système nerveux des organismes aquatiques (DeLorenzo *et al.* 2001). Les composés comme les métaux lourds et les polluants organiques persistants (*e.g.* les PCB) sont très peu dégradés dans l'environnement et souvent bioaccumulés par les organismes, affectant potentiellement la « santé » de toute la chaîne alimentaire, y compris l'homme (Ratte 1999; Geyer *et al.* 2000).

L'apport de nutriments, principalement d'azote et de phosphore en lien avec l'agriculture et les rejets domestiques, a pour effet de favoriser le développement d'algues et de « végétaux supérieurs » aboutissant à l'eutrophisation des milieux aquatiques. Il en résulte une diminution de la teneur en oxygène de l'eau, rendant celle-ci toxique pour certains organismes aquatiques mais aussi potentiellement impropre à la consommation, la baignade, l'irrigation ou encore la pêche (Carpenter *et al.* 1998; Dodds & Smith 2016).

- **Apparition d'espèces exotiques envahissantes** : On considère comme espèce exotique envahissante « une espèce allochtone (non indigène) dont l'introduction (volontaire ou fortuite) par l'Homme, l'implantation et la propagation menacent les écosystèmes, les habitats ou les espèces indigènes avec des conséquences négatives sur les services écosystémiques et/ou socio-économiques et/ou sanitaires» (Sarat *et al.* 2015). Ces espèces sont généralement introduites volontairement (*e.g.* commerce, aquaculture) ou involontairement (*e.g.* transport, tourisme) par l'homme dans l'environnement (Hulme 2009). Si elles n'ont pas toutes un effet négatif, elles n'en restent pas moins la 2^{ème} source de perte de biodiversité dans les écosystèmes d'eau douce (García-Berthou *et al.* 2005). En effet, ces espèces peuvent avoir un impact direct sur les communautés autochtones (*e.g.* compétition pour les ressources, prédation, vecteurs de pathogènes) ou indirect en modifiant leur habitat (*e.g.* altération physique, augmentation de la turbidité) (Crooks 2002).

A tous ces facteurs locaux viennent s'ajouter des facteurs environnementaux globaux, comme le changement climatique, qui complexifient davantage la compréhension des impacts sur les écosystèmes d'eau douce (Ormerod *et al.* 2010; Woznicki *et al.* 2016). Ainsi, l'accumulation de stress peut aboutir à des effets croisés de type antagonistes (stress AB << stress A + stress B), synergiques (stress AB >> stress A + stress B), additifs (stress AB = stress A + stress B) ou inversés (stress AB ≠ stress A + stress B) sur les écosystèmes d'eau douce, les effets antagonistes étant prépondérant selon Jackson *et al.* (2016). Par exemple, le réchauffement climatique permet de lutter contre les effets délétères des rayons ultraviolet (UV-B) chez les Daphnies en induisant un mécanisme de réparation photo-enzymatique (MacFadyen *et al.* 2004). Dans certains lacs, l'effet croisé du réchauffement climatique avec le phénomène d'acidification est supérieur à la somme de leurs effets individuels, aboutissant à une augmentation de la biomasse du zooplancton (effet synergique) tandis que le réchauffement semble amoindrir l'effet positif de l'acidification sur la croissance du phytoplancton (effet antagoniste) (Christensen *et al.* 2006).

Bien qu'il soit important de comprendre comment tous ces stress peuvent interagir entre eux et impacter les écosystèmes d'eau douce afin de mieux les protéger (Hering *et al.* 2015),

nos connaissances sur le sujet restent encore limitées et des politiques de gestion et de protection de ces écosystèmes ont d'ores et déjà été mises en place (e.g. Managing Aquatic ecosystems and water Resources under multiple Stress (MARS) project, <http://www.mars-project.eu/>).

1.2. Evolution des politiques de gestion de l'eau

1.2.1. Politiques de gestion en France

La mise en place d'une réelle politique de gestion de l'eau en France est assez récente et remonte aux années 1950-60 en lien avec la révolution industrielle et l'augmentation des pollutions des ressources en eau. La première loi sur l'eau du 16 décembre 1964, a mis en place une gestion par bassin hydrographique en divisant le territoire en 6 grands bassins, créant pour chacun une agence de l'eau et un comité de bassin (Narcy 2003). Cette action favorisait une gestion collective des ressources en eau à l'échelle des bassins, les ressources financières étant issues des redevances perçues auprès des usagers de l'eau en fonction de leur consommation d'eau et du niveau de pollution émis dans le milieu (principe de « pollueur-payeur ») (Feuillette 2004). Cette loi est l'une des premières actions de l'état en matière de protection de l'environnement, le ministère de l'environnement n'étant créé qu'en 1971. A partir des années 1990, les politiques de gestion de l'eau vont s'orienter de plus en plus vers des objectifs en lien avec le développement durable, suivant la « tendance » créée lors du « Sommet de la Terre » de Rio de Janeiro en 1992. Ainsi, la loi sur l'eau du 3 Janvier 1992 définit l'eau en tant que "patrimoine commun de la Nation" avec pour objectif la préservation des ressources en eau et des écosystèmes aquatiques (Girard 2012). Afin d'y parvenir, la loi propose la mise en place de nouveaux outils de gestion des eaux à l'échelle des bassins : les Schémas Directeurs d'Aménagement et de Gestion des Eaux (SDAGE).

Le 23 Octobre 2000, l'Union Européenne adopte la « Directive Cadre sur l'Eau » (DCE) qui propose une politique communautaire de gestion de l'eau à l'échelle de l'Europe. Elle est transposée en droit français le 21 avril 2004 et servira de patron pour la mise en place de la 3^{ème} Loi sur l'Eau et les Milieux Aquatiques (LEMA) le 31 Décembre 2006. De fait, la gestion de l'eau en France est actuellement régie par les législations française et européenne. Bien que la LEMA propose d'améliorer les conditions d'accès à l'eau, une meilleure transparence du fonctionnement du service public de l'eau (via la création de l'Office National de l'Eau et des Milieux Aquatiques – ONEMA qui fait dorénavant partie intégrante de l'Agence Française pour la Biodiversité – AFB) et

une meilleure organisation de la pêche en eau douce, l'objectif majeur est de veiller à l'atteinte des objectifs fixés par la DCE, à savoir le bon état des masses d'eau à l'échelle européenne.

1.2.2. La Directive Cadre sur l'Eau (DCE)

Les ressources en eau ainsi que les pollutions ne se bornant pas aux frontières entre états, la DCE a pour vocation d'harmoniser et de simplifier les politiques de gestion de l'eau entre les pays de la communauté européenne (Chave 2001). L'objectif principal étant d'atteindre un « bon état » des eaux, et ce avec obligation de résultats sous peine de sanctions financières. Pour ce faire, la DCE reprend le principe de gestion de l'eau par bassin versant, déjà initié en France en 1964, mais cette fois-ci à l'échelle de l'Europe avec 110 districts hydrographiques (Kallis 2001). En France, il existe 12 bassins hydrographiques dont 7 bassins métropolitains (Adour-Garonne, Artois-Picardie, Loire-Bretagne, Rhin-Meuse, Rhône-Méditerranée, Corse, Seine-Normandie) et 5 bassins d'outre-mer correspondant aux DOM (Guadeloupe, Guyane, Martinique, La Réunion et Mayotte). L'application de la DCE repose sur la mise en place de cycles de gestion sur chacun des bassins, chaque cycle comprenant 4 phases : (i) un état des lieux des ressources en eau ; (ii) la préparation des SDAGE qui indiquent les objectifs de gestion à mettre en œuvre (e.g. aménagements) ainsi que les objectifs en terme de qualité et de quantité d'eau ; (iii) les mesures à mettre en place pour atteindre ces objectifs ; (iv) leur application. Un calendrier d'application de la DCE a été mis en place en France comprenant 3 cycles de gestion successifs pour lesquels l'atteinte des objectifs fixés par les SDAGE sera évaluée à la fin de chaque cycle en 2015, 2021 et 2027.

Afin de simplifier les méthodes d'évaluation de la qualité des eaux, la DCE classe les milieux aquatiques par masses d'eau en différenciant les eaux souterraines des eaux de surface (cours d'eau, lacs, eaux de transition et eaux côtières). En France, des sites représentatifs des différentes masses d'eau ont été sélectionnés afin de mettre en place des réseaux de surveillance à l'échelle du territoire (Petit & Michon 2013) (**Figure 4**):

- Réseau de Contrôle de Référence (RCR) : composé de sites dits de référence, c'est-à-dire de sites représentatifs des différentes masses d'eau et faiblement impactés par les pressions anthropiques, ils servent à définir les limites du « très bon état ». Pour les cours d'eau, 393 sites ont ainsi été sélectionnés et sont régulièrement contrôlés (Mengin *et al.* 2010).

- Réseau de Contrôle et de Surveillance (RCS) : inclus des sites représentatifs de l'état général des eaux en France (dont certains sites du RCR) avec 2043 sites pour les eaux de surface

(dont 1669 stations sur les cours d'eau) et 1940 sites pour les eaux souterraines. Ces sites sont surveillés en permanence avec une évaluation annuelle de la qualité.

- Réseau de Contrôle et Opérationnel (RCO) : cible des sites nécessitant une attention particulière car susceptibles de ne pas atteindre « le bon état » requis par la DCE. Le RCO comprend 4 618 sites pour les eaux de surfaces (certains sites pouvant aussi faire partie du RCS) et 1446 sites pour les eaux souterraines (Petit & Michon 2015). Leur suivi peut être suspendu si la masse d'eau atteint le « bon état ».

Afin d'évaluer la qualité de l'eau sur chaque site, plusieurs paramètres sont contrôlés en fonction des masses d'eau. Le « bon état » des eaux souterraines est évalué en fonction de leur état chimique et quantitatif, tandis que les eaux de surface sont évaluées en fonction de leur état chimique et écologique.

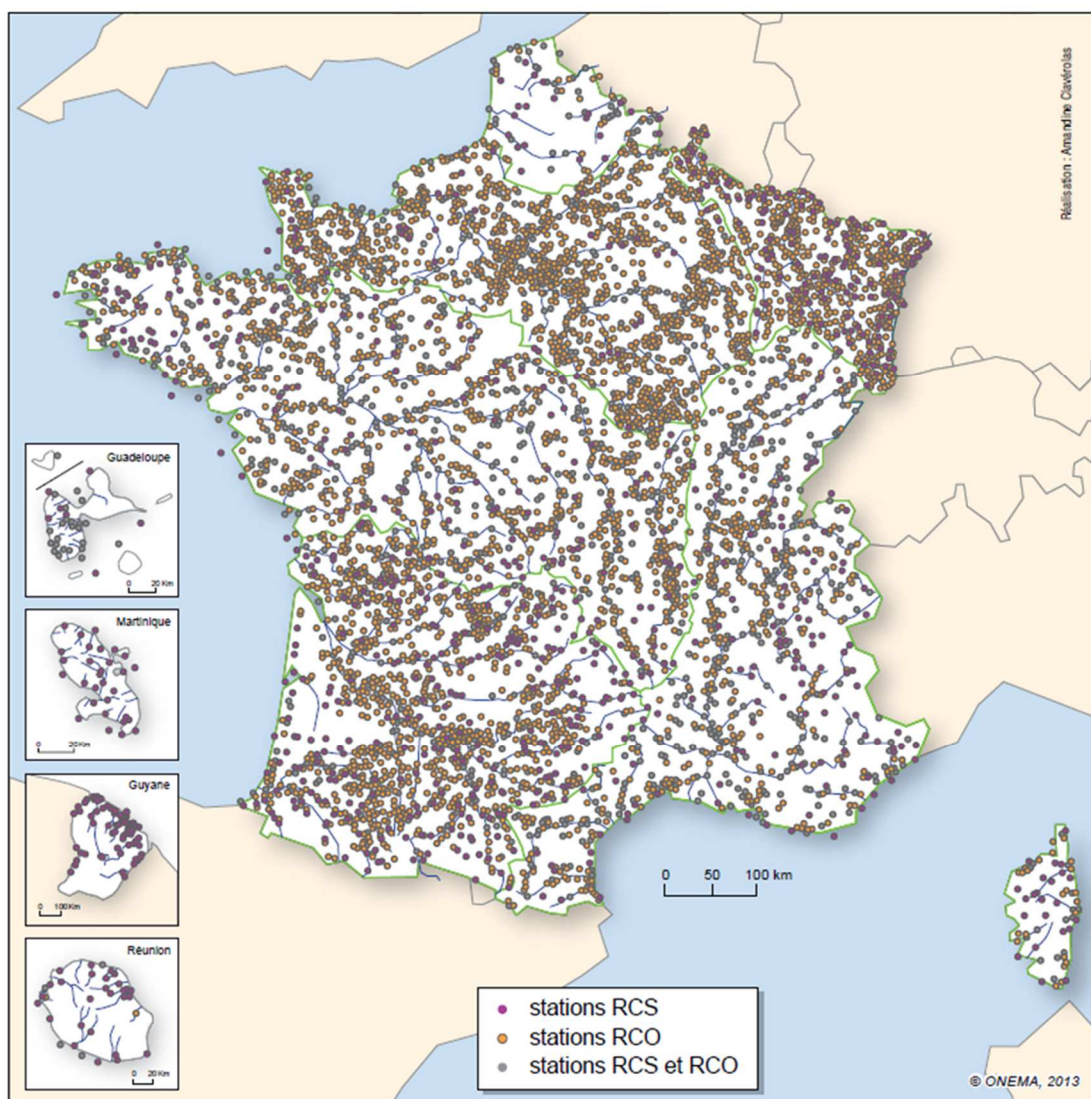


Figure 4 – Répartition des stations des réseaux de contrôle de surveillance (RCS) et de contrôle opérationnel (RCO) pour les eaux de surface.
(source : Petit & Michon 2013)

1.3. Evaluation de la qualité des cours d'eau

1.3.1. Etat chimique et écologique

L'état général de la qualité d'une station de prélèvement sur un cours d'eau est évalué en fonction de son état chimique (2 classes de qualité) et de son état écologique (5 classes de qualités), suivant les recommandations de la DCE traduites en droit français et résumées dans un guide (Ministère de l'Environnement de l'Énergie et de la Mer 2016) (**Figure 5**).

- **Etat chimique** : Pour attribuer l'état chimique d'une station, 45 paramètres correspondant à des substances ou groupes de substances sont contrôlés. Des normes de qualité environnementales (NQE) sont définies pour chaque paramètre, correspondant à des seuils de concentrations dans l'environnement. Ainsi, on attribue à chaque paramètre l'état « bon » ou « mauvais » en fonction du respect ou non des NQE, l'état chimique de la station étant considéré comme en « mauvais état » dès qu'un seul paramètre est classé « mauvais », selon le principe dit de l'élément déclassant (« one out, all out »).

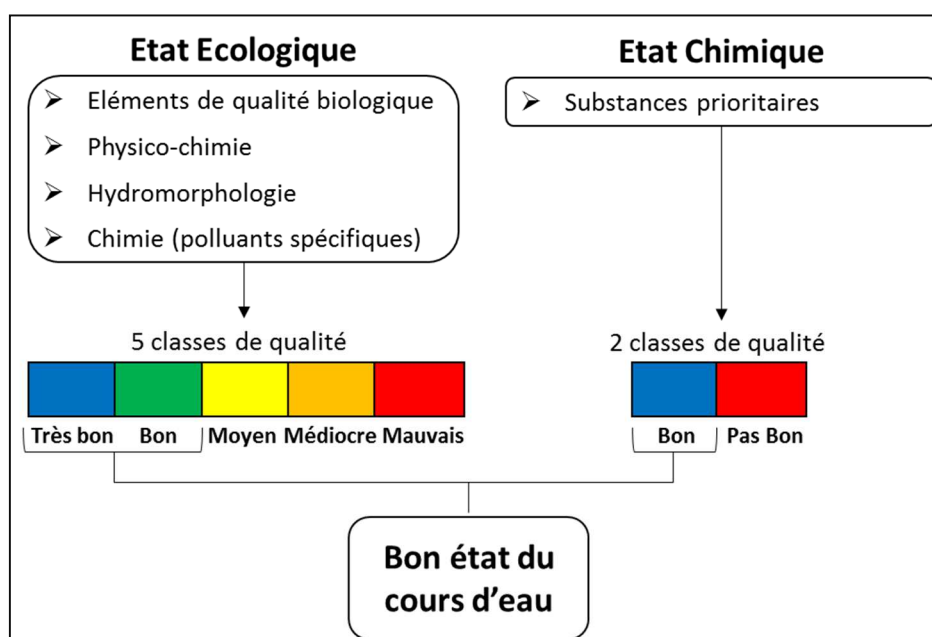


Figure 5 – Critères utilisés pour définir le bon état général des cours d'eau.

- **Etat écologique** : L'attribution de l'état écologique d'une station repose sur plusieurs éléments de qualité : biologiques (espèces végétales et animales), physico-chimiques (température, oxygène, salinité, pH, concentration en nutriments), hydromorphologiques (débit d'eau, substrat du lit, structure de la rivière) et chimiques (polluants spécifiques de l'état écologique). Des indicateurs basés sur des indices biologiques, des valeurs seuils (pour les paramètres

physico-chimiques) ou des NQE (polluants) sont mis en place pour définir l'état de chaque élément de qualité (**Annexe 1**).

1.3.2. Evaluation de la qualité biologique

Les organismes aquatiques étant en contact direct avec le milieu, ils sont directement impactés par des altérations ou modifications des conditions de celui-ci et reflètent donc son état de santé général, ce qui en fait d'excellents indicateurs biologiques de la qualité du site où ils se trouvent (Blandin 1986; Xu *et al.* 1999). L'évaluation de la qualité des cours d'eau par la bioindication est imposée par la DCE et repose sur la comparaison des communautés aquatiques rencontrées sur un site donné par rapport à celles rencontrées sur des sites de référence censées représenter le « très bon état » écologique (Bailey *et al.* 1998). Ces derniers sont choisis pour représenter au mieux l'état potentiel du site testé en l'absence de pressions anthropiques. Différentes métriques biologiques sont utilisées pour calculer des indices qui permettent d'évaluer l'écart de qualité entre le site testé et les sites de référence, permettant un classement dans l'une des 5 classes de qualité: très bon, bon, moyen, médiocre et mauvais (Birk *et al.* 2012; Kelly 2013) (**Annexe 2**). Concrètement, la classe de qualité du site testé est évaluée sur la base de « ratios de qualité écologique » (« Ecological Quality Ratios » ; EQR) correspondant au rapport entre la valeur d'indice du site testé et la valeur d'indice obtenu pour les sites de référence. Pour caractériser l'état de qualité biologique des cours d'eau, la DCE impose :

- l'utilisation de 4 ou 5 éléments de qualité biologique, correspondant à 5 groupes d'organismes aquatiques : poissons, macrophytes, invertébrés benthiques, phytobenthos (principalement les diatomées benthiques) et phytoplancton (seulement pour les grands cours d'eau) (Reyjol *et al.* 2013).
- que les indices de qualité créés pour ces 5 éléments de qualité biologique prennent en compte l'abondance et la composition des espèces retrouvées (ainsi que la structure de l'âge pour les poissons). Cependant la DCE n'impose ni l'utilisation d'indice de qualité en particulier, ni les valeurs seuils bornant les 5 classes de qualité, chaque état membre étant libre de développer et d'utiliser ses propres indices (Hering *et al.* 2010), dans les limites fixées par l'exercice d'intercalibration de l'Union Européenne (voir section 1.4.1 de l'annexe V de la DCE).

Les 5 groupes utilisés comme éléments de qualité biologique sont censés réagir différemment aux variations des conditions environnementales (*e.g.* physico-chimie, pollution) et sur des échelles de temps plus ou moins longues (Barbour *et al.* 1999) (**Figure 6**). Leur utilisation

conjointe permet ainsi de mieux caractériser l'état du milieu, et ce à différents niveaux de la chaîne trophique. Cependant, certains organismes comme les diatomées, qui représentent une part importante du phytoplancton et du phytobenthos, sont des bioindicateurs très étudiés car présentant de nombreux avantages pour évaluer la qualité des cours d'eau.

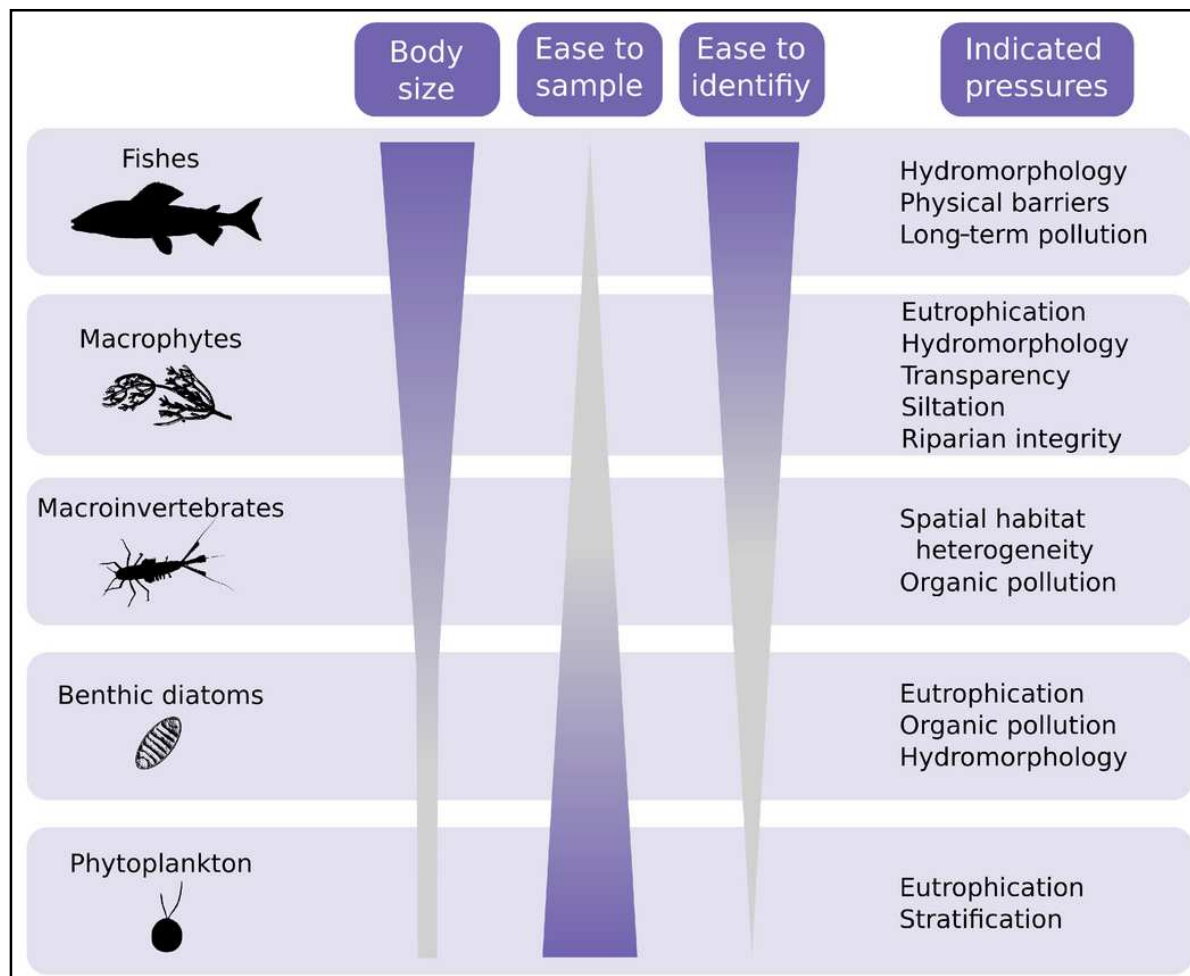


Figure 6 – Caractéristiques des 5 éléments de qualité biologique.
(source : Keck *et al.* 2017)

2. Les diatomées comme bioindicateurs

Les diatomées sont des microorganismes eucaryotes, unicellulaires et microscopiques qui sont apparus il y a environ 250 millions d'années (le plus vieux fossile de diatomée retrouvé datant de 190 millions d'années) (Sims *et al.* 2006; Sorhannus 2007). Elles font partie de l'infrarègne des *Heterokonta* et appartiennent plus précisément au phylum des *Bacillariophyta*, formant un groupe très diversifié avec environ 12 000 espèces décrites et potentiellement 100 000 espèces existantes (Guiry 2012; Mann & Vanormelingen 2013). Ces algues photosynthétiques sont retrouvées dans tous les écosystèmes aquatiques, marin et d'eau douce, et sont tellement abondantes qu'elles produisent environ 20 % de notre oxygène et réalisent 40 à 45 % de la production primaire des océans (Field *et al.* 1998; Mann 1999). Les premières représentations de diatomées obtenues grâce à la microscopie remontent à 1703 (Round *et al.* 1990a) et les recherches menées jusqu'au début du 20^{ème} siècle ont principalement permis « d'explorer » et de « systématiser » les connaissances sur les diatomées. Depuis les recherches ont permis de montrer que les diatomées sont d'excellents indicateurs qui permettent d'explorer et d'interpréter de nombreux problèmes écologiques mais aussi pratiques, dont l'évaluation de la qualité des cours d'eau (Smol & Stoermer 2010).

2.1. Biologie des diatomées

2.1.1. Principaux composants de la cellule

Comme tous les organismes eucaryotes, les cellules des diatomées contiennent différents organites nécessaires à leur bon fonctionnement : un noyau généralement positionné au centre de la cellule, des mitochondries, un complexe formé de plusieurs corps de Golgi, une vacuole qui occupe la majeure partie de la cellule ainsi que des gouttelettes lipidiques (Bedoshvili & Likhoshway 2012) (**Figure 7**). Les cellules des diatomées présentent 2 caractéristiques notables qui auront leur importance dans ces travaux de thèse :

- **Le frustule** : la cellule des diatomées est entourée par une paroi composée de silice (SiO_2) produite par la cellule et issue de la polycondensation du $\text{Si}(\text{OH})_4$ présent naturellement dans le milieu aquatique (Kröger & Poulsen 2008). Cette paroi forme une boîte protectrice en forme de « boîte de pétri » appelé le frustule. Les valves, qui correspondent aux deux couvercles de la boîte, sont entourées et connectées ensemble par une ou plusieurs ceintures connectives qui forment le cingulum (**Figure 8**). L'hypothèque est composée de la plus petite

des deux valves et de l'hypocingulum, tandis que l'épithèque inclut la plus grande valve et l'épicingulum (Cox 2014). Les valves peuvent avoir des stries, des pores ou encore un raphé (absent chez les diatomées centriques), donnant une ornementation caractéristique à chaque espèce et permettant leur classification sur la base de leur morphologie. Les diatomées sont ainsi groupées en 2 grandes classes : les centriques et les pennées (**Figure 8**).

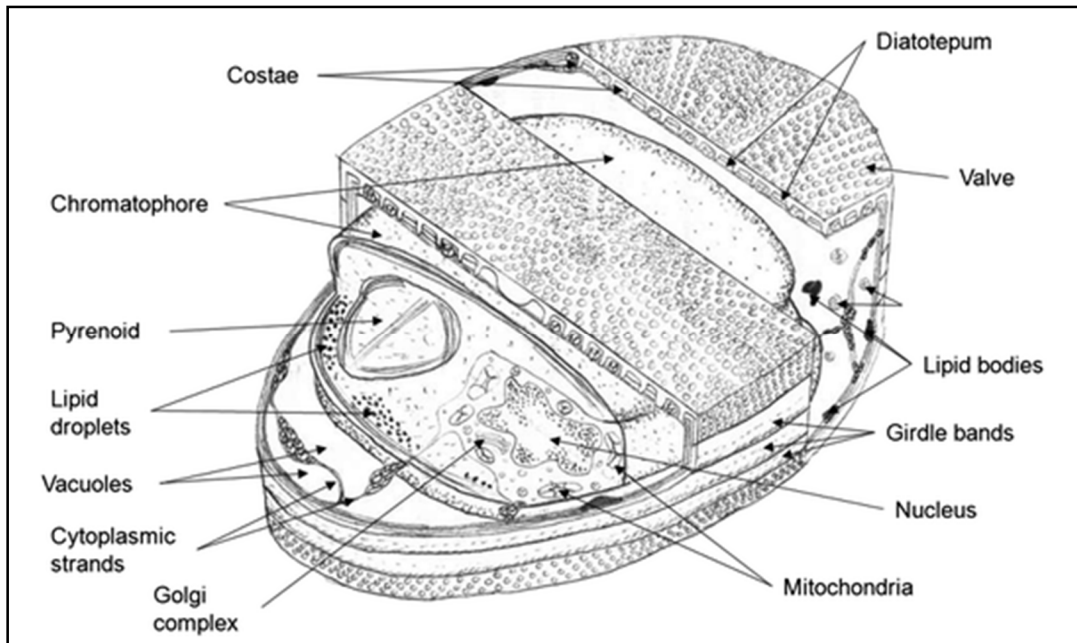


Figure 7 – Ultrastructure d'une cellule de diatomée avec ses principaux composants. (source : Yang *et al.* 2011)

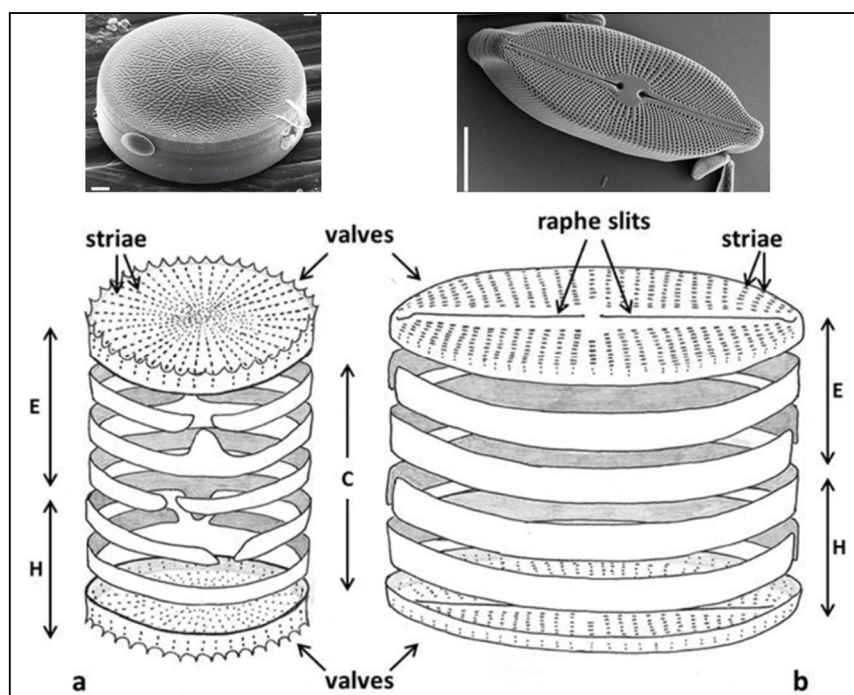


Figure 8 – Structure du frustule chez les diatomées centriques (a) et pennées (b). E = epitheca, H = hypotheca, C = cingulum. (adapté de Cox 2014, Mann *et al.* 2016).

- **Le chloroplaste** : la forme, la position et le nombre de chloroplastes dans la cellule varient d'une espèce à l'autre (Bedoshvili *et al.* 2009). Grossièrement, les diatomées pennées sont caractérisées par la présence de 1 à 8 chloroplastes plutôt de grande taille, tandis que les diatomées centriques sont le plus souvent caractérisées par la présence de plusieurs dizaines de chloroplastes de petite taille (Round *et al.* 1990a). Le chloroplaste a la particularité d'avoir 4 membranes, résultat de l'histoire évolutive des diatomées caractérisée par 2 endosymbioses successives : une endosymbiose primaire entre une cyanobactérie et une cellule eucaryote à l'origine de la création du chloroplaste retrouvé par exemple chez les plantes terrestres, les algues vertes et rouges ; une endosymbiose secondaire entre une algue rouge et une cellule eucaryote hétérotrophe hôte (Green 2011; Keeling 2013). Le reticulum endoplasmique n'est pas présent dans la cellule des diatomées car il a été intégré dans les membranes du chloroplaste (Gibbs 1981). Le chloroplaste contient de la chlorophylle *a* et *c* ainsi que d'autres pigments caroténoïdes, notamment la fucoxanthine, donnant leur couleur brune aux diatomées. Les chloroplastes possèdent leur propre matériel génétique constitué de plusieurs copies du génome circulaire (≈ 120 kbp) agrégées sous la forme d'un « anneau nucléoïde » (Coleman 1985; Brembu *et al.* 2014).

2.1.2. Cycle de vie et reproduction

Durant leur cycle de vie, les diatomées se multiplient la plupart du temps de manière végétative par mitoses successives, produisant à partir d'une cellule mère unique, deux cellules filles identiques d'un point de vue génétique. Ce style de division est partagé par tous les organismes eucaryotes (animaux, végétaux, levures) avec les mêmes étapes de division mitotique, bien que les mécanismes moléculaires de chaque modèle de cellule varient (De Martino *et al.* 2009). Durant la mitose chez les diatomées, chaque cellule fille hérite d'une des deux valves de silice de la cellule mère (**Figure 9**). Une fois la cytodierèse effectuée, chaque cellule fille va produire des vésicules de silice qui serviront à construire la deuxième valve manquante. Afin que le frustule définitif de la cellule fille conserve sa forme de « boîte de pétri », la nouvelle valve synthétisée est plus petite que la valve héritée. A cause de la différence de taille des valves de la cellule mère, on obtient alors deux cellules filles de tailles différentes. Ce phénomène, qui a pour effet d'entraîner une diminution progressive de la taille des cellules au fil des générations, suit le modèle décrit par Macdonald (1869) et Pfitzer (1869) (**Figure 9**).

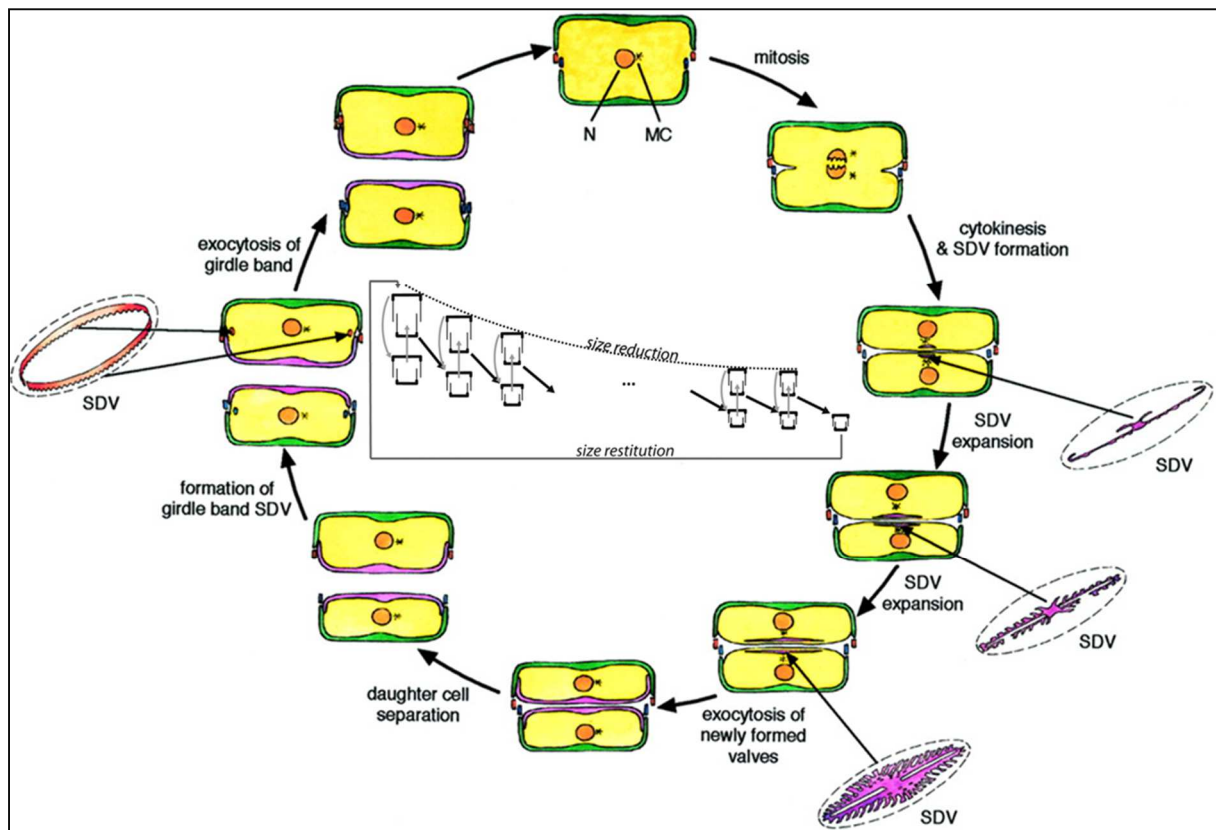


Figure 9 – Multiplication végétative induisant une réduction de la taille des cellules. N = Nucleus; MC = microtubule center; SDV = silica deposition vesicle (adapté de Zurzolo & Bowler 2001, Hense & Beckmann 2015)

Les diatomées sont les seules algues qui ont besoin de passer par la reproduction sexuée pour restaurer leur taille initiale, ce mode de reproduction étant généralement un facteur de dispersion ou de dormance (Edlund & Stoermer 1997). La reproduction sexuée ne peut être initiée que lorsque certaines conditions sont remplies. Premièrement, les cellules végétatives doivent avoir atteint une taille minimale de l'ordre de 30 à 40 % de la taille maximale de la cellule (Lewis 1983). Deuxièmement, les conditions environnementales doivent être favorables, par exemple en termes de température, de concentration en nutriments ou encore de luminosité (Edlund & Stoermer 1997). D'autres mécanismes complexes, comme la production de phéromones, semblent jouer un rôle important dans l'induction de la reproduction sexuée et permettent d'augmenter les chances de rencontre entre les partenaires (Moeys *et al.* 2016). De ce fait la reproduction sexuée est un phénomène plutôt rare dans l'environnement, les cycles de reproduction pouvant survenir après plusieurs années voire dizaines d'années de cycles végétatifs (Mann 1988). Il existe différents modes de reproduction en fonction des groupes de diatomées : principalement l'oogamie chez les centriques (gamète male petit et mobile via un flagelle / gamète femelle gros et immobile) et l'isogamie chez les pennées (gamètes mâles et femelles de grande taille et immobiles, difficiles à différencier) (Round *et al.* 1990a; Chepurnov *et al.* 2004). Quel que

soit le mode de reproduction sexuée, la fécondation des deux gamètes aboutit à la production de « l'auxospore » (**Figure 10**), une cellule qui possède une paroi de silice plus fine, permettant de restaurer la taille initiale et maximale de la cellule (Medlin & Kaczmarek 2004).

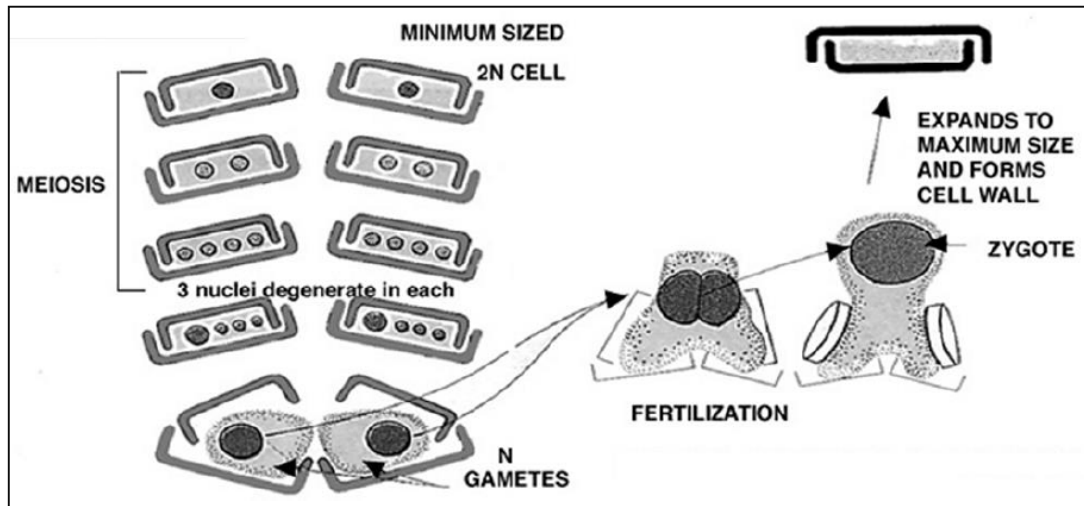


Figure 10 – Reproduction sexuée et formation de l'auxospore.
(source : Weir 1982)

2.1.3. Classification des diatomées

La classification des diatomées n'a eu de cesse d'évoluer depuis des siècles, le point de départ pouvant être attribué à Carl Adolf Agardh vers 1830-1832 (Williams & Kociolek 2007). A cette époque, les diatomées étaient classées en fonction de la forme et de la symétrie générales de leur frustule. Il faut attendre 1927, avec l'évolution des outils de microscopie, pour voir apparaître le développement d'un système de classification plus robuste proposé par Karsten (Karsten 1928). Celui-ci se base sur les motifs retrouvés sur les valves du frustule pour différencier, au sein du phylum *Bacillariophyta*, deux ordres : les Centrales et les Pennales. Les Centrales regroupent les diatomées centriques qui ont une symétrie radiale, tandis que les Pennales regroupent les diatomées dites pennées qui possèdent une symétrie bilatérale en forme de « plume » (**Figure 8**). Actuellement, c'est la classification proposée par Round *et al.* (1990) qui est la plus utilisée et comporte 3 classes : *Coscinodiscophyceae* (centriques), *Fragilariophyceae* (pennées sans raphé) et *Bacillariophyceae* (pennées avec raphé). Bien que cette classification évite au mieux les groupes polyphylétiques, de nombreux groupes paraphylétiques subsistent et empêchent d'avoir une bonne représentation de l'évolution des diatomées (Mann *et al.* 2016). Depuis, de nouvelles classifications ont été proposées prenant en compte des paramètres cytologiques (*e.g.* structure des pyrénoides dans les chloroplastes, paroi de l'auxospore) ou

utilisant la phylogénie des diatomées (Medlin & Kaczmarska 2004; Theriot *et al.* 2010). Encore une fois, ces classifications ne résolvent pas tous les problèmes de classification et sont sources de « vifs » débats entre diatomistes (Williams & Kociolek 2007; Medlin 2010). Récemment, Mann *et al.* (2016) se sont basés sur tous ces paramètres (morphologique, cytologique, phylogénétique) pour proposer un « compromis » en définissant de nouvelles subdivisions et classes au sein des *Bacillariophyta* (les modifications majeures concernant les diatomées centriques) tout en gardant les genres décrits par Round *et al.* (1990) (**Annexe 3**).

2.2. Ecologie générale et réponses aux pressions environnementales

2.2.1. Formes de vie et habitats

Les cellules de diatomées peuvent être mobiles, attachées ou suspendues, ce qui leur permet de coloniser une grande variété d'habitats dans les milieux aquatiques. On peut ainsi différencier les cellules planctoniques libres dans le milieu des cellules benthiques qui vivent à la surface de différents supports.

- **diatomées planctoniques** : étant des organismes photosynthétiques, les diatomées ont la possibilité de se développer dans la colonne d'eau des milieux aquatiques (*e.g.* rivières, lacs, océans) tant que la quantité de lumière disponible est suffisante. A cause de leur frustule de silice, les diatomées sont de nature plus lourdes que l'eau, ce qui fait qu'elles ont tendance à couler naturellement et à former des flocs encore appelés « neige » (Lund 1954). Comme les espèces planctoniques ne sont pas mobiles, elles ne peuvent rester en suspension que grâce aux mouvements d'eau créés par les courants et le vent. Ce mécanisme serait une stratégie adaptative qui apporterait de nombreux avantages : faciliter l'incorporation de nutriments, fuir les zones pauvres en silice ou encore éliminer les cellules les moins compétitives (Smetacek 1985; Sandgren 1988). Bien qu'étant des organismes unicellulaires, les diatomées planctoniques présentent différentes formes de vie, en cellule seule ou en colonies ayant des formes variées : étoile/zigzag, ruban ou filament (**Figure 11 a,b,c,d**). Ces formes varient d'une espèce à l'autre et en fonction des conditions de l'environnement ainsi que des interactions avec les autres espèces. Elles permettent par exemple de lutter contre la prédation ou d'apporter une meilleure flottabilité et ainsi éviter la sédimentation (Rimet & Bouchez 2012b; Nakov *et al.* 2015).

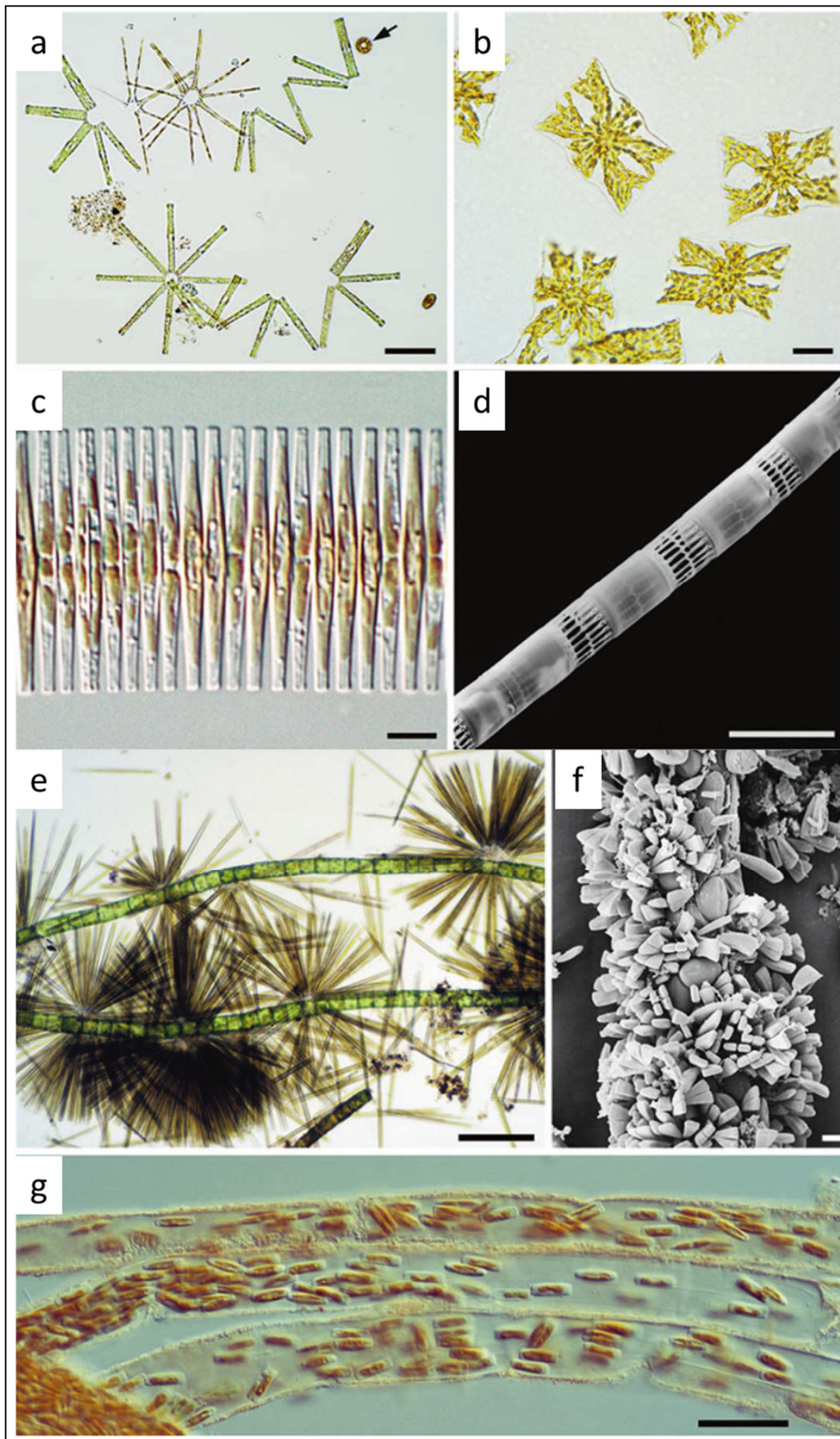


Figure 11 – Exemples de formes de vie observées chez les diatomées planctoniques (a,b,c,d) et benthiques (e,f,g).
 (adapté de Mann *et al.* 2016)

- **diatomées benthiques** : les diatomées benthiques peuvent être mobiles ou fixées sur toute sorte de support dans les milieux aquatiques. Seules les diatomées pennées ont la possibilité de se déplacer « en glissant » grâce au mucilage libéré au niveau du raphé (Edgar & Pickett-heaps 1983). La mobilité peut être induite lors de carence en silice (Bondoc *et al.* 2016) ou en fonction des conditions de luminosité (McLachlan *et al.* 2009). Avec d'autres organismes (*e.g.* bactéries, champignons, algues, protozoaires) elles se développent au sein d'un biofilm aquatique et forme le périphyton qui peut être catégorisé en fonction du support colonisé (Mann *et al.* 2016) : épilithon (substrats durs et inertes), épiphyton (végétaux aquatiques), épipélon / endopélon (à l'extérieur / intérieur du sédiment), épipsammon (grains de sable), métaphyton (mucilage produit par d'autres algues), épyzoon (macroorganismes comme les oiseaux ou encore les crustacés). Certaines espèces peuvent être retrouvées dans plusieurs habitats ou être spécifiques d'un seul type d'habitat. Les diatomées benthiques peuvent adopter des formes de vies beaucoup plus variées que les diatomées planctoniques (Round *et al.* 1990a), allant de la cellule unique, en passant par des colonies denses (**Figure 11 e,f**) et des structures plus complexes avec des colonies formant des filaments ou se développant à l'intérieur de cylindres composés de polysaccharides (**Figure 11 g**). En fonction de leur forme de vie, les diatomées seront plus ou moins sensibles aux stress présents dans l'environnement. Par exemple les formes prostrées (cellules fortement fixées sur le support) sont moins sensibles à la prédation par « broutage » mais sont limitées en terme d'accès aux ressources (Stevenson *et al.* 1996). A l'inverse, les formes filamenteuses accèdent plus facilement à la lumière et aux nutriments mais en contrepartie sont sensibles à la prédation et aux contraintes physiques de l'environnement. Les différentes formes de vie rencontrées correspondent donc à une adaptation des espèces aux contraintes de l'environnement et peuvent servir à classer les diatomées. Passy (2007a) a ainsi pu regrouper les diatomées en 3 guildes morphologiques qui répondent à des gradients de nutriments et de perturbation du courant : « low profile » (petites espèces fortement fixées ou peu mobiles), « high profile » (espèces plus grandes avec des formes de vies complexes) et « motile » (espèces très mobiles) (Rimet & Bouchez 2012b; Tapolczai *et al.* 2016).

L'évaluation de la qualité des cours d'eau dans le cadre de la DCE reposant principalement sur le phytobenthos (le phytoplancton étant utilisé pour les plans d'eau et les très grand cours d'eau), nous travaillerons principalement avec les communautés de diatomées benthiques.

2.2.2. Réponses des communautés benthiques aux pressions

Dès que des substrats sont immergés, la formation du biofilm aquatique se fait très rapidement en plusieurs étapes : (i) dépôt de matière organique sur le substrat, (ii) à l'échelle de quelques heures des bactéries se développent et produisent du mucilage pour se fixer, (iii) après quelques jours les premières diatomées pennées de petite taille (*e.g. Navicula*) colonisent le support, (iv) s'ensuit la colonisation de cellules plus grande avec des formes de vie plus complexes (Azim *et al.* 2005). Un biofilm aquatique mature peut contenir une grande diversité d'organismes comprenant aussi bien des insectes que des champignons, en passant par des ciliés et bien entendu des diatomées (**Figure 12**). On considère que la communauté de diatomées benthiques a atteint un équilibre lorsque 4-5 espèces contribuent à 80 % de la biomasse totale sur une durée de 4-5 semaines, la biomasse totale restant stable sur cette période (Lengyel *et al.* 2015). La richesse spécifique au sein des communautés de diatomées benthiques est directement liée à la taille des cellules et suit une loi unimodale, les espèces à petit biovolume sont abondantes alors que les espèces à gros biovolumes sont généralement peu abondantes (Passy 2007b). Ces dernières constituent cependant une part importante de la biomasse, parfois plus importante que celle des petites espèces (Lavoie *et al.* 2006). Cette distribution correspond à un compromis entre accès aux ressources et dispersion, les espèces les plus grosses captant mieux la lumière et les nutriments mais ayant un taux de croissance et une capacité de dispersion plus faible que les petites espèces (Passy 2007b). A cette diversité de biovolumes interspécifique vient aussi s'ajouter la variabilité intraspécifique liée à la reproduction végétative, la différence de taille entre les cellules les plus petites et les plus grandes d'une population pouvant atteindre un facteur 5 (Hense & Beckmann 2015). L'état physiologique des cellules au sein de la communauté est aussi fortement impacté par les pressions environnementales et la mortalité naturelle des diatomées. Celui-ci ne se résume plus à « vivant » ou « mort » mais correspond à un gradient d'états avec des cellules plus ou moins endommagées caractérisées par des frustules déformés (formes tératologiques), une perte d'intégrité membranaire, une dégradation des pigments photosynthétiques (réduisant la photosynthèse) ou encore une dégradation de l'ADN génomique (Veldhuis *et al.* 2001; Pandey *et al.* 2017).

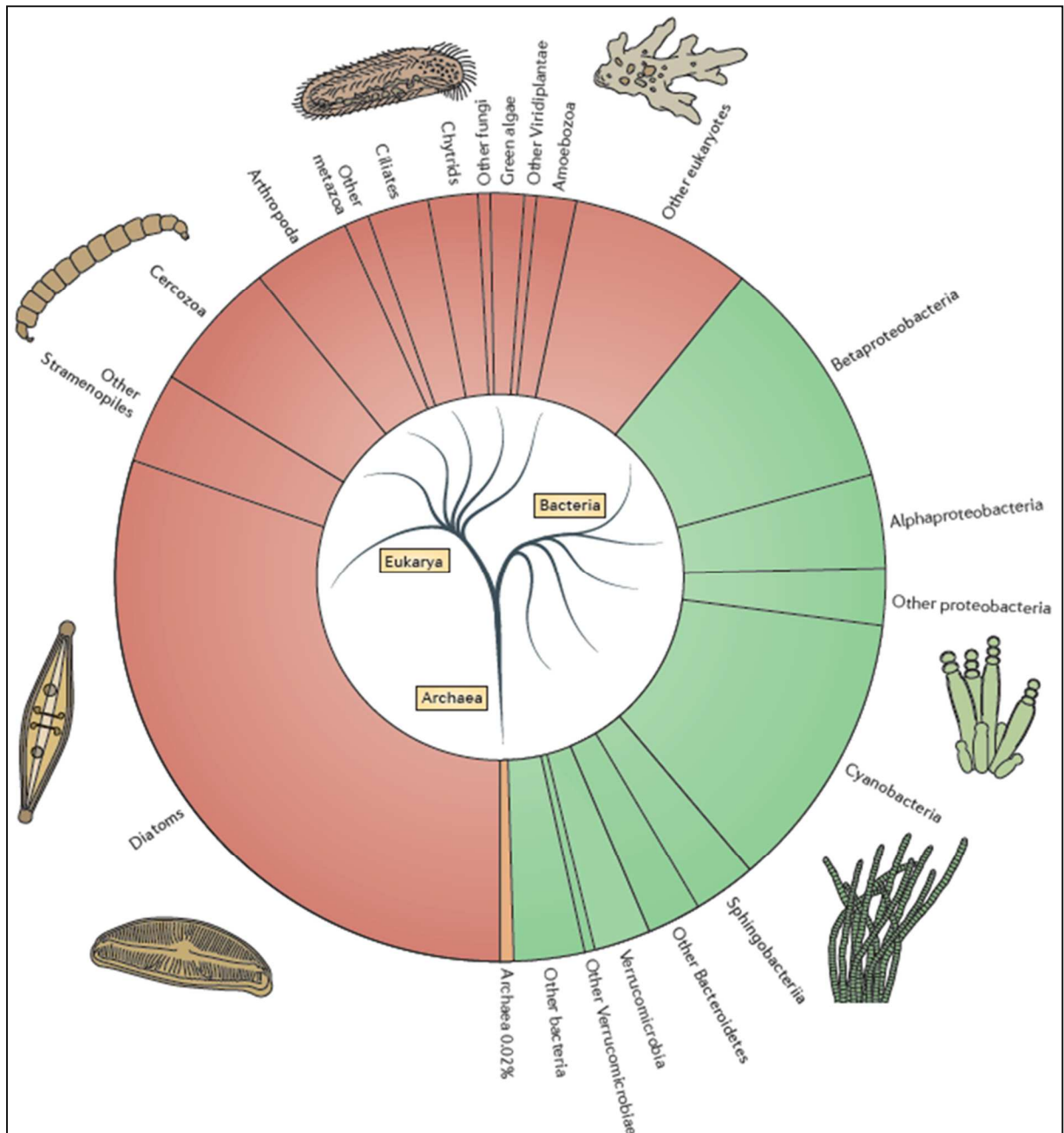


Figure 12 – Diversité biologique retrouvée au sein d'un biofilm aquatique de rivière.
(source : Battin *et al.* 2016)

Les diatomées se caractérisent par une grande diversité spécifique avec des préférences écologiques variées leur permettant de coloniser de nombreux milieux. La structure des communautés benthiques est donc déterminée par les conditions et pressions qui s'exercent sur le périphyton. Les cours d'eau étant par nature des écosystèmes très variables, de nombreux facteurs peuvent affecter les communautés benthiques : l'intensité lumineuse, la température de l'eau, le pH, les concentrations en silice et nutriments, les paramètres hydrologiques et hydromorphologiques (*e.g.* vitesse du courant, profondeur) ainsi que des facteurs biotiques (*e.g.* prédation) (Stevenson *et al.* 1996; Lengyel *et al.* 2015). Ces paramètres varient naturellement au

cours des saisons, si bien que des communautés différentes en termes d'abondance, de diversité, de biomasse totale et de formes de vie peuvent se succéder au cours de l'année (Andrus *et al.* 2013; Larras *et al.* 2014). L'apparition de nouvelles pressions qui s'exercent sur le periphyton (*e.g.* pollutions anthropiques) a pour effet de restructurer les communautés de diatomées en fonction de leurs préférences écologiques. Par exemple, certaines espèces sont très sensibles à la présence de polluants organiques et inorganiques (*e.g.* atrazine, phénols, métaux lourds, nutriments), tandis que d'autres sont plus ou moins affectées par la présence de minéraux (*e.g.* nitrates, phosphates) (Pandey *et al.* 2017). En fonction du type de pollution, la diversité et la richesse spécifique au sein de la communauté peuvent diminuer ou augmenter (Hillebrand & Sommer 2000; Izsak *et al.* 2002). On retrouve alors des assemblages spécifiques de diatomées qui sont caractéristiques du type et du niveau de pollution rencontrés. Ceci justifie l'utilisation des diatomées comme bioindicateurs de la qualité de l'eau dans le cadre de la DCE.

2.3. Evaluation de la qualité des cours d'eau via les diatomées

2.3.1. Développement des indices diatomiques

Le premier indice permettant l'évaluation de la qualité des cours d'eau en se basant sur la composition spécifique de microalgues (dont les diatomées), en prenant en compte leurs préférences écologiques ainsi que leur tolérance aux pressions, est attribué aux travaux de Kolkwitz & Marsson (1908). Ils montrent alors que les conditions du milieu, notamment la matière organique, déterminent la structure des communautés algales. Ils ont ainsi pu créer un indice saprobique qui permet de classer les cours d'eau en 4 classes de qualité en fonction du degré de pollution par la matière organique. Dès lors de nombreux indices basés sur les diatomées se sont développés (*e.g.* Butcher 1947; Fjerdingstad 1950; Zelinka & Marvan 1961) et on retrouve 3 catégories d'indice : les indices autécologiques, les indices multimétriques et les indices basés sur des modèles prédictifs. Les indices autécologiques, comme celui développé par Kolkwitz & Marsson en 1908, reposent généralement sur l'abondance relative des taxons présents et leurs préférences écologiques vis-à-vis de la pollution étudiée. Les indices multimétriques réunissent en un seul indice plusieurs métriques caractéristiques de la communauté qui répondent aux pressions de l'environnement (*e.g.* richesse spécifique, traits écologiques, fréquence de certains groupes de taxons) (Stevenson *et al.* 2010a). Les modèles prédictifs quant à eux, à la différence des indices autécologiques où la comparaison avec les sites de référence passe par le calcul de ratios d'indices (EQR), déterminent la qualité d'un site en

évaluant directement l'altération de sa communauté par rapport à une communauté de référence retrouvée dans des sites non altérés et présentant les meilleures conditions possibles, ce pour une région donnée (Feio *et al.* 2009). Cette dernière approche est assez complexe à mettre en place car elle nécessite un jeu de données suffisamment grand pour pouvoir définir la communauté de référence.

Les indices autécologiques étant les plus simples à mettre en place, la plupart des indices diatomiques actuels sont construits de cette manière et se basent généralement sur la formule développée par Zelinka & Marvan (1961) :

$$\text{Valeur d'indice} = \frac{\sum_{j=1}^n a_j \cdot s_j \cdot v_j}{\sum_{j=1}^n a_j \cdot v_j}$$

Celle-ci prend en compte pour chaque taxon (j) : son abondance ou son abondance relative dans la communauté (a), une valeur de sensibilité à une pollution ou à plusieurs polluants qui définissent son « optimum écologique » (s), une valeur indicatrice de la tolérance du taxon à la pollution, qui peut se définir comme sa capacité de résilience (v). Les indices basés sur cette formule diffèrent généralement entre eux par le nombre de taxons (j) rentrant dans le calcul de l'indice, ainsi que les valeurs indicatrices de tolérance (v) et de sensibilité (s) utilisées pour chaque taxon.

Afin de répondre aux exigences de la DCE, des indices basés sur les communautés de diatomées benthiques ont été développés dans toute l'Europe (Ács *et al.* 2004; Kelly 2013), principalement en adaptant des indices déjà existant comme l'Indice de Polluosensibilité Spécifique (IPS) (Cemagref 1982). La DCE imposant une harmonisation des méthodes entre les états membres, des exercices d'intercalibration sont réalisés pour comparer les méthodes d'échantillonnage, les conditions de références ou encore la définition des 5 classes de qualité (Poikane *et al.* 2014).

En France, le premier indice diatomique était basé sur 55 espèces et avait été développé pour évaluer la qualité des eaux dans le bassin de la Seine (Coste & Leynaud 1974). Depuis deux indices sont couramment utilisés :

- **Indice de Polluosensibilité Spécifique (IPS)** (Cemagref 1982): à l'origine l'IPS ne comprenait que 263 taxons, mais grâce à la complétion des bases de références il comprend 3143 taxons en 1999 et plus de 6214 en 2009. Les valeurs indicatrices de tolérance (v) et de sensibilité (s) de chaque taxa sont déterminées sur la base de différents paramètres physico-chimiques observés aux sites d'échantillonnage de ces taxa (*e.g.* température de l'eau, pH, conductivité, matières en suspension, DBO₅, DCO, O₂). La valeur de sensibilité (ou IPSS) est notée de 1 à 5 et indique si une espèce est indicatrice de mauvaise qualité d'eau (1,

polluorésistante) ou de bonne qualité d'eau (5, polluosensible). La valeur indicatrice de tolérance (ou IPSV) est notée de 1 à 3 et correspond à des espèces allant de très tolérantes (1, ubiquiste) à peu tolérantes (3).

- **Indice Biologique Diatomées (IBD)** (Lenoir & Coste 1996) : l'IBD a été développé de manière à simplifier le nombre de taxons utilisés dans le calcul de l'IPS, en enlevant les espèces considérées comme rares et en regroupant sous la forme de taxons appariés les espèces difficiles à identifier (Prygiel & Coste 2000). La création de l'IBD a été réalisée à partir d'un jeu de données de plus de 3000 sites ayant permis de : (i) définir les 1478 taxons utilisés dans le calcul (dont 838 espèces « vraies »), (ii) définir 7 classes de qualité d'eau sur la base de données physico-chimiques (pH, conductivité, DBO₅, O₂, NH₄, PO₄, NO₃) (Coste *et al.* 2009). Le calcul de l'indice prend ainsi en compte l'abondance relative des taxons dans l'échantillon, leur valeur écologique et leur probabilité de présence dans chacune des 7 classes de qualité. Cette méthode a été normalisée (AFNOR NFT 90-354 2000, 2007) et est actuellement utilisée dans le cadre de la DCE.

Les valeurs obtenues via l'IBD et l'IPS sont ensuite comparées aux valeurs d'indice obtenues pour les sites de référence afin de calculer les EQR. La valeur d'EQR est ensuite ramenée à une note sur 20, ce qui permet de catégoriser la qualité de l'eau étudiée en fonction des 5 classes de la DCE : très bonne (17 à 20), bonne (13 à 17), moyenne (9 à 13), médiocre (5 à 9), mauvaise (note inférieure à 5). Grâce à la prise en compte de nombreux paramètres physico-chimiques, l'IPS et l'IBD permettent de donner une information sur la qualité globale des eaux ; d'autres indices comme le Trophic Diatom Index sont plus des indicateurs de l'eutrophisation du milieu (Kelly & Whitton 1995). Bien que l'IBD soit requis par la DCE, l'IPS reste néanmoins plus précis car il prend en compte la totalité des espèces présentes dans le milieu et est un indice de référence dans les exercices européens d'intercalibration (Coste *et al.* 2009).

2.3.2. Application et limites des indices diatomiques

L'application des indices diatomiques requiert donc de décrire la communauté benthique présente sur le site étudié d'un point de vue qualitatif (liste des taxons présents) et quantitatif (abondance relative de chaque taxon). Ce travail, réalisé en microscopie par l'analyse morphologique des frustules, nécessite plusieurs étapes (**Figure 13**):



Figure 13 – Evaluation de la qualité écologique des cours d'eau via l'approche morphologique.

- **Prélèvement** : afin d'obtenir une communauté de diatomées benthiques la plus représentative possible du cours d'eau et de la station étudiée, l'échantillonnage suit un protocole standardisé au niveau européen qui définit : la zone du cours d'eau à prélever, le nombre et le type de substrats échantillonnés, le mode de conservation de l'échantillon (CEN - EN 13946 2004). Grossièrement, les diatomées benthiques sont récupérées en grattant avec une brosse à dent la surface de substrats durs immergés (pierres), le biofilm alors obtenu est collecté et conservé dans une solution de préservation (formaldéhyde, lugol ou éthanol).
- **Préparation de l'échantillon** : en vue de son observation au microscope, une partie du biofilm collecté précédemment est soumise à différents traitements chimiques afin de détruire la matière organique et les carbonates de calcium présents dans l'échantillon. A la fin, l'échantillon ne contient plus que les frustules de silice des diatomées qui sont ensuite montés entre lame et lamelle avec un résine à fort indice de réfraction, permettant une analyse en microscopie (CEN - EN 13946 2004).
- **Détermination des diatomées** : la détermination des taxons présents ainsi que leur abondance relative dans la communauté est réalisée en microscopie optique sur la base de l'identification et l'énumération de 400 valves (CEN EN 14407 2004). L'identification des espèces se fait via la reconnaissance morphologique des valves par un expert taxonomiste.
- **Calcul de l'indice** : afin de faciliter le calcul des indices de qualité (IPS ou IBD) ainsi que la gestion des inventaires d'espèces, un logiciel a été spécialement mis en place : OMNIDIA (Lecointe *et al.* 1993). Il centralise et met à jour régulièrement les informations relatives aux valeurs écologiques des taxons (valeurs de sensibilité et de tolérance), à la liste les taxons utilisés par les différents indices et permet le calcul en parallèle de nombreux indices diatomiques (*e.g.* IPS, IBD, TDI). Une fois le calcul de l'indice réalisé, la note obtenue permet de déterminer la classe de qualité (très bon, bon, moyen, ...) de la station étudiée.

Les indices diatomiques autécologiques, tels que l'IPS et l'IBD, utilisés dans le cadre de l'évaluation de la qualité des cours d'eau sont performants mais sont dépendants de la qualité des

données utilisées pour la conception de l'indice. Avant sa mise à jour en 2007, certaines difficultés méthodologiques affectaient le calcul de l'IBD et la note finale de qualité pour certains sites : l'indice était conçu sur la base d'un jeu de données trop faible (1332 relevés) faisant apparaître un manque de représentativité biogéographique; aucun élément ne permettait de prendre en compte les pollutions toxiques (*e.g.* métaux lourds) ; certaines espèces morphologiquement proches mais avec des préférences écologiques différentes étaient regroupées au sein du même taxon apparié faussant les notes (Coste *et al.* 2009). Les indices autécologiques, comme l'IBD et l'IPS, nécessitent de connaître les préférences écologiques de chaque espèce. Malheureusement ces informations sont manquantes pour de nombreuses espèces de diatomées, ce qui fausse le calcul de l'indice car ces espèces ne sont pas prises en compte (Passy & Bode 2004). Enfin, bien que la capacité des diatomées à répondre rapidement aux modifications de leur environnement en fasse d'excellents bioindicateurs, c'est aussi une source de variabilité et d'incertitude. Les assemblages de diatomées ayant une forte hétérogénéité spatiale et temporelle, ils peuvent par exemple varier d'une berge à l'autre de la rivière ou même d'un substrat à un autre (Fisher & Dunbar 2007). De ce fait, la manière dont le prélèvement est réalisé (*e.g.* choix du substrat) crée de l'incertitude et peut avoir un impact non négligeable sur l'indice final (Besse-Lototskaya *et al.* 2006; Kelly *et al.* 2009).

Au fil du temps, avec l'accumulation des connaissances et des expériences, des mesures ont été mises en place afin de limiter au mieux tous ces biais, notamment avec la mise en place de méthodes standardisées pour le prélèvement des diatomées ou encore la complétion des bases de références de valeurs autécologiques (nombre de sites étudiés, nouveaux taxons). Cependant, l'efficacité des indices diatomiques réside dans la méthode utilisée pour caractériser les assemblages de diatomées (composition et abondance relative d'espèces), à savoir la microscopie.

2.3.3. Limites liées à l'utilisation de la microscopie

Les indices diatomiques requièrent une identification précise des taxons au niveau spécifique. Pour y arriver on se base sur la reconnaissance morphologique des frustules en microscopie, ce qui permet d'identifier et de dénombrer les espèces qui composent la communauté représentative du site étudié. Cependant, l'approche morphologique présente des limites qui peuvent affecter le calcul d'indice et l'évaluation de qualité qui en découle.

- **Expertise taxonomique** : certaines espèces ont des morphologies très proches mais possèdent des préférences écologiques différentes (Vanelsländer *et al.* 2009). Ajouté au fait qu'on retrouve au sein d'une même population des individus de tailles variables en lien avec la

reproduction végétative (Hense & Beckmann 2015), certaines espèces ont une grande variabilité morphologique en réponse aux conditions environnementales (Kociolek & Stoermer 2010) ou en présence de composés toxiques (Pandey *et al.* 2017). De ce fait, l'identification des diatomées sur la base des caractéristiques morphologiques du frustule nécessite une expertise taxonomique élevée. Malgré cela, différents experts peuvent arriver à des identifications différentes, faussant les conclusions écologiques qui sont faites (Morales *et al.* 2001).

- **Evolution des connaissances taxonomiques** : la taxonomie des diatomées suit une évolution rapide et constante, fortement liée à l'évolution des méthodes de classification. Cela rajoute une difficulté supplémentaire car, pour reprendre les mots de David Mann, « ... the species taxonomy of diatoms is messy and lacks a satisfactory practical or conceptual basis ... » (Mann 1999). Les taxonomistes doivent donc actualiser régulièrement leurs connaissances, tenir compte des synonymies de chaque espèce et mettre à jour les bases de référence qui servent au calcul des indices de qualité (*e.g.* Omnidia). Toutes ces modifications sont une source de confusion pour les taxonomistes, ce qui peut faire apparaître des incohérences dans les identifications (Prygiel *et al.* 2002; Coste *et al.* 2009). De ce fait, des exercices d'harmonisation et d'intercalibration sont nécessaires pour limiter au maximum ces erreurs (Kahlert *et al.* 2009; Almeida *et al.* 2014).
- **Détermination d'individus morts** : la préparation des échantillons pour l'observation morphologique des frustules en microscopie ne permet pas de différencier les cellules qui étaient initialement vivantes des cellules mortes. En plus de comptabiliser les cellules mortes non informatives de la qualité du milieu, la présence de frustules provenant de communautés adjacentes et transportées par le courant peut être une source d'erreur (Round 1998). Bien que cela ne semble pas affecter la robustesse des comptages (Gillett *et al.* 2009), il est dommage de perdre cette information biologique qui peut avoir un intérêt pour évaluer l'état écologique des cours d'eau (Gillett *et al.* 2011).
- **Coût et temps d'analyse**: le coût et le temps d'analyse d'un échantillon sont directement dépendants de la complexité de l'assemblage et de la résolution taxonomique souhaitée, la détermination au niveau spécifique étant la plus onéreuse (Bennett *et al.* 2014). Des coûts supplémentaires peuvent s'ajouter si l'utilisation du microscope électronique à balayage est nécessaire pour l'identification de certaines espèces.

En soit, tous ces facteurs n'empêchent pas d'avoir une représentation qualitative et quantitative robuste des assemblages de diatomées. De plus, les indices de qualité ayant été

développés sur la base des comptages réalisés au microscope (*e.g.* IPS, IBD), l'approche morphologique est très performante pour l'évaluation de la qualité de l'eau. Cependant, son application devient limitante quand il s'agit d'évaluer rapidement et précisément l'état écologique de milliers de sites inclus dans les réseaux de surveillance (RCR, RCS, RCO). Le manque de taxonomistes, le manque de répétabilité entre les études, le temps d'analyse relativement long et le coût associé se révèlent alors difficiles à gérer. De nouvelles approches moléculaires basées sur l'ADN se sont donc développées afin de permettre une évaluation plus rapide, plus économique et plus répétable de la composition des assemblages de diatomées.

3. Metabarcoding et bioindication

Evaluer la diversité biologique d'une communauté consiste à déterminer sa richesse taxonomique (le nombre de taxons présents) et sa composition (abondance relative des taxons au sein de la communauté) (Frontier *et al.* 2008). Pour les organismes eucaryotes, cette évaluation repose généralement sur une approche basée sur la reconnaissance de caractéristiques morphologiques. Comme nous l'avons vu précédemment pour les diatomées, ces caractéristiques sont parfois loin d'être suffisantes pour différencier les espèces et nécessitent une expertise taxonomique élevée. Ce constat est d'autant plus vrai pour les organismes procaryotes comme les bactéries où les méthodes classiques, basée sur la mise en culture, ne permettent d'identifier dans l'environnement que la portion cultivables des bactéries, à savoir seulement 0.1 à 10 % des espèces (Theron & Cloete 2000). Afin de pallier les lacunes des approches classiques, de nouvelles méthodes d'identification des espèces ont été développées via des approches moléculaires basées sur l'ADN.

Chaque nucléotide d'une séquence ADN ayant la probabilité de correspondre à 4 bases (A, T, C, G), une séquence composée seulement de 15 nucléotides génère 4^{15} ($\approx 10^9$) possibilités de séquences uniques, 100 fois plus que le nombre d'espèces estimées sur Terre (Hawksworth & Kalin-Arroyo 1995). Cet exemple simpliste permet de visualiser le potentiel qu'à l'ADN à différencier les espèces non plus sur la base de leur variabilité morphologique, mais grâce à leur variabilité génétique. Depuis la première description de la structure de l'ADN en 1953 (Watson & Crick 1953) de nombreuses méthodes de biologie moléculaire ont été développées, les premières méthodes permettaient d'identifier les individus grâce à des sondes ADN spécifiques s'hybridant sur l'ADN (Gale & Crampton 1987). Il faudra attendre le développement du séquençage (Sanger *et al.* 1977) et de la PCR (Saiki *et al.* 1985; Mullis *et al.* 1986), pour voir apparaître les premières identifications basées sur le séquençage de produits PCR (Kocher *et al.* 1989). C'est par la suite, grâce au « Barcoding » proposé par Hebert *et al.* (2003), méthode qui permet d'identifier une espèce sur la base d'une séquence courte d'ADN ou « barcode », et à l'évolution récente des technologies de séquençage qu'il devient possible d'identifier l'ensemble des espèces retrouvées dans un échantillon environnemental : c'est le « Metabarcoding ».

3.1. Développement du metabarcoding

3.1.1. Du barcoding au metabarcoding

Le barcoding est une méthode moléculaire qui utilise un marqueur génétique, à savoir une courte séquence d'ADN ou « barcode », pour identifier à quelle espèce appartient un individu (Hebert *et al.* 2003). Pour cela, l'ADN est extrait de l'individu à identifier, le barcode ADN d'intérêt est ensuite amplifié par PCR et séquencé avant d'être comparé à une base de référence composée de séquences ADN issues d'espèces connues (**Figure 14**). Le séquençage reposait sur la technologie Sanger qui permet d'obtenir des séquences ADN de longue taille (> 1 kb), offrant ainsi la possibilité d'utiliser des gènes entiers comme barcode pour identifier les individus et ainsi de faire des analyses phylogénétiques solides (Shokralla *et al.* 2012). Par exemple, l'un des premiers barcodes utilisés correspondait à une séquence de 658 pb localisée sur le gène mitochondrial *cox1* codant pour l'enzyme cytochrome c oxydase I (Hebert *et al.* 2003). Ce gène présentait l'avantage d'être facilement amplifiable par PCR grâce à son grand nombre de copies, des amorces PCR universelles robustes étaient disponibles (*e.g.* Folmer *et al.* 1994) et l'évolution du gène suffisamment rapide pour discriminer des espèces proches (Hebert *et al.* 2003). Cependant ce barcode ne permettait pas une bonne identification au niveau spécifique des individus appartenant à certains groupes, comme les plantes et les champignons (Frézal & Leblois 2008; Kress & Erickson 2012). Depuis, différents critères ont été mis en avant, notamment par Valentini *et al.* (2009), pour caractériser un bon barcode ADN et optimiser au mieux le barcoding :

(i) la séquence de la région ciblée doit être proche entre individus d'une même espèce, mais différente entre les espèces.

(ii) elle doit pouvoir être standardisée, donc pouvoir être retrouvée chez des groupes taxonomiques différents.

(iii) la région ciblée doit contenir suffisamment d'information phylogénétique pour permettre d'assigner facilement des espèces inconnues, notamment grâce à leur positionnement dans les arbres phylogénétiques, ou non référencées aux bons groupes taxonomiques (*e.g.* genre, famille).

(iv) les zones de fixation des amorces PCR doivent être les plus conservées possibles afin de permettre une amplification la plus fiable et la plus robuste possible.

Comme il est fort probable qu'aucune séquence ADN ne puisse répondre à tous ces critères (Nielsen & Matz 2006), de nombreux barcodes ont été développés pour identifier les espèces appartenant à différents groupes d'organismes comme les animaux, les plantes, les

champignons, les protistes, les bactéries ou encore les archées (Ji *et al.* 2013). Malgré les biais inhérents au choix du barcode, le barcoding devient un outil d'identification puissant et complémentaire à la taxonomie classique (Hajibabaei *et al.* 2007), contribuant par exemple à différencier des espèces habituellement indiscernables sur la base de simples critères morphologiques (Hebert *et al.* 2004).

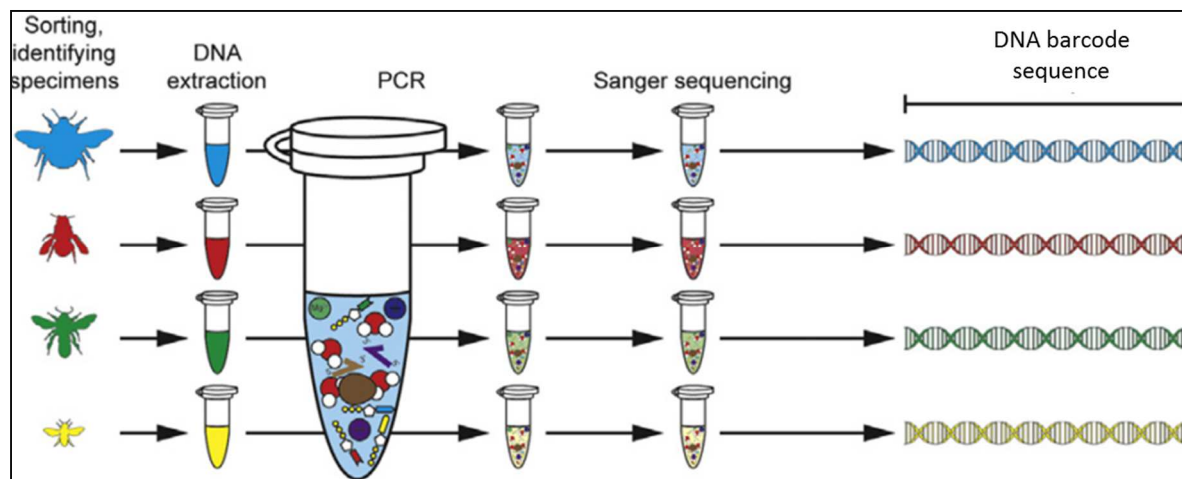


Figure 14 – Les différentes étapes du barcoding.
(adapté de Gill *et al.* 2016)

Par la suite, émerge le concept de « metabarcoding » (Sogin *et al.* 2006, Valentini *et al.* 2009), c'est-à-dire la possibilité d'identifier au niveau spécifique tous les organismes présents dans un échantillon environnemental. Le metabarcoding émergera réellement grâce à l'apparition des technologies de « séquençage de nouvelle génération » (NGS) aussi appelées technologies de « séquençage à haut débit » (HTS), qui permettent de séquencer en parallèle plusieurs millions de séquences longues de plusieurs centaines de bases. Le haut-débit permettra de s'affranchir des limites du clonage séquençage utilisé initialement.

Dès la sortie du premier séquenceur 454 en 2005 (Margulies *et al.* 2005), il devient dès lors possible d'évaluer la « totalité » de la diversité biologique d'une communauté environnementale à une échelle sans précédent, permettant par exemple pour les communautés microbienne de prendre en compte les taxons rares ou impossibles à cultiver (*e.g.* Sogin *et al.* 2006). Depuis, le terme « metabarcoding » est préféré à celui de barcoding environnemental et il désigne : « l'identification de l'ensemble des espèces (ou niveaux taxonomiques supérieurs) sur la base des barcodes amplifiés et séquencés à partir de l'ADN total extrait d'un échantillon environnemental (*e.g.* sol, eau, fèces) ou d'un mélange de plusieurs spécimens» (Taberlet *et al.* 2012a) (**Figure 15**). L'ADN environnemental peut être défini comme l'ADN extrait à partir

d'échantillons environnementaux, sans étape préalable d'isolement des organismes et inclut donc l'ADN retrouvé libre d'origine extracellulaire (e.g. ADN issus d'individus mort) ainsi que l'ADN intracellulaire des microorganismes (Keck *et al.* 2017).

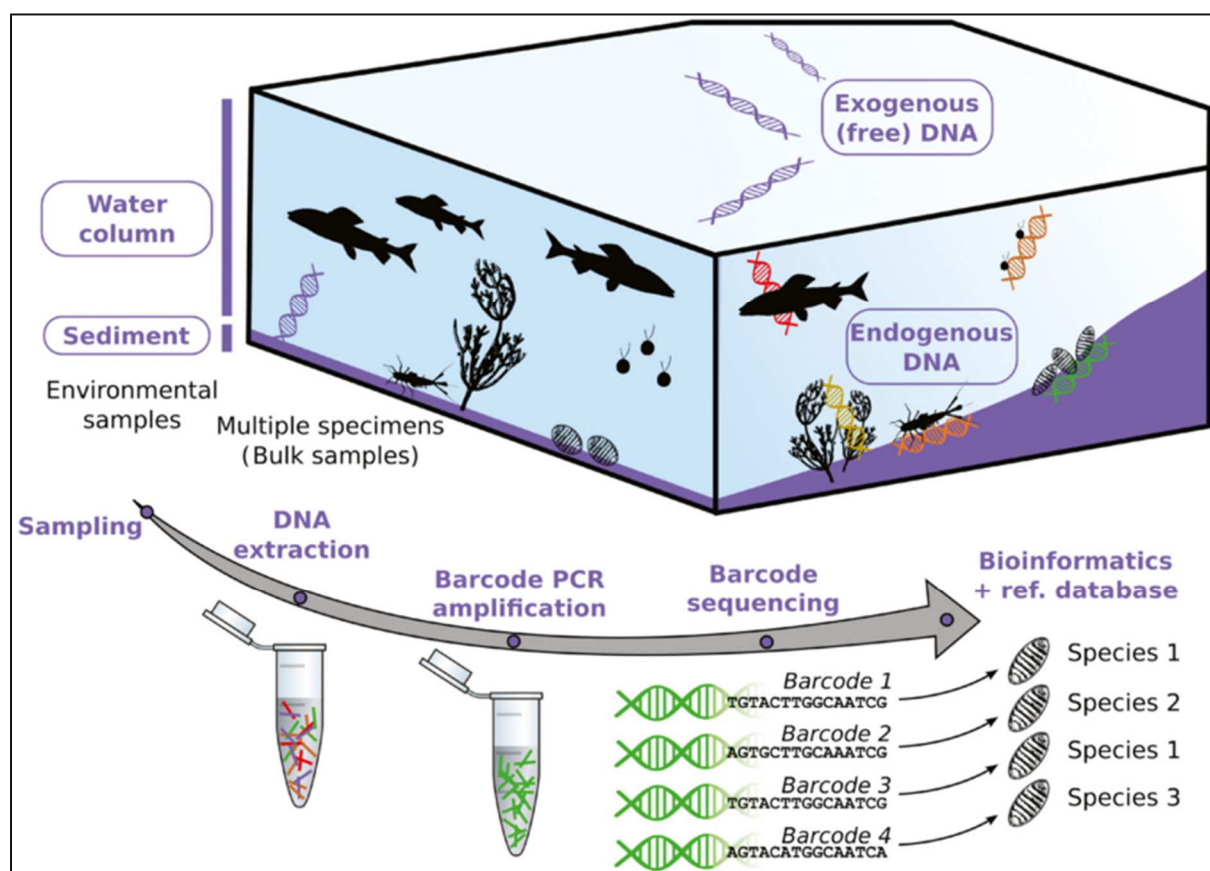


Figure 15 – Les différentes étapes du metabarcoding.
(source : Keck *et al.* 2017)

3.1.2. Contraintes liées au metabarcoding

Bien que le metabarcoding soit une approche puissante pour évaluer la diversité biologique d'un échantillon environnemental, plusieurs facteurs peuvent s'avérer problématiques.

- **Choix du barcode** : afin d'être suffisamment discriminant taxonomiquement pour permettre une identification à l'espèce, les barcodes développés pour le barcoding sont généralement d'une longueur ≥ 500 pb. Bien que les technologies HTS se soient rapidement développées ces dernières années (Goodwin *et al.* 2016), les séquences obtenues en sortie de séquençage sont en général plus courtes, de l'ordre de 300 pb. De plus, en fonction du type d'échantillon analysé, l'ADN peut être fortement dégradé et ne permettre l'amplification que de fragments courts ≈ 150 pb (Taberlet *et al.* 2012a). De ce fait, pour le metabarcoding il a fallu adapter les

barcodes existants, créés à l'origine via la technologie Sanger et trop longs pour le HTS, voire en créer de nouveaux en fonction des communautés ciblées. Face à la difficulté de trouver un barcode adapté, certaines études proposent d'utiliser plusieurs barcodes en parallèle pour augmenter les chances d'identifier correctement les espèces présentes dans un échantillon (Coward *et al.* 2015).

- **Base de référence de barcodes** : l'efficacité du metabarcoding à identifier les espèces réside dans la complétude de la base de référence utilisée. Celle-ci doit contenir suffisamment de séquences pour permettre de couvrir la diversité du groupe ciblé et ainsi permettre une identification au niveau spécifique. La création d'une telle base repose sur le séquençage « Sanger » de spécimens isolés (*e.g.* individu, culture pure) dont la taxonomie a été formellement établie (Taberlet *et al.* 2012b). La complétion des bases de référence est donc particulièrement chronophage et coûteuse, particulièrement pour les microorganismes, comme les diatomées, qui doivent préalablement être isolés et mis en culture.
- **Qualité des données obtenues** : de nombreux facteurs peuvent affecter la qualité des données, que ce soit d'un point de vue qualitatif (taxons identifiés) ou quantitatif (abondances relatives des taxons). Ceci est particulièrement dû au choix des méthodes et technologies utilisées pour réaliser le metabarcoding. Par exemple, les étapes d'amplification PCR et de préparation des librairies de séquençage sont connues pour avoir des effets sur les inventaires de diversité obtenus en HTS (*e.g.* Esling *et al.* 2015a; Kobschull & Zador 2015; Valentini *et al.* 2016). De la même manière, il existe actuellement plusieurs technologies HTS pour effectuer les séquençages qui produisent des données de qualité différente en terme de longueur des séquences ou encore d'incorporation d'erreurs dans les séquences (insertion et délétion de nucléotides) (*e.g.* Bragg *et al.* 2013; Schirmer *et al.* 2015; Goodwin *et al.* 2016). Enfin, les filtres de qualités et les divers traitements bio-informatiques réalisés sur les données HTS afin d'obtenir des données plus fiables affectent eux aussi les inventaires moléculaires (Bokulich *et al.* 2012; Pylro *et al.* 2014). D'un point de vue qualitatif, tous ces biais peuvent aboutir à la détection d'espèces absentes de l'échantillon (faux-positifs) ou à la non détection d'espèces présentes dans l'échantillon (faux négatifs). Pour ce qui est de l'aspect quantitatif, le metabarcoding ne permet qu'une quantification relative des taxons sur la base de proportion de séquences. De plus, à cause des biais, le lien entre proportion de séquences et abondance ou biomasse des spécimens est difficile à mettre en avant (Egge *et al.* 2013), si bien qu'en attendant de pouvoir exploiter pleinement les données quantitatives produites en

metabarcoding, certaines études préfèrent ne pas en tenir compte dans leurs analyses (*e.g.* Chariton *et al.* 2015).

- **Analyse des données** : les technologies HTS offrent la possibilité de séquencer en parallèle plusieurs échantillons à la fois (ce qu'on appelle le multiplexage), ce qui génère une grande quantité de données qui font entrer les analyses écologiques dans l'ère du « Big data » (Marx 2013, Keck *et al.* 2017). Afin de mieux visualiser cette évolution, en 2012 la technologie HTS « Hiseq 2000 » permettait de produire 6 millions de séquences, ce qui correspondrait à une pile de feuilles de 48 km de hauteur si on imprimait les résultats (Coissac *et al.* 2012). Ainsi de nouveaux outils d'analyse ont dû être développés pour permettre de produire et de traiter l'immense quantité de données issues des HTS, comme des programmes d'évaluation de la qualité des barcodes (Ficetola *et al.* 2010), de détection des séquences chimériques (Edgar *et al.* 2011), qui permettent d'aligner un grand nombre de séquences (Fonseca *et al.* 2012) ou de réaliser des assignations taxonomiques des séquences (Wang *et al.* 2007; Somervuo *et al.* 2017). Certains logiciels comme Mothur (Schloss *et al.* 2009), Qiime (Caporaso *et al.* 2010), HTseq (Anders *et al.* 2015) ou encore OBITools (Boyer *et al.* 2016) regroupent en une seule interface les différents outils permettant le traitement et l'analyse des données obtenues en metabarcoding. De plus, de nombreux pipelines fournissent des recommandations sur les différentes étapes de nettoyage et les paramètres à appliquer pour obtenir des données moléculaires fiables, certains étant mis à jour en ligne régulièrement (*e.g.* Kozich *et al.* 2013, <http://www.mothur.org/wiki/MiseqSOP>). De nombreuses études comparent les performances et limites de chaque approche (*e.g.* Majaneva *et al.* 2015).

Afin de pouvoir utiliser correctement le metabarcoding, il est nécessaire de mettre en place des règles adaptées qui tiennent compte de ces limitations, ce en fonction des espèces ciblées (*e.g.* détection des espèces aquatiques, Goldberg *et al.* 2016) ou encore de l'objectif de l'étude (*e.g.* détection d'espèces invasives, Trebitz *et al.* 2017). Une fois ces considérations prises en compte, le metabarcoding devient un outil puissant qui offre de nombreuses possibilités d'application.

