Nécessité, potentiel et limitations de l'approche en unités taxonomiques moléculaires pour analyser la biodiversité de l'ADN environnemental des poissons

Virginie Marques

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Écologie des Communautés

École doctorale GAIA

Unité de recherche UMR MARBEC

Nécessité, potentiel et limitations de l'approche en unités taxonomiques moléculaires pour analyser la biodiversité de l'ADN environnemental des poissons

Présentée par Virginie MARQUES
Le 26 Novembre 2020

Sous la direction de David MOUILLOT et Stéphanie MANEL

Devant le jury composé de

Louis BERNATCHEZ, Professeur Université de Laval, Département de Biologie, Québec	Rapporteur
Marie-Josée FORTIN, Professeure Université Toronto, Canada	Rapportrice
Jérôme MURIENNE, Chargé de recherches CNRS, UMR 5174, Toulouse	Examinateur
Emmanuel PARADIS, Directeur de recherches IRD, ISEM Montpellier	Examinateur
Stéphanie MANEL, Directrice d'études EPHE, CEFE/CNRS, Montpellier	Invitée
David MOUILLOT, Professeur Université de Montpellier, UMR MARBEC, Montpellier	Directeur de thèse
Claude MIAUD, Directeur d'études EPHE, CEFE/CNRS, Montpellier	Président du jury
Tony DEJEAN, Directeur SPYGEN, Le Bourget du Lac	Invité

UNIVERSITÉ DE MONTPELLIER

Résumé

La vitesse et l'intensité des changements globaux nécessitent de nouveaux moyens d'observations de la biodiversité qui soient rapides, non-destructifs, standardisés, déployables à large échelle et dans les écosystèmes les plus reculés (océan profond). Les méthodes de recensement classiques reposent sur l'identification morphologique ou sonore des espèces, mais celles-ci sont coûteuses en temps et en expertise. Au-delà de ces signaux, les animaux laissent aussi des traces d'ADN dans leur environnement sous la forme de cellules dermiques, de mucus ou de fèces. Le metabarcoding de cet ADN environnemental (ADNe) consiste à le collecter, l'amplifier et le séquencer pour identifier les espèces présentes grâce à des bases de séquences génétiques de référence. Or, ces bases de référence sont incomplètes, ce qui limite fortement le potentiel de l'ADNe pour révéler la biodiversité présente. Cette thèse a pour but de développer une approche alternative basée sur des unités taxonomiques moléculaires (MOTUs) pour analyser la biodiversité des macroorganismes aquatiques, et plus particulièrement celle des poissons osseux. J'ai tout d'abord réalisé une synthèse globale et spatialisée de la couverture taxonomique des bases de référence de séquences génétiques pour tous les poissons osseux, qui montre une sous-représentation des espèces de la zone tropicale ainsi que des lacunes concernant les espèces menacées et non-indigènes. Seules 13% des espèces de poisson sont séquencées pour le marqueur le plus commun, ce qui exclut toute ambition d'analyse exhaustive de la biodiversité par assignation aux espèces à court et moyen terme. En conséquence, j'ai développé un pipeline bio-informatique pour générer des estimations de la diversité en unités taxonomiques moléculaires (MOTUs) par famille de poissons. Les résultats démontrent que cette diversité en MOTUs représente un excellent substitut de la diversité en espèces à différentes échelles spatiales. Ensuite une application du metabarcoding de l'ADNe et de l'approche en MOTUs a permis d'estimer la diversité fonctionnelle, basée sur les traits des espèces, et la diversité phylogénétique, basée sur l'histoire évolutive des espèces, des poissons tropicaux de manière plus exhaustive que des méthodes traditionnelles (vidéos, plongées). Enfin, dans une première analyse globale de la diversité des récifs coralliens en ADNe, qui rassemble 251 échantillons récoltés depuis l'Océan Indien jusque dans les Caraïbes, l'approche en MOTUs permet de reconstruire les gradients biogéographiques des poissons mais aussi de révéler une hétérogénéité spatiale locale jusqu'alors sous-estimée. Alors qu'il est aujourd'hui crucial de mettre en place des méthodes de suivi efficaces, non dépendantes de spécialistes et à haute fréquence temporelle pour mieux comprendre les effets des changements globaux sur la biodiversité, ces travaux démontrent tout le potentiel de l'ADNe avec approche en MOTUs pour construire des indicateurs robustes de plusieurs facettes de la biodiversité à plusieurs échelles, mais aussi tester les hypothèses théoriques sous-jacentes à la distribution de cette biodiversité.

Mots-clés : ADNe, biodiversité, communauté, récifs coralliens, clustering, MOTU

Abstract

The speed and intensity of global change requires new means of observing biodiversity that are rapid, non-destructive, standardized, widely deployable and in remote ecosystems (deep sea). Conventional inventory methods are based on morphological or acoustic identification of species, which are costly in terms of time and expertise. Beyond these signals, animals also leave traces of DNA in their environment in the form of dermal cells, mucus or feces. The metabarcoding of this environmental DNA (eDNA) consists in collecting this DNA, amplifying and sequencing it to identify the species present using a genetic reference database. However, these reference databases are incomplete, which severely limits the potential of eDNA. The aim of this thesis is to develop an alternative approach based on molecular taxonomic units (MOTUs) to analyze the biodiversity of aquatic macroorganisms, and more particularly that of bony fish. I first performed a global and spatialized synthesis of the taxonomic coverage of the genetic reference database for all bony fishes, which shows an under-representation of species in the tropical zone as well as taxonomic gaps for endangered and non-indigenous species. Only 13% of fish species are sequenced for the most common marker, which excludes any ambition for an exhaustive analysis of biodiversity using only species-level assignments in the short or medium term. Consequently, I have developed a bioinformatics pipeline to generate estimates of diversity using molecular taxonomic units (MOTUs) by fish family. It shows how this MOTU diversity represents an excellent proxy for species diversity at different spatial scales. Then an application of eDNA metabarcoding and the MOTUs approach allowed to estimate the functional diversity, based on species traits, and the phylogenetic diversity, based on the evolutionary history of the species, of tropical fishes in a more exhaustive way than traditional methods (videos, dives). Finally, in a first global analysis of coral reef diversity in eDNA, which brings together 251 samples collected from the Indian Ocean to the Caribbean, the MOTUs approach allows the reconstruction of major trends in fish biogeography but also reveals local spatial heterogeneity hitherto underestimated. While it is now crucial to set up efficient, non-specialist dependent and high temporal frequency monitoring methods to better understand the effects of global changes on biodiversity, this work demonstrates the full potential of eDNA using a MOTUs approach to build robust indicators of several facets of biodiversity at several scales, but also to test theoretical hypotheses underlying the distribution of this biodiversity.

Keywords: eDNA, biodiversity, community, coral reef, clustering, MOTU

Remerciements

En premier lieu, je souhaite remercier les membres du jury Marie-Josée Fortin, Louis Bernatchez, Emmanuel Paradis et Jérôme Murienne d'avoir accepté d'évaluer ce travail de thèse.

Ensuite, un énorme merci à David Mouillot, de m'avoir donné ma chance il y a maintenant presque 4 ans en master, puis de m'avoir accompagnée tout au long de cette thèse. Ton optimisme à toute épreuve a certainement contribué au succès de ce travail. Merci pour ton implication et pour les discussions scientifiques toujours productives et motivantes même dans les moments de doute, merci d'être un aussi bon mentor. Un grand merci à Stéphanie Manel d'avoir accepté le rôle de co-directrice en cours de thèse et de m'avoir ouvert les portes du CEFE. Merci pour ton soutien sans faille et ta grande disponibilité. J'ai énormément appris pendant ces trois années et c'est en partie grâce à vous. C'est avec un grand plaisir que je continue l'aventure Montpelliéraine encore quelques mois avec vous.

Les missions de terrain sont une composante fondamentale de cette thèse, auxquelles j'ai eu la chance de participer à six reprises. Les histoires sont nombreuses et les missions jamais de tout repos. Un grand merci à toutes les personnes qui ont participé à ces aventures, j'en garde des souvenirs inoubliables. Merci également à toutes les personnes qui ont aidé à collecter des échantillons ADNe pour alimenter la base de données. Merci aux nombreux membres d'équipage des navires qui ont rendu ces missions possibles. Merci aux patchs anti mal de mer d'exister.

Au cours de ces expériences aux quatre coins du monde, merci à Tom d'avoir partagé ta grande expérience des embarquements et des missions, à Clara pour ton humour et ta bonne humeur constante malgré les odeurs de poisson pourri, à Jonathan pour la formation ADNe sous la houle, à Gaël pour m'avoir accompagnée en Polynésie Française, pas évident comme destination, à Andréa pour ta gentillesse, ta bonne humeur et pour nous avoir accueilli chez toi à Santa Marta, à Régis pour la gestion de mission, le partage de ton expérience et les nombreuses photos qui illustrent ce travail, à Eva pour l'organisation sans faille et l'assistance sur le terrain, à Laure et Nicolas pour les longs moments aux aéroports, en mer et coincés au port (je n'en dirais pas plus !).

Un énorme merci à JB, pour tous les moments sur le terrain à essuyer les galères et mais aussi à profiter de ces expériences hors du commun, pour ta grande expertise en nœuds marins et en marques de rhum. C'est un plaisir de travailler et de collaborer avec toi. Merci à MC pour la cohabitation

toujours joyeuse au cours des missions, pour être ma déménageuse officielle et pour les guacamoles. Un peu moins merci pour les odeurs d'intestins de poissons pourris. Merci à Camille A. pour les moments en mer et l'ambiance musicale au labo humide, c'est un plaisir de collaborer avec toi une fois revenue sur la terre ferme. Merci à Sebastien, pour les discussions poissons et fonctio qui ont amélioré la qualité de ce travail. Merci à Florine avec qui l'aventure ADNe a commencé, pour ta bonne humeur et le soutien moral apporté tout au long de cette première année qui fut remplie d'embûches.

Merci à l'équipe MARBEC et aux nombreuses personnes passées par le bureau des temporaires (où on n'est plus si temporaires que ça au bout de 3 ans). A Arthur pour ta bonne humeur, tes conseils et les échanges geeks, à Camille M. pour les moments détente devant les vidéos de poissons et pour nous rappeler d'avancer dans nos thèses (j'espère que le package est bientôt prêt), Amandine pour ton soutien et tes conseils. Merci à tous les doc longue durée ou de passage pour les pauses midi, Justine, Thomas, Criscely, Valentina, Tanguy, Raquel, Elsa, Laura, Angela et certainement d'autres noms que j'oublie.

Aux collègues du CEFE, merci pour votre accueil. A Émilie en particulier, ta bonne humeur est contagieuse et c'est toujours un plaisir d'échanger avec toi au labo et en dehors. A Pierre-Édouard, ton apport scientifique a été déterminant dans le déroulement de cette thèse, merci d'être toujours aussi positif et de m'avoir appris les rudiments de la bio-informatique. A Coline, Pauline, Laetitia et Laura, merci d'avoir toujours rendu mes visites au CEFE agréables.

Merci à Alice et Tony de SPYGEN, qui m'ont toujours accordé de leur temps malgré leurs plannings surchargés et qui ont transmis une partie de leur savoir sur l'ADNe. Merci à l'équipe de SPYGEN qui ont traité l'ensemble des échantillons au laboratoire et longuement préparé les kits d'échantillonnage.

Je tiens à adresser un remerciement particulier à Monnira, ta générosité envers la jeune étudiante désargentée que j'étais m'a donné des moyens matériels d'accéder à la recherche dans les récifs coralliens en master, je ne sais pas si j'en serais là aujourd'hui si je n'avais pas pu saisir cette opportunité.

Enfin, Flo un immense merci pour tout, ta présence, ton soutien et pour embellir mon quotidien.

Table des matières

Introduction Générale

1. Des changements rapides de la biodiversité

1.1. Changements globaux et perte de biodiversité

La biosphère comprend environ 1.9 millions d'espèces décrites, parmi lesquels 1.3 millions de métazoaires, avec un nombre total d'espèces estimé entre 3 et 100 millions (Mora et al. 2011, Grosberg et al. 2012, Costello et al. 2013, Larsen et al. 2017, Roskov et al. 2019). Parmi cette diversité, on compte près de 70,000 espèces de vertébrés décrites avec de nombreuses restantes à découvrir (Mora et al. 2011, Appeltans et al. 2012, Roskov et al. 2019). Cette biodiversité est à la base du fonctionnement des écosystèmes et essentielle à l'existence humaine. Les activités d'origine humaine exercent une pression croissante sur les écosystèmes et la biodiversité (Dirzo et al. 2014, Díaz et al. 2019a). Récemment, l'IPBES (Plateforme Intergouvernementale Scientifique et Politique sur la Biodiversité et les services écosystémiques) a réuni plus de 150 scientifiques pour rédiger le rapport de l'évaluation mondiale 2019 sur la biodiversité et les services écosystémiques à destination de la société civile (Díaz et al. 2019b), synthétisant plus de 15,000 publications scientifiques. Leurs conclusions révèlent la détérioration importante et globale des écosystèmes et de la biodiversité à travers le monde. Le rythme des changements ces 50 dernières années est inédit dans l'histoire de l'humanité avec 25% des espèces animales et végétales évaluées qui sont déjà menacées (Fig. 1) et des projections qui indiquent une forte augmentation de cette proportion si aucune mesure n'est prise pour réduire l'impact des facteurs anthropiques (Díaz et al. 2019a).

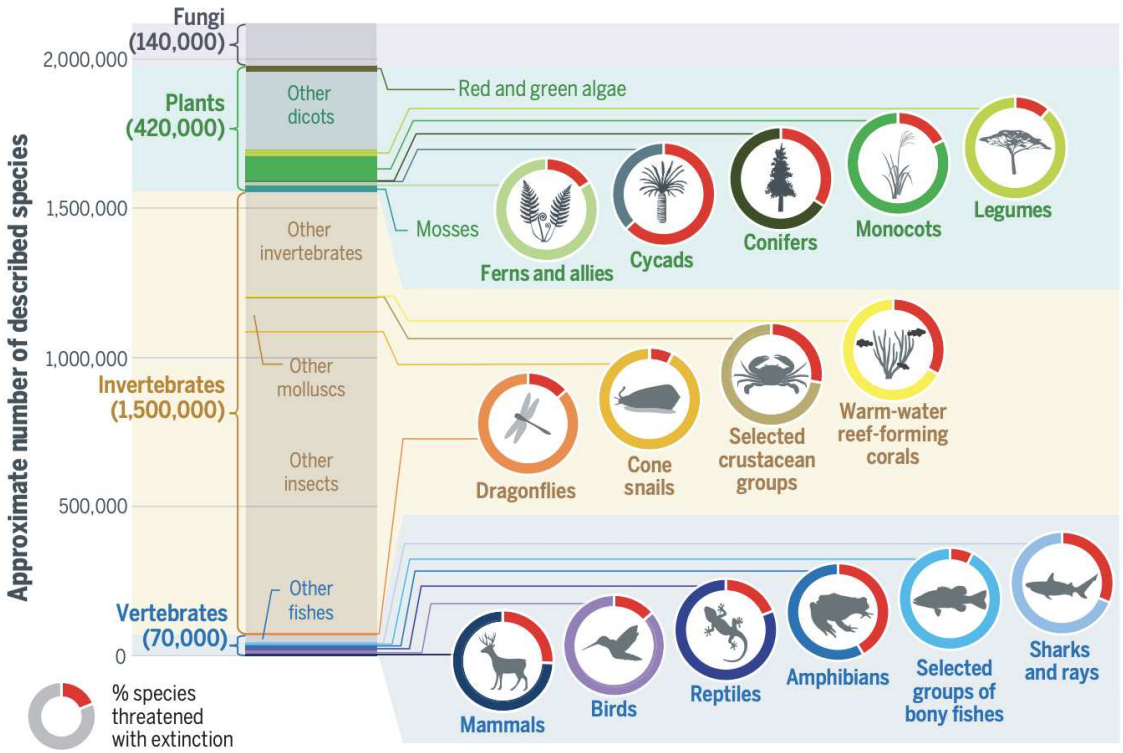


Fig. 1. Pourcentage d'espèces menacées d'extinction par groupe taxonomique. *Figure adaptée de Diaz et al. (2019a).*

Alors qu'aucun écosystème n'est épargné par ces impacts d'origine anthropique, tous ne sont pas sujets aux mêmes causes, aux mêmes intensités ou aux mêmes conséquences (Halpern et al. 2015). Dans les écosystèmes terrestres et d'eau douce, ce sont les changements d'utilisation des terres qui ont l'impact le plus important, notamment en détruisant ou en fragmentant les habitats naturels (>50% des zones humides ont disparu depuis 1900)(Davidson 2014), puis l'exploitation directe des organismes, à travers la collecte, la chasse ou la pêche. Dans les écosystèmes marins les facteurs sont inversés avec l'exploitation directe, donc la pêche, qui a le plus fort impact sur le milieu (Jackson et al. 2001). Le changement climatique est actuellement la troisième pression par ordre d'importance dans l'érosion de la biodiversité. Toutefois, les projections indiquent que dans les prochaines décennies ce sera le facteur le plus influant sur une biodiversité déjà extrêmement impactée par les activités humaines (Hughes et al. 2017, Thiault et al. 2019, Duarte et al. 2020). Il est relativement complexe d'estimer avec précision les conséquences prochaines des changements climatiques sur la biodiversité, notamment en raison de possibles boucles de rétroaction amplifiant les changements de température prévus par les modèles, tels que la fonte du permafrost ou encore le relargage de méthane depuis les zones humides (Steffen et al. 2018) et d'effets synergiques entre celui-ci et les autres impacts cumulés sur un même écosystème (Darling et al. 2010, Halpern et al. 2015, Holsman et al. 2020). La quantité de gaz à effet de

serre actuellement présente dans l'atmosphère (409.8 ppm en moyenne en 2019) suggère toutefois qu'une augmentation globale de 1.5°C est déjà un minimum d'ici les prochaines années, et que seuls d'énormes efforts pourraient permettre de ne pas dépasser les 2°C et de limiter les conséquences sur la biosphère (Leclère et al. 2020).

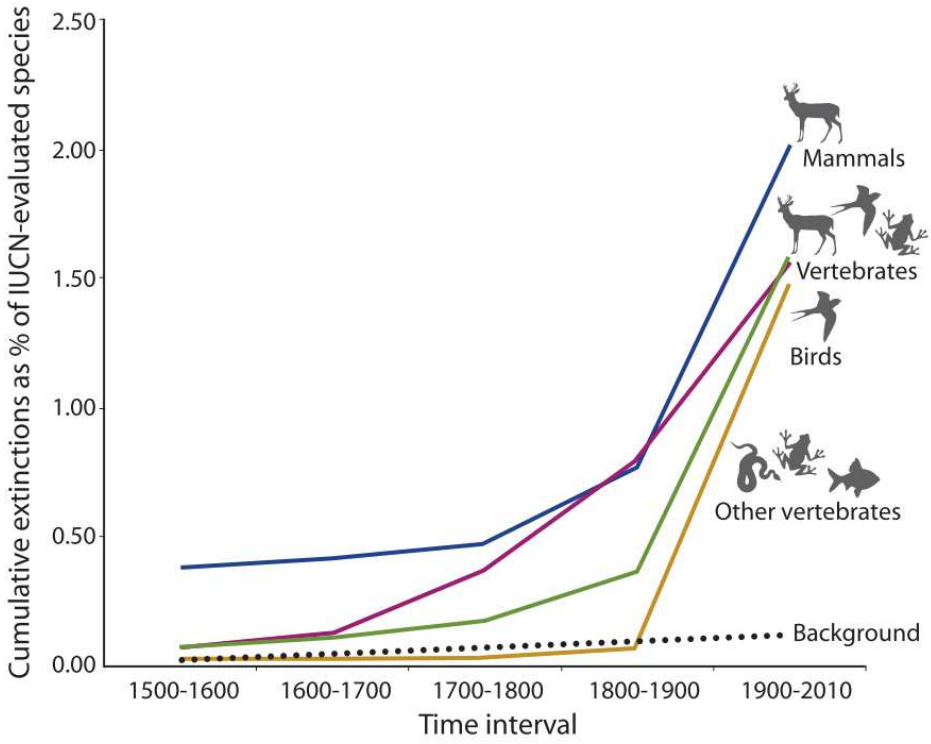


Fig. 2. Nombre cumulé d'espèces recensées comme éteintes dans la nature par l'IUCN en 2012 par groupe taxonomique. La ligne en pointillés représente le nombre d'extinctions attendu selon un taux naturel de 2 extinctions pour 10 000 espèces en 100 années (2E/MSY). *Figure adaptée de Ceballos et al. (2015).*

Les taux d'extinction d'espèces sont actuellement estimés entre 10 à 1000 fois plus élevés que lors de l'époque préindustrielle (Pimm et al. 2014, Ceballos et al. 2015, 2020), ce qui suggère que la biosphère pourrait être à l'aube d'une sixième extinction de masse (Barnosky et al. 2011)(Fig. 2), avec des espèces déjà déclarées éteintes localement due aux changements climatiques actuels (Wiens 2016). Toutefois, les actions de conservation contemporaines mises en place suite au placement des espèces sur liste rouge ont certainement permis d'éviter l'extinction de nombreuses espèces de vertébrés (Hoffmann et al. 2010, Bolam et al. 2020). Parmi les facteurs de risque d'extinction, la taille et la masse des organismes semblent jouer un rôle, avec les plus petites et plus grandes espèces plus susceptibles d'être menacées ou éteintes que les espèces de taille intermédiaire (Ripple et al. 2017). Typiquement, après contact avec des populations humaines, les espèces les plus grandes sont la cible de la chasse ou de la pêche, et subissent une forte diminution de leur abondance et risquent l'extinction (Jackson et al.

2001, He et al. 2018, Ripple et al. 2019). D'autres traits d'histoire de vie sont également corrélés aux risques d'extinction, tels que l'âge à maturité et la capacité de reproduction ou le régime alimentaire (Hutchings et al. 2012, Atwood et al. 2020). Les espèces ayant une durée de vie longue et se reproduisant tardivement sont naturellement plus impactées par des activités extractives, car les populations n'ont pas le temps de se régénérer (Reynolds et al. 2005). Le déclin des populations sauvages d'une espèce est un signe avant-coureur du risque d'extinction, où l'IUCN place les espèces ayant moins de 1000 individus restants dans la plus haute catégorie de classification (CR ; danger critique d'extinction)(Ceballos et al. 2020). Au moins 515 espèces de vertébrés terrestres (1.7% du total) sont considérées comme en danger critique d'extinction (Fig. 3). Beaucoup d'espèces ne sont pas évaluées par l'IUCN ou assignées en DD (*Data Deficient*) soulignant qu'il y a un manque de données sur leur répartition géographique et abondance pour correctement évaluer leur statut de conservation (Bland et al. 2015).

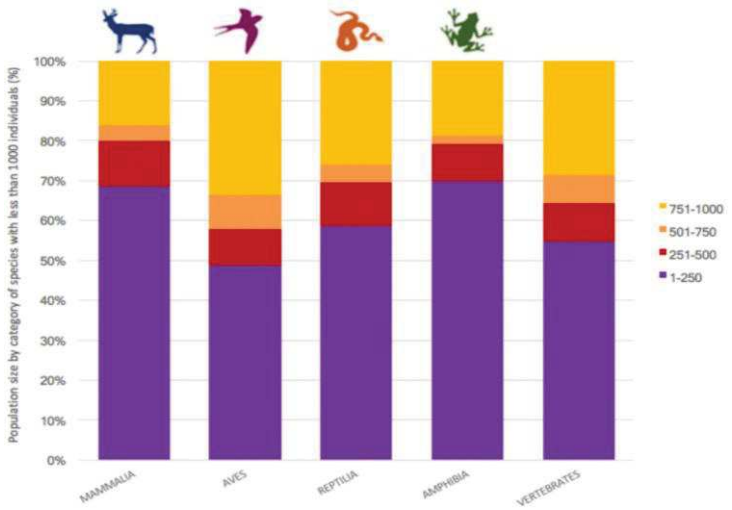


Fig. 3. Répartition de la taille des populations de vertébrés terrestres en danger critique d'extinction (moins de 1000 individus restant dans la nature). *Figure adaptée de Ceballos et al. (2020).*

La valeur économique accordée aux espèces peut également favoriser leur déclin, voire leur extinction. Cette valeur marchande peut même représenter le facteur principal de déclin d'une espèce, indépendamment de ses caractéristiques biologiques au-delà d'une valeur seuil (> 12 557$.kg^{-1}) (Purcell et al. 2014, McClenachan et al. 2016). Alors que le déclin en abondance des espèces pourrait décourager leur exploitation qui ne serait plus rentable et ainsi éviter l'extinction, la rareté peut en réalité provoquer une augmentation de la valeur marchande des espèces, et au contraire encourager l'exploitation d'espèces déjà fortement en déclin (Courchamp et al. 2006) (Fig. 4. A, B). Cet effet a été baptisé l'effet Allee anthropogénique. De nombreux exemples illustrent ce phénomène, notamment l'exploitation de l'abalone blanc (*Haliotis sorenseni*) dont la population a décliné de 99.99% suite à la surpêche et dont le prix a augmenté de façon inversement proportionnelle au déclin de l'espèce (Fig. 4. C).

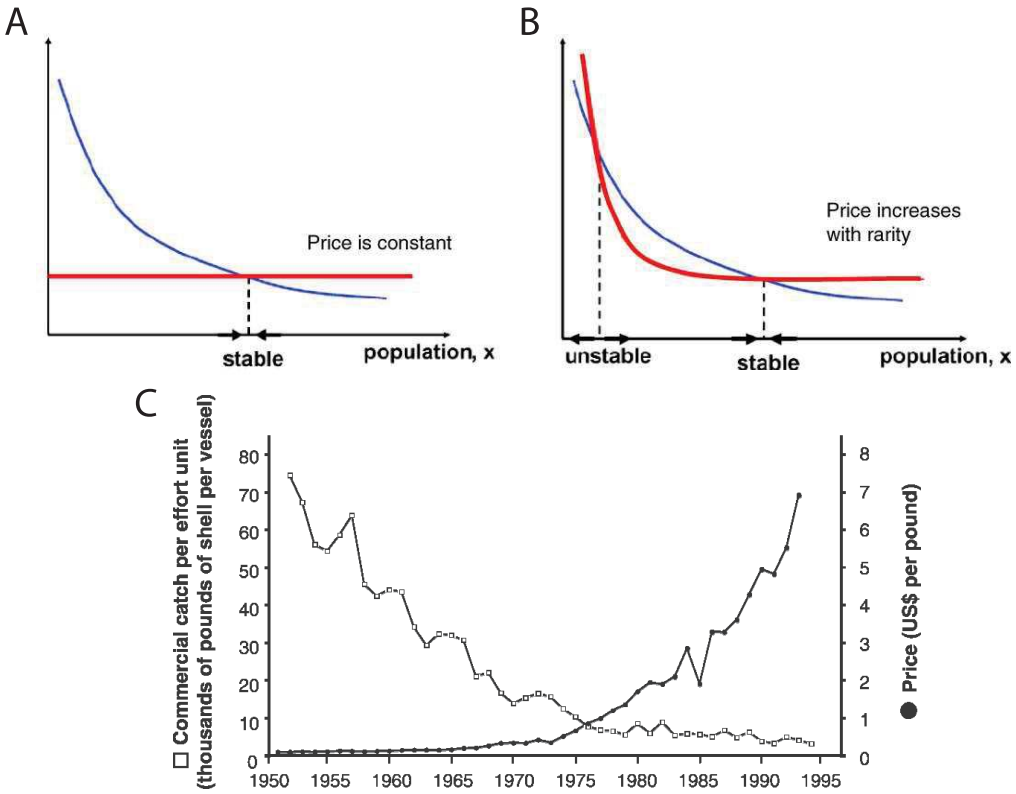


Fig. 4. Illustration de l'effet Allee anthropogénique avec un modèle d'espèce exploitée. Le prix (ligne rouge) et le coût par unité (ligne bleue) en fonction de la taille de population quand (A) le prix est constant et indépendant de la population et (B) le prix augmente quand la taille de la population diminue avec (C) un exemple du phénomène avec l'exploitation commerciale d'un gastéropode (*Haliotis sorenseni*), où les captures et les populations diminuent, mais le prix augmente exponentiellement entre 1972 et 1992. *Figure adaptée de Courchamp et al. (2006).*

1.2. Changements globaux et remplacement des espèces

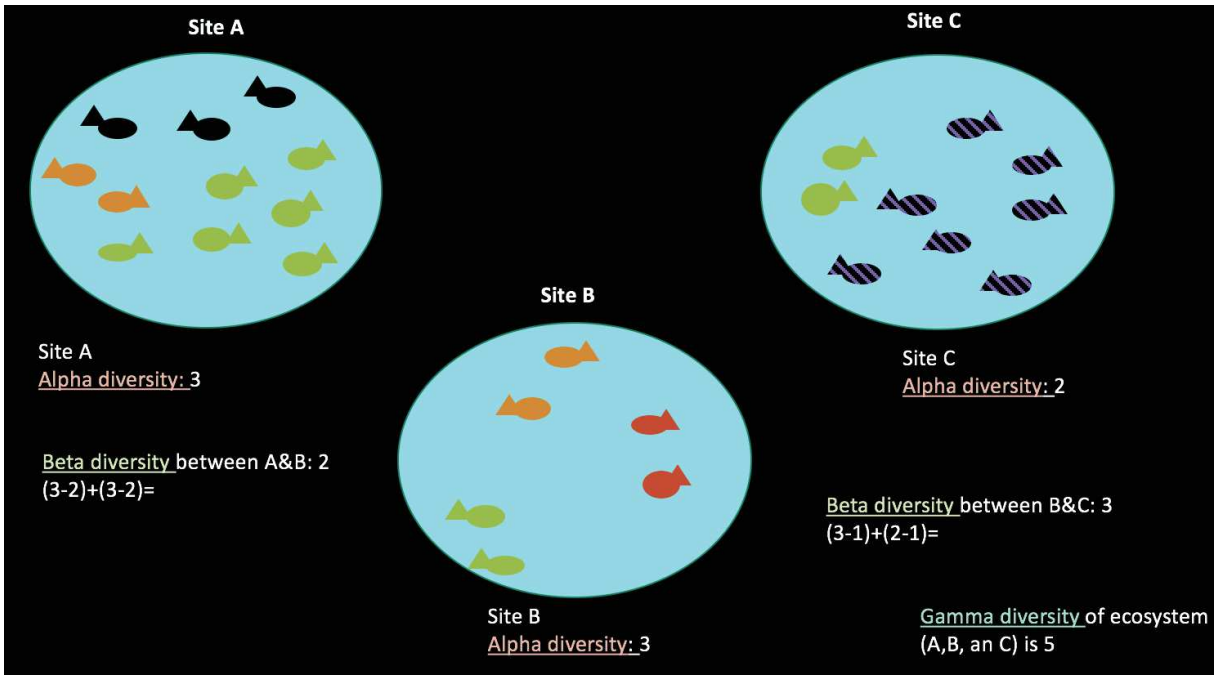


Fig. 5. Schéma de la partition spatiale de la diversité aux échelles alpha, beta additive selon Baselga (2010) et gamma. *Figurée adaptée de Socratic.org, crédit : Kate M.*

Toute mesure de biodiversité est à interpréter en termes d'échelles, notamment spatiales (Legendre and Fortin 1989). En écologie, la diversité est communément partitionnée spatialement en 3 composantes : la diversité **alpha** (échelle locale), la diversité **gamma** (échelle régionale) et la diversité **beta**, représentant la variation de composition entre des communautés (Whittaker 1972) (Fig. 5). Plusieurs indices de diversité existent à chacune de ces échelles, et peuvent prendre en compte (ou non) les abondances et propriétés des espèces au-delà de leur taxonomie. Ainsi, les changements de biodiversité ne sont pas identiques à chaque échelle de perception. Alors que la perte en biodiversité globale ainsi que les changements de communautés biotiques sont clairement documentés (Ceballos et al. 2020), les variations de richesse à l'échelle locale démontrent des résultats plus contrastés, avec des zones présentant des gains et d'autres des pertes de richesse (Dornelas et al. 2014, McGill et al. 2015)(Fig. 6). Les pertes d'espèces à l'échelle locale peuvent être compensées par l'arrivée de nouvelles espèces, créant ainsi de la diversité beta sans changement de richesse localement.

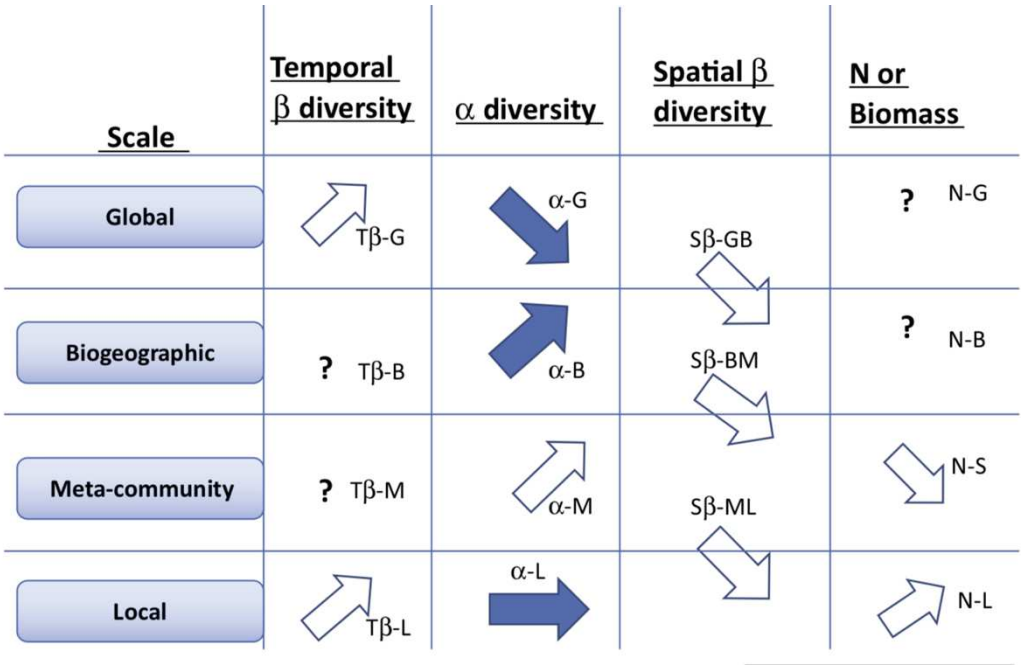


Fig. 6. Schéma identifiant les grandes tendances de la biodiversité aux 15 différentes échelles définies par McGill et al. 2015. Alors que certaines tendances sont bien documentées (flèches pleines, par exemple la diversité globale en diminution), d'autres manquent encore d'études empiriques pour valider les tendances (flèches vides, par exemple la beta-diversité spatiale à toutes les échelles) et d'autres encore sont trop peu étudiées pour émettre des hypothèses consensus (les points d'interrogation, par exemple la biomasse à l'échelle globale). *Figure adaptée de McGill et al. (2015).*

Une analyse de la plus large base de données de séries temporelles accessible, incluant 50 000 séries temporelles issues de 239 études rassemblées dans la base de données BIOTIME (Dornelas et al. 2018) révèle que les changements de biodiversité ne sont pas homogènes mais structurés spatialement (Blowes et al. 2019). Le nombre d'espèces à l'échelle globale diminue (diversité gamma) car les extinctions ne sont pas compensées par un nombre égal d'apparition d'espèces (évolution), mais à l'échelle locale (alpha), il n'y a en moyenne pas de signal homogène de perte ou de gain de richesse (Vellend et al. 2013, Dornelas et al. 2014, Yoccoz et al. 2018). Une décomposition de la diversité beta en ses composants de nestedness (« *emboîtement* ») et turnover (« *remplacement* ») (Baselga 2010) révèle que les assemblages ne deviennent pas de plus petits sous-échantillons d'une diversité régionale, mais qu'ils changent en composition car la composante de turnover (« *remplacement* ») est majoritaire. Les causes de ces remplacements peuvent être liées à l'introduction d'espèces non-indigènes (Kortz and Magurran 2019), l'arrivée d'espèces généralistes au détriment des espèces spécialistes plus sensibles aux perturbations (Nordberg and Schwarzkopf 2019), au changement climatique provoquant une migration des espèces (Feeley et al. 2013) ou bien encore aux modifications anthropiques des habitats ou l'exploitation des populations (Frank et al. 2018). Ces remplacements sont susceptibles d'entraîner une homogénéisation biotique entre les régions à travers une diminution globale de la beta-diversité

(Magurran et al. 2015, Kortz and Magurran 2019), comme cela a été montré sur une des communautés de plantes au Danemark au cours de 140 années, malgré une augmentation de la richesse liée à l'introduction d'espèces non indigènes (Nielsen et al. 2019).

Les changements climatiques forcent les espèces à migrer au-delà de leur aire de répartition originale (Vergés et al. 2014, Pecl et al. 2017, Lenoir et al. 2020). Cette migration s'effectue généralement vers des latitudes ou altitudes plus hautes lorsqu'elles ne sont pas adaptées pour tolérer ces changements de conditions abiotiques. L'ampleur des changements prédits est tel que la majorité des espèces devraient être amenées à migrer en réponse aux changements de conditions abiotiques (Lawler et al. 2013, Antão et al. 2020), malgré la mise en évidence de processus d'adaptation génétique aux contraintes environnementales (Razgour et al. 2019). Ces espèces migrantes étendent leur aire de répartition, ce qui questionne leur statut en tant qu'espèce non-indigène. Ce sont plutôt des espèces réfugiées climatiques ayant besoin de mesures de protection ou bien d'espèces néo-natives (Scheffers and Pecl 2019). Les migrations d'espèces sont susceptibles d'avoir des conséquences importantes et de provoquer une réorganisation fonctionnelle des écosystèmes concernant le cycle de l'eau ou les réseaux trophiques (Pecl et al. 2017, Nagelkerken et al. 2020). Le bouleversement de la répartition des espèces sur Terre s'accélère et reste extrêmement difficile à suivre et prédire à large échelle compte tenu notamment de l'ampleur géographique, de la forte diversité en espèces présentes et de la vitesse des changements.

1.3. Érosion des diversités phylogénétiques et fonctionnelles

L'érosion de la biodiversité est traditionnellement considérée dans sa dimension taxonomique, c'est à dire en termes de richesse et composition en espèces. Or il est de plus en plus reconnu que la seule identité des espèces n'est pas suffisante pour comprendre les changements et conséquences des restructurations biologiques, car toutes les espèces ne sont pas équivalentes et deux extinctions n'auraient pas les mêmes conséquences dans un écosystème selon les traits écologiques et l'histoire évolutive des deux espèces éteintes (Pollock et al. 2020). Ces deux facettes sont couramment proposées pour compléter la **diversité taxonomique** : il s'agit de la **diversité fonctionnelle**, qui considère le rôle des espèces à travers les fonctions effectuées dans leur environnent, et la **diversité phylogénétique**, qui considère la divergence évolutive entre espèces d'un assemblage.

La mesure de diversité fonctionnelle repose sur la valeur des traits fonctionnels d'espèces qui composent une communauté (Petchey and Gaston 2006). Ces traits peuvent être phénotypiques,

comportementaux ou physiologiques (Violle et al. 2007). Ces combinaisons de traits sont directement liées aux fonctions exercées par les organismes, par exemple la taille et le régime alimentaire conditionnent les flux trophiques entre espèces. Des mesures de traits fonctionnels permettent également de définir la niche écologique des espèces (Dehling et al. 2016), où un espace en 4 dimensions construit avec une combinaison de 9 traits morphologiques est suffisant pour extraire la niche trophique de la quasi-totalité des espèces d'oiseaux (Pigot et al. 2020). Les espèces peuvent ainsi être classées en groupes fonctionnels : un ensemble d'espèces partageant les mêmes valeurs de traits ou accomplissant un rôle considéré comme similaire dans l'environnement, par exemple les espèces de poissons récifaux herbivores de taille moyenne se nourrissant de gazon d'algues (« *turf* ») (Bellwood et al. 2019). Différents indices permettent de quantifier la diversité fonctionnelle, depuis la richesse en groupes fonctionnels (Mouillot et al. 2014) au placement dans un espace multidimensionnel pour en dériver la richesse fonctionnelle (FRic) (Villéger et al. 2017) ou la rareté fonctionnelle (Grenié et al. 2018). La diversité phylogénétique quantifie l'étendue de l'histoire évolutive présente dans une communauté. Plus une communauté présentera de taxa éloignés sur l'arbre phylogénétique, plus sa diversité phylogénétique sera importante ; elle correspond généralement à la somme des longueurs de branches composant la communauté (Faith and Baker 2006, Tucker et al. 2016).

Alors que la diversité phylogénétique a souvent été considérée comme un substitut de la diversité fonctionnelle sous l'hypothèse que plus deux espèces sont proches phylogénétiquement, plus leurs traits fonctionnels seront semblables, de récents travaux remettent en cause cette hypothèse (Mazel et al. 2018) ou révèlent une asynchronie dans la réponse de ces deux facettes à une même perturbation (Devictor et al. 2010, Monnet et al. 2014). Les mesures de biodiversité sont donc à considérer à travers une approche multifacette, où chacune est complémentaire et nécessaire afin d'évaluer à la fois les réponses de la biodiversité face aux pressions mais également l'influence de la biodiversité sur le fonctionnement des écosystèmes et les services associés (Pollock et al. 2020, Trindade-Santos et al. 2020).

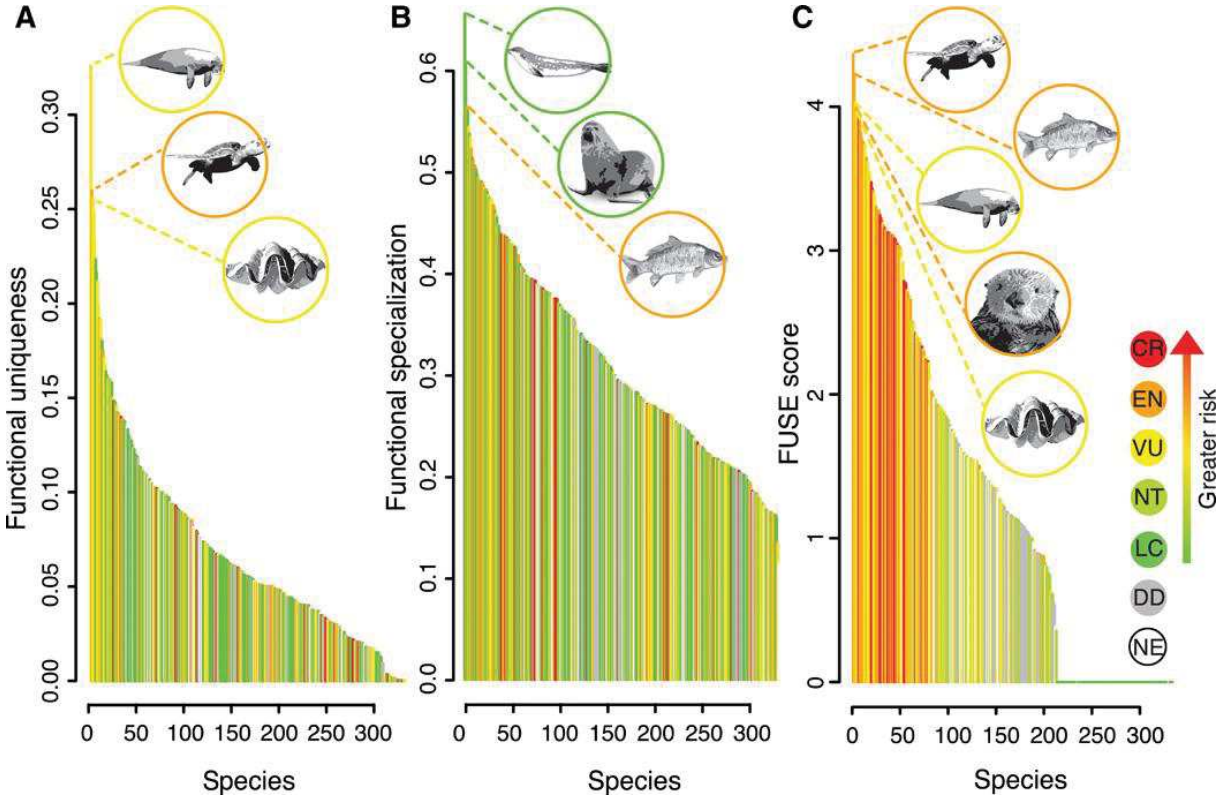


Fig. 7. Contribution des espèces à la diversité fonctionnelle et leur statut de conservation. *Figure adaptée de Pimiento et al. (2020).*

Cette approche multifacette permet de mieux définir les aires géographiques prioritaires en conservation afin de protéger non seulement la diversité des taxa, mais aussi la diversité des traits et des lignées phylogénétiques (Albouy et al. 2015, Pollock et al. 2017, Stein et al. 2018). En utilisant des scénarios de changements climatiques, il est aussi possible de modéliser les dynamiques des différentes facettes de la biodiversité (Monnet et al. 2014). En mer Méditerranée par exemple, sous un scénario d'extinction de 40 espèces de poissons marins, la perte de diversité fonctionnelle est projetée d'être limitée (3%) par rapport à perte de diversité phylogénétique (13%), grâce à la forte redondance fonctionnelle des espèces présentes dans ce bassin (Albouy et al. 2015). Ce résultat est également obtenu sur des communautés de plantes et vertébrés terrestres européens (Thuiller et al. 2011). Afin de prendre en compte la spécialisation fonctionnelle ainsi que le risque d'extinction, un nouvel indice a été mis au point : FUSE (*Functionally unique specialized and endangered*) afin d'identifier les espèces fonctionnellement importantes, non remplaçables et en même temps menacées d'extinction (Pimiento et al. 2020)(Fig. 7). L'application de cet indice à la mégafaune marine (>45 kg) révèle que si toutes les espèces menacées aujourd'hui disparaissent, 48% de la diversité fonctionnelle globale disparaîtrait avec elles (Pimiento et al. 2020). Étudier le remodelage des communautés passées face aux changements climatiques permet d'avoir un aperçu des évolutions à venir. Les analyses des variations spatiales de diversité phylogénétique lors de la dernière glaciation du Quaternaire révèlent ainsi une importante

homogénéisation phylogénétique à l'échelle de toute l'Europe, dont les conséquences sur les assemblages sont encore présentes après plusieurs millénaires (Saladin et al. 2020).

1.4. La vulnérabilité des poissons aux changements globaux

Les poissons osseux (Ostéichthyens, comprenant les classes Actinoptérygiens et Sarcoptérygiens, groupe non monophylétique) et cartilagineux (Elasmobranches, comprenant raies, requins et chimères) représentent près de la moitié des espèces de vertébrés avec environ ~32,000 espèces d'Actinoptérygiens et ~1,200 espèces d'Élasmobranches (Froese and Pauly 2000). Ces organismes provenant de lignées anciennes (530 millions d'années pour les Ostéichthyens et 450 millions d'années pour les Élasmobranches) ont colonisé la quasi-totalité des habitats aquatiques de la planète, y compris les abysses jusqu'à près de 8000m de profondeur pour les poissons osseux, et présentent une large variété de traits d'histoire de vie et de niches écologiques (Nelson et al. 2016, Gerringer et al. 2017). Les poissons osseux représentent environ 0.7 Gt (0.7 10^9 tonnes) de biomasse, et jouent un rôle clé dans le cycle du carbone (Wilson et al. 2009, Irigoien et al. 2014, Bar-On et al. 2018). Au-delà de leur importance écologique, ils remplissent également des rôles essentiels pour les sociétés humaines à travers les pêcheries mondiales, qui extraient plus de 100 millions de tonnes de poissons chaque année (Pauly and Zeller 2016). Ils assurent ainsi la sécurité alimentaire des populations de nombreux pays (Cisneros-Montemayor et al. 2016, Hicks et al. 2019).

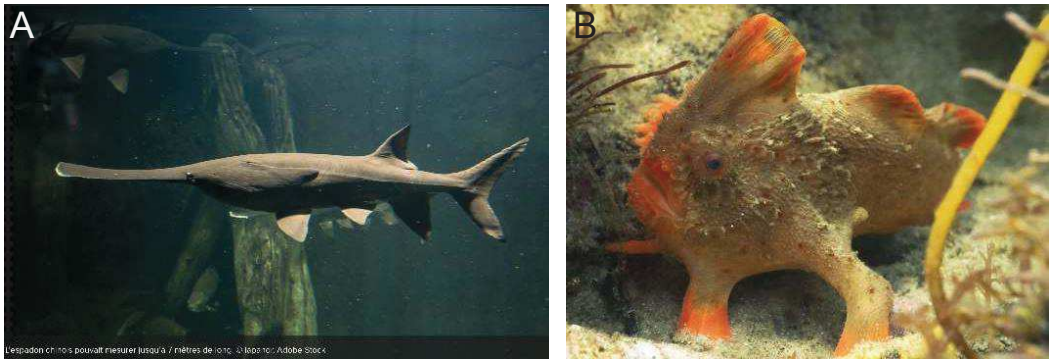


Fig. 8. Photos de *Psephurus gladius*, poisson d'eau douce déclaré éteint en 2020 (A) et de *Sympterichthys unipennis*, poisson marin déclaré éteint en 2018 (B). *Crédits : Lapandr, Adobe Stock (A) et Antonia Cooper (B).*

Les poissons d'eau douce représentent plus de 15 000 espèces vivantes sur l'équivalent de 1% de la surface terrestre (Collen et al. 2014), alors que le même nombre de poissons marins se répartissent

à travers tout l'Océan mondial (Froese and Pauly 2000). Ils sont aujourd'hui considérés comme faisant partie des vertébrés les plus menacés au monde, avec 12% des espèces menacées, contre 4% pour les poissons marins (Collen et al. 2014). De nombreux poissons d'eau douce sont déjà éteints, comme par exemple l'espadon chinois (*Psephurus gladius*) qui pouvait atteindre 3m de long (Fig. 8. A). Alors que 25 tonnes de ce poisson étaient pêchées annuellement dans les années 1970, la pression de pêche combinée à l'altération de son habitat par la construction d'un barrage hydraulique le mena à l'extinction en à peine 50 ans (Zhang et al. 2020a). Les extinctions contemporaines dans le domaine marin sont extrêmement rares (Dulvy and Polunin 2004), ce qui suggère que les espèces marines sont moins susceptibles de s'éteindre actuellement à l'échelle globale, ou bien que les extinctions ont déjà eu lieu mais n'ont pas été détectées.

L'unique extinction contemporaine d'un poisson marin a été prononcée en 2018, il s'agit du poisson-main lisse (*Sympterichthys unipennis*, Fig. 8. B), endémique de Tasmanie. Sa disparition est attribuée à une intense pêcherie de coquilles Saint-Jacques, dont la drague a détruit son unique habitat. 13 autres espèces de cette famille existent encore, mais parmi celles-ci 7 sont déjà menacées d'extinction. Si les poissons osseux marins sont relativement épargnés en termes d'extinction à l'échelle globale, ce n'est pas le cas des requins (« poissons » cartilagineux), pour lesquels 25% des espèces sont inscrites sur la liste rouge de l'IUCN. De nombreuses espèces ont perdu jusqu'à 99% de leur effectif en un siècle (Dulvy et al. 2014, Roff et al. 2018). Les prises accessoires ou la pêche ciblée pour la chaire et/ou surtout les ailerons sont les principales causes du déclin des requins, avec une mortalité estimée à plus de 100 millions d'individus tous les ans (Worm et al. 2013).

La quasi-absence d'extinction ainsi que le faible pourcentage de poissons osseux marins menacés pourraient faire croire que l'océan serait relativement épargné par les impacts anthropiques touchant la biodiversité. Mais ce serait ignorer que les écosystèmes terrestres sont fortement altérés depuis des centaines de milliers d'années, alors que l'altération des océans à large échelle a commencé plus tardivement, et se solde déjà par une diminution de biodiversité à tous les niveaux (Dulvy et al. 2003). Cette différence entre écosystèmes pourrait être la conséquence du décalage temporel du début de l'activité humaine à large échelle (pêche, tourisme,..) plutôt que d'une extrême résilience intrinsèque des écosystèmes marins (McCauley et al. 2015). Si aucune action urgente de conservation n'est mise en place pour diminuer et réguler les activités d'extraction des ressources, le déclin des espèces marines pourrait rattraper les taux d'extinction terrestres dans un futur proche (Fig. 9). De plus, certains taxa marins tels que les mammifères ou les requins sont déjà extrêmement menacés et ont subi plusieurs extinctions au cours des derniers siècles (Stein et al. 2018).

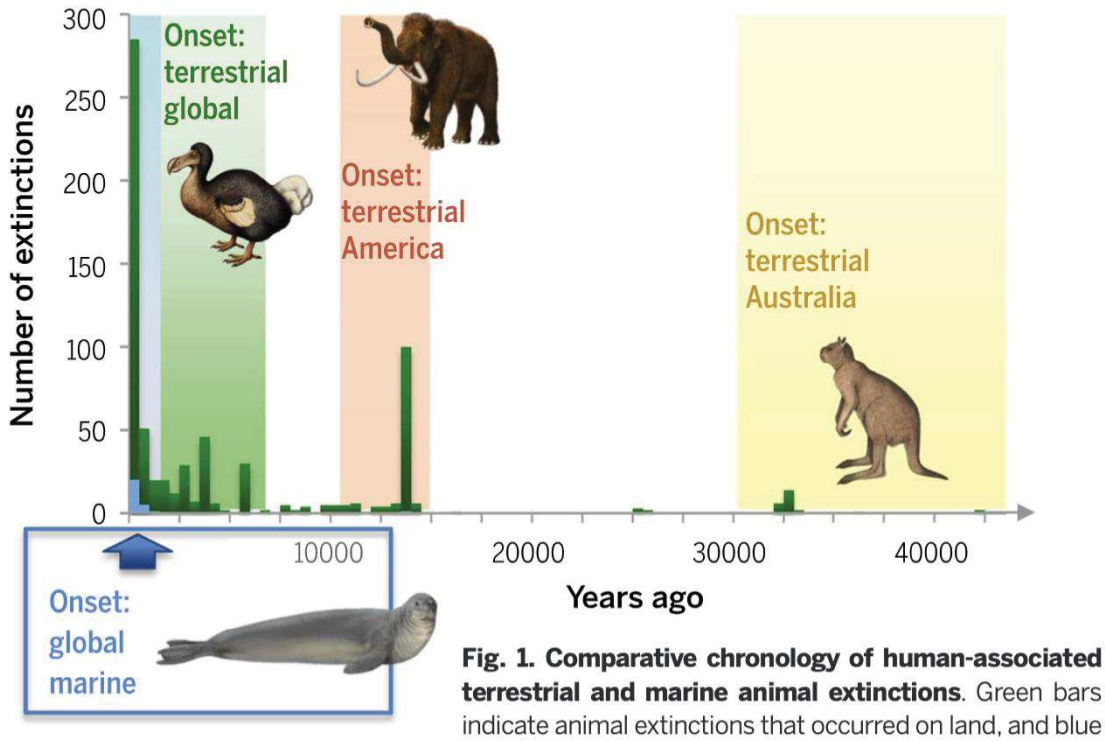


Fig. 9. Illustration du nombre d'extinctions dans chaque écosystème par rapport à la chronologie à partir de notre époque contemporaine. Les extinctions dans le milieu marin concernent majoritairement des mammifères marins. *Figure adaptée de McCauley et al (2015).*

Le nombre d'espèces menacées n'est pas l'unique métrique permettant de mesurer les impacts anthropiques. On estime ainsi que plus de 55% de la surface de l'océan est régulièrement pêchée de façon industrielle (Kroodsma et al. 2018), correspondant à une surface quatre fois plus importante que la totalité des terres cultivées, que ~33% des stocks de poissons marins sont surexploités et 60% exploités à pleine capacité (FAO 2018). Une espèce peut également échapper à l'extinction, mais présenter une abondance insuffisante localement pour remplir sa fonction dans l'écosystème, on parle alors d'extinction fonctionnelle (Säterberg et al. 2013). C'est le cas par exemple du poisson perroquet à bosse (*Bolmometopon muricatum*), autrefois abondante sur les récifs coralliens de l'Indo-Pacifique mais qui a désormais le statut IUCN « Vulnérable ». Cette espèce est le plus grand poisson perroquet au monde (1.5m, 75 kg), et par son activité de bio-érosion dans les récifs (5-6T corail consommé / an), influence la structure et le fonctionnement des récifs coralliens ainsi que le transport de sédiments (Bellwood et al. 2003, Donaldson and Dulvy 2004). Seuls quelques récifs éloignés des zones fréquentées par les humains abritent aujourd'hui des populations abondantes de Bolbometon (D'Agata et al. 2014). Aucune autre espèce ne présente la même combinaison de traits, ce qui signifie que sa fonction est perdue dès lors que l'espèce est éteinte ou fonctionnellement éteinte. De nombreuses espèces de requins sont également concernées par les extinctions fonctionnelles locales, avec des déclins

d'abondance mesurés jusqu'à 92% pour la famille des requins-marteaux sur la côte est australienne en 54 ans (Roff et al. 2018). Le plus large jeu de données regroupant les abondances de requins côtiers sur 3 années dans 58 pays a montré que 19% des 371 sites observés ne présentaient aucun requin (MacNeil et al. 2020), suggérant leur extinction fonctionnelle sur de nombreuses localités.

Les organismes aquatiques subissent également les conséquences des changements climatiques déjà enclenchés. Les espèces marines sont plus susceptibles de vivre à la limite de leur tolérance thermique, ce qui les rend plus vulnérables faces au réchauffement que les espèces terrestres (Sunday et al. 2012, Pinsky et al. 2019). Les changements de biodiversité sont ainsi plus importants dans le milieu marin que dans les milieux d'eau douce ou terrestres (Blowes et al. 2019). Ces observations sont corroborées par l'étude de la vitesse de déplacement des organismes pour suivre les isothermes car les organismes marins se déplacent en moyenne six fois plus rapidement que les organismes terrestres (Lenoir et al. 2020)(Fig. 10).

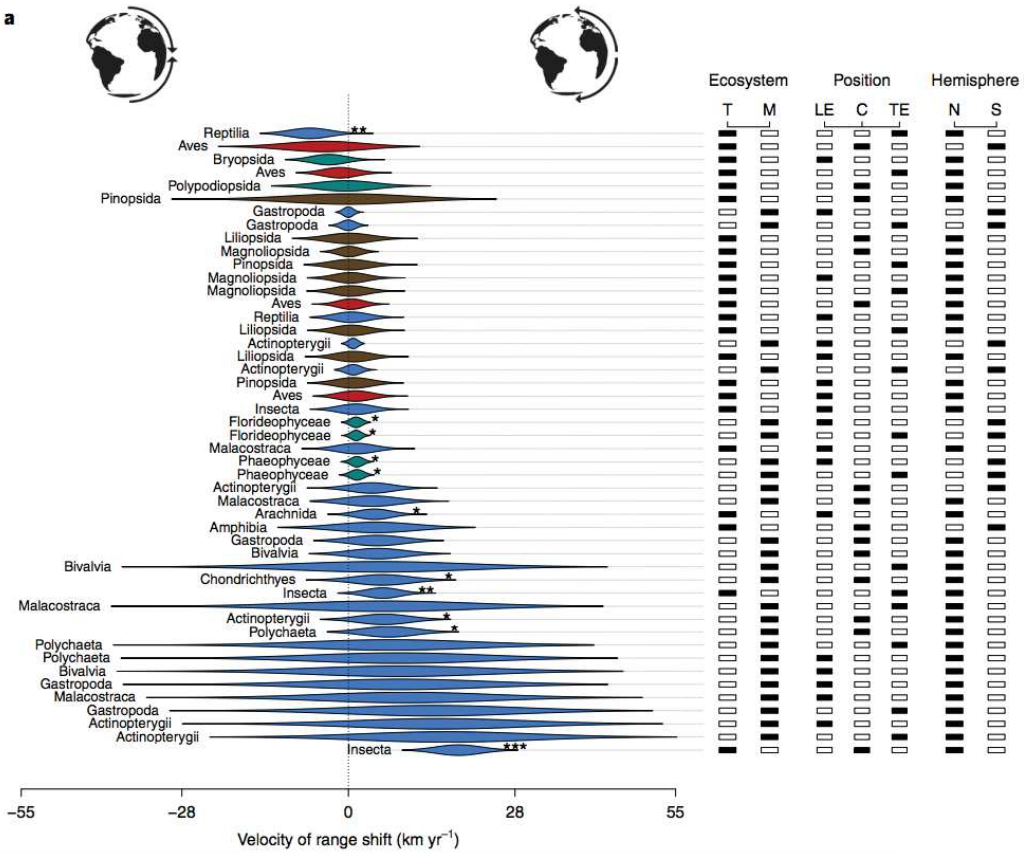


Fig. 10. Vitesses moyennes des déplacements latitudinaux d'espèces par classe taxonomique et par combinaison d'écosystème (T : terrestre, M : marin), de position géographique au sein de l'aire de répartition de chaque espèce, i.e. à la limite de l'aire ou bien au centre de celle-ci (LE : « leading edge », C : « centroid », TE : « trailing edge ») et d'hémisphère (N : nord, S : sud). *Figure adaptée de Lenoir et al. (2020).*

L'absence de barrières physiques à la dispersion dans les milieux marins est soulevée comme une possible explication de cette différence, car le milieu terrestre est globalement fragmenté du fait de l'occupation des sols par les humains (Haddad et al. 2015). Au sein des environnements marins, les changements les plus importants se situent dans les milieux polaires et tropicaux (Garciá Molinos et al. 2016, Antão et al. 2020). L'Arctique (polaire) se réchauffe deux fois plus vite que la moyenne terrestre (Hoegh-Guldberg and Bruno 2010), et cet environnement subit déjà l'arrivée de nombreuses espèces de plus basses latitudes, entraînant une « boréalisation » de l'Arctique (Fossheim et al. 2015). Le réchauffement des eaux polaires réduit l'aire de répartition des espèces qui y sont inféodées, et pourrait à terme menacer leur existence si le réchauffement dépasse leurs limites physiologiques (Dahlke et al. 2018). Les communautés de poissons marins tropicaux ont enregistré des changements de communautés en moyenne deux fois plus importants que leurs équivalents des biomes terrestres (Blowes et al. 2019). On observe déjà une « tropicalisation » de certaines zones, en Australie, au Japon ou en Méditerranée par exemple avec la présence de communautés de poissons tropicaux (Booth et al. 2011, Vergés et al. 2014). Alors que certains organismes typiquement assimilés aux eaux tropicales migrent vers des habitats plus tempérés, une baisse des captures de poissons dans toute la zone tropicale est anticipée avec le réchauffement des eaux, à hauteur de 40% dans certaines zones de l'Océan Pacifique (Cheung et al. 2010). La dépendance aux ressources marine étant plus élevée dans les tropiques qu'ailleurs dans le monde (Allison et al. 2009), ces projections sont alarmantes pour la sécurité alimentaire de nombreuses populations (Pinsky et al. 2018, Hicks et al. 2019).

2. Recensements des communautés ichtyologiques

2.1. Méthodes classiques et limitations

La compréhension des écosystèmes, la détection du déclin d'espèces et le suivi de leurs dynamiques temporelles nécessitent un recensement fiable et rapide, particulièrement dans cette période d'accélération des changements globaux. Les méthodes classiques de recensement des communautés ichtyologiques sont principalement basées sur une identification visuelle des organismes, ce qui nécessite une expertise taxonomique importante, et repose sur de longues heures de terrain pour collecter suffisamment de données. Ces méthodes sont diverses car spécifiques aux contraintes de chaque environnement et des organismes qui y vivent tant les combinaisons de traits d'histoire de vie chez les poissons sont nombreuses.

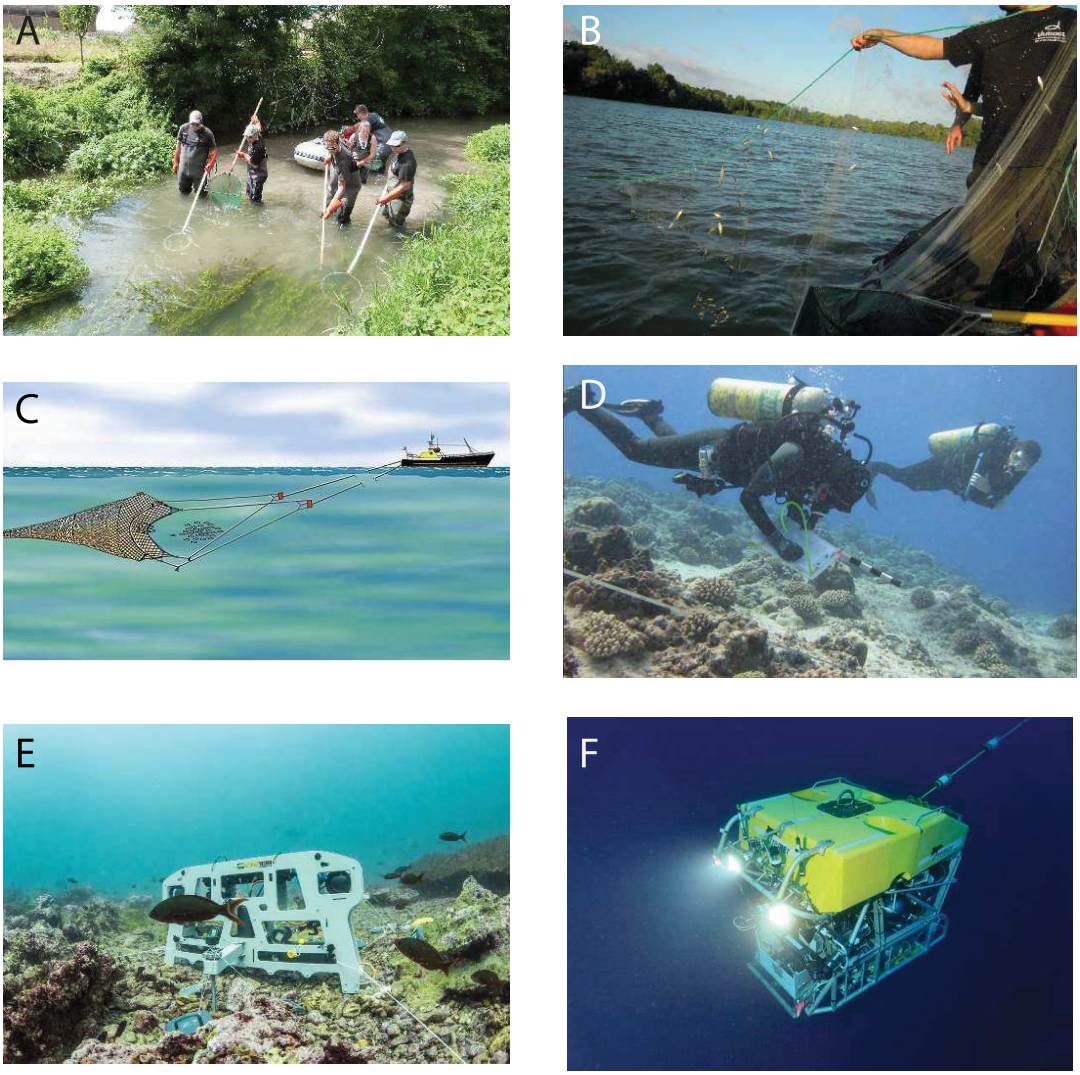


Fig. 11. Illustrations non exhaustives de méthodes d'échantillonnage des communautés ichtyologiques dans le milieu d'eau douce avec (A) la pêche électrique, (B) la pêche au filet et dans le milieu marin avec (C) la pêche au chalut, (D) les transects en plongée, (E) les vidéos sous-marines et (F) un véhicule sous-marin téléguidé (ROV). *Crédits : DUBOST environnement (A, B), Ifremer (C, F), NOAA Bernardo Vargas-Ángel (D) et O. Borde (E).*

En eau douce, la plupart des méthodes de recensement sont basées sur la capture des organismes pour leur identification. La pêche électrique est une méthode commune pour inventorier les communautés (Fig. 11). Un courant électrique est diffusé dans le cours ou plan d'eau afin de paralyser temporairement les organismes qui seront ainsi facilement capturables à la surface (Bain et al. 1985). Cette méthode est invasive car elle stresse et peut blesser les organismes, ce qui peut être inadapté pour des espèces menacées. De plus, la pêche électrique rate couramment les espèces les plus rares ou cachées dans des structures proches du benthos, car son champ d'action est très local (Reid and Haxton 2017). Il est alors nécessaire d'effectuer énormément d'inventaires afin d'avoir un échantillonnage représentatif des communautés présentes. Il n'est pas possible d'utiliser cette

méthode dans des cours d'eau trop profonds, ou lorsque la conductivité de l'eau est trop basse comme dans les cours d'eau de Guyane (Allard et al. 2014). Une alternative est alors l'utilisation de filets, qui ont l'avantage de récupérer plus d'individus, mais d'être encore plus invasifs et destructifs (Araújo et al. 2009). D'autres méthodes utilisent du poison comme la roténone, extrêmement efficace pour récupérer la totalité d'une communauté mais très destructeur (Allard et al. 2016). Il est parfois possible d'utiliser des vidéos sous-marines non invasives pour échantillonner les communautés d'eau douce, mais cette méthode est rarement applicable car la visibilité des eaux permet rarement d'avoir des images de bonne qualité (Nunes et al. 2020).

En mer, les méthodes d'inventaires dépendent du milieu échantillonné, notamment entre les écosystèmes côtiers peu profonds, les écosystèmes du large ou des profondeurs (Fig. 11). Sur la zone côtière, les inventaires se font généralement en plongée où une ligne de transect est déployée, et plusieurs plongeurs recensent et dénombrent les espèces rencontrées (Bosch et al. 2017). Les contraintes physiologiques humaines limitent le temps d'observation, la profondeur et donc l'étendue possible des inventaires, qui sont rares au-delà de 30m de profondeur (Pinheiro et al. 2016, Bongaerts et al. 2019). De fait, les écosystèmes mésophotiques (30-150m) sont parmi les moins bien échantillonnés et connus de l'océan (Rocha et al. 2018). L'exploration de ces zones plus profondes peut se faire en utilisant les avancées de la plongée dite « technique » en utilisant des mélanges gazeux et recycleur, des véhicules sous-marins téléguidés (« ROVs ») ou drones sous-marins, mais le niveau d'expertise, de complexité à déployer ainsi que le coût des méthodes ont historiquement limité leur utilisation. Une autre méthode couramment utilisée repose sur le déploiement de caméras, sans ou avec appâts à base de poissons gras pour attirer les prédateurs qui sont plus rares et difficiles à détecter (Schramm et al. 2020). Toutefois, les caméras restent généralement fixes sur la zone côtière, ne filment généralement pas au-delà de deux heures, et nécessitent du temps à posteriori pour identifier les organismes sur vidéos (Langlois et al. 2020). Les méthodes visuelles non-invasives sont également impactées par la visibilité qui diminue la qualité des recensements. Dans les zones estuariennes par exemple, il est impossible de mettre en place ce genre d'inventaire car les eaux sont rarement claires (Nunes et al. 2020). Dans le milieu pélagique, il est commun de pêcher au filet les organismes pour recenser les communautés (Trenkel et al. 2019). Ces méthodes destructrices ne sont pas adaptées aux espèces menacées. Une alternative est l'utilisation de caméras appâtées, cette fois dérivantes avec le courant, afin d'attirer la faune environnante devant les caméras (Letessier et al. 2019). Enfin les milieux profonds des océans (> 200m, au-delà de la zone crépusculaire) sont sous-échantillonnés due à leur grande difficulté d'accès, et aux coûts associés à leur exploration. Il est généralement possible d'inventorier ces communautés à l'aide de véhicules sous-marins téléguidés (« ROVs ») et pilotés depuis un navire océanographique, ou bien de sous-marins (Stefanoudis et al. 2019).

Ainsi, les problèmes associés à la plupart de ces méthodes sont notamment i) la nécessité d'avoir une bonne expertise taxonomique sur la région échantillonnée, ii) la durée souvent longue et/ou couteuse pour obtenir une bonne couverture spatiale et/ou temporelle des communautés iii) l'aspect invasif ou destructeur, nécessitant de stresser ou tuer des individus, iv) l'existence d'espèces cryptiques qui sont impossibles à différencier morphologiquement, v) la présence de biais de détectabilité en fonction de la méfiance des organismes face à l'humain. De fait, ces difficultés ont limité le nombre de recensements effectués globalement, avec un fort biais géographique car de nombreuses régions notamment tropicales et éloignées des centres de populations et des centres de recherche sont très peu, voire pas du tout étudiées (Boakes et al. 2010, Amano and Sutherland 2013, Reboredo Segovia et al. 2020). Il est cependant critique de mettre en place des inventaires efficaces, non destructifs, rapides et fiables avec un bon niveau de standardisation pour mieux comprendre les dynamiques des écosystèmes aquatiques et mettre en place des politiques de gestion adaptées à ces milieux.

2.2. Metabarcoding de l'ADN environnemental

L'étude de l'ADN environnemental, c'est à dire des molécules d'ADN extraites depuis l'environnement (Lacoursière-Roussel and Deiner 2019), est une technique d'inventaire moléculaire récente, appliquée pour la première fois en 2008 par Ficetola et al. (2008) sur les vertébrés à partir d'échantillons d'eau. Dans tous les milieux, et dans l'eau plus particulièrement, les organismes déposent continuellement des traces d'ADN sous formes de cellules par desquamation, par le mucus, les fèces, etc. Il est par la suite possible de récupérer cet ADN par filtration du milieu (eau, sol, sédiments), de l'extraire, puis de l'amplifier avec des amorces, puis de séquencer cet ADN au laboratoire (Fig. 12). Le choix des amorces sera conditionné par plusieurs facteurs, en particulier selon le choix de cibler une seule espèce, on parlera donc d'ADNe « *barcoding* » ou bien de cibler plusieurs espèces, on parlera alors d'ADNe « *metabarcoding* », qui est l'approche utilisée dans cette thèse. Les amorces de metabarcoding doivent être assez universelles pour être capable d'amplifier une large proportion des espèces d'un même groupe taxonomique cible, mais assez sélectives pour limiter l'amplification de groupes taxonomiques non cibles (Miya et al. 2015). Ces amorces sont situées sur des gènes mitochondriaux, car les mitochondries sont plus nombreuses que les noyaux au sein des cellules et sont ainsi plus susceptibles d'être récupérées en tant que fragment dans le milieu (Turner et al. 2014). Il est ensuite nécessaire de traiter les sorties de séquençage avec des pipelines bio-informatiques. Ces pipelines analysent les sorties brutes de séquençage, filtrent les données pour supprimer les séquences de mauvaise qualité, assignent chaque groupe de séquences à son échantillon (« *demultiplexage* »), puis assignent chaque séquence à une espèce ou taxon à l'aide d'une base de référence génétique. Si une séquence correspond à une espèce qui n'est pas présente dans la base de référence, il n'est toutefois

pas possible de lui assigner un nom d'espèce. On obtient par la suite une matrice de présence espèce/station, voire d'abondance en considérant le nombre de lectures de séquences (« *reads »)* mais son interprétation est encore sujette à débat (Doi et al. 2017).

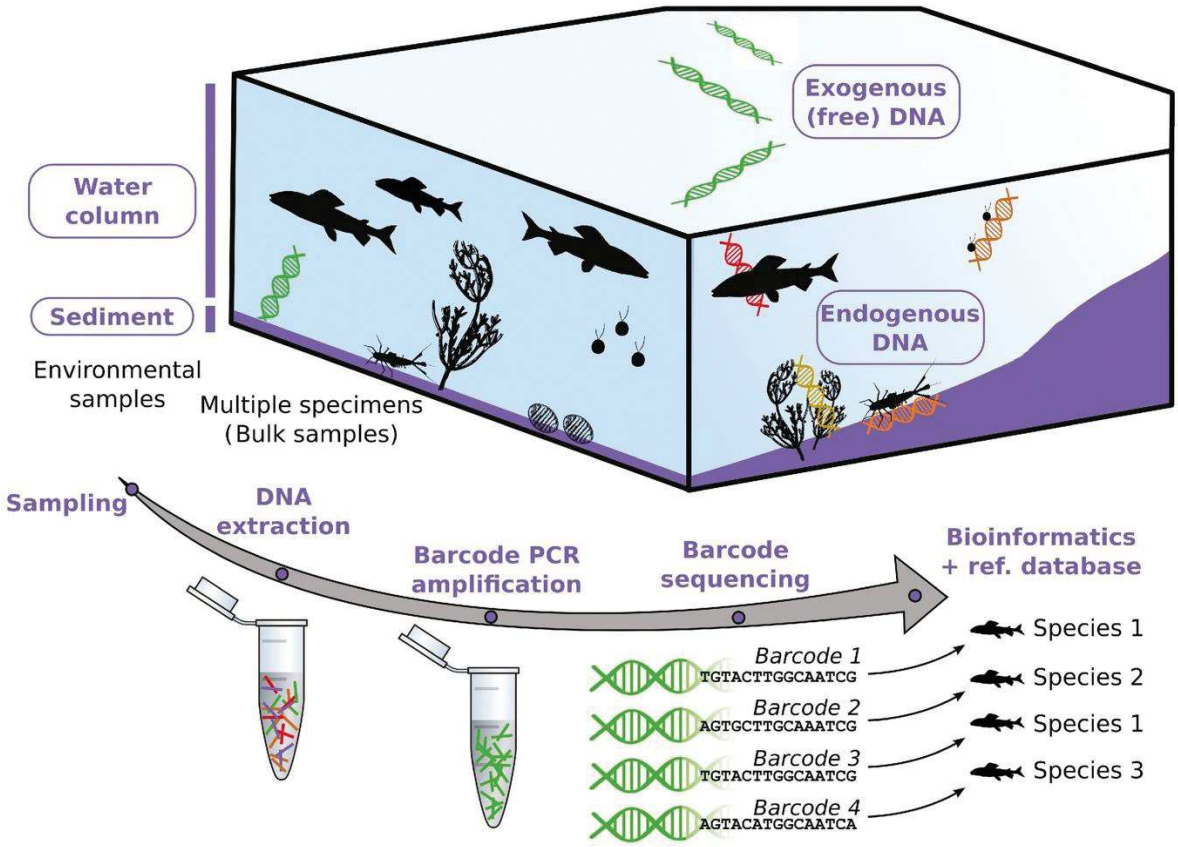


Fig. 12. Schéma de la procédure d'étude de l'ADNe par metabarcoding. *Figure modifiée depuis Keck et al. (2017).*

Parmi les premières applications de l'ADNe metabarcoding ciblant les poissons osseux dans le milieu naturel, Valentini et al. (2016) ont montré la faisabilité en milieu d'eau douce tempéré ainsi que la plus grande détectabilité de l'ADNe par rapport aux méthodes traditionnelles. Depuis, la méthode a été appliquée sur divers milieux, et a été comparée aux performances des méthodes classiques. De nombreuses études ont montré que l'ADNe avait des performances semblables ou supérieures aux méthodes traditionnelles telles que la pêche électrique ou au filet en milieu d'eau douce (Hinlo et al. 2017, Pont et al. 2018, McColl-Gausden et al. 2020) (Fig. 13). Dans un milieu tropical de rivière et à forte diversité en Amérique du Sud, Cilleros et al. (2019) ont montré que l'ADNe avait des performances proches par rapport au prélèvement par filet. Dans les rivières, l'information récupérée en ADNe recouvre en réalité le lieu d'échantillonnage, mais également les zones en amont de quelques kilomètres due au transport des molécules d'ADNe par le courant descendant (Deiner et al. 2016). L'utilisation de l'ADNe a aussi permis de révéler la présence d'espèces en danger critique d'extinction,

par exemple sur le fleuve du Rhône en France, où elles n'avaient jamais été détectées en 10 ans de pêche électrique (Pont et al. 2018). A travers une approche espèce-spécifique, Bellemain et al. (2016) ont permis la détection du poisson-chat géant du Mékong, malgré son statut IUCN en danger critique d'extinction et sa faible densité dans le milieu.

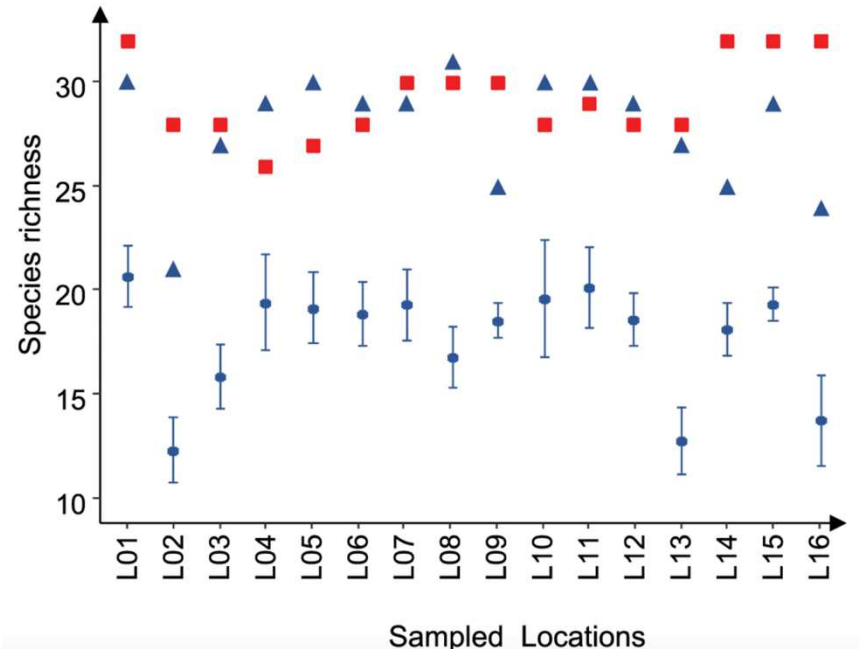


Fig. 13. Comparaison de la performance de 10 ans de campagnes de pêche électrique dans le Rhône (France) par rapport à une campagne d'ADNe metabarcoding. Les cercles bleus (+/- 95% intervalle de confiance) représentent le nombre annuel d'espèces échantillonnées avec la pêche, les triangles bleus le nombre total d'espèces échantillonnées en 10 ans (2006-2016) et les carrés rouges le nombre d'espèces détectées avec l'ADNe. *Figure adaptée de Pont et al. (2018).*

La première étude d'ADNe metabarcoding dans le milieu marin a permis de détecter 15 espèces de poissons à partir de quelques litres d'eau au Danemark (Thomsen et al. 2012). Plus tard, la même équipe a montré que l'ADNe avait les mêmes performances que des chalutages sur une zone profonde (~1 000m) au Groenland pour détecter les poissons présents (Thomsen et al. 2016). Récemment, il a été montré que malgré la possible dispersion grâce aux courants marins, l'ADNe permet de détecter une faune localisée par habitat à l'échelle d'un atoll tropical (West et al. 2020), ainsi que la détection d'un grand nombre d'espèces y compris de nouvelles détections locales étendant la liste d'espèces présentes localement. La méthode est également particulièrement adaptée à la détection des requins, qui a permis de mettre en évidence un fort gradient de présence des espèces en fonction des pressions anthropiques (Bakker et al. 2017) ainsi qu'une plus grande diversité d'espèces détectées avec 22 échantillons d'ADNe qu'avec 385 déploiements de caméras appâtées et 2758 plongées combinées

(Boussarie et al. 2018). En alternative à la filtration d'eau directement dans le milieu, il est également possible d'utiliser des tissus d'organismes filtreurs comme les éponges pour récupérer le signal ADNe de la faune de la zone à travers les larges quantités d'eau que ces organismes ont filtré (Mariani et al. 2019).

3. Enjeux

Les récents exemples d'application de l'ADNe suggèrent que la méthode est extrêmement prometteuse pour les milieux aquatiques, cependant certains verrous méthodologiques subsistent et limitent l'application plus généralisée de la méthode en complément ou remplacement des méthodes de recensement traditionnelles de la biodiversité.

L'exploitation du metabarcoding de l'ADNe pour les vertébrés nécessite d'identifier des espèces à partir des séquences ADN. En l'absence des références taxonomiques dans les bases de données génétiques, les séquences retrouvées sur les filtres ADNe ne sont pas identifiables et limitent considérablement le potentiel de la méthode. Or, les bases de données génétiques ne sont pas complètes pour la quasi-totalité de l'ichtyofaune, marine comme d'eau douce. Une étude multi-taxa à l'échelle de l'Europe a montré une forte hétérogénéité taxonomique et spatiale de la couverture des bases de références (Weigand et al. 2019), mais aucun bilan n'a été réalisé à plus large échelle. La complétude des bases de références pour identifier les espèces d'intérêt telles que les non-invasives ou menacées, pourtant souvent mises en avant dans le cadre des études ADNe (Rees et al. 2014, Larson et al. 2020), reste inconnue. On peut donc se demander quelle est la couverture taxonomique globale des bases de références pour les poissons osseux ? Quelle est l'hétérogénéité spatiale de cette couverture ? Quels marqueurs moléculaires offrent le plus de couverture ? Les bases de données sont-elles suffisamment complètes pour exploiter le potentiel de l'ADNe aux espèces non-indigènes et menacées ?

La couverture taxonomique incomplète des bases de références génétiques limite le potentiel du metabarcoding ADNe. Parvenir à séquencer la totalité des espèces sur la région de leur barcode ADNe est un objectif à long terme, mais n'est pas réaliste à court terme. Or, il est nécessaire d'exploiter dès aujourd'hui la totalité des informations contenues dans les filtres ADNe, même en l'absence de base de référence exhaustive. Les études en microbiologie, pionnières du séquençage haut-débit et du metabarcoding (de Vargas et al. 2015, Cordier et al. 2019), utilisent couramment des unités taxonomiques moléculaires (MOTUs) en tant que substituts (« proxys ») de la diversité réelle qui est

impossible à estimer autrement, compte tenu de leur grande diversité et de la pauvreté des bases de références génétiques mais également taxonomiques, car de nombreuses espèces procaryotes ou micro-eucaryotes ne sont pas décrites (de Vargas et al. 2015). Appliquer une telle approche aux vertébrés permettrait de bénéficier des avantages de la méthode ADNe sans dépendre d'une complétude totale des bases génétiques pour mieux comprendre les écosystèmes et potentiellement guider la mise en place de mesures de gestion de la biodiversité. Mais dans quelle mesure une approche par MOTUs peut-elle refléter la diversité spécifique des vertébrés réellement présente dans un écosystème ? Dans quels cas serait-il possible d'appliquer cette méthodologie par substitut moléculaire ?

Avant de généraliser l'utilisation des méthodes ADNe et de potentiellement les intégrer aux protocoles d'évaluation et de surveillance de la biodiversité, il est nécessaire d'effectuer des études comparatives avec les méthodes traditionnelles d'échantillonnage. Ces comparaisons sont notamment essentielles pour les milieux tropicaux qui présentent la plus forte biodiversité à l'échelle de la planète, avec de nombreuses espèces sous-étudiées ou inconnues, et qui concentrent une large proportion des enjeux de développements des sociétés humaines et de conservation de la biodiversité (Barlow et al. 2018). Nombre de ces écosystèmes sont éloignés des centres urbains et nécessitent un fort investissement sur le terrain pour y recenser la biodiversité. Il est donc important d'y généraliser l'approche ADNe avec des protocoles standardisés. Compte tenu de ces forts enjeux, quel est donc le potentiel de l'ADNe dans le recensement de la biodiversité des milieux tropicaux marins par rapport aux méthodes traditionnelles (plongées, vidéos) ? Au-delà de la richesse taxonomique, l'investigation des autres facettes de la diversité (fonctionnelle et phylogénétique) est fondamentale dans la caractérisation des écosystèmes. Dans les milieux tempérés d'eau douce, les études comparatives sont maintenant nombreuses et ont systématiquement révélé une performance égale mais complémentaire ou supérieure de l'ADNe par rapport aux techniques traditionnelles sur la facette taxonomique de la diversité (McElroy et al. 2020). Au-delà de diversité taxonomique, quelle est la capacité des recensements ADNe pour estimer la diversité fonctionnelle et phylogénétique des communautés ? Dans quelle mesure le metabarcoding de l'ADNe est-il complémentaire par rapport aux autres approches ?

Les pressions croissantes sur les environnements nécessitent d'être capable de suivre l'évolution de la biodiversité à large échelle spatiale et haute fréquence temporelle. Alors que les méthodes traditionnelles se heurtent à des difficultés en termes de logistique et de formation de personnel qualifié à l'identification morphologique des espèces, l'ADNe a le potentiel d'être à la hauteur de cet enjeu. La meilleure détectabilité des espèces par l'ADNe, notamment pour des groupes fonctionnellement importants mais traditionnellement ignorés comme les poissons crypto-benthiques

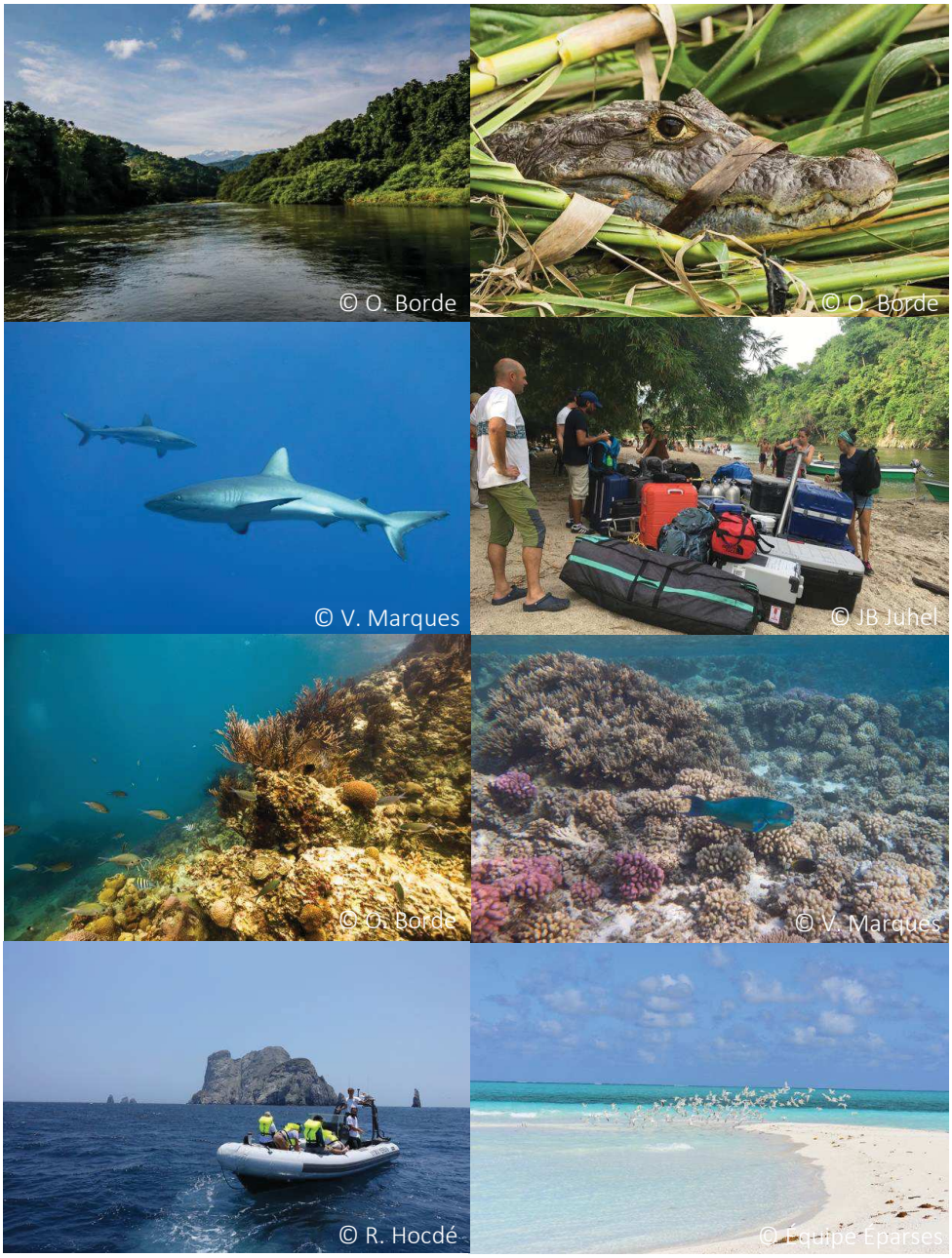
(Brandl et al. 2019), pourrait permettre d'affiner notre évaluation des assemblages de communautés hyper-diverses et sous-échantillonnées comme les récifs coralliens, avec des inventaires plus proches de l'exhaustivité. Pour cela, les études ADNe doivent encore démontrer leur efficacité pour étudier les assemblages d'espèces à l'échelle biogéographique. Alors que ces études ADNe sont limitées jusqu'ici à l'échelle régionale, est-il possible de dériver des règles d'assemblage biogéographiques à partir de séquences ADNe à large échelle sur des milieux riches en taxa ? Quelle est la capacité de détection de l'ADNe par rapport aux méthodes traditionnelles ? Que pourrait nous apprendre l'ADNe sur les règles d'assemblage des communautés grâce à une meilleure performance sur des groupes généralement négligés par les inventaires classiques ?

4. Objectifs

Cette thèse se concentre sur le verrou méthodologique posé par le manque de références génétiques, ainsi qu'à une meilleure compréhension de l'applicabilité de la méthode ADNe en milieu marin tropical pour lequel les études restent relativement rares. Ce travail est ainsi décliné en six chapitres, dont quatre ont fait l'objet de publications avec un total de 5 manuscrits en première ou co-première auteure. Parmi ces manuscrits, trois sont acceptés pour publication et deux sont soumis.

Dans un premier temps (chapitre 1), je présente la méthodologie de terrain et le cadre bio-informatique mis en place durant la thèse. Puis, le chapitre 2 propose une évaluation globale, spatiale et multi-marqueurs de la couverture taxonomique des bases de références génétiques pour les poissons d'eau douce et marins. Le chapitre 3 développe ensuite une alternative à l'assignement des séquences ADNe à des espèces en utilisant à la place des unités taxonomiques moléculaire (MOTUs), afin de s'affranchir de la nécessité d'avoir une base de référence complète pour quantifier la biodiversité détectée en ADNe metabarcoding. Dans le chapitre 4, deux articles utilisent l'approche développée au chapitre 3 pour explorer les performances de l'ADNe en milieu marin tropical par rapport à des méthodes classiques comme les vidéos sous-marines et les transects en plongée, dans une dimension multifacette en comparant diversité taxonomique, mais également fonctionnelle et phylogénétique des poissons. Le chapitre 5, présente la première étude à large échelle (3 océans) sur un écosystème donné (les récifs coralliens) de la biogéographie des fragments d'ADNe des poissons. Cette étude nous montre la capacité de l'ADNe à retrouver les grands patrons biogéographiques connus mais surtout à révéler de nouvelles règles d'assemblages des communautés de poissons récifaux. Enfin, le chapitre 6 dresse un bilan ainsi que les perspectives de ce travail.

Chapitre 1 – Échantillonnage et développements méthodologiques



1. Explorations de Monaco

Cette thèse s'inscrit dans le contexte des Explorations de Monaco initiées en 2017 dont une des ambitions soutenues par le programme est l'étude à l'échelle globale de la biodiversité des récifs coralliens en développant des outils de metabarcoding de l'ADN environnemental (https://www.monacoexplorations.org/). Ciblant plus spécifiquement les poissons osseux et cartilagineux (requins), ce projet est un partenariat entre l'entreprise à mission SPYGEN, spécialisée dans l'application des méthodes ADNe, l'université de Montpellier (MARBEC, CEFE), l'école polytechnique fédérale de Zurich (ETH), fédérés autour de l'application de l'ADNe à large échelle.

Les ambitions du projet sont de prélever des échantillons d'ADNe sur un large gradient spatial mais aussi d'anthropisation à travers le globe, incluant des sites éloignés des populations humaines, peu soumis aux pressions anthropiques et qui sont peu accessibles, mais également des sites proches des populations avec des impacts locaux tels que la pêche, la pollution ou la destruction des habitats. Un tel gradient permet d'avoir une idée de « l'état de référence » des écosystèmes marins avant leur forte dégradation par les activités humaines, et d'estimer quelles mesures de conservation peuvent être mises en place pour concilier développement des sociétés humaines avec la préservation des écosystèmes. Ce large échantillonnage standardisé a pour ambition d'avoir une couverture globale et ainsi cartographier la biodiversité en vertébrés, pour notamment assister la conservation des habitats marins. En parallèle des échantillons d'ADNe, des échantillons de tissus provenant d'un maximum d'espèces doivent être prélevés afin de constituer une banque génétique de la biodiversité des poissons marins, et ainsi permettre d'identifier les espèces dont les fragments d'ADN ont été récupérés sur les filtres.

