Etude du cryptobiome récifal des Mascareignes : apports des mini-récifs artificiels (ARMS) couplés aux approches d'écologie moléculaire

Marion Couëdel

Université de La Réunion

Ecole doctorale Sciences, Technologies et Santé

THÈSE

Présentée pour l'obtention du grade de docteur de l'Université de La Réunion

Discipline : Écologie marine

Etude du cryptobiome récifal des Mascareignes : apports des mini-récifs artificiels (ARMS) couplés aux approches d'écologie moléculaire

---

Marion COUËDEL

Soutenue le 27 juin 2023 devant le jury composé de :

Agnès BOUCHEZ	*DR, INRAE*	**Rapporteure**
Anne CHENUIL	*DR, CNRS*	**Rapporteure / Présidente du jury**
Erwan DELRIEU-TROTTIN	*MCF, EPHE*	**Examinateur**
Henrich BRUGGEMANN	*PR, Univ. de La Réunion*	**Directeur de thèse**
Agnès DETTAI	*MCMU, MNHN*	**Co-directrice de thèse**
Mireille M.M. GUILLAUME	*MCMU HDR, MNHN*	**Co-directrice de thèse**

La seule limite à nos réalisations de demain

sera nos doutes d'aujourd'hui.

F. D. Roosevelt

Résumé

Dans le contexte du changement global, la surveillance de la biodiversité est essentielle pour détecter les facteurs de perturbation et comprendre les réponses des communautés. Les récifs coralliens font partie des écosystèmes les plus riches et diversifiés de la planète, mais également des plus menacés. Pourtant, les organismes, souvent de petite taille, vivant cachés dans les anfractuosités du récif et constituant le cryptobiome, restent très peu étudiés alors qu'ils représentent la majeure partie de la diversité récifale et un maillon fondamental du réseau trophique. Si l'évaluation de la diversité est essentielle pour une gestion efficace des écosystèmes, les méthodes traditionnelles de taxonomie, basées sur la morphologie, sont peu adaptées au cryptobiome en étant chronophages, nécessitant des connaissances spécialisées et difficilement comparables entre sites. Ces petits taxons présentent des défis supplémentaires en étant plus difficiles à trouver et à identifier. Avec le métabarcoding d'ADN, les récents progrès des techniques de séquençage à haut-débit et de la bio-informatique offrent une alternative aux méthodes traditionnelles.

Cette thèse porte sur la diversité et l'écologie du cryptobiome récifal des Mascareignes (La Réunion et Rodrigues), collecté à l'aide de structures d'échantillonnage standardisées, les ARMS, et étudié par l'approche métabarcoding. Dans un premier temps, un pipeline bio-informatique adapté aux analyses métabarcoding de métazoaires a été élaboré pour évaluer les communautés échantillonnées. Les interprétations écologiques découlant de l'approche par métabarcoding sont très dépendantes des références moléculaires disponibles. C'est pourquoi, dans un second temps, ces travaux se sont attachés à initier un référentiel moléculaire pour documenter le cryptobiome des Mascareignes, en produisant des séquences pour trois marqueurs moléculaires, le 18S, le COI et le 16S, ainsi que le mitogénome complet pour l'ensemble des téléostéens échantillonnés. Dans un troisième temps, ces travaux ont évalué la diversité du cryptobiome et sa cinétique de colonisation à travers les saisons chaude et fraîche à La Réunion. Les taxons retrouvés correspondent aux taxons du cryptobiome observé dans d'autres régions du monde et reflète la grande diversité taxonomique observée dans les récifs coralliens. Ces travaux sont les premiers mettre en évidence des différences significatives dans la composition des communautés en fonction de la durée d'immersion des ARMS, de la saison de déploiement et de récolte. Ces facteurs influencent en particulier les taxons sessiles tels que les ascidies, les cnidaires, les porifères et les rhodophytes. Dès lors, ces travaux démontrent la nécessité de considérer ces variations dans l'analyse des résultats et leurs comparaisons avec d'autres études, ainsi que leurs implications dans l'utilisation des ARMS en tant qu'outil de suivi. Dans un dernier temps, la méthode des ARMS couplée aux méthodes d'identification moléculaire pour évaluer la diversité et la distribution des espèces a été examinée. Un regard particulier a été porté sur les patrons de distribution des téléostéens cryptobenthiques du genre *Cirripectes* trouvés dans les Mascareignes. En effet, de récentes études ont découvert l'endémisme de certaines espèces du genre à des zones géographiques restreintes, malgré la distribution généralement large des espèces de blennies tropicales et subtropicales. Parmi les trois espèces collectées, deux ont montré une distribution Indo-Pacifique et la troisième un endémisme aux Mascareignes.

Les travaux de cette thèse posent une base indispensable aux connaissances sur la diversité du cryptobiome récifal des Mascareignes. Ils apportent une meilleure compréhension des processus de colonisation des communautés cryptobenthiques et permettront d'améliorer l'utilisation des ARMS dans les récifs coralliens et l'emploi de méthodes de taxonomie moléculaire dans le Sud-Ouest de l'océan Indien.

Mots clés : cryptobiome, récifs coralliens, métabarcoding, écologie moléculaire, Sud-Ouest de l'océan Indien, diversité

Abstract

In the context of global change, monitoring biodiversity is essential to detect stressors and understand community responses. Coral reefs are among the most diverse ecosystems on the planet, but also among the most threatened. However, the small organisms living in the crevices of the reef, the cryptobiome, remain largely unstudied even though they account for the major share of the diversity and an essential part of the food web. Accurate assessment of diversity is essential for effective ecosystem management. However, traditional taxonomic methods, often based on morphology, are poorly adapted to the cryptobiome as they are time consuming, require specialist knowledge and difficult to compare between sites. Moreover, small taxa present additional challenges by being more difficult to find and identify. Recent advances in high-throughput sequencing techniques and bioinformatics offer an alternative to traditional methods with DNA metabarcoding.

This thesis focuses on the diversity of the Mascarene reef cryptobiome (Reunion and Rodrigues) sampled with standardised sampling structures, ARMS, and studied using a metabarcoding approach. Firstly, a bioinformatics pipeline adapted to metazoan metabarcoding analyses was to evaluate the reef communities found in ARMS. The ecological interpretations derived from the metabarcoding are dependent on the available molecular references. Secondly, this work aimed to augment a molecular reference database for the Mascarene cryptobiome with sequences for three molecular markers, 18S, COI and 16S, as well as the complete mitogenome for all teleosts sampled. Third, this work assessed the diversity of the cryptobiome and its colonisation patterns across the hot and cool seasons in Reunion. The season and the immersion time significantly affected the composition of the retrieved communities. Overall, the taxa retrieved were consistent with those observed in other regions of the world and reflected the great taxonomic diversity observed in coral reefs. This work is the first to show effect of ARMS immersion time, deployment and retrieval season on community composition. These factors influence particularly sessile taxa such as ascidians, cnidarians, sponges and rodophytes. This work demonstrates the need to consider these variations when analysing the results and comparing them with other studies, and has implications for the use of ARMS as a monitoring tool. Finally, the potential and limitations of the ARMS combined with molecular identification methods for assessing species diversity and distribution were investigated with the cryptobenthic teleost genus *Cirripectes*. Recent studies highlighted that some species of this genus are endemic to islands or archipelagos, despite the generally wide distribution of these tropical and subtropical blennies. Of the three species recovered in the Mascarenes, two showed and Indo-Pacific wide distribution while the third was endemic to the Mascarene Archipelago.

This thesis furthers our knowledge of the diversity of the Mascarene reef cryptobiome. It provides a better understanding of the colonisation processes of cryptobenthic communities and will improve the use of ARMS for monitoring coral reefs and the use of molecular taxonomy methods in the South West Indian Ocean.

Keywords: cryptobiome, coral reefs, metabarcoding, molecular ecology, Southwest Indian Ocean, diversity

Remerciements

Je souhaiterais remercier tout d'abord mes directeur.rices de thèse, Henrich Bruggemann, Agnès Dettaï et Mireille Guillaume. Henrich, pour m'avoir fait découvrir le monde de la recherche et m'avoir accompagnée pendant presque cinq ans, depuis mon stage de master jusqu'au bout de cette thèse. Agnès, de m'avoir ouverte au monde de la génétique et d'avoir répondu avec patience à mes questions de néophyte, ainsi que de m'avoir accompagnée et remotivée dans les moments difficiles, malgré la distance (je garde l'idée de GIF de chats pour mes futurs stagiaires et, qui sait un jour, pour mes futurs thésards). J'ai compris que même si c'était bien d'avoir de jolis poissons, c'est plus utile d'avoir des poissons que l'on sait identifier ! Mireille, pour ta rigueur et tes qualités rédactionnelles et de synthèse qui ont permis d'améliorer ces travaux. Merci à tous les trois de m'avoir fait confiance et de m'avoir permis de développer une grande autonomie.

Je remercie grandement les membres de mon jury, mes rapporteures Agnès Bouchez et Anne Chenuil, ainsi que mon examinateur, Erwan Delrieu-trottin, d'avoir accepté de porter leur regard sur mon travail.

Je tiens également à remercier Natacha Nikolic, Emmanuel Corse et Matthieu Leray pour avoir accepté de faire partie à trois reprises de mon comité de suivi de thèse et pour les discussions enrichissantes qui ont suivi.

Mes remerciements s'adressent également à l'ensemble de l'UMR ENTROPIE et l'Université de La Réunion pour m'avoir accueillie durant ces presque quatre années de thèse. Je voudrais également remercier la région Réunion et l'Union Européenne pour m'avoir attribuée une Allocation de Recherche (FEDER PO 2014-2020) pour les trois premières années de celle-ci.

Les données collectées durant cette thèse ont été le fruit de l'implication de nombreuses personnes. Je tiens à remercier particulièrement Sophie Bureau, Lionel Bigot, Fleur Bruggemann, Arnaud Guerbet, ainsi que les stagiaires que j'ai pu encadrer, Auriane Serval, Gwennaïs Fustemberg, Baptiste Frattini et Lisa Loze, pour votre aide dans l'échantillonnage. Merci à toutes les personnes qui ont aidé à la pose et à la collecte des ARMS ainsi qu'au tri de cette faune (trop?) diversifiée, de tôt le matin jusqu'à encore plus tôt le lendemain matin. Il est indéniable que pendant la thèse, nous apprenons de nombreuses nouvelles compétences, mais tout le monde n'a pas la chance de devenir spécialiste de la pêche à la crevette à la petite cuillère, alors merci petites crevettes de nous avoir fait tourner en bourriques !

Le terrain aurait pu être sponsorisé par une boulangerie dont je ne citerai pas le nom mais que toute personne ayant fait du labo à St-Gilles reconnaitra. Merci de nous avoir remonté le moral lors des journées de terrain sans fin (ou plutôt sans faim). Je remercie également les équipes de TMSOI, Kazabul et de la municipalité de St-Pierre pour leur soutien logistique dans la récolte des ARMS. Merci au département de Chimie de l'Université de la Réunion de nous avoir dépanné en filtres Nitex alors que les commandes étaient perdues dans un autre espace-temps dû à la crise COVID. Merci à Pascale Cuet pour m'avoir prêté sa centri et à Nico pour avoir été mes muscles pour la déplacer à travers la fac ! Merci au département de Biologie, en particulier Hafsa, pour m'avoir dépannée au pied levé lorsque le matériel est tombé en panne. Un grand merci à Agnès et Céline Bonillo pour avoir pris le relai pour les PCR, et de m'avoir aidée à rattraper le retard pris avec la crise COVID, entre confinements, absence d'avion, pénurie de matériel et pénurie de kits d'extraction ! Merci à l'OSU-Réunion pour la mise à disposition du serveur et à Guillaume Desprairies pour le débogage qui va avec. Je rigole en repensant à ma tête quand j'ai compris que je n'aurais rien de

plus que cet écran noir sans interface graphique et qu'il allait falloir que j'apprenne à dompter cette machine pour qu'elle veuille bien me créer un nouveau dossier.

Merci à Jo, de notre binôme au premier cours de R en M1 jusqu'à notre soutenance à trois jours d'écart sept ans plus tard, on en aura fait du chemin et j'espère que ça continuera <3, essuie cette petite larme, bien sûr qu'on se reverra ! Tu n'as pas le choix ! Merlène, merci pour ta bonne humeur infatigable et le futur gâteau qui j'en suis sûre sera délicieux ;) J'espère que les puffins arrêteront de se fracasser contre les falaises et les pétrels de s'y jeter pour que tu puisses les étudier encore longtemps ! Ils ne se rendent pas compte de la chance qu'ils ont de t'avoir ! Notre sauvetage de Plumeau aura été un des moments les plus stressants de cette thèse, mais on a géré, girl power ! Merci David d'avoir partagé avec moi notre placard électrique, tu entames la dernière ligne droite, tu vas gérer ! Merci Helena pour ton pragmatiste et de mettre toujours les deux pieds dans le plat XD, merci Chrystelle d'avoir instauré les p'tits dej' et les goûters. Merci, Justine, Felix, Diego et Romain pour les discussions et les encouragements. A tous, merci d'avoir été de supers collègues puis amis, excepté Romain *of course*, le couloir de l'amour n'aurait jamais existé sans vous <3. Merci pour les goûters et les dramas. Bravo à ceux qui sont déjà passés par là, courage pour les autres !

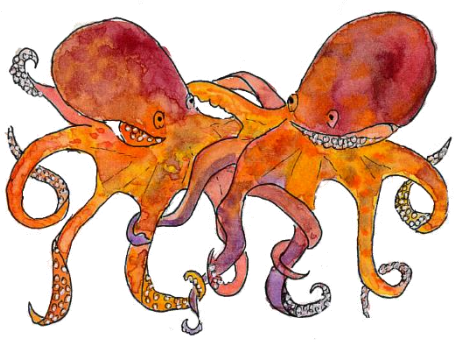
Merci à ma famille et en particulier à mes parents pour votre soutien et votre confiance tout au long de ma scolarité, des après-midis chez l'orthophoniste jusqu'à aujourd'hui. (Petit clin d'œil aux enseignants qui m'ont rabâché jusqu'au bac que je n'étais pas faite pour les grandes études). Je serai éternellement reconnaissante des efforts que vous avez faits pour que je puisse arriver jusqu'ici. Mais également, merci d'avoir cru en moi quand je suis partie en Irlande alors qu'on connaissait mon niveau d'anglais chaotique, et de m'avoir soutenue une nouvelle fois quand je vous ai dit que la Bretagne c'était bien mais que je m'envolais un an pour les tropiques. Merci à ma petite sœur, ma toute première stagiaire ;), d'avoir été là, je sais que je n'ai pas été très présente mais je compte me rattraper. Merci à la belle-famille pour votre soutien et les livraisons annuelles de fromages et de saucissons, la maison a paru moins loin.

Cohabiter avec un animal de compagnie aurait des bienfaits sur la santé, il parait que le sport aussi… Alors merci Olive d'avoir compensé ces séances de sport qui manquaient au maintien de ma santé mentale !

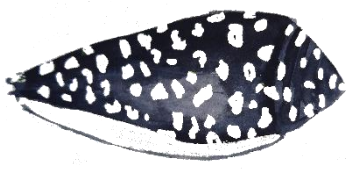
Un an à La Réunion s'est transformé en sept, des hauts, des bas, des moments inoubliables et d'autres encore plus, et aussi des rencontres ! S'il fallait n'en résumé qu'une seule ce serait ce Normand tout blanc qui est devenu mon mari. Merci pour ton soutien, des encouragements aux bons petits plats, merci de m'avoir poussée à prendre sur moi et à me dépasser afin de réaliser un de mes rêves. Merci pour tout, ces lignes, des mots, ne seront jamais assez pour exprimer toute ma reconnaissance d'être entré dans ma vie. J'ai hâte de voir ce que les récifs nous réservent <3.

Pour finir, merci à moi d'avoir tenu bon.

Table des matières



Liste des figures



Liste des tableaux

Liste des abréviations

ARMS : Autonomous Reef Monitoring Structures

ASU : Artificial sampling units (Plaques de recrutement artificielles)

ASV : Amplicon Sequence Variant

BLAST : Basic Local Alignment Search Tool

CBOL : Consortium for the Barcode of Life

COI : Cytochrome Oxydase sous unité 1

CoMF : Census of Marine Life

CRED : Coral Reef Ecosystem Division

CReefs : Census of Coral Reef Ecosystems

DADA : Divisive Amplicon Denoising Algorithm

eDNA : ADN environnemental

GCRMN : Global Coral Reef Monitoring Network

IFRECOR : Initiative Française pour les Récifs Coralliens

LCA : Last Common Ancestor (Plus proche ancêtre commun)

MACES : Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons Max score

mBRAVE : Multiplex Barcode Research And Visualization Environnement

mtADN : ADN mitochondriale

NCBI : National Center for Biotechnology Information

NGS : Séquençage de Nouvelle Génération

ORF : Open Reading Frame

OTU : Operational Taxonomic Unit (Unité Taxonomique Opérationnelle)

PCoA : Analyse en Coordonnées Principales

PCR : Polymerase Chain Reaction

POM : Particules de Matières Organiques

POC : Carbone Organique Particulaire

QIIME : Quantitative Insights Into Microbial Ecology software

RDP : Ribosomal Database Project

SBS : Sequencing By Synthesis

SOOI : Sud-Ouest de l'océan Indien

SST : Sea Surface Temperature (température des eaux de surface)

SWOT : Strengths (Forces), Weaknesses (Faiblesses), Opportunities (Opportunités), Threats (Menaces)

WoRMS : World Register of Marine Species

Liste des productions scientifiques

Les différentes valorisations scientifiques produites au cours de ce projet doctoral sont listées ci-dessous :

1. Publications

Couëdel M, Dettai A, Guillaume MMM, Bruggemann F, Bureau S, Frattini B, Verde Ferreira A, Azie JL, Bruggemann JH (2023). New insights into the diversity of cryptobenthic *Cirripectes* blennies in the Mascarene Archipelago sampled using Autonomous Reef Monitoring Structures (ARMS). *Ecology and Evolution*, 13(3), e9850. https://doi.org/10.1002/ece3.9850

Couëdel M, Dettai A, Guillaume MMM, Bonillo C, Frattini B, Bruggemann JH. Settlement patterns and temporal replacement of cryptobenthic coral reef communities and implication for evaluating diversity using ARMS (Autonomous Reef Monitoring Structures), *en préparation* pour Scientific Reports.

2. Communications orales

Couëdel M*, Dettaï A, Guillaume MMM, Bruggemann F, Bureau S, Frattini B, Fustemberg G, Loze L, Serval A, Verde Ferreira A, Bruggemann JH. How can the mini-reef ARMS method enhance knowledge of the coral reef cryptofauna? October 2022, the 12th WIOMSA, Port-Elizabeth, South Africa

Couëdel M, Dettaï A*, Guillaume MMM, Bruggemann F, Bureau S, Frattini B, Fustemberg G, Serval A, Bruggemann JH. Apports de la méthode des mini-récifs ARMS dans les Mascareignes à la connaissance des poissons crypto-benthiques récifaux. Mars 2022, RIF 2022 - 8e Rencontres de l'Ichtyologie en France, Paris, France

3. Posters

Frattini B*, Couëdel M, Guillaume MMM, Goberville Eric, Dettaï A, Bruggemann F, Bureau S, , Fustemberg G, Loze L, Serval A, Verde Ferreira A, Bruggemann JH. Communautés sessiles des mini-récifs artificiels (ARMS) sur les pentes externes des récifs coralliens de l'île de La Réunion : patrons spatiaux multi-échelles et possibles forçages environnementaux. Mai 2022, Journées scientifiques de BOREA, Dourdan, France. https://borea.mnhn.fr/fr/node/9198

4. Médiation scientifique

2022 : • La Nuit européenne des chercheurs : Exposition improbable, 30/09/2022 – Oratrice

• Biodiversité cachée des récifs coralliens : étude des dynamiques spatiale et temporelle du cryptobiome, Résultats préliminaires, UMR ENTROPIE, Université de La Réunion et Université de Nouvelle-Calédonie – Oratrice

2021 : • La Nuit européenne des chercheurs : Dans ma valise de chercheure, 24/09/2021 – Oratrice

• Tournier V. (2021) La biodiversité cachée de la petite faune marine. Le Journal de l'île de La Réunion, 15/02/2021, page 6.

• P.E. (2021) La petite faune marine à l'étude. Le Quotidien de La Réunion, 10/02/2021, page 9.

• Pasquier S (2021) Que cachent les récifs coralliens ? Webmag Recherche, Université de La Réunion, https://webmag-recherche.univ-reunion.fr/webmag-recherche-03/que-cachent-les-recifs-coralliens

2020 : • Biodiversité cachée des récifs coralliens : étude des dynamiques spatiale et temporelle du cryptobiome, Résultats préliminaires, (Séminaire Bioécotrop), UMR ENTROPIE, Université de La Réunion – Oratrice

• Biodiversité des récifs coralliens : étude des dynamiques spatiale et temporelle du cryptobiome, Résultats préliminaires, (Séminaire), UMR ISYEB et UMR BOREA, MNHN – Oratrice, https://www.youtube.com/watch?v=29oW5pQX4Jg&ab_channel=ISYEB

5. Rapports co-encadrés

Frattini B. (2021) Communautés sessiles des mini-récifs artificiels (ARMS) sur les pentes externes des récifs coralliens de l'île de La Réunion : patrons spatiaux multi-échelles et possibles forçages environnementaux. Master 2 Biodiversité, écologie et évolution, parcours Biodiversité et Ecosystèmes Tropicaux – Aquatiques Littoraux Insulaires, Université de La Réunion. Direction : H. Bruggemann et co-encadrement : M. Guillaume (MNHN) et M. Couëdel (UR), 35pp.

Fustemberg G., Serval A. (2021) Création d'un référentiel barcode de la faune récifale cryptique des Mascareignes. Master 1 Biodiversité, écologie et évolution, parcours Biodiversité et Ecosystèmes Tropicaux – Aquatiques Littoraux Insulaires, Université de La Réunion. Direction : H. Bruggemann et co-encadrement : M. Couëdel (UR), A. Dettai (MNHN) et M. Guillaume (MNHN), 41pp.

Loze L. (2022) Recherche de patrons de répartition de la petite faune récifale des Mascareignes. Master 2 Dynamique des Écosystèmes Aquatiques, Université de Pau et des pays de l'Adour. Direction : H. Bruggemann et co-encadrement : M. Couëdel (UR), A. Dettai (MNHN) et M. Guillaume (MNHN), 48pp.

Avant-propos

Ce travail de doctorat a été effectué à l'Université de La Réunion (UR) au sein de l'Unité Mixte de Recherche (UMR) ENTROPIE (Écologie mariNe TRopicale des Océans Pacifique et IndiEn). J'ai été encadrée par le Pr Henrich Bruggemann (UR, UMR ENTROPIE), ainsi que les Dr Agnès Dettaï (MNHN, UMR ISYEB) et Mireille Guillaume (MNHN, UMR BOrEA).

L'UMR ENTROPIE a pour objectif de mieux comprendre le fonctionnement des écosystèmes marins et insulaires tropicaux de l'Indo-Pacifique tropical dans le contexte du réchauffement climatique. L'UMR ISYEB a pour objectif d'étudier l'origine de la biodiversité, la diversification des espèces et l'établissement de communautés en relation avec l'évolution des taxons dans le temps et dans l'espace. L'UMR BOrEA a pour objectif l'étude de l'écologie et de la biologie des organismes et des habitats aquatiques dans des écosystèmes naturels et contraints.

J'ai bénéficié d'une Allocation Régionale de Recherche (ARR) d'une durée de trois ans, émanant de la Région Réunion et de l'Europe (Fonds Social Européen, FSE), puis d'un financement complémentaire de trois mois en raison de la crise COVID. Au cours des trois années, j'ai également été enseignante vacataire au sein du Département de Biologie de l'Université de La Réunion (74h TD). J'ai réalisé les 6 derniers mois de mon doctorat sur fonds personnels et en réalisant 37h TD dans ce même département.

La crise COVID-19 a fortement impacté la réalisation de ces travaux. Les mois de confinement ont décalé les manipulations en laboratoire, les séquençages, puis l'approvisionnement des consommables avec l'absence de transport aérien. La forte demande en matériels d'extraction et PCR pour la réalisation des tests anti-COVID ont conduit à des pénuries qui ont également retardé les manipulations en laboratoire.

Ces travaux s'inscrivent dans le programme FEDER-CALIBIOME, qui a pour objectif de proposer des méthodes calibrées permettant une meilleure connaissance et un suivi de la biodiversité marine côtière, tout en améliorant les connaissances sur les processus de maintien et les potentiels de résilience des communautés d'animaux marins. Quatre compartiments de la biodiversité ont été ciblés lors du programme : les larves de poissons, les méduses, les recrues coralliennes et le cryptobiome récifal. C'est dans ce quatrième compartiment que cette thèse a été développée avec objectif de mieux comprendre la diversité et les patrons de colonisation du cryptobiome récifal. Le projet CALIBIOME a été financé par le Fonds européen de développement régional (FEDER) via la Région Réunion (no. du projet 20171591-0002633 CALIBIOME 2017-2022) et fait suite au programme FEDER-Biodiversité (2014-2015), qui dans le cadre de son action 2 'Connaissances et outils pour la gestion de la biodiversité récifale régionale', a permis l'installation des ARMS à Rodrigues et à La Réunion en 2014 et la réalisation d'un atelier régional de formation aux méthodes de leur relève et le traitement des échantillons à La Reunion (mars 2015). Le déploiement des ARMS a été réalisé sous les autorisations de la Direction de l'environnement, de l'aménagement et du logement de la Réunion (DEAL ; n°2018-61 DEAL/SEB/UBIO et n°2020-09-DEAL/SEB/UBIO) et de la direction de la mer Sud océan Indien (n°2019-083 and n°2020-054). La partie terrain a bénéficié du

laboratoire humide de l'Université de La Réunion à Saint-Gilles et de la salle de la base nautique de la Mairie de Saint-Pierre, que nous avons adaptée aux besoins du démembrement des ARMS.

La relève des ARMS, structures d'échantillonnage du cryptobiome récifal, dépend d'un effort collectif de tri en laboratoire humide qui impliqué également deux étudiantes de M1 Guennaïs Fustemberg et Auriane Serval, un étudiant de M2 Baptiste Frattini, un étudiant post-M2 Arnaud Guerbet, deux ingénieures d'étude Fleur Bruggemann et Sophie Bureau, en plus de Mireille Guillaume et Henrich Bruggemann. Le tri en morpho-espèces des spécimens récoltés en 2018 a été effectué par moi-même dans le cadre de prestations de 3 mois antérieures à ma thèse (décembre 2018 à février 2019). Le tri de ceux récoltés par la suite a été partagé avec Fleur Bruggemann. Les manipulations de biologie moléculaire ont été réalisées en partie au MNHN avec l'aide de Céline Bonillo, Amélie Verde Ferreira et Agnès Dettaï.

Chapitre 1 :Introduction, étudier la biodiversité cachée

1. Les récifs coralliens peu profonds

1.1. Une biodiversité cruciale

Les récifs coralliens font partie des écosystèmes les plus riches et les plus diversifiés de la planète, accueillant près de 30 % de la biodiversité marine mondiale en terme d'espèces sur moins 1 % de la surface des océans (**Spalding et al. 2001 ; Fisher et al. 2015** ; Figure 1.1). Les coraux scléractiniaires, ou coraux durs, sont les principaux bioconstructeurs des récifs coralliens **(Salvat 1992)**. D'autres organismes au squelette calcaire tel que les hydrozoaires, les foraminifères, mais également les algues calcaires et les processus de calcification inorganique dans les interstices du récif contribuent également à la structure tridimensionnelle du récif **(Sorokin 1993)**.

Couvrant une superficie de 600 000 km^2 répartis dans les eaux marines entre 30°N et 30°S, les récifs coralliens sont des écosystèmes essentiels aux maintien des océans et aux sociétés humaines **(Wilkinson 2008)**. À l'instar des arbres d'une forêt, la structure tridimensionnelle du récif génère un habitat physique complexe qui est exploité par les espèces en offrant une surface de recrutement pour les organismes sessiles, un abri pour les organismes mobiles et une source de nutrition pour de nombreux organismes **(Steele 1999)**. Par ailleurs, la productivité élevée, associée à la biodiversité des récifs coralliens **(Odum and Odum 1955 ; Harrison and Booth 2007 ; Mora et al. 2011)**, fournit une source de revenus et assure la sécurité alimentaire des populations humaines riveraines **(Moberg & Folke 1999)**. Les récifs coralliens fournissent également des services écosystémiques vitaux aux sociétés et aux industries à travers la production halieutique, les matériaux de construction (sable), de nouveaux composants biochimiques, le tourisme, la protection côtière et représentent également un aspect culturel **(Hoegh-Guldberg et al. 2007, 2019).** Le dernier rapport

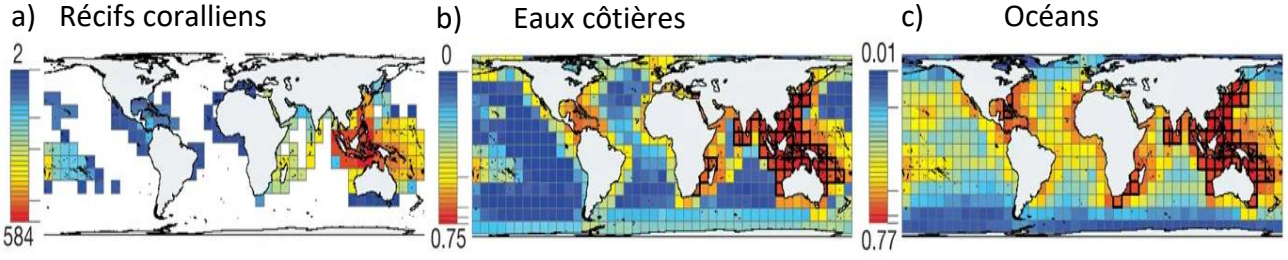


Figure 1.1: Répartition de la richesse spécifique en fonction des habitats (a) récifs coralliens, (b) eaux côtières et (c) milieu marin dans son ensemble. Traduit et adapté de Tittensor et al. 2010.

de l'**IFRECOR** (Initiative Française pour les Récifs Coralliens) **(2020)** a évalué la valeur totale annuelle des services rendus par les récifs coralliens en outre-mer français à 1,3 milliard d'euros.

1.2. Une biodiversité hétérogène

La biodiversité est repartie de façon hétérogène sur terre avec des zones géographiques ayant de fortes richesses spécifiques. En effet, à l'échelle de la province Indo-pacifique, la biodiversité des récifs coralliens est à son maximum dans la région de l'archipel Indo-Malay-Philippine et décline d'Est en Ouest dans les océans Indien et Pacifiques (Figure 1.1a ; **Bellwood 2001 ; Roberts et al. 2002 ; Tittensor et al. 2010**). Cependant, certaines des zones géographiques à forte richesse spécifique subissent un taux élevé de perte d'habitats naturels. Ces zones appelées point chaud de diversité (*hotspot* en anglais) sont définies par (1) un nombre élevé d'espèces, principalement endémiques, qu'elles hébergent et (2) les menaces anthropiques auxquelles elles doivent faire face. De tailles variées, les « hotspots » permettent de prioriser les actions de conservation, pour protéger un maximum d'espèces au coût moindre.

Hotspot de Madagascar et des îles de l'océan Indien

Parmi les 36 hotspots définis, on retrouve le Hotspot de Madagascar et des îles de l'océan Indien qui comprend Madagascar, les archipels des Mascareignes (Agalega, Cargados Carajos Shoals (Saint Brandon), La Réunion, Maurice, et Rodrigues), des Comores (Grande Comore, Anjouan, Mohéli et Mayotte) et des Seychelles (constitués de 115 îles) ainsi que les îles Éparses de l'ouest de l'océan Indien (Europa, Bassas da India, Juan de Nova, les îles Glorieuses et Tromelin). Si le hotspot a été principalement défini pour sa biodiversité végétale, sa biodiversité marine est également exceptionnelle notamment avec de forts taux d'endémisme (ex. coraux, espèces côtières) et fait partie des 10 hotspots marins définis par **Roberts et al. 2002**. Ce hotspot a souvent été considéré comme prioritaire au sein même des hotspots, de par son extrême diversité spécifique et haut niveau d'endémisme s'illustrant par de mécanismes évolutifs distincts liés à son isolement de longue date des grandes masses continentales (**BIOTOPE 2022**). Source image : **(CEPF 2015)**

1.3. Une biodiversité menacée

Les récifs coralliens sont l'un des écosystèmes les plus menacés **(Carpenter et al. 2008 ; Hoegh-Guldberg et al. 2019)**. À la fois sensibles aux changements globaux tels que le réchauffement des océans, l'augmentation du niveau marin et l'acidification des océans, les récifs aussi sont fortement impactés à l'échelle locale par les activités humaines (surpêche, sédimentation,

pollutions **(Sully et al. 2019)**). Au cours des dernières décennies, la perte de la couverture corallienne a été estimée entre 19 et 61 % et devrait continuer d'augmenter **(Knowlton & Jackson 2008 ; De'ath et al. 2012 ; Obura et al. 2022)**. Si les impacts directs de la diminution de la couverture coralliennes sont bien établis, les répercussions indirectes sur l'ensemble des écosystèmes associés restent encore peu commensurables **(Jones et al. 2004 ; Idjadi & Edmunds 2006 ; Glynn 2011)**. La perte des coraux et de la structure corallienne ont un impact direct sur les organismes vivants dans les récifs, en particulier ceux vivant dans les habitats cryptiques et les téléostéens (nommés par la suite dans ce manuscrit poissons) mettant ainsi en péril la pêche locale **(Rogers et al. 2014)**.

C'est dans ce contexte que les études se multiplient afin de mieux comprendre le fonctionnement des écosystèmes menacés, décrire les espèces et leur évolution dans le temps et l'espace afin de proposer des mesures pour limiter l'impact de l'anthropisation sur ces derniers. Face à ces taux de dégradation alarmants, il est urgent de documenter, d'étudier et de fournir des méthodes fiables d'estimation de la biodiversité à travers le temps et l'espace **(Plaisance et al. 2011)**.

2. L'évaluation de la diversité

Afin de décrire la distribution des espèces à différentes échelles spatiales, trois concepts sont employés en écologie des communautés, la diversité alpha (α), la diversité bêta (β) et la diversité gamma (γ) **(Whittaker 1960)**.

2.1. La diversité alpha

La diversité alpha (alpha index ; **Fisher et al. 1943**) réfère à la diversité des espèces d'une communauté d'un système délimité, c'est-à-dire, un site donné à un moment donné. La diversité alpha peut être évaluée en termes de richesse spécifique ou calculée à l'aide d'indices de diversité tels que ceux de Shannon, de Simpson ou de Chao.

La richesse spécifique correspond au nombre d'espèce observées dans des données de comptage. La richesse spécifique est la mesure la plus simple conceptuellement mais ne prends pas en compte l'abondance de ces espèces (diversité spécifique) comme le font l'indice de Shannon, aussi appelé indice de Shannon-Weaver ou Shannon-Wiener **(Shannon 1948)** et l'indice de Simpson. L'indice de diversité de Simpson donne plus de poids aux espèces abondantes qu'aux espèces rares. La présence d'espèces rares dans le peuplement ne modifie pratiquement pas la valeur de l'indice de diversité, contrairement à l'indice de Shannon beaucoup plus sensible. Ces mesures de la

diversité sont sujettes à des biais d'estimation **(Mouillot & Leprêtre 1999)**, notamment à cause des espèces non échantillonnées. En effet, il est généralement difficile de relever toutes les espèces rares, en particulier dans des systèmes très riches comme les récifs coralliens. C'est pour pallier à ce biais que l'estimateur de diversité Chao a été développé. Il estime le nombre d'espèces non observées à partir de celles observées 1 ou 2 fois (Chao1 ; **Chao 1984**). Dans un deuxième temps, Chao **(1987)** propose un estimateur du nombre d'espèces appliqué aux données de présence-absence (Chao2).

2.2. La diversité beta

La diversité β mesure la variation de la diversité des espèces (diversité α) entre deux communautés que ce soit à l'échelle spatiale (entre deux sites) ou à l'échelle temporelle (deux années, saisons différentes) **(Whittaker 1960)**. Ainsi, la diversité β mesure la dissimilarité entre communautés. Les indices de dissimilarité les plus connus sont ceux de Jaccard (**1912** ; cf. Figure 1.2: DJ = (b+c)/(a+b+c)) ou de Sørensen (**1948** ; cf. Figure 1.2: DS = (b+c)/(2a+b+c)), basés sur la présence-absence des espèces. Plus le nombre d'espèces partagées entre les deux communautés est faible, plus la dissimilarité et donc la diversité β est grande. La différence dans la composition des espèces peut être le résultat d'un remplacement de certaines des espèces par d'autres espèces (illustré en bleue dans la Figure 1.2 avec les gastéropodes du site 1 qui sont remplacés par les téléostéens sur le site 2), ou provenir d'une différence de richesse spécifique entre les deux communautés (illustré en vert par les échinodermes dans Figure 1.2; **Legendre 2014**).

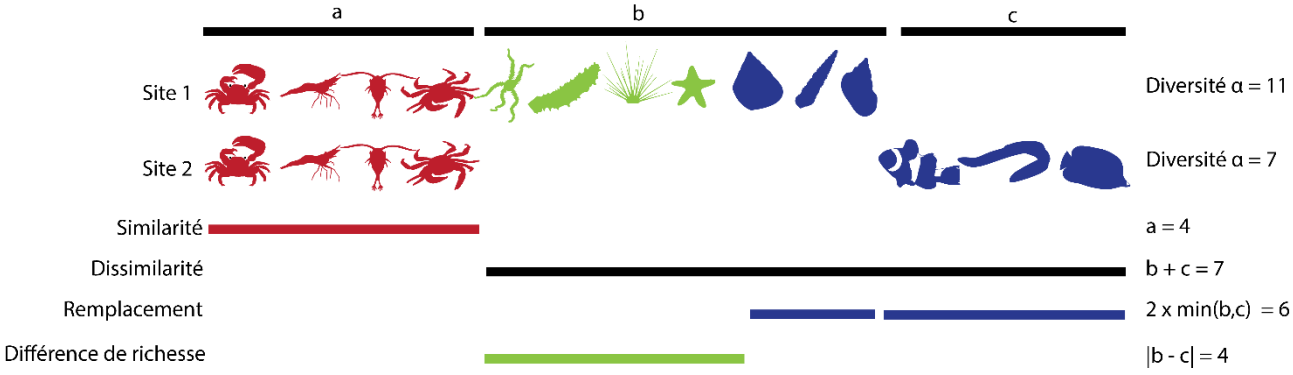


Figure 1.2 : Comparaison de la composition en espèce de deux communautés (Site 1 et Site 2). Chaque silhouette représente une espèce, a = représente le nombre d'espèces présentes sur les deux sites, b représente le nombre d'espèces présentes uniquement sur le site 1 et c uniquement sur le site 2. Les espèces d'arthropodes en rouge illustrent la similarité entre les deux communautés. Les échinodermes en vert illustrent la différence de richesse spécifique et les gastéropodes et téléostéens en bleu illustrent le remplacement des espèces entre les deux communautés. La diversité alpha du site 1 est de 11 espèces et celle du site 2 de 7 espèces. La diversité gamma (a+b+c) des deux sites est de 14 espèces. Dissimilarité de Jaccard : DJ = (b+c)/(a+b+c) =0.71); Dissimilarité de Sørensen DS = (b+c)/(2a+b+c) = 0.5. Adapté de Legendre (2014)

Pour finir, la diversité γ reflète la diversité spécifique d'ensemble de communautés échantillonnées (méta-communauté). Par exemple sur la Figure 1.2, la diversité gamma des deux sites est de 14 espèces.

Cependant, ces trois concepts et plus généralement l'évaluation de la biodiversité, reposent sur la notion d'espèce qui est depuis longtemps discutée en raison de la complexité de la délimitation des espèces et des hybridations interspécifiques **(Brown et al. 1996 ; Peterson & Navarro-Sigüenza 1999 ; De Queiroz 2007 ; Tang et al. 2012)** .

Espèce cryptique

Le terme d'espèce cryptique est employé pour désigner deux notions bien différentes :
(1) une espèce qui fait partie d'un complexe d'espèces cryptiques, c'est-à-dire, une espèce qui est difficilement différenciable, uniquement sur des critères morphologiques, des autres espèces du complexe d'espèce. C'est à cette notion que réfère « espèce cryptique » dans ces travaux ;
(2) les espèces composant la faune cryptique (cachée). L'emploi du terme espèce cryptique dans ce cas est un abus de langage.

2.3. Le problème de l'espèce

Évaluer la richesse spécifique suppose que les espèces soient clairement définies. Historiquement les espèces étaient délimitées et classées sur la base de critères morphologiques (espèce morphologique), puis sur le critère de compatibilité reproductive (espèce biologique ; **Dobzhansky 1937 ; Mayr 1948**). En fonction des groupes taxonomiques, de l'histoire des taxons et des caractères morphologiques étudiées, la délimitation des espèces peut être très complexe **(Gélin et al. 2017)**. L'utilisation des critères morphologiques était soumise à la subjectivité de l'observateur et difficile lorsque les espèces étaient sujettes à de la plasticité phénotypique (ex. coraux scléractiniaires **Todd 2008**) ou lorsque des espèces distinctes ne présentent pas de caractères morphologiques distinctifs (espèces cryptiques). Le concept d'espèce biologique basé sur l'isolement reproducteur ne permet pas de prendre en considération les hybridations interspécifiques et comporte des contraintes expérimentales faisant qu'il était rarement vérifiable.

Par la suite, l'apparition des études basées sur la génétique a permis de définir le concept d'espèce phylogénétique comme "le plus petit groupe identifiable d'individus avec un patron commun d'ancêtres et de descendants" **(Cracraft 1983)** et a remis en cause le concept de délimitation d'espèces uniquement sur des critères morphologiques en découvrant des complexes d'espèces cryptiques **(Knowlton 1993 ; Pfenninger & Schwenk 2007 ; Hoban & Williams 2020)**. Cependant, chaque gène a son histoire évolutive et donc en fonction du gène ciblé l'histoire reconstituée de l'espèce peut différer (arbres de taxons versus arbres de gènes ; **Maddison 1997**), surtout si les gènes ont une origine différente (*cf.* Chapitre 2). Chaque approche de délimitation

d'espèces ayant ses avantages et inconvénients, il est généralement admis que plusieurs approches doivent être combinées en menant des études de taxonomie intégrative **(Padial et al. 2010)** (ex. morphologie, moléculaire [ADN mitochondrial et nucléaire], écologie et distribution).

3. Le cryptobiome récifal

3.1. La majorité cachée

Les organismes vivant aux seins de écosystèmes coralliens peuvent être classés en trois compartiments : (1) les organismes benthiques mobiles et sessiles exposés ; (2) les organismes démersaux qui vivent dans la colonne d'eau et (3) le cryptobiome, qui désigne l'ensemble des organismes vivants cachés dans les anfractuosités du récif. Le cryptobiome inclut des organismes aux mode de vie très variés : certains annélides, siponcles et bivalves vivent enfouis dans le substrat meuble ; d'autres sont fixés voire encroûtants, tels que les éponges, les ascidies et les bryozoaires ; ou bien nichent dans les crevasses, à l'image de petits poissons, échinodermes, mollusques et crustacés (**Reaka-kudla 1997** ; *cf.* Chapitre 3). Malgré des années d'études sur les écosystèmes récifaux, le cryptobiome reste sous étudié **(Idjadi & Edmunds 2006 ; Glynn 2011)**, alors qu'il représente la majeure partie de la diversité et de la biomasse produite des récifs coralliens **(Reaka-kudla 1997 ; Pearman et al. 2016 ; Brandl et al. 2019)**. Les petits taxons présentent des défis supplémentaires en étant intrinsèquement plus difficiles à trouver et à identifier, et sont donc souvent omis des suivis visuels et des collections. Bien que sous étudié vis-à-vis des macro-organismes tel que les coraux et les poissons, le cryptobiome fait l'objet d'étude sur des taxons particuliers **(Winston 1986 ; Bouchet et al. 2016 ; Dick et al. 2020)** sans pour autant être inclus dans les suivis (*monitoring*) et l'évaluation de la santé des récifs coralliens (cf. *Global Coral Reef Monitoring Network* (**GCRMN**) ; **Chabanet et al. 2016** ; **Beger et al. 2007 ; Mellin et al. 2011**). Pourtant face au déclin des récifs tropicaux peu profonds, il est nécessaire et essentiel de prendre en compte l'ensemble des communautés récifales dans l'évaluation de la biodiversité **(Plotnick et al. 2016 ; Osman et al. 2020)**.

Les études qui se sont portées sur le cryptobiome montrent d'importants nombres d'espèces sur de petites unités de surface, suggérant une biodiversité récifale largement sous-détectée par les méthodes de suivis traditionnels, et par conséquent sous-estimée **(Plaisance et al. 2011)**. Certains travaux ont souligné l'importance de ces communautés cryptiques dans le fonctionnement des écosystèmes récifaux (ex. **Richter & Wunsch 1999 ; Goeij & Duyl 2007 ; Scheffers et al. 2010**). Le cryptobiome est composé de groupes trophiques qui ont un rôle fonctionnel clé dans le maintien des récifs coralliens tels que des herbivores **(Coen 1988)**, des suspensivores **(Scheffers et al. 2010)**,

des prédateurs **(Reaka 1987)** et des détritivores **(Rothans & Miller 1991)**. Les organismes composant le cryptobiome, comme les poissons cryptobenthiques, ont un taux de renouvellement important et représentent une production essentielle au réseau trophique et participe au maintien des communautés récifales et halieutiques **(Wolf et al. 1983 ; Brandl et al. 2019 ; Mihalitsis et al. 2022)**. En outre, certains organismes du cryptobiome peuvent intervenir dans la protection des coraux face aux prédateurs (ex. les crevettes *Alpheus lottini* et les crabes *Trapezia* spp. peuvent détecter les étoiles de mer *Acanthaster planci* et protéger le corail hôte ; **Glynn 1980 ; McKeon et al. 2012)** voir conférer une meilleure résistance et résilience aux coraux (ex. les crabes *Trapezia serenei* et *Tetralia nigrolineata* aident à nettoyer les sédiments **Stewart et al. 2006)**.

3.2. Vers une évaluation standardisée

L'étude du cryptobiome est complexe en raison de la nature cryptique et la taille des organismes qui le composent. L'utilisation de réplicats (ex. à travers des gradients environnementaux) est souvent difficile ou impraticable en raison de la complexité de l'habitat, de l'association étroite du cryptobiome avec des taxons sensibles aux variations environnementales ainsi qu'à sa grande variabilité en fonction des micro-habitats **(Enochs et al. 2011)**. La comparaison des études est également difficile avec des choix de taxons cibles et des méthodes d'échantillonnage qui diffèrent.

C'est dans ce contexte qu'ont été développés les *Autonomous Reef Monitoring Structures* (**ARMS**) ou mini-récifs à faune cryptique, en 2004 **(Zimmerman & Martin 2004)**. Ils ont été améliorés en 2006 par le *Coral Reef Ecosystem Division* (**CRED**), en partenariat avec le *Census of Marine Life* (CoML) et le *Census of Coral Reef Ecosystems* (**CREEFS**) **(Knowlton et al. 2010 ; Plaisance et al. 2011)**. Les ARMS sont des unités d'échantillonnage standardisées, conçues pour imiter la complexité structurale des récifs coralliens. Ils consistent en un assemblage de neuf plaques en PVC (22,5 x 22,5 cm) empilés avec des ouvertures (1,27 cm de haut) pour permettre aux organismes de venir s'abriter (volume : 0,005 m^3) ou de recruter (surface : 0,869 m^2 ; Figure 1.3). Les ouvertures sont en alternance obstruées par des barres en PVC allant des coins au centre de la plaque pour empêcher le flux d'eau (**Plaisance et al. 2011 ; Leray & Knowlton 2015a** ; Figure 1.3). Le traitement des ARMS suit un protocole standardisé, du déploiement des structures à l'extraction d'ADN des organismes **(Leray & Knowlton 2015)** et permet de collecter des organismes mobiles et sessiles de différentes tailles et de différents groupes taxonomiques **(Pearman et al. 2016)**. Le protocole de traitement des ARMS prévoit l'acquisition de données via trois méthodes d'analyse et ce pour 4

compartiments du cryptobiome (Figure 1.3 ; https://naturalhistory.si.edu/research/global-arms-program/protocols). Les organismes collectés supérieurs à 2 mm sont sous-échantillonnées pour des identifications morphologique et/ou moléculaire (*cf.* Chapitre 3). Les organismes mobiles de tailles inférieures sont filtrés en deux catégories de taille, 500-2000 µm et 106-500 µm, et conservées pour des analyses par métabarcoding (*cf.* Chapitre 4). Les organismes sessiles qui recouvrent les plaques des ARMS peuvent être analysés par deux méthodes : l'analyse du recouvrement (méthode *Coral Count Point* ; non abordé ici) et également par métabarcoding après le grattage des plaques (Figure 1.3).

L'emploi des ARMS et la multiplicité des méthodes possible pour l'étude du cryptobiome récifal a ses avantages et inconvénients. Ils ont été listés dans le tableau 1.1 sous forme de matrice **SWOT** (*Strengths Weaknesses Opportunities Threats*). Une matrice SWOT permet l'intégration des facteurs directement relatifs à la mise en œuvre des ARMS (Forces et Faiblesses) mais également

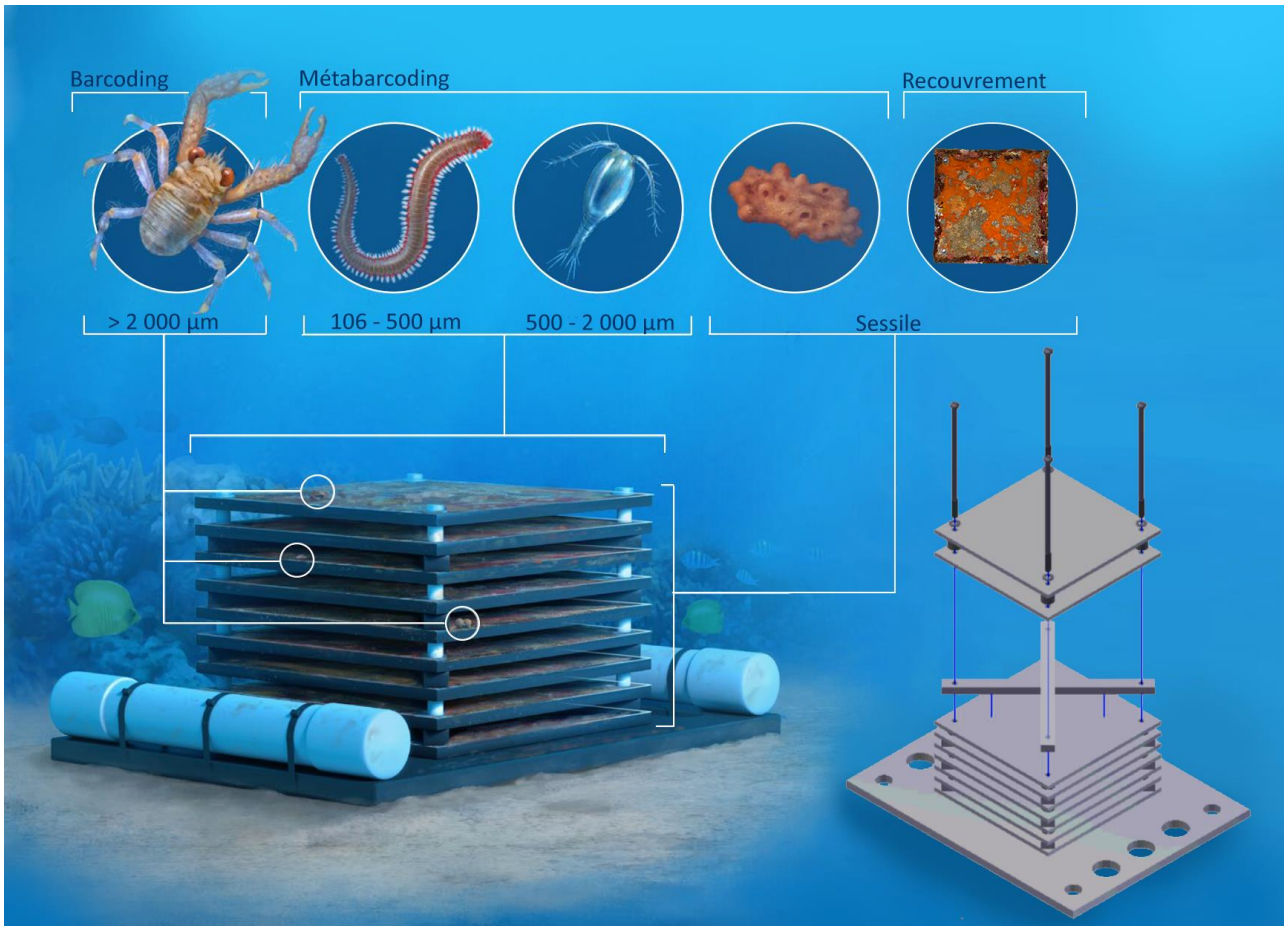


Figure 1.3 : ARMS utilisé dans l'échantillonnage du cryptobiome. Leray et al. 2013 recommandent de séparer les organismes mobiles supérieurs à 2 mm à des fins de barcoding, puis de séparer les organismes pour le métabarcoding en 3 catégories de taille. Adaptée de Pearman et al. 2018

Tableau 1.1 : Matrice SWOT de l'utilisation des ARMS pour l'étude du cryptobiome récifal

FORCES :	FAIBLESSES :
COLLECTE DE DONNÉES : • Volume et surface d'échantillonnage standardisés • Protocole de traitement standardisé • Méthode non destructrice • Collecte d'organismes difficilement accessibles par d'autres méthodes • Organismes collectés très divers (mobiles et sessiles) **TRAITEMENTS ET ANALYSES :** • Fournis de la donnée brute pérenne • Variété d'utilisation (barcoding, métabarcoding, analyses photos des plaques)	**COLLECTE DE DONNÉES :** • Ressource humaine importante pour le démantèlement des ARMS • Nécessite un financement adapté au temps de pose **BIAIS :** • Peut exclure des espèces (ex. espèces pionnières) ou colonisation biaisée par le type de substrat **(Pearman et al. 2020)**
OPPORTUNITES :	**MENACES :**
COLLECTE DE DONNÉES : • Probable nouvelles espèces • Pas d'expert taxonomique sur place • Pas d'expert écologique • Réutilisation des structures • Déployé dans différentes régions **TRAITEMENTS ET ANALYSES :** • Stockage de l'information (peut être utilisée et/ou ré-utilisée dans le futur) • Différents niveaux d'analyse en fonction des moyens financiers	**COLLECTE DE DONNÉES :** • Perte des structures suites aux tempêtes **(Villalobos et al. 2022)** • Déploiement nécessite des plongeurs **TRAITEMENTS ET ANALYSES :** • Coût financier des analyses de taxonomies moléculaires • Complétude des bases de données • Manque de standardisation des protocoles bio-informatiques

les facteurs externes et indirects tel que l'environnement ou les financements. Par exemple, la principale force de l'emploi des ARMS est l'échantillonnage d'un volume et d'une surface standardisés, et l'un des inconvénients est que cet échantillonnage peut ne pas être représentatif des communautés naturelles. En effet, l'utilisation d'un substrat artificiel affecte les communautés d'invertébrés benthiques échantillonnés en fonction du type de substrat utilisé **(Mallela et al. 2017 ; Siddik et al. 2019 ; Monroy-Velázquez et al. 2020)** et/ou de son orientation **(Glasby & Connell 2001 ; Siddik et al. 2019)**. En outre, une des opportunités qu'offrent les ARMS est qu'ils sont transposables dans de nombreuses régions du monde et écosystèmes **(Pearman et al. 2020)**. Cependant, une des limites actuelles, mais qui pourrait être levée dans le futur (*cf.* Menaces), est l'incomplétude des bases de données pour l'analyse par métabarcoding, en particulier des bases de données locales

spécifiques (*cf.* Chapitre 2). Ainsi pour répondre aux objectifs et aux contraintes des diverses études, certains auteurs ont pris en compte une seule méthode d'analyse comme le recouvrement des plaques **(David et al. 2019)** ou bien une seule catégorie de la diversité échantillonnée tel que les crustacés **(Plaisance et al. 2009 ; Hazeri et al. 2019)**.

Ces dernières années, les avancées dans les technologies de séquençage ont permis l'essor de l'identification moléculaire de très nombreuses espèces en même temps, le métabarcoding. Cette approche permet notamment la caractérisation des communautés biologiques à partir de l'ADN des organismes présents dans un échantillon (*cf.* Chapitre 2). Dans la majorité des études, les séquences ADN obtenues sont, par la suite, rassemblées en groupe de similarité, appelés **OTU** (Unité Taxonomique Opérationnelle) et assignées taxonomiquement à l'aide de bases de séquence de références (*cf.* Chapitre 2).

Les études employant la méthode des ARMS pour l'étude du cryptobiome récifal sont en plein essor et ont été répertoriées dans le tableau 1.2. **Leray and Knowlton (2015)** ont été les premiers à coupler les ARMS au métabarcoding, avec un marqueur moléculaire le COI, pour étudier le cryptobiome récifal avec 18 ARMS répartis sur deux sites de la côte est des Etats-Unis (océan Atlantique ; Tableau 1.2). Cependant, c'est au niveau de la mer Rouge que l'effort d'échantillonnage a été le plus important avec six études **(Al-Rshaidat et al. 2016 ; Pearman et al. 2016, 2018, 2019 ; Carvalho et al. 2019 ; Villalobos et al. 2022)** qui ont déployé entre 5 à 87 ARMS sur différents temps d'immersion allant de 12 à 36 mois. En revanche, aucune étude n'a employé des ARMS dans le Sud-Ouest de l'océan Indien, seulement dans l'archipel des Chagos au Nord-Ouest de l'océan Indien, toutefois les résultats publiés se sont limités aux recouvrement des plaques **(Steyaert et al. 2022a)** et la fluorescence des éponges **(Steyaert et al. 2022b)**. Ainsi cette étude doctorale s'intègre dans la première étude couplant ARMS et métabarcoding dans l'océan Indien. En outre, plusieurs études ont également combiné plusieurs marqueurs moléculaires complémentaires, le 18S et le COI, pour améliorer la résolution taxonomique comme c'est le cas ici (*cf.* Chapitre 2).

Les différentes études combinant ARMS et métabarcoding sur le cryptobiome récifal sont présentées dans le tableau 1.2. Cependant, la comparaison de ces études est complexe et fortement biaisée par l'utilisation de différents protocoles (ex. nombre d'ARMS, temps d'immersion) et différents pipelines bio-informatiques (cf. Chapitre 2). Par exemple, les ARMS récoltés à Singapour n'ont aucun OTU non assignés **(Ip et al. 2022)**, mais l'assignement des OTU a été réalisé à 85% de similarité, contrairement au 97% généralement utilisé. Des tendances générales peuvent être observées telles que :

(1) le nombre d'**OTU** trouvées augmente avec le nombre d'ARMS déployés et ne semble pas atteindre de plateau, ce qui présage une très forte biodiversité du cryptobiome (Tableau 1.2) ;

(2) les principaux taxons trouvés sont les annélides et les arthropodes, puis les mollusques et les éponges **(Leray & Knowlton 2015 ; Al-Rshaidat et al. 2016 ; Pearman et al. 2016, 2018 ; Ransome et al. 2017 ; Carvalho et al. 2019 ; Nichols et al. 2021 ; Ip et al. 2022 ; Villalobos et al. 2022)** ;

(3) les différentes fractions recueillies par les ARMS (500-2000 µm, 106-500 µm et sessile) ont des compositions différentes **(Leray & Knowlton 2015 ; Ip et al. 2022)**

(4) le cryptobiome récifal montre une très forte hétérogénéité spatiale des communautés avec la majorité des OTU trouvées sur un seul site et peu d'OTU partagées par l'ensemble des sites **(Carvalho et al. 2019)**.

Tableau 1.2 : Liste des publications combinant des ARMS et du métabarcoding pour étudier le cryptobiome des récifs coralliens et les différents protocoles de déploiement et traitement des OTU employés. (*) calculé à partir des données disponibles

Océan	# ARMS	# site	Temps de déploiement (mois)	Marqueur	Taille des OTU	# OTU	Moy. # OTU par ARMS	Seuil d'assignement	% OTU assignés au phylum	% OTU assignés à l'espèce	Référence
Atlantique	9	1	6	COI	NA - CROP	1 391	536 ± 30	97% blast 90% SAP	72%	12%	Leray & Knowlton 2015
Atlantique	9	1	6	COI	NA - CROP	1 204	434 ± 55	97% blast 90% SAP	59%	10%	
Mer rouge	5	2	16	COI	NA - CROP	1 197	609 ± 114	97% blast 80% SAP	63%	8%	Al-Rshaidat et al 2016
Mer rouge	9	3	12	COI	97%	1 700	1 297*	NA	NA	NA	Pearman et al. 2016
Pacifique	3	1	24	COI	NA	2 456	NA	97% blast 85% blast 90% SAP	55%	32%	Ransome et al. 2017
Mer rouge	33	11	24	COI	NA - CROP	3830	660 ± 151	97% blast SAP	58%	NA	Pearman et al. 2018
Mer rouge	33	11	24	18S	NA - CROP	5 420	750 ± 107	97% blast RDP	NA	NA	Pearman et al. 2018
Mer rouge	87	22	24	COI	-	10 416 (1 471 by site)	828	-	55%	NA	Carvalho et al. 2019
Pacifique	6	2	11 et 23	COI	97%	31 900	6 580 to 14 237	85% blast	51%	NA	Casey at al. 2021
Pacifique	6	2	11 et 23	18S	99%	25 994	7 113 to 11 237	90% blast	99%	NA	Casey at al. 2021
Pacifique	6	1	23	COI	97%	893	NA	97 % blast 85% LCA	NA	NA	Nichols et al. 2021
Mer de Chine	12	4	24	COI	97%	410	NA	RDP 80%	100%	54.60%	Ip et al. 2022
Mer de Chine	12	4	24	18S	1%	561	NA	RDP 60%	100%	NA	Ip et al. 2022
Mer rouge	33	4	24	COI	No (ASV)	33 832 ASV	NA	RDP	NA	NA	Villalobos et al. 2022

4. Contexte général

L'étude de la diversité globale des récifs coralliens est un enjeu majeur pour comprendre le fonctionnement de cet écosystème et notamment, sa capacité de résilience après des perturbations anthropiques ou naturelles. Le cryptobiome représente la majorité de la diversité associée aux récifs coralliens, une biomasse importante et une composante essentielle à leur fonctionnement. L'utilisation des ARMS couplés aux méthodes d'écologie moléculaire est de plus en plus plébiscitée pour documenter la composition et la structure de ce compartiment récifal. Cependant, malgré l'effort international pour standardiser cette approche, certains aspects restent variables entre les études. A l'heure actuelle, le renouvellement spatial (diversité β), l'effet du temps de déploiement, de la saison au moment de la pose et/ou de la récolte ne sont pas documentés, laissant une méconnaissance sur les dimensions spatiales et temporelles sur la structure des communautés du cryptobiome. La compréhension de ces deux dynamiques permettrait pourtant d'améliorer l'utilité des ARMS en tant qu'outil et dispositif de suivi, avec un intérêt en conservation et gestion en plus de l'intérêt écologique. Cette thèse s'inscrit dans cette perspective et se propose d'y contribuer.

Les îles des Mascareignes restent peu étudiées alors qu'elles font partie d'un des 36 hotspots de biodiversité, le hotspot de Madagascar et des îles de l'océan Indien, où l'on retrouve une biodiversité marine exceptionnelle avec de forts taux d'endémisme **(Goodman 2022)**. De plus, plusieurs études ont montré une faible connectivité entre l'éco-région des Mascareignes et de Madagascar, la zone continentale la plus proche, pour les coraux **(Oury 2022)**, les hydrozoaires **(Postaire et al. 2017)**, les échinodermes **(Hoareau et al. 2013)** ou bien les poissons **(Muths et al. 2015)**.

4.1. Les sites d'études

Pour inventorier et étudier le cryptobiome des Mascareignes, 54 ARMS (à La Réunion : 46 ; Rodrigues : 8), issus de trois campagnes d'échantillonnage, ont été utilisés (Figure 1.4). La première campagne de déploiement a été réalisée en 2014, avec l'installation de 12 ARMS à La Réunion et 9 ARMS à Rodrigues (Tableau 1.2) et les deux suivantes lors de cette étude doctorale. Durant la deuxième campagne 27 ARMS ont été déployés fin 2018 et début 2019. Neuf sites ont été sélectionnés le long de la côte Ouest et Sud-Ouest de La Réunion, allant de Pain de Sucre (commune de Saint-Paul) à l'Ouest de l'île jusqu'à Grande Anse (Petite-île) au Sud-Ouest, pour les patrons de distribution du cryptobiome récifale en fonction de la distance (diversité β). Sur chaque site, 3 ARMS ont été déployés permettant ainsi de créer un jeu de données avec distances entre les ARMS allant

de 2 m (intra-site), 20 m (sites les plus proches), jusqu'à 60 km entre les sites les plus éloignés (Figure 1.4). Si l'étude par barcoding de la répartition de certains taxons a pu être effectuée, l'étude spatiale par métabarcoding n'a pu être traitée ici en raison d'un problème de séquençage d'une partie des échantillons au niveau du prestataire. La troisième campagne a permis le déploiement de deux séries de 6 ARMS, l'une à la saison chaude et l'autre à la saison fraîche, pour caractériser la cinétique de la colonisation des substrats par le cryptobiome. Ces 12 ARMS ont été posés sur le site du sanctuaire de St-Gilles, déjà inclus dans le plan d'échantillonnage de la seconde campagne, permettant une comparaison avec une durée d'immersion de 2 ans (Tableau 1.2). La dernière relève des ARMS analysés ici a été effectuée fin août 2021.

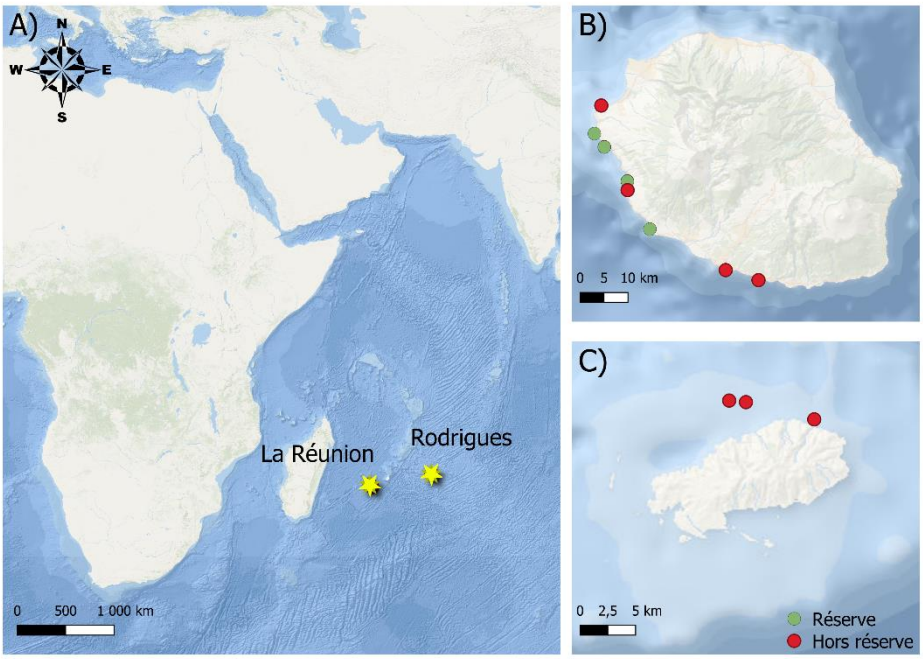


Figure 1.4: Localisation de la zone d'étude (A) et des ARMS déployés à La Réunion (B) et à Rodrigues (C)

Tableau 1.2 : ARMS considérés dans le cadre de ces travaux de thèse

Ile	Campagne	Année de déploiement	# ARMS déployés	# ARMS collectés	Année de collecte	Durée d'échantillonnage
La Réunion	1	2014	3	3	2015	7 mois
			9	4	2018	4 ans
	2	2018/2019	27	27	2020/2021	2 ans
	3	2020	12	12	2020/2021	7 mois, 1 an
Rodrigues	1	2014	9	8	2017	2 ans

4.2. Le démantèlement des ARMS

Avant leur récupération, les ARMS ont été couvertes par une boîte filtrante de 48 µm afin d'empêcher les organismes mobiles de s'échapper. Une fois récupérées, les structures ont été

maintenues immergées dans de l'eau de mer préalablement filtrée (48 µm). Chaque ARMS a ensuite été démontée plaque par plaque selon **Leray & Knowlton (2015)**. Les organismes mobiles ont été séparés en trois fractions à l'aide de tamis stérilisés : 106-500 µm, 500-2000 µm, et > 2 mm. Les deux fractions les plus petites ont été conditionnées pour les analyses de métabarcoding (*cf.* Chapitre 4). Les organismes d'une taille supérieure à 2 mm et les organismes sessiles des plaques ont été échantillonnés et conditionnés pour le barcoding (*cf.* Chapitre 3).

4.3. Problématique

Le premier objectif de cette thèse (Chapitre 2) consiste à établir un pipeline bio-informatique adapté aux analyses métabarcoding métazoaires (COI et 18S) des communautés récifales qui pourra ensuite être employé en routine locale. Les interprétations écologiques basées sur les méthodes d'identification moléculaire sont grandement dépendantes de la complétude des bases de références, en particulier locales. Le cryptobiome des Mascareignes étant peu connu, le second objectif de cette thèse a été de construire un référentiel moléculaire local pour une utilisation adaptée au Sud-Ouest de l'océan Indien (Chapitre 3). Ce référentiel contribuera à documenter le cryptobiome des Mascareignes tout en augmentant la résolution taxonomique de nos analyses, afin de pouvoir ensuite étudier les patrons de colonisation et l'évolution temporelle du cryptobiome récifal et leurs implications dans l'évaluation de la diversité avec les ARMS, troisième objectif de cette thèse (Chapitre 4). Par la suite, le chapitre 5 présente les potentiels et les limites de l'approche ARMS couplé aux méthodes identifications moléculaires pour évaluer la diversité et la distribution des espèces. Pour finir, l'ensemble de ces résultats et les perspectives à l'issu de ce travail sont discutés dans le chapitre 6.

Ainsi, cette thèse s'articule en 6 chapitres. Chaque chapitre peut être lu éventuellement de manière indépendante et possède figures et tableaux, ses propres références bibliographiques, annexes, dont la numérotation lui est propre. Cependant, pour limiter les répétitions, des renvois aux autres chapitres sont présents. Le chapitre 1 présente le contexte et les objectifs de ces travaux. Le chapitre 2 aborde la mise en place du pipeline bio-informatique et le chapitre 3 porte sur les organismes collectés et la création du référentiel barcode. Le chapitre 4 est composé d'un article scientifique en cours de préparation sur les patrons de colonisation du cryptobiome récifal. Le chapitre 5 est un article publié sur la diversité des poissons cryptobenthiques, *Cirripectes*, trouvés au sein des Mascareignes. Pour finir, le chapitre 6 propose une discussion générale synthétisant les implications et les perspectives de ces résultats dans l'étude du cryptobiome.

5. Références du chapitre 1

Al-Rshaidat MMD., Snider A., Rosebraugh S., Devine AM., Devine TD., Plaisance L., Knowlton N., Leray M. (2016) Deep COI sequencing of standardized benthic samples unveils overlooked diversity of Jordanian coral reefs in the northern Red Sea - Genome. Génome 59:724–737.

Beger M., McKENNA SA., Possingham HP. (2007) Effectiveness of surrogate taxa in the design of coral reef reserve systems in the Indo-Pacific. Conservation biology 21:1584–1593.

Bellwood DR. (2001) Regional-Scale Assembly Rules and Biodiversity of Coral Reefs. Science 292:1532–1535. DOI: 10.1126/science.1058635

BIOTOPE (2022) Profil d'écosystème du hotspot de biodiversité Madagascar et des îles de l'océan Indien. Conservation Internationa, VA, USA.

Bouchet P., Bary S., Héros V., Marani G. (2016) How many species of molluscs are there in the world's oceans, and who is going to describe them? Mémoires du Muséum national d'Histoire naturelle (1993).

Brandl SJ., Tornabene L., Goatley CHR., Casey JM., Morais RA., Côté IM., Baldwin CC., Parravicini V., Schiettekatte NMD., Bellwood DR. (2019) Demographic dynamics of the smallest marine vertebrates fuel coral reef ecosystem functioning. Science 364:1189–1192. DOI: 10.1126/science.aav3384

Brown JH., Stevens GC., Kaufman DM. (1996) The Geographic Range: Size, Shape, Boundaries, and Internal Structure. Annual Review of Ecology and Systematics 27:597–623.

Carpenter KE., Abrar M., Aeby G., Aronson RB., Banks S., Bruckner A., Chiriboga A., Cortes J., Delbeek JC., DeVantier L., Edgar GJ., Edwards AJ., Fenner D., Guzman HM., Hoeksema BW., Hodgson G., Johan O., Licuanan WY., Livingstone SR., Lovell ER., Moore JA., Obura DO., Ochavillo D., Polidoro BA., Precht WF., Quibilan MC., Reboton C., Richards ZT., Rogers AD., Sanciangco J., Sheppard A., Sheppard C., Smith J., Stuart S., Turak E., Veron JEN., Wallace C., Weil E., Wood E. (2008) One-Third of Reef-Building Corals Face Elevated Extinction Risk from Climate Change and Local Impacts. Science 321:560–563. DOI: 10.1126/science.1159196

Carvalho S., Aylagas E., Villalobos R., Kattan Y., Berumen M., Pearman JK. (2019) Beyond the visual: using metabarcoding to characterize the hidden reef cryptobiome. Proceedings of the Royal Society B: Biological Sciences 286:20182697. DOI: 10.1098/rspb.2018.2697

CEPF (2015) Madagascar et les Hotspots Iles de l'Océan Indien.

Chabanet P., Bigot L., Nicet J-B., Durville P., Massé L., Mulochau T., Russo C., Tessier E., Obura D. (2016) Coral reef monitoring in the Iles Eparses, Mozambique Channel (2011–2013). Acta Oecologica 72:62–71. DOI: 10.1016/j.actao.2015.10.010

Chao A. (1987) Estimating the Population Size for Capture-Recapture Data with Unequal Catchability. Biometrics 43:783–791. DOI: 10.2307/2531532

Chao A. (1984) Nonparametric Estimation of the Number of Classes in a Population. Scandinavian Journal of Statistics 11:265–270.

Coen LD. (1988) Herbivory by Caribbean majid crabs: feeding ecology and plant susceptibility. Journal of Experimental Marine Biology and Ecology 122:257–276. DOI: 10.1016/0022-0981(88)90127-X

Cracraft J. (1983) Species Concepts and Speciation Analysis. In: *Current Ornithology*. Current Ornithology, Johnston RF (ed) Springer US, New York, NY, p 159–187 DOI: 10.1007/978-1-4615-6781-3_6

David R., Uyarra MC., Carvalho S., Anlauf H., Borja A., Cahill AE., Carugati L., Danovaro R., De Jode A., Feral J-P., Guillemain D., Martire ML., D'Avray LTDV., Pearman JK., Chenuil A. (2019) Lessons from photo analyses of Autonomous Reef Monitoring Structures as tools to detect (bio-)geographical, spatial, and environmental effects. Marine Pollution Bulletin 141:420–429. DOI: 10.1016/j.marpolbul.2019.02.066

De Queiroz K. (2007) Species Concepts and Species Delimitation. Systematic Biology 56:879–886. DOI: 10.1080/10635150701701083

De'ath G., Fabricius KE., Sweatman H., Puotinen M. (2012) The 27-year decline of coral cover on the Great Barrier Reef and its causes. Proceedings of the National Academy of Sciences 109:17995–17999. DOI: 10.1073/pnas.1208909109

Dick MH., Ngai ND., Doan HD. (2020) Taxonomy and diversity of coelobite bryozoans from drift coral cobbles on Co To Island, northern Vietnam. Zootaxa 4747:201–252. DOI: 10.11646/zootaxa.4747.2.1

Dobzhansky Th. (1937) Genetic Nature of Species Differences. The American Naturalist 71:404–420.

Enochs IC., Toth LT., Brandtneris VW., Afflerbach JC., Manzello DP. (2011) Environmental determinants of motile cryptofauna on an eastern Pacific coral reef. Marine Ecology Progress Series 438:105–118. DOI: 10.3354/meps09259

Fisher R., O'Leary RA., Low-Choy S., Mengersen K., Knowlton N., Brainard RE., Caley MJ. (2015) Species Richness on Coral Reefs and the Pursuit of Convergent Global Estimates. Current Biology 25:500–505. DOI: 10.1016/j.cub.2014.12.022

Fisher RA., Corbet AS., Williams CB. (1943) The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. Journal of Animal Ecology 12:42–58. DOI: 10.2307/1411

Gélin P., Postaire B., Fauvelot C., Magalon H. (2017) Reevaluating species number, distribution and endemism of the coral genus Pocillopora Lamarck, 1816 using species delimitation methods and microsatellites. Molecular Phylogenetics and Evolution 109:430–446. DOI: 10.1016/j.ympev.2017.01.018

Glasby T., Connell S. (2001) Orientation and position of substrata have large effects on epibiotic assemblages. Marine Ecology Progress Series 214:127–135. DOI: 10.3354/meps214127

Glynn PW. (1980) Defense by symbiotic crustacea of host corals elicited by chemical cues from predator. Oecologia 47:287–290. DOI: 10.1007/BF00398518

Glynn PW. (2011) In tandem reef coral and cryptic metazoan declines and extinctions. Bulletin of Marine Science 87:767–794.