3.1.3. Applications du metabarcoding

Malgré ces limites, le metabarcoding est apparu comme une approche complémentaire aux méthodes d'identification classiques basées sur la morphologie. Grâce à l'utilisation de l'ADN environnemental, de nombreuses études ont été menées pour caractériser différentes

communautés d'organismes (*e.g.* bactéries, champignons, protistes, plantes, animaux) et ce dans des habitats très variés (*e.g.* sédiments, permafrost, rivières, lacs) (Taberlet *et al.* 2012c; Debroas *et al.* 2015; Thomsen & Willerslev 2015; Kammerlander *et al.* 2015; Pawlowski *et al.* 2016). L'utilisation du metabarcoding a ainsi pu s'étendre à de nombreuses applications parmi lesquelles :

- **Protection et conservation de la biodiversité** : afin de lutter contre le déclin permanent de la biodiversité sur Terre et mettre en place des moyens de protection efficaces, il est nécessaire de connaître l'état et la distribution des espèces (Ji *et al.* 2013). La diversité biologique étant sous-estimée par les approches classiques, le metabarcoding apparait comme un outil complémentaire pour la préservation de la biodiversité (Thomsen & Willerslev 2015).
- **Interactions trophiques** : au sein des réseaux trophiques, les espèces entretiennent des interactions de type proie/prédateur, hôte/parasite et herbivore/plantes qui sont clés pour les transferts de matière et le fonctionnement de l'écosystème (Pompanon *et al.* 2012). L'étude de ces interactions permet de comprendre le fonctionnement des écosystèmes et repose généralement sur l'analyse du régime alimentaire des espèces via l'étude de la composition des excréments ou des contenus stomacaux. Ces échantillons sont particulièrement difficiles à étudier via les approches morphologiques, tandis que l'ADN qu'ils contiennent, une fois analysé via le metabarcoding, permet de déterminer précisément le régime alimentaire d'un individu (*e.g.* chat léopard, Shehzad *et al.* 2012 ; poissons Albaina *et al.* 2016). Il est aussi possible de mettre en évidence des interactions jusqu'alors impossibles à observer en différenciant par exemple les types de pollen transportés par une abeille ou encore en identifiant les traces de salive retrouvées sur un fruit (Clare 2014).
- **Paléoécologie** : l'ADN stocké dans la glace et les sédiments (terrestres et aquatiques) peut persister sur des périodes de temps longues pouvant aller jusqu'à des centaines de milliers d'années en l'absence de lumière et de sources de dégradations (Willerslev *et al.* 2003). Cet ADN constitue une archive qui permet de reconstituer l'histoire des communautés et des processus écologiques passés. Différents exemples d'applications récentes existent, permettant de clarifier les liens existant entre la diversité des communautés lacustres et les forçages locaux (eutrophisation) et globaux (climat) à l'échelle de quelques décennies ou millénaires, par exemple pour les communautés microbiennes eucaryotes (Capo *et al.* 2016, 2017) ou encore pour les cyanobactéries (Monchamp *et al.* 2016).
- **Ecotoxicologie** : comme nous l'avons évoqué précédemment, les communautés biologiques retrouvées dans les écosystèmes d'eau douce répondent aux conditions de leur

environnement. Elles sont donc utilisées pour tracer les impacts environnementaux liés aux rejets d'origines anthropiques ou de contaminants (*e.g.* Relyea & Hoverman 2006; Clements & Rohr 2009). Par exemple, (Pascault *et al.* 2014) ont utilisé le metabarcoding pour évaluer l'effet du tebuconazole (pesticide) sur les communautés bactériennes de lacs et de rivières.

- **Surveillance environnementale** : le metabarcoding est de plus en plus utilisé en combinaison des approches classiques afin d'évaluer la qualité de l'environnement. Il permet par exemple de détecter et de suivre la prolifération d'espèces exotiques non indigènes plus efficacement que les méthodes de surveillance classique (observation directe) en cherchant des traces d'ADN dans l'environnement (*e.g.* prolifération de la grenouille taureau en France, Dejean *et al.* 2012). Il est aussi de plus en plus utilisé pour évaluer l'impact des activités humaines sur les écosystèmes comme l'impact des fermes à saumon sur l'enrichissement des sédiments benthiques (Pochon *et al.* 2015) ou encore la réponse des communautés microbiennes du sol à des contaminations par des nanoparticules (McGee *et al.* 2017).

Le metabarcoding présente de nombreux avantages pour la bioindication, mais son utilisation dans le cadre de l'évaluation de la qualité des cours d'eau a été particulièrement développée ces dernières années.

3.2. Intérêt pour la bioindication

3.2.1. « Biomonitoring 2.0 »

La structure des communautés biologiques utilisées pour l'évaluation de la qualité des cours d'eau (*e.g.* macroinvertébrés, poissons, diatomées) est influencée par de nombreux facteurs environnementaux et d'origine anthropique, ce à l'échelle globale, régionale et locale, (**Figure 16**). Afin de permettre une meilleure surveillance environnementale, il est nécessaire d'obtenir une quantité suffisante de données sur les différents paramètres et facteurs qui affectent la qualité du milieu. Pour cela, de nombreux outils d'acquisition de données à haut-débit ont été mis en place, comme par exemple l'utilisation de sondes ou d'outils de télédétection qui permettent l'acquisition de données physico-chimique du milieu à très haute fréquence (Keck *et al.* 2017). L'utilisation du metabarcoding répond donc à un besoin urgent de développer de nouvelles approches complémentaires aux méthodes classiques, permettant de produire suffisamment de données et de connaissances sur les communautés biologiques afin d'anticiper la réponse des écosystèmes à des stress multiples ainsi qu'aux changements environnementaux (Friberg *et al.* 2011; Dafforn *et al.* 2016).

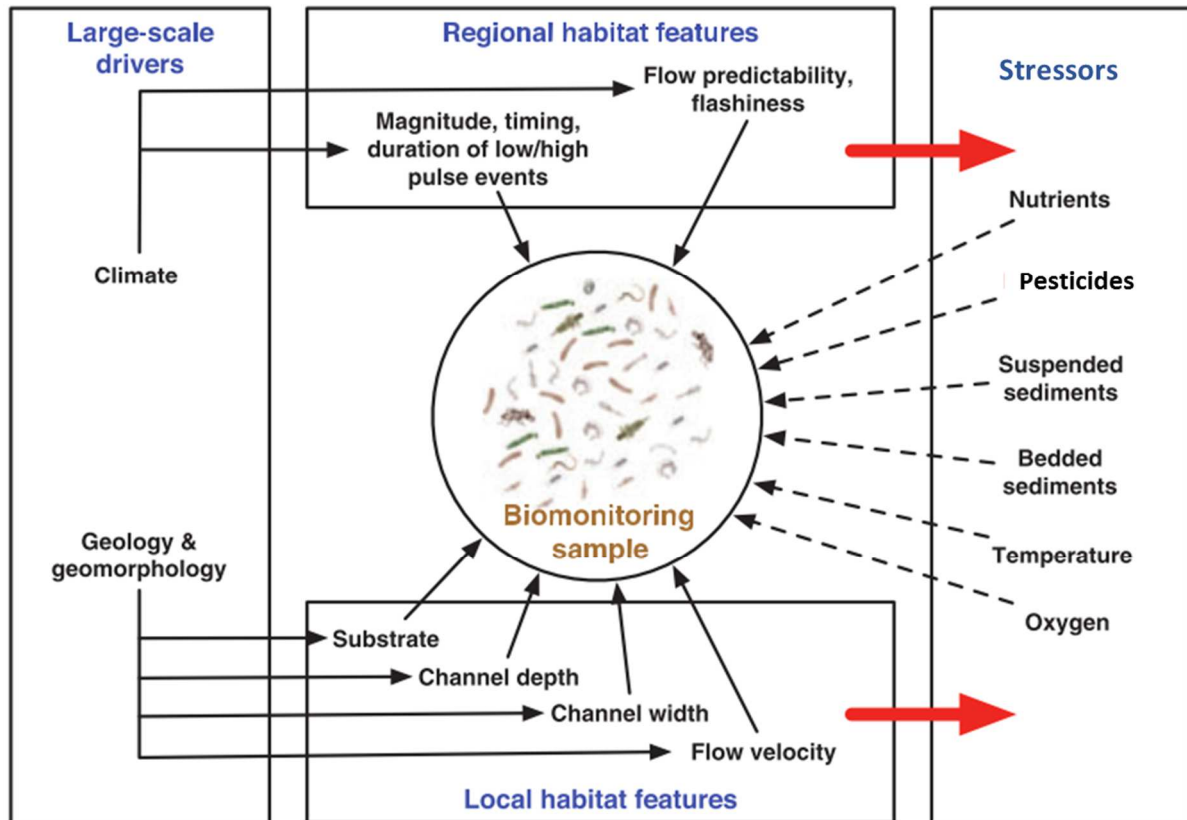


Figure 16 – Exemples de facteurs influençant les communautés biologiques utilisées en bioindication.
(adapté de Baird & Hajibabaei 2012)

Les données produites par le metabarcoding via les HTS offrent donc de nombreux avantages pour la bioindication (Keck *et al.* 2017) : (i) la grande quantité de données générées permet d'analyser plusieurs centaines d'échantillons en parallèle et offre ainsi la possibilité de surveiller plus de sites tout en permettant une évaluation précise de la biodiversité ; (ii) l'automatisation de l'identification des espèces a permis d'augmenter la vitesse d'acquisition des données, permettant d'évaluer plus rapidement l'état écologique des sites analysés ; (iii) il est possible d'obtenir une grande variété de données en ciblant plusieurs groupes indicateurs en parallèle (*e.g.* poissons, macroinvertébrés) ; (iv) les données moléculaires obtenues offrent une grande variabilité génétique ce qui permet de différencier précisément les espèces entre elles, mais aussi de différencier les communautés représentatives de sites impactés ou de référence (pour plus d'informations voir **Article annexe I**). Tous ces éléments liés à l'utilisation du metabarcoding comme outil pour la bioindication sont à l'origine de la notion de « Biomonitoring 2.0 » (Baird & Hajibabaei 2012).

3.2.2. Bioindication des cours d'eau

L'utilisation du metabarcoding pour la bioindication des cours d'eau a été développée plus ou moins rapidement en fonction des groupes d'organismes étudiés. Par exemple, pour les poissons, le metabarcoding a permis la détection des espèces grâce à leurs traces d'ADN retrouvées dans l'eau issues d'individus morts ou libérées par des individus vivants (*e.g.* fèces) (Taberlet *et al.* 2012a). Il suffit alors d'utiliser le metabarcoding sur l'ADN extrait à partir d'un échantillon d'eau, ce qui est plus simple que la mise en place des méthodes d'échantillonnages classiques (*e.g.* filets, pêches électriques) parfois lourdes et inefficaces pour certaines espèces (Valentini *et al.* 2016). De cette manière il est possible d'obtenir des inventaires d'espèces comparables à ceux obtenus via l'approche morphologique (Shaw *et al.* 2016; Civade *et al.* 2016). Cependant le manque de fiabilité des données quantitatives obtenues en metabarcoding (abondance relative de séquences) limite, dans ce cas particulier (poissons) et dans l'état actuel des techniques, le calcul des indices utilisés pour l'évaluation de la qualité de l'eau (Valentini *et al.* 2016; Shaw *et al.* 2016).

Le metabarcoding a été très rapidement testé pour identifier les communautés de macroinvertébrés benthiques dans les cours d'eau. Dès 2011, Hajibabaei *et al.* (2011) comparent l'efficacité des approches morphologique et moléculaire (metabarcoding) pour identifier les espèces de macroinvertébrés benthiques issues de 2 rivières, l'une située en zone urbaine et l'autre dans une zone protégée. Leurs résultats montrent qu'il est possible d'avoir des inventaires d'espèces via le metabarcoding aussi précis qu'avec l'approche morphologique. De plus, ils mettent en évidence la possibilité de discriminer des sites impactés par des activités anthropiques de sites non impactés. Dès lors, de nombreuses études ont permis de développer et d'optimiser cette méthode, comme la préservation des échantillons de macroinvertébrés dans l'éthanol (Hajibabaei *et al.* 2012) ou encore le choix du barcode et de primers adaptés (Elbrecht & Leese 2017), permettant d'avoir des inventaires d'espèces précis (*e.g.* Hajibabaei *et al.* 2012; Carew *et al.* 2013; Gibson *et al.* 2015). Récemment, une étude réalisée sur 18 cours d'eau Finlandais a comparé des indices basés sur des données de présence/absence des macroinvertébrés obtenues via les approches moléculaire et morphologique (Elbrecht *et al.* 2017b). Les valeurs d'indices obtenues étaient fortement corrélées entre les deux approches et permettaient une bonne évaluation de l'état écologique des cours d'eau. Cependant, comme indiqué précédemment pour les poissons, l'absence de relation claire entre abondance relative des séquences et abondances relative des taxons empêche encore l'utilisation des indices de qualité compatibles avec les exigences de la DCE (Elbrecht & Leese 2015).

En ce qui concerne les diatomées benthiques, l'application du metabarcoding pour l'évaluation de la qualité des cours d'eau n'a été développée que récemment, notamment dans le cadre du travail de thèse de Lenaïg Kermarrec (Kermarrec 2012), qui a montré le fort potentiel de cette approche pour l'évaluation de l'état écologique des cours d'eau dans le cadre de la DCE. C'est ce dernier point que nous allons aborder plus en détail dans la suite du manuscrit.

4. Metabarcoding des diatomées

Comme nous l'avons évoqué précédemment, l'identification morphologique des diatomées est particulièrement complexe par rapport à celle d'autres organismes bioindicateurs comme les macroinvertébrés ou les poissons. Le metabarcoding est donc apparu comme une solution prometteuse face aux limitations inhérentes aux méthodes classiques basées sur la reconnaissance morphologique des frustules. Bien que le principe du barcoding ait été proposé par Hebert *et al.* dès 2003, la recherche d'un barcode adapté à l'identification des diatomées ne démarrera qu'avec les travaux de Evans *et al.* (2007). Depuis, de nombreuses études ont proposé différents gènes comme barcode (Moniz & Kaczmarek 2009; Hamsher *et al.* 2011; Zimmermann *et al.* 2011; Stoof-Leichsenring *et al.* 2012). Les premiers travaux mettant en place le metabarcoding des diatomées se feront dans un premier temps avec des tests sur des communautés artificielles constituées de mélanges de cultures pures (Kermarrec *et al.* 2013b), puis sur des communautés naturelles benthiques de diatomées prélevées dans des rivières (Kermarrec *et al.* 2014; Zimmermann *et al.* 2015). Par la suite, l'utilisation du metabarcoding des diatomées comme outil pour l'évaluation de l'état écologique des cours d'eau va rapidement se développer.

4.1. Potentiel pour évaluer l'état écologique des cours d'eau

4.1.1. Développement du metabarcoding des diatomées

Le développement du metabarcoding des diatomées a tout d'abord nécessité de définir un barcode approprié pour identifier les espèces. Historiquement, 5 marqueurs génétiques, facilement amplifiables en PCR et offrant suffisamment de variabilité génétique avaient été identifiés comme barcodes potentiels : le gène mitochondrial *cox1* initialement proposé par (Hebert *et al.* 2003) ; le gène chloroplastique *rbcL* (Stoof-Leichsenring *et al.* 2012) ; différentes régions du génome codant pour les gènes ribosomiaux 18S et 28S, ainsi que la région ITS non codante (pour une revue voir Pawłowski *et al.* 2016). Bien qu'aucun de ces marqueurs ne soit parfait, les gènes 18S et *rbcL* semblent plus efficaces pour différencier les espèces de diatomées et sont donc plus adaptés pour une utilisation en bioindication (Kermarrec *et al.* 2013b; Guo *et al.* 2015; Pawłowski *et al.* 2016). Leur efficacité à identifier les espèces présentes dans des communautés environnementales de rivière via le metabarcoding a été mise en évidence pour la

région V4 du gène 18S (Zimmermann *et al.* 2015) et un fragment du gène *rbcL* (312 pb) (Kermarrec *et al.* 2014). Cependant, l'utilisation du gène *rbcL* comme barcode présente certains avantages par rapport au gène 18S : le gène *rbcL* code pour une protéine (l'enzyme Ribulose-1,5-bisphosphate carboxylase/oxygénase ou Rubisco) et de ce fait possède une variabilité intragénomique plus faible que le 18S (quelques insertion/délétion facilement détectables), rendant les séquences plus faciles à aligner et à comparer (Mann *et al.* 2010) ; l'utilisation du gène *rbcL* permet d'obtenir une meilleure résolution taxonomique au niveau spécifique que le 18S (Kermarrec *et al.* 2013b) ; ce gène étant localisé sur le génome chloroplastique, le risque de contamination et d'amplification non spécifique d'autres organismes (*e.g.* champignons) est fortement diminué par rapport à des gènes nucléiques (Evans *et al.* 2007). C'est pourquoi certaines études recommandent l'utilisation du gène *rbcL* comme barcode pour les diatomées, la région ITS pouvant être utilisée dans un deuxième temps, l'utilisation conjointe des deux barcodes devant en théorie permettre une identification plus robuste des espèces (Hamsher *et al.* 2011; Kermarrec *et al.* 2013b).

L'efficacité du metabarcoding pour identifier les espèces présentes dans un échantillon environnemental repose également sur la qualité de la base de référence utilisée pour l'assignation taxonomique des séquences. Ces bases de barcodes sont créées à partir de cultures « pures » de diatomées, population de cellules identiques issues d'une seule cellule (par mitoses successives) isolée à partir d'échantillons environnementaux, et doivent contenir au minimum : les informations concernant l'origine de la culture (*e.g.* lieu de prélèvement) ; le nom de l'espèce (taxonomie complète) identifiée sur la base de la morphologie ; la ou les séquences ADN représentatives de la culture, obtenues par séquençage Sanger (Zimmermann *et al.* 2014). Dans la dynamique des travaux de thèse de Lenaïg Kermarrec (Kermarrec 2012) et du réseau de systématique de l'INRA (RSyst, <http://www.rsyst.inra.fr/>), la base R-syst::diatom a été développée spécifiquement pour les diatomées, et donne désormais un libre accès aux séquences ADN de référence des diatomées pour les barcodes 18S et *rbcL* (Rimet *et al.* 2016). Cette base présente l'avantage (i) d'être complétée tous les 6 mois, prenant ainsi en compte les nouvelles séquences obtenues à partir de cultures pures et déposées dans les bases de données de séquences internationales ; (ii) d'utiliser un système de curation et de contrôle pour ne garder que les données fiables (*e.g.* identification des espèces correcte, qualité des séquences, disponibilité des photos en microscopie, identification des synonymies taxonomiques) ; (iii) être en accès libre pour la communauté internationale.

4.1.2. Limites et potentiel pour l'évaluation de l'état écologique des cours d'eau

La méthode proposée pour évaluer l'état écologique des cours d'eau par une approche moléculaire est la même que celle décrite précédemment (**Figure 13**), si ce n'est que l'on utilise le metabarcoding au lieu de la microscopie pour déterminer les taxons présents ainsi que leurs abondances relatives (**Figure 17**).

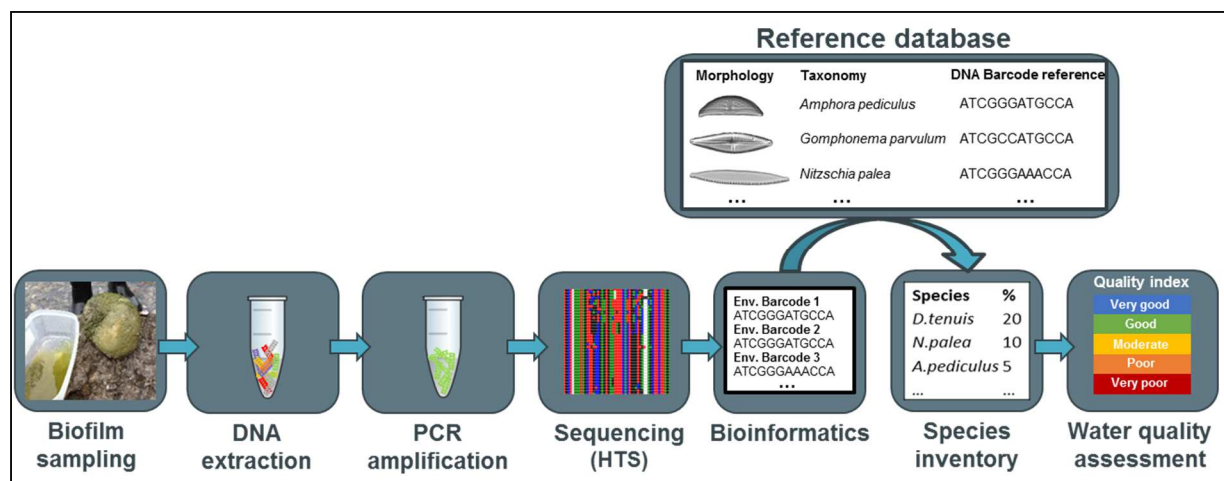


Figure 17 – Evaluation de la qualité écologique des cours d'eau via l'approche moléculaire (metabarcoding).

Cependant la mise en place de l'approche moléculaire peut être limitée par :

- **Incomplétude de la base de référence** : bien que les bases de référence de barcodes soient complétées régulièrement (*e.g.* R-syst::diatom), elles ne couvrent pas toute la diversité des diatomées. Par exemple, certaines espèces retrouvées dans des régions particulières (*e.g.* régions tropicales) et peu référencées dans les bases peuvent compliquer l'utilisation du metabarcoding dans ces régions (Kerमारrec *et al.* 2014). De ce fait, seule la partie assignée des données moléculaires peut être utilisée lors du calcul des indices, ce qui peut fausser les notes obtenues et l'évaluation de qualité qui en découle. La complétion des bases de référence est donc un facteur important mais complexe à mettre en œuvre car nécessitant d'isoler et de cultiver de nouvelles espèces pour pouvoir les séquencer (Zimmermann *et al.* 2014). Récemment, une nouvelle approche a été proposée pour compléter les bases de références (Rimet *et al.* 2018). Il s'agit d'utiliser directement comme barcodes des séquences environnementales de diatomées obtenues en metabarcoding dans des échantillons de faible diversité et présentant un taxon non référencé et dominant. L'identité taxonomique de ces barcodes fait l'objet d'une double confirmation à la fois morphologique (microscopie) et phylogénétique.

- **Divergence entre inventaires morphologique et moléculaire** : en plus des biais inhérents aux metabarcoding qui affectent l'identification et la quantification des taxons (décrits précédemment dans le paragraphe 3.1.2), d'autres biais liés à l'utilisation des diatomées peuvent être à l'origine des différences observées entre les inventaires taxonomiques obtenus avec les deux approches : **(i)** des erreurs d'identification morphologique des espèces lors de la détermination au microscope optique ; **(ii)** la présence de frustules appartenant à des individus morts dans l'échantillon naturel qui sont observés en microscopie mais non détectés en metabarcoding ; **(iii)** lors de l'extraction de l'ADN total à partir du biofilm aquatique, de l'ADN extracellulaire peut être présent dans l'échantillon (Dejean *et al.* 2011), ce qui peut aboutir à la détection de faux positifs en metabarcoding avec des espèces non représentatives de la communauté benthique comme par exemple des espèces planctoniques (Rivera *et al.* 2017) ; **(iv)** la présence d'espèces cryptiques pour lesquelles l'identification sur la base de critères génétiques peut permettre de différencier des individus morphologiquement identiques (Kermarrec *et al.* 2013a), cette diversité cryptique pouvant contribuer aux différences entre les inventaires moléculaire et morphologique (Visco *et al.* 2015); **(v)** à l'inverse, en fonction du barcode choisi et de sa résolution taxonomique, certaines espèces proches ne sont pas différenciables en metabarcoding, ce qui ne permet qu'une identification au niveau du genre (*e.g.* le barcode *rbcL* de 312 bp ne permet pas de différencier *Fragilaria vaucheriae* et *Fragilaria rumpens*) ; **(vi)** les comptages en microscopie reposent sur l'identification de 400 valves tandis qu'aucune limite n'est fixée au metabarcoding qui peut plus facilement mettre en évidence des espèces rares de par une forte profondeur de séquençage.

Malgré le faible pourcentage de séquences assignées au niveau spécifique (seulement 30% dans l'étude de Visco *et al.* 2015) et le faible nombre de taxons communs identifiés par les deux approches (de 22% à 75% dans l'étude de Kermarrec *et al.* 2014), les valeurs d'indices diatomiques calculés à partir des inventaires taxonomiques morphologiques et moléculaires sont généralement similaires (Kermarrec *et al.* 2014; Visco *et al.* 2015). Le calcul des indices diatomiques semble donc permettre une certaine souplesse dans la précision des inventaires taxonomiques moléculaires. Ceci est notamment dû au fait que les taxons non identifiés via le metabarcoding sont généralement peu abondants dans les échantillons et ont donc un faible poids dans le calcul des indices par rapport aux taxons abondants (Kermarrec *et al.* 2014; Visco *et al.* 2015). De plus, même si une identification au niveau spécifique n'est pas possible, il est possible d'utiliser des niveaux taxonomiques supérieurs (*e.g.* genre) dans le calcul des indices tout en

conservant une bonne évaluation de l'état écologique des cours d'eau (Rimet & Bouchez 2012a). Si des améliorations sont encore nécessaires, les données qualitatives fournies par le metabarcoding sont suffisamment fiables pour être utilisées dans le calcul des indices de qualité. Cependant, comme aucune corrélation fiable entre abondances relatives de séquences et abondances des taxons n'a encore pu être montrée, le plus grand potentiel d'amélioration des indices de qualité moléculaire réside dans l'optimisation de la quantification en metabarcoding.

4.2. Verrous majeurs liés à la quantification

4.2.1. Biais biologiques

En fonction des organismes étudiés, certains facteurs biologiques peuvent avoir un impact direct sur les proportions de séquences obtenues en metabarcoding. Un des problèmes récurrents est le fait que le nombre de copies du gène utilisé comme barcode par cellule n'est pas constant et peut varier fortement d'une espèce à l'autre. L'abondance relative des séquences obtenues en metabarcoding est donc liée à la fois à l'abondance du taxon dans l'échantillon, mais aussi au nombre de copies du gène qu'il possède. Cela peut aboutir à la surreprésentation ou la sous-représentation de certaines espèces dans les inventaires moléculaires, comme montré pour les communautés microbiennes (Kembel *et al.* 2012), les diatomées (Zimmermann *et al.* 2015), les macroinvertébrés (Elbrecht *et al.* 2017a) ou encore les poissons (Deagle *et al.* 2013). Le nombre de copies d'un gène dans une cellule dépend principalement (i) du nombre de copies du gène par génome, (ii) du nombre de copies du génome par cellule, (iii) du nombre d'organites ou de plastes présents dans la cellule pour certains gènes (*e.g.* nombre de chloroplastes pour le gène *rbcl*, nombre de mitochondries pour le *cox1*). Par exemple le gène 16S peut varier de 1 à 15 copies par cellules chez les bactéries et les archées (Lee *et al.* 2009) tandis que le gène 18S peut varier de 1 à plusieurs milliers de copies chez les organismes eucaryotes unicellulaires (Zhu *et al.* 2005; Weber & Pawlowski 2013; de Vargas *et al.* 2015). De plus, l'application du metabarcoding aux organismes multicellulaires tels que les poissons ou les macroinvertébrés pour lesquels la masse des individus ainsi que la densité des cellules peuvent être des biais supplémentaires qui favorisent la surreprésentation des plus grosses espèces (Deagle *et al.* 2013; Elbrecht *et al.* 2017a). Les diatomées présentent l'avantage d'être des organismes unicellulaires qui ne sont pas impactés par ces deux paramètres biologiques. Cependant, il semblerait que chez les diatomées, ainsi que pour la majorité du phytoplancton, la quantité d'ADN intracellulaire soit directement corrélée au biovolume cellulaire (Boucher *et al.* 1991; Zhu *et al.* 2005). Le nombre de copies de

certaines gènes retrouvés dans une cellule est donc directement lié à la taille de la cellule (comme montré pour le gène 18S, Godhe *et al.* 2008), ce qui peut aboutir à une surestimation en metabarcoding des diatomées avec un biovolume élevé par rapport aux espèces de plus petite taille.

Un autre biais biologique, qui peut aussi être considéré comme un biais technique, est lié au frustule des diatomées utilisé par l'approche morphologique pour identifier les espèces. Le frustule possède une très grande résistance mécanique et est capable de résister à des pressions allant jusqu'à 100-700 tonnes/m² (Hamm *et al.* 2003; Sumper & Brunner 2006). Il est aussi capable de résister à des traitements chimiques extrêmes comme ceux réalisés lors de la préparation des échantillons de biofilms avant la détermination au microscope (*e.g.* acide chlorhydrique). La plupart des méthodes d'extraction d'ADN reposent sur des traitements chimiques ou mécaniques qui ont pour objectif de lyser les cellules afin d'en libérer l'ADN, celles-ci n'étant pas forcément adaptées pour les diatomées (Nguyen *et al.* 2011; Yuan *et al.* 2015). Deiner *et al.* (2015) ont montré, pour les bactéries et les eucaryotes d'eau douce, que le choix de la méthode de lyse cellulaire affecte l'efficacité de l'extraction d'ADN ainsi que la détection de certaines espèces en metabarcoding. Pour les diatomées, il est donc possible que les séquences obtenues en metabarcoding, ainsi que leurs abondances relatives, soient affectées par la méthode d'extraction d'ADN employée.

4.2.2. Biais techniques

Outre l'étape d'extraction de l'ADN déjà évoquée précédemment, les différentes techniques et méthodes utilisées pour produire et traiter les données moléculaires peuvent introduire à toutes les étapes du metabarcoding de nombreux biais qui affectent les abondances relatives de séquences :

- **Amplification PCR** : les biais et erreurs inhérents à l'amplification par PCR de gènes à partir de mélanges complexes d'ADN extraits d'échantillons environnementaux sont connus depuis longtemps (Wagner *et al.* 1994; Polz & Cavanaugh 1998). En plus d'empêcher une identification correcte des taxons, ceux-ci peuvent affecter à la fois la qualité et les proportions des séquences obtenues en HTS (**Figure 18**): (i) en fonction de l'affinité d'appariement des primers aux séquences environnementales, de la température d'hybridation ou encore du nombre de cycles d'amplification, il peut y avoir une amplification préférentielle de certaines séquences qui fausse les abondances relatives des taxons (Sipos *et al.* 2007; Pinto & Raskin 2012); (ii) il peut y avoir une amplification stochastique des

séquences si l'efficacité de la PCR n'est pas optimale, ce qui se résume par le fait que certaines séquences ne sont pas dupliquées à chaque cycle, déséquilibrant les proportions d'amplicons obtenues en fin de PCR (Krebschull & Zador 2015) ; (iii) en fonction de l'état de l'ADN (Pääbo *et al.* 1990) et de l'enzyme polymérase utilisée (Odelberg *et al.* 1995), l'enzyme polymérase peut « sauter » d'une séquence à une autre lors de l'amplification, ce qui aboutit à la création de séquences hybrides ou « chimères » qui impactent elles aussi la quantification en metabarcoding (Amend *et al.* 2010). Plusieurs études ont proposé des solutions pour prévenir au maximum ces biais comme l'utilisation de cocktails de primers (Hong *et al.* 2009) ou de primers dégénérés (Elbrecht & Leese 2017), différentes stratégies d'amplifications (Krebschull & Zador 2015), l'utilisation de logiciel de détection des chimères (*e.g.* UCHIME, Edgar *et al.* 2011). Cependant, la meilleure option serait de pouvoir s'affranchir de la PCR et de séquencer directement l'ADN environnemental comme proposé par Liu *et al.* (2016).

- **Séquençage via les HTS** : depuis l'apparition de la technologie de séquençage 454 en 2005 (Margulies *et al.* 2005), de nouvelles technologies de séquençages HTS se sont développées et de nombreuses études ont comparé leurs caractéristiques (*e.g.* voir Loman *et al.* 2012; van Dijk *et al.* 2014; Laehnemann *et al.* 2016; Goodwin *et al.* 2016 pour les références les plus récentes). Les taux d'erreurs liés aux insertions et aux délétions de nucléotides dans les séquences varient grandement d'une technologie à l'autre, allant de 0,1 % pour la technologie Illumina HiSeq à 1 % pour le Ion Torrent PGM, et jusqu'à 13% pour le Pacific Biosciences RS II (Goodwin *et al.* 2016). En plus de poser des problèmes d'assignation taxonomique (*e.g.* Degnan & Ochman 2012), les séquences présentant trop d'erreurs risquent d'être éliminées lors des traitements bio-informatiques, affectant les abondances relatives des taxons (Schloss *et al.* 2011). L'évolution des technologies HTS ainsi que l'utilisation de certaines stratégies de séquençage, comme le séquençage bidirectionnel des séquences, permettent de réduire au maximum les erreurs incorporées durant le séquençage (Fox *et al.* 2014). A cela s'ajoute le fait que certaines technologies (*e.g.* Ion Torrent PGM) peuvent produire des séquences de longueurs variables, les séquences trop courtes risquant d'être éliminées lors des traitements bio-informatiques et non comptabilisées (Deagle *et al.* 2013).

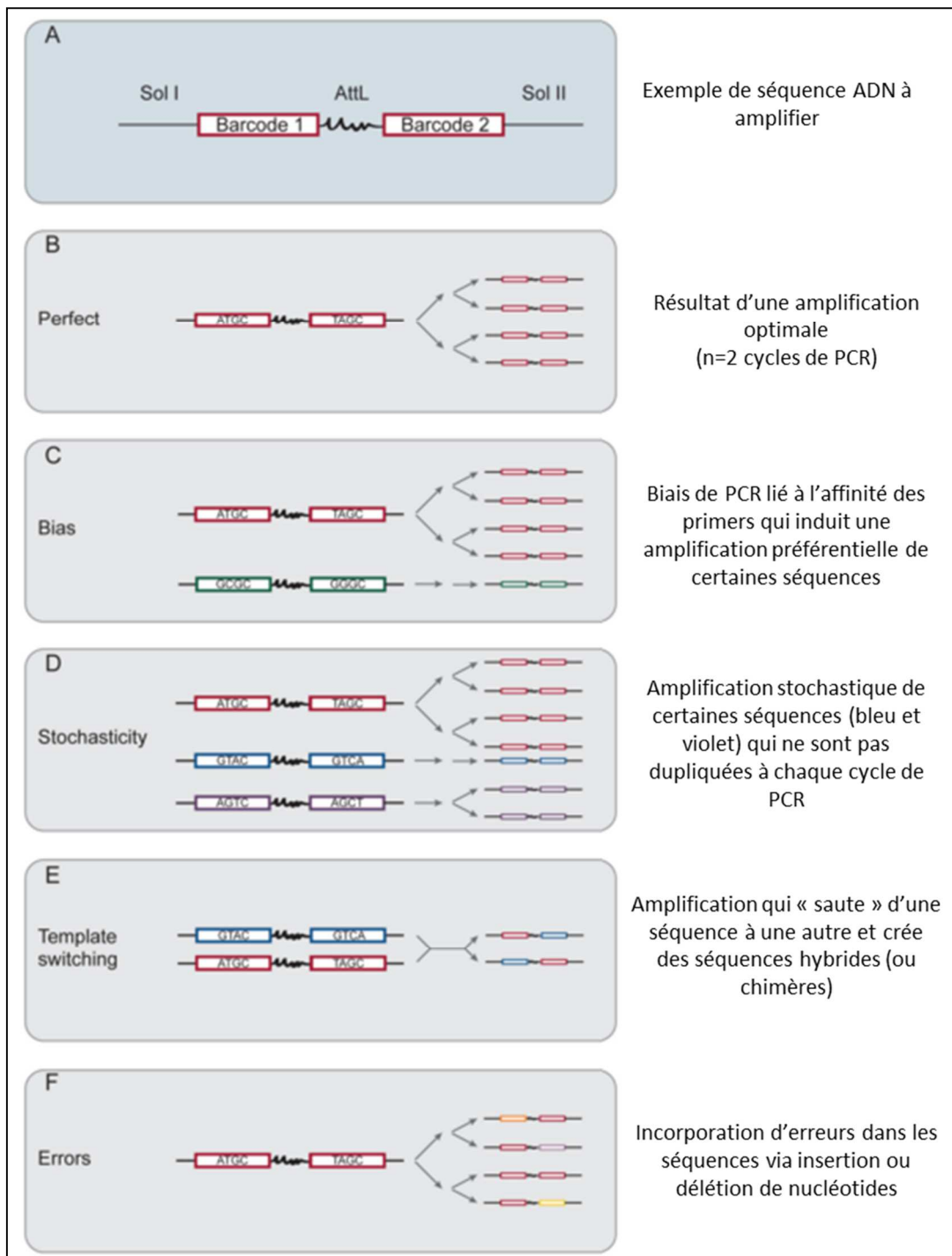


Figure 18– Erreurs et biais potentiellement introduits par la PCR.
(source : Kechschull & Zador 2015)

- **Traitements bio-informatiques** : compte tenu de la quantité de données générée par les HTS et le metabarcoding, de nombreux outils et logiciels de bio-informatique ont été développés pour faciliter leur analyse et sont accessibles via des programmes comme Mothur (Schloss *et al.* 2009) ou QIIME (Caporaso *et al.* 2010) (voir Bik *et al.* 2012 et Oulas *et al.* 2015 pour une liste plus complète). Plusieurs étapes sont généralement réalisées pour passer des séquences brutes en sortie de séquenceur à un inventaire taxonomique : (i) « pre-processing » : plusieurs

échantillons pouvant être séquencés simultanément, cette étape a pour but d'affilier chaque séquence à son échantillon d'origine et de les préparer pour les analyses suivantes ; (ii) « trimming »: afin d'éliminer les séquences douteuses, celles-ci sont filtrées sur la base de leur qualité/longueur et les séquences chimériques sont éliminées ; (iii) « OTU clustering » : les séquences sont regroupées en unité taxonomique opérationnelle (OTU) sur la base de leur similarité ; (iv) Assignation taxonomique : via les bases de référence, une taxonomie peut être assignée soit à chaque séquence, soit directement à chaque OTU. Il existe de nombreux programmes qui peuvent réaliser ces étapes, chacun utilisant des méthodes et des algorithmes différents avec autant de paramètres à régler (*e.g.* longueur et qualité minimales des séquences, % de similarité utilisé pour créer les OTUs) qui peuvent affecter la structure des communautés (*e.g.* méthode de clustering, Schmidt *et al.* 2015). Par exemple, Majaneva *et al.* (2015) ont utilisés 18 stratégies bio-informatiques différentes pour caractériser les communautés eucaryotes retrouvées dans différents échantillons (glace, neige, eau). Leurs résultats montrent qu'en fonction de la stratégie employée (*e.g.* choix de longueur de séquence, méthode de clustering, suppression des chimères), on obtient des compositions taxonomiques très différentes qui mènent parfois à des conclusions écologiques différentes.

L'effet de ces biais sur les données de metabarcoding obtenues à partir de communautés benthiques de diatomées a été très peu étudié. De ce fait, leur impact sur les calculs d'indices diatomiques et sur l'évaluation de qualité qui en découle reste indéterminé.

4.3. Implémentation à l'échelle de la DCE

Outre les considérations méthodologiques décrites précédemment (*e.g.* fiabilité de la quantification), d'autres questions relatives à la mise en pratique de l'approche moléculaire à l'échelle de réseaux de surveillances nationaux (*e.g.* RCS, RCO pour la France) doivent être élucidées avant d'envisager une implémentation dans la DCE :

- **Coût et temps d'analyse** : Stein *et al.* (2014) ont montré que le metabarcoding peut être une alternative économique viable par rapport à l'approche morphologique, les coûts étant similaires. De plus, avec la possibilité d'automatiser et de paralléliser le traitement de plusieurs centaines d'échantillons, le metabarcoding devrait permettre une vitesse d'analyse plus rapide et donc fournir un plus haut débit à la biosurveillance des milieux (Keck *et al.* 2017). Même si peu d'études ont été réalisées à grande échelle pour confirmer cela, il semblerait que les aspects de coût et de temps d'analyse ne soient pas problématiques pour une éventuelle implémentation dans le cadre de la DCE (Hunting *et al.* 2017).

- **Harmonisation des méthodes** : comme évoqué précédemment, la DCE impose une harmonisation des méthodes de bioindication entre les états membres. Comme l'approche moléculaire a été développée récemment, il n'existe pas encore de recommandations ou de standards pour les différentes étapes du metabarcoding (*e.g.* extraction ADN, séquençage, traitements des données, ...). A ce jour, chaque pays développe sa propre approche moléculaire avec des protocoles variés. Cependant, pour le metabarcoding des diatomées une action de standardisation a été initiée par un groupe de diatomistes européens auprès du Comité Européen de Normalisation dès 2012, ayant abouti à ce jour à 2 « Technical Specifications » décrivant le protocole d'échantillonnage et les critères d'établissement d'une base de référence. Plus récemment, le réseau européen COST DNAqua-Net (CA 15219 2016-2020, Leese *et al.* 2016) a été proposé afin de créer un réseau de discussion entre scientifiques de toute l'Europe dans le but de comparer, d'harmoniser et d'aller vers l'implémentation des méthodes : (i) protocoles d'échantillonnages ; (ii) bases de référence ADN et choix des barcodes ; (iii) plateformes de séquençages ; (iv) analyse et stockage des données ; (v) création de nouveaux indices moléculaires.
- **Communication entre les différents acteurs** : une éventuelle implémentation de l'approche moléculaire dans la DCE aura un impact direct sur tous les acteurs impliqués dans la surveillance et la protection de l'eau. Que ce soit les politiques, les gestionnaires, les organismes producteurs des données (*e.g.* Agence de l'eau, DREAL) ou les organismes privés (*e.g.* bureaux d'étude), il est primordial d'informer tous les acteurs de l'évolution des méthodes de bioindication. Il est donc nécessaire de transmettre les connaissances par l'intermédiaire de conférences ou de réunions d'informations et de co-construction, comme proposé dans le cadre du réseau COST DNAqua-Net. C'est aussi l'occasion de discuter du cadre législatif de l'implémentation des approches moléculaires au niveau européen et national.

5. Objectifs de la thèse

Les objectifs de ce travail de thèse consistaient, dans le cadre du metabarcoding des diatomées, à (i) identifier les principaux biais de quantification liés au metabarcoding, (ii) évaluer leur impact sur les inventaires taxonomiques et (iii) optimiser l'approche moléculaire afin de développer un outil d'évaluation de l'état écologique des cours d'eau à l'échelle du réseau de surveillance de Mayotte et plus globalement dans le cadre de la DCE. Afin de répondre au mieux à ces objectifs, la thèse se structure autour de 2 axes principaux :

- Axe 1 : étudier l'impact de différents biais (techniques et biologiques) sur le metabarcoding des diatomées afin d'optimiser l'approche moléculaire et produire des inventaires taxonomiques les plus fiables possibles d'un point de vue qualitatif (composition d'espèces) et surtout quantitatif (abondance relative des espèces).
- Axe 2 : tester la capacité de l'approche moléculaire, en comparaison de l'approche morphologique, pour l'évaluation de l'état écologique des cours d'eau par comparaison à l'approche classique basée sur la morphologie, en prenant comme cas d'étude les cours d'eaux de Mayotte et de France métropolitaine appartenant aux réseaux de surveillance inclus dans la DCE.

5.1. Stratégie

5.1.1. Base de travail

Les étapes initiales du développement du metabarcoding et de l'approche moléculaire pour les diatomées ont été réalisées durant les travaux de thèse de Lenaïg Kermarrec (Kermarrec 2012) dans le cadre d'une collaboration entre l'INRA de Thonon et le bureau d'étude ASCONIT consultant. Ces travaux ont permis de définir l'intérêt du gène *rbcL* (fragment de 312 pb) comme barcode pour étudier la structure des communautés de diatomées (Kermarrec *et al.* 2013b, 2014) et d'initier le développement d'une base de référence de barcodes (Rimet *et al.* 2016). Les travaux de thèse présentés dans ce manuscrit s'inscrivent donc dans la continuité de ce travail et répondent aux attentes de l'ONEMA-AFB qui souhaite tester des approches innovantes, dont l'approche moléculaire, dans le cadre de la mise en place de la bioindication DCE des cours d'eau à Mayotte.

Afin de permettre le traitement et l'analyse des données issues du séquençage d'un grand nombre d'échantillons, un protocole d'analyse utilisant le programme « Mothur » (Schloss *et al.*

2009) a été adapté au metabarcoding des diatomées au cours de ces travaux de thèse (données non publiées).

5.1.2. Axe 1 - Développement de l'approche moléculaire : Impact des biais techniques et biologiques sur la quantification

Comme indiqué précédemment, le metabarcoding ne permet pas d'obtenir des données quantitatives fiables, les abondances relatives de séquences n'étant pas bien corrélées aux abondances de taxons. Parmi les biais identifiés et décrits précédemment (voir la partie 4.2), nous avons décidé de nous concentrer principalement sur les biais biologiques qui affectent particulièrement le metabarcoding des diatomées et qui sont peu étudiés dans la littérature, à savoir : le choix de la méthode d'extraction d'ADN en lien avec les frustules des diatomées (**Chapitre II**) et les variations du nombre de copies du gène *rbcL* en lien avec le biovolume des cellules (**Chapitre III**). Concernant les biais techniques, nous avons mis en place une étude permettant l'évaluation des biais d'amplification et de séquençage pour lesquels nous avons peu de recul pour le metabarcoding des diatomées. Les résultats de cette dernière n'étant pas encore finalisés, elle sera seulement abordée dans la partie discussion et perspective du manuscrit (**Chapitre VI**). Bien que les biais liés aux traitements bio-informatiques des données soient importants, ceux-ci sont déjà largement étudiés dans la littérature et n'ont pas pu être abordés dans le laps de temps imparti à la thèse.

Les différentes études menées sur ces biais techniques et biologiques ont été réalisées à la fois avec l'approche moléculaire (metabarcoding) et l'approche morphologique (microscopie). Elles sont présentées ci-dessous et visent à répondre aux questions suivantes :

- Quels sont les impacts de ces biais techniques et biologiques sur les données de séquences produites et les inventaires taxonomiques qui en sont issus ?
 - Est-ce que ces impacts potentiels affectent le calcul d'indice et l'évaluation de qualité qui en découle ?
 - Comment limiter ces biais et leurs impacts ?
-
- **Chapitre II - "Application of high-throughput sequencing (HTS) metabarcoding to diatom biomonitoring: Do DNA extraction methods matter" (article publié dans *Freshwater Science*)**

En fonction de la méthode de lyse cellulaire employée, le frustule des diatomées est connu pour diminuer l'efficacité d'extraction de l'ADN, particulièrement pour les espèces possédant un

frustule épais et plus robuste (Eland *et al.* 2012). Afin d'évaluer à quel point ce biais peut fausser les données obtenues en metabarcoding, les travaux présentés dans ce chapitre ont été réalisés en plusieurs étapes : (i) l'efficacité de différentes méthodes pour extraire l'ADN a été testée sur des cultures pures de diatomées et sur des échantillons environnementaux (biofilms aquatiques), (ii) les échantillons environnementaux ont été séquencés afin d'estimer l'impact de la méthode d'extraction sur la structure des communautés en metabarcoding, (iii) les notes d'IPS calculées sur la base des inventaires moléculaires obtenus à partir des différentes méthodes d'extraction ont ensuite été comparées afin de voir l'impact de celles-ci sur les évaluations de qualité.

- **Chapitre III – “A correction factor inferred from cell biovolume improves quantification in diatom metabarcoding for Water Framework Directive monitoring” (*article soumis dans Methods in Ecology and Evolution*)**

Des travaux ont montré que le nombre de copies du gène 18S suit une corrélation linéaire avec les biovolumes des diatomées (Godhe *et al.* 2008). Si cette corrélation existe aussi pour le gène *rbcl*, il est possible d'appliquer un facteur de correction basé sur le biovolume cellulaire pour corriger les données moléculaires et obtenir des données quantitatives plus fiables. Pour tester cette hypothèse, nous avons d'abord vérifié l'existence d'une corrélation entre nombre de copies du gène *rbcl* et biovolume cellulaire à partir de données de qPCR obtenues sur des cultures pures. Puis nous avons évalué l'efficacité de ce facteur de correction sur des données moléculaires obtenues à partir de communautés artificielles contenant des proportions d'espèces connues, puis sur des communautés environnementales. Finalement, pour les échantillons environnementaux, les valeurs d'IPS moléculaire obtenues à partir des inventaires moléculaires corrigés ont été comparées aux valeurs d'IPS moléculaires non corrigées et morphologiques.

5.1.3. Axe 2 : Application de l'approche moléculaire pour évaluer l'état écologique des cours d'eau à l'échelle de réseaux de surveillance

Jusqu'à présent, peu d'études ont essayé de comparer les approches morphologiques et moléculaires à l'échelle d'un réseau de surveillance environnemental incluant beaucoup de sites avec un gradient de qualité marqué, allant de sites pollués jusqu'à des sites de référence. Ceci est essentiellement dû au fait que le metabarcoding est encore en développement, la majorité des études se concentrant sur l'optimisation de la méthode en effectuant des tests sur quelques sites (Kermarrec *et al.* 2014; Zimmermann *et al.* 2015; Visco *et al.* 2015) ou à une échelle régionale

sur un gradient de qualité limité (Apothéloz-Perret-Gentil *et al.* 2017). Une application à grande échelle de l'approche moléculaire permettrait notamment, outre une validation de l'approche, d'avoir (i) plus de recul sur les biais de la méthode, (ii) d'identifier de potentiels problèmes liés au traitement et à l'analyse de grands jeux de données, (iii) de sensibiliser les différents acteurs actuellement en charge de la protection de l'eau (*e.g.* politiques, gestionnaires), (iv) d'évaluer la faisabilité de l'approche dans le cadre de la DCE (*e.g.* temps, coûts). Nous avons donc décidé de travailler à l'échelle de deux réseaux de surveillance à Mayotte et en France métropolitaine, travaux qui seront présentés respectivement dans les deux chapitres résumés ci-dessous.

- **Chapitre IV – “*Avoiding quantification bias in metabarcoding: application of a cell biovolume correction factor in diatom molecular biomonitoring*” (article publié dans *Ecological Indicators*)**

L'île de Mayotte est un département d'outre-mer depuis 2011 et, à ce titre, la DCE doit être appliquée pour la surveillance de ses cours d'eau. Comme Mayotte est le seul DOM pour lequel aucun réseau de surveillance n'a encore été établi (voir **Figure 4**), un projet a été initié à l'INRA sous l'égide de l'ONEMA-AFB en 2013 afin de définir le futur réseau de surveillance et mettre en place un outil de bioindication basé sur les communautés de diatomées et utilisant l'approche morphologique classique pour l'évaluation de l'état écologique des cours d'eau. A la demande de l'ONEMA-AFB, ce fut aussi l'occasion de tester en parallèle sur ces cours d'eau une approche innovante de bioindication : le metabarcoding ADN. L'étude présentée ici a donc permis de comparer les deux approches en termes de fiabilité des inventaires taxonomiques produits et de notes de qualité obtenues à l'échelle du réseau de cours d'eau de Mayotte.

- **Chapitre V – Intégration de l'approche moléculaire à l'échelle du réseau DCE français de surveillance des cours d'eau (étude en cours)**

Contrairement à Mayotte, le réseau de surveillance des cours d'eau en France métropolitaine existe depuis plusieurs décennies et inclut plusieurs milliers de sites. Il existe donc un historique de données basées sur l'approche morphologique qui est assez conséquent et qui sert de base aux gestionnaires pour définir des plans de gestions à plus ou moins long terme. En plus de permettre une comparaison des approches moléculaire et morphologique à très grande échelle (*e.g.* coût, temps, efficacité), l'étude réalisée sur ce réseau a pour objectif de sensibiliser les différents acteurs sur l'approche moléculaire et de les impliquer dans la réflexion vis-à-vis d'une potentielle intégration dans la DCE.

5.2. Structure de la thèse

En complément du **Chapitre I** qui permet d'introduire le contexte général de la thèse, la thèse est constituée de différentes parties qui s'organisent de la manière suivante :

- L'axe 1 qui traite principalement des biais de quantification liés aux méthodes d'extraction d'ADN (**Chapitre II**) et aux variations du nombre de copies du gène *rbcL* (**Chapitre III**).
- L'axe 2 qui présente l'application de l'approche moléculaire à l'échelle des réseaux de surveillance des cours d'eau de Mayotte (**Chapitre IV**) et de France métropolitaine (**Chapitre V**)
- Le **Chapitre VI** qui présente les discussions et perspectives qui découlent de ce travail de thèse.

Les **Articles annexes** correspondent aux articles scientifiques préparés en parallèle de la thèse et incluent des informations complémentaires aux travaux présentés.

6. Annexes

		Indicateurs disponibles à utiliser pour l'évaluation au cours du 2 ^{ème} cycle DCE selon les territoires				
Eléments de qualité		Métropole	Antilles	Guyane	Réunion	Mayotte
Eléments de qualité biologique	Paramètres biologiques					
Phytoplancton	Composition, abondance et biomasse					
Macrophytes	Composition et abondance	IBMR				
Phytobenthos	Composition et abondance	IBD ₂₀₀₇	IDA	IPS	IDR	
Faune benthique invertébrée	Composition et abondance	IBGN	IBMA	SMEG	IRM	
Ichtyofaune	Composition, abondance et structure de l'âge	IPR		IPG _{global}	IRP	
Elément de qualité physico-chimique	Paramètres physico-chimiques					
Température de l'eau	-	Valeurs-seuils en annexe 5	Valeurs-seuils en annexe 5 adaptables aux spécificités locales			
Bilan d'Oxygène	Oxygène dissous	Valeurs-seuils en annexe 5	Valeurs-seuils en annexe 5 adaptables aux spécificités locales			
	Taux de saturation en O ₂					
	DBO ₅					
	Carbone organique dissous					
Salinité	Conductivité					
	Chlorures					
	Sulfates					
Etat d'acidification	pH _{min} et pH _{max}	Valeurs-seuils en annexe 5	Valeurs-seuils en annexe 5 adaptables aux spécificités locales			
Concentration en nutriment	PO ₄ ³⁻	Valeurs-seuils en annexe 5	Valeurs-seuils en annexe 5 adaptables aux spécificités locales			
	Phosphore total					
	NH ₄ ⁺					
	NO ₂ ⁻					
	NO ₃ ⁻					
Eléments de qualité hydromorphologique	Paramètres hydromorphologiques					
Régime hydrologique	Quantité et dynamique du débit d'eau					
	Connexion aux masses d'eau souterraines					
Continuité de la rivière	-					
Conditions morphologiques	Variation de la profondeur et de la largeur de la rivière					
	Structure et substrat du lit					
	Structure de la rive					
Vert : indicateurs disponibles pour le 2 ^{ème} cycle DCE / Jaune : indicateurs disponibles pour le 2 ^{ème} cycle mais devant être remplacés dès le 3 ^{ème} cycle DCE (adoption des indices I _{3M} et IPR+, détermination des valeurs-seuils des paramètres physico-chimiques soutenant la biologie) / Rouge : indicateurs à développer pour le 3 ^{ème} cycle DCE / En gris : indicateurs non pertinents						

Annexe 1 – Eléments de qualités et indicateurs utilisés pour l'évaluation de l'état écologique des cours d'eau en France (source : Ministère chargé de l'environnement 2016).

General		Macrophytes and phytobenthos
Class		
High	<p>There are no, or only very minor, anthropogenic alterations to the values of the physico-chemical and hydromorphological quality elements for the surface water type from those normally associated with that type under undisturbed conditions.</p> <p>The values of the biological quality elements for the surface water body reflect those normally associated with that type under undisturbed conditions, and show no, or only very minor, evidence of distortion.</p> <p>These are the type-specific conditions and communities.</p>	<p>The taxonomic composition corresponds totally or nearly totally to undisturbed conditions.</p> <p>There are no detectable changes in the average macrophytic and the average phytobenthic abundance.</p>
Good	<p>The values of the biological quality elements for the surface water body show low levels of distortion resulting from human activity, but deviate only slightly from those normally associated with the surface water body type under undisturbed conditions.</p>	<p>There are slight changes in the composition and abundance of macrophytic and phytobenthic taxa compared to the type-specific communities. Such changes do not indicate any accelerated growth of phytobenthos or higher forms of plant life resulting in undesirable disturbances to the balance of organisms present in the water body or to the physico-chemical quality of the water or sediment.</p> <p>The phytobenthic community is not adversely affected by bacterial tufts and coats present due to anthropogenic activity.</p>
Moderate	<p>The values of the biological quality elements for the surface water body type deviate moderately from those normally associated with the surface water body type under undisturbed conditions. The values show moderate signs of distortion resulting from human activity and are significantly more disturbed than under conditions of good status.</p>	<p>The composition of macrophytic and phytobenthic taxa differs moderately from the type-specific community and is significantly more distorted than at good status. Moderate changes in the average macrophytic and the average phytobenthic abundance are evident. The phytobenthic community may be interfered with and, in some areas, displaced by bacterial tufts and coats present as a result of anthropogenic activities.</p>
Poor	<p>Waters showing evidence of major alterations to the values of biological quality elements from those normally associated with the surface water body type under undisturbed conditions.</p>	<p>The taxonomic composition corresponds totally or nearly totally to undisturbed conditions.</p>
Bad	<p>Waters showing evidence of severe alterations to the values of the biological quality elements for the surface water body type and in which large portions of the relevant biological communities normally associated with the surface water body type under undisturbed conditions are absent.</p>	<p>The taxonomic composition corresponds totally or nearly totally to undisturbed conditions.</p>

Annexe 2 – Définitions normatives des 5 classes de qualité utilisés dans la DCE pour les cours d'eau et les rivières (source : Annexe 5 de la DCE, Kelly 2013).

<i>Division</i>	<i>Bacillariophyta</i>	Descriptions and subgroups	Examples of taxa
<p><i>Subdivision</i> Subdivision Coscinodiscophytina: monophyletic in Medlin and Kaczmarska (2005) (and then comprising the single class Coscinodiscophyceae), paraphyletic in Theriot et al. (2015). Contains several clades of radial centric diatoms whose interrelationships are unclear. Valves generally circular; pattern-center an annulus; sexual reproduction via oogamy; auxospores with scales only</p>	leptocylindriids	Chain-forming, delicate; valves circular, striae radiating from a central circular annulus; unique simple process present near the annulus; girdle bands segmental; auxospore forming a dormant resting stage (not present in other centric clades)	<i>Leptocylindrus</i> , <i>Tenuicylindrus</i>
	corethriids	Solitary; valves circular; radially symmetrical; articulating spines secreted from around the valve margin; rimoportulae absent; girdle bands segmental	<i>Corethron</i>
	melosiriids	Usually chain-forming, sometimes forming special "separation valves"; valves circular, radially symmetrical; rimoportulae small, scattered on the valve face or marginal; girdle bands hooplike or segmental	<i>Aulacoseira</i> , <i>Melosira</i> , <i>Podosira</i> , <i>Stephanopyxis</i>
	ellerbeckiids	= "paralids" of Mann in Adl et al. (2005); Chain-forming, heavily silicified; valves circular, radially symmetrical; small tube processes present, restricted to the mantle; girdle bands hooplike	<i>Ellerbeckia</i>
	arachnoidiscids	Solitary, heterovalvar; valves circular, radially symmetrical; one valve with its center surrounded by radial slits (apparently modified rimoportulae); girdle bands hooplike	<i>Arachnoidiscus</i>
	coscinodiscids	Solitary, isovalvar; valves usually circular, striae radiating from a central, subcentral, or submarginal circular annulus; rimoportulae central, scattered on the valve face or marginal; girdle bands hooplike	<i>Actinocyclus</i> , <i>Actinopychus</i> , <i>Coscinodiscus</i> , <i>Stellarima</i> , and many others
	rhizosoleniids	Chain-forming, with a long perivalvar axis, rarely solitary; valves circular, almost radially symmetrical or with the pattern-center displaced towards one side; rimoportula single, associated closely with the annulus, sometimes developed into a spine; girdle bands segmental	<i>Gainardia</i> , <i>Rhizosolenia</i>
	proboscids	Usually solitary, with a long perivalvar axis; valves circular, extended into an eccentric beak (proboscis); rimoportulae and other processes absent; girdle bands segmental	<i>Proboscia</i>

Annexe 3 – Nouvelles subdivisions et classes dans les Bacillariophyta (source : Mann *et al.* 2016).

II. Biais lié à la méthode d'extraction de l'ADN

“Application of high-throughput sequencing (HTS) metabarcoding to diatom biomonitoring: Do DNA extraction methods matter?”

(paru dans le journal *Freshwater Science*, 2017)

Valentin Vasselon¹, Isabelle Domaizon¹, Frédéric Rimet¹, Maria Kahlert², and Agnès Bouchez¹

¹CARTEL, INRA, Université de Savoie Mont Blanc, 74200, Thonon-les-bains, France

²Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, P.O. Box 7050, 75007, Uppsala, Sweden

1. Abstract

Current freshwater biomonitoring with diatoms is based on microscopic examination of the morphology of their silica skeleton. This standardized approach is time consuming and requires a high degree of taxonomic expertise. Metabarcoding combined with high-throughput sequencing (HTS) has great potential for next-generation biomonitoring applications but requires standardization. Molecular inventories are strongly influenced by the DNA extraction method used, but the effect of extraction protocols has not been tested to enable selection of the best DNA extraction method for HTS metabarcoding. We used 5 DNA extraction methods combining various types of cell lysis and DNA purification to extract DNA from 8 pure diatom cultures and 8 samples from streams and lakes with differing water quality. We compared the methods based on: 1) quality and purity of the extracted DNA, 2) community inventories obtained from HTS targeting the ribulose-1, 5-bisphosphate carboxylase (*rbcl*) barcode, and 3) similarity between molecular and microscopy-based inventories of community composition and the Specific Pollution-sensitivity Index [SPI]. A method based on GenElute™-LPA had higher extraction efficiency than the 4 commercial kits but had the highest polymerase chain reaction inhibition level. All 5 methods were efficient for HTS, and method did not affect operational taxonomic unit richness. We observed variations in the relative abundance of some taxa within *Nitzschia*, *Amphora*, *Encyonema*, *Gomphonema*, and *Navicula* between 2 of the 5 methods, but method did not affect global diatom community composition or SPI values. SPI values calculated from microscopy-based inventories and molecular inventories based on all 5 extraction methods were strongly correlated. For convenience purposes (high DNA quantity and low cost), we encourage standardization of HTS diatom biomonitoring based on the SA-Gen method.

2. Introduction

Diatoms are good bioindicators because of their high diversity, short life cycle, high sensitivity to environmental conditions, and widespread distribution in all freshwater ecosystems (Stevenson & Pan 1999). Therefore, diatom communities are used routinely for water-quality assessment in monitoring programs and by environmental agencies in many countries. Well-established guidelines like the Clean Water Act in USA (US CWA) or the Water Framework Directive in Europe (EU WFD) help to standardize methods across countries and laboratories. Classical diatom biomonitoring is based on the composition of environmental communities and relies on morphological identification at the species level with the aid of microscopes and specialized floristic books. Species identifications is challenging because of the large diversity of diatoms (Mann & Vanormelingen 2013) and the subtle differences in morphological features of their silica frustule (exoskeleton) used for taxonomy. Quite often, discrepancies in taxonomic inventories occur from one laboratory to another (Kahlert *et al.* 2012; Werner *et al.* 2016). Moreover, this approach is time consuming and costly. Increased demand for environmental assessment in recent years implies that the number of samples to be analyzed will increase, a trend that will become untenable if analysis is based on microscopic identification. Thus, fast and cost-effective alternatives must be developed. One promising alternative is application of environmental DNA metabarcoding.

The potential of DNA metabarcoding combined with high-throughput sequencing (HTS) for investigating benthic diatom community structure already has been demonstrated (Kermarrec *et al.* 2013b, 2014; Gibson *et al.* 2015; Zimmermann *et al.* 2015; Visco *et al.* 2015), opening the way to “next-generation biomonitoring”. However, these pioneer investigators used differing molecular methods and protocols, thereby hampering relevant comparison among studies. Factors ranging from the initial field sampling to the bioinformatics treatment of DNA sequences can affect the final molecular species inventory of diatom communities: 1) the DNA marker chosen, which affects species discriminatory power and the availability and completeness of a DNA reference database; 2) the methods used for various steps of molecular analyses (i.e., DNA extraction methods, sequencing technology); and 3) the bioinformatics workflow (i.e., data processing steps, clustering algorithms, and taxonomic assignment methods). HTS metabarcoding is still in its infancy, and guidelines need to be defined at each step to allow its standardization for biomonitoring purposes. Active investigations are under way to find the best DNA marker (Kermarrec *et al.* 2013b, 2014) or to optimize the HTS data sequence-processing using pipelines

(Schmidt *et al.* 2015; Majaneva *et al.* 2015) but little attention has been given to the DNA extraction from diatom samples and it requires further study.

Obtaining a molecular inventory relies on extraction of DNA representative of the indigenous diatom community composition. The quality and the quantity of the DNA extracted from environmental samples affect the investigator's ability to obtain a relevant taxonomic list. Several studies have been performed to evaluate the effect of extraction protocols on microbial DNA analysis. The studies have been focused mainly on bacterial communities (Willner *et al.* 2012; Rubin *et al.* 2014; Wesolowska-Andersen *et al.* 2014; Wagner Mackenzie *et al.* 2015) and freshwater microalgae (Eland *et al.* 2012) but rarely diatoms (Nguyen *et al.* 2011). Results of these studies show that the choice of the DNA extraction method, particularly the cell lysis type, affect quality and quantity of extracted DNA and inferences regarding community diversity and structure. However, authors of these studies generally depicted the diversity of the targeted biological groups based on fingerprinting methods (e.g., denaturing gradient gel electrophoresis [DGGE]), which provide only a very coarse view of community diversity.

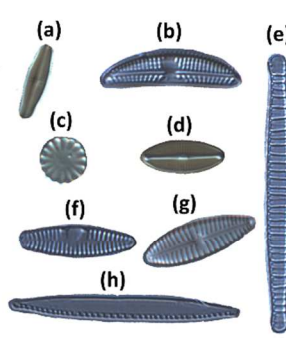
Our goal was to find the optimal method for DNA extraction when using HTS methods as a step toward standardizing the application of diatom metabarcoding. DNA extraction has 3 main requirements to: 1) obtain good quality DNA and sufficient DNA quantity, 2) obtain inhibitor-free DNA for subsequent molecular biological analyses, and 3) ensure representative lysis of all organisms (in our case, the different diatom species) in the sample. We compared 5 methods of DNA extraction in combination with HTS metabarcoding. These methods combined various types of cell lysis and DNA purification. We tested the 5 methods on 8 pure cultures of diatoms and 8 freshwater samples of benthic diatom communities from streams and lakes with differing water quality and geographical origin. We based our comparison on the following criteria: 1) DNA extraction efficiency (quantity of DNA), DNA quality, and presence of inhibitors in extracted DNA, 2) the diatom community structure as revealed by HTS sequencing of the ribulose-1, 5-bisphosphate carboxylase (*rbcL*) barcode (qualitative and quantitative comparisons were performed at different taxonomic levels), and 3) comparison of molecular and microscopy-based inventories in terms of community composition and inferred water-quality indices.

3. Methods

3.1. Diatom cultures

We selected 8 pure cultures of diatoms from the Thonon Culture Collection (TCC; http://www6.inra.fr/carrtel-collection_eng/) based on their contrasting morphological and phylogenetical features. These strains were cultured in 300 mL sterile DV media, as previously described (Rimet *et al.* 2014) (**Figure 19A**). From each diatom culture, we prepared a 20-mL aliquot containing 105 to 106 cells and froze the aliquot at -80°C until further analysis.

A)					
Code	Species	TCC code	Width range (μm)	Length range (μm)	Pictures
AM	<i>Achnantheidium minutissimum</i>	TCC 667	1.5-3.3	5.6-20.8	(a)
AP	<i>Amphora pediculus</i>	TCC 702	2.5-4	6-16	(b)
CMEN	<i>Cyclotella menegheniana</i>	TCC 690	diameter = 5-43		(c)
CMOL	<i>Craticula molestiformis</i>	TCC 459	3.4-4.9	12-15	(d)
DT	<i>Diatoma tenuis</i>	TCC 861	2.9-4.9	20-85	(e)
FP	<i>Fragilaria perminuta</i>	TCC 753	3-4	7-40	(f)
GP	<i>Gomphonema parvulum</i>	TCC 492	4-8	10-46	(g)
NP	<i>Nitzschia palea</i>	TCC 139-1	3-4	12-42	(h)



B)					
Code	Sampling Site	Sampling date	Geographical area	Physico-chemical characteristics	Site quality
Edian	Stream Edian	11-2014	France	Clear water, thick biofilm	good
Aire	Stream Aire	11-2014	France	Copper	polluted
Lake	Lake Geneva	11-2014	France	Thick biofilm	good
767	Stream Dammån	09-2013	Sweden	Acid pH, humic acid	good
M36	Agricultural stream	10-2014	Sweden	Pesticides, nutrients	polluted
P45	Lake Båtkåjåure	09-2009	Sweden	Clear water (mountain)	good
Ref7	Stream Dapani	11-2014	Mayotte	Clear water, thin biofilm	good
Pol2	Stream Majimbini	11-2014	Mayotte	Organic matter, detergent	highly polluted

Figure 19 – Characteristics of the diatom cultures from the Thonon Culture Collection (TCC) (A) and biofilm sampling sites (B).

Pictures transformed from the R-syst::diatom database (length not to scale).

3.2. Environmental community samples

Eight environmental community samples were collected from benthic biofilms at 6 streams and 2 lakes in 3 geographical areas (Sweden, France, and Mayotte, a French Tropical Island) (**Figure 19B**). We selected the sampling sites for their contrasting geographic origin, water-quality status (polluted to good quality), and physicochemical characteristics (concentration of organic matter and presence of metals or pesticides). These characteristics were chosen because

they can affect DNA extraction from the prevailing diatom assemblages. All environmental samples were collected following the European Water Framework Directive standards (AFNOR 2003) by scraping material from the surface of ≥ 5 submerged stones. The resulting material was transferred to 15 mL Falcon tubes and fixed by immediately adding 99% ethanol to reach a final ethanol concentration of ~ 70 – 80% . Ethanol fixation prevents grazing by metazooplankton and allows good preservation of DNA (Motwani & Gorokhova 2013). Fixed environmental samples were stored at room temperature under dark conditions until preparation for morphological analysis and DNA extraction.

We estimated diatom valve concentration in samples based on microscopic counts. Each diatom skeleton is composed of 2 valves. We used the formula:

$$N = \text{number of valves counted} \times \frac{R}{M} \quad (\text{Eq. 1})$$

$$\text{where } R = \frac{\text{cover slip area (mm}^2\text{)}}{\text{microscopic counting area (mm}^2\text{)}}, \quad (\text{Eq. 2})$$

N = number of valves/mg of sample, R = counting ratio, and M = quantity of sample fixed on slide (mg).

3.3. DNA extraction

We centrifuged environmental samples and pure culture subsamples at 13,000 rpm for 30 min and removed the supernatant. We used 25 mg of wet pellet as a starter for DNA extraction for each environmental sample. The quantity corresponded to the smallest amount of starting material recommended for the selected DNA extraction methods and is the usual environmental sample amount used for DNA extraction.

We extracted DNA in triplicate from each diatom culture and each environmental sample with 4 commercial DNA extraction kits : Macherey–Nagel (Düren, Germany) NucleoSpin® Soil kit (MN-Soil) Macherey–Nagel NucleoSpin® Plant II kit (MN-Plant), Stratec (Birkenfeld, Germany) Invisorb® Spin Plant Mini Kit (S-Plant); Qiagen (Hilden, Germany) DNeasy® Blood and Tissue kit (Q-Blood), and 1 non-kit protocol based on Sigma–Aldrich (St Louis, Missouri) GenElute™-LPA DNA precipitation (SA-Gen), which was used in previous studies (Kermarrec *et al.* 2013a; Chonova *et al.* 2016). These 5 DNA extraction methods have been used or recommended for use to extract DNA from freshwater algae (Nguyen *et al.* 2011; Eland *et al.* 2012; Kermarrec *et al.* 2013a; Zimmermann *et al.* 2015) and were chosen based on their various types of lysis (mechanical, enzymatic, thermal) and the use or not of columns to remove contaminants/co-extracted

molecules (**Figure 20**). SA-Gen was the only method that did not include a column purification step. We ran all protocols according to the manufacturer's instructions (**Figure 20**) with a single modification for MN-Plant where we changed incubation time at 65°C from 10 to 45 min (manufacturer's recommendation for difficult plant material).

The final elution volume was 40 µL for all DNA extraction methods. We conducted a total of 96 DNA extractions for diatom cultures (MN-Soil method not tested) and 120 for environmental samples (all 5 extraction methods tested).

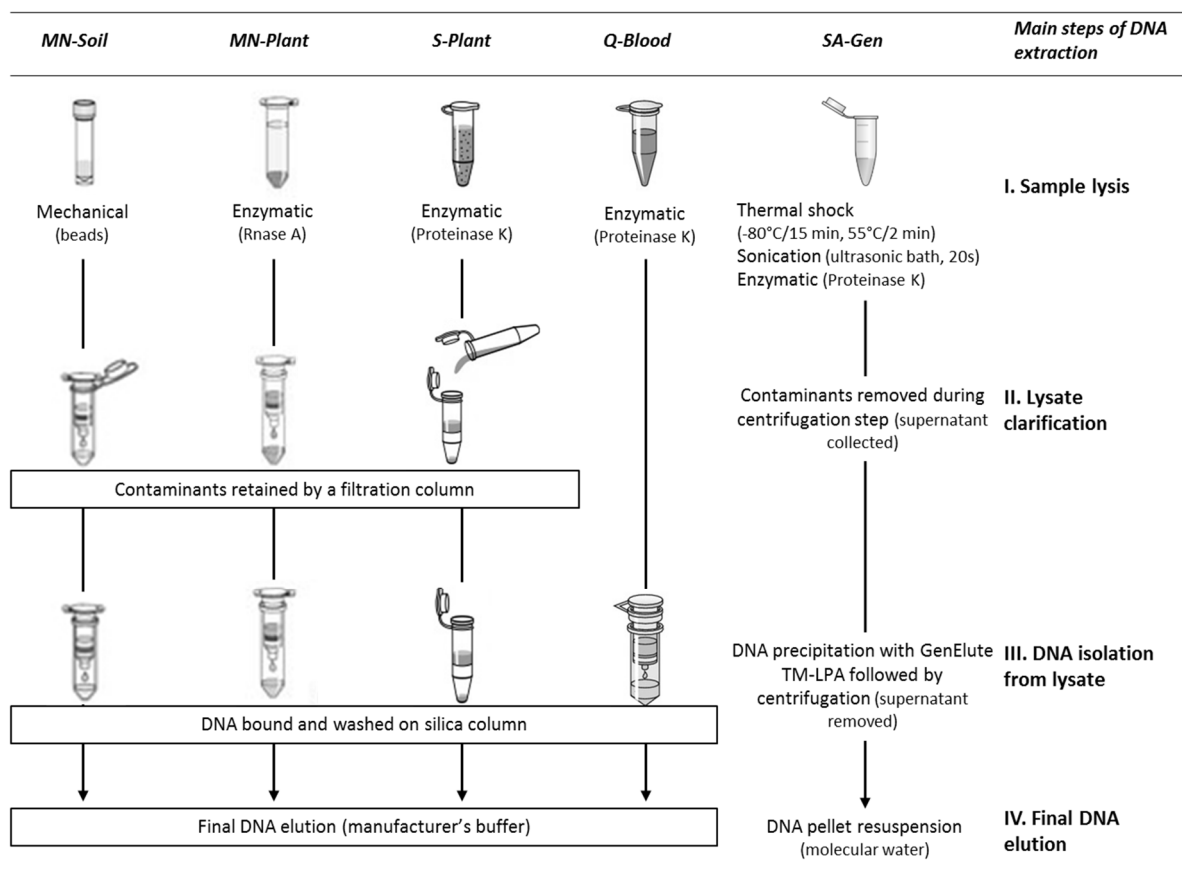


Figure 20 – The main steps of DNA extraction are presented for the 5 methods. Focus on sample lysis (I), lysate clarification (II), DNA isolation from lysate (III), and DNA elution (IV). Pictures modified from the manufacturers' web sites.

3.4. Evaluation of DNA extraction efficiency and DNA quality

For all samples, we quantified the extracted DNA with the Life Technologies (Carlsbad, California) Quant-iT™ PicoGreen® dsDNA assay kit using a microplate reader (Fluoroskan Ascent™ FL; Thermo Scientific, Waltham, Massachusetts) and following the manufacturer's instructions. To compare DNA extraction efficiency among methods, we normalized DNA concentrations as µg DNA/g wet biofilm for environmental samples and as µg DNA/10⁴ cells for

diatom cultures. We assessed DNA quality by spectrophotometry with 260/280 nm ratio with the Nanodrop®ND-1000 (Nanodrop Technologies, Wilmington, Delaware).

We compared mean values for DNA quantities and qualities based on the Kruskal–Wallis group test followed by the Mann–Whitney pairwise test to evaluate the effect of the different extraction methods on these parameters. These statistical analyses were performed in R (version 3.0.2; R Project for Statistical Computing, Vienna, Austria).

3.5. Polymerase Chain Reaction (PCR) inhibitor detection (quantitative PCR [qPCR])

We estimated the presence of inhibitors by making serial dilutions of the DNA extracts and estimating *rbcL* copy numbers via qPCR for every dilution (Gallup & Ackermann 2006; Lloyd *et al.* 2010). In this approach, inhibitors are assumed to be diluted with a log-linear relationship between cycle threshold (Ct) and the dilution factor (DF). Ct values obtained for a 10-fold dilution of the same sample have a theoretical difference of 3.3 cycles when considering 100% PCR efficiency. The presence of PCR inhibitors co-extracted with DNA reduces PCR efficiency and affects this expected value of 3.3, allowing detection of these inhibitors. We performed qPCR assay on serial dilutions (10^0 – 10^{-3}) of 1 DNA extraction replicate per environmental sample and DNA extraction method (corresponding to 40 environmental DNA extracts). The level of inhibition was estimated by calculating for each dilution level the dilution factor (DF) needed to remove all inhibition effects as $DF = 10^x$, where $x = (\text{theoretical Ct} - \text{measured Ct})/\text{standard curve slope}$ (transformed from Gibson *et al.* 2012), measured Ct = Ct obtained during assay for each dilution level, and theoretical Ct = expected Ct value for the dilution without inhibition.

We estimated the theoretical Ct for each assay, and it generally corresponded to the highest dilution (10^{-3}). We considered samples with $DF \leq 2$ as not inhibited, values with $2 < DF \leq 10$ as weakly inhibited, $10 < DF \leq 100$ as strongly inhibited, and $DF > 100$ as very strongly inhibited. We conducted qPCR targeting a short region of the *rbcL* plastid gene (312 base pairs [bp]; same region was used for HTS sequencing) in a Rotor Gene RG-3000 (Corbett Research, Sydney, Australia) with 2 replicates using the QuantiTect SYBR Green PCR Kit (Life Technologies). The mix (25 μ L final volume) contained: 12.5 μ L of master mix provided by the supplier, 1.25 μ L of 10 μ M forward primer Diat_ *rbcL*_708F (AGG TGA AGT TAA AGG TTC ATA CTT DAA) (Stoof-Leichsenring *et al.* 2012) and reverse primer R3 (CCT TCT AAT TTA CCA ACA ACT G) (Bruder & Medlin 2007), 1.25 μ L of 10 g/L bovine serum albumin (BSA), 2 μ L of extracted DNA, and 6.75 μ L H₂O (molecular

biology grade). Reaction conditions were: initial denaturation of DNA at 95°C for 15 min followed by 40 cycles with 45 s denaturation at 95°C, followed by 45 s annealing at 55°C and 45 s extension at 72°C. We used 1 no-template control (NTC) as a negative control.

We standardized qPCR assays by adding serial dilutions (7 points) of standard DNA with known [DNA] and known copy number of the *rbcL* fragment. This reference DNA was prepared with plastid DNA of *Nitzschia palea* following 4 main steps: 1) amplification with Diat_ *rbcL*_708F/R3 primers, 2) insertion of the *rbcL* 312 bp amplicon produced into TOPO plasmid and cloning into *Escherichia coli* bacteria using the TOPO TA cloning kit (Invitrogen, Carlsbad, California), 3) purification and extraction of plasmids with insert from positive clones using the QIAprep Spin Miniprep kit (Qiagen), 4) evaluation of plasmid DNA concentration using the Picogreen method (as described above); this concentration was considered as 100 dilution level.

We analyzed the data with Rotor-gene 6 (version 6.1; Corbett Research) with a fluorescence threshold of 0.3 for denoising and determining Ct. The results served both for detection of inhibitions and quantification of *rbcL* gene in environmental samples to provide a quantitative comparison between qPCR estimations and microscopic counts.

3.6. Preparation of the library of amplicons and HTS sequencing

For all environmental samples, we conducted HTS sequencing of the *rbcL* 312 bp fragment on 2 of the 3 DNA replicates from each extraction method. For each DNA sample, we ran the PCR amplification in triplicate on 1 µL of extracted DNA in a mixture (25 µL final volume) containing: 0.75 U of Takara LA Taq® polymerase (TaKaRa Bio, Sugats, Japan), 2.5 µL of 10× buffer, 1.25 µL of 10 µM of primers Diat_ *rbcL*_708F and R3, 1.25 µL of 10 g/L BSA, 2 µL of 2.5 mM deoxynucleotide (dNTP), and completed with 15.6 µL H₂O (molecular biology grade). PCR reaction conditions were the same as those used for qPCR (see above) with 30 cycles. Seventy-eight of the 80 DNA extracts were amplified successfully, and the 2 replicates extracted from Ref7 sample with SA-Gen method were not amplified.

For each DNA extract, we pooled the 3 replicates of PCR product and then cleaned with Agencourt AMPure beads (Beckman–Coulter, Brea, California) following the manufacturer's instructions except one modification regarding the beads/DNA ratio, which we adjusted to 1.5:1. We assessed purified amplicons for quality and quantified them using the 2200 TapeStation (Agilent Technologies, Santa Clara, California) with D1000 screen tape and reagents. We used the 78 purified amplicons to prepare 78 DNA libraries for HTS with Ion Torrent technology using the NEBNext® Fast DNA Library Prep set for Ion Torrent™ (BioLabs, Ipswich, Massachusetts),

following the manufacturer protocol for End repair, PCR amplification of adapter ligated DNA (7 cycles), and cleaning steps. Ligation of library adapters to purified amplicons was done with 2 μ L of P1 adapter (NEB kit) and 2 μ L of A-X tag adapter provided in Ion Express™ Barcode adapters (Life Technologies) using 1 tag per amplicon.

We checked the quality, size, and concentration of the libraries with the 2200 TapeStation with D1000 High Sensitivity screen tape and reagents. We diluted each library to 100 pM and pooled all of them together in a unique mixture that was sequenced using 1 Ion 318™ Chip Kit V2 (Life Technologies) on a PGM Ion Torrent machine by the Plateforme Génome Transcriptome (PGTB, Bordeaux, France).

3.7. Sequence data processing

Demultiplexing and adapter-removal steps were made by the Sequencing Platform, which provided a single fastq file for each of the 78 libraries. DNA reads were filtered for length and quality using Mothur software (Schloss *et al.* 2009) in every fastq file with the following settings: minimum length = 250 bp, Phred quality score >23 over a moving window of 25 bp, maximum of 1 mismatch in forward primer sequence, homopolymers <8 bp, and absence of ambiguous base. Reads that were not fully aligned with the *rbcl* barcode were removed. The 78 resulting files were analyzed together. Denoising of sequencing error was performed with the Precluster command by creating read clusters, allowing 1 nucleotide difference between DNA reads. Chimera removal was done using the Uchime algorithm (Edgar *et al.* 2011). The potential effect of the DNA extraction method or the sampling site on read abundances was assessed with 2-way analysis of variance.

We used the R-Syst::diatom database (Rimet *et al.* 2016, version updated in January 2015, <http://www.rsyst.inra.fr/en>) restricted to our 312-bp *rbcl* barcode as a reference database. Taxonomic assignment of DNA reads at the species level was made using this reference database and the naïve Bayesian method (Wang *et al.* 2007) with a confidence score threshold of 85%. Only DNA reads assigned to Bacillariophyta (diatoms) were used in further analysis.

We conducted a dereplication step and calculated uncorrected pairwise distances between aligned reads (alignment performed using the align.seqs command in Mothur with the algorithm proposed by Needleman & Wunsch 1970 and default setting) to generate a similarity distance matrix. Based on this distance matrix, reads were clustered in operational taxonomic units (OTUs) using the furthest-neighbor algorithm at a 95% similarity level. This similarity level was reported as a relevant cut-off threshold for OTU delineation that limits artificial inflation of eukaryote OTUs

(following recommendations by Mangot *et al.* 2013). Singletons were removed, and all samples were normalized to the smallest read abundance obtained among the 78 libraries for further analysis (**Fig. S1**).

Taxonomy was assigned to OTUs on the basis of the consensus taxonomy of reads (application of the `classify.otu` command from Mothur) (Schloss *et al.* 2009) with a stringent consensus confidence threshold (>80%) (**Fig. S1**). OTU α diversity was estimated in Mothur with the Chao1 estimator as a global richness estimator and Shannon index as diversity estimator.

3.8. Statistical analysis on community structure as revealed by HTS

We used the Kruskal–Wallis test to evaluate the effect of extraction method on Chao and Shannon indices. We compared community compositions of the 78 DNA extracts at the OTU and species levels. The OTU list represents the whole DNA reads that were clustered at 95% similarity level, whereas the species list takes into account only OTUs for which the taxonomic assignment was good enough to provide identification at the species level. We used the OTU or species lists to compute Bray–Curtis dissimilarity indices, which we visualized using nonmetric multidimensional scaling (NMDS). We used permutational ANOVA (PERMANOVA) (PRIMER-E, Plymouth, UK) to compare similarity between DNA extraction methods within and between the 8 environmental samples and similarity percentage (SIMPER) (PRIMER-E) analyses to detect which OTUs were the main contributors to the dissimilarity.

3.9. Comparison between molecular and morphological taxonomic inventories

We based morphological taxonomic inventories of environmental samples on diatom valves according to the European Committee for Standardization (AFNOR 2004). We counted a minimum of 400 valves with the aid of a light microscope with 1000× magnification and identified them based on classical European floras for French and Swedish samples (*e.g.*, Krammer & Lange-Bertalot 1986, 1988, 1991a; b, Krammer 2000, 2001, 2002, 2003) and specific literature for Mayotte tropical samples (*e.g.*, Bourrelly & Manguin 1952; Metzeltin & Lange-Bertalot 1998, 2007; Tudesque *et al.* 2008).

We compared diatom taxonomic inventories obtained by the molecular approach to those obtained by the morphological approach at the species and genus levels. We used OMNIDIA 5

software (Lecointe *et al.* 1993, library 5.3 2015) to calculate and compare the Specific Pollution-sensitivity Index (SPI) (Cemagref 1982) based on species lists (or genus if species level was not reached) obtained by PGM sequencing or by microscopy for each environmental sample.

We also calculated valve and *rbcl* gene copy numbers per mg of wet biofilm from microscopic count and qPCR assay and calculated the ratio [valve]/[*rbcl* copy].

4. Results

4.1. DNA extraction efficiency, quality, and PCR inhibition

DNA extraction efficiency differed significantly among methods either for diatom cultures ($p < 0.001$) and environmental samples ($p < 0.001$). The SA-Gen method yielded the highest quantity of DNA for both environmental samples and diatom cultures, whereas the lowest DNA quantities were obtained with the MN-soil method for environmental samples and S-Plant for diatom cultures (**Table 1**). All methods yielded good DNA quality (260/280 ratios: 1.7–2.0) with diatom cultures and environmental samples, but environmental samples extracted with MN-Plant method had a slightly lower value (1.5) (**Table 1**).

	DNA yield		260/280 ratio	
	Pure culture ($\mu\text{g}/10^4$ cells)	Environmental sample ($\mu\text{g}/15$ mg biofilm)	Pure culture	Environmental sample
MN-Soil	Ø	1.6 (1.5) ^{ab}	Ø	1.8 (0.5) ^a
MN-Plant	3.9 (5.7) ^a	2.1 (3.1) ^a	1.8 (0.2) ^a	1.5 (0.4) ^b
S-Plant	3.8 (4.9) ^{ab}	5.6 (6) ^c	2 (0.1) ^b	1.8 (0.3) ^a
Q-Blood	5.9 (7.1) ^{bc}	3.7 (4.6) ^b	1.7 (0.1) ^c	1.9 (0.4) ^a
SA-Gen	8.5 (16.8) ^c	20.8 (22.5) ^d	1.9 (0.1) ^d	ND

Table 1 – Mean (SD, n = 72) DNA extraction efficiency and 260/280 DNA ratios obtained for the 5 extraction methods with the pure culture and environmental samples.

ND = not determined (out of range), Ø = missing values. Values sharing the same letter are not statistically different.

Inhibition levels for each environmental sample and DNA extraction method were estimated from qPCR assays (**Table S1**). DNA samples extracted with the SA-Gen method presented the highest level of PCR inhibition (**Table 2**), whereas DNA obtained with the MN-Soil method was easily amplified without DNA dilution and was free of inhibition. MN-Plant, S-Plant,

and Q-Blood extracts were slightly inhibited, and a 10-fold dilution was sufficient to remove inhibition (with 1 exception for the MN-plant method on the Ref7 sample).

	MN-Soil	MN-Plant	S-Plant	Q-Blood	SA-Gen
Edian	-	+	-	-	+++
Aire	-	+	-	+	++
Lake	-	+	+	+	++
767	-	ND	ND	ND	ND
M36	-	-	-	-	+++
P45	ND	ND	ND	ND	ND
Ref7	-	++	+	+	+++
Pol2	-	+	+	-	++

Table 2 – Mean (SD, n = 72) estimation of the inhibition level for DNA extracted from the environmental samples.

– = not inhibited, + = weakly inhibited, ++ = strongly inhibited, +++ = very strongly inhibited, ND = not determined (out of range) (see Table S1 for corresponding details).

4.2. Comparison of diatom quantification: microscopy vs qPCR

rbcl copy number/mg sample was calculated during qPCR assay for all the environmental samples and DNA extraction methods and compared to valve number/mg sample (except for 767 and P45 samples, which were out of range for qPCR assay) (**Table S2**). The ratio between *rbcl* copy and valve numbers showed that the *rbcl* concentration was mostly (26 of 30 cases) below the valve concentration. *rbcl* concentrations obtained with the SA-Gen method provided the best correspondence compared to the valve concentration (**Table S2**).

4.3. Effect of extraction methods on richness, composition, and structure of diatom community

After DNA sequencing of the 78 libraries, a total of 4,711,673 of DNA reads was obtained with an average read length of 271 bp. After trimming and removing singletons, 967,089 DNA reads (20.5% of the initial number) were conserved and clustered into 3293 OTUs at a 95% similarity level (**Table S3**). DNA extraction method did not affect the total number of reads obtained after the bioinformatics process (2-way ANOVA, $p = 0.1$), but sampling site did ($p = 0.008$). The smallest average number of reads was obtained for Pol2 (5183 reads) and the

maximum was obtained for Ref7 (16,831 reads). We normalized read abundances for each sample to 4180 (lowest read number obtained for sample Pol2 with kit MN-plant; **Table S3**).

The values obtained for the Chao global richness estimator varied widely among environmental samples (94–436 OTUs). The Chao global richness estimator and Shannon diversity index values did not differ among DNA extraction methods (**Table S4, S5**).

The NMDS based on OTUs similarity showed that the 78 DNA samples were discriminated mainly according to sampling site (**Figure 21A**). Sampling site explained 90% of the total variance (PERMANOVA, $R^2 = 0.90$, $p < 0.001$), whereas DNA extraction method explained only 1.2% ($R^2 = 0.012$, $p < 0.001$). Site-by-site analysis with PERMANOVA showed that the DNA extraction method explained 72 to 91% of the total variance in 6 environmental samples, whereas no significant effect was assessed for 2 samples (**Figure 21B**).

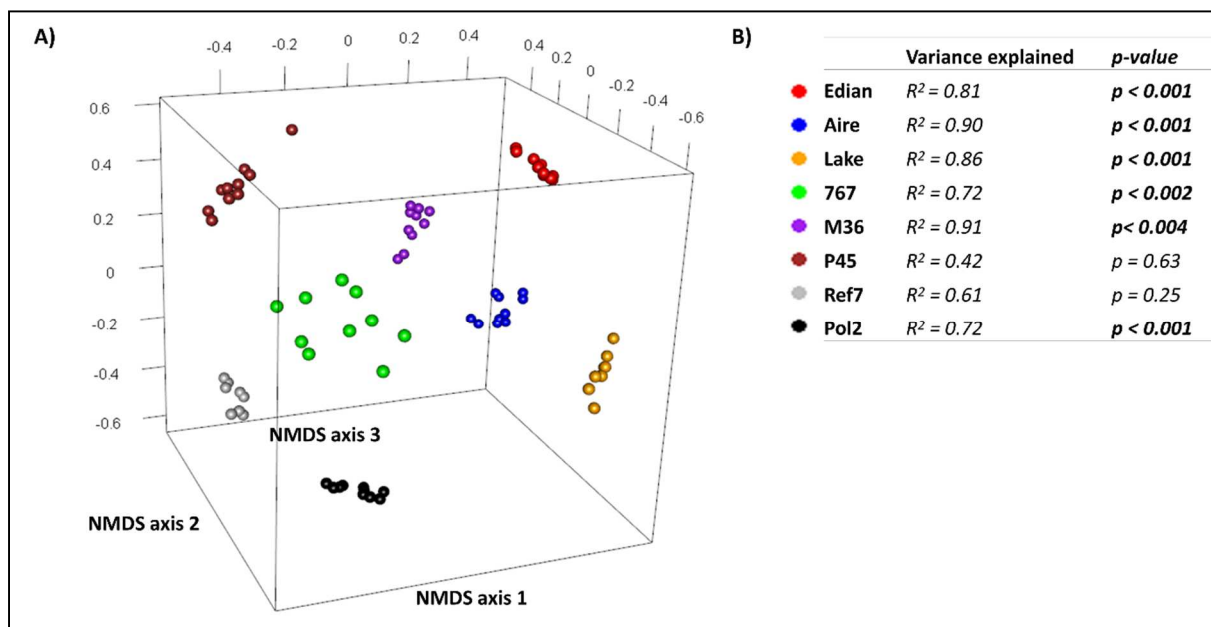
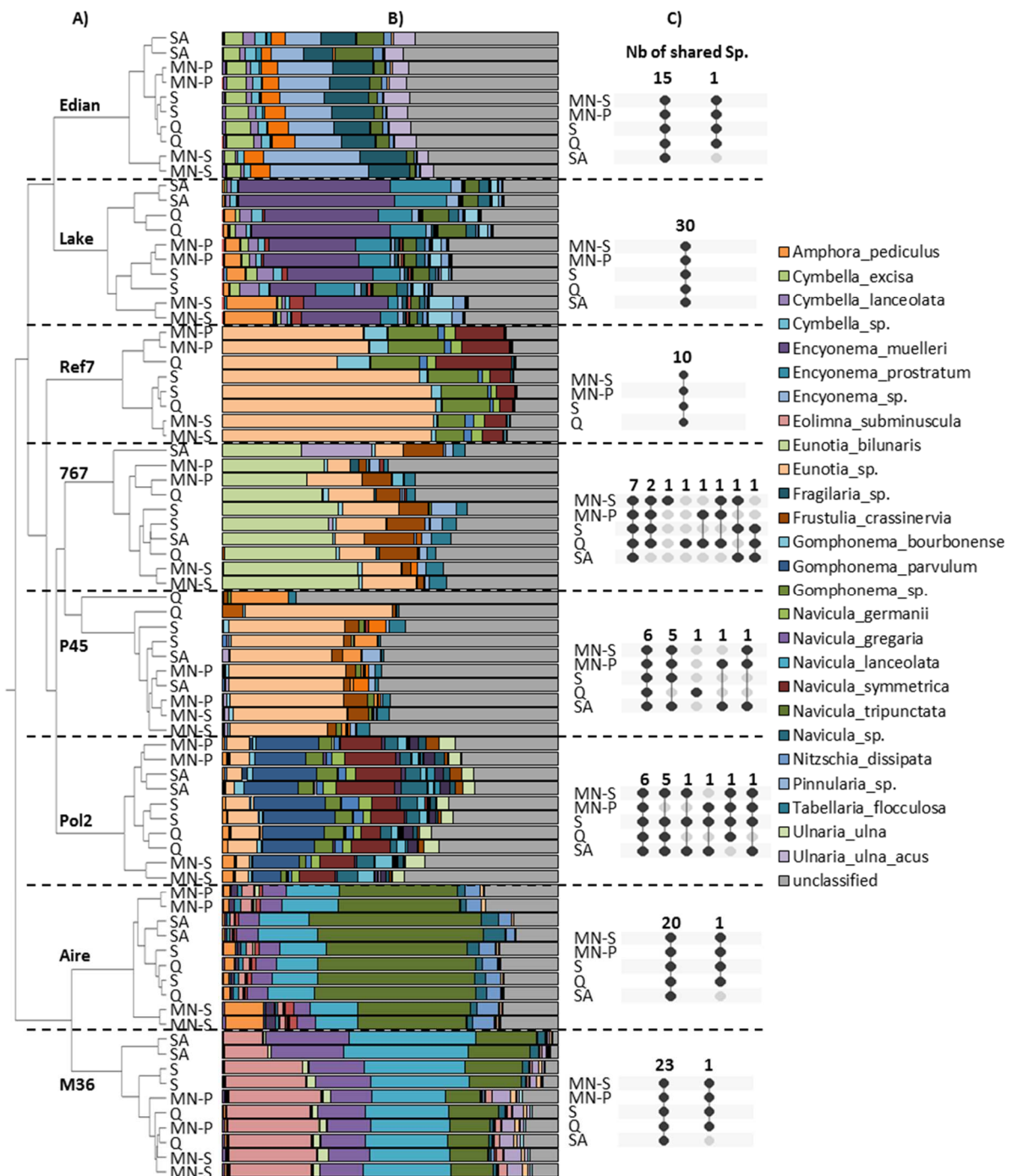


Figure 21 – Dissimilarity between molecular inventories obtained with the 5 DNA extraction methods at 8 sampling sites.

