L'ADN environnemental pour décrire les patrons de diversité des poissons à large échelle et informer la conservation

Laëtitia Mathon

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

En Ecologie et Biodiversité

École doctorale GAIA

Unité de recherche CEFE

L'ADN environnemental pour décrire les patrons de diversité des poissons à large échelle et informer la conservation

Présentée par Laëtitia MATHON
Le 26 janvier 2023

Sous la direction de Stéphanie MANEL

Devant le jury composé de

Simon BLANCHET, Directeur de recherche, CNRS	Rapporteur
Maud MOUCHET, Maître de conférence, Muséum National d'Histoire Naturelle	Rapportrice
Lucie ZINGER, Maître de conférence, Institut de Biologie de l'Ecole Normale Supérieure	Examinatrice
Anik BRIND'AMOUR, Chercheur, Ifremer	Examinatrice
Daniel BARTHELEMY, Directeur de recherche, CIRAD	Président du jury
Stéphanie MANEL, Directrice d'études, EPHE	Directrice
David MOUILLOT, Professeur, Université de Montpellier, UMR MARBEC	Encadrant
Laurent VIGLIOLA, Chargé de recherche, UMR ENTROPIE	Co-encadrant
Tony Dejean, Président, SPYGEN	Co-encadrant

UNIVERSITÉ DE MONTPELLIER

Résumé

L'accélération des changements globaux et les impacts humains menacent la survie des communautés de poissons à l'échelle mondiale. Or ces communautés sont indispensables au bon fonctionnement des écosystèmes marins et aux populations dépendantes de la pêche. Seul un suivi efficace et rapide des communautés de poissons à petite et grande échelle pour comprendre leurs distributions, les règles d'assemblage et les impacts des pressions humaines et environnementales peut permettre d'implémenter des mesures de conservation optimales. L'ADN environnemental (ADNe) est une méthode récente, dont l'efficacité a été démontrée à échelle locale et régionale pour étudier les communautés de poissons côtiers. Cette méthode permet de pallier à certains biais induits par les méthodes de suivi conventionnelles (pêche, plongée, caméras). Le but de cette thèse est d'utiliser l'ADNe pour étudier les distributions des poissons à diverses échelles spatiales, en réponse à des facteurs environnementaux, géographiques et socio-économiques, puis de nourrir les approches de planification de conservation. J'ai d'abord comparé les outils bio-informatiques fréquemment utilisés pour l'analyse des données d'ADNe, identifié les meilleurs programmes et pipelines et construit un pipeline optimal pour identifier les espèces contenues dans un échantillon, dans le cas d'une base de référence complète. Cependant, à large échelle, les bases de références sont largement incomplètes pour le gène mitochondrial 12S que nous utilisons et ne permettent donc pas l'assignation taxonomique de tous les fragments d'ADNe. Les autres études de cette thèse reposent ainsi sur une méthode de regroupement des séquences en unités taxonomiques moléculaires (MOTUs). À partir d'un jeu de données échantillonné à large échelle dans trois océans (Indien, Pacifique, Atlantique), j'ai comparé les estimations de diversité de poissons de récifs coralliens obtenues avec l'ADNe et avec des données de recensements visuels en plongée, à l'échelle de plusieurs biorégions. J'ai démontré que l'ADNe estimait une plus grande diversité de familles et de MOTUs que les plongées, tout en retrouvant les grands patrons de distributions connus (gradient longitudinal, isolation de la faune des Caraïbes). Puis, à l'échelle globale, j'ai étudié l'influence de facteurs environnementaux, géographiques et socio-économiques sur plusieurs indices de diversité alpha et beta, à partir de plus de 500 échantillons d'ADNe prélevés dans 11 régions du monde. Les résultats montrent un effet dominant de l'environnement (température et productivité) sur la diversité alpha et bêta, mais aussi une diminution de ces diversités dans les zones proches des populations humaines et notamment dans les pays dépendant des ressources marines. Enfin, à une échelle régionale en Nouvelle Calédonie, en

combinant l'ADNe avec des méthodes plus conventionnelles (caméras appâtées et échosondeur acoustique), j'ai estimé et modélisé plusieurs métriques de diversité des poissons sur les pentes externes et monts sous-marins, jusqu'à 600m de profondeur, que j'ai ensuite intégrées dans une planification de conservation en trois dimensions. Ces résultats indiquent de fortes richesses, abondances et biomasses sur les monts sous-marins peu profonds et isolés, ainsi que sur les pentes externes des îles et atolls éloignés des zones urbanisées, qui sont donc à prioriser dans les plans de conservation. L'ensemble des travaux de cette thèse démontrent l'utilité du metabarcoding de l'ADNe pour étudier la distribution des poissons à fine et large échelles spatiales, étudier l'impact des conditions environnementales et socio-économique sur la diversité et la distribution des communautés de poissons et informer les gestionnaires sur les zones de priorité de conservation.

Mots-clés : ADN environnemental, Distribution poissons, Diversité alpha, Diversité beta, Planification spatiale 3D

Abstract

Accelerating global changes and human impacts threaten the survival of fish communities worldwide, which are critical to the functioning of marine ecosystems and fisheries-dependent populations. Only an effective and rapid monitoring of fish communities at small and large scales to understand their distributions, assembly rules and impacts of human and environmental pressures can enable the implementation of optimal conservation measures. Environmental DNA (eDNA) is a recent method that has been demonstrated to be effective at local and regional scales for studying coastal fish communities. This method makes it possible to overcome some biases induced by conventional monitoring methods (fishing, diving, cameras). The goal of this thesis is to use eDNA to study fish distributions at various spatial scales in response to environmental, geographic, and socioeconomic factors, and then to feed conservation planning approaches. I first compared frequently used bioinformatics tools for eDNA data analysis, identified the best programs and pipelines, and constructed an optimal pipeline for identifying species contained in a sample, in the case of a complete reference database. However, at a large scale, the genetic reference databases are largely incomplete for the 12S mitochondrial gene we use, and thus do not allow for taxonomic assignment of all eDNA fragments. The other studies in this thesis therefore rely on a method of clustering sequences into molecular taxonomic units (MOTUs). From a large-scale dataset sampled in three oceans (Indian, Pacific, Atlantic), I compared estimates of coral reef fish diversity obtained with eDNA and with visual census data, at the scale of several bioregions. I demonstrated that eDNA estimated a higher diversity of families and MOTUs than visual census, while recovering the known major distribution patterns (longitudinal gradient, isolation of the Caribbean fauna). Then, on a global scale, I investigated the influence of environmental, geographical and socio-economic factors on several alpha and beta diversity indices, using more than 500 eDNA samples collected in 11 regions of the world. The results show a dominant effect of the environment (temperature and productivity) on alpha and beta diversity, but also a decrease of these diversities in areas close to human populations, and in particular in countries depending on marine resources. Finally, at a regional scale in New Caledonia, by combining eDNA with more conventional methods (baited cameras and acoustic echosounder), I estimated and modeled several metrics of fish diversity on deep outer slopes and seamounts, down to 600m depth, which I then integrated into three-dimensional conservation planning. These results indicate high richness, abundance and biomass on shallow and isolated seamounts, as

well as on the deep slopes of islands and atolls far from urbanized areas, which should be prioritized in conservation plans. All this thesis work demonstrates the utility of eDNA metabarcoding to study fish distribution at fine and large spatial scales, to study the impact of environmental and socio-economic conditions on the diversity and distribution of fish communities, and to inform managers on priority areas for conservation.

Keywords: Environmental DNA, Fish distribution, Alpha diversity, Beta diversity, 3D conservation spatial planning

Remerciements

Je souhaite tout d'abord remercier les membres de mon jury, Maud Mouchet, Simon Blanchet, Lucie Zinger, Annick Brind'Amour et Daniel Barthélémy d'avoir accepté d'évaluer ce travail de thèse.

J'aimerais ensuite remercier mes encadrants de thèse, Stéphanie Manel, David Mouillot et Laurent Vigliola, qui m'ont accompagnée tout au long de ce travail, m'ont transmis leurs connaissances et leur passion du monde marin, et m'ont permis de m'épanouir dans mes recherches. Stéphanie, merci pour ta confiance et ton soutien depuis le début. Ta grande disponibilité et le dynamisme que tu as insufflé à l'équipe ont créé une ambiance de travail très agréable et bienveillante. Merci de m'avoir permis de mener cette thèse dans les meilleures conditions. David, je te remercie pour ton optimisme sans faille, pour ton soutien et tes encouragements, même dans les moments plus difficiles. Merci d'avoir cru en moi dès le début de mon master et de m'avoir offert ces opportunités de recherche. Merci d'avoir toujours gardé ta porte ouverte pour des discussions toujours passionnantes et motivantes. Laurent, merci de m'avoir toujours accueillie à bras ouverts à Nouméa, d'avoir pris le temps de m'impliquer dans tes projets, et d'avoir partagé avec moi ta passion pour le Caillou et ses merveilles. Merci à vous trois de m'avoir formée et accompagnée dans le monde de la recherche académique.

Je remercie ensuite les membres de mon comité de thèse, Wilfried Thuiller, Laurent Pouyaud et Xavier Morin, pour votre bienveillance, vos encouragements et vos conseils toujours très pertinents.

Un grand merci à la région Occitanie et à SPYGEN, qui ont financé et rendu possible ces travaux de thèse. Merci également aux Explorations de Monaco d'avoir financé une partie du terrain et la collecte des données utilisé dans cette thèse.

Cette thèse n'aurait pu se faire sans les nombreuses personnes qui ont participé aux missions de terrain, à la collecte des données, et aux analyses en laboratoire. Je remercie donc Emilie pour avoir organisé et partagé avec moi la première mission de collecte d'ADN environnemental aux Baléares, qui fut assez épique. Et un grand merci à l'équipe de l'EIO pour leur accueil sur place. Merci à la team terrain de Nouvelle-Calédonie, Florian, Nadia, Romain, Mahé, Sam et Miguel pour la bonne ambiance sur les missions côtières. Et merci à l'équipe embarquée sur l'Alis de m'avoir supportée et d'avoir toujours gardé une bonne humeur, malgré mon mal de mer persistant. Je remercie particulièrement Florian d'avoir su endosser le rôle de chef de mission avec brio, Armelle d'avoir partagé mes galères, Claire pour ta bonne humeur

perpétuelle, Mahé pour ton endurance et tes compétences de pécheur, et tout l'équipage de l'Alis pour cette mission réussie ! Je remercie ensuite l'extraordinaire équipe de la mission en Tanzanie, Laure, Alicia, Loic et les sociologues Sébastien, Antoine et Julia, pour ces très bons moments passés ensemble à Zanzibar, malgré les galères. Je garde de merveilleux souvenirs de cette mission très *roots* dans ce pays magnifique. Enfin je remercie Laurent Millet, Clarisse Majorel et Claire Bonneville pour leur aide et pour m'avoir formée et aidée pour l'extraction et l'amplification des séquences de poissons pour la base de référence à Nouméa. Un immense merci également à Tony et Alice pour avoir partagé leur expertise sur l'ADNe, et à toute l'équipe de SPYGEN pour avoir traité au laboratoire l'ensemble des échantillons de cette thèse.

Les nombreuses collaborations réalisées au cours de cette thèse m'ont permis de développer mes compétences et de travailler avec des scientifiques de tous horizons qui m'ont transmis leur savoir-faire et leur passion. Je remercie donc Louis Bernatchez et Eric Normandeau de m'avoir accueillie à l'université de Laval, au Québec, au tout début de ma thèse, et de m'avoir accompagnée dans la conception de ma première étude. Virginie et Pierre-Edouard, je ne vous remercierai jamais assez pour votre aide, votre soutien, et la transmission de vos compétences en bio-informatique. Merci ensuite à tous les co-auteurs de mes articles, Loïc Pellissier, Nicolas Loiseau, Matthew McLean, Régis Hocdé, et tous les autres, pour votre aide, vos conseils, vos corrections et vos encouragements, et pour l'ouverture d'esprit que vous m'avez apportée.

Je tiens à remercier l'équipe BEV, passée et actuelle, pour la fantastique ambiance de travail au CEFE. Emilie, merci pour toutes les aventures ensemble (plongée, sport, spéléo..), pour ta sagesse et pour toutes les discussions et tes conseils qui m'ont permis d'avancer plus sereinement dans mes recherches. Merci à Laura, Coline, Pauline, Bobby, Thibaut, Jules… pour la bonne intégration lors de mon stage puis de mon début de thèse, pour les soirées et les randos. Et merci à Julia, Maurine, Letizia, Erwan, Manon, Morgane, pour l'ambiance chaleureuse dans le bureau 114 et pour votre soutien sur ma fin de thèse.

Une partie considérable de ma thèse s'étant déroulé à Nouméa, dans l'UMR Entropie, il m'est impossible de ne pas remercier les personnes qui ont croisé ma route et qui ont rendu chaque séjour en Nouvelle-Calédonie inoubliable. Je tiens d'abord à remercier Véronique Perrin. Merci pour ta douceur et ta joie de vivre, et pour avoir facilité chacune de mes arrivées. Merci à mes collègues de bureau, Florian, Laure, Thomas et Seb pour la meilleure ambiance de travail, si agréable et bienveillante. Merci aux différentes générations de stagiaires et doctorants que j'ai pu rencontrer (Bruce, Thib, Martin, Poé, Prisca, Federica, Léo, Romain, Marzac,

Sabine, Laure…), pour les soirées au loft, les weekends de vadrouille, la cohésion pendant le confinement, et tous les bons moments passés ensemble. Enfin, un immense merci à Laure et Bettina. Nos chemins se croisent depuis notre première rencontre à Nouméa en 2017, et je suis très heureuse de la belle amitié que nous avons construite. Merci pour tous les merveilleux moments partagés, aux 4 coins de la Nouvelle-Calédonie (et ailleurs), qui ont embelli et enrichi chacun de mes séjours.

Merci à ma famille de m'avoir soutenue tout au long de cette thèse et de mes études, et de m'avoir donné les moyens d'arriver jusque-là.

Hadrien, merci d'être à mes côtés depuis quasiment le début de ma thèse, depuis notre rencontre en Nouvelle-Calédonie en temps de covid. Merci de me soutenir, de m'écouter, de me conseiller, mais surtout d'ensoleiller mon quotidien.

Table des matières

Introduction générale

1. La biodiversité marine face aux changements globaux

1.1. Mesures de biodiversité

« La Terre abrite une extraordinaire diversité biologique, qui inclut non seulement les espèces qui habitent notre planète, mais aussi la diversité de leurs gènes, la multitude des interactions écologiques entre elles et avec leur environnement physique, et la variété des écosystèmes complexes qu'elles constituent » (Loreau, 2005). Cette définition de la biodiversité donnée par Michel Loreau à la *Conférence internationale Biodiversité, Science et Gouvernance* en 2005, implique la prise en compte de plusieurs niveaux de diversité et de leurs interactions, afin de décrire l'ensemble du monde vivant dans lequel nous évoluons. Mesurer la biodiversité, à l'aide de métriques comparables dans le temps et l'espace, est essentiel pour connaitre l'état de la biodiversité à un moment et un lieu donné, ainsi que d'être capable d'en suivre l'évolution.

La métrique la plus basique et la plus utilisée pour mesurer la diversité d'espèces est la richesse taxonomique, introduite par Mcintosh (1967). La richesse spécifique, aussi appelée richesse taxonomique, correspond simplement au nombre d'espèces, ou de taxa, présentes dans l'espace considéré. Cette diversité ne prend pas en compte les caractéristiques des espèces, mais seulement leur nombre et peut être estimée à partir des données d'inventaires taxonomiques. D'autres indices ont été introduits afin de décrire la répartition des individus et des espèces, tels que l'indice de Shannon, basé sur le nombre d'espèces et la répartition des individus au sein de ces espèces (Shannon, 1948), ou encore l'indice de Simpson, qui mesure la probabilité pour deux individus sélectionnés au hasard d'appartenir à la même espèce (Simpson, 1949). L'estimation de la richesse spécifique d'une communauté fait toujours l'objet de discussions et d'une abondante littérature, notamment sur l'utilisation d'estimateurs non-paramétriques (ne suivant pas une loi de probabilité définie) tels que ceux développés par Chao (1984).

Pour obtenir une description plus approfondie de la biodiversité, la diversité taxonomique n'est pas suffisante. Les fonctions écologiques et l'histoire évolutive sont des composantes importantes de la biodiversité qu'il est nécessaire de quantifier. La diversité phylogénétique – basée sur les distances phylogénétiques qui représentent le temps de divergence depuis le

dernier ancêtre commun entre deux espèces – est calculée comme la somme des longueurs de branches au sein d'une communauté, sur l'arbre du vivant (Figure 1.1, Rabosky et al. 2018). Cette diversité quantifie l'histoire évolutive d'une communauté.

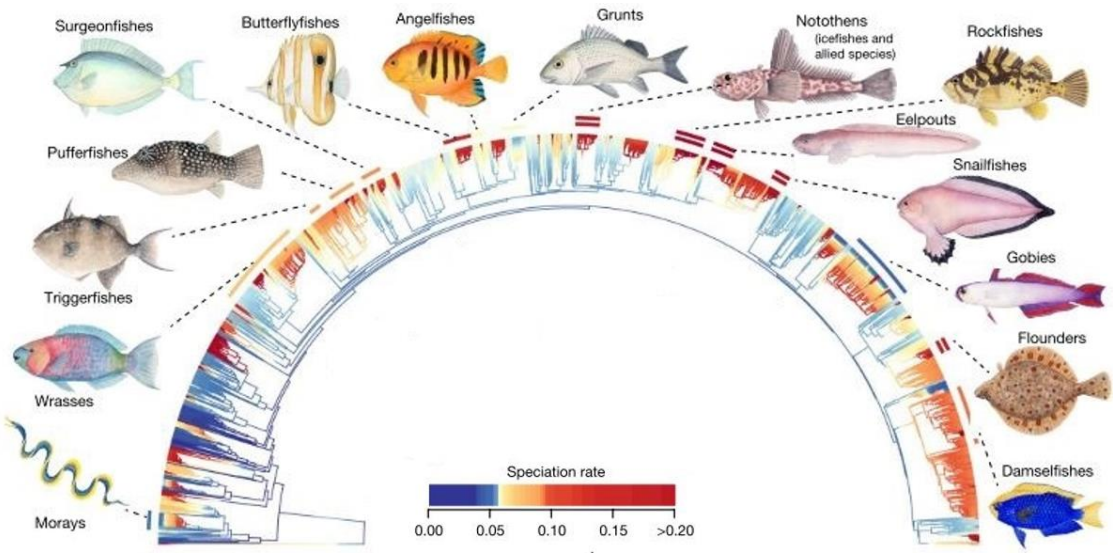


Figure 1.1. *Arbre phylogénétique et estimation des taux de spéciation, pour 5 223 espèces de poissons marins. Les clades de poissons coralliens sont identifiés par une simple bande extérieure. Les doubles bandes identifient les clades de hautes latitudes. Figure adaptée de Rabosky et al. (2018).*

La diversité fonctionnelle quantifie quant à elle la diversité de fonctions exercées par les espèces au sein d'une communauté. Ces fonctions sont décrites par des traits fonctionnels phénotypiques (ex : taille, forme…), éthologiques (ex : régime alimentaires, position dans la colonne d'eau, grégarisme…) ou encore physiologiques (ex : période d'activité, âge de maturité…). Chaque espèce ou individu est ainsi caractérisé par ses valeurs de traits qui sont considérées comme un proxy de sa niche écologique. Les valeurs de traits peuvent être représentées dans un espace multidimensionnel et les espèces proches dans cet espace sont considérées comme écologiquement proches. La dissimilarité fonctionnelle entre deux espèces peut être mesurée par la distance de Gower (Gower, 1971; Pavoine et al. 2009; Podani, 1999), qui moyenne les dissimilarités calculées pour chaque trait entre ces deux espèces. Les espèces exerçant les mêmes fonctions sont regroupées en groupes fonctionnels. De nombreuses métriques peuvent être calculées à partir de la composition en traits fonctionnels d'une communauté (Box 1), parmi lesquelles la diversité fonctionnelle, la plus utilisée, qui quantifie la répartition des espèces et leurs abondances dans l'espace fonctionnel d'une communauté (Mouillot et al. 2013).

Box 1. Métriques de diversité fonctionnelle

Diversité fonctionnelle : Répartition des espèces et de leurs abondances dans l'espace fonctionnel d'une communauté donnée.

Richesse fonctionnelle : Volume occupé par toutes les espèces d'une communauté dans l'espace fonctionnel.

Divergence fonctionnelle : Proportion de l'abondance totale supportée par les espèces avec les valeurs de traits les plus extrêmes au sein d'une communauté.

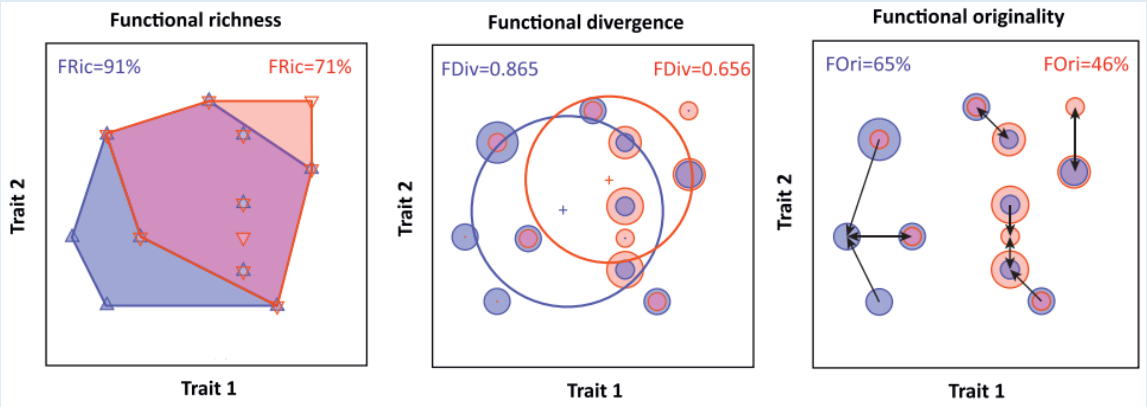
Régularité fonctionnelle : Homogénéité de la distribution et de l'abondance relative des espèces dans l'espace fonctionnel pour une communauté donnée.

Identité fonctionnelle : Valeur moyenne des traits fonctionnels, pondérée par l'abondance, pour toutes les espèces présentes dans une communauté donnée.

Originalité fonctionnelle : Isolement d'une espèce dans l'espace fonctionnel occupé par une communauté donnée.

Spécialisation fonctionnelle : Distance moyenne d'une espèce par rapport au reste du pool d'espèces dans l'espace fonctionnel.

Dissimilarité fonctionnelle : Dissimilarité dans l'espace fonctionnel occupé par deux communautés.



Schématisations de la richesse, divergence et originalité fonctionnelles. Illustrations extraites de Mouillot et al. (2013).

De même que pour la richesse spécifique, des indices ont été récemment développés pour mesurer la diversité phylogénétique et fonctionnelle, prenant également en compte l'abondance des espèces dans la communauté. Les nombres de Hill (Hill, 1973) mesurent le nombre d'espèces également abondantes (ou nombre d'espèces effectif) qui sont nécessaires pour obtenir une valeur égale d'une mesure de diversité. Ils sont de plus en plus utilisés pour

quantifier la diversité spécifique d'un assemblage et ont été récemment étendus à la diversité phylogénétique ainsi qu'à la diversité fonctionnelle. Les nombres de Hill (^{q}D) sont une famille d'indices paramétriques de diversité ne différant entre eux que par le paramètre *q* qui détermine la sensibilité à l'abondance relative des espèces. Lorsque *q* = 0, les abondances des espèces sont considérées comme toutes égales et $^{0}\boldsymbol{D} = \boldsymbol{S}$ (où D est la diversité de Hill et S est la richesse spécifique). Lorsque *q* = 1, les espèces sont pondérées proportionnellement à leurs fréquences et la mesure $^{1}\boldsymbol{D}$ peut être interprétée comme le nombre effectif d'espèces communes ou typiques dans l'assemblage et $^{1}\boldsymbol{D} = \mathbf{exp}(\boldsymbol{shannon})$. Lorsque *q* = 2, les espèces abondantes sont favorisées et les espèces rares défavorisées et $^{2}\boldsymbol{D}$ est l'inverse de l'indice de Simpson.

Chao et al. (2014) ont développé un cadre qui unifie les indices de diversité de Hill basés sur les entités taxonomiques, fonctionnelles et phylogénétiques :

- Nombre de Hill taxonomique : toutes les espèces sont traitées comme étant taxonomiquement également distinctes. Par conséquent, les nombres de Hill mesurent le nombre effectif d'entités taxonomiques.
- Nombre de Hill phylogénétique : la valeur (ou caractéristique) de l'attribut est la longueur de chaque segment de branche d'un arbre phylogénétique ; ainsi toutes les branches de longueur unitaire (ou entités phylogénétiques) sont traitées comme phylogénétiquement également distinctes. Les nombres de Hill mesurent la longueur de branche effective, en unité de longueur de branche.
- Nombre de Hill fonctionnel : la valeur de l'attribut est la distance fonctionnelle entre chaque paire d'espèces ; toutes les paires d'espèces présentant une unité de distance par paire (en tant qu'entités fonctionnelles) sont traitées comme étant fonctionnellement distinctes de manière égale. Les nombres de Hill mesurent la somme effective de la distance fonctionnelle entre les espèces, en unité de distance par paire d'espèces.

Toutes les métriques de diversité évoquées précédemment peuvent être mesurées à différentes échelles spatiales. La diversité alpha (α) est la diversité locale, mesurée à l'intérieur d'une communauté délimitée, tandis que la diversité gamma (γ) est la diversité globale ou régionale, sur l'ensemble du système étudié. La diversité bêta (β) reflète à quel point deux communautés sont différentes en termes de diversité taxonomique, fonctionnelle ou phylogénétique (Legendre & De Cáceres, 2013). Ces trois échelles de diversité sont liées par

une relation additive : la diversité γ est décomposable en la somme de la diversité α locale moyenne et de la diversité β inter-communautés (Crist & Veech, 2006).

Les indices les plus simples et probablement les plus connus pour calculer la diversité β sont les indices de Jaccard et de Sørensen. Ces indices sont basés sur la proportion d'espèces partagées entre deux communautés et la proportion d'espèces uniques à chacune des communautés. Plus la proportion d'espèces partagées est faible et plus la diversité β est élevée. Les indices de Jaccard et Sørensen se décomposent en deux types de dissimilarité (Baselga, 2010) : (i) l'emboîtement des compositions spécifiques (*nestedness*) – une communauté contient un sous-ensemble des espèces présentes dans l'autre communauté – et (ii) la substitution d'espèces (*turnover*) – remplacement d'espèces entre les deux communautés (Figure 1.2).

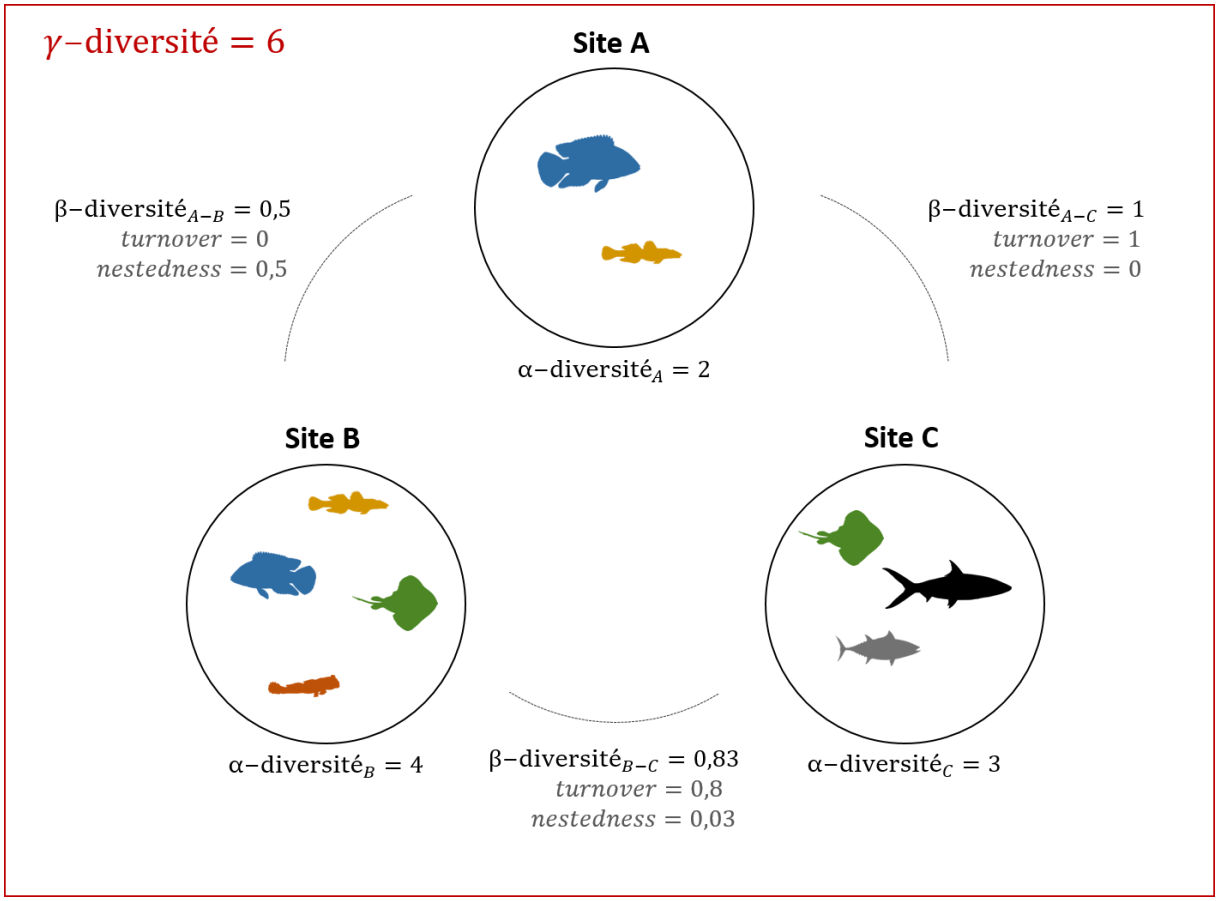


***Figure 1.2**. Schéma des diversités taxonomiques α, γ et β entre trois communautés. La β-diversité est calculée avec l'indice de Jaccard et décomposée entre emboitement (nestedness) et remplacement (turnover).*

1.2. Biodiversité des poissons marins à travers les océans

A l'échelle mondiale, les poissons représentent près de la moitié des espèces de vertébrés, avec ≃32000 espèces d'Actinoptérygiens (appartenant au groupe des poissons osseux Ostéichtyens) et ≃1200 espèces d'Elasmobranches (requins, raies, chimères) (Costello et al. 2015). Les poissons ont colonisé quasiment tous les milieux aquatiques de la planète, jusqu'aux abysses à 8000m de profondeur (Martinez et al. 2021), présentant ainsi une large variété de traits d'histoire de vie et de niches écologiques (Gerringer et al. 2017; Nelson et al. 2016). Dans le milieu marin, on dénombre ≃16000 espèces de poissons, dont 5000 à 8000 de poissons coralliens (Victor, 2015). Une large proportion de ces espèces sont des cryptobenthiques, groupes très diversifiés de poissons cryptiques et de petite taille vivant cachés dans le substrat (Brandl et al. 2018). Les poissons osseux représentent une biomasse de 0.7Gt de carbone (contre 0.06Gt pour les humains, Bar-On et al. 2018) et jouent un rôle clé dans le cycle du carbone (Wilson et al. 2009; Mariani et al. 2020) mais aussi dans la sécurité alimentaire de nombreux pays côtiers et insulaires (Batista et al. 2014; Eddy et al. 2021; Sing Wong et al. 2022).

Cependant, cette diversité de poissons n'est pas distribuée uniformément à travers les océans et les grands patrons de distribution des poissons sont étudiés depuis des décennies par les scientifiques (Bellwood & Hughes, 2001). Il est impératif de comprendre les mécanismes sous-jacents de ces patrons afin d'anticiper de potentiels changements de distribution et de diversité des poissons en réponse aux changements globaux.

Les principaux patrons observés concernent la variabilité de richesse spécifique entre les différentes régions et latitudes. La richesse spécifique des poissons suit un gradient latitudinal, avec une faible diversité aux pôles, qui augmente et atteint un pic proche de l'équateur. Ce gradient est bimodal car la richesse spécifique diminue à nouveau sur l'équateur (Chaudhary et al. 2016). Les théories et hypothèses, dont deux sont décrites ci-après, permettant d'expliquer ce gradient de richesse sont multiples, non exclusives et encore débattues aujourd'hui (Gaboriau et al. 2019). L'hypothèse de la stabilité climatique, largement supportée, stipule que les climats tropicaux, majoritaires au cours de l'histoire géologique de la terre, ont favorisé la diversification des espèces et la spéciation, en augmentant l'importance des interactions biotiques ou la vitesse d'évolution moléculaire (Mittelbach et al. 2007) et en diminuant les taux d'extinction (Fine, 2015). Ces régions tropicales ont connu une grande stabilité climatique au cours des temps géologiques et un développement de grandes surfaces d'habitats coralliens abritant une grande productivité. Cela a permis le maintien et la diversification des espèces dans

ces régions, qui sont considérées comme des sources et refuges pour la majorité des organismes (Chaudhary et al., 2016; Fine, 2015). Selon cette hypothèse de stabilité climatique, la température est significativement corrélée avec la surface d'habitat corallien disponible et donc indirectement avec la diversité de poissons (Parravicini et al. 2013). Le phénomène de conservation de niche implique que les espèces apparues dans les régions tropicales et adaptées à cette niche, peuvent se maintenir dans les tropiques actuels, mais s'éteignent dans les régions où le climat s'est modifié, réduisant de ce fait leur distribution. Seules les lignées et espèces adaptées aux climats plus froids se maintiennent dans les hautes latitudes (Romdal et al. 2013). L'hypothèse de contrainte de température stipule que les hautes températures accélèrent le métabolisme et les taux de mutation et de spéciation et donc la vitesse d'évolution des lignées (Trip et al. 2014). Par conséquent, les lignées évolueraient plus rapidement aux faibles latitudes. Certains groupes trophiques sont par ailleurs contraints par l'environnement, comme par exemple les herbivores, dont le taux de nutrition et digestion décroissent rapidement avec la diminution des températures, limitant ainsi la distribution et la diversité des poissons herbivores (Floeter et al. 2005). Cette théorie reste controversée, car une étude récente démontre que les lignées de poissons inféodées aux zones tempérées ou polaires (ex : Notothenioides, poissons plats, Liparidae) ont des taux de spéciation supérieurs dans les hautes latitudes (Rabosky et al. 2018).

La richesse spécifique des poissons marins suit un gradient longitudinal, avec un pic dans le centre du Triangle de Corail, entre l'Indonésie, les Philippines et la Papouasie-Nouvelle-Guinée (Veron et al. 2009) et une diminution uniforme à la distance au Triangle de Corail (Figure 1.3, exemple pour les Labridae). Là aussi, diverses hypothèses peuvent expliquer ce gradient de diversité. La stabilité du climat et des habitats dans cette région au cours du quaternaire en ont fait un refuge contre l'extinction et la perte d'habitat, une source d'espèces pour la recolonisation et un berceau d'évolution avec un fort taux de spéciation (Pellissier et al. 2014). La très grande surface d'habitats coralliens disponible et la continuité de ces habitats permettent d'accueillir un grand nombre d'espèces et une forte diversité génétique et donc d'augmenter les taux de spéciations sympatriques (par le cloisonnement de niches écologiques) et allopatriques (par la présence de barrières « douces » au flux de gènes) (Mellin et al. 2010; Tornabene et al. 2015). Une surface d'habitat limitée et une plus grande isolation peuvent donc mener à une plus faible colonisation et un plus fort taux d'extinction (Mora & Sale, 2011). Siqueira et al. (2021) ont mis en évidence que la proportion de planctivores dans le Triangle de Corail est largement supérieure aux autres régions. Ceci est dû à leur forte diversification

favorisée par la stabilité de l'habitat et la forte concentration de plancton, mais aussi de par la forte extinction en dehors du Triangle de Corail lors des variations climatiques au Quaternaire. Une seconde théorie permet d'expliquer ces patrons de distributions des poissons. Selon la théorie de l'effet de « milieu de domaine », une distribution aléatoire des aires de répartition des espèces produirait un pic de diversité au centre des limites géographiques de ce domaine (Colwell & Hurtt, 1994; Connolly et al. 2003). Ainsi, dans l'Indo-Pacifique, le Triangle de Corail serait proche du centre du milieu de domaine (Bellwood et al. 2012). Dans l'Atlantique, cependant, le hotspot de diversité situé dans les Caraïbes est distant du milieu de domaine et serait davantage corrélé à la température (Luiz et al. 2012).

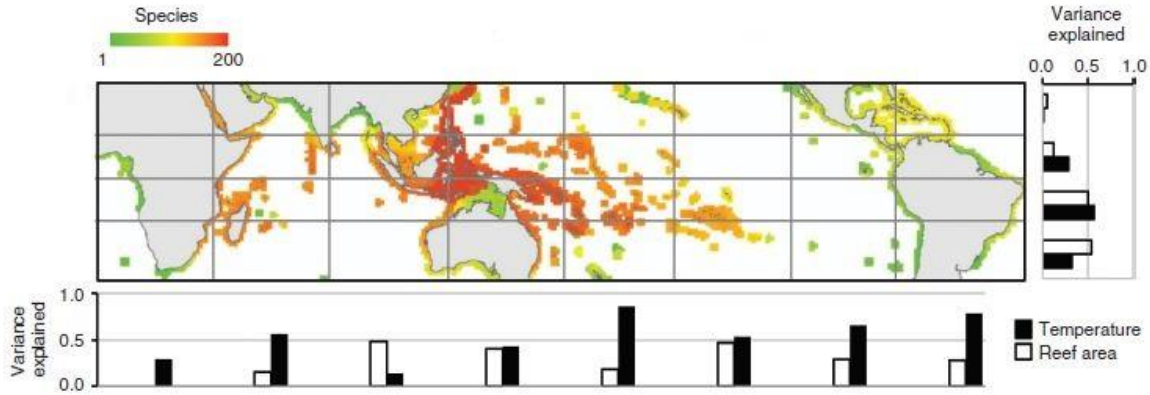


Figure 1.3. *Illustration du gradient longitudinal de richesse. Nombre d'espèces de Labridae dont les distributions se superposent, sur une grille de résolution à 100km. Le barplot représente les résulats d'un test de corrélation entre la richesse spécifique et la surface de récif coralliens et la température de surface, à différentes longitudes et latitudes. Figure extraite de Mora (2015).*

Ces deux gradients de diversité ne concernent que la richesse spécifique, mais d'autres patrons sont observés à l'échelle mondiale au regard d'autres métriques de diversité. À l'instar de la richesse taxonomique, les abondances et occurrences de chaque espèce sont variables à l'échelle globale. Rabinowitz (1981) a décrit 8 classes pour déterminer la rareté ou la banalité d'une espèce, selon sa distribution, la taille de la population et la taille de sa niche écologique. Une espèce est dite « commune » si sa distribution est large, ses populations grandes et sa niche écologique étendue. Tous les autres cas de figures font état d'au moins une forme de rareté. À l'échelle mondiale, plusieurs études ont démontré que la plupart des communautés sont composées de quelques espèces communes très abondantes et d'une majorité d'espèces « rares » représentées par très peu d'individus (Connolly et al. 2014; Enquist et al. 2019).

La composition des assemblages d'espèces diffère également entre les régions, pouvant représenter une combinaison entre emboitement et remplacement entre deux régions (Maxwell et al. 2022). Plusieurs zones sont ainsi identifiées dans les océans du monde et découpées en biorégions et provinces, qui se caractérisent par des faunes différentes, une certaine proportion d'endémisme et un environnement homogène (Spalding et al. 2007; Woolley et al. 2020). Les faunes de certaines biorégions sont distinctes taxonomiquement et phylogénétiquement des autres par la présence de barrières géographiques « dures » (ex : présence d'isthme) ou « douces » (ex : courants) (Maxwell et al. 2022). Par exemple, les assemblages de poissons dans la région Caraïbes sont très distincts des assemblages présents dans l'Indo-Pacifique, ce qui s'explique par la formation de l'Isthme de Panama durant le Pliocène, qui a formé une barrière géographique entre les océans Pacifique et Atlantique (Cowman & Bellwood, 2013; Leprieur et al. 2016; Bender et al. 2017). De même, la faune ichthyologique autour du continent Antarctique est unique et composée à ≃90% de poissons benthiques appartenant aux familles Notothenioidae, Liparidae et Zoarcidae, parmi lesquelles ≃90% des espèces sont endémiques (Eastman, 2005; Patarnello et al. 2011). Cette faune unique s'est diversifiée et adaptée à son environnement suite à l'isolement de l'Antarctique par l'ouverture du passage de Drake et le développement du courant circumpolaire (Crame, 2018). Dans l'Indo-Pacifique, les assemblages et environnements sont également découpés en biorégions et provinces, mais les faunes tropicales sont davantage similaires en raison de la connectivité historique au sein de l'archipel Indo-Australien et à la prolifération et à l'expansion des lignées de poissons depuis cette région durant le Miocène (Cowman & Bellwood, 2013). Ainsi, Bellwood & Hughes (2001) ont mis en évidence que la composition taxonomique des communautés de poissons est conservée et limitée dans une gamme étroite de configurations possibles, et que la représentation relative de chaque famille est similaire entre les sites au sein de l'Indo-Pacifique.

Cette différence de composition taxonomique (ou β-diversité taxonomique) entre communautés de différentes régions n'est pas retrouvée au niveau de la diversité fonctionnelle. En effet, les faibles valeurs de β-diversité fonctionnelle entre les différentes régions tropicales indiquent que la composition en traits fonctionnels des assemblages de poissons est similaire, même entre communautés géographiquement ou phylogénétiquement distinctes (Maxwell et al. 2022). À une échelle plus globale, McLean et al. (2021) ont identifié une structure commune de 21 combinaisons de traits fonctionnels partagées par toutes les communautés étudiées, dans des conditions environnementales variées. La composition en traits est d'autant plus similaire que les conditions environnementales des régions comparées sont proches : toutes les régions

tropicales partagent des compositions fonctionnelles semblables, et idem pour les régions tempérées, même si leurs compositions taxonomiques sont différentes. Il existe toutefois une variabilité dans la prévalence et la représentation des traits fonctionnels selon les habitats. Le long du gradient longitudinal de diversité, Parravicini et al. (2021) ont observé une plus forte représentation de poissons piscivores de grande taille sur les récifs isolés, impliquant une possible réorganisation fonctionnelle des communautés si les taux de fragmentation des habitats ne sont pas réduits. En effet, les communautés de poissons composées de traits fonctionnels sensibles aux perturbations (petites espèces, croissance rapide, corallivores) et peu redondants entre les espèces, seront plus vulnérables aux changements futurs (McLean et al. 2019). Or, au centre de l'Indo-Pacifique, un tiers des entités fonctionnelles sont représentées par plusieurs espèces et sont donc peu vulnérables, mais un tiers des entités fonctionnelles représentées par une seule espèce sont très vulnérables aux perturbations climatiques et humaines (Mouillot et al. 2014). De nouveaux indices développés récemment mesurent la rareté fonctionnelle selon la rareté des espèces et la distinction de leurs traits (Violle et al. 2017). À l'échelle globale, la rareté fonctionnelle est proportionnellement plus élevée dans les zones tempérées pour les poissons osseux et les Chondrichtyens (Figure 1.4, Trindade-Santos et al. 2022). Cela pourrait rejoindre la théorie de Rabosky et al. (2018) qui ont mesuré un taux de diversification plus élevé aux hautes latitudes. Ces indices de rareté taxonomique et rareté fonctionnelle peuvent être intégrés dans les plans de conservation (Albuquerque & Astudillo-Scalia, 2020).

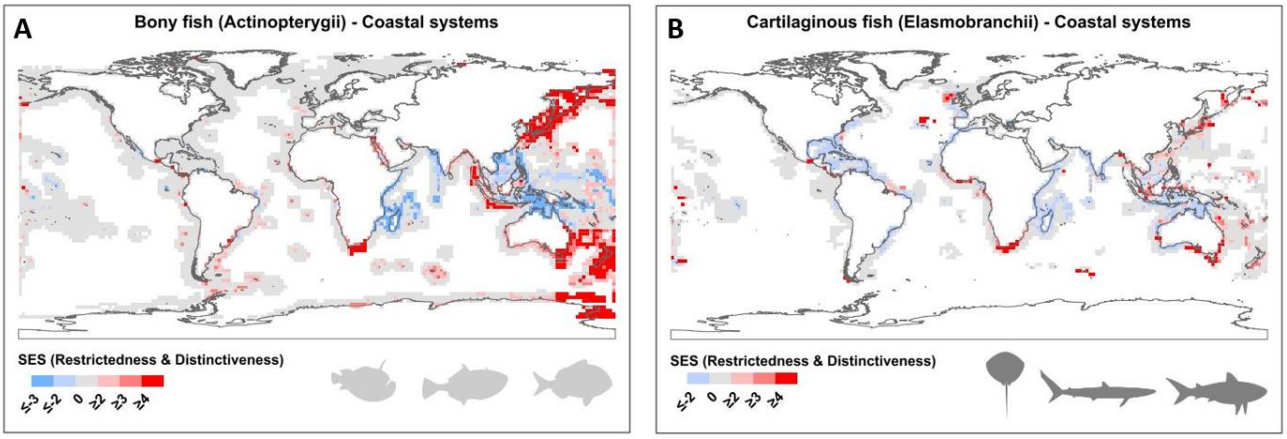


Figure 1.4. *Biogéographie de la rareté (taxonomique et fonctionnelle) des poissons osseux (A) et des poissons cartilagineux (B). Tailles d'effets standardisées, en contrôlant par la richesse spécifique de chaque cellule. Le rouge indique une rareté supérieure à celle attendue au hasard et le bleu indique une rareté inférieure à celle attendue au hasard. Figure adaptée de Trindade-Santos et al. (2022).*

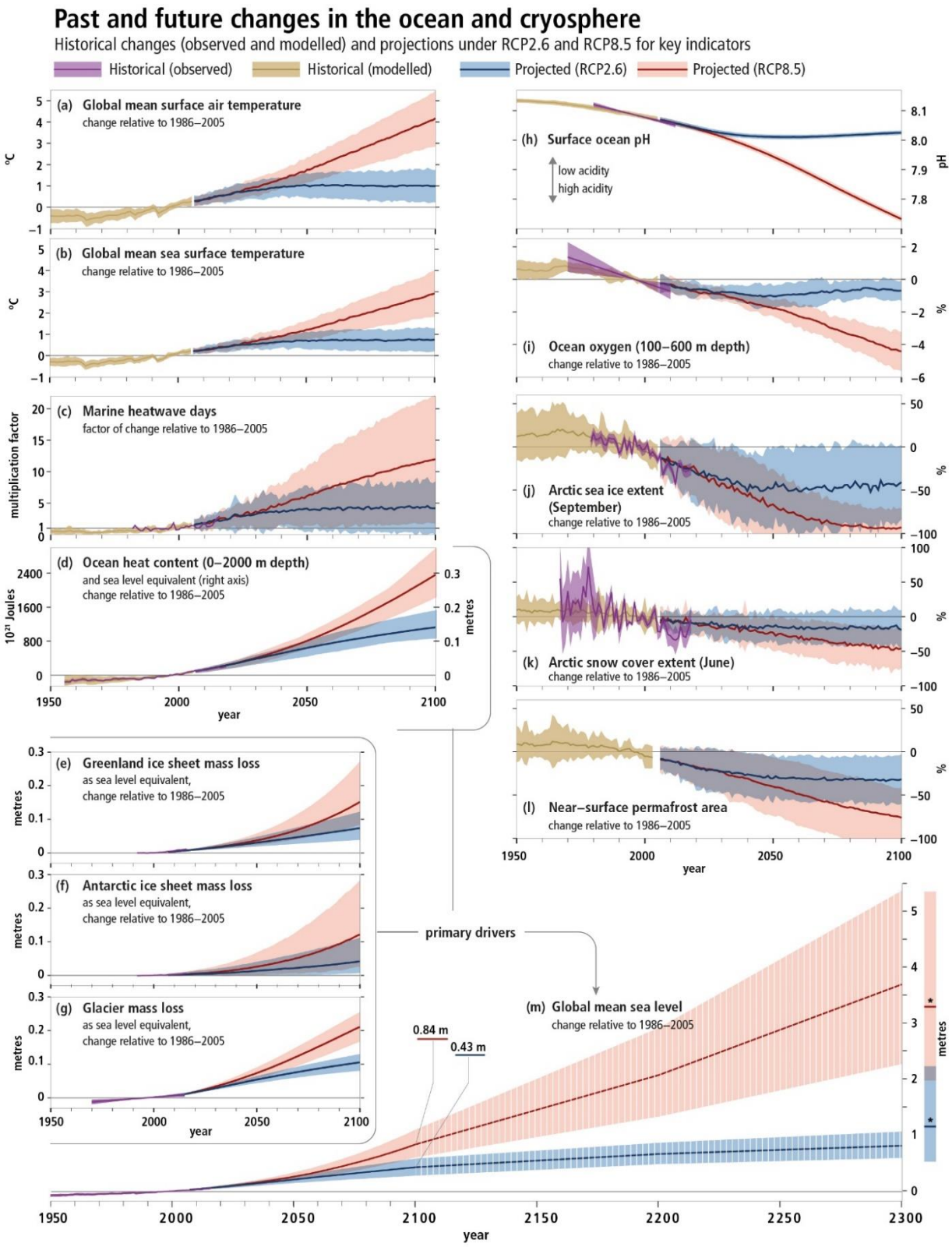
1.3. Influence des facteurs environnementaux et anthropiques sur la biodiversité

La biodiversité et les patrons de distribution sont cependant de plus en plus menacés par les changements actuels qui affectent notre planète (Díaz et al. 2019). En effet, les changements climatiques et les pressions anthropiques ont un impact grandissant sur la biodiversité (O'Hara et al. 2021; Worm & Lotze, 2021). Les taux d'extinctions sont aujourd'hui entre 10 et 1000 fois supérieurs à ceux de l'ère préindustrielle (Ceballos et al. 2020). La perte de biodiversité est particulièrement sévère dans les milieux d'eau douce, soumis aux pressions anthropiques depuis de longues périodes. Cependant, les océans ont connu une récente défaunation dans le milieu marin et la perte de biodiversité dans les océans continue de s'accélérer (Young et al. 2016). Les extinctions totales sont rares dans le milieu marin pour les poissons osseux, dont moins de 10% sont menacés, mais sont plus probables pour les espèces à longue histoire de vie, comme les requins dont ≃30% des espèces sont menacées, ou les espèces très localisées (Chichorro et al. 2019; Díaz et al. 2019). Actuellement, une seule espèce de poisson marin a été déclarée éteinte, en 2018, le poisson-main lisse (*Sympterichthys unipennis,* Stuart-smith et al. 2020), mais de nombreuses autres espèces s'éteignent localement, comme les poissons-scie (Yan et al. 2021).

Le réchauffement des températures est la principale conséquence du changement climatique et de l'augmentation du CO_2 atmosphérique, mais il s'accompagne également de l'élévation du niveau de la mer, de l'altération des habitats, de la modification de la circulation des courants et de la stratification, ou encore de l'acidification des océans (Worm & Lotze, 2021, Figure 1.5). Ces changements combinés, induits et aggravés par les pressions humaines, impactent et modifient la distribution de la biodiversité dans les océans. Les températures élevées entrainent généralement une accélération de la reproduction et de la croissance des poissons. Cependant, une augmentation de la température de l'océan au-delà des limites thermiques des espèces peut entrainer un ralentissement des fonctions vitales et la diminution des chances de survie de ces espèces (Zarco-Perello et al. 2012).

Figure 1.5. (Page suivante). *Changements historiques observés et modélisés dans l'océan et la cryosphère depuis 1950 et projections des changements futurs pour des scénarios d'émissions de gaz à effet de serre faibles (RCP2.6) et élevées (RCP8.5). (a) Evolution de la température moyenne de l'air en surface, (b) Evolution de la température moyenne de la mer en surface, (c) Variation des jours de vague de chaleur océanique, (d) Contenu thermique des océans et evolution du niveau de la mer associé, (e,f) Perte de masse de la calotte glaciaire du Groenland*

et de l'Antarctique, (g) Perte de masse glaciaire, (h) pH de surface moyen, (i) Variation moyenne de l'oxygène dans les océans, (j) Evolution de l'étendue de la banquise arctique en septembre, (k) Evolution de la couverture de neige arctique en juin, (l) Modification de la zone de pergélisol en surface dans l'hémisphère Nord et (m) Evolution du niveau moyen mondial de la mer. Figure adaptée du rapport de l'IPCC (2019).



Le succès de la reproduction de nombreuses espèces dépend de la température. Si cette dernière est trop élevée, cela peut réduire le nombre de couples, le nombre et la taille des œufs, ainsi que le nombre et la survie des larves. Ce sont les conséquences d'un taux d'oxygène insuffisant pour maintenir un métabolisme de base (Pörtner and Knust, 2007; Donelson et al. 2010). Les populations d'espèces aux pôles et à l'équateur sont plus vulnérables au réchauffement car les températures maximales d'été sont proches de leurs maximums tolérés (Enzor et al. 2013; Rummer et al. 2014; Rodgers et al. 2018). Le réchauffement climatique induit donc une migration des espèces, généralement vers les plus hautes latitudes, pour suivre leurs niches climatiques (Pinsky et al. 2020). On observe en effet un décalage de l'aire de distribution des espèces dans l'océan 6 fois plus rapide que sur terre, dû à la rapidité des changements climatiques dans les océans (Lenoir et al. 2020). Les espèces qui ne peuvent pas migrer en dehors de leur niche, ou s'adapter aux nouvelles conditions climatiques, seront plus vulnérables et menacées par des extinctions. Les écosystèmes tempérés (ex : Méditerranée, Australie, Japon) subissent des processus de tropicalisation (augmentation de l'abondance d'espèces adaptées aux eaux chaudes) et de déboréalisation (diminution de l'abondance d'espèces adaptées aux eaux froides, McLean, Mouillot, et al. 2021). De manière similaire, le réchauffement de l'Arctique et la modification de la circulation océanique entrainent une atlantification de l'Arctique : une transition physique et écologique des eaux arctiques, incluant un mélange renforcé de la couche supérieure de l'océan, un refroidissement atmosphérique diminué, une expansion vers le nord des espèces boréales et l'arrivée de nouvelles espèces pélagiques (Ingvaldsen et al. 2021). L'impact du réchauffement climatique sur les espèces vulnérables entraine la dominance d'espèces adaptées et la convergence des traits fonctionnels (petite taille, croissance rapide, habitat pélagique, haute tolérance thermique, McLean, Mouillot, et al. 2019).

L'augmentation du CO_2 atmosphérique entraine une augmentation du CO_2 dissous, puis une diminution du pH, ce qui résulte en l'acidification des océans. Cette acidification entraine plusieurs conséquences : i) sur la physiologie directe des poissons en augmentant les demandes énergétiques nécessaires au maintien d'un statut acido-basique normal, en altérant l'odorat, le comportement et les interactions proie-prédateur, ainsi que la sélection d'habitat (Munday et al. 2012; Munday et al. 2013), et ii) sur l'habitat des poissons, en diminuant la survie des coraux et dégradant la surface et la complexité topographique des habitats coralliens. La perte d'habitat affectera 60 à 70% des poissons coralliens, principalement les espèces spécialistes de cet habitat

(Stuart-Smith et al. 2021) et entrainera un changement de structure de l'écosystème (Fontoura et al. 2020; Magel et al. 2020; Stuart-Smith et al. 2022).

Les pressions anthropiques sur les écosystèmes marins ont grandement augmenté depuis le début de l'Anthropocène et continuent toujours de s'étendre et s'amplifier (Figure 1.6., Halpern et al. 2019; O'Hara et al. 2021). L'impact anthropique le plus important sur la diversité des poissons est l'exploitation directe par la pêche (O'Hara et al. 2021). En effet, 55% de la surface des océans est exploitée, en grande partie non durablement, car 60% des stocks de poissons sont exploités à pleine capacité et 33% sont surexploités (FAO, 2018).

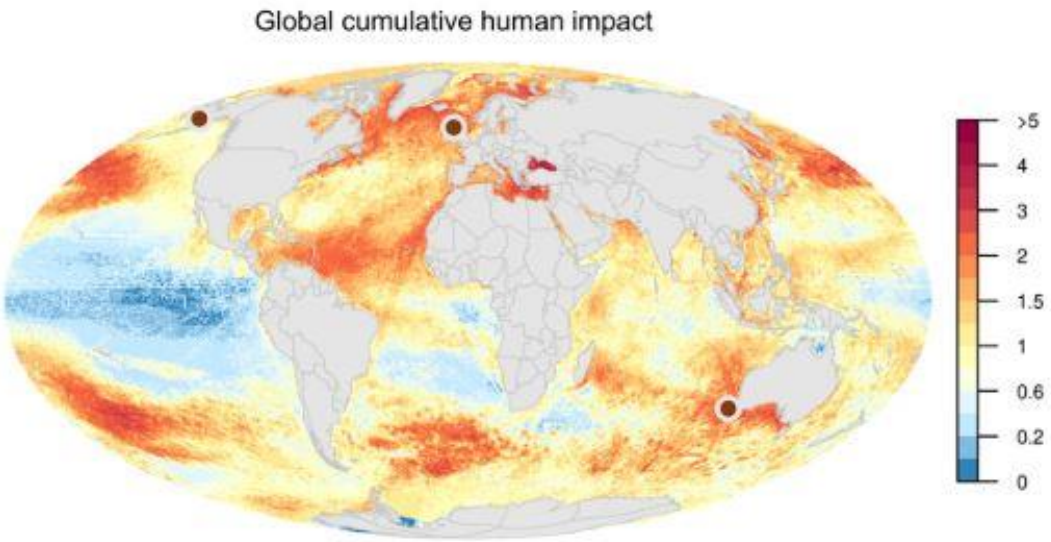


Figure 1.6. *Impact humain cumulé, calculé à partir de 14 pressions anthropiques (catégorisées en pêche, pollution, transport, réchauffement, acidification) sur l'année 2013. Figure extraite de Halpern et al. (2019).*

La pêche cible certains groupes de poissons, partageant souvent les mêmes traits fonctionnels (espèces de grande taille, mobiles, carnivores). Ces espèces ont des maturités tardives et de faibles taux de reproduction, ce qui les rend plus vulnérables à l'extinction si elles ne sont pas pêchées de façon durable (Sadovy de Mitcheson et al. 2013; Ceretta et al. 2020). La pêche ciblée de certains groupes et certaines tailles va entrainer une diminution de la diversité taxonomique, de l'abondance, mais aussi de la taille des individus restants (Tu et al. 2018; Lin et al. 2022). La diversité fonctionnelle des communautés est également diminuée par la pêche, en supprimant les espèces de grandes tailles et certains types de régimes alimentaires. Cependant, la perte de diversité fonctionnelle peut être amoindrie par la richesse taxonomique et la redondance des traits au sein de la communauté (Martins et al. 2012; McLean, Auber, et al. 2019). En mer Méditerranée, par exemple, une simulation d'extinction de 40 espèces de

poissons mène à une diminution de diversité fonctionnelle de 3% grâce à la forte redondance fonctionnelle entre espèces et à une diminution de 13% de la diversité phylogénétique (Albouy et al. 2015).

Les Elasmobranches sont particulièrement impactés par les activités humaines. Plus de 37% des espèces de requins et raies sont menacées et certaines espèces ont perdu plus de 99% de leurs effectifs (Dulvy et al. 2021; Pacoureau et al. 2021). L'abondance des requins océaniques a diminué de 71% entre 1970 et 2018 (Figure 1.7), principalement en raison de la surpêche et de l'augmentation de l'utilisation de palangres et seines, engins qui capturent le plus de requins océaniques (Pacoureau et al. 2021). La surpêche entraine un déclin mondial des populations de requins et autres gros poissons (ex : mérous) jusqu'à des extinctions locales (Sadovy de Mitcheson et al. 2013; MacNeil et al. 2020). Si toutes les espèces menacées de requins venaient à s'éteindre, on observerait une perte de diversité fonctionnelle de 80% (Pimiento et al. 2020). Si l'espèce n'est pas encore éteinte, il se peut que sa fonction le soit. On parle alors d'extinction fonctionnelle : l'espèce a une abondance trop faible dans la communauté pour remplir son rôle (Säterberg et al. 2013). Ce fut notamment le cas pour certaines espèces de requins, ou pour le perroquet à bosse (*Bolbometopon muricatum*), poisson bioérodeur autrefois abondant sur les récifs coralliens de l'Indo-Pacifique. La disparition des grands prédateurs entraine une modification de la chaine trophique et des interactions entre espèces, ce qui finit par déséquilibrer l'écosystème (Dulvy et al. 2021).

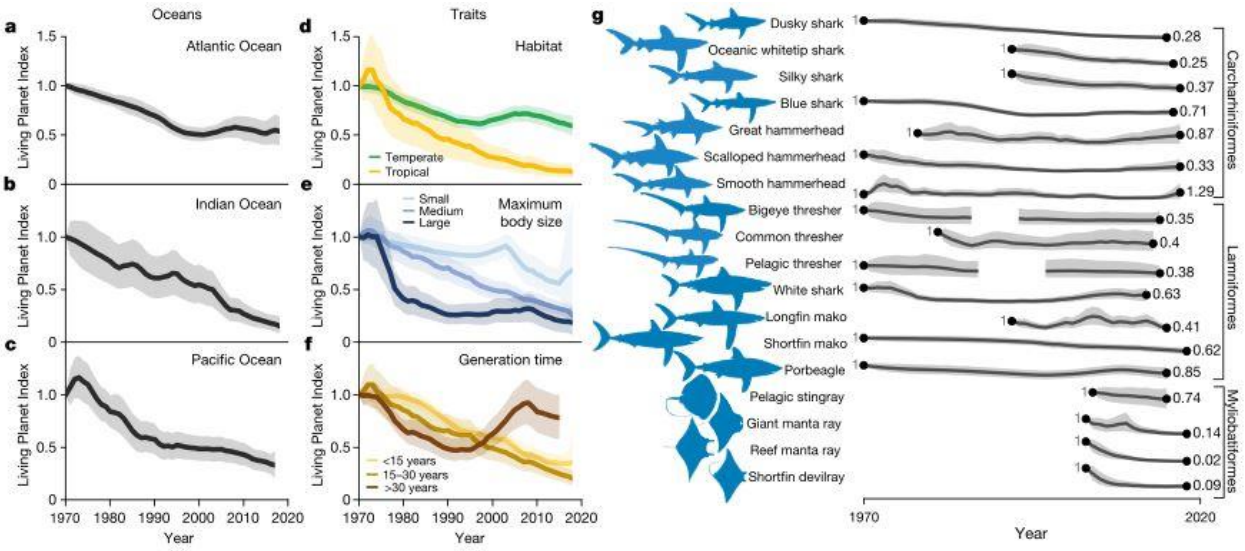


Figure 1.7. *Indice « Planète Vivante » (Living planet index) pour 18 espèces de requins océaniques, de 1970 à 2018, selon les océans et les traits : a) océan Atlantique, b) océan Indien, c) océan Pacifique, d) zone géographique, e) taille corporelle, d) temps de génération, f) détail pour chaque espèce. Figure extraite de Pacoureau et al. (2021).*

L'impact de la pêche est également très important pour d'autres animaux marins, notamment à travers les prises accessoires de mammifères marins, requins, tortues ou oiseaux (Roberson et al. 2022). Les engins de pêche tels que les dragues ou les chaluts entrainent une dégradation de l'habitat, qui affecte encore davantage les écosystèmes et les communautés marines (Das, 2018).

Outre la pêche, de nombreuses autres pressions anthropiques affectent la biodiversité marine (Figure 1.8). L'urbanisation des littoraux et le développement des villes côtières entrainent une destruction des habitats côtiers, comme les récifs, mangroves ou herbiers qui abritent une faune très diverse (Sharma et al. 2022; Turschwell et al., 2021) et sont essentiels pour le recrutement des juvéniles (Mercader et al. 2019). Le développement de l'activité anthropique s'accompagne également d'une augmentation des apports terrestres et chimiques dans les écosystèmes marins côtiers. Une augmentation de la matière particulaire en suspension entraine une altération du comportement des poissons, de leurs mouvements et de leur recherche de nourriture, une réduction de la vision et de la chimio-réception, une extension de la phase larvaire et une augmentation de la mortalité (DeMartini et al. 2013).

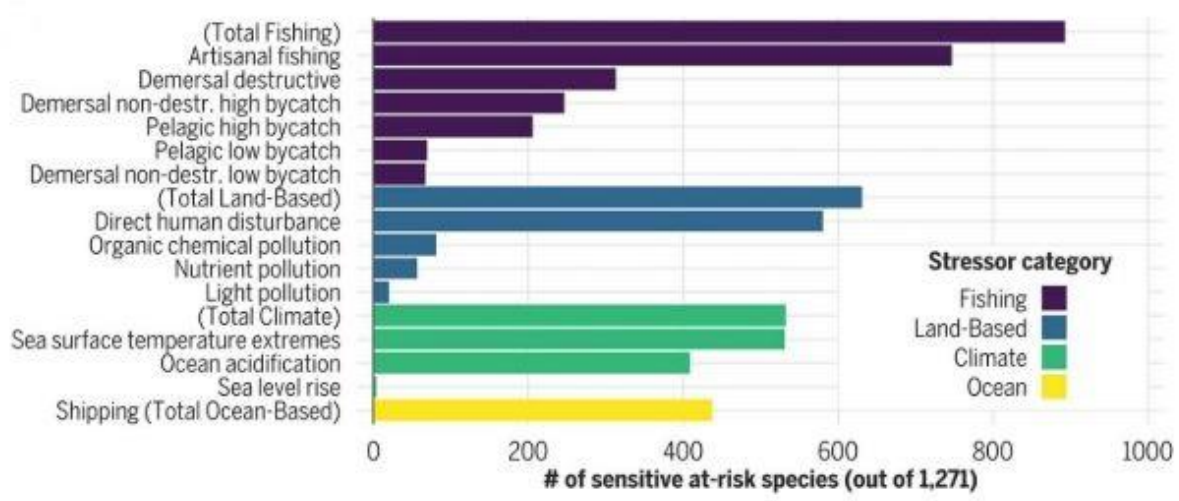


Figure 1.8. *Nombres d'espèces classées comme sensibles à chaque pression anthropique, parmi les 1271 espèces de poissons menacées étudiées. Figure extraite de O'Hara et al. (2021).*

L'apport de nutriments (azote et phosphore) peut conduire à un surdéveloppement des algues, qui entrent en compétition avec les coraux. Cette eutrophisation peut s'aggraver jusqu'à une anoxie dans le milieu, qui se transforme alors en « *dead zone* » (Monteil et al. 2020). Enfin, les pesticides et produits chimiques émis par l'agriculture et l'industrie et transportés par les réseaux fluviaux jusque dans l'océan, vont s'accumuler dans la chaine trophique et altérer les

fonctions reproductives et immunitaires, ainsi que le comportement des poissons (Brodie et al. 2012).

A l'échelle globale, la perte de diversité γ (diminution du nombre d'espèces) est bien documentée et ce pour plusieurs groupes taxonomiques. Cependant, à l'échelle locale, la perte d'espèces menacées ou fragiles peut être compensée par l'arrivée d'autres espèces plus résistantes ou adaptées. Ainsi la diversité α reste constante, mais la diversité β diminue du fait d'une homogénéisation des communautés, avec des espèces généralistes, non-indigènes (Stuart-Smith et al. 2021). Si le remplacement ou la perte d'espèces s'accompagne de la perte de fonctions rares, alors l'impact sur l'écosystème peut être très important et se refléter par une perte de productivité (Delalandre et al. 2022).

1.4. Zones refuges et aires marines protégées

En réponse à ces pressions environnementales et anthropiques, certaines communautés composées d'espèces résistantes et à rétablissement rapide peuvent se montrer résilientes face à des perturbations intermédiaires, c'est-à-dire qu'elles pourront maintenir leurs fonctions, leurs processus et leur structure (McLeod et al. 2021). Cependant, la grande majorité des espèces et communautés affectées par ces pressions et perturbations devront trouver refuge dans des zones reculées, éloignées des pressions environnementales et anthropiques, dans les profondeurs, ou dans des zones protégées où les perturbations sont contrôlées. En effet, plusieurs études à large échelle ont montré que l'éloignement à l'homme, et plus particulièrement aux marchés aux poissons et aux ports de débarquement, est un fort prédicteur de la biomasse des poissons sur les récifs coralliens, mais aussi de la taille des poissons et de l'abondance des prédateurs tels que les requins (Maire et al. 2016; Letessier et al. 2019). La biomasse de poissons sur les récifs éloignés (c.-à-d. à plus de 12h de navigation des marchés aux poissons) est supérieure à 500kg.ha^{-1}, tandis que proche des zones impactées par l'homme, la biomasse ne dépasse que rarement 100kg.ha^{-1} (Maire et al. 2016). Letessier et al. (2019) ont montré que la taille des individus et l'abondance des requins augmentent abruptement sur les récifs à plus de 1250 km des marchés aux poissons. Ainsi il semblerait que les sites éloignés et peu accessibles hébergent des communautés de poissons peu impactées. Cependant, ces récifs éloignés souffriront aussi des effets du changement climatique et peuvent donc être vulnérables à court terme (Baumann et al. 2022; Brown et al. 2022).

A proximité des zones impactées par l'homme (polluées, urbanisées ou pêchées), certaines populations résiduelles d'espèces menacées persistent à de plus grandes profondeurs. C'est le cas d'espèces ciblées par la pêche, qui déplacent leur niche en profondeur (Frank et al. 2018). Les écosystèmes mésophotiques, entre 30 et 150m de profondeur, peuvent devenir des refuges pour les espèces exploitées par la pêche et pour les espèces menacées sur les récifs peu profonds (Goetze et al. 2011; Soares et al. 2020). Les récifs mésophotiques tempérés peuvent également agir comme refuges lors des canicules marines (Giraldo-Ospina et al. 2020), mais à long terme, ces écosystèmes mésophotiques seront également touchés par le réchauffement climatique (Brito-Morales et al. 2020). De plus, la zone mésophotique ne peut pas servir de refuge à tous les groupes fonctionnels. Parmi les poissons herbivores par exemple, seules les espèces généralistes ou se nourrissant de macroalgues peuvent survivre à de plus grandes profondeurs (30-75m), alors que la majorité des espèces herbivores ne peuvent subsister que sur des récifs à moins de 30m de profondeur (Cure et al. 2021).

En haute mer, loin de toutes terres, il existe d'autres refuges et hotspots de biodiversité : les monts sous-marins. Il s'agit de montagnes immergées qui s'élèvent à plus de 1000 m au-dessus du fond océanique, sans percer la surface. Ces reliefs sont présents dans tous les océans et représentent une surface cumulée aussi grande que celle de l'Europe (Yesson et al. 2011), mais seulement 0,002% de ces monts ont été étudiés à des fins scientifiques (Rogers, 2018). À la différence des récifs isolés, ces habitats sont situés en plein océan, sont soumis aux courants et aux *up-welling* (remontée de courants froids profonds) et ne bénéficient pas des apports terrestres. Ils abritent cependant des habitats très riches, dans plusieurs gammes de profondeur. Les quelques études réalisées sur les monts sous-marins ont découvert qu'ils constituaient des hotspots pour la faune benthique et la vie fixée, telles que des coraux profonds, éponges, bryozoaires et crinoïdes et devraient ainsi être considérés comme des écosystèmes marins vulnérables (Figure 1.9, Watling & Auster, 2017).

Fiori et al. (2016) ont observé que les monts sous-marins exerçaient une forte attraction sur les mammifères marins, qui sont plus abondants dans un rayon de 10 à 20 km autours des monts. D'autres études récentes ont montré que les monts sous-marins peu profonds sont des refuges pour les prédateurs tels que les requins, thons, carangues, espadons (Morato et al. 2010; Letessier et al. 2019) et abritent une plus forte productivité (Campanella et al. 2021). Ces sommets jouent également le rôle de tremplin (*stepping stones*) pour la dispersion et diversification de certaines espèces de poissons (Mazzei et al. 2021; Simon et al. 2021). Plusieurs études ont également démontré que les monts sous-marins abritaient des

communautés de poissons fortement structurées en fonction de la profondeur (McClain & Lundsten, 2015; Muff et al. 2022).

Figure 1.9. *Représentation schématique d'un mont sous-marin et des communautés vivant sur son sommet et ses pentes. Illustration des potentiels courants influençant la colonisation et la dispersion d'individus entre les zones pélagiques de surface et benthiques. Figure extraite d'un rapport IUCN (2019).*

Les monts sous-marins, hébergeant de grandes abondances de poissons, sont les cibles des pêcheries industrielles qui menacent cette biodiversité. En effet, plus de 3 millions de tonnes de poissons sont prélevées chaque année sur les monts sous-marins, par des chaluts ou palangres qui dégradent le substrat et réalisent de nombreuses prises accessoires (Pitcher & Lam, 2014; Williams et al. 2020; Kerry et al. 2022).

Les différents refuges pour la biodiversité marine sont tout de même vulnérables aux pressions anthropiques et aux changements climatiques (Ariza et al. 2022). Il y a donc une

nécessité de protéger l'océan pour sa biodiversité et pour mitiger les effets du changement climatique, mais aussi pour la sécurité alimentaire qu'il procure grâce aux ressources vivrières (Sala et al. 2021).

Les aires marines protégées (AMP) sont le principal outil pour la protection et conservation de la biodiversité marine et pour amoindrir les impacts anthropogéniques (Woodcock et al. 2017). Les AMP sont définies par l'UICN comme « un espace géographique clairement défini, reconnu, dédié et géré par des moyens légaux ou autres moyens efficaces, pour assurer la conservation à long terme de la nature avec les services écosystémiques et les valeurs culturelles associés ». Depuis le début du 21ème siècle, la couverture des AMP a rapidement augmenté, passant de 2.5% en 2010 à 8.1% en 2022, mais cela reste insuffisant pour atteindre l'objectif mondial de 30% des océans protégés en 2030 (CBD, 2021; Dinerstein et al. 2019; Maxwell et al, 2020). De plus, il existe une large diversité de taille et de types de protection au sein des AMP, dont les efficacités varient. De nombreuses AMP sont proposées mais jamais implémentées et très peu sont strictement protégées (2.7% en 2021, Figure 1.10, Sala et al. 2021).

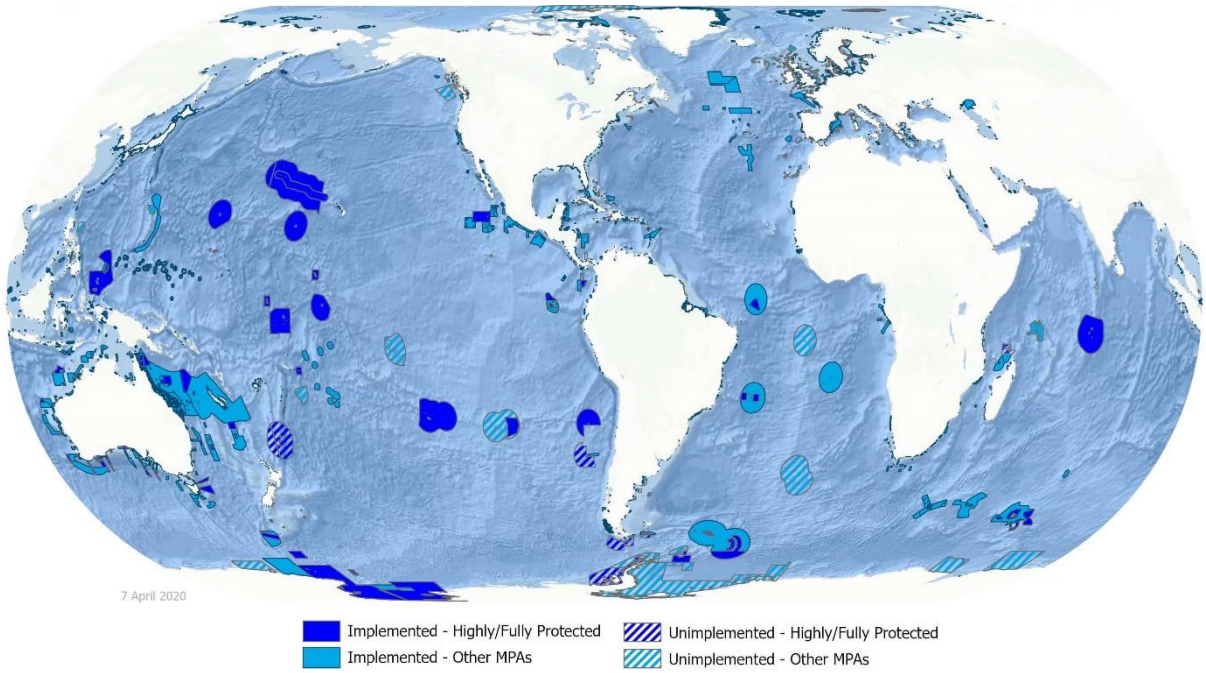


Figure 1.10. *Répartition des aires marines protégées en 2020 : réserves marines fortement protégées en bleu foncé (2.5% de couverture), autres aires protégées en bleu clair. Données extraites de Marine Protection Atlas (mpatlas.org).*

À l'échelle de la zone économique exclusive française, deuxième plus large au monde, 33.7% des eaux sont couvertes par des AMP, mais seulement 1.6% sont strictement et fortement

protégées (= réserves marines) (Claudet et al. 2021). Or, les réserves marines, AMP où toute activité humaine est interdite (Costello, 2014), sont les outils les plus efficaces pour protéger la biodiversité (Sala & Giakoumi, 2018). La mer Méditerranée, hotspot de diversité, n'est couverte qu'à 6% par des AMP et seulement à 0,23% par des réserves (Claudet et al. 2020).

Les bénéfices les plus documentés et reconnus des réserves consistent en l'augmentation de l'abondance, de la densité et de la biomasse des espèces exploitées (Giakoumi et al. 2017; Sala & Giakoumi, 2018; McClanahan, 2021). Les réserves permettent donc aux écosystèmes de se restaurer et de rétablir leur fonctionnement. Les prédateurs constituent le groupe qui bénéficie le plus de cette protection (Rojo et al. 2021). L'effet des réserves peut être variable sur les autres espèces, dont certaines restent neutres vis-à-vis de la protection, tandis que d'autres bénéficient de niveaux intermédiaires de protection ou de perturbations (Boulanger et al. 2021; Loiseau et al. 2021). L'effet des réserves marines sur la richesse spécifique reste discuté et les résultats sont contrastés entre les études : Giakoumi et al. (2017) trouvent un effet positif des AMP, Claudet et al. (2008) et Loiseau et al. (2021) ne trouvent pas d'effet significatif, tandis que Boulanger et al. (2021) trouvent un effet négatif. Cependant, les réserves procurent des bénéfices au-delà de leurs limites spatiales, par « débordement » (*spillover*) des populations de poissons aux alentours de la réserve (Di Lorenzo et al. 2020). Cet outil de conservation est donc également profitable pour les pêcheries (Cabral et al. 2019; Medoff et al. 2022). Une réserve seule et isolée n'aurait que peu de bénéfices, c'est pourquoi il est essentiel de développer un réseau de réserves ayant une connectivité importante, afin de préserver les processus qui maintiennent la biodiversité (Manel et al. 2019; Dedrick et al. 2021). Une review de la littérature par Jacquemont et al. (2022) démontre que les réserves marines, fortement protégées, augmentent la séquestration du carbone, et sont un bon outil pour l'atténuation des changements climatiques et l'adaptation des systèmes socio-écologiques.

Actuellement, on observe néanmoins un décalage entre les zones protégées et les hotspots de biodiversité et de productivité (Lindegren et al. 2018; Bellwood et al. 2019; Mouton et al. 2022). Par exemple, seulement 2% des monts sous-marins sont inclus dans des AMP (Yesson et al. 2011; Letessier et al. 2019). La création d'AMP se base principalement sur l'abondance d'espèces ou la biomasse, mais peu incluent les diversités fonctionnelles et phylogénétiques. Les hotspots de rareté fonctionnelle chez les poissons osseux sont couverts à 47% par des AMP, tandis que 67% des hotspots de rareté fonctionnelle chez les chondrichtyens sont protégés (Trindade-Santos et al. 2022). L'intégration plus fréquente de la diversité fonctionnelle et phylogénétique dans la planification spatiale est donc nécessaire (Henriques et al. 2020). Afin

de protéger également les refuges et hotspots situés dans les plus grandes profondeurs, il est primordial d'inclure cette dimension verticale dans les planifications de conservation (Levin et al. 2018; Venegas-Li et al. 2018; Brito-Morales et al. 2022; Doxa et al. 2022).

2. Recensement et suivi des communautés de poissons

2.1. Méthodes conventionnelles de suivi

Un recensement fiable et exhaustif des communautés de poissons est nécessaire pour décrire et comprendre les patrons de distribution des espèces et suivre l'évolution spatio-temporelle de ces dernières en réponse aux facteurs externes. La plupart des méthodes dites conventionnelles de recensement des poissons sont basées sur une identification visuelle des espèces. Ces méthodes nécessitent donc une expertise taxonomique et un échantillonnage conséquent pour parvenir à une description fiable et exhaustive des espèces présentes. Il existe une grande variabilité de méthodes conventionnelles, chacune adaptée aux caractéristiques du milieu ciblé (Figure 1.11).

En eau douce, la pêche électrique est communément utilisée pour caractériser les communautés. Cette méthode consiste à diffuser un courant électrique dans le cours d'eau étudié, paralysant temporairement les individus, qui sont ainsi facilement capturables à la surface (Bain et al. 1985). La pêche électrique peut cependant stresser ou blesser les animaux, augmenter la mortalité ou diminuer la survie des œufs (Huysman et al. 2018; Maahs et al. 2018). De plus, l'efficacité de cette méthode dépend de la conductivité et de la profondeur du cours d'eau. Cette méthode ayant une action très localisée, elle est moins adaptée pour la détection d'espèces rares ou benthiques (Allard et al. 2014; Pottier et al. 2020). L'utilisation de filets disposés en travers du cours d'eau permet de détecter davantage d'espèces mais est également très destructrice pour les communautés et pour certaines espèces emblématiques (Barko et al. 2004; Kelkar & Dey, 2020). Des inventaires plus exhaustifs peuvent être obtenus par l'utilisation de poisons tels que la roténone (Allard et al. 2016), mais ces méthodes sont évidemment très délétères pour l'environnement (Finlayson et al. 2010).

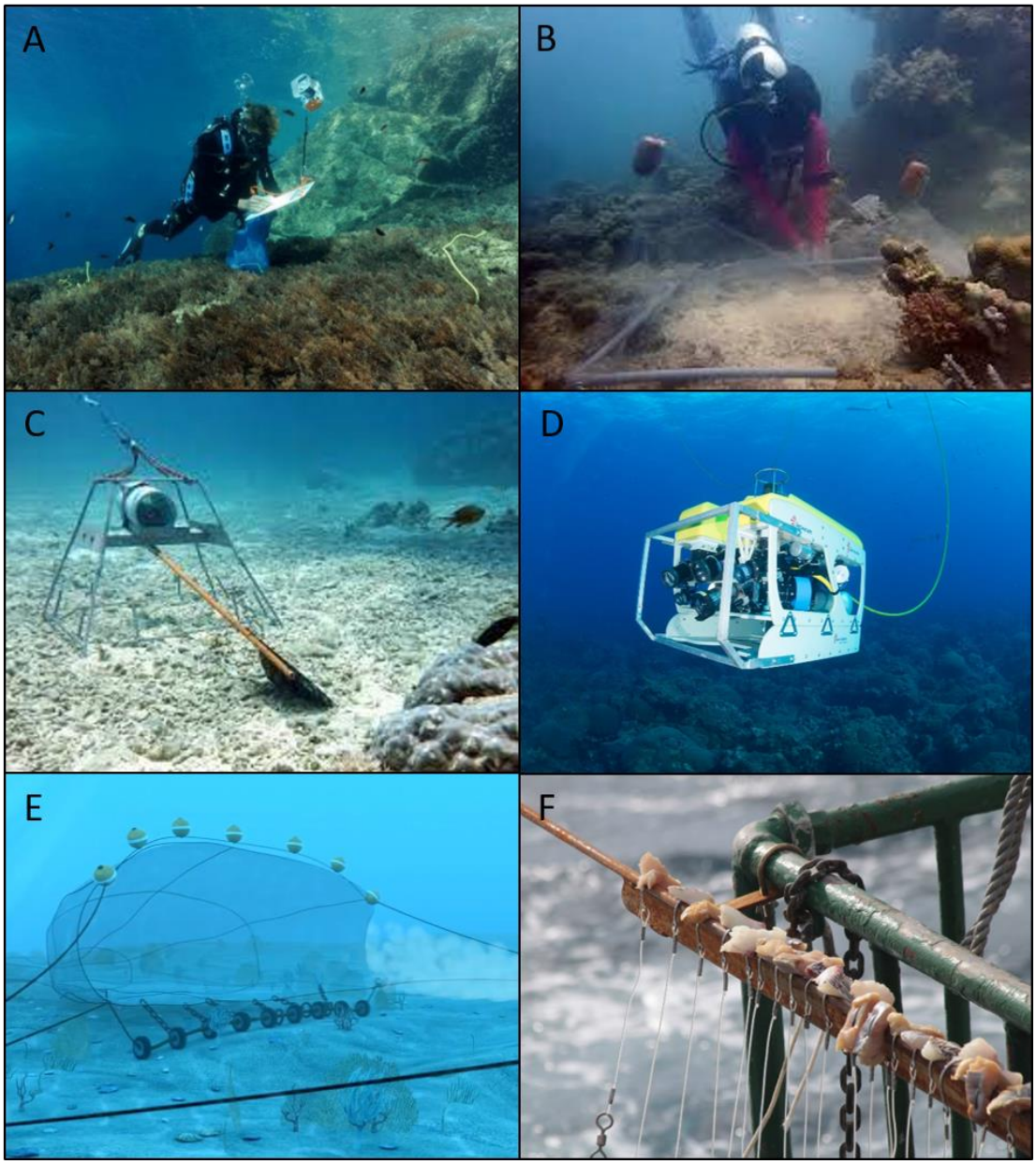


Figure 1.11. *Illustrations non exhaustives de méthodes d'échantillonnage des communautés de poissons en milieu marin. A) recensement visuel en plongée, B) capture des individus avec de la roténone, C) déploiement de caméras appâtées, D) déploiement d'un véhicule sous-marin téléguidé (ROV), E) pêche au chalut, F) pêche à la palangre.*

En milieu marin, la forte diversité des environnements à étudier (profondeur, luminosité, visibilité) implique l'utilisation de méthodologies adaptées. La méthode la plus communément utilisée pour recenser les communautés dans les zones peu profondes est le comptage visuel en plongée (Figure 1.11A). Le plongeur identifie et compte les poissons observés de chaque côté d'un transect (Irigoyen et al. 2018). Cette méthode permet d'obtenir des informations quantitatives et biologiques sur les espèces recensées. De très bonnes compétences taxonomiques sont requises pour pouvoir identifier les individus in-situ, ainsi les comptages

visuels peuvent comporter des biais observateurs dans les identifications (Bernard et al. 2013). La présence des plongeurs et le comportement des animaux peuvent également influencer la détection d'espèces, notamment des prédateurs et des poissons cryptobenthiques (Dickens et al. 2011; Asher et al. 2019). Brandl et al. (2018) ont ainsi démontré que la moitié des espèces de poissons cryptobenthiques ne sont pas détectées lors de recensements en plongée. Ces groupes fonctionnels sont pourtant d'une importance cruciale pour la dynamique des récifs, produisant près de 60% de la biomasse consommée par les autres espèces (Brandl et al. 2019). La roténone peut être utilisée pour étudier les communautés de poissons cryptobenthiques, très localement, ce qui permet de capturer les individus pour les identifier à posteriori (Figure 1.11B). Les méthodes de recensement en plongée sont toutefois contraintes par les capacités techniques et physiologiques humaines qui limitent la durée (≃1h) et la profondeur d'observation (max. 50m), afin de minimiser les risques pour les plongeurs. L'étude des zones mésophotiques (30-150m) et aphotique (>200m) repose donc sur d'autres méthodologies. Les caméras (appâtées ou non avec des poissons gras) peuvent être déployées jusqu'à 500m de fond, équipées de phares, mais sont le plus fréquemment déployées sur les zones côtières (Figure 1.11C). Cette méthode est efficace pour détecter et identifier les poissons proches de la caméra, ou de grandes tailles (prédateurs) (Schramm et al. 2020). Un dispositif en stéréo permet de mesurer la taille et ainsi la biomasse des individus. Toutefois, le temps de déploiement est restreint et l'annotation des vidéos pour l'identification des espèces est souvent très longue (Langlois et al. 2020). De plus, l'usage de ces caméras est limité par la visibilité du milieu et donc peu efficace en zone estuarienne par exemple. Des véhicules sous-marins téléguidés (ROV), drones sous-marins ou *gliders* (Figure 1.11D) peuvent être déployés à de plus grandes profondeurs, mais leur usage est contraint par leur coût et le niveau d'expertise requis pour les manipuler (Sward et al. 2019). Les zones pélagiques sont le plus couramment échantillonnées à l'aide d'engins de pêche (chalut, palangre, Figure 1.11E-F), très destructeurs pour l'écosystème et entrainant de nombreuses prises accidentelles (Trenkel et al. 2019; Roberson et al. 2022).

Les méthodes conventionnelles de recensement des communautés de poissons, bien que possédant de nombreux avantages, comportent tout autant de biais ou limitations concernant i) la nécessité d'une expertise taxonomique pour l'identification, ii) le coût et le temps nécessaires pour une bonne couverture spatiale et temporelle, iii) l'impact destructeur sur l'environnement et iv) la détection des espèces cryptiques, rares ou farouches (Figure 1.12). Du fait de l'utilisation de méthodes très diversifiées et peu transférables entre les différents milieux, il

n'existe qu'un faible nombre d'études standardisées à large échelle spatiale ou temporelle (G. J. Edgar & Stuart-Smith, 2014) alors que certaines régions ou milieux restent rarement étudiés (Rocha et al. 2018; Reboredo Segovia et al. 2020).

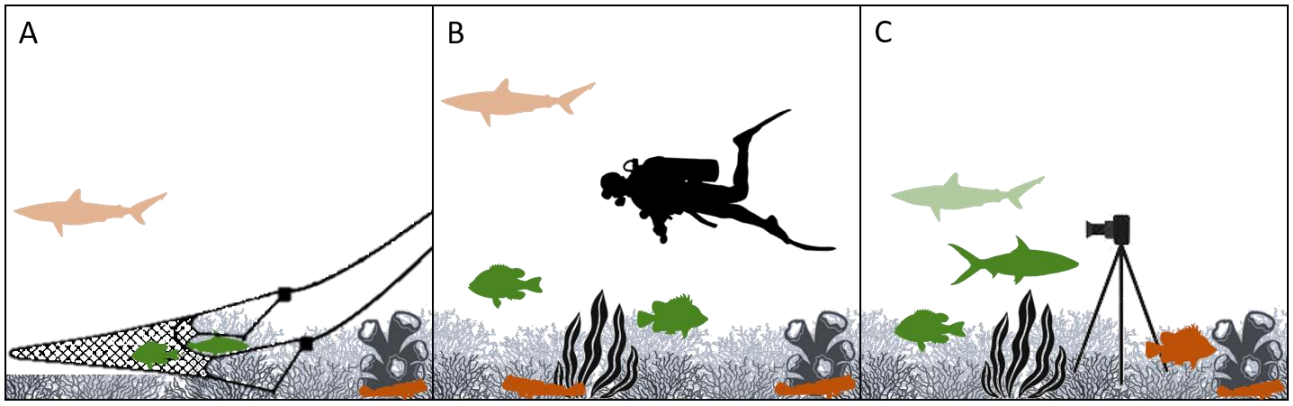


Figure 1.12. *Illustration théorique de biais de détection dépendants de certaines méthodes d'échantillonnage : A) la pêche au chalut, B) le recensement en plongée et C) le déploiement de caméras appâtées. Les individus en vert sont détectés, mais pas les individus en rouge.*

Il devient donc nécessaire de développer une méthodologie d'échantillonnage non invasive, non destructrice, rapide, fiable et standardisable pour obtenir des inventaires quasi-exhaustifs des communautés de poissons et étudier les patrons de diversité.

2.2. Metabarcoding de l'ADN environnemental

L'étude de l'ADN environnemental (ADNe) est une méthodologie récente consistant à analyser les fragments d'ADN libérés par les organismes dans leur milieu, pour détecter la présence d'espèces à enjeux ou étudier les communautés (Taberlet et al. 2018). L'ADNe est un mix de molécules polynucléotidiques, libres ou absorbées sur d'autres molécules, sécrétées par les individus dans leur environnement (Pawlowski et al. 2020; Lacoursière-Roussel & Deiner, 2021). Les poissons sécrètent de l'ADNe par desquamation, le mucus ou les fèces, qu'il est possible de collecter par filtration de l'eau. Cet ADNe peut ensuite être extrait, amplifié par des amorces et séquencé (Figure 1.13). L'ADNe a été utilisé la première fois par Ficetola et al. (2008) pour la détection d'un amphibien en eau douce alors que la première application en milieu marin a été réalisée par Thomsen et al. (2012).

Les avancées technologiques en écologie moléculaire et notamment dans les méthodes de séquençage pourraient grandement bénéficier au suivi de la biodiversité, pour caractériser rapidement et efficacement le statut des communautés et leur évolution (Goodwin et al. 2017).

Depuis les premières générations de séquenceurs « Sanger » permettant de traiter un échantillon à la fois (Sanger et al. 1977), de nombreux développements méthodologiques ont grandement augmenté les performances de séquençage et il est aujourd'hui possible de lire simultanément et rapidement des millions de fragments avec les technologies « Illumina », en réduisant le taux d'erreurs (Slatko et al. 2018).

Ces nouvelles capacités de séquençage ont permis de développer l'utilisation du *metabarcoding* de l'ADN environnemental (ADNe), qui est aujourd'hui appliqué à de nombreux organismes et environnements.

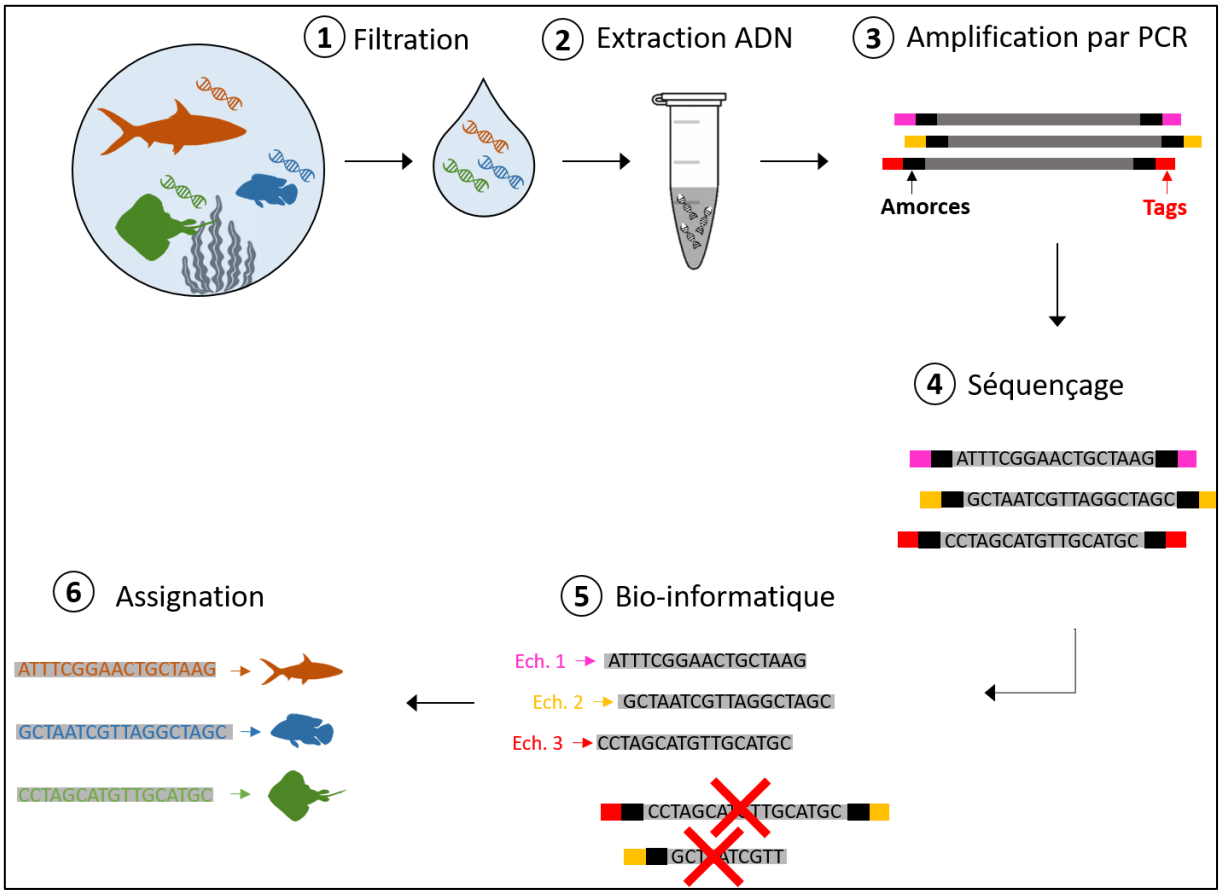


Figure 1.13. *Protocole du metabarcoding de l'ADNe : 1) Collection et filtration de l'eau, 2) Extraction de l'ADN, 3) amplification de l'ADN avec des amorces, 4) séquençage haut-débit, 5) nettoyage bio-informatique et 6) assignation taxonomique des séquences grâce à des bases de références.*

Il existe deux types d'analyses possible de l'ADNe : le *barcoding* permet de cibler une espèce unique par l'utilisation d'amorces d'amplification spécifiques et le *metabarcoding* qui permet de cibler un groupe taxonomique plus ou moins large (ex : eucaryotes, téléostéens), par l'utilisation d'amorces spécifiques à ce groupe. Le *barcoding* de l'ADNe est principalement

utilisé pour la détection et le suivi d'espèces menacées (Simpfendorfer et al. 2016; Weltz et al. 2017; Schweiss et al. 2019) ou invasives (Hatzenbuhler et al. 2017; Larson et al. 2020), ou encore pour étudier la génétique d'une population (Dugal et al. 2021). Le *metabarcoding* est quant à lui utilisé principalement pour étudier les assemblages d'espèces ou plus récemment, de populations (Sigsgaard et al. 2016) dans un milieu. Les amorces de *metabarcoding* sont le plus souvent sélectionnées sur des gènes mitochondriaux, car l'ADN mitochondrial est présent en plus grande quantité au sein des cellules que l'ADN nucléaire et est plus susceptible d'être collecté dans le milieu (Turner et al. 2014). Les amorces sont situées sur des régions conservées du gène au sein du groupe taxonomique ciblé, encadrant une région hypervariable permettant de discriminer les espèces au sein de ce groupe. Plusieurs paires d'amorces peuvent être conçues pour un même groupe taxonomique, avec chacune ses caractéristiques en termes de détectabilité, résolution spécifique, longueur de fragment. Pour l'amplification des séquences de poissons téléostéens, de nombreuses paires d'amorces ont été développées sur différents gènes (12S, 16S, COI, Cyt b). Zhang et al. (2020) ont montré que les amorces situées sur le gène mitochondrial 12S détectent une plus grande diversité de poissons (Ostéichtyens et Chondrichtyens) que celles situées sur les autres gènes (Figure 1.14).

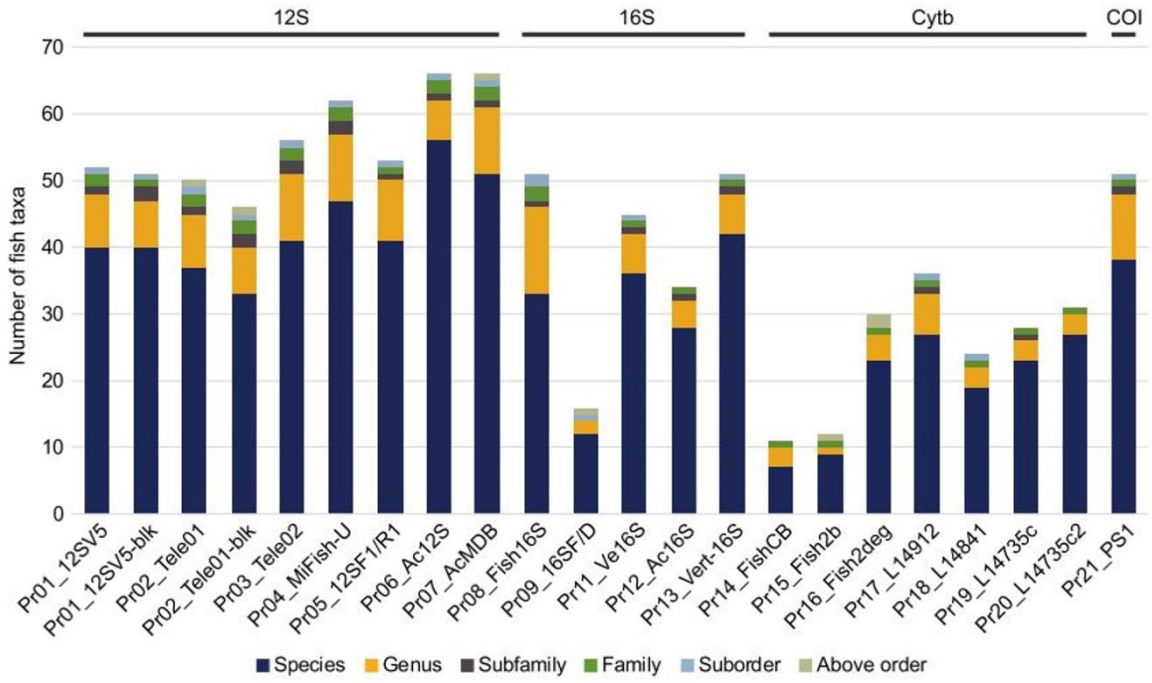


Figure 1.14. *Distribution des différents niveaux de classification taxonomique des taxa de poissons détectés avec diverses paires d'amorces, lors d'une analyse metabarcoding in-vitro effectuée par Zhang et al. (2020).*

Suite au séquençage, les séquences brutes obtenues sont triées par échantillon, nettoyées et assignées à un taxon lors d'un processus bio-informatique, aboutissant à une matrice de communauté par échantillon exploitable pour les analyses écologiques. Les choix des paramètres et des programmes utilisés pour le traitement bio-informatique peuvent toutefois influer sur les résultats obtenus (Pauvert et al. 2019). L'assignation des séquences à un taxon est réalisée par des algorithmes qui comparent les séquences à une base de référence, le plus souvent publique. L'assignation au niveau de l'espèce n'est pas toujours possible, en raison de l'incomplétude des bases de référence (Marques et al. 2020), auquel cas les algorithmes de type « dernier ancêtre commun » (*lowest common ancestor*, LCA) assigne la séquence au genre ou à la famille (ex : *ecotag* de la suite OBITools, Boyer et al. 2016).

Les premières applications de l'ADN environnemental en eau douce ont démontré son efficacité pour la détection des espèces (Valentini et al. 2016). Depuis, de nombreuses études ont révélé que l'ADNe détectait une diversité équivalente voire plus élevée que les méthodes conventionnelles d'échantillonnage, et ce dans différents milieux (Pont et al. 2018; Cilleros et al. 2019; Seymour et al. 2021). En environnement marin, les performances de l'ADNe égalent ou surpassent celles des autres méthodes, que ce soit pour l'étude des communautés (Polanco et al. 2021) ou pour la détection d'espèces emblématiques (Bakker et al. 2017; Boussarie et al. 2018). En raison de sa courte persistance et sa faible dispersion (Murakami et al. 2019), l'ADNe a été montré efficace pour décrire les communautés à fine échelle spatiale et étudier la distribution des poissons au sein de micro-habitats en milieu marin ou en eau douce (Berger et al. 2020; Dugal et al. 2022). À échelle régionale, l'ADNe permet également de décrire les patrons de distribution des poissons (Consuegra et al. 2021; West et al. 2021) et d'estimer la diversité régionale (Juhel et al. 2020, 2022). Outre la description taxonomique des communautés, l'ADNe peut renseigner sur la diversité fonctionnelle et phylogénétique d'un assemblage, lorsqu'il est possible d'assigner les séquences à l'espèce ou au genre (Figure 1.15, (Pont et al. 2019; Marques et al. 2021). Son efficacité prouvée, l'ADNe commence à être utilisé pour le développement d'indicateurs écologiques (Dalongeville et al. 2022; Polanco et al. 2022; Sanchez et al. 2022) et pour le suivi de l'efficacité des aires marines protégées (Boulanger et al. 2021).

L'estimation de la diversité à partir de données d'ADNe fait face à de nouveaux challenges, notamment pour la prise en compte de la détection d'espèces très rares (qui peuvent aussi être des erreurs résiduelles de PCR ou séquençage), ou encore dans le cas où l'état des bases de références ne permet pas d'identifier la totalité des séquences. Pour ce dernier cas, le

groupement de séquences en unités taxonomiques moléculaires (MOTUs) selon leur similarité permet d'obtenir une estimation de la diversité, même avec une identification taxonomique partielle. Marques et al. (2020) ont calibré un pipeline bio-informatique et déterminé les seuils de nettoyage des données nécessaires pour obtenir une diversité de MOTUs égale à la diversité d'espèces dans le milieu étudié. Les métriques de diversité taxonomique α, β et γ sont applicables directement sur ces MOTUs. La diversité peut également être mesurée en nombres de Hill, qui quantifient la diversité en nombres équivalents de MOTUs également abondants. Ces indices permettent de pondérer la diversité selon l'abondance relative des MOTUs et d'inclure une composante phylogénétique ou écologique dans les estimations de diversité (Chao et al. 2014). Alberdi & Gilbert (2019) ont conçu un guide pratique et conceptuel pour les diverses applications de ces nombres de Hill aux données ADNe, en fonction des objectifs souhaités.

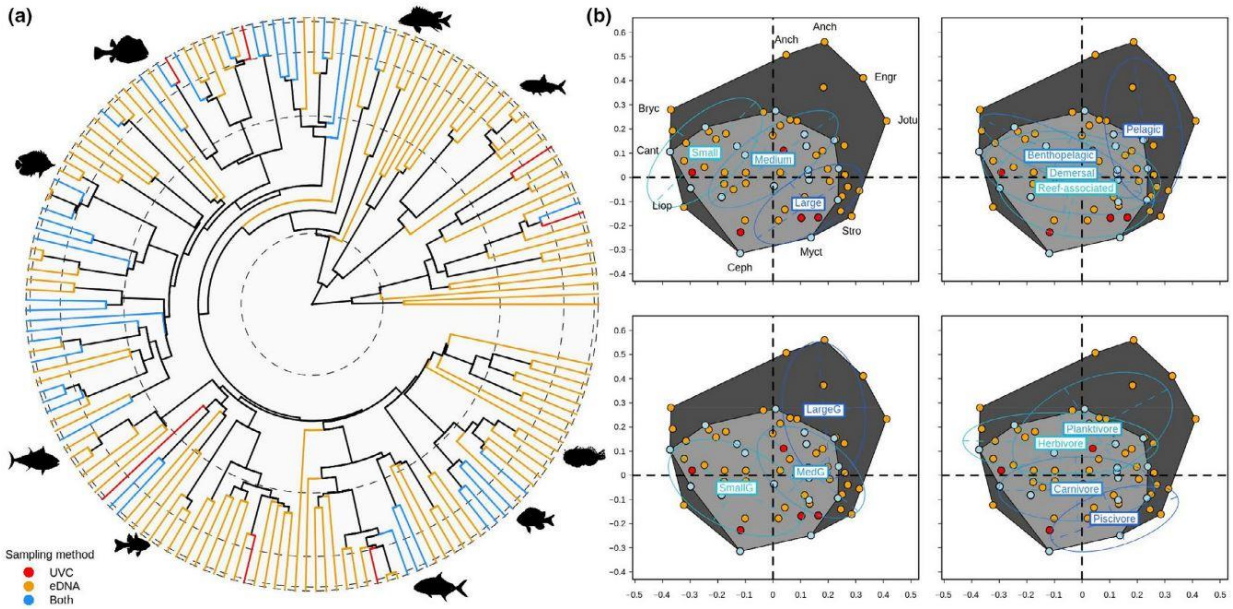


Figure 1.15. *Diversité phylogénétique et fonctionnelle détectées par ADNe et recensements visuels dans les Caraïbes par Polanco et al. (2021). A) Détections sur l'arbre phylogénétique de Rabosky et al. (2018) avec l'ADNe (jaune), les recensements visuels (rouge) ou les deux méthodes (bleu). b) Espace fonctionnel mesuré pour les genres identifiés par ADNe (gris foncé) et par recensements visuels (gris clair), pour 4 types de traits fonctionnels : taille (haut gauche), guilde trophique (haut droite), position dans la colonne d'eau (bas gauche) et grégarisme (bas droite). Figure extraite de Polanco et al. (2021).*

Le *metabarcoding* de l'ADNe est donc une méthode non-invasive et non-destructrice qui permet de caractériser les communautés de poissons à différentes échelles spatio-temporelles.

Cette méthode permet de s'exempter d'expertises taxonomiques pour l'identification et de limiter les biais de détectabilité. L'utilisation de cette méthode est en plein essor dans le domaine de la biologie marine et de nombreux développements et améliorations de protocoles sont en cours, concernant notamment les aspects quantitatifs et bio-informatiques (Yao et al. 2022). Le *metabarcoding* de l'ADNe est un outil qui peut informer la conservation, par l'estimation de zones de fortes diversité, par le suivi d'espèces invasives ou menacées (Bani et al. 2020; Van Oppen & Coleman, 2022).

3. Enjeux

De nombreuses études ont démontré la robustesse de l'ADNe pour l'étude des communautés de poissons, à échelle spatiale locale et régionale, et la détection des espèces menacées. Cependant, diverses incertitudes persistent quant à l'utilisation de l'ADNe pour l'étude de la diversité des poissons et des communautés à large échelle et sur ses capacités à décrire les mécanismes à l'origine des patrons de distribution.

Les estimations de diversité obtenues par l'ADNe dépendent fortement des traitements bio-informatiques et des paramètres de nettoyage appliqués sur les données brutes. Parmi les nombreuses études utilisant l'ADNe sur les poissons, à échelles locales ou régionales, différents programmes bio-informatiques sont utilisés (Pont et al. 2018; Berger et al. 2020; Marwayana et al. 2021), mais il existe peu de comparaisons de leurs performances (Prodan et al. 2020; Brandt et al. 2021), ni de recommandations pour la standardisation des protocoles bio-informatiques. Il persiste donc un enjeu méthodologique pour l'identification des meilleurs programmes bio-informatiques permettant d'obtenir les estimations de diversité les plus fiables et rapides. On peut donc se demander quels sont les programmes les plus performants à chaque étape du traitement bio-informatique ? Quelles étapes fournissent les résultats les plus divergents en fonction des programmes utilisés ? Un assemblage des programmes les plus performants serait-il plus efficace et fiable que les programmes complets existants ? Une étude comparative des performances des programmes bio-informatiques les plus utilisés dans la littérature, dans un cas théorique où les bases de références sont complètes localement, permettrait de suggérer une recommandation d'un ensemble de programmes à utiliser préférentiellement à chaque étape de traitement bio-informatique pour l'analyse de données ADNe de poissons.

A large échelle, les pressions environnementales et humaines croissantes impactent les communautés de poissons, indispensables au bon fonctionnement des écosystèmes et à l'alimentation humaine (Eddy et al. 2021; Mellin et al. 2022). Il est donc nécessaire de suivre efficacement et rapidement l'évolution des communautés et d'en comprendre les patrons de distribution. Les méthodes conventionnelles, limitées par le coût, le temps et l'expertise ne sont plus adaptées pour relever ce défi à large échelle et fournir un recensement quasi-exhaustif des communautés. Le *metabarcoding* de l'ADNe, couplé au protocole bio-informatique développé par Marques et al. (2020) estimant la diversité en MOTUs en cas de base de référence incomplète, pourrait être une méthode appropriée. Si plusieurs études à échelle locale et régionale ont détecté avec l'ADNe des diversités similaires ou supérieures aux méthodes conventionnelles (Boussarie et al. 2018; Polanco et al. 2021), l'efficacité de la méthode à large échelle sur les écosystèmes les plus riches, comme les récifs coralliens, reste à démontrer. Le *metabarcoding* de l'ADNe permet-il de détecter la même diversité que les méthodes conventionnelles à large échelle ? Est-il possible de retrouver les grands patrons biogéographiques connus chez les poissons coralliens avec l'ADNe ? L'ADNe permet-il de mettre en évidence de nouvelles règles d'assemblages des communautés ?

Plusieurs facteurs environnementaux influencent la distribution des poissons à l'échelle mondiale, mais l'impact anthropique devient de plus en plus prédominant dans la plupart des régions (Halpern et al. 2019; Yan et al. 2021). Le déclin continu de l'abondance des populations de poissons et la perte des principaux prédateurs ont été largement signalés (Cinner et al. 2018; Pacoureau et al. 2021). Cependant, l'effet des pressions environnementales et humaines sur la diversité des poissons et la composition des communautés à travers les échelles spatiales (diversité α locale et diversité β) doit encore être quantifiée à l'échelle mondiale. Evaluer la réponse des communautés de poissons côtiers à ces pressions humaines et environnementales, à l'échelle mondiale, est une urgente nécessité afin de pouvoir prendre les mesures de conservation adéquates. Nous pouvons donc nous demander quelle est l'influence relative des facteurs environnementaux, socio-économiques et géographiques sur la diversité des poissons ? Toutes les métriques de diversité sont-elles influencées de la même façon par les pressions locales ? Tous les groupes trophiques répondent-ils de la même manière à la modification de leur écosystème ?

L'ADNe permet de mesurer les diversités taxonomiques à travers différentes échelles spatiales, mais les séquences comportent également une information génétique qui peut être

exploitée. À partir de ces séquences, il est possible de calculer un proxy de la diversité génétique, mesuré par la dissimilarité entre les séquences composant une communauté. Bien que cette information génétique fournie par un petit fragment d'ADNe ne puisse pas être exploitée telle qu'elle car elle peut également représenter une part de diversité intraspécifique (Sigsgaard et al. 2016), il peut être possible d'étudier les corrélations avec la diversité phylogénétique et la diversité fonctionnelle d'une communauté. Ces diversités phylogénétique et fonctionnelle diminuent depuis le début de l'Anthropocène (Li et al. 2020) et il serait utile de pouvoir les estimer et suivre leur évolution en utilisant l'ADNe, sans nécessairement assigner toutes les séquences à un taxa. Des indices de diversité calculés à partir de l'information génétique fournie par l'ADNe peuvent-ils permettre d'estimer la diversité fonctionnelle ou phylogénétique d'une communauté ?

Plusieurs études ont suggéré que, si les écosystèmes à proximité des zones urbanisées sont dégradés et leur diversité diminuée (Korpinen et al. 2021; O'Hara et al. 2021), les communautés et populations de poissons peuvent se maintenir et prospérer dans les zones plus éloignées, intactes, ou trouver refuge dans les profondeurs (Frank et al. 2018; Pereira et al. 2018; Oliveira, 2019). La faible proportion de monts sous-marins étudiés dans le monde a tout de même permis de révéler qu'il s'agissait d'habitats très riches en organismes benthiques et de zones d'agrégation de la mégafaune (Rowden et al. 2010; Garrigue et al. 2015; Letessier et al. 2019; Muff et al. 2022). Si ces zones abritent de fortes diversités, abondances et biomasses de poissons, alors elles pourraient être des endroits stratégiques pour la mise en place de mesures de conservation. Cependant, très peu d'études se sont intéressées aux assemblages de poissons sur les monts sous-marins, ou à des comparaisons avec des écosystèmes côtiers de même profondeur. Les monts sous-marins abritent-ils une diversité de poissons plus faible, équivalente, ou plus élevée que les pentes externes côtières à des profondeurs égales ? Quelles métriques de diversité sont les plus sensibles à la profondeur et à l'« effet mont » ? Comment intégrer la profondeur dans les plans de conservation ? Comment trouver des solutions de conservation qui atteignent les objectifs internationaux (ex : 30% des océans en 2030) tout en protégeant des zones de fortes diversité ?

4. Hypothèses

On peut supposer que l'ADNe permettra de mieux détecter certaines espèces rares ou cryptiques, difficilement observables par les méthodes conventionnelles (Brandl et al. 2018;

Gaynor et al. 2018), et d'obtenir des inventaires taxonomiques plus exhaustifs. Par conséquent, cet outil pourrait permettre de redécouvrir des patrons de diversité et réévaluer les hypothèses d'assemblage ainsi que leurs mécanismes sous-jacents. Il est probable que l'estimation de diversité γ des poissons coralliens soit plus élevée qu'avec les méthodes conventionnelles et que les patrons de diversité à l'échelle locale soient plus marqués.

J'émets également l'hypothèse que les pressions humaines et environnementales pourraient avoir un impact négatif plus important sur les prédateurs et espèces de grandes tailles, alors que les petites espèces cryptiques pourraient plus facilement s'adapter à des milieux dégradés, ce qui corroborerait les observations de Boulanger et al. (2021) dans un autre environnement. Si tel est le cas alors je présume que les communautés pourraient subir un remplacement d'espèces, qui mènerait à une diminution de la diversité taxonomique de certains groupes trophiques, de la diversité phylogénétique et fonctionnelle, mais pas de la diversité taxonomique totale (Figure 1.16).

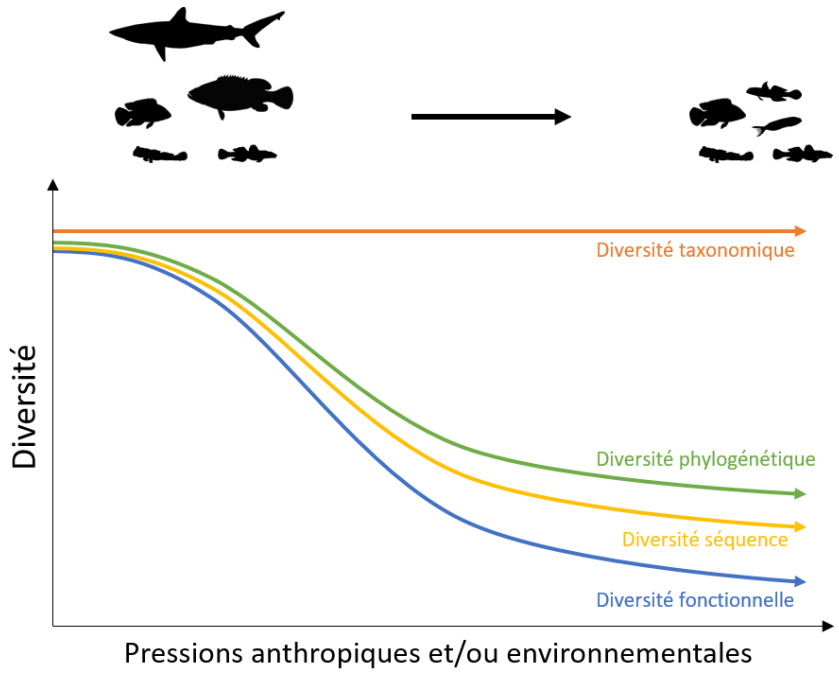


Figure 1.16. *Schéma théorique de l'évolution d'une communauté de poissons en réponse à une pression anthropique croissante, et de l'évolution des diversités taxonomique, phylogénétique, fonctionnelle et de séquences associée.*

Dans le cas des monts sous-marins, l'ADNe, combiné à d'autres méthodes d'échantillonnage, pourrait permettre selon moi d'étudier et modéliser différentes métriques de diversité autour des monts sous-marins et pentes externes. Cela permettrait ainsi d'informer la conservation sur les zones prioritaires à protéger, en intégrant la profondeur. Il est probable que

la profondeur du sommet aura une influence importante sur la diversité et la composition des communautés de poissons sur le mont. J'émets de ce fait une hypothèse, selon laquelle les monts sous-marins peu profonds, dont le sommet se trouve dans la zone euphotique, hébergeraient une diversité comparable à celle des pentes externes de récif côtiers. Les monts sous-marins dont le sommet se trouve dans la zone aphotique, quant à eux, hébergeraient une diversité moins importante, mais principalement constituée d'espèces profondes peu présentes près des côtes. Tous ces habitats seraient donc intégrés dans les plans de conservation, et cela permettrait de protéger une large surface, sur plusieurs couches de profondeur.

5. Objectifs de la thèse

L'objectif principal de cette thèse est d'utiliser le *metabarcoding* de l'ADNe pour décrire les grands patrons spatiaux de biodiversité des poissons marins à large échelle, étudier l'impact humain et environnemental sur divers aspects de la biodiversité marine et informer la conservation en incluant des zones potentiellement refuges telles que les monts sous-marins ou les pentes profondes.

Le premier chapitre est consacré à la description de la méthodologie employée pour la collecte des données de cette thèse, les analyses en laboratoire et le traitement bio-informatique. Dans le deuxième chapitre, je mène une comparaison de plusieurs outils bio-informatiques couramment utilisés en ADNe pour évaluer l'impact du choix de ces programmes sur les résultats obtenus et sélectionner les plus performants. Le chapitre 3 compare deux méthodes, l'ADNe et le comptage visuel en plongée, pour étudier les patrons de distribution des poissons coralliens à large échelle et attester de l'efficacité de l'ADNe. Dans le chapitre 4, j'utilise l'ADNe pour étudier l'impact de facteurs environnementaux, socio-économiques et géographiques sur la diversité taxonomique et la diversité en séquences α et β des poissons côtiers à l'échelle globale. Le cinquième et dernier chapitre est consacré à la modélisation de la biodiversité des poissons sur les monts sous-marins et pentes profondes de Nouvelle-Calédonie, en fonction de la profondeur et à partir de plusieurs méthodes d'échantillonnage complémentaires, afin de déterminer les enjeux de conservation et proposer une planification des futures aires marines protégées en trois dimensions. Enfin, la discussion générale, dernière partie de cette thèse, dresse un bilan des travaux réalisés et de leurs limitations. Ce chapitre propose également diverses perspectives pour la poursuite des recherches sur l'ADNe comme outil pour étudier la biodiversité des océans et informer la conservation.

Chapitre 1 – Méthodologie

Échantillonnage le long du récif corallien, Nouvelle-Calédonie

1. Échantillonnage

1.1. Campagnes d'échantillonnage

Les analyses réalisées dans cette thèse reposent sur 830 échantillons collectés dans 11 régions et 8 mers et océans au cours de diverses missions (Figure 2.1). Parmi les missions effectuées dans les environnements tropicaux, cinq sont portées par le projet MEGAFAUNA et les Explorations de Monaco (Caraïbes, Océan Indien, Pacifique Est). Les échantillons de Nouvelle-Calédonie collectés sur les récifs coralliens sont affiliés au projet REEF3.0 (partenariat entre l'IRD et l'entreprise Ginger Soproner) et ceux sur les monts sous-marins, à l'ANR SEAMOUNTS. L'échantillonnage en Indonésie est issu du projet Lengguru 2017 (collaboration entre l'IRD, l'université de Papouasie et l'Institut Indonésien des sciences). L'échantillonnage en Colombie est issu du projet Reefish, en collaboration avec l'INVEMAR. La collecte d'échantillons en Méditerranée a été financée par le projet ReserveBenefit (Biodiversa 2017-2020) et l'Agence de l'Eau. Les échantillons en Antarctique ont été prélevés au cours du projet Pole2Pole porté par Umweltstiftung Greenpeace. L'échantillonnage en Atlantique a été financé par l'Ifremer (projet FisheDNA) et la région Pays de la Loire. L'échantillonnage en Arctique est porté par le projet TOPtoTOP Global Climate Expedition. L'échantillonnage en mer de Chine a été réalisé par l'École de l'Environnement de Nanjing.

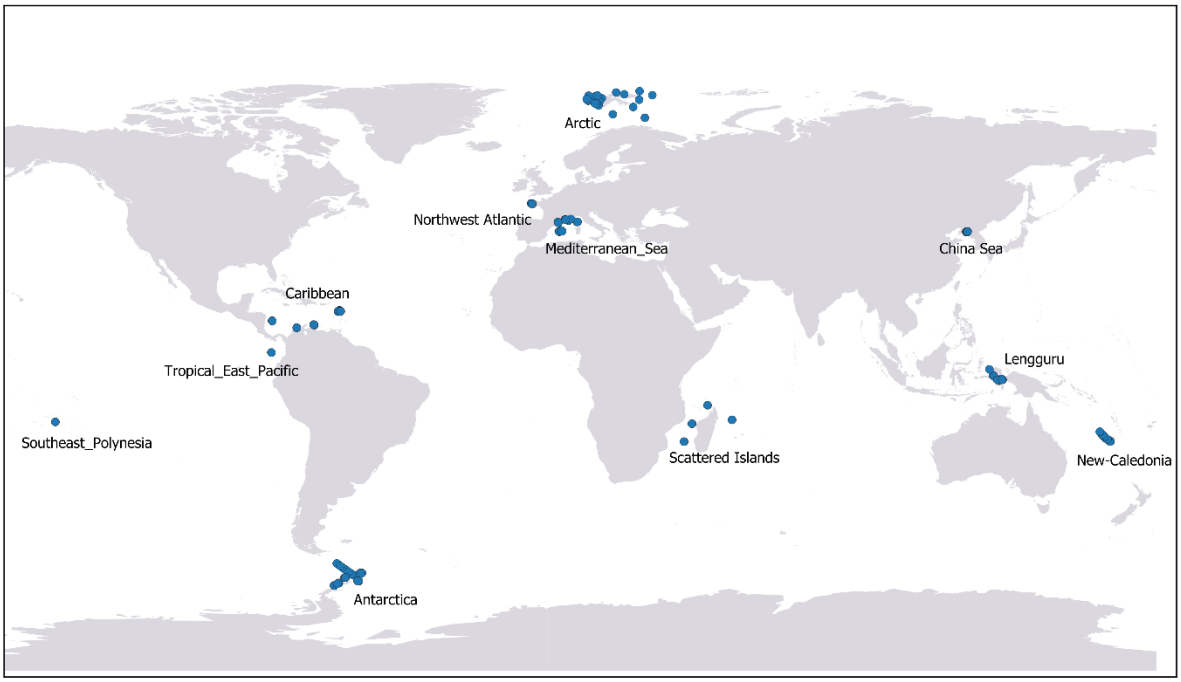


***Figure 2.1**. Carte des sites d'échantillonnage utilisés pour les analyses de cette thèse, répartis dans 11 régions et 8 mers et océans.*

Les données issues des régions tropicales seulement (Caraïbes, océan Indien, Pacifique ouest et central et Indonésie) ont été utilisées pour les analyses du manuscrit B (chapitre 3). L'ensemble des données est utilisé pour les analyses du manuscrit C (chapitre 4). Les données profondes de Nouvelle-Calédonie seulement ont été analysées dans le manuscrit D (chapitre 5). J'ai personnellement participé aux missions d'échantillonnage aux Baléares et à Banyuls en Méditerranée (2 semaines, en 2018), dans le lagon de Nouvelle-Calédonie (2 semaines, en 2019) et sur les monts sous-marins de Nouvelle-Calédonie (3 semaines, en 2020). J'ai également participé à une mission d'échantillonnage en Tanzanie (2 semaines, en 2021) dont les données ne sont pas analysées dans cette thèse.

1.2. Échantillonnage des zones côtières

L'ADN environnemental (ADNe) contenu dans l'eau de mer, du fait du large volume d'eau, de la géographie des fonds marins et des courants, est plus dilué que dans les milieux d'eau douce (Bessey et al. 2020). Les méthodes d'échantillonnage doivent donc être adaptées à ces milieux afin de garantir la meilleure collection des fragments d'ADNe présents et de détecter la diversité avec la plus grande fiabilité. Il n'existe actuellement pas de protocole standardisé pour la collecte d'ADNe en milieu marin, seulement des recommandations sur les tailles de filtres et volume à prélever (Bruce et al. 2021; Stauffer et al. 2021). L'ADNe en milieu marin peut être libre ou absorbé, est très dispersé et peu concentré et peut être dégradé rapidement (Saito & Doi, 2021). Afin de collecter le plus grand nombre de fragments d'ADNe présents dans le milieu, il est recommandé d'utiliser des filtres à pores de 0,2µm (Bessey et al. 2020) et de filtrer de grands volumes d'eau (>30L, Stauffer et al. 2021). L'échantillonnage ayant été réalisé sur une large période de temps entre les différentes régions et campagnes et les technologies évoluant rapidement, plusieurs méthodes différentes ont été utilisées au cours des campagnes. Les premiers échantillons, à Lengguru, en Indonésie, ont été collectés à l'aide de sacs plastiques stériles d'une contenance de 2L et filtrés sur des filtres Sterivex (pores de 0,22µm). Plusieurs réplicas par échantillon (2 à 4) ont été prélevés. Les échantillons de Chine ont également été collectés avec des bouteilles stérilisées, d'une contenance de 1L, avec 3 réplicas par échantillon. Les autres échantillons utilisés dans les études à échelle globale ont été collectés par filtration le long de transects de surface avec des pompes péristaltiques Athena (Proactive Environmental Products LLC, Florida, USA; débit de 1L/min, Figure 2.2).

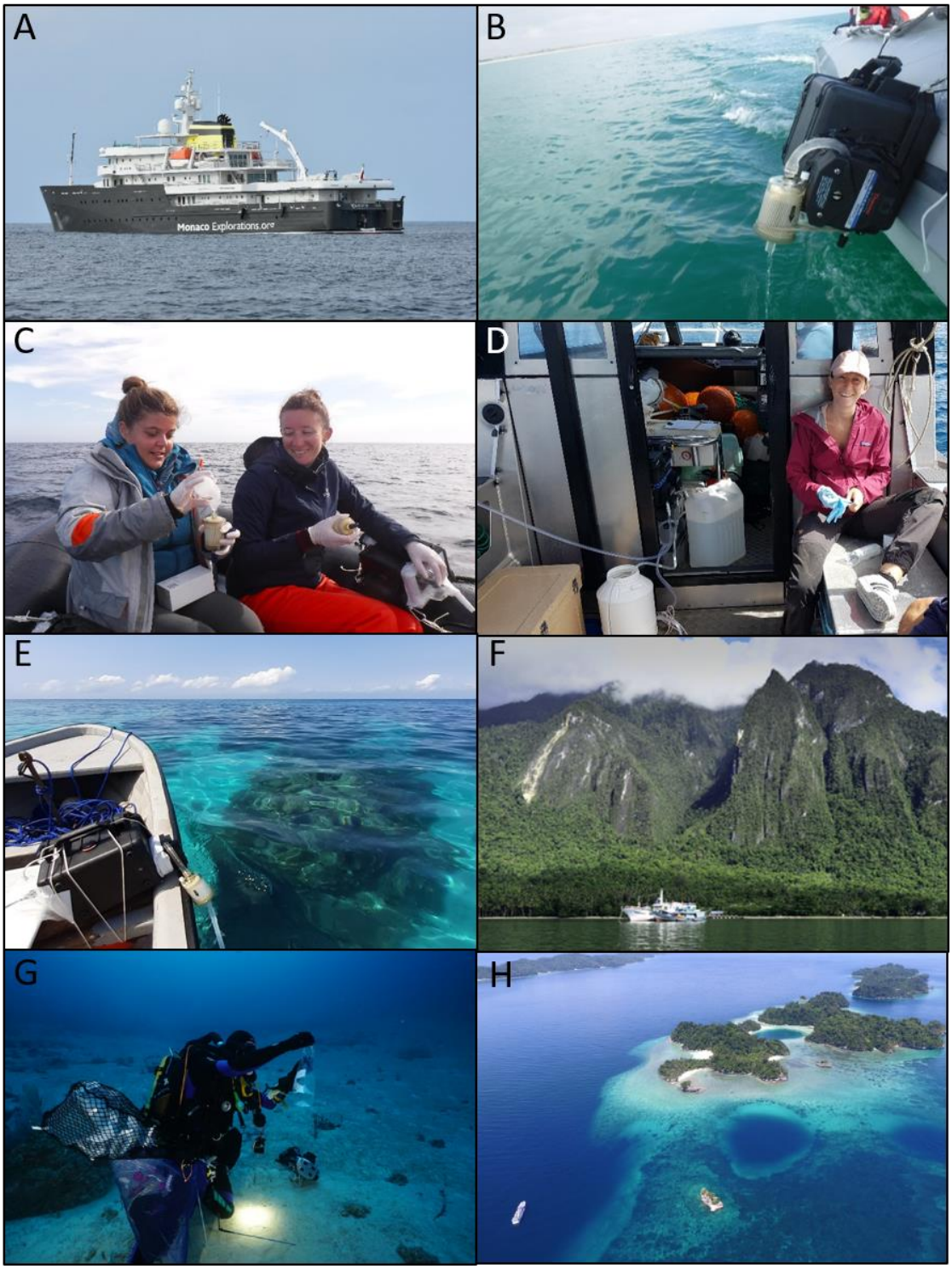


Figure 2.2. *Illustrations des missions d'échantillonnage. A) Navire océanographique de Monaco, le Yersin, à Malpelo (crédit : R. Hocdé, 2018). B) Positionnement de la pompe et de la capsule de filtration pour les filtrations en surface à bord d'un semi-rigide. C) Échantillonnage aux Baléares, en Méditerranée (crédit : S. Mallol, 2018). D) Échantillonnage en Nouvelle-Calédonie sur les récifs coralliens (crédit : F. Baletaud, 2019). E) Échantillonnage en Tanzanie, sur un récif corallien (crédit : A. Dalongeville, 2021). F) Échantillonnage en Indonésie, West Papua (crédit : R. Hocdé, 2017). G) Prélèvement d'eau de mer avec des sacs stériles, en plongée, en Indonésie (crédit : G. Diraimondo, 2017). H) Vue aérienne d'un site d'étude en Indonésie (crédit : G. Diraimondo, 2017).*

L'eau de mer est pompée dans un tuyau stérile et filtrée sur une capsule stérile avec un filtre de 0,2µm (VigiDNA®, SPYGEN, le Bourget du Lac, France). L'eau est ainsi pompée pendant 30min, afin de filtrer un volume de 30L. Deux réplicas sont filtrés en parallèle depuis les deux côtés du bateau sur lequel est réalisée la manipulation. À la fin de la filtration, les capsules sont remplies avec 80mL de tampon de conservation CL1 (SPYGEN, le Bourget du Lac, France). Les tuyaux utilisés pour la filtration sont stériles et changés entre chaque filtration. Tout le matériel réutilisable est désinfecté entre chaque manipulation. Les échantillons prélevés pendant les premières missions d'échantillonnage en Méditerranée et dans les Caraïbes ont été collectés le long de transects rectangulaires de 2km de long sur 0.5km de large, afin de collecter de l'ADNe côtier et pélagique (Figure 2.3A). Les résultats ont démontré que cette méthode ne permettait pas de détecter l'ADNe pélagique et diluait beaucoup l'ADNe côtier retrouvé. Par la suite, nous avons donc opté pour des transects aller-retour de 2km le long de la côte, au plus près des habitats d'intérêt (Figure 2.3B). Certains échantillons de Malpelo, Pacifique Est, ont été collectés sur des transects de surface circulaires, autour d'un point d'intérêt (Figure 2.3C). Sur les récifs de Nouvelle-Calédonie, j'ai filtré l'eau au plus près du substrat, en utilisant des tuyaux de 20 à 35m lestés permettant de filtrer à moins de 5m au-dessus du fond (Figure 2.3E). Ces tuyaux étaient désinfectés à la javel puis à l'eau de mer entre chaque filtration. Cette technique de filtration a été adoptée car de précédentes expérimentations ont démontré la diminution du signal avec l'éloignement au substrat.

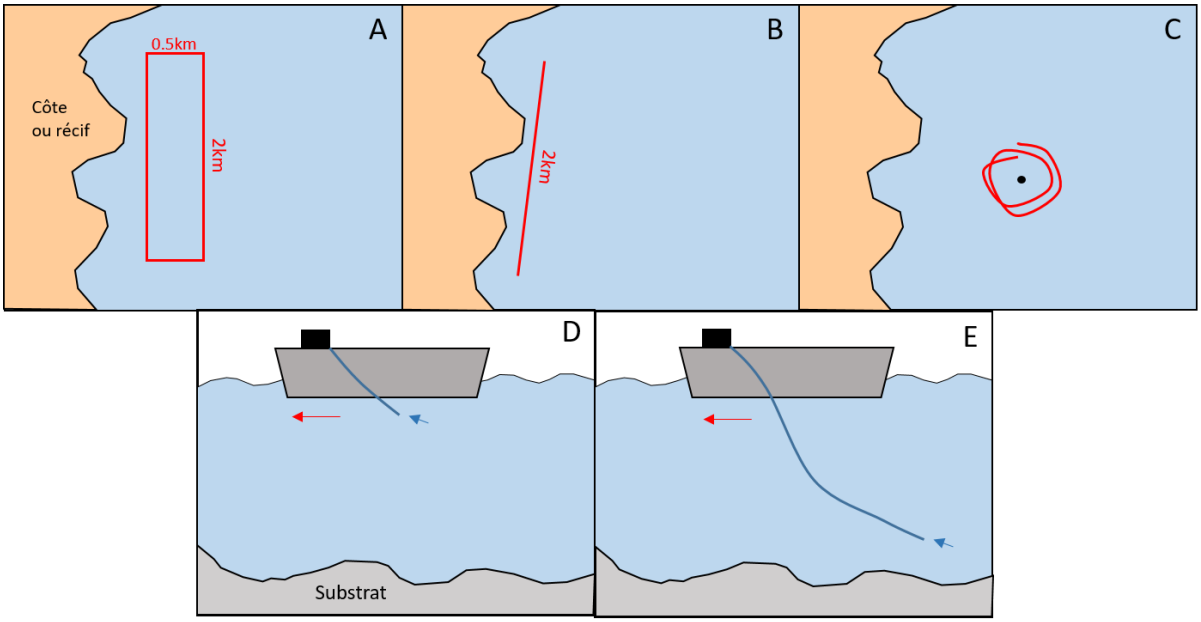


Figure 2.3. *Différentes méthodes d'échantillonnage par transect. A) Transect rectangulaire, B) Transect aller-retour le long de la côte, C) Transect circulaire autour d'un point d'intérêt, D) Transect en surface (utilisé pour les transects A, B et C), E) Transect profond avec un long tuyau (utilisé en Nouvelle-Calédonie sur des transects B).*

Les informations détaillées sur les dates d'échantillonnage, le nombre d'échantillons et de stations par région, ainsi que le volume filtré sont répertoriées dans le Tableau 2.1. Les méthodes d'analyses en laboratoire (extraction d'ADN et PCR) sont identiques pour tous les échantillons (voir section 4 ci-dessous). Cependant, les techniques de séquençage des échantillons ADNe évoluant rapidement, le type de séquenceur utilisé varie selon les campagnes et est donc également reporté dans le tableau 2.1.