2. Carte d'échantillonnage

Ce travail de thèse a nécessité de nombreuses missions de terrain, et les données récoltées par les différents membres du projet ont contribué à alimenter la base de données à large échelle en ADNe du projet. Toutes les données n'ont pas été utilisées dans le cadre de ces travaux de thèse, je présente donc ici uniquement celles qui ont fait l'objet d'une publication en lien avec ces travaux, et je détaille dans la partie suivante les missions auxquelles j'ai personnellement pris part.

La carte d'échantillonnage ADNe regroupe les données récoltées depuis Octobre 2017, jusqu'à Décembre 2019. Les échantillons sont classés selon :

(i) Leur **identifiant** de filtre unique

(ii) Leur **station**, c'est à dire les coordonnées d'un même transect

(iii) Leur **site**, la zone géographique associée regroupant plusieurs stations, par exemple une île, ou un ensemble d'écosystèmes similaires proches spatialement et écologiquement

(iv) Leur **région**, un ensemble biogéographique cohérent et regroupant potentiellement plusieurs sites

Elle comprend au total 435 échantillons (filtres), sur 303 stations, 33 sites et 6 régions (Fig. 1).

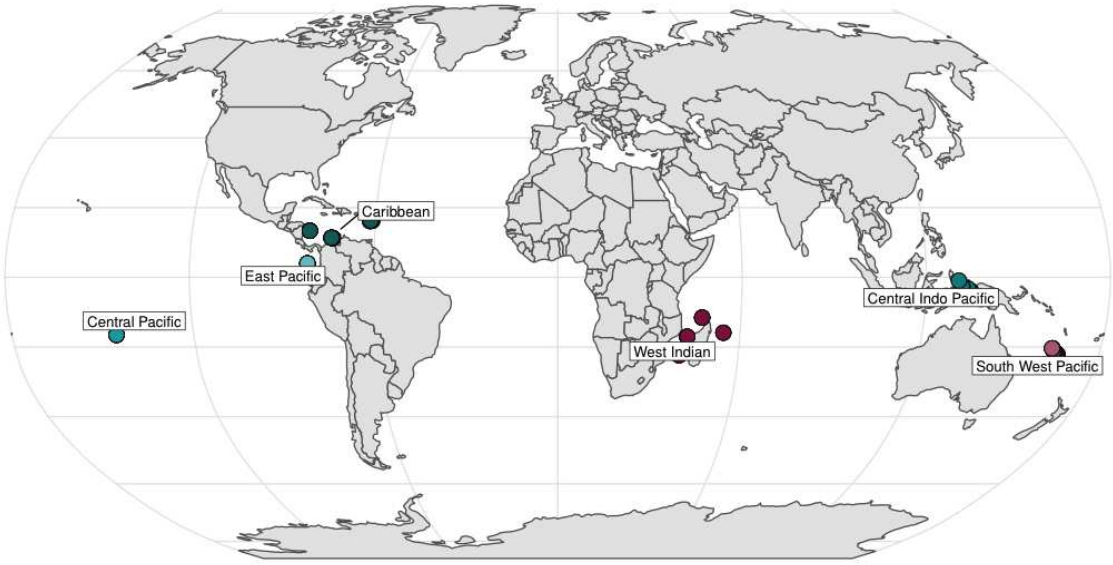


Fig. 1. Carte globale des échantillons ADNe récoltés à ce jour. Les points sont colorés selon leur appartenance à une région, et le nom de la région est indiqué autour des points.

3. Missions effectuées

Dans cette partie, je détaille les missions de terrain auxquelles j'ai participé, leur durée, le nombre de filtres ADNe collectés ainsi que le nombre de poissons pêchés pour la base de référence génétique ou projets annexes issus de collaborations.

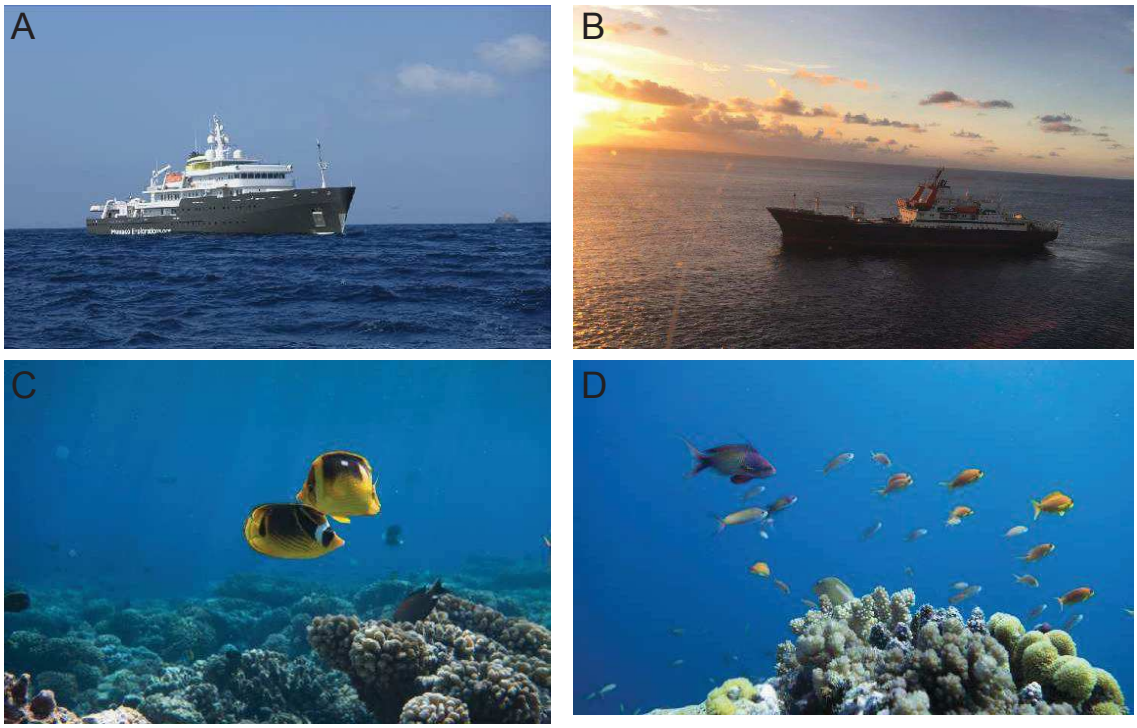


Fig. 2. Photos de terrain avec le navire Yersin au large de l'île de Malpelo (A), le navire Marion Dufresnes au large de l'île d'Europa (B), une image sous-marine sur l'atoll de Fakarava (C) et une image sous-marine sur l'île d'Europa (D). *Crédits : (A) R. Hocdé, (B) J-B. Juhel, (C,D) V. Marques.*

Au total, j'ai pris part à 6 missions de terrain (83 jours en mission dont 33 jours en embarquement, sur un période de 19 mois), dont 5 ayant permis la récolte d'échantillons ADNe, mélangeant des missions embarquées en mer avec des missions basées à terre et embarquements à la journée (Fig. 2). Les manipulations sur le terrain annexes à l'ADNe issues de projets partenaires sont mentionnées, qu'elles soient directement intégrées à ce travail ou non (Table. 1) (Fig. 3).

Table. 1. Tableau synthétique présentant les missions de terrain auxquelles j'ai personnellement pris part au cours de ce travail de thèse, le nom des campagnes, les dates, le nombre de jour, de filtres ADNe récoltés, le nombre de poissons et d'espèces différentes récupérées pour la base de référence et les projets annexes à l'ADNe.

Campagnes	Date début	Date fin	#jours	#filtres	#poissons (#espèces)	Projets annexes	Commentaires
Boa Vista (Cap Vert)	**23-09-2017**	**03-10-2017**	**13**	**0**	**-**	Déploiement caméras appâtées pélagiques (9 sites, 47 caméras, 94 heures)	7 jours en mer sur le Yersin. Aucun filtre ADNe (problème acheminement du matériel) Pas d'autorisation de pêche
Guadeloupe (France)	**25-02-2018**	**04-03-2018**	**8**	**28**	**-**	-	Mission à terre. Pas d'autorisation de pêche
Ile de Malpelo-LEG2 (Colombie)	**21-03-2018**	**02-04-2018**	**13**	**30**	**36 (22)**	Déploiement caméras longue durée non appâtées (1 site, 25 heures)	11 jours en mer sur le Yersin. 3 déploiements de caméras longue durée (25h) Participation au second leg de cette mission uniquement. Pas d'autorisation de pêche Poissons récupérés sur les marchés
Atoll de Fakarava (Polynésie Française)	**16-06-2018**	**30-06-2018**	**14**	**13**	**-**	-	Mission à terre Défaillance matérielle de la pompe Pas d'autorisation de pêche
Santa Marta (Colombie)	**13-10-2018**	**28-10-2018**	**15**	**32**	**181 (40)**	Déploiement caméras longue durée non appâtées	Mission à terre. 4 déploiements de caméras longue durée (48h)
Iles Éparses-LEG1 (France - TAAF)	**02-04-2019**	**21-04-2019**	**20**	**74**	**319 (53)**	**-**	15 jours en mer sur le navire Marion Dufresnes. Restriction sur le nombre d'espèces sur le permis de pêche. Premier leg uniquement.

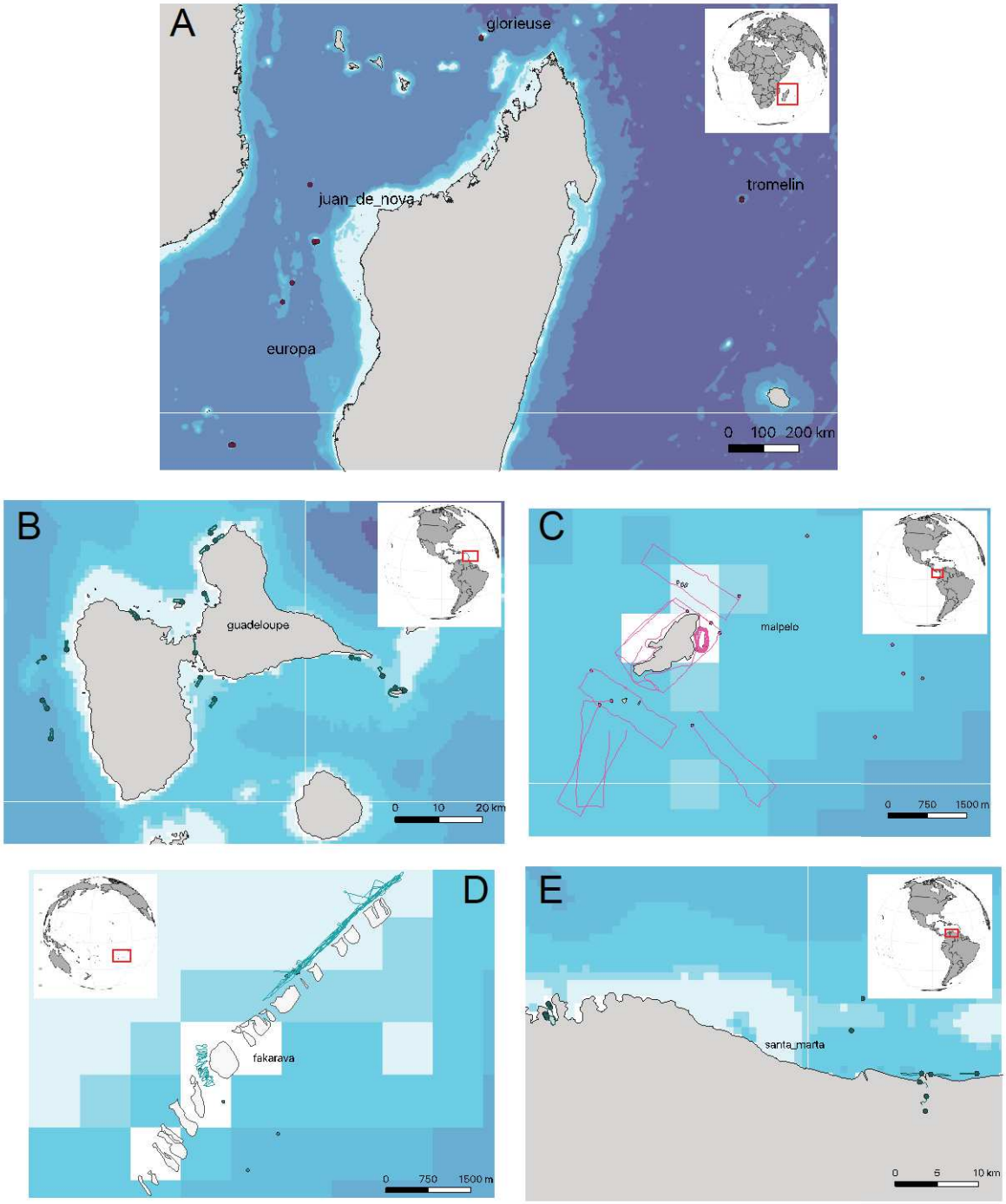


Fig. 3. Planche présentant la carte d'échantillonnage des missions auxquelles j'ai participé à l'échelle de la région (A) ou des sites (B-E).

4. Méthodes d'échantillonnage

Cette section présente les méthodes d'échantillonnage utilisées dans le cadre de cette thèse, et est divisée en deux parties. Dans un premier temps, je présente les méthodes de terrain concernant l'ADNe metabarcoding, la définition du protocole d'échantillonnage, les difficultés rencontrées ainsi que les évolutions méthodologiques mises en place. Dans un second temps, je présente les méthodes ayant permis d'amorcer la collecte de tissus d'espèces afin de remplir les bases de référence génétiques.

Protocole de terrain ADNe metabarcoding

Au début du projet en 2017, la littérature en ADNe appliquée aux vertébrés en milieu marin était encore très limitée (Tsuji et al. 2019), et peu d'informations étaient disponibles sur les méthodes d'échantillonnage dans l'océan. En revanche, l'ADNe était déjà mieux développé sur les milieux d'eau douce, nous sommes donc partis de cette base afin d'en adapter la méthodologie. L'ADN n'est pas distribué de façon homogène dans l'eau (Harrison et al. 2019, Laporte et al. 2020), et les spécificités du milieu conditionnent l'échantillonnage. Par exemple, dans une mare où la circulation de l'eau est très limitée, il est nécessaire de récolter de multiples échantillons d'eau sur une large surface pour correctement caractériser la communauté vivant dans le milieu (Furlan and Gleeson 2017, Harper et al. 2019). Pour cela, il est possible de faire des réplicas terrain à travers plusieurs prélèvements de petite quantité d'eau (~1L), ou bien de grouper tous les prélèvements sur un seul filtre, quand celui-ci est adapté pour limiter le colmatage grâce à une importante surface (Fig. 4). A l'inverse, en rivière l'eau circule permettant un brassage et un transport des molécule d'ADN, il n'est donc pas nécessaire de prélever des échantillons ou réplicas sur une large étendue spatiale. Cependant le signal ADN est toujours fragmentaire, nécessitant une filtration en continu ou de multiples sous-échantillons sur une zone proche.

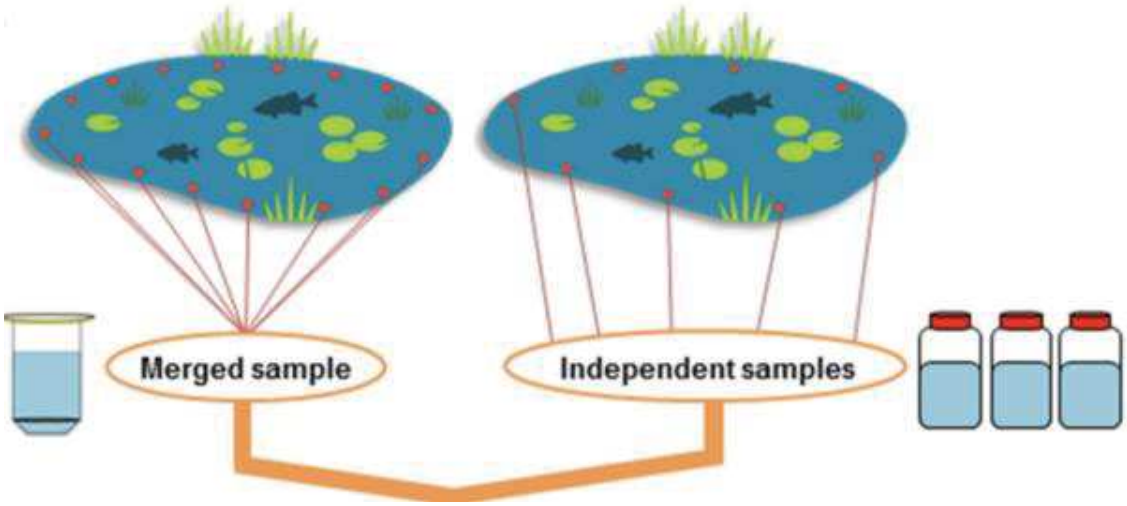


Fig. 4. Illustration de deux stratégies possibles d'échantillonnage dans une mare. *Figure modifiée depuis Harper et al. (2019).*

En milieu marin, on s'attend à ce que l'important volume lié aux dimensions en 3D de l'habitat ainsi que les courants diluent l'ADN de façon plus importante qu'en milieu d'eau douce, tout en ayant un risque réduit de colmatage due à l'oligotrophie des eaux. Il est donc nécessaire d'échantillonner une large surface, d'avoir des réplicas de terrain et de filtrer une quantité importante d'eau pour maximiser la force du signal ADN (Bessey et al. 2020). Pour cela, on a utilisé un filtre encapsulé à large surface similaire à certaines approches en rivière (Pont et al. 2018), mais avec une taille de pore réduite (0.20 µm au lieu de 0.45 µm). Afin d'être en mesure de filtrer directement dans le milieu tout en limitant la contamination, il est nécessaire que l'eau ne soit pas en contact avec les éléments internes de la pompe qui sont non nettoyables, ce qui justifie l'utilisation d'une pompe péristaltique. Ce type de pompe fonctionne grâce à une tête rotative, qui en appuyant sur le tuyau fait appel d'air et permet de faire circuler l'eau (Fig. 5. A). Il est ainsi aisé de changer de tuyau entre chaque prélèvement, de façon à ne pas contaminer les échantillons et de conserver la même pompe. Après filtration, cette capsule est remplie d'une solution tampon de conservation (CL1 SPYGEN, tampon de lyse : Tris-HCl 0.1 M, EDTA 0.1 M, NaCl 0.01 M and N-lauroyl sarcosine 1%, pH 7.5–8) permettant de conserver ces filtres à température ambiante pendant plusieurs mois (Fig. 5. B).

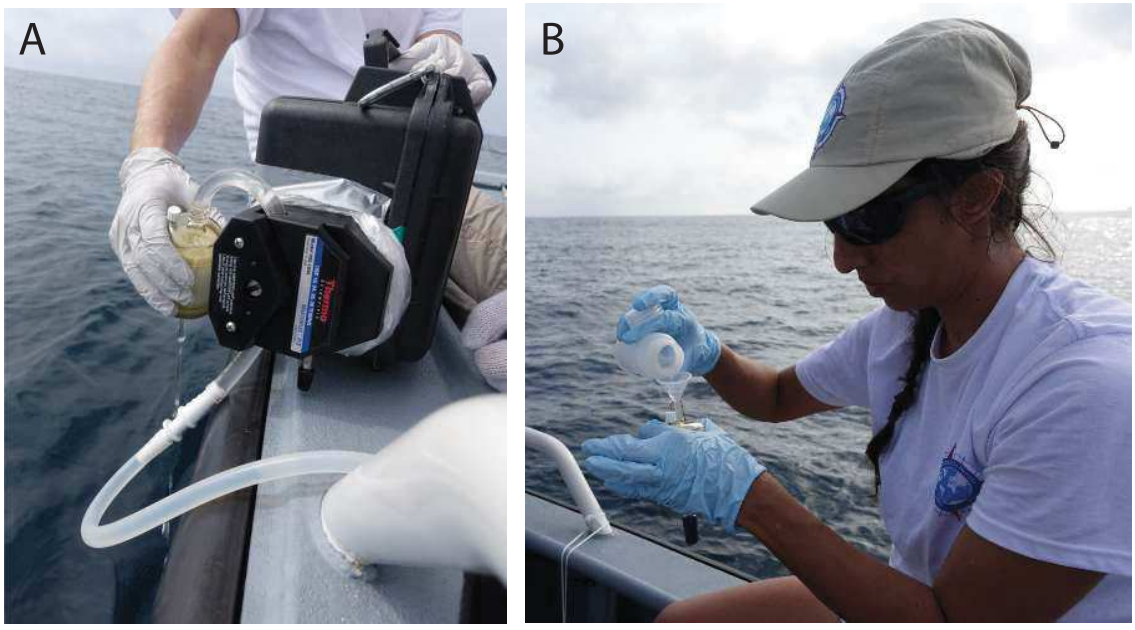


Fig. 5. Photos de terrain sur une annexe en mer, sur laquelle est montée une pompe péristaltique prélevant de l'eau en continu pendant le transect (A), une pompe est présente de chaque côté du bateau du façon à avoir deux réplicas de terrain et (B) Dr Andréa Polanco remplissant le filtre de solution tampon. *Crédits : R. Hocdé (A, B).*

Le protocole d'échantillonnage en surface nécessite l'accès à un bateau de type annexe, avec une pompe péristaltique (Athena®, Proactive Environmental Products LLC, Bradenton, Florida, USA), montée de chaque côté du bateau pour les réplicas de terrain. Les prélèvements se font le long d'un transect, afin de prendre en compte la distribution inégale des molécules d'ADN dans l'eau, pour un prélèvement total de 30L par filtre. Avec une pompe péristaltique réglée pour un débit de $1L.min^{-1}$, il est nécessaire de filtrer pendant 30 min. La vitesse minimale des bateaux étant rarement inférieure à 5 nœuds (~10km/h), la distance à parcourir avoisine les 5km et sera influencée par le vent, le courant et la houle. Pour limiter la longueur d'un même site, il a été décidé dans un premier temps de faire des transects en rectangle (2 km * 500 m) pour intégrer à la fois les compartiments côtier et pélagique (Fig. 6. A). Cependant, les résultats étaient décevants due à la dilution du signal ADNe au large. La stratégie a rapidement évolué vers des transects linéaires, longeant l'habitat, de 2km de long avec aller/retour (Fig. 6. B). Dans certains cas, cette méthodologie a été adaptée à la question posée. Par exemple, dans le chapitre 4, on effectue une comparaison entre les performances de l'ADNe et des caméras sous-marines. La caméra étant posée sur le fond, un transect de 2 km n'était pas adapté pour comparer les deux méthodes efficacement. L'adaptation de la méthode a consisté à naviguer autour du point où était posé la caméra pendant la durée classique de 30 min pour standardiser le prélèvement à 30L (Fig. 6. C). Une dernière méthodologie a été récemment développée pour limiter la distance à l'habitat causée par les prélèvements en surface. Il s'agit d'utiliser des tuyaux de prélèvement d'eau plus long afin d'atteindre des profondeurs classiques de récifs coralliens : entre 10 et 30m. Le coût unitaire d'un tuyau de cette taille contraint toutefois les équipes à les réutiliser entre différents sites et nécessite un protocole de décontamination en filtrant à vide dans la javel pour ne pas contaminer des sites distincts avec la signature génétique du précédent.

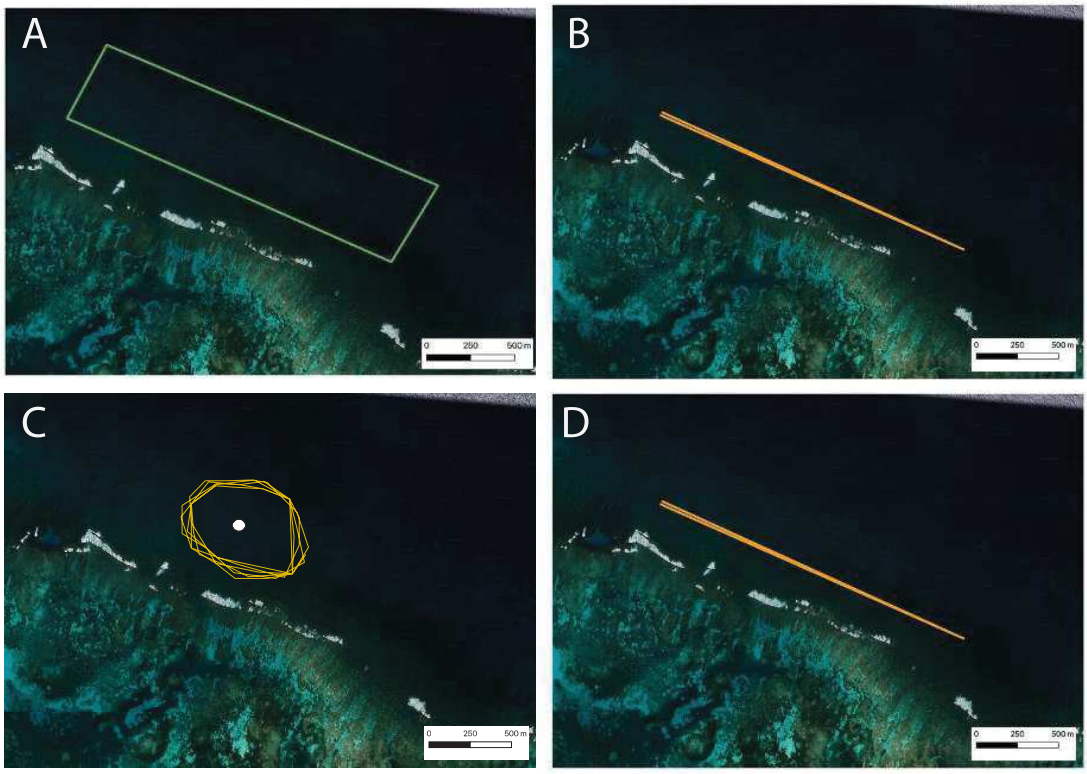


Fig. 6. Illustration de la stratégie d'échantillonnage initiale avec un rectangle de 2km*0.5 km (A) et modifiée par la suite en un aller-retour de 2km le long de l'habitat (B), utilisée en cas de comparaison avec des méthodes de recensement traditionnelles telles que les vidéos sous-marines (C) en utilisant un tuyau prélevant en sub-surface (<1m) pour ces trois stratégies, ou un tuyau assez long pour s'approcher de l'habitat benthique (D).

Le protocole de surface classique trouve ses limites par le fait que la pompe péristaltique n'est pas étanche et nécessite une étape de bricolage pour l'étanchéifier en cas de houle importante. Même avec ces précautions, plusieurs pompes ont cessé de fonctionner due aux contraintes de terrain auxquelles elles avaient été exposées. Pour travailler en mer plus efficacement, une première alternative de pompe résistante à l'eau a été trouvée avec un prototype de pompe péristaltique début 2018. Ce modèle n'a pas eu la résistance attendue et a cessé de fonctionner dès ses premières utilisations sur le terrain. La conception de la pompe était inadaptée, car le mécanisme rotatif chauffait au point de faire fondre les pignons de la tête péristaltique, rendant la pompe inutilisable (Fig. 7). Ce modèle a donc été abandonné.

Fig. 7. Organisation interne du prototype de pompe péristaltique abandonné suite à une trop importante surchauffe interne du mécanisme. *Crédit : V. Marques.*

La base complète ADNe regroupe des échantillons provenant de différentes méthodologies, reflétant l'évolution des méthodes de prélèvement. La liste complète des dates et méthodes d'échantillonnage de chacun des sites est présentée Table. 2.

Table. 2. Informations méthodologiques sur les échantillons de la base ADNe par campagne.

Campagne	Année	# sites	# stations	# filtres	Méthode échantillonnage	Filtre	Volume
Lengguru	2017	11	46	93	Sacs plastiques stériles (points)	Sterivex	2L
Guadeloupe	2018	1	17	28	Transects A	Capsule	30L
Malpelo	2018	1	18	30	Transects A et C	Capsule	30L
Fakarava	2018	1	15	23	Transects B	Capsule	30L
Providencia	2018	1	10	20	Transects B	Capsule	40L
Santa Marta	2018	1	13	32	Transects B et C	Capsule	30L
Iles Éparses	2019	4	56	74	Transects B	Capsule	30L
Nouvelle Calédonie	2019	6	65	65	Transects D	Capsule	32L

Au cours de la thèse, des prélèvements plus profonds dans les zones mésophotiques et mésopélagiques ont également été réalisés de manière opportuniste, entre 50m et 3000m (Table. 3). Pour atteindre ces profondeurs, il a fallu déployer des bouteilles Niskin pour prélever l'eau sur des points ponctuels. Dans la mesure du possible, des réplicas ou sous-échantillons issus de plusieurs bouteilles

ont été prélevés. Au cours de missions sur des navires océanographiques, le déploiement des Niskin s'est fait le long d'un câble en acier suite à la défaillance de la rosette (Fig. 8. A). Lorsque nous ne disposions que d'annexes, nous avons déployé et remonté des Niskins à l'aide de canne à pêche et moulinet de pêche électrique (Fig 8. B) ou à la main. Le moulinet s'est avéré de puissance trop légère pour remonter des bouteilles de 10L de façon optimale à forte cadence, mais fonctionne pour des déploiements ponctuels et peu profonds (< 200m, la longueur maximale de fil qu'il est possible d'enrouler tout en ayant un diamètre suffisant pour résister à la charge d'une bouteille pleine).

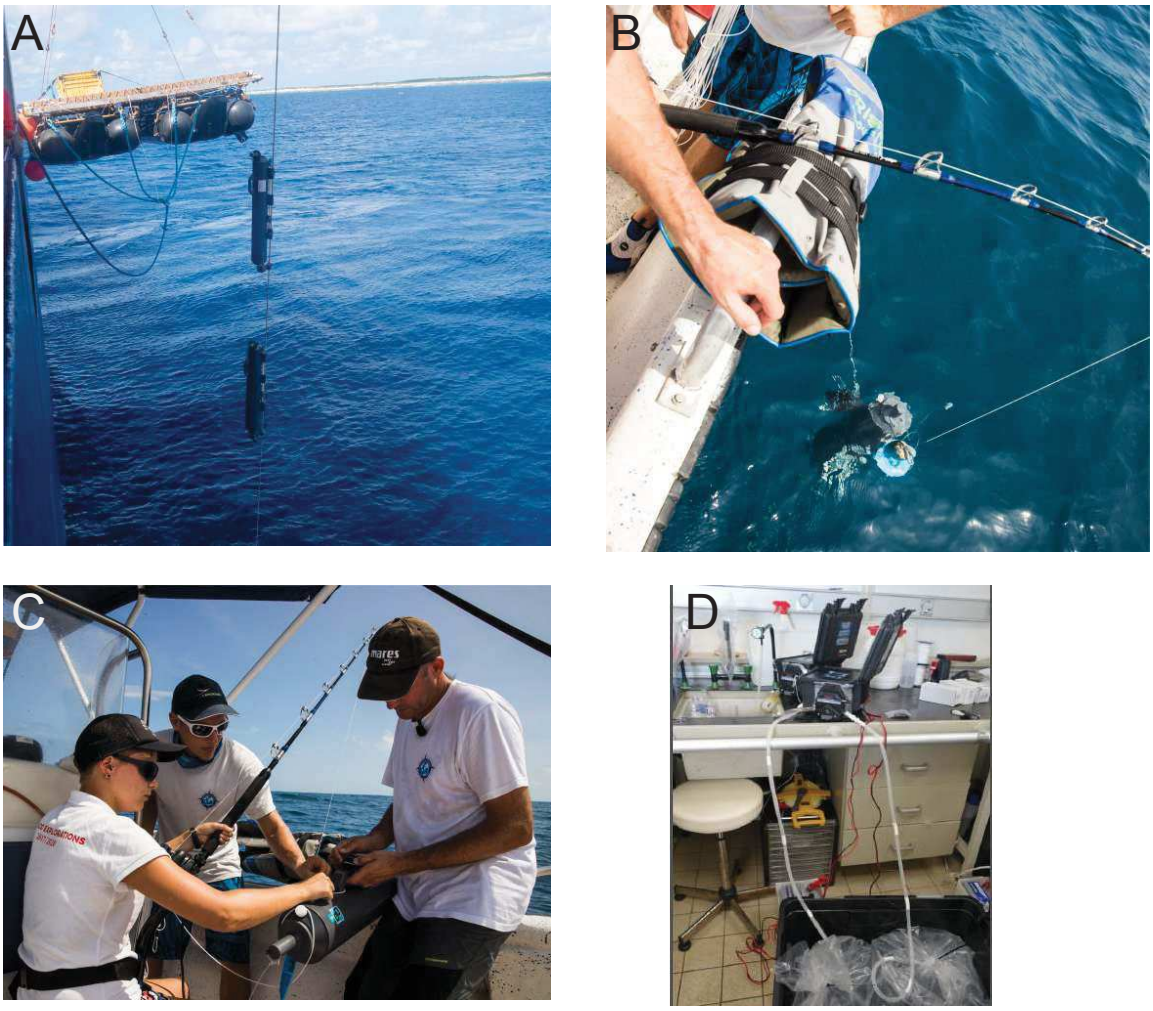


Fig. 8. Photos de terrain illustrant le déploiement de bouteilles Niskin depuis un navire océanographique jusqu'à 3000m de profondeur (A) ou depuis une annexe jusqu'à 200m de profondeur (B), ainsi que la récupération de la bouteille pour filtration sur l'annexe (C) ou filtration au laboratoire du contenu des bouteilles vidées dans des sacs stériles (D*). Crédits : V. Marques (A, D*), *O. Borde (B, C).*

Table. 3. Tableau récapitulatif des échantillons profonds (>50m) récoltés au cours de l'échantillonnage.

Campagne	Année	# stations	Profondeur max (m)	Distance min au substrat (approximative, m)	Méthode échantillonnage	Volume
Lengguru	**2017**	**13**	**270**	**2**	**Niskin et plongée recycleur**	**2L**
Ile de Malpelo	**2018**	**2**	**600**	**NA**	**Niskin**	**10L**
Fakarava	**2018**	**2**	**200**	**50**	**Niskin**	**10L**
Santa Marta	**2018**	**2**	**150**	**15**	**Niskin**	**10L**
Iles Éparses	**2019**	**9**	**3000**	**5**	**Niskin**	**10L**

Protocole récolte de tissus

Pour pouvoir constituer une base de référence génétique, il est nécessaire de récolter des échantillons de tissus des espèces susceptibles d'être détectées sur les filtres. Pour cela, notre démarche a été majoritairement opportuniste lors de déplacements internationaux afin de récupérer des tissus à travers la pêche autonome ciblée lorsque les permis de pêche ont pu être délivrés, ou sur les marchés aux poissons, en collaboration avec d'autres équipes utilisant ces mêmes individus pour leur tissus ou intestins (Fig. 9). Sur le site de l'île de Malpelo (Colombie) par exemple, qui est une réserve naturelle où vivent quelques espèces endémiques, il n'a pas été possible d'obtenir des permis de prélèvement de poissons au filet donc aucun individu n'a pu être prélevé. L'équipe du premier leg est toutefois parvenue à récolter quelques échantillons de tissus au marché aux poissons de Panama. Le protocole de pêche sous-marine consiste à prélever les individus à l'aide de filets à mailles de différentes tailles pour couvrir une large étendue d'espèces, ou bien à l'aide d'un fusil-harpon pour les espèces les plus grosses. Les filets lestés sont disposés au fond en plongée en scaphandre sur une zone adéquat et protégée des courants forts, puis les espèces d'intérêt sont poussées dans les filets et récupérées dans une nasse avant d'être remontée à bord. Au laboratoire humide, les individus sont identifiés par un code unique et identifiés à l'espèce lorsque c'est possible. Un morceau de nageoire ainsi que de muscle sont prélevés puis conservés dans des tubes uniques rempli d'ethanol à 96°, qui sera renouvelé au bout de 24-48h pour maximiser la conservation des échantillons.

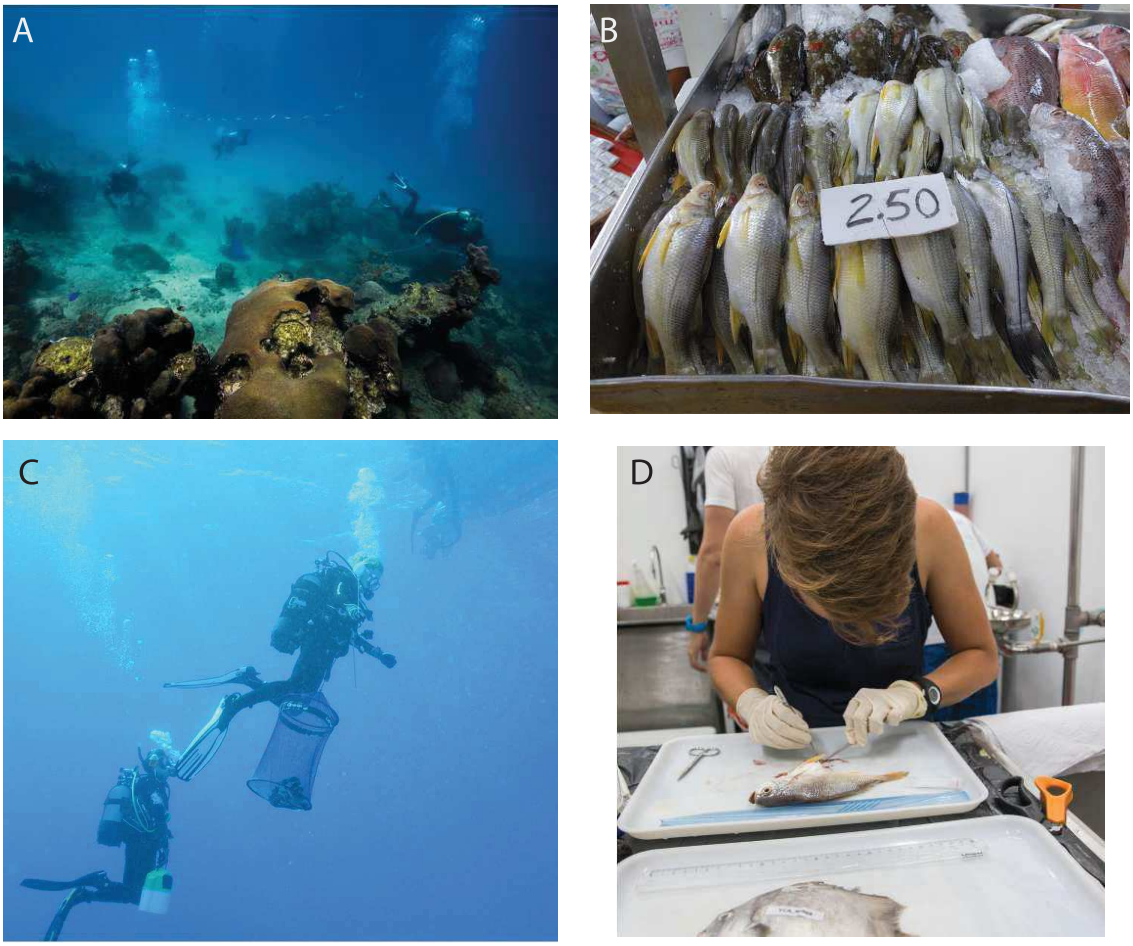


Fig. 9. Illustration de la collecte de tissus de poissons en plongée avec l'utilisation de filets maillants (A), sur les marchés (B), remontée des poissons dans les nasses (C), puis récupération des tissus au laboratoire (D). *Crédits : O. Borde (A,D). R. Hocdé (B) et C. Albouy (C).*

5. Analyses moléculaires et séquençage

Dans cette partie, deux points seront discutés. Dans un premier temps, la constitution de la base de référence génétique à partir des échantillons de tissus récoltés lors du travail de terrain, et dans un second temps, le traitement des échantillons d'ADNe metabarcoding.

Glossaire

Amorce : Courte séquence d'ADN ou ARN dont la composition en nucléotides permet d'amplifier une portion d'ADN cible avec de l'ADN polymérase.

Clustering : ou « *analyse de groupement* ». Méthode d'analyse de données permettant de diviser un ensemble de données en sous-ensembles partageant des caractéristiques communes.

HTS : Séquençage haut débit (« *High Throughput Sequencing* »). Ensemble de méthodes de séquençage apparues en 2005 et permettant le séquençage d'une grande quantité de données à bas coûts et rapidement.

MOTU : « *Molecular Operational Taxonomic Unit* », ou unité taxonomique moléculaire. Il s'agit d'un regroupement de séquences jugées similaires, généralement par le biais d'un algorithme de clustering.

PCR : Réaction en chaîne par polymérase (« *Polymerase Chain Reaction* »). Procédé moléculaire permettant la réplication en grand nombre de séquences amplifiées par des amorces à partir d'une faible quantité d'ADN initial.

Pipeline : Commandes de code informatique automatisé permettant l'exécution de plusieurs logiciels ou programmes, où chaque sortie constitue l'entrée de la tâche suivante pour limiter l'intervention humaine au cours du traitement. Par exemple, le traitement bio-informatique de séquences d'ADNe metabarcoding : le code permet de passer de sorties de séquençage (première entrée) à des tables de présence/absence des espèces/MOTUs (dernière sortie).

Traitement des échantillons de la base de référence

Les échantillons de tissus issus des campagnes de terrain sont rapatriés au laboratoire de biologie moléculaire et conservés au frigo en attendant les procédures d'extraction d'ADN. Une partie de ce travail a été réalisé au CEFE à Montpellier, et la majorité des échantillons ont été traités à l'ETH de Zurich par l'équipe du Prof. Dr. Loïc Pellissier. Ma contribution pour cette partie est faible et concerne quelques échantillons traités à Montpellier.

Les morceaux de nageoires conservés dans l'éthanol sont séchés à l'étuve, puis l'ADN est extrait en utilisant le kit « DNeasy Blood and Tissues kit » (QIAGEN®) selon les recommandations de QIAGEN®. L'ADN extrait est ensuite amplifié par PCR (voir Glossaire) avec les amorces V05F898 en forward (AAACTCGTGCCAGCCACC) et teleo en reverse (Table. 4), d'une longueur totale d'environ 700 paires de bases. Ce couple d'amorces permet de séquencer la presque totalité du gène 12S et d'inclure les 3 marqueurs utilisés dans le cadre de cette thèse (teleo, Vert01 et Chond01, voir Table. 4). Les sorties de PCR sont par la suite envoyées au séquençage par la méthode Sanger. Les chromatogrammes reçus (un par sens de lecture) sont alignés puis corrigés à la main pour extraire la séquence de chaque échantillon en utilisant le logiciel Geneious® (Fig. 10. A). De façon générale, la couverture du gène 12S était quasiment totale par ce couple d'amorces, tous les marqueurs ADNe situés sur le 12S sont couverts par la portion amplifiée (Fig. 10. B).

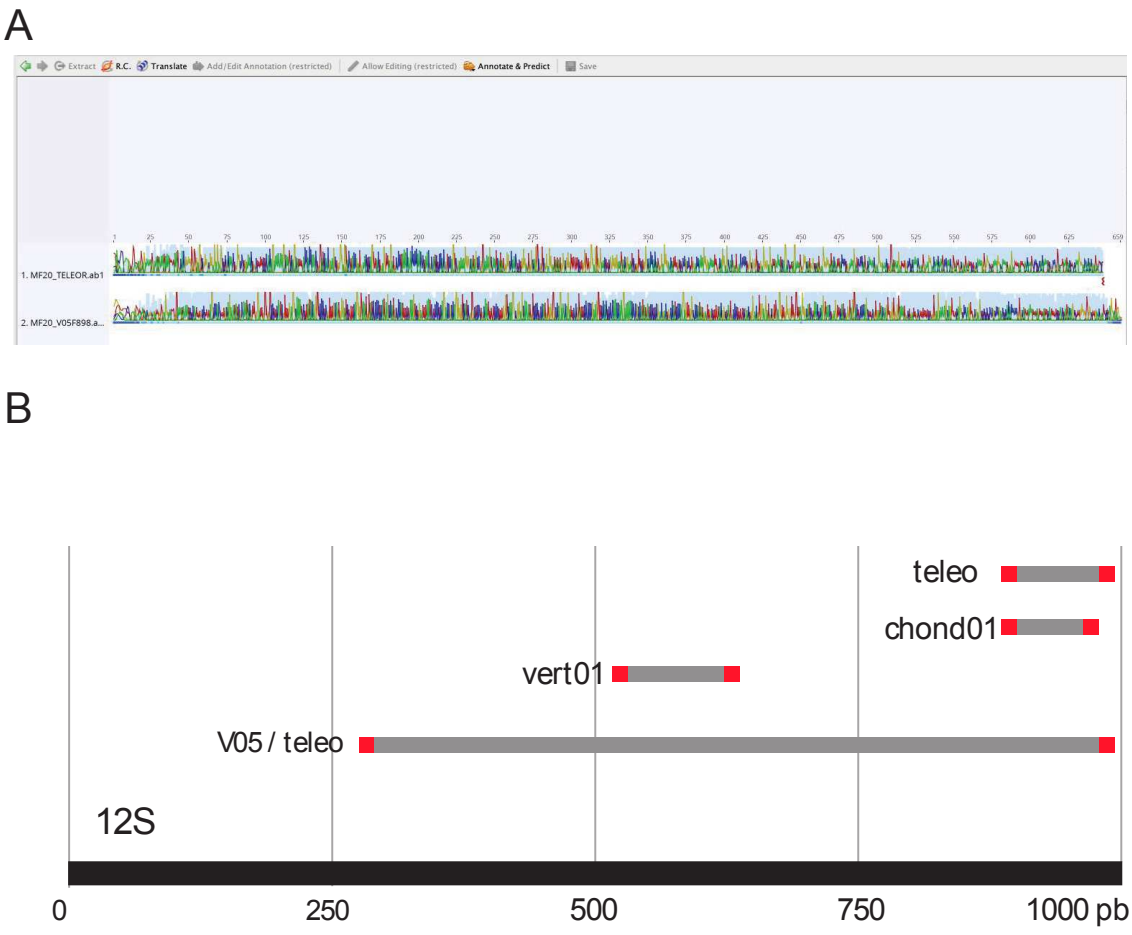


Fig. 10. Capture d'écran lors de l'alignement d'un chromatogramme dans le logiciel Geneious© (A) et illustration schématique du placement des marqueurs ou paires d'amorces sur le gène mitochondrial ARN ribosomal 12S (B). Les parties en rouge représentent les amorces de part et d'autre de la séquence d'intérêt. *Illustration (B) inspirée de Zhang et al. (2020).*

Lorsque l'identification taxonomique n'a pas pu être effectuée à l'espèce avec certitude sur le terrain, les individus seront séquencés pour le gène mitochondrial du cytochrome c oxydase codée I (COI) à l'aide d'amorces spécifiques pour cette région pour leur identification taxonomique via barcode génétique (Valentini et al. 2009).

Traitement des échantillons d'ADNe

Dans le cadre des analyses de l'ADNe metabarcoding, les filtres ADNe sont envoyés au laboratoire à l'issue de l'échantillonnage où ils seront extraits, amplifiés puis séquencés. Les principales étapes, données à titre général et indicative du procédé global, sont présentées Fig. 11., où la partie « Laboratoire moléculaire » est effectuée par l'entreprise SPYGEN par le personnel habilité, qui sous-traite la préparation de librairies et du séquençage au laboratoire Fasteris (Genève, Suisse). La salle d'extraction de l'ADN est dédiée à cette étape et équipée de pression positive et d'un renouvellement de l'air fréquent pour éviter les risques de contamination. Le personnel est équipé de protection à usage unique et les paillasses sont décontaminées à l'eau de javel avant et après toute manipulation. Les détails du protocole de laboratoire complet utilisé dans le cadre de ces travaux sont présentés dans l'article de Pont et al. (2018). Je n'ai pas participé à ces étapes de laboratoire et de séquençage.

Table. 4. Séquences des amorces forward et reverse des marqueurs génétiques ADNe metabarcording utilisés et leur taille moyenne en nombre de paires de bases.

Marqueur	Sequence forward	Sequence reverse	Taille (pb)
teleo	ACACCGCCCGTCACTCT	CTTCCGGTACACTTACCATG	~64
chond01	ACACCGCCCGTCACTCTC	CATGTTACGACTTGCCTCCTC	~45
vert01	TAGAACAGGCTCCTCTAG	TTAGATACCCCACTATGC	~100

Les étapes classiques de la partie laboratoire de traitement de l'ADNe par metabarcoding :

(1) **Extraction de l'ADN**. L'ADN est extrait des capsules de filtration hermétique suivant les recommandations d'un kit d'extraction commercial ou d'un protocole dédié. (SPYGEN)

(2) **Amplification par PCR**. L'ADN extrait est amplifié par réactions en chaine de polymérase (PCR, voir Glossaire) en utilisant l'amorce metabarcoding d'intérêt. Au cours de cette thèse, l'amorce teleo ciblant les poissons osseux et élasmobranches a été majoritairement utilisée, avec en complément les amorces chond01 et vert01, ciblant respectivement les élasmobranches plus spécifiquement et les organismes vertébrés. (SPYGEN)

(3) **Multiplexage**. Le multiplexage consiste à assigner une série de nucléotides connus et uniques (ici, 8 paires de bases) à un échantillon environnemental, appelé tag (« étiquette »). Puisqu'une librairie envoyée au séquençage contient de nombreux échantillons, le tag permet d'assigner à postériori les séquences à l'échantillon environnemental dont elles sont issues. (SPYGEN)

(4) **Séquençage**. La préparation des librairies est effectuée par la méthode de ligation (c'est le cas de ces travaux) ou bien par une seconde PCR. Le séquençage se fait couramment sur des plateformes Illumina MiSeq ou HiSeq en paired-end (« *double-lecture* »), c'est à dire que chaque séquence est lue en 5' – 3' et en 3' – 5' afin d'avoir une double lecture. (FASTERIS)

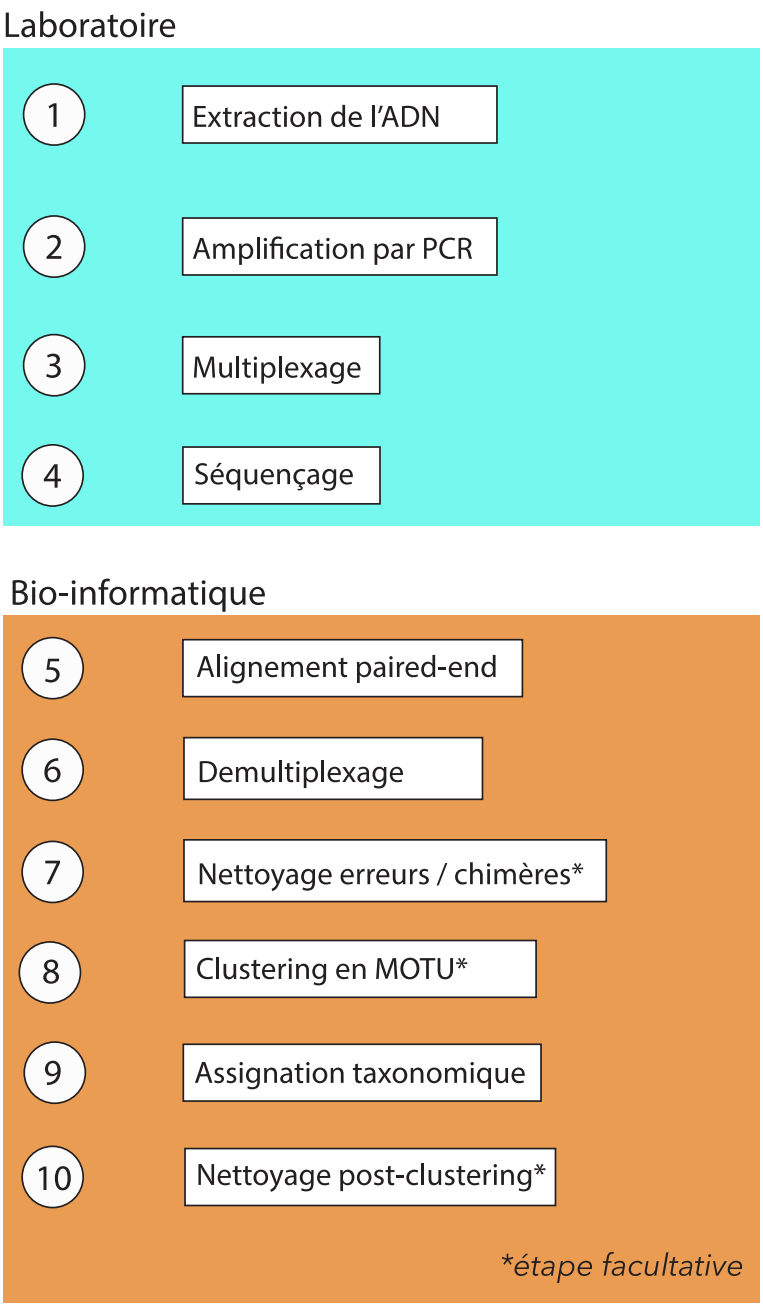


Fig. 11. Schéma du procédé d'ADNe en metabarcoding.

6. Bio-informatique

Principes généraux

Après l'étape de séquençage, il est nécessaire d'automatiser le traitement des données compte tenu de leur complexité et grande taille. Pour cela, les grandes étapes indicatives d'un traitement par metabarcoding sont indiquées Fig. 11.

(5) **Alignement paired-end (« *double lecture* »)**. Après séquençage, il est nécessaire d'assembler les deux fichiers (un par sens de lecture) afin de retirer les séquences non couvertes par les deux sens de lecture, ou avec une qualité de séquençage trop basse, indiquée par le score du FASTQ délivré par la plateforme de séquençage.

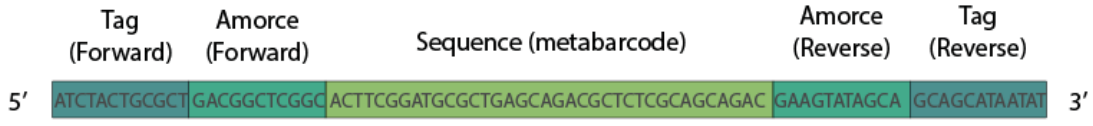


Fig. 12. Illustration schématique du contenu d'une lecture après le séquençage.

(6) **Demultiplexage**. L'étape de « demultiplexage » permet de séparer les séquences d'une même librairie afin de les attribuer à leur échantillon environnemental grâce au tag. A cette étape, on retire également les fragments de séquences correspondant aux amorces, voir Fig. 12, afin de ne conserver que le metabarcode d'intérêt.

(7) **Nettoyage erreurs / chimères**. Il est possible d'ajouter une ou plusieurs étapes de nettoyage basée sur la reconnaissance des séquences jugées comme erronées ou bien chimériques. Les chimères représentent un cas particulier d'erreur obtenue pendant l'étape de PCR où la séquence amplifiée correspond en réalité à un mélange de deux séquences réellement présentes. On peut également filtrer par la taille des séquences, où selon les marqueurs, les tailles extrêmes sont peu susceptibles de représenter autre chose que des artefacts. Certains algorithmes de nettoyage se basent sur la qualité des bases du fastq, ou sur les paramètres d'abondance et de similarité entre les séquences pour effectuer du nettoyage qualité (voir DENOISE, DADA2, etc)(Callahan et al. 2016, Edgar 2016, Amir et al. 2017).

(8) **Clustering**. Pour enlever les erreurs des séquences ou bien pour travailler avec des unités taxonomiques, une étape facultative consiste à utiliser un algorithme de clustering

(« *analyse de groupement* ») (Voir Glossaire) afin de regrouper des séquences jugées similaires. De nombreux algorithmes existent et fonctionnent généralement sur des principes de clustering distincts (lien simple, lien complet, algorithmes de voisinage), antérieurs aux techniques de séquençage haut-début (Legendre and Legendre 1998, Koskinen et al. 2015).

(9) **Assignation taxonomique**. Les séquences sont ensuite comparées à une base de référence afin d'assigner une taxonomie à chaque séquence. Si la base de référence est incomplète, les séquences non assignées à une espèce précise sont généralement assignées à un plus haut niveau taxonomique (genre, famille). Différents algorithmes permettent d'assigner une taxonomie à une séquence non représentée en utilisant notamment l'approche du plus proche ancêtre commun, LCA (« Lower Common Ancestor »), qui assigne le niveau taxonomique le plus précis en tenant compte des contraintes de remplissage de la base de référence.

(10) **Nettoyage post-clustering**. Il est possible de continuer le nettoyage de la qualité des séquences par l'utilisation d'algorithmes post-clustering, ou bien de seuils de qualité tels que retirer les séquences présentes dans un échantillon unique, trop peu abondantes, etc.

Par la suite, les données se présentent sous la forme de matrice espèce/MOTUs – station, qui permettent de procéder aux analyses écologiques.

La gestion des erreurs de séquençage

Un challenge important des technologies de séquençage à haut débit concerne la gestion des séquences erronées, c'est à dire ne reflétant pas la réalité biologique des échantillons environnementaux originaux. Ces erreurs trouvent majoritairement leur origine à deux étapes : i) lors de la PCR, où une erreur d'un ou plusieurs nucléotides lors de la réplication de séquences par l'enzyme polymérase peut se former, ou la production de chimères, une séquence hybride issue de deux séquences réelles, ii) lors du séquençage, où une base est mal lue par l'appareil.

La gestion des erreurs de séquençage est débattue depuis l'avènement du séquençage à haut débit (HTS, voir Glossaire) (Quince et al. 2009, Patin et al. 2013, Edgar and Flyvbjerg 2015). La microbiologie est un domaine pionnier des techniques de séquençage en ADN environnemental, qui utilisa cette méthode pour séquencer l'ADN des organismes non cultivables en laboratoire pour quantifier la quantité et la diversité de micro-organismes présents dans des environnements variés. L'étude de la « biosphère rare », c'est à dire présente en très faible abondance est compromise par la gestion des erreurs et du bruit généré par les technologies de séquençage haut-débit (Welch and Huse

2011). Des travaux démontrent que la surabondance d'unités taxonomiques rares détectées est à considérer avec précaution car elle est susceptible de représenter des erreurs de séquençage et de PCR et non une réalité biologique (Reeder and Knight 2009, Huse et al. 2010, Kunin et al. 2010). Comme la majorité des erreurs sont des occurrences rares dans les jeux de données, il est difficile de discerner les vraies occurrences d'espèces rares des erreurs, car ces deux catégories présentent généralement les mêmes caractéristiques (Patin et al. 2013).

Le clustering des séquences est une première approche pour corriger ces erreurs. Historiquement il s'agissait plutôt de regrouper des « souches » ou « espèces » de micro-organismes proches (97% de similarité génétique), regroupées sous forme d'unités taxonomiques (Edgar 2018). Ces unités taxonomiques n'avaient pas vocation à représenter des espèces uniques, mais des complexes d'espèces proches sous l'hypothèse qu'une proximité génétique entraînerait une proximité fonctionnelle, en sachant que le concept même d'espèce est plus difficile à poser pour les organismes procaryotes (Achtman and Wagner 2008). Ce dogme a depuis été sérieusement contesté (Callahan et al. 2017, Edgar 2018). Les nouvelles recommandations se basent sur (i) du clustering des séquences avec un seuil de similarité à 99%, afin de délimiter un maximum d'entités biologiques réelles, (ii) sur du clustering non basé sur une similarité génétique fixe (e.g. SWARM, (Mahé et al. 2015)), ou encore (iii) sur la définition de variants de séquence d'amplicon (ASV, « Amplicon Sequence Variant ») (Callahan et al. 2017). Les variants de séquences exactes (ESV, « Exact Sequence Variant ») ou OTU à rayon nul (zOTUs, « zero-radius OTUs ») sont des synonymes pour dénommer les ASV. Les méthodes de clustering ou « débruitage » (de l'anglais « denoising ») sont aujourd'hui largement utilisées pour nettoyer des séquences erronées (Fig. 13). Les méthodes de clustering sont variées, et parmi celles n'utilisant pas de seuil de similarité fixe on trouve l'algorithme SWARM (approche ii), qui a été développé sur les microorganismes par l'équipe de recherche des expéditions TARA (de Vargas et al. 2015, Mahé et al. 2015). SWARM construit des chaînes de séquences et définit chaque (M)OTUs en fonction d'un double paramètre : son abondance et sa proximité génétique avec d'autres séquences plus ou moins abondantes (Fig. 14. A). La plupart des algorithmes de « débruitage » génèrent un modèle d'erreur sur les sorties de séquençage, puis appliquent ce modèle pour trier les vraies séquences de celles représentant des erreurs et définir des ASVs (voir Callahan et al. (2016) pour le fonctionnement détaillé de DADA2) (Fig. 14. B). Parmi la pléthore d'outils et de philosophies de nettoyage et traitement des séquences disponibles, il s'agit surtout de trouver le compromis acceptable entre conserver un risque de faux positifs générés par les erreurs mais également garder le signal de rareté biologique. Cela revient à limiter les faux positifs au risque de créer des faux négatifs en supprimant des séquences réelles. Le débat autour de l'utilisation de (M)OTUs versus ASVs n'est pas tranché, et les utilisations semblent dépendre du marqueur moléculaire, taxon d'étude, tolérance aux faux-positifs et niveau de détail

biologique requis pour répondre à la question posée (Koskinen et al. 2015, Leger et al. 2015, Nearing et al. 2018). Plus particulièrement, ces recommandations sont basées sur l'application à la microbiologie, et sa transposition à d'autres contraintes et d'autres taxon tels que les métazoaires est incertaine.

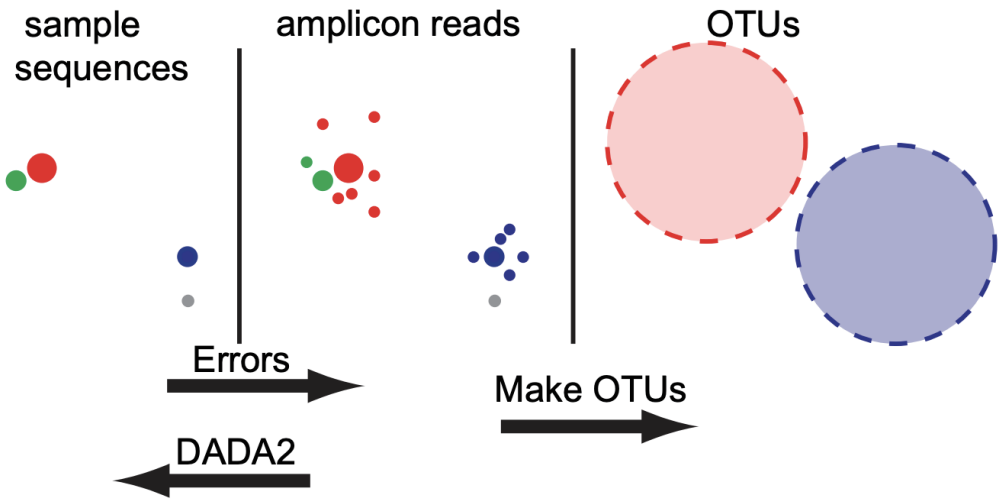


Fig. 13. Illustration schématique de la différence de philosophie entre le débruitage (« *denoising* ») par DADA2 et la création de (M)OTUs par le clustering. A noter que DADA2 n'est pas en mesure d'identifier la totalité des erreurs, et que les (M)OTUs ont des performances variables selon les algorithmes pour regrouper les séquences entre elles. *Figure adaptée de Callahan et al (2016).*

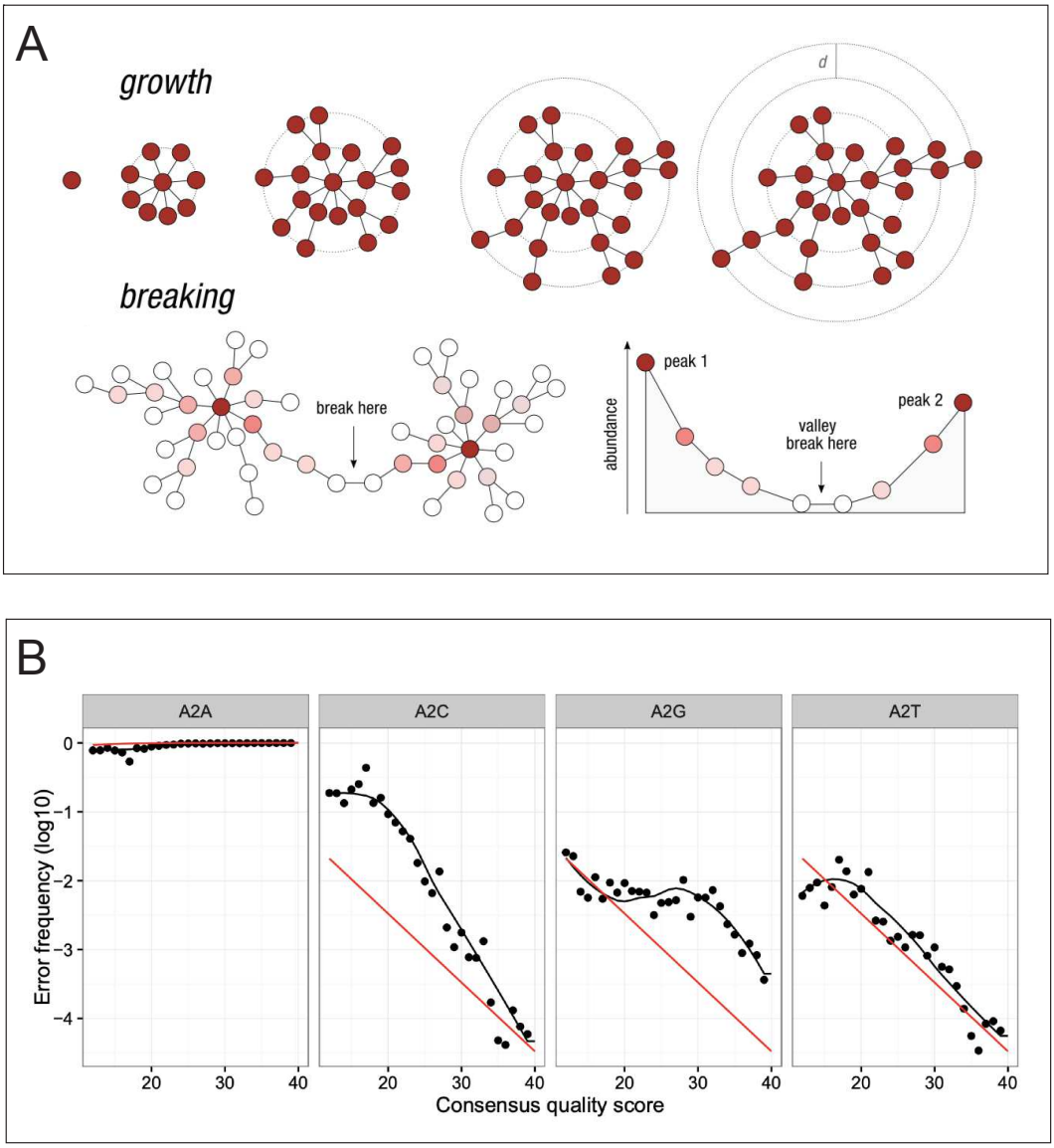


Fig. 14. Illustrations des fonctionnements de l'algorithme de SWARM (A), où des réseaux de séquences sont créés puis séparés en fonction de l'abondance relative et de la proximité des séquences et (B) d'un modèle d'erreur de DADA2 en fonction de la qualité des bases en sorties d'Illumina (2), où la base correcte est « A », l'axe x le score de qualité et l'axe y la fréquence de la transition d'un nucléotide autre que « A ». Les points représentent les fréquences observées et la ligne rouge le modèle d'erreur. *Figure adaptée de Mahé et al. (2014) (A) et Callahan et al. (2016).*

Développements au cours de la thèse

Mon travail de thèse a été guidé par la volonté de proposer des métriques de diversité réalistes quant au nombre d'espèces réellement présentes, sans dépendre d'une base de référence complète et intégrant la gestion des erreurs de PCR et séquençage. L'introduction de telles métriques a nécessité de développer un nouveau pipeline bio-informatique. Le **chapitre 3** présente ce pipeline en détails, ainsi que ses résultats lors d'une application en conditions réelles. Dans le but de définir un pipeline qui aurait

la possibilité d'estimer des unités taxonomiques fiables par rapport au nombre réel d'espèces, il était nécessaire de tester les outils développés par la communauté microbienne sur des marqueurs et taxa appartenant aux métazoaires et dans des conditions réelles. La plupart des tests de pipelines ou d'algorithmes sont effectués sur des taxons bactériens ou eucaryotes comme les diatomées (Apothéloz-Perret-Gentil et al. 2017, Vasselon et al. 2017, Keck et al. 2018, Pawlowski et al. 2018), où les marqueurs et caractéristiques génétiques des organismes sont très éloignés de ceux des métazoaires vertébrés. L'ADNe sur des vertébrés implique également un nombre plus élevé de cycles de PCR (50 dans le cas présent) car le signal est plus rare et dilué dans l'environnement, créant un plus grand nombre de séquences erronées. De plus, de nombreux tests de pipeline s'effectuent sur la base de résultats de « *mock community* » (assemblage in-vitro d'une composition de séquences connues), qui sont indispensables, mais ne représentent pas la complexité des conditions réelles. Je cite ici Koskinen et al. (2015) pour appuyer ce propos : « […] *testing these tools merely with mock communities may not result in representative results with real data, as the diversity and evenness of the natural microbial communities vary and the true performance of tested algorithms remains unclear*.".

En partant de l'organisation générale des étapes d'un pipeline bio-informatique de traitement des données metabarcoding et des contraintes liées à la gestion des erreurs, deux pipelines bio-informatiques ont été développés au cours de cette thèse. Ces deux pipelines utilisent des outils déjà existants dans la littérature (ex : Swarm, Obitools, Vsearch) et ont été modifiés et codés pour notre propre utilisation sur les poissons en collaboration avec P-E. Guérin, ingénieur bio-informaticien au CEFE. Les pipelines ont été optimisés grâce à l'utilisation conjointe du gestionnaire de pipeline snakemake (Köster and Rahmann 2012), de paquets conda (https://www.anaconda.com/) ou de containers singularity (Kurtzer et al. 2017) afin de garantir la bonne organisation, la portabilité et la reproductibilité des analyses. Les pipelines ont été conçus pour gérer le traitement des données issue d'un séquençage multi-marqueur avec un haut niveau de parallélisation pour optimiser les temps de calcul.

Pipeline 1 (« MOTU pipeline ») : Le cœur du pipeline est basé sur la combinaison des outils et algorithmes **Vsearch**, **Swarm** et **LULU** (Mahé et al. 2015, Rognes et al. 2016, Frøslev et al. 2017) et génère des unités taxonomiques (MOTUs). Ce pipeline est utilisé lorsque les bases de références sont non-exhaustives. **Vsearch** est une boîte à outils dont l'utilisation est complètement libre et accessible (contrairement à son analogue USEARCH) (Edgar 2010), **Swarm** est un algorithme de clustering et **LULU** est un algorithme de post-clustering qui utilise les scores de similarité avec les patrons de cooccurrence pour nettoyer les unités taxonomiques.

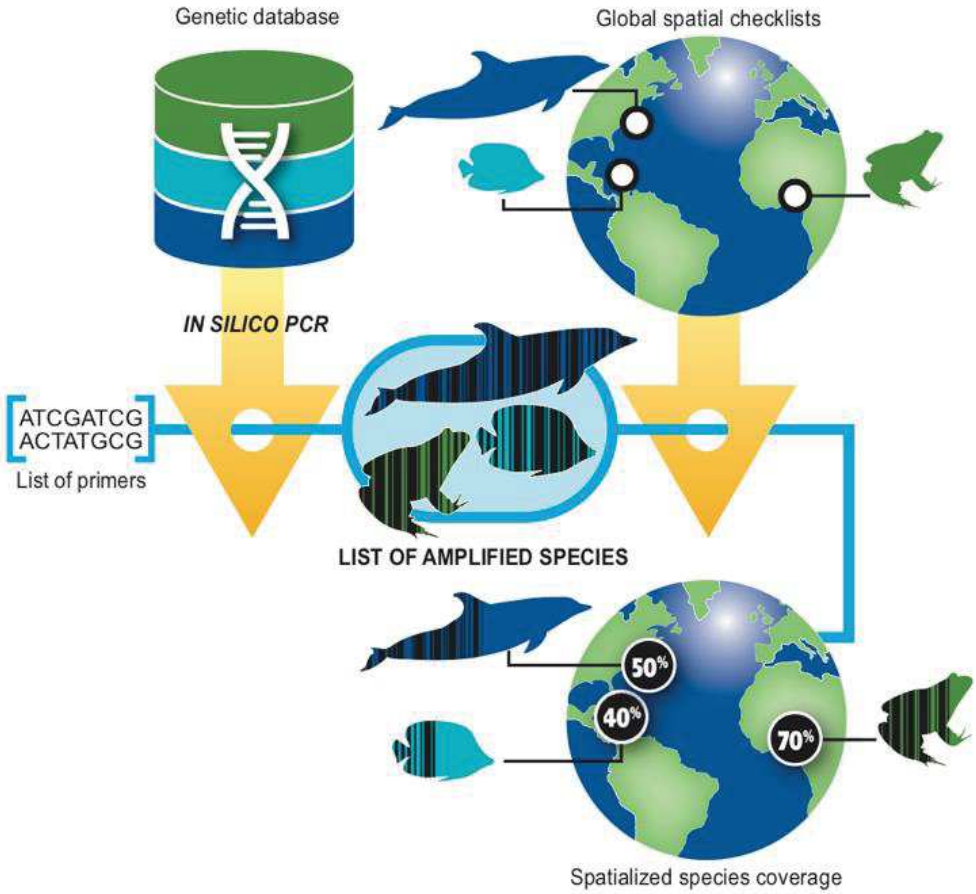
Pipeline 2 : Il est basé sur la boîte à outils Obitools, selon les recommandations initiales de ses auteurs (Boyer et al. 2016). Les erreurs éventuelles ne sont filtrées qu'avec la fonction *obiclean*, qui nettoie les données selon la proximité des séquences deux à deux et le ratio de leurs abondances avec des paramètres choisis. Ce pipeline s'utilise surtout dans le cas où les bases de référence sont proches de l'exhaustivité car aucune étape de clustering ni de débruitage par modèle d'erreur n'est appliquée.

Pour les deux pipelines, des seuils de qualité peuvent par la suite être appliqués pour (i) supprimer les lectures de séquence représentant du « *tag-jump* », c'est à dire une mauvaise assignation à un échantillon au moment du demultiplexage à cause des erreurs de lecture de nucléotide sur les tags (Schnell et al. 2015), (ii) les séquences en trop faible abondance, susceptibles de représenter des erreurs, (iii) les séquences présentes dans une unique PCR, car une même erreur de PCR a peu de chance d'être présente par chance dans plusieurs échantillons.

Chapitre 2 - GAPeDNA: Assessing and mapping global species gaps in genetic databases for eDNA metabarcoding

Manuscrit A

Marques V., Milhau T., Dejean T., Manel S., Mouillot D., Juhel J-B. GAPeDNA: Assessing and mapping global species gaps in genetic databases for eDNA metabarcoding. (2020). Diversity and Distributions. In Press.



1. Préface

Nous avons vu au chapitre d'introduction que les besoins en recensement et évaluation des communautés ichtyologiques sont très importants, notamment en milieu tropical, et que le metabarcoding de l'ADN environnemental est une méthode très prometteuse pour compléter ou remplacer les méthodes classiques. Cependant, une importante limitation concerne la couverture non exhaustive des dans les bases de référence génétiques. La détection de séquences sur des filtres ADNe sans correspondance taxonomique montre une forte sous-estimation de la richesse des espèces présentes. Alors que l'objectif à long terme est de séquencer la totalité des 32 000 espèces décrites de poissons osseux (Osteichthyes), il n'existe pas de base de données permettant de savoir quelles espèces ont été déjà séquencées, pour quel marqueur génétique et surtout de connaitre les régions où ces espèces manquantes se concentrent.

Une première étude a montré qu'à l'échelle de l'Europe, le remplissage des bases de références reste parcellaire avec une forte hétérogénéité spatiale et taxonomique (Weigand et al. 2019). Les espèces faisant l'objet de programmes de suivi dans plusieurs pays ont une meilleure couverture taxonomique que celles qui sont peu suivies, et les taxons plus charismatiques sont plus étudiés (i.e. meilleure couverture pour les poissons que les diatomées ou macroinvertébrés). Par contre l'hétérogénéité spatiale de la couverture taxonomique des bases de références génétiques pour les poissons n'est pas connue à l'échelle mondiale, alors que l'ADNe est fréquemment cité comme une méthode particulièrement adaptée à la détection d'espèces menacées ou non-indigènes. Combien d'entre elles sont présentement détectables compte tenu du remplissage des bases de référence ?

Ce chapitre vise à établir une référence de l'état de couverture taxonomique des bases de références pour (i) tous les marqueurs ADNe metabarcoding ciblant les poissons osseux, (ii) toutes les régions d'eau douce et marines. Une application web (R shiny) accompagne ce manuscrit afin d'explorer ce remplissage de manière interactive (https://shiny.cefe.cnrs.fr/GAPeDNA/).

Les résultats révèlent que la couverture taxonomique des bases de référence est très inégale spatialement pour tous les marqueurs, avec un fort déficit dans les régions tropicales, particulièrement pour les poissons d'eau douce. Les régions tempérées de l'hémisphère nord ont les bases de références les plus remplies pour tous les marqueurs et milieux. Les espèces menacées selon la liste rouge de l'IUCN sont plus représentées que les non-menacées, et les non-indigènes sont plus représentées que les non-indigènes pour tous les marqueurs.

Toutefois, aucun marqueur n'atteint un taux de couverture supérieur a 55% pour les espèces menacées et 70% pour les non-indigènes, ce qui montre la nécessité de se focaliser en priorité sur les espèces restant à séquencer au sein de ces catégories critiques pour exploiter pleinement la méthode ADNe en conservation. Trois bassins d'eau douce en particulier concentrent plus de 60 espèces menacées et non séquencées pour le marqueur 12S le plus utilisé dans la littérature : le Congo, le Mekong et le Mississipi. L'application GAPeDNA permet d'explorer la totalité de ces résultats de façon interactive et évolutive pour chacun des marqueurs génétiques en utilisant les poissons comme exemple. GAPeDNA a vocation à s'étendre pour inclure d'autres groupes taxonomiques si les données d'aires de répartition des espèces et les paires d'amorces sont disponibles et compilées.

2. Manuscrit A

GAPeDNA: Assessing and mapping global species gaps in genetic databases for eDNA metabarcoding

Virginie Marques[1,2*], Tristan Milhau[3], Camille Albouy[4], Tony Dejean[3], Stéphanie Manel[2], David Mouillot[1,5], Jean-Baptiste Juhel[1]

[1]: MARBEC, Univ. Montpellier, CNRS, Ifremer, IRD, Montpellier, France

[2]: CEFE, PSL Research University, EPHE, CNRS, UM, UPV, IRD, Montpellier, France

[3]: SPYGEN, 17 rue du Lac Saint-André Savoie Technolac - BP 274, Le Bourget-du-Lac, 73375, France

[4]: IFREMER, Unité Ecologie et Modèles pour l'Halieutique, Rue de l'Ile d'Yeu, BP21105, 44311 Nantes cedex 3, France

[5]: Australian Research Council Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, QLD 4811 Australia

*Corresponding author: virginie.marques01@gmail.com

Abstract

Aim:

Environmental DNA metabarcoding has recently emerged as a non-invasive tool for aquatic biodiversity inventories, frequently surpassing traditional methods for detecting a wide range of taxa in most habitats. The major limitation currently impairing the large-scale application of eDNA-based inventories is the lack of species sequences available in public genetic databases. Unfortunately, these gaps are still unknown spatially and taxonomically, hindering targeted future sequencing efforts.

Innovation:

We propose GAPeDNA, a user-friendly web-interface that provides a global overview of genetic database completeness for a given taxa across space and conservation status. As an application, we synthetized data from regional checklists for marine and freshwater fishes along with their IUCN conservation status to provide global maps of species coverage using the European Nucleotide Archive public reference database for 19 metabarcoding primers. This tool automatizes the scanning of gaps in these databases to guide future sequencing efforts and support the deployment of eDNA inventories at larger scale. This tool is flexible and can be expanded to other taxa and primers upon data availability.

Main conclusions:

Using our global fish case study, we show that gaps increase toward the tropics where species diversity and the number of threatened species were the highest. It highlights priority areas for fish sequencing like the Congo, the Mekong and the Mississippi freshwater basins which host more than 60 non-sequenced threatened fish species. For marine fishes, the Caribbean and East Africa host up to 42 non-sequenced threatened species. By presenting the global genetic database completeness for several primers on any taxa and building an open-access, updatable and flexible tool, GAPeDNA appears as a valuable contribution to support any kind of eDNA metabarcoding study.

Keywords: genetic markers, shiny, marine and freshwater fish, threatened species, IUCN, non-indigenous species, environmental DNA, reference database

Introduction

Aquatic ecosystems are increasingly impacted by human activities, threatening their biodiversity and causing major disruptions in their functioning (Cinner et al., 2016; Link & Watson, 2019; Reid et al., 2019). Marine systems are under severe defaunation with numerous local species extinctions (McCauley et al., 2015) and also experiencing the highest rates of biodiversity changes under the combined effects of climate change and direct human impacts (Blowes et al., 2019). Freshwater ecosystems are even more at risk, with fishes being among the most threatened vertebrates due to habitat degradation or exotic species introduction (Collen et al., 2014). In this context, efficient non-invasive methods are urgently needed to accurately monitor aquatic biodiversity including rare, highly mobile and elusive species in order to set appropriate conservation management.

Among the many ways to survey aquatic biodiversity, eDNA metabarcoding has recently emerged as a promising approach, frequently surpassing traditional inventory methods in detectability potential (Boussarie et al., 2018; Carraro, Hartikainen, Jokela, Bertuzzo, & Rinaldo, 2018; Stat et al., 2019; Valentini et al., 2016). Exogenous DNA released by animals in the environment, through shed skin, mucus or feces, can be retrieved by filtering water and amplified via Polymerase Chain Reactive (PCR) using universal primers (Ficetola, Miaud, Pompanon, & Taberlet, 2008). High-throughput sequencing of the amplified DNA fragments provides a list of sequences over which corresponding species can be assigned by comparison with available genetic databases like the European Nucleotide Archive (ENA) (Dickie et al., 2018; Kanz et al., 2005).

However, the major limitation currently impairing the large-scale application of eDNA inventories is the incompleteness of species sequences available in public genetic databases, considerably reducing the breadth of detected biodiversity. Historically, eDNA studies have primarily focused on well-known species-poor freshwater systems (Jerde, Wilson, & Dressler, 2019) but recently, eDNA biodiversity inventories have spread all over the globe, across a wide range of ecosystems encompassing less-studied and more diverse taxa and habitats (Cilleros et al., 2019; Jerde et al., 2019; Yamamoto et al., 2017). A recent study on European aquatic systems shows that genetic coverage varies widely among taxonomic groups, databases and the level of monitoring (Weigand et al., 2019) with for example European freshwater fish lacking genetic coverage on the 12S mitochondrial marker for 64% of the 627 species.

Teleostean fishes represent the largest group of vertebrates with more than 32,000 species ("www.fishbase.org," n.d.) and a total biomass of 0.7 Gt (Bar-On, Phillips, & Milo, 2018). They represent the most extensively studied taxonomic group using eDNA with up to 60% of the publications on vertebrates (Tsuji, Takahara, Doi, Shibata, & Yamanaka, 2019). Fish indeed play a significant role in carbon cycling (Wilson et al., 2009) and food security (Hicks et al., 2019). Despite their cultural,

commercial and ecological importance, fish populations are increasingly depleted or threatened due to overfishing (Anticamara, Watson, Gelchu, & Pauly, 2011) and habitat alterations (Collen et al., 2014). Surprisingly, the extent to which genetic reference databases cover fish biodiversity for the most widely used metabarcoding primers is unknown, while it ultimately determines the amount and the composition of species potential revealed by eDNA surveys. This kind of information is currently available, albeit scattered across different databases, but we still lack a tool facilitating the assessment and visualization of genetic species coverage for a given region, a given taxa and a given primer.

Here, we filled this gap by developing a user-friendly, flexible and interactive web-interface linking reference genetic databases to regional species lists. Using regional freshwater and marine fish checklists, we assessed geographic variations in species diversity coverage vs. gap for different metabarcoding primers. Then, we highlighted the geographical bias in genetic coverage and disparities according to the native and conservation status of species (IUCN), providing valuable recommendations for future eDNA investigations at global scale.

Methods

1. Interactive web-interface: GAPeDNA

To facilitate the global assessment and visualization of regional gaps in genetic databases for environmental DNA metabarcoding, we developed a user-friendly interactive web-interface called GAPeDNA (https://shiny.cefe.cnrs.fr/GAPeDNA/, Fig 1), using the shiny R package (Chang, Cheng, Allaire, Xie, & McPherson, 2019). This interface allows researchers and stakeholders to easily locate gaps in the reference genetic databases at global scale for a selection of fish metabarcoding primers. A virtual PCR using the selected primers is performed on a selected online genetic database. The list of the amplified species is then compared to a spatialized checklist to generate the percentage of species referenced in each spatial unit or area (e.g. basins and ecoregions for freshwater and marine fishes, respectively) (Fig 1A). This percentage is then displayed with an interactive global map in GAPeDNA. This interface is flexible and can display results for several primer pairs per taxon, several spatial units, and allows the user to choose between several options (Fig 1B). We present the application for fish but users are encouraged to suggest new taxa which requires to have (i) at least one primer pair targeting the taxa using metabarcoding and (ii) globally georeferenced species checklists. It also allows to visualize which species are actually sequenced for a given primer when clicking on the area of interest, under which conservation status (i.e. IUCN category) these species are, and extract this information as a Comma Separated Values (CSV) file. Users can thus quickly grasp information regarding sequencing priorities depending on their research interest.

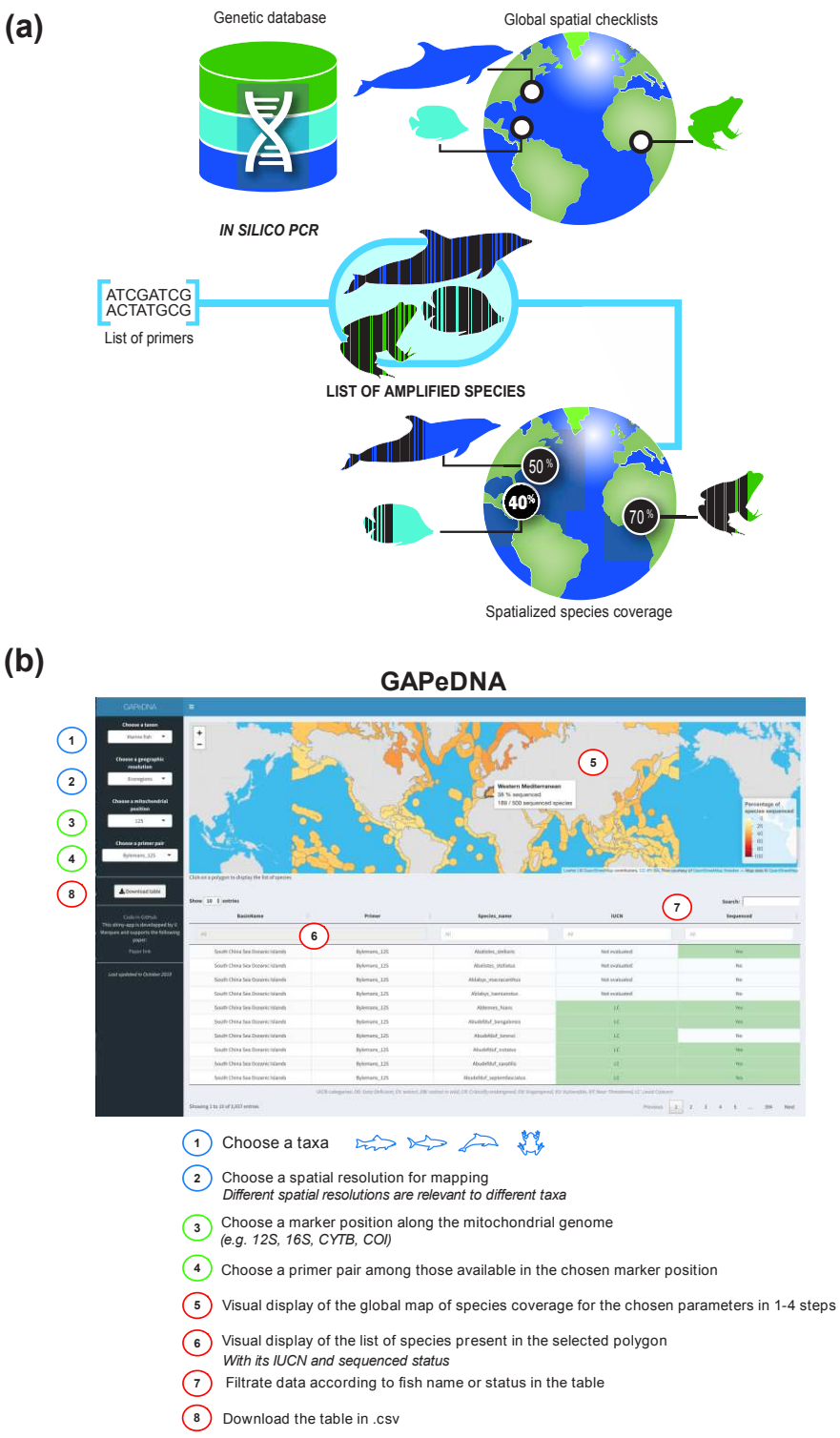


Fig. 1. Illustration of the process for generating map and data in the GAPeDNA web application (A) and details on the interface (B). User's spatial choices are in blue and green, genetic choices in green and visual displays are in red.

2. Genetic sequence database and genetic coverage by markers

To illustrate the distribution of species coverage, we used the European Nucleotide Archive (ENA) (Kanz et al., 2005) (release 138, downloaded in January 2019) as the genetic reference database for fish species. This database was formatted using obiconvert from the OBITOOLs toolkit (Boyer et al., 2016) to run in-silico PCRs (i.e. virtual PCR based on primer affinity to sequences). Yet, primer sequences need to be present within the sequence fragment deposited online to be detectable using this in silico approach.

An extensive literature search was conducted to identify the most commonly used primers targeting fish for metabarcoding on ISI Web of Science with the following keywords: “fish” AND “metabarcoding” AND “primer” AND “environmental DNA”. We discarded primer pairs not primarily targeting fish, only targeting a restricted group of fish or containing errors. Following this filtering, we retained 23 primer pairs from 18 papers (Supplementary, Table S1), from 5 regions in the mitochondrial genome (hereafter referred as markers): namely 12S, 16S, 18S, COI and CytB. All primer pairs were used individually to run in-silico PCRs using ecoPCR from OBITOOLS (Boyer et al., 2016), with 3 mismatches allowed. All species amplified by each primer were compared to the regional fish checklists of both marine and freshwater environments, to obtain the percentage of species coverage by spatial unit and by primer. Fish names obtained from GenBank were checked and updated using Fishbase as the sole reference. We further discarded 4 primer pairs with low performance (global fish coverage < 0.05 %) to avoid bias when comparing markers (Supplementary, Table S2), so we proceeded with a total of 19 primer pairs on 4 markers, as the only primer pair located on the 18S rDNA marker was discarded. The successful virtual amplification of a species by a primer pair is conditional to (i) species presence in the public genetic database and (ii) the primer ability to amplify the sequence. Hence, primer pairs lacking universality for fish sequence amplification show an overall low coverage, even if located on a genetic marker with a larger sequence coverage in online database, since they are unable to amplify those due to primer specificity.

3. Global species checklists and status

The checklist for freshwater fish was extracted from a global-scale database of fish diversity at the basin scale (Tedesco et al., 2017). The authors reviewed a large body of information from 1,436 distinct sources over 3,119 drainage basins, covering more than 80% of Earth surface and comprising 14,953 fish species, so 90% of all freshwater fishes recorded in Fishbase (www.fishbase.org). Although all biogeographic realms are well represented, some regional gaps remain in the database due to the scarcity of information or the probable low number of freshwater taxonomists in some regions like Southeast Asia. The global diversity of marine fishes was assembled using OBIS (Ocean Biogeographic Information System) and regional checklists (Albouy et al., 2019; Pellissier, Heine, Rosauer, & Albouy,

2018), including manual verification to remove taxonomic classification errors. It contains available occurrence data for all marine teleost and agnathan fishes, so a total of 14,202 species representing 82% of all marine fish species recorded in Fishbase. The original spatial resolution was a 1° grid for all marine environments. For visualization and interpretation purposes, this grid was then coerced at two supplementary biogeographic spatial scales according to Marine Ecoregions (Spalding et al., 2007) (i) at the province scale, with 62 distinct units and (ii) at the ecoregion scale, with 232 distinct units. Latitudes and longitudes were computed as the centroid of each polygon at the finest resolution for the both environments using the R package sf (Pebesma, 2016), and land areas were removed using polygons from Natural Earth data ("https://www.naturalearthdata.com/," n.d.). Areas were calculated using the Mollweide equal area projection and presented in figures using the Robinson projection.

For freshwater environments, a species is considered as non-indigenous in a given basin only if this species is able to complete its entire life cycle and harbors self-sustaining populations in that basin (Tedesco et al., 2017). A species is considered as indigenous when never occurring as non-indigenous in any basin following the original data (Tedesco et al., 2017). We acknowledge that some of the species classified as indigenous may have been introduced in another basin but have still not been identified, detected or been referred as such into global databases. However, our dataset represents currently the most recent and precise data on non-indigenous freshwater species at the global scale (Tedesco et al., 2017). For marine systems, we used the information supplied in Fishbase and only considered species flagged as 'introduced', excluding species categorized as 'questionable' or 'non-settled'.

Regarding the conservation status of species, we retrieved data from the redlist R package (Chamberlain, n.d.) to assign each species from both freshwater and marine environments into an IUCN red list category. The abbreviation 'DD' represents Data Deficient, 'LC' Least Concern and all threatened or near-threatened categories were grouped under the 'Threatened & NT' status. We excluded species identified as 'EX' for Extinct and 'EW' for Extinct in the Wild. Where no data was available, we assigned the value 'NA'.