A. Three-dimensional nonmetric multidimensional scaling (NMDS) plot of Bray–Curtis dissimilarity based on operational taxonomic unit (OTU) composition of the DNA extracts obtained from the 8 environmental samples (extraction performed in duplicate for the 5 extraction methods), stress value = 0.18. **B.** Results obtained by permutational analysis of variance (PERMANOVA) to reveal the effect of the DNA extraction method on dissimilarity values within sites.

We conducted further analyses focused on the SA-Gen and MN-Soil methods, which provided the most different results in terms of quantity, quality of DNA, and community structure (**Figure 22A**). We used SIMPER to identify the OTUs that were the main contributors (contribution

> 1%) to the dissimilarity between communities assessed using the SA-Gen and MN-Soil methods (Table 3).



Among the 6 environmental samples, 38 OTUs were identified as main contributors. Differences between the 2 methods were mostly (in 98% of cases) a result of variations in the read abundances of common OTUs rather than the presence or absence of OTUs detected by only 1 of the 2 methods. OTUs assigned to the genera *Nitzschia* and *Amphora* were represented by more reads when we used the MN-Soil method, whereas OTUs belonging to the genera *Encyonema*, *Gomphonema*, and *Navicula* were represented by more reads when we used the SA-Gen method (Table 3).

Family	Genus	OTU	Edian	Aire	Lake	767	M36	Pol2
			SA/MN reads (ratio)	SA/MN reads (ratio)	SA/MN reads (ratio)	SA/MN reads (ratio)	SA/MN reads (ratio)	SA/MN reads (ratio)
Achnanthidiaceae	<i>Planothidium</i>	n°043	-	-	-	-	31/163 (0.2)	-
Amphipleuraeae	<i>Frustulia</i>	n°031	-	-	-	382/52 (7.3)	-	-
Bacillariaceae	<i>Nitzschia</i>	n°021	-	115/226 (0.5)	-	-	-	-
		n°088	-	-	-	-	-	42/141 (0.3)
Catenulaceae	Unclassified	n°188	-	-	-	-	-	2/102 (0.2)
		n°027	-	-	21/564 (0.04)	-	-	-
		n°039	-	19/374 (0.05)	-	-	-	-
Cymbellaceae	<i>Encyonema</i>	n°113	-	-	-	-	-	22/146 (0.2)
		n°004	-	-	1840/992 (1.9)	-	-	-
		n°013	-	-	553/141 (3.9)	-	-	-
		n°065	53/163 (0.3)	-	-	-	-	-
		n°070	-	-	104/20 (5.2)	-	-	-
Eunotiaceae	<i>Eunotia</i>	n°128	-	-	88/1 (88.0)	-	-	-
		n°010	-	-	-	804/1199 (0.7)	-	-
		n°036	-	-	-	81/408 (0.2)	-	-
		n°038	-	-	-	217/68 (3.2)	-	-
		n°048	-	-	-	193/280 (0.7)	-	-
Fragilariaceae	<i>Fragilaria</i>	n°083	-	-	-	345/0	-	-
		n°017	298/423 (0.7)	-	-	-	-	-
Gomphonemataceae	<i>Pseudostauropis</i>	n°047	-	-	81/264 (0.3)	-	-	-
		n°012	-	-	-	-	-	129/38 (3.4)
		n°041	-	-	-	-	-	365/229 (1.6)
Naviculaceae	Unclassified	n°044	-	-	-	-	-	323/214 (1.5)
		n°025	-	-	-	542/826 (0.7)	-	-
		n°001	386/71 (5.4)	2008/1321 (1.5)	-	-	733/480 (1.5)	-
		n°003	-	581/454 (1.3)	-	-	1298/816 (1.6)	-
		n°008	-	-	-	-	623/361 (1.7)	-
		n°009	-	-	-	-	-	613/414 (1.5)
Pinnulariaceae	<i>Caloneis</i>	n°018	-	-	-	-	243/122 (2.0)	-
		n°046	-	195/15 (13.0)	-	-	-	-
Sellaphoraceae	<i>Eolimna</i>	n°063	-	22/114 (0.2)	-	-	-	-
Skeletonemataceae	Unclassified	n°005	-	-	-	-	467/1029 (0.5)	-
		n°085	-	-	21/151 (0.1)	-	-	-
Unclassified	Unclassified	n°022	-	-	90/298 (0.31)	-	-	-
		n°024	-	-	-	-	31/227 (0.1)	-
		n°034	-	-	-	-	-	171/567 (0.3)
		n°055	-	-	39/160 (0.2)	-	-	-
		n°067	-	-	-	-	-	69/257 (0.3)

Table 3 – Results of the similarity percentages (SIMPER) analysis performed to identify the operational taxonomic units (OTUs) contributing to >1% of the dissimilarity between diatom communities obtained from the SA-Gen (SA) and MN-Soil extraction methods (MN). The list of contributors is presented for each environmental sample. Read abundances obtained with the 2 methods (SA/MN reads) and the ratios of read abundances for each OTU are presented. – = no contribution was found or the OTU was absent from the sample.

Some OTUs could not be assigned at the species level. The proportion of DNA sequences that remained unclassified at the species level varied from 1.5% (sample M36) to 78% (sample P45). On average, considering all sampling site and extraction methods, 71% of the reads were assigned to the species level. The comparison of the species inventories from the 78 libraries (**Figure 22A**) showed, as previously observed for the OTUs, that samples clustered primarily by sampling site. Community structures based on species composition were similar among methods for each sample except samples 767 and P45 (**Figure 22B, C**) for which we suspected potential bias during the initial subsampling (small sample volume [52µL] used for DNA extraction with sample 767 and difficulty homogenizing sample P45).

4.4. Morphology vs molecular diatom community composition

The taxonomic lists obtained with the HTS approach for each environmental sample with each of the extraction methods were compared at genus and species levels to those obtained with the classical microscopy-based approach (**Figure 23**). In general, 43% of the genera (maximum = 61.5% for Edian sample) and 18% of the species (maximum = 34.5% for sample M36) were detected by both approaches. Sixty-three percent of species were detected only by microscopy (on average for all samples), whereas only 19% of specific-HTS species were observed. However, a very high number of OTUs could not be assigned to a precise species because the reference database was incomplete (68% of species detected only by microscopy were not represented in the database).

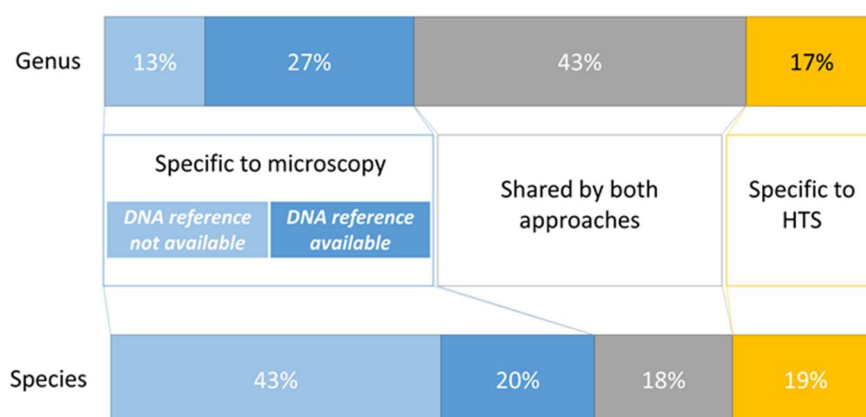


Figure 23 – Mean percentage of diatom genera and species detected by microscopy, by molecular inventory, or by both methods for the 8 sampling sites.

For all genera and species detected by only microscopy, the presence/absence of their DNA reference in the molecular database is specified. Unclassified operational taxonomic units were not used. HTS = high throughput sequencing.

SPI values calculated based on diatom lists identified by HTS and by microscopy were compared (**Figure 24**). SPI values were consistent with expected water-quality status (**Figure 19B**) for both French and Swedish sites. The SPI has not been adapted for Mayotte Island yet and cannot be used to infer quality status there. Different DNA extraction methods provided similar SPI values, which were close to SPI values obtained by microscopy except for Mayotte samples (Pol2 and Ref7).

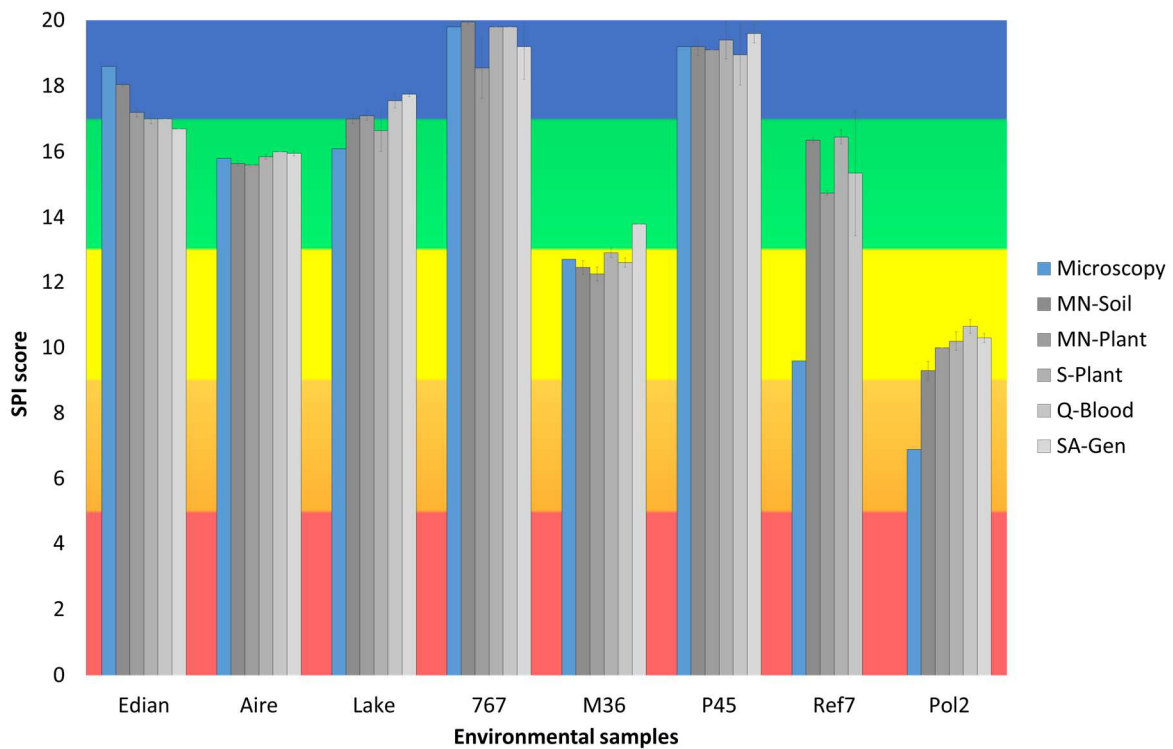


Figure 24 – Specific pollution-sensitivity index (SPI) values based on morphological inventories (counts obtained from microscopic observations of diatoms valves) and on molecular inventories (reads based on relative abundances estimated by high throughput sequencing [HTS]). When species taxonomic level was not reached, ecological values at the genus level were used for the calculation. Colors correspond to water-quality thresholds used inferred from the SPI, from red (worst quality) to blue (best quality).

5. Discussion

5.1. DNA extraction method affects quantity and quality of extracted DNA

The highest DNA extraction efficiency was observed for both diatom pure cultures and biofilm samples with the SA-Gen method, which outperformed the 4 commercial DNA extraction

kits in terms of DNA quantity. Elution parameters (*e.g.*, elution volume, temperature) are known to affect DNA yield when commercial kits are used, and yield loss can range from 20 to 30% (according to the manufacturer's specifications). However, the difference of efficiency between the commercial kits and the SA-Gen method is much higher than this % variation, and elution conditions alone fail to explain the low DNA concentrations obtained with the 4 commercial kits compared to the SA-Gen method.

We think the lysis method particularly affected DNA extraction efficiency, as previously suggested by Deiner *et al.* (2015) for eubacteria and freshwater eukaryotes. Various lysis methods including freezing–thawing (Fuhrman *et al.* 1988), enzymes (Somerville *et al.* 1989), liquid N (Bruckner *et al.* 2008), sonication (Chung *et al.* 2005), or bead beating (Yuan *et al.* 2015) can be used to disrupt the cell wall prior to DNA extraction. Bead beating, as presented in MN-Soil, has been used in metabarcoding studies of benthic eukaryotic communities, including diatoms, because it saves time and works with complex environmental matrices like sediments or biofilms (Zimmermann *et al.* 2015; Chariton *et al.* 2015). However, diatom cells are protected by a robust silica valve that limits the ability of these classic lysis methods, even bead beating (Eland *et al.* 2012), to disrupt the diatom cell and release DNA. The SA-Gen method combines different lysis mechanisms (sonication, enzyme, temperature variation) to recover high quantities of DNA from both pure cultures and environmental samples. For pure cultures, the quantity of DNA collected was >2× higher with the SA-Gen protocol than with the other methods tested. We also observed higher efficiency of SA-Gen for environmental samples, but we could not assess precisely which part of this total DNA was from diatoms because DNA from other organisms (bacteria or other microbes present in the biofilms) was co-extracted. However, DNA extracted by SA-Gen from environmental samples provided the highest quantity of *rbcl* copy (per mg of wet biofilm), as revealed by qPCR with our diatom-specific primers, and the best quantitative correlation to diatom valve counted by microscopy.

SA-Gen is an in-house method that does not include a silica column during the purification step. All the other protocols applied in our study include a silica column, so we assume that part of the DNA could have been lost by remaining fixed to the purification column. The effect of DNA purification methods on DNA recovery has been studied for soils and sediment samples (Miller *et al.* 1999). In these studies, use of a column reduced recovery to 80%, and in some cases (*e.g.*, sediment samples), as low as 40% of initial DNA concentrations. In our study, the 4 methods that included column purification (MN-Soil, MN-Plant, S-Plant, and Q-Blood) reached a maximum

efficiency of 69% with diatom pure cultures and only 27% with environmental samples relative to the SA-Gen method.

The DNA purification steps could be useful for DNA originating from environmental matrices that may contain a number of compounds that inhibit or decrease the sensitivity of PCR. PCR-inhibitor molecules in extracted DNA can come from residual compounds of the extraction process (*e.g.*, ethanol) or from molecules that are co-extracted with DNA (*e.g.*, protein, polysaccharides, humic acids) (Schrader *et al.* 2012). Based on the 260/280 ratios, we assume that the quality of our DNA extracts was not affected by protein contamination (for both pure culture and environmental samples) or by residual contamination of the extraction (for pure culture), regardless of which extraction method was used. The main source of inhibition was co-extracted environmental compounds. Based on real-time PCR results, we were able to estimate the level of inhibition present in all the DNA extracts from environmental samples. The SA-Gen method produced extracts with the highest inhibition level, whereas extracts produced with methods that included a column purification step were only slightly inhibited. Making serial dilutions of DNA extracts to reach a concentration of inhibitors that is low enough not to inhibit PCR reactions is an efficient strategy to overcome the problem of PCR inhibitors, but this approach requires a large initial amount of DNA so that diluting it does not affect its representativeness. SA-Gen produced a large-enough quantity of DNA to permit improvement of its quality by dilution. When low concentration of DNA is extracted, another option for overcoming the problem of inhibition is to use a column purification step to complete the SA-Gen method. We were able to purify DNA extracted with the SA-Gen method from Ref7 samples that could not be amplified because of inhibitors with the aid of a DNA purification column (Nucleospin® gDNA Clean-up, Macherey–Nagel). Despite the high loss of DNA during the purification (minimum = 40% loss), purified DNA quantities exceeded the quantities obtained with the MN-Soil kit, and we were able to use PCR amplification on the purified DNA without the need for dilution (data not shown). This result suggests that adding a column-purification step to the SA-Gen method could be a good solution for low DNA-concentration samples (<3 ng DNA/μL) containing very high levels of PCR inhibitors.

5.2. Diatom community composition is unchanged whatever the extraction method

No effect of DNA extraction methods on OTU richness and diversity was found, and the sample origin appeared to be the main source of variation in our study. This intersample variation

is consistent with the contrasting characteristics of our environmental samples (origin, quality status), which harbor different diatom community composition. Despite the presence of PCR inhibitors, SA-Gen provided a picture of diatom community similar to that provided by the other methods.

Regardless of taxonomic level (OTU or species), the taxonomic composition of the community represented in the extracts was not affected by DNA extraction methods. We observed 81.5% of shared species between the 5 methods, and when we removed the 2 samples with initial subsampling bias (767 and P45), this value increased to 93.8%. However, proportional reads did differ among extraction methods for some taxa. The observed intrasample variation (~27%) is consistent with variation observed in studies of bacterial community structure in water (Staley *et al.* 2015) or salivary samples (Lazarevic *et al.* 2013). These investigators found variations in relative abundance at the order and phylum/genus levels that were related to DNA extraction methods. Such variations are usually attributed to biases within the extraction process. Some diatom genera appeared to be preferentially detected with MN-Soil (*Nitzschia*, *Amphora*) or with SA-Gen methods (*Encyonema*, *Gomphonema*, *Navicula*), indicating that all methods did not extract DNA equally from all taxa. Depending on the diatom species, the skeleton can display different features (*e.g.*, shape, size, thickness) and different proportion of silica (Barker 1992). Mechanical resistance of diatom skeletons can vary from one species to another depending on factors, such as porosity or shape (Hamm *et al.* 2003; Moreno *et al.* 2015). Some diatom species are more resistant than others to mechanical lysis (bead beating) during the DNA extraction, and this resistance affects the relative abundances obtained (Koid *et al.* 2012; Manoylov *et al.* 2016). We hypothesize that diatom species with long and thin skeletons may be more easily broken by mechanical lysis than small species with thick skeletons, thereby affecting their relative representation in the molecular inventory. Considering the samples for which we obtained an efficient taxonomic assignment we can verify that small species (<20 μm length) were proportionally less represented in the molecular inventories than in the morphological ones, whereas the species >50 μm long appeared to be proportionally more abundant in the molecular inventories (data not shown).

5.3. Accuracy of molecular inventories and downstream quality indices

Diatom taxonomic inventories, based on assigned OTUs, were compared with taxonomic inventories based on morphological data for each DNA extraction method. Slight deviation between molecular and morphological diatom taxa was observed, but none of the DNA extraction methods provided a better match with microscopy than the others because they all shared similar diatom taxa. We found only 2 exceptional deviations in taxonomic composition data (from samples 767 and P45), and both were most probably caused by initial subsampling bias. We considered all observed taxa from the morphological inventories but only diatom taxa with robust taxonomic assignment from the molecular inventories when we compared molecular and morphological inventories. The molecular inventories were especially incomplete for Swedish and tropical samples because they encompassed taxonomic diversity that is not well represented in the R-syst::diatom database. This problem illustrates the need to continue updating the barcode reference database to provide more complete coverage of diatom diversity. The quality and completeness of the molecular reference database used to make the link between molecular data and diatom references is crucial in metabarcoding studies, as already pointed by Zimmermann *et al.* (2014) for diatoms. R-syst::diatom, the molecular reference database we used, was created to provide a reliable database curated by diatom taxonomist experts. The unsolved problem is the difficulty of enriching the molecular reference database with new taxa that can be identified unambiguously based on morphology. This requirement carries with it the need for the capacity to sample, isolate, accurately identify, and sequence new species. Single-cell PCR technique can help investigators obtain sequences from uncultured diatoms combined with their morphological identification (Hamilton *et al.* 2015). However, the efficiency of this approach is still low for diatoms and giving a precise taxonomic identification at the species level is often impossible with only one cell.

Despite the partial match between molecular and morphological data, molecular SPI values were highly correlated to morphological SPI values for the 6 European samples (Edian, Aire, Lake, 767, M36, and P45). The R-syst::diatom database was mainly populated with the DNA sequence of diatoms isolated from temperate regions, so it provided better molecular coverage for our European samples than the 2 tropical samples (Pol2 and Ref7). One of the mismatching taxa in the tropical samples was very abundant (*Nitzschia inconspicua*, 33.5 and 29.5% of total valves counts) only in the microscopy-based inventories, which added to the large differences between molecular and microscopy SPI values for these samples. This species is represented by

many sequences in the R-syst::diatom database but is paraphyletic species and a “taxonomic mess” (Rovira *et al.* 2015), which yields incomplete taxonomic assignment at genus/species level. *Nitzschia inconspicua* has a medium-sensitivity indicator value and is usually present at sites with medium or poor quality status. Consequently, its absence from the molecular SPI calculation tends to give higher SPI values with metabarcoding than with microscopy.

SPI values obtained were consistent with quality status of all environmental sites, even for Mayotte sites for which the SPI calculation is not yet adapted to infer quality status. SPI calculation is driven mainly by species with abundances >5% (Bigler *et al.* 2009), and we were able to detect most abundant genera (75.9%) and species (33.6%) in our DNA inventories. This feature of the SPI can explain why molecular and morphological SPI values were highly correlated for European samples despite some deviations in diatom taxonomic lists.

Taxa must be quantified before quality indices can be calculated. The correlation between relative abundances of sequences obtained by HTS and diatom specimens observed by microscopy was not constant and varied from one taxon to another in our data. For many biological groups, DNA metabarcoding for quantifying relative abundances is limited by biological and technical biases that might influence sequence read counts (Thomas *et al.* 2016). These biases can have multiple origins and include DNA extraction efficiency; variation of copy number of the targeted genes, primer specificity, and PCR amplification efficiency that may differ among species; DNA sequencing errors and bioinformatics filtering that may affect DNA sequence reliability (Jeon *et al.* 2008; Amend *et al.* 2010; Weber & Pawlowski 2013; Bragg *et al.* 2013; Schmidt *et al.* 2015; Deiner *et al.* 2015; Elbrecht & Leese 2015). In the case of diatoms, the number of copies of *rbcl* per genome, the number of genomes per chloroplast, and the number of chloroplasts per cell may influence the correlation between DNA sequences counts and morphological counts. However, we assume that the number of copies of *rbcl* genome per and number of genomes per chloroplast probably do not introduce major biases in DNA sequence counts. The sequenced plastid genomes currently available (*e.g.*, Ruck *et al.* 2014) reveal that the *rbcl* gene is present as 1 copy per plastid genome. The number of plastids per cell varies from 1 to 4 for benthic diatoms (Round *et al.* 1990a) indicating that a correction factor could be developed based on plastid genome number variation, as proposed by (Angly *et al.* 2014) to correct bacterial quantification based on 16S ribosomal RNA amplicons. Additional technical biases linked to primer efficiency, PCR amplification, and sequencing errors are not easily estimated and corrected unless some control material can be introduced as an internal standard, as proposed for quantification in fish (Thomas *et al.* 2016) or alien DNA to estimate the fraction of amplicons captured by the sequence library

(Gifford *et al.* 2011; Mangot *et al.* 2013). As an follow-up to our comparison of DNA extraction methods, further investigation could be done to estimate the importance of technical/biological biases and to test the feasibility of potential correction factors to improve quantification of HTS and to adapt calculations of quality indices.

6. Conclusion

Our results show that all of the DNA extraction methods tested provide DNA of sufficient quality and quantity to perform benthic diatom community analysis based on HTS to obtain reliable molecular inventories of diatoms. The composition of diatom assemblages obtained was not affected by the choice of DNA extraction method. The relative abundances of some taxa can vary with the efficiency of lysis methods to disrupt diatom cells, but this variability did not affect the SPI value.

The operating cost of following propositions of Kermarrec *et al.* (2014) to implement next-generation biomonitoring with diatom metabarcoding as an alternative to classical morphological approach has to be considered. The cost per sample may vary depending on the HTS technologies used (Loman *et al.* 2012), but (Stein *et al.* 2014) showed that DNA metabarcoding can be a valid economic solution for biomonitoring programs at the national scale. They showed for algal indicators (including diatoms) that the costs of molecular and classical methods for sampling and analysis are similar. The SA-Gen method, which is 24 to 39× times cheaper than other DNA extraction methods (including analysis and equipment costs), is an attractive choice to decrease the cost of next-generation biomonitoring. Moreover, this method provides a large quantity of DNA from environmental samples and the best correlation between *rbcl* copy number and valve observed by microscopy. The low quality DNA and the presence of PCR inhibitors in SA-Gen extracts did not affect diatom composition and SPI calculation, so we encourage the use of the SA-Gen method to perform DNA extraction for HTS diatom biomonitoring purposes.

Use of metabarcoding for biomonitoring is a complex workflow that requires standardization. We have provided a benchmark for the first step of this workflow. Further work is required for standardization of the full process, including reference database update, quantification, bioinformatics workflow, and adaptation of methods for calculating indices.

7. Acknowledgements

We thank Kalman Tapolczai for participating in the morphological identification and counting the diatoms in our environmental samples. We also thank Franck Salin, Christophe Boury, and Erwan Guichoux from the “Plateforme Génome Transcriptome” (PGTB, Bordeaux, France) who performed HTS sequencing and provided fastq files containing DNA reads. We extend special thanks to Alain Franc and Philippe Chaumeil from the “Biodiversité, Gènes et communautés” (INRA Biogeco) scientific team for helpful discussions. We thank the laboratory technical support of Cécile Chardon and Louis Jacas for their help and advice, and people from the INRA R-syst network from which the R-syst::diatom database was initiated. This work was funded by the French National Agency for Water and Aquatic Environments (ONEMA-AFB). The Swedish contribution was funded by the Swedish Agency for Marine and Water Management and by the SLU Environmental Monitoring and Assessment programme Lakes and Watercourses.

8. Author contributions

A. Bouchez, I. Domaizon, F. Rimet and V. Vasselon contributed to study conception and design. M. Kahlert, F. Rimet and V. Vasselon were involved in acquisition of data. V. Vasselon performed bioinformatics treatments and A. Bouchez, I. Domaizon, M. Kahlert, F. Rimet and V. Vasselon were involved in analysis and interpretation of data. A. Bouchez, I. Domaizon, M. Kahlert, F. Rimet and V. Vasselon participated in drafting the article and revising it critically.

9. Supplementary data

Table S1. Mean Ct values obtained for qPCR assays on DNA extracted from all environmental samples with the 5 DNA extraction methods (measurement performed for 4 levels of dilution). Standard deviation obtained for these measures are represented in brackets. The last column shows the theoretical dilution factor to apply in order to remove inhibition effects and the level of inhibition. (-) = not inhibited, (+) = weakly inhibited, (++) = strongly inhibited, (+++) = very strongly inhibited, - = Unamplified DNA and out of range values (below detection threshold compared to the standard curve), Ø = missing values.

Biofilm sample	Extraction method	Mean Ct values				Associated std curve slope	Theoretical	
		undil	10 fold dil	10 ² fold dil	10 ³ fold dil		dilution factor (DF)	
Edian	<i>MN-Soil</i>	22.4 (0.1)	25.4 (0.2)	28.9 (0.3)	-	-3.42	1.3	(-)
	<i>MN-Plant</i>	21 (0.4)	23.9 (0.3)	27.5 (0)	30.4 (0.5)	-3.54	2.2	(+)
	<i>S-Plant</i>	19.3 (0)	22.5 (0)	25.8 (0.1)	29.1 (0.2)	-3.54	1.6	(-)
	<i>Q-Blood</i>	20.1 (0.2)	22.7 (0.1)	26.8 (0.2)	30.4 (0.5)	-3.54	1.3	(-)
	<i>SA-Gen</i>	26.3 (0.7)	23.7 (0)	26.2 (0.2)	29.3 (0)	-3.54	143.7	(+++)
Aire	<i>MN-Soil</i>	19.7 (0.3)	23.1 (0.2)	27.1 (0.1)	31 (0.2)	-3.42	0.5	(-)
	<i>MN-Plant</i>	19.5 (0.1)	22.1 (0.1)	25.4 (0)	28.7 (0.3)	-3.54	2.4	(+)
	<i>S-Plant</i>	17.5 (0.3)	20.3 (0.2)	24 (0)	27.8 (0)	-3.54	1.2	(-)
	<i>Q-Blood</i>	17.7 (0.3)	20.5 (0.3)	24.1 (0.2)	27.2 (0.4)	-3.54	2.1	(+)
	<i>SA-Gen</i>	-	20.3 (0.4)	22.7 (0.6)	25.9 (0.3)	-3.54	26,0	(++)
Lake	<i>MN-Soil</i>	25.3 (0.3)	28.9 (0.3)	-	-	-3.42	0.9	(-)
	<i>MN-Plant</i>	26.3 (0.4)	28.9 (0.5)	32.3 (0.3)	-	-3.71	2.4	(+)
	<i>S-Plant</i>	24.4 (0.1)	27 (0.2)	30.2 (0)	-	-3.71	2.7	(+)
	<i>Q-Blood</i>	24.9 (0.3)	26.2 (0.3)	30.5 (0.7)	32.8 (0.4)	-3.71	7.3	(+)
	<i>SA-Gen</i>	-	25.2 (0.3)	27.1 (0.2)	30.5 (0.3)	-3.71	37.9	(++)
767	<i>MN-Soil</i>	27.7 (0.2)	31.9 (0.7)	-	-	-3.42	0.6	(-)
	<i>MN-Plant</i>	-	-	-	-	-3.63	Ø	Ø
	<i>S-Plant</i>	32 (0.5)	-	-	-	-3.63	Ø	Ø
	<i>Q-Blood</i>	32.4 (0.2)	-	-	-	-3.63	Ø	Ø
	<i>SA-Gen</i>	-	32.2 (0.3)	-	-	-3.63	Ø	Ø
M36	<i>MN-Soil</i>	22.8 (0)	26.2 (0.2)	30 (0.4)	-	-3.42	0.8	(-)
	<i>MN-Plant</i>	21.6 (0)	24.5 (0.2)	28 (0.3)	31.8 (0.6)	-3.54	1.4	(-)
	<i>S-Plant</i>	19.6 (0.1)	22.3 (0.1)	25.8 (0.2)	29.6 (0.5)	-3.54	1.5	(-)
	<i>Q-Blood</i>	19.9 (0.4)	22.8 (0.2)	26.7 (0.1)	30.1 (0.1)	-3.54	1.3	(-)
	<i>SA-Gen</i>	24.5 (1.5)	20.8 (0.8)	24 (0.1)	27.5 (0.1)	-3.54	142.2	(+++)
P45	<i>MN-Soil</i>	29.2 (0.4)	-	-	-	-3.42	Ø	Ø
	<i>MN-Plant</i>	28.6 (0)	-	-	-	-3.54	Ø	Ø
	<i>S-Plant</i>	-	-	-	-	-3.54	Ø	Ø
	<i>Q-Blood</i>	-	-	-	-	-3.54	Ø	Ø
	<i>SA-Gen</i>	-	30.6 (0.4)	-	-	-3.54	Ø	Ø
Ref7	<i>MN-Soil</i>	25.9 (0)	29.3 (0.8)	-	-	-3.42	1,0	(-)

	<i>MN-Plant</i>	28.7 (0.2)	29.6 (0.3)	32.2 (0.3)	-	-3.63	10.9 (++)
	<i>S-Plant</i>	23.1 (0.1)	24.7 (0.1)	27.9 (0.1)	31.6 (1)	-3.63	4.5 (+)
	<i>Q-Blood</i>	26.8 (0.1)	28.6 (0.3)	32.1 (0.3)	-	-3.63	3.2 (+)
	<i>SA-Gen</i>	-	26 (0.6)	25.1 (0.1)	28.6 (0.9)	-3.63	182.3 (+++)
Pol2	<i>MN-Soil</i>	28 (0.1)	31.9 (0.1)	-	-	-3.42	0.7 (-)
	<i>MN-Plant</i>	30 (0.5)	31.3 (0)	-	-	-3.71	4.5 (+)
	<i>S-Plant</i>	26.7 (0.4)	28.8 (0.8)	32.3 (0.6)	-	-3.71	3.1 (+)
	<i>Q-Blood</i>	28.5 (0.1)	30.9 (0.1)	-	-	-3.71	0.8 (-)
	<i>SA-Gen</i>	-	27 (0.2)	29.3 (0.1)	32.5 (0.5)	-3.71	31,0 (++)

Table S2. Ratio between valves number and *rbcl* copy number per mg of biofilm for each site. Positive and negative symbols show if *rbcl* copy number is higher or lower than valves number. "ND" for not determined values (out of range for qPCR assay).

	MN-Soil	MN-Plant	S-Plant	Q-Blood	SA-Gen
Edian	25 (-)	12 (-)	4 (-)	6 (-)	3 (-)
Aire	14 (-)	7 (-)	2 (-)	3 (-)	3 (+)
Lake	225 (-)	166 (-)	48 (-)	40 (-)	6 (-)
767	ND	ND	ND	ND	ND
M36	10 (-)	3 (-)	1 (+)	1 (-)	4 (+)
P45	ND	ND	ND	ND	ND
Ref7	56 (-)	40 (-)	2 (-)	31 (-)	3 (+)
Pol2	268 (-)	109 (-)	20 (-)	83 (-)	2 (-)

Table S3. Sequence data evolution from raw reads to OTU. DNA reads were trimmed (i) based on their length and sequence quality, (ii) reads which not fully aligned to *rbcl* barcode or were considered as chimera were removed, (iii) remaining reads that were assigned to non-diatoms taxa were removed. Clustering was made using a 95% similarity threshold and singletons were removed before the final standardization at 4180 reads per sample.

Biofilm sample	DNA extracion method	Raw reads	Trimmed reads			95% similarity OTU			
			length and quality	aligmnet and chimera	"non diatom" removal	Without singleton		After subsampling	
						Reads nb	OTU nb	Reads	OTU nb
Edian	<i>MN-Soil</i>	59460	14701	10044	10044	10034	255	4180	171
		53559	13344	9201	9201	9187	252	4180	185
	<i>MN-Plant</i>	72101	15875	9467	9464	9448	334	4180	234
		67302	16124	9581	9574	9557	337	4180	231
	<i>S-Plant</i>	77194	22540	14208	14208	14176	396	4180	233
		90338	24351	15160	15160	15140	403	4180	218
	<i>Q-Blood</i>	64475	17881	11012	11010	10982	332	4180	222
		66899	16052	9167	9166	9132	400	4180	296
<i>SA-Gen</i>	79408	23512	13211	13211	13186	386	4180	242	
	90819	28548	16078	16078	16051	443	4180	255	
Aire	<i>MN-Soil</i>	42510	16019	12662	12661	12656	263	4180	161
		43883	16603	12885	12885	12861	256	4180	164
	<i>MN-Plant</i>	108577	38647	29846	29846	29816	383	4180	160
		51288	18872	14854	14854	14840	275	4180	153
	<i>S-Plant</i>	54860	20542	15721	15721	15706	316	4180	174
		55299	20505	16336	16336	16319	323	4180	164
	<i>Q-Blood</i>	55876	22114	17433	17433	17405	313	4180	161
		45334	18145	14450	14450	14434	302	4180	173
<i>SA-Gen</i>	50163	19640	15208	15208	15193	275	4180	142	
	51805	18690	14391	14391	14378	259	4180	146	
Lake	<i>MN-Soil</i>	61452	24546	14032	14003	13982	304	4180	204
		54250	21305	11946	11920	11906	281	4180	207
	<i>MN-Plant</i>	44708	15284	9046	9046	9037	255	4180	195
		47727	18753	11228	11226	11216	287	4180	204
	<i>S-Plant</i>	35598	12397	7096	7096	7084	250	4180	205
		66673	23593	12891	12891	12865	329	4180	220
	<i>Q-Blood</i>	55428	18563	10278	10278	10253	299	4180	215
		61721	25613	15028	15027	15011	299	4180	185
<i>SA-Gen</i>	66259	34393	20716	20715	20692	311	4180	164	
	64375	33005	20117	20117	20095	309	4180	167	
767	<i>MN-Soil</i>	51331	14341	8730	8729	8716	182	4180	139
		50624	14041	8611	8611	8593	166	4180	129
	<i>MN-Plant</i>	56702	18199	12839	12710	12696	165	4180	113
		66441	18353	9063	9063	9047	158	4180	108
	<i>S-Plant</i>	46103	11803	5928	5928	5921	140	4180	124
		46184	11534	5644	5644	5635	151	4180	131
<i>Q-Blood</i>	51052	15498	7581	7573	7563	155	4180	126	

		57222	17731	8661	8661	8643	170	4180	118
	<i>SA-Gen</i>	65155	18464	9693	9693	9665	164	4180	121
		59390	16137	7383	7383	7364	151	4180	115
M36	<i>MN-Soil</i>	59140	20048	14715	14169	14159	232	4180	138
		55532	21702	16526	15866	15852	220	4180	140
	<i>MN-Plant</i>	85375	34016	24577	22890	22861	320	4180	145
		66222	26328	19506	17910	17901	269	4180	140
	<i>S-Plant</i>	65535	26216	19372	19323	19309	276	4180	134
		51301	20801	15477	15440	15436	249	4180	148
	<i>Q-Blood</i>	55748	18891	13884	13724	13709	269	4180	153
		57813	22035	16019	15935	15920	300	4180	178
	<i>SA-Gen</i>	61592	23930	18078	18053	18044	225	4180	124
		51635	18462	13881	13854	13840	242	4180	141
P45	<i>MN-Soil</i>	54006	18191	10670	10645	10626	188	4180	140
		69183	21251	11990	11979	11944	219	4180	139
	<i>MN-Plant</i>	51712	17906	10204	10190	10167	202	4180	137
		55278	17980	10604	10594	10561	190	4180	149
	<i>S-Plant</i>	64960	20945	11273	11270	11248	162	4180	117
		49523	15580	9027	9027	9018	125	4180	86
	<i>Q-Blood</i>	81627	31134	14999	14999	14987	89	4180	58
		73028	22122	13439	13439	13429	119	4180	78
	<i>SA-Gen</i>	67369	24826	14558	14525	14518	157	4180	102
		53193	17280	9219	9208	9198	153	4180	123
Ref7	<i>MN-Soil</i>	58460	21974	16635	16635	16612	331	4180	171
		55905	20788	15624	15624	15599	304	4180	162
	<i>MN-Plant</i>	59330	24846	17302	17302	17284	348	4180	167
		64824	26175	18460	18460	18421	376	4180	198
	<i>S-Plant</i>	52669	21192	14493	14493	14455	430	4180	246
		66371	24533	16511	16511	16476	429	4180	231
	<i>Q-Blood</i>	54749	21219	14957	14957	14921	372	4180	193
		68135	29225	20909	20909	20881	383	4180	194
Pol2	<i>MN-Soil</i>	58414	14010	6932	6929	6914	226	4180	192
		56632	11423	5785	5784	5769	215	4180	183
	<i>MN-Plant</i>	68003	14523	5048	5048	5043	174	4180	165
		62163	10897	4188	4188	4180	185	4180	185
	<i>S-Plant</i>	66065	14831	6189	6189	6175	205	4180	177
		44490	10612	4962	4960	4954	198	4180	181
	<i>Q-Blood</i>	67634	12835	4802	4802	4788	179	4180	169
		62604	11753	4475	4474	4468	168	4180	162
	<i>SA-Gen</i>	62058	12358	4282	4281	4269	179	4180	178
		63850	13854	5290	5288	5275	206	4180	186
	Total	4711673	1542950	972288	967089	965696	3293	326040	2776
		(100%)	(32.7%)	(20.6%)	(20.5%)	(20.5%)	OTU	(6.9%)	OTU

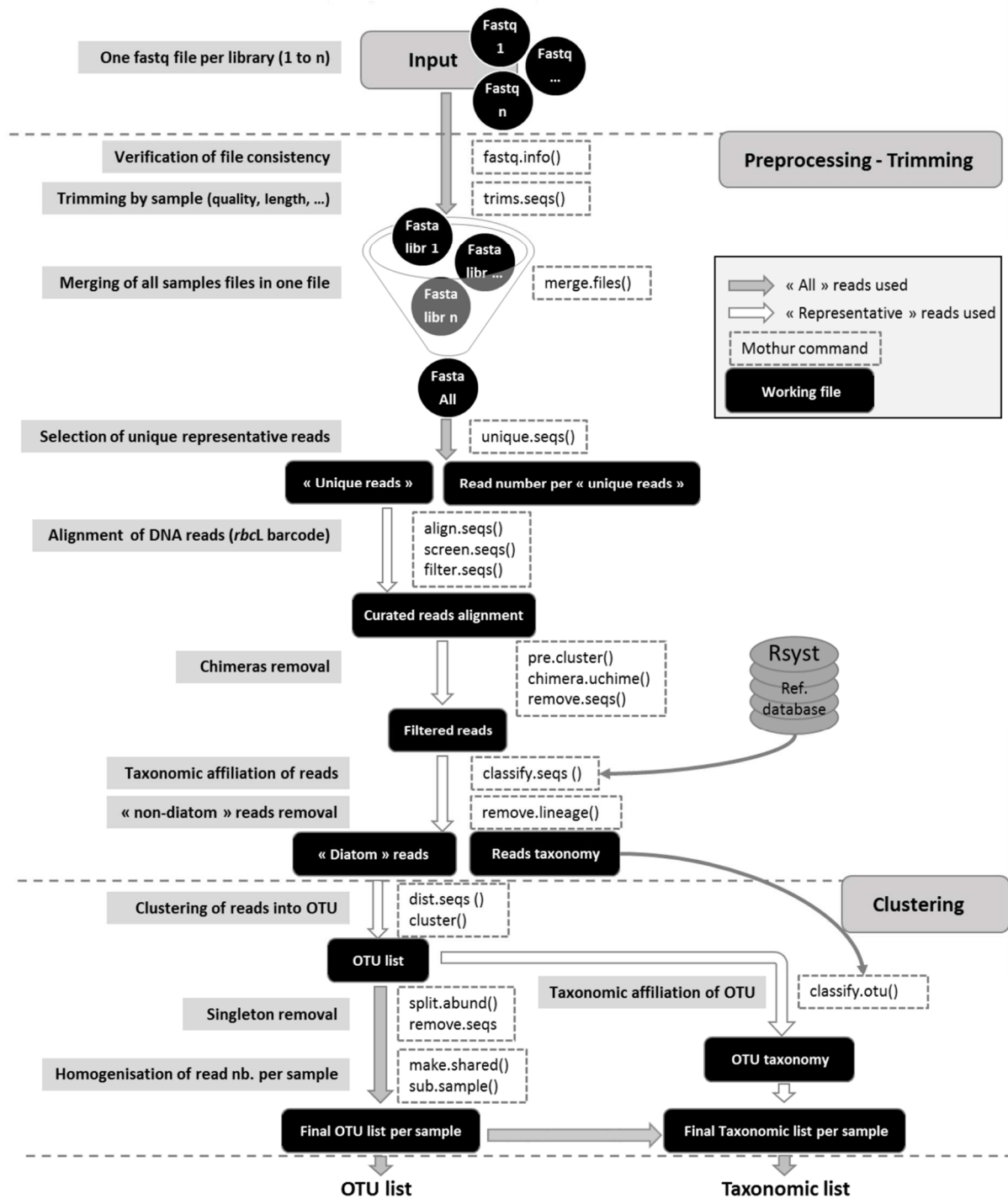
Table S4. Mean values of Chao richness estimator obtained for all environmental samples and DNA extraction methods. Standard deviations are presented in brackets. The effect of DNA extraction methods was assessed using Kruskal-Wallis test for each environmental sample, the p-values obtained for this test are presented in the last column (n=16 and $\alpha=0.05$). \emptyset : missing data.

Environmental sample	DNA extraction method					p-value
	<i>MN-Soil</i>	<i>MN-Plant</i>	<i>S-Plant</i>	<i>Q-Blood</i>	<i>SA-Gen</i>	
<i>Edian</i>	361 (94)	378 (24)	350 (9)	436 (122)	433 (13)	$p = 0.10$
<i>Aire</i>	274 (42)	304 (34)	319 (58)	284 (78)	265 (16)	$p = 0.64$
<i>Lake</i>	331 (12)	312 (15)	312 (36)	294 (13)	245 (2)	$p = 0.17$
<i>767</i>	198 (20)	160 (8)	193 (57)	201 (21)	229 (94)	$p = 0.70$
<i>M36</i>	263 (22)	244 (31)	249 (26)	278 (14)	205 (16)	$p = 0.14$
<i>P45</i>	184 (2)	210 (48)	138 (15)	94 (15)	173 (57)	$p = 0.17$
<i>Ref7</i>	318 (74)	299 (47)	420 (40)	292 (1)	\emptyset	$p = 0.25$
<i>Pol2</i>	287 (11)	263 (24)	274 (59)	287 (40)	308 (30)	$p = 0.66$

Table S5. Mean values of Shannon diversity index obtained for all environmental samples and DNA extraction methods. Standard deviations are presented in brackets. The effect of DNA extraction methods was assessed using Kruskal-Wallis test for each environmental sample, the p-values obtained for this test are presented in the last column (n=16 and $\alpha=0.05$). \emptyset : missing data.

Environmental sample	DNA extraction method					p-value
	<i>MN-Soil</i>	<i>MN-Plant</i>	<i>S-Plant</i>	<i>Q-Blood</i>	<i>SA-Gen</i>	
<i>Edian</i>	3.08 (0.05)	3.63 (0.01)	3.42 (0.09)	3.54 (0.25)	3.61 (0.01)	$p = 0.17$
<i>Aire</i>	2.94 (0.04)	2.82 (0.06)	2.69 (0.11)	2.64 (0.06)	2.29 (0.03)	$p = 0.08$
<i>Lake</i>	3.32 (0)	3.3 (0.04)	3.5 (0.15)	3.07 (0.23)	2.62 (0.02)	$p = 0.08$
<i>767</i>	2.75 (0.04)	3.07 (0.08)	3.29 (0.12)	3.06 (0.02)	3.02 (0.15)	$p = 0.13$
<i>M36</i>	2.72 (0.03)	2.75 (0.05)	2.56 (0.04)	2.81 (0.03)	2.39 (0.11)	$p = 0.09$
<i>P45</i>	2.91 (0)	2.81 (0.03)	2.8 (0.37)	2.08 (0.32)	2.68 (0.27)	$p = 0.20$
<i>Ref7</i>	2.25 (0.05)	2.62 (0.02)	2.41 (0.16)	2.41 (0.3)	\emptyset	$p = 0.32$
<i>Pol2</i>	3.62 (0.02)	3.58 (0.01)	3.43 (0.02)	3.44 (0)	3.61 (0.09)	$p = 0.13$

Fig. S1. Main steps of the bioinformatics process applied to perform parallel analysis of multiple samples, from raw reads to taxonomic inventories (using Mothur software).



III. Biais lié à la variation du nombre de copie de gène

“Avoiding quantification bias in metabarcoding: application of a cell biovolume correction factor in diatom molecular biomonitoring”

(version acceptée le 26 Septembre 2017 dans le journal *Methods in Ecology and Evolution*)

Valentin Vasselon¹, Agnès Bouchez¹, Frédéric Rimet¹, Stéphan Jacquet¹, Rosa Trobajo², Méline Corniquel¹, Kálmán Tapolczai¹, Isabelle Domaizon¹

¹CARTEL, INRA, Université de Savoie Mont Blanc, 74200, Thonon-les-bains, France

² Aquatic Ecosystems, Institute for Food and Agricultural Research and Technology (IRTA), Crta de Poble Nou Km 5.5, Sant Carles de la Ràpita, Catalunya, Spain

1. Abstract

In recent years, remarkable progress has been made in developing environmental DNA metabarcoding. However, its ability to quantify species relative abundance remains uncertain, limiting its application for biomonitoring. In diatoms, although the *rbcl* gene appears to be a suitable barcode for diatoms, providing relevant qualitative data to describe taxonomic composition, improvement of species quantification is still required.

Here, we hypothesized that *rbcl* copy number is correlated with diatom cell biovolume (as previously described for the 18S gene) and that a correction factor (CF) based on cell biovolume should be applied to improve taxa quantification. We carried out a laboratory experiment using pure cultures of 8 diatom species with contrasted cell biovolumes in order to (i) verify the relationship between *rbcl* copy numbers (estimated by qPCR) and diatom cell biovolumes, and (ii) define a potential CF. In order to evaluate CF efficiency, five mock communities were created by mixing different amounts of DNA from the 8 species, and were sequenced using HTS and targeting the same *rbcl* barcode.

As expected, the correction of DNA reads proportions by the CF improved the congruence between morphological and molecular inventories. Final validation of the CF was obtained on environmental samples (metabarcoding data from 80 benthic biofilms) for which the application of CF allowed differences between molecular and morphological water quality indices to be reduced by 47 %. Overall, our results highlight the usefulness of applying a CF factor, which is effective in reducing over-estimation of high biovolume species, correcting quantitative biases in diatom metabarcoding studies and improving final water quality assessment.

2. Introduction

DNA metabarcoding allows species present in an environmental sample to be detected using a short DNA marker specific for a particular taxonomic group (Taberlet et al. 2012a). Combined with High-Throughput Sequencing (HTS), hundreds of samples can be analyzed at the same time, offering an alternative to microscopy with higher resolution and accuracy, while being faster and cheaper (Stein et al. 2014). This is particularly interesting for freshwater biomonitoring, in which thousands of river samples have to be analyzed annually and management actions applied quickly (Keck et al. 2017). The European Water Framework Directive (WFD, European Council 2000) has implemented the use of benthic diatoms, among other biological indicators (fishes, macroinvertebrates, macrophytes and phytoplankton), for the assessment of aquatic

ecosystem integrity. The different biotic diatom indices that have been developed are based on the relative abundances and the ecological values (sensitivity and tolerance to pollutants) of the species observed in rivers and lakes systems (*e.g.* Rimet 2012). Different studies have already revealed the potential application of diatom metabarcoding in freshwater quality assessment (Kermarrec et al. 2014; Visco et al. 2015; Vasselon et al. 2017a; b; Apothéloz-Perret-Gentil et al. 2017). However, discrepancies between DNA metabarcoding and microscopy have been observed in species composition and relative abundance (Zimmermann et al. 2015). This drawback is likely to affect the congruence between morphological and DNA metabarcoding quality index values and, *in fine*, the ecological assessment.

With respect to qualitative aspects, the incompleteness of the reference databases, the choice of the DNA marker and the efficiency of the PCR primers have been identified as important biases affecting species detection using DNA metabarcoding (Pawłowski et al. 2016). For benthic diatoms, the *rbcL* gene has proved to be an appropriate taxonomic marker for biomonitoring (Kermarrec et al. 2013, 2014, Vasselon et al. 2017a,b) and a well-curated barcode reference library is already available in open-access to assign species names to *rbcL* sequences (R-Syst::diatom, Rimet et al. 2016). However, no clear relationship has yet been demonstrated between the relative species abundances obtained by DNA metabarcoding with the *rbcL* barcode and those obtained by morphological observations (Rimet et al. 2014). As quantification of diatom species is required by the WFD for quality index calculation, more investigation is needed to understand and correct biases affecting diatom quantification based on HTS data.

Species quantification based on HTS data can be estimated from the number of DNA sequences (*i.e.* reads) assigned to each species, from which relative abundances can be calculated. Previous studies have documented a variety of problems that may affect the proportions of DNA reads obtained with HTS (Amend et al. 2010; Deagle et al. 2013; Tan et al. 2015; Thomas et al. 2016; Pawłowski et al. 2016), including biological biases (*e.g.* gene copy number variation, tissue cell density, cell biovolume), technical biases (*e.g.* DNA extraction, PCR amplification), and biases linked to HTS itself (*e.g.* library construction, HTS technology used, bioinformatics treatments). Variation of gene copy number per cell constitutes a major bias known to affect the proportion of DNA-read found for each species present in complex assemblages; this has been demonstrated for macroinvertebrates (Elbrecht et al. 2017a), fish, amphibians (Evans et al. 2016), oligochaetes (Vivien *et al.* 2016), foraminifera (Weber & Pawłowski 2013), and microbial assemblages (Angly et al. 2014). However, to the best of our knowledge, no study has yet evaluated gene copy number variation bias on diatom metabarcoding quantification. While tissue cell density and species

biomass are major biases likely to affect DNA metabarcoding quantification of multicellular organisms like macroinvertebrates (Elbrecht & Leese 2015) or fish (Evans et al. 2016), diatoms are unicellular organisms for which gene copy number is mainly affected by the number of genomes and the number of gene copies per genome. This may be particularly true for non-nuclear markers like the chloroplast-encoded *rbcl* gene. Godhe et al. (2008) reported a clear correlation between the 18S gene copy number per cell with diatom cell length and biovolume, suggesting that the cell biovolume could be a proxy for the gene copy number. Keeping in mind that diatom biovolume varies from 10^1 to $10^9 \mu\text{m}^3$ (Snoeijs et al. 2002), gene copy number may vary greatly between the smallest and the biggest diatom species, affecting metabarcoding quantification.

For all the reasons mentioned above, we hypothesized that a quantification correction factor (CF) based on diatom cell biovolume should be necessary to correct DNA read proportions to provide species quantification more comparable to microscopical counts. In order to confirm this hypothesis, we firstly conducted experiments on 8 pure diatom cultures to examine whether variation in *rbcl* gene copy number per cell correlates with morphological characteristics (*e.g.* biovolume, cell length), from which a CF might be calculated. Secondly, the efficiency of the proposed CF was tested on (i) mock communities made by mixing known proportions of the 8 diatom species cultures, and (ii) environmental diatom assemblages from rivers previously sequenced (Vasselon et al. 2017b) and for which data are available online (Vasselon et al. 2017b dataset, <http://doi.org/10.5281/zenodo.400160>). Last, the capacity of the CF to improve the ecological assessment of rivers was tested by comparing water quality index values calculated from molecular data with corrected abundances to those calculated from classical morphological abundances.

3. Methods

3.1. Evaluation of the quantification bias and development of a quantification correction factor (CF)

To evaluate whether the *rbcl* copy number per cell varies between diatom species, strains from 8 freshwater diatom species were selected from the Thonon Culture Collection (TCC; http://www6.inra.fr/carrtel-collection_eng/) (**Table 4**). The 8 species were chosen for their

contrasted morphological (size and cell biovolume), cytological (e.g. chloroplast number) and phylogenetic characteristics (**Table 4**). Cell dimensions (width, length, thickness) of the 8 diatom species were measured under light microscopy (1000× magnification) using a minimum of 10 specimens per species. Then, appropriate geometrical models were applied to calculate their cell biovolume (Sun & Liu 2003) (**Table 4**).

Species	TCC code	Chloroplast nb./cell	Length (μm)	Width (μm)	Thickness (μm)	Biovolume (μm ³)
<i>Achnantheidium minutissimum</i> (Kützing) Czarnecki	TCC667	1	7.1	3.2	2.5	45
<i>Nitzschia palea</i> (Kützing) W.Smith	TCC139-1	2	22.7	4,0	4,0	183
<i>Ulnaria ulna</i> (Nitzsch) Compère	TCC670	2	54.6	7.9	9.5	4087
<i>Pinnularia viridiformis</i> (Nitzsch) Ehrenberg	TCC890	2	51.4	14.3	17.8	10282
<i>Diatoma tenuis</i> Kützing	TCC861	≈ 8	42.4	4.8	4.8	769
<i>Nitzschia inconspicua</i> Grunow	TCC488	2	8.1	4.3	3.6	98
<i>Fragilaria perminuta</i> (Grunow) Lange-Bertalot	TCC753	2	11.1	4.2	3.7	135
<i>Cyclotella meneghiniana</i> Kützing	TCC690	≈ 20	12.1		4.7	539

Table 4 – Characteristics of the 8 diatom species selected in the Thonon Culture Collection (TCC) and used in this study.

The 8 diatom cultures were cultivated in triplicate in 40 mL sterile DV medium (Rimet et al. 2014) using 50 mL Nunc™ EasYFlasks™ (Thermo Fisher Scientific, Waltham, Massachusetts). Flasks were placed on a rotating platter (4 rpm) in a controlled thermostatic room (21 ± 2°C, 14h light/10h dark cycle, light intensity of ca. 100 μmol quanta m⁻² s⁻¹). Flasks were inoculated in order to reach a concentration of ≈ 100 cells/mL at the beginning of the experiment for each species, except for *Ulnaria ulna* for which a concentration of ≈ 1000 cells/mL was used (due to its low growth rate). The growth of the 8 diatom cultures was followed during 40 days, except for *Pinnularia viridiformis* for which the survey lasted 73 days, due to its low growth rate. Cell concentrations, proportions of live/dead cells and *rbcL* gene copy concentrations per mL of media were measured for each culture at 7 sampling times (referred to as T0 to T6) (**Figure 25**).

Diatom cell concentrations and proportions of live/dead cells were obtained by counting at least 400 specimens using inverted microscopy (×1000 magnification) and the standard Utermöhl technique (European Committee for Standardization (CEN) 2006) (**Figure 25**). The proportion of live/dead cells was estimated by considering cells without visible intracellular contents as dead. Only living cells were taken into account to calculate the diatom cell concentration per mL of media. Flow cytometry using Sytox-Green was also used to confirm the microscopical data (not shown).

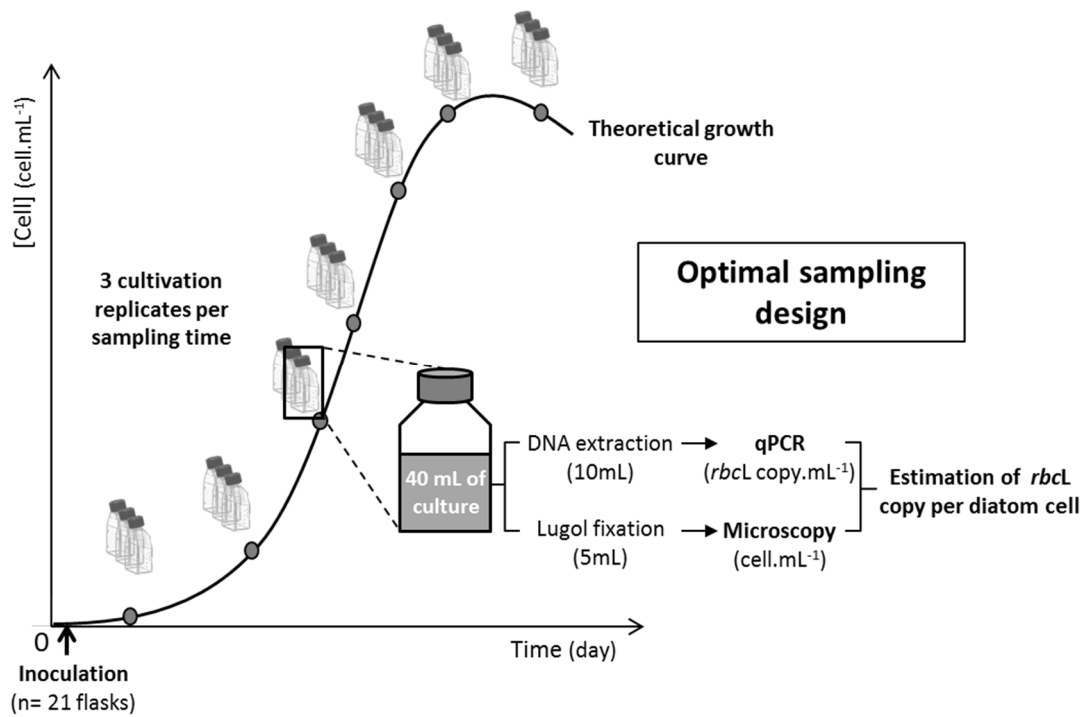


Figure 25 – Experimental design applied to the 8 diatom species.

After the inoculation of 21 flasks containing 40mL of DV media, diatom culture growth was followed at 7 sampling time (from T0 to T6) and analysis was performed in triplicate (3 flasks per sampling time).

rbcl copy number per mL was estimated by qPCR. From each cultivation replicate, 10 mL of culture was centrifuged at 17,000 x *g* for 30 min (**Figure 25**). Total DNA was extracted from the resulting pellet using a protocol based on GenElute™-LPA DNA precipitation (Sigma-Aldrich, St Louis, Missouri) as previously described (Vasselon *et al.* 2017a). Then, qPCR assays were performed for each of the 8 diatom species on DNA extracted at all 7 sampling times and with each of the 3 replicates, using the QuantiTect SYBR Green PCR Kit (Life Technologies, Carlsbad, USA) and the Rotor-Gene Q (Qiagen, Hilden, Germany). A short 312 bp region of the *rbcl* gene (the same as was used for HTS sequencing) was targeted using primers used by (Vasselon *et al.* 2017b) and described in **Table S1**. qPCR reactions were performed following the method used by Vasselon *et al.* (2017a), using a final volume of 25 μ L using mix preparation and reaction conditions as described in **Table S1**. A fluorescence threshold of 0.01 was used to allow comparison of qPCR assays, denoising and determination of the cycles' threshold (Ct). Data analysis was performed using the Rotor-Gene Q Series software (version 2.3.1) and the *rbcl* copy per mL of media was determined.

Finally, the number of *rbcl* gene copies per diatom cell was calculated for the 8 diatom species by dividing the *rbcl* concentration (qPCR data) by the living cell concentration (microscopy data). A Kruskal-Wallis test was performed using R (R Development core team 2013) to determine

if the *rbcl* gene copy number per diatom cell varied significantly between the 8 diatom species. Then, we tested the level of correlation between the number of *rbcl* gene copies per diatom cell and several morphological characteristics of the diatom cells (**Table 4**). Variables that did not approximate normal distributions were log transformed. Pearson correlation coefficients were calculated between the gene copy number per cell and the diatom cell morphological characteristics. This correlation was represented by a linear model.

3.2. Validation of the quantification CF to mock and environmental HTS data

3.2.1. Mock communities

The calculated CF was applied to metabarcoding data obtained from controlled diatom mock communities. 5 mock communities (M1 to M5) were created by mixing DNA extracted from each of the 8 diatom species sampled during their exponential growth phase, and for which the correspondence between cell abundances (microscopy) and qPCR counts was known. For each of the 5 mock communities, the volume of DNA used for 7 species was kept unchanged (1 μ L) and only the volume of DNA of *P. viridiformis* varied as followed: M1 = 0.2 μ L, M2 = 0.4 μ L, M3 = 0.8 μ L, M4 = 1.6 μ L, M5 = 3.2 μ L. This resulted in contrasted *rbcl* proportions of the 8 species among the 5 mock communities. Then, HTS sequencing of the *rbcl* 312 bp fragment was performed on 3 replicates of the 5 mock communities. The 15 corresponding libraries were prepared following the method described by Vasselon et al. (2017a) with the same primers and PCR reaction conditions as those used for *rbcl* qPCR (**Table S1**), changing only the cycle number to 30. Each library was diluted to 100 pm and all 15 were pooled together for one HTS run performed on the PGM Ion Torrent machine by the “Plateforme Génome Transcriptome” (PGTB, Bordeaux, France).

The sequencing platform provided a unique fastq file for each of the 15 libraries containing demultiplexed DNA reads without the sequencing adapters. Quality filtering of DNA reads was performed using the Mothur software (Schloss et al. 2009) and bioinformatics process described previously (Vasselon *et al.* 2017a; b). Finally, a taxonomy was assigned to each DNA read with the “classify.seqs” command (Mothur) using default parameters with a confidence threshold of 85% and the R-Syst::diatom library (Rimet et al. 2016, version updated in January 2015 and available upon request) as a *rbcl* reference library. A molecular taxonomic list with the associated read

numbers assigned to each of the 8 diatom species was obtained for each of the 5 mock communities and used for subsequent analysis.

The quantification CF defined for the *rbcL* gene was then applied to the molecular taxonomic lists for the 5 mock communities by dividing the read number for each species by its corresponding CF. Both the uncorrected and corrected HTS relative abundances of species from the 5 mock communities were then compared to the relative abundances obtained using microscopy.

3.2.2. Environmental diatom assemblages

To evaluate the efficiency of the CF to improve metabarcoding quantification from environmental samples, we used *rbcL* HTS data obtained from (Vasselon et al. 2017b), corresponding to 80 benthic diatom samples collected from rivers in tropical island of Mayotte, Indian Ocean (Vasselon et al. 2017b dataset, <http://doi.org/10.5281/zenodo.400160>). A CF was calculated for each species (or genus when the species level was not reached) detected in molecular inventories of the rivers of Mayotte island using a generalised average of the morphological information (*e.g.* biovolume, length) available in the R-Syst::diatom library and applied to HTS data. Corrected molecular inventories were produced for all the 80 river samples using the CF. The impact of the CF on diatom taxa abundance rank in the molecular inventories was assessed by comparing original and corrected molecular diatom inventories. Then, the Specific Pollution-sensitivity Index (SPI) used for ecological assessment was calculated for each sample based on the corrected diatom molecular inventories using the Omnidia 5 software (Lecointe, Coste & Prygiel 1993, library 5.3 2015) and compared to the morphological SPI values for all river samples (Vasselon et al. 2017b). Pearson correlation was used to evaluate the strength of correlations between original or corrected molecular SPI values and the morphological SPI values. Wilcoxon Signed Rank tests were conducted to determine whether the difference between the molecular and the morphological SPI (Δ SPI) varied significantly when using the original or the corrected molecular data for the molecular SPI calculation.

4. Results

4.1. Variation of *rbcL* gene copy number between diatom species

Cell and *rbcL* gene concentrations were measured, by inverted microscopy and qPCR respectively, for the 8 diatom species at different cultivation stages corresponding to 7 sampling points (T0 to T6). Information has been summarized in **Tables S2** and **S3**. As the 8 diatom species reached the beginning of the stationary phase at the sampling time T2 (*i.e.* between 13 and 31 days of cultivation), only the [cell] and the [gene copy] values obtained for the T0, T1 and T2 sampling times were used for further analysis. The calculated mean values of the *rbcL* gene copy number per cell for each diatom species varied between 0.5 and 130 copies per cell (**Figure 26**). The Kruskal-Wallis test revealed that the *rbcL* copy number per cell was significantly different ($p < 0.001$) between the 8 diatom species.

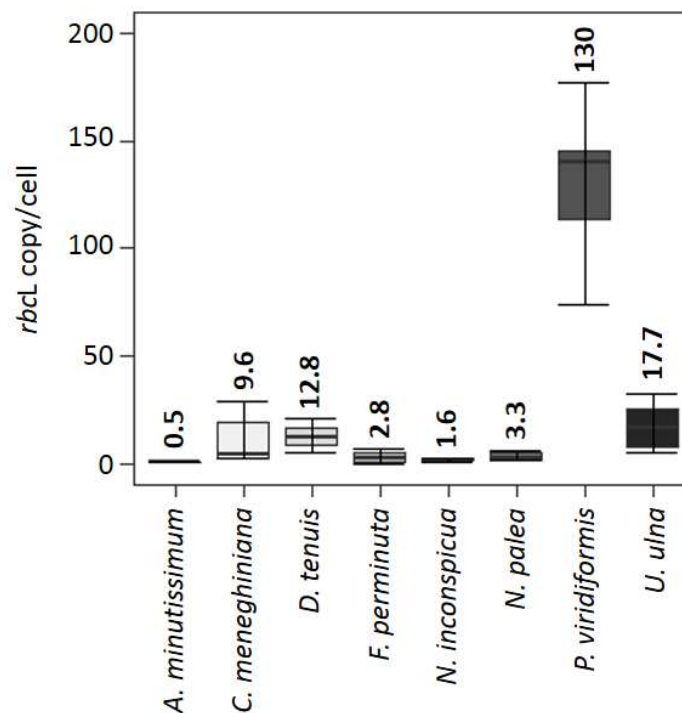


Figure 26 – Estimation of the *rbcL* copy number per diatom cell for the 8 diatom species. Mean values calculated using the gene and the diatom cell concentrations obtained respectively by qPCR and inverted microscopy at T0, T1 and T2 sampling points (n = 9).

4.2. Development of quantification CFs

The *rbcL* copy number per cell was highly correlated with cell biovolume ($r = 0.97$, $p < 0.001$), length ($r = 0.82$, $p < 0.001$), width ($r = 0.94$, $p < 0.001$) and thickness ($r = 0.96$, $p < 0.001$). The correlation between the *rbcL* copy number per cell and the cell biovolume followed a linear model (**Figure 27**). Assuming that this linear relation based on 8 diatom species is applicable to all diatom species, the equation of this model allows calculation of an estimate of the relative *rbcL* copy number per cell as soon as the biovolume of the cell is known, and thus to define a CF specific to each species.

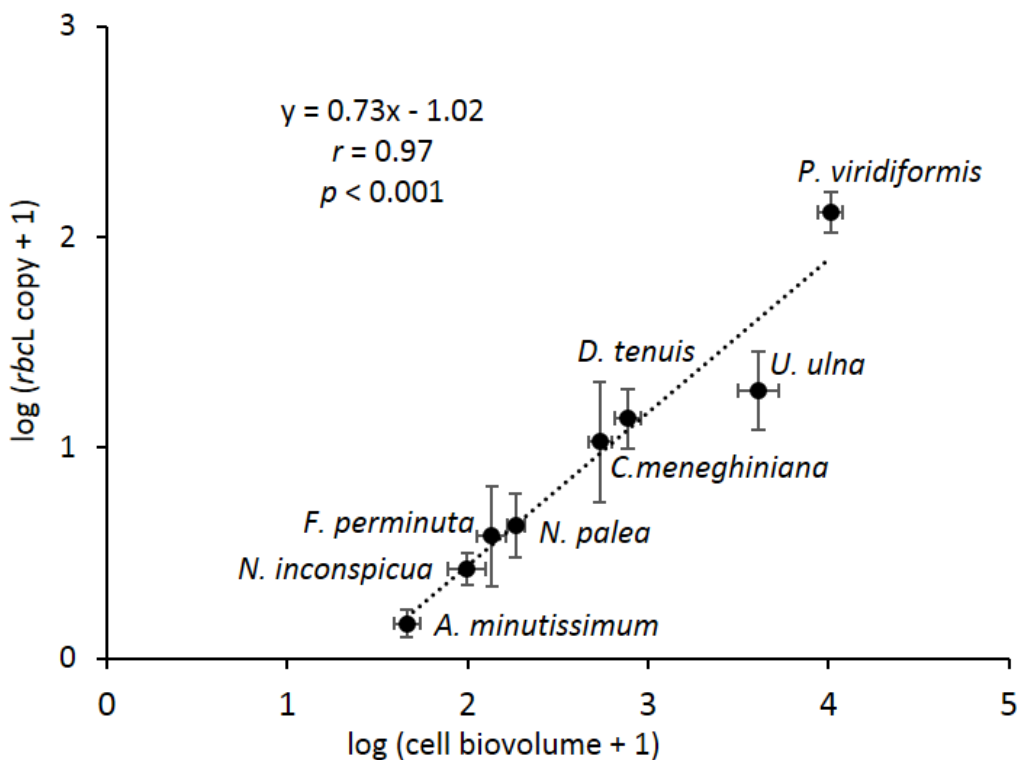


Figure 27 – Correlation between the diatom cell biovolume and the *rbcL* gene copy number per cell after $\log(x+1)$ transformation.

Such quantification CFs were calculated for each of the 8 diatom species of the mock communities (**Table 5**) and varied from 0.6 for *Achnantheidium minutissimum* to 78.5 for *P. viridiformis*. For each of the diatom taxa found in the environmental samples, CFs were also calculated using the biovolume information available for each taxa (from R-Syst::diatom library) (**Table S4**) and varied over a wider range, from 0.03 for *Fistulifera saprophila* to 649.8 for *Rhopalodia gibba*.

Species	Calculated CF
<i>A. minutissimum</i>	0.6
<i>N. inconspicua</i>	1.7
<i>N. palea</i>	3.3
<i>P. viridiformis</i>	78.5
<i>D. tenuis</i>	11.1
<i>F. perminuta</i>	2.4
<i>U. ulna</i>	39.6
<i>C. meneghiniana</i>	8.3

Table 5 – CF calculated for the 8 diatom species using their respective cell biovolume (Table 4) and the linear equation between the *rbcL* copy number and the cell biovolume (Figure 27).

4.3. Application of CFs to mock and environmental HTS data

953,082 DNA reads were produced from the 15 libraries corresponding to the 5 DNA mock communities (3 replicates per mock). Following the bioinformatics quality filtering steps, 385,367 DNA reads were retained. A molecular taxonomic list was then created by removing DNA reads which remained unclassified (0.43 % of the reads) or assigned to different taxa than the 8 diatom species present in the mock communities (0.004 % of the reads) (**Table S5**). The proportions of *P. viridiformis* reads in the 5 mock communities varied from 9 % in M1 to 57 % in M5 (**Figure 28A**) while observed cell proportions were lower; \approx 0.03 % in M1 and 0.55 % in M5 (**Figure 28B**). The application of the CF on DNA reads counts of the 8 species changed their relative abundances in the 5 mock communities (**Figure 28A**). The rank of the 8 species was also affected; for example, in M5 the application of the CF changed the proportion of *P. viridiformis* from 57 % to 4 % and the proportion of *A. minutissimum* from 4 % to 42 %. The correspondence between morphological and molecular relative abundances was highly improved by applying the CF on the HTS data (**Figure 28A, B**).

From the 80 environmental samples previously sequenced (Vasselon et al. 2017b), a molecular taxonomic list based on assigned DNA reads was produced including 23 families (75.1 % of total reads assigned), 39 genera (72 % of total reads assigned) and 66 diatom species (40.7 % of total reads assigned). From this list, 84 diatom taxa, including taxa assigned at the genus and the species level, were used to calculate the SPI freshwater quality index. CFs calculated from cell biovolumes for those 84 taxa were then applied to correct the quantification of the environmental molecular inventories (**Table S4**). The proportions and ranks of the dominant taxa were affected by the application of the CFs (**Figure 29**).

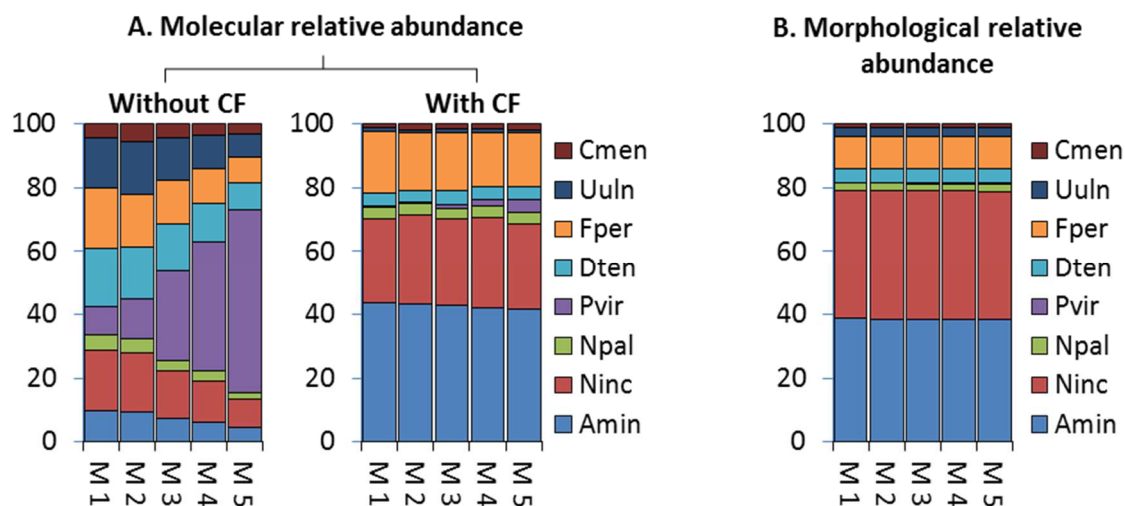


Figure 28 – Relative abundances of the 8 diatom species in the 5 DNA mock communities based (A) on mean of HTS DNA reads without (left) and with (right) correcting quantification using the biovolume correction factor and (B) on mean of morphological counts from inverted microscopy.

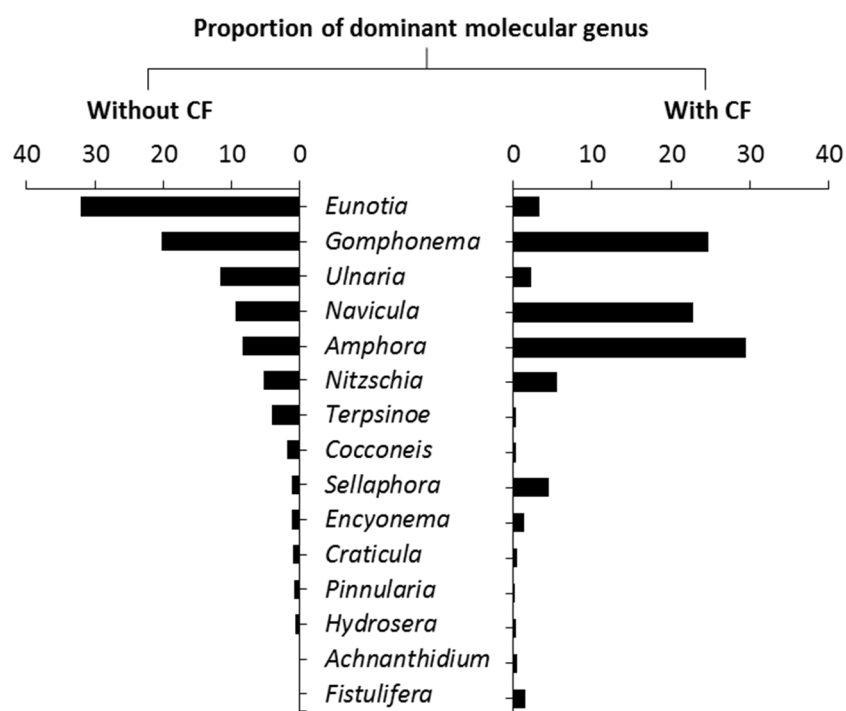


Figure 29 – Dominant taxa (relative abundance > 0.5 %) obtained in HTS Mayotte molecular inventories without (left) and with (right) application of the biovolume correction factor. All samples (n=80) are considered.

For example, the application of CFs reduced the relative abundances of *Eunotia* and *Ulnaria* from 31.9 % to 3.3 % and 11.7 % to 2.3 %, respectively, making them more congruent with cell proportions observed with microscopy (3.1% for *Eunotia* and 0.4 % for *Ulnaria*). The correlation between the morphological and the molecular SPI values for all river samples previously described

($r = 0.72$, $p < 0.001$) was slightly improved using SPI values based on inventories with corrected abundances ($r = 0.77$, $p < 0.001$). The application of the CF to correct the HTS quantification reduced significantly ($p < 0.001$) the differences between the molecular and morphological SPI values by 47 % (Δ SPI reduced to 1.9 on average compared to 3.6 before correction) (**Figure 30**).

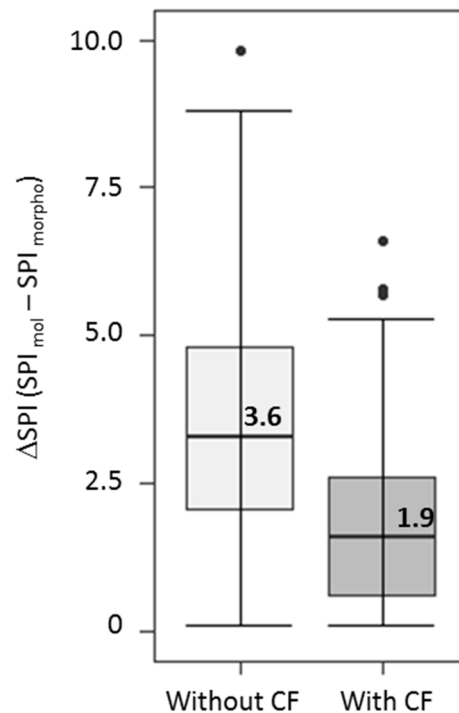


Figure 30 – Distribution of the differences between the molecular and the morphological SPI (Δ SPI) for all Mayotte samples using original molecular SPI values (left) and new molecular SPI values based on molecular inventories corrected with the biovolume CF (right).

5. Discussion

Species quantification based on DNA metabarcoding is challenging for most of taxonomic groups as technical and biological biases affect DNA reads proportions. In order to limit those biases, several attempts were done to apply a CF on metabarcoding data, as shown for fishes (Thomas et al. 2016), bacteria and archaea (Angly et al. 2014) or oligochaetes (Vivien *et al.* 2016). For those studies, application of the CF, whether for correcting single (Angly et al. 2014) or multiple sources of quantification biases (Thomas et al. 2016), improved taxa quantification from metabarcoding data compare to morphological one. The result is generally a change in the ranks of the dominant taxa which affect directly the community structure and can lead to different ecological interpretations. For example, the application of a CF on metabarcoding data obtained from aquatic oligochaetes samples improved the freshwater quality assessment based on

molecular index calculation (Vivien *et al.* 2016). However, the development of CF can be challenging depending on the organism studied, as it requires finding a clear relationship between DNA reads and specimen proportions. This may be impossible due to accumulation of quantification biases (*e.g.* cell density, cell biomass, gene copy number). Nevertheless, the use of CF can be advantageous for organisms with a high variation of the DNA reads proportions between taxa (*e.g.* several log) and where a limited number of biases are involved like diatoms.

5.1. Correlation between *rbcl* gene copy number and diatom cell biovolume: impacts on HTS quantification

The copy number of the *rbcl* gene present in one diatom cell is affected by 3 parameters: (i) the number of chloroplasts per cell, (ii) the number of genomes per chloroplast and (iii) the number of copies of the *rbcl* gene per chloroplast genome (Ersland *et al.* 1981; Treusch *et al.* 2012). (i) For benthic diatoms, the chloroplast number per cell is quite stable inside a single genus with variations ranging from 1 to \approx 8 chloroplast(s) per cell from a genus to another (Round *et al.* 1990b), even if some centric genera may have tens of chloroplasts (*e.g.* *Melosira*, *Cyclotella*). (ii) Regarding the chloroplast genome number per cell, higher plants can contain up to thousands of copies of chloroplast genome per cell (Bendich 1987; Rauwolf *et al.* 2010) while unicellular algae generally exhibit a lower number of copies. For example, *Olisthodiscus luteus* (Raphidophyceae), *Chlamydomonas reinhardtii* (Chlorophyceae), *Phaeodactylum tricorutum* (pennate diatom) and *Thalassiosira pseudonana* (centric diatom) contain respectively around 650, 80, 137 and 55 genome copies per cell (Ersland *et al.* 1981; Koop *et al.* 2007; Gruber 2008; von Dassow *et al.* 2008). (iii) Finally, there is only 1 copy of the *rbcl* gene per chloroplast genome (*e.g.* Sabir *et al.* 2014), as in higher plants (Gutteridge & Gatenby 1995).

Thus, the *rbcl* copy number may vary from tens to hundreds of copies per diatom cell. Our estimations are within this range with a maximum of 130 copies estimated for *P. viridiformis*. However, our method underestimates the *rbcl* gene copy number since 0.5 copy per cell was estimated for *A. minutissimum* (so implying that some cells have no *rbcl* copy). This may result from certain variability inherent to the estimation of gene copy number by qPCR and the quantification of cells by microscopical counts. Our results demonstrate, however, that the *rbcl* copy number varies significantly between the 8 diatoms species used in this study, according to the different diatom cell characteristics tested. In particular, we found a significant linear relationship between the *rbcl* copy number and the cell biovolume. Although the size of the

chloroplasts could not be estimated in this study, we assume that the increase of the cell biovolume is accompanied by an increase of the chloroplast biovolume (as shown by Okie, Smith & Martin-Cereceda 2016), inducing an increase of DNA quantity and chloroplast genome copies per chloroplast as shown by Rauwolf et al. (2010).

The correlation we found between the *rbcl* copy number and the diatom cell biovolume suggests that the relative abundance of diatom species with high cell biovolume is likely to be over-represented in metabarcoding data compared to microscopical counts. This is confirmed by the HTS data obtained for the mock communities, where diatom species with high cell biovolume are over-represented (e.g. *P. viridiformis*) and diatom species with low cell biovolume are under-represented (e.g. *A. minutissimum*). The relative abundance of *P. viridiformis* in the mock communities was negligible compared to other species, and doubling its proportion did not change its rank: the species remained the least abundant taxon within the morphological inventory. However, due to its high cell biovolume ($10^4 \mu\text{m}^3$) and relatively high *rbcl* copy number per cell, a marked over-representation of this species within the molecular inventory was observed. A CF was thus defined to correct these quantitative biases and was verified on mock communities and environmental samples.

5.2. Current potential and limits of the quantification CF

The use of the same *rbcl* primers for the qPCR assays and the HTS enabled us to generate a specific CF well suited to correct *rbcl* metabarcoding quantifications. Its application to the HTS data of the mock communities allowed us to obtain comparable species proportions in morphological and molecular based approaches of mock communities. This was also confirmed with the Mayotte river samples, for which the quantification CF resulted in a better congruence between DNA reads and cells proportions, reducing the over-representation of high biovolume *Eunotia* and *Ulnaria* species. Furthermore, SPI calculation based on corrected metabarcoding data gives SPI values more comparable to SPI values obtained from morphological data, suggesting that it may be possible to replace morphological by molecular monitoring for the ecological assessment of Mayotte rivers. In the same way, (Vivien *et al.* 2016) have shown that application of a CF to correct DNA reads proportions allows a more accurate estimation of oligochaete proportions, improving quality index calculation and quality assessment of watercourse sediments. Our results confirm that water quality index based on diatom metabarcoding and DNA read proportions are directly affected by gene copy number variation, and show the potential value of integrating CFs into molecular SPI calculation. However, as the biovolume–copy number relationship was based

on only 8 diatom species and the efficiency of the resulting CFs validated on only one HTS dataset, further experiments including more species and larger datasets will be required to develop and fully validate CFs for use in molecular biomonitoring.

The CF developed in the present study assumes that gene copy number is constant in each taxon. However, gene copy number may vary with the physiological status of the cell and stage of the life cycle, since in most diatoms cell volume decreases during the vegetative phase. The physiological status varies with cell cycle progression; additionally several factors may affect the physiological status of diatoms like changes in environmental conditions (*e.g.* nutrient availability, pollutants, temperature ...) (Pandey et al. 2017). Altered physiological status of a given population is generally characterized by a higher proportion of damaged cells. The compromised/damaged cells are characterized by alteration of membrane integrity, degradation of the photosynthetic pigments or fragmentation of genomic DNA (Zetsche & Meysman 2012; Znachor et al. 2015). Variations of DNA integrity and chloroplast physiology between cells of a given population can impact directly the *rbcl* gene copy number per cell and thus DNA metabarcoding quantification. (Eberhard *et al.* 2002) showed that chloroplast genome copy number is reduced when the green alga *Chlamydomonas reinhardtii* is cultivated under phototrophic conditions compared to cultivation in mixotrophic conditions. Limitation by mineral nutrients may also have an impact; for instance iron limitation can reduce the number of the chloroplast per cell (from 4 to 2) and their size in the marine diatom *Thalassiosira oceanica* (Hustedt) Hasle et Heimdal (Lommer et al. 2012). Variation of the cell physiological state was not taken into account in developing CFs for diatom metabarcoding. However, during our experiments we discriminated live and dead cells; we observed that their respective proportions did not affect significantly the correlation between the gene copy number per cell and the cell biovolume (Fig. S1). Further experiments should be performed to evaluate the impact on the final CFs of *rbcl* gene copy number variation linked to physiological status.

The biovolume of each diatom species is required to apply the CF and hence correct the quantification in metabarcoding datasets. Several reference databases provide biovolume information for a lot diatom species (*e.g.* Rimet et al. 2016), but they do not generally account for biovolume variability, which is a complicating factor in diatoms because of the peculiarities of the life cycle. Diatom cell size within a population is not constant due to the method of vegetative reproduction, which leads to a progressive cell size reduction of the population (Crawford 1981), followed by restoration of cell size via a sexual event. For this reason, different cell sizes can be observed in the same diatom population, either in pure cultures of (*e.g.* in the marine diatom *Thalassiosira weissflogii* Grunow: Armbrust & Chisholm 1992) or in environmental populations

(e.g. the freshwater species *Sellaphora pupula* (Kützing) Mereschk: (Mann *et al.* 1999). However, although the range of cell sizes within a given diatom population may vary by a factor of 2 to 5 in the environment (Hense & Beckmann 2015), natural populations usually have a rather narrow range of sizes and larger cells form a negligible fraction of the population (Mann 2011). Furthermore, the distribution of cell size within environmental populations is often close to being normal (Mann *et al.* 1999; Spaulding *et al.* 2012). The balance between small and big individuals in the same population will therefore limit errors associated with the use of a mean biovolume. Hence, we propose to use the mean of biovolume to calculate CFs; without considering other potential HTS quantification biases, its application to DNA reads of environmental material should allow a good correction of their proportions.

6. Acknowledgments

The authors declare no conflict of interest. Funding provided by the French National Agency for Water and Aquatic Environments (ONEMA-AFB) and supported by the European COST action DNAqua-Net (CA 15219). A special thanks to David G. Mann for the constructive discussions that helped to improve the manuscript.

7. Data accessibility

All PGM raw sequence data are available for the 15 libraries, corresponding to the 5 DNA mock communities with 3 replicates, on the Zenodo repository website (<http://doi.org/10.5281/zenodo.807178>).

8. Author contributions

V.V., A.B., F.R., S.J., M.C., K.T., I.D contributed to the study designed. V.V., M.C and S.J. conducted the laboratory work. V.V. analyzed the data and wrote the manuscript. All the authors contributed to the discussions and to manuscript editing.

9. Supplementary data

Table S1 – *rbcl* primers, qPCR reactions mix and condition used for the qPCR assays. Information is provided for 1 reaction in a final volume of 25 μ L.

Primer name		Primer sequence (5' - 3')	Length (bp)
<i>Forward</i>	Diat_ <i>rbcl</i> _708F_1	AGGTGAAGTAAAAGGTTTCWTTACTTAAA	27
	Diat_ <i>rbcl</i> _708F_2	AGGTGAAGTTAAAGGTTTCWTAYTTAAA	27
	Diat_ <i>rbcl</i> _708F_3	AGGTGAAACTAAAGGTTTCWTTACTTAAA	27
<i>Reverse</i>	R3_1	CCTTCTAATTTACWACWACTG	22
	R3_2	CCTTCTAATTTACWACAACAG	22

Reagents	Initial conc.	Final conc.	Volume (μ L)
Sybr MIX	2X	1X	12.5
H ₂ O molecular grade	-	-	6.75
Forward (Diat_ <i>rbcl</i> _708F_1 + _2 + _3)	10 μ M	0.5 μ M	1.25
Reverse (R3_1 + R3_2)	10 μ M	0.5 μ M	1.25
Bovine Serum Albumin (BSA)	10 mg/mL	0.5 mg/mL	1.25
DNA	25 ng/ μ L	2 ng/ μ L	2

Step	Time (s)	Temperature ($^{\circ}$ C)	Cycles
1	900	95	
2	45	95	X 40
3	45	55	
4	45	72	
5	1 $^{\circ}$ every 5s	60 to 95	

Table S2 – Estimation of the diatom cell concentration and the live/dead cell proportion per mL of media, based on microscopy counts, for the 8 diatom species at each sampling time and for the 3 replicates (A, B, C). Mean values of cell concentration per mL of media, which only take into account living cells, is provided and used for the calculation of *rbcL* copy number per diatom cell (bold values).