Goeij JM de., Duyl FC van. (2007) Coral cavities are sinks of dissolved organic carbon (DOC). Limnology and Oceanography 52:2608–2617. DOI: 10.4319/lo.2007.52.6.2608

Goodman SM. (2022) Updated estimates of biotic diversity and endemism for Madagascar—revisited after 20 years. Oryx:1–5. DOI: 10.1017/S0030605322001284

Harrison PL., Booth DJ. (2007) Marine Ecology. In: *Coral reefs: naturally dynamic and increasingly disturbed ecosystems*. SD Connell & BM Gillanders, Oxford University Press, South Melbourne, Vic., p 316–377

Hazeri G., Rahayu DL., Subhan B., Sembiring A., Anggoro AW., Ghozali AT., Madduppa HH. (2019) Latitudinal species diversity and density of cryptic crustacean (Brachyura and Anomura) in micro-habitat Autonomous Reef Monitoring Structures across Kepulauan Seribu, Indonesia. Biodiversitas Journal of Biological Diversity 20. DOI: 10.13057/biodiv/d200540

Hoareau TB., Boissin E., Paulay G., Bruggemann JH. (2013) The Southwestern Indian Ocean as a potential marine evolutionary hotspot: perspectives from comparative phylogeography of reef brittle-stars. Journal of Biogeography 40:2167–2179. DOI: 10.1111/jbi.12155

Hoban ML., Williams JT. (2020) *Cirripectes matatakaro*, a new species of combtooth blenny from the Central Pacific, illuminates the origins of the Hawaiian fish fauna. PeerJ 8:e8852. DOI: 10.7717/peerj.8852

Hoegh-Guldberg O., Mumby PJ., Hooten AJ., Steneck RS., Greenfield P., Gomez E. (2007) Coral reefs under rapid climate change and ocean acidification. Science 318:1737–1742. DOI: 0.1126/science.1152509

Hoegh-Guldberg O., Pendleton L., Kaup A. (2019) People and the changing nature of coral reefs. Regional Studies in Marine Science 30:100699. DOI: 10.1016/j.rsma.2019.100699

Idjadi JA., Edmunds PJ. (2006) Scleractinian corals as facilitators for other invertebrates on a Caribbean reef. Marine Ecology Progress Series 319:117–127.

IFRECOR (2020) Etat de santé des récifs coralliens, herbiers marins et mangroves des d'outre-mer.

Ip YCA., Chang JJM., Oh RM., Quek ZBR., Chan YKS., Bauman AG., Huang D. (2022) Seq' and ARMS shall find: DNA (meta)barcoding of Autonomous Reef Monitoring Structures across the tree of life uncovers hidden cryptobiome of tropical urban coral reefs. Molecular Ecology:1–20. DOI: 10.1111/mec.16568

Jaccard P. (1912) The Distribution of the Flora in the Alpine Zone.1. New Phytologist 11:37–50. DOI: 10.1111/j.1469-8137.1912.tb05611.x

Jones GP., McCormick MI., Srinivasan M., Eagle JV. (2004) Coral decline threatens fish biodiversity in marine reserves. Proceedings of the National Academy of Sciences 101:8251–8253. DOI: 10.1073/pnas.0401277101

Knowlton N. (1993) Sibling species in the sea. Annual review of ecology and systematics 24:189–216.

Knowlton N., Brainard RE., Fisher R., Moews M., Plaisance L., Caley MJ. (2010) Coral reef biodiversity. In: *Life in the World's Oceans: Diversity Distribution and Abundance*. Blackwell Publishing Ltd, Oxford, UK, p 65–74

Knowlton N., Jackson JBC. (2008) Shifting Baselines, Local Impacts, and Global Change on Coral Reefs. PLOS Biology 6:e54. DOI: 10.1371/journal.pbio.0060054

Legendre P. (2014) Interpreting the replacement and richness difference components of beta diversity. Global Ecology and Biogeography 23:1324–1334. DOI: 10.1111/geb.12207

Leray M., Knowlton N. (2015) DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. Proceedings of the National Academy of Sciences 112:2076–2081. DOI: 10.1073/pnas.1424997112

Maddison WP. (1997) Gene Trees in Species Trees. Systematic Biology 46:523–536. DOI: 10.1093/sysbio/46.3.523

Mallela J., Milne BC., Martinez-Escobar D. (2017) A comparison of epibenthic reef communities settling on commonly used experimental substrates: PVC versus ceramic tiles. Journal of Experimental Marine Biology and Ecology 486:290–295. DOI: 10.1016/j.jembe.2016.10.028

Mayr E. (1948) The Bearing of the New Systematics on Genetical Problems The Nature of Species. In: *Advances in Genetics*. Demerec M (ed) Academic Press, p 205–237 DOI: 10.1016/S0065-2660(08)60469-1

McKeon CS., Stier AC., McIlroy SE., Bolker BM. (2012) Multiple defender effects: synergistic coral defense by mutualist crustaceans. Oecologia 169:1095–1103. DOI: 10.1007/s00442-012-2275-2

Mellin C., Delean S., Caley J., Edgar G., Meekan M., Pitcher R., Przeslawski R., Williams A., Bradshaw C. (2011) Effectiveness of Biological Surrogates for Predicting Patterns of Marine Biodiversity: A Global Meta-Analysis. PLoS ONE 6:e20141. DOI: 10.1371/journal.pone.0020141

Mihalitsis M., Morais RA., Bellwood DR. (2022) Small predators dominate fish predation in coral reef communities. PLOS Biology 20:e3001898. DOI: 10.1371/journal.pbio.3001898

Moberg F., Folke C. (1999) Ecological goods and services of coral reef ecosystems. Ecological economics 29:215–233. DOI: 10.1016/S0921-8009(99)00009-9

Monroy-Velázquez LV., Rodríguez-Martínez RE., Blanchon P., Alvarez F. (2020) The use of artificial substrate units to improve inventories of cryptic crustacean species on Caribbean coral reefs. PeerJ 8:e10389. DOI: 10.7717/peerj.10389

Mora C., Aburto-Oropeza O., Ayala Bocos A., Ayotte PM., Banks S., Bauman AG., Beger M., Bessudo S., Booth DJ., Brokovich E., Brooks A., Chabanet P., Cinner JE., Cortés J., Cruz-Motta JJ., Cupul Magaña A., DeMartini EE., Edgar GJ., Feary DA., Ferse SCA., Friedlander AM., Gaston KJ., Gough C., Graham NAJ., Green A., Guzman H., Hardt M., Kulbicki M., Letourneur Y., López Pérez A., Loreau M., Loya Y., Martinez C., Mascareñas-Osorio I., Morove T., Nadon M-O., Nakamura Y., Paredes G., Polunin NVC., Pratchett MS., Reyes Bonilla H., Rivera F., Sala E., Sandin SA., Soler G., Stuart-Smith R., Tessier E., Tittensor DP., Tupper M., Usseglio P., Vigliola L., Wantiez L., Williams I., Wilson SK., Zapata FA. (2011) Global Human Footprint on the Linkage between Biodiversity and Ecosystem Functioning in Reef Fishes. PLoS Biology 9:e1000606. DOI: 10.1371/journal.pbio.1000606

Mouillot D., Leprêtre A. (1999) A comparison of species diversity estimators. Population Ecology 41:203–215. DOI: 10.1007/s101440050024

Muths D., Tessier E., Bourjea J. (2015) Genetic structure of the reef grouper *Epinephelus merra* in the West Indian Ocean appears congruent with biogeographic and oceanographic boundaries. Marine Ecology 36:447–461. DOI: 10.1111/maec.12153

Nichols PK., Timmers M., Marko PB. (2021) Hide 'n seq: Direct versus indirect metabarcoding of coral reef cryptic communities. Environmental DNA 4:93–107. DOI: 10.1002/edn3.203

Obura D., Gudka M., Samoilys M., Osuka K., Mbugua J., Keith DA., Porter S., Roche R., van Hooidonk R., Ahamada S., Araman A., Karisa J., Komakoma J., Madi M., Ravinia I., Razafindrainibe H., Yahya S., Zivane F. (2022) Vulnerability to collapse of coral reef ecosystems in the Western Indian Ocean. Nature Sustainability 5:104–113. DOI: 10.1038/s41893-021-00817-0

Odum HT., Odum EP. (1955) Trophic structure and productivity of a windward coral reef community on Eniwetok Atoll. Ecological monographs 25:291–320. DOI: 10.2307/1943285

Osman EO., Suggett DJ., Voolstra CR., Pettay DT., Clark DR., Pogoreutz C., Sampayo EM., Warner ME., Smith DJ. (2020) Coral microbiome composition along the northern Red Sea suggests high plasticity of bacterial and specificity of endosymbiotic dinoflagellate communities. Microbiome 8:8. DOI: 10.1186/s40168-019-0776-5

Oury N. (2022) De la délimitation des espèces à la diversité intra-coloniale : apport de la génomique chez les coraux du genre *Pocillopora* dans l'Indo-Pacifique. phdthesis, Université de la Réunion

Padial JM., Miralles A., De la Riva I., Vences M. (2010) The integrative future of taxonomy. Frontiers in Zoology 7:16. DOI: 10.1186/1742-9994-7-16

Pearman JK., Anlauf H., Irigoien X., Carvalho S. (2016) Please mind the gap – Visual census and cryptic biodiversity assessment at central Red Sea coral reefs. Marine Environmental Research 118:20–30. DOI: 10.1016/j.marenvres.2016.04.011

Pearman JK., Aylagas E., Voolstra CR., Anlauf H., Villalobos R., Carvalho S. (2019) Disentangling the complex microbial community of coral reefs using standardized Autonomous Reef Monitoring Structures (ARMS). Molecular Ecology 28:3496–3507. DOI: 10.1111/mec.15167

Pearman JK., Chust G., Aylagas E., Villarino E., Watson JR., Chenuil A., Borja A., Cahill AE., Carugati L., Danovaro R., David R., Irigoien X., Mendibil I., Moncheva S., Rodríguez-Ezpeleta N., Uyarra MC., Carvalho S. (2020) Pan-regional marine benthic cryptobiome biodiversity patterns revealed by metabarcoding Autonomous Reef Monitoring Structures. Molecular Ecology n/a. DOI: 10.1111/mec.15692

Pearman JK., Leray M., Villalobos R., Machida RJ., Berumen ML., Knowlton N., Carvalho S. (2018) Cross-shelf investigation of coral reef cryptic benthic organisms reveals diversity patterns of the hidden majority. Scientific Reports 8:1–17. DOI: 10.1038/s41598-018-26332-5

Peterson AT., Navarro-Sigüenza AG. (1999) Alternate Species Concepts as Bases for Determining Priority Conservation Areas. Conservation Biology 13:427–431. DOI: 10.1046/j.1523-1739.1999.013002427.x

Pfenninger M., Schwenk K. (2007) Cryptic animal species are homogeneously distributed among taxa and biogeographical regions. BMC Evolutionary Biology 7:121. DOI: 10.1186/1471-2148-7-121

Plaisance L., Caley MJ., Brainard RE., Knowlton N. (2011) The Diversity of Coral Reefs: What Are We Missing? PLOS ONE 6:e25026. DOI: 10.1371/journal.pone.0025026

Plaisance L., Knowlton N., Paulay G., Meyer C. (2009) Reef-associated crustacean fauna: biodiversity estimates using semi-quantitative sampling and DNA barcoding. Coral Reefs 28:977–986.

Plotnick RE., Smith FA., Lyons SK. (2016) The fossil record of the sixth extinction. Ecology Letters 19:546–553. DOI: 10.1111/ele.12589

Postaire B., Gélin P., Bruggemann JH., Magalon H. (2017) One species for one island? Unexpected diversity and weak connectivity in a widely distributed tropical hydrozoan. Heredity 118:385–394. DOI: 10.1038/hdy.2016.126

Ransome E., Geller JB., Timmers M., Leray M., Mahardini A., Sembiring A., Collins AG., Meyer CP. (2017) The importance of standardization for biodiversity comparisons: A case study using autonomous reef monitoring structures (ARMS) and metabarcoding to measure cryptic diversity on Mo'orea coral reefs, French Polynesia. PLOS ONE 12:e0175066. DOI: 10.1371/journal.pone.0175066

Reaka ML. (1987) Adult-juvenile interactions in benthic reef crustaceans. Bulletin of Marine Science 41:108–137.

Reaka-kudla ML. (1997) The global biodiversity of coral reefs: A comparison with rainforests. In: *Biodiversity II: Understandig and Protecting Our Natural Resources*. Joseph Henry / National Academy Press, Washington, D.C., p 83–108

Richter C., Wunsch M. (1999) Cavity-dwelling suspension feeders in coral reefs: A new link in reef trophodynamics. Marine Ecology Progress Series 188:105–116. DOI: 10.3354/meps188105

Roberts CM., McClean CJ., Veron JE., Hawkins JP., Allen GR., McAllister DE., Mittermeier CG., Schueler FW., Spalding M., Wells F. (2002) Marine biodiversity hotspots and conservation priorities for tropical reefs. Science 295:1280–1284.

Rogers A., Blanchard JL., Mumby PJ. (2014) Vulnerability of Coral Reef Fisheries to a Loss of Structural Complexity. Current Biology 24:1000–1005. DOI: 10.1016/j.cub.2014.03.026

Rothans TC., Miller AC. (1991) A link between biologically imported particulate organic nutrients and the detritus food web in reef communities. Marine Biology 110:145–150. DOI: 10.1007/BF01313101

Salvat B. (1992) Coral reefs - A challenging ecosystem for human societies. Global Environmental Change 2:12–18. DOI: 10.1016/0959-3780(92)90032-3

Scheffers SR., Van Soest RWM., Nieuwland G., Bak RPM. (2010) Coral Reef Framework Cavities: Is Functional Similarity Reflected in Composition of the Cryptic Macrofaunal Community? Atoll Research Bulletin 583:1–24. DOI: 10.5479/si.00775630.583.1

Shannon CE. (1948) A mathematical theory of communication. The Bell System Technical Journal 27:379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x

Siddik AA., Al-Sofyani AA., Ba-Akdah MA., Satheesh S. (2019) Invertebrate recruitment on artificial substrates in the Red Sea: role of substrate type and orientation. Journal of the Marine Biological Association of the United Kingdom 99:741–750. DOI: 10.1017/S0025315418000887

Sørensen TJ. (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. I kommission hos E. Munksgaard.

Sorokin YI. (1993) Coral Reef Ecology, Springer Science&Business Media. Springer, Berlin., 978-3-540-60532-4

Spalding M., Ravilious C., Green EP. (2001) World Atlas of Coral Reefs. University of California Press., 446 p., 978-0-520-23255-6

Steele MA. (1999) Effects of shelter and predators on reef fishes. Journal of Experimental Marine Biology and Ecology 233:65–79. DOI: 10.1016/S0022-0981(98)00127-0

Stewart HL., Holbrook SJ., Schmitt RJ., Brooks AJ. (2006) Symbiotic crabs maintain coral health by clearing sediments. Coral Reefs 25:609–615. DOI: 10.1007/s00338-006-0132-7

Steyaert M., Lindhart M., Khrizman A., Dunbar RB., Bonsall MB., Mucciarone DA., Ransome E., Santodomingo N., Winslade P., Head CEI. (2022a) Remote reef cryptobenthic diversity: Integrating autonomous reef monitoring structures and in situ environmental parameters. Frontiers in Marine Science 9:932375. DOI: 10.3389/fmars.2022.932375

Steyaert M., Mogg A., Dunn N., Dowell R., Head CEI. (2022b) Observations of coral and cryptobenthic sponge fluorescence and recruitment on autonomous reef monitoring structures (ARMS). Coral Reefs. DOI: 10.1007/s00338-022-02283-2

Sully S., Burkepile DE., Donovan MK., Hodgson G., van Woesik R. (2019) A global analysis of coral bleaching over the past two decades. Nature Communications 10:1264. DOI: 10.1038/s41467-019-09238-2

Tang CQ., Leasi F., Obertegger U., Kieneke A., Barraclough TG., Fontaneto D. (2012) The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. Proceedings of the National Academy of Sciences 109:16208–16212. DOI: 10.1073/pnas.1209160109

Tittensor DP., Mora C., Jetz W., Lotze HK., Ricard D., Berghe EV., Worm B. (2010) Global patterns and predictors of marine biodiversity across taxa. Nature 466:1098–1101. DOI: 10.1038/nature09329

Todd PA. (2008) Morphological plasticity in scleractinian corals. Biological Reviews 83:315–337. DOI: 10.1111/j.1469-185X.2008.00045.x

Villalobos R., Aylagas E., Pearman J., Cúrdia J., Lozano-Cortés D., Coker D., Jones B., Berumen M., Carvalho S. (2022) Inter-annual variability patterns of reef cryptobiota in the central Red Sea across a shelf gradient. Scientific Reports 12. DOI: 10.1038/s41598-022-21304-2

Whittaker RH. (1960) Vegetation of the Siskiyou Mountains, Oregon and California. Ecological Monographs 30:279–338. DOI: 10.2307/1943563

Wilkinson C. (2008) Status of Coral Reefs of the World: 2008. Australian Institute of Marine Science, Townsville, Australia.

Winston JE. (1986) An Annotated Checklist of Coral-Associated Bryozoans. AMERICAN MUSEUM NOVITATES:42.

Wolf NG., Bermingham EB., Reaka ML. (1983) Relationships between fishes and mobile benthic invertebrates on coral reefs. 11.

Zimmerman TL., Martin JW. (2004) Artificial Reef Matrix Structures (Arms): An Inexpensive and Effective Method for Collecting Coral Reef-Associated Invertebrates. Gulf and Caribbean Research 16:59–64. DOI: 10.18785/gcr.1601.08

Chapitre 2 : L'ADN pour évaluer et identifier la biodiversité

Résumé :

Les études des communautés à large échelle sont difficiles à mettre en place en raison de forts investissements financiers et temporels. Pourtant elles sont nécessaires pour comprendre la composition, la dynamique et la résilience des écosystèmes. Historiquement, la description des communautés était limitée aux taxons faciles à échantillonner et sur de petites surfaces en raison des quantités de données à expertiser face au faible nombre de taxonomistes **(Plaisance et al. 2009)**. Les récentes avancées en taxonomie moléculaire et les approches de métabarcoding ont révolutionné les capacités d'analyses des communautés, en particulier pour les petits organismes **(Plaisance et al. 2011)**. La composition et la diversité des communautés peuvent être évaluées avec ces méthodes en : (1) comparant les distances entre les séquences ADN, (2) regroupant les séquences proches en groupes de similarité, par exemple les OTU (Unité Taxonomique Opérationnelle), et (3) en assignant ces OTU à des taxons grâce aux bases de référence **(Valentini et al. 2009)**. Cependant, la majeure partie du travail en écologie moléculaire est le traitement du nombre important de séquences ADN disponible en fin de séquençage. L'amélioration des techniques de séquençage a permis une meilleure description de la diversité biologique intra et interspécifique mais également celle de tous les artéfacts moléculaires qui étaient restés jusqu'alors invisibles **(Taberlet et al. 2018)**. Par conséquent, le principal objectif du traitement bio-informatique est de filtrer ces artefacts et de produire des jeux de données robustes. A l'heure actuelle, il n'existe aucune méthode standardisée pour les études d'écologie moléculaire. Chaque pipeline, logiciel et paramètre employés a ses avantages et ses inconvénients qu'il convient de mettre en relation avec la diversité étudiée et les questions écologiques posées.

De ce fait, ce deuxième chapitre vise à décrire la sélection et la mise en place d'un pipeline bio-informatique adapté aux analyses métabarcoding métazoaires (18S et COI) des communautés récifales. Dans un premier temps, il établit l'état des lieux des connaissances sur les méthodes de taxonomie moléculaire, le *barcoding* et le métabarcoding, puis dans un second temps sur les pipelines bio-informatiques employés dans l'analyse écologique des communautés. Dans un troisième temps, il définit le protocole moléculaire et le pipeline bio-informatique mis en place dans cette étude doctorale.

1. La taxonomie moléculaire, le *barcoding*

1.1. Généralités

Les premiers travaux basés sur les différences entre des séquences nucléiques pour étudier les relations entre les espèces datent de **Woese & Fox** en 1977. Mais c'est au cours des 20 dernières années, que le *barcoding,* dans le cadre du projet Barcode of Life, est devenu populaire pour identifier rapidement les organismes et simplifier l'analyse des communautés **(Hebert et al. 2003a ; Hebert & Gregory 2005)**. Basé sur le séquençage d'une courte région standard d'ADN et spécifique aux espèces, il repose sur l'hypothèse que les distances génétiques au sein d'une même espèce sont plus faibles que les distances génétiques entre espèces différentes **(Moore 1995)**. Le gène utilisé pour l'identification va dépendre du règne des organismes ciblés. Ainsi c'est le **COI** (Cytochrome Oxydase sous unité 1) qui est employé pour les animaux (Metazoa ; **Hebert et al. 2003b**), utilisé dans cette étude ; le rbcL et matK pour les plantes (Plantae **; CBOL Plant Working Group et al. 2009**) ou l'ITS pour les champignons (Fungi ; **Schoch et al. 2012**), non utilisés dans cette étude.

Faisant partie du génome mitochondrial, le COI est hérité de la mère et n'est pas affecté par les recombinaisons **(Dawid & Blackler 1972 ; Avise et al. 1979).** Les mitochondries sont présentes dans de nombreux tissus et en de nombreux exemplaires rendant l'ADN mitochondrial (mtDNA) facilement accessibles pour l'amplification par **PCR** (*polymerase chain reaction*), même dans les échantillons dégradés **(Lansman et al. 1981)**. Le COI est un gène qui évolue rapidement comparé à l'ADN nucléaire **(Brown et al. 1979 ; Moore 1995)**, le rendant efficace pour distinguer des espèces qui ont divergé récemment **(Hajibabaei et al. 2007)**. Cette nouvelle méthode d'identification a suscité tellement d'enthousiasme **(Janzen 2004)** que certains auteurs ont suggéré que les méthodes traditionnelles devaient être remplacées par une taxonomie entièrement basée sur la divergence des séquences ADN **(Wiens & Penkrot 2002 ; Tautz et al. 2003 ; Hebert et al. 2004 ; Meierotto et al. 2019)**. Cependant, une séquence barcode récupérée d'un spécimen inconnu n'informe pas sur le statut de l'espèce (nouvelle espèce ou espèce non-référencée dans la base de référence ; **Moritz**

Le barcode idéal

Le barcode ADN idéal doit répondre à plusieurs critères :

1) La région du gène séquencé doit être quasiment identique pour les individus d'une même espèce mais différente entre les espèces
2) Il doit être standardisé, avec la même région ADN pour les différents groupes taxonomiques
3) La région ciblée doit contenir assez d'informations phylogénétiques pour assigner les espèces non référencées à leur groupe taxonomique
4) Il doit être robuste avec les sites d'amorces très conservés, s'amplifier et se séquencer correctement
5) La région ADN doit être assez courte pour permettre l'amplification d'ADN dégradé (Valentini et al. 2009).

& Cicero 2004 ; Rubinoff 2006). Le *barcoding* intervient en complément de la taxonomie traditionnelle pour faciliter l'identification de lignées différentes, comme pour les espèces cryptiques **(Ahrens et al. 2007 ; Wang et al. 2018)**. Ainsi la découverte et description des espèces est principalement le domaine des taxonomistes, sur des bases morphologiques et maintenant souvent également moléculaires, et l'identification des espèces, après l'établissement de la taxonomie, celui du *barcoding* **(De Salle 2006 ; Wägele et al. 2011)**. Les séquences obtenues sont comparées à une base de référence, qui regroupe des séquences de spécimens d'identification connue, pour assigner les spécimens **(Austerlitz et al. 2009)**. Les différences observées entre deux séquences peuvent être dues à la variation intra-spécifique (individus différents mais appartenant à même espèce) ou à la variation inter-spécifique (individus appartenant à deux espèces différentes). Toutefois, il n'existe pas de seuil universel pour différencier ces deux types de divergences **(Wiemers & Fiedler 2007)**. La différentiation de deux espèces doit reposer sur une approche de taxonomie intégrative avec l'emploi combiné de critères indépendants qui peuvent être la monophylie réciproque, une différentiation morphologique, écologique, phénotypique, etc.

1.2. Les limitations du *barcoding* et les approches pour réduire les biais

La principale limite du *barcoding* est sa dépendance aux bases de données de référence car elles restent incomplètes malgré les efforts de la communauté scientifiques durant la dernière décennie, notamment de la part du *Consortium for the Barcode of Life* (CBOL) **(Ratnasingham & Hebert 2007)**. Le manque de séquences de référence conduit souvent à un assignement taxonomique peu précis, tel qu'à l'ordre ou à la famille. Les deux principales raisons à ces lacunes dans les bases de données sont : (1) l'identification des espèces prend du temps et demande l'expertise de taxonomistes, qui sont surchargés et de moins en moins nombreux ; (2) de nombreuses espèces restent rares ou difficiles à échantillonner.

Une seconde limite au *barcoding* est l'utilisation d'amplicons mitochondriaux courts tel que le COI qui conduit à des biais d'estimation de la biodiversité **(Galtier et al. 2009 ; Krehenwinkel et al. 2017)**. La divergence des séquences mitochondriales ne représente pas nécessairement la divergence des espèces. En effet, la divergence des séquences mitochondriales dépend de nombreux facteurs, comme la dispersion des individus ou des stratégies de reproduction **(Mao et al. 2010)**. Par exemple, pour les espèces où seules les femelles sont philopatriques (se reproduisent à l'endroit où elles sont nées), les génomes mitochondriaux sont très divergents alors que les génomes nucléaires montrent peu de différenciation **(Prugnolle & de Meeus 2002)**. A l'inverse, les phénomènes d'introgression (transfert de gènes d'une espèce vers le pool génétique d'une autre

espèce) peuvent conduire à l'homogénéisation des gènes mitochondriaux, alors que les génomes nucléaires sont différenciés **(Irwin et al. 2009)**, allant jusqu'à des génomes mitochondriaux identiques entre espèces différentes (capture mitochondriale, **Perea et al. 2016**). Par ailleurs, des pseudogènes mitochondriaux (copie nucléaire d'un fragment mitochondrial, NUMTs, **Puertas & González-Sánchez 2020**) peuvent être amplifiés lors des PCR et conduire à un mauvais assignement taxonomique **(Bensasson et al. 2001 ; Song et al. 2008 ; Buhay 2009)**. Ces biais peuvent être limités en utilisant plusieurs gènes indépendants lors du *barcoding*, ainsi qu'en combinant l'approche moléculaire à des informations morphologiques, écologiques et géographiques **(Dupuis et al. 2012 ; Puillandre et al. 2012)**. L'utilisation combinée de gènes indépendants mitochondriaux/nucléaires facilite les assignements taxonomiques et permet de tester les hypothèses d'espèces cryptiques **(Dupuis et al. 2012)**. Plus généralement, la combinaison de marqueurs permet d'augmenter la résolution taxonomique lors du *barcoding*.

Le choix des marqueurs à utiliser dans l'analyse des communautés dépend des objectifs de l'étude (précision et échelle taxonomique recherchées) en fonction des avantages et inconvénients de chacun des marqueurs (Tableau 2.1). Les marqueurs complémentaires au COI les plus utilisés dans les études de biodiversité sont l'ADN ribosomique 16S et l'ADN ribosomique 18S. Le 16S présente les avantages découlant de l'ADN mitochondriale comme le COI, mais étant moins variable **(Knowlton & Weigt 1998)**, il permet d'amplifier et de détecter un plus large spectre taxonomique, tel que les procaryotes et les bactéries et a une meilleure précision pour certains taxons (Tableau 2.1). Le 18S issu de l'ADN nucléaire est un marqueur très conservé qui permet une large couverture taxonomique des eucaryotes **(Hillis & Dixon 1991)**, le rendant efficace pour les études de diversité à large échelle taxonomique **(Tang et al. 2012)**. L'utilisation combinée du COI avec le 16S et / ou du 18S permet de palier au principal biais du COI résultant de sa forte variabilité au niveau des amorces. Cette variabilité implique l'utilisation d'amorces spécifiques aux taxons ciblés **(Sanna et al. 2009)** ou bien celle d'amorces dégénérées (un mélange d'amorces correspondant à différentes séquences possibles ou comprenant des bases spécifiques comme l'inosine) pour minimiser les pertes de détection **(Rose et al. 2003 ; Boyce et al. 2009)**. Plusieurs études rapportent des difficultés d'amplification et de détection de certains taxons avec le COI tels que les nématodes, les plathelminthes et les hydrozoaires (Tableau 2.1 ; **Bhadury et al. 2006 ; Sanna et al. 2009 ; Peña Cantero et al. 2009**).

Tableau 2.1 : Avantages et inconvénients des marqueurs utilisés pour la taxonomie moléculaire. Les références sont indiquées en indice et listées en annexe 2.1.

Marqueurs	Avantages	Inconvénients
COI (mitochondrial)	- Très variable : bonne discrimination des espèces[1] - Optimisé pour les métazoaires[2] - Insertions et délétions rares : facilite l'alignement[3] - Code pour les protéines : différent taux d'évolution en fonction de la position des codons, source potentielle d'information du rang taxonomique[4] - Large base de référence disponible (Barcode of Life Database) - Bonne détection des arthropodes et annélides[5]	- Très variable, même aux amorces : nécessite des amorces spécifiques à certains taxons[6] - Légère surestimation de la biodiversité[7] - Mauvaise détections des : nématodes[8–10], plathelminthes[6] et hydrozoaires[11]
16S (mitochondrial)	- Large spectre taxonomique - Idéal pour les procaryotes et les études microbiennes[12] - Bonne détection des hydrozoaires[11]	
18S (nucléaire)	- Très conservé[13] : large couverture des taxons eucaryotes - Bonne détection des : nématodes[10,14] et échinodermes[5,15] - Efficace pour des études de diversité à large échelle taxonomique[7]	- Très conservé : faible discrimination des taxons - Sous-estimation importante des espèces[7] - Non efficace pour des études de biodiversité à fine échelle[7,16]

1.3. Le séquençage à haut débit

Ces 20 dernières années les techniques de séquençage ont connu de grandes améliorations, passant du séquençage manuel aux séquenceurs automatisés, puis à de nouvelles approches de séquencage. Le *barcoding* était traditionnellement basé sur le séquençage Sanger et ne pouvait séquencer qu'un seul échantillon à la fois. Avec un coût de 5 à 10 € l'échantillon, cette méthode était coûteuse pour les études de biodiversité **(Taberlet et al. 2012 ; Bohmann et al. 2014)**. De plus, les échantillons complexes devaient être individualisés par clonage, c'est-à-dire, insérés dans un plasmide ou BAC pour ensuite être intégrés dans des bactéries pour amplification, et suivi du séquençage pour de nombreuses colonies **(Lamoril et al. 2008)**. Le développement du Séquençage de Nouvelle Génération (**NGS**) offre une alternative plus rapide et moins onéreuse **(Kozich et al. 2013)** en traitant plusieurs milliers de spécimens simultanément dans les échantillons complexes **(Shokralla et al. 2015 ; Meier et al. 2016 ; Srivathsan et al. 2019)**. Les séquenceurs NGS, dits de 2ème génération, se basent sur l'une des 4 méthodes principales de séquençage de l'ADN : le séquençage par synthèse, le pyroséquençage, le séquençage par ligation ou le séquençage des semiconducteurs

ioniques. Un tableau comparatif de séquenceurs de 1ère, 2ème et 3ème génération est disponible en Tableau 2.2.

Tableau 2.2 : Tableau comparatif des séquenceurs basé sur : **Shokralla et al. 2012 ; Loman et al. 2012 ; Laver et al. 2015 ; Derocles et al. 2018 ; Bansal & Boucher 2019**.

Séquenceur	**Sanger**	**MiSeq**	**HiSeq 1000**	**Ion Torrent PGM (Chip 318 V2)**	**MinION**
Société	Applied Biosystems	Illumina	Illumina	Life Technologies	Oxford Nanopore Technologies
Génération	1	2	2	2	3
Méthode d'amplification	PCR	Bridge PCR	Bridge PCR	PCR en émulsion	Aucune ou PCR
Méthode de séquençage	Dideoxy chain termination	Synthèse	Synthèse	Polymerase synthesis	Nanopore
Capacité de séquençage par run		8 Gb	300 Gb	2 Gb	20 Gb
Taille moyenne des *reads*	1 000 b	2 x 300 b	2 x 150 b	400 b	Taille du brin ADN
Taux erreur (%)	0,001	0,1	0,1	1	38
Type d'erreur	Indel Substitution	Substitution	Substitution	Indel	Indel Substitution
Durée du run	3	27 h	8,5 j	7 h	48h
Nombre de *reads* par run	96	$3x10^8$(*paired*)	$8x10^9$ (*paired*)	$8,2x10^7$	$4.7x10^4$
Coût $/run		1 000	11 000	750	
Coût $ / Million de bases	500	0,03	0,15	0,10	6,44-17,90
Barcoding	✓	✓	✓	✓	✓
Métabarcoding	×	✓	✓	✓	✓
Gestion des homopolymères	-	✓	✓	×	×

Le séquençage par synthèse (**SBS** : *sequencing by synthesis*) est la méthode la plus répandue (80-90 %) et est utilisée par les séquenceurs Illumina qui dominent actuellement le marché **(van Dijk et al. 2014 ; Wang et al. 2018)**. Le séquençage Illumina utilise des réactifs de synthèse incluant les amorces, l'ADN polymérase et de 4 terminateurs réversibles marqués différemment. À chaque cycle de séquençage, un nucléotide marqué par fluorescence est incorporé au brin d'ADN répliqué, puis chaque *cluster* est excité par une source lumineuse pour récupérer les caractéristiques du signal fluorescent et identifier le nucléotide (Figure 2.1, étape 5 ; **Illumina Sequencing Technology**). Le nombre de cycles de séquençage va déterminer la longueur du *read* **(Buermans & den Dunnen 2014)**.

Les séquenceurs Illumina MiSeq produisent actuellement les *reads* parmi les plus longs pour des séquenceurs de 2ème génération avec 300 paires de base (bp) en *paired-end* **(Schirmer et al. 2016)** et sont utilisables pour le *barcoding* d'amplicons de moins de 590 pb **(Shokralla et al. 2015)**.

Le séquençage *paired-end* consiste à séquencer le fragment ADN depuis une extrémité (Figure 2.1, étape 5), puis de recommencer à partir de l'autre extrémité (Figure 2.1, étape 8). Les deux *reads* (R1 et R2) séquencés sont enregistrés dans des fichiers séparés et pourront être fusionnés ultérieurement lors des étapes bio-informatiques si une partie est chevauchante entre les deux (*cf.* 3.1.5).

La taille de lecture des fragments ADN reste la principale limitation des NGS car elle est généralement inférieure à celle du barcode COI (~680 bp ; **Leray et al. 2013**). Cette limite peut être contournée en séquençant plusieurs amplicons qui se chevauchent pour obtenir le barcode complet (Figure 2.2B ; **Kennedy et al. 2020)**, mais cela augmente le problème de l'adéquation des amorces pour l'amplification **(Shokralla et al. 2015)**.

L'avantage des séquenceurs Illumina réside dans la simplicité de la préparation des librairies à deux niveaux, où l'indexage peut se faire de trois façons (Figure 2.3). Le plus souvent, les PCR en deux étapes sont utilisées. La première étape permet d'amplifier la séquence ciblée, échantillon par échantillon. Lors de la deuxième étape, un index unique pour chaque échantillon (courte séquence distinctive), ainsi que les adaptateurs nécessaires au séquençage sont incorporés aux amplicons. Cette approche permet de passer de nombreux échantillons en un seul séquençage (Figure 2.3A). Le multiplexage de PCR avec des barcodes ciblant des gènes différents permet de réduire également les étapes de préparation **(Quick et al. 2017)**. L'utilisation d'amorces pré-taguées lors de la première PCR a été employé dans cette étude et permet de séquencer plusieurs échantillons dans une même banque (Figure 2.3B ; **Corse et al. 2017**). Il aurait été possible d'aller encore plus loin, avec l'utilisation d'amorces incluant le tag, l'adaptateur et l'index ce qui permet la création de librairie en une seule PCR (Figure 2.3C ; **Kozich et al. 2013 ; Fadrosh et al. 2014**). Le double indexage permet de regrouper les échantillons sous trois niveaux d'informations, le marqueur ciblé avec l'amorce, l'échantillon avec le tag, et le groupe d'échantillons avec l'*Index* (Figure 2.4).

Certaines approches visent à réduire les coûts, telles que la limitation du nombre d'extractions en regroupant des spécimens de lignées différentes **(Hinsinger et al. 2015 ; Kerdrel et al. 2020)**. Des auteurs suppriment l'étape d'extraction et effectuent directement les PCR **(Wong et al. 2014 ; Yeo et al. 2018)**. Ainsi, le coût de l'échantillon est réduit à 0,18 – 0,88 € **(Meier et al. 2016 ; Srivathsan et al. 2019)**. Cependant ces approches ne sont applicables qu'à certains types d'échantillons.

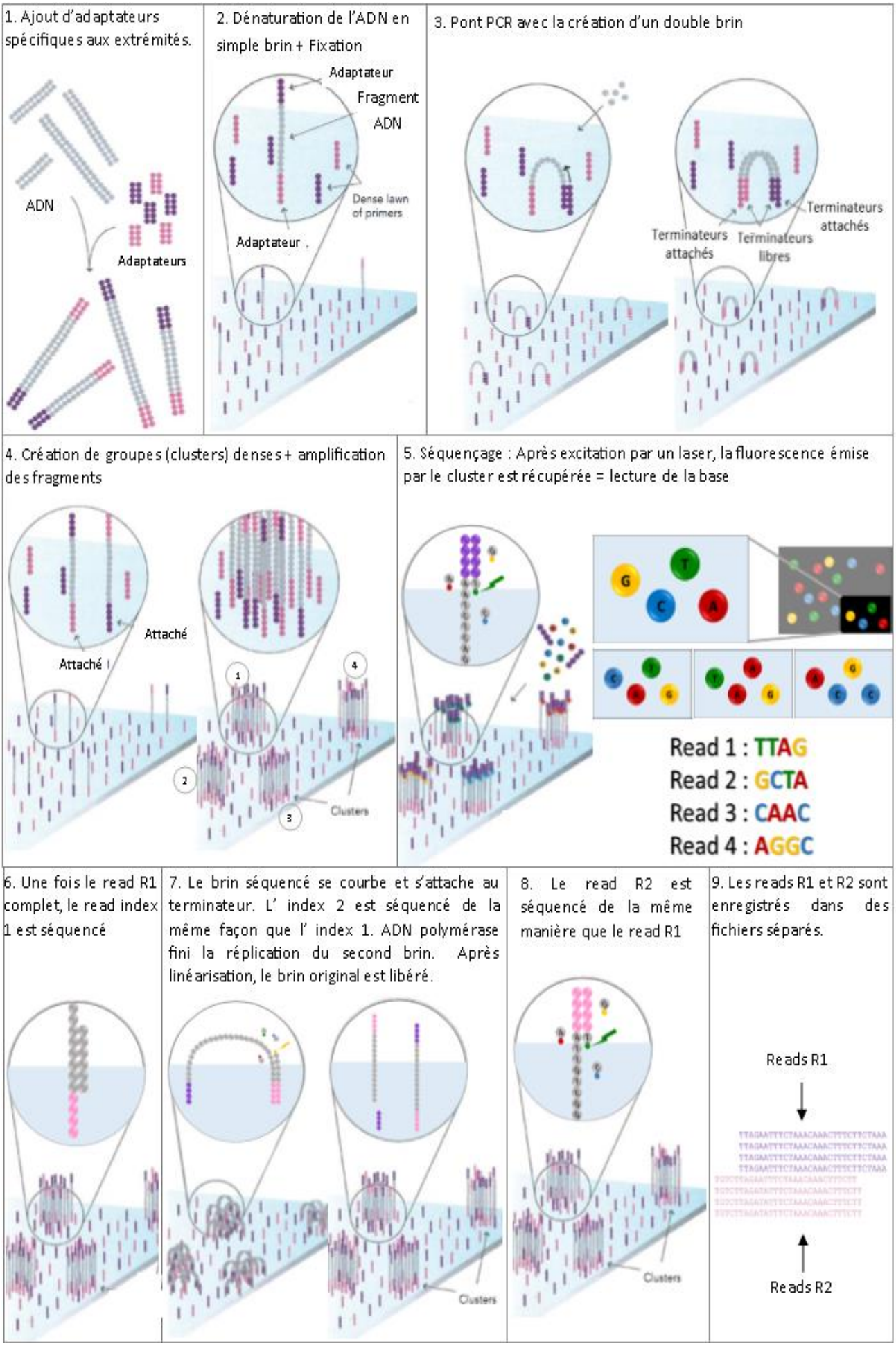


Figure 2.1 : Processus de séquençage par synthèse *paired-end* utilisé par les séquenceurs Illumina. Adapté de Illumina Sequencing Technology

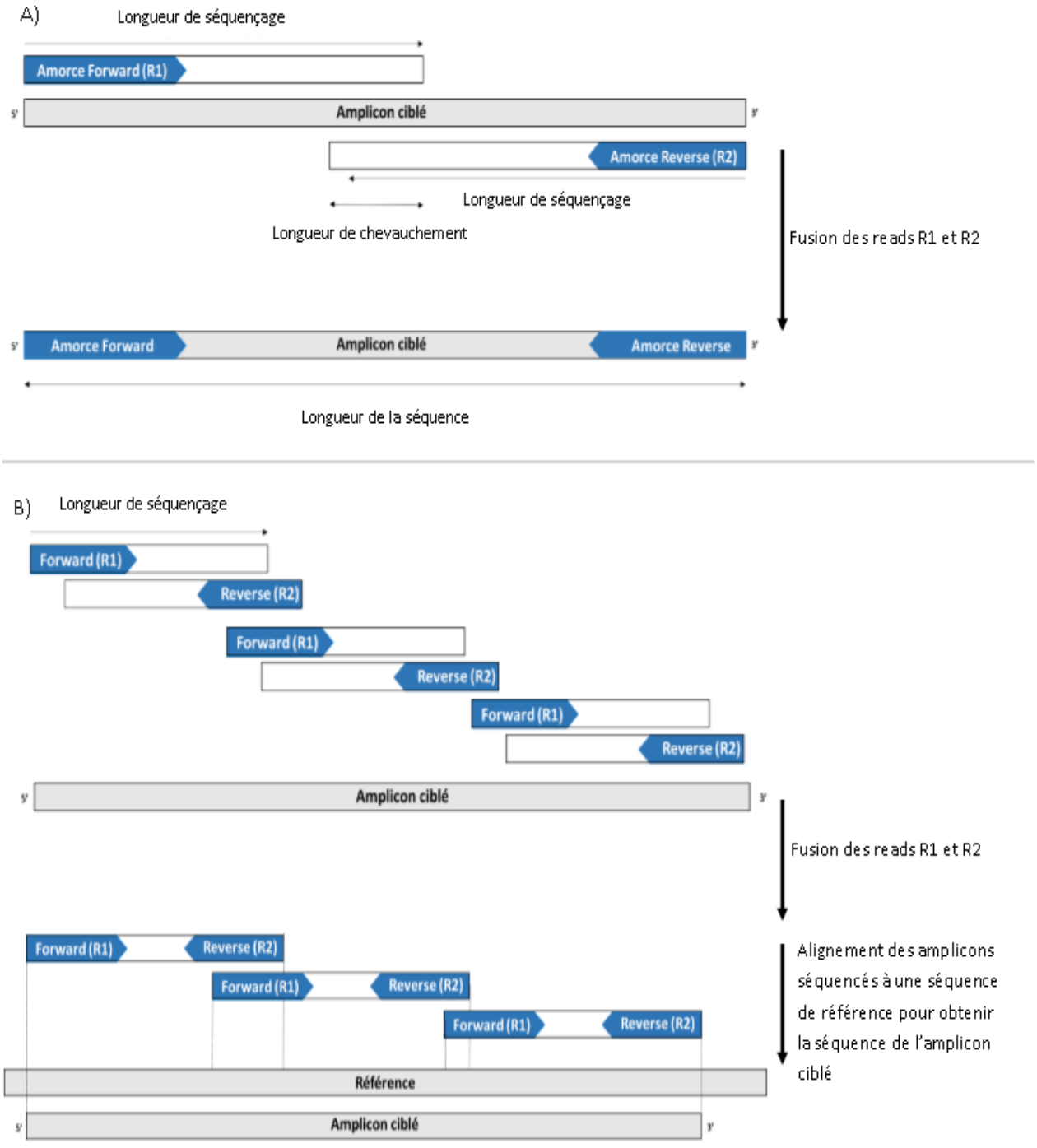


Figure 2.2 : Assemblage des *reads paired-end* en fonction de la taille de l'amplicon. A) l'amplicon ciblé est plus court que la taille de lecture des fragments x 2. Les reads R1 et R2 sont fusionnés pour reconstituer la séquence de l'amplicon ciblé ; B) l'amplicon cible est plus long que la taille de lecture (x2). L'amplicon initialement ciblé est divisé en plusieurs amplicons de taille inférieure à celle de lecture (x2). Une fois ces amplions séquencés, les reads R1 et R2 sont fusionnés et les séquences résultantes sont alignées à une séquence de référence pour reconstituer la séquence de l'amplicon initialement ciblé.

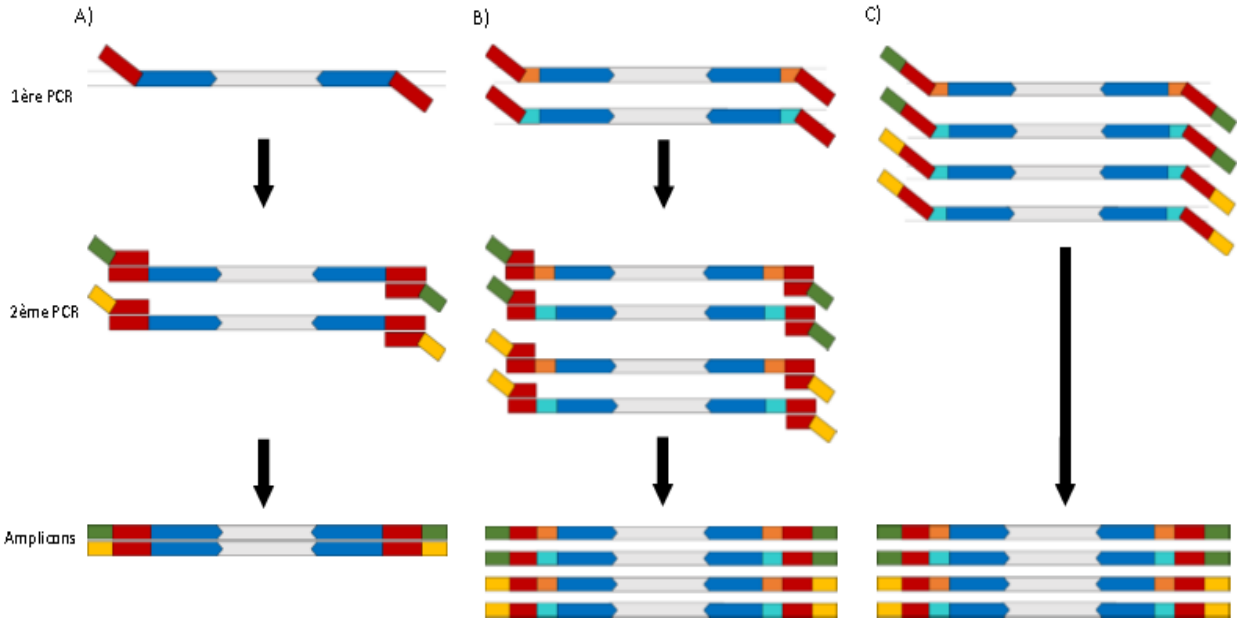


Figure 2.3 : Les différentes stratégies d'indexage des librairies Illumina. A) PCR en deux étapes ; B) PCR en deux étapes avec double indexage ; C) PCR en une étape avec double indexage. En bleu, les amorces ; en orange et turquoise, les Tag ; en rouge, les adaptateurs ; en vert et jaune les index.

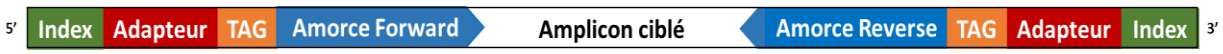


Figure 2.4 : Amplicon après double indexage et avant séquençage

2. Le *barcoding* de communauté, le métabarcoding

2.1. Généralités

La réduction des coûts et l'alternative rapide qu'offre le NGS a permis rendre accessible le *barcoding* pour les analyses de communautés à travers le métabarcoding. Le métabarcoding applique la technique du *barcoding* à un échantillon contenant plusieurs individus de différentes espèces **(Yu et al. 2012 ; Gibson et al. 2014)**. L'ensemble de l'ADN contenu dans cet échantillon est extrait et les barcodes ciblés vont être récupérés et regroupés en **OTU** (Unité Taxonomique Opérationnelle) qui pourront être identifiés et / ou servir pour les analyses de diversité. Comme pour le *barcoding*, il est recommandé d'utiliser plusieurs marqueurs pour récupérer l'ensemble des taxons présents dans l'échantillon.

Le métabarcoding regroupe trois approches différentes en fonction des finalités visées : (1) une approche basée sur la taxonomie, (2) une approche *de novo* et (3) une approche basée sur les groupes fonctionnels **(Cordier et al. 2020)**.

La première approche est la plus similaire aux suivis classiques de biodiversité et du *barcoding*. Son objectif principal est d'identifier les taxons présents dans l'échantillon et par la suite d'évaluer la congruence de cette méthode avec les inventaires morphologiques déjà existants **(Elbrecht et al. 2017)**. Cette approche est donc limitée aux taxons qui ont été préalablement identifiés morphologiquement et référencés dans les bases de données (*cf.* 1.2). Cette approche peut être utilisée pour identifier l'ADN environnemental (**eDNA**) présent en faible concentration et sous forme libre dans le milieu (eau, sédiment **; Thomsen et al. 2012 ; Leduc et al. 2019**) ou pour identifier de l'ADN présent en de fortes concentrations (*bulk sample*), tel que les biofilms, les pièges d'animaux de petites tailles ou les communautés sessiles encroûtantes **(Leray & Knowlton 2015 ; Taberlet et al. 2018)**.

La seconde approche a pour objectif de découvrir de nouveaux organismes ou ensembles d'organismes bioindicateurs sans qu'ils aient obligatoirement un assignement taxonomique. Cette méthode permet de considérer la biodiversité dans sa globalité et d'accéder à des espèces bioindicatrices, sous forme d'OTU, qui étaient jusqu'alors inaccessibles **(Cordier et al. 2020)**. En effet, les micro-organismes, les protistes ou bien la meiofaune, peuvent refléter rapidement et à différentes intensités les perturbations environnementales **(Creer et al. 2010 ; Payne 2013)**.

La troisième approche repose sur l'écologie fonctionnelle et la structure des communautés pour comprendre les processus qui régissent les assemblages et évaluer les impacts des perturbations. Différentes métriques peuvent être utilisées : celles relatives aux communautés, celles relatives à la phylogénie et celles relatives aux interactions biologiques. L'utilisation de données relatives à la composition des communautés, telle la diversité alpha, permet de mettre en évidence des perturbations anthropiques. Différentes études ont montré une corrélation positive entre la diversité des communautés et la distance à la perturbation anthropique, que ça soit pour les foraminifères **(Laroche et al. 2018)**, la macrofaune ou les communautés bactériennes des sédiments marins **(Stoeck et al. 2018)**. A l'inverse, certaines études montrent une augmentation de la diversité à l'approche de la perturbation pour des communautés bactériennes des sédiments **(Nogales et al. 2011)** ainsi que celles associées aux coraux, avec l'émergence d'espèces opportunistes **(Ziegler et al. 2016)**. Ces résultats démontrent que la diversité alpha seule n'est pas suffisante pour servir d'indicateur. L'utilisation des données relatives à la phylogénie repose sur le concept de conservatisme de niche écologique qui postule que les espèces apparentées ont des

Métabarcoding vs métagénomique

Le métabarcoding cible un fragment court et spécifique d'ADN alors que la métagénomique prend en compte l'ensemble des génomes contenus dans un échantillon.

fonctions écologiques similaires **(Webb et al. 2002 ; Pearman et al. 2008)**. Par extension, sous cette hypothèse, plus la diversité phylogénétique est importante, meilleure est la résilience de l'écosystème **(Cadotte et al. 2012 ; Oliver et al. 2015)**. Ainsi la communauté est caractérisée par les fonctions écologiques qu'elle fournit et non par les taxons qu'elle contient. Cependant ce concept est limité par le fait que tous les traits fonctionnels n'ont pas nécessairement un signal phylogénétique **(Srivastava et al. 2012)** biaisant l'évaluation des impacts par les méthodes de phylogénie. L'étude des interactions biologiques, telles les interactions trophiques, peuvent être utilisées pour déduire la stabilité jusqu'à la perte de biodiversité de l'écosystème **(Estrada 2007 ; Gilbert 2009)**, et les différents mécanismes d'assemblage ou d'interaction **(Vázquez 2005 ; Williams 2011)**. Les identifications issues des données de métabarcoding peuvent être utilisées pour reconstituer le réseau trophique **(Compson et al. 2019)**. Toutefois cette approche est récente et nécessite d'être développé pour fournir des données fiables **(Cordier et al. 2020)**.

2.2. Les limitations du métabarcoding

Si le métabarcoding permet d'étendre le *barcoding* à l'échelle de communautés écologiques, il reste toutefois sujet aux mêmes contraintes en ce qui concerne la dépendance aux bases de données (en fonction de l'approche choisie) et les biais de marqueurs (*cf.* 1.2). À ceux-là s'ajoute plusieurs limitations propres au métabarcoding :

(1) les séquences ADN obtenues ne peuvent pas être associées à un spécimen. Les protocoles actuels d'extraction ADN conduisent à la destruction des échantillons **(Wang et al. 2018)**. Effectuer un sous-échantillonnage (tissus, photos) avant l'extraction permet de limiter la perte d'informations morphologiques. Certains auteurs utilisent des protocoles d'extraction qui permettent de conserver l'intégrité morphologique de l'échantillon (lyse rapide entre 1h à 12h ; **Porco et al. 2010 ; Andersen & Mills 2012**). Ces méthodes non-destructrices, lorsqu'elles sont appliquées à des extractions de communautés, peuvent être biaisées en faveurs des groupes taxonomiques qui relâchent leur ADN plus rapidement que d'autres (lyse rapide des organismes aux corps mous ; **Carew et al. 2018 ; Marquina et al. 2019**). Selon le même principe que les séquences ne peuvent pas être reliées à un spécimen, les séquences de différents marqueurs d'un même spécimen ne peuvent pas être reliées entre elles ce qui limite la résolution taxonomique **(Kerdrel et al. 2020)**.

(2) le séquençage à grande échelle conduit à des traitements bio-informatiques conséquents pour analyser les séquences issues de l'échantillon. Or ils peuvent également biaiser les résultats en fonction de leurs ordres d'exécution et des paramètres choisis (*cf.* Chapitre 2.3). Par exemple, il est difficile de distinguer les *reads* similaires d'une séquence ou de séquences appartenant à des

espèces proches **(Gompert et al. 2014 ; Elbrecht et al. 2018)**. En fonction de la qualité du filtrage bio-informatique, le métabarcoding peut conduire à des sur ou sousestimations de la biodiversité.

- *Exemple de surestimation liée au processus bio-informatique :* Les séquences erronées qui ne sont pas retirées lors du filtrage, seront considérées comme des OTU. Il s'agit notamment des séquences des espèces non cibles (bactéries, champignons) et les chimères résultantes de la liaison de produits PCR incomplets de différents taxons **(Elbrecht et al. 2017)** ou bien d'erreurs d'assemblage des *reads* (*cf.* Chapitre 2.3.2.2). Les chimères peuvent être détectées avec les logiciels bio-informatiques appropriés et les espèces non-cibles avec les bases de référence (Chapitre 2.3.2.2). Cependant la filtration des séquences parasites reste difficile et n'est pas encore complétement résolue **(Kennedy et al. 2020)**.

-*Exemple de sous-estimation liée au processus bio-informatique :* Le seuil de regroupement des séquences sous un même OTU peut être supérieur à la variation intra-spécifique des organismes et ainsi regrouper deux espèces sous un seul OTU. Le barcode de certaines espèces contient des homopolymère (répétition d'une même base) qui sont détectés comme des erreurs de séquençages (ex. commun chez les ascidies ; **Griggio et al. 2014**).

(3) l'une des dernières limitations à prendre en compte est le manque actuel de précision dans l'assignation du métabarcoding. Pour l'instant les études de métabarcoding sont réalisées sur des compartiments de la biodiversité encore peu référencé moléculairement et conduisent à la génération d'une quantité importante de séquences non identifiables **(Wang et al. 2018)**.

Par ailleurs, même si le métabarcoding informe sur la composition des communautés. Il n'est pas encore possible d'obtenir des mesures précises de l'abondance des différents taxons **(Elbrecht & Leese 2015)**. Ceci est principalement dû aux différences d'efficacité des PCR en fonction des taxons conduisant à des biais dans l'abondance des séquences. Toutefois la réponse d'un taxon au cours d'une PCR peut être prévisible. L'abondance relative d'un taxon est corrélée linéairement aux nombres de *reads* récupérés **(Kennedy et al. 2020)**. La pente de cette corrélation diffère selon les taxons. Ainsi, si les pentes sont connues pour tous les taxons de l'échantillon, des corrections peuvent être appliquées pour estimer l'abondance relative des taxons. L'inconvénient est que les facteurs de correction doivent être développés individuellement pour chaque taxon **(Thomas et al. 2016)**, ce qui n'est pas faisable dans des échantillons très complexes aux taxons inconnus ou mal connus. La quantification des abondances étant nécessaire pour de nombreuses analyses de biodiversité, de nombreux efforts sont déployés pour optimiser les méthodes de métabarcoding quantitatif **(Saitoh et al. 2016 ; Krehenwinkel et al. 2017)**.

Malgré ces limites actuelles, le *barcoding* et le métabarcoding sont des approches complémentaires aux méthodes d'identification classiques basées sur la morphologie. Leur utilisation permet de mener de nombreuses études pour caractériser différentes communautés d'organismes (ex. bactéries, champignons, protistes, plantes, animaux) et ce dans des habitats très variés (ex. sédiments **(Pawlowski et al. 2021)**, permafrost **(Johnson et al. 2007)**, rivières **(Hajibabaei et al. 2011)**, lacs **(Rivera et al. 2018)**, récifs coralliens **(Carvalho et al. 2019)**. Aujourd'hui, l'utilisation du métabarcoding s'étend à de nombreuses applications parmi lesquelles la protection et la conservation de la biodiversité **(Thomsen & Willerslev 2015)**, les interactions trophiques **(Leray et al. 2015 ; Albaina et al. 2016)**, la paléoécologie **(Capo et al. 2017)**, l'écotoxicologie **(Pascault et al. 2014)** et la surveillance environnementale **(Pochon et al. 2015)**.

Par ailleurs, le développement des séquenceurs de 3ème génération permet de faire abstraction de la principale limite de *barcoding*, la longueur de *reads*. Oxford Nanopore Technologies (ONT) et Pacific Biosciences (PacBio) offrent des plateformes de séquençages qui semblent limitées par la seule longueur de l'amplicon **(Kennedy et al. 2020)**. Toutefois ces plateformes doivent encore être améliorées et souffrent de forts taux d'erreurs de séquençage, cependant en voie d'amélioration considérable **(Sahlin & Medvedev 2021)**.

2.3. Le protocole moléculaire mis en place pour l'étude du cryptobiome récifal par métabarcoding

Lors de cette étude doctorale, nous avons employé une combinaison de marqueurs nucléaires (18S) et mitochondriaux (COI) pour déterminer la diversité qui compose le cryptobiome récifal des Mascareignes par métabarcoding. Le plan d'échantillonnage et l'acquisition des échantillons sont détaillés dans le Chapitre 3.2.1.

2.3.1. L'extraction ADN des échantillons

Pour chaque échantillon, l'ADN a été extrait à partir de 10 g de fraction organique à l'aide du kit d'extraction ADN du sol DNeasy Powermax (Qiagen) en suivant le protocole standardisé établi par oceanarms.org et les recommandations de Leray & Knowlton (**2015**). Ainsi, les échantillons ont été centrifugés à 2 500 rcf pendant 10 min pour éliminer l'EtOH. 180 µl de solution ATL ont été ajouté et les tubes ont été vigoureusement vortexé 1 min. Comme recommandé par Leray & Knowlton (**2015**), 20 µl de protéinase K (10mg/mL) a été ajouté à la solution C1 et les échantillons ont été incubés sous agitation pendant une nuit à 56°C. Le reste du protocole d'extraction a suivi les instructions du kit DNeasy Powermax Soil. L'ADN extrait a été purifié à l'aide du kit de nettoyage DNeasy PowerClean Pro de Qiagen avant l'amplification par PCR. Les extractions ont été réalisées

en 5 lots de 31 échantillons. Pour chaque lot, des échantillons contrôles ont été réalisés avec (1) un contrôle négatif pour l'extraction (T_{ex}) composé de 10 ml d'eau *DNA free* soumis au protocole d'extraction (en position 17 du lot) et (2) un contrôle négatif pour les aérosols d'ADN (T_{pai}) composé de 10 ml d'eau *DNA free* qui est resté ouvert sur la paillasse pendant les étapes d'extraction et de purification, comme recommandé dans (**Corse et al. 2017** ; Figure 2.5). Pour un lot d'extraction, un témoin positif (T_{pos}) a été réalisé en première position de la série, où les 10 g de fraction organique ont été remplacés par un morceau de 1 g cher de dinde (*Meleagris gallopavo*) et 10 ml d'eau *DNA free*. Pour ce lot, trois T_{ex} ont été réalisés en position 5, 17 et 29 pour observer une potentielle contamination par l'ADN de dinde ainsi que sa persistance dans le temps.

2.3.2. Le multiplexage et les plans PCR mis en place pour maximiser la détection des erreurs

Quatre réplicats PCR ont été réalisées pour amplifier un le fragment COI de ~310 bp (*forward* : GGWACWGGWTGAACWGTWTAYCCYCC ; *reverse* : TAIACYTCIGGRTGICCRAARAAYCA ; **Leray et al. 2013**) et le fragment 18S de ~550 pb (*forward* : CTGGTGCCAGCAGCCGCGGYAA ; *reverse* : TCCGTCAATTYCTTTAAGTT ; **Machida & Knowlton 2012**). La taq Hotstart Multiplex de Qiagen a été utilisée avec les concentrations recommandées par le fabricant, les programmes ont été 94°C 15 min, puis 30 cycles de 94° 20s, 55° 1min, 72° 1 min pour le 18S et 94°C 15 min, puis 35 cycles de 94° 20s, 50° 1min, 72° 1 min pour le COI après optimisation. Afin de regrouper plusieurs échantillons en une seule préparation de banque, les amorces PCR présentées dans le tableau 2.3 ont été tagguées avec de courtes séquences nucléotidiques de 6 pb (Tableau 2.4). Chaque plaque PCR comprenait un contrôle négatif pour la PCR (T_{pcr}) qui consistait en 4 fois 12µl de mélange PCR et d'amorces tagguées. Les combinaisons tag-amorce et la position du témoin dans la plaque PCR ont été successivement changés (Figure 2.5). Le T_{pcr} indique le niveau de contamination croisée pendant la préparation du mélange PCR et des plaques (amorces marquées mais pas de matrice d'ADN). Nous n'avons pas besoin d'un contrôle négatif (mélange PCR sans amorce) pour évaluer le niveau de mistagging comme Corse et al. **(2017)** en raison de l'utilisation du même tag pour les amorces *forward* et *reverse* par échantillon, les mistagging seront identifiables par la présence de tags différents aux deux extrémités des fragments. Après amplifications, les réplicats PCR ont été regroupés puis multiplexés en fonction de leur tag et de leur amorce (Figure 2.5). La préparation des banques a été externalisée en juin 2022 à l'Institut du Cerveau et de la Moëlle Epinière (Hôpital de la Pitié-Salpêtrière, Paris) en PE250 avec un SP Reagent Kit (500 cycles) sur l'Illumina NovaSeq 6000 avec un kit de préparation des banques Illumina par ligation.

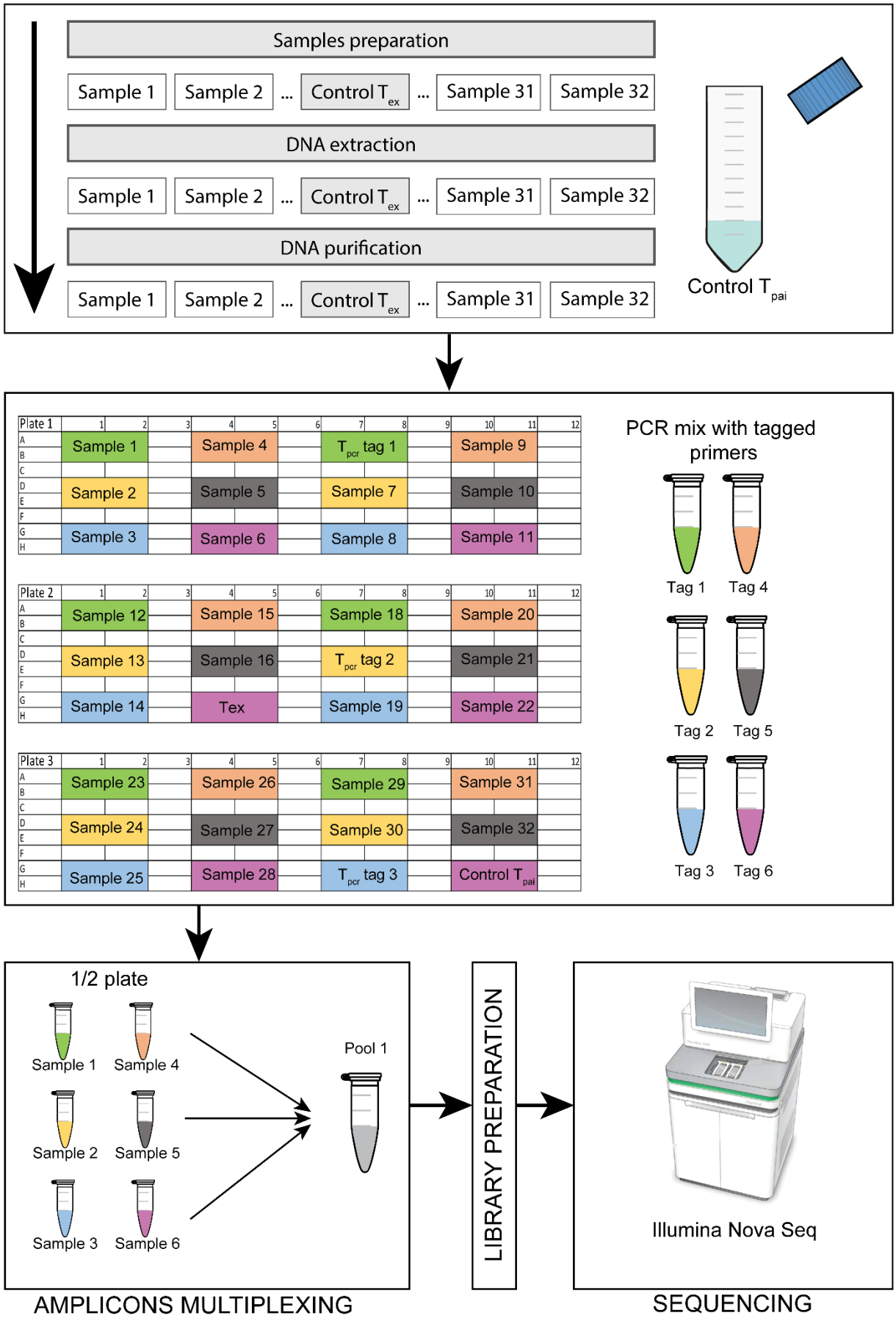


Figure 2.5 : Synthèse du protocole moléculaire

Tableau 2.3 : Amorces employées pour le métabarcoding

Gene	Nom de l'amorce	Séquence (5' - 3')	Direction	Taille de l'amplicon (bp)	Références
C01	mlCOlint	GGWACWGGWTGAACWGTWTAYCCYCC	forward	~310	Leray et al. 2013
	jgHCO2198	TAIACYTCIGGRTGICCRAARAAYCA	reverse		
18S	18Smk1F	CTGGTGCCAGCAGCCGCGGYAA	forward	~550	Machida & Knowlton 2012
	18Smk2R	TCCGTCAATTYCTTTAAGTT	reverse		

Tableau 2.4 : Tags employés dans ce projet doctoral. Référence : Leray & Knowlton **(2015)**

Nom du tag	Séquence (5' - 3')
Tag_01	AGACGC
Tag_02	AGTGTA
Tag_03	ACTAGC
Tag_04	ACAGTC
Tag_05	ATCGAC
Tag_06	ATGTCG
Tag_09	ACGTAT

3. Le traitement des données moléculaire par la bio-informatique

Aujourd'hui la majeure partie du travail en écologie moléculaire est le traitement du nombre important de séquences ADN disponible en fin de séquençage. En plus de la gestion de données en termes de stockage et capacité de traitement, un des aspects cruciaux du traitement des données de métabarcoding est la gestion des erreurs. L'amélioration de la couverture de séquençage des échantillons a permis une meilleure description de la diversité biologique intra et interspécifique mais également celle de tous les artefacts moléculaires qui étaient restés invisibles **(Taberlet et al. 2018)**. Dès lors, le principal objectif du traitement bio-informatique est de filtrer ces artefacts, conduisant ainsi à diminuer la taille du jeu de données et à produire une liste d'OTU à l'instar des listes d'espèce dans les inventaires taxonomiques. Par la suite ces résultats sont analysés statistiquement de façon analogue aux autres études écologiques. De nombreux pipelines bio-informatiques sont disponibles pour analyser les données de métabarcoding (Tableau 2.5 ; Tableau 2.6), tous sont essentiellement composés de trois étapes séquentielles : (1) le prétraitement des données avec le démultiplexage et le contrôle de la qualité des séquences, (2) le regroupement des séquences en OTU (*clustering*) et (3) l'assignement taxonomique de ces OTU. Toutefois la grande diversité des pipelines bio-informatiques disponibles pour analyser les données de séquençage posent certaines limites à la comparaison des études de métabarcoding. A l'heure actuelle, aucune méthode standardisée n'a fait l'unanimité pour les études d'écologies moléculaires **(Prodan et al. 2020)**. La diversité de plateformes de séquençage, de marqueurs moléculaires et de pipelines bio-informatiques rendent difficile la comparaison des études.

Tableau 2.5 : Principaux pipelines bio-informatiques pour analyser les données de métabarcoding. L'implémentation des différentes étapes est signalée par un code couleur : en vert : l'étape est implémentée ; en rouge : l'étape n'est pas implémentée ; en orange : l'étape n'est pas implémentée à l'origine mais le pipeline peut être modifié pour la rajouter. Références : **Schloss et al. 2009 ; Edgar 2013 ; McMurdie & Holmes 2013 ; Taberlet et al. 2018 ; Bolyen et al. 2018 ; Geneious Prime 2019.2.3 2019 ; Ratnasingham 2019 ; Schloss 2020**

	Qiime	Geneious	Mothur	UPARSE	OBItools	mBRAVE	R
Input							
Multiplexed files	🟢	🟢	🟠	🔴	🟠		🟢
Pair-end files	🟢	🟢	🟢		🟢		🟢
Single-end files	🟢	🟢			🔴		🟢
I^{ary} analyses : basic handling							
Demultiplexing	🟢	🟢	🟠	🔴	🟠		🟢
Make new files	🟢	🟢	🟠		🔴		🟢
Annotate sequences	🔴	🔴			🟢		🔴
Merge paired reads	🟢	🟢			🟢		
Allow differences in overlap region	🟢			🟢			🟢
Trim	🟢	🟢		🟢		🟢	
Left	🟢	🟢		🟢		🟢	🟢
Right	🔴			🔴		🟢	🟢
Lenght	🟢			🟢		🟢	
Quality	🟢	🟢	🟢				
Overlap			🟢				
Filtering	🟢	🟢		🟢	🟢		
Discard short reads		🟢	🟢		🟢	🟢	
Discard long reads			🟢		🟢		
Dereplication	🟢	🟢	🟢	🟢	🟢	Not available	
Discard singletons	🟢			🟢	🟢	🟢	
Low quality reads	🟠		🟢		🟠	🟢	
Homopolymers			🟢				
II^{ary} analyses : sequence classification							
OTU Clustering			🟢		🔴		
Discard chimeres	🟢	🟢	🟢		🔴		🟢
OTU Assignment							
BLAST	🟢	🟢	🟢		🔴		
LCA	🟢		🟢		🟢		
RDP			🟢				
Look reverse complement	🟢		🟢		🟢		
Custom database	🟠		🟢		🟢		

Tableau 2.6 : Aperçu des logiciels et pipelines disponibles, avec leurs avantages et inconvénients respectifs, pour les différentes étapes de l'analyse des données de séquençages.

Etape	Outils	Avantages	Inconvénients	Référence	Choix
Demultiplexage	Cutadapt	Algorithme le plus utilisé ; Implémenté dans QIIME 2 ; Démultiplexe dans les deux sens		Martin 2011	●
	QIIME-split-library.py	Paramètre : 100% match ; Elimine les reads de mauvaises qualités	Simple script python de QIIME1	DiBattista et al. 2020	●
Contrôle qualité	Fastqc	Le plus utilisé Bonne visualistation Complet	Non implementé dans QIIME2	Andrews 2010	●
	Qiime demux-summarize	Implementé dans Qiime	Visualisation difficile		
Débruitage	DADA2	Débruite les Forward et Reverse séparément, le plus adapté aux séquenceurs Illumina		Callahan et al. 2016	●
	UNOISE				
	Mothur		Ne peut être executé en dehors du pipeline Mothur	Prodan et al. 2020	
Déréplication	DADA2	Meilleure précision et moins de séquences incorrectes Inclus dans l'étape de débruitage	Le paramètre Maxee biaise le donnée ne pas l'utiliser	Prodan et al. 2020	●
	Qiime2-Deblur	Sensibilité et spécificité	Faible taux de conversion	Prodan et al. 2020	
	USEARCH-UNOISE3	Sensibilité et spécificité		Edgar (2016)	
Fusion des paired-end	FLASH	Maximise la longueur de chevauchement	Ne prend pas en compte le quality score	Magoc and Salzberg 2011	
	PANDASeq	Prends en compte le quality score de toutes les bases Le plus performant pour de petit overlap Inclut un filtrage de la qualité des séquences en fusionnant les reads	Assume que tous les reads peuvent être fusionnés	Masella et al. 2012	
	COPE	Prends en compte le quality score des bases non-concordantes		Liu et al. 2012	

Etape	Outils	Avantages	Inconvénients	Référence	Choix
	PEAR	Prend en compte la longueur de chevauchement et les quality scores		Zhang et al. 2013	
	Mothur		Ne peut être executé en dehors du pipeline Mothur	Prodan et al. 2020	
	DADA2	Actuellement l'algorithme le plus utilisé Inclus dans l'étape de débruitage			●
Chimera detection	UPARSE				
	DADA2	Inclus dans l'étape de débruitage			●
	UCHIME	Robuste en présence de mutation		Mysara et al. 2015	
	ChimeraSlayer	Robuste en présence d'indels et avec des faibles divergences entre les séquences		Mysara et al. 2015	
	DECIPHER	Bonne détection des petites chimères (short chimérique range)		Mysara et al. 2015	
Création des OTU	UCLUST / Qiime-UCLUST	Non adapté à la création d'OTU	OTU erroné et surestimation des OTU	Edgar 2010 Prodan et al. 2020	
	Vsearch de novo clustering	Analyse avec absence de base de référence Regroupe tous les reads	Lents	Rideout et al. 2014	●
	Vsearch à références fermées	Rapide, utile pour les gros jeux de données Meilleurs taxonomie et phylogénie	Ne détecte pas de nouveaux OTU Necessite une base de référence	Rideout et al. 2014	
	Paramètre :Vsearch	Regroupe tous les reads Rapide, fonctionne en partie en parallèle	Necessite une base de référence	Rideout et al. 2014	
	CROP		Non adapté aux jeux de données complexes et aux reads longs	Chen et al. 2013	
	Swarm	N'utilise pas de seuil de regroupement basé sur un choix arbitraire	Pas de seuil de regroupement ; mauvaise reproductibilité	Mahé et al. 2023	
	MOTHUR			Prodan et al. 2020	
	MeShClust		Lent	James et al. 2018	

3.1. Le prétraitement des séquences

3.1.1. Le format FASTQ

Lors d'un séquençage *paired-end* (*cf.* 1.3), les *reads* en de sorties de séquenceur Illumina (Miseq ou NovaSeq) sont fournies en deux fichiers R1 et R2 (Figure 2.1) au format FASTQ pour chaque banque (*Index*). Les fichiers FASTQ sont des fichiers textes qui comprennent pour chaque séquence 4 lignes : l'identifiant de la séquence, la séquence, l'identifiant de la ligne des scores de qualité (nommé ci-après *quality score*) et les informations de qualité pour chaque base de la séquence (Figure 2.6 ; **Institute for Integrative Genome Biology UC Riverside 2012**).

```
@M00801:583:000000000-CW85N:1:1101:14248:1469 1:N:0:14
AGACGCCGCCTGTTTATCAAAAACATAGCCTTCAGCGAACAACAAGTATTGAAGGTGATGCCTGCCCAGTGACCCCCA
+
BBBBBBBBBBBBEFGGGGGGGGHGHCHFHHHHGHGHGGGFCFHHHHFFHFGHHG3BFGHGHHBDFFGGHHHHHHHC
```

@<instrument>:<run number>:<fowcell ID>:<lane>:<tile>:<x- pos>:<y-pos> <read>:<isfitered>:<control>:<index>

Position	Element	Description
1	@	Début de ligne pour l'identifiant des séquences
2	<instrument>	Identifiant de l'instrument
3	<run-number>	Numéro de lecture
4	<fowcell-ID>	Identifiant de la cellule de lecture
5	<lane>	Numéro de la ligne de lecture
6	<tile>	Numéro du carreau de lecture
7	<x-pos>	Position X du cluster
8	<y-pos>	Position Y du cluster
9	<read>	Numéro du read. 1 pour le read forward et 2 pour le read reverse
10	<is-filtered>	Est-ce que le read est filtré ? (oui : Y ou non :N)
11	<control>	0 quand aucun bit de contrôle est activé, sinon un nombre pair supérieur à 0
12	<index>	Séquence de l'index (optionnel)

Figure 2.6 : Exemple de séquence contenue dans un fichier FASTQ Illumina et détail des informations contenues dans l'identifiant de cette séquence.

3.1.2. Le démultiplexage

Dans le cadre de cette thèse, les échantillons ont été multiplexés après amplifications PCR impliquant une étape de démultiplexage lors du traitement. Le démultiplexage consiste à séparer les *reads* d'une banque (*Index*) en sous-fichiers correspondant aux *reads* d'un échantillon. Les courtes séquences nucléotidiques utilisées pour démultiplexer (ci-après nommé étiquette) les banques correspondent aux nucléotides des tags et de l'amorce (Figure 2.4). La séquence de l'index n'est pas utilisée car elle sert au prestataire du séquençage pour démultiplexer les *reads* Illumina

en différents fichiers correspondant aux banques. L'index et l'adaptateur permettant le séquençage ont été retirés par les prestataires.