Results

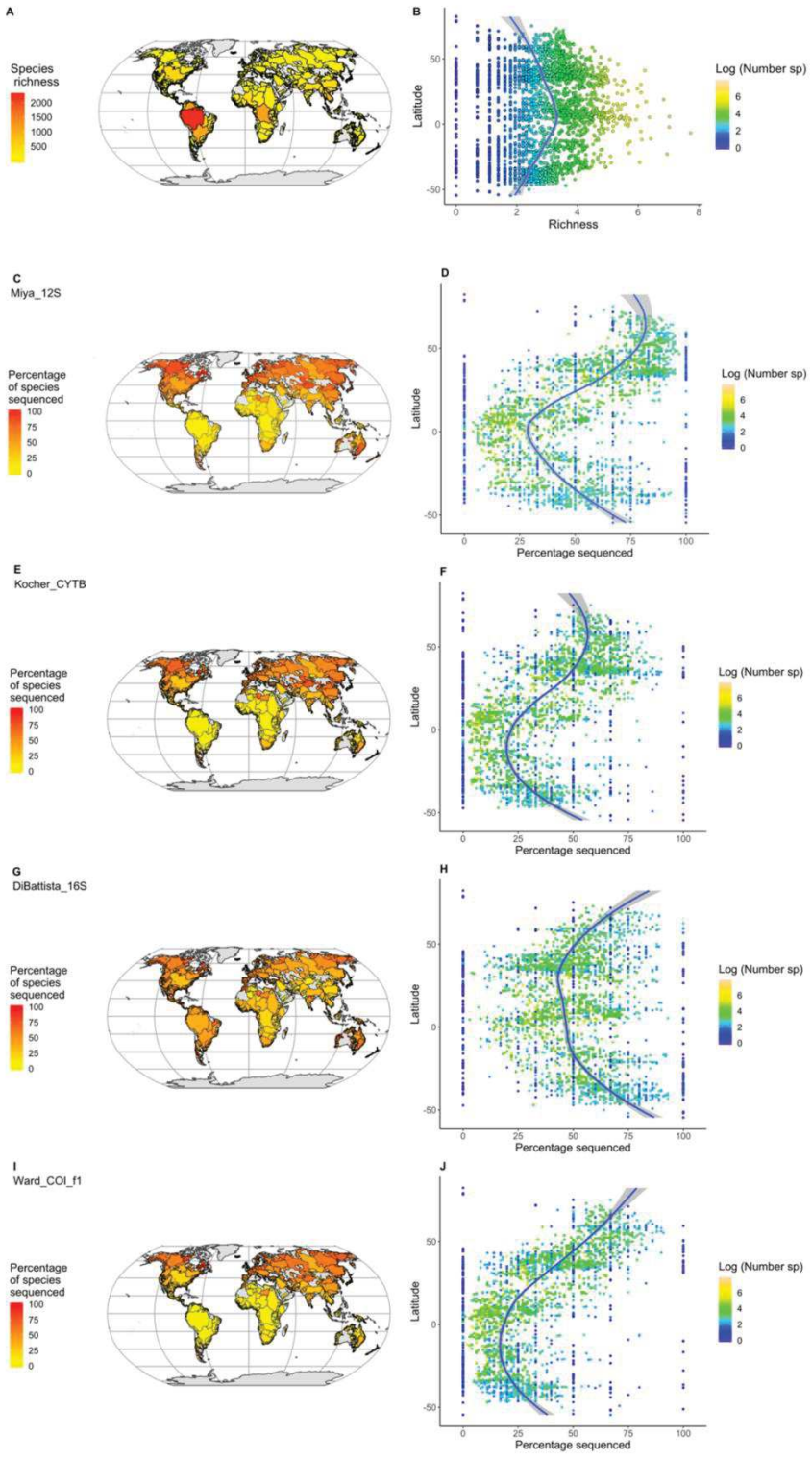


Fig. 2. Global and latitudinal distributions of freshwater fish species richness on log scale (A, B), coverage by online genetic database for the Miya primer targeting the 12S mitochondrial rDNA region (C, D), the Kocher primer targeting the Cytochrome B mitochondrial rDNA region (E, F), the DiBattista primer targeting the 16S rDNA region (G, H), and the Ward f2 primer targeting the COI mitochondrial region (I, J). The number of species along latitude (b) is log10 scaled and obtained from the finest resolution, here by basin. Global latitudinal patterns of all primer pairs are given in Supplementary, Fig S5 and S6, and the global distribution maps are reproducible and interactive using the web application. The displayed primers represent the most used per genetic marker in fish eDNA studies (Tsuji et al., 2019).

Global distribution of genetic database completeness and gaps

The 3,119 freshwater drainage basins, located across all continents (except the poles) largely varied in terms of surface, from 2 km^2 to 5,888,417 km^2 (Amazon) with a mean of 31,996 km^2 (SD = 20,9732 km^2). Their species richness ranged from 1 to 2,273 with a mean of 33 species (SD = 71), with an increasing number of species towards the equator following the classical latitudinal gradient (Figs 2A and 2B). Across the 232 marine ecoregions, species richness also greatly increased towards the equator, from 14 species (East Antarctic) to 3,937 species (South China Sea Oceanic Islands; Figs 4A and 4B). Marine ecoregion area varied from 19,000 km^2 (Puget Trough, Northern America) to 2,647,573 km^2 (Hawaii) with a mean of 588,862 km^2 (SD = 460,459 km^2) and no correlation between area and fish species richness was observed (Supplementary, Fig S1).

Global coverage of fish species in GenBank largely varied according to both the marker position along the mitochondrial genome and among primers for a given position (Figs 2 and 3), with a global coverage for freshwater species ranging between 7% for COI Ward and 26% for 16S McInnes, and a coverage for marine species between 4% for Thomsen Cytb cb and 30% for Shaw 16S (Supplementary, Table S2). For a given primer pair, species coverage also greatly varied along the latitudinal gradient, with a U-shaped relationship peaking in high absolute latitudes for most of the primers in freshwater systems. For example, the 16S McInnes primer had a mean coverage of 89% between 48° and 52° latitude (84 basins) and only 40% between -2° and 2° latitude (54 basins). This contrast was also marked for primers targeting the 12S mitochondrial rDNA region. For example, the 12S Miya primer pair covered 83% of the fish checklist in high latitudes (between 48° and 52°) but only 23% close to the equator (between -2° and 2° latitude, Fig. 2d). The Cytb from Thomsen 2cbl and 2deg (Supplementary, Figs S5 and S6) covered respectively 13% and 18% of the fish checklists, but showed no geographical gradient.

In marine ecosystems, the latitudinal gradient in species coverage was less pronounced with several primer pairs showing a steady decrease in coverage with decreasing latitude (Fig 3). Tropical fish assemblages along the equator were less sequenced than northern temperate assemblages, but were generally more sequenced than in negative latitude ecoregions towards the South Pole, as opposed to freshwater systems. Only the 12S Bylemans primer, covering 13% of marine fishes, showed no geographic pattern (Supplementary, Fig S6).

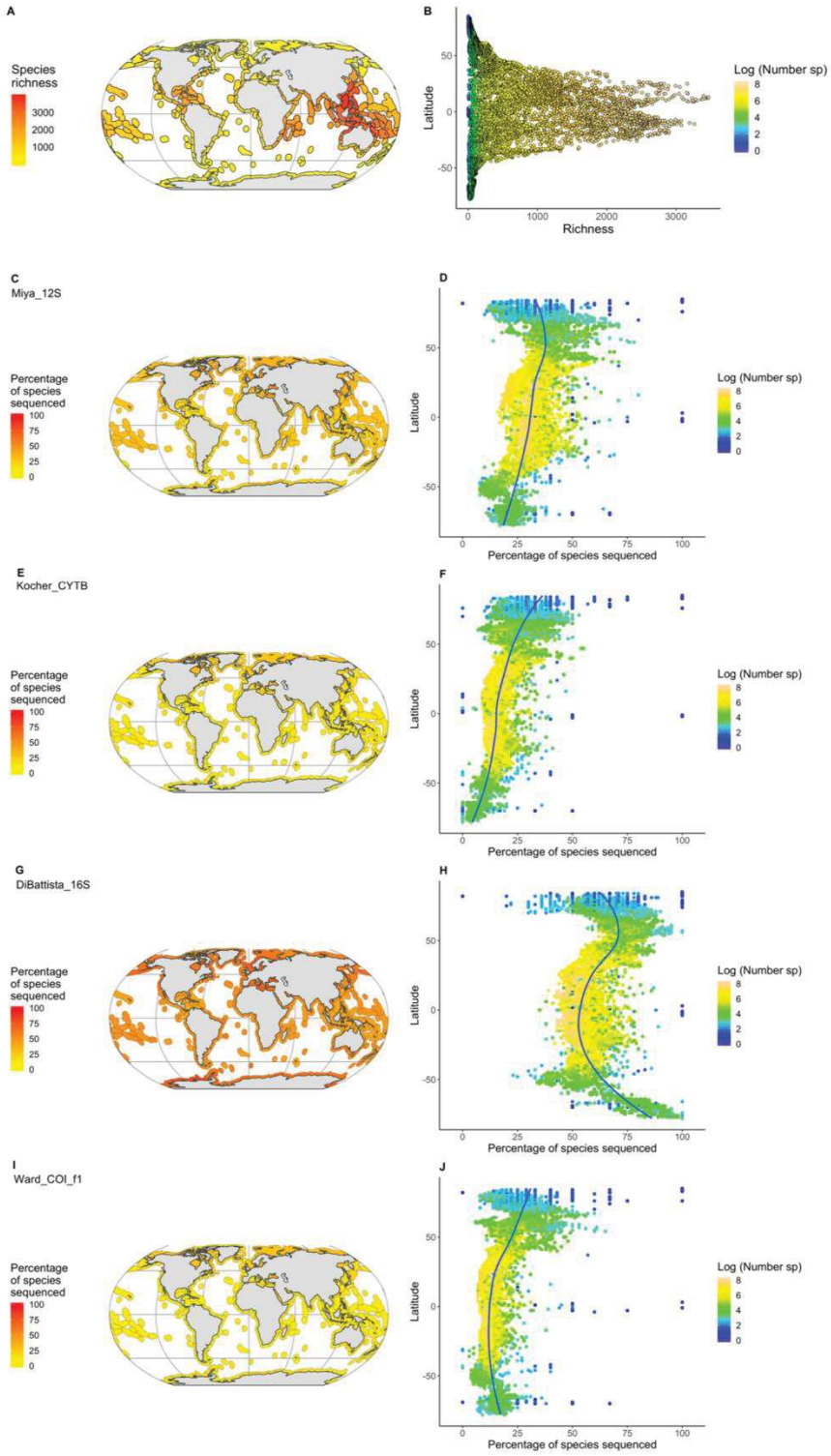


Fig. 3. Global and latitudinal distributions of marine fish species richness (A, B), coverage of online genetic database for the Miya primer targeting the 12S mitochondrial rDNA region (C, D), the Kocher primer targeting the Cytochrome B mitochondrial rDNA region (E, F), the DiBattista primer targeting the 16S rDNA region (G, H), and the Ward f2 primer targeting the COI mitochondrial region (I, J). The number of species along latitude (B) is log scaled and obtained from the finest resolution, here by a 1° grid. Global latitudinal patterns of all primer pairs are given in Supplementary, Fig S5 and S6, and the global distribution maps are reproducible and interactive using the web application (https://shiny.cefe.cnrs.fr/GAPeDNA/). The displayed primers represent the most used per genetic marker in fish eDNA studies (Tsuji et al., 2019).

Genetic coverage of native vs. non-indigenous species

Environmental DNA can be used to track non-indigenous species in ecosystems. However, only the primers located on the 12S and 16S had a mean species coverage superior to 50% for all 605 identified non-indigenous freshwater fishes (Fig 4A). For the primers on the COI and Cytb, less than half of all non-indigenous fishes were amplified and sequenced. Only two primers, both on the 16S, had a coverage for more than 60% of non-indigenous species, while none had a coverage above 57% for the 12S primers. However, these species still had an overall larger coverage in databases compared to native species, the maximum for native species being 31% for a 16S marker and 15% or 19% for 12S and Cytb markers, respectively.

For the marine fishes, we identified 196 species as non-indigenous in at least one region of the marine realm, two times less than the 605 species identified in freshwater. However, global patterns of coverage were similar (Fig 4B), albeit with a wider coverage of marine non-indigenous species compared to their freshwater counterparts (maximum 12S coverage of 69% *vs* 57%). Overall, for both categories, non-indigenous species were more sequenced than indigenous species, but 20% to 80% of fish species remain to be sequenced depending on the genetic marker.

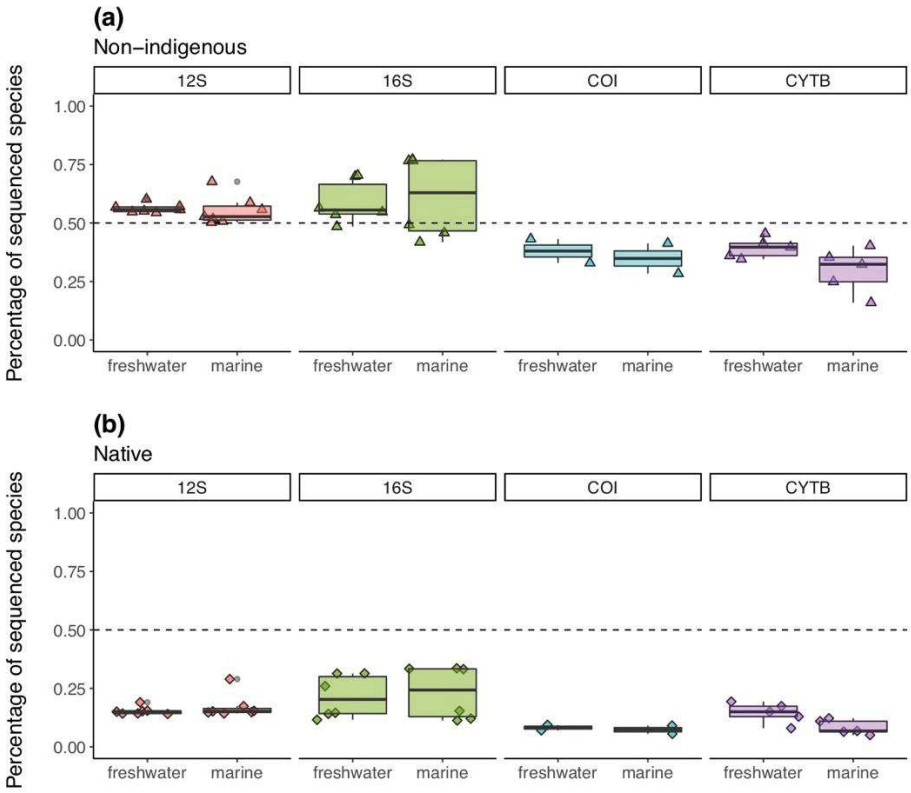


Fig. 4. Percentage of species coverage (A) in marine systems for non-indigenous (196) and native species (12,290) and (B) in freshwater for non-indigenous (605) and native species (14,348) depending on the marker position. Each triangle represents a primer pair.

Genetic coverage of fish species with different IUCN conservation status

Most of freshwater fish species were not evaluated (Not Applicable status, NA, 45.9% of total) or Least Concern (LC, 33.2%). However, 1,758 species (11.7%) were classified as threatened by including Vulnerable (VU), ENdangered (EN) and CRitically endangered (CR) species) or Near Threatened (NT) categories of the IUCN Red list (www.iucnredlist.org, Supplementary, Fig S2). The genetic database coverage of fish species according to their IUCN status showed consistent patterns for all markers (Figs 5A and 5C). Species classified as Least Concern (LC) were always more represented in genetics databases compared to non-evaluated (NA), data deficient species (DD) (Supplementary, Fig S3) or threatened species (T & NT). Freshwater basins where the most threatened species remain to be sequenced using the 12S Miya primer were mainly located around the equator with 79 species in the Congo basin and 63 species in the Mekong basin or in the Northern hemisphere with a maximum of 72 species in the Mississippi basin (Fig 5B). These basins also host the highest number of threatened species, independently of reference filling (Supplementary, Fig S4).

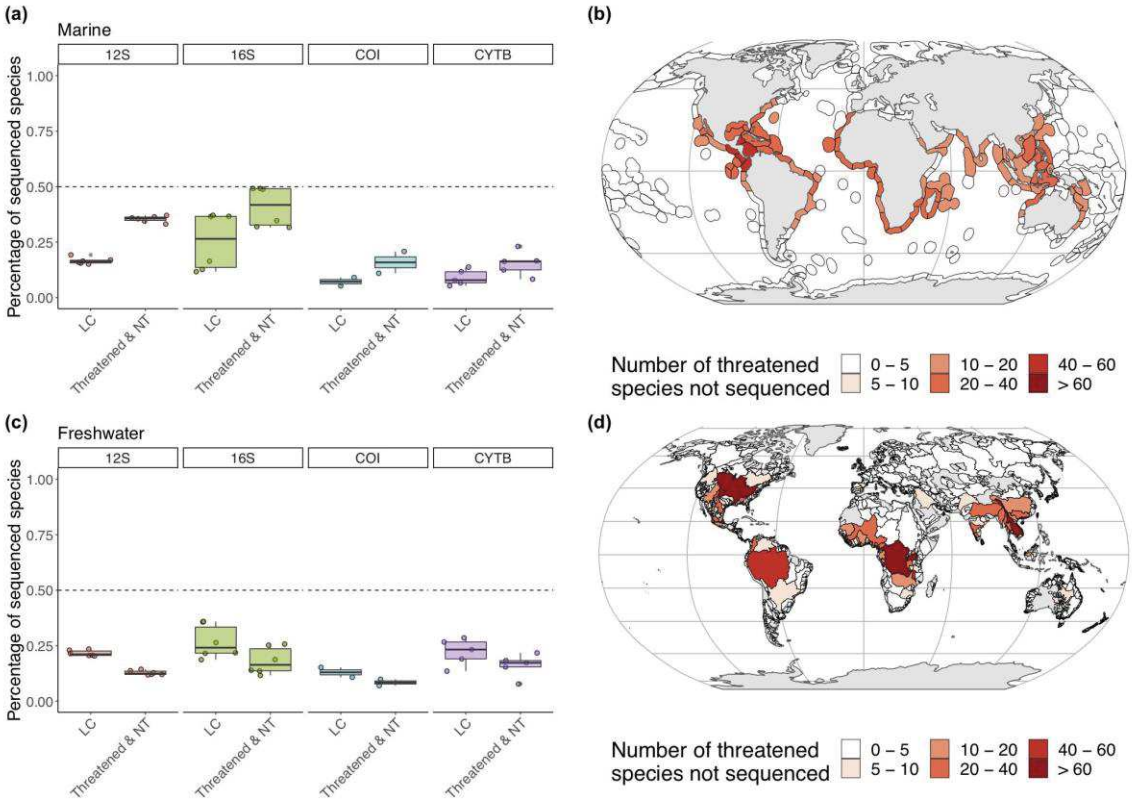


Fig. 5. Percentage of coverage according to two IUCN categories: Least Concern (LC) or Threatened and Near-Threated (NT) for all primer pairs and global gap in threatened species not sequenced illustrated for the Miya 12S primer in (A, B) marine systems and (C, D) freshwater systems. Each dot represents one primer pair. The threatened category includes the categories Vulnerable (VU), Endangered (E) and CRitically endangered (CR). The categories Not Applicable (NA) and Data Deficient (DD) were not represented. All the categories are displayed on Supplementary, Fig S3.

In marine environments, 3.5% of all species were classified under an IUCN red list status compared to 11.7% in freshwater systems (Supplementary, Fig S2), and around the same proportion of fishes were unevaluated or data deficient (49% *vs.* 55% for freshwater). Genetic coverage was systematically higher for threatened species compared to Least Concern (LC) species, albeit never exceeding 50% for any primer or ecoregion (Fig 5). Species listed as LC consistently had a higher coverage than unevaluated or data deficient species (Supplementary, Fig S3). Marine ecoregions hosting the most threatened species remaining to be sequenced using the 12S Miya primer were also located around the equator, particularly in the Caribbean with a maximum of 42 species in the Southwestern Caribbean ecoregion or in the Eastern Coast of Africa with a maximum of 32 species in the Delagoa ecoregion (Fig 5D).

Discussion

Genetic markers and primers selection

eDNA is currently limited by the scarcity of species sequences available in online public genetic databases. We provide here a spatialized global assessment of fish sequence coverage and gaps in databases, using published eDNA primers and displayed on an online, semi-automated and flexible application called GAPeDNA. Our study considers all existing markers and most primers capable of theoretically amplifying fish species by in-silico PCR, regardless of their performance, avoiding a bias in the choice of a genetic marker or primer. The marker and primer selection must be motivated by their efficiency to detect the targeted taxa owing to their specificity and sensitivity. A general consensus is emerging in fish eDNA studies toward the use of 12S primers (Collins et al., 2019; Weigand et al., 2019). Primers located on the 12S mitochondrial region have been recognized as the best to specifically amplify fishes, unlike COI primers which lack specificity, resulting in low fish detectability (Valentini et al., 2016). Unfortunately, we show that the 12S still has a very low species completeness in genetic databases, with strong spatial disparities. With the goal to sequence a maximum of species, it is crucial to reach a consensus in the genetic marker selection to join efforts towards a globally coordinated sampling strategy for this genetic marker. Once species gaps in the 12S sequences will be almost filled, it would pave the way to install eDNA metabarcoding as a robust and standard monitoring and inventory tool, capable of fish identification to the species level in every location.

Mapping species coverage gaps to improve eDNA monitoring

The global diversity of both freshwater and marine fishes is not well covered in public genetic databases. Globally, we show a higher coverage around high latitude in the northern hemisphere consistent across

the genetic markers and primers while tropical areas, which host more species, have more species gaps in public sequence databases (Figs 2 and 3). For freshwater fishes, the genetic species coverage exhibits a clear U-shaped pattern for almost all markers along the latitudinal diversity gradient (Hillebrand, 2004) (Fig 2), with a minimum percentage of sequenced species around the equator. For marine fishes, species coverage declines with declining latitude, and the minimum percentage of species sequenced is around the low latitudes of the southern hemisphere where marine fish diversity is the lowest.

The location of the gaps may drive the future sampling efforts required to fill them. Tropical environments are under-represented in public sequence databases and will require a costly, time consuming and globally coordinated efforts to both describe and sequence the numerous species left to be discovered, as well as sequence the numerous species already described (Juhel et al., 2020; Pinheiro, Moreau, Daly, & Rocha, 2019). Environmental DNA is settling as an efficient inventory tool that can overcome hurdles encountered when sampling in tropical ecosystems. In many large water bodies, such as the Mekong or the Amazon, water turbidity prevents visual census leaving the eDNA the only non-invasive monitoring method (Cilleros et al., 2019; Yamamoto et al., 2017). The need to fill species gaps is urgent in these environments since they are experiencing major turnover in species identities with unknown consequences on ecosystem functioning and resilience (Magurran et al., 2018).

Tropical marine ecosystems are biodiversity hotspots, particularly the Coral Triangle (Barlow et al., 2018; Myers, Mittermeier, Mittermeier, Da Fonseca, & Kent, 2000). Tropical countries also tend to have a high dependency to fish resources (Andrello et al., 2017; Barange et al., 2014), stressing the importance of securing a sustainable exploitation of fish which requires monitoring assessments and correct evaluations of biodiversity since these both aspects are intimately linked (Duffy, Lefcheck, Stuart-Smith, Navarrete, & Edgar, 2016; Lefcheck et al., 2019). For instance, crypto-benthic fishes (<5cm) have been recently shown to contribute massively to coral reef functioning (Brandl 2019), particularly by feeding fish consumed by humans, but they are still poorly inventoried. Tropical countries are also projected to undergo among the most severe environmental impacts related to human population expansion and climate change (Barlow et al., 2018), highlighting the importance of conducting ecological studies and setting appropriate conservation programs. For instance, mesophotic reefs (30-150 meters depth) are still poorly known while they potentially host very different species assemblages that can be also affected by climate change (Lesser, Slattery, Laverick, Macartney, & Bridge, 2019; Rocha, Pinheiro, Shepherd, & Papastamatiou, 2018). Their exploration will require new eDNA-based protocols (fish sampling for reference database and water filtering) that must complement visual surveys that remain limited at this depth. Yet, there is a clear publication bias with the most diverse ecosystems being the least studied in ecology (Hickisch, Hodgetts, Macdonald, & Tockner, 2019). So, the efforts to achieve genetic database completeness are massive but necessary in such highly diverse environments in order

to tackle major conservation challenges like the protection of vulnerable but still poorly described biodiversity.

Environmental DNA metabarcoding to monitor non-indigenous species

Among the numerous threats that all aquatic environments are currently facing lies non-indigenous species, which have the potential to disrupt entire ecosystems when declared as invasive (Albins & Hixon, 2013; Bax, Williamson, Aguero, Gonzalez, & Geeves, 2003; Clavero & García-Berthou, 2005). For example, the Nile perch (Lates nicotilus), introduced in the 1950s in the Lake Victoria, drove around half of the hundreds of native Cichlids fish species to extinction through predation and competition (McGee et al., 2015; Witte et al., 1992). As traditional methods struggle to detect those species at an early stage of installation, eDNA offers an important potential for early detection below the traditional detection threshold (Dougherty et al., 2016; Hunter et al., 2015). Yet, a successful detection of species introduction relies on database completeness for those species. We show that, even among fish species identified as non-indigenous in freshwater ecosystems, up to 30% are currently missing in the best curated 16S database (Fig 4). For the genetic marker 12S, a maximum of 55% of non-indigenous species are sequenced per basin, twice as much as native species. It was expected that more non-indigenous species would be genetically referenced compared to native ones since referencing species occurrence outside their native range necessarily assumes their observation, and a large proportion of introductions being intentional for recreational fishing (Leprieur, Beauchard, Blanchet, Oberdorff, & Brosse, 2008), making tissue for genetic sequencing easily available. We highlight here that despite a higher coverage for non-indigenous species (Fig 4), the potential of eDNA to detect invasion events and provide early warning signals is still limited while crucial for mitigating deleterious effects (Vander Zanden, Hansen, Higgins, & Kornis, 2010).

Sequencing threatened species to support their monitoring

Environmental DNA has a great potential in biodiversity conservation, addressing the constraints of detecting elusive or low-abundant species missed by traditional surveys. The proportions of threatened species estimated by the IUCN Red List (11% of freshwater and 3% of marine fishes) are likely underestimated since 48% of fish species are unevaluated while 7 to 9% are Data Deficient (Supplementary, Fig S2). Although the fate of Data Deficient species remains largely unexplored, they form the category with the least coverage in public genetic databases and are estimated to hide a large proportion of already threatened species (Bland, Collen, Orme, & Bielby, 2015). Even among threatened species, less than 50% have referenced sequences across all genetic markers, and surprisingly their coverage is lower than Least Concern species for freshwater fishes. This can be due to the high number of threatened freshwater fishes, mainly located in hard-to-explore tropical regions (Collen et al., 2014).

Most threatened freshwater fishes live in large tropical basins such as the Congo, the Mekong or the Amazon (Fig 5). However, the Mississippi basin, although located in a well-developed and science-leading country, the USA, where conservation measures and monitoring programs are well established, hosts 72 threatened species that are not sequenced for a 12S primer. So, efforts to complement genetic reference databases must be widespread and are not only related to the level of species richness or economic development, as often assumed.

Interactive online application to support eDNA metabarcoding studies

We developed the user-friendly web-app interface GAPeDNA to synthetize this large amount of information and make it easily accessible, even without any coding skills. It allows users to select a taxonomic group (at the moment, only freshwater and marine fish are available), the spatial unit or area, the genetic markers of interest and the corresponding primers to evaluate their global spatialized species coverage in public genetic databases, and have access to the corresponding list of species per spatial unit and status (IUCN). This permits the assessment of species remaining to be sequenced for a given spatial zone, and set priorities for sequencing. Although this study is focused on fish as an example, any new taxa can be added to GAPeDNA, providing necessary information are given: 1) primers suited for metabarcoding and 2) global spatialized species checklists. This can thus expand the reach and potential of this tool within the metabarcoding scientific community and managers using eDNA for ecological surveys.

As the adoption of eDNA metabarcoding as a standard and robust monitoring approach worldwide depends on its ability to identify organisms at the species level, we hope that our tool and its potential as demonstrated by the fish example included here will encourage researchers, managers, foundations and institutions to work towards a join effort for a global sequencing effort targeting taxa of interest to enhance eDNA metabarcoding inventories.

References

Albins, M. A., & Hixon, M. A. (2013). Worst case scenario: Potential long-term effects of invasive predatory lionfish (Pterois volitans) on Atlantic and Caribbean coral-reef communities. Environmental Biology of Fishes, 96(10–11), 1151–1157. https://doi.org/10.1007/s10641-011-9795-1

Albouy, C., Archambault, P., Appeltans, W., Araújo, M. B., Beauchesne, D., Cazelles, K., … Gravel, D. (2019). The marine fish food web is globally connected. Nature Ecology & Evolution, 3(8), 1153–1161. https://doi.org/10.1038/s41559-019-0950-y

Andrello, M., Guilhaumon, F., Albouy, C., Parravicini, V., Scholtens, J., Verley, P., … Mouillot, D. (2017). Global mismatch between fishing dependency and larval supply from marine reserves. Nature Communications, 8(May), 16039. https://doi.org/10.1038/ncomms16039

Anticamara, J. A., Watson, R., Gelchu, A., & Pauly, D. (2011). Global fishing effort (1950-2010): Trends, gaps, and implications. Fisheries Research, 107(1–3), 131–136. https://doi.org/10.1016/j.fishres.2010.10.016

Bar-On, Y. M., Phillips, R., & Milo, R. (2018). The biomass distribution on Earth. Proceedings of the National Academy of Sciences of the United States of America, 115(25), 6506–6511. https://doi.org/10.1073/pnas.1711842115

Barange, M., Merino, G., Blanchard, J. L., Scholtens, J., Harle, J., Allison, E. H., … Jennings, S. (2014). Impacts of climate change on marine ecosystem production in societies dependent on fisheries. Nature Climate Change, 4(3), 211–216. https://doi.org/10.1038/nclimate2119

Barlow, J., França, F., Gardner, T. A., Hicks, C. C., Lennox, G. D., Berenguer, E., … Graham, N. A. J. (2018). The future of hyperdiverse tropical ecosystems. Nature, 559(7715), 517–526. https://doi.org/10.1038/s41586-018-0301-1

Bax, N., Williamson, A., Aguero, M., Gonzalez, E., & Geeves, W. (2003). Marine invasive alien species: A threat to global biodiversity. Marine Policy, 27(4), 313–323. https://doi.org/10.1016/S0308-597X(03)00041-1

Bland, L. M., Collen, B., Orme, C. D. L., & Bielby, J. (2015). Predicting the conservation status of data-deficient species. Conservation Biology, 29(1), 250–259. https://doi.org/10.1111/cobi.12372

Blowes, S. A., Supp, S. R., Antão, L. H., Bates, A., Bruelheide, H., Chase, J. M., … Dornelas, M. (2019). The geography of biodiversity change in marine and terrestrial assemblages. Science, 366(6463), 339–345. https://doi.org/10.1126/science.aaw1620

Boussarie, G., Bakker, J., Wangensteen, O. S., Mariani, S., Bonnin, L., Juhel, J. B., … Mouillot, D. (2018). Environmental DNA illuminates the dark diversity of sharks. Science Advances, 4(5), eaap9661. https://doi.org/10.1126/sciadv.aap9661

Boyer, F., Mercier, C., Bonin, A., Bras, Y. Le, Taberlet, P., & Coissac, E. (2016). OBITOOLS: a UNIX-inspired software package for DNA metabarcoding. Molecular Ecology Resources, 16(4), 176–182. https://doi.org/10.1111/1755-0998.12428

Carraro, L., Hartikainen, H., Jokela, J., Bertuzzo, E., & Rinaldo, A. (2018). Estimating species distribution and abundance in river networks using environmental DNA. Proceedings of the National Academy of Sciences of the United States of America, 115(46), 11724–11729. https://doi.org/10.1073/pnas.1813843115

Chamberlain, S. (n.d.). rredlist: "IUCN" Red List Client. R package version 0.5.0.

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2019). shiny: Web Application Framework for R. R package version 1.3.2.

Cilleros, K., Valentini, A., Allard, L., Dejean, T., Etienne, R., Grenouillet, G., … Brosse, S. (2019). Unlocking biodiversity and conservation studies in high-diversity environments using environmental DNA (eDNA): A test with Guianese freshwater fishes. Molecular Ecology Resources, 19(1), 27–46. https://doi.org/10.1111/1755-0998.12900

Cinner, J. E., Huchery, C., MacNeil, M. A., Graham, N. A. J., McClanahan, T. R., Maina, J., … Mouillot, D. (2016). Bright spots among the world's coral reefs. Nature, 535(7612), 416–419. https://doi.org/10.1038/nature18607

Clavero, M., & García-Berthou, E. (2005). Invasive species are a leading cause of animal extinctions. Trends in Ecology and Evolution, 20(3), 110. https://doi.org/10.1016/j.tree.2005.01.003

Collen, B., Whitton, F., Dyer, E. E., Baillie, J. E. M., Cumberlidge, N., Darwall, W. R. T., … Böhm, M. (2014). Global patterns of freshwater species diversity, threat and endemism. Global Ecology and Biogeography, 23(1), 40–51. https://doi.org/10.1111/geb.12096

Collins, R. A., Bakker, J., Wangensteen, O. S., Soto, A. Z., Corrigan, L., Sims, D. W., … Mariani, S. (2019). Non-specific amplification compromises environmental DNA metabarcoding with COI. Methods in Ecology and Evolution, 10(11), 1985–2001. https://doi.org/10.1111/2041-210X.1

Dickie, I. A., Boyer, S., Buckley, H. L., Duncan, R. P., Gardner, P. P., Hogg, I. D., … Weaver, L. (2018). Towards robust and repeatable sampling methods in eDNA-based studies. Molecular Ecology Resources, 18(5), 940–952. https://doi.org/10.1111/1755-0998.12907

Dougherty, M. M., Larson, E. R., Renshaw, M. A., Gantz, C. A., Egan, S. P., Erickson, D. M., & Lodge, D. M. (2016). Environmental DNA (eDNA) detects the invasive rusty crayfish Orconectes rusticus at low abundances. Journal of Applied Ecology, 53(3), 722–732. https://doi.org/10.1111/1365-2664.12621

Duffy, J. E., Lefcheck, J. S., Stuart-Smith, R. D., Navarrete, S. A., & Edgar, G. J. (2016). Biodiversity enhances reef fish biomass and resistance to climate change. Proceedings of the National Academy of Sciences of the United States of America, 113(22), 6230–6235. https://doi.org/10.1073/pnas.1524465113

Ficetola, G. F., Miaud, C., Pompanon, F., & Taberlet, P. (2008). Species detection using environmental DNA from water samples. Biology Letters, 4(4), 423–425. https://doi.org/10.1098/rsbl.2008.0118

Hickisch, R., Hodgetts, T., Macdonald, D. W., & Tockner, K. (2019). Effects of publication bias on conservation planning, 33(5), 1151–1163. https://doi.org/10.1111/cobi.13326

Hicks, C. C., Cohen, P. J., Graham, N. A. J., Nash, K. L., Allison, E. H., D'Lima, C., … MacNeil, M. A. (2019). Harnessing global fisheries to tackle micronutrient deficiencies. Nature, 574(7776), 95–98. https://doi.org/10.1038/s41586-019-1592-6

Hillebrand, H. (2004). On the Generality of the Latitudinal Diversity Gradient. The American Naturalist, 163(2), 192–211. https://doi.org/10.1086/381004

https://www.naturalearthdata.com/. (n.d.).

Hunter, M. E., Oyler-McCance, S. J., Dorazio, R. M., Fike, J. A., Smith, B. J., Hunter, C. T., … Hart, K. M. (2015). Environmental DNA (eDNA) sampling improves occurrence and detection estimates of invasive Burmese pythons. PLoS ONE, 10(4), 1–17. https://doi.org/10.1371/journal.pone.0121655

Jerde, C. L., Wilson, E. A., & Dressler, T. L. (2019). Measuring global fish species richness with eDNA metabarcoding. Molecular Ecology Resources, 19(1), 19–22. https://doi.org/10.1111/1755-0998.12929

Juhel, J., Utama, R. S., Marques, V., Vimono, I. B., Sugeha, H. Y., Kadarusman, … Hocdé, R. (2020). Accumulation curves of environmental DNA sequences predict coastal fish diversity in the coral triangle. Proceedings of the Royal Society B, 20200248.

Kanz, C., Aldebert, P., Althorpe, N., Baker, W., Baldwin, A., Bates, K., … Apweiler, R. (2005). The EMBL Nucleotide Sequence Database. Nucleic Acids Research, 33(Database issue), D29-33. https://doi.org/10.1093/nar/gki098

Lefcheck, J. S., Brandl, S. J., Innes-Gold, A. A., Steneck, R. S., Torres, R. E., & Rasher, D. B. (2019). Response: Commentary: Tropical fish diversity enhances coral reef functioning across multiple scales. Frontiers in Ecology and Evolution, 7(AUG). https://doi.org/10.3389/fevo.2019.00303

Leprieur, F., Beauchard, O., Blanchet, S., Oberdorff, T., & Brosse, S. (2008). Fish invasions in the world's river systems: When natural processes are blurred by human activities. PLoS Biology, 6(2), 0404–0410. https://doi.org/10.1371/journal.pbio.0060028

Lesser, M. P., Slattery, M., Laverick, J. H., Macartney, K. J., & Bridge, T. C. (2019). Global community breaks at 60 m on mesophotic coral reefs. Global Ecology and Biogeography, 28(10), 1403–1416. https://doi.org/10.1111/geb.12940

Link, J. S., & Watson, R. A. (2019). Global ecosystem overfishing: Clear delineation within real limits to production. Science Advances, 5(6), 1–12. https://doi.org/10.1126/sciadv.aav0474

Magurran, A. E., Deacon, A. E., Moyes, F., Shimadzu, H., Dornelas, M., Phillip, D. A. T., & Ramnarine, I. W. (2018). Divergent biodiversity change within ecosystems. Proceedings of the National Academy of Sciences of the United States of America, 115(8), 1843–1847. https://doi.org/10.1073/pnas.1712594115

McCauley, D. J., Pinsky, M. L., Palumbi, S. R., Estes, J. a., Joyce, F. H., & Warner, R. R. (2015). Marine defaunation: Animal loss in the global ocean. Science, 347(6219), 247–254. https://doi.org/10.1126/science.1255641

McGee, M. D., Borstein, S. R., Neches, R. Y., Buescher, H. H., Seehausen, O., & Wainwright, P. C. (2015). A pharyngeal jaw evolutionary innovation facilitated extinction in Lake Victoria cichlids. Science, 350(6264), 1077–1079. https://doi.org/10.1126/science.aab0800

Myers, N., Mittermeier, R., Mittermeier, C., Da Fonseca, G., & Kent, J. (2000). Biodiversity hotspots for conservation priorities. Nature, 403(February), 853–858. Retrieved from www.nature.com

OBIS Ocean Biogeographic Information System. (n.d.). OBIS. Retrieved February 2, 2018, from https://obis.org/

Pebesma, E. (2016). GeoSPARQL (Perry and Herring, 2012), and open source libraries that empower the open source geospatial software landscape including GDAL (Warmerdam, 2008), GEOS (GEOS Development Team, 2017), and liblwgeom (a PostGIS component. The R Journal, 10(1), 439–446.

Pellissier, L., Heine, C., Rosauer, D. F., & Albouy, C. (2018). Are global hotspots of endemic richness shaped by plate tectonics? Biological Journal of the Linnean Society, 123(1), 247–261. https://doi.org/10.1093/biolinnean/blx125

Pinheiro, H. T., Moreau, S., Daly, M., & Rocha, L. A. (2019). Will DNA barcoding meet taxonomic needs?, 365(6456), 873–875.

Reid, A. J., Carlson, A. K., Creed, I. F., Eliason, E. J., Gell, P. A., Johnson, P. T. J., … Cooke, S. J. (2019). Emerging threats and persistent conservation challenges for freshwater biodiversity. Biological Reviews, 94(3), 849–873. https://doi.org/10.1111/brv.12480

Rocha, L. A., Pinheiro, H. T., Shepherd, B., & Papastamatiou, Y. P. (2018). Mesophotic coral ecosystems are threatened and ecologically distinct from shallow water reefs, 284(July), 281–284.

Spalding, M. D., Fox, H. E., Allen, G. R., Davidson, N., Ferdaña, Z. A., Finlayson, M., … Robertson, J. (2007). Marine Ecoregions of the World: A Bioregionalization of Coastal and Shelf Areas. BioScience, 57(7), 573. https://doi.org/10.1641/B570707

Stat, M., John, J., DiBattista, J. D., Newman, S. J., Bunce, M., & Harvey, E. S. (2019). Combined use of eDNA metabarcoding and video surveillance for the assessment of fish biodiversity. Conservation Biology, 33(1), 196–205. https://doi.org/10.1111/cobi.13183

Tedesco, P. A., Beauchard, O., Bigorne, R., Blanchet, S., Buisson, L., Conti, L., … Oberdorff, T. (2017). Data Descriptor: A global database on freshwater fish species occurrence in drainage basins. Scientific Data, 4, 1–6. https://doi.org/10.1038/sdata.2017.141

Tsuji, S., Takahara, T., Doi, H., Shibata, N., & Yamanaka, H. (2019). The detection of aquatic macroorganisms using environmental DNA analysis-A review of methods for collection, extraction, and detection. Environmental DNA, (April), 1–10. https://doi.org/10.1002/edn3.21

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., … Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. Molecular Ecology, 25(4), 929–942. https://doi.org/10.1111/mec.13428

Vander Zanden, M. J., Hansen, G. J. A., Higgins, S. N., & Kornis, M. S. (2010). A pound of prevention, plus a pound of cure: Early detection and eradication of invasive species in the Laurentian Great Lakes. Journal of Great Lakes Research, 36(1), 199–205. https://doi.org/10.1016/j.jglr.2009.11.002

Weigand, H., Beermann, A. J., Čiampor, F., Costa, F. O., Csabai, Z., Duarte, S., … Ekrem, T. (2019). DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. Science of the Total Environment, 678, 499–524. https://doi.org/10.1016/j.scitotenv.2019.04.247

Wilson, R. W., Millero, F. J., Taylor, J. R., Walsh, P. J., Christensen, V., Jennings, S., & Grosell, M. (2009). Contribution of Fish to the Marine Inorganic Carbon Cycle R. Science, 323(January), 359–362.

Witte, F., Goldschmidt, T., Wanink, J., van Oijen, M., Goudswaard, K., Witte-Maas, E., & Bouton, N. (1992). The destruction of an endemic species flock: quantitative data on the decline of the haplochromine cichlids of Lake Victoria. Environmental Biology of Fishes, 34(1), 1–28. https://doi.org/10.1007/BF00004782

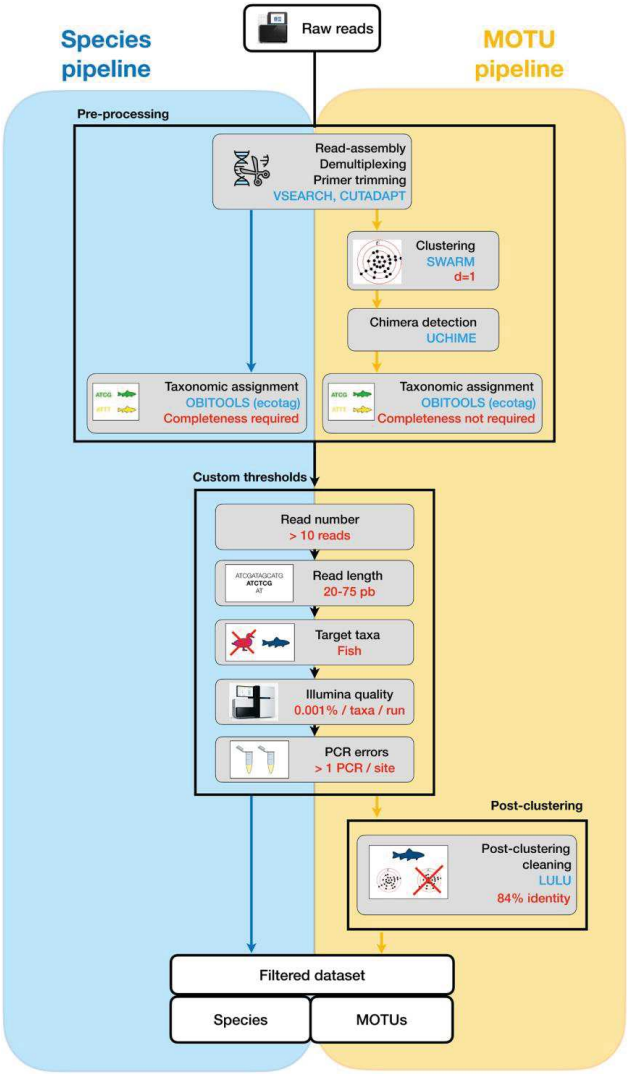
www.fishbase.org. (n.d.). Retrieved May 2, 2017, from http://www.fishbase.org/search.php

Yamamoto, S., Masuda, R., Sato, Y., Sado, T., Araki, H., Kondoh, M., … Miya, M. (2017). Environmental DNA metabarcoding reveals local fish communities in a species-rich coastal sea. Scientific Reports, 7, 40368. https://doi.org/10.1038/srep40368

Chapitre 3 - Blind assessment of vertebrate taxonomic diversity across spatial scales by clustering environmental DNA metabarcoding sequences

Manuscrit B

Marques V., Guérin P-E., Rocle M., Valentini A., Manel S., Mouillot D., Dejean T. Blind assessment of vertebrate taxonomic diversity across spatial scales by clustering environmental DNA metabarcoding sequences. (2020). Ecography. doi: 10.1111/ecog.05049



1. Préface

Le chapitre précédent a montré que la couverture des bases de références poissons est faible sur la majorité des marqueurs génétiques. Sur le gène 12S, reconnu comme le plus approprié pour les études en metabarcoding (Zhang et al. 2020b), la couverture mondiale n'excède pas 15% (sauf pour le marqueur 12S MiFish qui atteint environ 28% avec sa base de données interne). Cette couverture globale cache de fortes disparités spatiales avec une couverture taxonomique des écosystèmes tropicaux (faibles latitudes) et polaires (hautes latitudes) généralement plus faible que celle des milieux tempérés.

Ces chiffres montrent l'ampleur du travail à réaliser avant d'atteindre une complétude totale des bases de référence, ce qui est peu réaliste à court terme. Il est toutefois indispensable d'exploiter les données issues du metabarcoding de l'ADNe dès maintenant, en allant plus loin que la quantification de clades issues d'assignations taxonomiques parcellaires. Les études en microbiologie, qui sont pionnières dans l'utilisation des technologies de séquençage à base d'échantillons environnementaux, travaillent majoritairement sans assignation en utilisant des unités taxonomiques moléculaires (MOTUs) en tant que substitut des espèces. Toutefois, la transposition des outils de microbiologie pour le traitement de données ADNe ciblant les métazoaires reste incertaine tant les différences entre les marqueurs génétiques, taxons et méthodes sont fortes. En particulier, il est important de montrer que les estimations de MOTUs sont réalistes et fiables par rapport à la diversité en espèces réellement présentes dans l'écosystème. En effet l'utilisation de MOTUs pourrait générer une forte surestimation de richesse due aux nombreuses erreurs générées au cours du traitement moléculaire des données.

Ce chapitre vise à (i) proposer un traitement bio-informatique utilisant un algorithme de clustering de séquences d'ADNe de poissons ainsi que des seuils de qualité générant des MOTUs et (ii) tester si cette diversité en MOTUs est fiable par rapport à la diversité réelle *in situ* à trois échelles spatiales : alpha (locale), gamma (régionale) et beta (inter-site).

Les résultats révèlent qu'un traitement bio-informatique combinant l'algorithme de clustering SWARM, celui de post-clustering LULU avec le marqueur 12S teleo, et des filtres de qualité sévères (minimum 10 lectures par MOTUs, présence dans un au moins 2 PCR) peut estimer la diversité en poissons du Rhône de façon fiable. Ces estimations de diversité en utilisant ces substituts moléculaires sont fiables à trois niveaux de partitionnement de la diversité : alpha, la richesse locale estimée en MOTUs est proche de la diversité réelle en espèces (R = 0.98), gamma, la richesse régionale où le

nombre de MOTUs est proche du nombre réel d'espèces (67 MOTUs pour 63 espèces présentes) mais également beta, la différence de richesse inter-site qui est correctement estimée (R = 0.98). Ce travail ouvre la voie à l'estimation de la diversité en espèces dans les écosystèmes à partir d'échantillons d'ADNe même en l'absence de base de référence complète.

2. Manuscrit B

ECOGRAPHY

Research

Blind assessment of vertebrate taxonomic diversity across spatial scales by clustering environmental DNA metabarcoding sequences

Virginie Marques, Pierre-Édouard Guérin, Mathieu Rocle, Alice Valentini, Stéphanie Manel, David Mouillot and Tony Dejean

V. Marques (https://orcid.org/0000-0002-5142-4191) ✉ (virginie.marques01@gmail.com) and D. Mouillot, MARBEC, Univ. de Montpellier, CNRS, Ifremer, IRD, Montpellier, France. – P.-É. Guérin, S. Manel and VM, CEFE, Univ. Montpellier, CNRS, EPHE-PSL Univ., IRD, Univ. Paul Valéry Montpellier, Montpellier, France. – M. Rocle, Compagnie Nationale du Rhône, Direction de l'Ingénierie, Lyon, France. – A. Valentini and T. Dejean, SPYGEN, Le Bourget-du-Lac, France.

Ecography
43: 1–12, 2020
doi: 10.1111/ecog.05049

Subject Editor: Simon Creer
Editor-in-Chief: Miguel Araújo
Accepted 15 July 2020

Human activities impact all ecosystems on Earth, which urges scientists to better understand biodiversity changes across temporal and spatial scales. Environmental DNA (eDNA) metabarcoding is a promising non-invasive method to assess species composition in a wide range of ecosystems. Yet, this method requires the completeness of a reference database, i.e. a list of DNA sequences attached to each species of the regional pool, which is rarely met. As an alternative, molecular operational taxonomic units (MOTUs) can be extracted as clusters of sequences. However, the extent to which the diversity of MOTUs can predict the diversity of species across spatial scales is unknown. Here, we used 196 samples along the Rhone river (France) for which the reference database is complete to assess whether a blind eDNA approach can reliably predict the ground-truth number of species at different spatial scales. Using the 12S rDNA teleo primer, we curated and clustered 60 million sequences into MOTUs using a new assembled bioinformatic pipeline. We show that stringent quality filters were necessary to remove artefact noise, notably MOTUs present in a single PCR replicate, which represented 55% of MOTUs (103). Post-clustering cleaning also removed 19 additional erroneous MOTUs and only discarded one truly present species. We then show that the diversity of retained fish MOTUs accurately predicted the local (α, $r = 0.98$) and regional (γ) ground-truth species diversity (67 MOTUs versus 63 species), but also the species dissimilarity between samples (β-diversity, $r = 0.98$). This work paves the way towards extending the use of eDNA metabarcoding in community ecology and biogeography despite major gaps in genetic reference databases.

Keywords: 12S primer, α-β-δ-diversity, clustering, metabarcoding, MOTUs, reference database

NORDIC SOCIETY OIKOS

www.ecography.org

Introduction

In the new era of the Anthropocene, most ecosystems are experiencing severe human impacts and environmental changes with major consequences on species diversity (McCauley et al. 2015, Hughes et al. 2017, Isbell et al. 2019). Nevertheless, the ongoing reorganization of biodiversity is still poorly quantified and understood (but see Blowes et al. 2019) for two major reasons. First, the losses or gains of species are scale dependent with complex results emerging at the local or regional spatial scale (Vellend et al. 2013, Dornelas et al. 2014). For instance, several studies show that local species diversity is on average constant over time (Dornelas et al. 2014, Magurran et al. 2018), even under human impacts, while other studies report alarming species losses regionally or globally (Galetti et al. 2014, Doherty et al. 2016, Finderup Nielsen et al. 2019). Thus, any biodiversity monitoring should be spatially explicit (McGill et al. 2015) with three major components 1) local or α-diversity for the number of species within a given site, 2) spatial variation or β-diversity in species composition among sites and 3) regional or γ-diversity for the number of species within a geographical area containing all sites (Whittaker 1972). Second, biases and gaps in biodiversity inventories prevent accurate and comparable assessments across space and time (Hortal et al. 2015). This is particularly problematic when species are rare, small, cryptic or elusive or when ecosystems are either species-rich like in the tropics or hardly accessible like the deep sea (Mora et al. 2008, Menegotto and Rangel 2018). Hence, there is an urgent need for standardized and accurate biodiversity monitoring methods across spatial scales allowing reliable inter-study comparisons.

The metabarcoding of environmental DNA (eDNA) has the potential to fill this gap as it has been shown to surpass most traditional methods in species detection for both terrestrial and aquatic ecosystems (Bohmann et al. 2014, Valentini et al. 2016, Ruppert et al. 2019, Sales et al. 2020). Indeed, all organisms shed cells containing DNA in their environment, as intra or extra-cellular material, and can be retrieved for up to a few days (Dejean et al. 2011, Collins et al. 2018, Harrison et al. 2019). Amplification and high-throughput eDNA sequencing followed by bioinformatic analyses produce a list of sequences with the ultimate goal to assess species composition in a given site. This bioinformatic step requires the completeness of a reference database, i.e. a list of sequences attached to each species in the regional pool, to accurately assign each eDNA sequence to a given species. Yet, reference databases are often incomplete (Weigand et al. 2019). An estimated 91% of eukaryotic species inhabiting the ocean are yet to be described (Mora et al. 2011a) while only 13% of all described Teleostean fish species are referenced in public reference databases like the European Nucleotide Archive (ENA) (Leinonen et al. 2011) for the 12S ribosomal DNA fragment amplified by the teleo primers (Valentini et al. 2016), limiting the extent of species diversity revealed by eDNA metabarcoding.

Currently, completing reference databases would require massive sampling and sequencing efforts since many species still remain undiscovered due to their intrinsic nature (rare, small or elusive) or their unexplored habitat (e.g. deep sea) (Menegotto and Rangel 2018). Moreover, polymerase chain reaction (PCR) and sequencing generate numerous errors, overestimating the true number of species by several orders of magnitude (Edgar and Flyvbjerg 2015, Flynn et al. 2015). Thus, accurate methods able to assess biodiversity without complete reference databases while considering PCR and sequencing errors are urgently needed.

The microbial field pioneered methodological advances to infer biological diversity without a complete reference by clustering similar sequences into molecular operational taxonomic units (MOTUs) (Huse et al. 2010). However, these approaches focus mainly on fungi or unicellular organisms where the concept of species remains challenging (Pawlowski et al. 2018, Lladó Fernández et al. 2019). Even if clustering-based analyses are increasingly used in eDNA studies targeting vertebrates (Andruszkiewicz et al. 2017, Bakker et al. 2017, Closek et al. 2019, Sales et al. 2019), using the diversity of MOTUs as a reliable proxy for species diversity has yet not been evaluated. For instance, Closek et al. (2019) reported a large overestimation with more than 1300 MOTUs for 92 fish taxa only in the Californian Current upwelling ecosystem. The extent to which the metabarcoding of vertebrate eDNA can provide a reliable blind estimation of species diversity across spatial scales is unknown.

Here we evaluate how clusters of vertebrate eDNA sequences can predict species diversity across spatial scales. More precisely, we quantify how MOTUs can accurately predict local (α) and regional (γ) species diversity but also composition species dissimilarity between samples (β-diversity). For this, we focused on teleost fishes which are highly vulnerable to anthropogenic threats (Mora et al. 2011b) and represent the main group of vertebrates studied with eDNA (Tsuji et al. 2019). First, we highlight the geographic and taxonomic gaps in the reference database for the 12S mtDNA fragment, which is known to perform well with the teleo primer (Collins et al. 2019) designed by Valentini et al. (2016). Then, we assemble a metabarcoding bioinformatic pipeline based on sequence clustering using SWARM (Mahé et al. 2015), post-clustering using LULU (Frøslev et al. 2017) and stringent quality filters to analyze eDNA sequences from 196 samples along 500 km of the Rhône river (France). From the composition of MOTUs in each sample, we estimate α-, β- and γ-diversity and compare them to their analogs obtained with ground-truth assignment of all sequences using the complete reference database without clustering. Finally, we discuss strengths and weaknesses of this approach based on eDNA sequence clustering to assess taxonomic diversity across spatial scales, even when lacking exhaustive reference databases.

Material and methods

Global taxonomic and spatial gap analysis for fish

Recent fish metabarcoding studies indicate that primers located on the 12S ribosomal rRNA locus (12S rDNA)

perform better (i.e. detect more species, with less bias and more specific amplification) than primers based on alternatives loci (Ribosomal locus 16S, the cytochrome c oxidase I gene (COI)) (Collins et al. 2019, Weigand et al. 2019). Although the COI gene and associated primers might cover a larger proportion of fish species in the reference database and have a higher interspecific variability, their lack of suitable conserved region complicates the definition of taxa-specific primers. COI primers exhibit a clear lack of consistency across replicates, have a low specificity leading to a low amplification of target organisms with often less than 5% of cleaned reads assigned to fish (Collins et al. 2019) resulting in a low detectability power (Deagle et al. 2014, Bylemans et al. 2018, Collins et al. 2019). Among the fish eDNA 12S markers, we selected the teleo marker (forward primer-ACACCGCCCGTCACTCT, reverse primer-CTTCCGGTACACTTACCATG) (Valentini et al. 2016) given its high ability to detect fish species even in highly diverse ecosystems (Civade et al. 2016, Valentini et al. 2016, Bylemans et al. 2018, Pont et al. 2018, Cantera et al. 2019, Cilleros et al. 2019).

We first assessed the global taxonomic coverage of the teleo primers by performing in silico PCR using ecoPCR (Boyer et al. 2016) on the entire public database ENA (Leinonen et al. 2011) (release 138, January 2019). To build our reference database, we allowed a maximum of three mismatches and compared the results with the complete fish taxonomy from FishBase (Froese and Pauly 2019). For the spatial analysis, we extracted freshwater fish checklists of all drainage basins from the most recent and comprehensive data at the global scale (Tedesco et al. 2017), covering about 80% of inland waters. We obtained marine checklists from OBIS (OBIS Ocean Biogeographic Information System) at 1° resolution (Albouy et al. 2019), and used them to estimate fish composition within marine ecoregions globally (Spalding et al. 2007).

eDNA sampling and sequencing

We downloaded the sequence data from a previous study by Pont et al. (2018). The complete dataset encompasses 196 eDNA samples collected along 500 km of the Rhone River (France, Supplementary material Appendix 1 Fig. A1), corresponding to 103 distinct sites with field replicates (between 1 and 4 samples per site) in 2016. Among those, the original study used only 118 samples corresponding to 59 sites, but all samples were collected and processed in parallel. For each sample, 30 l of freshwater water were filtered, extracted, amplified and sequenced (Pont et al. 2018).

Clustering methods

Accurately delineating 'true' biological sequences from PCR and sequencing noise has been an ongoing challenge since the emergence of next generation sequencing (NGS) technologies. Clustering sequences into molecular operational taxonomic units (MOTUs) or defining exact sequence variants (ESVs) as proxies for species is a common practice in the prokaryote microbial field but also to study unicellular eukaryotes or fungi (Huse et al. 2010, Schmidt et al. 2013, Zimmermann et al. 2015, Callahan et al. 2017) and more recently eDNA of vertebrates (Closek et al. 2019, Sales et al. 2019).

While clustering has been historically limited to the creation of MOTUs based on a fixed similarity threshold, usually 97% (Stackebrandt and Goebel 2008, Edgar 2018), it poorly generalizes across markers or biological models (Edgar and Flyvbjerg 2015, Mahé et al. 2015, Nguyen et al. 2015, Callahan et al. 2017). As an alternative, new methods generate either ESV like the divisive amplicon denoising algorithm (DADA2) (Callahan et al. 2016) or MOTUs from de novo clustering algorithms based on sequence distribution and abundance to correct errors, like SWARM (Mahé et al. 2015). SWARM is an agglomerative unsupervised de novo single-linkage-clustering algorithm, building networks to define MOTUs based on sequence proximity and relative abundance (Mahé et al. 2015). While a threshold-based algorithm simply groups sequences together according to a fixed value, SWARM forms chains linking sequences based on their similarity and analyses the pattern to optimally break the network and delineate MOTUs (Mahé et al. 2014, 2015). So, the 'true' sequence is expected to be the most abundant while less abundant but close sequences are considered as erroneous as they are more likely to accumulate errors. This process avoids the dependence on a fixed value, which is not recommended in eDNA metabarcoding with short barcodes where only one mismatch can imply a different species (Miya et al. 2015).

Pipelines workflow

We based our analysis on two different pipelines: one where each unique sequence is independently assigned to a given species (called the Species pipeline) and the other one which clusters sequences into MOTUs using the SWARM algorithm (called the MOTU pipeline). In the Species pipeline, a complete reference database is required to assign a taxa to each sequence. In the MOTU pipeline, each MOTU also requires a taxonomic assignment but the completeness of the reference database is not required, as a partially complete reference database is sufficient to exclude MOTUs representing non-specific amplification, in our case, all non-fish taxa.

First, pre-processing steps were common for both pipelines (Fig. 1). Reads were assembled using VSEARCH (Rognes et al. 2016), demultiplexed at the PCR replicate level and primers trimmed using CUTADAPT (Martin 2011) adapted from an existing metabarcoding pipeline (<https://github.com/frederic-mahe/swarm/wiki/Fred's-metabarcoding-pipeline>). No mismatches were allowed in tags for demultiplexing while sequences containing ambiguous nucleotides were discarded. Two additional steps were applied in the pre-processing for the MOTU pipeline. First unsupervised clustering was performed with SWARM, using a minimum distance of one nucleotide between each MOTU ($d = 1$), as one mismatch can separate two distinct species with

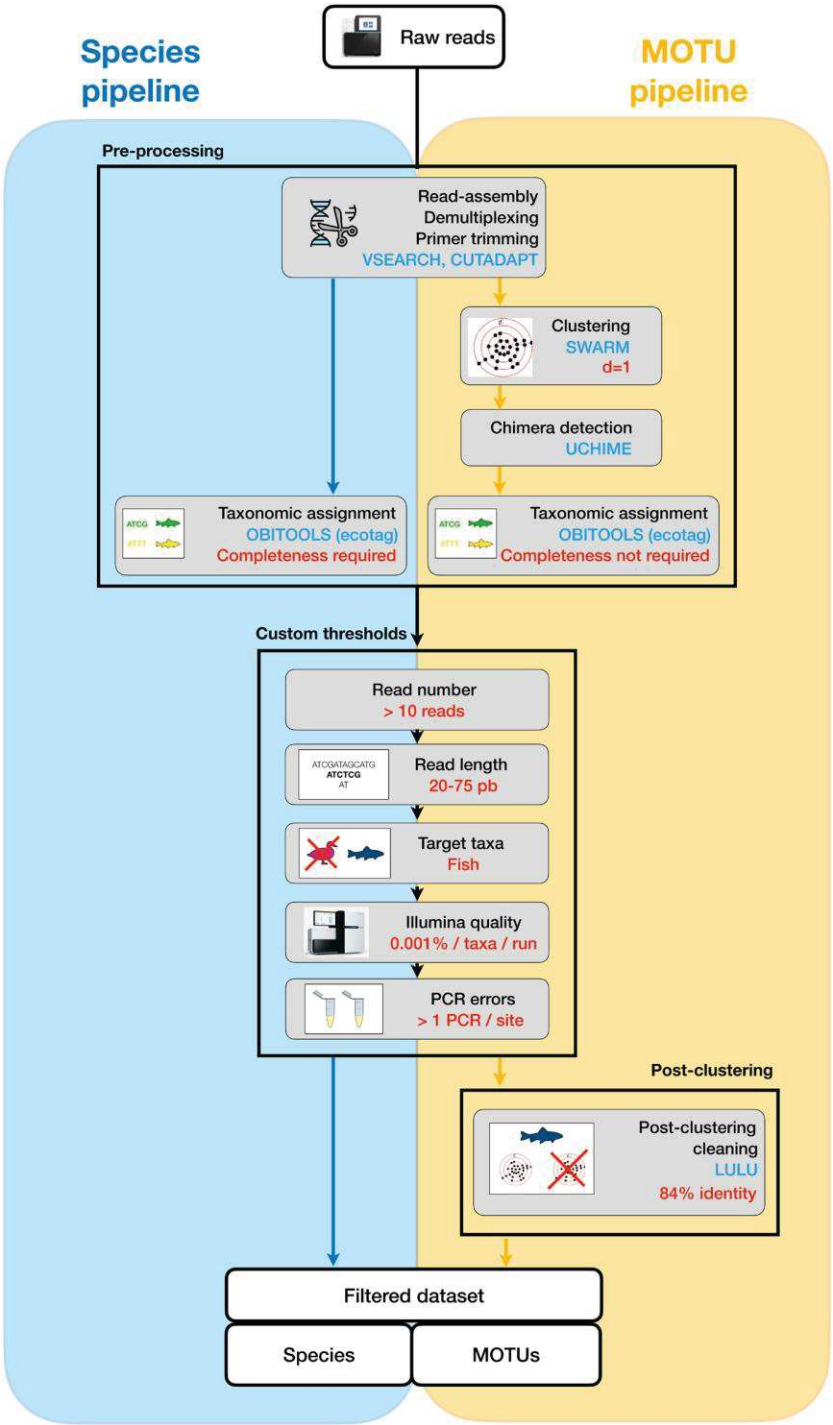


Figure 1. Illustration of the entire pipeline with three main steps: pre-processing, clustering, application of thresholds and post-clustering. Programs used are in blue and thresholds or requirements in red. Blue lines represent the classical alternative paths for the ground-truth method (Species pipeline), i.e. with the complete reference database and no clustering, whereas yellow lines represent the MOTU-based pipeline (MOTU pipeline), while black lines represent shared steps.

our primer. Taxonomic assignments of all unique sequences or MOTUs were then performed by ecotag, a lowest common ancestor (LCA) algorithm from the Obitools toolkit relying on the National Center for Biotechnology Information (NCBI) phylogeny tree (Boyer et al. 2016). Then, a set of custom and already published thresholds were applied on unique sequences for both the Species and MOTU pipelines (Fig. 1) (Valentini et al. 2016). All sequences or MOTUs with less than 10 reads, too short (< 20 bp), too long (> 75 bp) (Valentini et al. 2016) or not assigned to a fish phylum were discarded.

Each site usually has 2 samples as field replicates (except for 13 sites where the number of samples ranges from 1 to 4), and each sample has 12 PCR replicates, so most sites are represented by 24 individual PCRs (range: 12–48 PCRs replicates). For each site, we discarded all MOTUs or sequences present in only one PCR replicate (Civade et al. 2016). To avoid tag-jump noise (Schnell et al. 2015), all sequences with an abundance frequency of less than 0.001 per taxon/MOTU and per library were discarded. For the MOTU pipeline only, we then used the LULU algorithm (Frøslev et al. 2017) to clean MOTUs identified as erroneous based on sequence identity between MOTUs, abundances and patterns of co-occurrence. We used the blastn command line with the megablast algorithm to create the file matching all pairwise MOTUs to infer their similarity percentage. Then, to apply LULU, we used the 84% identity threshold (Frøslev et al. 2017) but also ran a sensitivity analysis with changes in the main parameters, i.e. the cross influence of identity threshold percentage and co-occurrence percentage (Supplementary material Appendix 1 Fig. A3).

Taxonomic assignments

For both pipelines, taxa assignments were performed on both our local database, exhaustive for resident species of the regional pool, and ENA (release 138, January 2019). For the Species pipeline, associating the local database with ENA (Leinonen et al. 2011) detected 24 extra species, among which 12 matches at 98% to our local database but at 100% in a public database to a foreign species (Supplementary material Appendix 1 Table A1). Those foreign species were unlikely to be present in the river, and most likely resulted from PCR or sequencing errors of local species randomly matching with foreign species. To avoid artificially inflating regional diversity from incorrect assignments, we only considered ENA assignments when our local database performed poorly (< 98% similarity). Among the 12 remaining species detected only by ENA and matching at < 98% to our local database, all were marine species from the Mediterranean Sea but 11 have records indicating a tolerance for brackish water while 6 were clearly known to enter estuaries (Supplementary material Appendix 1 Table A2). Most of those were also commonly consumed by humans, and DNA could have been transported into the river from sewage waters. Those extra species were hence kept for further analyses as they were unlikely to be errors generated at the PCR or sequencing step and they unlikely represent a methodological artefact.

Before analysis, assignments from ecotag were corrected to be more stringent as the algorithm can sometimes validate genus or family-level assignments to sequences with low similarity, which we chose to not trust blindly. This is due to the functioning of the ecotag algorithm (Boyer et al. 2016) and can happen in clades with a low species coverage in the reference database. We decided to add a level of standardization and only validate assignments at the species level for sequences matching at > 98% similarity, at 96–98% for the genus level, at 90–96% for the family level and at less than 90% similarity for the order or higher level for all sequences matching following a pilot study on phylogenetic signal for this marker (Supplementary material Appendix 1 Fig. A2).

Controlling taxonomic redundancy

When a sequence has a low percentage of similarity (< 98%), it can correspond to 1) a species absent from databases, 2) noise from PCR/sequencing errors from actual sequenced species or 3) rare but strong intra-specific variation at this non-coding locus which is prone to rapid mutations or insertions (Leinonen et al. 2011, Valentini et al. 2016). A common NGS metabarcoding issue is that for one species sequence matching at 100%, it can generate several noise variants matching at less than 100% (Frøslev et al. 2017). Hence, when counting the total number of taxa to infer the level of diversity, there is always a clear overestimation. For example, one *Salmo trutta* sequence with 100% similarity to a reference database would likely be accompanied by sequences matching at 97%, assigned at the *Salmo* genus and 95% assigned at the Salmonidae family. Where one species is present, the total taxa count can be three. To correct the number of taxa while being conservative, we created an estimated species count based on taxonomic correction for redundancy. A genus, family or order assignment can only be kept if there is no species already belonging to that rank, otherwise it would be more likely to be an error since the genetic databases are exhaustive for local resident species, the rest representing only a minority of rare sequences.

To evaluate the performance of LULU in the MOTU pipeline, we grouped taxa following this logic up to the family level. If a MOTU is assigned to a family for which a species representative is also detected, we assumed an error for this species and taxonomic redundancy. If a MOTU is assigned to an order only, it was not considered to represent an additional species.

Diversity comparison across scales

To assess the performance of our MOTU-based approach we calculated regional (γ) diversity, local (α) or sample diversity and dissimilarity between samples (β) with each pipeline. For the Species pipeline we retained all sequences matching at > 98% similarity cleaned for taxonomic redundancy to

count the number of distinct species. For the MOTU pipeline, we retained all MOTUs assigned to a fish taxa regardless of their similarity percentage. We used the software R ver. 3.6, where sample or α-diversity was computed as richness, i.e. plain species count. β diversity was computed using the Sorensen index, with the beta.temp and beta.multi functions from betapart package (Baselga and Orme 2012). A low value of dissimilarity between samples indicates similar communities, on a scale from 0 (identical) to 1 (totally dissimilar so no species or MOTU is common). We used the Mantel correlation test for pairwise sample comparisons.

Results

Global gaps in fish reference databases

Our analysis reveals that only 4243 out of 33 124 teleostean fish species (13%) are sequenced in the region amplified using the teleo primers, for both marine and freshwater environments (Fig. 2a). At higher taxonomic rank, we show that 38% of genera have at least one representative species sequenced for the 12S on the teleo fragment, this percentage reaching up to 80% for families. Next, we highlight a strong spatial heterogeneity between marine and freshwater environments but also among freshwater basins and marine ecoregions (Fig. 2b–c). For freshwater ecosystems, the proportion of fish species being referenced for the 12S fragment ranges from 0 to 100%, with tropical basins having an overall lower coverage than their temperate counterparts, except for Oceania where the proportion of sequenced species is among the highest. South America and Africa have by far the lowest coverage among all continents. For marine ecosystems, disparities are less pronounced but coverage varies between 10 and 53%. Ecoregions in Europe and Northern America have the highest coverage whereas tropical and southern ecoregions are the least covered.

γ-diversity assessment after filtering and clustering processes

In the 196 samples along the Rhone river, we obtain 60 689 053 reads of 299 225 distinct sequences with a mean of 309 617 reads per sample prior to any filtering (Table 1). First, we analyzed the eDNA metabarcoding data with the complete reference database (local database and ENA combined) with the Species pipeline (Fig. 1). We detect a total of 63 fish species (Table 1). Our new assembled MOTU pipeline applied on the same raw dataset identifies 67 MOTUs out of which 61 (91%) could be subsequently identified at the species level, i.e. matching at least at 98% of similarity with a species in the reference database.

We find that 98% of unique sequences and 96% of unique MOTUs correspond to either low abundant (< 10 reads) or non-fish species, so represent artefacts, noise or unspecific amplifications (Table 1), while only accounting for 12.5% and 4.4% of total reads, respectively. Sequence length filtering has a low influence, removing only 1 MOTU and no species. While removing only 0.004% of the total read count, our PCR filter removing all reads found in only one PCR replicate per site eliminates 45 MOTUs assigned to species (from 108 to 63) among which only 4 are possibly resident to the area (Supplementary material Appendix 1 Table A3). All other eliminated taxa are absent in the river and likely result from errors, contaminations from sewage waters or methodological artefacts. This PCR replicate filter also discards more than half of the detected MOTUs (86 out of 189, Table 1) representing mainly taxonomic redundancy and low-quality reads.

Following the PCR replicate filtering step, only 50 out of 86 MOTUs are represented by one taxon (Fig. 3), revealing either redundancy, with several MOTUs corresponding to the same taxa, or a lack of identification at the species level for the 36 remaining MOTUs. The application of LULU decreases the total number of MOTUs from 86 to 67 (Fig. 3). In particular, the number of taxa represented by more than 1 MOTU decreases from 15 (up to 6 MOTUs per taxa) to 8 after cleaning with LULU. Following this step, the lost MOTU representing a real taxa corresponds to a complex of two cyprinid fish species (*Ctenopharyngodon idella* and *Hypophthalmichthys molitrix*) for which teleo marker is not resolutive at the species level.

Finally, the regional pool (γ-diversity) of our fish Rhone dataset is comprised of 67 MOTUs among which 61 can be assigned to a species with 98% similarity while the ground-truth value is 63 fish species using the Species pipeline (Table 1).

Estimates of α and β species diversity using MOTUs

For each sample, we calculated the local (α-) diversity obtained by each pipeline so in terms of species and MOTUs. Overall the correlation between the number of MOTUs and the number of species is high and significant ($r = 0.98$; $p < 0.001$; Fig. 4a). The mean difference in local diversity across samples between the two pipelines is of 1.02 (SD = 1.5) with the MOTU-based approach underestimating true α-diversity. The maximum difference in local diversity is 5 (Fig. 4a), meaning that for one sample five less MOTUs are detected compared to the number of species identified with the reference database.

Since a similar value of α-diversity detected by the two pipelines does not necessary imply the same community composition, we performed a dissimilarity analysis (β-diversity) between samples pairs for both methods using the Sorensen index. We detect a high and significant correlation ($r = 0.98$, $p < 0.001$, Fig. 4b) between pairwise sample dissimilarity estimated with the Species and MOTU pipelines. We highlight no over or underestimation of dissimilarity by one pipeline compared to the other. Overall, the MOTU pipeline generates lower dissimilarity for 71% of pairs of samples compared to the Species pipeline but in 95% of all cases, the inferred level of dissimilarity has less than 0.1 difference between the two pipelines.

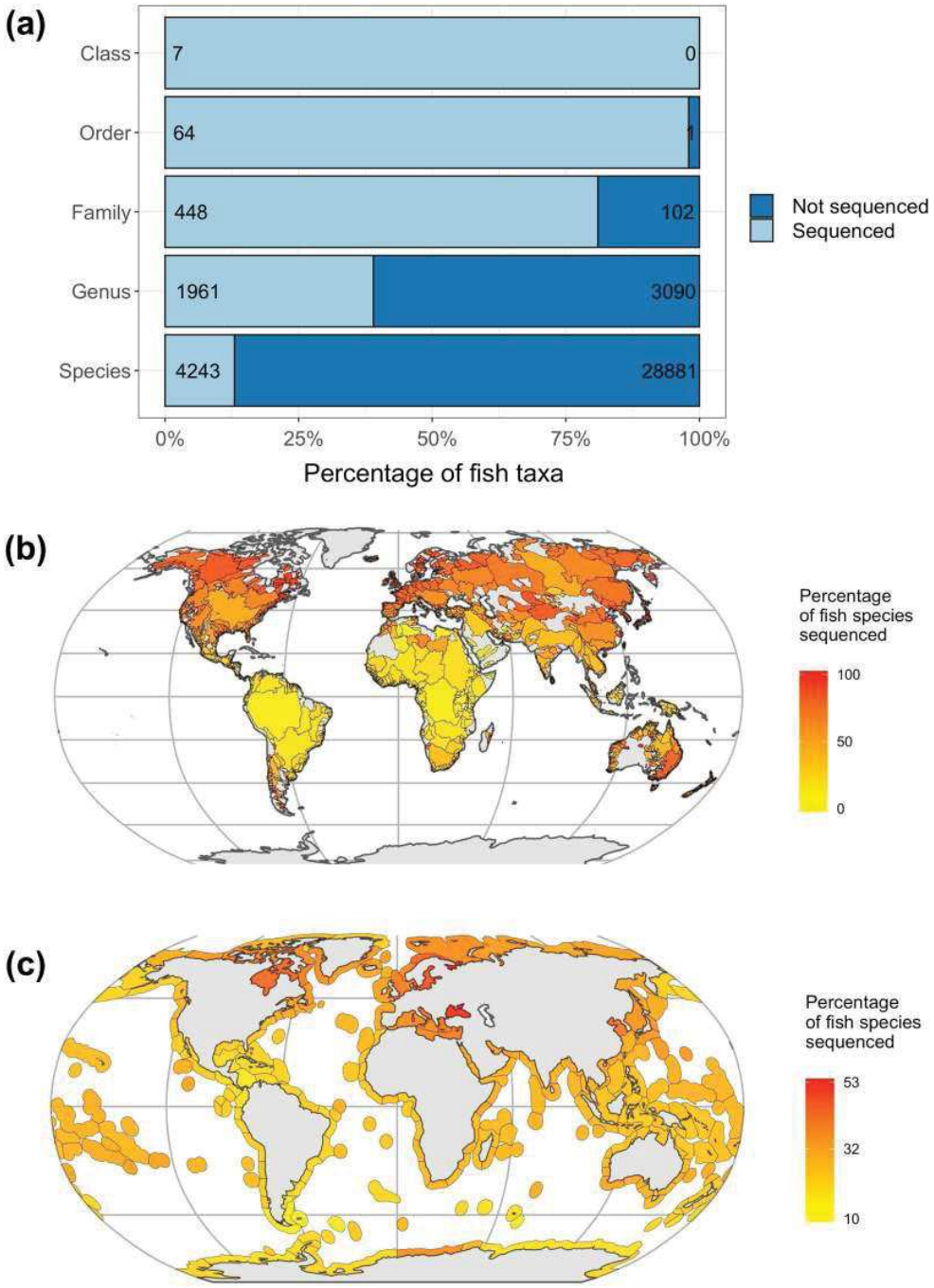


Figure 2. Percentage of sequenced freshwater and marine fish species using the teleo primer per taxonomic level (a), per freshwater basin (b) and per marine ecoregion (c).

Discussion

While eDNA metabarcoding represents a promising tool for scaling-up biodiversity inventories (Berry et al. 2019, Ruppert et al. 2019), its strong dependence on genetic reference databases limits its application in many regions of the world, as well as for some taxonomic groups or some habitats (Weigand et al. 2019). Indeed, even diverse yet well-studied ecosystems such as coral reefs do not have exhaustive genetic references for most lineages and the majority of commonly used primers in eDNA metabarcoding (DiBattista et al. 2017, West et al. 2020).

Table 1. Numbers (#) of reads, sequences, species and MOTUs identified and retained at each step of our Species or MOTUs pipelines (Fig. 1) with # Species representing the number of taxa corrected for taxonomic redundancy (see Methods). Details for each step are presented in Methods and Fig. 1.

	Species pipeline			MOTU pipeline	
Steps	# Reads	# Sequences	# Species	# Reads	# MOTUs
No filter	60 689 053	299 225	399	60 684 944	5375
> 10 reads	55 655 419	7819	227	60 593 926	568
Fish taxa	53 253 228	6424	108	57 988 700	190
Length filter	53 253 170	6422	108	57 988 623	189
> 1 PCR/site	53 021 739	6121	63	57 759 482	86
LULU	–	–	–	57 736 566	67

Some reference-free tools exist, but their application remains mostly limited to unicellular or fungi organisms, with different aims and constraints compared to eDNA studies targeting vertebrates. Moreover, such tools do not provide plausible diversity levels for most applications on vertebrate eDNA (Andruszkiewicz et al. 2017, Closek et al. 2019, Siegenthaler et al. 2019). Further, a proper testing of whether those approaches provide reliable diversity estimates is lacking (Pedrós-Alió 2006, Huse et al. 2010, Lladó Fernández et al. 2019) beyond controlled mock communities (Frøslev et al. 2017, Alberdi et al. 2018). In our study, we assembled a set of bioinformatic tools to generate fish MOTUs and assess the level of diversity across spatial scales based on the use of eDNA metabarcoding, using a well-known river system as a case study.

We show that, at the regional level, our MOTU-based pipeline provides a comparable estimate of species diversity with 67 MOTUs when 63 species are detected. However, some MOTUs represent either errors or unreferenced species, and 8 species remain undetected due to clustering and stringent filtering. Such weakness arises as many species have close sequences to each other and co-occur. So, it remains impossible for any algorithm to distinguish close species from errors. This dilemma – distinguishing rare MOTUs from errors – is inherent to clustering techniques (Huse et al. 2010, Frøslev et al. 2017, Pawlowski et al. 2018). Despite numerous attempts to solve this issue, there is still a trade-off between allowing false positives and creating false negatives (Reeder and Knight 2009). Among the MOTUs representing taxonomic redundancy, at least 3 taxa (*Gobio gobio*, *Alosa* sp.,

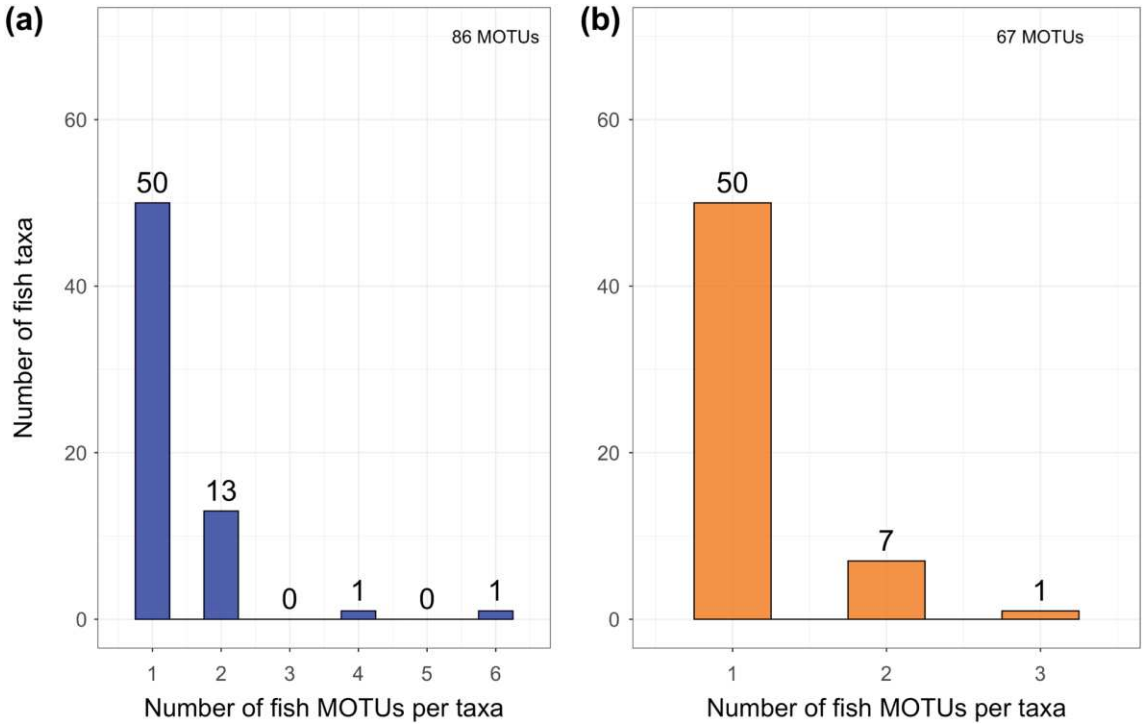


Figure 3. Distribution of the number of MOTUs per fish taxa (a) before LULU cleaning and (b) after LULU cleaning for taxonomic redundancy (Frøslev et al. 2017).

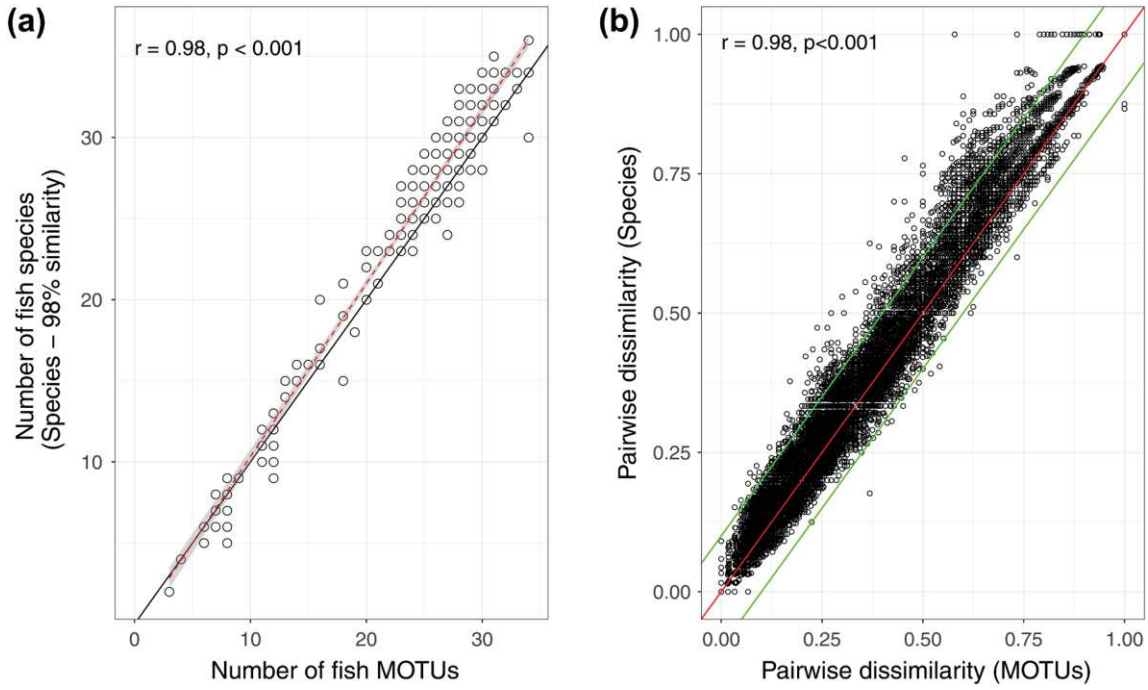


Figure 4. The Pearson linear correlation shows the strength of the relationship between the number of species and the number of MOTUs identified with our two pipelines for each sample (a). The black line represents the identity slope and the red line represents the linear regression between the number of species and that of MOTUs. (b) The Mantel correlation shows the relationship between the Species and the MOTU pipeline for pairwise sample dissimilarity. Each dot represents the β-diversity value for a pair of samples estimated by either one of our pipelines (Species versus MOTUs), red line represents the identity slope and green lines represent respectively the +0.1 and −0.1 limits.

Phoxinus phoxinus) are known to hybridize (Alexandrino et al. 2006) or are under taxonomic revision with the potential existence of multiple species displaying genetic variations (Kottelat and Persat 2005, Collin and Fumagalli 2011) while for one species (*Dicentrarchus labrax*), genetic public databases (Sayers et al. 2019) highlight a marked intra-specific variability.

Sequencing and PCR errors are common in metabarcoding datasets (Siegwald et al. 2017), but as eDNA barcodes are usually short to enhance detectability (Bohmann et al. 2014, Deiner et al. 2017), one mismatch generated randomly can easily correspond to a distinct but closely related species. This poses the risk of false-positive detection, like in the present study, where several foreign species were detected (Supplementary material Appendix 1 Table A2). Yet, none of the false positive species detected with the Species pipeline were retained as a MOTU after the clustering process, highlighting the strength of our clustering approach to clean false positive errors when they likely arise from PCR and sequencing errors. When using a classical metabarcoding pipeline without a stringent cleaning or clustering step to infer diversity from short sequences, those false positive species might remain in the global pool of detected species which would require special care to flag and exclude such errors (i.e. manual alignment of sequences and verification of species geographical distribution). We also show the extent to which SWARM is able to assign the correct sequence as the representative of each MOTU, since 61 (out of 67) MOTUs perfectly match to a species from the reference database. Our results stress the importance to combine post-clustering filters based on PCR replicates and a cleaning algorithm to remove spurious MOTUs.

Since our MOTU-based pipeline slightly overestimates regional diversity with 67 MOTUs obtained compared to 63 species identified, a key question is how it can impact local diversity assessment. We found a slight tendency for MOTUs to underestimate species richness, with less than 2 MOTUs of difference compared to the number of species for most samples. The underestimation of diversity stemming from missed species (8 species so 13% of the regional pool) is not totally compensated by the overestimation caused by taxonomic redundancy in the regional pool. Further, no outliers were identified over all 196 samples. We also show that most of mentioned pitfalls do not impact patterns of dissimilarity at the community scale, as results are similar whether they are based on blind MOTUs or species identification. In summary, the assessment of local diversity is nearly not impacted by the absence of a complete reference database, both estimates are highly correlated (98%) with a mean difference of one species between pipelines.

While these results are valid using the teleo marker (12S rDNA, ~60 bp long), we could not validate our pipeline using other primer sets due to time and financial constraints. This pipeline can still be applied to other markers, but it would require a marker with a similar level of taxonomic specificity and limited intra-specific variation, to avoid an over-estimation of taxonomic diversity due to haplotype diversity. An application with another primer would require more investigation to test if threshold adjustments are necessary to match its specificities (i.e. PCR replicates number, minimum number of reads, LULU parameters, minimum distance in SWARM clustering). We suggest the design of a small pilot study in a well-known system to validate its blind predictive power before larger-scale applications.

We show that our approach using MOTUs delivers robust estimates of species diversity at the three geographic scales, unlocking new potential for biodiversity monitoring through eDNA. With more than 75% of fish families potentially detectable, our approach can go beyond the simple delineation of sequences within clusters when further assigning taxonomy to our MOTUs. In particular, the use of assignment algorithms such as the Lowest Common Ancestor (LCA) algorithm (Boyer et al. 2016, Gao et al. 2017) is well suited for taxonomic assignment in eDNA studies with incomplete reference database. We can then estimate the potential number of species per family when the sequence coverage within families is sufficient for such assessment. While a family assignment has limitations, ecological characteristics are generally well conserved for species within a given family (Brandl et al. 2018) and allow relevant metrics of ecological analyses to be computed at this scale. As the minimum coverage within family necessary for robust detection using LCA is likely to vary across taxa and goes beyond the scope of this study, a complete coverage is not requested and our approach can provide an accurate estimate of species diversity within family for ecological studies. Yet, we highlight some limitations when it comes to conservation policies for which unnamed MOTUs will not be satisfying. As conservation programs usually focus on few taxa which are mostly rare, threatened, invasive or emblematic (Pimm et al. 2018, Enquist et al. 2019, Hannah et al. 2020), achieving the complete sequencing of those target species is urgent but realistic in the near future, as opposed to the sequencing of most vertebrates. The current filling of global DNA databases is sufficient for our approach to work globally and across scales. Diversity indices derived from this method are shown to be reliable at α, β and γ scales to infer similar ecological conclusions as those based on classical species identification.

Conclusion

While it has widely been reported that molecular biodiversity inventories outperform classical inventories (videos, acoustic) in the open environment (Thomsen et al. 2016, Boussarie et al. 2018), we demonstrate that, in the absence of a complete genetic reference database, a bioinformatic pipeline using Molecular Operational Taxonomic Units is able to provide robust estimates of species diversity across spatial scales. Even if some species cannot being distinguished after the clustering step, a common issue due to genetic proximity between close taxa (Fahner et al. 2016), the geographic biodiversity patterns are highly similar to those obtained with a species-based method. As false negatives are inherent to any inventory method in ecology (Field et al. 2007) and while false positives are rarer but to avoid at all cost (Chambert et al. 2015), we suggest a precautionary approach where some 'true' observations could be lost in order to reduce the risk of false observations. Given the current state of genetic database coverage, a species-based eDNA approach is only achievable in freshwater ecosystems located in the Northern hemisphere, where the coverage exceeds 50% of fish species (Fig. 2). For all other ecosystems, our study is the proof of concept demonstrating that, given an appropriate primer set as well as filtering and cleaning processes, MOTUs can be used to accurately assess the level of biodiversity at all scales: local, turnover and regional. We thus advocate the need to focus sequencing efforts in priority towards 1) families with no genetic coverage so presently virtually undetectable with our approach and 2) conservation-important like invasive species or IUCN Red List species for which unassigned MOTUs cannot substitute. This work paves the way towards extending the use of eDNA in community ecology and biogeography even for poorly known ecosystems or lineages, and install eDNA as a standard monitoring tool (Jarman et al. 2018). It also reinforces its initial goal of versatility and high comparability to monitor any kind of ecosystem and compare communities across wide environmental gradients.

Data and code availability

The Species (<https://gitlab.mbb.univ-montp2.fr/edna/bash_105vsearch_ecotag>) and MOTU pipelines (<https://gitlab.mbb.univ-montp2.fr/edna/bash_swarm>) are freely accessible in Gitlab. All sequencing data is already available on Dryad: <https://doi.org/10.5061/dryad.t4n42rr> (Pont et al. 2019).

Acknowledgements – We thank SPYGEN and CNR team for contributing to the field work and/or the laboratory analysis, and Franck Pressiat from the CNR for valuable comments on the manuscript.
Funding – Funding for this work was provided by the 'Compagnie Nationale du Rhône' (CNR) and SPYGEN.
Author contributions – VM, TD, DM, MR and SM contributed to the study design, PEG and VM wrote the bioinformatic pipeline, TD and MR conducted the fieldwork, VM and AV performed the analysis, and all authors contributed towards writing, reviewing and editing the article.
Conflicts of interest – MR is a research engineers of a French electricity generation companies, AV and TD are research scientists in a private company, specialized on the use of eDNA for species detection.

References

Alberdi, A. et al. 2018. Scrutinizing key steps for reliable metabarcoding of environmental samples. – Methods Ecol. Evol. 9: 134–147.

Albouy, C. et al. 2019. The marine fish food web is globally connected. – Nat. Ecol. Evol. 3: 1153–1161.

Alexandrino, P. et al. 2006. Interspecific differentiation and intraspecific substructure in two closely related clupeids with extensive hybridization, *Alosa alosa* and *Alosa fallax*. – J. Fish Biol. 69: 242–259.

Andruszkiewicz, E. A. et al. 2017. Biomonitoring of marine vertebrates in Monterey Bay using eDNA metabarcoding. – PLoS One 12: e0176343.

Bakker, J. et al. 2017. Environmental DNA reveals tropical shark diversity in contrasting levels of anthropogenic impact. – Sci. Rep. 7: 1–11.

Baselga, A. and Orme, C. D. L. 2012. Betapart: an R package for the study of beta diversity. – Methods Ecol. Evol. 3: 808–812.

Berry, T. E. et al. 2019. Marine environmental DNA biomonitoring reveals seasonal patterns in biodiversity and identifies ecosystem responses to anomalous climatic events. – PLoS Genet. 15: e1007943.

Blowes, S. A. et al. 2019. The geography of biodiversity change in marine and terrestrial assemblages. – Science 366: 339–345.

Bohmann, K. et al. 2014. Environmental DNA for wildlife biology and biodiversity monitoring. – Trends Ecol. Evol. 29: 358–367.

Boussarie, G. et al. 2018. Environmental DNA illuminates the dark diversity of sharks. – Sci. Adv. 4: eaap9661.

Boyer, F. et al. 2016. OBITOOLS: a UNIX-inspired software package for DNA metabarcoding. – Mol. Ecol. Resour. 16: 176–182.

Brandl, S. J. et al. 2018. The hidden half: ecology and evolution of cryptobenthic fishes on coral reefs. – Biol. Rev. 93: 1846–1873.

Bylemans, J. et al. 2018. Toward an ecoregion scale evaluation of eDNA metabarcoding primers: a case study for the freshwater fish biodiversity of the Murray–Darling Basin (Australia). – Ecol. Evol. 8: 8697–8712.

Callahan, B. J. et al. 2016. DADA2: high resolution sample inference from Illumina amplicon data. – Nat. Methods 13: 581–583.

Callahan, B. J. et al. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. – ISME J. 11: 2639–2643.

Cantera, I. et al. 2019. Optimizing environmental DNA sampling effort for fish inventories in tropical streams and rivers. – Sci. Rep. 9: 1–11.

Chambert, T. et al. 2015. Modeling false positive detections in species occurrence data under different study designs. – Ecology 96: 332–339.

Cilleros, K. et al. 2019. Unlocking biodiversity and conservation studies in high-diversity environments using environmental DNA (eDNA): a test with Guianese freshwater fishes. – Mol. Ecol. Resour. 19: 27–46.