Species	Sampling time	Days after inoculation	[cell.mL ⁻¹] per replicate			% of dead cell			Mean (living cells) (cell.mL ⁻¹)
			A	B	C	A	B	C	
Cmen	T0	5	3.7E+02	4.1E+02	4.2E+02	9.8	3.8	4.0	3.8E+02
	T1	10	8.2E+03	6.0E+03	8.5E+03	6.1	14.5	11.9	6.8E+03
	T2	13	1.2E+04	1.1E+04	2.0E+04	13.5	16.1	12.0	1.2E+04
	T3	20	1.2E+05	5.7E+04	1.3E+05	20.2	13.9	19.6	8.1E+04
	T4	25	2.6E+05	4.1E+05	3.2E+05	53.9	51.2	56.3	1.5E+05
	T5	31	2.0E+05	2.4E+05	2.3E+05	59.9	50.4	48.2	1.0E+05
	T6	38	4.6E+05	5.5E+05	2.6E+05	59.1	55.1	57.1	1.8E+05
Npal	T0	5	1.8E+04	2.1E+04	3.9E+04	0.0	0.0	1.0	2.6E+04
	T1	10	6.1E+05	4.9E+05	4.1E+05	0.0	0.0	0.0	5.1E+05
	T2	13	4.6E+05	4.8E+05	5.2E+05	0.9	1.9	1.0	4.8E+05
	T3	17	4.8E+05	3.9E+05	6.2E+05	5.9	4.0	6.7	4.7E+05
	T4	25	4.1E+05	4.3E+05	9.4E+05	15.4	9.9	8.1	5.3E+05
	T5	34	6.2E+05	7.2E+05	7.0E+05	23.5	30.3	25.0	5.0E+05
	T6	40	1.3E+06	1.1E+06	6.4E+05	46.6	38.5	54.1	5.6E+05
Uuln	T0	5	8.2E+03	7.9E+03	1.5E+04	3.8	2.8	0.7	1.0E+04
	T1	10	1.3E+04	1.2E+04	1.5E+04	5.4	7.8	7.2	1.2E+04
	T2	13	1.2E+04	3.4E+04	7.9E+03	14.3	13.6	10.5	1.5E+04
	T3	20	1.8E+04	1.6E+04	2.6E+04	27.1	23.8	23.9	1.5E+04
	T4	31	1.6E+04	1.1E+04	9.2E+03	83.3	74.4	63.9	3.0E+03
	T5	38	8.6E+03	9.2E+03	3.6E+04	82.8	84.8	82.2	3.1E+03
Ninc	T0	5	3.9E+03	8.7E+03	5.8E+03	0.5	1.0	0.0	6.1E+03
	T1	10	3.5E+05	3.9E+05	4.3E+05	0.0	0.2	0.2	3.9E+05
	T2	12	4.3E+05	2.6E+05	1.1E+06	0.6	0.2	0.7	5.9E+05
	T3	17	4.1E+05	6.4E+05	1.1E+06	6.9	7.0	5.1	6.8E+05
	T4	25	1.7E+06	1.4E+06	9.9E+05	11.6	10.4	12.6	1.2E+06
	T5	34	1.6E+06	1.3E+06	1.4E+06	7.2	9.9	6.7	1.3E+06
	T6	40	1.3E+06	1.9E+06	1.7E+06	21.4	28.9	30.8	1.2E+06
Dten	T0	12	1.2E+04	4.3E+04	2.6E+04	0.2	0.0	0.3	2.7E+04
	T1	17	1.1E+05	9.4E+04	1.0E+05	5.7	7.0	5.5	9.6E+04
	T2	20	1.8E+05	2.2E+05	1.3E+05	6.9	5.7	5.5	1.7E+05
	T3	25	4.9E+05	2.3E+05	1.4E+05	6.3	8.8	8.2	2.7E+05
	T4	34	2.7E+05	2.0E+05	2.1E+05	26.5	35.8	43.1	1.5E+05
	T5	38	4.1E+05	2.4E+05	1.6E+05	48.3	49.3	45.5	1.4E+05
Pvir	T0	13	6.0E+02	4.7E+02	4.1E+02	8.0	7.5	11.7	4.5E+02
	T1	20	9.6E+02	7.2E+02	1.1E+03	12.0	9.3	7.3	8.3E+02
	T2	31	1.5E+03	1.7E+03	3.1E+03	14.3	17.3	18.5	1.8E+03
	T3	34	2.0E+03	2.0E+03	2.4E+03	16.5	23.5	29.6	1.6E+03
	T4	40	2.7E+03	2.2E+03	3.6E+03	26.1	22.6	26.7	2.1E+03
	T5	73	4.9E+03	2.7E+03	2.6E+03	83.7	75.8	66.8	7.7E+02

Fper	T0	12	6.0E+04	3.4E+04	3.3E+04	0.7	0.7	1.3	4.2E+04
	T1	17	2.7E+05	1.1E+05	1.7E+05	14.6	12.6	11.6	1.6E+05
	T2	20	2.2E+05	1.6E+05	1.2E+05	23.4	24.1	19.7	1.3E+05
	T3	25	1.5E+05	1.8E+05	1.6E+05	62.2	65.4	62.3	6.0E+04
	T4	31	6.6E+05	3.0E+06	4.4E+05	69.3	73.8	65.5	3.8E+05
	T5	34	1.2E+06	3.2E+05	2.6E+05	78.5	74.8	76.8	1.3E+05
	T6	40	2.9E+05	5.8E+05	5.4E+05	82.5	75.8	73.5	1.1E+05
Amin	T0	12	1.8E+03	6.2E+03	3.7E+03	0.7	1.7	1.0	3.9E+03
	T1	17	3.0E+04	7.4E+04	8.4E+04	4.1	3.7	3.0	6.0E+04
	T2	25	5.5E+05	4.0E+05	1.4E+06	4.7	7.7	4.6	7.5E+05
	T3	31	1.3E+06	1.0E+06	5.2E+05	13.1	13.1	10.2	8.3E+05
	T4	34	2.1E+06	2.9E+06	6.7E+05	11.6	10.5	13.8	1.7E+06
	T5	38	2.7E+06	1.2E+06	5.6E+05	15.2	11.4	16.9	1.3E+06
	T6	40	2.8E+06	2.7E+06	1.7E+06	16.2	11.5	17.5	2.0E+06

Table S3 – Estimation of the *rbcl* copy number per mL of media determined by qPCR for the 8 diatom species at each sampling time and for the 3 replicates (A, B, C). Mean values of *rbcl* concentration per mL of media is provided and used for the calculation of *rbcl* copy number per diatom cell (bold values).

Species	Sampling time	Days after inoculation	[<i>rbcl</i>] (copy.mL ⁻¹)			Mean (copy.mL ⁻¹)
			A	B	C	
Cmen	T0	5	7.1E+03	7.3E+03	1.1E+04	8.4E+03
	T1	10	4.8E+04	2.2E+04	1.3E+04	2.8E+04
	T2	13	4.6E+04	2.6E+04	2.4E+04	3.2E+04
	T3	20	1.2E+05	1.1E+05	1.9E+05	1.4E+05
	T4	25	6.1E+05	6.2E+05	7.6E+05	6.6E+05
	T5	31	4.3E+05	2.3E+06	5.3E+05	4.8E+05
	T6	38	9.4E+05	1.0E+06	7.3E+05	9.1E+05
Npal	T0	5	3.8E+04	3.4E+04	7.0E+04	4.7E+04
	T1	10	1.3E+06	1.5E+06	9.1E+05	1.3E+06
	T2	13	2.5E+06	2.8E+06	2.7E+06	2.6E+06
	T3	17	2.5E+06	2.7E+06	2.6E+06	2.6E+06
	T4	25	3.3E+06	2.3E+06	3.0E+06	2.9E+06
	T5	34	1.7E+06	1.9E+06	2.2E+06	2.0E+06
	T6	40	1.1E+06	1.3E+06	8.4E+05	1.1E+06
Uuln	T0	5	1.4E+05	2.5E+05	1.6E+05	1.8E+05
	T1	10	4.0E+05	3.3E+05	3.1E+05	3.5E+05
	T2	13	1.2E+05	1.1E+05	7.5E+04	1.0E+05
	T3	20	4.9E+05	1.8E+05	2.6E+05	3.1E+05
	T4	31	1.2E+05	1.4E+05	2.2E+05	1.6E+05
	T5	38	7.5E+04	5.6E+04	5.7E+04	6.3E+04
Ninc	T0	5	1.1E+04	1.3E+04	1.5E+04	1.3E+04
	T1	10	3.5E+05	7.2E+05	5.2E+05	5.3E+05
	T2	12	8.1E+05	6.3E+05	1.1E+06	8.3E+05
	T3	17	9.9E+06	8.6E+06	7.5E+06	8.7E+06
	T4	25	4.7E+06	5.1E+06	6.3E+06	5.4E+06
	T5	34	7.3E+06	7.8E+06	8.1E+06	7.7E+06
	T6	40	4.8E+06	4.5E+06	3.2E+06	4.2E+06
Dten	T0	12	4.7E+05	2.1E+05	3.0E+05	3.3E+05
	T1	17	1.5E+06	2.0E+06	1.1E+06	1.6E+06
	T2	20	7.6E+05	1.4E+06	2.8E+06	1.7E+06
	T3	25	1.3E+06	5.0E+05	4.6E+05	7.5E+05
	T4	34	4.3E+05	2.3E+05	5.3E+05	4.0E+05
	T5	38	3.2E+05	4.5E+05	2.3E+05	3.4E+05
Pvir	T0	13	7.9E+04	4.6E+04	7.1E+04	6.6E+04
	T1	20	1.2E+05	1.2E+05	1.2E+05	1.2E+05
	T2	31	2.0E+05	1.3E+05	2.0E+05	1.8E+05
	T3	34	1.6E+05	2.6E+05	3.0E+05	2.4E+05
	T4	40	2.6E+05	2.2E+05	3.8E+05	2.9E+05
	T5	73	3.1E+05	5.5E+05	4.8E+05	4.5E+05
Fper	T0	12	1.4E+04	3.4E+03	9.4E+03	9.0E+03
	T1	17	3.0E+05	2.4E+05	3.6E+05	3.0E+05

	T2	20	8.3E+05	7.1E+05	9.2E+05	8.2E+05
	T3	25	8.1E+05	4.8E+05	1.4E+06	8.8E+05
	T4	31	4.4E+05	4.8E+05	4.4E+05	4.5E+05
	T5	34	6.7E+05	6.8E+05	1.0E+06	8.0E+05
	T6	40	4.0E+05	4.7E+05	4.5E+05	4.4E+05
Amin	T0	12	1.8E+03	2.8E+03	3.6E+03	2.7E+03
	T1	17	3.4E+04	1.5E+04	2.7E+04	2.6E+04
	T2	25	1.9E+05	1.7E+05	2.2E+05	1.9E+05
	T3	31	1.2E+05	1.6E+05	1.6E+05	1.5E+05
	T4	34	2.6E+05	3.6E+05	3.1E+05	3.1E+05
	T5	38	2.8E+05	3.2E+05	2.6E+05	2.9E+05
	T6	40	5.1E+05	3.6E+05	2.0E+05	3.6E+05

Table S4 – CF calculated for the 84 diatom taxa detected in Mayotte environmental samples. Calculation performed using the respective cell biovolume of each taxa (available in the R-Syst::diatom library) and the linear equation between the *rbcL* copy number and the cell biovolume produced in the Fig. 3.

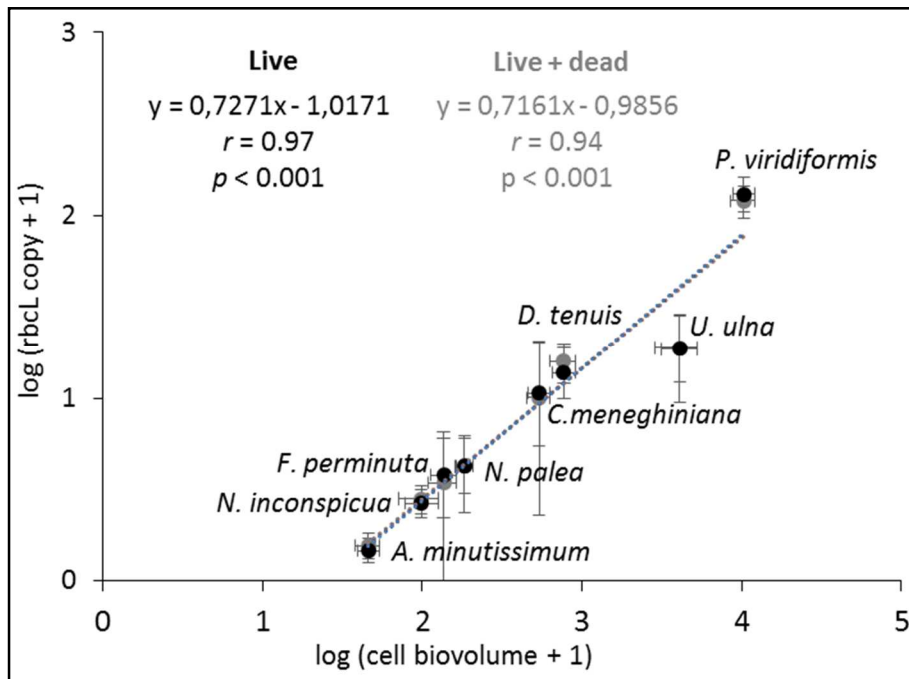
Diatom taxa	Biovolume (μm^3)	Calculated CF
<i>Achnanthes_coarctata</i>	53	0.7
<i>Achnantheidium_helveticum</i>	316	5.3
<i>Achnantheidium_minutissimum</i>	76	1.3
<i>Achnantheidium_sp.</i>	76	1.3
<i>Amphora_pediculus</i>	72	1.2
<i>Amphora_sp.</i>	20096	128.3
<i>Caloneis_silicula</i>	1994	23.1
<i>Caloneis_sp.</i>	523	8.1
<i>Cocconeis_placentula</i>	2963	31.2
<i>Craticula_cuspidata</i>	2850	30.3
<i>Craticula_molestiformis</i>	119	2.1
<i>Cyclotella_sp.</i>	328	5.5
<i>Cymbella_excisa</i>	520	8.1
<i>Cymbella_heterogibbosa</i>	5817	51.5
<i>Cymbella_sp.</i>	520	8.1
<i>Cymbopleura_naviculiformis</i>	1148	15.1
<i>Encyonema_minutum</i>	213	3.8
<i>Encyonema_muelleri</i>	12784	92.1
<i>Encyonema_silesiacum</i>	821	11.7
<i>Encyonema_sp.</i>	213	3.8
<i>Eolimna_subminuscula</i>	112	2.0
<i>Epithemia_sp.</i>	5967	52.5
<i>Eunotia_bilunaris</i>	617	9.3
<i>Eunotia_minor</i>	755	10.9
<i>Eunotia_pectinalis</i>	4219	40.6
<i>Eunotia_sp.</i>	15700	107.1
<i>Fallacia_pygmaea</i>	1229	16.0
<i>Fistulifera_saprophila</i>	14	0.03
<i>Fragilaria_sp.</i>	294	5.0
<i>Frustulia_vulgaris</i>	1625	19.8
<i>Frustulia_sp.</i>	1625	19.8
<i>Gomphonema_acuminatum</i>	1860	21.9
<i>Gomphonema_affine</i>	926	12.8
<i>Gomphonema_bourbonense</i>	270	4.6
<i>Gomphonema_cleveii</i>	484	7.6
<i>Gomphonema_parvulum</i>	331	5.5
<i>Gomphonema_sp.</i>	510	8.0
<i>Halamphora_montana</i>	161	2.9
<i>Halamphora_sp.</i>	161	2.9
<i>Hydrosera_sp.</i>	500	7.8
<i>Lemnicola_hungarica</i>	436	7.0
<i>Luticola_sparsipunctata</i>	176	3.1
<i>Mayamaea_permitis</i>	66	1.0
<i>Navicula_cryptocephala</i>	431	6.9
<i>Navicula_cryptotenella</i>	386	6.3
<i>Navicula_lanceolata</i>	1227	15.9
<i>Navicula_radiosa</i>	1852	21.9
<i>Navicula_rostellata</i>	854	12.0
<i>Navicula_sp.</i>	88	1.5
<i>Navicula_symmetrica</i>	818	11.6
<i>Navicula_tripunctata</i>	966	13.2

<i>Navicula_veneta</i>	279	4.8
<i>Neidium_sp.</i>	240	4.2
<i>Nitzschia_amphibia</i>	334	5.6
<i>Nitzschia_filiformis</i>	737	10.7
<i>Nitzschia_fonticola</i>	344	5.7
<i>Nitzschia_inconspicua</i>	89	1.5
<i>Nitzschia_lorenziana</i>	1362	17.3
<i>Nitzschia_palea</i>	391	6.4
<i>Nitzschia_sp.</i>	307	5.2
<i>Nitzschia_tubicola</i>	336	5.6
<i>Pinnularia_divergens</i>	3908	38.3
<i>Pinnularia_subanglica</i>	1188	15.6
<i>Pinnularia_subgibba</i>	3454	35.0
<i>Pinnularia_sp.</i>	1258	16.3
<i>Placoneis_clementis</i>	1123	14.9
<i>Placoneis_elginensis</i>	1266	16.3
<i>Planothidium_sp.</i>	267	4.6
<i>Rhopalodia_gibba</i>	185472	649.8
<i>Rhopalodia_sp.</i>	185472	649.8
<i>Sellaphora_minima</i>	88	1.5
<i>Sellaphora_pupula</i>	1183	15.5
<i>Sellaphora_seminulum</i>	69	1.1
<i>Sellaphora_sp.</i>	88	1.5
<i>Seminavis_robusta</i>	5308	48.1
<i>Staurosira_elliptica</i>	29	0.1
<i>Staurosira_sp.</i>	315	5.3
<i>Stephanodiscus_hantzschii</i>	670	9.9
<i>Surirella_sp.</i>	1034	14.0
<i>Tabellaria_flocculosa</i>	500	7.8
<i>Terpsinoe_musica</i>	10563	80.0
<i>Tryblionella_sp.</i>	655	9.7
<i>Ulnaria_ulna</i>	4724	44.1
<i>Ulnaria_sp.</i>	5260	47.8

Table S5 – Number of DNA reads assigned to the 8 species in each of the 5 DNA mock communities. A, B, and C represent the 3 replicates.

Species	Mock 1			Mock 2			Mock 3			Mock 4			Mock 5		
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
<i>A. minutissimum</i>	2828	1934	2410	1785	2129	2109	1837	1900	1882	2025	1342	1683	1202	1273	1332
<i>N. inconspicua</i>	5480	3484	4648	3673	4533	4083	3777	3622	3824	3920	3074	3741	2462	2588	2571
<i>N. palea</i>	1452	1059	1126	912	850	1037	718	896	904	899	715	888	695	567	634
<i>P. viridiformis</i>	2573	1966	2066	2372	2823	2999	6440	7861	7461	11586	10430	11722	18424	16703	14159
<i>D. tenuis</i>	5311	3423	4552	3286	4461	3172	4578	3377	3376	4013	2679	3442	2206	2861	2522
<i>F. perminuta</i>	5817	3796	4452	3484	3844	3549	3492	3569	3341	3427	2449	3083	2117	2318	2226
<i>U. ulna</i>	4486	3037	3863	3303	3893	3561	3259	3343	3449	3321	2412	2897	2395	2053	1992
<i>C. meneghiniana</i>	1360	844	984	1202	1344	1204	1129	1113	1235	1137	807	1126	994	869	779

Figure S1 – Correlation between the diatom cell biovolume and the *rbcl* gene copy number per cell after $\log(x+1)$ transformation based on live (black) or live/dead (grey) microscopical counts. Linear equation of the model and the Pearson correlation coefficient (r) with its associated p -value are indicated.



IV. Application aux cours d'eau de Mayotte

**“Assessing ecological status with diatoms DNA metabarcoding:
scaling-up on a WFD monitoring network (Mayotte island, France)”**

(paru dans le journal Ecological Indicators, 2017)

Valentin Vasselon, Frédéric Rimet, Kálmán Tapolczai, Agnès Bouchez

CARTELE, INRA, Université de Savoie Mont Blanc, 74200, Thonon-les-bains, France

1. Abstract

Diatoms are excellent ecological indicators of water quality because they are broadly distributed, they show high species diversity and they respond rapidly to human pressures. In Europe, the Water Framework Directive (WFD) gives the legal basis for the use of this indicator for water quality assessment and its management. Several quality indices, like the Specific Polluosensitivity Index (SPI), were developed to assess the ecological quality status of rivers based on diatom communities. It is based on morphological identifications and count of diatom species present in natural biofilms using a microscope. This methodology requires high taxonomic skills and several hours of analysis per sample as 400 individuals must be identified to species level. Since several years, a molecular approach based on DNA metabarcoding combined to High-Throughput Sequencing (HTS) is developed to characterize species assemblages in environmental samples which is potentially faster and cheaper. The ability of this approach to provide reliable diatom inventories has been demonstrated and its application to water quality assessment is currently being improved. Despite optimization of the DNA metabarcoding process with diatoms, few studies had yet extended it at the scale of a freshwater monitoring network and evaluated the reliability of its quality assessment compared to the classical morphological approach.

In the present study we applied DNA metabarcoding to the river monitoring network of the tropical Island Mayotte. This island is a French département since 2011 and the WFD has to be applied. This offered the opportunity to scale up the comparison of molecular and morphological approaches and their ability to produce comparable community inventories and water quality assessments. Benthic diatoms were sampled following WFD standards in 45 river sites in 2014 and 2015 (80 samples). All samples were submitted in parallel to the molecular and the morphological approaches. DNA metabarcoding was carried out using Genelute DNA extraction method, *rbcL* DNA barcode and PGM sequencing, while microscopic counts were carried out for the classical methodology. Diatom community structures in terms of molecular (OTUs) and of morphological (species) were significantly correlated. However, only 13% of the species was shared by both approaches, with qualitative and quantitative variation due to i) the incompleteness of the reference library (82% of morphological species are not represented in the database), ii) limits in taxonomic knowledge and iii) biases in the estimation of relative abundances linked to diatom cell biovolume. However, ecological quality status assessed with the molecular and morphological SPI values were congruent, and little affected by sequencing depth. DNA metabarcoding of diatom communities allowed a reliable estimation of the quality status for most of the rivers at the scale of the full biomonitoring network of Mayotte Island.

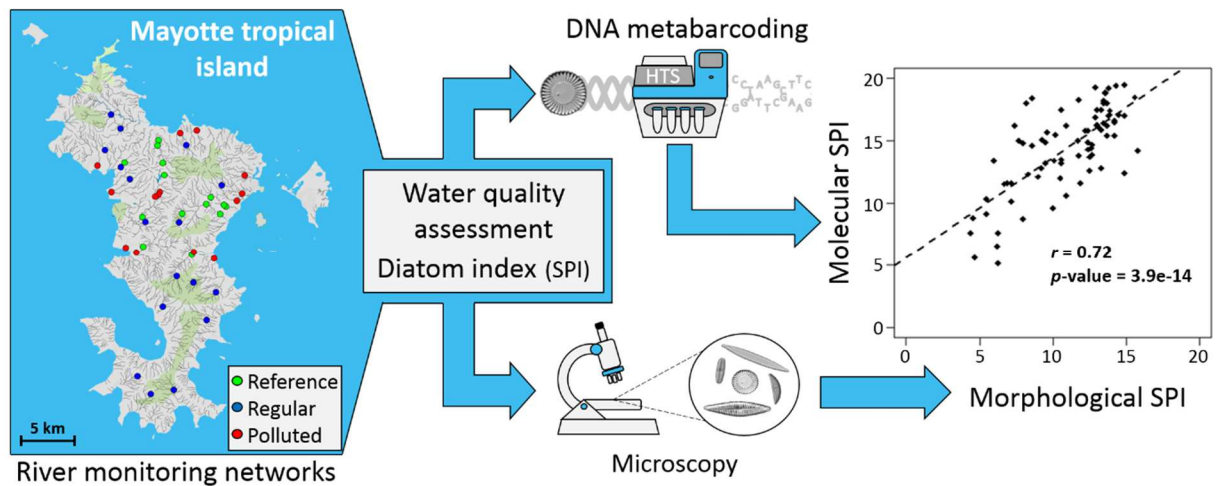


Figure 31 – Graphical abstract.

2. Introduction

Biological indicators are commonly used by environmental agencies for water quality assessment (Ibáñez *et al.* 2010; Birk *et al.* 2012). Diatoms, a group of microalgae, are known to be efficient indicators of river ecological quality and are required to be monitored in rivers by transnational directives as the Water Framework Directive in Europe (European Council 2000). Their indicator efficiency relies on their huge taxonomical diversity and their species ecological preferences to particular pollution levels (Pandey *et al.* 2017). After collecting natural diatom communities from benthic biofilms, the relative frequencies of diatom taxa are used together with their ecological optimum and tolerance values to compute biotic indices often derived from the Zelinka and Marvan formula (Zelinka & Marvan 1961).

Some indices are based on a restricted list of taxa, adapted to local diatom biodiversity and to type of pressures. For example, the French WFD index, BDI (Lenoir & Coste 1996), is based on a list of 800 taxa with their associated autecology. The Swiss DI-CH (Hürlimann & Niederhauser 2007) is based on a restricted list of only 188 taxa. At the opposite, the SPI (Specific Polluosensitivity Index, Cemagref 1982; Coste 1986) was developed to evaluate overall water quality in terms of organic pollution and nutrient levels and is though encountering a much larger set of diatom species (over 2000). Due to its large taxonomical and ecological base, its efficiency to assess ecological quality in a large range of rivers in Europe has been demonstrated (*e.g.* Kelly 2013). Though, SPI is the index currently used to apply WFD to rivers in several European countries like in Portugal, Belgium, Bulgaria, Netherlands, Sweden, Luxembourg or Spain (Kelly 2013). Even

in regions with a weaker taxonomical and ecological knowledge about diatom species, SPI is used as a reference index to reveal quality gradients in urban rivers of boreal regions (*e.g.* Teittinen *et al.* 2015) or in Chinese rivers (*e.g.* Yang *et al.* 2015).

However, applying such indices requires time and a high level of taxonomical expertise. Current standardized methods (*e.g.* CEN) call for the determination of diatoms until a minimum 400 individuals (diatom valves) at species or sub-species level using light microscopy and several tens of iconographical books. The development of DNA metabarcoding and High-Throughput Sequencing offered a solution, allowing to investigate prokaryote and eukaryote biodiversity present in environmental samples (Creer 2010). Assessing taxonomic inventories of environmental communities of macroinvertebrates based on DNA metabarcoding has been shown by Hajibabaei *et al.* (2011) as a promising alternative to morphological methodologies for biomonitoring (Keck *et al.* 2017). Same hopes were raised for diatoms testing DNA metabarcoding on mock communities (Kermarrec *et al.* 2013b) and environmental communities (Kermarrec *et al.* 2014). The SPI index values calculated for each diatom taxonomic inventory based on metabarcoding data enabled authors to assign quality classes to the environmental samples with the same ranking than using morphological methodologies. Later studies (Zimmermann *et al.* 2015; Visco *et al.* 2015; Vasselon *et al.* 2017a) confirmed the possibility of using DNA metabarcoding of diatom communities for environmental studies and biomonitoring. Current genes used for diatoms barcoding are the V4 region of the genomic gene 18S (Zimmermann *et al.* 2015; Visco *et al.* 2015) and the plastid gene *rbcl* (Kermarrec *et al.* 2014; Vasselon *et al.* 2017a). According to Kermarrec *et al.* (2013b, 2014) who compared both barcodes, *rbcl* polymorphism proved to be compatible with a detection at species level, while 18S was more efficient at genus level. The recent release of the R-Syst::diatom barcoding library (Rimet *et al.* 2016) in open-access (<http://www.rsyst.inra.fr/>) offers an expert and curated data on *rbcl* with more than 2500 *rbcl* sequences related to their taxonomic identity, which represents more than 900 species and 200 genera (20-03-2017: R-Syst::diatom v6). However, to date studies confirming the validation of the DNA metabarcoding approach for water quality assessment were done only at small scales (Kermarrec *et al.* 2014: 4 samples, Zimmermann *et al.* 2015: 7 samples, Visco *et al.* 2015: 27 samples, Vasselon *et al.* 2017b: 8 samples) and at a regional scale (Apothéloz-Perret-Gentil *et al.* 2017: 2 Switzerland cantons with 87 samples). We propose here to scale up the test to a large monitoring network, including larger gradients of ecological status.

Mayotte Island, a tropical island part of Comoros archipelago in Mozambique Channel became a French département in 2011, therefore it is now subject to the European regulations,

despite its distance to mainland Europe. We took advantage of the river monitoring network set up to develop WFD indices for Mayotte in order to scale up the DNA metabarcoding approach for diatoms. A total of 80 samples collected at 45 sites were used to compare the molecular and the classical morphological approaches. First, we compared both approaches through (i) taxonomic composition (using OTUs, species, genus and family levels), (ii) diversity (Shannon index), (iii) richness (Chao richness estimator), (iv) structure of diatom community (Bray-Curtis dissimilarity index). Second, biases that may affect the molecular SPI values were investigated, including i) the incompleteness of the reference library, ii) the importance of the taxonomic knowledge, iii) the diatom cell biovolume, and iv) the sequencing depth. Finally, the ability of DNA metabarcoding to produce congruent ecological quality status at the scale of a monitoring network was evaluated.

3. Material and methods

3.1. Mayotte island monitoring network

Mayotte is a French tropical island (374 km²), part of the Comoros archipelago located in the Indian Ocean, in the north-west of Madagascar, in the Mozambique Channel (12°50'35"S 45°08'18"E, **Figure 32**).

Large ecological gradients are encountered in Mayotte rivers where pollution originated from two main anthropogenic pressures. First, a major part of households is not connected to sewage, thus their wastewater is often released directly in the rivers, contributing to an elevated organic matter concentration. Second, clothing is usually washed directly in rivers resulting in high turbidity and suspended solids values due to washing powders. To that aim, a river monitoring network has been set up recently.

In Mayotte, a regular WFD monitoring network (Réseau de Contrôle de Surveillance – RCS) was monitored since 2008. This network groups 14 sites under intermediate conditions, generally just upstream agglomerations. Complementary sampling sites were added in 2014 in order to represent better the different environmental conditions in the rivers of Mayotte, based on their general conditions. Upstream sites (17 sites) under good conditions were grouped in the Reference sites network (REF). Downstream sites (14 sites) under degraded conditions and outside the zone of influence of tides were grouped in the Polluted sites network (POLL).

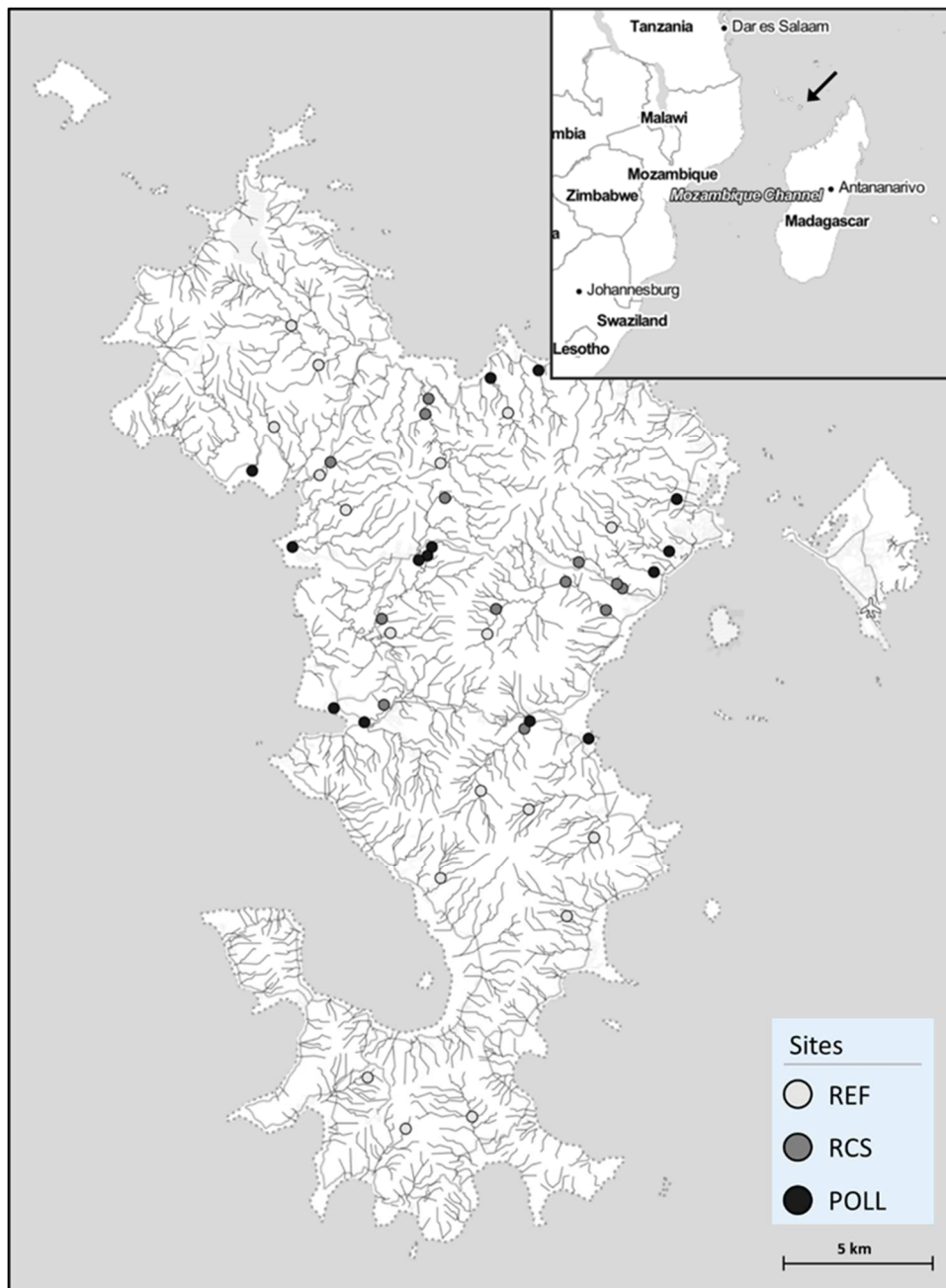


Figure 32 – Graphical abstract Location of Mayotte island (France) and the 45 river sites of the three monitoring networks: Reference sites network (REF - white), Regular WFD monitoring network (RCS – grey) and Polluted sites network (POLL – black).

Samples were taken from 45 sampling sites of 33 different rivers of the main island “Grande Terre” (363 km²) in 2014 and 2015 (**Figure 32**). The relevance of this a priori classification was confirmed by (Tapolczai *et al.* 2017) which showed a strong gradient of organic pollution, turbidity and suspended solids, increasing from REF to RCS and then to POLL networks (**Figure 33**).

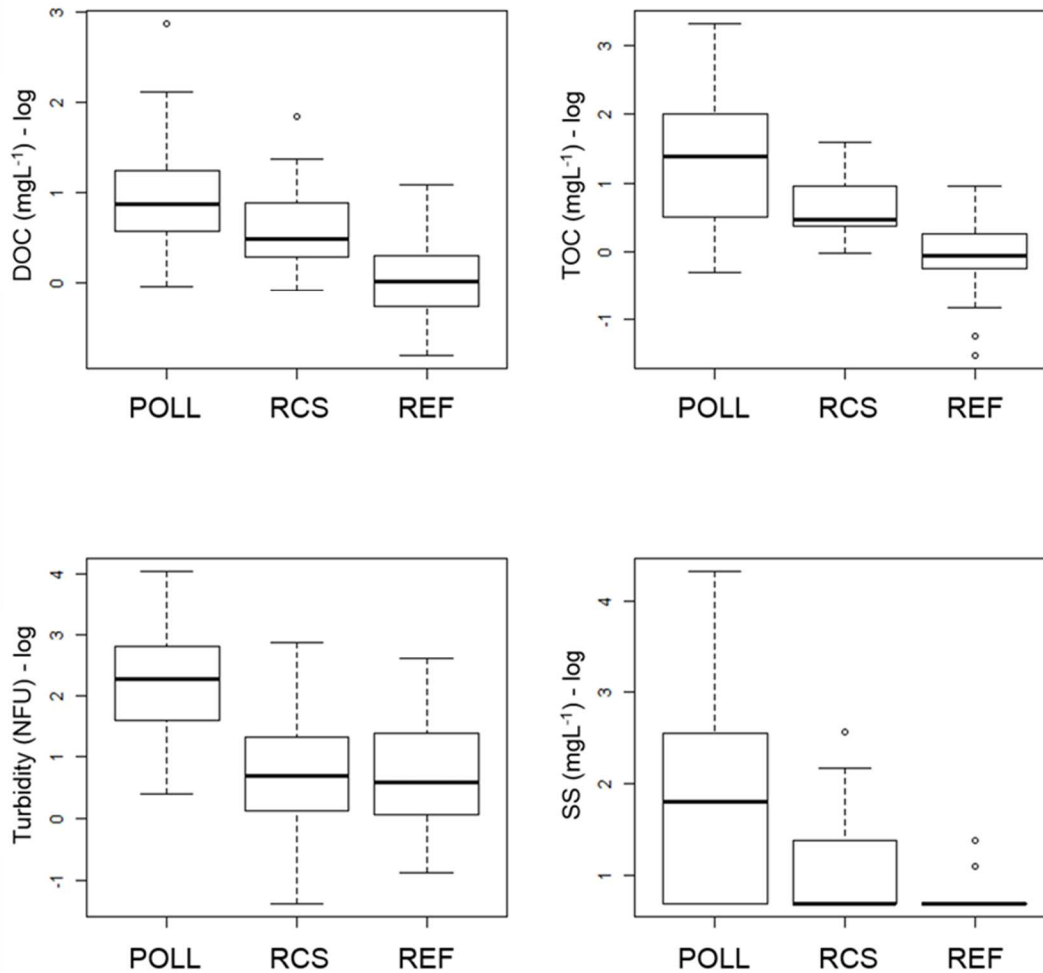


Figure 33 – Main environmental pressure gradients in rivers of Mayotte: dissolved organic carbon (DOC), total organic carbon (TOC), turbidity and suspended solids (SS). Boxplots present the log transformed values for each parameter at each of the three monitoring networks (REF, RCS, POLL).

3.2. Diatoms sampling

Diatoms were sampled following the French (AFNOR 2016) and European (AFNOR 2014a) standards and were carried out once a year during the dry season (July-August). Briefly, benthic diatoms were collected from at least 5 stones from the lotic parts of the sampling sites in order to limit local effect on diatom community (*e.g.* flow velocity, water depth) and mix into a unique vial. The upper surface of the stones was scrubbed with a clean toothbrush. The samples were preserved by adding 99% ethanol for a final ethanol concentration > 70%, in order to preserve DNA. For each site, 2 subsamples were taken from the vial with the pooled biofilm sample for the molecular and morphological approaches.

3.3. DNA Metabarcoding

3.3.1. DNA extraction

DNA extraction was performed using 2mL of the preserved sample. After centrifugation at 13,000 rpm during 30 min, supernatant containing ethanol was removed and the pellet used as starter for DNA extraction. Total genomic DNA was isolated using a non-commercial method based on Sigma-Aldrich GenElute™-LPA DNA precipitation, as described in previous studies (Kermarrec *et al.* 2013b; Chonova *et al.* 2016). This method combined various lysis mechanisms in order to disrupt diatom cell (mechanical, enzymatic, heat) and was recommended for diatom metabarcoding (Vasselon *et al.* 2017a).

3.3.2. PCR amplification

PCR amplification was performed on *rbcl* plastid gene targeting a 312bp barcode. For amplifying this region, the primer pair *Diat_rbcl_708F* (Stoof-Leichsenring *et al.* 2012) and R3 (Bruder & Medlin 2007) was slightly modified. Using an alignment of 1602 *rbcl* reference sequences from 638 diatom species, the degeneracy of the primers was increased in order to amplify a broader diversity of diatoms as follow: the forward primer combined an equimolar mix of *Diat_rbcl_708F_1* (AGGTGAAGTAAAAGGTTTCWTTACTTAAA), *Diat_rbcl_708F_2* (AGGTGAAGTTAAAGGTTTCWTAYTTAAA) and *Diat_rbcl_708F_3* (AGGTGAAACTAAAGGTTTCWTTACTTAAA); the reverse primer combined an equimolar mix of R3_1 (CCTTCTAATTTACWACWACTG) and R3_2 (CCTTCTAATTTACWACAACAG).

For each DNA sample, PCR amplification was performed in triplicate in a final volume of 25 µL. Each PCR mix was composed by 1 µL of extracted DNA, 0.75 U of Takara LA Taq® polymerase, 2.5 µL of 10X Buffer, 1.25 µL of 10 µM of primers *Diat_rbcl_708F_1_2_3* and R3_1_2, 1.25 µL of 10 g/L BSA, 2 µL of 2.5 mM dNTP, and completed with molecular biology grade water. The PCR reaction conditions were initiated by a denaturation step at 95 °C for 15 min followed by a total of 30 cycles of 95°C for 45s (denaturation), 55°C for 45s (annealing), and 72°C for 45s (final extension).

3.3.3. Sample libraries and HTS

PCR products of the 3 PCR replicates prepared for each DNA sample were pooled and cleaned with Agencourt AMPure beads (Beckman Coulter, Brea, USA). Quality and quantity of

purified amplicon were checked using the 2200 TapeStation (Agilent technologies, Santa Clara, USA). Ligation of tags to amplicons and library preparation were performed as described in Vasselon *et al.* (2017) using the NEBNext® Fast DNA Library Prep set for Ion Torrent™ (BioLabs, Ipswich, USA) and A-X tag adapter provided in Ion Express™ Barcode adapters (Life Technologies, Carlsbad, USA). Finally, 42 samples libraries (2014 campaign) and 38 samples libraries (2015 campaign) were pooled in 2 mix at a final concentration of 100pm per mix and sequenced independently. Each mix was sequenced on a Ion 318™Chip Kit V2 (Life Technologies, Carlsbad, USA) on a PGM Ion Torrent machine by the “Plateforme Génome Transcriptome” (PGTB, Bordeaux, France).

3.3.4. Bioinformatic processing

The sequencing platform performed demultiplexing and provided a fastq file for each of the 80 libraries. A first quality filtering step excluded DNA reads below 250 bp read length, with a Phred quality score below 23 over a moving window of 25 bp, with more than one mismatch in the primer sequence and homopolymer over 8 bp, or with ambiguous base. All the fastq files were then treated together following the bioinformatics process described in Vasselon *et al.* (2017b) using the Mothur software (Schloss *et al.* 2009). DNA reads were clustered in OTUs using a distance similarity threshold of 95 % (Mangot *et al.* 2013), and singletons were then removed. All samples were normalized to the same read number (using the smallest read abundance obtained for 1 sample) in order to allow inter-sample comparison. Diatom molecular inventories were obtained using the R-Syst::diatom library (Rimet *et al.* 2016, 13-02-2015: R-Syst::diatom v3, <http://www.rsyst.inra.fr/en>) for taxonomic assignment of OTUs, and the consensus taxonomy of DNA reads with a consensus confidence threshold over 80%. Fastq files with demultiplexed DNA reads, the final OTU (95%) list (including DNA reads proportion, DNA representative sequence of each OTU and the OTU taxonomic assignment) as well as the sampling site description are available for all the samples on the Zenodo repository website (<http://doi.org/10.5281/zenodo.400160>).

3.4. Morphological analysis

Parallel to their molecular analysis, the diatom benthic samples were treated for microscopy analysis, according to the European standard (AFNOR 2014a), using hot H₂O₂ and Naphrax to mount permanent slides. A minimum of 400 valves were counted and determined to

species level (or genus level when not possible) according to (AFNOR 2014b) using classical European floras (*e.g.* Krammer & Lange-Bertalot 1986, 1988, 1991a; b, Krammer 2000, 2001, 2002, 2003...) and literature dedicated to tropical areas (*e.g.* Bourrelly & Manguin 1952; Metzeltin & Lange-Bertalot 1998, 2007; Tudesque *et al.* 2008).

3.5. Morphological and molecular SPI

The ecological quality status of the different river sites was assessed based on the diatoms biological quality element, using the SPI (Cemagref 1982). Morphological and molecular SPI were determined using species taxonomic lists (or genus level if the species level was not reached) obtained by microscopy or HTS (relative abundance of DNA reads), respectively. The OMNIDIA 5 software (Lecointe *et al.* 1993, library 5.3 2015) was used for SPI calculation. As no water quality classes have not yet been defined for Mayotte rivers following WFD recommendation (Tapolczai *et al.* 2017), the general pollution gradient of freshwater rivers was divided into 5 ecological status corresponding to different water quality classes used by the French standard (AFNOR 2007): high (SPI: 17 to 20), good (SPI: 13 to 17), moderate (SPI: 9 to 13), poor (SPI: 5 to 9), bad (SPI: 1 to 5).

Effect of sequencing depth on the molecular SPI values was checked on all the samples by reducing *in silico* the number of reads used to obtain diatom taxonomic list, to a minimum of 50 reads per sample, using random subsampling. SPI values calculated with subsampling data were correlated to optimal SPI values obtained using all available DNA reads.

3.6. Statistical analysis

Relative abundance of each diatom species within each site was determined for molecular and morphological inventories.

The correlation coefficients and their associated *p*-value for the richness (Chao estimator), diversity (Shannon) and SPI indices comparisons were determined with the “Pearson correlation” available on R (R Development core team 2013). Correlations were visualized using linear regression. For each index (SPI, Chao, Shannon), the effect of DNA reads subsampling on index values calculation was checked by comparing together the index values obtained for all samples and all subsampling using one-way Anova analysis.

The molecular OTUs and morphological species lists were used to compute 2 separate Bray-Curtis distance matrices, which were compared together using Mantel test. Non-metric multidimensional scaling (NMDS) was used to visualize the Bray-Curtis matrices based on

molecular OTUs data and morphological data. An Anosim analyses was used to compare the similarity between the water quality status assessed by morphological and molecular SPI.

4. Results

4.1. Morphological analysis

A total of 24 families, 58 genera and 204 species of diatoms were identified among all the samples including 16.2 % of tropical species with 5.4 % of species endemic to Mayotte island. The number of diatoms species identified per sample varies from 6 to a maximum of 54 species (mean = 25 species per sample). The most abundant species (> 5 %) identified among all samples were *Cocconeis placentula* var. *euglypta*, *Gomphonema bourbonense*, *Gomphonema parvulum*, *Nitzschia inconspicua*, *Amphora pediculus* and the Mayotte endemic *Nitzschia* sp.1 (**Table A.1**).

4.2. HTS analysis

The PGM sequencing produced a total of 6,076,529 of DNA reads for the 80 libraries that were sequenced. Based on similar quality levels in the 2 sequencing runs, all sequence data could be analyzed together. After the first bioinformatics step of quality filtering, 1,562,321 reads were retained and clustered into 3381 OTUs (95 % similarity threshold) with a mean of 354 OTUs per sample. To allow inter-sample comparison, all samples were rarefied to 5710 reads (lowest read abundance obtained for one sample) for a total of 456,800 reads corresponding to 2754 OTUs with a mean of 233 OTUs per sample (min = 123, max = 432).

After taxonomic assignment of OTUs using R-Syst::diatom, 69.2 % of OTUs at Family level (75.1 % of total reads), 62.2 % at Genus level (72 % of total reads), and 35.7 % at Species level (40.7 % of total reads) were successfully assigned. Unclassified proportion of reads per sample at species level varied from 1.3 % to 97 %. Successful taxonomic assignment of OTUs resulted in a diatom taxonomic list of 23 families, 39 genera and 66 diatom species (**Table A.2**) with a mean of 16 species per sample (min = 6, max = 41). The most abundant species detected by HTS among all samples were *Ulnaria ulna*, *Amphora pediculus*, *Gomphonema parvulum* and *G. bourbonense*.

4.3. Comparison of the diatom communities obtained by molecular or morphological assignment

The morphological and the molecular taxonomic compositions of diatom communities were compared using Venn diagrams (**Figure 34**). The taxonomic composition between both inventories was similar at 80.8% for family level, at 59% for genus level, and at 13% for species level. 82% of the diatom species detected only by microscopy were absent from the molecular reference library R-Syst::diatom.

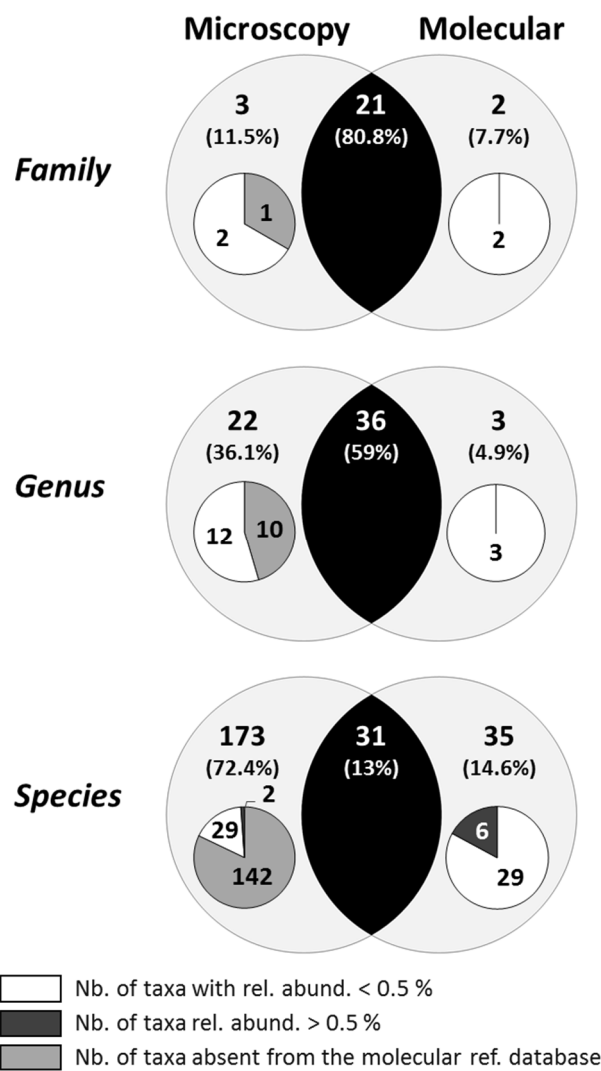


Figure 34 – Venn diagrams comparing the diatom inventories assigned at family, genus and species levels either by the molecular (right circles) or by the morphological (left circles) approach (80 river samples).

Taxa assigned by both approaches are represented by the overlapping region in the middle (black). For taxa detected only by one of the two methods, pie charts indicate the number of taxa with relative abundance < 0.5 % and > 0.5 %. For taxa only detected in microscopy, the number of taxa absent from the reference database is also indicated.

The correlation between Shannon indices calculated with the 2 approaches for all samples were highly significant and stable when comparing the molecular to morphological indices for families ($r = 0.37, p < 0.001$), for genera ($r = 0.33, p < 0.01$), and between molecular OTUs and morphological species ($r = 0.34; p < 0.01$). Only the molecular and morphological Shannon index based on species was not significantly correlated ($p = 0.16$). Regarding the Chao index, all the correlation factors were significant when comparing the family ($r = 0.40, p < 0.001$), genus ($r = 0.41, p < 0.001$), species ($r = 0.25, p = 0.02$) and OTU/species ($r = 0.26, p = 0.02$) levels.

Bray-Curtis dissimilarity indices were calculated based on OTU lists for the molecular approach and on species lists for the morphological one. Mantel's test revealed a significant correlation between the dissimilarity matrices obtained from HTS and morphological inventories ($r = 0.43, p = 0.01$).

4.4. Morphological and molecular SPI calculation

SPI index values calculated based on the morphological (205 taxa) and the molecular diatom (84 taxa) inventories ranged from 5.1 to 19.5 and from 4.4 to 15.8 for the molecular and the morphological approaches respectively. The mean SPI values obtained for the 3 monitoring networks were congruent with their expected water quality status with both the molecular (POLL = 11.7, RCS = 14.9, REF = 15.9) and the morphological (POLL = 8.1, RCS = 11.6, REF = 12.4) approaches (**Figure 35**). The POLL network was significantly different than the RCS and REF ones with both approaches ($p < 0.001$ in both cases). A significant correlation ($r = 0.72$) was observed between the SPI values for all samples obtained by both approaches (**Figure 36**).

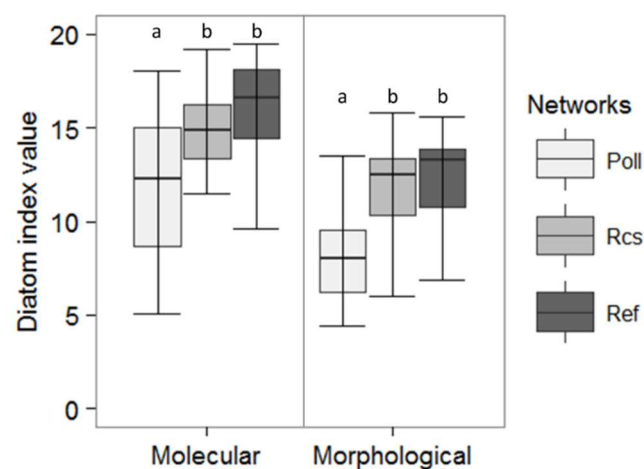


Figure 35 – Distribution of the values of the diatom Specific Pollution Index (SPI) based on molecular (left) and morphological (right) inventories for all 80 samples within the 3 monitoring networks (POLL, RCS, REF).

Different letters indicate significant difference between SPI means (T-test, $p < 0.05$).

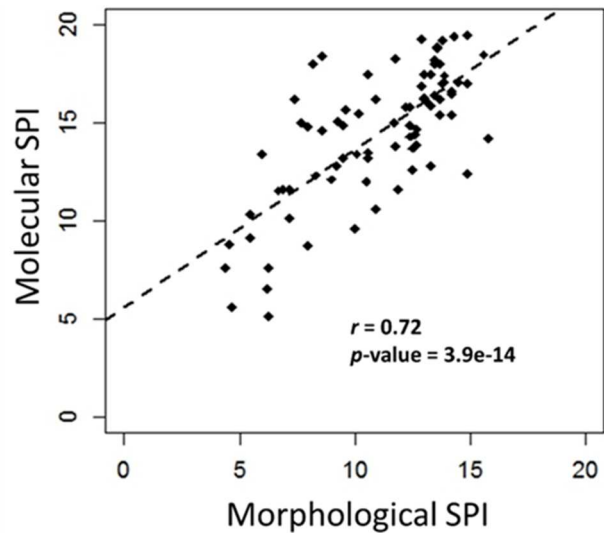


Figure 36 – Correlation between the diatom Specific Pollution Index (SPI) based on molecular (y axis) and morphological (x axis) inventories for all 80 samples. The linear regression model is represented by the dotted line, r and p -value are indicated. SPI values are in the range from 1 (bad quality status) to 20 (high quality status).

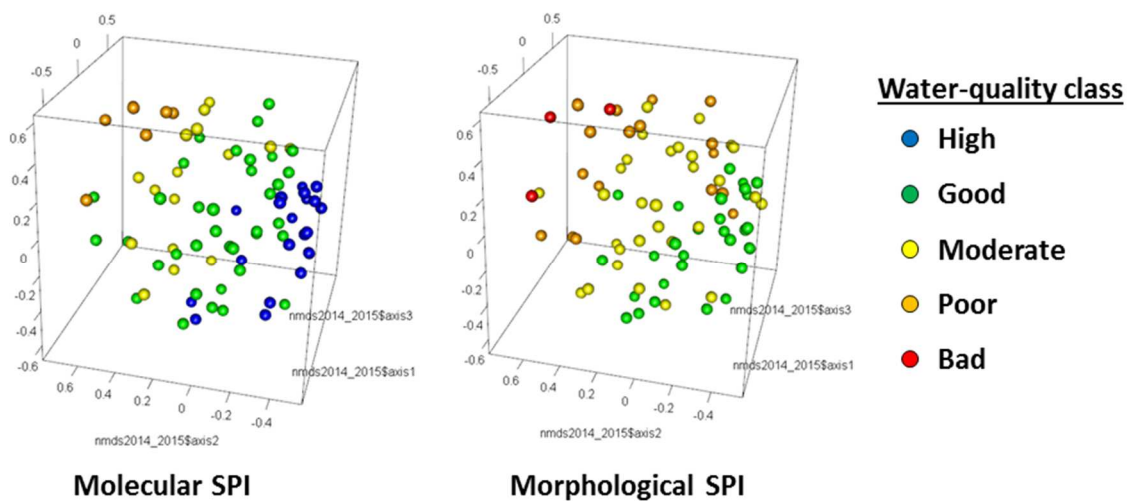


Figure 37 – Three-dimensional NMDS plots of Bray-Curtis dissimilarity based on OTU composition of all the 80 samples. Colours correspond to the quality class deduced from either molecular (left) or morphological (right) SPI values assessed to each sample. Quality classes: high (SPI: 17 to 20), good (SPI: 13 to 17), moderate (SPI: 9 to 13), poor (SPI: 5 to 9), bad (SPI: 1 to 5).

When comparing the water quality classes deduced from the SPI values of both approaches, 27.5% of all the samples shared an identical quality status, 60% had 1 class of difference, 10% had 2 classes of difference, and 2.5% had 3 classes of difference. The NMDS plots presented in **Figure 37** (based on OTUs similarity) showed that the samples distribution is driven

by their respective water quality level both for molecular and morphological approaches. Anosim analysis indicated that the water quality classes explained 22.7% ($p = 0.001$) and 29.0% ($p = 0.001$) of the total variance for the morphological and the molecular approaches respectively.

On average, the molecular SPI values were 3.6 points higher than the morphological SPI values (min difference = 0.1, max difference = 9.8) with a mean of 3.7 (sd = 2.1), 3.7 (sd = 2.4) and 3.4 (sd = 2.1) for the sites belonging to REF, POLL and RCS networks respectively. We observed positive correlations between the difference of SPI obtained by both approaches with the proportion of taxa from *Eunotia* genus and with the proportion of DNA reads that could not be classified at genus level in the molecular inventories (**Figure 38**).

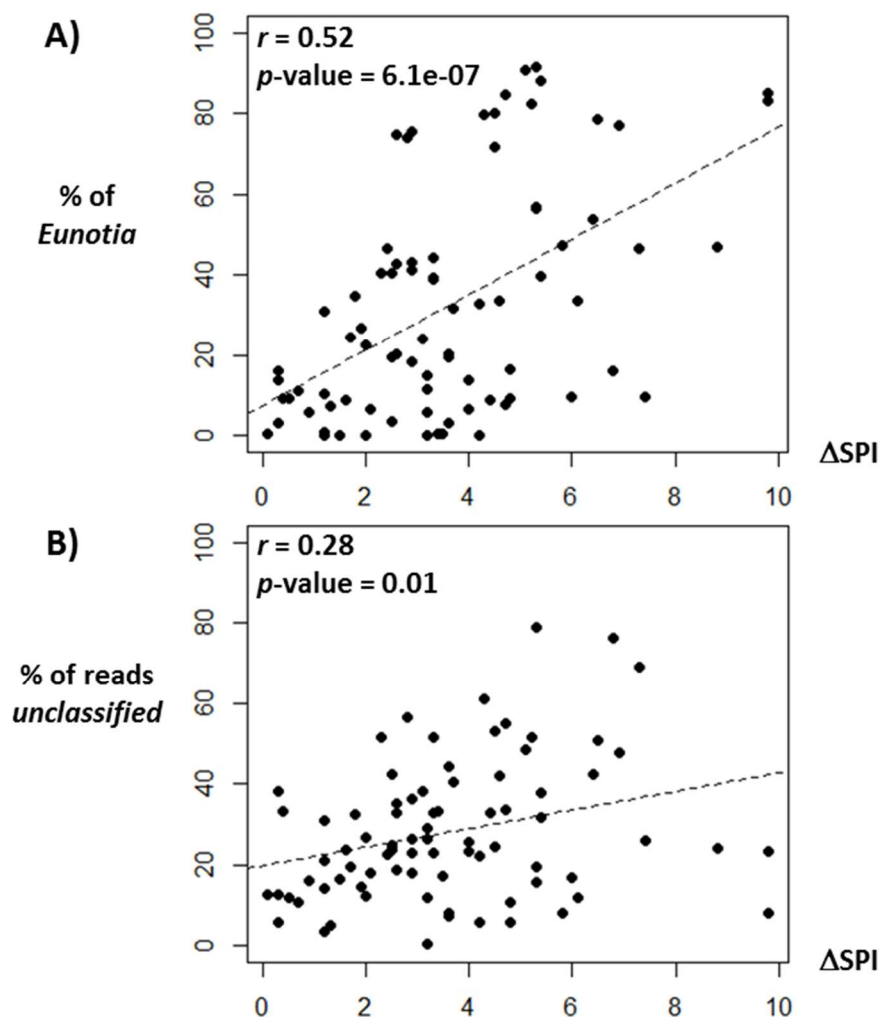


Figure 38 – Correlation between the Δ SPI (difference between the molecular SPI and the morphological SPI values) and the proportion in molecular inventories of (A) *Eunotia* taxa and (B) unclassified reads at Genus level, for all samples.

The linear regression model is represented by the dotted line, r and p -value are indicated.

4.5. Impact of sequencing depth on richness, diversity and the water quality index

Regarding the richness index (Chao), values were affected by the different subsampling with a drop of the correlation from 0.787 to 0.322 (“5710 reads vs all” and “50 reads vs all” respectively) (**Figure 39**). When all reads were used, an average of 555 OTUs per sample was estimated by the Chao index (min = 186; max = 937) while for 50 reads, the average of Chao index was only 45 OTUs per sample (min = 10; max = 111). The Anova analysis showed a significant effect of the subsampling on Chao index ($p < 0.001$) and this directly at the first subsampling of 5710 reads per sample ($p < 0.001$).

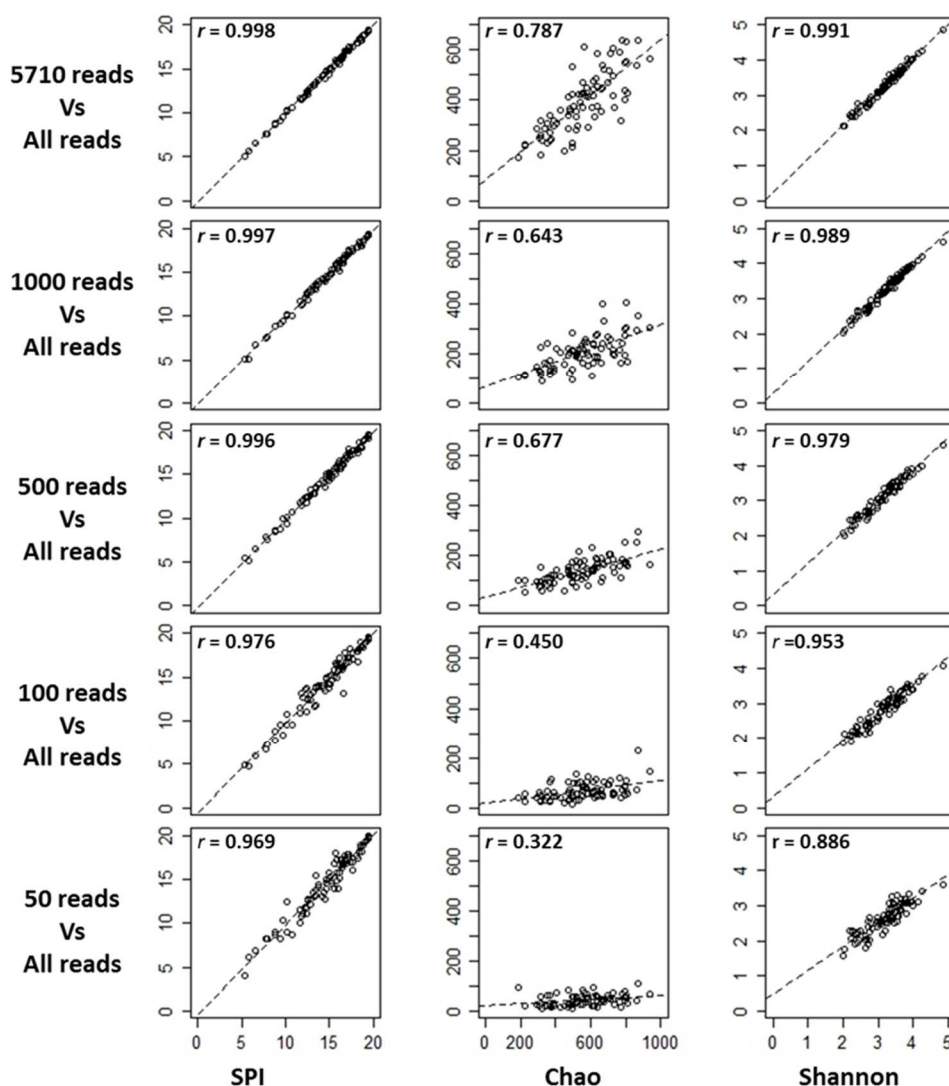


Figure 39 – Impact of the sequencing depth on molecular SPI values (left), OTUs richness (Chao estimator, middle), OTUs diversity (Shannon index, right) evaluated on all samples by performing random subsampling of DNA reads (subsampling decreasing from 5710 to 50 reads per sample). Correlations are presented between values obtained with subsamples of DNA reads (vertical axis) and values obtained using all available DNA reads for each sample (horizontal axis). All correlation values are significant.

The richness diversity index (Shannon) was affected in a similar manner by the different subsampling with a drop of the correlation from 0.99 to 0.89 (**Figure 39**). The Anova analysis did reveal a significant effect of the subsampling on the Shannon index ($p < 0.001$) when the subsampling was done with 100 reads and 50 reads.

The SPI values calculated with the molecular inventories based on all the DNA reads available per sample and with the different subsamples of reads per sample were all significantly correlated (**Figure 39**). An Anova analysis indicated that no significant effect of subsampling on SPI values was detected ($p = 0.99$), whatever the size of the tested subsamples. The average of the difference between the SPI values calculated with 50 reads per sample and those calculated with all reads was 0.66.

5. Discussion

5.1. Community structure and ecological quality status inferred by molecular and by morphological approaches are congruent

Mayotte is a small Island where rivers have the same typology due to their common geological substratum and their short length from the source to the sea pool and (≈ 10 km). Thus, diatom species diversity observed in morphological inventories (204 species for 58 genera) is comparable to other small tropical Islands as determined by Gassioles (doctoral dissertation, Gassiole 2014) for Reunion Island (343 species for 61 genera) and by (Gueguen *et al.* 2015) for Martinique (324 species for 59 genera) and Guadeloupe (352 species for 57 genera) islands. However, diatom community structures obtained with the DNA metabarcoding approach (based on OTU) and with the morphological approach (based on species) are correlated. This confirms previous observations that DNA metabarcoding combined to HTS is a good approach to evaluate diatom diversity in freshwater ecosystems (Zimmermann *et al.* 2015). Mayotte is a relatively complex field study for freshwater quality assessment, mainly due to the presence of endemic and tropical diatom species. Indeed, ecological preferences of these species have not been defined yet, thus making the applicability of many diatom indices (e.g. European diatom indices) uncertain in this part of the world. Previous studies showed that diatom indices developed in a particular geographical area are less effective when applied elsewhere (Rott *et al.* 2003; Potapova & Charles 2007). This is why a diatom index dedicated to Mayotte rivers quality assessment is currently developed in the framework of the WFD (Tapolczai *et al.* 2017). Despite that, Bellinger *et al.*

(2006) and (Bere 2016) showed that European indices can be applied to remote countries as an initial approach when undescribed diatom taxa are not the most abundant taxa. In our study, 50% of tropical and endemic species identified have unknown ecological preferences but correspond to low abundant taxa (< 1%) and therefore have a low impact on SPI index calculation. Furthermore, the dominant taxa were cosmopolitan diatoms (e.g. *Cocconeis placentula*, *Gomphonema bourbonense*, *G. parvulum*, *Nitzschia inconspicua*, *Amphora pediculus*) for which ecological preferences are well described in the literature and included in the SPI index calculation. Moreover, the SPI includes more than 2000 species, some being tropical. The results of the water quality status inferred by the SPI based on both morphological and molecular taxonomic inventories were congruent with the expected ecological status based on observed pressures (physical-chemical parameters) at the monitored sites. The lowest SPI values were obtained for river sites belonging to the polluted network (POLL) and the highest for river sites belonging to the reference network (REF). Even if the morphological approach allow a better discrimination between polluted and reference sites than the molecular one, the latter remains highly efficient to discriminate the 2 networks. Therefore, even if optimizations are required for an accurate and well-adapted water quality assessment, the SPI index is a good basis to compare the molecular (DNA metabarcoding) and the classical (morphological) approaches as monitoring tools for diatoms. Moreover, the river monitoring network developed in Mayotte offered the opportunity of an unprecedented large scale comparison of these two approaches. Our results showed a strong correlation between the molecular and the morphological based SPI indices, supporting previous smaller scales observations made by Kermarrec *et al.* (2014, 4 sites) in France and Visco *et al.* (2015, 27 sites) in a regions of Switzerland .

However, when comparing the taxonomic inventories at the species level between both approaches, important discrepancies appeared. This low correspondence was shown to be mainly due to the incompleteness of the DNA-barcode reference library. Despite those shortcomings, most of the abundant taxa were identified by both approaches at species level or at least at genus level. The SPI calculation is based on the Zelinka & Marvan (1961) formula where rare species have low impact on the final value, while abundant species drive it. This explains the good correlation between molecular and morphological based SPI. Indeed, Bigler *et al.* (2009) showed that removing taxa with relative abundance below 5% have low impact on diatom indices value, as well as (Lavoie *et al.* 2009) who showed that rare diatom taxa have little interest for ecological assessment. The presence of rare taxa can be considered as poorly informative for water quality assessment and for that reason they are often excluded from some indices calculation (e.g. Biological Diatom Index - Prygiel *et al.* 2002). Though, these results show that diatom DNA

metabarcoding could be a reliable tool to derive accurate diatom quality indices for regulatory use at the scale of a biomonitoring network. Another solution is to use a taxonomy-free approach to calculate molecular water quality index directly from metabarcoding data without any taxonomic assignment (Apothéloz-Perret-Gentil *et al.* 2017). Even if such approach prevent doing the link with ecological and historical knowledge based on morphological data, it would be suitable for Mayotte biomonitoring network where such knowledge is limited.

5.2. Biases explaining differences between molecular and morphological based diatom indices

DNA-based indices are currently mimicking conventional indices using two kinds of information about diatom communities : i) a qualitative information which is a list of species usually based on taxonomic assignment of OTUs, and ii) a quantitative information based on the proportion of DNA reads per taxa. Both kinds of information can be affected during the DNA metabarcoding workflow by biases linked to different steps : the DNA extraction method (Deiner *et al.* 2015; Vasselon *et al.* 2017a), the targeted gene for DNA-barcode (Kermarrec *et al.* 2013b; Valentini *et al.* 2016), the set of primers for DNA metabarcoding (Elbrecht & Leese 2015), the PCR amplification protocol (Kebschull & Zador 2015), the sequencing technology (Quail *et al.* 2012), the bioinformatics data processing (Schmidt *et al.* 2015), and the variation of cell biomass among taxa (Thomas *et al.* 2016). For diatoms DNA metabarcoding, previous studies started to identify the importance of those biases in order to optimize the choice of the DNA barcode (Kermarrec *et al.* 2013b) or the DNA extraction (Vasselon *et al.* 2017a). Recommendations of these 2 studies were taken into account to set up the experimental design of the present study: use of a 312 bp *rbcL* DNA-barcode and use of Genelute method for DNA extraction. Elbrecht & Leese (2017) shown that the use of well-developed primers, using specialized tool like PrimerMiner (Elbrecht & Leese 2016), reduces bias in macroinvertebrates metabarcoding inventories. The increase of degenerated bases in our *rbcL* primers allows to reduce primer bias, as discussed by Elbrecht & Leese (2017), however further investigation will have to be done to validate them for water quality assessment. In this study, the comparison between the morphological and molecular approaches was performed using 1 subsample of each sample per approach, which can be a source of variability. However, in the Vasselon *et al.* (2017b) study, 2 subsamples per environmental sample were sequenced and detection of abundant species and SPI calculation were not affected. Similar results were shown for the morphological approach

(Lavoie *et al.* 2005). Thus, we consider that variability linked to subsampling will have limited impact on quality index calculation and method comparison compared to other biases. We will thereafter focus on four other biases that could specifically affect qualitative and quantitative information obtained from DNA metabarcoding of diatom communities in Mayotte: the incompleteness of the reference library, the limits of current taxonomic knowledge, the variation of cell biovolume among diatom taxa and the sequencing depth.

5.2.1. Bias related to reference library incompleteness

Despite the good correlation between the SPI values obtained by both approaches, molecular based SPI values differed from morphological ones for all samples. The molecular diatom inventories used for quality index calculation were only based on the part of the molecular data to which a taxonomical identity could be assigned. For taxonomical assignment we used R-Syst::diatom, an expert library dedicated to diatoms (Rimet *et al.* 2016) which is up-dated with all available sequences allowing a reliable assignment of DNA reads at the family and genus levels. Hence, diversity indices obtained for these 2 taxonomic levels by both approaches were significantly correlated. However a major part of the sequences still remained unclassified at species level and consequently could not be included in the molecular SPI calculation, which leads to an absence of correlation for diversity indices at the species level. Among missing taxa are those often abundant which are present in most of the samples like *Navicula quasidisjuncta*, *Achnanthis subhudsonis*, *Amphora copulata*, *Planothidium rostratum*, *Navicula escambia* or *Gomphonema designatum* (17.7% of total microscopic counts). Such lacks in the reference library contribute to differences between molecular and morphological SPI values, despite their good correlation. The incompleteness of DNA barcode reference libraries is a recurrent bias for European diatom flora already discussed in the literature (Zimmermann *et al.* 2014; Visco *et al.* 2015; Vasselon *et al.* 2017a). This bias is getting even more acute for remote places like Mayotte island with tropical and endemic flora. Indeed, for diatoms, isolating living cells from fresh field samples, cultivating monoclonal strains and consequently identifying and sequencing them is the best way to add reliable new DNA barcode references in libraries. However, this time consuming approach suffers from a low success rate because it requires the isolation of living cells able to survive in culture, which depends on culture conditions (*e.g.* growth media composition, temperature, light). Although 24 strains from Mayotte samples had been isolated and sequenced in a previous study (Kermarrec *et al.* 2013a) and added to R-Syst::diatom library, many endemic and tropical taxa could not be cultivated and remained absent from the library. The single-cell PCR

method has been proposed to obtain sequences from uncultured diatoms (Hamilton *et al.* 2015; Khan-Bureau *et al.* 2016), however, taxonomic identification performed on living cells can lead to incomplete or inaccurate taxonomic identification. Alternative approaches based on OTU co-abundance networks (Irannia & Chen 2016) were proposed to predict the taxonomy of unknown OTU but predictions were limited to the phylum or class taxonomic levels. Finally, (Rimet *et al.* 2018) propose recently to use environmental sequences from HTS runs, to relate them with morphological observations and to integrate them into libraries after setting several quality criteria.

We also observed that increasing proportion of unclassified DNA-reads increases the difference between SPI values obtained from molecular and morphological approaches. Up to now, tests of water quality assessment based on diatom DNA metabarcoding has always mimicked classical morphological approach by i) using biotic indices initially developed for morphological inventories (Kermarrec *et al.* 2013b, 2014; Zimmermann *et al.* 2015; Visco *et al.* 2015; Lejzerowicz *et al.* 2015; Vasselon *et al.* 2017a) and ii) addressing ecological values of morphological species to OTUs through taxonomic assignment. A way to overcome this limit linked to the taxonomical assignment, as suggested by Pawlowski *et al.* (2016), is to connect directly OTUs with environmental data in order to determine their ecological preferences, thus morphological species in conventional indices could be replaced by OTUs. By this way, molecular indices will take into account all the molecular data, be it assigned or not, be it rare or not. This includes hidden diversity like cryptic diversity or unknown diatom taxa which have hardly been taken into account up to now. The development of such molecular indices can be a good option, especially for remote regions with recent biomonitoring initiatives, like Mayotte, suffering from a lack of taxonomic and environmental knowledge.

5.2.2. Bias related to limits in taxonomic knowledge (example of para/polyphyletic taxa)

The species *Nitzschia inconspicua*, indicator of poor quality rivers, was observed in morphological inventories in most of samples but was not detected by the molecular approach. Thus, this taxon is responsible for part of the divergence between molecular and morphological SPI values, molecular SPI being around 3 points higher than morphological SPI for all samples. The incompleteness of the reference library was not in cause in that case as R-Syst::diatom (v3) contains 9 sequences of *N. inconspicua*, among which 3 were obtained from strains isolated from

Mayotte. However, *N. inconspicua* is a paraphyletic species (Rovira *et al.* 2015), making precise taxonomic assignment very difficult at the genus/species level as discussed previously by (Vasselon *et al.* 2017a). Its absence from molecular inventories tends to produce higher SPI values with metabarcoding than with the morphological approach, the later including this low-quality taxon for SPI calculation while the former does not. As the DNA reference library construction is based on traditional taxonomy, the efficiency of the OTU taxonomic assignment relies on the reliability and extent of taxonomic knowledge's. However, nomenclature of diatoms is far from static and evolves over time (Cox 2009; Jahn & Kusber 2009; Kociolek & Williams 2015), which creates problems of taxonomic harmonization in the reference libraries and consequently biases in taxonomic assignation of OTUs. To improve the consistency and accuracy of reference libraries for metabarcoding purposes, it is crucial to up-date them with evolving taxonomical knowledge as well as with new DNA barcode references, which is done regularly for the open-access R-Syst::diatom library (Rimet *et al.* 2016).

5.2.3. Bias related to diatom biovolume

Read proportions were observed to differ from cell proportions in some cases, which may impact derived molecular and morphological SPI values respectively. This was the case for the *Eunotia* genus for which variations between molecular and morphological SPI values were high and positively correlated to the proportion of DNA reads assigned to this particular genus. The read proportion of a taxon can be affected by different technological (*e.g.* DNA extraction, PCR amplification, DNA sequencing, bioinformatics filtering) and biological (*e.g.* cell biomass) factors. The sum of those factors could result in a significant variation between DNA read and diatom cell proportions. In the case of Mayotte samples, the *Eunotia* genus was overrepresented in the molecular inventories compared to the morphological ones and was often a dominant genus. The different species of *Eunotia* observed in Mayotte, using microscope, are characterised by a large cell biovolume (around 19,000 μm^3). Previous studies showed the existence of a clear correlation between the SSU rDNA gene copy number and the diatom cell size and biovolume (Zhu *et al.* 2005; Godhe *et al.* 2008). If such a relationship exists between *rbcL* copy number and diatom cell size, it can explain why the *Eunotia* genus is overestimated in our molecular inventories and why its presence affects the SPI calculation. Should this hypothesis be confirmed, a general correction factor based on diatoms species biovolume could be envisaged and would help to improve the comparability between molecular and morphological indices. However, Tapolczai *et al.* (2017) have shown that the biovolume itself may be important to consider to assess quality status. They

proposed to use the biovolume to weight diatom counts as it is routinely done for quality assessment based on other ecological quality elements (*e.g.* phytoplankton). The theory behind is that the biomass partition of species shows better how resources are capitalized by the species (Kalff & Knoechel 1978; Reynolds 1980; Padisák *et al.* 2006). The bias we observed here due to biovolume may in the end prove to be interesting to improve bioassessment accuracy in future indices that may be developed directly from sequence data.

Angly *et al.* (2014) already proposed to use a correction factor based on 16S rRNA genome number variation to correct the molecular data and provide more reliable microbial community profiles. Further investigation is required to propose more reliable molecular inventories and adapted quality indices.

5.2.4. Bias related to sequencing depth

DNA metabarcoding combined to HTS allows multiplexing and sequencing hundreds of environmental samples at once, given sufficient unique tags are available for sample multiplexing. Increasing the number of samples in one sequencing run decreases the number of available DNA reads per sample. One important question for further accurate application of this approach to field biomonitoring is to know how many DNA reads per sample are required to have a good description of the community diversity and a good evaluation of the ecological status. Previous study showed that increasing the sequencing depth can improve the ecological inference from HTS data more than increasing the number of PCR replicates (Smith & Peay 2014). Of course, depending on the local community and its diversity, the required minimum number of DNA reads will vary from one sample to another. Lundin *et al.* (2012) shown that a minimum of 1000 and 5000 denoised DNA reads were sufficient to respectively describe trends in α and β diversity of bacterial communities from sediment and water samples.

In the present study we focused on benthic diatom diversity, which is known to be lower than diversity of bacterial community. Thus, the α diversity (Shannon index) was not significantly affected when using only a restricted amount of reads (500 DNA reads per sample). In the same manner, reducing drastically the number of DNA reads to 500 does not change SPI values and the final water quality assessment. Although a high sequencing depth is required to detect rare species (Valentini *et al.* 2016), this is not required for water quality assessment purposes where rare species have low impact on index values. This indicates that it may be possible to increase the number of samples in one HTS run, reducing the cost of the DNA metabarcoding approach which tends to make it economically suitable compared to the classical approach. Complementary to our

in silico subsampling, analysis in real laboratory condition are needed to confirm the minimum sequencing depth required for water quality assessment in order to propose a standard like it was done for the WFD morphological approach (with a minimum of 400 morphological counts, AFNOR 2014b).

6. Conclusion and perspectives

The use DNA metabarcoding and HTS for diatoms appears to be a promising approach for freshwater quality assessment. Our study confirmed at a larger scale previous observations that it is possible to infer ecological quality status of rivers based on molecular inventories (*e.g.* Kermarrec *et al.* 2014; Visco *et al.* 2015). However, this approach still requires optimisations and to define standards for key steps of the metabarcoding workflow in order to produce reliable and inter-comparable data between laboratories and platforms. Both HTS quantitative and qualitative information can be improved by working on the major biases like the use of a correction factor for the variation of *rbcl* copy number (based on diatom cell biovolume for example) or the completion of the DNA barcode reference library. While the creation of correction factor required new experiments and investigations, different solutions can be applied to complete the reference library. Availability of an expert and open-access reference library as R-Syst::diatom (Rimet *et al.* 2016) allows sharing and centralizing knowledge's from different laboratories, increasing the number of available references. The use of HTS data to complete DNA barcoding libraries, recently proposed by (Rimet *et al.* 2018), could be an efficient way to access to DNA barcode of uncultivable diatom taxa and improve database completion, especially for taxa that play an important part in water quality assessment.

In order to consider the progressive implementation of molecular approaches into large scale biomonitoring networks and to meet their requirements (*e.g.* WFD, CWA), it will be essential that scientists, environmental stakeholders and managers work hand in hand to validate new methods and to propose of implementation scenarios. The creation of international scientific networking groups, like the European DNAqua-Net Cost action (Leese *et al.* 2016, www.dnaqua.net), is an efficient way to improve scientific knowledge by addressing collectively current issues and to share knowledge with stakeholders in order to consider together the implementation of these new approaches. A better understanding of the DNA metabarcoding pros and cons, together with more scaling up, is still required prior any standardization and deployment. One major challenge will be to make the link between historical biomonitoring

methods applied over decades and new molecular methods, in order to have continuity in water quality assessment.

7. Acknowledgments

This paper was produced as part of the program for the development of biomonitoring network of Mayotte rivers and was funded by the French National Agency for Water and Aquatic Environments (ONEMA-AFB). This work was supported by the European COST action DNAqua-Net (CA 15219). We thank Philippe Chaumeil (INRA Biogeco) who helped us developing the *rbcl* primers, Sonia Lacroix who participates to the preparation of HTS libraries, Franck Salin and Christophe Boury who performed HTS sequencing (INRA-PGTB sequencing platform). We also thank people in BRGM, DEAL Mayotte, Asconit and ONEMA-AFB for their great support and contribution during the sampling campaigns in Mayotte. Special thanks to Gilles Gassiole who performed sampling and microscopy inventories for the RCS network.

8. Author contributions

V.V., A.B., F.R., K.T. contributed to the study designed. V.V. conducted the laboratory work. F.R. and K.T. performed microscopical observation and microscopy inventories. V.V. analyzed the data and wrote the manuscript. All the authors contributed to the discussions and to manuscript editing.

9. Supplementary data

Table A.1 – Valve proportion of diatom species identified by microscopy over all the sample. Only species with valve proportion > 0.05 % are shown.

Diatom species	% of valve
Achnanthydium_exiguum	1.26
Achnanthydium_minutissimum	0.51
Achnanthydium_subhudsonis	2.7
Adlafia_sp.	0.06
Amphora_copulata	2.57
Amphora_minutissima_var._africana	0.46
Amphora_pediculus	6.02
Amphora_sp.	1.66
Caloneis_aerophila	0.06
Cocconeis_placentula_var._euglypta	7.92
Cocconeis_placentula_var._placentula	0.12
Cocconeis_sp.	0.07
Cyclotella_meneghiniana	0.06
Diadesmis_confervacea_var._confervacea	0.26
Encyonema_neomesianum	0.05
Encyonema_silesiacum	0.1
Eolimna_minima	1.75
Eolimna_ruttneri	0.08
Eolimna_sp.	0.84
Eolimna_subminuscula	0.16
Eunotia_sp.	2.8
Eunotia_sp.1	0.06
Eunotia_sp.2	0.16
Fallacia_meridionalis	0.61
Fragilaria_capucina_var._capucina	0.11
Fragilaria_perminuta	0.06
Gomphonema_affine	0.38
Gomphonema_angustatum	2.1
Gomphonema_bourbonense	7.2
Gomphonema_brasiliense_ssp._pacificum	1.04
Gomphonema_clavatum	0.5
Gomphonema_cleveii	1.06
Gomphonema_designatum	1.79
Gomphonema_gracile	0.11
Gomphonema_minutum_f._minutum	0.12
Gomphonema_parvulum_var._parvulum_f._parvulum	6.82
Gomphonema_sp.	0.17
Gomphosphenia_lingulatiformis	0.06
Gomphosphenia_oahuensis	0.08
Gomphosphenia_sp.	0.8

Halamphora_ghanensis	0.88
Halamphora_montana	0.11
Humidophila_brekkaensis	0.07
Humidophila_contenta	2.04
Humidophila_pantropica	0.11
Hydrosera_triquetra	0.25
Karayevia_suchlandtii	0.07
Luticola_mutica	0.18
Mayamaea_permitis	0.07
Navicula_crassuliexigua	0.08
Navicula_cruimeridionalis	0.16
Navicula_cryptocephala	1.52
Navicula_erifuga	0.15
Navicula_escambia	2.12
Navicula_gregaria	0.11
Navicula_jacobii	0.07
Navicula_leptostriata	0.16
Navicula_notha	0.08
Navicula_quasidisjuncta	4.2
Navicula_reichardtiana_var._reichardtiana	0.05
Navicula_rostellata	0.14
Navicula_ruttneri_var._capitata	0.06
Navicula_simulata	0.17
Navicula_sp.	0.95
Navicula_vilaplanii	0.35
Naviculadicta_absoluta	0.24
Naviculadicta_nanogomphonema	0.09
Nitzschia_amphibia_f._amphibia	0.36
Nitzschia_amphibia_f._frauenfeldii	1.74
Nitzschia_clausii	0.07
Nitzschia_frustulum_var._frustulum	4.8
Nitzschia_inconspicua	6.27
Nitzschia_intermedia	0.07
Nitzschia_linearis_var._linearis	0.35
Nitzschia_palea	1.36
Nitzschia_sp.	0.48
Nitzschia_sp._1	5.75
Nitzschia_tropica	0.38
Nupela_sp.	0.21
Nupela_sp.1	0.07
Pinnularia_gibba	0.31
Pinnularia_subcapitata_var._elongata	0.05
Planothidium_biporumum	0.18
Planothidium_frequentissimum	0.44
Planothidium_robustus	0.43
Planothidium_rostratum	2.29
Platessa_hustedtii	0.1
Pleurosigma_sp.	0.06

Rhopalodia_musculus	0.12
Sellaphora_pupula	0.07
Sellaphora_seminulum	2.43
Seminavis_strigosa	0.67
Stauroneis_sp.	0.37
Terpsinoe_musica	0.25
Thalassiosira_weissflogii	0.13
Tryblionella_debilis	0.9
Ulnaria_biceps	0.06
Ulnaria_ulna	0.34

Table A.2 – Diatom species detected by HTS after taxonomic assignment of OTUs and their read proportion over all the sample. The proportion of reads which remained unclassified at the species level is also indicated.

Diatom species	% of read
Achnanthes_coarctata	0.03
Achnanthidium_helveticum	0.01
Achnanthidium_minutissimum	0.06
Amphora_pediculus	6.31
Caloneis_silicula	< 0.01
Cocconeis_placentula	1.4
Craticula_cuspidata	0.93
Craticula_molestiformis	0.01
Cymbella_excisa	0.02
Cymbella_heterogibbosa	< 0.01
Cymbopleura_naviculiformis	< 0.01
Encyonema_minutum	0.82
Encyonema_muelleri	< 0.01
Encyonema_silesiacum	0.09
Eolimna_subminuscula	< 0.01
Eunotia_bilunaris	< 0.01
Eunotia_minor	0.06
Eunotia_pectinalis	0.1
Fallacia_pygmaea	0.02
Fistulifera_saprophila	0.01
Frustulia_vulgaris	0.02
Gomphonema_acuminatum	0.06
Gomphonema_affine	0.16
Gomphonema_bourbonense	5.22
Gomphonema_cleveii	1.52
Gomphonema_parvulum	5.87
Halamphora_montana	0.01
Hydrosera_sp.	0.42
Lemnicola_hungarica	0.01
Luticola_sparsipunctata	0.01
Mayamaea_permitis	0.05
Navicula_cryptocephala	1.23
Navicula_cryptotenella	0.13
Navicula_lanceolata	0.07
Navicula_radiosa	0.01
Navicula_rostellata	0.05
Navicula_sp.	0.02
Navicula_symmetrica	0.2
Navicula_tripunctata	0.05
Navicula_veneta	0.01
Nitzschia_amphibia	1.24
Nitzschia_filiformis	0.13

Nitzschia_fonticola	< 0.01
Nitzschia_inconspicua	0.03
Nitzschia_lorenziana	< 0.01
Nitzschia_palea	1.05
Nitzschia_sp.	0.03
Nitzschia_tubicola	0.01
Pinnularia_divergens	0.03
Pinnularia_subanglica	0.01
Pinnularia_subgibba	0.58
Placoneis_clementis	< 0.01
Placoneis_elginensis	< 0.01
Rhopalodia_gibba	< 0.01
Rhopalodia_sp.	0.05
Sellaphora_minima	0.38
Sellaphora_pupula	0.11
Sellaphora_seminulum	0.28
Seminavis_robusta	0.13
Staurosira_elliptica	< 0.01
Stephanodiscus_hantzschii	< 0.01
Surirella_sp.	< 0.01
Tabellaria_flocculosa	< 0.01
Terpsinoe_musica	2.86
Tryblionella_sp.	< 0.01
Ulnaria_ulna	8.76
unclassified	59.29

V. Application au réseau de surveillance national

Résultats préliminaire obtenus pour la campagne de surveillance 2016 des cours d'eau de France métropolitaine.

Ce travail visant à tester l'application de l'approche moléculaire à l'échelle du réseau de surveillance des cours d'eau de France métropolitaine a été réalisé dans le cadre d'une action conjointe entre l'INRA et l'ONEMA-AFB. Sa mise en pratique a permis une collaboration des différents acteurs en charge de l'évaluation de la qualité des cours d'eau à l'échelle nationale, à savoir :

- Les **6 Agences de l'Eau** en charge de la gestion des ressources en eaux à l'échelle des bassins hydrographiques.
- Les **16 DREAL** en charge de la gestion de l'eau à l'échelle régionale, du travail de terrain et à l'acquisition des données.
- **6 bureaux d'étude** participant au travail de terrain et à l'acquisition des données.

1. Résumé

Depuis plusieurs années, de nombreuses études ont contribué au développement du metabarcoding des diatomées comme outil d'évaluation de la qualité des cours d'eau. Progressivement, les différents biais de l'approche moléculaire ont été identifiés et la méthode optimisée. Bien que des travaux complémentaires soient nécessaires pour accroître encore la fiabilité des données qualitatives et quantitatives obtenues en metabarcoding, les inventaires taxonomiques de diatomées actuellement produits via l'approche moléculaire permettent une évaluation de qualité des cours d'eau semblable à celle obtenue via l'approche morphologique. Cela a notamment été confirmé à l'échelle du réseau de surveillance des cours d'eau de Mayotte, où les évaluations de qualité obtenues à partir des données moléculaires étaient congruentes avec celles obtenues en morphologie. Cependant, le réseau de surveillance de Mayotte étant assez récent et incluant un nombre limité de sites, il ne permet pas d'avoir un recul suffisant quant à la faisabilité de l'implémentation de l'approche moléculaire comme outil d'évaluation de qualité des cours d'eau à grande échelle.

Dans le cadre de cette étude, nous avons appliqué l'approche moléculaire à l'échelle du réseau de surveillance des cours d'eau de France métropolitaine. Pour cela, 461 sites échantillonnés lors des campagnes de surveillance réalisées en 2016 et 2017 ont été utilisés afin de : (i) compléter les manques dans la base de référence du barcode *rbcl*, (ii) comparer les approches morphologique et moléculaire en terme d'évaluation de qualité (iii) ainsi qu'en termes de coût et de temps d'analyse. Lors de la campagne de prélèvement 2016, nous avons ainsi pu mettre en évidence que (i) la complétion de la base de référence de barcodes, en ciblant spécifiquement des espèces fréquemment abondantes dans les inventaires, permet de fortement

améliorer la congruence des notes de qualité entre les deux approches ; (ii) comme observé précédemment, les notes de qualité moléculaires sont congruentes avec les notes basées sur la morphologie ; (iii) l'approche moléculaire est plus rapide (4 X) et plus économique (X 5) que l'approche morphologique lorsqu'il s'agit de traiter plusieurs centaines d'échantillons.