Le démultiplexage est généralement réalisé avec Cutadapt **(Martin 2011)**, qui peut détecter les étiquettes (*adapters* dans Cutadapt) en début (5') et/ou en fin de séquence (3'), qu'ils soient aux extrémités de la séquence ou non. Une correspondance totale ou partielle aux étiquettes peut être utilisée pour démultiplexer les séquences. Bien que facultatif, le retrait des étiquettes peut être effectué avec Cutadapt lors du démultiplexage.

3.1.3. La qualité des séquences

Le *quality score* mesure la probabilité que la base appelée (étape 5 et 8 de la Figure 2.1) soit incorrecte. Lors du séquençage par synthèse, pour chaque base de la séquence lue, un *phred quality score* (*Q score*) est attribué par un algorithme *phred* similaire à celui développé pour le séquençage Sanger **(Ewing & Green 1998 ; Illumina Inc. 2011)**. Le *Q score* est attribué par l'équation suivante :

$\mathrm{Q} = -10\log_{10}(\mathrm{e})$ où *e* est la probabilité estimée que la base appelée est incorrecte.

Un *Q score* élevé indique une faible probabilité d'erreur, alors qu'un *Q score* faible indique qu'une portion de la séquence peut être inutilisable (Tableau 2.7). Les séquences de faible *Q score* peuvent augmenter la présence des faux-positifs et biaiser les conclusions de analyses de diversité.

Les *Q scores* Illumina sont calculés en deux étapes : (1) Pour chaque base, un nombre prédisant la qualité est déduit des différentes observations du *cluster*, tels que l'intensité du signal lumineux, le rapport signal/bruit et diverses mesures de la fiabilité de l'appel de la base ; (2) La valeur prédisant la qualité est calculée pour un nouvel appel de base et est comparée à un modèle de qualité (établi sur des données empiriques). Le *Q score* est ensuite enregistré dans des fichiers d'appels de base (.bcl) et est converti pour être restitué avec les séquences au format FASTQ.

Dans le cas du séquençage Illumina, les erreurs ne sont pas réparties aléatoirement au cours du séquençage **(Schirmer et al. 2015)**. Les régions avec des homopolymères ont des taux plus important d'insertions et de délétions **(Minoche et al. 2011)** et, de manière générale, le nombre d'erreurs augmente avec la longueur de la séquence (Figure 2.7 ; **Chaisson et al. 2009**). Pour augmenter la qualité des données, il est recommandé de réduire la longueur des *reads* aux extrémités 3' pour enlever les parties de séquence de moindre qualité. La qualité des *reads* peut être visualisée avec logiciel FastQC **(Andrews 2010)** pour déterminer la valeur de troncation optimale, qu'elle soit exprimée en longueur de séquences ou en *Q score* (Figure 2.6). La troncation

Tableau 2.7 : Relation entre le Quality Score et la précision de l'appel des bases

Quality Score (Q)	Probabilité d'une base incorrecte	Précision déduite des bases
10	1 sur 10	90 %
20	1 sur 100	99 %
30	1 sur 1000	99,9 %

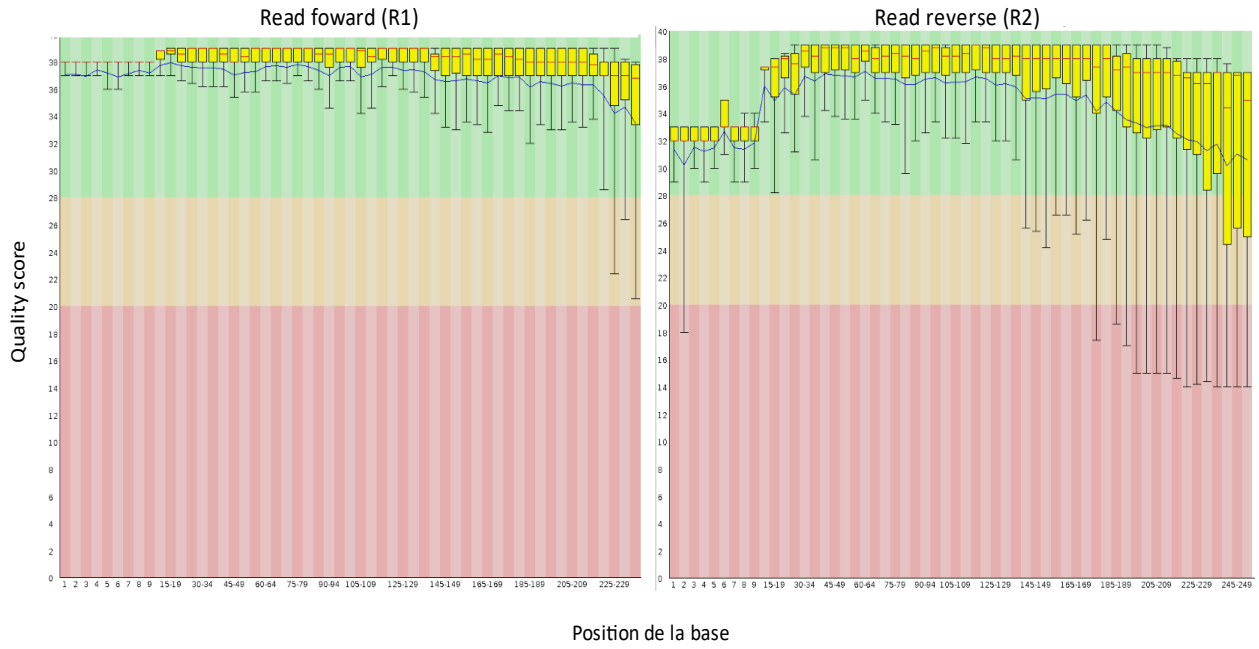


Figure 2.7 : Comparaison des sorties Fastqc pour les read Forward (R1) et Reverse (R2) séquencés par Miseq (2 x 250 bp ; reads du jeu de données test). La qualité (*score Phred*, en ordonnée) de chaque base (en abscisse) pour tous les *reads* du jeu de données. À chaque position du *read*, la qualité de tous les *reads* est représentée sous la forme d'un boxplot. La médiane est en rouge. Le code couleur vous indique les scores de très bonne qualité en vert (>28), bonne qualité en orange (entre 20 et 28) et mauvaise en rouge (<20). Généralement, la qualité baisse en fin de reads.

des séquences *via* les longueurs permet d'obtenir des *reads* de même longueur mais de qualité en fin de *read* non identique. La troncation basée sur le *Q score* coupe le *read* dès que le *Q score* est inférieur à la valeur donnée, cela permet d'avoir des *reads* de qualités semblables mais de longueurs différentes. Cela peut poser problème si les *read* sont en *paired-end* avec une troncation qui ne permet plus l'assemblage des *reads*, ou bien si le *read* comprend des homopolymères réduisant la qualité du *read* et qui est coupé trop court **(Taberlet et al. 2018)**. C'est pourquoi dans cette étude, les *reads* ont été coupés à une longueur donnée (*cf.* Chapitre 2.3.1.6). De nombreux logiciels existent pour tronquer les séquences **(Del Fabbro et al. 2013)**, les différences reposent sur le choix de la méthode de coupe et les paramètres proposés (troncation coté 5', 3', *Q score*).

3.1.4. La réduction du nombre de séquences

Le séquençage NGS permet d'obtenir des *reads* de longueurs connus (ex. 250 bp), toutefois des erreurs lors du séquençage conduisent à des *reads* plus courts que ceux désirés (Figure 2.78 : *reads* compris en 29 et 239 bp). Les *reads* courts doivent être supprimés. En fonction du jeu de données, un seuil de longueur minimum est établi (ex. 200 bp) et les *reads* de tailles inférieures sont supprimés. Il en est de même avec la qualité générale des séquences (Figure 2.7).

Le jeu de données est ensuite nettoyé des séquences erronées, avec le débruitage. Différents algorithmes et logiciels sont disponibles pour nettoyer les séquences, tels que DADA2 **(Callahan et al. 2016)**, UNOISE **(Edgar & Flyvbjerg 2015)**, MED **(Eren et al. 2015)**, **QIIME**, (UCLUST ; **Edgar 2010**) ou mothur (**Schloss et al. 2009 ; Schloss 2020** ; Tableau 2.6). **DADA** (*Divisive Amplicon Denoising Algorithm*) a été développé pour gérer les erreurs de séquençage Illumina avant l'assemblage des *reads* (*a contrario* de UNOISE) et est applicable au métabarcoding, contrairement à UCLUST (Clustering and Classification Inference with U-Statistics) qui n'est pas adaptés pour la production d'OTU. DADA2 montre une meilleure précision et moins de séquences incorrectes que les résultats des algorithmes cités précédemment **(Callahan et al. 2016)**. DADA2 est à l'origine un package R (https://github.com/benjjneb/dada2) amélioré par l'intégration de DADA dans certains environnement, tel que QIIME2 **(Bolyen et al. 2019)**. Le package R de DADA2 permet de réaliser l'ensemble du traitement des amplicons (filtration, déréplication, chimère, identification et assemblage des *reads*).

Finalement les *reads* identiques strictement identiques sont regroupés en **ASV** (*Amplicon Sequence Variant* aussi appelé ESV pour *exact sequence variants*) pour réduire la répétition (déréplication), la taille des fichiers et la mémoire nécessaire pour le reste des traitements bio-informatiques. L'abondance de *reads* associés à chaque ASV est gardée.

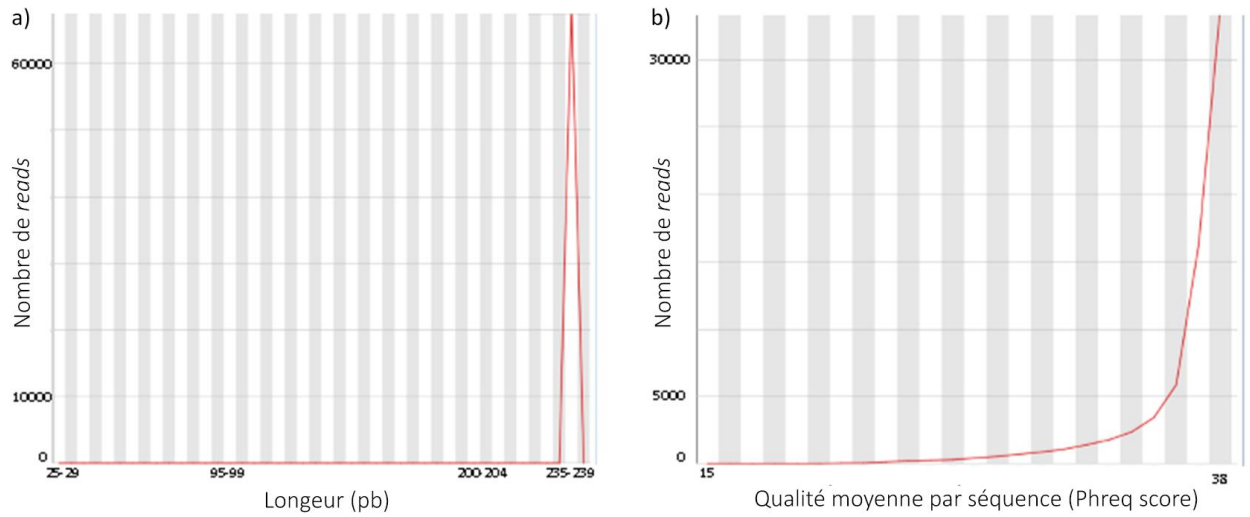


Figure 2.8 : Distribution des *reads* en fonction de leurs longueurs (a) et de leur qualité (b)

3.1.5. La fusion des séquences

Les *reads paired-end* peuvent être fusionnés pour reconstruire le barcode à analyser ou bien être analysés sans fusion, c'est-à-dire l'analyse du *read forward* est réalisée indépendamment de celle du *read reverse*. Quand les *reads paired-end* se chevauchent (Figure 2.2A), l'alignement puis la fusion des *reads forward* et *reverse* permettent une meilleure prédiction de la partie chevauchée **(Edgar & Flyvbjerg 2015)**. Plusieurs algorithmes d'alignement pour fusionner les *reads* sont disponibles, tel que : SHERA **(Rodrigue et al. 2010)**, DADA2 et FLASH **(Magoč & Salzberg 2011)**, qui maximisent la longueur du chevauchement et ne prennent pas en compte le *quality score* ; COPE **(Liu et al. 2012)**, prend en compte le *quality score* des bases non-concordantes et garde la base ayant le *quality score* le plus élevé ; et PANDASeq **(Masella et al. 2012)**, qui prend en compte les *quality scores* de toute les bases. Parmi ces algorithmes seul PANDASeq inclut un filtrage de la qualité des séquences en fusionnant les *reads* **(Taberlet et al. 2018)**. PANDASeq est également efficace pour fusionner les *reads* avec de faible chevauchement, mais le logiciel suppose que tous les *reads* peuvent être fusionnés **(Zhang et al. 2014 ; Edgar & Flyvbjerg 2015)**. PEAR **(Zhang et al. 2014)** et VSEARCH **(Rognes et al. 2016)** prennent en compte la longueur de chevauchement et les *quality scores*.

3.1.6. Le prétraitement des séquences mis en place dans ce projet

Dans la présente étude, les échantillons ont été démultiplexés avec cutadapt **(Martin 2011)** implémenté dans l'environnement QIIME2 (qiime cutadapt demux-paired) à partir de 6 pb du tag et des 6 premières paires de bases de l'amorce *forward* (Figure 2.11). L'algorithme de DADA2, implémenté dans QIIME2 (qiime dada2 denoise-*), a été retenu pour le débruitage et la fusion des

reads en raison de sa capacité à prendre en compte les erreurs de séquençage Illumina avant l'assemblage des reads. DADA2 appréhende différemment les *reads forward* et *reverse*, les *reads reverse* ayant généralement de plus faible qualité **(Callahan et al. 2016)**. Pour le marqueur 18S, qiime dada2 denoise-single a été utilisé car les *reads* R1 et R2 étaient non chevauchant (amplicons d'environ 550 bp), ainsi seul les *reads* R1 (de meilleur qualité) ont été utilisés. Pour le COI, les *reads* R1 et R2 étaient chevauchants (amplicons d'environ 310 bp), ainsi en sortie de qiime dada2 denoise-paired, les *reads forward* et *reverse* ont été fusionnées. Les *reads* résultants de qiime dada2 denoise-* strictement identiques (similarité de 100% et longueur de *reads* égale) ont été regroupés en ASV. En parallèle, l'abondance de *reads* de chaque ASV est stockée dans un tableau. Les ASV en sortie de DADA2 possèdent encore une (18S) ou deux (COI) amorces PCR qui sont supprimées avec cutadapt. Pour le 18S, le reste de l'amorce *forward* a été retiré. Pour le COI, les R1 et R2 ont été fusionnés à l'étape précédente, ainsi une partie de l'amorce *forward* et l'entièreté de l'amorce *reverse* étaient présentes. Seuls les ASV ayant les deux amorces à leur extrémité et ayant une longueur comprise en 300 et 320 pb ont été gardés.

3.2. Le traitement des séquences

3.2.1. Regroupement des séquences en OTU

Une fois le jeu de données nettoyé, les séquences sont regroupées en *cluster* pour former les OTU. En fonction du marqueur moléculaire employé, le seuil de regroupement varie. Généralement pour le COI, les OTU sont construit à un seuil de similarité de 97 % (Tableau 1.2 ; **Brown et al. 2015 ; Xiong & Zhan 2018**). En fonction des pipelines, les méthodes de création des OTU diffèrent que ce soit en termes de (1) méthodes de regroupement, (2) du choix de la séquence référente, ou bien (3) de la méthode d'alignement des séquences et des seuils de similarité.

En ce qui concerne les méthodes de regroupement en OTU, VSEARCH implémenté dans QIIME2 propose trois méthodes : (1) *de novo clustering*, les *reads* sont regroupés entre eux, à savoir chaque *read* est comparé aux autres *reads* du jeu de données ; (2) à références fermées, où les *reads* sont regroupés par rapport à une base de référence. Les *reads* qui ne correspondent pas à un *cluster* de la base de référence sont exclus des analyses futures ; (3) à références ouvertes, les *reads* sont regroupés par rapport à une base de référence et les *reads* qui n'ont pas été regroupés sont par la suite regroupés avec la méthode *de novo clustering* (**Rideout et al. 2014 ;** Tableau 2.9).

Pour ce qui est des différences dans le choix de la séquence représentative de l'OTU (centroïde), UNOISE2 **(Edgar 2016b)** détermine la séquence centroïde de l'OTU comme la séquence

du *read* le plus abondant du *cluster*, alors que VSEARCH la définit comme la première séquence qui n'appartient pas aux OTU précédemment créées **(Rognes et al. 2016)**. Le centroïde est défini comme étant la séquence correcte et les autres séquences comprises dans l'OTU étant des versions erronées du centroïde avec une ou plusieurs erreurs de séquençage **(Edgar 2016b)**. Contrairement à UNOISE2, DADA2 prend en compte les *quality scores* pour définir le centroïde.

Pour finir, concernant les différentes méthodes d'alignement pour comparer les *reads*, trois démarches sont couramment utilisées : le regroupement hiérarchique (mothur), le regroupement heuristique (**BLAST**, VSEARCH, UCLUST ; **Altschul et al. 1990 ; Rognes et al. 2016**) et le regroupement basé sur le jeu de données (CROP ; **Hao et al. 2011**). Pour les algorithmes de regroupements hiérarchique et heuristique, les seuils de dissimilarités sont définis par l'utilisateur (ex. 97 %, seuil dépendant du jeu de donnée ; **Edgar 2018**) alors que CROP regroupe les *reads* en utilisant l'organisation naturel du jeu de données **(Hao et al. 2011 ; Chen et al. 2013)**. Chen et al. (2013) montrent que l'estimation du nombre d'OTU dépend de l'algorithme employé (Tableau 2.8) et peut être aussi bien surestimé que sous-estimé. CROP semble mieux adapté aux jeux de données peu complexes avec des *reads* courts alors que les méthodes hiérarchiques ont de meilleures estimations du nombre d'OTU pour les jeux de données complexes composés de *reads* longs **(Chen et al. 2013)**.

Tableau 2.8 : Nombre d'OTU obtenus en fonction du seuil de regroupement et du logiciel utilisé (adapté de Chen et al. 2013)

Logiciel	**Jeu de données 1**			**Jeu de données 2**		
Seuil de similarité	OTU à 2 %	OTU à 3 %	OTU à 4 %	OTU à 2 %	OTU à 3 %	OTU à 4 %
Nombre OTU attendu		43			15	
Mothur	1882	720	369	63	41	20
UCLUST	2177	1883	597	80	75	51
CROP	339	133	62	21	15	15

3.2.2. Nettoyage et réduction du nombre d'OTU

Une fois les OTU obtenus, deux nouvelles étapes de nettoyage sont effectuées : le retrait des singletons et le retrait des chimères. Le retrait des singletons, les OTU représentés par un seul *read*, permet d'améliorer l'estimation du nombre d'OTU réels en retirant les séquences potentiellement erronées **(Edgar 2016a)**.

Les chimères sont des séquences formées à partir de deux ou plusieurs séquences originales, appelées séquences parents. Des amplicons chimériques peuvent être formés lors des étapes PCR précédent le séquençage, en particulier lorsque les séquences amplifiées sont proches **(Smyth et al. 2010)**. Le mécanisme le plus courant est l'extension incomplète d'une séquence (parent 1) qui est utilisée comme amorce pour la réplication d'un brin similaire mais différent (parent 2) au cycle

suivant (Figure 2.9 ; **Odelberg et al. 1995**). Pour le séquençage du 18S, seulement une petite fraction des *reads* est chimérique (1 à 5 %), mais cette fraction est beaucoup plus importante lorsque les *reads* sont regroupés en OTU (*cf.* Paradoxe de Tolstoy ; **Edgar 2016a**). Les chimères étant similaires à leurs parents, cela pose problème lors des analyses pour les différencier des réelles séquences **(Edgar 2016a)**. Certaines chimères sont créées lors du pré-traitement bio-informatique, lors de la fusion des *reads forward* et *reverse*. L'algorithme de UCHIME2 permet la détection de chimère *de novo* (sans base de référence de chimères) permettant une meilleure détection. UCHIME2 divise les séquences en plusieurs tronçons qui serviront de base de référence. Par la suite, UCHIME2 aligne les tronçons aux séquences du jeu de données, attribue un score d'alignement et identifie les deux séquences parentes. À noter qu'une séquence peut être considérée parente seulement si elle est deux fois plus abondante que la séquence fille **(Edgar et al. 2011 ; Edgar 2016a)**. Le retrait des singletons permet également de diminuer la présence des chimères.

Une troisième étape de nettoyage peut être implémentée pour les OTU ciblant des régions codantes, tel que le COI. La traduction en acides aminés des séquences codantes permet de mettre en évidence des erreurs dans les séquençages ou des séquences n'appartenant pas aux barcodes ciblés. Leray et Knowlton (2015) réconisent l'utilisation de **MACSE** (*Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons* ; **Ranwez et al. 2011)**, avec l'option « enrichAlignment », pour aligner les OTU à un jeu de référence et détecter les décalages du cadre de lecture (**ORF**, *Open Reading Frame*) due à l'insertion ou délétion de nucléotides, ainsi que la présence de codon stop **(Leray & Knowlton 2015)**. Cependant le code de traduction à utiliser dépend de l'origine taxonomique des séquences à traduire (ex. code mitochondrial des invertébrées, des ascidies, des plathelminthes, etc ; **Lavrov & Pett 2016)**.

Tableau 2.9 : Différentes méthodes de regroupement en OTU implémenté dans VSEARCH

	De novo clustering	**À références fermées**	**À références ouvertes**
Adapté à :	Analyse avec absence de base de référence	Analyse d'amplicons non-chevauchants	Analyse d'amplicons chevauchants avec une base de référence
Non adapté à :	Comparaison d'amplicons non-chevauchants Très grand jeu de données (Hiseq 2000)	Analyse avec absence de base de référence	Comparaison d'amplicons non-chevauchants Analyse avec absence de base de référence
Avantages :	Regroupe tous les *reads*	Rapide, utile pour les gros jeux de données Meilleurs identification et phylogénie	Regroupe tous les *reads* Rapide, fonctionne en partie en parallèle
Inconvénients :	Lent, ne fonctionne pas en parallèle	Ne détecte pas de nouveaux OTU	Lent, certaines étapes s'exécutent en série

Paradoxe de Tolstoy

Pour des reads de 250bp avec une qualité maximale de Q40, soit environ 1 erreur toutes les 1 000 bases, 4 reads de 250bp donneront 3 reads corrects (75%) et 1 read erroné (25%). Ainsi, en séquençant 100 fois la même séquence, 75 reads seront corrects et 25 reads erronés. La répartition des erreurs étant aléatoire, les 25 reads erronés seront probablement tous différents. Il y aura au final 26 séquences uniques, dont 1 seule correcte et 25 erronées, soit 4% des séquences correctes. Toutefois la séquence correcte sera présente en 75 exemplaires alors que les erronées en un seul exemplaire seront supprimés en retirant les singletons (Edgar 2016a).

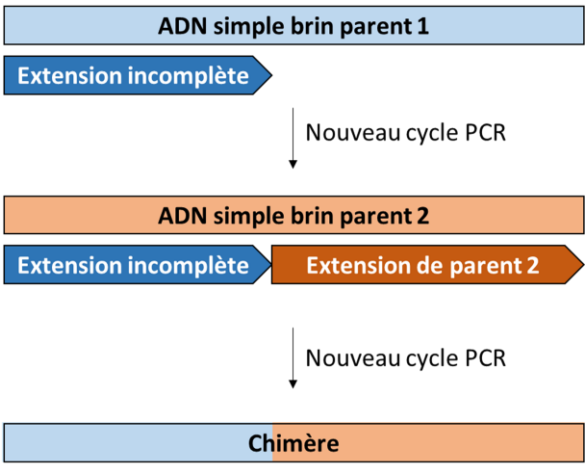


Figure 2.9 : Processus PCR conduisant à la création de chimère

3.2.3. Le traitement des séquences mis en place dans ce projet

Certains auteurs plébiscitent l'utilisation directe des ASV pour l'analyse des communautés **(Joos et al. 2020 ; González et al. 2020 ; Inglis et al. 2022)**. En effet, les ASV permettent de saisir une plus grande diversité que les OTU (qui reposent sur un regroupement des ASV avec un seuil de similarité défini par l'utilisateur). Leur utilisation augmente la précision de l'attribution taxonomique et la détection les taxons rares **(Amir et al. 2017 ; Callahan et al. 2017)**. Cependant, la technologie de séquençage employée ne permet pas de résoudre avec précision les séquences exactes et crées artificiellement des ASV. Le regroupement des ASV en OTU permet de limiter ce biais et reste efficace dans l'analyses de communautés à des niveaux taxonomiques plus élevés **(Edgar 2016b)**. Ainsi afin de réduire le nombre de séquences (ASV) artificiellement augmenté par le séquençage sans tomber dans les travers d'un regroupement trop large, les ASV ont été regroupés en OTU à 99% à l'aide de l'algorithme vsearch implémenté dans QIIME2 (qiime vsearch cluster-features-de-novo ; Figure 2.11). Deux types de filtration ont été réalisés, une basé sur l'abondance des reads au sein des OTU et la seconde sur le retrait des chimères. La filtration sur l'abondance des reads a consisté à retirer les singletons et à filtrer les OTU à partir de l'abondance des reads des OTU présents dans les échantillons témoins. Les OTU ayant un nombre de reads inférieurs au seuil déterminé sont considérés comme probables contaminants et sont retirés des jeux de données. Les chimères a été retiré à l'aide de la fonction qiime vsearch uchime-denovo implémenté dans QIIME2 et des fonctions lulu {lulu} **(Frøslev et al. 2017)** et isContaminant {decontam} **(Callahan 2021)** dans R (Figure 2.11).

3.3. L'assignement des OTU

3.3.1. Les méthodes d'assignement

La méthode d'assignement la plus simple est appelée « le plus proche voisin ou *nearest neighbor* ». Cette méthode compare une séquence à une base de référence et assigne à l'identification de la séquence la plus proche. Cette méthode est efficace lorsque le jeu de donnée est complet et quand le marqueur utilisé permet de différencier tous les taxons. Toutefois, les bases de données sont rarement complètes, et de nombreuses séquences ne peuvent être assignées à un taxon. Par ailleurs, si le marqueur utilisé ne permet pas de distinguer tous les taxons (plusieurs taxons partagent la même séquence), la séquence peut être assignée aléatoirement à un des taxons partageant la même séquence. Pour contourner ces limites, il est possible d'utiliser l'algorithme du « plus proche ancêtre commun ou *Last Commun Ancestor* (**LCA**) ». Cette méthode sélectionne un ensemble de séquences similaires et la séquence d'entrée est assignée à l'entité taxonomique partagée par l'ensemble des séquences. Le programme *classifer* de **RDP** **(Wang et al. 2007)** est également communément utilisé. Ce programme utilise une approche probabiliste bayésienne pour assigner les séquences. Pour finir, il existe les méthodes de placement phylogénétique, tel que EPA **(Kozlov et al. 2016)**, qui place la séquence à assigner au sein d'un arbre phylogénétique.

3.3.1.1. Méthode du plus proche voisin

L'assignement des OTU à une base de référence est réalisé à partir de mesures de similarité entre deux séquences, qui sont effectuées de façon locale ou globale **(Altschul et al. 1990)**. Les algorithmes d'alignements locaux, tels que BLAST, recherchent les tronçons de séquences conservés, une seule comparaison peut produire plusieurs alignements de tronçons distincts et les tronçons non conservés ne sont pas pris en compte dans la mesure de la similarité **(Altschul et al. 1990)**. Les algorithmes d'alignement globaux, tel que VSEARCH, optimisent l'alignement sur l'ensemble de la séquence, ce qui peut conduire à aligner deux tronçons de séquences peu similaires **(Needleman & Wunsch 1970)**.

BLAST (*Basic Local Alignment Search Tool*) est un programme qui recherche la similarité entre les séquences **(Altschul et al. 1990)** et peut être utilisé pour rechercher des séquences correspondantes dans une base de données. BLAST utilise une matrice de similarité afin de calculer pour chaque alignement un score d'alignement et donner une évaluation statistique de la pertinence de l'alignement. BLAST repose sur trois étapes principales : (1) la décomposition de la séquence d'entrée en portion chevauchantes de longueur *k* (fixée par l'utilisateur) appelés *k*-mers.

Pour chacun d'entre eux, BLAST cherche tous les autres *k*-mers possibles qui donneraient un score d'alignement supérieur à une valeur seuil (fixée elle-aussi par l'utilisateur). (2) BLAST analyse chaque séquence de la banque pour trouver des *k*-mers similaires et les étendre en amont et en aval, de sorte à augmenter le score d'alignement. (3) La recherche retourne plusieurs séquences, BLAST analyse ensuite la pertinence des alignements, pour déterminer si l'alignement obtenu est dû au hasard ou s'il correspond véritablement à une séquence homologue **(Altschul et al. 1990)**.

BLAST fournit les valeurs suivantes :

- *E-value* : reflète la probabilité que la séquence retournée soit due au hasard, plus la *e-value* est petite, plus cette probabilité est faible ;
- *Max score* : le score d'alignement le plus élevée ;
- *Total score* : la somme des scores d'alignement pour une même séquence.

Il existe différentes variations de BLAST qui permettent de comparer des séquences nucléiques ou protéiques à des bases de référence nucléiques ou protéiques. BLASTX compare des entrées protéiques à une base de données protéiques, alors que BLASTN est utilisé pour comparer des entrées nucléiques à une base de données nucléiques. BLASTN utilise de petits *k*-mers et laisse les *gaps* par défaut **(Madden 2013)**. MEGABLAST sert également à comparer des séquences nucléiques mais est optimisé pour des séquences proches (ex. recherche des erreurs de séquençage), en utilisant de longs *k*-mers ainsi que des pénalités pour les *gaps*. DISCONTIGUOUS MEGABLAST quant à lui permet des *gaps* dans les étapes initiales de BLAST **(Camacho et al. 2009)**.

VSEARCH conduit des alignements globaux et se décompose en deux étapes principales. La première étape est similaire à BLAST en cherchant les *k*-mers partagés entre les séquences. La deuxième étape effectue un alignement global de la séquence avec la séquence ayant maximum de *k*-mers en commun **(Rognes et al. 2016)**.

3.3.1.2. Méthode de l'ancêtre commun

L'assignement des OTU est réalisé à partir de mesures de similarité entre la séquence à identifier et un groupe de séquences. En premier lieu, l'algorithme détermine un groupe de séquences similaires à la séquence à assigner de la même manière que BLAST **(Huson et al. 2007)**. La taille du groupe peut être définie par l'utilisateur, soit par un entier fixe pour chaque séquence (ex. 10, fournit des groupes comprenant les 10 séquences les plus proches de la séquence à assigner), soit par un seuil de similarité (le nombre de séquences du groupe est alors variable), soit par les deux (ex. garde les 10 séquences les plus proches à plus de 97 % de similarité ; **Taberlet et**

al. 2018). Ensuite, l'algorithme recherche le nœud taxonomique qui regroupe l'ensemble de ces séquences et assigne ce rang taxonomique à la séquence. Si la séquence correspond à deux taxons différents A et B, et que A a une identification au rang taxonomique supérieur de B, seule la correspondance à B est gardée. Le logiciel utilisant la méthode de *Last Common Ancestor* (LCA) le plus repandu est MEGAN **(Huson et al. 2007)**,amélioré avec MEGAN4 **(Huson et al. 2011)** puis MEGAN6 **(Huson et al. 2016)**. Dans cette dernière version, une version pondérée de LCA peut être utilisée (*weighted* LCA). L'algorithme pondéré de LCA, dans un premier temps, attribue à chaque séquence référence « S », un poids qui correspond au nombre de *reads* à assigner « R » qui ne s'alignent que sur « S » (ou séquence ayant le même assignement taxonomique). Dans un second temps, chaque *read* « R » est placé au nœud taxonomique qui regroupe les séquences ayant un poids supérieur à 75 % (par défaut) du poids total de toutes les références sur lesquelles « R » a un alignement significatif. Cette pondération permet d'améliorer la précision de l'assignement mais nécessite un temps d'exécution plus long **(Huson et al. 2016)**.

3.3.1.3. Méthode probabiliste bayésienne

Les OTU peuvent être assignés par une approche bayésienne avec le programme *classifer* de RDP (Ribosomal Database Project II) **(Wang et al. 2007)**. La similarité des séquences est estimée à partir du nombre de *k*-mers partagés par les séquences. Cette estimation a l'avantage d'être rapide car elle ne nécessite pas d'alignement mais reste peu précise **(Taberlet et al. 2018)**. RDP utilise un jeu de données de référence pour entraîner le modèle de classification, puis détermine la probabilité d'assignement de l'OTU à chaque rang taxonomique **(Wang et al. 2007)**.

3.3.1.4. Méthode du placement phylogénétique

La méthode du placement phylogénétique est une approche alternative aux précédentes. Le principal argument en faveur de cette méthode est que la position d'une séquence au sein d'un arbre phylogénétique fournit des informations plus précises qu'un assignement à un simple nom de taxon. La première étape consiste à placer la séquence dans une région de l'arbre phylogénétique en se basant sur les distances entre les séquences. Ensuite, la séquence est placée plus précisément dans l'arbre en utilisant les algorithmes de similarité (*maximum likelyhood*). Cependant pour utiliser cette méthode, les séquences doivent contenir un signal phylogénétique, à savoir être assez longues et avoir un taux d'évolution modérée (ex. non adaptée au COI), ce qui est rarement le cas des séquences utilisées à des fins de *barcoding* et métabarcoding **(Taberlet et al. 2018)**.

Pour résumer l'assignement taxonomique, il n'existe pas de solution universelle optimale pour résoudre les problèmes de classification. Certains compromis sont inévitables lors du choix de stratégie d'analyse du jeu de données.

3.3.2. Les différentes bases de référence

Le *Barcode of Life DataSystems* (BOLD) et GenBank sont les principaux espaces de stockage des séquences barcodes ADN en libre accès. Idéalement, toutes les séquences présentes dans ces bases de données devraient être reliées à un spécimen identifié par un taxonomiste. Étant donnée leur nature collaborative, la présence d'erreurs est inévitable **(Meiklejohn et al. 2019)**. Les séquences incorrectes peuvent provenir d'erreurs telles qu'une mauvaise identification du matériel génétique ou bien du séquençage d'une espèce non-ciblée (ex. les endoparasites chez les insectes ou des porifères recouvrant des cnidaires ; **Valentini et al. 2009)**. Plusieurs études ont évalué la précision de ces bases de données mais seulement à certaines échelles taxonomiques. Ces études ont aussi mis en avant que 80 % des séquences présentes dans GenBank ont des références insuffisantes pour les relier à un spécimen ou proviennent d'échantillons environnementaux **(Meiklejohn et al. 2019)**.

3.3.2.1. GenBank

GenBank (www.ncib.nlm.nih.gov) est la base de données qui regroupe la majorité des séquences nucléiques publiquement disponibles. GenBank a été créée par le **NCBI** (*National Center for Biotechnology Information*) et regroupe des séquences produites par des laboratoires et des centres de séquençages du monde entier **(Benson et al. 2013)**. Le nombre de séquences déposées n'y cesse d'augmenter. Lors des nouvelles soumissions, GenBank réalise un contrôle basique de la qualité des séquences, tels que la bonne traduction des régions codantes et la taxonomie à celle de la base. Cependant l'identification du spécimen dont est issu la séquence ne peut pas être vérifiée.

3.3.2.2. BOLD

La base de données BOLD (www.Boldsystems.org) a été créée par le *Consortium for the Barcode of Life* (CBOL) et est en accès libre depuis 2003 à tout chercheur s'intéressant au *barcoding* **(Ratnasingham & Hebert 2007)**. BOLD ne répertorie que des séquences de COI et d'autres marqueurs de barcoding (>13 000 000) pour environ 344 000 espèces (début 2023). Dans BOLD

pour qu'une séquence soit considérée comme un barcode, l'auteur de la séquence doit fournir : le nom d'espèce, les informations propres à l'échantillon, le numéro de collection, etc., et des informations sur la qualité ce qui permet de réduire les erreurs et facilite la vérification des séquences. Par la suite les administrateurs de BOLD effectuent une vérification des données (ex. que la séquence n'appartient pas à un contaminant d'un autre taxon, est une copie fonctionnelle). Contrairement à GenBank qui stocke seulement les séquences, BOLD garde les données des collections et les photos des échantillons. Toutes les séquences enregistrées dans BOLD sont transférées automatiquement dans GenBank lorsque les auteurs le permettent. Toutefois 37 à 55 % des séquences enregistrées dans BOLD ne seraient pas présentes dans GenBank **(Porter & Hajibabaei 2018)**. Pour maximiser les assignements, il est conseillé d'utiliser les bases de données à tour de rôle ou de les fusionner **(Macher et al. 2017)**.

3.3.2.3. MIDORI et les bases de référence locales

Face aux erreurs présentes dans les bases de données en ligne, Machida et ses collaborateurs ont réalisé la base de référence MIDORI, en filtrant et vérifiant l'ensemble des gènes mitochondriaux des métazoaires de GenBank **(Machida et al. 2017)**. MIDORI est disponible en deux versions : MIDORI-UNIQUE qui contient tous les haplotypes uniques associés à chaque espèce et MIDORI-LONGEST qui contient une seule séquence, la plus longue, pour chaque espèce. Une nouvelle version de MIDORI, MIDORI2 est disponible depuis mars 2022 et inclus maintenant l'ensemble des eucaryotes **(Leray et al. 2022)**. La base de référence MIDORI2 est régulièrement mise à jour et les différentes versions sont disponibles au téléchargement au lien suivant : http://www.reference-midori.info/. A noter qu'il est maintenant possible de les télécharger sous différents formats, notamment QIIME, ce qui facilite leur intégration dans les pipelines.

La complexité et le temps nécessaire à nettoyer et filtrer les bases de référence publiques ont conduit de nombreux auteurs à développer leur propre base de référence. Ces bases locales reposent fréquemment sur une base de données publique taxon-spécifique filtrée et nettoyée à laquelle ont été rajoutées des séquences barcode effectuées par les auteurs. Ces bases locales permettent généralement un assignement plus rapide dû à leur taille réduite et de meilleure qualité. Cependant, la mise à disposition de ces bases de données locales par les auteurs reste rare et complexifie la comparaison des études.

3.3.2.4. SILVA

La base de données SILVA (www.arb-silva.de) permet d'assigner les séquences 16S de procaryotes et les séquences 18S des eucaryotes **(Yilmaz et al. 2014)**. La base de données est mise à jour une fois par an, la dernière version date du décembre 2019 et contient 1 983 534 séquences de bactéries, 69 198 séquences d'archées et 172 540 séquences d'eucaryotes.

3.3.3. La stratégie d'assignement mise en place dans ce projet

Les deux méthodes d'assignement au plus proche voisin blast et VSEARCH sont implémentés dans QIIME2 et ont été testées dans cette étude. Pour effectuer un assignement plus rapide, blast assigne à la première séquence de la base de référence qui répond au critère de sélection (ex : seuil de similarité) contrairement à VSEARCH qui sélectionne la séquence ayant la meilleure similarité. La stratégie d'assignement implémentée dans VSEARCH nécessite la comparaison de l'ensemble de la base de données de référence pour chaque séquence à assigner et ainsi implique un temps d'exécution plus important. En effet, blast assigne 2 500 OTU en 5 min environ contre environ 2,5 jours pour VSEARCH. Les deux méthodes montrent des résultats identiques, ainsi pour des raisons computationnelles, blast sous le plugin «*qiime feature-classifier classify-consensus-blast*» a été sélectionnée pour les analyses (Figure 2.10 ; Figure 2.11).

Pour déterminer la meilleure stratégie d'assignement, blast (sous QIIME2), LCA (blast + MEGAN) et RDP (http : //www.reference-midori.info/server.php) ont été testés sur un sous jeu de données tests (3 ARMS collecté en 2015 [*cf.* Tableau 1.2] ; COI ; seuil de similarité de 97% ; Figure 2.10). La méthode LCA permet d'assigner un plus grand nombre de séquences que ce soit à l'embranchement (blast N = 97 ; LCA N = 1 345 ; RDP = 186) ou à l'espèce (blast N = 53 ; LCA N = 178 ; RDP=64). En ce qui concerne l'assignement à l'espèce, 27 OTU ont été assignés par les trois méthodes dont 23 sont en accords jusqu'à l'espèce. Par ailleurs, certains OTU sont assignés uniquement par une seule méthode (Blast N=17 ; LCA N=134 ; RDP=17 ; Figure 2.10B).

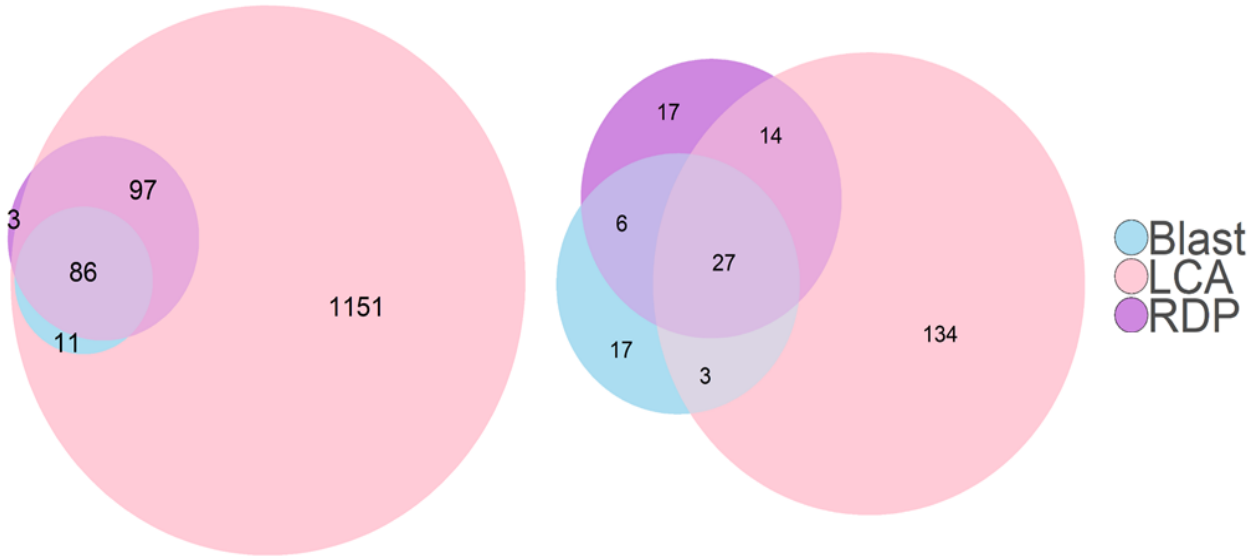


Figure 2.10 : Comparaison des OTU assignés A) au phylum et B) à l'espèce en fonction de la méthode d'assignement utilisée

A l'heure actuelle, aucune méthode d'assignement ne semble optimale en termes de qualité et nombre d'assignement si elle intervient seule. Les stratégies d'assignement au plus proche voisin sont efficaces lorsque le jeu de donnée est complet et quand le barcode utilisé permet de différencier tous les taxons. Toutefois, les bases de données sont rarement complètes, et de nombreuses séquences ne peuvent être assignées à un taxon. Par ailleurs, si la séquence utilisée ne permet pas de distinguer tous les taxons, la séquence sera assignée aléatoirement à un des taxons partageant la même séquence. Pour contourner ces limites, il est possible d'utiliser la méthode du plus proche ancêtre commun.

Pour augmenter la probabilité d'obtenir des assignements corrects, une combinaison des différentes méthodes d'assignement et des différentes bases de données a été mise en place. Pour le jeu de données 18S, l'identification a été attribuée par rapport aux bases de données locales et SILVA 138.1 (*cf.* Chapitre 3 ; **Yilmaz et al. 2014**) en utilisant des étapes hiérarchiques suivantes : (1) blast contre la base de données locale à 99% de similarité (qiime feature-classifier classify-consensus-blast) ; puis (2) LCA avec un seuil de 99% contre SILVA et enfin (3) LCA avec un seuil de 97% contre les bases de données locales et SILVA fusionnées (Figure 2.11). Pour le COI, l'assignement a été réalisé par rapport aux bases de données locales et MIDORI2 (sur GenBank 250 ; *cf.* Chapitre 3 ; **Leray et al. 2022**) en utilisant des étapes hiérarchiques : (1) blast contre la base de données locale à 99% de similarité (qiime feature-classifier classify-consensus-blast) ; puis (2) assignée avec la méthode LCA avec un seuil de 99% contre MIDORI2 ; (3) LCA avec un seuil de 97%

contre les bases de données locale et MIDORI fusionnées et enfin (4) LCA avec un seuil de 95% contre les bases de données fusionnées (Figure 2.11).

Le nombre de séquences dans les jeux de données finaux était trop important pour être assignées en LCA avec MEGAN6. Ainsi, après sélection de séquences (au seuil défini dans blast (blastn)), l'identification a été attribuée avec le programme BASTA (BAsic Sequence Taxonomy Annotation ; **Kahlke & Ralph 2019**).

Notes :
Pour des raisons de synthèse, l'ensemble des tests réalisés pour faire correspondre les différentes bases de données aux différents formats des logiciels ne sera pas developpé ici. Depuis le début de ce projet et le développement des scripts de reformatage des bases de données, différentes avancées ont été réalisées dans le domaine et sont maintenant disponibles pour faciliter leur intégration :
Le plugin RESCRIPt (REference Sequence annotation and CuRatIon Pipeline) à été développé en 2021 et permet de formater les différentes bases de données (SILVA, Genbank, BOLD ; **li et al. 2021**)
MIDORI2 est fourni au format QIIME **(Leray et al. 2022)**

3.4. Synthèse du pipeline bio-informatique utilisé dans ce projet

L'ensemble du pipeline bio-informatique mis en place lors de ce projet doctoral est synthétisé sur la Figure 2.11. Les scripts produits seront mis à disposition sur Github (https://github.com/Mcouedel) lors de la publication des articles.

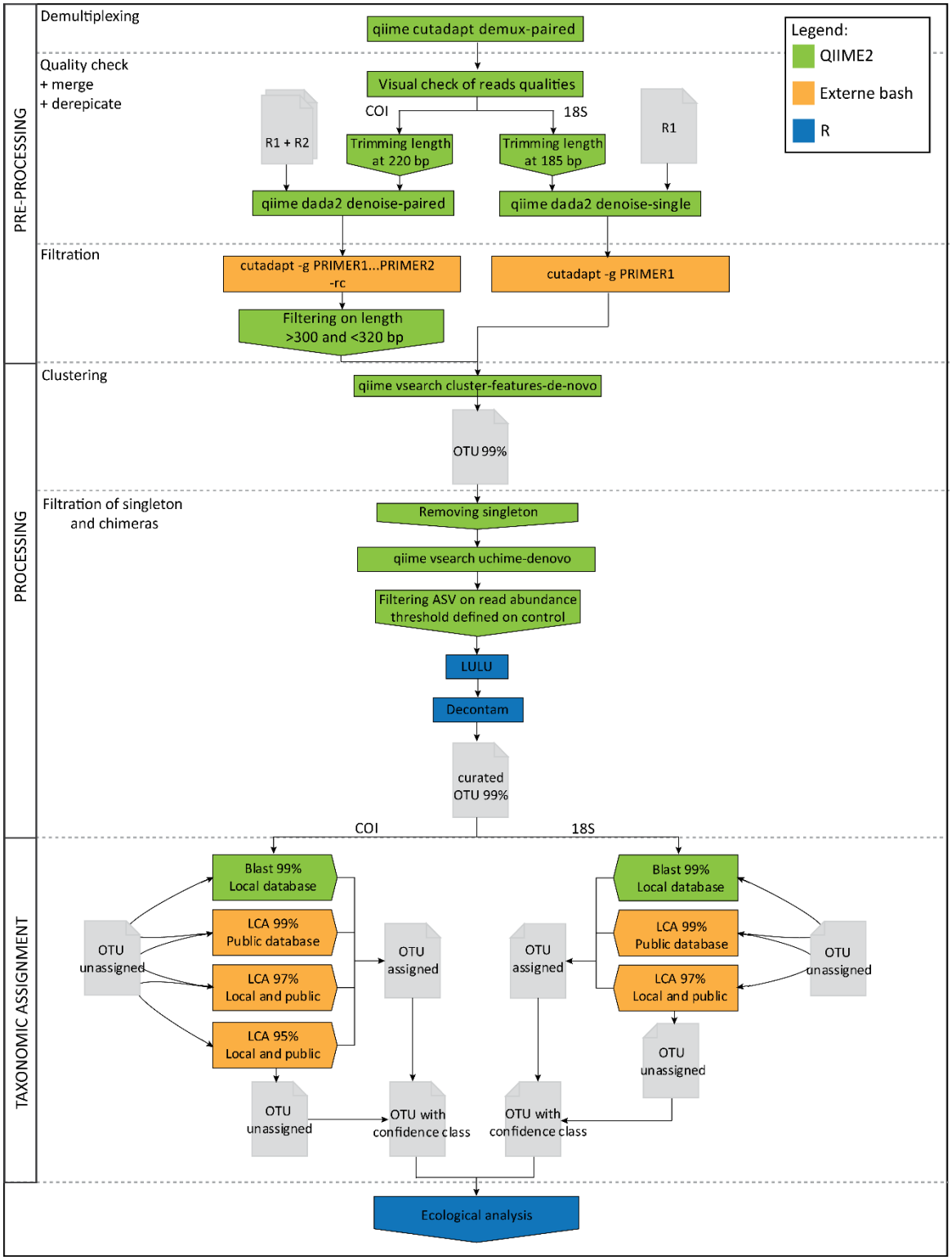


Figure 2.11 : Synthèse du processus bio-informatique mis en place lors de cette étude pour passer des reads en sortie de séquenceurs a des ASV assignées taxonomiquement pour les analyses écologiques

4. Références du chapitre 2

Ahrens D., Monaghan MT., Vogler AP. (2007) DNA-based taxonomy for associating adults and larvae in multi-species assemblages of chafers (Coleoptera: Scarabaeidae). Molecular Phylogenetics and evolution 44:436–449.

Albaina A., Aguirre M., Abad D., Santos M., Estonba A. (2016) 18S rRNA V9 metabarcoding for diet characterization: a critical evaluation with two sympatric zooplanktivorous fish species. Ecology and Evolution 6:1809–1824. DOI: 10.1002/ece3.1986

Altschul SF., Gish W., Miller W., Myers EW., Lipman DJ. (1990) Basic local alignment search tool. Journal of Molecular Biology 215:403–410. DOI: 10.1016/S0022-2836(05)80360-2

Amir A., McDonald D., Navas-Molina JA., Kopylova E., Morton JT., Zech Xu Z., Kightley EP., Thompson LR., Hyde ER., Gonzalez A., Knight R. (2017) Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. mSystems 2:e00191-16. DOI: 10.1128/mSystems.00191-16

Andersen JC., Mills NJ. (2012) DNA Extraction from Museum Specimens of Parasitic Hymenoptera. PLOS ONE 7:e45549. DOI: 10.1371/journal.pone.0045549

Andrews S. (2010) FastQC.

Austerlitz F., David O., Schaeffer B., Bleakley K., Olteanu M., Leblois R., Veuille M., Laredo C. (2009) DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. BMC bioinformatics 10:S10.

Avise JC., Lansman RA., Shade RO. (1979) The Use of Restriction Endonucleases to Measure Mitochondrial Dna Sequence Relatedness in Natural Populations. I. Population Structure and Evolution in the Genus Peromyscus. Genetics 92:279–295.

Bansal V., Boucher C. (2019) Sequencing Technologies and Analyses: Where Have We Been and Where Are We Going? iScience 18:37–41. DOI: 10.1016/j.isci.2019.06.035

Bensasson D., Zhang D-X., Hartl DL., Hewitt GM. (2001) Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends in Ecology & Evolution 16:314–321. DOI: 10.1016/S0169-5347(01)02151-6

Benson DA., Cavanaugh M., Clark K., Karsch-Mizrachi I., Lipman DJ., Ostell J., Sayers EW. (2013) GenBank. Nucleic Acids Research 41:D36–D42. DOI: 10.1093/nar/gks1195

Bhadury P., Austen M., Bilton D., Lambshead P., Rogers A., Smerdon G. (2006) Development and evaluation of a DNA-barcoding approach for the rapid identification of nematodes. Marine Ecology Progress Series 320:1–9. DOI: 10.3354/meps320001

Bohmann K., Evans A., Gilbert MTP., Carvalho GR., Creer S., Knapp M., Yu DW., de Bruyn M. (2014) Environmental DNA for wildlife biology and biodiversity monitoring. Trends in Ecology & Evolution 29:358–367. DOI: 10.1016/j.tree.2014.04.003

Bolyen E., Rideout JR., Dillon MR., Bokulich NA., Abnet C., Al-Ghalith GA., Alexander H., Alm EJ., Arumugam M., Asnicar F., Bai Y., Bisanz JE., Bittinger K., Brejnrod A., Brislawn CJ., Brown CT., Callahan BJ., Caraballo-Rodríguez AM., Chase J., Cope E., Da Silva R., Dorrestein PC., Douglas GM., Durall DM., Duvallet C., Edwardson CF., Ernst M., Estaki M., Fouquier J., Gauglitz JM., Gibson DL., Gonzalez A., Gorlick K., Guo J., Hillmann B., Holmes S., Holste H., Huttenhower C., Huttley G., Janssen S., Jarmusch AK., Jiang L., Kaehler B., Kang KB., Keefe CR., Keim P., Kelley ST., Knights D., Koester I., Kosciolek T., Kreps J., Langille MG., Lee J., Ley R., Liu Y-X., Loftfield E., Lozupone C., Maher M., Marotz C., Martin BD., McDonald D., McIver LJ., Melnik AV., Metcalf JL., Morgan SC., Morton J., Naimey AT., Navas-Molina JA., Nothias LF., Orchanian SB., Pearson T., Peoples SL., Petras D., Preuss ML., Pruesse E., Rasmussen LB., Rivers A., Robeson, II MS., Rosenthal P., Segata N., Shaffer M., Shiffer A., Sinha R., Song SJ., Spear JR., Swafford AD., Thompson LR., Torres PJ., Trinh P., Tripathi A., Turnbaugh PJ., Ul-Hasan S., van der Hooft JJ., Vargas F., Vázquez-Baeza Y., Vogtmann E., von Hippel M., Walters W., Wan Y., Wang M., Warren J., Weber KC., Williamson CH., Willis AD., Xu ZZ., Zaneveld JR.,

Zhang Y., Zhu Q., Knight R., Caporaso JG. (2018) QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. PeerJ Preprints. DOI: 10.7287/peerj.preprints.27295v2

Bolyen E., Rideout JR., Dillon MR., Bokulich NA., Abnet CC., Al-Ghalith GA., Alexander H., Alm EJ., Arumugam M., Asnicar F., Bai Y., Bisanz JE., Bittinger K., Brejnrod A., Brislawn CJ., Brown CT., Callahan BJ., Caraballo-Rodríguez AM., Chase J., Cope EK., Da Silva R., Diener C., Dorrestein PC., Douglas GM., Durall DM., Duvallet C., Edwardson CF., Ernst M., Estaki M., Fouquier J., Gauglitz JM., Gibbons SM., Gibson DL., Gonzalez A., Gorlick K., Guo J., Hillmann B., Holmes S., Holste H., Huttenhower C., Huttley GA., Janssen S., Jarmusch AK., Jiang L., Kaehler BD., Kang KB., Keefe CR., Keim P., Kelley ST., Knights D., Koester I., Kosciolek T., Kreps J., Langille MGI., Lee J., Ley R., Liu Y-X., Loftfield E., Lozupone C., Maher M., Marotz C., Martin BD., McDonald D., McIver LJ., Melnik AV., Metcalf JL., Morgan SC., Morton JT., Naimey AT., Navas-Molina JA., Nothias LF., Orchanian SB., Pearson T., Peoples SL., Petras D., Preuss ML., Pruesse E., Rasmussen LB., Rivers A., Robeson MS., Rosenthal P., Segata N., Shaffer M., Shiffer A., Sinha R., Song SJ., Spear JR., Swafford AD., Thompson LR., Torres PJ., Trinh P., Tripathi A., Turnbaugh PJ., Ul-Hasan S., van der Hooft JJJ., Vargas F., Vázquez-Baeza Y., Vogtmann E., von Hippel M., Walters W., Wan Y., Wang M., Warren J., Weber KC., Williamson CHD., Willis AD., Xu ZZ., Zaneveld JR., Zhang Y., Zhu Q., Knight R., Caporaso JG. (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology 37:852–857. DOI: 10.1038/s41587-019-0209-9

Boyce R., Chilana P., Rose TM. (2009) ICODEHOP: a new interactive program for designing COnsensus-DEgenerate Hybrid Oligonucleotide Primers from multiply aligned protein sequences. Nucleic Acids Research 37:W222–W228. DOI: 10.1093/nar/gkp379

Brown EA., Chain FJJ., Crease TJ., MacIsaac HJ., Cristescu ME. (2015) Divergence thresholds and divergent biodiversity estimates: can metabarcoding reliably describe zooplankton communities? Ecology and Evolution 5:2234–2251. DOI: 10.1002/ece3.1485

Brown WM., George M., Wilson AC. (1979) Rapid evolution of animal mitochondrial DNA. Proceedings of the National Academy of Sciences 76:1967–1971.

Buermans HPJ., den Dunnen JT. (2014) Next generation sequencing technology: Advances and applications. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease 1842:1932–1941. DOI: 10.1016/j.bbadis.2014.06.015

Buhay JE. (2009) "COI-like" Sequences Are Becoming Problematic in Molecular Systematic and DNA Barcoding Studies. Journal of Crustacean Biology 29:96–110. DOI: 10.1651/08-3020.1

Cadotte MW., Dinnage R., Tilman D. (2012) Phylogenetic diversity promotes ecosystem stability. Ecology 93:S223–S233. DOI: 10.1890/11-0426.1

Callahan B. (2021) Decontam.

Callahan BJ., McMurdie PJ., Holmes SP. (2017) Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME Journal 11:2639–2643. DOI: 10.1038/ismej.2017.119

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

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden TL. (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421. DOI: 10.1186/1471-2105-10-421

Capo E., Debroas D., Arnaud F., Perga M-E., Chardon C., Domaizon I. (2017) Tracking a century of changes in microbial eukaryotic diversity in lakes driven by nutrient enrichment and climate warming: Long-term dynamics of microbial eukaryotes. Environmental Microbiology 19:2873–2892. DOI: 10.1111/1462-2920.13815

Carew ME., Coleman RA., Hoffmann AA. (2018) Can non-destructive DNA extraction of bulk invertebrate samples be used for metabarcoding? PeerJ 6:e4980. DOI: 10.7717/peerj.4980

Carvalho S., Aylagas E., Villalobos R., Kattan Y., Berumen M., Pearman JK. (2019) Beyond the visual: using metabarcoding to characterize the hidden reef cryptobiome. Proceedings of the Royal Society B: Biological Sciences 286:20182697. DOI: 10.1098/rspb.2018.2697

CBOL Plant Working Group, Hollingsworth PM., Forrest LL., Spouge JL., Hajibabaei M., Ratnasingham S., van der Bank M., Chase MW., Cowan RS., Erickson DL. (2009) A DNA barcode for land plants. Proceedings of the National Academy of Sciences 106:12794–12797.

Chaisson MJ., Brinza D., Pevzner PA. (2009) De novo fragment assembly with short mate-paired reads: Does the read length matter? Genome Research 19:336–346. DOI: 10.1101/gr.079053.108

Chen W., Zhang CK., Cheng Y., Zhang S., Zhao H. (2013) A Comparison of Methods for Clustering 16S rRNA Sequences into OTUs. PLoS ONE 8. DOI: 10.1371/journal.pone.0070837

Compson ZG., Monk WA., Hayden B., Bush A., O'Malley Z., Hajibabaei M., Porter TM., Wright MTG., Baker CJO., Al Manir MS., Curry RA., Baird DJ. (2019) Network-Based Biomonitoring: Exploring Freshwater Food Webs With Stable Isotope Analysis and DNA Metabarcoding. Frontiers in Ecology and Evolution 7:395. DOI: 10.3389/fevo.2019.00395

Cordier T., Alonso Sáez L., Apotheloz-Perret-Gentil L., Aylagas E., Bohan DA., Bouchez A., Chariton A., Creer S., Fruhe L., Keck F., Keeley N., Laroche O., Leese F., Pochon X., Stoeck T., Pawlowski J., Lanzén A. (2020) Ecosystems Monitoring Powered by Environmental Genomics: A Review of Current Strategies with An Implementation Roadmap. BIOLOGY. DOI: 10.20944/preprints202001.0278.v1

Corse E., Meglécz E., Archambaud G., Ardisson M., Martin J-F., Tougard C., Chappaz R., Dubut V. (2017) A from-benchtop-to-desktop workflow for validating HTS data and for taxonomic identification in diet metabarcoding studies. Molecular Ecology Resources 17:e146–e159. DOI: 10.1111/1755-0998.12703

Creer S., Fonseca VG., Porazinska DL., Giblin-Davis RM., Sung W., Power DM., Packer M., Carvalho GR., Blaxter ML., Lambshead PJD., Thomas WK. (2010) Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. Molecular Ecology 19:4–20. DOI: 10.1111/j.1365-294X.2009.04473.x

Dawid IB., Blackler AW. (1972) Maternal and cytoplasmic inheritance of mitochondrial DNA in Xenopus. Developmental Biology 29:152–161. DOI: 10.1016/0012-1606(72)90052-8

De Salle ROB. (2006) Species Discovery versus Species Identification in DNA Barcoding Efforts: Response to Rubinoff. Conservation Biology 20:1545–1547. DOI: 10.1111/j.1523-1739.2006.00543.x

Del Fabbro C., Scalabrin S., Morgante M., Giorgi FM. (2013) An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. PLoS ONE 8. DOI: 10.1371/journal.pone.0085024

Derocles SAP., Bohan DA., Dumbrell AJ., Kitson JJN., Massol F., Pauvert C., Plantegenest M., Vacher C., Evans DM. (2018) Biomonitoring for the 21st Century: Integrating Next-Generation Sequencing Into Ecological Network Analysis. In: *Advances in Ecological Research*. Elsevier, p 1–62 DOI: 10.1016/bs.aecr.2017.12.001

van Dijk EL., Auger H., Jaszczyszyn Y., Thermes C. (2014) Ten years of next-generation sequencing technology. Trends in Genetics 30:418–426. DOI: 10.1016/j.tig.2014.07.001

Dupuis JR., Roe AD., Sperling F a. H. (2012) Multi-locus species delimitation in closely related animals and fungi: one marker is not enough. Molecular Ecology 21:4422–4436. DOI: 10.1111/j.1365-294X.2012.05642.x

Edgar RC. (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics (Oxford, England) 26:2460–2461. DOI: 10.1093/bioinformatics/btq461

Edgar RC. (2016a) UCHIME2: improved chimera prediction for amplicon sequencing. bioRxiv:074252. DOI: 10.1101/074252

Edgar RC. (2016b) UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv:081257. DOI: 10.1101/081257

Edgar RC. (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nature methods 10:996.

Edgar RC. (2018) Updating the 97% identity threshold for 16S ribosomal RNA OTUs. Bioinformatics 34:2371–2375. DOI: 10.1093/bioinformatics/bty113

Edgar RC., Flyvbjerg H. (2015) Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics 31:3476–3482. DOI: 10.1093/bioinformatics/btv401

Edgar RC., Haas BJ., Clemente JC., Quince C., Knight R. (2011) UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27:2194–2200. DOI: 10.1093/bioinformatics/btr381

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. DOI: 10.1371/journal.pone.0130324

Elbrecht V., Vamos EE., Meissner K., Aroviita J., Leese F. (2017) Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. Methods in Ecology and Evolution 8:1265–1275. DOI: 10.1111/2041-210X.12789

Elbrecht V., Vamos EE., Steinke D., Leese F. (2018) Estimating intraspecific genetic diversity from community DNA metabarcoding data. PeerJ 6:e4644. DOI: 10.7717/peerj.4644

Eren AM., Morrison HG., Lescault PJ., Reveillaud J., Vineis JH., Sogin ML. (2015) Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. The ISME journal 9:968–979. DOI: 10.1038/ismej.2014.195

Estrada E. (2007) Food webs robustness to biodiversity loss: The roles of connectance, expansibility and degree distribution. Journal of Theoretical Biology 244:296–307. DOI: 10.1016/j.jtbi.2006.08.002

Ewing B., Green P. (1998) Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. Genome Research 8:186–194. DOI: 10.1101/gr.8.3.186

Fadrosh DW., Ma B., Gajer P., Sengamalay N., Ott S., Brotman RM., Ravel J. (2014) An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. Microbiome 2:6. DOI: 10.1186/2049-2618-2-6

Frøslev TG., Kjøller R., Bruun HH., Ejrnæs R., Brunbjerg AK., Pietroni C., Hansen AJ. (2017) Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. Nature Communications 8:1188. DOI: 10.1038/s41467-017-01312-x

Galtier N., Nabholz B., Glémin S., Hurst GDD. (2009) Mitochondrial DNA as a marker of molecular diversity: a reappraisal. Molecular Ecology 18:4541–4550. DOI: 10.1111/j.1365-294X.2009.04380.x

Geneious Prime 2019.2.3 (2019)

Gibson J., Shokralla S., Porter TM., King I., van Konynenburg S., Janzen DH., Hallwachs W., Hajibabaei M. (2014) Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. Proceedings of the National Academy of Sciences 111:8007–8012.

Gilbert AJ. (2009) Connectance indicates the robustness of food webs when subjected to species loss. Ecological Indicators 9:72–80. DOI: 10.1016/j.ecolind.2008.01.010

Gompert Z., Lucas LK., Buerkle CA., Forister ML., Fordyce JA., Nice CC. (2014) Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. Molecular Ecology 23:4555–4573. DOI: 10.1111/mec.12811

González A., Dubut V., Corse E., Mekdad R., Dechatre T., Meglécz E. (2020) VTAM: A robust pipeline for validating metabarcoding data using internal controls. bioRxiv:2020.11.06.371187. DOI: 10.1101/2020.11.06.371187

Griggio F., Voskoboynik A., Iannelli F., Justy F., Tilak M-K., Xavier T., Pesole G., Douzery EJP., Mastrototaro F., Gissi C. (2014) Ascidian Mitogenomics: Comparison of Evolutionary Rates in Closely Related Taxa Provides Evidence of Ongoing Speciation Events. Genome Biology and Evolution 6:591–605. DOI: 10.1093/gbe/evu041

Hajibabaei M., Shokralla S., Zhou X., Singer GAC., Baird DJ. (2011) Environmental Barcoding: A Next-Generation Sequencing Approach for Biomonitoring Applications Using River Benthos. PLOS ONE 6:e17497. DOI: 10.1371/journal.pone.0017497

Hajibabaei M., Singer GA., Hebert PD., Hickey DA. (2007) DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. TRENDS in Genetics 23:167–172.

Hao X., Jiang R., Chen T. (2011) Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. Bioinformatics 27:611–618. DOI: 10.1093/bioinformatics/btq725

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

Hebert PD., Gregory TR. (2005) The promise of DNA barcoding for taxonomy. Systematic biology 54:852–859.

Hebert PD., Stoeckle MY., Zemlak TS., Francis CM. (2004) Identification of birds through DNA barcodes. PLoS biology 2.

Hebert PDN., Ratnasingham S., deWaard JR. (2003b) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proceedings of the Royal Society B: Biological Sciences 270:S96–S99. DOI: 10.1098/rsbl.2003.0025

Hillis DM., Dixon MT. (1991) Ribosomal DNA: molecular evolution and phylogenetic inference. The Quarterly review of biology 66:411–453.

Hinsinger DD., Debruyne R., Thomas M., Denys GPJ., Mennesson M., Utage J., Dettai A. (2015) Fishing for barcodes in the Torrent: from COI to complete mitogenomes on NGS platforms. DNA Barcodes 3:170–186. DOI: 10.1515/dna-2015-0019

Huson DH., Auch AF., Qi J., Schuster SC. (2007) MEGAN analysis of metagenomic data. Genome Research 17:377–386. DOI: 10.1101/gr.5969107

Huson DH., Beier S., Flade I., Górska A., El-Hadidi M., Mitra S., Ruscheweyh H-J., Tappu R. (2016) MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLOS Computational Biology 12:e1004957. DOI: 10.1371/journal.pcbi.1004957

Huson DH., Mitra S., Ruscheweyh H-J., Weber N., Schuster SC. (2011) Integrative analysis of environmental sequences using MEGAN4. Genome Research 21:1552–1560. DOI: 10.1101/gr.120618.111

Illumina Inc. (2010) Illumina Sequencing Technology. 5.

Illumina Inc. (2011) Quality Scores for Next-Generation Sequencing. 2.

Inglis C., Rahmani D., Yeung D., Mun D. (2022) A comparison of metabarcoding analysis between ASVs and OTUs - using data regarding the effects of chronic radiation on the bank vole gut microbiota. 27.

Institute for Integrative Genome Biology UC Riverside. (2012) Illumina 1.8 FASTQ Format.

Irwin DE., Rubtsov AS., Panov EN. (2009) Mitochondrial introgression and replacement between yellowhammers (Emberiza citrinella) and pine buntings (Emberiza leucocephalos)(Aves: Passeriformes). Biological Journal of the Linnean Society 98:422–438.

Janzen DH. (2004) Now is the time. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences 359:731–732.

Johnson SS., Hebsgaard MB., Christensen TR., Mastepanov M., Nielsen R., Munch K., Brand T., Gilbert MTP., Zuber MT., Bunce M., Rønn R., Gilichinsky D., Froese D., Willerslev E. (2007) Ancient bacteria show evidence of DNA repair. Proceedings of the National Academy of Sciences 104:14401–14405. DOI: 10.1073/pnas.0706787104

Joos L., Beirinckx S., Haegeman A., Debode J., Vandecasteele B., Baeyen S., Goormachtig S., Clement L., De Tender C. (2020) Daring to be differential: metabarcoding analysis of soil and plant-related microbial communities using amplicon sequence variants and operational taxonomical units. BMC Genomics 21:733. DOI: 10.1186/s12864-020-07126-4

Kahlke T., Ralph PJ. (2019) BASTA – Taxonomic classification of sequences and sequence bins using last common ancestor estimations. Methods in Ecology and Evolution 10:100–103. DOI: 10.1111/2041-210X.13095

Kennedy SR., Prost S., Overcast I., Rominger AJ., Gillespie RG., Krehenwinkel H. (2020) High-throughput sequencing for community analysis: the promise of DNA barcoding to uncover diversity, relatedness, abundances and interactions in spider communities. Development Genes and Evolution. DOI: 10.1007/s00427-020-00652-x

Kerdrel GA de., Andersen JC., Kennedy SR., Gillespie R., Krehenwinkel H. (2020) Rapid and cost-effective generation of single specimen multilocus barcoding data from whole arthropod communities by multiple levels of multiplexing. Scientific Reports 10:1–12. DOI: 10.1038/s41598-019-54927-z

Knowlton N., Weigt LA. (1998) New dates and new rates for divergence across the Isthmus of Panama. Proceedings of the Royal Society of London Series B: Biological Sciences 265:2257–2263.

Kozich JJ., Westcott SL., Baxter NT., Highlander SK., Schloss PD. (2013) Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. Applied and Environmental Microbiology 79:5112–5120. DOI: 10.1128/AEM.01043-13

Kozlov AM., Zhang J., Yilmaz P., Glöckner FO., Stamatakis A. (2016) Phylogeny-aware identification and correction of taxonomically mislabeled sequences. Nucleic Acids Research 44:5022–5033. DOI: 10.1093/nar/gkw396

Krehenwinkel H., Wolf M., Lim JY., Rominger AJ., Simison WB., Gillespie RG. (2017) Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. Scientific Reports 7:1–12. DOI: 10.1038/s41598-017-17333-x

Lamoril J., Ameziane N., Deybach J-C., Bouizegarène P., Bogard M. (2008) Les techniques de séquençage de l'ADN : une révolution en marche. Première partie. Immuno-analyse & Biologie Spécialisée 23:260–279. DOI: 10.1016/j.immbio.2008.07.016

Lansman RA., Shade RO., Shapira JF., Avise JC. (1981) The use of restriction endonucleases to measure mitochondrial DNA sequence relatedness in natural populations: III. Techniques and Potential Applications. Journal of Molecular Evolution 17:214–226. DOI: 10.1007/BF01732759

Laroche O., Pochon X., Tremblay LA., Ellis JI., Lear G., Wood SA. (2018) Incorporating molecular-based functional and co-occurrence network properties into benthic marine impact assessments. FEMS Microbiology Ecology 94. DOI: 10.1093/femsec/fiy167

Laver T., Harrison J., O'Neill PA., Moore K., Farbos A., Paszkiewicz K., Studholme DJ. (2015) Assessing the performance of the Oxford Nanopore Technologies MinION. Biomolecular Detection and Quantification 3:1–8. DOI: 10.1016/j.bdq.2015.02.001

Lavrov DV., Pett W. (2016) Animal Mitochondrial DNA as We Do Not Know It: mt-Genome Organization and Evolution in Nonbilaterian Lineages. Genome Biology and Evolution 8:2896–2913. DOI: 10.1093/gbe/evw195

Leduc N., Lacoursière-Roussel A., Howland KL., Archambault P., Sevellec M., Normandeau E., Dispas A., Winkler G., McKindsey CW., Simard N., Bernatchez L. (2019) Comparing eDNA metabarcoding and species collection for documenting Arctic metazoan biodiversity. Environmental DNA 1:342–358. DOI: 10.1002/edn3.35

Leray M., Knowlton N. (2015) DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. Proceedings of the National Academy of Sciences 112:2076–2081. DOI: 10.1073/pnas.1424997112

Leray M., Knowlton N., Machida RJ. (2022) MIDORI2: A collection of quality controlled, preformatted, and regularly updated reference databases for taxonomic assignment of eukaryotic mitochondrial sequences. Environmental DNA:edn3.303. DOI: 10.1002/edn3.303

Leray M., Meyer CP., Mills SC. (2015) Metabarcoding dietary analysis of coral dwelling predatory fish demonstrates the minor contribution of coral mutualists to their highly partitioned, generalist diet. PeerJ 3:e1047. DOI: 10.7717/peerj.1047

Leray M., Yang JY., Meyer CP., Mills SC., Agudelo N., Ranwez V., Boehm JT., Machida RJ. (2013) A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. Frontiers in Zoology 10:34. DOI: 10.1186/1742-9994-10-34

Liu L., Li Y., Li S., Hu N., He Y., Pong R., Lin D., Lu L., Law M. (2012) Comparison of Next-Generation Sequencing Systems. Journal of Biomedicine and Biotechnology 2012:e251364. DOI: https://doi.org/10.1155/2012/251364

Loman NJ., Misra RV., Dallman TJ., Constantinidou C., Gharbia SE., Wain J., Pallen MJ. (2012) Performance comparison of benchtop high-throughput sequencing platforms. Nature Biotechnology 30:434–439. DOI: 10.1038/nbt.2198

Macher J-N., Macher T-H., Leese F. (2017) Combining NCBI and BOLD databases for OTU assignment in metabarcoding and metagenomic datasets: The BOLD_NCBI _Merger. Metabarcoding and Metagenomics 1:e22262. DOI: 10.3897/mbmg.1.22262

Machida RJ., Knowlton N. (2012) PCR Primers for Metazoan Nuclear 18S and 28S Ribosomal DNA Sequences. PLoS ONE 7:e46180. DOI: 10.1371/journal.pone.0046180

Machida RJ., Leray M., Ho S-L., Knowlton N. (2017) Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. Scientific Data 4:1–7. DOI: 10.1038/sdata.2017.27

Madden T. (2013) The BLAST Sequence Analysis Tool. National Center for Biotechnology Information (US).

Magoč T., Salzberg SL. (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27:2957–2963. DOI: 10.1093/bioinformatics/btr507

Mao X., Zhang J., Zhang S., Rossiter SJ. (2010) Historical male-mediated introgression in horseshoe bats revealed by multilocus DNA sequence data. Molecular Ecology 19:1352–1366. DOI: 10.1111/j.1365-294X.2010.04560.x

Marquina D., Esparza-Salas R., Roslin T., Ronquist F. (2019) Establishing arthropod community composition using metabarcoding: Surprising inconsistencies between soil samples and preservative ethanol and homogenate from Malaise trap catches. Molecular Ecology Resources 19:1516–1530. DOI: 10.1111/1755-0998.13071

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

Masella AP., Bartram AK., Truszkowski JM., Brown DG., Neufeld JD. (2012) PANDAseq: paired-end assembler for illumina sequences. BMC Bioinformatics 13:31. DOI: 10.1186/1471-2105-13-31

McMurdie PJ., Holmes S. (2013) Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLOS ONE 8:e61217. DOI: 10.1371/journal.pone.0061217

Meier R., Wong W., Srivathsan A., Foo M. (2016) $1 DNA barcodes for reconstructing complex phenomes and finding rare species in specimen-rich samples. Cladistics 32:100–110. DOI: 10.1111/cla.12115

Meierotto S., Sharkey MJ., Janzen DH., Hallwachs W., Hebert PD., Chapman EG., Smith MA. (2019) A revolutionary protocol to describe understudied hyperdiverse taxa and overcome the taxonomic impediment. Deutsche Entomologische Zeitschrift 66:119.

Meiklejohn KA., Damaso N., Robertson JM. (2019) Assessment of BOLD and GenBank – Their accuracy and reliability for the identification of biological materials. PLOS ONE 14:e0217084. DOI: 10.1371/journal.pone.0217084

Minoche AE., Dohm JC., Himmelbauer H. (2011) Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. Genome Biology 12:R112. DOI: 10.1186/gb-2011-12-11-r112

Moore WS. (1995) Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees. Evolution 49:718–726.

Moritz C., Cicero C. (2004) DNA Barcoding: Promise and Pitfalls. PLOS Biology 2:e354. DOI: 10.1371/journal.pbio.0020354

Needleman SB., Wunsch CD. (1970) A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. J Mol Biol:443–453.