Civade, R. et al. 2016. Spatial representativeness of environmental DNA metabarcoding signal for fish biodiversity assessment in a natural freshwater system. – PLoS One 11: e0157366.

Closek, C. J. et al. 2019. Marine vertebrate biodiversity and distribution within the central California Current using environmental DNA (eDNA) metabarcoding and ecosystem surveys. – Front. Mar. Sci. 6: 732.

Collin, H. and Fumagalli, L. 2011. Evidence for morphological and adaptive genetic divergence between lake and stream habitats in European minnows (*Phoxinus phoxinus*, Cyprinidae). – Mol. Ecol. 20: 4490–4502.

Collins, R. A. et al. 2018. Persistence of environmental DNA in marine systems. – Commun. Biol. 1: 185.

Collins, R. A. et al. 2019. Non-specific amplification compromises environmental DNA metabarcoding with COI. – Methods Ecol. Evol. 10: 1985–2001.

Deagle, B. E. et al. 2014. DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. – Biol. Lett. 10: 20140562.

Deiner, K. et al. 2017. Environmental DNA metabarcoding: transforming how we survey animal and plant communities. – Mol. Ecol. 26: 5872–5895.

Dejean, T. et al. 2011. Persistence of environmental DNA in freshwater ecosystems. – PLoS One 6: e23398.

DiBattista, J. D. et al. 2017. Assessing the utility of eDNA as a tool to survey reef-fish communities in the Red Sea. – Coral Reefs 36: 1245–1252.

Doherty, T. S. et al. 2016. Invasive predators and global biodiversity loss. – Proc. Natl Acad. Sci. USA 113: 11261–11265.

Dornelas, M. et al. 2014. Assemblage time series reveal biodiversity change but not systematic loss. – Science 344: 296–299.

Edgar, R. C. 2018. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. – Bioinformatics 34: 2371–2375.

Edgar, R. C. and Flyvbjerg, H. 2015. Error filtering, pair assembly and error correction for next-generation sequencing reads. – Bioinformatics 31: 3476–3482.

Enquist, B. J. et al. 2019. The commonness of rarity: global and future distribution of rarity across land plants. – Sci. Adv. 5: 1–14.

Fahner, N. A. et al. 2016. Large-scale monitoring of plants through environmental DNA metabarcoding of soil: recovery, resolution and annotation of four DNA markers. – PLoS One 11: e0157505.

Field, S. A. et al. 2007. Improving precision and reducing bias in biological surveys: estimating false-negative error rates. – Ecol. Appl. 13: 1790–1801.

Finderup Nielsen, T. et al. 2019. More is less: net gain in species richness, but biotic homogenization over 140 years. – Ecol. Lett. 22: 1650–1657.

Flynn, J. M. et al. 2015. Toward accurate molecular identification of species in complex environmental samples: testing the performance of sequence filtering and clustering methods. – Ecol. Evol. 5: 2252–2266.

Froese, R. and Pauly, D. 2019. Fishbase. – <www.fishbase.org>.

Frøslev, G. T. et al. 2017. Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. – Nat. Commun. 8: 1188.

Galetti, M. et al. 2014. Defaunation in the Anthropocene. – Science 345: 401–406.

Gao, X. et al. 2017. A Bayesian taxonomic classification method for 16S rRNA gene sequences with improved species-level accuracy. – BMC Bioinform. 18: 1–10.

Hannah, L. et al. 2020. 30% land conservation and climate action reduces tropical extinction risk by more than 50%. – Ecography 43: 943–953.

Harrison, J. B. et al. 2019. Predicting the fate of eDNA in the environment and implications for studying biodiversity. – Proc. R. Soc. B 286: 20191409.

Hortal, J. et al. 2015. Seven shortfalls that beset large-scale knowledge of biodiversity. – Annu. Rev. Ecol. Evol. Syst. 46: 523–549.

Hughes, T. P. et al. 2017. Coral reefs in the Anthropocene. – Nature 546: 82–90.

Huse, S. M. et al. 2010. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. – Environ. Microbiol. 12: 1889–1898.
Isbell, F. et al. 2019. Deficits of biodiversity and productivity linger a century after agricultural abandonment. – Nat. Ecol. Evol. 3: 1533–1538.
Jarman, S. N. et al. 2018. The value of environmental DNA biobanking for long-term biomonitoring. – Nat. Ecol. Evol. 2: 1192–1193.
Kottelat, M. and Persat, H. 2005. The genus *Gobio* in France, with redescription of *G. gobio* and description of two new species (Teleostei: Cyprinidae). – Cybium 29: 211–234.
Leinonen, R. et al. 2011. The European nucleotide archive. – Nucleic Acids Res. 39: 44–47.
Lladó Fernández, S. et al. 2019. The concept of operational taxonomic units revisited: genomes of bacteria that are regarded as closely related are often highly dissimilar. – Folia Microbiol. 64: 19–23.
Magurran, A. E. et al. 2018. Divergent biodiversity change within ecosystems. – Proc. Natl Acad. Sci. USA 115: 1843–1847.
Mahé, F. et al. 2014. Swarm: robust and fast clustering method for amplicon-based studies. – PeerJ 2: e593.
Mahé, F. et al. 2015. Swarm v2: highly-scalable and high-resolution amplicon clustering. – PeerJ 3: e1420.
Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. – EMBnet.journal 17: 10.
McCauley, D. J. et al. 2015. Marine defaunation: animal loss in the global ocean. – Science 347: 247–254.
McGill, B. J. et al. 2015. Fifteen forms of biodiversity trend in the anthropocene. – Trends Ecol. Evol. 30: 104–113.
Menegotto, A. and Rangel, T. F. 2018. Mapping knowledge gaps in marine diversity reveals a latitudinal gradient of missing species richness. – Nat. Commun. 9: 4713.
Miya, M. et al. 2015. MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than 230 subtropical marine species. – R. Soc. Open Sci. 2: 150088.
Mora, C. et al. 2008. The completeness of taxonomic inventories for describing the global diversity and distribution of marine fishes. – Proc. R. Soc. B 275: 149–155.
Mora, C. et al. 2011a. How many species are there on earth and in the ocean? – PLoS Biol. 9: e1001127.
Mora, C. et al. 2011b. Global human footprint on the linkage between biodiversity and ecosystem functioning in reef fishes. – PLoS Biol. 9: e1000606.
Nguyen, N.-P. et al. 2015. A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. – Biofilms Microbiomes 1: 10–13.
Pawlowski, J. et al. 2018. The future of biotic indices in the ecogenomic era: integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. – Sci. Total Environ. 637–638: 1295–1310.
Pedrós-Alió, C. 2006. Marine microbial diversity: can it be determined? – Trends Microbiol. 14: 257–263.
Pimm, S. L. et al. 2018. How to protect half of earth to ensure it protects sufficient biodiversity. – Sci. Adv. 4: 1–9.
Pont, D. et al. 2018. Environmental DNA reveals quantitative patterns of fish biodiversity in large rivers despite its downstream transportation. – Sci. Rep. 8: 1–13.
Pont, D. et al. 2019. Data from: Environmental DNA reveals quantitative patterns of fish biodiversity in large rivers despite its downstream transportation. – Dryad, Dataset, <https://doi.org/10.5061/dryad.t4n42rr>.
Reeder, J. and Knight, R. 2009. The 'rare biosphere': a reality check. – Nat. Methods 6: 636–637.
Rognes, T. et al. 2016. VSEARCH: a versatile open source tool for metagenomics. – PeerJ 4: e2584.
Ruppert, K. M. et al. 2019. Past, present and future perspectives of environmental DNA (eDNA) metabarcoding: a systematic review in methods, monitoring and applications of global eDNA. – Global Ecol. Conserv. 17: e00547.
Sales, N. G. et al. 2019. Influence of preservation methods, sample medium and sampling time on eDNA recovery in a neotropical river. – Environ. DNA 1: 119–130.
Sales, N. G. et al. 2020. Fishing for mammals: landscape-level monitoring of terrestrial and semi-aquatic communities using eDNA from riverine systems. – J. Appl. Ecol. 57 707–716.
Sayers, E. W. et al. 2019. GenBank. – Nucleic Acids Res. 47: D94–D99.
Schmidt, P. A. et al. 2013. Illumina metabarcoding of a soil fungal community. – Soil Biol. Biochem. 65: 128–132.
Schnell, I. B. et al. 2015. Tag jumps illuminated – reducing sequence-to-sample misidentifications in metabarcoding studies. – Mol. Ecol. Resour. 15: 1289–1303.
Siegenthaler, A. et al. 2019. Metabarcoding of shrimp stomach content: harnessing a natural sampler for fish biodiversity monitoring. – Mol. Ecol. Resour. 19: 206–220.
Siegwald, L. et al. 2017. Assessment of common and emerging bioinformatics pipelines for targeted metagenomics. – PLoS One 12: e0169563.
Spalding, M. D. et al. 2007. Marine ecoregions of the world: a bioregionalization of coastal and shelf areas. – Bioscience 57: 573.
Stackebrandt, E. and Goebel, B. M. 2008. Taxonomic note: a place for DNA–DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. – Int. J. Syst. Evol. Microbiol. 44: 846–849.
Tedesco, P. A. et al. 2017. Data Descriptor: a global database on freshwater fish species occurrence in drainage basins. – Sci. Data 4: 1–6.
Thomsen, P. F. et al. 2016. Environmental DNA from seawater samples correlate with trawl catches of subarctic, deepwater fishes. – PLoS One 11: e0165252.
Tsuji, S. et al. 2019. The detection of aquatic macroorganisms using environmental DNA analysis – a review of methods for collection, extraction and detection. – Environ. DNA 1: 99–108.
Valentini, A. et al. 2016. Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. – Mol. Ecol. 25: 929–942.
Vellend, M. et al. 2013. Global meta-analysis reveals no net change in local-scale plant biodiversity over time. – Proc. Natl Acad. Sci. USA 110: 19456–19459.
Weigand, H. et al. 2019. DNA barcode reference libraries for the monitoring of aquatic biota in Europe: gap-analysis and recommendations for future work. – Sci. Total Environ. 678: 499–524.
West, K. M. et al. 2020. eDNA metabarcoding survey reveals fine-scale coral reef community variation across a remote, tropical island ecosystem. – Mol. Ecol. 29: 1069–1086.
Whittaker, R. H. 1972. Evolution and measurement of species diversity. – Taxon 21: 213–251.
Zimmermann, J. et al. 2015. Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. – Mol. Ecol. Resour. 15: 526–542.

Supplementary material (available online as Appendix ecog-05049 at <www.ecography.org/appendix/ecog-05049>). Appendix 1.

Chapitre 4 - Comparaison entre l'ADNe metabarcoding et des méthodes conventionnelles de recensement de la diversité en poissons marins tropicaux

Ce chapitre est composé de deux articles :

Manuscrit C

Marques V., Castagné P., Polanco A., Borrero G.H., Hocdé R., Guérin P-E., Juhel J-B., Velez L., Loiseau N., Letessier T., Bessudo S., Valentini A., Dejean T., Mouillot D., Pellissier L., Villéger S. Environmental DNA improves the assessment of fish functional and phylogenetic diversity on tropical reefs. (2020). Submitted.

Manuscrit D

Polanco A. [1]*, Marques V.* [1]*, Fopp F., Juhel J-B., Borrero G.H., Cheutin M-C., Dejean T., Gonzalès J.D.C., Acosta-Chaparro A., Hocdé R., Eme D., Maire E., Spescha M., Valentini A., Manel S., Mouillot D., Albouy C., Pellissier L. Comparing environmental DNA metabarcoding and underwater visual census to monitor tropical reef fishes. (2020). Environmental DNA. In Press.*
[1]*: shared first authorship*

1. Préface

Les besoins en suivi de la biodiversité sont urgents, et le metabarcoding de l'ADNe est une méthode très prometteuse pour complémenter voire remplacer certains recensements classiques de la biodiversité. Cependant, il est nécessaire d'évaluer la performance de l'ADNe par rapport aux outils traditionnels avant d'envisager une généralisation du recours à cette méthode. Or nous avons vu en introduction de cette thèse que les méthodes ADNe ont jusqu'ici été peu appliquées en milieu marin, et encore moins dans les milieux marins tropicaux pour lesquels très peu d'études comparatives existent, hormis Stat et al. (2019). L'application de l'ADNe dans ces écosystèmes reste cependant limitée par la faible couverture taxonomique des bases de références génétiques dans cette zone, comme exposé dans le chapitre 2.

Ce nouveau chapitre s'appuie sur les avancées réalisées au chapitre 3 qui permettent d'utiliser des substituts moléculaires (MOTUs) en guise d'unités taxonomiques à la place des traditionnelles assignations à l'espèce. Les études comparatives précédentes montraient une faible performance de l'ADNe par rapport aux méthodes traditionnelles notamment à cause de manquements dans la base de référence (McElroy et al. 2020), la réelle diversité récupérée sur les filtres ADNe n'étant que partiellement révélée en raison de ces bases de références lacunaires. L'utilisation de MOTUs permet ici d'éviter cet effet et de montrer le potentiel de l'ADNe pour étudier la diversité d'une zone ainsi que du nombre d'espèces potentiellement détectables une fois que les bases de référence seront plus renseignées.

De plus, la très large majorité des études comparatives entre ADNe et autres méthodes se concentrent sur la facette taxonomique de la diversité. Or les facettes phylogénétiques et fonctionnelles sont essentielles afin de décrire et comprendre la structure des écosystèmes, ainsi que proposer des mesures de conservation de la biodiversité (Pollock et al. 2020). Au lieu de considérer les espèces / MOTUs comme totalement similaires dans leur contribution à l'écosystème, une approche multi-facettes permet de grouper les unités taxonomiques en entités fonctionnelles partageant des valeurs de traits et considérées comme remplissant des rôles fonctionnels similaires. Il s'agit aussi de considérer la diversité de l'histoire évolutive dans une communauté composées d'unités taxonomiques à travers leurs positions sur un arbre phylogénétique.

A travers deux manuscrits, ce chapitre vise à (i) établir une comparaison entre l'ADNe metabarcoding et des méthodes conventionnelles (vidéos et transects en plongée) indépendamment

de la couverture taxonomique des bases de références sur deux régions marines tropicales, (ii) comparer les performances de chaque méthode sur trois facettes de la diversité : taxonomique, fonctionnelle et phylogénétique et (iii) comparer l'effort d'échantillonnage nécessaire pour atteindre la valeur maximale de chaque métrique de diversité.

Pour cela, nous avons échantillonné sur deux zones d'études sur le territoire Colombien. La première dans le Pacifique Oriental Tropical sur l'île de Malpelo, à 500km des côtes de la Colombie, seul point émergé d'une chaîne de montagnes sous-marines agrégeant une large faune récifale et pélagique. La deuxième zone se situe dans les Caraïbes, où un site comprend l'île de Providencia au large du Nicaragua et l'autre site le parc naturel de Tayrona à Santa Marta, proche du parc de la Sierra Nevada.

Les résultats révèlent que l'ADNe surpasse les méthodes traditionnelles pour chacune des trois facettes de diversité. L'ADNe détecte peu de MOTUs assignés au rang de l'espèce pour les deux régions (Pacifique Oriental Tropical et Caraïbes), mais l'approche par unités taxonomiques expose le potentiel de l'ADNe si les bases étaient plus complètes, en détectant systématique plus de genres et de familles que les vidéos et les plongées, ce qui induit un avantage certain pour l'estimation de la diversité phylogénétique. Cet avantage est également marqué pour les aspects fonctionnels, pour lesquels l'ADNe détecte plus d'entités fonctionnelles et plus de richesse fonctionnelle que les méthodes classiques. En termes d'efforts d'échantillonnage, l'ADNe détecte autant de diversité phylogénétique en un seul transect qu'en 25h de vidéos, avec une logistique beaucoup plus légère. Ces travaux apportent un argument supplémentaire pour l'application de l'ADNe dans les milieux marins tropicaux hyper-divers, où la méthode présente un fort potentiel pour améliorer les moyens de suivi de biodiversité traditionnels en élargissant la gamme de familles donc de traits et d'histoire évolutive qui devient détectable. La principale limite réside dans la faible couverture des bases de référence, qui peut être rapidement levée à l'échelle locale avec un effort ciblé pour les zones de suivi choisies.

2. Manuscrit C

Environmental DNA improves the assessment of fish functional and phylogenetic diversity on tropical reefs

Virginie Marques[1,2*], Paul Castagné[1], Andréa Polanco[3], Giomar Borrero[3], Régis Hocdé[1], Pierre-Édouard Guérin[2], Jean-Baptiste Juhel[1], Laure Velez[1], Nicolas Loiseau[1], Tom B Letessier[4], Sandra Bessudo[5], Alice Valentini[6], Tony Dejean6, David Mouillot[1,7], Loïc Pellissier[8], Sébastien Villéger[1]

[1:] MARBEC, Univ. Montpellier, CNRS, Ifremer, IRD, Montpellier, France

[2:] CEFE, Univ. Montpellier, CNRS, EPHE-PSL University, IRD, Univ Paul Valery Montpellier 3, Montpellier, France

[3:] Instituto de Investigaciones Marinas y Costeras-INVEMAR, Colombia. Museo de Historia Natural Marina de Colombia (MHNMC), Programa de Biodiversidad y Ecosistemas Marinos. Calle 25 No. 2 – 55 Playa Salguero, Santa Marta, Colombia

[4:] Institute of Zoology, Zoological Society of London, London, United Kingdom

[5:] Fundación Malpelo y otros ecosistemas marinos. Bogotá, Colombia

[6:] SPYGEN, 17 rue du Lac Saint-André Savoie Technolac - BP 274, Le Bourget-du-Lac, 73375, France

[7:] Institut Universitaire de France, France

[8:] Landscape Ecology, Institute of Terrestrial Ecosystems, Department of Environmental Systems Science, ETH Zürich, Zürich, Switzerland

*Corresponding author: virginie.marques01@gmail.com

Keywords: video, eDNA metabarcoding, biodiversity, functional traits, accumulation curves, Malpelo, Marine Protected Area

Authors contribution

D.M., L.P., S.V. and V.M. designed research; V.M., A.P., G.B., R.H., J.B.J., L.V., N.L., T.B.L., S.B. and L.P. collected samples and data on the field; S.B. organize expedition and permits. T.D. and A.V. coordinated biomolecular analyses; V.M. and A.V. performed the bioinformatics analyses; P.C. analyzed the videos, V.M., P.C. and S.V. analyzed the data; V.M. wrote the initial draft and designed the figures; all authors wrote the paper and approved the final draft.

Abstract

1- Monitoring marine fishes is a key step for assessing impact of global changes as well as protection effectiveness. Most traditional census methods are demanding or destructive. Non-disturbing and non-lethal approaches based on video and environmental DNA have been increasingly used as alternatives to underwater visual census or fishing. However, their ability to recover multiple facets of biodiversity beyond the traditional taxonomic diversity is still unknown.
2- Here, we compared the performance of eDNA metabarcoding targeting bony fishes and elasmobranches to that of long-duration remote videos in the assessment of species, phylogenetic and functional diversity. We used 10 eDNA samples from 30L of water each and 25 hours of underwater videos over three days on Malpelo island, on the pacific coast of Colombia, a remote marine protected area within a biodiversity hotspot.
3- eDNA detected 66% more Molecular Operational Taxonomic Units (MOTUs) than species on videos. We found 66 and 43 functional entities using a single eDNA marker and videos, respectively, and higher functional richness for eDNA than videos. Despite gaps in genetic reference databases, eDNA also recovered a higher fish phylogenetic diversity than videos, with accumulation curves revealing how a single eDNA transect can detect as much phylogenetic diversity as 25 hours of videos.
4- Synthesis and applications. Environmental DNA is an efficient and accurate method to census all biodiversity facets in marine systems. Although taxonomic assignments are still limited by species coverage in genetic reference databases, the use MOTUs highlights the potential of eDNA metabarcoding if efforts are made to sequence regional diversity. eDNA offers an affordable efficient opportunity to complement traditional monitoring methods in a context of ecosystem management.

Introduction

Species inventories are the building blocks of most ecological analyses, from biogeography to conservation or community ecology. In a context of global changes, monitoring species communities is essential for biodiversity assessment (Blowes et al., 2019) and the evaluation of management strategies (Cinner et al., 2020). Most biodiversity inventories focus on the taxonomic facet, where each species is considered independently of its evolutionary history or functional traits (Cardoso, Rigal, Borges, & Carvalho, 2014). Yet, species diversity alone is not sufficient to inform ecosystem states and processes since not all species are equivalent (Craven et al. 2018; Brun et al. 2019). Besides, a multifaceted approach of biodiversity is often required to better understand community changes and conservation outcomes (Kling, Mishler, Thornhill, Baldwin, & Ackerly, 2019; Mbaru, Graham, McClanahan, & Cinner,

2020; Monnet et al., 2014; Segovia et al., 2020; Trindade-Santos, Moyes, & Magurran, 2020). So far, few studies have compared the ability of inventory methods to recover all facets of biodiversity, focusing mostly either on taxonomic diversity alone or on a combination of two facets only (Devictor et al., 2010).

Taxonomic diversity (TD) represents the sum of species present in a given community and is the most widely used facet to describe biodiversity (Cardoso et al., 2014). Yet, TD ignores ecological differences among species (Cardoso et al., 2014; Jarzyna & Jetz, 2016). Two prominent approaches have been proposed to complement taxonomic information by accounting for species ecological features and evolutionary divergence (McGill, Enquist, Weiher, & Westoby, 2006; Webb, Ackerly, McPeek, & Donoghue, 2002). Phylogenetic diversity (PD) quantifies the extent of evolutionary history present in a given community, a key facet to biogeography, conservation and ecosystem functioning (Forest et al., 2007; Tucker et al., 2019, 2016). Functional diversity (FD), the extent of species trait values, sheds light on community assembly rules and ecosystem functioning (Mouillot, Graham, Villéger, Mason, & Bellwood, 2013). While PD has often been considered as a surrogate for FD, recent studies challenge this assumption (Mazel et al., 2018), or reveal an asynchrony in responses of both facets to disturbances (Devictor et al., 2010; Monnet et al., 2014). So, TD, FD and PD are thus complementary facets to be considered in parallel as part of a comprehensive assessment of biodiversity.

In marine coastal ecosystems, monitoring is traditionally performed using Underwater Visual Census (UVCs) (Cinner et al., 2020), Remote Underwater Video systems (RUVs) with or without bait (Wetz, Ajemian, Shipley, & Stunz, 2020) or more recently Environmental DNA (eDNA) metabarcoding, a molecular technic recovering DNA traces from organisms by water filtration (Deiner et al., 2017; Harrison, Sunday, & Rogers, 2019). While UVCs have known biases such as limited sampling time and space or diver avoidance (Dickens, Goatley, Tanner, & Bellwood, 2011; MacNeil et al., 2008; McClanahan et al., 2007), video-based assessments provide an alternative with a potential to film long hours without diver presence (Dickens et al., 2011). Indeed, remote videos can recover about the same TD as the most historical UVC methods, missing small benthic and low-range species but more frequently detecting large predators (Bosch, Gonçalves, Erzini, & Tuya, 2017; Colton & Swearer, 2010; Langlois et al., 2010). Environmental DNA has been shown to recover more or about the same TD than traditional methods like netting, UVC or RUVs (Boussarie et al., 2018; Nguyen et al., 2020), most often revealing a complementary inventory (Stat et al., 2019). Yet, to our knowledge, no study compared the ability of eDNA vs. video surveys to reveal all three biodiversity facets.

Here we used eDNA metabarcoding and long-duration videos to compare the biodiversity of marine fishes and sharks detected in Malpelo island, a marine protected area recognized as a World Heritage

Site by UNESCO. Malpelo is known to host a high diversity of underwater life, with around 400 fish and shark species recorded, including at least 5 endemic fish species (Chasqui Velasco, Gil Agudelo, & Nieto, 2016) and representing a potentially large functional diversity with a fauna composed of both reef-associated species and migrating pelagic species (Nalesso et al., 2019; Quimbayo, Mendes, Kulbicki, Floeter, & Zapata, 2017). We took advantage of this unique ecosystem features to compare the performance of two sampling methods, environmental DNA metabarcoding using three molecular markers and long-duration underwater videos, in the assessment of three biodiversity facets in light of corresponding sampling efforts.

Methods

1. Study site and sampling

We sampled around the Sanctuary of Fauna and Flora in Malpelo, a remote oceanic island located 490 km off the Colombian shore in the Eastern Tropical Pacific (Fig. 1) for three days over a four-day period (25-28 March 2018) on a single site, called El Arrecife. Malpelo is surrounded by deep-waters and fishing activities are prohibited over a surrounding area of 8,757 km2 (Edgar et al., 2011). The reef ecosystem around the island is influenced by major oceanic currents (Rodríguez-Rubio, Schneider, & del Río, 2003) and local upwelling, with a benthos dominated by bare rocks with a low coral cover (Quimbayo et al., 2017). We deployed a long-duration Remote Underwater Video system (RUVs hereafter) designed by Extrem-Vision (https://extrem-vision.com/en/; Rivesaltes, France) able to film for up to 12 hours (Supplementary, Fig. S1 for video screenshots). The cameras were set at 40 cm above seafloor and had wide field of view (90°) recording both benthic and pelagic habitats over an area of 10m2, with a resolution of 1920*1080 pixels and 30 frames per second. We set the camera at 13m depth, on a rocky and coral reef at the 'El recife' spot (04.00600°, -81.60433°). Three videos were recorded on the 25th (daylight, night) and 28th March (daylight) 2018 (Fig 1C).

We cumulated 24h50m of video, day and night and simultaneously collected 10 eDNA samples over five surface transects. During night recording, two dive lights illuminated the camera view. A GoPro Hero 5 camera was mounted on top on the RUVs to film in the opposite direction for the first 2h of deployment of each daylight recording. A total of 24h50min video with the long-duration camera and 3h30min with the GoPro in the opposite direction were recorded. Simultaneously, we filtered two water samples from a boat on surface transects for 30 min on top of the camera at each deployment (Fig 1), for a total of 10 filters on 5 transects. We used Athena® peristaltic pumps (Proactive Environmental Products LLC, Bradenton, Florida, USA; nominal flow of 1.0L.min−1) on each side of the boat, to filter water on-site for a total of c.a. 30L trough a VigiDNA® 0.20 µM cross flow filtration capsule (SPYGEN, le Bourget du Lac, France) with disposable sterile tubing for each filtration capsule. Immediately after, the filter units

were filled with CL1 Conservation buffer (SPYGEN, le Bourget du Lac, France) and stored at room temperature until the DNA extraction.

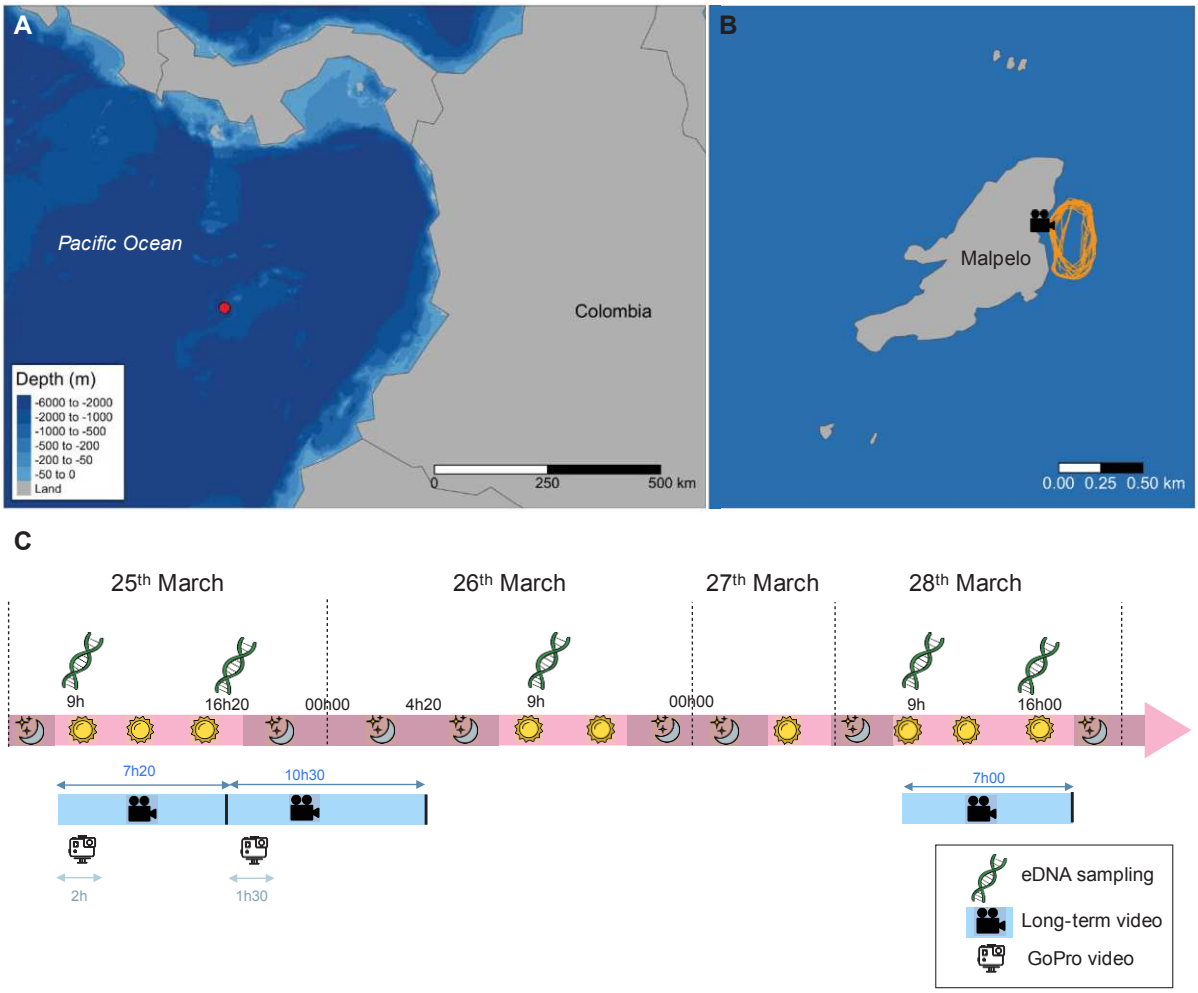


Fig. 1. Location of eDNA transects (orange tracks) and video recordings (black camera symbol) in Malpelo island, Colombia (A, B) and timing of video recording and eDNA sampling (C).

2. Video processing

Two frames per second were extracted from all videos. Fishes were identified at the lowest taxonomical level possible, following Fishbase taxonomy (Froese & Pauly, 2000), by a trained team who recorded the first occurrence of each species in each of the videos (i.e. the number of individuals per species was not accounted for).

3. Environmental DNA processing

The DNA extraction was performed in a dedicated controlled laboratory for environmental DNA extraction, equipped with positive air pressure, UV treatment and frequent air renewal and decontamination procedures conducted before and after all manipulation (Pont et al., 2018). We used

3 different primer pairs targeting distinct taxonomic groups: i) the teleo primer targeting teleost fishes and elasmobranches (Valentini et al., 2016), ii) the Chond01 primer, targeting elasmobranches most specifically and iii) the Vert01 primer (Taberlet, Bonin, Coissac, & Zinger, 2018), targeting vertebrates in general, see Table S2 for primer sequences. The PCR mixture was denatured at 95°C for 10 min, followed by 50 cycles of 30 s at 95°C, 30 s at 55°C for teleo and Vert01 and 58°C for Chon01 and 1 min at 72 °C and a final elongation step at 72°C for 7 min. Twelve replicates of PCRs were run per sample, i.e., 24 per transect as we have 2 field duplicates per transect. The primers were 5'-labeled with an eight-nucleotide tag with at least three differences between any pair of tags. The tag combinations were unique to each sample for Chond01 and Vert01 primer and unique to each PCR replicate for telo primer. The tagging system allows the assignment of each sequence to the corresponding sample during sequence analysis. After amplification, samples were titrated using capillary electrophoresis (QIAxcel; Qiagen GmbH, Hilden, Germany) and purified using a MinElute PCR purification kit (Qiagen GmbH, Hilden, Germany). The purified PCR products were pooled in equal volumes, to achieve a theoretical sequencing depth of 1,000,000 reads per sample per marker. Library preparation and sequencing were performed at Fasteris via a ligation protocol (Geneva, Switzerland). A total of three libraries were prepared using the MetaFast protocol (Fasteris, https://www.fasteris.com/dna/?q=content/metafast-protocol-amplicon-metagenomic-analysis). For all libraries a paired-end sequencing (2x125 bp) was carried out using an Illumina HiSeq 2500 sequencer on two HiSeq Rapid Flow Cell v2 using the HiSeq Rapid SBS Kit v2 (Illumina, San Diego, CA, USA). Library preparation and sequencing were performed at Fasteris (Geneva, Switzerland). Three negative extraction controls and one negative PCR controls (ultrapure water, 12 replicates) were amplified per primer pair and sequenced in parallel to the samples to monitor possible contaminants.

4. Bioinformatic pipeline

Following sequencing, reads were processed using clustering and post-clustering cleaning to remove errors and estimate the number of species using Molecular Operational Taxonomic Units (MOTUs) (J. Juhel et al., 2020; Marques et al., 2020). The estimation of species diversity using MOTU richness was only performed using the teleo marker, as other primers have not been extensively tested and their performance for correctly estimating species richness remains unassessed. First, reads were assembled using VSEARCH (Rognes, Flouri, Nichols, Quince, & Mahé, 2016), then cut using CUTADAPT (Martin, 2011) and clustering was performed using SWARM (Mahé, Rognes, Quince, de Vargas, & Dunthorn, 2015) with a minimum distance of 1 mismatch between clusters following Marques et al. (2020). Taxonomic assignation of MOTUs was carried out using ecotag from the OBITOOLS toolkit (Boyer et al., 2016) using the European Nucleotide Archive (ENA) (Leinonen et al., 2011) as a reference database (release 141, December 2019). We discarded all observations with less than 10 reads, present in only

one PCR in the entire dataset to avoid spurious MOTUs originating from a PCR error, as it unlikely for a same error to be generated several times in distinct PCRs and non-target taxa. We corrected for index-hoping (MacConaill et al., 2018) using a threshold empirically determined per sequencing batch using experimental blanks (combinations of tags not present in the libraries). In this case a threshold of 0.006 per MOTU for a given sequencing batch between libraries was used. We further corrected for tag-jump (Schnell, Bohmann, & Gilbert, 2015) using a threshold of 0.001 for a given MOTU within each library. Taxonomic assignments from ecotag were also corrected to avoid over-confidence in assignments: species-level assignments were validated only for an 100% sequence match, genus-level for a 90-99% match and family-level for an 85-90% match as suggested in Juhel et al. (2020). Fish names were then verified using the rfishbase R package (Boettiger, Lang, & Wainwright, 2012). All taxa assigned to deep-water or mesophotic species or lineages were flagged and not analyzed due to their impossible detection using shallow water cameras.

5. Trait-based analysis

The ecology of each fish species was described with 6 traits acting as proxies for their contribution to ecosystem functions: body size, mobility, period of activity, schooling behavior, vertical position in the water column and diet (Villéger, Brosse, Mouchet, Mouillot, & Vanni, 2017), coded as categories (Mouillot et al., 2014). When a recorded fish species was absent from this database, trait values were completed by literature when available, or replaced by the dominant trait value within its genera. Species sharing the same combination of trait values were grouped into Functional Entities (FE) (Mouillot et al., 2014). The functional distance between all pairs of fish species were computed using the Gower's distance which allows to account for several types of variables (Legendre & Legendre, 1998). In order to construct a multidimensional functional space, we performed a principal coordinate analysis (PCoA) on this distance matrix and kept the four first axes which provided a faithful representation of the initial trait-based distance between species according to the mSD quality index from (Maire, Grenouillet, Brosse, & Villéger, 2015). For the MOTUs not assigned at the species level due to gaps in the genetic reference database, trait values were assigned based on trait values from clades at higher taxonomic levels. More precisely, when a MOTU was only assigned at the genus level, we randomly sampled one species among all the species from the same genus occurring in the Tropical Eastern Pacific. If no species among the region had available trait data, we randomly sampled one species among all species from the genus. When a MOTU was assigned at the family level, the same methodology was applied among species from the same family. To evaluate the effect of assigning trait values to MOTUs based on one species from same genus or family, we ran the same analyzes based only on MOTUs assigned at the species level.

6. Comparing fish biodiversity estimates from the two inventory methods

We compared taxonomic, phylogenetic and functional diversity computed using video and eDNA data. Beyond the mere comparison of family identities and considering the limited number of species identified with eDNA, we sought to quantitatively compare the two methods using MOTUs generated with eDNA as a proxy for species. Since species identity is not accessible for most MOTUs, we used them as proxies for species using the teleo marker to make comparison at a higher taxonomic level, here the family-level. For each family, we can estimate the number of species detected by each method without requiring the assignment at the species level. Functional richness (FRic) was assessed as the proportion of the functional space occupied by a species assemblage (Villeger, Mason, & Mouillot, 2008). We generated accumulation curves on videos over recording time, that are combining all recordings according to a chronological order as if it was one long continuous video. For eDNA, we considered each of 10 filters individually, as field duplicates are not exactly similar, and arranged them in chronological order. For each pair of duplicates, so recovered as the same time, filters were randomly assigned at the first or second filter. Taxonomic richness and FE richness were computed to generate accumulation curves through time. We used the R package vegan to generate a randomized MOTU richness accumulation curve for richness and FEs, and the R package PDcalc to generate a rarefaction curve for PD. We then computed functional dissimilarity between the two census methods as the proportion of non-overlap in the functional space between the convex hulls shaping the taxa recorded by each method, as well as the contribution of turnover to this dissimilarity (Villéger, Grenouillet, & Brosse, 2013). These indices were calculated with the R package betapart (Baselga & Orme, 2012). We computed Faith's Phylogenetic Diversity (PD) of taxa censused by the two methods using the picante R package applied to 100 super-trees (Rabosky et al., 2018) pruned at the genus level for Teleostens (bony fish). All MOTUs assigned at genus or species levels were considered for PD analysis, but MOTUs assigned at family level only were not considered since the tree is pruned at genus level. Elasmobranchii taxa (sharks and rays) were not included in the phylogenetic diversity analysis.

Results

1. Biodiversity estimates

A total of 3.3 million DNA sequences passed our quality bioinformatic filters using the teleo marker, corresponding to 130 distinct MOTUs (Supplementary, table S1). Among those, 23 MOTUs were assigned to a taxonomic level higher than family (percentage of similarity < 85%) and were not included in our analyses. 22 MOTUs were assigned to deep-water fish taxa and were subsequently removed from analyses. Overall, among eDNA sequences belonging to a shallow-water taxa identified at least at the family level, 3 million sequences from 85 MOTUs were retained. As only 33 MOTUs could be assigned

to a species, we considered the lowest taxonomic assignation for each MOTU, and thus performed some analyses at the family level to allow a more representative comparison between our methods (Fig 2). Assigning FE to eDNA taxa revealed 66 distinct FEs, among which 52 were represented by only one taxa, 10 by two taxa, three by three taxa and one by four taxa (Fig 2D).

On videos, we identified 51 taxa, 50 at the species level and one at genus level (Mobula sp). Those 51 taxa were gathered into 43 functional entities (FE), with 37 FEs comprised of a single species, 5 FEs of two species and one FEs of 4 species. The combination of both methods generated 77 FEs, among which around half (33) were shared while 10 were unique to videos and 33 unique to eDNA (Fig. 2A). When considering only species-assigned MOTUs, 25 FEs were detected only on videos, 11 only by eDNA and 18 were detected by both methods (Supplementary, Fig S5).

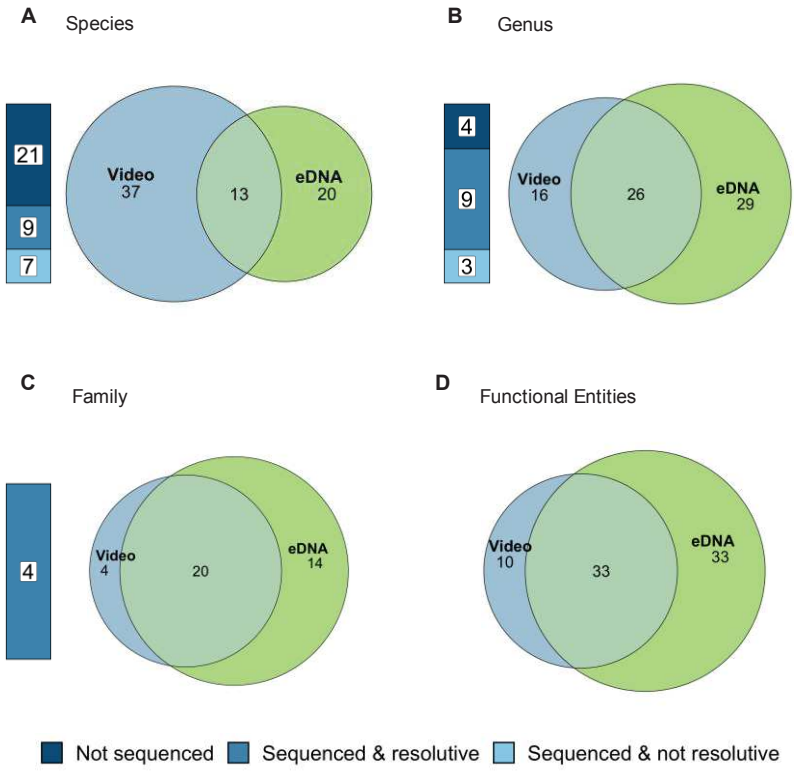


Fig. 2. Venn diagrams comparing the number of fish taxa detected by both methods at species (A), genus (B), family (C) level and the number of Functional Entities (D) for the teleo marker. The blue barplot on the left side represents the sequencing status of all species exclusively found on videos.

2. Taxonomic congruence between methods

The difference in rank-level taxa assignation between methods prevented a straightforward comparison of detected taxa, with 98% of taxa identified at the species level on videos, as opposed to 40% in eDNA.

At the species level, videos detected more diversity with 50 species against 33 with eDNA (13 shared), but only 24% (9/37) of species detected exclusively with videos were sequenced and detectable using eDNA (Fig 2). Environmental DNA with a single marker detected more genera (55) than videos (42) and also more families (34 compared to 24) (Fig. 2). For all 20 families detected with both eDNA and videos, eDNA systematically detected more or the same amount of species/MOTUs compared to videos (Fig. 3). For 13 families, the number of detected species using videos is the same as the number of detected MOTUs using eDNA. For the 7 remaining families, eDNA detected more MOTUs than species with videos. Among the 14 families detected exclusively using eDNA, we detected 5 MOTUs of Scombridae or 2 MOTUs of Gobiidae, where videos detected no species of these common families. Complementing the teleo marker with the Chond01 and Vert01 markers (Supplementary, Table S2) revealed 34 extra taxa, all bony fish (Supplementary, Table S3), including 15 at the species level. Considering all 3 markers changed the number of shared species with videos from 13 to 17, genera from 26 to 35 and families from 20 to 22 (Fig S2). In particular, multi-marker eDNA detected 46 families including 24 not detected on videos, while only 2 families were detected with videos exclusively (Scaridae and Aulostomidae). The Vert01 primer also detected 2 species of marine mammals and one unresolved taxa of marine bird (Sula sp.), which were not included in this study focusing only on bony and cartilaginous fishes.

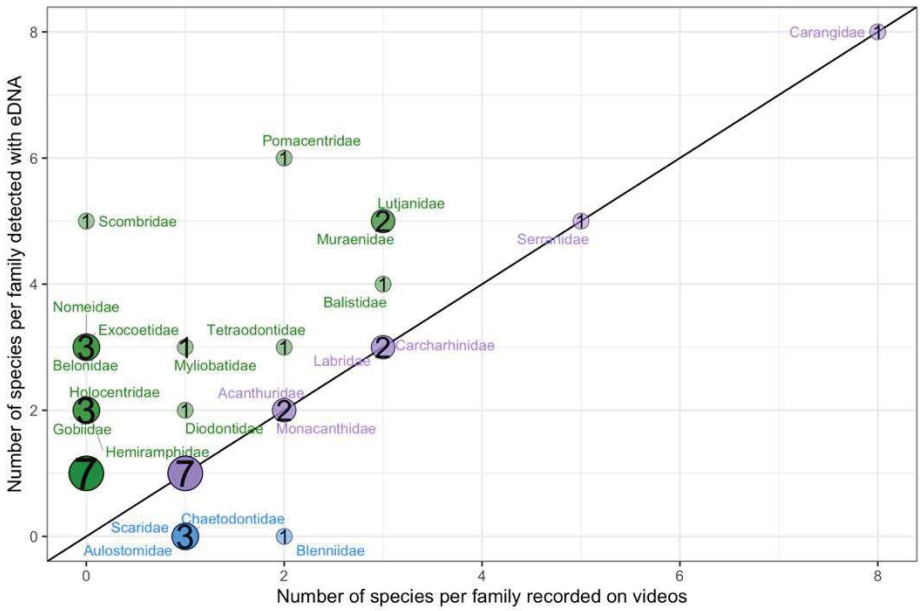


Fig. 3. Comparison of the number of MOTUs/species found at the family level, without the deep-water families using only the eDNA teleo marker. Numbers within bubbles represent the numbers of families on the same point, and colors indicate which method detects more species/MOTUs in green when eDNA performs better, blue when video performs better and purple when both methods have a similar performance. Family identities are indicated, except when more than 5 families co-occur on a same dot for graphic clarity purpose.

3. Functional and phylogenetic congruence between methods

Using both eDNA with the teleo marker and videos revealed a fish community of 71 genera, with 67 Actinopterygii genera (bony fish) representing a Faith's Phylogenetic Diversity of 4,603. The 37 genera detected with videos revealed a PD of 2,729 (so 59% of the total), while the 49 genera detected with eDNA revealed a PD of 3,767 (so 82% of total). However, 4 genera detected with videos only were not detectable using eDNA due to gaps in genetic reference database (Fig. 2). Extending the eDNA analysis to multi-markers revealed 14 extra genera, extending the entire assemblage PD to 5,322, with a PD of 4,971 for multi-marker eDNA alone (Supplementary, Fig S3), so 93% of the total PD detected using all methods and all markers, while for videos this amounts reached only 51%.

Species recorded on videos filled a smaller portion of the functional space (i.e. convex hull delimited by the most extreme combination of traits values) than the one filled by taxa recorded with eDNA (Fig. 4, Fig S4). The dissimilarity (□-diversity) between those two convex hulls equals 0.37, and turnover contributed to only 16% of this dissimilarity, highlighting that taxa recorded on videos filled mostly a subset of the space filled by taxa recorded with eDNA. The portion of the functional space filled only by eDNA is driven by a few taxa like Psene cyanophrys, Mobula tarapacana or Canthigaster jactator, whose traits correspond mainly to strict pelagic, planktivorous or small omnivorous species, respectively. The small portion of the functional space filled only by videos was due to the small invertivorous cryptobenthic blenny (Hypsoblennius maculipinna) not detected with eDNA. Further, including eDNA from all three markers revealed that the taxa recorded on videos filled a portion of the functional space completely embedded within the portion delineated by taxa detected with eDNA (Supplementary, Fig S3).

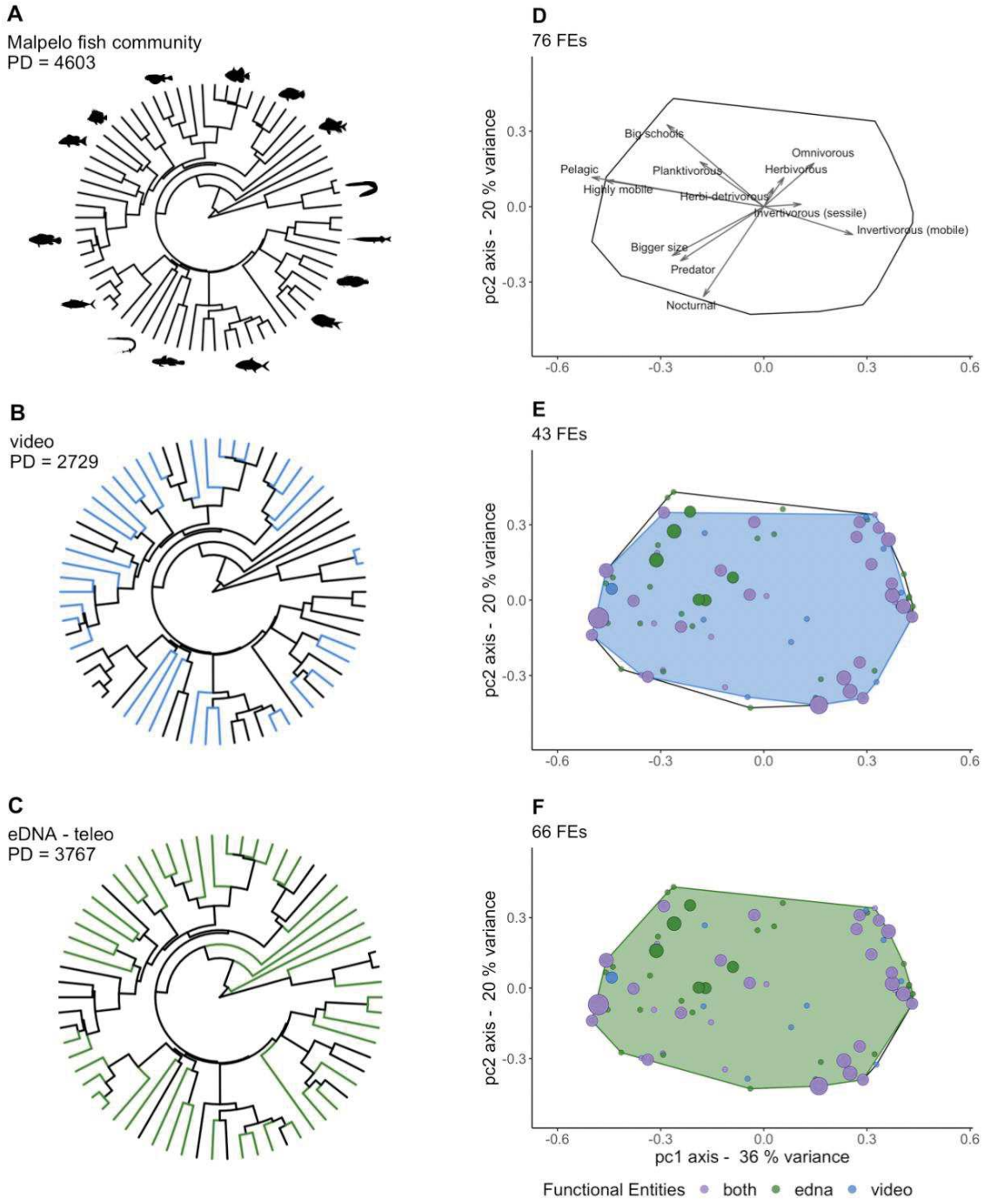


Fig. 4. Comparison of the Phylogenetic diversity and Functional diversity using eDNA and videos, with the phylogenetic tree of teleostean fish at the genus level of (A) the shallow water Malpelo assemblage using the combination of eDNA with teleo and video, (B) only the video and (C) eDNA using only the teleo marker and the functional space filling for the two first PCoA axes at the best taxonomical level for (D) the entire assemblage, (E) the video only and (F) the eDNA using only the teleo marker. Arrows indicate the traits driving each PCoA axis; they are scaled to the hull envelop extreme values. PD is reported using the Faith's Phylogenetic Diversity (PD).

4. Biodiversity accumulation curves and asymptotes

One hour of video resulted in the detection of 63% of species (32) and 70% of FEs (30) identified over the 25 hours of video (Fig 5a,c). After 2h, an additional seven species and four FEs were recorded, representing 76 % (39) of species and 81% (34) of FEs. Seven hours of video were necessary to detect 90% of all FEs detectable on videos (39 out of 43). After 25 hours of video, 56% of the complete number of FEs detected using both eDNA and videos were successfully detected (43 out of 77 in total). Six hours of video recording were necessary to detect 90% of total PD recorded on videos. For eDNA, a single transect (2 samples) was sufficient to detect 67% of MOTUs (57) and 70% of FEs (46) (Fig 5b,d). After four transects (8 samples), 93% of MOTUs (79) and 92% of FEs (61/66) were detected. Two eDNA samples detected as much Faith's PD as 25 hours of video (PD = 2,735) (Fig. 5e,f), but 10 eDNA samples using the teleo marker did not detect as much PD as 10 eDNA samples with the combination of all three markers (PD = 4,971) or the combination all methods, i.e. all eDNA primers and video combined (PD = 5322) (Supplementary, Fig. S3).

Discussion

Our study demonstrates that eDNA outperformed long-duration remote video recording for the estimation of all facets of reef fish biodiversity. In particular, eDNA revealed more MOTUs (species proxy), higher FE, FD and PD richness than 25 hours of videos. Fast and reliable estimations of biodiversity with eDNA are promising for conservation and monitoring programs, which could help to upscale current spatiotemporal extent of sampling and adopt a multifaceted perspective on reef fish biodiversity (Cinner et al., 2020).

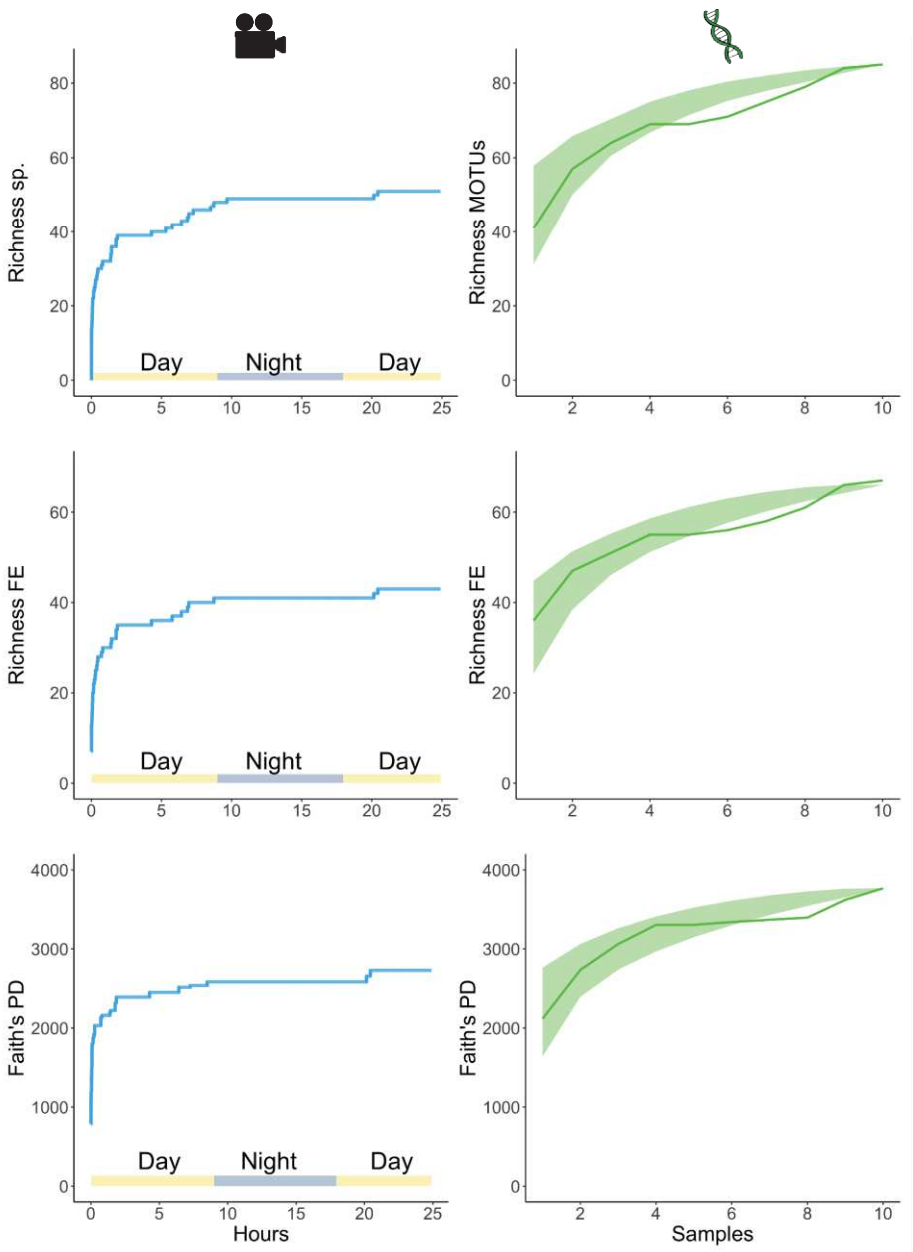


Fig. 5. Accumulation curves for (A) species richness on videos, (B) MOTU richness for eDNA, (C) richness of Functional Entities (FEs) for videos and (D) richness of FEs with eDNA and (E) Faith's Phylogenetic Diversity (PD) on videos and (F) with eDNA, using only the teleo marker for eDNA. The green ribbon represents the accumulation curve with standard deviation using a randomized saturation approach for eDNA, and the green line the real values ordered by sampling date.

Remote underwater videos and eDNA are complementary methods to census TD in reef ecosystems at the species level, but eDNA detects more genera and families. In particular, 14 families were only detected with eDNA, whereas only four were only detected on videos. This advantage is more pronounced when combining multiple primer pairs: we obtain more overlap between both methods and more taxa detected only using eDNA, with 24 families exclusive to eDNA and only two exclusives to videos. In terms of composition, both methods perform well to detect mobile yet elusive predators,

such as jacks and sharks with more taxa for eDNA. Un-baited cameras seem to perform exceptionally well for shark detection, but we argue this might be linked to the unique feature of Malpelo island being a shark gathering place (Bessudo et al., 2011; Ketchum et al., 2014). Hence, the detection probability on videos was certainly higher around Malpelo island than on average around a typical reef where sharks are scarcer and more cautious (J. B. Juhel et al., 2019), and thus where eDNA is expected to perform better than any other inventory method (Bakker et al., 2017; Boussarie et al., 2018). Our results contrast with the only other study comparing eDNA and camera-based fish census (Stat et al., 2019), where they found a clear complementarity in detection between both methods at the genus level. Such differences may come from the use of short (1h) baited cameras instead of long un-baited cameras (25h), or from a different eDNA protocol. They used 500mL point as samples with a 16S marker, whereas we used a 12S marker and sampled a larger water volume (30L) over a surface transect, which is expected to yield more detections due to eDNA particle dilution in marine environments (Thomsen et al., 2012). As expected, the reference database completeness currently impairs the use of eDNA, but our clustering approach allows to derive some TD metrics and reveals a strong potential for a fast TD census in marine fish communities once reference databases will be more populated.

We found that eDNA allows a better FD assessment compared to long-duration videos, with a higher number of FEs detected and higher Functional Richness. Despite disparities in taxonomic inventory, both methods revealed a close set of functional entities with 77% of FEs (33/43) detected on videos also detected with eDNA, and 50% (33/66) of FEs detected with eDNA also detected on videos. Using only species-assigned MOTUs revealed a total of 29 FEs, with 11 exclusives to eDNA despite a low number of species-level assignments. This hints how the higher number of FEs detected with eDNA using random assignments of traits at higher taxonomic levels is probably not an artefact but represents the potential of eDNA-based inventory if genetic reference databases were more populated. Filming for a long duration was necessary to capture most FEs as one hour of videos recovered 70% of all FEs detected with videos, seven hours were required to sample 90% of FEs. One eDNA transect with two filters detected as much FEs as 25 hours of videos, highlighting its efficiency for a fast inventory at the functional level. Environmental DNA also detected a higher Functional Richness (FRic), meaning that FEs detected exclusively using eDNA exhibited more extreme and distinct trait combinations compared to those detected only on videos. More specifically, the larger breadth of functional composition detected with eDNA is due to large pelagic piscivorous and planktivorous species which are vertices of the convex hull sharing the whole fish assemblage of the studied area. The recording of large pelagic taxa such as jack species (Carangidae) by cameras is probably due to the oceanic nature of Malpelo and the long duration video, which can capture rare events or mobile species with low abundance. Other studies on marine systems suggest that eDNA integrates a wider spatial signal at the scale of hundred meters,

enabling the detection of pelagic species rarely visiting coastal ecosystems and frequently missed by other census methods (Boussarie et al., 2018; Valdivia-Carrillo, Rocha-Olivares, Reyes-Bonilla, Domínguez-Contreras, & Munguia-Vega, 2019) but can still delineate distinct habitats (Nguyen et al., 2020; West et al., 2020). We show how eDNA inventory can go beyond TD and can measure FD on a given site without bias among functional entities, thus informing on the functioning of ecosystems with a low sampling effort.

PD complements previous facets of biodiversity by providing insight on the evolutionary history of communities. We show that even considering a single taxonomic group (bony fish) with a single marker, eDNA outperforms videos in terms of detected PD. Some lineages were detected by eDNA while completely missed by videos such as Thunnus or Lythrypnus (Gobiidae family). Further, additional markers targeting bony fishes expanded the PD detected with eDNA from 3767 to 4971, so almost two times the PD recovered on videos. As for the other biodiversity facets, we also highlight the limited sampling effort required to identify much of the PD from the community, with a single transect detecting as much diversity as the full 25 hours of video. Most video-based inventories do not film continuously for such long periods due to battery limitations and long processing time (Mallet & Pelletier, 2014), but we found that 6 hours of continuous recording is necessary to reach 90% of PD detected over 25 hours. So, shorter protocols are likely to miss rare and even mobile gregarious large species from underrepresented lineages, which can overcontribute to the overall PD. In fact, rare species are known to be more distinct both functionally and phylogenetically compared to their more common counterparts (Mi et al., 2012; Mouillot, Bellwood, et al., 2013), showing the importance of developing methods able to census them accurately. In a context of unprecedented global changes, phylogenetically diverse communities could host a stronger 'evolutionary potential' to better adapt to new conditions (Lavergne, Mouquet, Thuiller, & Ronce, 2010; Winter, Devictor, & Schweiger, 2013). It is then crucial for any monitoring method to accurately measure the full evolutionary diversity of a community, in order to rapidly implement conservation measures (Pollock, Thuiller, & Jetz, 2017) or track global change effects (Monnet et al., 2014).

The main limitations currently impairing the large-scale deployment of eDNA metabarcoding are the reference database coverage and marker resolution, as frequently mentioned by other studies (J. Juhel et al., 2020; Marques et al., 2020; Miya et al., 2015). Only up to 13% of all fish species are currently sequenced using our teleo 12S marker and alternative marker locations with larger reference sequences, on the COI for example, are currently not appropriate for fish inventory due to impossibility of designing fish-specific primers without amplification bias, which results in poor performances (Collins et al., 2019; Deagle, Jarman, Coissac, Pompanon, & Taberlet, 2014). Lack of taxonomical resolution

happens when distinct species share the same sequence, which can result in misidentification and underestimation of biodiversity. For example, in the present study, the detection of Carcharhinus obscurus is probably a misidentification of the species Carcharhinus galapagensis as they are phylogenetically close and C. galapengensis was seen on videos while its barcode sequence is still unavailable. In this study, TD overlap between methods increases with taxonomic level (i.e. more overlap at the family than the species level) due to gaps in genetic reference database, in accordance with previous eDNA studies (Valdivia-Carrillo et al., 2019). If a sequence does not match a referenced species, its genus or family can still be identified, explaining why we found a clear advantage for eDNA at higher taxonomic levels. It shows an important potential for a broader inventory which is currently impaired by a lack of reference sequences. Other biodiversity facets are also impacted by this limitation, where FD and PD could be better estimated if more MOTUs were identified at species level. Additional markers targeting the same taxonomical groups further expanded all measures of biodiversity, likely due to complementary reference database, although we can expect this advantage to fade in the medium term as reference databases expand. This finding reflects the potential of single-marker eDNA metabarcoding with larger genetic reference databases, although multiple-marker eDNA could still be of interest to overcome the limitation of marker resolution.

Conclusion and perspectives

Biodiversity measures should not focus solely on species richness, but also on ecological functions provided by organisms and the phylogenetic diversity supported by diverse evolutionary lineages (Cadotte, Dinnage, & Tilman, 2012; Diniz-Filho et al., 2013), as ecosystem functioning can be largely altered without a strong impact on taxonomic diversity (D'Agata et al., 2014). Our results suggest that a multifaceted approach of biodiversity is feasible using eDNA metabarcoding, which delivers a faster and more exhaustive inventory than long-duration videos for all three facets of fish diversity. However, we highlight that video-based and eDNA methods can be complementary given the current state of genetic database completeness which prevents most species-level assignments. Further, videos can record fish size and behavior thus ecological functions such as grazing pressure (Puk, Marshell, Dwyer, Evensen, & Mumby, 2020).

By better estimating biodiversity in its multifaceted nature, eDNA reveals an important potential for monitoring and conservation strategies, where fast and accurate measures of diversity are required. An earlier detection of erosion of each biodiversity facet would help to set appropriate protection measures on monitored ecosystems, as well as providing a better understanding of the structure and functioning of communities (Benkwitt, Wilson, & Graham, 2020).

Acknowledgments

We thank the Yersin crew for assistance with at-sea operations, SPYGEN staff for assistance in the eDNA laboratory, The National Parks of Colombia and the Navy of Colombia for the permits, The Malpelo Foundation for the coordination of the expedition, the Monaco explorations for funding fieldwork and sequencing, Dr Quimbayo for help in taxonomical identification on video, Camille Albouy and David Eme for help with phylogenetical trees pruning. **Conflicts of interests** A.V and T.D work for an eDNA species detection company.

Data availability

The metabarcoding clustering pipeline is freely available freely on GitHub https://gitlab.mbb.univ-montp2.fr/edna/snakemake_rapidrun_swarm v1.0.0, and sequencing data outputs will be deposited on Dryad.

References

Bakker, J., Wangensteen, O. S., Chapman, D. D., Boussarie, G., Buddo, D., Guttridge, T. L., … Mariani, S. (2017). Environmental DNA reveals tropical shark diversity in contrasting levels of anthropogenic impact. Scientific Reports, 7(1), 1–11. doi: 10.1038/s41598-017-17150-2

Baselga, A., & Orme, C. D. L. (2012). Betapart: An R package for the study of beta diversity. Methods in Ecology and Evolution, 3(5), 808–812. doi: 10.1111/j.2041-210X.2012.00224.x

Benkwitt, C. E., Wilson, S. K., & Graham, N. A. J. (2020). Biodiversity increases ecosystem functions despite multiple stressors on coral reefs. Nature Ecology and Evolution, 4(7), 919–926. doi: 10.1038/s41559-020-1203-9

Bessudo, S., Soler, G. A., Klimley, A. P., Ketchum, J. T., Hearn, A., & Arauz, R. (2011). Residency of the scalloped hammerhead shark (Sphyrna lewini) at Malpelo Island and evidence of migration to other islands in the Eastern Tropical Pacific. Environmental Biology of Fishes, 91(2), 165–176. doi: 10.1007/s10641-011-9769-3

Blowes, S. A., Supp, S. R., Antão, L. H., Bates, A., Bruelheide, H., Chase, J. M., … Dornelas, M. (2019). The geography of biodiversity change in marine and terrestrial assemblages. Science, 366(6463), 339–345. doi: 10.1126/science.aaw1620

Boettiger, C., Lang, D. T., & Wainwright, P. C. (2012). Rfishbase: Exploring, manipulating and visualizing FishBase data from R. Journal of Fish Biology, 81(6), 2030–2039. doi: 10.1111/j.1095-8649.2012.03464.x

Bosch, N. E., Gonçalves, J. M. S., Erzini, K., & Tuya, F. (2017). "How" and "what" matters: Sampling method affects biodiversity estimates of reef fishes. Ecology and Evolution, 7(13), 4891–4906. doi: 10.1002/ece3.2979

Boussarie, G., Bakker, J., Wangensteen, O. S., Mariani, S., Bonnin, L., Juhel, J. B., … Mouillot, D. (2018). Environmental DNA illuminates the dark diversity of sharks. Science Advances, 4(5), eaap9661. doi: 10.1126/sciadv.aap9661

Boyer, F., Mercier, C., Bonin, A., Bras, Y. Le, Taberlet, P., & Coissac, E. (2016). OBITOOLS: a UNIX-inspired software package for DNA metabarcoding. Molecular Ecology Resources, 16(4), 176–182. doi: 10.1111/1755-0998.12428

Brun, P., Zimmermann, N. E., Graham, C. H., Lavergne, S., Pellissier, L., Münkemüller, T., & Thuiller, W. (2019). The productivity-biodiversity relationship varies across diversity dimensions. Nature Communications, 10(1), 5691. doi: 10.1038/s41467-019-13678-1

Cadotte, M. W., Dinnage, R., & Tilman, D. (2012). Phylogenetic diversity promotes ecosystem stability. Ecology, 93(8 SPEC. ISSUE), 223–233. doi: 10.1890/11-0426.1

Cardoso, P., Rigal, F., Borges, P. A. V., & Carvalho, J. C. (2014). A new frontier in biodiversity inventory: A proposal for estimators of phylogenetic and functional diversity. Methods in Ecology and Evolution, 5(5), 452–461. doi: 10.1111/2041-210X.12173

Chasqui Velasco, L., Gil Agudelo, D. L., & Nieto, R. (2016). Endemic Shallow Reef Fishes From Malpelo Island: Abundance and Distribution. Bulletin of Marine and Coastal Research, 40(1096), 107–116. doi: 10.25268/bimc.invemar.2011.40.0.134

Cinner, J. E., Zamborain-Mason, J., Gurney, G. G., Graham, N. A. J., MacNeil, M. A., Hoey, A. S., … Mouillot, D. (2020). Meeting fisheries, ecosystem function, and biodiversity goals in a human-dominated world. Science, 368(6488), 307–311. doi: 10.1126/science.aax9412

Collins, R. A., Bakker, J., Wangensteen, O. S., Soto, A. Z., Corrigan, L., Sims, D. W., … Mariani, S. (2019). Non-specific amplification compromises environmental DNA metabarcoding with COI. Methods in Ecology and Evolution, 10(11), 1985–2001. doi: 10.1111/2041-210X.1

Colton, M. A., & Swearer, S. E. (2010). A comparison of two survey methods: Differences between underwater visual census and baited remote underwater video. Marine Ecology Progress Series, 400(February), 19–36. doi: 10.3354/meps08377

Craven, D., Eisenhauer, N., Pearse, W. D., Hautier, Y., Isbell, F., Roscher, C., … Manning, P. (2018). Multiple facets of biodiversity drive the diversity–stability relationship. Nature Ecology and Evolution, 2(10), 1579–1587. doi: 10.1038/s41559-018-0647-7

D'Agata, S., Mouillot, D., Kulbicki, M., Andréfouët, S., Bellwood, D. R., Cinner, J. E., … Vigliola, L. (2014). Human-mediated loss of phylogenetic and functional diversity in coral reef fishes. Current Biology, 24(5), 555–560. doi: 10.1016/j.cub.2014.01.049

Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., & Taberlet, P. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. Biology Letters, 10(9), 20140562–20140562. doi: 10.1098/rsbl.2014.0562

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., … Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. Molecular Ecology, 26(21), 5872–5895. doi: 10.1111/mec.14350

Devictor, V., Mouillot, D., Meynard, C., Jiguet, F., Thuiller, W., & Mouquet, N. (2010). Spatial mismatch and congruence between taxonomic, phylogenetic and functional diversity: The need for integrative conservation strategies in a changing world. Ecology Letters, 13(8), 1030–1040. doi: 10.1111/j.1461-0248.2010.01493.x

Dickens, L. C., Goatley, C. H. R., Tanner, J. K., & Bellwood, D. R. (2011). Quantifying relative diver effects in underwater visual censuses. PLoS ONE, 6(4), 6–8. doi: 10.1371/journal.pone.0018965

Diniz-Filho, J. A. F., Loyola, R. D., Raia, P., Mooers, A. O., & Bini, L. M. (2013). Darwinian shortfalls in biodiversity conservation. Trends in Ecology and Evolution, 28(12), 689–695. doi: 10.1016/j.tree.2013.09.003

Edgar, G. J., Banks, S. A., Bessudo, S., Cortés, J., Guzmán, H. M., Henderson, S., … Zapata, F. A. (2011). Variation in reef fish and invertebrate communities with level of protection from fishing across the Eastern Tropical Pacific seascape. Global Ecology and Biogeography, 20(5), 730–743. doi: 10.1111/j.1466-8238.2010.00642.x

Forest, F., Grenyer, R., Rouget, M., Davies, T. J., Cowling, R. M., Faith, D. P., … Savolainen, V. (2007). Preserving the evolutionary potential of floras in biodiversity hotspots. Nature, 445(7129), 757–760. doi: 10.1038/nature05587

Froese, R., & Pauly, D. (2000). FishBase 2000: concepts, design and data sources (Editors, ed.). ICLARM, Los Baños, Laguna, Philippines.

Harrison, J. B., Sunday, J. M., & Rogers, S. M. (2019). Predicting the fate of eDNA in the environment and implications for studying biodiversity. Proceedings of the Royal Society B: Biological Sciences, 286(1915), 20191409. doi: 10.1098/rspb.2019.1409

Jarzyna, M. A., & Jetz, W. (2016). Detecting the Multiple Facets of Biodiversity. Trends in Ecology and Evolution, 31(7), 527–538. doi: 10.1016/j.tree.2016.04.002

Juhel, J. B., Vigliola, L., Wantiez, L., Letessier, T. B., Meeuwig, J. J., & Mouillot, D. (2019). Isolation and no-entry marine reserves mitigate anthropogenic impacts on grey reef shark behavior. Scientific Reports, 9(1). doi: 10.1038/s41598-018-37145-x

Juhel, J., Utama, R. S., Marques, V., Vimono, I. B., Sugeha, H. Y., Kadarusman, … Hocdé, R. (2020). Accumulation curves of environmental DNA sequences predict coastal fish diversity in the coral triangle. Proceedings of the Royal Society B, 20200248.

Ketchum, J. T., Hearn, A., Klimley, A. P., Peñaherrera, C., Espinoza, E., Bessudo, S., … Arauz, R. (2014). Inter-island movements of scalloped hammerhead sharks (Sphyrna lewini) and seasonal connectivity in a marine protected area of the eastern tropical Pacific. Marine Biology, 161(4), 939–951. doi: 10.1007/s00227-014-2393-y

Kling, M. M., Mishler, B. D., Thornhill, A. H., Baldwin, B. G., & Ackerly, D. D. (2019). Facets of phylodiversity: Evolutionary diversification, divergence and survival as conservation targets. Philosophical Transactions of the Royal Society B: Biological Sciences, 374(1763). doi: 10.1098/rstb.2017.0397

Langlois, T. J., Harvey, E. S., Fitzpatrick, B., Meeuwig, J. J., Shedrawi, G., & Watson, D. L. (2010). Cost-efficient sampling of fish assemblages: Comparison of baited video stations and diver video transects. Aquatic Biology, 9(2), 155–168. doi: 10.3354/ab00235

Lavergne, S., Mouquet, N., Thuiller, W., & Ronce, O. (2010). Biodiversity and Climate Change: Integrating Evolutionary and Ecological Responses of Species and Communities. Annual Review of Ecology, Evolution, and Systematics, 41(1), 321–350. doi: 10.1146/annurev-ecolsys-102209-144628

Legendre, P., & Legendre, L. (1998). Numerical ecology (2nd Editio; E. Science, ed.). Amsterdam.

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., … Cochrane, G. (2011). The European nucleotide archive. Nucleic Acids Research, 39(SUPPL. 1), 44–47. doi: 10.1093/nar/gkq967

MacConaill, L. E., Burns, R. T., Nag, A., Coleman, H. A., Slevin, M. K., Giorda, K., … Thorner, A. R. (2018). Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. BMC Genomics, 19(1), 1–10. doi: 10.1186/s12864-017-4428-5

MacNeil, M. A., Graham, N. A. J., Conroy, M. J., Fonnesbeck, C. J., Polunin, N. V. C., Rushton, S. P., … McClanahan, T. R. (2008). Detection heterogeneity in underwater visual-census data. Journal of Fish Biology, 73(7), 1748–1763. doi: 10.1111/j.1095-8649.2008.02067.x

Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. PeerJ, 3, e1420. doi: 10.7717/peerj.1420

Maire, E., Grenouillet, G., Brosse, S., & Villéger, S. (2015). How many dimensions are needed to accurately assess functional diversity? A pragmatic approach for assessing the quality of functional spaces. Global Ecology and Biogeography, 24(6), 728–740. doi: 10.1111/geb.12299

Mallet, D., & Pelletier, D. (2014). Underwater video techniques for observing coastal marine biodiversity: A review of sixty years of publications (1952-2012). Fisheries Research, 154, 44–62. doi: 10.1016/j.fishres.2014.01.019

Marques, V., Guérin, P. É., Rocle, M., Valentini, A., Manel, S., Mouillot, D., & Dejean, T. (2020). Blind assessment of vertebrate taxonomic diversity across spatial scales by clustering environmental DNA metabarcoding sequences. Ecography, 1–12. doi: 10.1111/ecog.05049

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal, 17(1), 10. doi: 10.14806/ej.17.1.200

Mazel, F., Pennell, M. W., Cadotte, M. W., Diaz, S., Dalla Riva, G. V., Grenyer, R., … Pearse, W. D. (2018). Prioritizing phylogenetic diversity captures functional diversity unreliably. Nature Communications, 9(1). doi: 10.1038/s41467-018-05126-3

Mbaru, E. K., Graham, N. A. J., McClanahan, T. R., & Cinner, J. E. (2020). Functional traits illuminate the selective impacts of different fishing gears on coral reefs. Journal of Applied Ecology, 57(2), 241–252. doi: 10.1111/1365-2664.13547

McClanahan, T. R., Graham, N. A. J., Maina, J., Chabanet, P., Bruggemann, J. H., & Polunin, N. V. C. (2007). Influence of instantaneous variation on estimates of coral reef fish populations and communities. Marine Ecology Progress Series, 340, 221–234. doi: 10.3354/meps340221

McGill, B. J., Enquist, B. J., Weiher, E., & Westoby, M. (2006). Rebuilding community ecology from functional traits. Trends in Ecology and Evolution, 21(4), 178–185. doi: 10.1016/j.tree.2006.02.002

Mi, X., Swenson, N. G., Valencia, R., John Kress, W., Erickson, D. L., Pérez, Á. J., … Ma, K. (2012). The contribution of rare species to community phylogenetic diversity across a global network of forest plots. American Naturalist, 180(1), 17–30. doi: 10.1086/665999

Miya, M., Sato, Y., Fukunaga, T., Sado, T., Poulsen, J. Y., Sato, K., … Iwasaki, W. (2015). MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than 230 subtropical marine species. Royal Society Open Science, 2(7), 150088. doi: 10.1098/rsos.150088

Monnet, A. C., Jiguet, F., Meynard, C. N., Mouillot, D., Mouquet, N., Thuiller, W., & Devictor, V. (2014). Asynchrony of taxonomic, functional and phylogenetic diversity in birds. Global Ecology and Biogeography, 23(7), 780–788. doi: 10.1111/geb.12179

Mouillot, D., Bellwood, D. R., Baraloto, C., Chave, J., Galzin, R., Harmelin-Vivien, M., … Thuiller, W. (2013). Rare Species Support Vulnerable Functions in High-Diversity Ecosystems. PLoS Biology, 11(5). doi: 10.1371/journal.pbio.1001569

Mouillot, D., Graham, N. A. J., Villéger, S., Mason, N. W. H., & Bellwood, D. R. (2013). A functional approach reveals community responses to disturbances. Trends in Ecology and Evolution, 28(3), 167–177. doi: 10.1016/j.tree.2012.10.004

Mouillot, D., Villéger, S., Parravicini, V., Kulbicki, M., Arias-González, J. E., Bender, M., … Bellwood, D. R. (2014). Functional over-redundancy and high functional vulnerability in global fish faunas on tropical reefs. Proceedings of the National Academy of Sciences of the United States of America, 111(38), 13757–13762. doi: 10.1073/pnas.1317625111

Nalesso, E., Hearn, A., Sosa-Nishizaki, O., Steiner, T., Antoniou, A., Reid, A., … Arauz, R. (2019). Movements of scalloped hammerhead sharks (Sphyrna lewini) at Cocos Island, Costa Rica and between oceanic islands in the Eastern Tropical Pacific. PLoS ONE, 14(3), 1–16. doi: 10.1371/journal.pone.0213741

Nguyen, B. N., Shen, E. W., Seemann, J., Correa, A. M. S., O'Donnell, J. L., Altieri, A. H., … Leray, M. (2020). Environmental DNA survey captures patterns of fish and invertebrate diversity across a tropical seascape. Scientific Reports, 10(6729). doi: 10.1101/797712

Pollock, L. J., Thuiller, W., & Jetz, W. (2017). Large conservation gains possible for global biodiversity facets. Nature, 546(7656), 141–144. doi: 10.1038/nature22368

Pont, D., Rocle, M., Valentini, A., Civade, R., Jean, P., Maire, A., … Dejean, T. (2018). Environmental DNA reveals quantitative patterns of fish biodiversity in large rivers despite its downstream transportation. Scientific Reports, 8(1), 1–13. doi: 10.1038/s41598-018-28424-8

Puk, L. D., Marshell, A., Dwyer, J., Evensen, N. R., & Mumby, P. J. (2020). Refuge-dependent herbivory controls a key macroalga on coral reefs. Coral Reefs, 39(4), 953–965. doi: 10.1007/s00338-020-01915-9

Quimbayo, J. P., Mendes, T. C., Kulbicki, M., Floeter, S. R., & Zapata, F. A. (2017). Unusual reef fish biomass and functional richness at Malpelo , a remote island in the Tropical Eastern Pacific. Environmental Biology of Fishes, 149–162. doi: 10.1007/s10641-016-0557-y

Rodríguez-Rubio, E., Schneider, W., & del Río, R. A. (2003). On the seasonal circulation within the Panama Bight derived from satellite observations of wind, altimetry and sea surface temperature. Geophysical Research Letters, 30(7). doi: 10.1029/2002GL016794

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. PeerJ, 4, e2584. doi: 10.7717/peerj.2584

Schnell, I. B., Bohmann, K., & Gilbert, M. T. P. (2015). Tag jumps illuminated - reducing sequence-to-sample misidentifications in metabarcoding studies. Molecular Ecology Resources, 15(6), 1289–1303. doi: 10.1111/1755-0998.12402

Segovia, R. A., Pennington, R. T., Baker, T. R., Coelho de Souza, F., Neves, D. M., Davis, C. C., … Dexter, K. G. (2020). Freezing and water availability structure the evolutionary diversity of trees across the Americas. Science Advances, 6(19), eaaz5373. doi: 10.1126/sciadv.aaz5373

Stat, M., John, J., DiBattista, J. D., Newman, S. J., Bunce, M., & Harvey, E. S. (2019). Combined use of eDNA metabarcoding and video surveillance for the assessment of fish biodiversity. Conservation Biology, 33(1), 196–205. doi: 10.1111/cobi.13183

Taberlet, P., Bonin, A., Coissac, E., & Zinger, L. (2018). Environmental DNA: For Biodiversity Research and Monitoring (Oxford Uni).

Thomsen, P. F., Kielgast, J., Iversen, L. L., M??ller, P. R., Rasmussen, M., & Willerslev, E. (2012). Detection of a Diverse Marine Fish Fauna Using Environmental DNA from Seawater Samples. PLoS ONE, 7(8), 1–9. doi: 10.1371/journal.pone.0041732

Trindade-Santos, I., Moyes, F., & Magurran, A. E. (2020). Global change in the functional diversity of marine fisheries exploitation over the past 65 years. Proceedings. Biological Sciences, 287(1933), 20200889. doi: 10.1098/rspb.2020.0889

Tucker, C. M., Aze, T., Cadotte, M. W., Cantalapiedra, J. L., Chisholm, C., Díaz, S., … Mooers, A. O. (2019). Assessing the utility of conserving evolutionary history. Biological Reviews, 94(5), 1740–1760. doi: 10.1111/brv.12526

Tucker, C. M., Cadotte, M. W., Carvalho, S. B., Davies, T. J., Ferrier, S., Fritz, S. A., … Mazel, F. (2016). A guide to phylogenetic metrics for conservation, community ecology and macroecology. Biological Reviews, 92, 698–715. doi: 10.1111/brv.12252

Valdivia-Carrillo, T., Rocha-Olivares, A., Reyes-Bonilla, H., Domínguez-Contreras, J. F., & Munguia-Vega, A. (2019). Beyond traditional biodiversity fish monitoring: environmental DNA metabarcoding and simultaneous underwater visual census detect different sets of a complex fish community at a marine biodiversity hotspot. BioRxiv.

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., … Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. Molecular Ecology, 25(4), 929–942. doi: 10.1111/mec.13428

Villéger, S., Brosse, S., Mouchet, M., Mouillot, D., & Vanni, M. J. (2017). Functional ecology of fish: current approaches and future challenges. Aquatic Sciences, 79(4), 783–801. doi: 10.1007/s00027-017-0546-z

Villéger, S., Grenouillet, G., & Brosse, S. (2013). Decomposing functional β-diversity reveals that low functional β-diversity is driven by low functional turnover in European fish assemblages. Global Ecology and Biogeography, 22(6), 671–681. doi: 10.1111/geb.12021

Villeger, S., Mason, N. W. H., & Mouillot, D. (2008). NEW MULTIDIMENSIONAL FUNCTIONAL DIVERSITY INDICES FOR A MULTIFACETED FRAMEWORK IN FUNCTIONAL ECOLOGY. Ecology, 89(9), 2290–2301. doi: 10.1002/chin.200826189

Webb, C. O., Ackerly, D. D., McPeek, M. A., & Donoghue, M. J. (2002). Phylogenies and community ecology. Annual Review of Ecology and Systematics, 33, 475–505. doi: 10.1146/annurev.ecolsys.33.010802.150448

West, K. M., Stat, M., Harvey, E. S., Skepper, C. L., DiBattista, J. D., Richards, Z. T., … Bunce, M. (2020). eDNA metabarcoding survey reveals fine-scale coral reef community variation across a remote, tropical island ecosystem. Molecular Ecology, (February 2019), 1–18. doi: 10.1111/mec.15382

Wetz, J. J., Ajemian, M. J., Shipley, B., & Stunz, G. W. (2020). An assessment of two visual survey methods for documenting fish community structure on artificial platform reefs in the Gulf of Mexico. Fisheries Research, 225(January), 105492. doi: 10.1016/j.fishres.2020.105492

Winter, M., Devictor, V., & Schweiger, O. (2013). Phylogenetic diversity and nature conservation: Where are we? Trends in Ecology and Evolution, 28(4), 199–204. doi: 10.1016/j.tree.2012.10.015

3. Manuscrit D

Received: 4 April 2020 | Revised: 6 August 2020 | Accepted: 9 September 2020

DOI: 10.1002/edn3.140

SPECIAL ISSUE ORIGINAL ARTICLE

Environmental DNA
Dedicated to the study and use of environmental DNA for basic and applied sciences
WILEY

Comparing environmental DNA metabarcoding and underwater visual census to monitor tropical reef fishes

Andrea Polanco Fernández[1] | Virginie Marques[2,3] | Fabian Fopp[4,5] | Jean-Baptiste Juhel[2] | Giomar Helena Borrero-Pérez[1] | Marie-Charlotte Cheutin[2] | Tony Dejean[6] | Juan David González Corredor[1] | Andrés Acosta-Chaparro[1] | Régis Hocdé[2] | David Eme[7] | Eva Maire[2,8] | Manuel Spescha[4,5] | Alice Valentini[6] | Stéphanie Manel[3] | David Mouillot[2] | Camille Albouy[7] | Loïc Pellissier[4,5]

[1]Programa de Biodiversidad y Ecosistemas Marinos, Museo de Historia Natural Marina de Colombia (MHNMC), Instituto de Investigaciones Marinas y Costeras-INVEMAR, Santa Marta, Colombia

[2]MARBEC, CNRS, Ifremer, IRD, University of Montpellier, Montpellier, France

[3]EPHE, CNRS, UM, UM3, IRD, UMR5175 CEFE, PSL Research University, Montpellier, France

[4]Landscape Ecology, Institute of Terrestrial Ecosystems, Department of Environmental Systems Science, ETH Zürich, Zürich, Switzerland

[5]Unit of Land Change Science, Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Birmensdorf, Switzerland

[6]SPYGEN, Le Bourget-du-Lac, France

[7]Unité Ecologie et Modèles pour l'Halieutique, EMH, IFREMER, Nantes, France

[8]Lancaster Environment Centre, Lancaster University, Lancaster, UK

Correspondence
Andrea Polanco Fernández, Programa de Biodiversidad y Ecosistemas Marinos, Museo de Historia Natural Marina de Colombia (MHNMC), Instituto de Investigaciones Marinas y Costeras-INVEMAR, Calle 25 No.2—55 Playa Salguero, Santa Marta, Colombia.
Email: andrea.polanco@gmail.com

Abstract

Environmental DNA (eDNA) analysis is a revolutionary method to monitor marine biodiversity from animal DNA traces. Examining the capacity of eDNA to provide accurate biodiversity measures in species-rich ecosystems such as coral reefs is a prerequisite for their application in long-term monitoring. Here, we surveyed two Colombian tropical marine reefs, the island of Providencia and Gayraca Bay near Santa Marta, using eDNA and underwater visual census (UVC) methods. We collected a large quantity of surface water (30 L per filter) above the reefs and applied a metabarcoding protocol using three different primer sets targeting the 12S mitochondrial DNA, which are specific to the vertebrates Actinopterygii and Elasmobranchii. By assigning eDNA sequences to species using a public reference database, we detected the presence of 107 and 85 fish species, 106 and 92 genera, and 73 and 57 families in Providencia and Gayraca Bay, respectively. Of the species identified using eDNA, 32.7% (Providencia) and 18.8% (Gayraca) were also found in the UVCs. We further found congruence in genus and species richness and abundance between eDNA and UVC approaches in Providencia but not in Gayraca Bay. Mismatches between eDNA and UVC had a phylogenetic and ecological signal, with eDNA detecting a broader phylogenetic diversity and more effectively detecting smaller species, pelagic species and those in deeper habitats. Altogether, eDNA can be used for fast and broad biodiversity surveys and is applicable to species-rich ecosystems in the tropics, but improved coverage of the reference database is required before this new method could serve as an effective complement to traditional census methods.

Andrea Polanco Fernández and Virginie Marques shared first authorship.

Camille Albouy and Loïc Pellissier shared senior authorship.

This research was financed by the Monaco Explorations Foundation grant "Megafauna" to DM, SM, and LP. The research was further supported by the Swiss National Science Foundation grant n°310030E-164294 to LP.

Funding information
Monaco Explorations Foundation grant "Megafauna"; Swiss National Science Foundation grant n° 310030E-164294

KEYWORDS
biodiversity, biomonitoring, Caribbean Sea, environmental DNA, reef fishes, underwater visual census

1 | INTRODUCTION

Coral reefs represent the most diverse marine ecosystems on the planet (Fisher et al., 2015) and are also the most threatened (Williams et al., 2019). Due to their structural complexity, they host a large diversity of fish species, from tiny cryptic species to large migratory species (Collins et al., 2019; Darling et al., 2017). Because of this high species diversity, coral reefs have generally been difficult to inventory using traditional survey methods (Plaisance et al., 2011). Moreover, global changes, including exploitation, pollution, or climate change, are degrading biodiversity on reefs (Cinner et al., 2016; Descombes et al., 2015), but it is difficult to quantify and monitor these impacts because describing species diversity and composition is generally demanding (Costello et al., 2015; Mora et al., 2008). The monitoring of the biodiversity of coral reefs under global changes could benefit from novel solutions with lower costs and broader applicability complementing traditional methods (Thomsen et al., 2012; West et al., 2020).

Traditionally, monitoring fishes on coral reefs has been performed using underwater visual censuses (UVC) or video surveys (Stat et al., 2019), which offer a partial view of the dynamics of reef biodiversity, from their degradation under global changes to their recovery (Bozec et al., 2011; Cinner et al., 2016). These methods are limited in both spatial and temporal coverage and are biased toward certain categories of species (Boussarie et al., 2018). UVC is traditionally used to monitor fish diversity on coral reefs (Samoilys & Carlos, 2000). However, besides logistical difficulties to organize underwater sampling in remote locations, UVC can suffer from several observer biases, such as overlooking cryptobenthic (Bozec et al., 2011) or wideranged species such as sharks (Juhel et al., 2018). One of the most effective approaches to circumvent the limitations of traditional survey methods in highly diverse ecosystems is environmental DNA (eDNA) metabarcoding (Cilleros et al., 2019; Gomes et al., 2017). eDNA is a noninvasive method demonstrating higher detection capabilities and cost-effectiveness compared to traditional methods, especially when deployed in remote locations (Dejean et al., 2011; Kelly et al., 2014; Thomsen & Willerslev, 2015). Before it can effectively complement traditional sampling methods, the ability of eDNA to recover signals of diversity and composition of marine systems should be evaluated.

Animals leave DNA traces in the environment (Deiner et al., 2017), which may persist from hours to days and can be detected in water samples (Collins et al., 2019; Thomsen et al., 2012). Water filtering followed by a molecular protocol to amplify and sequence target DNA can be used to recover animal DNA present in a given site. Sequences are then taxonomically assigned using a genetic reference database, which provides an integrative inventory of species and composition in aquatic systems (Deiner et al., 2017; Harrison et al., 2019). A recent synthesis counted 54 papers on tropical eDNA, whereas only 15 focused on marine systems (Bakker et al., 2019; Huerlimann et al., 2020; Sigsgaard et al., 2019; West et al., 2020). Compared to freshwater systems, the marine environment has a larger water volume to fish biomass ratio, the movement of molecules in suspension is influenced by various currents, and reef systems can contain up to hundreds of species, which might challenge the detection of individual species (Collins et al., 2019; Hansen et al., 2018; Harrison et al., 2019). Several applications demonstrated that eDNA can recover multiple components of marine ecosystems, including species richness (Jerde et al., 2019), seasonal composition variation (Djurhuus et al., 2020), rare species (Weltz et al., 2017), abundance or biomass (Knudsen et al., 2019; Thomsen et al., 2016), and the occurrence of invasive species (Nevers et al., 2018). Nevertheless, a range of methodological challenges still hampers the broad use of eDNA for the reliable monitoring of marine ecosystems, linked to the choice of markers (Collins et al., 2019; Freeland, 2017), primers sets (Stat et al., 2017), laboratory and sequencing protocols (Deiner et al., 2017; Goldberg et al., 2016), and bioinformatic analyses (Calderón-Sanou et al., 2020; Juhel et al., 2020), which implies further testing of the eDNA methodology in situ.

Tropical ecosystems have historically been underrepresented in research (Collen et al., 2008), and increased monitoring efforts in these regions are urgently needed, particularly under ongoing global change (Barlow et al., 2018). Different abiotic conditions and high species richness might challenge the application of eDNA in the tropics (Huerlimann et al., 2020; Jerde et al., 2019). Studies of eDNA on coral reefs have shown a strong potential for biodiversity detection (Nguyen et al., 2020; Sigsgaard et al., 2019; West et al., 2020), but the scope of methodological testing remains narrow. Dibattista et al. (2017) used fish-specific 16S mitochondrial DNA to monitor fish diversity in the Red Sea, but captured only a fraction of the local fish species pool. Stat et al. (2019) compared the signal of eDNA with observations from baited videos and detected >30% more generic richness using the combination of approaches than when either method was used alone. Sigsgaard et al. (2019) used eDNA with fish-specific 12S mitochondrial DNA across a network of sites in the Gulf of Oman and recovered sequences from a diverse assemblage of marine vertebrates, which covered approximately one-third of the bony fish genera previously recorded in this area. Using a combination of markers, West et al. (2020) detected a wide range of organisms and showed that their composition varied significantly between habitats across an entire island in the Coral Sea. Hence, attempts to survey tropical marine fish assemblages using eDNA are yielding increasingly informative results, supporting the use of seawater to trace the molecular signatures of biodiversity for monitoring purposes.

Here, we compared the compositional patterns of the fish community using eDNA metabarcoding and UVCs in two different reef ecosystems in the Colombian Caribbean, the oceanic island of Providencia and Gayraca Bay in the Tayrona National Natural Park near Santa Marta. We investigated (a) whether the species recovered with three different sets of 12S primers are complementary and consistent with species recovered with UVC; (b) whether there is a correspondence between species richness within each genus and family recovered using both eDNA and UVC, as well as a correspondence between the number of reads within each genus and family and the number of individuals; and (c) whether the divergence between biodiversity recovered with eDNA and UVC has a phylogenetic or ecological component. Additionally, we explored (d) the signal of β diversity across eDNA samples by analyzing the compositional species dissimilarity between geographic locations.