2. Introduction

Les diatomées sont des algues unicellulaires retrouvées dans tous les écosystèmes aquatiques et représentent un groupe très diversifié avec potentiellement 100 000 espèces existantes (Mann & Vanormelingen 2013). En plus de posséder des préférences écologiques variées, les diatomées sont capables de répondre rapidement aux changements des conditions physico-chimiques de leur environnement, ce qui en fait d'excellents indicateurs de l'état de santé de leur milieu, notamment pour les cours d'eau (Stevenson *et al.* 2010a; Rimet 2012). Actuellement les diatomées font partie des éléments de qualité biologique surveillés dans le cadre de la DCE pour définir l'état écologique des cours d'eau (European Council 2000). L'évaluation de l'état écologique d'un cours d'eau repose sur le calcul d'indices biotiques, généralement dérivés de la formule développée par (Zelinka & Marvan 1961), qui prend en compte à la fois l'abondance relative et les préférences écologiques des taxons retrouvés au sein de la communauté benthique de diatomées. Les inventaires taxonomiques de diatomées utilisés dans le calcul des indices de qualité sont obtenus en identifiant 400 individus au niveau spécifique, identification basée sur la reconnaissance morphologique des frustules au microscope optique. La France est le 3^{ème} pays d'Europe, derrière le Danemark et la Suède, en terme de nombre de masses d'eau identifiées comme des cours d'eau avec plus de 10 000 cours d'eau référencés (European Environment Agency 2012), dont plus de la moitié sont inclus dans les réseaux de contrôle opérationnel (RCO) et de surveillance (RCS) suivis chaque année dans le cadre de la DCE (Petit & Michon 2013, 2015). Face à une telle demande, les limites inhérentes à l'approche morphologique (manque d'experts en taxonomie, temps et coût d'analyse élevés) font qu'il est nécessaire de développer de nouvelles méthodes d'identification mieux adaptées au volume d'analyse demandé par les gestionnaires.

Grâce au développement des techniques de séquençage à haut débit (HTS), de nouvelles méthodes d'identification des espèces basées sur l'ADN ont pu être développées. Le metabarcoding, méthode qui permet l'identification des différentes espèces retrouvées dans un échantillon environnemental (*e.g.* eau, sol, biofilm aquatique) sur la base de courtes séquences d'ADN (ou barcodes), a ainsi rapidement été perçu comme une alternative viable aux méthodes

d'identification classiques basées sur la reconnaissance morphologique des individus (Taberlet *et al.* 2012a). En plus d'être potentiellement plus rapide et moins cher que l'approche morphologique, le metabarcoding permet de traiter en parallèle plusieurs centaines d'échantillons. Ceci explique pourquoi l'application du metabarcoding comme outil d'évaluation de la qualité des cours d'eau s'est fortement développée ces dernières années. Concernant les diatomées, de nombreuses études ont contribué au développement de l'approche moléculaire en permettant : de choisir un barcode ADN adapté aux diatomées (fragment du gène *rbcL*, Kermarrec *et al.* 2013), la construction d'une base de référence complète et robuste (R-syst::diatom, Rimet *et al.* 2016), d'utiliser une méthode d'extraction d'ADN adaptée (Vasselon *et al.* 2017a) ou encore d'obtenir une quantification des espèces plus fiable grâce à l'utilisation de facteurs de correction (Vasselon *et al.* 2018). Bien qu'il soit nécessaire de continuer à développer et à optimiser le metabarcoding des diatomées, les données qualitatives (liste des taxons) et quantitatives (abondance relative des taxons) sont suffisamment fiables pour les utiliser dans le calcul des indices de qualité. En effet, plusieurs études, réalisées sur des échantillons de biofilm aquatique issus de cours d'eau, ont pu montrer que les notes de qualité calculées à partir des inventaires moléculaires étaient congruentes avec celles obtenues à partir des inventaires morphologiques (Kermarrec *et al.* 2014; Visco *et al.* 2015; Apothéloz-Perret-Gentil *et al.* 2017; Vasselon *et al.* 2017b). Cependant, toutes ces études ont testé l'approche moléculaire sur un nombre réduit d'échantillons avec au maximum 80 sites (cours d'eau de Mayotte, Vasselon *et al.* 2017c) et sur des gradients environnementaux également limités. Si l'objectif est de proposer le metabarcoding des diatomées comme une alternative aux méthodes d'identification morphologique utilisées dans le cadre de la DCE, il est nécessaire d'évaluer la faisabilité de l'approche moléculaire à grande échelle dans des conditions environnementales plus variées et dans les conditions d'application de la DCE.

Dans cette étude, nous avons appliqué l'approche moléculaire à 461 sites échantillonnés lors des campagnes de surveillance DCE des cours d'eau de France métropolitaine (campagnes 2016 et 2017). Ce projet, réalisé en collaboration avec les différents acteurs en charge de la surveillance de la qualité des cours d'eau, avait plusieurs objectifs : (i) continuer la complétion de la base de référence en incorporant des séquences d'espèces fréquemment observées dans les sites du réseau de surveillance ; (ii) évaluer la congruence entre les notes de qualité obtenues via les approches morphologique et moléculaire, ainsi que l'évaluation de qualité qui en découle ; (iii) évaluer la faisabilité de l'approche moléculaire en termes de coût et de temps d'analyse ; (iv) préparer le transfert potentiel en évaluant les avantages et limites avec les différents acteurs.

3. Matériels et méthodes

3.1. Sélection des sites

Le choix des sites étudiés a été fait en collaboration avec les hydrobiologistes des DREAL en charge de l'évaluation de la qualité des cours d'eau dans les différentes régions de France métropolitaine. Au total, 461 sites ont été sélectionnés sur l'ensemble du territoire (**Figure 40**) afin de répondre à deux grands objectifs :

(i) compléter la base de référence de barcodes : sur la base des inventaires morphologiques obtenus lors des campagnes d'évaluation de la qualité des cours d'eau précédentes (réalisées entre 1992 et 2014), la liste des taxons les plus identifiés a été établie. Sur les 100 taxons les plus fréquemment identifiés dans les inventaires morphologiques, 47 sont absents dans la base R-syst::diatom. Ceux-ci correspondent notamment à des espèces indicatrices de milieu de bonne qualité retrouvées dans des cours d'eau situés dans les Alpes, les Pyrénées ou encore le Massif Central et la Bretagne. Des sites incluant ces taxons en forte abondance ont donc été sélectionnés pour permettre de récupérer une séquence de référence à partir des données HTS. De plus, des sites incluant des taxons représentatifs de zones géographiques spécifiques comme la Corse et la plaine du Rhin ont aussi été choisis de par leur biodiversité particulière.

(ii) partir d'une grande diversité de situations environnementales pour comparer les évaluations de qualité obtenues par les approches morphologique et moléculaire : pour ce faire, des grands cours d'eau possédant un gradient de qualité d'eau marqué entre l'amont et l'aval ont été choisis et des sites échantillonnés sur tout leur linéaire, notamment sur la Loire, l'Adour, la Vienne, le Doubs, l'Allier, la Seine, la Somme, la Durance, l'Oise, la Dordogne et la Garonne. De plus, la comparaison entre les deux approches a pu être réalisée sur l'ensemble des sites du réseau RCS du bassin versant du Rhône (incluant les départements de l'Ain, Jura, Haute-Savoie, Savoie, Rhône et Loire), qui présente une grande diversité de milieux (*e.g.* zones agricoles, industrielles, urbaines, montagnardes).

3.2. Echantillonnage des biofilms aquatiques

Les 461 sites ont été échantillonnés lors des 2 campagnes de suivi des cours d'eau réalisées en France dans le cadre de la DCE en 2016 (166 sites) et 2017 (295 sites) (**Figure 40**).

L'échantillonnage a été réalisé par les organismes en charge de ces suivis (bureaux d'études, DREAL) en suivant les normes NFT 13946 (AFNOR 2014a) et NFT 90354 (AFNOR 2007). Brièvement, les communautés benthiques de diatomées ont été récupérées en frottant avec une brosse à dent la surface de 5 pierres immergées issues de la zone lotique des cours d'eau. Afin de préserver l'ADN, le biofilm aquatique ainsi obtenu a ensuite été transféré dans un falcon 50 mL dans lequel a été ajouté de l'éthanol jusqu'à atteindre une concentration finale d'au moins 70 % d'éthanol. Après homogénéisation des échantillons, les organismes en charge des prélèvements ont sous-échantillonné chacun des 461 échantillons de biofilm en 2 lots qui serviront aux analyses morphologique et moléculaire.

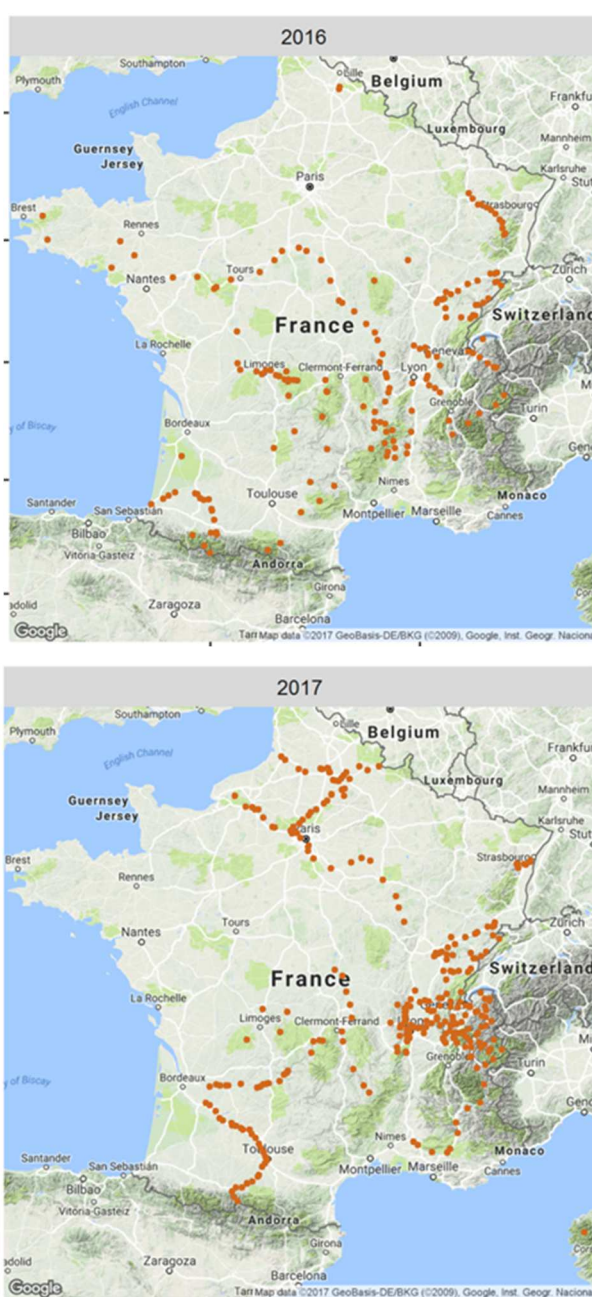


Figure 40 – Localisation des 461 sites utilisés dans cette étude.

3.3. Approche morphologique

La préparation des échantillons, la détermination et le dénombrement des diatomées en microscopie ont été réalisés pour tous les échantillons par les différents organismes préleveurs. Plusieurs traitements chimiques (H_2O_2 , HCl) ont été appliqués à chaque échantillon de biofilm en accord avec la norme européenne (AFNOR 2014a), ceci afin de ne conserver que les frustules des diatomées. Les échantillons ont ensuite été montés entre lame et lamelle dans une résine possédant un fort indice de réfraction (Naphrax[®]) pour permettre une observation au microscope optique. Finalement, les inventaires taxonomiques de diatomées ont été réalisés sur la base de la détermination d'au moins 400 valves au niveau de l'espèce (ou du genre lorsque c'était impossible) en utilisant les guides floristiques Européens (*e.g.* Krammer & Lange-Bertalot 1986, 1988, 1991a; b, Krammer 2000, 2001, 2002, 2003, Hoffman *et al.* 2011...).

3.4. Approche moléculaire

En accord avec les Agences de l'eau et les Dreal, les organismes préleveurs ont envoyé un sous-échantillon des biofilms fixés au laboratoire de l'INRA-CARTEL afin de mettre en place l'approche moléculaire. Les échantillons ont été centrifugés pendant 30 min à 17 000 g et l'extraction de l'ADN total effectuée à partir du culot en utilisant la méthode GenElute décrite précédemment (Vasselon *et al.* 2017a). Une amplification par PCR d'une partie du gène *rbcl* (312 pb), fragment recommandé comme barcode pour les diatomées (Kermarrec *et al.* 2014), a été réalisée en dupliqua sur l'ensemble des ADN extraits. Les « primers » (forward : Diat_*rbcl*_708F_1/708F_2/708F_3 ; reverse : R3_1/R3_2), le mix PCR ainsi que les conditions d'amplification étaient les mêmes que ceux décrits dans Vasselon *et al.* (2017c), à l'exception de l'amplification qui a été réalisée avec 33 cycles. Par la suite, les produits PCR ont été envoyés à la « Plateforme Génome et Transcriptome de Toulouse » (GeT-Plage) où a été réalisé : (i) une purification des produits PCR ; (ii) la préparation des bibliothèques de séquençage en ajoutant aux amplicons des tags spécifiques à chaque échantillon ainsi que les adaptateurs de séquençage ; (iii) la préparation des 2 « pools » finaux qui correspondent au mélange équimolaire des bibliothèques des échantillons des campagnes 2016 et 2017 ; (iv) le séquençage en « paired-end » de chaque pool réalisé avec la technologie Illumina Miseq et le kit V3 (2 x 250 pb).

Les étapes de « démultiplexage » et de « contigage » des séquences ont été réalisées par la plateforme de séquençage GeT-Plage qui a pu fournir 1 fichier fastq pour chaque échantillon séquencé. Des traitements bio-informatiques ont ensuite été réalisés avec le programme Mothur

(Schloss *et al.* 2009) pour éliminer les séquences de mauvaise qualité selon les critères suivant : score Phred < 23 sur une fenêtre de 25 pb, erreur dans la séquence des primers > 1, homopolymères > 8 pb, base ambiguë > 0. Le logiciel Uchime (Edgar *et al.* 2011) a ensuite été utilisé pour supprimer les séquences chimériques. Enfin, les séquences restantes ont été dérépliquées et seules les séquences uniques avec une abondance > 2 ont été conservées pour la suite des analyses.

Une taxonomie a été assignée à chaque séquence grâce à l'utilisation conjointe de la base de référence R-syst::diatom (Rimet *et al.* 2016) et de la méthode de classification bayésienne naïve (Wang *et al.* 2007) avec un seuil de confiance > 85 %. Les séquences assignées au phylum « Bacillariophyta » ont été regroupées en OTUs (95 % de similarité) pour lesquels ont été assignées une séquence ADN représentative et une taxonomie en suivant la méthode décrite par Vasselon *et al.* (2017c). A partir de ces informations, un inventaire taxonomique a été créé et utilisé pour calculer les indices de qualité moléculaire.

3.5. Estimation des coûts et temps d'analyse des deux approches

Deux critères ont été utilisés pour évaluer la faisabilité d'application du metabarcoding à l'échelle du réseau de surveillance par rapport à l'approche morphologique. Tout d'abord, le temps nécessaire pour traiter les échantillons, de la réception des échantillons jusqu'à la production des inventaires taxonomiques, a été estimé pour les deux approches (la méthode d'échantillonnage étant la même dans les deux cas, elle n'a pas été prise en compte). Dans un deuxième temps, les coûts relatifs aux deux approches ont été estimés en prenant en compte le coût des réactifs ainsi que celui de la main d'œuvre.

Pour l'approche moléculaire, les estimations de temps et coût ont été réalisées en considérant le travail d'un seul technicien pour gérer la totalité des étapes et des échantillons avec un même salaire horaire de 32 € / h (charges patronales incluses). Les temps nécessaires à la réalisation des extractions et des PCR ont été calculés en considérant 24 échantillons traités en parallèle (réplicas compris). Les temps nécessaires au séquençage HTS et au traitement bio-informatique ont été considérés comme incompressibles, même si le temps nécessaire au traitement d'1 ou de 288 échantillons (limite maximum pour un séquençage MiSeq Illumina) ne sont pas les mêmes. Les coûts des extractions et des PCR ont été calculés en incluant le coût des réactifs et des consommables (devis des fournisseurs), ainsi que le coût de la main d'œuvre. La préparation des bibliothèques et le séquençage ayant été réalisés par la plateforme de séquençage, le

devis fournit par celle-ci a été utilisé comme base fixe pour définir le coût du séquençage, coût permettant de séquencer en parallèle jusqu'à 288 échantillons.

Pour l'approche morphologique, les organismes préleveurs ayant réalisé l'ensemble des étapes (*e.g.* traitement chimique, préparation des lames, identification morphologiques), les coûts et temps d'analyse se basent sur leur devis. Pour simplifier les calculs, il a été considéré que pour chaque échantillon, la détermination taxonomique des diatomées au niveau spécifique était réalisée en ½ journée de travail plus le temps relatif à la préparation des échantillons et des lames pour la microscopie, pour un coût total de 200 € par échantillon.

3.6. Stratégie de complétion de la base de référence

Une nouvelle méthode a récemment été proposée pour compléter les bases de référence de barcodes en récupérant directement des séquences environnementales issues du metabarcoding (Rimet *et al.* 2018). Pour ce faire, il suffit de sélectionner et de séquencer en HTS des échantillons dans lesquels l'espèce ciblée (*i.e.* non séquencée et absente de la base de référence) est supposée présente en forte abondance. Si l'espèce ciblée est très abondante, il y a de fortes chances que les OTUs les plus abondants obtenus avec l'approche moléculaire correspondent à cette espèce. Après plusieurs étapes de vérification (*e.g.* morphologique, phylogénétique ; voir l'article de Rimet *et al.* (2018) dans l'**Article annexe II** pour plus de détails), il est alors possible d'utiliser la séquence environnementale de l'OTU comme séquence de référence pour l'espèce ciblée, permettant ainsi de compléter la base de référence. Cette stratégie a été utilisée dans cette étude afin de créer la nouvelle version de la base R-Syst::diatom (version de Mai 2017).

3.7. Evaluation de l'état écologique des cours d'eau

Les calculs des indices IBD et IPS ont été réalisés via le logiciel Omnidia (Lenoir & Coste 1996, version 6.0.2s) sur la base des inventaires taxonomiques produits. Concernant les inventaires moléculaire, il a été montré que le nombre de copies du gène *rbcL* est directement corrélé au biovolume des espèces, aboutissant à une surestimation des plus grosses espèces (Vasselon *et al.* 2018). Afin d'éviter ce biais et d'obtenir des abondances relatives d'espèces plus proches de celles obtenues en microscopie, un facteur de correction (CF) a été proposé par les auteurs pour chaque espèce de diatomée. Ces facteurs de correction ont été appliqués dans notre

étude à partir des données de biovolumes spécifiques disponibles dans la base Omnidia, ce afin de produire des inventaires taxonomiques corrigés, qui ont par la suite été utilisés pour calculer les indices IBD et IPS. Au final, 5 notes d'IPS et d'IBD ont été produites pour chaque échantillon : **1** sur la base de l'inventaire morphologique ; **2** sur la base des inventaires moléculaires créés à partir de la base de référence R-syst::diatom (version Mars 2017), corrigés et non corrigés par les CF ; **2** sur la base des inventaires moléculaires créés à partir de la base de référence R-syst::diatom (version enrichie de Mai 2017), corrigés et non corrigés par les CF .

4. Résultats préliminaires pour la campagne 2016

4.1. Mise en place de l'approche moléculaire

Sur les 166 échantillons de biofilm issus de la campagne de prélèvement 2016, 2 ont été réceptionnés tardivement (ils seront traités en 2017) et 22 n'ont pas pu être préparés pour le séquençage Illumina Miseq car :

- l'amplification PCR n'avait pas abouti ou n'était pas suffisante pour permettre la préparation des bibliothèques de séquençage (20 échantillons).
- les échantillons ont été fixés au formol (au lieu de l'éthanol) et n'ont donc pas pu être traités avec l'approche moléculaire (2 échantillons).

Concernant le premier cas, le fait que certains échantillons n'aient pu être amplifiés en PCR est directement lié à la qualité et à la quantité de l'ADN extrait. La méthode d'extraction GenElute n'inclue pas d'étape de purification sur colonne, si bien que de nombreux composés présents dans le biofilm (*e.g.* acides humiques, acides fulviques, ions métalliques) peuvent être co-extraits avec l'ADN et inhiber la PCR (Schradler *et al.* 2012). Comme les quantités d'ADN obtenues avec cette méthode sont généralement importantes, l'ADN est dilué afin d'être utilisé en PCR, ce qui suffit à lever l'inhibition dans la plupart des cas (Vasselon *et al.* 2017a). Cependant, lorsque les sites échantillonnés sont naturellement riches en composés inhibiteurs ou lorsque les communautés de diatomées benthiques sont peu développées et ne permettent pas d'obtenir des quantités d'ADN suffisantes, il devient difficile de réaliser les dilutions. Or, la majorité des échantillons non amplifiés sont soit issus de régions ayant un socle cristallin et dont les eaux sont chargées en acides humiques (Bretagne, Limousin), soit de sites où les communautés de diatomées sont peu développées (*e.g.* sites alpins). Pour ce genre d'échantillon, Vasselon *et al.* (2017a) avaient recommandé l'utilisation d'une colonne de purification d'ADN pour éliminer les

composés inhibiteurs. Cette solution sera appliquée aux 20 échantillons d'ADN non amplifiés afin de pouvoir les inclure lors du séquençage de ceux de la campagne 2017. En fonction des résultats obtenus, il pourrait être intéressant d'inclure systématiquement l'étape de purification sur colonne dans la méthode GenElute.

Dans le second cas (échantillons fixés au formol), la mise en place de l'approche moléculaire nécessite d'avoir des échantillons compatibles avec les techniques de biologie moléculaire (*e.g.* extraction ADN, amplification PCR). Il est donc nécessaire de mettre en place des méthodes et des pratiques d'échantillonnages adaptées, notamment vis-à-vis du stockage et de la préservation des échantillons de biofilm (*e.g.* fixation à l'éthanol). Ces pratiques étant nouvelles pour les acteurs en charge des prélèvements, elles ne sont pas encore un réflexe lors des campagnes d'échantillonnage sur le terrain. Ceci peut expliquer pourquoi certains échantillons ont été fixés au formol, produit habituellement utilisé pour fixer les échantillons mais qui est connu pour dégrader l'ADN (Stein *et al.* 2013). D'où l'importance de bien communiquer et échanger avec les différents acteurs, afin de leur expliquer les contraintes liées à l'approche moléculaire et les bonnes pratiques à mettre en place. Un compte rendu de la campagne 2016 a été fait en ce sens afin de préparer au mieux la campagne d'échantillonnage 2017 (Rimet *et al.* 2017).

4.2. Comparaison des coûts et temps d'analyse des deux approches

Les estimations des temps et coût d'analyses en fonction du nombre d'échantillons traités ont été calculées pour les deux approches (**Figure 41**). Lorsqu'il s'agit de traiter un nombre réduit d'échantillons (jusqu'à 50 environ) les coûts et temps relatifs à la mise en place des deux approches sont similaires. Lorsque le nombre maximum d'échantillons pouvant être séquencés dans 1 même run HTS est atteint, ce qui correspond à 288 échantillons dans les conditions de cette étude, l'approche moléculaire est environ 5 fois moins chère et 4 fois plus rapide que l'approche morphologique pour traiter la totalité des échantillons. Dans une précédente étude, il a été montré que le coût nécessaire à l'identification au niveau spécifique de 500 valves de diatomées était similaire pour les deux approches (Stein *et al.* 2014). Depuis, les technologies HTS se sont développées et les coûts liés au séquençage ont été fortement réduits (Goodwin *et al.* 2016). A titre d'exemple, la société Illumina développe une nouvelle série de séquenceurs HTS capables de séquencer un génome complet pour 100 \$ (Novaseq, <http://www.businesswire.com/news/home/20170109006363/en/>). Il n'est donc pas surprenant que l'approche moléculaire devienne moins chère que l'approche morphologique. Le temps d'analyse pour l'approche moléculaire peut être encore diminué car notre estimation prend en

compte un temps d'optimisation des amplifications PCR. De plus, l'extraction de l'ADN étant l'une des étapes la plus chronophage du metabarcoding, il est envisageable d'automatiser cette étape en utilisant des robots d'extraction, ce qui permettrait de réduire encore le temps de traitement des échantillons.

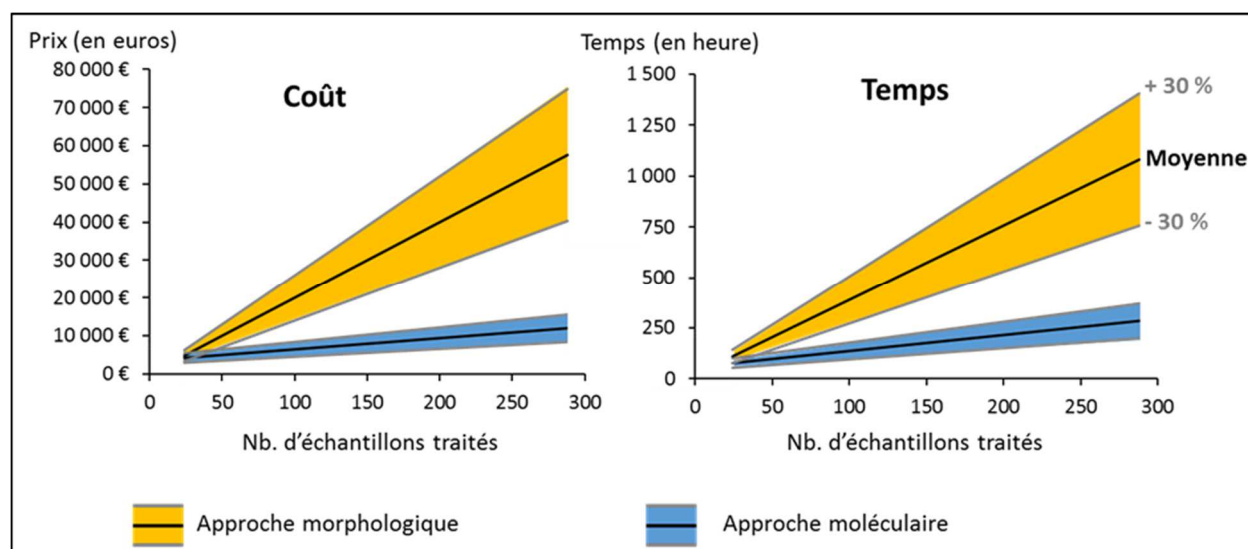


Figure 41 – Estimations des coûts et temps relatifs à la mise en place des deux approches d'identification des diatomées.

4.3. Résultats du séquençage et complétion de la base de référence

Le séquençage (technologie Illumina Miseq) réalisé sur les 142 échantillons de la campagne de prélèvement 2016 a permis d'obtenir 11 539 416 séquences (séquencées dans les deux sens). Suite aux différents filtres bio-informatiques, 3 119 226 séquences ont été conservées et groupées sur la base de leur similarité (seuil à 95 %) en 682 OTUs. L'assignation taxonomique des OTUs a été réalisée à partir de la base de référence R-syst::diatom (version mars 2017). L'assignation au niveau du genre a permis d'identifier 362 OTUs (représentant 77 % des séquences). L'assignation au niveau de l'espèce a permis d'identifier 205 OTUs (représentant 58 % des séquences). L'inventaire taxonomique global de l'ensemble des échantillons incluait 28 familles, 53 genres et 102 espèces de diatomées.

Pour compléter la base de référence, suffisamment d'échantillons présentaient les caractéristiques décrites par (Rimet *et al.* 2018) pour permettre d'identifier précisément la taxonomie associée à 61 OTUs, OTUs jusqu'alors non assignés au niveau du genre ou de l'espèce. Ceux-ci correspondaient à 21 espèces de diatomées fréquemment identifiées dans les rivières métropolitaines, notamment *Achnantheidium delmontii*, *Nitzschia costei* ou encore *Gomphonema*

rhombicum. Les séquences ADN représentatives des 61 OTUs ainsi que leur taxonomie ont été intégrées dans la nouvelle version de la base R-Syst::diatom (version mai 2017). Les inventaires taxonomiques moléculaires produits précédemment via la base R-Syst::diatom (version mars 2017) ont ainsi pu être complétés grâce aux nouvelles informations de la base R-Syst::diatom (version mai 2017).

4.4. Evaluation de la qualité des cours d'eau

Les valeurs d'indice de qualité calculées à partir des différents inventaires moléculaires ont été comparées aux valeurs obtenues à partir des inventaires morphologiques et les résultats résumés dans la **Figure 42** pour l'IBD et dans la **Figure 43** pour l'IPS. Quel que soit l'inventaire moléculaire utilisé (avec les versions de R-Syst::diatom de mars ou de mai 2017), les valeurs d'indices moléculaire et morphologique étaient significativement corrélées, aussi bien pour l'IPS que pour l'IBD. La meilleure corrélation a été obtenue pour l'indice IPS en utilisant l'inventaire moléculaire généré à partir de la base R-Syst::diatom complétée (version mai 2017) avec correction des abondances relatives de séquences par le facteur de correction basé sur le biovolume des espèces. Dans cette configuration, l'écart entre les notes moléculaire et morphologique était de 1,8 points pour l'IBD (min = 0, max = 13,9) et de 1,6 points pour l'IPS (min = 0, max = 9,2).

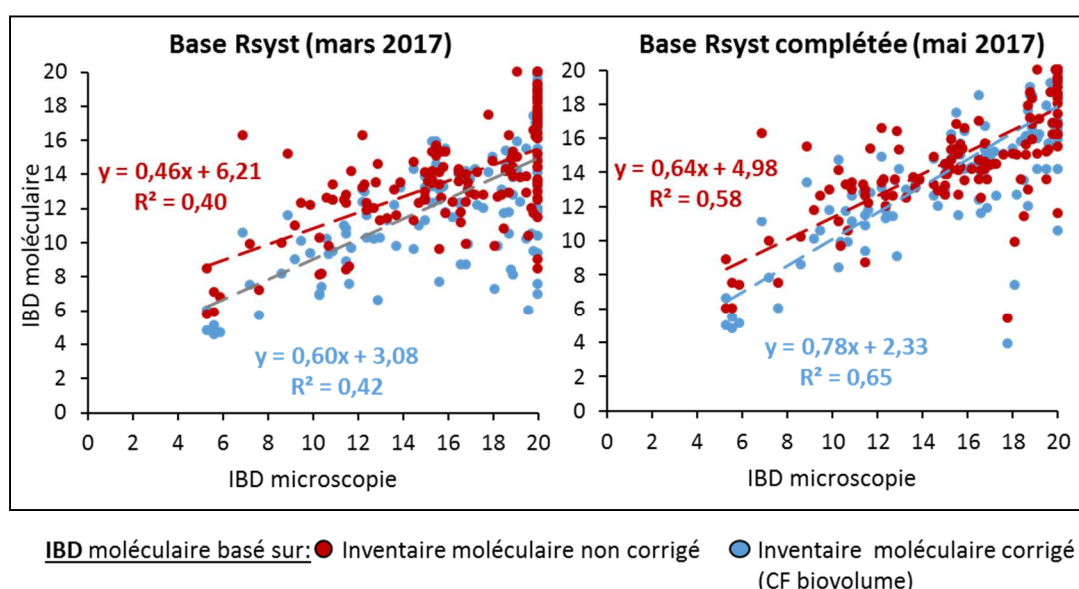


Figure 42 – Corrélation entre valeurs d'indices IBD moléculaire et IBD morphologique.

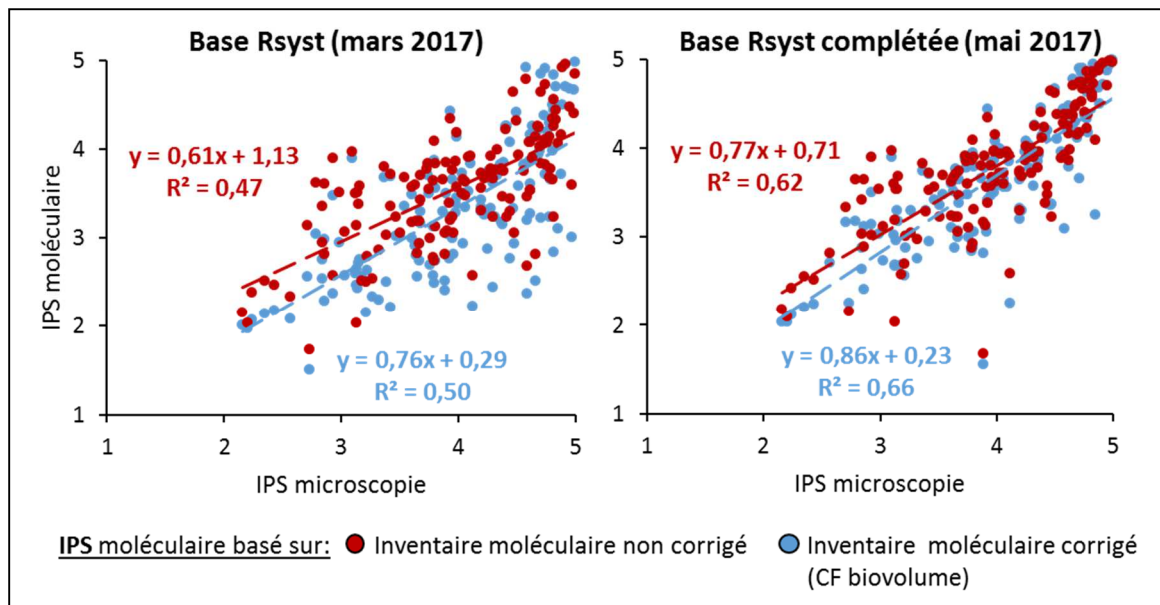


Figure 43 – Corrélation entre valeurs d’indices IPS moléculaire et IPS morphologique.

Ces premiers résultats montrent que les notes d’indice obtenues en metabarcoding sont fortement corrélées à celles obtenues avec l’approche morphologique pour les deux indices. Les corrélations obtenues sont soit similaires (Visco *et al.* 2015; Apothéloz-Perret-Gentil *et al.* 2017), soit inférieures à celles obtenues précédemment (Vasselon *et al.* 2017b, 2018). Malgré cela, ces résultats sont très encourageants car il s’agit d’une des premières études à utiliser le metabarcoding des diatomées comme outil pour évaluer la qualité des cours à grande échelle, sur des sites variés avec un gradient de qualité d’eau marqué. On remarque que l’IPS permet de mieux discriminer certains sites de bonne qualité que l’IBD pour lequel on obtient plus fréquemment des notes de 20 (voir **Figure 42**).

L’intégration des séquences environnementales dans la base de référence a particulièrement amélioré les corrélations entre les notes de qualité moléculaire et morphologique, de 15 % en moyenne, et ce pour les deux indices. On sait que les espèces de diatomée les plus abondantes ont un poids plus important dans le calcul de la note de qualité d’un site (Lavoie *et al.* 2009; Bigler *et al.* 2009). La méthode proposée par Rimet *et al.* (2018), et utilisée dans cette étude pour compléter la base R-Syst::diatom (version de mai 2017), permet de compléter la base de référence en ciblant des espèces retrouvées fréquemment et en forte abondance dans les inventaires morphologiques. Même si un nombre limité d’espèces a pu être intégré dans la base via cette approche (21 espèces ajoutées), celles-ci étant abondantes dans nos inventaires, leur prise en compte a amélioré le calcul des indices de qualité.

En plus de l’amélioration de la base de référence, le fait d’optimiser les abondances relatives des séquences en utilisant les facteurs de correction décrits dans (Vasselon *et al.* 2018)

a permis d'améliorer les corrélations entre indices morphologique et moléculaire en se rapprochant d'une pente égale à 1. Cependant, à cause d'incohérences dans la base de référence de biovolume, certaines espèces n'ont pas pu être corrigées correctement. C'est notamment le cas pour *Melosira varians*, pour laquelle un biovolume de 3267 μm^3 est référencé dans la base Omnidia alors qu'en réalité il est plutôt de l'ordre de 14515 μm^3 . La mise à jour des biovolumes de la base de référence devrait permettre d'aboutir à des inventaires moléculaires plus fiables et ainsi améliorer les notes de qualité obtenues.

5. Remerciements

Cette étude, financée par l'ONEMA-AFB, a été réalisée avec la contribution des membres des **Agences de l'Eau** Artois-Picardie (Christophe Lesniak), Rhône-Méditerranée et Corse (Loïc Imbert, Franck Repellini), Adour-Garonne (Majlis Durand, Jean-Pierre Rebillard, Margaux Saut), Rhin-Meuse (Jean-Luc Matte), Loire-Bretagne (Jacky Durocher), Seine-Normandie (Marie Berdoulay) ; des **DREAL** Aquitaine (Delphine Sagnet), Auvergne (Franck Véry), Bourgogne (Valérie Peeters), Bretagne (Gael Gicquiaud), Centre (Simon Saadat, Chafika Karabaghli), Corse (Isabelle Boulier), Franche-Comté (Eric Parmentier), Limousin (Jean Marc Vouters), Lorraine (David Heudre), Midi Pyrénées (Eléonore Seigneur), Pays de la Loire (Didier Guillard), Rhône-Alpes (Rémy Chavaux), Normandie (Frédéric Petel), Picardie (David Fouré), Nord Pas de Calais (Nathalie Zydek), Ile de France (Odile Courtial) ; des **bureaux d'étude** participant au travail de terrain et à l'acquisition des données: Aquabio (Rémy Marcel, Bruno Fontan), Aquascop (Jessica Vizinet), Asconit (Lénaig Kermarrec, Etienne Ponton), Sage (Anne Rolland, Jean-Philippe Vulliet, Carole Geret), GREBE (Philippe Prompt), Eurofins Expertises Environnementale (Léa Feret).

Des remerciements sont adressés en particulier à Cécile Chardon qui a réalisé la préparation des librairies de séquençage pour tous les échantillons.

VI. Discussion générale et perspectives

Ce travail de thèse avait pour objectif d'évaluer le potentiel du metabarcoding comme méthode alternative à l'approche morphologique pour l'identification des diatomées, dans le but de permettre l'évaluation de l'état écologique des cours d'eau. C'est dans cette optique et dans la continuité des travaux de L. Kermarrec (Kermarrec 2012), qui ont permis d'initier la mise en place de l'approche moléculaire (*e.g.* choix du barcode, création d'une base de référence) et de valider sa capacité à fournir à partir d'échantillons environnementaux des inventaires taxonomiques de diatomées fiables, que nous avons développé deux axes de travail. Dans un premier temps, nous avons poursuivi le développement du metabarcoding afin d'identifier et de réduire le plus possible les biais de quantification des espèces liés au choix de la méthode d'extraction de l'ADN (**Chapitre II**) et aux variations du nombre de copies du gène (**Chapitre III**), pour permettre le calcul d'indices diatomiques produisant des évaluations de qualité plus proches de celles obtenues via l'approche morphologique. Nous avons ensuite appliqué l'approche moléculaire optimisée à l'échelle des cours d'eau de Mayotte (**Chapitre IV**) et du réseau de surveillance des cours d'eau de France métropolitaine (**Chapitre V**), dans le but d'évaluer l'applicabilité de cette approche à grande échelle et de comparer son efficacité par rapport à l'approche morphologique. En parallèle de ce travail, nous avons aussi pu tester une nouvelle méthode permettant de compléter la base de référence à partir des données de metabarcoding obtenues sur des échantillons environnementaux (**Chapitre V, voir Article annexe II**). L'ensemble des travaux ont été réalisés en s'appuyant sur des échantillons biologiques variés (cultures pures, communautés artificielles, échantillons environnementaux) et ont permis de réaliser des ajustements techniques et méthodologiques sur l'approche moléculaire (voir **Figure 44**).

Cette discussion générale vise à apporter non seulement une synthèse des résultats obtenus mais également à aborder les perspectives qui en découlent.

	Chapitre II Biais extraction ADN	Chapitre III Biais nb. de copies de gène	Chapitre IV Application Mayotte	Chapitre V Application France (métropole)	Etude en cours Biais PCR et techno. HTS
Echantillons					
<i>Cultures pures</i>	✓	✓			
<i>Communautés artificielles</i>		✓			
<i>Biofilms environnementaux</i>	✓	✓	✓	✓	✓
<i>Communautés synthétiques</i>					✓
<i>Séquences aliènes</i>					✓
Extraction ADN					
<i>GenElute™-LPA</i>	✓	✓	✓	✓	✓
<i>MN - NucleoSpin Soil</i>	✓				
<i>MN - NucleoSpin Plant II</i>	✓				
<i>Q - DNeasy Blood and Tissue</i>	✓				
<i>I - Spin Plant Mini</i>	✓				
PCR quantitative					
<i>Primers rbcL (nb de cycles)</i>	Diat_rbcL_708F et R3 (30)	708F et R3 dégénérés (30)			
Amplification PCR					
<i>Primers rbcL (nb de cycles)</i>	Diat_rbcL_708F et R3 (30)	708F et R3 dégénérés (30)	708F et R3 dégénérés (30)	708F et R3 dégénérés (33)	708F et R3 dégénérés (33)
Séquençage HTS					
<i>PGM Ion Torrent</i>	✓	✓	✓		✓
<i>MiSeq Illumina (paired-end)</i>				✓	✓
Bio-informatique					
<i>Filtre qualité (long., qual., primer, chimères, ...)</i>	✓	✓	✓	✓	✓
<i>Filtre seq. unique (> 2 reads)</i>				✓	✓
<i>Regroupement en OTUs (95%)</i>	✓		✓	✓	✓
<i>Version base R-Syst</i>	01 - 2015	01 - 2015	01 - 2015	03 - 2017	05 - 2017

Figure 44 – Approches et techniques utilisées au cours de ces travaux de thèse.

1. Développement de l'approche moléculaire

1.1. Impact des biais techniques : avancées et perspectives

Comme évoqué précédemment, le metabarcoding est une méthode d'identification des espèces assez récente qui s'est surtout développée au cours des 10 dernières années (**Figure 45**). Il est important de bien comprendre que chaque étape du metabarcoding, depuis l'extraction de

l'ADN jusqu'à la méthode d'obtention de l'inventaire taxonomique, influence potentiellement les résultats acquis en termes de composition et d'abondance relative des taxons. Bien que de nombreuses études aient contribué à la compréhension des biais inhérents à chaque étape du metabarcoding, peu d'entre elles ont contribué spécifiquement au développement du metabarcoding des diatomées. Les travaux réalisés dans le cadre de ce travail de thèse apportent donc des éléments nouveaux quant aux impacts de différents biais techniques sur les inventaires taxonomiques de diatomées obtenus en metabarcoding.

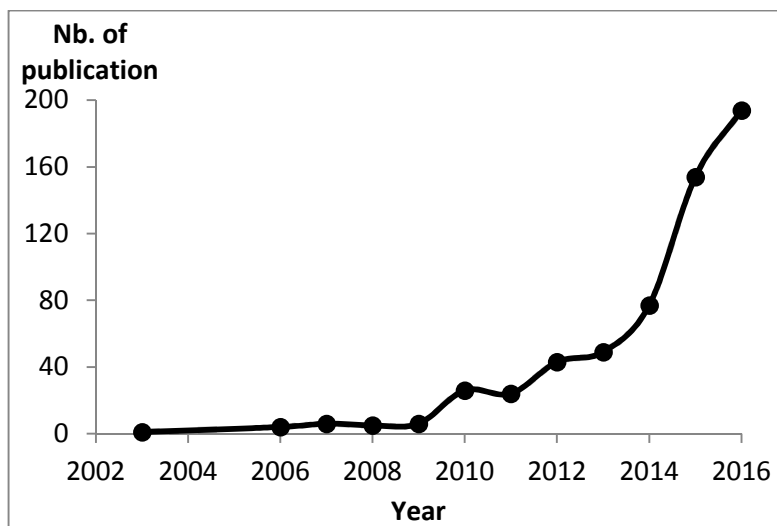


Figure 45 – Evolution au cours du temps du nombre d'articles incluant les termes « metabarcoding » OU « environmental barcoding ». Source: Web of Science Core Collection, base du 13/09/2017.

1.1.1. Extraction de l'ADN

Le protocole d'échantillonnage étant commun aux approches moléculaire et morphologique, la première étape spécifique au metabarcoding des diatomées consiste à extraire l'ADN d'échantillons de biofilms aquatiques. Le choix d'une méthode d'extraction appropriée dépend à la fois du type d'échantillon traité (*e.g.* tissu, plasma, eau, sol, biofilm) et du groupe d'organisme ciblé. Ainsi on trouve une grande diversité de méthodes et de protocoles commercialisés qui sont adaptés à différents types d'échantillons (voir Dhaliwal 2013 pour une liste détaillée). Cependant, la membrane cellulaire de la majorité des organismes étant généralement facile à lyser, les méthodes d'extraction de l'ADN ne sont pas forcément adaptées aux organismes possédant des structures fortifiées comme les diatomées avec leur frustule de silice. De par sa capacité à fixer l'ADN en présence de certains composés chimiques (Boom et al 1990), le frustule apparaît comme un obstacle pour l'extraction de l'ADN, ce qui est paradoxal quand on sait que la silice qui compose le frustule des diatomées est fréquemment utilisée dans

des kits d'extraction et de purification d'ADN (*e.g.* Geneclean® Kit For Ancient DNA, thèse Grunenwald 2014). La solidité du frustule varie d'une espèce à l'autre de diatomée en fonction de la forme, des motifs (*e.g.* pores, stries, raphé) et de la quantité de silice qui composent le frustule (Hamm *et al.* 2003; Moreno *et al.* 2015). Cette variabilité a pour effet de rendre le frustule plus ou moins facile à briser en fonction des espèces, et ce, même avec des méthodes de lyses considérées comme très efficaces et fréquemment utilisées comme le « bead beating » (Eland *et al.* 2012). Dans le **Chapitre II**, nous mettons en évidence l'effet de ce biais sur les inventaires taxonomiques de diatomées obtenus en metabarcoding. Bien que les compositions taxonomiques obtenues avec les différentes méthodes d'extraction testées soient similaires, des différences significatives en termes de quantification de certains taxons ont été observées indiquant que l'ADN de ces taxons n'est pas extrait de la même manière en fonction de la méthode d'extraction utilisée. C'est l'une des raisons pour laquelle nous avons décidé de travailler avec la méthode d'extraction d'ADN qui utilise le GenElute, car elle intègre plusieurs méthodes de lyse cellulaire (physique, thermique, enzymatique) et offre ainsi de meilleures possibilités d'extraction de l'ADN des différentes espèces de diatomées.

1.1.2. Amplification PCR et séquençage HTS

Comme nous l'avons évoqué dans le **Chapitre I** (partie 4.2.2), les étapes d'amplification de l'ADN par PCR et de séquençage HTS peuvent impacter fortement les inventaires taxonomiques.

En ce qui concerne l'amplification PCR, nous avons pu améliorer les « primers » utilisés pour amplifier le barcode *rbcL* en augmentant le nombre de dégénérescences dans leurs séquences (voir le **Chapitre IV** pour plus de détails). Bien que le risque d'amplification aspécifique soit augmenté, cela permet de limiter au mieux la fixation préférentielle des « primers » à certaines séquences, ce qui permet d'élargir le nombre d'espèces amplifiées comme montré *in silico* et sur des communautés de macroinvertébrés d'eau douce (Elbrecht & Leese 2017). Le changement de « primer » a ainsi permis de mieux détecter certains taxons des genres *Pinnularia*, *Achnanthisdium* ou encore *Cocconeis* jusqu'alors observés dans les inventaires morphologiques mais peu représentés dans les inventaires moléculaires.

Bien qu'il soit possible de limiter les biais liés à la PCR en utilisant des stratégies d'amplification adaptées (Kebschull & Zador 2015), il est difficile de contrôler la création de chimères et l'incorporation d'erreurs dans les séquences qui ne peuvent être détectées et éliminées que lors des traitements bio-informatiques (Schloss 2016). Afin de s'affranchir complètement des biais d'amplification, une nouvelle approche d'enrichissement basée sur la

capture de gènes a été intégrée avec succès au metabarcoding réalisé sur des macroinvertébrés d'eau douce (Dowle *et al.* 2016) et des arthropodes (Shokralla *et al.* 2016). Pour cela on utilise des sondes ADN qui se fixent par complémentarité à une région du génome des taxons ciblés (*e.g.* gène *rbcl*) et qui peuvent ensuite être récupérées grâce à des billes magnétiques, permettant d'isoler la région ciblée et de la séquencer directement en HTS sans amplification PCR. Une telle approche pourrait être appliquée aux metabarcoding des diatomées.

Au cours de ces travaux de thèse, deux technologies de séquençages HTS ont été utilisées : PGM Ion torrent (**Chapitre II, III, IV**) et MiSeq Illumina (**Chapitre V**). Ces technologies sont connues pour générer des erreurs de séquençage, généralement caractérisées par des insertions et des délétions de nucléotides dans les séquences de l'ordre de 0,06 à 1 % pour le PGM et moins de 0,01% pour le MiSeq (Bragg *et al.* 2013; Schirmer *et al.* 2015; Goodwin *et al.* 2016). Bien que la technologie MiSeq présente moins d'erreurs que le PGM, les filtres bio-informatiques que nous avons appliqués pour trier les séquences, notamment l'étape de « precluster », nous ont permis de limiter au maximum l'impact de ce biais. Cependant, le PGM a tendance à générer des séquences de longueurs variables en comparaison du MiSeq, comme déjà observé précédemment (Salipante *et al.* 2014). Sur l'ensemble des données de séquençage HTS issues du metabarcoding des diatomées ciblant le fragment *rbcl* de 312 pb, cela se traduit par environ 50 % des séquences avec une longueur inférieures à 250 pb en PGM, séquences qui sont donc inexploitées car trop courtes. La technologie MiSeq quant à elle permet de générer des séquences complètes, tout en permettant un séquençage bidirectionnel de chaque séquence. Elle permet également de séquencer plus d'échantillons en parallèle, ce qui rend cette technologie bien adaptée à l'avenir pour le metabarcoding des diatomées.

Afin d'évaluer directement comment les biais liés à la PCR et aux HTS affectent le metabarcoding des diatomées, un projet en collaboration avec Eric Pilgrim de l'Environmental Protection Agency (EPA, USA) a été initié dans le cadre de ces travaux de thèse. Les résultats de cette étude n'étant pas encore finalisés, ils n'ont pas été intégrés dans le manuscrit de thèse. Comme il est difficile de différencier les biais liés à la PCR de ceux dus au séquençage HTS, nous avons mis en place différentes stratégies dans le cadre de ce projet : (i) des échantillons environnementaux de biofilms aquatiques ont été séquencés en plusieurs répliques, à la fois avec le PGM et le MiSeq afin de comparer la structure des communautés de diatomées obtenue via les deux méthodes ; (ii) des séquences ADN correspondant à des barcodes *rbcl* (312 pb) de différentes espèces de diatomées ont été synthétisées *de novo* (par la société Eurofins Genomics). Ces séquences dites « synthétiques » ont ensuite été mélangées dans des proportions connues afin de créer des communautés synthétiques qui peuvent être séquencées directement en HTS, sans

étape de PCR. De cette manière il est possible de dissocier le biais du séquençage de celui de la PCR. (iii) des séquences « Alien » (*e.g.* séquence ADN de souris) amplifiable en même temps que l'ADN des diatomées, grâce à l'ajout des primers *rbcL* utilisés pour cibler les diatomées aux extrémités de ces séquences Alien, ont été incorporées en proportions connues dans les ADN extraits à partir des biofilms aquatiques, avant d'effectuer la PCR. De cette manière, il est possible de visualiser facilement dans les inventaires moléculaires la création de chimères ou des contaminations croisées entre échantillons. Sur ce dernier point, les résultats préliminaires ont montré qu'une séquence Alien initialement incorporée dans un seul échantillon avant PCR et séquençage HTS a été détectée dans d'autres échantillons après séquençage. Cela peut être dû à des problèmes lors de la préparation des libraires de séquençage.

1.1.3. Base de référence de barcodes

Le biais inhérent à la complétude de la base de référence est particulièrement important pour les diatomées, du fait de la difficulté à isoler et cultiver de nouvelles espèces pour pouvoir les séquencer. Bien que la base R-syst::diatom soit régulièrement mise à jour et complétée avec des séquences de nouveaux taxons, il reste des lacunes et des taxons représentatifs de certaines régions et de certaines qualité de milieux ne sont pas encore référencés. Dans le **Chapitre IV** on a pu voir pour les cours d'eau de Mayotte que 82 % des espèces identifiées en morphologie ne sont pas référencées ; de la même manière dans le **chapitre V**, 47 des 100 espèces les plus fréquemment identifiées en morphologie dans les cours d'eau de France métropolitaine ne sont pas référencées. Cela se traduit dans les inventaires taxonomiques moléculaires par des échantillons avec un fort pourcentage de séquences non assignées taxonomiquement (parfois jusqu'à 97 %, voir **Chapitre IV**), ce qui peut biaiser les notes de qualité obtenues. Il est donc nécessaire de poursuivre le travail de complétion de la base de référence afin de pouvoir utiliser l'ensemble de la donnée moléculaire et d'obtenir des inventaires taxonomiques complets. Pour cela, une solution est de réaliser l'amplification et le séquençage à partir de l'ADN extrait d'une seule cellule de diatomée, ce qui permet de s'affranchir de l'étape de mise en culture. Cette méthode a été appliquée avec succès sur des cultures pures de diatomées ainsi que des échantillons environnementaux de diatomées (benthiques et planctoniques) frais et fixés dans de l'éthanol (Lang & Kaczmarek 2011; Hamilton *et al.* 2015; Khan-Bureau *et al.* 2016). Cependant, afin d'associer une taxonomie à la séquence de référence, cette méthode n'est applicable que si il est possible d'identifier à quelle espèce appartient la cellule séquencée, ce qui n'est pas toujours évident sur la base d'une seule cellule vivante (les observations microscopiques étant réalisées sur

plusieurs cellules mortes). Dans le **Chapitre V**, nous avons pu mettre en place une nouvelle méthode de complétion décrite par (Rimet *et al.* 2018) permettant d'utiliser des séquences issues des données HTS d'échantillons environnementaux. Bien que cette approche ne permette de cibler que des taxons retrouvés en forte abondance dans les inventaires morphologiques, elle offre la possibilité opérationnelle de compléter rapidement et à moindre coût la base de référence de barcode. De plus, avec l'augmentation du nombre d'études utilisant le metabarcoding des diatomées et l'acquisition de jeux de données plus conséquents, il sera de plus en plus facile de mettre en place cette approche. Il est cependant important de bien conserver la traçabilité de l'origine de ces séquences environnementales obtenues en HTS afin de bien les identifier par rapport aux séquences obtenues à partir de culture pure et par séquençage Sanger.

1.2. Sources de divergences entre inventaires : deux visions d'une même réalité

Afin de valider les inventaires taxonomiques de diatomées obtenus en metabarcoding, on les compare à ceux obtenus via l'approche morphologique. Cependant il est important de rappeler que la correspondance entre ces deux méthodes ne pourra jamais être parfaite, en partie à cause des biais techniques inhérents à chacune d'elle, mais aussi parce qu'elles se basent sur des critères différents pour décrire la structure et la diversité des communautés de diatomées. Pour les diatomées, plusieurs points sont particulièrement importants car ils impactent la comparaison des approches morphologiques et moléculaires, aussi bien en termes d'identification que de quantification des taxons, à savoir :

- **Erreurs d'identification** : que ce soit avec l'une ou l'autre des approches, il peut y avoir des erreurs lors de l'identification des taxons. Pour l'approche morphologique, il s'agit d'erreurs qui surviennent lors de la détermination au microscope, principalement lorsque des espèces présentent des caractéristiques morphologiques similaires ou bien que les frustules observés sont de petite taille et difficiles à reconnaître (Bellinger & Sigee 2015). Avec l'évolution constante de la taxonomie des diatomées, il existe plusieurs noms pour une même espèce ce qui crée des problèmes de synonymie. Pour l'approche moléculaire, il s'agit d'erreurs qui surviennent lors de l'assignation taxonomique des séquences avec la base de référence, et dépendent principalement de la qualité des séquences obtenues en metabarcoding (*e.g.* erreurs incorporées lors de la PCR ou du séquençage), de la qualité de la base de référence et des algorithmes d'assignation utilisés (Schloss 2016; Balvočiūtė & Huson 2017). L'avantage

de l'approche moléculaire est que cette erreur est plus facilement répétable car automatisée, tandis qu'avec l'approche morphologique elle varie en fonction de la personne qui réalise l'identification et de son expertise au microscope (Kahlert *et al.* 2009).

- **Résolution taxonomique** : les deux approches ont la capacité d'identifier les taxons au niveau spécifique voire même sub-spécifique. Ceci dépend de la capacité des critères d'identifications utilisés, à savoir la variabilité morphologique du frustule pour l'approche morphologique et la variabilité génétique du barcode pour l'approche moléculaire, à discriminer les taxons entre eux, ce qui varie d'un genre voire d'une espèce de diatomée à une autre. Par exemple, Rimet *et al.* (2018) ont mis en évidence, sur la base des séquences *rbcl* obtenues en metabarcoding dans le **Chapitre IV**, plusieurs clades au sein de l'espèce *Gomphonema parvulum*, explicables par une diversité cryptique non visible en microscopie. A l'inverse, il nous est impossible de différencier les espèces *Fragilaria vaucheriae*, *F. rumpens* et *F. nanoides* car les séquences ADN de leur barcode *rbcl* (312 pb) sont identiques. Ce qui explique qu'on ne puisse les identifier dans les inventaires moléculaires obtenus dans le **Chapitre V**, bien qu'elles soient présentes dans les inventaires morphologiques correspondant.
- **Faux positifs** : que ce soit par la présence de frustules ou d'ADN issus d'individus morts, on détecte des faux positifs avec les deux approches. Ces biais ont déjà été discutés pour les diatomées (Kermarrec *et al.* 2014), et sont dus à la persistance des frustules et de l'ADN dans les écosystèmes d'eau douce (Dejean *et al.* 2011). Les contaminations proviennent souvent des diatomées planctoniques qui sédimentent et se déposent sur les biofilms aquatiques. Ce phénomène peut être important lorsqu'on étudie les communautés dans les biofilms aquatiques des lacs, avec une dominance de ces diatomées planctoniques dans les inventaires moléculaires (Rivera *et al.* 2017). Bien que moins problématique lorsqu'on travaille sur les communautés benthiques de rivières, on retrouve fréquemment des taxons planctoniques dans les inventaires moléculaires et morphologiques (*e.g.* genres *Cyclotella*, *Stephanodiscus*, **Chapitre IV et V**).
- **Profondeur d'analyse** : lorsqu'on compare les inventaires morphologiques et moléculaires, les comptages au microscope sont limités à 400 frustules par échantillon, tandis qu'aucune limite n'est fixée en termes de nombre de séquences par échantillon en metabarcoding. Ajouter à cela une capacité de détection du metabarcoding supérieure à l'approche morphologique, puisque par individu on a plusieurs séquences contre 1 frustule, le metabarcoding à la possibilité de mettre en évidence plus facilement des espèces peu abondantes et rares. Ceci explique en partie (avec le choix de la méthode de clustering et du

seuil de similarité utilisé pour définir les OTUs) pourquoi on obtient une richesse en OTU plus importante que la richesse spécifique obtenue avec l'approche morphologique, ce dans toutes les études menées dans ces travaux de thèse.

- **Méthode de quantification des taxons** : tandis que l'approche morphologique se base sur les abondances relatives de valves observées en microscopie, l'approche moléculaire utilise les abondances relatives de séquences pour quantifier les taxons. Les grosses espèces de diatomées étant généralement moins abondantes que les petites espèces dans l'environnement, elles sont plus facilement sous-estimées en microscopie (Snoeijs *et al.* 2002). A l'inverse, les travaux réalisés dans le **Chapitre III** montrent que les grosses espèces sont surestimées en metabarcoding, ceci étant dû au fait que le nombre de copies du gène *rbcL* est corrélé au biovolume cellulaire des diatomées. A cela s'ajoute le fait que l'approche morphologique ne prend pas en compte l'état physiologique des cellules, comme elle se base sur des comptages de frustules vidés de leur contenu cellulaire. A l'inverse, la quantité d'ADN dans une cellule pouvant varier en fonction de son état physiologique (Veldhuis *et al.* 2001) et donc impacter le nombre de copie du gène *rbcL*, l'état physiologique des cellules influence l'abondance relative des espèces obtenue en metabarcoding (voir discussion **Chapitre III**). Ces facteurs ont pour effet de contribuer fortement aux différences de quantification observées entre les inventaires morphologiques et moléculaires.

2. Application de l'approche moléculaire pour l'évaluation de la qualité des cours d'eau

2.1. Congruence entre approches morphologique et moléculaire

Bien que les approches morphologique et moléculaire fournissent des inventaires taxonomiques de diatomées qui ont en général une correspondance partielle, surtout au niveau spécifique, on observe généralement une forte congruence entre les évaluations de qualité basées sur ces 2 types d'inventaires. Nous avons pu faire ce constat dans les **Chapitre II, III, IV, V**, constat qui rejoint les résultats d'autres études réalisées sur l'utilisation du metabarcoding pour évaluer l'état écologique des cours d'eau (Kermarrec *et al.* 2014; Visco *et al.* 2015; Apothéloz-Perret-Gentil *et al.* 2017). Il n'est donc pas nécessaire que les inventaires taxonomiques obtenus avec les deux approches soient identiques pour permettre une évaluation correcte de la qualité des cours d'eau, ce qui peut s'expliquer par plusieurs facteurs.

Les indices diatomiques (*e.g.* IPS, IBD) utilisés pour évaluer l'état écologique des cours d'eau nécessitent d'identifier les taxons au niveau spécifique, ce afin d'utiliser les valeurs écologiques de chaque espèce dans le calcul de l'indice. Avec l'approche moléculaire, il est fréquent qu'on ne réussisse pas à assigner un nom d'espèce aux séquences environnementales, mais seulement un nom de genre ($\approx 40\%$ des séquences assignées à l'espèce contre 72% au genre dans le **Chapitre IV** ; 58% contre 77% dans le **Chapitre V**). On utilise alors les valeurs écologiques associées aux genres dans le calcul de l'indice, ce qui permet de prendre en compte plus de données dans le calcul de l'indice moléculaire, même si les valeurs écologiques utilisées sont plus moyennées. Plusieurs études ont montré qu'il n'est pas nécessaire de travailler au niveau spécifique pour avoir une bonne évaluation de l'état écologique des cours d'eau, l'utilisation de niveaux supérieurs comme le genre offrant une résolution suffisante (Kelly *et al.* 1995; Bigler *et al.* 2009; Rimet & Bouchez 2012a). Ceci a été confirmé récemment par (Keck *et al.* 2015), qui a mis en évidence la présence d'une relation entre la phylogénie et l'écologie des diatomées, montrant ainsi que des espèces phylogénétiquement proches partagent souvent des préférences écologiques similaires, ce, à un niveau taxonomique supérieur à celui de l'espèce. De ce fait, l'utilisation des valeurs écologiques des genres dans le calcul des indices de qualité moléculaire permet une évaluation de qualité qui est semblable à celle que nous aurions obtenue en travaillant au niveau spécifique. Cette approche a d'ailleurs été mise en place avec succès sur la base des inventaires morphologiques avec la création de l'Indice Diatomique Générique (IDG) qui prenait en compte seulement les valeurs écologiques des genres et donnait des résultats satisfaisants (Coste 1982; Rumeau & Coste 1988). Cependant de grands écarts de notes entre les indices IPS et IDG étaient observés pour certains sites (jusqu'à 7,5 points), principalement à cause de certains genres (*e.g.* *Navicula*, *Nitzschia*) qui renferment des espèces couvrant toute la gamme de sensibilité à la pollution (Prygiel & Coste 1993). Afin d'éviter ce problème, un indice basé sur l'identification de 45 genres et 91 espèces avait été proposé : l'Indice Diatomique Pratique (IDP) (Prygiel *et al.* 1996). Il est donc envisageable de créer un indice moléculaire utilisant dans son calcul à la fois les valeurs écologiques des genres et des espèces.

Bien que la possibilité de travailler au niveau du genre permette d'utiliser plus de données en metabarcoding, la correspondance entre les inventaires taxonomiques morphologiques et moléculaires reste globalement faible (*e.g.* 13% des espèces et 59% des genres communs aux deux approches dans le **Chapitre IV**). Avec si peu de similarité on devrait s'attendre à plus de divergence dans les notes de qualités obtenues avec les deux approches. Cependant, les espèces et les genres détectés par les deux approches sont généralement ceux retrouvés en forte abondance dans les inventaires, c'est-à-dire ceux qui influencent fortement la note de l'indice.

Bigler *et al.* (2009) ont montré que la suppression des taxons de diatomées avec une abondance relative < 5 % a un impact limité sur les valeurs d'indices obtenues. Comme les taxons peu abondants ont un faible impact sur le calcul de l'indice, le fait de ne pas détecter ces taxons avec l'une ou l'autre des approches affecte peu la congruence entre les indices de qualité morphologique et moléculaire. C'est pourquoi, certains indices comme l'IBD ne prennent pas en compte certains taxons rares (Prygiel *et al.* 2002). De plus, au vu des résultats obtenus dans le **Chapitre III**, il faudra prendre en compte les abondances relatives des taxons après application du facteur de correction pour définir si une espèce est abondante ou non dans les inventaires moléculaires.

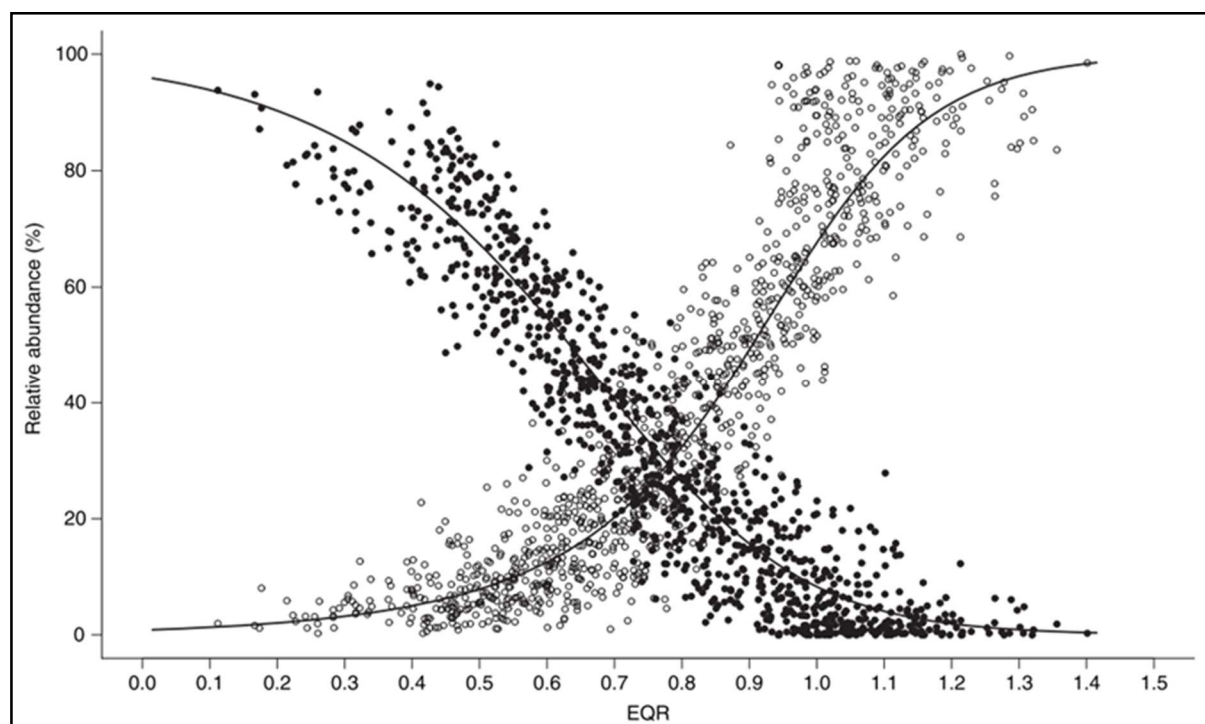


Figure 46 – Abondances relatives des taxons de diatomées sensibles (cercle blanc) et tolérants (cercle noir) aux nutriments en fonction du ratio de qualité écologique (EQR) obtenus pour des cours d'eau.

n= 1051 sites, source: Kelly *et al.* 2008.

Pour certains sites, on se retrouve dans une situation où la majorité des séquences sont non assignées au niveau de l'espèce et parfois même du genre (jusqu'à plus de 90 % pour les cas les plus extrêmes). Malgré cela, on obtient fréquemment des notes de qualité moléculaires qui sont assez proches des notes de qualité morphologiques. Une hypothèse est que les espèces abondantes retrouvées sur un même site partagent des préférences écologiques semblables, donc elles possèdent en théorie des valeurs indicatrices de tolérance et de sensibilité similaires. Kelly *et al.* (2008) ont pu montrer, à partir des inventaires morphologiques de diatomées obtenues sur un grand nombre de rivières anglaises, que l'abondance relative des taxons sensibles ou tolérants

aux nutriments varie en fonction de la qualité écologique du milieu (**Figure 46**). Ainsi pour les sites présentant des états écologiques les plus marqués, seuls des taxons sensibles (dans le cas de sites non impactés) ou des taxons tolérants (dans le cas de sites fortement impactés) sont présents dans les inventaires taxonomiques morphologiques. Pour ces sites avec un statut écologique marqué, la note de qualité obtenue lorsqu'on identifie toutes les espèces dominantes d'un échantillon devrait donc être assez proche de celle obtenue si l'on ne détecte qu'une partie des espèces dominantes. Cela peut expliquer pourquoi il est possible d'avoir une évaluation de qualité correcte à partir des inventaires moléculaires même si ceux-ci sont incomplets, l'important étant de détecter au moins une partie des taxons dominants. Il sera intéressant de vérifier cette hypothèse, qui peut aussi être testée pour simplifier les dénombrements morphologiques, sur un jeu de donnée important, comme celui qui sera obtenu dans le **Chapitre V** en 2018.

2.2. Potentiel d'amélioration et d'évolution des indices diatomiques

Afin d'évaluer le potentiel du metabarcoding comme outil d'évaluation de l'état écologique des cours d'eau, on calcul les indices diatomiques en utilisant les inventaires moléculaires puis on compare les valeurs obtenues avec celles calculées à partir des inventaires morphologiques. Ces indices (*e.g.* IPS, IBD) ont été développés à l'origine sur la base des inventaires morphologiques et ne sont donc pas originellement adaptés aux données moléculaires. Dans le **Chapitre III**, nous avons mis en évidence la nécessité d'appliquer un facteur de correction basé sur le biovolume des espèces pour produire des inventaires moléculaires proches des inventaires microscopiques en termes d'abondance relative. De la même manière, Vivien *et al.* (2016) avaient montré l'intérêt d'utiliser un facteur de correction afin d'obtenir des abondances relatives d'oligochètes plus proches des comptages microscopiques à partir des données de metabarcoding. Angly *et al.* (2014) ont proposé un programme (CopyRighter) qui corrige le biais de quantification dû aux variations du nombre de copies du gène 16S dans les inventaires moléculaires obtenus à partir de communautés microbiennes (bactéries, archées). Afin de pondérer le poids des espèces de diatomées à gros biovolume et d'améliorer le calcul des indices moléculaires, il pourra être intéressant d'inclure directement ce facteur de correction dans le calcul de l'indice (voir le détail de la formule dans le **Chapitre I**) :

$$\text{Valeur d'indice} = \frac{\sum_{j=1}^n a_j \cdot \mathbf{cf}_j \cdot s_j \cdot v_j}{\sum_{j=1}^n a_j \cdot \mathbf{cf}_j \cdot v_j}$$

Le fait d'appliquer ce facteur de correction permet de se placer dans les mêmes conditions d'évaluation de qualité que l'approche morphologique, à savoir que le biovolume des espèces de diatomées n'a pas d'importance dans le calcul des indices de qualité (Lavoie *et al.* 2009). Cependant, d'autres études prennent en compte le biovolume des diatomées dans le calcul d'indices servant à l'évaluation de l'état écologique des cours d'eau, afin de donner plus de poids aux grosses espèces généralement peu abondantes mais qui contribuent fortement à la biomasse totale et au fonctionnement (Lavoie *et al.* 2006). C'est notamment le cas des indices basés sur les traits écologiques développés pour évaluer les pollutions organiques et le niveau trophique des cours d'eau (Berthon *et al.* 2011), l'état écologique des cours d'eau de Mayotte (Tapolczai *et al.* 2017) ou encore d'un indice multi-métrique permettant d'évaluer l'impact de pressions multiples sur les assemblages de diatomées (Larras *et al.* 2017). A l'avenir, il pourrait donc être intéressant d'utiliser les inventaires moléculaires non corrigés, afin de prendre en compte à la fois l'abondance relative des taxons et leur biovolume.

Afin de pouvoir appliquer de manière optimale les indices diatomiques aux inventaires produits en metabarcoding moléculaire, il faudrait pouvoir assigner taxonomiquement toutes les séquences environnementales au niveau de l'espèce ou du genre. Comme nous l'avons vu précédemment, une forte proportion de séquences reste non assignée et n'est pas prise en compte dans le calcul des indices. Bien que les bases de références continuent d'être implémentées (Rimet *et al.* 2016) et que de nouvelles méthodes de complétions soient proposées (Rimet *et al.* 2018, voir **Chapitre V**), il est peu probable que l'on arrive à assigner la totalité des séquences environnementales. Cependant, si l'on se positionne en termes d'évaluation environnementale, l'assignation taxonomique sert principalement à attribuer des valeurs écologiques aux séquences environnementales afin de permettre le calcul des indices de qualité. Une alternative serait d'attribuer des valeurs écologiques aux séquences non plus en passant par l'assignation taxonomique, mais en utilisant une approche phylogénétique. Keck *et al.* (2015) ayant montré que des espèces phylogénétiquement proches partagent des préférences écologiques similaires, il est en théorie possible d'estimer les valeurs écologiques des séquences environnementales en fonction de leur remplacement dans une phylogénie de référence. C'est ce que nous avons mis en place, en parallèle des travaux de thèse présentés dans ce manuscrit, sur les inventaires moléculaires obtenus dans le **Chapitre V (voir Article annexe III)**. Pour cela nous avons (i) utilisé une phylogénie de référence pour le gène *rbcL* comprenant 604 espèces avec des valeurs écologiques connues, (ii) les séquences environnementales, correspondant aux séquences représentatives des OTUs (95 %) obtenus en metabarcoding, ont été placées dans la phylogénie

de référence, (iii) les valeurs de sensibilité (IPSS) et de tolérance (IPSV) de chaque séquence environnementale ont été estimées en fonction de leur distance phylogénétique vis-à-vis des séquences de référence, (iv) 3 indices moléculaires ont ainsi pu être calculés à partir des données obtenues en metabarcoding : l'indice IPS-DNAtaxo basé sur l'assignation taxonomique des séquences (valeurs présentées dans le **Chapitre V**), l'indice IPS-DNAphylo basé entièrement sur l'approche phylogénétique, ainsi que l'indice IPS-DNAhybrid utilisant les valeurs écologiques issues de l'assignation taxonomique pour les séquences assignées jusqu'au genre et les valeurs écologiques obtenues via l'approche phylogénétique pour les séquences non assignées au genre. Lorsqu'on compare les valeurs obtenues pour les 3 indices moléculaires aux valeurs d'IPS via les inventaires morphologiques, les indices IPS-DNAtaxo et IPS-DNAphylo sont corrélés de la même manière avec l'IPS-morpho ($r \approx 0.65$), tandis qu'on obtient une meilleure corrélation entre l'IPS-DNAhybrid et l'IPSmorpho ($r = 0.74$). L'utilisation conjointe de la taxonomie et de la phylogénie offre donc une perspective intéressante permettant d'exploiter toute la donnée moléculaire et d'améliorer la fiabilité des indices moléculaires produits, tout en bénéficiant de l'information écologique associée aux taxons.

2.3. Autre lecture des données HTS pour évaluer l'état écologique

Au lieu d'utiliser les valeurs écologiques définies pour les taxons sur la base de données morphologiques, il est possible de déterminer directement les valeurs écologiques des OTUs et de proposer, non plus un indice taxonomique, mais un indice OTU (Apothéloz-Perret-Gentil *et al.* 2017). Pour cela il est nécessaire d'avoir un jeu de données suffisamment grand et comprenant des échantillons avec un gradient de qualité marqué. En croisant les inventaires moléculaires (liste et abondance relative des OTUs) avec les données de physico-chimie collectées sur chaque site, il est possible de définir les valeurs de sensibilité et de tolérance de chaque OTU aux différents paramètres physico-chimiques testés. De cette manière on s'affranchit des problèmes liés à l'assignation taxonomique et on peut utiliser toute la donnée moléculaire produite en metabarcoding via l'indice OTU. Cette approche a été mise en place avec succès sur le jeu de données metabarcoding issu des rivières de Mayotte (**Chapitre V**), ce qui a permis d'obtenir une bonne évaluation de l'état écologique des cours d'eau (Tapolczai *et al.* 2017, in review). L'utilisation de cette approche est particulièrement appropriée lorsqu'on travaille avec des cours d'eaux pour lesquels il manque des taxons représentatifs dans la base de référence, ce qui est notamment le cas pour Mayotte. Cependant, une des limites de cette méthode est le fait de se déconnecter totalement de la réalité biologique, puisqu'on n'identifie plus les espèces retrouvées

dans l'échantillon, ce qui aboutit à une perte de diagnostic liée à l'utilisation d'éléments biologiques pour lesquels on possède une connaissance écologique de longue date.

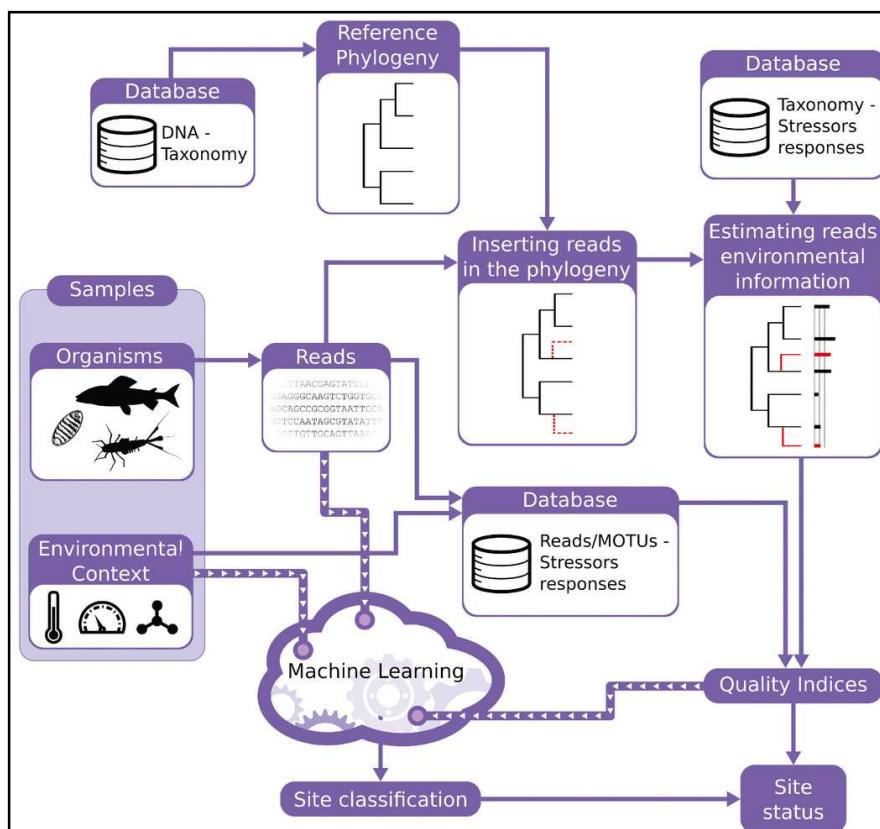


Figure 47 – Différentes stratégies d'évaluations de l'état écologique des cours d'eau à partir des données moléculaires obtenues en metabarcoding.

Source: Keck *et al.* 2017.

En plus des approches basées sur la phylogénie et l'utilisation des indices OTUs, on peut utiliser l'approche en « machine learning » pour s'affranchir de l'étape d'assignation taxonomique dans les inventaires moléculaires (**Figure 47**). L'objectif est d'utiliser des algorithmes capables d'analyser des jeux de données de séquences ADN issus du séquençage HTS obtenu en metabarcoding et de prédire, sur la base de motifs de séquences, l'état écologique des échantillons analysés. Pour cela il est nécessaire d'entraîner l'algorithme sur un jeu de données contenant plusieurs échantillons avec des états écologiques variés et de lui renseigner le plus d'informations possibles sur ces sites (*e.g.* paramètres physico-chimiques des sites, données environnementales, notes de qualité des sites). Une fois l'apprentissage réalisé, l'algorithme est capable de prédire l'état écologique d'un nouvel échantillon à partir des données de séquençage (Keck *et al.* 2017). Cette approche a été utilisée avec succès pour évaluer l'état écologique d'échantillons marins sur la base des données HTS obtenues à partir de communautés benthiques de foraminifères (Cordier *et al.* 2017). Il serait intéressant d'appliquer cette approche au metabarcoding des diatomées et voir son potentiel pour l'évaluation de la qualité des cours d'eau.

3. Vers une utilisation en routine de l'approche moléculaire ?

3.1. Fiabilisation des données moléculaires

Les perspectives d'amélioration de l'approche moléculaire proposées précédemment (approche phylogénétique, indice OTU, machine learning) nécessitent des études complémentaires. Cependant, des mesures peuvent être mises en place pour fiabiliser les indices moléculaires actuels basés sur l'assignation taxonomique.

Nous avons pu voir que l'approche moléculaire permet, dans l'ensemble, d'évaluer l'état écologique des cours d'eau de manière satisfaisante. De plus les études réalisées sur les cours d'eau de Mayotte (**Chapitre IV**) et sur le réseau de surveillance des cours d'eau de France métropolitaine (**Chapitre V**) nous ont permis de valider l'application de l'approche moléculaire à grande échelle. Nous avons ainsi confirmé que l'approche moléculaire est plus rapide et plus économique que l'approche morphologique pour traiter et analyser un grand nombre d'échantillon. Comme le traitement des échantillons et l'analyse des données se fait de manière automatique, les biais introduits lors de la création des inventaires moléculaires sont répétables et les mêmes pour tous les échantillons, contrairement à l'approche morphologique où les inventaires morphologiques peuvent varier d'un observateur à l'autre (Kahlert *et al.* 2009).

En termes d'optimisation, les diverses études qui ont été menées pour développer le metabarcoding des diatomées ont généralement été réalisées en laboratoire dans des conditions contrôlées (*e.g.* utilisation de communautés artificielles, cultures pures) avant d'être appliquées sur un faible nombre d'échantillons environnementaux (**Chapitre II et III**, Kermarrec *et al.* 2013, 2014; Zimmermann *et al.* 2015; Apothéloz-Perret-Gentil *et al.* 2017). Ces études sont bien loin d'avoir pu confronter et tester l'approche moléculaire dans toutes les conditions rencontrées dans l'environnement (*e.g.* conditions physico-chimiques, communautés de diatomées). L'application à grande échelle a ainsi permis de mettre en évidence les limites de l'approche moléculaire à évaluer l'état écologique de certains cours d'eau localisés dans des régions présentant des caractéristiques physiques et géochimiques particulières (*e.g.* régions à socle cristallins) ou pour lesquelles les espèces dominantes de diatomées sont très peu référencées (*e.g.* cours d'eau alpins, Mayotte). Il reste donc nécessaire de déployer l'approche moléculaire à une plus large échelle, comme celle des réseaux de surveillance DCE, afin de continuer à tester et à optimiser la méthode mais aussi afin de mieux définir ses limites d'application.

Avant de pouvoir envisager d'utiliser l'approche moléculaire en routine, il faudrait pouvoir standardiser les différentes étapes du metabarcoding, notamment en termes de :

- **Validation des inventaires moléculaires** : comme nous l'avons vu précédemment, toutes les étapes du metabarcoding peuvent avoir un impact sur les inventaires moléculaires produits. Ajouter au fait qu'il existe une grande diversité de méthodes et de protocoles, il devient vite délicat de comparer les résultats obtenus avec différentes approches. Afin de d'évaluer la qualité des inventaires moléculaires produits en metabarcoding et aussi permettre de comparer des inventaires issus de différentes études, il a été proposé d'intégrer systématiquement dans chaque séquençage une communauté artificielle dont la composition et la proportion des taxons sont connues (Kozich *et al.* 2013). De cette manière il est possible d'appliquer les traitements bio-informatiques adaptés à chaque run de séquençage, en fonction des erreurs observées sur cette communauté artificielle, et par la même, d'avoir un « standard » pour valider si les inventaires produits sont de bonne qualité ou non. Les communautés synthétiques ainsi que les séquences « Aliens » proposées précédemment (voir **Chapitre VI, section 1.1.2**) pourraient d'ailleurs être utilisées comme standards moléculaires en étant incluses systématiquement dans chaque séquençage et ainsi servir de référence pour valider la qualité des données moléculaires obtenues.
- **Profondeur de séquençage** : comme pour l'approche morphologique où l'on considère qu'il faut déterminer au minimum 400 valves pour produire un inventaire taxonomique suffisamment fiable pour permettre une bonne évaluation de l'état écologique du site étudié (AFNOR 2014a), il faudrait déterminer le nombre de séquence minimum requis par échantillon pour considérer que l'évaluation de qualité basée sur l'inventaire moléculaire est correcte. Dans le **Chapitre IV**, nous avons mis en évidence *in silico* que l'augmentation de la profondeur de séquençage ne change pas la note de qualité obtenue et qu'il est possible d'avoir des notes de qualité correctes avec un nombre très réduit de séquences (500 séquences). Il faudra évaluer les risques liés à l'utilisation de seuils aussi bas et cette observation devra être confirmée expérimentalement sur des sites présentant des communautés de diatomées variées en termes de diversité et de richesse afin de proposer un seuil adapté. Cela pourra être réalisé sur les données obtenues dans le **Chapitre V**.
- **Seuil de fiabilité de l'indice moléculaire** : jusqu'à présent, nous avons calculé les indices diatomiques à partir de tous les inventaires taxonomiques obtenus en metabarcoding. Cependant, comme évoqué précédemment, on utilise seulement la partie assignée des séquences, la partie non assignée n'étant pas prise en compte dans le calcul de l'indice et l'évaluation de qualité qui en découle. Dans la mesure où pour certains sites la proportion de séquences non assignées peut atteindre plus de 90 %, on est en droit de se demander qu'elle est la fiabilité que l'on peut accorder à la note de qualité produite. A l'avenir il pourra être

intéressant de donner, en plus de la valeur de l'indice moléculaire, un seuil de confiance basé sur la proportion de séquences non assignées de l'échantillon, ce afin de donner une indication sur la fiabilité de l'indice produit. L'autre option serait de définir un seuil, toujours basé sur la proportion des séquences non assignées, à partir duquel on considère qu'on ne peut pas calculer l'indice moléculaire pour l'échantillon.

- **Seuil de détection des taxons** : la plupart des études sur le metabarcoding des diatomées ont mis en place un filtre bio-informatique qui permet d'éliminer les séquences uniques qui sont présentes en faible nombre de copies (< 10 séquences en général) dans les données de séquençage HTS (**Chapitre V**, Visco *et al.* 2015; Apothéloz-Perret-Gentil *et al.* 2017). La mise en place d'un tel seuil permet d'éliminer les séquences qui correspondent à des artefacts de séquençage qui contiennent beaucoup d'erreurs, ce qui améliore la qualité des inventaires produits (Bokulich *et al.* 2012). Bien entendu, le risque est d'éliminer des taxons peu abondants, mais nous avons vu précédemment qu'ils ont peu de poids dans le calcul des indices de qualité. Cependant, au vue des résultats obtenus dans le **Chapitre III**, il est important de se méfier de la notion de taxon peu abondant dans les inventaires moléculaires car après l'application du facteur de correction de biovolume, celui-ci peut devenir abondant et avoir une contribution non négligeable dans le calcul de l'indice moléculaire. Il serait donc intéressant de mettre en place 2 filtres : (i) un filtre sur l'abondance des séquences uniques qui ne soit pas trop élevé (< 3 reads), (ii) un filtre sur les inventaires taxonomiques moléculaires corrigés afin de ne conserver que les taxons représentés par plus de 10 reads.

3.2. Intégration dans le paysage actuel de la bioindication et de la DCE

Au vu des résultats obtenus au cours de ces travaux de thèse et de la rapidité à laquelle le metabarcoding des diatomées se développe, il paraît évident que son utilisation comme outil d'évaluation de la qualité des cours d'eau n'est qu'une question de temps. Les discussions et échanges qui ont été initiés entre les chercheurs européens dans le cadre du réseau COST DNAqua-Net devraient permettre d'accélérer le processus d'optimisation de l'approche moléculaire (Leese *et al.* 2016). Cependant, dans l'hypothèse où cette approche serait entièrement opérationnelle, cela ne signifierait pas pour autant qu'elle serait implémentée rapidement à l'échelle des réseaux de surveillance. Il sera nécessaire de prendre en compte d'autres facteurs politiques, économiques et sociaux car l'intégration d'une nouvelle approche de

bioindication nécessiterait des modifications importantes en termes de compétences et d'infrastructures requises pour sa mise en œuvre. La prochaine étape est donc de savoir comment sera déployée l'approche moléculaire et qui sera en charge de l'appliquer, ce qui inclue les échantillonnages, le travail en laboratoire, les traitements bio-informatiques jusqu'à la validation des notes de qualité produites. Afin de répondre à ces questions, il est nécessaire de réunir les différents acteurs qui sont actuellement en charge de la bioindication (*e.g.* décideurs, chercheurs, gestionnaires, organismes en charge des suivis) afin de choisir la stratégie d'implémentation la plus adéquate, démarche déjà initiée au niveau européen dans le cadre du COST DNAqua-Net et aux acteurs français dans le cadre du projet sur les réseaux de surveillance des cours d'eau de France métropolitaine (**Chapitre V**).

Au vu des capacités de l'approche moléculaire pour l'évaluation de la qualité des cours d'eau, il paraît plus logique d'utiliser cette approche en complément de l'approche morphologique plutôt qu'en remplacement. L'approche moléculaire, grâce à sa capacité à traiter rapidement et à moindre coût un nombre important d'échantillons, pourrait être utilisée pour évaluer l'état écologique de cours d'eau présentant peu de risques (*e.g.* sites inclus dans le RCS). Dans les cas où une évaluation de qualité précise est requise, par exemple pour des sites à la limite du bon état écologique (*e.g.* sites inclus dans le RCO), ayant des particularités limitant l'utilisation du metabarcoding ou pour lesquels des actions de gestion ont été mises en place depuis plusieurs années, il sera alors plus intéressant de conserver l'approche morphologique (Kelly *et al.* 2015). Chacune des deux approches possède des points forts et des points faibles qui les caractérisent, mais elles ont toutes les deux leur place dans le paysage de la bioindication, l'objectif final étant toujours la préservation des cours d'eau. L'implémentation progressive de l'approche moléculaire dans le cadre de la DCE et de la surveillance des cours d'eau pourrait donc être envisagée en deux étapes calées sur les cycles de gestion DCE : (i) cycle de gestion 2015-2021 : appliquer les approches morphologique et moléculaire sur l'ensemble des réseaux de surveillance du territoire dans le but d'intercaler les 2 approches et de standardiser l'approche moléculaire tout en définissant ses limites d'application ; (ii) cycle de gestion 2022-2027 : mettre en place une double surveillance, morphologique et moléculaire, qui pourra jeter les bases d'un nouveau mode de gestion des cours d'eau au-delà de ces échéances.