Tableau 2.1. *Informations sur l'échantillonnage côtier, de surface, dans chaque région. En bleu, les régions où j'ai participé à l'échantillonnage.*

Région	**Année d'échantil-lonnage**	**Méthode d'échantil-lonnage**	**Nombre d'échant.**	**Nombre de stations**	**Volume moyen filtré par station (L)**	**Volume total filtré (L)**	**Filtre**	**Séquenceur**
Indonésie	2017	Sacs plastiques stériles	64	32	27.2 ±2.2	1005	Sterivex	Hiseq
Pacifique Est	Mars 2018	Transect A & C	26	13	44	572	Capsule	Hiseq
Polynésie Française	Juin 2018	Transect B	12	3	68 ±81.3	272	Capsule	Hiseq
Méditerra-née	2018 2019 2020	Transect A & B	108	35	67.8 ±22.2	2376	Capsule	Miseq - Hiseq
Caraïbes	2018 2020	Transect A & B	89	41	45.7 ±24.4	1876	Capsule	Hiseq - Miseq
Océan Indien	2019	Transect B	31	16	44	704	Capsule	Miseq
Atlantique Est	2019 2020	Transect B	23	13	44	572	Capsule	Miseq
Pacifique Sud-Ouest	2019 2020	Transect D	71	26	69	1380	Capsule	NextSeq - Miseq
Arctique	2020	Transect B	19	19	22	418	Capsule	Miseq
Mer de Chine	2020	Bouteille Stérile	68	24	12	288	Sterivex	IonTorrent
Antarctique	Janvier-Février 2020	Transect B	79	41	32.7 ±17.9	1344	Capsule	NextSeq - Miseq

1.3. Échantillonnage des monts sous-marins

Les données utilisées pour les analyses du manuscrit D (chapitre 5) ont été collectées dans le cadre du projet SEAMOUNTS, un partenariat entre l'UMR Entropie (IRD), le CEFE et l'UMR MARBEC. Ce projet avait pour but de caractériser les communautés de poissons sur les monts sous-marins autour de la Nouvelle-Calédonie, en fonction de leur profondeur, de leur éloignement à l'homme et de plusieurs paramètres physico-chimiques. Quatre pentes externes de récif et 11 monts sous-marins ont été échantillonnés : 4 dont le sommet était à moins de 200m sous la surface, 4 entre 200 et 320m et 3 entre 320 et 500m sous la surface (Figure 2.4). Ces classes de profondeurs ont été choisies pour représenter les différentes strates de luminosité selon le gradient euphotique – aphotique. Ces monts sous-marins et pentes externes, répartis dans le parc marin de la mer de Corail, à différentes distances des zones côtières habitées, ont été échantillonnés au cours de quatre campagnes océanographiques, deux en 2019 et deux en 2020, à bord du navire ALIS. J'ai personnellement participé à une de ces quatre campagnes, en 2020 et échantillonné les sites 8 à 12.

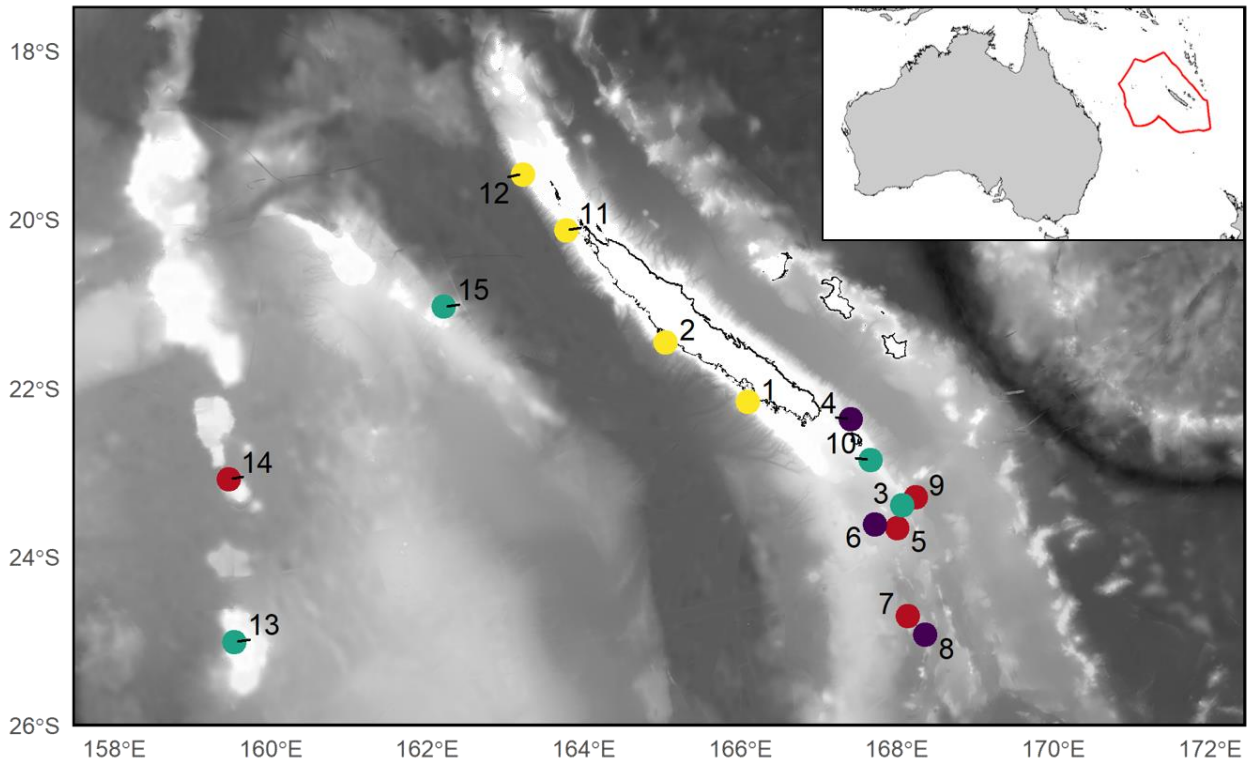


Figure 2.4. *Echantillonnage pour le projet SEAMOUNTS. 4 pentes externes de récifs (jaune), 4 monts sous-marins < 200m (vert), 4 monts sous-marins entre 200 et 320m (rouge) et 3 monts sous-marins entre 320 et 500m (violet).*

Afin de caractériser le plus fidèlement possible la biodiversité sur ces sites, plusieurs techniques d'échantillonnage ont été utilisées. Nous avons déployé des stations stéréo de caméras appâtées sur chaque site. Ces systèmes sont composés de 2 caméras GoPro Hero 5 montées sur une structure métallique, orientées avec un angle permettant d'analyser les vidéos en stéréo et de mesurer les individus observés. Sur la structure métallique sont également fixés un phare pour éclairer dans la zone aphotique et une perche avec une boite à appât contenant 1kg de sardines fraichement broyées (Figure 2.5). Nous avons déployé simultanément entre cinq et huit systèmes sur les sommets des monts sous-marins et les pentes externes, à l'aide d'un treuil et les avons récupérés après 2h30 d'enregistrement vidéo.

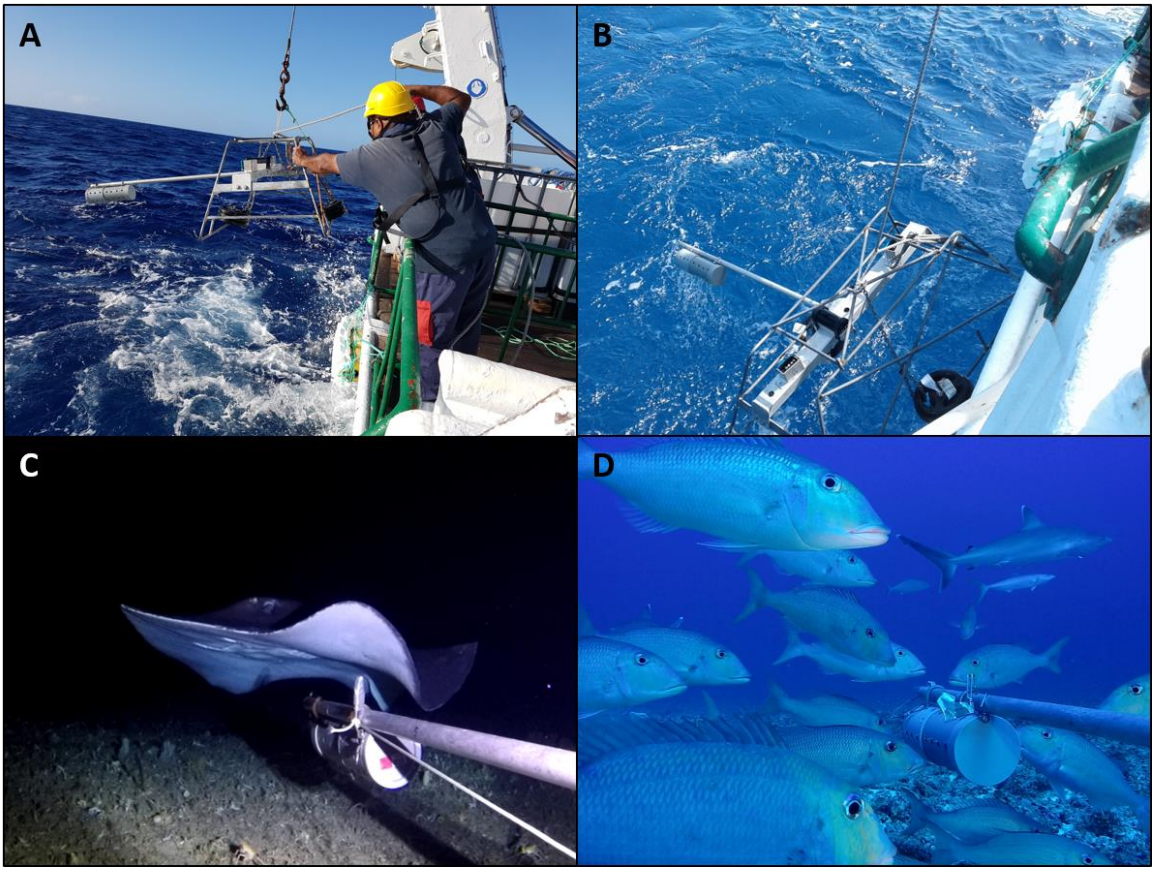


Figure 2.5. *A) et B) Déploiement des BRUVS depuis le navire Alis, C), Dasyatidae sur le mont Eponge à 500m de profondeur, D) Lethrinidae et Carcharhinidae sur le mont Fairway à 60m de profondeur (crédits : L. Mathon, H. Bidenbach et F. Baletaud, 2020).*

Les vidéos obtenues sur les caméras appâtées sont ensuite analysées avec le logiciel EventMeasure, qui permet d'identifier les individus observés, de les mesurer et de calculer le nombre maximum d'individus d'une espèce observés sur une frame (maxN). Il est donc possible d'extraire des indices de richesse spécifique, d'abondance, de biomasse et de diversité fonctionnelle des poissons.

Nous avons également prélevé de l'eau de mer sur chacun des sites afin de filtrer l'ADNe présent sur et autour des monts sous-marins. Les zones d'échantillonnage étant situées entre 50 et 500m de profondeur, la filtration par transect n'était pas envisageable. Lors de ces campagnes d'échantillonnage, nous avons utilisé une rosette équipée de 12 bouteilles Niskin d'une contenance de 8L chacune. La rosette est descendue sur chaque sommet de mont sous-marin ou pente externe à l'aide du treuil du navire, jusqu'à 5m au-dessus du substrat. Quatre bouteilles sont refermées afin d'emprisonner 32L d'eau puis la rosette est remontée. La rosette est également équipée d'une sonde CTD permettant de mesurer la température, la profondeur, l'oxygène dissous et la conductivité. Les bouteilles Niskin sont remontées sur le pont, l'extérieur est rincé à l'eau distillée, puis les tuyaux stériles de Spygen sont branchés dessus, passés dans la pompe péristaltique Athéna et branchés sur les capsules stériles récoltant l'ADNe à l'autre extrémité. Les quatre bouteilles contenant les 32L d'un échantillon sont filtrées sur la même capsule. Nous avons prélevé dix réplicas de 32L par site, répartis sur la surface de chaque sommet et pente externe. Nous avons également collecté six échantillons le long d'un profil de profondeur, à proximité du mont ou du récif. Une première rosette est descendue à 1000m où 4 bouteilles sont refermées, puis 4 autres bouteilles sont fermées à 500m et 4 autres à 250m de profondeur. Ces trois échantillons sont filtrés comme expliqué précédemment. Une deuxième rosette est ensuite descendue à la même localisation pour collecter les échantillons à 150, 80 et 20m de profondeur (Figure 2.6). Dans l'étude du chapitre 5 je n'ai analysé que les échantillons jusqu'à 500m, car les échantillons prélevés à 1000m n'ont pas tous été séquencés.

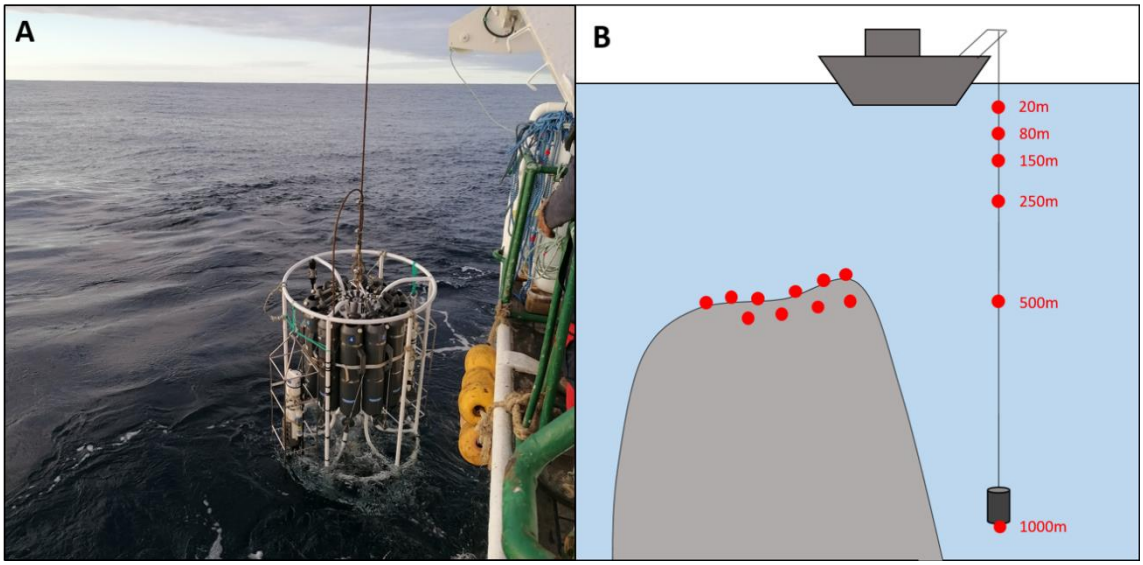


Figure 2.6. *A) Mise à l'eau de la rosette depuis le navire Alis. B) Schéma échantillonnage ADNe sur un mont sous-marin (crédits : L. Mathon, 2020).*

Les capsules contenant l'ADNe sont ensuite envoyées à l'entreprise Spygen, spécialisée dans le traitement des échantillons d'ADNe, pour extraire, amplifier et séquencer l'ADNe (voir section 3 ci-dessous). Les données issues du séquençage doivent être nettoyées, filtrées et assignées (voir section 4) et permettront d'obtenir des indices de diversité taxonomiques.

Lors de ces 4 campagnes océanographiques, le sondeur acoustique du navire ALIS a enregistré le signal acoustique jusqu'à 800m de profondeur, sur tous les sites échantillonnés et pendant les périodes de transit. Ces données permettent d'extraire des indices de biomasse par tranche de profondeur.

2. Constitution de la base de référence

2.1. Échantillonnage des tissus de poissons

Pour ne garder que l'ADNe des poissons dans nos échantillons, nous ciblons un fragment du gène mitochondrial 12S. Afin d'identifier les espèces auxquelles appartiennent ces fragments d'ADNe, les séquences obtenues à la fin du traitement bio-informatique sont comparées aux séquences contenues dans les bases génétiques publiques telles que NCBI. Malheureusement, ces bases de références sont très largement incomplètes pour les espèces de poissons tropicaux et/ou profonds (Marques, Milhau, et al. 2020) et ces bases seules ne nous permettent pas d'assigner précisément nos séquences au niveau de l'espèce. Il est donc nécessaire de compléter ces bases en extrayant et séquençant l'ADN des espèces manquantes à partir d'individus prélevés sur nos lieux d'échantillonnage. Ainsi, lors de chaque mission et lorsque les permis de pêche nous étaient accordés, nous avons pêché à la ligne sur les récifs coralliens ou à la palangre profonde au-dessus des monts sous-marins pour prélever des espèces peu représentées dans les bases génétiques publiques. Les individus prélevés sont pourvus d'un code identifiant unique, photographiés et identifiés taxonomiquement. Un morceau de nageoire est ensuite prélevé avec des outils désinfectés et placé dans un tube ependorf rempli d'alcool à 96°qui sera renouvelé après 24h pour optimiser la conservation de l'ADN (Figure 2.7). J'ai également collecté des échantillons de poissons dont la séquence n'était pas référencée, sur le marché aux poissons de Nouméa et sur des individus pêchés lors de précédentes missions et conservés au congélateur. Lors des autres missions auxquelles je n'ai pas participé, des poissons absents des bases de référence étaient également péchés ou échantillonnés sur les marchés aux poissons.

Figure 2.7. *Prélèvement de spécimens pour séquencer leur ADN. A) Marché au poisson en Indonésie, West Papua (crédit : R. Hocdé, 2017), B) Beauclaire longue-aile (Cookeolus japonicus) pêché lors d'une campagne du projet SEAMOUNTS, C) Conservation des échantillons de tissus dans les tubes d'éthanol (crédits : L. Mathon & A. Brouquier, 2020).*

2.2. Extraction, amplification, séquençage

Les échantillons de tissus de poissons collectés sont conservés dans l'éthanol 96° jusqu'à leur traitement en laboratoire. Les échantillons sont découpés et séchés à l'étuve. L'ADN est ensuite extrait avec le kit DNeasy Blood & Tissue de Qiagen, selon les recommandations du fournisseur. L'ADN est ensuite amplifié par PCR (réaction en chaine par polymérase : procédé moléculaire de réplication de l'ADN). L'ADN est amplifié en utilisant les amorces « sens » VO5F898 (AAACTCGTGCCAGCCACC) et « anti-sens » Teleo-R (CTTCCGGTAC ACTTACCATG), qui ciblent un fragment d'environ 700 paires de bases sur le gène 12S. Ce fragment est choisi car il contient plusieurs marqueurs d'intérêts pour les poissons (tele01, tele02 et MiFish), chondrichtyens (Chond01) et vertébrés (Vert01) (Figure 2.8).

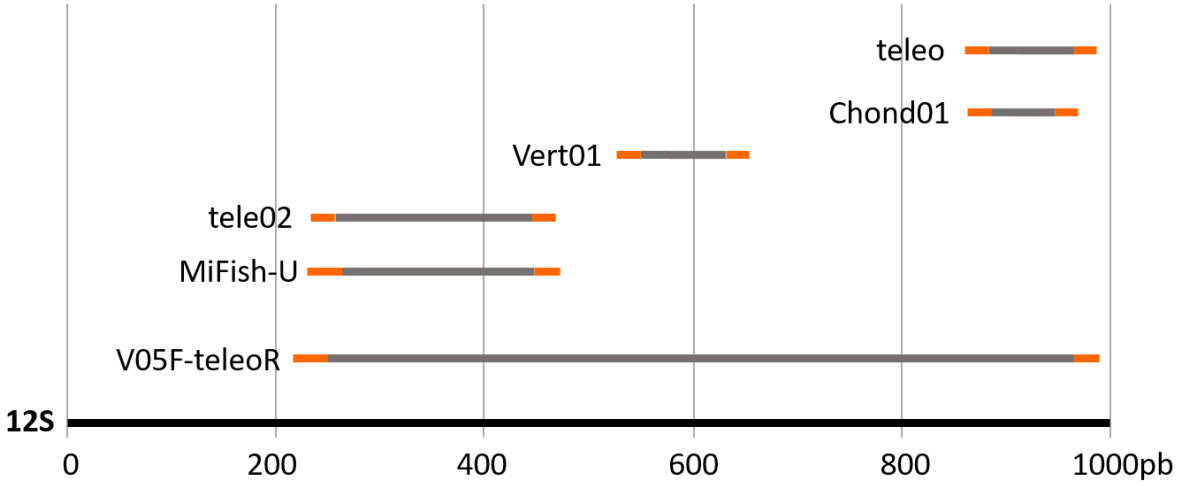


***Figure 2.8**. Schéma de l'emplacement des marqueurs d'intérêt pour cibler les poissons sur le gène 12S. Les extrémités oranges indique l'emplacement des amorces. Figure inspirée de Zhang et al. (2020).*

La quantité d'ADN et la qualité de l'amplification sont vérifiées par NanoDrop et électrophorèse sur gel, avant d'envoyer les échantillons à Genoscreen, pour les séquencer par la méthode Sanger. Les chromotogrammes de chaque séquence (un par sens de lecture) sont ensuite nettoyés et alignés pour extraire la séquence correspondant à chaque échantillon, en utilisant le logiciel Geneious. J'ai personnellement effectué l'extraction, l'amplification et le nettoyage des séquences de plus de 150 échantillons collectés en Nouvelle-Calédonie. En parallèle les échantillons prélévées dans les autres sites subissent le même traitement dans les laboratoires du CEFE à Montpellier et de l'ETH à Zurich afin de compléter globalement la base de référence. J'ai ensuite centralisé toutes les nouvelles séquences obtenues, je les ai nettoyées et j'ai constitué une base de référence propre.

3. Metabarcoding : extraction, amplification, séquençage des échantillons ADNe

Les échantillons d'ADNe collectés dans les capsules stériles sont envoyés à Spygen pour le traitement en laboratoire. La première étape consiste à extraire l'ADN contenu sur les filtres (Pont et al. 2018). Cette étape est réalisée par Spygen, dans une salle blanche dédiée, désinfectée avant et après toute manipulation. L'amplification, également réalisée par le personnel de Spygen, consiste à répliquer en très grand nombre les séquences d'ADN contenues dans les échantillons, par réaction en chaîne par polymérase (PCR) en utilisant les amorces ciblant le fragment d'intérêt. Dans le cadre de ces travaux de thèse, le marqueur choisi est teleo, ciblant les poissons osseux et les Elasmobranches. Le fragment d'ADN amplifié avec l'amorce « sens »

Teleo-F (ACACCGCCCGTCACTCT) et « anti-sens » Teleo-R (CTTCCGGTACACTT-ACCATG) mesure ≃65pb (Figure 2.8) (Valentini et al. 2016). L'amplification de chaque échantillon est répliquée 12 fois, pour maximiser la détection d'espèces rares.

L'étape suivante, le multiplexage, consiste à ajouter à chaque séquence un tag (séquence de 8 nucléotides connus) unique à chaque réplica PCR de chaque échantillon terrain (Figure 2.9). Ainsi, lorsque plusieurs échantillons sont mélangés dans une librairie pour le séquençage, il est par la suite possible de réassigner chaque séquence à son échantillon d'origine.

La préparation des librairies et le séquençage sont réalisés par la compagnie FASTERIS en Suisse. La préparation des librairies est réalisée par ligation des index de librairies et adaptateurs de séquençage (Figure 2.9, Bohmann et al. 2021)). Le séquençage des échantillons utilisés dans cette thèse a été réalisé sur plusieurs plateformes : Illumina MiSeq, HiSeq ou NextSeq. Le séquençage est réalisé en *paired-end* (« double-lecture »), chaque séquence est lue dans le sens 5'-3' puis dans le sens 3'-5', afin d'avoir une double lecture des séquences.

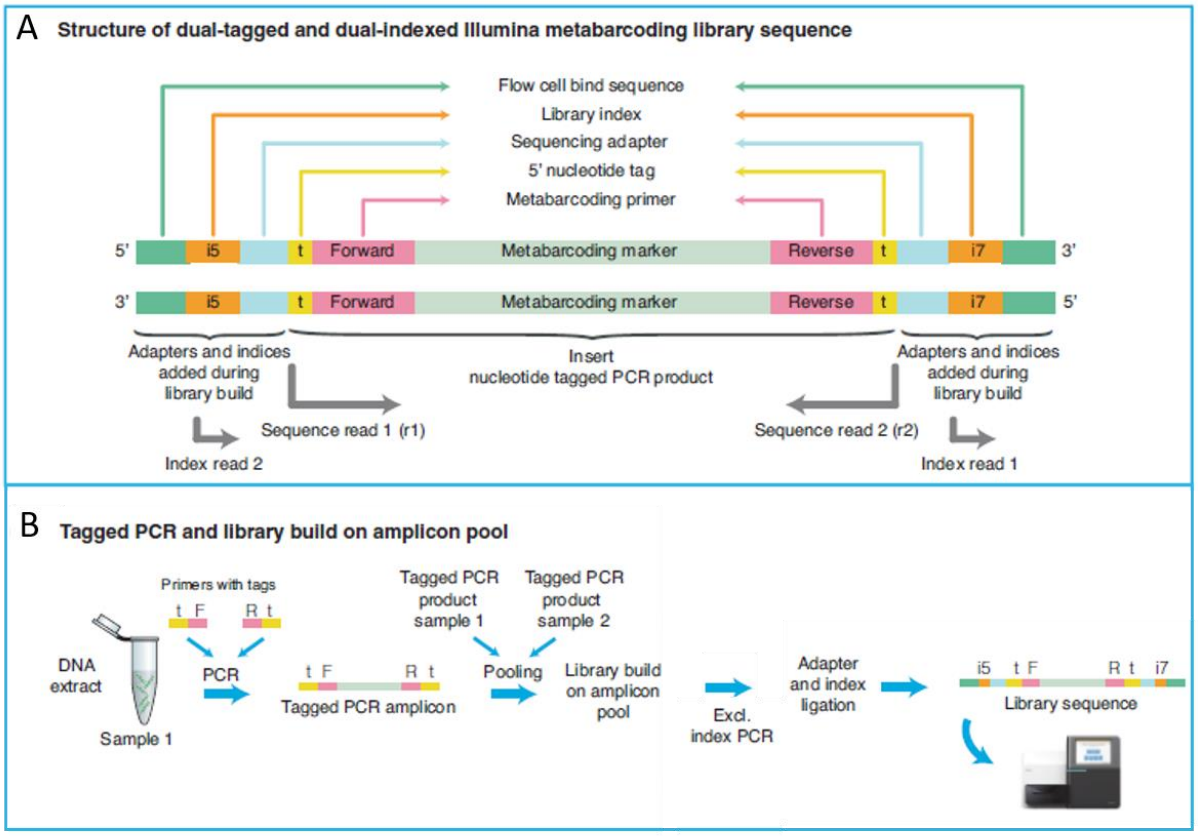


Figure 2.9. *Schéma de la composition d'une séquence de librairie Illumina, avec deux tags et deux index (A), Protocole d'ajout des index et adapteurs par ligation (B). Figure adaptée de Bohmann et al. (2021).*

4. Bio-informatique

Les données brutes issues du séquençage doivent ensuite subir plusieurs traitements bio-informatiques avant de pouvoir être exploitées pour des analyses écologiques. Ces traitements sont automatisés pour être standardisés et diminuer le temps de traitement des données brutes très volumineuses. Les traitements bio-informatiques sont composés de plusieurs étapes permettant d'obtenir une grande fiabilité dans les données :

1- L'alignement des séquences paired-end (« double-lecture ») : cette étape permet d'assembler les séquences « sens » et « anti-sens » et d'appliquer certains filtres sur la longueur de la séquence assemblée, ou sur la qualité du score FASTQ fourni par la plateforme de séquençage.

2- Le démultiplexage : cette étape consiste à identifier les tags uniques à chaque échantillon, à les retirer et ajouter le nom de l'échantillon dans l'entête de la séquence et à retirer les amorces afin de ne conserver que la séquence d'intérêt.

3- La déréplication : cette étape consiste à dérépliquer toutes les séquences 100% identiques au sein de chaque échantillon, à n'en garder qu'une et renseigner le nombre de séquences identiques comptées dans l'échantillon.

4- Le nettoyage des erreurs : au cours de cette étape, on cherche à éliminer les séquences issues d'erreurs de PCR ou de séquençage. Il est possible de filtrer les séquences selon leur longueur, leur score qualité ou selon leur abondance ou fréquence dans le jeu de données. Certaines séquences, appelées chimères, sont des erreurs liées à un assemblage de deux séquences différentes lors de la PCR. Certains programmes sont dédiés à la détection de ces erreurs et à leur nettoyage (DENOISE, DADA2... Callahan et al. 2016; Edgar, 2016).

5- L'assignation taxonomique : cette étape consiste à comparer les séquences à une base de référence afin d'assigner un taxon à chaque séquence. Plusieurs algorithmes permettent de réaliser cette étape, dont l'algorithme du plus proche ancêtre commun (*Lower Common Ancestor*, LCA, Aho et al. 1976). Dans le cas où la base de référence n'est pas complète et/ou certaines séquences peuvent être assignées à plusieurs espèces, cet algorithme assignera le taxon correspondant au rang le plus précis possible (genre, famille).

De nombreux programmes existent pour réaliser ces différentes étapes et très peu d'études ont comparé leurs performances. Pour les analyses du chapitre 2, j'ai sélectionné dans la

littérature les programmes et pipelines les plus fréquemment utilisés dans le cadre de l'analyse de données brutes d'ADNe de poissons et comparé leurs performances sur un jeu de données simulé et sur un jeu de données réel issus de l'échantillonnage en mer Méditerranée. À l'issue de cette comparaison, j'ai proposé un nouveau pipeline composé des programmes les plus performants pour chaque étape du traitement bio-informatique. Cependant, cette comparaison a été effectuée dans un contexte de base de référence complète pour l'assignation taxonomique. Les autres analyses de cette thèse comportant des données collectées à large échelle ou dans des écosystèmes tropicaux très divers, l'approche bio-informatique a dû être adaptée afin de travailler avec des MOTUs (= *molecular operational taxonomic unit*, regroupement de séquences selon un paramètre de similarité) plutôt que des espèces, car les bases de références sont largement incomplètes (Marques, Milhau, et al. 2020).

Pour cela, j'ai formaté, nettoyé et analysé les données des chapitres 3, 4 et 5 avec un pipeline bio-informatique développé par Marques, Guérin, et al. (2020) (Figure 2.10).

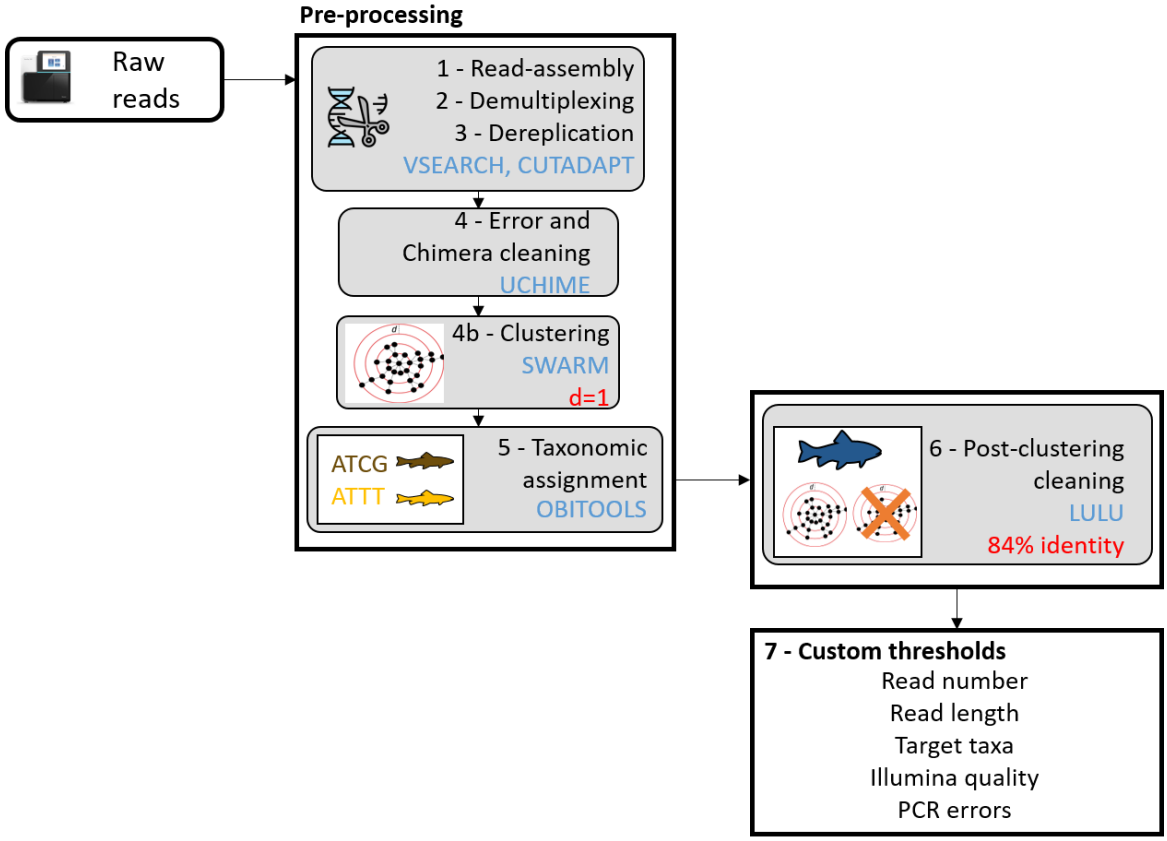


Figure 2.10. *Schéma des grandes étapes de bio-informatique et nettoyage des séquences incluses dans le pipeline de Marques et al 2020, utilisé pour les analyses en OTUs. Les programmes utilisés sont identifiés en bleu et les paramètres importants en rouge. Figure adaptée de Marques, Guérin, et al. (2020).*

Ce pipeline comporte les étapes décrites ci-dessus auxquelles sont ajoutées les étapes suivantes :

4b - Le clustering : cette étape a lieu avant l'assignation taxonomique et consiste à regrouper des séquences similaires dans le but de travailler avec des unités taxonomiques moléculaires (MOTUs) ou de nettoyer des erreurs.

6 - Le nettoyage post-clustering : réalisée après l'assignation taxonomique, cette étape consiste à finaliser le nettoyage des données en utilisant un algorithme spécialisé, ou en fixant des filtres sur l'abondance ou la fréquence des séquences dans les échantillons ou les réplicas PCR.

Lors du *metabarcoding* de l'ADNe, de nombreuses erreurs peuvent se produire et il est important de supprimer ces séquences erronées qui ne reflètent pas la réalité biologique de l'échantillon. Ces erreurs peuvent se produire lors de la PCR, où la combinaison de deux séquences peut former des chimères, ou des erreurs de réplication sur plusieurs nucléotides peuvent survenir et également lors du séquençage, où les nucléotides peuvent être mal lus par le séquenceur. La majorité des erreurs représentent des occurrences rares dans le jeu de données. Il est ainsi difficile de différencier les erreurs des vraies occurrences d'espèces rares (Patin et al. 2013). Il s'agit donc de trouver un compromis pour limiter le nombre de faux-positifs, sans supprimer trop d'occurrences réelles. Ajouter une étape de clustering peut permettre de corriger ces erreurs. Différents types de clustering existent : le clustering des séquences avec un seuil de similarité afin de délimiter les entités biologiques réelles avec un clustering non basé sur une similarité génétique fixe comme avec l'algorithme SWARM (Mahé et al. 2015) ou un clustering basé sur la définition de variants de séquences d'amplicon (*Amplicon Sequence Variant*, ASV) (Callahan et al. 2017).

Dans le pipeline de Marques et al 2020, le clustering est réalisé avec l'algorithme SWARM (Mahé et al. 2015). SWARM construit des chaines de séquences et définit chaque unité taxonomique (MOTU) en fonction d'un double paramètre : son abondance et sa proximité génétique avec d'autres séquences plus ou moins abondantes (Figure 2.11).

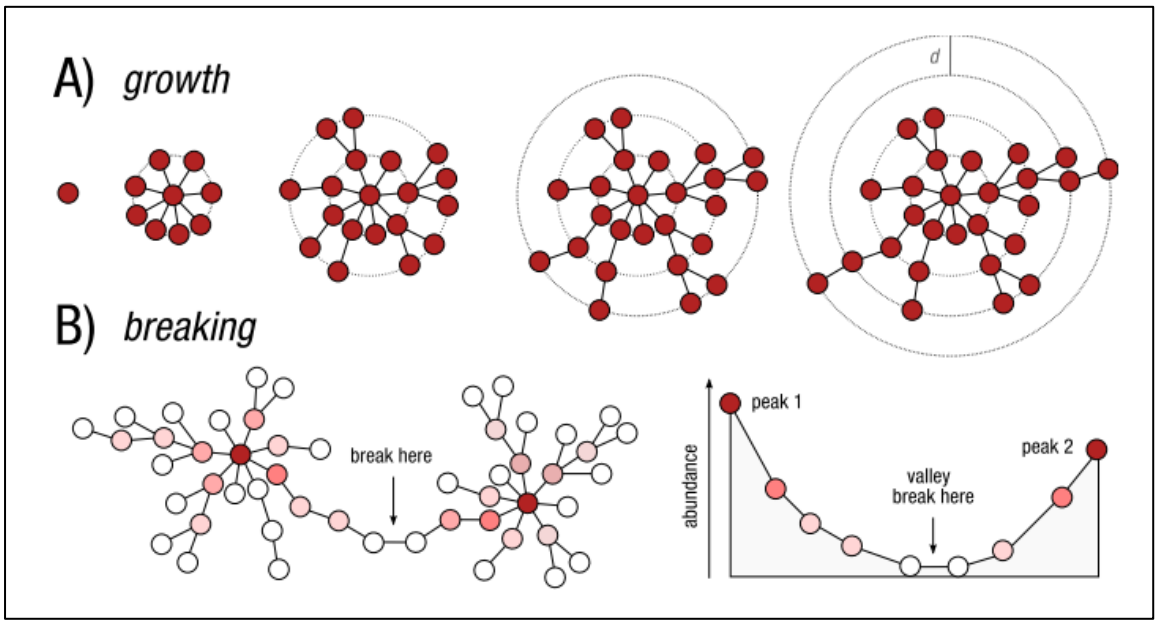


Figure 2.11. *Schéma du clustering par SWARM. (A) Swarm regroupe les amplicons par itérations, selon le seuil d définit par l'utilisateur, jusqu'à ce qu'aucun autre amplicon ne puisse être ajouté au MOTU. (B) Swarm coupe les MOTUs en fonction de leur abondance. Figure adaptée de Mahé et al. (2014).*

Le nettoyage post-clustering est d'abord réalisé avec l'algorithme LULU (Frøslev et al. 2017), qui utilise les scores de similarité avec les patrons de co-occurrence pour nettoyer les unités taxonomiques. Cet algorithme cherche si les MOTUs les plus abondants peuvent être combinés avec des MOTUs « parents » et cherche ainsi à réduire le nombre de MOTUs correspondants à la même entité taxonomique (Figure 2.12).

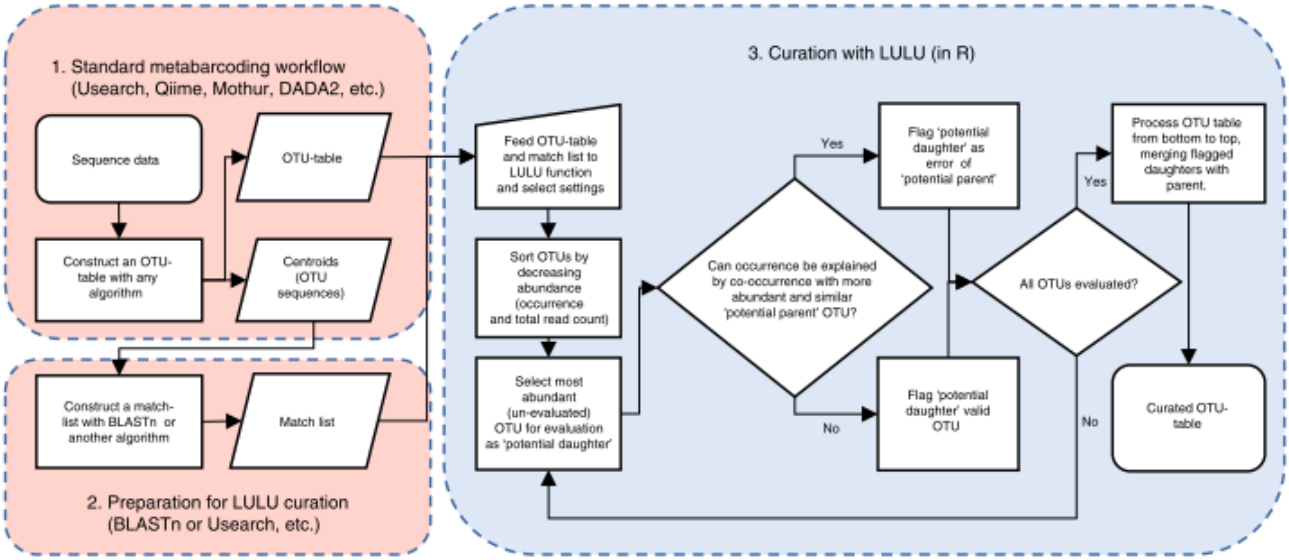


Figure 2.12. *Fonctionnement de l'algorithme LULU : 1) Construction de la table d'OTUs par l'utilisateur, 2) Construction de la liste d'assignation par l'utilisateur, 3) LULU cherche si les MOTUs les plus abondants peuvent être combinés avec des MOTUs « parents » selon les patrons de co-occurrence et abondance. Figure extraite de Frøslev et al. (2017).*

Puis j'applique d'autres filtres sur les données pour nettoyer plusieurs types d'erreurs (étape 7) :

- Les séquences issues de *tag-jump*, c'est-à-dire une assignation à un mauvais échantillon au moment du démultiplexage, due à une contamination des tags entre les puits lors du séquençage (Schnell et al. 2015),
- Les séquences en trop faible abondance, susceptibles de représenter des erreurs,
- Les séquences présentes dans une unique PCR, représentant probablement une erreur. Alors qu'une séquence présente dans au moins deux réplicas PCR est plus probablement une vraie occurrence d'espèce rare.

5. Analyses statistiques et modélisation

Dans le chapitre 2, les différents programmes bio-informatiques sont comparés en se basant sur les compositions de communautés attendues selon un jeu de donnée simulé. À partir des listes d'espèces obtenues par le traitement par chaque programme, les nombres de vrais positifs (espèce détectée et réellement présente dans l'échantillon), faux positifs (espèce détectée mais pas présente dans l'échantillon) et de faux négatifs (espèce non détectée mais présente dans l'échantillon) sont calculés et inclus dans le calcul de plusieurs indices (sensibilité, F-mesure et erreur quadratique moyenne – RMSE – sur les abondances de *reads*). Les différences significatives entre ces indices pour chacun des programmes permettent d'identifier les programmes les plus performants.

Dans le chapitre 3, j'ai estimé la diversité en MOTUs et en espèces de poissons coralliens, avec l'ADNe et les recensements visuels en plongée, à l'aide de courbes d'accumulations. J'ai également étudier l'importance de la partition de la diversité α et β entre les stations, sites et régions. Un des principaux challenges pour la modélisation des données d'ADNe à large échelle fut de prendre en compte les différences d'échantillonnage, tant au niveau spatial (nombre d'échantillons et réplicas par station, site, région) qu'au niveau méthodologique (méthode de filtration, volume filtré) dans les modèles. Les analyses de ce chapitre ont été réalisées en intégrant tous les échantillons d'ADNe, mais aussi en sous-échantillonnant le nombre d'échantillons par station, site et région selon les localités les moins échantillonnées, afin d'obtenir un échantillonnage équivalent et comparable entre les différentes régions. J'ai également étudié les proportions de MOTUs par famille dans chaque site, en fonction des régions, à l'aide d'une dbRDA (*distance-based Redundancy Analysis*). Dans cette analyse, j'ai pris en compte l'autocorrélation spatiale entre les échantillons en calculant des vecteurs de

corrélation (dbMEM, *distance-based Moran Eigenvectors Maps*) et en les intégrant comme variables conditionnelles dans le modèle.

Dans le chapitre 4, j'ai analysé les données d'ADNe à échelle mondiale, en calculant la diversité α de tous les poissons, des cryptobenthiques et des poissons de grande taille, mais aussi une diversité de séquences basée sur les distances génétiques et les nombres de Hill. J'ai modélisé ces métriques en fonction de variables environnementales, géographiques et socio-économique, à l'aide de modèles GLS (*generalized least-square models*). J'ai intégré les variables d'échantillonnage (volume et méthode) dans ces modèles afin de mesurer l'effet partiel de chacun des facteurs en contrôlant par les variables d'échantillonnage. Dans ce chapitre j'ai également étudié la β-diversité de composition en MOTUs et séquences entre les différents sites, à l'aide d'une dbRDA dans laquelle j'ai inclus les vecteurs de corrélations dbMEM. Toutes les analyses de ce chapitre ont aussi été réalisées sur un sous-échantillonnage standardisé entre régions.

Enfin, dans le chapitre 5, j'ai analysé les données ADNe provenant de Nouvelle-Calédonie seulement, ainsi que des données de caméras appâtées et acoustique, prélevées sur les monts-sous-marins et pentes externes jusqu'à 600m de profondeur. J'ai calculé plusieurs métriques de richesse taxonomique, abondance et biomasse, que j'ai modélisé en fonction de la profondeur, de variables physico-chimique et d'isolement, à l'aide de modèles BRT (*boosted regression trees*) qui prennent en compte les interactions entres variables (Elith et al. 2008). J'ai également modélisé les abondances d'espèces et MOTUs individuels, présents dans un grand pourcentage de stations de chaque habitat et chaque profondeur, à l'aide de modèle GJAM (*generalized joint-attribute models*) (Clark et al. 2017). Toutes les métriques de diversité ont ensuite été prédites à l'échelle de l'archipel Néo-Calédonien, sur 3 couches de profondeur (0-200, 200-400 et 400-600m). À partir de ces couches, j'ai paramétré un algorithme de priorisation spatiale en 3 dimensions pour sélectionner les zones à inclure prioritairement dans un plan de conservation protégeant 30% de chaque métrique de diversité.

GLOSSAIRE

Amorce : Courte séquence d'ADN ou ARN, complémentaire du début de la séquence d'ADN, servant de point de départ pour la synthèse du brin complémentaire de la séquence, par une ADN polymérase. L'utilisation d'une amorce « sens » et « anti-sens » permet de définir la séquence à amplifier.

Amplicon : fragment d'ADN amplifié par PCR.

Clustering : (= Analyse de groupement). Analyse permettant de subdiviser un jeu de données en entités, selon un paramètre de similarité. Chaque entité regroupe des séquences partageant des caractéristiques similaires.

MOTU : Unité taxonomique moléculaire opérationnelle (« *Molecular Operational Taxonomic Unit* »). Regroupement de séquences selon un paramètre de similarité, par 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 : Ensemble de programmes bio-informatiques traitant une tâche les uns après les autres. La sortie d'un programme est donnée en entrée au programme suivant. L'enchainement entre les programmes est automatisé, afin de ne fournir qu'un fichier d'entrée et récupérer les fichiers de sortie à la fin du traitement.

Read : lecture unique d'une séquence de paires de bases correspondant à tout ou partie d'un fragment d'ADN unique.

Chapitre 2 - Comparaison des outils bio-informatiques pour le traitement des données ADNe

Communauté de poissons sur le récif corallien de la passe de Boulari, Nouvelle-Calédonie

1. Préface

Recenser efficacement les communautés ichtyologiques, de la manière la plus fiable et exhaustive possible est nécessaire pour comprendre les patrons de distributions des poissons, étudier les impacts environnementaux et humains et suivre la dynamique des communautés. Le *metabarcoding* de l'ADNe est une méthode prometteuse pour réaliser ces recensements. Les inventaires taxonomiques obtenus par le *metabarcoding* dépendent cependant grandement des méthodes d'échantillonnage et des techniques utilisées en laboratoire, mais également des traitements bio-informatiques effectués sur les données issues du séquençage. S'il existe un consensus sur les étapes de traitement et de nettoyage à réaliser depuis les données brutes jusqu'à l'assignation des séquences à un taxon, de nombreux programmes ont été développés et sont couramment utilisés pour réaliser chacune de ces étapes et très peu d'études ont comparé les performances de ces programmes à la fois en terme de fiabilité et de vitesse.

Ce chapitre vise à comparer une sélection de programmes et pipelines bio-informatiques, en termes de diversité détectée (richesse taxonomique, identité des espèces et abondance des *reads*) et de temps d'exécution. Une recherche bibliographique a permis d'identifier les six étapes clés du traitement bio-informatique des séquences brutes issues du séquençage des échantillons ADNe (assemblage des séquences, démultiplexage, déréplication, filtre qualité, suppression des erreurs de PCR et séquençage et assignation taxonomique) et les programmes les plus fréquemment utilisés dans la littérature. Les programmes et pipelines sélectionnés pour cette comparaison sont compatibles avec les données de cette thèse : ADN mitochondrial de poisson, sur un fragment court du gène 12S. Grâce à l'utilisation de données simulées (espèces et nombres de *reads* connus dans chaque échantillons), d'une base de référence complète et d'une calibration avec le pipeline complet OBITools, j'ai calculé des indices de sensibilité et F-measure basés sur le nombre de vrais positifs (espèces détectées et réellement présentes dans l'échantillon), faux positifs (espèces détectées mais non présentes dans l'échantillon) et faux négatifs (espèces non détectées malgré leur présence dans l'échantillon) retrouvés dans les listes d'espèces obtenues par chaque programme, ainsi qu'un indice comparant les abondances relatives de *reads* attendues et observées et le temps d'exécution de chaque programme. Les meilleurs programmes pour chaque étape sont assemblés en un nouveau pipeline comparé à d'autres pipelines complets existants. Cette comparaison a ensuite été validée sur des données réelles provenant de la mer Méditerranée.

Les résultats montrent des performances similaires entre tous les programmes pour les cinq premières étapes. Le temps d'exécution est le seul facteur permettant de sélectionner le meilleur programme pour ces étapes. La différence maximale entre les temps d'exécution est de l'ordre de 400 (Tableau 3.1). Des différences significatives entre les programmes ont été obtenues pour l'étape d'assignation taxonomique seulement, où le programme Vsearch obtient de meilleurs résultats qu'OBITools et Sintax.

Tableau 3.1. *Synthèse des résultats obtenus par chaque programme à chaque étape de traitement. Les programmes en bleu sont ceux sélectionnés dans le pipeline assemblé.*

Etape	Programme	Temps d'exécution (min)	Sensibilité	F-measure	RMSE abondance
Assemblage des *reads*	OBITools	1234	0,953 ± 0,05	0,973 ± 0,03	1,06 ± 1,36
	Pear	62,7	0,953 ± 0,05	0,973 ± 0,03	1,05 ± 1,37
	Casper	22	0,958 ± 0,05	0,974 ± 0,03	1,04 ± 1,35
	Fastq-join	4,6	0,953 ± 0,05	0,973 ± 0,03	1,06 ± 1,37
	Flash	3,7	0,953 ± 0,05	0,973 ± 0,03	1,05 ± 1,37
	Fastp	3,5	0,952 ± 0,05	0,968 ± 0.03	1,76 ± 1,4
	Vsearch	3,1	0,953 ± 0,05	0,973 ± 0,03	1,05 ± 1,39
Démultiplexage	OBITools	488	0,953 ± 0,05	0,973 ± 0,03	1,06 ± 1,36
	Cutadapt	30	0,956 ± 0,05	0,975 ± 0,02	1,22 ± 1,58
Déréplication	OBITools	198	0,953 ± 0,05	0,973 ± 0,03	1,06 ± 1,36
	Vsearch	0,8	0,953 ± 0,05	0,973 ± 0,03	1,06 ± 1,37
Filtre qualité	Fastp	2,25	0,953 ± 0,05	0,973 ± 0,03	1,06 ± 1,37
	OBITools	2,2	0,953 ± 0,05	0,973 ± 0,03	1,06 ± 1,36
	Prinseq	1,3	0,953 ± 0,05	0,973 ± 0,03	1,06 ± 1,37
	Cutadapt	1,1	0,953 ± 0,05	0,973 ± 0,03	1,06 ± 1,37
	Vsearch	0,2	0,953 ± 0,05	0,973 ± 0,03	1,06 ± 1,37
	Flexbar	0,2	0,953 ± 0,05	0,973 ± 0,03	1,06 ± 1,37
Suppression d'erreur	OBITools	17,4	0,953 ± 0,05	0,973 ± 0,03	1,06 ± 1,36
	Swarm	0,45	0,948 ± 0,05	0,971 ± 0,03	1,06 ± 1,37
	Vsearch	0,4	0,937 ± 0,06	0,966 ± 0,03	1,07 ± 1,37
Assignation taxonomique	OBITools	58	0,953 ± 0,05	0,973 ± 0,03	1,06 ± 1,36
	Sintax	1,8	0,575 ± 0,17	0,715 ± 0,13	3,26 ± 3,15
	Vsearch	0,14	0,97 ± 0,05	0,981 ± 0,03	0,39 ± 0,67

L'assemblage des meilleurs programmes de chaque étape produit des résultats similaires ou meilleurs que les autres pipelines complets existants. Cette étude montre que le choix de programmes influence les estimations de diversité et les listes taxonomiques obtenues. Cela

démontre également la nécessité d'appliquer ces comparaisons à d'autres programmes, groupes taxonomiques et marqueurs génétiques.

La vitesse de traitement et la précision de l'assignation taxonomique sont donc deux paramètres à optimiser pour l'analyse de larges jeux de données ADNe. L'application de nouvelles méthodes d'intelligence artificielle pourrait permettre de répondre à ce challenge. Une étude réalisée sur des données ADNe collectées dans plusieurs rivières tropicales de Guyane, a évalué la capacité des réseaux de neurones convolutifs (CNN) à traiter les courtes séquences d'ADNe 12S et à les assigner à un taxon (Manuscrit A1, voir Annexes). La vitesse et la précision des CNN ont été comparées avec celles du pipeline bio-informatique OBITools. Les assignations taxonomiques obtenues avec les CNN sont comparables à celles d'OBITools, avec de fortes corrélations entre les compositions spécifiques obtenues.

Les CNN ont permis le traitement de fichiers bruts à un rythme d'environ 1 million de séquences par minute, ce qui est environ 150 fois plus rapide qu'avec OBITools. Compte tenu de la bonne performance des CNN dans l'écosystème très diversifié considéré dans cette étude, le développement de CNN plus élaborés pourrait permettre une application rapide et facile pour les futurs inventaires de biodiversité utilisant l'ADNe.

2. Manuscrit A

Publié dans *Molecular Ecology Resources* :

Mathon, L., Valentini, A., Guérin, P-E., Normandeau, E., Noel, C., Lionnet, C., Boulanger, E., Thuillier, W., Bernatchez, L., Mouillot, D., Dejean, T., & Manel, S. (2021). Benchmarking bioinformatic tools for fast and accurate eDNA metabarcoding species identification. *Molecular Ecology Resources*, *21*(7), 2565-2579.

Received: 22 January 2021 | Revised: 10 May 2021 | Accepted: 11 May 2021

DOI: 10.1111/1755-0998.13430

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Benchmarking bioinformatic tools for fast and accurate eDNA metabarcoding species identification

Laetitia Mathon[1,2] | Alice Valentini[2] | Pierre-Edouard Guérin[1] | Eric Normandeau[3] | Cyril Noel[4] | Clément Lionnet[5] | Emilie Boulanger[1,6] | Wilfried Thuiller[5] | Louis Bernatchez[3] | David Mouillot[6,7] | Tony Dejean[2] | Stéphanie Manel[1]

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

[2]SPYGEN, Savoie Technolac, Le Bourget du Lac, France

[3]Université Laval, IBIS (Institut de Biologie Intégrative et des Systèmes), Québec, QC, Canada

[4]IFREMER - IRSI - Service de Bioinformatique (SeBiMER), Plouzané, France

[5]Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LECA, Laboratoire d'Ecologie Alpine, Grenoble, France

[6]MARBEC, Univ. Montpellier, CNRS, IRD, Ifremer, Montpellier, France

[7]Institut Universitaire de France, IUF, Paris, France

Correspondence
Laetitia Mathon, CEFE, Univ. Montpellier, CNRS, EPHE-PSL University, IRD, Montpellier, France.
Email: laetitia.mathon@gmail.com

Funding information
PIA grant managed by the ADEME, Grant/Award Number: 1882C0046; French Agence Nationale de la Recherche (ANR) through the GlobNets, Grant/Award Number: ANR-16-CEO2-0009; 'Investissement d'Avenir' grants managed by the ANR, Grant/Award Number: ANR-15-IDEX-02 and ANR-10-LAB-56

Abstract

Bioinformatic analysis of eDNA metabarcoding data is a crucial step toward rigorously assessing biodiversity. Many programs are now available for each step of the required analyses, but their relative abilities at providing fast and accurate species lists have seldom been evaluated. We used simulated mock communities and real fish eDNA metabarcoding data to evaluate the performance of 13 bioinformatic programs and pipelines to retrieve fish occurrence and read abundance using the 12S mt rRNA gene marker. We used four indices to compare the outputs of each program with the simulated samples: sensitivity, F-measure, root-mean-square error (RMSE) on read relative abundances, and execution time. We found marked differences among programs only for the taxonomic assignment step, both in terms of sensitivity, F-measure and RMSE. Running time was highly different between programs for each step. The fastest programs with best indices for each step were assembled into a pipeline. We compared this pipeline to pipelines constructed from existing toolboxes (OBITools, Barque, and QIIME 2). Our pipeline and Barque obtained the best performance for all indices and appear to be better alternatives to highly used pipelines for analysing fish eDNA metabarcoding data when a complete reference database is available. Analysis on real eDNA metabarcoding data also indicated differences for taxonomic assignment and execution time only. This study reveals major differences between programs during the taxonomic assignment step. The choice of algorithm for the taxonomic assignment can have a significant impact on diversity estimates and should be made according to the objectives of the study.

KEYWORDS
benchmark, bioinformatics, eDNA, metabarcoding, sensitivity, species identification

1 | INTRODUCTION

Environmental DNA (eDNA) metabarcoding is a promising approach to identify species within communities and can be used to evaluate biodiversity through a variety of estimators (Boulanger et al., 2021; Deiner et al., 2020; Pawlowski et al., 2018). The approach is based on the collection of environmental samples (e.g., soil, air or water) that contain the target organisms' DNA. After DNA extraction, DNA amplification with primers designed for a specific taxonomic group is performed and submitted to high-throughput sequencing (Deiner

et al., 2017; Taberlet et al., 2018). The resulting sequencing data typically contains millions of amplicon DNA fragments. Bioinformatic programs are then used to (i) clean the data, (ii) associate fragments to samples when amplicons are pooled in a single library and (iii) produce either a matrix of read counts per species occurring within each sample using a reference database or a matrix of operational taxonomic units (OTUs) occurring within each sample. With decent completeness of the genetic reference database, eDNA metabarcoding can provide accurate representation of the taxonomic composition within samples (Djurhuus et al., 2020; Marques et al., 2020; Minamoto et al., 2012). Nevertheless, many biases can reduce the performance of such approaches which need to be controlled for, such as PCR and sequencing errors, gaps in reference databases, different species with identical sequences in the amplified region, etc. (Kwok & Higuchi, 1989; Schnell et al., 2015; Zinger et al., 2019). Several of these biases can be mitigated with posterior bioinformatic analyses by implementing appropriate filters. Yet, comparative quantitative analyses are still lacking on the different key steps of bioinformatic pipelines.

A literature search with the keywords "environmental DNA", "metabarcoding", "community", "bioinformatics analysis" (Method S1) identified six steps from raw sequence fragments to final identifications: paired-end reads merging, demultiplexing, dereplication, quality filtering, removal and correction of PCR/sequencing errors and taxonomic assignment (Figure 1). The order of those steps can vary depending on the pipeline being used. Some of these steps can have a strong impact on the resulting taxonomic composition and consequently on biodiversity estimation (Bonder et al., 2012; Calderón-Sanou et al., 2020). Therefore, choosing the most appropriate bioinformatics program producing accurate, fast and sensitive taxa identifications is a crucial step (Pauvert et al., 2019). Until now, no consensus among existing bioinformatic programs and pipelines has emerged to choose the most appropriate for analysing eDNA metabarcoding data (Bazinet & Cummings, 2012; Gardner et al., 2019; Lindgreen et al., 2016; Peabody et al., 2015; Prodan et al., 2020; Sczyrba et al., 2017; Siegwald et al., 2017). In particular, there is no standard or recommendation related to existing programs, based on quantitative comparisons, in the context of aquatic eDNA data and particularly so for Teleostean fishes.

With more than 32,000 species, Teleostean fishes are the largest group of vertebrates (www.fishbase.org). Worldwide, a growing number of fish species and populations are threatened and decreasing in size due to overfishing as well as habitat degradation (Yan et al., 2021). With more than 60% of the publications on eDNA dealing with vertebrate monitoring, fishes represent the most studied group using eDNA approaches (Tsuji et al., 2019). As a result, they represent a relevant candidate taxonomic group to compare eDNA metabarcoding programs and pipelines.

eDNA metabarcoding based on water samples was first applied to monitor fish species both in marine (Thomsen et al., 2012) and freshwater environments (Robson et al., 2016). Currently, there is increasing interest in this technique for characterizing fish diversity

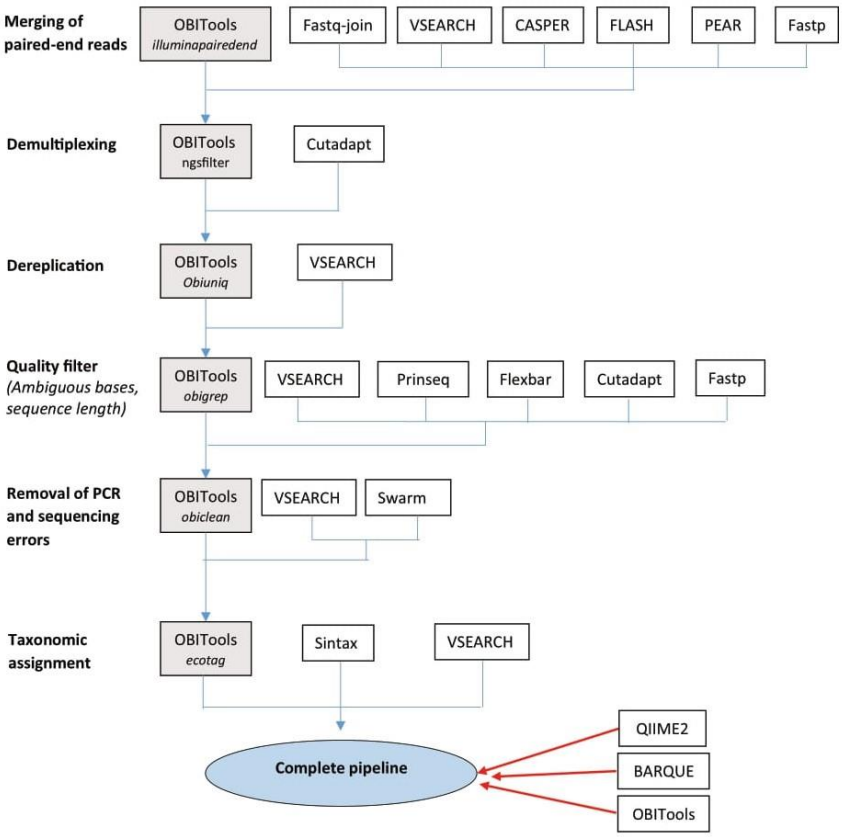


FIGURE 1 Protocol of the step-by-step program comparisons. The comparison of each program (white boxes) is arranged sequentially according to the suite of steps in the OBITools pipeline (grey boxes)

(Berger et al., 2020; Jerde et al., 2019; Juhel et al., 2020; McElroy et al., 2020; Sigsgaard et al., 2017) in particular when classical methods are too invasive or do not perform well, as is the case for rare species or those that inhabit the deep sea. Many primer pairs have been developed to amplify different mitochondrial DNA fragments of fish DNA. The primer pair used in this study (teleo 12S mt rRNA; Valentini et al., 2016) is one of the most frequently used and has been proven effective in detecting rare biodiversity and discriminating species in European fluvial ecosystems (Civade et al., 2016; Collins et al., 2019; Pont et al., 2019). Since teleo 12S mt rRNA is a widely used primer set, it is a good candidate to explore the efficiency of different bioinformatic programs.

In this context, the goals of this study were to: (i) explore common bioinformatic tools including one-step programs used in the different steps of metabarcoding data analysis and integrated pipelines, (ii) assess the ability of these programs to accurately and rapidly retrieve the species composition (occurrences and abundances) of representative fish communities, and (iii) assemble the best-performing programs for each step into a custom pipeline, and compare its performance to three other pipelines designed for metabarcoding analysis (Barque, OBITools and QIIME2-based) using both simulated and real data.

2 | MATERIALS AND METHODS

2.1 | Simulated fish communities

We simulated 29 different species assemblages, hereafter designed as "samples", using the program Grinder (Angly et al., 2012). In these samples, the presence of 18 to 51 fish species were included (among a variety of cartilaginous and bony fishes: Actinopteri, Chondrichthyes, Cladistia, Cyclostomata and Sarcopterygii). Each sample was composed of random species from various classes, orders and families. Some samples contained several species belonging to the same genus (samples 3, 4, 6, 14, 19, 21 and 28). Relative abundances were attributed to each sequence in a given sample to represent real data sets. In sample 1, the most abundant species represented 50% of the reads and the other species had decreasing read abundances with each having half as many sequences as the previous one. For six samples (2 to 7), species relative abundance were based on real samples from both large and small rivers (Cantera et al., 2019; Milhau et al., 2019; Pont et al., 2018) and marine ecosystems (Polanco Fernández et al., 2020). In samples 9 and 25, all fish species (30 and 18, respectively) had equal abundance. For the other samples, abundances were attributed such that some sequences were very abundant and others rare (see Table S1). We simulated amplicons of the 26 to 60 bp 12S mt rRNA region containing the teleo primer. Species composing each sample were selected randomly from a custom reference database of 2,070 sequences of the fish mitochondrial 12S rRNA gene (downloaded from GenBank, on the 23/01/2019). Each simulated assemblage was replicated 12 times, with the same species and abundance composition, to mimic PCR replicates variability. To obtain a data set similar to those obtained from high-throughput sequencing, we applied an Illumina error model with 98% substitutions and 2% insertions/deletions. A total of 45,000 reads were simulated in each sample replicate, for a total of 15,660,000 reads in the complete data set. All the Grinder FASTQ files containing the interleaved amplicon sequences and quality scores for each simulated assemblage were concatenated to obtain output files similar to those obtained after a Miseq sequencing of one library of pooled PCR samples. Grinder also produced a text file describing the abundance of each sequence in the simulated replicates for each sample. These files were used as expected species composition and relative abundances to compute sensitivity, F-measure and RMSE on reads abundance using the outputs of each program involved in the comparative analysis.

The input and output data as well as the code written to construct the simulated data set are available as a GitHub repository at: https://github.com/lmathon/metabarcoding_data_simulation.

2.2 | Selected programs and steps

We selected some of the most cited programs for each step through a literature review with the keywords "bioinformatics", "metabarcoding" and the name of the analysis step (see Table S2 and Method S1). The most cited programs we considered are listed in Table S3. Here, we use the word "program" to define an independent binary or package dedicated to one of the six steps of eDNA metabarcoding analysis. The word "pipeline" refers to a program or set of programs that proceeds to analysing all the steps from raw read assembly to species identification.

The characteristics of our simulated data set (Illumina sequencing, mitochondrial 12S mt rRNA gene region, very short amplicons, fish DNA) excluded many programs from our comparisons since they were not compatible with such data. For example, Mothur is specialized in analysing 16S rRNA gene sequences, TagCleaner and DeML do not support paired-end reads, and Kaiju is specialized in protein-level assignment (see Table S3). QIIME is no longer maintained and some studies have shown that it requires a long execution time (Bonder et al., 2012) and the plugins used give a F-measure worse than that of QIIME2 (Gardner et al., 2019). As a result, QIIME2 was tested instead of QIIME. USEARCH (Edgar, 2010) is a widely used program but is only available as open-source in its memory-limited 32-bit version which does not meet our open-source requirement. Instead, we used VSEARCH, the open-source equivalent (Rognes et al., 2016).

To compare and identify the best programs, each bioinformatic step was tested successively by changing the program that performs this part of the analysis and by maintaining all others fixed. We decided to use the OBITools pipeline (Boyer et al., 2016) as a backbone for the performance tests (Figure 1), because this pipeline generates reliable results for fish eDNA metabarcoding data (Bylemans et al., 2018; Pont et al., 2018; Sales et al., 2019). All OBITools programs composing the fixed pipeline were evaluated in parallel with

the other programs tested for each step. The steps were defined as follows: (i) Merging, where forward and reverse reads were aligned to create a single consensus sequence, (ii) Demultiplexing, which assigned each sequence to its sample and removed the primers, (iii) Dereplication, or keeping only one representing sequence and count for strictly identical sequences, (iv) Quality filtering, which removed sequences that were too short or contained ambiguous bases, (v) Removing variants/PCR errors, so taking into account that real haplotypes and variants due to intractable sequencing/PCR errors should be grouped to avoid overestimating species richness, and (vi) Identifying taxa, where a taxon was assigned to each sequence. For this last step, we used the same reference database as the one used for the Grinder simulations. Only sequences assigned to the species level with more than 98% similarity were considered for species identification. The 10 programs compared in this study are listed in Table 1. Parameters chosen at each step for each program being compared can be found in Table 2. Since Grinder cannot simulate chimeras, the chimera removal step was not tested in this study. Each program was run on a cluster using Ubuntu 18.04.3 LTS with 128GB RAM and 1 CPU to obtain comparable execution times. Data and software commands necessary to reproduce this study are available at: https://github.com/lmathon/eDNA--benchmark_pipelines.

2.3 | Performance evaluation

Each program was evaluated by calculating indices that quantify its ability to produce accurate species lists (sensitivity and F-measure) and relative read abundances expected from the known or ground-truth simulated communities (RMSE). The execution time of each program was also recorded. Details on the computation of execution times can be found in Method S2.

After taxonomic assignment, sequence counts were aggregated by species and by replicate. For each tested program, the number of false positives (FP, species present in the output of the program but not in the initial community), true positives (TP, species present in both the output of the program and the initial community) and false negatives (FN, species present in the initial community but not in the output of the program) were calculated. We then computed the sensitivity (equation 1) and F-measure (equation 2) indices for each replicate of each sample from FP, TP and FN, as suggested in Gardner et al. (2019):

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (1)$$

$$F - \text{measure} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \quad (2)$$

These indices present complementary advantages. Sensitivity is relevant to identify programs missing rare taxa while the F-measure highlights programs detecting false positives. For each sample, we derived the mean and the standard error of these indices for each replicate. The standard error among replicates represents intrasample variability for a given program. The mean and standard error among the 29 samples represent the intersample variability for the results of each program. We compared the indices averaged across all samples between programs to determine if there was a significant difference of performance between programs. Since the data were not normally distributed, the mean comparison was carried out with a nonparametric Kruskal-Wallis test followed by a Dunn post-hoc test. Sensitivity

TABLE 1 Bioinformatic programs and pipelines used for the comparison

Program	Step	References	Source code
OBITools	Pipeline	Boyer et al. (2016)	https://git.metabarcoding.org/obitools/obitools/wikis/home
Barque	Pipeline	–	https://github.com/enormandeau/barque
QIIME2	Pipeline	Bolyen et al. (2019)	https://docs.qiime2.org
VSEARCH	M, Dr, QF, E, T[†]	Rognes et al. (2016)	https://github.com/torognes/VSEARCH
Pear	M	Zhang et al. (2014)	http://www.exelixis-lab.org/web/software/pear
FLASh	M	Magoč and Salzberg (2011)	https://sourceforge.net/p/flashpage
CASPER	M	Kwon et al. (2014)	http://best.snu.ac.kr/casper/
Fastq-join	M	Aronesty (2013)	https://github.com/brwnj/fastq-join
Fastp	M, QF	Chen et al. (2018)	https://github.com/OpenGene/fastp
Cutadapt	Dm, QF	Martin (1994)	https://github.com/marcelm/cutadapt https://cutadapt.readthedocs.io
Prinseq	QF	Schmieder and Edwards (2011)	http://prinseq.sourceforge.net
Flexbar	QF	Dodt et al. (2012)	https://github.com/seqan/flexbar/wiki
Swarm	E	Mahé et al. (2015)	https://github.com/torognes/swarm
Sintax	T	Edgar (2016)	https://www.drive5.com/usearch/manual/cmd_sintax.html

Abbreviations: Dm, demultiplexing; Dr, dereplication; E, PCR/sequencing error removal; M, merging; QF, quality filtering; T, taxonomic assignment.

TABLE 2 Description of the six analyses steps, their objectives, and the parameters set for each program compared

Analysis step	Objective	Program	Parameters
Merging reads	Assemble forward and reverse reads. Min. overlap = 10 Max. overlap = 150 (Max mismatch = 25%)	illuminapairedend	--
		VSEARCH	--fastq_mergepairs –threads 1 –fastq_maxdiffpct 25
		Flash	-m 10 –M 150 –x 0.25 –t 1
		Fastq-join	-m 10 –p 25
		CASPER	-w 10 –g 0.25 –t 1 -j
		Pear	-v 10 –c 0 –n 0 –j 1
		Fastp	--merge –overlap_len_require 10 --overlap_diff_limit 15 –w 1 --overlap_diff_limit_percent 25
Demultiplexing	Assign reads to sample and remove primers. 0 mismatch on tags, max. 2 mismatches on primers	ngsfilter	-e 2
		Cutadapt	-g –j 1 –e 0 (or 0.12) –O 8 (or 15) --revcomp
Dereplication	Gather strictly identical sequences and keep count of reads abundance	obiuniq	-m sample
		VSEARCH	--derep_fulllength –sizeout –fasta_width 0 –threads 1 --minseqlength 1
Quality filtering	Filter sequences shorter than 20 bp and/or containing ambiguous bases	obigrep	-s'[ATCG]+$'–l 20
		VSEARCH	--fastx_filter –fastq_maxn 0 –fastq_minlength 20 –threads 1
		Cutadapt	-m 20 –max_n 0 –j 1
		Flexbar	--max-uncalled 0 –n 1 –min-read-length 20
		Prinseq	-min_length 20 –ns_max_n 0 -noniupac
		Fastp	-n 0 –l 20 –w 1
Error removal	Identify and remove PCR and sequencing errors (by clustering with proportion of variants/parents)	obiclean	-r 0.05 –H
		VSEARCH	--cluster_unoise --sizein –minsize 1 –sizeout –threads 1 –unoise_alpha 2 --minseqlength 20
		Swarm	-z –f –t 1
Taxonomic assignment	Assign reads to a species, with a 98% similarity threshold with the best match	ecotag	–m 0.98
		VSEARCH	--usearch_global –id 0.98 –fasta_width 0 –dbmask none --maxaccepts 20 --maxrejects 20 –blast6out –maxhits 20 --top_hits_only --qmask none --minseqlength 1 –threads 1 –dbmatched --matched
		Sintax	-sintax_cutoff 0.98 –threads 1

and F-measure were also calculated after removing singletons from the data set (sequences with only 1 read in the pipeline outputs).

Here, we are referring to species abundance as the number of reads assigned to a given species in each sample replicate. For each replicate of each sample, the relative abundance of each species was calculated from the total number of reads and compared to the expected relative abundances for each species (simulated by the Grinder program). The root mean square error (RMSE) was then calculated for each abundance comparison. This index quantifies the level of dissimilarity between two lists of abundances: the lower the RMSE is, the more similar the observed and expected relative abundances are. The mean RMSE per sample was calculated as well as the associated standard error. Sensitivity, F-measure and RMSE were also calculated per sample, after summing the species counts in the twelve replicates. Statistical analyses were carried out with R v3.5.3.

2.4 | Comparing assembly of best programs to full pipelines

A complete pipeline was built by assembling the most performant programs for each step based on the performance indices to detect

species occurrence, to retrieve the relative read abundances and the execution time. Formatting scripts were written when necessary to facilitate the transition between programs. This pipeline was compared to other pipelines, namely BARQUE v1.6.2, OBITOOLS v1.2.13 and QIIME2 (Bolyen et al., 2019). Since QIIME2 is a toolbox, the results will be dependent on the plugins used. Here, we used demux for demultiplexing, cutadapt for primer removal, DADA2 (Callahan et al., 2016) for error removal, dbotu-q2 for ASV clustering and sklearn-classifier for taxonomic assignment. BARQUE uses trimmomatic for filtering, Flash for merging, its own python script to split amplicons, and VSEARCH for taxonomic assignment. The steps, programs, and parameters used by BARQUE and QIIME2-based pipelines can be found in Table 3. Because BARQUE takes demultiplexed reads as inputs, the demultiplexing was performed upstream with Cutadapt using the same parameters as for our assembled pipeline. Each pipeline was run using 16 CPUs. The same performance indices (sensitivity, F-measure and RMSE on reads abundance) were calculated for the outputs of each pipeline and compared.

2.5 | Illustrating the benchmark on real data

The same comparison process was run on an empirical data set obtained from the Mediterranean Sea. The eDNA samples were collected on the 5 June 2018 in four replicates within the no-take reserve of Carry-le-Rouet, at 5 km, and at 10 km outside the reserve, for a total of 12 samples (Boulanger et al., 2021). For each sample, 30 L of seawater were continuously collected along a 2 km transect from approximately 1m below the surface. Transects were conducted close to the coastline and the substrate to ensure the sampling of coastal organisms. Seawater samples were filtered on site using a VigiDNA 0.2 µM cross flow filtration capsule (SPYGEN). Immediately after filtration, the capsule was drained by filtering air, filled with 80 ml of CL1 buffer (SPYGEN) and stored at room temperature until the extraction. DNA extraction, amplification (12 replicates per sample) and sequencing followed the protocol described in Polanco Fernández et al. (2020). The different programs and pipelines were run on the raw sequences obtained after sequencing. The reference database used for the taxonomic assignment was built by performing in silico PCR with teleo primers using ecoPCR (Boyer et al., 2016) on the entire public database ENA (Leinonen et al., 2011; release 141) and by adding sequences from Mediterranean species sequenced by our group (Boulanger et al., 2021). Since the information about actual read abundances in the environment is unknown, it was not possible to measure the RMSE index. Hence, only the sensitivity and F-measure indices were measured before and after removing singletons in samples. To do so, fish species lists obtained by each program or pipeline were compared to lists of fish species identified by underwater visual census in Carry-le-Rouet reserve and outside, during several campaigns in 2018 (Charbonnel et al., 2020). Those lists obtained by independent sampling methods were considered as the expected species occurrences. To measure comparable execution time, each individual program was run using 1 CPU, and each pipeline was run using 16 CPUs.

3 | RESULTS

3.1 | Sensitivity, F-measure and RMSE on abundances

For each program tested, a mean index was estimated by averaging raw values of indices across replicates and samples.

TABLE 3 Programs and parameters used in the complete pipelines compared

Pipeline	Step	Program used	Parameters
QIIME2	Demultiplexing	demux emp-paired	--p-golay-error-correction FALSE
	Primer removal	cutadapt trim-paired	--p-error-rate 0.12 --p-overlap 16
	Filtering and denoising	Dada2 denoise-paired	--p-trunc-len-f 0 --p-trunc-len-r 0 --p-trunc-left-f 0 --p-trunc-left-r 0 --p-max-ee-f 2 –p-max-ee-r 2 --p-trunc-q 2 --p-chimera-method none
	OTU calling	dbotu-q2 call-otus	--p-gen-crit 0.1 --p-abund-crit 0 --p-pval-crit 0.005
	Taxonomic assignment	Feature-classifier classify-sklearn	--p-confidence 0.7
Barque	Filter and trim raw reads	Trimmomatic	Min_hit_length 16 Crop_length 80
	Merge paired-end reads	Flash	-t 1 –z –m 20 –M 280
	Split amplicon	Python script split_amplicons_one_file.py	Max_primer_diff 8
	Taxonomic assignment	VSEARCH –usearch_global	--qmask none –dbmask none –id 0.98 –maxaccepts 20 –maxrejects 20 –maxhits 20 –minseqlength 20 –query_cov 0.6 –fasta_width 0

For the merging, demultiplexing, dereplication, read filtering and error removal steps, the sensitivity, F-measure and RMSE were not significantly different between the programs (Figures S1–S3). The mean sensitivity obtained with the full OBITools pipeline was 0.94 and ranged from 0.78 to 1 with a mean standard error per sample of 0.004. The mean F-measure obtained with OBITools was 0.97 (ranged between 0.88 and 1). The mean RMSE between the relative abundances obtained for each replicate and the expected relative abundances was 1.1 with OBITools (ranged between 0.09 and 4.6).

We found significant differences between programs only for the taxonomic assignment step (Figure 2). Taxonomic assignment with Sintax produced significantly lower sensitivity (0.57, Figure 2a, p = 4.8e-13) and F-measure (0.71, Figure 2b, p = 7.1e-13) and higher RMSE (3.2, Figure 2c, p = 2e-08) than VSEARCH –usearch_global and ecotag. VSEARCH –usearch_global provided a significantly higher mean sensitivity than ecotag (0.97), and significantly lower mean RMSE (0.4). The assignment program VSEARCH was therefore more accurate when evaluating community composition and read abundances than ecotag.

Sensitivity and F-measures, after removing singletons, showed the same pattern but were lower due to less TP and more FN (Figures S4–S5). Sensitivity and F-measure per sample were higher and RMSE lower due to the increased detection of rare species when pooling the 12 replicates (Figures S6–S8), but the difference between programs showed the same patterns.

3.2 | Execution time

Execution time varied importantly between programs (Figure 3). For all but one step, OBITools programs were the slowest, sometimes by a factor of more than 200. The fastest program for merging was VSEARCH (3.1 min, Figure 3). Demultiplexing with Cutadapt was faster than with ngsfilter (30 min and 488 min respectively, Figure 3). Execution time of the sequence dereplication step was 198 min with obiuniq, and 0.8 min only with VSEARCH (Figure 3). VSEARCH and Flexbar were faster than other programs to filter reads (0.2 min). Execution time of the PCR and sequencing error removal step lasted 17.4 min with obiclean, while Swarm and VSEARCH –cluster_unoise ran in 0.4 min (Figure 3). Sintax and VSEARCH –usearch_global executed the assignment in 1.8 and 0.14 min respectively while ecotag (OBITools) ran in 58 min on our simulated data set (Figure 3).

3.3 | Comparison between pipelines

From the step comparison results between programs, we selected the best ones in terms of sensitivity, F-measure, RMSE on the abundance, and execution time. Since indices varied in the same direction, the selection was straightforward. These selected programs were integrated in a pipeline following the order of the steps shown in Figure 1. This custom pipeline was composed of VSEARCH –fastq_mergepairs for assembling the reads, Cutadapt for demultiplexing, VSEARCH –derep_fulllength for the dereplication, VSEARCH –fastx_filter for the quality filtering, Swarm for the suppression of PCR and sequencing errors and VSEARCH –usearch_global for the taxonomic assignment.

Our assembled pipeline obtained a mean sensitivity of 0.97, the same as Barque (0.97), higher than OBITools (0.94) and significantly higher than QIIME2-based (0.9) (Figure 4a, p = 6.4e-03). The mean F-measure was the same for the assembled pipeline and Barque (0.98) and significantly higher than OBITools (0.97, Figure 4b, p = 0.05) and QIIME2-based (0.94). Barque and our assembled pipeline were also the best pipelines to recover relative abundances, and mean RMSE were not significantly different, with 0.31 for Barque and 0.44 for our pipeline, while mean RMSE were significantly higher for OBITools (1.1) and QIIME2-based (1.24) (Figure 4c, p = 2.5e-07). Sensitivity and F-measures after removing singletons showed the same pattern but were lower due to less TP and more FN (Figure S9).

The execution times of the four pipelines were very different. Barque alone ran in 2 min 25 s and in 30 min when the demultiplexing with Cutadapt was added. With 16 CPUs used where possible, our assembled pipeline ran in 46 min and QIIME2 in 95 min. OBITools was the longest and ran in 1,010 min so 40 times slower than Barque (Figure 4d).

The percentage of reads assigned to the species level with 98% similarity also differed between pipelines. Barque was able to assign a species name to 98.7% of the raw demultiplexed reads (15,458,570) whereas our pipeline assigned 95.6% of the reads (14,970,256 reads), OBITools 94.4% (14,783,635), and QIIME2 91.5% (14,316,059).

3.4 | Illustration from real data

The comparison of the program performances on empirical data provided results identical to those obtained with the simulated data set. The only significant difference was found for the assignment step where Sintax obtained a significantly lower F-measure. The sensitivity and F-measure showed slight variations between programs, but these were not significant (Figures S10–S11). After removing singletons, sensitivity and F-measures showed the same variation which were lower due to less TP and more FN (Figures S12–S13). Execution time on real data confirmed that VSEARCH was the fastest program for merging, dereplicating, filtering and assigning (along with Sintax), Cutadapt was fastest for demultiplexing, and Swarm was fastest for cleaning errors (Figure S14). The performance comparison between our assembled pipeline and the other pipelines provided results concordant with the analyses on simulated data. QIIME2-based pipeline was significantly less performant than Barque, OBITools and our assembled pipeline for sensitivity (Figure 5a, p = 5.7e-06) and F-measure (Figure 5b, p = 2.2e-06), also when removing singletons (Figure 5c–d). Barque was only significantly less performant than OBITools and the assembled pipeline for F-measure, due to a slightly higher number of FP. Execution times were much shorter for Barque and the assembled pipeline (53 and 155 min, respectively, Figure 5e).

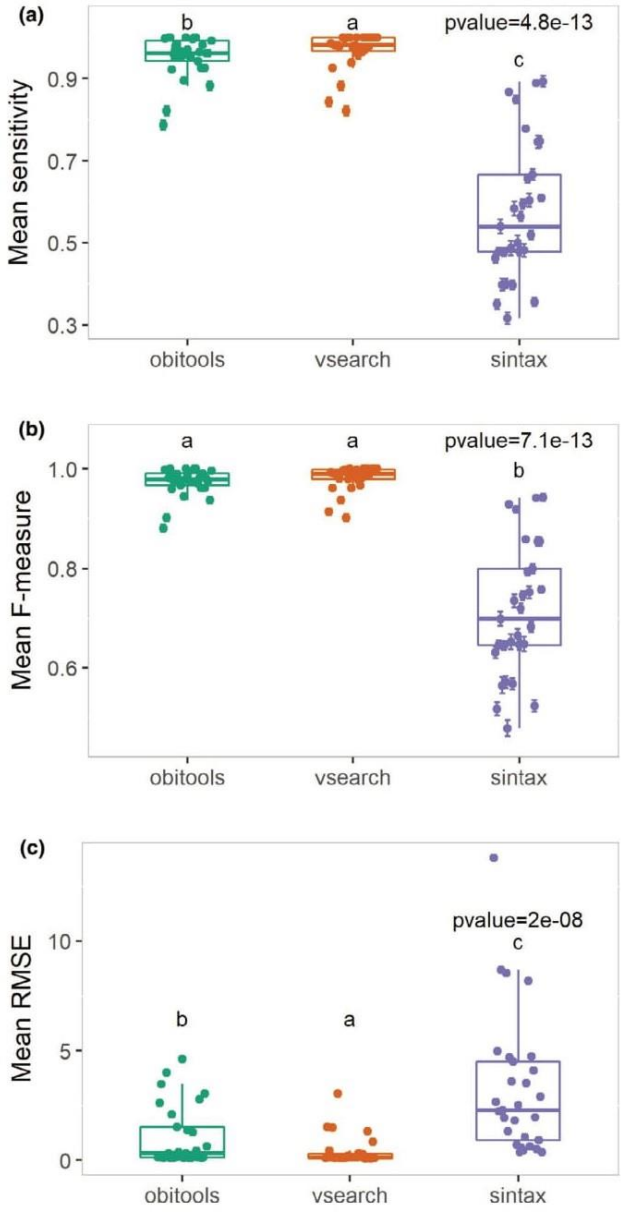


FIGURE 2 Compared performance indices of each program tested on the simulated data set, for the taxonomic assignment step. The dots represent the mean index for the 12 replicates of each sample, with the standard error; the boxplot represents the median of the performance index and the first and third quartiles for the 29 samples. (a) Sensitivity calculated on the raw outputs of each pipeline. (b) F-measure calculated on the raw outputs of each pipeline. (c) RMSE calculated between observed abundances and expected abundances for each replicate of each sample. Letters indicate significant differences

The lower sensitivity and F-measure values obtained with all programs tested on the empirical data set were due to a high number of FN (species seen by divers and not found with eDNA). However, many of these species were identified with eDNA with a similarity to the reference sequence that was lower than our threshold of 98% and were thus discarded from further analyses.

4 | DISCUSSION

4.1 | A step-by-step comparison between programs

The results of our program comparison allowed us to select the best programs for retrieving the initial community composition and abundance structure of both simulated and real fish communities. For five out of six steps, execution time was the most discriminant factor. OBITools programs obtain high sensitivities and F-measures but require much longer execution times than the other programs. Results obtained with the real data set are similar to those obtained with the simulated data. The assignment step was the only one showing significant differences between programs indices, and time was the deciding factor for all other steps.

We thus provide recommendations for programs to use at each step. For merging reads, we recommend to use VSEARCH –fastq_mergepairs, which is the fastest program. For demultiplexing we suggest Cutadapt, which has similar performance as OBITools' ngsfilter but is much faster. VSEARCH –derep_fulllength and --fastx_filter are retained for dereplication and read filtering respectively, because they obtain similar performances as other programs tested, but are faster. For error removal, we recommend using Swarm, which produces results as good as obiclean and VSEARCH --cluster_unoise but is faster. VSEARCH –usearch_global provides significantly better results than Sintax and ecotag for taxonomic assignment with a complete reference database, both in terms of sensitivity, F-measure and RMSE on relative abundances.

4.2 | Comparison of complete pipelines

The comparison of complete pipelines shows that Barque obtains sensitivity and F-measures as high as those of the assembled pipeline made of the best individual programs. These two pipelines also report the most accurate estimates of relative abundances. Barque is the fastest pipeline while our pipeline takes 1.5 times longer to analyse the same simulated data set. QIIME2-based pipeline is slightly slower than our pipeline while taxonomic and abundance results are significantly worse than with the other pipelines. OBITools requires more than 30 times the running time of Barque, the fastest pipeline. It also returns significantly worse results for abundances RMSE but provides good sensitivity and F-measures. It is noteworthy to consider that the intentions behind the design of these pipelines differ. For example, Barque aims to be exhaustive in the detection of species while minimizing the risk of not detecting a rare species of potential interest, such as an invasive species, so omits a denoising step. As a result, Barque annotates a higher proportion of the raw reads, and also produces slightly more false positives than our pipeline. In contrast, while the goal of our pipeline is also to provide species abundances that are as close to reality as possible, it controls more stringently for false positives by using a denoising step, which could lead to the removal of some very rare species and to

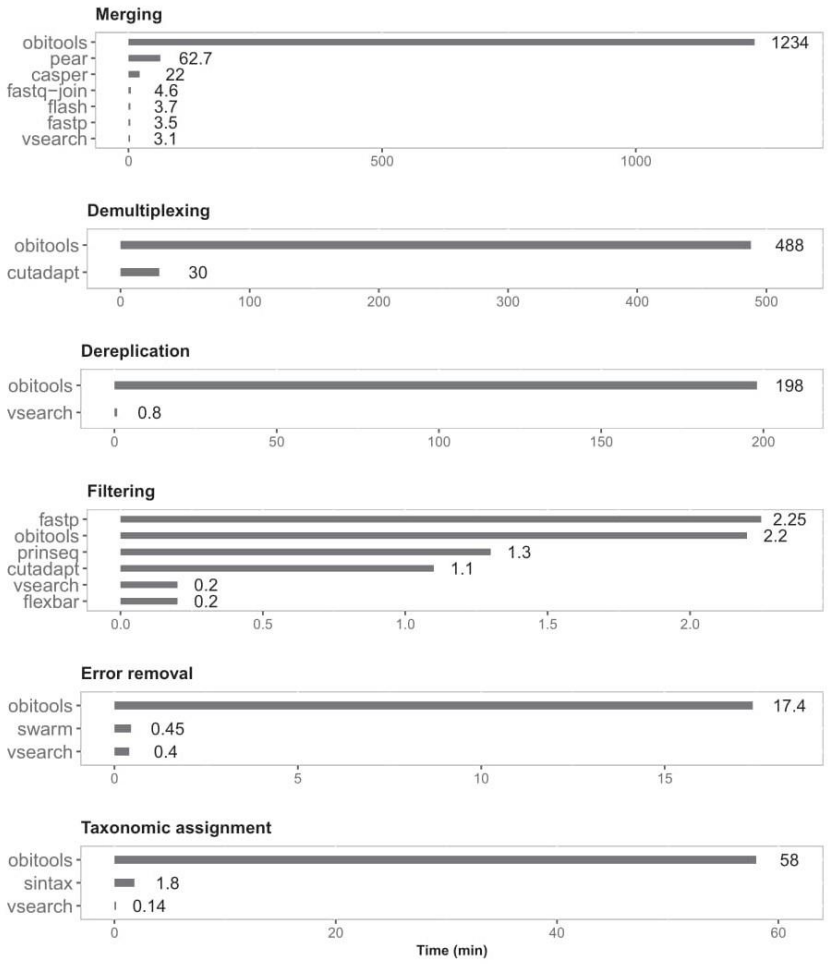


FIGURE 3 Execution time in minutes of each program tested for each step on the simulated data set. Programs compared for the assembly, demultiplexing, dereplication, filtering, error removal, and assignment steps

less annotated reads. Despite these different designs, Barque and our pipeline give almost identical results on all three indices. QIIME2 is a toolbox with many different steps where several plugins are available. The plugins chosen in the QIIME2-based pipeline were the most suitable for our data given the goal of our study, and the most comparable to the tools comprised in the other pipelines. However, many other possibilities exist and choosing other plugins and other treatment steps would result in as many different pipelines, each with a different outcome. QIIME2 offers numerous possibilities, and choosing the most appropriate tools for the purpose of a given study is important as this will influence the results and interpretation.

Analyses on the empirical data set also revealed that the new pipeline is the most performant. Barque appears slightly less performant on the real data due to an important number of what was classified as false positives in the data set. However, many of these species were observed by divers in different years. These species were thus probably present in the area during eDNA sampling even though they were not spotted by divers at the time of their campaign. As a result, it is important to be critical towards the species identified by eDNA and keep in mind that they could be real occurrences even if they were not reported using conventional observation methods. Therefore, in the context of Teleostean metabarcoding based on primer teleo, we recommend using Barque or the assembled pipeline. However, it is important to keep in mind that each pipeline considered here is a bespoke solution to the questions we aimed to address. Moreover, the differences observed in the performance of each pipeline depends on the choice of tools composing each pipeline.

4.3 | Taxonomic assignment

The three taxonomic assignment algorithms we tested differ in many aspects. OBITools's *ecotag* searches the reference database to find the reference sequence with the highest similarity to the query sequence. It then searches for all other potential reference sequences with a similarity to the first reference sequence equal or higher than the similarity to the query sequence. Ecotag then assigns the query sequence to the lowest common ancestor (LCA) of reference sequences. Ecotag provides a taxon name at the family, genus or species level as well as information about all matching reference sequences (Boyer et al., 2016). VSEARCH –usearch_global filters reference sequences that share the highest number of k-mers with the query sequence and then computes the optimal global alignment

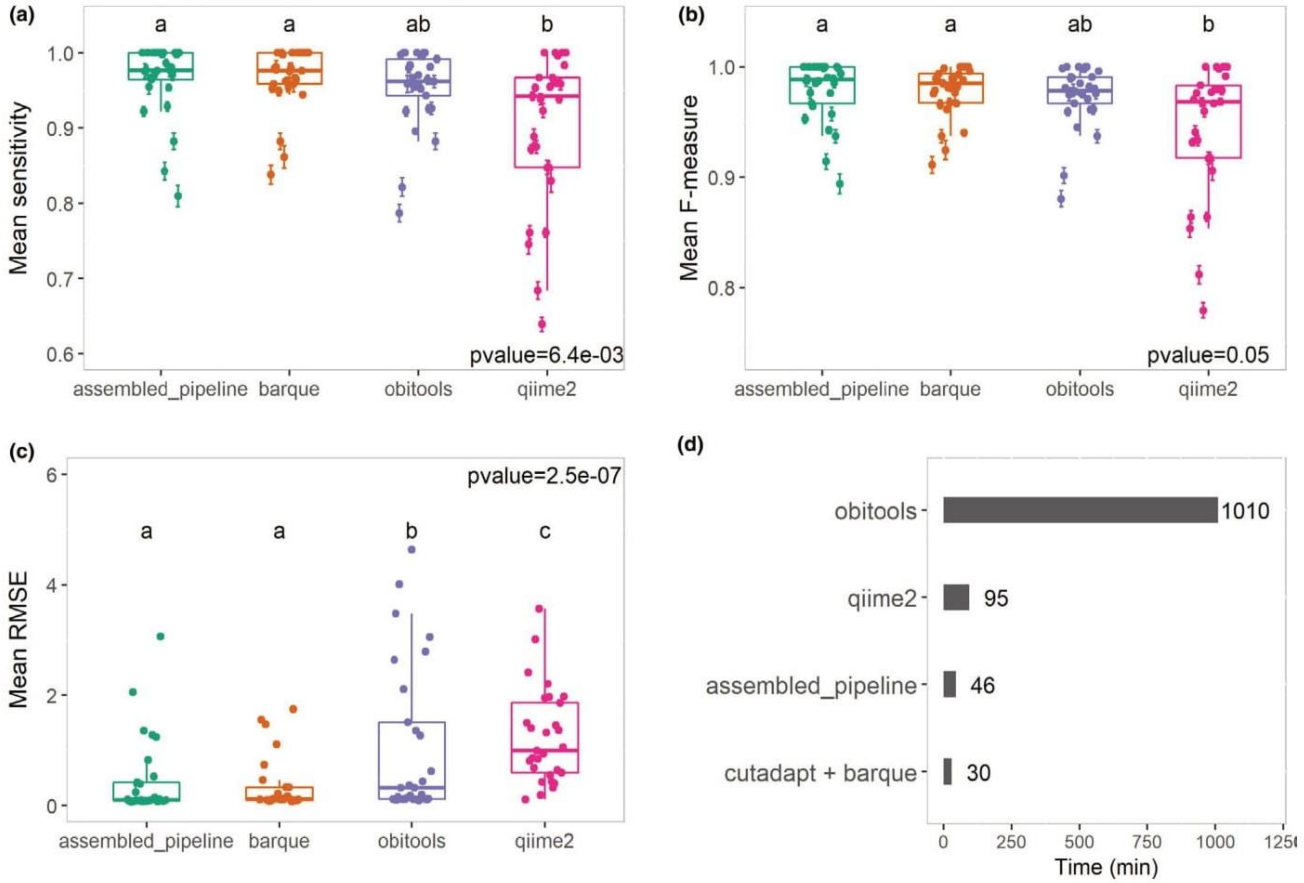


FIGURE 4 Performance indices of each pipeline on the simulated data set. The dots represent the mean index for the 12 replicates of each sample, with the standard error; the boxplots represent the median of the index and the first and third quartiles for the 29 samples. (a) Sensitivity calculated on the raw outputs of each pipeline. (b) F-measure calculated on the raw outputs of each pipeline. (c) RMSE calculated between observed abundances and expected abundances for each replica of each sample. (d) Execution time of each pipeline

between the query sequence and these reference sequences (Rognes et al., 2016). The taxonomic assignment contains the list of species matching to the query sequence with equal similarity. Sintax also proceeds with a k-mers search, among 100 iterations (Edgar, 2016). For each iteration, a subsample of k-mers contained in the query sequence is extracted. The reference sequence that has the maximum k-mers in common is retained and the taxonomy is taken from this sequence. After the 100 iterations, the species name that occurs most often is identified and its frequency is reported, along with the frequency of the genus and family identifications. If these frequencies are lower than the chosen identity threshold, then the assignment is not retained. QIIME2-based pipeline uses the plugin classify-sklearn that apply a machine learning classifier from the SciKit-learn algorithm (Pedregosa et al., 2011). The method is based on k-mer counts extraction from the reference sequences (up to 32-mers) and training of the scikit-learn multinomial naive Bayes classifier (Bokulich et al., 2018). Barque also uses VSEARCH --usearch_global and provides assignments to species, genus, and group level (which can be anything above the genus, for example the family) and uses different similarity thresholds for assignment at each of these three levels. However, assignments to taxonomic levels above the species level were not analysed here, as we focused on species name assignments.

4.4 | Sources of variation in species detection

Some samples from the simulated data set obtained much poorer sensitivity and F-measure values regardless of the program used. This is due to the presence of false negatives of several origins. First, some species of the custom reference database used for simulation and assignment have 100% identical sequences for this portion of the 12S rRNA gene and are therefore not distinguishable at the species level (e.g., *Neosalanx taihuensis* and *N. tangkahkeii*). Second, some species have a very low abundance and are therefore not found in each of the 12 replicates of the samples in the pipeline outputs. The presence of these false negatives also influences the RMSE between the expected and obtained abundances. For each of these false negatives, the observed relative abundance is 0 and the RMSE is thus higher for these replicates. The computation of the performance indices per sample, after pooling the observations in the 12 replicates, shows better results, as the full consideration

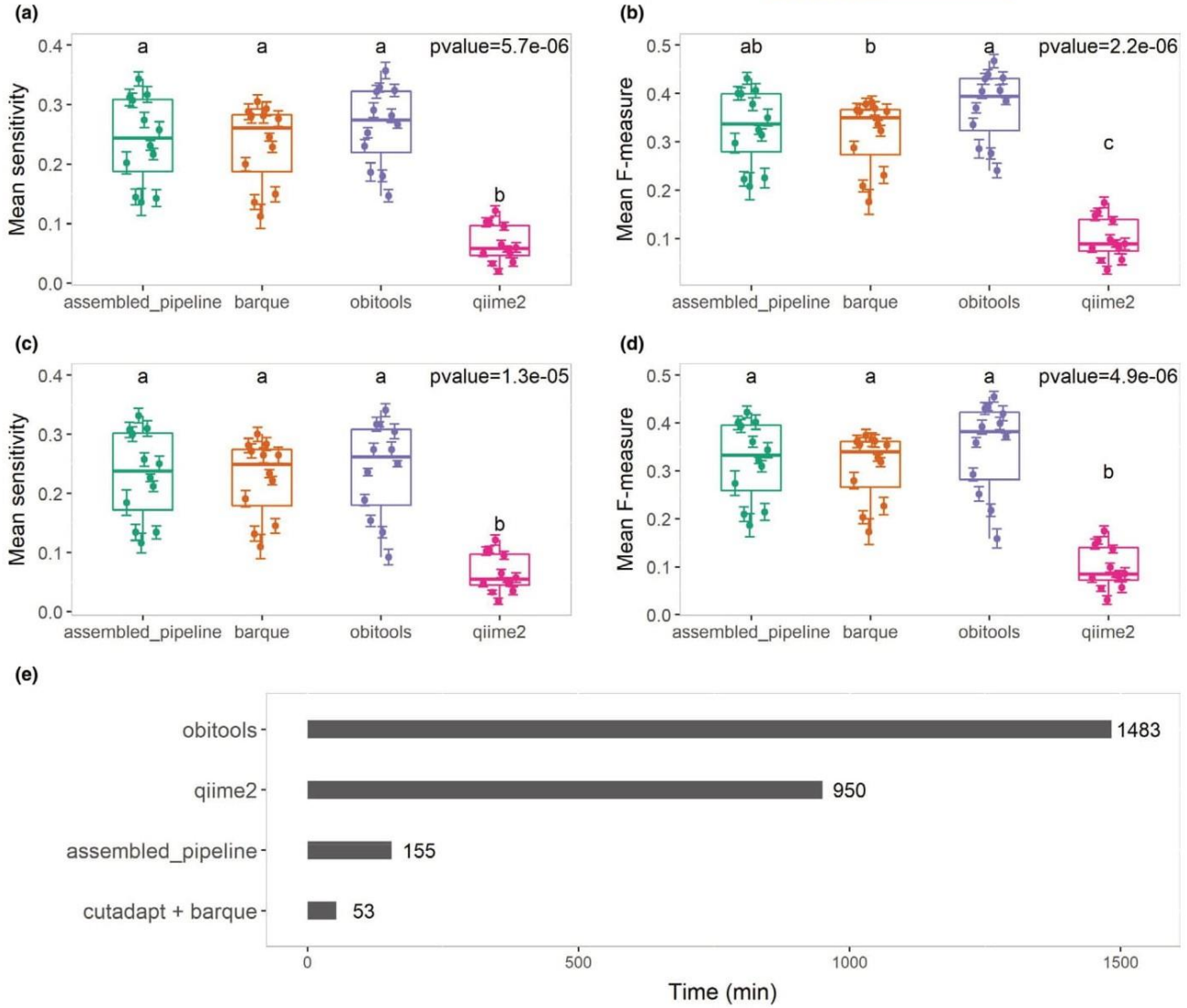


FIGURE 5 Performance index of the four pipelines on the real data set. (a) Mean sensitivity. (b) Mean F-measure. (c) Mean sensitivity, after removing singletons from the data set. (d) Mean F-measure, after removing singletons from the data set. (e) Execution time of each pipeline

of rare species decreases the number of false negatives. Only a few false positives are observed in the outputs of the different programs and pipelines; they are due to wrong taxonomic assignments either caused by sequencing errors introduced during the amplicon simulations or represent residual errors in the real data set.

Performance indices for the empirical data set are lower than the ones on the simulated data due to a high number of false negatives (some due to the too stringent similarity threshold) and some false positives. It is likely that with an eDNA sampling limited to one day, all of the species present on site were not detected and that divers did not detect elusive or hidden species (Aglieri et al., 2020; Polanco et al., 2020). In order to reduce the number of false negatives and false positives, it would be necessary to extend the eDNA sampling of each site to several seasons and to consider similarity thresholds adapted for the taxonomic assignment.

For all programs tested in this study, as well as for both data sets, we looked at the impact of the removal of singletons. This results in a slight decrease in sensitivity and F-measure (see Figures S4–S6 and S9–S10). After removing the sequences with a very weak representation, some of the false positives are removed. However, removing the singletons also removes species with real but very low abundances, thus increasing the number of false negatives, which can also lead to bad interpretations of species presence/absence. In real eDNA samples, singletons represent rare taxa of high interest, like invasive or threatened species, but also contamination and PCR or sequencing errors, such as tag or index jumps (Kwok & Higuchi, 1989; Schnell et al., 2015; Taberlet et al., 2018). In the real data from Carry-le-Rouet, the removal of singletons led to the loss of true positives, indicating that eDNA can detect rare and low-abundance species. The decision to remove singletons from a data set should then

depend on the objectives and preferences of the study, or aimed at finding a balance between removing all contaminations and errors and retaining higher chances to detect rare species.

4.5 | Perspectives

In this study we focused on a specific fish 12S mitochondrial gene region but our benchmark process could be extended to other taxa and barcodes with only slight modifications. When using different markers, depending for example on the size of the barcode and the completeness of reference databases, some parameters will have to be updated, but the general bioinformatic treatment will be similar, and the same programs can be used.

The same comparison could be extended to recent bioinformatic programs, or programs not considered in this study, such as MeFit for merging and filtering (Parikh et al., 2016), or DUDE-seq for denoising and correction of sequencing errors (Lee et al., 2017). We could also apply our comparison approach on other pipelines, such as eDNAFlow which produces ZOTUs and uses a LCA assignment method (Mousavi-Derazmahalleh et al., 2021) or CoMa (Hupfauf et al., 2020).

The similarity threshold set at 98% for assigning a sequence to a species is equivalent, on our short amplicons, to a maximum of either zero or one mismatch between the query and the reference sequences, depending on the length of the amplicon, which varies from one species to another. This can result, as in the case of the real data analysed here, in the removal of a number of sequences that would have been correctly assigned with a lower confidence and thus lead to some false negatives. Therefore, it would be relevant to consider adaptive thresholds.

In this study, we focused on taxonomic assignment at the species level. As a result, we did not explore the ability of the algorithm to provide taxonomic assignment above the species level. Nevertheless, it would be worthwhile for ecological applications to consider higher taxonomic assignment using an algorithm with such abilities, especially in the case of incomplete reference databases. PROTAX, for example, is a probabilistic method for taxonomic assignment that uses outputs of other classifiers (BLAST, RDP classifier, Wang et al., 2007) as predictors (Somervuo et al., 2016). The Anacapa Toolkit (Curd et al., 2019) includes the Anacapa Classifier module that aligns ASVs to a reference database using Bowtie2 and assigns taxonomy with a Bayesian lowest common ancestor (BLCA) method (Gao et al., 2017). These two approaches might provide relevant results for taxonomic assignments at higher levels, with probability and confidence scores.

5 | CONCLUSION

The main finding of this study is that the choice of a given program for eDNA metabarcoding analysis depends mostly on the taxonomic assignment step and the resulting diversity estimates. For all other steps, the only difference between programs standardized with the same parameters is in the execution time. This study provides some guidance for the choice of the best bioinformatics tools or the best pipeline to use for analysis of eDNA metabarcoding data. Most importantly, this study highlights the need for more efficient and accurate tools for eDNA metabarcoding taxonomic assignments, especially when only incomplete reference databases are available.

ACKNOWLEDGEMENTS

This study was part of the ALIVe project (funding PIA / ADEME / SPYGEN). The simulation and program comparison algorithms benefited from the Montpellier Bioinformatics Biodiversity platform supported by the LabEx CeMEB, an ANR "Investissements d'avenir" program (ANR-10-LABX-04-01). WT and CL received funding from the French Agence Nationale de la Recherche (ANR) through the GlobNets (ANR-16-CE02-0009) project and from 'Investissement d'Avenir' grants managed by the ANR (Trajectories: ANR-15-IDEX-02; Montane: OSUG@2020: ANR-10-LAB-56). We are grateful for the discussions and advice given by Lucie Zinger and Frédéric Boyer. We also thank Lucie Zinger for the comments and suggestions on the manuscript. Finally we are grateful to Benjamin Sibbet and Carla Lopes and anonymous referees for their constructive comments on a previous version of this manuscript.

AUTHOR CONTRIBUTIONS

L.M., A.V., C.L., S.M. and T.D. designed the experiment. L.M. and E.N. simulated the data. L.M., P.-E.G., E.N., C.N. and C.L. ran the programs. C.L., P.-E.G. and L.M. wrote the formatting scripts. L.M. analysed the data and wrote the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

DATA AVAILABILITY STATEMENT

The input data used for the simulated and real analysis are available on Dryad at https://doi.org/10.5061/dryad.15dv41nx6. The code for the simulation protocol, inputs and outputs are accessible at https://github.com/lmathon/metabarcoding_data_simulation. The codes for the benchmark study can be found at: https://github.com/lmathon/eDNA-benchmark_pipelines. The version of BARQUE used in this paper is available at: https://github.com/enormandeau/barque/releases/tag/v1.6.2.

ORCID

Laetitia Mathon https://orcid.org/0000-0001-8147-8177
Alice Valentini https://orcid.org/0000-0001-5829-5479
Pierre-Edouard Guérin https://orcid.org/0000-0001-7909-3729
Eric Normandeau https://orcid.org/0000-0003-2841-9391
Cyril Noel https://orcid.org/0000-0002-7139-4073
Emilie Boulanger https://orcid.org/0000-0002-6446-7342
Wilfried Thuiller https://orcid.org/0000-0002-5388-5274
Louis Bernatchez https://orcid.org/0000-0002-8085-9709
David Mouillot https://orcid.org/0000-0003-0402-2605
Tony Dejean https://orcid.org/0000-0002-5115-4902
Stéphanie Manel https://orcid.org/0000-0001-8902-6052

REFERENCES

Aglieri, G., Baillie, C., Mariani, S., Cattano, C., Calò, A., Turco, G., Spatafora, D., Di Franco, A., Di Lorenzo, M., Guidetti, P., & Milazzo, M. (2020). Environmental DNA effectively captures functional diversity of coastal fish communities. *Molecular Ecology*, *00*, 1–13. https://doi.org/10.1111/mec.15661

Angly, F. E., Willner, D., Rohwer, F., Hugenholtz, P., & Tyson, G. W. (2012). Grinder: A versatile amplicon and shotgun sequence simulator. *Nucleic Acids Research*, *40*(12), e94. https://doi.org/10.1093/nar/gks251

Aronesty, E. (2013). Comparison of sequencing utility programs. *The open bioinformatics journal*, *7*(1), 1–8.

Bazinet, A. L., & Cummings, M. P. (2012). A comparative evaluation of sequence classification programs. *BMC Bioinformatics*, *13*(1), 1–13. https://doi.org/10.1186/1471-2105-13-92

Berger, C. S., Hernandez, C., Laporte, M., Côté, G., Paradis, Y., Kameni T., D. W., Normandeau, E., & Bernatchez, L. (2020). Fine-scale environmental heterogeneity shapes fluvial fish communities as revealed by eDNA metabarcoding. *Environmental DNA*, *2*(4), 647–666. https://doi.org/10.1002/edn3.129

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., Huttley, G. A., & Gregory Caporaso, J. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome*, *6*(1), 1–17. https://doi.org/10.1186/s40168-018-0470-z

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C., Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Titus Brown, C., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J. … Caporaso, J. G. (2019). QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. *Nature Biotechnology*, *32*, 852–857. https://doi.org/10.7287/peerj.preprints.27295

Bonder, M. J., Abeln, S., Zaura, E., & Brandt, B. W. (2012). Comparing clustering and pre-processing in taxonomy analysis. *Bioinformatics*, *28*(22), 2891–2897. https://doi.org/10.1093/bioinformatics/bts552

Boulanger, E., Loiseau, N., Valentini, A., Arnal, V., Boissery, P., Dejean, T., Deter, J., Guellati, N., Holon, F., Juhel, J.-B., Lenfant, P., Manel, S., & Mouillot, D. (2021). Environmental DNA metabarcoding reveals and unpacks a biodiversity conservation paradox in Mediterranean marine reserves, Dryad, Dataset. *Proceedings of the Royal Society B*, *288*(20210112). 1–10. https://doi.org/10.1098/rspb.2021.0112

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. https://doi.org/10.1111/1755-0998.12428

Bylemans, J., Furlan, E. M., Gleeson, D. M., Hardy, C. M., & Duncan, R. P. (2018). Does size matter? An experimental evaluation of the relative abundance and decay rates of aquatic environmental DNA. *Environmental Science and Technology*, *52*(11), 6408–6416. https://doi.org/10.1021/acs.est.8b01071

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*(1), 193–206. https://doi.org/10.1111/jbi.13681

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Cantera, I., Cilleros, K., Valentini, A., Cerdan, A., Dejean, T., Iribar, A., Taberlet, P., Vigouroux, R., & Brosse, S. (2019). Optimizing environmental DNA sampling effort for fish inventories in tropical streams and rivers. *Scientific Reports*, *9*(1), 1–11. https://doi.org/10.1038/s41598-019-39399-5

Charbonnel, E., Monin, M., & Bachet, F. (2020). *Suivi des peuplements de poissons de la réserve marine de Carry-le-Rouet (Parc Marin de la Côte Bleue)*.

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

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), 1–19. https://doi.org/10.1371/journal.pone.0157366

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

Curd, E. E., Gold, Z., Kandlikar, G. S., Gomer, J., Ogden, M., O'Connell, T., & Meyer, R. S. (2019). Anacapa Toolkit: An environmental DNA toolkit for processing multilocus metabarcode datasets. *Methods in Ecology and Evolution*, *10*(9), 1469–1475. https://doi.org/10.1111/2041-210X.13214

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. https://doi.org/10.1111/mec.14350

Deiner, K., Yamanaka, H., & Bernatchez, L. (2020). The future of biodiversity monitoring and conservation utilizing environmental DNA. *Environmental DNA*, *3*(1), 3–7. https://doi.org/10.1002/edn3.178

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., Otis, D., Muller-Karger, F. E., Chavez, F. P., Boehm, A. B., & Breitbart, M. (2020). Environmental DNA reveals seasonal shifts and potential interactions in a marine community. *Nature Communications*, *11*(254), 1–9. https://doi.org/10.1038/s41467-019-14105-1

Dodt, M., Roehr, J., Ahmed, R., & Dieterich, C. (2012). FLEXBAR—Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. *Biology*, *1*(3), 895–905. https://doi.org/10.3390/biology1030895

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

Edgar, R. (2016). *SINTAX: A simple non-Bayesian taxonomy classifier for 16S and ITS sequences*. BioRxiv, 1–20, 10.1101/074161.

Gao, X., Lin, H., Revanna, K., & Dong, Q. (2017). A Bayesian taxonomic classification method for 16S rRNA gene sequences with improved species-level accuracy. *BMC Bioinformatics*, *18*(1), 1–10. https://doi.org/10.1186/s12859-017-1670-4

Gardner, P. P., Watson, R. J., Stott, M. B., Morales, S. E., Morgan, X. C., Finn, R. D., & Draper, J. L. (2019). Identifying accurate metagenome and amplicon software via a meta-analysis of sequence to taxonomy benchmarking studies. *PeerJ*, *7*, 1–19. https://doi.org/10.7717/peerj.6160

Hupfauf, S., Etemadi, M., Juárez, M. F. D., Gómez-Brandón, M., Insam, H., & Podmirseg, S. M. (2020). CoMA – an intuitive and user-friendly pipeline for amplicon-sequencing data analysis. *PLoS One*, *15*(12 December), 1–28. https://doi.org/10.1371/journal.pone.0243241

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.-B., Utama, R. S., Marques, V., Vimono, I. B., Sugeha, H. Y., Kadarusman Pouyaud, L., Dejean, T., Mouillot, D., & Hocdé, R. (2020). Accumulation curves of environmental DNA sequences predict coastal fish diversity in the coral triangle. *Proceedings. Biological Sciences*, *287*(1930), 1–10. https://doi.org/10.1098/rspb.2020.0248

Kwok, S., & Higuchi, R. (1989). Avoiding false positives with PCR. *Nature*, *339*(6221), 237–238. https://doi.org/10.1038/339237a0

Kwon, S., Lee, B., & Yoon, S. (2014). CASPER: context-aware scheme for paired-end reads from high-throughput amplicon sequencing Sequencing. *BMC Bioinformatics*, *15*(Suppl 9), 1–11. https://doi.org/10.1186/1471-2105-15-S9-S10

Lee, B., Moon, T., Yoon, S., & Weissman, T. (2017). DudE-Seq: Fast, flexible, and robust denoising for targeted amplicon sequencing. *PLoS One*, *12*(7), 1–25. https://doi.org/10.1371/journal.pone.0181463

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tarraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Jang, M., Pakseresht, N., Plaister, S., Radhakrishnan, R., Reddy, K., Sobhany, S., Ten Hoopen, P., Vaughan, R., … Cochrane, G. (2011). The European nucleotide archive. *Nucleic Acids Research*, *39*(Suppl. 1), 44–47. https://doi.org/10.1093/nar/gkq967

Lindgreen, S., Adair, K. L., & Gardner, P. P. (2016). An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific Reports*, *6*, 1–14. https://doi.org/10.1038/srep19233

Magoč, T., & Salzberg, S. L. (2011). FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, *27*(21), 2957–2963. https://doi.org/10.1093/bioinformatics/btr507

Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ*, *3*, 1–12. Retrieved from https://peerj.com/articles/1420

Martin, M. (1994). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10–12.

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*, *43*, 1–12. https://doi.org/10.1111/ecog.05049

McElroy, M. E., Dressler, T. L., Titcomb, G. C., Wilson, E. A., Deiner, K., Dudley, T. L., Eliason, E. J., Evans, N. T., Gaines, S. D., Lafferty, K. D., Lamberti, G. A., Li, Y., Lodge, D. M., Love, M. S., Mahon, A. R., Pfrender, M. E., Renshaw, M. A., Selkoe, K. A., & Jerde, C. L. (2020). Calibrating environmental DNA metabarcoding to conventional surveys for measuring fish species richness. *Frontiers in Ecology and Evolution*, *8*, 1–12. https://doi.org/10.3389/fevo.2020.00276

Milhau, T., Valentini, A., Poulet, N., Roset, N., Jean, P., Gaboriaud, C., & Dejean, T. (2019). Seasonal dynamics of riverine fish communities using eDNA. *Journal of Fish Biology* Accepted Author Manuscript., *98*, 387–398, https://doi.org/10.1111/1744-1633.12020

Minamoto, T., Yamanaka, H., Takahara, T., Honjo, M. N., & Kawabata, Z. (2012). Surveillance of fish species composition using environmental DNA. *Limnology*, *13*(2), 193–197. https://doi.org/10.1007/s10201-011-0362-4

Mousavi-Derazmahalleh, M., Stott, A., Lines, R., Peverley, G., Nester, G., Simpson, T., Zawierta, M. De La Pierre, M., Bunce, M., & Christophersen, C. T. (2021). eDNAFlow, an automated, reproducible and scalable workflow for analysis of environmental DNA (eDNA) sequences exploiting Nextflow and Singularity. *Molecular Ecology Resources*, (August 2020), 1–8. https://doi.org/10.1111/1755-0998.13356

Parikh, H. I., Koparde, V. N., Bradley, S. P., Buck, G. A., & Sheth, N. U. (2016). MeFiT: Merging and filtering tool for illumina paired-end reads for 16S rRNA amplicon sequencing. *BMC Bioinformatics*, *17*(1), 1–6. https://doi.org/10.1186/s12859-016-1358-1

Pauvert, C., Buée, M., Laval, V., Edel-Hermann, V., Fauchery, L., Gautier, A., Lesur, I., Vallance, J., & Vacher, C. (2019). Bioinformatics matters: The accuracy of plant and soil fungal community data is highly dependent on the metabarcoding pipeline. *Fungal Ecology*, *41*, 23–33. https://doi.org/10.1016/j.funeco.2019.03.005

Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A., Borja, A., Bouchez, A., Cordier, T., Domaizon, I., Feio, M. J., Filipe, A. F., Fornaroli, R., Graf, W., Herder, J., van der Hoorn, B., Iwan Jones, J., Sagova-Mareckova, M., Moritz, C., … Kahlert, M. (2018). The future of biotic indices in the ecogenomic era: Integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. *Science of the Total Environment*, *637–638*, 1295–1310. https://doi.org/10.1016/j.scitotenv.2018.05.002

Peabody, M. A., Van Rossum, T., Lo, R., & Brinkman, F. S. L. (2015). Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. *BMC Bioinformatics*, *16*(1), 1–19. https://doi.org/10.1186/s12859-015-0788-5

Pedregosa, F., Varoquaux, G., Thirion, B., Gramfort, A., Michel, V., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Polanco F., A., Fopp, F., Albouy, C., Brun, P., Boschman, L., & Pellissier, L. (2020). Marine fish diversity in Tropical America associated with both past and present environmental conditions. *Journal of Biogeography*, *47*(12), 2597–2610. https://doi.org/10.1111/jbi.13985

Polanco Fernández, A., Marques, V., Fopp, F., Juhel, J.-B., Borrero-Pérez, G. H., Cheutin, M.-C., Dejean, T., González Corredor, J. D., Acosta-Chaparro, A., Hocdé, R., Eme, D., Maire, E., Spescha, M., Valentini, A., Manel, S., Mouillot, D., Albouy, C., & Pellissier, L. (2020). Comparing environmental DNA metabarcoding and underwater visual census to monitor tropical reef fishes. *Environmental DNA*, *3*(1), 142–156. https://doi.org/10.1002/edn3.140

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

Pont, D., Valentini, A., Rocle, M., Delaigue, O., Jean, P., & Dejean, T. (2019). The future of fish-based ecological assessment of European rivers : from traditional EU Water Framework Directive compliant methods to eDNA metabarcoding-based approaches. *Journal of Fish Biology*. Accepted Author Manuscript., *98*, 354–366, https://doi.org/10.1111/1744-1633.12020

Prodan, A., Tremaroli, V., Brolin, H., Zwinderman, A. H., Nieuwdorp, M., & Levin, E. (2020). Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PLoS One*, *15*(1), 1–19. https://doi.org/10.1371/journal.pone.0227434

Robson, H. L. A., Noble, T. H., Saunders, R. J., Robson, S. K. A., Burrows, D. W., & Jerry, D. R. (2016). Fine-tuning for the tropics: Application of eDNA technology for invasive fish detection in tropical freshwater ecosystems. *Molecular Ecology Resources*, *16*(4), 922–932. https://doi.org/10.1111/1755-0998.12505

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, *4*, 1–22. https://doi.org/10.7717/peerj.2584

Sales, N. G., Wangensteen, O. S., Carvalho, D. C., & Mariani, S. (2019). Influence of preservation methods, sample medium and sampling time on eDNA recovery in a neotropical river. Environmental. *DNA*, *1*(2), 119–130. https://doi.org/10.1002/edn3.14

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, *27*(6), 863–864. https://doi.org/10.1093/bioinformatics/btr026

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. https://doi.org/10.1111/1755-0998.12402

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., Bremges, A., Fritz, A., Garrido-Oter, R., Jørgensen, T. S., Shapiro, N., Blood, P. D., Gurevich, A., Bai, Y., Turaev, D., … McHardy, A. C. (2017). Critical Assessment of Metagenome Interpretation - A benchmark of metagenomics software. *Nature Methods*, *14*(11), 1063–1071. https://doi.org/10.1038/nmeth.4458

Siegwald, L., Touzet, H., Lemoine, Y., Hot, D., Audebert, C., & Caboche, S. (2017). Assessment of common and emerging bioinformatics

pipelines for targeted metagenomics. *PLoS One*, *12*(1), 1–26. https://doi.org/10.1371/journal.pone.0169563

Sigsgaard, E. E., Nielsen, I. B., Carl, H., Krag, M. A., Knudsen, S. W., Xing, Y., Holm-Hansen, T. H., Møller, P. R., & Thomsen, P. F. (2017). Seawater environmental DNA reflects seasonality of a coastal fish community. *Marine Biology*, *164*(6), 1–15. https://doi.org/10.1007/s00227-017-3147-4

Somervuo, P., Koskela, S., Pennanen, J., Henrik Nilsson, R., & Ovaskainen, O. (2016). Unbiased probabilistic taxonomic classification for DNA barcoding. *Bioinformatics*, *32*(19), 2920–2927. https://doi.org/10.1093/bioinformatics/btw346

Taberlet, P., Bonin, A., Coissac, E., & Zinger, L. (2018). *Environmental DNA: For biodiversity research and monitoring*. Retrieved from https://books.google.fr/books?hl=fr&lr=&id=1e9IDwAAQBAJ&oi=fnd&pg=PP1&dq=taberlet+et+al+2018&ots=UX8Vj4tfnO&sig=saiG_Z_TcrrzgtDDsKWizkCwYwc

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), 1–9. https://doi.org/10.1371/journal.pone.0041732

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*, *1*(2), 99–108. https://doi.org/10.1002/edn3.21

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

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, *73*(16), 5261–5267. https://doi.org/10.1128/AEM.00062-07

Yan, H. F., Kyne, P. M., Jabado, R. W., Leeney, R. H., Davidson, N. K., Derrick, D. H., Dulvy, N. K. (2021). Overfishing and habitat loss drives range contraction of iconic marine fishes to near extinction. *Science Advances*, in press., *7*, 1–10.

Zinger, L., Bonin, A., Alsos, I. G., Bálint, M., Bik, H., Boyer, F., Chariton, A. A., Creer, S., Coissac, E., Deagle, B. E., De Barba, M., Dickie, I. A., Dumbrell, A. J., Ficetola, G. F., Fierer, N., Fumagalli, L., Gilbert, M. T. P., Jarman, S., Jumpponen, A., … Taberlet, P. (2019). DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. *Molecular Ecology*, *28*(8), 1857–1862. https://doi.org/10.1111/mec.15060

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, *30*(5), 614–620. https://doi.org/10.1093/bioinformatics/btt593

SUPPORTING INFORMATION

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

How to cite this article: Mathon, L., Valentini, A., Guérin, P.-E., Normandeau, E., Noel, C., Lionnet, C., Boulanger, E., Thuillier, W., Bernatchez, L., Mouillot, D., Dejean, T., & Manel, S. Benchmarking bioinformatic tools for fast and accurate eDNA metabarcoding species identification. *Mol Ecol Resour.* 2021;00:1–15. https://doi.org/10.1111/1755-0998.13430

Chapitre 3 – L'ADN environnemental pour étudier les patrons de diversité des poissons de récifs coralliens

Poisson-clown à collier (Amphiprion perideraion), caché dans son anémone (Heteractis magnifica), à l'approche du plongeur. Passe de Boulari, Nouvelle-Calédonie

1. Préface

L'ampleur et la vitesse des changements environnementaux et des perturbations anthropiques menacent la biodiversité mondiale et en particulier celle des poissons récifaux coralliens. Une meilleure compréhension des patrons de biodiversité et de ses processus à grande échelle sur les récifs coralliens est essentielle pour anticiper et prévenir le déclin de la biodiversité des poissons. Cette compréhension nécessite un suivi des communautés plus rapide et plus exhaustif que les suivis actuels, et employant des méthodes standardisées et applicables à large échelle. Le prélèvement d'ADNe ne nécessitant pas de connaissances taxonomiques (excepté pour la constitution de la base de référence) ni de compétences spécifiques et pouvant être réalisé sans mise en place d'une logistique complexe, pourrait permettre un suivi de la biodiversité à large échelle spatiale et temporelle.

Les récifs coralliens sont les écosystèmes marins les plus riches, abritant entre 2400 et 8000 espèces de poissons et sont principalement étudiés par des recensements visuels en plongée. La majorité des études menées à une échelle locale ou régionale et comparant l'ADNe avec des méthodes conventionnelles ont démontré l'efficacité de cette méthode (Marques et al. 2021; Polanco et al. 2021), mais très peu d'études ont été menées à large échelle. La capacité du *metabarcoding* de l'ADNe à décrire les patrons de distribution des poissons coralliens à large échelle reste donc à démontrer. **À** une telle échelle spatiale, les bases de référence sont très incomplètes (Marques, Milhau, et al. 2020) et le pipeline assemblé au chapitre précédent n'est donc pas applicable. Cependant, la méthodologie en unités taxonomiques moléculaires développée par Marques, Guérin, et al. (2020) permet de s'affranchir de ce biais et de produire des estimations de diversité à de larges échelles géographiques.

Ce chapitre vise à i) étudier la capacité du *metabarcoding* de l'ADNe à décrire la diversité des poissons sur les récifs coralliens à large échelle, ii) comparer cette diversité avec celle obtenue par la plus grande base de données de recensements visuels en plongée (Reef Life Survey) et iii) détecter de nouveaux patrons d'assemblages de communautés, par l'étude de la rareté et de la partition de la diversité entre les échelles spatiales.

L'analyse de 226 échantillons d'ADNe collectés dans 100 stations dans cinq régions tropicales (Caraïbes, Pacifique central et sud-ouest, Triangle de corail et océan Indien occidental) a permis de détecter 16% de taxons et 25% de familles en plus qu'avec 2047 recensements visuels.

L'ADNe détecte davantage d'espèces cryptobenthiques et pélagiques, peu observables en plongée et confirme les grands patrons connus de la répartition biogéographique des poissons tropicaux : gradient longitudinal de diversité avec un pic dans le Triangle de Corail, barrière géographique différenciant la faune des Caraïbes. L'étude de la partition de la diversité à différentes échelles spatiales montre que l'ADNe détecte une plus grande β-diversité entre stations, ce qui indique une potentielle capacité de l'ADNe à détecter des changements de communautés à fine échelle. Ces résultats montrent le fort potentiel de l'ADNe pour étudier et suivre les communautés de poissons à large échelle.

2. Manuscrit B

Publié dans *Proceedings of the Royal Society B*:

Mathon, L., Marques, V., Mouillot, D., Albouy, C., Andrello, M., Baletaud, F., ... & Manel, S. (2022). Cross-ocean patterns and processes in fish biodiversity on coral reefs through the lens of eDNA metabarcoding. *Proceedings of the Royal Society B, 289*(1973), 20220162.

PROCEEDINGS B

royalsocietypublishing.org/journal/rspb

Research

Cite this article: Mathon L *et al*. 2022 Cross-ocean patterns and processes in fish biodiversity on coral reefs through the lens of eDNA metabarcoding. *Proc. R. Soc. B* **289**: 20220162.
https://doi.org/10.1098/rspb.2022.0162

Received: 27 January 2022
Accepted: 24 March 2022

Subject Category:
Ecology

Subject Areas:
biotechnology, ecology

Keywords:
eDNA metabarcoding, coral reef fish, biogeographic patterns, visual census

Author for correspondence:
Laetitia Mathon
e-mail: laetitia.mathon@gmail.com

†These authors contributed equally as first author to this work.
‡These authors contributed equally as senior author to this work.

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.5933156.

THE ROYAL SOCIETY PUBLISHING

Cross-ocean patterns and processes in fish biodiversity on coral reefs through the lens of eDNA metabarcoding

Laetitia Mathon[1,2,†], Virginie Marques[1,3,†], David Mouillot[3,4], Camille Albouy[5], Marco Andrello[3,17], 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], 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,‡] and Stéphanie Manel[1,‡]

[1]CEFE, Univ. Montpellier, CNRS, EPHE-PSL University, IRD, Montpellier, France
[2]ENTROPIE, Institut de Recherche pour le Développement (IRD), Univ. Réunion, UNC, CNRS, Q1 IFREMER, Nouméa, New Caledonia, France
[3]MARBEC, Univ Montpellier, CNRS, IFREMER, IRD, Montpellier, France
[4]Institut Universitaire de France, France
[5]DECOD (Ecosystem Dynamics and Sustainability), IFREMER, INRAE, Institut Agro - Agrocampus Ouest, Nantes, France
[6]SOPRONER, groupe GINGER, 98000 Noumea, New Caledonia, France
[7]Programa de Biodiversidad y Ecosistemas Marinos, Museo de Historia Natural Marina de Colombia (MHNMC), Instituto de Investigaciones Marinas y Costeras- INVEMAR, 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, National Research and Innovation Agency, Jl. Pasir Putih 1, Ancol Timur, Jakarta Utara 14430, 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
[17]Institute for the study of Anthropic Impacts and Sustainability in the marine environment, National Research Council (CNR-IAS), Rome, Italy

LM, 0000-0001-8147-8177; DM, 0000-0003-0402-2605; MA, 0000-0001-7590-2736; AV, 0000-0001-5829-5479; LP, 0000-0002-2289-8259; SM, 0000-0001-8902-6052

Increasing speed and magnitude of global change threaten the world's biodiversity and particularly coral reef fishes. A better understanding of large-scale patterns and processes on coral reefs is essential to prevent fish biodiversity decline but it requires new monitoring approaches. Here, we use environmental DNA metabarcoding to reconstruct well-known patterns of fish biodiversity on coral reefs and uncover hidden patterns on these highly diverse and threatened ecosystems. We analysed 226 environmental DNA (eDNA) seawater samples from 100 stations in five tropical regions (Caribbean, Central and Southwest Pacific, Coral Triangle and Western Indian Ocean) and compared those to 2047 underwater visual censuses from the Reef Life Survey in 1224 stations. Environmental DNA reveals a higher (16%) fish biodiversity, with 2650 taxa, and 25% more families than underwater visual surveys. By identifying more pelagic, reef-associated and crypto-benthic species, eDNA offers a fresh view on assembly rules across spatial scales. Nevertheless, the reef life survey identified more species than eDNA in 47 shared families, which can be due to incomplete

sequence assignment, possibly combined with incomplete detection in the environment, for some species. Combining eDNA metabarcoding and extensive visual census offers novel insights on the spatial organization of the richest marine ecosystems.