2 | METHODS

2.1 | Study areas

The study focuses on two regions of Colombia, the island of Providencia and the Tayrona National Natural Park, with extensive coral reef habitats (Figure 1, Table S1). Providencia is located in the southwestern Caribbean Sea and is included in the UNESCO Seaflower Biosphere Reserve of Colombia. This island, which is part of the San Andres, Providencia, and Santa Catalina Archipelago, comprises a complex barrier reef on a calcareous platform surrounding an extinct Miocene volcano (Sánchez et al., 1998). The high habitat diversity provides a wide range of substratum types and coral reefs (Geister, 1992; Márquez, 1987), which shape the diversity, abundance, and distribution of coral reef fishes (Mejía & Garzón-Ferreira, 2000). The Tayrona National Natural Park is located along the continental Colombian Caribbean coast bordering the Sierra Nevada de Santa Marta. Tayrona Park has a heterogeneous coastal topography composed of metamorphic rocks, with numerous rocky headlands, islets, and bays (Garzón-Ferreira & Díaz, 2003). Coral and other hard-bottom communities are distributed along the coast, mainly as fringing reefs, while seagrass beds, mangroves, and coral reefs have developed to some extent in sheltered conditions within the bays (Garzón-Ferreira & Cano, 1991). The study was carried out in Gayraca Bay, where corals on the exposed side exhibit mainly massive to encrusting growth forms with colonies and a reef-like structure.

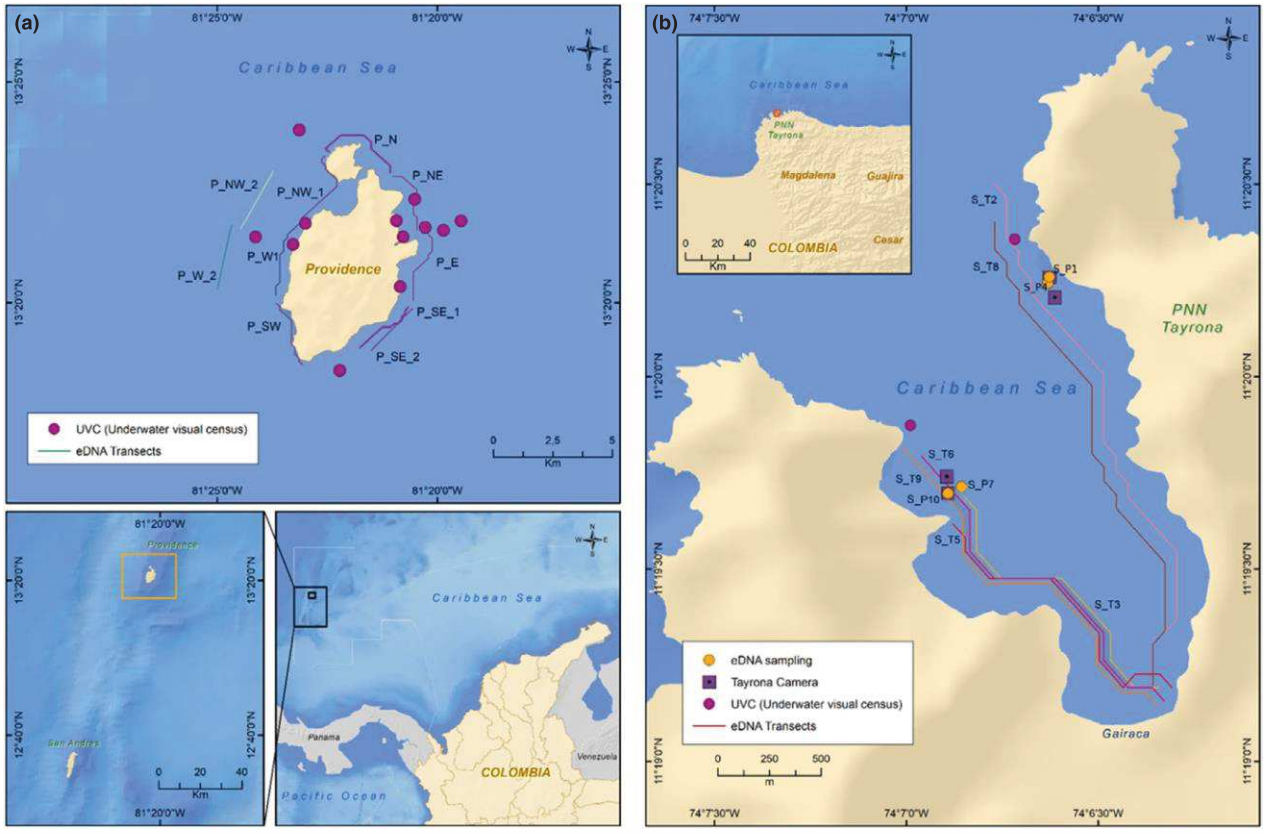


FIGURE 1 Area of eDNA sampling and underwater visual census (UVC) observations in (a) Providencia and (b) Tayrona National Natural Park. The magenta points indicate the sampling locations for the UVC at each of the chosen localities. The lines, yellow points, and purple squares indicate the transects filtered at each of the chosen localities. Source: Laboratorio de Sistemas de Información LabSIS, INVEMAR, Claudia Correa

2.2 | Underwater visual censuses

Divers conducted underwater visual censuses, using scuba equipment to survey the composition and abundance of fishes in Providencia and in Gayraca Bay. The surveys were performed during multiple years: 2000–2003, 2006–2007, and 2017 in Providencia and 1999–2011, 2013, and 2017 in Tayrona National Natural Park. Data were collected using the 30-min timed roving diver fish survey method for the established depths, 4–10 m in Providencia, and 8–14 m in Gayraca, inventorying all the observed species and estimating abundances in categories following the Coral Reef Monitoring System (SIMAC) methodology (CARICOMP, 1994, 1997, 2001; Garzón-Ferreira et al., 2002). In cases of fish schools abundance was estimated in tens. Four censuses per station were implemented, resulting in a total of 120 min of sampling in each monitoring event. In Providencia, the survey was performed in eight different habitats within the reef complex (Figure 1) and included a total of 4,200 min of sampling. Furthermore, seagrass habitats were also sampled in four 30-min roving diver visual surveys within a predefined area of 2,500 m^2. In Tayrona National Natural Park, the survey was performed in two different habitats comprising the exposed and protected reefs of Gayraca Bay (Figure 1), and it included a total of 3,600 min of sampling. Scientific names of species follow the Catalog of Fishes (Fricke et al., 2020), classification follows Fricke et al. (2020) for Elasmobranchii and Betancur et al. (2017) for Actinopterygii. To obtain a representative level of species diversity and abundance in the two regions (Providencia and Gayraca), we pooled values over multiple years and averaged abundances.

2.3 | eDNA field sampling, in situ filtration and treatment

For Providencia, we sampled two filtration replicates from each of 10 stations near the island, for a total of 20 water samples, from 29 to 15 July 2018. In Gayraca Bay, we sampled two filtration replicates from each of six stations, for a total of 12 water samples, from 23 to 26 October 2018. We sampled eDNA in situ using a filtration device composed of an Athena® peristaltic pump (Proactive Environmental Products LLC; nominal flow of 1.1 L/min), a VigiDNA® 0.22 μM cross-flow filtration capsule (SPYGEN) making it possible to filter a large water volume, and disposable sterile tubing for each filtration capsule. Two filtration replicates were performed in parallel, one on each side of the boat, at each station for 30 min, corresponding to a water volume of 30 L. At the end of each filtration, the water inside the capsules was emptied and the capsules were filled with 80 ml of CL1 conservation buffer (SPYGEN) and stored at room temperature. We followed a strict contamination control protocol in both field and laboratory stages (Goldberg et al., 2016; Valentini et al., 2016). Each water sample was processed using disposable gloves and single-use filtration equipment.

2.4 | DNA extraction, amplification, and high-throughput sequencing

DNA extraction, amplification, and sequencing were performed in separate dedicated rooms equipped with positive air pressure, UV treatment, and frequent air renewal. Two extractions per filter were performed following the protocol of Pont et al. (2018). For DNA extraction, each filtration capsule, containing the CL1 buffer, was agitated for 15 min on an S50 shaker (cat Ingenieurbüro™) at 800 rpm. The buffer was then emptied into two 50-ml tubes before being centrifuged for 15 min at 15,000 *g*. The supernatant was removed with a sterile pipette, leaving 15 ml of liquid at the bottom of each tube. Subsequently, 33 ml of ethanol and 1.5 ml of 3 M sodium acetate were added to each 50-ml tube and stored for at least one night at −20°C. The tubes were then centrifuged at 15,000 *g* for 15 min at 6°C, and the supernatants were discarded. After this step, 720 μl of ATL buffer from the DNeasy Blood & Tissue Extraction Kit (Qiagen GmbH) was added to each tube. Each tube was then vortexed, and the supernatant was transferred to a 2-ml tube containing 20 μl of Proteinase K. The tubes were finally incubated at 56°C for 2 hr. Subsequently, DNA extraction was performed using NucleoSpin® Soil (MACHEREY-NAGEL GmbH & Co.) starting from step 6 and following the manufacturer's instructions, and two DNA extractions were carried out per filtration capsule. The elution was performed by adding 100 μl of SE buffer twice. The two DNA samples were pooled before the amplification step. After the DNA extraction, the samples were tested for inhibition following the protocol described in Biggs et al. (2015). If a sample was considered inhibited, it was diluted fivefold before the amplification. DNA amplifications were performed in a final volume of 25 μl, using 3 μl of DNA extract as the template. The amplification mixture contained 1 U of AmpliTaq Gold DNA Polymerase (Applied Biosystems), 10 mM Tris-HCl, 50 mM KCl, 2.5 mM $MgCl_2$, 0.2 mM of each dNTP, 0.2 μM of each primer, 4 μM human blocking primer for the “teleo” primers (Civade et al., 2016), and 0.2 μg/μl bovine serum albumin (BSA, Roche Diagnostic).

We used three different primer sets, targeting chondrichthyans (Chon01, forward: -ACACCGCCCGTCACTCTC, reverse: -CATGTTACGACTTGCCTCCTC), teleosteans (teleo/Tele01, forward: -ACACCGCCCGTCACTCT, reverse: -CTTCCGGTACACTTACCATG) and more generally vertebrates (Vert01, forward: -TAGAACAGGCTCCTCTAG, reverse: -TTAGATACCCCACTATGC) (Taberlet et al., 2018; Valentini et al., 2016). Mean markers lengths were 44 bp for Chond01, 64 bp for teleo, and 97 for Vert01. These three primer sets were 5′-labeled with an eight-nucleotide tag unique to each PCR replicate for teleo and unique to each sample for the other two primer pairs (with at least three differences between any pair of tags), allowing the assignment of each sequence to the corresponding sample during sequence analysis. The tags for the forward and reverse primers were identical. The PCR mixture was denatured at 95°C for 10 min, followed by 50 cycles of 30 s at 95°C, 30 s at 55°C for teleo and Vert01 and 58°C for Chon01, 1 min at 72°C, and a final elongation step at 72°C for 7 min. Twelve PCR replicates were

run per filtration, that is, 24 per sampling site. After amplification, the samples were titrated using capillary electrophoresis (QIAxcel; Qiagen GmbH) and purified using the MinElute PCR purification kit (Qiagen GmbH). Before sequencing, purified DNA was titrated again using capillary electrophoresis. The purified PCR products were pooled in equal volumes to achieve a theoretical sequencing depth of 1,000,000 reads per sample. Three libraries were prepared using the MetaFast protocol (Fasteris). For two libraries, a paired-end sequencing (2 × 125 bp) was carried out using an Illumina HiSeq 2500 sequencer on a HiSeq Rapid Flow Cell v2 using the HiSeq Rapid SBS Kit v2 (Illumina) and on a MiSeq (2 × 125 bp) with the MiSeq Flow Cell Kit v3 (Illumina), following the manufacturer's instructions. Library preparation and sequencing were performed at Fasteris. Four negative extraction controls and two negative PCR controls (ultrapure water, 12 replicates) were amplified per primer pair and sequenced in parallel to the samples to monitor possible contaminants.

2.5 | OBITools filtering analyses for taxonomic assignments

Following sequencing, reads were processed to remove errors and analyzed using programs implemented in the OBITools package (http://metabarcoding.org/obitools, Boyer et al., 2016) based on a previous protocol (Valentini et al., 2016). The forward and reverse reads were assembled with the ILLUMINAPAIREDEND program, using a minimum score of 40 and retrieving only joined sequences. The reads were then assigned to each sample using the NGSFILTER software. A separate data set was created for each sample by splitting the original data set into several files using OBISPLIT. After this step, each sample was analyzed individually before merging the taxon list for the final ecological analysis. Strictly identical sequences were clustered together using OBIUNIQ. Sequences shorter than 20 bp, or with fewer than 10 occurrences were excluded using the OBIGREP program. The OBICLEAN program was then run within a PCR product. All sequences labeled "internal," which most likely correspond to PCR substitutions and indel errors, were discarded. Taxonomic assignment of the remaining sequences was performed using the ECOTAG program with the NCBI reference sequence (www.ncbi.nlm.nih.gov, release 233, downloaded on 11 October 2019). Considering the assignment of a few sequences to the wrong samples due to tag jumps (Schnell et al., 2015) and index hopping (MacConaill et al., 2018), all sequences with a frequency of occurrence <0.001 per taxon and per library and all sequences with an occurrence of <0.0006 per taxon in the RapidRun were discarded. Sequences with <100 reads in each sample were also discarded. These thresholds were empirically determined to clear all reads from blanks and controls and were included in our global data production procedure as suggested in De Barba et al. (2014). After the filtering pipeline, the extraction and PCR negative controls were completely clean, and no sequence reads remained in those samples.

2.6 | SWARM clustering analyses for MOTU identification

For the teleo primer set only, we used a second bioinformatics workflow based on sequence clustering using SWARM, an algorithm that groups multiple variants of sequences into MOTUs (Molecular Operational Taxonomic Units; Mahé et al., 2014; Rognes et al., 2016). Reads were assembled using VSEARCH (Rognes et al., 2016), then trimmed using CUTADAPT (Martin, 2013) and clustered using SWARM (Mahé et al., 2014). The clustering algorithms use sequence similarity and co-occurrence patterns to delineate meaningful entities, by grouping together sequence variants generated due to PCR and sequencing errors. Sequences were first merged using VSEARCH. CUTADAPT was then used for demultiplexing and primer trimming, and sequences containing ambiguities were removed with VSEARCH. SWARM was run with a minimum distance of one mismatch to make clusters. Once MOTUs were generated, the most abundant sequence within each cluster was used as a representative sequence for taxonomic assignment. A postclustering curation algorithm (LULU; Frøslev et al., 2017) was then applied to curate the data. The taxonomic assignment was performed using the ECOTAG program against the NCBI database. The taxonomic level of assignment was determined based on the results of the ECOTAG algorithm program and the percentage of similarity between the sequences in the sample and those in the reference database. After the clustering, bioinformatic filters were applied to remove PCR- or sequencing-related errors and nonspecific amplifications: (a) removal of amplicons with <10 reads per PCR, (b) removal of the nonspecific amplifications (nonfish), (c) removal of the amplicons whose size was not in the range of the targeted sequence (50–75 bp), (d) removal of all sequences found in only one PCR in the entire data set, and (e) cross-sample contamination cleaning by removing amplicons with <1/1,000 reads per PCR run (i.e., tag jumps; Schnell et al., 2015) and occurring in only one PCR run from a single sample (Ficetola et al., 2015). We corrected for tag jumps following the same procedure as for the OBITools workflow.

2.7 | Taxonomic comparison of eDNA and underwater visual censuses

For both pipelines, taxonomic assignments were corrected to avoid over-confident assignment outputs from ECOTAG: We only validated identification for 100% (species level), 90%–99% (genus level), or 85%–99% (family level) identity matches, when possible. Using the outputs of the OBITools pipeline, we compared the species, genera, and families recovered by eDNA to those recorded by UVC in Providencia and Gayraca Bay. We first compared the overlap in the list of species, genera, and families recovered with each of the three 12S primers targeting vertebrates, Actinopterygii and Elasmobranchii. Second, we evaluated whether the species, genera, and families recovered with the three eDNA primers matched the species recorded by

UVC. Complementary to the UVCs, we used the checklist of Bolaños-Cubillos et al. (2015) for Providencia and the compilation of species of SIMAC for Tayrona National Natural Park. These surveys were not performed at the same time as the eDNA, but represent in-depth, up-to-date knowledge of the species in the two regions. We further compared the species recorded by eDNA with other species distribution sources, including a compiled set of species distribution maps for the Caribbean region (Robertson & Van Tassell, 2019).

We analyzed whether detection differences between eDNA and UVC represented a phylogenetic signal and were associated with ecological traits. We performed this analysis at the genus level because the coverage of the reference database at the species level was sparser. We excluded all genera not represented in the reference database (10 genera were not detected with eDNA, were not in the reference database, but were detected in UVCs). We classified the remaining genera into (a) detected in eDNA only, (b) detected in UVCs only, and (c) detected in both. Because eDNA detection can be influenced by ecological features that are phylogenetically conserved, we first computed the phylogenetic signal of taxa recovered from eDNA and UVCs using the *D*-statistic (Fritz & Purvis, 2010) as implemented in the R package "caper." A negative value indicates that the phylogenetic pattern in the binary trait is extremely clumped on the tree, whereas a positive value indicates an overdispersed signal. A value around zero means that the trait is distributed on the tree as if it had evolved following a Brownian model (Fritz & Purvis, 2010). We used the distribution of 100 super-trees (Rabosky et al., 2018) pruned at the genus level. Next, we related detection classes to a set of ecological traits assembled for each species and aggregated at the genus level. Ecological traits were gathered from FishBase (Froese & Pauly, 2018) and included body size (small <15 cm, medium and large >40 cm), trophic guild (carnivore, herbivore, piscivore, planktivore), position in the water column (benthopelagic, demersal, pelagic, reef-associated, pelagic), home range mobility (sedentary, mobile, highly mobile), and schooling behavior (of a single of two individual, schools of 3–20 individuals, schools of >20 individuals). Based on these traits, we calculated a gower distance matrix between genera and constructed a trait space using a Principal Coordinates Analysis (PCoA). We mapped and estimated the trait volume recovered by each method to identify the differences between eDNA and UVCs. We plotted trait modalities as ellipses encompassing 90% of the genera of each modality.

2.8 | Diversity, abundance, and spatial variation in eDNA samples

We used the MOTU outputs from the SWARM protocol to perform diversity and composition analyses that did not strictly depend on completeness of the reference database. For the UVCs, we pooled species composition across multiple censuses and averaged the number of individuals per species and per region across the different sampling years. We evaluated the correspondence in species richness and abundance between eDNA and UVC. We performed a spearman correlation between the number of MOTUs per genus and per family and the number of species per genus and per family recorded by UVC. Next, we performed a spearman correlation between the number of reads per genus and per family and the number of individuals per genus and per family estimated by UVC. To perform the comparison between the number of reads and the number of individuals, values were scaled to between 0 and 1 before the analyses.

We also investigated the differences in eDNA composition between the sampling stations in Providencia and Gayraca Bay together and within Providencia separately. From the MOTU presence–absence matrix, we calculated a Jaccard distance matrix. To ordinate the compositional differences between the eDNA samples collected in both sampling sites, we performed a PCoA on this distance matrix. Using the same method, we performed a second PCoA analysis to investigate the compositional difference between the eDNA samples collected in the Providencia sampling stations. Sampling around this island covered multiple sites, following a gradient from sheltered locations to very exposed areas to marine currents. For each PCoA, we reported the explained deviance of each axis and mapped the ordination values in the geographic space.

We further calculated the pairwise Jaccard's dissimilarity index (Anderson et al., 2011; β_{jac}) of the compositional difference in MOTUs between (a) Providencia and Gayraca Bay and (b) between the west and east coast of Providencia. This index is expressed as: $\beta_{jac} = b + c/a + b + c$, where a is the number of MOTUs present in both sites, b is the number of MOTUs present in Providencia but not in Gayraca, and c is the number of MOTUs present in Gayraca Bay but not in Providencia. β_{jac} ranges from 0 (MOTU composition does not differ between sites) to 1 (MOTU composition is completely different between sites). We applied the partitioning framework proposed by Baselga (2012), which consists of decomposing β_{jac}, into two additive components, replacement and nestedness. The MOTU replacement component describes MOTU replacement without the influence of a difference in MOTU richness between sites ($\beta_{jtu} = 2 \min(b, c)/a + 2 \min(b, c)$). The nestedness component ($\beta_{jne} = \beta_{jac} - \beta_{jtu}$) accounts for the fraction of dissimilarity caused by a difference in MOTU richness.

3 | RESULTS

3.1 | Comparison between eDNA primers using OBITools

For Providencia, we detected a total of 107 different species when all three primer sets were used. We detected 53 species using the teleo primers, 74 species using the Vert01 primers, and five species exclusively of Elasmobranchii using the Chon01 primers. Using the teleo and Vert01 primers together we detected all 107 species, whereas we detected 53 species when the teleo and Chon01 primers were used together and 80 when the Vert01 and Chon01 primers were used together. We detected 19 species in common

between the teleo and Vert01 primers, five between the teleo and Chon01 primers, and none between the Vert01 and Chon01 primers. The identified families included Chaenopsidae, Gobiesocidae, Labrisomidae, Blenniidae and Gobiidae, which constitute the majority of cryptobenthic species. Among the detected species, we found the Caribbean reef shark (*Carcharhinus perezii*; Figure 2a), the Atlantic sharpnose shark (*Rhizoprionodon terraenovae*) and the great hammerhead shark (*Sphyrna mokarran*), all of which are characterized by elusive behavior. We further found species such as *Erotelis smaragdus*, a demersal dweller of brackish and marine waters that has not been reported before in the Archipelago.

In Gayraca Bay, we detected 85 species using the teleo and Vert01 primer sets. No species were identified with the Chon01 primers. Out of the 85 detected species, we identified 18 with both primer sets, 39 with teleo only and 65 with Vert01 only. In particular, we identified the family Narcinidae, which was the only chondrichthyan family detected using the teleo primer and was not identified using the Chon01 primers. We additionally detected cryptobentonic families such as the Bleniidae, Gobiesocidae, Labrisomidae, Apogonidae, and Gobiidae. At the genus level, *Entomacrodus*, a monospecific (*Entomacrodus nigricans*) cryptobenthic genus in the Caribbean, was among those detected with the Vert01 primers. At the species level, notable detected species included the goldspot goby (*Gnatholepis thompsoni*) and the rusty goby (*Priolepis hipoliti*; Figure 2d).

3.2 | Comparison of species detection between eDNA and UVC

A total of 113 species were recorded in the UVCs around Providencia. Using all three primers together, with eDNA we detected 35 (31%) of the 113 species that were observed in the UVCs. Out of these species, we detected 20 with the teleo primers, 25 with the Vert01 primers and 2 with the Chon01 primers. On the other hand, we detected 72 species with eDNA that were not observed during the UVCs. Overall, 41 out of 106 genera detected with eDNA were also recorded by UVC. We recorded some reef-associated species, such as the yellowhead wrasse (*Halichoeres garnoti*) and the blue chromis (*Chromis cyanea*), with both UVC and eDNA, while we detected typical cryptobenthic species, such as the dwarf blenny (*Starksia nanodes*), the island goby (*Lythrypnus nesiotes*), and the mimic cardinalfish (*Apogon phenax*) only with eDNA. The detection of these species or other taxa by eDNA is supported by their known occurrence in Providencia based on species range maps and a local checklist (Tables S2–S4).

A total of 57 species were recorded during the UVCs in Gayraca Bay. Using all three primers together, we detected 16 (28%) of these 57 species. Out of these species found with both UVC and eDNA, we detected 7 with the teleo primer, 14 with the Vert01 primer, and none with the Chon01 primer. On the other hand, we detected 85 species with eDNA that were not observed during the UVCs. Out of the 92 genera detected by eDNA, 24 were also observed during

FIGURE 2 Montage of pictures of emblematic species detected using eDNA but not observed in the underwater visual surveys. (a) The Caribbean reef shark (*Carcharhinus perezii*), (b) the goldentail moray (*Gymnothorax miliaris*), (c) the bigeye scad (*Selar crumenophthalmus*), (d) the rusty goby (*Priolepis hipoliti*), (e) the orangespotted goby (*Nes longus*), (f) the green razorfish (*Xyrichthys splendens*). Pictures: Juan David González Corredor

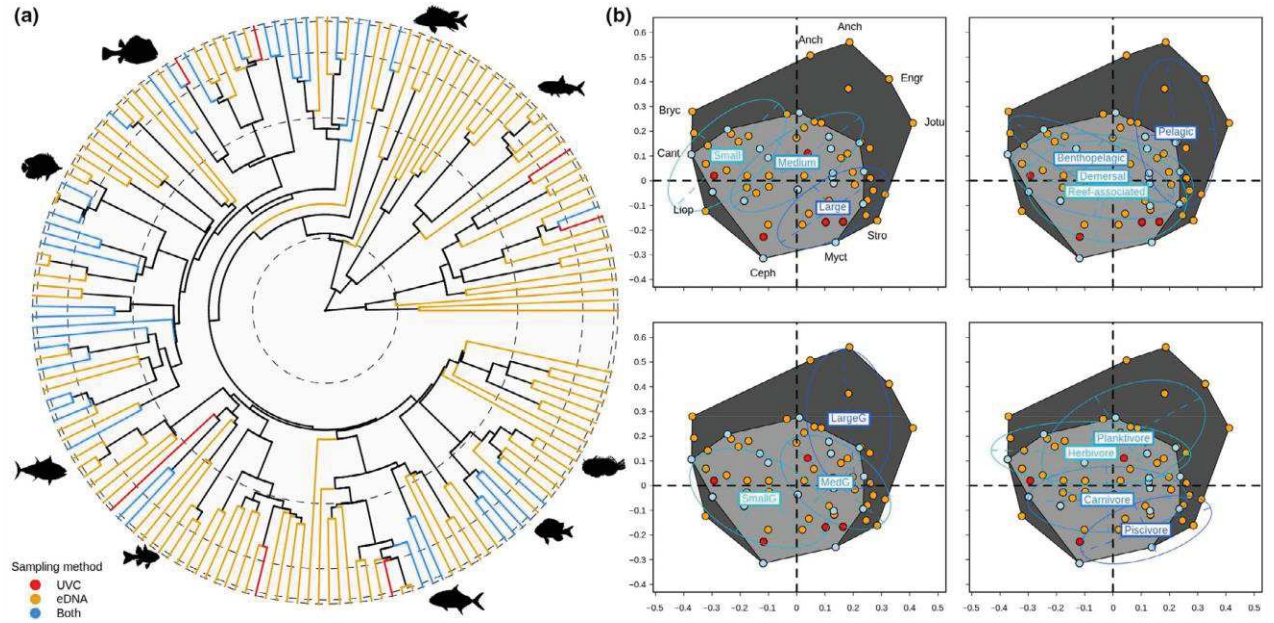


FIGURE 3 Phylogenetic and functional bias detection using underwater visual census and eDNA. (a) one of the 100 phylogenetic trees pruned at the genus level from the super-trees of Rabosky et al. (2018). (b) The trait space obtained by performing a PCoA (percentage of inertia, axis 1: 24.4% and axis 2: 15.4%) on a set of ecological traits assembled for each species and aggregated at the genus level. The dark grey polygon represents the trait space covered by genera sampled by eDNA, whereas the light grey polygon represents the trait space covered by genera sampled by UVC. On the trait space, we drew an ellipse representing 90% of the points belonging to a trait category for the following traits: body size (small <15 cm, medium and large >40 cm), trophic guilds (carnivore, herbivore, piscivore, planktivore), position in the water column (benthopelagic, demersal, pelagic, reef-associated, pelagic), schooling behaviour (small groups of 1 or 2 individuals, medium groups of species gathering in schools of 3–20 individuals, schooling species of >20 individuals). In all plots, orange circles represent genera detected by the eDNA sampling method only, red by UVC only, and blue by both methods

the UVCs. Of these, 8 and 16 were detected using the teleo and Vert01 primers, respectively. We found some reef-associated species with both UVC and eDNA, such as the Spanish hogfish (*Bodianus rufus*), the yellowtail damselfish (*Microspathodon chrysurus*), and the yellow goatfish (*Mulloidichthys martinicus*), while we detected typical cryptobenthic species, such as the rusty goby (*P. hipoliti*; Figure 2d), the dusky cardinalfish (*Phaeoptyx pigmentaria*), and the spotfin goby (*Oxyurichthys stigmalophius*) only with eDNA. The detection of these species or other taxa by eDNA is supported by their known occurrence in Tayrona Park based on species range maps and a local checklist (Tables S5–S7).

3.3 | Comparison of species richness and abundances between eDNA MOTUs and underwater visual surveys

We performed the aggregation into MOTUs using the teleo primers, as the bioinformatics clustering pipeline using SWARM has only been developed and fully tested with this primer (Juhel et al., 2020; Marques et al., 2020). In Providencia, the eDNA clustering pipeline identified 227 distinct MOTUs, and we detected an average of 26.2 ± 12.6 MOTUs per filter. Altogether, we detected 53 species, 76 genera, and 50 families by comparing MOTUs to the reference database. In Gayraca Bay, the eDNA clustering pipeline identified 189 distinct MOTUs. We detected an average of 12.9 ± 6.9 MOTUs per sample. Altogether, we detected 35 species, 52 genera, and 42 families by comparing MOTUs to the reference database.

We tested the correlation between species richness and numbers of MOTUs in Providencia and Gayraca Bay (Figure S1). In Providencia, we found a significant correlation between the number of species per genus and the number of MOTUs per genus (Spearman correlation test, $n = 30$, $\rho = .37$, $p = .04$). The genera *Urobatis*, *Scarus*, and *Hypoplectrus* were identified as outliers in these correlations. We found a weaker correlation between the number of species per family and the number of MOTUs per family ($n = 23$, $\rho = .33$, $p = .13$). The number of individuals was also correlated with the number of MOTU reads per genus ($n = 30$, $\rho = .4$, $p = .03$, Figure S2). The genera *Canthigaster*, *Halichoeres*, *Scarus*, and *Sparisoma* were outliers in this relationship. The number of individuals and the number of MOTU reads per family also showed a significant positive correlation ($n = 23$, $\rho = .45$, $p = .03$), with Tetraodontidae and Labridae as outliers. In Gayraca Bay, we found no correlation between the number of species per genus and the number of MOTUs per genus ($n = 13$, $\rho = .1$, $p = .75$). The number of species per family versus the number of MOTUs per family showed no correlation ($n = 12$, $\rho = -.04$, $p = .91$). We also found no correlation between the number of individuals and the number of MOTU reads per genus ($n = 13$, $\rho = .04$,

$p = .9$). Finally, there was not a significant correlation between the number of individuals and the number of MOTU reads per family in this region ($n = 12$, $\rho = .28$, $p = .38$).

3.4 | Ecological and phylogenetic distribution of species detection

We investigated the ecological and phylogenetic distributions of detection considering all genera recorded by either eDNA or UVC and also included in the reference database. We examined the phylogenetic signal of the detection in either eDNA, UVC, or both. For the genera detected by UVC, we found an average *D*-statistic of 0.18 ± 0.1 across the 100 trees, indicating that the clustering of genera identified by this monitoring technique is not different than expected under a Brownian model ($p = .28 \pm .12$; Figure 3a). In contrast, for the genera detected by eDNA, we found an average *D*-statistic of 1.16 ± 0.15, indicating that these taxonomic units detected by eDNA are widely distributed across the phylogenetic tree, as expected under a model of random phylogenetic signal ($p = .66 \pm .18$; Figure 3a).

We related the detection classes to ecological traits using PCoA. The percentage of inertia of the first axis of the PCoA was 24.4%, while the percentage of inertia of the second axis was 15.4%. We found that a large proportion of ecological traits was covered by the two sampling methods, even if UVC detected a smaller number of genera than eDNA. eDNA was better at detecting large piscivore and pelagic species belonging to genera such as *Istiophorus*, *Euthynnus*, *Decapterus*, *Acanthocybium*, and *Strongylura*, but also smaller planktivorous species of *Sardinella*, *Cetengraulis*, *Lycengraulis*, and *Engraulis* (Figure 3b). eDNA further detected more small and bottom-associated species represented by the genera *Liopropoma*, *Hypsoblennius* and *Arcos*.

3.5 | Spatial variation in eDNA MOTUs

We investigated MOTU composition dissimilarity among samples and found marked differences between the eDNA samples collected in Providencia and those from Gayraca Bay, but also between samples from opposite sides of the island of Providencia. The PCoA performed on both Providencia and Gayraca Bay explained a large fraction of the total inertia (50%), with 41.2% for the first axis and 8.8% for the second axis (Figure 4), and it showed a marked difference in composition between the two Caribbean sites. The pairwise Jaccard's dissimilarity index calculated between Providencia and Gayraca Bay reached a value of 0.71, meaning that the two sites present a high dissimilarity. The two regions had only 93 MOTUs in common out of the total of 323 identified. The difference in MOTU composition between the two regions was mainly explained by turnover ($\beta_{jtu} = 0.67$), while the nestedness was low ($\beta_{jne} = 0.04$). The second PCoA, focusing on samples collected off the west and east coasts of Providencia, explained 42.6% of the total data set inertia, with 25.6% for the first axis and 17% for the second axis. We found marked differences in eDNA composition between the eastern and western sides of the island (Figure S3). When exploring the difference between the west and east coast of Providencia, we found that the MOTU composition differed moderately ($\beta_{jac} = 0.27$) and 97.6% of the β_{jac} was turnover ($\beta_{jtu} = 0.267$; $\beta_{jne} = 0.006$). The two sides of the island had 165 MOTUs in common out of the total of 227 identified. With more taxa, the western side included some species typically associated with complex habitats of seagrasses and reef patches, such as the hogfish (*Lachnolaimus maximus*) and *Sygnathus* sp.

4 | DISCUSSION

We showed that eDNA metabarcoding can provide a comprehensive overview of fish composition in two highly diverse tropical marine reefs of Colombia. UVC is traditionally used to monitor fish diversity on coral reefs (Samoilys & Carlos, 2000). However, besides logistical difficulties to organize underwater sampling in remote locations, UVC can suffer from several observer biases, such as overlooking cryptobenthic (Bozec et al., 2011) or wideranged species such as sharks (Juhel et al., 2018). Compared with UVCs performed over two decades (1999–2017), the eDNA surveys from one year detected a large fraction of the fish species diversity, including many species that were not recorded during UVCs, and covered a wider fraction of the phylogeny and ecological space of the ichthyofauna. Moreover, we showed that eDNA has a marked spatial signal, both between the two investigated regions and within the Providencia region, supporting future local habitat monitoring of reefs using eDNA (West et al., 2020). Together, our analyses support the use of eDNA as an approach for the fast monitoring of highly diverse tropical marine ecosystems. In an eDNA study using a different marker (CO1) to detect fish, Nguyen et al. (2020) likewise showed that eDNA methods are efficient in detecting small taxa that would be undetected in traditional surveys, while also accurately describing biodiversity patterns in adjacent tropical habitats.

Environmental DNA detected many species recorded by UVC, as well as cryptic species known to occur regionally. The majority of the species detected by eDNA in Providencia and Gayraca Bay, 67.2% and 81.2%, respectively, were not detected by UVC. Similarly, 61.7% and 81.7% of genera and 59.0% and 78.9% of families detected by eDNA were absent from UVC records in Providencia and Gayraca Bay, respectively. The species occurrences detected by eDNA but not by UVC are most likely genuine, as those species are known from complementary sources to occur in Providencia or in Tayrona National Natural Park (Table S2; Bolaños-Cubillos et al., 2015; Robertson & Van Tassell, 2019). While both methods jointly detected some abundant reef fishes, such as the brown chromis (*Chromis multilineata*), the bicolor damselfish (*Stegastes partitus*), and the yellow goatfish (*M. martinicus*), eDNA alone detected species within the Chaenopsidae, Labrisomidae and Gobiidae, mainly cryptobenthic clades that are

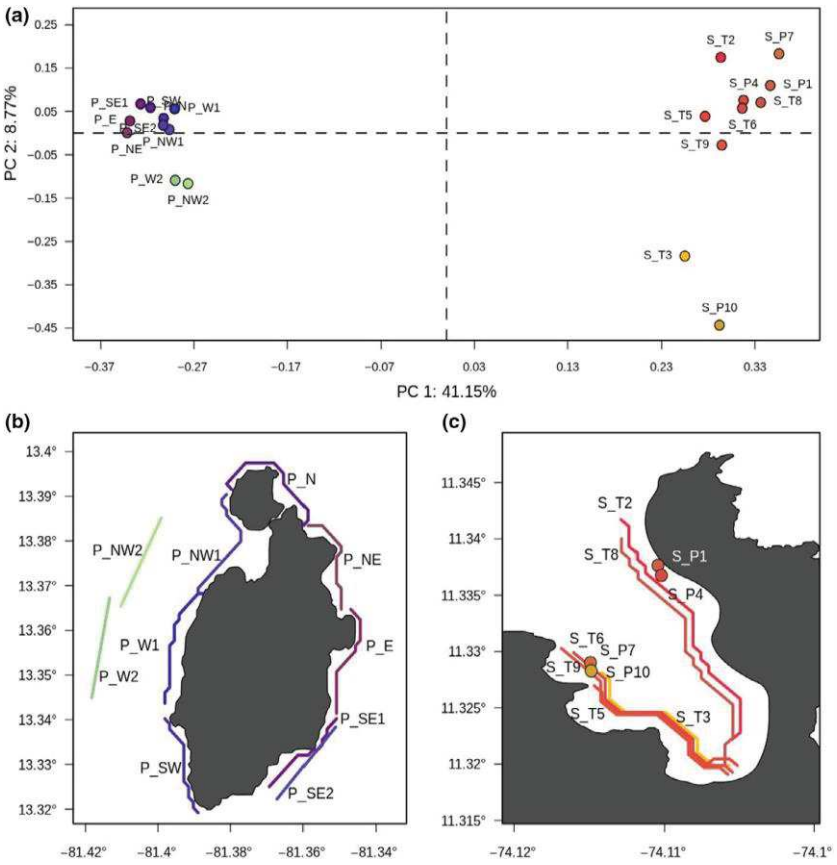


FIGURE 4 (a) Compositional differences (PCoA) from the MOTUs presence–absence matrix between the eDNA sampling stations in Providencia and Gayraca Bay. (b) Transects maps of the island of Providencia, where colors correspond to the position of the transect in the ordination space. (c) Map of the transects realized in the Tayrona National Natural Park, where colors correspond to the position of the transects in the ordination space

difficult to observe with UVC (Brandl et al., 2018). Further, eDNA sampling delivered potential new records of species for the studied areas. In particular, the eDNA detection of the blue hamlet (*Hypoplectrus gemma*) constitutes the first record of the species in the south of the Greater Caribbean, and the detection of the smooth-eye blenny (*Starksia atlantica*) the first record in the northwestern Caribbean. The redeye parrotfish (*Sparisoma axillare*) has previously mainly been reported in the southeastern Caribbean but was detected with eDNA in both sample sites of this study, expanding the known distribution range of the species. While these records require further validation, our results suggest that, beyond providing a comprehensive assessment of local biodiversity, eDNA offers a novel approach to document more accurately the biogeographic range of species.

Because some taxa were detected by eDNA but not by UVC, and vice versa, we further analyzed the difference in detection between the two approaches. As the most obvious cause of discrepancy, species and genera found in the UVCs but not detected in the eDNA were missing from the reference database. We found that 60% of the genera that were recorded during UVCs but not detected by eDNA were not in the reference database extracted from NCBI, highlighting that the reference database is central to effective eDNA monitoring (DiBattista et al., 2017). Overall, eDNA analysis led to the recovery of a larger number of genera, covering a larger fraction of the phylogenetic tree and of the ecological space of fishes (Figure 3). The fish on coral reefs tend to be phylogenetically diverse, with representatives of multiple families (Leprieur et al., 2016). We found that the genera detected using eDNA had a wide spread across the fish phylogenetic tree, while the genera observed during UVCs were phylogenetically clumped. Our results suggest that eDNA surveys are more representative than UVCs of the entire phylogenetic diversity of fishes on coral reefs. We found a positive correlation in diversity and abundance between the two sampling approaches in Providencia but not in Gayraca Bay. While the UVC sampling effort was high in Providencia, with eight UVCs targeting different habitats, the effort was lower in Gayraca Bay, where only two sites were sampled, which could explain the difference in signal between regions. Together, this indicates a general limitation of the comparison proposed in this study, that we do not know the true compositions and abundances, as both sampling approaches involve some level of bias. Longer term, synchronous eDNA sampling and video recording could provide further validation of eDNA (Stat et al., 2019).

Besides species diversity, eDNA is also expected to provide information on the spatial distribution of species assemblages across different habitats (Nguyen et al., 2020; West et al., 2020). In agreement with findings from previous studies (Closek et al., 2019; Nguyen et al., 2020) and in contrast to the idea that eDNA would be largely redistributed in a more open marine system (Díaz-Ferguson & Moyer, 2014), we found a clear spatial structure in the eDNA

composition. Indeed, our approach captured marked differences between Gayraca Bay and Providencia, but also more locally between the east and west coasts of Providencia, corresponding to variation in habitat. The island of Providencia is composed of various habitats, and the eastern side is more exposed than the western one (Coralina-Invemar, 2012). Geomorphological diversity of the coral reef system, added to the combination of oceanic influences and terrigenous contributions from the island, lead to high variety in underwater environments and coastlines (Díaz et al., 2000). We found that the eastern side of the island has a species composition dominated by species associated with reef habitats, such as the blackear wrasse (*Halichoeres poeyi*) and the redtail parrotfish (*Sparisoma chrysopterum*); the western side is characterized by species associated with lagoon complexes covered with extensive patches of seagrass meadows alternating with small coral reef patches, such as the seagrass eel (*Chilorhinus suensonii*) and the blackfin cardinalfish (*Astrapogon puncticulatus*). Our results align with those of West et al. (2020), who observed marked eDNA compositional differences between habitats in the Cocos Islands of Australia, and suggest that coastal eDNA can be localized in marine environments.

eDNA metabarcoding is now widely employed in various aquatic ecosystems (Deiner et al., 2017), but some uncertainties remain as regard to sampling design (Valentini et al., 2016) and the choice of markers (Collins et al., 2019; Stat et al., 2017) and bioinformatics pipeline (Calderón-Sanou et al., 2020; Juhel et al., 2020). We tested three different primer sets for the 12S region looking for fish taxa, but we did not find a universal marker able to detect all taxa. The teleo primer generally performed best, as it was able to retrieve many teleost species, as well as five of the six species of Elasmobranchii also detected with the Chon01 primer in Providencia and one taxa of the same group at the family level in Gayraca. Nevertheless, the teleo primer did not recover some of the species that were recovered by the Vert01 primer (54 vs. 74 in Providencia and 39 vs. 64 in Gayraca), while the Vert01 primer did not recover a few species only found with the teleo primer (33 and 21 for Providencia and Gayraca, respectively). Hence, as this stage of primer development and testing, it appears that a multiprimer approach is required to capture of the entire diversity of a site (West et al., 2020). Moreover, because we found many Elasmobranchii with the teleo primer, a specialized primer for Elasmobranchii might not be needed and could be replaced by the more ubiquitous teleo primer. In that regard, teleo is an exception among eDNA primers because other sets, such as the MiFish primers, do not amplify Elasmobranchii (Bylemans et al., 2018; Miya et al., 2015).

A mayor limitation of eDNA is the lack of completeness of the reference database. Yet, a high coverage of the reference database is crucial to allow future accurate identification of species assemblages. In fact, many species recorded by UVC were not recovered with eDNA simply because they were not represented in the reference database. In order to fully exploit the potential detection power of eDNA metabarcoding, a vast effort is needed to improve taxonomic coverage of reference databases (Schenekar et al., 2020; Weigand et al., 2019). Addressing these important database gaps requires analyses that are not based solely on species assignment. We generated MOTUs using SWARM to get an indication of the expected overall biodiversity. However, while some MOTUs perfectly delineate true biological species without the need of a reference sequence, a fraction of these MOTUs also represent errors stemming from PCR and sequencing, overestimating true diversity (Morgan et al., 2013; Reeder & Knight, 2009), while clustering might also bind together distinct closely related species, underestimating true diversity (Huse et al., 2010). Thus, procuring a taxonomically comprehensive database with high-quality sequences and accurate data curation steps is crucial for producing robust and reproducible ecological conclusions from eDNA metabarcoding methods (Collins et al., 2019; Weigand et al., 2019).

Alternative ways to survey marine biodiversity beyond UVCs and unbiased evaluations of the ecosystem components are needed, as these provide a baseline for the management of marine protected areas (Stat et al., 2019). eDNA metabarcoding is becoming a more accessible method that generates reliable information for ecosystem surveillance and could prove valuable in marine monitoring programs (Lacoursière-Roussel et al., 2016). Here, we show that eDNA quickly provides a detailed picture of fish diversity and composition in two marine protected areas of Colombia, which can be used for future monitoring and management of these sites (Bálint et al., 2018). Despite water exchange in coastal marine systems, eDNA signals are localized on coral reefs, which is promising for monitoring the health status of these ecosystems. Repeated observations of eDNA measurements at multiple stations in these areas will facilitate assessment of the status and ultimately trends in biodiversity, particularly in response to disturbance events associated with climate change (Berry et al., 2019) or pollution (Bagley et al., 2019). Our results further highlight the importance of establishing a complete reference database for eDNA analyses, as many of the sequences could not be attributed to a particular genus or species. As shown for lake ecosystems (Hänfling et al., 2016), eDNA could become an important complement to traditional UVCs for monitoring coral reef biodiversity.

ACKNOWLEDGMENTS

This project was supported by the ETH Global grant and the Monaco Explorations Foundation, CORALINA, which provided support for entrance to the island of Providencia for the development of the project and maritime support for the field sampling. We express special thanks to Nicasio and Casimiro for providing guidance at sea. LP received funding from the Swiss National Science Foundation for the project Reefish (grant number 310030E-164294). We thank Claudia Correa Rojas (Information Systems Laboratory of INVEMAR) for her support in cartography and SPYGEN staff for technical support in the laboratory. We are grateful to PE Guerin for his support in bioinformatics pipeline development. This study is contribution number 1269 of the Instituto de Investigaciones Marinas y Costeras—INVEMAR.

CONFLICT OF INTEREST

All authors declare that there is no conflict of interest regarding the publication of this article.

AUTHOR CONTRIBUTIONS

LP, CA, and APF jointly designed this study, APF, VM, JBJ, GHB, MCC, JDGC, AAC, RH, EM, MS, and CA participated in the field work. FF, AV, VM, and CA analyzed the data. All the authors APF, VM, FF, JBJ, GHB, MCC, TD, JDGC, AAC, RH, DE, EM, MS, AV, SM, DM, CA, and LP contributed to writing the manuscript.

DATA AVAILABILITY STATEMENT

Summary data are presented in the Supplementary Material. All the raw reads can be found following https://doi.org/10.5061/dryad.mcvdncjz9. Code for the clustering bioinformatics pipeline can be found in Github: https://gitlab.mbb.univ-montp2.fr/edna/snakemake_rapidrun_swarm. Coding for the OBITools pipeline can be found at: https://gitlab.mbb.univ-montp2.fr/edna/snakemake_rapidrun_obitools/-/tree/master (Albouy et al., 2020).

ORCID

Andrea Polanco Fernández https://orcid.org/0000-0001-6121-5214
Fabian Fopp https://orcid.org/0000-0003-0648-8484
Régis Hocdé https://orcid.org/0000-0002-5794-2598

REFERENCES

Albouy, C., Polanco, A., & Pellissier, L. (2020). Environmental DNA metabarcoding to monitor tropical reef fishes in Providencia island. *Dryad Dataset*. https://doi.org/10.5061/dryad.mcvdncjz9

Anderson, M. J., Crist, T. O., Chase, J. M., Vellend, M., Inouye, B. D., Freestone, A. L., Sanders, N. J., Cornell, H. V., Comita, L. S., Davies, K. F., & Harrison, S. P. (2011). Navigating the multiple meanings of β diversity: A roadmap for the practicing ecologist. *Ecology Letters*, *14*(1), 19–28. https://doi.org/10.1111/j.1461-0248.2010.01552.x

Bagley, M., Pilgrim, E., Knapp, M., Yoder, C., Santo Domingo, J., & Banerji, A. (2019). High-throughput environmental DNA analysis informs a biological assessment of an urban stream. *Ecological Indicators*, *104*, 378–389. https://doi.org/10.1016/j.ecolind.2019.04.088

Bakker, J., Wangensteen, O. S., Baillie, C., Buddo, D., Chapman, D. D., Gallagher, A. J., Guttridge, T. L., Hertler, H., & Mariani, S. (2019). Biodiversity assessment of tropical shelf eukaryotic communities via pelagic eDNA metabarcoding. *Ecology and Evolution*, *9*(24), 14341–14355. https://doi.org/10.1002/ece3.5871

Bálint, M., Pfenninger, M., Grossart, H. P., Taberlet, P., Vellend, M., Leibold, M. A., Englund, G., & Bowler, D. (2018). Environmental DNA time series in ecology. *Trends in Ecology & Evolution*, *33*(12), 945–957. https://doi.org/10.1016/j.tree.2018.09.003

Barlow, J., França, F., Gardner, T. A., Hicks, C. C., Lennox, G. D., Berenguer, E., Castello, L., Economo, E. P., Ferreira, J., Guénard, B., & Leal, C. G. (2018). The future of hyperdiverse tropical ecosystems. *Nature*, *559*(7715), 517–526.

Baselga, A. (2012). The relationship between species replacement, dissimilarity derived from nestedness and nestedness. *Global Ecology and Biogeography*, *21*(12), 1223–1232. https://doi.org/10.1111/j.1466-8238.2011.00756.x

Berry, T. E., Saunders, B. J., Coghlan, M. L., Stat, M., Jarman, S., Richardson, A. J., Davies, C. H., Berry, O., Harvey, E. S., & Bunce, M. (2019). Marine environmental DNA biomonitoring reveals seasonal patterns in biodiversity and identifies ecosystem responses to anomalous climatic events. *PLoS Genetics*, *15*(2), e1007943. https://doi.org/10.1371/journal.pgen.1007943

Betancur-R, R., Wiley, E. O., Arratia, G., Acero, A., Bailly, N., Miya, M., Lecointre, G., & Orti, G. (2017). Phylogenetic classification of bony fishes. *BMC Evolutionary Biology*, *17*(1), 162. https://doi.org/10.1186/s12862-017-0958-3

Biggs, J., Ewald, N., Valentini, A., Gaboriaud, C., Dejean, T., Griffiths, R. A., Foster, J., Wilkinson, J. W., Arnell, A., Brotherton, P., & Williams, P. (2015). Using eDNA to develop a national citizen science-based monitoring programme for the great crested newt (*Triturus cristatus*). *Biological Conservation*, *183*, 19–28. https://doi.org/10.1016/j.biocon.2014.11.029

Bolaños-Cubillos, N., Abril-Howard, A., Bent-Hooker, H., Caldas, J. P., & Acero, A. (2015). Lista de peces conocidos del archipiélago de San Andrés y Providencia, Caribe occidental colombiano. *Boletín de Investigaciones Marinas y Costeras*, *44*, 127–162.

Boussarie, G., Bakker, J., Wangensteen, O. S., Mariani, S., Bonnin, L., Juhel, J. B., Kiszka, J. J., Kulbicki, M., Manel, S., Robbins, W. D., & Vigliola, L. (2018). Environmental DNA illuminates the dark diversity of sharks. *Science Advances*, *4*(5), eaap9661. https://doi.org/10.1126/sciadv.aap9661

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., & Coissac, E. (2016). obitools: A unix-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, *16*(1), 176–182.

Bozec, Y. M., Kulbicki, M., Laloë, F., Mou-Tham, G., & Gascuel, D. (2011). Factors affecting the detection distances of reef fish: Implications for visual counts. *Marine Biology*, *158*(5), 969–981. https://doi.org/10.1007/s00227-011-1623-9

Brandl, S. J., Goatley, C. H., Bellwood, D. R., & Tornabene, L. (2018). The hidden half: Ecology and evolution of cryptobenthic fishes on coral reefs. *Biological Reviews*, *93*(4), 1846–1873. https://doi.org/10.1111/brv.12423

Bylemans, J., Gleeson, D. M., Hardy, C. M., & Furlan, E. (2018). Toward an ecoregion scale evaluation of eDNA metabarcoding primers: A case study for the freshwater fish biodiversity of the Murray-Darling Basin (Australia). *Ecology and Evolution*, *8*(17), 8697–8712. https://doi.org/10.1002/ece3.4387

Calderón-Sanou, I., Münkemüller, T., Boyer, F., Zinger, L., & Thuiller, W. (2020). From environmental DNA sequences to ecological conclusions: How strong is the influence of methodological choices? *Journal of Biogeography*, *47*, 193–206. https://doi.org/10.1111/jbi.13681

CARICOMP (1994). *CARICOMP manual de métodos nivel 1: Manual de métodos para el mapeo y monitoreo de parámetros físicos y biológicos en la zona costera del Caribe* (68 p.). DMC CARICOMP, Univ. West Indies.

CARICOMP (1997). *CARICOMP monitoring of coral reefs*. Proc. 8th Int. Coral Reef Symp. 1 (pp. 651–656).

CARICOMP (2001). *Methods manual levels 1 and 2: Manual of methods for mapping and monitoring of physical and biological parameters in the coastal zone of the Caribbean* (85 p.). CARICOMP Data Management Center, Univ. West Indies.

Cilleros, K., Valentini, A., Allard, L., Dejean, T., Etienne, R., Grenouillet, G., Iribar, A., Taberlet, P., Vigouroux, R., & Brosse, S. (2019). Unlocking biodiversity and conservation studies in high-diversity environments using environmental DNA (eDNA): A test with Guianese freshwater fishes. *Molecular Ecology Resources*, *19*(1), 27–46. https://doi.org/10.1111/1755-0998.12900

Cinner, J. E., Huchery, C., MacNeil, M. A., Graham, N. A., McClanahan, T. R., Maina, J., Maire, E., Kittinger, J. N., Hicks, C. C., Mora, C., & Allison, E. H. (2016). Bright spots among the world's coral reefs. *Nature*, *535*(7612), 416–419.

Civade, R., Dejean, T., Valentini, A., Roset, N., Raymond, J. C., Bonin, A., Taberlet, P., & Pont, D. (2016). Spatial representativeness of environmental DNA metabarcoding signal for fish biodiversity assessment in a natural freshwater system. *PLoS One*, *11*(6), e0157366. https://doi.org/10.1371/journal.pone.0157366

Closek, C. J., Santora, J. A., Starks, H. A., Schroeder, I. D., Andruszkiewicz, E. A., Sakuma, K. M., Bograd, S. J., Hazen, E. L., Field, J. C., & Boehm,

A. B. (2019). Marine vertebrate biodiversity and distribution within the central California Current using environmental DNA (eDNA) metabarcoding and ecosystem surveys. *Frontiers in Marine Science*, *6*, 732. https://doi.org/10.3389/fmars.2019.00732

Collen, B., Ram, M., Zamin, T., & McRae, L. (2008). The tropical biodiversity data gap: Addressing disparity in global monitoring. *Tropical Conservation Science*, *1*(2), 75–88. https://doi.org/10.1177/194008290800100202

Collins, R. A., Bakker, J., Wangensteen, O. S., Soto, A. Z., Corrigan, L., Sims, D. W., Genner, M. J., & Mariani, S. (2019). Non-specific amplification compromises environmental DNA metabarcoding with COI. *Methods in Ecology and Evolution*, *10*(11), 1985–2001. https://doi.org/10.1111/2041-210X.13276

CORALINA-INVEMAR (2012). *Atlas de la Reserva de Biósfera Seaflower*. D. I. Gómez López, C. Segura-Quintero, P. C. Sierra-Correa, & J. Garay-Tinoco (Eds), (180 p.). Archipiélago de San Andrés, Providencia y Santa Catalina. Instituto de Investigaciones Marinas y Costeras "José Benito Vives De Andréis" -INVEMAR- y Corporación para el Desarrollo Sostenible del Archipiélago de San Andrés, Providencia y Santa Catalina -CORALINA-. Serie de Publicaciones Especiales de INVEMAR # 28.

Costello, M. J., Lane, M., Wilson, S., & Houlding, B. (2015). Factors influencing when species are first named and estimating global species richness. *Global Ecology and Conservation*, *4*, 243–254. https://doi.org/10.1016/j.gecco.2015.07.001

Darling, E. S., Graham, N. A., Januchowski-Hartley, F. A., Nash, K. L., Pratchett, M. S., & Wilson, S. K. (2017). Relationships between structural complexity, coral traits, and reef fish assemblages. *Coral Reefs*, *36*(2), 561–575. https://doi.org/10.1007/s00338-017-1539-z

De Barba, M., Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E., & Taberlet, P. (2014). DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: Application to omnivorous diet. *Molecular Ecology Resources*, *14*(2), 306–323. https://doi.org/10.1111/1755-0998.12188

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., De Vere, N., & Pfrender, M. E. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, *26*(21), 5872–5895. https://doi.org/10.1111/mec.14350

Dejean, T., Valentini, A., Duparc, A., Pellier-Cuit, S., Pompanon, F., Taberlet, P., & Miaud, C. (2011). Persistence of environmental DNA in freshwater ecosystems. *PLoS One*, *6*(8), e23398. https://doi.org/10.1371/journal.pone.0023398

Descombes, P., Wisz, M. S., Leprieur, F., Parravicini, V., Heine, C., Olsen, S. M., Swingedouw, D., Kulbicki, M., Mouillot, D., & Pellissier, L. (2015). Forecasted coral reef decline in marine biodiversity hotspots under climate change. *Global Change Biology*, *21*(7), 2479–2487. https://doi.org/10.1111/gcb.12868

Díaz, J. M., Barrios, L. M., Cendales, M. H., Garzón-Ferreira, J., Geister, J., López-Victoria, M., Ospina, G. H., Parra-Velandia, F., Pinzón, J., Zapata, B. & Zea, S. (2000). *Áreas coralinas de Colombia*. INVEMAR, Serie Publicaciones Especiales No. 5, Santa Marta (176 p.).

Díaz-Ferguson, E. E., & Moyer, G. R. (2014). History, applications, methodological issues and perspectives for the use environmental DNA (eDNA) in marine and freshwater environments. *Revista de Biología Tropical*, *62*(4), 1273–1284. https://doi.org/10.15517/rbt.v62i4.13231

DiBattista, J. D., Coker, D. J., Sinclair-Taylor, T. H., Stat, M., Berumen, M. L., & Bunce, M. (2017). Assessing the utility of eDNA as a tool to survey reef-fish communities in the Red Sea. *Coral Reefs*, *36*(4), 1245–1252. https://doi.org/10.1007/s00338-017-1618-1

Djurhuus, A., Closek, C. J., Kelly, R. P., Pitz, K. J., Michisaki, R. P., Starks, H. A., Walz, K. R., Andruszkiewicz, E. A., Olesin, E., Hubbard, K., & Montes, E. (2020). Environmental DNA reveals seasonal shifts and potential interactions in a marine community. *Nature Communications*, *11*(1), 1–9. https://doi.org/10.1038/s41467-019-14105-1

Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., Gielly, L., Lopes, C. M., Boyer, F., Pompanon, F., & Rayé, G. (2015). Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources*, *15*(3), 543–556. https://doi.org/10.1111/1755-0998.12338

Fisher, R., O'Leary, R. A., Low-Choy, S., Mengersen, K., Knowlton, N., Brainard, R. E., & Caley, M. J. (2015). Species richness on coral reefs and the pursuit of convergent global estimates. *Current Biology*, *25*(4), 500–505. https://doi.org/10.1016/j.cub.2014.12.022

Freeland, J. R. (2017). The importance of molecular markers and primer design when characterizing biodiversity from environmental DNA. *Genome*, *60*(4), 358–374. https://doi.org/10.1139/gen-2016-0100

Fricke, R., Eschmeyer W. N., & Van der Laan R. (Eds.) (2020). *Eschmeyer's catalog of fishes: Genera, species, references*. Retrieved from http://researcharchive.calacademy.org/research/ichthyology/catalog/fishcatmain.asp

Fritz, S. A., & Purvis, A. (2010). Selectivity in mammalian extinction risk and threat types: A new measure of phylogenetic signal strength in binary traits. *Conservation Biology*, *25*(4), 1042–1051. https://doi.org/10.1111/j.1523-1739.2010.01455.x

Froese, R., & Pauly, D. (Eds.) (2018). *FishBase*. World Wide Web electronic publication. Retrieved from www.fishbase.org, version (06/2018).

Frøslev, T. G., Kjøller, R., Bruun, H. H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., & Hansen, A. J. (2017). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nature Communications*, *8*(1), 1–11. https://doi.org/10.1038/s41467-017-01312-x

Garzón-Ferreira, J., & Cano, M. (1991). *Tipos, distribución, extensión y estado de conservación de los ecosistemas marinos costeros del Parque Nacional Natural Tayrona* (82 p). Instituto de Investigaciones Marinas y Costeras "José Benito Vives de Andréis" (Invemar).

Garzón-Ferreira, J., & Díaz, J. M. (2003). The Caribbean coral reefs of Colombia. In J. Cortés (Ed.), *Latin American coral reefs* (pp. 275–301). Elsevier Science.

Garzón-Ferreira, J., Reyes-Nivia, M., & Rodríguez-Ramírez, A. (2002). *Manual de métodos del SIMAC: Sistema Nacional de Monitoreo de Arrecifes Coralinos en Colombia* (102 p.). INVEMAR.

Geister, J. (1992). Modern reef development and Cenozoic evolution of an oceanic island/reef complex: Isla de Providencia (Western Caribbean Sea, Colombia). *Facies*, *27*(1), 1–69. https://doi.org/10.1007/BF02536804

Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., Spear, S. F., McKee, A., Oyler-McCance, S. J., Cornman, R. S., & Laramie, M. B. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution*, *7*(11), 1299–1307.

Gomes, G. B., Hutson, K. S., Domingos, J. A., Chung, C., Hayward, S., Miller, T. L., & Jerry, D. R. (2017). Use of environmental DNA (eDNA) and water quality data to predict protozoan parasites outbreaks in fish farms. *Aquaculture*, *479*, 467–473.

Hänfling, B., Lawson Handley, L., Read, D. S., Hahn, C., Li, J., Nichols, P., Blackman, R. C., Oliver, A., & Winfield, I. J. (2016). Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Molecular Ecology*, *25*(13), 3101–3119. https://doi.org/10.1111/mec.13660

Hansen, B. K., Bekkevold, D., Clausen, L. W., & Nielsen, E. E. (2018). The sceptical optimist: Challenges and perspectives for the application of environmental DNA in marine fisheries. *Fish and Fisheries*, *19*(5), 751–768. https://doi.org/10.1111/faf.12286

Harrison, J. B., Sunday, J. M., & Rogers, S. M. (2019). Predicting the fate of the eDNA in the environment and implications for studying

biodiversity. *Proceedings of the Royal Society B: Biological Sciences*, *286*, 20191409. https://doi.org/10.1098/rspb.2019.1409

Huerlimann, R., Cooper, M. K., Edmunds, R. C., Villacorta-Rath, C., Le Port, A., Robson, H. L. A., Strugnell, J. M., Burrows, D., & Jerry, D. R. (2020). Enhancing tropical conservation and ecology research with aquatic environmental DNA methods: An introduction for non-environmental DNA specialists. *Animal Conservation*. https://doi.org/10.1111/acv.12583

Huse, S. M., Welch, D. M., Morrison, H. G., & Sogin, M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environmental Microbiology*, *12*(7), 1889–1898. https://doi.org/10.1111/j.1462-2920.2010.02193.x

Jerde, C. L., Wilson, E. A., & Dressler, T. L. (2019). Measuring global fish species richness with eDNA metabarcoding. *Molecular Ecology Resources*, *19*(1), 19–22.

Juhel, J. B., Utama, R. S., Marques, V., Vimono, I. B., Sugeha, H. Y., Kadarusman, K., Pouyaud, L., Dejean, T., Mouillot, D., & Hocdé, R. (2020). Accumulation curves of environmental DNA sequences predict coastal fish diversity in the coral triangle. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1930), 20200248. https://doi.org/10.1098/rspb.2020.0248

Juhel, J. B., Vigliola, L., Mouillot, D., Kulbicki, M., Letessier, T. B., Meeuwig, J. J., & Wantiez, L. (2018). Reef accessibility impairs the protection of sharks. *Journal of Applied Ecology*, *55*(2), 673–683. https://doi.org/10.1111/1365-2664.13007

Kelly, R. P., Port, J. A., Yamahara, K. M., & Crowder, L. B. (2014). Using environmental DNA to census marine fishes in a large mesocosm. *PLoS One*, *9*(1), e86175. https://doi.org/10.1371/journal.pone.0086175

Knudsen, S. W., Ebert, R. B., Hesselsøe, M., Kuntke, F., Hassingboe, J., Mortensen, P. B., Thomsen, P. F., Sigsgaard, E. E., Hansen, B. K., Nielsen, E. E., & Møller, P. R. (2019). Species-specific detection and quantification of environmental DNA from marine fishes in the Baltic Sea. *Journal of Experimental Marine Biology and Ecology*, *510*, 31–45. https://doi.org/10.1016/j.jembe.2018.09.004

Lacoursière-Roussel, A., Rosabal, M., & Bernatchez, L. (2016). Estimating fish abundance and biomass from eDNA concentrations: Variability among capture methods and environmental conditions. *Molecular Ecology Resources*, *16*(6), 1401–1414. https://doi.org/10.1111/1755-0998.12522

Leprieur, F., Colosio, S., Descombes, P., Parravicini, V., Kulbicki, M., Cowman, P. F., Bellwood, D. R., Mouillot, D., & Pellissier, L. (2016). Historical and contemporary determinants of global phylogenetic structure in tropical reef fish faunas. *Ecography*, *39*(9), 825–835. https://doi.org/10.1111/ecog.01638

MacConaill, L. E., Burns, R. T., Nag, A., Coleman, H. A., Slevin, M. K., Giorda, K., Light, M., Lai, K., Jarosz, M., McNeill, M. S., & Ducar, M. D. (2018). Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics*, *19*(1), 1–10. https://doi.org/10.1186/s12864-017-4428-5

Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2014). Swarm: Robust and fast clustering method for amplicon-based studies. *PeerJ*, *2*, e593. https://doi.org/10.7717/peerj.593

Marques, V., Guérin, P.-E., Rocle, M., Valentini, A., Manel, S., Mouillot, D., & Dejean, T. (2020). Blind assessment of vertebrate taxonomic diversity across spatial scales by clustering environmental DNA metabarcoding sequences. *Ecography*. https://doi.org/10.1111/ecog.05049

Márquez, G. (1987). *Las islas de Providencia y Santa Catalina* (110 p.). Fondo Fen-Univ. Nacional de Colombia.

Martin, M. (2013). Cutadapt removes adapter sequences from high-throughput sequencing reads 2011. *EMBnet.Journal*, *17*(1):10. https://doi.org/10.14806/ej.17.1.200

Mejía, L. S., & Garzón-Ferreira, J. (2000). Estructura de comunidades de peces arrecifales en cuatro atolones del Archipiélago de San Andrés y Providencia (Caribe sur occidental). *Revista de Biología Tropical*, *48*(4), 883–896.

Miya, M., Sato, Y., Fukunaga, T., Sado, T., Poulsen, J. Y., Sato, K., Minamoto, T., Yamamoto, S., Yamanaka, H., Araki, H., & Kondoh, M. (2015). MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: Detection of more than 230 subtropical marine species. *Royal Society Open Science*, *2*(7), 150088. https://doi.org/10.1098/rsos.150088

Mora, C., Tittensor, D. P., & Myers, R. A. (2008). The completeness of taxonomic inventories for describing the global diversity and distribution of marine fishes. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1631), 149–155. https://doi.org/10.1098/rspb.2007.1315

Morgan, M. J., Chariton, A. A., Hartley, D. M., Court, L. N., & Hardy, C. M. (2013). Improved inference of taxonomic richness from environmental DNA. *PLoS One*, *8*(8), e71974. https://doi.org/10.1371/journal.pone.0071974

Nevers, M. B., Byappanahalli, M. N., Morris, C. C., Shively, D., Przybyla-Kelly, K., Spoljaric, A. M., Dickey, J., & Roseman, E. F. (2018). Environmental DNA (eDNA): A tool for quantifying the abundant but elusive round goby (*Neogobius melanostomus*). *PLoS One*, *13*(1), e0191720. https://doi.org/10.1371/journal.pone.0191720

Nguyen, B. N., Shen, E. W., Seemann, J., Correa, A. M., O'Donnell, J. L., Altieri, A. H., Knowlton, N., Crandall, K. A., Egan, S. P., McMillan, W. O., & Leray, M. (2020). Environmental DNA survey captures patterns of fish and invertebrate diversity across a tropical seascape. *Scientific Reports*, *10*(1), 1–14. https://doi.org/10.1038/s41598-020-63565-9

Plaisance, L., Caley, M. J., Brainard, R. E., & Knowlton, N. (2011). The diversity of coral reefs: What are we missing? *PLoS One*, *6*(10), e25026. https://doi.org/10.1371/journal.pone.0025026

Pont, D., Rocle, M., Valentini, A., Civade, R., Jean, P., Maire, A., Roset, N., Schabuss, M., Zornig, H., & Dejean, T. (2018). Environmental DNA reveals quantitative patterns of fish biodiversity in large rivers despite its downstream transportation. *Scientific Reports*, *8*(1), 1–13. https://doi.org/10.1038/s41598-018-28424-8

Rabosky, D. L., Chang, J., Title, P. O., Cowman, P. F., Sallan, L., Friedman, M., Kaschner, K., Garilao, C., Near, T. J., Coll, M., & Alfaro, M. E. (2018). An inverse latitudinal gradient in speciation rate for marine fishes. *Nature*, *559*(7714), 392–395.

Reeder, J., & Knight, R. (2009). The 'rare biosphere': A reality check. *Nature Methods*, *6*(9), 636–637. https://doi.org/10.1038/nmeth0909-636

Robertson, D. R., & Van Tassell, J. V. (2019). *Shorefishes of the Greater Caribbean: Online information system. Version 2.0*. Smithsonian Tropical Research Institute.

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584

Samoilys, M. A., & Carlos, G. (2000). Determining methods of underwater visual census for estimating the abundance of coral reef fishes. *Environmental Biology of Fishes*, *57*(3), 289–304.

Sánchez, J. A., Zea, S. V. E. N., & Díaz, J. M. (1998). Patterns of octocoral and black coral distribution in the oceanic barrier reef-complex of Providencia Island, Southwestern Caribbean. *Caribbean Journal of Science*, *34*(3–4), 250–264.

Schenekar, T., Schletterer, M., Lecaudey, L. A., & Weiss, S. J. (2020). Reference databases, primer choice, and assay sensitivity for environmental metabarcoding: Lessons learnt from a re-evaluation of an eDNA fish assessment in the Volga headwaters. *River Research and Applications*.*36*(7), 1004–1013.

Schnell, I. B., Sollmann, R., Calvignac-Spencer, S., Siddall, M. E., Douglas, W. Y., Wilting, A., & Gilbert, M. T. P. (2015). iDNA from terrestrial haematophagous leeches as a wildlife surveying and monitoring tool–prospects, pitfalls and avenues to be developed. *Frontiers in Zoology*, *12*(1), 24. https://doi.org/10.1186/s12983-015-0115-z

Sigsgaard, E. E., Torquato, F., Frøslev, T. G., Moore, A. B., Sørensen, J. M., Range, P., Ben-Hamadou, R., Bach, S. S., Møller, P. R., & Thomsen, P.

F. (2019). Using vertebrate environmental DNA from seawater in biomonitoring of marine habitats. *Conservation Biology*, *34*(3), 697–710. https://doi.org/10.1111/cobi.13437

Stat, M., Huggett, M. J., Bernasconi, R., DiBattista, J. D., Berry, T. E., Newman, S. J., Harvey, E. S., & Bunce, M. (2017). Ecosystem biomonitoring with eDNA: Metabarcoding across the tree of life in a tropical marine environment. *Scientific Reports*, *7*(1), 12240. https://doi.org/10.1038/s41598-017-12501-5

Stat, M., John, J., DiBattista, J. D., Newman, S. J., Bunce, M., & Harvey, E. S. (2019). Combined use of eDNA metabarcoding and video surveillance for the assessment of fish biodiversity. *Conservation Biology*, *33*(1), 196–205. https://doi.org/10.1111/cobi.13183

Taberlet, P., Bonin, A., Coissac, E., & Zinger, L. (2018). *Environmental DNA: For biodiversity research and monitoring*. Oxford University Press.

Thomsen, P. F., Kielgast, J., Iversen, L. L., Møller, P. R., Rasmussen, M., & Willerslev, E. (2012). Detection of a diverse marine fish fauna using environmental DNA from seawater samples. *PLoS One*, *7*(8), e41732. https://doi.org/10.1371/journal.pone.0041732

Thomsen, P. F., Møller, P. R., Sigsgaard, E. E., Knudsen, S. W., Jørgensen, O. A., & Willerslev, E. (2016). Environmental DNA from seawater samples correlate with trawl catches of subarctic, deepwater fishes. *PLoS One*, *11*(11), e0165252. https://doi.org/10.1371/journal.pone.0165252

Thomsen, P. F., & Willerslev, E. (2015). Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation*, *183*, 4–18. https://doi.org/10.1016/j.biocon.2014.11.019

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., Bellemain, E., Besnard, A., Coissac, E., Boyer, F., Gaboriaud, C., Jean, P., Poulet, N., Roset, N., Copp, G. H., Geniez, P., Pont, D., Argillier, C., Baudoin, J.-M., … Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, *25*(4), 929–942. https://doi.org/10.1111/mec.13428

Weigand, H., Beermann, A. J., Čiampor, F., Costa, F. O., Csabai, Z., Duarte, S., Geiger, M. F., Grabowski, M., Rimet, F., Rulik, B., Strand, M., Szucsich N., Weigand A. M., Willassen E., Wyler S. A., Bouchez A., Borja A., Čiamporová-Zaťovičová Z., Ferreira S., Dijkstra K. B., … Ekrem T. (2019). DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. *Science of The Total Environment*, *678*, 499–524. https://doi.org/10.1016/j.scitotenv.2019.04.247

Weltz, K., Lyle, J. M., Ovenden, J., Morgan, J. A., Moreno, D. A., & Semmens, J. M. (2017). Application of environmental DNA to detect an endangered marine skate species in the wild. *PLoS One*, *12*(6), e0178124. https://doi.org/10.1371/journal.pone.0178124

West, K. M., Stat, M., Harvey, E. S., Skepper, C. L., DiBattista, J. D., Richards, Z. T., Travers, M. J., Newman, S. J., & Bunce, M. (2020). eDNA metabarcoding survey reveals fine-scale coral reef community variation across a remote, tropical island ecosystem. *Molecular Ecology*, *29*(6), 1069–1086. https://doi.org/10.1111/mec.15382

Williams, G. J., Graham, N. A., Jouffray, J. B., Norström, A. V., Nyström, M., Gove, J. M., Heenan, A., & Wedding, L. M. (2019). Coral reef ecology in the Anthropocene. *Functional Ecology*, *33*(6), 1014–1022.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Polanco Fernández A, Marques V, Fopp F, et al. Comparing environmental DNA metabarcoding and underwater visual census to monitor tropical reef fishes. *Environmental DNA*. 2020;00:1–15. https://doi.org/10.1002/edn3.140

Chapitre 5 – Circumglobal distribution of fish environmental DNA on coral reefs

Manuscrit E

Mathon, L., Marques, V.*, Mouillot, D., Albouy, C., Andrello, M., Baletaud, F., Borrero, G.H., Dejean, T., Edgar, G., Grondin, J., Guérin, P-E., Hocdé, R., Juhel, J-B., Kadarusman, Maire, E., Mariani, G., McLean, M., Polanco, A., Pouyaud, L., Stuart-Smith, R., Sugeha, H.Y., Valentini, A., Vigliola, L., Vimono, I.B., Pellissier, L., Manel, S. Circumglobal distribution of fish environmental DNA on coral reefs. (2020). Submitted.*

**shared first authorship*

1. Préface

La vitesse et l'ampleur du changement global requièrent une capacité de suivi de la biodiversité fortement supérieure aux efforts fournis actuellement à la fois dans la rapidité et l'exhaustivité. L'utilisation d'une méthode ne nécessitant pas de personnel qualifié pour les prélèvements ainsi que d'une logistique légère sur le terrain comme avec l'ADNe metabarcoding a le potentiel de permettre une échelle d'observation de la biodiversité jamais atteinte auparavant. Le chapitre 4 a montré que l'ADNe est une méthode adaptée même aux milieux marins tropicaux hyper-divers, avec une capacité de recensement des poissons supérieure aux transects en plongée et aux observations par vidéo, où deux échantillons ADNe détectent autant de diversité spécifique que 25h de vidéos. L'ADNe détecte aussi des groupes taxonomiques et fonctionnels couramment ratés par les méthodes traditionnelles, apportant des informations importantes sur la structure des communautés dans les écosystèmes depuis les organismes les plus abondants jusqu'aux espèces les plus rares et souvent occultées par les méthodes de recensements classiques.

Les applications de l'ADNe en milieu marin se limitent généralement à l'échelle locale, parfois régionale. Aucune étude n'a été menée à grande échelle ou n'a utilisé les avantages de l'ADNe pour tester des hypothèses écologiques d'assemblage des communautés. La méthodologie générant des unités taxonomiques moléculaires développée au chapitre 3 pour s'affranchir de la couverture taxonomique des bases de référence offre une alternative permettant de répondre à des questions à l'échelle biogéographique. Les récifs coralliens sont des écosystèmes très riches en espèces de poissons (2 400 à 8 000 espèces) dont la diversité est très structurée spatialement, permettant de tester le potentiel du metabarcoding de l'ADNe à large échelle.

Ce chapitre vise à (i) présenter une étude pilote sur la distribution des fragments d'ADNe des poissons à large échelle, (ii) comparer la capacité de détection de l'ADNe avec la plus grande base de données de transects en plongées sur les récifs corallins et (iii) tester des hypothèses d'assemblages des communautés à travers l'étude de la rareté et du partitionnement de la diversité à différentes échelles spatiales.

On montre ici que les fragments d'ADNe sont capables de retrouver les patrons bien connus de la biogéographie des poissons coralliens, avec une diversité plus élevée au niveau du triangle de corail et décroissante en s'en éloignant. Avec un nombre de familles et de MOTUs détectés plus élevé avec l'ADNe qu'avec les transects, l'ADNe retrouve les mêmes patrons de stabilité des proportions de familles

au sein des récifs coralliens : moins de 10% des MOTUs sont assignés à des Caranguidae ou Serranidae et entre 10 et 20% sont assignés à des Labridae qu'importe la richesse de chaque site, ce qui cohérent avec des études précédentes à l'échelle de l'Indo-Pacifique (Bellwood and Hughes 2001). L'étude du partitionnement de la diversité inter-échelles montre que l'ADNe révèle plus de beta-diversité inter-stations que les plongées, ce qui indique que l'ADNe est potentiellement plus apte à discerner des différences de communautés à fine échelle grâce à une capacité de détection plus élevée pour les espèces rares. Ces résultats révèlent le fort potentiel de l'ADNe pour les suivis de biodiversité et pour l'intégration de ces données standardisées dans un observatoire global.

2. Manuscrit E

Circumglobal distribution of fish environmental DNA on coral reefs

Laetitia Mathon[1,2,8*], Virginie Marques[1,3*], David Mouillot[3,4] , Camille Albouy[5] , Marco Andrello[3] , Florian Baletaud[2,3,6], Giomar H. Borrero-Pérez[7], Tony Dejean[8], Graham J. Edgar[9], Jonathan Grondin[8], Pierre-Edouard Guerin[1], Régis Hocdé[3], Jean-Baptiste Juhel[3], Kadarusman[10], Eva Maire[3,11], Gael Mariani[3] , Matthew McLean[12,13], Andrea Polanco F.[7], Laurent Pouyaud[13], Rick D. Stuart-Smith[9], Hagi Yulia Sugeha[14], Alice Valentini[8], Laurent Vigliola[2], Indra B Vimono[14], Loïc Pellissier[15,16**], Stéphanie Manel[1**]

[1] CEFE, Univ. Montpellier, CNRS, EPHE-PSL University, IRD, Univ. Paul Valéry Montpellier 3, Montpellier, France

[2] ENTROPIE, Institut de Recherche pour le Développement (IRD), Univ. Réunion, UNC, CNRS, IFREMER, Centre IRD de Nouméa, Nouméa, New Caledonia, France

[3] MARBEC, Univ Montpellier, CNRS, IFREMER, IRD, Montpellier, France

[4] Institut Universitaire de France

[5] IFREMER, unité Écologie et Modèles pour l'Halieutique, rue de l'Ile d'Yeu, BP21105, 44311 Nantes cedex 3, France.

[6] SOPRONER, groupe GINGER, 98000 Noumea, New Caledonia, France

[7] Instituto de Investigaciones Marinas y Costeras-INVEMAR, Colombia. Museo de Historia Natural Marina de Colombia (MHNMC), Programa de Biodiversidad y Ecosistemas Marinos. Calle 25 No. 2 – 55 Playa Salguero, Santa Marta, Colombia.

[8] SPYGEN, Le Bourget-du-Lac, France

[9] Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, Tasmania, Australia

[10] Politeknik Kelautan dan Perikanan Sorong, KKD BP Sumberdaya Genetik, Konservasi dan Domestikasi, Papua Barat, Indonesia

[11] Lancaster Environment Centre, Lancaster University, Lancaster, LA1 4YQ, UK

[12] Department of Biology, Dalhousie University, Halifax, NSB3H4R2, Canada

[13] ISEM, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France

[14] Research Center for Oceanography, Indonesian Institute of Sciences Jl. Pasir Putih 1, Ancol Timur, Jakarta Utara 14430 Jakarta Pusat – Indonesia

[15] Landscape Ecology, Institute of Terrestrial Ecosystems, Department of Environmental Systems Science, ETH Zürich, Zürich, Switzerland

[16] Unit of Land Change Science, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

* These authors contributed equally as first author to this work

** These authors contributed equally as senior author to this work

Abstract

The increasing speed and magnitude of global change requires effective approaches to monitor the world's biodiversity. While remote sensing and automated receptors allow instantaneous observations of habitats and physical environments, similar biodiversity monitoring over equivalent scales has not been possible to date, particularly in the vast ocean. Here, we demonstrate the capacity of environmental DNA (eDNA) metabarcoding from water samples to reconstruct well-known patterns of fish diversity on one of the most diverse and threatened ecosystems on the planet, coral reefs. Additionally, eDNA reveals a higher fish species (16%) and family (50%) diversity than estimates obtained with a dataset of standardized underwater visual surveys carried out at 20x more sites. eDNA also shows that fish species dissimilarity among adjacent coral reefs is higher than detected with visual surveys and that rarity is more prevalent than previously thought. Our study demonstrates how sequencing DNA from water samples provides a rapid and effective approach to characterize and monitor coral reef fish diversity, and to uncover hidden patterns. eDNA approaches can be applied by large research consortia and non-experts, providing the foundation for a global observatory network for efficiently tracking the effects of global change and providing a new lens for testing ecological paradigms.

Main text

Decrypting the diversity of life on Earth from organisms to ecosystems requires multiple layers of information across scales, only now becoming possible using high throughput approaches from novel molecular methods[1,2]. Our present understanding of the world's changing biodiversity is minute compared to that of the earth's physical environment obtained through remote sensing and automated monitoring systems[3,4]. Environmental DNA (eDNA) metabarcoding, a method retrieving and analyzing DNA naturally released by organisms in their environment[5], has the potential to revolutionize the monitoring of biodiversity[6]. eDNA has proven particularly useful for aquatic ecosystems[7] and is now well established for microorganisms[2,8]. By contrast, its potential to provide an integrated biodiversity assessment of macro-organisms, including vertebrates of high trophic level, has not yet emerged at large spatial scale.

Intensifying global changes, such as overexploitation, pollution and climate warming, are impacting the distribution and diversity of vertebrates more rapidly than our monitoring capacity can cover using classical surveys[9], particularly in the under investigated global ocean[10]. Marine vertebrates are the main targets of fisheries[11] and are especially sensitive to pollution[12] and climate change with many species on the brink of extinction[13,14]. We thus need more efficient, large-scale, replicable, real-time biodiversity monitoring to inform rapid transition from knowledge to conservation action[2]. High-throughput eDNA

metabarcoding may provide the only means to obtain global synoptic snapshots of biodiversity over the sub-yearly time scales that have only been possible previously for physical variables. However, validation is needed before broad-scale systematic application of eDNA protocols, to better appreciate how these new approaches compare with classical biodiversity observation systems.

Coral reefs host the highest fish diversity on Earth despite covering less than 0.1% of the ocean's surface but are also severely threatened[15]. Data syntheses over decades of surveys estimate the total number of coral reef fishes from 2,400 to 8,000 species[16,17], distributed among approximately 100 families[18]. Typically, this diversity of coral reef fishes displays clear spatial patterns, including longitudinal and latitudinal gradients which peak in the Indo-Australian Archipelago[19,20] also known as the 'coral triangle' with the world's highest level of marine diversity[21]. The proportions of fish species among families are strongly conserved across the Indo-Pacific19. The spatial diversity gradient in coral reef fishes is also marked by strong variations in taxonomic composition (species turnover or β diversity), dominated by species replacement across space[22], with many species on coral reefs being rare and geographically localized, yet sometimes locally abundant[23].

Coral reef fishes have also evolved in a physically complex environment and present a wide range of forms and functions[24,25]. Small cryptic species that live inside the reef structure can be very difficult to sample or survey using non-destructive methods. Such 'cryptobenthic' fishes represent half of fish diversity on coral reefs[26] but are missed by most classical surveys[27]. Even though fish are among the best-studied taxa inhabiting coral reefs[28], our knowledge of this biodiversity is only partial[29]; the taxonomy is complex and uncertain for many described species[16] and countless species remain to be described. eDNA represents an opportunity to not only better understand classical biodiversity patterns, but also uncover novel ones hidden by our incomplete taxonomic and biogeographic knowledge.

Here, we investigated how a global snapshot of hierarchically sampled eDNA could describe the distribution of fish diversity on coral reefs. We generated 504,457,267 raw 12S ribosomal DNA (rDNA) sequence reads from 251 samples (2,693 PCR replicates) collected at 25 sites in 145 stations covering five regions across the Indian, Pacific and Atlantic Oceans (Extended Data Fig. 1). Bioinformatic quality control[30] produced a final dataset of 335,223,143 sequence reads (see Methods), clustered into 2,160 molecular operational taxonomic units (MOTUs), which were taxonomically assigned to Actinopterygii (bony fishes) and Chondrichthyes (cartilaginous fishes) taxa (Extended Data Table 1) using a public genetic reference database (see Methods). We then compared diversity patterns obtained from eDNA to those observed in the most extensive global data set of standardized visual surveys of reef fishes (Reef Life Survey, visual census data[31]). Visual census data include observations of all fish species

recorded on shallow coral reefs, including crypto-benthic species, and have previously been used to describe global patterns in reef fish biodiversity[32].

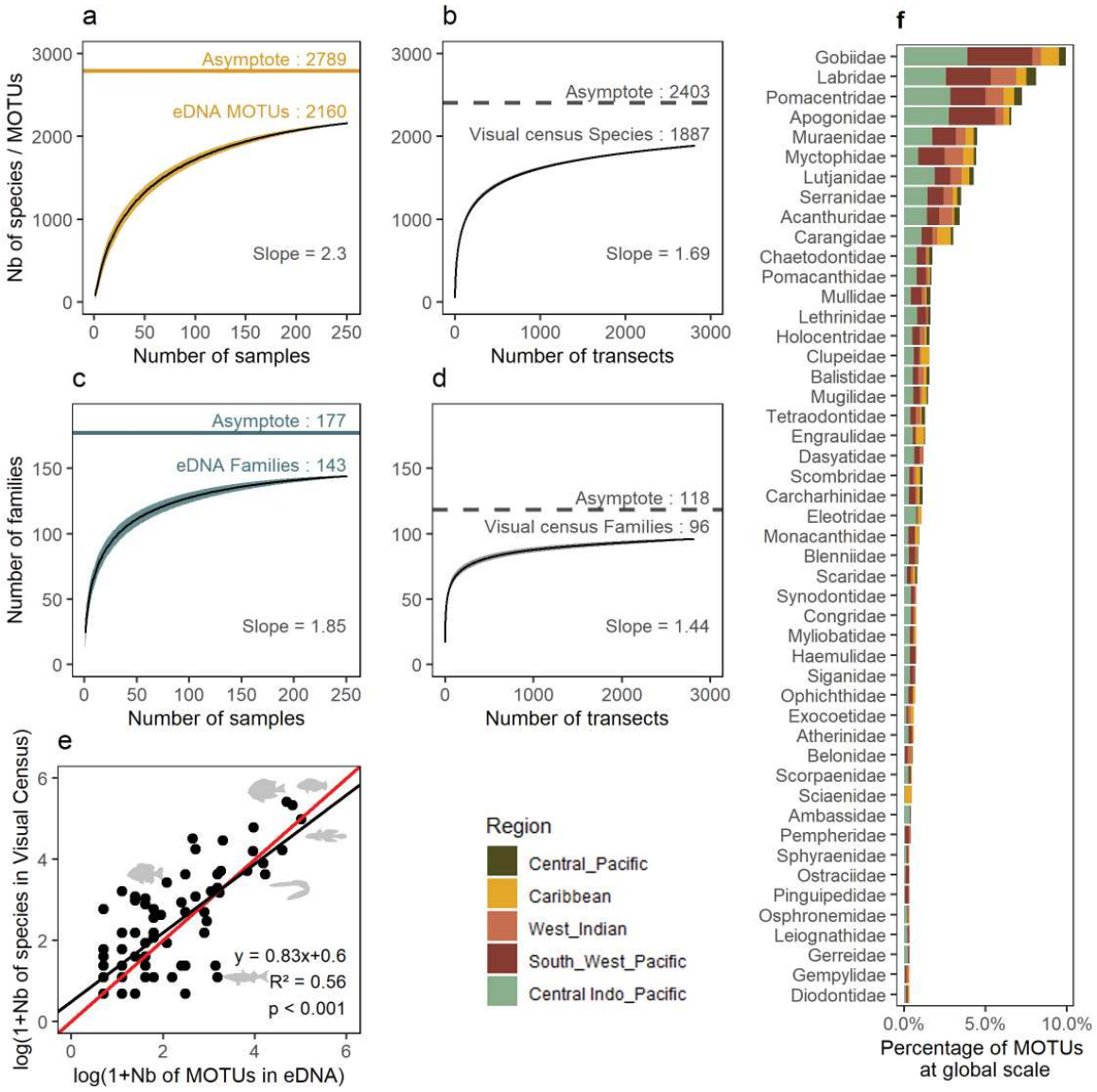


Figure 1 | Estimations of overall fish richness from environmental DNA (eDNA) and visual census. a, accumulation curve of molecular operational taxonomic units from eDNA (eDNA MOTUs), b, accumulation curve of species from the visual census database, c, accumulation curve of eDNA families, d, accumulation curve of visual census families. For a-d, Species accumulation model is fitted according to Lomolino method (see methods). e, linear regression (black line) between the number of species per family in visual census data and the number of MOTUs per family in eDNA (log(x+1) transformation) over n = 77 families. Each point is a family. Red line is x=y. f, percentage of MOTUs assigned to each family at global scale, and proportion in each region.

Estimates of fish biodiversity on coral reefs

We estimated fish diversity from the asymptote of a multi-model accumulation curve for both eDNA MOTUs[33] and visual census species (see Methods). Estimates obtained from 145 eDNA stations sampled over a 28-month period show that detectable fish diversity is 16% higher (asymptote at 2,789 MOTUs, Fig. 1a) than the equivalent estimate from the visual census data, which reaches an asymptote at 2,403 fish species from 2813 transects surveyed during 13 years (Fig. 1b). The asymptotic estimation of family richness obtained with eDNA reaches 177 families, 50% more than the asymptotic number of families estimated with visual census data (118 families, Fig. 1 c,d). Among the 71 families shared between both datasets, 25 have a higher number of MOTUs from eDNA than species from visual census (Fig. 1e). These families with more taxa detected using eDNA include those often associated with reef-adjacent habitats such as mangroves or soft sediments like Mugilidae, Elopidae and Gerreidae[34], and cryptobenthic species that live hidden in crevices like Gobiescocidae or nocturnal species like Congridae[35]. eDNA also reveals higher richness of pelagic and wide-ranging species, with 10x more of these taxa observed than with visual census. These include members of Scombridae, Clupeidae, Carcharhinidae and Belonidae (Extended Data Fig. 2). Many pelagic fish likely avoid divers or are not resident on the reef for long enough to be consistently detected during visual surveys[36]. As eDNA is rapidly degraded in tropical inshore waters[7,37], we assume the eDNA signal comes from individuals present in close proximity to the station. Thus, the detection of species not typically considered as coral reef fishes reveals their use of reef habitats from time to time[38]. Nevertheless, some evidence suggests that eDNA from pelagic fishes degrades slower than from inshore species[37], and the detection of some deep-water families (e.g. Myctophidae) additionally suggests that eDNA might disperse sufficiently with sea currents such that species immediately adjacent to reef habitats are detected. Whether and how such taxa (under-detected by classical surveys) contribute to reef ecology, through pelagic larval stages or nocturnal migration up the reef slope[39–41] represents an interesting avenue of further studies.

MOTU richness per family retrieved with eDNA closely matches fish species richness within families recorded in visual census data (Pearson correlation = 0.84, $p < 0.001$, $n = 71$, Fig. 1e). Highly diverse families seen on coral reefs are also well represented in eDNA, with Gobiidae, Labridae and Pomacentridae containing more than 100 MOTUs each, together representing about 18% of MOTUs (Fig. 1f, Extended Data Figs. 3 and 4). Both approaches recover highly congruent taxonomic diversity gradients. However, our eDNA-based survey allows a faster compilation of fish diversity on coral reefs, and removes some limitations of traditional visual surveys.