Références bibliographiques

- Abell R (2002) Conservation biology for the biodiversity critic: A freshwater follow-up. *Conservation Biology*, **16**, 1435–1437.
- Ács É, Szabo K, Tóth B, Kiss KT (2004) Investigation of benthic algal communities, especially diatoms of some hungarian streams in connection with reference conditions of the water framework directives. *Acta Botanica Hungarica*, **46**, 255–278.
- AFNOR (2000) *Norme Française NFT 90-354. Détermination de l'Indice Biologique Diatomées (IBD)*.
- AFNOR (2003) Water quality - Guidance standard for the routine sampling and pretreatment of benthic diatoms from rivers. *European Standard EN 13946:1–15*.
- AFNOR (2004) Water quality: guidance standard for the identification, enumeration and interpretation of benthic diatom samples from running waters. *European Standard EN 14407:1–13*.
- AFNOR (2007) *Norme Française NFT 90-354. Détermination de l'Indice Biologique Diatomées (IBD) - mise à jour*.
- AFNOR (2014a) *NF EN 13946 – Qualité de l'eau – Guide pour l'échantillonnage en routine et le prétraitement des diatomées benthiques de rivières et de plans d'eau*.
- AFNOR (2014b) *NF EN 14407 – Qualité de l'eau – Guide pour l'identification et le dénombrement des échantillons de diatomées benthiques de rivières et de lacs*.
- AFNOR (2016) *Norme française NF T90 354 – Qualité de l'eau – Échantillonnage, traitement et analyse de diatomées benthiques en cours d'eau et canaux*.
- Albaina A, Aguirre M, Abad D, Santos M, Estonba A (2016) 18S rRNA V9 metabarcoding for diet characterization: a critical evaluation with two sympatric zooplanktivorous fish species. *Ecology and Evolution*, **6**, 1809–1824.
- Almeida SFP, Elias C, Ferreira J *et al.* (2014) Water quality assessment of rivers using diatom metrics across Mediterranean Europe: A methods intercalibration exercise. *Science of The Total Environment*, **476–477**, 768–776.
- Amend AS, Seifert KA, Bruns TD (2010) Quantifying microbial communities with 454 pyrosequencing: does read abundance count? *Molecular Ecology*, **19**, 5555–5565.
- Anders S, Pyl PT, Huber W (2015) HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)*, **31**, 166–9.
- Andrus JM, Winter D, Scanlan M *et al.* (2013) Seasonal synchronicity of algal assemblages in three Midwestern agricultural streams having varying concentrations of atrazine, nutrients, and sediment. *Science of The Total Environment*, **458–460**, 125–139.
- Angly FE, Dennis PG, Skarszewski A *et al.* (2014) CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. *Microbiome*, **2**, 11.
- Apothéloz-Perret-Gentil L, Cordonier A, Straub F *et al.* (2017) Taxonomy-free molecular diatom index for high-throughput eDNA biomonitoring. *Molecular ecology resources*, (**in press**).
- Armbrust EV, Chisholm SW (1992) Patterns of cell size change in a marine centric diatom: variability evolving from clonal isolates. *Journal of Phycology*, **28**, 146–156.
- Azim ME, Verdegem MCJ, van Dam AA, Beveridge MC (2005) *Periphyton: ecology, exploitation and management*. CABI Publishing.
- Bailey RC, Kennedy MG, Dervish MZ, Taylor ARM (1998) Biological assessment of freshwater ecosystems using a reference condition approach: comparing predicted and actual

- benthic invertebrate communities in Yukon streams. *Freshwater Biology*, **39**, 765–774.
- Baird DJ, Hajibabaei M (2012) Biomonitoring 2.0: A new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Molecular Ecology*, **21**, 2039–2044.
- Balian E V., Segers H, Lévêque C, Martens K (2008) The Freshwater Animal Diversity Assessment: an overview of the results. *Hydrobiologia*, **595**, 627–637.
- Balvočiūtė M, Huson DH (2017) SILVA, RDP, Greengenes, NCBI and OTT - how do these taxonomies compare? *BMC genomics*, **18**, 114.
- Barbour MT, Gerritsen J, Snyder BD, Stribling JB (1999) Rapid bioassessment protocols for use in streams and wadeable rivers: periphyton, benthic macroinvertebrates, and fish. EPA 841-B-99-002. U. S. Environmental Protection Agency, Office of Water, Washington, D.C., USA.
- Barker P (1992) Growth and reproductive strategies of freshwater phytoplankton. *Regulated Rivers: Research & Management*, **7**, 308–309.
- Battin TJ, Besemer K, Bengtsson MM, Romani AM, Packmann AI (2016) The ecology and biogeochemistry of stream biofilms. *Nature Reviews Microbiology*, **14**, 251–263.
- Bedoshvili YD, Likhoshway YV (2012) The Cell Ultrastructure of Diatoms - Implications for Phylogeny? In: *The Transmission Electron Microscope*, pp. 147–160. InTech.
- Bedoshvili YD, Popkova TP, Likhoshway Y V. (2009) Chloroplast structure of diatoms of different classes. *Cell and Tissue Biology*, **3**, 297–310.
- Bellinger BJ, Cocquyt C, O'Reilly CM (2006) Benthic diatoms as indicators of eutrophication in tropical streams. *Hydrobiologia*, **573**, 75–87.
- Bellinger EG, Sigeo DC (2015) *Freshwater Algae: Identification, Enumeration and Use as Bioindicators* (Wiley-Blackwell, Ed.).
- Bendich AJ (1987) Why do chloroplasts and mitochondria contain so many copies of their genome? *BioEssays*, **6**, 279–282.
- Bennett JR, Sisson DR, Smol JP *et al.* (2014) Optimizing taxonomic resolution and sampling effort to design cost-effective ecological models for environmental assessment (M Cadotte, Ed.). *Journal of Applied Ecology*, **51**, 1722–1732.
- Bere T (2016) Challenges of diatom-based biological monitoring and assessment of streams in developing countries. *Environmental Science and Pollution Research*, **23**, 5477–5486.
- Berger SA, Krompass D, Stamatakis A (2011) Performance, Accuracy, and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood. *Systematic Biology*, **60**, 291–302.
- Berthon V, Bouchez A, Rimet F (2011) Using diatom life-forms and ecological guilds to assess organic pollution and trophic level in rivers: a case study of rivers in south-eastern France. *Hydrobiologia*, **673**, 259–271.
- Besse-Lototskaya A, Verdonshot PF, Coste M, Van de Vijver B (2011) Evaluation of European diatom trophic indices. *Ecological Indicators*, **11**, 456–467.
- Besse-Lototskaya A, Verdonshot PFM, Sinkeldam JA (2006) Uncertainty in Diatom Assessment: Sampling, Identification and Counting Variation. *Hydrobiologia*, **566**, 247–260.
- Bigler C, Gälman V, Renberg I (2009) Numerical simulations suggest that counting sums and taxonomic resolution of diatom analyses to determine IPS pollution and ACID acidity indices can be reduced. *Journal of Applied Phycology*, **22**, 541–548.
- Bik HM, Porazinska DL, Creer S *et al.* (2012) Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in ecology & evolution*, **27**, 233–43.
- Birk S, Bonne W, Borja A *et al.* (2012) Three hundred ways to assess Europe's surface waters: An almost complete overview of biological methods to implement the Water Framework Directive. *Ecological Indicators*, **18**, 31–41.

- Blandin P (1986) Bioindicateurs et diagnostic des systèmes écologiques. *Bulletin d'écologie*, **17**, 215–307.
- Blomberg SP, Garland T, Ives AR (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, **57**, 717–745.
- Bokulich NA, Subramanian S, Faith JJ *et al.* (2012) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods*, **10**, 57–59.
- Bonada N, Prat N, Resh VH, Statzner B (2006) Developments in aquatic insect biomonitoring: a comparative analysis of recent approaches. *Annual Review of Entomology*, **51**, 495–523.
- Bondoc KG V, Heuschele J, Gillard J, Vyverman W, Pohnert G (2016) Selective silicate-directed motility in diatoms. *Nature communications*, **7**, 10540.
- Boucher N, Vaultot D, Partensky F (1991) Flow cytometric determination of phyto-plankton DNA in cultures and oceanic populations. *Marine Ecology Progress Series*, **71**, 75–84.
- Bourrelly P, Manguin E (1952) Algues d'eau douce de la Guadeloupe et dépendances. *SEDES, Paris, France*.
- Boyer F, Mercier C, Bonin A *et al.* (2016) obitools: a unix-inspired software package for DNA metabarcoding. *Molecular ecology resources*, **16**, 176–82.
- Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW (2013) Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLoS computational biology*, **9**, e1003031.
- Brembu T, Winge P, Tooming-Klunderud A *et al.* (2014) The chloroplast genome of the diatom *Seminavis robusta*: New features introduced through multiple mechanisms of horizontal gene transfer. *Marine Genomics*, **16**, 17–27.
- Britton T, Anderson CL, Jacquet D, Lundqvist S, Bremer K (2007) Estimating Divergence Times in Large Phylogenetic Trees. *Systematic Biology*, **56**, 741–752.
- Bruckner CG, Bahulikar R, Rahalkar M, Schink B, Kroth PG (2008) Bacteria Associated with Benthic Diatoms from Lake Constance: Phylogeny and Influences on Diatom Growth and Secretion of Extracellular Polymeric Substances. *Applied and Environmental Microbiology*, **74**, 7740–7749.
- Bruder K, Medlin LK (2007) Molecular assessment of phylogenetic relationships in selected species/genera in the naviculoid diatoms (Bacillariophyta). I. The genus *Placoneis*. *Nova Hedwigia*, **85**, 331–352.
- Bruggeman J, Heringa J, Brandt BW (2009) PhyloPars: estimation of missing parameter values using phylogeny. *Nucleic Acids Research*, **37**, W179–W184.
- Butcher RW (1947) Studies in the ecology of rivers. IV. The algae of organically enriched water. *Journal of Ecology*, **35**, 86–91.
- Capo E, Debroas D, Arnaud F *et al.* (2016) Long-term dynamics in microbial eukaryotes communities: a palaeolimnological view based on sedimentary DNA. *Molecular ecology*, **25**, 5925–5943.
- Capo E, Debroas D, Arnaud F *et al.* (2017) Tracking a century of changes in microbial eukaryotic diversity in lakes driven by nutrient enrichment and climate warming. *Environmental Microbiology*, **19**, 2873–2892.
- Caporaso JG, Kuczynski J, Stombaugh J *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nature methods*, **7**, 335–6.
- Carew ME, Pettigrove VJ, Metzeling L, Hoffmann A a (2013) Environmental monitoring using next generation sequencing: rapid identification of macroinvertebrate bioindicator species. *Frontiers in zoology*, **10**, 45.
- Carlisle DM, Wolock DM, Meador MR (2011) Alteration of streamflow magnitudes and potential ecological consequences: A multiregional assessment. *Frontiers in Ecology and*

- the Environment*, **9**, 264–270.
- Carpenter SR, Caraco NF, Correll DL *et al.* (1998) Nonpoint pollution of surface waters with phosphorus and nitrogen. *Ecological Applications*, **8**, 559–568.
- Cemagref (1982) Étude des méthodes biologiques quantitative d’appréciation de la qualité des eaux. *Bassin Rhône-Méditerranée-Corse. Centre National du Machinisme Agricole, du Génie rural, des Eaux et des Forêts, Lyon, France.*
- CEN - EN 13946 (2004) Water quality - Guidance standard for the routine sampling and pretreatment of benthic diatoms from rivers. *European Standard.*
- CEN EN 14407 (2004) Water quality – Guidance standard for the identification, enumeration and interpretation of benthic diatom samples from running waters. *European standard.*
- Chapman D V. (1996) *Water quality assessments: a guide to the use of biota, sediments and water in environmental monitoring.* E & F Spon London.
- Chariton AA, Stephenson S, Morgan MJ *et al.* (2015) Metabarcoding of benthic eukaryote communities predicts the ecological condition of estuaries. *Environmental pollution (Barking, Essex : 1987)*, **203**, 165–74.
- Chave P (2001) *The EU water framework directive.* IWA publishing.
- Chepurnov VA, Mann DG, Sabbe K, Vyverman W (2004) Experimental Studies on Sexual Reproduction in Diatoms. In: *International Review of Cytology*, pp. 91–154.
- Chonova T, Keck F, Labanowski J *et al.* (2016) Separate treatment of hospital and urban wastewaters: A real scale comparison of effluents and their effect on microbial communities. *Science of The Total Environment*, **542**, 965–975.
- Christensen MR, Graham MD, Vinebrooke RD *et al.* (2006) Multiple anthropogenic stressors cause ecological surprises in boreal lakes. *Global Change Biology*, **12**, 2316–2322.
- Chung C-C, Hwang S-PL, Chang J (2005) Cooccurrence of ScDSP Gene Expression, Cell Death, and DNA Fragmentation in a Marine Diatom, *Skeletonema costatum*. *Applied and Environmental Microbiology*, **71**, 8744–8751.
- Civade R, Dejean T, Valentini A *et al.* (2016) Spatial Representativeness of Environmental DNA Metabarcoding Signal for Fish Biodiversity Assessment in a Natural Freshwater System (C Garcia de Leaniz, Ed.). *Plos One*, **11**, e0157366.
- Clare EL (2014) Molecular detection of trophic interactions: emerging trends, distinct advantages, significant considerations and conservation applications. *Evolutionary Applications*, **7**, 1144–1157.
- Clavel J, Escarguel G, Merceron G (2015) mvmorph: an r package for fitting multivariate evolutionary models to morphometric data. *Methods in Ecology and Evolution*, **6**, 1311–1319.
- Clements WH, Rohr JR (2009) Community responses to contaminants: using basic ecological principles to predict ecotoxicological effects. *Environmental toxicology and chemistry*, **28**, 1789–800.
- Coissac E, Riaz T, Puillandre N (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology*, **21**, 1834–1847.
- Coleman AW (1985) Diversity of plastid DNA configuration among classes of eukaryote algae. *Journal of Phycology*, **21**, 1–16.
- Collen B, Whitton F, Dyer EE *et al.* (2014) Global patterns of freshwater species diversity, threat and endemism. *Global Ecology and Biogeography*, **23**, 40–51.
- Cordier T, Esling P, Lejzerowicz F *et al.* (2017) Predicting the Ecological Quality Status of Marine Environments from eDNA Metabarcoding Data Using Supervised Machine Learning. *Environmental Science & Technology*, **51**, 9118–9126.
- Coste M (1982) Étude des méthodes biologiques d’appréciation quantitative de la qualité des eaux. Cemagref.

- Coste M (1986) *Les méthodes microfloristiques d'évaluation de la qualité des eaux*. Cemagref, Bordeaux.
- Coste M, Boutry S, Tison-Rosebery J, Delmas F (2009) Improvements of the Biological Diatom Index (BDI): Description and efficiency of the new version (BDI-2006). *Ecological Indicators*, **9**, 621–650.
- Coste M, Leynaud G (1974) Etudes sur la mise au point d'une méthode biologique de détermination de la qualité des eaux en milieu fluvial. *Rapport C.T.G.R.E.F. et A.F.B.S.N., Paris*, 79 p.
- Cowart DA, Pinheiro M, Mouchel O *et al.* (2015) Metabarcoding Is Powerful yet Still Blind: A Comparative Analysis of Morphological and Molecular Surveys of Seagrass Communities (S Mazzuca, Ed.). *Plos One*, **10**, e0117562.
- Cox EJ (2009) What's in a name? Diatom classification should reflect systematic relationships. *Acta Botanica Croatica*, **68**, 443–454.
- Cox EJ (2014) Diatom identification in the face of changing species concepts and evidence of phenotypic plasticity. *Journal of Micropalaeontology*, **33**, 111–120.
- Crawford RM (1981) The Siliceous Components of the Diatom Cell Wall and Their Morphological Variation. In: *Silicon and Siliceous Structures in Biological Systems*, pp. 129–156. Springer New York, New York, NY.
- Creer S (2010) Second-generation sequencing derived insights into the temporal biodiversity dynamics of freshwater protists. *Molecular Ecology*, **19**, 2829–2831.
- Cressie NAC (1993) *Statistics for Spatial Data*. Wiley, New York.
- Crooks JA (2002) Characterizing ecosystem-level consequences of biological invasions: the role of ecosystem engineers. *Oikos*, **97**, 153–166.
- Dafforn KA, Johnston EL, Ferguson A *et al.* (2016) Big data opportunities and challenges for assessing multiple stressors across scales in aquatic ecosystems. *Marine and Freshwater Research*, **67**, 393.
- Daremborg C (1855) *Oeuvres choisies d'Hippocrate* (1855 Labé, Ed.).
- von Dassow P, Petersen TW, Chepurinov VA, Virginia Armbrust E (2008) Inter- and Intraspecific relationships between nuclear DNA content and cell size in selected members members of the centric diatom genus *Thalassiosira* (Bacillariophyceae). *Journal of Phycology*, **44**, 335–349.
- Deagle BE, Thomas AC, Shaffer AK, Trites AW, Jarman SN (2013) Quantifying sequence proportions in a DNA-based diet study using Ion Torrent amplicon sequencing: which counts count? *Molecular Ecology Resources*, **13**, 620–633.
- Debroas D, Hugoni M, Domaizon I (2015) Evidence for an active rare biosphere within freshwater protists community. *Molecular Ecology*, **24**, 1236–1247.
- Degnan PH, Ochman H (2012) Illumina-based analysis of microbial community diversity. *The ISME journal*, **6**, 183–94.
- Deiner K, Walser J-C, Mächler E, Altermatt F (2015) Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental DNA. *Biological Conservation*, **183**, 53–63.
- Dejean T, Valentini A, Duparc A *et al.* (2011) Persistence of environmental DNA in freshwater ecosystems. *Plos One*, **6**, e23398.
- Dejean T, Valentini A, Miquel C *et al.* (2012) Improved detection of an alien invasive species through environmental DNA barcoding: the example of the American bullfrog *Lithobates catesbeianus*. *Journal of Applied Ecology*, **49**, 953–959.
- DeLorenzo ME, Scott GI, Ross PE (2001) Toxicity of pesticides to aquatic microorganisms: A review. *Environmental Toxicology and Chemistry*, **20**, 84–98.
- Dhaliwal A (2013) DNA Extraction and Purification. *Material and Methods*, **3**.

- van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C (2014) Ten years of next-generation sequencing technology. *Trends in Genetics*, **30**, 418–426.
- Dodds WK, Perkin JS, Gerken JE (2013) Human Impact on Freshwater Ecosystem Services: A Global Perspective. *Environmental Science & Technology*, **47**, 9061–9068.
- Dodds W, Smith V (2016) Nitrogen, phosphorus, and eutrophication in streams. *Inland Waters*, **6**, 155–164.
- Dowle EJ, Pochon X, C. Banks J, Shearer K, Wood SA (2016) Targeted gene enrichment and high-throughput sequencing for environmental biomonitoring: a case study using freshwater macroinvertebrates. *Molecular Ecology Resources*, **16**, 1240–1254.
- Dudgeon D, Arthington AH, Gessner MO *et al.* (2006) Freshwater biodiversity: importance, threats, status and conservation challenges. *Biological reviews of the Cambridge Philosophical Society*, **81**, 163–82.
- Eberhard S, Drapier D, Wollman F-A (2002) Searching limiting steps in the expression of chloroplast-encoded proteins: relations between gene copy number, transcription, transcript abundance and translation rate in the chloroplast of *Chlamydomonas reinhardtii*. *The Plant Journal*, **31**, 149–160.
- Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797.
- Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, **27**, 2194–2200.
- Edgar L a., Pickett-heaps JD (1983) The mechanism of diatom locomotion. *Proceedings of the Royal Society B: Biological Sciences*, **218**, 331–343.
- Edlund MB, Stoermer EF (1997) Ecological, evolutionary, and systematic significance of diatom life histories. *Journal of Phycology*, **33**, 897–918.
- Egge E, Bittner L, Andersen T *et al.* (2013) 454 pyrosequencing to describe microbial eukaryotic community composition, diversity and relative abundance: a test for marine haptophytes. *PloS one*, **8**, e74371.
- Eland LE, Davenport R, Mota CR (2012) Evaluation of DNA extraction methods for freshwater eukaryotic microalgae. *Water Research*, **46**, 5355–5364.
- Elbrecht V, Leese F (2015) Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer Bias and Biomass—Sequence Relationships with an Innovative Metabarcoding Protocol (M Hajibabaei, Ed.). *Plos One*, **10**, e0130324.
- Elbrecht V, Leese F (2016) PrimerMiner : an R package for development and in silico validation of DNA metabarcoding primers (M Bunce, Ed.). *Methods in Ecology and Evolution*, **8**, 622–626.
- Elbrecht V, Leese F (2017) Validation and Development of COI Metabarcoding Primers for Freshwater Macroinvertebrate Bioassessment. *Frontiers in Environmental Science*, **5**, 1–11.
- Elbrecht V, Peinert B, Leese F (2017a) Sorting things out: Assessing effects of unequal specimen biomass on DNA metabarcoding. *Ecology and Evolution*, **7**, 6918–6926.
- Elbrecht V, Vamos EE, Meissner K, Aroviita J, Leese F (2017b) Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring (D Yu, Ed.). *Methods in Ecology and Evolution*, **(in press)**.
- Ersland DR, Aldrich J, Cattolico R a (1981) Kinetic Complexity, Homogeneity, and Copy Number of Chloroplast DNA from the Marine Alga *Olisthodiscus luteus*. *Plant Physiology*, **68**, 1468–1473.
- Esling P, Lejzerowicz F, Pawlowski J (2015) Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic acids research*, **43**, 2513–24.
- European Committee for Standardization (CEN) (2006) EN 15204 - Water quality - Guidance

- standard on the enumeration of phytoplankton using inverted microscopy (Utermöhl technique). *European Standard*, 1–42.
- European Council (2000) Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 Establishing a Framework for Community Action in the Field of Water Policy. *Office for official publications of the European Communities, Brussels*.
- European Environment Agency (2012) *European Waters—Assessment of Status and Pressures (EEA Report No 8/2012; EEA: Copenhagen, Denmark)*.
- Evans NT, Olds BP, Renshaw MA *et al.* (2016) Quantification of mesocosm fish and amphibian species diversity via environmental DNA metabarcoding. *Molecular Ecology Resources*, **16**, 29–41.
- Evans KM, Wortley AH, Mann DG (2007) An Assessment of Potential Diatom “Barcode” Genes (cox1, rbcL, 18S and ITS rDNA) and their Effectiveness in Determining Relationships in Sellaphora (Bacillariophyta). *Protist*, **158**, 349–364.
- Feio MJ, Almeida SFP, Craveiro SC, Calado AJ (2009) A comparison between biotic indices and predictive models in stream water quality assessment based on benthic diatom communities. *Ecological Indicators*, **9**, 497–507.
- Felsenstein J (1985) Phylogenies and the Comparative Method. *The American Naturalist*, **125**, 1–15.
- Feuillette S (2004) L’eau en France : entre subsidiarité et gestion spatiale. *Cybergeo*.
- Ficetola G, Coissac E, Zundel S *et al.* (2010) An In silico approach for the evaluation of DNA barcodes. *BMC Genomics*, **11**, 434.
- Field CB, Behrenfeld MJ, Randerson JT, Falkowsky P (1998) Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science*, **281**, 237–240.
- Fisher J, Dunbar MJJ (2007) Towards a representative periphytic diatom sample. *Hydrology and Earth System Sciences*, **11**, 399–407.
- Fjerdingstad E (1950) The microflora of the river Molleaa with special reference to the relation of benthic algae to pollution. *Folia Limnologica Scandinavica*, **5**, 1–123.
- Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, **3**, 294–299.
- Fonseca NA, Rung J, Brazma A, Marioni JC (2012) Tools for mapping high-throughput sequencing data. *Bioinformatics*, **28**, 3169–3177.
- Fox EJ, Reid-Bayliss KS, Emond MJ, Loeb LA (2014) Accuracy of Next Generation Sequencing Platforms. *Next generation, sequencing & applications*, **1**, 1–9.
- Frézal L, Leblois R (2008) Four years of DNA barcoding: Current advances and prospects. *Infection, Genetics and Evolution*, **8**, 727–736.
- Friberg N, Bonada N, Bradley DC *et al.* (2011) Biomonitoring of Human Impacts in Freshwater Ecosystems. In: *Advances in Ecological Research*, pp. 1–68.
- Frontier S, Pichod-Viale D, Leprêtre A, Davoult D, Luczak C (2008) *Ecosystèmes. Structure, fonctionnement, évolution*.
- Fuhrman JA, Comeau DE, Hagström A, Chan AM (1988) Extraction from natural planktonic microorganisms of DNA suitable for molecular biological studies. *Applied and environmental microbiology*, **54**, 1426–9.
- Gale KR, Crampton JM (1987) DNA probes for species identification of mosquitoes in the *Anopheles gambiae* complex. *Medical and Veterinary Entomology*, **1**, 127–136.
- Gallup JM, Ackermann MR (2006) Addressing fluorogenic real-time qPCR inhibition using the novel custom excel file system “FocusField2-6GallupqPCRSet-upTool-001” to attain consistently high fidelity qPCR reactions. *Biological Procedures Online*, **8**, 87–153.
- García-Berthou E, Alcaraz C, Pou-Rovira Q *et al.* (2005) Introduction pathways and

- establishment rates of invasive aquatic species in Europe. *Canadian Journal of Fisheries and Aquatic Sciences*, **62**, 453–463.
- Gassiole G (2014) Diatomées epilithiques des cours d'eau pérennes de l'île de la Réunion: taxinomie - écologie (PhD thesis).
- Geyer HJ, Rimkus GG, Scheunert I *et al.* (2000) Bioaccumulation and Occurrence of Endocrine-Disrupting Chemicals (EDCs), Persistent Organic Pollutants (POPs), and Other Organic Compounds in Fish and Other Organisms Including Humans. In: *Bioaccumulation -- New Aspects and Developments* (ed Beek B), pp. 1–166. Springer Berlin Heidelberg.
- Gibbs SP (1981) The chloroplasts of some algal groups may have evolved from endosymbiotic eukaryotic algae. *Annals of the New York Academy of Sciences*, **361**, 193–208.
- Gibson KE, Schwab KJ, Spencer SK, Borchardt MA (2012) Measuring and mitigating inhibition during quantitative real time PCR analysis of viral nucleic acid extracts from large-volume environmental water samples. *Water Research*, **46**, 4281–4291.
- Gibson JF, Shokralla S, Curry C *et al.* (2015) Large-Scale Biomonitoring of Remote and Threatened Ecosystems via High-Throughput Sequencing. *Plos One*, **10**, e0138432.
- Gifford SM, Sharma S, Rinta-Kanto JM, Moran MA (2011) Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. *The ISME journal*, **5**, 461–472.
- Gill RJ, Baldock KCR, Brown MJF *et al.* (2016) Protecting an Ecosystem Service. In: *Advances in Ecological Research*, pp. 135–206. Elsevier Ltd.
- Gillett ND, Pan Y, Manoylov KM, Stevenson RJ (2011) The role of live diatoms in bioassessment: a large-scale study of Western US streams. *Hydrobiologia*, **665**, 79–92.
- Gillett N, Pan Y, Parker C, Pan ÆY, Parker C (2009) Should only live diatoms be used in the bioassessment of small mountain streams? *Hydrobiologia*, **620**, 135–147.
- Girard S (2012) La territorialisation de la politique de l'eau est-elle gage d'efficacité environnementale ? : Analyse diachronique de dispositifs de gestion des eaux dans la vallée de la Drôme (1970-2011). Géographie. Ecole normale supérieure de lyon - ENS LYON.
- Godhe A, Asplund ME, Härnström K *et al.* (2008) Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples by real-time PCR. *Applied and environmental microbiology*, **74**, 7174–82.
- Goldberg CS, Turner CR, Deiner K *et al.* (2016) Critical considerations for the application of environmental DNA methods to detect aquatic species (M Gilbert, Ed.). *Methods in Ecology and Evolution*, **7**, 1299–1307.
- Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, **17**, 333–351.
- Goolsby EW, Bruggeman J, Ané C (2017) Rphylopars: fast multivariate phylogenetic comparative methods for missing data and within-species variation. *Methods in Ecology and Evolution*, **8**, 22–27.
- Gosselain V, Coste M, Campeau S *et al.* (2005) A large-scale stream benthic diatom database. *Hydrobiologia*, **542**, 151–163.
- Green BR (2011) Chloroplast genomes of photosynthetic eukaryotes. *Plant Journal*, **66**, 34–44.
- Grove RH (1995) *Green Imperialism: Colonial Expansion, Tropical Island Edens and the Origins of Environmentalism*. Cambridge University Press, New York, NY.
- Grove RH, Damodaran V (2006) Imperialism, Intellectual Networks, and Environmental Change Origins and Evolution of Global Environmental History, 1676-2000: Part I. *Economic and political weekly*, **41**, 4345–4354.
- Gruber A (2008) Molecular Characterisation of Diatom Plastids (PhD thesis). University of Konstanz.
- Grunenwald A (2014) Etude de l'interaction entre ADN et apatites analogues au minéral

- osseux et dentaire – Implications pour la préservation de l’ADN ancien, son extraction, son analyse (PhD thesis). Université Paul Sabatier, Toulouse.
- Gueguen J, Eulin A, Lefrançois E *et al.* (2015) *Production of an Improved Version of the Indice Diatomique Antilles (IDA-2), Use for the Evaluation of the Ecological Status of Rivers in the French Caribbean: Final Version, 2015-03-12. IRSTEA Scientific Report. Pages 185. <http://cemadoc.irstea.fr/cemoa>.*
- Guiry MD (2012) How many species of algae are there ? *Journal of Phycology*, **48**, 1057–1063.
- Guo L, Sui Z, Zhang S, Ren Y, Liu Y (2015) Comparison of potential diatom “barcode” genes (the 18S rRNA gene and ITS, COI, rbcL) and their effectiveness in discriminating and determining species taxonomy in the Bacillariophyta. *International journal of systematic and evolutionary microbiology*, **65**, 1369–80.
- Gutteridge S, Gatenby A (1995) Rubisco Synthesis, Assembly, Mechanism, and Regulation. *The Plant Cell Online*, **7**, 809–819.
- Hajibabaei M, Shokralla S, Zhou X, Singer G a C, Baird DJ (2011) Environmental Barcoding: A Next-Generation Sequencing Approach for Biomonitoring Applications Using River Benthos (CR Voolstra, Ed.). *Plos One*, **6**, e17497.
- Hajibabaei M, Singer GAC, Hebert PDN, Hickey DA (2007) DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends in genetics : TIG*, **23**, 167–72.
- Hajibabaei M, Spall JL, Shokralla S, van Konynenburg S (2012) Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. *BMC Ecology*, **12**, 28.
- Halpern BS, Walbridge S, Selkoe KA *et al.* (2008) A Global Map of Human Impact on Marine Ecosystems. *Science*, **319**, 948–952.
- Hamilton PB, Lefebvre KE, Bull RD (2015) Single cell PCR amplification of diatoms using fresh and preserved samples. *Frontiers in Microbiology*, **6**, 1084.
- Hamm CE, Merkel R, Springer O *et al.* (2003) Architecture and material properties of diatom shells provide effective mechanical protection. *Nature*, **421**, 841–843.
- Hamsher SE, Evans KM, Mann DG, Poulíčková A, Saunders GW (2011) Barcoding Diatoms: Exploring Alternatives to COI-5P. *Protist*, **162**, 405–422.
- Hawksworth DL, Kalin-Arroyo MT (1995) Magnitude and distribution of biodiversity. In: *Global biodiversity assessment*, pp. 107–191. Cambridge University Press.
- Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings. Biological sciences / The Royal Society*, **270**, 313–21.
- Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004) Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proceedings of the National Academy of Sciences*, **101**, 14812–14817.
- Hense I, Beckmann A (2015) A theoretical investigation of the diatom cell size reduction–restitution cycle. *Ecological Modelling*, **317**, 66–82.
- Hering D, Borja A, Carstensen J *et al.* (2010) The European Water Framework Directive at the age of 10: A critical review of the achievements with recommendations for the future. *Science of The Total Environment*, **408**, 4007–4019.
- Hering D, Carvalho L, Argillier C *et al.* (2015) Managing aquatic ecosystems and water resources under multiple stress — An introduction to the MARS project. *Science of The Total Environment*, **503–504**, 10–21.
- Hillebrand H, Sommer U (2000) Diversity of benthic microalgae in response to colonization time and eutrophication. *Aquatic Botany*, **67**, 221–236.
- Ho T, Si L, Ané C (2014) A Linear-Time Algorithm for Gaussian and Non-Gaussian Trait Evolution Models. *Systematic Biology*, **63**, 397–408.

- Hoffman G, Werum M, Lange-Bertalot H (2011) *Diatomeen im Süßwasser-benthos von Mitteleuropa*. A.R.G.
- Hong S, Bunge J, Leslin C, Jeon S, Epstein SS (2009) Polymerase chain reaction primers miss half of rRNA microbial diversity. *The ISME Journal*, **3**, 1365–1373.
- Hulme PE (2009) Trade, transport and trouble: managing invasive species pathways in an era of globalization. *Journal of Applied Ecology*, **46**, 10–18.
- Hunting ER, De Jong S, Vijver MG (2017) *Assessment of monitoring tools and strategies safeguarding aquatic ecosystems within the European water framework directive*.
- Hürlimann J, Niederhauser P (2007) *Méthodes d'Analyse et d'Appréciation des Cours d'Eau. Diatomées Niveau R (région); Etat de l'environnement no 0740*. Office Fédéral de l'Environnement, Berne 132p.
- Ibáñez C, Caiola N, Sharpe P, Trobajo R (2010) Ecological Indicators to Assess the Health of River Ecosystems. In: pp. 447–464.
- Irannia ZB, Chen T (2016) TACO: Taxonomic prediction of unknown OTUs through OTU co-abundance networks. *Quantitative Biology*, **4**, 149–158.
- Isaaks EH, Srivastava RM (1989) *An Introduction to Applied Geostatistics*. Oxford University Press, New York.
- Izsak C, Price ARG, Hardy JT, Basson PW (2002) Biodiversity of periphyton (diatoms) and echinoderms around a refinery effluent, and possible associations with stability. *Aquatic Ecosystem Health & Management*, **5**, 233–242.
- Jackson MC, Loewen CJG, Vinebrooke RD, Chimimba CT (2016) Net effects of multiple stressors in freshwater ecosystems: a meta-analysis. *Global Change Biology*, **22**, 180–189.
- Jahn R, Kusber W-H (2009) A key to diatom nomenclature. *Diatom Research*, **24**, 101–111.
- Jeon S, Bunge J, Leslin C *et al.* (2008) Environmental rRNA inventories miss over half of protistan diversity. *BMC Microbiology*, **8**, 222.
- Ji Y, Ashton L, Pedley SM *et al.* (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding (M Holyoak, Ed.). *Ecology Letters*, **16**, 1245–1257.
- Kahlert M, Albert R-L, Anttila E-L *et al.* (2009) Harmonization is more important than experience—results of the first Nordic–Baltic diatom intercalibration exercise 2007 (stream monitoring). *Journal of Applied Phycology*, **21**, 471–482.
- Kahlert M, Kelly M, Albert R-L *et al.* (2012) Identification versus counting protocols as sources of uncertainty in diatom-based ecological status assessments. *Hydrobiologia*, **695**, 109–124.
- Kalff J, Knoechel R (1978) Phytoplankton and their Dynamics in Oligotrophic and Eutrophic Lakes. *Annual Review of Ecology and Systematics*, **9**, 475–495.
- Kallis G (2001) The EU water framework directive: measures and implications. *Water Policy*, **3**, 125–142.
- Kammerlander B, Breiner H-W, Filker S *et al.* (2015) High diversity of protistan plankton communities in remote high mountain lakes in the European Alps and the Himalayan mountains. *FEMS microbiology ecology*, **91**, 1–10.
- Karsten G (1928) Bacillariophyta (Diatomaceae). In: *In: Die Natürlichen Pflanzenfamilien, 2nd ed.*, pp. 105–203.
- Kebschull JM, Zador AM (2015) Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic acids research*, **43**, e143.
- Keck F, Rimet F, Franc A, Bouchez A (2015) Phylogenetic signal in diatom ecology: perspectives for aquatic ecosystems biomonitoring. *Ecological Applications*, 14–1966.1.
- Keck F, Vasselon V, Tapolczai K, Rimet F, Bouchez A (2017) Freshwater biomonitoring in the Information Age. *Frontiers in Ecology and the Environment*, **15**, 266–274.

- Keeling PJ (2013) The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. *Annual review of plant biology*, **64**, 583–607.
- Kelly M (2013) Data rich, information poor? Phytobenthos assessment and the Water Framework Directive. *European Journal of Phycology*, **48**, 437–450.
- Kelly M, Bennion H, Burgess A *et al.* (2009) Uncertainty in ecological status assessments of lakes and rivers using diatoms. *Hydrobiologia*, **633**, 5–15.
- Kelly M, Juggins S, Guthrie R *et al.* (2007) Assessment of ecological status in U.K. rivers using diatoms. *Freshwater Biology*, **53**, 403–422.
- Kelly MG, Penny CJ, Whitton BA (1995) Comparative performance of benthic diatom indices used to assess river water quality. *Hydrobiologia*, **302**, 179–188.
- Kelly MG, Schneider SC, King L (2015) Customs, habits, and traditions: the role of nonscientific factors in the development of ecological assessment methods. *Wiley Interdisciplinary Reviews: Water*, **2**, 159–165.
- Kelly M, Urbanic G, Acs E *et al.* (2014) Comparing aspirations: intercalibration of ecological status concepts across European lakes for littoral diatoms. *Hydrobiologia*, **734**, 125–141.
- Kelly MG, Whitton BA (1995) The Trophic Diatom Index: a new index for monitoring eutrophication in rivers. *Journal of Applied Phycology*, **7**, 433–444.
- Kembel SW, Wu M, Eisen JA, Green JL (2012) Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS computational biology*, **8**, e1002743.
- Kermarrec L (2012) Apport des outils de la biologie moléculaire pour l'utilisation des diatomées comme bioindicateurs de la qualité des écosystèmes aquatiques lotiques et pour l'étude de leur taxonomie (PhD thesis). Grenoble.
- Kermarrec L, Bouchez A, Rimet F, Humbert J-F (2013a) First evidence of the existence of semi-cryptic species and of a phylogeographic structure in the *Gomphonema parvulum* (Kützing) Kützing complex (Bacillariophyta). *Protist*, **164**, 686–705.
- Kermarrec L, Franc A, Rimet F *et al.* (2013b) Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: a test for freshwater diatoms. *Molecular Ecology Resources*, **13**, 607–619.
- Kermarrec L, Franc A, Rimet F *et al.* (2014) A next-generation sequencing approach to river biomonitoring using benthic diatoms. *Freshwater Science*, **33**, 349–363.
- Khan-Bureau DA, Morales EA, Ector L, Beauchene MS, Lewis LA (2016) Characterization of a new species in the genus *Didymosphenia* and of *Cymbella janischii* (Bacillariophyta) from Connecticut, USA. *European Journal of Phycology*, **262**, 1–14.
- Kikyo M, Tanaka K, Kamei T *et al.* (1999) An FH domain-containing Bnr1p is a multifunctional protein interacting with a variety of cytoskeletal proteins in *Saccharomyces cerevisiae*. *Oncogene*, **18**, 7046–54.
- Kocher TD, Thomas WK, Meyer A *et al.* (1989) Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. *Proceedings of the National Academy of Sciences*, **86**, 6196–6200.
- Kociolek JP, Stoermer EF (2010) Variation and polymorphism in diatoms: The triple helix of development, genetics and environment. A review of the literature. *Vie et Milieu*, **60**, 75–87.
- Kociolek JP, Williams DM (2015) How to define a diatom genus? Notes on the creation and recognition of taxa, and a call for revisionary studies of diatoms. *Acta Botanica Croatica*, **74**, 195–210.
- Koid A, Nelson WC, Mraz A, Heidelberg KB (2012) Comparative Analysis of Eukaryotic Marine Microbial Assemblages from 18S rRNA Gene and Gene Transcript Clone Libraries by Using Different Methods of Extraction. *Applied and Environmental Microbiology*, **78**, 3958–

3965.

- Kolkwitz R, Marsson M (1908) 59. R. Kolkwitz und M. Marsson: Ökologie der pflanzlichen Saprobien. *Berichte der Deutschen Botanischen Gesellschaft*, **26**, 505–519.
- Koop H-U, Herz S, Golds TJ, Nickelsen J (2007) The genetic transformation of plastids. In: *Stress-Activated Protein Kinases*, pp. 457–510.
- Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD (2013) Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Applied and environmental microbiology*, **79**, 5112–20.
- Krammer K (2000) *The Genus Pinnularia*.
- Krammer K (2001) *Navicula Sensu Stricto, 10 Genera Separated from Navicula Sensu Stricto, Frustulia*.
- Krammer K (2002) *Cymbella*.
- Krammer K (2003) *Cymbopleura, Delicata, Navicymbula, Gomphocymbellopsis, Afrocybella*.
- Krammer K, Lange-Bertalot H (1986) Bacillariophyceae 1. Teil: Naviculaceae. Süßwasserflora von Mitteleuropa. In: *Süßwasserflora von Mitteleuropa*, p. 876 pages.
- Krammer K, Lange-Bertalot H (1988) Bacillariophyceae 2. Teil: Bacillariaceae, Epithemiaceae, Surirellaceae. In: *Süßwasserflora von Mitteleuropa*, p. 610.
- Krammer K, Lange-Bertalot H (1991a) Bacillariophyceae 3. Teil: Centrales, Fragilariaceae, Eunotiaceae. Süßwasserflora von Mitteleuropa. In: *Süßwasserflora von Mitteleuropa*, p. 598.
- Krammer K, Lange-Bertalot H (1991b) Bacillariophyceae 4. Teil: Achnanthaceae. Kritische Ergänzungen zu Navicula (Lineolatae) und Gomphonema. Gesamtliteraturverzeichnis Teil 4. Süßwasserflora von Mitteleuropa. In: *Süßwasserflora von Mitteleuropa*, p. 437 pages.
- Kress WJ, Erickson DL (2012) *DNA Barcodes: methods and protocols*.
- Kröger N, Poulsen N (2008) Diatoms—From Cell Wall Biogenesis to Nanotechnology. *Annual Review of Genetics*, **42**, 83–107.
- Laehnmann D, Borkhardt A, McHardy AC (2016) Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. *Briefings in bioinformatics*, **17**, 154–79.
- Lang I, Kaczmarska I (2011) A protocol for a single-cell PCR of diatoms from fixed samples: method validation using *Ditylum brightwellii* (T. West) Grunow. *Diatom Research*, **26**, 43–49.
- Larras F, Coulaud R, Gautreau E *et al.* (2017) Assessing anthropogenic pressures on streams: A random forest approach based on benthic diatom communities. *Science of The Total Environment*, **586**, 1101–1112.
- Larras F, Montuelle B, Rimet F, Chèvre N, Bouchez A (2014) Seasonal shift in the sensitivity of a natural benthic microalgal community to a herbicide mixture: impact on the protective level of thresholds derived from species sensitivity distributions. *Ecotoxicology*, **23**, 1109–1123.
- Lavoie I, Campeau S, Fallu M-A, Dillon PJ (2006) Diatoms and biomonitoring: should cell size be accounted for? *Hydrobiologia*, **573**, 1–16.
- Lavoie I, Dillon PJ, Campeau S (2009) The effect of excluding diatom taxa and reducing taxonomic resolution on multivariate analyses and stream bioassessment. *Ecological Indicators*, **9**, 213–225.
- Lavoie I, Somers KM, Paterson AM, Dillon PJ (2005) Assessing scales of variability in benthic diatom community structure. *Journal of Applied Phycology*, **17**, 509–513.
- Lazarevic V, Gaïa N, Girard M, François P, Schrenzel J (2013) Comparison of DNA Extraction Methods in Analysis of Salivary Bacterial Communities. *Plos One*, **8**, e67699.

- Lecoite C, Coste M, Prygiel J (1993) "Omnidia": software for taxonomy, calculation of diatom indices and inventories management. *Hydrobiologia*, **269–270**, 509–513.
- Lee ZM-P, Bussema C, Schmidt TM (2009) rrnDB: documenting the number of rRNA and tRNA genes in bacteria and archaea. *Nucleic acids research*, **37**, D489–93.
- Leese F, Altermatt F, Bouchez A *et al.* (2016) DNAqua-Net: Developing new genetic tools for bioassessment and monitoring of aquatic ecosystems in Europe. *Research Ideas and Outcomes*, **2**, e11321.
- Lehner B, Liermann CR, Revenga C *et al.* (2011) High-resolution mapping of the world's reservoirs and dams for sustainable river-flow management. *Frontiers in Ecology and the Environment*, **9**, 494–502.
- Lejzerowicz F, Esling P, Pillet L *et al.* (2015) High-throughput sequencing and morphology perform equally well for benthic monitoring of marine ecosystems. *Scientific Reports*, **5**, 13932.
- Lengyel E, Padisák J, Stenger-Kovács C (2015) Establishment of equilibrium states and effect of disturbances on benthic diatom assemblages of the Torna-stream, Hungary. *Hydrobiologia*, **750**, 43–56.
- Lenoir A, Coste M (1996) Development of a practical diatom index of overall water quality applicable to the French National Water Board Network. In: *Use of Algae for Monitoring Rivers II. International symposium, Volksbildungsheim Grilhof Vill, AUT, 17-19 September 1995.*, pp. 29–43.
- Levkov Z (2009) *Amphora sensu lato* (H Lange-Bertalot, Ed.). Gantner Verlag.
- Lewis WM (1983) The diatom sex clock and its evolutionary significance. *The American Naturalist*, **123**, 73–80.
- Liu S, Wang X, Xie L *et al.* (2016) Mitochondrial capture enriches mito-DNA 100 fold, enabling PCR-free mitogenomics biodiversity analysis. *Molecular ecology resources*, **16**, 470–9.
- Lloyd KG, MacGregor BJ, Teske A (2010) Quantitative PCR methods for RNA and DNA in marine sediments: maximizing yield while overcoming inhibition. *FEMS Microbiology Ecology*, **72**, 143–151.
- Loman NJ, Misra R V, Dallman TJ *et al.* (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, **30**, 434–439.
- Lommer M, Specht M, Roy A-S *et al.* (2012) Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. *Genome biology*, **13**, R66.
- Lund JWG (1954) The seasonal cycle of the plankton diatom, *Melosira italica* (Erh.) Kutz. subsp. *Subarctica* O. Mull. *Journal of Ecology*, **42**, 151–179.
- Lundin D, Severin I, Logue JB *et al.* (2012) Which sequencing depth is sufficient to describe patterns in bacterial α - and β -diversity? *Environmental Microbiology Reports*, **4**, 367–372.
- Macdonald JD (1869) I.— On the structure of the Diatomaceous frustule, and its genetic cycle. *Journal of Natural History Series 4*, **3**, 1–8.
- MacFadyen EJ, Williamson CE, Grad G *et al.* (2004) Molecular response to climate change: temperature dependence of UV-induced DNA damage and repair in the freshwater crustacean *Daphnia pulex*. *Global Change Biology*, **10**, 408–416.
- Majaneva M, Hyytiäinen K, Varvio SL, Nagai S, Blomster J (2015) Bioinformatic Amplicon Read Processing Strategies Strongly Affect Eukaryotic Diversity and the Taxonomic Composition of Communities (G Berg, Ed.). *Plos One*, **10**, e0130035.
- Mangot J-F, Domaizon I, Taib N *et al.* (2013) Short-term dynamics of diversity patterns: evidence of continual reassembly within lacustrine small eukaryotes. *Environmental Microbiology*, **15**, 1745–1758.
- Mann DG (1988) Why didn't Lund see sex in *Asterionella*? A discussion of the diatom life cycle in nature. *Algae and the aquatic environment*, 385–412.

- Mann DG (1999) The species concept in diatoms. *Phycologia*, **38**, 437–495.
- Mann DG (2011) Size and Sex. In: *The Diatom World* (ed J Seckbach & JP Kociolek E), pp. 145–166. Springer, Dordrecht.
- Mann DG, Chepurnov VA, Droop SJM (1999) Sexuality, incompatibility, size variation, and preferential polyandry in natural populations and clones of *Sellaphora pupula* (Bacillariophyceae). *Journal of Phycology*, **35**, 152–170.
- Mann DG, Crawford RM, Round FE (2016) Bacillariophyta. In: *Handbook of the Protists* (eds Archibald JM, Simpson AGB, Slamovits CH, et al.), pp. 1–62. Springer International Publishing, Cham.
- Mann DG, Sato S, Trobajo R, Vanormelingen P, Souffreau C (2010) DNA barcoding for species identification and discovery in diatoms. *Cryptogamie*, **31**, 557–577.
- Mann DG, Vanormelingen P (2013) An Inordinate Fondness? The Number, Distributions, and Origins of Diatom Species. *Journal of Eukaryotic Microbiology*, **60**, 414–420.
- Manoylov K, France Y, Geletu A, Dominy J (2016) Algal Community Membership of Estuarine Mudflats from the Savannah River, United States. *Journal of Marine Science and Engineering*, **4**, 11.
- Margulies M, Egholm M, Altman WE *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
- De Martino A, Amato A, Bowler C (2009) Mitosis in diatoms: rediscovering an old model for cell division. *BioEssays: news and reviews in molecular, cellular and developmental biology*, **31**, 874–84.
- Marx V (2013) Biology: The big challenges of big data. *Nature*, **498**, 255–260.
- McGee CF, Storey S, Clipson N, Doyle E (2017) Soil microbial community responses to contamination with silver, aluminium oxide and silicon dioxide nanoparticles. *Ecotoxicology*, **26**, 449–458.
- McLachlan DH, Brownlee C, Taylor AR, Geider RJ, Underwood GJC (2009) Light-induced motile responses of the estuarine benthic diatoms *Navicula perminuta* and *Cylindrotheca closterium* (bacillariophyceae). *Journal of Phycology*, **45**, 592–599.
- Medlin LK (2010) Pursuit of a natural classification of diatoms: An incorrect comparison of published data. *European Journal of Phycology*, **45**, 155–166.
- Medlin LK, Kaczmarska I (2004) Evolution of the diatoms: V. Morphological and cytological support for the major clades and a taxonomic revision. *Phycologia*, **43**, 245–270.
- Mengin N, Bougon N, Chandesris A *et al.* (2010) *Réseau de référence des eaux douces de surface – cours d'eau (rapport final ONEMA-AFB/CEMAGREF)*.
- Metzeltin D, Lange-Bertalot H (1998) Tropical diatoms of South America I. *Iconographia Diatomologica*, **5**, 1–695.
- Metzeltin D, Lange-Bertalot H (2007) Tropical diatoms of South America II. *Iconographia Diatomologica*, **18**, 1–877.
- Miller DN, Bryant JE, Madsen EL, Ghiorse WC (1999) Evaluation and Optimization of DNA Extraction and Purification Procedures for Soil and Sediment Samples Evaluation and Optimization of DNA Extraction and Purification Procedures for Soil and Sediment Samples. *Applied and Environmental Microbiology*, **65**, 4715–4724.
- Ministère de l'Environnement de l'Énergie et de la Mer (2016) *Guide relatif à l'évaluation de l'état des eaux de surface continentales (cours d'eau, canaux, plans d'eau) - Français*.
- Moeys S, Frenkel J, Lembke C *et al.* (2016) A sex-inducing pheromone triggers cell cycle arrest and mate attraction in the diatom *Seminavis robusta*. *Scientific reports*, **6**, 19252.
- Monchamp M-E, Walser J-C, Pomati F, Spaak P (2016) Sedimentary DNA Reveals Cyanobacterial Community Diversity over 200 Years in Two Perialpine Lakes. *Applied and environmental microbiology*, **82**, 6472–6482.

- Moniz MBJ, Kaczmarska I (2009) Barcoding diatoms: Is there a good marker? *Molecular ecology resources*, **9**, 65–74.
- Morales EA, Siver P, Trainor F (2001) Identification of diatoms (Bacillariophyceae) during ecological assessments: Comparison between Light Microscopy and Scanning Electron Microscopy techniques. *Proceedings of the Academy of Natural Sciences of Philadelphia*, **151**, 95–103.
- Moreno MD, Ma K, Schoenung J, Dávila LP (2015) An integrated approach for probing the structure and mechanical properties of diatoms: Toward engineered nanotemplates. *Acta Biomaterialia*, **25**, 313–324.
- Motwani NH, Gorokhova E (2013) Mesozooplankton Grazing on Picocyanobacteria in the Baltic Sea as Inferred from Molecular Diet Analysis (E Sotka, Ed.). *Plos One*, **8**, e79230.
- Mullis K, Faloona F, Scharf S *et al.* (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor symposia on quantitative biology*, **51**, 263–73.
- Nakov T, Ashworth M, Theriot EC (2015) Comparative analysis of the interaction between habitat and growth form in diatoms. *The ISME Journal*, **9**, 246–255.
- Narcy J (2003) La politique de l'eau face à la gestion des espaces: les Agences de l'Eau aux limites de la modernité. *Espaces et sociétés*, **115**, 179.
- Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**, 443–453.
- Nguyen TNM, Berzano M, Gualerzi CO, Spurio R (2011) Development of molecular tools for the detection of freshwater diatoms. *Journal of microbiological methods*, **84**, 33–40.
- Nielsen R, Matz M (2006) Statistical Approaches for DNA Barcoding. *Systematic Biology*, **55**, 162–169.
- Nilsson C (2005) Fragmentation and Flow Regulation of the World's Large River Systems. *Science*, **308**, 405–408.
- Odelberg SJ, Weiss RB, Hata A, White R (1995) Template-switching during DNA synthesis by *Thermus aquaticus* DNA polymerase I. *Nucleic acids research*, **23**, 2049–57.
- Oki T, Kanae S (2006) Global hydrological cycles and world water resources. *Science*, **313**, 1068–72.
- Okie JG, Smith VH, Martin-Cereceda M (2016) Major evolutionary transitions of life, metabolic scaling and the number and size of mitochondria and chloroplasts. *Proceedings. Biological sciences*, **283**, 20160611.
- Ormerod SJ, Dobson M, Hildrew AG, Townsend CR (2010) Multiple stressors in freshwater ecosystems. *Freshwater Biology*, **55**, 1–4.
- Oulas A, Pavloudi C, Polymenakou P *et al.* (2015) Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinformatics and biology insights*, **9**, 75–88.
- Pääbo S, Irwin DM, Wilson AC (1990) DNA damage promotes jumping between templates during enzymatic amplification. *The Journal of Biological Chemistry*, **265**, 4718–4721.
- Padisák J, Borics G, Grigorszky I, Soróczki-Pintér É (2006) Use of Phytoplankton Assemblages for Monitoring Ecological Status of Lakes within the Water Framework Directive: The Assemblage Index. *Hydrobiologia*, **553**, 1–14.
- Pandey LK, Bergey EA, Lyu J *et al.* (2017) The use of diatoms in ecotoxicology and bioassessment: Insights, advances and challenges. *Water Research*, **118**, 39–58.
- Pascualt N, Roux S, Artigas J *et al.* (2014) A high-throughput sequencing ecotoxicology study of freshwater bacterial communities and their responses to tebuconazole. *FEMS microbiology ecology*, **90**, 563–74.
- Passy SI (2007a) Diatom ecological guilds display distinct and predictable behavior along

- nutrient and disturbance gradients in running waters. *Aquatic Botany*, **86**, 171–178.
- Passy SI (2007b) Differential cell size optimization strategies produce distinct diatom richness–body size relationships in stream benthos and plankton. *Journal of Ecology*, **95**, 745–754.
- Passy SI, Bode RW (2004) Diatom Model Affinity (DMA), a New Index for Water Quality Assessment. *Hydrobiologia*, **524**, 241–252.
- Pawlowski J, Lejzerowicz F, Apotheloz-Perret-Gentil L, Visco J, Esling P (2016) Protist metabarcoding and environmental biomonitoring: Time for change. *European Journal of Protistology*, **55**, 12–25.
- Petit K, Michon J (2013) La surveillance des milieux aquatiques et des eaux souterraines. *Les Synthèses de l'ONEMA*, **8**, 1–12.
- Petit K, Michon J (2015) L'état des eaux de surface et des eaux souterraines. *Les Synthèses de l'ONEMA*, **12**, 1–12.
- Pfitzer E (1869) Über den Bau und die Zellteilung der Diatomeen. *Botanische Zeitung Berlin*, **27**, 774–776.
- Pimentel D, Houser J, Preiss E *et al.* (1997) the Resources : Agriculture , and Society. *Society*, **47**, 97–106.
- Pinto AJ, Raskin L (2012) PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *Plos One*, **7**, e43093.
- Pochon X, Wood SA, Keeley NB *et al.* (2015) Accurate assessment of the impact of salmon farming on benthic sediment enrichment using foraminiferal metabarcoding. *Marine Pollution Bulletin*, **100**, 370–382.
- Poikane S, Zampoukas N, Borja A *et al.* (2014) Intercalibration of aquatic ecological assessment methods in the European Union: Lessons learned and way forward. *Environmental Science & Policy*, **44**, 237–246.
- Polz MF, Cavanaugh CM (1998) Bias in template-to-product ratios in multitemplate PCR. *Applied and environmental microbiology*, **64**, 3724–30.
- Pompanon F, Deagle BE, Symondson WOC *et al.* (2012) Who is eating what: diet assessment using next generation sequencing. *Molecular ecology*, **21**, 1931–50.
- Potapova M, Charles DF (2007) Diatom metrics for monitoring eutrophication in rivers of the United States. *Ecological Indicators*, **7**, 48–70.
- Prygiel J, Carpentier P, Almeida S *et al.* (2002) Determination of the biological diatom index (IBD NF T 90-354): Results of an intercomparison exercise. *Journal of Applied Phycology*, **14**, 27–39.
- Prygiel J, Coste M (1993) Utilisation des diatomeés benthiques pour la mesure de la qualité des eaux du bassin Artois-Picardie: bilan et perspectives. *Annales De Limnologie*, **29**, 255–267.
- Prygiel J, Coste M (2000) *Guide méthodologique pour la mise en oeuvre de l'Indice Biologique Diatomées (NF T 90-354)*.
- Prygiel J, Leveque L, Iserentant R (1996) Un nouvel indice Diatomique Pratique por l'évaluation de la qualité des eaux en réseau de surveillance. *Revue des Sciences de L'Eau*, **1**, 97–113.
- Pylro VS, Roesch LFW, Morais DK *et al.* (2014) Data analysis for 16S microbial profiling from different benchtop sequencing platforms. *Journal of microbiological methods*, **107**, 30–7.
- Quail M, Smith ME, Coupland P *et al.* (2012) A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics*, **13**, 341.
- R Development core team (2013) R: a language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*.
- Ratte HT (1999) Bioaccumulation and toxicity of silver compounds: A review. *Environmental*

- Toxicology and Chemistry*, **18**, 89–108.
- Rauwolf U, Golczyk H, Greiner S, Herrmann RG (2010) Variable amounts of DNA related to the size of chloroplasts III. Biochemical determinations of DNA amounts per organelle. *Molecular Genetics and Genomics*, **283**, 35–47.
- Relyea R, Hoverman J (2006) Assessing the ecology in ecotoxicology: a review and synthesis in freshwater systems. *Ecology letters*, **9**, 1157–71.
- Reyjol Y, Spyrtatos V, Basilico L (2013) *Bioindication : des outils pour évaluer l'état écologique des milieux aquatiques Perspectives en vue du 2e cycle DCE – Eaux de surface continentales*.
- Reynolds CS (1980) Phytoplankton assemblages and their periodicity in stratifying lake systems. *Ecography*, **3**, 141–159.
- Rimet F (2012) Recent views on river pollution and diatoms. *Hydrobiologia*, **683**, 1–24.
- Rimet F, Abarca N, Bouchez A *et al.* (2018) The potential of high throughput sequencing (HTS) of natural samples as a source of primary taxonomic information for reference libraries of diatom barcodes. *Fottea*, **(in press)**.
- Rimet F, Bouchez A (2012a) Biomonitoring river diatoms: Implications of taxonomic resolution. *Ecological Indicators*, **15**, 92–99.
- Rimet F, Bouchez A (2012b) Life-forms, cell-sizes and ecological guilds of diatoms in European rivers. *Knowledge and Management of Aquatic Ecosystems*, **1**.
- Rimet F, Chaumeil P, Keck F *et al.* (2016) R-Syst::diatom: an open-access and curated barcode database for diatoms and freshwater monitoring. *Database*, **2016**, baw016.
- Rimet F, Trobajo R, Mann DG *et al.* (2014) When is Sampling Complete? The Effects of Geographical Range and Marker Choice on Perceived Diversity in *Nitzschia palea* (Bacillariophyta). *Protist*, **165**, 245–259.
- Rimet F, Vasselon V, Chardon C *et al.* (2017) *Bioindication diatomées : comparaison microscopie / barcoding ADN - Premiers résultats - juillet 2017*.
- Rivera SF, Vasselon V, Jacquet S *et al.* (2017) Metabarcoding of lake benthic diatoms: from structure assemblages to ecological assessment. *Hydrobiologia*, **(in press)**.
- Rott E, Pipp E, Pfister P (2003) Diatom methods developed for river quality assessment in Austria and a cross-check against numerical trophic indication methods used in Europe. *Algological Studies*, **110**, 91–115.
- Round FE (1998) A problem in algal ecology: contamination of habitats from adjacent communities. *A problem in algal ecology: contamination of habitats from adjacent communities*, **19**, 49–55.
- Round FE, Crawford RM, Mann DG (1990a) *The diatoms: biology and morphology of the genera*. Cambridge University Press, Cambridge, UK.
- Round FE, Crawford RM, Mann DG (1990b) *Diatoms: Biology and Morphology of the Genera* (Cambridge University Press, Ed.).
- Rovira L, Trobajo R, Sato S, Ibáñez C, Mann DG (2015) Genetic and Physiological Diversity in the Diatom *Nitzschia inconspicua*. *Journal of Eukaryotic Microbiology*, **62**, 815–832.
- Rubin BER, Sanders JG, Hampton-Marcell J *et al.* (2014) DNA extraction protocols cause differences in 16S rRNA amplicon sequencing efficiency but not in community profile composition or structure. *MicrobiologyOpen*, **3**, 910–921.
- Ruck EC, Nakov T, Jansen RK, Theriot EC, Alverson AJ (2014) Serial gene losses and foreign DNA underlie size and sequence variation in the plastid genomes of diatoms. *Genome biology and evolution*, **6**, 644–54.
- Rumeau A, Coste M (1988) Initiation à la systématique des diatomées d'eau douce pour l'utilisation pratique d'un indice diatomique générique. *Bulletin français de la Pêche et de la pisciculture*, **309**, 1–69.

- Sabir JSM, Yu M, Ashworth MP *et al.* (2014) Conserved gene order and expanded inverted repeats characterize plastid genomes of Thalassiosirales. *Plos One*, **9**, e107854.
- Saiki R, Scharf S, Faloona F *et al.* (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, **230**, 1350–1354.
- Salipante SJ, Kawashima T, Rosenthal C *et al.* (2014) Performance comparison of Illumina and ion torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. *Applied and environmental microbiology*, **80**, 7583–91.
- Sandgren CD (1988) *Growth and Reproductive Strategies of Freshwater Phytoplankton*.
- Sanger F, Nicklen S, Coulson R (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, **74**, 5463–5467.
- Sarat E, Mazaubert E, Dutartre A, Poulet N, Soubeyran Y (2015) *Les espèces exotiques envahissantes. Connaissances pratiques et expériences de gestion. Volume 1 - Connaissances pratiques. Onema*.
- Schirmer M, Ijaz UZ, D'Amore R *et al.* (2015) Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic acids research*, **43**, e37.
- Schloss PD (2016) Application of a Database-Independent Approach To Assess the Quality of Operational Taxonomic Unit Picking Methods. *mSystems*, **1**, e00027-16.
- Schloss PD, Gevers D, Westcott SL (2011) Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *Plos One*, **6**, e27310.
- Schloss PD, Westcott SL, Ryabin T *et al.* (2009) Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, **75**, 7537–7541.
- Schmidt TSB, Matias Rodrigues JF, von Mering C (2015) Limits to robustness and reproducibility in the demarcation of operational taxonomic units. *Environmental Microbiology*, **17**, 1689–1706.
- Schrader C, Schielke A, Ellerbroek L, Johne R (2012) PCR inhibitors - occurrence, properties and removal. *Journal of Applied Microbiology*, **113**, 1014–1026.
- Shaw JLA, Clarke LJ, Wedderburn SD *et al.* (2016) Comparison of environmental DNA metabarcoding and conventional fish survey methods in a river system. *Biological Conservation*, **197**, 131–138.
- Shehzad W, Riaz T, Nawaz MA *et al.* (2012) Carnivore diet analysis based on next-generation sequencing: application to the leopard cat (*Prionailurus bengalensis*) in Pakistan. *Molecular ecology*, **21**, 1951–65.
- Shokralla S, Gibson J, King I *et al.* (2016) Environmental DNA barcode sequence capture: targeted, PCR-free sequence capture for biodiversity analysis from bulk environmental samples. *bioRxiv preprint*, 1–28.
- Shokralla S, Spall JL, Gibson JF, Hajibabaei M (2012) Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology*, **21**, 1794–1805.
- Sims PA, Mann DG, Medlin LK (2006) Evolution of the diatoms: insights from fossil, biological and molecular data. *Phycologia*, **45**, 361–402.
- Sipos R, Székely AJ, Palatinszky M *et al.* (2007) Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targeting bacterial community analysis. *FEMS microbiology ecology*, **60**, 341–50.
- Smetacek VS (1985) Role of sinking in diatom life-history cycles: ecological, evolutionary and geological significance. *Marine Biology*, **84**, 239–251.
- Smith DP, Peay KG (2014) Sequence Depth, Not PCR Replication, Improves Ecological Inference from Next Generation DNA Sequencing (CA Kellogg, Ed.). *Plos One*, **9**, e90234.
- Smol JP, Stoermer EF (2010) *The Diatoms: Applications for the Environmental and Earth*

- Sciences*. Cambridge University Press, Cambridge, United Kingdom.
- Snoeijs P, Busse S, Potapova M (2002) The importance of diatom cell size in community analysis. *Journal of Phycology*, **38**, 265–281.
- Sogin ML, Morrison HG, Huber JA *et al.* (2006) Microbial diversity in the deep sea and the underexplored “rare biosphere.” *Proceedings of the National Academy of Sciences*, **103**, 12115–12120.
- Somerville C, Knight I, Straube W, Colwell R (1989) Simple, rapid method for direct isolation of nucleic acids from aquatic environments. *Applied and Environmental Microbiology*, **55**, 548–554.
- Somervuo P, Yu DW, Xu CCY *et al.* (2017) Quantifying uncertainty of taxonomic placement in DNA barcoding and metabarcoding (D Warton, Ed.). *Methods in Ecology and Evolution*, **8**, 398–407.
- Sorhannus U (2007) A nuclear-encoded small-subunit ribosomal RNA timescale for diatom evolution. *Marine Micropaleontology*, **65**, 1–12.
- Spaulding S a., Jewson DH, Bixby RJ, Nelson H, McKnight DM (2012) Automated measurement of diatom size. *Limnology and Oceanography: Methods*, **10**, 882–890.
- Staley C, Gould TJ, Wang P *et al.* (2015) Evaluation of water sampling methodologies for amplicon-based characterization of bacterial community structure. *Journal of Microbiological Methods*, **114**, 43–50.
- Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.
- Stein ED, Martinez MC, Stiles S, Miller PE, Zakharov E V (2014) Is DNA barcoding actually cheaper and faster than traditional morphological methods: results from a survey of freshwater bioassessment efforts in the United States? (M Casiraghi, Ed.). *Plos One*, **9**, e95525.
- Stein ED, White BP, Mazor RD, Miller PE, Pilgrim EM (2013) Evaluating ethanol-based sample preservation to facilitate use of DNA barcoding in routine freshwater biomonitoring programs using benthic macroinvertebrates. *Plos One*, **8**, e51273.
- Stevenson RJ, Bothwell ML, Lowe RL (1996) *Algal ecology: Freshwater benthic ecosystem*.
- Stevenson RJ, Pan Y (1999) Assessing environmental conditions in rivers and streams with diatoms. *The diatoms: applications for the environmental and earth sciences*, **1**, 4.
- Stevenson RJ, Pan Y, van Dam H (2010a) Assessing environmental conditions in rivers and streams with diatoms. In: *The Diatoms* (eds Smol JP, Stoermer EF), pp. 57–85. Cambridge University Press, Cambridge.
- Stevenson RJ, Yangdong P, Van Dam H (2010b) Assessing environmental conditions in rivers and streams with diatoms. In: *The Diatoms: Applications for the Environmental and Earth Sciences* (eds Smol JP, Stoermer EF), pp. 55–85. Cambridge University Press.
- Stoof-Leichsenring KR, Epp LS, Trauth MH, Tiedemann R (2012) Hidden diversity in diatoms of Kenyan Lake Naivasha: a genetic approach detects temporal variation. *Molecular Ecology*, **21**, 1918–1930.
- Sumper M, Brunner E (2006) Learning from Diatoms: Nature’s Tools for the Production of Nanostructured Silica. *Advanced Functional Materials*, **16**, 17–26.
- Sun J, Liu D (2003) Geometric models for calculating cell biovolume and surface area for phytoplankton. *Journal of Plankton Research*, **25**, 1331–1346.
- Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH (2012a) Environmental DNA. *Molecular Ecology*, **21**, 1789–1793.
- Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012b) Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, **21**, 2045–2050.

- Taberlet P, Prud'homme SM, Campione E *et al.* (2012c) Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. *Molecular Ecology*, **21**, 1816–1820.
- Tan B, Ng C, Nshimiyimana JP *et al.* (2015) Next-generation sequencing (NGS) for assessment of microbial water quality: current progress, challenges, and future opportunities. *Frontiers in microbiology*, **6**, 1027.
- Tapolczai K, Bouchez A, Stenger-Kovács C, Padisák J, Rimet F (2016) Trait-based ecological classifications for benthic algae: review and perspectives. *Hydrobiologia*, **776**, 1–17.
- Tapolczai K, Bouchez A, Stenger-Kovács C, Padisák J, Rimet F (2017) Taxonomy- or trait-based ecological assessment for tropical rivers? Case study on benthic diatoms in Mayotte island (France, Indian Ocean). *Science of The Total Environment*, **607–608**, 1293–1303.
- Tapolczai K, Vasselon V, Bouchez A *et al.* Optimization of OTU-based water quality index with high throughput sequencing. *Molecular Ecology Resources*.
- Tedersoo L, Tooming-Klunderud A, Anslan S (2017) PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. *New Phytologist*.
- Teittinen A, Taka M, Ruth O, Soininen J (2015) Variation in stream diatom communities in relation to water quality and catchment variables in a boreal, urbanized region. *Science of The Total Environment*, **530–531**, 279–289.
- Theriot EC, Ashworth MP, Nakov T, Ruck E, Jansen RK (2015) Dissecting signal and noise in diatom chloroplast protein encoding genes with phylogenetic information profiling. *Molecular Phylogenetics and Evolution*, **89**, 28–36.
- Theriot EC, Ashworth M, Ruck E, Nakov T, Jansen RK (2010) A preliminary multigene phylogeny of the diatoms (Bacillariophyta): challenges for future research. *Plant Ecology and Evolution*, **143**, 278–296.
- Theriot EC, Ruck E, Ashworth M, Nakov T, Jansen RK (2011) Status of the pursuit of the diatom phylogeny: Are traditional views and new molecular paradigms really that different? In: *The Diatom World* (eds Seckbach J, Kociolek JP), pp. 119–142. Springer, New York, USA.
- Theron J, Cloete TE (2000) Molecular Techniques for Determining Microbial Diversity and Community Structure in Natural Environments. *Critical Reviews in Microbiology*, **26**, 37–57.
- Thomas AC, Deagle BE, Eveson JP, Harsch CH, Trites AW (2016) Quantitative DNA metabarcoding: improved estimates of species proportional biomass using correction factors derived from control material. *Molecular Ecology Resources*, **16**, 714–726.
- Thomsen PF, Willerslev E (2015) Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation*, **183**, 4–18.
- Trebitz AS, Hoffman JC, Darling JA *et al.* (2017) Early detection monitoring for aquatic non-indigenous species: Optimizing surveillance, incorporating advanced technologies, and identifying research needs. *Journal of Environmental Management*, **202**, 299–310.
- Treusch AH, Demir-Hilton E, Vergin KL *et al.* (2012) Phytoplankton distribution patterns in the northwestern Sargasso Sea revealed by small subunit rRNA genes from plastids. *The ISME Journal*, **6**, 481–492.
- Tryon T (1684) Friendly Advice to the Gentlemen Planters of the East and West Indies. In: *Versions of Blackness* (ed Hughes D), pp. 349–352. Cambridge University Press, Cambridge.
- Tudesque L, Rimet F, Ector L (2008) A new taxon of the section *Nitzschiae lanceolatae* Grunow: *Nitzschia costei* sp. nov. compared to *N. fonticola* Grunow, *N. macedonica* Hustedt, *N. tropica* Hustedt and related species. *Diatom Research*, **23**, 483–501.
- Valentini A, Pompanon F, Taberlet P (2009) DNA barcoding for ecologists. *Trends in Ecology & Evolution*, **24**, 110–117.

- Valentini A, Taberlet P, Miaud C *et al.* (2016) Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, **25**, 929–942.
- Vanellander B, Créach V, Vanormelingen P *et al.* (2009) Ecological Differentiation Between Sympatric Pseudocryptic Species in the Estuarine Benthic Diatom *Navicula Phyllepta* (Bacillariophyceae)1. *Journal of Phycology*, **45**, 1278–1289.
- de Vargas C, Audic S, Henry N *et al.* (2015) Eukaryotic plankton diversity in the sunlit ocean. *Science*, **348**, 1261605–1261605.
- Vasselon V, Bouchez A, Rimet F *et al.* (2018) Avoiding quantification bias in metabarcoding: application of a cell biovolume correction factor in diatom molecular biomonitoring. *Methods in Ecology and Evolution*, **in press**.
- Vasselon V, Domaizon I, Rimet F, Kahlert M, Bouchez A (2017a) Application of high-throughput sequencing (HTS) metabarcoding to diatom biomonitoring: Do DNA extraction methods matter? *Freshwater Science*, **36**, 162–177.
- Vasselon V, Rimet F, Tapolczai K, Bouchez A (2017b) Assessing ecological status with diatoms DNA metabarcoding: Scaling-up on a WFD monitoring network (Mayotte island, France). *Ecological Indicators*, **82**, 1–12.
- Veldhuis M, Kraay G, Timmermans K (2001) Cell death in phytoplankton: correlation between changes in membrane permeability, photosynthetic activity, pigmentation and growth. *European Journal of Phycology*, **36**, 167–177.
- Visco JA, Apothéloz-Perret-Gentil L, Cordonier A *et al.* (2015) Environmental Monitoring: Inferring the Diatom Index from Next-Generation Sequencing Data. *Environmental Science & Technology*, **49**, 7597–7605.
- Vitousek PM, Mooney H a, Lubchenco J, Melillo JM (1997) Human Domination of Earth' s Ecosystems. *Science*, **277**, 494–499.
- Vivien R, Lejzerowicz F, Pawlowski J (2016) Next-generation sequencing of aquatic oligochaetes: Comparison of experimental communities. *Plos One*, **11**, 1–14.
- Vörösmarty CJ, Green P, Salisbury J, Lammers RB (2000) Global water resources: vulnerability from climate change and population growth. *Science (New York, N.Y.)*, **289**, 284–8.
- Vörösmarty CJ, McIntyre PB, Gessner MO *et al.* (2010) Global threats to human water security and river biodiversity. *Nature*, **468**, 334–334.
- Voulvoulis N, Georges K (2016) Industrial and Agricultural Sources and Pathways of Aquatic Pollution. In: *Impact of Water Pollution on Human Health and Environmental Sustainability*, pp. 29–54.
- Wagner A, Blackstone N, Cartwright P *et al.* (1994) Surveys of Gene Families Using Polymerase Chain Reaction: PCR Selection and PCR Drift. *Systematic Biology*, **43**, 250–261.
- Wagner Mackenzie B, Waite DW, Taylor MW (2015) Evaluating variation in human gut microbiota profiles due to DNA extraction method and inter-subject differences. *Frontiers in Microbiology*, **6**, 1–11.
- Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, **73**, 5261–5267.
- Water quality - Guidance standard for the routine sampling and pretreatment of benthic diatoms from rivers* (2016) Brussels.
- Watson JD, Crick FHC (1953) Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, **171**, 737–738.
- Weber A a-T, Pawlowski J (2013) Can abundance of protists be inferred from sequence data: a case study of foraminifera. *PloS one*, **8**, e56739.
- Weir ET (1982) *Botany : An Introduction to Plant Biology*. Wiley and Sons, Incorporated, John.
- Werner P, Adler S, Dreßler M (2016) Effects of counting variances on water quality

- assessments: implications from four benthic diatom samples, each counted by 40 diatomists. *Journal of Applied Phycology*, **28**, 2287–2297.
- Wesolowska-Andersen A, Bahl M, Carvalho V *et al.* (2014) Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. *Microbiome*, **2**, 19.
- Willerslev E, Hansen AJ, Binladen J *et al.* (2003) Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science*, **300**, 791–5.
- Williams DM, Kociolek JP (2007) Pursuit of a natural classification of diatoms: History, monophyly and the rejection of paraphyletic taxa. *European Journal of Phycology*, **42**, 313–319.
- Willner D, Daly J, Whiley D *et al.* (2012) Comparison of DNA Extraction Methods for Microbial Community Profiling with an Application to Pediatric Bronchoalveolar Lavage Samples (RK Aziz, Ed.). *Plos One*, **7**, e34605.
- Withgott J, Brenan SR (2007) *Environment: The science behind the stories*. Pearson.
- Woznicki SA, Nejadhashemi AP, Tang Y, Wang L (2016) Large-scale climate change vulnerability assessment of stream health. *Ecological Indicators*, **69**, 578–594.
- Xu F-L, Jørgensen SE, Tao S (1999) Ecological indicators for assessing freshwater ecosystem health. *Ecological Modelling*, **116**, 77–106.
- Yang Y, Cao J-X, Pei G-F, Liu G-X (2015) Using benthic diatom assemblages to assess human impacts on streams across a rural to urban gradient. *Environmental Science and Pollution Research*, **22**, 18093–18106.
- Yang W, Lopez PJ, Rosengarten G (2011) Diatoms: Self assembled silicananostructures, and templates for bio/chemical sensors and biomimetic membranes. *The Analyst*, **136**, 42–53.
- Yuan J, Li M, Lin S (2015) An Improved DNA Extraction Method for Efficient and Quantitative Recovery of Phytoplankton Diversity in Natural Assemblages. *PloS one*, **10**, e0133060.
- Zarfl C, Lumsdon AE, Berlekamp J, Tydecks L, Tockner K (2015) A global boom in hydropower dam construction. *Aquatic Sciences*, **77**, 161–170.
- Zelinka M, Marvan P (1961) Zur präzisierung der biologischen klassifikation der reinheit fließender gewässer. *Archiv für Hydrobiologie*, **57**, 389–407.
- Zetsche E-M, Meysman FJR (2012) Dead or alive? Viability assessment of micro- and mesoplankton. *Journal of Plankton Research*, **34**, 493–509.
- Zhu F, Massana R, Not F, Marie D, Vaultot D (2005) Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS microbiology ecology*, **52**, 79–92.
- Zimmermann J, Abarca N, Enk N *et al.* (2014) Taxonomic Reference Libraries for Environmental Barcoding: A Best Practice Example from Diatom Research (B Schierwater, Ed.). *Plos One*, **9**, e108793.
- Zimmermann J, Glöckner G, Jahn R, Enke N, Gemeinholzer B (2015) Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. *Molecular Ecology Resources*, **15**, 526–542.
- 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**, 173–192.
- Znachor P, Rychtecký P, Nedoma J, Visocká V (2015) Factors affecting growth and viability of natural diatom populations in the meso-eutrophic Římov Reservoir (Czech Republic). *Hydrobiologia*, **762**, 253–265.
- Zurzolo C, Bowler C (2001) Exploring Bioinorganic Pattern Formation in Diatoms. A Story of Polarized Trafficking. *Plant Physiology*, **127**, 1339–1345.

Contributions scientifiques

Articles scientifiques

2017 **Vasselon V.**, Bouchez A., Rimet F., Jacquet S., Trobajo R., Corniquel M., Tapolczai K., Domaizon I. "Avoiding quantification bias in metabarcoding: application of a cell biovolume correction factor in diatom molecular biomonitoring". *Methods in Ecology and Evolution*. In press.

Vasselon V., Domaizon I., Rimet F., Kahlert M., Bouchez A. "Application of high-throughput sequencing (HTS) metabarcoding to diatom biomonitoring: Do DNA extraction methods matter?" *Freshwater Science*.

Vasselon V., Rimet F., Tapolczai K., Bouchez, A. "Assessing ecological status with phytoplankton DNA metabarcoding : scaling-up on a WFD monitoring network (Mayotte island, France)". *Ecological indicators*.

Keck F., **Vasselon V.**, Tapolczai K., Rimet F., Bouchez A. "Freshwater biomonitoring in the Information Age: Challenges and perspectives." *Frontiers in Ecology and the Environment*.

Rivera S. F., **Vasselon V.**, Jacquet S., Bouchez A., Ariztegui D., Rimet, F. "Metabarcoding of lake benthic diatoms: from structure assemblages to ecological assessment." *Hydrobiologia*.

Rimet F., Abarca N., Bouchez A., Kusber W.H., Jahn R., Kahlert M., Keck F., Kelly M., Mann D.G., Piuze A., Trobajo R., Tapolczai K., **Vasselon V.**, Zimmermann J. "The potential of high throughput sequencing (HTS) of natural samples as a source of primary taxonomic information for reference libraries of diatom barcodes". *Fottea*, à paraître en 2018.

Chonova T., Labanowski J., Cournoyer B., Chardon C., Keck F., Laurent E., Marjolet L., Marti R., Mondamert L., Montuelle B., Rimet F., **Vasselon V.**, Bouchez, A. "Biofilm communities in river impacted by pharmaceutical loads from a wastewater treatment plant". *Environmental Science and Pollution Research*.

2016 Rimet F., Chaumeil P., Keck F., Kermarrec L., **Vasselon V.**, Kahlert M., Franc A., Bouchez A. "R-Syst::diatom: An open-access and curated barcode database for diatoms and freshwater monitoring." *Database: The Journal of Biological Databases and Curation*.

2015 Kahlert M., Bouchez A., Chaumeil P., Franc A., Frigerio J.M., Rimet F., Salin F., **Vasselon V.** "Gaps to fill when analyzing freshwater diatom diversity with DNA barcoding – notes from a boreal region." *European Journal of Phycology*, actes de colloque.

Articles scientifiques soumis

Keck F., **Vasselon V.**, Rimet F., Bouchez A., Kahlert M. "Enhancing DNA metabarcoding for biomonitoring with phylogenetic estimation of ecological profiles for unclassified OTUs". *Methods in Ecology and Evolution*.

Tapolczai K., **Vasselon V.**, Bouchez A., Stenger-Kovács C., Padisák J., Rimet F. "Optimization of OTU-based water quality index with high throughput sequencing". *Molecular Ecology Resources*.

Congrès internationaux avec comité de lecture

2017 Bouchez A., Kermarrec L., Reyjol Y., Tapolczai K., **Vasselon V.**, Rimet F. "On the way to implementation of ecogenomic indices for river biomonitoring: a French progress report for diatoms. " 10th Symposium for European Freshwater Sciences, Oulomouc (République Tchèque), 2 au 7 Juillet 2017.

Tapolczai K., Bouchez A., Stenger-Kovács C., Padisák J., **Vasselon V.**, Rimet F. "Diatom-based ecological assessment on the rivers of the tropical island, mayotte (france) using different approaches. " 10th Symposium for European Freshwater Sciences, Oulomouc (Check Republic), 2 au 7 Juillet 2017.

Vasselon V., Domaizon I., Rimet F., Tapolczai K., Bouchez A. "Optimization of diatom DNA metabarcoding : application to Mayotte streams monitoring network." COST DNAqua-net kick-off meeting, Essen (Allemagne), 7 au 8 Mars 2017.

Rimet F., **Vasselon V.**, Tapolczai K., Bouchez A. "R-Syst::diatom, a reference library for diatoms: overview, uses and perspectives." COST DNAqua-net kick-off meeting, Essen (Allemagne), 7 au 8 Mars 2017.

Bouchez A., Franc A., Blancher P., Chaumeil P., Frigerio J.M., Keck F., Kermarrec L., Monnier O., Reyjol Y., Salin F., Tapolczai K., **Vasselon V.**, Rimet F. "Diatom DNA metabarcoding & WFD : where are we ?" COST DNAqua-net kick-off meeting, Essen (Allemagne), 7 au 8 Mars 2017.

2016 Tapolczai K., Bouchez A., **Vasselon V.**, Keck F., Stenger-Kovács C., Padisák J., Rimet F., "Species- and trait-based quality evaluation methods for the rivers of Mayotte (France, Southeast Africa)." 10th Central European Diatom Meeting, Budapest (Hongrie), 20 au 23 Avril 2016.

Rimet F., **Vasselon V.**, Keck F., Chardon C., Tapolczai K., Piuz A., Bouchez A. "Diatom DNA-barcoding databases: how to fill them quickly at low cost?" 10th Central European Diatom Meeting, Budapest (Hongrie), 20 au 23 Avril 2016.

Chonova T., Labanowski J., Chardon C., Keck F., Laurent E., Mondamert L., Montuelle B., Rimet F., **Vasselon V.**, Bouchez A. "High Throughput Sequencing to highlight bacterial community changes in river biofilms linked to pharmaceutical loads from a wastewater treatment plant." Society of environmental toxicology and chemistry, Nantes (France), du 22 au 26 Mai 2016.

Bouchez A., Chardon C., Keck F., Rimet F., Tapolczai K., **Vasselon V.** “Metabarcoding and High-Throughput Sequencing for assessing river ecological quality with diatom indices at the scale of a regular monitoring network.” Society for Freshwater Science 2016 annual meeting, Sacramento (USA), 22 au 26 Mai 2016.

Vasselon V., Domaizon I., Kahlert M., Rimet F., Bouchez A. “Towards standardization of DNA extraction for next-generation biomonitoring with diatoms.” Society for Freshwater Science 2016 annual meeting, Sacramento (USA), 22 au 26 Mai 2016.

Chonova T., Chardon C., Keck F., Labanowski J., Laurent E., Mondamert L., Montuelle B., Rimet F., **Vasselon V.**, Bouchez A. “Next-Generation Sequencing to highlight community changes in river biofilms linked to pharmaceutical loads from a wastewater treatment plant.” 1st International Conference on Risk Assessment of Pharmaceuticals in the Environment, Paris (France), 8 au 9 Septembre 2016.

2015 Rimet F., Bouchez A., Keck F., **Vasselon V.**, Frigerio J.M., Chaumeil P., Franc A. “ Use of metabarcoding for water quality assessment using diatoms.” Dutch-Flemish Society of Diatomists meeting, Mont Rigi (Belgique), 4 au 6 Juin 2015.

Rimet F., Chaumeil P., Frigerio J.M., Tapolczai K., Keck F., **Vasselon V.**, Franc A., Bouchez A. “ Potential of diatom metabarcoding and phylogeny for ecological assessment.” 9th Use of Algae for Monitoring Rivers and comparable habitats. Trento (Italie), 15 au 19 Juin 2015.

Vasselon V., Rimet F., Salin F., Franc A., Bouchez A. “Temporal evolution of benthic diatom community in Lake Geneva using Next-Generation Sequencing approach.” Symposium for European Freshwater Sciences, Genève (Suisse), 5 au 10 Juillet 2015.

Franc A., Rimet F., Chaumeil P., Frigerio J.M., Laizet Y., Keck F., **Vasselon V.**, Kahlert M., Bouchez A. “A pipeline for building molecular based inventories and computing diversity indices of communities.” Symposium for European Freshwater Sciences, Genève (Suisse), 5 au 10 Juillet 2015.

Rimet F., Chaumeil P., Frigerio J.M., Keck F., **Vasselon V.**, Franc A., Bouchez A. “R-Syst::diatom: An open-access and curated barcode database for diatoms.” Symposium for European Freshwater Sciences, Genève (Suisse), 5 au 10 Juillet 2015.

Kahlert M., Bouchez A., Chaumeil P., Franc A., Frigerio J.M., Rimet F., Salin F., **Vasselon V.** “Gaps to fill when analyzing freshwater diatom diversity with DNA barcoding – notes from a boreal region.” 6th European Phycological Congress, Londres (UK), 23 au 28 Août 2015.

Congrès francophones avec comité de lecture

2016 Chonova T., Chardon C., Keck F., Labanowski J., Laurent E., Mondamert L., Montuelle B., Rimet F., **Vasselon V.**, Bouchez A. “L'impact d'effluents de STEP sur la structure de communautés microbiennes de biofilms dépend de l'origine des effluents et de leur charge résiduelle en médicaments.” 3^{èmes} Journées d'Écotoxicologie Microbienne, Rovaltain (France), 16 au 18 Mars 2016.

Tapolczai K., Bouchez A., **Vasselon V.**, Keck F., Stenger-Kovács C., Padisák J., Rimet F. "L'évaluation de la qualité des cours d'eau de Mayotte basé sur un indice classique et un indice trait." 35^{ème} Colloque de l'Association des Diatomistes de Langue Française, Belvaux (Luxembourg), 13 au 15 Septembre 2016.

Rimet F., **Vasselon V.**, Keck F., Chardon C., Tapolczai K., Piuze A., Bouchez A. "Bases de référence de barcodes-ADN diatomées : comment les compléter rapidement à faible coût ?" 35^{ème} Colloque de l'Association des Diatomistes de Langue Française, Belvaux (Luxembourg), 13 au 15 Septembre 2016.

2015 Bouchez A., Rimet F., Chaumeil P., Frigerio J.M., Keck F., Tapolczai K., **Vasselon V.**, Franc A. "Potentiel du metabarcoding et de la phylogénie des diatomées pour la bioindication." 34^{ème} Colloque de l'Association des Diatomistes de Langue Française, Bordeaux (France), 7 - 10 Septembre 2015.

Posters

2017 **Vasselon V.**, Bouchez A., Corniquel M., Jacquet S., Rimet F., Tapolczai K., Domaizon I. "Quantitative diatom metabarcoding : a correction factor inferred from cell biovolume." COST DNAqua-net kick-off meeting, Essen (Allemagne), 7 au 8 Mars 2017.

2016 Chonova T., Chardon C., Keck F., Labanowski J., Laurent E., Mondamert L., Montuelle B., Rimet F., **Vasselon V.**, Bouchez A. "Next-Generation Sequencing to highlight community changes in river biofilms linked to pharmaceutical loads from a wastewater treatment plant." Society for Freshwater Science 2016 annual meeting, Sacramento (USA), 22 au 26 Mai 2016.

Abonyi-Keszte B., **Vasselon V.**, Feret L., Jacas L., Birck C., Bouchez A., Rimet F. "Lacs Alpains d'altitude : comparaison des approches microscopie et barcoding ADN." 35^{ème} Colloque de l'Association des Diatomistes de Langue Française, Belvaux (Luxembourg), 13 au 15 Septembre 2016.

Rivera S., Ariztegui D., Frossard V., **Vasselon V.**, Jacquet S., Bouchez A., Rimet F. "Evaluation de la qualité environnementale du lac du Bourget : Une comparaison entre les approches de microscopie et de barcoding." 35^{ème} Colloque de l'Association des Diatomistes de Langue Française, Belvaux (Luxembourg), 13 au 15 Septembre 2016.

Chonova T., Labanowski J., Chardon C., Keck F., Laurent E., Mondamert L., Montuelle B., Rimet F., **Vasselon V.**, Bouchez A. "Biofilm communities in river impacted by pharmaceutical loads from a wastewater treatment plant." Forum ECO-TOX Rovaltain, Alixan (France), 11 au 13 Octobre 2016.

Articles annexes

Article annexe I

“Freshwater biomonitoring in the Information Age”

(paru dans le journal *Frontiers in Ecology and the Environment*, 2017)

Keck F., Vasselon V., Tapolczai K., Rimet F., Bouchez A.

CARTELE, INRA, Université de Savoie Mont Blanc, 74200, Thonon-les-bains, France

Freshwater biomonitoring in the Information Age

François Keck^{1,2*}, Valentin Vasselon¹, Kálmán Tapolczai¹, Frédéric Rimet¹, and Agnès Bouchez¹

Freshwaters worldwide face serious threats, making their protection increasingly important. Freshwater monitoring has historically produced valuable data and continues to develop. Rapid improvements to biomolecular techniques are revolutionizing the way scientists describe biological communities and are bringing about major changes in biomonitoring. Combined with high-throughput sequencing, DNA metabarcoding is fast and cost-effective, generating massive amounts of data. In a world with numerous ecological threats, “big data” constitute a tremendous opportunity to improve the efficiency of biological monitoring. These fundamental changes in biomonitoring will require freshwater ecologists and environmental managers to reconsider how they handle large amounts of data.

Front Ecol Environ 2017; 15(5): 266–274, doi:10.1002/fee.1490

Human activities have broadly affected freshwater ecosystems, especially since the Industrial Revolution. Over the past 50 years, however, policy makers and citizens have become more attuned to environmental issues. This has led to the development of important governmental programs to assess and limit ecological impacts of human activities (Figure 1). In this context, one objective of environmental managers is to evaluate how water quality changes over time. Bioindicator organisms are commonly used for this purpose, based on the premise that the presence or absence of certain biological communities at a given site reflects its environmental quality.

Freshwater biomonitoring has a long tradition in the field of ecology. A century of research has led to substantial improvements in understanding how human disturbances can shape biological communities. Based on this knowledge, many approaches have been developed to estimate environmental quality from the richness, diversity, structure, and functioning of these communities

(Jørgensen *et al.* 2010). These widely used methods are based on solid theoretical grounds and are known to perform quite well. Most of them commonly require a taxonomical description of the community. Hence, freshwater biomonitoring essentially consists of collecting individual organisms, performing taxonomic identification, and using inventories to estimate the environmental condition of a given site. However, traditional biomonitoring also faces recurrent criticisms, mainly related to taxonomic identification relying on morphological criteria, a process that is time-consuming, complex, and technically demanding (Mandelik *et al.* 2010). These limits inevitably restrict the number of sites that can be monitored and the frequency of controls.

During the past decade, the idea arose that DNA analyses (Figure 2) could advantageously replace morphological methods to identify species (Hebert *et al.* 2003). Metabarcoding was developed as a set of techniques to identify multiple taxa simultaneously from an environmental sample with standard genetic markers (Taberlet *et al.* 2012; Panel 1 and Figure 2). This has led to the idea of “Biomonitoring 2.0”, which offers novel perspectives for monitoring environmental communities (Baird and Hajibabaei 2012). In this paper, we explain why and how metabarcoding will profoundly change the nature of data produced by biomonitoring. We examine these changes in the general context of massive data production – so-called “big data”, a topic that is the subject of increasing interest in biology (Marx 2013). We show why this big data revolution holds promise for ecological assessment purposes. Finally, we highlight three challenges posed by big data for metabarcoding and propose a framework that takes them into account. We illustrate our point with examples taken from freshwater monitoring, where metabarcoding is developing rapidly (Hajibabaei *et al.* 2011; Kermarrec *et al.* 2014). Nevertheless, the ideas discussed could be extended and applied to a broader context.