1. Introduction

Coral reefs host the highest fish diversity on earth despite covering less than 0.1% of the ocean's surface [1,2]. They are also severely threatened [3], with near-future outlooks predominantly pessimistic [4]. Data syntheses over decades of surveys estimate the total number of coral reef fishes to be 2400 to 8000 species [5,6], distributed among approximately 100 families [7]. Typically, coral reef biodiversity displays clear spatial patterns, including longitudinal and latitudinal gradients outwards the Indo-Australian Archipelago [8,9], also known as the 'Coral Triangle', hosting the world's highest level of marine biodiversity [10]. The exceptional biodiversity in the Coral Triangle has recently been suggested to strongly relate to higher diversity among fish families that feed on plankton [11]. Other trophic groups are also very important on coral reefs but are often undetected because they are transient or hidden [12,13]. Intriguingly, the proportions of fish species among families are shown to be strongly conserved across the Indo-Pacific [8]. The spatial patterns of coral reef fishes are also marked by strong variations in taxonomic composition (species turnover or β diversity), often due to isolation [14]. Many species on coral reefs are geographically localized, but can sometimes be locally abundant, while others are widespread [15].

Coral reef fishes have evolved in a physically complex environment and present a wide range of forms and functions [16]. Small cryptic species, hereafter called crypto-benthic, that live inside the reef structure, can be very difficult to sample or survey using non-destructive methods [17], yet represent half of the fish diversity on coral reefs [13]. Even though fishes are among the best-studied taxa inhabiting coral reefs [18], our knowledge of their biodiversity is only partial [19], the taxonomy is complex, uncertain for many species [5], and countless species remain undescribed.

Environmental DNA (eDNA) metabarcoding, a method retrieving and analysing DNA naturally released by organisms in their environment [20], provides an opportunity to not only better understand classical biodiversity patterns, but also uncover novel ones hidden by our incomplete taxonomic and biogeographic coverage [21]. Environmental DNA is particularly powerful in aquatic ecosystems [22] and is now well established for marine microorganisms [23,24]. By contrast, its potential to provide an integrated biodiversity assessment of macroorganisms, including vertebrates of all trophic levels (from crypto-benthic to large pelagic fish species), is only shown at local [25] and regional [26–30] scales, but not yet at spatial scales including more than one biogeographic region or multiple ocean basins.

Here, we investigate how a cross-ocean basin snapshot of eDNA sampling could describe the distribution of fish biodiversity on coral reefs, reveal unknown patterns and challenge well-established assembly rules. From 226 eDNA seawater samples (2712 PCR replicates) collected in 100 stations at 26 sites covering five tropical regions (Southeast Polynesia, Tropical Northwestern Atlantic, Tropical Southwestern Pacific, Western Indian Ocean and Western Coral Triangle) across the Indian, Pacific and Atlantic Oceans (electronic supplementary material, figures S1 and S2), we produced a final dataset of 189 350 273 mitochondrial 12S rRNA gene sequence reads (see Methods), clustered into 2023 molecular operational taxonomic units (MOTUs) and assigned to Actinopterygii (bony fishes) and Chondrichthyes (cartilaginous fishes) taxa (electronic supplementary material, tables S1 and S2). We then compared fish biodiversity patterns obtained from eDNA to those observed from 2047 standardized visual surveys of reef fishes in 1224 stations at 219 sites within 24 tropical regions [31].

2. Results

(a) Global estimates of fish biodiversity on coral reefs

We estimated total fish diversity on coral reefs using the asymptote of a multi-model accumulation curve for both eDNA MOTUs [32] and visual census species (see Methods). The asymptote estimated from 100 eDNA stations distributed in five regions sampled over a 28-month period reaches 2650 MOTUs (figure 1*a*). This detectable fish MOTU diversity, including also MOTUs unassigned at the species level, is 16% higher than the estimate from visual census data, which reaches an asymptote at 2268 fish species from 2047 tropical transects surveyed during 13 years (figure 1*b*). The asymptotic estimation of family richness obtained with eDNA reaches 147 families, 25% more than the asymptotic number of families estimated with visual census data (118 families, figure 1*c*,*d*). Among the 71 families shared between both datasets, 24 have a higher number of MOTUs from eDNA survey than species from the visual survey while 47 have more species from visual survey than MOTUs from eDNA survey (figure 1*e*). Families with more taxa identified using eDNA include those often associated with reef-adjacent habitats such as mangroves or soft sediments like Mugilidae (e.g. *Mugil rubrioculus*), Elopidae and Gerreidae [33] (e.g. *Gerres oyena*), and crypto-benthic species that live hidden in crevices (e.g. Gobiidae) or nocturnal fish species [34] (e.g. Congridae). Families with more taxa with visual census include Acanthuridae, Chaetodontidae, Blenniidae, Labridae, Pomacentridae and Scaridae. Fifty-five families are detected only with eDNA, including Myctophidae, Engraulidae, Atherinidae and Exocoetidae, while 24 families are detected only by the visual census, including Caesionidae, Chaenopsidae, Labrisomidae and Microdesmidae. Environmental DNA estimates a diversity of crypto-benthic species 13% higher than with visual census, and, among many others, includes species such as the elegant firefish (*Nemateleotris decora*), which lives on the outer reef slope between 25 and 70 m (figure 2*a*). Yet, the difference in fish diversity assessment between the two methods is the strongest for pelagic and wide-ranging species, for which eDNA reveals more than seven times higher richness than with visual census. These species mainly belong to Scombridae (e.g. *Katsuwonus pelamis*), Clupeidae, Carcharhinidae (e.g. *Carcharhinus leucas*, *Sphyrna lewini*) and Belonidae (figure 2*b*).

MOTU richness per fish family retrieved with eDNA is strongly correlated with fish species richness within families recorded in visual census data (Pearson correlation = 0.84, $p < 0.001$, $n = 71$; figure 1*e*). Highly diverse families seen on coral reefs are also well represented in eDNA samples, with

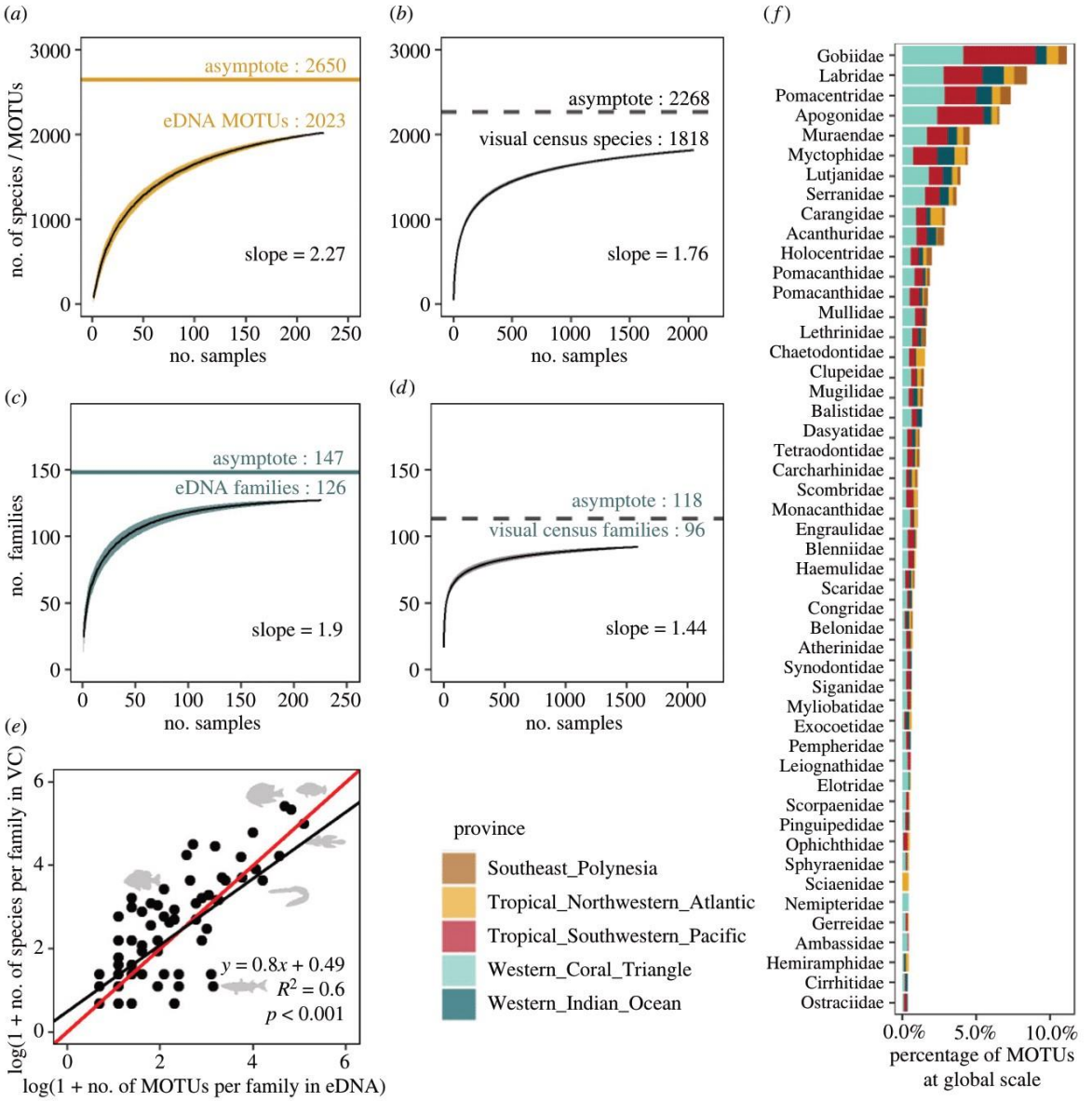


Figure 1. Estimates of overall fish richness from environmental DNA (eDNA) and visual census. (*a*) Accumulation curve of MOTUs from eDNA (eDNA MOTUs), (*b*) accumulation curve of species from the visual census database, (*c*) accumulation curve of eDNA families and (*d*) accumulation curve of visual 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. (no. = number of) (Online version in colour.)

Gobiidae, Labridae and Pomacentridae containing more than 100 MOTUs each, together representing about 20% of MOTUs (figure 1*f*; electronic supplementary material, figures S3 and S4). The slope of the log–log relationship between MOTUs richness per family and species richness per family is equal to 0.8 showing that the relationship is not proportional but saturating. The richest fish families contain more MOTUs detected with eDNA than species detected with visual surveys.

(b) Biogeography of eDNA sequences

The spatial distribution of MOTUs follows clear biogeographic patterns, with a peak in the Coral Triangle and lower values of MOTU richness toward Southeast Polynesia (electronic supplementary material, figure S5). The richest region (West Papua, Indonesia, Western Coral Triangle) contains approximately 50% of the global pool of fish MOTUs while the poorest region (Fakarava, French Polynesia, Southeast Polynesia) contains only 9% of the global pool (electronic supplementary material, figures S6 and S7, and table S2). Distance-based redundancy analysis (dbRDA) was performed on fish family proportions at each site (i.e. number of MOTUs or species assigned to each family in each site, see Methods) for eDNA and visual surveys with the region and the site MOTU/species richness as explanatory variables, including their interaction (figure 3; electronic supplementary material, table S3). For eDNA, the dbRDA explains up to 42% of variation in family proportions between pairs of sites with region and MOTU/species richness both having significant effects ($F = 4.1$ and 5.7, respectively, $p < 0.001$), but no significant interaction ($F = 1.99$, $p > 0.05$). The partial dbRDA on eDNA showed a significant effect of region while controlling for MOTU

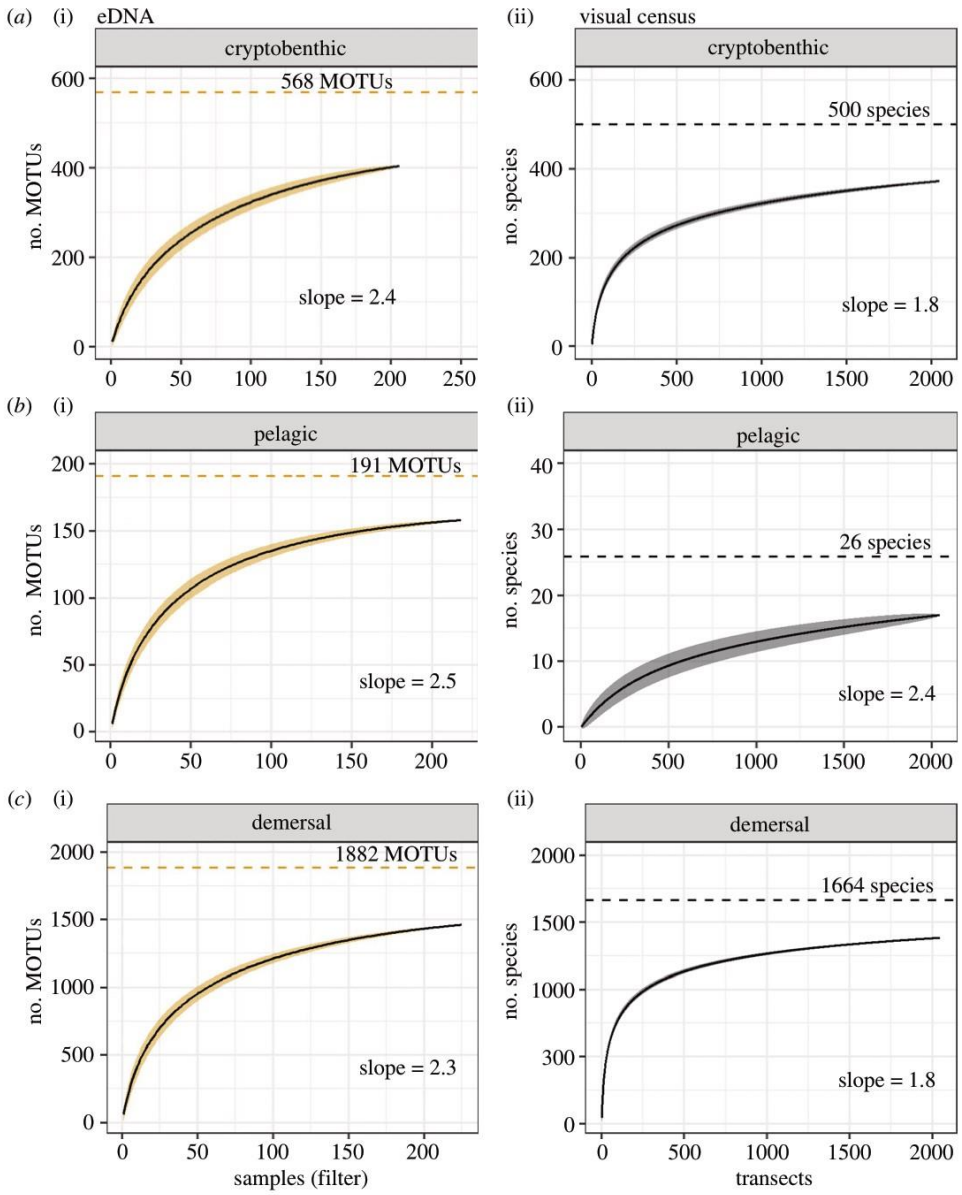


Figure 2. Estimates of overall fish richness from eDNA and visual census across habitat categories. (*a*) Accumulation curve of crypto-benthic eDNA MOTUs (i) and visual census species (ii). (*b*) Accumulation curve of pelagic MOTUs (i) and visual census species (ii). (*c*) Accumulation curve of demersal MOTUs (i) and visual census species (ii). Accumulation model is fitted with a nonlinear Lomolino model (see Methods). (Online version in colour.)

richness ($F = 2.79$, $p < 0.001$). The first axis explains 17.2% of variation in family proportions and separates the Western Coral Triangle from other regions (figure 3*a*,*b*). The first axis shows a higher proportion of Lutjanidae but lower proportions of Labridae and Gobiidae in sites of the Western Coral Triangle. It also confirms the longitudinal diversity gradient from the Coral Triangle. The second axis explains 11.2% of variation and discriminates the Tropical Northwestern Atlantic from the Western Indian Ocean, due to a higher proportion of Clupeidae and Carangidae in the Atlantic Ocean and a higher proportion of Acanthuridae in the Indian Ocean. The dbRDA performed on visual census data explained greater variation ($R^2 = 0.5$, $p < 0.001$) and the region also had a significant, albeit weaker than for MOTUs, effect on fish family proportions ($F = 17.7$, $p < 0.01$), while species richness and interaction between the two variables also had significant effects ($F = 6.28$ and 2, $p < 0.01$, respectively). The first axis explains 41.6% of variance in family proportions and separates the Tropical Northwestern Atlantic from the other regions with a higher proportion of Gobiidae and Serranidae. The second axis explains 5.7% of variance in family proportions and separates the Southeast Polynesia from Indo-Pacific regions, and is mostly driven by the higher proportion of Pomacentridae in the Indo-Pacific (figure 3*c*,*d*).

(c) Global patterns of fish turnover and rarity

Our eDNA survey shows that a majority of MOTUs are geographically restricted, with 85% of the MOTUs detected in only one region (figure 4*a*), and 35% in only one site (electronic supplementary material, figure S8). Geographic restriction is one aspect of species rarity but is shown to play a primary role in determining extinction risk while local abundance and habitat specialization have secondary

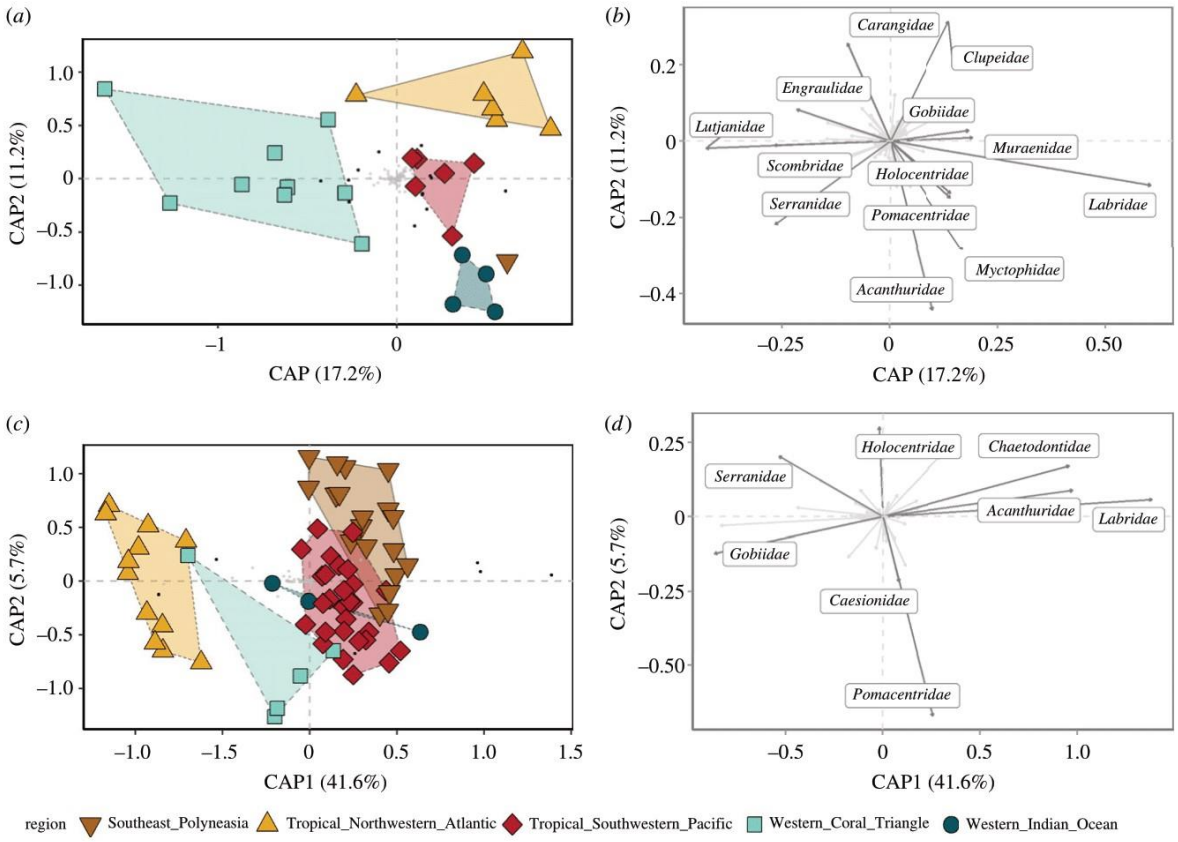


Figure 3. Partial dbRDA of MOTU proportions of each family in each site. (*a*) dbRDA on eDNA dataset, with 133 families in 26 sites ($R^2 = 0.21$, $F = 3.11$, $p = 0.001$). (*b*) Families with scores greater than 95% of scores distribution on each axis for eDNA. (*c*) dbRDA on a subset of visual census dataset to select only the sites in the same regions as in the eDNA dataset, with 76 families in 68 sites ($R^2 = 0.5$, $F = 15.8$, $p = 0.001$). (*d*) Families with scores greater than 95% of scores distribution on each axis for visual census. Axis labels indicate the percentage of variance explained by the 2 first dbRDA dimensions (CAP1 and CAP2). (Online version in colour.)

roles [35]. We hierarchically partitioned the global MOTU diversity (γ_{global}) into additive diversity components (i.e. dissimilarity) due to difference between regions ($\beta_{inter\text{-}region}$), mean difference between sites within regions ($\overline{\beta}_{inter}$), mean difference between stations within sites ($\overline{\beta}_{inter\text{-}station}$) and mean station diversity ($\overline{\alpha}_{station}$) [36]. As a consequence of the geographic restriction of most MOTUs to one region, the total fish MOTU (γ) diversity is mainly due to inter-region β-diversity (approx. 74%) followed by inter-site (14.8%) and inter-station (5.9%) β-diversity (figure 4*b*). The same partitioning using different site delineations (10 and 20 km) provides similar results (electronic supplementary material, table S4). Diversity partitioning of crypto-benthic fish MOTUs only or pelagic fish MOTUs only reveals similar patterns (electronic supplementary material, table S5). The partitioning diversity of species detected by visual census also revealed similar patterns but with a stronger effect of $\beta_{inter\text{-}region}$ (84%) and lower (3x) $\overline{\beta}_{inter\text{-}site}$ and $\overline{\beta}_{inter\text{-}station}$ (electronic supplementary material, table S5 and figure S9).

Beyond the hierarchical partitioning of diversity, we compared the distribution of fish MOTUs and species visual occurrences independently of the survey method and sampling effort using global species abundance distributions (gSAD) [37]. We fitted the fish MOTU and species visual occurrences to three distributions (log-series, Pareto and Pareto with exponential finite adjustment, i.e. Pareto Bended; see Methods) and estimated the parameters by maximum likelihood. For the visual census gSAD, the best fit was obtained with the log-series and Pareto distributions (electronic supplementary material, table S6) with a slope of −0.95 (confidence interval at 95% [−0.98; −0.92]) (electronic supplementary material, figure S10). This suggests a distribution of geographically restricted or rare species close to the neutral theory (β close to −1). By contrast, the best fit for fish MOTUs was obtained with the Pareto Bended distribution with a slope $\beta = -0.76$ (confidence interval at 95% [−0.85; −0.65]) and then with the log-series distribution, suggesting a lower prevalence of rarity than under the neutral theory, in agreement with previous tests based on species distributions on coral reefs [38].

3. Discussion

Environmental DNA allows the detection and identification of more taxa than traditional techniques [26,39], but further offers novel insights on the spatial organization of the richest marine ecosystem at a large scale. Over a timespan of 2.3 years, in major tropical ocean basins, eDNA metabarcoding reveals a higher proportion of crypto-benthic, pelagic and soft-sediment-associated fishes on coral reefs than detected in the most extensive visual census over 13 years. We found a high local MOTU turnover, but we were not able to conclude if it is due to an insufficient sampling at the station level, or if it suggests that differences in fish species composition may exist between adjacent reefs that are not detected

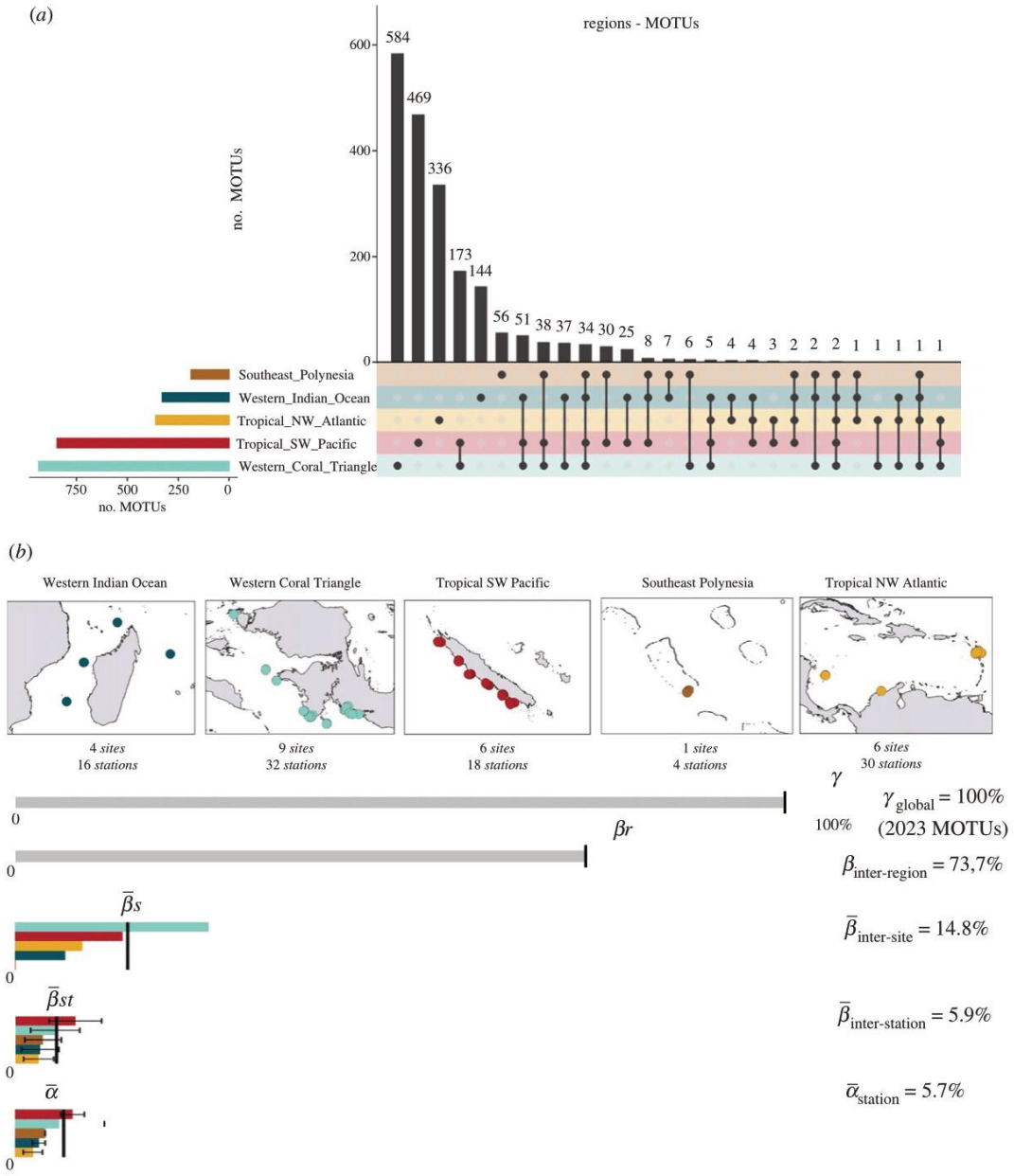


Figure 4. Hierarchical partitioning of MOTU occurrences across spatial scales. (*a*) Number of MOTUs found in only one region, or shared between 2, 3, 4 or all 5 regions. Histograms indicate the number of MOTUs present in all the regions identified by the dots in the lower part. (*b*) Global fish diversity (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 (coloured bars) with the standard errors. $\beta_{inter\text{-}region}$ contributes the highest to gamma global (73.7%). (Online version in colour.)

by visual surveys [26], so that fish biodiversity is more patchy than previously thought on coral reefs.

We were also able to retrieve well-known patterns of fish diversity on coral reefs such as the biogeographic boundaries between the Atlantic and Pacific oceans, the longitudinal diversity gradient from the centre of the Coral Triangle, with Southeast Polynesia being the least diverse region and Western Coral Triangle the richest, and that Gobiidae, Labridae, Pomacentridae and Apogonidae are the most diverse fish families on coral reefs [8]. We found a lower proportion of rare MOTUs than expected under the neutral theory with eDNA, which is in agreement with the findings of a previous study from coral reefs in the Indo-Pacific [38], while visual census data suggest higher rarity close to that predicted from the neutral theory. More surprising, our study calls into question the pattern of fish family stability composition across the Indo-Pacific that was revealed more than 20 years ago [8], and the recent finding that planktivore families drive fish biodiversity patterns on coral reefs [11]. We found significant effects of species richness and region on family composition, which appears less stable than previously thought.

Environmental DNA identified many pelagic, deep-water and crypto-benthic species not seen by divers. Among the pelagic species identified with eDNA, many belong to the Scombridae and Carcharhinidae families, which likely avoid

divers or are not permanent residents on coral reefs so can be missed in visual surveys [40]. Some crypto-benthic or reef-associated species, hidden in the reef, can also be missed by divers so were also more represented in eDNA than in visual surveys. Crypto-benthic species also have a crucial role for coral reef functioning, by promoting biomass production and fuelling the reef trophodynamics [41], but their diversity has been underestimated so far [13]. Transient, pelagic and deep-water species may be very important for reef functioning, through pelagic larval stages or nocturnal migration up the reef slope [12,42,43], but their presence and role need further investigation. By contrast, visual census also detected many families not detected, or not identified, by eDNA, such as Acanthuridae, Blenniidae, Caesionidae, Chaenopsidae, Chaetodontidae, Labrisomidae, Labridae or Microdesmidae. This limited identification by eDNA can be due to the very low representation of these families in 12S reference databases (between 0 and 12%), or to the low resolution of the teleo marker for species of these families, so several species can share the same sequence and be grouped under the same MOTU. Environmental DNA may also be inappropriate to detect these species in the environment.

The finding of a strong regional effect on both species composition (figure 3) and species differentiation (figure 4) at a large scale is in agreement with visual surveys and previous knowledge [44], while the suggestion of a strong turnover at the local scale may be an unexpected result for coral reef fishes. This predominant role of large-scale bioregional differentiation explains the exceptional fish diversity on coral reefs, probably associated with long-term geological isolation [2]. Overall, the Tropical Northwestern Atlantic region has a very distinct MOTU composition compared to the four other regions (figure 3) with only 1.2% of MOTUs being shared between the Tropical Northwestern Atlantic and any other region, while 20% of MOTUs are shared between at least two Indo-Pacific regions (figure 4*a*). The isolation of the Tropical Northwestern Atlantic region can be explained by the hard vicariant barrier of the Isthmus of Panama [14,45], and a limited suitable area for coral reefs during the past quaternary glaciation. By contrast, the Indo-Pacific maintained extensive coral reef refuges that have served as centres of survival during ice-age periods [9].

The greater local compositional dissimilarity of reef fishes among adjacent stations with eDNA than with visual census may correspond to local environmental or habitat differences, to stochastic or random processes [46], or may be due to an insufficient sampling at the station level (electronic supplementary material, analyses, figure S6). A higher number of replicates per station would be necessary to characterize exhaustively the diversity at the station level and more confidently conclude on the local turnover hypothesis.

While our results confirm the potential of eDNA to monitor biodiversity in marine ecosystems, some limitations should be addressed in the future to fully exploit this potential. Completing public reference databases would improve the accuracy of taxonomic assignment, which is essential for a better estimation of biodiversity patterns. At such a large spatial scale, reference databases are far from exhaustive with only up to 13% of fish species sequenced on our marker [47], preventing assignment to the species level for 81% of our eDNA sequences. Using multiple markers is an alternative to the database limitation [48,49], but it is much more expensive. For these reasons, we used MOTUs curated by a combination of a clustering algorithm and conservative abundance-based post-clustering filters. While uncurated MOTUs are prone to overestimate real diversity [50] and a given MOTU can represent several species within one cluster or several MOTUs belonging to one species, MOTUs with conservative curation have been shown to reflect the true level of fish diversity across scales in streams [51,52]. Additionally, some species share the same barcode sequence due to insufficient genetic differentiation on such a small mitochondrial marker [49]. This lack of taxonomic resolution combined with a conservative curated MOTUs pipeline can underestimate MOTUs richness. Moreover, some crypto-benthic or rare fish families are still underrepresented in public databases, and their diversity is potentially underestimated with eDNA (i.e. Blenniidae, Gobiescocidae, Chaenopsidae and Aploactinidae).

Differences in sampling method and in sample size might influence the detected biodiversity with eDNA. The lower volume of water sampled in the Western Coral Triangle region (2 l per sample, so 4 l per station using point-sampling instead of 2 km transect with 30 l elsewhere), could underestimate fish biodiversity. However, previous studies show that MOTU accumulation curves based on this dataset were close to the total fish diversity reported in this region [32]. Furthermore, β-diversity between samples within stations in each region indicates that dissimilarity between samples is not greater in the Western Coral Triangle than in other regions (electronic supplementary material, figure S11). To account for differences in sample size and obtain a balanced design, we performed sensitivity analyses by rarefying our complete dataset to (i) four stations for all sites and (ii) four sites per region after removing the lowest sampled region (Southeast Polynesia) (electronic supplementary material, analyses, figures S1–S4). We obtained similar patterns even after subsampling stations or sites. However, our site-based and station-based accumulation curves do not reach plateaus suggesting that our sampling effort was not sufficient to exhaustively estimate fish biodiversity for each site (electronic supplementary material, analyses, figure S5) and station (electronic supplementary material, analyses, figure S6). Twenty-five replicates (so, 12 stations in case of field duplicates) could accurately estimate biodiversity regionally due to high local turnover [53]. A higher number of eDNA samples would be necessary here to reach MOTU accumulation per site and station.

The transport and degradation of eDNA can also impact species detection. As some evidence suggests that eDNA from pelagic fishes degrades slower than from inshore species [54], we cannot exclude that eDNA from pelagic and deep-water families (e.g. Myctophidae) might disperse sufficiently with sea currents such that species living close to reef habitats are detected. Environmental DNA transport could also explain the detection of some freshwater fish families (i.e. Centrarchidae, Osphronemidae or Channidae) in a few samples located near an estuary or in an enclosed bay with freshwater inputs.

Better understanding and anticipating the effects of multiple threats to the marine environment depends on the temporal and spatial extent of our monitoring capacity in the vast ocean. Environmental DNA is a powerful tool to investigate biodiversity patterns at large scale and monitor biodiversity, but still benefits from the combination with complementary approaches as visual methods for an exhaustive biodiversity survey across space and time to keep pace with ongoing changes.

4. Methods

(a) Environmental DNA collection and sample processing

Environmental DNA seawater samples were collected between 2017 and 2019, following a hierarchical pattern. A total of 226 eDNA samples (filters) were collected in 100 stations (gathering of replicates at the same location) located in 26 sites (groups of stations separated by at least 35 km) distributed across five tropical regions (electronic supplementary material, figures S1 and S2). Three different sampling methods were used comprising a 2 km-long sampling transect of 30 l (surface or bottom depth) or point samples of 2 l (electronic supplementary material, table S7 and Methods S1), and between 12 and 64 samples were collected by region. Filtration was performed with polyethersulfone filters, 0.2 µm pore size. For each sampling campaign, a strict contamination control protocol was followed in both field and laboratory stages [39]. Negative field controls were performed in multiple sites, and revealed no contamination from the boat or samplers.

(b) eDNA extraction, amplification and sequencing

DNA extraction was performed in a dedicated DNA laboratory (SPYGEN, www.spygen.com) equipped with positive air pressure, UV treatment and frequent air renewal. Decontamination procedures were conducted before and after all manipulations. Detailed protocols of DNA extraction, amplification and sequencing can be found in the electronic supplementary material, Method S2 and in [32,39]. A teleost-specific 12S mitochondrial rRNA primer pair (teleo, forward primer—ACACCGCCCGTCACTCT, reverse primer—CTTCCGGTACACTTACCATG [39]) was used for the amplification of metabarcode sequences. As we analysed our data using MOTUs as a proxy for species to overcome genetic database limitations, we chose to amplify only one marker. Teleo marker has been shown to be the most appropriate for fish, owing to its high interspecific variability, and its short size allowing us to detect rare and degraded DNA reliably [39,49,55,56]. Twelve DNA amplifications PCR per sample were performed.

(c) Bioinformatic analysis

Following sequencing, reads were processed using clustering and post-clustering cleaning to remove errors and estimate the number of species using MOTUs [51]. First, reads were assembled using *VSEARCH* [57], then demultiplexed and trimmed using *CUTADAPT* [58] and clustering was performed using *SWARM* v.2 [59] with a minimum distance of 1 mismatch between clusters. Taxonomic assignment of MOTUs was carried out using the lower common ancestor (LCA) algorithm *ecotag* implemented in the OBITOOLS toolkit [60] and the European nucleotide archive as a reference database (release 143, March 2020). Details on the bioinformatics analysis can be found in the electronic supplementary material, Methods S3. Taxonomic assignments obtained from the LCA algorithm at the species level were accepted if the percentage of similarity with the reference sequence was 100%, at the genus level if the similarity was between 90 and 99%, and at the family level if the similarity was greater than 85% following previous studies [32,61]. If these criteria were not met, the MOTU was left unassigned. Only 21% of assigned MOTUs are assigned to the family level with a similarity between 85 and 90% (electronic supplementary material, table S8).

(d) Visual census data

The visual census survey data used here is a subset (2047 transects, in 219 sites, electronic supplementary material, figure S1) of the complete visual census data (3027 transects) provided by the RLS [31] and comprises all species observed on standardized 50 m surveys at sites in tropical biogeographic realms between 2006 and 2017 (electronic supplementary material, methods S4) [62]. We selected only the most recent survey for each station and only transects with more than five per cent of coral cover. Two different sampling protocols were adapted to detect both reef and crypto-benthic fishes.

(e) Statistical analysis

More details on the statistical analysis are available in the electronic supplementary material, methods S5.

Accumulation curves were calculated for species per 500 m^2 transect, MOTUs per eDNA sample, and families per transect and sample. We used the functions 'specaccum' and 'fitspecaccum' from the R package 'vegan' which calculates the expected species accumulation curve using a sample-based rarefaction method and fit a nonlinear accumulation model. In order to assess the impact of the irregular sampling on the estimates measured with accumulation curves, we subset randomly half of the transects in the three most sampled regions in Australia and calculated again the accumulation curves for species and families (electronic supplementary material, figure S12). The results were unchanged.

Linear regression models were fitted between the number of MOTUs per family in the eDNA dataset and the number of species per family in the visual census dataset, after $\log(x+1)$ transformation (figure 1*e*).

Accumulation curves were also calculated by sub-setting MOTUs belonging to crypto-benthic orders, or to pelagic families, for both datasets (figure 2). The asymptote was calculated as described above.

We performed dbRDA on family proportions, with *region* and *site richness* as explanatory variables, using the function *capscale* from the *vegan* package. We subset the visual census to select only the 68 sites that fell into the five regions in common with the eDNA dataset. Total dbRDA provided the effects of each of the variables and their interaction. We then calculated partial dbRDA to measure the effect of the region while correcting for the effect of site richness (figure 3; electronic supplementary material, table S3).

We applied an additive partitioning framework [63] to separate the total MOTUs diversity at the global scale (γ global) into contributions at smaller scales from regions to local richness: $\gamma_{\text{global}} = \beta_{\text{inter-region}} + \text{mean}\ \beta_{\text{inter-site}} + \text{mean}\ \beta_{\text{inter-station}} + \text{mean}\ \bar{\alpha}_{\text{station}}$. In this additive framework, the three levels of biodiversity [64] (i.e. α, β and γ) are expressed with the same unit and consequently the contribution of α and β diversity to total diversity (γ) can be directly compared [65].

We analysed the distribution of fish MOTU and species occurrences using gSAD which plots, on a log–log scale, the number of species as a function of the number of observations [37].

Data accessibility. All eDNA data (except New Caledonia) and visual census data are available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.3xsj3txj2. New Caledonia eDNA data are available in Zenodo, https://doi.org/10.5281/zenodo.6381130. Code used for the analyses is available at https://github.com/virginiemarques/Global_eDNA. The bioinformatic pipeline used to analyse the metabarcoding data has been published in [51].

Authors' contributions. L.M.: conceptualization, data curation, formal analysis, methodology, visualization, writing—original draft and writing—review and editing; V.M.: conceptualization, data curation, formal analysis, methodology, visualization, writing—original draft and writing—review and editing; D.M.: conceptualization, funding acquisition, project administration, resources, supervision, validation, writing—original draft and writing—review and editing; C.A.: resources and writing—review and editing; M.A.: resources and writing—review and editing; F.B.: resources and writing—review and editing; G.H.B.-P.: resources and writing—review and editing; T.D.:

data curation, formal analysis and writing—review and editing; G.J.E.: resources and writing—review and editing; J.G.: resources and writing—review and editing; P.-E.G.: data curation, methodology, software and writing—review and editing; R.H.: funding acquisition, resources and writing—review and editing; J.-B.J.: resources and writing—review and editing; K.: resources and writing—review and editing; E.M.: resources and writing—review and editing; G.M.: resources and writing—review and editing; M.M.: resources and writing—review and editing; A.P.F.: resources and writing—review and editing; L.P.: resources and writing—review and editing; R.D.S.-S.: resources and writing—review and editing; H.Y.S.: resources and writing—review and editing; A.V.: data curation, methodology, resources and writing—review and editing; L.V.: resources and writing—review and editing; I.B.V.: resources and writing—review and editing; L.P.: conceptualization, funding acquisition, investigation, project administration, supervision, validation, writing—original draft and writing—review and editing; S.M.: conceptualization, funding acquisition, investigation, project administration, supervision, validation, writing—original draft and writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

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

Funding. The sampling in the Caribbean, Indian Ocean and Polynesia and the sequencing were funded by Monaco Explorations. Fieldwork in Indonesia and laboratory activities were supported by the Lengguru 2017 Project (www.lengguru.org), conducted by the French National Research Institute for Sustainable Development (IRD), National Research and Innovation Agency (BRIN) with the Research Center for Oceanography (RCO), the Politeknik Kelautan dan Perikanan Sorong), the University of Papua (UNIPA) with the help of the Institut Français in Indonesia (IFI), funding from Monaco Explorations, and corporate sponsorship from the Total Foundation and TIPCO company. Fieldwork and laboratory activities in New-Caledonia were supported by the projects ANR SEAMOUNTS and CIFRE REEF 3.0 conducted by the French National Institute for Sustainable Development (IRD) and GINGER-BURGEAP-SOPRONER company with funding from Monaco Explorations. Fieldwork and laboratory activities in Colombia were supported by Monaco Explorations, ETH Global grant and the project Reefish, conducted in collaboration with the Instituto de Investigaciones Marinas y Costeras – INVEMAR. Monaco Explorations supported also sampling and sequencing in the Caribbean. Fieldwork in the French Scattered islands was supported by the Terres Australes et Antartiques Françaises (TAAF).

Acknowledgements. The authors thank all staff and students involved in fieldwork and acknowledge SPYGEN staff for the technical support in the eDNA laboratory.

References

1. Parravicini V *et al.* 2013 Global patterns and predictors of tropical reef fish species richness. *Ecography (Cop)* **36**, 1254–1262. (doi:10.1111/j.1600-0587.2013.00291.x)
2. Cowman PF, Bellwood DR. 2013 The historical biogeography of coral reef fishes: global patterns of origination and dispersal. *J. Biogeogr.* **40**, 209–224. (doi:10.1111/jbi.12003)
3. Cinner JE *et al.* 2020 Meeting fisheries, ecosystem function, and biodiversity goals in a human-dominated world. *Science* **368**, 307–311. (doi:10.1126/science.aax9412)
4. Hoegh-Guldberg O *et al.* 2019 The human imperative of stabilizing global climate change at 1.5°C. *Science* **365**, eaaw6974. (doi:10.1126/science.aaz4390)
5. Victor BC. 2015 How many coral reef fish species are there? Cryptic diversity and the new molecular taxonomy. In *Ecology of fishes on coral reefs*, pp. 76–88. Cambridge, UK: Cambridge University Press.
6. Siqueira AC, Morais RA, Bellwood DR, Cowman PF. 2020 Trophic innovations fuel reef fish diversification. *Nat. Commun.* **11**, 1–11. (doi:10.1038/s41467-020-16498-w)
7. Bellwood D, Wainwright P. 2002 The history and biogeography of fishes on coral reefs. In *Coral reef fishes: dynamics and diversity in a complex ecosystem* (ed. P Sale). San Diego, CA: Academic Press.
8. Bellwood DR, Hughes TP. 2001 Regional-scale assembly rules and biodiversity of coral reefs. *Science* **292**, 1532–1534. (doi:10.1126/science.1058635)
9. Pellissier L *et al.* 2014 Quaternary coral reef refugia preserved fish diversity. *Science* **344**, 1016–1020. (doi:10.1126/science.1249853)
10. Veron J, Devantier LM, Turak E, Green AL, Kininmonth S, Stafford-Smith M, Peterson N. 2009 Delineating the coral triangle. Galaxea. *J. Coral Reef Stud.* **11**, 91–100. (doi:10.3755/galaxea.11.91)
11. Siqueira AC, Morais RA, Bellwood DR, Cowman PF. 2021 Planktivores as trophic drivers of global coral reef fish diversity patterns. *Proc. Natl Acad. Sci. USA* **118**, e2019404118. (doi:10.1073/pnas.2019404118)
12. Morais RA, Bellwood DR. 2019 Pelagic subsidies underpin fish productivity on a degraded coral reef. *Curr. Biol.* **29**, 1521–1527. (doi:10.1016/j.cub.2019.03.044)
13. 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.12423)
14. Bender MG, Leprieur F, Mouillot D, Kulbicki M, Parravicini V, Pie MR, Barneche DR, Oliveira-Santos LGR, Floeter SR. 2017 Isolation drives taxonomic and functional nestedness in tropical reef fish faunas. *Ecography (Cop)* **40**, 425–435. (doi:10.1111/ecog.02293)
15. Hughes TP, Bellwood DR, Connolly SR, Cornell HV, Karlson RH. 2014 Double jeopardy and global extinction risk in corals and reef fishes. *Curr. Biol.* **24**, 2946–2951. (doi:10.1016/j.cub.2014.10.037)
16. Mouillot D, Villéger S, Parravicini V, Kulbicki M, Arias-gonzález JE. 2014 Functional over-redundancy and high functional vulnerability in global fish faunas on tropical reefs. *Proc. Natl Acad. Sci. USA* **111**, 13 757–13 762. (doi:10.1073/pnas.1317625111)
17. Alzate A, Zapata FA, Giraldo A. 2014 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. (doi:10.15517/rbt.v62i0.16361)
18. Bellwood D, Renema W, Rosen BB. 2012 Biodiversity hotspots, evolution and coral reef biogeography: a review. In *Biotic evolution and environmental change in Southeast Asia*, pp. 216–245. Cambridge, UK: Cambridge University Press.
19. Mora C. 2015 Ecology of fishes on coral reefs. In *Ecology of fishes on coral reefs*, pp. 1–374. Cambridge, UK: Cambridge University Press.
20. Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH. 2012 Environmental DNA. *Mol. Ecol.* **21**, 1789–1793. (doi:10.1111/j.1365-294X.2012.05542.x)
21. Boulanger E *et al.* 2021 Environmental DNA metabarcoding reveals and unpacks a biodiversity conservation paradox in Mediterranean marine reserves [Dryad Dataset]. *Proc. R. Soc. B* **288**, 20210112. (doi:10.1098/rspb.2021.0112)
22. Harrison JB, Sunday JM, Rogers SM. 2019 Predicting the fate of eDNA in the environment and implications for studying biodiversity. *Proc. R. Soc. B* **286**, 1–9. (doi:10.1098/rspb.2019.1409)
23. Cordier T *et al.* 2020 Ecosystem monitoring powered by environmental genomics: a review of current strategies with an implementation roadmap. *Mol. Biol. Evol.* **30**, 2937–2958. (doi:10.1111/mec.15472 1–22)
24. De Vargas C *et al.* 2015 Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1–11. (doi:10.1126/science.1261605)
25. Valdivia-Carrillo T, Rocha-Olivares A, Reyes-Bonilla H, Domínguez-Contreras JF, Munguia-Vega A. 2021 Integrating eDNA metabarcoding and simultaneous underwater visual surveys to describe complex fish communities in a marine biodiversity hotspot. *Mol. Ecol. Resour.* **21**, 1558–1574. (doi:10.1111/1755-0998.13375)

26. West K *et al.* 2021 Large-scale eDNA metabarcoding survey reveals marine biogeographic break and transitions over tropical north-western Australia. *Divers. Distrib.* **27**, 1942–1957. (doi:10.1111/ddi.13228)
27. Kume M *et al.* 2021 Factors structuring estuarine and coastal fish communities across Japan using environmental DNA metabarcoding. *Ecol. Indic.* **121**, 107216. (doi:10.1016/j.ecolind.2020.107216)
28. Fraija-Fernández N, Bouquieaux MC, Rey A, Mendibil I, Cotano U, Irigoien X, Santos M, Rodríguez-Ezpeleta N. 2020 Marine water environmental DNA metabarcoding provides a comprehensive fish diversity assessment and reveals spatial patterns in a large oceanic area. *Ecol. Evol.* **10**, 7560–7584. (doi:10.1002/ece3.6482)
29. Aglieri G *et al.* 2020 Environmental DNA effectively captures functional diversity of coastal fish communities. *Mol. Ecol.* **30**, 3127-3139. (doi:10.1111/mec.15661 1–13)
30. DiBattista JD *et al.* 2021 Environmental DNA reveals a multi-taxa biogeographic break across the Arabian Sea and Sea of Oman. *Environ. DNA* **4**, 206–221. (doi:10.1002/edn3.252)
31. Edgar GJ, Stuart-Smith RD. 2014 Systematic global assessment of reef fish communities by the Reef Life Survey program. *Sci. Data* **1**, 1–8. (doi:10.1038/sdata.2014.7)
32. Juhel JB *et al.* 2020 Accumulation curves of environmental DNA sequences predict coastal fish diversity in the coral triangle. *Proc. Biol. Sci.* **287**, 1–10. (doi:10.22541/au.158082947.75232617)
33. Castellanos-Galindo GA, Krumme U, Rubio EA, Saint-Paul U. 2013 Spatial variability of mangrove fish assemblage composition in the tropical eastern Pacific Ocean. *Rev. Fish Biol. Fish.* **23**, 69–86. (doi:10.1007/s11160-012-9276-4)
34. Willis TJ, Anderson MJ. 2003 Structure of cryptic reef fish assemblages: relationships with habitat characteristics and predator density. *Mar. Ecol. Prog. Ser.* **257**, 209–221. (doi:10.3354/meps257209)
35. Harnik PG, Simpson C, Payne JL. 2012 Long-term differences in extinction risk among the seven forms of rarity. *Proc. R. Soc. B* **279**, 4969–4976. (doi:10.1098/rspb.2012.1902)
36. Crist TO, Veech JA. 2006 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. (doi:10.1111/j.1461-0248.2006.00941.x)
37. Enquist BJ *et al.* 2019 The commonness of rarity: global and future distribution of rarity across land plants. *Sci. Adv.* **5**, 1–14. (doi:10.1126/sciadv.aaz0414)
38. Dornelas M, Connolly SR, Hughes TP. 2006 Coral reef diversity refutes the neutral theory of biodiversity. *Nature* **440**, 80–82. (doi:10.1038/nature04534)
39. 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)
40. Boussarie G *et al.* 2018 Environmental DNA illuminates the dark diversity of sharks. *Sci. Adv.* **4**, 1–8. (doi:10.1126/sciadv.aap9661)
41. Brandl SJ *et al.* 2019 Demographic dynamics of the smallest marine vertebrates fuel coral-reef ecosystem functioning. *Science,* **364**(6446), 1189–1192.
42. Kimmerling N *et al.* 2018 Quantitative species-level ecology of reef fish larvae via metabarcoding. *Nat. Ecol. Evol.* **2**, 306–316. (doi:10.1038/s41559-017-0413-2)
43. Beckley LE, Holliday D, Sutton AL, Weller E, Olivar MP, Thompson PA. 2019 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. (doi:10.1016/j.dsr2.2018.03.008)
44. McLean M, Stuart-Smith RD, Villéger S, Auber A, Edgar GJ, MacNeil MA, Loiseau N, Leprieur F, Mouillot D. 2021 Trait similarity in reef fish faunas across the world's oceans. *Proc. Natl Acad. Sci. USA* **118**, e2012318118. (doi:10.1073/pnas.2012318118)
45. Gaboriau T, Leprieur F, Mouillot D, Hubert N. 2018 Influence of the geography of speciation on current patterns of coral reef fish biodiversity across the Indo-Pacific. *Ecography (Cop)* **41**, 1295–1306. (doi:10.1111/ecog.02589)
46. Ahmadia GN, Tornabene L, Smith DJ, Pezold FL. 2018 The relative importance of regional, local, and evolutionary factors structuring cryptobenthic coral-reef assemblages. *Coral Reefs* **37**, 279–293. (doi:10.1007/s00338-018-1657-2)
47. Marques V, Milhau T, Albouy C, Dejean T, Manel S, Mouillot D, Juhel JB. 2021 GAPeDNA: assessing and mapping global species gaps in genetic databases for eDNA metabarcoding. *Divers. Distrib.* **27**, 1880–1892. (doi:10.1111/ddi.13142)
48. Ruppert KM, Kline RJ, Rahman MS. 2019 Past, present, and future perspectives 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)
49. Polanco A *et al.* 2021 Comparing the performance of 12S mitochondrial primers for fish environmental DNA across ecosystems. *Environ. DNA* **3**, 1113–1127. (doi:10.1002/edn3.232)
50. Brandt MI, Trouche B, Quintric L, Günther B, Wincker P, Poulain J, Arnaud-Haond S. 2021 Bioinformatic pipelines combining denoising and clustering tools allow for more comprehensive prokaryotic and eukaryotic metabarcoding. *Mol. Ecol. Resour.* **21**, 1904–1921. (doi:10.1111/1755-0998.13398)
51. Marques V, Guérin PÉ, 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 (Cop)* **43**, 1779–1790. (doi:10.1111/ecog.05049)
52. Sales NG, Wangensteen OS, Carvalho DC, Deiner K, Praebel I, McDevitt A, Mariani S. 2020 Space-time dynamics in monitoring neotropical fish communities using eDNA metabarcoding. *bioRxiv*.
53. Stauffer S *et al*. 2021 How many replicates to accurately estimate fish biodiversity using environmental DNA on coral reefs? *bioRxiv*.
54. Collins RA, Wangensteen OS, O'Gorman EJ, Mariani S, Sims DW, Genner MJ. 2018 Persistence of environmental DNA in marine systems. *Commun. Biol.* **1**, 1–12. (doi:10.1038/s42003-018-0192-6)
55. 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)
56. Zhang S, Zhao J, Yao M. 2020 A comprehensive and comparative evaluation of primers for metabarcoding eDNA from fish. *Methods Ecol. Evol.* **2020**, 1609–1625. (doi:10.1111/2041-210X.13485)
57. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016 VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, 1–22. (doi:10.7717/peerj.2584)
58. Martin M. 1994 Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12. (doi:10.14806/ej.17.1.200)
59. Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. 2015 Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ* **3**, 1–12. (doi:10.7717/peerj.1420)
60. Boyer F, Mercier C, Bonin A, Le Bras Y, Taberlet P. 2016 Coissac E. obitools: a unix-inspired software package for DNA metabarcoding. *Mol. Ecol. Resour.* **16**, 176–182. (doi:10.1111/1755-0998.12428)
61. Polanco Fernández A *et al.* 2020 Comparing environmental DNA metabarcoding and underwater visual census to monitor tropical reef fishes. *Environ. DNA* **3**, 142–156. (doi:10.1002/edn3.140)
62. Spalding MD *et al.* 2007 Marine ecoregions of the World: a bioregionalization of coastal and shelf areas. *Bioscience* **57**, 573–583. (doi:10.1641/B570707)
63. Escalas A *et al.* 2017 Functional diversity and redundancy across fish gut, sediment and water bacterial communities. *Environ. Microbiol.* **19**, 3268–3282. (doi:10.1111/1462-2920.13822)
64. Whittaker RH. 1972 Evolution and measurement of species diversity. *Taxon* **21**, 213–251. (doi:10.2307/1218190)
65. Veech JA, Summerville KS, Crist TO, Gering JC. 2002 The additive partitioning of species diversity: recent revival of an old idea. *Oikos* **99**, 3–9. (doi:10.1034/j.1600-0706.2002.990101.x)

Chapitre 4 – Influence des facteurs environnementaux et humains sur la diversité des poissons à l'échelle globale

Demoiselle sur le récif de la passe de Uitoé, Nouvelle-Calédonie (crédit : H. Bidenbach)

1. Préface

Les poissons côtiers jouent un rôle fondamental dans le fonctionnement des écosystèmes marins et la sécurité alimentaire de plus d'un milliard de personnes dans le monde, mais sont confrontés à des menaces croissantes dues au changement climatique, à la perte d'habitat et à la surexploitation des ressources (Mellin et al. 2022; Sing Wong et al. 2022). Les conditions environnementales ont historiquement déterminé la composition des communautés de poissons (Pellissier et al. 2014; Stein et al. 2014; Pecuchet et al. 2016), mais les fortes pressions humaines de l'Anthropocène modifient de plus en plus leur structure originelle et leur fonctionnement. Le déclin continu de l'abondance des poissons et la perte des principaux prédateurs ont été largement documentés (Dulvy et al. 2021; Pacoureau et al. 2021). Cependant, la part relative des pressions environnementales et humaines affectant la biodiversité des poissons et la composition des communautés à travers les échelles spatiales (α-diversité locale et β-diversité) doit encore être quantifiée à l'échelle mondiale (Hadj-Hammou et al. 2021). L'augmentation des températures de l'océan entraine des changements de communautés sur les récifs tempérés et tropicaux, avec un remplacement des espèces les plus sensibles par des espèces généralistes (Stuart-Smith et al. 2022; Brown et al. 2022). Le déclin des espèces et les changements de composition suite à la pression humaine sont principalement liés à la dégradation de l'habitat associée à une forte pression de pêche, entraînant souvent la perte locale d'espèces spécialisées (Stuart-Smith et al. 2021; Yan et al. 2021). La disparition locale d'espèces de poissons marins due à la pression humaine ou au changement climatique reste difficile à évaluer car des populations résiduelles peuvent persister sans être détectées en raison de leur rareté ou de leur comportement modifié (Boussarie et al. 2018). Depuis le début de l'Anthropocène, la diversité phylogénétique et la diversité fonctionnelle des poissons ont également diminué dans les endroits les plus impactés (Li et al. 2020; Pimiento et al. 2020). De récentes études ont démontré la capacité de l'ADNe à estimer les diversités fonctionnelles et phylogénétiques d'un assemblage, dans le cas où les séquences sont assignées à une espèce (Aglieri et al. 2020; Marques et al. 2021), mais peu se sont intéressées à l'information génétique fournie par l'ADNe.

Au chapitre précédent, j'ai démontré que le *metabarcoding* de l'ADNe permettait d'identifier de nombreux taxa non détectés par les méthodes conventionnelles, tels que les poissons cryptobenthiques et pélagiques, mais aussi d'étudier les patrons de distribution des poissons à large échelle.

Dans ce chapitre, j'analyse les données issues d'un échantillonnage d'ADNe des écosystèmes côtiers dans tous les océans, incluant des zones polaires, tempérées et tropicales et des écosystèmes impactés ou quasi-vierges. Je modélise la distribution globale de la diversité taxonomique et génétique des séquences de poissons en fonction de facteurs géographiques (bathymétrie, distance à la côte, distance au Triangle de Corail), environnementaux (température de surface, pH, productivité) et socio-économiques (IDH, dépendance aux ressources marines, gravité). Enfin, j'évalue si un indice de diversité des séquences de poissons, mesuré par les nombres de Hill à partir de la dissimilarité entre les séquences composant une communauté, peut être un bon indicateur de la diversité fonctionnelle et phylogénétique.

Les résultats montrent que les facteurs environnementaux sont les principaux déterminants des α- et β-diversité taxonomiques et génétiques de l'ADNe de poisson dans les écosystèmes côtiers du monde entier. La diversité est plus élevée dans les zones les plus chaudes et les plus productives. Les modèles révèlent également une érosion de la biodiversité là où la dépendance humaine aux écosystèmes marins est élevée, c'est-à-dire là où les hommes vivent de la pêche (Figure 4.1) et bénéficient des autres services écosystémiques fournis par l'océan. Les résultats suggèrent que de courtes séquences d'ADNe fournissent un proxy fiable de la diversité phylogénétique et fonctionnelle et peuvent être utilisées pour évaluer les impacts climatiques et humains sur la biodiversité marine à l'échelle mondiale.

Figure 4.1. *Pêcheur remontant son filet sur un récif côtier en Martinique. (crédit : L. Mathon)*

2. Manuscrit C

Soumis à Global Ecology and Biogeography, en révision.

The global distribution of environmental DNA sequences from coastal fishes in the Anthropocene

Laetitia Mathon[1,2]*, Virginie Marques[1], Stéphanie Manel[1], Camille Albouy[3], Marco Andrello[4,5], Emilie Boulanger[6], Julie Deter[4,7], Régis Hocdé[4], Fabien Leprieur[4], Tom B. Letessier[8], Nicolas Loiseau[4], Eva Maire[9], Alice Valentini[10], Laurent Vigliola[2], Florian Baletaud[4,2,11], Sandra Bessudo[12], Tony Dejean[10], Nadia Faure[1], Pierre-Edouard Guerin[1], Meret Jucker[13,14], Jean-Baptiste Juhel[4], Kadarusman[15], Andrea Polanco F.[16,17], Laurent Pouyaud[18], Dario Schwörer[14], Kirsten F. Thompson[19,20], Marc Troussellier[4], Hagi Yulia Sugeha[21], Laure Velez[4], Xiaowei Zhang[22], Wenjun Zhong[22], Loïc Pellissier[13,23,‡], David Mouillot[4,24,‡]

[1] CEFE, Univ. Montpellier, CNRS, EPHE-PSL University, IRD, Montpellier, France
[2] ENTROPIE, Institut de Recherche pour le Développement (IRD), Univ. Réunion, UNC, CNRS, IFREMER, Nouméa, New Caledonia, France
[3] DECOD (Ecosystem Dynamics and Sustainability), IFREMER, INRAE, Inst. Agro – Agrocampus Ouest, Nantes, France
[4] MARBEC, Univ Montpellier, CNRS, IFREMER, IRD, Montpellier, France
[5] Institute for the study of Anthropic Impacts and Sustainability in the marine environment, National Research Council (CNR-IAS), Rome, Italy
[6] Aix-Marseille Université, Université de Toulon, CNRS, IRD, Mediterranean Institute of Oceanography (MIO), UM 110, 13288, Marseille, France
[7] Andromède océanologie, place cassan, Mauguio, France
[8] Institute of Zoology, Zoological Society of London, Regent's Park, NW1 4RY, London UK
[9] Lancaster Environment Centre, Lancaster University, Lancaster, LA1 4YQ, UK
[10] SPYGEN, Le Bourget-du-Lac, France
[11] SOPRONER, groupe GINGER, 98000 Noumea, New Caledonia, France
[12] Fundación Malpelo y otros ecosistemas marinos, Colombia
[13] Landscape Ecology, Institute of Terrestrial Ecosystems, Department of Environmental Systems Science, ETH Zürich, Zürich, Switzerland
[14] TOPtoTOP Global Climate Expedition, Switzerland.
[15] Politeknik Kelautan dan Perikanan Sorong, KKD BP Sumberdaya Genetik, Konservasi dan Domestikasi, Papua Barat, Indonesia
[16] Beauval Nature association, Saint Aignan sur Cher, France
[17] Programa de Biodiversidad y Ecosistemas Marinos, Museo de Historia Natural Marina de Colombia (MHNMC), Instituto de Investigaciones Marinas y Costeras- INVEMAR, Santa Marta, Colombia
[18] ISEM IRD, Univ Montpellier, Montpellier, France.
[19] University of Exeter, Exeter, Devon EX4 4PS, UK
[20] Greenpeace Research Laboratories, College of Life and Environmental Sciences, University of Exeter, Devon EX4 4RN, UK
[21] Research Center for Oceanography, National Research and Innovation Agency, Jl. Pasir Putih 1, Ancol Timur, Jakarta Utara 14430, Indonesia
[22] School of the Environment, Nanjing University, Nanjing, 210023, P. R. China
[23] Unit of Land Change Science, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland
[24] Institut Universitaire de France

* Corresponding author: laetitia.mathon@gmail.com, +33 (0)6 65 75 46 90, CEFE 1919 route de Mende, 34090 Montpellier, France

‡**These authors contributed equally as senior authors to this work.**

Keywords: Environmental DNA, coastal fish communities, environmental factors, socioeconomic factors, α-and β-diversity

Abstract

Aim: Coastal fishes have a fundamental role in marine ecosystem functioning and food security, but face increasing threats due to climate change, habitat loss and overexploitation of resources. Environmental conditions determine the composition of fish communities, but high human pressures of the Anthropocene are increasingly modifying their original structure. Here we investigate the relationship between coastal fish biodiversity and environmental and socio-economic factors, at large spatial scale, and study global scale biodiversity patterns of fish distribution.

Location: Global

Time period: Present day

Major Taxa Studied: Marine bony and cartilaginous fish

Methods: We analyzed fish environmental DNA in 263 stations across the world's oceans, in polar, temperate, and tropical regions. We modeled the effect of environmental, geographic and socioeconomic factors on α-and β-diversity, within and between assemblages. We computed partial effect of each pressure on fish community composition, at the taxonomic (taxonomic molecular units, MOTU) and sequence level. We investigated the correlation between genetic diversity measured from our barcodes, and the functional and phylogenetic diversities.

Results: We show that fish eDNA MOTU and sequence α-and β-diversity have the strongest correlation with environmental factors on coastal ecosystems worldwide. However, our models also reveal biodiversity erosion correlated with high human dependence on marine ecosystems. In areas with high fishing dependence, diversity of all fish MOTUs, cryptobenthic fish MOTUs and large fish MOTUs declined steeply. Finally, we show that a sequence diversity index, accounting for genetic diversity, within and between communities is a reliable proxy of phylogenetic and functional diversity.

Main Conclusions: Together, our results demonstrate that short eDNA sequences can be used to assess climate and direct human impacts on marine biodiversity at a global scale, and can further be used to investigate biodiversity in its phylogenetic and functional dimensions.

Introduction

Species and their habitats are under increasing threats worldwide (Andrello et al. 2022), and the biodiversity crisis is particularly acute in coastal ecosystems that provide well-being and socioeconomic benefits to over one billion people globally (Eddy et al. 2021). The ongoing decline of fish abundance and the loss of top predators have been widely reported (Pacoureau et al. 2021). However, the extent to which environmental and human pressures affect fish biodiversity and community composition across spatial scales (local α-diversity and turnover β-diversity) are yet to be quantified globally (Loiseau et al. 2021).

Global patterns of marine fish biodiversity are predominantly related to environmental factors (Pecuchet, Törnroos, & Lindegren, 2016), but the increasing human footprint across the ocean can modify the natural structure of such assemblages (O'Hara, Frazier, & Halpern, 2021). Sea surface temperature, which is inversely correlated with oxygen concentration, is the main determinant of fish species distribution (Lenoir et al. 2020) and fish trait composition (McLean, et al. 2021). High temperatures increase metabolic and reproductive rates so promote speciation rates and ultimately species richness (Fine, 2015) while Quaternary climate refugia preserved marine species from extinction (Mittelbach et al. 2007; Pellissier et al. 2014). Together with other factors (Leprieur et al. 2016; Zinke et al. 2018), past and current natural environmental gradients have structured fish assemblages across coastal bioregions (Parravicini et al. 2021).

Marine biodiversity decline and compositional shift following human pressure are predominantly linked to habitat degradation coupled with high fishing pressure, often leading to the local loss of specialist species (Stuart-Smith, Mellin, Bates, & Edgar, 2021; Yan et al. 2021). The local extirpation of marine fish species due to human pressure or climate change remains challenging to assess since residual populations may persist without being detected owing to their rarity or modified behavior. For example, 32.6% of chondrichthyan species (391 shark and ray species) are globally threatened with extinction (Dulvy et al. 2021) but some are elusive and remain unseen when using classical survey techniques like visual census or baited cameras, giving a false signal of local extirpation (Boussarie et al. 2018).

Environmental DNA (eDNA) metabarcoding improves the assessment of marine biodiversity by collecting, sequencing and analyzing small fragments of intra- and extra-cellular DNA released by organisms in their proximate environment (Miya, 2021). Environmental DNA has provided fish assemblage-wide scans in both temperate and tropical seas (Boulanger et al. 2021; Juhel et al. 2020; Polanco et al. 2021) and is less prone to false absences than classical

surveys, particularly for elusive, rare and cryptobenthic species (Boussarie et al. 2018; Mathon et al. 2022). Thus, eDNA can be a powerful tool to reveal global biodiversity patterns and their drivers. Yet, eDNA sampling is often local or regional (Valdivia-Carrillo, Rocha-Olivares, Reyes-Bonilla, Domínguez-Contreras, & Munguia-Vega, 2021; West et al. 2021) while genetic reference databases are notoriously incomplete (Marques et al. 2020) to assign eDNA sequences to known taxa preventing large scale biodiversity assessments.

Here, we take advantage of a large-scale eDNA sampling of coastal marine ecosystems across all the oceans, including pristine, tropical, temperate and polar areas, to model the global distribution of fish biodiversity according to geographic, environmental and socio-economic factors. We used the diversity of Molecular Operational Taxonomic Units (MOTUs) and of genetic sequences as alternatives to taxonomic diversity, hereafter called MOTU and sequence diversity, respectively. We hypothesize that fish MOTU diversity, is mainly shaped by the environment following the well-known latitudinal diversity gradient with a peak in the Coral Triangle (Bellwood & Hughes, 2001; Parravicini et al. 2013). However, fish MOTU diversity may decrease when human activities intensify, at least for large fish (Edgar & Stuart-Smith, 2014), revealing 'true' biodiversity erosion until local or regional extirpation occurs. Human activities may also impact sequence diversity since some families are heavily targeted by fisheries and can be locally extirpated close to humans (Cinner et al. 2018) while other evolutionary lineages, such as cryptobenthic fishes, can thrive in a human-dominated seascape (Boulanger et al. 2021; Loiseau et al. 2021). Last, we evaluate whether fish sequence diversity can be considered as a good proxy for fish functional and phylogenetic diversity across space.

Methods

Environmental DNA collection

Environmental DNA (eDNA) samples of seawater were collected between surface and 40m deep, at 263 stations, in 68 sites, across 11 marine regions covering the global ocean from pole to pole (2 polar, 3 temperate and 6 tropical regions, Figure 1). Four different sampling methods were used: (i) collection of 2L of water in DNA-free sterile plastic bags on the surface water from a small boat as well as close circuit rebreather diving (depths between 10 – 40 m) as close as possible to the habitat (Juhel et al. 2020); (ii) collection of 1L of water in sterilized bottle, from the surface; (iii) 2-km long filtration transect with two replicates (one on each side of a boat at each station for 30 min), for a total of 30L ± 15% of water just under the surface; (iv) 2

km-long filtration of water along a transect, approximately 5 m above the substrate, using a long pipe, from the boat. Details on the filtration device and storing methods can be found in Supporting Information Method S1 and Tables S1-2. For each sampling campaign, a strict contamination control protocol was followed in both field and laboratory stages (Valentini et al. 2016), and each water sample processing included the use of disposable gloves and single-use filtration equipment. Negative field controls were performed in multiple sites across all sampling locations and revealed no contamination from the boat or samplers.

eDNA extraction, amplification and sequencing

Environmental DNA extractions were performed following the protocols in (Juhel et al. 2020; Pont et al. 2018). As we analyzed our data using MOTUs as a proxy for species to overcome genetic database limitations, we chose to amplify only one marker. The teleo barcode, on the 12S mitochondrial rRNA gene (forward primer – ACACCGCCCGTCACTCT, reverse primer – CTTCCGGTACACTTACCATG (Valentini et al. 2016)) has been shown to be one of the most appropriate for fishes, owing to its high interspecific variability and its short size allowing the detection of rare and degraded DNA reliably (Collins et al. 2019; Kumar, Reaume, Farrell, & Gaither, 2022; Polanco et al. 2021; Zhang, Zhao, & Yao, 2020). The primers were 5' labeled with a unique eight-nucleotide tag (with at least three differences between tags) allowing the assignment of sequences to the respective samples during the sequence analysis. Tags for forward and reverse primers were identical for each sample. Twelve DNA amplifications PCR per sample were performed in a final volume of 25 µL, using 3 µL of DNA extract as the template (Pont et al. 2018). Details on the extraction, amplification and sequencing can be found in Supporting Information Method S2. An average of 624,468 sequence reads (paired-end Illumina or Ion Torrent) were generated per sample.

Bioinformatic analysis

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) (Marques, Guérin, et al. 2020). First, reads were assembled using vsearch (Rognes, Flouri, Nichols, Quince, & Mahé, 2016), then demultiplexed and trimmed using CUTADAPT (Martin., 1994) and clustering was performed using SWARM v.2 (Mahé, Rognes, Quince, de Vargas, & Dunthorn, 2015) with d = 1, which corresponds to a maximum of 1

mismatch between neighboring pairs of sequences within each cluster. Taxonomic assignment of MOTUs was carried out using the Lower Common Ancestor (LCA) algorithm ecotag implemented in the Obitools toolkit (Boyer et al. 2016) and the European Nucleotide Archive (ENA) as a reference database (release 143, March 2020), supplemented by our custom reference database, containing approximately 800 sequences. We discarded all observations with less than 10 reads, and present in only one PCR replicate to avoid spurious MOTUs originating from a PCR error. Then, errors generated by index-hopping (MacConaill et al. 2018) were filtered using a threshold empirically determined per sequencing batch using experimental blanks (Taberlet, Bonin, Coissac, & Zinger, 2018). Tag-jumps (Schnell, Bohmann, & Gilbert, 2015) were corrected by removing sequences with unmatching tags on the forward and reverse primers, and tolerating zero mismatch on tag sequences. An additional threshold removing all sequences with a frequency of occurrence <0.001 per MOTU and per library was implemented to clear all reads from the blanks. 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, with an identity threshold of 84% (Marques, Guérin, et al. 2020). Details on the bioinformatic processes can be found in Supporting Information Method S3. Number of reads, MOTUs and species after each cleaning step are available in Supporting Information Tables S3-4.

Explanatory factors

Environmental factors included sea surface temperature (mean SST), degree heating weeks (mean DHW), pH, net primary productivity (mean NPP), and salinity (mean SSS) that use a variety of satellite and in-situ observations, optimal interpolations and ocean system models (as documented in Supporting Information Method S4 and Table S5).

Socioeconomic factors included the Human development Index of the sovereign country in 2019 (HDI), which is a synthetic measure capturing elements of life expectancy, education and wealth (http://hdr.undp.org); an index of marine ecosystem dependence which quantifies nutritional, economic, and coastal protection dependence on marine ecosystems at the country scale (Selig et al. 2019); and an index of the human impact gravity. We calculated gravity of a sampling station as the human population size divided by the travel time between the station and this population center (in minutes). Total gravity is the sum of gravities in a buffer of 500 km around a station (Cinner et al. 2018).

Geographic factors included the bathymetry (measured directly on site with a sounder, or extracted from GEBCO_2020 Esri ASCII raster on a 15 arc-second interval grid), the depth of sampling (measured on site), the distance to shore (computed as the minimum distance between the sampling point and all shoreline points, using the function *gDistance* from the "rgeos" package) and the distance to the Coral Triangle (calculated as the geographic distance from the sampling point to the center of the Coral Triangle (longitude = 133.679826, latitude = -1.307436), using the function *pointDistance* from the "raster" package). The Coral Triangle hosts the highest fish diversity due to the development of complex reef habitats in the Miocene and the persistence of these habitats during the Quaternary climate change periods (Cowman & Bellwood, 2013; Pellissier et al. 2014). The distance to this refugia has been demonstrated to shape the traits structure and family richness in reef fishes (Parravicini et al. 2021), and can thus explain the variation of alpha and beta diversity across oceans.

Sampling factors considered included the sample method (transect or point), and the total volume filtered per station.

Selection of factors

To select only uncorrelated explanatory factors, we first computed the pairwise correlation between factors within each group of variables. For each pair of factors with a correlation higher than |0.7|, we removed the factor involved in several high correlations and kept the other one, until all pairwise correlations were inferior to |0.7|. The same process was carried out after assembling all factors from all types in one dataset (Supporting Information Figures S1-2). Factors with a large amplitude were transformed as log10(x+1).

Maps of SST, gravity and marine ecosystem dependence can be found in Supporting Information Figures S3-7.

Statistical analyses

All statistical analyses were run at the station level, pooling reads from samples and PCR replicates. All analyses were run in R version 4.1.1. Details on statistical analyses can be found in Supporting Information Method S5.

MOTU diversity

Fish MOTU diversity, expressed as the number of distinct MOTUs, was calculated at each station, as well as the MOTU diversity of fish from large fish families (n = 479 MOTUs) and cryptobenthic families (n = 539 MOTUs). MOTU diversities were log-transformed. The selection of cryptobenthic MOTUs was made according to the definition of cryptobenthic families (Brandl, Goatley, Bellwood, & Tornabene, 2018), so families characterized by the high prevalence (> 10%) of small-bodied species (< 50 mm). To select the large fish MOTUs, we extracted the length of all fish species from FishBase, computed the mean and 5th and 95th quantiles for each family and order, and selected species belonging to families and orders with a 5th quantile superior to 20 cm. MOTU α-diversity corresponds to numbers of fish MOTUs per station, and is independent of the taxonomic assignment which was only used to select the MOTUs belonging to cryptobenthic and large fish families.

Sequence α-diversity

To compute the sequence α-diversity for each station, we first computed the genetic distances between each pair of sequences with the function *dist.gene* from package "ape". We then applied the unifier framework based on generalizations of Hill number to measure sequence diversity. Hill numbers have been recommended to produce reliable diversity assessments from molecularly characterized samples (Alberdi & Gilbert, 2019; Mächler, Walser, & Altermatt, 2021). We used the function *alpha.fd.hill* from package "mFD" (Magneville et al. 2022), with parameters q = 0, which gives equal weight to all sequences, and τ as equal to the mean genetic distance (Chao et al. 2019).

Modeling MOTU and sequence α-diversity

We investigated the relationship between fish MOTU and sequence diversity at each station and all explanatory factors with a generalized least square model (GLS) that considers the spatial autocorrelation between samples. A variance inflation factor (VIF) approach was used to identify and remove residual collinear factors (factors with VIF > 10). We tested for spatial autocorrelation in the model residuals using the Moran's index I. Standardized effect sizes of each explanatory factor were extracted with the function *effectsize* from the "effectsize" package. Partial relationships between response variables and each explanatory factor while controlling for all the other factors were visualized with the function *visreg* from the "visreg"

package. The same procedures were repeated for cryptobenthic and large fish MOTU α-diversity within stations. The volume filtered at the station and the sampling method were included in the model to account for the heterogeneity in our sampling design and effort.

Sensitivity analyses were performed on 10 subset datasets after randomly removing 20% of the stations, to assess the robustness of our models, and after removing samples from polar regions (Scotia Sea and Arctic), to control for the influence of these extreme regions.

Modeling β-diversity

The Jaccard dissimilarity index was computed between stations using fish MOTU composition (presence/absence) with the function *vegdist* from package "vegan". Similarly, we computed the dissimilarity in sequence β-diversity between each pair of stations using the Hill number framework. The sequence β-diversity was calculated with the function *beta.fd.hill* from the "mFD" package, with parameter q = 0 and tau = "mean".

We then performed a distance-based redundancy analysis (dbRDA) on the sequence β-diversity and MOTU β-diversity matrices. To account for spatial autocorrelation in our samples, we first computed distance-based Moran Eigenvectors Maps (dbMEM) with the function *dbmem* from the "adespatial" package, which returned 15 dbMEM. We then ran the dbRDA on the full model, with all explanatory factors and 5 most explicative dbMEMs. Factors with VIF > 10 were removed and a final partial dbRDA was run with all selected explanatory factors, and with sampling factors and dbMEMs as conditional variables. Partial R^2 for each group of factors were obtained with the *varpart* function of the "vegan" package.

Functional and phylogenetic diversity

We explored the relationship between the pairwise sequence, phylogenetic and functional distances, by selecting only the MOTUs assigned to the species level in our dataset (n = 787). We computed genetic pairwise distances for these species with the function *dist.gene* from package "ape". We computed functional Gower distance based on functional traits extracted from fishbase, available for 685 of our species (habitat, substrate, depth range, longevity, vulnerability, length, weight, position in the water column, diet, interaction) using the function *compute_dist_matrix* from package "funrar". The phylogenetic distance between species was computed using the functions *fishtree_phylogeny* from the "fishtree" package and

cophenetic.phylo from the package "ape", and the phylogeny from (Rabosky et al. 2018). These distance matrices were compared with a mantel test, and by calculating the area under the curve (AUC) criterion, based on Somer's D statistic. AUC varies between 0 (no correlation) and 1 (identical matrices), and is computed with the functions coranking, R_NX and AUC_ln_K from package "coRanking".

We applied the α- and β-diversity Hill number framework for sequence, functional and phylogenetic diversity (q = 0 and τ equal to the mean genetic, functional or phylogenetic distance), using the functions *alpha.fd.hill* and *beta.fd.hill* from "mFD" package, for sequence and functional diversities, and function *ChaoPD* from package "entropart" for phylogenetic diversity. Alpha diversity indices were compared with a Pearson correlation test, and the β-diversity matrices were compared with AUC and Mantel tests.

Results

Global biodiversity patterns

From the 584 seawater eDNA samples collected at 263 stations across 11 marine regions (Figure 1A), we found a global MOTU diversity of 2,888 MOTUs, of which 2,276 were assigned at least to the family level (539 MOTUs belong to cryptobenthic families and 479 to large fish families), and we identified 791 distinct fish species.

The regions with the highest detected fish diversity were Lengguru in the Western Coral Triangle (1,145 MOTUs) and New-Caledonia in the Tropical Southwestern Pacific (917 MOTUs), followed by the Caribbean (Tropical Northwestern Atlantic, 452 MOTUs), the Scattered Islands (Western Indian Ocean, 357 MOTUs), the Mediterranean Sea (249 MOTUs), Southeast Polynesia (197 MOTUs) and the Tropical East Pacific (153 MOTUs). The lowest fish MOTU diversity was found in the Northeast Atlantic Ocean (Lusitanian, 96 MOTUs), the Yellow Sea (Cold Temperate Northwest Pacific Ocean, 42 MOTUs), the Antarctic Ocean (Scotia Sea, 40 MOTUs) and the Arctic Ocean (33 MOTUs).

Local or α-diversity ranged between 2 and 414 fish MOTUs per station (Supporting Information Figure S8). Cryptobenthic fish diversity ranged between 0 and 95 MOTUs per station (Supporting Information Figure S9). Large fish diversity ranged between 0 and 67 MOTUs per station while 16% of stations had no MOTU belonging to large fish (Supporting Information Figure S10). Yet, the high variability in MOTU diversity among stations cannot be

directly interpreted, partly due to the heterogeneity in the sampling design and effort that are accounted for in further analyses.

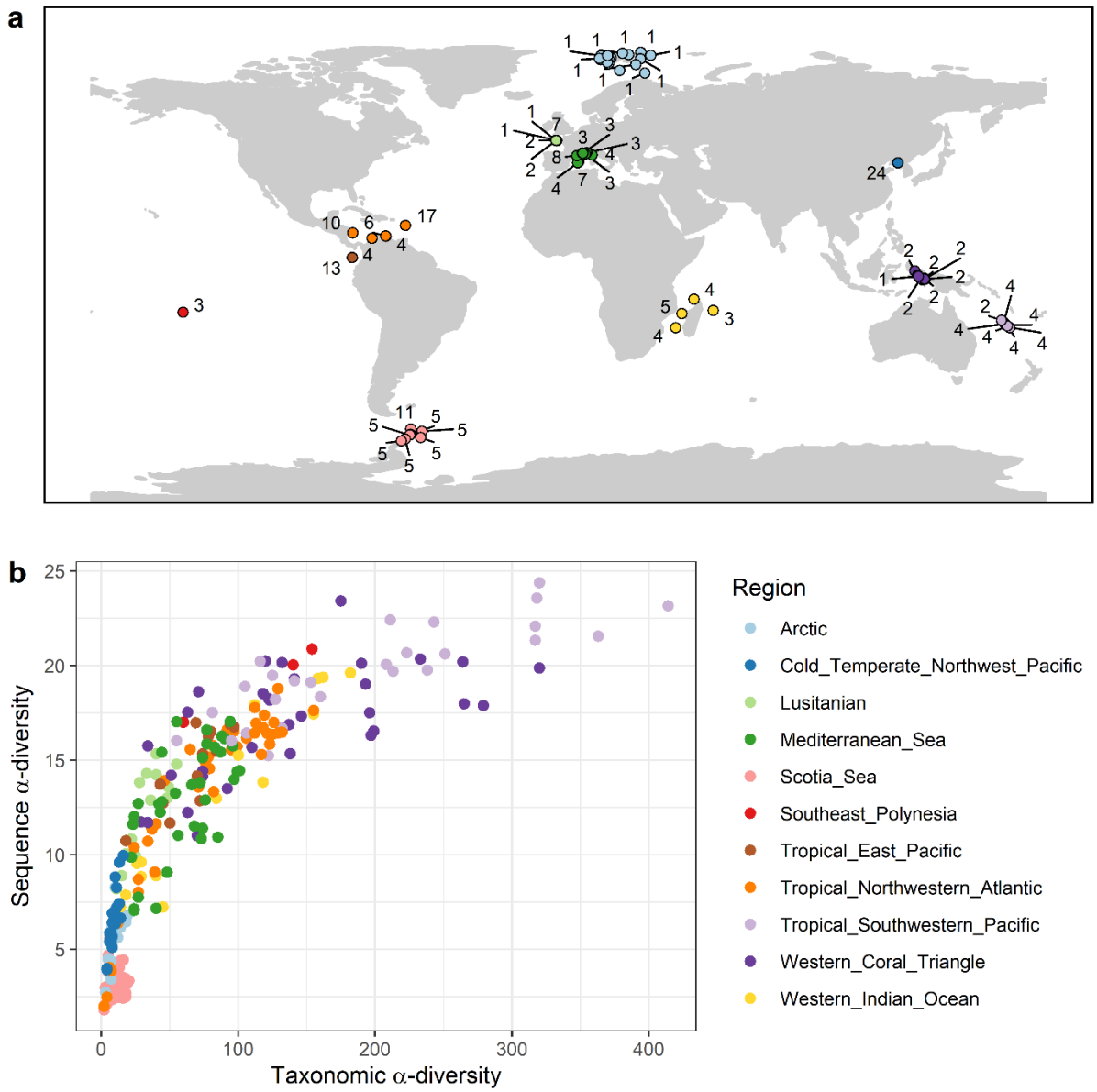


Figure 1. *Sampling locations and patterns of α-diversity. (A) Map of the sampling sites, with the number of stations per site and (B) relationship between fish (molecular operational taxonomic unit (MOTU) α-diversity (number of MOTUs per station) and sequence α-diversity (expressed as Hill number with genetic relatedness between MOTUs). The Spearman correlation between MOTU and sequence α-diversity is rho = 0.92 (p<0.001).*

Sequence α-diversity in our stations ranged between 1.7 and 16.3 (Figure 1B, Supporting Information Figure S11). The highest values of sequence α-diversity were found in the Western Coral Triangle and Tropical Southwestern Pacific. The lowest sequence α-diversity was observed at the poles (Arctic and Scotia Sea) while temperate regions showed an intermediate

level of sequence α-diversity. MOTU and sequence α-diversity were significantly and positively correlated (Spearman's rho = 0.92, $p < 0.001$), indicating that richer stations are composed of more genetically differentiated sequences (Figure 1B).

Modeling α-diversity patterns

The GLS model fitted on fish MOTU and sequence α-diversity both revealed high explanatory power, with adjusted R^2 of 0.81 and 0.78, respectively. Both responses were primarily related to environmental factors, then to geographic and socio-economic factors (Supporting Information Figures S12-13). The partitioning R^2 (see Methods) showed that fish MOTU diversity was mainly linked to environmental factors ($R^2 = 0.32$), then to geographic ($R^2 = 0.21$), sampling ($R^2 = 0.14$) and socio-economic factors ($R^2 = 0.09$). Sequence α-diversity was also primarily related to environmental factors ($R^2 = 0.36$), then to geographic ($R^2 = 0.18$), socio-economic ($R^2 = 0.11$) and sampling factors ($R^2 = 0.09$).

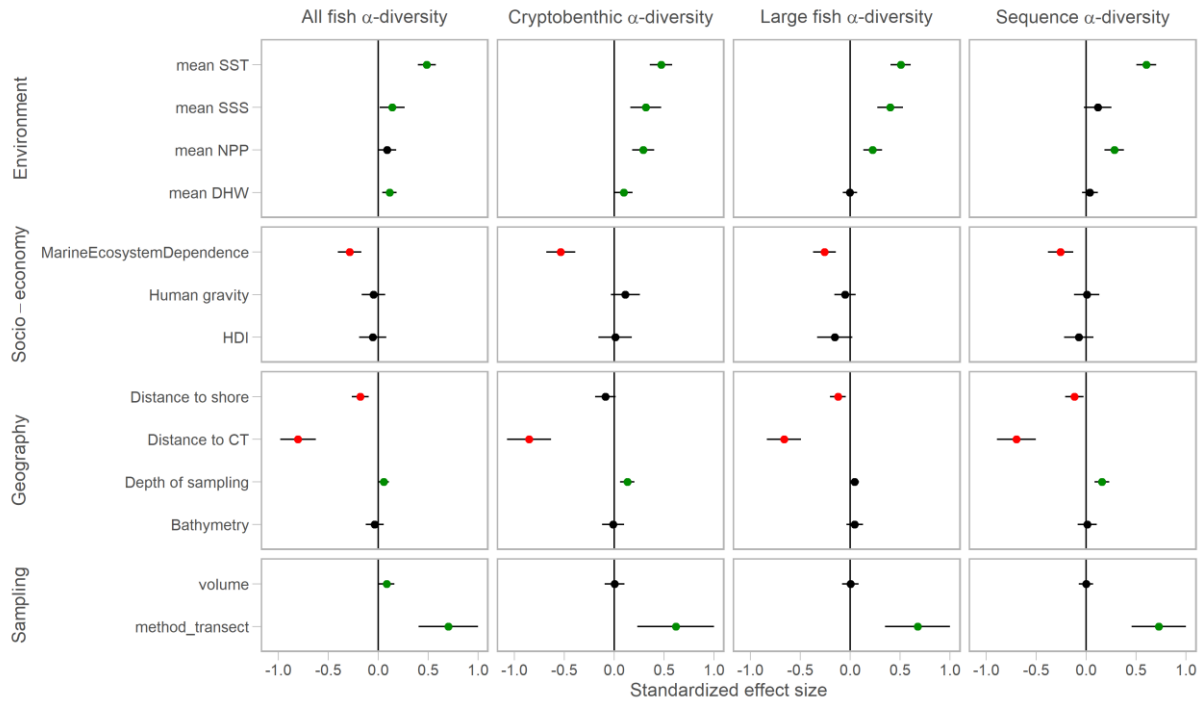


Figure 2. *Effect size of factors in GLS models predicting the level of fish molecular operational taxonomic unit (MOTU) and sequence α-diversity, but also cryptobenthic (n = 539 MOTUs) and large fish MOTU α-diversity (n = 479 MOTUs). Segments indicate 95% confidence intervals. Red dots indicate significant negative effects and green dots indicate significant positive effects while black dots are for non-significant effects. All factors and their acronyms are presented in the Methods.*

Most environmental factors showed a significant and positive relationship with MOTU and sequence α-diversity (Supporting Information Tables S6-7). MOTU α-diversity increased with sea surface temperature and salinity while sequence α-diversity increased with SST and NPP (Figure 2). Both MOTU and sequence α-diversity were significantly and negatively related to the distance to the shore and the distance to the Coral Triangle. Sequence α-diversity was positively related to the depth of sampling. Among the socio-economic factors, MOTU and sequence α-diversity were significantly and negatively related to marine ecosystem dependence while MOTU or sequence α-diversity showed no significant relationship with human gravity.

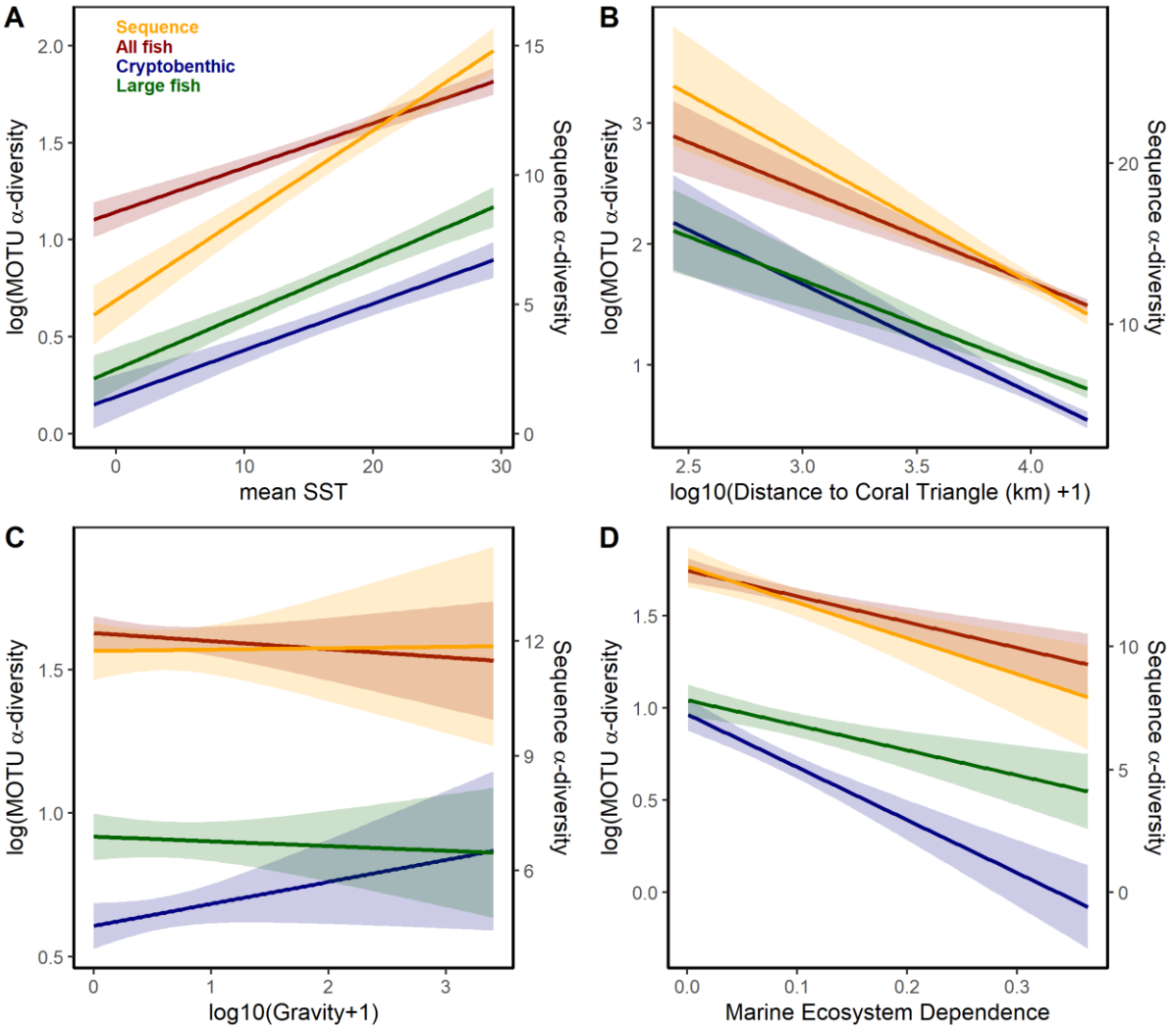


Figure 3. *Partial regression plots showing the relationships between the α-diversity of all fish molecular operational taxonomic units (MOTU) (red), large fish MOTUs (green, n = 479 MOTUs), cryptobenthic MOTUs (blue, n = 539 MOTUs) and all fish sequences (yellow, right y axis), and the four main factors conditioned on the median value of all other retained factors. Factors were (A) Sea Surface Temperature, (B) Distance to Coral Triangle, (C) gravity and (D) marine ecosystem dependence. The colored shaded areas are the 95% confidence intervals of the relationships.*

The GLS model was then fitted on cryptobenthic and large fish MOTU diversity and both showed high explanatory power (R^2 = 0.71 and 0.83, respectively). Both MOTU diversities were primarily associated to environmental factors, and then to geographic and socio-economic factors (Supporting Information Figures S14-15 and Tables S8-9). Cryptobenthic and large fish MOTU diversity increased with SST and NPP (Figure 2). Both cryptobenthic and large fish diversity significantly decreased with distance to the shore and distance to the Coral Triangle. Only cryptobenthic MOTU diversity was positively related to the depth of sampling. Cryptobenthic and large fish diversity both decreased significantly with increasing country's dependence on marine resources, but none was significantly related to human gravity (Figure 3). However, the combined effects of gravity and marine ecosystem dependence amplified the decrease of all fish and large fish MOTU diversity (Figure 4). Sensitivity analyses provided similar results (Supporting Information Figures S16-17).

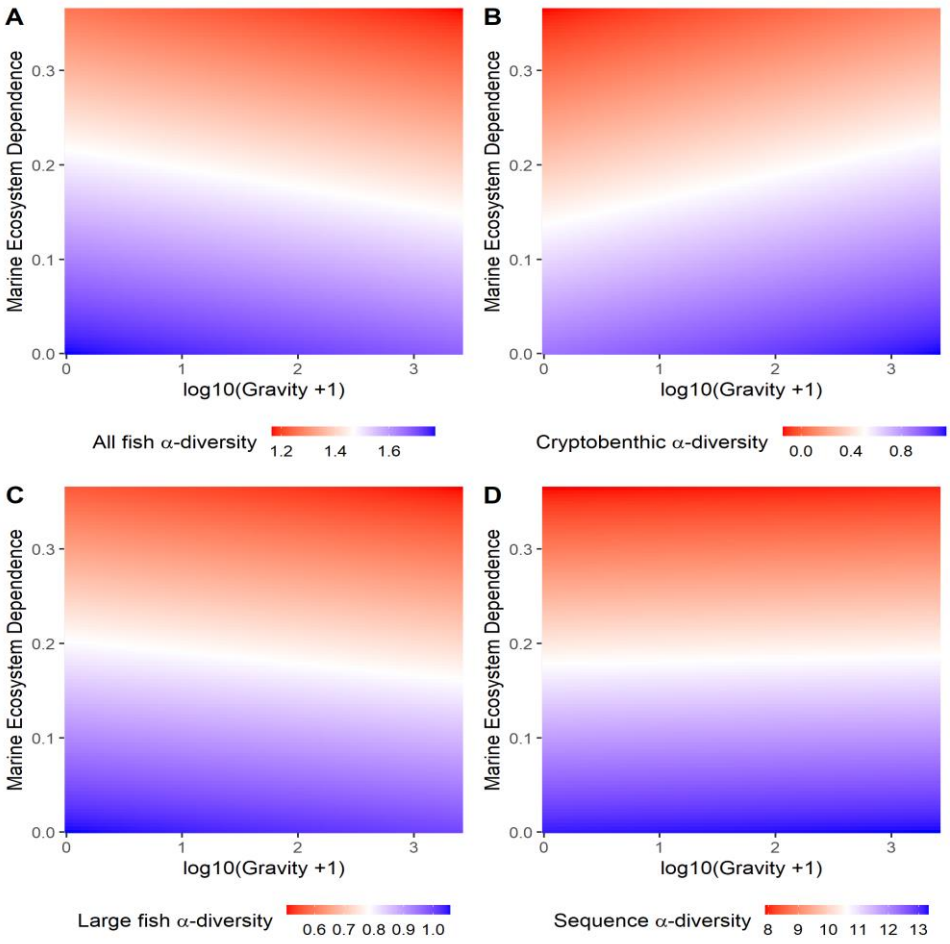


Figure 4. *Partial regression plots showing the relationships between the α-diversity of (A) all fish molecular operational taxonomic unit (MOTU), (B) large fish MOTUs (n = 479 MOTUs), (C) cryptobenthic MOTUs (n = 539 MOTUs), and (D) all fish sequences, and the combined influence of human gravity and marine ecosystem dependence. Blue indicates high diversity values and red indicates low diversity values.*

Modeling β-diversity patterns

The dbRDAs on MOTU and sequence β-diversity between stations showed a marked dissimilarity but with a low to moderate explanatory power (R^2 = 0.13 and 0.35 respectively) (Figure 5). MOTU β-diversity was related to environmental (R^2 = 0.07), socio-economic (R^2 = 0.05) and geographic factors (R^2 = 0.03) (Supporting Information Figure S18). Sequence β-diversity was mainly related to environmental (R^2 = 0.26) and geographic factors (R^2 = 0.16), then to socio-economic factors (R^2 = 0.06) (Supporting Information Figure S19). The Antarctic substantially differed from all other regions on the first axis of both dbRDAs, indicating a distinct fish MOTU and sequence composition from other regions. The second axis of the MOTU dbRDAs differentiated the Mediterranean temperate region from the East Pacific and the Caribbean, while all tropical regions were grouped together. Fish MOTU composition was more similar when considering the Atlantic, China and Arctic regions. The MOTU composition of the Caribbean and East Pacific regions were distinct from the other tropical stations.

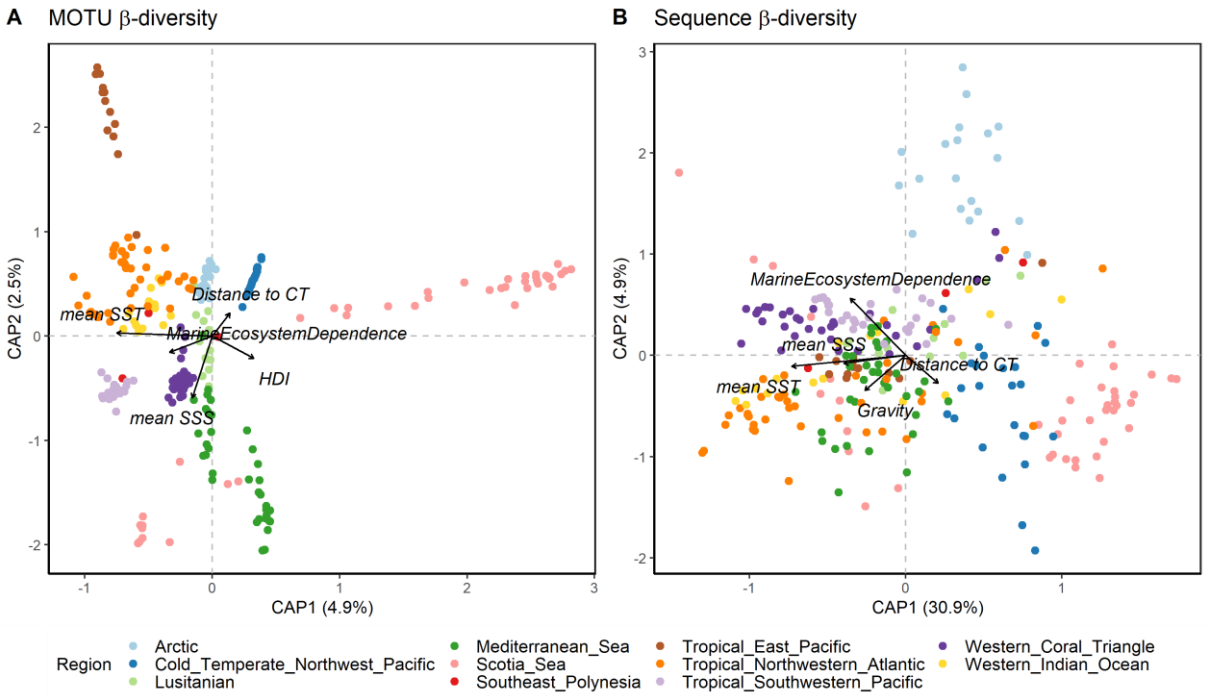


Figure 5. *Distance-based redundancy analysis (dbRDA) showing the variation in fish (A) molecular operational taxonomic unit (MOTU) composition and (B) sequence composition between stations, according to 12 factors, with sampling factors and distance-based Moran Eigenvectors Maps (dbMEMs) as conditional variables. Stations are colored by marine region and only the main factors are shown.*

The pattern of sequence β-diversity differed from that of MOTU β-diversity, with the second axis differentiating the Arctic from the tropics and temperate regions. Even though MOTU composition was similar between the Arctic and temperate regions, the sequence β-diversity was high, due to fewer species in the Arctic and a few very distant species. The Caribbean and East Pacific, however, had similar sequence composition in comparison with the other tropical regions, while the MOTU composition greatly differed. The factors showing the strongest relation with fish MOTU β-diversity were SST, NPP, marine ecosystem dependence and distance to the Coral Triangle (Supporting Information Table S10). Sequence β-diversity between stations was mainly related to SST, SSS, bathymetry, marine ecosystem dependence and distance to the shore (Supporting Information Table S11).

Sequence diversity as a proxy for functional and phylogenetic diversity

The sequence pairwise distance between assigned species pairs was significant and positively but weakly correlated to phylogenetic (mantel = 0.23, $p < 0.001$, AUC = 0.18,) or functional distance (Mantel = 0.04, $p < 0.001$, AUC = 0.027) between these species pairs (Supporting Information Figure S20). In contrast, the sequence α-diversity was positively and strongly correlated to phylogenetic α-diversity (Pearson's correlation coefficient $r = 0.94$, $p < 0.001$) and to functional α-diversity (Pearson's $r = 0.91$, $p < 0.001$) (Figure 6). The sequence β-diversity between samples was also strongly correlated to phylogenetic (Mantel test $r = 0.91$, $p < 0.001$; AUC = 0.58) and functional β-diversity (Mantel test r = 0.87, $p < 0.001$; AUC = 0.55). These results indicate that, although the genetic distance based on a short raw sequence is weakly representative of the phylogenetic or functional distance between species pairs, the sequence-based diversity indices computed from genetic distances can accurately represent the level of phylogenetic and functional diversity of fishes within and between stations.

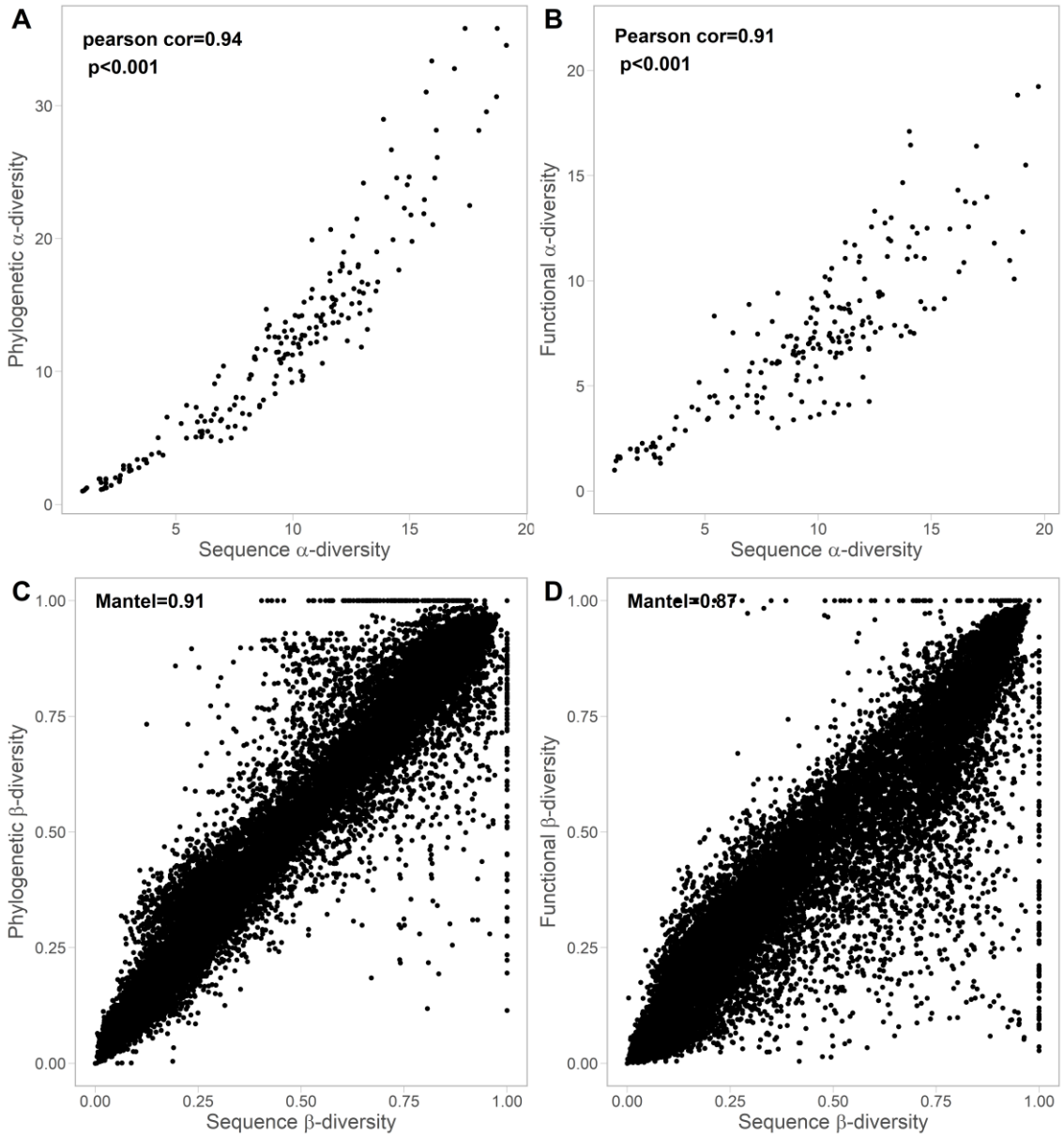


Figure 6. *Correlation between fish sequence diversity and phylogenetic or functional diversity for all stations, considering only the 787 MOTUs assigned to the species level. (A) Phylogenetic and sequence a-diversity, (B) Functional and sequence a-diversity, (C) Phylogenetic and sequence β-diversity, and (D) Functional and sequence β-diversity.*

Discussion

Through the analysis of 584 eDNA samples distributed in 263 stations across the world's oceans, we show that fish MOTU and sequence diversity show the strongest relation with environmental factors. However, we also reveal a negative relation between coastal fish diversity and human pressures, suggesting that the global distribution of environmental DNA sequences released by coastal fishes is, at least partly, shaped by human activities.

Our regional fish biodiversity estimates obtained with eDNA are close to those of several available regional checklists (Allen & Erdmann, 2009; Coll et al. 2010; Ronald Fricke, Kulbicki, & Wantiez, 2011), and to local or regional biodiversity estimations with classical methods (Friedlander et al. 2020; Johannesen et al. 2021; Polanco et al. 2021; Siu et al. 2017). Yet, we detected species that are rarely reported in classical inventories like the Greenland shark (Somniosus microcephalus) in the Arctic, the elegant firefish (Nemateleotris decora) in the Coral Triangle, or the Antarctic escolar (Paradiplospinus antarcticus). We also find strong significant correlations between the number of MOTUs and the number of species belonging to each family in the checklists (Supporting Information Figure S21). Our diversity estimates confirm the ability of eDNA metabarcoding and of the teleo primer, to recover regional fish diversity with few samples (Juhel et al. 2020; Marques et al. 2021; Polanco et al. 2021).

MOTU diversity patterns display the expected fish species richness gradient from the Coral Triangle to the Caribbean (Bellwood & Hughes, 2001; Parravicini et al. 2013) and the poles (Freeman & Pennell, 2021). The sequence α-diversity, capturing species genetic relatedness, follows the same gradients. The lowest sequence α-diversity is observed at the poles, where most fish MOTUs are close relatives, the highest diversity lies in the tropics, where MOTU richness was the highest with a large number of families and genera (Juhel et al. 2020), while temperate regions present intermediate values (Figure 1B). Some stations in the Yellow Sea and polar regions, where the MOTU richness is the lowest, reveals intermediate sequence α-diversity due to very distinct genera (i.e. Lycodes, Liparis and Somniosus in the Arctic).

The increase in MOTU and sequence α-diversity of marine fishes with increasing temperature can be related to the ‘evolutionary speed’ and ‘climate stability’ hypotheses (Manel et al. 2020). According to the ‘evolutionary speed’ hypothesis (Fine, 2015), high temperatures promote metabolic, reproduction and speciation rates, and ultimately increase species richness (Harmelin-Vivien, 2002), yet fish richness gradients worldwide do not correlate with recent speciation rates (Rabosky et al. 2018). The ‘climate stability’ hypothesis posits that warmer areas in the tropics have experienced less historical variability in climatic conditions, whereas colder areas were highly unstable, leading to species diversity declines along temperature gradients (Mittelbach et al. 2007; Pellissier et al. 2014). Owing to environmental niche conservatism (Gaboriau et al. 2019), closely related species or entire lineages can be extirpated under climatic filtering while only a limited number of lineages are adapted to the extreme climatic conditions of the poles (Mittelbach et al. 2007). This second ‘climatic stability’ hypothesis is better supported by our fish MOTU and sequence α-diversity patterns.

Fish biodiversity is also related to human pressures, as indicated by the significant decrease of MOTU and sequence α-diversity in response to countries' dependence on marine resources (Figure 2). Human populations highly depending on marine resources for food and incomes may use non-selective fishing gears catching even small or cryptobenthic species (Batista, Fabré, Malhado, & Ladle, 2014; Munro, 1996) leading to a deterioration of key habitats such as coral reefs. Indeed, artisanal fishing methods such as blast, trawling or poison can destroy the habitat and affect the whole trophic spectrum (Fox, Pet, Dahuri, & Caldwell, 2003). Fishing pressure could thus impact all fish size classes in those countries highly dependent on marine resources, removing entire parts of the food web, and decreasing the genetic diversity of the remaining species pool. More specifically, the diversity of large-bodied fishes is negatively associated to human pressure (Figure 2). Species abundance and richness of sharks, jacks, groupers and snappers are known to decrease with increasing human population density (Dulvy et al. 2021), affecting top down control in overexploited ecosystems (McClure, Hoey, Sievers, Abesamis, & Russ, 2020). As human pressures, such as release of nutrient and chemicals and habitat degradation, affect the most sensitive species, often belonging to the same evolutionary lineages or families (Cinner et al. 2018; Dulvy et al. 2021), we observe an overall decrease of fish diversity with increasing human pressure (Figure 3) and the remaining species may be more closely related, decreasing even more sequence α-diversity than MOTU diversity as hypothesized.

Geography is also significantly related to the distribution of fish eDNA sequences in the Antarctic, with MOTU composition and sequence β-diversity differing from all other regions (Figure 5). This result is coherent with previous studies showing that the Antarctic region was progressively isolated and cooled during the Cenozoic period by the opening of the Drake Passage and the development of the Antarctic Circumpolar Current (Crame, 2018). Antarctic marine fauna is therefore evolutionarily isolated and dominated by a few families of highly specialized benthic fishes. Indeed, of the 22 MOTUs identified in our Antarctic samples, 15 were unique to this region, among which Nototheniidae, Zoarcidae and Liparidae species, and the endemic Antarctic silverfish (Pleuragramma Antarctica). Most of the known fish fauna in Antarctica comprises benthic species from these three families, with 97% of endemism among the Nototheniidae (Eastman, 2005).

Arctic fauna is similar to temperate faunas in terms of fish MOTUs composition but distinct in terms of sequence β-diversity, which is also expected (Bluhm et al. 2011; Fauchald et al. 2021). A few fish MOTUs are shared between the Arctic and Atlantic, which can be explained

by the connectivity between these regions due to currents from the Atlantic and the Pacific flowing into the Arctic Ocean, and species range shifts due to the ongoing Atlantification occurring in the Arctic (McLean, Mouillot, et al. 2021). The Arctic also contains some very distinct MOTUs belonging to Anarhichadidae, Cyclopteridae, Stichaeidae, Somniosidae or Zoarcidae for example, of which several are adapted to cold waters or are endemic to the Arctic (Bluhm et al. 2011).

The Caribbean and East Pacific faunas also show distinct fish MOTU compositions compared to other tropical regions but similar sequence β-diversity patterns. The unique fauna composition of the Caribbean is well-known and explained by a strong geographic barrier (i.e. Isthmus of Panama) and a limited suitable area for coral reefs during the past quaternary glaciation (Bender et al. 2017; Pellissier et al. 2014). Our study reveals a low sequence β-diversity between stations of tropical regions, suggesting that the high fish MOTU dissimilarity between these regions is due to close relative species. In other words, most fish species in the Caribbean and other tropical regions belong to the same evolutionary lineages.

The pairwise species genetic distances, computed only for the MOTUs assigned at the species level (~30% of the dataset), were positively but weakly correlated to the phylogenetic and functional pairwise distances for these same species pairs. On such a short barcode (~60bp), several species can have the same sequence (Polanco et al. 2021) so a null pairwise genetic distance while they diverge in terms of phylogeny and traits. In contrast, a few species may show intraspecific variability, resulting in a positive genetic distance within the same species. However, this intraspecific signal is likely very low since the “teleo” sequence that we use is small and not variable enough to be a good marker for estimating intraspecific variability (Marques, Guérin, et al. 2020). Our results reveal, however, that the overall sequence α- and β-diversity computed from these pairwise distances are significantly correlated to the corresponding phylogenetic and, to a lesser extent, functional α- and β-diversity. Therefore, sequence α- and β-diversity within and between stations represent good proxies for phylogenetic and functional α- and β-diversity, which is not trivial given the length of our barcode. This new finding offers major perspectives for the use of short eDNA barcodes with low production costs to quantify phylogenetic and functional diversity and thus monitor ecosystem functioning (Duffy, Lefcheck, Stuart-Smith, Navarrete, & Edgar, 2016), evolutionary history (McLean, Stuart-Smith, et al. 2021), or environmental and human impacts (Trindade-Santos, Moyes, & Magurran, 2020). A longer barcode would bring even more

reliable information on the genetic structure and, therefore, evolutionary lineages composing fish assemblages, but may detect less species.

Since our study covers an extensive spatial scale, there are some limitations in terms of data collection or analysis. Our sampling design was not balanced among regions (Supporting Information Table S1), which can affect biodiversity estimates. We found a positive relationship between the number of detected MOTUs and the volume of seawater filtered indicating a species-area relationship (Figure 2). The models used in our study also find a significant effect of the sampling method. Sampling along transects retrieves higher MOTU and sequence diversity than sampling on points (Supporting Information Figure S22). Ideally, sampling would be standardized across regions. However, including the sampling information in our models ensures that, at least partly, these differences are taken into account when estimating the importance of other factors.

Due to the incompleteness of available reference databases at the global scale (Marques, Milhau, et al. 2020), the assignment to the species level remains impossible for more than 70% of sequences. For this reason, we used MOTUs curated by a conservative bioinformatic pipeline (Marques, Guérin, et al. 2020). Non-curated MOTUs often overestimate real diversity (Brandt et al. 2021) since a given MOTU can represent several species within one cluster or several MOTUs can correspond to the same species (Polanco et al. 2021). A conservative curation of MOTUs better reflects the true level of fish diversity (Marques, Guérin, et al. 2020; Sales et al. 2021) by decreasing the number of MOTUs representing the same taxa. Our conservative MOTU pipeline may, however, underestimate fish diversity of some cryptobenthic or rare fish groups that are more poorly represented in public databases, and of families with low taxonomic resolution (Supporting Information Figure S23). Yet, several studies in species-rich marine ecosystems, using the same barcode and the same MOTU pipeline, provide diversity estimates similar to those obtained with traditional methods (Juhel et al. 2020; Polanco et al. 2021). Using multiple markers could be an alternative to overcome the incompleteness of genetic reference database and the lack of primer resolution (Polanco et al. 2021; Ruppert, Kline, & Rahman, 2019), but this approach would be much more expensive. Improving the accuracy of taxonomic assignment and completing genetic reference databases are thus urgently needed to improve estimates of large-scale biodiversity patterns (Marques, Milhau, et al. 2020) and local monitoring.

From an extensive eDNA survey from pole to pole across all oceans, our study associates the global distribution of eDNA sequences released by coastal fishes to environmental,

geographic and human factors. Our study reveals that eDNA metabarcoding, beyond providing reliable fish diversity estimations and distributions at large spatial scale, despite incomplete genetic reference databases, can inform on the relative correlates of environmental, geographic and human factors on fish diversity, including crypto-benthic, rare and elusive species. As expected, the environment shows the strongest relation to fish biodiversity, but human activities are also at play. Fish sequence diversity, reflecting species relatedness, strongly decreases with human pressures suggesting a strong environmental but also human filtering on coastal ecosystems. Furthermore, our study highlights that sequence diversity from eDNA metabarcoding is a robust indicator of human impact and a reliable proxy of phylogenetic and functional diversity which are essential to ecosystem functioning. We recommend that eDNA monitoring should be considered in future conservation management plans.

References

Alberdi, A., & Gilbert, M. T. P. (2019). A guide to the application of Hill numbers to DNA-based diversity analyses. Molecular Ecology Resources, 19(4), 804–817.

Allen, G. R., & Erdmann, M. V. (2009). Reef fishes of the Bird's Head Peninsula, West Papua, Indonesia. Check List, 5(3), 587–628.

Andrello, M., Darling, E. S., Wenger, A., Suárez-Castro, A. F., Gelfand, S., & Ahmadia, G. N. (2022). A global map of human pressures on tropical coral reefs. Conservation Letters, 15(e12858), 1–12.

Batista, V. S., Fabré, N. N., Malhado, A. C. ., & Ladle, R. J. (2014). Tropical Artisanal Coastal Fisheries: Challenges and Future Directions. Reviews in Fisheries Science and Aquaculture, 22(1), 1–15.

Bellwood, D. R., & Hughes, T. P. (2001). Regional-scale assembly rules and biodiversity of coral reefs. Science, 292(5521), 1532–1534.

Bender, M. G., Leprieur, F., Mouillot, D., Kulbicki, M., Parravicini, V., Pie, M. R., … Floeter, S. R. (2017). Isolation drives taxonomic and functional nestedness in tropical reef fish faunas. Ecography, 40(3), 425–435.

Bluhm, B., Gebruk, A., Gradinger, R., Hopcroft, R., Huettmann, F., Kosobokova, K., … Weslawski, M. (2011). Arctic Marine Biodiversity: An Update of Species Richness and Examples of Biodiversity Change. Oceanography, 24(3), 232–248.

Boulanger, E., Loiseau, N., Valentini, A., Arnal, V., Boissery, P., Dejean, T., … Mouillot, D. (2021). Environmental DNA metabarcoding reveals and unpacks a biodiversity conservation paradox in Mediterranean marine reserves. Proceedings of the Royal Society B, 288(20210112), 1–10.

Boussarie, G., Kiszka, J. J., Mouillot, D., Bonnin, L., Manel, S., Kulbicki, M., … Mariani, S. (2018). Environmental DNA illuminates the dark diversity of sharks. Science Advances, 4(5), 1–8.

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.

Brandl, S. J., Goatley, C. H. R., Bellwood, D. R., & Tornabene, L. (2018). The hidden half : ecology and evolution of cryptobenthic fishes on coral reefs. Biological Reviews, 93, 1846–1873.

Brandt, M. I., Trouche, B., Quintric, L., Günther, B., Wincker, P., Poulain, J., & Arnaud-Haond, S. (2021). Bioinformatic pipelines combining denoising and clustering tools allow for more comprehensive prokaryotic and eukaryotic metabarcoding. Molecular Ecology Resources, 21(6), 1904–1921.

Chao, A., Chiu, C. H., Villéger, S., Sun, I. F., Thorn, S., Lin, Y. C., … Sherwin, W. B. (2019). An attribute-diversity approach to functional diversity, functional beta diversity, and related (dis)similarity measures. Ecological Monographs, 89(2), 1–29.

Cinner, J. E., Maire, E., Huchery, C., MacNeil, M. A., Graham, N. A. J., Mora, C., … Mouillot, D. (2018). Gravity of human impacts mediates coral reef conservation gains. Proceedings of the National Academy of Science, 115(27), 6116–6125. Retrieved from

Coll, M., Piroddi, C., Steenbeek, J., Kaschner, K., Lasram, F. B. R., Aguzzi, J., … Voultsiadou, E. (2010). The biodiversity of the Mediterranean Sea: Estimates, patterns, and threats. PLoS ONE, 5(8).

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.

Cowman, P. F., & Bellwood, D. R. (2013). The historical biogeography of coral reef fishes: Global patterns of origination and dispersal. Journal of Biogeography, 40(2), 209–224.

Crame, J. A. (2018). Key stages in the evolution of the Antarctic marine fauna. Journal of Biogeography, 45(5), 986–994.

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.

Dulvy, N. K., Pacoureau, N., Rigby, C. L., Hilton-taylor, C., Fordham, S. V, & Simpfendorfer, C. A. (2021). Overfishing drives over one-third of all sharks and rays toward a global extinction crisis. Current Biology, 31(21), 4773–4787.

Eastman, J. T. (2005). The nature of the diversity of Antarctic fishes. Polar Biology, 28(2), 93–107.

Eddy, T. D., Lam, V. W. Y., Reygondeau, G., Cisneros-Montemayor, A. M., Greer, K., Palomares, M.-L. D., … Cheung, W. W. L. (2021). Global decline in capacity of coral reefs to provides ecosystem services. One Earth, 4, 1278–1285.

Edgar, G. J., & Stuart-Smith, R. D. (2014). Systematic global assessment of reef fish communities by the Reef Life Survey program. Scientific Data, 1, 1–8.

Fauchald, P., Arneberg, P., Debernard, J. B., Lind, S., Olsen, E., & Hausner, V. H. (2021). Poleward shifts in marine fisheries under Arctic warming. Environmental Research Letters, 16(7).

Fine, P. V. A. (2015). Ecological and Evolutionary Drivers of Geographic Variation in Species Diversity. Annual Review of Ecology, Evolution, and Systematics, 46(October), 369–392.

Fox, H. E., Pet, J. S., Dahuri, R., & Caldwell, R. L. (2003). Recovery in rubble fields: Long-term impacts of blast fishing. Marine Pollution Bulletin, 46(8), 1024–1031.

Freeman, B. G., & Pennell, M. W. (2021). The latitudinal taxonomy gradient. Trends in Ecology and Evolution, 36(9), 778–786.

Fricke, R, Durville, P., Bernardi, G., Borsa, P., Mou-Tham, G., & Chabanet, P. (2013). Checklist of the shore fishes of Europa Island, Mozambique Channel, southwestern Indian Ocean, including 302 new records. Stuttgarter Beiträge Zur Naturkunde A, Neue Serie 6, (1952), 247–276.

Fricke, Ronald, Kulbicki, M., & Wantiez, L. (2011). Checklist of the fishes of New Caledonia, and their distribution in the Southwest Pacific Ocean (Pisces). Stuttgarter Beitrage Zur Naturkunde, Serie A (Biologie), 2011(4), 341–463.

Friedlander, A. M., Goodell, W., Salinas-De-León, P., Ballesteros, E., Berkenpas, E., Capurro, A. P. … Sala, E. (2020). Spatial patterns of continental shelf faunal community structure along the Western Antarctic Peninsula. PLoS ONE, 15(10 October), 1–19.

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).

Gaboriau, T., Albouy, C., Descombes, P., Mouillot, D., Pellissier, L., & Leprieur, F. (2019). Ecological constraints coupled with deep- time habitat dynamics predict the latitudinal diversity gradient in reef fishes. Proceedings of the Royal Society B, (286).

Harmelin-Vivien, M. L. (2002). Energetics and fish diversity. In The Ecology of Fishes on Coral Reefs (pp. 265-274.).

Johannesen, E., Wienerroither, R., Mørk, H. L., Husson, B., Holmin, A. J., Johnsen, E., … Prokhorova, T. (2021). Fish diversity data from the Barents Sea Ecosystem Survey 2004-2019. In Rapport fra havforskningen.

Juhel, J. B., 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, 287(20200248), 1–10.

Kumar, G., Reaume, A. M., Farrell, E., & Gaither, M. R. (2022). Comparing eDNA metabarcoding primers for assessing fish communities in a biodiverse estuary. PLoS ONE, 17(6), 1–20.

Lenoir, J., Bertrand, R., Comte, L., & … L. B. (2020). Species better track climate warming in the oceans than on land. Nature Ecology & Evolution, 4, 1044–1059.

Leprieur, F., Descombes, P., Gaboriau, T., Cowman, P. F., Parravicini, V., Kulbicki, M., … Pellissier, L. (2016). Plate tectonics drive tropical reef biodiversity dynamics. Nature Communications, 7(May), 1–8.

Loiseau, N., Thuiller, W., Stuart-smith, R. D., Devictor, V., Edgar, G. J., Velez, L., … Mouillot, D. (2021). Maximizing regional biodiversity requires a mosaic of protection levels. PLOS Biology, 19(5), 1–18.

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.

Mächler, E., Walser, J., & Altermatt, F. (2021). Decision making and best practices for taxonomy-free eDNA metabarcoding in biomonitoring using Hill numbers. Molecular Ecology, 30, 3326–3339.

Magneville, C., Loiseau, N., Albouy, C., Casajus, N., Claverie, T., Escalas, A., … Villéger, S. (2022). mFD: an R package to compute and illustrate the multiple facets of functional diversity. Ecography, 1, 1–15.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. PeerJ, 3, 1–12.

Manel, S., Guerin, P. E., Mouillot, D., Blanchet, S., Velez, L., Albouy, C., & Pellissier, L. (2020). Global determinants of freshwater and marine fish genetic diversity. Nature Communications, 11(1), 1–9.

Marques, V., Castagné, P., Polanco, A., Borrero-Pérez, G. H., Hocdé, R., Guérin, P. É., … Villéger, S. (2021). Use of environmental DNA in assessment of fish functional and phylogenetic diversity. Conservation Biology, (May), 1–13.

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, 43, 1–12.

Marques, V., Milhau, T., Albouy, C., Dejean, T., Manel, S., Mouillot, D., & Juhel, J. (2020). GAPeDNA : Assessing and mapping global species gaps in genetic databases for eDNA metabarcoding. Conservation Biology, 27(10), 1880–1892.

Martin., M. (1994). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal, 17(1), 10–12.

Mathon, L., Marques, V., Mouillot, D., Albouy, C., Andrello, M., Baletaud, F., … Vigliola, L. (2022). Cross-ocean patterns and processes in fish biodiversity on coral reefs through the lens of eDNA metabarcoding. Proceedings of the Royal Society B, 289, 20220162.

McClure, E. C., Hoey, A. S., Sievers, K. T., Abesamis, R. A., & Russ, G. R. (2020). Relative influence of environmental factors and fishing on coral reef fish assemblages. Conservation Biology, 0(0), 1–14.

McLean, M., Mouillot, D., Maureaud, A. A., Hattab, T., MacNeil, M. A., Goberville, E., … Auber, A. (2021). Disentangling tropicalization and deborealization in marine ecosystems under climate change. Current Biology, 31(21), 4817-4823.e5.

McLean, M., Stuart-Smith, R. D., Villéger, S., Auber, A., Edgar, G. J., MacNeil, M. A., … Mouillot, D. (2021). Trait similarity in reef fish faunas across the world's oceans. Proceedings of the National Academy of Science, 118(12), e2012318118.

Mittelbach, G. G., Schemske, D. W., Cornell, H. V., Allen, A. P., Brown, J. M., Bush, M. B., … Turelli, M. (2007). Evolution and the latitudinal diversity gradient: Speciation, extinction and biogeography. Ecology Letters, 10(4), 315–331.

Miya, M. (2021). Environmental DNA Metabarcoding: A Novel Method for Biodiversity Monitoring of Marine Fish Communities. Annual Review of Marine Science, Vol 8, 14(6), 1–25.

Munro, J. L. (1996). The scope of tropical reef fisheries and their management. Reef Fisheries, 1–14.

O'Hara, C. C., Frazier, M., & Halpern, B. S. (2021). At-risk marine biodiversity faces extensive, expanding, and intensifying human impacts. Science, 372(6537), 84–87.

Pacoureau, N., Rigby, C. L., Kyne, P. M., Sherley, R. B., Winker, H., Carlson, J. K., … Dulvy, N. K. (2021). Half a century of global decline in oceanic sharks and rays. Nature, 589(7843), 567–571.

Parravicini, V., Bender, M. G., Villéger, S., Leprieur, F., Pellissier, L., Donati, F. G. A., … Kulbicki, M. (2021). Coral reef fishes reveal strong divergence in the prevalence of traits along the global diversity gradient. Proceedings of the Royal Society B: Biological Sciences, 288(1961), 1–7.

Parravicini, V., Kulbicki, M., Bellwood, D. R., Friedlander, A. M., Arias-Gonzalez, J. E., Chabanet, P., … Mouillot, D. (2013). Global patterns and predictors of tropical reef fish species richness. Ecography, 36(12), 1254–1262.

Pecuchet, L., Törnroos, A., & Lindegren, M. (2016). Patterns and drivers of fish community assembly in a large marine ecosystem. Marine Ecology Progress Series, 546, 239–248.

Pellissier, L., Leprieur, F., Parravicini, V., Cowman, P. F., Kulbicki, M., Litsios, G., … Mouillot, D. (2014). Quaternary coral reef refugia preserved fish diversity. Science, 344(6187), 1016–1020.

Polanco, A., Richards, F. E., Flück, B., Valentini, A., Altermatt, F., Jean-, S. B., … Pellissier, L. (2021). Comparing the performance of 12S mitochondrial primers for fish environmental DNA across ecosystems. Environmental DNA, 3(6), 1113–1127.

Polanco Fernández, A., Marques, V., Fopp, F., Juhel, J., Borrero-Pérez, G. H., Cheutin, M., … Pellissier, L. (2021). Comparing environmental DNA metabarcoding and underwater visual census to monitor tropical reef fishes. Environmental DNA, 3(1), 142–156.

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.

Rabosky, D. L., Chang, J., Title, P. O., Cowman, P. F., Sallan, L., Friedman, M., … Alfaro, M. E. (2018). An inverse latitudinal gradient in speciation rate for marine fishes. Nature, 559(7714), 392–395.

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. PeerJ, 4, 1–22.

Ruppert, K. M., Kline, R. J., & Rahman, M. 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.

Sales, N. G., Wangensteen, O. S., Carvalho, D. C., Deiner, K., Præbel, K., Coscia, I., … Mariani, S. (2021). Space-time dynamics in monitoring neotropical fish communities using eDNA metabarcoding. Science of the Total Environment, 754, 142096.

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.

Selig, E. R., Hole, D. G., Allison, E. H., Arkema, K. K., McKinnon, M. C., Chu, J., … Zvoleff, A. (2019). Mapping global human dependence on marine ecosystems. Conservation Letters, 12(2), 1–10.

Siu, G., Bacchet, P., Bernardi, G., Brooks, A. J., Carlot, J., Causse, R., … Galzin, R. (2017). Shore fishes of French Polynesia. Cybium, 41(3), 245–278.

Stuart-Smith, R. D., Mellin, C., Bates, A. E., & Edgar, G. J. (2021). Habitat loss and range shifts contribute to ecological generalization among reef fishes. Nature Ecology and Evolution, 5(5), 656–662.

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

Trindade-Santos, I., Moyes, F., & Magurran, A. E. (2020). Global change in the functional diversity of marine fisheries exploitation over the past 65 years: Fisheries and Functional Diversity. Proceedings of the Royal Society B: Biological Sciences, 287(1933).

Valdivia-Carrillo, T., Rocha-Olivares, A., Reyes-Bonilla, H., Domínguez-Contreras, J. F., & Munguia-Vega, A. (2021). Integrating eDNA metabarcoding and simultaneous underwater visual surveys to describe complex fish communities in a marine biodiversity hotspot. Molecular Ecology Resources, (March), 1558–1574.

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.

West, K., Travers, M. J., Stat, M., Harvey, E. S., Richards, Z. T., Dibattista, J. D., … Bunce, M. (2021). Large-scale eDNA metabarcoding survey reveals marine biogeographic break and transitions over tropical north- western Australia. Diversity and Distributions, (00), 1–16.

Yan, H. F., Kyne, P. M., Jabado, R. W., Leeney, R. H., Davidson, N. K., Derrick, D. H., … Dulvy, N. K. (2021). Overfishing and habitat loss drives range contraction of iconic marine fishes to near extinction. Science Advances, 7, 1–10.

Zhang, S., Zhao, J., & Yao, M. (2020). A comprehensive and comparative evaluation of primers for metabarcoding eDNA from fish. Methods in Ecology and Evolution, 2020(January), 1609–1625.

Zinke, J., Gilmour, J., Fisher, R., Puotinen, M., Maina, J., Darling, E. S., … Wilson, S. K. (2018). Gradients of disturbance and environmental conditions shape coral community structure for south-eastern Indian Ocean reefs. Diversity and Distributions, 24, 65–620.