Nogales B., Lanfranconi MP., Piña-Villalonga JM., Bosch R. (2011) Anthropogenic perturbations in marine microbial communities. FEMS Microbiology Reviews 35:275–298. DOI: 10.1111/j.1574-6976.2010.00248.x

Odelberg SJ., Weiss RB., Hata A., White R. (1995) Template-switching during DNA synthesis by Thermus aquaticus DNA polymerase I. Nucleic Acids Research 23:2049–2057. DOI: 10.1093/nar/23.11.2049

Oliver TH., Heard MS., Isaac NJB., Roy DB., Procter D., Eigenbrod F., Freckleton R., Hector A., Orme CDL., Petchey OL., Proença V., Raffaelli D., Suttle KB., Mace GM., Martín-López B., Woodcock BA., Bullock JM. (2015) Biodiversity and Resilience of Ecosystem Functions. Trends in Ecology & Evolution 30:673–684. DOI: 10.1016/j.tree.2015.08.009

Pascault N., Roux S., Artigas J., Pesce S., Leloup J., Tadonleke RD., Debroas D., Bouchez A., Humbert J-F. (2014) A high-throughput sequencing ecotoxicology study of freshwater bacterial communities and their responses to tebuconazole. FEMS Microbiology Ecology 90:563–574. DOI: 10.1111/1574-6941.12416

Pawlowski J., Bruce K., Panksep K., Aguirre FI., Amafitano S., Apothéloz-Perret-Gentil L., Baussant T., Bouchez A., Carugati L., Cermakova K., Cordier T., Corinaldesi C., Costa FO., Danovaro R., Dell'Anno A., Duarte S., Eisendle U., Ferrari BJD., Frontalini F., Frühe L., Haegerbaeumer A., Kisand V., Krolicka A., Lanzén A., Leese F., Lejzerowicz F., Lyautey E., Maček I., Sagova-Marečková M., Pearman JK., Pochon X., Stoeck T., Vivien R., Weigand A., Fazi S. (2021) Environmental DNA metabarcoding for benthic monitoring: A review of sediment sampling and DNA extraction methods. Science of The Total Environment:151783. DOI: 10.1016/j.scitotenv.2021.151783

Payne RJ. (2013) Seven reasons why protists make useful bioindicators. Acta Protozoologica 52.

Pearman PB., Guisan A., Broennimann O., Randin CF. (2008) Niche dynamics in space and time. Trends in Ecology & Evolution 23:149–158.

Peña Cantero ÁL., Sentandreu V., Latorre A. (2009) Phylogenetic relationships of the endemic Antarctic benthic hydroids (Cnidaria, Hydrozoa): what does the mitochondrial 16S rRNA tell us about it? Polar Biology 33:41. DOI: 10.1007/s00300-009-0683-5

Perea S., Vukić J., Šanda R., Doadrio I. (2016) Ancient Mitochondrial Capture as Factor Promoting Mitonuclear Discordance in Freshwater Fishes: A Case Study in the Genus Squalius (Actinopterygii, Cyprinidae) in Greece. PLOS ONE 11:e0166292. DOI: 10.1371/journal.pone.0166292

Plaisance L., Caley MJ., Brainard RE., Knowlton N. (2011) The Diversity of Coral Reefs: What Are We Missing? PLOS ONE 6:e25026. DOI: 10.1371/journal.pone.0025026

Plaisance L., Knowlton N., Paulay G., Meyer C. (2009) Reef-associated crustacean fauna: biodiversity estimates using semi-quantitative sampling and DNA barcoding. Coral Reefs 28:977–986.

Pochon X., Wood SA., Keeley NB., Lejzerowicz F., Esling P., Drew J., Pawlowski J. (2015) Accurate assessment of the impact of salmon farming on benthic sediment enrichment using foraminiferal metabarcoding. Marine Pollution Bulletin 100:370–382. DOI: 10.1016/j.marpolbul.2015.08.022

Porco D., Rougerie R., Deharveng L., Hebert P. (2010) Coupling non-destructive DNA extraction and voucher retrieval for small soft-bodied Arthropods in a high-throughput context: the example of Collembola. Molecular Ecology Resources 10:942–945. DOI: 10.1111/j.1755-0998.2010.2839.x

Porter TM., Hajibabaei M. (2018) Over 2.5 million COI sequences in GenBank and growing. PLoS ONE 13. DOI: 10.1371/journal.pone.0200177

Prodan A., Tremaroli V., Brolin H., Zwinderman AH., Nieuwdorp M., Levin E. (2020) Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. PLOS ONE 15:e0227434. DOI: 10.1371/journal.pone.0227434

Prugnolle F., de Meeus T. (2002) Inferring sex-biased dispersal from population genetic tools: a review. Heredity 88:161–165. DOI: 10.1038/sj.hdy.6800060

Puertas MJ., González-Sánchez M. (2020) Insertions of mitochondrial DNA into the nucleus—effects and role in cell evolution. Genome 63:365–374. DOI: 10.1139/gen-2019-0151

Puillandre N., Lambert A., Brouillet S., Achaz G. (2012) ABGD, Automatic Barcode Gap Discovery for primary species delimitation. Molecular Ecology 21:1864–1877. DOI: 10.1111/j.1365-294X.2011.05239.x

Quick J., Grubaugh ND., Pullan ST., Claro IM., Smith AD., Gangavarapu K., Oliveira G., Robles-Sikisaka R., Rogers TF., Beutler NA., Burton DR., Lewis-Ximenez LL., de Jesus JG., Giovanetti M., Hill SC., Black A., Bedford T., Carroll MW., Nunes M., Alcantara LC., Sabino EC., Baylis SA., Faria NR., Loose M., Simpson JT., Pybus OG., Andersen KG., Loman NJ. (2017) Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nature Protocols 12:1261–1276. DOI: 10.1038/nprot.2017.066

Ranwez V., Harispe S., Delsuc F., Douzery EJP. (2011) MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. PLOS ONE 6:e22594. DOI: 10.1371/journal.pone.0022594

Ratnasingham S. (2019) MBRAVE: The Multiplex Barcode Research And Visualization Environment. Biodiversity Information Science and Standards 3:e37986. DOI: 10.3897/biss.3.37986

Ratnasingham S., Hebert PDN. (2007) BOLD: The Barcode of Life Data System. Molecular Ecology Notes. DOI: 10.1111/j.1471-8286.2006.01678.x

Rideout JR., He Y., Navas-Molina JA., Walters WA., Ursell LK., Gibbons SM., Chase J., McDonald D., Gonzalez A., Robbins-Pianka A., Clemente JC., Gilbert JA., Huse SM., Zhou H-W., Knight R., Caporaso JG. (2014) Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. PeerJ 2:e545. DOI: 10.7717/peerj.545

Rivera SF., Vasselon V., Jacquet S., Bouchez A., Ariztegui D., Rimet F. (2018) Metabarcoding of lake benthic diatoms: from structure assemblages to ecological assessment. Hydrobiologia 807:37–51. DOI: 10.1007/s10750-017-3381-2

Robeson II MS., O'Rourke DR., Kaehler BD., Ziemski M., Dillon MR., Foster JT., Bokulich NA. (2021) RESCRIPt: Reproducible sequence taxonomy reference database management. PLOS Computational Biology 17:e1009581. DOI: 10.1371/journal.pcbi.1009581

Rodrigue S., Materna AC., Timberlake SC., Blackburn MC., Malmstrom RR., Alm EJ., Chisholm SW. (2010) Unlocking Short Read Sequencing for Metagenomics. PLOS ONE 5:e11840. DOI: 10.1371/journal.pone.0011840

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

Rose TM., Henikoff JG., Henikoff S. (2003) CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design. Nucleic Acids Research 31:3763–3766. DOI: 10.1093/nar/gkg524

Rubinoff D. (2006) Utility of Mitochondrial DNA Barcodes in Species Conservation. Conservation Biology 20:1026–1033. DOI: 10.1111/j.1523-1739.2006.00372.x

Sahlin K., Medvedev P. (2021) Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis. Nature Communications 12:2. DOI: 10.1038/s41467-020-20340-8

Saitoh S., Aoyama H., Fujii S., Sunagawa H., Nagahama H., Akutsu M., Shinzato N., Kaneko N., Nakamori T. (2016) A quantitative protocol for DNA metabarcoding of springtails (Collembola). The 6th International Barcode of Life Conference 01:705–723. DOI: 10.1139/gen-2015-0228@gen-iblf.issue01

Sanna D., Lai T., Francalacci P., Curini-Galletti M., Casu M. (2009) Population structure of the Monocelis lineata (Proseriata, Monocelididae) species complex assessed by phylogenetic analysis of the mitochondrial Cytochrome c Oxidase subunit I (COI) gene. Genetics and Molecular Biology 32:864–867. DOI: 10.1590/S1415-47572009005000076

Schirmer M., D'Amore R., Ijaz UZ., Hall N., Quince C. (2016) Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC Bioinformatics 17:125. DOI: 10.1186/s12859-016-0976-y

Schirmer M., Ijaz UZ., D'Amore R., Hall N., Sloan WT., Quince C. (2015) Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Research 43:e37–e37. DOI: 10.1093/nar/gku1341

Schloss PD. (2020) Reintroducing mothur: 10 Years Later. Applied and Environmental Microbiology 86. DOI: 10.1128/AEM.02343-19

Schloss PD., Westcott SL., Ryabin T., Hall JR., Hartmann M., Hollister EB., Lesniewski RA., Oakley BB., Parks DH., Robinson CJ., Sahl JW., Stres B., Thallinger GG., Horn DJV., Weber CF. (2009) Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Applied and Environmental Microbiology 75:7537–7541. DOI: 10.1128/AEM.01541-09

Schoch CL., Seifert KA., Huhndorf S., Robert V., Spouge JL., Levesque CA., Chen W., Consortium FB. (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proceedings of the National Academy of Sciences 109:6241–6246.

Shokralla S., Porter TM., Gibson JF., Dobosz R., Janzen DH., Hallwachs W., Golding GB., Hajibabaei M. (2015) Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. Scientific Reports 5:9687. DOI: 10.1038/srep09687

Shokralla S., Spall JL., Gibson JF., Hajibabaei M. (2012) Next-generation sequencing technologies for environmental DNA research. Molecular Ecology 21:1794–1805. DOI: 10.1111/j.1365-294X.2012.05538.x

Smyth RP., Schlub TE., Grimm A., Venturi V., Chopra A., Mallal S., Davenport MP., Mak J. (2010) Reducing chimera formation during PCR amplification to ensure accurate genotyping. Gene 469:45–51. DOI: 10.1016/j.gene.2010.08.009

Song H., Buhay JE., Whiting MF., Crandall KA. (2008) Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. Proceedings of the National Academy of Sciences 105:13486–13491. DOI: 10.1073/pnas.0803076105

Srivastava DS., Cadotte MW., MacDonald AAM., Marushia RG., Mirotchnick N. (2012) Phylogenetic diversity and the functioning of ecosystems. Ecology Letters 15:637–648. DOI: 10.1111/j.1461-0248.2012.01795.x

Srivathsan A., Hartop E., Puniamoorthy J., Lee WT., Kutty SN., Kurina O., Meier R. (2019) Rapid, large-scale species discovery in hyperdiverse taxa using 1D MinION sequencing. BMC Biology 17:96. DOI: 10.1186/s12915-019-0706-9

Stoeck T., Frühe L., Forster D., Cordier T., Martins CIM., Pawlowski J. (2018) Environmental DNA metabarcoding of benthic bacterial communities indicates the benthic footprint of salmon aquaculture. Marine Pollution Bulletin 127:139–149. DOI: 10.1016/j.marpolbul.2017.11.065

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

Taberlet P., Coissac E., Pompanon F., Brochmann C., Willerslev E. (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular ecology 21:2045–2050.

Tang CQ., Leasi F., Obertegger U., Kieneke A., Barraclough TG., Fontaneto D. (2012) The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. Proceedings of the National Academy of Sciences 109:16208–16212. DOI: 10.1073/pnas.1209160109

Tautz D., Arctander P., Minelli A., Thomas RH., Vogler AP. (2003) A plea for DNA taxonomy. Trends in ecology & evolution 18:70–74.

Thomas AC., Deagle BE., Eveson JP., Harsch CH., Trites AW. (2016) Quantitative DNA metabarcoding: improved estimates of species proportional biomass using correction factors derived from control material. Molecular Ecology Resources 16:714–726. DOI: 10.1111/1755-0998.12490

Thomsen PF., Kielgast J., Iversen LL., Møller PR., Rasmussen M., Willerslev E. (2012) Detection of a diverse marine fish fauna using environmental DNA from seawater samples. PLoS one 7.

Thomsen PF., Willerslev E. (2015) Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. Biological Conservation 183:4–18. DOI: 10.1016/j.biocon.2014.11.019

Valentini A., Pompanon F., Taberlet P. (2009) DNA barcoding for ecologists. Trends in ecology & evolution 24:110–117.

Vázquez DP. (2005) Degree distribution in plant–animal mutualistic networks: forbidden links or random interactions? Oikos 108:421–426. DOI: 10.1111/j.0030-1299.2005.13619.x

Wägele H., Klussmann-Kolb A., Kuhlmann M., Haszprunar G., Lindberg D., Koch A., Wägele JW. (2011) The taxonomist - an endangered race. A practical proposal for its survival. Frontiers in Zoology 8:25. DOI: 10.1186/1742-9994-8-25

Wang Q., Garrity GM., Tiedje JM., Cole JR. (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology 73:5261–5267. DOI: 10.1128/AEM.00062-07

Wang WY., Srivathsan A., Foo M., Yamane SK., Meier R. (2018) Sorting specimen-rich invertebrate samples with cost-effective NGS barcodes: Validating a reverse workflow for specimen processing. Molecular ecology resources 18:490–501.

Webb CO., Ackerly DD., McPeek MA., Donoghue MJ. (2002) Phylogenies and Community Ecology. Annual Review of Ecology and Systematics 33:475–505. DOI: 10.1146/annurev.ecolsys.33.010802.150448

Wiemers M., Fiedler K. (2007) Does the DNA barcoding gap exist? – a case study in blue butterflies (Lepidoptera: Lycaenidae). Frontiers in Zoology 4:8. DOI: 10.1186/1742-9994-4-8

Wiens JJ., Penkrot TA. (2002) Delimiting species using DNA and morphological variation and discordant species limits in spiny lizards (Sceloporus). Systematic biology 51:69–91.

Williams RJ. (2011) Biology, Methodology or Chance? The Degree Distributions of Bipartite Ecological Networks. PLOS ONE 6:e17645. DOI: 10.1371/journal.pone.0017645

Woese CR., Fox GE. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proceedings of the National Academy of Sciences 74:5088–5090.

Wong WH., Tay YC., Puniamoorthy J., Balke M., Cranston PS., Meier R. (2014) 'Direct PCR'optimization yields a rapid, cost-effective, nondestructive and efficient method for obtaining DNA barcodes without DNA extraction. Molecular Ecology Resources 14:1271–1280.

Xiong W., Zhan A. (2018) Testing clustering strategies for metabarcoding-based investigation of community–environment interactions. Molecular Ecology Resources 18:1326–1338. DOI: 10.1111/1755-0998.12922

Yeo D., Puniamoorthy J., Ngiam RWJ., Meier R. (2018) Towards holomorphology in entomology: rapid and cost-effective adult–larva matching using NGS barcodes. Systematic entomology 43:678–691.

Yilmaz P., Parfrey LW., Yarza P., Gerken J., Pruesse E., Quast C., Schweer T., Peplies J., Ludwig W., Glöckner FO. (2014) The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Research 42:D643–D648. DOI: 10.1093/nar/gkt1209

Yu DW., Ji Y., Emerson BC., Wang X., Ye C., Yang C., Ding Z. (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods in Ecology and Evolution 3:613–623. DOI: 10.1111/j.2041-210X.2012.00198.x

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

Ziegler M., Roik A., Porter A., Zubier K., Mudarris MS., Ormond R., Voolstra CR. (2016) Coral microbial community dynamics in response to anthropogenic impacts near a major city in the central Red Sea. Marine Pollution Bulletin 105:629–640. DOI: 10.1016/j.marpolbul.2015.12.045

5. Annexes du chapitre 2

Annexe 2.1. Références du tableau 2.1 : Avantages et inconvénients des marqueurs utilisés pour la taxonomie moléculaire

1. Yu, D. W. et al. Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods Ecol. Evol. 3, 613–623 (2012).
2. Folmer, O., Black, M., Hoeh, W., Lutz, R. & Vrijenhoek, R. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. 7 (1994).
3. Machida, R. J. & Knowlton, N. PCR Primers for Metazoan Nuclear 18S and 28S Ribosomal DNA Sequences. PLoS ONE 7, e46180 (2012).
4. Mueller, R. L. Evolutionary Rates, Divergence Dates, and the Performance of Mitochondrial Genes in Bayesian Phylogenetic Analysis. Syst. Biol. 55, 289–300 (2006).
5. Leduc, N. et al. Comparing eDNA metabarcoding and species collection for documenting Arctic metazoan biodiversity. Environ. DNA 1, 342–358 (2019).
6. Sanna, D., Lai, T., Francalacci, P., Curini-Galletti, M. & Casu, M. Population structure of the Monocelis lineata (Proseriata, Monocelididae) species complex assessed by phylogenetic analysis of the mitochondrial Cytochrome c Oxidase subunit I (COI) gene. Genet. Mol. Biol. 32, 864–867 (2009).
7. Tang, C. Q. et al. The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. Proc. Natl. Acad. Sci. 109, 16208–16212 (2012).
8. Bhadury, P. et al. Development and evaluation of a DNA-barcoding approach for the rapid identification of nematodes. Mar. Ecol. Prog. Ser. 320, 1–9 (2006).
9. Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E. Towards next-generation biodiversity assessment using DNA metabarcoding. Mol. Ecol. 21, 2045–2050 (2012).
10. Holovachov, O., Haenel, Q., Bourlat, S. J. & Jondelius, U. Taxonomy assignment approach determines the efficiency of identification of OTUs in marine nematodes. R. Soc. Open Sci. 4, 170315 (2017).
11. Peña Cantero, Á. L., Sentandreu, V. & Latorre, A. Phylogenetic relationships of the endemic Antarctic benthic hydroids (Cnidaria, Hydrozoa): what does the mitochondrial 16S rRNA tell us about it ? Polar Biol. 33, 41 (2009).
12. Moran, M. A. The global ocean microbiome. Science 350, (2015).
13. Hillis, D. M. & Dixon, M. T. Ribosomal DNA: molecular evolution and phylogenetic inference. Q. Rev. Biol. 66, 411–453 (1991).
14. Creer, S. et al. Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. Mol. Ecol. 19, 4–20 (2010).
15. Drummond, A. J. et al. Evaluating a multigene environmental DNA approach for biodiversity assessment. GigaScience 4, (2015).
16. Gibson, J. et al. Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. Proc. Natl. Acad. Sci. 111, 8007–8012 (2014)

Chapitre 3 : Création d'un référentiel moléculaire pour les Mascareignes

Résumé :

Les méthodes d'identification moléculaire sont de plus en plus plébiscitées pour étudier la biodiversité. Ces approches nécessitent des bases de référence taxonomiquement diversifiées et spécifiques à la région géographique étudiée. Cependant, la grande diversité des taxons présents dans les récifs coralliens complexifie l'exhaustivité de ces bases de référence, or leur incomplétude peut entraîner des assignations erronées. Pour limiter ces erreurs, il est recommandé de travailler à partir de bases de données vérifiées et restreintes aux taxons ciblés d'un point de vue taxonomique et géographique. Cependant, le cryptobiome des Mascareignes reste très peu documenté et peu de séquences barcodes locale sont disponibles.

Ainsi, ce troisième chapitre s'attache à documenter le cryptobiome des Mascareignes et à poser les bases d'un référentiel moléculaire local qui permettra d'augmenter la résolution taxonomique des futures études dans le Sud-Ouest de l'océan Indien. Plus de 4 500 spécimens, issus de 54 ARMS, ont été collectés, photographiés et conditionnés dans l'éthanol afin de documenter morphologiquement et génétiquement le cryptobiome récifal de deux îles du Sud-Ouest de l'océan Indien, La Réunion et Rodrigues. Des séquences ont été produites pour un marqueur nucléaire, le 18S (190 séquences) et deux marqueurs mitochondriaux, le COI (200 séquences) et le 16S (100 séquences, non utilisés dans la suite des analyses). L'ajout de ces séquences locales aux bases de référence publiques ne représente que 0,04 % de leur taille respective mais a permis d'améliorer de 1,8 % et 16,4 % le nombre d'OTU assignés pour le 18S et le COI, respectivement.

De plus, la diversité au sein des ARMS déployés dans les Mascareignes concorde avec celle observée dans d'autres régions du monde, présentant une majorité d'arthropodes, d'annélides, de mollusques et de porifères. Cependant, malgré les efforts déployés, seule une petite fraction du cryptobiome a pu être identifiée. Ces résultats soulignent le caractère unique du cryptobiome des Mascareignes et la nécessité de poursuivre les efforts pour compléter les bases de référence locales.

1. Introduction

L'un des défis du métabarcoding est la production de jeux de données robustes, à la fois en termes de conception expérimentale, avec par exemple l'utilisation de réplicats, de contrôles négatifs et de différents marqueurs moléculaires et la sélection minutieuse des amorces **(Cristescu & Hebert 2018)**, mais également d'analyses robustes en termes de pipelines bio-informatiques appropriés aux communautés étudiées **(Zinger et al. 2019)**. Ainsi, une des étapes cruciales dans la production d'OTU assignés est la qualité des bases de référence employées. La base de référence idéale doit être taxonomiquement diversifiée avec une couverture large des phylums ciblés, mais également être spécifique à la région géographique étudiée et être exempte de séquences mal identifiées et/ou erronées **(Mugnai et al. 2023)**. Toutefois, pour de nombreux groupes taxonomiques, et notamment pour les taxons marins, cet objectif est rarement atteint **(Duarte et al. 2021 ; Mugnai et al. 2021)** en particulier pour certaines régions géographiques sous étudiées **(Monchamp et al. 2023)**, telles que l'archipel des Mascareignes.

Une proportion élevée de taxons manquants dans les bases de référence entraîne l'absence d'assignement pour de nombreuses séquences. Cette absence peut être due au fait que que certaines n'appartiennent pas aux groupes taxonomiques ciblés. Par exemple, les amorces de COI, qui sont généralement utilisées pour l'étude des métazoaires, amplifient également les algues, les diatomées et les bactéries. Il peut donc être judicieux d'inclure des séquences de taxons non ciblés dans les bases de référence pour retrouver les OTU correspondants et les retirer. Cependant, il est très probable que de nombreux OTU ne soient pas assignés en raison de l'absence de séquences de référence appropriées **(Ransome et al. 2017 ; Mugnai et al. 2023)** Des bases de référence généralistes plus complètes, comprenant l'ensemble des taxons disponibles, semble être la solution à privilégier pour maximiser les assignements. Idéalement, la présence des taxons non ciblés dans les bases de référence ne devrait pas interférer dans l'assignement des OTU, mais les méthodes d'assignement employées, telles que blast, assignent à la séquence la plus similaire. Or, s'il n'y a pas de taxon proche présent dans les références, la probabilité que la séquence la plus proche corresponde par hasard à un taxon non-cible augmente. Ainsi, pour réduire les assignements erronés, il est recommandé d'utiliser une base de référence restreinte **(Mugnai et al. 2023)** et des seuils de similarités stricts, par exemple 97 % pour le COI. C'est pourquoi les assignements de cette étude ont été réalisés (1) en combinant différentes méthodes d'assignement (plus proche voisin et ancêtre commun), (2) avec différents seuils de similarité et (3) à partir de bases de référence nettoyées (cf. Chapitre 2).

Par ailleurs, la grande diversité taxonomique des récifs coralliens complexifie la construction de bases de référence exhaustives. En effet, (1) ils incluent 31 des 35 embranchements métazoaires (Tableau 3.1) ; (2) même si aucun phylum n'est inféodé aux récifs coralliens, il est commun que certaines familles et de nombreuses espèces le soient, comme les bénitiers (Cardiidae) et certaines familles d'octocoralliaires (ex. Helioporidae) ; (3) certaines familles y atteignent leur plus forte diversité, comme les gastéropodes Mitridae, Cypraeidae et Conidae, les poissons Scarinae et Chaetodontidae, les crustacés Gonodactylidae , ou encore la plupart des familles d'anthozoaires **(Paulay 1997)**.

L'inclusion de séquences locales aux bases de référence en amont des études de métabarcoding améliore la quantité et la qualité des assignements **(Mugnai et al. 2023)**. L'élaboration de ces bases de référence locales peut paraître simple, mais représente un véritable défi sur les plans taxonomique (identifier précisément les spécimens alors que les taxonomistes se font de plus en plus rares), moléculaire (par exemple choisir des amorces adaptées aux taxons) et bio-informatique (adapter et nettoyer les bases de données).

Dans cette thèse, nous avons employé une combinaison de marqueurs nucléaires (18S) et mitochondriaux (COI) pour déterminer la diversité qui compose le cryptobiome récifal des Mascareignes par métabarcoding. Les interprétations écologiques basées sur les méthodes

Tableau 3.1 : Liste des phylums métazoaires retrouvés ou non dans les récifs coralliens (adapté de Paulay 1997). Les phylums avec + ont été rajoutés à la description de Paulay de 1997, car découverts a posteriori. Les phylums en gras ont été retrouvés dans cette étude.

Largement trouvés (N=13)	Trouvés (N=18)	Non trouvés (N=4)
Annelida	Acanthocephala	Cycliophora
Arthropoda	Brachipoda	Micrognathozoa^{+}
Bryozoa	**Chaetognatha**	Onychophora (non marin)
Chordata	Ctenophora	Orthonectida
Cnidaria	Dicyemida	
Echinodermata	**Echiura**	
Mollusca	**Entoprocta**	
Nematoda	Gastrotricha	
Nemerta	Gnathostomulida	
Platyhelminthes	Hemichordata	
Porifera	**Kinorhyncha**	
Rotifera	Loricifrea	
Sipuncula	Nematomorpha	
	Phoronida	
	Placozoa	
	Priapula	
	Tardigrada	
	Xenacoelomorpha^{+}	

d'identification moléculaire sont grandement dépendantes de la qualité des identifications, et donc des bases de données. En outre, le cryptobiome des Mascareignes reste relativement peu connu. Ainsi, ce troisième chapitre s'attache à documenter le cryptobiome des Mascareignes et à poser les bases d'un référentiel moléculaire local qui permettra d'augmenter la résolution des approches moléculaires des futures études dans le Sud-Ouest de l'océan Indien.

2. Matériel et Méthodes

2.1. La classification et le séquençage barcoding des spécimens récoltés

L'ensemble des organismes récoltés ont été triés collégialement selon une connaissance généralistes des taxons, puis à partir de différents critères morphologiques tels que : la taille, la couleur, la forme, la texture et les motifs pour constituer des groupes de morpho-espèces hypothétiques. Tous les spécimens ont été photographiés vivants par H. Bruggemann. À partir des photos, le rang taxonomique le plus bas possible a été déterminé. Une partie des spécimens a été identifiée ultérieurement par des spécialistes (les ascidies par F. Monniot, les bryozoaires par J-L. D'Hondt) et dont certains conservés au MNHN. Les organismes ont été immergés dans l'éthanol à 95 % et conservés au réfrigérateur (4°C ; **Rimet et al. 2021**). Pour chaque morpho-espèce, deux individus ont été sélectionnés et leur ADN a été extrait à l'aide du kit Qiagen DNeasy Blood and Tissue en suivant le protocole du fournisseur. Lorsque la taille du spécimen le permettait, un morceau de tissu a été prélevé en parallèle et stocké à -20° C dans de l'éthanol à 99 % afin de réaliser une banque de tissus conservée à l'Université de La Réunion. Le sous-échantillonnage a été réalisé de façon à maximiser la préservation des zones utiles à une future identification morphologique. Trois marqueurs moléculaires ont été utilisés lors de l'amplification par PCR : le 18S, le COI et le 16S. Les amorces utilisées sont indiquées dans le tableau 3.2 **(Palumbi et al. 1991 ; Stefaniak et al. 2009 ; Heimeier et al. 2010 ; Machida & Knowlton 2012 ; Geller et al. 2013)** et les programmes PCR employés dans le tableau 3.3.

Dans la majorité des cas, les produits PCR ont été multiplexés (cf. **Hinsinger et al. 2015**) et séquencés par NGS (séquençage Illumina MiSeq 2 × 250 pb) par le Service de Systématique Moléculaire du MNHN (Muséum national d'Histoire Naturelle). Le reste des amplifiats a été envoyé à Genoscreen (Lille, France) et séquencés par séquençage Sanger.

Tableau 3.2 : Amorces oligonucleotidiques utilisées pour les amplifications. Le sens des amorces est indiqué par (F) pour *forward* et (R) pour *reverse*.

Cible	Amorces	Séquence (5' - 3')	Taille de la séquence	Référence
CO1	jgLCO1490 (F)	TNTCNACNAAYCAYAARGAYATTGG	~658	Geller et al. 2013
	jdHCO2198 (R)	TAAACYTCNGGRTGNCCRAARAAYCA		
CO1	Echino-COI-F (F)	TKTCDACDAAYCAYAAGGAYATTGG	~658	Heimeier et al. 2010
	Echino-COI-R (R)	TGRTTCTTCGGHCACCCVGARGTTTA		
CO1	Tun-*forward* (F)	TCGACTAATCATAAAGATATTA	~586	Stefaniak 2009
	Tun-reverse2 (R)	AACTTGTATTTAAATTACGATC		
18S	18Smk1F (F)	CTGGTGCCAGCAGCCGCGGYAA	~550	Machida & Knowlton 2012
	18Smk2R (R)	TCCGTCAATTYCTTTAAGTT		
16S	16S-Ar-L (F)	CGCCTGTTTATCAAAAACAT	~490	Palumbi 1991
	16S-Br-H (R)	CCGGTCTGAACTCAGATCACGT		

Tableau 3.3 : Programmes PCR en fonction des paires d'amorces

Marqueur	Amorces	Dénaturation initiale	Nombre de cycle	Dénaturation	Hybridation	Elongation	Elongation finale
COI	jgHCO2198 jgLCO1490	1 min 94°C	50	20s 94°C	30s 46°C	1 min 72°C	2 min 72 °C
COI	Echino-COI-F Echino-COI-R	1 min 94°C	60	20s 94°C	30s 48°C	1 min 72°C	2 min 72 °C
COI	Tun-forward Tun-reverse2	1 min 94°C	60	20s 94°C	30s 40°C	1 min 72°C	2 min 72 °C
18S	18Smk1F 18Smk2R	1 min 94°C	40	20s 94°C	30s 55°C	1 min 72°C	2 min 72 °C
16S	JB-16Sar + 16S-forward 16Sbr-H + 16S-reverse	1 min 94°C	50	20s 94°C	30s 50°C	1 min 72°C	2 min 72 °C

2.2. La reconstitution des séquences barcodes

Le choix de multiplexer les produits PCR pour créer le référentiel barcode permet de réduire le coût du séquençage mais implique une importante partie de traitements bio-informatiques pour reconstituer les séquences barcodes. En effet, pour récupérer la séquence correspondant au spécimen séquencé, les *reads* doivent être démultiplexés à l'aide la combinaison tag + amorce (cf. Chapitre 2 et **Hinsinger et al. 2015**) pour ne garder que les *reads* correspondant aux extrémités des PCRs du spécimen, puis les *reads* sont assemblés de façon à reconstituer la séquence complète (séquençage Illumina MiSeq 2 x 250 bp pour des séquences finales >650 bp).

Une partie des séquences ont été reconstruites manuellement à l'aide du logiciel Geneious Prime (Annexe 3.1). L'acquisition des séquences finales est très chronophage avec environ quatre

séquences reconstruites en 30 minutes, voire une heure en fonction de la complexité (qualité de la séquence, contaminations, etc.). Aucun logiciel ou pipeline ne permet aujourd'hui la reconstitution automatisée des séquences barcodes produites par NGS d'après la littérature. En raison du nombre important de séquences à acquérir pour construire le référentiel barcode, j'ai décidé de développer un pipeline bio-informatique ayant pour objectif de reconstruire les séquences barcodes pour les trois marqueurs moléculaires ciblés (18S, COI et 16S). Les scripts de ce pipeline ont été élaborés pour être efficaces sur les différents taxons et les différents marqueurs moléculaires tout en incluant des contrôles pour limiter les erreurs. Les séquences ne passant pas les étapes de vérification sont contrôlées et/ou reconstruites manuellement sur Geneious Prime.

Les procédures de vérification automatiques et manuelles ont été basées sur : (1) la concordance entre l'identification moléculaire et l'identification morphologique, (2) le pourcentage de dissimilarité entre la séquence barcode et la séquence de référence la plus proche ainsi que (3) l'homogénéité des assignements obtenus au sein d'une même base de référence (afin de vérifier la présence de séquences erronées dans les bases de référence, ainsi que la résolution taxonomique du marqueur, par exemple l'ensemble du genre de scléractiniaire *Pocillopora* a 100% de similarité pour le COI) et (4) entre les deux bases de référence GenBank et BOLD. Pour les séquences COI, une étape supplémentaire a été réalisée en alignant l'ensemble des séquences entre elles afin de vérifier la concordance des longueurs et des alignements.

Ci-dessous est détaillée la stratégie employée pour l'automatisation de la reconstitution des séquences barcodes. Les différentes étapes, représentées par les numéros, sont synthétisées dans la Figure 3.1.

1. En sortie de séquençage, la puce est démultiplexée par le prestataire en N banques (ici 40 banques) correspondant aux pools d'origine.
2. Chaque banque est ensuite démultiplexée à l'aide de cutadapt **(Martin 2011)** pour produire un fichier par combinaison tag + amorce. Les fichiers produits contiennent ainsi l'ensemble des *reads* qui possèdent la combinaison tag + amorce et sont classés dans des dossiers selon l'arborescence suivante : marqueur/tag/paire d'amorces
3. Dans chacun de ces dossiers, le fichier correspondant à l'amorce *forward* est utilisé pour regrouper les *reads* en contigs à l'aide de vsearch (de novo) **(Rognes et al. 2016)**.
4. La liste des contigs produite va permettre la sélection de la séquence de référence pour reconstruire le barcode. Pour cela, l'ensemble des contigs, un à un, vont être assignés à l'aide de blast **(Camacho et al. 2009)** à partir d'une base de référence locale (par exemple, MIDORI

pour le COI). Les contigs sont ensuite classés par nombres de *reads* décroissants. Si le premier contig n'obtient pas de correspondance avec la base de référence, alors le second est utilisé et ainsi de suite. Lorsqu'un contig obtient une ou des correspondances dans la base de référence, la taxonomie des séquences de référence est récupérée. Le phylum ainsi obtenu doit correspondre au phylum attendu (sur la base des identifications morphologiques) du spécimen séquencé pour que le contig soit retenu. Si aucune correspondance n'est trouvée pour l'ensemble des contigs, une deuxième tentative d'assignement est réalisée à partir de GenBank (à distance). Si à nouveau aucune correspondance n'est trouvée, les étapes 3 et 4 sont réalisées sur le fichier de l'amorce reverse.

5. Plusieurs tests ont été réalisés pour déterminer le meilleur critère dans le choix du contig qui servira de référence (c'est-à-dire, le contig ayant le consensus qui correspond le plus à l'extrémité de la séquence barcode ~250 bp). La sélection sur le nombre de *reads* par contig ou bien le contig le plus long montre des taux de reconstitution correcte plus faible qu'une sélection sur la longueur de l'alignement. Ainsi la séquence consensus du contig qui permet le plus long alignement (en autorisant l'insertion et la délétion ; basé sur le CIGAR) est sélectionnée comme référence
6. En parallèle, un assemblage *de novo* est exécuté sur l'ensemble de la banque avec megahit pour obtenir une liste de l'ensemble des contigs contenus dans la banque.
7. La séquence sélectionnée à l'étape 5, sert de référence pour l'alignement des contigs produits à l'étape 6. Cette étape, réalisée avec bowtie2, permet d'allonger la taille de la séquence de référence (similaire à *map to reference with iteration* dans Geneious Prime). Ainsi, on obtient une nouvelle liste de contigs ayant des séquences consensus plus longues que précédemment et qui correspondent toujours à l'extrémité de la séquence barcode.
8. Le consensus du contig ayant le plus de *reads* est retenu comme référence pour un dernier alignement avec l'ensemble des *reads* contenus dans la banque (bowtie2). Cette étape permet d'obtenir une séquence consensus déterminée à partir de l'ensemble des *reads* de la banque et ainsi de limiter le risque d'erreur en ayant une profondeur de séquençage/couverture de *reads* maximale.
9. La séquence consensus produite passe une étape de PCR *in silico* qui permet de vérifier la présence des amorces et de les retirer, de vérifier le sens de la séquence produite et de vérifier la longueur de la séquence barcode. Si besoin, la séquence est retournée dans le sens 5'-3'.

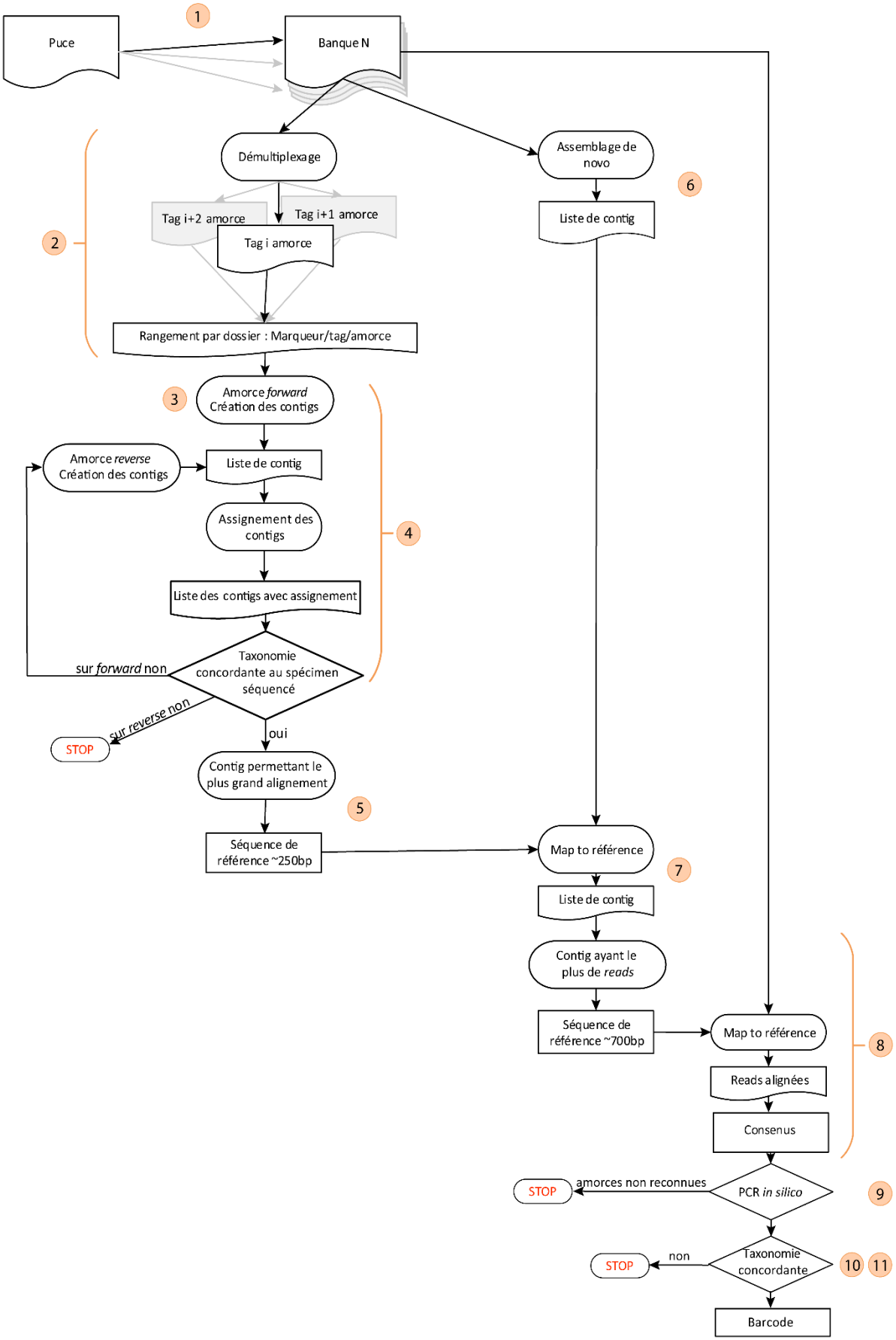


Figure 3.1 : Stratégie employée pour la reconstitution des séquences barcodes après séquençage NGS multiplexé. Les numéros correspondent aux étapes détaillées dans le texte ci-dessus.

10. La séquence barcode est ensuite assignée à l'aide des bases de référence GenBank et BOLD, et si la taxonomie obtenue avec les deux bases de références est congruente, la séquence barcode est considérée comme valide.
11. Une dernière étape de vérification est effectuée manuellement pour évaluer la concordance de la taxonomie associée avec celle du spécimen séquencé (jusqu'alors basée sur la morphologie).

2.3. La construction des bases de référence

GenBank est le principal espace de stockage de séquences ADN et ARN où l'on peut trouver des séquences barcode pour différents marqueurs moléculaires comme le 18S et le COI. De nombreux barcodes sont également disponibles sur SILVA pour le 18S et la plateforme BOLD pour le COI (cf. Chapitre 2). Cependant, étant donné l'assemblage collaboratif de ces bases, la présence d'erreurs est inévitable et nécessite l'implémentation d'étapes de curation et de mise en forme qui peuvent être fastidieuses **(Meiklejohn et al. 2019)**.

Pour l'assignement des séquences COI, trois bases de référence avaient été pressenties au début de ce projet : Genbank pour la quantité de séquences recensées, mais contenant beaucoup d'erreurs, BOLD pour la qualité des séquences répertoriées (moins de séquences erronées que Genbank) et MIDORI car étant une base de référence métazoaire COI contenant des séquences uniques et nettoyées **(Machida et al. 2017)**. Les données de ces différentes bases sont totalement (MIDORI avec GenBank) ou partiellement (GenBank et BOLD) partagées. Depuis, MIDORI a été amélioré en MIDORI2 pour intégrer l'ensemble des séquences eucaryotes ainsi que celles n'ayant pas d'identification jusqu'au rang de l'espèce **(Leray et al. 2022)**. De plus, MIDORI2 est mis à jour à chaque version de Genbank, ce qui n'était pas le cas avec MIDORI, et a donc conforté le choix de l'utiliser comme base de référence. Concernant l'assignement des séquences 18S, la base de données SILVA SSU (16S/18S) a été retenue **(Yilmaz et al. 2014)**.

Un travail conséquent de reformatage, *via* des scripts bash, a été nécessaire pour intégrer ces bases de référence au pipeline de QIIME2. Toutefois, au vu des difficultés de l'intégration des différentes bases de données, Robeson II et collaborateurs **(2021)** ont développé le plugin QIIME2 RESCRIPt (REference Sequence annotation and CuRatIon Pipeline) en 2021. RESCRIPt permet d'effectuer diverses opérations de gestion et de curation des bases de données de référence, que cela soit en termes de séquences ADN/ARN ou de données taxonomiques. De plus, la nouvelle version de MIDORI, MIDORI2, est disponible directement au format employé par QIIME2 (*.qza). Cependant, MIDORI2 n'est pas entièrement exempt d'erreurs, comme par exemple le décalage des rangs taxonomiques pour la taxonomie associée à certaines séquences, qui peuvent conduire au

non-assignement de certains OTU avec la méthode LCA (cf. Chapitre 2). Par ailleurs, même si l'intégration des bases de données publiques a été facilitée ces dernières années, les difficultés rencontrées restent présentes lors de l'intégration de séquences produites localement.

3. Résultats

3.1. La diversité collectée par les ARMS

À l'aide de sept à huit personnes, la relève des 39 ARMS effectuée au cours de la thèse a permis de trier, photographier et informatiser 2 989 échantillons. En incluant les ARMS posés en 2014, le nombre total d'organismes échantillonnés s'élève à 4 584 spécimens, dont 3 843 spécimens à La Réunion et 741 spécimens à Rodrigues (Figure 3.2 ; Figure 3.3). Six phylums principaux ont été retrouvés. Les arthropodes étaient les plus abondants avec 1 659 spécimens (36,2 %) dont 766 crevettes caridées (16,7 %) et 461 brachyures (10,1 %), suivi par les mollusques (N=897), majoritairement composés de gastéropodes (N = 525), les annélides (N=538), les ophiures (échinoderme ; N=302) et les chordés et les porifères (Figure 3.4).

La majorité des spécimens a été classée en 1 291 morpho-espèces, qui regroupent de un à 61 spécimens. Parmi les 6 phylums principaux, ce sont les mollusques qui présentent la plus grande diversité en morpho-espèces avec 308 morpho-espèces, suivi par les arthropodes avec 292 morpho-espèces alors que ce phylum présentait le plus de spécimens (Figure 3.5). Les ARMS relevés à Rodrigues ont permis de répertorier en moyenne 92 ± 13 (écart-type) spécimens par ARMS contre 84 ± 29 spécimens pour La Réunion.

Figure 3.2 : Aperçu de la diversité taxonomique mobile échantillonnée à l'aide des ARMS déployés dans les Mascareignes. Annelida (a et b) ; Arthropoda : Paguroidea (c), Brachyura (d), Caridea (e), Galathea (f) et Pycnogonida (g); Chordata Actinopterygii (h) ; Echinodermata : Asteroidea (i), Holothuroidea (j), Ophiuroidea (k) et Echinoidea (l) ; Mollusca : Bivalvia (m), Cephalopoda (n), Gasteropoda (o) et Polyplacophora (p) ; Nematoda (q) ; Nemerta (r) ; Platyhelminthes (s) et Sipuncula (t). Photographies de H. Bruggemann.

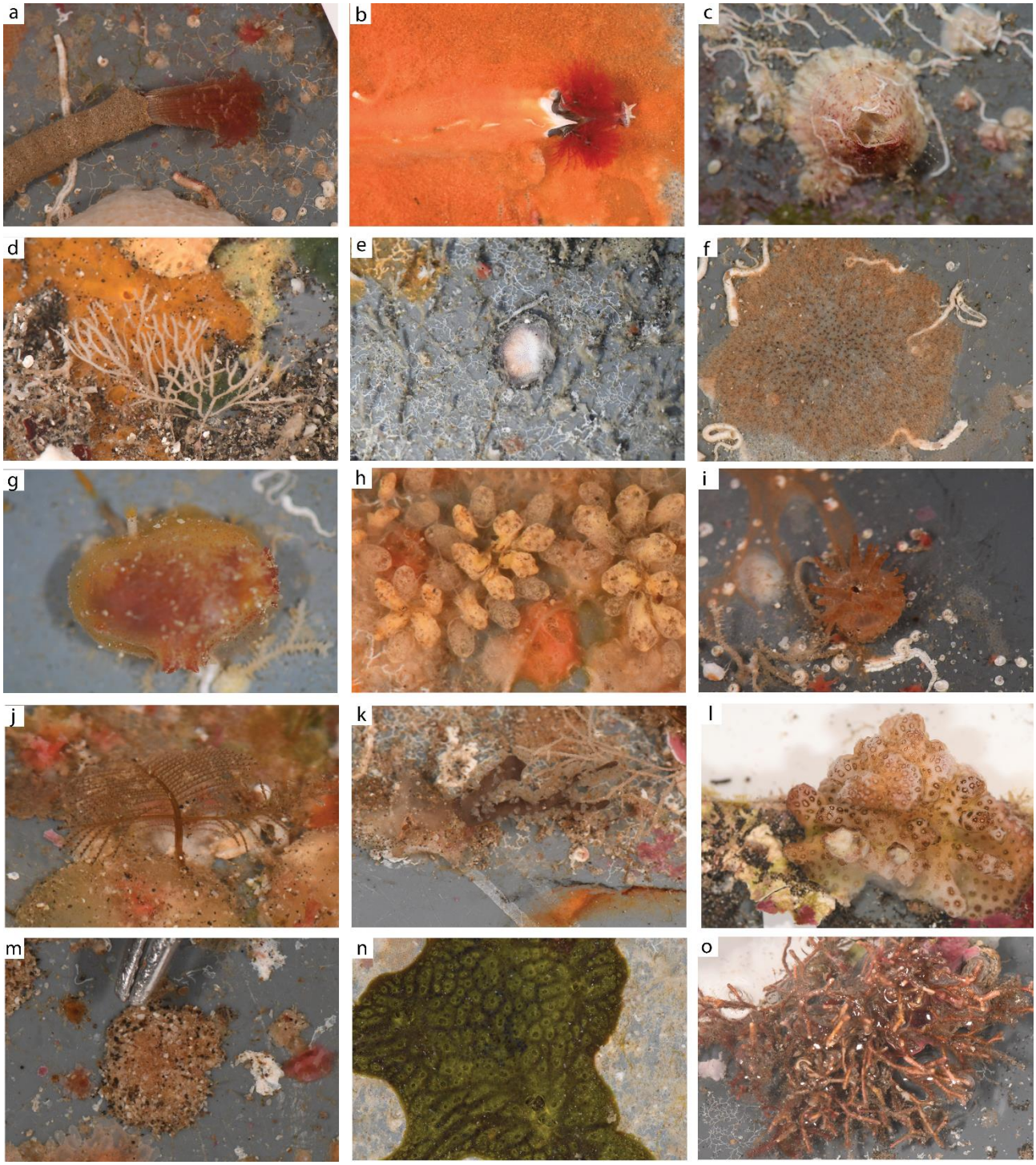


Figure 3.3 : Aperçu de la diversité taxonomique sessile échantillonnée à l'aide des ARMS déployés dans les Mascareignes. Annelida Sabellidae (a) et Serpulidae (b) ; Arthropoda Cirripedia (c) ; Bryozoa (d, e et f), Chordata Ascidiacea : solitaire (g) et coloniale (h) ; Cnidaria Actinaria (i), Hydrozoa (j), Octocorallia (k) et Scleractinia (l) ; Foraminifera (m) ; Porifera (n) et Rhodophyta (o) Photographies de H. Bruggemann.

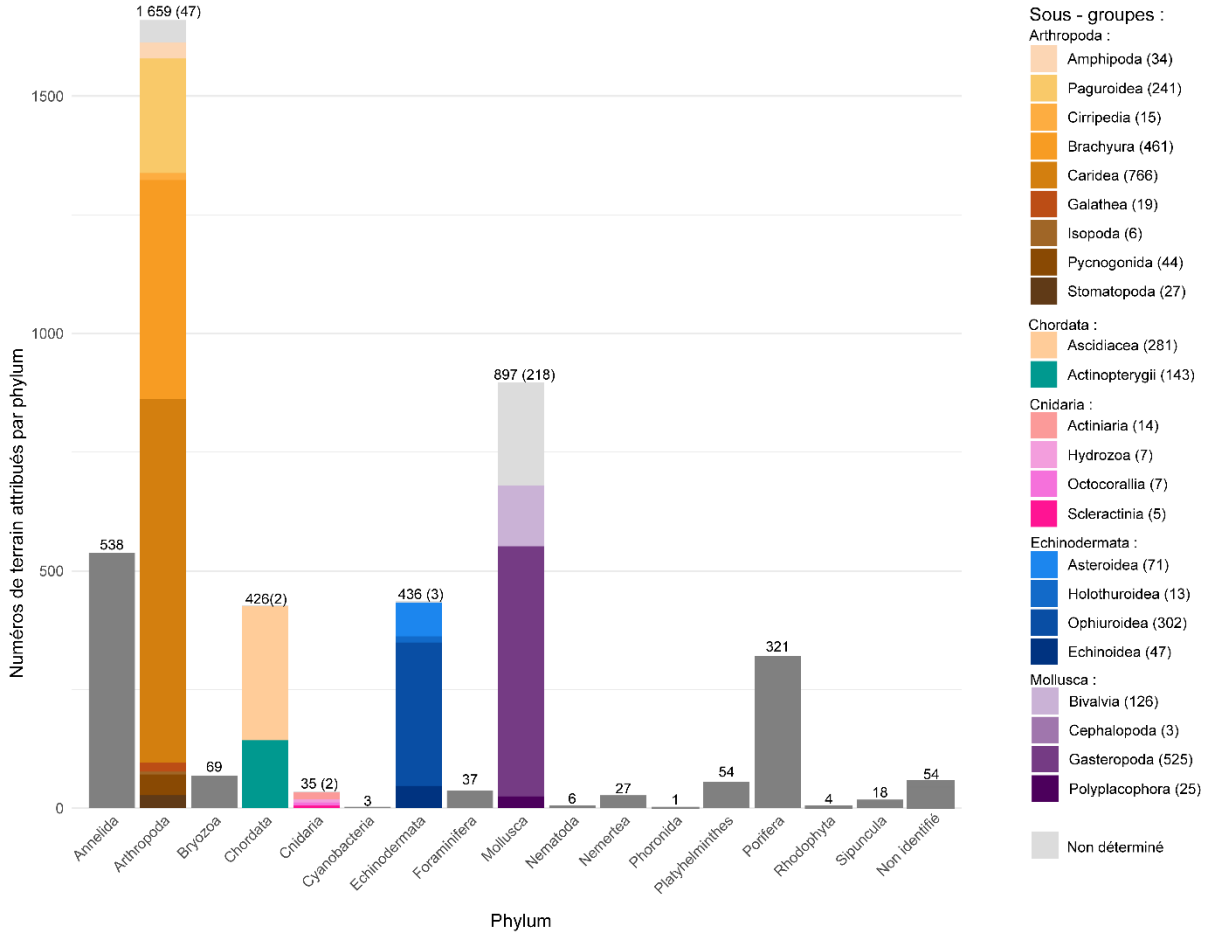


Figure 3.4 : Numéros de terrain attribués (N=4 584) lors de la collecte de l'ensemble des ARMS déployés dans les Mascareignes. Certains phylums ont été subdivisés en sous-groupes (ex. Chordata en Ascidiacea et Actinopterygii). Le nombre d'identifiants uniques par sous-groupe est indiqué entre parenthèses dans la légende. Le nombre total d'identifiants uniques attribués par phylum est indiqué en haut de chaque barre de l'histogramme, avec entre parenthèses le nombre de spécimens n'ayant pas de sous-groupe déterminés (représenté en gris clair).

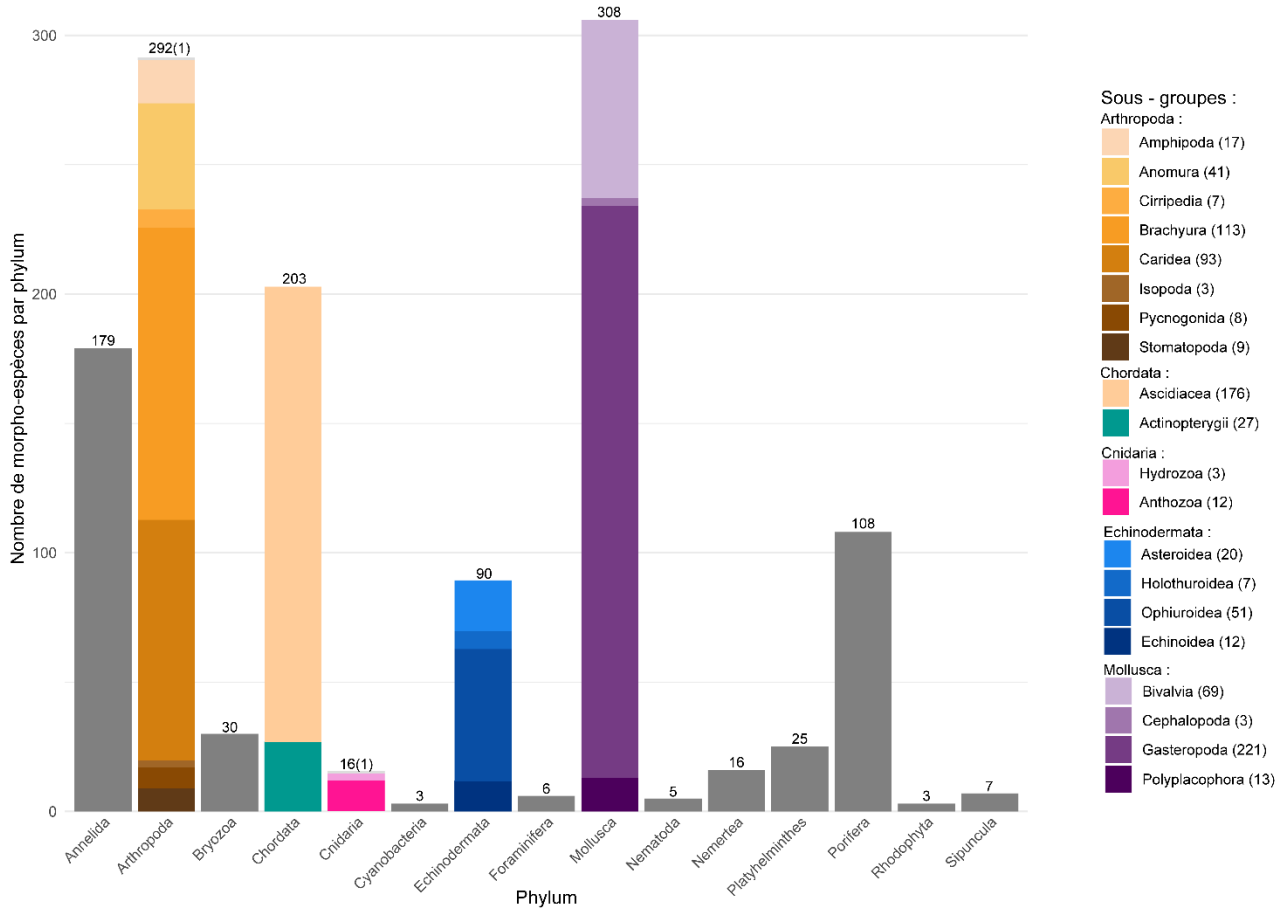


Figure 3.5 : Nombre de morpho-espèces individualisées lors de la collecte de l'ensemble des ARMS déployés dans les Mascareignes. Certains phylums ont été subdivisés en sous-groupe (ex. Chordata en Ascidiacea et Actinopterygii). Le nombre de morpho-espèces par sous-groupe est indiqué entre parenthèses dans la légende. Le nombre total de morpho-espèces par phylum est indiqué en haut de chaque barre de l'histogramme, avec entre parenthèses le nombre de morpho-espèces n'ayant pas de sous-groupe déterminé (représenté en gris clair).

3.2. Les séquences produites et les bases de référence utilisées

Pendant cette étude doctorale, l'ADN de 1 441 spécimens (31,4 %) ont été extraits s'ajoutant aux 198 spécimens dont l'ADN a été extrait lors de la campagne de 2014. Environ 80 % des spécimens extraits ont été séquencés pendant la thèse et 50 % ont été traités bio-informatiquement pour reconstituer les séquences. Ainsi, ces travaux ont permis d'acquérir 490 nouvelles séquences barcodes, 190 pour le 18S, 200 pour le COI, s'ajoutant aux 196 produites lors de la campagne de 2014, et 100 pour le 16S (Tableau 3.4). Les séquences ont été analysées afin de déterminer le nombre de séquences uniques produites, soit 164 pour le 18S, 328 pour le COI et 88 pour le 16S (Tableau 3.4).

Ces séquences uniques ont ensuite été ajoutées aux bases de référence internationales. Pour le 18S, la base de référence était constituée de 441 792 séquences provenant de SILVA 138.1 **(Yilmaz et al. 2014)** et de 164 séquences locales ; pour le COI elle était constituée de 788 530 séquences provenant de MIDORI2 (GenBank version 250 ; **Leray et al. 2022**) et de 328 séquences locales. Dans un souci de reproductibilité et pour s'inscrire dans les principes de la science ouverte, les deux jeux de données de référence employés seront disponibles dès la publication des articles.

Tableau 3.4 : Nombre de séquences barcode produites à partir des ARMS déployés dans les Mascareignes. Pour le COI, entre parenthèses le nombre de séquences produites au cours de cette thèse.

Marqueur	# séquences disponibles avant ces travaux	# séquences uniques avant ces travaux	# séquences disponibles après ces travaux	# séquences uniques après ces travaux
COI	196	156	396 (200)	328 (172)
18S	0	0	190	164
16S	0	0	100	88

3.3. L'amélioration des assignements des OTU à l'aide d'un référentiel local

Pour les deux marqueurs étudiés (18S et COI), la création d'un référentiel barcode local a permis d'améliorer l'assignement des OTU obtenus lors des analyses métabarcoding des ARMS. Pour le 18S, le nombre d'OTU obtenus après traitement bio-informatique s'élevait à 6 203 OTU. Par la suite, la base de données SILVA a permis d'identifier 2 645 OTU au moins au rang taxonomique du Domaine (en anglais *Kingdom*) ; l'ajout de 164 séquences locales a permis d'identifier 19 (0,7 %) OTU supplémentaires (Tableau 3.5). En ce qui concerne la diversité des eucaryotes, la base de données SILVA a permis d'identifier 1 348 OTU et les séquences locales ont permis d'identifier 25 (1,8 %) OTU supplémentaires. Au total, 61 OTU ont été assignés à 99 % de similarité à des séquences locales, soit 8,8 % des OTU eucaryotes assignés à 99 %. Pour le COI, le nombre d'OTU obtenus après traitement bio-informatique s'élevait à 7 701 OTU. La base de données MIDORI2 a permis d'identifier 629 OTU au moins au rang taxonomique du Domaine, et l'ajout de 328 séquences locales a permis d'identifier 123 (1,6 %) OTU supplémentaires (Tableau 3.5). Sur l'ensemble des OTU assignés à 99 % (N = 379), 103 (27,2 %) OTU ont été assignés à des séquences locales.

Tableau 3.5 : Nombre et pourcentage d'OTU totaux assignés au moins au rang taxonomique du Domaine et du Phylum en fonction des séquences de référence employées et nombre d'OTU similaires à 99% aux séquences produites localement

Base de référence	# séquences locales	# total OTU	Assignés au Domaine	Assignés au Phylum	Assignés aux séquences locales
MIDORI2	0	7 701	629 (8.17%)	625 (8.12%)	0
MIDORI2 + locale	328 (0.04%)	7 701	752 (9.76%)	747 (9.7%)	103
Apport séquences locales	-	-	123	122	103
SILVA	0	6 203	2 645(42.6%)	1 848 (29.8%)	0
SILVA + locale	164 (0.04%)	6 203	2 664 (42.9%)	1 871 (30.2%)	61
Apport séquences locales	-	-	19	23	61
SILVA assigné aux Eukaryota	0	1 348	1 348 (21.7%)	1 215 (19.6%)	0
SILVA + locale assigné aux Eukaryota	164 (0.04%)	1 373	1 373 (22.1%)	1 238 (20.0%)	61
Apport séquences locales	-	-	25	23	61

3.4. La diversité retrouvée par métabarcoding

L'utilisation des bases de référence actuelles n'a permis d'assigner qu'une partie des OTU produits par métabarcoding. Pour le 18S, 42,9 % des OTU (N=2 664) ont été assignés au Domaine, pour le COI cette proportion n'est que de 9,76 % (N=752). Le métabarcoding a permis de détecter 29 phylums dont 16 phylums métazoaires (Tableau 3.1). Les phylums majoritairement retrouvés étaient similaires entre les deux marqueurs utilisés, à savoir Annelida, Arthropoda, Mollusca et Porifera, mais avec des proportions différentes (Figure 3.6 et Figure 3.7). Pour le 18S, la majorité des OTU a été assignée aux Annelida (N=189), Porifera (N=147), Rhodophyta (N=144), Arthropoda (N=134) et Mollusca (N=123 ; Figure 3.6), alors que pour le COI, la majorité des OTU a été assignée aux Porifera (N=201), Mollusca (N=118), Arthropoda (N=109) et Annelida (N=68 ; Figure 3.7).

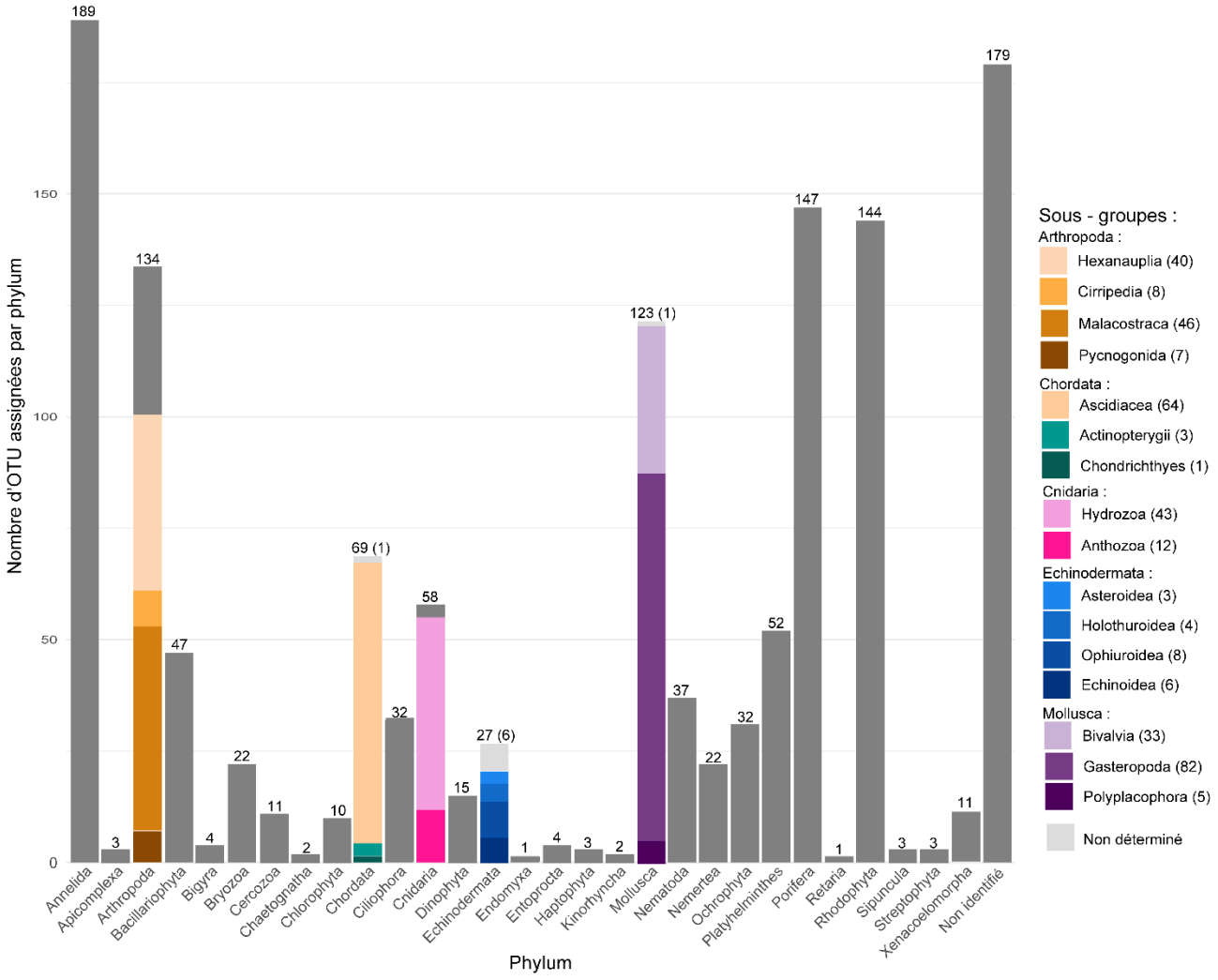


Figure 3.6 : Nombre d'OTU assignées par phylum eucaryote pour les analyses métabarcoding avec le 18S

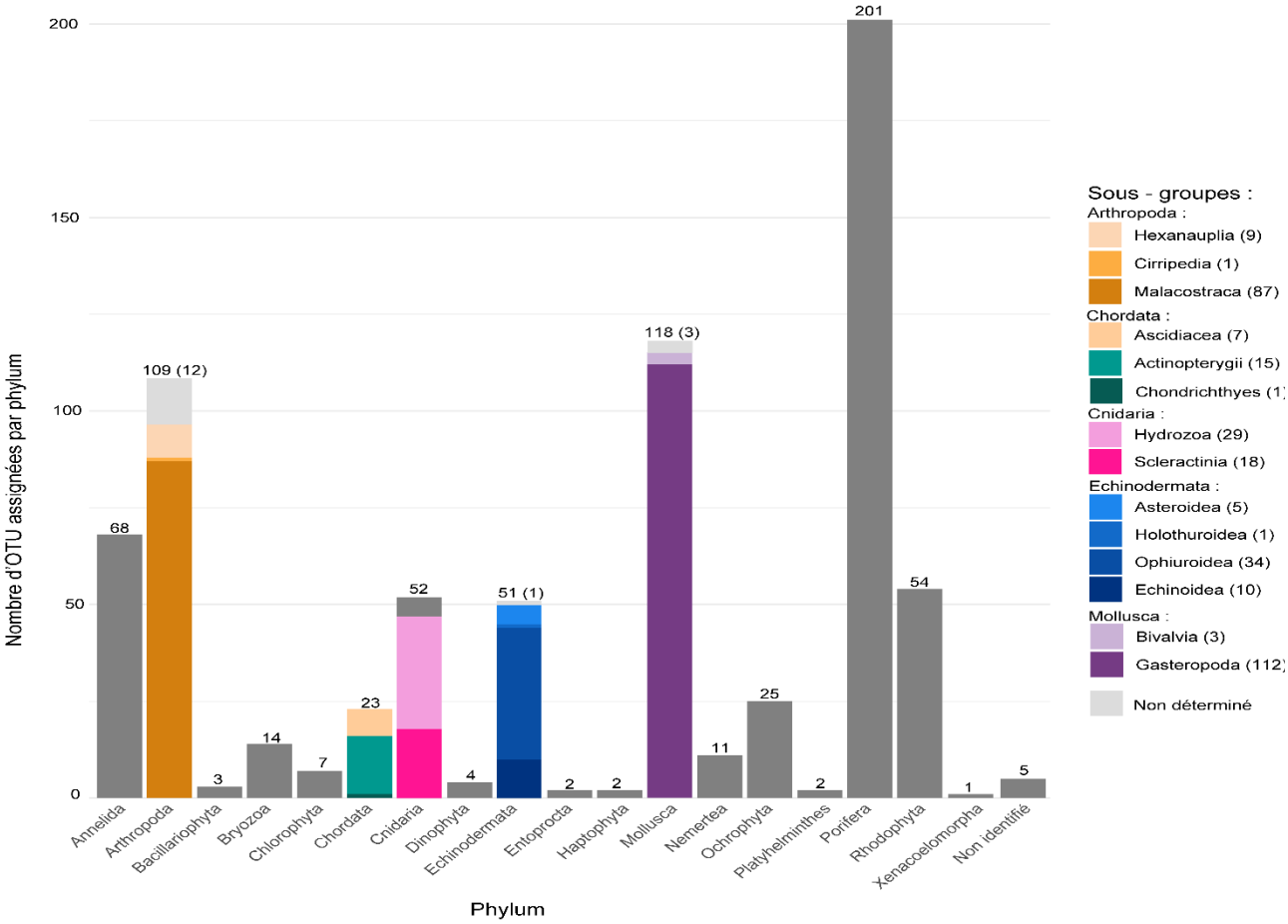


Figure 3.7 : Nombre d'OTU assignées par phylum eucaryote pour les analyses métabarcoding avec le COI

4. Discussion et perspectives

Ces travaux ont permis d'initier la construction d'une base de référence moléculaire du cryptobiome des Mascareignes. La récolte de 54 ARMS à La Réunion et à Rodrigues a permis d'effectuer un échantillonnage conséquent du cryptobiome, de bancariser 4 584 spécimens et d'acquérir 490 nouvelles séquences de référence. Les spécimens collectés et les séquences pourront être utilisés dans de futures études génétiques (ex. étude de phylogéographie comme dans le Chapitre 5) et/ou morphologiques. De plus, l'établissement du protocole moléculaire et du pipeline bio-informatique pour l'acquisition des séquences permet une utilisation en routine et ainsi de continuer à consolider le référentiel local. Lors de cette étude, seule une partie de la diversité échantillonnée a pu être retranscrite en séquences de référence utilisables. Toutefois, l'ajout de ces séquences a permis d'améliorer nettement l'assignement de la diversité trouvée par métabarcoding. L'ajout des 164 séquences de 18S à la base de données SILVA et des 328 séquences de COI à MIDORI2 reste négligeable et ne représente que 0,04 % de leur taille initiale. Néanmoins, ces ajouts ont permis d'augmenter le nombre d'OTU eucaryotes assignés de 25 (1,8 %) pour le 18S

et de 123 (16,4 %) pour le COI. Ces résultats confirment la pertinence d'inclure des séquences locales dans les bases de référence et montrent une amélioration du nombre d'OTU assignés pour les deux marqueurs. Malgré les efforts déployés pour créer une base adaptée au Sud-Ouest de l'océan Indien et l'utilisation de différentes méthodes d'assignement, seule une petite fraction du cryptobiome a pu être identifiée au Domaine (18S : 42.9 % ; COI : 9,76 %). Ces proportions sont inférieures à celles trouvées dans les autres études du cryptobiome. Par exemple, Ip et collaborateurs **(2022)** ont réussi à assigner 100 % de leurs OTU en COI. Cependant, notre processus d'assignement était intentionnellement plus strict pour limiter les identifications taxonomiques erronées (*cf.* Chapitre 2 ; seuil de 80 % de similarité dans Ip et al. 2022 contre 95 % dans la présente étude).

La diversité retrouvée au sein des ARMS déployés sur les pentes externes récifales des Mascareignes concorde avec la grande diversité taxonomique observée dans les récifs coralliens. Nous avons pu échantillonner 17 phylums pour les organismes mobiles de taille supérieure à 2 mm et 29 phylums par les analyses de métabarcoding. Les principaux taxons correspondent à ceux observés dans les ARMS pour d'autres régions comme en Mer Rouge **(Carvalho et al. 2019 ; Villalobos et al. 2022)**, mais également aux taxons observés dans le cryptobiome corallien collecté par d'autres méthodes : substrats artificiels **(Enochs et al. 2011)**, colonie corallienne morte **(Enochs & Manzello 2012 ; Pisapia et al. 2020)**, ou encore débris coralliens **(Moran & Reaka-Kudla 1991)**.

La majorité des spécimens collectés, des morpho-espèces associées, et des OTU trouvés appartenait aux arthropodes, annélides, mollusques, et porifères, ce qui concorde avec les estimations de la biodiversité du cryptobiome mobile des récifs coralliens **(Enochs 2012)**. Les échinodermes, les chordés (principalement des ascidies) et les bryozoaires sont également largement retrouvés, en accord avec les précédentes observations de la diversité du cryptobiome récifal **(Enochs 2012 ; Enochs & Manzello 2012)**. Les proportions relatives des différents phylums changent en fonction de l'approche employée (c'est-à-dire, classement par morpho-espèces, métabarcoding en 18S ou en COI), mais ces différences peuvent s'expliquer par au moins les deux facteurs suivants : (1) les communautés du cryptobiome échantillonnées ne sont pas les mêmes, avec d'un côté pour les morpho-espèces, des organismes supérieurs à 2 mm, majoritairement mobiles en raison de la méthode d'échantillonnage qui repose principalement sur la filtration de l'eau récupérée lors du démantèlement de l'ARMS, et d'un autre côté, pour les analyses métabarcoding de organismes mobiles <2 mm et sessile, (2) les assignements sont potentiellement biaisés en raison de l'incomplétude des bases de référence. En effet, le taux de substitution des acides nucléiques étant plus grand chez les Arthropoda que chez les Porifera **(Huang et al. 2008 ; Gissi et al. 2008)**, la probabilité d'avoir des références génétiquement proches dans les bases de

référence pour les Arthropoda en est réduite, et par conséquent le nombre d'OTU assignés aux Arthropoda l'est aussi. Par exemple, dans notre étude, 134 OTU ont été assignés aux Arthropoda pour le 18S, contre seulement 109 pour le COI. Cependant, pour ce taxon, le COI fournit une meilleure résolution taxonomique que le 18S **(Yu et al. 2012)**. Dès lors, avec un référentiel complet, on devrait retrouver au moins autant, voire plus d'OTU assignés avec le COI qu'avec le 18S, comme chez les Porifera (147 et 201 OTU assignés avec le 18S et le COI, respectivement).

Ces résultats soulignent le caractère unique du cryptobiome des Mascareignes et la nécessité de poursuivre les efforts pour compléter la base de référence locale. En outre, il est vraisemblable que l'augmentation du nombre de séquences de référence pour les crustacés récoltés localement améliorera significativement la diversité des arthropodes détectée.