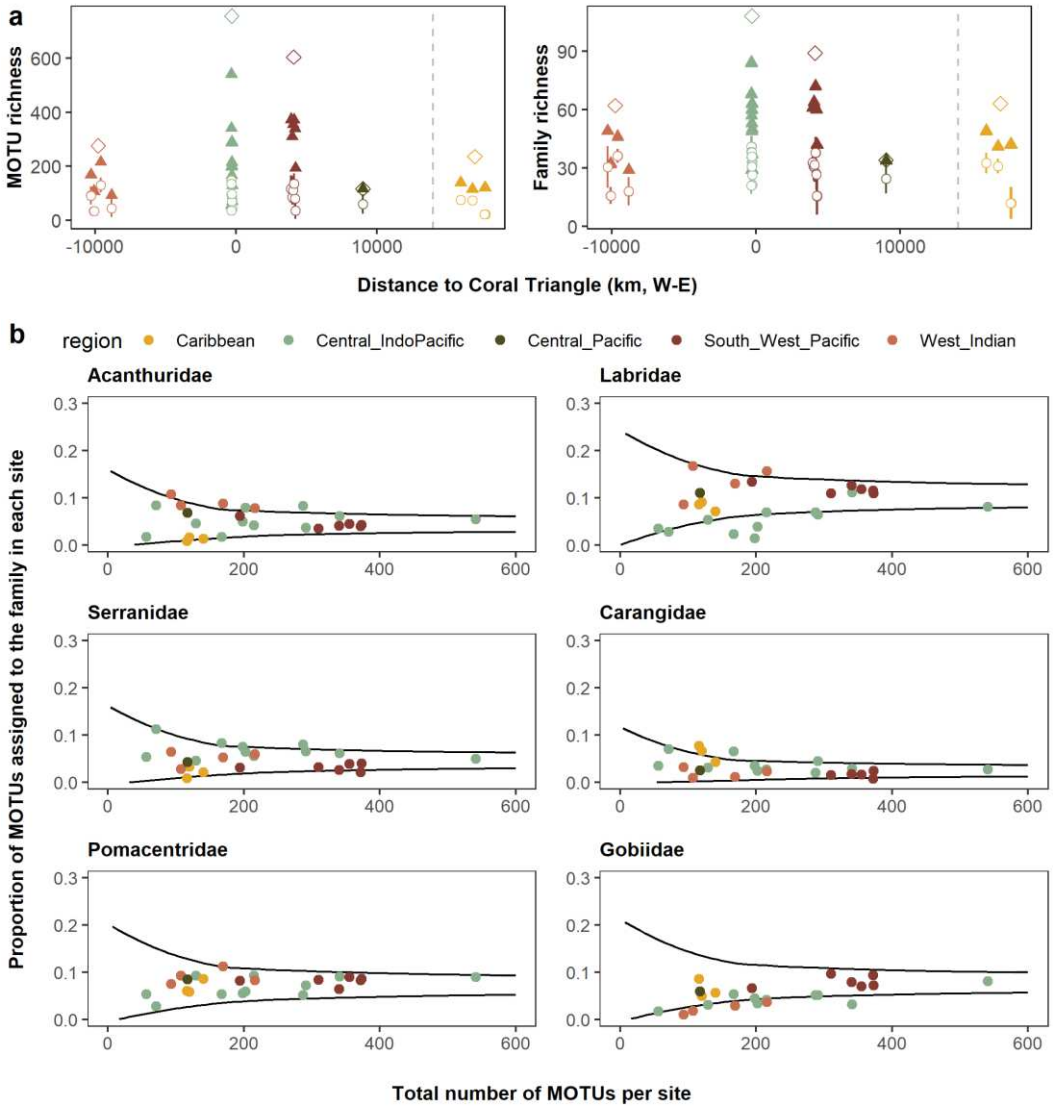


Figure 2 | Geographic patterns of fish MOTUs. a, mean MOTUs (left) and mean Family (right) richness per station in each site ± standard deviation (empty circles and vertical bars), total site richness (filled triangles) and total region richness (empty diamonds) as a function of the distance from the center of the coral triangle (in km); the vertical dashed line represents the delimitation between the Indo-Pacific and the Atlantic Ocean basins. Kruskal-Wallis test showed significant differences in site MOTU richness between regions (Dunn post-hoc test showed Central Indo-Pacific and South-West Pacific richest than the three other regions). b, proportion of MOTUs assigned to some of the most represented families, in each site, as a function of total MOTU richness of the sites. Families were selected among most represented strictly reef-associated, pelagic and cryptobenthic families. Upper and lower lines are bootstrapped 95% confidence limits based on random selection of species from the total species pool.

Biogeography of eDNA metabarcoding sequences

The spatial distribution of MOTUs follows clear and well-known biogeographic patterns, with a peak in the coral triangle and lower values of MOTU richness toward the Central Pacific (Fig. 2a, Kruskal Wallis test among sites: $p < 0.001$, n=25, Dunn test of pairwise comparisons: $p < 0.001$). The richest region

sampled in this study (Lengguru, Indonesia, Central Indo-Pacific) contains ~50% of the global pool of fish MOTUs while the poorest regions (Fakarava, French Polynesia, Central Pacific) contains only 11% of the global pool (Extended Data Fig. 5a and Table 2). The proportion of MOTUs belonging to the best-represented families is similar among the sites, suggesting family stability across space, regardless of local diversity (Fig. 2b). The proportion of MOTUs per family lies within confidence intervals predicted by random allocation of species in each site from the global pool of MOTUs (Fig. 2b). The pattern of family stability shown across the Indo-Pacific[19] can thus be generalized towards the Atlantic Ocean (Caribbean) hosting the similar richest families: Gobiidae, Labridae, Pomacentridae and Apogonidae (Fig. 1f, Extended Data Fig. 6). This global pattern of family stability across space supports the hypothesis that large-scale assembly rules determine species composition of coral reef fishes[19,29] as the proportion of MOTUs within family in each site could be simply predicted from a random allocation of MOTUs from the global pool. This pattern also indicates a complementarity for essential ecosystem functions performed by these families on the reefs in each site[19], and also a highly variable level of redundancy in the functions performed by species within families among sites with species-poor sites potentially more vulnerable to functional extinction[24]. Our results suggest that eDNA can recover the signature of ecological and evolutionary processes on the spatial organization of diversity on coral reefs globally and can also inform conservation strategies.

Global patterns of fish rarity and turnover

Our eDNA sampling shows that a majority of MOTUs are rare and geographically localized, with 80% of the MOTUs detected in only one region (Fig. 3a), and 30% in only one site (Extended Data Fig. 7). We hierarchically partition the global MOTU diversity (γ_{global}) into additive diversity components (i.e. dissimilarity) due to differences between regions ($\beta_{inter\text{-}region}$), mean differences between sites within regions ($\bar{\beta}_{inter\text{-}site}$), mean differences between stations within sites ($\bar{\beta}_{inter\text{-}station}$) and mean station diversity ($\bar{\alpha}_{station}$)[42]. As a consequence of the rarity of most MOTUs, the total fish MOTU (γ) diversity is mainly due to inter-regions β-diversity (~74.5%) followed by inter-sites (14%) and inter-stations (7%) β-diversity (Fig. 3b). This result reflects a predominant role of large-scale bioregional differentiation which explains the exceptional fish diversity on coral reefs, probably associated with long-term geological isolation[43]. For example, the Caribbean region has a very distinct MOTU composition compared to the four other regions with only 1.5% of MOTUs being shared between the Caribbean and any other region while 21% of MOTUs are shared between at least two Indo-Pacific regions. The isolation of the Caribbean region can be explained by hard vicariant barriers (continents) from the other regions, and a limited suitable area for coral reefs during the past quaternary glaciation, while the Indo-Pacific maintained extensive coral reef refuges that have served as centres of survival during ice periods[20].

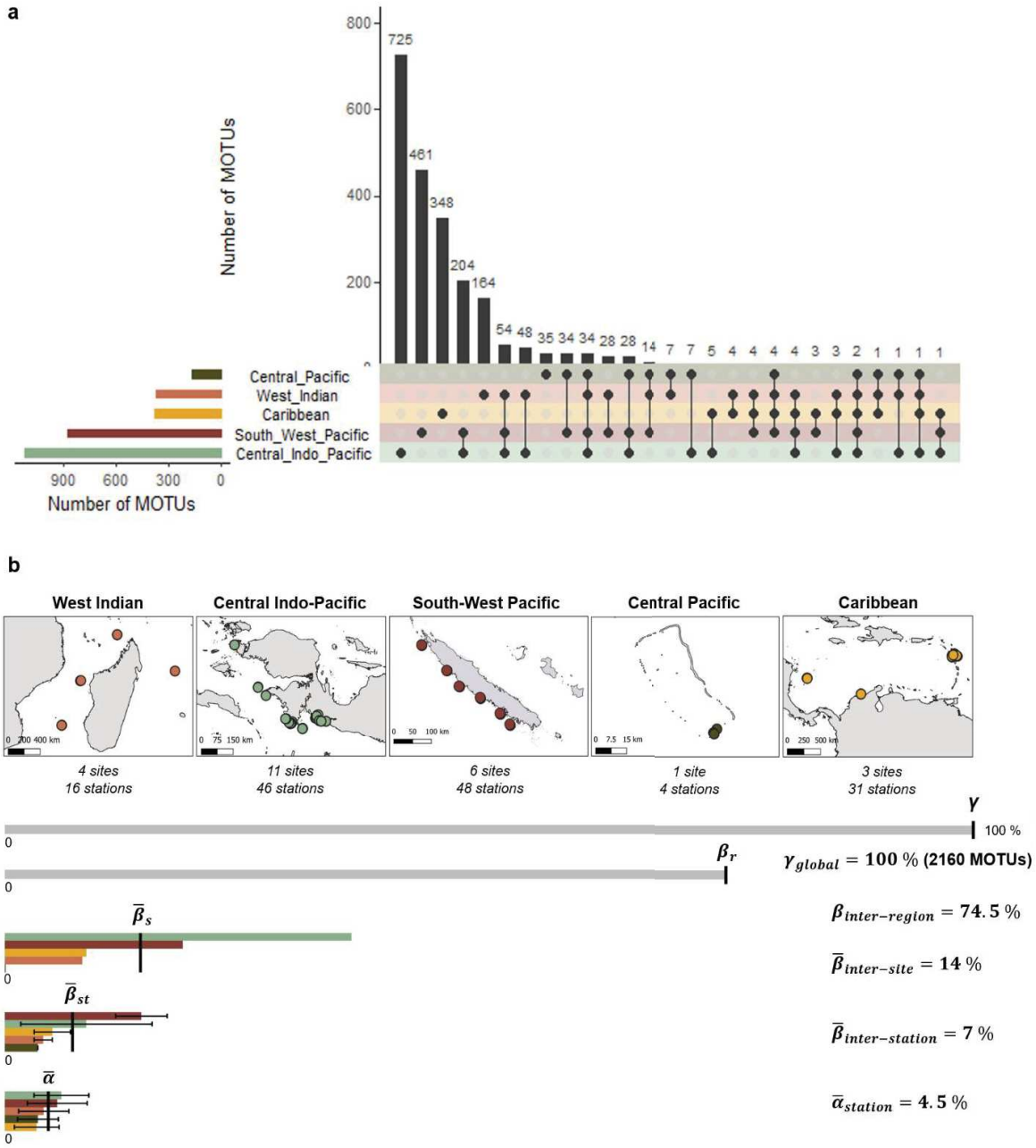


Figure 3 | Hierarchical partitioning of MOTU occurrences across spatial scales. a, Number of MOTUs found in only one region, or shared between 2, 3, 4 or all 5 regions. b, Global fish diversity (γ_{global}) is partitioned into $\beta_{inter\text{-}region}$ + mean $\beta_{inter\text{-}site}$ + mean $\beta_{inter\text{-}station}$ + mean $\bar{\alpha}_{station}$. Mean values at global scales are indicated with the black vertical segments. For $\beta_{inter\text{-}site}$, $\beta_{inter\text{-}station}$ and $\bar{\alpha}_{station}$, mean values are given for each region (colored bars) with the standard errors. $\beta_{inter\text{-}region}$ contributes the highest to gamma global (74.5%).

Spatial patterns of dissimilarity based on eDNA data generally match those obtained with visual census data, except for the inter-station β-diversity, where eDNA reveals higher values (Table 1, Extended Data Fig. 8c-d). This result implies that visual census tends to detect similar subsets of species among multiple transects within a single reef site, while eDNA reveals greater differences in composition among stations within sites. This local compositional dissimilarity of reef fishes among adjacent stations is greater than previously found and may correspond to local environmental or habitat differences, but also stochastic or random processes[44,45] or potential differences in the spatial signal of eDNA. This marked local dissimilarity promotes regional diversity and is critical to coral reef resilience at the site level in the context of ever-increasing disturbances[46]. If the positive influence of local species diversity on coral reef functioning is well-demonstrated[47], that of local β-diversity remains to be investigated, but could be more important than anticipated from patterns based on classical non-destructive survey methods.

Table 1 | Partitioning of different subsets across spatial scales. Partitioning for MOTUs assigned to cryptobenthic families, pelagic families and to species level.

	γ_{global}	$\beta_{inter-region}$	$\overline{\beta}_{inter-site}$	$\overline{\beta}_{inter-station}$	$\overline{\alpha}_{station}$
Cryptobenthic MOTUs	275	77%	14.7%	5.5%	2.8%
Pelagic MOTUs	171	74.3%	13.5%	7.7%	4.5%
eDNA MOTU	388	69.5%	14.1%	9.4%	7%
Visual census Species	1786	88.2%	8.3%	0.9%	2.6%

Beyond the hierarchical partitioning of diversity, we compared the distribution of fish MOTUs and species visual occurrences independently of the survey method and effort using global species abundance distributions (gSAD)[48]. gSADs provide a way to test several theories of community assembly rules at large scale. For example, the unified neutral theory of biogeography (UNTB)[49], stipulating that each individual's prospects of death and reproduction are equivalent whatever the species it belongs to within a given trophic group, would produce a gSAD converging towards a log-series distribution at large scale but also a power or Pareto distribution with a slope β = -150. These two distributions can be combined into a generalized Pareto distribution with exponential finite adjustment where the slope β is allowed to vary (see Methods). We fitted the fish MOTU and species visual occurrences to these three distributions (log-series, Pareto and Pareto with exponential finite adjustment (i.e. Pareto Bended) and estimated the parameters by maximum likelihood. For the visual census gSAD, the best fit was obtained with the Pareto and Pareto Bended distributions (Extended Data Table 3) with a slope significantly higher than -1 in both cases (Fig. 4). This suggests fewer rare species than under the neutral theory

which is rejected as in previous tests based on species abundances on coral reefs[51]. By contrast, the best fit was obtained with the log-series and the Pareto Bended distributions for fish MOTUs, the latter with a slope β = -1.03 (confidence interval at 95% [-1.12 ; -0.94]), indicating the same prevalence of rarity as under the UNTB. Therefore, MOTU gSAD suggests that neutral processes within trophic groups, underpinned by per capita equivalence among species and individuals, cannot be excluded as potential legitimate explanations for the assembly of fishes on the world's coral reefs.

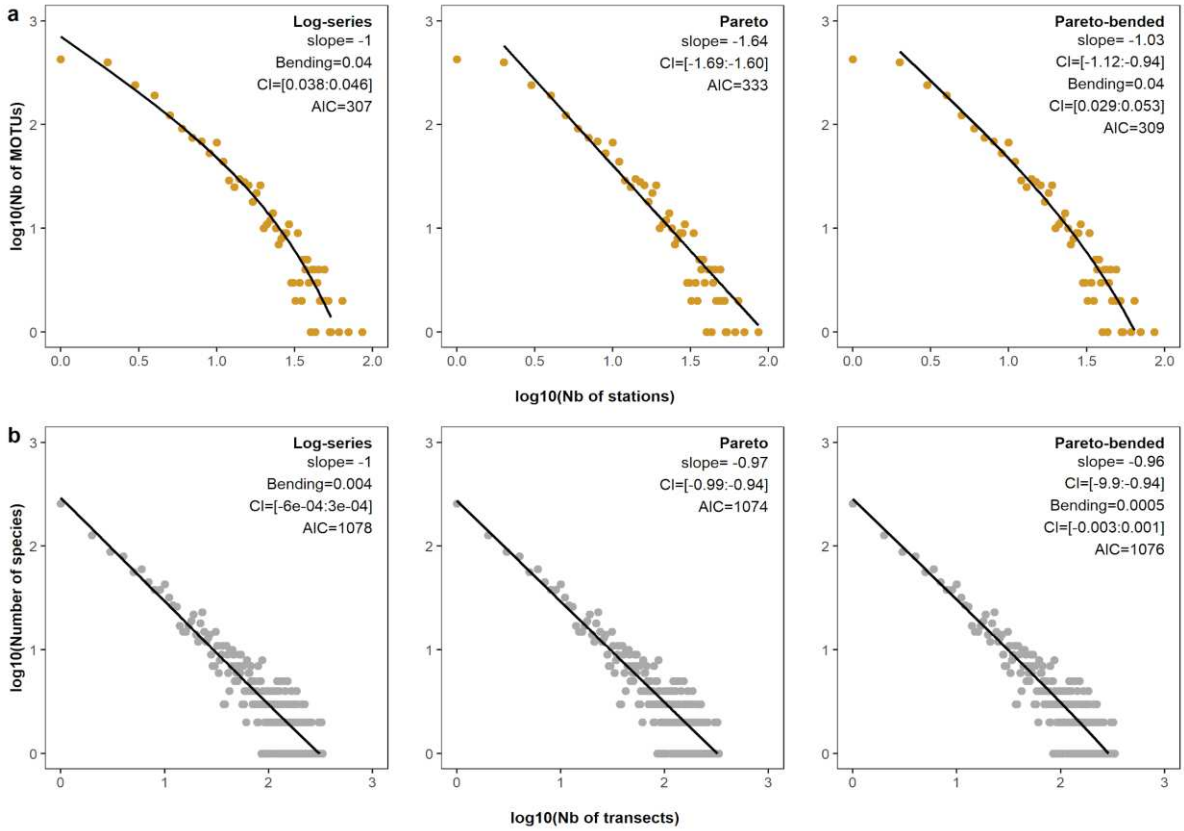


Figure 4 | The distribution of the total number of global observations per fish species. a, Distribution of MOTU occurrences across stations, log-transformed (yellow points). b, Distribution of visual census occurrences across transects (black points), log-transformed. For both distributions, three abundance distribution models were fitted: Log-series (left), Pareto (middle) and Pareto-bended (with exponential finite adjustment) (right). Slope, confidence interval of the slope (CI) and AIC of the models are given.

Conclusion and biogeography implication

Over a timespan of 28 months, eDNA sampling in major tropical ocean basins has allowed validation of classic biogeographic patterns and theories of species and family diversity in coral reef fishes, which previously required decades of previous surveys to uncover. This first circumglobal test of marine eDNA confirms the important future role that eDNA can play in mapping biodiversity at short time scales over large spatial scales. This includes establishing biodiversity monitoring for coral reefs on time-scales that are more compatible with the increasing speed and magnitude of global change, and the potential to better understand and account for seasonal variability[52]. Our eDNA approach has also revealed diversity beyond the current state of taxonomic knowledge of coral reef fishes. If census survey observations are limited to shallow coral reef habitats, eDNA allows the study of greater depths including mesophotic reef ecosystems[33]. By capturing DNA signals as MOTUs, rather than being constrained to separating taxa based on taxonomic descriptions and visual distinctions, the approach may further unveil new aspects of ecology and evolution. Our study generalizes our ability to use eDNA not only for ubiquitous microorganisms with widespread, abundant cells in seawater[8], but also for more elusive and less abundant vertebrates, opening application toward the monitoring of the entire food webs. Obtaining occurrence data on more taxa covering all levels of food webs will allow ecosystem modelers to build more complete networks of interactions among trophic levels and to move from an incomplete node-MOTU based biodiversity description to a node-link description, enabling the reconstruction of networks of ecological interactions[53].

Our study provides strong impetus for the development of a global marine observatory based on massive water sampling for eDNA, as it can easily be sampled by non-experts[54]. Improving knowledge of the effects of threats to the marine environment depends on temporal and global monitoring of biodiversity using low cost data collection methods, ideally with wide scientific consortia and public engagement in research[55], as now offered by eDNA technology[54,56].

Data availability

The data necessary to reproduce figures and results in this study will be publicly archived following best-practice guidelines, and can be available to editors and reviewers at the time of submission upon request.

Code availability

All code used to conduct the study will be available in a GitHub repository if accepted and is available to editors and reviewers at the time of submission upon request.

Main references

1. Bork, P. et al. Tara Oceans studies plankton at Planetary scale. Science. 348, 873–875 (2015).
2. Cordier, T. et al. Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. Mol. Ecol. May, 1–22 (2020).
3. Zellweger, F., De Frenne, P., Lenoir, J., Rocchini, D. & Coomes, D. Advances in Microclimate Ecology Arising from Remote Sensing. Trends Ecol. Evol. 34, 327–341 (2019).
4. Bohan, D. A. et al. Next-Generation Global Biomonitoring: Large-scale, Automated Reconstruction of Ecological Networks. Trends Ecol. Evol. 32, 477–487 (2017).
5. Taberlet, P., Coissac, E., Hajibabaei, M. & Rieseberg, L. H. Environmental DNA. Mol. Ecol. 21, 1789–1793 (2012).
6. Beng, K. C. & Corlett, R. T. Applications of environmental DNA (eDNA) in ecology and conservation: opportunities, challenges and prospects. Biodivers. Conserv. 29, 2089–2121 (2020).
7. Harrison, J. B., Sunday, J. M. & Rogers, S. M. Predicting the fate of eDNA in the environment and implications for studying biodiversity. Proc. R. Soc. B Biol. Sci. 286, 1–9 (2019).
8. De Vargas, C. et al. Eukaryotic plankton diversity in the sunlit ocean. Science (80-.). 348, 1–11 (2015).
9. Makiola, A. et al. Key Questions for Next-Generation Biomonitoring. Front. Environ. Sci. 7, 1–14 (2020).
10. Lenoir, J., Bertrand, R., Comte, L. & … L. B. Species better track climate warming in the oceans than on land. Nat. Ecol. Evol. 4, 1044–1059 (2020).
11. Palomares, M.-L. D. & Pauly, D. Chapter 32 - Coastal Fisheries: The Past, Present, and Possible Futures. Coasts and Estuaries (2019).
12. Besson, M. et al. development and survival via thyroid disruption. Nat. Commun. 11, 1–5 (2020).
13. Young, H. S., Mccauley, D. J., Galetti, M. & Dirzo, R. Patterns, Causes, and Consequences of Anthropocene Defaunation. Annu. Rev. Ecol. Evol. Syst. 47, 333–358 (2016).
14. Albouy, C. et al. Global vulnerability of marine mammals to global warming. Sci. Rep. 10, 1–13 (2020).
15. Cinner, J. E. et al. Meeting fisheries, ecosystem function, and biodiversity goals in a human-dominated world. Science. 368, 307–311 (2020).
16. Victor, B. C. How many coral reef fish species are there? Cryptic diversity and the new molecular taxonomy. in Ecology of fishes on coral reefs 76–88 (Cambridge University Press, Cambridge, 2015).
17. Siqueira, A. C., Morais, R. A., Bellwood, D. R. & Cowman, P. F. Trophic innovations fuel reef fish diversification. Nat. Commun. 11, 1–11 (2020).
18. Bellwood, D. & Wainwright, P. The history and biogeography of fishes on coral reefs. in Coral Reef Fishes: Dynamics and Diversity in a Complex Ecosystem (ed. Sale, P.) (2002).
19. Bellwood, D. R. & Hughes, T. P. Regional-scale assembly rules and biodiversity of coral reefs. Science. 292, 1532–1534 (2001).
20. Pellissier, L. et al. Quaternary coral reef refugia preserved fish diversity. Science. 344, 1016–1020 (2014).
21. VERON, J. E. N. et al. Delineating the Coral Triangle. Galaxea, J. Coral Reef Stud. 11, 91–100 (2009).
22. Bender, M. G. et al. Isolation drives taxonomic and functional nestedness in tropical reef fish faunas. Ecography. 40, 425–435 (2017).
23. Hughes, T. P., Bellwood, D. R., Connolly, S. R., Cornell, H. V. & Karlson, R. H. Double jeopardy and global extinction risk in corals and reef fishes. Curr. Biol. 24, 2946–2951 (2014).
24. Mouillot, D., Villéger, S., Parravicini, V., Kulbicki, M. & Arias-gonzález, J. E. Functional over-redundancy and high functional vulnerability in global fish faunas on tropical reefs. PNAS 111, 13757–13762 (2014).
25. Bellwood, D. R., Goatley, C. H. R. & Bellwood, O. The evolution of fishes and corals on reefs: Form, function and interdependence. Biol. Rev. 92, 878–901 (2017).
26. Brandl, S. J., Goatley, C. H. R., Bellwood, D. R. & Tornabene, L. The hidden half : ecology and evolution of cryptobenthic fishes on coral reefs. Biol. Rev. 93, 1846–1873 (2018).

27. Alzate, A., Zapata, F. A. & Giraldo, A. A comparison of visual and collection-based methods for assessing community structure of coral reef fishes in the Tropical Eastern Pacific. Rev. Biol. Trop. 62, 359–371 (2014).
28. Bellwood, D., Renema, W. & Rosen, B. Biodiversity hotspots, evolution and coral reef biogeography: a review. in Biotic Evolution and Environmental Change in Southeast Asia. 216–245 (Cambridge University Press, 2012).
29. Mora, C. Large-scale patterns and processes in reef fish richness. in Ecology of fishes on coral reefs (Cambridge University Press, Cambridge, 2015).
30. Marques, V. et al. Blind assessment of vertebrate taxonomic diversity across spatial scales by clustering environmental DNA metabarcoding sequences. Ecography. 43, 1–12 (2020).
31. Edgar, G. J. & Stuart-Smith, R. D. Systematic global assessment of reef fish communities by the Reef Life Survey program. Sci. Data 1, 1–8 (2014).
32. Edgar, G. J. et al. Abundance and local-scale processes contribute to multi-phyla gradients in global marine diversity. Sci. Adv. 3, 1–12 (2017).
33. Juhel, J. B. et al. Accumulation curves of environmental DNA sequences predict coastal fish diversity in the coral triangle. Proceedings. Biol. Sci. 287, 1–10 (2020).
34. Castellanos-Galindo, G. A., Krumme, U., Rubio, E. A. & Saint-Paul, U. Spatial variability of mangrove fish assemblage composition in the tropical eastern Pacific Ocean. Rev. Fish Biol. Fish. 23, 69–86 (2013).
35. Willis, T. J. & Anderson, M. J. Structure of cryptic reef fish assemblages: Relationships with habitat characteristics and predator density. Mar. Ecol. Prog. Ser. 257, 209–221 (2003).
36. Boussarie, G. et al. Environmental DNA illuminates the dark diversity of sharks. Sci. Adv. 4, 1–8 (2018).
37. Collins, R. A. et al. Persistence of environmental DNA in marine systems. Commun. Biol. 1, 1–12 (2018).
38. Sambrook, K. et al. Beyond the reef: The widespread use of non-reef habitats by coral reef fishes. Fish Fish. 20, 903–920 (2019).
39. Kimmerling, N. et al. Quantitative species-level ecology of reef fish larvae via metabarcoding. Nat. Ecol. Evol. 2, 306–316 (2018).
40. Beckley, L. E. et al. Structuring of larval fish assemblages along a coastal-oceanic gradient in the macro-tidal, tropical Eastern Indian Ocean. Deep. Res. Part II 161, 105–119 (2019).
41. Morais, R. A. & Bellwood, D. R. Pelagic Subsidies Underpin Fish Productivity on a Degraded Coral Reef. Curr. Biol. 29, 1521–1527 (2019).
42. Crist, T. O. & Veech, J. A. Additive partitioning of rarefaction curves and species-area relationships: Unifying α-, β- and γ-diversity with sample size and habitat area. Ecol. Lett. 9, 923–932 (2006).
43. Cowman, P. F. & Bellwood, D. R. The historical biogeography of coral reef fishes: Global patterns of origination and dispersal. J. Biogeogr. 40, 209–224 (2013).
44. Ahmadia, G. N., Tornabene, L., Smith, D. J. & Pezold, F. L. The relative importance of regional, local, and evolutionary factors structuring cryptobenthic coral-reef assemblages. Coral Reefs 37, 279–293 (2018).
45. MacNeil, M. A. et al. Hierarchical drivers of reef-fish metacommunity structure. Ecology 90, 252–264 (2009).
46. Wang, S. & Loreau, M. Biodiversity and ecosystem stability across scales in metacommunities. Ecol. Lett. 19, 510–518 (2016).
47. Benkwitt, C., Wilson, S. & Graham, N. A. J. Biodiversity increases ecosystem functions despite multiple stressors on coral reefs. Nat. Ecol. Evol. 4, 916–926 (2020).
48. Enquist, B. J. et al. The commonness of rarity: Global and future distribution of rarity across land plants. Sci. Adv. 5, 1–14 (2019).
49. Hubbell, S. P. The unified neutral theory of biodiversity and biogeography. vol. 32 (2001).
50. Pueyo, S. Diversity: Between neutrality and structure. Oikos 112, 392–405 (2006).
51. Dornelas, M., Connolly, S. R. & Hughes, T. P. Coral reef diversity refutes the neutral theory of biodiversity. Nature 440, 80–82 (2006).

52. Djurhuus, A. et al. Environmental DNA reveals seasonal shifts and potential interactions in a marine community. Nat. Commun. 11, 1–9 (2020).
53. Albouy, C. et al. The marine fish food web is globally connected. Nat. Ecol. Evol. 3, 1153–1161 (2019).
54. Biggs, J. et al. Using eDNA to develop a national citizen science-based monitoring programme for the great crested newt (Triturus cristatus). Biol. Conserv. 183, 19–28 (2015).
55. Irwin, A. CITIZEN SCIENCE COMES OF AGE Efforts to engage the public in research are bigger and more diverse than ever. But how much more room is there to grow ? Nature 532, 480–482 (2018).
56. Larson, E. R. et al. From eDNA to citizen science: emerging tools for the early detection of invasive species. Front. Ecol. Environ. 18, 194–202 (2020).

Chapitre 6 - Discussion

1. Synthèse

1.1. La limitation des bases de référence génétiques

La complétude des bases de références génétiques est l'une des principales limitations dans l'utilisation du metabarcoding ADNe (Wangensteen et al. 2018, Weigand et al. 2019). Il est cependant difficile de mesurer l'ampleur du travail restant à accomplir pour atteindre un taux de remplissage satisfaisant ou complet pour les poissons osseux. Au cours de ces travaux de thèse, j'ai proposé une première évaluation spatialisée à l'échelle mondiale de la couverture taxonomique des bases de références génétiques pour 19 marqueurs moléculaires ciblant les poissons marins et d'eau douce (chapitre 2). Une application *online* interactive (R shiny) complète cette évaluation et facilite l'extraction des données par zone biogéographique, avec le potentiel d'être étendu à n'importe quel groupe taxonomique et n'importe quel marqueur.

L'analyse globale de la couverture taxonomique des poissons osseux des bases de référence génétiques révèle que les zones tropicales sont moins séquencées, surtout en milieu d'eau douce, alors que ce sont les zones où la diversité en espèces est la plus élevée avec les besoins en conservation les plus importants (Barlow et al. 2018). Ainsi, moins de 15% des ~32 000 espèces de poissons décrites sont séquencées et amplifiées par la plupart des paires d'amorces sur le gène 12S, reconnu comme le plus performant et approprié pour les études de metabarcoding des poissons (Collins et al. 2019, Zhang et al. 2020b). Le marqueur génétique MiFish fait toutefois exception grâce à un fort remplissage des bases de référence ces 5 dernières années, portant à 8375 espèces (28% du total, version 38) le nombre d'espèces séquencées sur cette portion du 12S (Miya et al. 2020). Une partie de ce remplissage ciblé n'apparait pas dans le chapitre 2 car la méthodologie d'extraction des données publiques nécessite la présence des amorces sur les séquences partagées. L'extension récente de la base de référence pour le marqueur MiFish effectuée par les équipes est déposée en ligne mais ne présente que la séquence d'intérêt, sans les amorces, et n'est donc pas détectable en utilisant une PCR virtuelle. Ce pourcentage reste toujours limité pour des études avec une ambition globale, et ne tient pas compte des nombreuses espèces de poissons osseux qui restent à découvrir et à décrire. Pour un taxon aussi charismatique que le Cœlacanthe, deux espèces sont présentement décrites mais une récente analyse génétique émet l'hypothèse qu'une troisième espèce pourrait exister, celle-ci aurait divergée de son plus proche parent il y a 13 millions d'années (Kadarusman et al. 2020). Sur la période 2005 à 2014, c'est environ 400

nouvelles espèces de poissons osseux par an qui sont décrites (Nelson et al. 2016), et les musées regorgent de spécimens qu'il reste à examiner (Pinheiro et al. 2019). Ces arguments mettent en évidence l'important chemin qu'il reste à parcourir pour compléter les bases de références des poissons avant de pouvoir utiliser l'ADNe à large échelle sur ce groupe taxonomique, et la nécessité de disposer d'outils permettant de quantifier ces progrès.

L'ampleur du travail à réaliser pourrait décourager les écologues d'utiliser le metabarcoding ADNe tant que ces bases ne sont pas plus remplies et que les limites associées sont importantes. Je cite ici Jerde et al. (2019), qui déclare « […] *but if we wait for the methods to be further improved before deploying in areas of conservation concern with greater species richness, we will miss global opportunities to motivate the protection of rare species and prevent fishery collapses.* ». Mais si nous attendons que des méthodes innovantes comme l'ADNe soient parfaitement opérationnelles avant de les appliquer, nous risquons de manquer des occasions d'avancer notre compréhension des systèmes écologiques et de faire progresser les stratégies de conservation. Cette vision est en accord avec la récente décision du GBIF (Août 2020) d'implémenter des occurrences basées sur des séquences génétiques, qu'elles soient ou non agrémentées d'une assignation taxonomique précise à l'espèce (Andersson et al. 2020). Alors que la complétion totale (ou même quasi-totale) des bases de références poissons est relativement irréaliste à court terme et que les besoins de recensement de la biodiversité et du suivi des écosystèmes sont pressants (Díaz et al. 2019a), il est nécessaire d'utiliser des métriques de biodiversité fiables tout en intégrant le fait que les bases de références sont incomplètes. C'était surement l'enjeu principal de ma thèse.

On a montré dans le chapitre 3 qu'il est possible d'estimer le nombre d'espèces présentes à l'aide d'unités taxonomiques moléculaires (MOTUs), sans nécessairement parvenir à associer un nom d'espèce sur chaque MOTU, donc partiellement « en aveugle ». Cette avancée est cruciale car le chapitre 2 a montré de fortes disparités géographiques en termes de couverture taxonomique des bases de référence, ce qui biaise fortement les comparaisons du nombre d'espèces ou taxa identifiés entre régions. Une plus grande richesse en taxa identifiée sur un site par rapport à un autre pourrait aussi bien représenter une réalité biologique qu'être un artefact lié à une meilleure complétion de la base de référence sur un site. Plus du trois quarts des familles de poissons ont au moins une espèce séquencée avec le marqueur teleo, ce qui rend théoriquement détectables les espèces non séquencées à ce niveau taxonomique. De nombreuses familles sont extrêmement diversifiées en termes de nombre d'espèces, notamment dans les récifs coralliens (Brandl et al. 2018, Siqueira et al. 2020). Par exemple, la famille des *Gobiidae* représente près de 1 900 espèces décrites réparties dans plus de 200 genres (Froese and Pauly 2000). En l'absence de mesure quantitative en termes d'unités taxonomiques et sans base de

référence exhaustive, dénombrer les taxa (nombre de genres, familles, ou ordres) peut donc se révéler extrêmement réducteur (Annexe 1). On peut aisément trouver sur un même site plus de 20 espèces d'une même famille qui seraient comptabilisées comme un seul taxa par une approche sans unités taxonomiques et sans base exhaustive. Par une approche utilisant des unités taxonomiques, il est possible d'approximer le nombre d'espèces réellement présentes pour chaque genre ou famille sans la nécessité d'avoir des identifications à l'espèce, et ainsi obtenir de meilleures estimations de la diversité présente. Cette méthodologie, basée sur le clustering de séquences, ouvre ainsi la voie à l'application des méthodes de metabarcoding ADNe dans virtuellement toutes les régions du monde, pour proposer des estimateurs de diversité et de richesse peu biaisés par l'absence de complétude des bases de référence.

L'analyse des unités taxonomiques moléculaires est également particulièrement adaptée à l'étude des patrons de biodiversité. Il n'est pas toujours nécessaire de connaître l'identité de toutes les espèces d'un assemblage pour analyser la répartition de la diversité entre communautés. En particulier, le partitionnement de la diversité à différentes échelles spatiales (alpha, beta, gamma) est possible grâce à une approche en unités taxonomiques (MOTUs) comme démontré dans le chapitre 3. En utilisant toutes les données ADNe collectée sur 25 sites sur les récifs coralliens des 3 bassins océaniens tropicaux dans le chapitre 5, on a retrouvé les patrons de diversité des poissons marins attendus à l'échelle biogéographique. On observe un gradient de diversité longitudinal avec une richesse qui décroît avec la distance de part et d'autre du triangle de corail (zone géographique située entre la Malaisie, les Philippines, les Iles Salomon et l'Indonésie). L'analyse de la diversité beta à différentes échelles spatiale (échantillon, récif, site, région, global) met en évidence que la diversité globale en MOTUs est principalement influencée par la diversité beta inter-régionale, ce qui reflète le rôle prédominant des processus régionaux dans la diversification et le maintien de la diversité en espèces. La quantification d'unités taxonomiques moléculaires permet d'accéder à des mesures de biodiversité fiables en s'affranchissant de la nécessité d'assigner la totalité des séquences au niveau de l'espèce.

1.2. Détecter les changements de la biodiversité plus efficacement

Un des arguments les plus fréquemment employés pour justifier l'emploi de méthodes basées sur l'ADNe repose sur son aptitude à pouvoir détecter efficacement les changements de biodiversité qui s'accélèrent, notamment en milieu marin (McLean et al. 2018, Lenoir et al. 2020). Pour que cet argument soit valable, il faut que les inventaires par méthodes moléculaires présentent une plus-value par rapport aux méthodes de recensements conventionnelles : de meilleurs résultats, une complémentarité, ou une performance égale contrebalancée par une facilité opérationnelle. La plupart

des études convergent pour indiquer que les méthodes ADNe sont au moins identiques, complémentaires ou supérieures aux méthodes conventionnelles dans leur capacité à détecter l'ensemble de la communauté sous sa facette taxonomique (McColl-Gausden et al. 2020). Une récente méta-analyse de toutes les études comparatives entre ADNe metabarcoding et méthodes traditionnelles renforce ce postulat (McElroy et al. 2020). Les tendances sont très claires en milieu d'eau douce pour lesquels 24 études ont examiné plus de 100 sites au total pour comparer les méthodes. Le milieu marin bénéficie de beaucoup moins d'intérêt dans la littérature avec seulement 14 études examinant 17 sites dans des études comparatives. Pour les deux milieux, les études comparatives dans les zones tropicales très riches (> 100 espèces) sont extrêmement rares, avec seulement 3 sites par système. D'après ces quelques exemples, on observe que les capacités de détectabilité de l'ADNe en metabarcoding diminuent au-delà d'une certaine richesse de la communauté (Fig. 1). Ce désavantage semble lié aux manquements dans les bases de références, qui sont moins complètes dans les régions très riches en diversité, plus qu'à une incapacité intrinsèque de la méthode à fonctionner dans les milieux hyper-diverses (Jerde et al. 2019). Alors que les besoins en suivi de communautés sont extrêmement importants dans la zone tropicale, aucun consensus n'a encore émergé sur l'utilité de la méthode ADNe dans ces écosystèmes en raison d'un trop faible nombre d'études.

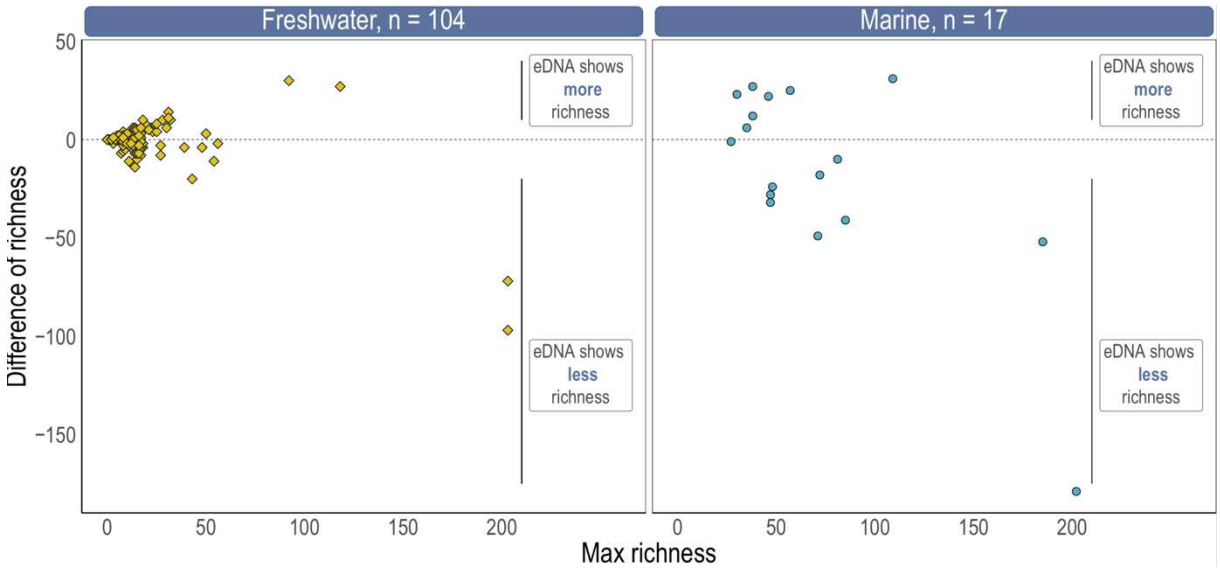


Fig. 1. Comparaison de la performance de l'ADNe metabarcoding avec des méthodes traditionnelles d'après les données d'une méta-analyse de McElroy et al (2020). Au-dessus de 0, l'ADNe a enregistré plus de richesse spécifique, alors qu'en dessous de la ligne 0, les méthodes traditionnelles ont enregistré plus de richesse. La richesse est ici considérée comme le nombre d'espèces identifiées, et les méthodes traditionnelles recouvrent une forte disparité de méthodes et peuvent être composées d'inventaires historiques ou bien de mise en œuvre concomitante sur une zone.

Nous avons montré que le metabarcoding ADNe est extrêmement prometteur pour les inventaires taxonomiques marins dans les sites récifaux et tropicaux (chapitre 4). Notre étude comparative utilise des caméras longue durée (~25h de vidéo au total) en parallèle d'un échantillonnage ADNe afin de capturer la biodiversité présente au même moment pour s'affranchir de possibles biais saisonniers ou même journaliers (Sales et al. 2021). Nous avons montré que l'ADNe détecte moins d'espèces identifiées formellement que les caméras (33 contre 50), mais que le nombre d'unités taxonomiques moléculaires est estimé à 85, ce qui surpasse largement l'inventaire effectué avec les caméras. De plus, l'avantage de l'ADNe par rapport aux vidéos augmente avec le niveau taxonomique. Ainsi l'ADNe détecte moins d'espèces identifiées par rapport aux vidéos mais plus de genres et de familles. Ce constat suggère que la faiblesse de l'ADNe au niveau des espèces réside principalement dans le manque de séquences référencées. A une plus large échelle, on a également montré que le metabarcoding ADNe permet de détecter énormément de diversité taxonomique avec peu d'effort d'échantillonnage par rapport aux méthodes classiques (chapitre 5). Avec 251 échantillons ADNe (en 28 mois), on a détecté 2160 MOTUs alors que 2813 transects (en 13 ans) ont été nécessaires pour détecter 1887 espèces en utilisant les recensements en plongée.

Alors que les changements climatiques agissent à une échelle temporelle extrêmement rapide (Lenoir et al. 2020), il est absolument nécessaire de disposer de méthodes permettant de recenser les communautés rapidement, efficacement et sans destruction. On montre ainsi que l'ADNe est capable de répondre à ce défi, avec un effort d'échantillonnage limité et une couverture taxonomique, notamment en genres et familles, qui est très large. Par rapport aux transects en plongée, le metabarcoding ADNe détecte également de plus fines variations de diversité taxonomique aux échelles locales (chapitre 5). Cela signifie que l'ADNe détecte un signal plus localisé d'espèces couramment ratées en plongée. Ce type de biais engendre une vision des récifs comme étant plus homogènes qu'ils ne le seraient en réalité. Ces résultats, associés à d'autres éléments de littérature trop récents pour avoir été intégrés à la meta-analyse citée précédemment (Nguyen et al. 2019, Valdivia-Carrillo et al. 2019), suggèrent que l'ADNe est une méthode fiable pour détecter la biodiversité taxonomique présente, plus efficace en termes d'effort d'échantillonnage que les méthodes classiques et dont la principale limitation est la couverture taxonomique en séquences. Dans le cadre d'études de suivi de biodiversité, la méthode d'ADNe metabarcoding est une option pertinente même dans les milieux marins hyper-riches, mais nécessite la mise en place d'une base de référence de séquences locales en amont pour être exploitée et interprétée à son plein potentiel.

Il est de plus en reconnu que l'approche taxonomique n'est pas suffisante pour détecter efficacement des changements de structure dans les communautés (Devictor et al. 2010, McLean et al.

2018). Les diversités fonctionnelles et phylogénétiques complètent couramment la facette taxonomique de la diversité, en prenant en compte les fonctions effectuées par les espèces dans l'écosystème et l'histoire évolutive qu'elles représentent. Alors que les méthodes ADNe estiment correctement la facette taxonomique de la diversité, la capacité d'évaluation des autres facettes demeure incertaine. Très peu d'études comparent le potentiel ou les biais des méthodes ADNe sur plus d'une facette de diversité, et aucune n'a comparé les trois facettes de façon simultanée. Au-delà de la diversité taxonomique, le **chapitre 4** a montré que l'ADNe détecte également une diversité fonctionnelle et phylogénétique plus étendue qu'une approche traditionnelle par vidéo. Cet avantage s'accompagne d'une efficacité plus importante en termes d'effort d'échantillonnage, puisqu'un transect ADNe (2 filtres) détecte autant de diversité phylogénétique que 25 heures de vidéo. L'ADNe détecte plus efficacement des espèces appartenant aux groupes fonctionnels pélagiques par rapport aux vidéos, mais aucune méthode n'exclut totalement un groupe fonctionnel particulier : gros prédateurs, poissons de récifs ainsi que crypto-benthiques sont tous détectés avec les deux méthodes. Toutefois, l'île de Malpelo présente un contexte particulier et agrège une faune diverse sur une petite aire géographique, surestimant probablement la probabilité de détection d'espèces beaucoup plus difficiles à filmer sur des récifs plus classiques, comme les gros poissons pélagiques ou les requins. Les résultats d'autres études s'accordent avec ce constat selon lequel l'ADNe détecte mieux les taxons pélagiques, qui sont couramment manqués par les méthodes visuelles (Valdivia-Carrillo et al. 2019). L'ADNe est donc capable d'identifier des espèces présentant une large gamme de traits fonctionnels à l'échelle d'un récif, y compris les espèces crypto-benthiques vivant sur le fond depuis une filtration en surface, et de recenser les communautés plus exhaustivement. L'ADNe metabarcoding est donc une méthode adaptée au recensement rapide de la diversité d'une zone permettant de couvrir toutes les facettes de la diversité d'une communauté.

1.3. L'ADNe pour traquer les espèces menacées ou non-indigènes

Au-delà des estimations de diversité, l'ADNe est particulièrement adapté à la détection des espèces d'intérêt, souvent rares, notamment celles inscrites sur la liste rouge de l'IUCN ou les espèces non-indigènes susceptibles de perturber un écosystème. Alors qu'un MOTU assigné au genre ou à la famille peut être suffisant pour étudier la diversité, le suivi d'espèces nécessite une identification taxonomique précise et ne peut pas bénéficier de ce type d'approche à plus haut niveau taxonomique. Le **chapitre 2** a mis en évidence que même si les espèces menacées sont proportionnellement plus représentées que les espèces non menacées dans les bases de données, de nombreuses restent non détectables car non séquencées. L'application GAPeDNA permet de visualiser pour chaque marqueur

de façon interactive la liste des espèces séquencées par statut IUCN et par région biogéographique. Ce constat est un point de départ indispensable et permet de cibler les efforts de séquençage dans une région donnée où un suivi des espèces menacées est envisagé.

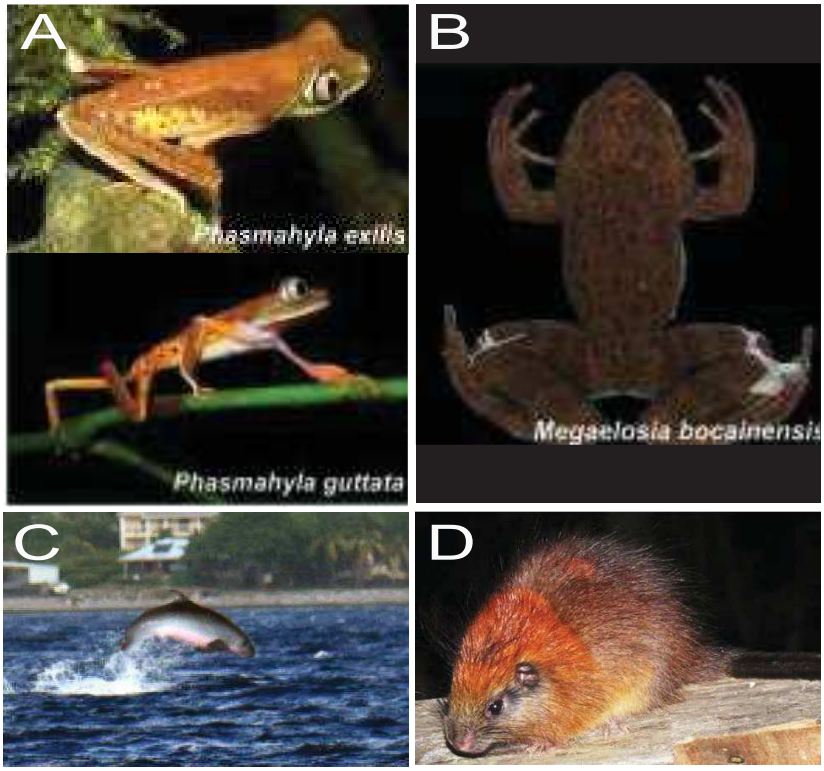


Fig. 2. Photos d'espèces ayant été détectées en ADNe alors qu'elles sont élusives ou étaient considérées disparues localement avec des amphibiens *Phasmahyla exilis* et *P. guttata* (A), *Megaelosia bocainensis* (B), le mammifère marin *Kogia sima* (C) et le rat à crête rousse *Santamartamys rufodorsalis* (D). *Crédits : Lopes et al. (2020) (A, B), Camille Albouy (C) et l'association Proaves (D).*

Les exemples issus de la littérature et de mes travaux démontrent un fort potentiel de détection des espèces rares par l'ADNe, qui pourrait bénéficier aux gestionnaires d'écosystèmes et aux organismes travaillant en conservation. Grâce à la constitution d'une importante base de référence génétique locale au Brésil (138 espèces d'amphibiens), une équipe a démontré la présence de deux espèces (*Phasmahyla exilis* et *Phasmahyla guttata*) considérées comme disparues localement et d'une espèce qui n'avait pas été détectée depuis 1968 (*Megaelosia bocainensis*) (Fig. 2. A, B) (Lopes et al. 2020). Mes travaux de thèse ont également permis de détecter des espèces menacées ou extrêmement élusives : trois occurrences d'un mammifère marin très furtif : le cachalot pygmé (*Kogia sima,* Fig. 2 C) sur l'île de Malpelo (Annexe 2) qui fréquente préférentiellement les eaux plus profondes (50-100m) et n'a été vu que 4 fois en 20 ans sur des transects ciblant les mammifères marins dans tout le Pacifique

Est colombien (> 22 000 km parcourus) (Palacios et al. 2012). Il est de plus quasiment impossible de discerner *K. sima* de son con-spécifique *K. breviceps* par des méthodes visuelles sur le terrain, alors que leur barcode génétique permet de les différencier. Jusqu'à très récemment, l'espèce était inscrite comme « Data Deficient » par l'IUCN (Kiszka and Braulik 2020). L'exemple le plus remarquable issue de mes travaux est la détection de *Santamartamys rufodorsalis*, le rat arboricole à crête rousse, une espèce endémique de la région de Santa Marta (Colombie) qui est inscrite sur la liste des 100 espèces les plus menacées au monde (Velazco et al. 2017) (Annexe 3, Fig. 2.D). Cette espèce n'a été aperçue que 3 fois dans l'histoire, et a été redécouverte en 2011, lorsque des scientifiques l'ont photographié par hasard furtivement dans la région, la dernière détection remontant à 1913 (Velazco et al. 2017). L'espèce a été détectée sur un filtre provenant d'un échantillon de la rivière Don Diego, et indique la présence de l'espèce au niveau du cours d'eau ou en amont de quelques kilomètres due au transport du matériel génétique avec le courant (Deiner et al. 2016, Pont et al. 2018).

L'ADNe permet de détecter des espèces considérées comme rare, qui sont parfois restées invisibles pendant des décennies. Un taux de détectabilité très faible (i.e. probabilité de détecter l'espèce alors qu'elle est présente) mène à de fréquents faux négatifs et peut concerner deux scénarios : (i) l'espèce est tellement rare qu'il est presque impossible de la détecter par des méthodes moins sensibles, ou (ii) l'espèce est présente mais les méthodes sont peu efficaces ou non adaptées pour la détecter (par exemple : espèces craintives de l'humain en mer qui fuient sa présence). Dans le cadre de protocoles de conservation de la biodiversité, il est important de discerner ces deux cas afin de proposer la mesure de gestion la plus adaptée. Une espèce réellement absente ne peut recoloniser un écosystème que par dispersion naturelle ou translocation par l'activité humaine. Des espèces rares peuvent avoir subi une diminution de leur population si importante qu'elles sont maintenant éteintes fonctionnellement, toutefois l'indication de leur présence même résiduelle est fondamentale. Dans ce cas et contrairement à une extinction locale complète, il est possible de proposer une gestion de l'espace pour permettre à l'espèce de rétablir des effectifs plus importants. Ces mesures peuvent passer par exemple par l'interdiction des activités de pêche dans une zone donnée ou l'arrêt de projets de modification du territoire si l'espèce occupe un habitat destiné à être détruit. Les bénéfices d'une méthode bien plus sensibles pour la détection des espèces rares sont donc très importants pour la compréhension de la structure des écosystèmes, mais également pour appuyer la conservation des espaces naturels et accompagner des propositions de gestion adaptées.

2. Limitations

2.1. Estimation de la biodiversité par MOTUs

Estimer la richesse des communautés à partir de substituts moléculaires est une alternative nécessaire lorsque les bases de références sont peu renseignées comme c'est le cas en milieu tropical. Toutefois, les substituts moléculaires sont à considérer comme des approximations de la richesse réelle. Il est hautement improbable qu'une méthode parvienne un jour à estimer parfaitement des richesses à partir d'unités moléculaires sans erreurs en utilisant les procédés moléculaires classiques de metabarcoding aujourd'hui (PCR et séquençage Illumina). Les approximations liées à la combinaison des sous-estimations et surestimations selon les groupes taxonomiques sont inévitables. Alors que les études en microbiologie sont pionnières dans ce domaine, elles restent confrontées à ce type de problèmes malgré des années d'expérience et de recul (Reeder and Knight 2009, Callahan et al. 2016). La combinaison d'erreurs stochastiques de PCR ainsi que d'erreurs de séquençages sont difficiles à détecter et se confondent avec des occurrences de *vraies* espèces rares. Typiquement, une vraie espèce rare dont la séquence est proche d'une autre espèce également présente mais plus abondante sera identifiée comme une erreur. Ajuster les seuils afin de conserver ces deux espèces est possible, mais susceptible d'entraîner une importante surestimation de la richesse car les erreurs sont nombreuses et ne seraient plus écartées. A l'inverse, une espèce rare mais très distincte des autres espèces abondantes a une plus haute probabilité d'être retenue. La structure des communautés en termes d'occurrence, d'abondance et de phylogénie joue donc un rôle dans la qualité de l'estimation de richesse par les unités taxonomiques.

L'approche proposée dans le chapitre 2 n'a été testée que sur un seul marqueur moléculaire et la fiabilité des estimations n'est pas garantie avec d'autres marqueurs, certains ajustements de seuils peuvent s'avérer nécessaires. Pour les marqueurs présentant une importante variabilité intra-spécifique, il est possible que cette approche surestime la richesse réelle. Pour le cas des espèces très rares, cette approche risque de sous-estimer leur diversité car un des seuils requiert la présence d'un MOTU dans au moins deux réplicas PCRs sur un jeu de données. Cette étape est pour le moment indispensable, car de nombreuses erreurs de PCR ne se produisent qu'une fois et ce filtre permet de limiter drastiquement la quantité de faux-positifs. Notre pipeline bio-informatique a le potentiel d'être amélioré en combinant d'autres algorithmes ou filtres qualités afin de limiter la quantité de faux positifs tout en augmentant la part de vrais positifs. Brandt et al. (2020) ont proposé une approche très similaire pour d'autres taxons (métazoaires des écosystèmes profonds), avec notamment la combinaison de l'algorithme de nettoyage DADA2, de clustering puis de nettoyage post-classification, pour tendre vers

un ratio 1:1 entre les estimations de MOTUs et la réelle diversité en espèces présentes. Enfin des approches de mesure de diversité réduisant le poids des unités les plus rares en occurrence ou bien en abondance telles que la diversité de Hill ont déjà été appliquées au metabarcoding de l'ADNe et des fécès pour la caractérisation des régimes alimentaires (Alberdi et al. 2018, Mächler et al. 2020). L'utilisation de ces indices pondérés permet la comparaison entre sites même si une partie des MOTUs ne représentent pas une réalité biologique. La méthode est toutefois sensible à la valeur de q, paramètre qui contrôle le poids donné à la communauté la plus rare et ne permet pas une réelle comparaison avec des méthodes traditionnelles basées sur des relevés taxonomiques. L'estimation de diversité avec des MOTUs fonctionne avec une base incomplète, mais nécessite tout de même un remplissage partiel permettant d'assigner les MOTUs au moins au niveau de la famille pour certaines analyses écologiques. Certaines familles n'ont actuellement aucun représentant séquencé et ne sont donc pas identifiables, d'autres ont si peu d'espèces séquencées que les algorithmes ne parviendraient pas à les assigner à la bonne famille. Si le remplissage complet des bases est un objectif à plus long terme, il est possible de cibler la collecte d'échantillons de référence afin de maximiser la couverture de certaines familles dont la détection est actuellement très aléatoire ou impossible.

2.2. Détection d'espèces à fort intérêt en gestion ou conservation

Pour valider la détection d'une espèce en particulier, sa séquence doit être présente dans une base de données mais le barcode génétique doit également être unique à cette espèce. Or la résolution taxonomique des barcodes ADNe est un réel problème pour les inventaires biologiques, car de nombreuses espèces partagent leur séquence avec un conspécifique. C'est notamment le cas de groupes taxonomiques ayant subi une diversification récente, dont les séquences mitochondriales n'ont pas encore assez divergé pour présenter de la variabilité génétique sur les barcodes ADNe. Gold et al. (2020) ont remarqué que la quasi-totalité les espèces du genre *Sebastes* de la Californie partageaient la même séquence sur le marqueur 12S MiFish. Alors que ces espèces ont une importance commerciale, il est impossible de les distinguer génétiquement entre elles avec les marqueurs ADNe classiques et universels. Doble et al. (2020) ont aussi été confrontés à ce problème dans le lac Tanganyika (Tanzanie), où de nombreuses espèces de poissons cichlides partagent la même séquence sur le 12S due à une récente radiation évolutive, ce qui biaisait fortement les estimations de richesse avec des marqueurs universels.

Dans le cas où une seule espèce d'intérêt partage sa séquence avec une autre, il est possible de valider sa présence sur les échantillons concernés à l'aide d'approches plus spécifiques, telles que

les qPCR, ddPCR ou CRISPR-Cas9 (Baker et al. 2018, Williams et al. 2019, Postaire et al. 2020). Ces approches sont généralement moins chères et plus faciles à mettre en place qu'une étude en metabarcoding. Dans le cas où c'est un clade complet qui manque de résolution taxonomique, il est nécessaire de cibler un barcode alternatif et de concevoir une paire d'amorces spécifiques de ce groupe taxonomique (genre, famille). Il est alors nécessaire doubler le nombre de PCR et la quantité de séquençages afin d'avoir les résultats de plusieurs marqueurs en metabarcoding. Toutefois, il n'y a pas de solution parfaite car un barcode ADNe semble plus efficace s'il est plus court (Bylemans et al. 2018a) due à la dégradation rapide des fragments long dans le milieu, mais les barcodes courts ont généralement moins de résolution taxonomique. D'après deux études *in-silico*, le marqueur 12S AcMDB07 (Bylemans et al. 2018b) semble avoir le meilleur compromis entre longueur (~300 pb) et résolution taxonomique, mais requiert d'avantages d'études *in-vivo* afin de confirmer sa performance en milieu réel (Zhang et al. 2020b). Les méthodes traditionnelles ne sont toutefois pas exemptes de ce type de limitation, bien qu'elles soient moins fréquentes et moins abordées. Il existe de nombreux cas d'espèces qui ont été classées à tort sous un même nom d'espèce alors qu'il s'agit en réalité de complexe d'espèces cryptiques, c'est à dire pas facilement discernables par des critères morphologiques (Kon et al. 2007, Melo et al. 2016).

Une autre limitation réside dans le fait que l'ADNe ne permet pas de récupérer d'informations biologiques sur les espèces dont la détection est validée. Par exemple, il n'est pas possible de connaître le stade ontogénique (larve, juvénile, adulte), le sexe ni la taille des espèces détectées alors que ces caractéristiques sont essentielles en dynamique des populations et en conservation. Concernant les estimations de biomasse, il n'y a pas de consensus en metabarcoding sur la manière d'inférer des notions d'abondance par rapport à la quantité de lectures de séquences. Le procédé de PCR est extrêmement stochastique et les amorces peuvent présenter des affinités différentes entre espèces ce qui biaise le nombre de copies d'ADN réalisées à cette étape (Elbrecht and Leese 2015). Certaines applications semblent prometteuses et montrent une corrélation entre biomasse et nombre de lectures alors que d'autres ne trouvent pas de résultats convaincants (Lacoursière-Roussel et al. 2016, Thomsen et al. 2016, Stoeckle et al. 2017). La méthode de PCR quantitative, qui est espèce-spécifique contrairement au metabarcoding, semble plus adaptée à l'estimation de la biomasse d'une espèce (Mauvisseau et al. 2017). Pour certains aspects, il est donc nécessaire de coupler les approches ADNe à des méthodes plus conventionnelles. La complémentarité des méthodes peut ainsi apporter une meilleure estimation de la structure des communautés au-delà de l'inventaire, particulièrement au niveau biologique ou comportemental.

3. Perspectives

La complétion des bases de référence génétique est un travail à long terme, qui risque d'être confronté à la difficulté de récupérer des individus appartement à des espèces rares ou peu accessibles lié à l'isolement géographique de leur habitat ou bien dans des régions du monde instables politiquement. Comment récupérer l'ADN d'espèces qui n'ont été aperçues qu'une fraction de fois dans l'histoire, vivant seulement dans un pays en guerre ou au fond des océans où les campagnes océanographies sont rares ? Les spécimens présents dans les musées ont une valeur scientifique considérable, mais leur âge et conditions de conservation entravent fortement la capacité des protocoles de laboratoire à extraire et amplifier cet ADN. Les poissons sont notamment fréquemment conservés dans du formol, ce qui complique les procédures moléculaires. Une amélioration future des méthodes moléculaires liées à l'ADN ancien ou des méthodes alternatives telles que la capture par hybridation dérivée de sondes RAD (hyRAD-X , « *hybridization capture from RAD-derived probes obtained from a reduced eXome template* ») sont prometteuses pour parvenir à compléter la base pour des espèces rares à partir des échantillons de musées (Schmid et al. 2017). Cette méthode nécessiterait toutefois l'ADN d'espèces con-spécifiques pour concevoir une sonde suffisamment spécifique pour permettre l'hybridation, mais pourrait se révéler importante pour récupérer l'ADN d'espèces particulièrement difficiles à trouver ou à amplifier avec les méthodes traditionnelles selon l'état de conservation des spécimens.

La détection d'une espèce dans un échantillon ne permet pas de savoir si un individu adulte est présent, si l'ADN a été transporté par les mouvements d'eau d'une rivière, par la défécation du prédateur l'ayant ingéré ou s'il s'agit d'un flux larvaire non résident. Une méthode prometteuse a récemment été proposé pour permettre de distinguer la détection d'individus ou de larves et gamètes. En utilisant le ratio d'abondance entre ADN nucléaire et ADN mitochondrial, il serait possible de savoir si les traces d'ADN proviennent d'une ponte ou d'une agrégation d'adultes (Bylemans et al. 2017). Cette découverte permettrait d'élargir la gamme d'application de l'ADNe à la détection d'aire et de période de reproduction des espèces.

La collecte de données ADNe à large échelle initiée lors de cette thèse se poursuit avec de nouveaux partenariats. Depuis Décembre 2019, de nouvelles campagnes ont été effectuées en Méditerranées dans les profondeurs mésophotiques (50-120m), sur des monts sous-marins de l'Indien et du Pacifique, sur la Péninsule Antarctique avec Greenpeace mais également sur l'archipel de Svalbard en Arctique. Ces nouveaux échantillons étendent drastiquement la couverture géographique initiale qui

devient quasiment globale ainsi que les types d'écosystèmes, en particulier grâce aux échantillons des pôles, où la campagne Antarctique a eu lieu au moment du record chaleur historique sur la péninsule (20°C). En complément du chapitre 5 réalisé uniquement avec les bases de références génétiques publiques, une ré-analyse de tous les filtres de pôle à pôle est prévue avec l'ajout de plus de 1000 espèces récupérées et séquencées au cours de ces 3 dernières années. On y cherchera spécifiquement des signes de présence d'espèces d'intérêt telles que les espèces menacées localement ou globalement, ou encore d'espèces jamais enregistrées dans la zone afin de déterminer leur statut : espèces potentiellement non-indigènes transportées via les eaux de ballast, ou migrations liées aux changements climatiques. Ce large échantillonnage permettra de mieux comprendre, à travers la modélisation statistique, les effets conjugués des conditions environnementales et socio-économiques connues pour influencer les peuplement ichtyologiques, tels que les antécédents en termes de blanchissement corallien, mise en place d'aires protégées ou de réserves marines (Hughes et al. 2017, Cinner et al. 2020).

Dans ces travaux de thèse, seul le compartiment des vertébrés a été exploré avec les marqueurs moléculaires spécifiques à ce groupe. L'ADNe a un fort potentiel pour étudier l'ensemble des niveaux trophiques afin de caractériser les communautés de manière plus intégrative, en utilisant une méthode standardisée. Permettant également un suivi temporel, cette approche a un potentiel extrêmement important pour mieux comprendre les dynamiques temporelles et les interactions entre organismes tout le long de la chaine trophique (Djuurhus 2020).

Les avantages du metabarcoding de l'ADNe en font une méthode particulièrement indiquée pour être utilisée dans le cadre de la mise en place d'observatoire de la biodiversité à large échelle et multi-trophique (Sutherland 2015). Il est ainsi possible de caractériser la totalité des composants biologiques d'un écosystème de manière non-destructive, facile et rapide, sans nécessiter d'expertise pour la récolte de données sur le terrain. L'inclusion de méthodes ADNe dans des programmes de sciences participatives a déjà montré son fort potentiel à l'échelle d'un pays pour un programme de suivi d'une espèce menacée (Royaume-Uni) (Biggs et al. 2015), ou dans la lutte contre l'établissement d'espèces non-indigènes (Larson et al. 2020). A l'échelle du continent Américain, une initiative tente de mettre en place un réseau d'observation de la biodiversité en mer à travers de multiples approches telles que les satellites, les mesures océanographiques et intègre déjà la technologie ADNe pour faciliter la récolte et la standardisation des données de biodiversité (Canonico et al. 2019). Instaurer des consignes de bonnes pratiques et de standardisation de récolte et de traitement des données est toutefois fondamental pour imaginer l'intégration de la méthode à des réseaux de suivi de la biodiversité à large échelle. En Europe, le collectif DNAquanet (https://dnaqua.net/) rassemble une

large partie des personnels de recherche et de l'industrie en lien avec l'ADNe, afin de décider de standards à appliquer dans la discipline pour garantir la qualité des résultats obtenus (Leese et al. 2016). Plus récemment en Amérique du Nord, le collectif PISCeS s'est formé dans le même but de rassembler les connaissances existantes et d'élaborer des standards (Loeza-Quintana et al. 2020). Le potentiel de la méthode ADNe pour intégrer ou créer des réseaux d'observations de la biodiversité est immense, compte tenu de ses performances sur de nombreux écosystèmes et sa facilité de déploiement même par des personnels non spécialistes, ce qui pourrait permettre d'étendre drastiquement les capacités de suivi des communautés à large échelle en ces périodes de bouleversements rapides.

Les méthodes d'échantillonnage évoluent rapidement pour être utilisées dans un maximum de systèmes différents le plus simplement possible. Un nouveau prototype de pompe péristaltique immergeable sans limite de profondeur et légère a été mis au point lors d'une collaboration entre SPYGEN, MARBEC et la société ANDROMEDE (Fig. 3. A). Cette pompe a déjà été utilisée pour filtrer en surface sans risque de noyer le matériel et en plongée autonome à des profondeurs mésophotiques (120m). Une intervention humaine est pour le moment nécessaire pour enclencher le début de filtration de la pompe, ce qui limite le potentiel d'automatisation et de prélèvement à distance avec ce modèle. Les développements futurs incluent toutefois une programmation possible du déclenchement, sous forme de programmateur temporel ou à une certaine profondeur grâce à des capteurs de pression. Une équipe américaine a aussi développé un robot sous-marin capable de coupler la réalisation de vidéos sous-marines et le prélèvement d'échantillons ADNe dans des profondeurs mésophotiques (Fig. 3. B). Pour aller plus loin dans les perspectives d'automatisation de la récolte de données, le couplage de drones sous-marins ou de planeurs sous-marins avec des dispositifs de prélèvements d'ADNe aurait un fort potentiel pour faciliter un échantillonnage marin avec moins de logistique (Fig 3. C, D). Les planeurs sous-marins (« *gliders* ») sont des engins prévus pour effectuer de longs séjours en mer, équipés de nombreux capteurs permettant la récolte à haute fréquence de nombreux paramètres océanographiques. Ces engins sont toujours à l'état de prototype, mais permettent de réduire les coûts des embarquements en mer via leur autonomie et capacité à se déplacer. Les développements technologiques des méthodes d'exploration des océans pourraient ainsi permettre d'augmenter l'automatisation des relevés d'échantillons et la mise en place d'un véritable réseau de surveillance de la biodiversité même sur des zones éloignées et peu visitées par les instituts de recherche ou le public.

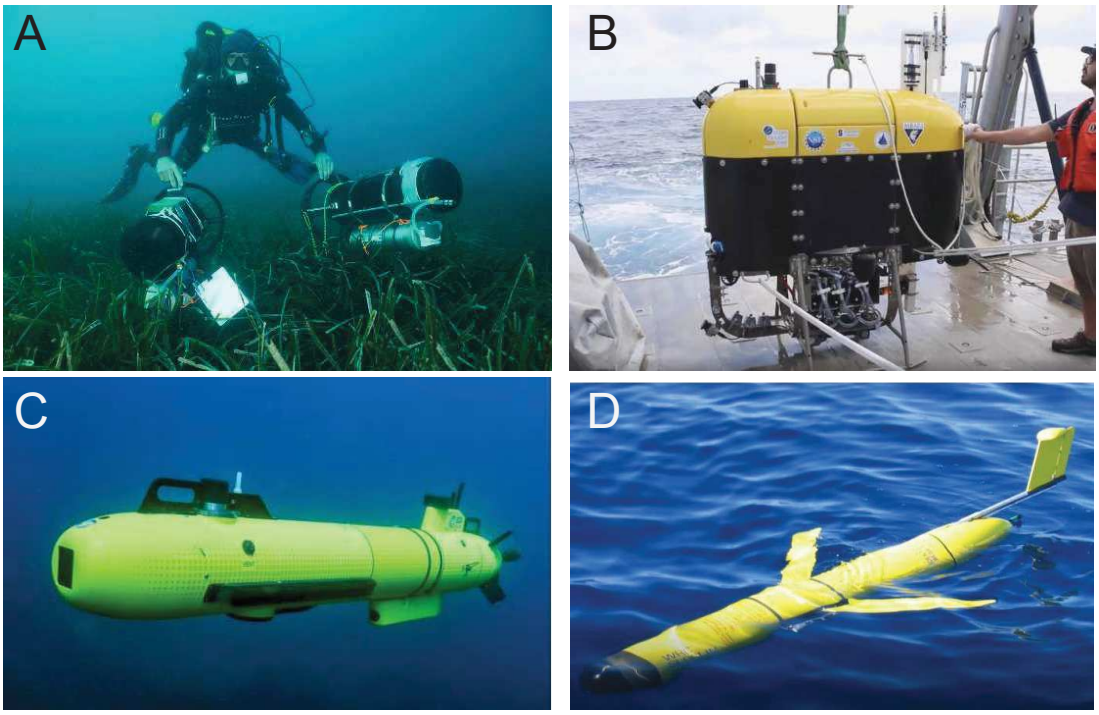


Fig. 3. Photos de dispositifs d'échantillonnage d'ADNe, avec (A) le prototype de la pompe péristaltique immergeable utilisée dans les études les plus récentes, ici monté sur un scooter sous-marin, (B) un robot sous-marin développé par une équipe américaine du WHOI (Woods Hole Oceanographic Institution) pour associer vidéo sous-marine et prélèvement ADNe dans les profondeurs mésophotiques, (C) un drone sous-marin, (D) un planeur sous-marin océanographique sur lequel pourrait être couplé un dispositif de prélèvement d'ADNe. *Crédits : (A) Andromède, (B) Allison Albritton, (C) ECA GROUP, (D) Ifremer.*

La collecte d'échantillons ADNe dès aujourd'hui est fondamentale malgré les limites posées par les bases de références ou la résolution taxonomique des marqueurs car les filtres ADNe agissent comme une sauvegarde historique, une *time capsule* permettant de remonter dans le temps pour y chercher de nouvelles espèces avec d'autres protocoles ou pour figer un moment particulier. Par exemple nous avons filtré pendant le confinement du printemps 2020 dans toute la Méditerranée française. Ces prélèvements ADNe constituent de véritables témoins d'un événement extrême (la mer sans présence humaine : https://www.umontpellier.fr/articles/la-mer-sans-les-hommes) qui a le potentiel de fournir un nouvel état de référence pour nos indicateurs de l'état des écosystèmes côtiers. D'abord, les séquences retrouvées qui sont présentement non identifiées pourraient l'être dans un avenir proche, et il est aisé de relancer une analyse bio-informatique sur une base de référence comprenant plus de séquences. Ensuite, il est possible de faire plusieurs extractions d'ADN à partir d'un seul filtre ADNe, où les portions non utilisées initialement peuvent être archivées à la manière d'une bio-banque (Jarman et al. 2018). Cette précaution permettrait une ré-analyse des jeux de données plus

anciens avec de potentielles nouvelles technologies moins limitantes que celles disponibles actuellement. Il est ici important de rappeler que l'avènement du séquençage à haut-débit, qui a révolutionné une partie du monde de la génétique, ne date que des années 2005 avec l'avènement des plateformes Roche 454 puis Illumina. Il est dont impossible d'anticiper avec certitude le développement de nouvelles technologies d'ici les prochaines décennies. Dans un contexte de changements climatiques et de dégradation des écosystèmes toujours croissant, l'accès à un suivi temporel de l'évolution des communautés est important pour mieux comprendre et anticiper les bouleversements à venir afin de produire une réponse de gestion adaptée.

Références

Achtman, M. and Wagner, M. 2008. Microbial diversity and the genetic nature of microbial species. - Nat. Rev. Microbiol. 6: 431–440.

Alberdi, A. et al. 2018. Scrutinizing key steps for reliable metabarcoding of environmental samples. - Methods Ecol. Evol. 9: 134–147.

Albouy, C. et al. 2015. Projected impacts of climate warming on the functional and phylogenetic components of coastal Mediterranean fish biodiversity. - Ecography 38: 681–689.

Allard, L. et al. 2014. Electrofishing efficiency in low conductivity neotropical streams: Towards a non-destructive fish sampling method. - Fish. Manag. Ecol. 21: 234–243.

Allard, L. et al. 2016. Effect of reduced impact logging and small-scale mining disturbances on Neotropical stream fish assemblages. - Aquat. Sci. 78: 315–325.

Allison, E. H. et al. 2009. Vulnerability of national economies to the impacts of climate change on fisheries. - Fish Fish. 10: 173–196.

Amano, T. and Sutherland, W. J. 2013. Four barriers to the global understanding of biodiversity conservation: Wealth, language, geographical location and security. - Proc. R. Soc. B Biol. Sci. 280: 20122649.

Amir, A. et al. 2017. Deblur Rapidly Resolves Single-. - Am. Soc. Microbiol. 2: 1–7.

Andersson, A. F. et al. 2020. Publishing sequence-derived data through biodiversity data platforms [Community review draft]. - Copenhagen GBIF Secr. in press.

Antão, L. H. et al. 2020. Temperature-related biodiversity change across temperate marine and terrestrial systems. - Nat. Ecol. Evol. 4: 927–933.

Apothéloz-Perret-Gentil, L. et al. 2017. Taxonomy-free molecular diatom index for high-throughput eDNA biomonitoring. - Mol. Ecol. Resour. 17: 1231–1242.

Appeltans, W. et al. 2012. The magnitude of global marine species diversity. - Curr. Biol. 22: 2189–2202.

Araújo, F. G. et al. 2009. Longitudinal patterns of fish assemblages in a large tropical river in southeastern Brazil: Evaluating environmental influences and some concepts in river ecology. - Hydrobiologia 618: 89–107.

Atwood, T. B. et al. 2020. Herbivores at the highest risk of extinction among mammals, birds, and reptiles. - Sci. Adv. 6: eabb8458.

Bain, M. et al. 1985. A Quantitative Method for Sampling Riverine Microhabitats by Electrofishing. - North Am. J. Fish. Manag. 5: 489–493.

Baker, C. S. et al. 2018. Environmental DNA (eDNA) from the wake of the whales: Droplet digital PCR for detection and species identification. - Front. Mar. Sci. 5: 1–11.

Bakker, J. et al. 2017. Environmental DNA reveals tropical shark diversity in contrasting levels of anthropogenic impact. - Sci. Rep. 7: 1–11.

Bar-On, Y. M. et al. 2018. The biomass distribution on Earth. - Proc. Natl. Acad. Sci. U. S. A. 115: 6506–6511.

Barlow, J. et al. 2018. The future of hyperdiverse tropical ecosystems. - Nature 559: 517–526.

Barnosky, A. D. et al. 2011. Has the Earth's sixth mass extinction already arrived? - Nature 471: 51–57.

Baselga, A. 2010. Partitioning the turnover and nestedness components of beta diversity. - Glob. Ecol. Biogeogr. 19: 134–143.

Bellemain, E. et al. 2016. Trails of river monsters: Detecting critically endangered Mekong giant catfish Pangasianodon gigas using environmental DNA. - Glob. Ecol. Conserv. 7: 148–156.

Bellwood, D. R. and Hughes, T. P. 2001. Regional-Scale Assembly Rules and Biodiversity of Coral Reefs. - Science 292: 1532–1535.

Bellwood, D. R. et al. 2003. Limited functional redundancy in high diversity systems: Resilience and ecosystem function on coral reefs. - Ecol. Lett. 6: 281–285.

Bellwood, D. R. et al. 2019. The meaning of the term 'function' in ecology: A coral reef perspective. - Funct. Ecol. 33: 948–961.

Bessey, C. et al. 2020. Maximizing fish detection with eDNA metabarcoding. - Environ. DNA: 1–12.
Biggs, J. et al. 2015. Using eDNA to develop a national citizen science-based monitoring programme for the great crested newt (Triturus cristatus). - Biol. Conserv. 183: 19–28.
Bland, L. M. et al. 2015. Predicting the conservation status of data-deficient species. - Conserv. Biol. 29: 250–259.
Blowes, S. A. et al. 2019. The geography of biodiversity change in marine and terrestrial assemblages. - Science 366: 339–345.
Boakes, E. H. et al. 2010. Distorted views of biodiversity: Spatial and temporal bias in species occurrence data. - PLoS Biol. 8: e1000385.
Bolam, F. et al. 2020. How many bird and mammal extinctions has recent conservation action prevented?: 1–11.
Bongaerts, P. et al. 2019. Mesophotic.org: a repository for scientific information on mesophotic ecosystems. - Database (Oxford). 2019: baz140.
Booth, D. J. et al. 2011. Detecting range shifts among Australian fishes in response to climate change. - Mar. Freshw. Res. 62: 1027–1042.
Bosch, N. E. et al. 2017. “How” and “what” matters: Sampling method affects biodiversity estimates of reef fishes. - Ecol. Evol. 7: 4891–4906.
Boussarie, G. et al. 2018. Environmental DNA illuminates the dark diversity of sharks. - Sci. Adv. 4: eaap9661.
Boyer, F. et al. 2016. OBITOOLS: a UNIX-inspired software package for DNA metabarcoding. - Mol. Ecol. Resour. 16: 176–182.
Brandl, S. J. et al. 2018. The hidden half: ecology and evolution of cryptobenthic fishes on coral reefs. - Biol. Rev. 93: 1846–1873.
Brandl, S. J. et al. 2019. Demographic dynamics of the smallest marine vertebrates fuel coral reef ecosystem functioning. - Science 364: 1189–1192.
Bylemans, J. et al. 2017. An environmental DNA-based method for monitoring spawning activity: a case study, using the endangered Macquarie perch (Macquaria australasica). - Methods Ecol. Evol. 8: 646–655.
Bylemans, J. et al. 2018a. Does Size Matter? An Experimental Evaluation of the Relative Abundance and Decay Rates of Aquatic Environmental DNA. - Environ. Sci. Technol. 52: 6408–6416.
Bylemans, J. et al. 2018b. Monitoring riverine fish communities through eDNA metabarcoding: determining optimal sampling strategies along an altitudinal and biodiversity gradient. - Metabarcoding and Metagenomics 2: 1–12.
Callahan, B. J. et al. 2016. DADA2: High resolution sample inference from Illumina amplicon data. - Nat Methods 13: 581–583.
Callahan, B. J. et al. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. - ISME J. 11: 2639–2643.
Canonico, G. et al. 2019. Global observational needs and resources for marine biodiversity. - Front. Mar. Sci. 6: 367.
Ceballos, G. et al. 2015. Accelerated modern human–induced species losses: Entering the sixth mass extinction. - Sci. Adv. 1: e1400253.
Ceballos, G. et al. 2020. Vertebrates on the brink as indicators of biological annihilation and the sixth mass extinction. - Proc. Natl. Acad. Sci. U. S. A. 117: 13596–13602.
Cheung, W. W. L. et al. 2010. Large-scale redistribution of maximum fisheries catch potential in the global ocean under climate change. - Glob. Chang. Biol. 16: 24–35.
Cilleros, K. et al. 2019. Unlocking biodiversity and conservation studies in high-diversity environments using environmental DNA (eDNA): A test with Guianese freshwater fishes. - Mol. Ecol. Resour. 19: 27–46.
Cinner, J. E. et al. 2020. Meeting fisheries, ecosystem function, and biodiversity goals in a human-dominated world. - Science 368: 307–311.
Cisneros-Montemayor, A. M. et al. 2016. A global estimate of seafood consumption by coastal indigenous peoples. - PLoS One 11: 1–16.

Collen, B. et al. 2014. Global patterns of freshwater species diversity, threat and endemism. - Glob. Ecol. Biogeogr. 23: 40–51.
Collins, R. A. et al. 2019. Non-specific amplification compromises environmental DNA metabarcoding with COI. - Methods Ecol. Evol. 10: 1985–2001.
Cordier, T. et al. 2019. Multi-marker eDNA metabarcoding survey to assess the environmental impact of three offshore gas platforms in the North Adriatic Sea (Italy). - Mar. Environ. Res. 146: 24–34.
Costello, M. J. et al. 2013. Can we name earth's species before they go extinct? - Science 339: 413–416.
Courchamp, F. et al. 2006. Rarity value and species extinction: The anthropogenic allee effect. - PLoS Biol. 4: 2405–2410.
D'Agata, S. et al. 2014. Human-mediated loss of phylogenetic and functional diversity in coral reef fishes. - Curr. Biol. 24: 555–560.
Dahlke, F. T. et al. 2018. Northern cod species face spawning habitat losses if global warming exceeds 1.5°C. - Sci. Adv. 4: 1–11.
Darling, E. S. et al. 2010. Combined effects of two stressors on Kenyan coral reefs are additive or antagonistic, not synergistic. - Conserv. Lett. 3: 122–130.
Davidson, N. C. 2014. How much wetland has the world lost? Long-term and recent trends in global wetland area. - Mar. Freshw. Res. 65: 934–941.
de Vargas, C. et al. 2015. Eukaryotic plankton diversity in the sunlit ocean. - Science 348: 1261605–1261605.
Dehling, D. M. et al. 2016. Morphology predicts species' functional roles and their degree of specialization in plant–Frugivore interactions. - Proc. R. Soc. B Biol. Sci. 283: 20152444.
Deiner, K. et al. 2016. Environmental DNA reveals that rivers are conveyer belts of biodiversity information. - Nat. Commun. 7: 12544.
Devictor, V. et al. 2010. Spatial mismatch and congruence between taxonomic, phylogenetic and functional diversity: The need for integrative conservation strategies in a changing world. - Ecol. Lett. 13: 1030–1040.
Díaz, S. et al. 2019a. Pervasive human-driven decline of life on Earth points to the need for transformative change. - Science 366: eaax3100.
Díaz, S. et al. 2019b. Summary for policymakers of the global assessment report on biodiversity and ecosystem services.
Dirzo, R. et al. 2014. Defaunation in the Anthropocene. - Science 345: 401–406.
Doble, C. J. et al. 2020. Testing the performance of environmental DNA metabarcoding for surveying highly diverse tropical fish communities: A case study from Lake Tanganyika. - Environ. DNA 2: 24–41.
Doi, H. et al. 2017. Environmental DNA analysis for estimating the abundance and biomass of stream fish. - Freshw. Biol. 62: 30–39.
Donaldson, T. J. and Dulvy, N. K. 2004. Threatened fishes of the world: Bolbometopon muricatum (Valenciennes 1840) (Scaridae). - Environ. Biol. Fishes 70: 373.
Dornelas, M. et al. 2014. Assemblage time series reveal biodiversity change but not systematic loss. - Science 344: 296–299.
Dornelas, M. et al. 2018. BioTIME: A database of biodiversity time series for the Anthropocene. - Glob. Ecol. Biogeogr. 27: 760–786.
Duarte, C. M. et al. 2020. Rebuilding marine life. - Nature 580: 39–51.
Dulvy, N. K. and Polunin, N. V. C. 2004. Using informal knowledge to infer human-induced rarity of a conspicuous reef fish. - Anim. Conserv. 7: 365–374.
Dulvy, N. K. et al. 2003. Extinction vulnerability in marine populations. - Fish Fish. 4: 25–64.
Dulvy, N. K. et al. 2014. Extinction risk and conservation of the world's sharks and rays. - Elife 3: e00590.
Edgar, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. - Bioinformatics 26: 2460–2461.
Edgar, R. C. 2016. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. - bioRxiv: 81257.
Edgar, R. C. 2018. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. - Bioinformatics

34: 2371–2375.
Edgar, R. C. and Flyvbjerg, H. 2015. Error filtering, pair assembly and error correction for next-generation sequencing reads. - Bioinformatics 31: 3476–3482.
Elbrecht, V. and Leese, F. 2015. Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass-sequence relationships with an innovative metabarcoding protocol. - PLoS One 10: 1–16.
Faith, D. P. and Baker, A. M. 2006. Phylogenetic Diversity (PD) and Biodiversity Conservation: Some Bioinformatics Challenges. - Evol. Bioinforma. 2: 117693430600200.
FAO 2018. The State of World Fisheries and Aquaculture 2018 - Meeting the sustainable development goals.
Feeley, K. J. et al. 2013. Compositional shifts in costa rican forests due to climate-driven species migrations. - Glob. Chang. Biol. 19: 3472–3480.
Ficetola, G. F. et al. 2008. Species detection using environmental DNA from water samples. - Biol. Lett. 4: 423–425.
Fossheim, M. et al. 2015. Recent warming leads to a rapid borealization of fish communities in the Arctic. - Nat. Clim. Chang. 5: 673–677.
Frank, K. T. et al. 2018. Exploitation drives an ontogenetic-like deepening in marine fish. - Proc. Natl. Acad. Sci. U. S. A. 115: 6422–6427.
Froese, R. and Pauly, D. 2000. FishBase 2000: concepts, design and data sources (Editors, Ed.).
Frøslev, G. T. et al. 2017. Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. - Nat. Commun. 8: 1188.
Furlan, E. M. and Gleeson, D. 2017. Improving reliability in environmental DNA detection surveys through enhanced quality control. - Mar. Freshw. Res. 68: 388–395.
Garciá Molinos, J. et al. 2016. Climate velocity and the future global redistribution of marine biodiversity. - Nat. Clim. Chang. 6: 83–88.
Gerringer, M. E. et al. 2017. Pseudoliparis swirei sp. Nov.: A newly-discovered hadal snailfish (Scorpaeniformes: Liparidae) from the Mariana Trench. - Zootaxa 4358: 161–177.
Gold, Z. et al. 2020. FishCARD : Fish 12S California Current Specific Reference Database for Enhanced Metabarcoding Efforts. - Authorea: 1–14.
Grenié, M. et al. 2018. Functional rarity of coral reef fishes at the global scale: Hotspots and challenges for conservation. - Biol. Conserv. 226: 288–299.
Grosberg, R. K. et al. 2012. Biodiversity in water and on land. - Curr. Biol. 22: 900–903.
Haddad, N. M. et al. 2015. Habitat fragmentation and its lasting impact on Earth's ecosystems. - Sci. Adv. 1: 1–10.
Halpern, B. S. et al. 2015. Spatial and temporal changes in cumulative human impacts on the world's ocean. - Nat. Commun. 6: 1–7.
Harper, L. R. et al. 2019. Prospects and challenges of environmental DNA (eDNA) monitoring in freshwater ponds. - Hydrobiologia 826: 25–41.
Harrison, J. B. et al. 2019. Predicting the fate of eDNA in the environment and implications for studying biodiversity. - Proc. R. Soc. B Biol. Sci. 286: 20191409.
He, F. et al. 2018. Freshwater megafauna diversity: Patterns, status and threats. - Divers. Distrib. 24: 1395–1404.
Hicks, C. C. et al. 2019. Harnessing global fisheries to tackle micronutrient deficiencies. - Nature 574: 95–98.
Hinlo, R. et al. 2017. Environmental DNA monitoring and management of invasive fish : comparison of eDNA and fyke netting. - Manag. Biol. Invasions 8: 89–100.
Hoegh-Guldberg, O. and Bruno, J. 2010. The Impact of Climate Change on the World's Marine Ecosystems. - Science 328: 1523–1528.
Hoffmann, M. et al. 2010. The impact of conservation on the status of the world's vertebrates. - Science 330: 1503–1509.
Holsman, K. K. et al. 2020. Ecosystem-based fisheries management forestalls climate-driven collapse. - Nat. Commun. 11: 4579.