Chapitre 5 - Modélisation tridimensionnelle de la biodiversité et implications pour la conservation

Navire océanographique Alis, en mer de Corail lors des campagnes SEAMOUNTS (crédit : F. Baletaud)

1. Préface

Le chapitre précédent montre que la diversité locale et globale des poissons est impactée par l'activité humaine. Il est donc important de protéger cette diversité. Un des outils les plus efficaces pour la protection de la biodiversité et des habitats est la mise en place d'aires marines protégées (Edgar et al. 2014). Cependant, la localisation de ces réserves ou AMP doit prendre en compte plusieurs aspects de la diversité (richesse, abondance, biomasse) et également l'activité humaine, pour que leur efficacité soit optimale. En effet, Cinner et al. (2018) ont montré que dans les réserves implémentées dans des zones fortement impactées par l'homme, l'abondance des tops prédateurs et la biomasse n'étaient pas aussi élevée que dans les réserves plus éloignées. Il est donc important d'étudier la distribution de l'abondance, la biomasse et la richesse des poissons, ainsi que leurs réponses aux impacts humains, afin de définir les zones à protéger en priorité. Les océans sont vastes et profonds, or la majorité des aires marines protégées se trouvent en zones côtières peu profondes et n'intègrent pas de dimension verticale (Zhao et al. 2020). Dans l'époque actuelle où le changement climatique s'intensifie et menace les profondeurs des océans (Brito-Morales et al. 2020; Ariza et al. 2022) et où les pressions humaines dégradent toujours plus la diversité des océans et des profondeurs, notamment par la pêche profonde et bientôt les prospections minières (Lins et al. 2021; Long et al. 2021; Good et al. 2022), il semble important et urgent d'étudier la diversité de ces zones profondes et de les considérer dans les plans de conservation (Epstein & Roberts, 2022). L'idée de la nécessité d'intégrer une troisième dimension dans les plans de conservation commence à s'imposer dans la littérature (Venegas-Li et al. 2018; Manea et al. 2020; Doxa et al. 2022), mais encore trop peu d'études se sont intéressées à la biodiversité dans les zones profondes des océans, telles que les plaines abyssales, les monts sous-marins ou les pentes profondes des îles océaniques. Les quelques études sur la diversité sur les monts sous-marins ont pourtant démontré qu'il s'agissait d'habitats importants pour les mammifères marins, les prédateurs et la faune sessile benthique (Rogers, 1994; Morato et al. 2016; Watling and Auster, 2017; Letessier et al. 2019), et que les communautés de poissons y étaient fortement structurées en fonction de la profondeur (McClain & Lundsten, 2015; Muff et al. 2022; Proud et al. 2018), mais davantage d'investigations sont nécessaires.

Dans ce chapitre j'étudie la diversité sur 11 monts sous-marins et 4 pentes externes profondes dans l'archipel de Nouvelle-Calédonie, en combinant l'ADNe, les caméras appâtées et l'acoustique. Je mesure la richesse spécifique, l'abondance et la biomasse de l'ensemble des poissons et de certaines espèces les plus fréquentes, sur les sommets des monts sous-marins et

pentes externes, ainsi que dans la zone pélagique avoisinante, jusqu'à 600m de profondeur. Je modélise ensuite ces métriques en fonction de facteurs géographiques, environnementaux et humains, puis je les prédis sur l'ensemble de l'archipel Néo-Calédonien. Ces prédictions sont ensuite découpées en trois couches de profondeur (0-200m, 200-400m et 400-600m) et entrées dans un algorithme de priorisation spatiale en trois dimensions, cherchant à protéger 30% de chaque métrique de diversité et 30% de l'abondance des espèces les plus fréquentes dans chaque habitat.

Les résultats indiquent un fort effet de la profondeur, de la température et de la distance à l'homme sur l'abondance, la richesse et la biomasse des poissons. La diversité prédite est ainsi plus élevée sur les pentes externes des atolls et îles les plus éloignées de Nouméa et sur les monts sous-marins dont le sommet est peu profond. Le scénario de priorisation autorisant la plus grande fragmentation sélectionne de nombreuses zones sur les pentes externes des atolls et îles éloignés de la Grande-Terre, sur les sommets et pentes des monts sous-marins peu profonds, mais aussi des zones sur les sommets et pentes profonds. Près de la moitié des cellules sont priorisées sur les trois couches de profondeur (44%), mais 24% des cellules sont protégées seulement dans la couche la moins profonde, 8% seulement dans la couche de profondeur intermédiaire et 10% seulement dans la couche la plus profonde. Cette méthodologie permet de démontrer que certaines zones hébergent une forte diversité seulement à de grande profondeurs, et qu'il est donc important d'inclure une troisième dimension, la profondeur, dans les plans de conservation. De plus, en protégeant 30% de chaque métrique de diversité, la meilleure solution de planification protège une surface correspondant à 30% de la surface considérée dans notre étude. Ainsi, en sélectionnant les zones à protéger en se basant sur les indices de diversité il est possible d'atteindre les objectifs internationaux de protection des océans (CBD, 2021; Dinerstein et al., 2019).

2. Manuscrit D

3D conservation planning of multiple marine fish biodiversity metrics reveals 30x30 CBD target in the deep Coral Sea

Running title: 3D conservation planning of deep Coral Sea fish biodiversity

Laetitia Mathon[1,2], Florian Baletaud[2,3,4], Anne Lebourges-Dhaussy[5], Gaël Lecellier[2], Christophe Menkes[2], Céline Bachelier[6], Claire Bonneville[2], Tony Dejean[7], Mahé Dumas[2], Sylvie Fiat[2], Jacques Grelet[8], Jérémie Habasque[5], Stéphanie Manel[1], Laura Mannocci[3], David Mouillot[3], Maëlis Peran[2], Gildas Roudaut[5], Christine Sidobre[2], David Varillon[6], Laurent Vigliola[2]

Affiliations:

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

[2] ENTROPIE, IRD, CNRS, Ifremer, Université de la Réunion, Université de la Nouvelle-Calédonie, Nouméa, New Caledonia, France

[3] MARBEC, Univ. Montpellier, CNRS, Ifremer, IRD, Montpellier, France

[4] Soproner, groupe GINGER, 98000 Nouméa, New Caledonia

[5] LEMAR, UBO, CNRS, IRD, Ifremer, Plouzané, France.

[6] IMAGO, IRD, Nouméa, New Caledonia

[7] SPYGEN, Le Bourget-du-Lac

[8] IRD DR-OUEST, US191 IMAGO, Technopole de Brest-Iroise—Site de la Pointe du Diable, Plouzané, France

Keywords: eDNA, BRUVS, acoustic, 3D conservation planning, seamounts

Corresponding author: Laetitia Mathon, laetitia.mathon@gmail.com, CEFE, Montpellier, France.

Abstract

Accelerating rate of human impact and environmental change severely affect marine biodiversity and increase the urgency to implement the 30x30 CBD plan for conserving 30% of biodiversity by 2030. However, area-based conservation targets are complex to define in a 3-dimensional ocean where deep-sea features such as seamounts have been seldom studied. To fill this gap, we collected environmental DNA, acoustics, and video data through the water column to study fish biodiversity at 15 seamounts and deep island slopes across the Coral Sea. We modelled 7 fish community metrics (*i.e.* species richness, biomass, abundance, MOTUs richness) and 45 individual species and MOTUs abundances in benthic and pelagic waters, down to 600m deep, using boosted regression trees (BRTs) and generalized joint attribute models (GJAMs). We predicted these 52 metrics on other seamounts and deep slopes across the New-Caledonian marine waters. Finally, we prioritized conservation units (1x1km and 200m deep) in a three-dimensional space, toward the goal of protecting at least 30% of each metric. Depth, temperature and remoteness from fish market influenced most fish communities. Different compactness parameters led to different conservation planning scenarios, with the best solution allowing to preserve 30% of the fish biodiversity in ~30% of the considered spatial domain while accounting for the 3D distribution of biodiversity and offering relative compactness to implement management plans. Our study paves the way for a new methodology in the selection of conservation areas in 3D structured environments, taking depth and its biodiversity into account while providing solutions compatible with area-based targets agreed by the international community.

Introduction

Human disturbances and climate change strongly affect biodiversity worldwide (Andrello et al., 2022; Yan et al., 2021) with severe impacts on the well-being, food security and socioeconomic situation of billions of people in the world (Eddy et al., 2021). This biodiversity crisis is particularly acute in the oceans (Dulvy et al., 2017; Pacoureau et al., 2021), including on seamounts and other deep-sea environments where overfishing, destructive fishing, marine mining, pollution and changes in physical and chemical properties of water increasingly threaten species and associated ecosystem services (Rogers, 2018). To safeguard ecosystems, a key response of governments, through the Convention on Biological Diversity (CBD), is to set area-based conservation targets, i.e., the “30x30” call for protecting 30 percent of sea areas by

2030 (CBD, 2021). However, achieving these targets seems challenging in the deep-seas, where more than 99% of the environments remain unexplored by science (Rogers, 2018). This makes it difficult to prioritize conservation areas and allows these environments be threatened without even realizing it. The three-dimensional distribution of biodiversity in the deep oceans further complicates the achievement of area-based conservation objectives, which is not the case on shallow marine ecosystems such as coral reefs where protection can reasonably cover the full water column. While deep-sea environments cover an immense area, seamounts alone spanning a territory as vast as Europe (Kvile et al., 2014), they are virtually left unprotected (Letessier et al., 2019; Yesson, Clark, Taylor, & Rogers, 2011). We thus face the urgent need to document the 3D distribution of marine biodiversity across depths in order to set up appropriate conservation measures including the underexplored deep sea.

Seamounts are ubiquitous features of the deep-sea, and among the least studied habitats in the ocean. More than 170,000 seamounts have been identified worldwide, of which 33,000 are greater than 1000m in height (Yesson et al., 2011), but only 0,002% of world's seamounts have been sampled for scientific purposes (Rogers, 2018). Very little is known about fish biodiversity on seamounts, and no studies have yet compared seamounts and other deep-sea habitats such as continental or island slopes. Although being hotspots and refuges for many different taxa (Letessier et al., 2019; Morato et al., 2010; Morato et al., 2008), seamounts are presently under-represented within no-take marine protected areas (MPAs), with only 2% of the world seamounts included in the global MPA network (Letessier et al., 2019; Yesson et al., 2011). Seamounts are however highly targeted by fisheries (Kerry et al., 2022; Pitcher et al., 2010), implying particular concern for fish biodiversity.

Biodiversity metrics such as species richness, abundance or biomass are all important indicators of ecosystem health, but their distribution may vary across space and time (Gust et al., 2001; Maureaud et al., 2019) and may respond differently to threats (Lin et al., 2021; Pacoureau et al., 2021). In order to efficiently protect fish biodiversity in depth-structured ecosystems, such as seamounts, island slopes and pelagic waters, it is therefore essential to model several diversity metrics and implement conservation planning in three dimensions, taking depth into account (Dambach & Rödder, 2011; Duffy & Chown, 2017). No-take MPAs are the most efficient tools to protect many aspects of biodiversity (Giakoumi et al., 2017; McClanahan, 2021; Sala et al., 2018). Most MPAs are implemented in national coastal waters, without explicit consideration of the different depth domains (Levin et al., 2018). Such two-dimensional spatial prioritization schemes may fail to protect species habitats and refugia to

climate change (Brito-Morales et al., 2022; Doxa et al., 2022). Some recent studies have advocated the need to incorporate a third dimension to inform conservation actions (Levin et al., 2018; Manea et al., 2020; Venegas-Li et al., 2018), enabling both vertical and horizontal spatial prioritization simultaneously.

Until now, the main obstacle to 3D conservation planning has been the lack of comparable data along the depth gradient from the surface to the abyss, a problem that can be overcome thanks to recent technological advances. Environmental DNA (eDNA) metabarcoding, video and echosounder surveys allow quantitative data to be collected in a standardized way at any depth. The metabarcoding of eDNA is based on the retrieval and analysis of genetic material naturally released by organisms in their environments. It was recently shown to outperform dive and video surveys for estimating marine biodiversity (Mathon et al., 2022), and to be efficient to study fish assemblages on seamounts (Muff et al., 2022). Yet, the drawback of eDNA metabarcoding is the lack of knowledge about organism size, abundance and biomass, and the relative narrowness of the sampled surface with barely a few liters of filtered water in the vast ocean. Stereo baited remote underwater video stations (BRUVS) can efficiently estimate species abundance and biomass on any marine habitat (Langlois et al., 2020), and acoustic echosounders can estimate fish biomass continuously across vast oceanic areas (Proud et al., 2018). These three methods seem complementary for the survey of marine biodiversity and prioritization of its conservation in 3D.

In this study, we first collect eDNA, BRUVS and acoustic echosounder data on the benthic and pelagic habitats of 15 seamounts and island slopes, down to 600m depth, across the vast (1.4 million km2) archipelago of New-Caledonia, a South Pacific biodiversity hotspot listed as a UNESCO World Heritage Site since 2008. Using this unique and complementary dataset, we model and predict fish community metrics (richness, abundance and biomass) and individual species abundances in three-dimensions on island slopes, seamounts and pelagic surroundings, using boosted regression trees (BRTs), taking complex non-linear interactions into account (Elith et al., 2008) and generalized joint attribute models (GJAMs), taking species co-occurrences into account (Clark et al., 2017). We then compute spatial prioritization both horizontally and vertically, across three depths layers, using a customized version of prioritizr (Hanson et al., 2022), at the scale of the archipelago, to protect 30% of fish diversity in various scenarios of connectivity. We then measure the area effectively covered by our 3D prioritization and contrast the result with the 30x30 CBD target. This study provides the first 3D spatial

conservation planning of a world biodiversity hotspot and reconcile area-based conservation planning with the 3 dimensional nature of oceans.

Methods

Survey area and data collection

New Caledonia is a South Pacific archipelago, east of Australia, in the Coral Sea. The 400 km-long main island of "Grande Terre" is surrounded by the second longest barrier reef in the world cumulating 1,600km. Beyond the barrier reef is the island's deep slope habitat, also surrounding the remote atolls (Chesterfield, Bellona, Entrecasteaux…), the three Loyalty Islands and the other smaller islands and remote reefs of the archipelago (Figure 1). Two third of the archipelago's population lives in Grand Terre's southwest, in and around Nouméa (~180,000 people) (ISEE, 2019), creating a gradient of human pressure from densely populated areas to wilderness areas located at > 10-20h travel time from the capital (Januchowski-hartley et al., 2020).

Data was collected during 4 cruises on the R/V Alis, in April and June 2019 and August and September 2020. We sampled 11 seamounts across the archipelago, and 4 deep island slopes along the west coast of Grande Terre (Figure 1, Figures S1-15). Samples were collected on the summit of the seamounts and on the bottom along the deep slopes (benthic samples), as well as in the pelagic waters, 2-7 miles away from the seamounts or island slopes (pelagic samples). The seamounts were chosen to have different summit depths corresponding to euphotic, intermediate and aphotic zones: 4 seamounts had their summit shallower than 200m deep, 4 seamounts had their summits between 200 and 320m deep, and 3 seamounts had their summits between 320 and 500m deep. Deep slopes were sampled between 100 and 220m deep (Table S1).

Benthic environmental DNA samples were collected with 4 x 8L Niskin bottles at each station, 5m above the seafloor, with 10 stations on each seamount summit or deep island slope (total of 148 stations). The shallowest benthic eDNA sample was collected at 45m deep and the deepest at 570m deep (Table S1). Pelagic eDNA samples were collected at two vertical profiles per site, with samples of 32L collected at 6 depths: 20, 80, 150, 250, 500 and 1000m. Details on the filtration and storage can be found in Method S1.

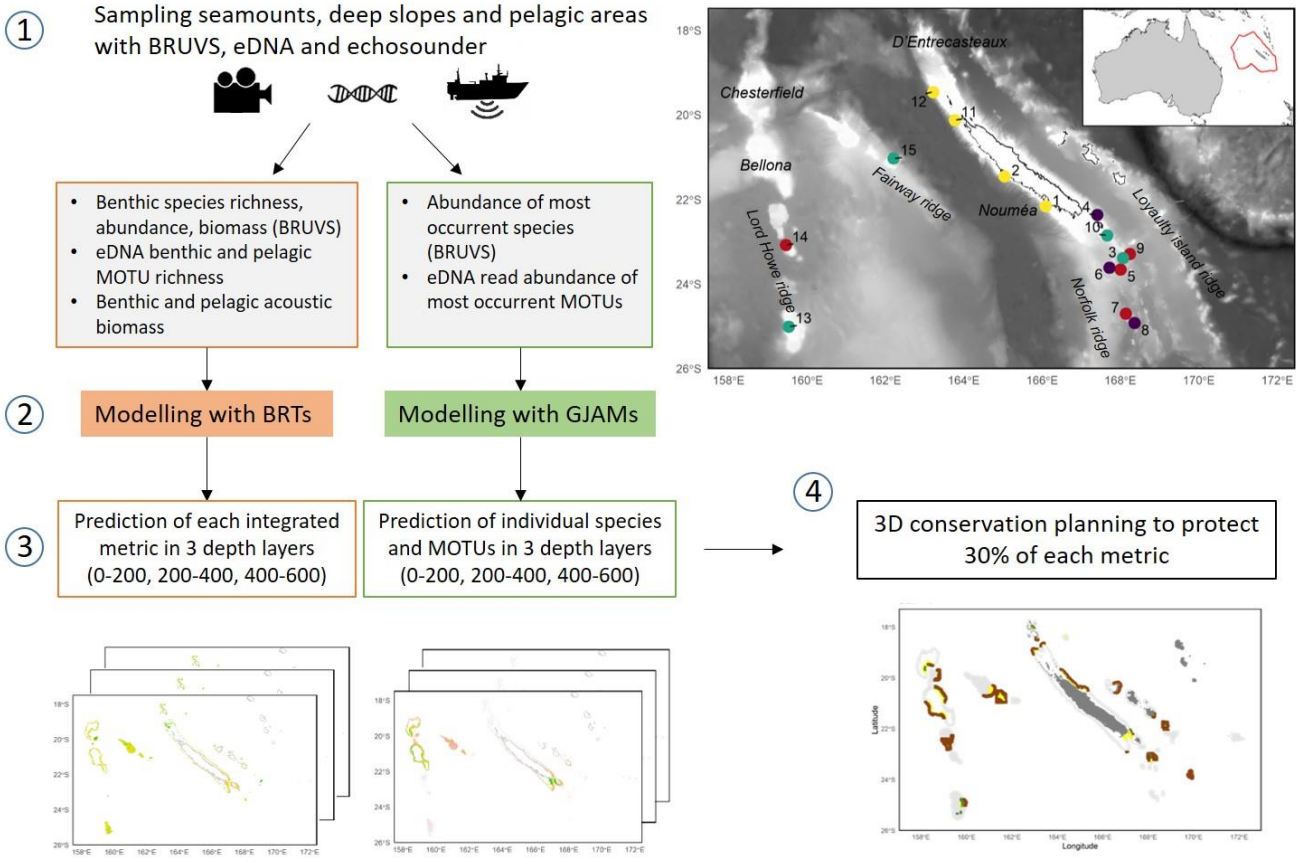


Figure 1. Framework of our study. *1) Sampling : 4 sites on deep island slopes ~150m (yellow), 4 sites on seamounts with summit ~50m (green), 4 sites on seamounts with summit ~250m (red) and 3 sites on seamounts with summit ~500m (purple). 2) Modelling of diversity metrics with BRTs and GJAMs. 3) Predictions of each diversity metric. 4) 3D conservation planning.*

On the same sites (summits and deep slopes) BRUVS were deployed on the seafloor, at 5 to 10 stations per site (total of 120 stations, Table S1). Stations were separated by at least 1km to avoid individuals from appearing on multiple videos and to assume independence of samples (Langlois et al., 2020). The BRUVS were composed of 2 cameras aligned horizontally on a metallic structure, a bait of 1kg of crushed sardines at the end of a 1.5m bar facing the cameras and a spotlight (Method S1). The shallowest BRUVS was deployed at 47m deep and the deepest at 552m deep (Table S1).

Acoustic data were recorded continuously during the cruises, using an EK60 echosounder (SIMRAD Kongsberg Maritime AS, Horten, Norway) connected to four split-beam transducers at 38, 70, 120 and 200 kHz. EK60 calibration was performed according to Foote (1987) for each cruise. In the present study, we used 38 kHz only as only that frequency allowed to cover

all depth ranges considered. Acoustic data collection started at 10 m below the surface. The maximum detection range was 800 m for all the surveys.

Data processing

eDNA extraction, amplification and sequencing

DNA extraction was performed in a dedicated DNA laboratory (SPYGEN, www.spygen.com) equipped with positive air pressure, UV treatment and frequent air renewal. Decontamination procedures were conducted before and after all manipulations. Detailed protocols of DNA extraction, amplification and sequencing can be found in Method S2 and in Polanco et al. (2022). A teleost-specific 12S mitochondrial rRNA primer pair (teleo, forward primer - ACACCGCCCGTCACTCT, reverse primer – CTTCCGGTACACTTACCATG, Valentini et al., 2016) was used for the amplification of fish metabarcode sequences, with 12 PCR replicates per sample. The amplified DNA were then sequenced using Illumina MiSeq or NextSeq sequencers (Illumina, San Diego, CA, USA) at Fasteris (Geneva, Switzerland).

eDNA bioinformatic analyses

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) (Marques, Guérin, et al., 2020). First, reads were assembled using vsearch (Rognes et al., 2016), then demultiplexed and trimmed using cutadapt (Martin, 1994) and clustering was performed using Swarm v.2 (Mahé et al., 2015) with d=1, which corresponds to a maximum of 1 mismatch between neighboring pairs of sequences within each cluster. Taxonomic assignment of MOTUs was carried out using the Lower Common Ancestor (LCA) algorithm ecotag implemented in the Obitools toolkit (Boyer et al., 2016) with the European Nucleotide Archive (ENA, Leinonen et al., 2011) as a reference database (release 143, March 2020). Details on bioinformatic analyses can be found in Method S3.

Video analysis

Calibration of stereo videos was done using the software CAL and fish were counted using the EvenMeasure software (www.seagis.com.au). We used the MaxN metric (corresponding to the maximum number of individuals of a particular species seen in any one video frame across

the duration of the video record), which is until now the standard and most used method (Cappo et al., 2007; Langlois et al., 2020; Whitmarsh et al., 2017). Fork length of individual fish was measured, when possible, up to a limit of 10 individuals per BRUVS per species to optimize video processing time. Biomass was calculated for each species of each BRUVS using the length-weight relationship (Taylor & Willis, 1998). Details on length measurements and estimations can be found in Method S4.

Acoustic data processing

All raw acoustic data were processed with the open-source Matecho software (Perrot et al., 2018). The first steps of the processing consist of several cleaning steps to remove (i) ghost bottom echoes, (ii) acoustic device interference, (iii) attenuated signals, (iv) elevated signals and (iv) reduced background noise (De Robertis & Higginbottom, 2007). Details of filter parameters can be found in Béhagle et al., (2016) and Perrot et al., (2018). After data cleaning, we integrated acoustic data in 500m distance by 10m depth bins, providing the nautical area scattering coefficient (NASC or S_A, in $m^2.nm^{-2}$), a proxy for marine organisms' biomass (Dornan et al., 2019; Irigoien et al., 2014). The final dataset was composed of 5,064 vertical profiles ranging from 10 to 800 m depth with S_A integrated in 10m vertical bins and 500m horizontal resolution.

Statistical analysis

Fish biodiversity metrics

We pooled the eDNA reads from the 12 PCR replicates per station and calculated the MOTU richness per station, for both the benthic and pelagic samples. Additionally, we extracted the individual MOTU read number for MOTUs present in at least 30% of stations per habitat. From the BRUVS data, we computed the fish species richness, abundance and biomass per station and we also extracted individual species abundance per station for species present in at least 30% of the stations of each habitat and depth class (summit ~50, deep slope ~150, summit ~250 and summit ~500). Acoustic data were divided into benthic and pelagic compartments. The benthic data consisted of the mean of S_A in the 20m above the seafloor in each cell. The pelagic data consisted of vertical profiles of S_A from 10m to the start of the benthic layer, for each cell.

Environmental explanatory variables

Seventeen variables were collected as potential explanatory variables for fish biodiversity patterns. At each station, we recorded the latitude and longitude, the sampling depth, the bottom depth, the habitat (seamount or deep island slope) and the depth of the summit. Using a bathymetry at 100m resolution (Roger, 2020), we calculated the summit area (km^2) and the summit rugosity as the standard deviation of depth in the cells of the summit area. For deep slope stations, the summit depth was set at 0, as the land is considered to be the summit, and the summit area was calculated as the area of cells with depth < 60m. For each station, we extracted maximum and mean sea surface temperature (SST), mean surface salinity, eastward and northward current velocity, surface suspended particulate matter, seafloor potential temperature and chlorophyll a over the last 10 years from available rasters. Details on the sources and resolution of each variable can be found in Table S2. We also calculated the travel time from Nouméa to our stations as a proxy for human pressure and habitat remoteness (Januchowski-hartley et al., 2020; Maire et al., 2016) and the minimum distances from our stations to reefs and land, using the New-Caledonia Millennium Geomorphology (Andréfouët et al., 2006).

Modelling abundance, richness and biomass

Boosted regression trees (BRTs, Elith et al., 2008) were used to model the 7 diversity metrics against the 17 available explanatory variables (Method S5). BRTs are able to cope with strongly interacting factors and nonlinear relationships, and are particularly suited for predictive modelling of complex biodiversity metrics. The function *gbm.step* from the package "dismo" (Hijmans et al., 2017) was used to find the combination of parameters producing the best fit. Parameters were the tree complexity (from 1 to 5), the learning rate (0.01, 0.005 or 0.001) and the bag fraction (0.5 or 0.75). All possible combinations of tree complexity, learning rate and bag fraction were run. The combination with the lowest deviance and standard error (evaluated over a 10-fold cross-validation) provided the set of best parameters. Then, models were computed again with these best parameters and fixed number of trees with the function *gbm.fixed*. The predictors contributing the most (>5%) to the models were selected, and a gbm.step followed by a gbm.fixed were computed with this reduced predictors selection to fit the final BRT model. See Method S5 for details on predictor contribution computation.

GJAMs were computed on the BRUVS species abundance matrix, and the benthic and pelagic matrix of eDNA read numbers per sample, using the function *gjam* from the package "gjam". The explanatory variables were selected with a step-by-step process, analyzing the sensitivity of the response variables to each variable and their quadratic and cubic terms to include non-linear responses. Only variables with a variable inflation factor (VIF) < 20 were kept in the final models. The response variable type was set to 'discrete abundance' for species abundance and 'count composition' for read numbers, and the models were run with parameters ng=2500 and burnin=500.

Predictions at the scale of New-Caledonian EEZ

Using the best models, we predicted the 7 biodiversity metrics and the individual species and MOTUs abundances at the scale of the New-Caledonian EEZ, at a resolution of 1x1km, according to the explanatory variables. To remain in the validity range of our data, we selected seamounts in the EEZ with a summit shallower than 600m deep (n=22, Allain et al., 2008), and all the deep slopes surrounding islands, atolls, banks or drowned atolls, where the bottom depth did not exceed the maximum bottom depth on which we sampled (2175m). For the benthic data, we predicted diversity from below the surface to 600m on seamounts, and from 60 to 600m on deep slopes. For pelagic data, we made predictions by layers of 20m down to 600m deep, on top and around seamounts and deep slopes. So for deep slopes, predictions were extrapolated down to 600m. Predictions from the 7 BRTs were carried out with the function *predict* from package "dismo", and predictions from the 3 GJAMs were made with the function *gjamPredict* from the package "gjam".

Spatial conservation planning in 3D

The benthic prediction rasters were divided in 3 depth layers (bottom between 0-200m, 200-400m and 400-600m, approximately corresponding to euphotic, intermediate and aphotic zones, respectively), and the pelagic predictions were aggregated within these 3 depth layers (sum of S_A, and mean of MOTU richness). We used the spatial prioritization package *prioritizr* (Hanson et al., 2022), with the Gurobi optimizer, to identify conservation priority areas across the 3 depth layers. In order to perform the 3D spatial prioritization, we modified the input data required for the 2D prioritization, to consider each 1 km x 1km (horizontal resolution) x 200 m (vertical resolution) unit as a volume (details can be found in Method S6). The relative conservation targets were set to 0.3 (30%) for each biodiversity metric and individual species

or MOTU abundance, to prioritize the effective protection of areas of high diversity (Devillers et al., 2020; Huang et al., 2022). We set equal cost to all the planning units of 1km side x 200m deep.

The formulation of the problem with prioritizr includes a factor referred to as boundary length modifier (BLM), which controls the compactness of selected units. BLM is equal to the perimeter to surface ratio and balances the aggregation and patchiness of reserve solutions. Low BLM values allow high fragmentation of the solution, while high BLM values emphasize compact solutions. We computed 8 iterations of our prioritization problem with BLM values between 0 and 10 (0, 10-5, 10-4, 10-3, 10-2, 10-1, 1, 10), computed the total cost and the total boundary length of each solution, and used the TOPSIS method (Hwang & Yoon, 1981) to rank the solutions. The prioritization with the greatest TOPSIS score was considered to represent the best trade-off between total cost and total boundary length.

Results

Diversity overview

Environmental DNA metabarcoding analyses produced 113,421,197 sequence reads, clustered into 596 MOTUs, belonging to at least 93 families. The most occurrent families (>100 MOTUs) were the Myctophidae, Scombridae, Serranidae, Lutjanidae, Lethrinidae, Mullidae and Labridae. MOTUs also were associated with different habitats (Figure S16). Mean richness per sample was 12.1 MOTUs ± 12, and mean number of reads per sample was 527,540 ± 721,560 (Table S3). The highest MOTU richness was recorded on deep slopes, while the lowest MOTU richness was recorded on deep seamount. Twelve benthic MOTUs of which 4 belong to Myctophidae (Table S4), and 10 pelagic MOTUs belonging to Myctophidae, Scombridae, Lutjanidae and Lethrinidae (Table S5) were frequent enough to be retained in the GJAM modelling.

A total of 190 species were identified from BRUVS, belonging to 53 families. The most represented families (>10 species) were the Labridae, Serranidae, Lutjanidae, Acanthuridae and Carangidae. The most frequent species (> 40 observations) were *Seriola rivoliana*, *Carcharhinus albimarginatus*, *Pristipomoides filamentosus* and *Gymnocranius euanus*. Twenty-four percent of species were found only on deep slopes, 28% only on shallow seamounts, 2.6% only on intermediate seamounts and 8% only on deep seamounts, while 30% were common to both deep slopes and seamounts (Figure S17). Mean richness per BRUVS was

6.84 species ± 8.5, mean biomass was 113.28 kg ± 121 and mean abundance was 35.3 individuals ± 80. The highest species richness and biomass were recorded on shallow seamounts, while the lowest richness and biomass were recorded on deep seamounts. A total of 23 species were present in a least 30% of BRUVS stations in one habitat (Table S6). These belong to 10 families, including Lutjanidae (*Aphareus rutilans, Aprion virescens, Etelis coruscans, Pristipomoides flavipinnis, P. filamentosus, P. argyrogrammicus*) and Carcharinidae (*Carcharhinus albimarginatus, C. plumbeus*).

Modelling abundance, richness and biomass

The modelling of species richness, abundance and biomass, MOTU richness and acoustic biomass with BRTs all reached moderate to high cross-validation accuracy (between 0.44 and 0.85) while including different sets of explanatory variables (Table S7).

The relative contribution of each variable varied largely among models (Figure 2), but depth (seafloor depth and summit depth), remoteness (travel time to Nouméa, distance to the reef and land), and temperature (SST and seafloor temp.) appeared most important factor to explain the diversity metrics (Figure 2). Among all models, seafloor depth ranked first with a mean contribution of 20.7% ± 25%, further justifying the need for 3-dimension planning. Seafloor depth contributed the most in explaining BRUVS fish abundance and biomass, while the sampling depth influenced the pelagic acoustic biomass. Travel time had a mean contribution of 13.8% ± 11%, so was a major factor explaining eDNA benthic MOTU richness, acoustic biomass and BRUVS species richness. Finally, surface temperature had a mean contribution of 12.7% ± 9.7%, and explained most acoustic biomass, BRUVS fish abundance and eDNA MOTU richness. All the diversity metrics showed varying relationships with the explanatory factors (Figures S18-24).

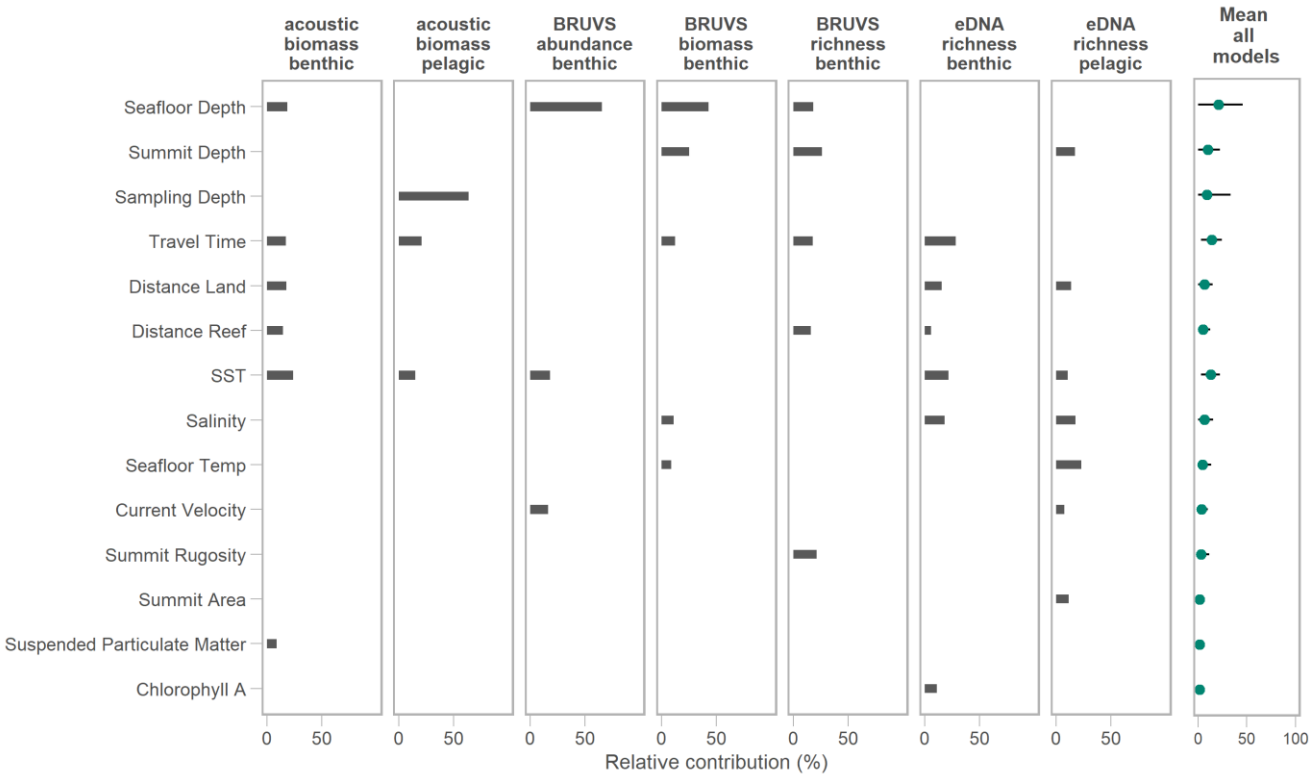


Figure 2. Relative contribution of each variable in the BRTs. *Relative contribution of each variable in percentage within each model. A contribution of 0% indicates that the variable was not included in the model. Mean contribution of each variable among models (last panel, blue points), error bars indicate standard deviation among models.*

The modelling of individual species abundances and MOTU read numbers with GJAMs also provided moderate to high goodness-of-fit, with Pearson's r correlation between observed and predicted values ranging from 0.62 ($p < 0.001$) to 0.68 ($p < 0.001$). The details on each model parameters and outputs can be found in Table S8. The GJAM on BRUVS species abundance revealed 3 clusters, mostly explained by salinity, habitat and distance to land (Figure 3, Figure S25). The first cluster included species associated with great depth, either in seamount or deep slope habitat (*Etelis coruscans*, *Seriola lalandi*, *Pentaceros richardsoni*, *Pristipomoides argyrogrammicus*, *Squalus megalops* and *Polymixia japonica*). The second one clustered species associated to seamounts and shallow depth (*Pseudocaranx dentex*, *Carcharhinus plumbeus*, *Aprion virescens*, *Lethrinus sp* and *Gymnocranius euanus*). The third cluster grouped species associated to the deep slope habitat and low salinity, some with shallow depth (*Epinephelus sp*, *Pristipomoides flavipinnis*, *Aphareus rutilans*, *Gymnosarda unicolor* and *Carcharhinus albimarginatus)*, and others with high depth (*Wattsia mossambica*, *Seriola rivoliana* and *Pristipomoides filamentosus*).

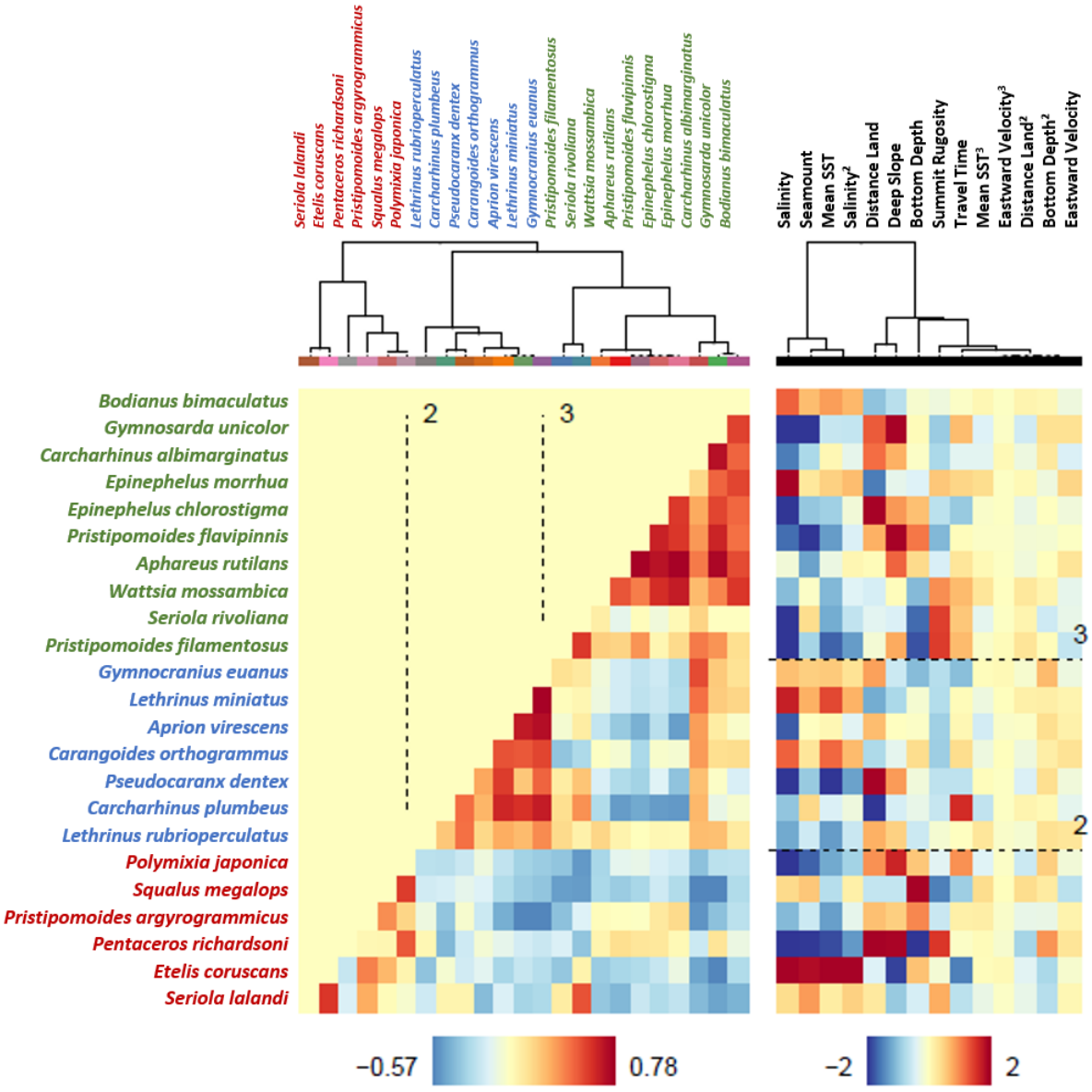


Figure 3. Cluster and grid plot of fitted BRUVS abundances and predictors. *Left panel represents correlation among species in terms of their responses to predictors ($\hat{E}$). Right panel represents the correlation of each species with predictors (B). Dotted lines represent the 3 species clusters. Red indicate strong positive correlation, and blue indicate strong negative correlation.*

The GJAM on benthic MOTU reads revealed 3 clusters mostly influenced by SST, suspended particulate matter and distance to reef (Figures S26-27), while the GJAM on pelagic MOTU reads highlighted 2 clusters influenced by salinity, SST and habitat (Figures S28-29).

Archipelago-wide predictions

From the fitted values of all BRTs and GJAMs, we predicted the 7 biodiversity metrics, the 23 selected species abundances, and the 22 selected MOTUs read numbers at the scale of the New-Caledonian archipelago. Predictions reflected well the influence of each explanatory variable and showed spatial heterogeneity both horizontally and vertically (Figures S30-33). For example, fish species richness predicted from the BRUVS dataset was higher in the first depth layer (seafloor between 0 and 200m deep) of shallow seamounts and remote atoll slopes, and on the third depth layer (seafloor between 400 and 600m deep) of summits and slopes of deep seamounts. The spatial difference in species richness distribution among depth layers was especially visible on the deep seamounts south of Bellona atoll, south of Grande Terre and south of the Loyalty islands, and near the shallow atolls and reefs in the northern and southern lagoons, and midway between Grande Terre and Chesterfields (Figure 4.AB). Likewise, strong spatial differences existed in the distribution of modelled species and MOTUs. For example, the commercially important deep-water snapper *Pristipomoides filamentosus* was more abundant on relatively shallow atoll and island slopes when the dogfish shark *Squalus megalops* showed higher abundance on deepest slopes and seamounts (Figure 4.CD). In order to account for such an horizontally and vertically heterogenous seascape, benthic and pelagic predictions were divided into three depth layers (0-200m, 200-400m and 400-600m) and included in the prioritization computation in three dimensions (Figures S34-35).

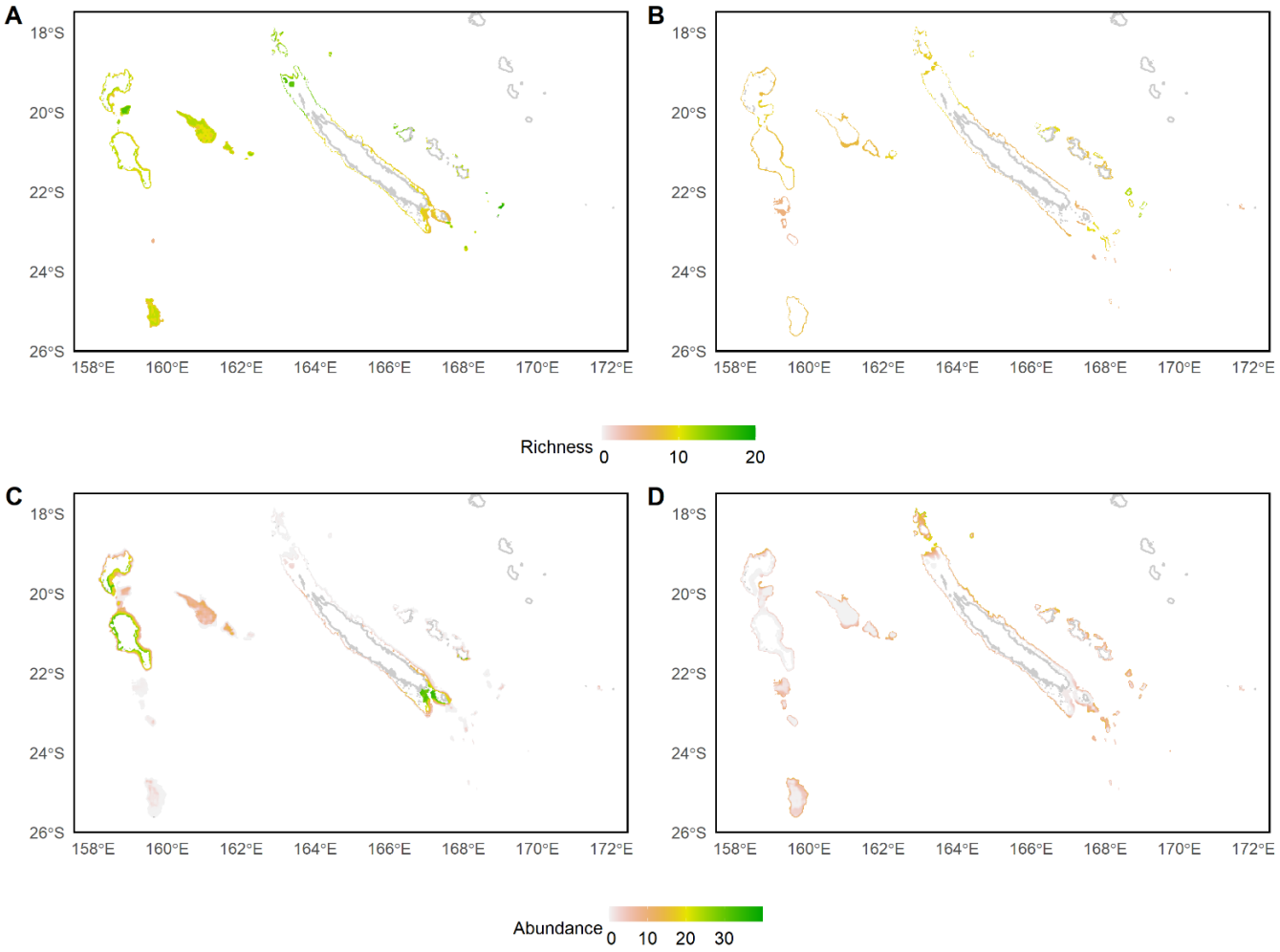


Figure 4. Prediction of fish species richness and individual species abundance measured by BRUVS, *from the fitted values of the BRTs and GJAMs, in all seamounts and deep slopes of the New-Caledonian EEZ, down to 600m deep. A) species richness in cells shallower than 200m deep, B) species richness in cells with seafloor between 400 and 600m deep, C) abundance of* Pristipomoides filamentosus *down to 600m deep, D) abundance of* Squalus megalops *down to 600m deep.*

Spatial conservation planning in 3D

To identify the best 3D conservation planning solution, 8 scenarios were computed with different boundary length modifier values (BLM). Setting no penalty on fragmentation (BLM=0) led to a highly fragmented solution of 60,241 prioritized planning units, with a large spatial extent (29,788 km²), and units prioritized only in one or two depth layers, where the biodiversity is the highest (Figure 5.A). With this solution, 44% of planning units were prioritized across all depth layers, mostly located on remote deep slopes and on several

seamounts. Twenty-four percent of planning units were prioritized only in the first depth layer (0-200m), mostly on the shallow slope around the Chesterfield and on shallow seamounts, while 8% of planning units were prioritized only in the intermediate depth layer (200-400m), and 10% only in the deepest depth layer (400-600m). Nine percent of planning units were prioritized both in the first and second depth layers (0-400m), 4% were prioritized both in the second and third depth layers (200-600m), and 1% both in the first and third depth layers (0-200 and 400-600m).

The ranking of the different prioritization solutions with the TOPSIS method based on the total cost and total boundary length of each solution identified the best solution with a BLM value of 1 (Table S9, Figure S36). This solution had a total cost of 63,636 planning units, total boundary length of 39,008,162 m (versus 91,070,476 m for BLM0), and total surface of 24,931 km², including 11,668 km² on slopes and 13,263 km² on seamounts (Figure 5.B). This corresponded to 29.3% of the total spatial domain (total area of 85,247 km^2), 27.0% of the slopes (43,181km^2) and 31.5% of the seamounts (42,066 km^2) considered in the study down to 600m depth. More than 70% of planning units prioritized by this solution are across all depth layers, 16% only in the shallowest depth layer, 13% both in the first and second depth layers, <1% only in the second depth layer, and none were selected in the deepest depth layer alone. This conservation solution comprised 17 main areas: 6 areas located on slopes and seamounts of the Chesterfield-Bellona alignment (Lord Howe ridge), 1 area on the Fairway ridge, 2 areas on the slopes of Entrecasteaux and the Great Northern Lagoon, 1 area on the north-east of Grande Terre, 3 areas on the slopes and seamounts of the Loyalty islands ridge, and 4 areas on slopes in the south of Grande Terre and on southern seamounts (Norfolk ridge). Thus, this solution allows to preserve 30% of the fish biodiversity in ~30% of the considered spatial domain while accounting for the 3D distribution of biodiversity and offering relative compactness to implement management plans.

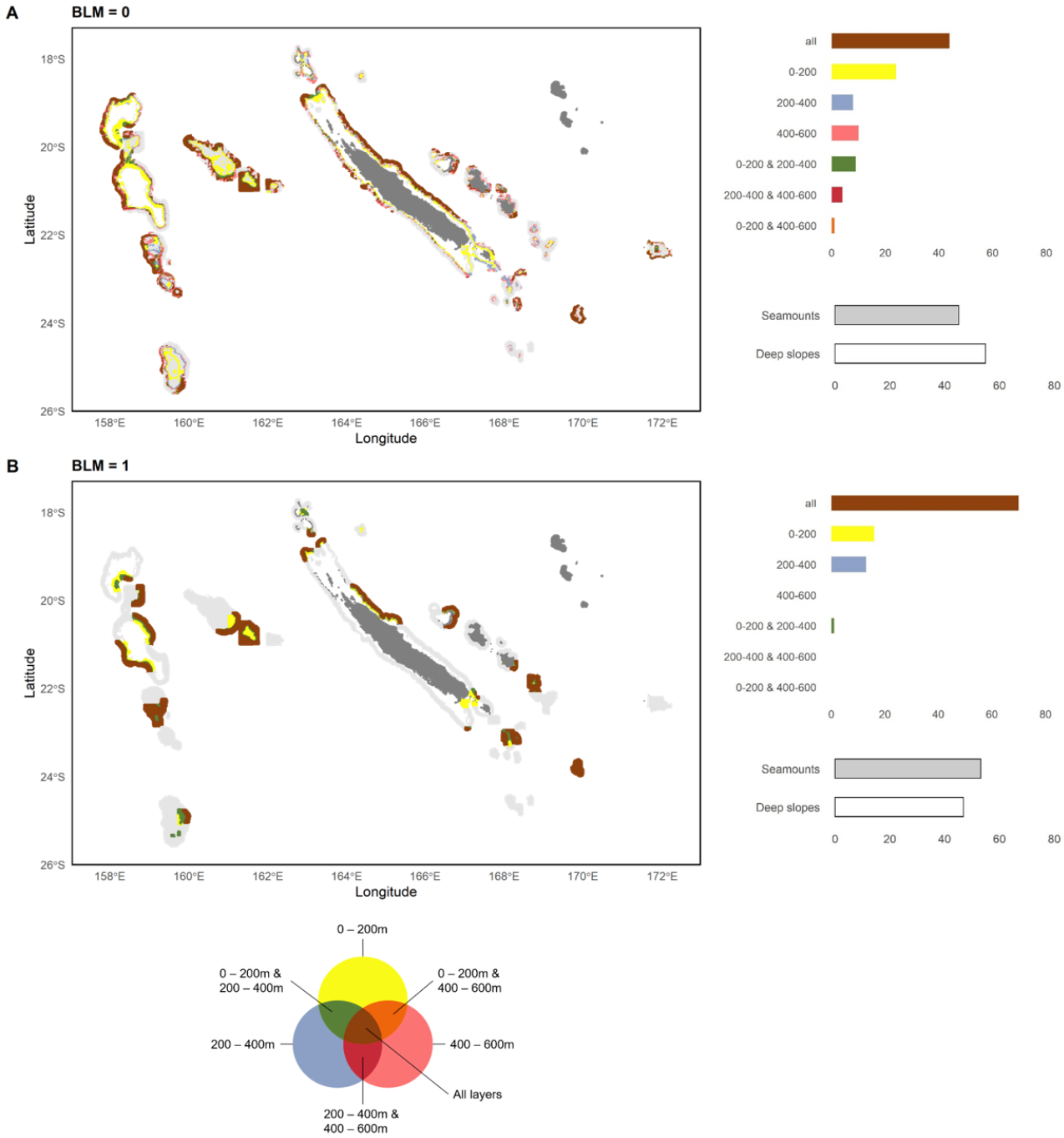


Figure 5. Prioritization maps across space and depth, along a compactness gradient. *A) Solution with boundary length modifier (BLM) value = 0, (i.e. fragmented solution), B) solution with BLM = 1 (best solution identified by the TOPSIS score). Colors indicate for which depth each planning unit is prioritized, as shown by the Venn diagram. Dark grey areas indicate land. Light grey areas indicate planning units not prioritized by the solution. Histograms indicate the percentage of planning units across depth layers, and habitats.*

Discussion

To address the global biodiversity crisis, governments are implementing strategic plans with area-based conservation targets as the cornerstone. In 2010, the UN Convention on Biological Diversity was signed by 168 countries with an explicit target (Aichi Target 11) of at least 10 % sea areas protected by 2020, with a special attention to areas of particular importance for biodiversity and ecosystem services. The post-2020 UN strategic plan, released in 2021, targets a protection of 30% of the sea areas by 2030 (Target 3), with a special attention to areas of particular importance for biodiversity and its contribution to people (CBD, 2021). Between 2010 and 2019, protected areas expanded from covering 2.9% to 7.5% of the global marine realm (Maxwell et al., 2020). However, deep-sea environments remain largely unprotected with only 2% of world seamounts inside MPAs (Letessier et al., 2019). The difficulty to define areas of particular importance for biodiversity in deep-sea environments, and the complexity to apply an area-based approach in a 3-dimensional oceanic environment, may partly explain this failure. In this study, we collected and modelled a set of biodiversity metrics in 15 seamounts and deep slopes of the archipelago of New Caledonia and conducted spatial conservation planning in 3-dimensions with a 30% target for each of our 52 modelled metrics. The conservation solution, focusing on areas of high importance for biodiversity, included ~30% of the area covered by the studied spatial domain, which is coherent with the 30x30 CBD target. This solution was sufficiently compact to provide a reasonable first draft of an area-based management plan, while accounting for the 3D distribution of species and biodiversity down to 600 m depth. This approach seems therefore promising to implement area-based strategic plans in poorly known 3D structured deep-sea environments.

The prioritizing solution of 3D marine planning with no penalty on compactness was highly fragmented both horizontally and vertically, with many planning units selected in only one or two depth layer. Setting a boundary length modifier of 1 not only resulted in an horizontally compact solution but also aligned vertically the selected planning unit, de facto resulting in a 2D solution able to protect the 3D distribution of biodiversity by favoring the protection of a few large contiguous areas, including all habitat types (seamounts and slopes) across the whole water column. The protection of a few areas of particular importance for biodiversity is coherent with the CBD recommendations. In our study these few important areas were derived from the 3D distribution of 52 modelled biodiversity metrics. To some extent, they correspond to the concepts of SAC (Special Area for Conservation), KBA (Key Biodiversity Area), EBSAs (Ecologically or Biologically Significant Areas) that are widely used in conservation planning.

The archipelago of New-Caledonia hosts 1/3rd of global wilderness coral reefs, mostly located at Chesterfield, Bellona, and D'Entrecasteaux reefs (Januchowski-hartley et al., 2020). These remotes reefs are now included in highly protected areas with depth limit to 1000m (Claudet et al., 2021). The 17 areas selected by our best planning solution include several deep-slopes of these wilderness reefs, which are therefore already protected. Most other selected areas are on seamounts in the Coral Sea Marine Park where human pressure is low and where protection may be easier to achieve than on populated areas along the Grande Terre and the Loyalty islands.

Our study is one of the few comparing biodiversity from seamounts and deep slopes, and combining several novel sampling technologies (Letessier et al., 2019; Mazzei et al., 2021; Salvetat et al., 2022). Most studies on deep water diversity only focus on one habitat and use no more than one or two sampling methods (Annasawmy et al., 2019; Cherel et al., 2020; Mejía-Mercado et al., 2019; Quattrini et al., 2017). Using various sampling methods enabled us to investigate 52 biodiversity metrics. The modelling framework revealed that each metric was explained by a different combination of environmental variables, hence had a rather unique distribution. The 3D optimization was therefore essential to select the planning units that could protect all metrics within such an heterogeneous seascape. Despite the heterogeneity, depth, remoteness (travel time, distance to land and reefs), and temperature were the strongest predictors of our biodiversity metrics. All metrics decreased with increasing summit depth or seafloor depth, which is consistent with previous studies of fish diversity along depth gradients (McLean et al., 2018; Quattrini et al., 2017; Smith & Brown, 2002). Pelagic acoustic biomass showed a second peak around 500m deep, which corresponds to the region of micronektonic concentration during day-time (Annasawmy et al., 2018; Ariza et al., 2016). Increasing distance to land and reefs had a negative effect on all biodiversity metrics. Areas close to land and reefs, and with a low human impact, benefit from higher habitat diversity, terrigenous influence (Carassou et al., 2010), and reef areas supporting large populations and individuals (Gove et al., 2016; Kulbicki et al., 2015). Most biodiversity metrics increased with moderate travel time, before showing a decrease. A similar relationship had previously been shown for reef biodiversity (D'Agata et al., 2014; Maire et al., 2016), where biomass but also functional and phylogenetic diversity increased with travel time. Sea surface temperature (SST) had a divergent effect on biodiversity metrics: benthic MOTU richness decreased at high SST, while benthic acoustic biomass was the highest at low and high SST.

Our biodiversity predictions were mostly higher on deep slopes around the Grande-Terre and remote atolls (Chesterfield, Entrecasteaux), and on shallow seamounts (Capel, Fairway). Deeper seamounts hosted high abundances of species associated with great depth, but lower biodiversity in general. These results are coherent with recent studies on seamounts showing that deep-sea fish assemblages were strongly correlated with depth, salinity, rugosity and chlorophyll-a (Mejía-Mercado & Baco, 2022; Muff et al., 2022), and that shallow seamounts are refuges for marine predators such as sharks, jacks, tunas and billfish (Letessier et al., 2019; Morato et al., 2010; Morato et al., 2008).

This study of seamounts fish biodiversity provides one of the first complete framework, from sampling to spatial conservation planning in three dimensions. We sampled 11 seamounts and 4 deep slopes, down to 600m deep, across the New-Caledonian EEZ, using various cutting-edge technologies (eDNA, BRUVS and ship-born acoustically-derived micronekton). New-Caledonian marine domain counts 83 seamounts, among which 22 have their summit above 600m deep (Allain et al., 2008). We have thus sampled half of the seamounts in this category. Despite our massive sampling effort, we could not sample seamounts deeper than 600m or deep-sea environments such as abyssal valleys between seamounts. While sampling on the seafloor of the valleys would be technically challenging, it may bring some evidence for seamounts as hotspots of fish diversity in the open ocean (Campanella et al., 2021). We sampled deep slopes between 100 and 220m deep, and extrapolated the predictions down to 600m to have similar depth range as seamounts data. While other studies confirmed that deep island slopes harbor diverse fish and coral assemblages (Cruz-Acevedo et al. 2018; Etnoyer et al., 2022), it would be necessary to add some samples from deep slopes at 600m to increase confidence in our interpretation. In this study, we only sampled, modelled and predicted fish biodiversity. While our prioritization solution reflects the most important areas for the conservation of 30% of fish biodiversity, some other areas may be more important for other taxa. For example, seamounts, have been identified as oasis of epibenthic megafauna (Rowden et al., 2010), and some as important feeding areas for cetaceans due to modified oceanography around seamounts and increased prey biomass (Romagosa et al., 2020; Wagner et al., 2021). Each of the sampling methods we used had its own limitations. The accuracy of diversity estimates with BRUVS can be limited by the visibility at the station (Langlois et al., 2020). The acoustic estimates obtained with the 38kHz echosounder correspond to micronekton, so include mostly fish but also crustaceans and cephalopods. Fish estimates could be incomplete, as a large biomass of mesopelagic fish without gas-filled swim bladders may be present but hidden by

stronger scatterers to the acoustic signal (Davison et al., 2015; Foote, 1980). One way to improve our acoustic estimates could be to use multi-frequency or wide-band acoustic vertical profilers. For the analysis of eDNA metabarcoding data, we used MOTUs curated by a conservative pipeline decreasing the number of MOTUs representing the same taxa and thus better reflecting the true level of fish diversity (Brandt et al., 2021; Marques, Guérin, et al., 2020; Sales et al., 2021). This methodology may, however, underestimate fish diversity of some rare fish species that are more poorly represented in public databases. Improving the accuracy of taxonomic assignment and completing genetic reference databases are thus urgently needed to improve local deep-sea monitoring (Marques et al., 2020).

Our 3D conservation planning solution remains theoretical. We applied the same cost to all planning units, to be equally considered in the solution. We did not take constraints such as artisanal or industrial fishing grounds, or seafloor mineral resources into account. To compute a realistic prioritization plan, we should apply higher cost to areas that are used for fishing activities, resources exploration or with high cultural values, and lower costs to remote areas and areas already protected. This mapping of the costs should be made by consulting stakeholders, users of the sea, scientists and governments to find a solution that satisfies all the needs while protecting biodiversity effectively (André et al., 2021; Venegas-Li et al., 2018). It would also be interesting to consider the climate refugia and predicted climate change in our prioritization, to meet conservation targets for future marine biodiversity distribution (Brito-Morales et al., 2022; Doxa et al., 2022).

From an extensive deep-sea sampling down to 600m, on 15 seamounts and deep slopes across the New-Caledonian archipelago, combining high technology methods, we provide a framework of biodiversity modelling, prediction and marine prioritization across space and depth. Our results suggest that fish biodiversity is strongly structured across depth, although each metric had its own spatial distribution within a highly heterogeneous seascape. 3D conservation planning allowed defining a conservation solution comprising few areas with special importance for biodiversity across all depth, and covering 30% of the investigated spatial domain. Our study paves the way for a new methodology in the selection of conservation areas in 3D structured environments, taking depth and its biodiversity into account while providing solutions compatible with area-based targets agreed by the international community.

Author contributions: LM and LV conceived the study. LV designed the sampling. FB, LV, MD, LM, CM, CB, CB, CS, JG, CS and DV collected the data. TD supervised the eDNA laboratory analyses. LM and GL performed eDNA bioinformatic analyses, FB and MD analyzed the BRUVS data. ALD, JH, MP, LV and GR analyzed the acoustic data. LM performed the modeling and statistical analyses, and interpretation of outputs. LM wrote the first draft of the manuscript, and all authors contributed substantially to revisions.

Data accessibility: The data used for this article will be made available in an appropriate public repository (Dryad or Zenodo) upon publication, and the data DOI will be included in the article. The codes used for the analyses will be available at https://github.com/lmathon/Seamounts_3Dmodelling upon publication.

Acknowledgements: We acknowledge the help of Jeffrey Hanson, developer of the *prioritizr* R package, in the conception of the 3D prioritization algorithm.

Conflict of interest: The authors declare no conflict of interest.

References

Allain, V., Kerandel, J. A., Andréfouët, S., Magron, F., Clark, M., Kirby, D. S., & Muller-Karger, F. E. (2008). Enhanced seamount location database for the western and central Pacific Ocean: Screening and cross-checking of 20 existing datasets. Deep-Sea Research Part I: Oceanographic Research Papers, 55(8), 1035–1047. https://doi.org/10.1016/j.dsr.2008.04.004

André, L. V., Van Wynsberge, S., Chinain, M., Gatti, C. M. I., Dempsey, A., & Andréfouët, S. (2021). A framework for mapping local knowledge on ciguatera and artisanal fisheries to inform systematic conservation planning. ICES Journal of Marine Science, 78(4), 1357–1371. https://doi.org/10.1093/icesjms/fsab016

Andréfouët, S., Muller-karger, F. E., Robinson, J. A., Kranenburg, C. J., Torres-pulliza, D., Spraggins, S. A., & Murch, B. (2006). Global assessment of modern coral reef extent and diversity for regional science and management applications : a view from ... Proceedings of 10th International Coral Reef Symposium, 1732–1745.

Andrello, M., Darling, E. S., Wenger, A., Suárez-Castro, A. F., Gelfand, S., & Ahmadia, G. N. (2022). A global map of human pressures on tropical coral reefs. Conservation Letters, 15(e12858), 1–12. https://doi.org/10.1111/conl.12858

Annasawmy, P., Ternon, J. F., Marsac, F., Cherel, Y., Béhagle, N., Roudaut, G., … Ménard, F. (2018). Micronekton diel migration, community composition and trophic position within two biogeochemical provinces of the South West Indian Ocean: Insight from acoustics and stable isotopes. Deep-Sea Research Part I: Oceanographic Research Papers, 138(July), 85–97. https://doi.org/10.1016/j.dsr.2018.07.002

Annasawmy, Pavanee, Ternon, J. F., Cotel, P., Cherel, Y., Romanov, E. V., Roudaut, G., … Marsac, F. (2019). Micronekton distributions and assemblages at two shallow seamounts of the south-western Indian Ocean: Insights from acoustics and mesopelagic trawl data. Progress in Oceanography, 178(May), 102161. https://doi.org/10.1016/j.pocean.2019.102161

Ariza, A., Landeira, J. M., Escánez, A., Wienerroither, R., Aguilar de Soto, N., Røstad, A., … Hernández-León, S. (2016). Vertical distribution, composition and migratory patterns of acoustic

scattering layers in the Canary Islands. Journal of Marine Systems, 157, 82–91. https://doi.org/10.1016/j.jmarsys.2016.01.004

Béhagle, N., Cotté, C., Ryan, T. E., Gauthier, O., Roudaut, G., Brehmer, P., … Cherel, Y. (2016). Acoustic micronektonic distribution is structured by macroscale oceanographic processes across 20-50°S latitudes in the South-Western Indian Ocean. Deep-Sea Research Part I: Oceanographic Research Papers, 110, 20–32. https://doi.org/10.1016/j.dsr.2015.12.007

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. https://doi.org/10.1111/1755-0998.12428

Brandt, M. I., Trouche, B., Quintric, L., Günther, B., Wincker, P., Poulain, J., & Arnaud-Haond, S. (2021). Bioinformatic pipelines combining denoising and clustering tools allow for more comprehensive prokaryotic and eukaryotic metabarcoding. Molecular Ecology Resources, 21(6), 1904–1921. https://doi.org/10.1111/1755-0998.13398

Brito-Morales, I., Schoeman, D. S., Everett, J. D., Klein, C. J., Dunn, D. C., García Molinos, J., … Richardson, A. J. (2022). Towards climate-smart, three-dimensional protected areas for biodiversity conservation in the high seas. Nature Climate Change, 12(4), 402–407. https://doi.org/10.1038/s41558-022-01323-7

Campanella, F., Collins, M. A., Young, E. F., Laptikhovsky, V., Whomersley, P., & van der Kooij, J. (2021). First Insight of Meso- and Bentho-Pelagic Fish Dynamics Around Remote Seamounts in the South Atlantic Ocean. Frontiers in Marine Science, 8(June). https://doi.org/10.3389/fmars.2021.663278

Cappo, M., Harvey, E., & Shortis, M. (2007). Counting and measuring fish with baited video techniques -- an overview. Australian Society for Fish Biology Workshop Proceedings, Hobart, Tasmania, August 2006, (August), 101–114.

Carassou, L., Le Borgne, R., Rolland, E., & Ponton, D. (2010). Spatial and temporal distribution of zooplankton related to the environmental conditions in the coral reef lagoon of New Caledonia, Southwest Pacific. Marine Pollution Bulletin, 61(7–12), 367–374. https://doi.org/10.1016/j.marpolbul.2010.06.016

CBD. (2021). First Draft of the Post-2020 Global Biodiversity Framework. Secretariat of the United Nations Convention on Biological Diversity. In Cbd/Wg2020/3/3.

Cherel, Y., Romanov, E. V, Annasawmy, P., Thibault, D., & Ménard, F. (2020). Micronektonic fish species over three seamounts in the southwestern Indian Ocean. Deep-Sea Research Part II, (March 2019). https://doi.org/10.1016/j.dsr2.2020.104777

Clark, J. S., Nemergut, D., Seyednasrollah, B., Turner, P. J., & Zhang, S. (2017). Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data. Ecological Monographs, 87(1), 34–56. https://doi.org/10.1002/ecm.1241

Claudet, J., Loiseau, C., & Pebayle, A. (2021). Critical gaps in the protection of the second largest exclusive economic zone in the world. Marine Policy, 124(November 2020), 104379. https://doi.org/10.1016/j.marpol.2020.104379

Cruz-Acevedo, E., Tolimieri, N., & Aguirre-Villaseñor, H. (2018). Deep-sea fish assemblages (300-2100 m) in the eastern Pacific off northern Mexico. Marine Ecology Progress Series, 592, 225–242. https://doi.org/10.3354/meps12502

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. https://doi.org/10.1016/j.cub.2014.01.049

Dambach, J., & Rödder, D. (2011). Applications and future challenges in marine species distribution modeling. Aquatic Conservation: Marine and Freshwater Ecosystems, 21(1), 92–100. https://doi.org/10.1002/aqc.1160

Davison, P., Koslow, J. A., & Kloser, R. (2015). Acoustic biomass estimation of mesopelagic fish: backscattering from individuals, populations and communities. ICES Journal of Marine Science, 75(5), 1413–1424.

De Robertis, A., & Higginbottom, I. (2007). A post-processing technique to estimate the signal-to-noise ratio and remove echosounder background noise. ICES Journal of Marine Science, 64(6), 1282–1291. https://doi.org/10.1093/icesjms/fsm112

Devillers, R., Pressey, R. L., Ward, T. J., Grech, A., Kittinger, J. N., Edgar, G. J., & Watson, R. A. (2020). Residual marine protected areas five years on: Are we still favouring ease of establishment over need for protection? Aquatic Conservation: Marine and Freshwater Ecosystems, 30(9), 1758–1764. https://doi.org/10.1002/aqc.3374

Dornan, T., Fielding, S., Saunders, R. A., & Genne, M. J. (2019). Swimbladder morphology masks Southern Ocean mesopelagic fish biomass. Proceedings of the Royal Society B, 286, 1–8.

Doxa, A., Almpanidou, V., Katsanevakis, S., Queirós, A. M., Kaschner, K., Garilao, C., … Mazaris, A. D. (2022). 4D marine conservation networks: Combining 3D prioritization of present and future biodiversity with climatic refugia . Global Change Biology, (May), 1–12. https://doi.org/10.1111/gcb.16268

Duffy, G. A., & Chown, S. L. (2017). Explicitly integrating a third dimension in marine species distribution modelling. Marine Ecology Progress Series, 564, 1–8. https://doi.org/10.3354/meps12011

Dulvy, N. K., Simpfendorfer, C. A., Davidson, L. N. K., Fordham, S. V., Bräutigam, A., Sant, G., & Welch, D. J. (2017). Challenges and Priorities in Shark and Ray Conservation. Current Biology, 27(11), R565–R572. https://doi.org/10.1016/j.cub.2017.04.038

Eddy, T. D., Lam, V. W. Y., Reygondeau, G., Cisneros-Montemayor, A. M., Greer, K., Palomares, M.-L. D., … Cheung, W. W. L. (2021). Global decline in capacity of coral reefs to provides ecosystem services. One Earth, 4, 1278–1285.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. Journal of Animal Ecology, 77(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Etnoyer, P. J., Messing, C. G., Stanley, K. A., Baumiller, T. K., Lavelle, K., & Shirley, T. C. (2022). Diversity and time-series analyses of Caribbean deep-sea coral and sponge assemblages on the tropical island slope of Isla de Roatán, Honduras. Marine Biodiversity, 52(1), 1–17. https://doi.org/10.1007/s12526-021-01255-z

Foote, K. G. (1980). Importance of the swimbladder in acoustic scattering by fish: A comparison of gadoid and mackerel target strengths. Journal of Acoustical Society of America, 67(6), 2084–2089.

Foote, K. G. (1987). Fish target strengths for use in echo integrator surveys. Journal of the Acoustical Society of America, 82(3), 981–987. https://doi.org/10.1121/1.395298

Giakoumi, S., Sciann, C., Plass-johnson, J., Micheli, F., Grorud-colvert, K., Thiriet, P., … Lubchenco, J. (2017). Ecological effects of full and partial protection in the crowded Mediterranean Sea : a regional meta-analysis. Scientific Reports, (November 2016), 1–12.

Gove, J. M., McManus, M. A., Neuheimer, A. B., Polovina, J. J., Drazen, J. C., Smith, C. R., … Williams, G. J. (2016). Near-island biological hotspots in barren ocean basins. Nature Communications, 7(10581), 1–8. https://doi.org/10.1038/ncomms10581

Gust, N., Choat, J. H., & McCormick, M. I. (2001). Spatial variability in reef fish distribution, abundance, size and biomass: A multi-scale analysis. Marine Ecology Progress Series, 214, 237–251. https://doi.org/10.3354/meps214237

Hanson, J., Schuster, R., Morrell, N., Strimas-Mackey, M., Edwards, B., Watts, M., … Possingham, H. (2022). prioritizr: Systematic Conservation Prioritization in R. https://prioritizr.net, https://github.com/prioritizr/prioritizr.

Hijmans, R. J., Phillips, S., Leathwick, J., Elith, J., & Hijmans, M. R. J. (2017). Package 'dismo.' Circles, 9(1), 1–68. https://doi.org/10.1017/CBO9781107415324.004

Huang, S.-L., Wu, H., Li, Q., Jefferson, T. A., Chen, M., Peng, C., & Wang, X. (2022). Conservation planning for threatened marine megafauna:Moving forward with a better approach. Aquatic Conservation: Marine and Freshwater Ecosystems, 1–13.

Hwang, C.-L., & Yoon, K. (1981). Methods for Information on Attribute Given. Lecture Notes in Economics and Mathematical Systems, 58–191.

Irigoien, X., Klevjer, T. A., Røstad, A., Martinez, U., Boyra, G., Acuña, J. L., … Kaartvedt, S. (2014). Large mesopelagic fishes biomass and trophic efficiency in the open ocean. Nature Communications, 5(May 2013), 3271. https://doi.org/10.1038/ncomms4271

Januchowski-hartley, F. A., Vigliola, L., Maire, E., Kulbicki, M., & Mouillot, D. (2020). Low fuel cost and rising fish price threaten coral reef wilderness. Conservation Letters, 13, 1–9.

Kerry, C. R., Exeter, O. M., & Witt, M. J. (2022). Monitoring global fishing activity in proximity to seamounts using automatic identification systems. Fish and Fisheries, (January), 1–17. https://doi.org/10.1111/faf.12647

Kulbicki, M., Parravicini, V., & Mouillot, D. (2015). Patterns and processes in reef fish body size. In C. Mora (Ed.), Ecology of Fishes on Coral Reefs (pp. 104–115). https://doi.org/10.1017/CBO9781316105412.013

Kvile, K. O., Taranto, G. H., Pitcher, T. J., & Morato, T. (2014). A global assessment of seamount ecosystems knowledge using an ecosystem evaluation framework. Biological Conservation, 173, 108–120. https://doi.org/10.1016/j.biocon.2013.10.002

Langlois, T. J., Goetze, J., Bond, T., Monk, J., Abesamis, R. A., Asher, J., … Harvey, E. S. (2020). A field and video annotation guide for baited remote underwater stereo-video surveys of demersal fish assemblages. Methods in Ecology and Evolution, 11, 1401–1409.

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. https://doi.org/10.1093/nar/gkq967

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(8), e3000366. https://doi.org/10.1371/journal.pbio.3000366

Levin, N., Kark, S., & Danovaro, R. (2018). Adding the Third Dimension to Marine Conservation. Conservation Letters, 11(3), 1–14. https://doi.org/10.1111/conl.12408

Lin, T., Akamatsu, T., Sinniger, F., & Harii, S. (2021). Exploring coral reef biodiversity via underwater soundscapes. Biological Conservation, 253(June 2020), 108901. https://doi.org/10.1016/j.biocon.2020.108901

Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. PeerJ, 3, 1–12.

Maire, E., Cinner, J., Velez, L., Huchery, C., Mora, C., Dagata, S., … Mouillot, D. (2016). How accessible are coral reefs to people? A global assessment based on travel time. Ecology Letters, 19(4), 351–360.

Manea, E., Bianchelli, S., Fanelli, E., Danovaro, R., & Gissi, E. (2020). Towards an Ecosystem-Based Marine Spatial Planning in the deep Mediterranean Sea. Science of the Total Environment, 715, 136884. https://doi.org/10.1016/j.scitotenv.2020.136884

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, 43, 1–12.

Marques, V., Milhau, T., Albouy, C., Dejean, T., Manel, S., Mouillot, D., & Juhel, J. (2020). GAPeDNA : Assessing and mapping global species gaps in genetic databases for eDNA metabarcoding. Diversity and Distributions, 00, 1–13. https://doi.org/10.1111/ddi.13142

Martin., M. (1994). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal, 17(1), 10–12.

Mathon, L., Marques, V., Mouillot, D., Albouy, C., Andrello, M., Baletaud, F., … Vigliola, L. (2022). Cross-ocean patterns and processes in fish biodiversity on coral reefs through the lens of eDNA metabarcoding. Proceedings of the Royal Society B, 289, 20220162.

Maureaud, A. A., Hodapp, D., van Denderen, P. D., Hillbrand, H., Gislason, H., Spaanheden Dencker, T., … Lindegren, M. (2019). Biodiversity - ecosystem functionning relationships in fish communities: biomass is related to evenness and the environment, not to species richness. Proceedings of the Royal Society B, 286, 1–9.

Maxwell, S. L., Cazalis, V., Dudley, N., Hoffmann, M., Rodrigues, A. S. L., Stolton, S., … Watson, J. E. M. (2020). Area-based conservation in the twenty-first century. Nature, 586(7828), 217–227. https://doi.org/10.1038/s41586-020-2773-z

Mazzei, E. F., Pinheiro, H. T., Simon, T., Moura, R. L., Macieira, R. M., Pimentel, C. R., … Joyeux, J.-C. (2021). Mechanisms of dispersal and establishment drive a stepping stone community assembly on seamounts and oceanic islands. Marine Biology, 168(7). https://doi.org/10.1007/s00227-021-03919-7

McClanahan, T. R. (2021). Marine reserve more sustainable than gear restriction in maintaining long-term coral reef fisheries yields. Marine Policy, 128(November 2020), 104478. https://doi.org/10.1016/j.marpol.2021.104478

McLean, D. L., Taylor, M. D., Partridge, J. C., Gibbons, B., Langlois, T. J., Malseed, B. E., … Bond, T. (2018). Fish and habitats on wellhead infrastructure on the north west shelf of Western Australia. Continental Shelf Research, 164, 10–27. https://doi.org/10.1016/j.csr.2018.05.007

Mejía-Mercado, B. E., & Baco, A. R. (2022). Characterization and spatial variation of the deep-sea fish assemblages on Pioneer Bank, Northwestern Hawaiian Islands. Marine Ecology Progress Series, 692, 99–118. https://doi.org/10.3354/meps14071

Mejía-Mercado, B. E., Mundy, B., & Baco, A. R. (2019). Variation in the structure of the deep-sea fish assemblages on Necker Island, Northwestern Hawaiian Islands. Deep Sea Research Part I: Oceanographic Research Papers, (January), 103086. https://doi.org/10.1016/j.dsr.2019.103086

Morato, T., Hoyle, S. D., Allain, V., & Nicol, S. J. (2010). Seamounts are hotspots of pelagic biodiversity in the open ocean. Proceedings of the National Academy of Sciences, 107(21), 9707–9711. https://doi.org/10.1073/pnas.0910290107

Morato, T., Varkey, D. A., Damaso, C., Machete, M., Santos, M., Prieto, R., … Pitcher, T. J. (2008). Evidence of a seamount effect on aggregating visitors. Marine Ecology Progress Series, 357(Fonteneau 1991), 23–32. https://doi.org/10.3354/meps07269

Muff, M., Jaquier, M., Marques, V., Ballesta, L., Deter, J., Bockel, T., … Pellissier, L. (2022). Environmental DNA highlights fish biodiversity in mesophotic ecosystems. Environmental DNA, 00(August), 1–17. https://doi.org/10.1002/edn3.358

Pacoureau, N., Rigby, C. L., Kyne, P. M., Sherley, R. B., Winker, H., Carlson, J. K., … Dulvy, N. K. (2021). Half a century of global decline in oceanic sharks and rays. Nature, 589(7843), 567–571. https://doi.org/10.1038/s41586-020-03173-9

Perrot, Y., Brehmer, P., Habasque, J., Roudaut, G., Behagle, N., Sarré, A., & Lebourges-Dhaussy, A. (2018). Matecho: An Open-Source Tool for Processing Fisheries Acoustics Data. Acoustics Australia, 46(2), 241–248. https://doi.org/10.1007/s40857-018-0135-x

Pitcher, T. J., Clark, M. R., Morato, T., & Watson, R. (2010). Seamount fisheries: Do they have a future? Oceanography, 23(1), 134–144. https://doi.org/10.5670/oceanog.2010.66

Polanco F, A., Waldock, C., Keggin, T., Marques, V., Valentini, A., Dejean, T., … Vermeij, M. (2022). Ecological indices from environmental DNA to contrast coastal reefs under different anthropogenic pressures. Ecology and Evolution, 12, 1–17.

Proud, R., Cox, M. J., Le Guen, C., & Brierley, A. S. (2018). Fine-scale depth structure of pelagic communities throughout the global ocean based on acoustic sound scattering layers. Marine Ecology Progress Series, 598, 35–48.

Quattrini, A. M., Demopoulos, A. W. J., Singer, R., Roa-Varon, A., & Chaytor, J. D. (2017). Demersal fish assemblages on seamounts and other rugged features in the northeastern Caribbean. Deep-Sea Research Part I, 123(March), 90–104. https://doi.org/10.1016/j.dsr.2017.03.009

Roger, J. (2020). Données bathymétriques et topographiques de Nouvelle- Calédonie : Réalisation d'un MNT terre- mer pour l'étude de l'aléa tsunami (projet TSUCAL).

Rogers, A. D. (2018). The Biology of Seamounts: 25 Years on. In Advances in Marine Biology (1st ed., Vol. 79). https://doi.org/10.1016/bs.amb.2018.06.001

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. PeerJ, 4, 1–22. https://doi.org/10.7717/peerj.2584

Romagosa, M., Lucas, C., Pérez-Jorge, S., Tobeña, M., Lehodey, P., Reis, J., … Silva, M. A. (2020). Differences in regional oceanography and prey biomass influence the presence of foraging odontocetes at two Atlantic seamounts. Marine Mammal Science, 36(1), 158–179. https://doi.org/10.1111/mms.12626

Rowden, A. A., Schlacher, T. A., Williams, A., Clark, M. R., Stewart, R., Althaus, F., … Dowdney, J. (2010). A test of the seamount oasis hypothesis: Seamounts support higher epibenthic megafaunal biomass than adjacent slopes. Marine Ecology, 31(SUPPL. 1), 95–106. https://doi.org/10.1111/j.1439-0485.2010.00369.x

Sala, E., Lubchenco, J., Grorud-colvert, K., Novelli, C., Roberts, C., & Sumaila, U. R. (2018). Assessing real progress towards effective ocean protection. Marine Policy, 91, 11–13.

Sales, N. G., Wangensteen, O. S., Carvalho, D. C., Deiner, K., Praebel, I., McDevitt, A., & Mariani, S. (2021). Space-time dynamics in monitoring neotropical fish communities using eDNA metabarcoding. Science of the Total Environment, 754, 1–14.

Salvetat, J., Bez, N., Habasque, J., Lebourges-Dhaussy, A., Lopes, C., Roudaut, G., … Bertrand, A. (2022). Comprehensive spatial distribution of tropical fish assemblages from multifrequency acoustics and video fulfils the island mass effect framework. Scientific Reports, 12(1), 1–24. https://doi.org/10.1038/s41598-022-12409-9

Smith, K. F., & Brown, J. H. (2002). Patterns of diversity, depth range and body size among pelagic fishes along a gradient of depth. Global Ecology and Biogeography, 11(4), 313–322. https://doi.org/10.1046/j.1466-822X.2002.00286.x

Taylor, R. B., & Willis, T. J. (1998). Relationships amongst length, weight and growth of north-eastern New Zealand reef fishes. Marine Freshwater Resource, 49, 255–260.

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

Venegas-Li, R., Levin, N., Possingham, H., & Kark, S. (2018). 3D spatial conservation prioritisation: Accounting for depth in marine environments. Methods in Ecology and Evolution, 9(3), 773–784. https://doi.org/10.1111/2041-210X.12896

Wagner, D., van der Meer, L., Gorny, M., Sellanes, J., Gaymer, C. F., Soto, E. H., … Morgan, L. E. (2021). The Salas y Gómez and Nazca ridges: A review of the importance, opportunities and challenges for protecting a global diversity hotspot on the high seas. Marine Policy, 126(December 2020), 104377. https://doi.org/10.1016/j.marpol.2020.104377

Whitmarsh, S. K., Fairweather, P. G., & Huveneers, C. (2017). What is Big BRUVver up to? Methods and uses of baited underwater video. Reviews in Fish Biology and Fisheries, 27(1), 53–73. https://doi.org/10.1007/s11160-016-9450-1

Yan, H. F., Kyne, P. M., Jabado, R. W., Leeney, R. H., Davidson, N. K., Derrick, D. H., … Dulvy, N. K. (2021). Overfishing and habitat loss drives range contraction of iconic marine fishes to near extinction. Science Advances, 7, 1–10.

Yesson, C., Clark, M. R., Taylor, M. L., & Rogers, A. D. (2011). The global distribution of seamounts based on 30 arc seconds bathymetry data. Deep-Sea Research Part I: Oceanographic Research Papers, 58(4), 442–453. https://doi.org/10.1016/j.dsr.2011.02.004

Discussion générale

1. Synthèse des résultats

1.1. Assemblage d'un pipeline bio-informatique performant

Les estimations de diversité et de composition spécifique des communautés obtenues par *metabarcoding* de l'ADN environnemental dépendent grandement des traitements et nettoyages bio-informatiques appliqués aux données brutes (Calderón-Sanou et al. 2020). Or, de très nombreux programmes et pipelines ont été développés au cours des dernières années, sans qu'une comparaison approfondie de leurs performances ait été réalisée (Gardner et al. 2019; Prodan et al. 2020). Dans le **chapitre 2** de ma thèse j'ai ainsi comparé les performances des programmes et pipelines les plus fréquemment utilisés pour l'analyse de données ADNe de poissons à partir de données simulées et réelles. Cette performance a été évaluée selon 3 critères : (i) la sensibilité et la *F-measure* se rapportant aux espèces détectées, (ii) l'erreur quadratique moyenne (*root mean square error*, RMSE) entre abondances attendues et observées et (iii) le temps d'exécution.

Les résultats de cette comparaison m'ont amenée à sélectionner les meilleurs programmes permettant de retrouver la composition initiale de chaque communauté ainsi que les structures d'abondance de *reads* des communautés de poissons simulées et réelles. Les programmes de la suite OBITools (Boyer et al. 2016) obtiennent des sensibilités et des *F-measure* élevées mais nécessitent des temps d'exécution beaucoup plus longs que les autres. L'étape d'assignation taxonomique est la seule à présenter des différences significatives entre les indices pour chaque programme, tandis que le temps d'exécution est le facteur déterminant pour toutes les autres étapes de traitement. Les résultats sont similaires entre les jeux de données simulés et réels. Cette étude a donc abouti à des recommandations sur les programmes à utiliser pour chacune des étapes : VSEARCH (Rognes et al. 2016) pour l'assemblage des *reads*, la déréplication des séquences identiques ainsi que pour l'assignation taxonomique, Cutadapt (Martin, 1994) pour le démultiplexage et Swarm (Mahé et al. 2015) pour le nettoyage des erreurs de PCR et séquençage. La comparaison des pipelines complets montre que Barque obtient une sensibilité et une *F-measure* aussi élevées que celles du pipeline assemblé constitué des meilleurs programmes individuels. Ces deux pipelines produisent également les estimations les plus précises des abondances relatives de *reads*. Cette étude permet donc de recommander

l'utilisation de Barque ou d'un pipeline constitué de programmes performant, dans le contexte considéré, soit des données d'ADNe de poissons, sur le fragment teleo du gène mitochondrial 12S, avec une base de référence complète. De nouveaux programmes bio-informatiques sont régulièrement développés, par exemple eDNAFlow (Mousavi-Derazmahalleh et al. 2021) ou CoMa (Hupfauf et al. 2020). Il serait intéressant de les considérer dans une future comparaison plus exhaustive.

Afin d'optimiser la vitesse et la sensibilité du traitement des séquences brutes en s'affranchissant du choix de programmes et de paramètres, il est possible d'avoir recours aux réseaux neuronaux convolutifs (CNN). Une étude de Flück et al. (2022), à laquelle j'ai collaboré (voir Annexe 1), a évalué la capacité des CNN à analyser de courtes séquences d'ADNe (60pb sur le gène mitochondrial 12S) collectées dans des fleuves d'Amérique du Sud et à leur attribuer une assignation taxonomique. Le traitement des séquences par les CNN a fourni des résultats très comparables à ceux obtenus avec le pipeline OBITools, en termes d'assignation taxonomique et de sensibilité, mais en un temps d'analyse 150 fois plus rapide. Étant donné les bonnes performances des CNN dans l'écosystème très diversifié considéré ici, le développement de CNN plus élaborés pourrait permettre un déploiement rapide pour les futurs inventaires de biodiversité utilisant l'ADNe. Cette méthode est cependant toujours dépendante d'une base de référence suffisamment complète, ce qui limite actuellement son utilisation à large échelle.

En l'absence de base de référence génétique exhaustive, d'autres méthodes ont été développées. Le regroupement des séquences en unités taxonomiques moléculaires (MOTU) permet d'estimer le nombre d'espèces présentes, sans nécessairement parvenir à associer un nom d'espèce sur chaque MOTU. Cette méthode est couramment utilisée pour l'étude des bactéries ou plancton (De Vargas et al. 2015). Le pipeline bio-informatique paramétré par Marques, Guérin, et al. (2020) permet d'obtenir des estimations de diversité de MOTUs très proches du nombre réel d'espèces. Cette approche permet de réaliser des études de biodiversité à large échelle, comparant la diversité et la composition des communautés entre régions, sans avoir à mettre un nom sur chaque espèce.

1.2. L'ADNe pour décrire la biodiversité à large échelle

En milieu marin, les changements de biodiversité en réponse aux changements environnementaux et aux perturbations humaines sont plus rapides qu'en milieu terrestre

(Herbert-Read et al. 2022; McLean et al. 2018). Les déplacements des populations de poissons pour suivre leurs niches sont également plus rapide dans l'océan que sur terre (Lenoir et al. 2020). Les méthodes conventionnelles (pêche, caméras, plongée…) ne sont pas adaptées pour étudier et surveiller ces changements d'écosystèmes car elles sont trop longues ou trop onéreuses pour être déployées à large échelle spatio-temporelle. Le *metabarcoding* de l'ADNe peut être une solution adaptée à de tels suivis, si son avantage par rapport aux autres méthodes est démontré (rapidité, meilleure détection). Bien que de nombreuses études aient comparé les performances de l'ADNe et des méthodes conventionnelles en milieu d'eau douce (Cilleros et al. 2019; Gehri et al. 2021; McColl-Gausden et al. 2021) et en milieu marin tempéré (He et al. 2022; Stoeckle et al. 2021), très peu se sont intéressées au milieu marin tropical. Ces précédentes recherches ont démontré l'efficacité du *metabarcoding* de l'ADNe, la plupart détectant une diversité de poisson similaire ou plus élevée avec l'ADNe qu'avec les méthodes conventionnelles (Keck et al. 2022). En milieu marin tropical, quelques études récentes au niveau régional montrent que les estimations de diversité obtenues par le *metabarcoding* de l'ADNe surpassent celles des méthodes conventionnelles et permettent même de détecter de nouveaux taxa (Juhel et al. 2022; Nguyen et al. 2020; Polanco et al. 2021; Valdivia-Carrillo et al. 2021). La faible complétude des bases de référence pour ces environnements très divers ne permet l'identification que d'un nombre limité d'espèces détectées, mais l'ADNe permet tout de même la détection de davantage de MOTUs, genres et familles. Cependant, aucune étude à large échelle n'a été menée pour valider l'efficacité du *metabarcoding* de l'ADN environnemental en milieu marin tropical.

L'étude présentée au **chapitre 3** apporte une première comparaison entre l'ADNe et une méthode conventionnelle pour l'étude des poissons coralliens à large échelle. Les données obtenues de 251 échantillons d'ADNe collectés dans 5 régions tropicales et 3 océans, sur une durée de 28 mois sont comparées avec celles provenant de plus de 2000 transects de recensements visuels en plongée, dans 20 régions tropicales, sur une durée de 13 ans. L'ADNe détecte une diversité de 2023 MOTUs et 126 familles, tandis que les recensements visuels détectent 1818 espèces réparties dans 96 familles. Parmi les taxa mieux détectés par l'ADNe on retrouve de nombreux poissons cryptobenthiques (vivant sur le fond ou cachés dans les coraux) ainsi que des poissons pélagiques tels que des Scombridae et des requins. Ces taxa sont moins facilement observables par les plongeurs car ils vivent cachés et sont souvent furtifs et farouches. Or, les espèces cryptobenthiques jouent un rôle crucial dans le fonctionnement des récifs coralliens, en favorisant la production de biomasse et en alimentant la dynamique

trophique du récif (Brandl et al. 2019). Les espèces transitoires, pélagiques et d'eaux profondes peuvent également être très importantes pour le fonctionnement des récifs, à travers les stades larvaires pélagiques ou la migration nocturne le long de la pente récifale (Beckley et al. 2019; Morais & Bellwood, 2019), mais leur présence et leur rôle doivent encore être étudiés plus en profondeur. Les familles les plus diverses détectées avec l'ADNe sont les mêmes que celles détectées en plongée : les Gobiidae, Labridae et Pomacentridae. La famille des Gobiidae contient ≃1900 espèces décrites, dans plus de 200 genres (Froese & Pauly, 2000). Ces espèces étant très peu représentées dans les bases de référence, il est difficile de les identifier avec certitude avec l'ADNe, mais l'utilisation de MOTU permet cependant d'estimer le nombre d'espèces de chaque famille dans nos échantillons. Les recensements visuels ont identifié de nombreuses familles non détectées, ou non identifiées, par l'ADNe, telles que les Acanthuridae, Blenniidae, Caesionidae, Chaenopsidae, Chaetodontidae, Labrisomidae, Labridae ou Microdesmidae. Cette détection limitée par l'ADNe peut être due à la très faible représentation de ces familles dans les bases de référence 12S (entre 0 et 12%), ou à la faible résolution du marqueur teleo pour les espèces de ces familles, de sorte que plusieurs espèces peuvent partager la même séquence et être regroupées sous le même MOTU.

L'estimation de la diversité en MOTU permet également de revisiter les patrons de distribution biogéographique des poissons, à différentes échelles. On observe ainsi un gradient longitudinal où la richesse en MOTU décroit en s'éloignant du centre du Triangle de Corail (entre la Malaisie, l'Indonésie, les Philippines et les Iles Salomon). La partition de la β-diversité entre différentes échelles spatiales a mis en évidence que la diversité globale en MOTUs était principalement due à la β-diversité inter-régionale. Ce résultat démontre le rôle prédominant des processus régionaux dans la diversification et le maintien de la diversité d'espèces. Cette étude a également détecté un patron bien connu : la différenciation de la faune des Caraïbes par rapport aux autres zones tropicales, qui peut s'expliquer par l'isolation de la zone Caraïbes par l'Isthme de Panama et par l'activité tectonique (Bender et al. 2017; Leprieur et al. 2016). À plus fine échelle, l'ADNe a également permis de détecter des variations de composition de communautés entre récifs ou îles voisines. Ce résultat pourrait témoigner d'une plus grande hétérogénéité entre les récifs à échelle locale par rapport à ce qui était connu jusqu'à présent. Ces patrons de diversité et distribution des poissons sont également retrouvés dans l'étude du **chapitre 4**, réalisée à échelle globale (584 échantillons d'ADNe répartis dans 263 stations à travers les océans du monde). Dans cette étude nous mesurons la diversité de chaque communauté en termes de nombre de MOTUs et par un indice de diversité de séquences, dérivé

des nombres de Hill (Alberdi & Gilbert, 2019; Chao et al. 2019). Cet indice, original, capture la similarité entre séquences, à partir de la distance génétique. La diversité en MOTUs suit, ici aussi, le gradient de diversité longitudinal, du centre du Triangle de Corail aux Caraïbes (Bellwood & Hughes, 2001; Parravicini et al. 2013), mais aussi un gradient latitudinal, des tropiques vers les pôles (Freeman & Pennell, 2021). Bien que la diversité en séquences montre globalement les mêmes gradients, certaines stations de zones tempérées ou polaires révèlent une diversité en séquences intermédiaire du fait de la présence de genres très distincts (ex : Liparis, Lycodes et Somniosus dans l'Arctique). Les analyses de β-diversité de cette étude produisent également des patrons attendus. La faune de l'Antarctique est très différente des autres régions, du fait de l'isolement du continent par le courant circumpolaire et du très fort taux d'endémisme d'espèces benthiques (Crame, 2018; Eastman, 2005). De même, on détecte une différenciation de la faune des Caraïbes et du Pacifique Est par rapport aux autres régions tropicales, cependant, la composition en séquences de la région Caraïbes est proche des autres régions tropicales, car bien que les espèces soient différentes, elles sont issues des mêmes lignées. En concordance avec les résultats du **chapitre 3**, les estimations de diversités régionales mesurées avec l'ADNe dans cette étude sont très similaires aux estimations des listes régionales ou aux estimations obtenues par d'autres méthodes plus conventionnelles, ce qui conforte l'interprétation des résultats.

Décrire la biodiversité seulement en termes de nombre et d'identité des espèces peut fournir une bonne vision de l'état de santé des écosystèmes et permet de suivre les évolutions dans le temps et l'espace, mais cela n'est pas suffisant pour étudier le fonctionnement d'une communauté et la résilience d'un écosystème. C'est pourquoi il est important de considérer également la diversité fonctionnelle et phylogénétique lors des suivis de biodiversité et de la mise en place de stratégies de conservation (Auber et al. 2022). Mesurer ces diversités est possible avec l'ADNe si toutes les séquences sont assignées à une espèce (Marques et al. 2021). Or cela n'est pas le cas à large échelle car les bases de références sont encore incomplètes. Dans le **chapitre 4** j'ai donc étudié la corrélation entre les indices de diversité de séquences α et β, calculés à partir des nombres de Hill et les mêmes indices calculés sur les diversités fonctionnelle et phylogénétique. En ne me basant que sur les séquences assignées jusqu'à l'espèce, j'ai montré que les indices de diversité de séquences α et β, même sur de si courtes séquences, sont de bons proxys de la diversité phylogénétique et dans une moindre mesure, de la diversité fonctionnelle. Ce nouveau résultat offre des perspectives majeures pour l'utilisation de barcodes d'ADNe pour quantifier la diversité phylogénétique et fonctionnelle et ainsi suivre

le fonctionnement des écosystèmes (Duffy et al. 2016) et l'histoire évolutive des communautés (McLean, Stuart-Smith, et al. 2021).

Les études des **chapitres 3 et 4** suggèrent que le *metabarcoding* de l'ADNe est donc une méthode fiable pour détecter la biodiversité présente dans des zones très riches à large échelle ainsi que pour étudier les changements de composition des communautés (remplacement d'espèces ou familles par d'autres). Cette méthode est plus efficace en termes d'effort d'échantillonnage que les méthodes conventionnelles et sa principale limitation reste la couverture taxonomique des bases de références.

1.3. Impact des pressions humaines sur la diversité des poissons

De nombreuses études ont démontré l'impact des conditions environnementales sur la diversité et la composition des communautés (Antão et al. 2020; Brandl et al. 2020; Brown et al. 2022; Jouffray et al. 2019; Stuart-Smith et al. 2022), mais de plus en plus, l'activité humaine impacte la biodiversité, dans toutes les régions océaniques (Andrello et al. 2022; Halpern et al. 2019; O'Hara et al. 2021). Le **chapitre 3** a démontré l'efficacité du *metabarcoding* de l'ADNe pour étudier les changements de diversité à de larges échelles spatiales. L'analyse des facteurs qui influencent ces changements permettra de cibler les zones à surveiller attentivement et celles à protéger prioritairement.

L'étude du **chapitre 4** est la première réalisée à une échelle globale pour étudier les patrons de diversité α et β des poissons marins en réponse à des facteurs environnementaux, socio-économiques et géographiques. Les résultats montrent que la diversité en MOTUs et en séquences de poissons est influencée en premier lieu par les facteurs environnementaux. Ces facteurs environnementaux et géographiques incluent la température de surface, la productivité et la distance au Triangle de Corail. Cette relation peut s'expliquer par l'hypothèse de « stabilité climatique » qui stipule que les zones les plus chaudes, sous les tropiques, ont connu une variabilité historique moindre des conditions climatiques. Ceci aurait permis de conserver une large surface d'habitats coralliens productifs abritant de nombreuses espèces et de fort taux de spéciation, constituant ainsi un refuge au cours du Quaternaire. Les zones plus froides, quant à elles, auraient été très instables et auraient subi des déclins de diversité le long des gradients de températures (Mittelbach et al. 2007; Pellissier et al. 2014). Cette relation température – diversité peut également s'expliquer par l'hypothèse de « vitesse d'évolution », selon laquelle les températures élevées favorisent les taux de métabolisme, reproduction et spéciation (Fine,

2015; Harmelin-Vivien, 2002), mais cette hypothèse est actuellement débattue (Rabosky et al. 2018).

On observe également une relation négative entre la diversité des poissons côtiers et les pressions humaines, ce qui suggère que la distribution des séquences d'ADNe libérées par les poissons côtiers est, au moins en partie, façonnée par les activités humaines. La dépendance des pays aux ressources marines (pour la pêche ou les services écosystémiques) est le facteur socio-économique influençant le plus la diversité des poissons, aussi bien des cryptobenthiques que des poissons de grandes tailles. En effet, les populations humaines fortement dépendantes des ressources marines pour leur alimentation et leurs revenus utilisent des engins de pêche non sélectifs tels que le chalut, la dynamite, ou le poison, capturant même des espèces de petites tailles ou cryptobenthiques, menant à une détérioration des habitats clés tels que les récifs coralliens ou les herbiers (Batista et al. 2014; Fox et al. 2003; Munro, 1996). Les pressions de pêche pourraient donc avoir un impact sur toutes les classes de taille de poissons dans les pays fortement dépendants des ressources marines, en supprimant des parties entières du réseau trophique et en diminuant la diversité génétique du pool d'espèces restant. Concernant les poissons de grandes tailles, il a déjà été démontré que l'abondance et la richesse des espèces de requins, carangues, mérous et vivaneaux diminuent avec l'augmentation de la densité de population humaine (Dulvy et al. 2021), ce qui affecte les relations trophiques dans les écosystèmes surexploités (McClure et al. 2020). D'autres pressions humaines, telles que le rejet de nutriments ou de produits chimiques, la pollution et la dégradation de l'habitat, affectent les espèces les plus sensibles, qui appartiennent souvent aux mêmes lignées ou familles évolutives (Cinner et al. 2018; Dulvy et al. 2021). Nous observons donc une diminution globale de la diversité des poissons avec l'augmentation de la pression humaine, même si localement des tendances inverses peuvent émerger (Boulanger et al. 2021).

Cette étude démontre l'efficacité du *metabarcoding* de l'ADNe à détecter l'influence des facteurs humains et environnementaux sur la diversité des poissons marins ainsi qu'à localiser les zones où cette diversité est affectée. A l'aide de l'indice de diversité de séquences, corrélé aux indices de diversité phylogénétique et fonctionnelle, nous pourrions également estimer les impacts humains et environnementaux sur la diversité des fonctions et des lignées au sein des communautés. Cette méthode peut donc être utilisée pour apporter une expertise à la conservation et déterminer les zones à protéger en priorité, au regard de plusieurs métriques de diversité.

1.4. Modélisation et conservation de la biodiversité en 3D

Le **chapitre 4** démontre donc une influence de l'impact humain et notamment de la dépendance aux écosystèmes marins sur la diversité des poissons côtiers. Les réserves sont le meilleur outil pour protéger la biodiversité marine, mais l'étude de la diversité des habitats et des écosystèmes marins, y compris des zones profondes, permettrait d'évaluer leur emplacement optimal. Par exemple, il a été démontré que dans une région où les zones à proximité de l'homme sont les plus impactées, les grands prédateurs trouvent refuge sur les récifs éloignés des activités humaines et sur les monts sous-marins (Letessier et al. 2019). Dans le **chapitre 5**, nous étudions donc la diversité des poissons et Elasmobranches sur des habitats profonds, dans l'archipel Néo-Calédonien, avec diverses méthodes d'échantillonnage, afin de tester la nécessité ou non d'inclure ces habitats dans les plans de conservation.

Onze monts sous-marins et 4 pentes externes profondes ont été échantillonnés, par des prélèvements d'ADNe, des enregistrements vidéos et des enregistrements acoustiques sur chaque sommet de mont et pente, et des enregistrements acoustiques et prélèvement d'ADNe dans la zone pélagique autour des monts et pentes. A partir de ces données, j'ai mesuré la richesse taxonomique, la biomasse et l'abondance des poissons. La modélisation de ces métriques de diversité en fonction de facteurs environnementaux, géographiques et humains a permis de prédire la diversité des poissons à l'échelle de l'archipel Néo-Calédonien et de réaliser une planification spatiale en trois dimensions, en intégrant la profondeur.

Cette étude met en évidence que les différentes métriques de diversité mesurées sont principalement influencées par la profondeur, la température et l'éloignement à l'homme. Si l'influence de la température est variable entre les métriques, la proximité de l'homme a un effet négatif sur toutes les métriques mesurées. La richesse, la biomasse et l'abondance sont plus élevées dans les zones éloignées de la capitale, Nouméa. Les zones peu profondes telles que les monts sous-marins, dont le sommet se trouve dans la zone euphotique (0-200m), ou encore les pentes externes jusqu'à 200m de profondeur abritent la plus grande diversité de poissons et on y retrouve une faune identique à celle présente sur les récifs coralliens. Cependant, les monts sous-marins plus profonds abritent également une diversité remarquable, due à la présence d'espèces inféodées aux habitats plus profonds (ex : *Squalus megalops*). L'algorithme de priorisation spatiale en trois dimensions sélectionnant les cellules permettant de protéger 30% de chaque métrique de diversité a fourni plusieurs scénarios selon le degré de fragmentation de la solution. Le scénario avec la solution la plus fragmentée a sélectionné de

nombreuses cellules sur les monts sous-marins peu profonds et les pentes externes éloignées, mais aussi de nombreuses cellules dans les zones plus profondes. De nombreuses unités de planification ne sont sélectionnées que dans une ou deux couches de profondeur, dont 10% seulement dans la couche la plus profonde. Bien que cette solution puisse être difficile à mettre en œuvre dans un véritable plan de gestion, elle permet d'identifier les zones et les couches de profondeur qui abritent une grande biodiversité. L'intégration de l'abondance d'espèces individuelles dans l'algorithme de priorisation favorise la sélection de zones peu profondes et profondes et la sélection de tous les types d'habitats. La meilleure solution de planification en 3D, incluant une pénalité sur la fragmentation, privilégie la protection de grandes zones contiguës, dispersées sur tous les types d'habitats (pentes profondes autour de la Grande-Terre loin de Nouméa, pentes profondes autour des atolls et petites îles, sommets et pentes des monts sous-marins peu profonds et profonds) et sur toute la colonne d'eau, pour limiter la fragmentation verticale. La prise en compte de la continuité lors de la hiérarchisation des zones à conserver constitue la meilleure stratégie pour la persistance de la biodiversité (Goetze et al. 2021; Magris et al. 2018), mais dans notre cas d'étude, elle élimine de nombreuses zones, notamment les zones profondes des autres monts sous-marins et les pentes des îles. Dans les deux solutions, le Grand Lagon Nord, le récif d'Entrecasteaux, les atolls de Chesterfield et de Bellona, les monts sous-marins peu profonds (Capel et Fairway) et les monts sous-marins plus profonds (au sud de Grande-Terre) sont priorisés. La plupart de ces zones sont connues pour abriter une grande biodiversité (Letessier et al. 2019), mais ne sont pas encore fortement protégées (Claudet et al. 2021). En protégeant 30% de chaque métrique de diversité, la meilleure solution de planification protège une surface correspondant à 30% de la surface considérée dans notre étude. Cela démontre qu'en sélectionnant les zones à protéger en se basant sur les indices de diversité, il est possible d'atteindre des objectifs internationaux de protection des océans (CBD, 2021; Dinerstein et al., 2019).

Cette étude est l'une des rares à explorer conjointement la biodiversité sur les monts sous-marins et les pentes profondes, en combinant plusieurs méthodes d'échantillonnage (Annasawmy et al. 2019; Cherel et al. 2020; Mazzei et al. 2021; Mejía-Mercado et al. 2019; Quattrini et al. 2017). L'utilisation de ces diverses méthodes nous a permis de mesurer sept métriques de biodiversité complémentaires, qui ont chacune leurs zones de grande diversité, certaines communes à toutes les métriques mais d'autres propres à une seule métrique. De plus, notre plan d'échantillonnage et notre méthodologie nous ont permis de montrer que la biodiversité diffère entre les zones benthiques et pélagiques et que la profondeur est le facteur

le plus structurant, démontrant la nécessité d'intégrer une troisième dimension dans les planifications de conservation (Brito-Morales et al. 2022; G. A. Duffy & Chown, 2017; Manea et al. 2019; Venegas-Li et al., 2018). Une récente étude réalisée sur divers écosystèmes mésophotiques a également démontré l'efficacité de l'ADNe pour décrire les communautés de poissons en fonction de la profondeur (Muff et al. 2022). Etudier la diversité dans les zones profondes au préalable de la constitution d'un plan de conservation permettrait de définir la nécessité d'intégrer ou non ces zones dans les aires protégées. Il serait de ce fait possible de réaliser des plans de conservation stratifiés verticalement, qui doivent nous amener à couvrir plus de 30% des océans d'ici 2030 (Dinerstein et al. 2019) ou même 40% pour les espèces marines menacées (Jefferson et al. 2021). Protéger les zones profondes en haute mer permettrait de conserver des refuges de biodiversité, peu exposés au changement climatique, tout en limitant l'impact humain. Cependant, ce type de législation n'est pas encore mis en œuvre (Brito-Morales et al. 2022).

En vue d'approfondir cette étude pour obtenir une solution de planification plus réaliste, il serait nécessaire d'attribuer des coûts à chaque cellule, selon la facilité (zone éloignée, zone déjà protégée) ou la difficulté (zone de pêche ou de loisir) de les inclure dans la planification et de pondérer ainsi les solutions calculées par l'algorithme de priorisation. Prendre en compte les zones les plus menacées par les activités humaines, telle que la pêche au chalut, afin de les inclure prioritairement dans la solution de conservation serait également plus efficace (Epstein & Roberts, 2022). Il serait aussi intéressant d'étendre l'échantillonnage aux plaines abyssales et aux pentes plus profondes afin de mieux comprendre le rôle des monts sous-marins dans les vastes océans.

2. Limitations

2.1. Biais d'échantillonnage

L'étude de la biodiversité à de grandes échelles spatiales, comme c'est le cas dans les études des **chapitres 3 et 4**, requiert du temps pour la mise en place de collaborations, la conception des missions d'échantillonnage, ou encore l'obtention des permis de prélèvement. Ainsi, les missions de collecte des données analysées dans les **chapitres 3 et 4** ont été réalisées sur plusieurs années, entre 2017 et 2020. Durant cette période, les technologies d'échantillonnage et d'analyse de l'ADNe ont évolué, nos connaissances sur la distribution et la persistance de l'ADNe en milieu marin se sont améliorées et la littérature s'est étoffée et a

apporté de nouvelles recommandations pour le prélèvement et l'analyse de l'ADNe (Stauffer et al. 2021). Il est donc naturel que nos stratégies d'échantillonnage aient évolué au fil des missions, afin d'obtenir les inventaires et les indices de diversité les plus exhaustifs et vraisemblables. Cependant, il en résulte un déséquilibre et une inconsistance dans les méthodes d'échantillonnage, le nombre de réplicas par station, le volume filtré, etc. Durant les premières missions, en Indonésie, l'eau de mer a été prélevée ponctuellement, par des plongeurs, dans des sacs stériles d'une contenance de 2L et filtrée sur des filtres Sterivex. La plupart des échantillons des autres localités ont été prélevés à l'aide de pompes péristaltiques filtrant 30L le long d'un transect en surface, ou à l'aide bouteille Niskin pour les prélèvements en profondeur. Enfin, les échantillonnages les plus récents ont été réalisés par transects en profondeur, proche du substrat pour capturer l'ADNe au plus près des organismes. Cette méthode, facilitée par le développement de pompes submersibles et autonomes, permet de détecter un plus grand nombre d'organismes benthiques et vivant en profondeur (Muff et al. 2022).

Cette irrégularité de méthodologie a été intégrée dans nos modèles pour prendre en compte et corriger son effet sur les estimations de diversité et mesurer les effets partiels des autres facteurs en jeu. Cependant, il serait plus judicieux pour les études futures de mettre en place un échantillonnage standardisé, utilisant les mêmes designs et outils pour la filtration de l'eau, de grands volumes et le même nombre de réplicas par station, pour obtenir un effort d'échantillonnage comparable entre les différentes localités. De l'expérience acquise des échantillonnages précédents, je conseillerais de filtrer au moins 30L par réplica à l'aide de pompes submersibles, au plus près du substrat (même pour les zones profondes), et de réaliser plusieurs réplicas par station pour s'assurer de détecter la biodiversité avec fiabilité.

2.2. Base de référence incomplète – estimation de la diversité en MOTUs

A l'échelle globale, les bases de référence publiques sont très incomplètes (Marques, Milhau, et al. 2020), notamment pour le gène mitochondrial 12S. Moins de 15% des ≃32000 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. 2020). Le pipeline performant assemblé au **chapitre 2** n'est donc pas approprié pour les études à large échelle, car il n'a été testé qu'en présence d'une base de référence complète. J'ai donc utilisé le pipeline développé par Marques, Guérin, et al. (2020), groupant les séquences en unités taxonomiques moléculaires (MOTUs). La richesse en MOTUs est cependant à considérer comme une approximation de la

richesse spécifique réelle, car elle dépend de la divergence moléculaire inter- et intraspécifique. Un MOTU étant construit en regroupant des séquences différentes d'au plus un nucléotide, il peut contenir des séquences appartenant à des espèces différentes ayant une faible variabilité interspécifique sur ce marqueur. A l'inverse, si la variabilité génétique intraspécifique est grande, plusieurs MOTUs peuvent correspondre à la même espèce. De plus, le calcul de la richesse en MOTUs est soumis à sous-estimation d'espèces rares et à la surestimation dues aux erreurs de PCR et de séquençage. Il est difficile de différencier une séquence très peu abondante appartenant à une espèce rare d'une erreur de PCR ou de séquençage. Une espèce rare représentée par une faible abondance de *reads* dans l'échantillon pourra être traitée comme une erreur dérivée d'une autre séquence très proche et très abondante. La richesse en MOTUs est donc influencée par les rapports d'abondance et d'occurrence entre les séquences ainsi que par leur proximité phylogénétique. L'ajustement des seuils de nettoyage permettrait de conserver plus d'espèces rares, mais augmenterait aussi le nombre de faux-positifs dus aux erreurs. Actuellement, nous utilisons un seuil de présence dans deux réplicas PCR minimum pour conserver une séquence, ce qui peut sous-estimer les espèces rares mais permet d'éliminer la plupart des erreurs PCR, qui ne se produisent rarement plus d'une fois. A partir des séquences assignées à l'espèce, il serait possible d'étudier la correspondance entre le nombre d'espèces et le nombre de MOTUs, de compter combien de MOTUs correspondent à la même espèce et de mesurer la variation moyenne entre les séquences rassemblées dans un MOTU.

Cette méthode d'estimation de la diversité en MOTUs fonctionne avec une base de référence incomplète, mais il est toujours préférable d'avoir une base la plus complète possible, afin de pouvoir assigner un maximum de séquences au minima à la famille et de pouvoir réaliser des analyses écologiques. Cependant, de nombreuses familles sont très peu, voire pas du tout, représentées dans les bases de références (ex : Bleniidae, Chaetodontidae, Liparidae), ce qui empêche l'assignation des séquences à la bonne famille. Il serait donc intéressant de compléter les bases de références en commençant par les familles les moins séquencées, dont l'identification reste pour l'instant quasi-impossible.

Ce pipeline bio-informatique basé sur les MOTUs n'a pour l'instant été développé et validé que sur le marqueur teleo qui cible les poissons, mais son efficacité sur d'autres marqueurs n'est pas garantie. Des tests et des ajustements des seuils seraient surement nécessaires. Ce pipeline peut également être amélioré en ajoutant ou remplaçant certains des algorithmes ou filtres. Des approches très similaires ont été utilisées pour étudier d'autres taxons et obtenir les estimations de diversité de MOTUs les plus proches de la réalité, comme par exemple la combinaison de

l'algorithme de nettoyage DADA2, de clustering avec Swarm et de nettoyage post-classification avec LULU, proposée par Brandt et al. (2021). Une autre approche consiste à estimer la diversité par les nombres de Hill qui donnent une mesure en unité d'espèces (ou unités de MOTUs dans notre cas). Ces indices peuvent être modulés pour donner plus ou moins de poids aux espèces rares ou abondantes, grâce au paramètre *q* (Alberdi & Gilbert, 2019; Mächler et al. 2021). Le nombre de Hill d'ordre 0 donne un poids égal à tous les MOTUs et revient donc à une richesse de MOTUs, donnant une importance égale aux MOTUs rares et aux MOTUs abondants. Le nombre de Hill d'ordre 1 prend en compte les abondances relatives exactes des MOTUs, tandis que le nombre de Hill d'ordre 2 donne plus de poids aux MOTUs les plus abondants. Ainsi, différents traitements bio-informatiques peuvent être appliqués, en fonction de la diversité que l'on cherche à mettre en valeur.

2.3. Biais de résolution taxonomique sur certains clades

Au-delà des possibles biais de sous-estimations ou surestimations du nombre de MOTUs dûs aux espèces rares et erreurs de PCR et de séquençage, il existe également des biais liés à la résolution taxonomique du marqueur choisi pour l'amplification des séquences. En effet, pour estimer avec fiabilité un nombre de MOTUs correspondant au nombre d'espèces présentes, il est nécessaire que chaque séquence ne corresponde qu'à une seule espèce. Or, pour certaines familles, le marqueur teleo que nous avons choisi n'est pas résolutif et de nombreuses espèces partagent la même séquence, comme pour les Cichlidae par exemple (Doble et al. 2019; Taberlet et al. 2018), ou encore pour le genre *Etelis*. Chaque marqueur possède des biais de résolution, ainsi le marqueur MiFish de Miya et al. (2020) est très peu résolutif pour le genre *Sebastes* (Gold et al. 2020). Ces biais de résolution sont notamment rencontrés dans des groupes taxonomiques ayant connu une diversification récente et dont les séquences mitochondriales n'ont pas suffisamment divergé pour présenter de la variabilité interspécifique. Cette faible résolution empêche donc de discerner la présence d'une espèce particulière par le *metabarcoding* et entraine des sous-estimations de diversité.

Des approches plus spécifiques sont possibles, notamment par le barcoding, si l'on souhaite cibler une espèce particulière, ou bien par l'utilisation de multiples marqueurs en *metabarcoding*. Cette dernière solution permet, pour un même échantillon, d'amplifier et séquencer plusieurs barcodes afin de détecter plusieurs groupes taxonomiques (ex : Chondrichtyens, Téléostéens, Vertébrés (Polanco et al. 2021), ou de compenser les biais de résolutions de chaque primer. Bylemans et al. (2018) ont comparé plusieurs paires de primers

fréquemment utilisées pour le *metabarcoding* des poissons et ont identifié que les primers les plus courts (MiFish, teleo) sont les plus efficaces car ils permettent de retrouver le plus d'espèces, mais sont également les moins résolutifs. Leur étude teste et valide un nouveau primer (12S AcMDB07) qui serait plus résolutif pour les poissons. Un étude plus récente a comparé plusieurs paires de primers sur les gènes 12S, 16S et 18S pour décrire les communautés de poissons en milieu estuarien très riche (Kumar et al. 2022). Cette étude démontre que les marqueurs Riaz_12S et Berry_16S détectent la plus grande richesse de poissons et de chondrichtyens. Il serait donc recommandé d'utiliser plusieurs marqueurs afin d'améliorer la détection des espèces. L'utilisation de multi-markers nécessite cependant de multiplier le nombre de PCR et de séquençages, ce qui augmente par conséquent le coût d'analyse.

2.4. ADNe peu informatif sur la biologie des espèces

Il faut garder à l'esprit que l'ADNe collecté dans l'environnement ne nous informe en aucun cas sur la biologie ou la physiologie des espèces. Il n'est pas possible d'obtenir des informations sur le sexe des individus, leur taille, leur stade de vie, leur comportement, etc. Il n'est pas non plus possible de savoir si l'individu adulte est présent dans le milieu, ou si seulement des larves sont détectées dû au transport larvaire. Toutes ces informations sont pourtant très importantes dans le cadre des études de dynamique des populations et pour la conservation. Un individu adulte n'a pas la même fonction au sein de la communauté et de l'écosystème, ni le même habitat, qu'une larve de la même espèce et cela influe donc sur le réseau trophique (Compaire et al. 2021; Guerreiro et al. 2021; Huang et al. 2021).

Le fonctionnement et la résilience d'un écosystème ne dépendent pas seulement du nombre d'espèces qu'il contient, mais également du nombre d'individus représentant chaque espèce. Une espèce représentée par un nombre insuffisant d'individus ne pourra plus remplir ses fonctions dans la communauté (Säterberg et al. 2013). Il est donc crucial d'intégrer un aspect quantitatif aux inventaires de diversité. En utilisant une PCR quantitative (qPCR) sur une espèce unique, en barcoding, il est possible d'établir une corrélation entre la concentration d'ADNe et l'abondance et la biomasse des individus de cette espèce (Nevers et al. 2018; Spear et al. 2020; Tillotson et al. 2018). Récemment, une méthodologie de comptage d'haplotypes sur la D-loop, à partir d'échantillon d'ADNe a permis de détecter 94% des haplotypes d'une population de thons rouges du Pacifique, en aquarium (Yoshitake et al. 2021). La méthodologie pour estimer l'abondance de plusieurs espèces à partir d'échantillons de *metabarcoding* d'ADNe est

cependant encore en développement (Nakagawa et al. 2022; Yao et al. 2022). Les caractéristiques physiologiques, métaboliques ou comportementales de chaque espèce peuvent affecter le taux de sécrétion d'ADNe (Sassoubre et al. 2016). Les amorces peuvent présenter des affinités différentes entre espèces durant la PCR, ce qui biaise le nombre de copies d'ADN réalisées à cette étape entre les espèces (Elbrecht & Leese, 2015). De plus, durant la PCR, les séquences d'ADN sont amplifiées de façon exponentielle et une légère stochasticité dans les premiers cycles de PCR peut résulter en de grandes variations dans la composition finale du nombre d'amplicons (Kelly et al. 2019). Sato et al. (2021) ont cependant identifié une corrélation entre le nombre total de copies d'ADNe dans un échantillon et l'intensité du signal acoustique relevé à la même localisation par un échosondeur. De même, Mariani et al. (2021) ont trouvé une corrélation entre l'abondance de *reads* et la fréquence d'occurrence pour plusieurs espèces de Chondrichtyens et ont donc pu estimer la proportion relative de chaque espèce dans le milieu, seulement pour des espèces taxonomiquement et physiologiquement similaires. Dans une review, Rourke et al. (2022) ont trouvé que sur 12 articles étudiant l'aspect quantitatif du *metabarcoding* de l'ADNe, certaines en milieu contrôlé et d'autres en milieu naturel, 11 rapportaient une corrélation positive entre le nombre de *reads* d'ADNe et l'abondance ou la biomasse des espèces de poissons. Depuis, d'autres études ont démontré les applications quantitatives de l'ADNe. En utilisant un séquençage sur une plateforme NovaSeq et une inférence de séquences exactes basée sur le modèle DADA2, Skelton et al. (2022) ont mis en évidence une forte corrélation entre le nombre de *reads* et l'aire totale représentée par la population de chaque espèce. Cette relation, cependant, est fortement biaisée par les espèces les plus abondantes. Pont et al. (2022) ont combiné *metabarcoding* et qPCR d'ADNe collecté en rivière et ont trouvé une corrélation entre la concentration d'ADNe par taxon et l'abondance spécifique absolue.

En attendant de futurs développements permettant de calculer les abondances exactes de chaque espèce à partir d'échantillons d'ADNe, il est nécessaire de combiner l'ADNe avec d'autres outils d'échantillonnage, tels que les transects visuels ou caméras, comme c'est le cas dans le **chapitre 5**, afin de pouvoir mesurer plusieurs aspects de la diversité (richesse, biomasse, abondance).

3. Perspectives

3.1. Innovations technologiques

De nombreuses innovations technologiques sont en train de voir le jour et contribueront à l'amélioration de l'efficacité de l'analyse de l'ADNe. Par exemple, de nouvelles approches d'échantillonnage passif ou naturel sont en cours de développement et permettraient de limiter le coût et la logistique de l'échantillonnage par filtration (Figure 6.1). C'est le cas des échantillonneurs passifs imprimés en 3D, faits d'hydroxyapatite, un minéral naturel ayant de fortes capacités d'absorption de l'ADNe, développé par (Verdier et al. 2022). Une récente étude a comparé l'efficacité de plusieurs types d'échantillonneurs passifs (filtre en nitrate de cellulose, filtres en nylon, filtres en nylon chargés positivement, éponges artificielles) associés à un protocole d'extraction in-situ (PDQeX) (Jeunen et al. 2022). Les résultats démontrent que les techniques de filtration passives utilisant des substrats poreux sont aussi efficaces que les filtrations actives en systèmes contrôlés, mais que les performances du système d'extraction in-situ sont moins bonnes que les protocoles d'extraction classiques en laboratoire. D'autres études utilisant des échantillonneurs passifs de ce type (substrats poreux, membranes filtrantes, biofilms, etc.) ont également démontré leurs bonnes performances (Bessey et al. 2021; Chen et al. 2022; Kirtane et al. 2020; Rivera et al. 2022). Les éponges commencent également à être utilisées pour la collecte d'ADNe. Grâce à leur rôle de filtreur naturel, elles permettent de réduire les coûts d'échantillonnage et de capter l'ADNe des poissons associés aux récifs coralliens, mais aussi de poissons pélagiques, profonds et migrateurs (Mariani et al. 2019; Turon et al. 2020).

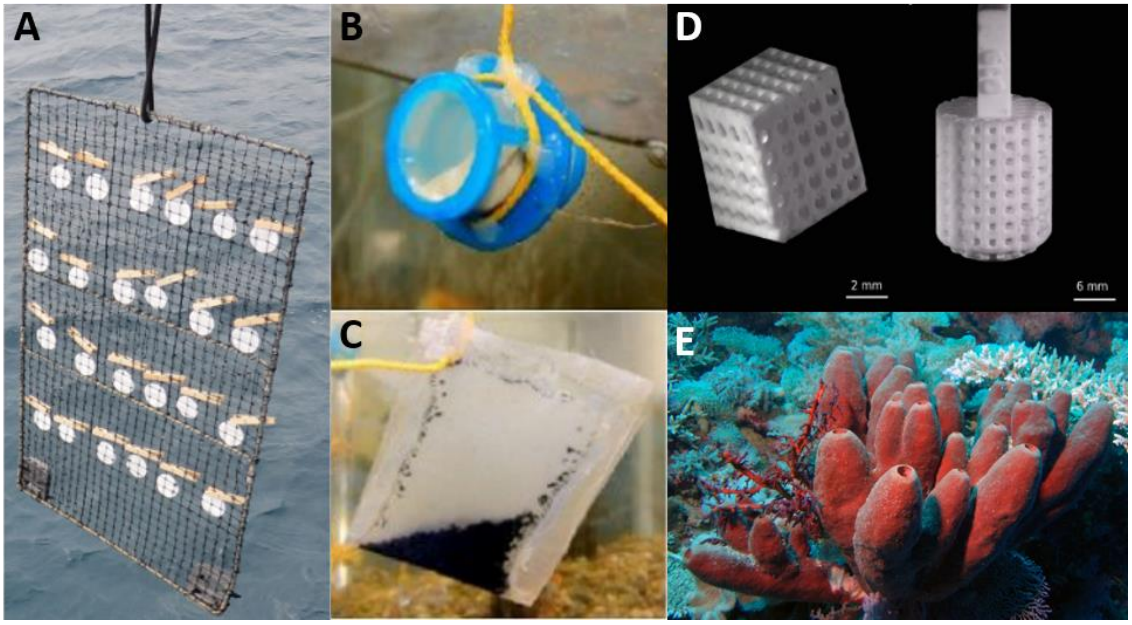


Figure 6.1. *Exemples d'échantillonneurs passifs et naturels. A) Filtres d'ester de cellulose non chargés et filtres de nylon chargés (figure extraite de Bessey et al. 2021), B) Echantillonneur contenant de l'argile Montmorillonite, C) Echantillonneur contenant du charbon actif (Figures extraites de Kirtane et al. 2020), D) Echantillonneurs en hydroxyapatite (Figure extraite de Verdier et al. 2022), F) Eponge du genre Porifera, utilisée comme filtreur naturel.*

En intégrant les facteurs relatifs à l'échantillonnage dans mes modèles au **chapitre 4**, j'ai mis en évidence que la richesse en MOTUs détectée était plus élevée avec des échantillonnages par transects que par bouteilles Niskin, très ponctuels. Les transects permettent de couvrir de plus grandes surfaces et sont donc plus intégrateurs dans la détection du signal ADNe. De même, j'ai montré que la profondeur d'échantillonnage avait une influence sur la richesse en MOTUs et notamment sur la détection des poissons cryptobenthiques. Ces espèces vivant cachées dans la structure du substrat, se déplacent peu dans la colonne d'eau et leur ADNe n'est que peu transporté. Il faut donc filtrer au plus près du substrat pour les détecter efficacement. Pour les futures missions d'échantillonnage, l'idéal serait donc de filtrer de grands volumes, le long de transects, en surface et près du substrat et d'effectuer plusieurs réplicas par localité. Les transects de surface sont faciles à réaliser depuis un bateau, mais les transects profonds demandent plus de logistique et de développements technologiques. Il est possible d'utiliser des pompes submersibles, développées dans le cadre d'une collaboration entre Spygen, Marbec et la société Andromède (Figure 6.2). Ces pompes peuvent être manipulées par des plongeurs, si la profondeur le permet, ou programmées pour se déclencher à distance. D'autres technologies ont été récemment développées, notamment un robot sous-marin permettant de coupler l'enregistrement de vidéos et la filtration d'ADNe, à de grandes profondeurs (McLean et al.

2020). De futurs développements technologiques permettront d'automatiser et faciliter les prélèvements en profondeur, avec par exemple l'utilisation de drones ou planeurs sous-marins auxquels on peut coupler des dispositifs de filtration d'ADNe. Ces engins permettent de collecter de nombreuses données océanographiques et sont autonomes sur de longs déplacements, ce qui allègerait la logistique pour les prélèvements en mer à de grandes profondeurs (Truelove et al. 2022).

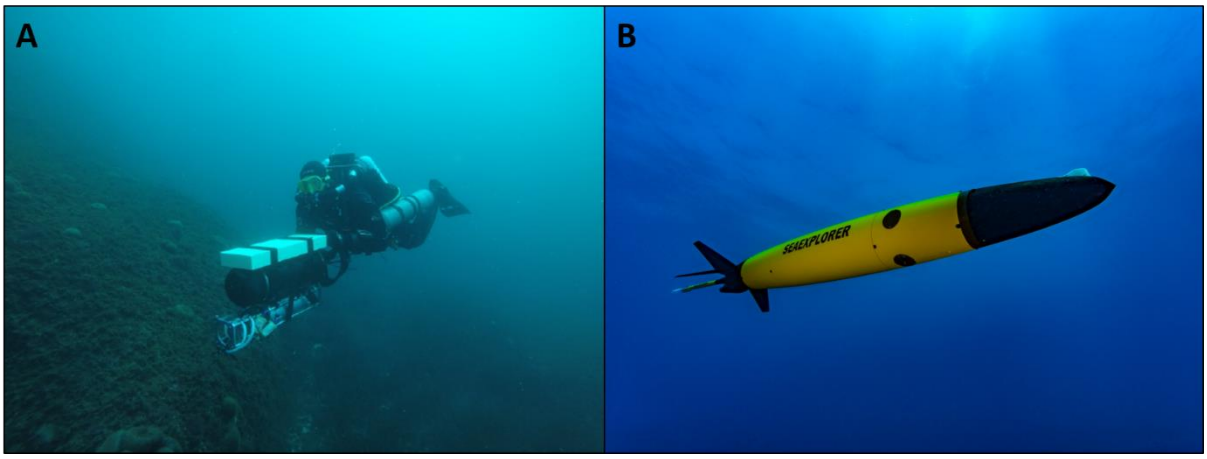


Figure 6.2. *Nouvelles technologies d'échantillonnage d'ADNe profond. A) Pompe submersible fixée sur un scooter sous-marin et dirigée par un plongeur, dans la zone mésophotique des calanques de Marseille (crédit : R. Hocdé, 2022). B) Drone sous-marin SeaExplorer (crédit : Alseamar).*

Concernant le traitement des échantillons en laboratoire, de nombreux développements sont également en cours, notamment concernant la quantification de l'ADN et la relation entre quantité d'ADNe et abondance des individus. Approfondir les études sur la corrélation entre le nombre de *reads* de chaque espèce dans un échantillon de *metabarcoding* et la biomasse ou l'abondance de ces espèces dans le milieu permettrait d'étudier les aspects quantitatifs de la diversité avec l'ADNe, jusqu'alors seulement possible par des méthodes conventionnelles. Une technique innovante consiste à ajouter un étalon interne d'ADN (molécule d'ADN biologique ou synthétique étrangère, *ISD*) dans les échantillons d'ADNe en quantité absolue égale, au début du traitement en laboratoire. En intégrant une combinaison de différents ISD de quantités différentes dans les échantillons d'ADNe et en comparant l'abondance des *reads* de chaque espèce à l'abondance de *reads* des ISD, les abondances absolues de chaque espèce peuvent être calculées et comparées entre échantillons (Harrison et al. 2021; Sato et al. 2021).

Les processus de traitements bio-informatiques des données brutes peuvent également être optimisés en testant de nouvelles méthodes. Les premiers essais de traitements et assignation des séquences avec l'intelligence artificielle ont fourni des résultats prometteurs, en détectant les mêmes espèces qu'avec le pipeline OBITools et que celles connues dans les inventaires de rivières tropicales (Flück et al. 2022). Ces méthodes permettent d'accélérer grandement les temps de calculs et de prendre en compte les variations inter- et intraspécifique dans l'apprentissage de l'assignation, mais sont cependant dépendantes d'une base de référence complète avec un large nombre de séquences pour chaque espèce. Une solution pour l'assignation des séquences en cas de base de référence incomplète serait le placement phylogénétique (Matsen et al. 2010). Cette méthodologie a émergé du fait de l'impossibilité d'inférer un arbre robuste directement à partir des courtes séquences qui ne contiennent qu'une faible information phylogénétique. Le placement phylogénétique place les séquences sur les branches d'un arbre phylogénétique de référence construit à partir de génomes de référence (Czech et al. 2022), et assigne une probabilité de placement sur chaque branche (Figure 6.3).

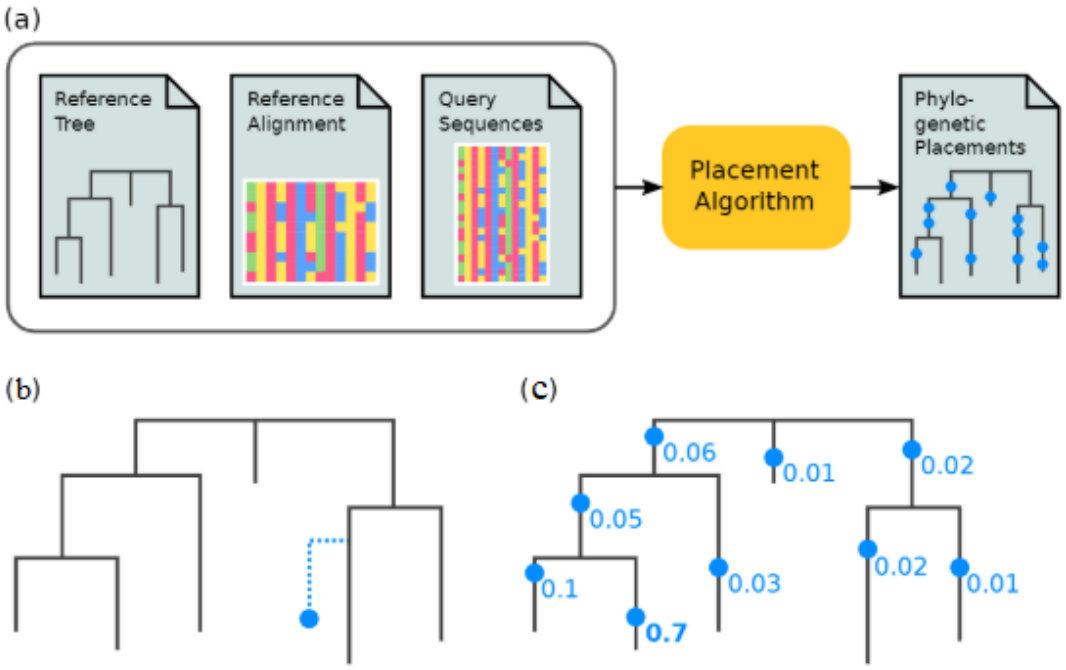


Figure 6.3. *Schématisation du placement phylogénétique. (a) données d'entrée : arbre de référence (RT), alignement de référence (RA), et les séquences d'interrogation (QS), (b) Une QS est placée le long d'une branche existante, avec une longueur de branche propre. Le placement basé sur le maximum de vraisemblance calcule la vraisemblance du RT avec le QS comme branche supplémentaire. (c) Une fois que les probabilités de placer la QS sur chaque branche ont été calculées, le ratio de pondération de la vraisemblance (LWR) pour cette QS sont calculés. Figure extraite de Czech et al. (2022).*

Par conséquent, il serait possible d'identifier la taxonomie d'une courte séquence, et potentiellement de déterminer si une séquence correspond à une nouvelle branche de l'arbre (espèce dont la séquence ne serait pas encore connue). Plusieurs algorithmes de placement existent, basés sur le maximum de vraisemblance, sur la distance entre séquences, ou sur la reconstruction ancestrale (Czech et al. 2022). De nouveaux algorithmes, dont certains basés sur les k-mers (Linard et al. 2019), sont en cours de développement pour accélérer ce processus et s'affranchir de l'alignement des séquences (Blanke & Morgenstern, 2020; Linard et al. 2020).