5. Références du chapitre 3

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden TL. (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421., DOI: 10.1186/1471-2105-10-421

Carvalho S., Aylagas E., Villalobos R., Kattan Y., Berumen M., Pearman JK. (2019) Beyond the visual: using metabarcoding to characterize the hidden reef cryptobiome. Proc R Soc B Biol Sci 286:20182697., DOI: 10.1098/rspb.2018.2697

Cristescu ME., Hebert PDN. (2018) Uses and Misuses of Environmental DNA in Biodiversity Science and Conservation. Annu Rev Ecol Evol Syst 49:209–230., DOI: 10.1146/annurev-ecolsys-110617-062306

Duarte S., Leite BR., Feio MJ., Costa FO., Filipe AF. (2021) Integration of DNA-Based Approaches in Aquatic Ecological Assessment Using Benthic Macroinvertebrates. Water 13:331., DOI: 10.3390/w13030331

Enochs IC. (2012) Motile cryptofauna associated with live and dead coral substrates: implications for coral mortality and framework erosion. Mar Biol 159:709–722., DOI: 10.1007/s00227-011-1848-7

Enochs IC., Manzello DP. (2012) Species richness of motile cryptofauna across a gradient of reef framework erosion. Coral Reefs 31:653–661., DOI: 10.1007/s00338-012-0886-z

Enochs IC., Toth LT., Brandtneris VW., Afflerbach JC., Manzello DP. (2011) Environmental determinants of motile cryptofauna on an eastern Pacific coral reef. Mar Ecol Prog Ser 438:105–118., DOI: 10.3354/meps09259

Geller J., Meyer C., Parker M., Hawk H. (2013) Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. Mol Ecol Resour 13:851–861., DOI: 10.1111/1755-0998.12138

Gissi C., Iannelli F., Pesole G. (2008) Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. Heredity 101:301–320., DOI: 10.1038/hdy.2008.62

Heimeier D., Lavery S., Sewell MA. (2010) Using DNA barcoding and phylogenetics to identify Antarctic invertebrate larvae: Lessons from a large scale study. Mar Genomics 3:165–177., DOI: 10.1016/j.margen.2010.09.004

Hinsinger DD., Debruyne R., Thomas M., Denys GPJ., Mennesson M., Utage J., Dettai A. (2015) Fishing for barcodes in the Torrent: from COI to complete mitogenomes on NGS platforms. DNA Barcodes 3:170–186., DOI: 10.1515/dna-2015-0019

Huang D., Meier R., Todd P., Chou L. (2008) Slow Mitochondrial COI Sequence Evolution at the Base of the Metazoan Tree and Its Implications for DNA Barcoding. J Mol Evol 66:167–74., DOI: 10.1007/s00239-008-9069-5

Ip YCA., Chang JJM., Oh RM., Quek ZBR., Chan YKS., Bauman AG., Huang D. (2022) Seq' and ARMS shall find: DNA (meta)barcoding of Autonomous Reef Monitoring Structures across the tree of life uncovers hidden cryptobiome of tropical urban coral reefs. Mol Ecol:1–20., DOI: 10.1111/mec.16568

Leray M., Knowlton N., Machida RJ. (2022) MIDORI2: A collection of quality controlled, preformatted, and regularly updated reference databases for taxonomic assignment of eukaryotic mitochondrial sequences. Environ DNA:edn3.303., DOI: 10.1002/edn3.303

Machida RJ., Knowlton N. (2012) PCR Primers for Metazoan Nuclear 18S and 28S Ribosomal DNA Sequences. PLoS ONE 7:e46180., DOI: 10.1371/journal.pone.0046180

Machida RJ., Leray M., Ho S-L., Knowlton N. (2017) Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. Sci Data 4:1–7., DOI: 10.1038/sdata.2017.27

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

Monchamp M-E., Taranu Z., Garner R., Re T., Morissette O., Iversen L., Fugère V., Littlefair J., Barbosa da Costa N., Desforges J., Schacht J., Derry A., Cooke S., Barrett R., Walsh D., Ragoussis J., Albert M., Cristescu M., Gregory-Eaves I. (2023) Prioritizing taxa for genetic reference database development to advance inland water conservation. Biol Conserv 280., DOI: 10.1016/j.biocon.2023.109963

Moran DP., Reaka-Kudla ML. (1991) Effects of disturbance: disruption and enhancement of coral reef cryptofaunal populations by hurricanes. Coral Reefs 9:215–224., DOI: 10.1007/BF00290425

Mugnai F., Costantini F., Chenuil A., Leduc M., Ortega JMG., Meglécz E. (2023) Be positive: customized reference databases and new, local barcodes balance false taxonomic assignments in metabarcoding studies. PeerJ 11:e14616., DOI: 10.7717/peerj.14616

Mugnai F., Meglécz E., Abbiati M., Bavestrello G., Bertasi F., Bo M., Capa M., Chenuil A., Colangelo MA., De Clerck O., Gutiérrez JM., Lattanzi L., Leduc M., Martin D., Matterson KO., Mikac B., Plaisance L., Ponti M., Riesgo A., Rossi V., Turicchia E., Waeschenbach A., Wangensteen OS., Costantini F. (2021) Are well-studied marine biodiversity hotspots still blackspots for animal barcoding? Glob Ecol Conserv 32:e01909., DOI: 10.1016/j.gecco.2021.e01909

Palumbi SR., Martin A., Romano S., McMillan EO., Stice L., Grabowski G. (1991) The simple fool's guide to PCR, Version 2.0. Dept. of Zoology and Kewalo Marine Laboratory, University of Hawaii, Honolulu, HI., 45 p.

Paulay G. (1997) Diversity and Distribution of Reef Organisms. In: *Life and Death of Coral Reefs*. Birkeland C (ed) Springer US, Boston, MA, p 298–353, DOI: 10.1007/978-1-4615-5995-5_14

Pisapia C., Stella J., Silbiger NJ., Carpenter R. (2020) Epifaunal invertebrate assemblages associated with branching Pocilloporids in Moorea, French Polynesia. PeerJ 8:e9364., DOI: 10.7717/peerj.9364

Ransome E., Geller JB., Timmers M., Leray M., Mahardini A., Sembiring A., Collins AG., Meyer CP. (2017) The importance of standardization for biodiversity comparisons: A case study using autonomous reef monitoring structures (ARMS) and metabarcoding to measure cryptic diversity on Mo'orea coral reefs, French Polynesia. PLOS ONE 12:e0175066., DOI: 10.1371/journal.pone.0175066

Rimet F., Aylagas E., Borja Á., Bouchez A., Canino A., Chauvin C., Chonova T., Ciampor Jr F., Costa FO., Ferrari BJD., Gastineau R., Goulon C., Gugger M., Holzmann M., Jahn R., Kahlert M., Kusber W-H., Laplace-Treyture C., Leese F., Leliaert F., Mann DG., Marchand F., Méléder V., Pawlowski J., Rasconi S., Rivera S., Rougerie R., Schweizer M., Trobajo R., Vasselon V., Vivien R., Weigand A., Witkowski A., Zimmermann J., Ekrem T. (2021) Metadata standards and practical guidelines for specimen and DNA curation when building barcode reference libraries for aquatic life. Metabarcoding Metagenomics 5:e58056., DOI: 10.3897/mbmg.5.58056

Robeson II MS., O'Rourke DR., Kaehler BD., Ziemski M., Dillon MR., Foster JT., Bokulich NA. (2021) RESCRIPt: Reproducible sequence taxonomy reference database management. PLOS Comput Biol 17:e1009581., DOI: 10.1371/journal.pcbi.1009581

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

Stefaniak L., Lambert G., Gittenberger A., Zhang H., Lin S., Whitlatch R. (2009) Genetic conspecificity of the worldwide populations of Didemnum vexillum Kott, 2002. Aquat Invasions 4.

Villalobos R., Aylagas E., Pearman J., Cúrdia J., Lozano-Cortés D., Coker D., Jones B., Berumen M., Carvalho S. (2022) Inter-annual variability patterns of reef cryptobiota in the central Red Sea across a shelf gradient. Sci Rep 12., DOI: 10.1038/s41598-022-21304-2

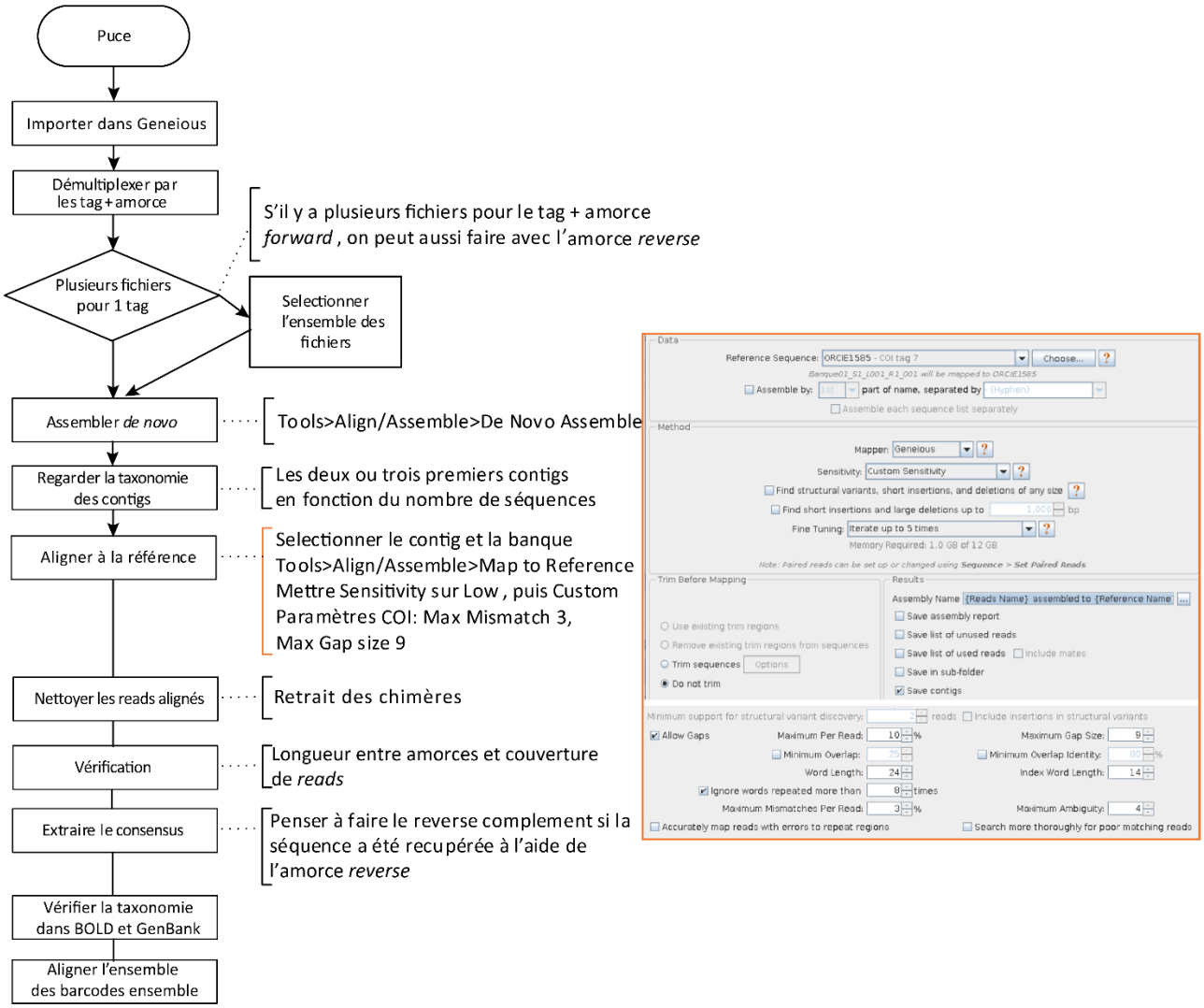
Yilmaz P., Parfrey LW., Yarza P., Gerken J., Pruesse E., Quast C., Schweer T., Peplies J., Ludwig W., Glöckner FO. (2014) The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. Nucleic Acids Res 42:D643–D648., DOI: 10.1093/nar/gkt1209

Yu DW., Ji Y., Emerson BC., Wang X., Ye C., Yang C., Ding Z. (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods Ecol Evol 3:613–623., DOI: 10.1111/j.2041-210X.2012.00198.x

Zinger L., Bonin A., Alsos IG., Bálint M., Bik H., Boyer F., Chariton AA., Creer S., Coissac E., Deagle BE., De Barba M., Dickie IA., Dumbrell AJ., Ficetola GF., Fierer N., Fumagalli L., Gilbert MTP., Jarman S., Jumpponen A., Kauserud H., Orlando L., Pansu J., Pawlowski J., Tedersoo L., Thomsen PF., Willerslev E., Taberlet P. (2019) DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. Mol Ecol 28:1857–1862., DOI: 10.1111/mec.15060

6. Annexes du chapitre 3

Annexe 3.1 : Schéma illustrant la reconstitution manuelle des séquences barcodes après un séquençage NGS avec multiplexage des amplicons sous Geneious Prime



Chapitre 4 : Variabilité temporelle du cryptobiome récifal collecté par les ARMS

Résumé :

Les ARMS sont utilisées dans le monde entier, en particulier sur les récifs coralliens, pour évaluer les communautés cryptobenthiques. Ils ont été développés comme un outil standardisé, mais certains aspects de leur déploiement varient d'une étude à l'autre, tel que le temps d'immersion et la saison au moment du déploiement et de collecte, avec peu d'informations sur l'incidence de ces changements sur les résultats. Dans ce chapitre, nous avons étudié la variabilité temporelle et saisonnière des communautés du cryptobiome récifal échantillonnées à l'aide d'ARMS dans le Sud-Ouest de l'océan Indien. Pour ce faire, 15 ARMS ont été déployés sur la pente externe d'un récif corallien de La Réunion, afin d'étudier les schémas de peuplement et le remplacement temporel des communautés du cryptobiome entre trois durées d'immersion (six mois, un an et deux ans) et deux saisons d'immersion (fraîche et chaude) en employant deux marqueurs moléculaires (18S et COI). Les deux marqueurs ont détecté différents taxons avec des résolutions différentes, mais ont montré des patrons de composition et de structure des communautés similaires. Alors qu'aucune variation du nombre d'OTU dans la communauté (diversité α) n'a été observée entre les différent temps d'immersion et les saisons, les structures et les compositions des communautés ont, quant à elles, été significativement affectées à la fois par le temps d'immersion et la saison. Nos résultats ont mis en évidence une diminution de la similarité des communautés entre les réplicats avec le temps d'immersion, en raison d'un remplacement des OTU plus important aux premiers stades de la colonisation, puis d'une augmentation des différences de richesse en OTU. De plus, seule une petite fraction du cryptobiome semble être stable dans le temps.

Ces résultats font l'objet d'un manuscrit en cours de préparation pour *Scientific Reports*.

Settlement patterns and temporal successions of coral reef cryptic communities: implications for evaluating diversity using Autonomous Reef Monitoring Structures (ARMS)

Marion Couëdel[1], Agnes Dettai[2], Mireille M. M. Guillaume[3,4], Céline Bonillo[3], Baptiste Frattini[3,1], J. Henrich Bruggemann[1,4]

1: Université de La Réunion, UMR 9220 ENTROPIE (Université de La Réunion, IRD, IFREMER, Université de Nouvelle-Calédonie, CNRS), Saint-Denis, La Réunion, 97400, France
2: Muséum national d'Histoire naturelle (MNHN), UMR 7205 ISYEB (MNHN, CNRS, Sorbonne Université, EPHE, Université des Antilles), Paris, 75005, France
3: Muséum national d'Histoire naturelle (MNHN), UMR 8067 BOrEA (MNHN, CNRS 2030, Sorbonne Université, IRD 207, Uni Caen-Normandie, Université des Antilles), Paris, 75005, France
4: LabEx CORAIL, Université de Perpignan, Perpignan, 66860, France

Abstract

Autonomous Reef Monitoring Structures (ARMS) are used worldwide especially on coral reefs to assess cryptic diversity. They were developed as standardised tools, yet conditions of deployment, such as immersion time and/or deployment and collection times, vary between studies, with little information on the incidence of these changes on the results. Here, we studied temporal and seasonal variability in coral reef cryptic communities sampled with ARMS in the southwest Indian Ocean. We deployed 15 ARMS in one site on a coral reef slope at Reunion Island to investigate the settlement patterns and temporal successions of coral reef cryptic communities among three immersion times (six months, one year and two years), two immersion seasons (cool vs. hot) and three fractions (500-2000 µm, 106-500 µm and sessile) using two genetic markers (18S and COI). Both markers detected different taxa with different resolutions, but the patterns of community composition and structure were broadly similar. While no variation in OTUs numbers was observed in communities depending on the immersion time and season, community structures and compositions were significantly affected by both the immersion time and season. Our results evidenced a decrease of the similarity of ARMS communities with immersion time, due to higher turnover in early colonisation stages and then an increase in OTUs richness differences. Thus, only a small fraction of the community appears to be stable over time. Finally, the small fraction of diversity that can be assigned at phylum level highlights the uniqueness of the Mascarene cryptobiome and the need for further sampling and sequencing efforts on these communities.

1. Introduction

In the context of global change, monitoring biodiversity is essential to detect stressors and understand community responses. Accurately quantifying biodiversity is crucial for effective ecosystem management **(Cinner et al. 2020)**, but traditional methods of taxonomy based on morphology are time consuming and require specialized knowledge, particularly for small and cryptic organisms. A powerful alternative to traditional methods is DNA metabarcoding, which involves the extraction of bulk DNA of environmental samples, DNA mass amplification and sequencing using universal genetic markers followed by taxonomic assignation **(Yu et al. 2012 ; Gibson et al. 2014 ; Elbrecht et al. 2017 ; Taberlet et al. 2018)**. This approach is now widely employed to evaluate biodiversity in various contexts **(Thomsen et al. 2012)**, including microbiome studies **(Gibson et al. 2014)**, food web reconstructions **(Leray et al. 2015 ; Albaina et al. 2016)**, and water-based environmental DNA (eDNA) analyses **(Alexander et al. 2019 ; Antich et al. 2020)**.

Standardisation of methods is essential to address specific questions and scale up results from local to global scales. In an effort to minimize sampling biases when estimating cryptic biodiversity, the Autonomous Reef Monitoring Structures (ARMS) were developed in the framework of the Census of Marine Life (CoML; **(Zimmerman & Martin 2004)**). Each ARMS unit consists of a stack of nine PVC plates (22.5 cm × 22.5 cm) that are alternatively separated by 1cm thick spacers and cross-bars, mimicking the structural complexity of reef habitats with different levels of exposure to light and water flow. ARMS have been used in various ways and environments, from tropical and temperate marine ecosystems **(Pennesi & Danovaro 2017 ; David et al. 2019)** to Antarctica (pers. comm. C. Gallut), combining standardised genetic analysis and image processing **(David et al. 2019)**. Some studies focused on specific taxa such as arthropods **(Plaisance et al. 2011 ; Hazeri et al. 2019)**, sponges **(Vicente et al. 2021 ; Steyaert et al. 2022)** or fishes **(Couëdel et al. 2023)**, or included the entire mobile **(Villalobos et al. 2022)** or sessile communities **(Palomino-Alvarez et al. 2021)**. ARMS are usually deployed for two years on site **(Ransome et al. 2017 ; Pearman et al. 2018 ; Carvalho et al. 2019 ; Nichols et al. 2021 ; Villalobos et al. 2022)** but immersion times vary among studies, from six months **(Leray & Knowlton 2015)** to three years **(Plaisance et al. 2011)**. However, no study has systematically investigated the effects of immersion duration on the recovered communities and how it may affect the ecological inferences from such surveys.

The composition and abundance of colonising taxa depend on the presence of source populations, their reproductive cycles and the dispersal capacity of propagules as well as on local hydrodynamics. These circumstances vary with season which may affect the diversity and growth of

settlers **(Ateweberhan et al. 2006 ; Astudillo et al. 2016 ; Larkin et al. 2017)**. Benthic communities show reproductive seasonality synchronised with temperature **(Tanner 1996 ; Gaudron et al. 2008)**, nutrient availability **(Shenkar et al. 2008)** or light intensity **(Glasby 1999 ; Muthiga & Jaccarini 2005)**, shaping recruitment success **(Zea 1993 ; Astudillo et al. 2016)**. Currents, which have a seasonal component **(Matano et al. 2002)**, may affect larval dispersal, especially in organisms with short larval life spans such as sponges **(Zea 1993)** and ascidians **(Shanks et al. 2003 ; Weersing & Toonen 2009)**. Therefore, seasonality represents an important determinant of recruitment and community dynamics and must be considered in ecological surveys **(Astudillo et al. 2016)**.

In this study, we investigated settlement and succession patterns of cryptic coral reef communities after three different immersion times (six months, one year and two years). Considering that seasonal variations in physico-chemical conditions and biological aspects drive the spatio-temporal dynamics of biological communities, we also explored the colonisation patterns for two seasons (hot and cool) in Reunion, a tropical island of the Mascarene archipelago (Southwest Indian Ocean). We hypothesize the alpha diversity increases with immersion time **(Martens et al. 2006)**. Moreover, we presume that short immersion times (6 months) reflect the early colonizers community (thus high similarity among ARMS replicates), while the random arrival of subsequent colonizers drives community succession **(Connell & Slatyer 1977)**, which may lead community composition into different directions (lower similarity among ARMS replicates). Therefore, we expected to observe decreasing similarity among ARMS replicates with increasing immersion times. In contrast, if primary succession **(Connell & Slatyer 1977)** is the main process driving changes in community composition, then a greater similarity between ARMS replicates is expected through time. Finally, numerous reef species have reproductive seasonality, with spawning in hot season, when temperature and light reach their annual maxima. Therefore, we expected to observe a higher similarity among ARMS replicates retrieved in hot season.

2. Material and methods

2.1. Deployment, recovery, and processing of ARMS

Five sets of three replicate ARMS units were deployed on top of a spur at 10-12 m depth on the outer coral reef slope at La Saline (21.10401° S; 55.23598° E) on the West coast of Reunion Island, between 2019 and 2022 (ESM 1). All ARMS units were affixed to the sea bed during SCUBA dives, with distances among ARMS varying from 2 to 8 m. Three sets were deployed during the hot season

and immersed for six months, one year and two years; two sets were deployed during the cool season and immersed for six months and one year (Fig. 1). A temperature logger (HOBO Water Temp Pro v2, ONSET) attached to an ARMS base plate recorded temperatures hourly from 23 October 2019 to 19 February 2021. Before retrieval, ARMS were covered with crates lined with 48 µm mesh to prevent the escape of mobile fauna. During transport and processing in the laboratory, ARMS and plates were kept submerged in 48 µm filtered seawater. Plates were gently brushed to remove motile organisms, photographed on both sides, after which sessile organisms were removed by scraping the plates and homogenized using a kitchen blender. The water that held the ARMS was filtered through 2-mm, 500-µm, and 106-µm sieves. The sessile fraction and the two filtered fractions, 500-2000 µm and 106-500 µm, were rinsed in 106 µm mesh cloth, first with seawater, then with 95% EtOH before being wrung to remove excess liquid. Each sample was preserved in 95% EtOH and conserved at -80°C until DNA extraction. This processing protocol thus allowed to recover three bulk samples per ARMS, one for each fraction (i.e. 500-2000 µm, 106-500 µm and sessile). The mobile organisms >2 mm size were processed separately and will not be included here.

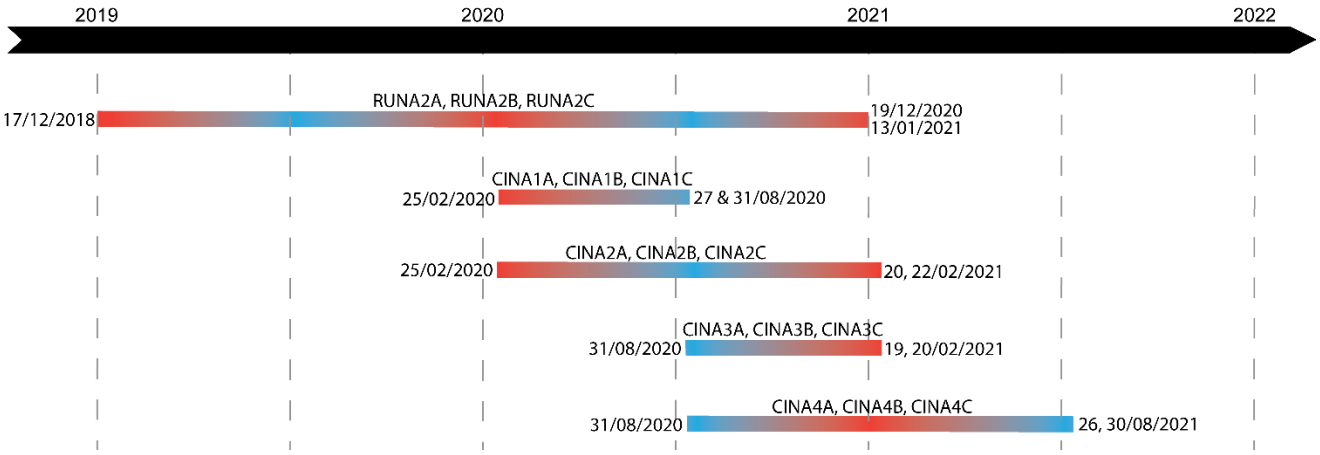


Figure 1: Timeline of deployment and retrieval of ARMS used in this study. The red/blue colour gradient on the bars represents the seasonality and thus the seawater temperature: red for hot season and blue for cool season and purple for inter-seasons

2.2. DNA extraction, amplification and sequencing

DNA was extracted from 10 g of each sample using the DNeasy Powermax Soil DNA isolation kit (Qiagen), with the standardised protocol established by the Smithsonian Global ARMS Program (https://naturalhistory.si.edu/research/global-arms-program/protocols) and following the recommendations of **(Leray & Knowlton 2015)**. Samples were centrifuged at 2,500 rcf for 10 min to discard EtOH. Then, 15 mL of PowerBead Solution was added and vortexed vigorously for 1 min. As recommended by **(Leray & Knowlton 2015)**, 405 µL of proteinase K at 20 mg/mL was added to the solution and samples were incubated in a shaking incubator at 56°C overnight. The DNeasy Powermax Soil protocol was followed for the rest of the extraction procedure. Extracted DNA was purified using the DNeasy PowerClean Pro Cleanup kit from Qiagen before PCR amplification. Extraction controls were made with a negative control for extraction that consisted of 10 ml of DNA-free water subjected to the DNA extraction protocol and a negative control for DNA aerosols consisting of a 10-mL tube containing 10 mL of DNA-free water that remained open but otherwise untouched during the extraction and purification protocols **(Corse et al. 2017)**. Four 12µl replicate PCR assays were performed to amplify a ca. 310-bp COI (Forward: 5'-GGWACWGGWTGAACWGTWTAYCCYCC-3'; Reverse: 5'-TAIACYTCIGGRTGICCRAARAAYCA-3'; **Leray et al. 2013**) and a ca. 550-bp 18S (Forward: 5'-CTGGTGCCAGCAGCCGCGGYAA-3'; Reverse: 5'-TCCGTCAATTYCTTTAAGTT-3'; **Machida & Knowlton 2012**) fragments for each of the 44 samples. The Qiagen Hotstart Multiplex taq was used with the concentrations recommended by the manufacturer, programs were 94°C 15 min, then 30 cycles of 94° 20s, 55° 1min, 72° 1 min for the 18S and 94°C 15 min, then 35 cycles of 94° 20s, 50° 1min, 72° 1 min for the COI after optimisation. In order to pool six samples for each marker per sequencing library, PCR primers were tailed with 6-bp tags **(Leray & Knowlton 2015)**. Forward and reverse primers were both tagged using the same tag for each sample to assess the extent of tag crossing **(Corse et al. 2017)**. Amplicons replicates were pooled and then multiplexed according to their tag and primer. Illumina library preparation and paired end sequencing (PE250) on a Illumina NovaSeq 6000 were performed at the platform iGenSeq (ICM, Paris, France).

2.3. Bioinformatics

Most of the following steps were performed using QIIME2 **(Bolyen et al. 2018)** and implemented functions. Reads were first demultiplexed into individual samples and markers according to their tag and primers using *qiime cutadapt demux-paired*. For the COI, forward (R1) and

reverse (R2) reads were trimmed at 220 bp, denoised and merged with *qiime dada2 denoise-paired*. For 18S, since R1 and R2 were not overlapping (the total length of the amplicon is around 550 bp depending on species), only R1 was used. Reads were trimmed at 185 bp and cleaned with *qiime dada2 denoise-single*. Then, for both markers, reads were merged into 99% similarity OTU (Operational Taxonomic Unit) clusters. Singletons and chimeras were discarded. OTUs were filtered based on their read abundance. The filtration threshold was determined from the OTU read abundance of control samples (18S threshold = 500 reads; COI threshold = 25 reads). OTUs with reads numbers under these thresholds and occurring in less than two samples were discarded. Two supplementary filtering steps were performed with LULU (**Frøslev et al. 2017**; chimeras detection) and decontam with default parameters (**Callahan 2021**; contaminants detection) in R 4.1.1 **(R Core Team 2021)**.

For 18S, taxonomy was assigned against both local and SILVA 138.1 databases using hierarchical steps: 1/ blast against local database of 164 unique sequences at 99% similarity (*qiime feature-classifier classify-consensus-blast*); 2/ LCA with a threshold of 99% similarity against MIDORI; 3/ LCA method with a threshold of 97% similarity against the local and SILVA 138.1 databases merged.

For COI, two datasets were created, one with OTUs grouped at 99% similarity (COI99) and another with OTUs grouped at 97% similarity (COI97). The 97% similarity dataset was reconstructed for the purpose of comparison with other studies. However, ecological analyses were performed on 99% similarity to take advantage of finer resolution, on questions about the relevance of the lower threshold from the scientific community **(Callahan et al. 2017)**. Taxonomy was assigned against both local cryptobiome sequences and the MIDORI V2 (GenBank 250) database, using hierarchical steps to improve the accuracy of assignments: (1) blast against a local database of 328 unique sequences at 99% similarity (*qiime feature-classifier classify-consensus-blast*); (2) identification of the Last Common Ancestor (LCA) with a 99% similarity threshold against MIDORI V2; (3) LCA method with a threshold of 97% similarity against the local and MIDORI V2 merged databases; (4) LCA method with a threshold of 95% similarity against the merged databases. We focused here on the Eukaryota cryptobiome; therefore all OTUs without an assignation to Eukaryota were removed from the analyses. Deployment and sequencing information of the samples included were provided in ESM 2. OTU files and reference databases used will be available at Mcouedel github.

2.4. Data analysis.

In order to examine the variations in community composition related to fractions, analyses were performed on each of the three datasets (i.e., COI99, COI97 and 18S), distinguishing four sub-sets: each fraction (i.e., 500-2000 µm, 106-500 µm and sessile) separately and one where all fractions of the same ARMS were pooled (ARMS level). All further analyses were conducted in R 4.1.1. Four factors were considered for statistical comparisons: immersion time, deployment season, retrieval season, and modalities, which correspond to the combination of all three factors inherent at each deployment batch (3 ARMS replicates). Variation in alpha-diversity and dissimilarity indices among immersion times and seasons were tested using one-way ANOVA when the conditions for parametric tests could be met; otherwise non-parametric Kruskal-Wallis tests (KW) were used. When significant effects were detected, post-hoc Tukey tests with single-step adjustment of probability were conducted. OTU accumulation curves were estimated using the {iNext} package **(Hsieh & Chao 2022)**, and numbers of unique and shared OTUs among immersion times and seasons were visualized with Euler diagrams using the {eulerr} package **(Larsson et al. 2022)**. Moreover, for each sub-dataset, beta-diversity patterns across ARMS and immersion time series were explored by calculating Jaccard-binary dissimilarity indices **(Jaccard 1912)** and running permutational multivariate analyses of variance (PERMANOVA; adonis {vegan}; pairwise.adonis {pairwiseAdonis}). Dissimilarities were visualized using a nonmetric multidimensional scaling (NMDS) through the {phyloseq} and {vegan} packages. The mean contributions of OTUs to the dissimilary between factors were computed by SIMPER analyses in PAST4 **(Hammer et al. 2001)**. OTUs were considered as discriminant when they were involved in 50% of the observed difference. The beta diversity may result from OTUs replacing others across communities (replacement) or due to communities differing in richness (richness difference) **(Legendre 2014)**. Here, beta diversity was expressed as Jaccard similarity and examined with beta.div.comp {adespatial} and plotted with {ggtern}.Composition plots in relative abundance for each time series were plotted using {phyloseq} and {ggplot2} packages.

2.5. Environmental parameters

Hourly *in situ* temperature records were averaged by day. To obtain *in situ* temperature variations for the missing periods of our study (01-12-2018 to 23-10-2019 and 13-02-2021 to 01-09-2021), NOAA SST (Sea Surface Temperature) data over the entire period was downloaded (NOAA virtual station Reunion-Tromelin) and compared to the measured *in situ* temperatures. For this

purpose, a 7-days sliding mean was applied on the *in situ* and SST temperatures over the overlapping period (23-10-2019 to 19-02-2021), and the average difference between both sources of temperature was used to correct the NOAA SST record for our study site. Rainfall and daily global radiation over the study period were retrieved from the Météo France meteorological station of Trois Bassins (located at sea level 2.35 km distance from the study site). Chlorophyll, particulate organic and inorganic carbon concentrations (POC and PIC, respectively) were retrieved from NASA's Oceancolor website (https://oceancolor.gsfc.nasa.gov/) derived from the MODIS A satellites at 4 km resolution. Seasonal trends of all environmental parameters were modelled with gam relations in R (geom_smooth {ggplot2}).

3. Results

3.1. Seasonal variation of environmental parameters

In situ temperatures varied from 29.3 °C in March 2019 to 24.0 °C for August 2020 (Fig. 2). The mean temperature for the warmest month in 2019 (March) and 2021 (February) was 0.4 °C higher (28.7°C) than in 2020 (March; 28.3°C). The mean temperature for the coolest month (August) in 2019 was 0.5 °C higher (24.2°C) than for 2020 and 2021 (23.7°C). Temperature variations highlighted four periods: the hot season from January to April; the cooling months of May and June; the cool season from July to October and the warming months of November and December. The seasonal pattern of daily global radiation preceded and paralleled that of temperature (Fig. 2). In contrast, chlorophyll and POC concentrations suggested an inverted pattern with maximum concentrations during the cool season and minimum concentrations during the hot season (Fig. 2). Monthly precipitations demonstrated maximum rainfall during the hot season and minimum during the cool season (Fig. 2). For the intermediate periods (warming and cooling), rainfall showed inter-annual variations, such as a wet warming season in 2019 contrasting with a dry warming season in 2020. No seasonal variations were observed for wind force and PIC (data not shown).

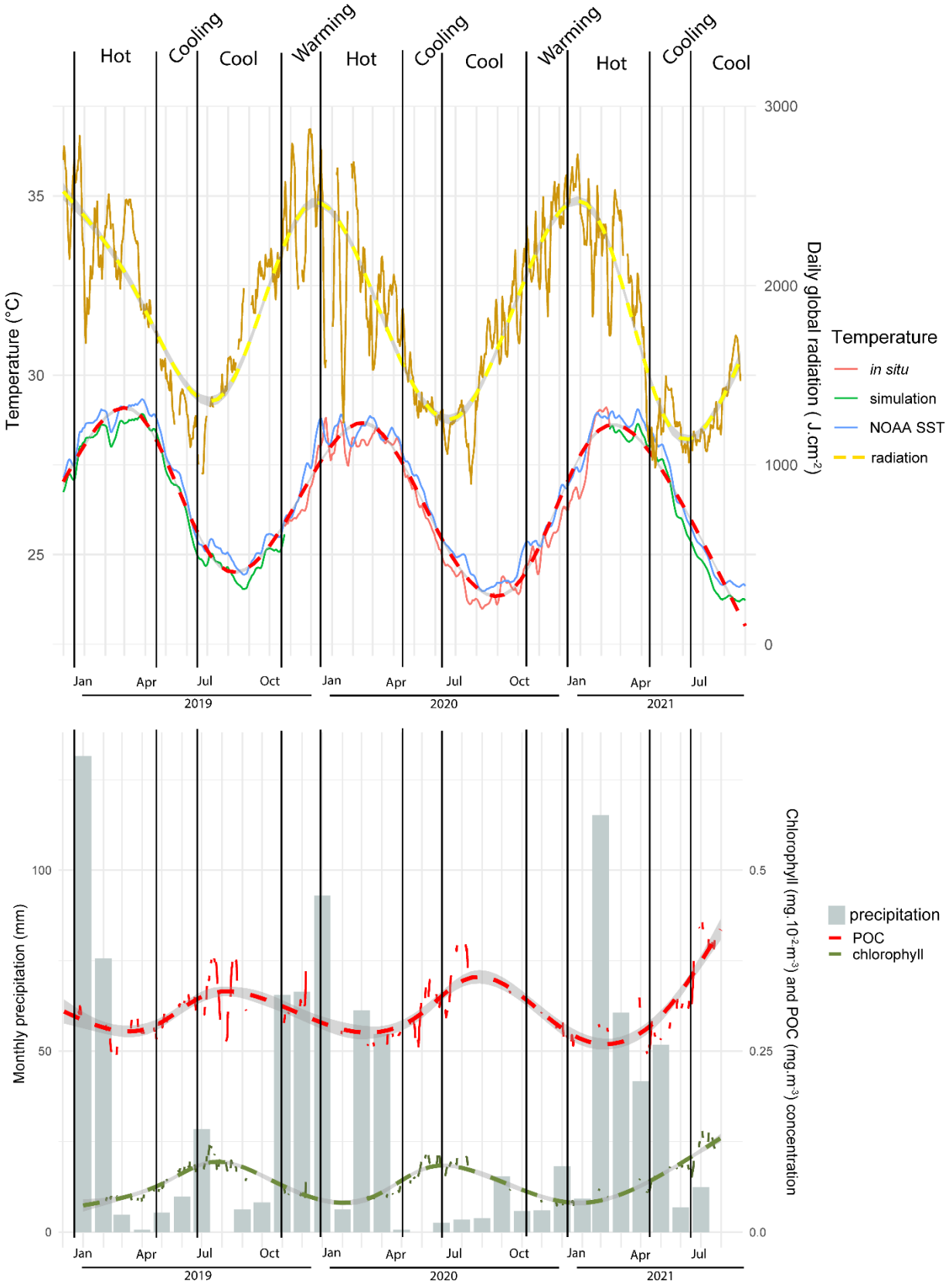


Figure 2: Environmental parameters during the study period.

3.2. Alpha diversity

After processing and cleaning, the 18S dataset contained a total of 5,621 OTUs, of which 959 OTUs (17%) were assigned to Eukaryota (Tab. 1). The COI datasets contained a total of 4,722 (OTU99) and 4,870 (OTU97) OTUs, but only 438 (OTU99; 9.3%) and 316 (OTU97; 6.4%) were assigned to Eukaryota, and only half of these could be assigned to species level.

For the 18S, OTU richness per ARMS did not vary significantly across modalities (KW, $X^2(4) = 3.20$, $p = 0.525$) and reached on average (± standard deviation) 463.6 ± 57.6 OTUs (Tab. 2). However, significant richness variations were observed within some fractions. Thus, in the fraction of 106-500 µm, significantly more OTUs were found in ARMS immerged during one year than those immerged for six months (Tukey, $p = 0.006$), and especially in ARMS that were deployed in the hot season (Tukey, $p = 0.03$). For the sessile fraction, OTU numbers were higher in ARMS immerged during one year than those immerged for two years (Tukey, $p = 0.02$). Similar to 18S, OTU richness of COI was not different between modalities and was on average (± standard deviation) 114.1 ± 16.7 OTUs (OTU99; KW, $X^2(4) = 2.32$, $p = 0.676$) and 90.9 ± 13.4 OTUs (OTU97; KW, $X^2(4) = 1.81$, $p = 0.771$; Tab. 2).

Accumulation curves for each dataset showed that three ARMS per modality were insufficient to sample total eukaryote richness. Moreover, although deploying 15 ARMS over different seasons and immersion times, the total eukaryote diversity of the site was not sampled, but represented 93.7% and 88.0% of the 18S and COI OTU richness respectively (ESM 3 and 4; Tab. 1).

Table 1: Number of total and Eukaryota (after removal of contaminants) reads and OTUs retrieved for each genetic markers. The percentages in brackets represent the proportion of eukaryote reads and OTUs in the overall dataset.

Dataset	**18S**	**COI OTU99%**	**COI OTU97%**
# reads	20,291,985	907,217	827,284
# OTUs	5,621	4,722	4,870
# Eukaryota reads	8,710,230 (42.9%)	122,571	104,517
# Eukaryota OTUs	959 (17%)	438 (9.3%)	316 (6.4)%
# Eukaryota OTUs estimated (Chao)	1024	498	369
Proportion of site diversity sampled with 15 ARMS	93.7%	88.0%	87%

Table 2: Mean number of OTUs retrieved by samples depending on the fraction or the immersion time for each dataset and genetic markers. sd: standard deviation.

Dataset		# samples	18S	COI	
			OTU mean ± sd	**OTU 99%** mean ± sd	**OTU 97%** mean ± sd
Fraction	100-500 µm	15	308.9 ± 39.6	55.2 ± 8.8	46.3 ± 7.5
	500-2000 µm	14	285.8 ± 40.2	59.1 ± 12.0	49.0± 10.6
	Sessile	15	277.8 ± 49.5	57.9 ± 10.3	46.9 ± 9.4
Immersion time	6 months	6	273.9 ± 23.8	57.5 ± 9.5	47.5 ± 8.6
	1 year	6	315.9 ± 42.5	55.5 ± 7.0	45.6 ± 5.6
	2 years	3	269.0 ± 54.6	61.2 ± 16.7	50.9 ± 14.9
OTU per ARMS (all fractions merged)		15	463.6 ± 57.6	114.1 ± 16.8	90.9 ± 13.4

3.3. Communities composition

The marker 18S detected 25 phyla including 15 metazoan phyla while the COI marker recovered 18 phyla comprising 11 metazoan phyla. In the 18S dataset, Annelida represented the highest proportion of OTUs and reads assigned at the ARMS level, and the 500-2000 µm and 106-500 µm fractions (Fig. 3; ESM 5). For the sessile fraction, the highest proportion of OTUs were assigned to Rhodophyta, however, Ascidiacea accounted for a highest proportion of reads (27.26%). In the COI99 dataset, Annelida were not dominant in terms of OTUs, although they contributed to the highest proportion of assigned reads at the ARMS level, and the 500-2000 µm and 106-500 µm fractions. Porifera contributed to the highest proportion of OTUs in all fractions (>18.26%) and accounted for a higher proportion of reads (26.35%) in the sessile fraction. NMDS analysis indicated that the composition of the ARMS communities differed with immersion time and deployment season. This trend was observed regardless of the genetic marker or the fraction analysed (ESM 6 and 7).

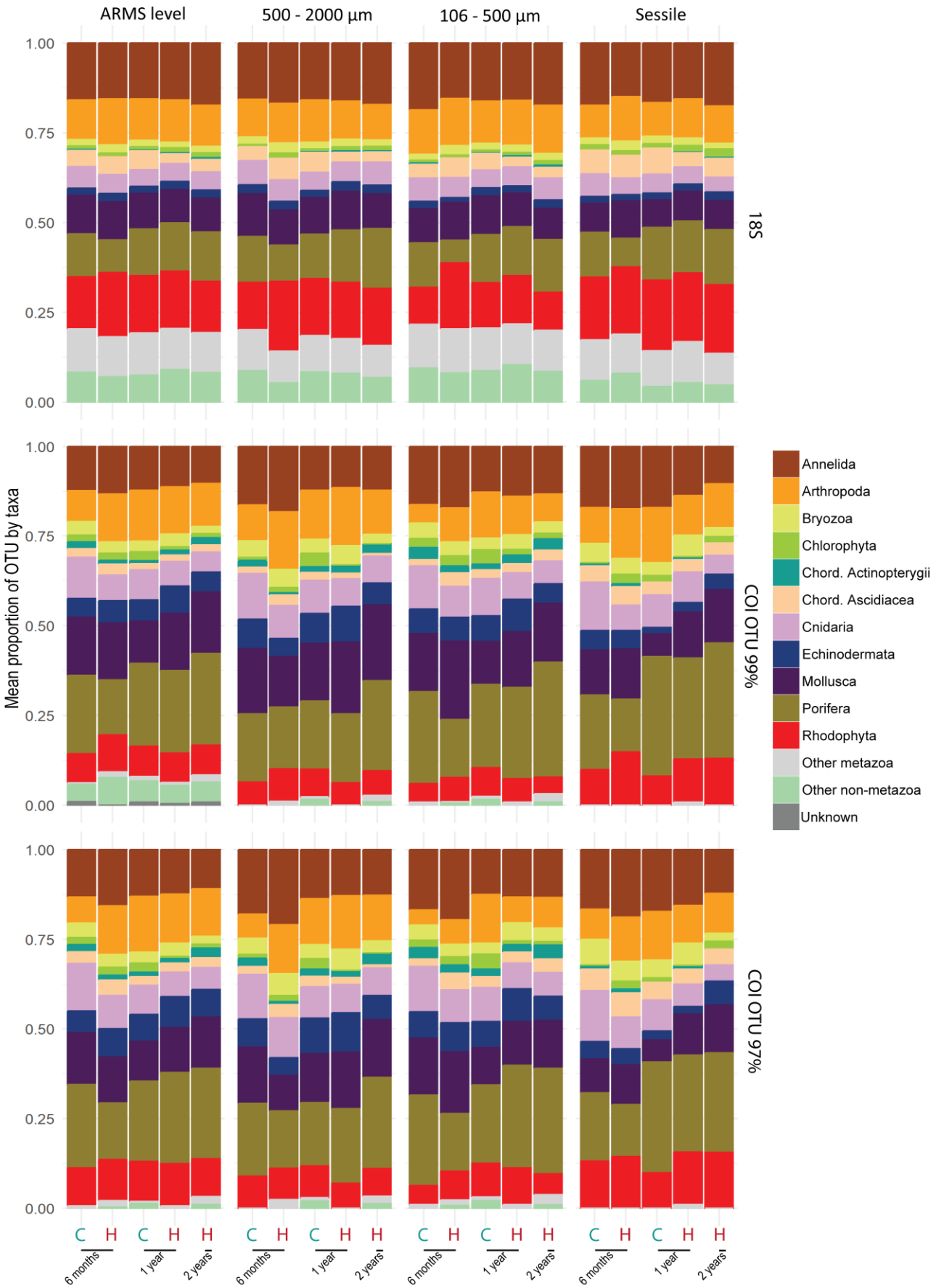

ARMS level
500 - 2000 µm
106 - 500 µm
Sessile
18S
COI OTU 99%
COI OTU 97%
Mean proportion of OTU by taxa
1.00
0.75
0.50
0.25
0.00
C H C H H
6 months
1 year
2 years
Annelida
Arthropoda
Bryozoa
Chlorophyta
Chord. Actinopterygii
Chord. Ascidiacea
Cnidaria
Echinodermata
Mollusca
Porifera
Rhodophyta
Other metazoa
Other non-metazoa
Unknown

Figure 3: Composition of the assemblages at taxa category level for each primer (horizontal blocks) and each fraction (vertical blocks). Letter codes below plots indicate the season of deployment: C = cool, H = hot.

Effects of immersion time

For both genetic markers, the immersion time of ARMS significantly affected the cryptobiome communities retrieved. For 18S, the communities were significantly different among the three immersion times at the ARMS level (pairwise.adonis, 18S: $p < 0.033$). For the fraction 500-2000 µm, the communities retrieved after one year of immersion differed significantly from those retrieved after shorter or longer immersion times (pairwise.adonis, $p < 0.036$). For the fractions 106-500 µm and sessile, the cryptobiome community after six months of immersion was significantly different from the one recovered after one year (pairwise.adonis, 106-500 µm: $p = 0.018$, sessile: $p = 0.027$). For COI99, the cryptobiome communities present at the ARMS level after six months' immersion differed significantly from the community present after immersion for one year (pairwise.adonis, $p = 0.012$). For the fraction 500-2000 µm, like for the 18S, the communities recovered after one year were significantly different from those recovered after immersion for shorter or longer periods (pairwise.adonis, $p < 0.042$). Finally, for the fractions 106-500 µm and sessile, the communities were significantly different between all three immersion times (pairwise.adonis, 106-500 µm: $p < 0.048$, sessile: $p < 0.042$).

The immersion time affected phyla differently. For Annelida, the proportion of OTUs decreased with increasing immersion times (Fig. 3). This trend was also reflected in a decrease of read abundance of Annelida discriminant OTUs (OTUs which were involved in 50% of the observed difference): sedentary annelids decreased from one to two years of immersion and appeared to be replaced partly by errant annelids in the two mobile fractions (ESM 8). For Arthropoda, little variation in OTU proportions among modalities were observed with the 18S dataset. This contrasts with the COI99 dataset which identified a higher proportion of Arthropoda in ARMS immersed for one year and deployed in the hot season (Fig. 3). The simper analyses of both markers suggested that the difference between Arthropoda communities after six months and one year of immersion was related to a decrease in the abundance of sessile arthropods (cirripeds) and an increase of mobile arthropods (ESM 8). For the sessile taxa, the proportion of Porifera and Rhodophyta OTUs and reads of the discriminant OTUs increased with immersion time (Fig. 3; ESM 8). The proportion of ascidians

and Cnidaria OTUs decreased with immersion time (Fig. 3). Between ARMS sets immerged for one year and for two years, the Simper analyses highlighted a decrease of solitary ascidians (Class: Stolidobranchia) and an increase in colonial ascidians (Class: Aplousobranchia) in 106-500 µm and sessile fraction of the 18S marker.

Seasonal effects

The 18S dataset highlighted significant differences in communities among season of deployment for all fractions (PERMANOVA, ARMS level: p = 0.005, 500-2000 µm: p = 0.003, 106-500 µm: p = 0.001, sessile: p = 0.01). These differences were also observed in the COI99 dataset at ARMS level (PERMANOVA, p=0.008**) and for the 500-2000 µm (PERMANOVA, p=0.03*) and sessile (PERMANOVA, p=0.003**) fractions. The encrusting taxa like Ascidiacea, Bryozoa, Cnidaria and Porifera had higher proportions of OTUs in ARMS deployed in the cool season (Fig. 3). Furthermore, the discriminant OTUs of these taxa had more reads in ARMS deployed in the cool season compared to those deployed in the hot season. In contrast, Rhodophyta and Mollusca were represented by a higher proportion of OTUs in ARMS deployed in the hot season (Fig. 3). The 18S marker also showed a general trend of higher abundance (reads numbers) of Rhodophyta (but not Ceramiales) in the hot season.

More comparisons among deployment seasons showed significant differences than comparisons among seasons of retrieval. However, for both markers, significant differences were observed among seasons of retrieval at the ARMS level (PERMANOVA, COI99: p=0.002, 18S: p=0.001), and for the fractions 500-2000 µm (PERMANOVA, COI99: p=0.03*, 18S: p=0.002) and 106-500 µm (PERMANOVA, COI99 & 18S: p=0.001). Moreover, the Ascidiacea represented a higher proportion of the OTUs in ARMS retrieved during the cool season (Fig. 3).

3.4. Community structure

For both genetic markers, most of the OTUs were exclusive to their immersion time (Fig. 4, ESM 9 and 10) and replacement was a higher component of beta diversity than richness differences (Fig. 5, ESM 11). The community structures provided a consistent pattern across markers, therefore only results for the COI99, due its higher taxonomic precision, are presented here. Results for the 18S are available in the ESM.

Less than 10% of the cryptobiome community was shared among the three immersion times (all season included; Fig. 4 and ESM 10) and the decomposition of beta diversity highlighted that between 39% (ARMS level) and 28% (106-500 µm) of the cryptobiome community was shared among two ARMS of different immersion times (Fig. 5). The highest proportion of unique OTUs was found in ARMS immerged for one year, followed by ARMS immerged for two years and those immerged for six months (Fig. 4 and ESM 10). In addition, the similarity among ARMS replicates was higher than among ARMS of different modalities. For example, at the ARMS level, Jaccard similarity among modalities reached 35% against 29% across all 15 ARMS. This trend was also present within the same immersion times, but blurred by the variability induced by seasonality among modalities.

Sequential immersion times (i.e., six months compared to one year and one year compared to two years) were more similar than the more distant immersion times (i.e., six months compared to two years. In fact, successive immersion times shared more OTUs (Fig. 5 and ESM 9) and showed smaller values of replacement (KW, $p<0.03$; Fig. 5 and ESM 11) than the more distant immersion times. A closer examination of the ARMS were retrieved during the same period (hot season: dec-2020 / jan-2021) confirmed that successive immersion times shared more also OTUs than the more distant immersion times (ESM 12, 13 and 14) and the proportions of shared OTUs were not different. In addition, increasing immersion times resulted in decreasing similarity among ARMS replicates, and among ARMS batches with the same immersion time (ESM 15). The decrease of similarity was explained by two process: an increasing species turnover between six months and one year, and then an increasing richness difference between one and two years of immersion.

Considering the season, ARMS deployed or retrieved in the cool season demonstrated greater similarities than those deployed or retrieved in the hot season (ESM 16 and 17). This trend was clearest in the mobile fractions, however no significant difference between Jaccard similarities was found (KW, $p > 0.05$).

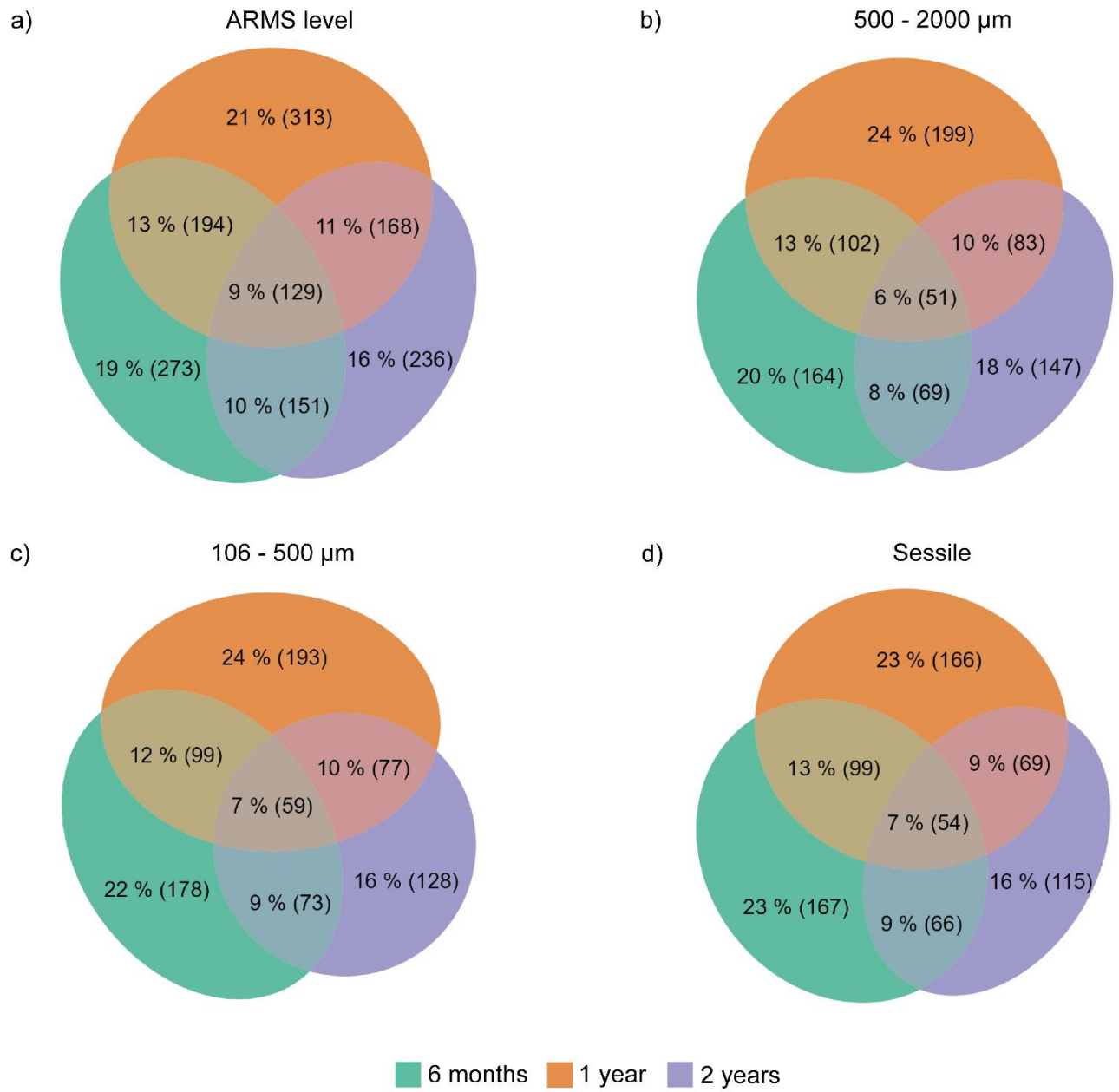


Figure 4: Number and proportion of unique and shared OTU99% for the COI among the three immersion times for the ARMS level dataset and the three fractions datasets. Ellipses are proportional.

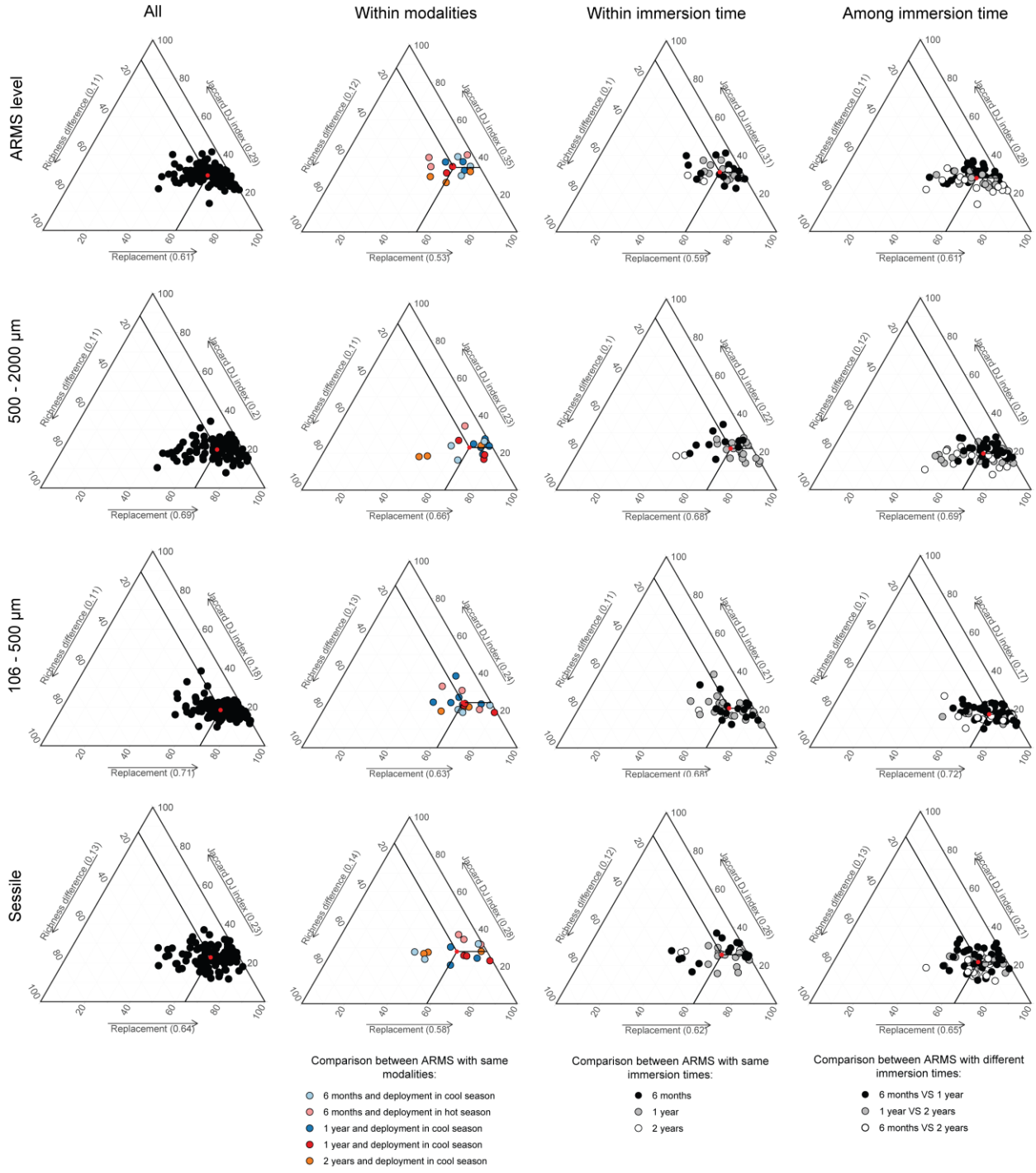


Figure 5: Ternary plots of Jaccard similarity and the partitions of beta diversity (replacement and richness) for the ARMS unit (ARMS level) and the three fractions obtained from the OTU99% of the COI metabarcoding. Ternary plots are shown for the total experiment (All) as well as within modalities and within and among immersion times. Red dot and numbers in brackets on the axis labels represent the mean value.

4. Discussion

Despite efforts to build a local database, combine public databases and use different assignation methods, only a small fraction of the diversity could be taxonomically assigned (18S: 17%; COI: 9.3%), compared to COI OTUs that were taxonomically resolved in other studies (e.g., 100% in **(Ip et al. 2022)**, 51% in **(Casey et al. 2021)**). However, our assignment process was intentionally more stringent to limit false taxonomic assignments. Indeed, some studies used an 85% threshold for the COI against 95% in the present study (Table 3). This low rate of assigned OTUs was also reflected in the average number of OTUs found by ARMS, as the proportions of OTUs considered here (Eukaryota) were lower than if a less stringent threshold was employed. For this reason, previous studies using ARMS to investigate the coral reef cryptobiome found higher numbers of OTUs, ranging from fourfold **(Leray & Knowlton 2015)** to eightfold **(Carvalho et al. 2019)** higher OTU numbers compared to the present study. Several other factors need to be considered to compare our results with previous studies, such as (1) the OTU clustering threshold (OTU size in Table 3); (2) the filtration step (removing bacteria or keeping only a subset of taxa, like metazoans); (3) the completeness of the local reference database and (4) the number of sites studied.

In Reunion Island, OTU assignments were dominated by Annelida and Porifera, regardless of the fractions, comparable to results from the Red Sea **(Pearman et al. 2018)**. Other cryptobiome inventories using ARMS found a high proportion of Arthropoda and few Porifera **(Villalobos et al. 2022)**. The difference in assigned proportions could be explain by higher substitution rate of nucleic acids in COI for Arthropoda than for Porifera **(Huang et al. 2008)** which decreases the probability of having genetically close references for Arthropoda and thus decreases the numbers of OTUs assigned to Arthropoda for the COI. These results further highlight the uniqueness of the Mascarene cryptobiome and the need for further sampling and sequencing efforts on these communities to complement a local reference database.

Table 3: Summary of the parameters employed for ARMS deployment and OTUs processing for reef cryptobiome studies. An extended version is available in ESM 18. * represent filtration based on abundance of reads. 1: only metazoans and marcroalgae; 2: only metazoans.

Site	# site	# ARMS	Imm. time (month)	Immersion period	OTU size	Filtration				Assignment threshold	% OTU with phylum assignement	# OTU	Mean # OTU by ARMS	Reference
						Cod	Sin	Bac	Tax					
Florida (USA)	1	9	6	Nov - May	NA (CROP)	V	V	V		97% blast 90% SAP	72%	1 391	536 ± 30	Leray & Knowlton 2015
Virginie (USA)	1	9	6	Sep - May	NA (CROP)	V	V	V		97% blast 90% SAP	59%	1 204	434 ± 55	
Gulf of Aqaba (Jordan)	2	5	16	Oct - Feb	NA (CROP)	V	V	V		97% blast 80% SAP	63%	1 197	609 ± 114	Al-Rshaidat et al 2016
Saudi Arabian coast	3	9	12	Apr - May	97%		V			NA	NA	1 700	1 297*	Pearman et al. 2016
Thuwal (Saudi Arabia)	11	33	24	Feb, May, Jun - May, Jul	NA (CROP)	V	V			97% blast SAP	58%	3830	660 ± 151	Pearman et al. 2018
Saudi Arabian coast	22	87	24	Feb, May ,Aug - May, Jun, Jul, Nov	ESM not available					ESM not available	55%	10 416 (1 471 by site)	828	Carvalho et al. 2019
Saudi Arabian coast	4	33	24	May	100% (ASV)	V	V			RDP	NA	33 832 ASV	NA	Villalobos et al. 2022
Mo'orea (French Polynesia)	1	3	24	Jan	NA		V	V		97% blast 85% blast 90% SAP	55%	2 456	NA	Ransome et al. 2017
Bali (Indonesia)	2	6	11 and 23	Jul - Jun	97%	V	V			85% blast	51%	31 900	6 580 to 14 237	Casey at al. 2021
Hawai'i	1	6	23	Jul	97%	V	V*		1	97 % blast 85% LCA	NA	893	NA	Nichols et al. 2021
Singapore	4	12	24	Jun –Jun, Jul	97%	V	V*		2	RDP 80% confidence	100%	410	NA	Ip et al. 2022

Marine communities are generally dominated by a few taxa, with most of the diversity is represented by rare species whose presence may vary in space and time **(Logares et al. 2014 ; Lindh et al. 2017)**. Our results are congruent with this observation, as a small part of the OTU diversity composed the core community of the reef cryptobiome across immersion times, as was also the case in the Red Sea **(Carvalho et al. 2019 ; Villalobos et al. 2022)**. Low similarity values among ARMS replicates indicate considerable changes in the species composition at small spatial scales (i.e., < 5 m). The partitioning of beta diversity provided similar values to those found in the Red Sea **(Pearman 2018)** with an average OTU replacement rate of 60% between two ARMS at the same site. The partitioning of beta diversity also indicated that replacement was higher for the mobile fractions than for the sessile fraction, indicating a greater stochasticity in mobile fauna. Moreover, although deploying 15 ARMS over different seasons and immersion times, this sampling effort failed to recover the total estimated eukaryote diversity of that site, highlighting the overwhelming diversity of cryptic species on coral reefs **(Plaisance et al. 2011 ; Pearman et al. 2016)**.

Deployment of ARMS over different seasons and immersion times revealed significant temporal effects on the cryptobiome communities retrieved. While the alpha diversity of OTUs sampled with ARMS did not depend on the immersion time or the deployment and/or retrieval seasons, this was not the case for the composition of the cryptic communities. Moreover, the different fractions collected by ARMS (i.e., 500-2000 µm, 106-500 µm and sessile) emphasised different community compositions **(Leray & Knowlton 2015 ; Ip et al. 2022)**, but showed similar trends in community structure, as detected by both 18S and COI markers.

4.1. Succession in communities across immersion times

The cryptobiome in ARMS differed according to the immersion time. Contrary to the initial hypothesis, the number of OTUs collected did not increase over time. Instead, the cryptobiome showed a strong temporal turnover, with an average replacement of over 60% of the OTUs between two immersion times. At the level of encrusting organisms, we observed a replacement of taxonomic groups with immersion time. Reflecting their ability to be early colonisers, ascidians, cirripeds and cnidarians were more abundant in ARMS immersed for six months. Their diversity decreased with immersion time in favour of Porifera and Rodophyta OTU that became more abundant. The latter taxon includes crustose coralline algae that are known to undergo species successions after colonising newly available substrates **(Adey & Vassar 1975)**. The lack of OTU taxonomic assignations did not allow to consistently analyse such species successions. However, the decrease in ascidians diversity appeared mainly due to a decline of solitary ascidians. For the mobile cryptobiome, we observe a decrease of Annelida diversity with immersion time while arthropods reached

their maximum diversity in ARMS that were immerged during one year (especially for those deployed and retrieved in the hot season).

The decline in similarity with immersion time may relies on two ecological processes may underpin. In the early stages of ARMS colonisation, communities among replicates were quite similar, suggesting settlement by a suite of pioneer species ("early succession" species in **Connell & Slatyer (1977)**). After one year of immersion, similarity decreased as species turnover among ARMS replicates increased. We hypothesise that a pool of pioneer species first establishes itself in the ARMS and is later replaced by a pool of subsequent colonizers of the mature community which arrive randomly. After two years of immersion, species replacement decreased in contrast to the increase in richness difference suggesting the maintenance of part of the communities (the core communities) and the overgrowth by some species that become dominant in certain ARMS replicates.

4.2. Season shapes communities

The study site on Reunion's outer reef slope was subject to seasonal variations in environmental conditions. The two main seasons, hot and cool, each lasted about four months while the two inter-seasons lasted two months **(Kolasinski et al. 2011)**. The hot season (January to April) was marked by strong solar radiation and higher SST as well as high but variable rainfall, related to cyclonic conditions. Higher concentrations of chlorophyll and particulate organic matter characterized the cool season. The spatio-temporal dynamics of cryptic communities may be linked to this seasonal variability. To our best knowledge, this study is the first to highlight the role played by the season in community pattern composition sampled by ARMS.

Similarity values within season indicated greater changes in the species composition in the hot season than in the cool season. The increase in dissimilarity in summer is hypothesized to be related to the reproduction of organisms. Thus, these ARMS may have collected additional taxa that reproduce around this time of year (e.g. eggs of gastropods, fishes) **(Aranda et al. 2003)**. However, alpha diversity was constant and the richness difference values remained low between hot and cool seasons.

The season of ARMS deployment seemed to have significant effect on community composition at the retrieval season than the retrieval season itself (fewer comparisons were significant). Moreover, the season seemed to affect sessile taxa more than mobile ones. The sessile taxa such as Bryozoa, Cnidaria and Porifera demonstrated higher proportions of OTUs in ARMS deployed in the cool season. Conversely, Rhodophyta demonstrated greater proportions in hot season. This is consistent with the fact that coral reef macroalgal

biomass increases with temperature at Reunion **(Naim 1993)**. Furthermore, the OTU proportion of Mollusca increased during the hot season, which could be explained by an increase of food resources such as Rhodophyta **(Morton & Blackmore 2009 ; Larkin et al. 2017)**. However, ascidians found in ARMS appeared to be influenced by the retrieval season. Our results showed a higher OTU proportion of ascidians in ARMS collected during the cool season. Several other studies have shown a seasonality in the abundance of ascidians, which is generally correlated with the supply of nutrients in the environment **(Shenkar et al. 2008)**. Thus, our results were consistent with the peak of chlorophyll, POC and the reef waters δ13C enrichment during the cool and dry season **(Kolasinski et al. 2011)**. Nevertheless, the ascidian recruitment patterns seem to be species-specific with variations varying between seasons, orientation and position on the substrata **(Shenkar et al. 2008)**. Moreover, given their short pelagic duration and limited larval swimming ability, ascidians generally have a relatively localized dispersal **(Shanks et al. 2003 ; Weersing & Toonen 2009)**. The dispersal of larvae and thus the colonisation of ARMS may thus be favoured by the stronger hydrodynamic conditions that prevail during the cool season **(Naim 1993 ; Chazottes et al. 2008)**. Overall, the seasonal variations in coral reef communities include complex interactions of environmental factors such as water temperature, daily irradiance, rainfall, and nutrient availability. Taxon-specific studies are needed to better understand the implications of seasonal variations.

4.3. Implications for future studies

To the best of our knowledge, this was the first study to highlight the effects of immersion time and season on the communities sampled by ARMS. Our results showed that both factors need to be taken into account in designing the sampling plan and when analysing the results. Depending on the study scope, reducing immersion times could provide more frequent information and detect faster changes in communities. Indeed, after six months of immersion, the same number of collected species was reached compared to one and two years of immersion. Our results also highlighted effects of season of deployment and of retrieval. We therefore suggest to deploy and retrieve ARMS during the same time of year / season. Thus, ARMS could be used on a year-by-year basis to monitor the evolution of communities over time in the context of global change. Given the lack of available base-line data, such information could be very useful as benthic recruitment is highly variable from year to year **(Bento et al. 2017)** and it is unknown whether the results of any short-term study are representative of longer-term patterns or of other reefs. Carrying out a short pilot study aiming to evaluate possible seasonal effects, before starting a longer-term monitoring program, would probably improve the interpretation of the results.

5. Conclusion

The effects of immersion time and season on the communities collected by ARMS were analysed systematically for the first time here. Our results show that both factors need to be taken into account in monitoring or quantifying cryptobentic diversity patterns using ARMS, and probably other standardized approaches. Furthermore, depending on the scope of the study, the conventional immersion of two years could be shortened to provide more frequent information, detect changes in the communities more quickly and reduce the risk of losing ARMS units (e.g., due to cyclonic events). Indeed, the number of OTUs collected with ARMS does not seem to depend on the immersion time or the deployment and/or retrieval season. However, differences were observed in the composition of the communities. Our results suggest an initial colonisation of ARMS by pool of pioneer taxa, which subsequently partly disappear due to the random arrival of later succession taxa and overgrowth by competitors. As a result, only a small proportion of the community remains stable over time. In addition, the season in which ARMS was deployed seems to have a greater effect on the taxa found at the retrieval season than the retrieval season itself. The sessile organisms appear to be more sensitive to the seasonal effect, such as ascidians that were most abundant in the cool season, reflecting enrichment of coastal waters with Chl-a and POM. Finally, although we deployed 15 ARMS in a single site at a single depth on the outer slopes of the coral reef in Reunion, this sampling effort did not suffice to recover the cryptobiome diversity of the site, underscoring the overwhelming diversity of cryptic species present in the reef. Furthermore, the uniqueness of the Mascarene cryptobiome and the need for further sampling and sequencing of these communities is highlighted by the small fraction of diversity that can presently be assigned.

6. References

Adey WH., Vassar JM. (1975) Colonization, succession and growth rates of tropical crustose coralline algae (Rhodophyta, Cryptonemiales). Phycologia 14:55–69. DOI: 10.2216/i0031-8884-14-2-55.1

Albaina A., Aguirre M., Abad D., Santos M., Estonba A. (2016) 18S rRNA V9 metabarcoding for diet characterization: a critical evaluation with two sympatric zooplanktivorous fish species. Ecology and Evolution 6:1809–1824. DOI: 10.1002/ece3.1986

Alexander JB., Bunce M., White N., Wilkinson SP., Adam AAS., Berry T., Stat M., Thomas L., Newman SJ., Dugal L., Richards ZT. (2019) Development of a multi-assay approach for monitoring coral diversity using eDNA metabarcoding. Coral Reefs. DOI: 10.1007/s00338-019-01875-9

Antich A., Palacín C., Cebrian E., Golo R., Wangensteen OS., Turon X. (2020) Marine biomonitoring with eDNA: can metabarcoding of water samples cut it as a tool for surveying benthic communities? Molecular Ecology n/a. DOI: 10.1111/mec.15641

Aranda DA., Cárdenas EB., Martínez I., Zárate AZ., Brulé T. (2003) A Review of the Reproductive Patterns Of Gastropod Mollusks from Mexico. BULLETIN OF MARINE SCIENCE 73.

Astudillo JC., Leung KMY., Bonebrake TC. (2016) Seasonal heterogeneity provides a niche opportunity for ascidian invasion in subtropical marine communities. Marine Environmental Research 122:1–10. DOI: 10.1016/j.marenvres.2016.09.001

Ateweberhan M., Bruggemann JH., Breeman AM. (2006) Effects of extreme seasonality on community structure and functional group dynamics of coral reef algae in the southern Red Sea (Eritrea). Coral Reefs 25:391–406. DOI: 10.1007/s00338-006-0109-6

Bento R., Feary D., Hoey A., Burt J. (2017) Settlement Patterns of Corals and other Benthos on Reefs with Divergent Environments and Disturbances Histories around the Northeastern Arabian Peninsula. Frontiers in Marine Science 4:305. DOI: 10.3389/fmars.2017.00305

Bolyen E., Rideout JR., Dillon MR., Bokulich NA., Abnet C., Al-Ghalith GA., Alexander H., Alm EJ., Arumugam M., Asnicar F., Bai Y., Bisanz JE., Bittinger K., Brejnrod A., Brislawn CJ., Brown CT., Callahan BJ., Caraballo-Rodríguez AM., Chase J., Cope E., Da Silva R., Dorrestein PC., Douglas GM., Durall DM., Duvallet C., Edwardson CF., Ernst M., Estaki M., Fouquier J., Gauglitz JM., Gibson DL., Gonzalez A., Gorlick K., Guo J., Hillmann B., Holmes S., Holste H., Huttenhower C., Huttley G., Janssen S., Jarmusch AK., Jiang L., Kaehler B., Kang KB., Keefe CR., Keim P., Kelley ST., Knights D., Koester I., Kosciolek T., Kreps J., Langille MG., Lee J., Ley R., Liu Y-X., Loftfield E., Lozupone C., Maher M., Marotz C., Martin BD., McDonald D., McIver LJ., Melnik AV., Metcalf JL., Morgan SC., Morton J., Naimey AT., Navas-Molina JA., Nothias LF., Orchanian SB., Pearson T., Peoples SL., Petras D., Preuss ML., Pruesse E., Rasmussen LB., Rivers A., Robeson, II MS., Rosenthal P., Segata N., Shaffer M., Shiffer A., Sinha R., Song SJ., Spear JR., Swafford AD., Thompson LR., Torres PJ., Trinh P., Tripathi A., Turnbaugh PJ., Ul-Hasan S., van der Hooft JJ., Vargas F., Vázquez-Baeza Y., Vogtmann E., von Hippel M., Walters W., Wan Y., Wang M., Warren J., Weber KC., Williamson CH., Willis AD., Xu ZZ., Zaneveld JR., Zhang Y., Zhu Q., Knight R., Caporaso JG. (2018) QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. PeerJ Preprints. DOI: 10.7287/peerj.preprints.27295v2

Callahan B. (2021) Decontam.

Callahan BJ., McMurdie PJ., Holmes SP. (2017) Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME Journal 11:2639–2643. DOI: 10.1038/ismej.2017.119

Carvalho S., Aylagas E., Villalobos R., Kattan Y., Berumen M., Pearman JK. (2019) Beyond the visual: using metabarcoding to characterize the hidden reef cryptobiome. Proceedings of the Royal Society B: Biological Sciences 286:20182697. DOI: 10.1098/rspb.2018.2697

Casey JM., Ransome E., Collins AG., Mahardini A., Kurniasih EM., Sembiring A., Schiettekatte NMD., Cahyani NKD., Wahyu Anggoro A., Moore M., Uehling A., Belcaid M., Barber PH., Geller JB., Meyer CP. (2021) DNA metabarcoding marker choice skews perception of marine eukaryotic biodiversity. Environmental DNA 3:1229–1246. DOI: 10.1002/edn3.245

Chazottes V., Reijmer JJG., Cordier E. (2008) Sediment characteristics in reef areas influenced by eutrophication-related alterations of benthic communities and bioerosion processes. Marine Geology 250:114–127. DOI: 10.1016/j.margeo.2008.01.002

Cinner JE., Zamborain-Mason J., Gurney GG., Graham NAJ., MacNeil MA., Hoey AS., Mora C., Villéger S., Maire E., McClanahan TR., Maina JM., Kittinger JN., Hicks CC., D'agata S., Huchery C., Barnes ML., Feary DA., Williams ID., Kulbicki M., Vigliola L., Wantiez L., Edgar GJ., Stuart-Smith RD., Sandin SA., Green AL., Beger M., Friedlander AM., Wilson SK., Brokovich E., Brooks AJ., Cruz-Motta JJ., Booth DJ., Chabanet P., Tupper M., Ferse SCA., Sumaila UR., Hardt MJ., Mouillot D. (2020) Meeting fisheries, ecosystem function, and biodiversity goals in a human-dominated world. Science 368:307–311. DOI: 10.1126/science.aax9412

Connell JH., Slatyer RO. (1977) Mechanisms of Succession in Natural Communities and Their Role in Community Stability and Organization. The American Naturalist 111:1119–1144. DOI: 10.1086/283241

Corse E., Meglécz E., Archambaud G., Ardisson M., Martin J-F., Tougard C., Chappaz R., Dubut V. (2017) A from-benchtop-to-desktop workflow for validating HTS data and for taxonomic identification in diet metabarcoding studies. Molecular Ecology Resources 17:e146–e159. DOI: 10.1111/1755-0998.12703

Couëdel M., Dettai A., Guillaume MMM., Bruggemann F., Bureau S., Frattini B., Verde Ferreira A., Azie J-L., Bruggemann JH. (2023) New insights into the diversity of cryptobenthic *Cirripectes* blennies in the Mascarene Archipelago sampled using Autonomous Reef Monitoring Structures (ARMS). Ecology and Evolution 13:e9850. DOI: 10.1002/ece3.9850

David R., Uyarra MC., Carvalho S., Anlauf H., Borja A., Cahill AE., Carugati L., Danovaro R., De Jode A., Feral J-P., Guillemain D., Martire ML., D'Avray LTDV., Pearman JK., Chenuil A. (2019) Lessons from photo analyses of Autonomous Reef Monitoring Structures as tools to detect (bio-)geographical, spatial, and environmental effects. Marine Pollution Bulletin 141:420–429. DOI: 10.1016/j.marpolbul.2019.02.066

Elbrecht V., Vamos EE., Meissner K., Aroviita J., Leese F. (2017) Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. Methods in Ecology and Evolution 8:1265–1275. DOI: 10.1111/2041-210X.12789

Frøslev TG., Kjøller R., Bruun HH., Ejrnæs R., Brunbjerg AK., Pietroni C., Hansen AJ. (2017) Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. Nature Communications 8:1188. DOI: 10.1038/s41467-017-01312-x

Gaudron S., KOHLER S., Conand C. (2008) Reproduction of the sea cucumber Holothuria leucospilota in the Western Indian Ocean: Biological and ecological aspects. Invertebrate Reproduction & Development 51:19–31. DOI: 10.1080/07924259.2008.9652253

Gibson J., Shokralla S., Porter TM., King I., van Konynenburg S., Janzen DH., Hallwachs W., Hajibabaei M. (2014) Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. Proceedings of the National Academy of Sciences 111:8007–8012.