Hughes, T. P. et al. 2017. Coral reefs in the Anthropocene. - Nature 546: 82–90.
Huse, S. M. et al. 2010. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. - Environ. Microbiol. 12: 1889–1898.
Hutchings, J. et al. 2012. Life-history correlates of extinction risk and recovery potential. - Ecol. Appl. 22: 1061–1067.
Irigoien, X. et al. 2014. Large mesopelagic fishes biomass and trophic efficiency in the open ocean. - Nat. Commun. 5: 3271.
Jackson, J. B. C. et al. 2001. Historical overfishing and the recent collapse of coastal ecosystems. - Science 293: 629–637.
Jarman, S. N. et al. 2018. The value of environmental DNA biobanking for long-term biomonitoring. - Nat. Ecol. Evol. 2: 1192–1193.
Jerde, C. L. et al. 2019. Measuring global fish species richness with eDNA metabarcoding. - Mol. Ecol. Resour. 19: 19–22.
Kadarusman et al. 2020. A thirteen-million-year divergence between two lineages of Indonesian coelacanths. - Sci. Rep. 10: 1–9.
Keck, F. et al. 2017. Freshwater biomonitoring in the Information Age. - Front. Ecol. Environ. 15: 266–274.
Keck, F. et al. 2018. Boosting DNA metabarcoding for biomonitoring with phylogenetic estimation of operational taxonomic units' ecological profiles. - Mol. Ecol. Resour. 18: 1299–1309.
Kiszka, J. and Braulik, G. 2020. Kogia sima. - IUCN Red List Threat. Species 2020 e: T11048A50359330.
Kon, T. et al. 2007. DNA sequences identify numerous cryptic species of the vertebrate: A lesson from the gobioid fish Schindleria. - Mol. Phylogenet. Evol. 44: 53–62.
Kortz, A. R. and Magurran, A. E. 2019. Increases in local richness (a-diversity) following invasion are offset by biotic homogenization in a biodiversity hotspot. - Biol. Lett. 15: 20190133.
Koskinen, K. et al. 2015. Inconsistent Denoising and Clustering Algorithms for Amplicon Sequence Data. - J. Comput. Biol. 22: 743–751.
Köster, J. and Rahmann, S. 2012. Snakemake-a scalable bioinformatics workflow engine. - Bioinformatics 28: 2520–2522.
Kroodsma, D. A. et al. 2018. of Fisheries. - Science 908: 904–908.
Kunin, V. et al. 2010. Wrinkles in the rare biosphere: Pyrosequencing errors can lead to artificial inflation of diversity estimates. - Environ. Microbiol. 12: 118–123.
Kurtzer, G. M. et al. 2017. Singularity: Scientific containers for mobility of compute. - PLoS One 12: 1–20.
Lacoursière-Roussel, A. et al. 2016. Improving herpetological surveys in eastern North America using the environmental DNA method [1]. - Genome 59: 991–1007.
Lacoursière-Roussel, A. and Deiner, K. 2019. Environmental DNA is not the tool by itself. - J. Fish Biol.: 1–4.
Langlois, T. et al. 2020. A field and video-annotation guide for baited remote underwater stereo-video surveys of demersal fish assemblages. - Methods Ecol. Evol. 2020: 2041–210X.13470.
Laporte, M. et al. 2020. Caged fish experiment and hydrodynamic bidimensional modeling highlight the importance to consider 2D dispersion in fluvial environmental DNA studies. - Environ. DNA 2: 362–372.
Larsen, B. et al. 2017. INORDINATE FONDNESS MULTIPLIED AND REDISTRIBUTED: THE NUMBER OF SPECIES ON EARTH AND THE NEW PIE OF LIFE. - Q. Rev. Biol. 92: 298–301.
Larson, E. R. et al. 2020. From eDNA to citizen science: emerging tools for the early detection of invasive species. - Front. Ecol. Environ.: 1–9.
Lawler, J. J. et al. 2013. Projected climate-driven faunal movement routes. - Ecol. Lett. 16: 1014–1022.
Leclère, D. et al. 2020. Bending the curve of terrestrial biodiversity needs an integrated strategy. - Nature 585: 551–556.
Leese, F. et al. 2016. DNAqua-Net: Developing new genetic tools for bioassessment and monitoring of aquatic ecosystems in Europe. - Res. Ideas Outcomes 2: e11321.
Legendre, P. and Fortin, M.-J. 1989. Spatial pattern and ecological analysis Pierre. - Vegetatio 80: 107–

138.
Legendre, P. and Legendre, L. 1998. Numerical ecology (E Science, Ed.).
Leger, J. B. et al. 2015. Clustering methods differ in their ability to detect patterns in ecological networks. - Methods Ecol. Evol. 6: 474–481.
Lenoir, J. et al. 2020. Species better track climate warming in the oceans than on land. - Nat. Ecol. Evol. 4: 1044–1059.
Letessier, T. B. et al. 2019. Remote reefs and seamounts are the last refuges for marine predators across the Indo-Pacific. - PLoS Biol. 17: 1–20.
Loeza-Quintana, T. et al. 2020. Pathway to Increase Standards and Competency of eDNA Surveys (PISCeS)—Advancing collaboration and standardization efforts in the field of eDNA. - Environ. DNA 2: 255–260.
Lopes, C. M. et al. 2020. Lost and found: Frogs in a biodiversity hotspot rediscovered with environmental DNA. - Mol. Ecol.: 1–10.
Mächler, E. et al. 2020. Decision making and best practices for taxonomy-free eDNA metabarcoding in biomonitoring using Hill numbers. - bioRxiv: 2020.03.31.017723.
MacNeil, M. A. et al. 2020. Global status and conservation potential of reef sharks. - Nature 583: 801–806.
Magurran, A. E. et al. 2015. Rapid biotic homogenization of marine fish assemblages. - Nat. Commun. 6: 2–6.
Mahé, F. et al. 2015. Swarm v2: highly-scalable and high-resolution amplicon clustering. - PeerJ 3: e1420.
Mariani, S. et al. 2019. Sponges as natural environmental DNA samplers. - Curr. Biol. 29: R401–R402.
Mauvisseau, Q. et al. 2017. On the way for detecting and quantifying elusive species in the sea: The Octopus vulgaris case study. - Fish. Res. 191: 41–48.
Mazel, F. et al. 2018. Prioritizing phylogenetic diversity captures functional diversity unreliably. - Nat. Commun. 9: 2888.
McCauley, D. J. et al. 2015. Marine defaunation: Animal loss in the global ocean. - Science 347: 247–254.
McClenachan, L. et al. 2016. Rethinking Trade-Driven Extinction Risk in Marine and Terrestrial Megafauna. - Curr. Biol. 26: 1640–1646.
McColl-Gausden, E. F. et al. 2020. Multi-species models reveal that eDNA metabarcoding is more sensitive than backpack electrofishing for conducting fish surveys in freshwater streams. - Mol. Ecol. in press.
McElroy, M. E. et al. 2020. Calibrating Environmental DNA Metabarcoding to Conventional Surveys for Measuring Fish Species Richness. - Front. Ecol. Evol. 8: 1–12.
McGill, B. J. et al. 2015. Fifteen forms of biodiversity trend in the anthropocene. - Trends Ecol. Evol. 30: 104–113.
McLean, M. et al. 2018. A Climate-Driven Functional Inversion of Connected Marine Ecosystems. - Curr. Biol. 28: 3654-3660.e3.
Melo, B. F. et al. 2016. Cryptic species in the Neotropical fish genus Curimatopsis (Teleostei, Characiformes). - Zool. Scr. 45: 650–658.
Miya, M. et al. 2015. MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than 230 subtropical marine species. - R. Soc. Open Sci. 2: 150088.
Miya, M. et al. 2020. MiFish metabarcoding: a high-throughput approach for simultaneous detection of multiple fish species from environmental DNA and other samples. - Fish. Sci. in press.
Monnet, A. C. et al. 2014. Asynchrony of taxonomic, functional and phylogenetic diversity in birds. - Glob. Ecol. Biogeogr. 23: 780–788.
Mora, C. et al. 2011. How many species are there on earth and in the ocean? - PLoS Biol. 9: 1–8.
Mouillot, D. et al. 2014. Functional over-redundancy and high functional vulnerability in global fish faunas on tropical reefs. - Proc. Natl. Acad. Sci. U. S. A. 111: 13757–13762.
Nagelkerken, I. et al. 2020. Trophic pyramids reorganize when food web architecture fails to adjust to ocean change. - Science 369: 829–832.
Nearing, J. T. et al. 2018. Denoising the Denoisers: an independent evaluation of microbiome sequence

error-correction approaches. - PeerJ 6: e5364.
Nelson, J. S. et al. 2016. Fishes of the world (Wiley & Sons, Ed.).
Nguyen, B. et al. 2019. Environmental DNA survey captures patterns of fish and invertebrate diversity across a tropical seascape. - Sci. Rep. 10: 6729.
Nielsen, T. F. et al. 2019. More is less: net gain in species richness, but biotic homogenization over 140 years. - Ecol. Lett. 22: 1650–1657.
Nordberg, E. J. and Schwarzkopf, L. 2019. Reduced competition may allow generalist species to benefit from habitat homogenization. - J. Appl. Ecol. 56: 305–318.
Nunes, L. T. et al. 2020. Habitat and community structure modulate fish interactions in a neotropical clearwater river. - Neotrop. Ichthyol. 18: 1–20.
Palacios, D. M. et al. 2012. Cetacean distribution and relative abundance in Colombia's Pacific EEZ from survey cruises and platforms of opportunity. - J. Cetacean Res. Manag. 12: 45–60.
Patin, N. V. et al. 2013. Effects of OTU Clustering and PCR Artifacts on Microbial Diversity Estimates. - Microb. Ecol. 65: 709–719.
Pauly, D. and Zeller, D. 2016. Catch reconstructions reveal that global marine fisheries catches are higher than reported and declining. - Nat. Commun. 7: 10244.
Pawlowski, J. et al. 2018. The future of biotic indices in the ecogenomic era: Integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. - Sci. Total Environ. 637–638: 1295–1310.
Pecl, G. T. et al. 2017. Biodiversity redistribution under climate change: Impacts on ecosystems and human well-being. - Science 355: eaai9214.
Petchey, O. L. and Gaston, K. J. 2006. Functional diversity: Back to basics and looking forward. - Ecol. Lett. 9: 741–758.
Pigot, A. L. et al. 2020. Macroevolutionary convergence connects morphological form to ecological function in birds. - Nat. Ecol. Evol. 4: 230–239.
Pimiento, C. et al. 2020. Functional diversity of marine megafauna in the Anthropocene. - Sci. Adv. 6: eaay7650.
Pimm, S. L. et al. 2014. The biodiversity of species and their rates of extinction, distribution, and protection. - Science 344: 1246752.
Pinheiro, H. T. et al. 2016. Upper and lower mesophotic coral reef fish communities evaluated by underwater visual censuses in two Caribbean locations. - Coral Reefs 35: 139–151.
Pinheiro, H. T. et al. 2019. Will DNA barcoding meet taxonomic needs? 365: 873–875.
Pinsky, M. L. et al. 2018. Preparing ocean governance for species on the move. - Science 360: 1189–1191.
Pinsky, M. L. et al. 2019. Greater vulnerability to warming of marine versus terrestrial ectotherms. - Nature 569: 108–111.
Pollock, L. J. et al. 2017. Large conservation gains possible for global biodiversity facets. - Nature 546: 141–144.
Pollock, L. J. et al. 2020. Protecting Biodiversity (in All Its Complexity): New Models and Methods. - Trends Ecol. Evol. in press.
Pont, D. et al. 2018. Environmental DNA reveals quantitative patterns of fish biodiversity in large rivers despite its downstream transportation. - Sci. Rep. 8: 1–13.
Postaire, B. D. et al. 2020. Environmental DNA detection tracks established seasonal occurrence of blacktip sharks in a semi - enclosed subtropical bay. - Sci. Rep.: 1–8.
Purcell, S. W. et al. 2014. The cost of being valuable: Predictors of extinction risk in marine invertebrates exploited as luxury seafood. - Proc. R. Soc. B Biol. Sci. 281: 20133296.
Quince, C. et al. 2009. Accurate determination of microbial diversity from 454 pyrosequencing data. - Nat. Methods 6: 639–641.
Razgour, O. et al. 2019. Considering adaptive genetic variation in climate change vulnerability assessment reduces species range loss projections. - Proc. Natl. Acad. Sci. U. S. A. 116: 10418–10423.
Reboredo Segovia, A. L. et al. 2020. Who studies where? Boosting tropical conservation research where

it is most needed. - Front. Ecol. Environ. 18: 159–166.
Reeder, J. and Knight, R. 2009. The “rare biosphere”: a reality check. - Nat. Methods 6: 636–637.
Rees, H. C. et al. 2014. The application of eDNA for monitoring of the Great Crested Newt in the UK. - Ecol. Evol. 4: 4023–4032.
Reid, S. M. and Haxton, T. J. 2017. Backpack electrofishing effort and imperfect detection: Influence on riverine fish inventories and monitoring. - J. Appl. Ichthyol. 33: 1083–1091.
Reynolds, J. D. et al. 2005. Biology of extinction risk in marine fishes. - Proc. R. Soc. B Biol. Sci. 272: 2337–2344.
Ripple, W. J. et al. 2017. Extinction risk is most acute for the world’s largest and smallest vertebrates. - Proc. Natl. Acad. Sci.: 201702078.
Ripple, W. J. et al. 2019. Are we eating the world’s megafauna to extinction? - Conserv. Lett.: 1–10.
Rocha, L. A. et al. 2018. Mesophotic coral ecosystems are threatened and ecologically distinct from shallow water reefs. 284: 281–284.
Roff, G. et al. 2018. Decline of coastal apex shark populations over the past half century. - Commun. Biol. 1: 1–11.
Rognes, T. et al. 2016. VSEARCH: a versatile open source tool for metagenomics. - PeerJ 4: e2584.
Roskov, Y. et al. 2019. Species 2000 & ITIS Catalogue of Life, 2019 Annual Checklist.
Saladin, B. et al. 2020. Rapid climate change results in long-lasting spatial homogenization of phylogenetic diversity. - Nat. Commun. 11: 1–8.
Sales, N. G. et al. 2021. Space-time dynamics in monitoring neotropical fish communities using eDNA metabarcoding. - Sci. Total Environ. 754: 142096.
Säterberg, T. et al. 2013. High frequency of functional extinctions in ecological networks. - Nature 499: 468–470.
Scheffers, B. R. and Pecl, G. 2019. Persecuting, protecting or ignoring biodiversity under climate change. - Nat. Clim. Chang. 9: 581–586.
Schmid, S. et al. 2017. HyRAD-X, a versatile method combining exome capture and RAD sequencing to extract genomic information from ancient DNA. - Methods Ecol. Evol. 8: 1374–1388.
Schnell, I. B. et al. 2015. Tag jumps illuminated - reducing sequence-to-sample misidentifications in metabarcoding studies. - Mol. Ecol. Resour. 15: 1289–1303.
Schramm, K. D. et al. 2020. A comparison of stereo-BRUV, diver operated and remote stereo-video transects for assessing reef fish assemblages. - J. Exp. Mar. Bio. Ecol. 524: 151273.
Siqueira, A. C. et al. 2020. Trophic innovations fuel reef fish diversification. - Nat. Commun. 11: 2669.
Stat, M. et al. 2019. Combined use of eDNA metabarcoding and video surveillance for the assessment of fish biodiversity. - Conserv. Biol. 33: 196–205.
Stefanoudis, P. V. et al. 2019. Depth-dependent structuring of reef fish assemblages from the shallows to the rariphotic zone. - Front. Mar. Sci. 6: 1–16.
Steffen, W. et al. 2018. Trajectories of the Earth System in the Anthropocene. - Proc. Natl. Acad. Sci. 115: 8252–8259.
Stein, R. W. et al. 2018. Global priorities for conserving the evolutionary history of sharks, rays and chimaeras. - Nat. Ecol. Evol. 2: 288–298.
Stoeckle, M. Y. et al. 2017. Aquatic environmental DNA detects seasonal fish abundance and habitat preference in an urban estuary. - PLoS One 12: 1–15.
Sunday, J. M. et al. 2012. Thermal tolerance and the global redistribution of animals. - Nat. Clim. Chang. 2: 686–690.
Thiault, L. et al. 2019. Escaping the perfect storm of simultaneous climate change impacts on agriculture and marine fisheries. - Sci. Adv. 5: eaaw9976.
Thomsen, P. F. et al. 2012. Detection of a Diverse Marine Fish Fauna Using Environmental DNA from Seawater Samples. - PLoS One 7: 1–9.
Thomsen, P. F. et al. 2016. Environmental DNA from seawater samples correlate with trawl catches of subarctic, deepwater fishes. - PLoS One 11: 1–22.
Thuiller, W. et al. 2011. Consequences of climate change on the tree of life in Europe. - Nature 470: 531–534.

Trenkel, V. M. et al. 2019. We can reduce the impact of scientific trawling on marine ecosystems. - Mar. Ecol. Prog. Ser. 609: 277–282.
Trindade-Santos, I. et al. 2020. Global change in the functional diversity of marine fisheries exploitation over the past 65 years. - Proceedings. Biol. Sci. 287: 20200889.
Tsuji, S. et al. 2019. The detection of aquatic macroorganisms using environmental DNA analysis-A review of methods for collection, extraction, and detection. - Environ. DNA: 1–10.
Tucker, C. M. et al. 2016. A guide to phylogenetic metrics for conservation, community ecology and macroecology. - Biol. Rev. 92: 698–715.
Turner, C. R. et al. 2014. Particle size distribution and optimal capture of aqueous macrobial eDNA. - Methods Ecol. Evol. 5: 676–684.
Valdivia-Carrillo, T. et al. 2019. Beyond traditional biodiversity fish monitoring: environmental DNA metabarcoding and simultaneous underwater visual census detect different sets of a complex fish community at a marine biodiversity hotspot. - bioRxiv in press.
Valentini, A. et al. 2009. DNA barcoding for ecologists. - Trends Ecol. Evol. 24: 110–117.
Vasselon, V. et al. 2017. Assessing ecological status with diatoms DNA metabarcoding: Scaling-up on a WFD monitoring network (Mayotte island, France). - Ecol. Indic. 82: 1–12.
Velazco, P. M. et al. 2017. Santamartamys rufodorsalis (Rodentia: Echimyidae). - Mamm. Species 49: 63–67.
Vellend, M. et al. 2013. Global meta-analysis reveals no net change in local-scale plant biodiversity over time. - Proc. Natl. Acad. Sci. 110: 19456–19459.
Vergés, A. et al. 2014. The tropicalization of temperate marine ecosystems: Climate-mediated changes in herbivory and community phase shifts. - Proc. R. Soc. B Biol. Sci. 281: 20140846.
Villéger, S. et al. 2017. Functional ecology of fish: current approaches and future challenges. - Aquat. Sci. 79: 783–801.
Violle, C. et al. 2007. Let the concept of trait be functional! - Oikos 116: 882–892.
Wangensteen, O. S. et al. 2018. DNA metabarcoding of littoral hard-bottom communities: high diversity and database gaps revealed by two molecular markers. - PeerJ 6: e4705.
Weigand, H. et al. 2019. DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. - Sci. Total Environ. 678: 499–524.
Welch, D. B. M. and Huse, S. M. 2011. Microbial Diversity in the Deep Sea and the Underexplored “Rare Biosphere.” - Handb. Mol. Microb. Ecol. II Metagenomics Differ. Habitats: 243–252.
West, K. M. et al. 2020. eDNA metabarcoding survey reveals fine-scale coral reef community variation across a remote, tropical island ecosystem. - Mol. Ecol. 29: 1069–1086.
Whittaker, R. H. 1972. Evolution and Measurement of Species Diversity. - Taxon 21: 213–251.
Wiens, J. J. 2016. Climate-Related Local Extinctions Are Already Widespread among Plant and Animal Species. - PLoS Biol. 14: 1–18.
Williams, M. A. et al. 2019. The application of CRISPR-Cas for single species identification from environmental DNA. - Mol. Ecol. Resour. 19: 1106–1114.
Wilson, R. W. et al. 2009. Contribution of fish to the marine inorganic carbon cycle. - Science 323: 359–362.
Worm, B. et al. 2013. Global catches, exploitation rates, and rebuilding options for sharks. - Mar. Policy 40: 194–204.
Yoccoz, N. G. et al. 2018. Biodiversity may wax or wane depending on metrics or taxa. - Proc. Natl. Acad. Sci. U. S. A. 115: 1681–1682.
Zhang, H. et al. 2020a. Extinction of one of the world’s largest freshwater fishes: Lessons for conserving the endangered Yangtze fauna. - Sci. Total Environ. 710: 1–7.
Zhang, S. et al. 2020b. A comprehensive and comparative evaluation of primers for metabarcoding eDNA from fish. - Methods Ecol. Evol. in press.

Annexes

1. Accumulation curves of environmental DNA sequences predict coastal fish diversity in the coral triangle

PROCEEDINGS B

royalsocietypublishing.org/journal/rspb

Research

Cite this article: Juhel J-B *et al*. 2020 Accumulation curves of environmental DNA sequences predict coastal fish diversity in the coral triangle. *Proc. R. Soc. B* **287**: 20200248. http://dx.doi.org/10.1098/rspb.2020.0248

Received: 6 February 2020
Accepted: 18 June 2020

Subject Category:
Ecology

Subject Areas:
ecology

Keywords:
eDNA metabarcoding, sequence clustering, Operational Taxonomic Unit, diversity assessment, detectability

Author for correspondence:
Jean-Baptiste Juhel
e-mail: jeanbaptiste.juhel@gmail.com

†Joint last authorship.

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.5047669.

THE ROYAL SOCIETY PUBLISHING

Accumulation curves of environmental DNA sequences predict coastal fish diversity in the coral triangle

Jean-Baptiste Juhel[1], Rizkie S. Utama[2], Virginie Marques[1], Indra B. Vimono[2], Hagi Yulia Sugeha[2], Kadarusman[3], Laurent Pouyaud[4], Tony Dejean[5], David Mouillot[1,6,†] and Régis Hocdé[1,†]

[1]MARBEC, Univ. Montpellier, CNRS, Ifremer, IRD, Montpellier, France
[2]Research Center for Oceanography, Indonesian Institute of Sciences, Jl. Pasir Putih 1, Ancol Timur, Jakarta Utara, Indonesia
[3]Politeknik Kelautan dan Perikanan Sorong, KKD BP Sumberdaya Genetik, Konservasi dan Domestikasi, Papua Barat 98411, Indonesia
[4]Institut des Sciences de l'Evolution de Montpellier, Montpellier, France
[5]SPYGEN, 73370 Le Bourget-du-Lac, France
[6]ARC Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, Australia

J-BJ, 0000-0003-2627-394X; VM, 0000-0002-5142-4191; K, 0000-0003-2312-2417; LP, 0000-0003-4415-9198; DM, 0000-0003-0402-2605; RH, 0000-0002-5794-2598

Environmental DNA (eDNA) has the potential to provide more comprehensive biodiversity assessments, particularly for vertebrates in species-rich regions. However, this method requires the completeness of a reference database (i.e. a list of DNA sequences attached to each species), which is not currently achieved for many taxa and ecosystems. As an alternative, a range of operational taxonomic units (OTUs) can be extracted from eDNA metabarcoding. However, the extent to which the diversity of OTUs provided by a limited eDNA sampling effort can predict regional species diversity is unknown. Here, by modelling OTU accumulation curves of eDNA seawater samples across the Coral Triangle, we obtained an asymptote reaching 1531 fish OTUs, while 1611 fish species are recorded in the region. We also accurately predict ($R^2 = 0.92$) the distribution of species richness among fish families from OTU-based asymptotes. Thus, the multi-model framework of OTU accumulation curves extends the use of eDNA metabarcoding in ecology, biogeography and conservation.

1. Introduction

Providing accurate biodiversity assessments is a critical goal in ecology and biogeography with estimations being constantly revised for some species-rich groups [1]. This issue is increasingly important, given the accelerating human footprint on Earth. The ongoing worldwide defaunation, characterized by massive population declines, may trigger the local or even global extinction of rare, elusive and cryptic species that are still unknown or poorly documented [2,3]. Such biodiversity losses directly impact ecosystem functioning, but also human health, well-being and livelihood [4,5]. This urges scientists to improve the accuracy and extend the breadth of biodiversity inventories and monitoring.

In the marine realm, the detection of species occurrences is particularly challenging due to the vast volume to monitor, the high diversity of habitats, the inaccessibility of some areas (e.g. deep sea) and the behaviour of some species (cryptobenthic or elusive) [6,7]. Environmental DNA (eDNA) metabarcoding is an emerging tool that can provide more accurate and wider biodiversity assessments than classical census methods, particularly for rare and elusive species [8–10]. This non-invasive method is based on retrieving DNA naturally released by organisms in their environment, amplified by polymerase chain reaction (PCR) and then sequenced to ultimately identify corresponding species [11]. However, inventorying and monitoring biodiversity using eDNA metabarcoding

requires the completeness of a reference database to accurately assign each sequence to a given species (e.g. [9]).

By now, only a minority of fish species are present in online DNA databases for mitochondrial regions targeted by metabarcoding markers, limiting the extent to which species diversity can be revealed by eDNA. This proportion of sequenced species is even lower in species-rich regions and poorly sampled habitats or taxa, while the effort to complete genetic reference databases is long and costly. As an alternative, a range of operational taxonomic units (OTUs) can be extracted from eDNA metabarcoding through filtering and clustering techniques [12]. Even if environmental genomics approaches have a long tradition of using OTU-based bioindicators [13], the extent to which the diversity of OTUs from a limited number of eDNA samples can reveal or predict the diversity of vertebrate species in a given biodiversity hotspot has not yet been investigated. This is particularly challenging for cryptobenthic fish species, which are key for reef ecosystems [14] but usually missed by classical surveys [7]. We thus urgently need a regional case study with a wide breadth of fish families and traits to test the potential of OTU-based assessment of biodiversity.

The Bird's Head Peninsula of West Papua (eastern Indonesia) is located in the centre of the Coral Triangle, which is known to host the world's richest marine biodiversity [15,16]. The current checklist of coastal fishes in the Bird's Head Peninsula identifies 1611 species belonging to 508 genera and 112 families [15,17], among which some are still poorly described or under severe threat [18–20]. Providing a blind but accurate assessment of the level and composition of a well-known vertebrate diversity from eDNA OTUs is thus a critical step in conservation, biogeography and ecology, particularly in such biodiversity hotspots.

Here, using eDNA metabarcoding from 92 seawater samples across the Bird's Head Peninsula, we (i) assessed the diversity of coastal fish species based on an online reference database for the teleo primers region of the 12S mitochondrial rDNA gene [21], (ii) estimated the diversity of fish OTUs based on a custom filtering and clustering bioinformatic pipeline, and (iii) tested the capacity of OTU accumulation curves to predict the level and composition of regional fish diversity.

2. Methods

(a) Sampling area and protocol

A total of 92 water samples were collected during October and November 2017 along the south coast of the Bird's Head region of West Papua (500 km) across different habitats but mainly coral reefs (electronic supplementary material, figure S1). Samples were collected in DNA-free plastic bags at the surface from a dinghy boat, at depths between 10 and 100 m during close circuit rebreather dives and at depths between 100 and 300 m using Niskin water samplers. A pressure and temperature sensor was coupled to the Niskin bottle to control the sampling depth and characterize the water mass via the vertical temperature profile. For each sample, 2 l of seawater was filtered with sterile Sterivex filter capsules (Merck Millipore; pore size 0.22 µm) using disposable sterile syringes. Immediately after, the filter units were filled with lysis conservation buffer (CL1 buffer SPYGEN) and stored in 50 ml screw-cap tubes at −20°C. A contamination control protocol was followed in both field and laboratory stages [21,22]. Water sample processing included the use of disposable gloves and single-use filtration equipment, and the bleaching (50% bleach) of Niskin water sampler.

(b) DNA extraction, amplification and high-throughput sequencing

The DNA extraction and amplification were performed following the protocol of [23], including 12 separate PCR amplifications per sample (see electronic supplementary material for more details on the protocol). A teleost-specific 12S mitochondrial rDNA primer (teleo, forward primer-ACACCGCCCGTCACTCT, reverse primer -CTTCCGGTACACTTACCATG [21]) was used for the amplification of metabarcoding sequences, generating 63 ± 3pb (mean ± s.d.) long amplicons for all fish species referenced in EMBL database (European Molecular Biology Laboratory, www.ebi.ac.uk, v. 138, downloaded on January 2019) [24]. Eight negative extraction controls and two negative PCR controls (ultrapure water) were amplified (with 12 replicates as well) and sequenced in parallel to the samples to monitor possible contaminations. The teleo primers were 5′-labelled with an eight-nucleotide tag unique to each PCR replicate with at least three differences between any pair of tags, allowing the assignment of each sequence to the corresponding sample during sequence analysis. The tags for the forward and reverse primers were identical for each PCR replicate.

The purified PCR products were pooled in equal volumes, to achieve a theoretical sequencing depth of 1 000 000 reads per sample. Library preparation and sequencing were performed at Fasteris (Geneva, Switzerland). A total of five libraries were prepared using the MetaFast protocol (Fasteris, https://www.fasteris.com/dna/?q=content/metafast-protocol-amplicon-metagenomic-analysis), a ligation-based PCR-free library preparation. A paired-end sequencing (2 × 125 bp) was carried out using an Illumina HiSeq 2500 sequencer on three HiSeq Rapid Flow Cell v. 2 using the HiSeq Rapid SBS Kit v. 2 (Illumina, San Diego, CA, USA) following the manufacturer's instructions.

(c) Sequence analyses and taxonomic assignment

To evaluate the current completeness of the online database for the teleo region of the 12S mitochondrial DNA, an *in silico* PCR with 3 allowed mismatches using the teleo primers sequences was performed with ecoPCR [25] on the EMBL database. The generated list of sequenced species was compared with the checklists of fish species present in the Bird's Head region of Papua, provided by courtesy of Kulbicki *et al.* [17].

The amplified DNA sequences from the water samples were processed following two metabarcoding workflows. The first workflow used the OBITools software package [26] based on direct taxonomic assignment of the sequences using the ecotag lower common ancestor algorithm in EMBL database as a reference (see details in electronic supplementary material).

The ecotag algorithm can sometimes wrongly assign sequences to a given species or genus, despite a low-similarity percentage due to the incompleteness of reference database. We thus set the following similarity thresholds, 100–98, 90–98, 85–90 and 80–85% bp to assign sequences at the species, genus, family and order level, respectively. All the assignments with a similarity percentage lower than 80% were discarded from the analyses.

We evaluated the database completeness for the marker by running an *in silico* PCR on all fish mitochondrial DNA present in EMBL online database. A total of 394 species are sequenced in the Bird's Head region (24.5%, electronic supplementary material, table S1).

The second metabarcoding workflow was based on the SWARM clustering algorithm that groups multiple variants of sequences into OTUs [12]. Then, a post-clustering curation algorithm (LULU) was performed to curate data (see details in electronic supplementary material).

The SWARM clustering workflow was used to investigate the taxa present in the samples but not revealed by the taxonomic assignment process because of gaps in the EMBL database. The

number of taxa assigned in each family was corrected to avoid taxonomical redundancy assignment. For instance, the combined assignments to the genus *Zanclus* and the species *Zanclus cornutus* were considered as one taxa as potential PCR error may have produced two different assignment levels from the same sequence. These corrected numbers of taxa were then compared to the number of OTUs from the SWARM workflow in each family to evaluate the magnitude of the diversity missed by the direct assignment method. In the SWARM workflow, a family-level assignment was performed as well to remove the taxa that were not fish from non-specific amplifications and investigate the intrafamily diversity.

(d) Statistical analyses

To evaluate the number of taxa/OTUs present in the study area, a multi-model approach was implemented to fit asymptotes on the species and OTU accumulation curves. This approach considered five different accumulation models (Lomolino, Michaelis–Menten, Gompertz, asymptotic regression and logistic curve) and weighted them using the Akaike information criterion (AIC) [29]. For each curve, the accumulation model with the lowest AIC was selected. Accumulation curves and associated asymptotes were generated using the vegan R package. To estimate the sampling effort required to achieve a given proportion of asymptotes, we considered the model selected for accumulation curves. Then, we extracted the predicted number of samples producing a number of taxa/OTUs that outreached 90% and 95% of the asymptotes.

3. Results

(a) High heterogeneity of fish species detection among families

A total of 299 479 007 reads were produced using the OBITools pipeline over the 92 eDNA samples corresponding to 14 423 unique sequences with a mean of 307 unique sequences per sample (± 134 s.d.). In a conservative approach, stringent bioinformatic filters retained 9345 unique sequences, so 65% of the total. These 9345 unique sequences were then assigned to different taxonomic levels using the following genetic similarity thresholds: 98–100% for species, 90–98% for genus, 85–90% for family and 80–85% for order. This set of thresholds retained 7389 unique sequences resulting in 678 taxonomic assignments (electronic supplementary material, table S2).

A total of 310 species were detected, including 211 coastal fish species present in the checklist of the Bird's Head Peninsula and 99 fish species present in other regions but absent from this checklist (figure 1*a*). Conversely, 183 sequenced fish species which are present in the Bird's Head Peninsula were not detected in our eDNA samples using our stringent filters, representing 53.6% of the sequenced species present in the checklist. Since 75.5% of fish species in the checklist of the Bird's Head Peninsula were not sequenced for the 12S rDNA, the largest part of fish species diversity remained hidden through direct assignment (electronic supplementary material, table S1).

A total of 282 genera and 128 families of fish were detected compared with the regional checklist of 508 genera and 112 families out of which 46.1% and 72.3% are sequenced, respectively (electronic supplementary material, table S1). The number of fish species per family varied from 1 to 191 in the Bird's Head checklist (figure 1*b*), the richest family being the Gobiidae. Only 12 species of Gobiidae were detected in our 92 samples. Meanwhile, the most represented family in the eDNA samples was the Labridae with 48 species (15.5% of the species found in the samples) out of 136 in the checklist (figure 1*b*).

The percentage of fish species sequenced per family varied between 0 and 100% with a mean of 40.3% (±31% s.d) in the Bird's Head Peninsula checklist while the percentage of detected species per family varied between 0 and 100% with a mean of 27.1% (±30.2% s.d.) in eDNA samples (figure 1*b*). These two percentages were significantly and strongly related ($p < 0.001$) with the percentage of species sequenced per family explaining 85% of variation in the percentage of detected species per family (figure 1*c*).

(b) High but underestimated diversity of operational taxonomic units

Given that the low percentage of fish species sequenced for the 12S in the region is the main limitation to detect taxonomic diversity (figure 1*c*), we used an alternative approach based on unique clusters of genetic sequences called OTUs.

From the 331 839 591 initial reads, 4012 OTUs were generated using the SWARM clustering algorithm. After a series of post-clustering curation processes, 972 fish OTUs were filtered among which 819 were assigned to a family (electronic supplementary material, table S3). The number of detected OTUs varied from 1 to 54 among fish families (figure 2*a*), the richest families (greater than 50 OTUs) being the Gobiidae, Labridae and Pomacentridae. Overall, the number of OTUs was superior to the number of assigned taxa (genus and species) in 64.7% of the families found in the samples (mean $\Delta = 4 \pm 6.7$ s.d., figure 2*a*). This richness difference was null in 31.4% of the families and negative in 3.9% of them (figure 2*a*). This difference was notably high in some rich families such as the Gobiidae and Pomacentridae where the number of OTUs was more than 2 times and 1.5 times higher than the number of assigned taxa, respectively. By contrast, only 7 OTUs were produced compared with 11 assigned taxa for the Scombridae so $\Delta = -4$ units or −66.7% of this family richness.

The discrepancy between the two approaches (taxa and OTUs) was not significantly explained neither by the species richness of the family in the checklist ($R^2 < 0.01$, $p = 0.08$, figure 2*b*) nor by the percentage of sequenced fish species within each family in the checklist ($R^2 = 0.09$, $p = 0.05$, figure 2*c*).

On average, the number of OTUs underestimated the total number of coastal fish species in the Bird's Head Peninsula checklist with a mean net difference of 40.2% per family (± 38.8% s.d., figure 2*d*). For most families, this difference was high, reaching the maximum value of 95% for the Pseudochromidae. However, this difference could also be negative with more OTUs detected than species present in the checklist as for the Dasyatidae, Leiognathidae and Orectolobidae for which this difference reached −50%. Overall, the difference was marginally but significantly explained by the species richness of the family in the regional checklist ($R^2 = 0.09$, $p = 0.04$, figure 2*d*), suggesting that the bias is not proportional to the species richness of the family with species-rich families being more underestimated by OTUs than species-poor families.

(c) Prediction of fish species diversity from operational taxonomic unit accumulation curves

Since the two approaches (taxa and OTUs) underestimated the level of taxonomic diversity within fish families with a high

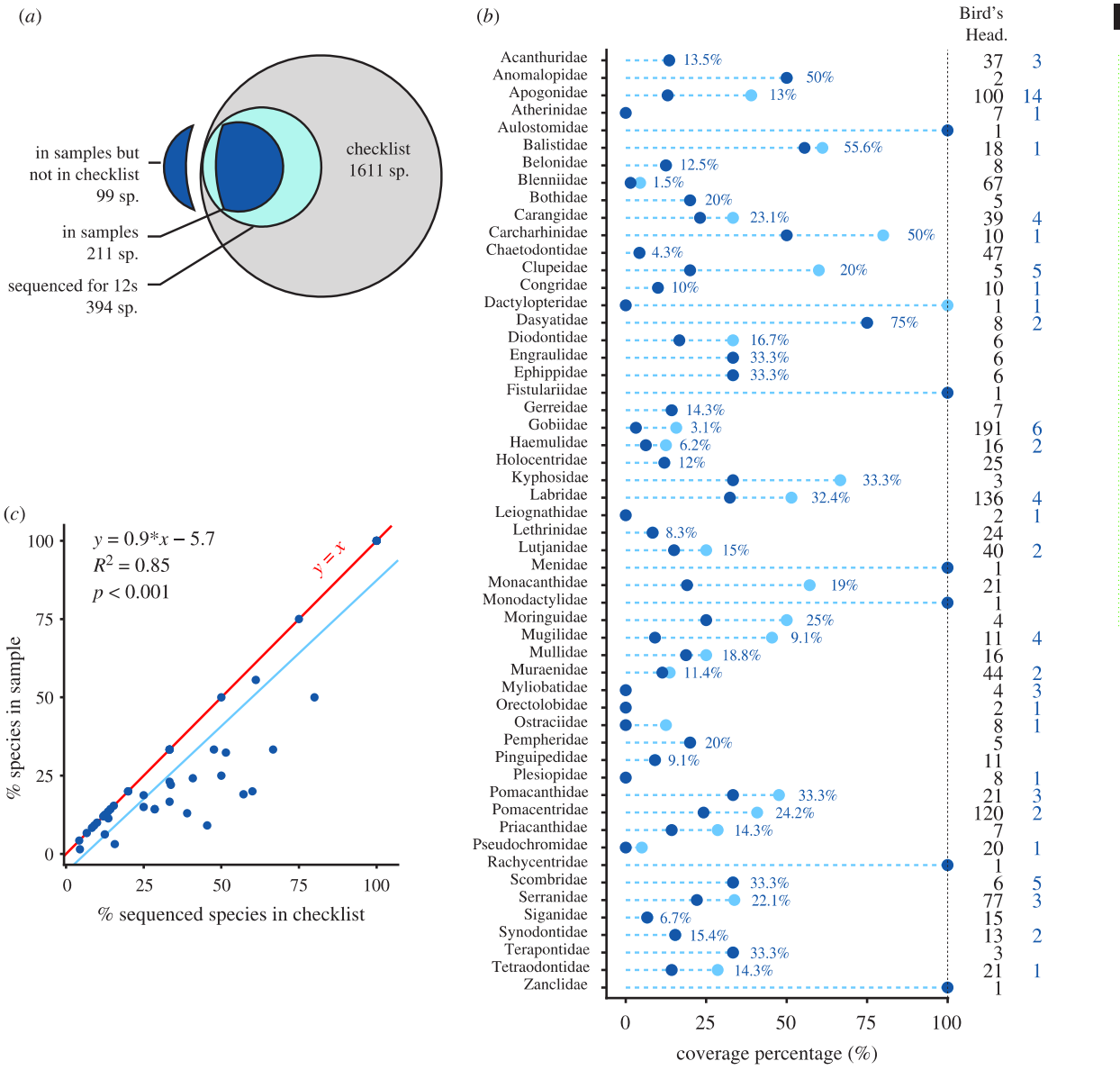


Figure 1. Number of fish species present in the checklist of the Bird's Head region (grey), sequenced in the European Molecular Biology Laboratory database (EMBL) (light blue) and detected in the eDNA samples (dark blue) (*a*); percentage of species detected in the samples (dark blue), sequenced in EMBL (light blue) in each family of species (*b*); percentage of species detected in the samples as a function of the percentage of sequenced species in EMBL (*c*). (*b*) The percentages of the species detected in the eDNA samples compared with the species present in the Bird's Head region are displayed next to the points. The number of species per family in the checklist and the number of species detected in the samples but not present in the checklist are both on the right of the figure in black and dark blue, respectively. Only the sequences assigned to species using ecotag program (similarity >98%) are used in this figure. (*c*) Each point corresponds to a fish family. (Online version in colour.)

uncertainty, we modelled accumulation curves from the diversity of species and OTUs found across our 92 samples. The modelled asymptote of the assigned species reached 429 species, a value very close to the 394 sequenced species present in the Bird's Head Peninsula, but 3.7 times lower than the 1611 species in the regional checklist (figure 3*a*). Meanwhile, the OTU accumulation curve reached an asymptote of 1531; a value close (95%) to the number of fish species (1611) referenced in the checklist of the Bird's Head Peninsula.

Applying this method to the 15 fish families which counted more than 10 OTUs and 10 species in the checklist permitted to assess the ability of eDNA-based accumulation curves to predict regional fish richness. For instance, the OTU accumulation curves for the Gobiidae, Labridae and Pomacentridae, the three richest families (51, 54 and 53 OTUs, respectively), produced asymptotes and thus predictions of fish diversity much lower than those in the regional checklists with 107.5, 66.1 and 76.2 OTUs (i.e. 47.5%, 81.7% and 69.6% of the checklist richness respectively; figure 3*b–d*).

We then tested the ability of the assigned taxa, the OTUs and the OTU accumulation curve approaches to predict fish species richness within families of the regional checklist, so the predictive power of linear or proportional relationships. The total number of assigned taxa per family in our samples was a significant but weak predictor of the number of fish species per family in the checklist ($R^2 = 0.60$, $p < 0.001$,

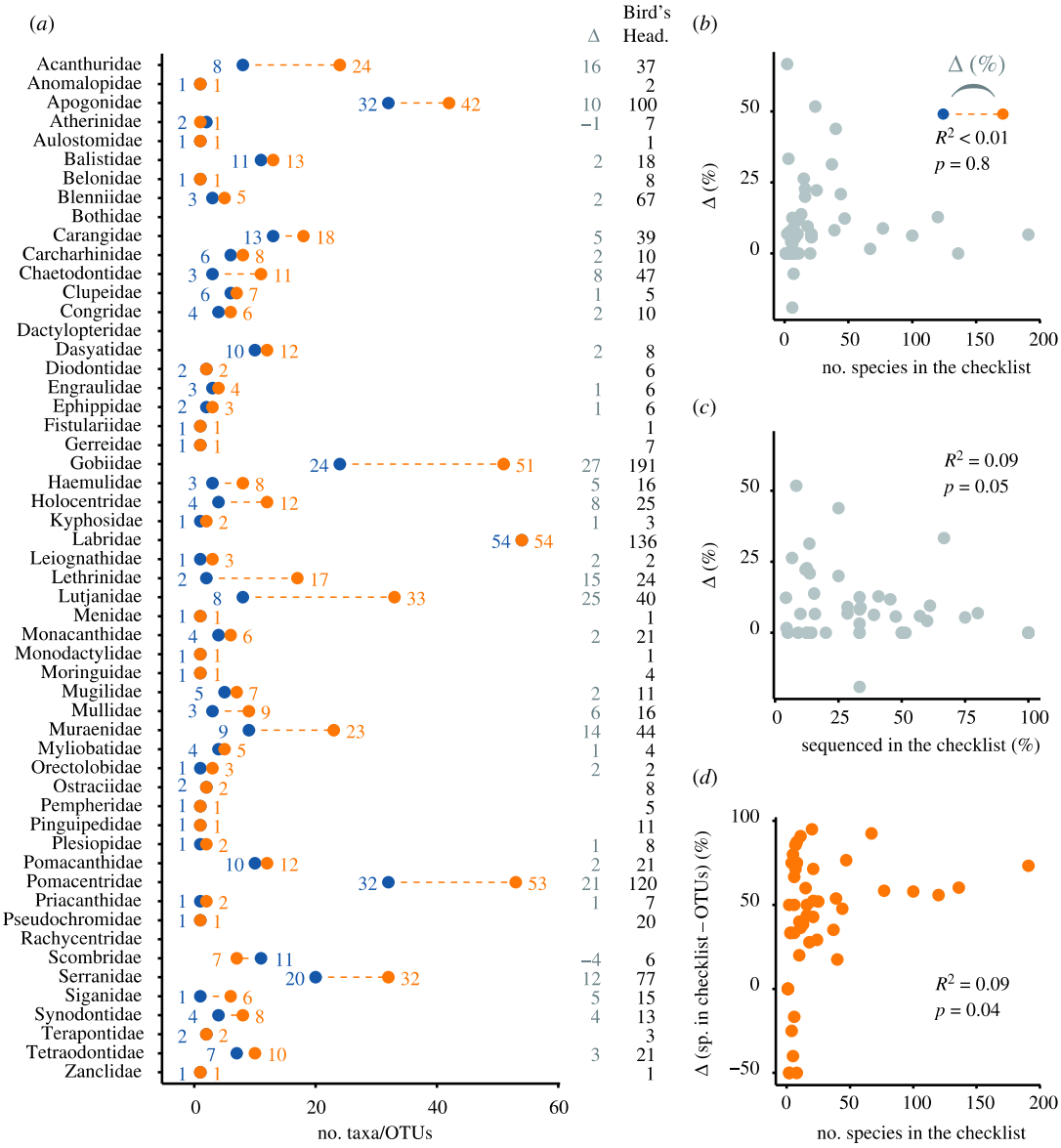


Figure 2. Number of taxa assigned by the OBITools workflow (blue) and number of OTUs generated by the SWARM workflow (orange) in the different fish families (*a*); distribution of the differences between the two workflows as a function of family richness (*b*) and as a function of family sequencing coverage (*c*); distribution of the differences between OTUs and the number of species in the checklist as a function of family richness (*d*). (*a*) The difference of taxa/OTUs between the two methods (noted Δ) and the number of species in the checklist of the Bird's Head region are on the right of the figure in grey and black, respectively. For the OBITools workflow, only the sequences assigned to species and genus using ecotag program (similarity greater than 98% and greater than 90% respectively) are used in this figure. For the SWARM workflow, only the OTUs curated by LULU and assigned to family (similarity greater than 85%) are used in this figure. (Online version in colour.)

figure 4*a*) with the richness of some families being largely underestimated (e.g. 87.4% of net difference with the checklist for the Gobiidae, figure 4*a*,*d*). The number of OTUs per family was a better predictor of the family species richness in the checklist ($R^2 = 0.80$, $p < 0.001$) but left 20% of unexplained variation among families with still a marked underestimation (73.3% of net difference with the checklist for Gobiidae, figure 4*b*,*e*). Using the asymptotes of OTU accumulation curves, we obtained a high predictive accuracy of $R^2 = 0.92$ ($p < 0.001$) for the species richness within families with less bias for the Gobiidae (43.7% of net difference with the checklist) (figure 4*c*,*f*).

In addition, we observed that the net difference between the number of assigned taxa per family and the number of species per fish family in the checklist is not related to the number of species of the families (figure 4*d*), suggesting an absence of systematic bias towards the underestimation of species-rich families. By contrast, the net difference between the number of OTUs per fish family and the number of species per family in the checklist significantly increased ($R^2 = 0.35$, $p = 0.02$) with the number of species per family (figure 4*e*). This bias towards the underestimation of species richness within species-rich families is nonetheless avoided when using the asymptotes of OTU accumulation curves ($p = 0.24$, figure 4*f*). Thus, asymptotes of OTU accumulation curves are most accurate and least biased eDNA-based predictors of fish species diversity within families in this marine biodiversity hotspot.

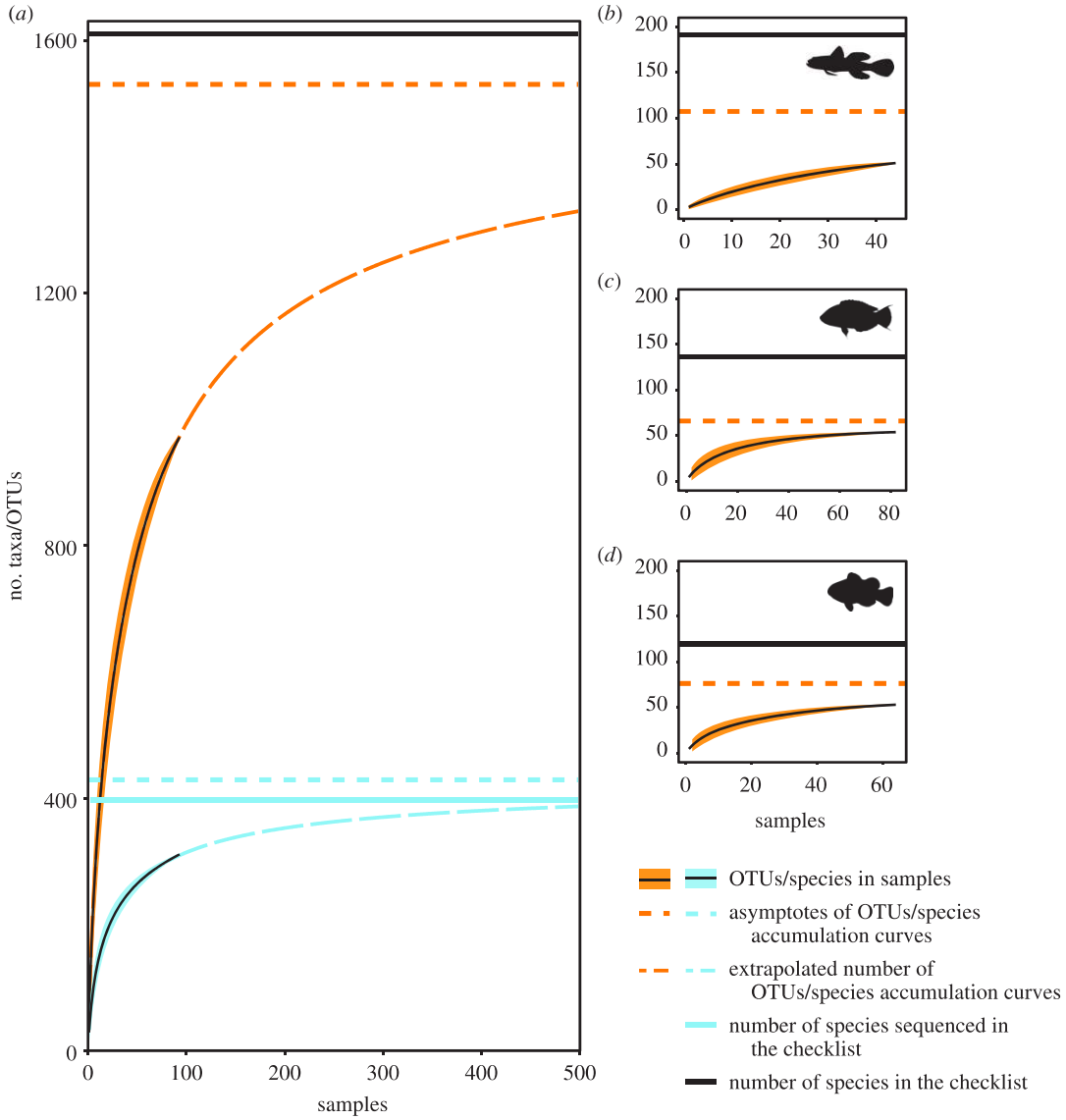


Figure 3. Accumulation curves of species assigned (blue) and the OTUs (orange) obtained in the whole sampling (*a*) and within the three most diverse families: Gobiidae (*b*), Labridae (*c*) and Pomacentridae (*d*). The detection of species and OTUs was randomized 100 times and the results were used to generate the confidence intervals. The asymptotes were modelled by a multi-model approach weighted by the Akaike information criterion (AIC). Fish silhouettes are from phylopic.org (Kent Sorgon & Lily Hughes). (Online version in colour.)

(d) Sampling efforts necessary to achieve regional fish diversity inventory

Not only the OTU accumulation curves and their asymptotes provide diversity estimates, they also provide crucial insights into the sampling effort needed to achieve a more complete census. Here, using the asymptote on the OTU accumulation curve for all fish species (figure 3*a*), we found that our 92 cumulated samples (representing 0.2 m^3) achieved up to 63.5% of the potential fish OTU diversity in the Bird's Head Peninsula (figure 5). To collect 90% of this regional fish diversity, we should have filtered seawater in 735 samples, so eight times the effort of our sampling campaign, representing an aggregated sampled water volume of 1.5 m^3. This sampling effort would reach 1883 samples (an aggregated water volume of 3.8 m^3) to collect 95% of the regional fish OTU richness (figure 5).

On average across fish families, our sampling effort achieved the detection of 77.1% (±14.9 s.d.) of OTUs predicted by the asymptote of the accumulation curve with a variation among families ranging from 42.2% (Muraenidae) and 47.5% (Gobiidae) to 93.9% (Balistidae) (figure 5). The sampling effort needed to achieve 90% of the asymptotic number of OTUs in the region varied greatly among families, ranging from 37 samples for Chaetodontidae to 494 samples for Gobiidae, with a mean of 164 samples (±123 s.d.). The estimated additional sampling effort to reach 95% from 90% of the OTU richness ranged from 20 more samples (Tetraodontidae) to 593 more samples (Gobiidae).

4. Discussion

(a) Overcoming incompleteness of genetic reference databases

Environmental DNA metabarcoding has the potential to surpass most classical survey methods to assess biodiversity in both terrestrial and aquatic systems [30]. Yet, genetic reference

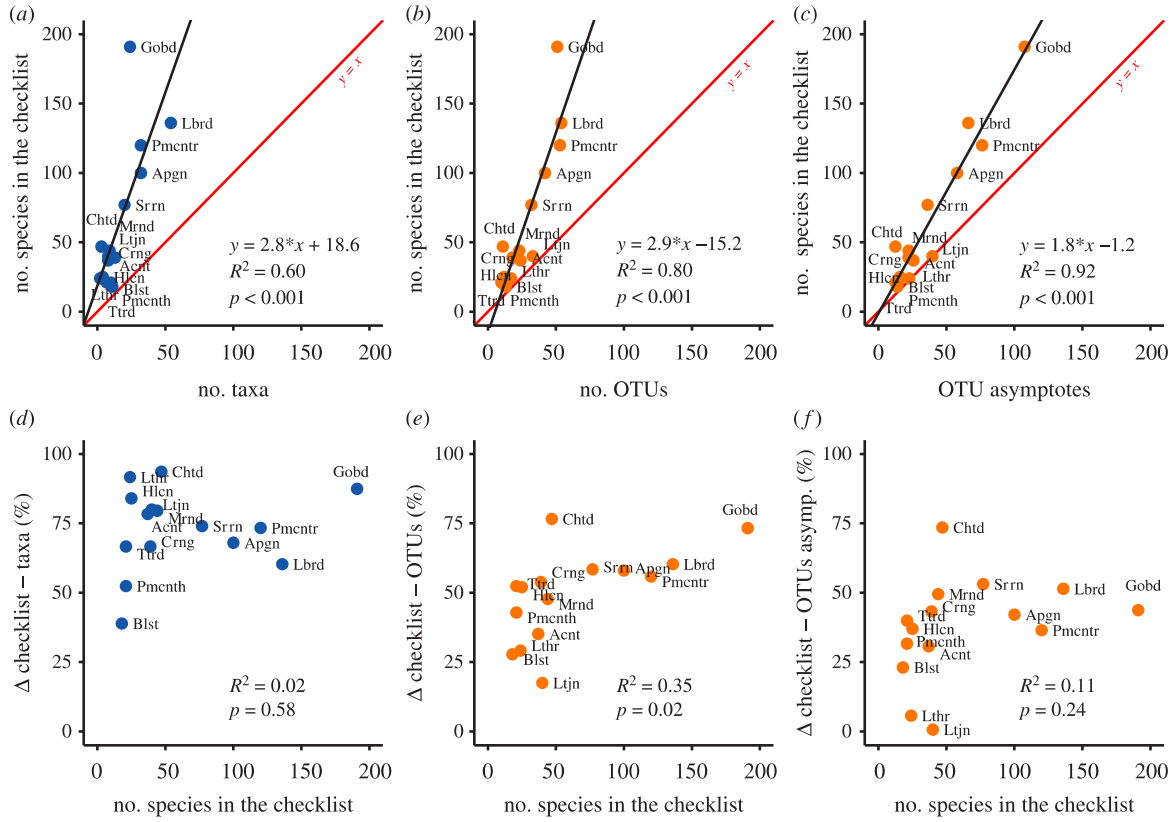


Figure 4. Linear regression of the diversity of the most diverse families as a function of the number taxa assigned (*a*), the number of OTU (*b*), the asymptotes of the OTU accumulation curves (*c*) and differences between the number of taxa assigned (*d*), the number of OTUs (*e*), the asymptotes of OTU accumulation curves (*f*) and the number of species in the checklist as a function of the number of species in the checklist. Only the families with a number of OTU and a number of species in the checklist greater than or equal to 10 are presented to provide accurate estimations. (Online version in colour.)

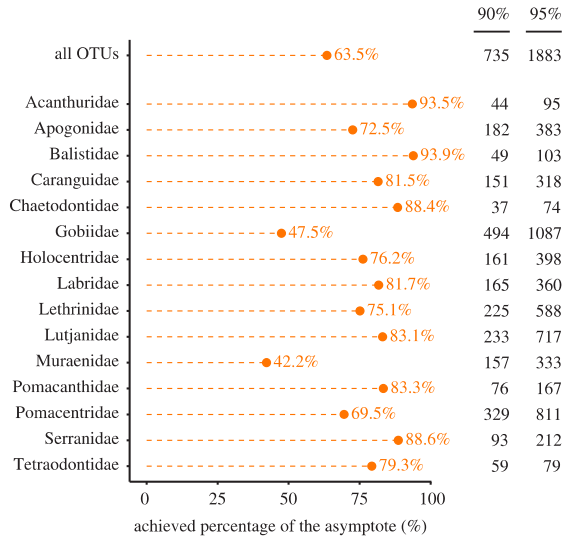


Figure 5. Percentage of the OTUs diversity covered by the current sampling effort ($n = 92$) in the families of fish (orange) and the estimated sampling effort required to achieve both 90% and 95% of the diversity. Only the families with a number of OTU and a number of species in the checklist greater than or equal to 10 are presented to provide accurate estimations. (Online version in colour.)

databases are often incomplete, especially for species-rich ecosystems such as the Coral Triangle, a global marine biodiversity hotspot [14]. For instance, the current completeness of the 12S rDNA online databases for the teleo primer covers only 24.5% of fish species in the Bird's Head Peninsula. Meanwhile, this cover reaches 77.3% for the COI (mitochondrial cytochrome *c* oxidase subunit I), but fish COI primers still perform poorly in comparison to 12S markers [31].

With around 28% of families, 54% of the genera and 76% of species not sequenced for the 12S rDNA teleo primers region, the largest part of fish diversity in the Bird's Head Peninsula remains thus hidden through direct assignment. Additionally, sequences present in the reference online databases may have been collected from individuals not located in the region of interest. This can induce assignment errors due to biogeographic-related genetic variation (e.g. [32]). The lack of sequencing coverage highlights the immense gap to be filled for online databases to be exhaustive, while numerous species still remain to be described [33]. This limitation prevents metabarcoding approaches from characterizing entire fish assemblages through direct species assignment. Yet, the taxa-assignment method reveals the presence of 211 fish species referenced in the checklist of coastal fishes in the Bird's Head Peninsula (figure 1*a*). Conversely, 99 assigned species were absent from this checklist. These 99 detections can either be true presences extending the distribution of some species and revisiting the regional checklist or false presences due to wrong assignments or possible contaminations. For instance, the Atlantic salmon (*Salmo salar*), probably a laboratory kit contaminant, was found in our study and removed from the analyses (see Methods). The large number of species present

in the samples but absent from the regional checklist suggests that inventories of some families are still incomplete. On average, 2.5 detected species per family (±2.6 s.d., figure 1*b*) are absent from the checklist, ranging from 0 to 14 species (Apogonidae). This mismatch allows us to target future sampling efforts towards families and their habitats to complete the regional checklist.

As an alternative to species assignment, the use of OTUs as species proxy units is an option that has not yet been tested for vertebrates in species-rich ecosystems while currently used when the concept of species is debatable like for fungi or unicellular organisms [34,35].

Here, using a conservative and stringent bioinformatic pipeline, we show that the diversity of OTUs is a weak and biased estimator of species diversity with species-rich families being strongly underrepresented. To overcome this limitation, we propose to rely on OTU accumulation curves which provide an unbiased estimate of regional fish diversity and fish richness within families. The asymptotes underestimate the regional fish species richness, but the bias is highly consistent among families (figure 4*f*). We thus propose to extend this method for taxonomic inventories in poorly sampled ecosystems like the deep sea to estimate the diversity at different taxonomic levels.

(b) Revealing the potential and limitation of eDNA metabarcoding inventories

Fishes are the most diverse group of vertebrates on Earth with varying body sizes, environmental niches and diets. Monitoring fish assemblages in marine biodiversity hotspots like the Coral Triangle is a great challenge, particularly for small, rare, cryptobenthic or elusive species. Here, we show that the percentage of sequenced species is highly variable among families preventing any robust estimation of species richness. Instead, OTUs have the potential to reveal the presence of a broad range of fish species (i.e. from different lineages and with contrasted life-history traits). For instance, cryptobenthic families have been poorly documented and are often ignored in traditional visual censuses [7], while they strongly influence ecosystem functioning [13]. Similarly, traditional visual censuses often miss highly mobile and elusive species such as sharks [9].

Among the 310 assigned fish species, we detected the presence of small cryptobenthic species such as *Gobiodon histrio* or *Ostorhinchus selas*, a goby and a cardinalfish with a maximum length below 40 mm, respectively. We also detected large pelagic fish such as the dogtooth tuna (*Gymnosarda unicolor*) or the thresher shark (*Alopias pelagicus*) reaching over 2 m and 4 m long, respectively. Flagship species for conservation were also present in our DNA samples such as the over-exploited Napoleon wrasse (*Cheilinus undulatus*, Endangered, IUCN Red List, www.iucnredlist.org), the Scalloped hammerhead shark (*Sphyrna lewini*, Endangered) and several shark species being classified as Near Threatened (NT) (*C. brevipinna*, *C. Leucas*, *C. sorrah*, *C. melanopterus*, *T. obesus*).

Even if not assigned at species level, OTUs can be defined as distinct entities for which their distribution and temporal variability can be assessed and monitored [36]. Moreover, the OTUs and their associated sequences can remain in public repositories until they are assigned to a species, subspecies or complex as databases improve [37]. However, the major caveat of using OTUs for diversity inventories is that they cannot be directly considered as species with complete certainty. Species with intra-specific genetic variability can produce two separate OTUs, overestimating species diversity. Conversely, two species phylogenetically close to each other with low genetic variability can be grouped into a single OTU, thus underestimating species diversity. The accuracy of diversity inventories using eDNA metabarcoding is thus directly based on the taxonomic resolution of the barcode used and genetic variability among families but also the number of samples.

Here, we also reveal the gap of biodiversity that remains to be detected using OTU accumulation curves. The effort can be massive for some families (figure 5) and more ambitious eDNA sampling campaigns should be on the agenda in species-rich regions like the Coral Triangle. OTU accumulation curves can also serve to evaluate the efficiency of a sampling method (e.g. punctual filtration, transect filtration), the sampled area or the diversity of habitats that are required (e.g. depth, complexity, distance from the seafloor) and their location (e.g. proximity of reefs, hotspots) especially when targeting rare, elusive, highly mobile or cryptobenthic families of fish.

The contrasts between assigned taxa diversity, OTU diversity and OTU asymptote diversity show that the detectability varies strongly among fish families. These contrasts can be related to the ecology of the species but also to the state of the retrieved DNA fragments (intra or extracellular), their sources (e.g. gametes, larvae, faeces), their release rate, their diffusion in the water column (limited or wide) and their transportation [38]. For instance, benthic fish species such as gobies with a small movement range would release DNA fragments through skin and faeces on a small area. However, such species could release a massive number of gametes carried through the water column [13] so may appear highly detectable during breeding season. Further comparative works are urgently needed between visual, camera and eDNA metabarcoding surveys to better estimate the level of detectability of each species or family in order to provide reliable biodiversity assessments. For instance, coupling eDNA metabarcoding and video surveillance allows the detection of 82 fish genera from 13 orders on reefs and seagrass with only 24 genera in common [39]. Investigating biodiversity should also consider its multiple components including functional and phylogenetic diversity that are key for reef ecosystem functioning [40]. Associating OTUs to species might allow us to fill this gap, but it will require massive sampling and sequencing efforts.

Data accessibility. The metadata and bioinformatic outputs are available in the Dryad Digital Repository [41]. The metabarcoding pipelines are available in GitLab (https://gitlab.mbb.univ-montp2.fr/edna/snakemake_only_obitools and https://gitlab.mbb.univ-montp2.fr/edna/bash_swarm).

Authors' contributions. J.-B.J., I.B.V., K., L.P., D.M. and R.H. designed research; J.-B.J and R.H. design the specific research methods of data collection and the sampling strategy; J.-B.J, R.S.U., K. and R.H. collected samples and data; T.D. coordinated the biomolecular analyses; J.-B.J., R.S.U. and V.M. performed the bioinformatics analyses; J.-B.J, R.S.U., V.M., T.D., L.P., D.M. and R.H. defined sequencing strategy, analysed and interpreted data; J.-B.J wrote the initial draft and designed the figures; J.-B.J, R.S.U., V.M., I.B.V., H.Y.S., K., T.D., L.P., D.M. and R.H. wrote the paper and approved the final draft; and L.P., D.M. and R.H. acquired funding to conduct the study.

Competing interests. We declare we have no competing interests.

Funding. Fieldwork and laboratory activities were supported by the Lengguru 2017 Project (www.lengguru.org), conducted by the French National Research Institute for Sustainable Development

(IRD), the Indonesian Institute of Sciences (LIPI) with the Research Center for Oceanography (RCO, the Politeknik KP Sorong), the University of Papua (UNIPA) with the help of the Institut Français in Indonesia (IFI) and with corporate sponsorship from the Total Foundation and TIPCO company. The eDNA sequencing was funded by Monaco Explorations.

Acknowledgements. We thank the Indonesian Institute of Sciences (LIPI) for promoting our collaboration and the Sorong Polytechnic of Marine and Fisheries (Politeknik KP Sorong, West Papua) for providing the vessel *Airaha 02* that we used in this campaign. We thank the crew of the *Aihara 02* for assisting us during the operations and the SPYGEN staff for the technical support in the laboratory.

References

1. Costello MJ, Chaudhary C. 2017 Marine biodiversity, biogeography, deep-sea, and conservation. *Curr. Biol.* **27**, R511–R527. (doi:10.1016/j.cub.2017.04.060)
2. Barlow J *et al.* 2018 The future of hyperdiverse tropical ecosystems. *Nature* **559**, 517–526. (doi:10.1038/s41586-018-0301-1)
3. Lees AC, Pimm SL. 2015 Species, extinct before we know them. *Curr. Biol.* **5**, R177-R180. (doi:10.1016/j.cub.2014.12.017)
4. Díaz S *et al.* 2018 Assessing nature's contributions to people. *Science* **359**, 270–272. (doi:10.1126/science.aap8826)
5. Duffy JE, Godwyn CM, Cardinale BJ. 2017 Biodiversity effects in the wild are common and as strong as key drivers of productivity. *Nature* **549**, 261–264. (doi:10.1038/nature23886)
6. Juhel JB, Vigliola L, Wantiez L, Letessier TB, Meeuwig JJ, Mouillot D. 2019 Isolation and no-entry marine reserves mitigate anthropogenic impacts on grey reef shark behavior. *Sci. Rep.* **9**, 2897. (doi:10.1038/s41598-018-37145-x)
7. Brandl SJ, Goatley CHR, Bellwood DR, Tornabene L. 2018 The hidden half: ecology and evolution of cryptobenthic fishes on coral reefs. *Biol. Rev.* **93**, 1846–1873. (doi:10.1111/brv.124233)
8. Garlapati D, Charankumar B, Ramu K, Madeswaran P, Ramana Murthy MV. 2019 A review on the applications and recent advances in environmental DNA (eDNA) metagenomics. *Rev. Environ. Sci. Bio.* **18**, 389. (doi:10.1007/s11157-019-09501-4)
9. Boussarie G *et al.* 2018 Environmental DNA illuminates the dark diversity of sharks. *Sci. Adv.* **4**, eaap9661. (doi:10.1126/sciadv.aap9661)
10. Fukumoto S, Ushimaru A, Minamoto T. 2015 A basin-scale application of environmental DNA assessment for rare endemic species and closely related exotic species in rivers: a case study of giant salamanders in Japan. *J. Appl. Ecol.* **52**, 358–365. (doi:10.1111/1365-2664.12392)
11. Ruppert KM, Kline RJ, Rahman MDS. 2019 Past, present, and future of environmental DNA (eDNA) metabarcoding: a systematic review in methods, monitoring, and applications of global eDNA. *Glob. Ecol. Conserv.* **17**, e00547. (doi:10.1016/j.gecco.2019.e00547)
12. Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. 2014 Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ* **2**, e593. (doi:10.7717/peerj.593)
13. Cordier T *et al.* 2019 Multi-marker eDNA metabarcoding survey to assess the environmental impact of three offshore gas platforms in the North Adriatic Sea (Italy). *Mar. Environ. Res.* **146**, 24–34. (doi:10.1016/j.marenvres.2018.12.009)
14. Brandl SJ, Rasher DB, Côté IM, Casey JM, Darling ES, Lefcheck JS, Duffy JE. 2019 Coral reef ecosystem functioning: eight core processes and the role of biodiversity. *Front. Ecol. Environ.* **17**, 445–454. (doi:10.1002/fee.2088)
15. Veron JEN, Devantier LM, Turak E, Green AL, Kininmonth S, Stafford-Smith M, Peterson N. 2009 Delineating the coral triangle. *Galaxea, JCRS* **11**, 91–100. (doi:10.3755/galaxea.11.91)
16. Allen GR, Erdmann MV. 2012 *Reef fishes of the East Indies*. Volumes I–III. Perth, Australia: Tropical Reef Research.
17. Kulbicki M *et al.* 2013 Global biogeography of reef fishes: a hierarchical quantitative delineation of regions. *PLoS ONE* **8**, e81847.
18. Exton DA *et al.* 2019 Artisanal fish fences pose broad and unexpected threats to the tropical coastal seascape. *Nat. Commun.* **10**, 2100. (doi:10.1038/s41467-019-10051-0)
19. Jones LA, Mannion PD, Farnsworth A, Valdes PJ, Kelland S-J, Allison PA. 2019 Coupling of palaeontological and neontological reef coral data improves forecasts of biodiversity responses under climatic change. *R. Soc. Open Sci.* **6**, 182111. (doi:10.1098/rsos.182111)
20. Ainsworth CH, Pitcher TJ, Rotinsulu, C. 2008 Evidence of fishery depletions and shifting cognitive baselines in Eastern Indonesia. *Biol. Conserv.* **141**, 848–859. (doi:10.1016/j.biocon.2008.01.006)
21. Valentini A *et al.* 2016 Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Mol. Ecol.* **25**, 929–942. (doi:10.1111/mec.13428)
22. Goldberg CS *et al.* 2016 Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods Ecol. Evol.* **7**, 1299–1307. (doi:10.1111/2041-210X.12595)
23. Pont D *et al.* 2018 Environmental DNA reveals quantitative patterns of fish biodiversity in large rivers despite its downstream transportation. *Sci. Rep.* **8**, 10361. (doi:10.1038/s41598-018-28424-8)
24. Baker W, van den Broek A, Camon E, Hingamp P, Sterk P, Stoesser G, Tuli MA. 2000 The EMBL nucleotide sequence database. *Nucleic Acids Res.* **28**, 19–23. (doi:10.1093/nar/gki098)
25. Ficetola GT, Coissac E, Zundel S, Riaz T, Shehzad W, Bessière J, Taberlet P, Pompanon F. 2010 An *in silico* approach for the evaluation of DNA barcodes. *BMC Genomics* **11**, 434. (doi:10.1186/1471-2164-11-434)
26. Boyer F, Mercier C, Bonin A, Le Bras Y, Taberlet, P, Coissac E. 2016 OBITOOLS: a UNIX-inspired software package for DNA metabarcoding. *Mol. Ecol. Res.* **16**, 176–182. (doi:10.1111/1755-0998.12428)
27. Larkin MA *et al.* 2007 Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948. (doi:10.1093/bioinformatics/btm404)
28. Kearse M *et al.* 2012 Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649. (doi:10.1093/bioinformatics/bts19)
29. Aho K, Derryberry D, Peterson T. 2014 Model selection for ecologists: the worldviews of AIC and BIC. *Ecology* **95**, 631–636. (doi:10.1890/13-1452.1)
30. Deiner K *et al.* 2017 Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Mol. Ecol.* **26**, 5872–5895. (doi:10.1111/mec.14350)
31. Collins RA, Bakker J, Wangensteen OS, Soto AZ, Corrigan L, Sims DW, Genner MJ, Mariani S. 2019 Non-specific amplification compromises environmental DNA metabarcoding with COI. *Methods Ecol. Evol.* **10**, 1985–2001. (doi:10.1111/2041-210X.13276)
32. Wadrop E, Hobbs J-P, Randall JE, DiBattista JD, Rocha LA, Kosaki RK, Berumen ML, Bowen BW. 2016 Phylogeography, population structure and evolution of coral-eating butterflyfishes (Family Chaetodontidae, genus *Chaetodon*, subgenus *Corallochaetodon*). *J. Biogeogr.* **43**, 1116–1129. (doi:10.1111/jbi.12680)
33. Pinheiro HT, Moreau S, Daly M, Rocha LA. 2019 Will DNA barcoding meet taxonomic needs? *Science* **365**, 873–875. (doi:10.1126/science.aay7174)
34. Pawlowski J *et al.* 2018 The future of biotic indices in the ecogenomic era: integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. *Sci. Total Environ.* **637–638**, 1295–1310. (doi:10.1016/j.scitotenv.2018.05.002)
35. Lladó FS, Větrovský T, Baldrian P. 2019 The concept of operational taxonomic units revisited: genomes of bacteria that are regarded as closely related are often highly dissimilar. *Folia Microbiol.* **64**, 19–23. (doi:10.1007/s12223-018-0627-y)
36. 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. *Environ. Sci. Technol.* **51**, 9118–9126. (doi:10.1021/acs.est.7b01518)

37. Wangensteen O, Palacín C, Guardiola M, Turon X. 2018 DNA metabarcoding of littoral hard-bottom communities: high diversity and database gaps revealed by two molecular markers. *PeerJ* **6**, e4705. (doi:10.7717/peerj.4705)
38. Harrison JB, Sunday JM, Rogers SM. 2019 Predicting the fate of eDNA in the environment and implications of studying biodiversity. *Proc. R. Soc. B* **286**, 20191409. (doi:10.1098/rspb.2019.1409)
39. Stat M, Jeffrey J, DiBattista JD, Newman SJ, Bunce M, Harvey ES. 2018 Combined use of eDNA metabarcoding and video surveillance for the assessment of fish biodiversity. *Conserv. Biol.* **33**, 196–205. (doi:10.1111/cobi.13183)
40. Duffy JE, Lelcheck JS, Stuart-Smith RD, Navarrete SA, Edgar GJ. 2016 Biodiversity enhances reef fish biomass and resistance to climate change. *Proc. Natl Acad. Sci. USA* **113**, 6230–6235. (doi:10.1073/pnas.1524465113)
41. Juhel J-B *et al.* 2020 Data from: Accumulation curves of enviornmental DNA sequences predict coastal fish diversity in the coral triangle. Dryad Digital Repository. (doi:10.5061/dryad.t1g1jwt05)

2. Detection of the elusive Dwarf sperm whale (Kogia sima) using environmental DNA at Malpelo island (Eastern Pacific, Colombia)

Manuscrit en révison dans *Ecology and Evolution*

Jean-Baptiste Juhel[1*], Virginie Marques[1,2], Andrea Polanco F.[3], Giomar H. Borrero-Pérez[3], Maria Mutis Martinezguerra[3], Alice Valentini[4], Tony Dejean[4], Stéphanie Manel[2], Nicolas Loiseau[1], Laure Velez[1], Régis Hocdé[1], Tom B. Letessier[5], Eilísh Richards[6], Florine Hadjadj[1], Sandra Bessudo[7], Felipe Ladino[7], Camille Albouy[8], David Mouillot[1], Loïc Pellissier[6, 9]

[1] MARBEC, Univ. Montpellier, CNRS, Ifremer, IRD, Montpellier, France

[2] CEFE, Univ. Montpellier, CNRS, EPHE-PSL University, IRD, Univ Paul Valéry Montpellier 3, Montpellier, France

[3] Instituto de Investigaciones Marinas y Costeras-INVEMAR, Museo de Historia Natural Marina de Colombia (MHNMC), Programa de Biodiversidad y Ecosistemas Marinos, Santa Marta, Colombia

[4] SPYGEN, Le Bourget-du-Lac, France

[5] Institute of Zoology, Zoological Society of London, London, UK

[6] Landscape Ecology, Institute of Terrestrial Ecosystems, Department of Environmental Systems Science, ETH, Universität Zürich, 8092 Zürich, Switzerland

[7] Fundación Malpelo, Bogotá, Colombia

[8] IFREMER, Unité Ecologie et Modèles pour l'Halieutique, EMH, Nantes, France

[9] Unit of Land Change Science, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

[*]Correspondence: Université de Montpellier, 839 Rue du Truel, 34095 Montpellier Cedex 5, France. jeanbaptiste.juhel@gmail.com

KEYWORDS - eDNA, mobile species, megafauna, pelagic

ABSTRACT

1. Monitoring large marine mammals is challenging due to their low abundances in general, an ability to move over large distances and wide geographical range sizes.
2. The distribution of the pygmy (Kogia breviceps) and dwarf (Kogia sima) sperm whales is informed by relatively rare sightings, which does not permit accurate estimates of their distribution ranges. Hence, their conservation status has long remained Data Deficient (DD) in the Red list of the International Union for Conservation of Nature (IUCN), which prevent appropriate conservation measures.
3. Environmental DNA (eDNA) metabarcoding uses DNA traces left by organisms in their environments to detect the presence of targeted taxon, and is here proved to be useful to increase our knowledge on the distribution of rare but emblematic megafauna.
4. Retrieving eDNA from filtered surface water provides the first detection of the Dwarf sperm whale (Kogia sima) around the remote Malpelo island (Colombia).
5. Environmental DNA collected during oceanic missions can generate better knowledge on rare but emblematic animals even in regions that are generally well sampled for other taxa.

INTRODUCTION

Marine mammals are among the most threatened vertebrates on earth with 37% of them being considered as endangered by the IUCN (e.g. Albouy et al., 2020). Yet, the monitoring of marine mammals is still challenging, generally due to their low abundances, their ability to move over large distances, their wide geographical range sizes and their elusive behavior (Hays et al., 2015). Most studies focusing on the distribution of relatively common marine animals rely on telemetry, passive acoustic surveys, or visual observations performed from the coast, during aerial surveys or during boat-based surveys (e.g. Palacios et al., 2012, Balmer et al., 2014, Mannocci et al., 2015). By contrast, the distribution of rare or elusive mammal species are mainly investigated using compilations of scarce observations, fisheries bycatch and strandings (Plön, 2004; Palacios et al., 2012 ; Coombs et al., 2019). As a result, only a limited knowledge has been accumulated on the distribution of those species, which limits our capacity to set effective protection measures (Davidson et al., 2012). Developing complementary and effective tools for detecting and monitoring threatened, rare or elusive marine mammal species is key to better guide their conservation (Pikitch, 2018).

Environmental DNA (eDNA) metabarcoding is increasingly used to detect micro- and macro-organisms in aquatic environments (Ruppert et al., 2019), but more case studies are needed to demonstrate its ability to detect unseen species that are elusive, threatened and rare in marine ecosystems. The eDNA metabarcoding approach is based on retrieving DNA naturally released by organisms in their

environment. This genetic material is then amplified by polymerase chain reaction (PCR), sequenced using high-throughput DNA sequencing systems and assigned to species based on a reference database (Taberlet et al., 2012). Most recent studies confirm the greater detectability of species using eDNA compared with traditional survey approaches in marine environments, especially those with a behavior that impede their direct observation (Simpfendorfer et al., 2016 ; Boussarie et al., 2018 ; Pikitch, 2018). For example, Thomsen et al. (2012) found eDNA to detect more species than nine conventional sampling methods of fish surveys in marine environments. Environmental DNA detection of cetaceans has been validated (Baker et al., 2018; Parsons et al., 2018) and can be used when direct observations are limited. For instance, the long-finned pilot whale (Globicephala melas) was successfully detected in unexpected locations (Foote et al., 2012). The time sensitive nature of eDNA means that its detection is limited to a restricted area from where it was first shed, and can be influenced by environmental factors such as currents and tides (Collins et al., 2018; Harrison et al., 2019).

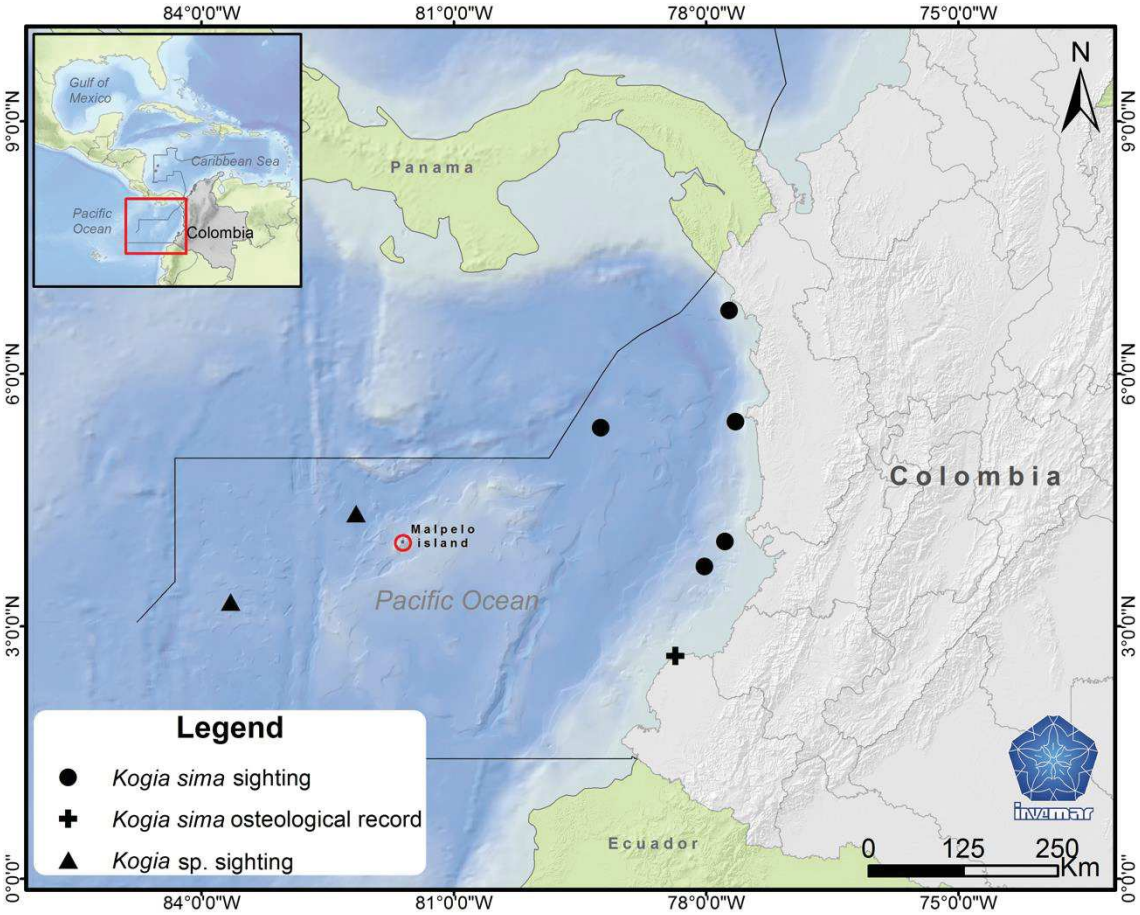


Fig. 1. Map of Dwarf sperm whale (*Kogia sima*) et *Kogia sp.* sightings in the Colombian eastern Pacific.

The pygmy (Kogia breviceps) and dwarf (Kogia sima) sperm whales are porpoise-like shaped odontocetes smaller than 4m (Plön, 2004) that are able to travel long distances (e.g. 255 nautical miles in 4 days, Scott et al., 2001). They occur worldwide in tropical and temperate waters including Colombia (Rice, 1998) and count 1,931 records (1,627 at the species level) of opportunistic sightings and strandings referenced in OBIS (Ocean Biogeographic Information System, www.obis.org, January 2020) and 2, 503 records (2,223 at species level) in GBIF (Global Biodiversity Information Facility, www.gbif.org, e.g. Mora-Pinto et al., 1995). Their relatively scarce sightings prevent an accurate estimation of their distribution ranges and abundances while their conservation status has long remained Data Deficient (DD) in the Redlist of the IUCN. Kogia sima has been sighted only recently in the Colombian Caribbean (Mutis-Martinezguerra and Polanco, 2019) and only six occurrences have been documented in the Colombian Pacific, including five sightings and one stranding (Fig. 1). Two sightings of Kogia sp. have been reported in the vicinity of the Malpelo volcanic island between 1986 and 2006 during line transect surveys, one was near the Island (Wade and Gerrodette, 1993) and the other one was 230 km WSW (Muñoz-Hincapié et al., 1998; Palacios et al., 2012; Fig. 1, Table 1). The Malpelo island (3°58'N, 81°37'W), covering an area of 1.2km², is located 490km off the coast of Colombia, on the top of the submerged Malpelo ridge. This island is composed of barren rocks and steep edges with several underwater habitats including coral formations, vertical rock walls, sands and gravel, tunnels and caves. It is surrounded by deep waters with strong currents where at least nine cetacean species are present (Herrera et al., 2007; Ávila et al., 2013). These deep waters support important populations of large predators and pelagic species including giant grouper, billfish, short-nosed ragged-toothed shark, deepwater sharks and pelagic sharks (Unesco, 2005). Here we document the first detection of the uncommon Dwarf sperm whale (Kogia sima) around the remote Malpelo island (Colombia) using eDNA.

METHODS

During an oceanographic expedition (March 2018, Fig. 2A, B) seawater samples were collected in a 2km radius around the island to investigate the marine vertebrate diversity. A total of 13 non-overlapping 5km-long transects, either rectangular or circular, were performed. During each transect, duplicates of 30L of subsurface seawater (between 0 and 40cm) were simultaneously filtered using two peristaltic pumps placed on each side of the boat (Fig. 2C) and two sterile filter capsules (VigiDNA 0.2µm, SPYGEN). Immediately after, the filters were filled with conservation buffer (CL1 buffer, SPYGEN) and stored in the dark at ambient temperature. A contamination control protocol was carried out at both field and laboratory stages including the use of disposable gloves and single-use filtration equipment (Goldberg et al., 2016; Valentini et al., 2016). The laboratory and equipment were not in contact with cetaceans

or cetacean tissue, before or during the operations, and was cleaned with bleach before each sampling event and before each sample processing.

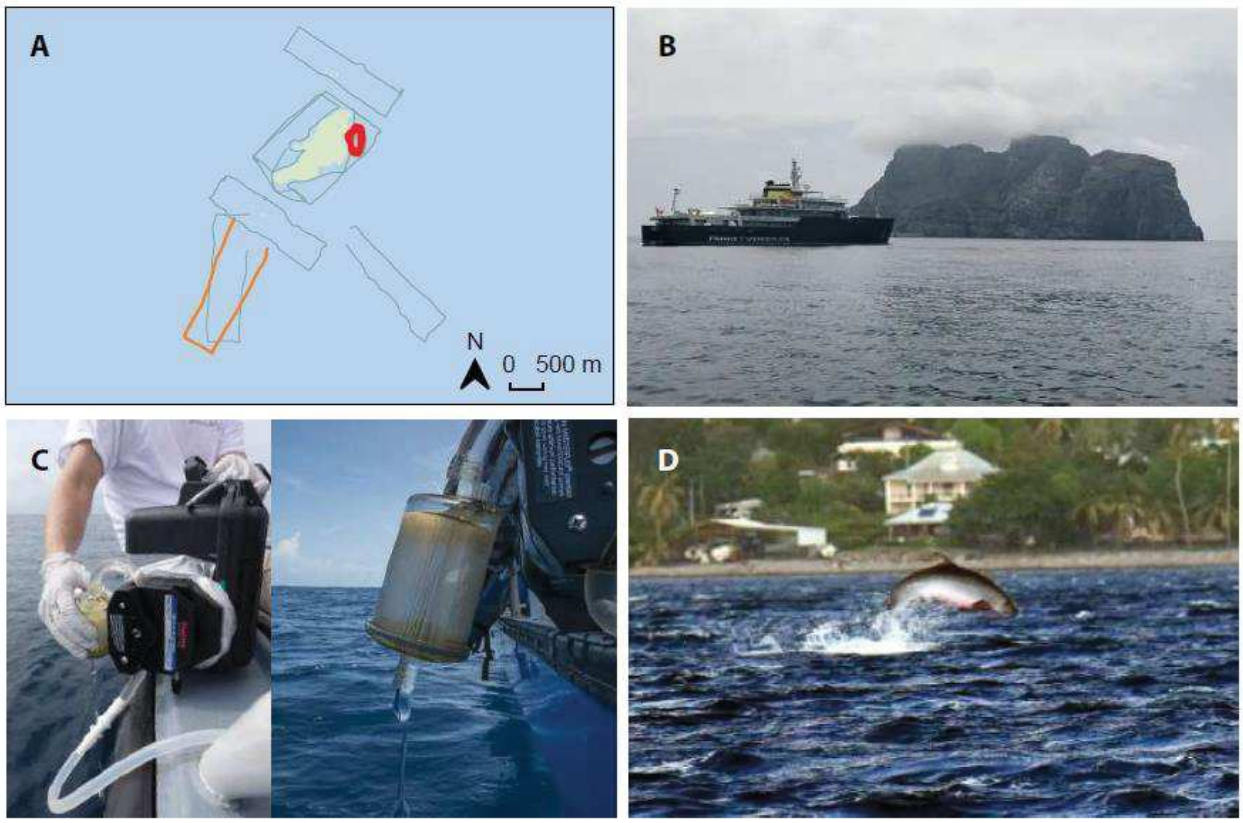


Fig. 2. Location of *Kogia sima* detections around Malpelo Island using environmental DNA (red and orange track) (A); Malpelo Island seascape and our oceanographic vessel (B); eDNA filtration equipment (C) and Opportunistic sighting of *Kogia sp*. around Martinique (French West Indies) (D). *Kogia sima* was detected with both the Vert01 and Mamm01 primer pairs on the circular red track and detected with the Mamm01 primer pair on the rectangular orange track. It was not detected on the grey transects. Credit Photo R. Hocdé, C. Albouy, Megafauna project).

DNA extraction was performed in a dedicated eDNA laboratory equipped with separate clean rooms, positive air pressure, UV treatment and frequent air renewal. Decontamination procedures were conducted before and after all manipulations. Two extractions per filter were performed, following the protocol of Pont et al. (2018), and pooled after the amplification process. Two primer pairs were used for the amplification of metabarcoding sequences, a universal vertebrate 12S mitochondrial rDNA primer pair (Vert01, 5'-TAGAACAGGCTCCTCTAG, 3'-TTAGATACCCCACTATGC) and a mammal 12S mitochondrial rDNA primer pair (Mamm01, 5' -CCGCCCGTCACYCTCCT, 3'-GTAYRCTTACCWTGTTACGAC). Both were used with a human blocking primer pair (5'-CTATGCTTAGCCCTAAACCTCAACAGTTAAATCAACAAAACTGCT -3') (De Barba et al., 2014; Pont et al.,

2018). The amplification primers were 5'-labeled with an eight-nucleotide tag unique to each sample (with at least three differences between any pair of tags), so all 12 PCRs from a single sample shared the same tag. The tags for the forward and reverse primers were identical for each sample. Twelve PCR replicates were run per filter. Three negative extraction controls and two negative PCR controls (ultrapure water) were amplified and sequenced in parallel to monitor possible contaminations. Two libraries were prepared using the MetaFast protocol (Fasteris, www.fasteris.com) and a paired-end sequencing (2x125 bp) was carried out using an Illumina HiSeq 2500 sequencer on two HiSeq Rapid Flow Cell v2 using the HiSeq Rapid SBS Kit v2 (Illumina, San Diego, CA, USA) following the manufacturer's instructions at Fasteris (Geneva, Switzerland). All sequences with a frequency of occurrence below 0.1% per taxon and library were discarded to avoid index cross-talk (MacConaill et al., 2018) and tag-jumps (Schnell et al., 2015). Additionally, sequences with less than 10 reads were removed. The metabarcoding workflow was based on the VSEARCH toolkit and the clustering algorithm SWARM that groups multiple sequence variants into OTUs (Operational Taxonomic Units) to clean errors from PCR and sequencing (Mahé et al., 2014, Rognes et al., 2016). The SWARM clustering algorithm uses single linkage clustering, in which sequence similarity and co-occurrence patterns are used to group sequences together. It allows the removal of erroneous sequences and most reliable detections. Taxonomic assignment was performed using the ecotag program (lower common ancestor algorithm) from the OBITOOLS software package (Boyer et al., 2016) against the global and public EMBL genetic database (European Molecular Biology Laboratory, www.ebi.ac.uk, release 141 downloaded on 11th oct. 2019, Baker et al., 2000). Sequences assigned to common laboratory contaminants such as human, pig or dog were remove from analysis. Sequences were aligned using MUltiple Sequence Comparison by Log-Expectation (MUSCLE) on MEGA software (www.megasoftware.net).

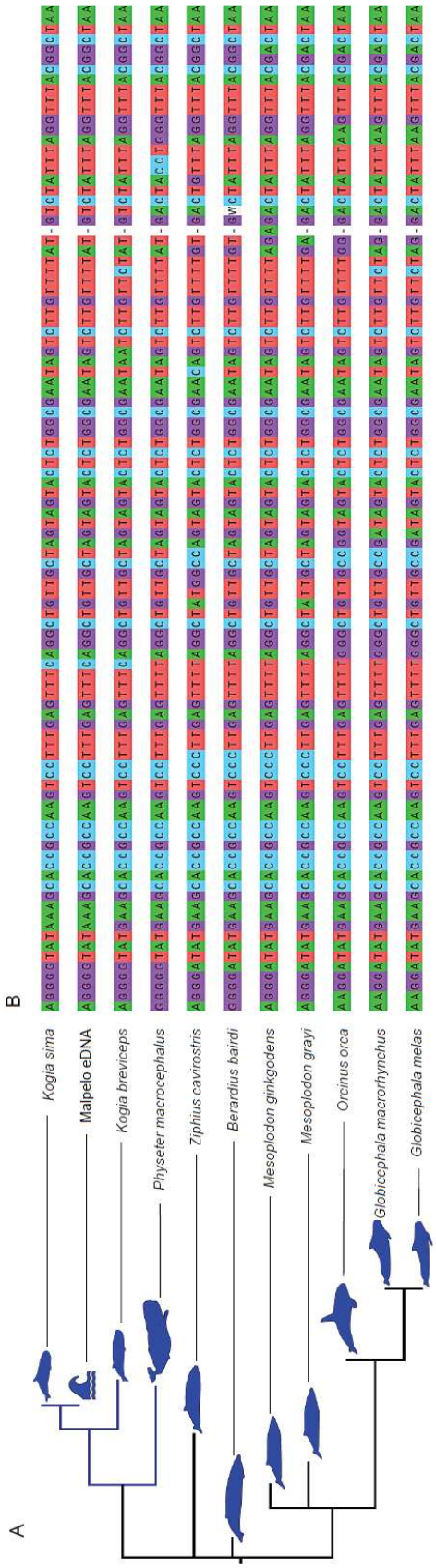


Fig. 3. Phylogenetic tree of the toothed-whales species (A) and their aligned sequences including the sequence from the eDNA sampling (B). The silhouettes were retrieved from phylopic.org (Chris Huh, Creative Commons Attribution-ShareAlike 3.0 Unported).

RESULTS

From the 13 seawater samples, a total of 20,092,190 reads were produced with the vertebrate specific primer pair Vert01 and 4,321,072 reads with the mammal specific primer pair Mamm01. From these reads, 18,007,106 and 2,784,180 passed the bioinformatic cleaning process, respectively. Among the retained reads produced with the Vert01 primer pair, 469 reads corresponding to a unique 99 bp sequence (Fig. 3) matched at 100% similarity with the dwarf sperm whale 12S rDNA (Kogia sima, complete mitochondrial genome, Shan et al., 2019, NC_041303.1), at 97% similarity with the pygmy sperm whale (Kogia breviceps, Arnason et al., 2004, AJ554055.1), while only at <96.4% with other phylogenetically close cetacean species referenced in EMBL (Gatesy et al., 2013; Fig. 3). This sequence was detected on a single transect performed on the 25th of March 2018 at 17 PM (local time UTC -5, Fig. 2A, circular transect). Among the retained reads produced with the Mamm01 primer pair, 3,042 reads corresponding to a unique 63 bp sequence matched at 100% similarity with the same dwarf sperm whale 12S rDNA sequence and at 92.1% similarity with the pygmy sperm whale (Kogia breviceps, Arnason et al., 2004, AJ554055.1) referenced in EMBL. This sequence was detected on two transects performed on the 25th of March 2018 at 17 pm and the 27th of march 2018 at 10:30 AM (local time UTC -5, Fig. 2A circular and rectangular transects) where 1,371 and 1,671 reads were respectively retrieved. Sea surface temperature, measured by the Naval Oceanographic Office (NAVOCEANO) and retrieved from the French Institute for Ocean Science repository (http://www.ifremer.fr/co-argoFloats/float?ptfCode=3901263) was 26.0°C and consistent with the thermal range of the dwarf sperm whale (10°C- 30°C; www.obis.org).

DISCUSSION

Cetaceans include many threatened and difficult-to-study species for which eDNA is expected be a highly effective approach. Despite extensive efforts conducted over the span of 30 years, there are many gaps in the distribution records of those species (Fig. 1). Environmental DNA metabarcoding can provide additional detections without visual observations (Boussarie et al., 2018). The two species K. breviceps and K. sima are very similar and very difficult to separate in the field leaving uncertain identifications in sighting records (Palacios et al., 2012). In contrast, environmental DNA can detect and identify accurately the species, avoiding observer related errors in records.

Table 1. Observation of *Kogia sima* et *Kogia* sp. in the Colombian Pacific. *confusing record assumed to be *Kogia sima* by Wade and Gerrodette, 1993.

Species	Location			Author
	Lat	Long	Geographical reference	
Kogia sp.	3.291776°	-83.653121°	230 km WSW of Malpelo Island	Wade and Gerrodette, 1993; Palacios et al., 2012
Kogia sp.*	4.339118°	-82.159164°	Near Malpelo Island	Wade and Gerrodette, 1993; Palacios et al., 2012
Kogia sima	5.359326°	-79.247856°	225 km W off Cabo Corrientes	Muñoz-Hincapie et al., 1998
Kogia sima	6.753827°	-77.721361°	Near shore Cabo Marzo	Palacios et al., 2012
Kogia sima	5.432542°	-77.647060°	Near shore Cabo Corrientes	Vidal, 1990; Wade and Gerrodette, 1993; Muñoz-Hincapie et al., 1998; Palacios et al., 2012
Kogia sima	4.008406°	-77.770061°	Off Bahía Málaga	Vidal, 1990; Wade and Gerrodette, 1993; Muñoz-Hincapie et al., 1998; Palacios et al., 2012
Kogia sima	3.713174°	-78.013709°	Off Bahía Málaga	Vidal, 1990; Wade and Gerrodette, 1993; Muñoz-Hincapie et al., 1998; Palacios et al., 2012
Kogia sima	2.650000°	-78.360000°	Stranded animal between the communities of La Vigía and Mulatos	Muñoz-Hincapie et al., 1998

These results highlight the promises of eDNA as an alternative to standard monitoring methods for cetaceans, without requiring a close approach of a vessel. For example, Baker et al. (2018) show that eDNA of killer whales has been detected in seawater samples taken up to several hours after their passage and despite marine current circulation. Given its greater sensitivity and the fact that samples can be obtained from a wide variety of platforms (Harrison et al., 2019), eDNA has the potential to rapidly fill data gaps for cetaceans. Studies using this census method are usually limited by the completeness of genetic databases to taxonomically assign the retrieved sequences (Marques et al., 2020). However, strandings of cetaceans along the shores provide a valuable source of genetic material that can be sequenced on eDNA genetic markers to complete reference databases and investigate within species genetic diversity.

Opportunistic detections or targeted sampling in hotspots (e.g. Letessier et al., 2019) are expected to provide valuable new information on the occurrence of uncommon marine vertebrates and better define conservation plans. Malpelo island harbors a wide diversity of marine predators and presents all the characteristics of the last refuges for marine megafauna (Letessier et al., 2019). Thus, it deserves to be a priority for conservation and be placed under appropriate protection from human activities. Marine megafauna plays unique and irreplaceable functional roles in the ocean ecosystem such as the regulation of prey populations, removal of diseased individuals, transport of nutrients between habitats and over vast distances, and protection of blue carbon stocks (Higgs et al., 2014; Atwood et al., 2015; Estes et al., 2016).

Environmental DNA is a method that is easily applicable in the field and can benefit from the thousands of marine sampling operations that can take place regularly around the globe. These novel detections through eDNA will be crucial for Data Deficient species that can include a large proportion of threatened species (Bland et al., 2014; Parson 2016). Building on existing sampling efforts, filling reference database gaps and developing a large-scale observatory network using environmental DNA from water collected in oceanic missions would contribute to a broader knowledge on those rare but emblematic animals.