In a nutshell:

- DNA metabarcoding and high-throughput sequencing methods produce massive quantities of data and will markedly change freshwater biomonitoring
- Molecular methods propel biomonitoring into the Information Age and bring exciting new opportunities to make ecological monitoring more effective and relevant
- Genetic “big data” challenge scientists to think differently about the way that biological monitoring information is analyzed; we propose and discuss alternatives to the classical taxonomic affiliation approach to process bioassessment metabarcoding data

¹UMR CARTELE, Institut National de la Recherche Agronomique, Université Savoie Mont Blanc, Thonon-Les-Bains, France;

²Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden * (francois.keck@gmail.com)

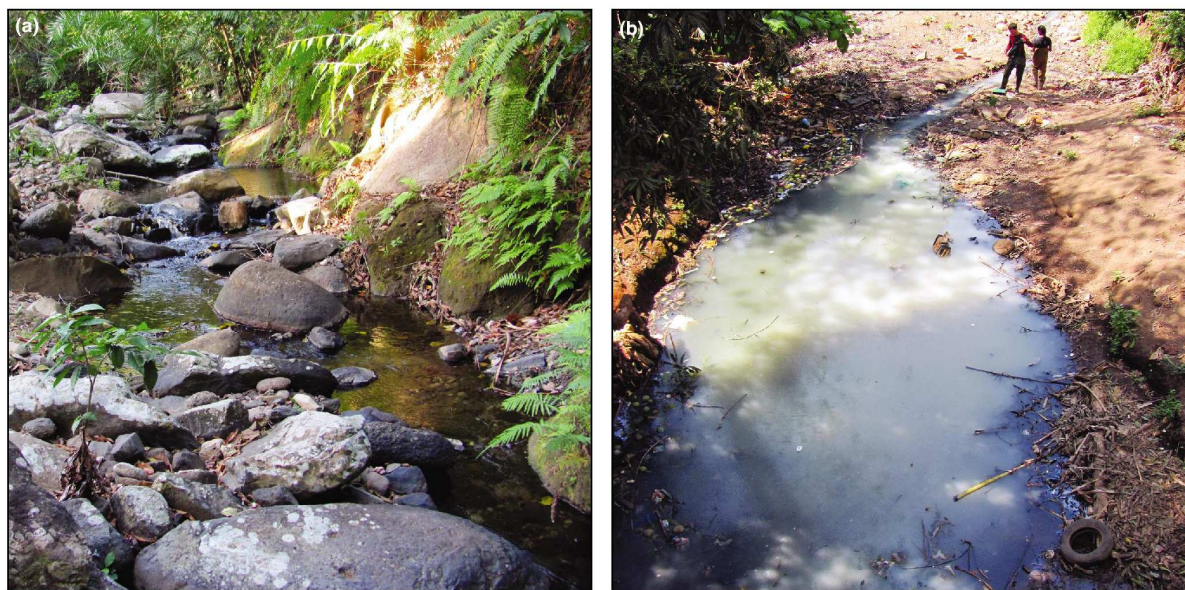


Figure 1. Two streams included in the river monitoring network of Mayotte Island, France. (a) A pristine upstream site (Longoni River) and (b) a polluted site located downstream of village waste (Majimbini River). The biological assessment of Mayotte's rivers currently relies on benthic diatom communities studied using both classical morpho-taxonomical and metabarcoding approaches.

■ Biomonitoring as a source of massive data

Characterizing ecological quality from biological entities has produced important sources of data since the first attempts to do so at the beginning of the 20th century. This is because biomonitoring largely consists of sampling, identifying, enumerating, and reporting biological organisms. The saprobic system for organic pollution assessment developed by Kolkwitz and Marsson (1908, 1909) is often cited as the first bioassessment tool in freshwaters and uses 298 plant species and 527 animal species as indicator organisms. Methods soon diversified thereafter, and specific biological groups (fishes, macroinvertebrates, algae) have been employed. Increasing stringency in precision requirements has led to more powerful and sophisticated tools, based on hundreds of families and thousands of species.

The amount of data produced has increased rapidly because biomonitoring is rarely done in isolation, but instead is replicated across space (through a network of sites; eg along a river, within a watershed) and over time (long-term monitoring). Since the 1970s, general awareness of ecological issues has grown, and biomonitoring has been increasingly implemented and incorporated into legal frameworks for fresh waters, such as the Clean Water Act (CWA, 1972) in the US and the Water Framework Directive (WFD, 2000) in Europe. This guarantees the abundant production of data with respect to recognized standards.

However, biomonitoring methods are expected to change considerably in coming years. After a century of classifying taxa based on morphological criteria, species

can now be identified through the use of DNA barcodes (Hebert *et al.* 2003); for definitions of selected specialist terms used throughout, see Panel 1. The introduction of high-throughput sequencing (HTS; Shokralla *et al.* 2012) coupled with the development of extended reference databases (Ratnasingham and Hebert 2007; Benson *et al.* 2008) and efficient bioinformatics tools (eg Schloss *et al.* 2009) have enabled the production of reliable and cost-effective community inventories from environmental DNA (Chariton *et al.* 2015; Gibson *et al.* 2015; Pawlowski *et al.* 2016). While numerous issues and technical limitations remain (DNA spatial transfer and persistence over time, polymerase chain reaction [PCR] amplification biases, sequencing errors, chimeras, quantification; see also Coissac *et al.* 2012 and Shokralla *et al.* 2012), methods are improving quickly and metabarcoding is expected to be an increasingly important component of biomonitoring in the future.

The progressive adoption of metabarcoding for taxonomical identification will substantially increase the volume of data produced by biomonitoring activities and modify the characteristics of these data (Dafforn *et al.* 2016). It is often stated that characteristics of big data fulfill five “Vs”: volume, velocity, variety, variability, and value (Fan and Bifet 2013). Biomonitoring data will likely meet these five criteria in unprecedented ways in the coming years.

Volume

The amount of data acquired from biomonitoring is expected to increase very quickly. HTS techniques are

Panel 1. Biomonitoring and metabarcoding

The biological monitoring of freshwater systems is traditionally based on the morphological identification of indicator species, which provides information on the ecological status of their environment. Instead of relying on morphological features (eg size, shape) to perform species identification, which requires specialized knowledge of taxonomic groups, small DNA fragments – about 300 base pairs in length, known as DNA barcodes – can be used (Hebert *et al.* 2003). This identification approach is termed DNA barcoding. Existing DNA barcode reference databases are based on different genes (including *CO1*, *18S*, and *rbcl*) and link species taxonomy to DNA barcodes. While DNA barcoding is useful for identifying individual specimens, its application to community-level samples (ie multiple species) was difficult because it required sorted samples or even isolating and cultivating individuals. This challenge was overcome through a metagenomic method called metabarcoding, which allows for the detection of all species found in one sample directly from their DNA barcode sequences using a single workflow. The DNA is extracted directly from the sample, followed by the amplification and sequencing of the targeted DNA barcode (Figure 2). Using bioinformatics tools, DNA barcodes are compared to those contained in a reference database to identify the species composition within the sample.

Environmental DNA was defined by Taberlet *et al.* (2012) as the “DNA that can be extracted from environmental samples (such as soil, water, or air), without first isolating any target organisms”. This includes DNA from microorganisms and free DNA. The free part of environmental DNA may be used to detect the presence of invasive species (Ficetola *et al.* 2008) or to monitor rare and indicator species (Mächler *et al.* 2014). Microorganisms present in environmental samples (eg bacteria, fungi, and diatoms) enable the use of longer DNA barcodes

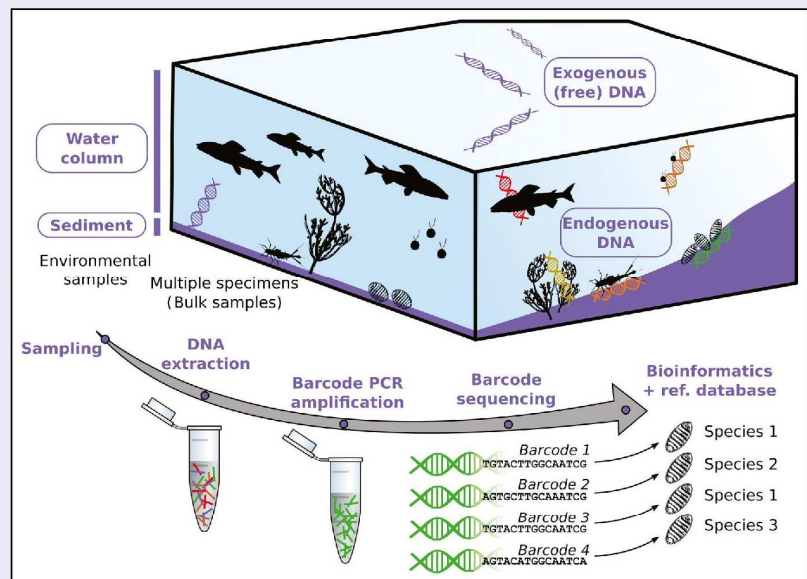


Figure 2. Several steps are required to perform DNA metabarcoding: (i) the sampling of environmental samples (eg sediment, biofilm, water) or the creation of bulk samples (mix of individual specimens); (ii) the extraction of the DNA; (iii) the amplification of a DNA barcode specific to the targeted community using polymerase chain reaction (PCR) techniques; (iv) the sequencing of the amplicons (amplified DNA barcodes); and (v) the taxonomical assignment of the DNA reads (amplicon sequences) using bioinformatics and a reference database (database connecting DNA barcode sequences to their taxonomic identity). Total environmental DNA comprises “endogenous” DNA from living organisms and “exogenous” free DNA.

(Taberlet *et al.* 2012) and facilitate access to uncultured taxa. For example, diatom molecular inventories can be used to calculate a quality index that indicates the ecological status of the sampled river (Kerमारrec *et al.* 2014; Visco *et al.* 2015). Precision and reliability of the species list obtained from DNA metabarcoding depend on the completeness and reliability of the reference database.

The development of high-throughput sequencing (HTS) enables the rapid and inexpensive sequencing of hundreds of environmental samples at a time, making the incorporation of the DNA metabarcoding into biomonitoring programs possible.

developing rapidly and have extremely high-throughput (Figure 3d). With the development of standardized protocols, the processing rate will also probably increase considerably and allow more sites to be surveyed and with greater frequency. Finally, assessments that rely on morphological criteria alone tend to underestimate species diversity, whereas the level of diversity detected by genetic methods tends to be much higher, especially for microbial communities (Caron *et al.* 2009), leading to larger inventory tables.

Velocity

Traditional monitoring requires experts to undertake a long and laborious process of taxonomically identifying collected biota. Consequently, one site is typically monitored seasonally or yearly. With metabarcoding and HTS techniques, however, the identification process is automated and faster. This will allow sites to be monitored at a finer time scale and to approach real-time monitoring.

Variety

Biomonitoring elicits multiple types of data. Community inventories generally come in the form of presence–absence or count data tables. Environmental managers often prefer to rely on multiple biological indicators (eg fishes and macroinvertebrates) to monitor multiple sources of impairment. Moreover, assessment methods commonly integrate physical and chemical data, which may also constitute big data, especially when recorded with remote sensors and with high frequency. Metabarcoding will also make it possible to work with genetic data and phylogenies (Hajibabaei *et al.* 2007).

Variability

Biomonitoring data are valuable when there is variability in community structures between reference and impacted sites (Jørgensen *et al.* 2010). With the use of DNA, finer-scale taxonomic characterization of communities can be achieved. Thus, with appropriate analyses, it will be possible to differentiate communities in a subtler way (Stein *et al.* 2014a) and to gain capacity in distinguishing the effects of various pressures.

Value

Data produced by biomonitoring are used to assess environmental quality. Many applications could be enhanced with big data, including monitoring over space and time; examining multi-trophic food web structure; and assessing the effects of pollution, environmental restoration, and invasive species. Moreover, biomonitoring data are often exploited by ecologists for purposes other than environmental assessment, such as studying biodiversity patterns or validating theoretical models (Lovett *et al.* 2007; Lindenmayer and Likens 2010).

Modern techniques and big data

Increasing the number of indicators

The modern concept of biomonitoring – as implemented in the WFD and CWA – is to use biological indicators accompanied by hydromorphological and physicochemical measurements (Ibáñez *et al.* 2010). For example, the WFD's bioindicators (or biological quality elements [BQEs]) are fishes, macrophytes, macroinvertebrates, benthic diatoms, and phytoplankton. Each of these indicators

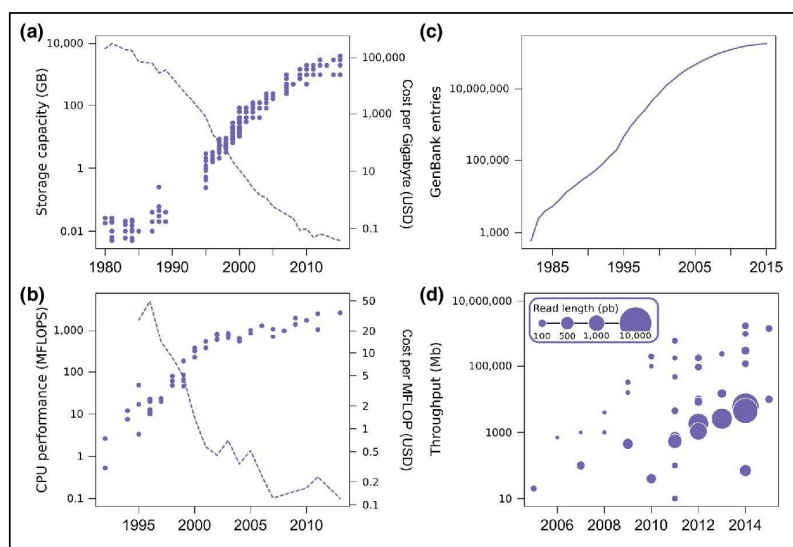


Figure 3. The Information Age is characterized by rapid technological developments exponentially increasing scientists' capacity to produce, store, and process data. (a) Storage capacity of commercialized computer hard drives in gigabytes (dots) and average price of a gigabyte (dashed line). (b) Microprocessor performance (dots) in millions of floating-point operations per second (MFLOPS) and average price of MFLOPS (dashed line). (c) Number of entries in the open-access nucleotide sequence database GenBank. (d) The throughput and read length evolution of high-throughput sequencing technologies.

presents advantages (eg diversity, ubiquity, ecological importance) and disadvantages (sampling difficulties, lack of metrics) (Resh 2008). Each BQE can indicate different pressures and provide complementary information (Passy *et al.* 2004; Figure 4). Thus, the overall quality assessment of an aquatic ecosystem is based on the results of all BQEs. In the WFD, the “one out all out” (OOAO) rule states that the worst status of the BQEs used in the assessment determines the final status of the ecosystem. However, in practice, using all BQEs for a sampled site is seldom or only partly achieved because of both financial and logistical constraints (Birk *et al.* 2012).

There is a trade-off between the ease of sampling and the ease of identifying organisms with respect to the average size of different BQEs (Figure 4). Groups of organisms with larger individual body size (typically fishes) are more difficult to sample representatively and collect, whereas smaller or microscopic organisms such as macroinvertebrates or benthic diatoms are relatively easy to collect by sampling the substrate directly. On the other hand, larger organisms are easier to manipulate and identify. For fishes and macrophytes, identification is performed in situ, whereas macroinvertebrates, benthic diatoms, and phytoplankton require arduous laboratory-based work (chemical treatment, microscopy). Modern molecular techniques appear to offer a promising solution to the trade-off between the ease of sampling and identifying organisms.

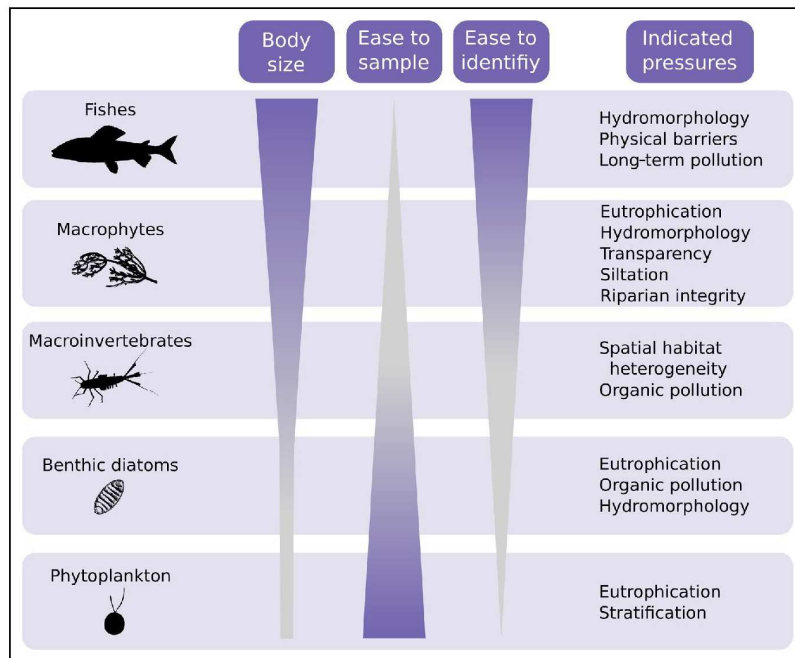


Figure 4. Gradients, trade-offs, and complementarity between body size, ease of sampling, ease of identification, and the indicated pressures of the five indicators included in the Water Framework Directive.

Covering a larger diversity

In traditional biomonitoring, taxonomical identification is rarely performed at the most precise levels of specificity because doing so is cost-prohibitive. DNA metabarcoding, however, could reveal diversity at the finest level for a fraction of that cost. With appropriate libraries, DNA barcodes can be linked to a Linnaean taxonomic name. The precision of taxonomic affiliation depends on the selected barcode and the availability of data in the reference libraries. By using correctly populated libraries, it is possible to reach the species level (eg Hajibabaei *et al.* 2011; Kermarrec *et al.* 2014) with less ambiguity and discrepancy than with classical microscopy, where species-level identification is often extremely laborious and even impossible at some development stages. However, data derived from DNA carry much more information than taxonomic names alone. Baird and Hajibabaei (2012) emphasized that genetic techniques have far more potential for identifying taxa than the traditional approach of relying on morphological characteristics. DNA-based techniques should facilitate working at the infra-species level and ultimately at the nucleotide level. It will therefore be possible to disentangle cryptic species complexes and to perform population-level analyses. Having the capacity to monitor diversity at so many levels should also promote the development of very sensitive tools to monitor the effects of specific types of pollution on various biota.

Enforcing and extending monitoring networks

High-throughput sequencing and the evolution of laboratory methods have made metabarcoding much more cost-effective (Stein *et al.* 2014b), and prices continue to decrease as technologies develop (van Dijk *et al.* 2014). DNA-based methods are also much faster than traditional methods. Sample processing can be serialized and automated with the aid of robots (Chapman 2003). Reductions in cost and processing time should boost sampling efforts by making it possible to increase the number of sites being monitored and the sampling frequency. This is an advantageous consequence of using metabarcoding, because biomonitoring often lacks spatial and temporal representativeness.

One specific site will poorly represent an entire ecosystem, particularly when habitats therein are heterogeneous and when bioindicators are micro-habitat dependent. To obtain

an improved and integrated view of environmental quality, researchers must augment the number of sampling sites to account for the spatial heterogeneity of the broader area. This increases the resolution of the grid of sampled sites and enables better interpolations among the nodes of the monitored network. For a given site, the frequency of sampling is also important. A more frequent sampling protocol gives a more reliable picture of the temporal evolution of the site's environmental quality. This is especially relevant for microscopic communities, which change extremely quickly with changes in the environment. Thus, sampling plans with higher spatial and temporal resolution should enable the development of more complex spatiotemporal models and increase the capacity to detect the effects of local and diffuse pollution.

■ Taking advantage of the data deluge: a proposed framework

From morphology to genetics: beyond the classical concept of species

Conventional taxonomy aims to classify biological organisms in different groups based on shared traits. These groups correspond to the different taxonomic levels, with the species level as a central unit. Even if still under debate (De Queiroz 2007), the concept and definition of species provides scientists with a unit of reference for ecological studies. With the rise of molecular methods,

the DNA sequence has appeared as a promising alternative unit. Scientists have tried to integrate genetic sequences in the classical taxonomy, with varying degrees of success (Padial *et al.* 2010). However, in the context of biomonitoring, the question remains, whether the traditional Linnaean binomial species name affiliation still makes sense within a full molecular approach.

Typically, DNA reads provided by HTS are clustered into molecular operational taxonomic units (MOTUs), which are in turn converted to species units through the use of a bioinformatic workflow and a DNA reference database. The conversion from DNA reads to species units is not without drawbacks: for instance, selected barcodes may be associated with incorrect taxonomic affiliations, genetic information may be lost (unaffiliated reads are discarded), and rare species are often insufficiently studied. This approach is suitable if the reference database is sufficiently comprehensive, but this is rarely the case because of the high species diversity and the time and effort required to sequence organisms' barcodes. Previously undescribed species are also frequently detected from genetic data, while formal taxonomic description can be a very long process (Goldstein and DeSalle 2011). Moving to full molecular biomonitoring will allow for much more data to be used, beyond that limited strictly to taxonomic assignments. The greatest challenge is to develop new, high-quality indices based on DNA reads and environmental information. Three alternative but complementary approaches are described below and are represented in Figure 5.

Developing MOTU-based indices

Biomonitoring assumes that the presence or absence of particular taxa at a site of interest is indicative of distinct environmental conditions at that site. Thus, in traditional biological assessments, an ecological profile associated with each taxon is required. Pawlowski *et al.* (2016) suggested calibrating MOTU-based indices with traditional indices computed from simultaneously conducted morphology-based identifications. However, the traditional indices could be easily adapted to the new molecular approach by computing the indices directly from the reads clustered in MOTUs (Steele *et al.* 2011). This approach would require databases associating reads, MOTUs, and their

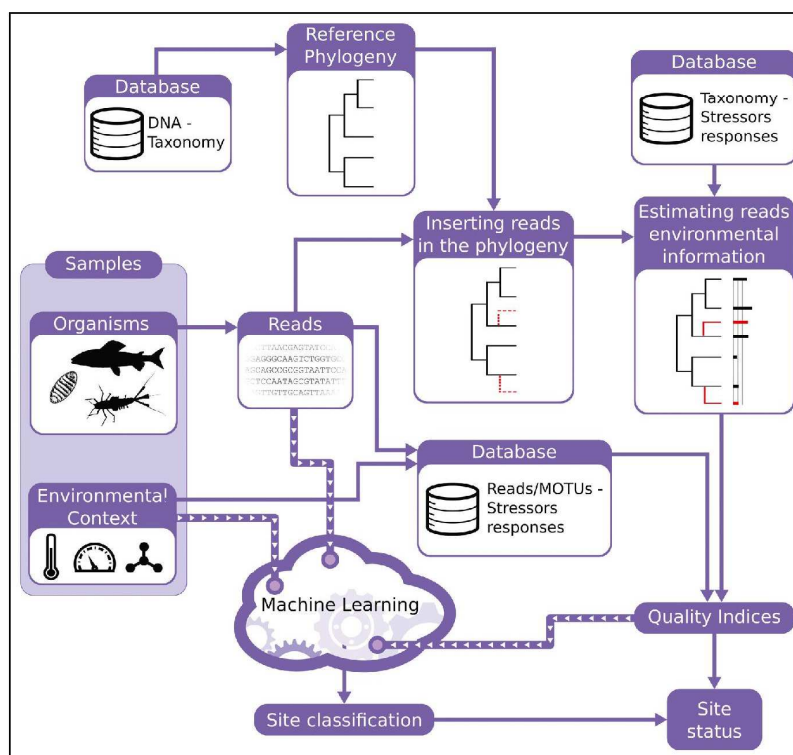


Figure 5. Flowchart introducing a new framework to process bioassessment metabarcoding data. Genetic reads can be interpreted without taxonomic affiliation using reads/MOTUs-based indices, phylogenetic modeling, or machine learning.

responses to environmental stressors (Figure 5). Thus, the MOTU-based indices approach is expected to be fully functional when ecological profiles for clusters of reads are estimated directly from previous molecular inventories; this will require substantial work in addition to data compilation and sharing. As a first step, known ecological profiles for taxa can be transferred to MOTUs.

Using phylogeny to include rare species

DNA metabarcoding can reveal a wealth of diversity, but the lack of taxon–stressor response libraries is problematic. Given that ecological profiles are usually estimated from in situ observations of general disturbances or from laboratory bioassays for specific substances, such libraries are restricted to common species and to a few types of disturbances. Rare species are often ignored (Guénard *et al.* 2011), and the effects of specific compounds remain poorly understood (Schwarzenbach *et al.* 2006).

One elegant way to solve these problems could involve phylogenetic methods harnessing the principle that species' tolerances are the legacy of evolution (Keck *et al.* 2016). The increasing availability of DNA sequences and computational power (Figure 3) should allow for the establishment of large and robust phylogenies. Then, if adequately long and informative (thereby excluding short

fragments and degraded DNA), reads can be inserted in the reference phylogeny using a posteriori replacement algorithms (Matsen *et al.* 2010; Berger *et al.* 2011). Finally, recent approaches to predict species' tolerances based on information available from other species and their respective phylogenetic positions (Guénard *et al.* 2013) could be used to estimate an ecological profile for a given read (Figure 5). Routine inclusion of such phylogenetic-based methods in biomonitoring would help to account for the immense diversity uncovered by DNA barcoding and the thousands of toxicants in the environment.

Machine learning techniques for ecological assessment

Analyzing and extracting valuable information from massive datasets can be extremely challenging. This has encouraged the development of machine learning methods, which use a set of statistical algorithms designed to recognize complex patterns in vast quantities of data. These methods include modern algorithms for classification, such as random forest, gradient boosting, support vector machines, and neural networks (Hastie *et al.* 2009). Machine learning approaches are fully data-driven and do not rely on any theoretical models (Breiman 2001). This system fits particularly well with the goals of biomonitoring, where the first aim is not necessarily to understand and explain the ecological processes leading to a given observation. In an applied context, correlation approaches are interesting because the final aim is to assess the state of the environment. This does not imply that machine learning should be used indiscriminately, but that these techniques are fully compatible with the ecological monitoring philosophy.

Machine learning methods have a broad range of applications. In biomonitoring, they may be used with different kinds of inputs for site classification, analyses of spatial networks of sites, and time-series forecasting. However, the most anticipated application of machine learning for biomonitoring is the processing of genetic

data. The ultimate aim is for algorithms to classify a new site directly from the bulk of DNA reads just by identifying genetic patterns learned from previous experience.

The same data can be interpreted in various ways if analyzed by different algorithms programmed with different training for different purposes (eg detection of eutrophication, effects of toxicants, or changes in flow regime). A set of sophisticated algorithms should enable scientists to monitor the effects of complex combinations of stressors on the environment. Such approaches are needed in view of multiple global threats (Vörösmarty *et al.* 2010). Furthermore, these methods should be implemented for massive datasets and communicate with holistic and integrative algorithms for automated and autonomous monitoring systems. In contrast to other more established fields in biology (Marx 2013), bioassessment is just beginning to face the problems associated with massive datasets. Scientists will need to begin collaborating more closely with experts in computer science and applied mathematics to benefit from big data, and to develop new ways to communicate results to managers (Panel 2).

Conclusions

With the development of DNA metabarcoding, traditional environmental monitoring is experiencing a period of transformation, one outcome of which will be the need to deal with unprecedented amounts of data. Ascertaining the technical requirements to obtain and analyze data is just a part of the challenge. In contrast to scientists from other disciplines, ecologists have a relatively poor culture of data sharing, despite opportunities for making big data more accessible (Reichman *et al.* 2011; Hampton *et al.* 2013). However, there are signs that this is starting to change. Making biomonitoring big data freely available will potentially allow a range of new applications such as meta-analyses and large-scale analyses of biodiversity. Metabarcoding data are particularly relevant in this case because genetic data are highly comparable. Scientists and resource

Panel 2. Communication with managers

Molecular methods constitute a new paradigm in freshwater ecosystem assessment. Environmental managers who are accustomed to traditional biological assessments and who are not familiar with genetics and molecular methods may be initially reluctant to adopt these approaches or may need training in order to do so. The widespread use of metabarcoding in biomonitoring depends on how these new tools will be implemented in future environmental assessment programs. Thus, new ways to communicate with resource managers must be developed. Communication should emphasize the benefits of metabarcoding, as well as explain the basics of genetics and the vocabulary of metabarcoding and HTS to managers in order to empower them to understand, interpret, communicate,

and benefit from the results of metabarcoding. However, we must also acknowledge difficulties, such as the challenges associated with machine learning. Although it is important that biomonitoring tools are derived from sound theoretical concepts in ecology, because machine learning often operates as a black box (ie the user does not understand how the algorithm works), it might be hard to relate results to environmental health and key stressors. The implementation of such new environmental assessment frameworks will therefore take time and require a close collaboration between scientists and managers. Knowledge and experience gained over many years must not be lost and traditional approaches should continue to be used, at least for the purposes of comparison and discussion.

managers must work together to create effective networks and to develop dedicated sharing platforms. Indeed, the technical solutions discussed in this paper require substantial quantities of data and supporting infrastructures. Sharing platforms should be accessible to citizens and ecologists and would provide both raw and processed data as well as metadata. Raw data can be re-used with new bioinformatic workflows and statistical methods, while processed data are important for non-specialists and to help inform citizens (Soranno *et al.* 2015). If we can make public – and make sense of – the terabytes of data that ecological assessments will produce in the foreseeable future, the entry of biomonitoring into the Information Age will be a genuine success.

Acknowledgements

We thank A Franc for constructive comments and I Domaizon for insightful discussion on metabarcoding terminology.

References

- Baird DJ and Hajibabaei M. 2012. Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Mol Ecol* 21: 2039–44.
- Benson DA, Karsch-Mizrachi I, Lipman DJ, *et al.* 2008. GenBank. *Nucleic Acids Res* 36: D25–30.
- Berger SA, Krompass D, and Stamatakis A. 2011. Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Systems Biol* 60: 291–302.
- Birk S, Bonne W, Borja A, *et al.* 2012. Three hundred ways to assess Europe's surface waters: an almost complete overview of biological methods to implement the Water Framework Directive. *Ecol Indic* 18: 31–41.
- Breiman L. 2001. Statistical modeling: the two cultures. *Stat Sci* 16: 199–231.
- Caron DA, Countway PD, Savai P, *et al.* 2009. Defining DNA-based operational taxonomic units for microbial-eukaryote ecology. *Appl Environ Microb* 75: 5797–808.
- Chapman T. 2003. Lab automation and robotics: automation on the move. *Nature* 421: 661–66.
- Chariton AA, Stephenson S, Morgan MJ, *et al.* 2015. Metabarcoding of benthic eukaryote communities predicts the ecological condition of estuaries. *Environ Pollut* 203: 165–74.
- Coissac E, Riaz T, and Puillandre N. 2012. Bioinformatic challenges for DNA metabarcoding of plants and animals. *Mol Ecol* 21: 1834–47.
- Dafforn KA, Johnston EL, Ferguson A, *et al.* 2016. Big data opportunities and challenges for assessing multiple stressors across scales in aquatic ecosystems. *Mar Freshwater Res* 67: 393–413.
- De Queiroz K. 2007. Species concepts and species delimitation. *Syst Biol* 56: 879–86.
- Fan W and Bifet A. 2013. Mining big data: current status, and forecast to the future. *SIGKDD Explorations* 14: 1–5.
- Ficetola GF, Miaud C, Pompanon F, and Taberlet P. 2008. Species detection using environmental DNA from water samples. *Biol Lett* 4: 423–25.
- Gibson JF, Shokralla S, Curry C, *et al.* 2015. Large-scale biomonitoring of remote and threatened ecosystems via high-throughput sequencing. *PLoS ONE* 10: e0138432.
- Goldstein PZ and DeSalle R. 2011. Integrating DNA barcode data and taxonomic practice: determination, discovery, and description. *Bioessays* 33: 135–47.
- Guénard G, Legendre P, and Peres-Neto P. 2013. Phylogenetic eigenvector maps: a framework to model and predict species traits. *Methods Ecol Evol* 4: 1120–31.
- Guénard G, von der Ohe PC, de Zwart D, *et al.* 2011. Using phylogenetic information to predict species tolerances to toxic chemicals. *Ecol Appl* 21: 3178–90.
- Hajibabaei M, Shokralla S, Zhou X, *et al.* 2011. Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE* 6: e17497.
- Hajibabaei M, Singer GAC, Hebert PDN, and Hickey DA. 2007. DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends Genet* 23: 167–72.
- Hampton SE, Strasser CA, Tewksbury JJ, *et al.* 2013. Big data and the future of ecology. *Front Ecol Environ* 11: 156–162.
- Hastie T, Tibshirani R, and Friedman J. 2009. The elements of statistical learning: data mining, inference, and prediction. 2nd edn. New York, NY: Springer.
- Hebert PDN, Cywinska A, Ball SL, and deWaard JR. 2003. Biological identifications through DNA barcodes. *P Roy Soc Lond B Bio* 270: 313–21.
- Ibáñez C, Caiola N, Sharpe P, and Trobajo R. 2010. Ecological indicators to assess the health of river ecosystems. In: Jørgensen SE, Xu F-L, and Costanza R (Eds). Handbook of ecological indicators for assessment of ecosystem health. Boca Raton, FL: CRC Press.
- Jørgensen SE, Xu F-L, Salas F, and Marques JC. 2010. Application of indicators for the assessment of ecosystem health. In: Jørgensen SE, Xu F-L, and Costanza R (Eds). Handbook of ecological indicators for assessment of ecosystem health. Boca Raton, FL: CRC Press.
- Keck F, Rimet F, Franc A, and Bouchez A. 2016. Phylogenetic signal in diatom ecology: perspectives for aquatic ecosystems biomonitoring. *Ecol Appl* 26: 861–72.
- Kermarrec L, Franc A, Rimet F, *et al.* 2014. A next-generation sequencing approach to river biomonitoring using benthic diatoms. *Freshwater Sci* 33: 349–63.
- Kolkwitz R and Marsson M. 1908. Ökologie der pflanzlichen Saprobien. *Ber Deut Bot Ges* 26: 505–19.
- Kolkwitz R and Marsson M. 1909. Ökologie der tierischen Saprobien. Beiträge zur Lehre von der biologischen Gewässerbeurteilung. *Int Rev Ges Hydrobiol Hydrogr* 2: 126–52.
- Lindenmayer DB and Likens GE. 2010. The science and application of ecological monitoring. *Biol Conserv* 143: 1317–28.
- Lovett GM, Burns DA, Driscoll CT, *et al.* 2007. Who needs environmental monitoring? *Front Ecol Environ* 5: 253–60.
- Mächler E, Deiner K, Steinmann P, and Altermatt F. 2014. Utility of environmental DNA for monitoring rare and indicator macroinvertebrate species. *Freshwater Sci* 33: 1174–83.
- Mandelik Y, Roll U, and Fleischer A. 2010. Cost-efficiency of biodiversity indicators for Mediterranean ecosystems and the effects of socio-economic factors. *J Appl Ecol* 47: 1179–88.
- Marx V. 2013. Biology: the big challenges of big data. *Nature* 498: 255–60.
- Matsen F, Kodner R, and Armbrust EV. 2010. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* 11: 538.
- Padial JM, Miralles A, la Riva ID, and Vences M. 2010. The integrative future of taxonomy. *Front Zool* 7: 1–14.
- Passy SI, Bode RW, Carlson DM, and Novak MA. 2004. Comparative environmental assessment in the studies of benthic diatom, macroinvertebrate, and fish communities. *Int Rev Hydrobiol* 89: 121–38.
- Pawlowski J, Lejzerowicz F, Apotheloz-Perret-Gentil L, *et al.* 2016. Protist metabarcoding and environmental biomonitoring: time for change. *Eur J Protistol* 55: 12–25.

- Ratnasingham S and Hebert PDN. 2007. BOLD: the Barcode of Life Data system (www.barcodinglife.org). *Mol Ecol Notes* 7: 355–64.
- Reichman OJ, Jones MB, and Schildhauer MP. 2011. Challenges and opportunities of open data in ecology. *Science* 331: 703–05.
- Resh VH. 2008. Which group is best? Attributes of different biological assemblages used in freshwater biomonitoring programs. *Environ Monit Assess* 138: 131–38.
- Schloss PD, Westcott SL, Ryabin T, *et al.* 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75: 7537–41.
- Schwarzenbach RP, Escher BI, Fenner K, *et al.* 2006. The challenge of micropollutants in aquatic systems. *Science* 313: 1072–77.
- Shokralla S, Spall JL, Gibson JF, and Hajibabaei M. 2012. Next-generation sequencing technologies for environmental DNA research. *Mol Ecol* 21: 1794–805.
- Steele JA, Countway PD, Xia L, *et al.* 2011. Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *ISME J* 5: 1414–25.
- Stein ED, White BP, Mazor RD, *et al.* 2014a. Does DNA barcoding improve performance of traditional stream bioassessment metrics? *Freshwater Sci* 33: 302–11.
- Stein ED, Martinez MC, Stiles S, *et al.* 2014b. Is DNA barcoding actually cheaper and faster than traditional morphological methods? Results from a survey of freshwater bioassessment efforts in the United States. *PLoS ONE* 9: e95525.
- Soranno PA, Cheruvilil KS, Elliott KC, and Montgomery GM. 2015. It's good to share: why environmental scientists' ethics are out of date. *BioScience* 65: 69–73.
- Taberlet P, Coissac E, Hajibabaei M, and Rieseberg LH. 2012. Environmental DNA. *Mol Ecol* 21: 1789–93.
- van Dijk EL, Auger H, Jaszczyszyn Y, and Thermes C. 2014. Ten years of next-generation sequencing technology. *Trends Genet* 30: 418–26.
- Visco JA, Apothéloz-Perret-Gentil L, Cordonier A, *et al.* 2015. Environmental monitoring: inferring the diatom index from next-generation sequencing data. *Environ Sci Technol* 49: 7597–605.
- Vörösmarty CJ, McIntyre PB, Gessner MO, *et al.* 2010. Global threats to human water security and river biodiversity. *Nature* 467: 555–61.

Article annexe II

“The potential of high throughput sequencing (HTS) of natural samples as a source of primary taxonomic information for reference libraries of diatom barcodes”

(accepté dans le journal *Fottea*, à paraître en 2018)

Rimet F.^{1,2}, Abarca N.³, Bouchez A.^{1,2}, Kusber W.H.³, Jahn R.³, Kahlert M.⁴, Keck F.⁴, Kelly M.⁵, Mann D.G.⁶, Piuz A.⁷, Trobajo R.⁸, Tapolczai K.^{1,2}, Vasselon V.^{1,2}, Zimmermann J.³

¹ INRA – UMR Carrel, 75 av. de Corzent - BP 511, FR-74203 Thonon-les-Bains cedex, France

² Université de Savoie, UMR CARREL, 73011 Chambéry, France

³ Botanischer Garten und Botanisches Museum Berlin-Dahlem, Freie Universität Berlin, Königin-Luise-Str. 6-8, 14195 Berlin, Germany

⁴ Swedish University of Agricultural Sciences, Department of Aquatic Sciences and Assessment, PO Box 7050, SE- 750 07 Uppsala, Sweden

⁵ Bowburn Consultancy, 11 Monteigne Drive, Bowburn, Durham DH6 5QB, UK

⁶ Royal Botanic Garden Edinburgh, Edinburgh, EH3 5LR, Scotland, UK

⁷ Muséum d’Histoire Naturelle, Route de Malagnou 1, Case postale 6434, CH-1211 Genève 6, Switzerland

⁸ Aquatic Ecosystems, Institute for Food and Agricultural Research and Technology (IRTA), Crta de Poble Nou Km 5.5, Sant Carles de la Ràpita, Catalonia, Spain

Abstract

Diatoms are used routinely to assess pollution level from rivers and lakes. Current methods are based on identification by light microscopy, which is laborious. An alternative is to identify species based on short DNA fragments and high-throughput sequencing (HTS). However a potential limitation is the incomplete coverage of species in reference barcode libraries. Usually these libraries are compiled by isolating cells, before culturing and sequencing them, which is tedious and often unsuccessful. Here we propose the use of *rbcL* sequences from environmental samples analysed by HTS. We set several criteria to ensure good sequence quality and correspondence with the target species observed in microscopy: the sequence needed to be abundant in the sample, and with no insertions nor deletions or stop codon, phylogenetic neighbour taxa had to correspond to neighbour taxonomic taxa expected from morphological observations. Four species from tropical rivers are given, including one that is new to science.

Key words

Algae, Bacillariophyta, biomonitoring, data traceability, DNA barcoding, eDNA, ecosystem assessment, metabarcoding, pollution, Water Framework Directive.

Introduction

Human activities have had an impact on the environment and in particular on freshwater ecosystems for a long time. Increased fish mortality in the Rhine and Thames in the late 17th and 18th centuries drew attention to the impact of pollution on aquatic ecosystem health (MARKERT *et al.* 2003) and, by the mid-19th century, several authors had observed that the composition of microscopic organisms in polluted aquatic ecosystems was different to that in unpolluted habitats (COHN 1853). In 1908, KOLKWITZ and MARSSON demonstrated a clear relationship between water quality and organisms, including microalgae, in freshwaters (KOLKWITZ & MARSSON 1908). Microalgae are often the dominant primary producers of aquatic ecosystems. They display huge taxonomic diversity, and diatoms alone are estimated to have over 100,000 extant species (MANN & VANORMELINGEN 2013). This diversity, coupled with a high sensitivity to their chemical environment and wide distribution makes them excellent ecological indicators (STEVENSON 2014). In the 1950s several authors (e.g. HUSTEDT 1957; ZELINKA & MARVAN 1961; BUTCHER 1947) started to use diatoms for practical assessment of pollution. Interest grew over subsequent decades and diatoms are now widely used for ecological assessment in both Europe (e.g. RIMET 2012; KELLY *et al.* 2014) with the Water Framework Directive and the US (e.g. BARBOUR *et al.* 1999; POTAPOVA & CHARLES 2007; HAUSSMANN *et al.* 2016) with the Clean Water Act.

Until now standard methods for freshwater ecological assessment based on diatoms (e.g. EUROPEAN COMMITTEE FOR STANDARDISATION 2014, 2016) have required the biofilms on submerged surfaces to be sampled and then the species present in this biofilm to be identified and counted using their siliceous exoskeleton under light. This requires time and highly trained analysts with good knowledge of the taxonomic literature. This limits output per analyst to no more than 200–300 analyses per year for the most taxonomically-intensive methods, which is a bottleneck for ecological monitoring (HAJIBABAEI *et al.* 2016). Moreover, there can be considerable inter-analyst variation (e.g. BESSE-LOTOSKAYA *et al.* 2006; KAHLERT *et al.* 2009, 2012).

So-called “DNA-Barcoding” (HEBERT *et al.* 2003) has been proposed as an alternative to microscopical identification, using DNA sequencing to recognise diatom species. This concept was expanded to environmental samples, “DNA-metabarcoding” (POMPANON *et al.* 2011), where species are identified in natural samples using their DNA. Several studies have demonstrated that this approach may be applicable to river diatoms (KERMARREC *et al.* 2013b; ZIMMERMANN *et al.* 2015;

Visco *et al.* 2015). However, these studies also showed that a potential limitation of metabarcoding was an insufficient coverage of existing taxa of the reference barcoding library. This needs to be as complete as possible and must be curated to maintain its quality (i.e. taxonomic homogeneity of assignments, sequence quality and traceability of data and metadata; (RIMET *et al.* 2016; KUSBER *et al.* 2012). Barcodes in such libraries are obtained in several ways. Firstly, through single-cell isolations, culturing, and Sanger sequencing (e.g. EVANS *et al.* 2007; TROBAJO *et al.* 2009; ZIMMERMANN *et al.* 2014a; ABARCA *et al.* 2014). However, cell isolation is a long process which can in many cases be unsuccessful since some species appear not to thrive under culture conditions. Moreover, some species in cultures show deformations of their frustule, which can make them difficult to identify especially if they have been cultivated for a long time. Another drawback is the rapid cell size reduction of some species in culture which also complicates identification (e.g. KI *et al.* 2008). As a result of a combination of these factors, many species that are important for ecological assessment have not yet been sequenced.

A second means of obtaining barcodes is single-cell PCR which can be used to obtain nucleotide sequence for taxa that can not be cultured (TAKANO & HORIGUCHI 2006; GOMEZ *et al.* 2012). Single-cell extraction/PCR has recently been applied to diatoms by HAMILTON *et al.* (2015) and KHAN-BUREAU *et al.* (2016), but in these cases, identifications were carried out only on living cells, which, in most cases, may prevent correct species identification (HAMILTON *et al.* 2015). Moreover, majority of sequenced diatoms were relatively large and small-celled taxa (10–20 µm) were excluded.

A final means of obtaining barcodes for reference libraries is direct sequencing of environmental samples in order to avoid the laborious procedures involved in the first two methods (isolation, culturing, single cell sorting). Examples of two different approaches have already been published. One is simple direct Sanger sequencing of samples presenting very low species diversity, as in some Chilean rivers with *Didymosphaenia geminata* (LYNGBYE) MART. SCHMIDT blooms (JARAMILLO *et al.* 2015). However, this method is relatively imprecise and several *rbcl* sequences published do not correspond to the targeted species announced in the paper (e.g. NCBI accession numbers: KR066780, KR066784). Another approach is a PCR of the environmental sample followed by cloning (e.g. KHAN-BUREAU *et al.* 2016).

In this paper, we present another means of enriching barcode reference libraries. In this case environmental samples were sequenced using High-Throughput Sequencing (HTS) and the outputs compared with the results of light and electron microscope analyses of the same samples. There are several potential benefits compared to previous studies. First, a greater number of sequences per sample than Sanger sequencing should be accessible with HTS. Second, a much bigger number of samples can be sequenced with HTS and thereby reduce costs. The challenge lies in selecting sequences from environmental samples that correspond to the species of interest. The questions we want to address are:

1. Is it possible to relate environmental sequences to the target species observed by microscopy and to do so with high reliability?
2. What are the advantages/disadvantages associated with the use of sequences from uncultured diatoms?
3. Which material, data and metadata must be stored with these sequences to ensure good traceability?

Several examples will be given from Mayotte island, a French tropical island of the Comoros archipelago situated in the Mozambique Channel (Africa), where 98 environmental samples from rivers were sequenced using a HTS technology, Ion-Torrent PGM, and also observed with light and scanning electron microscopy. The open-access barcoding library R-Syst::diatom (RIMET *et al.* 2016) is used as host database to store the data (www.rsyst.inra.fr/) presented in this paper and the Thonon Culture Collection (TCC) to store the material (www6.inra.fr/cartel-collection_eng/), as well as the Botanischer Garten und Botanisches Museum Berlin-Dahlem of Berlin (Germany) and the Conservatoire et Jardin Botaniques of Geneva (Switzerland).

Materials and Methods

Study area and sampling methodology: Mayotte (France) is a tropical island with a surface of 374 km², located in the Indian Ocean to the northwest of Madagascar and to the east of Mozambique (12°50'35"S 45°08'18"E) (Fig. 1). Geologically it is of volcanic origin and is part of the Comoros archipelago. It consists of two main parts, the smaller Petite-Terre (11 km²) and the Grande-Terre (363 km²) where the study was performed. The main pressures on its rivers are related to the fast growing population (226,915 inhabitants in 2015; INSEE 2016). Samples were collected as part of river pollution assessments; results are reported to the European Commission as part of France's obligations under the Water Framework Directive (EUROPEAN COMMISSION, 2000). For this study, samples were collected during the dry season in July and August 2015.

The sampling procedure followed European standards (EUROPEAN COMMITTEE FOR STANDARDISATION 2014 A,B). Benthic diatoms were collected from at least five stones from the fast-flowing parts of sampling sites. The upper surfaces of the stones were scrubbed with a toothbrush in order to collect the biofilms. The samples were then fixed with ethanol to give a final concentration of at least 70%.

Preparation for microscopy: Diatom valves were cleaned using 40% H₂O₂ and 40% HCl. Cleaned valves were mounted in a resin (Naphrax®, Brunel Microscopes: <http://www.brunelmicroscopes.co.uk/>) with a high refractive index to create permanent slides. For scanning electron microscopy (SEM), dried cleaned diatom valves were coated with gold using a Cressington 108 auto sputter coater® and examined with a Zeiss DSM940A © in the Natural History Museum of Geneva (Switzerland). Samples were analysed by light microscopy as part of a parallel study (TAPOLCZAI *et al.* 2016). Four samples with the low number of species were selected.

DNA extraction: These four environmental samples were centrifuged at 13,000 rpm (equivalent to 18000 × *g*) for 30 min, the supernatant was then removed. 25 mg of wet pellet was used for each sample. DNA extraction was based on the Sigma–Aldrich GenElute™-LPA DNA protocol which was used in previous studies (KERMARREC *et al.* 2014; CHONOVA *et al.* 2016; VASSELON *et al.* 2017). The final elution volume was 40 µL.

Preparation of the library of amplicons and HTS sequencing: For all four samples, HTS sequencing of a 312 bp fragment of *rbcL* (exact region is situated between primers given below) was performed. For each DNA sample, PCR amplification was performed in three replicates on 1 µL of extracted DNA in a mix (25µL final volume) containing: 0.75 Unit of Takara LA Taq® polymerase (for 25 µL of final volume, 0.15 µL of Taq at 5Unit/L), 2.5 µL of 10× LA PCR Buffer II (Mg² plus), 1.25 µL of 10 µM of forward and reverse primers, 1.25 µL of 1.51.10⁻⁴ mol/L BSA (Bovine Serum Albumin), 2 µL of 2.5mM dNTP and completed with 15.6 µL H₂O (molecular biology grade). The primer pair *Diat_rbcL_708F* (STOOF-LEICHSENRING *et al.* 2012) and R3 (BRUDER & MEDLIN 2007) was modified to amplify a broader diversity of diatom as follows: forward primer combine an equimolar mix of *Diat_rbcL_708F_1* (AGGTGAAGTAAAAGGTTTCWTACTTAAA), *Diat_rbcL_708F_2* (AGGTGAAGTTAAAGGTTTCWTAYTTAAA) and *Diat_rbcL_708F_3* (AGGTGAACTAAAGGTTTCWTACTTAAA); reverse primer combine an equimolar mix of R3_1 (CCTTCTAATTTACCWACTG) and R3_2 (CCTTCTAATTTACCWACAACAG). PCR reaction conditions were as follows: initial denaturation of DNA at 95°C for 15 min followed by 40 cycles with 45 s denaturation at 95°C, followed by 45 s annealing at 55°C and 45 s extension at 72°C. One no-template control (NTC) was used as a negative control.

The 3 replicates of PCR amplicon for each sample were then pooled and cleaned with Agencourt AMPure beads (Beckman Coulter, Brea, USA) following the manufacturer's instructions, except that a 1.5:1 beads:DNA ratio was used specifically to purify the 312 bp fragment. Purified

amplicons were assessed for quality and quantified using the 2200 TapeStation (Agilent technologies, Santa Clara, USA) with D1000 screen tape and reagents. The purified amplicons were used to prepare four DNA libraries for HTS with Ion Torrent technology using the NEBNext® Fast DNA Library Prep set for Ion Torrent™ (BioLabs, Ipswich, USA) following the manufacturer protocols for End repair, PCR amplification of adapter ligated DNA (7 cycles) and cleaning steps. Ligation of library adapters to purified amplicons was done using 2µL of P1 adapter (NEB kit) and 2µL of A-X tag adapter provided in Ion Express™ Barcode adapters (Life Technologies, Carlsbad, USA) using 1 tag per amplicon.

The quality, size and concentration of the libraries were checked using the 2200 TapeStation with D1000 High Sensitivity screen tape and reagents. Each library was diluted to 100pM and all were pooled together with 50 libraries from other environmental samples from Mayotte in a unique mix sequenced using 1 Ion 318™ Chip Kit V2 (Life Technologies, Carlsbad, USA) on a PGM Ion Torrent machine by the "Plateforme Génome Transcriptome" (PGTB, Bordeaux, France).

Sequence data processing: Demultiplexing and adapter removal steps were made by the Sequencing Platform which provided a single fastq file for each of the 55 libraries. DNA reads were filtered for length and quality using mothur software (SCHLOSS *et al.* 2009) in every fastq file with the following settings: a minimum length of 250bp, a Phred quality score higher than 23 over a moving window of 25 bp, a maximum of 1 mismatch in the forward primer sequence, homopolymers shorter than 8 bp, and absence of ambiguous bases. Reads which were not fully aligned with the *rbcl* barcode were removed. The resulting files were analysed together. Denoising of sequencing error was performed with the Precluster command by creating read clusters allowing one nucleotide difference between DNA reads. Chimera removal was done with UCHIME algorithm (EDGAR *et al.* 2011).

The R-Syst::diatom database (RIMET *et al.* 2016) (database version v5, <http://www.rsyst.inra.fr/en>), restricted to our 312bp *rbcl* barcode, was used as the reference database. Taxonomic assignment of DNA reads at species level was made using this reference database and the Naïve Bayesian method (WANG *et al.* 2007) with a confidence score threshold of 85%. Only DNA reads assigned to the Bacillariophyta phylum (diatoms) were used in further analyses.

After dereplication, uncorrected pairwise distances were calculated between aligned reads to generate a similarity distance matrix. Based on this distance matrix, reads were clustered in Operational Taxonomic Units (OTUs) using the Furthest Neighbour algorithm at 100% similarity level in order to have each OTU represented by a single sequence. Singletons were removed.

Then, for each of the four low-diversity samples, a Blastn was run on the entire NCBI database with each of the 15 to 20 most abundant 312-bp sequences. 312-bp sequences showing a BLAST result congruent with the microscopical identification of targeted species were kept for subsequent phylogenetic analyses.

Phylogenetic analyses: For each of the four samples, the selected 312-bp sequences were aligned with a selection of sequences from R-Syst::diatom database using Muscle (EDGAR 2004) in Seaview (GOUY *et al.* 2010). This selection of sequences was done based on their taxonomic proximity to the 312-bp sequences. The lengths of the Sanger sequences from R-Syst::diatom were at least 1000 bp. The best substitution model was then tested in MEGA7 (KUMAR *et al.* 2016). A first phylogenetic tree was calculated following the best substitution model with raxmlGUI (SILVESTRO & MICHALAK 2012) and the sequence selection of R-Syst::diatom, after which we calculated a second phylogenetic tree, adding the 312-bp sequences in the phylogeny and enforcing its topology with the topology of the first tree. This analysis is also available in raxmlGUI under the "enforce constraint menu" and "define topological constraint". Trees were drawn in MEGA7.

Material and data accessibility: All material is accessible through the Thonon Culture Collection (TCC: https://www6.inra.fr/carrtel-collection_eng/) and at the Botanischer Garten und Botanisches Museum Berlin-Dahlem, Freie Universität Berlin (B) and also at the Conservatoire et Jardin Botaniques of Geneva (G) (for the newly described species, all nomenclatural acts have been submitted to PhycoBank for registration of new scientific names; <http://phycobank.org>). TCC culture collection hosts algal cultures, but also uncultured samples containing species of interest (as permanent slides, and raw and treated material). All metadata are stored in the open-access R-Syst::diatom reference database; a detailed description of this database and its management is given in RIMET *et al.* (2016). Taxonomy, sequences, photos, sampler names, phenotypic data, etc., can be consulted and downloaded at: <http://www.rsyst.inra.fr/>

Results

Table 1 and Fig. 1 give the sample locations, their environmental characteristics and the dominant target species.

Morphology and ecology

Halamphora ghanensis LEVKOV (Fig. 2)

This sample is registered and conserved in the TCC (TCC956 - uncultured sample) and in the Botanischer Garten und Botanisches Museum Berlin-Dahlem, Freie Universität Berlin (B 40 0041830).

The morphology of the valves from the Gouloué river (Fig. 1, site 2) fitted the description of *H. ghanensis* (LEVKOV 2009): valve length from 25.2 to 26.4 μm (24 to 27 μm in LEVKOV), width 5.6 to 6.4 μm (5 to 5.6 μm in LEVKOV 2009), 13 to 14 dorsal striae/10 μm (14 to 16 in LEVKOV 2009). We do not think the small differences between measured width and stria density are sufficiently different from the original description to suggest a different species. Morphological features observable in SEM also correspond to the description given in LEVKOV (2009): dorsal ledge crenulated, dorsal striae biseriate and interrupted by longitudinal bars near the dorsal margin. Striae also appear biseriate internally. There is a poorly developed helictoglossa in the distal raphe endings and fused central helictoglossae at the proximal raphe endings, as in the species description (LEVKOV 2009). A single row of dorsal areola was observed near the raphe in internal view (a feature not present in *H. acutiuscula* (KÜTZING) LEVKOV, which is otherwise morphologically close to *H. ghanensis*).

The sample was collected from Gouloué river, near Passamainty city (Table 1), a polluted river surrounded by a village where many houses discard their wastewater directly into the river (2.5 mg/L of dissolved organic carbon, 0.08 mg N/L of NH_4^+ , 113 mg/L O_2 chemical oxygen demand). This section of river is subject to tidal influence, which explains the high conductivity (2260 $\mu\text{S}/\text{cm}$) and chloride concentration (706 mg/L Cl^-) measured on the day of sampling. We would emphasise, however, that *H. ghanensis* is regularly observed in samples from Mayotte island, in river stretches that are not tidal and where the conductivity is < 300 $\mu\text{S}/\text{cm}$. Usually, it forms < 10% of the diatom assemblage (based on 400 valves counted per sample) whereas in the Gouloué river it was abundant (38% of the valves counted). We therefore suspect this species prefers brackish waters rich in organic matter. In Levkov (2009), *H. ghanensis* was reported from a river in Ghana (West Africa), but no chemical measurements were given.

Gomphonema clavatuloides RIMET, D.G. MANN, TROBAJO et N. ABARCA, *sp. nov.* (Fig 3)

Description: Valve lanceolate-clavate, with the broadest portion of the valve at the central nodule; apex and base rounded. Axial area narrow, linear. Central area small, transversely elongated,

made by slight shortening of central striae on both valves sides. Stigma present at the end of a shortened central striae. Length from 25 to 41 μm , breadth from 5.5 to 7.0 μm , striae density from 7 to 9 in 10 μm . Transapical striae moderately radiate, more parallel towards the poles. Areolae lineolate in external view: in internal view, the areolae have the same shape and lie in a furrow. The central punctum has a round opening in external view and is lineolate in internal view. The internal central raphe ends are hooked towards the primary side of the valve (i.e. in the opposite direction to the external distal ends). The external central raphe ends are slightly deflected towards the primary side of the valve and terminate in a drop shaped expansion. Distally, the raphe ends in terminal fissures that are slightly bent towards the secondary side. Both internal distal raphe ends terminate straight in helictoglossae. Four pores are present on the mantle in extension of each valve stria.

Holotype: B 40 0041829 (Botanischer Garten und Botanisches Museum Berlin-Dahlem, Freie Universität Berlin, Germany) represented by Fig. 3.

Isotype: Conservatoire et Jardin Botaniques, Geneva, Switzerland, reference number: G00260989 This sample (slides, raw material) is also registered and conserved in the Thonon-Culture-Collection of the INRA: TCC955 - uncultured sample deposited in Thonon-les-Bains, France.

Type locality: Songaro Mbili river near Dembeni city (France, Mayotte Island, Northern Mozambique Channel), sampled 24 July 2015 by F. RIMET, coordinates: $-12^{\circ}50'23.3''$ $45^{\circ}10'28.3''$ (Fig. 1 site 3).

Name registration: <http://phycobank.org/100011>

Etymology: The specific epithet refers to the close resemblance to *G. clavatum* EHRENBERG

Similar taxa: This taxon resembles the *G. clavatum* species complex, which also includes similar species and varieties such as *G. clavatum* E. REICHARDT and *G. subclavatum* (GRUNOW) GRUNOW (KRAMMER & LANGE-BERTALOT 1986; REICHARDT 1999). *Gomphonema clavatuloides* has the same general clavate shape as *G. clavatum* and *G. subclavatum* but differs in its breadth (5.5–7.0 μm : compare 4.7–5.7 μm for *G. clavatum* and 8–10 for *G. subclavatum*) and striae density (7–9 in 10 μm ; compare 10.5–14 μm for *G. clavatum* and 9–13 for *G. subclavatum*). It might also be confused with small or medium-sized valves of *G. paludosum* E. REICHARDT but differs in striation pattern and number of striae in 10 μm (REICHARDT 1999).

Ecology: This sample was collected from Songaro Mbili river (site 3, Fig. 1), near Dembeni city, a polluted stream surrounded by subsistence farming and villages; several houses discard their wastewaters directly in the river. The water was almost stagnant with traces of soap (women wash clothes at this place). Conductivity was 280 $\mu\text{S}/\text{cm}$, chemical oxygen demand was 66 mg/L O_2 (measured in August 2014), but the dissolved phosphorus concentration was low (0.014 mg of P $\text{PO}_4^{2-}/\text{L}$). In this sample, *G. clavatuloides* was the dominant species (85.6% of the counted valves).

***Epithemia hirudiniformis* (O. MÜLLER) RIMET, D.G. MANN, R. TROBAJO, J. ZIMMERMANN et R. JAHN, comb. nov. (Fig. 4)**

Basionym: *Rhopalodia hirudiniformis* O. MÜLLER, Botanische Jahrbücher für Systematik, Pflanzengeschichte und Pflanzengeographie 22, p. 67, pl. I: figs 40-46, 51, 52; pl. II: figs 15-17 (1895).

Name registration: <http://phycobank.org/100012>

Taxonomy and morphology: This sample is registered and conserved in TCC (TCC954 – uncultured sample) and in the Botanischer Garten und Botanisches Museum Berlin-Dahlem, Freie Universität Berlin (B 40 0041831).

We propose a new combination for *Rhopalodia hirudiniformis* in *Epithemia* since, according to RUCK *et al.* (2016), *Rhopalodia* is paraphyletic with respect to *Epithemia*. When

Rhopalodia and *Epithemia* are combined in a single monophyletic genus, the correct name is *Epithemia*.

A few species and varieties belonging to former *Rhopalodia* are heteropolar, in particular, *R. hirudiniformis* and its varieties var. *parva* O. MÜLLER, var. *turgida* FRICKE, and also the species *R. rhopala* (EHRENBERG) HUSTEDT. All these taxa were described from east African lakes (COCQUYT 1998).

The length of the valves found in the Soulou waterfall (site 1, Fig. 1, Fig. 4) ranged from 73.6 to 215 µm and the maximum width from 7 to 14 µm. However, the smallest width measurements may be erroneous as the valves were not always flat during light microscopy. In our sample, the striae and fibulae densities were 11-12 and 5–6 in 10 µm, respectively. These densities correspond to the description of *R. rhopala*, *R. hirudiniformis* and its varieties. However, the length of the frustules in our sample corresponds to *R. rhopala* and *R. hirudiniformis*. Indeed, according to the valve sizes given in COCQUYT (1998) the largest valves (220 µm) correspond to *R. rhopala* while most of the other valves fall into the size range of *R. hirudiniformis* (58–113 µm). The distribution of the length and width measures followed a normal distribution ($p > 5\%$, measurements carried out on 28 valves) so we cannot consider this population as belonging to two different species. Finally, the clear constriction observed on *R. hirudiniformis* var. *turgida* does not correspond to the morphology observed in this sample. For these reasons we decided to identify the valves observed in our sample as *R. hirudiniformis*.

Ecology: The Soulou waterfall (site 1, Fig. 1) comes from the Chirini river and falls onto a beach in the north west part of Mayotte Island. The sample was taken on the vertical wall of the waterfall where thick aerial biofilms were visible. *E. hirudiniformis* was associated with several species of cyanobacteria (*Oscillatoria* sp., *Pseudanabaena* sp., *Chroococcus* sp., *Plectonema* sp.) and was the only diatom species observed. Moreover, four to eight cells of endosymbiotic cyanobacteria could be observed in each cell of *E. hirudiniformis*. There is a sampling station on the Chirini river 500 m upstream of the waterfall, which shows a quite good water quality (230 µS/cm conductivity, 30 mg/L O₂ chemical oxygen demand, less than 0.1 mg/L of N-NO₃²⁻, 0.01 mg/L of P-PO₄²⁻).

***Gomphonema parvulum* (KÜTZING) KÜTZING *sensu lato* (Fig. 5)**

Taxonomy and morphology: This sample is registered and conserved in TCC (TCC958 - uncultured sample) and in the Botanischer Garten und Botanisches Museum Berlin-Dahlem, Freie Universität Berlin (B 40 0041832).

Several samples in Mayotte island contained high abundances of *G. parvulum sensu lato* (encompassing *G. narodoense* R. JAHN, N. ABARCA, J. ZIMMERMANN et ENKE, *G. lagenula* KÜTZING, *G. parvulum*, and varieties). One of these (Mouala river near Mirereni city) showed significant morphological and genetic diversity. Its morphology is given in Figure 5. Figs. 5.1 to 5.7 correspond to the morphology of *G. lagenula*, indeed, the shape of the head pole is consistently more rostrate to capitate than in *G. parvulum sensu stricto* and the general shape of the frustule is lanceolate, as described in ABARCA *et al.* (2014). Figs. 5.8 to 5.16 have a rather different shape; indeed the linear to lanceolate valve is more slender and much less clavate than *G. lagenula*. This second morphodeme could fit *G. parvulum* var. *parvulum* morphodeme *exilissimum* (strain D12_022) according to the information given by ABARCA *et al.* (2014). Nevertheless, the stria density of the valves ranged from 9 to 12 in 10 µm and does not correspond to the density in *G. exilissimum* (GRUNOW) LANGE-BERTALOT (12–14 in 10 µm, in HOFMANN *et al.* 2011).

In this sample, another *Gomphonema* species was observed under microscope, *G. bourbonense* E. REICHARDT.

Ecology: This sample was collected from the Mouala river (site 4, Fig. 1) near Mirereni city, a polluted river receiving wastewaters from several houses. When the diatoms were sampled, the water was deoxygenated (24% O₂ saturation); conductivity was 120 µS/cm and chemical oxygen demand was 76 mg/L O₂.

Phylogeny

***Halamphora ghanensis* (Fig. 6)**

1751 different environmental sequences were obtained from the sample from the Gouloué river, near Passamainty city, of which 432 were singletons. The 15 most abundant sequences were blasted on NCBI. Two sequences had high similarity with *Halamphora montana* (KRASSKE) LEVKOV (96%) and their length was 312 bp. These sequences were selected to build a phylogeny. The list of sequences used as constraints for the 312 bp sequences is given in Supplementary data S1 and the alignment used in the phylogeny is in Supplementary data S5. No indels or stop codon appeared in the 312 bp sequences after alignment. The best model selection was the GTR+G+I model (General Time Reversible model with gamma distribution and Invariable sites) which showed the lowest AICs (Akaike Information Criterion, corrected). A maximum likelihood constrained phylogeny with GTR+G+I model and 1000 rapid bootstraps was calculated and is drawn in Fig. 6.

The two 312 bp sequences are included in a group supported by a high bootstrap value (97%). This clade is composed of representatives of *Amphora* and *Halamphora*. Three of the *Amphora* species included were described by WACHNICKA & GAISER (2016), and STEPANEK & KOCIOLEK (2014) have already suggested that they should be transferred into *Halamphora*. These are *Amphora semperpalorum* WACHNICKA ET GAISER, *A. subtropica* WACHNICKA ET GAISER and *A. caribaea* WACHNICKA ET GAISER. *Amphora hyalina* KÜTZING, also included in this group, was described by KÜTZING (1844). The results of our phylogeny confirm the proposition of STEPANEK & KOCIOLEK (2014) and the morphological features of these species correspond to the description of *Halamphora* by LEVKOV (2009). This clade is therefore composed only by *Halamphora* species and we consider that the two 312 bp sequences can be kept in R-Syst::diatom. Moreover we suggest the following new taxonomic combinations:

Halamphora semperpalorum (WACHNICKA ET GAISER) RIMET ET R. JAHN, *comb. nov.*

Basionym: *Amphora semperpalorum* WACHNICKA ET GAISER, Diatom Res. 22: 403, figs 44–48 (2007).

Name registration: <http://phycobank.org/100014>

Halamphora hyalina (KÜTZING) RIMET ET R. JAHN, *comb. nov.*

Basionym: *Amphora hyalina* KÜTZING. Die Kieselalgen Bacillarien oder Diatomeen. W. Köhne, Nordhausen. p. 108, fig. 30/18. (1844).

Name registration: <http://phycobank.org/100016>

Halamphora subtropica (WACHNICKA ET GLAISER) RIMET ET R. JAHN, *comb. nov.*

Basionym: *Amphora subtropica* WACHNICKA ET GLAISER, Diatom Res. 22: 407, figs 64–70 (2007).

Name registration: <http://phycobank.org/100018>

Halamphora caribaea (WACHNICKA ET GAISER) RIMET ET R. JAHN, *comb. nov.*

Basionym: *Amphora caribaea* WACHNICKA ET GAISER, Diatom Res. 22: 399, figs 35–37 (2007).

Name registration: <http://phycobank.org/100020>

***Gomphonema clavatuloides* (Fig. 7)**

1876 different environmental sequences were obtained from the sample from Songaro Mbili, near Dembeni city, of which 153 were singletons. The 20 most abundant sequences were blasted on NCBI and R-Syst::diatom. 10 sequences had high similarity with *Gomphonema acuminatum* EHRENBERG (95–96%) and their length was 312 bp. Three other sequences had also high similarities with species of *Gomphonema* genus: *G. bourbonense* (one sequences matching with *rbcl*

sequences of strains TCC441, TCC460, TCC513, TCC450, TCC514, TCC453, TCC452, TCC451) and *G. lagenula* (two sequences matching with *rbcl* sequences of strains TCC470, TCC500, TCC440, TCC432, TCC431, TCC429 and NCBI sequences HG530055, HG530054) ; these sequences were rejected from the further analyses. Sequences showing high similarities with *G. acuminatum* were selected to build a phylogeny. The list of sequences used as constraint for the 312 bp sequences is given in Supplementary data S2 and the alignment used in the phylogeny is in Supplementary data S6. No indels and no stop codon appeared in the 312 bp sequences after alignment. The best model was the GTR+G+I model which showed the lowest AICs. A maximum likelihood constrained phylogeny with GTR+G+I and 1000 rapid bootstraps is given in Fig. 7.

The ten 312 bp sequences form a monophyletic clade supported by a high bootstrap value (93%). The nearest species is *G. subclavatum* var. *mexicanum* (GRUNOW) R.M. PATRICK. Therefore these ten 312 bp sequences can be kept for R-Syst::diatom.

***Epithemia hirudiniformis* (Fig. 8)**

959 different environmental sequences were obtained from the sample from Soulou waterfall, none of which were singletons. The 15 most abundant sequences were blasted on NCBI. 14 sequences had high similarity with *Epithemia gibba* KÜTZING (94–96%) and their length was 290 bp. The 15th had high similarity with *Ulnaria ulna* (NITZSCH) COMPÈRE. This last sequence accounted for 0.5% of the total abundance of the sequences. The list of sequences used as constrain for the 290 bp sequences is given in Supplementary data S3 and the alignment used in the phylogeny is in Supplementary data S7. After alignment, 9 of the 290 bp sequences showed deletions and were removed from the following analyses. 5 sequences were selected and showed no stop codon. The best model selection was the GTR+G+I model which showed the lowest AICs. A maximum likelihood constrained phylogeny with GTR+G+I model and 1000 rapid bootstraps is given in Fig. 8.

The five 290 bp sequences form a monophyletic clade supported by a high bootstrap value (100%). They are inside a larger clade composed of other *Epithemia* species and supported by a high bootstrap value (79%). Therefore only five sequences out of the fourteen 290 bp sequences were kept for R-Syst::diatom.

***Gomphonema parvulum* (Fig. 9)**

1053 different environmental sequences were obtained from the sample from the Mouala river near Mirereni city of which 516 were singletons. The 20 most abundant were blasted on NCBI. 14 sequences had high similarity with *Gomphonema parvulum* sensu lato and their length was 275 bp. These were used to build a phylogeny. The list of sequences used as constraint for the 275 bp sequences is given in Supplementary data S4 and the alignment used in the phylogeny is in Supplementary data S8. No indels and no stop codon appeared in the 275 bp sequences after alignment. The best model selection was the GTR+G+I model which showed the lowest AICs. A maximum likelihood constrained phylogeny with GTR+G+I model and 1000 rapid bootstraps is given in Fig. 9.

Gomphonema parvulum sensu lato (including *G. parvulum* sensu stricto, *G. saprophilum* (LANGE-BERTALOT et E. REICHARDT) ABARCA, R. JAHN, J. ZIMMERMANN et ENKE, *G. narodoense*, *G. lagenula*) forms a large group supported by 61% bootstrap value. Inside this large group several well supported smaller groups are present: a group with *G. lagenula* strains, a group with *G. saprophilum* strains, a group with *G. parvulum* sensu stricto, a group with *G. narodoense*. However, the position in the tree of several sequences (apart those of 275 bp) is not well supported.

The 275 bp sequences are present in several groups. Nine sequences are included in the *G. lagenula* group with a bootstrap support of 52%; however, the position of three 275 bp sequences inside the *G. parvulum* sensu lato group is not well supported. Therefore, given the low

support of the position of these sequences, we decided not to keep these sequences for R-Syst::diatom.

Discussion

1. Is it possible to relate sequences from HTS analyses of biofilms to a target species observed by light microscopy with high reliability?

Relating sequences to morphological features with high reliability is not an issue in the case of monoclonal cultures since Sanger sequencing delivers a single sequence. Such certainty is, however, harder to obtain when using HTS to sequence biofilm samples collected from the natural environment.

The simplest example presented here is *Epithemia hirudiniformis*, since this was the only species found in microscopy and the 14 most abundant sequences corresponded to this genus. The 15th most abundant sequence matched to *Ulnaria ulna*. It was represented by 0.5% of the sequences in the sample which explains why *U. ulna* was not detected by light microscopy. Such a sample presenting a very low species diversity is similar to that described for *Didymosphenia geminata* in Chilean rivers (JARAMILLO *et al.* 2015) and it is consequently easy to relate the sequences found with high reliability to the target taxon.

Two other examples, *Halamphora ghanensis* and *Gomphonema clavatuloides*, represent a slightly more complicated situation as each sample included several other species. In both samples six different species were identified using microscopy, while for metabarcoding seven taxa were detected for the *H. ghanensis* sample (3 species and 4 generic assignments where the species could not be determined) and 22 taxa for the *G. clavatuloides* sample (16 species and 6 generic assignments where the species could not be determined). Retrieving the sequences of *H. ghanensis* was possible because it was the only member of the Catenulaceae MERESCHKOWSKY in this sample and several reference barcodes of species belonging to this family are present in R-Syst::diatom. Results of BLAST and of the phylogeny confirmed that the sequences are belonging to this family. For *Gomphonema clavatuloides*, two other species belonging to *Gomphonema* were present in the sample (*G. bourbonensis* E. REICHARDT and *G. parvulum*). But here again, identifying the sequences of *G. clavatuloides* was possible, even if this species is new to science, because reference sequences of the two other *Gomphonema* species were available in the reference library and results of BLAST and phylogenies could confirm their membership.

In the case of *G. parvulum* (in the Mouala river) it was much harder to relate sequences to identified target species. This species complex has been studied for a long time (e.g. GEITLER 1972), it is found in rivers worldwide (e.g. MURAKAMI & KASUYA 1993; SILVA-BENAVIDES 1996; NDIRITU *et al.* 2006; RIMET 2009), and it shows significant phenotypic plasticity even in monoclonal culture (ROSE & COX 2014), which may explain why so many varieties and forms have been described. More recent studies integrating morphological and molecular data have clarified this species complex (KERMARREC *et al.* 2013a; ABARCA *et al.* 2014). In our case, several sequences belonging to several separate clades were present in a single sample and after a careful examination two different morphodemes could be distinguished, which may fit the descriptions of *G. lagenula* and *G. parvulum* var. *parvulum* [morphodeme *exilissimum*]. It was difficult, however, to relate these two morphodemes to the different clades which were, moreover, not supported with high bootstrap values in the *G. parvulum* phylogeny. The presence of several species belonging to the *G. parvulum* species complex in a single sample has already been observed by KERMARREC *et al.* (2013). This example shows the limits of the method we are proposing. Therefore these sequences were not included in the reference database R-Syst::diatom.

2. What are the advantages/disadvantages associated with use of uncultured diatom HTS sequences to enrich barcode reference libraries?

Unlike macroorganisms such as aquatic insects, where the DNA barcodes of individuals can be easily Sanger sequenced from a part of their body (e.g. insect legs in SWEENEY *et al.* 2011), diatom cells need to be isolated and cultured to get their barcodes, which is laborious and not always successful. Using HTS for natural samples encompassing several millions of diatom cells and several species makes it potentially possible to obtain thousands of barcode sequences in one HTS run.

One advantage of using HTS is shown by the examples of *Epithemia hirudiniformis*, *Halamphora ghanensis* and *Gomphonema clavatuloides* reported here. In each case, from 2 to 10 sequences were kept for each species in the reference database R-Syst::diatom. This gives an idea of the intraspecific genetic diversity of species. Obtaining such information with cultures is possible and has already been done (e.g. *Nitzschia palea* (KÜTZING) W. SMITH studied by RIMET *et al.* 2014; *Pseudo-nitzschia pungens* (GRUNOW ex CLEVE) HASLE by CASTELEYN *et al.* 2010) but requires intensive effort to isolate clones representing many clades.

The ease with which large numbers of sequences are generated using HTS is, however, balanced by some drawbacks. In particular, HTS produces more sequencing mistakes than classical Sanger sequencing (PAPARINI *et al.* 2015) and comparisons of the performances of different HTS technologies show different results. LOMAN *et al.* (2012) showed that MiSeq (Illumina) had the lowest error rates compared to 454 GS Junior (Roche) and Ion Torrent PGM (Life Technologies), but Ion Torrent had the highest throughput (80–100 Mb/h, compared to MiSeq 60 Mb/h and 454 9Mb/h) and 454 GS Junior gives the longest reads (up to 600 bases, compared to MiSeq with 150 bases and Ion Torrent PGM 300 bases). Such differences must be taken into account during data processing. When integrating sequences from HTS in reference databases, sequence quality criteria have to be defined. In this case, we set four criteria:

Criterion 1: only the 15–20 most often sequenced reads were selected because such sequences are more likely to be the best representatives of the population and also are likely to be sequences showing the lowest probabilities of sequencing mistakes (BRAGG *et al.* 2013).

Criterion 2: since *rbcl* is a coding region for a gene, no indels were accepted when aligning the HTS sequences with Sanger sequences from R-Syst::diatom. Sanger sequencing is, until now, considered to be the reference technology in terms of sequencing quality (e.g. MONTOYA *et al.* 2016; KHALIFA *et al.* 2016). Such problems were particularly common in the case of *Epithemia hirudiniformis*, where many sequences showed deletions, and were therefore not included in the reference database.

Criterion 3: since *rbcl* is a coding region, after translating the nucleotide sequence into proteins, no stop codon should appear. Those containing stop codons must be discarded. However, this problem did not occur for the most abundant sequences in our examples.

Criterion 4: the HTS sequences should show phylogenetic neighbours corresponding to the same neighbour taxa expected from morphological observations.

3. Which material data and metadata must be stored with these sequences in barcoding libraries to ensure good traceability?

The fundamental requirement suggested by ZIMMERMANN *et al.* (2014a) is that reliable identification of a taxon via DNA barcodes needs unambiguous agreement between genotype and phenotype/morphodeme with a valid binomial. Since this is not possible for HTS generated sequences, in contrast to sequences generated from unialgal cultures, we have to adapt and modify the requirements in order to guarantee a high standard for depositing these sequences.

In the case of HTS generated sequences from biofilm samples, the material data as well as the linked metadata have to be deposited in curated collections (herbaria such as B, BM, G, P, etc. see index of herbaria in HOLMGREN *et al.* 1990) and have to be available through public scientific databases (e.g. R-Syst, AlgaTerra, GGBN, GBIF, BOLD, INSDC). This includes eDNA material

deposition (e.g. DNA Bank Network / GGBN) and possible linked voucher material, information concerning the HTS methodology (e.g. DNA extraction, primers, PCR, library preparation, HTS chemistry, HTS platform, paired-end reads or not), details of the bioinformatics pipeline and the algorithms used (e.g. sequence trimming, chimera treatment, thresholds for OTU clustering, modus operandi of species assignment), availability of the raw reads. As for reference sequences from unialgal diatom cultures, ZIMMERMANN *et al.* (2016) suggest that metadata should include sampling localities and collectors, basic environmental data, high-resolution LM pictures, morphometrics, taxonomy and nomenclature, maps, literature as well as references to databases where this data is stored.

Even though a 100% unambiguous identification could not be made in the case of *Epithemia hirudiniformis*, the requirements for sound documentation are still applicable for a reference library, but the environmental origin of the sequence also has to be clearly highlighted. This also applies to the other examples, but they have to be treated with great care in order to ensure that a correlation between the presence of frustules and sequences is not just a coincidence. It has been shown in a number of cases that frequency of occurrence of a species within a sample analysed by light microscopy does not coincide with the frequency of occurrences of sequences in an environmental sample (e.g. JAHN *et al.* 2007). It is therefore essential that data are critically checked, the details of their creation are transparent and the pitfalls of the method discussed here are pointed out. We also recommend that sequences identified using these approaches are distinguished from those obtained from unialgal cultures in barcode libraries. Table 2 suggests terminology to prevent any confusion when such sequences are used in metabarcoding studies.

Conclusions: limits of the proposed methodology and perspectives

The methods and examples given here are clearly related to a particular application, which is routine assessment of aquatic ecosystems where species identification from natural samples is needed. We have shown that it is possible to enrich a reference barcoding library at low cost, taking advantage of sequencing data from routine samples collected for ecological assessment. Given the HTS technology used (PGM Ion Torrent), the sequence length was 312 bp. Such sequence length has been shown to be long enough for applied topics such as the one addressed here: diatom species identification for ecological assessment (e.g. KERMARREC *et al.* 2014; ZIMMERMANN *et al.* 2015; VISCO *et al.* 2015). It will be straightforward to expand such methodology to wider monitoring networks, if taxonomic experts are available who can follow the recommendations we propose here. Given that current reference barcoding libraries (and R-Syst::diatom in particular) cover only a small part of the diversity of diatoms in freshwater ecosystems, such HTS approaches may be essential if the libraries are to be completed quickly in order to be used for routine assessments.

On the other hand, one must make no mistake about the objective of this method. DNA barcodes are often inappropriate for phylogenetic studies, especially for defining deep nodes of classification trees (HAJIBABAEI *et al.* 2007). For such studies, much longer sequences are required and indeed, multigene approaches are now commonly recommended and are providing a much better understanding of diatom evolution (e.g. RUCK *et al.* 2016; THERIOT *et al.* 2015; NAKOV *et al.* 2014).

Finally, a possibility emerging from studies accumulating numerous DNA barcodes from particular species is that it can be a starting point of population genetics studies or can give indications of genetic diversity at an intra-specific level (HAJIBABAEI *et al.* 2007). Coupled with ecological, physiological and geographical information, diatom species boundaries could be re-evaluated, as was done recently for some green microalgae (DARIENKO *et al.* 2015). Diatom species descriptions have until recently been based only on morphological features. Only a few studies

(e.g. ROVIRA 2013; TROBAJO *et al.* 2013; KELLY *et al.* 2015, for the *Nitzschia inconspicua* GRUNOW species complex) follow the precepts of integrative taxonomy which aims to delimit species on sets of different criteria such as morphology, DNA, physiology, ecology and biogeography (DAYRAT 2005). This should prompt reconsideration of the boundaries of many poorly delimited diatom species and, potentially, enhance their value for ecological assessment.

In conclusion, the approach described here offers a pragmatic and universally applicable way to enrich existing diatom reference barcode libraries with HTS generated barcodes, especially when the sequences are as well documented as classical voucher specimens, no matter which region/gene is used. Nonetheless, we believe that unialgal diatom cultures should still be the backbone of reference libraries, because this is still the method with lowest amount of error.

Acknowledgements

Sampling, microscopy and sequencing were funded by the French ONEMA (Office National de l'Eau et des Milieu Aquatiques). Sequencing was carried out in the Genome Transcriptome Facility of Bordeaux (INRA Pierroton) and we thank Alain Franc, Philippe Chaumeil, Jean-Marc Frigerio and Franck Salin for helpful discussions. We also thank DNAqua-Net (European Cost Action CA15219) which helped discussion between some of the authors.

References

- ABARCA, N.; JAHN, R.; ZIMMERMANN, J. & ENKE, N. (2014): Does the cosmopolitan diatom *Gomphonema parvulum* (Kützing) Kützing have a biogeography? – PLoS ONE 9: 1–18.
- BARBOUR, M.T.; GERRITSEN, J.; SNYDER, B.D. & STRIBLING, J.B. (1999): Rapid bioassessment protocols for use in streams and wadeable rivers: periphyton, benthic macroinvertebrates and fish. Second edition. – EPA 841-B-99-002. US Environmental Protection Agency, Office of Water, Washington, DC.
- BESSE-LOTOSKAYA, A.; VERDONSCHOT, P. & SINKELDAM, J. (2006): Uncertainty in diatom assessment: sampling, identification and counting variation. – Hydrobiologia 566: 247–260.
- BRAGG, L.M.; STONE, G.; BUTLER, M.K.; HUGENHOLTZ, P. & TYSON, G.W. (2013): Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. – PLoS Comput. Biol. 9: 1–18.
- BRUDER, K. & MEDLIN, L.K. (2007): Molecular assessment of phylogenetic relationships in selected species/genera in the naviculoid diatoms (Bacillariophyta). I. The genus *Placoneis*. – Nova Hedwigia 85: 331–352.
- BUTCHER, R.W. (1947): Studies in the ecology of rivers. IV. The algae of organically enriched water. – J. Ecol. 35: 186–191.
- CASTELEYN, G.; LELIAERT, F.; BACKELJAU, T.; DEBEER, A.E.; KOTAKI, Y.; RHODES, L.; LUNDHOLM, N.; SABBE, K. & VYVERMAN, W. (2010): Limits to gene flow in a cosmopolitan marine planktonic diatom. – PNAS 107: 12952–12957.
- CHONOVA, T.; KECK, F.; LABANOWSKI, J.; MONTUELLE, B.; RIMET, F. & BOUCHEZ, A. (2016): Separate treatment of hospital and urban wastewaters: a real scale comparison of effluents and their effect on microbial communities. – Sci. Total Environ. 542: 965–975.
- COCQUYT, C. (1998): Diatoms from Northern Basin of Lake Tanganyika. – Bibliotheca Diatomologica, vol. 39, J. Cramer, Berlin and Stuttgart.
- COHN, F. (1853): Über lebendige Organismen im Trinkwasser. – Z. klin. Med. 4: 229–237.
- DARIENKO, T.; GUSTAVS, L.; EGGERT, A.; WOLF, W. & PRÖSCHOLD, T. (2015): Evaluating the species boundaries of green microalgae (*Coccomyxa*, Trebouxiophyceae, Chlorophyta) using integrative taxonomy and DNA Barcoding with further implications for the species identification in environmental samples. – PLoS ONE 10: 1–31.
- DAYRAT, B. (2005): Towards integrative taxonomy. – Biol. J. Linn. Soc. 85: 407–415.
- EDGAR, R.C.; HAAS, B.J.; CLEMENTE, J.C.; QUINCE, C. & KNIGHT, R. (2011): UCHIME improves sensitivity and speed of chimera detection. – Bioinformatics 27: 2194–2200.
- EDGAR, R.S. (2004): MUSCLE: multiple sequence alignment with high accuracy and high throughput. – Nucleic Acids Res. 32: 1792–1797.
- EUROPEAN COMMISSION (2000): Directive 2000/60/EC of the European Parliament and of the Council of 23rd October 2000 establishing a framework for Community action in the field of water policy. – Official Journal of the European Communities 327: 1-72.

EUROPEAN COMMITTEE FOR STANDARDISATION (2014a): EN 13946 – Water quality - Guidance for the routine sampling and preparation of benthic diatoms from rivers and lakes. – 18 pp., Afnor, La Plaine St Denis, France.

EUROPEAN COMMITTEE FOR STANDARDISATION (2014b): EN 14407 – Water quality - Guidance for the identification and enumeration of benthic diatom samples from rivers and lakes. – 13 pp., Afnor, La Plaine St Denis, France.

EVANS, K.M.; WORTLEY, A.H. & MANN, D. G. (2007): An assessment of potential diatom "barcode" genes (*cox1*, *rbcl*, 18S and ITS rDNA) and their effectiveness in determining relationships in *Sellaphora* (Bacillariophyta). – Protist 158: 349–364.

GEITLER, L. (1972): Sippen von *Gomphonema parvulum*, Paarungsverhalten und Variabilität pennater Diatomeen. – Österr. Bot. Z. 120: 257–268.

GOMEZ, F.; LOPEZ-GARCIA, P.; DOLAN, J.R. & MOREIRA, D. (2012): Molecular phylogeny of the marine dinoflagellate genus *Heterodinium* (Dinophyceae). – Eur. J. Phycol. 47: 95–104.

GOUY, M.; GUINDON, S. & GASCUEL, O. (2010): SeaView Version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. – Mol. Biol. Evol. 27: 221–224.

HAJIBABAEI, M.; SINGER, G.A.C.; HEBERT, P. & HICKEY, D.A. (2007): DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. – Trends Genet. 23: 167–172.

HAJIBABAEI, M.; BAIRD, D.J.; FAHNER, N.A.; BEIKO, R. & GOLDING, G.B. (2016): A new way to contemplate Darwin's tangled bank: how DNA barcodes are reconnecting biodiversity science and biomonitoring. – Phil. Trans. R. Soc., B 371: 20150330.

HAMILTON, P B.; LEFEBVRE, K. & BULL, R. (2015): Single cell PCR amplification of diatoms using fresh and preserved samples. – Front. Microbiol. 6: 1084.

HAUSSMANN, S.; CHARLES, D.F.; GERRITSEN, J. & BELTON, T.J. (2016): A diatom-based biological condition gradient (BCG) approach for assessing impairment and developing nutrient criteria for streams. – Sci. Total Environ. 562: 914–927.

HEBERT, P.; CYWINSKA, A.; BALL, S.L. & DEWAARD, J.R. (2003): Biological identifications through DNA barcodes. – Proc. R. Soc. Lond., B 270: 313–321.

HOFFMAN, G.; WERUM, M. & LANGE-BERTALOT, H. (2011): Diatomeen im Süßwasser-Benthos von Mitteleuropa. – A.R.G. Gantner, Ruggell, Liechtenstein.

HOLMGREN, P.K.; HOLMGREN, N.H. & BARNETT, L.C. (eds) (1990): Index herbariorum, ed. 8. Part 1. The herbaria of the world. – 704 pp., New York Botanical Garden.

HUSTEDT, F. (1957): Die Diatomeenflora des Flusssystemes der Weser im Gebiet der Hansestadt Bremen. – Abh. naturwiss. Ver. Bremen 34: 181-440.

INSEE (2016): Estimation de la population au 1er janvier par région, département, sexe et âge de 1975 à 2015. – Internet Communication, consulted at www.insee.fr, 9 September 2016.

JAHN, R.; ZETSCHE, H.; REINHARDT, R. & GEMEINHOLZER, B. (2007): Diatoms and DNA barcoding: A pilot study on an environmental sample. In: KUSBER, W.H. & JAHN, R. (eds): Proceedings of the 1st Central

European Diatom Meeting. – pp. 63–68. Botanic Garden and Botanical Museum Berlin-Dahlem, Freie Universität Berlin.

JARAMILLO, A.; OSMAN, D.; CAPUTO, L. & CARDENAS, L. (2015): Molecular evidence of a *Didymosphenia geminata* (Bacillariophyceae) invasion in Chilean freshwater systems. – *Harmful Algae* 49: 117–123.

KAHLERT, M.; ALBERT, R.L.; ANTTILA, E.L.; BENGTSSON, R.; BIGLER, C.; ESKOLA, T.; GALMAN, V.; GOTTSCHALK, S.; HERLITZ, E.; JARLMAN, A.; KASPEROVICIENE, J.; KOKOCINSKI, M.; LUUP, H.; MIETTINEN, J.; PAUNKSNYTE, I.; PIIRSOO, K.; QUINTANA, I.; RAUNIO, J.; SANDELL, B.; SIMOLA, H.; SUNDBERG, I.; VILBASTE, S. & WECKSTROM, J. (2009): Harmonization is more important than experience – results of the first Nordic–Baltic diatom intercalibration exercise 2007 (stream monitoring). – *J. Appl. Phycol.* 21: 471–482.

KELLY, M.; URBANIC, G.; ACS, E.; BENNION, H.; BERTRIN, V.; BURGESS, A.; DENYS, L.; GOTTSCHALK, S.; KAHLERT, M.; KARJALAINEN, S.M.; KENNEDY, B.; KOSI, G.; MARCHETTO, A.; MORIN, S.; PICINSKA-FALTYNOWICZ, J.; POIKANE, S.; ROSEBERY, J.; SCHOENFELDER, I.; SCHOENFELDER, J. & VARBIRO, G. (2014): Comparing aspirations: intercalibration of ecological status concepts across European lakes for littoral diatoms. – *Hydrobiologia* 734: 125–141.

KELLY, M.G.; TROBAJO, R.; ROVIRA, L. & MANN, D.G. (2015): Characterizing the niches of two very similar *Nitzschia* species and implications for ecological assessment. – *Diatom Res.* 30: 27–33.

KERMARREC, L.; BOUCHEZ, A.; RIMET, F. & HUMBERT, J.F. (2013a): First evidence of the existence of semi-cryptic species and of a phylogeographic structure in the *Gomphonema parvulum* (Kützing) Kützing complex (Bacillariophyta). – *Protist* 164: 686–705.

KERMARREC, L.; FRANC, A.; RIMET, F.; CHAUMEIL, P.; HUMBERT, J.F. & BOUCHEZ, A. (2013b): Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: a test for freshwater diatoms. – *Mol. Ecol. Res.* 13: 607–619.

KERMARREC, L.; FRANC, A.; RIMET, F.; CHAUMEIL, P.; FRIGERIO, J.M.; HUMBERT, J.F. & BOUCHEZ, A. (2014): A next-generation sequencing approach to river biomonitoring using benthic diatoms. – *Freshw. Sci.* 33: 349–363.