L'une des principales perspectives de ces travaux de thèse serait de parvenir à mieux identifier les espèces détectées dans nos échantillons, en vue d'affiner les inventaires de biodiversité et de réaliser des analyses écologiques plus poussées à l'échelle des communautés. Compléter les bases de référence pour le marqueur que nous utilisons serait la meilleure solution pour mieux identifier les espèces dans nos échantillons, mais c'est une solution à long terme. En effet, il est fastidieux d'identifier les espèces dont la séquence n'est pas renseignée dans les bases de données, de se rendre dans l'aire de répartition de ces espèces, de les pêcher et séquencer leur ADN. D'autant plus que beaucoup de ces espèces sont très rares ou menacées, ou vivent dans des zones difficiles d'accès, comme les grandes profondeurs ou les zones polaires. Une solution pour compléter ces bases de référence est de faire appel aux aquariums et musées, afin de récupérer de l'ADN d'espèces difficiles à trouver autrement. Cependant la qualité de l'ADN dans les musées dépend de l'âge et de la méthode de conservation de l'échantillon. Hahn et al. (2022) ont démontré qu'une lyse alcaline à chaud couplée à une extraction phénol-chloroforme donne de meilleurs résultats que la digestion par la protéinase K pour la récupération de l'ADN sur des tissus préservés dans du formol. D'autres études récentes sont parvenues à extraire et séquencer des génomes mitochondriaux quasi-complets, à partir d'ADN d'anciens spécimens de musées, en convertissant l'ADN en librairie simple-brins et en utilisant des sondes sur mesures pour la capture (Straube et al. 2021). L'augmentation des bases de référence à partir de spécimens de musées commence à voir le jour. En Australie, l'Organisation de recherche scientifique et industrielle du Commonwealth (CSIRO), en partenariat avec la fondation Minderoo, lance une initiative de séquençage de tous les vertébrés marins à partir des collections de musées. En attendant que les bases de références soient complétées, il est possible d'amplifier et séquencer de multiples marqueurs sur nos échantillons d'ADNe afin de mieux assigner les espèces en croisant les informations, ou de cibler différents groupes taxonomiques (Polanco et al. 2021; West et al. 2021), mais le coût d'analyse en est alors multiplié.

3.2. Développement des connaissances en biogéographie

L'échantillonnage réalisé dans cette thèse permet d'étudier les patrons de distribution des poissons à échelle locale, régionale et globale, comme montré dans les **chapitres 3 à 5**. En complétant cet échantillonnage spatialement et en profondeur et en affinant l'identification des espèces, il serait possible d'étudier davantage de métriques de diversité et de mieux décrire ces patrons de distribution dans l'espace tridimensionnel. L'échantillonnage à l'échelle globale s'est poursuivi depuis mes dernières analyses, avec de nouvelles données collectées en Tanzanie, Arctique, Norvège et Méditerranée et de nouvelles missions sont prévues pour les années à venir. Ces nouvelles données pourront être analysées et intégrées dans une prochaine étude à large échelle, pour affiner nos connaissances sur les patrons de distribution des poissons et étudier de nouveaux aspects de la diversité. Certaines régions bénéficieraient d'un échantillonnage supplémentaire pour obtenir une estimation de diversité plus exhaustive (ex : Polynésie, Pacifique central). Afin d'obtenir des données comparables en termes de méthodologie et d'effort d'échantillonnage, il serait nécessaire de proposer une méthode d'échantillonnage standardisée à utiliser dans les futures études. Plusieurs études ont démontré que de gros volumes filtrés et un grand nombre de réplicas par localités permettaient de détecter un plus grand nombre d'espèces et d'obtenir des estimations de diversité semblables aux checklists locales et régionales (Stauffer et al. 2021). Nous pourrions donc appliquer ces méthodes d'échantillonnage lors des prochaines campagnes.

L'ensemble des données utilisées dans cette thèse et les nouvelles données collectées depuis, pourraient être analysées à nouveau pour étudier d'autres aspects de la diversité et de la distribution des poissons, en précisant l'assignation taxonomique à l'aide d'une base de référence plus complète. En effet, les échantillons d'ADNe se conservent facilement et peuvent être réanalysés plusieurs fois, si l'on souhaite cibler différents groupes taxonomiques, ou si les méthodes de laboratoire et bio-informatiques deviennent plus performantes dans les prochaines années et permettent d'obtenir des résultats plus précis qu'actuellement. Une nouvelle analyse des échantillons à l'échelle globale est en cours, en utilisant une approche de factorisation de matrices éparses non-négatives (*sparse non-negative matrix factorization approaches – sNMF*, Frichot et al. 2014). Cette analyse permet d'étudier la structure spatiale et taxonomique d'un assemblage et d'inférer la structure d'une population et les coefficients d'ascendance individuelle, à partir d'un jeu de données génétiques. Les résultats de cette analyse permettront d'approfondir nos connaissances sur les patrons d'assemblage et la structure des communautés de poissons à large échelle. En analysant à nouveau nos filtres, il serait également possible

d'investiguer si les séquences d'ADNe détectées dans nos échantillons proviennent d'individus adultes occupant l'habitat ou de larves transportées par le courant. Bylemans et al. (2017) ont démontré qu'en mesurant la concentration d'ADNe nucléaire et mitochondrial dans le temps, par PCR quantitative, il est possible d'identifier si l'ADN provient de larves (pic de concentration d'ADNe nucléaire) ou d'individus adultes. Cette application pourrait permettre de confirmer la présence d'individus adultes d'espèces pélagiques, profondes ou de fonds meubles identifiées sur les récifs coralliens dans l'étude du **chapitre 3**. Ainsi nous pourrions affiner nos connaissances sur la distribution de ces espèces et sur leur usage de ces habitats. Différencier l'ADNe provenant de larves ou d'adultes permettrait également d'identifier des zones de pontes d'espèces à fort enjeux de conservation, afin de suivre et protéger ces zones (Ratcliffe et al. 2020). Il serait également possible d'identifier des transports larvaires d'espèces invasives dans des zones encore non colonisées, et ainsi informer rapidement les gestionnaires.

En se focalisant seulement sur la richesse taxonomique, on manque les aspects cruciaux de diversité des fonctions et diversité des lignées, qui sont pourtant largement menacées par les changements globaux actuels (D'Agata et al. 2014; Frainer et al. 2017; Pimiento et al. 2020). Il est possible d'associer des traits fonctionnels et une position phylogénétique aux séquences d'ADNe assignées à l'espèce (voire au genre, en moyennant les traits et phylogénie entre les espèces au sein du genre). Ainsi on peut mesurer les indices de diversité fonctionnelle d'un assemblage (ou rareté, divergence, etc.) et sa diversité phylogénétique (Marques et al. 2021; Polanco et al. 2022). En revanche, si les séquences d'ADNe sont assignées seulement à la famille ou à des niveaux taxonomiques supérieurs, il est difficile d'estimer les caractéristiques fonctionnelles et phylogénétiques de ces séquences, car ces caractéristiques peuvent être largement variables au sein d'une famille. Il serait alors intéressant d'approfondir les études sur les relations entre les indices de diversité de séquences et de diversité fonctionnelle et phylogénétique, mesurés avec les nombres de Hill. Nous pourrions alors étudier ces métriques de diversité à l'échelle globale, directement à partir des MOTUs, et étudier leurs réponses aux pressions humaines et environnementales.

Les poissons jouent un rôle primordial pour la santé des écosystèmes, mais les interactions avec les autres taxons au sein du réseau trophique sont également indispensables à leur maintien et au bon fonctionnement des océans (Cline & Allgeier, 2022). En effet, Gaüzère et al. (2022) ont démontré que la diversité d'interactions entre plusieurs groupes trophiques n'était corrélée ni à la diversité fonctionnelle ni à la diversité phylogénétique, et informait donc sur un autre aspect du fonctionnement des écosystèmes. Etudier d'autres groupes taxonomiques permettrait

donc d'avoir une vision plus générale sur les réseaux trophiques et les différents niveaux d'organisation dans l'écosystème. Nous pourrions ainsi réanalyser tous nos échantillons d'ADNe pour amplifier et séquencer d'autres taxa, tels que les mammifères marins, les invertébrés, les coraux etc. En effet, Cline & Allgeier (2022) ont mis en évidence que la structure et la dynamique des communautés de poissons ne permettaient pas d'expliquer la dynamique des récifs coralliens, et donc qu'une protection contre la pêche seule ne renforcerait pas la résilience des récifs. Il serait donc intéressant d'étudier les autres groupes trophiques. En se basant sur la littérature concernant les interactions trophiques, il serait possible de reconstruire les réseaux trophiques au sein de chacune de nos communautés, comme l'ont fait D'Alessandro & Mariani (2021). Djurhuus et al. (2020) ont ainsi pu construire un réseau d'interactions et mettre en évidence des interactions proies-prédateurs, des liens trophiques et des changements saisonniers au sein du réseau, en calculant les corrélations entre les occurrences de plusieurs taxa. A la manière de Gaüzère et al. (2022), nous pourrions également, à partir d'un réseau trophique connu, mesurer un indice de diversité d'interactions entre nos séquences d'ADNe identifiées à l'espèce. Cet indice pourrait ensuite être étudié à large échelle, au même titre que les indices de diversité taxonomique, fonctionnelle et phylogénétique, pour obtenir une meilleure compréhension du fonctionnement des communautés.

Enfin, lors des prochaines campagnes d'échantillonnage, il serait intéressant d'obtenir des données d'ADNe et des estimations de diversité dans les zones mésophotiques et aphotiques. Ces données permettraient d'étudier de nouveaux patrons de distribution des poissons, de mieux comprendre leur répartition dans la colonne d'eau et d'apporter de nouvelles connaissances sur leur niche écologique (Muff et al. 2022). Réétudier la distribution des poissons à la lumière de nouvelles connaissances permettrait sûrement de redéfinir la délimitation des régions et provinces biogéographiques, basés principalement sur l'endémicité des espèces et l'homogénéité des assemblages (Costello et al. 2017), ou de remettre en questions d'autres patrons comme les gradients de diversité (Chaudhary et al. 2016). En effet, si de nouvelles données dans les zones profondes permettent de redéfinir les aires de répartition des espèces, alors ces patrons de distribution pourraient en être bouleversés, comme l'illustre l'observation d'un requin du Groenland dans les eaux profondes des Caraïbes (Kasana et al. 2022). Plusieurs études ont déjà montré des différences d'assemblages de communautés, de richesse et divergence fonctionnelle en fonction de la profondeur, à l'aide de méthodes conventionnelles (McClain & Lundsten, 2015; Mindel et al. 2016). L'utilisation de l'ADNe serait plus adaptée pour étudier ces écosystèmes profonds, car plus facile et plus rapide à mettre en place. Comme

démontré dans le **chapitre 5** ainsi que dans une étude récente (Muff et al. 2022), le prélèvement d'ADNe sur les zones côtières profondes ou sur les monts sous-marins fournit des informations inédites et localisées sur les assemblages profonds des poissons, jusqu'à de grandes profondeurs. Obtenir des informations sur ces habitats profonds dans un plus grand nombre de localisations permettrait d'étudier ces patrons de distribution en profondeur à large échelle et d'appliquer une méthodologie de planification spatiale en trois dimensions (Duffy & Chown, 2017). En effet, plusieurs études ont démontré le rôle des zones profondes et des monts sous-marins comme refuges contre les changements climatiques et perturbations humaines pour les grands prédateurs (Letessier et al. 2019), les coraux (Baird et al. 2018; Bongaerts & Smith, 2019; Montgomery et al. 2021) et les poissons (Brito-Morales et al. 2022).

3.3. Perspectives pour la conservation

Mes travaux ont démontré que le *metabarcoding* de l'ADNe peut être utilisé pour étudier les patrons de diversité des poissons à plusieurs échelles spatiales. Cette méthode est également prometteuse pour réaliser des suivis temporels de la diversité. En effet, l'échantillonnage de l'ADNe est rapide à effectuer et peut donc être répété sur de faibles pas de temps, afin d'obtenir des séries temporelles de plusieurs indices de diversité, sur de larges zones spatiales. Des suivis spatio-temporels seraient particulièrement intéressants pour étudier la propagation d'espèces invasives, comme le poissons-lion *(Pterois miles*) et le poissons-lapin (*Siganus luridus*) en Méditerranée, ou encore le crabe bleu (*Callinectes sapidus*) sur les côtes du golfe du lion (Daniel et al. 2009). Ces analyses de *barcoding*, ciblant ici une espèce particulière, pourraient permettre la détection précoce de la présence de ces espèces dans des nouvelles localisations, et la mise en place de mesures de gestion (Van Oppen & Coleman, 2022). Une récente étude a développé des primers spécifiques pour 69 espèces de poissons et invertébrés invasives, qui ont permis de détecter ≃98% des espèces ciblées, même en très faible abondance (Wu et al. 2022). Le *barcoding* de l'ADNe pourrait également être utilisé pour cibler des espèces menacées et en danger d'extinction, comme l'Ange de mer (*Squatina squatina*) en Méditerranée. Cela permettrait d'apporter de nouvelles informations sur les populations résiduelles de ces espèces, et sur leurs habitats, afin de pouvoir mieux les protéger (Bonfil et al. 2021; Plough et al. 2021). Une étude en cours, utilisant l'ADNe, a permis de détecter des individus d'ange de mer autour de la Corse, dans des localisations où il n'avait plus été détecté par les méthodes traditionnelles. Il serait également intéressant de réanalyser nos échantillons d'ADNe en ciblant un groupe taxonomique particulier (ex : Elasmobranches), ou les espèces inscrites sur la Liste Rouge de

l'UICN, afin d'étudier leur distribution à l'échelle globale et suivre l'évolution de leur aire de répartition. Un suivi spatio-temporel de ces groupes ou espèces permettrait d'obtenir de précieuses informations sur leurs aires de répartition, leurs habitats occupés, leurs migrations potentielles et d'étudier le déplacement de leur niche écologique en réponse aux changements climatiques et aux perturbations humaines (Ariza et al. 2022; Lenoir et al. 2020). Les espèces menacées étant encore peu représentées dans les bases de référence publiques (Marques, Milhau, et al. 2020), nous pourrions cibler nos efforts de séquençage sur ces espèces prioritairement. Il serait également envisageable de se focaliser sur les espèces dites « *data deficient* », pour lesquelles les données sont insuffisantes pour estimer leur statut de conservation, et ainsi apporter plus d'informations sur leur distribution et leur niche écologique. Une récente étude a estimé grâce au *machine learning* qu'au moins la moitié des espèces « *data deficient* » serait en danger d'extinction (Borgelt et al. 2022). L'analyse de l'ADN environnemental peut donc être un outil puissant pour la conservation des espèces et la gestion des espaces et notamment pour surveiller les espèces invasives et étudier les changements de composition des communautés dans le temps et l'espace.

Les données issues du *metabarcoding* de l'ADNe peuvent également être utilisées pour évaluer les stratégies de gestion et calculer des indicateurs écologiques reflétant la santé de l'écosystème (Pawlowski et al. 2018; Van Oppen & Coleman, 2022). Ainsi, Boulanger et al. (2021) ont démontré que les réserves marines de Méditerranée bénéficiaient aux espèces prédatrices et de grandes tailles, bien qu'à l'extérieur la richesse spécifique soit plus élevée du fait d'un grand nombre d'espèces cryptobenthiques. En mesurant plusieurs indices de diversité, de récentes études sur les mêmes réserves de Méditerranée ont montré une plus grande proportion d'espèces démersales et pélagiques ainsi que de plus grandes diversités fonctionnelles et phylogénétiques dans les réserves qu'à l'extérieur (Dalongeville et al. 2022; Sanchez et al. 2022) (Figure 6.4). Ces indicateurs peuvent être utilisés pour l'évaluation du bon état écologique des écosystèmes marins. Par exemple, à l'échelle nationale, la Directive Cadre Stratégie pour le Milieu Marin (DCSMM) définit le bon état écologique selon une dizaine de critères, comme la biodiversité (distribution et abondance des espèces), l'absence ou la régulation des espèces non-indigènes, l'exploitation durable des stocks d'espèces commerciales, le maintien des réseaux trophiques, la préservation des fonds marins etc. L'ADNe peut donc être utilisé pour mesurer une partie de ces indicateurs et informer sur l'état de santé des écosystèmes et sur les mesures à prendre pour leur conservation.

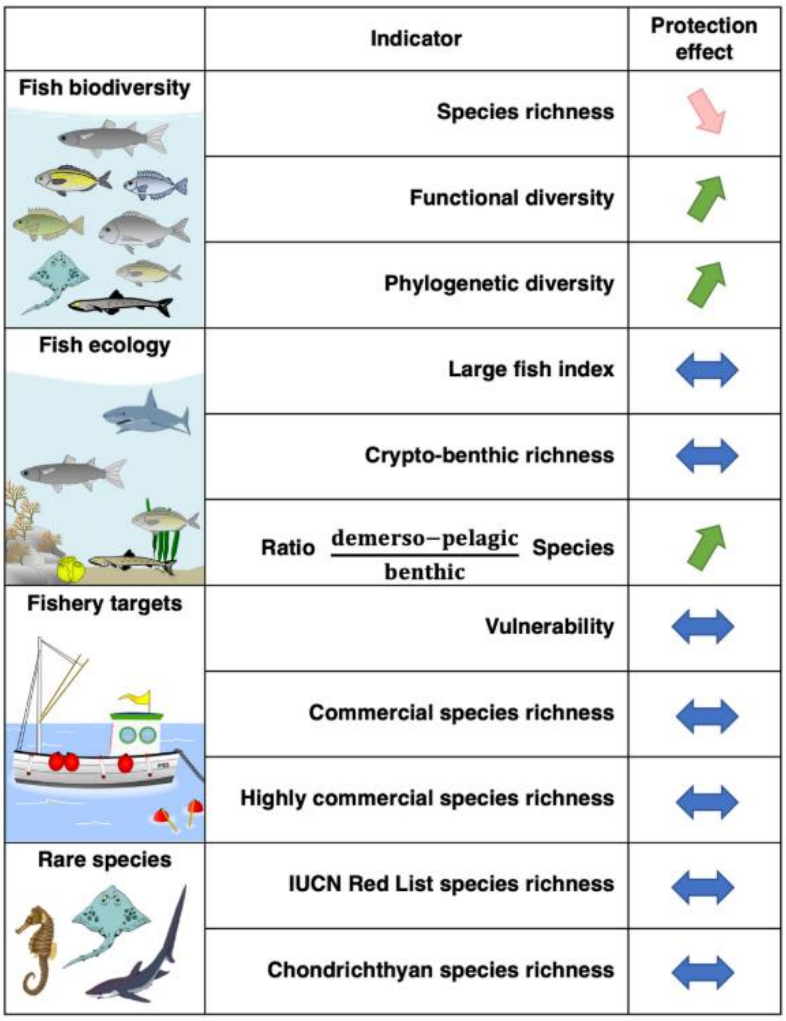


Figure 6.4. *Réponses des indicateurs de biodiversité à la protection en réserve, en Méditerranée française, mesurées par des modèles linéaires généralisés (GLM). Figure extraite de Dalongeville et al. (2022).*

En combinant l'ensemble de ces nouvelles données et nouveaux indices (diversité fonctionnelle, diversité phylogénétique, richesse de chondrichtyens, richesse d'espèces menacées, présence d'espèce invasive, interactions trophiques, influence des facteurs humains et environnementaux etc.), nous obtenons une vision plus complète des écosystèmes marins et de leur fonctionnement. Toutes ces informations, obtenues par l'ADNe ou par combinaison avec d'autres méthodes, permettront d'informer la conservation pour la mise en place de plan de gestion. En effet, comme réalisé au **chapitre 5**, il est possible d'identifier les zones hébergeant une forte diversité (d'espèces, de fonctions, de lignées etc.) ou une biodiversité menacée, et de les inclure dans des plans de gestion correspondant à des objectifs de conservation concrets et réalisables (Bani et al. 2020). Afin de pouvoir considérer l'océan dans toutes ses dimensions, il serait important d'obtenir davantage d'échantillons prélevés sur les

habitats profonds et d'étudier la diversité des poissons dans ces milieux encore méconnus. Ainsi, les plans de conservation pourront être réalisés en trois dimensions afin de protéger les zones profondes abritant une faune unique (Manea et al. 2020; Venegas-Li et al. 2018). De tels plans de conservation pourraient par exemple interdire ou limiter le chalutage de fond ou les forages sous-marins sur les habitats profonds les plus sensibles ou les plus riches, ou alors limiter la pêche démersale pour protéger le compartiment pélagique. Ainsi, Brito-Morales et al. (2022) ont développé une approche de planification et proposent un scénario de conservation de la haute mer (océans se trouvant en dehors des juridictions nationales) intégrant 4 couches de profondeur. Ce scénario permettrait de protéger des refuges de biodiversité sur 12% de la surface des océans dans la zone pélagique, et 6% sur toute la colonne d'eau jusqu'au fond océanique, dans des zones encore peu impactées par le changement climatique, et en limitant les conflits avec la pêche (Figure 6.5).

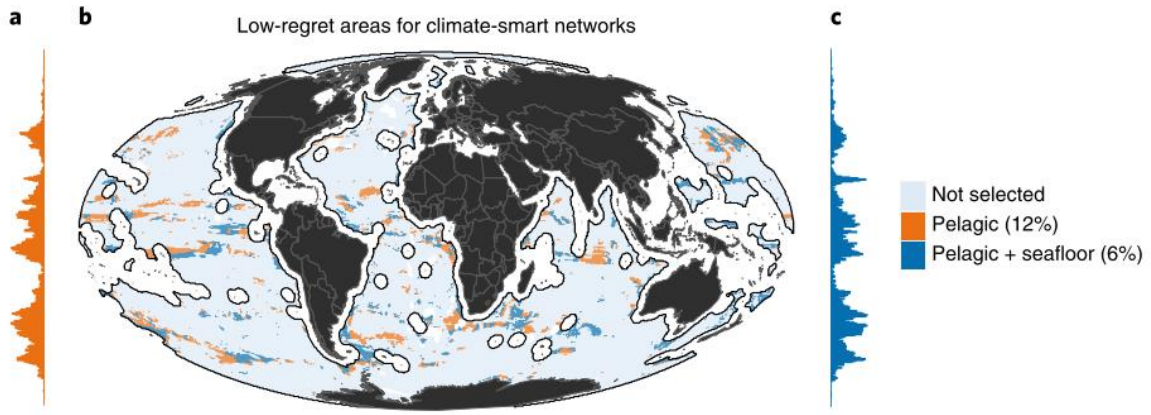


Figure 6.5. *Réseaux de priorisation des zones de conservation dans toute la colonne d'eau de l'océan pour les domaines pélagiques (épipélagique, mésopélagique et bathyabyssopélagique) et pour les domaines pélagiques et du fond marin (b). Les graphes latéraux montrent la distribution latitudinale du réseau de priorisation en tant que proportion de la surface de l'océan pour les domaines pélagiques (a) et les domaines pélagiques plus le plancher océanique (c). Figure extraite de Brito-Morales et al. (2022).*

Afin d'œuvrer efficacement pour la conservation des écosystèmes marins à l'échelle globale, il serait très bénéfique de centraliser les efforts et standardiser les méthodes d'échantillonnage et d'analyse (Bruce et al. 2021). C'est déjà le cas à l'échelle Européenne avec le collectif DNAqua-Net (Leese et al. 2016) et en Amérique du Nord avec le collectif PISCeS (Loeza-Quintana et al. 2020). La mise en place d'initiatives internationales à large échelle

permet de créer de larges banques de données pour la surveillance et l'évaluation de l'état de santé des écosystèmes. L'Observatoire du Vivant, crée par Vigilife vise ainsi à étudier la diversité dans les fleuves et aires marines sentinelles, par des méthodes d'ADN environnemental standardisées. De la même façon, eBioAtlas, issu d'un partenariat entre NatureMetrics et l'IUCN, a pour objectif de créer un atlas de la biodiversité dans les grands fleuves du monde afin d'informer la conservation. Enfin, l'UNESCO lance également un programme de suivi de la biodiversité marine dans les sites du patrimoine mondial de l'UNESCO, par ADN environnemental, afin de mesurer les effets du changement climatique.

Références

Aglieri, G., Baillie, C., Mariani, S., Cattano, C., Calò, A., Turco, G., … Milazzo, M. (2020). Environmental DNA effectively captures functional diversity of coastal fish communities. *Molecular Ecology*, (August), 1–13. https://doi.org/10.1111/mec.15661

Aho, A. V., Hopcroft, J. E., & Ullman, J. D. (1976). On finding lowest common ancestors in trees. *SIAM Journal on Computing*, *5*(1), 1–18. https://doi.org/10.1007/978-3-642-27848-8_630-1

Alberdi, A., & Gilbert, M. T. P. (2019). A guide to the application of Hill numbers to DNA-based diversity analyses. *Molecular Ecology Resources*, *19*(4), 804–817. https://doi.org/10.1111/1755-0998.13014

Albouy, C., Leprieur, F., Le Loc'h, F., Mouquet, N., Meynard, C. N., Douzery, E. J. P., & Mouillot, D. (2015). Projected impacts of climate warming on the functional and phylogenetic components of coastal Mediterranean fish biodiversity. *Ecography*, *38*(7), 681–689. https://doi.org/10.1111/ecog.01254

Albuquerque, F., & Astudillo-Scalia, Y. (2020). The role of rarity as a surrogate of marine fish species representation. *PeerJ*, *2020*(2). https://doi.org/10.7717/peerj.8373

Allard, L., Grenouillet, G., Khazraie, K., Tudesque, L., Vigouroux, R., & Brosse, S. (2014). Electrofishing efficiency in low conductivity neotropical streams: Towards a non-destructive fish sampling method. *Fisheries Management and Ecology*, *21*(3), 234–243. https://doi.org/10.1111/fme.12071

Allard, Luc, Popée, M., Vigouroux, R., & Brosse, S. (2016). Effect of reduced impact logging and small-scale mining disturbances on Neotropical stream fish assemblages. *Aquatic Sciences*, *78*(2), 315–325. https://doi.org/10.1007/s00027-015-0433-4

Andrello, M., Darling, E. S., Wenger, A., Suárez-Castro, A. F., Gelfand, S., & Ahmadia, G. N. (2022). A global map of human pressures on tropical coral reefs. *Conservation Letters*, *15*(e12858), 1–12. https://doi.org/10.1111/conl.12858

Annasawmy, P., Ternon, J. F., Cotel, P., Cherel, Y., Romanov, E. V., Roudaut, G., … Marsac, F. (2019). Micronekton distributions and assemblages at two shallow seamounts of the south-western Indian Ocean: Insights from acoustics and mesopelagic trawl data. *Progress in Oceanography*, *178*(May), 102161. https://doi.org/10.1016/j.pocean.2019.102161

Antão, L. H., Bates, A. E., Blowes, S. A., Waldock, C., Supp, S. R., Magurran, A. E., … Schipper, A. M. (2020). Temperature-related biodiversity change across temperate marine and terrestrial systems. *Nature Ecology and Evolution*, *4*(7), 927–933. https://doi.org/10.1038/s41559-020-1185-7

Ariza, A., Lengaigne, M., Menkes, C., Lebourges-dhaussy, A., Receveur, A., Gorgues, T., … Maury, O. (2022). Global decline of pelagic fauna in a warmer ocean. *Nature Climate Change*, (September). https://doi.org/10.1038/s41558-022-01479-2

Asher, J., Williams, I. D., & Harvey, E. S. (2019). Is seeing believing? Diver and video-based censuses reveal inconsistencies in roving predator estimates between regions. *Marine Ecology Progress Series*, *630*, 115–136. https://doi.org/10.3354/meps13107

Auber, A., Waldock, C., Maire, A., Goberville, E., Albouy, C., Algar, A. C., … Mouillot, D. (2022). A functional vulnerability framework for biodiversity conservation. *Nature Communications*, *13*(4774), 1–13.

Bain, M. B., Finn, J. T., & Booke, H. E. (1985). A Quantitative Method for Sampling Riverine Microhabitats by Electrofishin. *North American Journal of Fisheries Management*, *5*, 475–479. https://doi.org/10.1577/1548-8659(1985)5<475:rtabtm>2.0.co;2

Baird, A. H., Madin, J. S., Álvarez-Noriega, M., Fontoura, L., Kerry, J. T., Kuo, C. Y., … Hughes, T. P. (2018). A decline in bleaching suggests that depth can provide a refuge from global warming in most coral taxa. *Marine Ecology Progress Series*, *603*, 257–264.

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), 16886.

Bani, A., De Brauwer, M., Creer, S., Dumbrell, A. J., Limmon, G., Jompa, J., … Beger, M. (2020). Informing marine spatial planning decisions with environmental DNA. In *Advances in Ecological Research* (1st ed., Vol. 62). https://doi.org/10.1016/bs.aecr.2020.01.011

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

Barko, V. A., Briggler, J. T., & Ostendorf, D. E. (2004). Passive Fishing Techniques: a Cause of Turtle Mortality in the Mississippi River. *Journal of Wildlife Management*, *68*(4), 1145–1150. https://doi.org/10.2193/0022-541x(2004)068[1145:pftaco]2.0.co;2

Baselga, A. (2010). Partitioning the turnover and nestedness components of beta diversity. *Global Ecology and Biogeography*, *19*(1), 134–143.

Batista, V. S., Fabré, N. N., Malhado, A. C. ., & Ladle, R. J. (2014). Tropical Artisanal Coastal Fisheries: Challenges and Future Directions. *Reviews in Fisheries Science and Aquaculture*, *22*(1), 1–15.

Baumann, J. H., Zhao, L. Z., Stier, A. C., & Bruno, J. F. (2022). Remoteness does not enhance coral reef resilience. *Global Change Biology*, *28*(2), 417–428. https://doi.org/10.1111/gcb.15904

Beckley, L. E., Holliday, D., Sutton, A. L., Weller, E., Olivar, M. P., & Thompson, P. A. (2019). Structuring of larval fish assemblages along a coastal-oceanic gradient in the macro-tidal, tropical Eastern Indian Ocean. *Deep-Sea Research Part II*, *161*(March 2018), 105–119. https://doi.org/10.1016/j.dsr2.2018.03.008

Bellwood, D. R., Pratchett, M. S., Morrison, T. H., Gurney, G. G., Hughes, T. P., Álvarez-Romero, J. G., … Cumming, G. S. (2019). Coral reef conservation in the Anthropocene: Confronting spatial mismatches and prioritizing functions. *Biological Conservation*, *236*(June), 604–615. https://doi.org/10.1016/j.biocon.2019.05.056

Bellwood, D., Renema, W., & Rosen, B. B. (2012). Biodiversity hotspots, evolution and coral reef biogeography: a review. In *Biotic Evolution and Environmental Change in Southeast Asia.* (Cambridge, pp. 216–245). Cambridge.

Bellwood, David, & Hughes, T. P. (2001). Regional-scale assembly rules and biodiversity of coral reefs. *Science*, *292*(5521), 1532–1534. https://doi.org/10.1126/science.1058635

Bender, M. G., Leprieur, F., Mouillot, D., Kulbicki, M., Parravicini, V., Pie, M. R., … Floeter, S. R. (2017). Isolation drives taxonomic and functional nestedness in tropical reef fish faunas. *Ecography*, *40*(3), 425–435.

Berger, C. S., Hernandez, C., Laporte, M., Côté, G., Paradis, Y., Kameni T., D. W., … Bernatchez, L. (2020). Fine-scale environmental heterogeneity shapes fluvial fish communities as revealed by eDNA metabarcoding. *Environmental DNA*, *2*(4), 647–666. https://doi.org/10.1002/edn3.129

Bernard, A. T. F., Götz, A., Kerwath, S. E., & Wilke, C. G. (2013). Observer bias and detection probability in underwater visual census of fish assemblages measured with independent double-

observers. *Journal of Experimental Marine Biology and Ecology*, *443*, 75–84. https://doi.org/10.1016/j.jembe.2013.02.039

Bessey, C., Jarman, S. N., Berry, O., Olsen, Y. S., Bunce, M., Simpson, T., … Keesing, J. (2020). Maximizing fish detection with eDNA metabarcoding. *Environmental DNA*, *2*(4), 493–504. https://doi.org/10.1002/edn3.74

Bessey, C., Jarman, S. N., Simpson, T., Miller, H., Stewart, T., Keesing, J. K., & Berry, O. (2021). Passive eDNA collection enhances aquatic biodiversity analysis. *Communications Biology*, *4*(236), 1–12. https://doi.org/10.1038/s42003-021-01760-8

Blanke, M., & Morgenstern, B. (2020). Phylogenetic placement of short reads without sequence alignment. *BioRxiv*.

Bohmann, K., Elbrecht, V., Carøe, C., Bista, I., Leese, F., Bunce, M., … Creer, S. (2021). Strategies for sample labelling and library preparation in DNA metabarcoding studies. *Molecular Ecology Resources*, (April), 1–16. https://doi.org/10.1111/1755-0998.13512

Bonfil, R., Palacios-Barreto, P., Mendoza Vargas, O. U., Ricaño-Soriano, M., & Diaz-Jaimes, P. (2021). Detection of critically endangered marine species with dwindling populations in the wild using eDNA gives hope for sawfish. *Marine Biology*, *168*(60), 1–12.

Bongaerts, P., & Smith, T. B. (2019). Beyond the "Deep Reef Refuge" Hypothesis: A Conceptual Framework to Characterize Persistence at Depth. In Y. Loya (Ed.), *Mesophotic Coral Ecosystems* (Springer N, pp. 881–895). https://doi.org/10.1007/978-3-319-92735-0_45

Borgelt, J., Dorber, M., Hoiberg, M. A., & Verones, F. (2022). More than half of data deficient species predicted to be threatened by extinction. *Communications Biology*, *5*(679), 1–10.

Boulanger, E., Loiseau, N., Valentini, A., Arnal, V., Boissery, P., Dejean, T., … Mouillot, D. (2021). Environmental DNA metabarcoding reveals and unpacks a biodiversity conservation paradox in Mediterranean marine reserves. *Proceedings of the Royal Society B*, *288*(20210112), 1–10.

Boussarie, G., Kiszka, J. J., Mouillot, D., Bonnin, L., Manel, S., Kulbicki, M., … Mariani, S. (2018). Environmental DNA illuminates the dark diversity of sharks. *Science Advances*, *4*(5), 1–8. 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. https://doi.org/10.1111/1755-0998.12428

Brandl, S. J., Goatley, C. H. R., Bellwood, D. R., & Tornabene, L. (2018). The hidden half : ecology and evolution of cryptobenthic fishes on coral reefs. *Biological Reviews*, *93*, 1846–1873.

Brandl, S. J., Johansen, J. L., Casey, J. M., Tornabene, L., Morais, R. A., & Burt, J. A. (2020). Extreme environmental conditions reduce coral reef fish biodiversity and productivity. *Nature Communications*, *11*(1), 1–14. https://doi.org/10.1038/s41467-020-17731-2

Brandl, S. J., Tornabene, L., Goatley, C. H. R., Casey, J. M., Morais, R. A., Côté, I. M., … Bellwood, D. R. (2019). Demographic dynamics of the smallest marine vertebrates fuel coral-reef ecosystem functioning. *Science*, (May), 799–802. https://doi.org/10.1038/45533

Brandt, M. I., Trouche, B., Quintric, L., Günther, B., Wincker, P., Poulain, J., & Arnaud-Haond, S. (2021). Bioinformatic pipelines combining denoising and clustering tools allow for more comprehensive prokaryotic and eukaryotic metabarcoding. *Molecular Ecology Resources*, *21*(6), 1904–1921. https://doi.org/10.1111/1755-0998.13398

Brito-Morales, I., Schoeman, D. S., Everett, J. D., Klein, C. J., Dunn, D. C., García Molinos, J., … Richardson, A. J. (2022). Towards climate-smart, three-dimensional protected areas for biodiversity conservation in the high seas. *Nature Climate Change*, *12*(4), 402–407.

https://doi.org/10.1038/s41558-022-01323-7

Brito-Morales, I., Schoeman, D. S., Molinos, J. G., Burrows, M. T., Klein, C. J., Arafeh-Dalmau, N., … Richardson, A. J. (2020). Climate velocity reveals increasing exposure of deep-ocean biodiversity to future warming. *Nature Climate Change*, *10*(6), 576–581. https://doi.org/10.1038/s41558-020-0773-5

Brodie, J. E., Kroon, F. J., Schaffelke, B., Wolanski, E. C., Lewis, S. E., Devlin, M. J., … Davis, A. M. (2012). Terrestrial pollutant runoff to the Great Barrier Reef: An update of issues, priorities and management responses. *Marine Pollution Bulletin*, *65*(4–9), 81–100. https://doi.org/10.1016/j.marpolbul.2011.12.012

Brown, S. C., Mellin, C., Molinos, J. G., Lorenzen, E. D., & Fordham, D. A. (2022). Faster ocean warming threatens richest areas of marine biodiversity. *Global Change Biology*, *28*, 5849–5858.

Bruce, K., Blackman, R. C., Bourlat, S. J., Hellström, M., Bakker, J., Bista, I., … Deiner, K. (2021). *A practical guide to DNA- based methods for biodiversity assessment A practical guide to DNA-based methods for biodiversity assessment*.

Bylemans, J., Furlan, E. M., Hardy, C. M., McGuffie, P., Lintermans, M., & Gleeson, D. M. (2017). An environmental DNA-based method for monitoring spawning activity: a case study, using the endangered Macquarie perch (Macquaria autralasica). *Methods in Ecology and Evolution*, *8*, 646–655.

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

Cabral, R. B., Halpern, B. S., Lester, S. E., White, C., Gaines, S. D., & Costello, C. (2019). Designing MPAs for food security in open-access fisheries. *Scientific Reports*, *9*(1), 1–10. https://doi.org/10.1038/s41598-019-44406-w

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*, (July), 1–14. https://doi.org/10.1111/jbi.13681

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME Journal*, *11*(12), 2639–2643. https://doi.org/10.1038/ismej.2017.119

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Campanella, F., Collins, M. A., Young, E. F., Laptikhovsky, V., Whomersley, P., & van der Kooij, J. (2021). First Insight of Meso- and Bentho-Pelagic Fish Dynamics Around Remote Seamounts in the South Atlantic Ocean. *Frontiers in Marine Science*, *8*(June). https://doi.org/10.3389/fmars.2021.663278

CBD. (2021). First Draft of the Post-2020 Global Biodiversity Framework. Secretariat of the United Nations Convention on Biological Diversity. In *Cbd/Wg2020/3/3*.

Ceballos, G., Ehrlich, P. R., & Raven, P. H. (2020). Vertebrates on the brink as indicators of biological annihilation and the sixth mass extinction. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(24), 13596–13602. https://doi.org/10.1073/pnas.1922686117

Ceretta, B. F., Fogliarini, C. O., Giglio, V. J., Maxwell, M. F., Waechter, L. S., & Bender, M. G. (2020). Testing the accuracy of biological attributes in predicting extinction risk. *Perspectives in Ecology and Conservation*, *18*(1), 12–18. https://doi.org/10.1016/j.pecon.2020.01.003

Chao, Anne. (1984). Nonparametric Estimation of the Number of Classes in a Population. *Scandinavian Journal of Statistics*, *11*(4), 265–270.

Chao, A., Chiu, C.-H., & Jost, L. (2014). Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers. *Annual Review of Ecology, Evolution, and Systematics*, *45*(1), 297–324. https://doi.org/10.1146/annurev-ecolsys-120213-091540

Chao, A., Chiu, C. H., Villéger, S., Sun, I. F., Thorn, S., Lin, Y. C., … Sherwin, W. B. (2019). An attribute-diversity approach to functional diversity, functional beta diversity, and related (dis)similarity measures. *Ecological Monographs*, *89*(2), 1–29. https://doi.org/10.1002/ecm.1343

Chaudhary, C., Saeedi, H., & Costello, M. J. (2016). Bimodality of Latitudinal Gradients in Marine Species Richness. *Trends in Ecology and Evolution*, *31*(9), 670–676. https://doi.org/10.1016/j.tree.2016.06.001

Chen, X., Kong, Y., Zhang, S., Zhao, J., Li, S., & Yao, M. (2022). Comparative Evaluation of Common Materials as Passive Samplers of Environmental DNA. *Environmental Science & Technology*, *56*(15), 10798–10807.

Cherel, Y., Romanov, E. V, Annasawmy, P., Thibault, D., & Ménard, F. (2020). Micronektonic fish species over three seamounts in the southwestern Indian Ocean. *Deep-Sea Research Part II*, (March 2019). https://doi.org/10.1016/j.dsr2.2020.104777

Chichorro, F., Juslén, A., & Cardoso, P. (2019). A review of the relation between species traits and extinction risk. *Biological Conservation*, *237*(July), 220–229. https://doi.org/10.1016/j.biocon.2019.07.001

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., Maire, E., Huchery, C., MacNeil, M. A., Graham, N. A. J., Mora, C., … Mouillot, D. (2018). Gravity of human impacts mediates coral reef conservation gains. *Proceedings of the National Academy of Science*, *115*(27), 6116–6125. Retrieved from https://www.pnas.org/content/115/27/E6116.short

Clark, J. S., Nemergut, D., Seyednasrollah, B., Turner, P. J., & Zhang, S. (2017). Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data. *Ecological Monographs*, *87*(1), 34–56. https://doi.org/10.1002/ecm.1241

Claudet, J., Loiseau, C., & Pebayle, A. (2021). Critical gaps in the protection of the second largest exclusive economic zone in the world. *Marine Policy*, *124*(November 2020), 104379. https://doi.org/10.1016/j.marpol.2020.104379

Claudet, J., Loiseau, C., Sostres, M., & Zupan, M. (2020). Underprotected Marine Protected Areas in a Global Biodiversity Hotspot. *One Earth*, *2*(4), 380–384. https://doi.org/10.1016/j.oneear.2020.03.008

Claudet, J., Osenberg, C. W., Benedetti-Cecchi, L., Domenici, P., Pérez-Ruzafa, A., Badalamenti, F., … Planes, S. (2008). Marine reserves : size and age do matter. *Ecology Letters*, *11*, 481–489.

Cline, T. J., & Allgeier, J. E. (2022). Fish community structure and dynamics are insufficient to mediate coral resilience. *Nature Ecology & Evolution*, 1–10.

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.13276

Colwell, R. K., & Hurtt, G. C. (1994). Nonbiological Gradients in Species Richness and a Spurious Rapoport Effect. *The American Naturalist*, *144*(4), 570–595.

Compaire, J. C., Pérez-Brunius, P., Adelheid, J.-R. S. P., Rodriguez Outerelo, J., del Pilar Echeverri Garcia, L., & Herzka, S. Z. (2021). Connectivity of coastal and neritic fish larvae to the deep waters. *Limnology and Oceanography Letters*, *66*, 2423–2441.

Connolly, S. R., Bellwood, D. R., & Hughes, T. P. (2003). Indo-pacific biodiversity of coral reefs: Deviations from a mid-domain model. *Ecology*, *84*(8), 2178–2190. https://doi.org/10.1890/02-0254

Connolly, S. R., MacNeil, M. A., Caley, M. J., Knowlton, N., Cripps, E., Hisano, M., … Wilson, R. S. (2014). Commonness and rarity in the marine biosphere. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8524–8529. https://doi.org/10.1073/pnas.1406664111

Consuegra, S., O'Rorke, R., Rodriguez-Barreto, D., Fernandez, S., Jones, J., & de Leaniz, C. G. (2021). Impacts of large and small barriers on fish assemblage composition assessed using environmental DNA metabarcoding. *Science of The Total Environment*, *790*, 148054. https://doi.org/10.1016/j.scitotenv.2021.148054

Costello, M. J. (2014). Long live Marine Reserves: A review of experiences and benefits. *Biological Conservation*, *176*, 289–296. https://doi.org/10.1016/j.biocon.2014.04.023

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

Costello, M. J., Tsai, P., Wong, P. S., Kwok Lun Cheung, A., Basher, Z., & Chaudhary, C. (2017). Marine biogeographic realms and species endemicity. *Nature Communications*, *8*(1057), 1–11.

Cowman, P. F., & Bellwood, D. R. (2013). The historical biogeography of coral reef fishes: Global patterns of origination and dispersal. *Journal of Biogeography*, *40*(2), 209–224. https://doi.org/10.1111/jbi.12003

Crame, J. A. (2018). Key stages in the evolution of the Antarctic marine fauna. *Journal of Biogeography*, *45*(5), 986–994. https://doi.org/10.1111/jbi.13208

Crist, T. O., & Veech, J. A. (2006). Additive partitioning of rarefaction curves and species-area relationships: Unifying α-, β- and γ-diversity with sample size and habitat area. *Ecology Letters*, *9*(8), 923–932. https://doi.org/10.1111/j.1461-0248.2006.00941.x

Cure, K., Ronen, L. C., & Ben, G. (2021). Depth gradients in abundance and functional roles suggest limited depth refuges for herbivorous fishes. *Coral Reefs*. https://doi.org/10.1007/s00338-021-02060-7

Czech, L., Stamatakis, A., Dunthorn, M., & Barbera, P. (2022). Metagenomic Analysis using Phylogenetic Placement - A review of the First Decade. *ArXiv Preprint*, (March), 1–19.

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. https://doi.org/10.1016/j.cub.2014.01.049

D'Alessandro, S., & Mariani, S. (2021). Sifting environmental DNA metabarcoding data sets for rapid reconstruction of marine food webs. *Fish and Fisheries*, *00*(March), 1–12. https://doi.org/10.1111/faf.12553

Dalongeville, A., Boulanger, E., Marques, V., Charbonnel, E., Hartmann, V., Santoni, M. C., … Mouillot, D. (2022). Benchmarking eleven biodiversity indicators based on environmental DNA surveys : ore diverse functional traits and evolutionary lineages inside marine reserves. *Journal of*

Applied Ecology, *00*, 1–11.

Daniel, B., Piro, S., Charbonnel, E., Francour, P., & Letourneur, Y. (2009). Lessepsian rabbitfish Siganus luridus reached the French Mediterranean coasts. *Cybium*, *33*(2), 163–164.

Das, M. (2018). Environmental impact of trawling on the continental shelf of Bay of Bengal. *Sustainable Water Resources Management*, *4*(4), 1091–1104. https://doi.org/10.1007/s40899-018-0247-3

De Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., … Karsenti, E. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, *348*(6237), 1–11. https://doi.org/10.1007/s13398-014-0173-7.2

Dedrick, A. G., Catalano, K. A., Stuart, M. R., White, J. W., Montes, H. R., & Pinsky, M. L. (2021). Persistence of a reef fish metapopulation via network connectivity: theory and data. *Ecology Letters*, *24*(6), 1121–1132. https://doi.org/10.1111/ele.13721

Delalandre, L., Gaüzère, P., Thuiller, W., Cadotte, M., Mouquet, N., Mouillot, D., … Violle, C. (2022). Functionally distinct tree species support long-term productivity in extreme environments. *Proceedings of the Royal Society B: Biological Sciences*, *289*(1967). https://doi.org/10.1098/rspb.2021.1694

DeMartini, E., Jokiel, P., Beets, J., Stender, Y., Storlazzi, C., Minton, D., & Conklin, E. (2013). Terrigenous sediment impact on coral recruitment and growth affects the use of coral habitat by recruit parrotfishes (F. Scaridae). *Journal of Coastal Conservation*, *17*(3), 417–429. https://doi.org/10.1007/s11852-013-0247-2

Di Lorenzo, M., Guidetti, P., Di Franco, A., Calò, A., & Claudet, J. (2020). Assessing spillover from marine protected areas and its drivers: A meta-analytical approach. *Fish and Fisheries*, *21*(5), 906–915. https://doi.org/10.1111/faf.12469

Díaz, S., Settele, J., Brondízio, E. S., Ngo, H. T., Agard, J., Arneth, A., … Zayas, C. N. (2019). Pervasive human-driven decline of life on Earth points to the need for transformative change. *Science*, *366*(6471). https://doi.org/10.1126/science.aax3100

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).

Dinerstein, E., Vynne, C., Sala, E., Joshi, A. R., Fernando, S., Lovejoy, T. E., … Wikramanayake, E. (2019). A Global Deal for Nature: Guiding principles, milestones, and targets. *Science Advances*, *5*(4), 1–18. https://doi.org/10.1126/sciadv.aaw2869

Djurhuus, A., Closek, C. J., Kelly, R. P., Pitz, K. J., Michisaki, R. P., Starks, H. A., … Breitbart, M. (2020). Environmental DNA reveals seasonal shifts and potential interactions in a marine community. *Nature Communications*, *11*(254), 1–9. https://doi.org/10.1038/s41467-019-14105-1

Doble, C. J., Hipperson, H., Salzburger, W., Horsburgh, G., Mwita, C., Murrell, D. J., & Day, J. J. (2019). Testing the performance of environmental DNA metabarcoding for surveying highly diverse tropical fish communities: A case study from Lake Tanganyika. *Environmental DNA*, (October), 1–18. https://doi.org/10.1002/edn3.43

Donelson, J. M., Munday, P. L., McCormick, M. I., Pankhurst, N. W., & Pankhurst, P. M. (2010). Effects of elevated water temperature and food availability on the reproductive performance of a coral reef fish. *Marine Ecology Progress Series*, *401*, 233–243. https://doi.org/10.3354/meps08366

Doxa, A., Almpanidou, V., Katsanevakis, S., Queirós, A. M., Kaschner, K., Garilao, C., … Mazaris, A. D. (2022). 4D marine conservation networks: Combining 3D prioritization of present and future biodiversity with climatic refugia . *Global Change Biology*, (May), 1–12. https://doi.org/10.1111/gcb.16268

Duffy, G. A., & Chown, S. L. (2017). Explicitly integrating a third dimension in marine species distribution modelling. *Marine Ecology Progress Series*, *564*, 1–8. https://doi.org/10.3354/meps12011

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

Dugal, L., Thomas, L., Meenakshisundaram, A., Simpson, T., Lines, R., Colquhoun, J., … Meekan, M. (2022). Distinct coral reef habitat communities characterized by environmental DNA metabarcoding. *Coral Reefs*. https://doi.org/10.1007/s00338-022-02301-3

Dugal, L., Thomas, L., Reinholdt Jensen, M., Sigsgaard, E. E., Simpson, T., Jarman, S., … Meekan, M. (2021). Individual haplotyping of whale sharks from seawater environmental DNA. *Molecular Ecology Resources*. https://doi.org/10.1111/1755-0998.13451

Dulvy, N. K., Pacoureau, N., Rigby, C. L., Hilton-taylor, C., Fordham, S. V, & Simpfendorfer, C. A. (2021). Overfishing drives over one-third of all sharks and rays toward a global extinction crisis. *Current Biology*, *31*(21), 4773–4787. https://doi.org/10.1016/j.cub.2021.08.062

Eastman, J. T. (2005). The nature of the diversity of Antarctic fishes. *Polar Biology*, *28*(2), 93–107. https://doi.org/10.1007/s00300-004-0667-4

Eddy, T. D., Lam, V. W. Y., Reygondeau, G., Cisneros-Montemayor, A. M., Greer, K., Palomares, M.-L. D., … Cheung, W. W. L. (2021). Global decline in capacity of coral reefs to provides ecosystem services. *One Earth*, *4*, 1278–1285.

Edgar, G. J., & Stuart-Smith, R. D. (2014). Systematic global assessment of reef fish communities by the Reef Life Survey program. *Scientific Data*, *1*, 1–8. https://doi.org/10.1038/sdata.2014.7

Edgar, G. J., Stuart-Smith, R. D., Willis, T. J., Kininmonth, S., Baker, S. C., Banks, S., … Thomson, R. J. (2014). Global conservation outcomes depend on marine protected areas with five key features. *Nature*, *506*(7487), 216–220. https://doi.org/10.1038/nature13022

Edgar, R. C. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *BioRxiv*, 081257. Retrieved from https://www.biorxiv.org/content/10.1101/081257v1%0Ahttps://www.biorxiv.org/content/10.1101/081257v1.abstract

Elbrecht, V., & Leese, F. (2015). Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass-sequence relationships with an innovative metabarcoding protocol. *PLoS ONE*, *10*(7), 1–16. https://doi.org/10.1371/journal.pone.0130324

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Enquist, B. J., Feng, X., Boyle, B., Maitner, B., Newman, E. A., Jørgensen, P. M., … McGill, B. J. (2019). The commonness of rarity: Global and future distribution of rarity across land plants. *Science Advances*, *5*(11), 1–14. https://doi.org/10.1126/sciadv.aaz0414

Enzor, L. A., Zippay, M. L., & Place, S. P. (2013). High latitude fish in a high CO2 world: Synergistic effects of elevated temperature and carbon dioxide on the metabolic rates of Antarctic notothenioids. *Comparative Biochemistry and Physiology - A Molecular and Integrative Physiology*, *164*(1), 154–161. https://doi.org/10.1016/j.cbpa.2012.07.016

Epstein, G., & Roberts, C. M. (2022). Identifying priority areas to manage mobile bottom fishing on seabed carbon in the UK. *PLOS Climate*, *1*(9), 1–21.

FAO. (2018). The State of World Fisheries and Aquaculture 2018. In *Meeting the sustainable*

development goals.

Ficetola, G. F., Miaud, C., Pompanon, F., & Taberlet, P. (2008). Species detection using environmental DNA from water samples. *Biology Letters*, *4*(4), 423–425.

Fine, P. V. A. (2015). Ecological and Evolutionary Drivers of Geographic Variation in Species Diversity. *Annual Review of Ecology, Evolution, and Systematics*, *46*(October), 369–392. https://doi.org/10.1146/annurev-ecolsys-112414-054102

Finlayson, B., Somer, W. L., & Vinson, M. R. (2010). Rotenone Toxicity to Rainbow Trout and Several Mountain Stream Insects. *North American Journal of Fisheries Management*, *30*(1), 102–111. https://doi.org/10.1577/m09-078.1

Fiori, C., Paoli, C., Alessi, J., Mandich, A., & Vassallo, P. (2016). Seamount attractiveness to top predators in the southern Tyrrhenian Sea (central Mediterranean). *Journal of the Marine Biological Association of the United Kingdom*, *96*(3), 769–775. https://doi.org/10.1017/S002531541500171X

Floeter, S. R., Behrens, M. D., Ferreira, C. E. L., Paddack, M. J., & Horn, M. H. (2005). Geographical gradients of marine herbivorous fishes: Patterns and processes. *Marine Biology*, *147*(6), 1435–1447. https://doi.org/10.1007/s00227-005-0027-0

Flück, B., Mathon, L., Manel, S., Valentini, A., Dejean, T., Albouy, C., … Pellissier, L. (2022). Applying convolutional neural networks to speed up environmental DNA annotation in a highly diverse ecosystem. *Scientific Reports*, 1–13. https://doi.org/10.1038/s41598-022-13412-w

Fontoura, L., Zawada, K. J. A., D'agata, S., Álvarez-Noriega, M., Baird, A. H., Boutros, N., … Madin, E. M. P. (2020). Climate-driven shift in coral morphological structure predicts decline of juvenile reef fishes. *Global Change Biology*, *26*(2), 557–567. https://doi.org/10.1111/gcb.14911

Fox, H. E., Pet, J. S., Dahuri, R., & Caldwell, R. L. (2003). Recovery in rubble fields: Long-term impacts of blast fishing. *Marine Pollution Bulletin*, *46*(8), 1024–1031. https://doi.org/10.1016/S0025-326X(03)00246-7

Frainer, A., Primicerio, R., Kortsch, S., Aune, M., Dolgov, A. V., Fossheim, M., & Aschan, M. M. (2017). Climate-driven changes in functional biogeography of Arctic marine fish communities. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(46), 12202–12207. https://doi.org/10.1073/pnas.1706080114

Frank, K. T., Petrie, B., Leggett, W. C., & Boyce, D. G. (2018). Exploitation drives an ontogenetic-like deepening in marine fish. *PNAS*, *115*, 6422–6427.

Freeman, B. G., & Pennell, M. W. (2021). The latitudinal taxonomy gradient. *Trends in Ecology and Evolution*, *36*(9), 778–786. https://doi.org/10.1016/j.tree.2021.05.003

Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., & François, O. (2014). Fast and Efficient Estimation of Individual Ancestry Coefficients. *Genetics*, *196*, 973–983.

Froese, R., & Pauly, D. (2000). FishBase 2000: concepts, design and data sources. *ICLARM, Los Banos Laguna*.

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). https://doi.org/10.1038/s41467-017-01312-x

Gaboriau, T., Albouy, C., Descombes, P., Mouillot, D., Pellissier, L., & Leprieur, F. (2019). Ecological constraints coupled with deep- time habitat dynamics predict the latitudinal diversity gradient in reef fishes. *Proceedings of the Royal Society B*, (286).

Gardner, P. P., Watson, R. J., Stott, M. B., Morales, S. E., Morgan, X. C., Finn, R. D., & Draper, J. L. (2019). Identifying accurate metagenome and amplicon software via a meta-analysis of sequence to taxonomy benchmarking studies. *PeerJ*, *7*, 1–19. https://doi.org/10.7717/peerj.6160

Garrigue, C., Clapham, P. J., Geyer, Y., Kennedy, A. S., & Zerbini, A. N. (2015). Satellite tracking reveals novel migratory patterns and the importance of seamounts for endangered south pacific humpback whales. *Royal Society Open Science*, *2*(11). https://doi.org/10.1098/rsos.150489

Gaüzère, P., O'Connor, L., Botella, C., Poggiato, G., Münkemüller, T., Pollock, L. J., … Thuiller, W. (2022). The diversity of biotic interactions complements functional and phylogenetic facets of biodiversity. *Current Biology*, *32*(9), 2093-2100.e3. https://doi.org/10.1016/j.cub.2022.03.009

Gaynor, K. M., Hojnowski, C. E., Carter, N. H., & Brashares, J. S. (2018). The influence of human disturbance on wildlife nocturnality. *Science*, *360*(June), 1232–1235.

Gehri, R. R., Larson, W. A., Gruenthal, K., Sard, N. M., & Shi, Y. (2021). eDNA metabarcoding outperforms traditional fisheries sampling and reveals fine-scale heterogeneity in a temperate freshwater lake. *Environmental DNA*, *3*(5), 912–929. https://doi.org/10.1002/edn3.197

Gerringer, M. E., Linley, T. D., Jamieson, A. J., Goetze, E., & Drazen, J. C. (2017). Pseudoliparis swirei sp. nov.: A newly-discovered hadal snailfish (Scorpaeniformes: Liparidae) from the Mariana Trench. *Zootaxa*, *4358*(1), 161–177.

Giakoumi, S., Sciann, C., Plass-johnson, J., Micheli, F., Grorud-colvert, K., Thiriet, P., … Lubchenco, J. (2017). Ecological effects of full and partial protection in the crowded Mediterranean Sea : a regional meta-analysis. *Scientific Reports*, (November 2016), 1–12.

Giraldo-Ospina, A., Kendrick, G. A., & Hovey, R. K. (2020). Depth moderates loss of marine foundation species after an extreme marine heatwave: Could deep temperate reefs act as a refuge?: Deep thermal refuge. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1928). https://doi.org/10.1098/rspb.2020.0709rspb20200709

Goetze, J. S., Langlois, T. J., Egli, D. P., & Harvey, E. S. (2011). Evidence of artisanal fishing impacts and depth refuge in assemblages of Fijian reef fish. *Coral Reefs*, *30*(2), 507–517. https://doi.org/10.1007/s00338-011-0732-8

Goetze, Jordan S., Wilson, S., Radford, B., Fisher, R., Langlois, T. J., Monk, J., … Harvey, E. S. (2021). Increased connectivity and depth improve the effectiveness of marine reserves. *Global Change Biology*, *27*(15), 3432–3447. https://doi.org/10.1111/gcb.15635

Gold, Z., Choi, E., Kacev, D., Frable, B., Burton, R., Thompson, A., & Barber, P. (2020). FishCARD : Fish 12S California Current Specific Reference Database for Enhanced Metabarcoding Efforts. *Authorea*, 1–14.

Good, E., Holman, L., Pusceddu, A., Russo, T., Rius, M., & Iacono, C. Lo. (2022). Detection of community-wide impacts of bottom trawl fishing on deep-sea assemblages using environmental DNA metabarcoding. *Marine Pollution Bulletin*, *183*(April), 114062. https://doi.org/10.1016/j.marpolbul.2022.114062

Goodwin, K. D., Thompson, L. R., Duarte, B., Kahlke, T., Thompson, A. R., Marques, J. C., & Caçador, I. (2017). DNA Sequencing as a Tool to Monitor Marine Ecological Status. *Frontiers in Marine Science*, *4*. https://doi.org/10.3389/fmars.2017.00107

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, *27*(4), 857–871.

Guerreiro, M. A., Martinho, F., Baptista, J., Costa, F., Pardal, M. Â., & Primo, A. L. (2021). Function of estuaries and coastal areas as nursery grounds for marine fish early life stages. *Marine Environmental Research*, *170*(June). https://doi.org/10.1016/j.marenvres.2021.105408

Hadj-Hammou, J., McClanahan, T. R., & Graham, N. A. J. (2021). Decadal shifts in traits of reef fish communities in marine reserves. *Scientific Reports*, *11*(1), 1–12. https://doi.org/10.1038/s41598-021-03038-9

Hahn, E. E., Alexander, M. R., Grealy, A., Stiller, J., Gardiner, D. M., & Holleley, C. E. (2022). Unlocking inaccessible historical genomes preserved in formalin. *Molecular Ecology Resources*, *22*, 2130–2147.

Halpern, B. S., Frazier, M., Afflerbach, J., Lowndes, J. S., Micheli, F., O'Hara, C., … Selkoe, K. A. (2019). Recent pace of change in human impact on the world's ocean. *Scientific Reports*, *9*(1), 1–8. https://doi.org/10.1038/s41598-019-47201-9

Harmelin-Vivien, M. L. (2002). Energetics and fish diversity. In *The Ecology of Fishes on Coral Reefs* (pp. 265-274.).

Harrison, J. G., Calder, W. J., Shuman, B., & Buerkle, C. A. (2021). The quest for absolute abundance : The use of internal standards for DNA-based community ecology. *Molecular Ecology Resources*, *21*(August 2020), 30–43. https://doi.org/10.1111/1755-0998.13247

Hatzenbuhler, C., Kelly, J. R., Martinson, J., Okum, S., & Pilgrim, E. (2017). Sensitivity and accuracy of high- throughput metabarcoding methods for early detection of invasive fish species. *Scientific Reports*, *7*(March), 1–10.

He, X., Stanley, R. R. E., Rubidge, E. M., Jeffery, N. W., Hamilton, L. C., Westfall, K. M., … Abbott, C. L. (2022). Fish community surveys in eelgrass beds using both eDNA metabarcoding and seining: implications for biodiversity monitoring in the coastal zone. *Canadian Journal of Fisheries and Aquatic Sciences*, *1346*(February), 1–12. https://doi.org/10.1139/cjfas-2021-0215

Henriques, R., Mann, B. Q., Nielsen, E. S., Hui, C., & von der Heyden, S. (2020). Extending biodiversity conservation with functional and evolutionary diversity: a case study of South African sparid fishes. *African Journal of Marine Science*, *42*(3), 215–221. https://doi.org/10.2989/1814232X.2020.1798282

Herbert-read, J. E., Thornton, A., Amon, D. J., Birchenough, S. N. R., Côté, I. M., Dias, M. P., … Thompson, P. M. (2022). A global horizon scan of issues impacting marine and coastal biodiversity conservation. *Nature Ecology & Evolution*. https://doi.org/10.1038/s41559-022-01812-0

Hill, M. O. (1973). Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology*, *54*(2), 427–432. https://doi.org/10.2307/1934352

Huang, S., Deng, Z., Tang, G., Li, H., & Yu, T. (2021). Numerical study on blue mackerel larval transport in East China Sea. *Journal of Marine Systems*, *217*(February), 103515. https://doi.org/10.1016/j.jmarsys.2021.103515

Hupfauf, S., Etemadi, M., Juárez, M. F. D., Gómez-Brandón, M., Insam, H., & Podmirseg, S. M. (2020). CoMA – an intuitive and user-friendly pipeline for amplicon-sequencing data analysis. *PLoS ONE*, *15*(12 December), 1–28. https://doi.org/10.1371/journal.pone.0243241

Huysman, N., Voorhees, J. M., Meyer, H., Krebs, E., & Barnes, M. E. (2018). Electrofishing of Landlocked Fall Chinook Salmon Broodstock Negatively Impacts Egg Survival. *North American Journal of Aquaculture*, *80*(4), 411–417. https://doi.org/10.1002/naaq.10058

Ingvaldsen, R. B., Assmann, K. M., Primicerio, R., Fossheim, M., Polyakov, I. V., & Dolgov, A. V. (2021). Physical manifestations and ecological implications of Arctic Atlantification. *Nature Reviews Earth and Environment*, *2*(12), 874–889. https://doi.org/10.1038/s43017-021-00228-x

Irigoyen, A. J., Rojo, I., Calò, A., Trobbiani, G., Sánchez-Carnero, N., & García-Charton, J. A. (2018). The "Tracked Roaming Transect" and distance sampling methods increase the efficiency of underwater visual censuses. *PLoS ONE*, *13*(1), 1–15. https://doi.org/10.1371/journal.pone.0190990

IUCN. (2019). *Thematic Report - Conservation overview of biodiversity deep-sea Mediterranean: A strategic assessment*. Gland, Switzerland and Malaga, Spain.

Jacquemont, J., Blasiak, R., Le Cam, C., Le Gouellec, M., & Claudet, J. (2022). Ocean conservation boosts climate change mitigation and adaptation. *One Earth*, *5*, 1–13.

Jefferson, T., Costello, M. J., Zhao, Q., & Lundquist, C. J. (2021). Conserving threatened marine species and biodiversity requires 40% ocean protection. *Biological Conservation*, *264*, 109368. https://doi.org/10.1016/j.biocon.2021.109368

Jeunen, G. J., von Ammon, U., Cross, H., Ferreira, S., Lamare, M. D., Day, R., … Stanton, J.-A. L. (2022). Moving environmental DNA (eDNA) technologies from benchtop to the field using passive sampling and PDQeX extraction. *Environmental DNA*, *00*, 1–14.

Jouffray, J. B., Wedding, L. M., Norström, A. V., Donovan, M. K., Williams, G. J., Crowder, L. B., … Nyström, M. (2019). Parsing human and biophysical drivers of coral reef regimes. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1896), 1–10. https://doi.org/10.1098/rspb.2018.2544

Juhel, J. B., 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*, *287*(20200248), 1–10. https://doi.org/10.1098/rspb.2020.0248

Juhel, J., Marques, V., Utama, R. S., Vimono, I. B., Sugeha, H. Y., Kadarusman, K., … Pouyaud, L. (2022). Estimating the extended and hidden species diversity from environmental DNA in hyper-diverse regions. *Ecography*, 1–11. https://doi.org/10.1111/ecog.06299

Kasana, D., Martinez, H. D., Faux, O., Monzon, N., Guerra, E., & Chapman, D. D. (2022). First report of a sleeper shark (Somniosus sp.) in the western Caribbean, off the insular slope of a coral atoll. *Marine Biology*, *169*(8), 1–6. https://doi.org/10.1007/s00227-022-04090-3

Keck, F., Blackman, R. C., Bossart, R., Brantschen, J., Couton, M., Hürlemann, S., … Altermatt, F. (2022). Meta-analysis shows both congruence and complementarity of DNA and eDNA metabarcoding to traditional methods for biological community assessment. *Molecular Ecology*, *31*(6), 1820–1835. https://doi.org/10.1111/mec.16364

Kelkar, N., & Dey, S. (2020). Mesh mash: Legal fishing nets cause most bycatch mortality of endangered South Asian river dolphins. *Biological Conservation*, *252*(September), 108844. https://doi.org/10.1016/j.biocon.2020.108844

Kelly, R. P., Shelton, A. O., & Gallego, R. (2019). Understanding PCR Processes to Draw Meaningful Conclusions from Environmental DNA Studies. *Scientific Reports*, *9*(1), 1–14. https://doi.org/10.1038/s41598-019-48546-x

Kerry, C. R., Exeter, O. M., & Witt, M. J. (2022). Monitoring global fishing activity in proximity to seamounts using automatic identification systems. *Fish and Fisheries*, (January), 1–17. https://doi.org/10.1111/faf.12647

Kirtane, A., Atkinson, J. D., & Sassoubre, L. (2020). Design and Validation of Passive Environmental DNA Samplers Using Granular Activated Carbon and Montmorillonite Clay. *Environmental Science and Technology*, *54*(19). https://doi.org/10.1021/acs.est.0c01863

Korpinen, S., Laamanen, L., Bergstro, L., Nurmi, M., Andersen, J. H., Haapaniemi, J., … Reker, J. (2021). Combined effects of human pressures on Europe ’ s marine ecosystems. *Ambio*. https://doi.org/10.1007/s13280-020-01482-x

Kumar, G., Reaume, A. M., Farrell, E., & Gaither, M. R. (2022). Comparing eDNA metabarcoding primers for assessing fish communities in a biodiverse estuary. *PLoS ONE*, *17*(6), 1–20. https://doi.org/10.1371/journal.pone.0266720

Lacoursière-Roussel, A., & Deiner, K. (2021). Environmental DNA is not the tool by itself. *Journal of Fish Biology*, *98*(2), 383–386. https://doi.org/10.1111/jfb.14177

Langlois, T. J., Goetze, J., Bond, T., Monk, J., Abesamis, R. A., Asher, J., … Harvey, E. S. (2020). A field and video annotation guide for baited remote underwater stereo-video surveys of demersal fish assemblages. *Methods in Ecology and Evolution*, *11*, 1401–1409.

Larson, E. R., Graham, B. M., Achury, R., Coon, J. J., Daniels, M. K., Gambrell, D. K., … Suarez, A. V. (2020). From eDNA to citizen science: emerging tools for the early detection of invasive species. *Frontiers in Ecology and the Environment*, *18*(4), 194–202.

Leese, F., Altermatt, F., Bouchez, A., Ekrem, T., Hering, D., Meissner, K., … Zimmermann, J. (2016). DNAqua-Net: Developing new genetic tools for bioassessment and monitoring of aquatic ecosystems in Europe. *Research Ideas and Outcomes*, *2*, e11321. https://doi.org/10.3897/rio.2.e11321

Legendre, P., & De Cáceres, M. (2013). Beta diversity as the variance of community data: Dissimilarity coefficients and partitioning. *Ecology Letters*, *16*(8), 951–963. https://doi.org/10.1111/ele.12141

Lenoir, J., Bertrand, R., Comte, L., & … L. B. (2020). Species better track climate warming in the oceans than on land. *Nature Ecology & Evolution*, *4*, 1044–1059.

Leprieur, F., Descombes, P., Gaboriau, T., Cowman, P. F., Parravicini, V., Kulbicki, M., … Pellissier, L. (2016). Plate tectonics drive tropical reef biodiversity dynamics. *Nature Communications*, *7*(May), 1–8. https://doi.org/10.1038/ncomms11461

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*(8), e3000366. https://doi.org/10.1371/journal.pbio.3000366

Levin, N., Kark, S., & Danovaro, R. (2018). Adding the Third Dimension to Marine Conservation. *Conservation Letters*, *11*(3), 1–14. https://doi.org/10.1111/conl.12408

Li, D., Olden, J. D., Lockwood, J. L., Record, S., McKinney, M. L., & Baiser, B. (2020). Changes in taxonomic and phylogenetic diversity in the Anthropocene: Changes in biodiversity. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1929). https://doi.org/10.1098/rspb.2020.0777rspb20200777

Lin, Q., Zhang, Y., & Zhu, J. (2022). Simulating the impacts of fishing on central and eastern tropical Pacific ecosystem using multispecies size-spectrum model. *Acta Oceanologica Sinica*, *41*(3), 34–43. https://doi.org/10.1007/s13131-021-1902-3

Linard, B., Romashchenko, N., Pardi, F., & Rivals, E. (2020). PEWO: A collection of workflows to benchmark phylogenetic placement. *Bioinformatics*, *36*(21), 5264–5266. https://doi.org/10.1093/bioinformatics/btaa657

Linard, B., Swenson, K., & Pardi, F. (2019). Rapid alignment-free phylogenetic identification of metagenomic sequences. *Bioinformatics*, *35*(18), 3303–3312. https://doi.org/10.1093/bioinformatics/btz068

Lindegren, M., Holt, B. G., MacKenzie, B. R., & Rahbek, C. (2018). A global mismatch in the protection of multiple marine biodiversity components and ecosystem services. *Scientific Reports*, *8*(1), 1–8. https://doi.org/10.1038/s41598-018-22419-1

Lins, L., Zeppilli, D., Menot, L., Michel, L. N., Bonifacio, P., Brandt, M. I., … Vanreusel, A. (2021). Toward a reliable assessment of potential impact of deep-sea polymetallic nodules mining on abyssal fauna. *Limnology and Oceanography : Methods*, *19*, 626–650.

Loeza-Quintana, T., Abbott, C. L., Heath, D. D., Bernatchez, L., & Hanner, R. H. (2020). Pathway to Increase Standards and Competency of eDNA Surveys (PISCeS) - Advancing collaboration and standardization efforts in the field of eDNA. *Environmental DNAvir*, *2*, 255–260.

Loiseau, N., Thuiller, W., Stuart-smith, R. D., Devictor, V., Edgar, G. J., Velez, L., … Mouillot, D.

(2021). Maximizing regional biodiversity requires a mosaic of protection levels. *PLOS Biology*, *19*(5), 1–18. https://doi.org/10.1371/journal.pbio.3001195

Long, S., Blicher, M. E., Hammeken Arboe, N., Fuhrmann, M., Darling, M., Kemp, K. M., … Yesson, C. (2021). Deep-sea benthic habitats and the impacts of trawling on them in the offshore Greenland halibut fishery, Davis Strait, west Greenland. *ICES Journal of Marine Science*, *78*(8), 2724–2744. https://doi.org/10.1093/icesjms/fsab148

Loreau, M. (2005). *Discours de clôture. Actes de la Conférence internationale Biodiversité Science et Gouvernance. Sous la dir. de R. Barbault et J.-P. Le Duc. Paris, France.*

Luiz, O. J., Madin, J. S., Ross Robertson, D., Rocha, L. A., Wirtz, P., & Floeter, S. R. (2012). Ecological traits influencing range expansion across large oceanic dispersal barriers: Insights from tropical Atlantic reef fishes. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1730), 1033–1040. https://doi.org/10.1098/rspb.2011.1525

Maahs, B., Meyer, H., Huysman, N., Voorhees, J. M., & Barnes, M. E. (2018). Mortality of Landlocked Fall Chinook Salmon Broodstock After Electrofishing or Ascending a Fish Ladder. *Jacobs Journal of Aquaculture and Research*, *3*(1), 1–6.

Mächler, E., Walser, J., & Altermatt, F. (2021). Decision making and best practices for taxonomy-free eDNA metabarcoding in biomonitoring using Hill numbers. *Molecular Ecology*, *30*, 3326–3339.

MacNeil, M. A., Chapman, D. D., Heupel, M., Simpfendorfer, C. A., Heithaus, M., Meekan, M., … Cinner, J. E. (2020). Global status and conservation potential of reef sharks. *Nature*, *583*(7818), 801–806. https://doi.org/10.1038/s41586-020-2519-y

Magel, J. M. T., Dimoff, S. A., & Baum, J. K. (2020). Direct and indirect effects of climate change-amplified pulse heat stress events on coral reef fish communities. *Ecological Applications*, *30*(6), 1–15.

Magris, R. A., Andrello, M., Pressey, R. L., Mouillot, D., Dalongeville, A., Jacobi, M. N., & Manel, S. (2018). Biologically representative and well-connected marine reserves enhance biodiversity persistence in conservation planning. *Conservation Letters*, *11*(4), 1–10. https://doi.org/10.1111/conl.12439

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.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ*, *3*, 1–12.

Maire, E., Cinner, J., Velez, L., Huchery, C., Mora, C., Dagata, S., … Mouillot, D. (2016). How accessible are coral reefs to people? A global assessment based on travel time. *Ecology Letters*, *19*(4), 351–360.

Manea, E., Bianchelli, S., Fanelli, E., Danovaro, R., & Gissi, E. (2020). Towards an Ecosystem-Based Marine Spatial Planning in the deep Mediterranean Sea. *Science of the Total Environment*, *715*, 136884. https://doi.org/10.1016/j.scitotenv.2020.136884

Manea, E., Di Carlo, D., Depellegrin, D., Agardy, T., & Gissi, E. (2019). Multidimensional assessment of supporting ecosystem services for marine spatial planning of the Adriatic Sea. *Ecological Indicators*, *101*(December 2018), 821–837. https://doi.org/10.1016/j.ecolind.2018.12.017

Manel, S., Loiseau, N., Andrello, M., Fietz, K., Goñi, R., Forcada, A., … Mouillot, D. (2019). Long-Distance Benefits of Marine Reserves: Myth or Reality? *Trends in Ecology and Evolution*, *34*(4), 342–354. https://doi.org/10.1016/j.tree.2019.01.002

Mariani, G., Cheung, W. W. L., Lyet, A., Sala, E., Mayorga, J., Velez, L., … Mouillot, D. (2020). Let more big fish sink: Fisheries prevent blue carbon sequestration-half in unprofitable areas. *Science*

Advances, *6*(44), 1–9. https://doi.org/10.1126/sciadv.abb4848

Mariani, S., Baillie, C., Colosimo, G., & Riesgo, A. (2019). Sponges as natural environmental DNA samplers. *Current Biology*, *29*(11), R401–R402. https://doi.org/10.1016/j.cub.2019.04.031

Mariani, S., Fernandez, C., Baillie, C., Magalon, H., & Jaquemet, S. (2021). Shark and ray diversity, abundance and temporal variation around an Indian Ocean Island, inferred by eDNA metabarcoding. *Conservation Science and Practice*, (September 2020), 1–10. https://doi.org/10.1111/csp2.407

Marques, V., Castagné, P., Polanco, A., Borrero-Pérez, G. H., Hocdé, R., Guérin, P. É., … Villéger, S. (2021). Use of environmental DNA in assessment of fish functional and phylogenetic diversity. *Conservation Biology*, (May), 1–13. https://doi.org/10.1111/cobi.13802

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*, *43*, 1–12.

Marques, V., Milhau, T., Albouy, C., Dejean, T., Manel, S., Mouillot, D., & Juhel, J. (2020). GAPeDNA : Assessing and mapping global species gaps in genetic databases for eDNA metabarcoding. *Diversity and Distributions*, *00*, 1–13. https://doi.org/10.1111/ddi.13142

Martin., M. (1994). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10–12.

Martinez, C. M., Friedman, S. T., Corn, K. A., Larouche, O., Price, S. A., & Wainwright, P. C. (2021). The deep sea is a hot spot of fish body shape evolution. *Ecology Letters*, *24*(9), 1788–1799. https://doi.org/10.1111/ele.13785

Martins, G. M., Arenas, F., Neto, A. I., & Jenkins, S. R. (2012). Effects of Fishing and Regional Species Pool on the Functional Diversity of Fish Communities. *PLoS ONE*, *7*(8), 1–9. https://doi.org/10.1371/journal.pone.0044297

Marwayana, O. N., Gold, Z., & Barber, P. H. (2021). Environmental DNA in a Global Biodiversity Hotspot: Lessons from Coral Reef Fish Diversity Across the Indonesian Archipelago. *Environmental DNA*, (00), 1–17. https://doi.org/10.1002/edn3.257

Matsen, F. A., Kodner, R. B., & Armbrust, E. V. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, *11*(1), 538. https://doi.org/10.1186/1471-2105-11-538

Maxwell, M. F., Leprieur, F., Quimbayo, J. P., Floeter, S. R., & Bender, M. G. (2022). Global patterns and drivers of beta diversity facets of reef fish faunas. *Journal of Biogeography*, *49*, 954–967. https://doi.org/10.1111/jbi.14349

Maxwell, S. L., Cazalis, V., Dudley, N., Hoffmann, M., Rodrigues, A. S. L., Stolton, S., … Watson, J. E. M. (2020). Area-based conservation in the twenty-first century. *Nature*, *586*(7828), 217–227. https://doi.org/10.1038/s41586-020-2773-z

Mazzei, E. F., Pinheiro, H. T., Simon, T., Moura, R. L., Macieira, R. M., Pimentel, C. R., … Joyeux, J.-C. (2021). Mechanisms of dispersal and establishment drive a stepping stone community assembly on seamounts and oceanic islands. *Marine Biology*, *168*(7). https://doi.org/10.1007/s00227-021-03919-7

McClain, C. R., & Lundsten, L. (2015). Assemblage structure is related to slope and depth on a deep offshore Pacific seamount chain. *Marine Ecology*, *36*(2), 210–220. https://doi.org/10.1111/maec.12136

McClanahan, T. R. (2021). Marine reserve more sustainable than gear restriction in maintaining long-term coral reef fisheries yields. *Marine Policy*, *128*(November 2020), 104478.

https://doi.org/10.1016/j.marpol.2021.104478

McClure, E. C., Hoey, A. S., Sievers, K. T., Abesamis, R. A., & Russ, G. R. (2020). Relative influence of environmental factors and fishing on coral reef fish assemblages. *Conservation Biology*, *0*(0), 1–14. https://doi.org/10.1111/cobi.13636

McColl-Gausden, E. F., Weeks, A. R., Coleman, R. A., Robinson, K. L., Song, S., Raadik, T. A., & Tingley, R. (2021). Multispecies models reveal that eDNA metabarcoding is more sensitive than backpack electrofishing for conducting fish surveys in freshwater streams. *Molecular Ecology*, *30*(13), 3111–3126. https://doi.org/10.1111/mec.15644

Mcintosh, R. P. (1967). An Index of Diversity and the Relation of Certain Concepts to Diversity. *Ecology*, *48*(3), 392–404.

McLean, D. L., Taylor, M. D., Partridge, J. C., Gibbons, B., Langlois, T. J., Malseed, B. E., … Bond, T. (2018). Fish and habitats on wellhead infrastructure on the north west shelf of Western Australia. *Continental Shelf Research*, *164*, 10–27. https://doi.org/10.1016/j.csr.2018.05.007

McLean, Dianne L., Parsons, M. J. G., Gates, A. R., Benfield, M. C., Bond, T., Booth, D. J., … Jones, D. O. B. (2020). Enhancing the Scientific Value of Industry Remotely Operated Vehicles (ROVs) in Our Oceans. *Frontiers in Marine Science*, *7*(April). https://doi.org/10.3389/fmars.2020.00220

McLean, M., Auber, A., Graham, N. A. J., Houk, P., Villéger, S., Violle, C., … Mouillot, D. (2019). Trait structure and redundancy determine sensitivity to disturbance in marine fish communities. *Global Change Biology*, *25*(10), 3424–3437. https://doi.org/10.1111/gcb.14662

McLean, M., Mouillot, D., Lindegren, M., Villéger, S., Engelhard, G., Murgier, J., & Auber, A. (2019). Fish communities diverge in species but converge in traits over three decades of warming. *Global Change Biology*, *25*(11), 3972–3984. https://doi.org/10.1111/gcb.14785

McLean, M., Mouillot, D., Maureaud, A. A., Hattab, T., MacNeil, M. A., Goberville, E., … Auber, A. (2021). Disentangling tropicalization and deborealization in marine ecosystems under climate change. *Current Biology*, *31*(21), 4817-4823.e5. https://doi.org/10.1016/j.cub.2021.08.034

McLean, M., Stuart-Smith, R. D., Villéger, S., Auber, A., Edgar, G. J., MacNeil, M. A., … Mouillot, D. (2021). Trait similarity in reef fish faunas across the world's oceans. *Proceedings of the National Academy of Science*, *118*(12), e2012318118. https://doi.org/10.1073/pnas.2012318118

McLeod, E., Shaver, E. C., Beger, M., Koss, J., & Grimsditch, G. (2021). Using resilience assessments to inform the management and conservation of coral reef ecosystems. *Journal of Environmental Management*, *277*, 111384. https://doi.org/10.1016/j.jenvman.2020.111384

Medoff, S., Lynham, J., & Raynor, J. (2022). Spillover benefits from the world ' s largest fully protected MPA. *Science*, *378*(October), 313–316.

Mejía-Mercado, B. E., Mundy, B., & Baco, A. R. (2019). Variation in the structure of the deep-sea fish assemblages on Necker Island, Northwestern Hawaiian Islands. *Deep Sea Research Part I: Oceanographic Research Papers*, (January), 103086. https://doi.org/10.1016/j.dsr.2019.103086

Mellin, C., Hicks, C. C., Fordham, D. A., Golden, C. D., Kjellevold, M., MacNeil, M. A., … Graham, N. A. J. (2022). Safeguarding nutrients from coral reefs under climate change. *Nature Ecology & Evolution*, 1–10. https://doi.org/10.1038/s41559-022-01878-w

Mellin, C., Huchery, C., Caley, M. J., Meekan, M. G., & Bradshaw, C. J. A. (2010). Reef size and isolation determine the temporal stability of coral reef fish populations. *Ecology*, *91*(11), 3138–3145.

Mercader, M., Blazy, C., Di Pane, J., Devissi, C., Mercière, A., Cheminée, A., … Lenfant, P. (2019). Is artificial habitat diversity a key to restoring nurseries for juvenile coastal fish? Ex situ experiments on habitat selection and survival of juvenile seabreams. *Restoration Ecology*, *27*(5), 1155–1165.

https://doi.org/10.1111/rec.12948

Mindel, B. L., Webb, T. J., Neat, F. C., Trueman, C. N., & Blanchard, J. L. (2016). Functional, size and taxonomic diversity of fish along a depth gradient in the deep sea. *PeerJ*, *4*, e2387. https://doi.org/10.7717/peerj.2387

Mittelbach, G. G., Schemske, D. W., Cornell, H. V., Allen, A. P., Brown, J. M., Bush, M. B., … Turelli, M. (2007). Evolution and the latitudinal diversity gradient: Speciation, extinction and biogeography. *Ecology Letters*, *10*(4), 315–331. https://doi.org/10.1111/j.1461-0248.2007.01020.x

Miya, M., Gotoh, R. O., & Sado, T. (2020). MiFish metabarcoding : a high - throughput approach for simultaneous detection of multiple fish species from environmental DNA and other samples. In *Fisheries Science*. https://doi.org/10.1007/s12562-020-01461-x

Monteil, Y., Teo, A., Fong, J., Bauman, A. G., & Todd, P. A. (2020). Effects of macroalgae on coral fecundity in a degraded coral reef system. *Marine Pollution Bulletin*, *151*(December 2019), 110890. https://doi.org/10.1016/j.marpolbul.2020.110890

Montgomery, A. D., Fenner, D., Donahue, M. J., & Toonen, R. J. (2021). Community similaruty and species overlap between habitats provide insight into the deep reef refuge hypothesis. *Scientific Reports*, *11*(23787), 1–16.

Mora, C. (2015). Large-scale patterns and processes in reef fish richness. In *Ecology of fishes on coral reefs*. Cambridge University Press, Cambridge.

Mora, Camilo, & Sale, P. F. (2011). Ongoing global biodiversity loss and the need to move beyond protected areas: A review of the technical and practical shortcomings of protected areas on land and sea. *Marine Ecology Progress Series*, *434*, 251–266. https://doi.org/10.3354/meps09214

Morais, R. A., & Bellwood, D. R. (2019). Pelagic Subsidies Underpin Fish Productivity on a Degraded Coral Reef. *Current Biology*, *29*(9), 1521–1527. https://doi.org/10.1016/j.cub.2019.03.044

Morato, T., Hoyle, S. D., Allain, V., & Nicol, S. J. (2010). Seamounts are hotspots of pelagic biodiversity in the open ocean. *Proceedings of the National Academy of Sciences*, *107*(21), 9707–9711. https://doi.org/10.1073/pnas.0910290107

Morato, T., Miller, P. I., Dunn, D. C., Nicol, S. J., Bowcott, J., & Halpin, P. N. (2016). A perspective on the importance of oceanic fronts in promoting aggregation of visitors to seamounts. *Fish and Fisheries*, 1–16.

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. https://doi.org/10.1016/j.tree.2012.10.004

Mouillot, D., Villéger, S., Parravicini, V., Kulbicki, M., & Arias-gonzález, J. E. (2014). Functional over-redundancy and high functional vulnerability in global fish faunas on tropical reefs. *PNAS*, *111*(38), 13757–13762.

Mousavi-Derazmahalleh, M., Stott, A., Lines, R., Peverley, G., Nester, G., Simpson, T., … Christophersen, C. T. (2021). eDNAFlow, an automated, reproducible and scalable workflow for analysis of environmental DNA (eDNA) sequences exploiting Nextflow and Singularity. *Molecular Ecology Resources*, (August 2020), 1–8. https://doi.org/10.1111/1755-0998.13356

Mouton, T. L., Stephenson, F., Torres, L. G., Rayment, W., Brough, T., McLean, M., … Leprieur, F. (2022). Spatial mismatch in diversity facets reveals contrasting protection for New Zealand's cetacean biodiversity. *Biological Conservation*, *267*(May 2021). https://doi.org/10.1016/j.biocon.2022.109484

Muff, M., Jaquier, M., Marques, V., Ballesta, L., Deter, J., Bockel, T., … Pellissier, L. (2022).

Environmental DNA highlights fish biodiversity in mesophotic ecosystems. *Environmental DNA*, *00*(August), 1–17. https://doi.org/10.1002/edn3.358

Munday, P. L., McCormick, M. I., & Nilsson, G. E. (2012). Commentary impact of global warming and rising CO2 levels on coral reef fishes: What hope for the future? *Journal of Experimental Biology*, *215*(22), 3865–3873. https://doi.org/10.1242/jeb.074765

Munday, P. L., Pratchett, M. S., Dixson, D. L., Donelson, J. M., Endo, G. G. K., Reynolds, A. D., & Knuckey, R. (2013). Elevated CO2 affects the behavior of an ecologically and economically important coral reef fish. *Marine Biology*, *160*(8), 2137–2144. https://doi.org/10.1007/s00227-012-2111-6

Munro, J. L. (1996). The scope of tropical reef fisheries and their management. In N. V. C. Polunin & C. M. Roberts (Eds.), *Reef Fisheries* (Chapman &, pp. 1–14). https://doi.org/10.1007/978-94-015-8779-2_1

Murakami, H., Yoon, S., Kasai, A., Minamoto, T., Yamamoto, S., Sakata, M. K., … Masuda, R. (2019). Dispersion and degradation of environmental DNA from caged fish in a marine environment. *Fisheries Science*, *85*(2), 327–337. https://doi.org/10.1007/s12562-018-1282-6

Nakagawa, H., Fukushima, K., Sakai, M., Wu, L., & Minamoto, T. (2022). Relationships betweenthe eDNA concentration obtained from metabarcoding and stream fish abundance estimated by the removal method under field conditions. *Environmental DNA*, *00*, 1–12. Retrieved from papers3://publication/uuid/80664BD0-2317-4AE9-8101-2848F3291F1D

Nelson, J. S., Grande, T. C., & Wilson, M. V. H. (2016). *Fishes of the World* (John Wiley).

Nevers, M. B., Byappanahalli, M. N., Morris, C. C., Shively, D., Przybyla-Kelly, K., Spoljaric, A. M., … Roseman, E. F. (2018). Environmental DNA (eDNA): A tool for quantifying the abundant but elusive round goby (Neogobius melanostomus). *PLoS ONE*, *13*(1), 1–22. https://doi.org/10.1371/journal.pone.0191720

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*(1), 1–14. https://doi.org/10.1038/s41598-020-63565-9

O'Hara, C. C., Frazier, M., & Halpern, B. S. (2021). At-risk marine biodiversity faces extensive, expanding, and intensifying human impacts. *Science*, *372*(6537), 84–87. https://doi.org/10.1126/science.abe6731

Oliveira, M. De. (2019). Marginal reef paradox : A possible refuge from environmental changes ? *Ocean and Coastal Management*, (November), 105063. https://doi.org/10.1016/j.ocecoaman.2019.105063

Pacoureau, N., Rigby, C. L., Kyne, P. M., Sherley, R. B., Winker, H., Carlson, J. K., … Dulvy, N. K. (2021). Half a century of global decline in oceanic sharks and rays. *Nature*, *589*(7843), 567–571. https://doi.org/10.1038/s41586-020-03173-9

Parravicini, V., Bender, M. G., Villéger, S., Leprieur, F., Pellissier, L., Donati, F. G. A., … Kulbicki, M. (2021). Coral reef fishes reveal strong divergence in the prevalence of traits along the global diversity gradient. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1961), 1–7. https://doi.org/10.1098/rspb.2021.1712

Parravicini, V., Kulbicki, M., Bellwood, D. R., Friedlander, A. M., Arias-Gonzalez, J. E., Chabanet, P., … Mouillot, D. (2013). Global patterns and predictors of tropical reef fish species richness. *Ecography*, *36*(12), 1254–1262. https://doi.org/10.1111/j.1600-0587.2013.00291.x

Patarnello, T., Verde, C., di Prisco, G., Bargelloni, L., & Zane, L. (2011). How will fish that evolved at constant sub-zero temperatures cope with global warming? Notothenioids as a case study. *BioEssays*, *33*(4), 260–268. https://doi.org/10.1002/bies.201000124

Patin, N. V., Kunin, V., Lidström, U., & Ashby, M. N. (2013). Effects of OTU Clustering and PCR Artifacts on Microbial Diversity Estimates. *Microbial Ecology*, *65*(3), 709–719. https://doi.org/10.1007/s00248-012-0145-4

Pauvert, C., Buée, M., Laval, V., Edel-Hermann, V., Fauchery, L., Gautier, A., … Vacher, C. (2019). Bioinformatics matters: The accuracy of plant and soil fungal community data is highly dependent on the metabarcoding pipeline. *Fungal Ecology*, *41*, 23–33. https://doi.org/10.1016/j.funeco.2019.03.005

Pavoine, S., Vallet, J., Dufour, A.-B., Gachet, S., & Daniel, H. (2009). On the challenge of treating various types of variables: application for improving the measurement of functional diversity. *Oikos*, *118*(3), 391–402. https://doi.org/10.1111/j.1600-0706.2009.16668.x

Pawlowski, J., Apothéloz-Perret-Gentil, L., & Altermatt, F. (2020). Environmental DNA: What's behind the term? Clarifying the terminology and recommendations for its future use in biomonitoring. *Molecular Ecology*, *29*(22), 4258–4264. https://doi.org/10.1111/mec.15643

Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A., … Kahlert, M. (2018). The future of biotic indices in the ecogenomic era: Integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. *Science of the Total Environment*, *637–638*, 1295–1310. https://doi.org/10.1016/j.scitotenv.2018.05.002

Pecuchet, L., Törnroos, A., & Lindegren, M. (2016). Patterns and drivers of fish community assembly in a large marine ecosystem. *Marine Ecology Progress Series*, *546*, 239–248. https://doi.org/10.3354/meps11613

Pellissier, L., Leprieur, F., Parravicini, V., Cowman, P. F., Kulbicki, M., Litsios, G., … Mouillot, D. (2014). Quaternary coral reef refugia preserved fish diversity. *Science*, *344*(6187), 1016–1020.

Pereira, P. H. C., Macedo, C. H., Nunes, J. de A. C. C., Marangoni, L. F. de B., & Bianchini, A. (2018). Effects of depth on reef fish communities: Insights of a "deep refuge hypothesis" from Southwestern Atlantic reefs. *PLoS ONE*, *13*(9), 1–20. https://doi.org/10.1371/journal.pone.0203072

Pimiento, C., Leprieur, F., Silvestro, D., Lefcheck, J. S., Albouy, C., Rasher, D. B., … Griffin, J. N. (2020). Functional diversity of marine megafauna in the Anthropocene. *Science Advances*, *6*(16), eaay7650. https://doi.org/10.1126/sciadv.aay7650

Pinsky, M. L., Selden, R. L., & Kitchel, Z. J. (2020). Climate-Driven Shifts in Marine Species Ranges: Scaling from Organisms to Communities. *Annual Review of Marine Science*, *12*, 153–179. https://doi.org/10.1146/annurev-marine-010419-010916

Pitcher, T. J., & Lam, M. E. (2014). Fish commoditization and the historical origins of catching fish for profit. *Maritime Studies*, *14*(1). https://doi.org/10.1186/s40152-014-0014-5

Plough, L. V., Bunch, A. J., Lee, B. B., Fitzgerald, C. L., Stence, C. P., & Richardson, B. (2021). Development and testing of an environmental DNA (eDNA) assay for endangered Atlantic sturgeon to assess its potential as a monitoring and management tool. *Environmental DNA*, *3*, 800–814.

Podani, J. (1999). Extending Gower's general coefficient of similarity to ordinal characters. *Taxon*, *48*(2), 331–340. https://doi.org/10.2307/1224438

Polanco F, A., Waldock, C., Keggin, T., Marques, V., Valentini, A., Dejean, T., … Vermeij, M. (2022). Ecological indices from environmental DNA to contrast coastal reefs under different anthropogenic pressures. *Ecology and Evolution*, *12*, 1–17.

Polanco Fernández, A., Marques, V., Fopp, F., Juhel, J., Borrero-Pérez, G. H., Cheutin, M., … Pellissier, L. (2021). Comparing environmental DNA metabarcoding and underwater visual census to monitor tropical reef fishes. *Environmental DNA*, *3*(1), 142–156. https://doi.org/10.1002/edn3.140

Pont, D., Meulenbroek, P., Bammer, V., Dejean, T., Eros, T., Jean, P., … Valentini, A. (2022). Quantitative monitoring of diverse fish communities on a large scale combining eDNA metabarcoding and qPCR. *Molecular Ecology Resources*.

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. https://doi.org/10.1038/s41598-018-28424-8

Pont, D., Valentini, A., Rocle, M., Delaigue, O., Jean, P., & Dejean, T. (2019). The future of fish-based ecological assessment of European rivers : from traditional EU Water Framework Directive compliant methods to eDNA metabarcoding-based approaches. *J Fish Biol. Accepted Author Manuscript.*, 1–50. https://doi.org/10.1111/1744-1633.12020

Pörtner, H. O., & Knust, R. (2007). Climate Change Affects Marine Fishes Through the Oxygen Limitation of Thermal Tolerance. *Science*, *315*, 95–97. https://doi.org/10.1259/0007-1285-53-633-920-b

Pottier, G., Beaumont, W. R., Marchand, F., Le Bail, P. Y., Azam, D., Rives, J., … Roussel, J. M. (2020). Electrofishing in streams of low water conductivity but high biodiversity value: Challenges, limits and perspectives. *Fisheries Management and Ecology*, *27*(1), 52–63. https://doi.org/10.1111/fme.12384

Prodan, A., Tremaroli, V., Brolin, H., Zwinderman, A. H., Nieuwdorp, M., & Levin, E. (2020). Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PLoS ONE*, *15*(1), 1–19. https://doi.org/10.1371/journal.pone.0227434

Proud, R., Cox, M. J., Le Guen, C., & Brierley, A. S. (2018). Fine-scale depth structure of pelagic communities throughout the global ocean based on acoustic sound scattering layers. *Marine Ecology Progress Series*, *598*, 35–48.

Quattrini, A. M., Demopoulos, A. W. J., Singer, R., Roa-Varon, A., & Chaytor, J. D. (2017). Demersal fish assemblages on seamounts and other rugged features in the northeastern Caribbean. *Deep-Sea Research Part I*, *123*(March), 90–104. https://doi.org/10.1016/j.dsr.2017.03.009

Rabinowitz, D. (1981). Seven forms of rarity. In H. Synge (Ed.), *The biological aspects of rare plant conservation* (pp. 205–217). Chichester: John Wiley & Sons.

Rabosky, D. L., Chang, J., Title, P. O., Cowman, P. F., Sallan, L., Friedman, M., … Alfaro, M. E. (2018). An inverse latitudinal gradient in speciation rate for marine fishes. *Nature*, *559*(7714), 392–395. https://doi.org/10.1038/s41586-018-0273-1

Ratcliffe, F. C., Uren Webster, T. M., Garcia de Leaniz, C., & Consuegra, S. (2020). A drop in the ocean: Monitoring fish communities in spawning areas using environmental DNA. *Environmental DNA*, (March), 1–12. https://doi.org/10.1002/edn3.87

Reboredo Segovia, A. L., Romano, D., & Armsworth, P. R. (2020). Who studies where ? Boosting tropical conservation research where it is most needed. *Frontiers in Ecology and the Environment*, *18*(3), 159–166.

Rivera, S. F., Rimet, F., Vasselon, V., Vaultier, M., Domaizon, I., & Bouchez, A. (2022). Fish eDNA metabarcoding from aquatic biofilm samples: Methodological aspects. *Molecular Ecology Resources*, *22*, 1440–1453.

Roberson, L., Wilcox, C., Boussarie, G., Dugan, E., Garilao, C., Gonzalez, K., … Kiszka, J. J. (2022). Spatially explicit risk assessment of marine megafauna vulnerability to Indian Ocean tuna fisheries. *Fish and Fisheries*, (May), 1–22. https://doi.org/10.1111/faf.12676

Rocha, L. A., Pinheiro, H. T., Shepherd, B., Papastamatiou, Y. P., Luiz, O. J., Pyle, R. L., & Bongaerts, P. (2018). Mesophotic coral ecosystems are threatened and ecologically distinct from shallow water reefs. *Science*, *361*(6399), 281–284. https://doi.org/10.1126/science.aaq1614

Rodgers, G. G., Donelson, J. M., McCormick, M. I., & Munday, P. L. (2018). In hot water: sustained ocean warming reduces survival of a low-latitude coral reef fish. *Marine Biology*, *165*(4), 1–10. https://doi.org/10.1007/s00227-018-3333-z

Rogers, A. D. (1994). The Biology of Seamounts. *Advances in Marine Biology*, *30*(C), 305–350. https://doi.org/10.1016/S0065-2881(08)60065-6

Rogers, Alex D. (2018). The Biology of Seamounts: 25 Years on. In *Advances in Marine Biology* (1st ed., Vol. 79). https://doi.org/10.1016/bs.amb.2018.06.001

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, *4*, 1–22. https://doi.org/10.7717/peerj.2584

Rojo, I., Anadon, J. D., & Garcia-Charton, J. A. (2021). Exceptionally high but still growing predatory reef fish biomass after 23 years of protection in a Marine Protected Area. *PLoS ONE*, *16*(2), 1–20.

Romdal, T. S., Araújo, M. B., & Rahbek, C. (2013). Life on a tropical planet: Niche conservatism and the global diversity gradient. *Global Ecology and Biogeography*, *22*(3), 344–350. https://doi.org/10.1111/j.1466-8238.2012.00786.x

Rourke, M. L., Fowler, A. M., Hughes, J. M., Broadhurst, M. K., Dibattista, J. D., Fielder, S., … Elise, W. (2022). Environmental DNA (eDNA) as a tool for assessing fish biomass : A review of approaches and future considerations for resource surveys. *Environmental DNA*, *4*(July 2020), 9–33. https://doi.org/10.1002/edn3.185

Rowden, A. A., Schlacher, T. A., Williams, A., Clark, M. R., Stewart, R., Althaus, F., … Dowdney, J. (2010). A test of the seamount oasis hypothesis: Seamounts support higher epibenthic megafaunal biomass than adjacent slopes. *Marine Ecology*, *31*(SUPPL. 1), 95–106. https://doi.org/10.1111/j.1439-0485.2010.00369.x

Rummer, J. L., Couturier, C. S., Stecyk, J. A. W., Gardiner, N. M., Kinch, J. P., Nilsson, G. E., & Munday, P. L. (2014). Life on the edge: Thermal optima for aerobic scope of equatorial reef fishes are close to current day temperatures. *Global Change Biology*, *20*(4), 1055–1066. https://doi.org/10.1111/gcb.12455

Sadovy de Mitcheson, Y., Craig, M. T., Bertoncini, A. A., Carpenter, K. E., Cheung, W. W. L., Choat, J. H., … Sanciangco, J. (2013). Fishing groupers towards extinction: A global assessment of threats and extinction risks in a billion dollar fishery. *Fish and Fisheries*, *14*(2), 119–136. https://doi.org/10.1111/j.1467-2979.2011.00455.x

Saito, T., & Doi, H. (2021). Degradation modeling of water environmental DNA: Experiments on multiple DNA sources in pond and seawater. *Environmental DNA*, 1–28. https://doi.org/10.1002/edn3.192

Sala, E., & Giakoumi, S. (2018). No-take marine reserves are the most effective protected areas in the ocean. *ICES Journal of Marine Science*, *75*(3), 1166–1168. https://doi.org/10.1093/icesjms/fsx059

Sala, E., Mayorga, J., Bradley, D., Cabral, R. B., Atwood, T. B., Auber, A., … Mouillot, D. (2021). Protecting the global ocean for biodiversity , food and climate. *Nature*. https://doi.org/10.1038/s41586-021-03371-z

Sanchez, L., Boulanger, E., Boissery, P., Dalongeville, A., Dejean, T., Deter, J., … Mouillot, D. (2022). Ecological indicators based on quantitative eDNA metabarcoding : the case of marine reserves. *Ecological Indicators*, *140*(May). https://doi.org/10.1016/j.ecolind.2022.108966

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Sassoubre, L. M., Yamahara, K. M., Gardner, L. D., Block, B. A., & Boehm, A. B. (2016).

Quantification of environmental DNA (eDNA) shedding and decay rates for three marine fish. *Environmental Science & Technology*. https://doi.org/10.1021/acs.est.6b03114

Säterberg, T., Sellman, S., & Ebenman, B. (2013). High frequency of functional extinctions in ecological networks. *Nature*, *499*(7459), 468–470. https://doi.org/10.1038/nature12277

Sato, M., Inoue, N., Nambu, R., Furuichi, N., Imaizumi, T., & Ushio, M. (2021). Quantitative assessment of multiple fish species around artificial reefs using environmental DNA metabarcoding. *Scientific Reports*, (0123456789), 6. https://doi.org/10.1038/s41598-021-98926-5

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. https://doi.org/10.1111/1755-0998.12402

Schramm, K. D., Marnane, M. J., Elsdon, T. S., Jones, C., Saunders, B. J., Goetze, J. S., … Harvey, E. S. (2020). A comparison of stereo-BRUVs and stereo-ROV techniques for sampling shallow water fish communities on and off pipelines. *Marine Environmental Research*, *162*(October), 105198. https://doi.org/10.1016/j.marenvres.2020.105198

Schweiss, K. E., Lehman, R. N., Drymon, J. M., & Phillips, N. M. (2019). Development of highly sensitive environmental DNA methods for the detection of Bull Sharks , Carcharhinus leucas (Müller and Henle , 1839), using Droplet Digital ™ PCR. *Environmental DNA*, (September), 1–10. https://doi.org/10.1002/edn3.39

Seymour, M., Edwards, F. K., Cosby, B. J., Bista, I., Scarlett, P. M., Brailsford, F. L., … Creer, S. (2021). Environmental DNA provides higher resolution assessment of riverine biodiversity and ecosystem turnover partitioning. *Communications Biology*, *4*(512). https://doi.org/10.1038/s42003-021-02031-2

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Sharma, D., Rao, K., & Ramanathan, A. (2022). A Systematic Review on the Impact of Urbanization and Industrialization on Indian Coastal Mangrove Ecosystem. In *Coastal Ecosystems* (pp. 175–199). https://doi.org/10.1007/978-3-030-84255-0_8

Sigsgaard, E. E., Nielsen, I. B., Bach, S. S., Lorenzen, E. D., Robinson, D. P., Knudsen, S. W., … Thomsen, P. F. (2016). Population characteristics of a large whale shark aggregation inferred from seawater environmental DNA. *Nature Ecology & Evolution*, *1*(1), 0004. https://doi.org/10.1038/s41559-016-0004

Simon, T., Pinheiro, H. T., Santos, S., Macieira, R. M., Ferreira, Y. S. S., Bernardi, G., … Joyeux, J.-C. (2021). Comparative phylogeography of reef fishes indicates seamounts as stepping stones for dispersal and diversification. *Coral Reefs*. https://doi.org/10.1007/s00338-021-02178-8

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*(1), 109–116.

Simpson, E. (1949). Measurment of Diversity. *Nature*, *163*(1943), 688. Retrieved from https://doi.org/10.1038/163688a0

Sing Wong, A., Vrontos, S., & Taylor, M. L. (2022). An assessment of people living by coral reefs over space and time. *Global Change Biology*, *00*, 1–15.

Siqueira, A. C., Morais, R. A., Bellwood, D. R., & Cowman, P. F. (2021). Planktivores as trophic drivers of global coral reef fish diversity patterns. *PNAS*, *118*(9). https://doi.org/10.1073/pnas.2019404118

Skelton, J., Cauvin, A., & Hunter, M. E. (2022). Environmental DNA metabarcoding read numbers and

their variability predict species abundance but weakly in non-dominant species. *Environmental DNA*, *00*, 1–13.

Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next-Generation Sequencing Technologies. *Current Protocols in Molecular Biology*, *122*(1), 1–11. https://doi.org/10.1002/cpmb.59

Soares, M. de O., Araújo, J. T. de, Ferreira, S. M. C., Santos, B. A., Boavida, J. R. H., Costantini, F., & Rossi, S. (2020). Why do mesophotic coral ecosystems have to be protected? *Science of the Total Environment*, *726*, 138456. https://doi.org/10.1016/j.scitotenv.2020.138456

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), 573–583. https://doi.org/10.1641/b570707

Spear, M. J., Embke, H. S., Krysan, P. J., & Vander Zanden, M. J. (2020). Application of eDNA as a tool for assessing fish population abundance. *Environmental DNA*, (March), 1–9. https://doi.org/10.1002/edn3.94

Stauffer, S., Jucker, M., Keggin, T., Marques, V., Andrello, M., Bessudo, S., … Mouillot, D. (2021). How many replicates to accurately estimate fish biodiversity using environmental DNA on coral reefs ? Authors : *Ecology and Evolution*, *11*, 14630–14643.

Stein, A., Gerstner, K., & Kreft, H. (2014). Environmental heterogeneity as a universal driver of species richness across taxa, biomes and spatial scales. *Ecology Letters*, *17*(7), 866–880. https://doi.org/10.1111/ele.12277

Stoeckle, M. Y., Adolf, J., Charlop-Powers, Z., Dunton, K. J., Hinks, G., & VanMorter, S. M. (2021). Trawl and eDNA assessment of marine fish diversity, seasonality, and relative abundance in coastal New Jersey, USA. *ICES Journal of Marine Science*, *78*(1), 293–304.

Straube, N., Lyra, M. L., Paijmans, J. L. A., Preick, M., Basler, N., Penner, J., … Hofreiter, M. (2021). Successful application of ancient DNA extraction and library construction protocols to museum wet collection specimens. *Molecular Ecology Resources*, *21*(7), 2299–2315. https://doi.org/10.1111/1755-0998.13433

Stuart-smith, J., Edgar, G. J., Last, P., Linardich, C., Lynch, T., Barrett, N., … Stuart-smith, R. D. (2020). Conservation challenges for the most threatened family of marine bony fishes (handfishes : Brachionichthyidae). *Biological Conservation*, *252*(September), 108831. https://doi.org/10.1016/j.biocon.2020.108831

Stuart-Smith, R. D., Mellin, C., Bates, A. E., & Edgar, G. J. (2021). Habitat loss and range shifts contribute to ecological generalization among reef fishes. *Nature Ecology and Evolution*, *5*(5), 656–662. https://doi.org/10.1038/s41559-020-01342-7

Stuart-Smith, R. D., Edgar, G. J., Clausius, E., Oh, E. S., Barrett, N. S., Emslie, M. J., … Mellin, C. (2022). Tracking widespread climate-driven change on temperate and tropical reefs. *Current Biology*, *32*, 1–11.

Sward, D., Monk, J., & Barrett, N. (2019). A systematic review of remotely operated vehicle surveys for visually assessing fish assemblages. *Frontiers in Marine Science*, *6*(APR), 1–19. https://doi.org/10.3389/fmars.2019.00134

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), 1–9.

Tillotson, M. D., Kelly, R. P., Duda, J. J., Hoy, M., Kralj, J., & Quinn, T. P. (2018). Concentrations of environmental DNA (eDNA) reflect spawning salmon abundance at fine spatial and temporal scales. *Biological Conservation*, *220*(July 2017), 1–11. https://doi.org/10.1016/j.biocon.2018.01.030

Tornabene, L., Valdez, S., Erdmann, M., & Pezold, F. (2015). Support for a "Center of Origin" in the Coral Triangle: Cryptic diversity, recent speciation, and local endemism in a diverse lineage of reef fishes (Gobiidae: Eviota). *Molecular Phylogenetics and Evolution*, *82*(PA), 200–210. https://doi.org/10.1016/j.ympev.2014.09.012

Trenkel, V. M., Vaz, S., Albouy, C., Amour, A. B., Duhamel, E., Laffargue, P., … Lorance, P. (2019). We can reduce the impact of scientific trawling on marine ecosystems. *Marine Ecology Progress Series*, *609*(January), 277–282. https://doi.org/10.3354/meps12834

Trindade-Santos, I., Moyes, F., & Magurran, A. E. (2022). Global patterns in functional rarity of marine fish. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-28488-1

Trip, E. D. L., Clements, K. D., Raubenheimer, D., & Choat, J. H. (2014). Temperature-related variation in growth rate, size, maturation and life span in a marine herbivorous fish over a latitudinal gradient. *Journal of Animal Ecology*, *83*, 866–875.

Truelove, N. K., Patin, N. V., Min, M., Pitz, K. J., Preston, C. M., Yamahara, K. M., … Chavez, F. P. (2022). Expanding the temporal and spatial scales of environmental DNA research with autonomous sampling. *Environmental DNA*, 1–13.

Tu, C. Y., Chen, K. T., & Hsieh, C. H. (2018). Fishing and temperature effects on the size structure of exploited fish stocks. *Scientific Reports*, *8*(1), 1–10. https://doi.org/10.1038/s41598-018-25403-x

Turner, C. R., Barnes, M. A., Xu, C. C. Y., Jones, S. E., Jerde, C. L., & Lodge, D. M. (2014). Particle size distribution and optimal capture of aqueous macrobial eDNA. *Methods in Ecology and Evolution*, *5*(7), 676–684. https://doi.org/10.1111/2041-210X.12206

Turon, M., Angulo-Preckler, C., Antich, A., Praebel, K., & Wangensteen, O. S. (2020). More Than Expected From Old Sponge Samples : A Natural Sampler DNA Metabarcoding Assessment of Marine Fish Diversity in Nha Trang Sponge Samples and Study Site. *Frontiers in Marine Science*, *7*(December), 1–14. https://doi.org/10.3389/fmars.2020.605148

Turschwell, M. P., Connolly, R. M., Dunic, J. C., Sievers, M., Buelow, C. A., Pearson, R. M., … Brown, C. J. (2021). Anthropogenic pressures and life history predict trajectories of seagrass meadow extent at a global scale. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(45), 1–11. https://doi.org/10.1073/pnas.2110802118

Valdivia-Carrillo, T., Rocha-Olivares, A., Reyes-Bonilla, H., Domínguez-Contreras, J. F., & Munguia-Vega, A. (2021). Integrating eDNA metabarcoding and simultaneous underwater visual surveys to describe complex fish communities in a marine biodiversity hotspot. *Molecular Ecology Resources*, (March), 1558–1574. https://doi.org/10.1111/1755-0998.13375

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

Van Oppen, M. J. H., & Coleman, M. A. (2022). Advancing the protection of marine life through genomics. *PLOS Biology*, *20*(10), 1–17. https://doi.org/10.1371/journal.pbio.3001801

Venegas-Li, R., Levin, N., Possingham, H., & Kark, S. (2018). 3D spatial conservation prioritisation: Accounting for depth in marine environments. *Methods in Ecology and Evolution*, *9*(3), 773–784. https://doi.org/10.1111/2041-210X.12896

Verdier, H., Konecny-Dupre, L., Marquette, C., Reveron, H., Tadier, S., Grémillard, L., … Lefébure, T. (2022). Passive sampling of environmental DNA in aquatic environments using 3D-printed

hydroxyapatite samplers. *Molecular Ecology Resources*, (May 2021), 1–13. https://doi.org/10.1111/1755-0998.13604

Veron, J. E. N., Devantier, L. M., Turak, E., Green, A. L., Kininmonth, S., Stafford-Smith, M., & Peterson, N. (2009). Delineating the Coral Triangle. *Galaxea, Journal of Coral Reef Studies*, *11*(2), 91–100. https://doi.org/10.3755/galaxea.11.91

Victor, B. C. (2015). How many coral reef fish species are there? Cryptic diversity and the new molecular taxonomy. In *Ecology of fishes on coral reefs. Cambridge University Press, Cambridge* (pp. 76–88).

Violle, C., Thuiller, W., Mouquet, N., Munoz, F., Kraft, N. J. B., Cadotte, M. W., … Mouillot, D. (2017). Functional Rarity: The Ecology of Outliers. *Trends in Ecology and Evolution*, *32*(5), 356–367. https://doi.org/10.1016/j.tree.2017.02.002

Watling, L., & Auster, P. J. (2017). Seamounts on the high seas should be managed as vulnerable marine ecosystems. *Frontiers in Marine Science*, *4*(JAN), 1–4. https://doi.org/10.3389/fmars.2017.00014

Weltz, K., Lyle, J. M., Ovenden, J., Morgan, J. A. T., 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), 1–16. https://doi.org/10.1371/journal.pone.0178124

West, K., Travers, M. J., Stat, M., Harvey, E. S., Richards, Z. T., Dibattista, J. D., … Bunce, M. (2021). Large-scale eDNA metabarcoding survey reveals marine biogeographic break and transitions over tropical north- western Australia. *Diversity and Distributions*, (00), 1–16. https://doi.org/10.1111/ddi.13228

Williams, A., Althaus, F., Maguire, K., Green, M., Untiedt, C., Alderslade, P., … Schlacher, T. A. (2020). The Fate of Deep-Sea Coral Reefs on Seamounts in a Fishery-Seascape: What Are the Impacts, What Remains, and What Is Protected? *Frontiers in Marine Science*, *7*(September). https://doi.org/10.3389/fmars.2020.567002

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. *Science*, *323*(January), 359–362.

Woodcock, P., O'Leary, B. C., Kaiser, M. J., & Pullin, A. S. (2017). Your evidence or mine? Systematic evaluation of reviews of marine protected area effectiveness. *Fish and Fisheries*, *18*(4), 668–681. https://doi.org/10.1111/faf.12196

Woolley, S. N. C., Foster, S. D., Bax, N. J., Currie, J. C., Dunn, D. C., Hansen, C., … Dunstan, P. K. (2020). Bioregions in Marine Environments: Combining Biological and Environmental Data for Management and Scientific Understanding. *BioScience*, *70*(1), 48–59. https://doi.org/10.1093/biosci/biz133

Worm, B., & Lotze, H. K. (2021). Marine biodiversity and climate change. In *Climate change* (pp. 445–464). Elsevier.

Wu, Y., Colborne, S. F., Charron, M. R., & Heath, D. D. (2022). Development and validation of targeted environmental DNA (eDNA) metabarcoding for early detection of 69 invasive fishes and aquatic invertebrates. *Environmental DNA*, *00*, 1–12.

Yan, H. F., Kyne, P. M., Jabado, R. W., Leeney, R. H., Davidson, N. K., Derrick, D. H., … Dulvy, N. K. (2021). Overfishing and habitat loss drives range contraction of iconic marine fishes to near extinction. *Science Advances*, *7*, 1–10.

Yao, M., Zhang, S., Lu, Q., Chen, X., Zhang, S.-Y., Kong, Y., & Zhao, J. (2022). Fishing for fish environmental DNA: Ecological applications, methodological considerations, surveying designs, and ways forward. *Molecular Ecology*, (00), 1–33.

Yesson, C., Clark, M. R., Taylor, M. L., & Rogers, A. D. (2011). The global distribution of seamounts

based on 30 arc seconds bathymetry data. *Deep-Sea Research Part I: Oceanographic Research Papers*, *58*(4), 442–453. https://doi.org/10.1016/j.dsr.2011.02.004

Yoshitake, K., Fujiwara, A., Matsuura, A., Sekino, M., Yasuike, M., Nakamura, Y., … Watabe, S. (2021). Estimation of tuna population by the improved analytical pipeline of unique molecular identifier-assisted HaCeD-Seq (haplotype count from eDNA). *Scientific Reports*, *11*(1), 1–12. https://doi.org/10.1038/s41598-021-86190-6

Young, H. S., Mccauley, D. J., Galetti, M., & Dirzo, R. (2016). Patterns, Causes, and Consequences of Anthropocene Defaunation. *Annual Review of Ecology, Evolution, and Systematics*, *47*(1), 333–358. https://doi.org/10.1146/annurev-ecolsys-112414-054142

Zarco-Perello, S., Pratchett, M., & Liao, V. (2012). Temperature-growth performance curves for a coral reef fish, Acanthochromis polyacanthus. *Galaxea, Journal of Coral Reef Studies*, *14*(1), 97–103. https://doi.org/10.3755/galaxea.14.97

Zhang, S., Zhao, J., & Yao, M. (2020). A comprehensive and comparative evaluation of primers for metabarcoding eDNA from fish. *Methods in Ecology and Evolution*, *2020*(January), 1609–1625. https://doi.org/10.1111/2041-210X.13485

Zhao, Q., Stephenson, F., Lundquist, C., Kaschner, K., Jayathilake, D., & Costello, M. J. (2020). Where Marine Protected Areas would best represent 30% of ocean biodiversity. *Biological Conservation*, *244*(March), 108536. https://doi.org/10.1016/j.biocon.2020.108536

Annexes

1. Manuscrit A1

Publié dans Scientific Reports :

Flück, B., Mathon, L., Manel, S. *et al.* Applying convolutional neural networks to speed up environmental DNA annotation in a highly diverse ecosystem. *Sci Rep* 12, 10247 (2022). https://doi.org/10.1038/s41598-022-13412-w

www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

Applying convolutional neural networks to speed up environmental DNA annotation in a highly diverse ecosystem

Benjamin Flück[1,2]✉, Laëtitia Mathon[3], Stéphanie Manel[3], Alice Valentini[4], Tony Dejean[4], Camille Albouy[5], David Mouillot[6,7], Wilfried Thuiller[8], Jérôme Murienne[9], Sébastien Brosse[9] & Loïc Pellissier[1,2]✉

High-throughput DNA sequencing is becoming an increasingly important tool to monitor and better understand biodiversity responses to environmental changes in a standardized and reproducible way. Environmental DNA (eDNA) from organisms can be captured in ecosystem samples and sequenced using metabarcoding, but processing large volumes of eDNA data and annotating sequences to recognized taxa remains computationally expensive. Speed and accuracy are two major bottlenecks in this critical step. Here, we evaluated the ability of convolutional neural networks (CNNs) to process short eDNA sequences and associate them with taxonomic labels. Using a unique eDNA data set collected in highly diverse Tropical South America, we compared the speed and accuracy of CNNs with that of a well-known bioinformatic pipeline (OBITools) in processing a small region (60 bp) of the 12S ribosomal DNA targeting freshwater fishes. We found that the taxonomic labels from the CNNs were comparable to those from OBITools, with high correlation levels for the composition of the regional fish fauna. The CNNs enabled the processing of raw fastq files at a rate of approximately 1 million sequences per minute, which was about 150 times faster than with OBITools. Given the good performance of CNNs in the highly diverse ecosystem considered here, the development of more elaborate CNNs promises fast deployment for future biodiversity inventories using eDNA.

Effective ecosystem governance and management require an increase in speed, accuracy and ease of collecting and processing of biodiversity data[31,49]. Biodiversity data collection requires a shift in focus from expert monitoring towards high-throughput data acquisition technology[24]. Conventional biodiversity monitoring approaches are labor intensive, depend on expert knowledge—resulting in long delays between sampling and results[53], and miss many species that are either small, rare, cryptic or elusive[41], which in turn hinders accurate ecological interpretations. Fortunately, our ability to rapidly generate inventories of whole species communities is improving with the emergence of environmental genomics, specifically environmental DNA (eDNA)[6,25,27,75]. All organisms living in an ecosystem shed tissue material, which can be detected through eDNA metabarcoding[74], offering an integrative view of the ecosystem composition[27,33]. Coupled with high-throughput DNA sequencing methods, eDNA metabarcoding can help with the rapid assessment and monitoring of biodiversity across all levels of life, from prokaryotes to eukaryotes[40], with a higher detection capacity and cost-effectiveness than traditional methods[59]. The reads from high-throughput amplicon sequencing of eDNA can be compared with reference barcode libraries, enabling the establishment of taxonomic lists directly from environment samples[74]. Ultimately, these lists can be used to assess ecosystem functioning and health status[25]. With an increasing number of initiatives proposing

[1]Department of Environmental System Science, ETH Zürich, 8092 Zurich, Switzerland. [2]Swiss Federal Research Institute WSL, 8903 Birmensdorf, Switzerland. [3]CEFE, Univ. Montpellier, CNRS, EPHE-PSL University, IRD, Montpellier, France. [4]SPYGEN, Le Bourget-du-Lac, France. [5]DECOD (Ecosystem Dynamics and Sustainability), IFREMER, INRAE, Institut Agro - Agrocampus Ouest, Rue de l'Ile d'Yeu, BP21105, 44311 Nantes Cedex 3, France. [6]MARBEC, Univ. Montpellier,CNRS, IRD, Ifremer, Montpellier, France. [7]Institut Universitaire de France, IUF, 75231 Paris, France. [8]CNRS, LECA, Laboratoire d'Écologie Alpine, Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, 38000 Grenoble, France. [9]Laboratoire Evolution et Diversité Biologique (UMR5174), CNRS, IRD, Université Paul Sabatier, Toulouse, France. ✉email: benjamin.flueck@usys.ethz.ch; loic.pellissier@usys.ethz.ch

the use of eDNA metabarcoding routinely and globally to monitor ecosystems[5], the processing of such massive sequencing data will require novel automated bioinformatic solutions that are both fast and accurate.

As the laboratory molecular steps of eDNA metabarcoding have gained in efficiency[69,75], the major bottleneck and technical challenge has shifted from the development of efficient sampling and laboratory protocols to the processing of the produced large set of raw sequencing data into taxonomic lists[32]. In particular, eDNA metabarcoding amplifies small DNA sequences ('barcodes'), typically 60–300 bp long, from the mitochondrial genome for use with Illumina sequencing technology[71]. This sequencing process generates a huge quantity of small sequence reads that require fast and accurate bioinformatic processing to be interpreted[34,35]. The bioinformatic processing includes several steps (merging the forward and reverse reads, demultiplexing, dereplicating, filtering by quality, removing errors), after which the retained and cleaned sequences are assigned to a taxonomic label[32,50,56]. Taxonomic labelling then involves transforming sequence reads from eDNA into lists of taxa that can be used by experts and scientists to understand biodiversity patterns, structures and dynamics of assemblages. They can additionally be used for management decisions[68], based on the detection of rare[9,63], endangered[37], or invasive species[68]. Given that most existing pipelines are time consuming to apply[52], efficient algorithms transforming eDNA reads into accurate taxonomic lists using machine learning could potentially enable efficient and parallel automatization on cloud infrastructure for a broad application of eDNA technology[65].

Compared with traditional bioinformatic approaches[52], machine learning could increase the efficiency and capacity of the taxonomic labelling of eDNA reads[55]. Deep learning has revolutionized object classifications in various biological applications, from identifying species on images[38] to modelling species distributions in habitats[28]. Taxonomic groups represent discrete classes that can be related to sequence features, including the composition and distribution of nucleobases within DNA sequences[14,39]. For example, k-mer summarizes the counts of nucleotides within sub-sequences of length k and, in combination with machine classifications, have been used to label sequences from bacteria, archaea, fungi and viruses[58]. The association between k-mer features and taxonomic labelling can be trained in a neural network from a reference genetic database[58,60] to predict the label of any new sequence. Alternatively, a convolutional neural network (CNN) can self-learn a broader range of spatially organized DNA base-motif features existing in the DNA sequences[39]. The neural structure subsets signals from a restricted region of the input data known as the receptive field and responds to localized patterns in the sequence data. The numeric encoding of the four DNA bases makes it possible for the spatial placements of nucleotides to be interpreted by the CNN. In particular, Busia et al. developed a CNN[14] which trains a deep neural network to predict database-derived taxonomic labels directly from query sequences. Hence, preliminary use of machine learning with DNA sequence data shows the potential of this approach for taxonomic labelling[14,44], but so far it has mainly been used to label relatively long amplicons such as the full 16S gene, in fragments up to 250 bp long[14]. It remains to be determined how it performs in the taxonomic labelling of short sequences from eDNA metabarcoding.

The most computationally costly step in the processing of eDNA metabarcoding is data cleaning[52], and a large computational gain from machine learning could be achieved if a CNN can be applied directly on raw sequencing data that can contain many errors, including PCR substitutions or insertions or deletions of bases[66,67,73]. Existing eDNA bioinformatic pipelines apply a computationally demanding process of sequence processing and cleaning[52], conserving only high-quality reads[10] before the taxonomic labelling of DNA reads. To circumvent this data cleaning procedure, CNNs should be able to either identify low-quality sequences or accommodate noisy data in the taxonomic labelling. CNNs with data augmentation have been used to render networks more robust to noisy data, for example by adding random variation in the training data[70]. Busia et al.[14] artificially introduced variation into sequences within the reference database to build a more robust CNN, adding between 0.5 and 16% of mutations by switching DNA bases randomly[14]. While the authors found that moderate artificial noise rendered the network more robust to potential sequencing errors, setting an excessive value decreased the CNN performance. Furthermore, the CNN should be trained to tolerate the library tags and the PCR primers present in raw metabarcoding data, but these aspects have remained largely unexplored. The CNN could then be used to process and identify the sequences from raw metabarcoding files, independently of the processing step in which they are demultiplexed to each sample. If reliable, a CNN pipeline serves as a revolutionary tool to process the exponentially growing quantity of eDNA metabarcoding data used to characterize ecosystems.

Here, we used a comprehensive eDNA data set collected in tropical South America to evaluate the ability of CNNs to rapidly and accurately process eDNA metabarcoding files into taxonomic labels. We built CNNs that allow the processing of short sequences produced by eDNA metabarcoding and tested whether the accuracy and speed of CNNs are comparable to those of OBITools[10], a widely used pipeline to process eDNA data. As a case study we used one of the largest standardized eDNA data sets currently available for fishes, corresponding to a multi-year campaign effort to sample the tropical South American rivers of French Guiana[54] (Fig. 1). This eDNA data set is associated with a quasi-exhaustive reference database covering most of the known species of the region for the 'teleo' region of the 12S rRNA mitochondrial gene[22,26]. The raw data set contains nearly 700 million sequences, with about 205 million sequences belong to the samples of interest here. The freshwater ecosystems of French Guiana are among the most species-rich ecosystems for riverine fishes globally[3], and among the rivers the least impacted by humans[72]. Good performance of an approach in a complex ecosystem provides a robust proof of concept for further applications in any other ecosystem with a simpler species assemblage. Within this general processing framework and using this case study, we asked the following questions: (1) How does a CNN approach perform in the training of eDNA sequence classification for labels of the reference database? (2) How robust is the classification of a CNN applied directly to raw Illumina metabarcoding short sequences? (3) How do a classical metabarcoding pipeline and our CNN approach compare with the pre-existing information about biodiversity composition within two river catchments with a long history of traditional sampling?

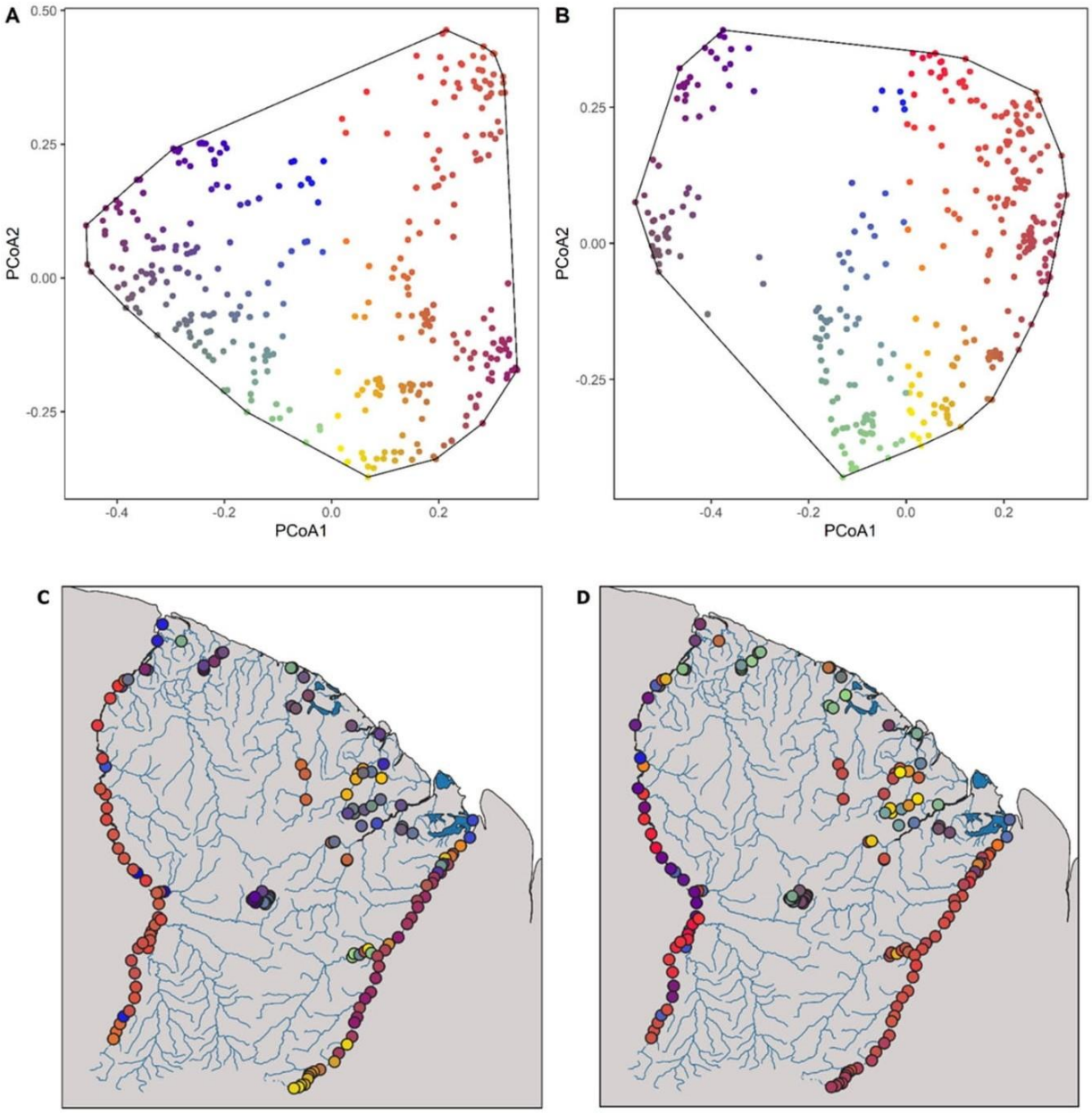


Figure 1. Principal coordinate analysis (PCoA) of species composition dissimilarity between filters. (**A**) Ordination of filter species composition dissimilarity in the outputs of OBITools. (**B**) Ordination of filter species composition dissimilarity in the outputs of the CNN applied to raw reads. Dissimilarity matrices were built with Bray–Curtis distances on read abundance per species per filter. (**C**) Maps of the filter locations, coloured according to the position of the filters in the PCoA space for OBITools outputs. (**D**) Maps of the filter locations, coloured according to the position of the filters in the PCoA space for the CNN applied to raw reads outputs. The maps were created with QGIS 3.6.1.

Results

CNN training and evaluation with split sampling. CNNs learned features of the 60 bp teleo sequence reads with good internal and external predictive power. Larger networks did not necessarily produce better results, indicating low overfitting. A CNN of moderate complexity learned the full structure contained in the training sequence data. The training and evaluation of the CNN with split sampling considered 156 species (out of 368) which had at least two unique sequences. The optimal CNN consisted of a 150×4 unit input layer, one convolutional layer of 4 filters with a 7×4 extent, 3 dense layers with 128 neurons each, and an output layer 156 neurons wide. On the training data, the networks achieved 92% accuracy, with small differences between the networks trained on the base reference data and those trained on the augmented reference data (i.e. with added tags, primers and reverse complements). When applying the CNN to the hold-out data (316 sequences

from the 156 species), we found an accuracy of 91% on the base data and 89% on the augmented data. When an optimized 0.9 binarization threshold was used with the F-beta metric, the accuracy rose to 98% for both CNNs, at the cost of 16–26% of the predictions being discarded for the base and augmented data, respectively. We then used the entire data set in the training process, using all 368 species, and repeated the analyses for the base and augmented data. The optimal CNN was similar to the previously chosen networks, with a single convolutional layer of 4 filters with a 7×4 extent, followed by 2 dense layers each 384 neurons wide. With these networks, training accuracy was similar to that from the split evaluation at 92%. Validating the networks on the reference sequences yielded higher accuracies of 96% and 94% for the base and augmented CNNs, respectively. With a binarization threshold of 0.9, the accuracy rose to 99% for both the base and augmented data sets, at the cost of rejecting 9–13% of all sequences evaluated (Supplementary Material Fig. 1). We used a binarization threshold of 0.9 for all further evaluations.