Glasby TM. (1999) Effects of shading on subtidal epibiotic assemblages. Journal of Experimental Marine Biology and Ecology 234:275–290. DOI: 10.1016/S0022-0981(98)00156-7

Hammer Ø., Harper DAT., Ryan PD. (2001) PAST: Paleontological Statistics software package for education and data analysis. Palaeontological Electronica 4:9.

Hazeri G., Rahayu DL., Subhan B., Sembiring A., Anggoro AW., Ghozali AT., Madduppa HH. (2019) Latitudinal species diversity and density of cryptic crustacean (Brachyura and Anomura) in micro-habitat Autonomous Reef Monitoring Structures across Kepulauan Seribu, Indonesia. Biodiversitas Journal of Biological Diversity 20. DOI: 10.13057/biodiv/d200540

Hsieh TC., Chao KHM and A. (2022) INEXT: Interpolation and Extrapolation for Species Diversity.

Huang D., Meier R., Todd P., Chou L. (2008) Slow Mitochondrial COI Sequence Evolution at the Base of the Metazoan Tree and Its Implications for DNA Barcoding. Journal of molecular evolution 66:167–74. DOI: 10.1007/s00239-008-9069-5

Ip YCA., Chang JJM., Oh RM., Quek ZBR., Chan YKS., Bauman AG., Huang D. (2022) Seq' and ARMS shall find: DNA (meta)barcoding of Autonomous Reef Monitoring Structures across the tree of life uncovers hidden cryptobiome of tropical urban coral reefs. Molecular Ecology:1–20. DOI: 10.1111/mec.16568

Jaccard P. (1912) The Distribution of the Flora in the Alpine Zone.1. New Phytologist 11:37–50. DOI: 10.1111/j.1469-8137.1912.tb05611.x

Kolasinski J., Rogers K., Cuet P., Barry B., Frouin P. (2011) Sources of particulate organic matter at the ecosystem scale: a stable isotope and trace element study in a tropical coral reef. Marine Ecology Progress Series 443:77–93. DOI: 10.3354/meps09416

Larkin M., Smith S., Willan R., Davis T. (2017) Diel and seasonal variation in heterobranch sea slug assemblages within an embayment in temperate eastern Australia. Marine Biodiversity. DOI: 10.1007/s12526-017-0700-9

Larsson J., Godfrey AJR., Gustafsson P., algorithms) DHE (geometric., code) EH (root solver., Privé F. (2022) Eulerr: Area-Proportional Euler and Venn Diagrams with Ellipses.

Legendre P. (2014) Interpreting the replacement and richness difference components of beta diversity. Global Ecology and Biogeography 23:1324–1334. DOI: 10.1111/geb.12207

Leray M., Knowlton N. (2015) DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. Proceedings of the National Academy of Sciences 112:2076–2081. DOI: 10.1073/pnas.1424997112

Leray M., Meyer CP., Mills SC. (2015) Metabarcoding dietary analysis of coral dwelling predatory fish demonstrates the minor contribution of coral mutualists to their highly partitioned, generalist diet. PeerJ 3:e1047. DOI: 10.7717/peerj.1047

Leray M., Yang JY., Meyer CP., Mills SC., Agudelo N., Ranwez V., Boehm JT., Machida RJ. (2013) A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. Frontiers in Zoology 10:34. DOI: 10.1186/1742-9994-10-34

Lindh MV., Sjöstedt J., Ekstam B., Casini M., Lundin D., Hugerth LW., Hu YOO., Andersson AF., Andersson A., Legrand C., Pinhassi J. (2017) Metapopulation theory identifies biogeographical patterns among core and satellite marine bacteria scaling from tens to thousands of kilometers. Environmental Microbiology 19:1222–1236. DOI: 10.1111/1462-2920.13650

Logares R., Audic S., Bass D., Bittner L., Boutte C., Christen R., Claverie J-M., Decelle J., Dolan JR., Dunthorn M., Edvardsen B., Gobet A., Kooistra WHCF., Mahé F., Not F., Ogata H., Pawlowski J., Pernice MC., Romac S., Shalchian-Tabrizi K., Simon N., Stoeck T., Santini S., Siano R., Wincker P., Zingone A., Richards TA., de Vargas C., Massana R. (2014) Patterns of Rare and Abundant Marine Microbial Eukaryotes. Current Biology 24:813–821. DOI: 10.1016/j.cub.2014.02.050

Machida RJ., Knowlton N. (2012) PCR Primers for Metazoan Nuclear 18S and 28S Ribosomal DNA Sequences. PLoS ONE 7:e46180. DOI: 10.1371/journal.pone.0046180

Martens K., Queiroga H., Cunha MR., Cunha A., Moreira MH., Quintino V., Rodrigues AM., Seroôdio J., Warwick RM. (eds) (2006) Marine Biodiversity: Patterns and Processes, Assessment, Threats, Management and Conservation. Springer Netherlands, Dordrecht., 978-1-4020-4321-5 DOI: 10.1007/1-4020-4697-9

Matano RP., Beier EJ., Strub PT., Tokmakian R. (2002) Large-Scale Forcing of the Agulhas Variability: The Seasonal Cycle. Journal of Physical Oceanography 32:1228–1241. DOI: 10.1175/1520-0485(2002)032<1228:LSFOTA>2.0.CO;2

Morton B., Blackmore G. (2009) Seasonal variations in the density of and corallivory by *Drupella rugosa* and *Cronia margariticola* (Caenogastropoda: Muricidae) from the coastal waters of Hong Kong: 'plagues' or 'aggregations'? Journal of the Marine Biological Association of the United Kingdom 89:147–159. DOI: 10.1017/S002531540800218X

Muthiga NA., Jaccarini V. (2005) Effects of seasonality and population density on the reproduction of the Indo-Pacific echinoid *Echinometra mathaei* in Kenyan coral reef lagoons. Marine Biology 146:445–453. DOI: 10.1007/s00227-004-1449-9

Naim O. (1993) Seasonal responses of a fringing reef community to eutrophication (Reunion Island, Western Indian Ocean). Marine Ecology Progress Series 99:137–151. DOI: 10.3354/meps099137

Nichols PK., Timmers M., Marko PB. (2021) Hide 'n seq: Direct versus indirect metabarcoding of coral reef cryptic communities. Environmental DNA 4:93–107. DOI: 10.1002/edn3.203

Palomino-Alvarez LA., Vital XG., Castillo-Cupul RE., Suárez-Mozo NY., Ugalde D., Cervantes-Campero G., Muciño-Reyes MR., Homá-Canché P., Hernández-Díaz YQ., Sotelo-Casas R., García-González M., Avedaño-Peláez YA., Hernández-González A., Paz-Ríos CE., Lizaola-Guillermo JM., García-Venegas M., Dávila-Jiménez Y., Ortigosa D., Hidalgo G., Tello-Musi JL., Rivera-Higueras M., Moreno Mendoza R., Wicksten MK., Rocha RM., Vieira L., Mendoza-Garfias MB., Simões N., Guerra-Castro EJ. (2021) Evaluation of the Use of Autonomous Reef Monitoring Structures (ARMS) for Describing the Species Diversity of Two Coral Reefs in the Yucatan Peninsula, Mexico. Diversity 13:579. DOI: 10.3390/d13110579

Pearman JK., Anlauf H., Irigoien X., Carvalho S. (2016) Please mind the gap – Visual census and cryptic biodiversity assessment at central Red Sea coral reefs. Marine Environmental Research 118:20–30. DOI: 10.1016/j.marenvres.2016.04.011

Pearman JK., Leray M., Villalobos R., Machida RJ., Berumen ML., Knowlton N., Carvalho S. (2018) Cross-shelf investigation of coral reef cryptic benthic organisms reveals diversity patterns of the hidden majority. Scientific Reports 8:1–17. DOI: 10.1038/s41598-018-26332-5

Pennesi C., Danovaro R. (2017) Assessing marine environmental status through microphytobenthos assemblages colonizing the Autonomous Reef Monitoring Structures (ARMS) and their potential in coastal marine restoration. Marine Pollution Bulletin 125:56–65. DOI: 10.1016/j.marpolbul.2017.08.001

Plaisance L., Caley MJ., Brainard RE., Knowlton N. (2011) The Diversity of Coral Reefs: What Are We Missing? PLOS ONE 6:e25026. DOI: 10.1371/journal.pone.0025026

R Core Team (2021) R: A language and environment for statistical computing.

Ransome E., Geller JB., Timmers M., Leray M., Mahardini A., Sembiring A., Collins AG., Meyer CP. (2017) The importance of standardization for biodiversity comparisons: A case study using autonomous reef monitoring structures (ARMS) and metabarcoding to measure cryptic diversity on Mo'orea coral reefs, French Polynesia. PLOS ONE 12:e0175066. DOI: 10.1371/journal.pone.0175066

Shanks AL., Grantham BA., Carr MH. (2003) Propagule Dispersal Distance and the Size and Spacing of Marine Reserves. Ecological Applications 13:159–169. DOI: 10.1890/1051-0761(2003)013[0159:PDDATS]2.0.CO;2

Shenkar N., Bronstein O., Loya Y. (2008) Population dynamics of a coral reef ascidian in a deteriorating environment. Marine Ecology Progress Series 367:163–171. DOI: 10.3354/meps07579

Steyaert M., Mogg A., Dunn N., Dowell R., Head CEI. (2022) Observations of coral and cryptobenthic sponge fluorescence and recruitment on autonomous reef monitoring structures (ARMS). Coral Reefs. DOI: 10.1007/s00338-022-02283-2

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

Tanner JE. (1996) Seasonality and lunar periodicity in the reproduction of Pocilloporid corals. Coral Reefs 15:59–66. DOI: 10.1007/BF01626077

Thomsen PF., Kielgast J., Iversen LL., Møller PR., Rasmussen M., Willerslev E. (2012) Detection of a diverse marine fish fauna using environmental DNA from seawater samples. PLoS one 7.

Vicente J., Webb MK., Paulay G., Rakchai W., Timmers MA., Jury CP., Bahr K., Toonen RJ. (2021) Unveiling hidden sponge biodiversity within the Hawaiian reef cryptofauna. Coral Reefs. DOI: 10.1007/s00338-021-02109-7

Villalobos R., Aylagas E., Pearman J., Cúrdia J., Lozano-Cortés D., Coker D., Jones B., Berumen M., Carvalho S. (2022) Inter-annual variability patterns of reef cryptobiota in the central Red Sea across a shelf gradient. Scientific Reports 12. DOI: 10.1038/s41598-022-21304-2

Weersing K., Toonen R. (2009) Population genetics, larval dispersal, and connectivity in marine systems. Marine Ecology Progress Series 393:1–12. DOI: 10.3354/meps08287

Yu DW., Ji Y., Emerson BC., Wang X., Ye C., Yang C., Ding Z. (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods in Ecology and Evolution 3:613–623. DOI: 10.1111/j.2041-210X.2012.00198.x

Zea S. (1993) Recruitment of Demosponges (Porifera, Demospongiae) in Rocky and Coral Reef Habitats of Santa Marta, Colombian Caribbean. Marine Ecology 14:1–21. DOI: 10.1111/j.1439-0485.1993.tb00361.x

Zimmerman TL., Martin JW. (2004) Artificial Reef Matrix Structures (Arms): An Inexpensive and Effective Method for Collecting Coral Reef-Associated Invertebrates. Gulf and Caribbean Research 16:59–64. DOI: 10.18785/gcr.1601.08

Acknowledgements

This study was supported by the research program *Fonds européen de développement régional* (FEDER) 20171591-0002633 CALIBIOME 2017-2022. ARMS deployments at Reunion were conducted under permit n°2020-09-DEAL/SEB/UBIO of the *Direction de l'environnement, de l'aménagement et du logement de La Réunion*, and permit n°2020-054 of the *Direction de la mer Sud océan Indien.* Sampling was in conformity to the Nagoya protocol (declaration n°3040030). The participation to the processing of ARMS samples of Sophie Bureau, Fleur Bruggemann, the Master student Gwennaïs Fustemberg and Auriane Serval, and the BSc student Amélie Verde Ferreira was greatly appreciated. Marion Couëdel has a PhD fellowship provided by the European Union FSE programme. The authors thank all the persons who helped during field sampling and lab work. Laboratory work was made possible by the *Service de Systématique Moléculaire* of the MNHN (UAR 2700 2AD). We are grateful to Emmanuel Corse and Matthieu Leray for their advice on lab work and bioinformatics process and Eric Goberville for his comments on the statistical analyses.

Author contributions

MC, MG, and HB conceptualized the project and AD, MG, HB acquired the funding for the project. MC, MG, BF, and HB performed the field work. MC, AD and CB performed the lab work. MC performed the analyses, wrote the main manuscript text and prepared figures. All authors reviewed the manuscript.

Supplementary materials

ESM 1: ARMS deployed during this study

ARMS	Island	Region	Depth (m)	Latitude	Longitude	Deployment	Retrieval	Immersion time (months)	Deployment season	Retrieval season
RUNA2A	Reunion	La Saline	11	-21.10401	55.23598	17/12/2018	19/12/2020	24	Hot	Hot
RUNA2B	Reunion	La Saline	11	-21.10401	55.23598	17/12/2018	13/1/2021	24	Hot	Hot
RUNA2C	Reunion	La Saline	11	-21.10401	55.23598	17/12/2018	13/1/2021	24	Hot	Hot
CINA1A	Reunion	La Saline	11	-21.10401	55.23598	25/2/2020	27-31/08/2020	6	Hot	Cool
CINA1B	Reunion	La Saline	11	-21.10401	55.23598	25/2/2020	27-31/08/2020	6	Hot	Cool
CINA1C	Reunion	La Saline	11	-21.10401	55.23598	25/2/2020	27-31/08/2020	6	Hot	Cool
CINA2A	Reunion	La Saline	11	-21.10401	55.23598	25/2/2020	20/2/2021	12	Hot	Hot
CINA2B	Reunion	La Saline	11	-21.10401	55.23598	25/2/2020	22/2/2021	12	Hot	Hot
CINA2C	Reunion	La Saline	11	-21.10401	55.23598	25/2/2020	22/2/2021	12	Hot	Hot
CINA3A	Reunion	La Saline	11	-21.10401	55.23598	31/08/2020	19/2/2021	6	Cool	Cool
CINA3B	Reunion	La Saline	11	-21.10401	55.23598	31/08/2020	19/2/2021	6	Cool	Cool
CINA3C	Reunion	La Saline	11	-21.10401	55.23598	31/08/2020	20/2/2021	6	Cool	Cool
CINA4A	Reunion	La Saline	11	-21.10401	55.23598	31/08/2020	26-30/08/2021	12	Cool	Hot
CINA4B	Reunion	La Saline	11	-21.10401	55.23598	31/08/2020	26-30/08/2021	12	Cool	Hot
CINA4C	Reunion	La Saline	11	-21.10401	55.23598	31/08/2020	26-30/08/2021	12	Cool	Hot

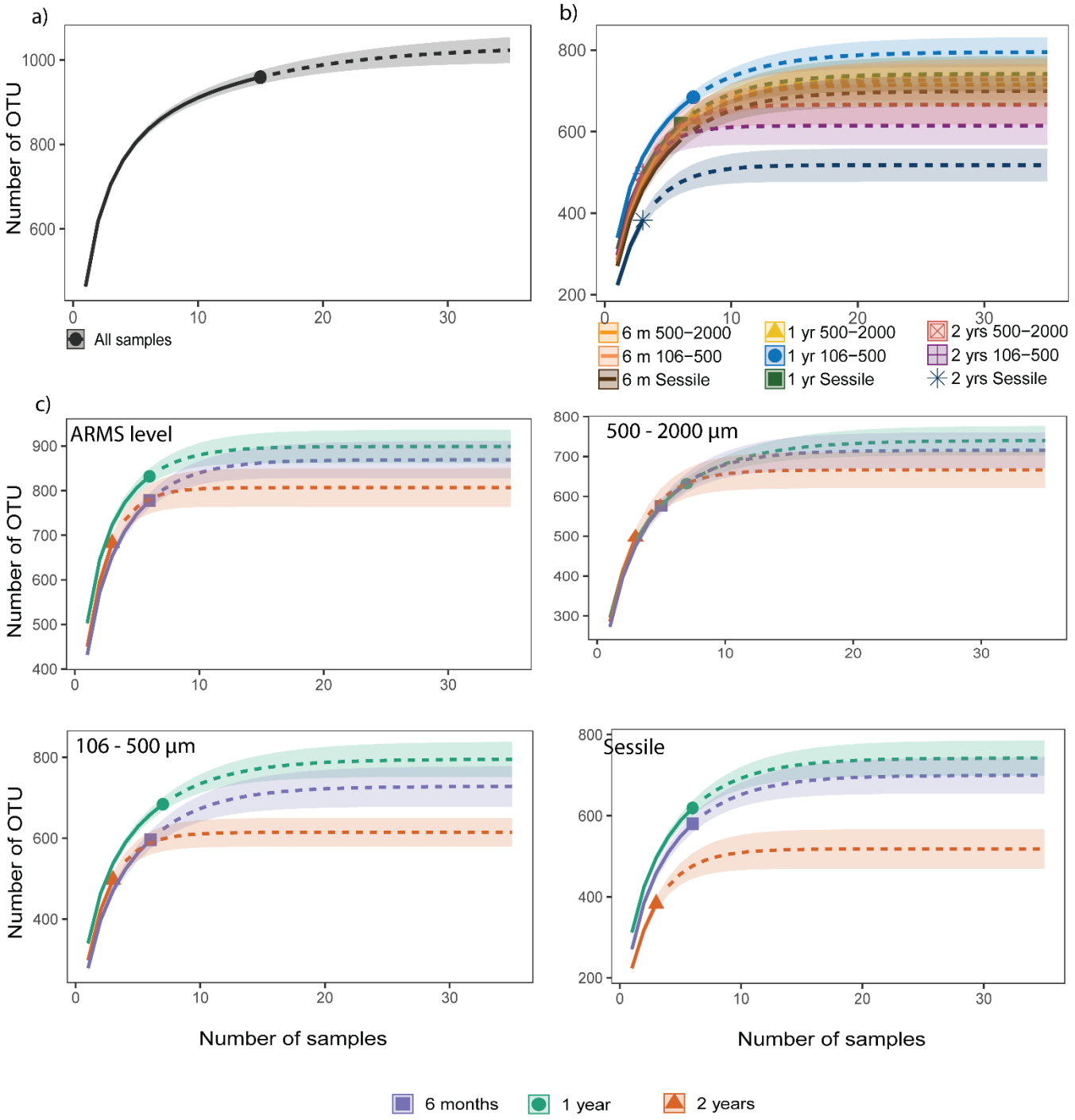
ESM 2: Deployment and sequencing information of the samples included.

SampleID	ARMS	Fraction	Marker	Deployment season	Retrieval season	Immersion time	Index	Tag	Pcr plate
REU_S1_100-500_COI	CINA1A	106-500	COI	Hot	Cool	0.5	7	1	4
REU_S1_Sessile_COI	CINA1A	Sessile	COI	Hot	Cool	0.5	7	2	4
REU_S1_100-500_18S	CINA1A	106-500	18S	Hot	Cool	0.5	7	1	4
REU_S1_Sessile_18S	CINA1A	Sessile	18S	Hot	Cool	0.5	7	2	4
REU_10_100-500_COI_doublon1	CINA4A	106-500	COI	Cool	Hot	1	2	2	1
REU_10_500-2000_COI_doublon1	CINA4A	500-2000	COI	Cool	Hot	1	2	3	1
REU_10_Sessile_COI	CINA4A	Sessile	COI	Cool	Hot	1	3	4	2
REU_10_500-2000_COI_doublon2	CINA4A	500-2000	COI	Cool	Hot	1	3	5	2
REU_10_100-500_COI_doublon2	CINA4A	106-500	COI	Cool	Hot	1	4	5	2
REU_10_100-500_18S_doublon1	CINA4A	106-500	18S	Cold	Cool	1	2	2	1
REU_10_500-2000_18S_doublon1_t3	CINA4A	500-2000	18S	Cold	Cool	1	2	3	1
REU_10_500-2000_18S_doublon1_t9	CINA4A	500-2000	18S	Cold	Cool	1	2	9	1
REU_10_Sessile_18S	CINA4A	Sessile	18S	Cold	Cool	1	3	4	2
REU_10_500-2000_18S_doublon2	CINA4A	500-2000	18S	Cold	Cool	1	3	5	2
REU_10_100-500_18S_doublon2	CINA4A	106-500	18S	Cold	Cool	1	4	5	2
REU_11_100-500_COI	CINA4B	106-500	COI	Cool	Hot	1	2	4	1
REU_11_500-2000_COI	CINA4B	500-2000	COI	Cool	Hot	1	2	5	1
REU_11_Sessile_COI	CINA4B	Sessile	COI	Cool	Hot	1	2	6	1
REU_11_100-500_18S	CINA4B	106-500	18S	Cold	Cool	1	2	4	1
REU_11_500-2000_18S	CINA4B	500-2000	18S	Cold	Cool	1	2	5	1
REU_11_Sessile_18S	CINA4B	Sessile	18S	Cold	Cool	1	2	6	1
REU_12_100-500_COI	CINA4C	106-500	COI	Cool	Hot	1	3	1	2
REU_12_500-2000_COI	CINA4C	500-2000	COI	Cool	Hot	1	3	2	2
REU_12_Sessile_COI	CINA4C	Sessile	COI	Cool	Hot	1	3	3	2
REU_12_100-500_18S	CINA4C	106-500	18S	Cold	Cool	1	3	1	2
REU_12_500-2000_18S	CINA4C	500-2000	18S	Cold	Cool	1	3	2	2
REU_12_Sessile_18S_t9	CINA4C	Sessile	18S	Cold	Cool	1	3	9	2

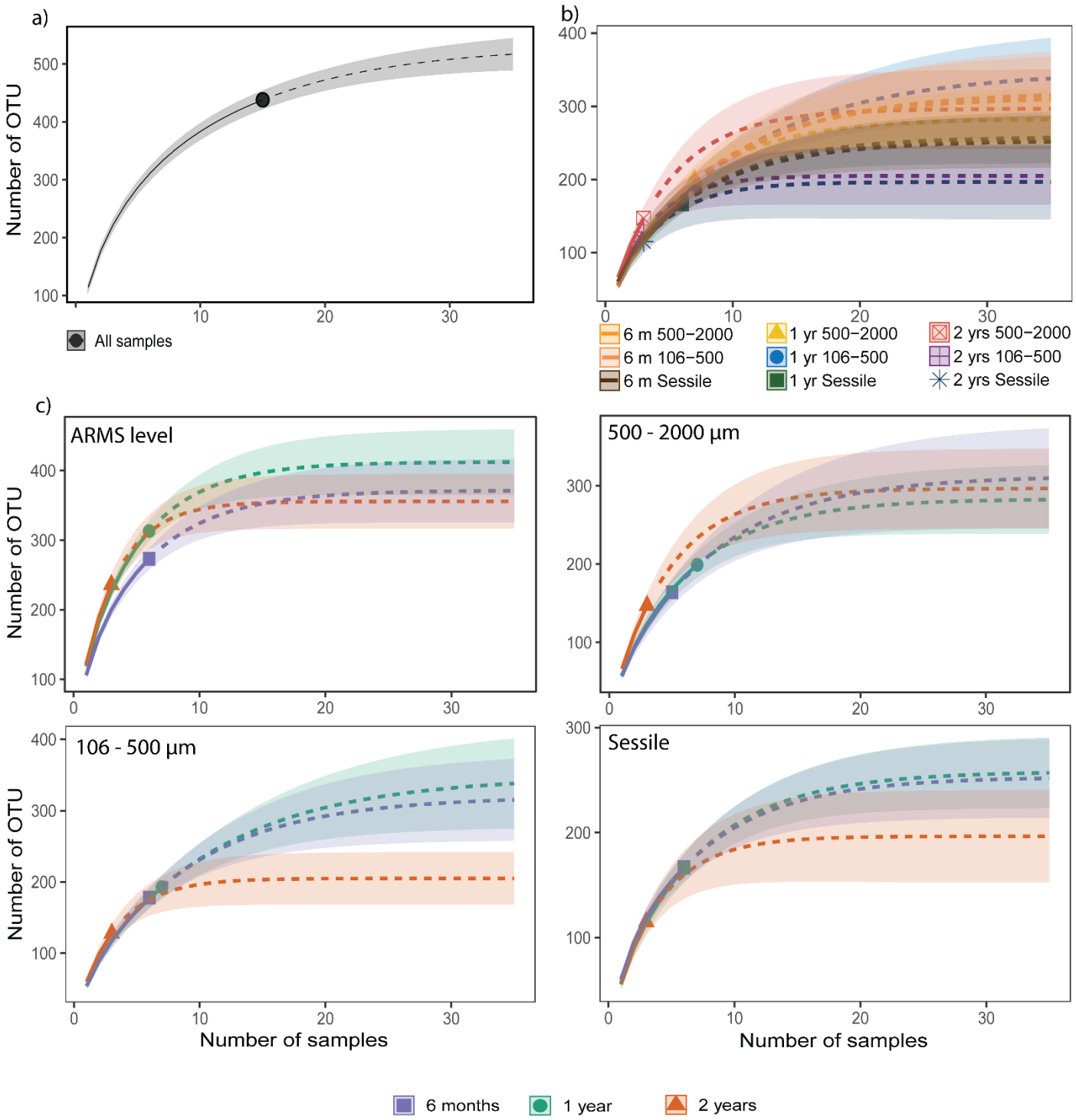
SampleID	ARMS	Fraction	Marker	Deployment season	Retrieval season	Immersion time	Index	Tag	Pcr plate
REU_S2_100-500_COI	CINA1B	106-500	COI	Hot	Cool	0.5	7	3	4
REU_S2_500-2000_COI	CINA1B	500-2000	COI	Hot	Cool	0.5	7	5	4
REU_S2_Sessile_COI	CINA1B	Sessile	COI	Hot	Cool	0.5	7	6	4
REU_S2_100-500_18S_t9	CINA1B	106-500	18S	Hot	Cool	0.5	7	9	4
REU_S2_500-2000_18S	CINA1B	500-2000	18S	Hot	Cool	0.5	7	5	4
REU_S2_Sessile_18S	CINA1B	Sessile	18S	Hot	Cool	0.5	7	6	4
REU_S3_Sessile_COI	CINA1C	Sessile	COI	Hot	Cool	0.5	8	1	4
REU_S3_100-500_COI	CINA1C	106-500	COI	Hot	Cool	0.5	8	2	4
REU_S3_Sessile_18S	CINA1C	Sessile	18S	Hot	Cool	0.5	8	1	4
REU_S3_100-500_18S	CINA1C	106-500	18S	Hot	Cool	0.5	8	2	4
REU_S3_500-2000_18S	CINA1C	500-2000	18S	Hot	Cool	0.5	14	5	8
REU_S4_Sessile_COI	CINA2A	Sessile	COI	Hot	Hot	1	5	3	3
REU_S4_500-2000_COI	CINA2A	500-2000	COI	Hot	Hot	1	5	4	3
REU_S4_100-500_COI	CINA2A	106-500	COI	Hot	Hot	1	5	5	3
REU_S4_Sessile_18S_t9	CINA2A	Sessile	18S	Hot	Hot	1	5	9	3
REU_S4_500-2000_18S	CINA2A	500-2000	18S	Hot	Hot	1	5	4	3
REU_S4_100-500_18S	CINA2A	106-500	18S	Hot	Hot	1	5	5	3
REU_S5_100-500_COI	CINA2B	106-500	COI	Hot	Hot	1	5	2	3
REU_S5_500-2000_COI	CINA2B	500-2000	COI	Hot	Hot	1	5	6	3
REU_S5_Sessile_COI	CINA2B	Sessile	COI	Hot	Hot	1	6	1	3
REU_S5_100-500_18S	CINA2B	106-500	18S	Hot	Hot	1	5	2	3
REU_S5_500-2000_18S	CINA2B	500-2000	18S	Hot	Hot	1	5	6	3
REU_S5_Sessile_18S	CINA2B	Sessile	18S	Hot	Hot	1	6	1	3
REU_S6_500-2000_COI	CINA2C	500-2000	COI	Hot	Hot	1	6	2	3
REU_S6_Sessile_COI	CINA2C	Sessile	COI	Hot	Hot	1	6	4	3
REU_S6_100-500_COI	CINA2C	106-500	COI	Hot	Hot	1	6	5	3
REU_S6_500-2000_18S	CINA2C	500-2000	18S	Hot	Hot	1	6	2	3
REU_S6_Sessile_18S	CINA2C	Sessile	18S	Hot	Hot	1	6	4	3
REU_S6_100-500_18S	CINA2C	106-500	18S	Hot	Hot	1	6	5	3

SampleID	ARMS	Fraction	Marker	Deployment season	Retrieval season	Immersion time	Index	Tag	Pcr plate
REU_S7_100-500_COI	CINA3A	106-500	COI	Cool	Cool	0.5	8	3	4
REU_S7_500-2000_COI	CINA3A	500-2000	COI	Cool	Cool	0.5	8	4	4
REU_S7_Sessile_COI	CINA3A	Sessile	COI	Cool	Cool	0.5	8	5	4
REU_S7_100-500_18S_t9	CINA3A	106-500	18S	Cold	Hot	0.5	8	9	4
REU_S7_500-2000_18S	CINA3A	500-2000	18S	Cold	Hot	0.5	8	4	4
REU_S7_Sessile_18S	CINA3A	Sessile	18S	Cold	Hot	0.5	8	5	4
REU_S8_Sessile_COI	CINA3B	Sessile	COI	Cool	Cool	0.5	4	6	2
REU_S8_500-2000_COI	CINA3B	500-2000	COI	Cool	Cool	0.5	5	1	3
REU_S8_100-500_COI	CINA3B	106-500	COI	Cool	Cool	0.5	8	6	4
REU_S8_Sessile_18S	CINA3B	Sessile	18S	Cold	Hot	0.5	4	6	2
REU_S8_500-2000_18S	CINA3B	500-2000	18S	Cold	Hot	0.5	5	1	3
REU_S8_100-500_18S	CINA3B	106-500	18S	Cold	Hot	0.5	8	6	4
REU_S9_100-500_COI	CINA3C	106-500	COI	Cool	Cool	0.5	4	1	2
REU_S9_Sessile_COI	CINA3C	Sessile	COI	Cool	Cool	0.5	4	3	2
REU_S9_500-2000_COI	CINA3C	500-2000	COI	Cool	Cool	0.5	4	4	2
REU_S9_100-500_18S	CINA3C	106-500	18S	Cold	Hot	0.5	4	1	2
REU_S9_Sessile_18S_t9	CINA3C	Sessile	18S	Cold	Hot	0.5	4	9	2
REU_S9_500-2000_18S	CINA3C	500-2000	18S	Cold	Hot	0.5	4	4	2
REU_2A_100-500_COI	RUNA2A	106-500	COI	Hot	Hot	2	10	6	5
REU_2A_500-2000_COI	RUNA2A	500-2000	COI	Hot	Hot	2	11	1	6
REU_2A_Sessile_COI	RUNA2A	Sessile	COI	Hot	Hot	2	11	2	6
REU_2A_100-500_18S	RUNA2A	106-500	18S	Hot	Hot	2	10	6	5
REU_2A_500-2000_18S	RUNA2A	500-2000	18S	Hot	Hot	2	11	1	6
REU_2A_Sessile_18S	RUNA2A	Sessile	18S	Hot	Hot	2	11	2	6
REU_2B_100-500_COI	RUNA2B	106-500	COI	Hot	Hot	2	1	1	1
REU_2B_500-2000_COI	RUNA2B	500-2000	COI	Hot	Hot	2	1	2	1
REU_2B_Sessile_COI	RUNA2B	Sessile	COI	Hot	Hot	2	1	3	1
REU_2B_100-500_18S	RUNA2B	106-500	18S	Hot	Hot	2	1	1	1
REU_2B_500-2000_18S	RUNA2B	500-2000	18S	Hot	Hot	2	1	2	1

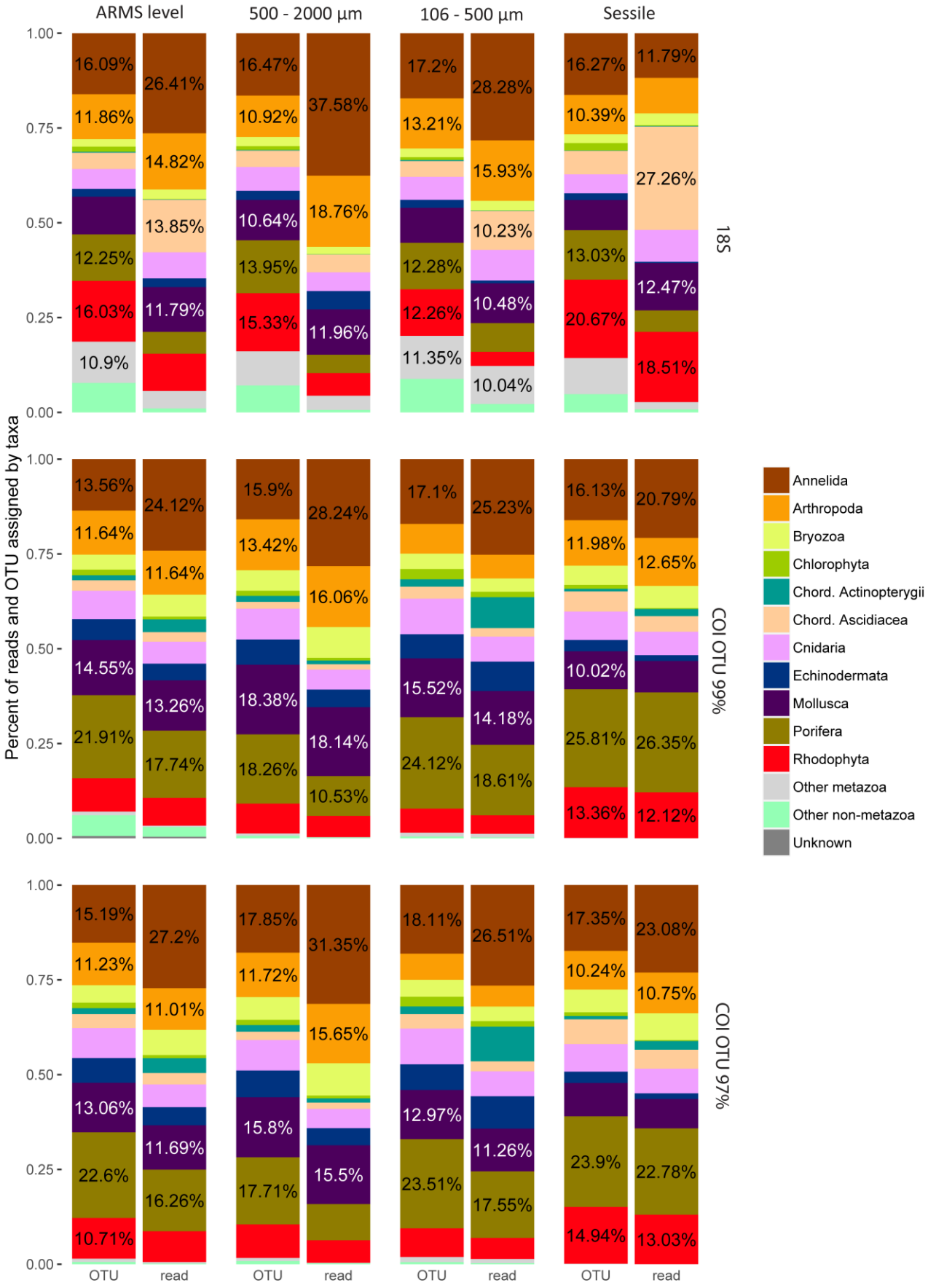
SampleID	ARMS	Fraction	Marker	Deployment season	Retrieval season	Immersion time	Index	Tag	Pcr plate
REU_2B_Sessile_18S_t9	RUNA2B	Sessile	18S	Hot	Hot	2	1	9	1
REU_2C_100-500_COI	RUNA2C	106-500	COI	Hot	Hot	2	1	4	1
REU_2C_500-2000_COI	RUNA2C	500-2000	COI	Hot	Hot	2	1	5	1
REU_2C_Sessile_COI	RUNA2C	Sessile	COI	Hot	Hot	2	1	6	1
REU_2C_100-500_18S	RUNA2C	106-500	18S	Hot	Hot	2	1	4	1
REU_2C_500-2000_18S	RUNA2C	500-2000	18S	Hot	Hot	2	1	5	1
REU_2C_Sessile_18S	RUNA2C	Sessile	18S	Hot	Hot	2	1	6	1



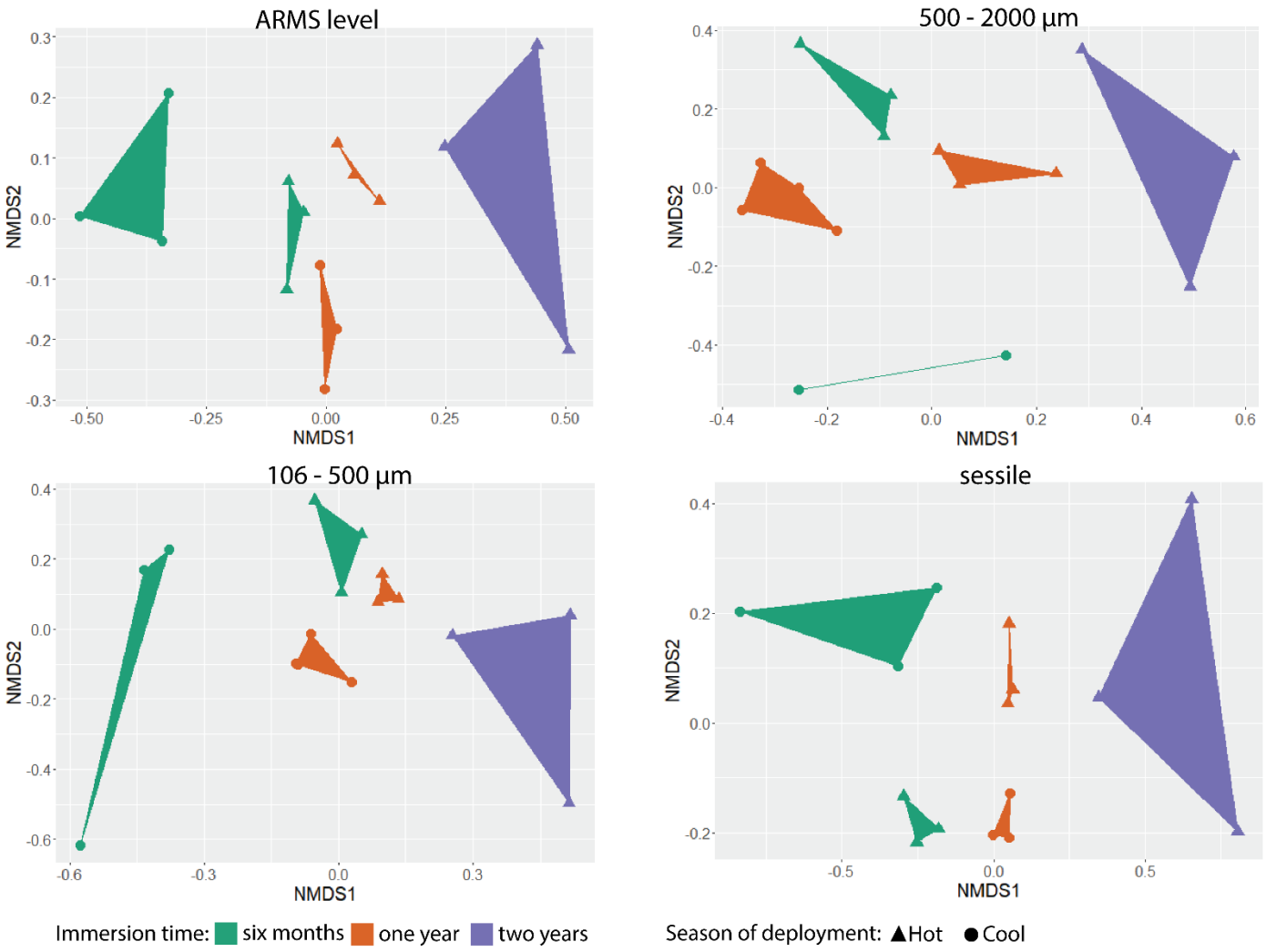
ESM 3: OTUs accumulation curves for the 18S for a): all samples; b) samples by modalities and c) samples by times for each fraction datasets.



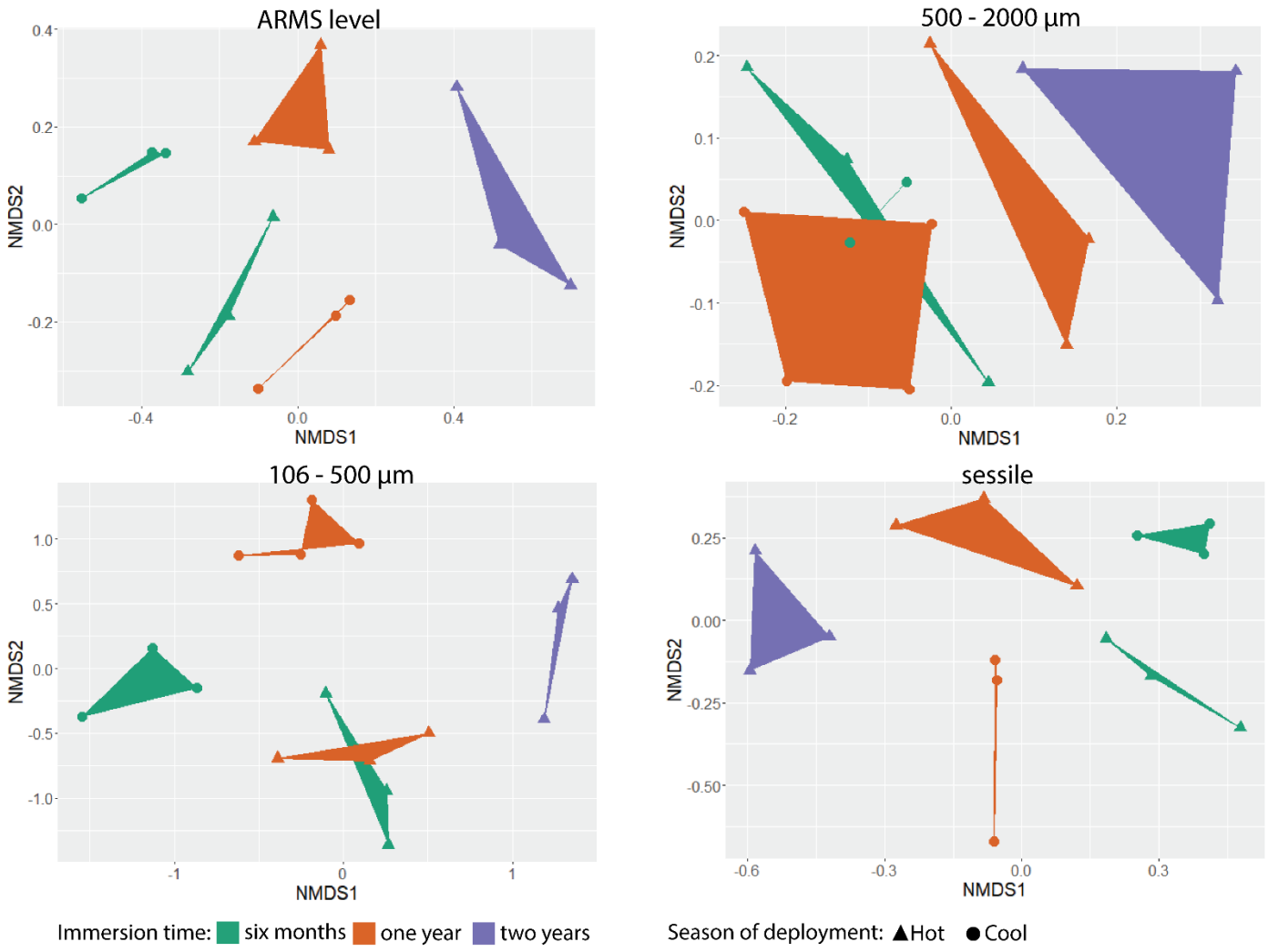
ESM 4 : OTUs accumulation curves for the COI OTU99% for a): all samples; b) samples by modalities and c) samples by times for each fraction datasets.



ESM 5: Percent of reads and OTU assigned by taxa category for each primer (horizontal blocks) and each fraction (vertical blocks).



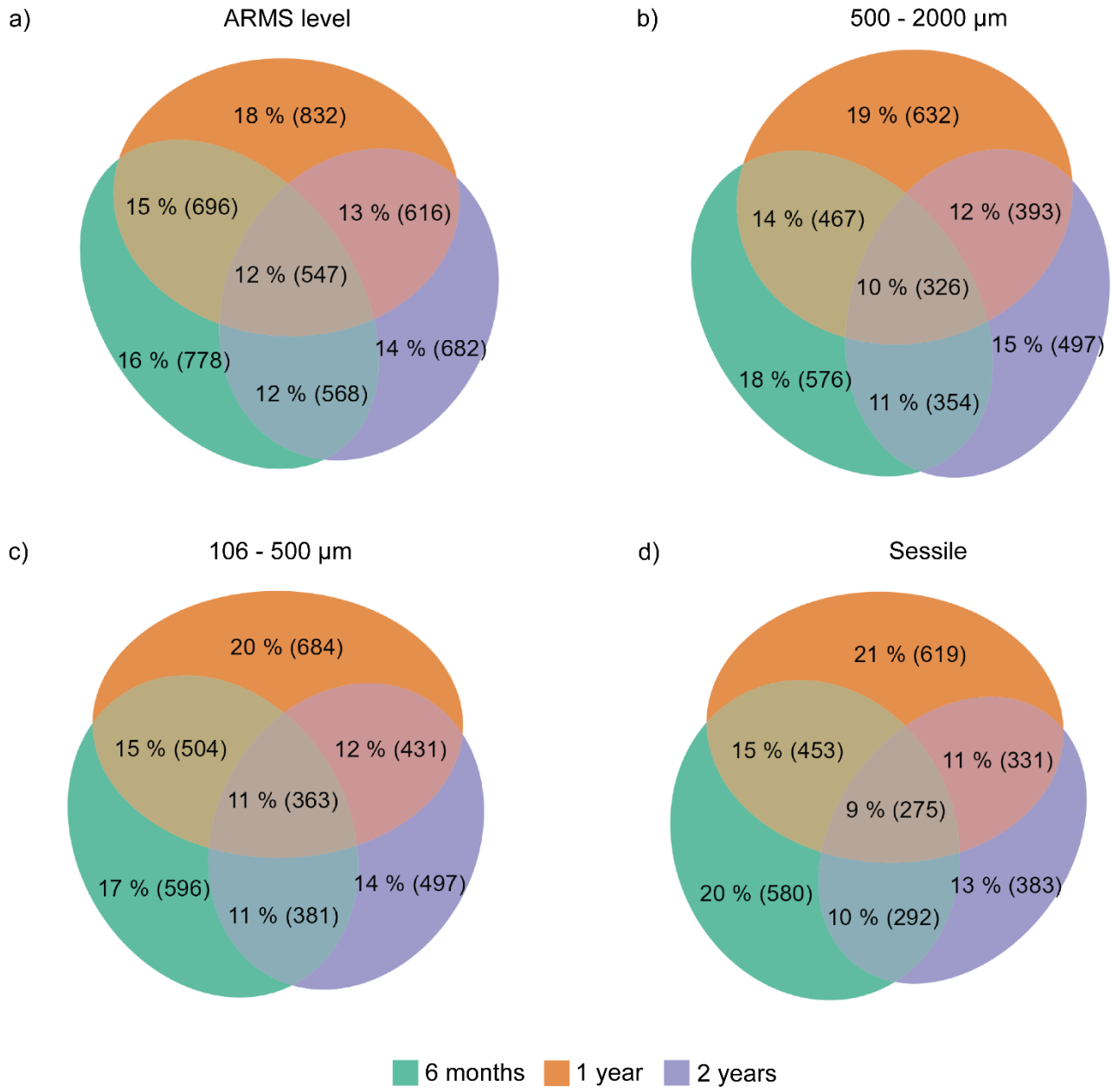
ESM 6: Non-metric multidimensional scaling (nMDS) ordination plots dissimilarities in community composition based Chao estimation of 18S. Analysis was undertaken on the full ARMS as well as the different fractions (500 - 2000 µm, 106 - 500 µm, sessile). Points are coloured according to season of deployment.



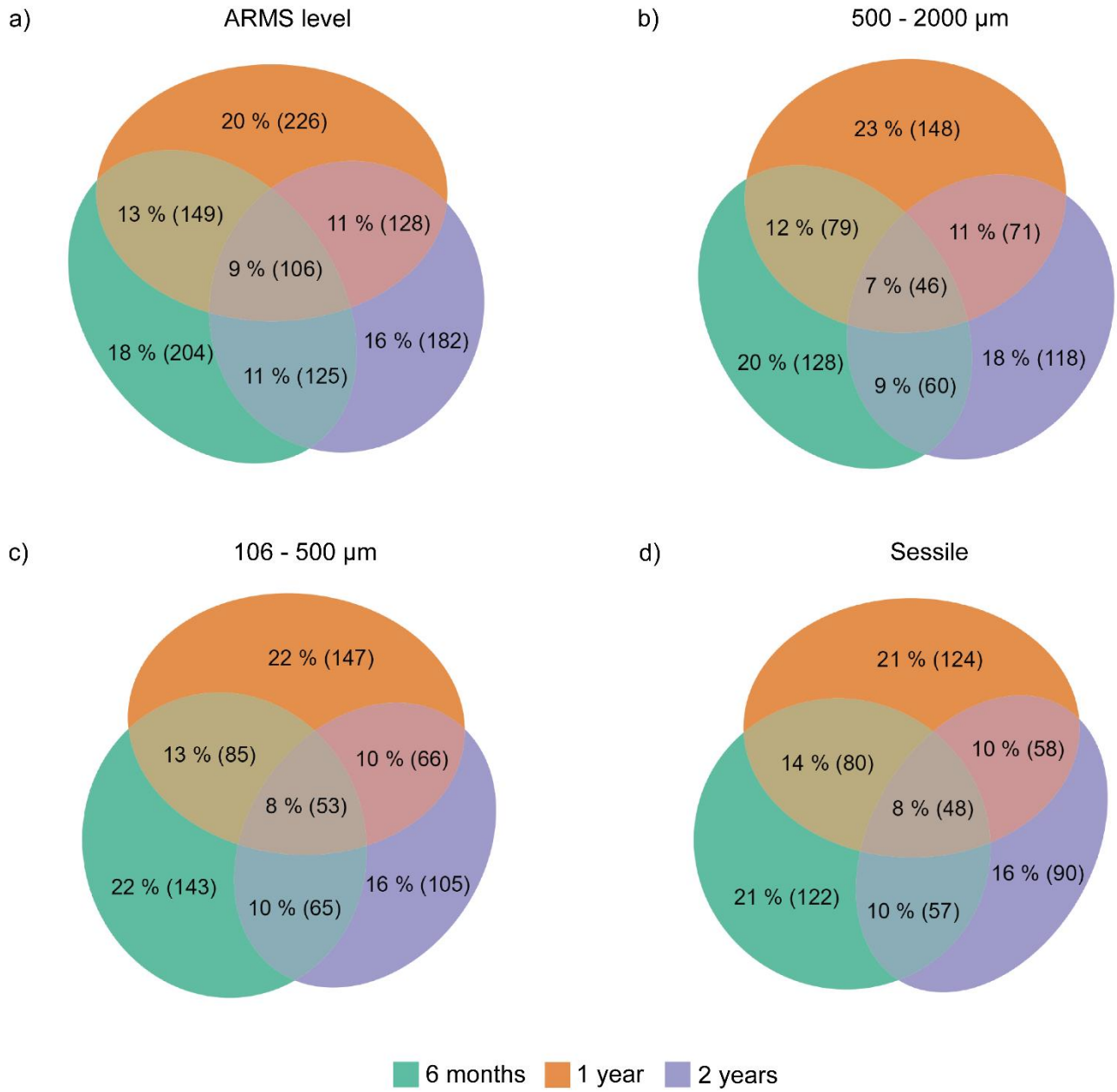
ESM 7: Non-metric multidimensional scaling (nMDS) ordination plots dissimilarities in community composition based Chao estimation of COI99. Analysis was undertaken on the full ARMS as well as the different fractions (500 - 2000 µm, 106 - 500 µm, sessile). Points are coloured according to season of deployment.

ESM 8: Summary of the SIMPER analyses performed by marker for each comparison (in column) and for each dataset (in row). For each phylum, the number of OTU contributing to 50% of the difference is indicated. Coloured backgrounds represent the evolution of read abundance for the corresponding OTU between the two compared levels (green: increase; yellow: no distinct pattern; red: decrease).

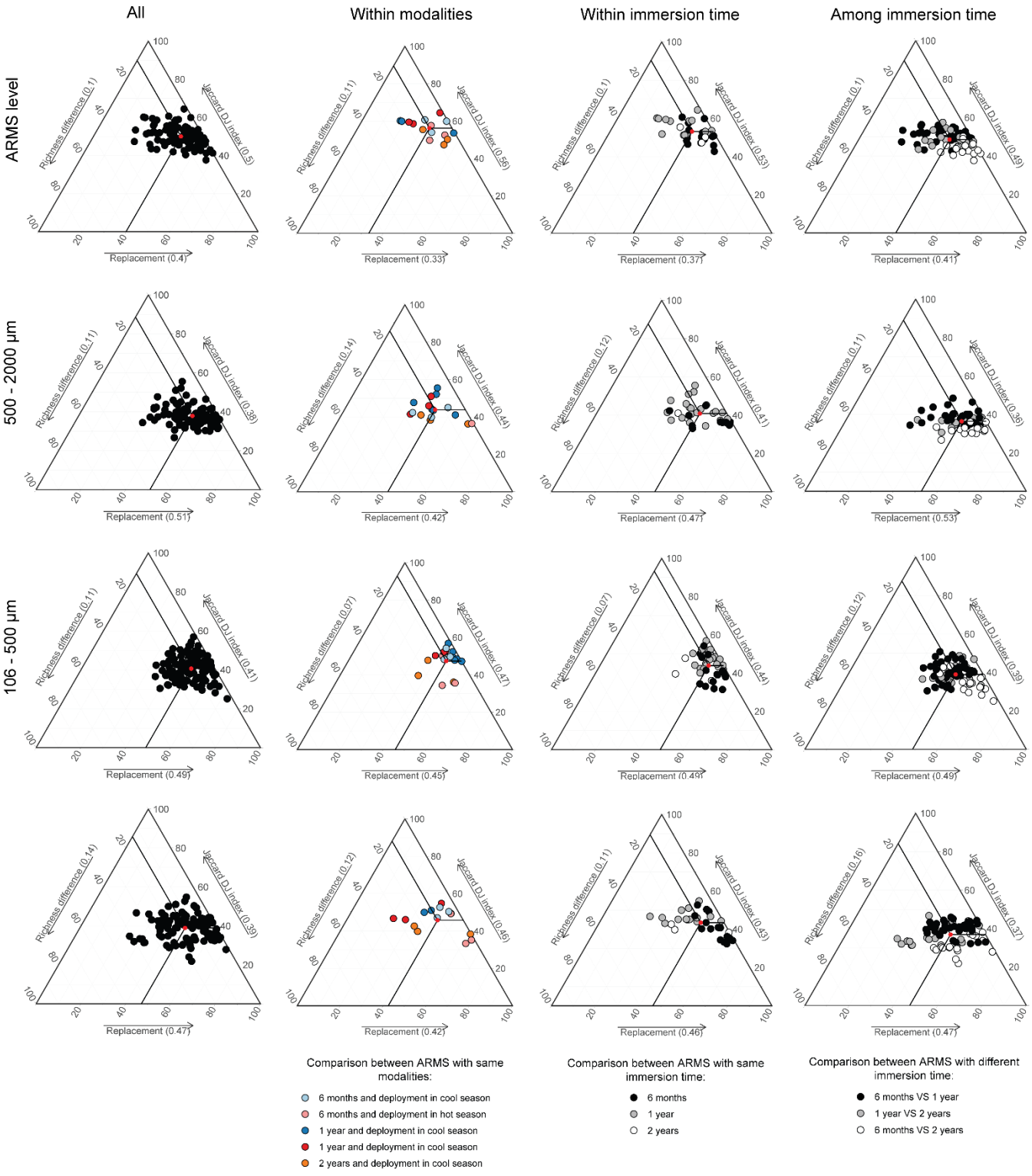
		18S					COI				
		Time			Season (Cold VS Warm)		Time			Season (Cold VS Warm)	
	Phylum	0,5-1	0,5-2	1-2	of collect	of deployement	0,5-1	0,5-2	1-2	of collect	of deployement
ARMS level	Annelida	5	3	3	**6**	7	**12**	**12**	10	14	13
	Arthropoda	3	**4**	**5**	4	4	7	6	6	6	7
	Bryozoa	1	0	0	1	1	3	2	3	3	3
	Chlorophyta	0	0	0	0	0	0	0	1	1	1
	Chordata	**7**	2	3	5	5	2	4	4	3	2
	Cnidaria	2	2	1	2	2	3	3	2	3	4
	Echinodermata	0	1	1	1	1	2	1	2	3	2
	Mollusca	2	3	4	3	3	5	7	9	7	7
	Nematoda	0	0	0	0	0	0	0	0	0	0
	Nemertea	0	0	0	0	0	0	0	0	0	0
	Porifera	1	0	0	1	1	6	9	14	6	7
	Rhodophyta	2	**4**	4	3	2	4	2	4	3	3
	# OTU which contribute for 50% of the difference	23	19	21	26	26	44	46	45	35	36
	Overall average dissimilarity:	69.18	79.18	72.70	71.05	70.03	66.73	77.07	70.44	69.26	69.75
	p-value	0.003**	0.033*	0.03*	0.001***	0.005**	0.015*	0.045*	0.03*	0.009**	0.007**
		Time			Season (Cold VS Warm)		Time			Season (Cold VS Warm)	
	Phylum	0,5-1	0,5-2	1-2	of collect	of deployement	0,5-1	0,5-2	1-2	of collect	of deployement
500 -2000 um	Annelida	4	2	2	5	4	9	7	10	9	9
	Arthropoda	3	3	4	3	3	4	5	5	5	5
	Bryozoa	0	0	0	0	0	2	2	2	3	3
	Chlorophyta	0	0	0	0	0	0	0	0	1	0
	Chordata	2	0	0	2	2	1	0	1	1	1
	Cnidaria	1	0	1	1	1	2	1	0	1	1
	Echinodermata	0	1	1	1	1	1	0	1	1	2
	Mollusca	1	2	2	2	3	7	8	9	8	8
	Nematoda	0	0	0	0	0					
	Nemertea	0	0	0	0	0					
	Porifera	1	0	0	0	0	1	3	4	2	2
	Rhodophyta	1	1	0	1	1	2	2	1	3	3
	# OTU which contribute for 50% of the difference	13	9	10	15	15	29	28	23	25	25
	Overall average dissimilarity:	72.79	84.78	80.64	75.84	75.12	70.52	78.96	76.02	72.04	72.57
	p-value	0.036*	0.252 NS	0.027*	0.002**	0.003**	0.012*	0.063*	0.045*	0.031*	0.026*
		Time			Season (Cold VS Warm)		Time			Season (Cold VS Warm)	
	Phylum	0,5-1	0,5-2	1-2	of collect	of deployement	0,5-1	0,5-2	1-2	of collect	of deployement
106 - 500 um	Annelida	10	8	8	9	9	11	9	9	12	12
	Arthropoda	1	4	2	2	2	1	3	4	1	1
	Bryozoa	0	1	1	0	1	1	0	1	1	1
	Chlorophyta	0	0	0	0	0	1	0	1	1	1
	Chordata	5	3	3	6	6	2	2	2	2	2
	Cnidaria	3	2	1	3	3	2	2	2	3	4
	Echinodermata	0	0	0	0	0	3	5	4	4	4
	Mollusca	1	1	1	1	1	4	2	4	5	5
	Nematoda	0	1	1	0	0	0	0	0	0	0
	Nemertea	2	2	2	2	2	0	0	0	0	0
	Porifera	3	1	2	3	3	3	7	11	5	5
	Rhodophyta	0	0	0	0	0	2	2	2	4	4
	# OTU which contribute for 50% of the difference	25	23	21	26	27	23	26	32	30	31
	Overall average dissimilarity:	69.82	80.44	73.52	72.97	70.67	81.33	87.81	84.44	83.46	81.77
	p-value	0.018*	0.078 NS	0.159 NS	0.001***	0.001***	0.018*	0.054 NS	0.075 NS	0.001***	0.146 NS
		Time			Season (Cold VS Warm)		Time			Season (Cold VS Warm)	
	Phylum	0,5-1	0,5-2	1-2	of collect	of deployement	0,5-1	0,5-2	1-2	of collect	of deployement
Sessile	Annelida	0	1	1	1	1	11	8	6	13	11
	Arthropoda	1	2	1	1	1	4	4	5	4	4
	Bryozoa	1	0	0	1	1	2	1	1	2	2
	Chlorophyta	0	0	0	0	0	0	0	0	0	0
	Chordata	8	3	4	7	6	3	3	2	3	2
	Cnidaria	2	1	1	2	2	3	3	2	3	3
	Echinodermata	0	0	0	0	0	0	0	0	0	0
	Mollusca	1	2	2	2	1	3	3	1	1	2
	Nematoda	0	0	0	0	0	0	0	0	0	0
	Nemertea	0	0	0	0	0	0	0	0	0	0
	Porifera	0	0	0	0	0	7	12	14	11	12
	Rhodophyta	2	5	4	3	3	3	3	3	3	3
	# OTU which contribute for 50% of the difference	15	14	13	17	15	36	37	34	40	39
	Overall average dissimilarity:	76.23	85.12	82.11	78.26	78.96	74.26	84.05	76.45	75.05	76.01
	p-value	0.027*	0.063 NS	0.063 NS	0.063 NS	0.01*	0.039*	0.03*	0.051 NS	0.27 NS	0.003**



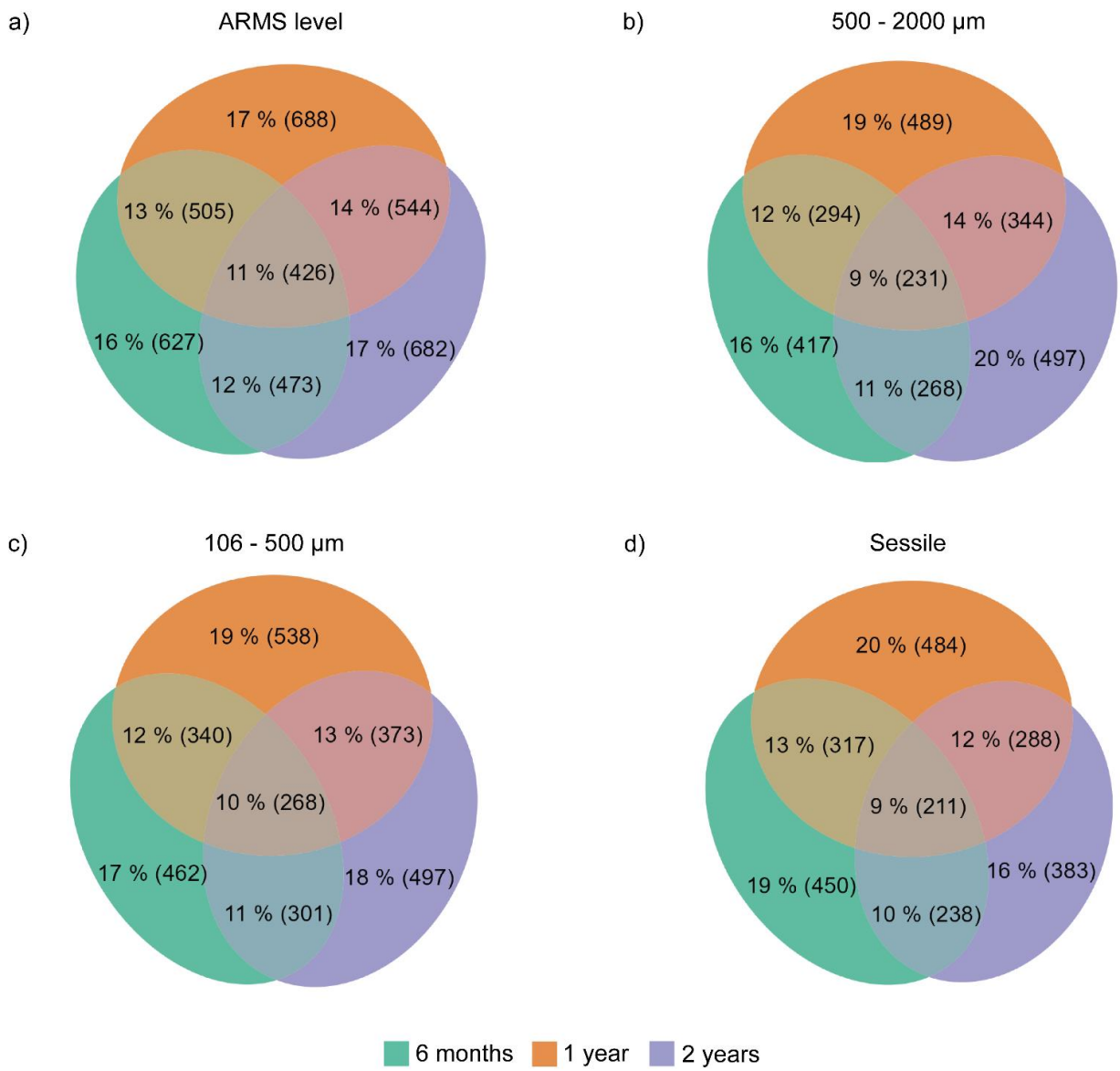
ESM 9: Number and proportion of unique and shared 18S marker amongst the three immersion times for the ARMS level dataset and the three fractions datasets. Ellipses are proportional.



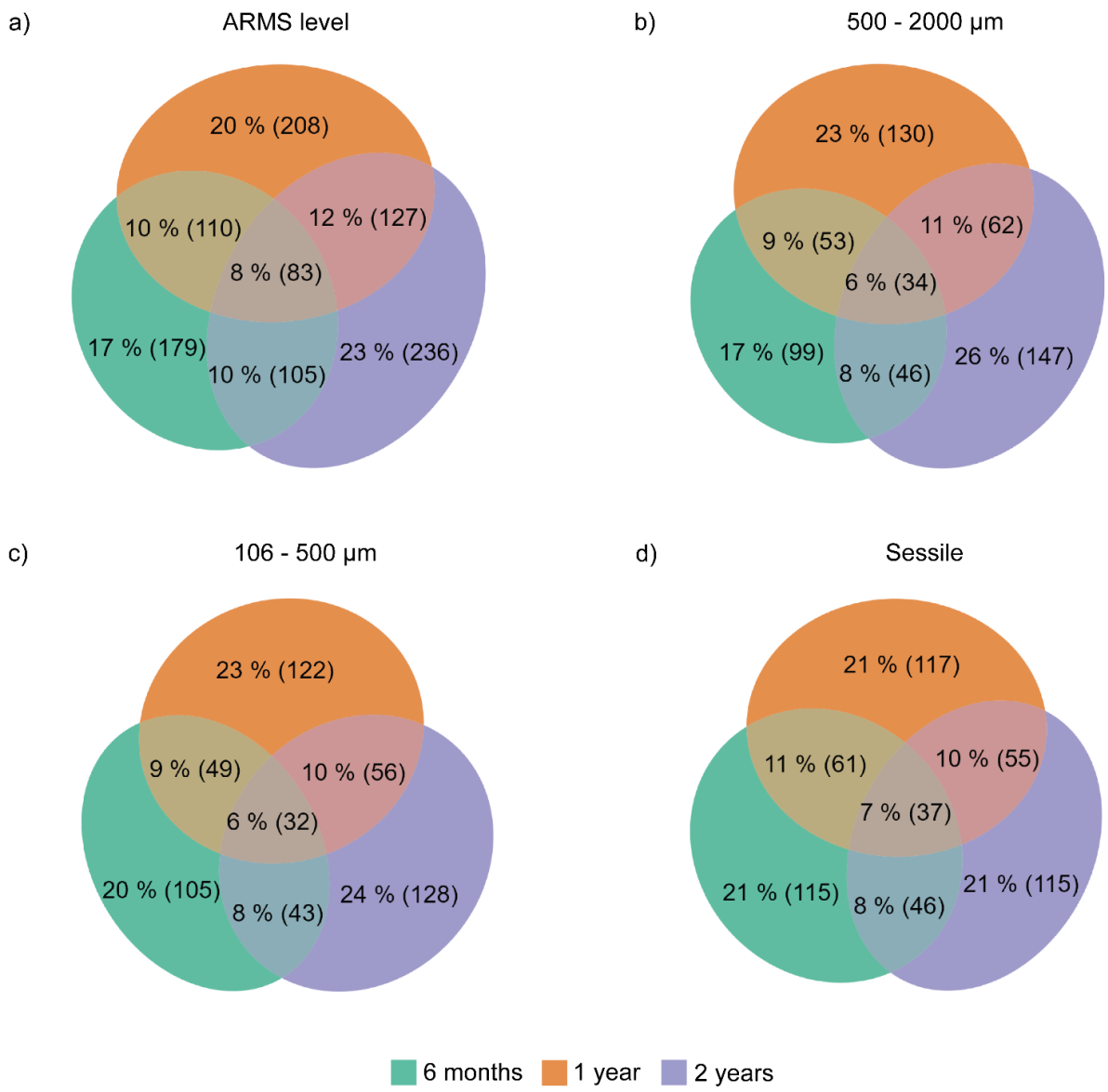
ESM 10: Number and proportion of unique and shared OTU97% for the COI marker amongst the three immersion times for the ARMS level dataset and the three fractions datatsets. Ellipses are proportional.



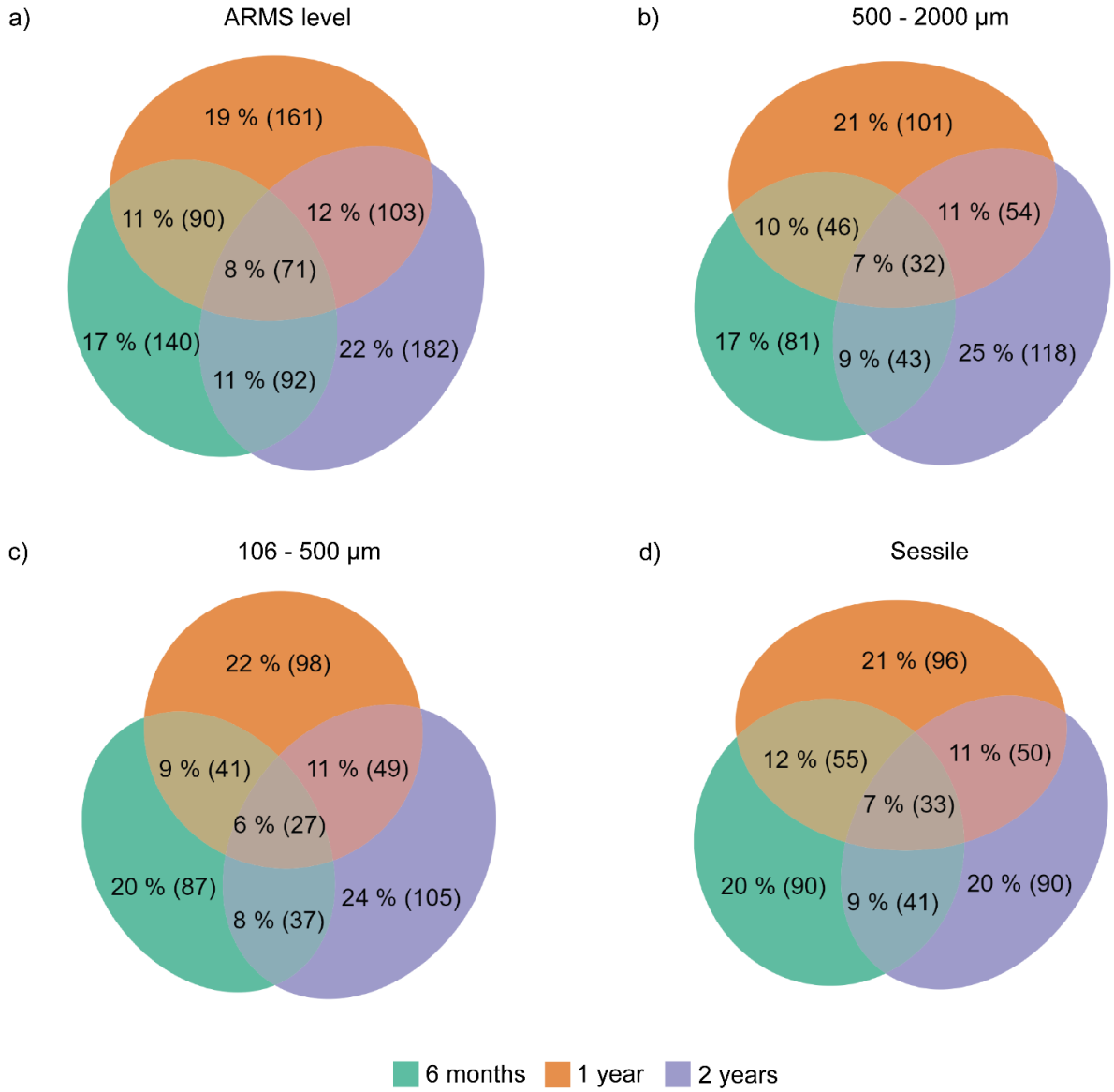
ESM 11: Ternary plots of Jaccard similarity and the partitions of beta diversity (replacement and richness) for the full ARMS unit (ARMS level) and the three fractions obtained from 18S metabarcoding. Ternary plots are shown for the total experiment (All) as well as within modalities and within and among immersion times. Red dot and numbers in brackets on the axis labels represent the mean value.



ESM 12: Number and proportion of unique and shared for the 18S marker amongst the three immersion times for ARMS retrieved at the same period (Summer December 2020 and January 2021) for the ARMS level dataset and the three fractions datasets. Ellipses are proportional.



ESM 13: Number and proportion of unique and shared OTU99% for the COI marker amongst the three immersion times for ARMS retrieved at the same period (Summer December 2020 and January 2021) for the ARMS level dataset and the three fractions datasets. Ellipses are proportional.



ESM 14: Number and proportion of unique and shared OTU97% for the COI marker amongst the three immersion times for ARMS retrieved at the same period (Summer December 2020 and January 2021) for the ARMS level dataset and the three fractions datasets. Ellipses are proportional.

ESM 15: Jaccard similarity and the partitions of beta diversity (replacement and richness) for the full ARMS unit (ARMS level) and the three fractions obtained from the OTU99% of the COI and 18S.