KHALIFA, M.E.; VARSANI, A.; GANLEY, A.R.D. & PEARSON, M.N. (2016): Comparison of Illumina *de novo* assembled and Sanger sequenced viral genomes: A case study for RNA viruses recovered from the plant pathogenic fungus *Sclerotinia sclerotiorum*. – *Virus Res.* 219: 51–57.

KHAN-BUREAU, D.A.; MORALES, E.A.; ECTOR, L.; BEAUCHENE, M.S. & LEWIS, L.A. (2016): Characterization of a new species in the genus *Didymosphenia* and of *Cymbella janischii* (Bacillariophyta) from Connecticut, USA. – *Eur. J. Phycol.* 51: 203–216.

Ki, J.S.; Cho, S.Y.; Katano, T.; Jung, S.W.; Lee, J.; Park, B.S.; Kang, S.H. & Han, M.S. (2009) Comprehensive comparisons of three pennate diatoms, *Diatoma tenuae*, *Fragilaria vaucheriae*, and *Navicula pelliculosa*, isolated from summer Arctic reservoirs (Svalbard 79°N), by Wne-scale morphology and nuclear 18S ribosomal DNA. – *Polar Biol.* 32: 147–159.

KOLKWITZ, R. & MARSSON, M. (1908): Ökologie der pflanzliche Saprobien. – *Ber. Deutsch. Bot. Ges.* 26: 505–519.

KRAMMER, K. & LANGE-BERTALOT, H. (1986): Bacillariophyceae 1. Teil: Naviculaceae. – In: Ettl, H.; Gerloff, J.; Heynig, H. & Mollenhauer, D. (eds): Süßwasserflora von Mitteleuropa – Vol. 2/1, 876 pp., G. Fischer, Stuttgart & New York.

- KUMAR, S.; STECHER, G., & TAMURA, K. (2016): MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. – *Mol. Biol. Evol.* 33: 1870–1874.
- KUSBER, W.H.; ABARCA, N.; SKIBBE, O.; ZIMMERMANN, J. & JAHN, R. (2012): Reference library of DNA-barcoded diatoms - A use case for publishing data via the GBIF database AlgaTerra. – In: SABBE, K.; VAN DE VIJVER, B. & VYVERMAN, W. (eds): Abstracts. 22nd International Diatom Symposium, Aula Academica, Ghent – p. 65, VLIZ Special Publication 58 (available at <http://www.vliz.be/events/ids2012/ABSTRACTBOOK%20IDS%202012.pdf>).
- KÜTZING, F.T. (1844): Die kieselschaligen Bacillarien oder Diatomeen. – 152 pp., W. Köhne, Nordhausen.
- LEVKOV, Z. (2009): *Amphora* sensu lato. – In: LANGE-BERTALOT, H. (ed.) Diatoms of the European Inland Waters and Comparable Habitats. – Vol. 5, 916 pp., A.R.G. Gantner, Ruggell, Liechtenstein.
- LOMAN, N.J.; MISRA, R.V.; DALLMAN, T.J.; CONSTANTINIDOU, C.; GHARBIA, S.E.; WAIN, J. & PALLAN, M.J. (2012): Performance comparison of benchtop high-throughput sequencing platforms. – *Nat. Biotechnol.* 30: 434–439.
- MANN, D.G. & VANORMELINGEN, P. (2013): An inordinate fondness? The number, distributions and origins of diatom species. – *J. Euk. Microbiol.* 60: 414–420.
- MARKERT, B.A.; BREURE, A.M. & ZECHMEISTER, H.G. (2003): Definitions, strategies and principles for bioindication/biomonitoring of the environment. – In: MARKERT, B.A. BREURE, A.M. & ZECHMEISTER, H.G. (eds): *Bioindicators & Biomonitoring Principles, Concepts and Applications*. – pp. 3–39, Elsevier, Amsterdam.
- MONTOYA, V.; OLMSTEAD, A.; TANG, P.; COOK, D.; JANJUA, N.; GREBELY, J.; JACKA, B.; POON, A.F.Y. & KRAJEN, M. (2016): Deep sequencing increases hepatitis C virus phylogenetic cluster detection compared to Sanger sequencing. – *Infect. Gen. Evol.* 43: 329–337.
- MURAKAMI, T. & KASUYA, M. (1993): Teratological variations of *Gomphonema parvulum* Kützing in heavily polluted drainage channel. – *Diatom* 8: 7–10.
- NAKOV, T.; RUCK, E.; GALACHYANTS, Y.; SPAULDING, S.A. & THERIOT, E.C. (2014): Molecular phylogeny of the Cymbellales (Bacillariophyceae, Heterokontophyta) with a comparison of models for accommodating rate variation across sites. – *Phycologia* 53: 359–373.
- NDIRITU, G.G.; GICHUKI, N.N. & TRIEST, L. (2006): Distribution of epilithic diatoms in response to environmental conditions in an urban tropical stream, Central Kenya. – *Biodivers. Conserv.* 15: 3267–3293.
- PAPARINI, A.; GOFTON, A.; YANG, R.; WHITE, N.; BUNCE, M. & RYAN, U.M. (2015): Comparison of Sanger and next generation sequencing performance for genotyping *Cryptosporidium* isolates at the 18S rRNA and actin loci. – *Exp. Parasitol.* 151: 21–27.
- POMPANON, F.; COISSAC, E. & TABERLET, P. (2011): Metabarcoding, une nouvelle façon d’analyser la biodiversité. – *Biofutur* 319: 30–32.
- POTAPOVA, M. & CHARLES, D.F. (2007): Diatom metrics for monitoring eutrophication in rivers of the United States. – *Ecol. Indic.* 7: 48–70.

REICHARDT, E. (1999): Zur revision der Gattung *Gomphonema*. Die Arten um *G. affine/insigne*, *G. angustatum/micropus*, *G. acuminatum* sowie gomphonemoide Diatomeen aus dem Oberoligozän in Böhmen. – In: Lange-Bertalot, H. (ed.) Iconographia Diatomologica. – Vol. 8, 203 pp., A.R.G. Gantner, Ruggell, Liechtenstein.

RIMET, F. (2009): Benthic diatom assemblages and their correspondence with ecoregional classifications: case study of rivers in north-eastern France. – *Hydrobiologia* 636: 137–151.

RIMET, F. (2012): Recent views on river pollution and diatoms. – *Hydrobiologia* 683: 1–24.

RIMET, F.; TROBAJO, R.; MANN, D.G.; KERMARREC, L.; FRANC, A.; DOMAIZON, I. & BOUCHEZ, A. (2014): When is sampling complete? The effects of geographical range and marker choice on perceived diversity in *Nitzschia palea* (Bacillariophyta). – *Protist* 165: 245–259.

RIMET, F.; CHAUMEIL, P.; KECK, F.; KERMARREC, L.; VASSELON, V.; KAHLERT, M.; FRANC, A. & BOUCHEZ, A. (2016): R-Syst::diatom: An open-access and curated barcode database for diatoms and freshwater monitoring. – *Database (Oxford)* 2016: baw016: 1–21.

ROSE, D. & COX, E.J. (2014): What constitutes *Gomphonema parvulum*? Long-term culture studies show that some varieties of *G. parvulum* belong with other *Gomphonema* species. – *Plant Ecol. Evol.* 147: 366–373.

ROVIRA, L. (2013): The ecology and taxonomy of estuarine benthic diatoms and their use as bioindicators in a highly stratified estuary (Ebra Estuary, NE Iberian Peninsula): a multidisciplinary approach. – 295 pp., PhD dissertation, University of Barcelona.

RUCK, E.; NAKOV, T.; ALVERSON, A.J. & THERIOT, E.C. (2016): Phylogeny, ecology, morphological evolution, and reclassification of the diatom orders Surirellales and Rhopalodiales. – *Mol. Phylogenet. Evol.* 103: 155–171.

SCHLOSS, P.D.; WESTCOTT, S.L.; RYABIN, T.; HALL, J.R.; HARTMANN, M.; HOLLISTER, E.B.; LESNIEWSKI, R.A.; OAKLEY, B.B.; PARKS, D.H.; ROBINSON, C.J.; SAHL, J.W.; STRES, B.; THALLINGER, G.G.; VAN HORN, D.J., & WEBER, C.F. (2009): Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. – *Appl. Environ. Microbiol.* 75: 7537–7541.

SILVA-BENAVIDES, A.M. (1996): The epilithic diatom flora of a pristine and a polluted river in Costa Rica, Central America. – *Diatom Res.* 11: 105–142.

SILVESTRO, D. & MICHALAK, I. (2012): raxmlGUI: a graphical front-end for RAxML. – *Org. Divers. Evol.* 12: 335–337.

STEPANEK, J.G. & KOCIOLEK, J.P. (2014): Molecular phylogeny of *Amphora* sensu lato (Bacillariophyta): an investigation into the monophyly and classification of the amphoroid diatoms. – *Protist* 165: 177–195.

STEVENSON, R.J. (2014): Ecological assessments with algae: a review and synthesis. – *J. Phycol.* 50: 437–461.

STOOF-LEICHSENDRING, K.R.; EPP, L.S.; TRAUTH, M.H. & TIEDEMANN, R. (2012): Hidden diversity in diatoms of Kenyan Lake Naivasha: a genetic approach detects temporal variation. – *Mol. Ecol.* 21: 1918–1930.

- SWEENEY, B.W.; BATTLE, J.M.; JACKSON, J.K. & DAPKEY, T. (2011): Can DNA barcodes of stream macroinvertebrates improve descriptions of community structure and water quality? – J. North Am. Benthol. Soc. 30: 195–216.
- TAKANO, Y. & HORIGUCHI, T. (2006): Acquiring scanning electron microscopical, light microscopical and multiple gene sequence data from a single dinoflagellate cell. – J. Phycol. 42: 251–256.
- TAPOLCZAI, K.; BOUCHEZ, A.; STENGER-KOVÁCS, C.; PADISÁK, J. & RIMET, F. (2016): Species- or trait-based ecological assessment for tropical rivers? Case study of benthic diatoms in Mayotte island (France, northern Mozambique Channel). *Submitted*.
- THERIOT, E.C.; ASHWORTH, M.P.; NAKOV, T.; RUCK, E. & JANSEN, R.K. (2015): Dissecting signal and noise in diatom chloroplast protein encoding genes with phylogenetic information profiling. – Mol. Phylogenet. Evol. 89: 28–36.
- TROBAJO, R.; CLAVERO, E.; CHEPURNOV, V.; SABBE, K.; MANN, D.G.; ISHIHARA, S., & COX, E.J. (2009): Morphological, genetic and mating diversity within the widespread bioindicator *Nitzschia palea* (Bacillariophyceae). – Phycologia 48: 443–459.
- TROBAJO, R.; ROVIRA, L.; ECTOR, L.; WETZEL, C.E.; KELLY, M., & MANN, D.G. (2013): Morphology and identity of some ecologically important small *Nitzschia* species. – Diatom Res. 28: 37–59.
- VASSELON V.; DOMAIZON I.; RIMET F.; KAHLERT M. & BOUCHEZ A. (2017, online): Application of high-throughput sequencing (HTS) metabarcoding to diatom biomonitoring: Do DNA extraction methods matter? – Freshw. Sci. 36:162–177.
- VISCO, J.; APOTHE-LOZ-PERRET-GENTIL, L.; CORDONIER, A.; ESLING, P.; PILLET, L. & PAWLOWSKI J. (2015): Environmental monitoring: inferring the diatom index from next-generation sequencing data. – Environ. Sci.Technol. 49: 7597–7605.
- WACHNICKA, A.H. & GAISER, E.E. (2016): Characterization of *Amphora* and *Seminavis* from South Florida, U.S.A. – Diatom Res. 22: 387–455.
- WANG, Q. G.; GARRITZ, G.M.; TEIDJE, J.M. & COLE, J.R. (2007): Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. – Appl. Environ. Microbiol. 73: 5261–5267.
- ZELINKA, M. & MARVAN, P. (1961): Zur Präzisierung der biologischen Klassifikation der Reinheit fließender Gewässer. – Arch. Hydrobiol. 57: 389–407.
- ZIMMERMANN, J.; ABARCA, N.; ENKE, N.; SKIBBE, O.; KUSBER, W.H. & JAHN, R. (2014): Taxonomic reference libraries for environmental barcoding: a best practice example from diatom research. – PLoS ONE 9: 1–24.
- ZIMMERMANN, J.; GLÖCKNER, G.; JAHN, R.; ENKE, N. & GEMEINHOLZER, B. (2015): Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. – Mol. Ecol. Res. 15: 526–542.
- ZIMMERMANN, J.; KUSBER, W.H.; DROEGE, G. & JAHN R. (2016): GBOL2 – Increasing the accessibility of eDNA barcoding data. – GGBN Newsletter 5: 7–8.

Tables

Table 1. Sampling sites location, habitat description, and collectors names. Sites acronyms used by the local authorities are given in brackets.

Sampling site	Sampling date	Habitat	Coordinates	Collectors	Dominant target species
Soulou river waterfall	25/07/2015	Unpolluted waterfall	- 12°46'48.3"S 45°6'6.5"E	Rimet F. and Tapolczai K.	<i>Epithemia hirundiformis</i> (O. Müller) comb. nov.
Gouloué river, near Passamainty city (Poll7)	20/07/2015	Polluted river, marine influence	- 12°47'57.9"S 45°12'37.9"E	Rimet F. and Tapolczai K.	<i>Halamphora ghanensis</i> Levkov
Songaro Mbili river near Dembeni city (Poll29)	24/07/2015	Polluted river	- 12°50'23.3"S 45°10'28.3"E	Rimet F. and Tapolczai K.	<i>Gomphonema clavatuloides</i> sp. nov.
Mouala river near Mirereni city (Poll5)	23/07/2015	Polluted river	- 12°47'25.9"S 45°08'21.6"E	Tapolczai K. and Vasselon V.	<i>G. parvulum</i> (Kützing) Kützing sensu lato

Table 2. Proposed terminology for the naming of sequences from different origins.

Cases	Newly described	Original material	Morphological data	Molecular data	Terminology examples
1	Yes	Authentic ¹ strain of the type (epitype, holotype)	Unialgal culture	Unialgal culture (Sanger sequencing)	<i>Planothidium caputium</i> authentic ¹ strain D06_014
2	No	No	Unialgal culture	Unialgal culture (Sanger sequencing)	<i>Planothidium frequentissimum</i> strain D06_138
3	Yes	Yes	Environmental sample	Environmental sample (HTS)	<i>Gomphonema clavatuloides</i> authentic ¹ uncultured isolate TCC955 inferred via eDNA
4	No	No	Environmental sample	Environmental sample (HTS)	<i>Epithemia hirundiformis</i> uncultured isolate TCC955 inferred via eDNA <i>Halamphora ghanensis</i> uncultured isolate TCC956 inferred via eDNA
5	No	No	No	Environmental sample (via cloning/Sanger)	<i>Planothidium</i> sp. uncultured isolate TF-2014 05DB5_12 inferred via cloning ²
6	No	No	No	Environmental sample (HTS)	<i>Genus species</i> uncultured inferred via eDNA ³

¹term for a type-bearing strain (unialgal culture) or isolate (object isolated from environmental sample) because a nomenclatural type is usually a preparation, a slide etc.

²sometimes named as clone

³only applicable for 100% match with DNA reference library

Figure caption

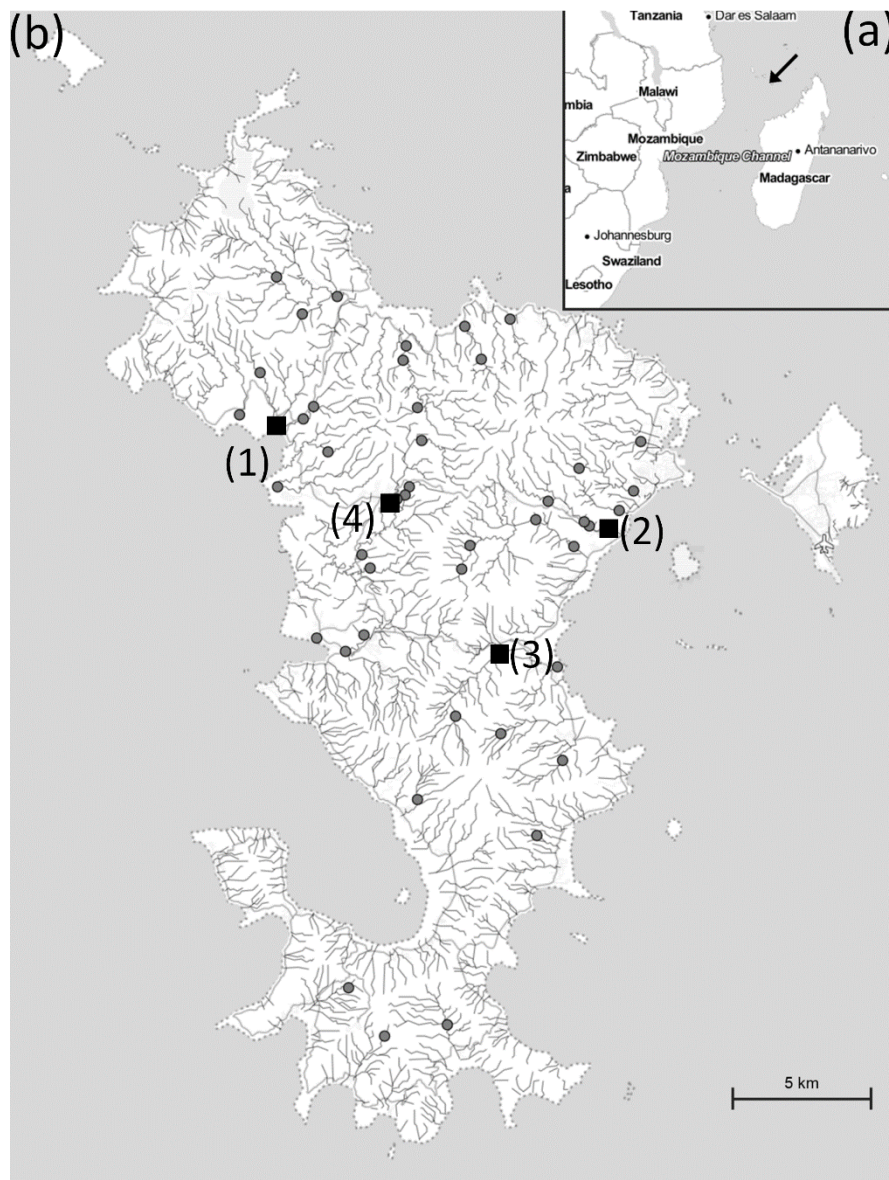


Figure 1: Study site location. (a) General location of Mayotte in Mozambique channel (arrow). (b) Location of rivers (grey lines), sampling sites in Mayotte (grey dots) and the selected sampling sites discussed in this paper (black squares): (1) Soulou river waterfall (2) downstream the Gouloué river (Poll 7) (3) downstream the Songaro Mbili river (Poll 29) (4) Mouala river near Mirereni city (Poll5).

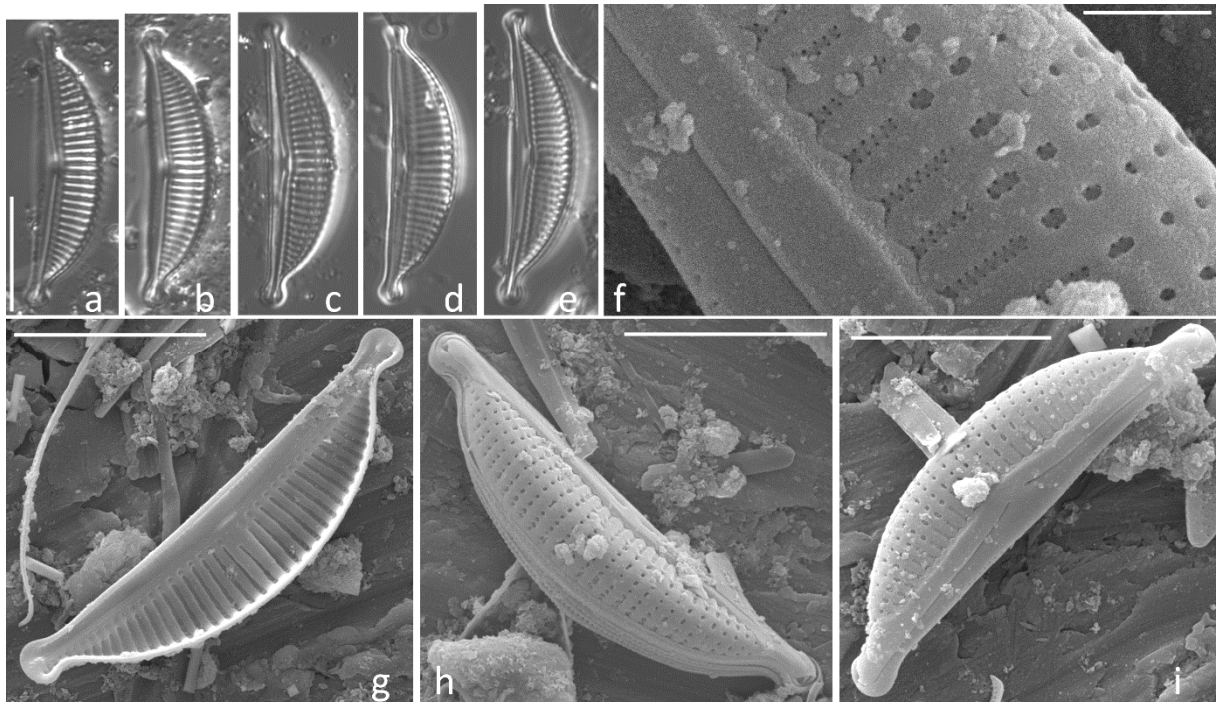


Figure 2: *Halamphora ghanensis* in downstream Gouloué river. (a-e) Light microscopy, valve views. (f-i) Scanning electron microscopy. (g) internal view. (f, h, i) external views. (f) detail of areola structure. Scale bar 10 μm .

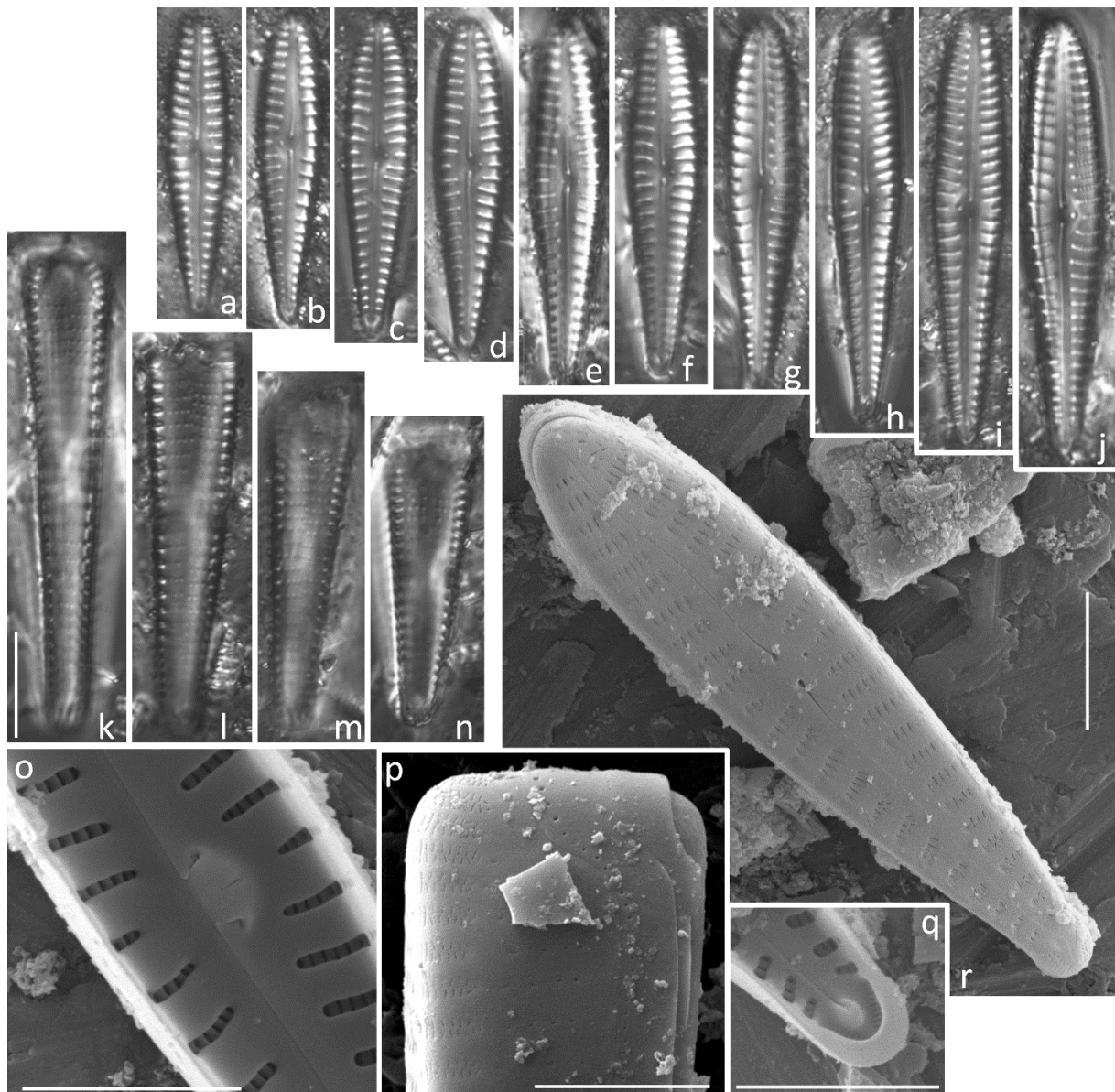


Figure 3: *Gomphonema clavatuloides*, downstream the Songaro Mbili river. (a-j) light microscopy, valve view. (k-n) light microscopy, connective view. (o-r) Scanning electron microscopy. (o) internal view, detail of proximal raphe ending and areola structure. (p) external view of connective view (head pole). (q) internal raphe ending at the foot pole, (r) general external view. Scale bar 10 μm light microscopy photos, 5 μm for scanning electron microscopy.

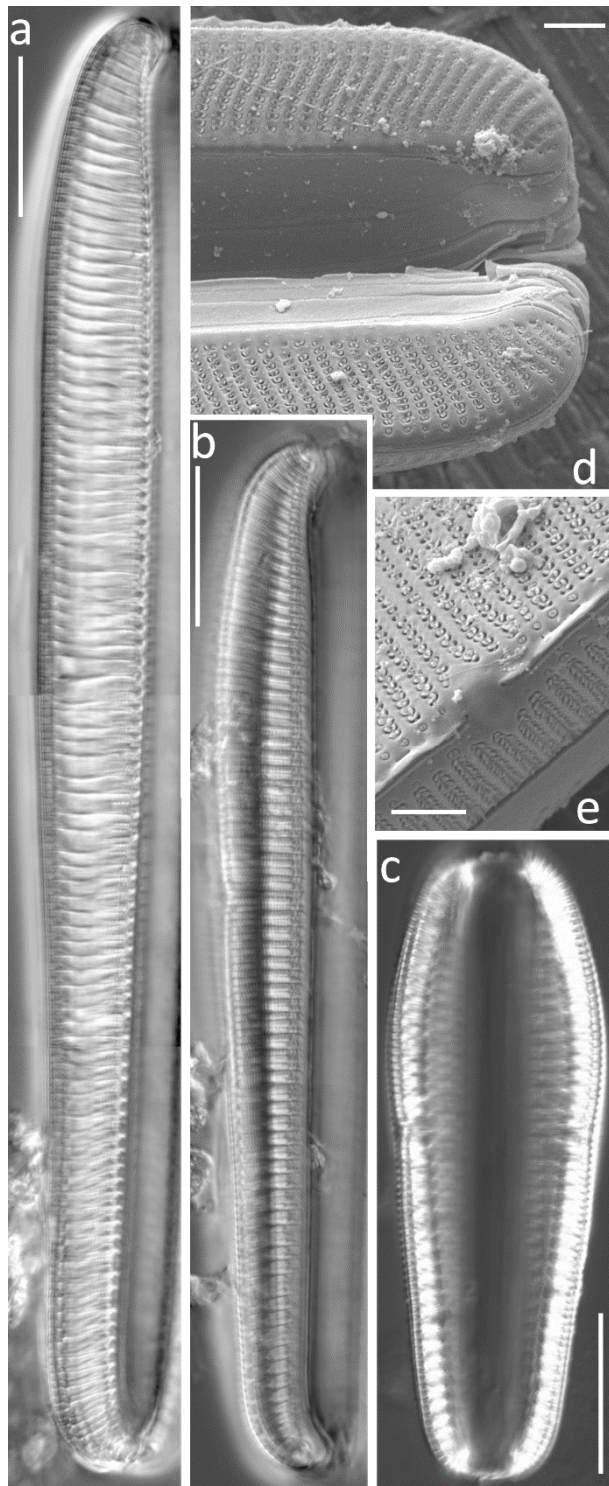


Figure 4: *Epithemia hirudiniformis*, Waterfall of Soulou river. (a-c) light microscopy. (d, e) Scanning electron microscopy. (d) detail of the foot pole. (e) detail of the proximal ending of the raphe and of the areola structure. Scale bar 20 μm (4.1-4.3), 3 μm (1.4-4.5).

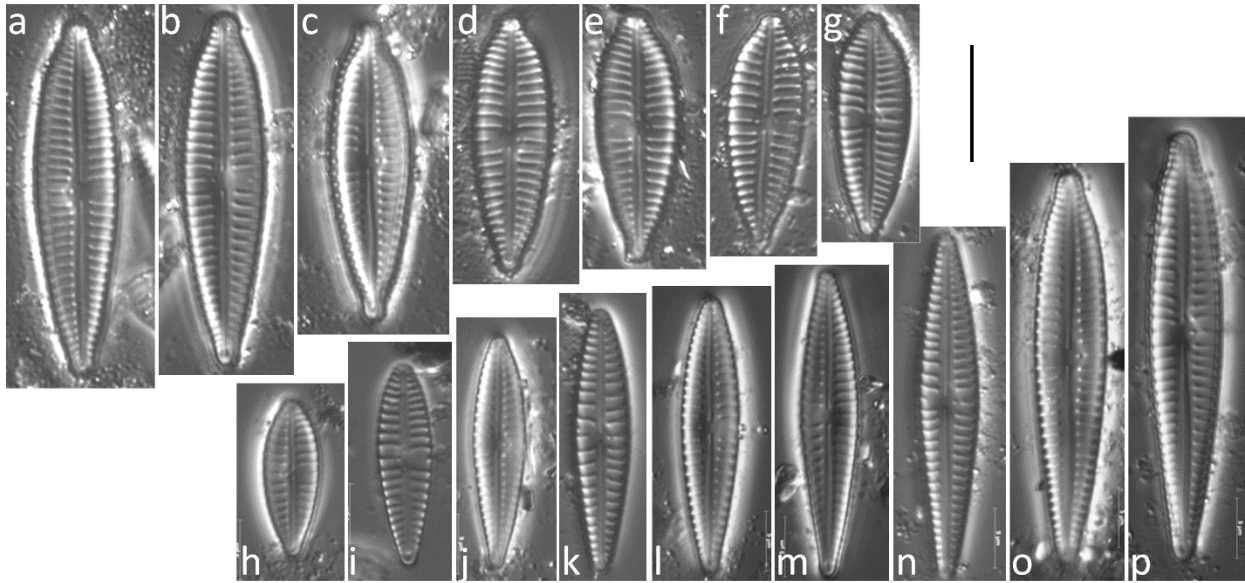


Figure 5: *Gomphonema parvulum* sensu lato downstream the Mouala river. Light microscopy photos showing the morphological heterogeneity of this taxon in this sample. (a-g) morphology corresponding to *G. lagenula*. (h-p) morphology corresponding to *G. parvulum* sensu lato. Scale bar 10 μ m.

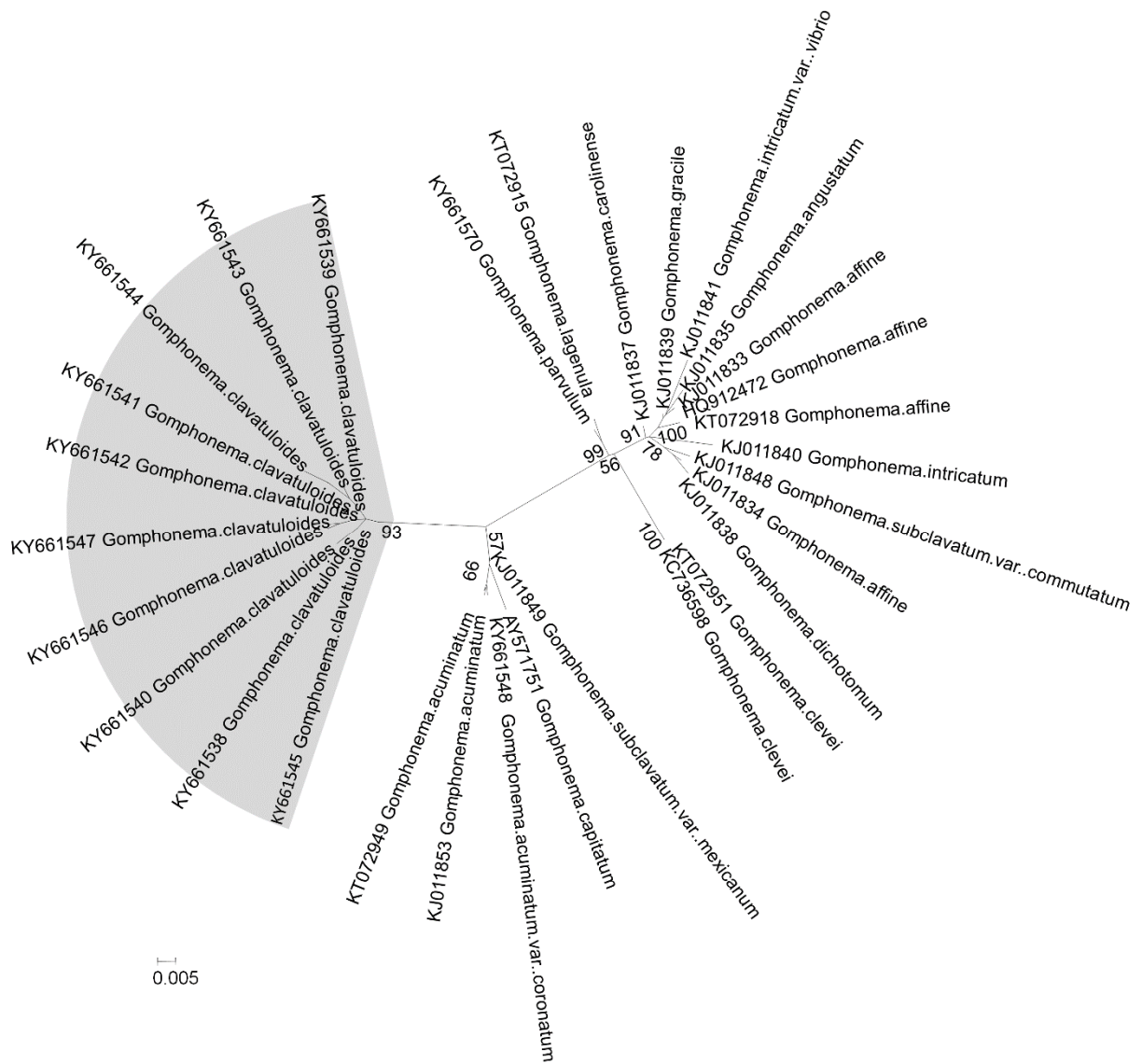


Figure 7: Constrained unrooted phylogeny of *Gomphonema clavatuloides*. The ten *G. clavatuloides* sequences (312 bp) were constrained by the phylogeny of the 87 other sequences (1093 bp). Maximum likelihood tree with rapid bootstrap and GT Gamma, RaxML. 1000bootstraps. Bootstrap values above 50% are given for each node. Scale bar: number of substitution per site.

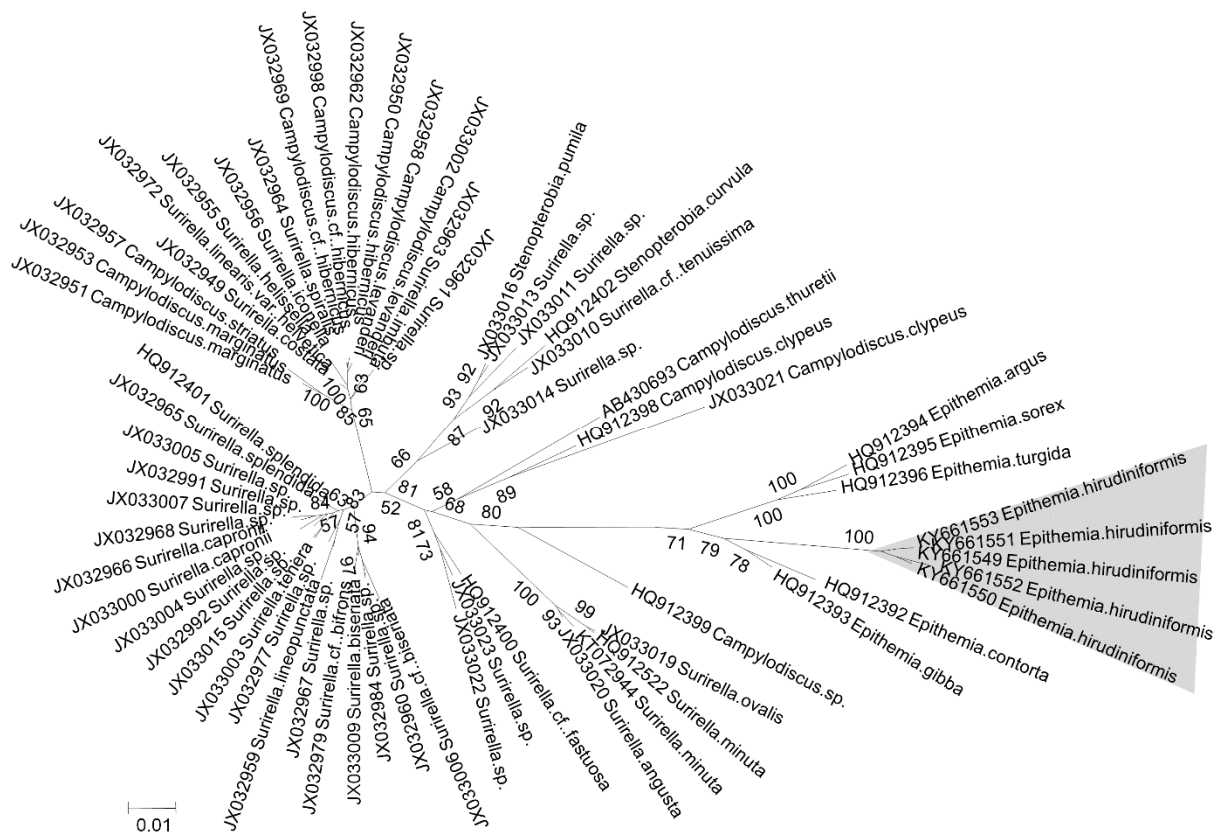


Figure 8: Constrained unrooted phylogeny of *Epithemia hirudiniformis*. The five *E. hirudiniformis* sequences (290 bp) were constrained by the phylogeny of the 87 other sequences (1362 bp). Maximum likelihood tree with rapid bootstrap and GT Gamma, RaxML. 1000bootstraps. Bootstrap values above 50% are given for each node. Scale bar: number of substitution per site.

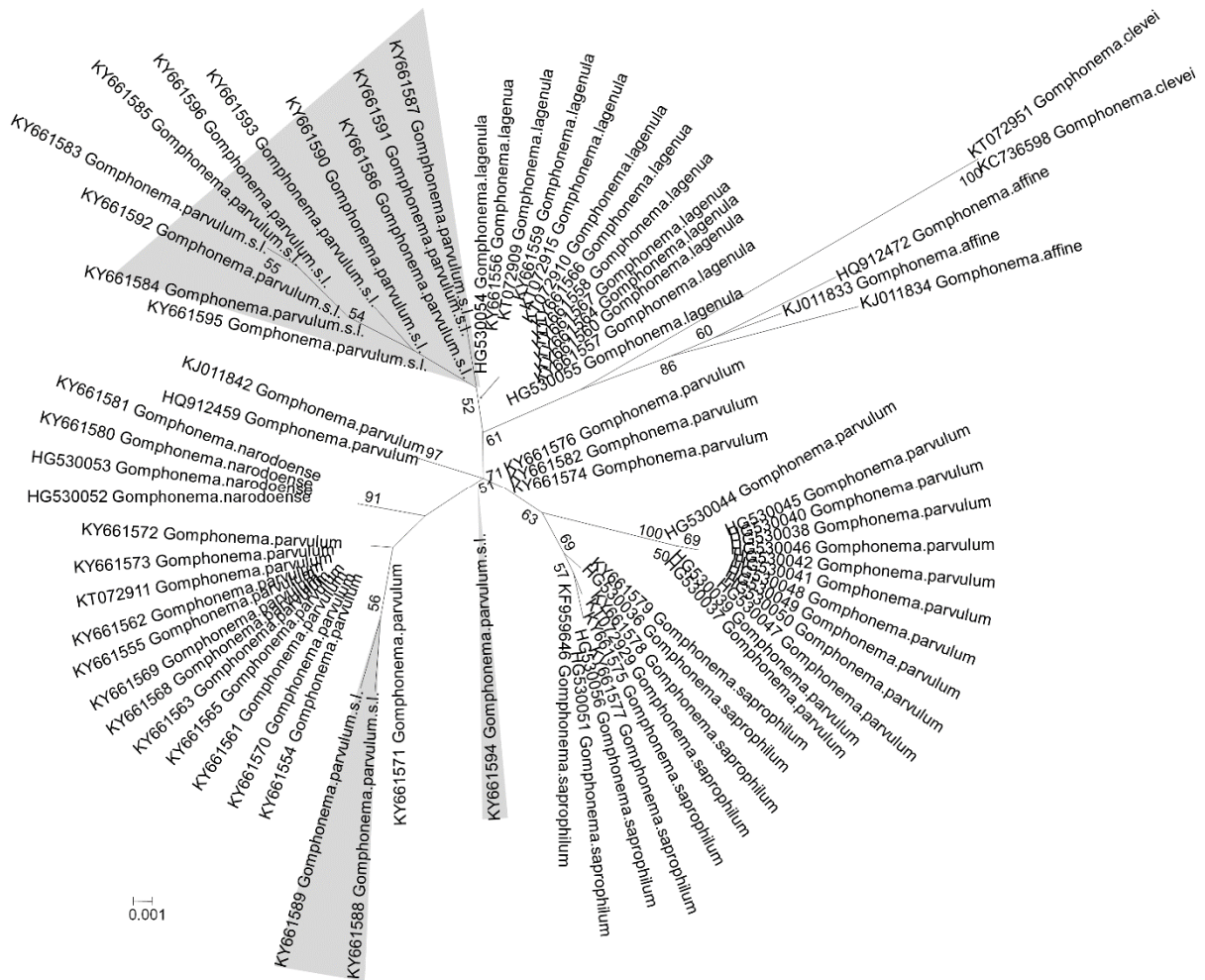


Figure 9: Constrained unrooted phylogeny of *Gomphonema parvulum* sequences of the downstream sampling site of the Mouala river. Fourteen *G. parvulum* sequences (275 bp) were constrained by the phylogeny of the 62 other sequences (916 bp). Maximum likelihood tree with rapid bootstrap and GT Gamma, RaxML. 1000 bootstraps. Bootstrap values above 50% are given for each node. Scale bar: number of substitution per site.

Article annexe III

“Enhancing DNA metabarcoding for biomonitoring with phylogenetic estimation of ecological profiles for unclassified OTUs”

(soumis dans le journal *Methods in Ecology and Environment*, 2017)

Keck F.^{1,2}, Vasselon V.², Rimet F.², Bouchez A.², Kahlert M.¹

¹ Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, P. O. Box 7050, 750 07 Uppsala, Sweden

² UMR CARRETEL, INRA, Université Savoie Mont Blanc, F-74200 Thonon, France

Abstract

1. DNA metabarcoding has been introduced as a revolutionizing way to identify organisms and monitor ecosystems. However, the potential of this approach for biomonitoring remains partially unfulfilled because a significant part of the sampled DNA cannot be affiliated to species due to incomplete reference libraries. Thus, biotic indices which are based on the estimated abundances of the species in the community and their ecological profiles can be inaccurate.
2. We propose to compute biotic indices using phylogenetic imputation of OTUs' ecological profiles (OTU-PITI approach). First, OTUs sequences are inserted within a reference phylogeny. Second, OTUs' ecological profiles are estimated on the basis of their phylogenetic relationships with reference species whose ecology is known. Based on these ecological profiles, biotic indices can be computed, using all available OTUs.
3. Using freshwater diatoms as a case study, we show that short DNA barcode can be placed accurately within a phylogeny and their ecological preferences can be estimated with a satisfying degree of precision. In light of these results, we tested the approach with a dataset of 139 environmental samples of river diatoms for which the same biotic index (IPS) was calculated using (i) traditional microscopy, (ii) OTUs with taxonomic assignment approach, (iii) OTUs with phylogenetic estimation of ecological profiles (OTU-PITI), and (iv) OTU with taxonomic assignment completed by the phylogenetic approach (OTU-PITI) for unclassified OTUs. Using traditional microscopy as reference, we found that the combination of the OTUs' taxonomic assignment completed by the phylogenetic method performed well and substantially better than the other methods.
4. Phylogenetic estimation of ecological profiles for unclassified OTUs is a promising solution allowing to benefit from all available DNA material when computing biotic indices. The method can be easily extended to other biotic indices and groups of biological indicator.

Keywords: Metabarcoding, Biomonitoring, Environmental DNA, Diatoms, Phylogenetic signal

Introduction

The protection and conservation of ecosystems requires to accurately assess the quality of the environment over time (Ibáñez *et al.* 2010). Ecologists have developed a wide set of biotic indices to monitor ecological impacts of human activities, based on the principle that anthropic pressures shape biological communities (Chapman 1996). Hence, a large variety of indices are available to estimate environmental quality from the richness, diversity, structure, and functioning of biological communities.

Diatoms are unicellular eukaryotic algae encompassing a large taxonomic diversity (Round *et al.* 1990a). Because they have a relatively short generation time and communities respond strongly to changes in habitat quality, diatoms are recognized as powerful bioindicators of freshwater quality (Stevenson *et al.* 2010b; Rimet 2012). Most of the diatom biotic indices are based on species autecology and are usually derived from the equation of Zelinka and Marvan (1961). For example, the IPS index (Coste 1982) which is used in this paper as a case study is defined in Equation 1, where a_i is the relative abundance of species i in the sample, $IPSV_i$ its indicator value (tolerance) and $IPSS_i$ its pollution sensitivity (optimum).

$$IPS = \frac{\sum_{i=1}^n a_i \times IPSV_i \times IPSS_i}{\sum_{i=1}^n a_i \times IPSV_i} \quad (1)$$

The estimation of diatom indices like the IPS index require an accurate taxonomic inventory of the community (Besse-Lototskaya *et al.* 2011). Diatom inventories are traditionally based on the morphological identification of several hundred individuals under microscope (Prygiel *et al.* 2002). Given the diversity of diatoms (Mann & Vanormelingen 2013), this step is time-consuming, requires highly qualified staff and is prone to errors (Besse-Lototskaya *et al.* 2006). However, the development of methods to identify multiple taxa simultaneously from an environmental sample with standard genetic markers (DNA metabarcoding), combined with high-throughput sequencing technologies (HTS) have enabled the production of fast and cost-effective taxonomical inventories of communities (Taberlet *et al.* 2012a). Therefore, metabarcoding has been promoted as an attractive alternative to the traditional identification under microscope for biomonitoring (Baird & Hajibabaei 2012). Recent studies have shown that molecular inventories of diatom communities can be used to calculate various biotic indices (Kermarrec *et al.* 2014; Visco *et al.* 2015; Rivera *et al.* 2017; Vasselon *et al.* 2017b).

The classical approach to compute ecological indices with metabarcoding data consists in clustering DNA reads into operational taxonomic units (OTUs) and assign them a taxonomic name using a reference library (Kermarrec *et al.* 2014; Zimmermann *et al.* 2015). Once the list of OTUs converted into a taxonomic list, one can compute traditional bioassessment indices based on the ecological preferences of the species (IPSS and IPSV values in the case of IPS). However, this approach has some identified drawbacks, the most significant being that an important proportion of OTUs cannot be adequately classified into species because reference library are not sufficiently comprehensive. Hence, a large part of the biological diversity unraveled by DNA methods is discarded and cannot be used for bioassessment purposes.

To circumvent this problem, it has been suggested to skip the conversion from DNA reads to taxonomic entities and work directly on molecular data (Keck *et al.* 2017). In this respect, different strategies have been considered, including OTU-based indices (Apothéloz-Perret-Gentil *et al.* 2017) or the use of supervised machine learning algorithms to process genetic inventories (Cordier *et al.* 2017). Alternatively, Keck *et al.* (Keck *et al.* 2015, 2017) have suggested an approach based on the relationships existing between the phylogenetic position of species and their ecology (ie. the phylogenetic signal, Blomberg *et al.* 2003). The central idea is to combine an algorithm to place OTUs within a reference phylogeny and an algorithm to phylogenetically impute OTUs' ecological profiles (i.e. autecological values, here IPSS and IPSV) based on information available from neighbor species. However, implementing this approach for ecological assessment with metabarcoding requires to assess if DNA reads produced by HTS are long and informative enough to be accurately placed in a reference phylogeny and, once inserted, if the phylogenetic signal is strong enough to estimate precisely their ecological profiles.

In this paper, we aim to implement and test the phylogenetic approach (termed OTU-PITI, i.e. OTU Phylogenetic Insertion and Trait Imputation) for ecological assessment with metabarcoding data. We first compared the placement accuracy of short *rbcL* DNA reads (312 bp) as produced by HTS technologies with full length sequences of the *rbcL* gene. Second, we performed a cross-validation procedure to test whether phylogenetic imputation of species ecological profiles can be estimated from their phylogenetic positions. Finally, we tested the method with a dataset of 139 environmental river samples for which diatoms communities were analyzed using both microscopy and DNA metabarcoding. IPS indices based on the OTU-

PITI approach and on taxonomically assigned OTUs were then compared to IPS indices based on classical microscopy.

Material and Methods

Reference phylogenetic tree reconstruction

The reference phylogenetic tree was reconstructed from the chloroplast *rbcl* gene coding for the RuBisCO enzyme. This gene is recognized for its good performances to differentiate diatoms species and is a popular marker both for phylogenetic and metabarcoding studies. However, *rbcl* may have a limited ability to recover deep phylogenetic relationships within diatom clades (Theriot *et al.* 2011). Therefore, we used the phylogeny of diatoms published by (Theriot *et al.* 2015) and based on seven genes (*SSU*, *atpB*, *psaA*, *psaB*, *psbA*, *psbC* and *rbcl*) as a fixed guide in the reconstruction process. We extracted 1380 *rbcl* sequences from the curated library R-Syst::diatom (Rimet *et al.* 2016). The sequences were aligned using MUSCLE (Edgar 2004) and consensus sequences were computed using per-basis majority rule for 550 species. The new set of 550 sequences was then merged and re-aligned against the 208 sequences alignment of Theriot *et al.* (Theriot *et al.* 2015). Duplicated species were dropped, giving a final reference alignment of 604 species. The phylogeny was then reconstructed with RAxML 8.2.11 (Stamatakis 2014) using the phylogenetic tree of Theriot *et al.* (Theriot *et al.* 2015) as a topological constraint, a substitution model GTR+G+I, 200 runs and 1000 bootstraps (Fig S1, Supporting information). The tree was dated in relative time using PATHd8 (Britton *et al.* 2007).

Testing for short sequences placement

To test if a 312 bp *rbcl* barcode is sufficient to recover the phylogenetic position of the species, we sequentially dropped species from the reference phylogenetic tree and placed them using their reference barcode of 312 bp. The phylogenetic placement was performed using the Evolutionary Placement Algorithm (EPA; Berger *et al.* 2011) implemented in RAxML. To assess the quality of the barcode placement, we measured the distance between the insertion point of the full-length reference sequence (~1500 bp) edge and the insertion point of the placed barcode sequence (312 bp). This distance is expressed as the number of nodes located on the path that connects the two insertion points. Ideally, the barcode sequence is placed at the same location as the full-length sequence and the node distance is zero.

Testing for phylogenetic estimation of autecological values

We used a leave-one-out cross-validation (LOOCV) procedure to test whether the pollution sensitivity values (IPSS) and the indicator values (IPSV) of the species can be estimated accurately from their phylogenetic position. The analysis was performed on a subset of 237 species which were found both in the reference phylogenetic tree and the IPS data base. We sequentially estimated the IPSS and IPSV values of each species, given its phylogenetic position (as estimated using the barcode sequence; see above), and the known autecological values of the other species in the tree. The prediction was done using the framework introduced by Bruggeman, Heringa, and Brandt (2009) which estimates the phylogenetic covariance matrix parameters under a given evolution model and use it to impute the missing data as the best linear unbiased predictions (Ho *et al.* 2014). We tested 6 different phylogenetic models: Brownian motion (BM), Ornstein-Uhlenbeck (OU), Early-Burst (EB), Lambda, Delta and Kappa (see Goolsby *et al.* 2017 for details on the tested models).

Additionally we used an ad-hoc non-phylogenetic model (star) which assumes that the best estimate for a missing value is given by the mean of all observations. The performances of the different models were assessed and compared using the LOOCV mean squared error (MSE).

Sample collection

A total of 142 benthic water samples were collected from rivers as part of the 2016 French monitoring campaign for water quality assessment (Fig. S2, Supporting information). Benthic diatoms communities were collected by scraping at least 5 submerged stones using a toothbrush, as recommended by the European standard (*Water quality - Guidance standard for the routine sampling and pretreatment of benthic diatoms from rivers* 2016). Immediately after collection, each sampled biofilm was homogenized and divided into 2 subsamples to perform the molecular and morphological approaches. Each subsample was transferred into 50 mL Falcon tubes and preserved with a final concentration of at least 70 % of ethanol.

Morphological approach

Sample preparation, species identification and counting were performed by offices in charge of ecological assessment of French rivers in the context of the Water Framework Directive. Benthic samples were treated using 40 % H₂O₂ and HCl according to the European standard (AFNOR 2004). Resulting diatom samples were mounted in Naphrax and used to obtain permanent slides for microscopical analysis. A minimum of 400 diatoms valves were determined using classical European floras (AFNOR 2004).

Molecular laboratory methods

The preserved biofilm samples were centrifuged at 17,000 g during 30 minutes and the supernatant containing ethanol discarded. Total genomic DNA was extracted from the pellet using a non-commercial method based on Sigma-Aldrich GenElute™-LPA DNA precipitation, as described and recommended previously for diatom metabarcoding (Chonova *et al.* 2016; Vasselon *et al.* 2017a). In order to have technical replicates, two subsamples of each DNA extracts were used for subsequent PCR amplification and HTS, for a total of 284 DNA samples (142 x 2) sequenced. To enable the sequencing of all samples in a single Illumina run, 2 successive PCR were performed to prepare HTS libraries. (i) PCR1: DNA extracts were amplified in triplicate using the equimolar mixes of *Diat_rbcL_708F_1*, *708F_2*, *708F_3* and *R3_1*, *R3_2* as forward and reverse primers respectively (Vasselon *et al.* 2017b), allowing to focus a short fragment of the *rbcL* plastid gene (312 bp). Half of the P5 (CTTCCCTACACGACGCTCTCCGATCT) and P7 (GGAGTTCAGACGTGTGCTCTCCGATCT) Illumina adapters were included to the 5' part of the *rbcL* forward and reverse primers respectively. PCR1 amplifications were performed in a final volume of 25µL following mix and reaction conditions used in Vasselon, Rimet, *et al.* (Vasselon *et al.* 2017a), except the number of amplification cycles which was set to 33. (ii) PCR2: the 3 PCR1 replicates prepared for each DNA sample were pooled and sent to the "GenoToul Genomics and Transcriptomics" facility (GeT-PlaGe, Auzerville, France) where subsequent laboratory preparation were performed. PCR1 amplicons were purified and used as templates in the PCR2 which used Illumina-tailed primers targeting the half of P5 and P7 sequences. Finally, all generated 284 PCR2 amplicons were dual indexed and pooled into a single tube. Final pool was sequenced on an Illumina Miseq platform using the V3 paired-end sequencing kit (250 bp x 2).

HTS data analyses

Demultiplexed and overlapped Miseq data were delivered by the GeT-PlaGe sequencing platform (paired sequences overlap > 140 bp and mismatches < 0.1 %), resulting in 284 fastq files. A quality filtering was performed using Mothur software (Schloss *et al.* 2009) to remove DNA reads with: Phred quality score < 23 over a moving window = 25 bp, primer sequence mismatch > 1, homopolymer > 8 bp, ambiguous base > 0. Chimeras were removed using the Uchime algorithm (Edgar *et al.* 2011) available in Mothur. Then, all the fastq files were combined and de-replicated in order to keep only unique sequences with DNA read abundance > 2. Using the R-Syst::diatom library (Rimet *et al.* 2016) and the naïve Bayesian method (Wang *et al.* 2007), taxonomy was assigned to each DNA read with a confidence threshold > 85 %. DNA reads assigned to Bacillariophyta phylum were clustered into OTUs using a distance similarity threshold of 95 % as described in (Vasselon *et al.* 2017a). For each sample, the 2 replicates were merged and only the OTUs shared by both replicates were conserved in order to remove unrepresentative and spurious OTUs. Taxonomy of OTUs was defined as the consensus taxonomy of DNA reads (threshold > 80). A DNA representative sequence was determined for each OTU using the *Get.oturep* command in Mothur.

Biotic indices

For each site we computed four biotic indices, all based on the IPS index (Coste 1982). The first index IPS-MICROTAXO was computed from the relative abundances of the species estimated using classical microscopy. The second, IPS-DNATAXO was computed from the relative abundance of the OTUs after they were classified into species using Mothur. Since the IPSS and IPSV values are inherited from the taxonomical affiliation, the fraction of unclassified OTUs cannot be used for this index. The third index, IPS-DNAPHYLO takes into account all the OTUs. For this index the IPSS and IPSV values are phylogenetically imputed. OTUs were placed within the reference phylogenetic tree using their representative sequence (most abundant sequence) and the EPA algorithm. The IPSS and IPSV values of each OTU were estimated using rphylopars with the best evolution model selected at the cross-validation step (see above). Finally, the fourth index, IPS-DNAHYBRID is a combination of IPS-DNATAXO and IPS-DNAPHYLO: species IPSS and IPSV are used for OTUs which can be classified into species using Mothur, while the unclassified fraction of OTUs is used with phylogenetically imputed IPSS and IPSV values.

Results

Quality of read placements

Overall, barcode sequences allowed to place species accurately within the reference phylogeny (Fig. 1, Table S1, Supporting information). About 45% (272) of the species were placed exactly at the same location as the full-length sequence. Most of the species (508; 84%) were placed at a short distance, ≤ 3 nodes from the reference target. Only a few species were not placed correctly within the reference phylogeny (35 species; 5.8% at ≥ 10 nodes from the reference targets).

Quality of autecological values estimation

For IPSS, all phylogenetic models produced better predictions (lower MSE) than the non-phylogenetic star model (Fig. 2, Table S2, Supporting information). The best model with the lowest MSE was the lambda model which exhibited a 30% decrease of MSE compared to

the star model. For IPSV, only the OU model performed better than the star model but the difference was marginal (4%).

The estimated IPSS values for each species are mapped onto the reference phylogenetic tree in Fig. 3 and can be compared with the true IPSS values. For 150 (63%) of the species represented in green in Fig. 3, the absolute error (i.e. the absolute value of the difference between the estimated and the true IPSS) was found to be low (≤ 1), indicating a good prediction. The absolute error was ranging from 1 to 2 for 79 (33%) of the species (represented in orange), indicating a poor prediction quality. Finally, for a few species (8; 3%) the prediction quality was found to be very poor (absolute error > 2).

Morphological analysis

A total of 534 species were determined using microscopy. The dominant species were *Achnantheidium minutissimum* (14% of the valves determined), *Achnantheidium pyrenaicum* (6%), *Amphora pediculus* (5%), *Achnantheidium delmontii* (5%) and *Eolimna minima* (4%). For these 5 dominant taxa a reference barcode is present in the R-Syst::diatom library, except for *Achnantheidium delmontii*. Among the 100 most frequently determined species, 38 have a barcode in R-Syst::diatom and among the 534 species, only 114 species have a DNA barcode.

HTS analysis

The Illumina Miseq sequencing produced a total of 11,539,416 x 2 DNA reads. After all the bioinformatics processes, the OTU list obtained for the 142 samples included 682 OTUs composed by 3,119,226 DNA reads. After the taxonomic assignment, 362 OTUs were identified at the genus level (77 % of DNA reads) and 205 at the species level (58 % of DNA reads). Final molecular taxonomic list contained 28 families, 53 genera and 102 diatom species. The final list of OTUs with their taxonomic assignment and DNA representative sequence is available in the Supporting information (Table S3).

Performances of phylogenetic indices vs taxonomic indices

The distribution of IPS-MICROTAXO scores was positively skewed with a majority of high rated sites. DNA-based indices scores exhibited unimodal distribution with a restricted variability (few sites low rated and high rated). This was particularly true for IPS-DNAPHYLO which showed a variance of 0.12, much lower than the variance of IPS-MICROTAXO ($s^2 = 0.5$). The indices were strongly correlated with each other (all correlations > 0.49 ; Fig. 4). When comparing DNA-based indices with IPS-MICROTAXO, IPS-DNAHYBRID appeared to be the best index with the highest correlation ($r = 0.74$) and the lowest MSE (0.33). IPS-DNATAXO and IPS-DNAPHYLO exhibited similar correlation with IPS-MICROTAXO ($r = 0.69$ and $r = 0.70$, respectively) and similar MSE (0.45 and 0.43).

Discussion

DNA metabarcoding appears like a promising alternative to the traditional ways to characterize biodiversity and assess environmental quality. However, the massive quantities of genetic data produced by HTS challenge ecologists to think differently about the way biotic indices are computed (Keck *et al.* 2017). In this paper, we have introduced a new method based on phylogeny to compute biotic indices from DNA reads generated by metabarcoding workflows. The phylogenetic method is in line with the recent developments in taxonomy-free approaches for bioassessment which aim to bypass taxonomic reference libraries in order to

maximize the genetic information taken into account (Apothéloz-Perret-Gentil *et al.* 2017; Cordier *et al.* 2017). The phylogenetic method OTU-PITI has sound theoretical grounds. Indeed, phylogenetic imputation of missing values is based on the phylogenetic signal (i.e. the non-independence among species trait values because of their phylogenetic relatedness), a direct consequence of Darwin's principle of descent with modification (Felsenstein 1985).

The OTU-PITI approach is based on two main steps: first, the placement of DNA reads within the phylogeny and second, the estimation of their ecological values. We found that short *rbcL* marker (312 bp) gives satisfying results with most of the species barcodes placed exactly or very close to their reference position. This is consistent with the benchmark results obtained by Berger *et al.* (Berger *et al.* 2011) on short sequences (200 ± 60 bp) in the original publication of the Evolutionary Placement Algorithm. However, some species could not be placed correctly by the EPA (Fig. 1; see Table S1, Supporting information for the detailed list). The performances of the EPA for a given sequence may depend on many factors like the choice of genetic marker, the length of the sequence, and the presence of closely related taxa in the reference tree. In our case, it seems that wrong placements often involve species isolated in the phylogeny. Thus, increasing the phylogenetic coverage of underrepresented taxa may help to improve the placement of these species. Obviously, longer DNA reads capture more historical signal. Hence, the quality of reads insertion is also expected to improve as read lengths produced by HTS will increase (Tedersoo *et al.* 2017). Finally, it should be noted that most of the species which were wrongly placed are marine (e.g. *Guinardia striata*, *Stephanopyxis turris*) and therefore will not impede the computation of freshwater biotic indices like the IPS index.

Diatoms pollution sensitivity (IPSS) was much better predicted using a phylogenetic model (Pagel) than using the ad-hoc non-phylogenetic star model. This result is consistent with the presence of phylogenetic signal for ecological optima and pollution sensitivity (Keck *et al.* 2015). Nonetheless, some species were very poorly predicted (e.g. *Nitzschia soratensis*, *Terpsinoë musica*). Incorrect estimation can be the result of a wrong placement of the species within the phylogeny. For example, the high absolute error found for *Lemnicola hungarica* (2.55) could be explained by a rough phylogenetic placement of this species (node distance = 11). It is also clear that trait imputation is less effective when closely related species exhibit very contrasted trait values and therefore strongly depart from the underlying model of evolution (overdispersion). For example, *Halamphora oligotraphenta* and *Halamphora veneta* are two closely related species with very different ecological preferences, the former living in oligotrophic freshwater, while the latter is found in eutrophic habitats (Levkov 2009). As a result, the pollution sensitivity values of these two species are incorrectly predicted (Fig. 3). Overdispersion can be the consequence of recent evolutionary events and selection under active constraints like convergent evolution or character displacement. Unlike IPSS, the species tolerance value IPSV was poorly predicted by the phylogenetic models. The fact that the ad-hoc model (star) performed as good as the best phylogenetic model (OU) reflects a low phylogenetic signal, or a signal which cannot be appropriately modelled with the tested phylogenetic methods. A weak signal can be the result of trait instability and lability over time (Blomberg *et al.* 2003). IPSV being an approximate and partial measure of diatoms realized niche volume, its variability may be more related to interspecific interactions and non-genetic effects.

The two IPS-derived indices implementing the OTU-PITI approach (IPS-DNAPHYLO and IPS-DNAHYBRID) were strongly correlated to the index estimated from microscopy. However, IPS-DNAPHYLO had a restricted range of values, with a tendency to overestimate the score of bad quality sites and underestimate the score of good quality sites. This tendency is likely to be caused by the phylogenetic imputation algorithm which has been shown to be a form of

kriging (Cressie 1993) in a phylogenetic context (Ho *et al.* 2014). As an inverse distance weighting method, kriging is subject to a smoothing effect and do not reproduce the histogram of the sample data (Isaaks & Srivastava 1989). One solution could be to estimate the strength of smoothing from the LOOCV data and apply a correction factor to the OTUs estimated ecological values. The true ecological values being more reliable than the phylogenetically imputed ones, we advocate for the use of the DNAHYBRID which benefits from the ecological values of the assigned species if available while using 100% of the OTUs, thanks to the OTU-PITI approach. In our study, the IPS-DNAHYBRID is the molecular index correlating the best with the index based on microscopy.

The OTU-PITI approach solves two important problems that scientists and environmental managers recurrently face when using metabarcoding data to compute biotic indices. The first problem is the incompleteness of reference libraries connecting DNA barcode sequences to taxonomic names. An incomplete library strongly limits the proportion of OTUs which can be taxonomically assigned and used for indices calculation. In this study, the reference library covered 21% of the species detected using microscopy. As a consequence, the proportion of OTUs assigned at species level was only 30%, similar to the proportions obtained in previous diatom studies (e.g. (Apothéloz-Perret-Gentil *et al.* 2017; Rivera *et al.* 2017; Vasselon *et al.* 2017b) respectively 35%, 23% and 35.7%). Additionally, the OTU-PITI approach offers a convenient solution to the incompleteness of ecological libraries connecting taxa and ecological values. Some species, detected either by microscopy or DNA, do not have IPSS and IPSV values. Therefore, they cannot be used to compute the IPS index. For example, in this study, 9 OTUs were assigned to species which were not found in the IPS library. This is often the result of taxonomic names discrepancies among libraries (synonyms, misspellings) but in some cases autecological information can simply be missing. The OTU-PITI allowed to estimate IPSS and IPSV values for these OTUs and include them in the calculation of IPS. This feature is particularly interesting for the implementation of biotic indices including a restricted number of taxa, or to extend the use of well-established indices to new habitats and new regions with endemic taxa.

Two other taxonomy-free approaches to compute molecular indices have been recently introduced in the literature. First, Apothéloz *et al.* (Apothéloz-Perret-Gentil *et al.* 2017) proposed to assign ecological values directly to OTUs. Second, Cordier *et al.* (Cordier *et al.* 2017) investigated the use of supervised machine learning regression to infer indices values from lists of OTUs. The main advantage of the OTU-PITI over these two approaches is that it does not require the collection of chemical and physical measurements to train or calibrate the model which makes it a ready-to-use tool, not restricted to the geographical area of the training data. Conversely, a well-trained machine learning classifier used within its geographical scope will probably outperformed the OTU-PITI approach. As advocated by Keck *et al.* (Keck *et al.* 2017), OTU-PITI, OTU-based indices and machine learning are complementary tools which should make it possible to make better use of genetic data in the future.

Our knowledge of biodiversity is very unbalanced. Microscopic organisms, which include diatoms, are extremely diversified and largely unknown. Thus, OTU-PITI can be a very interesting way to fill the gaps, pending the availability of comprehensive taxonomic and ecological libraries. Here we have shown that this approach can be successfully applied to use the unclassified DNA material which is normally discarded from biotic indices computation. The range of applications of the OTU-PITI is very large: the method can be applied to any biotic index and any group of biological indicator, provided that an accurate phylogeny is available. Moreover, traits values can be modeled and estimated within phylogenetic multivariate frameworks (Clavel *et al.* 2015; Goolsby *et al.* 2017). Multiple biological traits and functional

groups come with several advantages compared to biotic autecological indices (Bonada *et al.* 2006; Tapolczai *et al.* 2016). Thus, the OTU-PITI approach could be the way to integrate the immense diversity revealed by metabarcoding and a step towards a functional biomonitoring.

Acknowledgment

This article is based upon work from COST Action DNAqua-Net (CA15219), supported by the COST (European Cooperation in Science and Technology) program. We thank Cécile Chardon who performed the DNA extractions, amplifications, and the preparation of libraries. Samplings and microscopic analyses were carried out by the Dreal (Direction Régionale de l'Environnement de l'aménagement et du Logement) Aquitaine (D. Sagnet), Auvergne (F. Véry), Bourgogne (V. Peeters), Bretagne (G. Gicquiaud), Centre (S. Saadat, C. Karabaghli), Franche-Comté (E. Parmentier), Limousin (J.-M. Vouters), Lorraine (D. Heudre), Midi Pyrénées (E. Seigneur), Pays de la Loire (D. Guillard), Rhône-Alpes (R. Chavaux), Normandie (F. Petel), Nord Pas de Calais (N. Zydek), and the private offices Aquabio (R. Marcel, B. Fontan), Asconit (L. Kermarrec, E. Ponton), Sage (A. Rolland, J.-P. Vulliet, C. Geret). We thank the French Water Agencies, Artois-Picardie (C. Lesniak), Rhône-Méditerranée et Corse (L. Imbert, F. Repellini), Adour-Garonne (M. Durand, J.-P. Rebillard, M. Saut), Rhin-Meuse (J.-L. Matte, G. Demortier), Loire-Bretagne (J. Durocher), Seine-Normandie (M. Berdoulay) who funded the microscopical analyses. The AFB (Agence Française de la Biodiversité) funded the sequencing, which was realized in the GeT-PlaGe sequencing platform. We thank the Swedish Agency for Marine and Water Management for a contribution via the program Environmental monitoring (project 2014-16, Development of the diatom barcoding method for freshwater).

Author Contributions

FK and VV conceived the study and performed the data analyses. FK wrote the paper with significant contributions from all authors. All authors gave final approval for publication.

References

- Apothéloz-Perret-Gentil, L., Cordonier, A., Straub, F., Iseli, J., Esling, P. & Pawlowski, J. (2017). Taxonomy-free molecular diatom index for high-throughput eDNA biomonitoring. *Molecular Ecology Resources*.
- Baird, D.J. & Hajibabaei, M. (2012). Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Molecular Ecology*, 21, 2039–2044.
- Berger, S.A., Krompass, D. & Stamatakis, A. (2011). Performance, Accuracy, and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood. *Systematic Biology*, 60, 291–302.
- Besse-Lototskaya, A., Verdonschot, P.F., Coste, M. & Van de Vijver, B. (2011). Evaluation of European diatom trophic indices. *Ecological Indicators*, 11, 456–467.
- Besse-Lototskaya, A., Verdonschot, P.F.M. & Sinkeldam, J.A. (2006). Uncertainty in diatom assessment: sampling, identification and counting variation. *Hydrobiologia*, 566, 247–260.
- Blomberg, S.P., Garland, T. & Ives, A.R. (2003). Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, 57, 717–745.
- Bonada, N., Prat, N., Resh, V.H. & Statzner, B. (2006). Developments in aquatic insect biomonitoring: a comparative analysis of recent approaches. *Annual Review of Entomology*, 51, 495–523.

- Britton, T., Anderson, C.L., Jacquet, D., Lundqvist, S. & Bremer, K. (2007). Estimating Divergence Times in Large Phylogenetic Trees. *Systematic Biology*, 56, 741–752.
- Bruggeman, J., Heringa, J. & Brandt, B.W. (2009). PhyloPars: estimation of missing parameter values using phylogeny. *Nucleic Acids Research*, 37, W179–W184.
- Chapman, D.V. (1996). *Water quality assessments: a guide to the use of biota, sediments and water in environmental monitoring*. E & Fn Spon London.
- Chonova, T., Keck, F., Labanowski, J., Montuelle, B., Rimet, F. & Bouchez, A. (2016). Separate treatment of hospital and urban wastewaters: A real scale comparison of effluents and their effect on microbial communities. *Science of The Total Environment*, 542, 965–975.
- Clavel, J., Escarguel, G. & Merceron, G. (2015). mvmorph: an r package for fitting multivariate evolutionary models to morphometric data. *Methods in Ecology and Evolution*, 6, 1311–1319.
- Cordier, T., Esling, P., Lejzerowicz, F., Visco, J., Ouadahi, A., Martins, C., Cedhagen, T. & Pawlowski, J. (2017). Predicting the Ecological Quality Status of Marine Environments from eDNA Metabarcoding Data Using Supervised Machine Learning. *Environmental Science & Technology*, 51, 9118–9126.
- Coste, M. (1982). *Étude des méthodes biologiques d'appréciation quantitative de la qualité des eaux*. Cemagref.
- Cressie, N.A.C. (1993). *Statistics for Spatial Data*. Wiley, New York.
- Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32, 1792–1797.
- Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C. & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, 27, 2194–2200.
- European Committee for Standardization. (2014). *Water quality - Guidance standard for the identification, enumeration and interpretation of benthic diatom samples from running waters*. Brussels.
- European Committee for Standardization. (2016). *Water quality - Guidance standard for the routine sampling and pretreatment of benthic diatoms from rivers*. Brussels.
- Felsenstein, J. (1985). Phylogenies and the Comparative Method. *The American Naturalist*, 125, 1–15.
- Goolsby, E.W., Bruggeman, J. & Ané, C. (2017). Rphylopars: fast multivariate phylogenetic comparative methods for missing data and within-species variation. *Methods in Ecology and Evolution*, 8, 22–27.
- Ho, T., Si, L. & Ané, C. (2014). A Linear-Time Algorithm for Gaussian and Non-Gaussian Trait Evolution Models. *Systematic Biology*, 63, 397–408.
- Ibáñez, C., Caiola, N., Sharpe, P. & Trobajo, R. (2010). Ecological indicators to assess the health of river ecosystems. *Handbook of Ecological Indicators for Assessment of Ecosystem Health* (eds S.E. Jørgensen, F.-L. Xu & R. Costanza), pp. 447–464. CRC Press, Boca Raton, Florida.
- Isaaks, E.H. & Srivastava, R.M. (1989). *An Introduction to Applied Geostatistics*. Oxford University Press, New York.
- Keck, F., Rimet, F., Franc, A. & Bouchez, A. (2016). Phylogenetic signal in diatom ecology: Perspectives for aquatic ecosystems biomonitoring. *Ecological Applications*, 26, 861–872.
- Keck, F., Vasselon, V., Tapolczai, K., Rimet, F. & Bouchez, A. (2017). Freshwater biomonitoring in the Information Age. *Frontiers in Ecology and the Environment*, 15, 266–274.
- Kermarrec, L., Franc, A., Rimet, F., Chaumeil, P., Frigerio, J.-M., Humbert, J.-F. & Bouchez, A. (2014). A next-generation sequencing approach to river biomonitoring using benthic diatoms. *Freshwater Science*, 33, 349–363.
- Levkov, Z. (2009). *Amphora sensu lato* (H. Lange-Bertalot, Ed.). Gantner Verlag.
- Mann, D.G. & Vanormelingen, P. (2013). An inordinate fondness? The number, distributions, and origins of diatom species. *Journal of Eukaryotic Microbiology*, 60, 414–420.

- Prygiel, J., Carpentier, P., Almeida, S., Coste, M., Druart, J.-C., Ector, L., Guillard, D., Honoré, M.-A., Iserentant, R. & Ledeganck, P. (2002). Determination of the biological diatom index (IBD NF T 90–354): results of an intercomparison exercise. *Journal of Applied Phycology*, 14, 27–39.
- Rimet, F. (2012). Recent views on river pollution and diatoms. *Hydrobiologia*, 683, 1–24.
- Rimet, F., Chaumeil, P., Keck, F., Kermarrec, L., Vasselon, V., Kahlert, M., Franc, A. & Bouchez, A. (2016). R-Syst::diatom: an open-access and curated barcode database for diatoms and freshwater monitoring. Database, 2016, baw016.
- Rivera, S.F., Vasselon, V., Jacquet, S., Bouchez, A., Ariztegui, D. & Rimet, F. Metabarcoding of lake benthic diatoms: from structure assemblages to ecological assessment. *Hydrobiologia*.
- Round, F.E., Crawford, R.M. & Mann, D.G. (1990). *The diatoms: biology and morphology of the genera*. Cambridge University Press, Cambridge, UK.
- Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., Horn, D.J.V. & Weber, C.F. (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology*, 75, 7537–7541.
- Stamatakis, A. (2014). RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics*, btu033.
- Stevenson, R.J., Yangdong, P. & Van Dam, H. (2010). Assessing environmental conditions in rivers and streams with diatoms. *The Diatoms: Applications for the Environmental and Earth Sciences* (eds J.P. Smol & E.F. Stoermer), pp. 55–85. Cambridge University Press.
- Taberlet, P., Coissac, E., Hajibabaei, M. & Rieseberg, L.H. (2012). Environmental DNA. *Molecular Ecology*, 21, 1789–1793.
- Tapolczai, K., Bouchez, A., Stenger-Kovács, C., Padisák, J. & Rimet, F. (2016). Trait-based ecological classifications for benthic algae: review and perspectives. *Hydrobiologia*, 776, 1–17.
- Tedersoo, L., Tooming-Klunderud, A. & Anslan, S. (2017). PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. *New Phytologist*.
- Theriot, E.C., Ashworth, M.P., Nakov, T., Ruck, E. & Jansen, R.K. (2015). Dissecting signal and noise in diatom chloroplast protein encoding genes with phylogenetic information profiling. *Molecular Phylogenetics and Evolution*, 89, 28–36.
- Theriot, E.C., Ruck, E., Ashworth, M., Nakov, T. & Jansen, R.K. (2011). Status of the pursuit of the diatom phylogeny: Are traditional views and new molecular paradigms really that different? *The Diatom World* (eds J. Seckbach & J.P. Kociolek), pp. 119–142. Springer, New York, USA.
- Vasselon, V., Domaizon, I., Rimet, F., Kahlert, M. & Bouchez, A. (2017a). Application of high-throughput sequencing (HTS) metabarcoding to diatom biomonitoring: Do DNA extraction methods matter? *Freshwater Science*, 36, 162–177.
- Vasselon, V., Rimet, F., Tapolczai, K. & Bouchez, A. (2017b). Assessing ecological status with diatoms DNA metabarcoding: Scaling-up on a WFD monitoring network (Mayotte island, France). *Ecological Indicators*, 82, 1–12.
- Visco, J.A., Apothéoz-Perret-Gentil, L., Cordonier, A., Esling, P., Pillet, L. & Pawlowski, J. (2015). Environmental Monitoring: Inferring the Diatom Index from Next-Generation Sequencing Data. *Environmental Science & Technology*, 49, 7597–7605.
- Wang, Q., Garrity, G.M., Tiedje, J.M. & Cole, J.R. (2007). Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied and Environmental Microbiology*, 73, 5261–5267.
- Zelinka, M. & Marvan, P. (1961). Zur präzisierung der biologischen klassifikation der reinheit fließender gewässer. *Archiv für Hydrobiologie*, 57, 389–407.

Zimmermann, J., Glöckner, G., Jahn, R., Enke, N. & Gemeinholzer, B. (2015). Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. *Molecular Ecology Resources*, 15, 526–542.

Figures

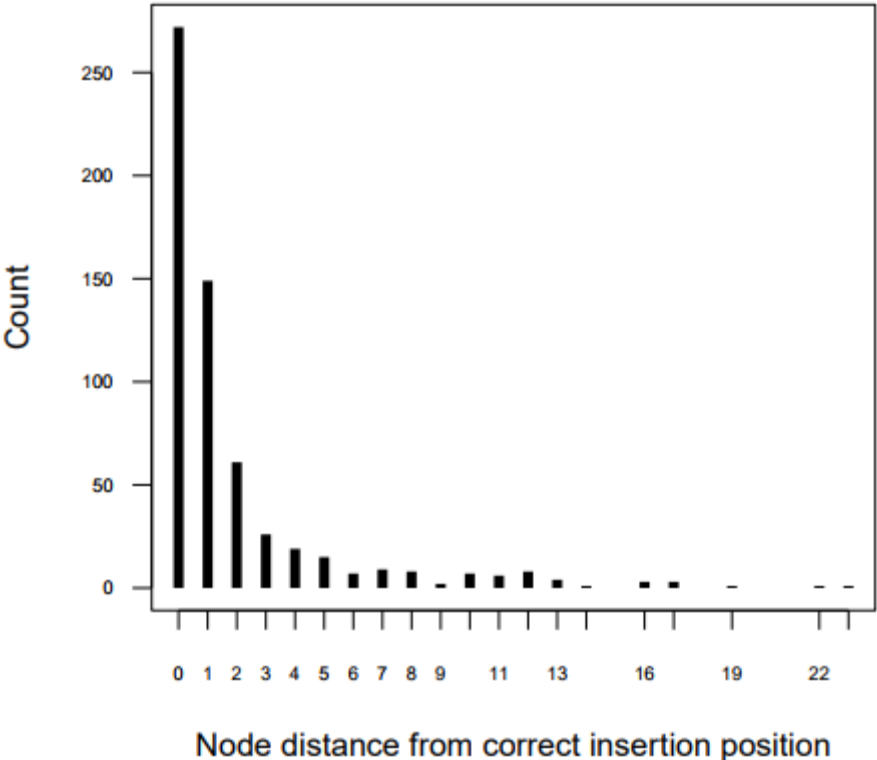


Fig. 1 Histogram showing the placement accuracy of the 604 species from the reference tree using 312 bp *rbcL* barcode sequences and the EPA algorithm.

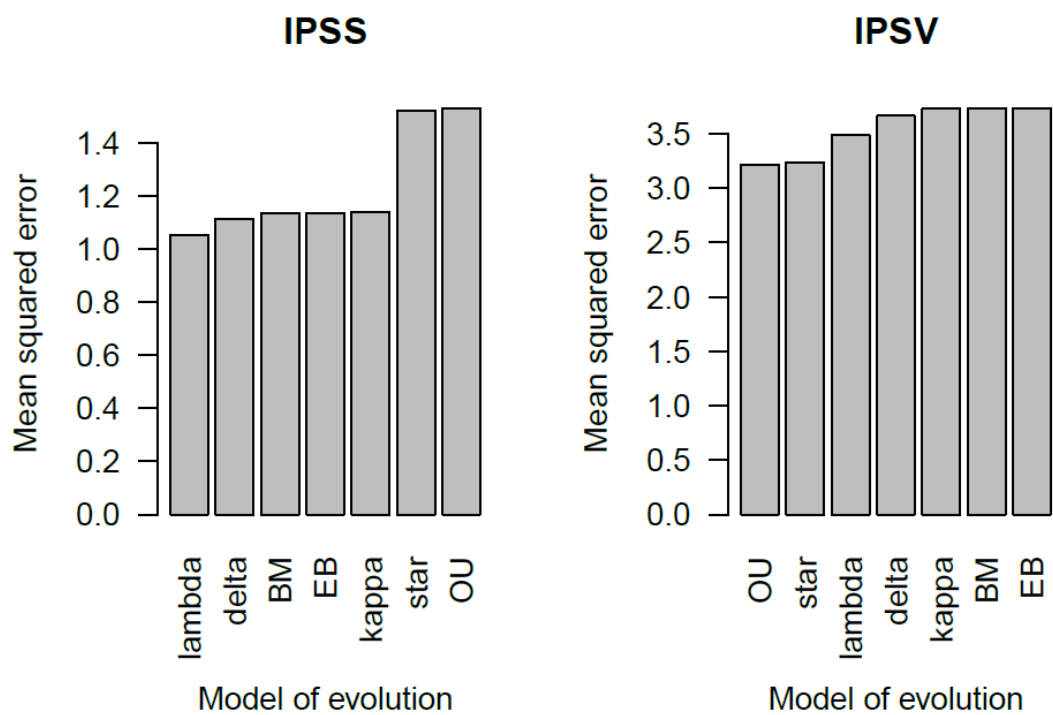


Fig. 2 Barplots showing the LOOCV mean squared error of each model for IPSS and IPSV estimation.

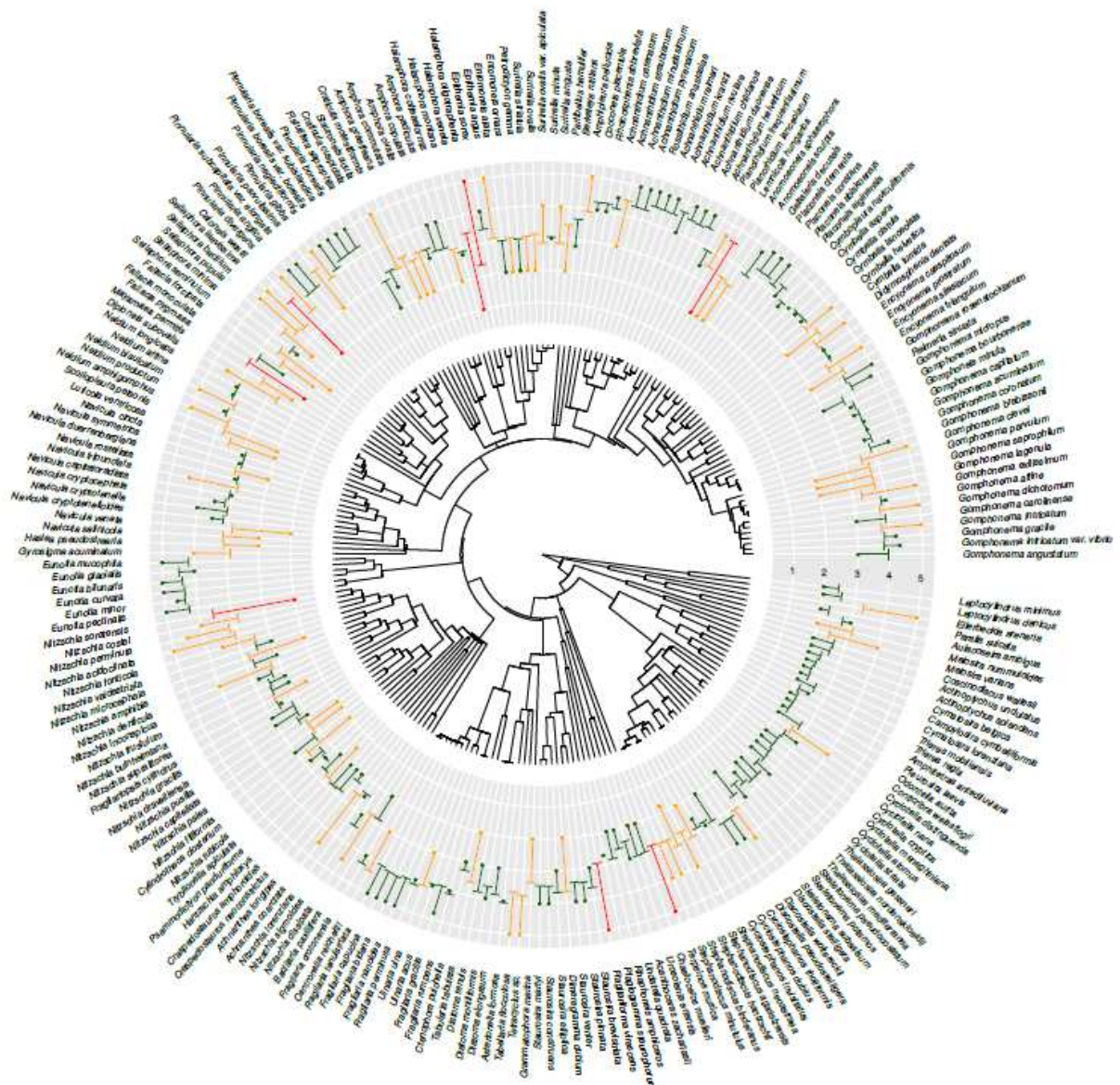


Fig. 3 Phylogenetic tree representing 236 diatoms species for which both phylogenetic position and IPSS value were available. For each species, true IPSS value is represented as a point, while its estimated IPSS value is represented as a dash. Low absolute errors (≤ 1) are represented in green, medium absolute errors (> 1 and ≤ 2) in orange and high absolute errors (> 2) in red.

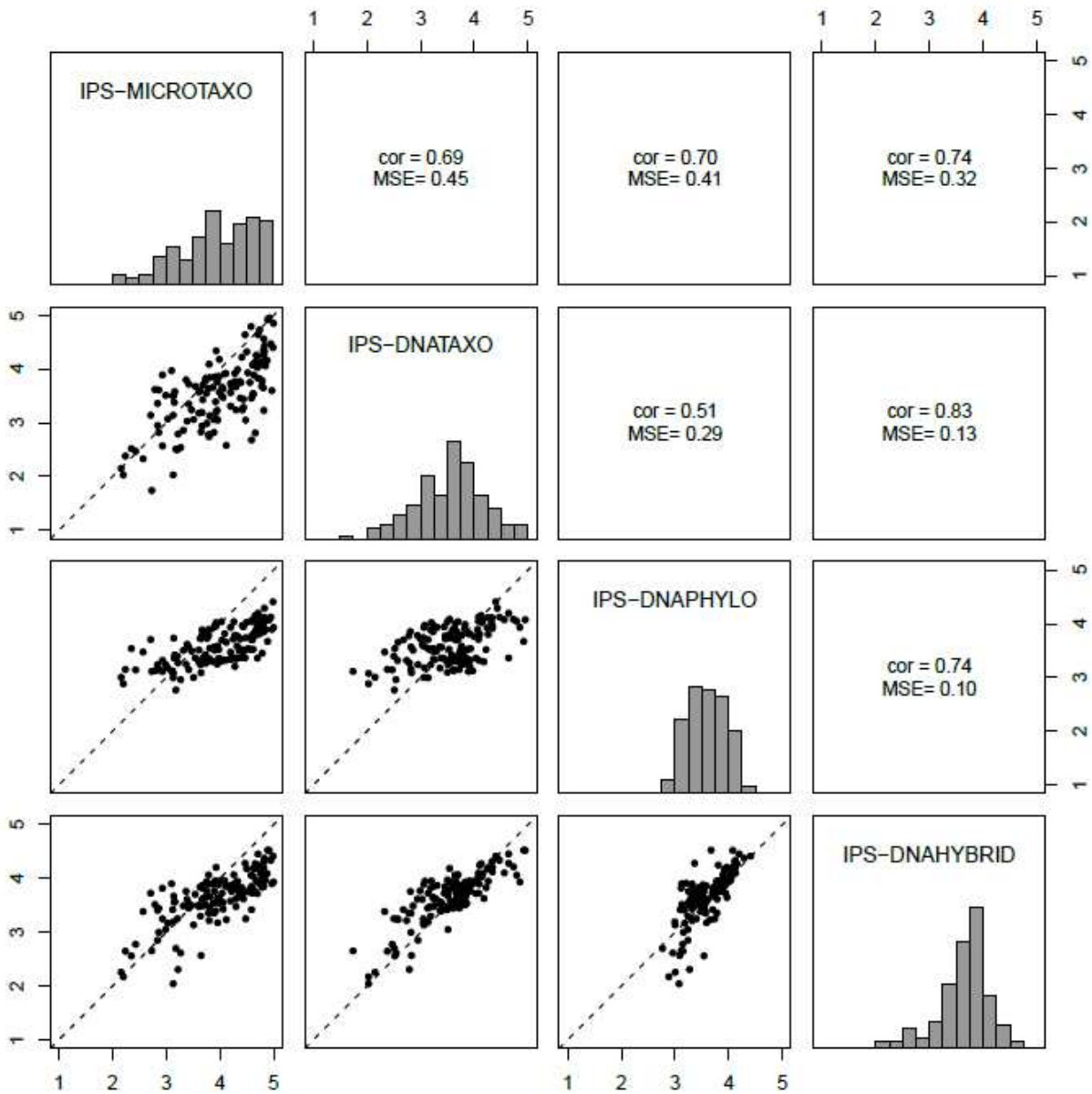


Fig. 4 Distributions and relationships between the 4 indices computed for 139 environmental samples. Diagonal: Histograms of the distribution of each index expressed as frequencies. Lower triangle: Scatterplots showing the relationships between the indices. The dashed lines represent the full equivalence between the indices. Upper triangle: correlation (cor) and mean squared error (MSE) between the indices.

Supporting Information (non fournies dans le manuscrit)

Fig. S1 Phylogenetic tree with bootstrap support values.

Fig. S2 Map of sampling sites.

Table S1 Node distances between references and barcode placements, detailed per species.

Table S2 Leave one out cross-validation results for IPSS and IPSV phylogenetic imputation (best model), detailed per species.

Table S3 List of OTUs, number of copies, taxonomic affiliations and DNA representative sequences