CNN application on the raw and cleaned eDNA data set. We found that there were limited differences in the output between the CNNs trained on the raw sequence data compared to those trained on cleaned data. To attenuate sequencing noise in the analysis, we considered a second threshold of the minimum number of reads required for a species to be retained. We compared the number of reads per species needed for each CNN with that needed for OBITools and observed that the median Kendall Tau-b correlation increased when a more stringent threshold on the minimum number of reads per species was applied to all levels of sample aggregation. An optimal threshold of 50 reads per species resulted a slightly better correlation for clean (median Kendall Tau-b = 0.77, range 0.22–0.94) than for raw reads (median Kendall Tau-b = 0.84, range = 0.2–1, Fig. 2) at the filter level. The same effect persisted on the PCR replicate and river levels. We considered only species with more than 50 reads within a PCR replicate in the following analyses. We repeated the analysis using the kappa similarity measurement (Fig. 3). The CNN applied to the clean reads (after assembling and demultiplexing) had a slightly higher composition similarity (median kappa value 0.96, range 0.83–1.0) than that applied to the raw reads directly from the Illumina outputs (median kappa value 0.93, range 0.79–0.99). The kappa values are based on the predicted presence and absence of species. Hence, the results were slightly better than those from the Kendall Tau b values, as those take the relative abundance of the predictions into account. All approaches recovered similar gradients of composition, differentiating between coastal and upstream assemblages (Fig. 1). The composition difference between methods resulted from a slightly larger number of species predicted by the CNN (median species number 63) than by OBITools (median species number 56). Furthermore, the CNNs still lacked feature parity with OBITools with regard to ambiguous sequences, which can result in more pronounced differences in the OBITools output.

Validation with the known species list of the region. We found a major overlap between historical records and the species composition recovered from the CNN. The data synthesis across historical fish surveys yielded a total of 351 species in the Maroni and Oyapock rivers, 293 of which were present in the reference database and thus potentially detectable with eDNA. For both rivers combined, the CNN applied to raw reads assigned 319 species, 264 of which were known from the historical records, while 55 had never been recorded before (Fig. 4a). The CNN and OBITools detected 274 species in common, while the CNN retrieved 21 species known from the historical surveys in these rivers that were not retrieved with OBITools but identified 24 species not known from the survey synthesis or identified with OBITools. The species detected only with the CNN mainly belong to the Loricariidae, Cichlidae, Characidae and Callichthyidae families. The 23 species known from historical records and not detected by either eDNA method mainly belong to the Loricariidae, Characidae, Apteronotidae and Anostomidae families. The two species detected only with OBItools are from the Cichlidae and Aspredinidae families (Fig. 4b). The CNN applied to clean reads detected 293 species, 254 of which were present in the Maroni and Oyapock synthesis, 276 of which were also found in the outputs of OBITools, 9 of which were found only with the CNN and in the synthesis, and 8 of which were found only with the CNN. In the case of OBITools, 282 species were detected, 249 of which were included in the historical synthesis and 33 of which had never been recorded in the Maroni or Oyapock rivers (Fig. 4c). The species detected only with the CNN mainly belong to the Characidae family. The species known from historical records but not detected with either eDNA method belong to the Loricariidae, Characidae and Apteronotidae families. The two species detected only with OBITools are from the Loricariidae and Cichlidae families (Fig. 4d). The same analysis at the single river scale provided similar results (Supplementary Material Figs 4, 5). Hence, while both methods detected species not found in the historical records, the CNN generally recovered more species than OBITools, which could correspond to either new true observations or commission errors. The CNN applied to raw reads retrieved more species that were not in historical records nor found with OBITools. For the Maroni river, the CNN applied to raw reads and the CNN applied to clean reads retrieved 232 species in common, while 48 were found only with the raw reads and 16 only with the clean reads. For the Oyapock river, 185, 66 and 18 species were found in common, only with the raw reads, and only with the clean reads, respectively (Supplementary Material Fig. 6).

Computation time. Overall, the CNN processed approximately 1 million input sequences per minute, compared with 20,000 input sequences per minute for OBITools. For the CNN, we distinguished between two computational efforts, which were measured independently: (1) network training, which needed to be performed once per reference database, and (2) the application on field data. Training a network on the augmented and complete reference database currently took around 10 min on an Nvidia Titan RTX GPU. Training a network on the clean reference database was faster and takes 6 min on the same GPU. The training and application time is dependent on the size of the input data and the network size. A large part of the computational time for

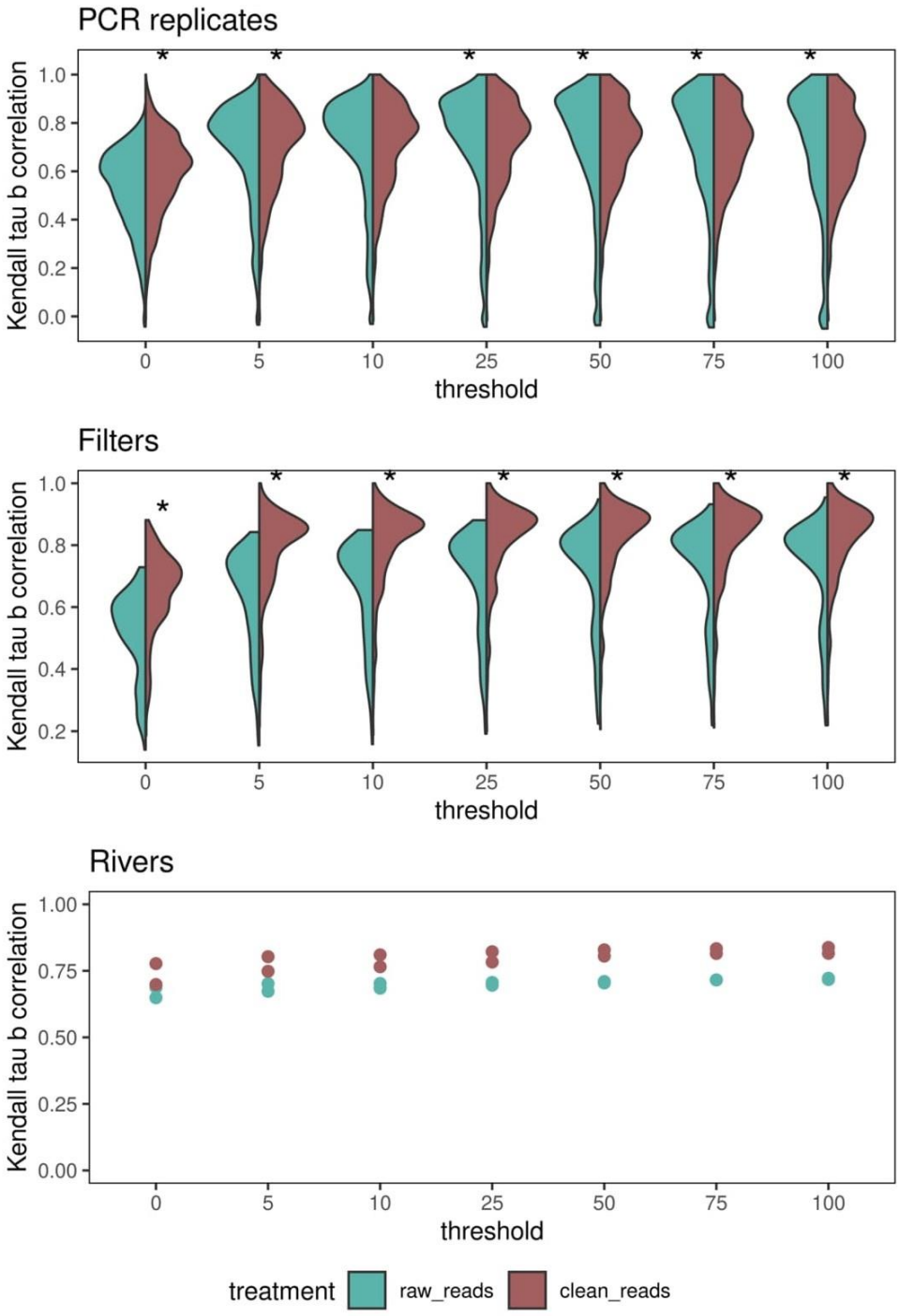


Figure 2. Kendall Tau-b correlation coefficient between the outputs of the CNN and OBITools. The left side of the violin plots (blue) displays correlation values between OBITools and the CNN applied to raw reads. The right side of the violin plots (red) displays correlation values between OBITools and the CNN applied to clean reads. The x-axis represents the threshold of the minimum read number per species for the species to be considered present. Stars represent a significant difference between the two correlations. The analysis was made at three levels: PCR replicates (top), eDNA filters (middle), and rivers (bottom).

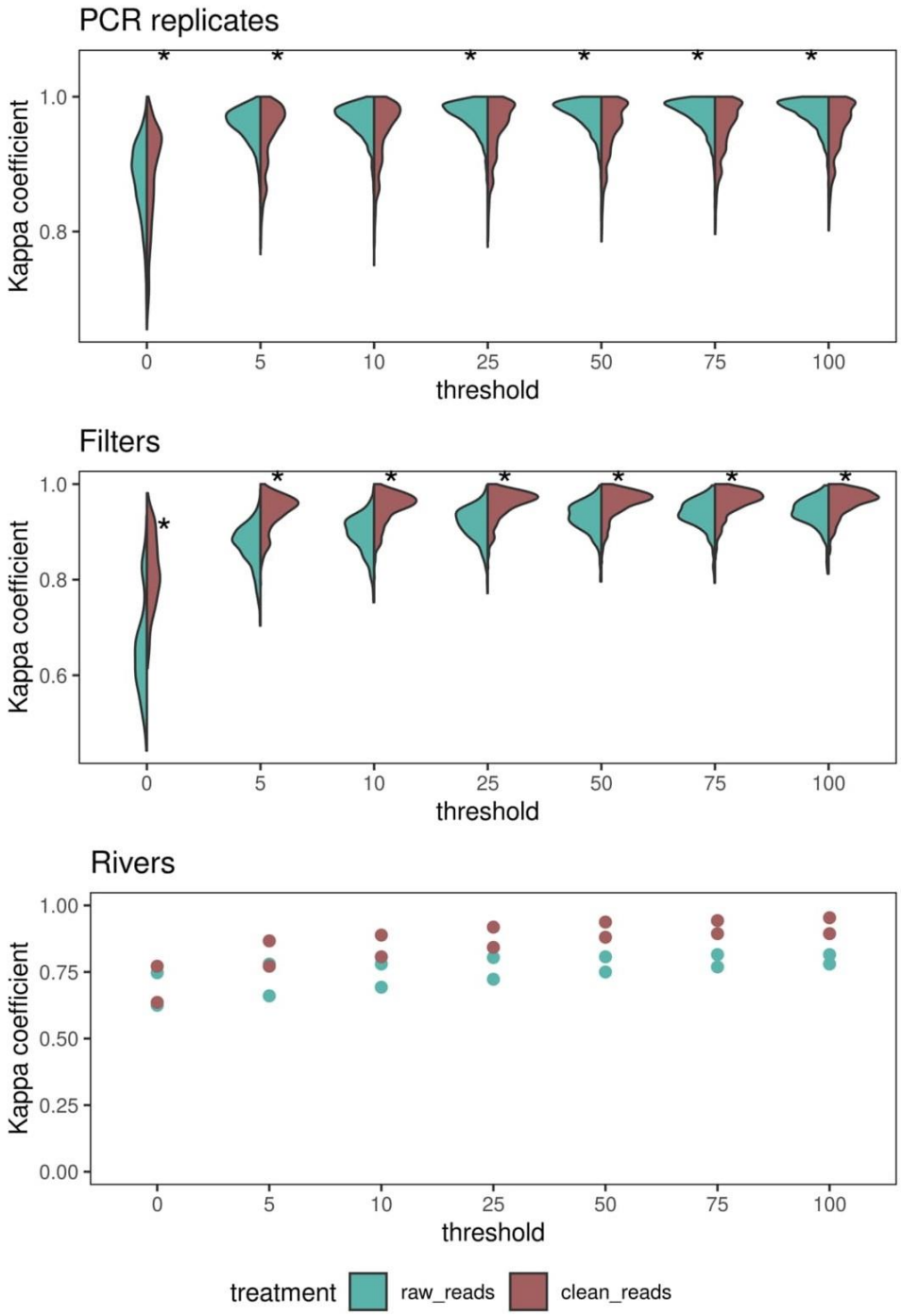


Figure 3. Kappa correlation coefficient between the outputs of the CNN and OBITools. The left side of the violin plots (blue) displays correlation values between OBITools and the CNN applied to raw reads. The right side of the violin plots (red) displays correlation values between OBITools and the CNN applied to clean reads. The x-axis represents the threshold of the minimum read number per species for the species to be considered present. Stars represent a significant difference between the two correlations.

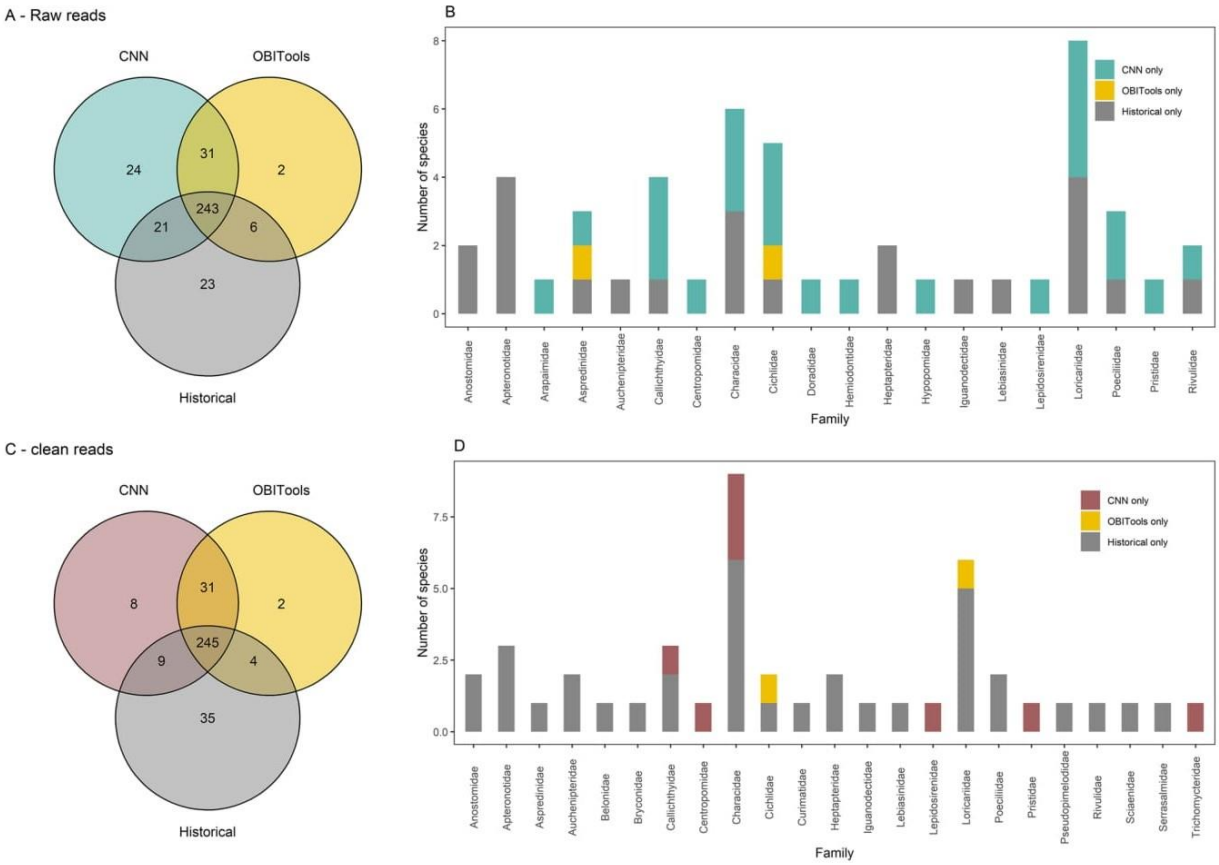


Figure 4. Species detections with the CNN approach, with OBITools, and in historical records in the combined Maroni and Oyapock rivers. (**A**) Overlap of species detections between the CNN applied to raw reads (blue), OBITools (yellow) and historical records (grey). (**B**) Number of species per family, detected with only one method (CNN applied to raw reads, OBITools or historical records). (**C**) Overlap of species detections between the CNN applied to clean reads (red), OBITools (yellow) and historical records (grey). (**D**) Number of species per family that were detected with only one method (CNN applied to clean reads, OBITools or historical records).

the OBITools pipeline is dedicated to the alignment (up to 80%) and demultiplexing (up to 15%) steps. By training and applying a convolutional neural network directly on raw reads, we could sidestep this issue completely and achieve significantly faster processing times and lower power consumption at the cost of more marked differences in the recovered compositions overall.

Discussion

The monitoring of biodiversity in highly species-rich ecosystems has generally been challenging, with gaps in biodiversity data existing in the tropics[23]. eDNA metabarcoding is a revolutionary method that can enhance the monitoring of species in complex ecosystems[48], but is associated with the challenge of rapidly processing large data sets. Our study demonstrates the application of a CNN to process short eDNA sequence reads directly from raw sequencing Illumina outputs. We show that the CNN approach delivers species compositions comparable to those from OBITools and historical records. Fish assemblages retrieved using OBITools and CNN were consistent with the current knowledge on Guianese fish fauna, with marked differences between coastal and inland sites[29]. Fish homogeneity in coastal areas was explained by a historical connectivity between the coastal basins during the Miocene[21], but also by the salt tolerance of a substantial number of the fishes inhabiting coastal streams[45]. Composition analysis further highlighted sites with a markedly different fauna, corresponding to the areas heavily disturbed by gold mining, forestry and agriculture[4]. In only a few minutes, the software transformed a raw fastq sequence data set into a species list associated with each eDNA sample collected in the field, which can serve further biodiversity analyses. Overall, our findings indicate that machine learning offers new possibilities for the taxonomic labelling of short DNA sequences and can transform rapidly collected eDNA data samples into interpretable taxonomy-based biodiversity indicators[25,30,68].

In classical bioinformatic pipelines, the processing from raw sequence reads to taxonomic identifications includes seven steps (paired-end read merging, demultiplexing, dereplication, quality filtering, removal and correction of PCR/sequencing errors, and taxonomic labelling) expected to be essential to generate high-quality results from metabarcoding studies, but which can be computationally demanding[8,16] and challenging to

articulate[50]. We show that a CNN can embed all these steps in a single process applied directly to the raw Illumina reads when the CNN is trained to handle noisy data. Moreover, for relatively short eDNA markers (e.g. 60 bp for the 'teleo' marker used here), merging paired-end reads is not necessary, which leads to a significant computational gain[52]. While still offering results roughly comparable to those of OBITools, the CNN decreases the processing time of the whole data set analysis by a factor of around 150. In a recent comparison, Barque (https://github.com/enormandeau/barque) combined with a fast demultiplexing module was able to process over 15 million reads in 30 min, while it took 17 h for OBITools V1[52]. Assuming the same rate as found for our CNN, i.e. 1 million read per minute, to this data set, the application of the CNN would be two times faster than the fastest existing bioinformatic pipeline in a single model[52]. Our study represents a first successful adaption of CNN to the processing of eDNA metabarcoding data, but we foresee several avenues of optimization to gain speed and accuracy, making it a promising tool for scaling-up biodiversity inventories via eDNA[5,64].

The training of CNNs leads to an efficient adjustment to the reference database, avoiding the need to explore a large number of parameters and arbitrary thresholds, as required in classical bioinformatic pipelines. Existing bioinformatic pipelines contain a variety of modules (i.e. QIIME2, DADA2, Vsearch), each with its own set of parameters[7,17,62]. Selecting the appropriate modules and parameters requires advanced knowledge of the functioning of the program, since changes in those parameters can considerably modify the outputs[8,13,36]. The absence of an appropriate and automated method for parameter optimization[2] often limits the use of those pipelines by non-specialists. In contrast, the application of a CNN only includes a first step of training, where the optimization of the network is nearly automated, and two independent steps for applying the CNN and demultiplexing the reads to reach to final taxonomic outputs per sample. During the learning step of a CNN, only three parameters have to be set by the user: the network size (number of layers, filters and units), the learning rate, and the augmentation values. During the application step, two parameters are optimized, the binarization threshold and the minimum number of reads per sample to be considered. We expect that these steps can be nearly automated within a user-friendly software, as developed for other machine learning applications[76]. Given the relative ease of the training process and application of CNN, the approach could be transformed into an application with a user-friendly interface demanding only a minimum amount of interaction. Hence, CNNs could make eDNA metabarcoding data processing accessible even to less trained users and provide an overview of biodiversity more rapidly.

A CNN trained on a complete reference database produced species composition outputs congruent with the outputs of a popular bioinformatic pipeline, but showed a tendency to predict more species than those of OBITools and historical records. Compositional differences in the outputs of pipelines have already been highlighted (e.g.[11]) and have mainly resulted from the detection of several false positives and false negatives[52]. With a binarization threshold of 0.9 optimized during the training phase, we found congruent but slightly divergent results between OBITools and the CNN applied to either the raw or clean reads. While the CNN and OBITools shared most of their recovered species, each method detected a few species not detected by the other approach (Fig. 4). However, the CNN showed a general tendency of overprediction compared with OBITools and the historical records, especially when it was applied directly to the raw sequencing data. Using the historical records as a baseline, the CNN applied to clean reads reduced the detection of species only found with the CNN, without decreasing the number of species shared with OBITools or historical records, suggesting false positives resulting from noisy inputs. Specifically, the CNN applied to raw reads detected more species from the Loricariidae, Cichlidae and Characidae families that were not found with OBITools, which may have been the result of sequencing errors that were not denoised by the CNN. In the case of the Cichlidae family, the short barcode we used is known to be poorly resolved[73], with many species sharing the same sequence[61], and our CNN did not perform well in this situation, like all other pipelines. Moreover, Loricaridae and Characidae are the two most speciose families of the Guianese fish fauna, with more than 50 species per family[45] and with several new species occurrences recorded each year in Guianese rivers (e.g.[12]). These two families, together with Cichlidae, are also known to host cryptic and still unnamed species, as shown by Papa et al. for the Maroni river[57]. This could also contribute to species misdetections. Finally, we found that the correlation between OBITools and CNN was lower at the sample level than at the level of the PCR replicates when the CNN was applied to raw reads. Hence, appropriately combining the PCR replicates could confer more robustness to the final outputs of the CNN. Refinement of the network could be added, so that the detection across multiple PCR replicates could be used to compute the final likelihood.

In our study we proposed a novel application of a CNN approach to eDNA metabarcoding data, but several improvements are required before broad-scale applications to large eDNA data sets can be considered. The CNN trained in this study learns from the species class and is forced to assign the sequences to that taxonomic level. Thus, when presented with conflicting sequences, the network might assign all of them to a single species, or may split the probabilities across several species, which might then be discarded given the use of the 0.9 binarization threshold. In contrast, in the case of a conflict, OBITools can assign sequences to higher taxonomic levels, thus keeping information related to these species with identical sequences. In this case study, we had an ideal situation where the reference database was almost complete for the territory. The CNN could be improved to handle incomplete reference databases and to be able to assign a read to another taxonomic level or to an unknown class, rather that forcing a species-level identification and relying on the binarization threshold to reject unknown sequences. Further, we expect that it is possible to improve the CNN by implementing more stringent filters that would reduce the number of false detection and prediction errors. For instance, a filter for tag-jump handling, included in previous pipelines for eDNA metabarcoding for fish (e.g.[22]) could be considered. Finally, while the computational speed was already faster than existing traditional pipelines, specific optimizations, such as network pruning or lower precision computations, could improve the performance further, making this approach even more attractive for applications in future broad-scale eDNA projects.

Conclusion and perspectives. We have demonstrated that we can use deep learning to increase the speed and decrease the energy consumption required for processing eDNA metabarcoding data, with a high accuracy when applied to clean reads and a slightly lower accuracy for raw reads. The largest part of the computation time for the CNN is for the training phase; once trained, the CNN can be used as a computationally efficient tool for applications in the cloud, facilitating analyses of the mass of eDNA data expected to be collected in future biodiversity surveys. eDNA data are being collected at an exponentially increasing rate. Owing to its easy application—due to the reduced number of processing steps and the automated learning of best-suited parameters, a CNN approach contrasts with other widely-used bioinformatic pipelines. Our work paves the way towards computationally efficient and user-friendly online processing pipelines that will contribute to the democratization of bioinformatic analyses of eDNA samples. Our work is a major complement to the recent development and standardization of eDNA in the laboratory; together, they will make it possible to extend the use of eDNA in community ecology and biogeography, even for poorly understood ecosystems or lineages[43], and they will help to install eDNA as a standard monitoring tool[42]. Our findings also reinforce the initial goal of quick and efficient eDNA application for biodiversity monitoring. We expect that the results from this study will be scaled up to help CNNs become a major toolkit for ecological analyses of eDNA data, possibly associated with a cloud infrastructure and parallel computation on GPUs.

Material and methods

eDNA data collection and reference database. As a test data set we used data collected in French Guiana, a *c.* 80,000 km^2 South American territory almost entirely covered by dense primary forest (Supplementary Material Fig. 2). The equatorial climate, associated with abundant rainfall, has created a dense hydrographic network consisting of six major watersheds and several coastal rivers that host a highly diverse fish fauna with at least 368 strictly freshwater fish species[45]. eDNA field collection was initiated in 2014 and continued until 2020. We sampled over 200 sites (see Murienne et al.[54] for details), where we filtered 30 litres of river water across a flow filtration capsule using a peristaltic pump. For the purposes of this study, we analysed only the filters collected in both the Maroni and Oyapock rivers.

At each site we collected one to ten filtration capsules, but at most sites two capsules were used (2×34 l), using a previously established protocol[20,26]. We used a peristaltic pump (Vampire sampler, Burlke, Germany) and disposable sterile tubing to pump the water through the encapsulated filtering cartridges (VigiDNA 0.45 μM, SPYGEN, France). We held the input part of the tube a few centimetres below the surface in rapid hydromorphologic units to facilitate homogenization of DNA in the water column. When the filters began to clog, we decreased the pump speed to avoid material damage. To minimize DNA contamination, the operators remained downstream from the filtration site, either on the boat or on emerging rocks. After filtration, we filled the capsules with a preservation buffer and stored them in the dark at room temperature for less than 1 month before DNA extraction. We applified the 12S rRNA 'teleo' gene fragment[77] using PCR and sequenced it on an Illumina platform, generating an average of 500,000 paired-end sequence reads per sample. The DNA extraction, amplification and sequencing protocol have been described previously[19].

We generated an eDNA reference database by combining fish specimens caught using various types of fishing gear. These data were complemented by fish collections carried out by environmental management agencies (DGTM Guyane, Office de l'eau Guyane, Hydreco laboratory), fish hobbyists (Guyane Wild Fish), and Museum tissue collections (MHN Geneva). Although rare for Guianese fishes, we also included existing sequence data from online databases (Genbank, Mitofish). We extracted and sequenced the 12S ribosomal gene from the collected species. The local reference database has improved over the years[20,22] and now covers over 368 species out of 380 estimated to occur in the region. This almost full coverage is exceptional considering the many gaps globally[51]. Sample collection was authorized by both the French Ministry of Environment (DEAL) and the Guyanese National Park (PAG). The samples comply with the international rules of the Nagoya protocol for access and benefit sharing (project refs ABSCH-IRCC-FR-246820-1 and ABSCH-IRCC-FR-245902-1).

OBITools bioinformatic pipeline. As a standard processing pipeline we selected OBITools[10], which is commonly used in eDNA metabarcoding studies[15,47,78]. We processed the reads from the sequencing following Valentini et al.[77]. In short, we assembled the forward and reverse reads using `illuminapairedend` with a minimum score of 40, retrieving only joined sequences. We then assigned the reads to each sample using `ngsfilter`. We then created a separate data set for each eDNA sample by splitting the original data set into several files using `obisplit`. After this step, we analysed each sample individually before generating the taxonomic list. We clustered strictly identical sequences together using `obiuniq`. Further, we excluded sequences shorter than 20 bp using `obigrep`. We then ran `obiclean` within each PCR product for clustering. We discarded all sequences labelled as 'internal', corresponding most likely to PCR substitutions and indel errors. We performed taxonomic labelling of the remaining sequences using `ecotag` with the custom genetic reference database relevant for the eDNA samples. Finally, we applied an empirical threshold to account for tag-jumps and spurious errors.

Reference data augmentation and training data set. The reference database has a full species coverage, but the number of DNA replicate sequences for each species was limited because there were only 683 sequences for 368 species. This makes training a CNN challenging for several reasons. The number of sequences per species is not balanced, there are not enough sequences to capture the entire inter- and intraspecific variation, and the noise from the sequencing process is not accounted for. To balance the data set using data augmentation procedures, we oversampled the underrepresented species before training. To increase the sequence variation, we implemented an inline sequence mutation step similar to that applied by Busia et al.[14]. During each

training epoch all sequences were randomly mutated. We added between zero and two random insertions and deletions each, as well as noise in the form of a 5% mutation rate. This procedure further reduced overfitting, as no training sample was likely to be repeated twice. For the evaluations, we either added no augmentation or 2% noise and singular insertions and deletions, as we expected the PCR amplification and sequencing to be better than the 5% noise considered during the training phase.

For the direct application on the raw reads, another data transformation step was required. All sequences processed in an Illumina machine retain the selected primers, and were tagged with 8-bp-long tags. During the sequencing two bases from the plate attachment sequence were often read as well. We therefore pre- and appended the forward and reverse primers, and the combined tags and attachment bps to the sequences from the reference database. Specifically, we added 10 bp of unknown bases to each reference sequence, represented by the IUPAC code 'N'. This shifted the sequences to a position in the training input similar to where they would occur in the Illumina data. While there is a canonical read direction for DNA, the read direction during the sequencing randomly occurs on either DNA strand. Therefore, we added the reverse complement of all sequences to the final data set. As a last step we truncated all sequences to a read length of 150 bps, as fixed by the field metabarcoding data.

Convolutional networks. CNNs play a key role in modern computer vision applications and date back to the emergence of artificial neural networks in the 1950s and 1960s. Some of the first applications of CNNs and their training method include digit recognition for handwritten ZIP codes[46]. Each convolutional layer in a CNN consists of a number of convolutional kernels often called filters. These filters can be thought of as feature detectors each responding to a specific feature in the input data. Compared with fully connected dense layers, the small extent of these filters drastically reduces the number of free parameters to train. Intuitive examples in image processing are edge or corner detectors. By arranging the DNA sequences as two-dimensional inputs, the convolutional layer can learn and exploit abstract features in the sequences.

CNN training and evaluation using split sampling. We investigated the performance of a CNN approach trained on the reference database at the species level. To encode DNA sequence information, each canonical base (A, C, T, G) and each IUPAC ambiguity code was translated to an appropriate four-dimensional probability distribution over the four canonical bases (A, T, C, G), including uncertain base reads (e.g. W and S). For example 'A' became [1,0,0,0] and 'W' became [0.5, 0, 0, 0.5]. The neural network was designed and optimized through a series of tests that allow the optimal set of correct DNA features to be selected. In particular, we explored an exhaustive number of model sizes, including one to three layers of 2D (depth-wise separable) convolutions with 4–16 filters each, one to three fully-connected layers with varying numbers of neurons each, and a softmax activated output layer which produces a probability distribution over all possible taxonomic labels. We applied dropout regularization and used leaky rectified-linear activation for all but the last layers.

We used TensorFlow[1] to train the CNNs with all the aforementioned data augmentations. Due to the sparse data set, we characterized and evaluated the performance of the neural networks using three different methods. First, we applied random split-sampling from the reference database. This established a proper separation between the training and validation data, but less than half the species in the reference data set had two or more sequences, resulting in a reduced range of species that could be included. Specifically, only 156 out of 368 species possessed two or more unique sequences and were considered for the split data set. Next, we trained several networks on the full reference data set with all 368 species and validated them using the original non-augmented reference data. We derived more synthetic data from the reference sequences similar to the training augmentations and evaluated them with the chosen network. We evaluated whether there were systematic errors in the CNN performance. We further investigated whether a binarization threshold, requiring the probability of the most likely prediction to be above a certain value, improved the classification performance. As we prioritized the absence of errors, i.e. fewer false positives, over the presence of correct predictions, we evaluated the effects of such a binarization threshold using the F-beta measure, which uses a weighted trade-off between these errors. We chose a small beta value of 0.3 to heavily discourage false positives at the cost of discarding some correct results.

CNN application on demultiplexed and cleaned samples. We tested the best trained CNN on the curated eDNA reads after the application of the main cleaning steps of the OBITools pipeline. In particular, from the Illumina raw output, we assembled the forward and reverse reads using the `illuminapairedend` algorithm from the OBITools package, after which we kept only high-quality reads and demultiplexed them across the different eDNA samples. We applied the best-trained CNN at the species level to these curated eDNA samples. We compared the taxonomic labelling performed by the CNN to classic labelling using `ecotag` from OBITools. We evaluated and applied different thresholds for accepting species detection as a way to remove spurious errors and wrong assignments (0, 5, 10, 25, 50, 75 and 100 reads in at least one PCR replicate). For each eDNA PCR replicate and filter, and for the whole rivers, we ranked the taxonomic groups by the number of reads recovered by each method and performed a Kendall rank correlation. We ran one rank correlation per eDNA sample and reported the median rank correlation across all samples. In addition, we compared the presence–absence using the kappa statistic, which measured the general agreement between the methods for each sample. We calculated the median percentages and median kappa values across the samples. Then, across all eDNA samples, we correlated the species richness obtained via CNN with that obtained with OBITools. Each analysis was performed at three different scales: the PCR replicate, the filtration capsule and the river. Finally, we ordinated the species composition of each filtration capsule for both methods using a principal coordinate analysis (PCoA), to compare differences in recovered compositions among the methods.

CNN application on the raw illumina sequences. We applied the best CNN directly to the raw outputs from the Illumina sequencing, where we omitted all the preprocessing steps from OBITools. The CNN was expected to learn how to ignore the primer, as it was constant for all presented sequences. Furthermore, the output sequences from the Illumina sequencer were fixed in length (150 bp), so we fixed the input width of the CNN to this size. We systematically zero-padded or truncated the input sequences to this length during training, evaluation, and application. After the training phase with the reference database and the application on fastq, we developed a custom code for the fast demultiplexing of the reads. By focusing on the tag information in the first few positions of the sequence and not considering read errors in tags, we reduced the demultiplexing to a few simple look-ups in a hash table (currently 5), therefore reducing computation time with limited information loss. As in the previous test, we obtained a list of taxonomic labels for each eDNA sample, which could be compared with species composition information obtained with the OBITools pipeline. We further applied a threshold approach, obtaining a predicted composition per sample for any threshold tested. As done previously, for each eDNA sample, we ranked the taxonomic groups by the number of reads recovered with each method and performed a rank correlation. We calculated the median rank correlations across all the eDNA samples. In addition, we compared the presence–absence at the species level using the overall kappa statistic. We further evaluated whether differences between methods were more frequent in specific taxonomic families than others. Then, across all eDNA samples, we correlated the species richness obtained via CNN with that obtained with OBITools, and ordinated species composition of each filter for both methods on a PCoA. We evaluated the change in accuracy between the CNN applied to curated reads compared with the CNN applied to raw fastq files.

Validation with existing biodiversity knowledge on the region. We compared the species composition recovered in the eDNA samples by CNN and OBITools to the species, genus and family checklists of each river catchment. Species lists for each catchment were obtained from an updated version of the catchment-scale species lists[45] provided in Le Bail et al. From this list, we updated the taxonomy and added novel occurrences of known species based on fish catches by several research and management organizations (see 'Material and methods' section). Only collected specimens with a validated taxonomy were considered when updating this list, and detections using eDNA were not considered. We specifically quantified the number of matching species, false presences and false absences from each method, taking the checklists as references.

Data availability

Partial data is available through Cilleros et al.[22]. The full data set is available from the corresponding author upon request.

Code availability

The code is available in the supplementary material and released under the AGPLv3 license.

Received: 13 October 2021; Accepted: 24 May 2022
Published online: 17 June 2022

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G., Davis, A., Dean, J., Devin, M., Ghemawat, S. & Zheng, X. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. (2015).
2. Alberdi, A., Aizpurua, O., Gilbert, M. T. P. & Bohmann, K. Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods Ecol. Evol.* **9**, 134–147 (2018).
3. Albert, J. S. & Reis, R. E. One. Introduction to Neotropical freshwaters. In Historical biogeography of Neotropical freshwater fishes (pp. 3-20). University of California Press. (2011).
4. Allard, L., Popée, M., Vigouroux, R. & Brosse, S. Effect of reduced impact logging and small-scale mining disturbances on Neotropical stream fish assemblages. *Aquat. Sci.* **78**, 315–325 (2016).
5. Berry, O. *et al.* Making environmental DNA (eDNA) biodiversity records globally accessible. *Environ. DNA* **3**(4), 699–705 (2020).
6. Bohmann, K. *et al.* Environmental DNA for wildlife biology and biodiversity monitoring. *Trends Ecol. Evol.* **29**(6), 358–367 (2014).
7. Bolyen, E. *et al.* QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. *Nat. Biotechnol.* **32**, 852–857 (2019).
8. Bonder, M. J., Abeln, S., Zaura, E. & Brandt, B. W. Comparing clustering and pre-processing in taxonomy analysis. *Bioinformatics* **28**(22), 2891–2897 (2012).
9. Boussarie, G. *et al.* Environmental DNA illuminates the dark diversity of sharks. *Sci. Adv.* **4**, eaap9661 (2018).
10. Boyer, F. *et al.* obitools: A unix-inspired software package for DNA metabarcoding. *Mol. Ecology Resour.* **16**(1), 176–182 (2016).
11. Brandt, M.I., Trouche, B., Quintric, L., Günther, B., Wincker, P., Poulain, J. & Arnaud-Haond, S. Bioinformatic pipelines combining denoising and clustering tools allow for more comprehensive prokaryotic and eukaryotic metabarcoding. Molecular Ecology Resources. Accepted (2021).
12. Brosse, S., Melki, F. & Vigouroux, R. Fishes from the Mitaraka mountains (French Guiana). *Zoosystema* **41**, 131–151 (2019).
13. Brown, E. A., Chain, F. J., Crease, T. J., MacIsaac, H. J. & Cristescu, M. E. Divergence thresholds and divergent biodiversity estimates: can metabarcoding reliably describe zooplankton communities?. *Ecol. Evol.* **5**(11), 2234–2251 (2015).
14. Busia, K., George, D. E., Fannjiang, C., Alexander, D.H., Dorfman, E., Poplin, R., Chang, P., & DePris, M. A deep learning approach to pattern recognition for short DNA sequences. BioRxiv (2020).
15. Bylemans, J., Gleeson, D. M., Hardy, C. M. & Furlan, E. 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**(17), 8697–8712 (2018).
16. Calderón-Sanou, I., Münkemüller, T., Boyer, F., Zinger, L. & Thuiller, W. From environmental DNA sequences to ecological conclusions: How strong is the influence of methodological choices?. *J. Biogeogr.* **47**(1), 193–206 (2020).
17. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**(7), 581–583 (2016).
18. Cantera, I., Coutant, O., Jézéuel, C., Decotte, J.B., Dejean, T., Vigouroux, R., Valentini, A. Murienne, J. & Brosse S. Slight deforestation causes harsh biodiversity decline in Amazonian rivers (submitted)
19. Cantera, I., Decotte, J. B., Dejean, T., Murienne, J., Vigouroux, R., Valentini, A., & Brosse, S. Characterizing the spatial signal of environmental DNA in river systems using a community ecology approach. BioRxiv (2020).

20. Cantera, I. *et al.* Optimizing environmental DNA sampling effort for fish inventories in tropical streams and rivers. *Sci. Rep.* **9**(1), 1–1 (2019).
21. Cardoso, Y. P. & Montoya-Burgos, J. I. Unexpected diversity in the catfish Pseudancistrus brevispinis reveals dispersal routes in a Neotropical center of endemism: The Guyanas Region. *Mol. Ecol.* **18**, 947–964 (2009).
22. Cilleros, K. *et al.* Unlocking biodiversity and conservation studies in high-diversity environments using environmental DNA (eDNA): A test with Guianese freshwater fishes. *Mol. Ecol. Resour.* **19**(1), 27–46 (2019).
23. Collen, B., Ram, M., Zamin, T. & McRae, L. The tropical biodiversity data gap: Addressing disparity in global monitoring. *Trop. Conserv. Sci.* **1**(2), 75–88 (2008).
24. Cordier, T., Lanzén, A., Apothéloz-Perret-Gentil, L., Stoeck, T. & Pawlowski, J. Embracing environmental genomics and machine learning for routine biomonitoring. *Trends Microbiol.* **27**(5), 387–397 (2019).
25. Cordier, T. *et al.* Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. *Mol. Ecol.* **30**(13), 2937–2958 (2020).
26. Coutant, O. *et al.* Detecting fish assemblages with environmental DNA: Does protocol matter? Testing eDNA metabarcoding method robustness. *Environ. DNA* **3**(3), 619–630 (2020).
27. Deiner, K. *et al.* Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Mol. Ecol.* **26**(21), 5872–5895 (2017).
28. Deneu, B., Servajean, M., Bonnet, P., Botella, C., Munoz, F., & Joly, A. Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. PLoS Comput. Biol. (in press) (2021).
29. de Mérona, B., Tejerina-Garro, F. L. & Vigouroux, R. Fish-habitat relationships in French Guiana rivers: A review. *Cybium* **36**, 7–15 (2012).
30. DiBattista, J. D. *et al.* Environmental DNA can act as a biodiversity barometer of anthropogenic pressures in coastal ecosystems. *Sci. Rep.* **10**(1), 1–15 (2020).
31. Dornelas, M., Madin, E. M., Bunce, M., DiBattista, J. D., Johnson, M., Madin, J. S., Magurran, A. E., McGill, B. J., Pettorelli, N., Pizarro, O. & Williams, S. B. Towards a macroscope: Leveraging technology to transform the breadth, scale and resolution of macroecological data. Glob. Ecol. Biogeogr. (2019).
32. Dufresne, Y., Lejzerowicz, F., Perret-Gentil, L. A., Pawlowski, J. & Cordier, T. SLIM: A flexible web application for the reproducible processing of environmental DNA metabarcoding data. *BMC Bioinform.* **20**(1), 1–6 (2019).
33. Ficetola, G. F., Miaud, C., Pompanon, F. & Taberlet, P. Species detection using environmental DNA from water samples. *Biol. Lett.* **4**(4), 423–425 (2008).
34. Ficetola, G. F., Taberlet, P. & Coissac, E. How to limit false positives in environmental DNA and metabarcoding?. *Mol. Ecol. Resour.* **16**(3), 604–607 (2016).
35. Ficetola, G. F. *et al.* Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Mol. Ecology Resour.* **15**(3), 543–556 (2015).
36. Flynn, J. M., Brown, E. A., Chain, F. J., MacIsaac, H. J. & Cristescu, M. E. Toward accurate molecular identification of species in complex environmental samples: Testing the performance of sequence filtering and clustering methods. *Ecol. Evol.* **5**(11), 2252–2266 (2015).
37. Gold, Z. *et al.* eDNA metabarcoding bioassessment of endangered fairy shrimp (Branchinecta spp.). *Conserv. Genet. Resour.* **12**, 685–690 (2020).
38. Grünig, M., Razavi, E., Calanca, P., Mazzi, D., Wegner, J. D., & Pellissier, L. Applying deep neural networks to predict incidence and phenology of plant pests and diseases. Ecosphere (accepted) (2021).
39. Helaly, M. A., Rady, S., & Aref, M. M. Convolutional neural networks for biological sequence taxonomic classification: A comparative study. In *International Conference on Advanced Intelligent Systems and Informatics* (pp. 523–533). Springer, Cham (2019).
40. Holman, L. E. *et al.* Animals, protists and bacteria share marine biogeographic patterns. *Nat. Ecol. Evol.* **5**(6), 738–746 (2021).
41. Iknayan, K. J., Tingley, M. W., Furnas, B. J. & Beissinger, S. R. Detecting diversity: Emerging methods to estimate species diversity. *Trends Ecol. Evol.* **29**(2), 97–106 (2014).
42. Jarman, S. N., Berry, O. & Bunce, M. The value of environmental DNA biobanking for long-term biomonitoring. *Nat. Ecol. Evol.* **2**(8), 1192–1193 (2018).
43. Juhel, J. B., Utama, R. S., Marques, V., Vimono, I. B., Sugeha, H. Y., Kadarusman, Pouyaud, L., Dejean, T., Mouillot, D. & Hocdé, R. Accumulation curves of environmental DNA sequences predict coastal fish diversity in the coral triangle. *Proc. R. Soc. B* **287**(1930), 20200248 (2020).
44. Kopp, W., Monti, R., Tamburrini, A., Ohler, U. & Akalin, A. Deep learning for genomics using Janggu. *Nat. Commun.* **11**(1), 1–7 (2020).
45. Le Bail, P. Y. *et al.* Updated checklist of the freshwater and estuarine fishes of French Guiana. *Cybium* **36**(1), 293–319 (2012).
46. LeCun, Y. *et al.* Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1**(4), 541–551 (1989).
47. Li, W. *et al.* Validating eDNA measurements of the richness and abundance of anurans at a large scale. *J. Anim. Ecol.* **90**(6), 1466–1479 (2021).
48. Lopes, C. M. *et al.* eDNA metabarcoding: A promising method for anuran surveys in highly diverse tropical forests. *Mol. Ecol. Resour.* **17**(5), 904–914 (2017).
49. Makiola, A. *et al.* Key questions for next-generation biomonitoring. *Front. Environ. Sci.* **7**, 197 (2020).
50. Marques, V. *et al.* Blind assessment of vertebrate taxonomic diversity across spatial scales by clustering environmental DNA metabarcoding sequences. *Ecography* **43**(12), 1779–1790 (2020).
51. Marques, V. *et al.* GAPeDNA: Assessing and mapping global species gaps in genetic databases for eDNA metabarcoding. *Divers. Distrib.* **27**(10), 1880–1892 (2020).
52. Mathon, L. *et al.* Benchmarking bioinformatic tools for fast and accurate eDNA metabarcoding species identification. *Mol. Ecol. Resour.* **21**(7), 2565–2579 (2021).
53. McGee, K. M., Robinson, C. & Hajibabaei, M. Gaps in DNA-based biomonitoring across the globe. *Front. Ecol. Evol.* **7**, 337 (2019).
54. Murienne, J. *et al.* Aquatic eDNA for monitoring French Guiana biodiversity. *Biodivers. Data J.* **7**, e37518 (2019).
55. Nugent, C. M. & Adamowicz, S. J. Alignment-free classification of COI DNA barcode data with the Python package Alfie. *Metabarcoding Metagenomics* **4**, e55815 (2020).
56. Pagni, M. *et al.* Density-based hierarchical clustering of pyro-sequences on a large scale-the case of fungal ITS1. *Bioinformatics* **29**(10), 1268–1274 (2013).
57. Papa, Y., Le Bail, P. Y. & Covain, R. Genetic landscape clustering of a large DNA barcoding dataset reveals shared patterns of genetic divergence among freshwater fishes of the Maroni Basin. Authorea Preprints (2020).
58. Piro, V. C., Dadi, T. H., Seiler, E., Reinert, K. & Renard, B. Y. ganon: Precise metagenomics classification against large and up-to-date sets of reference sequences. *Bioinformatics* **36**(Supplement 1), i12–i20 (2020).
59. Polanco Fernández, A., Marques, V., Fopp, F., Juhel, J. B., Borrero-Pérez, G. H., Cheutin, M. C., Eme, D. & Pellissier, L. Comparing environmental DNA metabarcoding and underwater visual census to monitor tropical reef fishes. *Environ. DNA* **3**, 142–156 (2021).
60. Polanco, A. *et al.* Comparing the performance of 12S mitochondrial primers for fish environmental DNA across ecosystems. *Environ. DNA* **3**(6), 1113–1127 (2021).

61. Polanco Fernández, A., Martinezguerra, M. M., Marques, V., Francisco Villa-Navarro, Borrero-Pérez, G. H., Cheutin, M. C., Dejean, T., Hocdé, R., Juhel, J. B., Maire, E., Manel, S. & Pellissier, L. Recovering aquatic and terrestrial biodiversity in a tropical estuary using environmental DNA. *Biotropica* **53**(6), 1606–1619 (2021).
62. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: A versatile open source tool for metagenomics. *PeerJ* **4**, 1–22 (2016).
63. Rojahn, J., Gleeson, D. M., Furlan, E., Haeusler, T. & Bylemans, J. Improving the detection of rare native fish species in environmental DNA metabarcoding surveys. *Aquat. Conserv. Mar. Freshw. Ecosyst.* **31**(4), 990–997 (2021).
64. Ruppert, K. M., Kline, R. J. & Rahman, M. S. Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Glob. Ecol. Conserv.* **17**, e00547 (2019).
65. Sato, Y., Miya, M., Fukunaga, T., Sado, T. & Iwasaki, W. MitoFish and MiFish pipeline: A mitochondrial genome database of fish with an analysis pipeline for environmental DNA metabarcoding. *Mol. Biol. Evol.* **35**(6), 1553–1555 (2018).
66. Schirmer, M. *et al.* Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* **43**(6), e37 (2015).
67. Schnell, I. B., Bohmann, K. & Gilbert, M. T. P. Tag jumps illuminated–reducing sequence-to-sample misidentifications in metabarcoding studies. *Mol. Ecol. Resour.* **15**(6), 1289–1303 (2015).
68. Sepulveda, A. J., Nelson, N. M., Jerde, C. L. & Luikart, G. Are environmental DNA methods ready for aquatic invasive species management?. *Trends Ecol. Evol.* **35**, 668–678 (2020).
69. Shokralla, S., Spall, J. L., Gibson, J. F. & Hajibabaei, M. Next-generation sequencing technologies for environmental DNA research. *Mol. Ecol.* **21**(8), 1794–1805 (2012).
70. Shorten, C. & Khoshgoftaar, T. A survey on image data augmentation for deep learning. *J. Big Data* **6**, 60 (2019).
71. Singer, G. A. C., Fahner, N. A., Barnes, J. G., McCarthy, A. & Hajibabaei, M. Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology: A case study of eDNA metabarcoding seawater. *Sci. Rep.* **9**(1), 1–12 (2019).
72. Su, G. *et al.* Human impacts on global freshwater fish biodiversity. *Science* **371**(6531), 835 (2021).
73. Taberlet, P., Bonin, A., Coissac, E. & Zinger, L. *Environmental DNA: For Biodiversity Research and Monitoring* (Oxford University Press, Oxford, 2018).
74. Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E. Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol. Ecol.* **21**(8), 2045–2050 (2012).
75. Thomsen, P. F. & Willerslev, E. Environmental DNA-An emerging tool in conservation for monitoring past and present biodiversity. *Biol. Conserv.* **183**, 4–18 (2015).
76. Thuiller, W., Lafourcade, B., Engler, R. & Araújo, M. B. BIOMOD–A platform for ensemble forecasting of species distributions. *Ecography* **32**(3), 369–373 (2009).
77. Valentini, A. *et al.* Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Mol. Ecol.* **25**(4), 929–942 (2016).
78. West, K. *et al.* Large-scale eDNA metabarcoding survey reveals marine biogeographic break and transitions over tropical northwestern Australia. *Divers. Distrib.* **27**(10), 1942–1957 (2021).

Author contributions

B.F. and L.P. conceived the idea, study design, and analytic methods. B.F. developed the neural networks and ran the computational study. J.M. and S.B. collected the eDNA samples. A.V. and T.D. completed the eDNA laboratory and data preparation work. B.F. and L.M. analysed the results. L.P., L.M. and B.F. led the writing of the manuscript, with the support of S.M., A.V., T.D., C.A., D.M., W.T., J.M. and S.B.

Competing Interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-13412-w.

Correspondence and requests for materials should be addressed to B.F. or L.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

2. Suppléments du manuscrit A : *Benchmarking bioinformatic tools for fast and accurate eDNA metabarcoding species identification*

2.1. Méthodes supplémentaires

Methods S1 : Literature search

The programs compared in this study were identified through a literature search, from March to June 2019. We searched Google Scholar with the words "environmental DNA", "metabarcoding", "community" first. We extracted the aim of the studies, the taxa of interest, the methodology of amplification and sequencing (barcode size and location, PCR protocol, sequencer) and the bioinformatics processes used in each study. For each sequence treatment step, we noted the program used, and its function and parameters. This first literature search allowed us to identify the programs that were most frequently used by eDNA studies. Then we searched the literature with the word "bioinformatics tools", "bioinformatics programs", "bioinformatics pipeline" and the name of the steps. We thus identified other bioinformatics programs that could be useful for our data. Here also, we noted the name of each programs, the steps it performs, its function and parameters, and if available, the results of performance comparison with other programs. Sixty papers were analyzed in total, and the detail of the search is available in Supplementary Information Table S2.

Methods S2: Computation of programs execution times

Every programs compared were run on the same queue of the same cluster. We made sure that all the nodes of this queue had the same characteristics. We used only 1CPU to run all the programs, to obtain comparable times. Execution times of each program were obtained with the command `time`, reporting the "user time". For the steps of dereplication, quality filtering and error removal, the dataset was split in files for each sample, resulting in 348 files containing ~45,000 sequences each. These small files were analyzed in less than a second by many programs. Access to the disk and writing of the output files ("system" time) seem to take more time than execution of the programs ("user" time), but it is difficult to be certain, with such short reported times. To be sure that "user" measures the execution time, we tested these programs on bigger files (concatenation of 10 sample files), and verified that we observed the same pattern between "user" and "system" times. Since the "system" time was also longer on the big file, we were confident to interpret the times on the small files. We thus added the "user" times of the 348 samples for each programs tested in the steps of dereplication, quality filtering and error removal.

2.2. Tableaux supplémentaires

Table S3. *All bioinformatics programs identified with the literature search*

Program	Step	Reference	Source code
OBITools	Pipeline	Boyer et al. 2016	https://git.metabarcoding.org/obitools/obitools/wikis/home
Barque	Pipeline	--	https://github.com/enormandeau/barque
QIIME2	Pipeline	Bolyen et al. 2018	https://docs.qiime2.org
VSEARCH	M, Dr, QF, E, T†	Rognes et al. 2016	https://github.com/torognes/VSEARCH
Pear	M	Zhang et al. 2014	http://www.exelixis-lab.org/web/software/pear
FLASh	M	Magoč & Salzberg, 2011	https://sourceforge.net/p/flashpage
CASPER	M	Kwon et al. 2014	http://best.snu.ac.kr/casper/
Fastq-join	M	Aronesty, 2013	https://github.com/brwnj/fastq-join
Fastp	M, QF	Chen et al. 2018	https://github.com/OpenGene/fastp
Cutadapt	Dm, QF	Martin., 1994	https://github.com/marcelm/cutadapt https://cutadapt.readthedocs.io
Prinseq	QF	Schmieder & Edwards, 2011	http://prinseq.sourceforge.net
Flexbar	QF	Dodt et al. 2012	https://github.com/seqan/flexbar/wiki
Swarm	E	Mahé et al. 2015	https://github.com/torognes/swarm
Sintax	T	Edgar, 2016	https://www.drive5.com/usearch/manual/cmd_sintax.html
Programs excluded from the comparison, because not compatible with our dataset			
Trimmomatic	Trim	Bolger et al. 2014	http://www.usadellab.org/cms/index.php?page=trimmomatic
TagCleaner	Dm	Schmieder et al. 2011	http://tagcleaner.sourceforge.net/manual.html
DeML	Dm	Renaud et al. 2015	https://github.com/grenaud/deml
Tally	Dr	Davis et al. 2013	http://wwwdev.ebi.ac.uk/enright-dev/kraken/reaper/src/reaper-latest/doc/tally.html
CATCh	C	Mysara et al. 2015	https://science.sckcen.be/en/Institutes/EHS/MCB/MIC/Bioinformatics/CATCh
Perseus	C	Quince et al. 2011	https://code.google.com/archive/p/ampliconnoise/
Kaiju	T	Menzel et al. 2016	http://kaiju.binf.ku.dk/
Mothur	Pipeline	Schloss et al. 2009	https://github.com/mothur/mothur/releases/tag/v.1.43.0
QIIME	Pipeline	Caporaso et al. 2010	http://qiime.org/
SLIM	Pipeline	Dufresne et al. 2019	https://github.com/yoann-dufresne/SLIM/
USEARCH	Pipeline	Edgar 2010	https://drive5.com/usearch/download.html

† M: merging, Dm: demultiplexing, Dr: dereplication, QF: quality filtering, E: PCR/sequencing error removal, T: taxonomic assignment

2.3. Figures supplémentaires

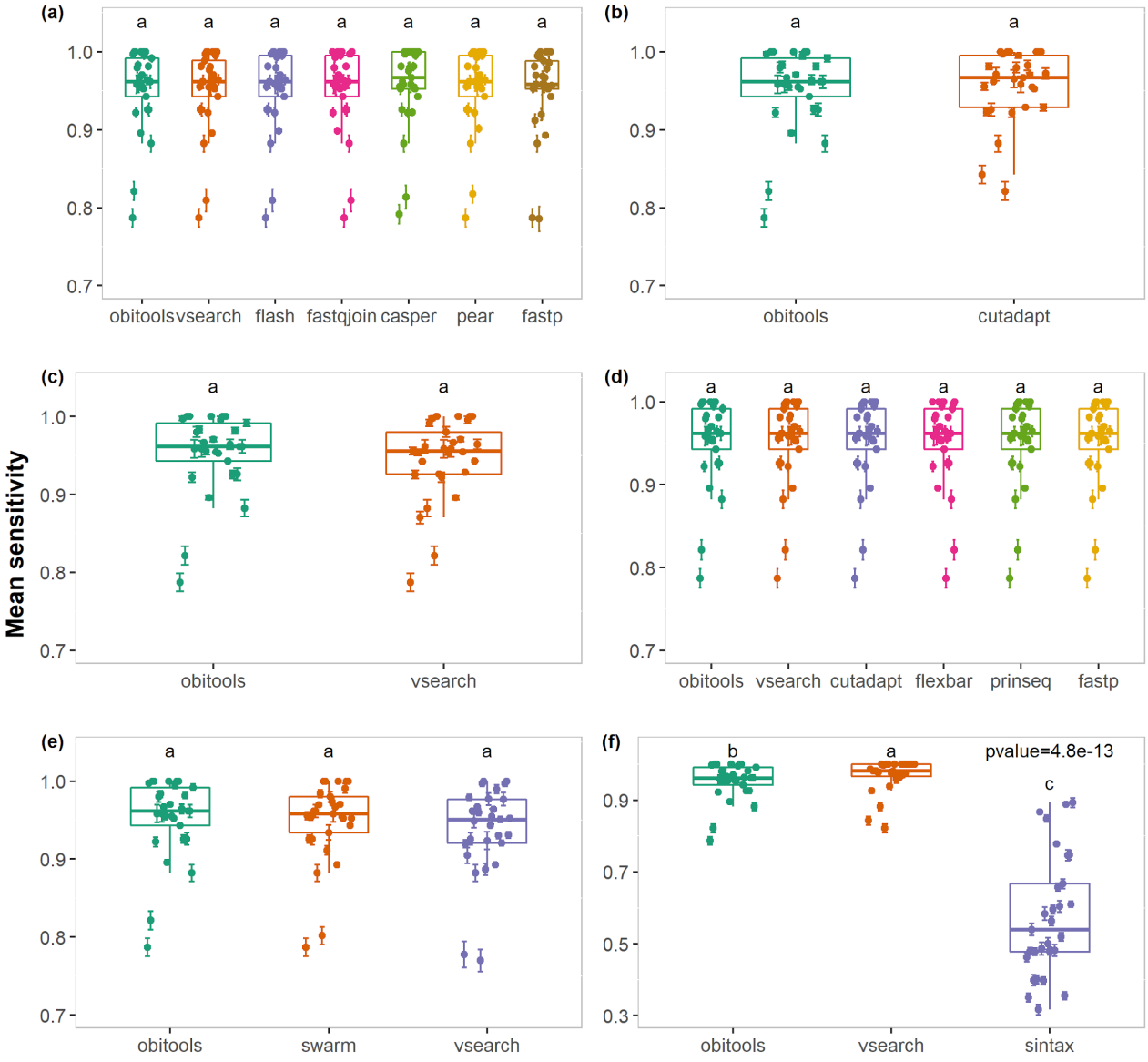


Figure S1. Mean sensitivity obtained with each program tested for each step on the simulated dataset. *The dots represent the mean sensitivity for the 12 replicas of each sample, with the standard error; the boxplot represents the median and the first and third quartiles of sensitivity for the 29 samples.* ***(a)*** *programs compared for the assembly step,* ***(b)*** *programs compared for the demultiplexing step,* ***(c)*** *programs compared for the dereplication step,* ***(d)*** *programs compared for the filtering step,* ***(e)*** *programs compared for the error removal step,* ***(f)*** *programs compared for the assignment step.*

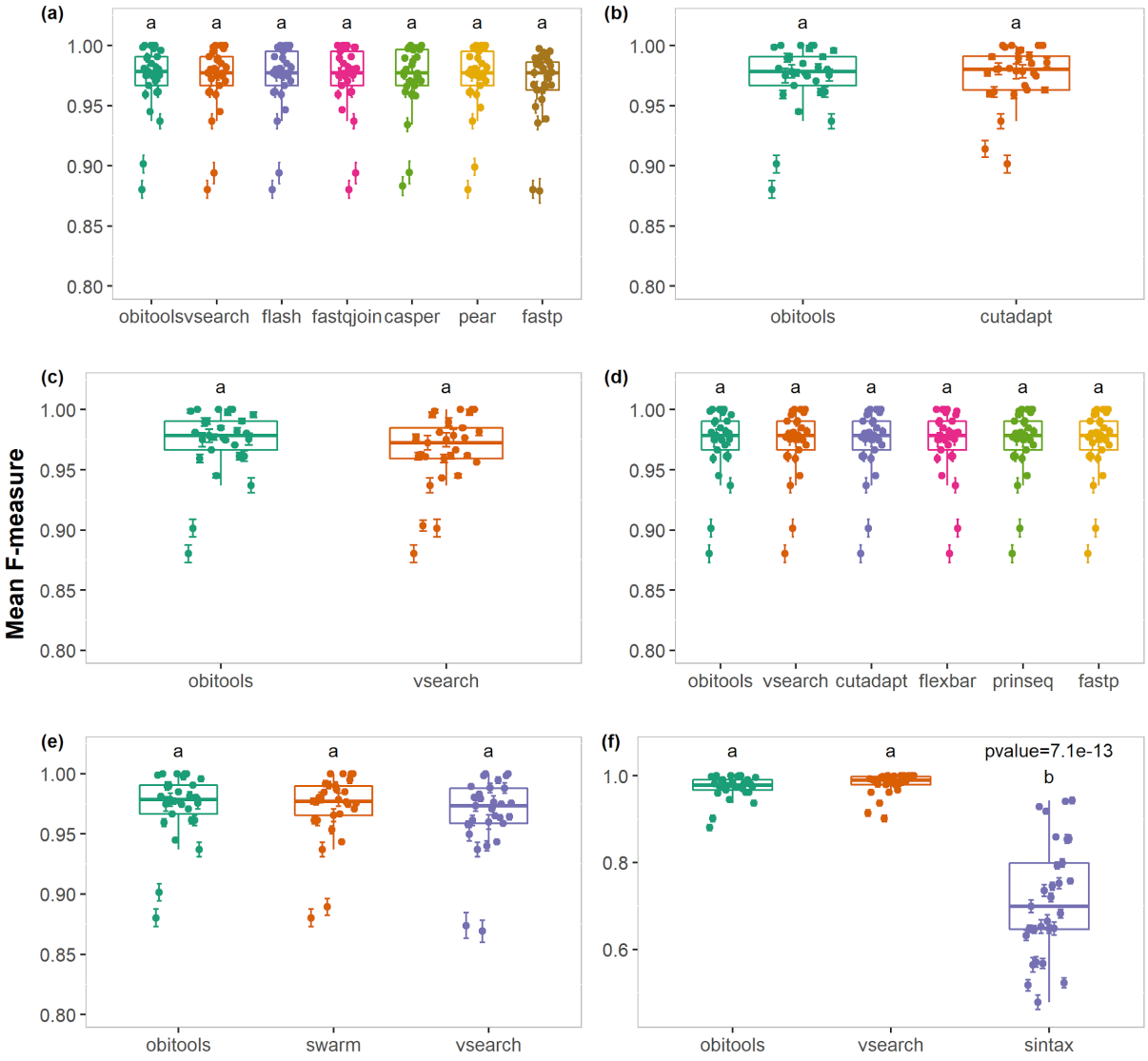


Figure S2. Mean F-measure obtained with each program tested for each step on the simulated dataset. *The dots represent the mean F-measure for the 12 replicas of each sample, with the standard error; the boxplot represents the median and the first and third quartiles of F-measure for the 29 samples.* ***(a)*** *programs compared for the assembly step,* ***(b)*** *programs compared for the demultiplexing step,* ***(c)*** *programs compared for the dereplication step,* ***(d)*** *programs compared for the filtering step,* ***(e)*** *programs compared for the error removal step,* ***(f)*** *programs compared for the assignation step.*

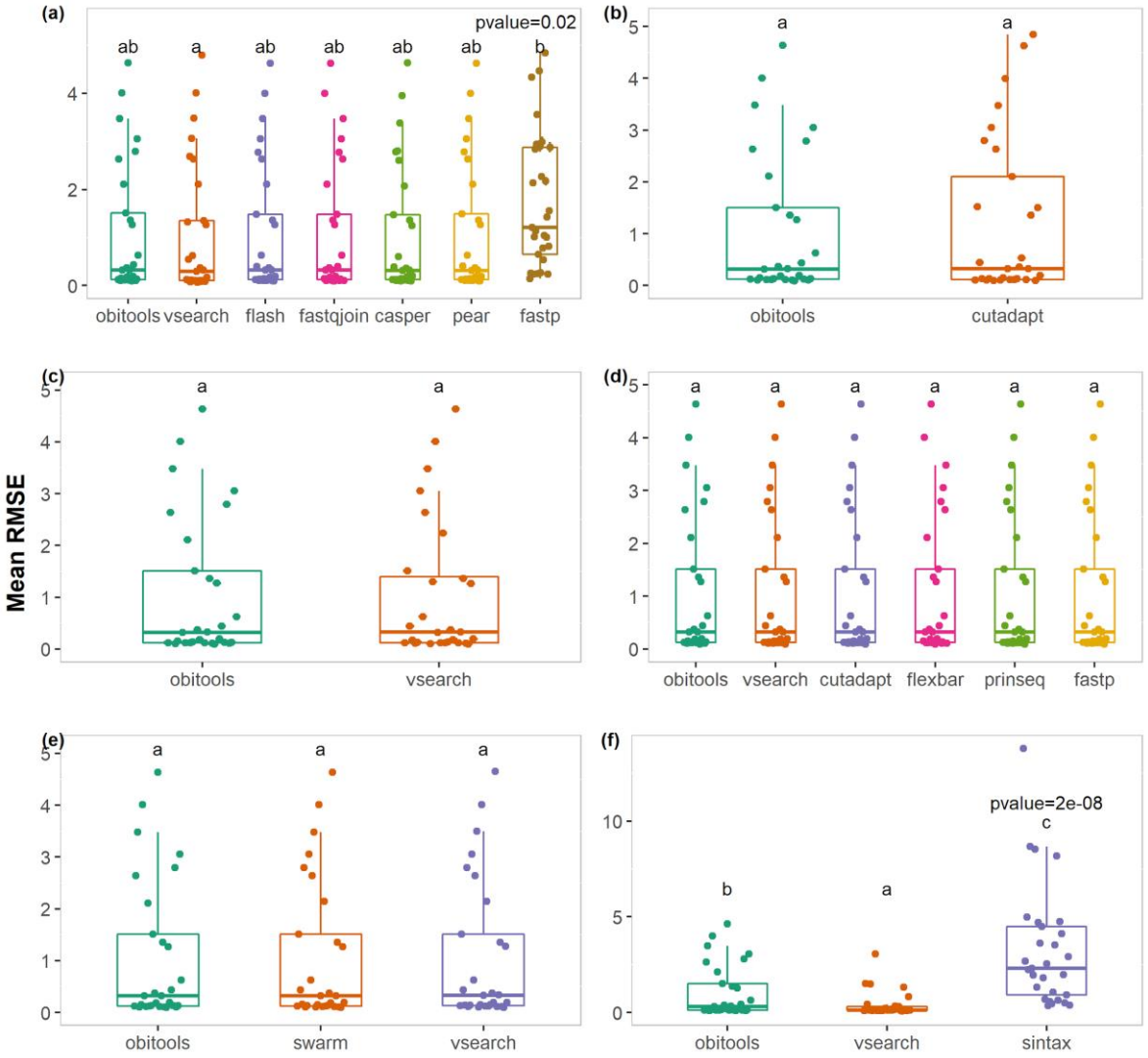


***Figure S3**. **Mean RMSE between observed and expected relative abundances obtained with each program tested for each step on the simulated dataset.** The dots represent the mean RMSE for the 12 replicas of each sample, with the standard error; the boxplot represents the median and the first and third quartiles of RMSE for the 29 samples. **(a)** Programs compared for the assembly step, **(b)** programs compared for the demultiplexing step, **(c)** programs compared for the dereplication step, **(d)** programs compared for the filtering step, **(e)** programs compared for the error removal step, **(f)** programs compared for the assignment step.*

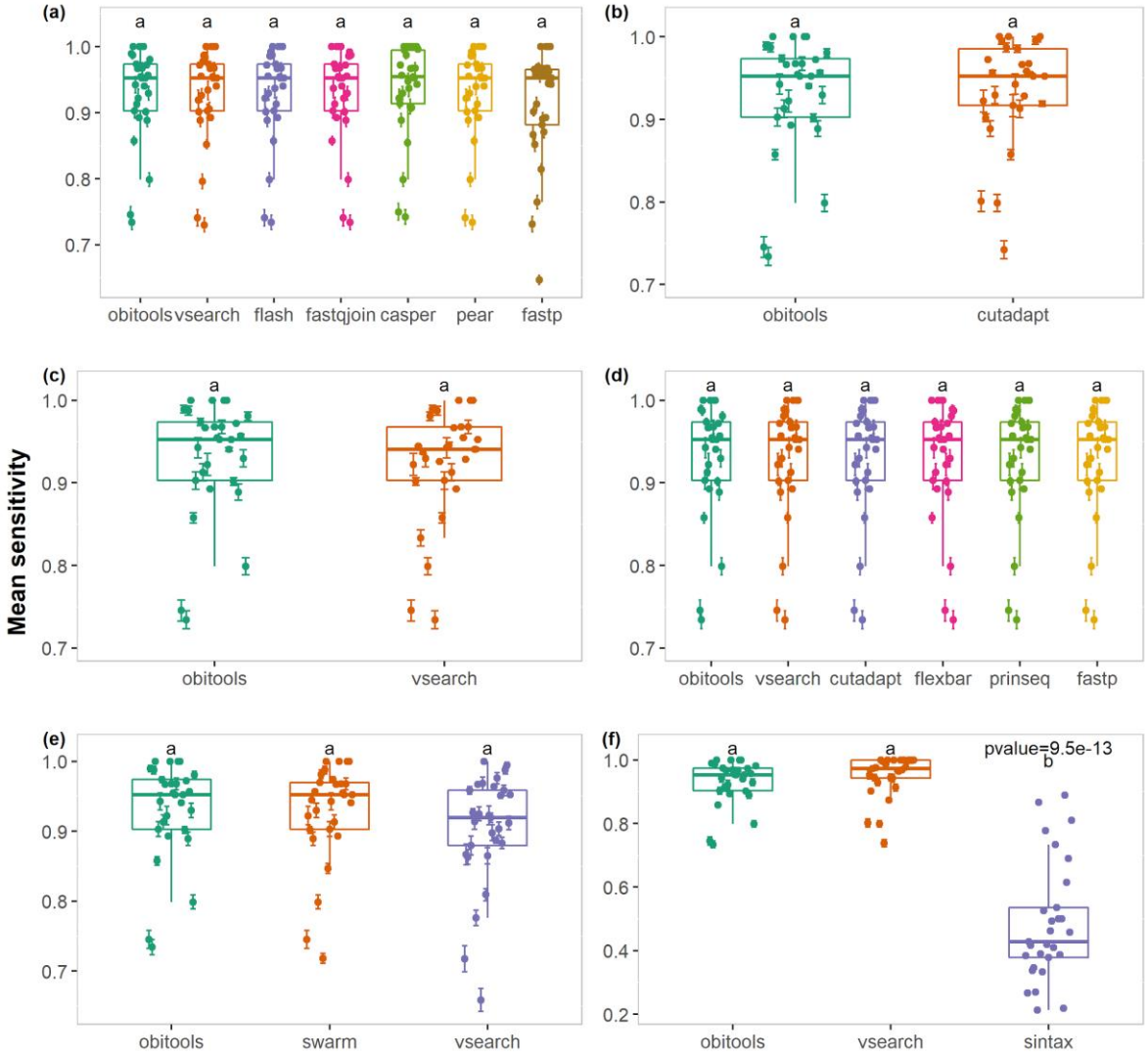


Figure S4. Mean sensitivity obtained with each program tested for each step on the simulated dataset, after removing the singletons in the outputs of each pipeline. *The dots represent the mean sensitivity for the 12 replicas of each sample, with the standard error; the boxplot represents the median and the first and third quartiles of sensitivity for the 29 samples.* ***(a)*** *programs compared for the assembly step,* ***(b)*** *programs compared for the demultiplexing step,* ***(c)*** *programs compared for the dereplication step,* ***(d)*** *programs compared for the filtering step,* ***(e)*** *programs compared for the error removal step,* ***(f)*** *programs compared for the assignation step.*

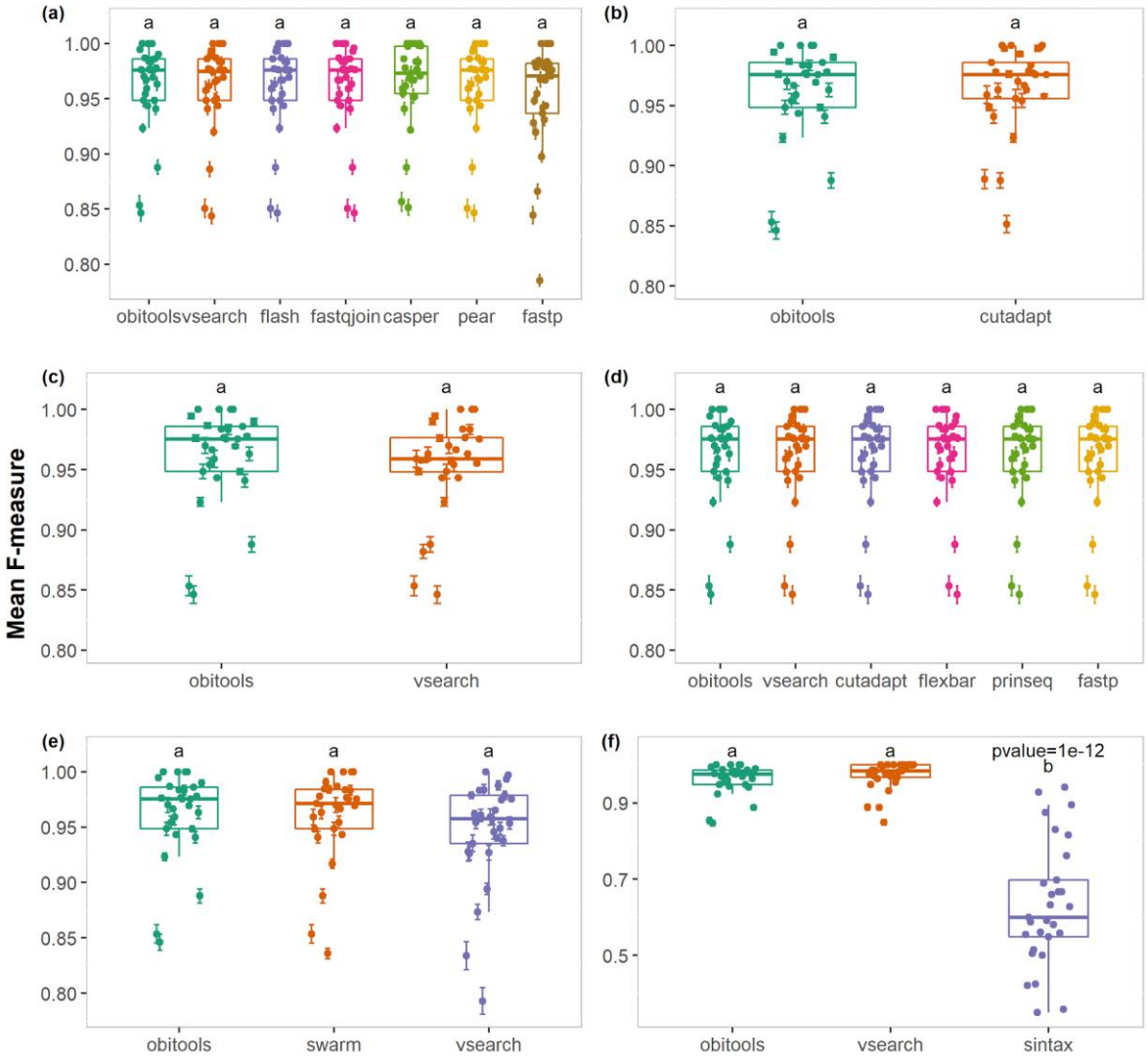


Figure S5. Mean F-measure obtained with each program tested for each step on the simulated dataset, after removing the singletons in the outputs of each pipeline. *The dots represent the mean F-measure for the 12 replicas of each sample, with the standard error; the boxplot represents the median and the first and third quartiles of F-measure for the 29 samples.* ***(a)*** *programs compared for the assembly step,* ***(b)*** *programs compared for the demultiplexing step,* ***(c)*** *programs compared for the dereplication step,* ***(d)*** *programs compared for the filtering step,* ***(e)*** *programs compared for the error removal step,* ***(f)*** *programs compared for the assignation step.*

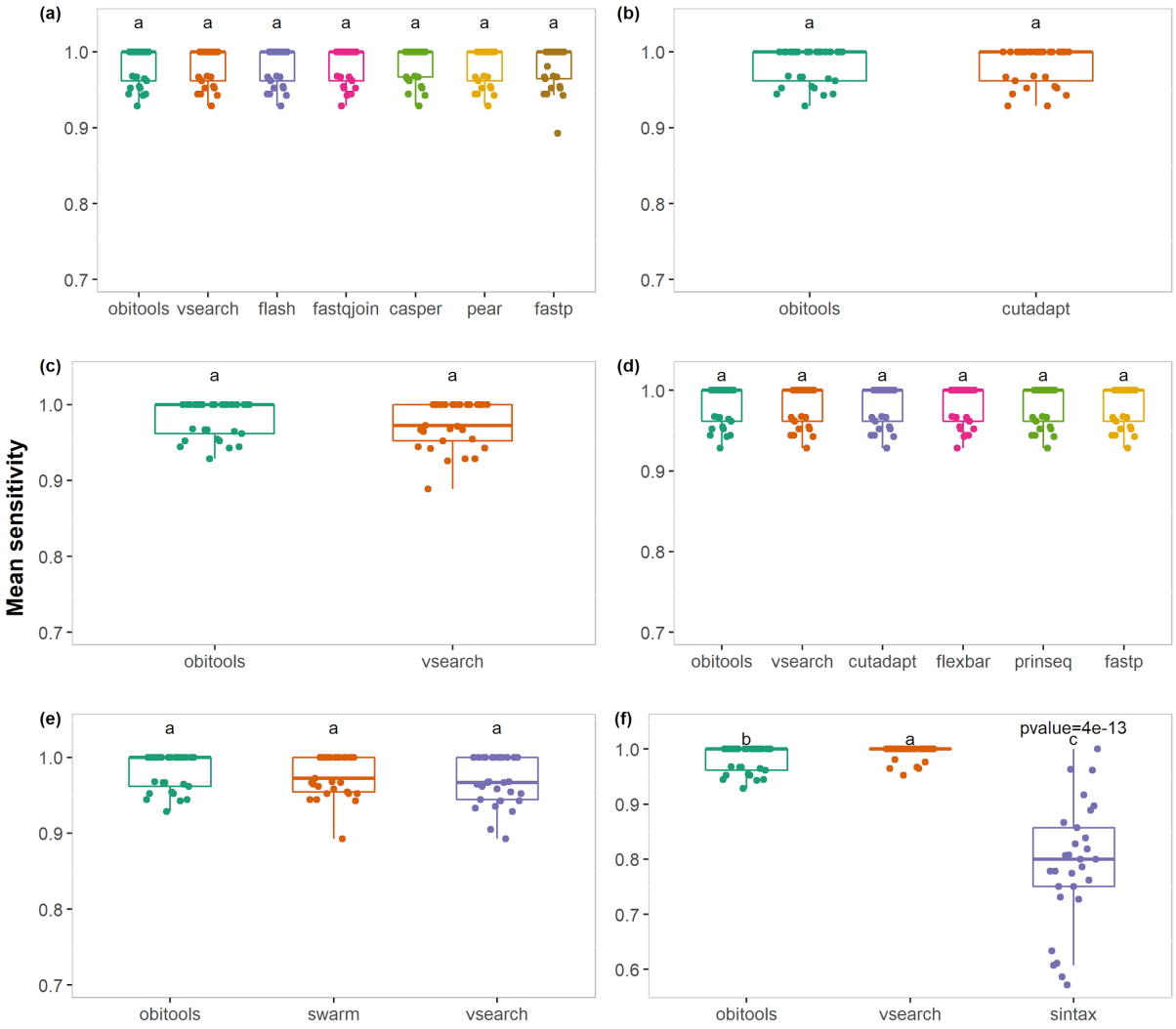


Figure S6. Mean sensitivity obtained with each program tested for each step on the simulated dataset, after summing the 12 replicates per sample in the outputs of each pipeline. *The dots represent the sensitivity of each sample; the boxplot represents the median and the first and third quartiles of sensitivity for the 29 samples.* ***(a)*** *programs compared for the assembly step,* ***(b)*** *programs compared for the demultiplexing step,* ***(c)*** *programs compared for the dereplication step,* ***(d)*** *programs compared for the filtering step,* ***(e)*** *programs compared for the error removal step,* ***(f)*** *programs compared for the assignation step.*

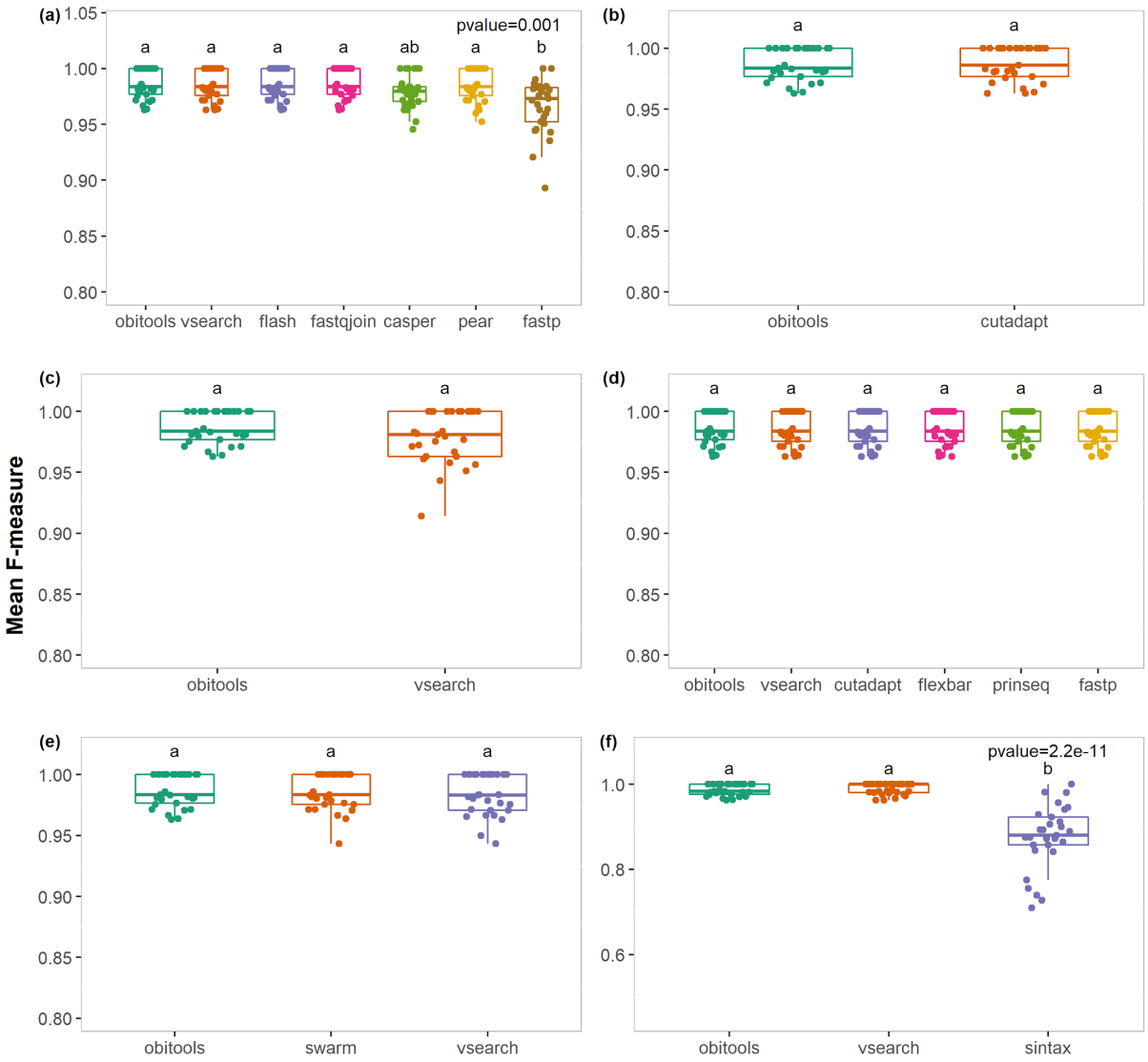


Figure S7. Mean F-measure obtained with each program tested for each step on the simulated dataset, after summing the 12 replicates per sample in the outputs of each pipeline. *The dots represent the F-measure of each sample; the boxplot represents the median and the first and third quartiles of F-measure for the 29 samples.* ***(a)*** *programs compared for the assembly step,* ***(b)*** *programs compared for the demultiplexing step,* ***(c)*** *programs compared for the dereplication step,* ***(d)*** *programs compared for the filtering step,* ***(e)*** *programs compared for the error removal step,* ***(f)*** *programs compared for the assignation step.*

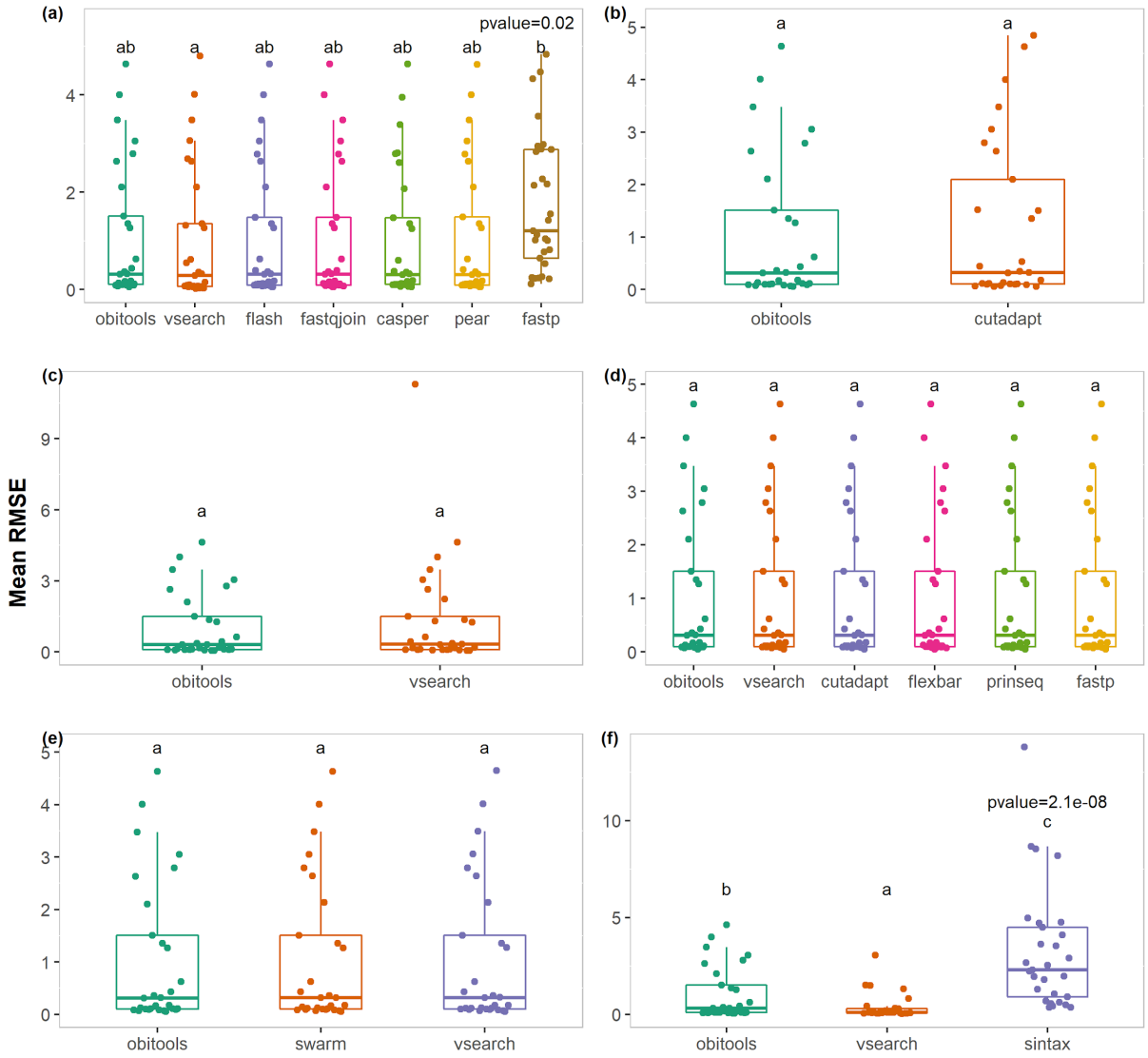


Figure S8. Mean RMSE obtained with each program tested for each step on the simulated dataset, after summing the 12 replicates per sample in the outputs of each pipeline. *The dots represent the RMSE on abundances of each sample; the boxplot represents the median and the first and third quartiles of RMSE on abundances for the 29 samples.* ***(a)*** *programs compared for the assembly step,* ***(b)*** *programs compared for the demultiplexing step,* ***(c)*** *programs compared for the dereplication step,* ***(d)*** *programs compared for the filtering step,* ***(e)*** *programs compared for the error removal step,* ***(f)*** *programs compared for the assignation step.*

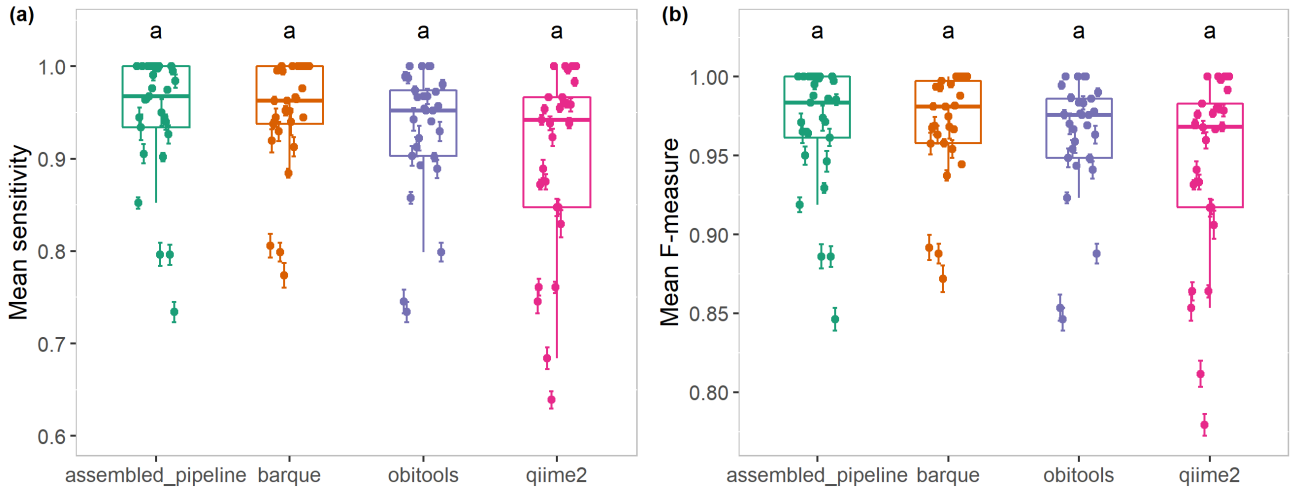


Figure S9. Performance indices of each complete pipeline on the simulated dataset, after removing singletons. *The dots represent the mean index for the 12 replicas of each sample, with the standard error; the boxplot represents the median of the index and the first and third quartiles for the 29 samples.* ***(a)*** *Sensitivity calculated on the raw outputs of each pipeline,* ***(b)*** *F-measure calculated on the raw outputs of each pipeline.*

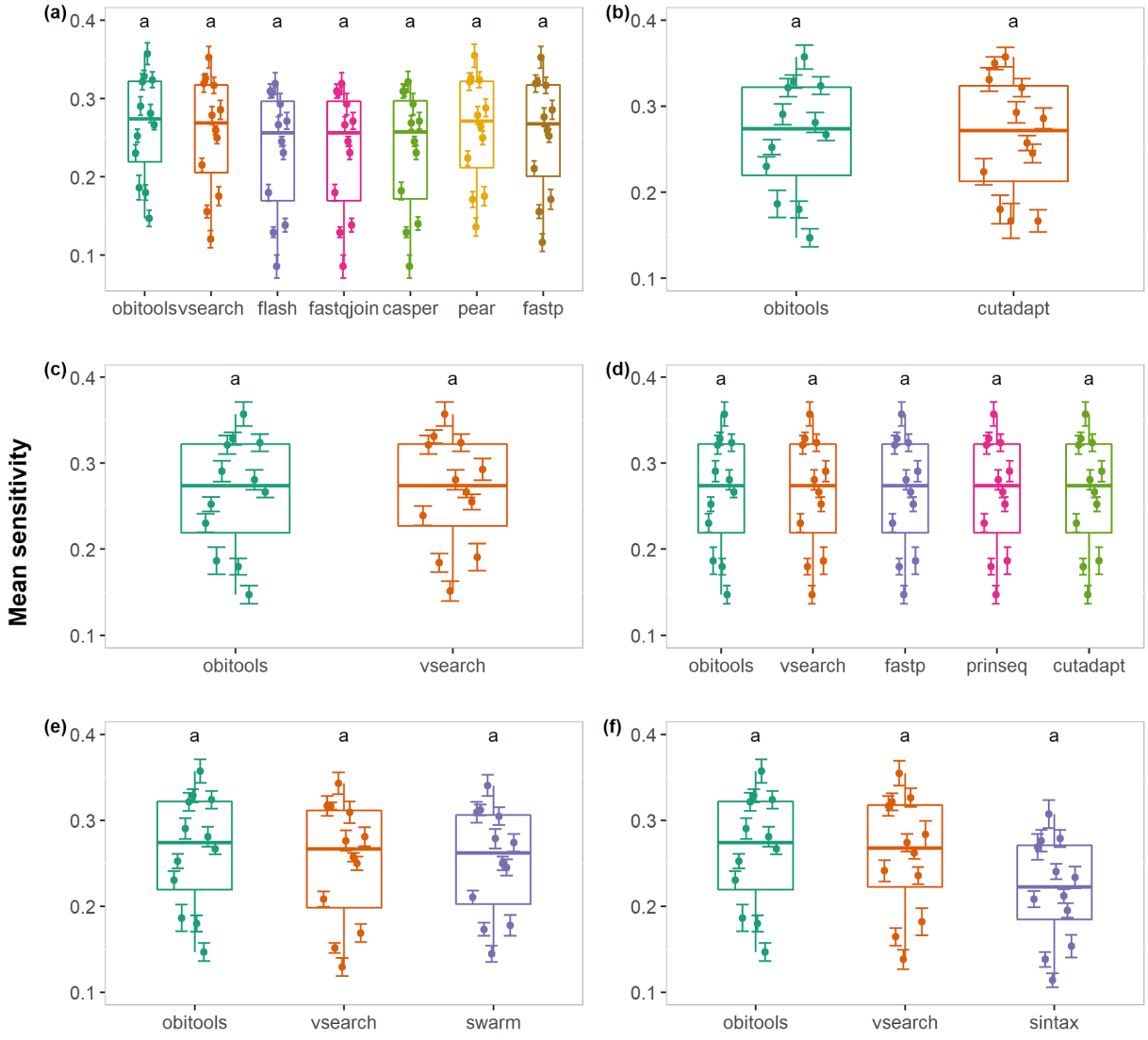


Figure S10. Mean sensitivity obtained with each program tested for each step, on the real data from Carry. *The dots represent the mean sensitivity for the 12 replicas of each sample, with the standard error; the boxplot represents the median and the first and third quartiles of sensitivity for the 12 samples.* ***(a)*** *programs compared for the assembly step,* ***(b)*** *programs compared for the demultiplexing step,* ***(c)*** *programs compared for the dereplication step,* ***(d)*** *programs compared for the filtering step,* ***(e)*** *programs compared for the error removal step,* ***(f)*** *programs compared for the assignment step.*

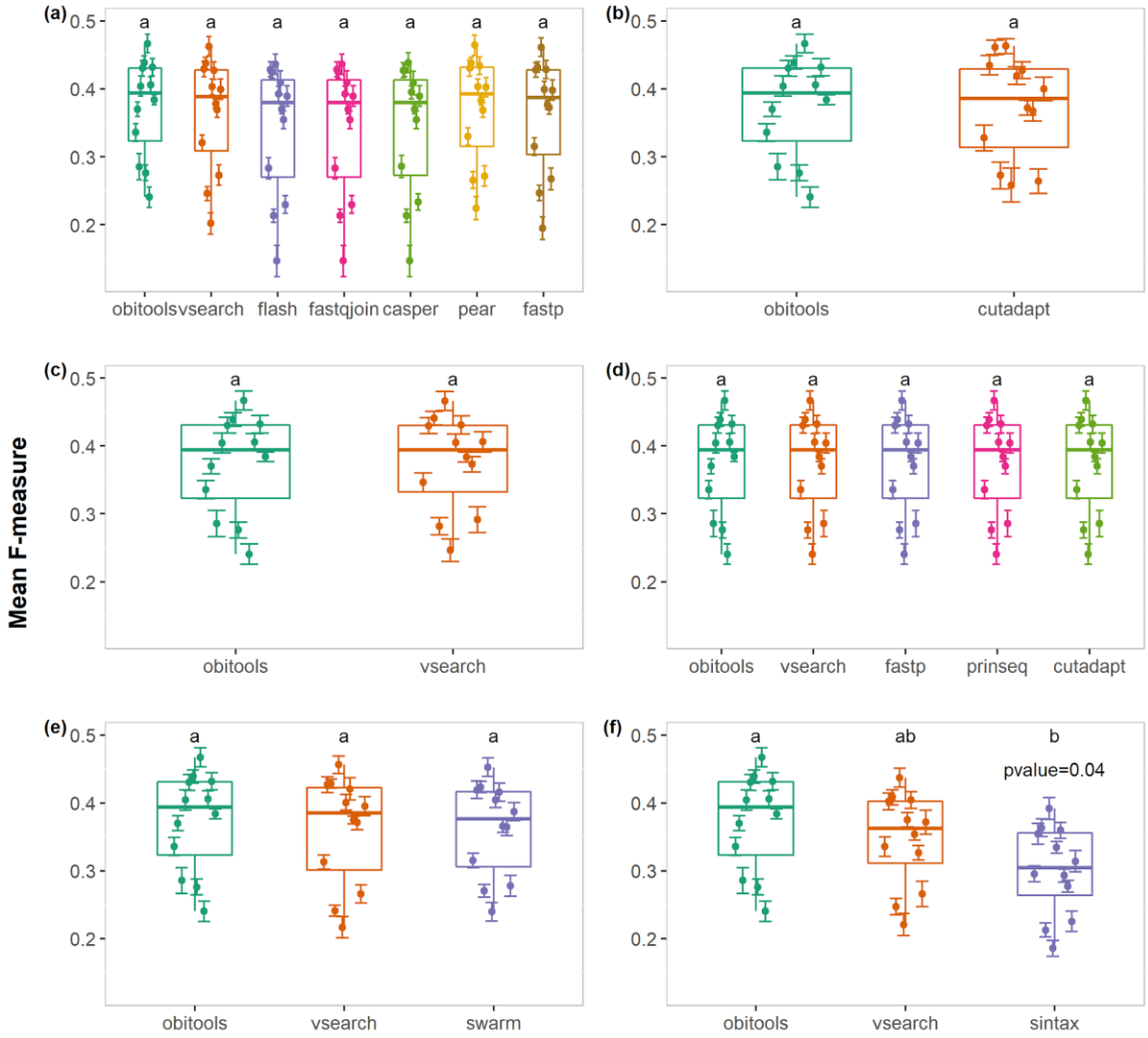


Figure S11. Mean F-measure obtained with each program tested for each step, on the real data from Carry. *The dots represent the mean sensitivity for the 12 replicas of each sample, with the standard error; the boxplot represents the median and the first and third quartiles of F-measure for the 12 samples.* ***(a)*** *programs compared for the assembly step,* ***(b)*** *programs compared for the demultiplexing step,* ***(c)*** *programs compared for the dereplication step,* ***(d)*** *programs compared for the filtering step,* ***(e)*** *programs compared for the error removal step,* ***(f)*** *programs compared for the assignment step.*

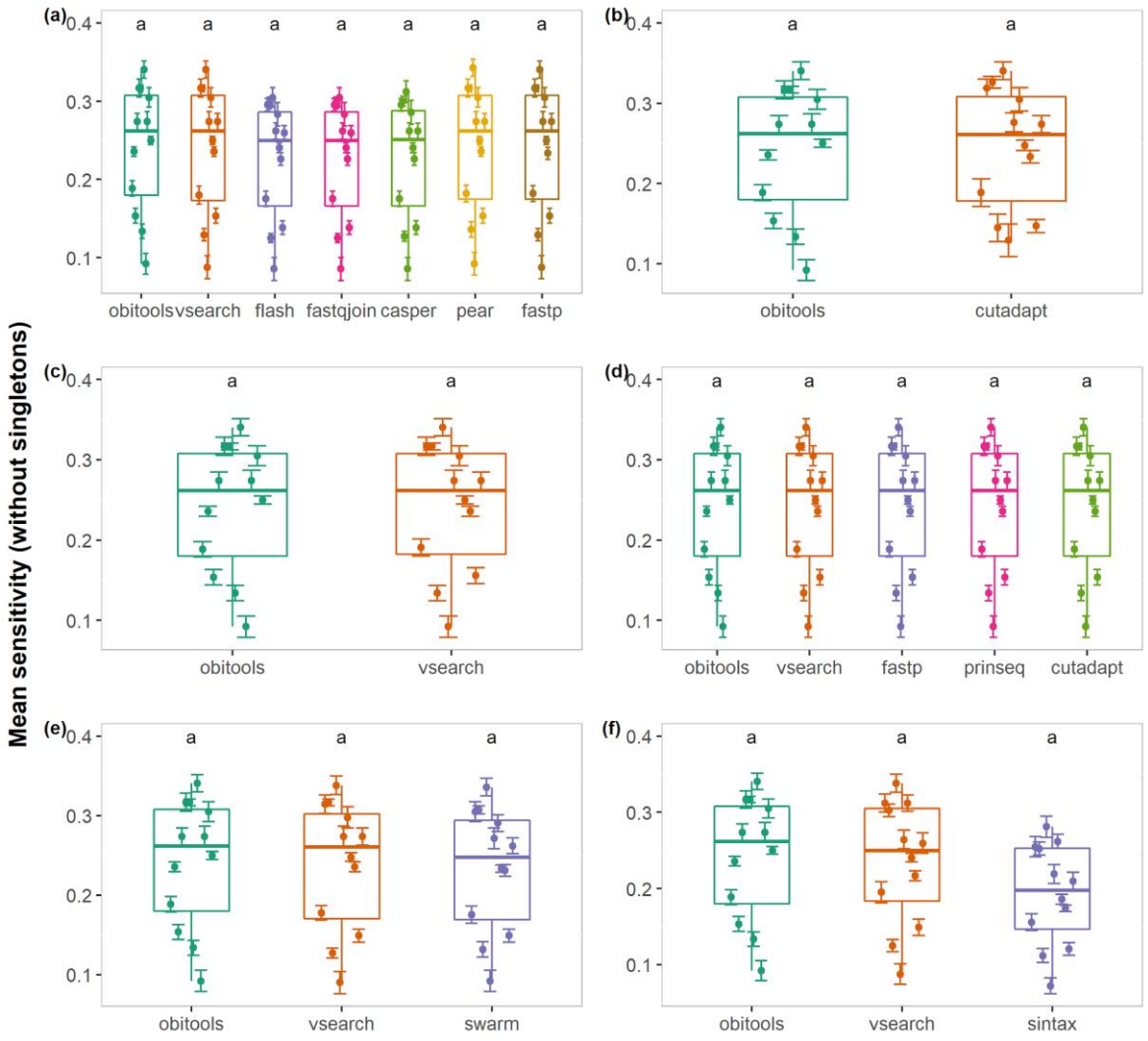


Figure S12. Mean sensitivity obtained with each program tested for each step, on the real data from Carry, after removing singletons. *The dots represent the mean sensitivity for the 12 replicas of each sample, with the standard error; the boxplot represents the median and the first and third quartiles of sensitivity for the 12 samples.* ***(a)*** *programs compared for the assembly step,* ***(b)*** *programs compared for the demultiplexing step,* ***(c)*** *programs compared for the dereplication step,* ***(d)*** *programs compared for the filtering step,* ***(e)*** *programs compared for the error removal step,* ***(f)*** *programs compared for the assignment step.*

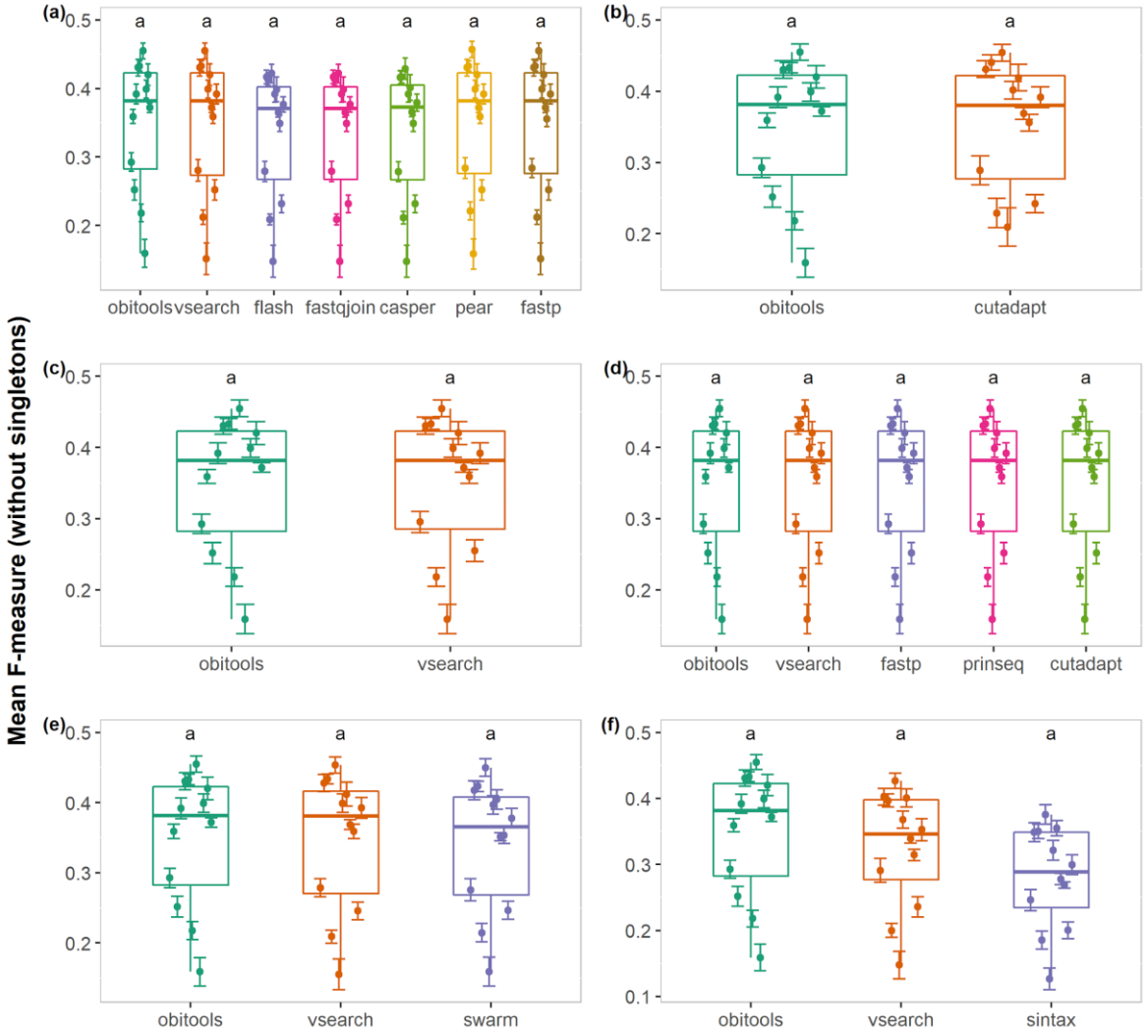


Figure S13. Mean F-measure obtained with each program tested for each step, on the real data from Carry, after removing singletons. *The dots represent the mean sensitivity for the 12 replicas of each sample, with the standard error; the boxplot represents the median and the first and third quartiles of F-measure for the 12 samples.* ***(a)*** *programs compared for the assembly step,* ***(b)*** *programs compared for the demultiplexing step,* ***(c)*** *programs compared for the dereplication step,* ***(d)*** *programs compared for the filtering step,* ***(e)*** *programs compared for the error removal step,* ***(f)*** *programs compared for the assignment step.*

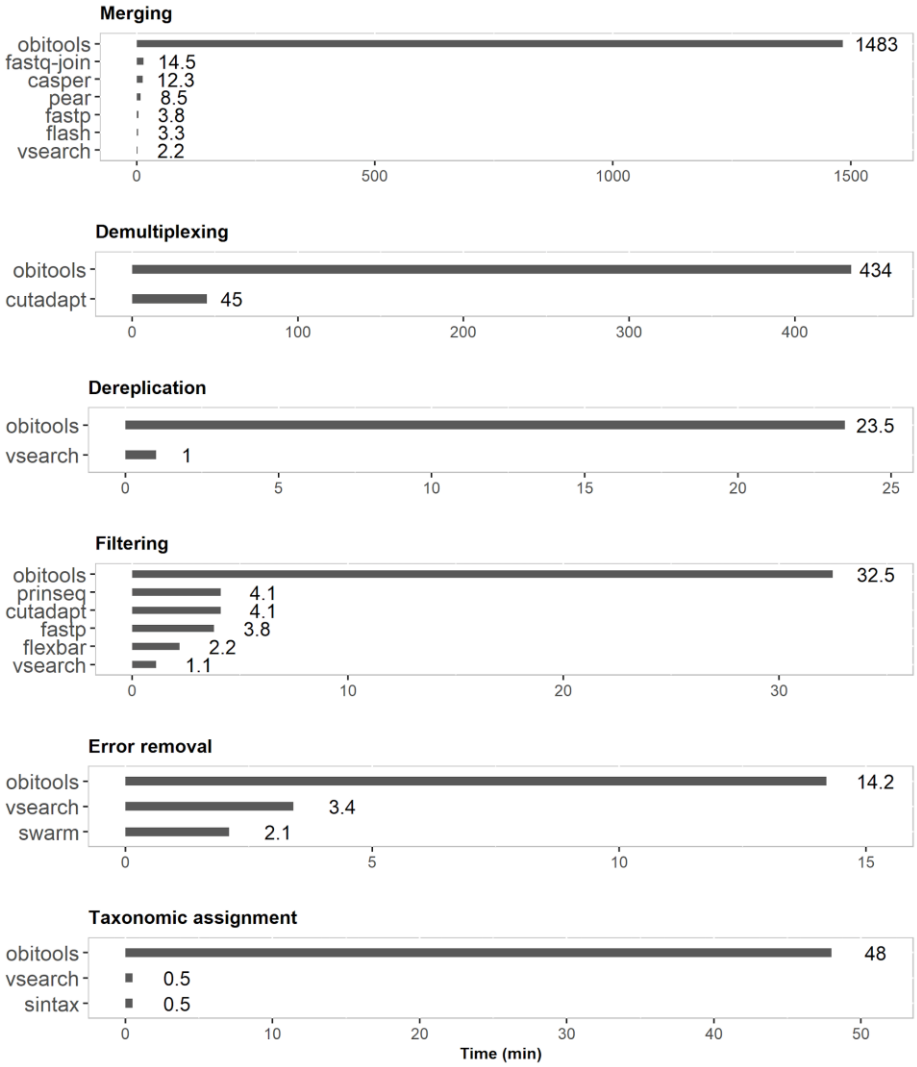


Figure S14. Execution time in minutes of each program tested for each step, on the real data from Carry. *Programs compared for the assembly step, demultiplexing step, dereplication step, filtering step, error removal step, and assignment step.*

3. Suppléments du manuscrit B : *Cross-ocean patterns and processes in fish biodiversity on coral reefs through the lens of eDNA metabarcoding*

3.1. Méthodes supplémentaires

Method S1: Environmental DNA collection and sample processing

Environmental DNA (eDNA) samples of seawater were collected in five marine regions, encompassing 26 sites (defined as groups of stations separated by at least 35 km), 100 stations and 226 samples (figure S1-S2, (1)). Three different sampling methods were used: collection of 2 L of water in DNA-free sterile plastic bags on the surface water from a dinghy as well as close circuit rebreather diving (depths between 10 – 40 m) as close as possible to the habitat (1); 2-km long filtration transect with two replicates (one on each side of a boat at each station for 30 min), for a total of 30 of water just under the surface; 2 km-long filtration of water along a transect, approximately 5 m above the substrate, using a long pipe, from the boat. Details on which sampling method was used in each region are provided in table S7. For each sample collected with the first sampling protocol, 2 L of seawater were filtered with sterile Sterivex filter capsules (Merck© Millipore; pore size 0.22µm) using disposable sterile syringes. Immediately after, the filter units were filled with CL1 Conservation buffer (SPYGEN, le Bourget du Lac, France) and stored in 50 mL screw-cap tubes at room temperature. The eDNA filtration device for the other two sampling protocols was composed of an Athena® peristaltic pump (Proactive Environmental Products LLC, Bradenton, Florida, USA; nominal flow of 1.0 $L.min^{-1}$), a VigiDNA® 0.2µM cross flow filtration capsule with a polyethersulfone membrane (SPYGEN, le Bourget du Lac, France) and disposable sterile tubing for each filtration capsule. At the end of each filtration, the water inside the capsules were 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 each sampling campaign, a strict contamination control protocol was followed in both field and laboratory stages (2,3), and each water sample processing included the use of disposable gloves and single-use filtration equipment. Negative field controls were performed in multiple sites across all sampling locations, and revealed no contamination from the boat or samplers. A large number of extraction and amplification negative controls were performed for each sample (see next section).

Method S2: eDNA extraction, amplification and sequencing

DNA extraction was performed in a dedicated DNA laboratory (SPYGEN, www.spygen.com) equipped with positive air pressure, UV treatment and frequent air renewal. Decontamination procedures were conducted before and after all manipulations. Each filtration capsule was agitated for 15 min on a S50 Shaker (Cat Ingenieurbüro™) at 800 rpm. For sterivex filters, the buffer was retrieved using a 3 mL BD Disposable Syringe with Luer-Lok™ tips, emptied into a 50 mL tube containing 33 mL of ethanol and 1.5 mL of 3M sodium acetate and, finally, stored for at least one night at -20°C. The tubes were 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) 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 two hours. Subsequently, DNA extraction was performed using NucleoSpin® Soil (MACHEREY-NAGEL GmbH & Co., Düren Germany) starting from step 6 and following the manufacturer's instructions. The elution was performed by adding 100 µL of SE buffer twice. For VigiDNA 0.2 µM filters, each capsule, containing the CL1 buffer, was agitated for 15 min on an S50 shaker (cat Ingenieurbüro™) at 800 rpm and then the buffer was emptied into two 50-mL tube 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 3M sodium acetate were added to each 50-mL tube and stored for at least one night at -20°C. The DNA extraction was performed as described above except that the two 50 mL tubes per filtration capsule were extracted separately then the two DNA samples were pooled before the amplification step. A teleost-specific 12S mitochondrial rRNA primer pair (teleo, forward primer - ACACCGCCCGTCACTCT, reverse primer – CTTCCGGTACACTTACCATG (2)) was used for the amplification of metabarcode sequences. As we analysed our data using MOTUs as a proxy for species to overcome genetic database limitations, we chose to amplify only one marker. Twelve DNA amplifications PCR per sample 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" primers (2) and 0.2 µg/µL bovine serum albumin (BSA, Roche Diagnostic, Basel, Switzerland). 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, 1 min at 72 °C and a final elongation step at 72°C for 7 min. The teleo primers were 5'-labeled 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. Negative extraction controls and negative PCR controls (ultrapure water) were amplified (with 12 replicates as well) and sequenced in parallel to the samples to monitor possible contaminations. 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. Library preparation and sequencing were performed at Fasteris (Geneva, Switzerland). A total of 18 libraries were prepared using MetaFast protocol. A paired-end sequencing (2x125 bp) was carried out using an Illumina HiSeq 2500 sequencer with the HiSeq Rapid Flow Cell v2 using the HiSeq Rapid SBS Kit v2 (Illumina, San Diego, CA, USA) or a MiSeq (2x125 bp, Illumina, San Diego, CA, USA) using the MiSeq Flow Cell Kit v3 (Illumina, San Diego, CA, USA) or a NextSeq sequencer (2x125 bp, Illumina, San Diego, CA, USA) with the NextSeq Mid kit following the manufacturer's instructions. This generated an average of 1,335,896 sequence reads (paired-end Illumina) per sample.

Methods S3: Bioinformatic analysis

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) (4). First, reads were assembled using *vsearch* (5), then demultiplexed and trimmed using *cutadapt* (6) and clustering was performed using *Swarm* v.2 (7) with a minimum distance of 1 mismatch between clusters. Taxonomic assignment of MOTUs was carried out using the Lower Common Ancestor (LCA) algorithm *ecotag* implemented in the Obitools toolkit (8) and the European Nucleotide Archive (ENA (9)) as a reference database (release 143, March 2020). It assigns a taxonomy to sequences even when the sequence match is not perfect, based on NCBI taxonomic tree of species to consider the current knowledge on molecular diversity per branch and assign a taxonomy at the lowest possible rank. If the sequence matches several identifications with equal percentages of similarity, *ecotag* assigns to the upper taxonomic level common between all possible matches. We then applied quality filters to be conservative in our estimates. We discarded all observations with less than 10

reads, and present in only one PCR per site to avoid spurious MOTUs originating from a PCR error. Then, errors generated by index-hopping (10) were filtered using a threshold empirically determined per sequencing batch using experimental blanks (combinations of tags not present in the libraries) (11), and tag-jump (12) was corrected using a threshold of 0.001 of occurrence for a given MOTU within a library. Taxonomic assignments at the species level were accepted if the percentage of similarity with the reference sequence was 100%, at the genus level if the similarity was between 90 and 99%, and at the family level if the similarity was > 85%. If these criteria were not met, the MOTU was left unassigned. The post-LCA algorithm correction threshold of 85% similarity for family assignment was chosen to include a maximum of correct family assignment while minimizing the risk of adding wrong family assignments in the family detections, and only 21% of assigned MOTUs are assigned to the family level with a similarity between 85 and 90% (Table S8).

Methods S4: Visual Census data

The visual census survey data used here is a subset (2047 transects, figure S1)) of the complete visual census data (3027 transects) provided by the Reef Life Survey (13), and comprises all species observed on standardized 50 m surveys at sites in tropical biogeographic realms (14). We selected only the most recent survey for each transect and only transects with more than five percent of coral cover. The visual census method involves divers surveying duplicate 5-m-wide blocks along each 50 m transect in which all fish species sighted are recorded, and then in duplicate 1-m-wide blocks in which the divers closely search the substrate (including in crevices) for smaller crypto-benthic fishes (13). The full list of fish species for each survey from both methods was used for this study. Full details of the methods are provided in an online methods manual at www.reeflifesurvey.com.

Methods S5. Statistical analysis

Accumulation curves were calculated for species per 500 m^2 transect, MOTUs per eDNA sample, and families per transect and sample. We used the function “specaccum” from the R package “vegan” v.2.5-6, with the "exact" method, which calculates the expected species accumulation curve using a sample-based rarefaction method. We then used the function “fitspecaccum” to fit five nonlinear species accumulation models (Lomolino, Michaelis-Menten, Gompertz, Asymp and Logis). The best model was selected based on AIC, and its asymptote recorded. Sampling effort varied between regions in the Visual Census dataset, with Australia having twice as many transects as other regions. In order to assess the impact of this irregular sampling on the estimates measured with accumulation curves, we randomly subset half of the transects in the 3 most sampled regions in Australia, and calculated again the accumulation curves for species and families (figure S12). The results were unchanged.

Pearson’s correlation coefficient was calculated between the number of MOTUs per family in the eDNA dataset and the number of species per family in the visual census dataset. Linear regression models were fitted between the number of MOTUs per family in the eDNA dataset and the number of species per family in the visual census dataset, after log(x+1) transformation (figure 1*e*).

Accumulation curves were also calculated by sub-setting MOTUs belonging to crypto-benthic orders, or to pelagic families, for both datasets (figure 2). The asymptote was calculated as described above.

MOTU proportions of each fish family (i.e. family proportions) were calculated as the number of MOTUs assigned to each family in each site for eDNA and species assigned to each family in each site for the Visual Census. We performed distance-based Redundancy Analysis (dbRDA) on these family proportions, with *region* and *site richness* as explanatory variables, using the function *capscale* from the *vegan* package. We subset the Visual Census to select only the 68 sites that fell into the 5 regions in common with the eDNA dataset. Total dbRDA provided the effects of each of the variables and their interaction. We then calculated partial dbRDA to measure the effect of the Region while correcting for the effect of site richness (figure 3, table S3).

As eDNA is rapidly degraded in tropical inshore waters (15,16), and based on caged experiments in marine ecosystems (17), we assume the eDNA signal comes from individuals present in close proximity to the filtering station. Thus, the detection of species not typically considered as coral reef fishes may reveal their use of reef habitats from time to time (18).

We applied an additive partitioning framework (19,20) to separate the total MOTUs diversity at the global scale (ɣ global) into contributions at smaller scales from regions to local richness. More precisely, global MOTUs diversity was expressed as the sum of inter-region difference, the mean inter-site difference, the mean inter-station difference and mean station MOTUs diversity with: $\gamma_{global} = \beta_{inter\text{-}region}$ + mean $\beta_{inter\text{-}site}$ + mean $\beta_{inter\text{-}station}$ + mean $\bar{\alpha}_{station}$. In this additive framework, the three levels of biodiversity (21) (i.e. α, β and γ) are expressed with the same unit and consequently the contribution of α and β diversity to total diversity (ɣ) can be directly compared (22,23). The diversity partitioning in figure 4 has been calculated with sites defined as groups of stations distant from 35km. In order to assess the influence of the spatial scale in site definition on the diversity partition, we repeated the diversity analysis with sites defined as groups of stations distant from 10 or 20 km (table S4). The results were similar.

We analyzed the distribution of fish MOTU and species occurrences using global species abundance distribution (gSAD) which plots, on a log-log scale, the number of species as a function of the number of observations (24). This representation has the advantage of being comparable between datasets sampled with different methods and allowing the testing of several species assembly rules and models at large scale. For example, the unified neutral theory of biogeography (UNTB) (25) would produce a gSAD following a log series model or Pareto model with a slope β = -1 while niche-based processes would provide β values indicating more or less rare species than under the UNTB if β values are respectively higher or lower than -1. Testing whether the gSAD is best fit by a log series or a Pareto distribution (where β is allowed to vary) provides a test of neutral dynamics. Additionally, a third model, coined the Pareto with exponential finite adjustment, adds an exponential "bending" parameter to the Pareto model allowing the right tail to drop down because of finite sample size. Thus, fitting the Pareto or the Pareto with exponential finite adjustment provides a test of neutral or niche dynamics with a β value $\neq$ -1 rejecting the neutral theory while a β value <-1 indicates more rare species than under neutrality and >-1 fewer. We summed all fish MOTUs and species observations across all samples obtained with eDNA and visual census data to build gSAD that were fitted with a log series, Pareto and Pareto with exponential finite adjustment (Pareto bended) distribution using maximum likelihood estimation.