ACKNOWLEDGEMENTS

We are grateful to the crew of the Yersin and the Malpelo foundation for assisting us during the operations. We thank the National Parks of Columbia for granting us the access to the study area. We thank SPYGEN staff and Véronique Arnal (UMR5175 CEFE, France) for the laboratory analyses. Fieldwork and laboratory activities were supported by Monaco Explorations. Thanks to Janneth Andrea Beltrán (Information Systems Laboratory of INVEMAR, Colombia) for her support in cartography. Contribution No. 1279c of the Instituto de Investigaciones Marinas y Costeras – INVEMAR, Colombia.

REFERENCES

Albouy, C., Delattre, V., Donati, G., Frölicher, T.L., Albouy-Boyer, S., Rufino, M. et al. (2020) Global vulnerability of marine mammals to global warming. Scientific Reports, 10, 1-12. https://doi.org/10.1038/s41598-019-57280-3.

Arnason, U., Gullberg, A. & Janke, A. (2004) Mitogenomic analyses provide new insights into cetacean origin and evolution. Gene, 333: 27-34. DOI: 10.1016/j.gene.2004.02.010.

Atwood, T.B., Connolly, R.M., Ritchie, E.G., Lovelock, C.E., Heithaus, M.R., Hays, G.C., Fourqurean, J.W. & Macreadie, P.I. (2015) Predators help protect carbon stocks in blue carbon ecosystems. Nature Climate Change, 5: 1038-1045. DOI: 10.1038/nclimate2763.

Ávila, I., García, C., Palacios, D. & Caballero, S. Mamíferos acuáticos de la región del Pacífico colombiano. In: Trujillo, F., Gartner, A., Caicedo, D. and Diazgranados, M.C. (Eds.) (2013) Diagnóstico del estado de conocimiento y conservación de los mamíferos acuáticos en Colombia. Ministerio de Ambiente y Desarrollo Sostenible, Fundación Omacha, Conservación Inernacional and WWF. Bogotá, 312 pp.

Baker, C.S., Steel, D., Nieukirk, S. & Klinck, H. (2018) Environmental DNA (eDNA) from the wake of the whales: droplet digital PCR for detection and species identification. Frontiers in Marine Science, 5: 133. DOI: 10.3389/fmars.2018.00133.

Baker, W., van den Broek, Camon, E., Hingamp, P., Sterk, P., Stoesser, G. & Tuli, M.A. (2000) The EMBL nucleotide sequence database. Nucleic Acids Research, 28: 19-23. DOI: 10.1093/nar/gki098.

Balmer, B.C., Wells, R.S., Howle, L.E., Barleycorn, A.A., McLellan, W.A., Pabst, D.A., … Zolman, E.S. (2014) Advances in cetacean telemetry: A review of single-pin transmitter attachment techniques on small cetaceans and development of a new satellite-linked transmitter design. Marine Mammal Science, 30: 656-673. DOI: 10.1111/mms.12072

Bland, L.M., Collen, B., Orme, C.D.L. & Bielby, J. (2014) Predicting the conservation status of data-deficient species. Conservation Biology, 1: 250-259. DOI: 10.1111/cobi.12372.

Boussarie, G., Bakker, J., Wangensteen, O.S., Mariani, S., Bonin, L., Juhel, J.-B., … Mouillot, D. (2018) Environmental DNA illuminates the dark diversity of sharks. Science Advances, 4: eaap9661. DOI: 10.1126/sciadv.aap9661.

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P. & Coissac, E. (2016) OBITOOLS: a UNIX-inspired software package for DNA metabarcoding. Molecular Ecology Resources, 16: 176-182. DOI: 10.1111/1755-0998.12428.

Collins, R.A., Wangesteen, O.S., O'Gorman, E.J., Mariani, S., Sims, D.W. & Genner, J. (2018) Persistence of environmental DNA in marine systems. Communications Biology, 1: 185. DOI: 10.1038/s42003-018-0192-6.

Coombs, E.J., Deaville, R., Sabin, R.C., Allan, L., O'Connell, M., Berrow, S., …, Cooper, N. (2019) What can cetacean stranding records tell us ? Marine Mammal Science, 35: 1527-1555. DOI: 10.1111/mms.12610.

Davidson, A.D., Boyer, A.G., Kim, H., Pompa-Mansilla, S., Hamilton, … Brown, J.H. (2012) Drivers and hotspots of extinction risk in marine mammals. Proceedings of the national academy of sciences, 109: 3395-3400. DOI: 0.1073/pnas.112146910.

De Barba, M., Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E. & Taberlet, P. (2014) DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet. Molecular Ecology Resources, 14: 306-323. DOI: 10.1111/1755-0998.12188.

Estes, J.A., Heithaus, M., McCauley, D.J., Rasher, D.B. & Worm, B. (2016) Megafaunal impacts on structure and function of ocean ecosystems. Annual Review of Environment and Resources, 41: 83-116. DOI: 10.1146/annurev-environ-110615-085622.

Foote, A.D., Thomsen, P.F., Sveegaard, S., Wahlberg, M., Kielgast, J., … Gilbert, M.T.P. (2012) Investigating the potential use of environmental DNA (eDNA) for genetic monitoring of marine mammals. PLoS ONE, 7: e41781. DOI: 10.1371/journal.pone.0041781.

Gatesy, J., Geisler, J.H., Chang, J., Buell, C., Berta, A., Meredith, R.W., Springer, M.S. & McGowen, M.R. (2013) A phylogenetic blueprint for a modern whale. Molecular Phylogenetics and Evolution, 66: 479-506. DOI: 10.1016/j.ympev.2012.10.012.

Goldberg, C.S., Turner, C.R., Deiner, K., Klymus, K.E., Thomsen, P.F., Murphy, M.A., … Taberlet, P. (2016) Critical considerations for the application of environmental DNA methods to detect aquatic species. Methods in Ecology and Evolution, 7: 1299-1307. DOI: 10.1111/2041-210X.12595.

Harrison, J.B., Sunday, J.M. & Rogers, S.M. (2019) Predicting the fate of the eDNA in the environment and implications for studying biodiversity. Proceedings of the Royal Society B, 286: 20191409. DOI: 10.1098/rspb.2019.1409.

Hays, G.C., Ferreira, L.C., Sequeira, A.M., Meekan, M.G., Duarte, C.M., Bailey, H., … Thums, M. (2016) Key questions in marine megafauna movement ecology. Trends in Ecology and Evolution, 31: 463–475. DOI: 10.1016/j.tree.2016.02.015.

Herrera, J., Ávila, I., Falk, P., Soler, G., García, C., Tobón, I. & J. Capella (2007) Los mammiferos marinos en el santuario de fauna y flora Malpelo y aguas hacia el continente, Pacífico Colombiano. In: DIMARCCCP y UAESPNN. Santuario de Fauna y Flora Malpelo: descubrimiento en marcha, (Ed.) DIMAR, Bogotá, pp. 113-130.

Higgs, N.D., Gates, A.R. & Jones, D.O. (2014) Fish food in the deep sea: Revisiting the role of large food-falls. PLoS ONE, 9: e96016. DOI: 10.1371/journal.pone.0096016.

Letessier, T.B., Mouillot, D., Bouchet, P.J., Vigliola, L., Fernandes, M.C., Thompson, C., … Meeuwig, J.J. (2019) Remote reefs and seamounts are the last refuges for marine predators across the Indo-Pacific. PLoS Biology, 17: e3000366. DOI: 10.1371/journal.pbio.3000366.

MacConaill, L.E., Burns, R.T., Nag, A., Coleman, H.A., Slevin, M.K., Giorda, K., … Thorner A.R. (2018) Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. BMC Genomics, 19, 30. DOI:10.1186/s12864-017-4428-5.

Mahé, F., Rognes, T., Quince, C., de Vargas, C. & Dunthorn, M. (2014) Swarm: robust and fast clustering method for amplicon-based studies. PeerJ, 2: e593. DOI: 10.7717/peerj.593.

Mannocci, L., Monestiez, P., Spitz, J. & Ridoux, V. (2015) Extrapolating cetacean densities beyond surveyed regions: habitat-based predictions in the circumtropical belt. Journal of Biogeography, 42: 1267-1280. DOI: 10.1111/jbi.12530.

Marques, V., Milhau, T., Albouy, C., Dejean, T., Manel, S., Mouillot, D. & Juhel, J.B. (2020) GAPeDNA: Assessing and mapping global species gaps in genetic databases for eDNA metabarcoding. Diversity and Distribution, in press. DOI: 10.1111/ddi.13142.

Mora-Pinto, D. M., Muñoz-Hincapié, M.F., Mignucci-Giannoni, A.A. & Acero-Pizarro, A. (1995) Marine mammal mortality and strandings along the Pacific Coast of Colombia. Reports of the International Whaling Commission, 45: 427-429.

Muñoz-Hincapié, M.F., Mora-Pinto, D.M., Palacios, D.M., Secchi, E.R. & Mignucci-Giannoni, A.A. (1998) First osteological record of the dwarf sperm whale in Colombia, with notes on the zoogeography of Kogia in South America. Revista de la Academia Colombiana de Ciencias, 22: 433-444.

Mutis-Martinezguerra, M.A. & Polanco F., A. (2019) First stranding record of Kogia sima (Owen, 1866) in the Colombian Caribbean. Latin American Journal of Aquatic Mammals, 14: 18-26. DOI: 10.5597/lajam00250.

Palacios, D.M., Herrera, J.C., Gerrodette, T., Garcia, C., Soler, G.A., Avila, I.C., … Kerr, I. (2012) Cetacean distribution and relative abundance in Columbia's Pacific EEZ from survey cruises and platforms of opportunity. Journal of Cetacean Research and Management, 12: 45-60.

Parsons, E.C.M. (2016) Why IUCN should replace “Data Deficient” conservation statut with a precautionary “Assume Threatened” status – A cetacean case study. Frontiers in Marine Science, 3:193. DOI: 10.3389/fmars.2016.00193.

Parsons, K.M., Everett, M., Dahlheim, M. & Park, L. (2018) Water, water everywhere: environmental DNA can unlock population structure in elusive marine species. Royal Society open science, 5: 180537. DOI: 10.1098/rsos.180537.

Pikitch, E.K. (2018) A tool for finding rare marine species. Science, 360: 1180-1182. DOI: 10.1126/science.aao3787.

Plön, S. (2004) The status and natural history of pygmy (Kogia breviceps) and dwarf (K. sima) sperm whales off Southern Africa. PhD thesis. Rhodes University, Grahamstown.

Pont, D., Rocle, M., Valentini, A., Civade, R., Jean, P., Maire, A., … Dejean, T. (2018) Environmental DNA reveals quantitative patterns of fish biodiversity in large rivers despite its downstream transportation. Scientific Reports, 8: 10361. DOI: 10.1038/s41598-018-28424-8.

Rice, D.W. (1998) Marine mammals of the world: systematics and distribution. Society for Marine Mammalogy, Special Publication Number 4. Lawrence, KS. USA: Editor Wartzok D.

Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ, 4: e2584; DOI: 10.7717/peerj.2584.

Ruppert, K.M., Kline, R.J. & Rahman, Md.S. (2019) Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. Global Ecology and Conservation, 17: e00547. DOI: 10.1016/j.gecco.2019.e00547.

Schnell, I.B., Bohmann, K. & Gilbert, T.P. (2015) Tag jumps illuminated – reducing sequence-to-sample misidentifications in metabarcoding studies. Molecular Ecology Resources, 15: 1289–1303. DOI: 10.1111/1755-0998.12402.

Scott, M.D., Hohn, A.A., Westgate, A.J., Nicolas, J.R., Whitaker, B.R. & Campbell, A. (2001) A note on the release and tracking of a rehabilitated pygmy sperm whale (Kogia breviceps). Journal of Cetacean Resources and Management, 3: 87-94.

Shan, L., Tian, R. & Liu, Y. (2019) The complete mitochondrial genome of the dwarf sperm whale Kogia sima (Cetacea: Kogiidae). Mitochondrial DNA part B, 4: 72-73. DOI: 10.1080/23802359.2018.1536464.

Simpfendorfer, C.A., Kyne, P.M., Noble, T.H., Goldsbury, J., Basiita, R.K., Lindsay, R., … Jerry, D.R. (2016) Environmental DNA detects Critically Endangered largetooth sawfish in the wild. Endangered Species Research, 30: 109-116. DOI: 10.3354/esr00731.

Taberlet, P., Coissac, E., Hajibabaei, M. & Rieseberg, L.H. (2012) Environmental DNA. Molecular Ecology, 21: 1789-1793. DOI: 10.1111/j.1365-294X.2012.05542.x.

Thomsen, P.F., Kielgast, J., Iversen, L.L., Møller, P.R., Rasmussen, M., Willerslev E. (2012) Detection of a diverse marine fish fauna using environmental DNA from seawater samples. PLoS ONE, 7: e41732. doi: 10.1371/journal.pone.0041.

UNESCO (2005) Gorgona and Malpelo Islands, Coastal and Oceanic National Marine Parks of Colombia's Eastern Tropical Pacific. Nomination Document World Heritage Site, 1- 84.

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P.F., … Dejean, T. (2016) Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. Molecular Ecology, 25, 929–942. DOI: 10.1111/mec.13428.

Vidal, O. (1990) Lista de los mamíferos acuáticos de Colombia. Informe del Museo del Mar, Bogotá, Colombia, 37, 1-18.

Wade, P.R. and Gerrodette, T. (1993) Estimates of cetacean abundance and distribution in the eastern tropical Pacific. Report of the International Whaling Commission, 43: 477-493.

3. Recovering aquatic and terrestrial biodiversity in a tropical estuary using environmental DNA

Manuscrit en révison dans *Biotropica*

Andrea Polanco F.[1], Maria Mutis Martinezguerra [1], Virginie Marques[2,6], Francisco Villa-Navarro[3], Giomar Helena Borrero[1], Marie-Charlotte Cheutin[2], Tony Dejean[4], Régis Hocdé[2], Jean-Baptiste Juhel[2], Eva Maire[2,10], Stéphanie Manel[2,6], Manuel Spescha[8], Alice Valentini[4], David Mouillot[2], Camille Albouy[1,7], Loïc Pellissier[1,8,9]

[1] Instituto de Investigaciones Marinas y Costeras-INVEMAR, Colombia. Museo de Historia Natural Marina de Colombia (MHNMC), Programa de Biodiversidad y Ecosistemas Marinos. Calle 25 No. 2 – 55 Playa Salguero, Santa Marta, Colombia.

[2] MARBEC, Univ. Montpellier, CNRS, Ifremer, IRD, Montpellier, France

[3] Grupo de Investigación en Zoología, Facultad de Ciencias, Universidad del Tolima, Barrio Santa Helena, Ibagué, Tolima, Colombia.

[4] SPYGEN, Le Bourget-du-Lac, France

[6] CEFE, Univ Montpellier, CNRS, EPHE-PSL University, IRD, Univ Paul Valéry Montpellier 3, Montpellier, France

[7] IFREMER, unité Écologie et Modèles pour l'Halieutique, rue de l'Ile d'Yeu, BP21105, 44311 Nantes cedex 3, France.

[8] Landscape Ecology, Institute of Terrestrial Ecosystems, Department of Environmental Systems Science, ETH Zürich, Zürich, Switzerland

[9] Unit of Land Change Science, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

[10] Lancaster Environment Centre, Lancaster University, Lancaster, UK

[1] Shared senior authorship

Abstract

Biodiversity is declining globally as a result of combined direct human exploitation and climate change. Monitoring of animal biodiversity conventionally relies on taxonomic identification from direct observations or trapping, which represents a challenging effort in very species-rich tropical ecosystems. Estuaries are characterized by a tidal regime and are strongly influenced by hydrodynamics and sedimentary flows. They host much contrasted and highly dynamic habitats, from freshwater, brackish or saltwater to terrestrial, which are especially difficult to monitor. Here, we investigated the potential of environmental DNA (eDNA) metabarcoding, with three different primers targeting different regions of the mitochondrial DNA 12S ribosomal RNA gene, to detect vertebrate diversity in the estuary of the Don Diego River in the Sierra Nevada de Santa Marta in Colombia. We show that eDNA recovers not only aquatic organisms including fishes, amphibians and reptiles, but also a large diversity of terrestrial, arboricole and flying vertebrates including mammals and birds living in the estuary surroundings. We further show that the eDNA signal remains relatively localized along the watercourse. A transect from the deep outer section of the estuary, across the river mouth toward the inner section of the river, showed a marked taxonomic turnover from typical marine to freshwater fishes, while eDNA of terrestrial and arboricolous species was mainly found in the inner section of the estuary. Together, our results indicate that eDNA allows detecting a large diversity of vertebrates and can become an important tool for biodiversity monitoring in estuaries, where water integrates information across the entire ecosystem.

Keywords: Biodiversity, Caribbean Sea, environmental DNA, biomonitoring, Vertebrates, tropical ecosystems.

1. INTRODUCTION

Biodiversity is declining globally due to a combination of global changes including human exploitation and climate warming (Díaz et al., 2019). Monitoring species diversity and composition in space and time is the cornerstone to document biodiversity erosion and identify where conservation measures must be applied (e.g. Dixon et al., 2019; Blowes et al., 2019). Monitoring plant and animal diversity is conventionally based on visual methods. Besides being invasive, expensive, and highly dependent on a declining taxonomic expertise which are generally in decline (Paknia et al., 2015), conventional biodiversity surveys have shortcomings in the detection of discrete, elusive or cryptic species. A shortage of taxonomic skills, and time consuming monitoring programs cause limited biodiversity information for conservationists (Mace, 2004), while a solid scientific documentation of biodiversity loss is necessary to trigger conservation? action. The problem is accentuated in lower-income countries, which are paradoxically usually rich in biodiversity (Collen et al., 2008; Barlow et al., 2018). In tropical ecosystems, the complex structure and diversity of ecosystems are often summarized through a few indicator species, which might provide only a partial ecosystem health assessment (Hilty & Merenlender, 2000; Müller & Geist, 2016). The limited systematic biodiversity sampling and resampling time series precludes the detection of phase shifts (Folke et al., 2004), so that usually slow biodiversity erosion can remain detected (Potapov et al., 2017). Information gaps on negative biodiversity trends limit support in favour of appropriate action to prevent further declines (Dornelas et al., 2013). We thus need to reinforce our capacity of monitoring long term changes in species diversity and composition in complex tropical ecosystems (Schmeller et al., 2017; Barlow et al., 2018).

Environmental DNA (eDNA) metabarcoding can retrieve and sequence species DNA from the environment (water, soil, sediment) and does not require any visual observation of the target species. Monitoring a wide array of organisms (Deiner et al., 2017; Taberlet et al., 2012) with a single method could provide a simplified ecosystem-wide quantification of biodiversity. Species leave DNA footprints via faeces, urine and epidermal cells in the environment detectable during a limited period of time in aquatic ecosystems (Dejean et al., 2011) which, after amplification and sequencing, can be processed into species composition information (Deiner et al., 2017). The biodiversity signal retrieved from an eDNA sample could be trans-kingdom (Stat et al., 2017), as multiple primers can be developed specifically to target taxonomic groups of interest, from microorganisms (Gilbert et al., 2012; Pellissier et al., 2014) to very large vertebrates (Boussarie et al., 2018; Stat et al. 2017; Djurhuus et al., 2020). Combined with high-throughput sequencing, environmental DNA metabarcoding allows large-scale and multi-taxa surveys from material that is fast to collect in the field. Hence, eDNA method is non-invasive, demonstrates higher detection capabilities and cost-effectiveness compared to traditional methods for

environmental monitoring and biodiversity assessments (e.g. Dejean et al., 2012; Valentini et al., 2016; Taberlet et al., 2018). Recent aquatic applications demonstrate its potential to assess freshwater (e.g. Valentini et al., 2016; Pont et al., 2018) and marine species composition (e.g. West et al., 2020; Polanco et al., in review), indicating that filtering water might be particularly efficient to monitor animal eDNA. Moreover, water can transport eDNA from both aquatic and terrestrial organisms and thus integrate information across several ecosystems (Deiner et al., 2017). For example, Sales et al. (2020) compared eDNA with camera traps and found that terrestrial mammals recorded with cameras were also detected in eDNA. The signal of terrestrial mammals may be weaker and less stable than for aquatic organisms (Harper et al., 2019), but remains sufficient for ecosystem monitoring. Water eDNA metabarcoding could allow large-scale, multi-species monitoring of an entire ecosystem, especially those that are difficult to sample using traditional methods (Beng & Corlett, 2020).

Ecotones represent an interface between multiple contiguous habitats, where neighboring species occupancy generates high levels of biodiversity (Smith et al., 1997). These complex habitats are particularly difficult to monitor using traditional methods, but need to receive particular attention due to their important ecosystem functions (Basset et al., 2013). Estuaries are critical transition zones between land, wetlands, freshwater habitats, and the sea and host a very diverse biodiversity of both terrestrial and aquatic species (Lachavanne et al., 1997; Levin et al., 2001). Estuaries contain a variety of both immerged and emerged habitats with clines in salinity associated with sharp species compositional turnover (Reizopoulou et al., 2014). Assessing the status of biodiversity in such a dynamic environment is difficult because each habitat generally requires different types of taxonomic sampling or indicator organisms. Moreover, sampling aquatic organisms in brackish water using traditional methods can be especially challenging due to low visibility. Hence, eDNA metabarcoding could be particularly efficient to measure biodiversity in those interface aquatic systems and even more if it integrates the detection of both aquatic and terrestrial organisms (Sales et al., 2020). Estuaries also serve as vital nurseries for many marine species. Amphihaline or migratory species pass through estuaries (e.g., Beck et al., 2001). Estuaries attract terrestrial animals for a variety of reasons, including feeding or drinking water (Greenberg, 2012) and are critical transition zones of water fluxes from terrestrial to aquatic ecosystems (Wall et al., 2001). As a result of the direct animal contact with water or the indirect fluxes of water, the terrestrial animal DNA could be transferred in the water and the signal of their presence should be recovered using eDNA (Harper et al., 2019). So far, there is only a limited number of studies that investigated the signal of terrestrial vertebrates in aquatic environments (Leempoel et al., 2020, Sales et al., 2020) and few have investigated an entire estuary (but see Stoeckle et al., 2017), including the signal of both terrestrial and aquatic animals.

Here, we assessed the biodiversity in the estuary of the Don Diego River and its adjacent marine waters in the Natural National Park Sierra Nevada de Santa Marta in Colombia using eDNA metabarcoding. The terrestrial habitat along the Don Diego river is a protected area expected to host many vertebrate species (Jiménez-Alvarado et al., 2015) including mammals (Alberico et al., 2000; Torné-Salas, 2013; Pineda-Guerrero et al., 2015), birds (Strewe & Navarro, 2003; 2004), reptiles and amphibians (Ruthven & Carriker, 1922; Pérez-Gonzáles et al., 2016), while the river should contain a large set of freshwater species including some endemic ones, as sampled in previous visual monitoring (Villa-Navarro et al., 2016). In contrast, the marine habitats and species composition in front of the Don Diego River are less known, due to turbid waters and open coast, which limits the knowledge to fisheries catches (Rueda et al., 2011). eDNA metabarcoding can reveal the biodiversity of a wide range of vertebrate clades using either general or clade specific primers, and is becoming an established tool to monitor diversity for conservation (Bohmann et al., 2014). In this study, we investigated the capacity of eDNA metabarcoding applied on the freshwater and marine waters to provide an integrative measure of estuarine biodiversity using three primers targeting all vertebrates, bony fishes and chondrichthyes. We collected and filtered 18 water samples of 30L along a transect from the marine part of the estuary, the shallow brackish waters of the river mouth, and into the freshwater environments surrounded by tropical dry forests. We asked the following questions:

1) Does a multimarker eDNA metabarcoding survey discriminate biodiversity (taxa composition) between connected, but ecologically dissimilar habitats across a tropical estuary?

2) Does eDNA metabarcoding applied to aquatic samples not only detect aquatic species, but also further integrate the signal of terrestrial and arboricolous species surrounding the river?

3) Does eDNA metabarcoding highlight higher connectivity between pairs of habitat, especially between downstream and upstream or between marine and brackish environments?

We further used the results to evaluate the eDNA capacity to detect the presence of endemic species in the Natural National Park Sierra Nevada de Santa Marta. Together, we appraise the capacity of different primers to recover the biodiversity using eDNA in estuaries, which could provide a much-needed approach to monitor species in these highly dynamic and rich ecosystems.

2. METHODS

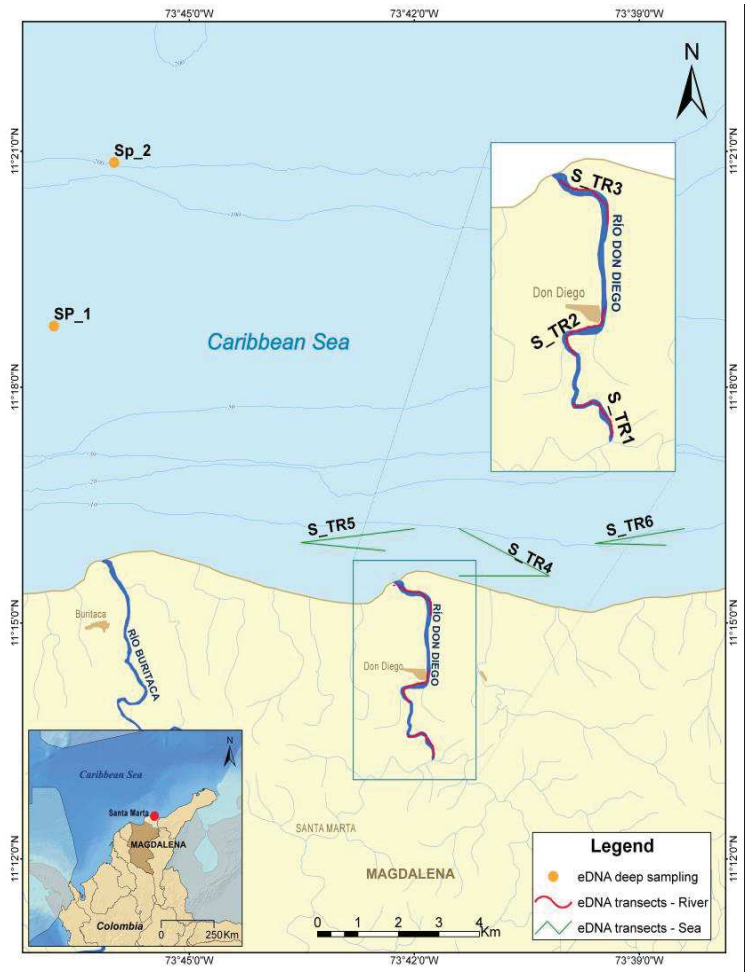


Figure 1. Maps of the sampled sites. 1) the marine surface sampling, in green, corresponding to the eDNA transects performed in three different areas in front of the river mouth, 2) the marine deep-water sampling, in orange, corresponding to the eDNA sampled with Niskin bottles at three different depths in each site and 3) the freshwater sampling in red, corresponding to the eDNA transects performed in three different areas of the Don Diego river.

2.1. Study area

The Don Diego River is one of the 18 basins over the northern flank of the Sierra Nevada de Santa Marta (SNSM) that flow into the Caribbean Sea (Figure 1). The SNSM (5775 m) is the highest coastal mountain in the world, located in the North of Colombia, Atlantic Coast (10° 10' and 10° 20' N and 72° 30'and 74° 15' W) and declared a biosphere reserve by UNESCO. It is a hotspot of biodiversity due to its geographic isolation and the climatic conditions of its recent geological past that have favored a surprising diversity of fauna and flora and the development of a high level of endemism (Almeda et al., 2013; Roach et al., 2020). The SNSM is composed of 34 main hydrographic basins supplying the populations of three main departments in northern Colombia (Magdalena, Cesar and La Guajira). The rainfall regime is largely

defined by the movement of the Intertropical Convergence Zone (ZCIT), which determines two rainy periods, from April to June and from August to November, alternated by two dry seasons from December to March and from June to August, the latter known as the “veranillo de San Juan” (Villa-Navarro et al., 2016). In the Don Diego River, the increase in flows are progressive from April, with a maximum in November, and then decline in December (INGEOMINAS et al., 2008). The river meets with the sea in a dynamic river mouth that depends on the river water regime influenced by climatic conditions and leading to a high energy open shore entering a plain of sandy bottoms in the sea. The river mouth of this bar-built estuary varies according to the sand spit. Due to its ecosystem heterogeneity and its strategic location in the foothills of the mountain and owing to the critical transition zone between the land and marine environments, the estuarine area of the Don Diego is expected to represent a biodiversity hotspot.

2.2. Field sampling

We collected a total of 18 samples from 8 different sites (Figure 1, Table S1), each of a large water volume, with two filtration replicates per site from October 16th to October 18th, 2018. Using a boat, we sampled water in (i) two sites at three different depths to comprise different layers of the water column, farthest from the coast using niskin bottles, (ii) in three sites of surface water in the marine environment close to the river mouth and (iii) in three sites inside the river in the freshwater environment (Figure 1). For the surface water, eDNA sampling was performed using a filtration device composed of a Athena® peristaltic pump (Proactive Environmental Products LLC, Bradenton, Florida, USA; nominal flow of 1.0 L.min−1), a VigiDNA® 0.2 µM cross flow filtration capsule (SPYGEN, le Bourget du Lac, France) and disposable sterile tubing for each filtration capsule. For the three freshwater sites (S_TR1, S_TR2, S_TR3; Figure 1) we used a VigiDNA® 0.45 µM cross flow filtration capsule (SPYGEN, le Bourget du Lac, France) to limit the risks of clogging. Two filtration replicates were performed in parallel on each side of a small boat, at each station, during approximately 30 minutes corresponding to a water volume of approximately 30 L of water. At the end of each filtration, the water inside the capsules was emptied, and the capsules were filled with 80 mL of CL1 Conservation buffer (SPYGEN, le Bourget du Lac, France) and stored at room temperature. For the two deeper water sites, we collected 10 L of water using a disinfected niskin bottle at three different layers of the water column: surface, mid-depth and deep as follow:0, 35, and 53 m for the sampling point S_P1 with an estimated depth of 70 m and 0, 58, 115 m for the sampling point S_P2 with an estimated depth of 150 m. We transferred the water from the niskin bottle into a sterilized bag placed in a container and filtered it with the same protocol as surface marine waters. We followed a strict contamination control protocol in both field and laboratory

stages (Goldberg et al., 2016; Valentini et al., 2016). Each water sample processing included the use of disposable gloves and single-use filtration equipment.

2.3. DNA extraction, amplification and high-throughput sequencing

The DNA extraction, amplification and sequencing were performed in separate dedicated rooms, equipped with positive air pressure, UV treatment and frequent air renewal. Two extractions per filter were performed following the protocol of Pont et al. (2018), the two DNA samples per filtration capsule were pooled together before the amplification step. After the DNA extraction the samples were tested for inhibition following the protocol described in Biggs et al. (2015). If the sample was considered inhibited it was diluted 5-fold before the amplification. We used three different primer sets, targeting chondrichthyans (Chon01), teleosteans (teleo/Tele01) and all vertebrates (Vert01) (Valentini et al., 2016; Taberlet et al., 2018). DNA amplifications were performed in a final volume of 25 µL, using 3 µL of DNA extract as the template. The amplification mixture contained 1 U of AmpliTaq Gold DNA Polymerase (Applied Biosystems, Foster City, CA), 10 mM Tris-HCl, 50 mM KCl, 2.5 mM MgCl2, 0.2 mM each dNTP, 0.2 µM of each primers, 4 µM human blocking primer (for the teleo and Chond01 primers Civade et al., 2016; for Vert01 De Barba et al., 2014) and 0.2 µg/µL bovine serum albumin (BSA, Roche Diagnostic, Basel, Switzerland).

The three primer pairs were 5'-labeled with an eight-nucleotide tag unique to each PCR replicate for teleo and unique to each sample for the other two primer pairs (with at least three differences between any pair of tags), allowing the assignment of each sequence to the corresponding sample during sequence analysis. The tags for the forward and reverse primers were identical. The PCR mixture was denatured at 95°C for 10 min, followed by 50 cycles of 30 s at 95°C, 30 s at 55°C for teleo and Vert01 and 58°C for Chon01 and 1 min at 72 °C and a final elongation step at 72°C for 7 min. Twelve replicates of PCRs were run per filtration for each primer pair. After amplification, the samples were titrated using capillary electrophoresis (QIAxcel; Qiagen GmbH) and purified using the MinElute PCR purification kit (Qiagen GmbH). Before sequencing, purified DNA was titrated again using capillary electrophoresis. The purified PCR products were pooled in equal volumes to achieve a theoretical sequencing depth of 1,000,000 reads per sample. Three libraries were prepared using the MetaFast protocol (Fasteris). Out of the 3 libraries prepared, 2 of them were sequenced using Illumina HiSeq and 1 using Illumina MiSeq. The Hiseq for two libraries a paired-end sequencing (2x125 bp) was carried out using an Illumina HiSeq 2500 sequencer using a Rapidrun mode on a HiSeq Rapid Flow Cell v2 using the HiSeq Rapid SBS Kit v2 (Illumina, San Diego, CA, USA) and on a MiSeq (2x125 bp, Illumina, San Diego, CA, USA) and the MiSeq Flow Cell Kit Version3 (Illumina, San Diego, CA, USA) were used following the manufacturer's instructions. Library preparation and sequencing were performed at Fasteris (Geneva, Switzerland).

Four negative extraction controls and two negative PCR controls (ultrapure water, 12 replicates) were amplified per primer pair and sequenced in parallel to the samples to monitor possible contaminants.

2.4. Obitools filtering, taxonomic assignments

Following the sequencing, reads were processed to remove errors and analyzed using programs implemented in the OBITools package (http://metabarcoding.org/obitools, Boyer et al., 2016) following a previous protocol (Valentini et al., 2016). The forward and reverse reads were assembled using the ILLUMINAPAIREDEND program using a minimum score of 40 and retrieving only joined sequences. The reads were then assigned to each sample using the NGSFILTER software. A separate dataset was created for each sample by splitting the original data set into several files using OBISPLIT. After this step, each sample was analyzed individually before merging the taxon list for the final ecological analysis. Strictly identical sequences were clustered together using OBIUNIQ. Sequences shorter than 20 bp, or with less than 10 reads were excluded using the OBIGREP program. The OBICLEAN program was then run within a PCR product. All sequences labelled 'internal' that correspond most likely to PCR substitutions and indel errors were discarded. Taxonomic assignment of the remaining sequences was performed using the program ECOTAG with the NCBI reference sequence (www.ncbi.nlm.nih.gov, release 233, downloaded on 11th oct. 2019). Taxonomic assignments were corrected as follows to be more conservative: for an identification match > 99% identity, we assigned at the species level, for a 90-99% match, genus level if available and for an 85-90% match, family level if possible. Considering the wrong assignment of a few sequences to the sample due to tag-jumps (Schnell et al., 2015), all sequences with a frequency of occurrence below 0.001 per taxon and per library were discarded. We further corrected for index-hoping (MacConaill et al., 2018) with a threshold empirically determined per sequencing batch using experimental blanks (i.e. combinations of tags not present in the libraries), for a given sequencing batch between libraries.

2.5. Comparison of eDNA species identification to local faunal lists

A current limitation of the eDNA approach lies in the incompleteness of the genetic reference database, which can bias taxonomic assignments. In order to provide additional constraints on taxonomic assignments from the eDNA, we validated the recovered eDNA taxonomic assignments with the known lists of the regional species pools. In particular for fishes, we used Robertson and Van Tassel (2019), Villa-Navarro et al. (2016) and unpublished personal databases of one of the authors (FV-N). For mammals, we used the Mammal Species of the World Checklist data set (National Museum of Natural History, Smithsonian Institution, 2020), the ASM Mammal Diversity Database (Mammal Diversity Database, 2020), and for specific distribution of the species we used Alberico et al. (2000), Torné Salas

(2013) and Pineda-Guerrero et al. (2015). For birds, we used Strewe and Navarro (2003, 2004), Ayerbe-Quiñones (2018), Verhelst-Montenegro and Salaman (2019) and Clements et al. (2019), for amphibians and reptiles we used Ruthven and Carriker (1922), Pèrez-Gonzales et al. (2016). We matched regional lists with eDNA records, and checked whether the species, genus or family found in eDNA are known in the area. This was carried out for all three 12S primers targeting Vertebrates, bony fishes and chondrichthyans respectively. We discarded taxonomic identifications that are not present in the Caribbean Sea or the surrounding continental waters. Genera or species identified from other regions have been considered in their immediately higher taxa if they exist in the area (i.e. the genera Argyrosomus restricted to the Tropical Eastern Pacific, were considered as the detection of the Sciaenidae family).

2.6. eDNA clustering using Swarm

We applied a second bioinformatic workflow on the teleo marker only using sequence clustering to provide realistic diversity estimation in the absence of a complete reference database (Marques et al., 2020). The ability of the pipeline to provide realistic estimates has only been assessed using the teleo marker, so other markers (Vert01 and Chond01) were not processed using clustering. We used the clustering algorithm SWARM, which uses sequence similarity and abundance patterns to cluster multiple variants of sequences into MOTU (Molecular Operational Taxonomic Units; Mahé et al., 2014; Rognes et al., 2016). First, sequences were merged using vsearch (Rognes et al., 2016), we used CUTADAPT (Martin, 2013) for demultiplexing and primer trimming and vsearch to remove sequences containing ambiguities. SWARM was then run with a minimum distance of one mismatch to make clusters. Once the MOTUs are generated, the most abundant sequence within each cluster is used as a representative sequence for taxonomic assignment. Then, a post-clustering curation algorithm (LULU, Frøslev et al., 2017) was applied to curate the data. The taxonomic assignment was performed using the ECOTAG program against the NCBI database. Ecotag outputs were validated using the same thresholds as for the obitools one. Further quality cleaning was identical to the obitools pipeline (minimum number of reads, remove non-target taxa, tag-jump cleaning), with the addition of a single step removing all MOTUs present in only PCR within the entire dataset. This additional step is necessary as PCR errors are unlikely to be present in more than one PCR occurrence and it removes spurious MOTUs inflating diversity estimates (see Marques et al. 2020).

2.7. β diversity from marine to freshwater environments

We used the outputs from the SWARM pipeline in the form of MOTUs to perform diversity and composition analyses that do not strictly depend on the coverage of the reference database. We

performed an ordination of the 18 filters and eight sampling sites to investigate species composition. Next, we ranked the sampled sites along a salinity gradient from the upper part of the river, freshwater, brackish, shallow marine to deeper marine. We tested for an association between the vertebrate composition and the gradient from freshwater to marine composition. From the MOTUs presence-absence matrix, we calculated a Jaccard distance matrix between samples. To ordinate the compositional differences between the eDNA samples, we performed a PCoA on this distance matrix and reported the explained deviance of each axis. We mapped the ordination values in the geographic space. We tested for the effect of habitat on species composition by performing a permanova using the "adonis" function of the R package Vegan (Oksanen et al., 2019).

We tested whether assemblage in the same type of habitat (freshwater brackish and marine) had a more similar species assemblage. We created a presence-absence matrix based on the MOTUs at the habitat level, and we further calculated the pairwise Jaccard dissimilarity between sites (Anderson et al., 2011; βjac). This index is expressed as: βjac = (b+c)/(a+b+c); where a is the number of MOTUs present at both sites, b is the number of MOTUs present in first but not in the second site, and c is the number of MOTUs present in second, but not in the first site. βjac ranges from 0 (MOTUs composition does not change between sites) to 1 (MOTUs composition completely changes between sites). Second, we applied the partitioning framework proposed by Baselga (2012), which consists in decomposing the βjac, in two additive components, the replacement (or turnover) and the nestedness. The MOTUs replacement component describes MOTUs replacement without the influence of MOTUs richness difference between sites (βjtu= 2min(b,c)/a+2min(b,c)). The nestedness component (βjne = βjac – βjtu) accounts for the fraction of dissimilarity due to MOTUs richness difference.

We also quantified β diversity at the site level and applied the same partitioning of the β diversity and explored the relationship between species composition pairwise dissimilarity and geographical distance between sampled sites. We fitted exponential and power-law models, which describe the increase in species dissimilarity with spatial distance (Nekola & White, 1999). Following the procedure of Gómez-Rodríguez and Baselga (2018), we fitted a GLM where dissimilarity is explained by spatial distance. We selected a log link and Gaussian error for the exponential model or log-transformed for the power-law model. Then, we assessed the goodness of fit of the two models by calculating pseudo-r2. The significance of the relationships were assessed by randomizing spatial distances 999 times and computing the proportion of times in which the model deviance was smaller than the randomized model deviance (Gómez-Rodríguez & Baselga, 2018). We tested which model best fitted our data between the negative exponential or power-law models by comparing the AIC values.

3. RESULTS

3.1. Faunal list

We detected 253 different taxa for a total of 16,771,150 reads, but only 95 taxa (37.5%) could be identified to the species level. The remaining 160 were assigned to a higher taxonomic level. When filtering this taxa list by species and genus which have been reported in regional checklists, we excluded 15 taxa representing a total of 5,159,591 reads. 66 taxa were assigned at the species level spanning five vertebrate taxonomic groups, fishes, birds, amphibians, mammals and reptiles (Table S2-S5). Among those 66 species, 31 were fishes (26 detected in the marine environment and 10 in freshwater), and 35 were other vertebrate species. The teleo primer pair only detected 17 fish species (15 marine and 8 freshwater species). Using the Chond01 primer pair, two additional taxa were detected, the silky shark (Carcharhinus falciformis) in brackish waters and the genera Carcharhinus in both freshwater and marine environments. The spotted eagle ray (Aetobatus narinari) was the second chondrichthyan, detected in marine waters. There was an overlap in the fish species (8) recovered with the vertebrate primer and the fish specific primer sets (teleo). In contrast, other species such as the bigeye scad (Selar crumenophthalmus), the shortfinger anchovy (Anchoa lyolepis) and the Caitipa mojarra (Diapterus rhombeus) were detected only using the Vert01 primer, while the river goby (Awaous banana) and the Tarpon (Megalops atlanticus) were detected only using the teleo primer.

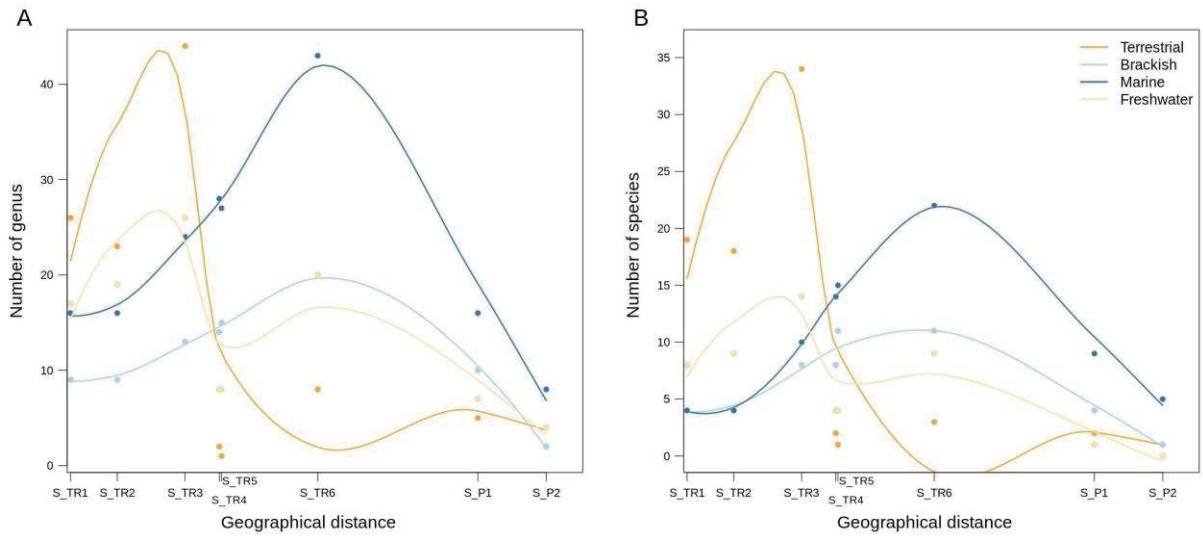


Figure 2. Relationship between a linear gradient representation from river (S_TR1 site) to the outer sea (S_P2 site) and the number of genus (A) and the species richness (B) of organisms recovered by eDNA and assigned taxonomically using obitools. The lines show the evolution of the species or genus number along a salinity gradient for terrestrial (dark orange), freshwater (light orange), brackish (light blue) and marine (dark blue) taxonomic groups. The linear representations were obtained by fitting a local polynomial regression.

The detected marine fishes mainly belonged to the families Pristigasteridae, Sciaenidae and Ariidae, which are mostly associated with pelagic habitats or with sandy bottoms. Hence, while the monitoring of this turbid habitat was previously difficult, the detected species suggest that there are no reefs at that location and the fauna is dominated by sandy bottom fauna. Closer to the river mouth, the samples contained typical brackish waters species, while in the river, it was dominated by typical freshwater species (Figure 2A). Across the four different depths, we found different compositions of taxa, with for example the detection of the families Carangidae in all the four depths (0, 35, 53-58, 115 m), pelagic families such as Hemiramphidae (Hemiramphus sp.), Carangidae (Selar crumenophthalmus) and Clupeidae (Ophistonema oglinum) in the surface samples; families as Engraulidae, Clupeidae and Gerreidae at 35 m; Elopidae (Elops sp.), Carangidae (Caranx sp.) and Myctophidae around 53-58m and Carangidae (Caranx sp.), Myctophidae (Diaphus sp.) and Ophidiidae at 115 m. The Myctophidae family was detected in the station located furthest from the coast.

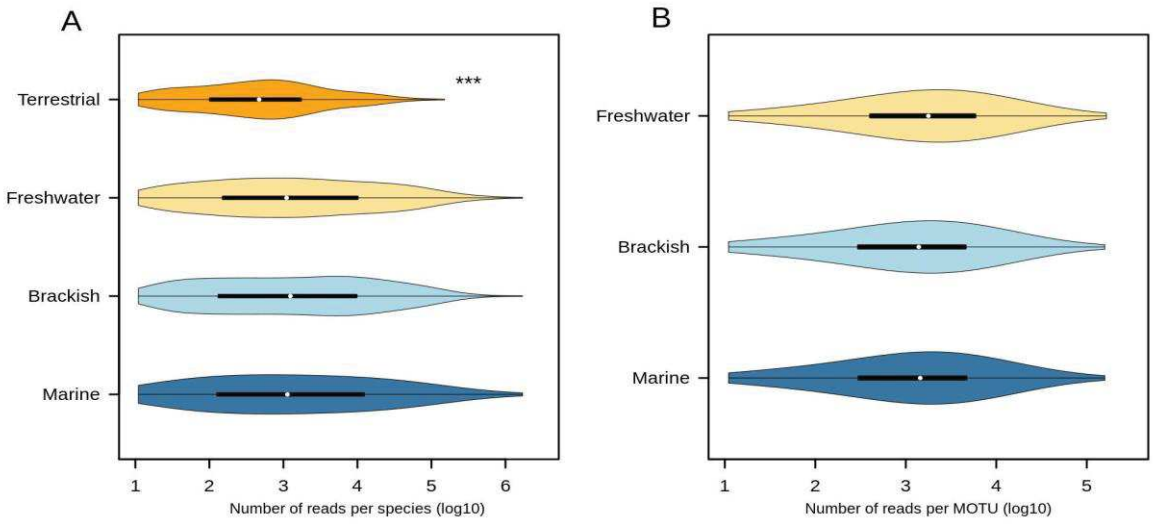


Figure 3. Number of reads per species and per MOTU in each habitat. Shown are (A) number of reads per species processed with the obitools bioinformatic pipeline (log10) and (B) number of reads per MOTU recovered from the Swarm bioinformatic pipeline (log10), represented in logarithmic scale. Habitat classification is based on the taxonomy recovered when comparing the reads to the reference database.

The vertebrate primers recovered not only fishes, but also species of many vertebrate clades, while the teleo primer showed the expected specification and did not recover any non-fish vertebrate species. The Vert01 primer set was particularly performing in detecting many species of amphibians, reptiles, birds, and mammals surrounding the upper section of the sampled river. Two species of amphibians (two species in marine and one species in freshwater) and one species, one genus and two families of reptiles were detected in freshwater, besides 18 species of birds (three species in marine and 17 in

freshwater) and 14 species of mammals (two species in marine and 13 species in freshwater). Among the mammals, we detected the brown-eared woolly opossum (Caluromys lanatus) and the tapir (Tapirus terrestris). In addition, we detected the small endemic red-crested tree rat (Santamartamys rufodorsalis), but with a low number of reads. Moreover, we detected a very important number of bat species, with nine genera and five species within four families. Among the birds, we detected endemic species such as the Santa Marta toucanet (Aulacorhynchus albivitta lautus) and the masked trogon (Trogon personatus sanctaemartae), as well as neotropical migrant birds, the spotted sandpiper (Actitus macularius), the greater yellowlegs (Tringa melanoleuca) and the belted kingfisher (Megaceryle alcyon). Among the amphibians, we detected the South American white-lipped grassfrog (Leptodactylus fuscus) and among the reptiles, we only detected the crocodile Caiman crocodilus. While we detected terrestrial species using eDNA, the number of reads per species remains significantly lower than for strictly aquatic species (marine versus terrestrial, t.test = 6.63, p.value < 0.001; Brackish vs Terrestrial, t.test = 5.44, p.value <0.001; Freshwater vs Terrestrial, t.test = 5.94, p.value < 0.001; Figure 3A).

3.2. Beta diversity from marine to freshwater environments

We used the outputs from the SWARM pipeline in the form of MOTUs to perform diversity and composition analyses that do not strictly depend on the coverage of the reference database. We focused only on the teleo primer since the clustering pipeline was only validated for this short marker. We detected 145 different MOTUs for a total of 12,682,925 reads, but only 25 sequences could be associated to a given species and 120 remained unassigned at the species level. We detected on average 29.11 ± 18.5 MOTUs per filter and there were no differences in detection between habitats when considering the number of reads per MOTU (Figure 3B).

We investigated the compositional differences in fish eDNA composition between the sampling stations by calculating MOTUs composition dissimilarity among samples. The PCoA ordination showed that the composition of the assemblages recovered from eDNA were grouped into their original habitats. The PCoA explained a large fraction of the total inertia (43.4%) with 24% for the first axis and 19.4% for the second axis and showed a marked difference in composition (Figure 4). We identified three different clusters presenting marked differences in MOTUs composition and that are related to habitat structuration (Permanova n=11, F = 3.3, R2 = 0.423, p.value = 0.001). The first axis of the PCoA discriminated freshwater sites from sites with marine influence whereas the second axis discriminates brackish from marine sites.

We observed a high βjac diversity between the three types of habitats (μβjac = 0.83 ± 0.063) mainly due to a high rate of MOTUs turnover in species composition (Figure S1). The value of βjtu was particularly high between freshwater and marine environments (βjtu = 0.823) and between freshwater and brackish environments (βjtu = 0.69) indicating a high rate of species replacement. Howeverconsidering the brackish and marine environments the nestedness component was more important highlighting a higher proportion of species shared between these habitats (βjne = 0.32; βjtu = 0.5; Figure S1).

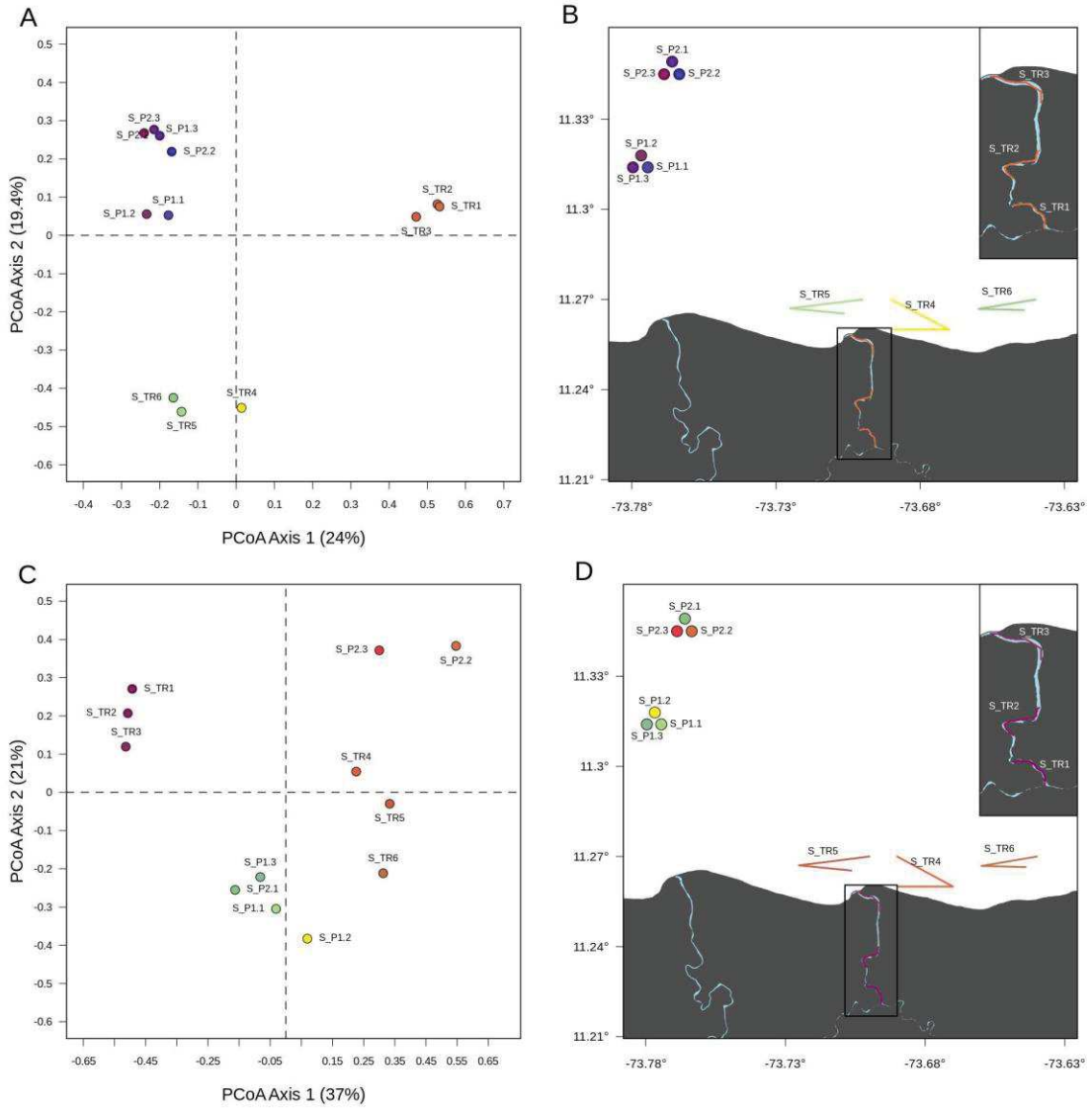


Figure 4. Ordination of the composition of the 18 eDNA samples using a Principal Coordinate Analysis (PCoA) on a Jaccard distance matrix computed from differences in fish MOTUs (A) in the outer estuary (S_P1.1, S_P1.2, S_P1.3 and S_P2.1, S_P2.2, S_P2.3), in proximity to the river mouth (S_TR4, S_TR5, S_TR6) and in the river (S_TR1, S_TR2, S_TR3) (B). Ordination of the composition of the 18 eDNA samples using a Principal Coordinate Analysis (PCoA) on the turnover component of the Jaccard dissimilarity metric computed from difference in fish MOTUs (C) and its associated geographical distribution (D).

We explored the relation between MOTUs compositional dissimilarity (βjac) and geographical distance between sampled sites by fitting exponential and power-law models. The best model was the exponential model that presents the lowest AIC with a value of -16.44, the highest pseudo-r2 = 0.22 associated with a significant p.value (p=0.01, Table 1, Figure S2A). The exponential model showed an increasing significant dissimilarity with the increasing distance between sites (Table 1, Figure S2A). However, the compositional dissimilarity between geographically close sites also presented a high rate of turnover leading to a non-significant fit of the exponential model (pseudo-r2 = 0.08; p.value = 0.13; Figure S2B), which indicate local composition heterogeneity within habitat.

Table 1 Adjusted GLM with dissimilarity as variable to explain and spatial distance as explanatory variable, we assessed the goodness of fit of these two models by calculating pseudo-r2 while the significant of the relations were assessed by randomizing spatial distances 999 times and computing the proportion of times in which the model deviance was smaller than the randomized model deviance. We tested which model best fitted our data between the negative exponential or power-law models by comparing the AIC values.

	Model type	Pseudo r squared	intercept	Slope	P.value	AIC
Bjac	Power	0.17	0.94	0.4	0.04	-14.83
Bjac	Exponential	0.22	0.64	10.61	0.01	-16.44
Bjtu	Exponential	0.08	0.57	5.58	0.13	-
Bjnes	Exponential	0.016	0.087	0.31	0.52	-

4. DISCUSSION

Estuaries provide critical goods and services for both local and worldwide populations and are known for example for their delivery of services such as coastal protection or fish nurseries (Barbier et al., 2011; Beck et al., 2001). However, estuaries are also heavily used and deteriorated globally (Lotze et al., 2006), which impacts the biodiversity and services that they provide (Barbier et al., 2011). Yet, monitoring biodiversity and possible associated ecosystem services in those complex habitats remains challenging. Our study demonstrates that eDNA metabarcoding represents a useful tool for the monitoring of

biodiversity in an estuary located in the Natural National Park Sierra Nevada de Santa Marta in Colombia and could provide a key technology for quantifying essential biodiversity variables (Proenca et al., 2017). We show that (i) eDNA allows a clear distinction in vertebrate composition among the three habitats inventoried, freshwater, brackish and marine (ii) the river habitat also carries a signal of terrestrial environment, which provides an integrator of biodiversity information. Moreover, while the region of Santa Marta has a high rate of deforestation and many of the forests surrounding estuaries have been largely impacted by human exploitation over the last few decades (Cavelier et al., 1998), we show that the estuary of the Don Diego River still contains a large diversity of vertebrate species and that the existing protection of the park is valuable to preserve the local biodiversity.

Our results indicate that water is a good environment to obtain an integrative view on the composition of biodiversity in estuary ecosystems, which include aquatic, but also terrestrial, and arboricol species (Figura 2; Figure 5). Samples filtered from the river contain fishes and aquatic reptiles, but also terrestrial vertebrates including mammals, amphibians and birds. Sampling tropical systems to find eDNA traces of vertebrates is difficult. It has been argued that soil samples are unlikely to be the most relevant material for sampling the diversity of aboveground animals because eDNA is poorly transported in soils and thus more patchily distributed (Levy-Booth et al., 2007, Nagler et al., 2018). Hence, the use of indirect medium has been tested, for example via owl pellets to detect smaller mammals (Rocha et al., 2015), or extract DNA from blood feeding invertebrates (Lynggaard et al., 2019; Rodgers et al., 2017). Alternatively, rivers could integrate the signal of both aquatic and terrestrial vertebrates, since water can transport material from the whole catchment and accumulate eDNA within water bodies (Sales et al., 2019; Leempoel et al., 2020). In our results, some of the species detected using eDNA from water samples belong to strictly terrestrial species such as bats or anteaters, which can be explained by the transport of the DNA into the river, even if how the DNA diffuses from the surrounding terrestrial surface into the water remains to be investigated. In agreement with our results, Sales et al. (2020) also detected eDNA from both aquatic and terrestrial mammals when sampling water in the Amazon's mainstream and tributaries, in addition to a river of the Brazilian Atlantic forest. Comparing these results with camera trapping data confirms the congruence between the methods (Sales et al., 2010). Hence, water transport of animal DNA makes rivers a good integrator of biodiversity information (Deiner et al., 2015).

Figure 5. Montage of pictures of the view of the Don Diego River and the Sierra Nevada de Santa Marta from the river mouth (a) and one example of terrestrial and arboricol species detected using eDNA. In the middle (b), the Spectacled caiman (*Caiman crocodilus*) and to the right (c), the Venezuelean red howler (*Alouatta seniculus*) detected as *Alouatta* sp.

The detection of species, that represent important conservation targets, emphasizes the relevance of eDNA metabarcoding as a useful tool for biodiversity assessment (Bohmann et al., 2014). As regards to vertebrates, we detected one critical endangered endemic species, the red-crested tree rat (Santamartamys rufodorsalis) that is listed as the top 100 most endangered species in the world and was not seen since 1898 and rediscovered recently in 2011 in the Sierra Nevada de Santa Marta (Velazco et al., 2017). We also detected two endemic subspecies of birds, the Santa Marta toucanet (Aulacorhynchus albivitta lautus) and the masked trogon (Trogon personatus sanctaemartae). The presence of the great tinamou (Tinamus major), listed as a near threatened (NT) species by the IUCN Red List and three neotropical migrant birds, the spotted sandpiper (Actitus macularius), the greater yellowlegs (Tringa melanoleuca) and the belted kingfisher (Megaceryle alcyon) represent, as well, important records for the region and help us to understand the migration behaviour of those animals. Some records were interesting from a biogeographic perspective. For example, the detection of the South American white-lipped grassfrog (Leptodactylus fuscus) represents the most septentrional record for the species, however, this finding requires further investigation as the detected sequences may come from a closely related species. Finally, we detected multiple fish taxa including the genus of Anguilla with the American eel (Anguilla rostrata), previously registered for some other river mouths of the Sierra Nevada de Santa Marta and some introduced species like the widespread Guppy (Poecilia reticulata). The presence of the grey triggerfish (Balistes capriscus), listed as a near threatened (NT) species by the IUCN Red List using both ecosystems possibly in different stages of life, found the juveniles in the brackish waters or big predators as Carcharhinus that was also detected in both ecosystems possibly more in search of food. Altogether, eDNA allows delivering novel information on the local distribution of vertebrates in a protected area, which include many species relevant for conservation.

Despite the diffusion of eDNA in the water environment (Harrison et al., 2019), the signal is not homogenized and a clear compositional gradient can be detected from the river into the shallow to the deeper part of the marine ecosystem (Figure 4). The analysis of fish MOTUs composition showed a strong turnover from the marine to the freshwater environment (Figure 4). The increase in compositional dissimilarity with geographical distance between sampled sites is due to species-specific niche differences in responses to the main environmental gradient from freshwater to marine habitats. The limited species turnover between marine and brackish sites suggest more permeability to exchange of organisms between those habitats (Figure 4C, D). Moreover, our results indicate that despite the movement of water in estuary, there is a localized eDNA signal which allows targeting specific habitat sampling (Jeunen et al., 2019). Among the families detected, the Myctophidae was detected in the three sampled strata from the farthest station on the coast and corresponds to the most diverse family of the mesopelagic zone and can be sometimes present in the epipelagic zone due to its vertical migrations. Closer to the coast, we found the typical Ariidae and Sciaenidae families, while in the river we found Eleotridae and Gobiidae with species such as the large scaled spinycheek sleeper (Eleotris amblyopsis) and the river goby (Awaous banana) respectively. In agreement with our results, West et al. (2020) sampled multiple sites in a tropical island ecosystem and showed that species assemblage composition varied significantly between habitats at small spatial scale, indicating the localisation of eDNA signals, despite extensive oceanic water movements. In agreement with other studies, our results suggest that eDNA represents one promising non-invasive alternatives to traditional sampling for small streams, rivers, lakes and the sea. Cilleros et al. (2019) compared eDNA and traditional sampling both in small streams and rivers across French Guiana. Not only did they find that species assemblages were congruent between eDNA and traditional records, but also that eDNA results were more efficient in distinguishing the fauna from different river drainages. One advantage of eDNA over other methods is that it enables the study of fish communities at cryptic life stages, which is critical for estuaries that serve as nurseries for many species (Huges et al., 2014).

Our study has several limitations associated with the limited number of samples collected and identification of the eDNA sequences, which were carried without a dedicated regional reference database. First, estuaries are complex habitats that show not only spatial but also temporal variations. In our case study, we only sampled during one specific period and did not investigate the seasonal variations in biodiversity. The second main limitation is the lack of a reference database, and thus the use of available sequences for vertebrates, with many species expected to be missing from the database and some others included but wrongly identified. For example, we detected 253 different taxa, but only 95 could be identified to the species level. Hence, a large number of unique sequences could not be associated with species, which highlights the need for a more detailed reference database. Moreover,

we did not find many species of amphibians while many are known for the region (Pérez-Gonzales et al., 2016) suggesting that the performance of the vertebrate primers that we used is less high for amphibians compared with other vertebrates. To optimize the detection of amphibian species, a more targeted primer could be necessary (Valentini et al., 2016). A further limitation of eDNA is that it is not possible to assess the age classes of individuals detected so that eDNA information does not allow us to directly indicate whether we detected adults or larvae. Yet, the detected presence of the Parassi mullet (Mugil incilis) and the Bobo mullet (Joturus pichardi), which are usually using a different habitat at the adult stage, suggest that the estuary is acting as a nursery for those species. Here, we took care to document the entire protocol used in the future, but standardized protocols are currently not available, and researchers should bring together expertise in eDNA to develop fast standard approaches that are comparable through further testing and efforts.

Here, we assessed the biodiversity in an estuary associated with the Sierra Nevada de Santa Marta National Natural Park. Our eDNA analyses in different aquatic environments of the estuary recovered a large number of species both aquatic and terrestrial. Monitoring the biodiversity in such a complex ecosystem as estuaries remains challenging, the turbidity and strong water currents make species inventories difficult, a limit that could be circumvented with eDNA (Belle et al., 2019). Our study allowed recovering a large biodiversity in the interface between the Don Diego River and the surrounding environments. Our results pave the way toward a broader application across estuaries of Colombia and in the Neotropics. The next step is to analyse a temporal signal to demonstrate temporal biodiversity dynamics, which would support the use of eDNA technology for future monitoring of estuaries. Assessing the fate of biodiversity changes within the context of global changes and supporting management policies rely largely on the measurement of biological diversity. We expect that expanding from our approach will help model biodiversity, challenge previously drawn ecological patterns and document biodiversity decline, which will support better defined conservation plans (Juhel et al., 2020). The slow degradation of estuaries and the decline in biodiversity (Thrush et al., 2004) could be better monitored using eDNA as demonstrated in our study. Given existing efforts to rehabilitate estuaries (Botero & Salzwedel, 1999), eDNA is expected to represent a major new tool to monitor the efficiency of those efforts.

Conflict of interest

All authors declare that there is no conflict of interest regarding the publication of this article.

Author Contribution Statement

LP, CA, APF jointly designed this study, APF, MMM, VM, JBJ, MCC, RH, EM, MS participated in the field work, AV, CA developed the analysis of the data; and all the authors APF, MMM, VM, FAV, GHB, MCC, TD, RH, JBJ, JDGC, EM, SM, MS, AV, DM, CA, LP contributed to write the manuscript.

Data Availability Statement:

Data are presented in supplemental material. All the sequence reads will be published after the acceptance of the manuscript.

Ethical guidelines:

According to Paragraph 1, Article 2.2.2.8.1.2., Section 1 (Permits), Chapter 8 (Scientific Research), of Decree 1076 of 2015 "The Ministry of Environment and Sustainable Development of Colombia, its affiliated entities, National Natural Parks of Colombia, the Autonomous Regional and / or Sustainable Development Corporations and the Large Urban Centers will not require the Specimen Collection Permit covered by this decree (…) "; therefore, the INVEMAR, being an entity attached to the Ministry of Environment and Sustainable Development (MADS) (see Article 1.2.2.1., Title 2, of Decree 1076 of 2015), does not require permission to collect specimens of wild life.

Acknowledgment

This project was supported by the foundation "Monaco Explorations". Thanks to the boats local community association for the transport services during the field work and to National Natural Parks especially to Tito Rodriguez, SNSM National Natural Park Chief. Thanks to Janeth Andrea Beltrán (Information Systems Laboratory of INVEMAR) for her support in cartographie, to Olivier Borde (photographer Explorations of Monaco) for the photographs taken during the expedition and SPYGEN staff for its support in eDNA laboratory. Contribution number xxx of the Instituto de Investigaciones Marinas y Costeras – INVEMAR, Colombia.

Reference

Alberico, M., Cadena, A., Camacho, J. H., & Saba, Y. M. (2000). Mamíferos (Synapsida: Theria) de Colombia. Biota colombiana, 1(1), 43-75.

Almeda, F., Alvear, M., & Mendoza-Cifuentes, H. (2013). Colombia, biodiversity hotspot and major center of diversity for Melastomataceae. In Scientific Abstracts (No. 276, pp. 27-31).

Anderson, M.J., Crist, T.O., Chase, J.M., Vellend, M., Inouye, B.D., Freestone, A.L., Sanders, N.J., … & Harrison, S.P. (2011). Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. Ecology letters, 14(1), 19-28.

Ayerbe-Quiñones, F. (2018). Guía ilustrada de la avifauna colombiana. Wildlife Conservation Society, Bogotá, Colombia.

Barbier, E. B., Hacker, S. D., Kennedy, C., Koch, E. W., Stier, A. C., & Silliman, B. R. (2011). The value of estuarine and coastal ecosystem services. Ecological monographs, 81(2), 169-193.

Barlow, J., França, F., Gardner, T. A., Hicks, C. C., Lennox, G. D., Berenguer, E., Castello, L., Economo, E. P., Ferreira, J., Guénard, B., Gontijo Leal, C., Isaac, V., Lees, A. C., Parr, C. L., Wilson, S. K., Young, P. J., & Graham, N. A. J. (2018). The future of hyperdiverse tropical ecosystems. Nature, 559(7715), 517–526. https://doi.org/10.1038/s41586-018-0301-1

Baselga, A. (2012). The relationship between species replacement, dissimilarity derived from nestedness and nestedness. Global ecology and biogeography, 21(12) 1223-1232.

Basset, A., Elliott, M., West, R. J., & Wilson, J. G. (2013). Estuarine and lagoon biodiversity and their natural goods and services.

Beck, M. W., Heck, K. L., Able, K. W., Childers, D. L., Eggleston, D. B., Gillanders, B. M., ... & Orth, R. J. (2001). The identification, conservation, and management of estuarine and marine nurseries for fish and invertebrates: a better understanding of the habitats that serve as nurseries for marine species and the factors that create site-specific variability in nursery quality will improve conservation and management of these areas. Bioscience, 51(8), 633-641., https://doi.org/10.1641/0006-3568(2001)051[0633:TICAMO]2.0.CO;2

Belle, C. C., Stoeckle, B. C., & Geist, J. (2019). Taxonomic and geographical representation of freshwater environmental DNA research in aquatic conservation. Aquatic Conservation: Marine and Freshwater Ecosystems, 29(11), 1996-2009.

Beng, K. C., & Corlett, R. T. (2020). Applications of environmental DNA (eDNA) in ecology and conservation: opportunities, challenges and prospects. Biodiversity and Conservation, 1-33.

Biggs, J., Ewald, N., Valentini, A., Gaboriaud, C., Dejean, T., Griffiths, R. A., ... & Williams, P. (2015). Using eDNA to develop a national citizen science-based monitoring programme for the great crested newt (Triturus cristatus). Biological Conservation, 183, 19-28.

Blowes, S. A., Supp, S. R., Antão, L. H., Bates, A., Bruelheide, H., Chase, J. M., ... & Winter, M. (2019). The geography of biodiversity change in marine and terrestrial assemblages. Science, 366(6463), 339-345. https://doi.org/10.1126/science.aaw1620

Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., ... & De Bruyn, M. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. Trends in ecology & evolution, 29(6), 358-367.

Botero, L., & Salzwedel, H. (1999). Rehabilitation of the Ciénaga Grande de Santa Marta, a mangrove-estuarine system in the Caribbean coast of Colombia. Ocean & Coastal Management, 42(2-4), 243-256.

Boussarie, G., Bakker, J., Wangensteen, O. S., Mariani, S., Bonnin, L., Juhel, J. B., ... & Vigliola, L. (2018). Environmental DNA illuminates the dark diversity of sharks. Science advances, 4(5), eaap9661.

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., & Coissac, E. (2016). obitools: A unix-inspired software package for DNA metabarcoding. Molecular ecology resources, 16(1), 176-182

Cantera, I., Cilleros, K., Valentini, A., Cerdan, A., Dejean, T., Iribar, A., ... & Brosse, S. (2019). Optimizing environmental DNA sampling effort for fish inventories in tropical streams and rivers. Scientific reports, 9(1), 1-11.

Cavelier, J., Aide, T., Santos, C., Eusse, A., & Dupuy, J. (1998). The savannization of moist forests in the Sierra Nevada de Santa Marta, Colombia. Journal of Biogeography, 25(5), 901-912.

Cilleros, K., Valentini, A., Allard, L., Dejean, T., Etienne, R., Grenouillet, G., Iribar, A., Taberlet, P., Vigouroux, R. & Brosse, S. (2019). Unlocking biodiversity and conservation studies in high-

diversity environments using environmental DNA (eDNA): A test with Guianese freshwater fishes. Molecular ecology resources, 19(1), 27-46.

Civade, R., Dejean, T., Valentini, A., Roset, N., Raymond, J. C., Bonin, A., ... & Pont, D. (2016). Spatial representativeness of environmental DNA metabarcoding signal for fish biodiversity assessment in a natural freshwater system. PloS one, 11(6), e0157366.

Clements, J. F., T. S. Schulenberg, M. J. Iliff, S. M. Billerman, T. A. Fredericks, B. L. Sullivan, and C. L. Wood. 2019. The eBird/Clements Checklist of Birds of the World: v2019. Retrieved from https://www.birds.cornell.edu/clementschecklist/download/

Collen, B., Ram, M., Zamin, T., & McRae, L. (2008). The tropical biodiversity data gap: addressing disparity in global monitoring. Tropical Conservation Science, 1(2), 75-88.

De Barba, M., Adams, J. R., Goldberg, C. S., Stansbury, C. R., Arias, D., Cisneros, R., & Waits, L. P. (2014). Molecular species identification for multiple carnivores. Conservation Genetics Resources, 6(4), 821-824.

Dejean, T., Valentini, A., Duparc, A., Pellier-Cuit, S., Pompanon, F., Taberlet, P., & Miaud, C. (2011). Persistence of environmental DNA in freshwater ecosystems. PloS one, 6(8), e23398.

Dejean, T., Valentini, A., Miquel, C., Taberlet, P., Bellemain, E., & Miaud, C. (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(4), 953-959.

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., ... & Pfrender, M. E. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. Molecular ecology, 26(21), 5872-5895.

Díaz, S., Settele, J., Brondízio, E. S., Ngo, H. T., Agard, J., Arneth, A., ... & Garibaldi, L. A. (2019). Pervasive human-driven decline of life on Earth points to the need for transformative change. Science, 366(6471).

Dixon, K. M., Cary, G. J., Worboys, G. L., Banks, S. C., & Gibbons, P. (2019). Features associated with effective biodiversity monitoring and evaluation. Biological Conservation, 238, 108221.

Dornelas, M., Magurran, A. E., Buckland, S. T., Chao, A., Chazdon, R. L., Colwell, R. K., ... & McGill, B. (2013). Quantifying temporal change in biodiversity: challenges and opportunities. Proceedings of the Royal Society B: Biological Sciences, 280(1750), 20121931.

Djurhuus, A., Closek, C. J., Kelly, R. P., Pitz, K. J., Michisaki, R. P., Starks, H. A., ... & Montes, E. (2020). Environmental DNA reveals seasonal shifts and potential interactions in a marine community. Nature communications, 11(1), 1-9. https://doi.org/10.1038/s41467-019-14105-1

Ficetola, G.F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., Gielly, L., Lopes, C.M., Boyer, F., Pompanon, F. & Rayé, G. (2015). Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. Molecular ecology resources, 15(3), 543-556.

Folke, C., Carpenter, S., Walker, B., Scheffer, M., Elmqvist, T., Gunderson, L., & Holling, C. S. (2004). Regime shifts, resilience, and biodiversity in ecosystem management. Annual review of ecology, evolution, and systematics, 35.

Frøslev, T. G., Kjøller, R., Bruun, H. H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., & Hansen, A. J. (2017). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. Nature communications, 8(1), 1-11.

Gilbert, S. F., Sapp, J., & Tauber, A. I. (2012). A symbiotic view of life: we have never been individuals. The Quarterly review of biology, 87(4), 325-341.

Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., ... & Laramie, M. B. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. Methods in ecology and evolution, 7(11), 1299-1307.

Gómez-Rodríguez, C., & Baselga, A. (2018). Variation among European beetle taxa in patterns of distance decay of similarity suggests a major role of dispersal processes. Ecography, 41(11), 1825-1834.

Greenberg, R. (2012). The ecology of estuarine wildlife. Estuarine ecology, 357-380.

INGEOMINAS, ECOPETROL ICP, INVEMAR, 2008. Evolución Geohistórica de la Sierra Nevada de Santa Marta. Caracterización climática de la SNSM y su efecto regulador en el clima regional. Bogotá, Colombia: Servicio Geológico Colombiano.

Harper, L. R., Handley, L. L., Carpenter, A. I., Ghazali, M., Di Muri, C., Macgregor, C. J., ... & McDevitt, A. D. (2019). Environmental DNA (eDNA) metabarcoding of pond water as a tool to survey conservation and management priority mammals. Biological Conservation, 238, 108225.

Harrison, J. B., Sunday, J. M., & Rogers, S. M. (2019). Predicting the fate of eDNA in the environment and implications for studying biodiversity. Proceedings of the Royal Society B, 286(1915), 20191409. https://doi.org/10.1098/rspb.2019.1409

Hilty, J., & Merenlender, A. (2000). Faunal indicator taxa selection for monitoring ecosystem health. Biological conservation, 92(2), 185-197.

Hughes, B. B., Levey, M. D., Brown, J. A., Fountain, M. C., Carlisle, A. B., Litvin, S. Y., ... & Gleason, M. G. (2014). Nursery functions of US West Coast estuaries: the state of knowledge for juveniles of focal invertebrate and fish species. Arlington, Virginia: Nature Conservancy.

Jeunen, G. J., Knapp, M., Spencer, H. G., Lamare, M. D., Taylor, H. R., Stat, M., ... & Gemmell, N. J. (2019). Environmental DNA (eDNA) metabarcoding reveals strong discrimination among diverse marine habitats connected by water movement. Molecular Ecology Resources, 19(2), 426-438.

Jiménez-Alvarado, J. S., Rodríguez, C., Valencia-Mazo, J. D., Velandia, O., Fajardo, S., Morelo, L., Moreno-Díaz, C., Vela-Vargas, I.M., González-Maya, J. F. (2015). Planeación ambiental para la conservación de la biodiversidad en las áreas operativas de ecopetrol: informe final ventana SNSM, Ciénaga, Magdalena. Informe Técnico Final. Proyecto de Conservación de Aguas y Tierras – ProCAT Colombia, The Sierra To Sea Institute, Instituto de investigación de Recursos Biológicos Alexander von Humboldt. Retrieved from http://repository.humboldt.org.co/bitstream/handle/20.500.11761/9344/10_ProCAT_El_Congo.pdf?sequence=1&isAllowed=y.

Juhel, J.B., Utama, R.S., Marques, V., Vimono, I.B., Sugeha, H.Y., Kadarusman, Pouyaud, L., Dejean, T., Mouillot, D., Hocdé, R. (in press) Accumulation curves of environmental DNA predict coastal fish diversity in the Coral Triangle. Proceedings of the Royal Society B: Biological Sciences.

Lachavanne, J. B., & Juge, R. (Eds.). (1997). Biodiversity in land-inland water ecotones (Vol. 18). Taylor & Francis.

Leempoel, K., Hebert, T., & Hadly, E. A. (2020). A comparison of eDNA to camera trapping for assessment of terrestrial mammal diversity. Proceedings of the Royal Society B, 287(1918), 20192353.

Levin, L. A., Boesch, D. F., Covich, A., Dahm, C., ErsÚus, C., Ewel, K. C., ... & Strayer, D. (2001). The function of marine critical transition zones and the importance of sediment biodiversity. Ecosystems, 4(5), 430-451.

Levy-Booth, D. J., Campbell, R. G., Gulden, R. H., Hart, M. M., Powell, J. R., Klironomos, J. N., ... & Dunfield, K. E. (2007). Cycling of extracellular DNA in the soil environment. Soil Biology and Biochemistry, 39(12), 2977-2991.

Lotze, H. K., Lenihan, H. S., Bourque, B. J., Bradbury, R. H., Cooke, R. G., Kay, M. C., ... & Jackson, J. B. (2006). Depletion, degradation, and recovery potential of estuaries and coastal seas. Science, 312(5781), 1806-1809.

Lynggaard, C., Nielsen, M., Santos-Bay, L., Gastauer, M., Oliveira, G., & Bohmann, K. (2019). Vertebrate diversity revealed by metabarcoding of bulk arthropod samples from tropical forests. Environmental DNA, 1(4), 329-341.

Mace, G. M. (2004). The role of taxonomy in species conservation. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 359(1444), 711-719.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2014). Swarm: robust and fast clustering method for amplicon-based studies. PeerJ, 2, e593.

[dataset] Mammal Diversity Database. (2020). Retrieved from www.mammaldiversity.org. American Society of Mammalogists. Accessed 2020-06-16.

Marques, V., Guérin, P. E., Rocle, M., Valentini, A., Manel, S., Mouillot, D., Dejean, T. (In review) Blind assessment of vertebrate taxonomic diversity across spatial scales by clustering environmental DNA metabarcoding sequences. Ecography.

Martin, M. (2013). Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011. 2011; 17 (1): 3% J EMBnet. journal. Epub 2011-08-02. https://doi. org/10.14806/ej. 17.1. 200.

Mueller, M., & Geist, J. (2016). Conceptual guidelines for the implementation of the ecosystem approach in biodiversity monitoring. Ecosphere, 7(5), e01305.

Nagler, M., Insam, H., Pietramellara, G., & Ascher-Jenull, J. (2018). Extracellular DNA in natural environments: features, relevance and applications. Applied microbiology and biotechnology, 102(15), 6343-6356.

[dataset] National Museum of Natural History, Smithsonian Institution. Mammal Species of the World. Checklist dataset Retrieved from https://doi.org/10.15468/csfquc accessed via GBIF.org on 2020-06-15.

Nekola, J. C., & White, P. S. (1999). The distance decay of similarity in biogeography and ecology. Journal of biogeography, 26(4), 867-878.

Oksanen,J., F. Guillaume Blanchet, Michael Friendly, Roeland Kindt,Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin, L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs and Helene Wagner (2019). vegan: Community Ecology Package. R package version 2.5-6. Retrieved from https://CRAN.R-project.org/package=vegan

Paknia, O., Sh, H. R., & Koch, A. (2015). Lack of well-maintained natural history collections and taxonomists in megadiverse developing countries hampers global biodiversity exploration. Organisms Diversity & Evolution, 15(3), 619-629.

Pellissier, L., Niculita-Hirzel, H., Dubuis, A., Pagni, M., Guex, N., Ndiribe, C., ... & Guisan, A. (2014). Soil fungal communities of grasslands are environmentally structured at a regional scale in the A lps. Molecular ecology, 23(17), 4274-4290.

Pérez-Gonzales J.L., Mejía-Quintero L.R., Jiménez-López L.C., Rocha-Usiaga, A., Rueda-Solano, L.A. (2016). Anfibios y reptiles de Santa Marta y sus alrededores Colombia. Editorial Unimagdalena. Santa marta, Colombia.

Pineda-Guerrero, A., González-Maya, J. F., & Zárrate-Charry, D. (2015). Inventario preliminar de mamíferos de las Reservas privadas Námaku y el Jardín de Las Delicias, estribaciones de la Sierra Nevada de Santa Marta, Colombia. Mammalogy Notes, 2(1), 40-43.

Polanco, F., A., Fopp, F., Albouy, C., Brun, P., Boschman, L., & Pellissier, L. (In review) Past and present environmental conditions are associated with marine fish diversity in Tropical America. Journal of Biogeography.

Pont, D., Rocle, M., Valentini, A., Civade, R., Jean, P., Maire, A., ... & Dejean, T. (2018). Environmental DNA reveals quantitative patterns of fish biodiversity in large rivers despite its downstream transportation. Scientific reports, 8(1), 1-13.

Potapov, P., Hansen, M. C., Laestadius, L., Turubanova, S., Yaroshenko, A., Thies, C., ... & Esipova, E. (2017). The last frontiers of wilderness: Tracking loss of intact forest landscapes from 2000 to 2013. Science advances, 3(1), e1600821.

Proença, V., Martin, L. J., Pereira, H. M., Fernandez, M., McRae, L., Belnap, J., ... & Honrado, J. P. (2017). Global biodiversity monitoring: from data sources to essential biodiversity variables. Biological Conservation, 213, 256-263.

Reizopoulou, S., Simboura, N., Barbone, E., Aleffi, F., Basset, A., & Nicolaidou, A. (2014). Biodiversity in transitional waters: steeper ecotone, lower diversity. Marine ecology, 35, 78-84.

Roach, N. S., Urbina-Cardona, N., & Lacher Jr, T. E. (2020). Land cover drives amphibian diversity across steep elevational gradients in an isolated neotropical mountain range: Implications for community conservation. Global Ecology and Conservation, 22, e00968.

Rocha, E., Soares, K., & Pereira, I. (2015). Medium-and large-sized mammals in Mata Atlântica State Park, southeastern Goiás, Brazil. Check List, 11, 1.

[dataset] Robertson, D. R. & Tassell, J. Van. (2019). Shorefishes of the Greater Caribbean: online information system. Version 2.0 Smithsonian Tropical Research Institute, Balboa, Panamá. Retrieved from https://biogeodb.stri.si.edu/caribbean/es/pages

Rodgers, T. W., Xu, C. C., Giacalone, J., Kapheim, K. M., Saltonstall, K., Vargas, M., ... & Jansen, P. A. (2017). Carrion fly-derived DNA metabarcoding is an effective tool for mammal surveys: Evidence from a known tropical mammal community. Molecular Ecology Resources, 17(6), e133-e145.

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. PeerJ, 4, e2584.

Rueda, M., Doncel, O., Viloria, E. A., Mármol, D., García, C., Girón, A., ... & Barreto, C. (2011). Atlas de pesca marino-costera de Colombia:(2010-2011). Tomo Caribe. Santa Marta, Colombia: Serie de Publicaciones Generales del Invemar.

Ruthven, A. G., & Carriker, M. A. (1922). The amphibians and reptiles of the Sierra Nevada de Santa Marta, Colombia. University of Michigan, Museum of Zoology, Miscellaneous Publications N°8. 121 pp.

Sales, N. G., McKenzie, M. B., Drake, J., Harper, L. R., Browett, S. S., Coscia, I., ... & Ochu, E. (2020). Fishing for mammals: Landscape-level monitoring of terrestrial and semi-aquatic communities using eDNA from riverine systems. Journal of Applied Ecology, 57(4), 707-716. https://doi.org/10.1111/1365-2664.13592

Schmeller, D. S., Böhm, M., Arvanitidis, C., Barber-Meyer, S., Brummitt, N., Chandler, M., ... & Gill, M. (2017). Building capacity in biodiversity monitoring at the global scale. Biodiversity and conservation, 26(12), 2765-2790.

Schnell, I. B., Sollmann, R., Calvignac-Spencer, S., Siddall, M. E., Douglas, W. Y., Wilting, A., & Gilbert, M. T. P. (2015). iDNA from terrestrial haematophagous leeches as a wildlife surveying and monitoring tool–prospects, pitfalls and avenues to be developed. Frontiers in zoology, 12(1), 24.

Smith, T. B., Wayne, R. K., Girman, D. J., & Bruford, M. W. (1997). A role for ecotones in generating rainforest biodiversity. Science, 276(5320), 1855-1857.

Stat, M., Huggett, M. J., Bernasconi, R., DiBattista, J. D., Berry, T. E., Newman, S. J., ... & Bunce, M. (2017). Ecosystem biomonitoring with eDNA: metabarcoding across the tree of life in a tropical marine environment. Scientific Reports, 7(1), 1-11. https://doi.org/10.1038/s41598-017-12501-5

Stoeckle, M. Y., Soboleva, L., & Charlop-Powers, Z. (2017). Aquatic environmental DNA detects seasonal fish abundance and habitat preference in an urban estuary. PloS one, 12(4), e0175186. https://doi.org/10.1371/journal.pone.0175186

Strewe, R., & Navarro, C. (2003). New distributional records and conservation importance of the San Salvador Valley, Sierra Nevada de Santa Marta, northern Colombia. Ornitología Colombiana, 1, 29-41.

Strewe, R., & Navarro, C. (2004). New and noteworthy records of birds from the Sierra Nevada de Santa Marta region, north-eastern Colombia. Bulletin-British Ornithologists Club, 124(1), 38-50.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular ecology, 21(8), 2045-2050.

Taberlet, P., Bonin, A., Coissac, E., & Zinger, L. (2018). Environmental DNA: For biodiversity research and monitoring. Oxford University Press.

Thrush, S. F., Hewitt, J. E., Cummings, V. J., Ellis, J. I., Hatton, C., Lohrer, A., & Norkko, A. (2004). Muddy waters: elevating sediment input to coastal and estuarine habitats. Frontiers in Ecology and the Environment, 2(6), 299-306.

Torné Salas, A.C. (2013). Evaluación del ensamblaje de carnívoros por medio de metodologías comparativas en la reserva natural Ñámaku, Sierra Nevada de Santa Marta. (Bachelor dissertation Facultad de Ciencias básicas, Programa de Biología, Universidad del Magdalena). Retrieved from http://repositorio.unimagdalena.edu.co/jspui/handle/123456789/65

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., ... & Gaboriaud, C. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. Molecular ecology, 25(4), 929-942.

Velazco, P. M., Vargas, L. M., & Ramírez-Chaves, H. E. (2017). Santamartamys rufodorsalis (Rodentia: Echimyidae). Mammalian Species, 49(948), 63-67.

[dataset] Verhelst-Montenegro, J.C. & Salaman, P. (2019) Checklist of the Birds of Colombia / Lista de las Aves de Colombia. Electronic list, version '15 February 2019'. Atlas of the Birds of Colombia. Retrieved from https://sites.google.com/site/haariehbamidbar/atlas-of-the-birds-of-colombia [Accessed 15/03/2020].

Villa-Navarro, F.A., Sánchez-Duarte, P., Acero P., A., & Lasso, C., (2016) Composición y estructura de la ictiofauna de ríos y arroyos costeros de la Sierra Nevada de Santa Marta, Caribe colombiano. In: Bolaños, N., Barriga, R., Lira, E., Lasso-Alcalá, Ó. M., Lasso, C. A., Morales-Betancourt, M. A., ... & Espino, J. (2016). Cuencas pericontinentales de Colombia. Bogotá, Colombia: IAvH

Wall, D.H., Palmer, M.A., & Snelgrove, P.V. (2001). Biodiversity in critical transition zones between terrestrial, freshwater, and marine soils and sediments: processes, linkages, and management implications. Ecosystems, 4(5), 418-420.

West, K. M., Stat, M., Harvey, E. S., Skepper, C. L., DiBattista, J. D., Richards, Z. T., ... & Bunce, M. (2020). eDNA metabarcoding survey reveals fine-scale coral reef community variation across a remote, tropical island ecosystem. Molecular Ecology, 29(6), 1069-1086.

Résumé

La vitesse et l'intensité des changements globaux nécessitent de nouveaux moyens d'observations de la biodiversité qui soient rapides, non-destructifs, standardisés, déployables à large échelle et dans les écosystèmes les plus reculés (océan profond). Les méthodes de recensement classiques reposent sur l'identification morphologique ou sonore des espèces, mais celles-ci sont coûteuses en temps et en expertise. Au-delà de ces signaux, les animaux laissent aussi des traces d'ADN dans leur environnement sous la forme de cellules dermiques, de mucus ou de fèces. Le metabarcoding de cet ADN environnemental (ADNe) consiste à le collecter, l'amplifier et le séquencer pour identifier les espèces présentes grâce à des bases de séquences génétiques de référence. Or, ces bases de référence sont incomplètes, ce qui limite fortement le potentiel de l'ADNe pour révéler la biodiversité présente. Cette thèse a pour but de développer une approche alternative basée sur des unités taxonomiques moléculaires (MOTUs) pour analyser la biodiversité des macroorganismes aquatiques, et plus particulièrement celle des poissons osseux. J'ai tout d'abord réalisé une synthèse globale et spatialisée de la couverture taxonomique des bases de référence de séquences génétiques pour tous les poissons osseux, qui montre une sous-représentation des espèces de la zone tropicale ainsi que des lacunes concernant les espèces menacées et non-indigènes. Seules 13% des espèces de poisson sont séquencées pour le marqueur le plus commun, ce qui exclut toute ambition d'analyse exhaustive de la biodiversité par assignation aux espèces à court et moyen terme. En conséquence, j'ai développé un pipeline bio-informatique pour générer des estimations de la diversité en unités taxonomiques moléculaires (MOTUs) par famille de poissons. Les résultats démontrent que cette diversité en MOTUs représente un excellent substitut de la diversité en espèces à différentes échelles spatiales. Ensuite une application du metabarcoding de l'ADNe et de l'approche en MOTUs a permis d'estimer la diversité fonctionnelle, basée sur les traits des espèces, et la diversité phylogénétique, basée sur l'histoire évolutive des espèces, des poissons tropicaux de manière plus exhaustive que des méthodes traditionnelles (vidéos, plongées). Enfin, dans une première analyse globale de la diversité des récifs coralliens en ADNe, qui rassemble 251 échantillons récoltés depuis l'Océan Indien jusque dans les Caraïbes, l'approche en MOTUs permet de reconstruire les gradients biogéographiques des poissons mais aussi de révéler une hétérogénéité spatiale locale jusqu'alors sous-estimée. Alors qu'il est aujourd'hui crucial de mettre en place des méthodes de suivi efficaces, non dépendantes de spécialistes et à haute fréquence temporelle pour mieux comprendre les effets des changements globaux sur la biodiversité, ces travaux démontrent tout le potentiel de l'ADNe avec approche en MOTUs pour construire des indicateurs robustes de plusieurs facettes de la biodiversité à plusieurs échelles, mais aussi tester les hypothèses théoriques sous-jacentes à la distribution de cette biodiversité.

Mots-clés : ADNe, biodiversité, communauté, récifs coralliens, clustering, MOTU

Abstract

The speed and intensity of global change requires new means of observing biodiversity that are rapid, non-destructive, standardized, widely deployable and in remote ecosystems (deep sea). Conventional inventory methods are based on morphological or acoustic identification of species, which are costly in terms of time and expertise. Beyond these signals, animals also leave traces of DNA in their environment in the form of dermal cells, mucus or feces. The metabarcoding of this environmental DNA (eDNA) consists in collecting this DNA, amplifying and sequencing it to identify the species present using a genetic reference database. However, these reference databases are incomplete, which severely limits the potential of eDNA. The aim of this thesis is to develop an alternative approach based on molecular taxonomic units (MOTUs) to analyze the biodiversity of aquatic macroorganisms, and more particularly that of bony fish. I first performed a global and spatialized synthesis of the taxonomic coverage of the genetic reference database for all bony fishes, which shows an under-representation of species in the tropical zone as well as taxonomic gaps for endangered and non-indigenous species. Only 13% of fish species are sequenced for the most common marker, which excludes any ambition for an exhaustive analysis of biodiversity using only species-level assignments in the short or medium term. Consequently, I have developed a bioinformatics pipeline to generate estimates of diversity using molecular taxonomic units (MOTUs) by fish family. It shows how this MOTU diversity represents an excellent proxy for species diversity at different spatial scales. Then an application of eDNA metabarcoding and the MOTUs approach allowed to estimate the functional diversity, based on species traits, and the phylogenetic diversity, based on the evolutionary history of the species, of tropical fishes in a more exhaustive way than traditional methods (videos, dives). Finally, in a first global analysis of coral reef diversity in eDNA, which brings together 251 samples collected from the Indian Ocean to the Caribbean, the MOTUs approach allows the reconstruction of major trends in fish biogeography but also reveals local spatial heterogeneity hitherto underestimated. While it is now crucial to set up efficient, non-specialist dependent and high temporal frequency monitoring methods to better understand the effects of global changes on biodiversity, this work demonstrates the full potential of eDNA using a MOTUs approach to build robust indicators of several facets of biodiversity at several scales, but also to test theoretical hypotheses underlying the distribution of this biodiversity.

Keywords: eDNA, biodiversity, community, coral reefs, clustering, MOTU