ESM 15.1: For intra-modalities comparisons

	Immersion time	Deployement season	COI			18S		
			Similarity	Replacement	Richness	Similarity	Replacement	Richness
ARMS level	6 months	Cool	0.352	0.574	0.0746	0.582	0.346	0.0723
	6 months	Hot	0.387	0.459	0.154	0.533	0.374	0.0929
	1 year	Cool	0.36	0.544	0.0961	0.579	0.274	0.147
	1 year	Hot	0.335	0.551	0.115	0.608	0.269	0.123
	2 years	Hot	0.294	0.536	0.169	0.51	0.399	0.0907
500 - 2000 µm	6 months	Cool	0.22	0.645	0.135	0.424	0.408	0.168
	6 months	Hot	0.343	0.586	0.0707	0.365	0.628	0.00719
	1 year	Cool	0.237	0.722	0.0407	0.473	0.402	0.125
	1 year	Hot	0.207	0.704	0.089	0.461	0.355	0.183
	3 years	Hot	0.2	0.554	0.246	0.385	0.469	0.146
106 - 500 µm	6 months	Cool	0.207	0.679	0.114	0.508	0.441	0.0514 ab
	6 months	Hot	0.28	0.601	0.119	0.4	0.493	0.106 bc
	1 year	Cool	0.268	0.588	0.145	0.503	0.479	0.0176 a
	1 year	Hot	0.218	0.693	0.0896	0.517	0.411	0.0718 ac
	4 years	Hot	0.208	0.619	0.172	0.412	0.426	0.162 c
Sessile	6 months	Cool	0.278	0.499	0.222	0.501	0.422	0.0775
	6 months	Hot	0.343	0.599	0.0574	0.393	0.572	0.0354
	1 year	Cool	0.251	0.606	0.143	0.501	0.391	0.107
	1 year	Hot	0.248	0.674	0.0787	0.492	0.294	0.214
	5 years	Hot	0.275	0.525	0.2	0.403	0.434	0.163

ESM 15.2: for intra-time comparisons

	Immersion time	COI			18S		
		Similarity	Replacement	Richness	Similarity	Replacement	Richness
ARMS level	6 months	0.317	0.585	0.0976	0.511 a	0.415 a	0.0742
	1 year	0.311	0.597	0.0923	0.557 b	0.322 b	0.12
	2 years	0.294	0.536	0.169	0.51 ab	0.399 ab	0.0907
500 - 2000 µm	6 months	0.247	0.616 a	0.136 a	0.375 a	0.522	0.104
	1 year	0.205	0.733 b	0.0615 b	0.427 b	0.444	0.129
	2 years	0.2	0.554 a	0.246 a	0.385 ab	0.469	0.146
106 - 500 µm	6 months	0.203	0.702	0.0954	0.405 a	0.526 a	0.0686 a
	1 year	0.21	0.675	0.115	0.468 b	0.479 ab	0.0528 a
	2 years	0.208	0.619	0.172	0.412 ab	0.426 b	0.162 b
Sessile	6 months	0.269	0.605	0.126	0.407 a	0.538 a	0.0544 a
	1 year	0.239	0.662	0.0994	0.469 b	0.386 b	0.146 b
	2 years	0.275	0.525	0.2	0.403 ab	0.434 ab	0.163 ab

ESM 15.3: For inter-time comparisons

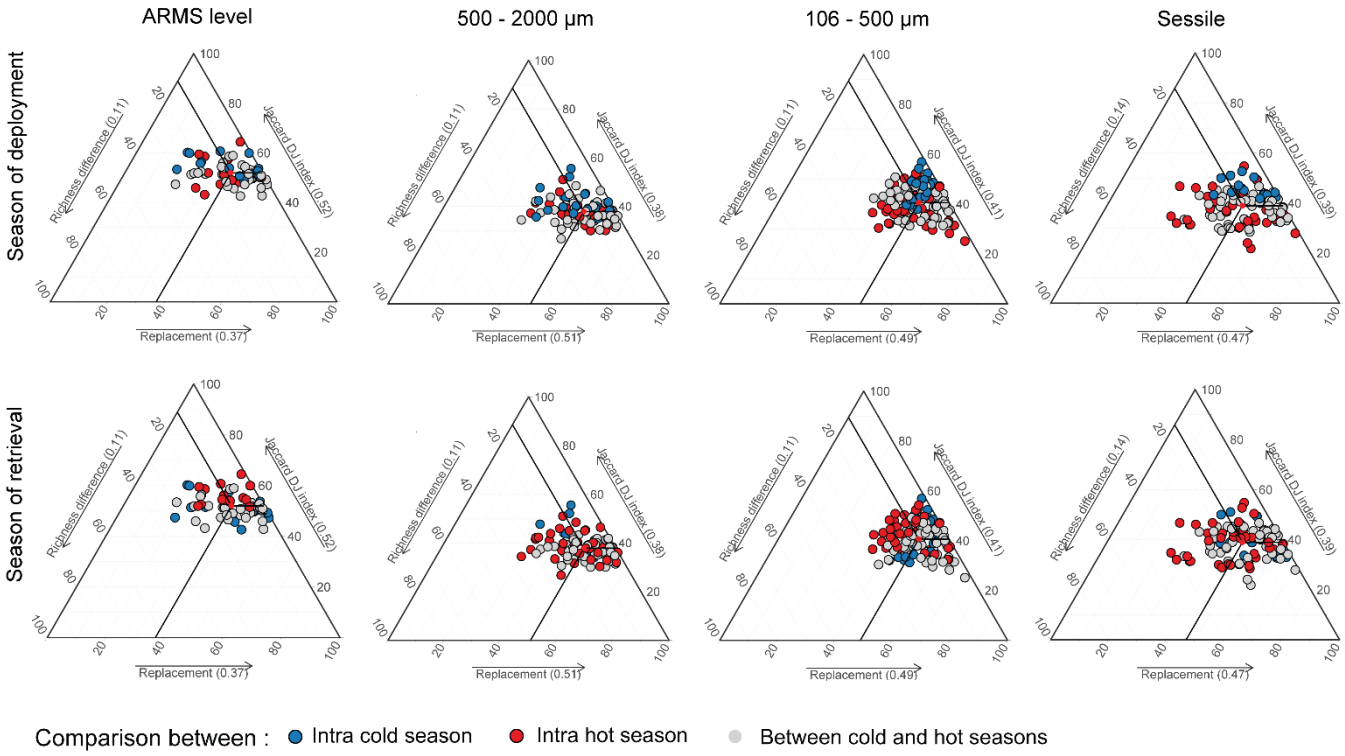
	Comparisons	COI			18S		
		Similarity	Replacement	Richness	Similarity	Replacement	Richness
ARMS level	6 months VS 1 year	0.301 a	0.597	0.101	0.508 a	0.37 a	0.122
	1 year VS 2 years	0.268 b	0.622	0.11	0.489 a	0.406 a	0.105
	6 months VS 2 years	0.243 b	0.641	0.117	0.439 b	0.496 b	0.0649
500 - 2000 µm	6 months VS 1 year	0.208 a	0.698	0.0942 a	0.385 a	0.502 a	0.113
	1 year VS 2 years	0.175 b	0.689	0.136 ab	0.359 b	0.532 ab	0.109
	6 months VS 2 years	0.162 b	0.673	0.165 b	0.325 c	0.571 b	0.103
106 - 500 µm	6 months VS 1 year	0.181	0.727	0.0917	0.407 a	0.455 a	0.138
	1 year VS 2 years	0.167	0.713	0.119	0.395 a	0.484 a	0.122
	6 months VS 2 years	0.159	0.73	0.111	0.343 b	0.562 b	0.0953
Sessile	6 months VS 1 year	0.232 a	0.646	0.122	0.4 a	0.475 a	0.125 a
	1 year VS 2 years	0.213 ab	0.661	0.125	0.367 b	0.4 b	0.233 b
	6 months VS 2 years	0.183 b	0.66	0.158	0.309 c	0.543 c	0.148 a

ESM 15.4: For season of deployment comparisons

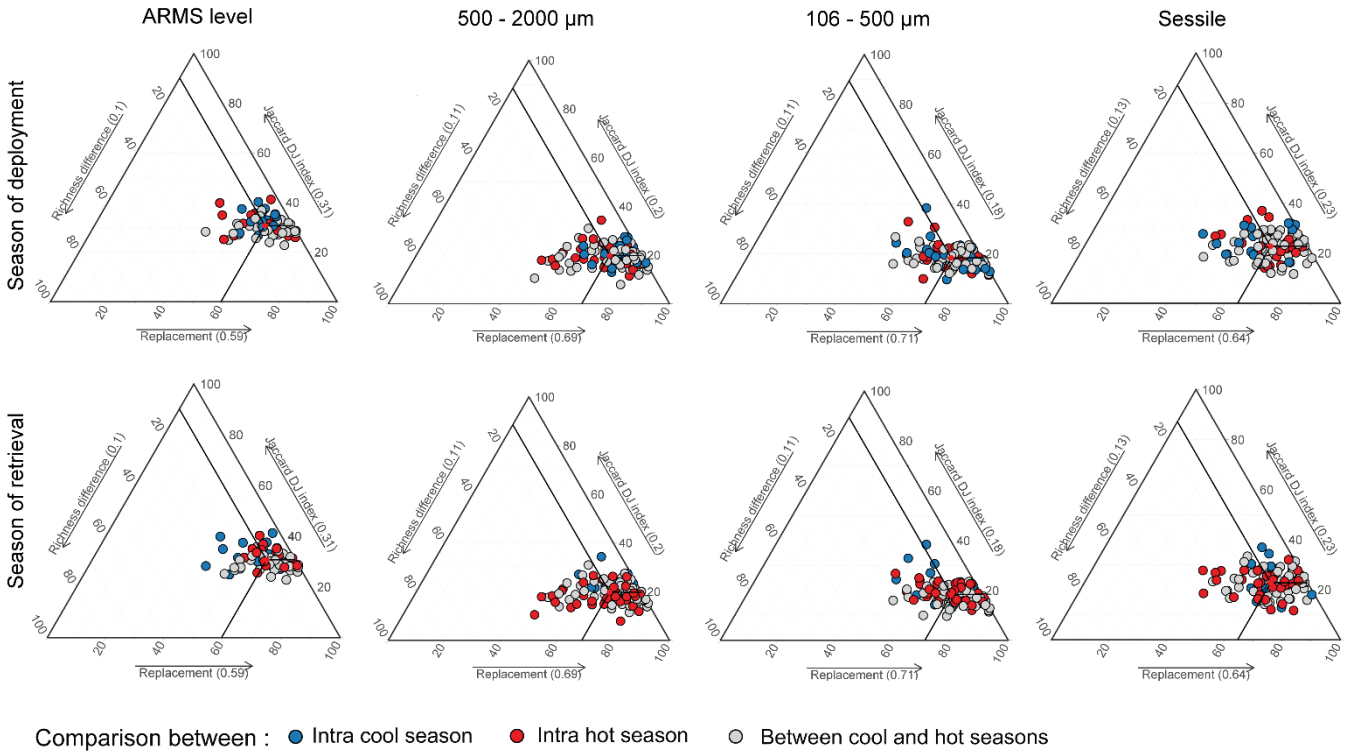
	Comparisons	COI			18S		
		Similarity	Replacement	Richness	Similarity	Replacement	Richness
ARMS level	Cool	0.332 a	0.58	0.0884	0.546 a	0.349	0.105
	Hot	0.317 ab	0.562	0.121	0.521 ab	0.351	0.128
	Cool VS hot	0.293 a	0.614	0.0934	0.508 b	0.385	0.106
500 - 2000 µm	Cool	0.212	0.712	0.0763	0.416 a	0.458 a	0.126
	Hot	0.2	0.663	0.138	0.374 b	0.51 ab	0.116
	Cool VS hot	0.191	0.691	0.117	0.367 ab	0.524 b	0.109
106 - 500 µm	Cool	0.197	0.691	0.112	0.46 a	0.469	0.0704 a
	Hot	0.183	0.713	0.104	0.388 b	0.478	0.134 b
	Cool VS hot	0.18	0.715	0.105	0.398 b	0.501	0.101 a
Sessile	Cool	0.256 a	0.581 a	0.163	0.463 a	0.45	0.0865 a
	Hot	0.233 ab	0.65 b	0.117	0.368 b	0.461	0.171 b
	Cool VS hot	0.217 b	0.657 b	0.126	0.384 b	0.479	0.137 ab

ESM 15.5: For season of retrieval

	Comparisons	COI			18S		
		Similarity	Replacement	Richness	Similarity	Replacement	Richness
ARMS level	Cool	0.324 a	0.549 a	0.127	0.512 a	0.364	0.124
	Hot	0.323 a	0.594 ab	0.0827	0.56 b	0.329	0.11
	Cool VS hot	0.293 b	0.614 b	0.0931	0.506 a	0.388	0.106
500 - 2000 µm	Cool	0.239 a	0.675	0.0858	0.407 a	0.504 ab	0.0889
	Hot	0.184 b	0.674	0.142	0.387 ab	0.474 a	0.139
	Cool VS hot	0.196 b	0.7	0.104	0.365 b	0.53 b	0.105
106 - 500 µm	Cool	0.22 a	0.666 a	0.114	0.414 ab	0.497 a	0.0884
	Hot	0.192 b	0.699 ab	0.109	0.431 b	0.446 b	0.124
	Cool VS hot	0.167 c	0.731 b	0.102	0.389 a	0.51 a	0.101
Sessile	Cool	0.24	0.649	0.111	0.406	0.503 a	0.0906 a
	Hot	0.224	0.626	0.15	0.391	0.43 b	0.179 b
	Cool VS hot	0.227	0.655	0.119	0.384	0.485 a	0.131 ab



ESM 16: Ternary plots of Jaccard similarity and the partitions of beta diversity (replacement and richness) for the full ARMS unit (ARMS level) and the three fractions obtained from the 18S metabarcoding. Comparisons were computed from ARMS deployed six months and one year. Ternary plots are shown for comparison within and among season of deployment and season of retrieval. Red dot and numbers in brackets on the axis labels represent the mean value.



ESM 17: Ternary plots of Jaccard similarity and the partitions of beta diversity (replacement and richness) for the full ARMS unit (ARMS level) and the three fractions obtained from the COI (OTU99%) metabarcoding. Comparisons were computed from ARMS deployed six months and one year. Ternary plots are shown for comparison within and among season of deployment and season of retrieval. Red dot and numbers in brackets on the axis labels represent the mean value.

ESM 18: Extended version of the table 3 which summarise of the parameters employed for ARMS deployment and OTUs processing for cryptobiome reef studies. The table is provided in two pages. * represent filtration based on abundance of reads. 1: only metazoans and marcroalgae; 2: only metazoans.

Reference	Site	Ocean / sea	# ARMS	# site	Immersion time (month)	Immersion period	Marker	OTU size	Fraction
Leray & Knowlton 2015	Florida (USA)	Atlantic ocean	9	1	6	November - May	COI	NA - CROP	all
	Virginie (USA)	Atlantic ocean	9	1	6	September - May	COI	NA - CROP	all
Al-Rshaidat et al 2016	Gulf of Aqaba (Jordan)	Red Sea	5	2	16	October - February	COI	NA - CROP	all
Pearman et al. 2016	Saudi Arabian coast	Red Sea	9	3	12	April - May	COI	97%	all
Ransome et al. 2017	Mo'orea (French Polynesia)	Pacific ocean	3	1	24	January	COI	NA	all
Pearman et al. 2018	Thuwal (Saudi Arabia)	Red Sea	33	11	24	February and May/June - May/July	COI	NA - CROP	all
Pearman et al. 2018	Thuwal (Saudi Arabia)	Red Sea	33	11	24	February and May/June - May/July	18S	NA - CROP	all
Carvalho et al. 2019	Saudi Arabian coast	Red Sea	87	22	24	February,May,August - May, June, July, November	COI	ESM not available	Mobile and sessile
Casey at al. 2021	Bali (Indonesia)	Pacific ocean	6	2	11 and 23	July - June	COI	97%	all
Casey at al. 2021	Bali (Indonesia)	Pacific ocean	6	2	11 and 23	July - June	18S	99%	all
Nichols et al. 2021	Hawai'i	Pacific ocean	6	1	23	July	COI	97%	all
Ip et al. 2022	Singapore	Indian ocean	12	4	24	June -June/July	COI	97%	all
Ip et al. 2022	Singapore	Indian ocean	12	4	24	June -June/July	18S	1%	all
Villalobos et al. 2022	Saudi Arabian coast	Red Sea	33	4	24	May	COI	No (ASV)	merged mobile

* calculed in this study

Reference	# reads	# OTU	Mean # OTU by ARMS	Assignment threshold	% OTU with phylum assignement	% OTU with species assignement	# phylum (Metazoa)	Sequencing	Filtration				Database
									Cod	Sin	Bac	Tax	
Leray & Knowlton 2015	409 613	1 391	536 ± 30	97% blast 90% SAP	72%	12%	32 (22)	Ion Torrent PMG	V	V	V		BOLD; Genbank
	572 290	1 204	434 ± 55	97% blast 90% SAP	59%	10%			V	V	V		
Al-Rshaidat et al 2016	152 604	1 197	609 ± 114	97% blast 80% SAP	63%	8%	NA (15)	Ion Torrent PMG	V	V	V		BOLD; Genbank
Pearman et al. 2016	69 000	1 700	1 297*	NA	NA	NA	50 (NA)	Illumina MiSeq		V			PR2
Ransome et al. 2017	1 227 154	2 456	NA	97% blast 85% blast 90% SAP	55%	32%	28 (17)	Ion Torrent PMG		V	V		BOLD; Genbank; Mo'orea Biocode
Pearman et al. 2018	34 000	3830	660 ± 151	97% blast SAP	58%	NA	NA	Illumina MiSeq	V	V			BOLD; MIDORI
Pearman et al. 2018	19 750	5 420	750 ± 107	97% blast RDP	NA	NA	NA	Illumina MiSeq	V	V			Silva; PR2
Carvalho et al. 2019	NA	10 416 (1 471 by site)	828	ESM not available	55%	NA	20 (14)	Illumina MiSeq					ESM not available
Casey at al. 2021	3 964 674	31 900	6 580 to 14 237	85% blast	51%	NA	38	Illumina MiSeq	V	V			Local; Mo'orea Biocode; Genbank
Casey at al. 2021	3 696 915	25 994	7 113 to 11 237	90% blast	99%	NA	51	Illumina MiSeq		V			Silva; PR2
Nichols et al. 2021	NA	893	NA	97 % blast 85% LCA	NA	NA	NA	Illumina MiSeq	V	V*		1	Local; Mo'orea Biocode; BOLD; Genbank
Ip et al. 2022	157 941	410	NA	RDP 80% confidence	100%	54.60%	(11)	Illumina HiSeq2500 and MiSeq		V*	V	2	MIDORI
Ip et al. 2022	NA	561	NA	RDP 60% confidence	100%	NA	32 (13)	Illumina HiSeq2500 and MiSeq		V*			Silva; PR2
Villalobos et al. 2022	NA	33 832 ASV	NA	RDP	NA	NA	NA	Illumina MiSeq	V	V			BOLD; MIDORI

Chapitre 5 : Evaluation de la distribution spécifique du cryptobiome

Résumé :

Ce chapitre s'intéresse à évaluer la diversité et la distribution spatiale des espèces d'un principaux genre de poissons cryptobenthiques retrouvés dans les ARMS déployés dans l'archipel des Mascareignes. Dans un premier temps, les patrons de répartitions retrouvés au sein des espèces du genre *Cirripectes* ont discuté dans ce chapitre. Dans un second temps, les résultats observés sur les variations intra et inter-spécifiques ont été mis en perspective du pipeline d'analyses de données de métabarcoding dans le Chapitre 6.

Les ARMS (*Autonomous Reef Monitoring Structures*) sont des mini-récifs artificiels conçus pour échantillonner de façon standardisée les organismes cryptobenthiques sessiles et mobiles de petite taille. Ils permettent également la collecte de petits poissons cryptobenthiques, tels que les blennies du genre *Cirripectes*. Des études récentes ont permis de découvrir plusieurs espèces de *Cirripectes* endémiques de zones géographiques restreintes (îles ou archipels), malgré la distribution généralement large des espèces de blennies tropicales et subtropicales. Ainsi, pour évaluer la diversité et la distribution des espèces de *Cirripectes* dans l'archipel des Mascareignes, une région sous étudiée, mais un haut lieu de biodiversité important, le génome mitochondrial complet ainsi que le gène nucléaire de la rhodopsine ont été séquencés pour les 39 spécimens collectés avec des ARMS déployés sur les pentes externes des récifs coralliens de La Réunion et de Rodrigues. Les séquences mitochondriales COI ont été analysées en intégrant les spécimens des principales bases de données de séquences publiques. Trois espèces ont été trouvées dans l'archipel des Mascareignes, *Cirripectes castaneus*, *Cirripectes randalli* et *Cirripectes stigmaticus*. *Cirripectes castaneus* et *C. stigmaticus* ont toutes les deux une distribution indopacifique avec plusieurs haplotypes partagés entre des localités éloignées. En accord avec la littérature, *C. randalli* présente un endémisme restreint aux Mascareignes. Nos résultats ont confirmé la présence de *C. castaneus*, *C. randalli*, et *C. stigmaticus* à Rodrigues, ainsi que la présence de *C. stigmaticus* à La Réunion. Cette étude contribue à combler les lacunes dans les connaissances taxonomiques et moléculaires du cryptobiome récifal dans le sud-ouest de l'océan Indien, et fournit les premiers mitogénomes complets pour le genre, une étape cruciale pour les futurs inventaires conduits en biologie moléculaires (ex. eDNA).

Ces résultats sont publiés dans *Ecology and Evolution* et ont été présentés à travers une communication orale lors du 12ème symposium scientifique du WIOMSA (Western Indian Ocean Marine Science Association).

Received: 17 October 2022 | Revised: 25 January 2023 | Accepted: 30 January 2023

DOI: 10.1002/ece3.9850

RESEARCH ARTICLE

Ecology and Evolution Open Access WILEY

New insights into the diversity of cryptobenthic *Cirripectes* blennies in the Mascarene Archipelago sampled using Autonomous Reef Monitoring Structures (ARMS)

Marion Couëdel[1] | Agnes Dettai[2] | Mireille M. M. Guillaume[3,4] | Fleur Bruggemann[1] | Sophie Bureau[1] | Baptiste Frattini[1,3] | Amélie Verde Ferreira[2] | Jean-Lindsay Azie[5] | J. Henrich Bruggemann[1,4]

[1]Université de La Réunion, UMR 9220 ENTROPIE (Université de La Réunion, IRD, IFREMER, Université de Nouvelle-Calédonie, CNRS), La Réunion, Saint-Denis, France

[2]Muséum national d'Histoire naturelle (MNHN), UMR 7205 ISYEB (MNHN, CNRS, Sorbonne Université, EPHE, Université des Antilles), Paris, France

[3]Muséum national d'Histoire naturelle (MNHN), UMR 8067 BOrEA (MNHN, CNRS 2030, Sorbonne Université, IRD 207, Uni Caen-Normandie, Université des Antilles), Paris, France

[4]LabEx CORAIL, Université de Perpignan, Perpignan, France

[5]Rodrigues Regional Assembly, Port Mathurin, Rodrigues, Mauritius

Correspondence
Marion Couëdel, Université de La Réunion, UMR 9220 ENTROPIE (Université de La Réunion, IRD, IFREMER, Université de Nouvelle-Calédonie, CNRS), Saint-Denis, La Réunion, 97400, France.
Email: marion.couedel@gmail.com

Funding information
Fonds européen de développement régional (FEDER), Grant/Award Number: 20171591-0002633 CALIBIOME 2017-2022

Abstract

Autonomous Reef Monitoring Structures (ARMS) are artificial mini-reefs designed for standardized sampling of sessile and small motile cryptobenthic organisms. ARMS are also effective for collecting small cryptobenthic fishes, such as the combtooth blennies of the genus *Cirripectes*. Recent studies discovered several *Cirripectes* species endemic to islands or archipelagos, in spite of the generally broad distributions of tropical and subtropical blennies. Thus, to evaluate the diversity and distribution of *Cirripectes* species in the Mascarene Archipelago, a little-studied region but an important biodiversity hotspot, complete mitochondrial genomes, and nuclear rhodopsin genes were sequenced for 39 specimens collected with ARMS deployed on outer reef slopes at Reunion and Rodrigues islands. Mitochondrial COI sequences were analyzed to integrate these specimens within the largest dataset of publicly available sequences. Three species were found in the Mascarene Archipelago, *Cirripectes castaneus*, *Cirripectes randalli*, and *Cirripectes stigmaticus*. *C. castaneus* and *C. stigmaticus* both have an Indo-Pacific distribution with several haplotypes shared among distant localities. In agreement with the literature, *C. randalli* shows a small-range endemism restricted to the Mascarenes. We confirmed the presence of *C. castaneus*, *C. randalli*, and *C. stigmaticus* in Rodrigues, and the presence of *C. stigmaticus* in Reunion. This study contributes to filling the gaps in taxonomic and molecular knowledge of the reef cryptobiome in the South-West Indian Ocean, and provides the first complete mitogenomes for the genus, a crucial step for future molecular-based inventories (e.g., eDNA).

KEYWORDS
barcoding, coral reefs, cryptic teleosts, mitogenome, molecular species delineation, South-West Indian Ocean

TAXONOMY CLASSIFICATION
Biogeography

1 | INTRODUCTION

Understanding coral reef ecosystem functioning requires knowledge of the distribution and abundance of reef-associated fishes. Due to their accessibility to visual observation, the larger reef fishes have been intensively studied for decades (Knowlton et al., 2010). Although their diversity and taxonomy appeared relatively well resolved (Allen, 2015; Fisher et al., 2015; Mora et al., 2008), molecular studies revealed extensive hidden diversity (Hubert et al., 2017; Steinke et al., 2009). The smaller taxa present additional challenges, as they are inherently more difficult to find and identify, and are therefore often omitted from visual surveys and collections (Bellwood et al., 2019; Brandl et al., 2018; Pearman et al., 2018). Although they are often overlooked, their distinctive demographic dynamics may make them a cornerstone of ecosystem functioning in modern coral reefs (Brandl et al., 2019).

Cryptobenthic reef fishes are small, bottom-dwelling, morphologically, or behaviorally cryptic species. They comprise families such as combtooth blennies (Blenniidae), gobies (Gobiidae), triplefins (Tripterygiidae), and cardinalfishes (Apogonidae). Despite being the ocean's smallest vertebrates, they contribute disproportionately to coral reef food webs through their high abundance, rapid somatic growth, and high predation mortality, producing almost 60% of reef fish biomass consumed within the ecosystem (Brandl et al., 2019). Furthermore, cryptic fishes represent approximately 10% of vertebrate diversity on coral reefs and can exhibit high levels of endemism (Bellwood et al., 2019; Brandl et al., 2018).

The genus *Cirripectes* Swainson, 1839 (Family Blenniidae, Order Blenniiformes) comprises 24 recognized species of combtooth blennies, broadly distributed in the Indo-Pacific from East Africa to Rapa Nui in the eastern Pacific (Hastings & Springer, 2009; Hoban & Williams, 2020; Williams, 1988). Currently, 14 species are recorded with COI DNA sequences in the BOLD database (Ratnasingham & Hebert, 2007), plus three genetically divergent groups that possibly represent not yet described new species. Most *Cirripectes* species are smaller than 100 mm. They are herbivorous and/or detrivorous cryptobenthic teleosts that primarily inhabit rocky or coral substrates in shallow (<5 m depth) high-surge fore reef habitats (Williams, 1988). However, individuals of *Cirripectes matatakaro* and *Cirripectes castaneus* may be encountered deeper (more than 20 m and over 32 m depth, respectively; Williams, 1988, Hoban & Williams, 2020). Species show considerable variation in geographic range sizes, from small area endemism (e.g., *Cirripectes heemstraorum* endemic to the East coast of South Africa at Cape Vidal; Williams, 2010) to Indo-Pacific-wide distributions (e.g., *C. castaneus*; Williams, 1988). Blennies' eggs are demersal and attached to the substratum with a filamentous, adhesive pad or pedestal (Breder & Rosen, 1966; Watson, 2009). Larvae are planktonic and abundant in shallow, coastal waters (Watson, 2009). While *Cirripectes* specimens are common in museum collections, incomplete knowledge of sexual dimorphism and geographic color variations, combined with a lack of adequate species identification keys, have resulted in numerous misidentifications and undetected cryptic species (Williams, 2010). In many cases, color morphs were considered distinct species, even though sexual polychromatism has been described for several *Cirripectes* species (Williams, 1988).

The latest fish checklist of Reunion reported six species of *Cirripectes* (Wickel et al., 2020): *C. castaneus*, *Cirripectes filamentosus*, *Cirripectes polyzona*, *Cirripectes quagga*, *Cirripectes randalli*, and *Cirripectes stigmaticus* (Table 1). For Rodrigues, four *Cirripectes* species have been listed: *C. castaneus*, *C. filamentosus*, *Cirripectes gilberti* and *C. stigmaticus* (Heemstra et al., 2004). *Cirripectes auritus* was not recorded in Reunion and Rodrigues but was described as present in Mauritius (Debelius, 1993; Fricke, 1999). Most of these *Cirripectes* species are considered to be widely distributed. *Cirripectes filamentosus*, *Cirripectes polyzona*, *Cirripectes quagga*, and *C. stigmaticus* occur throughout the Indo-Pacific, while the distribution of *C. auritus* and *C. castaneus* ranges from East Africa to the western Pacific. *C. gilberti* has a more restricted distribution and occurs throughout the Indian Ocean (Williams, 1988). However, recent studies discovered several *Cirripectes* species endemic to islands or archipelagos (Delrieu-Trottin et al., 2018; Hoban & Williams, 2020). For the Mascarenes, only *C. randalli* is known to have a distribution limited to the archipelago (Williams, 1988). Ecological and geographical distributions of these species are summarized in Table 1.

Cryptobenthic fishes are often under-sampled due to their hidden habits. Therefore approaches focused on sampling small and cryptobenthic fauna, such as artificial mini-reefs ARMS, represent alternative sampling techniques. ARMS, for Autonomous Reef Monitoring Structures, are stacks of nine PVC plates spaced at a 12 mm distance, designed to mimic the complexity of coral reef habitats. Each ARMS represents slightly over 4.5 L of habitat volume. Affixed to the seabed, they are left to be colonized by a diversity of marine species, then collected and dismantled to study the associated biota (see Zimmerman & Martin, 2004 for more details).

To evaluate the diversity and distribution of *Cirripectes* species in the Mascarene Archipelago, a little-studied region with high endemism and source of type material for this genus, we reconstructed the phylogeographic relationships within *Cirripectes* collected using ARMS. We conducted a multi-marker approach by sequencing mitochondrial and nuclear genes and integrated newly collected specimens within the largest dataset of publicly available sequences. We sequenced complete mitochondrial genomes as part of the current effort to complete the inventory of teleosts in French territories led by the *Muséum national d'Histoire naturelle* (MNHN). Having available complete mitogenomes enables the construction of more robust phylogenies, but the current paucity of mitogenomes in public databases makes such analyses premature. Finally, we explored the nucleotide divergences between *Cirripectes* species found in the Mascarene Islands to assess the performance of current mini-barcodes used to detect teleosts species in eDNA studies.

TABLE 1 *Cirripectes* species reported from the Mascarene Archipelago with their ecological and geographical distributions.

Species	Habitat	Depth	Polymorphisms	Distribution
C. auritus (Carlson, 1981)	Coral reefs	<10 m; max 20 m	S+G	Indo-West Pacific
C. castaneus (Valenciennes, 1836)	Rocky and coralline substrates; wave-swept algal ridges	<10 m; max 30 m	S+G*	Indo-West Pacific
C. filamentosus (Alleyne & Macleay, 1877)	Coral and rocky reefs; tolerate a wider range of environmental conditions than other *Cirripectes* spp	<7 m; 20 m	S (+G)	Indo-West Pacific
C. gilberti (Williams, 1988)	Rocky and coralline substrates	<8 m	S	Indian Ocean
C. polyzona (Bleeker, 1868)	Algal ridges and crests between surge channels of exposed seaward reefs	Usually <3 m; <20 m	S	Indo-Pacific
C. quagga (Fowler & Ball, 1924)	Algal ridges and crests between surge channels of exposed seaward reefs	<10 m; max 19 m	S+G*	Indo-Pacific
C. randalli (Williams, 1988)	Coral patches in surge channels of rocky reefs with light surf	<8 m	S	Mascarenes
C. stigmaticus (Strasburg & Schultz, 1953)	Upper edge of seaward reef slopes. Adults inhabit coastal reef flats with rich corals and algae. Among *Acropora* and *Pocillopora* corals of wave-swept algal ridges	<20 m	S+G	Indo-Pacific

Note: The types of polymorphism were indicated as 'S' for sexual and 'G' for geographical. Species with more than one sympatric sexual color pattern are indicated by *. Information was synthetized from Williams (1988), Letourneur et al. (2004), and Allen et al. (2013).

2 | MATERIALS AND METHODS

2.1 | Specimen collection

Specimens were collected from 54 ARMS deployed between September 2014 and August 2021 at two islands of the Mascarene Archipelago (South-West Indian Ocean): along the western and south-western coasts of Reunion Island (10 sites) and along the North coast of Rodrigues (3 sites). At each site, three replicate ARMS units were deployed on spurs of outer coral reef slopes at 10–12 m depth, with immersion times varying from 6 months to 4 years. The ARMS deployment recovered 144 fishes, of which 39 specimens were *Cirripectes* (Appendices S1 and S2). These samples comprised at least four teleost families, including 85 Gobiidae (mainly *Eviota* and *Enneapterygius*), 40 Blenniidae (39 *Cirripectes*, 1 *Aspidontus*), 2 Pomacentridae (*Pycnochromis nigrurus*), 2 Muraenidae (including one *Gymnothorax*), and 15 specimens of undetermined affinity. Most individuals were photographed alive, identified to the lowest taxon level possible based on morphology, individually preserved in 90% ethanol (EtOH) and stored at 4°C. For *Cirripectes* specimens, the total length was measured with a ruler (Appendix S2).

2.2 | DNA extraction and sequencing

Muscle tissue was used for total genomic DNA extraction using DNeasy Blood & Tissue Kit (Qiagen), following the manufacturer's instructions. PCR amplification and sequencing were performed for the entire mitochondrial genome and the partial retro-rhodopsin nuclear gene (Rh193 and Rh1039r; Chen et al., 2003). The mitogenome was amplified in three overlapping fragments. The first fragment from the end of the 16S to the end of COI was amplified with 16SAR (Kocher et al., 1989) and MtH7061 (Hinsinger et al., 2015). The second fragment from the beginning of COI to the end of ND4 was amplified with F5231cha (Hinsinger et al., 2015) and MtH11944 (Hinsinger et al., 2015). Additionally, we developed three new primers, R11944cha 5′-CATAGCTNCTACTTGGATTTGCACCA-3′ and two specific primers designed for the genus *Cirripectes*: F5231Cirri 5′-TAGRCAGGCAGGCCTCGATCCTRCA-3′ and R11944Cirri 5′-CATAGTTTCTGCTTGGAGTTGCACCA-3′ to improve amplification success. The third fragment from ND5 to the end of 16S was amplified with MtL11910 (Hinsinger et al., 2015) and 16SBR (Kocher et al., 1989). The amplicons were pooled with other PCR amplicons following Hinsinger et al. (2015) for cost efficiency. Library preparation followed Meyer and Kircher (2010) and Illumina MiSeq sequencing (PE250) was performed at the *Service de Systématique Moléculaire* of the MNHN at Concarneau and at the *Institut du Cerveau et de la Moëlle Epinière* (Pitié-Salpêtrière Hospital, Paris). Two rhodopsin samples were sequenced in both directions through Sanger sequencing by Eurofins Genomic, France.

2.3 | Sequences processing and public sequences retrieval

Reads were processed with Geneious Prime 2019.2.3. Paired-end reads were merged, the primers were used as barcodes to recover the fragment ends and the merged reads were de novo assembled. The resulting contigs were checked in BOLD and Genbank databases before being used as references for elongation through

repeated mapping of reads with a maximum of 1% mismatch and three bases gap allowed. Mapping was repeated until no further reads could be mapped. Complete linear mitochondrial consensuses were transformed into circular sequences and overlapping sections were manually inspected and adjusted. Last, reads were mapped back against the obtained circular sequences to check coverage and final assembly. The mitogenome sequences were annotated using online MitoFish (Iwasaki et al., 2013). Sanger sequences were assembled and checked in Geneious Prime 2019.2.3 (2019). To increase taxonomic and spatiotemporal coverage, our datasets were extended with publicly available sequences from BOLD and/or GenBank (Figure 1; ESM 1). In some instances, sequences were renamed according to the conclusions of the papers in which they were published, even if the names were not corrected in the database: sequences within BOLD BIN:AAU0601 were renamed from *C. castaneus* to *C. randalli* (Hoban & Williams, 2020), and MH932003 to MH932007 were renamed from *C. alboapicalis* to *C. patuki* sensu Delrieu-Trottin et al., 2018 (Delrieu-Trottin et al., 2018). Two sequences, GBMNB4802-20 and KX223895.1, respectively from BOLD and Genbank, probably belong to a different genus and were removed from the analyses.

2.4 | Molecular and phylogenetic analyses

All analyses were performed on three datasets: (i) mitochondrial cytochrome oxidase I (further called COI dataset; $N = 296$; ESM 1), (ii) nuclear rhodopsin (further called Rho dataset; ESM 2) for which 53 sequences were used corresponding to 17 homozygotes and 2×18 heterozygotes (NGS reads were checked for gene versions); for haplotype analyses, 70 sequences were used corresponding to 35 specimens with two alleles, and (iii) a dataset consisting of the concatenated two rRNA (12S and 16S) and 13 coding DNA sequence (CDS: ND1, ND2, COI, COII, ATPase 8, ATPase 6, COIII, ND3, ND4L, ND4, ND5, ND6, and Cyt b) from the complete

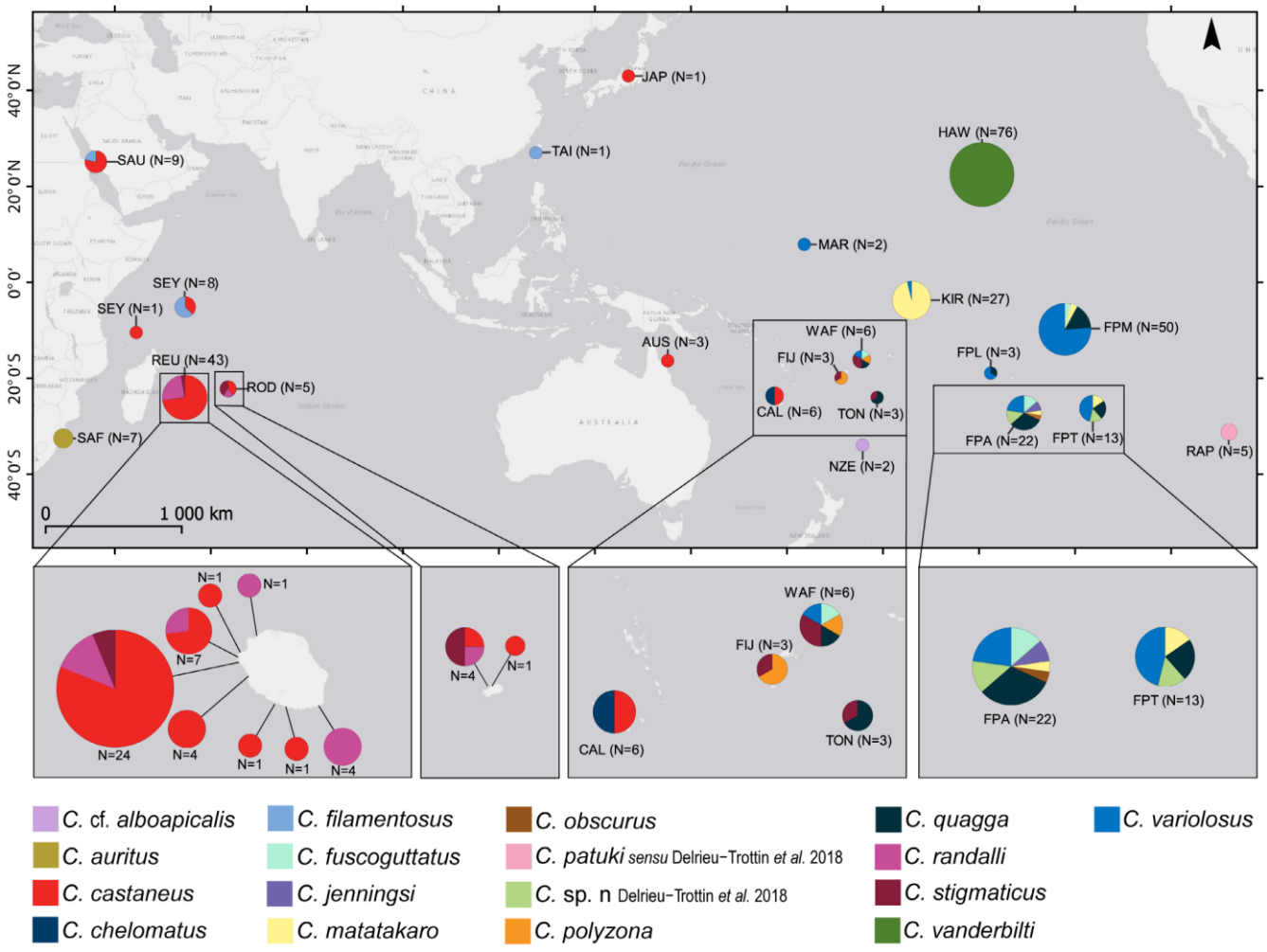


FIGURE 1 *Cirripectes* species distribution inferred from the COI dataset based on molecular delimitation analyses. AUS, Australia; CAL, New Caledonia; FIJ, Fiji; FPA, French Polynesia, Austral Islands; FPL, French Polynesia, Leeward Islands; FPM, French Polynesia, Marquesas and Society Islands; FPT, French Polynesia, Tuamotu-Gambier; HAW, Hawaii; JAP, Japan; KIR, Kiribati; MAR, Marshall Islands; NZE, New Zealand; RAP, Rapa Nui; REU, Reunion; ROD, Rodrigues; SAF, South Africa; SAU, Saudi Arabia; SEY, Seychelles; TAI, Taiwan; TON, Tonga; WAF, Wallis and Futuna. The size of pie charts is proportional to the number of samples.

mitogenome (henceforth called mt dataset; $N = 24$). Based on the Blenniidae phylogeny of Lin and Hastings (2013), we selected sequences from multiple outgroup species (COI: $N = 5$; Rho: $N = 4$; mt: $N = 4$) from the recognized tribes *Ophioblennius macclurei*, *Omobranchus elegans*, *Omobranchus obliquus*, *Petroscirtes breviceps*, *Salarias fasciatus*, and *Ecsenius bicolor* (Appendix S3). Sequences of each marker were aligned separately via Muscle 3.8.425 (Edgar, 2004), implemented in Geneious Prime 2019.2.3 using default parameters, and manually trimmed to maximize the shared length among the sequences (COI = 506 bp, Rho = 737 bp). For the mt dataset, complete mitogenomes were aligned and translated into protein to check the codon position for the coding genes. The two tRNA and 13 CDS were kept according to coding position and overlapping CDS portions were trimmed at the end to prevent a shift in the translation frame.

ModelFinder 1.6.8 (Kalyaanamoorthy et al., 2017) in PhyloSuite (Zhang et al., 2020) was used to select the evolutionary model with the best Akaike Information Criterion (AIC; Akaike, 1974) for each dataset. Maximum-likelihood (ML) and Bayesian inference (BI) analyses were conducted. ML tree searching was conducted in W-IQ-TREE (Trifinopoulos et al., 2016) for 1000 ultrafast bootstraps (Hoang et al., 2018). Multiple independent Bayesian inference searches used strict or uncorrelated log-normal relaxed molecular clock models with a Yule or birth rate ratio tree prior in BEAST2 2.1.2 (Bouckaert et al., 2014). Codon partitions gene and chains were run for 10 million generations, using Tracer (Rambaut et al., 2018) to confirm stationarity and mixing. Maximum Clade Credibility (MCC) tree was obtained through TreeAnnotator 1.10.4 and the generations before reaching stability were discarded as burn-in. For the mt dataset, TreeAnnotator did not allow to recover an MCC tree, which is consistent with the lack of convergence observed in Tracer. Therefore Bayesian inference was computed with MrBayes 3.2.6 (Ronquist & Huelsenbeck, 2003) in PhyloSuite (Zhang et al., 2020). The reconstruction was visualized in R 4.1 (R Core Team, 2021).

Three molecular delineation approaches were run separately on the COI, rhodopsin, and mitochondrial concatenated datasets: (i) Assemble Species by Automatic Partitioning (ASAP; Puillandre et al., 2021), (ii) the multi-rate Poisson Tree Processes (mPTP) method (Zhang et al., 2013), and (iii) a single threshold General Mixed Yule-Coalescent (GMYC) approach (Fujisawa & Barraclough, 2013) implemented with the R 4.1 package "splits" (Ezart et al., 2009; R Core Team, 2021). Additionally, Refined Single Linkage (RESL) analysis was performed on the COI dataset. RESL is used for the barcode identification numbers (BINs) system implemented in BOLD. The relationships among haplotypes from distinct localities were inferred from haplotype networks built using median-joining (Bandelt et al., 1999) methods implemented in popART 1.7 (Leigh & Bryant, 2015). Nucleotide diversity, Tajima's D, and an AMOVA were conducted in Arlequin 3.5 (Excoffier & Lischer, 2010), following the different clades observed in phylogenetic trees, to assess the level of genetic differentiation within and between the groupings obtained.

2.5 | Sliding window analyses

Sliding window analyses were performed to explore nucleotide divergence between *Cirripectes* species found in the Mascarene Islands and to determine the performance of current mini-barcodes used to detect teleosts species in eDNA (MiFish [163–185 bp; Miya et al., 2015]; Teleo [65 bp; Valentini et al., 2016]). These analyses were performed using the R package "SPIDER" 1.1.2 (Brown et al., 2012) with a window size of 160 and 65 bp for the MiFish and the Teleo markers, respectively, and with a step size of 1 bp. Divergence analyses were also performed on complete COI and partial COI fragments from Geller et al., 2013 (jgLCO1490-jdHCO2198) which are mostly used for metabarcoding of broad taxonomic groups.

3 | RESULTS

3.1 | Sequence analysis

The specimens were deposited in the MNHN fish collections under collection numbers MNHN-IC-2023-0216 to MNHN-IC-2023-0254. The sequences produced in this study were deposited in the BOLD dataset "Cryptic fishes from IO ARMS" (IOACT) and in Genbank under accession numbers presented in Appendix S2; the specimens were deposited in the collections of the MNHN. We generated 34 complete COI sequences representing 22 haplotypes. Including public sequences from BOLD and GenBank databases, the COI dataset was composed of 296 sequences of 506 bp from 15 species names and two undescribed groups, including 15 public BOLD BINs and 18 localities (Figure 1). The COI dataset comprised 135 haplotypes (ambiguities or missing data were not taken into account; if considered $N = 159$) with 170 parsimony informative sites. Out of 22 haplotypes, 15 were new while the remaining seven were identical to previously published sequences. A total of 176 bp of the 506 bp of the COI region were variable (34.78%; Appendix S4). The third codon position of COI sequences provided significantly more parsimony informative sites (89.88% were informative) than the first (11.31%) and the second (0%) codon positions.

The 27 rhodopsin sequences obtained by NGS were checked for heterozygotes. This yielded 14 distinct sequences. The complete dataset including public sequences comprised 70 sequences for 35 individuals (corresponding to six species names and five localities) and contained a total of 25 haplotypes, 11 from public databases and 14 from this study (13 new). The rhodopsin dataset had 30 variable sites (4.07%) and 19 parsimony informative characters (2.58%) over 737 bp ($\pi = 0.00479$).

The mitogenome dataset comprised complete mitochondrial genomes of 18 *C. castaneus*, 2 *C. randalli*, and 4 *C. stigmaticus*. For 10 *C. castaneus*, mitogenomes were not recovered completely. Out of the five partitions, the third codon position of the CDS partition provided significantly more parsimony informative sites (49.39%) than

the first and second positions, or the two tRNA partitions (15.69% and 26.84%). The 12S sequence had the lowest proportion of both variable characters (29.21%) and parsimony informative characters (15.69%).

3.2 | Phylogenic reconstruction of *Cirripectes*

For the COI dataset, the nucleotide substitution model with codon partition had less good fit than the single model; hence, the GTR+F+I+G4 model was selected for the analysis (Appendix S4). For the mitochondrial concatenated dataset sequences, the best-fit partitioning scheme was GTR+I+G4. For rhodopsin, different partition models were used. While TVM+F+G was the best-fit partitioning scheme for the first codon position, this scheme was unfortunately not implemented in BEAST. However, it was equivalent to the GTR+F+I+G4 model with fixed AG rate parameters (1.0; Bagley, 2018). For the second and third codon positions, F81+F+G4 and GTR+F+G4 were respectively selected.

Maximum-likelihood and BI-based phylogenetic reconstructions for COI resulted in tree topologies with marked similarities (Figure 2). For the 17 species in the dataset, the COI trees recovered well-supported clades within the genus. However, the branching order of these clades presented differences among methods. Clades 6 and 7 were sister species in the BI analysis, while clade 6 was closer to clade 8 within the ML approach. These incongruent results were probably related to the short branch lengths at the base of clades 1–8. *Cirripectes* appear to be a monophyletic genus relative to outgroups (Figure 2). *Cirripectes* collected in the Mascarenes were present in three clusters within two separate larger clades of the trees (Figure 2). For the mitochondrial dataset, both ML and BI trees strongly supported the grouping of our specimens in three clades (Appendix S5). For the rhodopsin dataset, both BI and ML trees had poorly supported clades (not shown) probably because of the low variability of sequences (1.63–15.92% of parsimony informative sites; Appendix S4).

3.3 | Molecular species delimitation

Based on the COI dataset, the ASAP molecular species delimitation approach divided the sequences into 16 groups, while the GMYC and BIN subdivided the COI dataset into 17 groups, and mPTP subdivided into 19 groups (Figure 2; ESM 1). For the sequences from specimens identified as *C. filamentosus* and *Cirripectes chelomatus*, GMYC, BIN, and mPTP methods divided sequences according to geographic origin (the western Indian and western Pacific oceans, respectively) and distinct morphological characters (Williams, 1988). The mPTP method was the only one that subdivided specimens identified as *Cirripectes matakaro* into two groups, as was the case for those identified as *Cirripectus variolosus*. These two subdivisions were not supported by the other methods and they will not be considered further here. Therefore, a final delimitation scheme was established based on a majority-rule consensus among the different delimitation analyses which included the 16 clades from the ASAP approach and the *C. chelomatus* clade delimited by the three other methods. Most of these can be unequivocally associated with *Cirripectes* species names from public databases and literature, except for clade 17 (*Cirripectes* sp. n. Delrieu-Trottin et al., 2018; Figure 2; ESM 1). According to Hoban and Williams (2020), clade 11 corresponded to *C. castaneus* and clade 9 to *C. stigmaticus* (ESM 1). However, sequences from the group named *C. castaneus* contained samples identified as *C. castaneus* and others as *C. stigmaticus* in public databases.

3.4 | Sequence variability and haplotype relationship

Nucleotide diversity for the COI marker was low and ranged from 0.000 to 0.014 (Table 2). *Cirripectes* showed an average difference of 2.07 bp (0.41%; 0–1.38%) within species and 61.43 bp (11.98%; 2.60–18.77%) among species (Appendix S6). *C. castaneus* and *C. stigmaticus* had a larger difference average of 5.38% (27.2 bp).

The AMOVA supported the hypothesis of the above species grouping and high levels of genetic structure, with Fst = 0.94 ($p < .001$) for COI. Moreover, a high percentage variation (94.59%) was observed for COI among species while a low percentage variation (5.41%) occurred within species (Appendix S7). Population pairwise tests also revealed that COI haplotypes were significantly genetically different in 113 of 136 comparisons ($p < .05$; Appendices S8 and S9). Most non-significant values were obtained for comparison with *Cirripectes obscurus* due to the insufficient number of sequences ($N = 1$; Table 2). For rhodopsin, the AMOVA results also supported the above species grouping (for species present in the dataset: Fst = 0.86; $p < .001$; Appendix S7).

The haplotype networks based on COI sequences were computed for each of the 17 species from the molecular delineation analyses (Figure 2). *Cirripectes* species presented several different haplotype network structures. *Cirripectes vanderbilti* had a star-shaped network, with one abundant haplotype surrounded by less common haplotypes, indicative of a possible recent population expansion or high demographic turnover within Hawaii (Grant & Bowen, 1998; Hoban & Williams, 2020). In contrast, *C. castaneus* had a complex star network with several abundant haplotypes surrounded by less common ones. Only the haplotype networks from species collected in this study are detailed here. The *C. castaneus* group comprised a higher number of specimens (Ns = 52) and haplotypes (Nh = 30), compared to the *C. randalli* (Ns = 10; Nh = 6) and *C. stigmaticus* (Ns = 8; Nh = 4; Table 2) groups. The *C. castaneus* network showed a unique haplotype for specimens collected in the Red Sea. *C. castaneus* haplotypes had wide geographic ranges with one haplotype shared among western Indian Ocean localities (Reunion, Rodrigues, Seychelles) and three shared among Indo-South Pacific localities (two by Reunion, Seychelles,

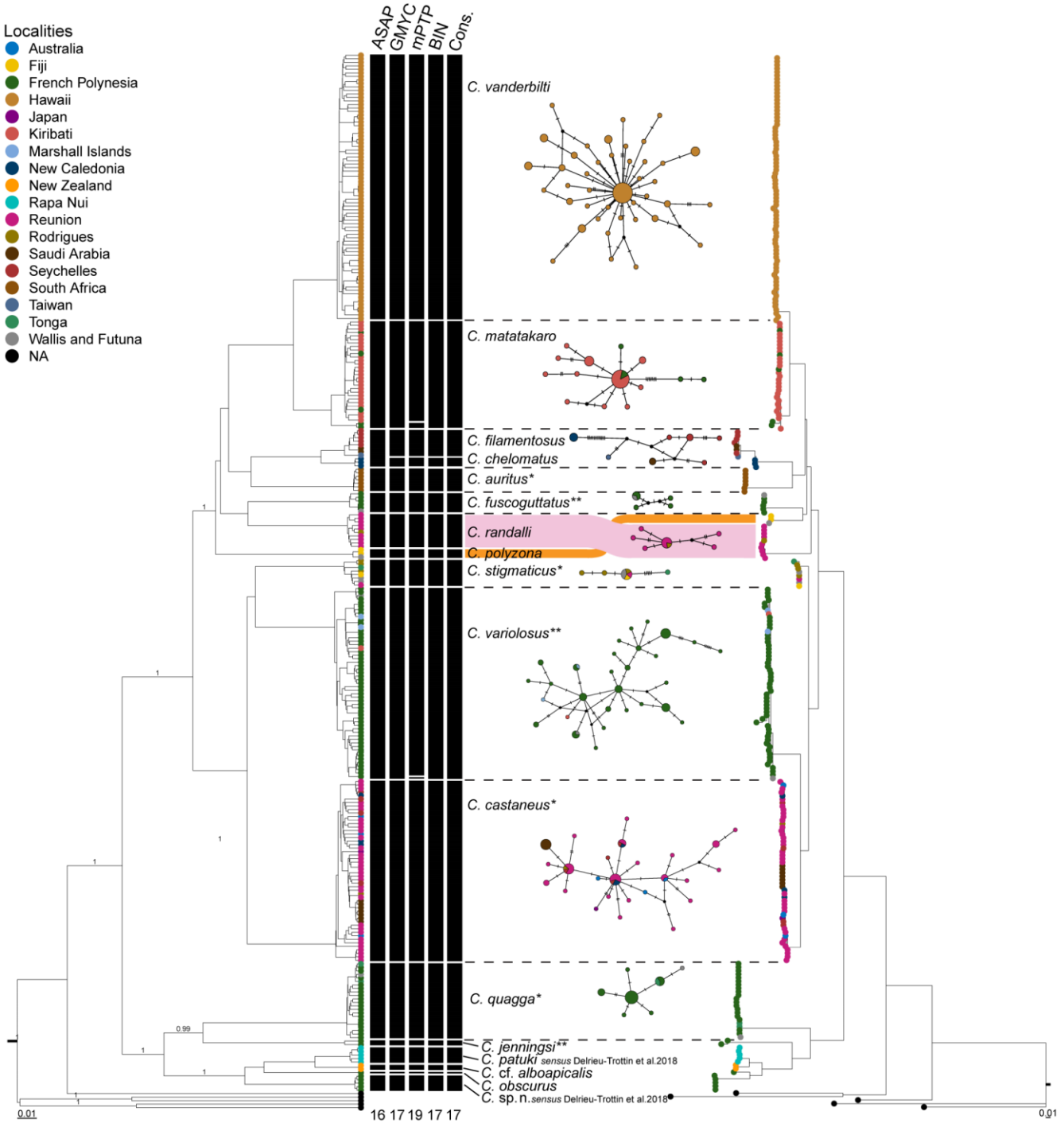


FIGURE 2 Ultrametric BEAST Bayesian Inference (left) and Maximum-Likelihood (right) trees of COI mitochondrial dataset, with a graphical representation of the results of the ASAP, GMYC, and mPTP molecular species delimitation analyses, BIN clustering from BOLD (RESL), and the resulting consensus clusters (Cons.). Between trees, median-joining networks for the consensus clusters. Groups with ≤3 haplotypes are not shown, except for *C. chelomatus* which is very close to *C. filamentosus*. Each circle represents a haplotype with size proportional to total frequency. Black crossbars on branches represent single nucleotide changes, black nodes indicate non-sampled probable haplotypes, while colors denote collection location as indicated in the embedded key. *Indicates species described as having Indo-Pacific ranges while **denotes species with Pacific ranges (from Williams, 1988).

New Caledonia and one by Reunion and Australia; Figure 2). The *C. randalli* network displayed one haplotype common to both Mascarene localities (only one specimen for Rodrigues) and one *C. stigmaticus* haplotype was shared among Indo-South Pacific localities (Reunion, Rodrigues, Fiji, Wallis & Futuna, and Tonga). Compared to other species included in the dataset, only *C. castaneus* and *C. stigmaticus* contained specimens from Indo-Pacific localities. The haplotype network computed on the rhodopsin

TABLE 2 Genetic diversity measures of *Cirripectes* species.

Species	# Sequences		# Haplotypes		# Polymorphic sites		# Nucleotide diversity		Differentiation for COI (%)		Tajima's D	
	COI	Rho	COI	Rho	COI	Rho	COI	Rho	Intra-specific	Inter-specific	COI	Rho
C. cf. *alboapicalis**	2	–	1	–	0	–	0	–	–	4.09	0	–
*C. auritus**	7	–	3	–	2	–	0.00111	–	0.11	8.60	−1.23716	–
C. castaneus	51	52	30	15	25	15	0.005411	0.001325	0.55	4.91	−1.80603*	2.15061*
*C. chelomatus**	3	6	2	3	1	2	0.00129	0.001447	0.13	2.60	0.00000	1.03194
C. filamentosus	8	2	7	2	11	1	0.00758	0.001357	0.75	3.04	−0.95806	0
C. fuscoguttatus	6	–	4	–	5	–	0.00414	–	0.42	8.04	−0.14427	–
*C. jenningsi**	2	–	2	–	7	–	0.01359	–	1.38	8.85	0.00000	–
C. matatakaro	31	–	14	–	21	–	0.00431	–	0.44	3.92	−2.01227*	–
*C. obscurus**	1	–	1	–	0	–	–	–	–	7.09	0.00000	–
C. patuki sensu Delrieu-Trottin et al., 2018*	5	–	5	–	7	–	0.005929	–	0.59	4.39	−0.74682	–
C. polyzona	3	–	2	–	2	–	0.00259	–	0.26	8.10	0.00000	–
C. quagga	22	2	7	2	7	1	0.00247	0.001357	0.25	12.35	−1.07910	0
C. randalli	10	2	6	1	8	0	0.00341	0	0.35	8.71	−1.63600*	0
C. stigmaticus	8	6	4	3	7	4	0.00374	0.001809	0.38	5.38	−1.35929	−1.29503
*C. vanderbilti**	76	–	40	–	42	–	0.00410	–	0.41	4.34	−2.45250*	–
C. variolosus	55	–	29	–	35	–	0.00827	–	0.83	6.46	−1.49064*	–
C. sp. n. Delrieu-Trottin et al., 2018*	5	–	2	–	1	–	0.000791	–	0.08	7.75	−0.81650	–
Global	296	70	159	26		30						

Note: Species with * have been sampled from one locality only.

dataset generated similar but less detailed information in terms of haplotype diversity, due to the lack of representation by available sequences (Appendix S10). No haplotypes were shared by species among the specimens studied.

3.5 | Extension of DNA barcode library and mitochondrial genome

The 24 complete mitochondrial genome sequences produced in this study belong to three species, *C. castaneus* ($N = 19$), *C. stigmaticus* ($N = 4$), and *C. randalli* ($N = 1$). These represent the first complete mitogenome sequences for the genus *Cirripectes*. The complete mitogenomes of *C. castaneus*, *C. randalli*, and *C. stigmaticus* are circular with sizes ranging from 16,476 to 16,512 bp (43.3% GC mean), 16,482 bp (43.6% GC), and 16,482 to 16,532 bp (43.8 % GC mean, Appendix S11.1), respectively. More detailed information is provided in Appendix S11.

3.6 | Sliding window analyses

Two pairwise comparisons between the most closely related species pair (*C. castaneus* with *C. stigmaticus*) and the more distant one (*C. stigmaticus* with *C. randalli*), were chosen for sliding window analysis. The distribution of divergent sites is shown in Figure 3. The alignment of the complete mitogenome of *C. castaneus* (Accession number OP749996) with that of *C. stigmaticus* (OP575312) was 16,485 bp long and contained 1162 variable characters. The uncorrected pairwise distance overall between this pair of taxa was 7.1% (p-distance; Srivathsan & Meier, 2012). The alignment of the complete mitogenome of *C. stigmaticus* with that of *C. randalli* (OP749983) was 16,489 bp long and contained 1795 variable characters corresponding to an uncorrected pairwise distance of 10.9% (p-distance). The divergence between species was calculated specifically for four fragments amplified by several commonly used primers pairs (Table 3). The PCR in silico failed for the MiFish primers pair (Miya et al., 2015) as no binding site was found for the

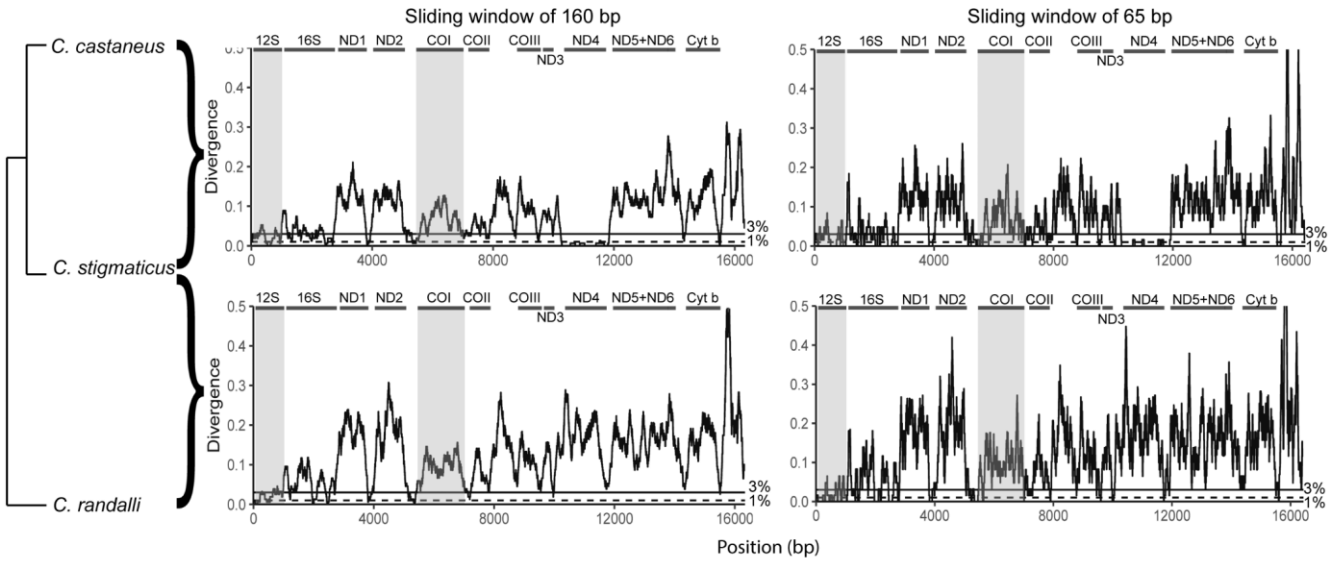


FIGURE 3 Sliding window analyses of the alignment between *C. castaneus* and *C. stigmaticus* (up) and *C. stigmaticus* and *C. randalli* (bottom). Distances are in K2P (Kimura, 1980). The horizontal lines show divergence threshold: dashed: 1%, full: 3%. Gene boundaries are indicated. Gray box indicates the regions of 12S and COI genes usually targeted for amplification. To improve visibility, values greater than 0.5 (only present in the control region) were cut.

TABLE 3 Nucleotide divergence between *C. castaneus* and *C. stigmaticus*, and between *C. stigmaticus* and *C. randalli*.

Primers	Reference	Length (bp)	% div (p-distance)	
			C. castaneus vs. *C. stigmaticus*	*C. stigmaticus* vs. *C. randalli*
Complete COI	–	1560	6.3% (99 bp)	9% (140 bp)
jgLCO1490-jdHCO2198 (COI)	Geller et al. (2013)	658	4.9% (32 bp)	8.8% (58 bp)
MiFish (12S)	Miya et al. (2015)	163–185	4.5%[a] (7 bp)	2.5%[a] (4 bp)
Teleo (12S)	Valentini et al. (2016)	63	6.3% (4 bp)	6.3% (4 bp)

[a]PCR in silico failed to find the biding site for the reverse primer of MiFish. Divergence was estimated from the end of the forward primer to the 160 following pb.

reverse primer. The complete COI fragment, partial COI from Geller et al. (2013) (jgLCO1490-jdHCO2198), and Teleo primers were efficiently amplified in silico and discriminated species with a divergence threshold of 3% (in p or K2P distances: minimal differences between models which rarely affect the identification success rates [Collins et al., 2012]). However, prior to the novel sequences presented here, only one reference sequence was available for the 12S fragment targeted by the Teleo primers, therefore only the COI barcodes could be used for species identification. Moreover, this sequence attributed to *C. polyzona* [LC278140.2] was identical to our *C. castaneus* sequences.

3.7 | *Cirripectes* species of Reunion and Rodrigues

Following the results of the delimitation analyses, from the 34 new COI sequences generated in this study, 28 were assigned to *C. castaneus*, 4 to *C. stigmaticus*, and two to *C. randalli*. Among the 26 specimens sequenced for rhodopsin, 22 were assigned to *C. castaneus*, 3 to *C. stigmaticus*, and one to *C. randalli*. Of the 29 specimens sampled in Reunion, 27 were assigned to *C. castaneus*, one to *C. stigmaticus*, and one to *C. randalli*. Of the 5 *Cirripectes* samples from Rodrigues, one was assigned to *C. castaneus*, three to *C. stigmaticus*, and one to *C. randalli*. Photographs of the three species are presented in Appendix S12. Each species has a distinct color pattern: *C. castaneus* has orange to red vertical bars on the head and anterior body half, and orange-red colored upper and lower caudal fin rays, *C. randalli* has orange-red dots on the head and body, and *C. stigmaticus* has orange-red vertical bars on the head and dots on the anterior body half (Appendix S12). Their sizes varied from 29 to 67 mm with a median size of 37 mm for *C. castaneus*, 41–44 mm for *C. randalli*, and 34–47 mm for *C. stigmaticus* (Appendix S2).

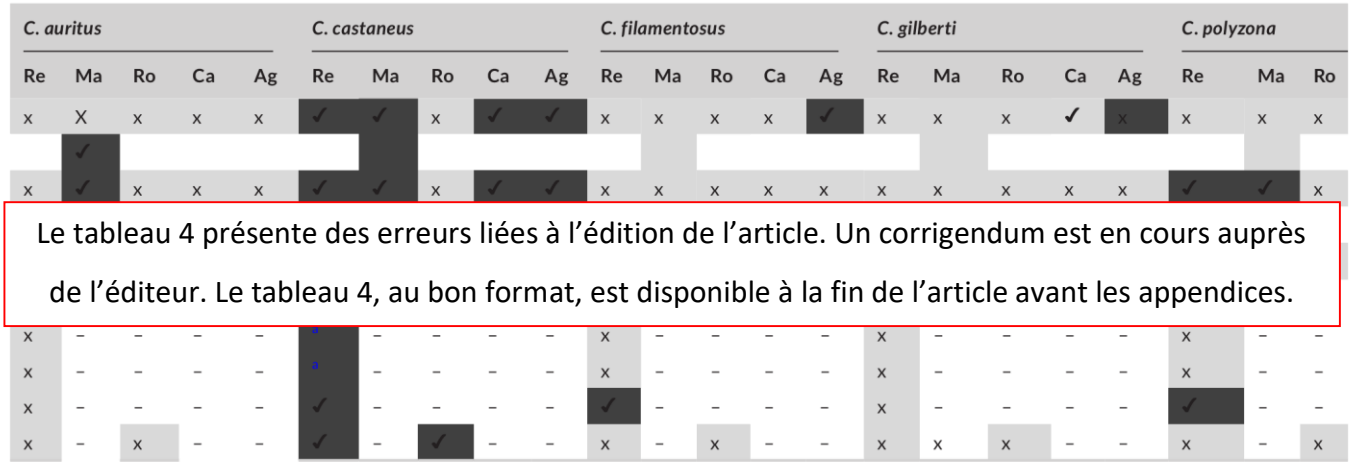
Cirripectes species previously reported from Reunion and Rodrigues are shown in Table 4, as based on visual surveys (Fricke et al., 2009; Letourneur et al., 2004), morphological (Fricke, 1999; Heemstra et al., 2004; Williams, 1988), and molecular studies (Collet et al., 2017; Hubert et al., 2015). Some of the original reference sequences for *Cirripectes* samples from Reunion were misidentified: it turned out that some *C. stigmaticus* were *C. castaneus* (BOLD:AAE2835) and all *C. castaneus* were *C. randalli* (BOLD:AAU0601; Hoban & Williams, 2020). Thus, *C. stigmaticus* (BOLD:AAE2834) was not genetically identified from Reunion prior to the present study (Table 4).

4 | DISCUSSION

4.1 | Phylogeny of *Cirripectes*

Phylogenetic trees constructed on the COI gene were the most informative since the availability of the sequences of the alternative genes for the other species and localities was limited. The present study supplements the range of available markers for future analyses. Phylogenetic trees corroborated the monophyly of the genus *Cirripectes*. Despite some differences in internal branching order, the COI trees were generally congruent with the morphological phylogeny established by Williams (1988) and recent single marker phylogenies also based on the COI gene (Delrieu-Trottin et al., 2018; Hoban & Williams, 2020). The differences observed between our results and phylogenies produced by Hoban and Williams (2020) were among the short branches and poorly supported clades in both studies. The low resolution among these branches is probably due to multiple rapid divergence events with little time to accumulate shared mutations, even in a fast-evolving marker (Avise, 2009; Douzery, 2010). The use of additional genetic markers with greater

TABLE 4 *Cirripectes* species reported from Mascarene Islands.



Le tableau 4 présente des erreurs liées à l'édition de l'article. Un corrigendum est en cours auprès de l'éditeur. Le tableau 4, au bon format, est disponible à la fin de l'article avant les appendices.

Note: Dark gray for the present, light gray for not observed, and white for not evaluated. Methods of identification were synthetized as follows (Meth.): M for morphometric analysis, V for visual survey, and G for molecular identification.

Abbreviations: Ag: Agalega Islands; Ca: Cargados Carajos Shoals (Saint Brandon); Ma: Mauritius; Re: Reunion; Ro: Rodrigues.

[a]Misidentification between *C. stigmaticus* and *C. castaneus*.

[b]Misidentification between *C. randalli* and *C. castaneus*.

variability, and a better taxon coverage for the genetic markers available, may resolve the remaining topological uncertainties or confirm the existence of fast diversification events.

4.2 | Molecular delimitation of *Cirripectes*

The number of clusters recovered by molecular species delineation approaches using the COI gene depends on the method used. In our case, the number of recovered clusters varied from 16 to 19. The ASAP delimitation approach generated less clusters than the tree-based methods GMYC and mPTP, which is consistent with the literature (Dvořák et al., 2022; Kekkonen & Hebert, 2014; Puillandre et al., 2021). These results highlight the importance of performing multiple molecular approaches to determine the congruent clusters. However, using multiple molecular markers (ideally from mitochondrial and nuclear DNA) combined with geographically wide sampling is needed to resolve remaining uncertainties. In addition, independent evidence, such as ecological data or morphological characters must be examined on all the specimens for integrative taxonomy and validation of the molecular species hypothesis (Puillandre et al., 2021). Our phylogeny and delimitation analyses also support the previous molecular distinction between *Cirripectes alboapicalis*, *Cirripectes patuki* sensu Delrieu-Trottin et al. (2018), and a new species (Delrieu-Trottin et al., 2018). Therefore, the use of the COI gene alone as a barcode could be sufficient for *Cirripectes* identification.

4.3 | *Cirripectes* in the Mascarene Islands

The type localities of two *Cirripectes* species are in the Mascarene Islands. The holotype of *Salarias castaneus* Valenciennes in Cuvier and Valenciennes (1836) was collected at Isle de France (Mauritius). This species was later assigned to *Cirripectes* and, according to Williams (1988), several times erroneously synonymized with *C. variolosus* (e.g., Smith, 1959). The holotype of *C. randalli* Williams, 1988 originates from Cargados Carajos Shoals and paratypes come from Mauritius. Moreover, one paratype of *C. gilberti* Williams, 1988 is from Agalega. However, to date, few studies comprehensively assessed the diversity and distribution of *Cirripectes* species in the Mascarene archipelago.

Cirripectes occurrence at Agalega, Cargados Carajos Shoals, Mauritius, and Reunion was first described by Williams in Williams, 1988 (Table 3). *C. castaneus* and *C. quagga* were listed for Reunion, with the checklist subsequently completed based on visual records by Letourneur (1992) with *C. polyzona* by Fricke (1999) with *C. stigmaticus* and *C. randalli*, and by Wickel et al. (2020) with *C. filamentosus*. Subsequent checklists were based on the previous ones, sometimes without the addition of new *Cirripectes* species (Fricke et al., 2009; Letourneur et al., 2004). Heemstra et al. (2004) provided a preliminary fish list for Rodrigues and referred to the earlier studies of Gunther (1879), de Baissac (1968) and Fricke (1999) as problematic, due to misidentifications and undocumented sight records (Heemstra et al., 2004). The *Cirripectes* sequences found in public databases were produced by studies not focused on *Cirripectes* but on taxonomically broad barcoding efforts of Indo-Pacific coral-reef fishes (Hubert et al., 2012) and post-larvae (Collet et al., 2017).

Both on Reunion and Rodrigues, *C. castaneus*, *C. randalli*, and *C. stigmaticus* were collected using ARMS. This means that ARMS sampled three out of the 6 *Cirripectes* species listed for Reunion and three out of four species listed for Rodrigues (Table 4). These results were in agreement with the previous molecular studies available for Reunion, which reported *C. castaneus* and *C. randalli* (Collet et al., 2017; Hubert et al., 2015). To the best of our knowledge, we

		C. quagga					*C. stigmaticus*					*C. randalli*						
Ca	Ag	Re	Ma	Ro	Ca	Ag	Re	Ma	Ro	Ca	Ag	Re	Ma	Ro	Ca	Ag	Meth.	Studies
X	x	✓	x	x	✓	✓	x	✓	x	✓	x	x	✓	x	✓	x	M	Williams (1988)
																	M	Debelius (1993)
–	–	x	–	–	–	–	[a]	–	–	–	–	[b]	–	–	–	–	G	Hubert et al. (2015)
–	–	x	–	–	–	–	[a]	–	–	–	–	[b]	–	–	–	–	G	Collet et al. (2017)
–	–	✓	–	–	–	–	✓	–	–	–	–	✓	–	–	–	–	V	Wickel et al. (2020)
–	–	x	–	x	–	–	✓	–	✓	–	–	✓	–	✓	–	–	G	Present study

Le tableau 4 présente des erreurs liées à l'édition de l'article. Un corrigendum est en cours auprès de l'éditeur. Le tableau 4, au bon format, est disponible à la fin de l'article avant les appendices.

provide the first record of *C. randalli* for Rodrigues and the first sequence of *C. stigmaticus* for Reunion.

In spite of the extensive sampling using ARMS at Reunion (46 ARMS deployed at 10 sites in the course of 7 years), the absence of the other three species recorded for the island by visual censuses (*C. filamentosus*, *C. polyzona*, and *C. quagga*) can be explained by several hypotheses: (i) their habitats were not sampled, (ii) ARMS are not suitable to sample these species, (iii) these species are not present on the island, or (iv) were not present at the time of sampling.

The first possibility is that the sampling sites did not include the habitats of the species not sampled. Indeed, we sampled only on the top of spurs at 10–12 m depth, whereas some species (*C. polyzona* and *C. quagga*) are reported to prefer algal ridges and crests in the surf zone (Williams, 1988). Also, the public sequences of *C. randalli* were from specimens collected in tide pools or on shallow reef flats (JHB, pers. comm.). The difference in the habitat sampling may explain the low number of *C. randalli* recovered using ARMS on outer reef slopes.

The second explanation pertains to the sampling ability of ARMS for certain species. The ARMS are deployed on the reef surface, whereas other methods, such as rotenone or clove oil sampling, can potentially access cryptic species that live deeper within the reef matrix. In spite of earlier rotenone sampling at Reunion at various locations and depths (Hubert et al., 2012), only *C. castaneus* and *C. randalli* were recovered. In contrast, ARMS sampling allowed recovering an additional species, *C. stigmaticus*.

The third explanation rests on the hypothesis that the other three species do not occur on Reunion. From the first observation of *Cirripectes*, it has been reported that the species can be highly variable, leading to numerous misidentifications, even when morphology could be studied in detail (Smith, 1959; Williams, 1988). Thus, visual surveys without detailed examination of the morphological characters may not permit to reliably identify all *Cirripectes* species. The species *C. filamentosus*, *C. polyzona*, and *C. quagga* listed for Reunion based exclusively on visual records might correspond to erroneous identifications. Ideally, the visual records of the "missing" species need to be confirmed by genetic analyses.

Fourth, *Cirripectes*, like other cryptobenthic reef fishes, are considered to have an abundant larval supply (Brandl et al., 2019). It is therefore surprising that the "missing" three species have not been recovered during several years of sampling of post-larvae at Reunion (Adeline Collet, pers. comm.). Moreover, the ARMS sampling were carried out over several years and at different times of the year, including hot and cool seasons. In view of this extensive and diverse sampling, it is unlikely that the three species were missed due to a seasonal effect, and this further supports the hypothesis that these species may in fact not be present at Reunion. Moreover, *C. castaneus* collected with ARMS had smaller sizes (median of 37 mm) than rotenone-collected specimens from shallow reef flats (median of 71 mm; JHB, pers. comm.), suggesting that ARMS provide habitat for juvenile individuals following recruitment.

Molecular studies are not limited by morphological identification and allow reliable identification of specimens and matching among studies. For this reason, we can confirm with certitude the presence of *C. castaneus*, *C. stigmaticus*, and *C. randalli* in Reunion and Rodrigues. In the same way the COI marker is highly useful for *Cirripectes* identification, the mitochondrial dataset provided here should allow the detection of the three 'missing' species with an eDNA metabarcoding study and resolve the uncertainty about their occurrence in the Mascarene.