3.2. Figures supplémentaires

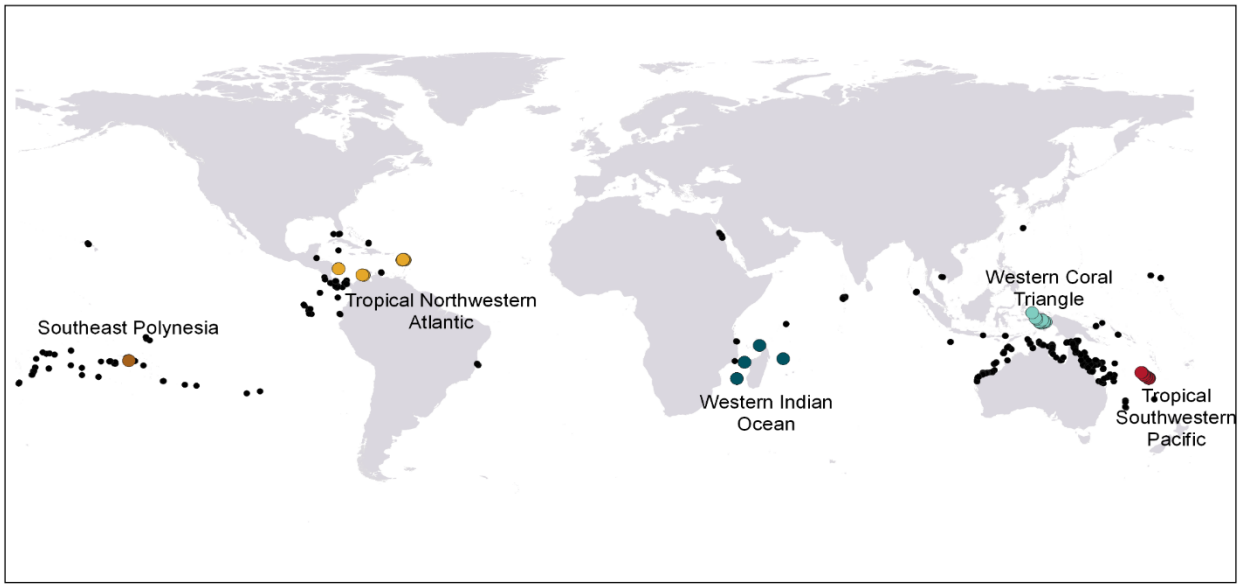


Figure S1. Global map of the sampling locations. *The 26 eDNA sampling sites (including 100 stations) are represented by colored dots (colors represent regions). The 219 UVC sampling sites (including 2,047 transects) are represented by the black dots.*

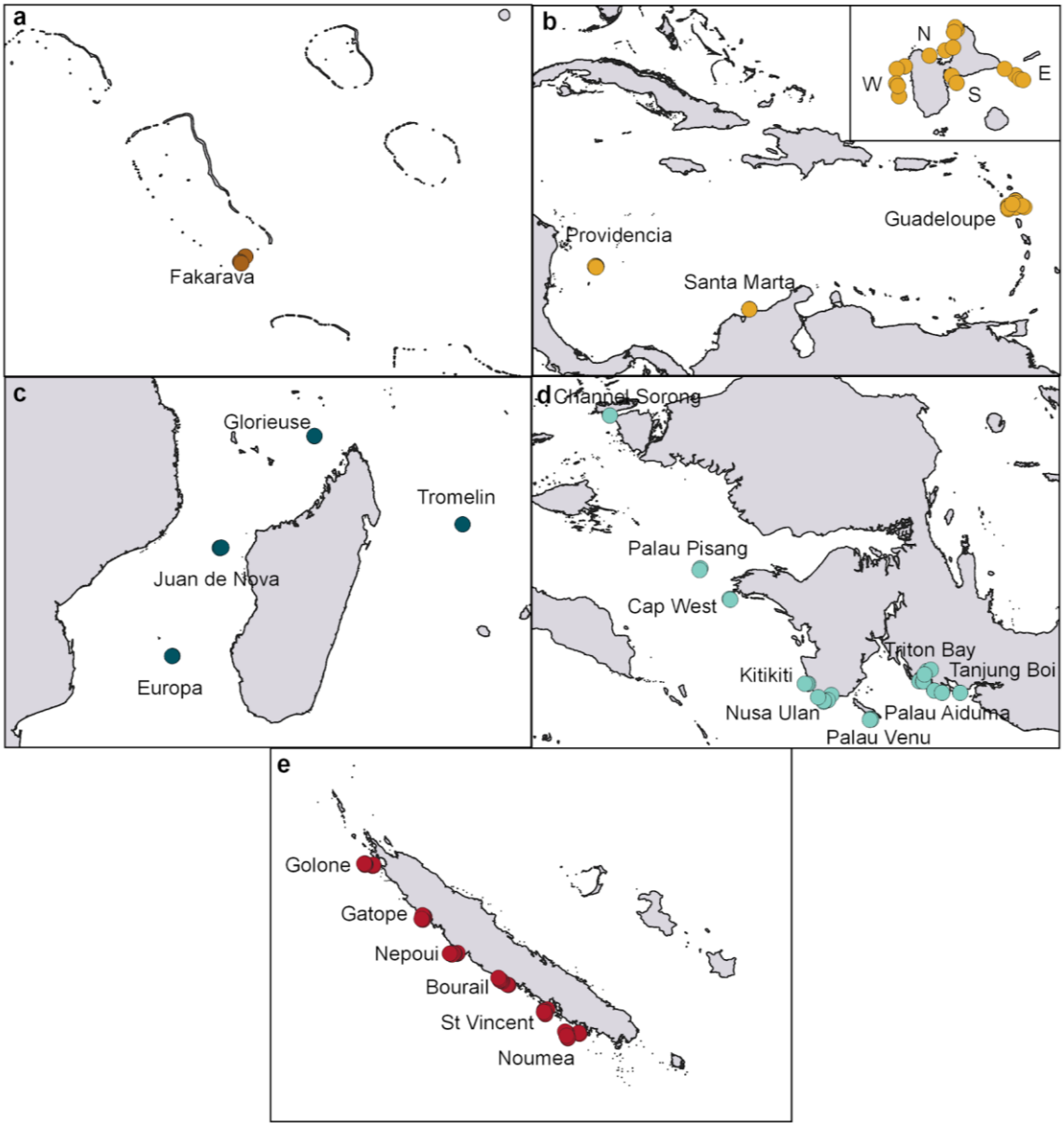


Figure S2. Sampling locations. *Map of our sampling in the 5 regions including 26 sites: (a) 1 site in Southeast Polynesia, (b) 4 sites in Western Indian Ocean, (c) 6 sites in the Tropical Northwestern Atlantic, (d) 9 sites in Western Coral Triangle and (e) 6 sites in Tropical Southwestern Pacific. The 100 eDNA stations are overlapping at each site.*

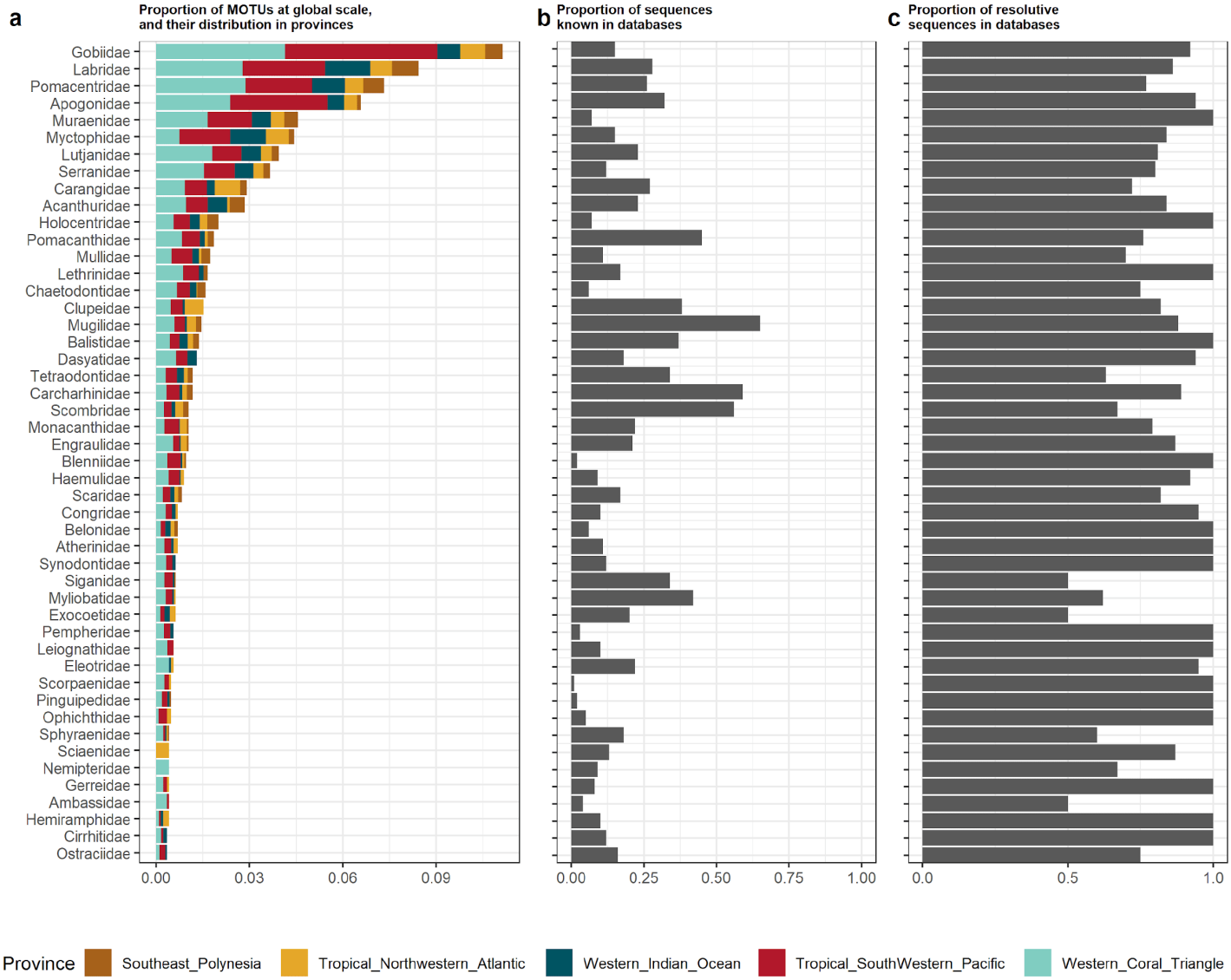


Figure S3. Characteristics of the 48 families identified in our study, with more than 5 MOTUs. *(a) Proportion of MOTUs assigned to each family at global scale, and proportion in each region (b) proportion of 12S sequences in databases for each family. (c) Proportion of resolutive sequences in the reference database of our barcode for each family (= distinguishable species). All families with less than 100% of resolution (45% of all families), might result in an underestimated MOTU richness, due to a perfect genetic match of the 12S rRNA teleo marker between some species within these families.*

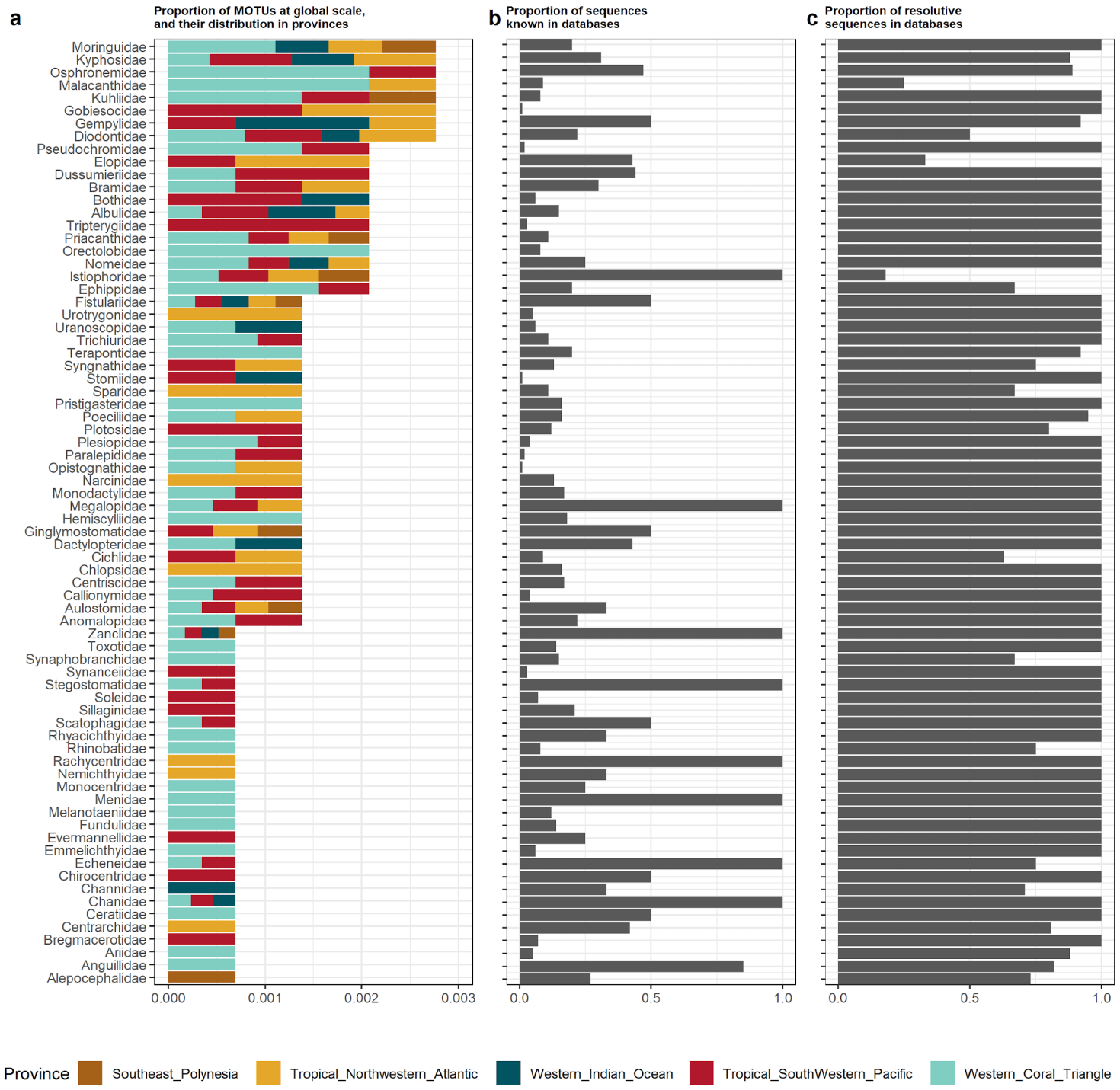


Figure S4. Characteristics of the 78 families identified in our study, with less than 5 MOTUs. *(a) Proportion of MOTUs assigned to each family at global scale, and proportion in each region. (b) proportion of 12S sequences in databases for each family. (c) Proportion of resolutive sequences in the reference database of our barcode for each family (= distinguishable species). All families with less than 100% of resolution (45% of all families), might result in an underestimated MOTU richness, due to a perfect genetic match of the 12S rRNA teleo marker between some species within these families.*

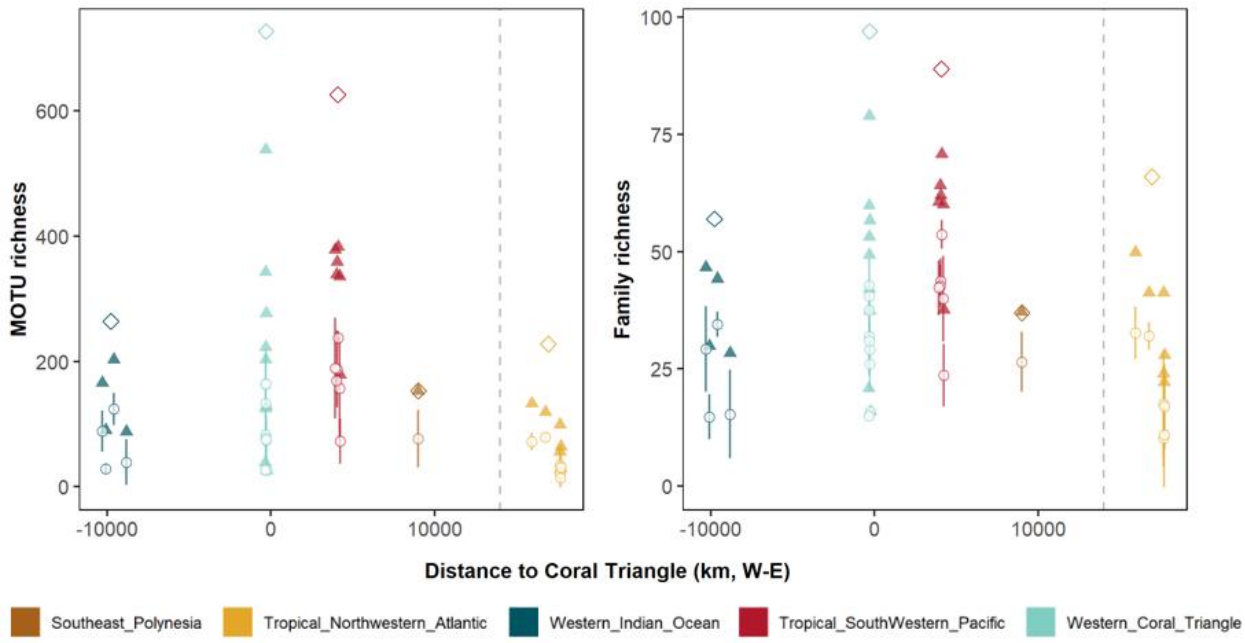


Figure S5. MOTUs and Family richness according to the distance to the coral triangle. *Mean MOTUs (left) and mean Family (right) richness per station in each site ± standard deviation (empty circles and vertical bars), total site richness (filled triangles) and total region richness (empty diamonds) 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 Western Coral Triangle and Tropical Southwestern Pacific richest than the three other regions). (Kruskal Wallis test among sites: $p < 0.001$, $n=26$, Dunn test of pairwise comparisons: $p < 0.001$).*

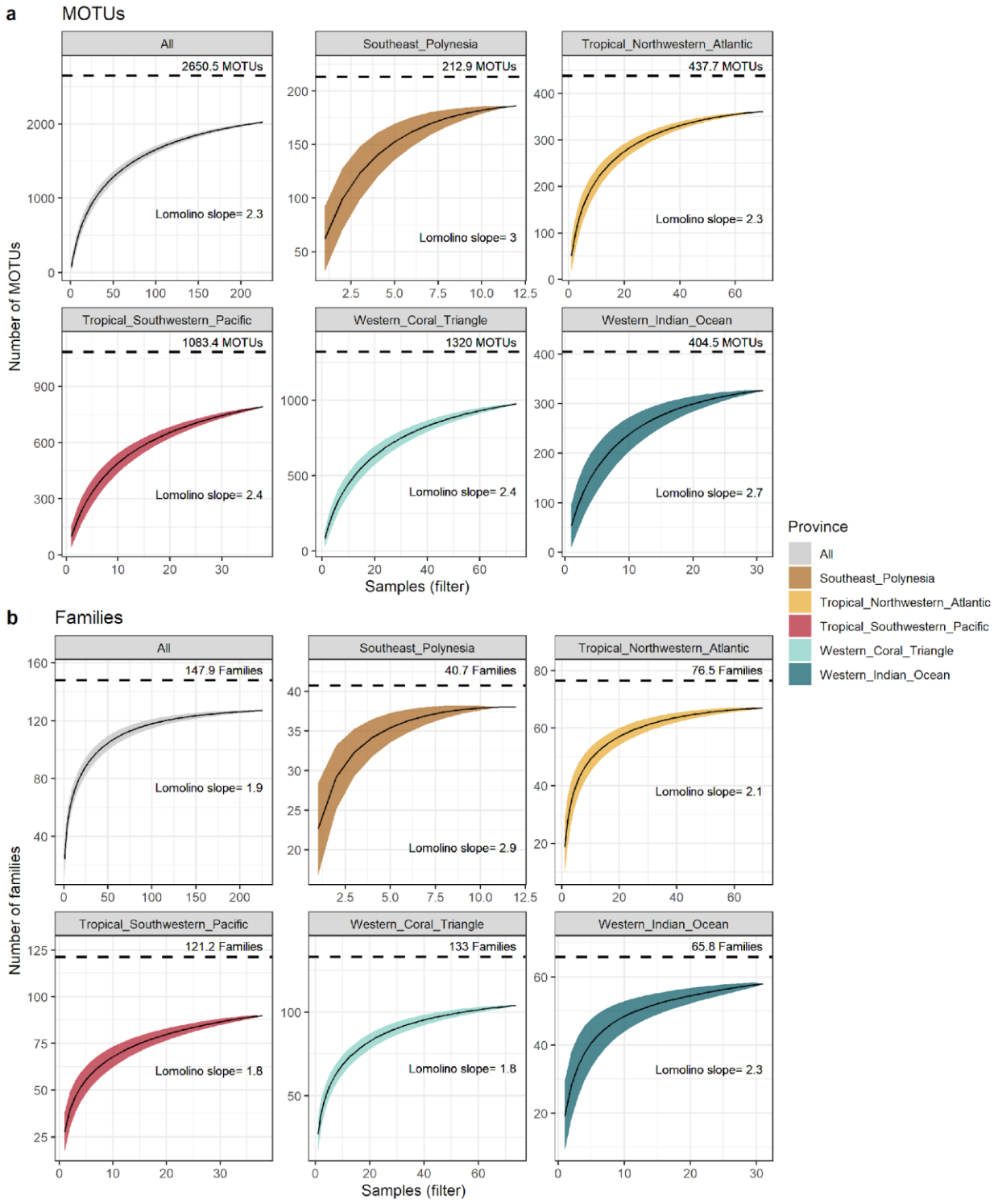


***Figure S6. Accumulation curves per region**. (a) of MOTUs and (b) of families in each region, according to the number of samples. Accumulation model is fitted with a nonlinear lomolino model (see methods).*

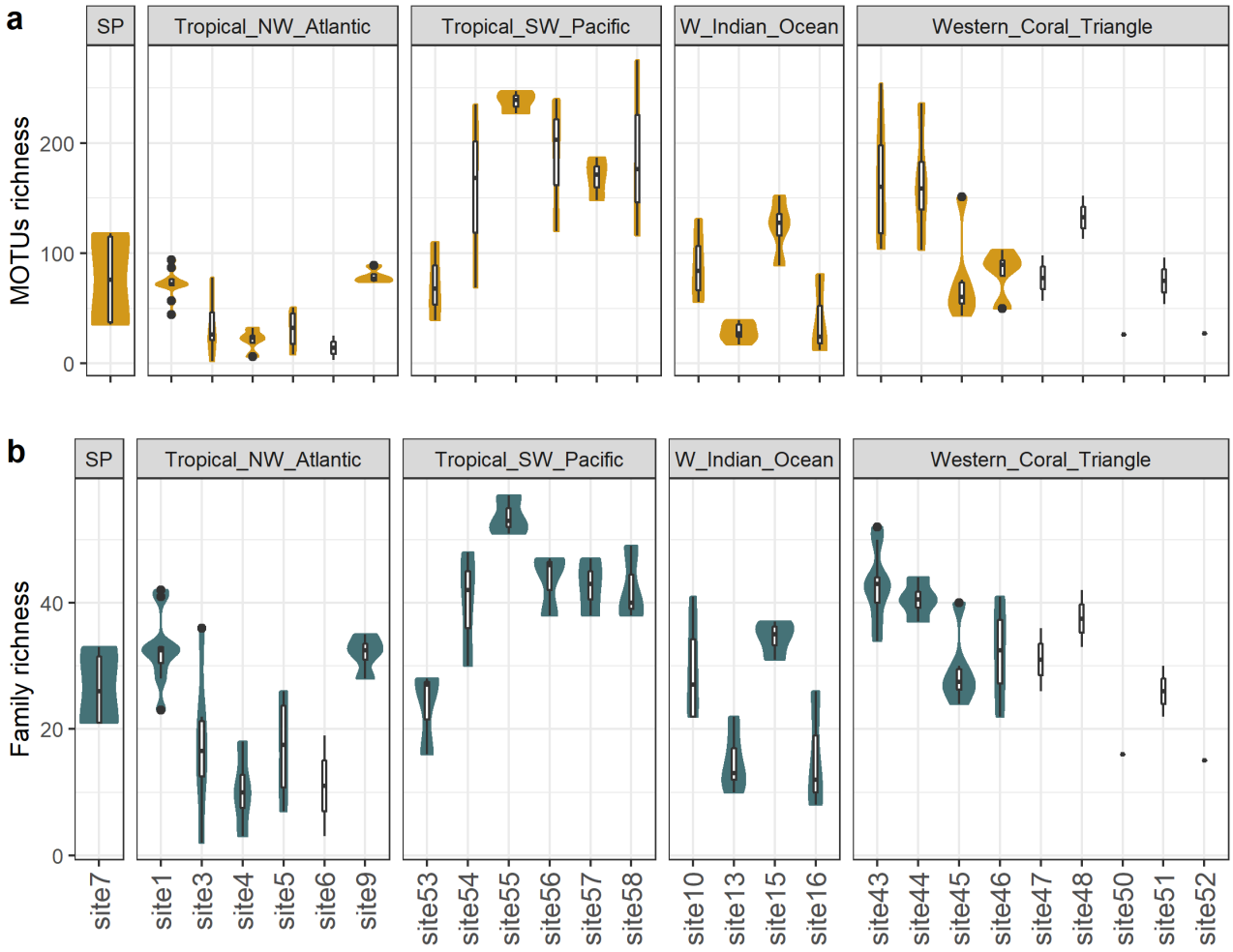


Figure S7. Richness per site in each region. *(a) MOTU richness, (b) family richness. Boxplots represent median and quartiles of richness per station. Violin plots represent the density of probabilities of richness values among stations.*

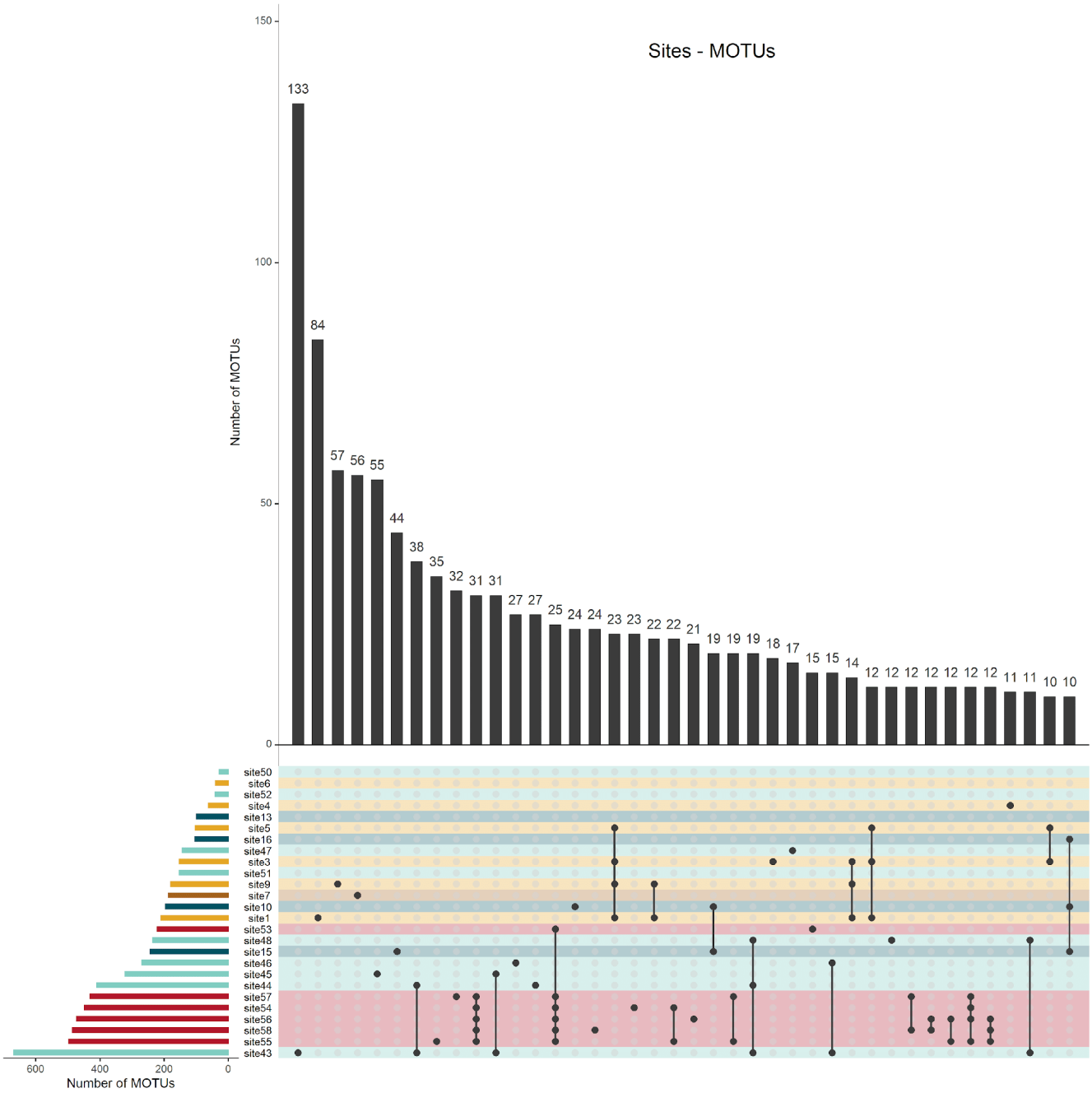


Figure S8. Distribution of MOTUs across sites. *Upset plot representing the number of MOTUs found in only one site, or shared between 2 to all 25 sites. Histograms in the upper part and numbers on top indicate the number of MOTUs present in all the sites identified by the dots in the lower part. The black lines in the lower part link the sites where the MOTUs are present, for visual simplicity. Colors show regions of each site. Horizontal histograms in the lower part indicate the MOTU richness of each site.*

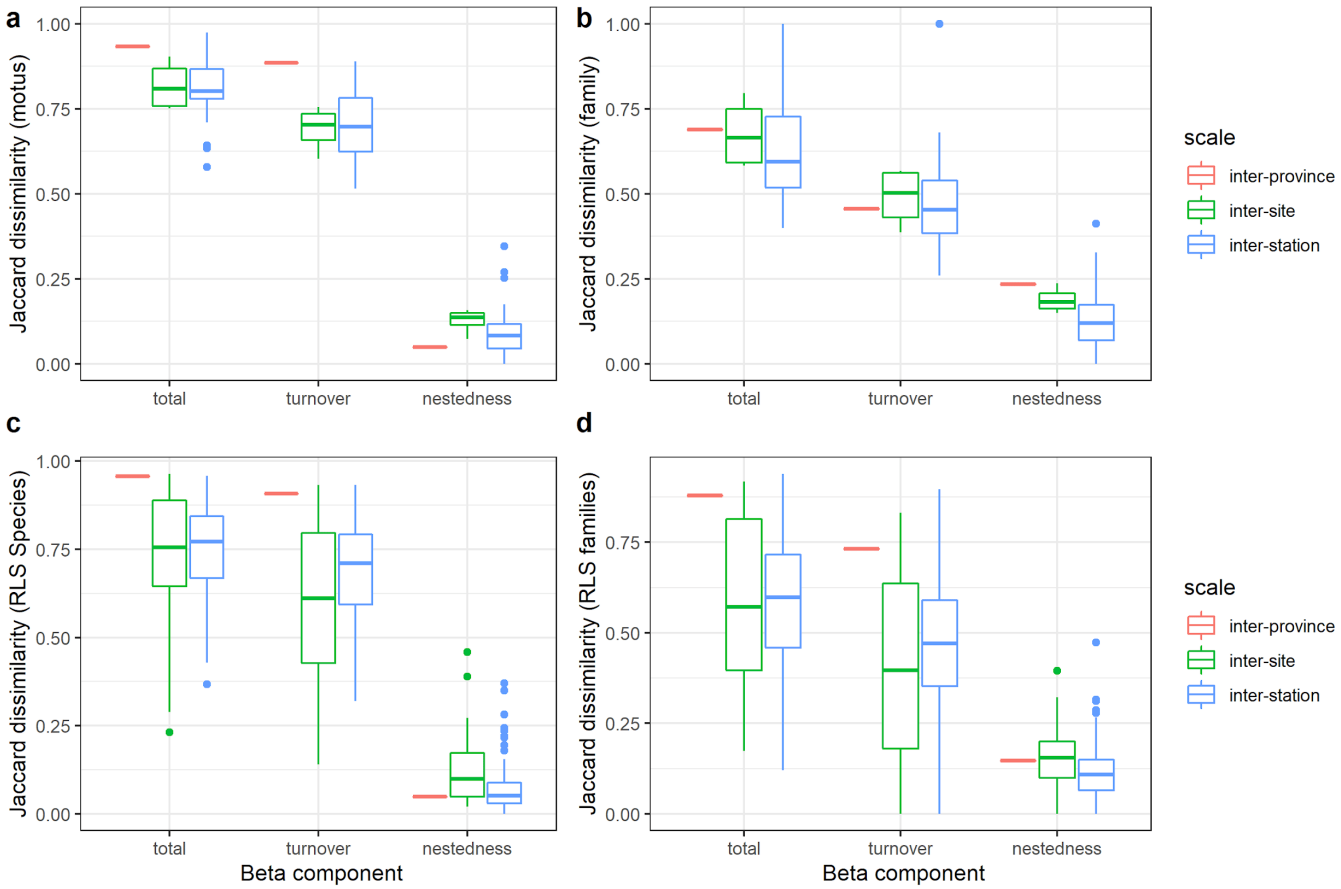


Figure S9. Beta diversity decomposition in turnover and nestedness. *(a) for eDNA MOTUs, (b) for eDNA families, (c) for visual census species and (d) for visual census families. Beta diversity is measured across spatial scales: between regions, between sites within regions, and between stations/transects within sites. Boxplots represent the median, the 1st and 3rd quartiles, and 1st and 9th deciles.*

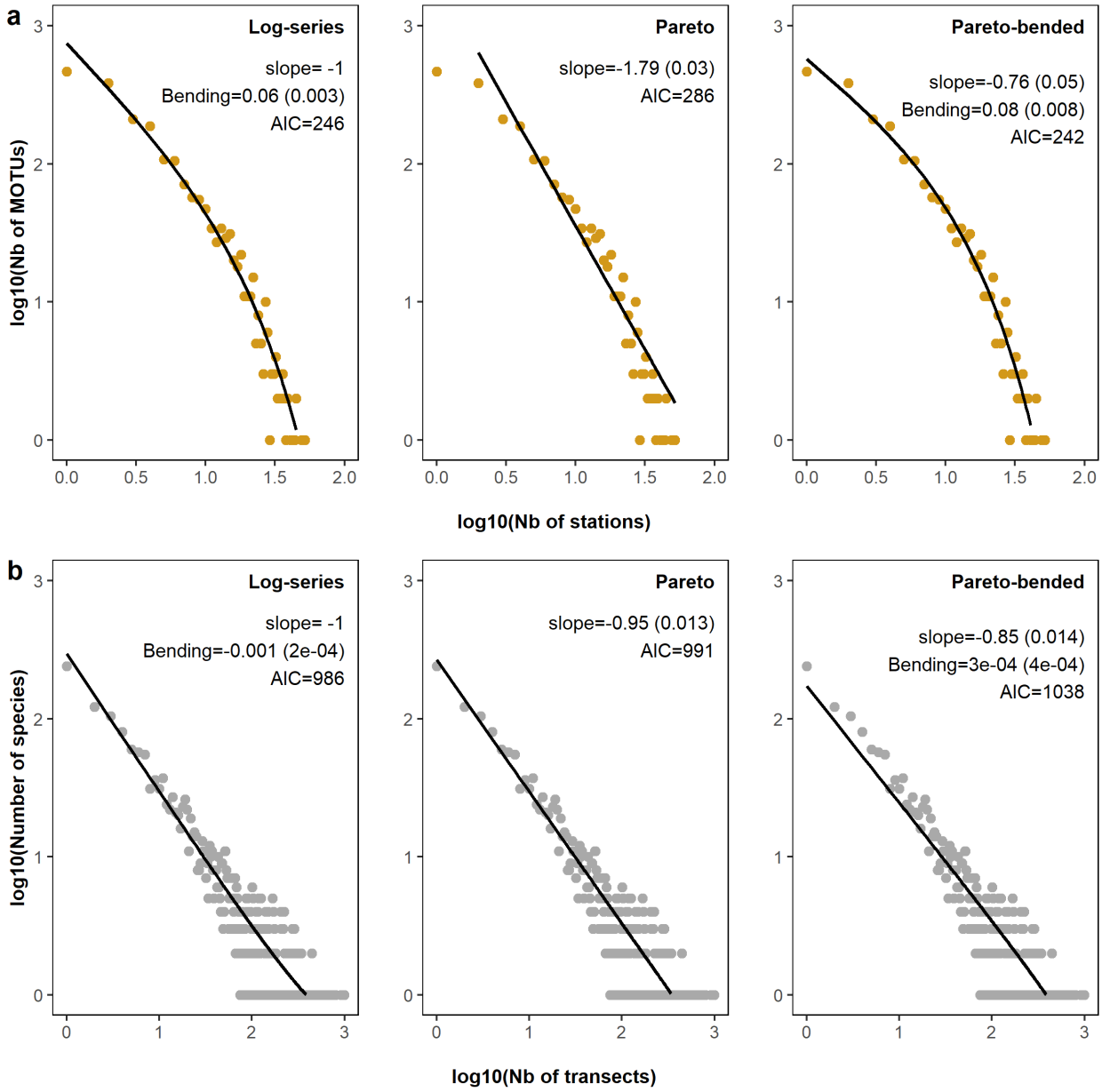


Figure S10. Distribution of the total number of global observations per fish species. *(a) Distribution of MOTU occurrences across stations, log-transformed (yellow points). (b) Distribution of visual census occurrences across transects (black points), log-transformed. For both distributions, three abundance 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.*

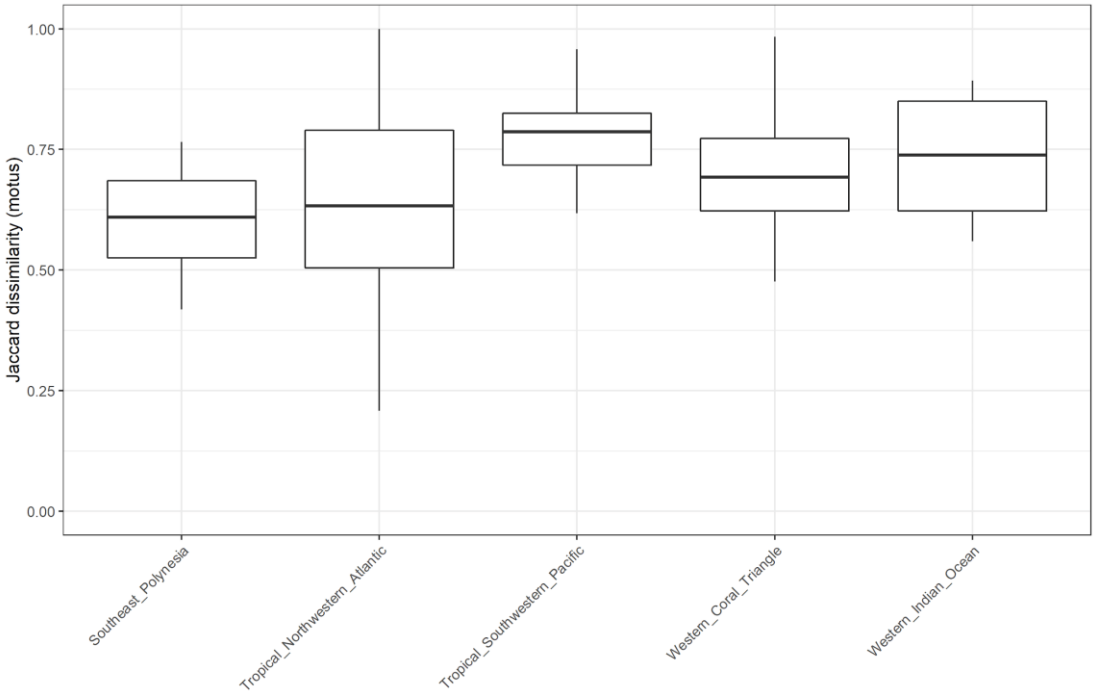


Figure S11. Beta diversity calculated between replicates of each station. *The boxplots represent the median, 1st and 3rd quartiles, and 1st and 9th deciles of beta in each region.*

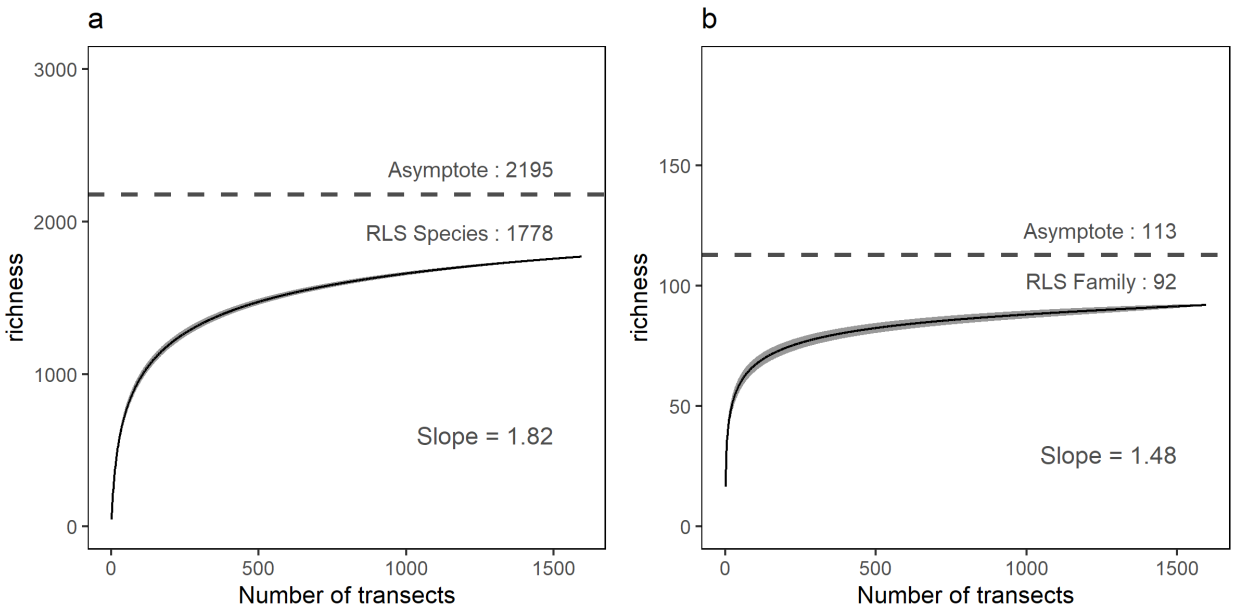


Figure S12. Accumulation curve for (a) species and (b) families in RLS transects, *after a random subset of 169 transects in the 3 regions the most sampled. 169 is the number of transects sampled in the 4th most sampled region.*

3.3. Tableaux supplémentaires

Table S1. Number of reads, MOTUs or assignment to species in the global dataset *after each bioinformatic treatment: 1) without any treatment, 2) after removing sequences with less than 10 reads per sample, and sequences being identified as chimeras, 3) after removing MOTUs found in PCR blanks, 4) after removing MOTUs that do not belong to fish taxa, 5) after removing reads outside the size limits of 30-100bp, 5) after removing MOTUs found in only one PCR in the total dataset, 6) after cleaning with LULU (ie = total MOTUs richness in our study), and 7) number of species detected. As only 16% of 12S rDNA reference barcodes from reef-associated fish species are currently available, only 382 of the 2,023 MOTUs (19%) could be assigned to particular species. Of the remaining MOTUs, 1446 (71%) could be assigned to a particular family, representing 126 families in total.*

Step	Reads	MOTUs	Species
Before	238,322,711	77,065	474
Tenreads	238,120,827	5,595	449
Blanks	238,101,674	5,212	449
Fishonly	199,261,204	3,900	442
Readlength	199,258,587	3,891	442
PCR_all	189,436,754	2,375	382
LULU	189,350,273	2,023	382
LULU_family	157,425,418	1,446	382

Table S2. Number of reads, MOTUs or assignment to species in each region after each bioinformatic treatment*: 1) without any treatment, 2) after removing sequences with less than 10 reads per sample, and sequences being identified as chimeras, 3) after removing MOTUs found in PCR blanks, 4) after removing MOTUs that do not belong to fish taxa, 5) after removing reads outside the size limits of 30-100bp, 5) after removing MOTUs found in only one PCR in the total dataset, 6) after cleaning with LULU, and 7) number of species detected*

Region	Step	Reads	MOTUs	Species
Southeast_Polynesia	before	7,450,199	2,308	86
	tenreads	7,445,477	370	75
	PCR_blanks_chimeras	7,445,443	367	75
	fishonly	7,132,341	306	74
	readlength	7,132,341	306	74
	PCR_all	6,806,780	195	61
	LULU	6,801,947	186	61
	LULU_family	6,174,191	153	61
Tropical_Northwestern_Atlantic	before	27,106,054	7,952	102
	tenreads	27,078,315	827	79
	PCR_blanks_chimeras	27,073,034	785	79
	fishonly	24,495,949	634	76
	readlength	24,495,509	633	76
	PCR_all	22,881,429	402	65
	LULU	22,866,586	361	65
	LULU_family	17,871,554	228	65
Tropical_Southwestern_Pacific	before	36,064,726	10,614	218
	tenreads	36,039,813	1,370	214
	PCR_blanks_chimeras	36,039,240	1,352	214
	fishonly	32,637,229	1,181	211
	readlength	32,637,157	1,179	211
	PCR_all	31,372,335	873	189
	LULU	31,368,868	843	188
	LULU_family	23,448,388	626	188
Western_Coral_Triangle	before	149,448,618	51,069	293
	tenreads	149,318,822	3,314	279
	PCR_blanks_chimeras	149,306,439	3,022	279
	fishonly	119,045,200	2,097	273
	readlength	119,043,095	2,091	273
	PCR_all	113,385,473	1,210	240
	LULU	113,323,213	1,035	237
	LULU_family	96,458,866	787	237
Western_Indian_Ocean	before	18,253,114	7,011	106
	tenreads	18,238,400	702	104
	PCR_blanks_chimeras	18,237,518	670	104
	fishonly	15,950,485	543	101
	readlength	15,950,485	543	101
	PCR_all	14,990,737	349	86
	LULU	14,989,659	327	86
	LULU_family	13,472,419	264	86

Table S3. Summary of ANOVA on Distance-based Redundancy Analysis models.

		eDNA			Visual Census		
		Df	SS	F	Df	SS	F
Total dbRDA	Region	4	1.02	4.1***	4	1.83	17.7***
	Richness	1	0.35	5.77***	1	0.16	6.28**
	Region*Richness	3	0.25	1.99	1	0.21	2**
	Residuals	17	1.05		58	1.49	
Partial dbRDA with Regions	Region	4	0.74	2.79***	4	1.74	15.7**
	Residuals	20	1.32		62	1.71	
Partial dbRDA with Richness	Richness	1	0.35	5.38***	1	0.16	5.9**
	Residuals	20	1.32		62	1.71	

Table S4. eDNA MOTUs and RLS species diversity partitioning across scales*, with 2 different sites definition : groups of stations distinct from 10km and groups of stations distinct from 20km.*

Site definition	Beta scale	eDNA MOTUs	RLS species
Site = Groups of station distinct from 10km	$\beta_{inter-region}$	73.9%	84.5%
	$\overline{\beta}_{inter-site}$	16.6%	10.2%
	$\overline{\beta}_{inter-station}$	4%	2.2%
	$\overline{\alpha}_{station}$	5.5%	3.1%
Site = Groups of station distinct from 20km	$\beta_{inter-region}$	73.9%	84.6%
	$\overline{\beta}_{inter-site}$	15.6%	9.5%
	$\overline{\beta}_{inter-station}$	5.5%	2.9%
	$\overline{\alpha}_{station}$	5%	3%

Table S5. Partitioning of different subsets across spatial scales. *Partitioning for all MOTUs and Visual Census species, and for MOTUs assigned to crypto-benthic families, pelagic families and to species level.*

	γ_{global}	$\beta_{inter-region}$	$\overline{\beta}_{inter-site}$	$\overline{\beta}_{inter-station}$	$\overline{\alpha}_{station}$
All MOTUs	2023	73.7%	14.8%	5.9%	5.7%
Crypto-benthic MOTUs	404	76.7%	14.3%	4.8%	4.2%
Pelagic MOTUs	158	73%	15%	6%	6%
Demersal MOTUs	1461	73%	14.9%	6.1%	6%
eDNA species	396	67.4%	15.6%	8.2%	8.8%
Visual census Species	1818	84.6%	8.9%	3.7%	2.8%

Table S6. Fit of the three global abundance distribution models on fish species observed in visual census and fish MOTUs detected with eDNA. *For each model, parameter values (standard deviation) are provided (intercept, slope, bending) along with the degree of freedom (df) and the Akaike's information criterion (AIC).*

		Visual census (Species)				eDNA (MOTUs)			
Model	**df**	**AIC**	**Intercept**	**Slope**	**Bending**	**AIC**	**Intercept**	**Slope**	**Bending**
Log series	3	986	295 (0.01)	-1	0.001 (2.10^{-4})	246	792 (1.10^{-4})	-1	0.06 (0.003)
Pareto	3	991	267 (11)	-0.95 (0.01)	0	286	2213 (9.10^{-5})	-1.79 (0.03)	0
Pareto Bended	4	1038	173 (0.007)	-0.85 (0.01)	0.005 (4.10^{-4})	242	623 (2.10^{-4})	-0.76 (0.05)	0.08 (0.008)

Table S7. Environmental DNA sampling information in each of the five regions. *Dates of sampling, number of sites per region, number of stations per region, number of filters (samples) per region, the sampling method used, the volume filtered per sample, and the total volume filtered in the region.*

Region	**Date**	**Nb sites**	**Nb stations**	**Nb filters**	**Method**	**Volume per sample**	**Total Volume filtered**
Western Coral Triangle	2017	9	32	64	DNA-free plastic bags and Sterivex filters	2L	128L
Tropical Northwestern Atlantic	2018	6	30	67	Surface filtration along transect and VigiDNA 0.2 filters	30L	2010L
Western Indian Ocean	2019	4	16	31	Surface filtration along transect and VigiDNA 0.2 filters	30L	930L
Southeast Polynesia	2018	1	4	12	Surface filtration along transect & DNA-free plastic bags and VigiDNA 0.2 filters	30L	330L
Tropical Southwestern Pacific	2019	6	18	52	Bottom filtration along transects and VigiDNA 0.2 filters	32L	1664L

Table S8. Number and percentage of MOTUs assigned to their taxa, per class of percentage of similarity with the reference sequence.

Percentage_identity	Number_of_MOTUs	Percentage_of_MOTUs
85-87%	104	7.2
87-89%	140	9.7
89-91%	169	11.7
91-93%	109	7.5
93-95%	96	6.6
95-97%	227	15.7
97-99%	148	10.2
>99%	453	31.3

3.4. Analyses supplémentaires

In order to verify that our unbalanced eDNA sampling did not bias our results and patterns, we ran the analyses after performing two types of rarefaction:

- We randomly sampled 4 stations in all the other sites (random sampling repeated 50 times), as there is only 4 stations in the site in Southeast Polynesia, the lowest sampled region
- We removed the lowest sampled region, Southeast Polynesia,from our dataset and we rarefied the dataset according to the second least sampled regions (Western Indian, 4 sites) by randomly sampling 4 sites in all other regions (random sampling repeated 50 times)

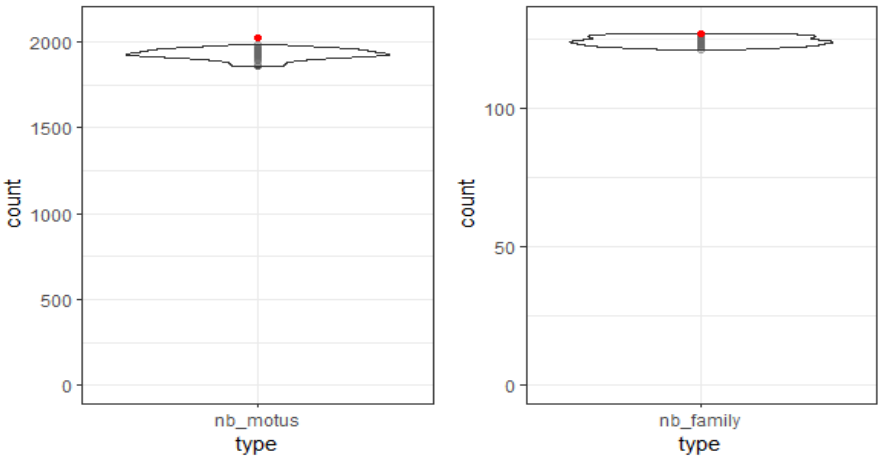


***Fig. 1** Richness of MOTUs (left) and families (right) in the 50 datasets rarefied by sites. Red dots are the richness in the original dataset.*

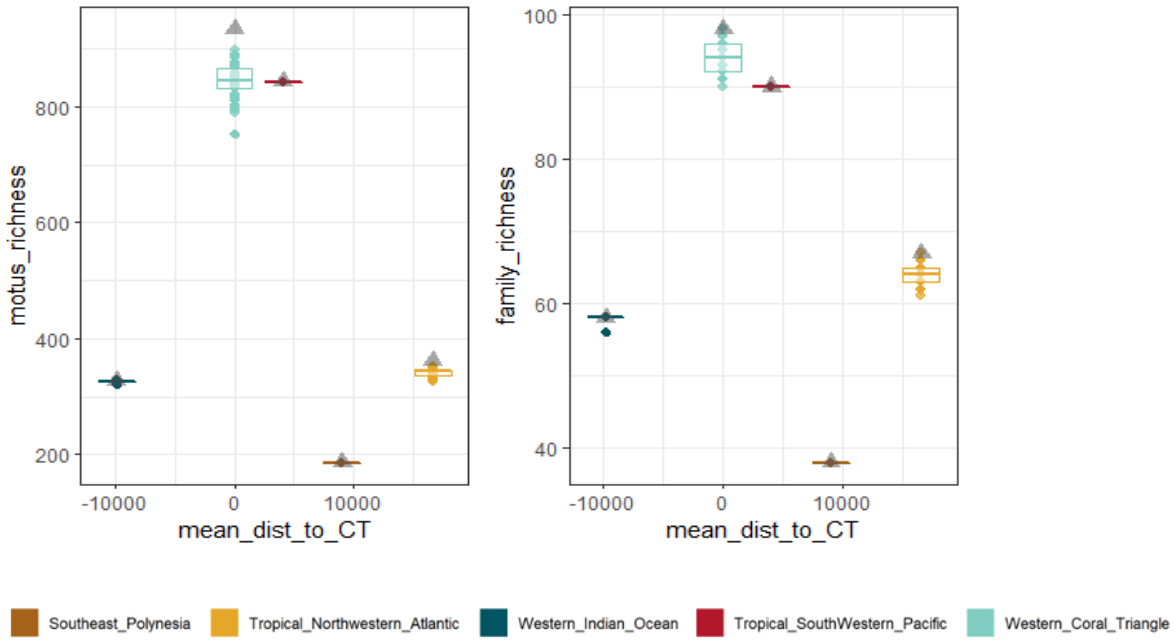


Fig. 2 *Regional MOTUs richness (left) and family richness (right). Boxplots represent the distribution across the 50 rarefied datasets, sampled randomly. Gray triangles represent regional richness in the original dataset.*

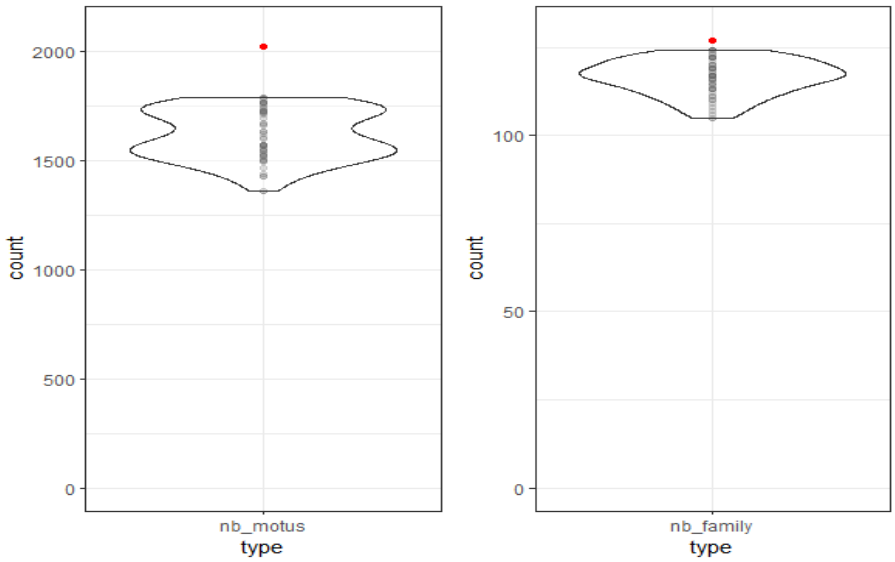


Fig. 3 *Richness of MOTUs (left) and families (right) in the 50 datasets rarefied by regions. Red dots are the richness in the original dataset.*

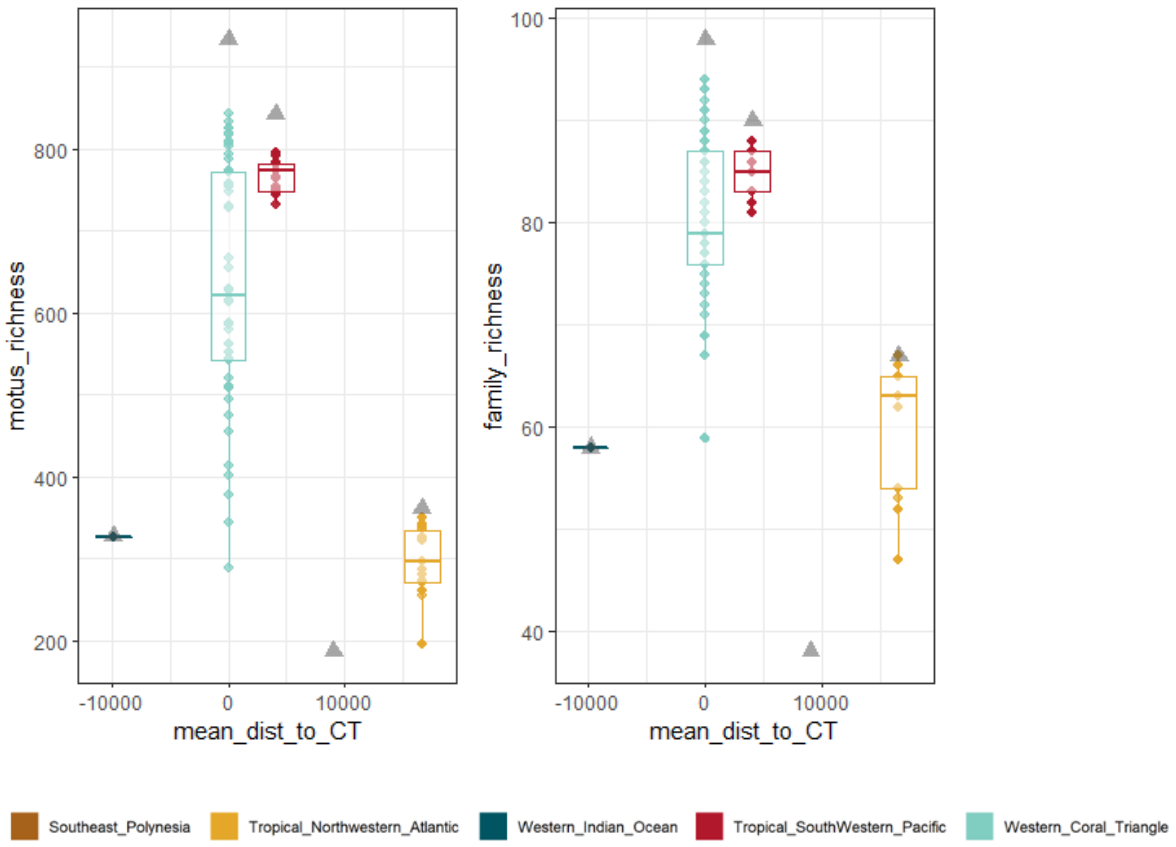


Fig. 4 *Regional MOTUs richness (left) and family richness (right). Boxplots represent the distribution across the 50 rarefied datasets, sampled randomly. Gray triangles represent regional richness in the original dataset.*

Table 1. *Diversity partitioning across scales calculated for the two types of rarefactions.*

component	Full dataset	Rarefaction 4 stations per site (%)	Rarefaction 4 sites per region(%)
$\bar{\alpha}_{station}$	5.7 %	6 ±0.08	7.16 ±0.39
$\bar{\beta}_{inter-station}$	5.9 %	5.77 ±0.06	7.69 ±0.45
$\bar{\beta}_{inter-site}$	14.8 %	14.84 ±0.07	16.41 ±0.51
$\beta_{inter-region}$	73.7 %	73.67 ±0.09	68.73 ±0.51

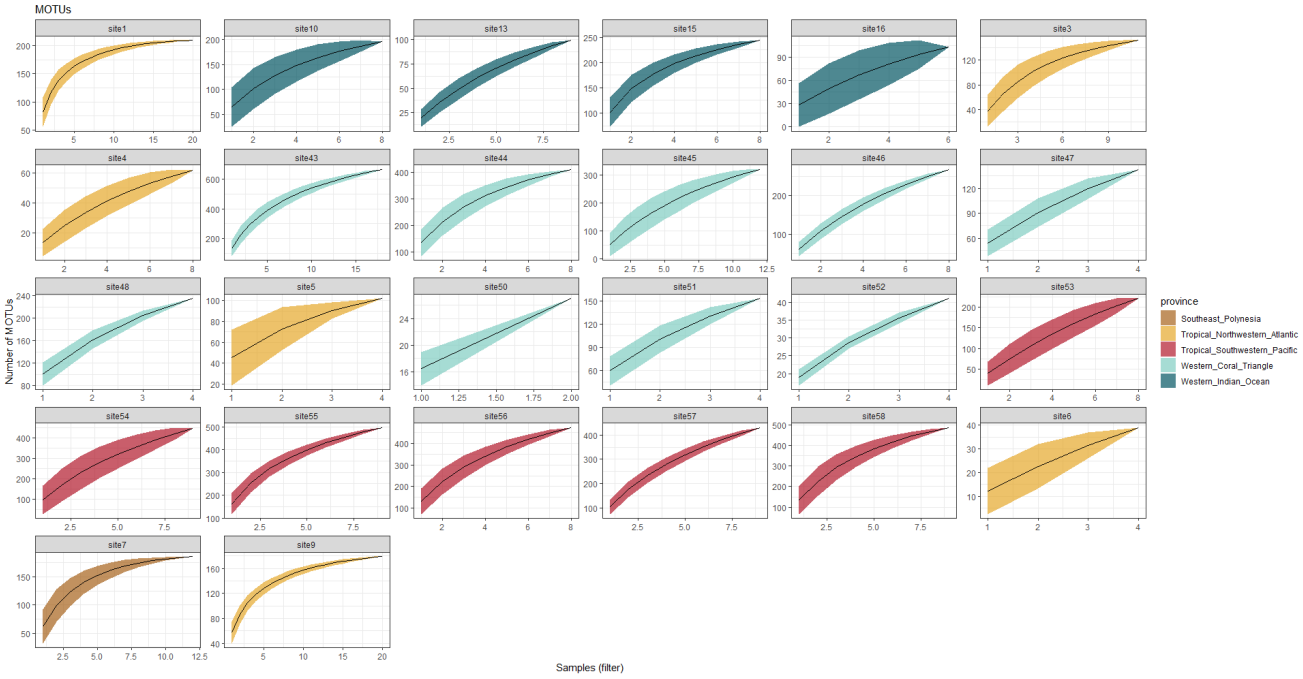


Fig. 5 *Accumulation curve of molecular operational taxonomic units from eDNA at the site level (colors indicate the region of each site). Species accumulation model is fitted according to Lomolino method.*

A)

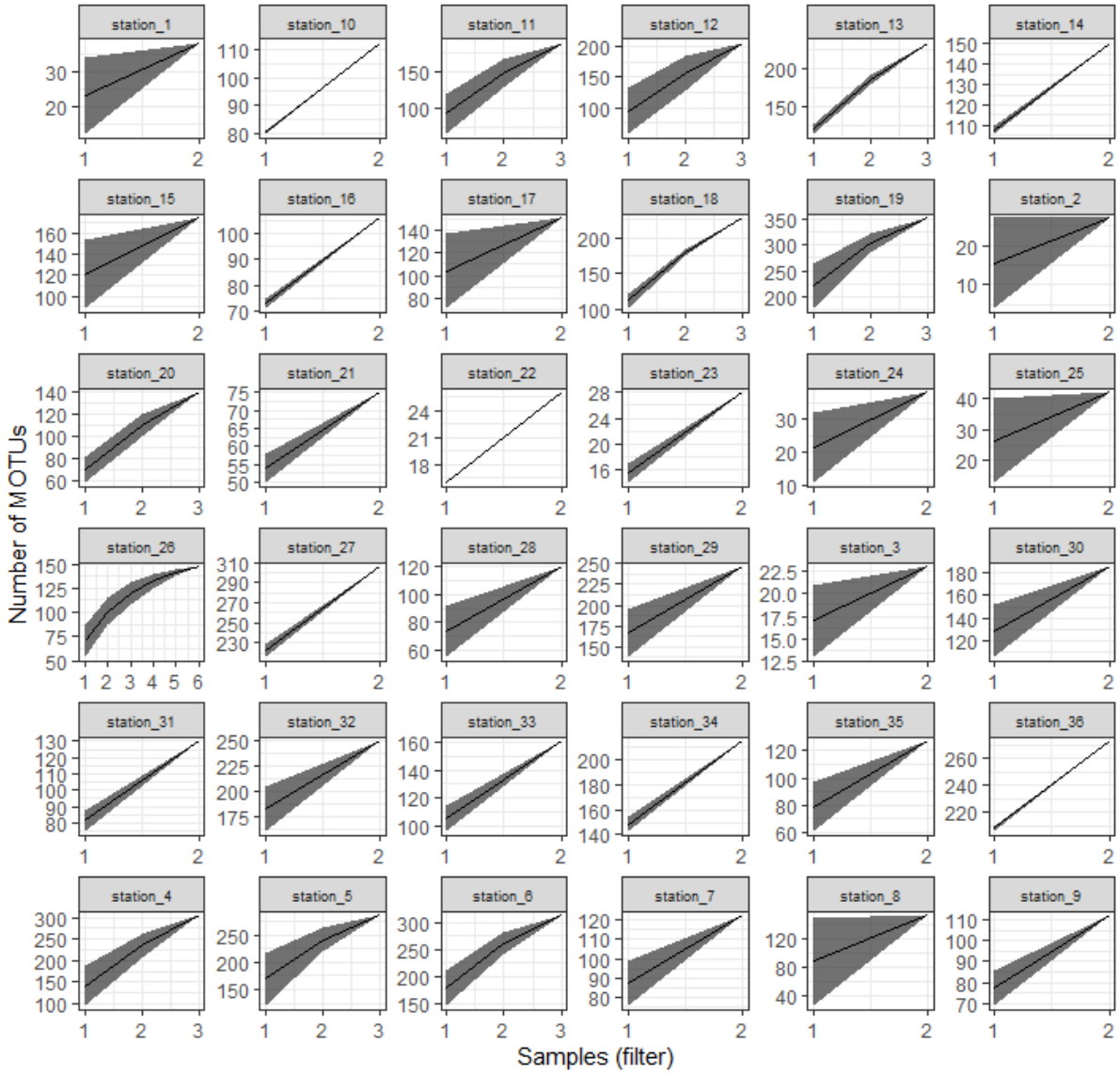
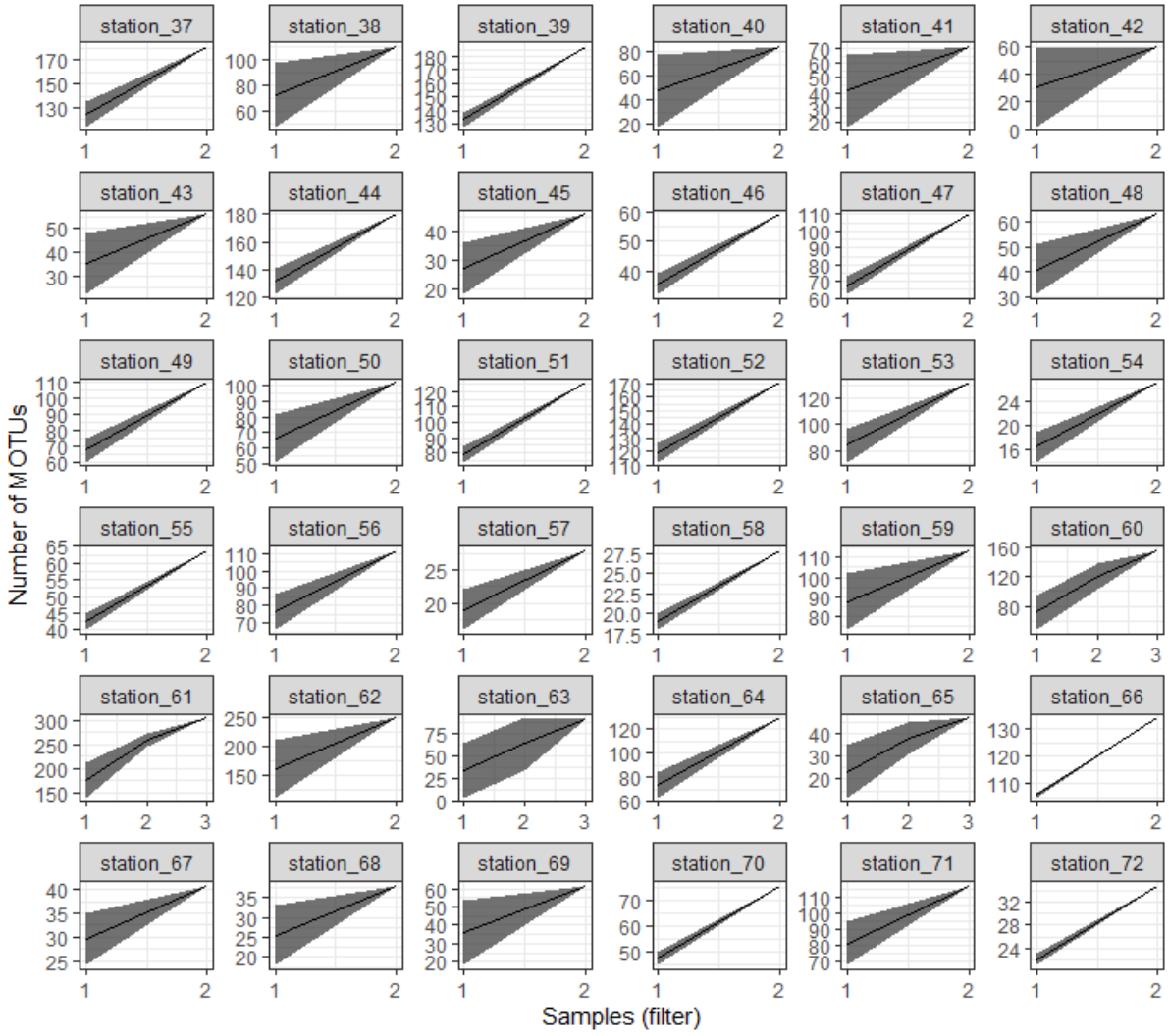
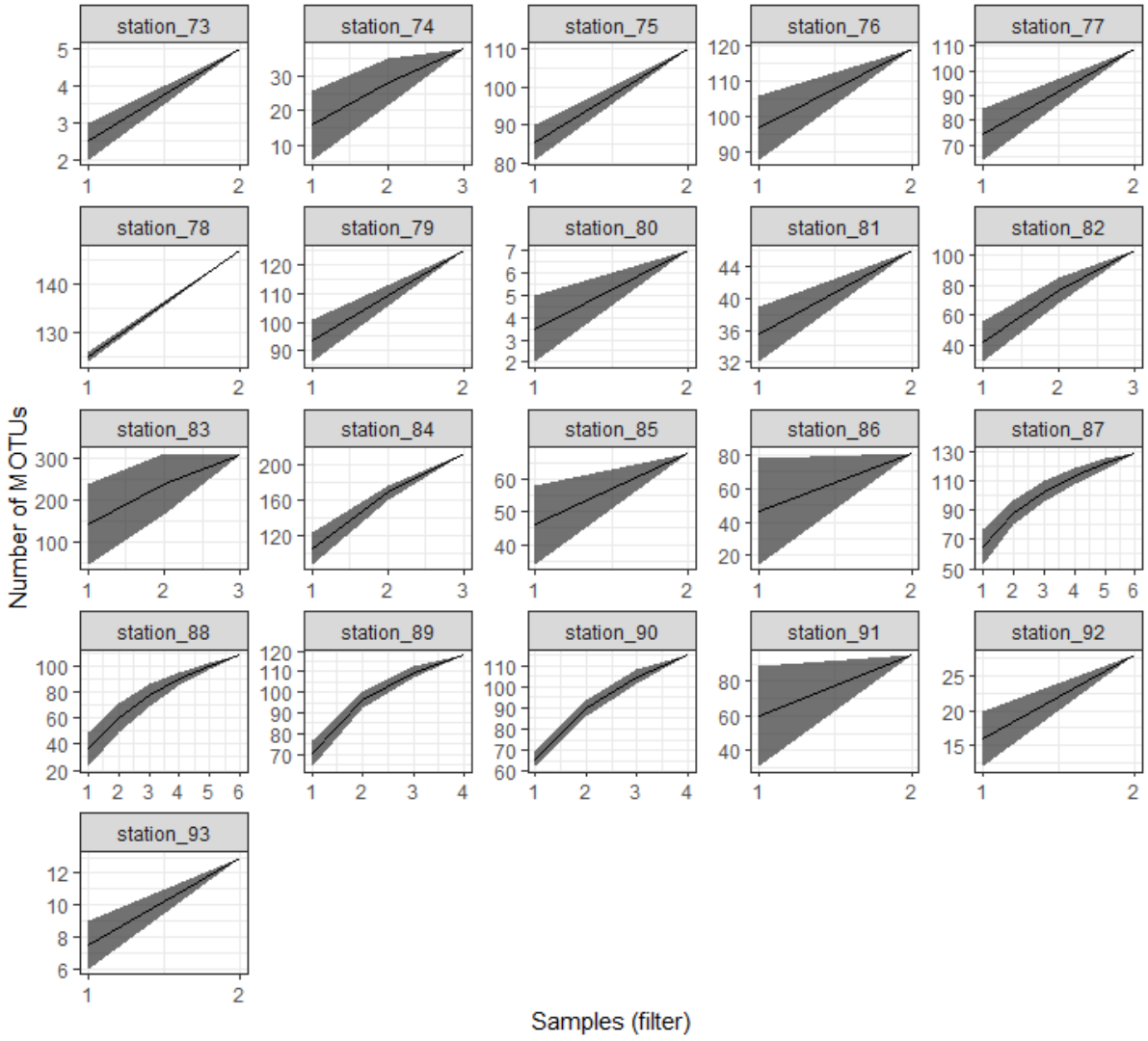


Fig. 6 *Accumulation curve of molecular operational taxonomic units from eDNA at the station level. Species accumulation model is fitted according to Lomolino method. A) stations 1 to 36, B) stations 37 to 72, C) stations 73 to 93*

B)


station_37
station_38
station_39
station_40
station_41
station_42
station_43
station_44
station_45
station_46
station_47
station_48
station_49
station_50
station_51
station_52
station_53
station_54
station_55
station_56
station_57
station_58
station_59
station_60
station_61
station_62
station_63
station_64
station_65
station_66
station_67
station_68
station_69
station_70
station_71
station_72
Number of MOTUs
Samples (filter)

C)


station_73
station_74
station_75
station_76
station_77
station_78
station_79
station_80
station_81
station_82
station_83
station_84
station_85
station_86
station_87
station_88
station_89
station_90
station_91
station_92
station_93
Number of MOTUs
Samples (filter)

4. Suppléments du manuscrit C : *The global distribution of environmental DNA sequences from coastal fishes in the Anthropocene*

4.1. Méthodes supplémentaires

Method S1. Environmental DNA collection

Environmental DNA (eDNA) samples of seawater were collected at 263 stations, in 68 sites, across 11 marine regions covering the global ocean from pole to pole (2 polar, 3 temperate and 6 tropical regions, Fig. 1). Between 1 and 4 replicates were sampled at each station, so the total number of samples is 584. Only samples filtered between 0 and 40m deep were considered in this study. Four different sampling methods were used: (*i*) collection of 2 L of water in DNA-free sterile plastic bags on the surface water from a small boat as well as close circuit rebreather diving (depths between 10 – 40 m) as close as possible to the habitat (Juhel et al., 2020); (*ii*) collection of 1L of water in sterilized bottle, from the surface; (*iii*) 2-km long filtration transect with two replicates (one on each side of a boat at each station for 30 min), for a total of 30L ± 15% of water just under the surface; (*iv*) 2 km-long filtration of water along a transect, approximately 5 m above the substrate, using a long pipe, from the boat. For each sample collected with the two first sampling protocols, the seawater was filtered with sterile Sterivex filter capsules (Merck© Millipore; pore size 0.22µm) using disposable sterile syringes. The eDNA filtration device for the other two sampling protocols was composed of an Athena® peristaltic pump (Proactive Environmental Products LLC, Bradenton, Florida, USA; nominal flow of 1.0 L.min ± 15%), a VigiDNA® 0.2µM cross flow filtration capsule (SPYGEN, le Bourget du Lac, France) and disposable sterile tubing for each filtration capsule. At the end of each filtration, the filter units were filled with CL1 Conservation buffer (SPYGEN, le Bourget du Lac, France) and stored in 50 mL screw-cap tubes at room temperature. The water inside the capsules were emptied, and the capsules were filled with 80 mL of CL1 Conservation buffer (SPYGEN, le Bourget du Lac, France) and stored at room temperature. Details on which sampling method was used in each region are provided in *Supporting Information* Table S1. For each sampling campaign, a strict contamination control protocol was followed in both field and laboratory stages (Valentini et al., 2016), and each water sample processing included the use of disposable gloves and single-use filtration equipment. Negative field controls were performed in multiple sites across all sampling locations and revealed no contamination from the boat or samplers. The list of permits delivered for eDNA sampling can be found in Supporting Information Table S2.

Method S2. eDNA extraction, amplification and sequencing

eDNA extraction was performed in dedicated DNA laboratories. Decontamination procedures were conducted before and after all manipulations. Environmental DNA extractions were performed following the protocols in (Pont et al., 2018) for SPYGEN capsules, and in (Juhel et al., 2020) for the sterivex filters. A teleost-specific 12S mitochondrial rRNA primer pair (teleo,

forward primer – ACACCGCCCGTCACTCT, reverse primer – CTTCCGGTACACTTACCATG (Valentini et al., 2016)) was used for the amplification of metabarcode sequences. As we analyzed our data using MOTUs as a proxy for species to overcome genetic database limitations, we chose to amplify only one marker. The teleo barcode has been shown to be one of the most appropriate for fishes, owing to its high interspecific variability and its short size allowing the detection of rare and degraded DNA reliably (Collins et al., 2019; Kumar, Reaume, Farrell, & Gaither, 2022; Polanco et al., 2021; Zhang, Zhao, & Yao, 2020), even though it presents some amplification bias (Bylemans, Gleeson, Hardy, & Furlan, 2018). The primers were 5' labeled with a unique eight-nucleotide tag (with at least three differences between tags) allowing the assignment of sequences to the respective samples during the sequence analysis. Tags for forward and reverse primers were identical for each sample. Twelve DNA amplifications PCR per sample were performed in a final volume of 25 µL, using 3 µL of DNA extract as the template (Pont et al., 2018). 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 45 libraries were prepared using MetaFast protocol for Illumina sequencing platforms. A paired-end sequencing (2x125 bp) was carried out using an Illumina HiSeq 2500 sequencer with the HiSeq Rapid Flow Cell v2 using the HiSeq Rapid SBS Kit v2 (Illumina, San Diego, CA, USA) or a MiSeq (2x125 bp, Illumina, San Diego, CA, USA) using the MiSeq Flow Cell Kit v3 (Illumina, San Diego, CA, USA) or a NextSeq sequencer (2x125 bp, Illumina, San Diego, CA, USA) with the NextSeq Mid kit following the manufacturer's instructions. This generated an average of 624,468 sequence reads (paired-end Illumina or Ion Torrent) per sample. Samples from the Cold_Temperate_Northwest_Pacific region were filtered with 0.45µm pore size and then treated at the State Key Laboratory of Pollution Control & Resource Reuse in Nanjing University, China. DNA was extracted with the DNeasy Blood & Tissue Kit (Qiagen, Germany), with 3 negative controls, and the teleo fragment was amplified with one PCR replicate per sample. The library was prepared with a VAHTS® Universal DNA Library Prep Kit for Ion Torrent (Vazyme, China) and sequenced on Ion Torrent S5 sequencer (Life Technologies, USA).

Method S3. Bioinformatic analysis

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) (Marques, Guérin, et al., 2020). First, reads were assembled using vsearch (Rognes, Flouri, Nichols, Quince, & Mahé, 2016), then demultiplexed and trimmed using CUTADAPT (Martin., 1994) and clustering was performed using SWARM v.2 (Mahé, Rognes, Quince, de Vargas, & Dunthorn, 2015) with d = 1, which corresponds to a maximum of 1

mismatch between neighboring pairs of sequences within each cluster. The iterative process of SWARM leads to clusters composed of many sequences with more than d mismatches. Further, we used the -f (fastidious) option, which creates virtual sequences within clusters to link more dissimilar sequences together, hence limiting alpha-diversity inflation by joining low abundant MOTUs within larger ones. The minimum distance between clusters is 2 mismatches (d+1). Taxonomic assignment of MOTUs was carried out using the Lower Common Ancestor (LCA) algorithm ecotag implemented in the Obitools toolkit (Boyer et al., 2016) and the European Nucleotide Archive (ENA) as a reference database (release 143, March 2020), supplemented by our custom reference database, containing approximately 800 sequences. It assigns a taxonomy to sequences even when the sequence match is not perfect. The assignment was based on NCBI taxonomic tree of species to consider the current knowledge on molecular diversity per branch and assign a taxonomy at the lowest possible rank. If the sequence matches several identifications with equal percentages of similarity, ecotag assigns to the upper taxonomic level common between all possible matches. We then applied quality filters to be conservative in our estimates. We discarded all observations with less than 10 reads, and present in only one PCR replicate to avoid spurious MOTUs originating from a PCR error. Then, errors generated by index-hopping (MacConaill et al., 2018) were filtered using a threshold empirically determined per sequencing batch using experimental blanks (Taberlet, Bonin, Coissac, & Zinger, 2018). Tag-jumps (Schnell, Bohmann, & Gilbert, 2015) were corrected by removing sequences with unmatching tags on the forward and reverse primers, and tolerating zero mismatch on tag sequences. An additional threshold removing all sequences with a frequency of occurrence <0.001 per MOTU and per library was implemented to clear all reads from the blanks. 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, with an identity threshold of 84% (Marques, Guérin, et al., 2020). The goal is to reduce the bias induced by intra-specific variability and the potential over-estimation of alpha-diversity due to d=1 in SWARM. Taxonomic assignments obtained from the LCA algorithm were further selected to ensure more conservative assignments, following results from previous studies (Juhel et al., 2020; Polanco et al., 2021). Assignments were accepted at the species level, as putative species, if the percentage of similarity with the reference sequence was 100%, at the genus level if the similarity was between 90 and 99%, and at the family level if the similarity was > 85%. If these criteria were not met, the MOTU was left unassigned. The post-LCA algorithm correction thresholds of 85% similarity for family and 90% for genus assignments were chosen to include a maximum of correct family and genus assignments while minimizing the risk of adding wrong assignments in the detections. Number of reads, MOTUs and species after each cleaning step are available in Supporting Information Tables S3-4.

Method S4. Explanatory factors

Environmental factors included sea surface temperature (mean SST), degree heating weeks (mean DHW), pH, net primary productivity (mean NPP), and salinity (mean SSS) that use a variety of satellite and in-situ observations, optimal interpolations and ocean system models (as documented in Supporting Information Table S5). We extracted environmental factors at their native spatial resolutions, over one year prior to the date of sampling and calculated the mean. In this way, we ensured that the highest spatial resolution of data was used in combination with temporally relevant metrics that describe the recent environmental conditions prior to sampling. Socioeconomic factors included an index of marine ecosystem dependence in the sovereign country, the Human development Index of the sovereign country in 2019 (HDI), and an index of the human impact gravity. HDI is a synthetic measure capturing elements of life expectancy, education and wealth. We used HDI values for 2019 from the Human Development Indicators and Indices (http://hdr.undp.org). Marine Ecosystem Dependence quantifies nutritional, economic, and coastal protection dependence on marine ecosystems at the country scale (Selig et al., 2019). We calculated gravity of a sampling station as the human population size divided by the travel time between the station and this population center (in minutes). Total gravity is the sum of gravities in a buffer of 500 km around a station (Cinner et al., 2018). We extracted the population size from the UN WPP-adjusted population count v4.11 for year 2020 at 30 arc-second resolution (https://sedac.ciesin.columbia.edu). Geographic factors included the bathymetry, the depth of sampling, the distance to shore and the distance to the Coral Triangle. The Coral Triangle hosts the highest fish diversity due to the development of complex reef habitats in the Miocene and the persistence of these habitats during the Quaternary climate change periods (Cowman & Bellwood, 2013; Pellissier et al., 2014). The distance to this refugia has been demonstrated to shape the traits structure and family richness in reef fishes (Parravicini et al., 2021), and can thus explain the variation of alpha and beta diversity across oceans. We estimated the bathymetry from different sources, directly on site with a sounder, or extracted from GEBCO_2020 Esri ASCII raster on a 15 arc-second interval grid, so approximatively 500m (www.gebco.net). The depth of sampling was measured on site. We calculated the distance to the Coral Triangle as the geographic distance from the sampling point to the center of the Coral Triangle (longitude = 133.679826, latitude = -1.307436), using the function pointDistance from the “raster” package. We computed the distance to shore as the minimum distance between the sampling point and all shoreline points, using the function gDistance from the “rgeos” package. Sampling factors considered included the sample method (transect or point), and the total volume filtered per station**.**

Method S5. Statistical analyses

All statistical analyses are run at the station level, pooling reads from samples and PCR replicates. All analyses were run in R version 4.1.1.

- MOTU diversity

Fish MOTU diversity, expressed as the number of distinct MOTUs, was calculated at each station, as well as the MOTU diversity of fish from large fish families (n = 479 MOTUs) and cryptobenthic families (n = 539 MOTUs). MOTU diversities were log-transformed. The selection of cryptobenthic MOTUs was made according to the definition of cryptobenthic families (Brandl, Goatley, Bellwood, & Tornabene, 2018), so families characterized by the high prevalence (> 10%) of small-bodied species (< 50 mm). To select the large fish MOTUs, we extracted the length of all fish species from FishBase, computed the mean and 5th and 95th quantiles for each family and order, and selected species belonging to families and orders with a 5th quantile superior to 20 cm. MOTU α-diversity corresponds to numbers of fish MOTUs per station, and is independent of the taxonomic assignment which was only used to select the MOTUs belonging to cryptobenthic and large fish families.

- Sequence α-diversity

To compute the sequence α-diversity for each station, we first computed the genetic distances between each pair of sequences with the function dist.gene from package "ape". We then applied the unifier framework based on generalizations of Hill number to measure sequence diversity. Hill numbers provide a set of diversity indices, differing by a parameter "q", which determines their sensitivity to relative abundances, and τ, their sensitivity to distance between sequences. Hill numbers have been recommended to produce reliable diversity assessments from molecularly characterized samples (Alberdi & Gilbert, 2019; Mächler, Walser, & Altermatt, 2021). We used the function alpha.fd.hill from package "mFD" (Magneville et al., 2022), with parameters q = 0, which gives equal weight to all sequences, and τ as equal to the mean genetic distance (Chao et al., 2019).

- Modeling MOTU and sequence α-diversity

We investigated the relationship between fish MOTU and sequence diversity at each station and all explanatory factors with a generalized least square model (GLS) that considers the spatial autocorrelation between samples. The model with Gaussian spatial correlation (function corGaus) had the lowest Akaike Information Criterion (AIC) compared to other correlation functions. A variance inflation factor (VIF) approach was used to identify and remove residual collinear factors (factors with VIF > 10). We tested for spatial autocorrelation in the model residuals using the Moran's index I. The adjusted R^2 of the model was computed with the function r2 from the package "performance". Standardized effect sizes of each explanatory factor were extracted with the function effectsize from the "effectsize" package. Partial R^2 for

each group of factors were obtained with the function calc.relimp from the package "relaimpo". Partial relationships between response variables and each explanatory factor while controlling for all the other factors were visualized with the function visreg from the "visreg" package. The same procedures were repeated for cryptobenthic and large fish MOTU α-diversity within stations. Sensitivity analyses were performed on 10 subset datasets after randomly removing 20% of the stations, to assess the robustness of our models, and after removing samples from polar regions (Scotia Sea and Arctic), to control for the influence of these extreme regions.

- Modeling β-diversity

The Jaccard dissimilarity index was computed between stations using fish MOTU composition (presence/absence) with the function vegdist from package "vegan". Similarly, we computed the dissimilarity in sequence β-diversity between each pair of stations using the Hill number framework. The sequence β-diversity was calculated with the function beta.fd.hill from the "mFD" package, with parameter q = 0 and tau = "mean". We then performed a distance-based redundancy analysis (dbRDA) on the sequence β-diversity and MOTU β-diversity matrices. To account for spatial autocorrelation in our samples, we first computed distance-based Moran Eigenvectors Maps (dbMEM) with the function dbmem from the "adespatial" package, which returned 15 dbMEM. To select dbMEMs to include in our dbRDA, we ran a dbRDA with the response variable and all dbMEMs as explanatory factors and selected the dbMEMs explaining most of the model's variance (MEM1 to MEM5). We then ran the dbRDA on the full model, with all explanatory factors and selected dbMEMs. Factors with VIF > 10 were removed and a final partial dbRDA was run with all selected explanatory factors, and with sampling factors and dbMEMs as conditional variables. Partial R^2 for each group of factors were obtained with the varpart function of the "vegan" package.

- Functional and phylogenetic diversity

We explored the relationship between the pairwise sequence, phylogenetic and functional distances, by selecting only the MOTUs assigned to the species level in our dataset (n = 787). We computed genetic pairwise distances for these species with the function dist.gene from package "ape". We computed functional Gower distance based on functional traits extracted from fishbase, available for 685 of our species (habitat, substrate, depth range, longevity, vulnerability, length, weight, position in the water column, diet, interaction) using the function compute_dist_matrix from package "funrar". The phylogenetic distance between species was computed using the functions fishtree_phylogeny from the "fishtree" package and cophenetic.phylo from the package "ape", and the phylogeny from (Rabosky et al., 2018). These distance matrices were compared with a mantel test, and by calculating the area under the curve (AUC) criterion, based on Somer's D statistic. AUC varies between 0 (no correlation) and 1 (identical matrices), and is computed with the functions coranking, R_NX and AUC_ln_K from package "coRanking". We applied the α- and β-diversity Hill number

framework for sequence, functional and phylogenetic diversity (q = 0 and τ equal to the mean genetic, functional or phylogenetic distance), using the functions alpha.fd.hill and beta.fd.hill from "mFD" package, for sequence and functional diversities, and function ChaoPD from package "entropart" for phylogenetic diversity. Alpha diversity indices were compared with a Pearson correlation test, and the β-diversity matrices were compared with AUC and Mantel tests.

4.2. Tableaux supplémentaires

Table S1. *Information on sampling in each region*

Region	**Date of sampling**	**Sampling method**	**number of stations**	**mean volume filtered per station (L)**	**Total volume filtered (L)**	**Sequencer**
Arctic	July-August 2020	Transect	19	30	570	Miseq
Cold_Temperate_ Northwest_Pacific	September 2020	Bottle	24	2.8 ±0.3	69	IonTorrent
Lusitanian	October-November 2019 September 2020	Transect	13	60	780	Miseq
Mediterranean_Sea	March-June 2018 March & July 2019 July 2020	Transect	35	70.3 ±23	2460	Miseq - Hiseq
Scotia_Sea	January-February 2020	Transect	41	32.7 ±17.9	1344	NextSeq - Miseq
Southeast_Polynesia	June 2018	Transect & Bag	3	128 ±76	384	Hiseq
Tropical_East_Pacific	March 2018	Transect	13	60	780	Hiseq
Tropical Northwestern Atlantic	February, March & July 2018 January 2020	Transect	41	45 ±12.7	1844	Hiseq - Miseq
Tropical Southwestern Pacific	October-December 2019 August 2020	Benthic transect	26	92.3 ±10.4	2400	NextSeq - Miseq
Western_Coral_Triangle	October-November 2017	Bag	32	3.95±0.1	126.5	Hiseq
Western_Indian_Ocean	April 2019	Transect	16	60	960	Miseq

Table S2. *Permits issued by the relevant authorities in each sampling location requiring governmental authorization.*

REGION	PERMIT DETAIL
Scattered Islands	Data collected by a team of students and researchers from French institutions aboard a French ship belonging to the TAAF fleet, under permit 2019-45 from April 1st 2019, delivered by the Administration of the French Southern and Antarctic Lands.
New-Caledonia	Data collected by students and researchers from French institutions working in New-Caledonia, under permit N° 3066-2019/ARR/DENV delivered by the Southern Province of New-Caledonia, and permit N° 609011/2019/DEPART/JJC delivered by the Northern Province of New-Caledonia.
Indonesia	Fieldwork conducted by students and researchers from a French institution (IRD) and an Indonesian institution (BRIN) within the Lengguru project according to relevant guidelines by the government of the Republic of Indonesia and under research permit issued by RISTEK (Indonesia) (3179/FRP/E5/Dit.KI/IX/2017) and relevant Indonesian government collecting permit.
Arctic	Data collected by a team of student and researchers from French and Swiss institutions, under the research permit n° 20/01465-7 for research application RiS-ID 11544, delivered by the Governor of Svalbard.
Antarctic	Samples collected under a permit (RWS-2019/40813) provided by Rijkswaterstaat Zee en Delta, Government of the Netherlands, as per the Protection of Antarctica Act, to Stichting Greenpeace Council to carry out such research in the Antarctic area.
Colombia (Malpelo, Santa Marta, Providencia)	Samples collected by the INVEMAR, an entity attached to the Ministry of Environment and Sustainable Development. The INVEMAR does not require permit to sample as it is clarified in Paragraph 1, Article 2.2.2.2.8.1.2, Section 1 (Permits), Chapter 8 (Scientific Research) of Decree 1076 of 2015, Sole Regulatory Decree of the Environment and Sustainable Development Sector. "The Ministry of Environment and Sustainable Development, its affiliated entities, the National Natural Parks of Colombia, Regional Autonomous Corporations and/or Sustainable Development Corporations and the Large Urban Centers will not require the Specimen Collection Permit referred to in this decree (...)."

Table S3. *Read, MOTU and species counts after each cleaning and treatment step in the whole dataset. Numbers in parenthesis for LULU steps correspond to reads, MOTUs and species added with the China dataset.*

step	Reads	MOTUs	Species
before	368436373	124922	945
tenreads	368099042	9297	902
blanks	368073065	8792	902
fishonly	317176168	4844	874
readlength	317162303	4817	874
PCR_all	308854365	2983	786
LULU	271694868 (+12014)	2862 (+40)	778 (+15)
LULU_family	245262610 (+11520)	2239 (+37)	778 (+15)

Table S4. *Read, MOTU and species counts after each cleaning and treatment step for each region.*

region	step	Reads	MOTUs	Species
Arctic	before	6346919	2557	21
	tenreads	6340103	327	10
	PCR_blanks_chimeras	6339817	314	10
	fishonly	5059617	57	3
	readlength	5059591	56	3
	PCR_all	5037842	33	2
	LULU	5037842	33	2
	LULU_family	4945738	28	2
Cold Temperate Northwest Pacific	LULU	12014	42	18
	LULU_family	11520	39	18
Lusitanian	before	6320079	2079	60
	tenreads	6314698	206	58
	PCR_blanks_chimeras	6314653	204	58
	fishonly	6160427	110	55
	readlength	6160427	110	55
	PCR_all	6153448	96	52
	LULU	6153448	96	52
	LULU_family	4362220	92	52
Mediterranean Sea	before	69838075	15412	134
	tenreads	69778918	1079	129
	PCR_blanks_chimeras	69776913	1032	129
	fishonly	64681216	394	120
	readlength	64671959	384	120
	PCR_all	64462922	249	100
	LULU	64462922	249	100
	LULU_family	63876182	227	100
Scotia Sea	before	20412024	5326	19
	tenreads	20396150	484	15
	PCR_blanks_chimeras	20396065	480	15
	fishonly	16405013	62	9
	readlength	16404333	61	9
	PCR_all	16041636	40	8

	LULU	16041636	40	8
	LULU_family	11005415	24	8
Southeast Polynesia	before	7442809	2408	118
	tenreads	7438309	373	112
	PCR_blanks_chimeras	7438275	370	112
	fishonly	7179074	307	111
	readlength	7179074	307	111
	PCR_all	6859662	197	83
	LULU	6859662	197	83
	LULU_family	6293496	169	83
Tropical East Pacific	before	4957433	2055	66
	tenreads	4950268	380	65
	PCR_blanks_chimeras	4950242	379	65
	fishonly	4739633	221	62
	readlength	4739588	220	62
	PCR_all	4676658	153	48
	LULU	4676658	153	48
	LULU_family	4401537	126	48
Tropical Northwestern Atlantic	before	32286077	12387	164
	tenreads	32243468	975	135
	PCR_blanks_chimeras	32237820	927	135
	fishonly	29328813	734	132
	readlength	29328137	731	132
	PCR_all	28831276	466	116
	LULU	28825577	452	116
	LULU_family	23452276	314	116
Tropical Southwestern Pacific	before	52564185	20134	352
	tenreads	52523440	2392	335
	PCR_blanks_chimeras	52519753	2341	335
	fishonly	47897730	1328	331
	readlength	47896969	1327	331
	PCR_all	45546471	954	295
	LULU	34292060	917	288
	LULU_family	30246965	742	288
Western Coral Triangle	before	149939309	57015	472
	tenreads	149800467	3693	447
	PCR_blanks_chimeras	149787194	3385	447
	fishonly	119647080	2120	436
	readlength	119644760	2112	436
	PCR_all	115574720	1228	397
	LULU	89675333	1145	390
	LULU_family	82359410	941	390
Western Indian Ocean	before	18329463	8294	162
	tenreads	18313221	713	159
	PCR_blanks_chimeras	18312333	681	159
	fishonly	16077565	550	156
	readlength	16077465	548	156
	PCR_all	15669730	357	133
	LULU	15669730	357	133
	LULU_family	14319371	301	133

Table S5. *Information and sources of environmental variables*

Variable	Source	Unit	Original spatial resolution	Temporal resolution	Temporal duration	Web link	Reference
Temperature	CoralReefWatch v3.1	Celcius	0.05°	Daily	1985-2021	https://coralreefwatch.noaa.gov/product/5km/index.php	Liu, G., et al. (2014)
Degree heating weeks	CoralReefWatch v3.1	Daily temperature 1°C above the maximum monthly mean SST from 1985-1993, over a 12 week period	0.05°	Daily	1985-2021	https://coralreefwatch.noaa.gov/product/5km/index.php	Liu, G.,et al. (2014)
pH	Norwegian Earth System Model forced ocean simulation (NorESM2)	pH	1°	Monthly	1980-2018	https://www.noresm.org/resources/	Norwegian Earth System Model forced ocean simulation (NorESM2)
Net primary productivity	Standard Vertically Generalized Production Model	mgC m^{-2} day^{-1}	0.083°	Monthly	2003-2021	http://orca.science.oregonstate.edu/2160.by.4320.monthly.hdf.vgpm.m.chl.m.sst.php	Behrenfeld, M. J., & Falkowski, P. G. (1997).
Chlorophyll-A	CMEMS GlobColour	Mg m^{-3}	0.041°	Monthly	1997-2021	Copernicus product ID: OCEANCOLOUR_GLO_CHL_L4_REP_OBSERVATIONS_009_082	Garnesson, P., Mangin, A., D'Andon, O. F., Demaria, J., & Bretagnon, M. (2019).
Salinity	Global SSS/SSD L4 Reprocessed dataset	PSU	0.25°	Monthly	1994-2021	https://resources.marine.copernicus.eu/product-detail/MULTIOBS_GLO_PHY_S_SURFACE_MYNRT_015_013	Buongiorno Nardelli, B., 2012

Table S6. *Anova of GLS predicting fish taxonomic α-diversity within stations*

	numDF	F-value	p-value
(Intercept)	1	10.47266	0.0014
mean_DHW_1year	1	9.41162	0.0024
mean_sss_1year	1	4.72381	0.0307
mean_SST_1year	1	109.54371	<.0001
mean_npp_1year	1	3.65093	0.0572
HDI2019	1	0.64457	0.4228
Gravity	1	0.59028	0.4430
MarineEcosystemDependence	1	22.62229	<.0001
dist_to_CT	1	77.42021	<.0001
bathy	1	0.59233	0.4422
depth_sampling	1	4.13457	0.0431
distCoast	1	17.38010	<.0001
volume	1	4.46858	0.0355
sample_method	1	21.37818	<.0001

Table S7. *Anova of GLS predicting sequence α-diversity within stations*

	numDF	F-value	p-value
(Intercept)	1	1.02278	0.3128
mean_DHW_1year	1	0.86087	0.3544
mean_sss_1year	1	2.73074	0.0997
mean_SST_1year	1	139.48278	<.0001
mean_npp_1year	1	31.30070	<.0001
HDI2019	1	1.00210	0.3178
Gravity	1	0.00607	0.9380
MarineEcosystemDependence	1	15.37866	0.0001
dist_to_CT	1	48.79961	<.0001
bathy	1	0.04562	0.8311
depth_sampling	1	17.16315	<.0001
distCoast	1	6.49684	0.0114
volume	1	0.00431	0.9477
sample_method2	1	27.39906	<.0001

Table S8. *Anova of GLS predicting cryptobenthic taxonomic α-diversity within stations*

	numDF	F-value	p-value
(Intercept)	1	2.40720	0.1220
mean_DHW_1year	1	4.45024	0.0359
mean_sss_1year	1	15.95809	0.0001
mean_SST_1year	1	66.31302	<.0001
mean_npp_1year	1	25.78901	<.0001
HDI2019	1	0.01149	0.9147
Gravity	1	2.42389	0.1208
MarineEcosystemDependence	1	52.39797	<.0001
dist_to_CT	1	57.76497	<.0001
bathy	1	0.04206	0.8377
depth_sampling	1	13.17270	0.0003
distCoast	1	2.70875	0.1011
volume	1	0.00693	0.9337
sample_method	1	10.11247	0.0017

Table S9. *Anova of GLS predicting large fish taxonomic diversity within stations*

	numDF	F-value	p-value
(Intercept)	1	2.56717	0.1104
mean_DHW_1year	1	0.05015	0.8230
mean_sss_1year	1	26.65398	<.0001
mean_SST_1year	1	82.81727	<.0001
mean_npp_1year	1	14.31105	0.0002
HDI2019	1	2.01311	0.1572
Gravity	1	0.16557	0.6844
MarineEcosystemDependence	1	14.82563	0.0001
dist_to_CT	1	45.73543	<.0001
bathy	1	0.78522	0.3764
depth_sampling	1	3.06234	0.0814
distCoast	1	9.12800	0.0028
volume	1	0.00010	0.9920
sample_method	1	14.00773	0.0002

Table S10. *Anova of dbRDA predicting taxonomic β-diversity between stations*

	Df	SumOfSqs	F	Pr(>F)	
mean_DHW_1year	1	0.997	2.4709	0.01	**
mean_SST_1year	1	2.754	6.8218	0.01	**
mean_sss_1year	1	1.442	3.5709	0.01	**
mean_npp_1year	1	1.908	4.7268	0.01	**
HDI2019	1	1.494	3.7002	0.01	**
Gravity	1	1.061	2.6271	0.01	**
MarineEcosystemDependence	1	1.741	4.3135	0.01	**
dist_to_CT	1	1.753	4.3430	0.01	**
bathy	1	1.118	2.7697	0.01	**
depth_sampling	1	0.683	1.6924	0.01	**
distCoast	1	0.760	1.8821	0.01	**
Residual	249	100.521			

Table S11. *Anova of dbRDA predicting sequence β-diversity between stations*

	Df	SumOfSqs	F	Pr(>F)	
mean_DHW_1year	1	0.2400	4.0580	0.01	**
mean_SST_1year	1	2.0332	34.3722	0.01	**
mean_sss_1year	1	0.3325	5.6213	0.01	**
mean_npp_1year	1	0.3097	5.2363	0.01	**
HDI2019	1	0.1139	1.9253	0.09	.
Gravity	1	0.0807	1.3646	0.23	
MarineEcosystemDependence	1	0.4128	6.9791	0.02	*
dist_to_CT	1	0.2741	4.6330	0.01	**
bathy	1	0.4450	7.5222	0.01	**
depth_sampling	1	0.0448	0.7578	0.50	
distCoast	1	0.3732	6.3089	0.01	**
Residual	248	14.6698			

4.3. Figures supplémentaires

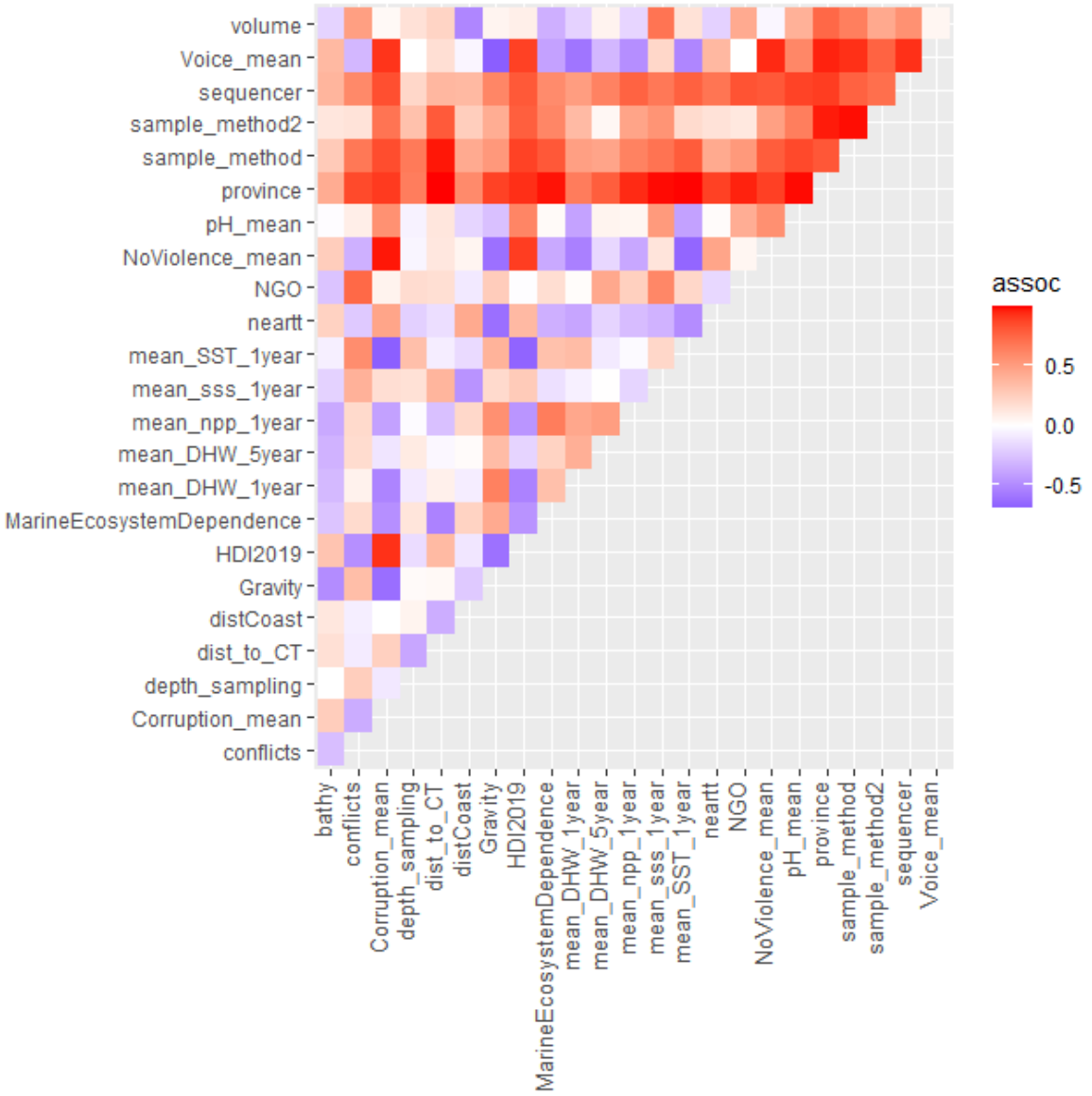


Fig. S1. *Correlation matrix between explanatory variables before the selection of variables correlation inferior to |0.7|.*

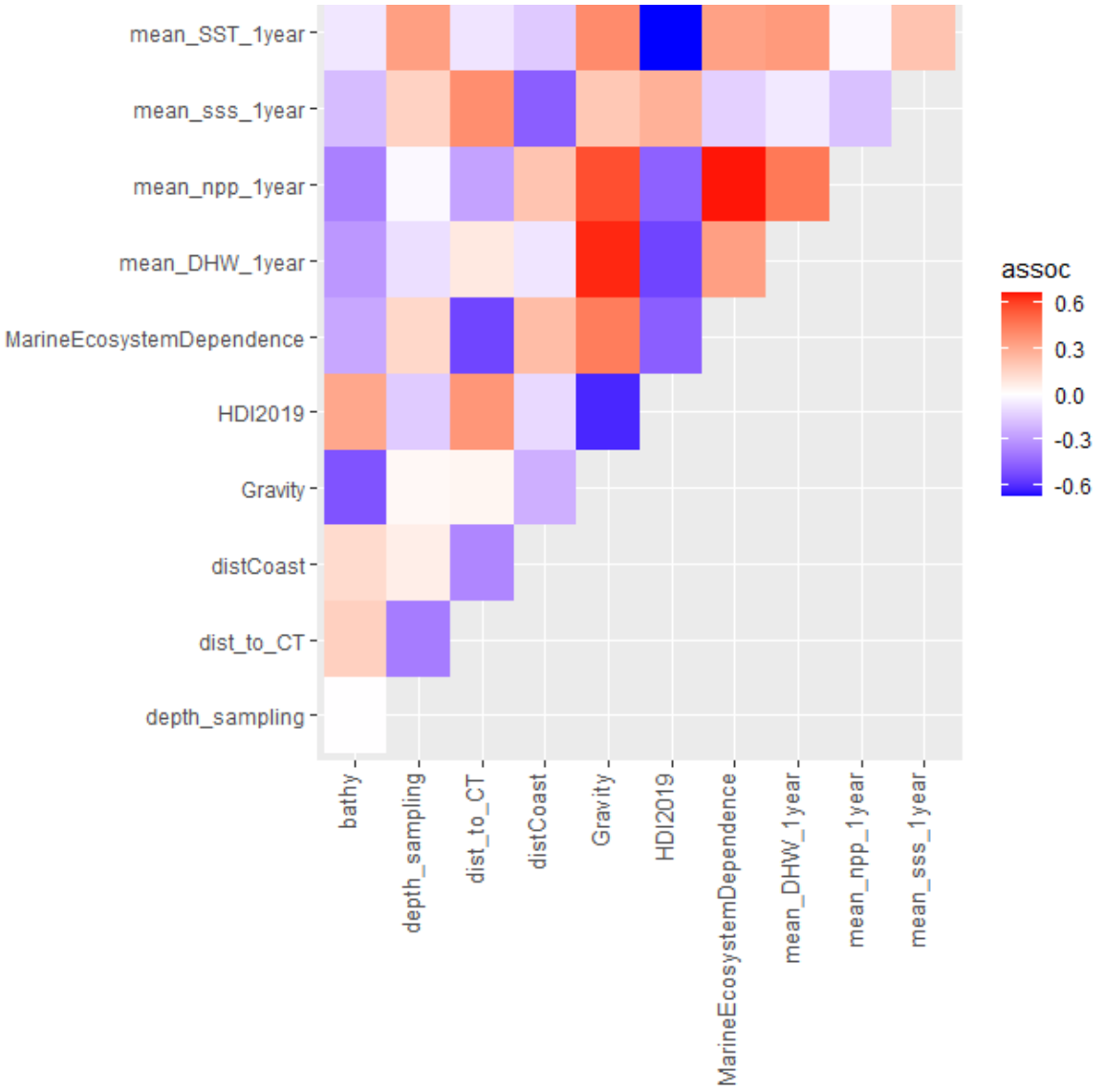


Fig. S2. *Correlation matrix between explanatory variables after the selection of variables correlation inferior to |0.7|.*

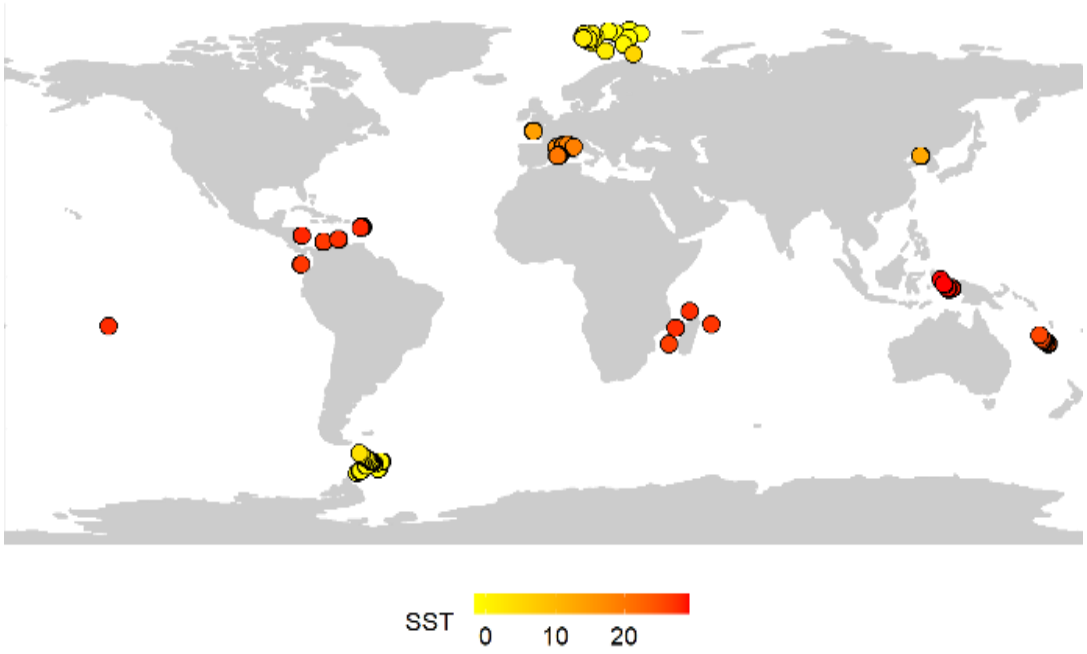


Fig. S3. *Map of sea surface temperature (SST)*

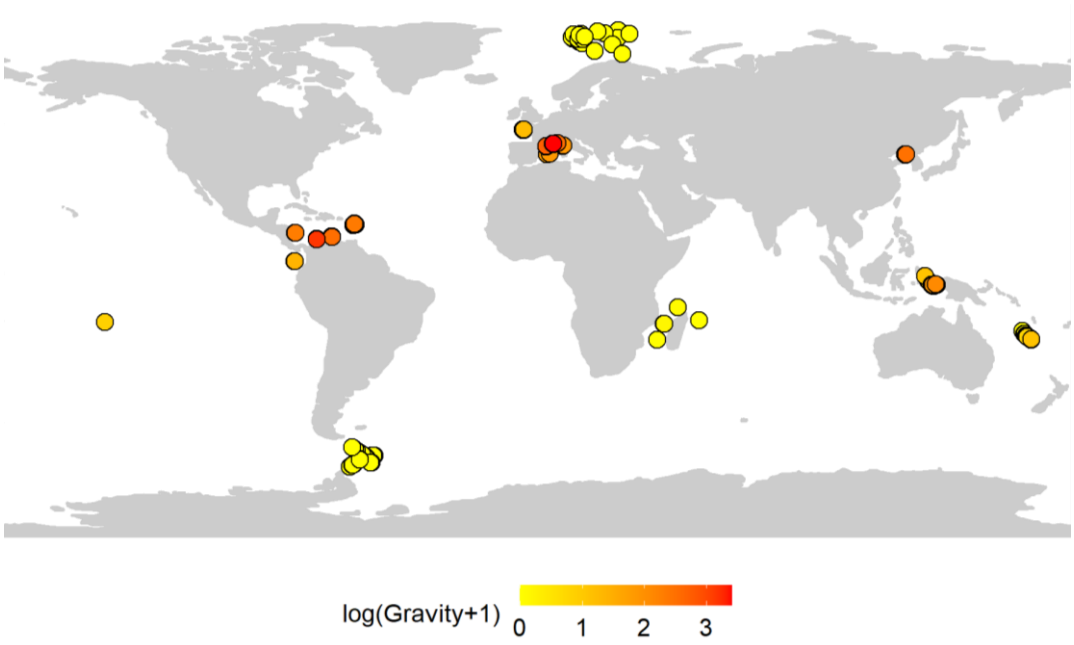


Fig. S4. *Map of gravity per station (log transformed)*

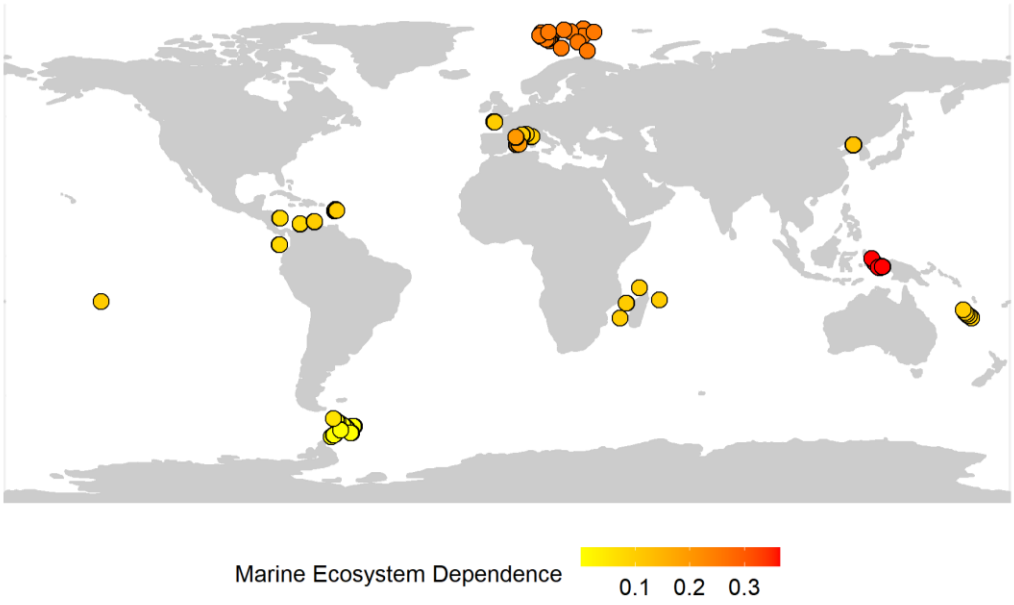


Fig. S5. *Map of marine ecosystem dependence per station*

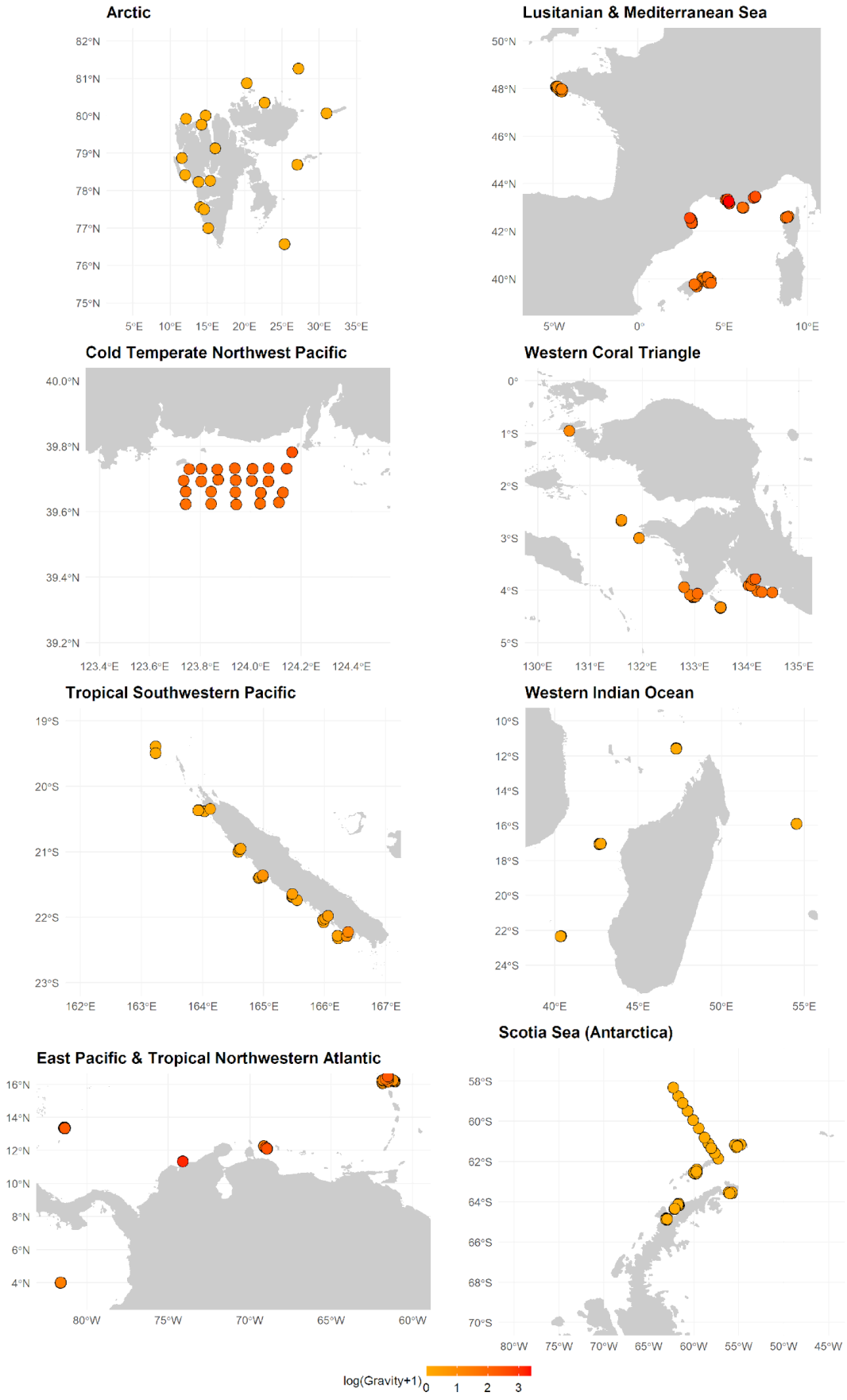


Fig. S6. *Maps of gravity (log transformed) by region*

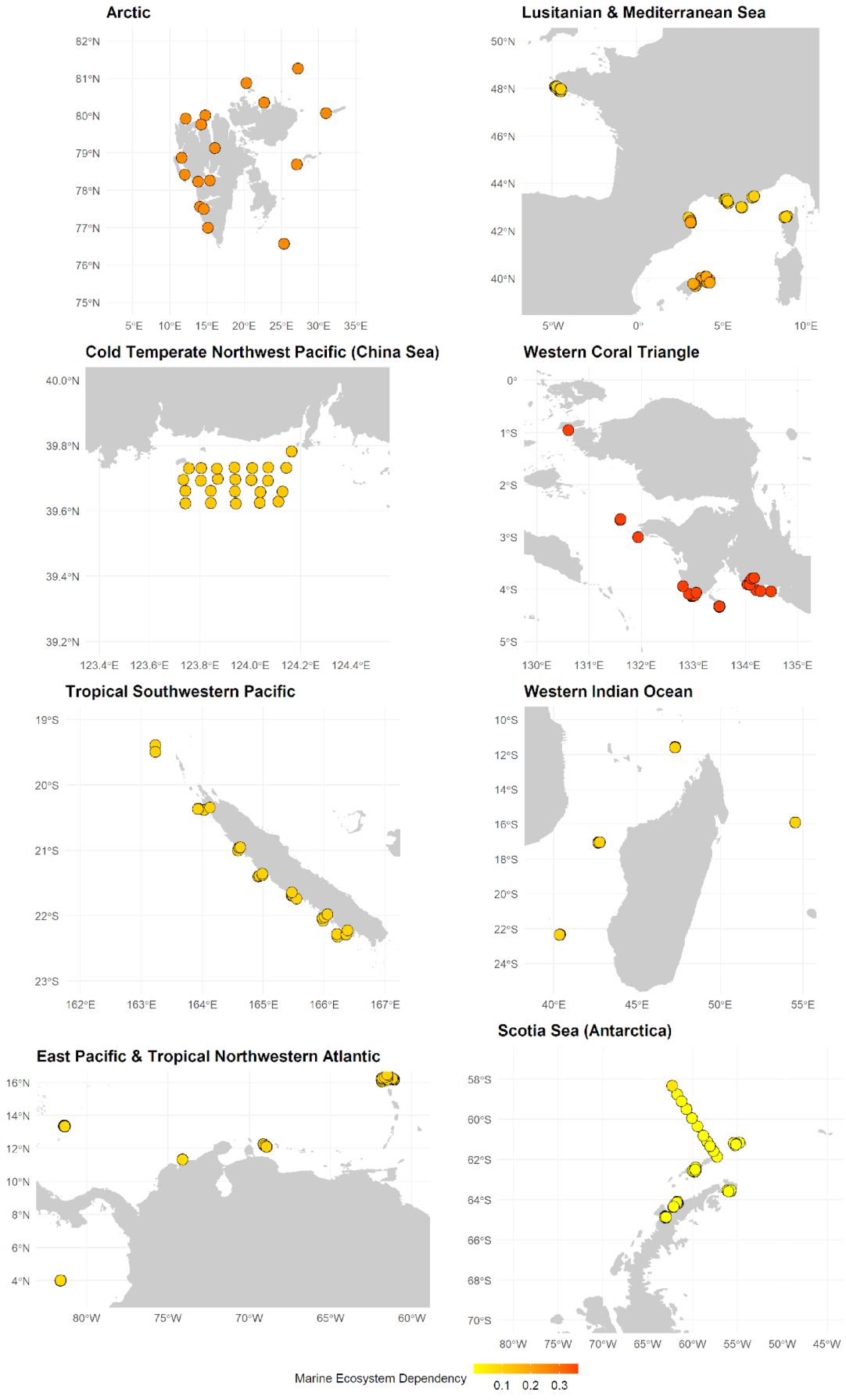


Fig. S7. *Maps of marine ecosystem dependence by region*

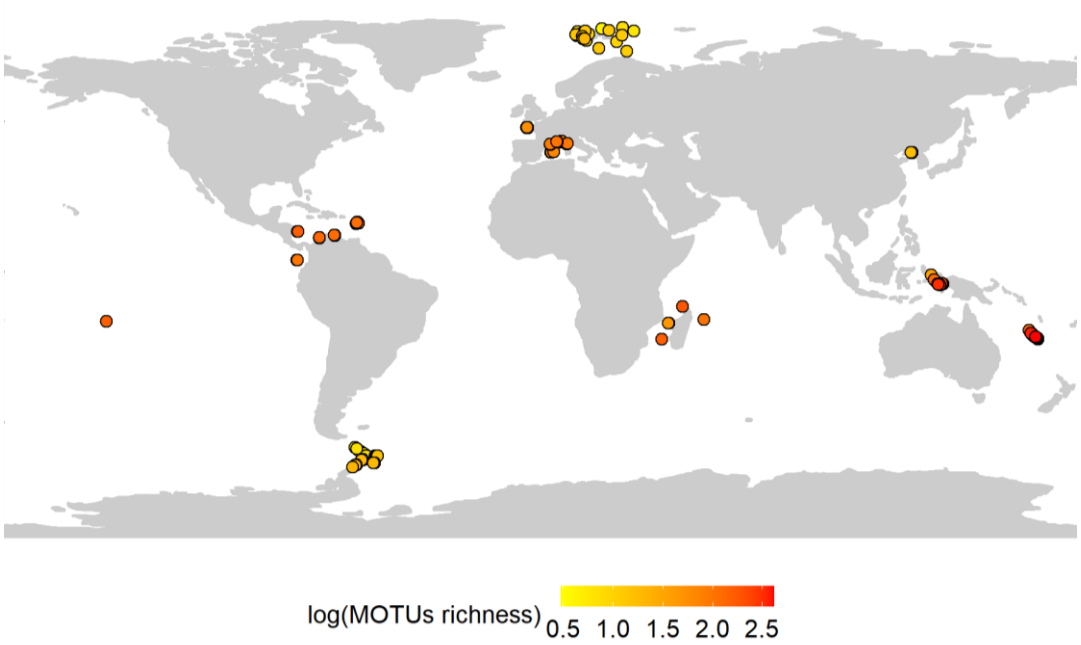


Fig. S8. *Map of all fish taxonomic α-diversity per station (log_{10} transformed). Raw values range from 2 to 414 MOTUs per station.*

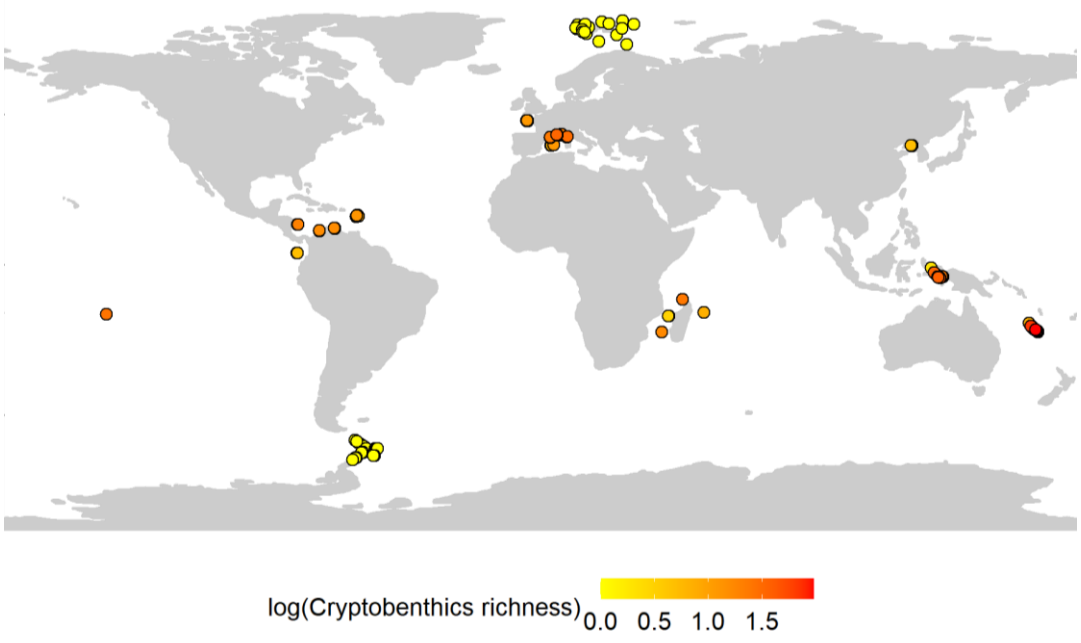


Fig. S9. *Map of Cryptobenthic fish taxonomic α-diversity per station (log_{10} transformed). Raw values range from 0 to 95 MOTUs per station.*

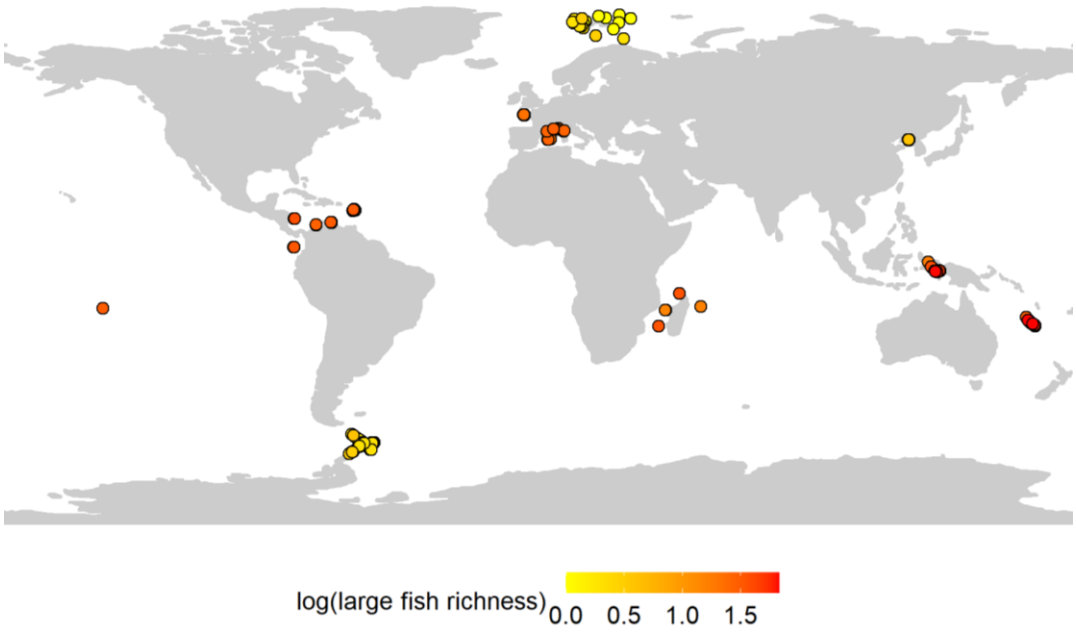


Fig. S10. *Map of large fish taxonomic α-diversity per station (log_{10} transformed). Raw values range from 0 to 67 MOTUs per station.*

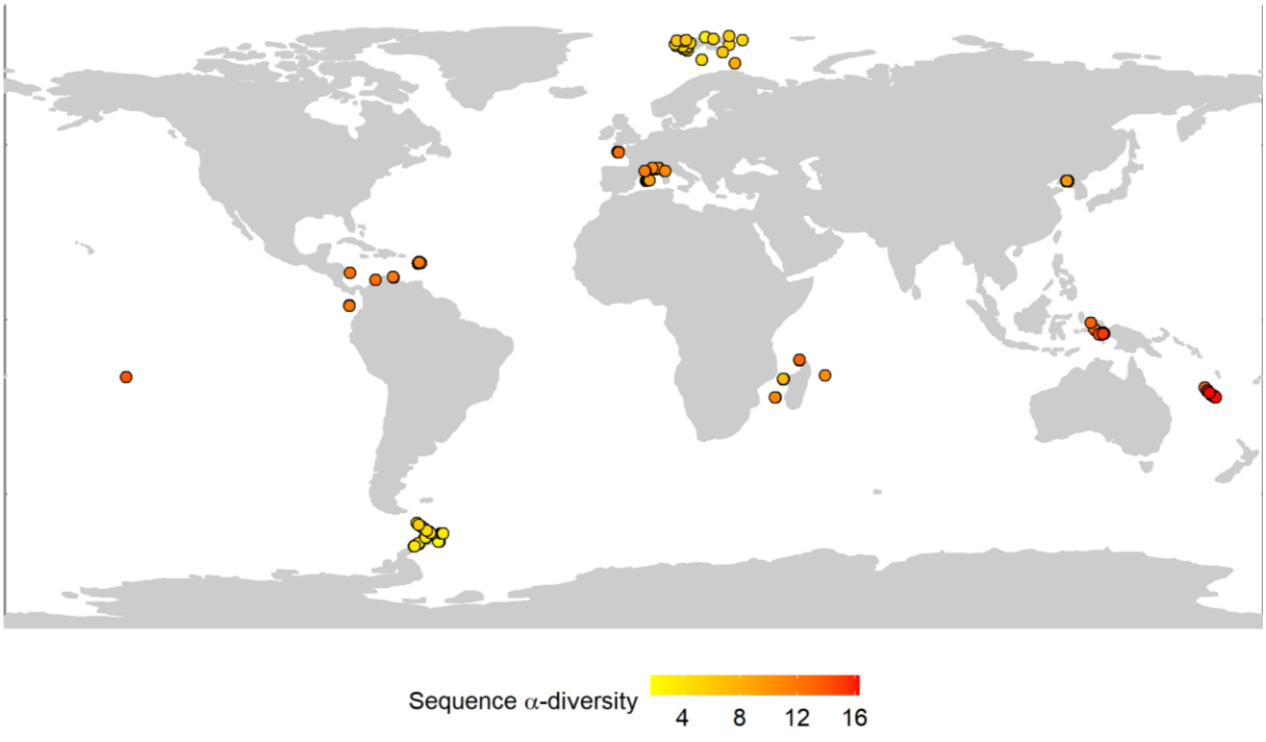


Fig. S11. *Map of sequence α-diversity per station.*

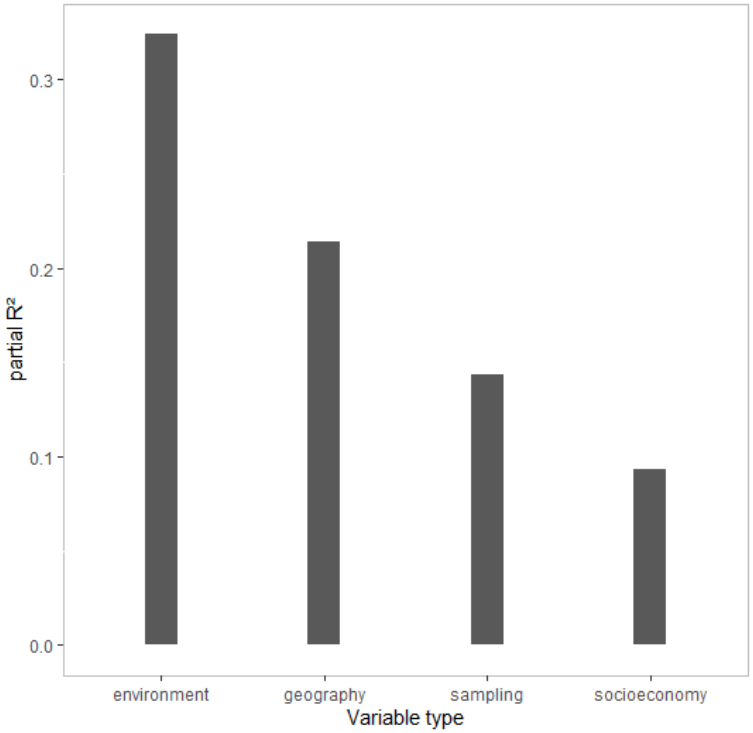


Fig. S12. *R² partitioned by variable groups in GLS predicting all fish taxonomic α-diversity.*

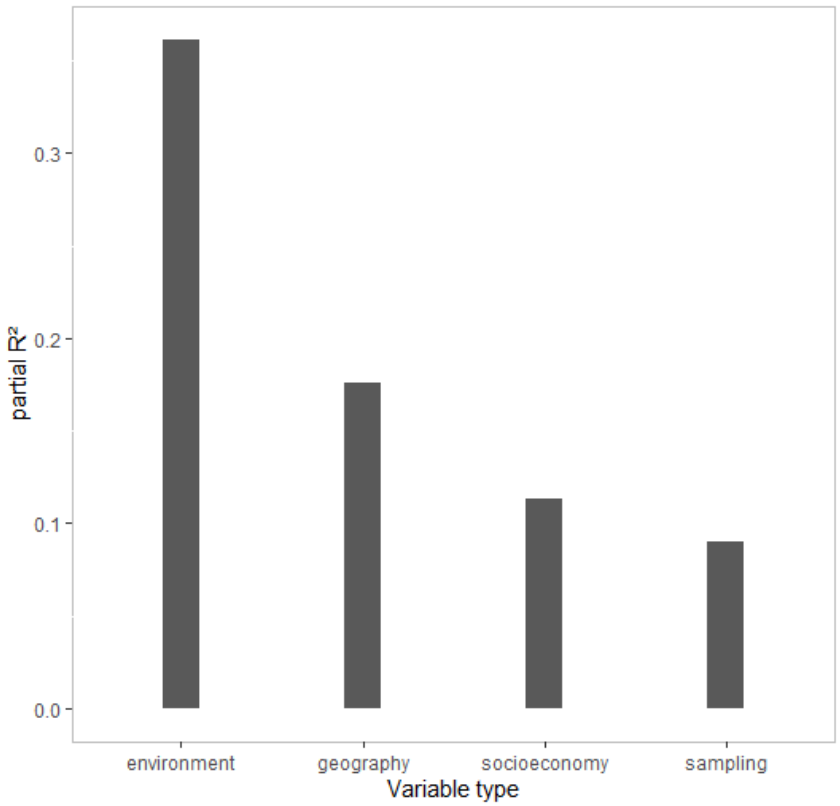


Fig. S13. *R² partitioned by variable groups in GLS predicting fish sequence α-diversity.*

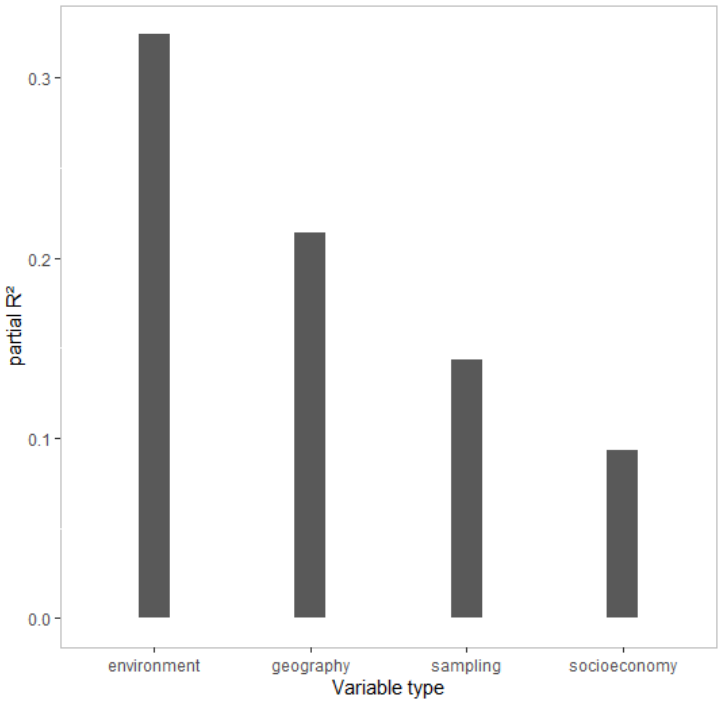


Fig. S14. *R² partitioned by variable groups in GLS predicting cryptobenthic taxonomic α-diversity (n=539 MOTUs).*

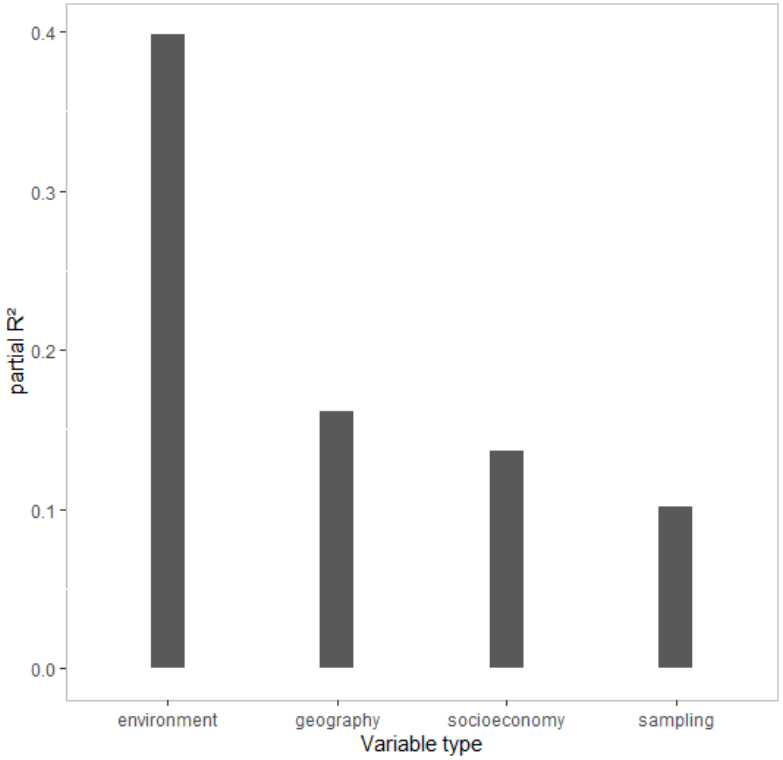


Fig. S15. *R² partitioned by variable groups in GLS predicting large fish taxonomic α-diversity (n=479 MOTUs).*

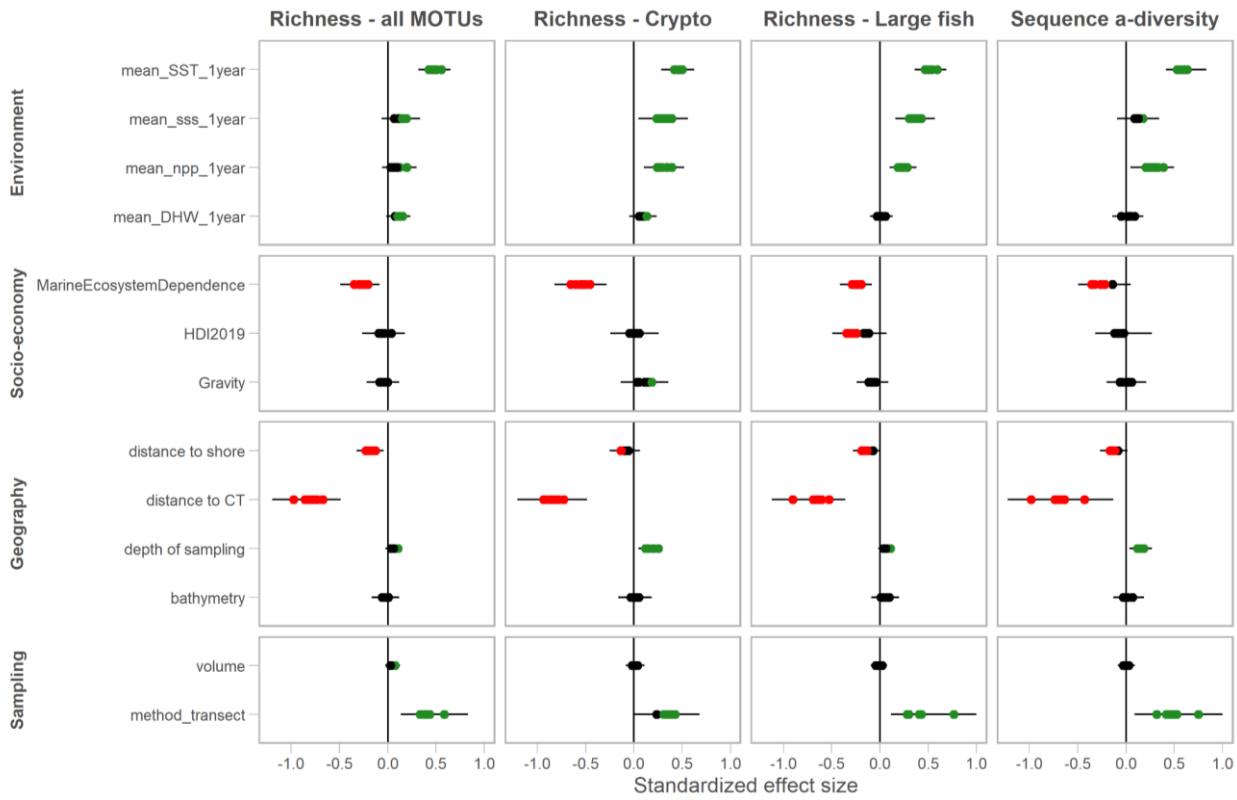


Fig. S16. *Sensitivity analysis: Effect size of 10 GLS with random subset of 80% of stations. (Cryptobenthic reduced dataset n=539 MOTUs, Large fish reduced dataset n=479 MOTUs).*

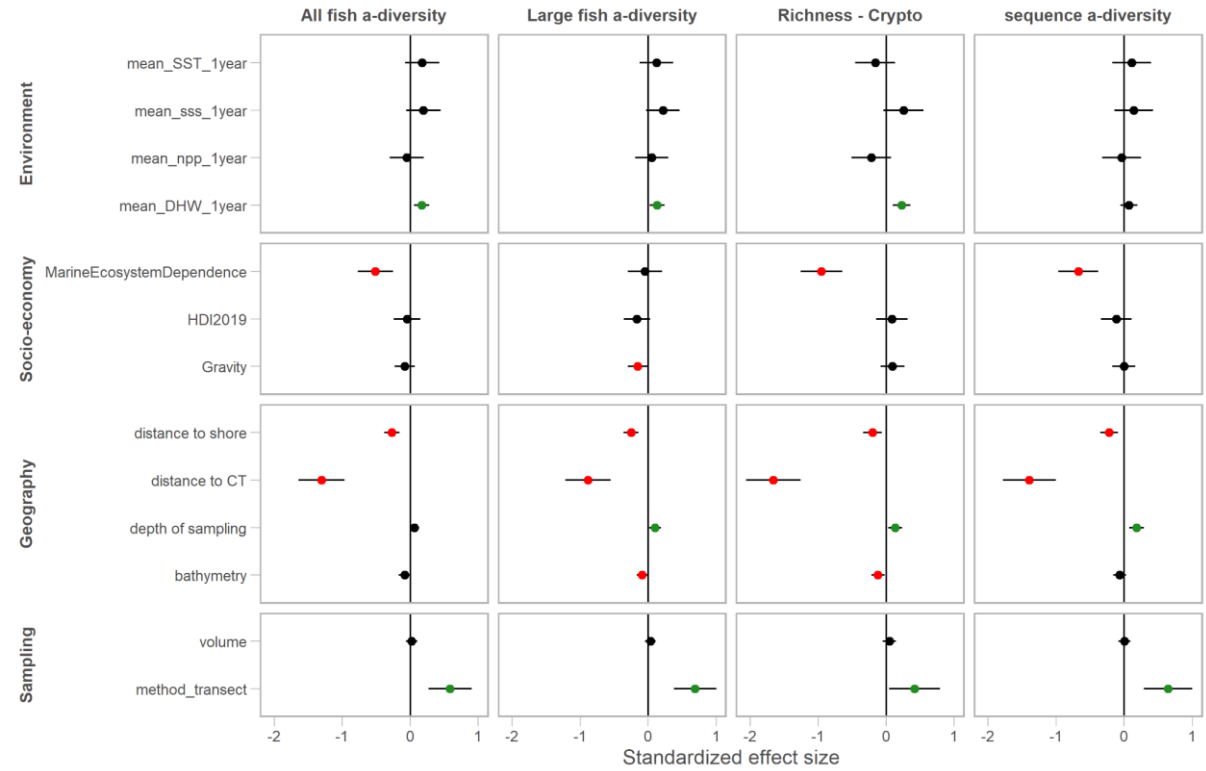


Fig. S17. *Effect sizes of variables in GLS on all stations without polar regions (Scotia sea and Arctic). (Cryptobenthic reduced dataset n=539 MOTUs, Large fish reduced dataset n=479 MOTUs).*

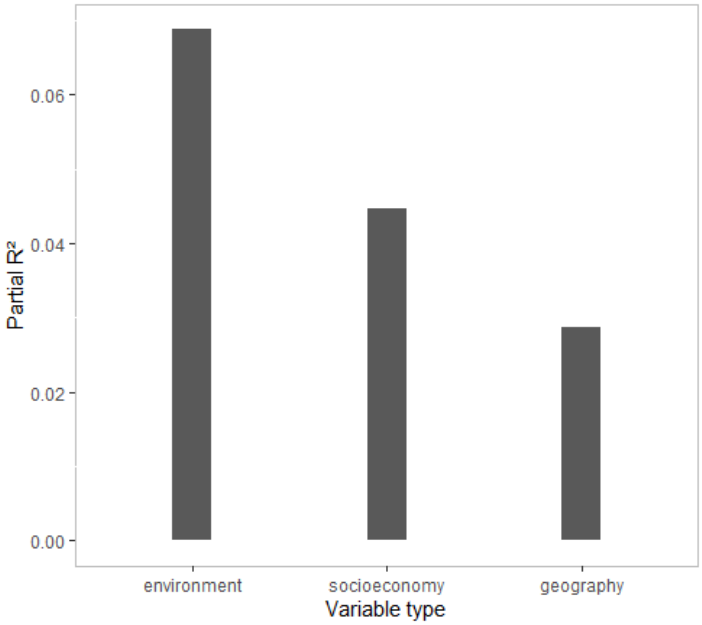


Fig. S18. *Partial R² by variable groups in dbRDA predicting taxonomic β-diversity between stations*

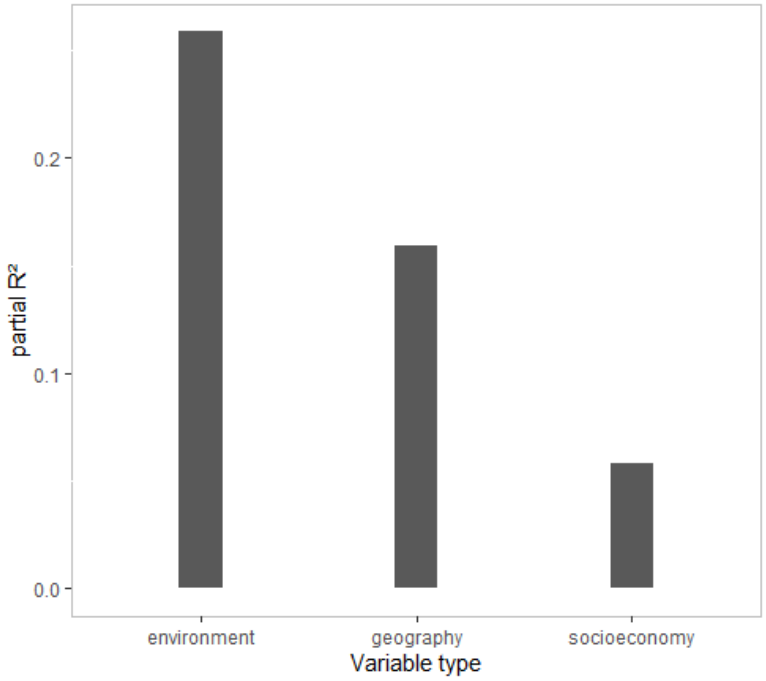


Fig. S19. *Partial R² by variable groups in dbRDA predicting sequence β-diversity between stations*

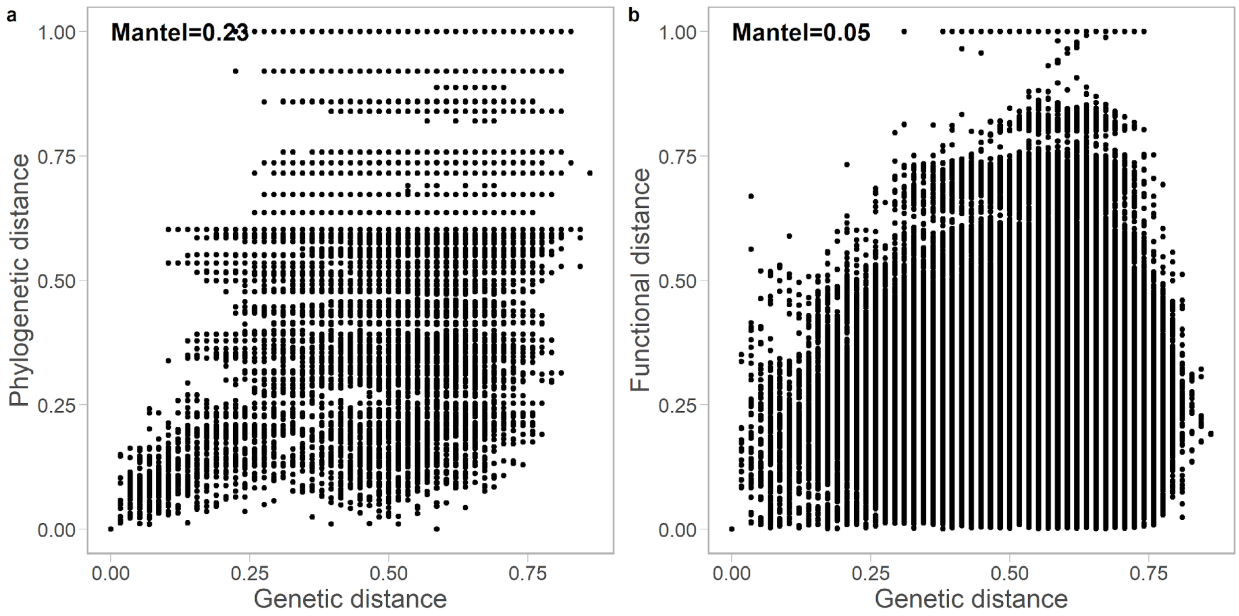


Fig. S20. *Correlation between* ***a,*** *phylogenetic pairwise distance and genetic pairwise distance, and* ***b****, between functional pairwise distance and genetic pairwise distance.*

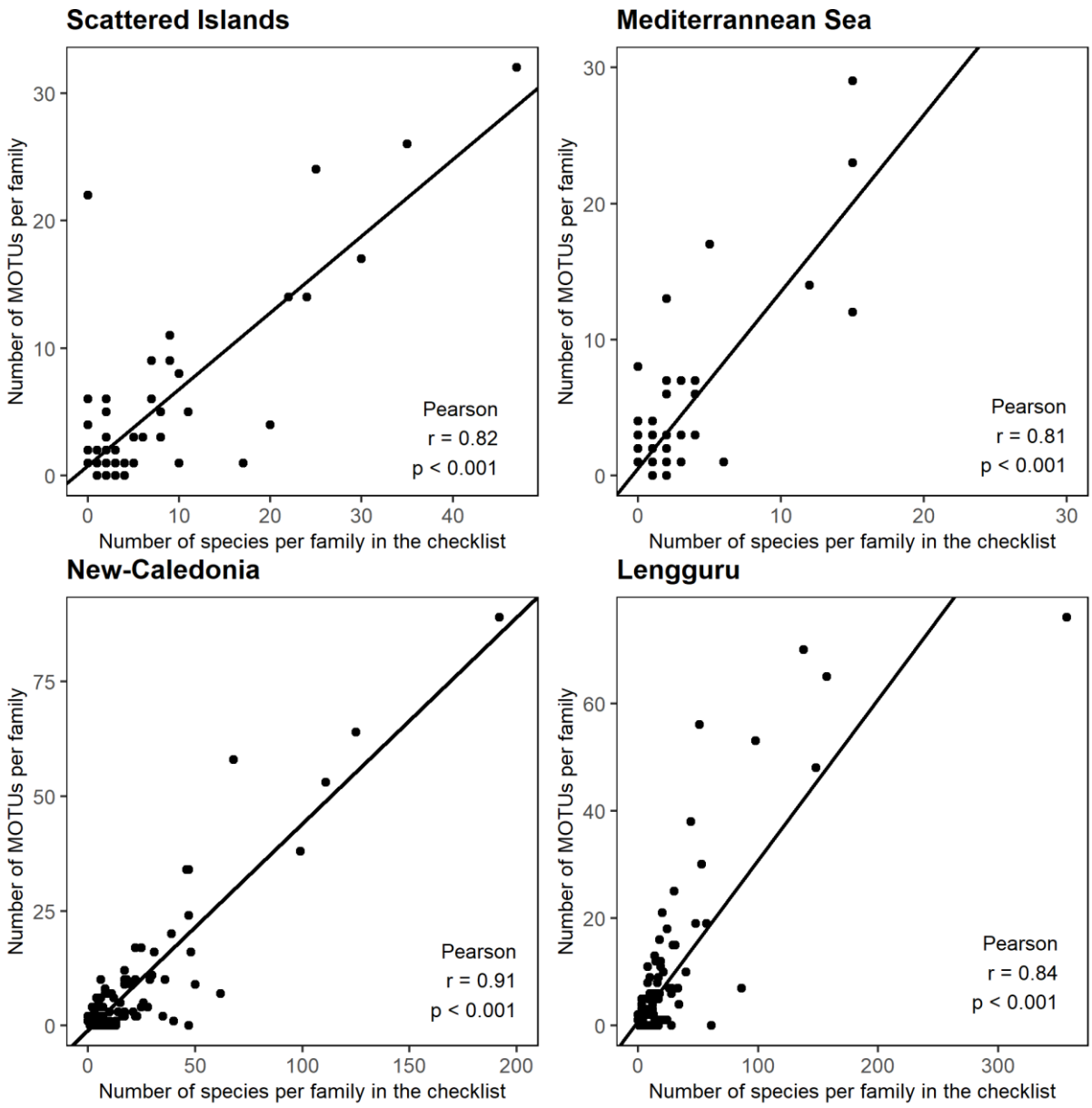


Fig. S21. *Correlation between number of MOTUs per family in our dataset and number of species per family in the regional checklists for the scattered islands, Mediterranean Sea, New-Caledonia and Lengguru.*

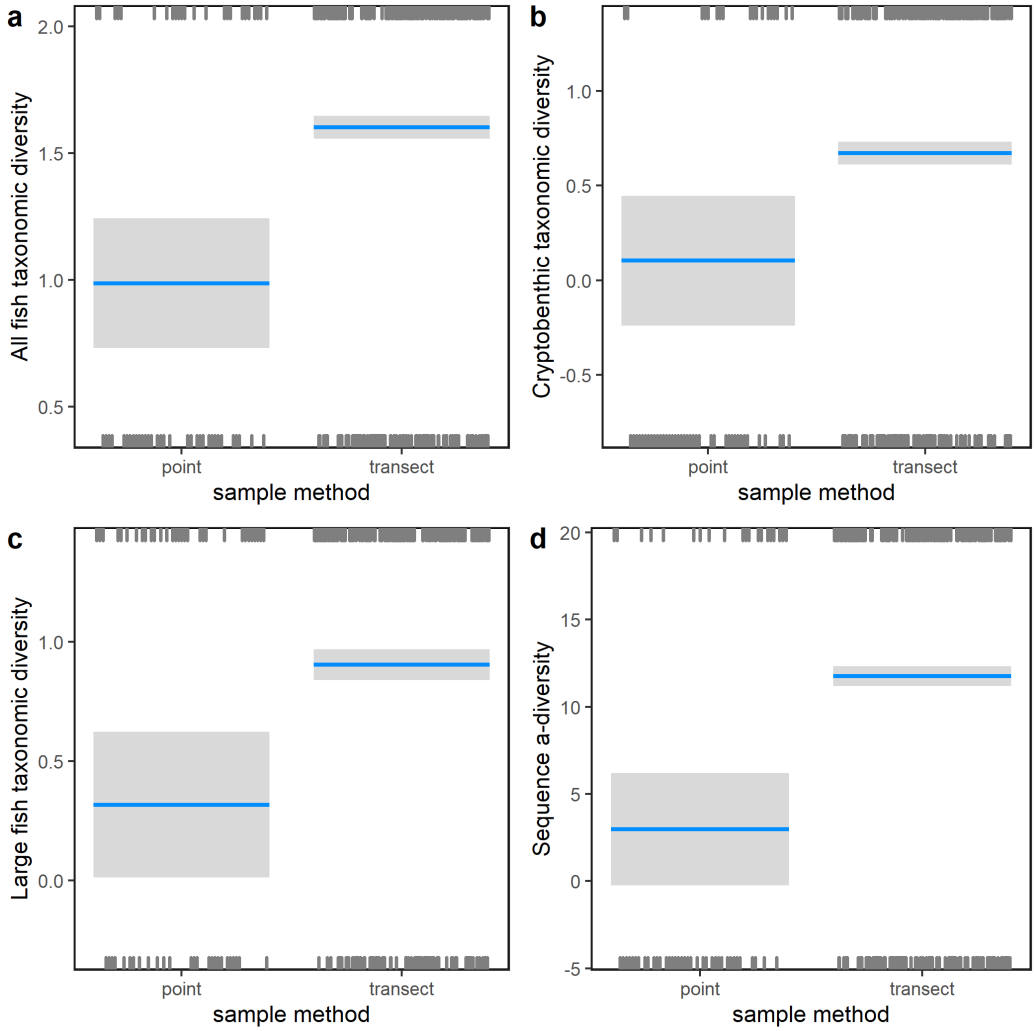


Fig. S22. *Partial regression plots showing the influence of the sampling method on* ***a,*** *all fish taxonomic α-diversity,* ***b,*** *Cryptobenthic taxonomic α-diversity (n=539 MOTUs),* ***c,*** *Large fish taxonomic α-diversity (n=479 MOTUs), and* ***d,*** *sequence α-diversity, conditioned on the median value of all other retained factors.*

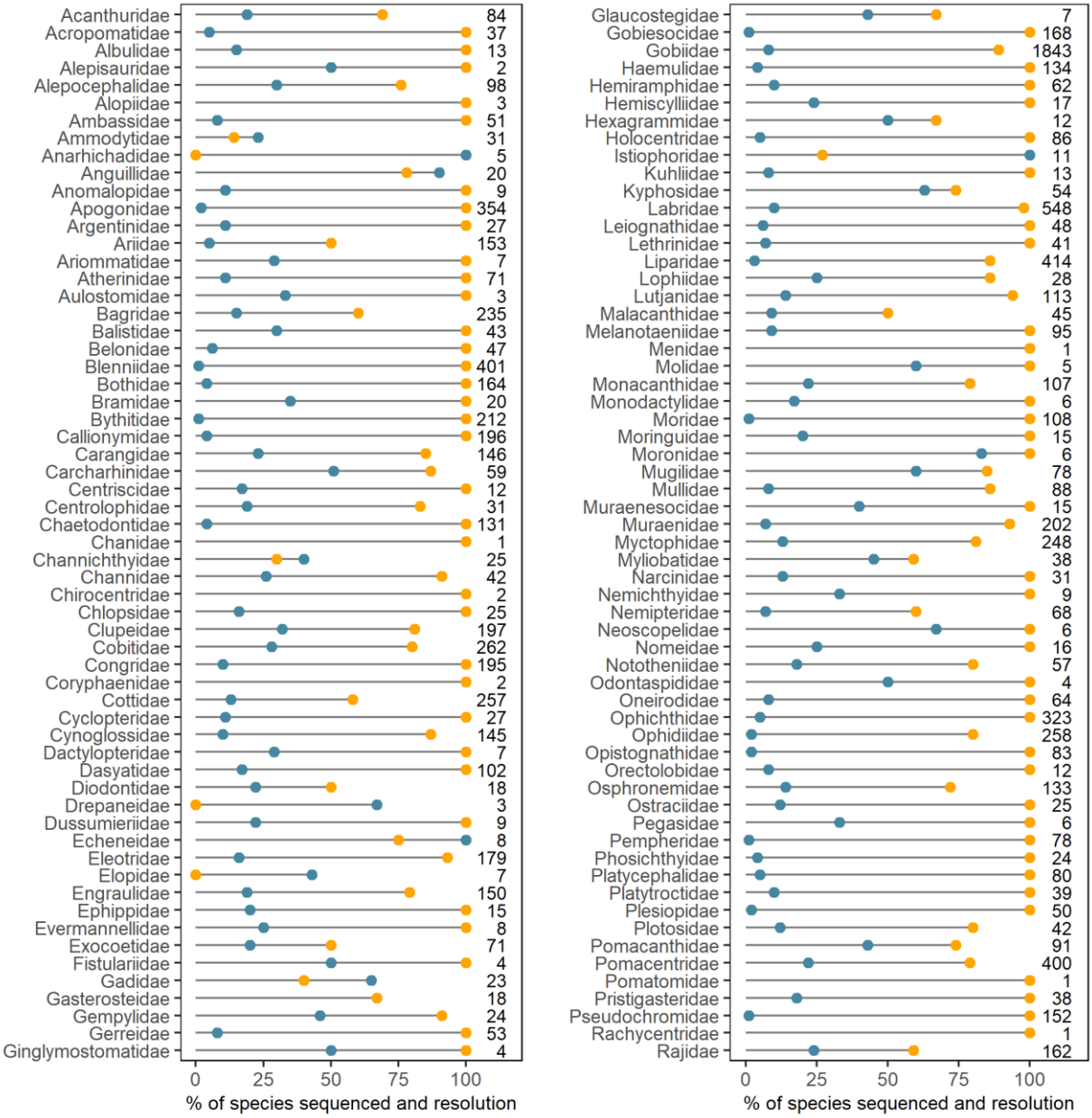


Fig. S23. *Percentage of species sequenced within each family (blue) and percentage of exclusive taxonomic resolution of the teleo barcode (BE = Proportion of species whose amplified sequences are unambiguously identified. Considering repeated species labels as ambiguity, Marquina et al (2018)) within each family (orange). Total number of species in the family is reported on the graph.*

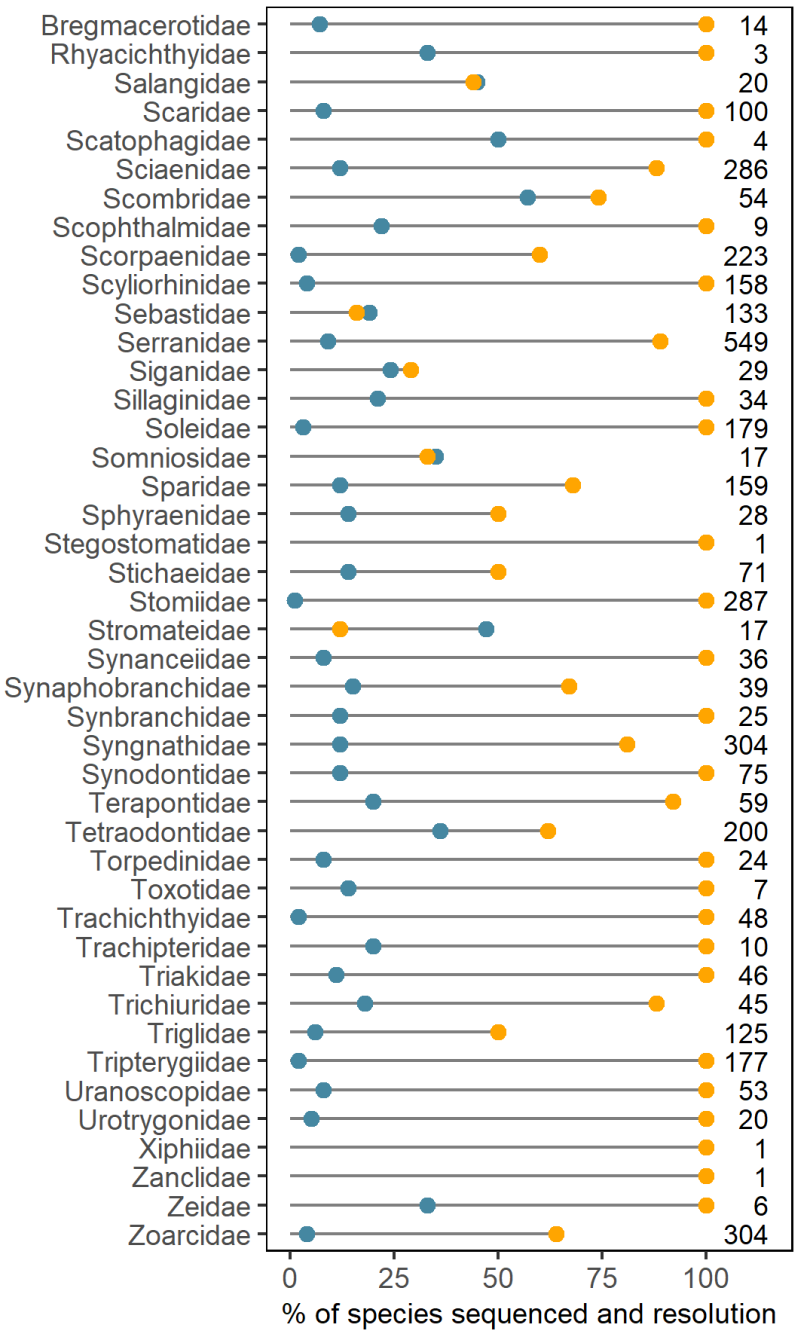


Fig. S23 (suite). *Percentage of species sequenced within each family (blue) and percentage of exclusive taxonomic resolution of the teleo barcode within each family (orange). Total number of species in the family is reported on the graph.*

5. Suppléments du manuscrit D : *3D conservation planning of multiple marine fish biodiversity metrics reveals 30x30 CBD target in the deep Coral Sea*

5.1. Méthodes supplémentaires

Data collection

New Caledonia is a South Pacific archipelago, east of Australia, in the Coral Sea. The 400 km-long main island of "Grande Terre" is surrounded by the second longest barrier reef in the world cumulating 1,600 km. Beyond the barrier reef is the island's deep slope habitat, also surrounding the remote atolls (Chesterfield, Bellona, Entrecasteaux…), the three Loyalty Islands and the other smaller islands and remote reefs of the archipelago (Fig. 1). Two third of the archipelago's population lives in Grand Terre's southwest, in and around Nouméa (~180,000 people) (ISEE, 2019), creating a gradient of human pressure from densely populated areas to wilderness areas located at > 10-20 h travel time from the capital (Januchowski-hartley et al., 2020).

Data was collected during four cruises on the R/V Alis, in April and June 2019 and August and September 2020 (https://doi.org/10.17600/18000889, https://doi.org/10.17600/18001119). We sampled 11 seamounts across the archipelago, and four deep island slopes along the west coast of Grande Terre (Figure 1, Figures S1-15). Samples were collected on the summit of the seamounts or the seafloor along the deep slopes (benthic samples), as well as in the pelagic waters, 2-7 miles away from the seamounts or island slopes (pelagic samples). The seamounts were chosen to have different summit depth corresponding to euphotic, intermediate and aphotic zones: four seamounts had their summit around 50 m deep (Seamount 50), four seamounts had their summits between 200 and 320 m deep (Seamount 250), and three seamounts had their summits between 320 and 500 m deep (Seamount 500). Deep slopes were sampled between 100 and 220 m deep (Table S1). Sampling sites were spread throughout the archipelago in order to obtain a wide range of human and environmental conditions found in New Caledonia.

Benthic environmental DNA samples were collected with 4 x 8 L Niskin bottles at each station, 5 m above the seafloor, with 10 stations on each seamount summit or deep island slope (total of 150 stations). The shallowest benthic eDNA sample was collected at 45 m deep and the deepest at 570 m deep (Table S1). Pelagic eDNA samples were collected at one vertical profiles per site, with samples of 32 L collected at 6 depths: 20, 80, 150, 250, 500 and 1000 m. The Niskin bottles were lifted on the ship and the water from the four bottles per station was filtered on a single filtration capsule. The eDNA filtration was done with an Alexis® 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 with a polyethersulfone membrane (SPYGEN, le Bourget du Lac, France) and disposable sterile tubing for each filtration capsule. At the end of each filtration, the water inside the capsules were 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 each sampling campaign, a strict contamination control protocol was followed in both field and laboratory stages (Goldberg et al., 2016; Valentini et al., 2016), and

each water sample processing included the use of disposable gloves and single-use filtration equipment. The shallowest benthic eDNA sample was collected at 45 m deep and the deepest at 570 m deep (Table S1).

On the same sites (summits and deep slopes), baited remote underwater videos stations (BRUVS) were deployed on the seafloor, at 5 to 10 stations per site (total of 120 stations, Table S1). Stations were separated by at least 1 km to avoid individuals from appearing on multiple videos and assume independence of samples (Langlois et al., 2020). The BRUVS were composed of two cameras aligned horizontally on a metallic structure, a bait of 1kg of crushed sardines at the end of a 1.5 m bar facing the cameras, a spotlight, and 20 kg of weight to hold the system still on the floor. The stereo pair of cameras were separated by 800 mm, with a convergent angle of 8 °. GoPro Hero 4 cameras were used and set to a medium field of view (FOV) in 1920 x 1080 pixel format running at 60 frames per second. Soaking times were calculated from the time the BRUVS reached the seabed (t0) to t0 + 120 mns. The shallowest BRUVS was deployed at 47 m deep and the deepest at 552 m deep (Table S1).

Acoustic data were recorded in-situ continuously during the cruises, using an EK60 echosounder (SIMRAD Kongsberg Maritime AS, Horten, Norway) connected to four split-beam transducers at 38, 70, 120 and 200 kHz. EK60 calibration was performed according to Foote (1987) for each cruise. In the present study, we used 38 kHz only as only that frequency allowed to cover all depth ranges considered. The hull-mounted transducer was 4 m below the surface and detections shallower than 6 m below the transducer face were deleted from the records to avoid surface noise. Thus, acoustic data collection started at 10 m below the surface. The water column was sampled down to 800 m depth for all the surveys.

eDNA extraction, amplification and sequencing

DNA extraction was performed in a dedicated DNA laboratory (SPYGEN, www.spygen.com) equipped with positive air pressure, UV treatment and frequent air renewal. Decontamination procedures were conducted before and after all manipulations. Each filtration capsule was agitated for 15 min on a S50 Shaker (Cat Ingenieurbüro™) at 800 rpm and then the buffer was emptied into two 50-mL tube 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 3M sodium acetate were added to each 50-mL tube and stored for at least one night at -20 °C. The DNA extraction was performed using NucleoSpin® Soil (MACHEREY-NAGEL GmbH & Co., Düren Germany) starting from step 6 and following the manufacturer's instructions. The elution was performed by adding 100 µL of SE buffer twice. The two 50 mL tubes per filtration capsule were extracted separately then the two DNA samples were pooled before the amplification step. A teleost-specific 12S mitochondrial rRNA primer pair (teleo, forward primer - ACACCGCCCGTCACTCT, reverse primer – CTTCCGGTACACTTACCATG, Valentini et al 2016) was used for the amplification of metabarcode sequences. As we analysed our data using MOTUs as a proxy for species to overcome genetic database limitations, we chose to amplify only one marker. Twelve DNA amplifications PCR per sample 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" primers and 0.2 µg/µL bovine serum albumin (BSA, Roche Diagnostic, Basel, Switzerland). 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, 1 min at 72 °C and a final elongation step at 72 °C for 7 min. The teleo primers were 5'-labeled 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. Negative extraction controls and negative PCR controls (ultrapure water) were amplified (with 12 replicates as well) and sequenced in parallel to the samples to monitor possible contaminations. 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. Library preparation and sequencing were performed at Fasteris (Geneva, Switzerland). A total of 24 libraries were prepared using MetaFast protocol. A paired-end sequencing (2x125 bp) was carried out using an Illumina MiSeq (2x125 bp, Illumina, San Diego, CA, USA) using the MiSeq Flow Cell Kit v3 (Illumina, San Diego, CA, USA) or a NextSeq sequencer (2x125 bp, Illumina, San Diego, CA, USA) with the NextSeq Mid kit following the manufacturer's instructions. This generated an average of 527,540 (± 721,560) sequence reads (paired-end Illumina) per sample.

eDNA bioinformatic analyses

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) (Marques et al., 2020). First, reads were assembled using vsearch (Rognes et al., 2016), then demultiplexed and trimmed using cutadapt (Martin., 1994) and clustering was performed using Swarm v.2 (Mahé et al., 2015) with d = 1, which corresponds to a maximum of one mismatch between neighboring pairs of sequences within each cluster. The iterative process of SWARM leads to clusters composed of many sequences with more than d mismatches. Further, we used the -f (fastidious) option, which creates virtual sequences within clusters to link more dissimilar sequences together, hence limiting alpha-diversity inflation by joining low abundant MOTUs within larger ones. The minimum distance between clusters is 2 mismatches (d+1).Taxonomic assignment of MOTUs was carried out using the Lower Common Ancestor (LCA) algorithm ecotag implemented in the Obitools toolkit (Boyer et al., 2016) and the European Nucleotide Archive (ENA, Leinonen et al., 2011) as a reference database (release 143, March 2020). It assigns a taxonomy to sequences even when the sequence match is not perfect, based on NCBI taxonomic tree of species to consider the current knowledge on molecular diversity per branch and assign a taxonomy at the lowest possible rank. If the sequence matches several identifications with equal percentages of similarity, ecotag assigns to the upper taxonomic level common between all possible matches. We then applied quality filters to be conservative in our estimates. We discarded all observations with less than 10 reads, and present in only one PCR per site to avoid spurious MOTUs originating from a PCR error.

Then, errors generated by index-hopping (MacConaill et al., 2018) were filtered using a threshold empirically determined per sequencing batch using experimental blanks (combinations of tags not present in the libraries) (Taberlet et al., 2018), and tag-jump (Schnell, Bohmann, & Gilbert, 2015) was corrected using a threshold of 0.001 of occurrence for a given MOTU within a library. Taxonomic assignments at the species level were accepted if the percentage of similarity with the reference sequence was 100%, at the genus level if the similarity was between 90 and 99%, and at the family level if the similarity was > 85%. If these criteria were not met, the MOTU was left unassigned. The post-LCA algorithm correction threshold of 85% similarity for family assignment was chosen to include a maximum of correct family assignment while minimizing the risk of adding wrong family assignments in the family detections.

Video analysis

Stereo measurement was made available with the recording of three claps before deployment to synchronize frames. Calibration was done using the software CAL and fish were counted using the EventMeasure software (www.seagis.com.au). We used the MaxN metric (corresponding to the maximum number of a particular species seen in any one video frame across the duration of the video record), which is until now the standard and most used method (Cappo et al., 2007; Langlois et al., 2020; Whitmarsh et al., 2017). Fork length of individual fish was measured, when possible, up to a limit of 10 individuals per BRUVS per species to optimize video processing time. Biomass was calculated for each species of each BRUVS using the length-weight relationship:$Weight(g) = a * Length(cm)^b$ (Taylor & Willis, 1998) with a and b being the allometric coefficient of the length-weight relationship retrieved from FishBase (www.fishbase.se). As we did not seek highly accurate length structure among replicates but rather the general patterns of biomass, the length used for each estimation was the average length of all measured individuals per species (up to ten) in a single BRUVS. When particular species could not be measured on a single BRUVS, the missing species length was estimated by data imputation using the MissForest algorithm with 999 trees. Missing length were imputed using measured length records of other samples, but also family, genus, maximum species size, and size type from Fishbase. Latitude and longitude of available species lengths were also used to account for geographic proximity of measured lengths. The MissForest accuracy was tested with a k-fold cross validation procedure by predicting 5% of the lengths each time by training the missForest on the 95% left of the data and looking at the linear fit between the original and predicted. We also ensured that imputed length did not exceed Fishbase's max reported length.

Acoustic data cleaning

All raw acoustic data were processed with the open-source Matecho software (Perrot et al., 2018). A first manual cleaning step removed ghost bottom echoes. Then, four semi-automatic cleaning filters were applied to: (i) remove acoustic device interference ('un-parasite' Matecho filter), (ii) remove attenuated signals ('white pings' filter), (iii) remove elevated signals ('deep spike' filter) and (iv) reduce background noise (De Robertis & Higginbottom, 2007). Details of filter parameters can be found in Béhagle *et al.*, (2016) and Perrot *et al.*, (2018). After data

cleaning, the echo-integration was done on cells of 1m deep and 0.1nm long, providing volume backscattering strength S_v (S_v, in $dB.re.1.m^{-1}$), and the nautical area scattering coefficient s_A (NASC or s_A, in $m^2.nm^{-2}$), a proxy for the fish biomass for each cell (Irigoien et al., 2014; Maclennan et al., 2002). Vertical profiles were smoothed using a locally polynomial quantile regression (Koenker, 2004) to remove high-frequency peaks (e.g. interferences or very small schools that create peaks in an acoustic profile) that were considered non-interpretable in the present study. The final dataset was composed of 5,064 vertical profiles ranging from 10 to 800 m depth with s_A integrated in 10 m vertical bins and 500 m horizontal resolution.

Modelling abundance, richness and acoustic biomass

Seventeen variables were collected as potential explanatory variables for fish biodiversity patterns. At each station, we recorded the sampling depth, the bottom depth, the habitat (seamount or deep island slope) and the depth of the summit. Using a bathymetry at 100 m resolution (Roger, 2020), we calculated the summit area (km²) and the summit rugosity as the standard deviation of depth in the cells of the summit area. For deep slope stations, the summit depth was set at 0, as the land is considered to be the summit, and the summit area was calculated as the area of cells with depth < 60 m. For each station, we extracted maximum and mean sea surface temperature (SST), mean surface salinity, eastward and northward current velocity, surface suspended particulate matter, seafloor potential temperature and chlorophyll a over the last 10 years from available rasters. Details on the sources and resolution of each variable can be found in Table S2. We also calculated the travel time from Noumea to our stations as a proxy for human pressure and habitat remoteness (Januchowski-hartley et al., 2020; Maire et al., 2016) and the minimum distances from our stations to reefs and land, using the New-Caledonia Millennium Geomorphology (Andréfouët et al., 2006).

Boosted regression trees (BRT, Elith et al., 2008) were used to model total species richness and biomass (BRUVS), benthic and pelagic MOTU richness per sample (eDNA), and benthic and pelagic acoustic biomass. BRUVS biomass and abundance, acoustic biomass and pelagic MOTU richness were log-transformed. The MOTU and species richness were modeled with a Poisson distribution, while biomass, abundance and acoustic biomass were modeled with a Gaussian distribution. The function *gbm.step* from the package *dismo* (Hijmans et al., 2017) was used to find the combination of parameters producing the best fit. Parameters were the tree complexity (from 1 to 5), the learning rate (0.01, 0.005 or 0.001) and the bag fraction (0.5 or 0.75). All possible combinations of tree complexity, learning rate and bag fraction were run. The number of folds was set to 10. The initial number of trees was set to 700, and the step size was 25. The combination with the lowest deviance and standard error (evaluated over a 10-fold cross-validation) was then selected to identify best parameters. Models were computed again with the best parameters and fixed number of trees with the function *gbm.fixed*. The predictors contributing the most to the models were selected, removing variables contributing less than 5%, and a *gbm.step* was computed with this reduced predictors selection to fit the final models.

The contributions reflect the relative influence of each predictor and are computed based on the number of times a variable is selected for splitting, weighted by the squared improvement to the model as a result of each split, and averaged over all trees (Friedman & Meulman, 2003).

The relative influence (or contribution) of each variable is scaled so that the sum adds to 100, with higher numbers indicating stronger influence on the response.

The generalized joint attribute models (GJAM) on the BRUVS species abundance matrix, and the benthic and pelagic matrix of eDNA read number per sample, were computed using the function *gjam* from the package *gjam* (Clark et al., 2017). The explanatory variables to include in each model were selected with a step-by-step process, analyzing the sensitivity of the response variables to each variable and their quadratic and cubic terms. Only variables with a VIF < 20 were kept in the final model. The response variable type was set to 'discrete abundance' for species abundance and 'count composition' for read number, and the models were run with parameters *ng* = 2500 and *burnin* = 500. Pearson's correlation coefficient R was computed between observed and predicted values, to estimate the model's goodness-of-fit.

Spatial conservation planning in 3D

A principal component analysis on acoustic vertical profiles and a classification by k-means of these profiles allowed us to identify three depth layers with differentiated acoustic signals: 0-200 m, 200-400 m and 400-600 m, approximately corresponding to euphotic, intermediate and aphotic zones, respectively. We thus decided to use these three depths layers for our planning in three dimensions. We divided our benthic predictions between these three depth layers, and aggregated our pelagic predictions within these three depth layers (sum of acoustic biomass, and mean of MOTU richness).

We used the spatial prioritization package prioritizr (Hanson et al., 2022) to identify conservation priority areas across the three depth layers. In order to perform the 3D spatial prioritization, we modified the input data required for the 2D prioritization. Each planning unit was identified by a unique identifier, a 2D identifier and a depth identifier. Then, we computed the boundary length between each pair of planning units, so that each planning unit shares boundaries with its 2D neighbors, but also with units in the upper and lower depth layers. The relative conservation targets were set to 0.3 (30%) for each community metric and individual species or MOTU abundance, to reflect the 30% target of the Convention on Biological Diversity's global framework aiming at protecting 30% of sea areas by 2030 (CBD, 2021). We set equal cost to all the planning units.

The formulation of the problem with *prioritizr* includes a factor referred to as boundary length modifier (BLM), which controls the compactness of selected sites. Lower BLM values minimize the total cost of the solutions, albeit more spatially fragmented; higher BLM values emphasize compact solutions, but the total cost becomes greater. We computed 8 iterations of our prioritization problem with BLM values between 0 and 10 (0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10), and assessed the efficiency of each solution. For each solution, we computed the total cost and the total boundary length (a proxy of fragmentation). We used the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) method (Hwang & Yoon, 1981). The candidate prioritization with the greatest TOPSIS score was considered to represent the best trade-off between total cost and total boundary length.

Table S1. *Information on benthic eDNA and BRUVS sampling on deep slopes and seamounts. Pelagic eDNA and acoustic sampling can be seen in Figures S1-S15.*

Environment	Site name	Latitude	Longitude	Summit depth (m)	Summit height (m)	eDNA		BRUVS	
						Number of samples	Sample depth range (m)	Number of stations	Sample depth range (m)
Seamount_50	Torche	-22,87503	167,6631	45	1318	10	45 - 58	5	47 - 62
	Antigonia	-23,42824	168,0752	54	1330	10	54 - 70	8	56 - 66
	Capel	-25,03758	159,5323	60	3054	10	60 - 70	10	65 - 69
	Fairway	-21,04964	162,255	62	2964	10	62 - 67	10	63 - 67
Deep slope_150	St Vincent	-22,12531	166,0376	-	-	10	80 - 219	7	120 - 180
	Nepoui	-21,42488	164,9774	-	-	10	88 - 218	10	105 - 150
	Poum	-20,15007	163,7853	-	-	10	100 - 185	8	100 - 220
	Grand Lagon Nord	-19,45239	163,2159	-	-	10	85 - 235	10	118 - 150
Seamount_250	Crypthelia	-23,3078	168,2498	195	1627	10	201 - 236	9	200 - 244
	Kaimon Maru	-24,74137	168,1411	236	1799	10	238 - 325	8	238 - 340
	Jumeau Ouest	-23,68215	168,0081	239	1119	10	242 - 313	8	245 - 339
	Argo	-23,09286	159,463	299	2251	10	299 - 313	8	301 - 312
Seamount_500	Stylaster	-23,6461	167,7134	434	801	10	439 - 488	8	444 - 491
	Ile Des Pins	-22,38325	167,407	470	818	10	469 - 488	5	480 - 506
	Eponge	-24,91183	168,363	511	1932	10	518 - 570	7	520 - 552

Table S2. *Information and sources of environmental variables*

Variable	Source	Unit	Original spatial resolution	Temporal resolution	Temporal duration	Web link
Sea Surface Temperature	GLOBAL observed high resolution Sea Surface Temperatures.	Celcius degrees	1 km		2019-2021	https://thredds.jpl.nasa.gov/thredds/ncss/grid/OceanTemperature/MUR-JPL-L4-GLOB-v4.1.nc/dataset.html
Sea surface Salinity	Global observed sea surface salinity	PSU	8 km	Monthly	2009-2019	https://resources.marine.copernicus.eu/product-detail/MULTIOBS_GLO_PHY_S_SURFACE_MYNRT_015_013
Surface Chlorophyll-A	CMEMS GlobColour : observed satellite chlorophyll	Mg m^{-3}	4 km	Monthly	2010-2020	https://resources.marine.copernicus.eu/product-detail/OCEANCOLOUR_GLO_CHL_L4_REP_OBSERVATIONS_009_082
Eastward Velocity	CMEMS Mercator model reanalyses at the surface	m. s^{-1}	8 km	Monthly	2009-2019	https://resources.marine.copernicus.eu/product-detail/GLOBAL_REANALYSIS_PHY_001_031
Northward Velocity	CMEMS Mercator model reanalyses at the surface	m. s^{-1}	8 km	Monthly	2009-2019	https://resources.marine.copernicus.eu/product-detail/GLOBAL_REANALYSIS_PHY_001_031
Suspended particulate matter	CMEMS GlobColour : observed satellite data	g. m^{-3}	4 km	Monthly	2010-2020	https://resources.marine.copernicus.eu/product-detail/OCEANCOLOUR_GLO_OPTICS_L4_NRT_OBSERVATIONS_009_083
Potential seafloor temperature	CMEMS Mercator model reanalyses at the surface	Celcius degrees	8 km	Monthly	2009-2019	https://resources.marine.copernicus.eu/product-detail/GLOBAL_REANALYSIS_PHY_001_031

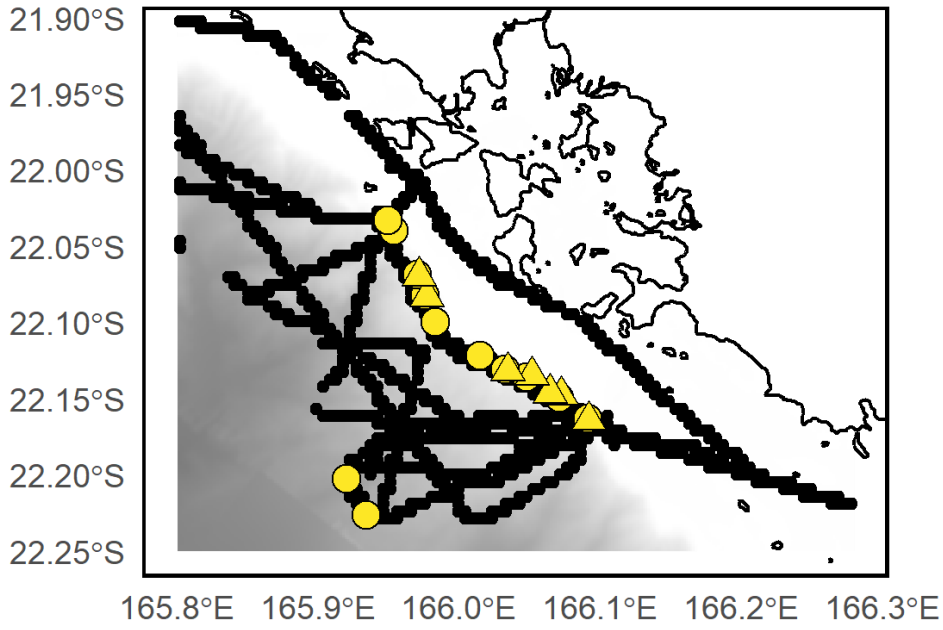


Figure S1. *Sampling design in site 1 (Nouméa). Grey shading indicate the depth category. BRUVS sampling stations (triangles), eDNA sampling stations (circles), and acoustic recordings (black dots).*

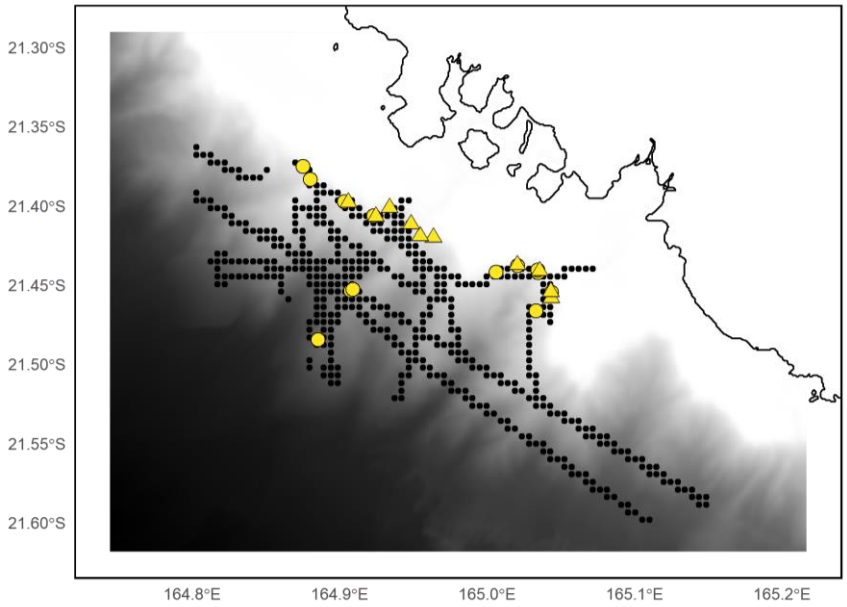


Figure S2. *Sampling design in site 2 (Poya-Nepoui). Grey shading indicate the depth category. BRUVS sampling stations (triangles), eDNA sampling stations (circles), and acoustic recordings (black dots).*

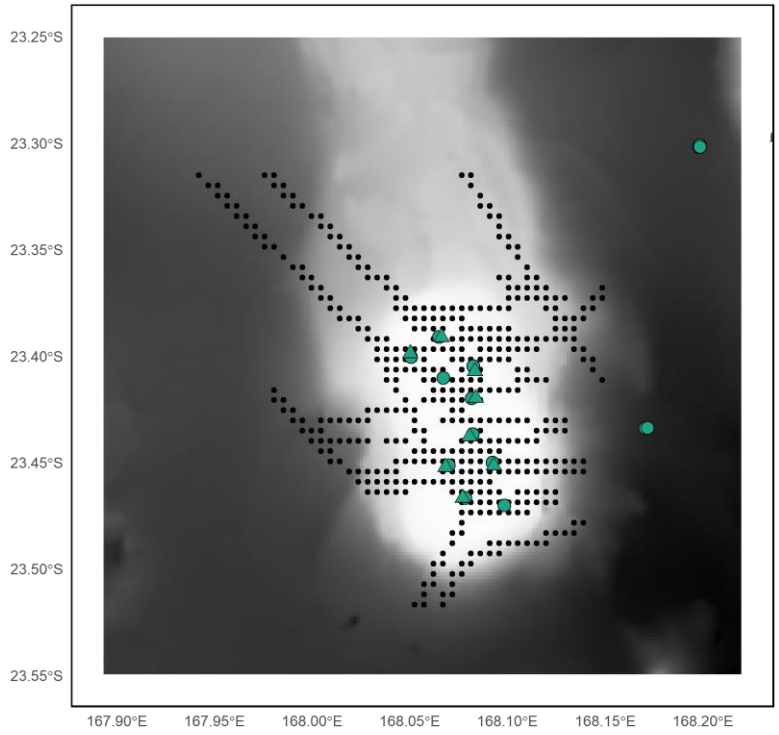


Figure S3. *Sampling design in site 3 (Antigonia). Grey shading indicate the depth category. BRUVS sampling stations (triangles), eDNA sampling stations (circles), and acoustic recordings (black dots).*

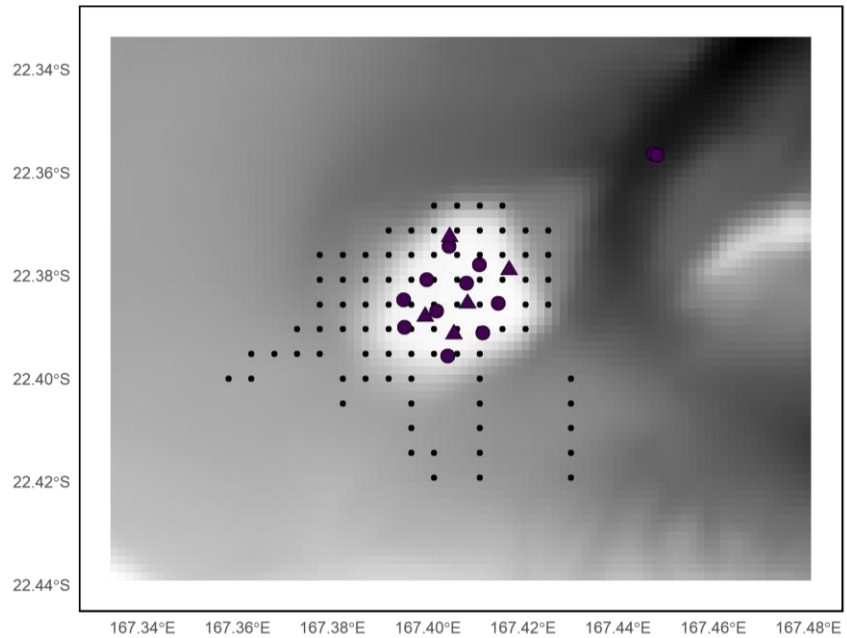


Figure S4. *Sampling design in site 4 (Ile des Pins). Grey shading indicate the depth category. BRUVS sampling stations (triangles), eDNA sampling stations (circles), and acoustic recordings (black dots).*

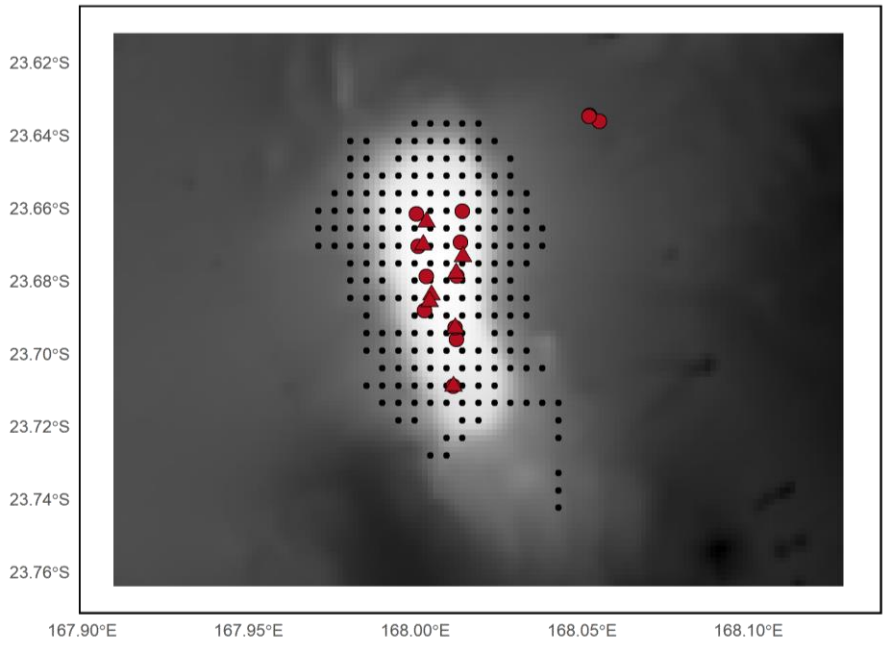


Figure S5. *Sampling design in site 5 (Jumeau West). Grey shading indicate the depth category. BRUVS sampling stations (triangles), eDNA sampling stations (circles), and acoustic recordings (black dots).*

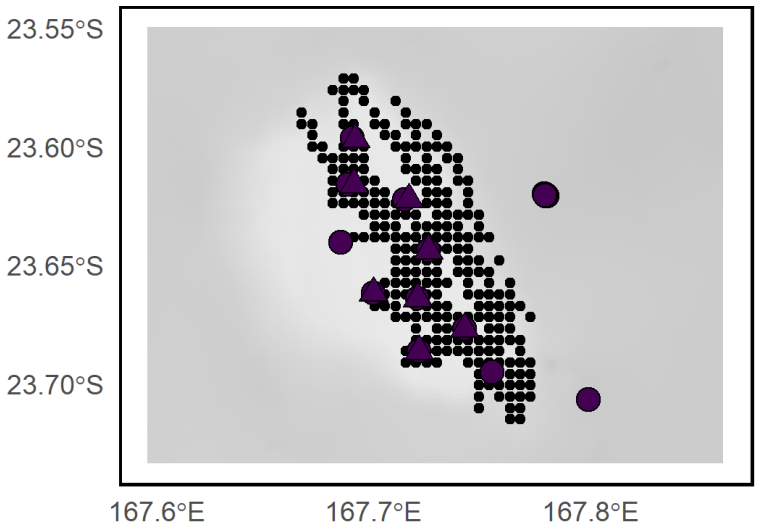


Figure S6. *Sampling design in site 6 (Stylaster). Grey shading indicate the depth category. BRUVS sampling stations (triangles), eDNA sampling stations (circles), and acoustic recordings (black dots).*

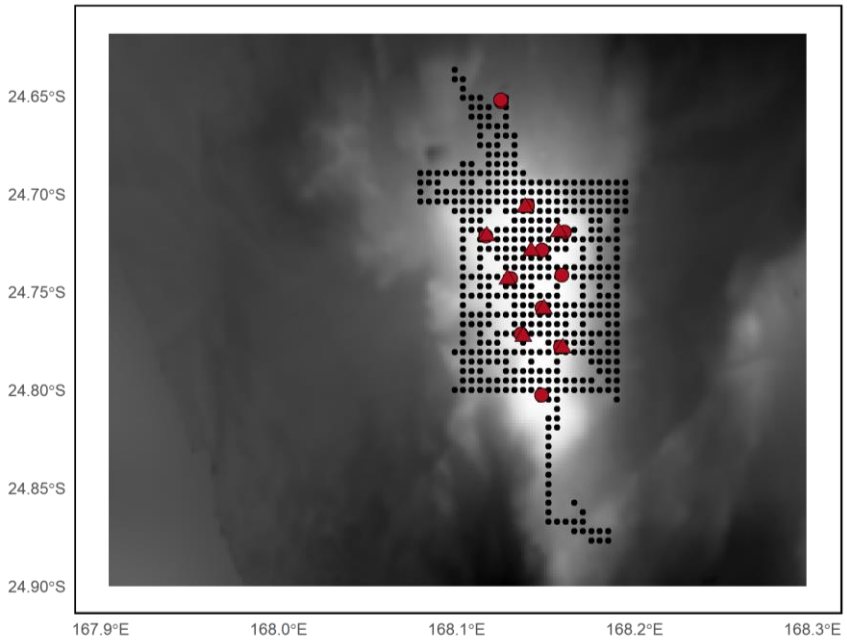


Figure S7. *Sampling design in site 7 (Kaimon Maru). Grey shading indicate the depth category. BRUVS sampling stations (triangles), eDNA sampling stations (circles), and acoustic recordings (black dots).*

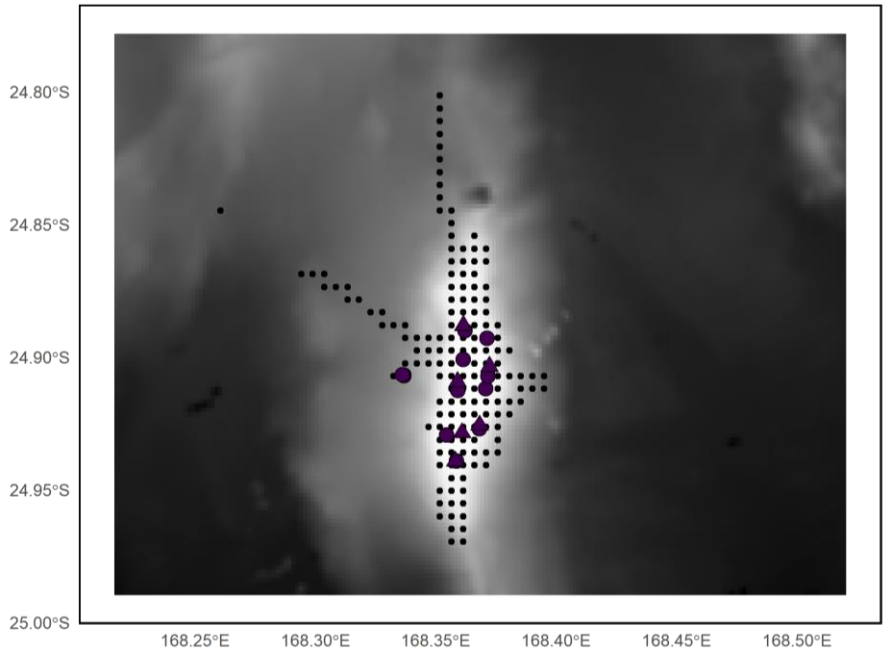


Figure S8. *Sampling design in site 8 (Eponge). Grey shading indicate the depth category. BRUVS sampling stations (triangles), eDNA sampling stations (circles), and acoustic recordings (black dots).*

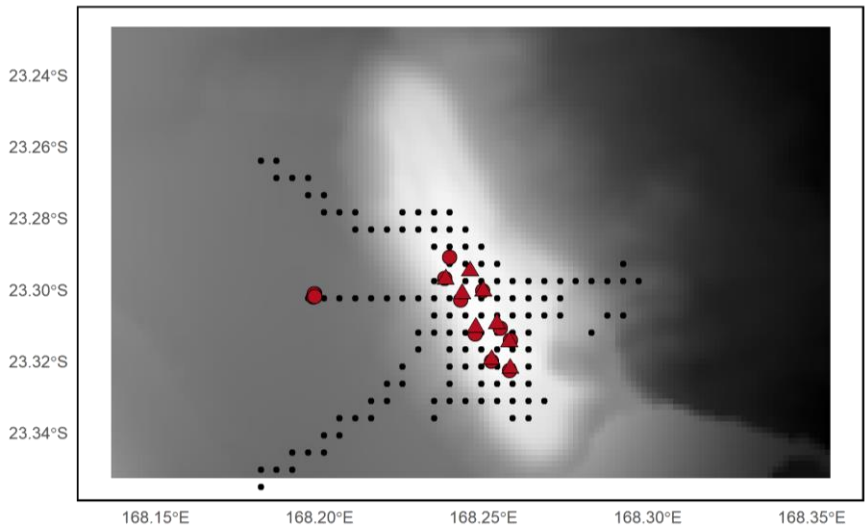


Figure S9. *Sampling design in site 9 (Crypthelia). Grey shading indicate the depth category. BRUVS sampling stations (triangles), eDNA sampling stations (circles), and acoustic recordings (black dots).*

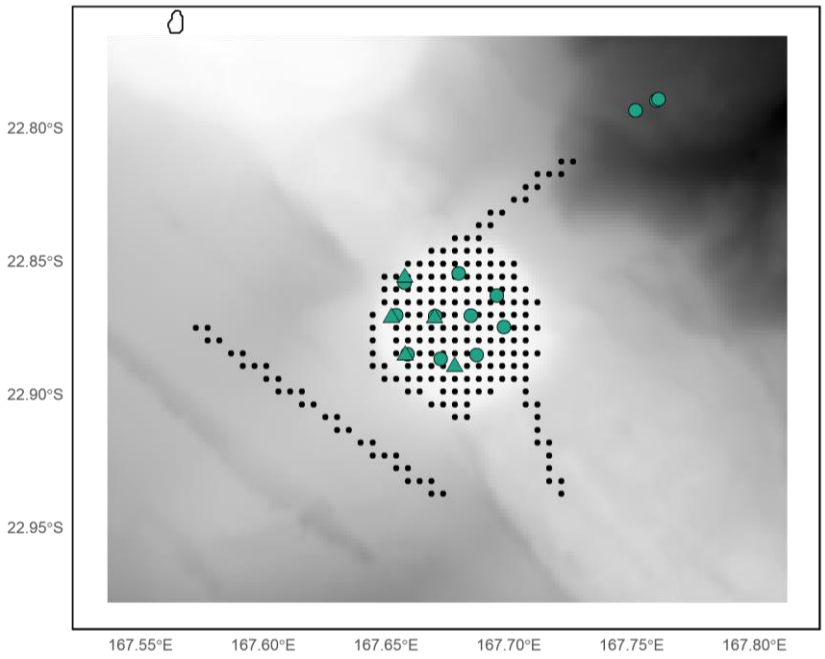


Figure S10. *Sampling design in site 10 (Torche). Grey shading indicate the depth category. BRUVS sampling stations (triangles), eDNA sampling stations (circles), and acoustic recordings (black dots).*

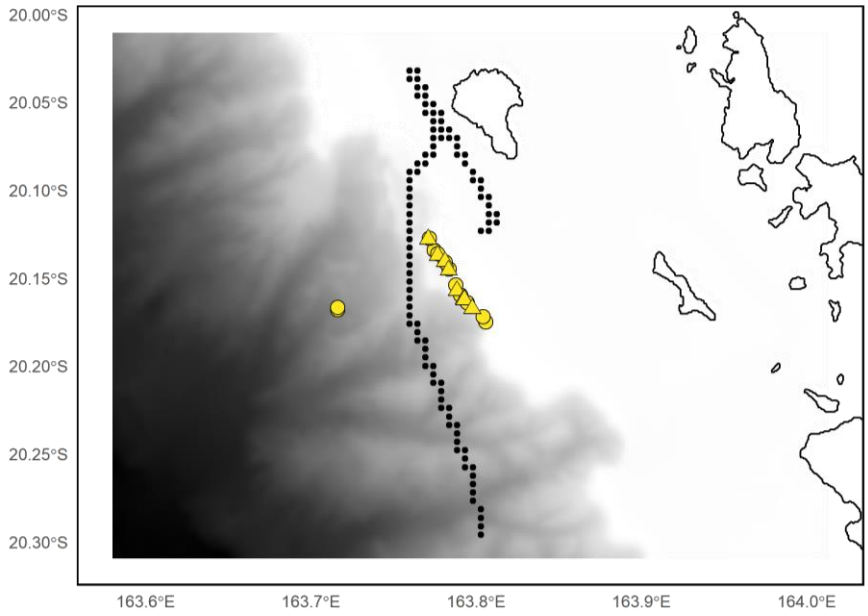


Figure S11. *Sampling design in site 11 (Poum). Grey shading indicate the depth category. BRUVS sampling stations (triangles), eDNA sampling stations (circles), and acoustic recordings (black dots).*

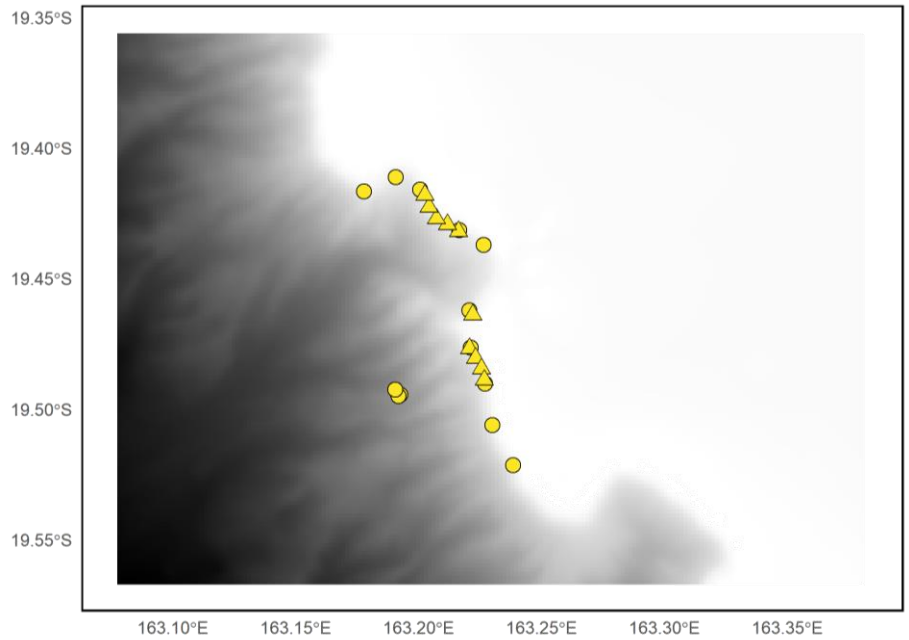


Figure S12. *Sampling design in site 12 (Great Northern Lagoon). Grey shading indicate the depth category. BRUVS sampling stations (triangles), eDNA sampling stations (circles), and acoustic recordings (black dots).*

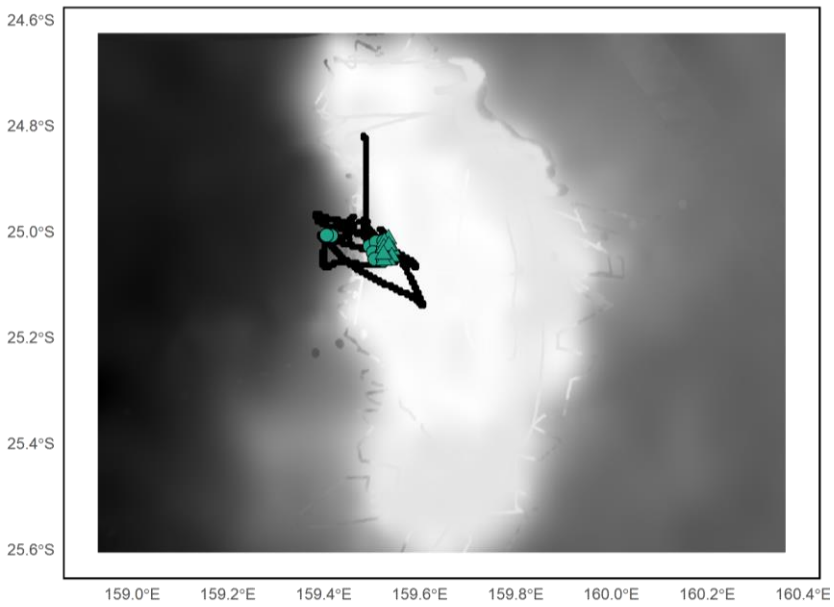


Figure S13. *Sampling design in site 13 (Capel). Grey shading indicate the depth category. BRUVS sampling stations (triangles), eDNA sampling stations (circles), and acoustic recordings (black dots).*

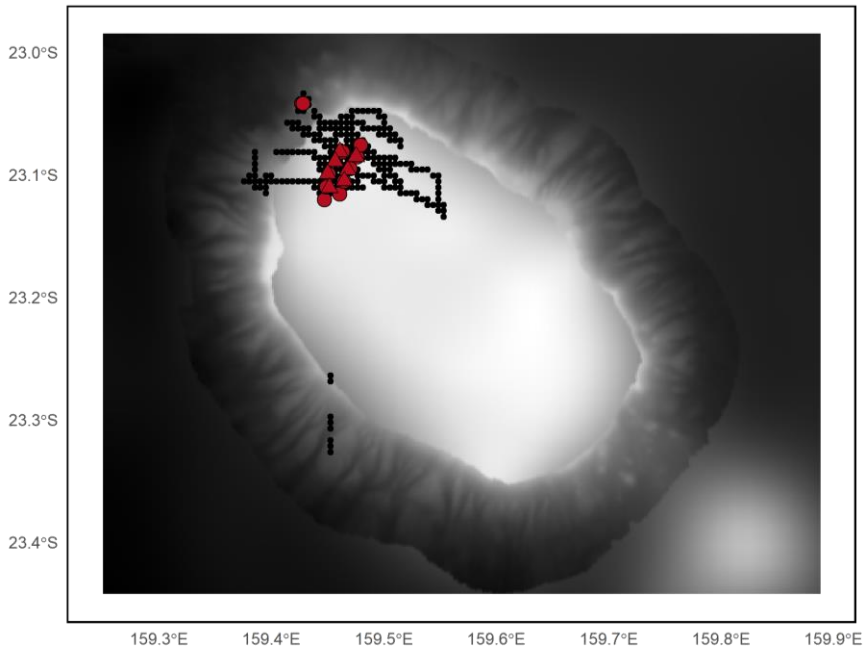


Figure S14. *Sampling design in site 14 (Argo). Grey shading indicate the depth category. BRUVS sampling stations (triangles), eDNA sampling stations (circles), and acoustic recordings (black dots).*

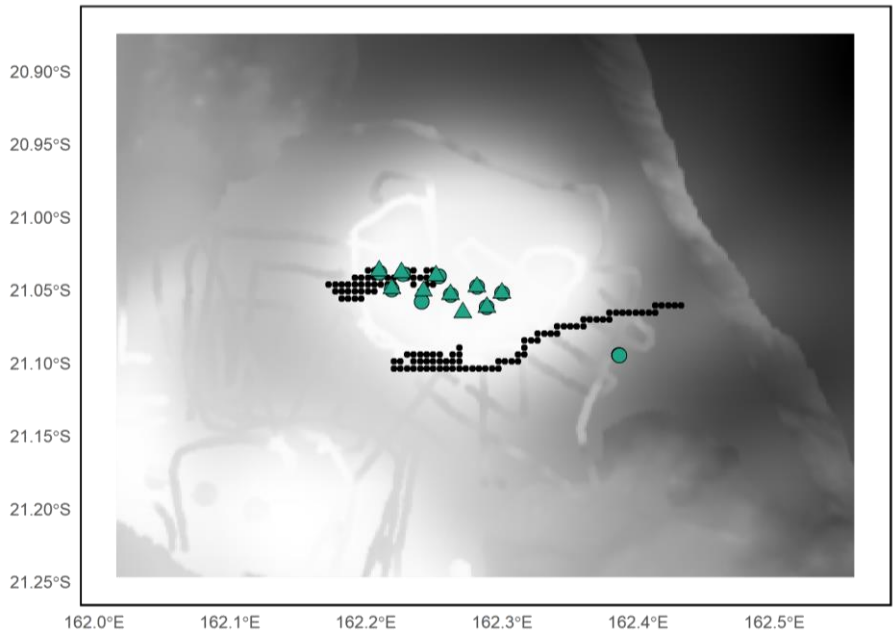


Figure S15. *Sampling design in site 15 (Fairway). Grey shading indicate the depth category. BRUVS sampling stations (triangles), eDNA sampling stations (circles), and acoustic recordings (black dots).*

5.2. Résultats supplémentaires

Table S3. Values (min, mean and max) of the 7 fish community metrics measured by the 3 sampling methods.

Method	Metric	Min	Mean	Max
BRUVS	Benthic fish richness	1	6.84 ± 8.46	42
	Benthic fish biomass	0.180 kg	113.28 ± 121	677 kg
	Benthic fish abundance	2	35.3 ± 79.9	754
eDNA	Benthic MOTU richness	0	13.26 ± 13.35	71
	Pelagic MOTU richness	0	7.9 ± 7.7	42
Acoustic	Benthic biomass	5.34	11.46 ± 19.72	29.48
	Pelagic biomass	1.62	2.85 ± 5.7	4.34

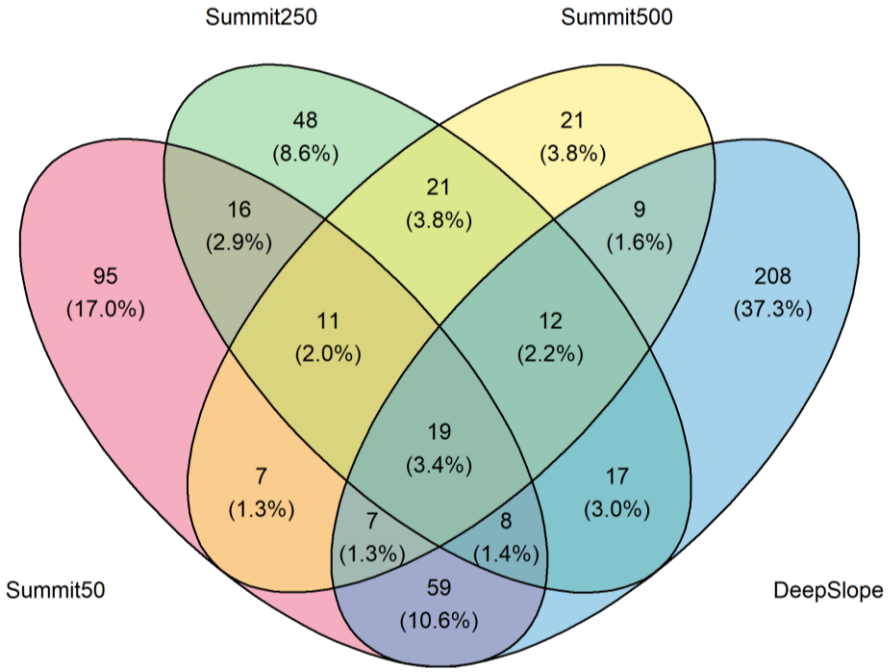


Figure S16. Venn diagram of MOTUs identified with eDNA. *Number and percentage of MOTUs unique to each habitat, and shared between two, three or four habitats.*

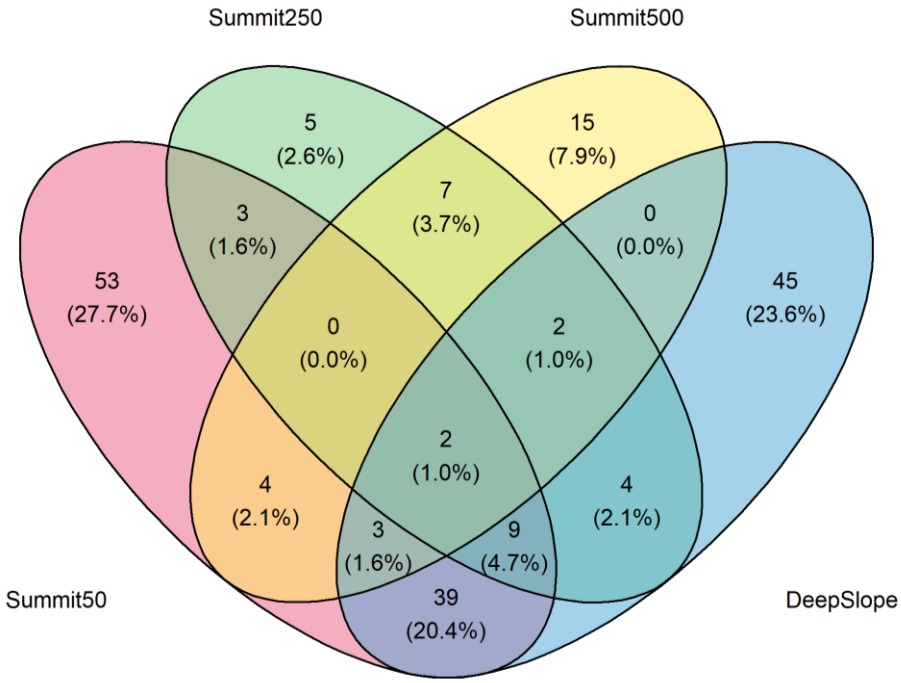


Figure S17. Venn diagram of species identified with BRUVS. *Number and percentage of species unique to each habitat, and shared between two, three or four habitats.*

Table S4. *MOTUs selected for the GJAM model on benthic MOTU read abundance by eDNA*

MOTU	Assignment	Min read number	Max read number	Mean read number
3da8f1176b92c07364bc3896edf66759fb0ab965	Myctophiformes	0	528526	15462.55
85278160ab227ef795d7d34b3b7f143ce48f85d0	Myctophidae	0	368637	10921.37
d6945c710869f4b37aa5130f8c13f2fffbf7b7d3	Clupeocephala	0	2264954	25560.08
ef40c96064831034fabfc6cf9c0c1e073e3e3515	Percomorphaceae	0	625223	11211.06
966b82c7f5deb4d86bb7245414bacad68917cfa8	Diaphus	0	1197125	16485.50
278fada4a105bc0ca041da419686a7486f08a66a	Percomorphaceae	0	458036	14024.04
89237aa306b52b84a560cc510d820aff82bc22aa	Naso	0	940494	9741.21
9eb94e56c2f5b5fadac929dc12d3520f642a1aa1	Beryx	0	1711722	35497.61
8ad697ca257e80c0e9a1fb4d180766d76fed37fc	Eupercaria	0	266410	7169.08
bc50ec564902f2a348db5a327290cdf47bd2ed15	Diogenichthys	0	299446	6718.53
d83357c786e0943ffd8df9fedfb0f11fd5f193cf	Parupeneus	0	97586	1843.98
0a68e7af285e1c8a09c4b3a1d1eb8358ae9993e9	Scomberomorus	0	672106	12110.83

Table S5. *MOTUs selected for the GJAM model on pelagic MOTU read abundance by eDNA*

MOTU	Assignment	Min read number	Max read number	Mean read number
198f251983a2c07127b0627155c2bfd1cb395018	Myctophidae	0	74986	1726.90
64bd1d960e40022a1495514b003ec2d6d6eff0d0	Diaphus	0	63446	2158.40
3ea7ca11fd53b49737dfe737333399f3fd6f3ba5	Lethrinus	0	138913	4381
88f21a755cd3ada743d261a320b227d3296c497a	Percomorphaceae	0	60721	1011.27
4da9175453088622f374a4c218d51abb164dce98	Thunnus	0	26575	1583.50
61be9061b5785556dd08d30ea76efb00d167bc3a	Lutjanus	0	46122	987.48
3b0c030da83544658146617d798382ff6ffa1aa0	Sardina pilchardus	0	78994	1638.37
278fada4a105bc0ca041da419686a7486f08a66a	Percomorphaceae	0	57919	2776.81
4138a9e4b074a38af19d74bfb1035179c89c9ede	Tylosurus	0	14785	408.62
0a68e7af285e1c8a09c4b3a1d1eb8358ae9993e9	Scomberomorus	0	8024	399.70

Table S6. *Species selected for the GJAM model on individual abundance by BRUVS*

Species	Min abundance	Max abundance	Mean abundance
Aphareus rutilans	0	12	0.65
Carcharhinus albimarginatus	0	6	0.616
Pristipomoides flavipinnis	0	8	0.616
Pristipomoides filamentosus	0	26	2.116
Seriola rivoliana	0	25	1.816
Carcharhinus plumbeus	0	4	0.308
Gymnosarda unicolor	0	2	0.125
Lethrinus miniatus	0	24	1.375
Lethrinus rubrioperculatus	0	13	0.366
Gymnocranius euanus	0	13	1.208
Bodianus bimaculatus	0	3	0.116
Epinephelus chlorostigma	0	8	0.408
Wattsia mossambica	0	10	0.341
Aprion virescens	0	5	0.383
Carangoides orthogrammus	0	50	0.608
Pseudocaranx dentex	0	41	1.291
Seriola lalandi	0	10	0.325
Squalus megalops	0	10	0.866
Epinephelus morrhua	0	2	0.158
Etelis coruscans	0	12	0.425
Pristipomoides argyrogrammicus	0	4	0.325
Polymixia japonica	0	5	0.225
Pentaceros richardsoni	0	3	0.075

Table S7. *Details on boosted regression trees (BRT) model parameters and goodness-of-fit*

Response variable	**Explanatory variables**	**Tree complexity**	**Learning rate**	**Bag Fraction**	**Number of trees**	**Cross-validation correlation**
BRUVS species richness	SummitRugosity, BottomDepth, SSTmean, Salinity	1	0.005	0.75	1050	0.70 ± 0.05
BRUVS biomass	TravelTime, SSTmax, BottomDepth, Salinity, seafloorTemp, Suspended Particulate Matter	5	0.005	0.75	1600	0.85 ± 0.02
BRUVS abundance	BottomDepth, SSTmax, EastwardVelocity	5	0.001	0.75	2350	0.62 ± 0.05
eDNA MOTU richness benthic	TravelTime, SSTmax, Salinity, Chla	2	0.001	0.5	1425	0.42 ± 0.07
eDNA MOTU richness pelagic	SummitRugosity, EastwardVelocity, NorthwardVelocity, Sampling_Depth, Salinity, seafloorTemp, LandMinDist	2	0.005	0.75	1525	0.60 ± 0.09
Benthic acoustic biomass	LandMinDist, SSTmean, BottomDepth, Chla, EastwardVelocity, NorthwardVelocity	5	0.01	0.5	3650	0.59 ± 0.01
Pelagic acoustic biomass	Sampling_Depth, TravelTime, SSTmean	5	0.01	0.75	10000	0.76 ± 0.001

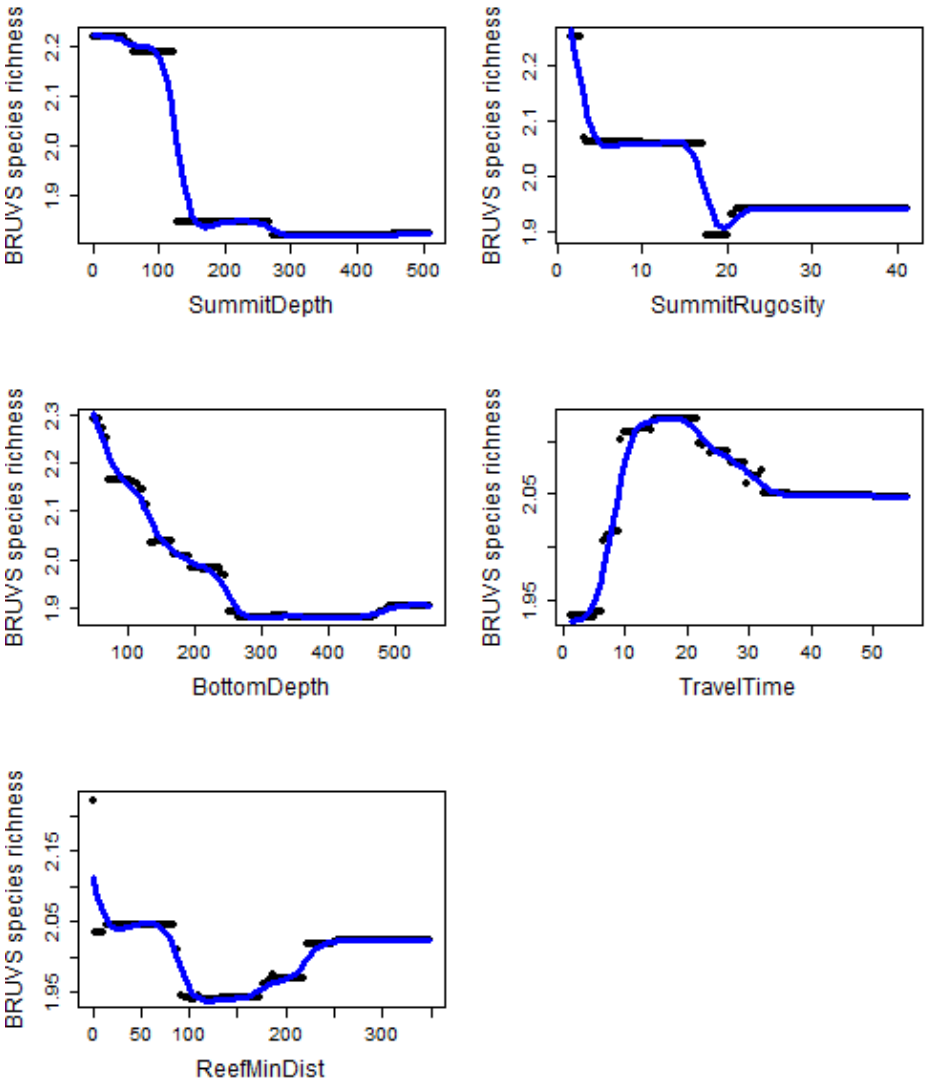


Figure S18. *Partial relationships between explanatory variables and BRUVS species richness in BRTs.*

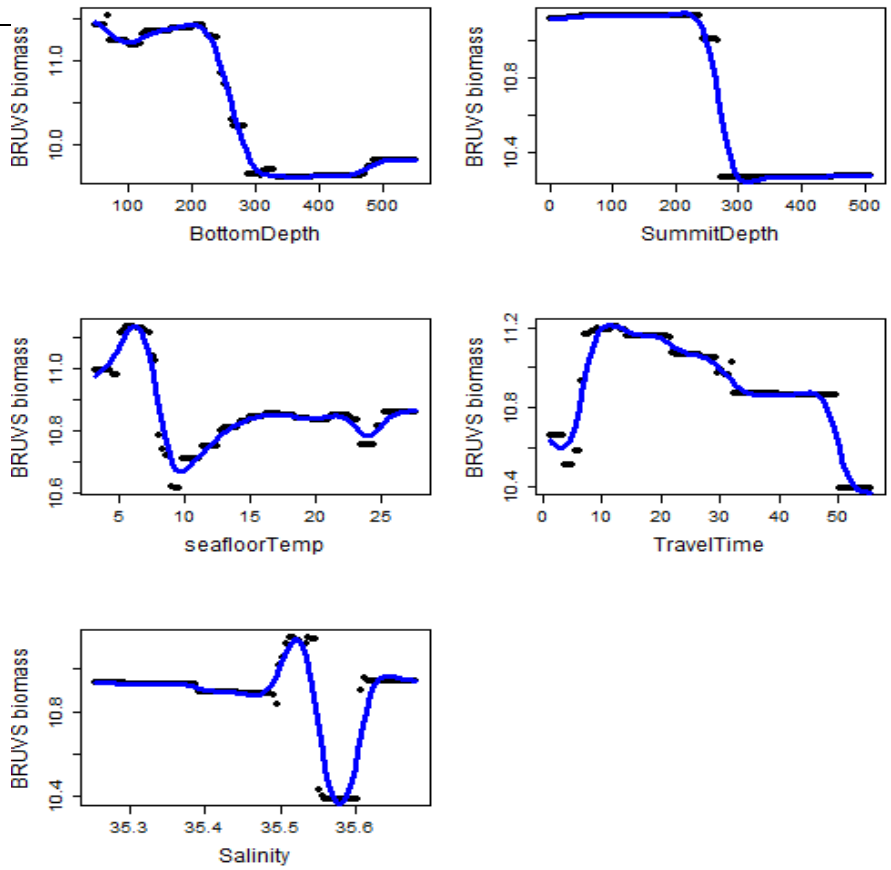


Figure S19. *Partial relationships between explanatory variables and BRUVS biomass (log-transformed) in BRTs.*

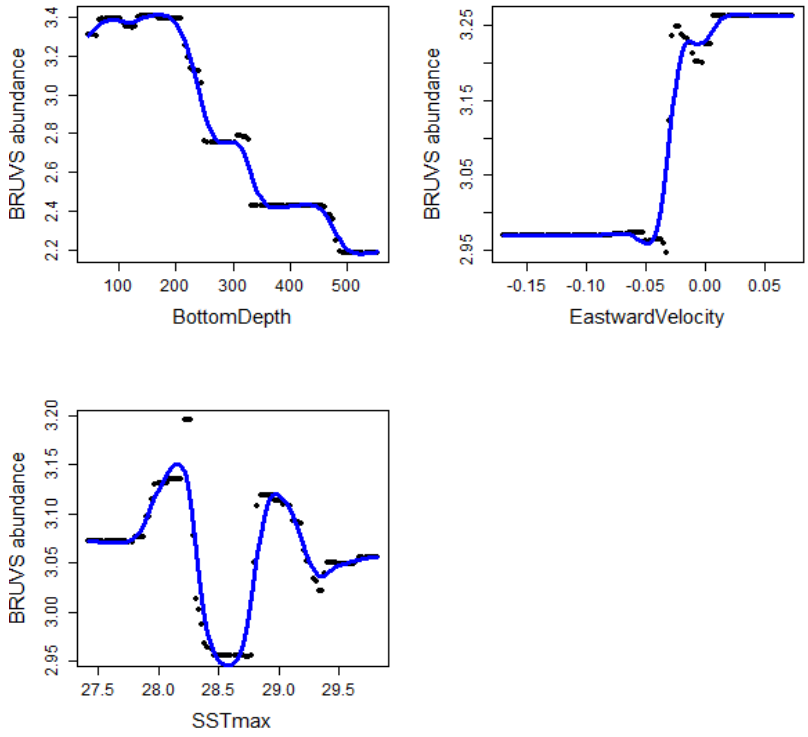


Figure S20. *Partial relationships between explanatory variables and BRUVS abundance (log-transformed) in BRTs.*

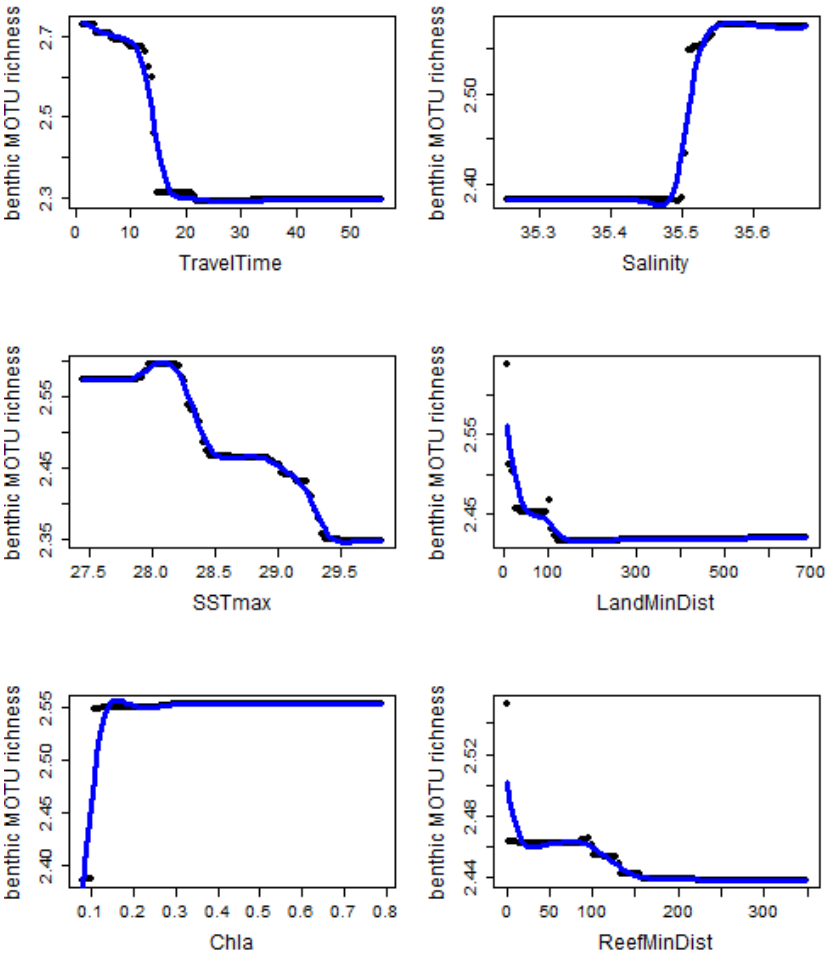


Figure S21. *Partial relationships between explanatory variables and benthic MOTU richness in BRTs.*

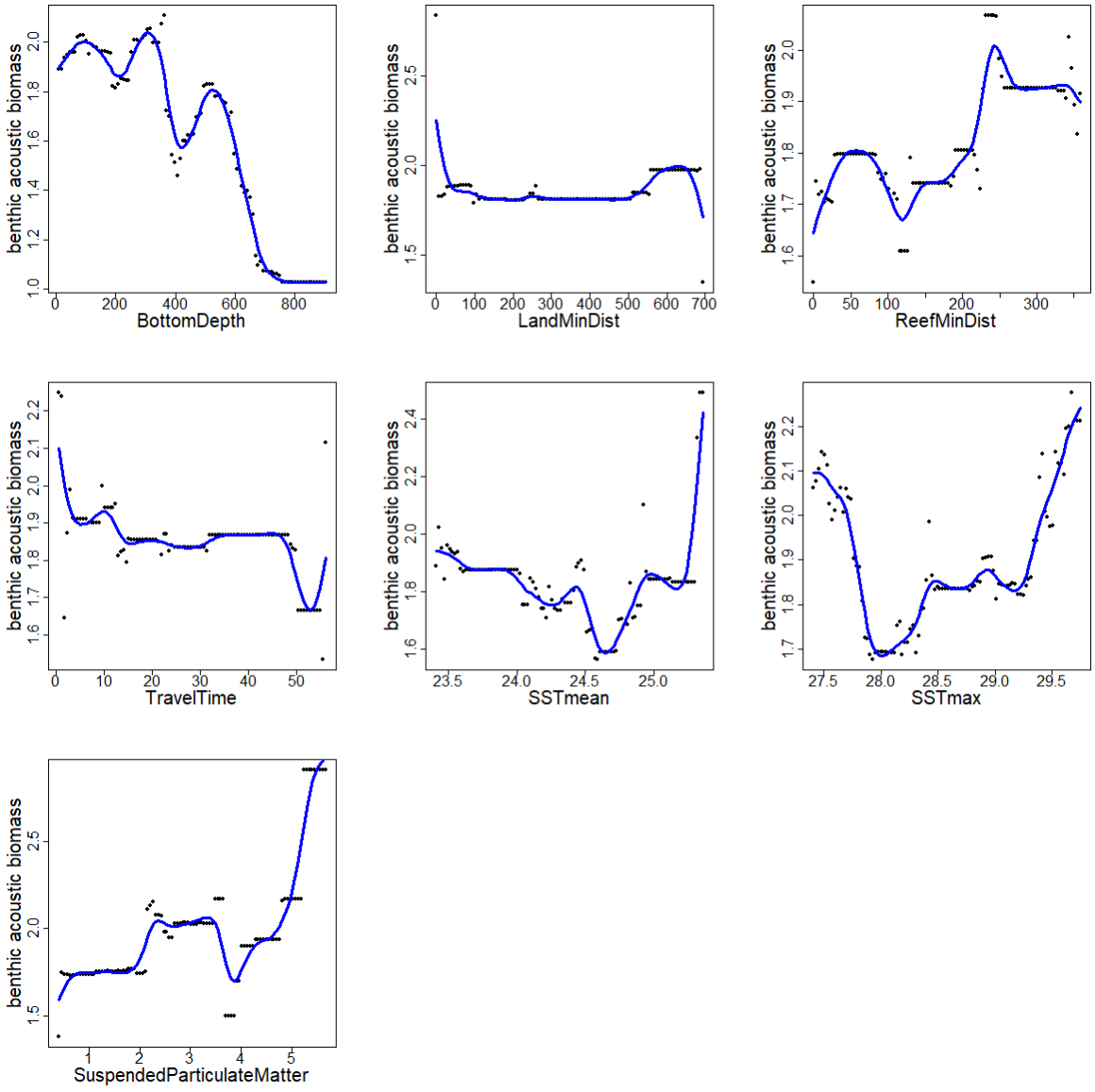


Figure S22. *Partial relationships between explanatory variables and benthic acoustic biomass in BRTs.*

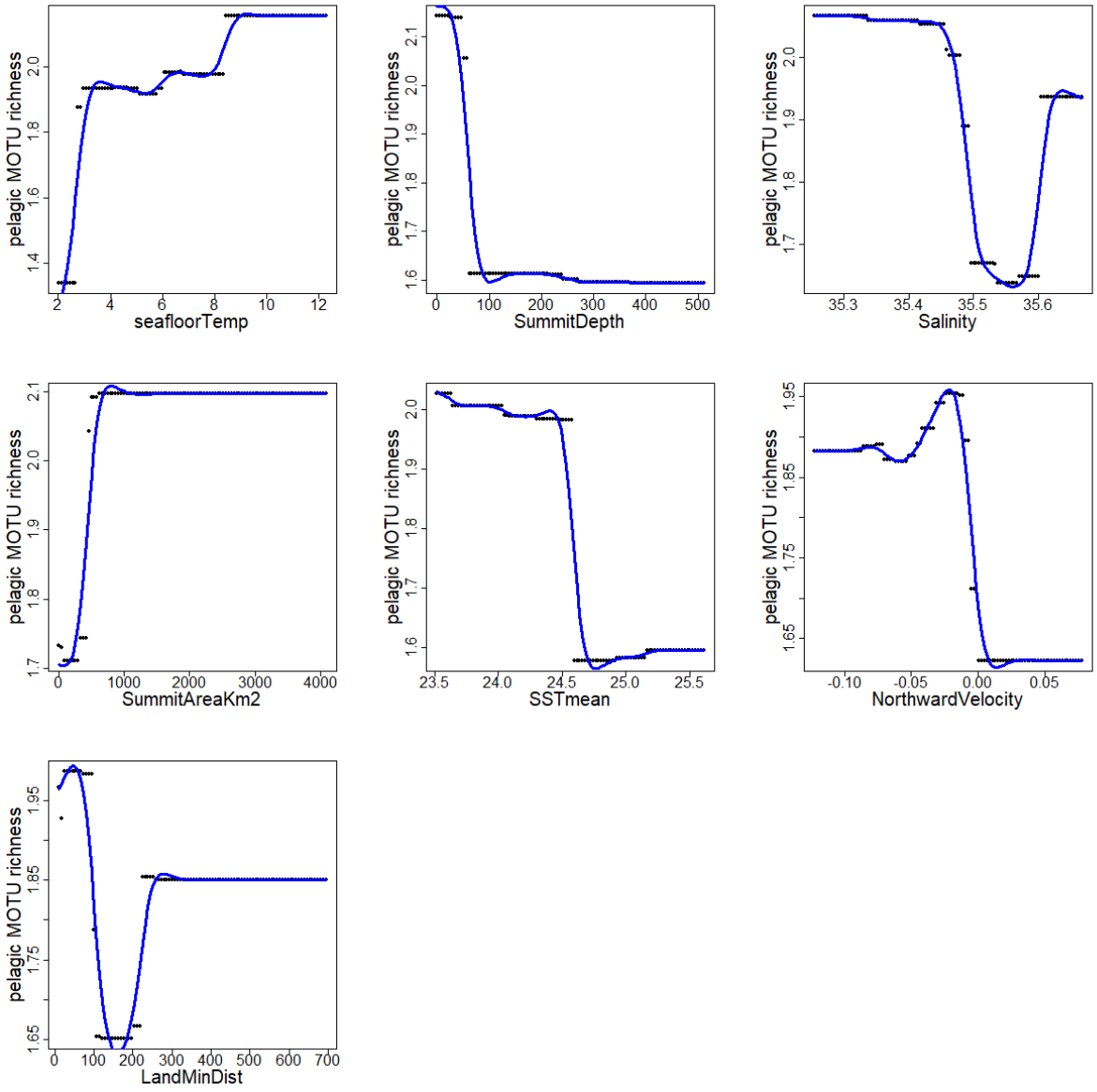


Figure S23. *Partial relationships between explanatory variables and pelagic MOTU richness in BRTs.*

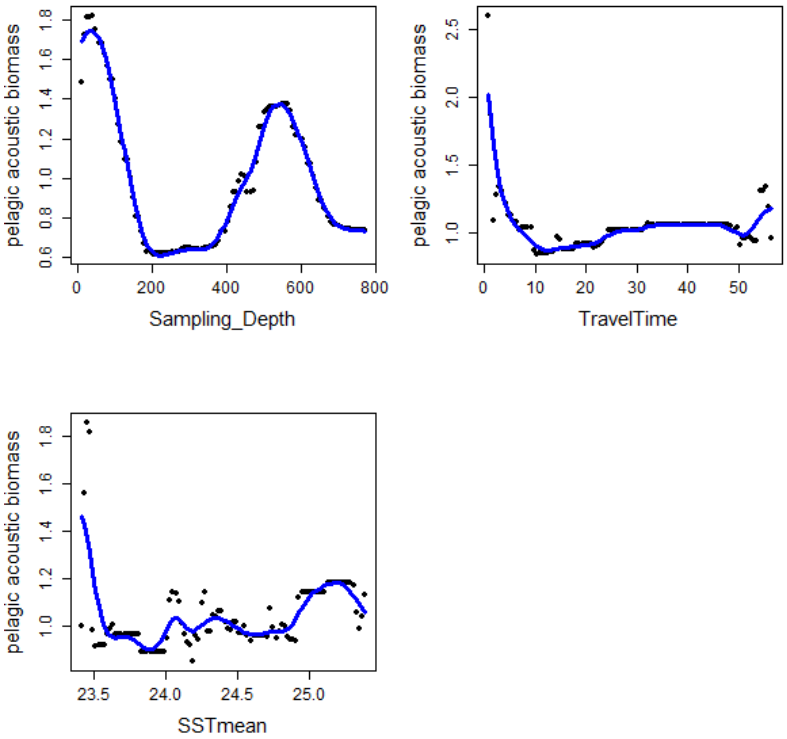


Figure S24. *Partial relationships between explanatory variables and pelagic acoustic biomass in BRTs.*

Table S8. *Details on generalized joint attribute models (GJAM) goodness-of-fit*

Response variable	Explanatory variables	DIC	Pearson r
BRUVS species abundances	SSTmean^3, BottomDepth^2, EastwardVelocity^3, LandMinDist^2, Salinity^2, TravelTime, Habitat, SummitRugosity	22469	0.62
eDNA benthic MOTU read number	SSTmean^2, BottomDepth^2, ReefMinDist, TravelTime, SuspendedParticulateMatter^2, Salinity^3, NorthwardVelocity^3, Habitat, SummitRugosity^2	4207	0.63
eDNA pelagic MOTU read number	SSTmean, SuspendedParticulateMatter, Habitat, Salinity^2, TravelTime, ReefMinDist, seafloorTemp^3, SummitRugosity^2, SummitAreaKm², NorthwardVelocity^2, EastwardVelocity, SamplingDepth, BottomDepth	1491	0.67

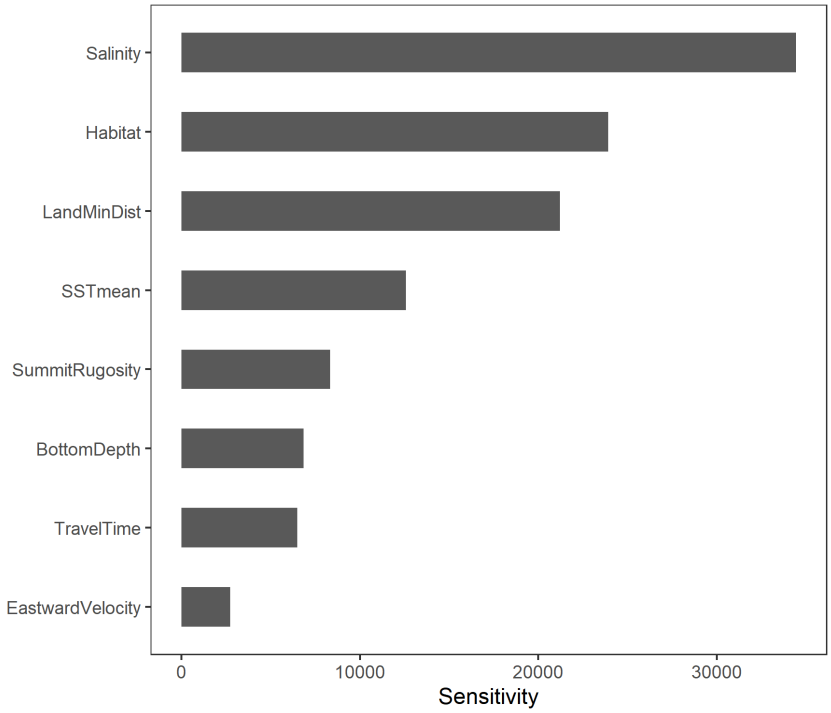


Figure S25. *Sensitivity of species abundance measured by BRUVS to each variable (sum of linear, quadratic and cubic terms when applicable) in the GJAM model.*

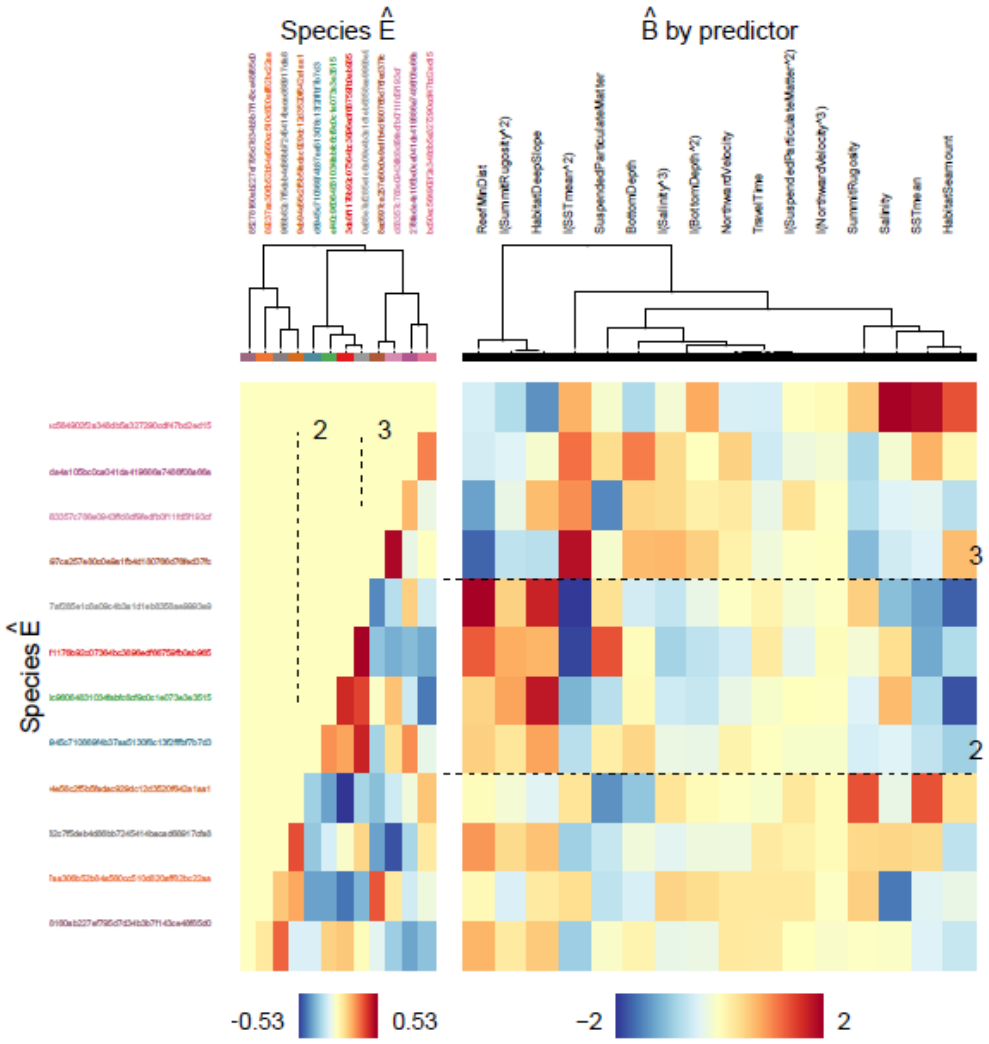


Figure S26. Cluster and grid plot of fitted benthic MOTU read number and predictors. *Left panel represents correlation of MOTUs with one another and clustering of species highly correlated. Right panel represents the correlation of each MOTU with predictors, and clustering in three groups. Red indicate strong positive correlation, and blue indicate strong negative correlation.*

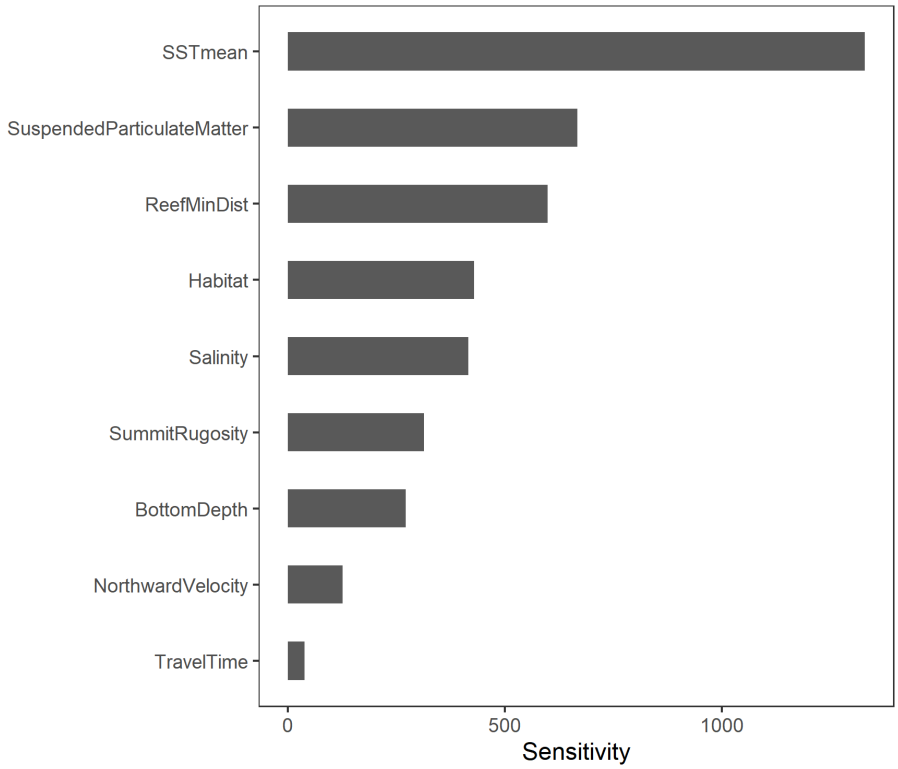


Figure S27. *Sensitivity of benthic MOTU read number measured by eDNA to each variable (sum of linear, quadratic and cubic terms when applicable) in the GJAM model.*

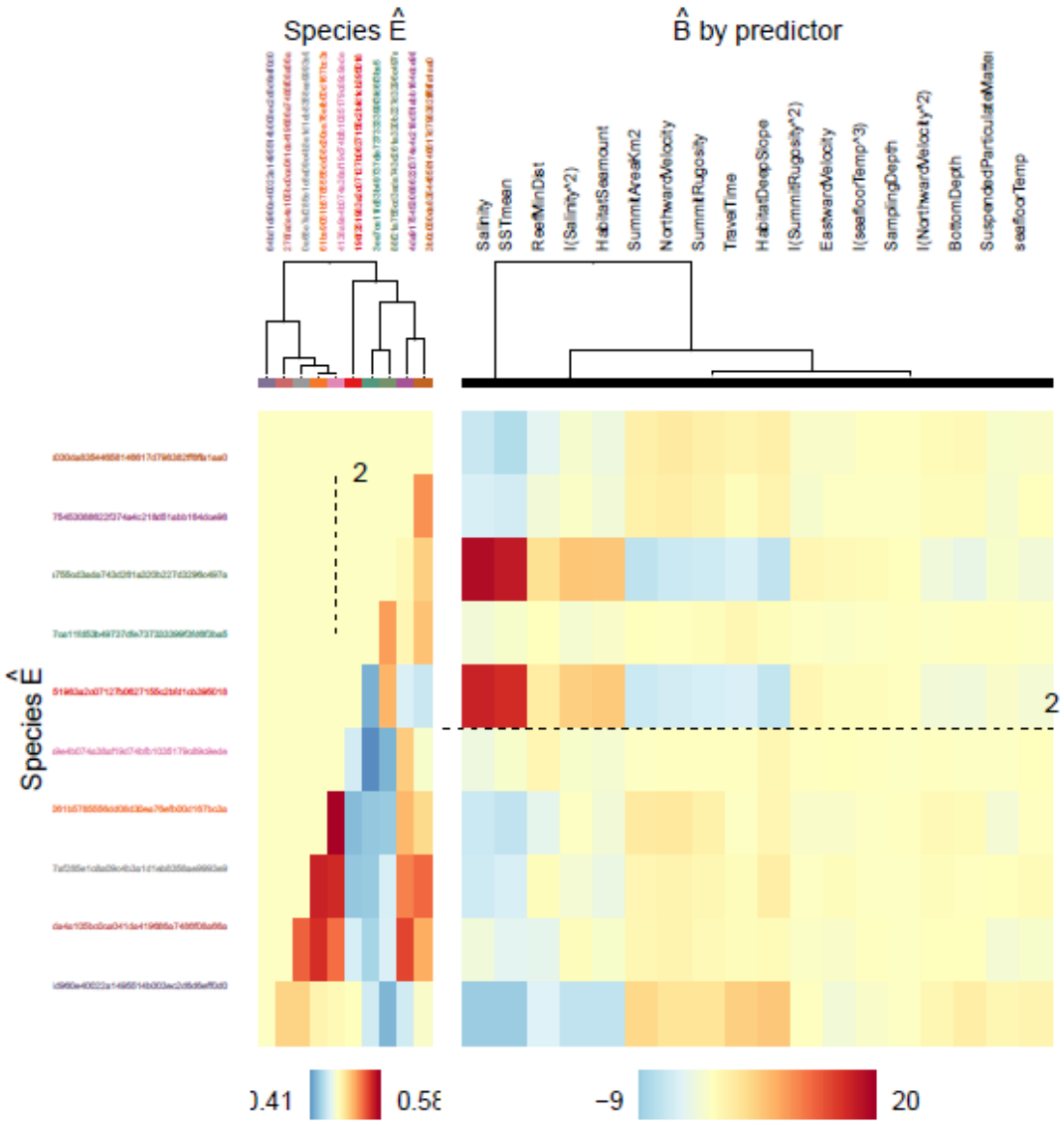


Figure S28. Cluster and grid plot of fitted pelagic MOTU read number and predictors. *Left panel represents correlation of MOTUs with one another and clustering of species highly correlated. Right panel represents the correlation of each MOTU with predictors, and clustering in two groups. Red indicate strong positive correlation, and blue indicate strong negative correlation.*

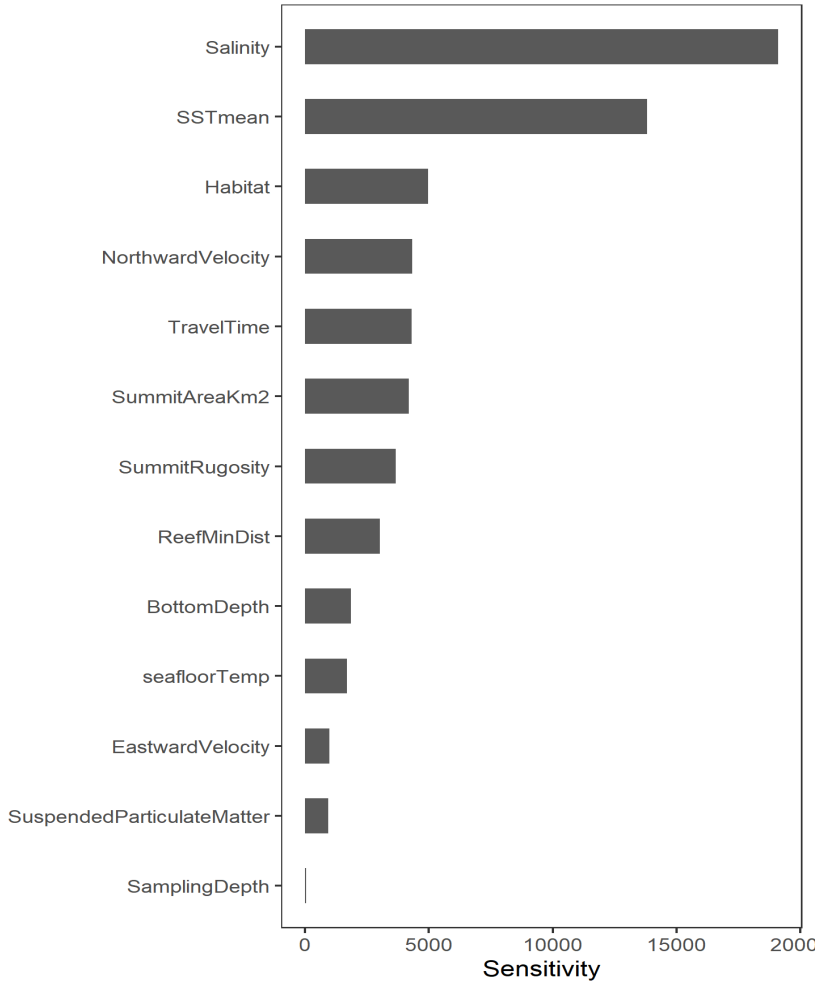


Figure S29. *Sensitivity of pelagic MOTU read number measured by eDNA to each variable (sum of linear, quadratic and cubic terms when applicable) in the GJAM model.*

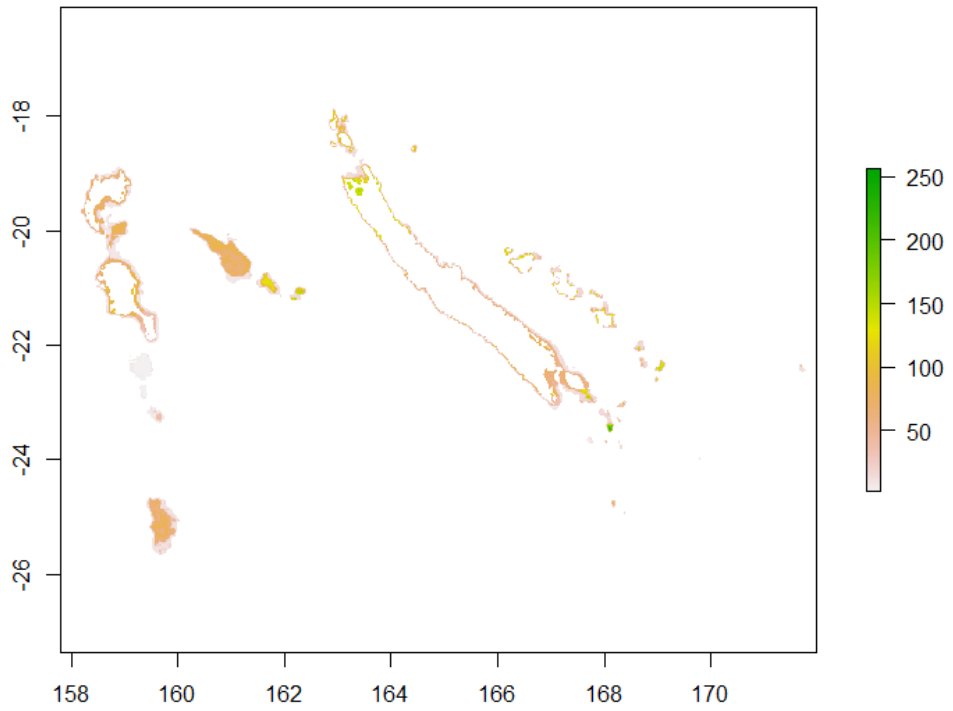


Figure S30. Prediction of total fish biomass measured by BRUVS (in Kg), *from the fitted values of the BRT model, in all seamounts and deep slopes of the New-Caledonian ZEE, down to 600 m deep.*

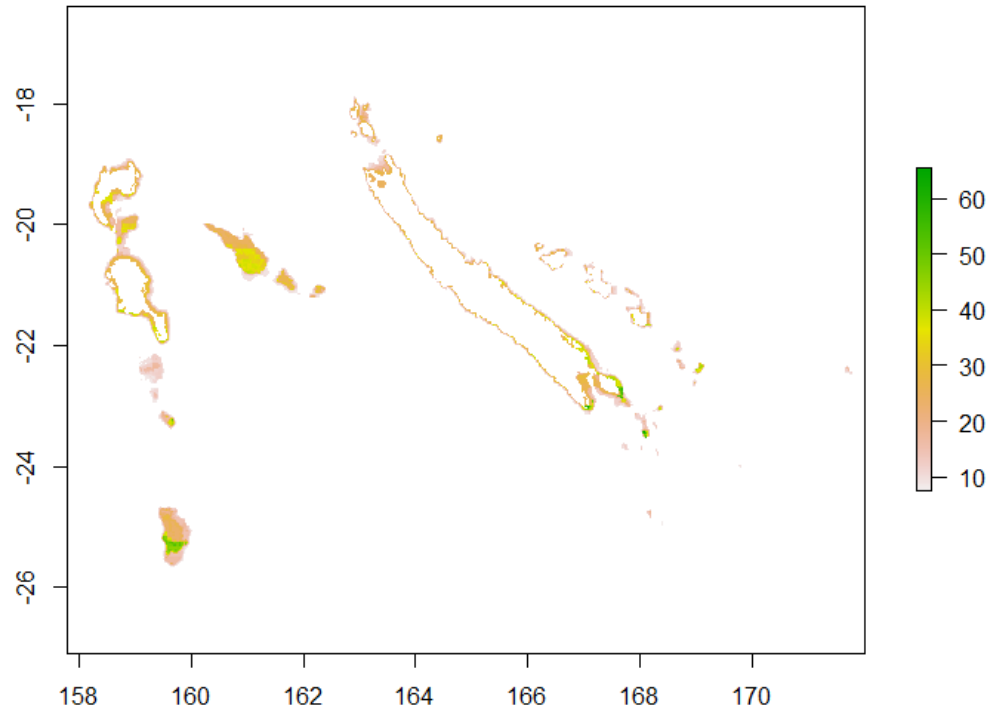


Figure S31. Prediction of total fish abundance measured by BRUVS, *from the fitted values of the BRT model, in all seamounts and deep slopes of the New-Caledonian ZEE, down to 600 m deep.*

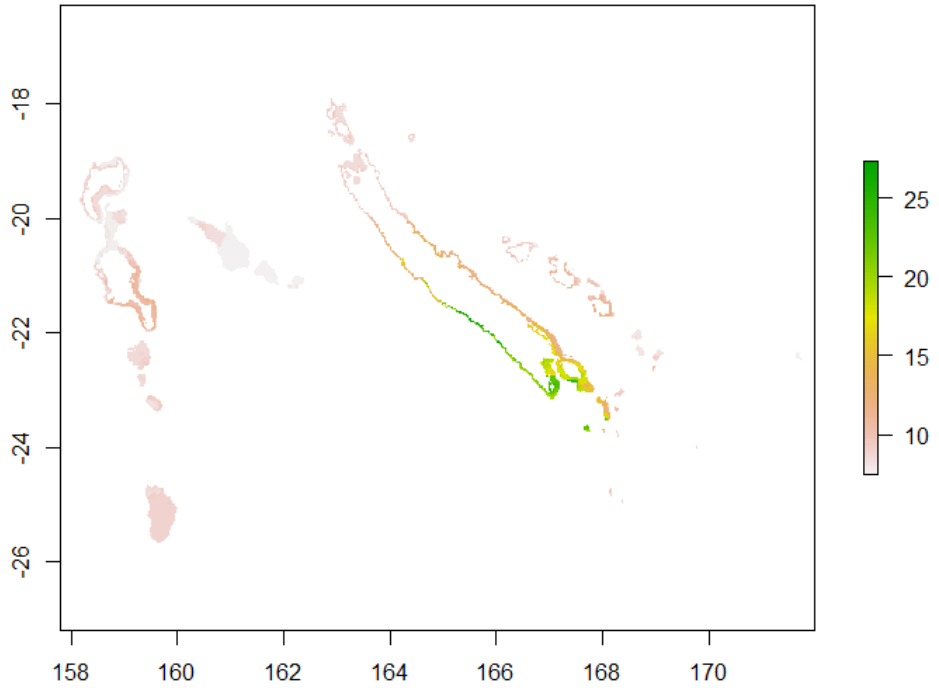


Figure S32. Prediction of benthic MOTU richness measured by eDNA, *from the fitted values of the BRT model, in all seamounts and deep slopes of the New-Caledonian ZEE, down to 600 m deep.*

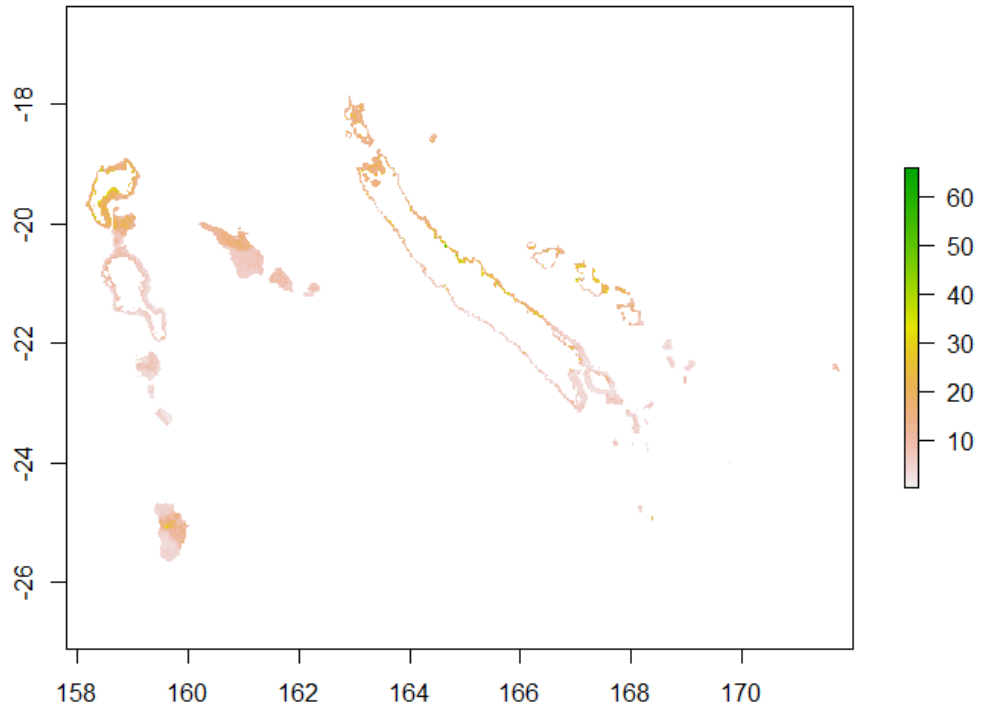


Figure S33. Prediction of benthic acoustic biomass, *from the fitted values of the BRT model, in all seamounts and deep slopes of the New-Caledonian ZEE, down to 600 m deep.*

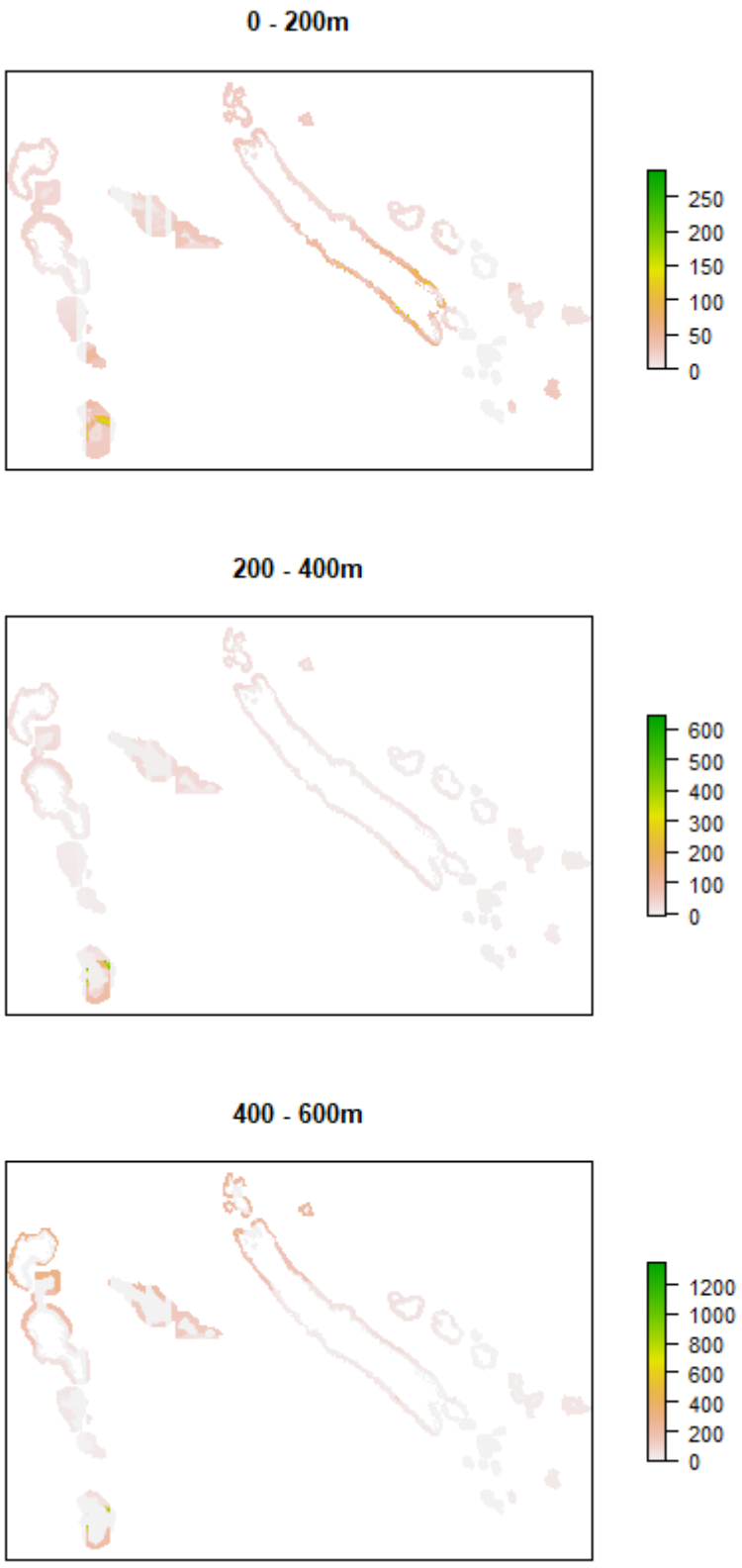


Figure S34. Prediction of pelagic acoustic biomass, *from the fitted values of the BRT model, in all seamounts and deep slopes of the New-Caledonian ZEE, down to 600 m deep, in the three depth layers: 0-200 m, 200-400 m and 400-600 m.*

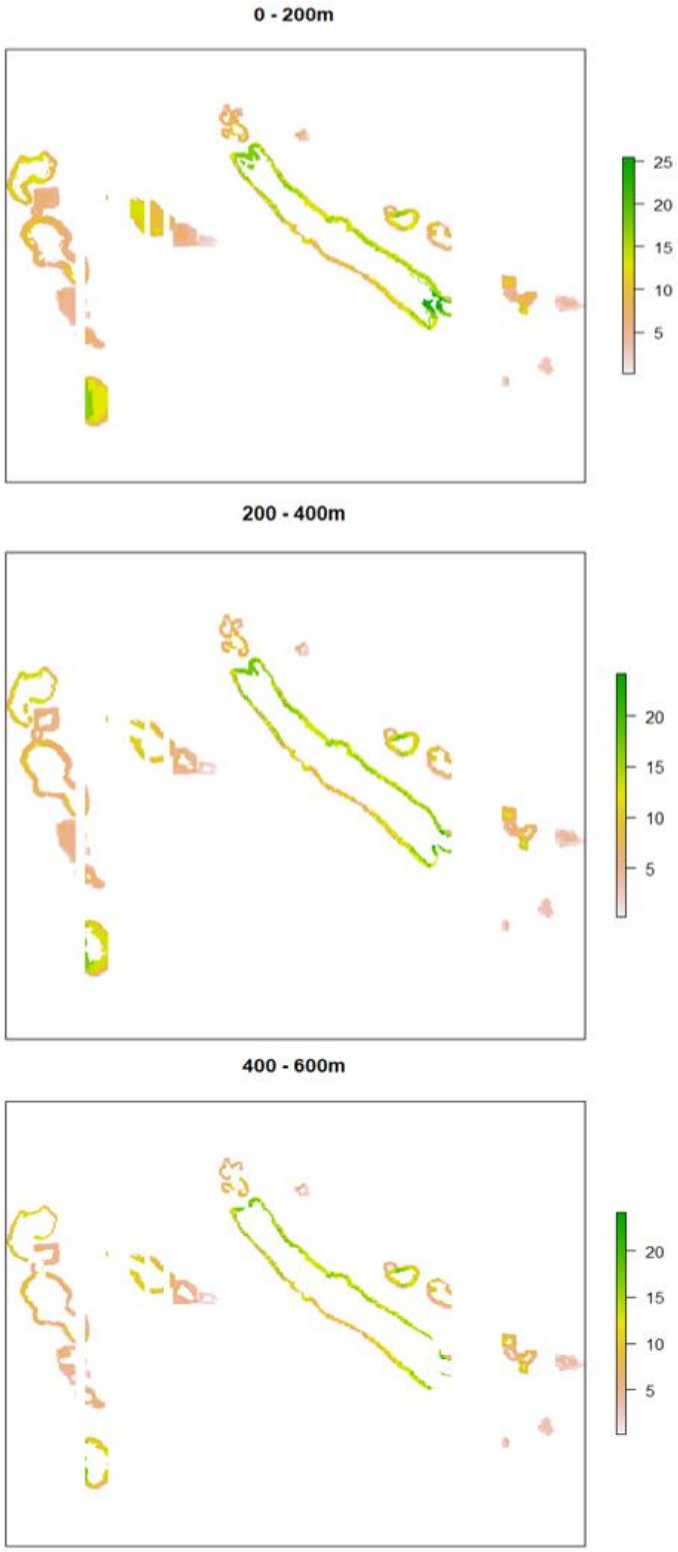


Figure S35. Prediction of pelagic MOTU richness, *from the fitted values of the BRT model, in all seamounts and deep slopes of the New-Caledonian ZEE, down to 600 m deep, in the three depth layers: 0-200 m, 200-400 m and 400-600 m.*

Table S9. *Ranking of the different solutions of prioritization computed with different boundary length modifier values (blmval), by the TOPSIS method based on their total cost and total boundary length. Ranking with wide BLM values range.*

blmval	Total boundary length	Total cost	score	rank
0	91,070,476	60,241	0.05492	8
1,00E-05	61,704,676	60,258	0.56533	7
1,00E-04	54,572,208	61,537	0.70081	6
1,00E-03	39,640,752	63,590	0.94621	2
1,00E-02	39,210,864	63,791	0.94475	5
1,00E-01	39,066,082	63,753	0.94556	3
1	39,008,162	63,636	0.94734	1
10	39,075,004	63,757	0.94549	4

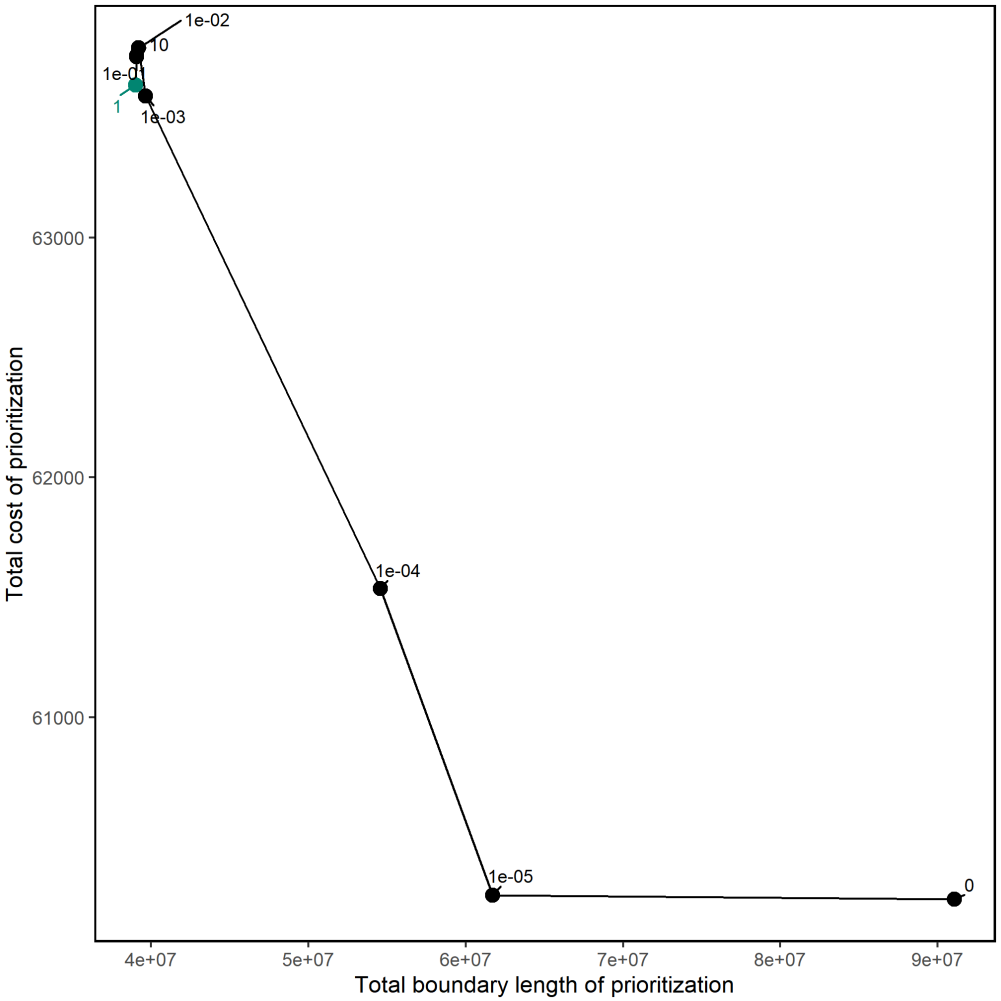


Figure S36. *Prioritization solutions acording to their total cost and total boundary length. Labels indicate the boundary length modifier values (BLM). Green dot indicate the best solution.*