4.4 | Geographical ranges of Mascarene *Cirripectes*

The geographic ranges of the three *Cirripectes* species that we collected in the Mascarene Islands, *C. castaneus*, *C. stigmaticus*, and *C. randalli*, show different patterns. *C. randalli* was reported as an endemic species of the Mascarene Islands and listed in Mauritius, Cargados Carajos Shoals, and Reunion (Fricke, 1999; Williams, 1988). The presence of *C. randalli* in Reunion confirmed in this study is consistent with the literature and previous molecular studies. To the best of our knowledge, this is the first report of *C. randalli* from Rodrigues, which is congruent with its known geographical range. *C. castaneus* and *C. stigmaticus* are both reported to have an Indo-Pacific wide distribution. Our analyses confirm this and showed a lack of divergence and genetic structuration among sequences from various locations. Moreover, for both species, several haplotypes were shared among distant localities such as Seychelles, Reunion, and New Caledonia for *C. castaneus*, and the Mascarene Islands to Fiji and Wallis & Futuna for *C. stigmaticus*. Surprisingly, for other species reported to have very wide distributions (*C. auritus*, *C. filamentosus*, *C. polyzona*, *C. quagga*, and *C. variolosus*) no sequences have been recovered from geographically distant sites. These results may be an artifact of uneven sampling efforts among localities. Alternatively, given the recent highlight of cryptic species with small endemism areas in the genus (Delrieu-Trottin et al., 2018; Hoban & Williams, 2020), this could be the outcome of multiple cryptic lineages in the genus and in need of further investigation. Indeed, the presence of cryptic species could have implications for understanding mechanisms driving biodiversity patterns (Eme et al., 2018). Cryptic species have distinct evolutionary patterns and, in some cases, a restricted geographical range with specialized behavior or higher threat susceptibility. Therefore, the detection of cryptic species is crucial for the conservation and/or management of marine biodiversity (Bickford et al., 2007).

4.5 | Extending and using DNA databases for specimen identification

It is of utmost importance to share molecular datasets that can be cross-checked and used openly for taxonomic or identification purposes. Private datasets limit error detection and correction, and the comparability of many current metabarcoding studies is limited because of the inaccessibility of the datasets used for taxon assignment. The newly added sequences expand the reference

public DNA database of *Cirripectes* to 85 individual sequences. The new sequences are all attached to a specimen deposited in a registered collection and correspond therefore to genseq-3 (collection-vouchered non-types) according to the nomenclature of Chakrabarty et al. (2013). Of these, 24 represent the first mitogenome sequences for the genus and these three species. These additions will support future studies using mitochondrial markers other than COI for specimen identification or phylogenetic reconstruction. Available complete mitogenomes enable the selection of the most appropriate marker to respond to the objectives of such studies. Likewise, we deposited the first *Cirripectes* sequences from Rodrigues for three species and added 23 sequences to the 14 *Cirripectes* sequences from Reunion. Finally, this study added 45 new sequences for the rhodopsin gene (27 specimens with 18 heterozygotes) to the existing 12 public sequences (8 specimens with 3 heterozygotes). There are still very few *Cirripectes* sequences for nuclear markers, and the rhodopsin gene was one of the most represented with eight sequences.

Sequences previously deposited in the BOLD and GenBank databases and previous studies allowed us to assign the specimens we collected to two taxonomic names. Further analyses revealed the existence of misidentified specimens in these databases. Some of these misidentifications have already been corrected and published but were not yet corrected in the BOLD database (BOLD:ADB2362 probably do not correspond to *Cirripectes* specimens in Chu et al., 2019; BOLD:AAU0601 is *C. randalli* in Hoban & Williams, 2020). The presence and problem of misidentified specimens in reference databases are well documented, even for Indo-Pacific fishes (Leis, 2015; Pentinsaari et al., 2020). In fact, even when authors are aware of this problem, the wrong assignation may occur. Indeed, if a specimen is assigned to a BOLD BIN that contains specimens assigned to different species with no obvious errors about the geographic distribution, deciding which name is correct is problematic. Similar problems arise when a species corresponds to several BOLD BINs. To resolve these uncertainties, a new morphological examination of the specimens must be performed and, if needed, an investigation of phylogenetic relationships, possibly with added specimens and species (Ward et al., 2009).

4.6 | Primer selection and perspectives

Among the technical issues in molecular ecology, the choice of primer for PCR amplification is one of the most important factors affecting the probability of species detection. The eDNA approach led to the design of primers to respond to the new constraints, such as short ID sequences (<200 bp) to improve PCR success with degraded eDNA (Bylemans et al., 2018; Freeland, 2017). In the case of *Cirripectes*, two widely used primer pairs targeting the 12S were tested for comparison with results from the longer COI barcode. Unlike the proprietary Teleo primers (Valentini et al., 2016), the MiFish primer set failed to amplify species *in sillico*, which contrasts with the results of Zhang et al. (2020) that showed that the latter primers had a larger detection range than the Teleo primer set. The genome areas targeted by COI and Teleo primers had a divergence >3% and thus allow automated species identification using the commonly used threshold for discriminating potential species. However, for the 12S targeted by the Teleo primers, only one sequence was available as a reference for this genus before our study, therefore, only the COI barcodes could truly be used for species identification.

In the future, primer limitations for eDNA surveys will be overcome with the development of shotgun sequencing for the eDNA, by direct sequencing of total eDNA and bypassing the PCR limitations associated with metabarcoding to provide insights into community composition (Taberlet et al., 2012; Tringe & Rubin, 2005). Currently, this approach is still limited by the completeness of the reference databases, in terms of species diversity (as barcoding) and in terms of reference genomes coverage (e.g., complete mitochondrial DNA).

Finally, multiplexed NGS sequencing of long amplicons produced cost-efficient complete mitogenomes, providing sequences for markers with greater variability than available previously. In the case of *Cirripectes*, ND1 and ND2 may be used instead of COI to resolve the remaining uncertainties. Moreover, using reads from NGS sequencing enables easy differentiation of alleles for nuclear markers.

5 | CONCLUSION

In this study, the geographic distribution and species relationships within genus *Cirripectes* were examined. The major conclusions are summarized below:

1. At Reunion, numerous replicates of ARMS allowed sampling of one additional species in comparison with rotenone or light-trap captures, repeated in space and time.
2. The presence of *C. castaneus*, *C. randalli*, and *C. stigmaticus* in Reunion and Rodrigues was confirmed. However, the species *C. filamentosus*, *C. polyzona*, and *C. quagga*, listed as present in Reunion based on visual identifications, were not found and may correspond to erroneous identifications.
3. The generated data contribute to filling the gaps in taxonomic and molecular knowledge of reef cryptobiome for the South-West Indian Ocean, and the first complete mitogenomes of three *Cirripectes* species are provided.
4. Both the COI gene and the target of the eDNA Teleo primer set have interspecific divergence >3% and allow species identification for Mascarene *Cirripectes*.
5. COI sequences are not sufficient to clearly resolve the relationships among *Cirripectes* species.

AUTHOR CONTRIBUTIONS

Marion Couëdel: Conceptualization (equal); data curation (lead); formal analysis (lead); investigation (lead); project administration (equal); visualization (lead); writing – original draft (lead); writing – review and editing (lead). **Agnes Dettai:** Conceptualization (equal); formal analysis (supporting); funding acquisition (supporting); investigation (equal); methodology (lead); project administration (equal);

validation (lead); writing – review and editing (equal). **Mireille M. M. Guillaume:** Conceptualization (equal); investigation (equal); project administration (equal); writing – review and editing (equal). **Fleur Bruggemann:** Investigation (equal). **Sophie Bureau:** Investigation (equal). **Baptiste Frattini:** Investigation (equal). **Amélie Verde Ferreira:** Investigation (equal). **Jean-Lindsay Azie:** Investigation (equal). **J. Henrich Bruggemann:** Conceptualization (lead); funding acquisition (lead); investigation (equal); project administration (lead); writing – review and editing (equal).

ACKNOWLEDGMENTS

This study was supported by the research program *Fonds européen de développement régional* (FEDER) 20171591-0002633 CALIBIOME 2017–2022. ARMS deployments at Reunion were conducted under permit n°2018-61 DEAL/SEB/UBIO and n°2020-09-DEAL/SEB/UBIO of the *Direction de l'environnement, de l'aménagement et du logement de La Réunion*, and permit n°2019-083 and n°2020-054 of the *Direction de la mer Sud océan Indien*. The campaign in Rodrigues was approved by the Rodrigues Regional Assembly (Ref: RA 402/17 Vol II) and the Ministry of Foreign Affairs of the Republic of Mauritius (Ref: 50/1/45), and funded by the FEDER Biodiversity Program in Dec. 2014 and UMR ENTROPIE, University of Reunion in Jan. 2017. Sampling was in conformity with the Nagoya protocol (declaration n°3040030). The Rodrigues Hospital kindly provided ethanol for DNA preservation. Sabrina Meunier and Runolph Raffaut let us use the facilities of the NGO Shoals Rodrigues, while Michel Fontaine and Philippe Faconnier permitted using the Nautical Base of the Municipality of Saint-Pierre. The participation in the processing of ARMS samples of high school students Jean Jefferson Albert, Marie Olivia Babet, Daphné Bruggemann, Sinedia Emilien, Robert Jolicoeur, of BSc student Bryan Young, and of Master students Gwennaïs Fustemberg and Auriane Serval, is greatly appreciated. Marion Couëdel has a PhD fellowship provided by the European Union FSE programme. The authors thank all the persons who helped during field sampling and lab work. Christine Carrau, *Bibliothèque Théodore Monod* at the MNHN, was helpful in getting the old literature. Laboratory work was made possible by access to the *Service de Systématique Moléculaire* of the MNHN (UAR 2700 2AD) and help from its team. We are grateful to Gael Denys for his comments on the manuscript. Finally, we thank the reviewers for constructive comments on an earlier version of the manuscript.

CONFLICT OF INTEREST STATEMENT

The authors state that they have no conflicting interests.

DATA AVAILABILITY STATEMENT

All sequences produced in this study were deposited on GenBank and Bold databases. Accession numbers are provided in Appendix S2 and at the following DOI: 10.5281/zenodo.7599910. R code to reproduce findings is available on MC (Mcouedel) github.

ORCID

Marion Couëdel https://orcid.org/0000-0002-6959-1422
Agnes Dettai https://orcid.org/0000-0002-8496-9737
Mireille M. M. Guillaume https://orcid.org/0000-0001-7249-0131
J. Henrich Bruggemann https://orcid.org/0000-0001-8764-3452

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Allen, G. R. (2015). Review of Indo-Pacific coral reef fish systematics: 1980 to 2014. *Ichthyological Research*, *62*(1), 2–8. https://doi.org/10.1007/s10228-014-0411-1

Allen, G. R., Erdmann, M. V., Randall, J. E., Ching, P., Rauzon, M. J., Hayashi, L. A., Thomas, M. D., Robertson, D. R., Taylor, L., & Coste, M. (2013). Reef fishes of the east indies. *Philosophy East and West*, *63*(2), 1292.

Alleyne, H. G., & Macleay, W. (1877). The ichthyology of the Chevert expedition. *Proceedings of the Linnean Society of New South Wales* v. 1 (pts 3–4): 261–281, 321–359, Pls. 3–9, 10–17. [pp. 261–281, Pls. 3–9 published Feb. 1877; pp. 321–359, Pls. 10–17 published Mar. 1877].

Avise, J. C. (2009). Phylogeography: Retrospect and prospect. *Journal of Biogeography*, *36*(1), 3–15. https://doi.org/10.1111/j.1365-2699.2008.02032.x

Bagley, J. C. (2018). *Setting DNA substitution models in BEAST - blog*. Justin Bagley. https://justinbagley.rbind.io/2016/10/11/setting-dna-substitution-models-beast/

Bandelt, H. J., Forster, P., & Röhl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, *16*(1), 37–48. https://doi.org/10.1093/oxfordjournals.molbev.a026036

Bellwood, D. R., Streit, R. P., Brandl, S. J., & Tebbett, S. B. (2019). The meaning of the term 'function' in ecology: A coral reef perspective. *Functional Ecology*, *33*, 948–961.

Bickford, D., Lohman, D. J., Sodhi, N. S., Ng, P. K. L., Meier, R., Winker, K., Ingram, K. K., & Das, I. (2007). Cryptic species as a window on diversity and conservation. *Trends in Ecology & Evolution*, *22*(3), 148–155. https://doi.org/10.1016/j.tree.2006.11.004

Bleeker, P. (1868). Description de deux espèces nouvelles de Blennioïdes de l'Inde archipélagique. *Verslagen en Mededeelingen der Koninklijke Akademie van Wetenschappen. Afdeeling Natuurkunde (Ser. 2)*, *2*, 278–280.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., & Drummond, A. J. (2014). BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, *10*(4), e1003537. https://doi.org/10.1371/journal.pcbi.1003537

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*(4), 1846–1873. https://doi.org/10.1111/brv.12423

Brandl, S. J., Tornabene, L., Goatley, C. H. R., Casey, J. M., Morais, R. A., Côté, I. M., Baldwin, C. C., Parravicini, V., Schiettekatte, N. M. D., & Bellwood, D. R. (2019). Demographic dynamics of the smallest marine vertebrates fuel coral reef ecosystem functioning. *Science*, *364*(6446), 1189–1192. https://doi.org/10.1126/science.aav3384

Breder, C. M., & Rosen, D. E. (1966). Reproduction in fishes: Modes of reproduction in fishes. *Science*, *155*(3763), 684. https://doi.org/10.1126/science.155.3763.684.b

Brown, S. D. J., Collins, R. A., Boyer, S., Lefort, M.-C., Malumbres-Olarte, J., Vink, C. J., & Cruickshank, R. H. (2012). Spider: An R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. *Molecular Ecology Resources*, *12*(3), 562–565. https://doi.org/10.1111/j.1755-0998.2011.03108.x

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 & Technology*, *52*(11), 6408–6416. https://doi.org/10.1021/acs.est.8b01071

Carlson, B. A. (1981). A new Indo-Pacific fish of the genus *Cirripectes* (Blenniidae, Salariini). *Pacific Science*, *34*(4), 407–414.

Chakrabarty, P., Warren, M., Page, L. M., & Baldwin, C. C. (2013). GenSeq: An updated nomenclature and ranking for genetic sequences from type and non-type sources. *ZooKeys*, *346*, 29–41. https://doi.org/10.3897/zookeys.346.5753

Chen, W.-J., Bonillo, C., & Lecointre, G. (2003). Repeatability of clades as a criterion of reliability: A case study for molecular phylogeny of Acanthomorpha (Teleostei) with larger number of taxa. *Molecular Phylogenetics and Evolution*, *26*(2), 262–288. https://doi.org/10.1016/S1055-7903(02)00371-8

Chu, C., Loh, K. H., Ng, C. C., Ooi, A. L., Konishi, Y., Huang, S.-P., & Chong, V. C. (2019). Using DNA barcodes to aid the identification of larval fishes in tropical estuarine waters (Malacca Straits, Malaysia). *Zoological Studies*, *58*, e30. https://doi.org/10.6620/ZS.2019.58-30

Collet, A., Durand, J.-D., Desmarais, E., Cerqueira, F., Cantinelli, T., Valade, P., & Ponton, D. (2017). DNA barcoding post-larvae can improve the knowledge about fish biodiversity: An example from La Reunion, SW Indian Ocean. *Mitochondrial DNA Part A*, *29*(6), 905–918. https://doi.org/10.1080/24701394.2017.1383406

Collins, R. A., Boykin, L. M., Cruickshank, R. H., & Armstrong, K. F. (2012). Barcoding's next top model: An evaluation of nucleotide substitution models for specimen identification. *Methods in Ecology and Evolution*, *3*(3), 457–465. https://doi.org/10.1111/j.2041-210X.2011.00176.x

Cuvier, G., & Valenciennes, A. (1836). *Histoire naturelle des poissons (Chez Pitois Levrault, Vol. 14)* (p. 586). Bertrand.

de Baissac, J. B. (1968). Some notes on the fish species of Rodrigues. *Proceedings of the Royal Society of Arts and Science. Mauritius*, *3*(1), 45–63.

Debelius, H. (1993). *Indian Ocean: Tropical fish guide* (1st ed., p. 321). Aquaprint Verlags GmbH.

Delrieu-Trottin, E., Liggins, L., Trnski, T., Williams, J. T., Neglia, V., Rapu-Edmunds, C., Planes, S., & Saenz-Agudelo, P. (2018). Evidence of cryptic species in the blenniid *Cirripectes alboapicalis* species complex, with zoogeographic implications for the South Pacific. *ZooKeys*, *810*, 127–138. https://doi.org/10.3897/zookeys.810.28887

Douzery, E. J. P. (2010). Phylogénie moléculaire. In *Biologie evolutive* (814 pp.). De Boeck CNRS.

Dvořák, T., Šlechtová, V., & Bohlen, J. (2022). Using species groups to approach the large and taxonomically unresolved freshwater fish family Nemacheilidae (Teleostei: Cypriniformes). *Biology*, *11*(2), 175. https://doi.org/10.3390/biology11020175

Edgar, R. C. (2004). MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*(1), 113. https://doi.org/10.1186/1471-2105-5-113

Eme, D., Zagmajster, M., Delić, T., Fišer, C., Flot, J.-F., Konecny-Dupré, L., Pálsson, S., Stoch, F., Zakšek, V., Douady, C. J., & Malard, F. (2018). Do cryptic species matter in macroecology? Sequencing European groundwater crustaceans yields smaller ranges but does not challenge biodiversity determinants. *Ecography*, *41*(2), 424–436. https://doi.org/10.1111/ecog.02683

Excoffier, L., & Lischer, H. E. L. (2010). Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and windows. *Molecular Ecology Resources*, *10*, 564–567.

Ezart, T., Fusjisawa, T., & Barraclough, T. (2009). Splits: SPecies' LImits by threshold statistics version 1.0-20 from R-forge. https://rdrr.io/rforge/splits/

Fowler, H. W., & Ball, S. C. (1924). Descriptions of new fishes obtained by the Tanager Expedition of 1923 in the Pacific islands west of Hawaii. *Proceedings of the Academy of Natural Sciences of Philadelphia*, *76*, 269–274.

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

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

Fricke, R. (1999). *Fishes of the Mascarene Islands (Réunion, Mauritius, Rodriguez): An annotated checklist, with descriptions of new species* (Vol. *31*, p. 759). Koeltz Scientific Books, Koenigstein, Theses Zoologicae.

Fricke, R., Mulochau, T., Durville, P., Chabanet, P., & Tessier, E. (2009). Annotated checklist of the fish species (Pisces) of La Réunion. Including a red list of threatened and declining species. *Stuttgarter Beiträge Zur Naturkunde A Neue Serie*, *2*, 1–168.

Fujisawa, T., & Barraclough, T. G. (2013). Delimiting species using Single-Locus Data and the Generalized Mixed Yule Coalescent Approach: A revised method and Evaluation on Simulated Data Sets. *Systematic Biology*, *62*(5), 707–724.

Geller, J., Meyer, C., Parker, M., & Hawk, H. (2013). Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. *Molecular Ecology Resources*, *13*(5), 851–861. https://doi.org/10.1111/1755-0998.12138

Geneious Prime 2019.2.3. (2019). https://www.geneious.com

Grant, W. S., & Bowen, B. W. (1998). Shallow population histories in deep evolutionary lineages of marine fishes: Insights from sardines and anchovies and lessons for conservation. *The Journal of Heredity*, *89*(5), 415–426. https://doi.org/10.1093/jhered/89.5.415

Gunther, A. C. L. G. (1879). Fishes. In *An account of the botanical collections made in Kerguelen's Island during the transit of Venus expedition in the years 1874–75: Vol. zoology of Rodriguez* (pp. 470–472). Royal Society.

Hastings, P. A., & Springer, V. G. (2009). 3: Systematics of the Blenniidae (Combtooth blennies). In *The biology of blennies* (Vol. *24*, 1st ed., pp. 494). CRC Press.

Heemstra, E., Heemstra, P., Smale, M., Hooper, T., & Pelicier, D. (2004). Preliminary checklist of coastal fishes from the Mauritian Island of Rodrigues. *Journal of Natural History*, *38*(23–24), 3315–3350. https://doi.org/10.1080/00222930410001695088

Hinsinger, D. D., Debruyne, R., Thomas, M., Denys, G. P. J., Mennesson, M., Utage, J., & Dettai, A. (2015). Fishing for barcodes in the torrent: From COI to complete mitogenomes on NGS platforms. *DNA Barcodes*, *3*(1), 170–186. https://doi.org/10.1515/dna-2015-0019

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, *35*(2), 518–522. https://doi.org/10.1093/molbev/msx281

Hoban, M. L., & Williams, J. T. (2020). *Cirripectes matatakaro*, a new species of combtooth blenny from the Central Pacific, illuminates the origins of the Hawaiian fish fauna. *PeerJ*, *8*, e8852. https://doi.org/10.7717/peerj.8852

Hubert, N., Dettai, A., Pruvost, P., Cruaud, C., Kulbicki, M., Myers, R., & Borsa, P. (2017). Geography and life history traits account for the accumulation of cryptic diversity among Indo-West Pacific coral reef fishes. *Marine Ecology Progress Series*, *583*, 179–193. https://doi.org/10.3354/meps12316

Hubert, N., Espiau, B., Meyer, C., & Planes, S. (2015). Identifying the ichthyoplankton of a coral reef using DNA barcodes. *Molecular Ecology Resources*, *15*(1), 57–67. https://doi.org/10.1111/1755-0998.12293

Hubert, N., Meyer, C. P., Bruggemann, H. J., Guérin, F., Komeno, R. J. L., Espiau, B., Causse, R., Williams, J. T., & Planes, S. (2012). Cryptic diversity in Indo-Pacific coral-reef fishes revealed by DNA-barcoding

provides new support to the Centre-of-overlap hypothesis. *PLoS One*, *7*(3), e28987. https://doi.org/10.1371/journal.pone.0028987

Iwasaki, W., Fukunaga, T., Isagozawa, R., Yamada, K., Maeda, Y., Satoh, T. P., Sado, T., Mabuchi, K., Takeshima, H., Miya, M., & Nishida, M. (2013). MitoFish and MitoAnnotator: A mitochondrial genome database of fish with an accurate and automatic annotation pipeline. *Molecular Biology and Evolution*, *30*(11), 2531–2540. https://doi.org/10.1093/molbev/mst141

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587–589. https://doi.org/10.1038/nmeth.4285

Kekkonen, M., & Hebert, P. D. N. (2014). DNA barcode-based delineation of putative species: Efficient start for taxonomic workflows. *Molecular Ecology Resources*, *14*(4), 706–715. https://doi.org/10.1111/1755-0998.12233

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, *16*(2), 111–120. https://doi.org/10.1007/BF01731581

Knowlton, N., Brainard, R. E., Fisher, R., Moews, M., Plaisance, L., & Caley, M. J. (2010). Coral reef biodiversity. In *Life in the World's oceans: Diversity distribution and abundance* (pp. 65–74). Blackwell Publishing Ltd.

Kocher, T. D., Thomas, W. K., Meyer, A., Edwards, S. V., Pääbo, S., Villablanca, F. X., & Wilson, A. C. (1989). Dynamics of mitochondrial DNA evolution in animals: Amplification and sequencing with conserved primers. *Proceedings of the National Academy of Sciences of the United States of America*, *86*(16), 6196–6200. https://doi.org/10.1073/pnas.86.16.6196

Leigh, J. W., & Bryant, D. (2015). Popart: Full-feature software for haplotype network construction. *Methods in Ecology and Evolution*, *6*(9), 1110–1116. https://doi.org/10.1111/2041-210X.12410

Leis, J. M. (2015). Taxonomy and systematics of larval Indo-Pacific fishes: A review of progress since 1981. *Ichthyological Research*, *62*(1), 9–28. https://doi.org/10.1007/s10228-014-0426-7

Letourneur, Y. (1992). *Dynamique des peuplements ichtyologiques des platiers récifaux de l'île de La Réunion* [PhD Thesis]. Aix-Marseille 2.

Letourneur, Y., Chabanet, P., Durville, P., Taquet, M., Teissier, E., Parmentier, M., Quero, J.-C., & Pothin, K. (2004). An updated checklist of the marine fish fauna of Reunion Island, South-Western Indian Ocean. *Cybium*, *28*(3), 199–216.

Lin, H.-C., & Hastings, P. A. (2013). Phylogeny and biogeography of a shallow water fish clade (Teleostei: Blenniiformes). *BMC Evolutionary Biology*, *13*(1), 210. https://doi.org/10.1186/1471-2148-13-210

Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, *2010*, pdb.prot5448. https://doi.org/10.1101/pdb.prot5448

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

Mora, C., Tittensor, D. P., & Myers, R. A. (2008). The completeness of taxonomic inventories for describing the global diversity and distribution of marine fishes. *Proceedings: Biological Sciences*, *275*(1631), 149–155.

Pearman, J. K., Leray, M., Villalobos, R., Machida, R. J., Berumen, M. L., Knowlton, N., & Carvalho, S. (2018). Cross-shelf investigation of coral reef cryptic benthic organisms reveals diversity patterns of the hidden majority. *Scientific Reports*, *8*(1), 1–17. https://doi.org/10.1038/s41598-018-26332-5

Pentinsaari, M., Ratnasingham, S., Miller, S. E., & Hebert, P. D. N. (2020). BOLD and GenBank revisited – Do identification errors arise in the lab or in the sequence libraries? *PLoS One*, *15*(4), e0231814. https://doi.org/10.1371/journal.pone.0231814

Puillandre, N., Brouillet, S., & Achaz, G. (2021). ASAP: Assemble species by automatic partitioning. *Molecular Ecology Resources*, *21*(2), 609–620. https://doi.org/10.1111/1755-0998.13281

R Core Team. (2021). *R: A language and environment for statistical computing (4.1)*. R Foundation for Statistical Computing. https://www.R-project.org/

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics using tracer 1.7. *Systematic Biology*, *67*(5), 901–904. https://doi.org/10.1093/sysbio/syy032

Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The barcode of life data system. *Molecular Ecology Notes*, *7*, 355–364. https://doi.org/10.1111/j.1471-8286.2006.01678.x

Ronquist, F., & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, *19*(12), 1572–1574. https://doi.org/10.1093/bioinformatics/btg180

Smith, J. L. B. (1959). Fishes of the families Blenniidae and Salariidae of the Western Indian Ocean. *Ichtyological Bulletin*, *14*, 229–252.

Srivathsan, A., & Meier, R. (2012). On the inappropriate use of Kimura-2-parameter (K2P) divergences in the DNA-barcoding literature. *Cladistics*, *28*(2), 190–194. https://doi.org/10.1111/j.1096-0031.2011.00370.x

Strasburg, D. W., & Schultz, L. P. (1953). The blenniid fish genera *Cirripectus* and Exallias with descriptions of two new species from the tropical Pacific. *Journal of the Washington Academy of Sciences*, *43*(4), 128–135.

Steinke, D., Zemlak, T. S., & Hebert, P. D. N. (2009). Barcoding nemo: DNA-based identifications for the ornamental fish trade. *PLoS One*, *4*(7), e6300. https://doi.org/10.1371/journal.pone.0006300

Swainson, W. (1839). On the natural history and classification of fishes, amphibians and reptiles. *Longman, Orme, Brown, Green & Longmans*, *386*. https://doi.org/10.5962/bhl.title.62140

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*(8), 2045–2050.

Trifinopoulos, J., Nguyen, L.-T., von Haeseler, A., & Minh, B. Q. (2016). W-IQ-TREE: A fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Research*, *44*(W1), W232–W235. https://doi.org/10.1093/nar/gkw256

Tringe, S. G., & Rubin, E. M. (2005). Metagenomics: DNA sequencing of environmental samples. *Nature Reviews Genetics*, *6*(11), 805–814. https://doi.org/10.1038/nrg1709

Valenciennes. (1836). Tome onzième. Livre treizième. De la famille des Mugiloïdes. Livre quatorzième. De la famille des Gobioïdes. In G. Cuvier & A. Valenciennes (Eds.), *Histoire naturelle des poissons* (506 pp.). Levrault.

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

Ward, R. D., Hanner, R., & Hebert, P. D. N. (2009). The campaign to DNA barcode all fishes, FISH-BOL. *Journal of Fish Biology*, *74*(2), 329–356. https://doi.org/10.1111/j.1095-8649.2008.02080.x

Watson, W. (2009). 4: Larval Development in Blennies. In *The biology of blennies* (pp. 321–362). CRC Press. https://doi.org/10.1201/b10301-19

Wickel, J., Fricke, R., Durville, P., Chabanet, P., Dumestre, M., Mulochau, T., Pinault, M., & Tessier, E. (2020). Updated checklist of the fish species of Reunion Island. In Dumestre & Wickel (Ed.), *Évaluation de statuts de conservation pour les poisons récifaux à la Réunion* (pp. 1, 974–25). Marex/Deal.

Williams, J. T. (1988). Revision and phylogenetic relationships of the blenniid fish genus *Cirripectes*. *Indo-Pacific Fishes*, *17*, 1–78.

Williams, J. T. (2010). A new species of blenny, *Cirripectes heemstraorum*, from cape Vidal, South Africa (family Blenniidae). *Smithiana Bulletin*, *12*, 3–7.

Zhang, D., Gao, F., Jakovlić, I., Zou, H., Zhang, J., Li, W. X., & Wang, G. T. (2020). PhyloSuite: An integrated and scalable desktop platform for streamlined molecular sequence data management and evolutionary phylogenetics studies. *Molecular Ecology Resources*, *20*(1), 348–355. https://doi.org/10.1111/1755-0998.13096

Zhang, J., Kapli, P., Pavlidis, P., & Stamatakis, A. (2013). A general species delimitation method with applications to phylogenetic placements. *Bioinformatics (Oxford, England)*, *29*(22), 2869–2876. https://doi.org/10.1093/bioinformatics/btt499

Zimmerman, T. L., & Martin, J. W. (2004). Artificial reef matrix structures (Arms): An inexpensive and effective method for collecting coral reef-associated invertebrates. *Gulf and Caribbean Research*, *16*(1), 59–64. https://doi.org/10.18785/gcr.1601.08

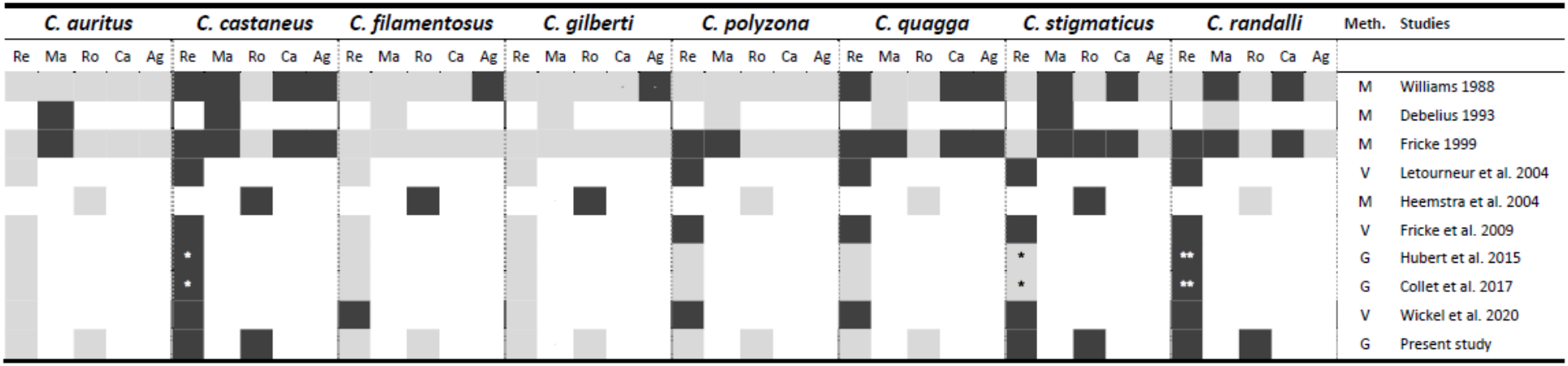
SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Couëdel, M., Dettai, A., Guillaume, M. M. M., Bruggemann, F., Bureau, S., Frattini, B., Verde Ferreira, A., Azie, J.-L., & Bruggemann, J. H. (2023). New insights into the diversity of cryptobenthic *Cirripectes* blennies in the Mascarene Archipelago sampled using Autonomous Reef Monitoring Structures (ARMS). *Ecology and Evolution*, *13*, e9850. https://doi.org/10.1002/ece3.9850

Tableau 4 :

Table 4: *Cirripectes* species reported from Mascarene Islands. Dark grey for present, light grey for not observed and white for not evaluated. Methods of identification were synthetized as following (Meth.): M for morphometric analysis, V for visual survey, and G for molecular identification. Ag: Agalega Islands; Ca: Cargados Carajos Shoals (Saint Brandon); Ma: Mauritius; Ro: Rodrigues; Re: Reunion.



* Misidentification between *C. stigmaticus* and *C. castaneus*

** Misidentification between *C. randalli* and *C. castaneus*

Appendices

Appendix 1: ARMS deployments and *Cirripectes* spp collected. * designates sites located inside no-fishing zones.

Island	Site	Soak time (yrs)	Deployment	Recovery	ARMS recovered	Latitude (N)	Longitude (E)	*Cirripectes castaneus*	*Cirripectes randalli*	*Cirripectes stigmaticus*	Unid.	Total
Reunion	RUNA2*	0.6	08/09/2014	22-24/04/2015	3	-21.103883	55.235917	2				2
	RUNA2*	4.5	08/04/2014	17/12/2018	2	-21.103883	55.235917	1				1
	GILA1*	4	01/11/2014	18, 19/12/2018	2	-21.077183	55.215967	2				2
	RUNA1	2	23/01/2019	14, 17/12/2020	3	-21.02060	55.22983	1				1
	RUNA2*	2	17/12/2018	19/12/2020, 13/01/2021	3	-21.10401	55.23598	5	1	1	1	8
	RUNA3*	2	23/01/2019	15, 17/12/2020	3	-21.10411	55.23604	7				7
	RUNA4*	2	18/12/2018	19/12/2020, 08/01/2021	3	-21.17256	55.28228					
	RUNA5*	2	18/12/2018	18/12/2020, 15/01/2021	3	-21.17352	55.28249	3				3
	RUNA6	2	23/01/2019	18/12/2020, 14/01/2021	3	-21.191040	55.282821					
	RUNA7*	2	25/01/2019	19, 20/01/2021	3	-21.269565	55.327891				1	1
	RUNA8	2	25/01/2019	18/01/2021, 08/02/2021	3	-21.35209	55.48074	1			1	2
	RUNA9	2	25/01/2019	20/01/2021, 09/02/2021	3	-21.37236	55.54692	1			1	2
	RUNA2*	0.5	25/02/2020	27, 31/08/2020	3	-21.10401	55.23598	1				1
	RUNA2*	1	25/02/2020	20, 22/02/2021	3	-21.10401	55.23598				1	1
	RUNA2*	0.5	31/08/2020	19, 20/02/2021	3	-21.10401	55.23598	3				3
	RUNA2*	1	31/08/2020	26, 30/08/2021	3	-21.10401	55.23598					
Rodrigues	RODA1	2	08/12/2015	11,12/01/2017	3	-19.66878	63.47387					
	RODA2	2	10/12/2015	13, 16/01/2017	3	-19.65262	63.40073	1	1	2		4
	RODA3	2	11/12/2015	16, 17/01/2017	2	-19.65397	63.41513			1		1
Total					54			28	2	4	5	39

Appendix 2: Sequences generated during this study and specimen total length

Field ID	Collection number (MNHN-IC-2023-)	Species	GenBank Accession Number			BOLD	Collection Date	Country	Site	ARMS	Soak time (yrs)	Season	Length (mm)
			Complete mitogenome	COI	Rho								
RUNA_0272	0248	*castaneus*	OP820448			IOACT113-22	2018	Reunion	GILA1	GILA1A	4	Hot	35
RUNA_0370	0247	*castaneus*	OP820446			IOACT112-22	2018	Reunion	GILA1	GILA1B	4	Hot	39
RUNA_2168	0222	*castaneus*	OP749990		OP776358 / OP776359	IOACT040-21	2020	Reunion	RUNA1	RUNA1B	2	Hot	36
ORCIE1027	0250	*castaneus*	OP820445			IOACT115-22	2018	Reunion	RUNA2	SALA1A	4	Hot	62
ORCIE1125	0251	*castaneus*	OP820447			IOACT116-22	2018	Reunion	RUNA2	SALA1B	4	Hot	56
RUNA_0022	0249	*castaneus*	OP820444			IOACT114-22	2018	Reunion	RUNA2	SALA1D	4	Hot	35
RUNA_0737	0219	*castaneus*	OP749987		OP776352 / OP776353	IOACT025-21	2020	Reunion	RUNA2	CINA1C	0,5	Cool	38
RUNA_2899	0229	*castaneus*	OP749997			IOACT053-21	2020	Reunion	RUNA2	RUNA2A	2	Hot	NA
RUNA_3154	0230	*castaneus*	OP749998		OP776378 / OP776379	IOACT057-21	2021	Reunion	RUNA2	RUNA2B	2	Hot	67
RUNA_3155	0231	*castaneus*	OP749999		OP776380 / OP776381	IOACT058-21	2021	Reunion	RUNA2	RUNA2B	2	Hot	30
RUNA_3176	0232	*randalli*	OP749984*		OP776382 / OP776383	IOACT059-21	2021	Reunion	RUNA2	RUNA2B	2	Hot	44
RUNA_3214	0233	*castaneus*	OP750000		OP776384 / OP776385	IOACT061-21	2021	Reunion	RUNA2	RUNA2C	2	Hot	29
RUNA_3262	0242	*stigmaticus*	OP575312			IOACT088-22	2021	Reunion	RUNA2	RUNA2C	2	Hot	NA
RUNA_3895	0235	*castaneus*	OP750001		OP776388 / OP776389	IOACT075-21	2021	Reunion	RUNA2	CINA3B	0,5	Hot	32
RUNA_4001	0236	*castaneus*	OP750002		OP776390 / OP776391	IOACT078-21	2021	Reunion	RUNA2	CINA3C	0,5	Hot	47
RUNA_4002	0237	*castaneus*	OP750003		OP776392 / OP776393	IOACT079-21	2021	Reunion	RUNA2	CINA3C	0,5	Hot	34
RUNA_3256	0234	*castaneus*	OP749986		OP776386 / OP776387	IOACT063-21	2021	Reunion	RUNA2	RUNA2C	2	Hot	54
RUNA_2344	0224	*castaneus*	OP749991		OP776360 / OP776361	IOACT043-21	2020	Reunion	RUNA3	RUNA3A	2	Hot	48
RUNA_2346	0241	*castaneus*	OP749992		OP776362 / OP776363	IOACT087-22	2020	Reunion	RUNA3	RUNA3A	2	Hot	29
RUNA_2347	0225	*castaneus*	OP749993		OP776364 / OP776365	IOACT044-21	2020	Reunion	RUNA3	RUNA3A	2	Hot	34

RUNA_2348	0240	*castaneus*		OP787979	OP776366 / OP776367	IOACT086-22	2020	Reunion	RUNA3	RUNA3A	2	Hot	42
RUNA_2379	0226	*castaneus*	OP749994		OP776368 / OP776369	IOACT045-21	2020	Reunion	RUNA3	RUNA3B	2	Hot	59
RUNA_2449	0227	*castaneus*	OP749995		OP776370 / OP776371	IOACT047-21	2020	Reunion	RUNA3	RUNA3C	2	Hot	30
RUNA_2450	0228	*castaneus*	OP749996		OP776372 / OP776373	IOACT048-21	2020	Reunion	RUNA3	RUNA3C	2	Hot	30
RUNA_2717	0244	*castaneus*		OP787978	OP776374 / OP776375	IOACT090-22	2020	Reunion	RUNA5	RUNA5A	2	Hot	NA
RUNA_2718	0243	*castaneus*		OP787977	OP776376 / OP776377	IOACT089-22	2020	Reunion	RUNA5	RUNA5A	2	Hot	33
RUNA_2719	0239	*castaneus*	OP820449			IOACT084-22	2020	Reunion	RUNA5	RUNA5A	2	Hot	31
RUNA_0854	0220	*castaneus*	OP749988		OP776354 / OP776355	IOACT026-21	2021	Reunion	RUNA8	RUNA8A	2	Hot	67
RUNA_1208	0221	*castaneus*	OP749989		OP776356 / OP776357	IOACT034-21	2021	Reunion	RUNA9	RUNA9A	2	Hot	44
ORCIE1808	0245	*castaneus*	OP749985		OP776344 / OP776345	IOACT091-22	2017	Rodrigues	RODA2	RODA2A	2	Hot	38
ORCIE1889	0246	*stigmaticus*	OP575310		OP776346 / OP776347	IOACT092-22	2017	Rodrigues	RODA2	RODA2B	2	Hot	46
ORCIE1890	0217	*stigmaticus*	OP575311		OP776348 / OP776349	IOACT016-21	2017	Rodrigues	RODA2	RODA2B	2	Hot	34
ORCIE1990	0218	*randalli*	OP749983		OP776350 / OP776351	IOACT021-21	2017	Rodrigues	RODA2	RODA2C	2	Hot	41
1583	0216	*stigmaticus*	OP575309		OP776342 / OP776343	IOACT001-21	2017	Rodrigues	RODA3	RODA3A	2	Hot	47

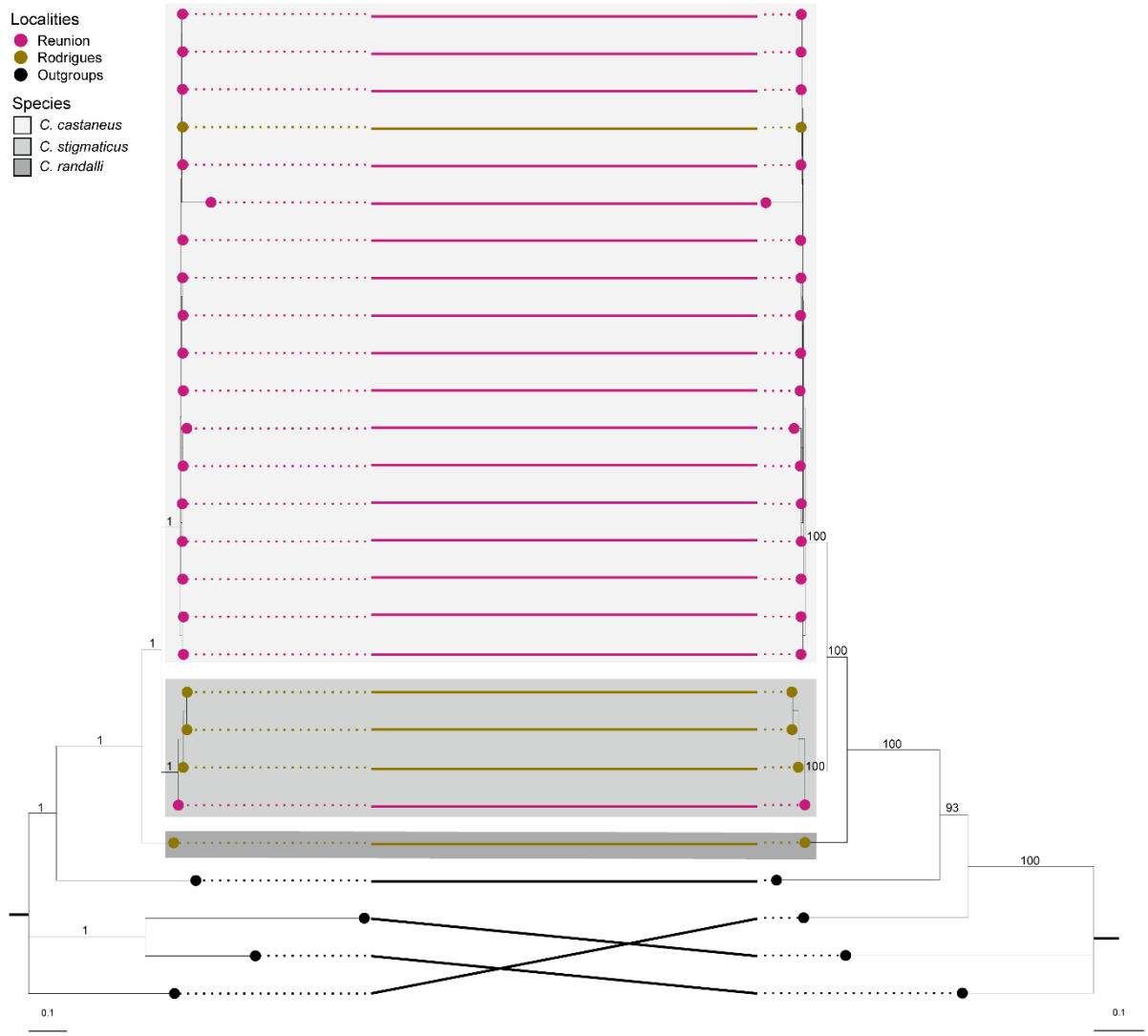
Appendix 3: Outgroups for each dataset.

Species	GenBank Accession Number		
	COI	**Rho**	**Complete mitogenome (mt)**
Ophioblennius macclurei	HQ168577.1	HQ168928 .1	
Omobranchus elegans	KT284893.1		KT284893.1
Omobranchus obliquus		HQ168927.1	
Petroscirtes breviceps	NC_004411.1	KF265117.1	NC_004411.1
Salarias fasciatus	NC_004412.1	HQ168942.1	NC_004412.1
Ecsenius bicolor	NC_028295.1		NC_028295.1

Appendix 4: Proportions of variable and parsimony informative characters and best nucleotide substitution model for each partition of the datasets.

Dataset	Partition	Codon position	#seq	Length (bp)	Variable characters	Parsimony informative characters	Best/fit partitioning scheme
COI	COI	1st	296	168	20 (11.90%)	19 (11.31%)	SYM+G4
		2nd	296	168	0	0	F81+F
		3rd	296	168	154(91.67%)	151 (89.88%)	GTR+I+G4
	COI		292	506	176 (34.78%)	170 (33.60%)	GTR+F+I+G4
mt	12S		24	969	283 (29.21%)	152 (15.69%)	GTR+I+G4
	16S		24	1708	675 (39,52%)	378 (22.13%)	GTR+I+G4
	CDS	1st	24	3802	1377 (36.22%)	845 (22.23%)	GTR+I+G4
		2nd	24	3801	1418 (37.31%)	1020 (26.84%)	GTR+I+G4
		3rd	24	3801	2623 (69,01%)	1877 (49.38%)	GTR+I+G4
Rho	Rho	1st	70	245	18 (7.35%)	6 (2.45%)	TVM+F+G4
		2nd	70	245	8 (3.27%)	4 (1.63%)	F81+F+G4
		3rd	70	245	81 (33.06%)	39 (15.92%)	GTR+F+G4

Appendix 5: Bayesian Inference (left) and Maximum Likelihood (right) trees of concatenated two rRNA (12S and 16S) and coding DNA sequence (CDS) from complete mitogenomes, with the same samples in both trees linked by a line. Intraspecific support values are not shown.



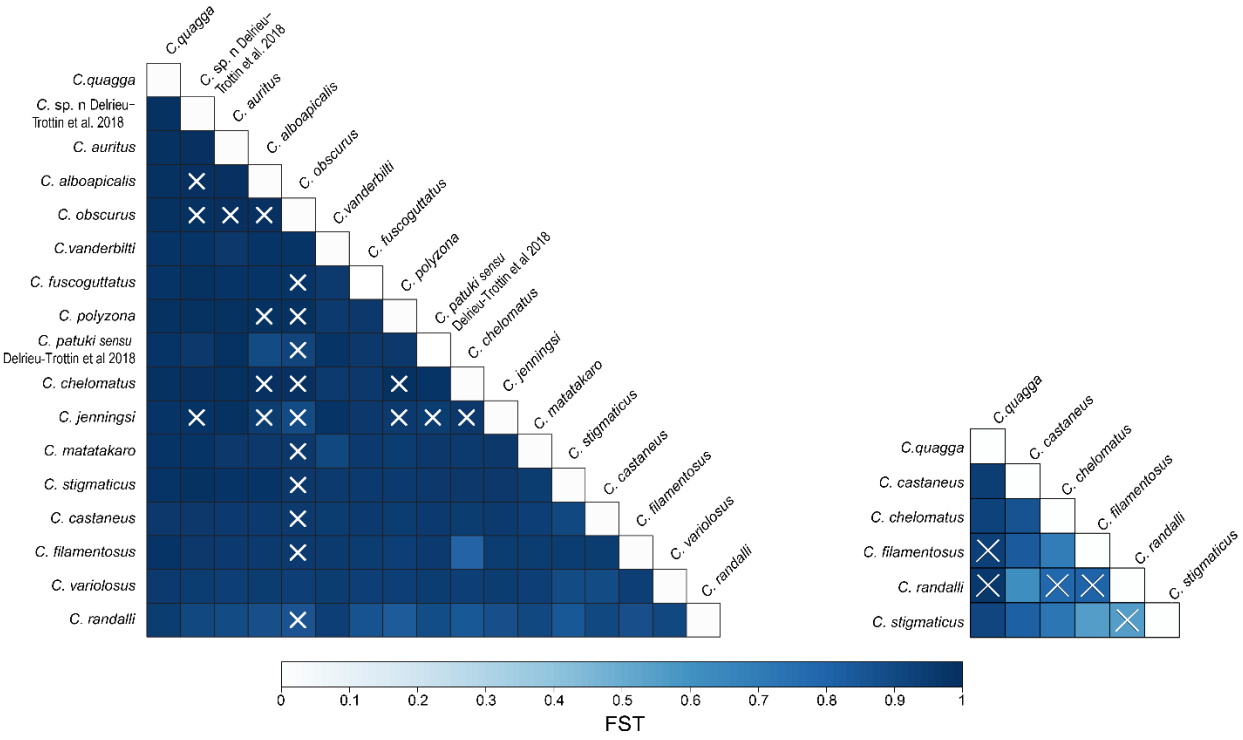
Appendix 6: Population average pairwise difference in bp. Above diagonal: Average number of pairwise differences between species (PiXY); Diagonal elements in grey: Average number of pairwise differences within species (PiX); Below diagonal: Corrected average pairwise difference (PiXY/(PiX+PiY)/2). * for significant p-values.

	C. alboapicalis	*C. auritus*	*C. castaneus*	*C. chelomatus*	*C. filamentosus*	*C. fuscoguttatus*	*C. jenningsi*	*C. matatakaro*	*C. obscurus*	*C. patuki sensu* Delrieu-Trottin et al. 2018	*C. polyzona*	*C. quagga*	*C. randalli*	*C.* sp.n. Delrieu-Trottin	*C. stigmaticus*	*C. vanderbilti*	*C. variolosus*
C. alboapicalis	0	72.86	62.13*	75.33	79.63	75.17*	63.5	69.71*	42	22.2*	69	66.77*	71*	47.8*	71.63*	77.91*	69.55*
C. auritus	72.57*	0.57	51.69*	45.67*	55.38*	58.67*	84.36	49.47*	80.14	82.86*	57.29*	81.08*	61.54*	87.94*	55.66*	52.67*	61.56*
C. castaneus	60.74*	50.01*	2.79	46.23*	50.78*	56.53*	69.07*	43.29*	61.22	67.65*	46.41*	62.43*	48.17*	67.9*	27.2*	46.34*	32.7*
C. chelomatus	75*	45.05*	44.51*	0.67	15.38*	50*	81.33	44.54*	74.33	78.33*	51	69.79*	56.07*	82.53*	49.04*	47.22*	58.18*
C. filamentosus	77.73*	53.2*	47.5*	13.15*	3.79	54.67*	81.75*	45.49*	79.63	80.93*	53.33*	71.07*	59.58*	79.08*	52.81*	45.35*	60.98*
C. fuscoguttatus	74.1*	57.31*	54.07*	48.6*	51.71*	2.13	81.5*	47.33*	72.33	77.33*	52.33*	71.68*	53.6*	80.7*	48.96*	47.71*	62.19*
C. jenningsi	60*	80.57*	64.17*	77.5*	76.36*	76.93*	7	77.89*	68.5	71.7*	75.33	68.55*	76.9*	69.7*	70.63*	81.13*	75.42*
C. matatakaro	68.6*	48.07*	40.79*	43.1*	42.49*	45.16*	73.28*	2.22	69.16	72.01*	42.11*	67.06*	46.73*	75.3*	39.92	21.96*	55.98*
C. obscurus	42*	79.86*	59.82*	74*	77.73*	71.27*	65*	68.05	0	37.4	68*	66.32*	71	39.2	59.5*	80.29*	65.25*
***C. patuki sensu* Delrieu-Trottin et al. 2018**	20.7*	81.07*	64.76*	76.5*	77.53*	74.77*	66.7*	69.4*	35.9*	3	76.07*	65.83*	73.04*	49.8*	75.9*	78.4*	77.63*
C. polyzona	68.33*	56.33*	44.35*	50*	50.77*	50.6*	71.17*	40.33*	67.33*	73.9*	1.33	78.09*	53.1*	70.47*	42.63*	45.52*	57.95*
C. quagga	66.14*	80.16*	60.4*	68.82*	68.54*	69.98*	64.41*	65.31*	65.68*	63.69*	76.79*	1.27	72.49*	68.25*	71.24*	74.13*	68.71*
C. randalli	70.12*	60.38*	45.89*	54.86*	56.8*	51.66*	72.52*	44.75*	70.12*	70.66*	51.56*	70.98*	1.76	78.9*	45.93*	53.28*	59.2*
***C.* sp.n. Delrieu-Trottin et al. 2018**	47.6*	87.46*	66.3*	82*	76.98*	79.43*	66*	73.99*	39*	48.1*	69.6*	67.41*	77.82*	0.4	72.58*	74.49*	75.97*
C. stigmaticus	70.66*	54.41*	24.84*	47.74*	49.96*	46.93*	66.16*	37.84*	58.54*	73.44*	40.99*	69.64*	44.08*	71.41*	1.93	42.95*	35.69*
C. vanderbilti	76.87*	51.35*	43.91*	45.85*	42.41*	45.6*	76.59*	19.81*	79.25*	75.86*	43.81*	72.46*	51.36*	73.25*	40.95*	2.05	58.83*
C. variolosus	67.44*	59.16*	29.2*	55.73*	56.98*	59.02*	69.81*	52.76*	63.14*	74.02*	55.17*	65.96*	56.21*	73.66*	32.61*	55.68*	4.22

Appendix 7: AMOVA results for among species and within species. s.s.: sum of squares, v.c.: variance components, % var: % of variation, FST: fixation index.

AMOVA results for COI					
Source of variation	**s.s**	**v.c.**	**% var**	**FST**	**p-value**
Among species	6323.217	25.07971	94.59	0.94589	0.000
Within species	400.317	1.43483	5.41		
AMOVA results for rhodopsin					
Source of variation	**s.s**	**v.c.**	**% var**	**FST**	**p-value**
Among species	93.47	3.02	85.81	0.858	0.000
Within species	31.90	0.50	14.19		

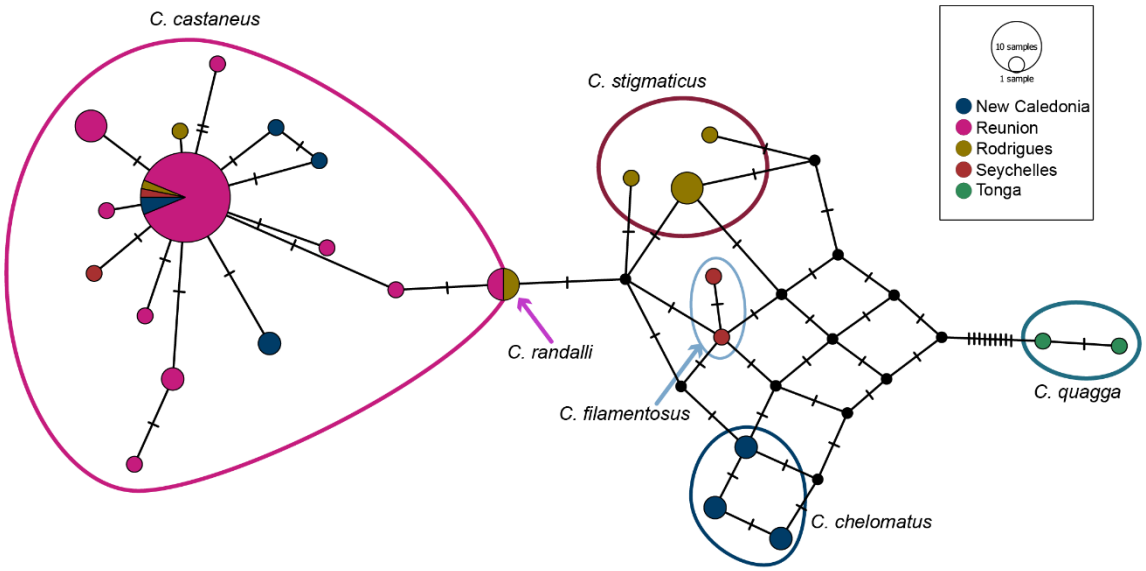
Appendix 8: FST matrix between *Cirripectes* species for COI dataset (left) and rhodopsin dataset (right). Blue scale represents FST values (darkest blues for highest FST). Crosses represent single non-significant p-value of the pairwise FST (p-value >0.05).



Appendix 9: Pairwise FST values for the mtDNA COI of the 17 *Cirripectes* species and associated p-values (below diagonal).

	C. alboapicalis	*C. auritus*	*C. castaneus*	*C. chelomatus*	*C. filamentosus*	*C. fuscoguttatus*	*C. jenningsi*	*C. matatakaro*	*C. obscurus*	*C. patuki sensu* Delrieu-Trottin et al. 2018	*C. polyzona*	*C. quagga*	*C. randalli*	*C.* sp.n. Delrieu-Trottin	*C. stigmaticus*	*C. vanderbilti*	*C. variolosus*
C. alboapicalis		0.99329	0.95542	0.9941	0.95872	0.97644	0.94488	0.96944	1	0.89362	0.98713	0.98188	0.87097	0.99331	0.97654	0.97387	0.94125
C. auritus	0.01802		0.95028	0.98696	0.95849	0.97814	0.98213	0.96108	0.99287	0.98134	0.98667	0.98625	0.89276	0.99428	0.97662	0.96309	0.93859
C. castaneus	0.00000	0.00000		0.94098	0.94122	0.95088	0.95692	0.93973	0.95334	0.95759	0.94039	0.96196	0.90256	0.96106	0.90066	0.94851	0.89152
C. chelomatus	0.11712	0.01802	0.00000		0.80526	0.96583	0.9657	0.95286	0.99103	0.97171	0.98039	0.98256	0.8488	0.99407	0.96653	0.95721	0.93091
C. filamentosus	0.02703	0.00000	0.00000	0.00901		0.94345	0.94844	0.94423	0.95246	0.95679	0.93969	0.97305	0.87216	0.96783	0.9459	0.95033	0.93176
C. fuscoguttatus	0.02703	0.00000	0.00000	0.01802	0.00000		0.96357	0.95343	0.97051	0.96742	0.96367	0.97988	0.86296	0.98312	0.95885	0.95639	0.93571
C. jenningsi	0.29730	0.02703	0.00901	0.05405	0.01802	0.01802		0.96912	0.89781	0.94664	0.95705	0.97722	0.87057	0.97504	0.96328	0.97322	0.94287
C. matatakaro	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00901		0.96794	0.96783	0.94896	0.97277	0.90559	0.97358	0.94592	0.90351	0.93768
C. obscurus	0.99099	0.99099	0.99099	0.99099	0.99099	0.99099	0.99099	0.99099		0.91979	0.98039	0.98081	0.85403	0.9898	0.96759	0.97415	0.93532
***C. patuki sensu* Delrieu-Trottin et al. 2018**	0.04505	0.00000	0.00000	0.03604	0.00000	0.00000	0.05405	0.00000	0.10811		0.96793	0.97631	0.88916	0.96586	0.96942	0.97281	0.947
C. polyzona	0.12613	0.00000	0.00000	0.15315	0.00000	0.01802	0.15315	0.00000	0.30631	0.02703		0.98363	0.83894	0.9899	0.95795	0.95505	0.93
C. quagga	0.00901	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.03604	0.00000	0.00000		0.94077	0.98345	0.97979	0.97444	0.95102
C. randalli	0.00901	0.00000	0.00000	0.00000	0.00000	0.00901	0.00000	0.00000	0.08108	0.00000	0.00901	0.00000		0.90649	0.84802	0.93917	0.90611
***C.* sp. n. Delrieu-Trottin et al. 2018**	0.06306	0.00901	0.00000	0.02703	0.00000	0.00901	0.06306	0.00000	0.09910	0.01802	0.00901	0.00000	0.00000		0.98112	0.97348	0.94878
C. stigmaticus	0.02703	0.00000	0.00000	0.00000	0.00000	0.00000	0.02703	0.00000	0.09009	0.00000	0.01802	0.00000	0.00000	0.00000		0.95202	0.8914
C. vanderbilti	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000		0.9493
C. variolosus	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.01802	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	

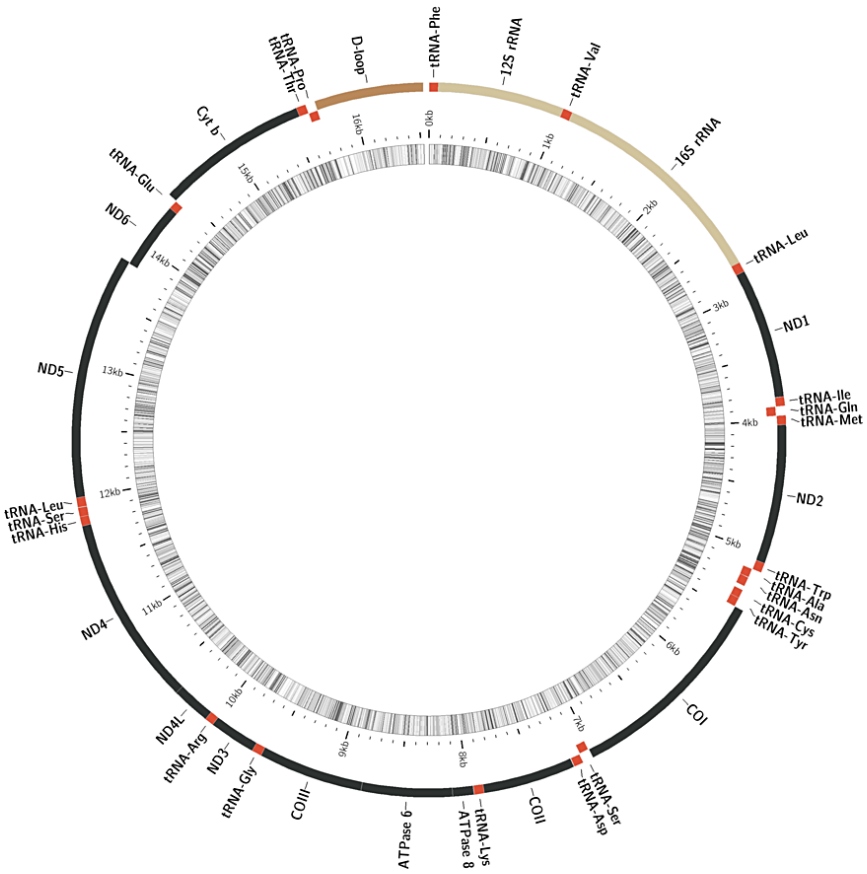
Appendix 10: Median/joining networks showing the relationships among nuclear rhodopsin haplotypes in *Cirripectes* species. Each circle represents a haplotype with its size proportional to its total frequency. Black crossbars on branches indicate single nucleotide changes, black nodes represent unsampled probable haplotypes, colours indicate collection location.



Appendix 11: Detailed description of *Cirripectes* mitogenome.

The mitogenome lengths are within the ranges of other teleost mitogenomes. As in other vertebrates (Miya et al. 2001), they contained 13 protein-coding genes, 2 rRNA genes (12S and 16S), 22 tRNA, and a control region (Appendix 11.1). Among the tRNAs, two forms of tRNA-Leu and tRNA-Ser were identified as in other bony fishes (Prosdocimi et al., 2012). Likewise, most mitochondrial genes were encoded on the H-strand, with only ND6 and eight tRNA (Gln, Ala, Asn, Cys, Tyr, Ser [only one of the two tRNA-Ser], Glu, and Pro) genes encoded on the L-strand (Wang et al., 2021). The ATPase 6 and ATPase 8 overlapped by 10 nucleotides, and ND4 and ND4L shared 7 nucleotides. ND5 and ND6 overlapped by 9 nucleotides (but only 4 nucleotides for RUNA_0737) on the opposite strand. The 12S ribosomal genes are 948 bp long for the three species (Appendix 11.2). The 16S ribosomal RNA genes showed some individual length variation among *C. castaneus* with 1,684 bp in 18 individuals, but 1,685 bp in one (RUNA_4002; T insertion at 1,343 bp). For *C. randalli* and *C. stigmaticus*, the 16S ribosomal sequences are 1,681 bp and 1,682 bp long, respectively. They were located between tRNA-Phe and tRNA-Leu, and were separated by tRNA-Val, as they were in other vertebrates (Miya et al., 2003) The 22 tRNA genes were interspersed in the genome and ranged in size from 65 to 75 bp.

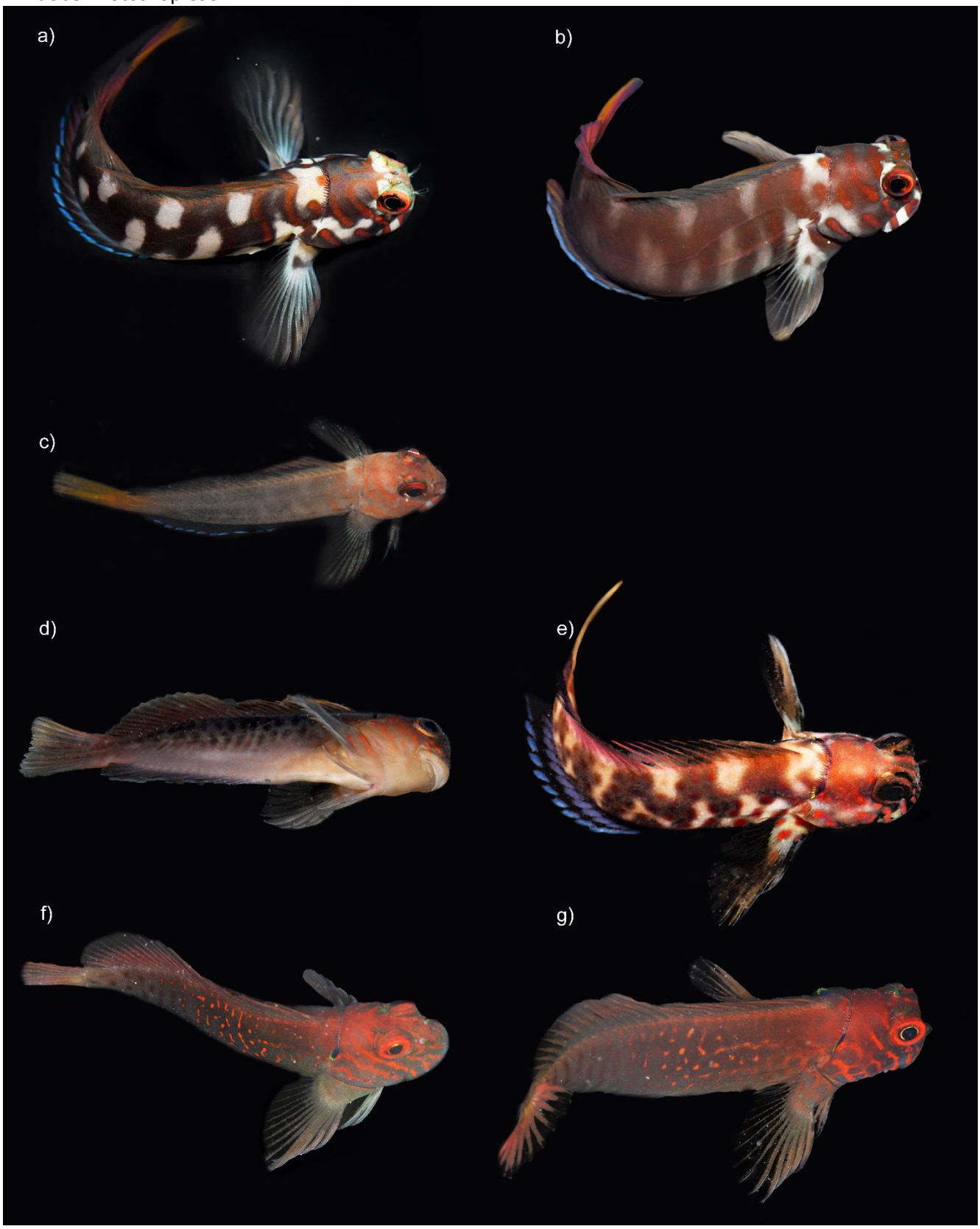
Appendix 11.1: Visualization and annotation using Mitofish (Sato et al., 2018) of the mitochondrial genome of *Cirripectes stigmaticus*. Black = protein coding sequences, Red = tRNAs, Tan = rRNAs, Brown = control region.



Appendix 11.2: Length of mitochondrial genes found in *C. castaneus*, *C. stigmaticus* and *C. randalli*. In grey, genes with intra- and inter-species length variability.

Gene in position order	Type	Strand	Length in bp for each species		
			***C. castaneus* (N=19)**	***C. stigmaticus* (N=4)**	***C. randalli* (N=1)**
tRNA/Phe	transfer RNA	H		69	
12S	ribosomal RNA	H		948	
tRNA/Val	transfer RNA	H		72	
16S	ribosomal RNA	H	1,684 (18) ; 1,685 (1)	1,682	1,681
tRNA/Leu	transfer RNA	H		75	
ND1	protein coding	H		975	
tRNA/Ile	transfer RNA	H		69	
tRNA/Gln	transfer RNA	L		71	
tRNA/Met	transfer RNA	H		70	
ND2	protein coding	H		1,042	
tRNA/Trp	transfer RNA	H		73	
tRNA/Ala	transfer RNA	L		69	
tRNA/Asn	transfer RNA	L		73	
tRNA/Cys	transfer RNA	L		65	
tRNA/Tyr	transfer RNA	L		70	
COI	protein coding	H		1,572	
tRNA/Ser	transfer RNA	L		71	
tRNA/Asp	transfer RNA	H		73	
COII	protein coding	H		691	
tRNA/Lys	transfer RNA	H		74	
ATPase8	protein coding	H		168	
ATPase6	protein coding	H		683	
COIII	protein coding	H		785	
tRNA/Gly	transfer RNA	H	72	73	73
ND3	protein coding	H		349	
tRNA/Arg	transfer RNA	H		68	
ND4L	protein coding	H		297	
ND4	protein coding	H		1,381	
tRNA/His	transfer RNA	H		69	
tRNA/Ser	transfer RNA	H		68	
tRNA/Leu	transfer RNA	H		73	
ND5	protein coding	H	1,827 (1); 1,842 (14);1,863(4)	1,842	1,842
ND6	protein coding	L		522	
tRNA/Glu	transfer RNA	L		68	
Cyt b	protein coding	H		1,141	
tRNA/Thr	transfer RNA	H		72	
tRNA/Pro	transfer RNA	L		70	
Dloop		H	820/825	818/820	818
Total Length			16,476 / 16,512	16,482 / 16,532	16,482

Appendix 12: Pictures of *Cirripectes* species sampled in the Mascarene Archipelago using ARMS: *C. castaneus*: a: RUNA_3256 (MNHN-IC-2023-0234), b: RUNA_0854 (MNHN-IC-2023-0220) and c: RUNA_2719 (MNHN-IC-2023-0239); *C. randalli*: d: ORCIE1990 (MNHN-IC-2023-0218) and e: RUNA_3176 (MNHN-IC-2023-0232); and *C. stigmaticus*: f: 1583 (MNHN-IC-2023-0216) and g: ORCIE1889 (MNHN-IC-2023-0246). Pictures were edited in Adobe Photoshop CS6.



Chapitre 6 : Synthèse générale et perspectives

En tant que haut lieu de biodiversité, Madagascar et les îles du Sud-Ouest de l'océan Indien font partie des zones géographiques fortement menacées par la perte d'habitats et les changements globaux. Au sein de cette écorégion, plusieurs études ont montré que les récifs coralliens des Mascareignes sont faiblement connectés avec les plus proches, ceux de Madagascar. De ce fait, la faible connectivité observée pour les macro-organismes tels que les coraux **(Oury 2022)**, les hydrozoaires **(Postaire et al. 2017)**, les échinodermes **(Hoareau et al. 2013)** et les poissons **(Muths et al. 2015)** présage une spécificité du cryptobiome des Mascareignes.

Dans le contexte des changements globaux, la surveillance de la biodiversité est essentielle pour détecter les facteurs de perturbation et comprendre les réponses des communautés. La quantification de la biodiversité est cruciale pour une gestion efficace des écosystèmes. Le métabarcoding apparaît comme une méthode prometteuse pour évaluer la diversité des récifs coralliens **(Plaisance et al. 2011 ; Leray & Knowlton 2015 ; Ransome et al. 2017)**. L'emploi des ARMS comme collecteurs passifs de la diversité du cryptobiome benthique permet dans standardiser les volumes et surfaces échantillonnés afin minimiser les biais d'échantillonnage et encourager leur déploiement à l'échelle plantaire **(Cahill et al. 2018 ; Pearman et al. 2020)**. Cependant à l'heure actuelle, la comparaison des communautés échantillonnées par les ARMS reste limitée par une grande variabilité (1) des temps et saison de déploiement de ces mini-récifs lors l'échantillonnage et (2) des traitements bio-informatiques des séquences. Pourtant, jusqu'à présent, aucune étude ne s'était penchée sur l'influence des facteurs temporels liés au déploiement de ces unités d'échantillonnage : la durée d'immersion et les variations saisonnières.

À ce titre, ces travaux de thèse avaient pour objectif d'évaluer la diversité et les patrons de répartition du cryptobiome des Mascareignes en utilisant d'une part des unités d'échantillonnage standardisées, les ARMS, et d'autre part une approche moléculaire par métabarcoding, faisant de cette étude la première à utiliser cette méthodologie pour étudier le cryptobiome dans l'océan Indien.

Les résultats, synthétisés et discutés ci-dessous, permettent une meilleure compréhension de la diversité du cryptobiome des Mascareignes et des dimensions spatiales et temporelles de la structure des communautés du cryptobiome récifal. Ces travaux permettront à terme d'améliorer les performances des ARMS en tant que dispositif de suivi de la biodiversité en proposant (1) une

standardisation temporelle du déploiement des ARMS, (2) un pipeline bio-informatique adapté au cryptobiome récifal et (3) un référentiel moléculaire adapté au Sud-Ouest de l'océan Indien.

1. Le cryptobiome des Mascareignes

1.1. La composition et les variations du cryptobiome récifal

Les ARMS relevés au cours de cette thèse ont permis de collecter 2 989 organismes dans 39 ARMS, supplémentaires à ceux déjà échantillonnés, portant le nombre de spécimens collectés à 4 584 pour un total de 54 ARMS récoltés dans les Mascareignes. L'approche métabarcoding a fourni 6 203 OTU pour le 18S et 7 701 OTU pour le COI. Les organismes mobiles de taille supérieure à 2 mm comprennent 17 phylums et les analyses de métabarcoding ont identifié 29 phylums. Les principaux taxons retrouvés correspondent aux taxons du cryptobiome observés dans d'autres régions du monde (ex. en Mer Rouge ; **Carvalho et al. 2019**), avec une majorité d'arthropodes, d'annélides, de mollusques, et de porifères, tandis que les échinodermes, les chordés et les bryozoaires ont également été largement retrouvés. Ces résultats sont également cohérents avec les précédentes évaluations de la diversité du cryptobiome récifal par différentes méthodes **(Enochs 2012 ; Enochs & Manzello 2012)** et concorde avec la grande diversité taxonomique observée dans les récifs coralliens.

1.2. Une majorité qui reste à référencer

Malgré les efforts déployés pour créer une base de référence adaptée au Sud-Ouest de l'océan Indien (Chapitre 3) et la combinaison de différentes méthodes et seuils d'assignement (Chapitre 2), seule une petite fraction des OTU produits par métabarcoding a pu être identifiée au Domaine (18S : 42.9 % ; COI : 9,76 % ; Chapitre 3). Ces proportions sont inférieures à celles retrouvées dans les autres études du cryptobiome et peuvent s'expliquer par au moins les deux facteurs suivants qui interviennent conjointement :

(1) les différences dans les pipelines d'assignement employés. En effet, les pourcentages d'assignement des études sur le cryptobiome viennent d'études employant des paramètres bio-informatiques extrêmement variables avec : différents seuils de regroupement (*clustering*) des ASV en OTU (ex. sans (ASV ; **Villalobos et al. 2022**), similarité à 97% **(Pearman et al. 2016 ; Nichols et al. 2021 ; Casey et al. 2021 ; Ip et al. 2022)** ou à partir d'un algorithme de prédiction bayésien **(Leray & Knowlton 2015 ; Al-Rshaidat et al. 2016 ; Pearman et al. 2018)**) et différents méthode et seuil d'assignement (par exemple pour le COI : blast : 97 % **(Leray & Knowlton 2015 ; Al-Rshaidat et al. 2016 ; Ransome et al. 2017 ; Pearman et al. 2018 ; Nichols et al. 2021)**, 85% **(Ransome et al. 2017 ; Casey et al. 2021)** ; RDP : 80% **(Ip et al. 2022)** ; LCA 85% **(Nichols et al. 2021)**).

(2) les séquences 18S et COI des OTU retrouvés dans la région du Sud-Ouest de l'océan Indien sont éloignées génétiquement des espèces référencées dans les bases de données publiques. En effet, le faible taux d'OTU assignés reflète généralement le manque de séquences de référence dans les bases de données **(Gaither et al. 2022)** et en particulier pour certaines régions géographiques sous-étudiées **(Monchamp et al. 2023)**.

Par exemple, si l'on prend les cartes de répartition des sites d'échantillonnage des séquences déposés dans BOLD, on observe très peu de points d'échantillonnages dans les Mascareignes (Figure 6.1). Certaines études ont bénéficié de référentiels locaux faisant suite à un échantillonnage intensif du cryptobiome **(Ransome et al. 2017 ; Nichols et al. 2021)**.Plusieurs études mettent en avant une différenciation génétique des espèces présentes dans les Mascareignes. En effet, notre analyse phylogénétique du poisson cryptobenthique du genre *Cirripectes* (Chapitre 5), confirme qu'une des trois espèces retrouvées à La Réunion est endémique à l'archipel des Mascareignes et diverge de 8.71 % de l'espèce la plus proche. Une faible connectivité des organismes des Mascareignes a

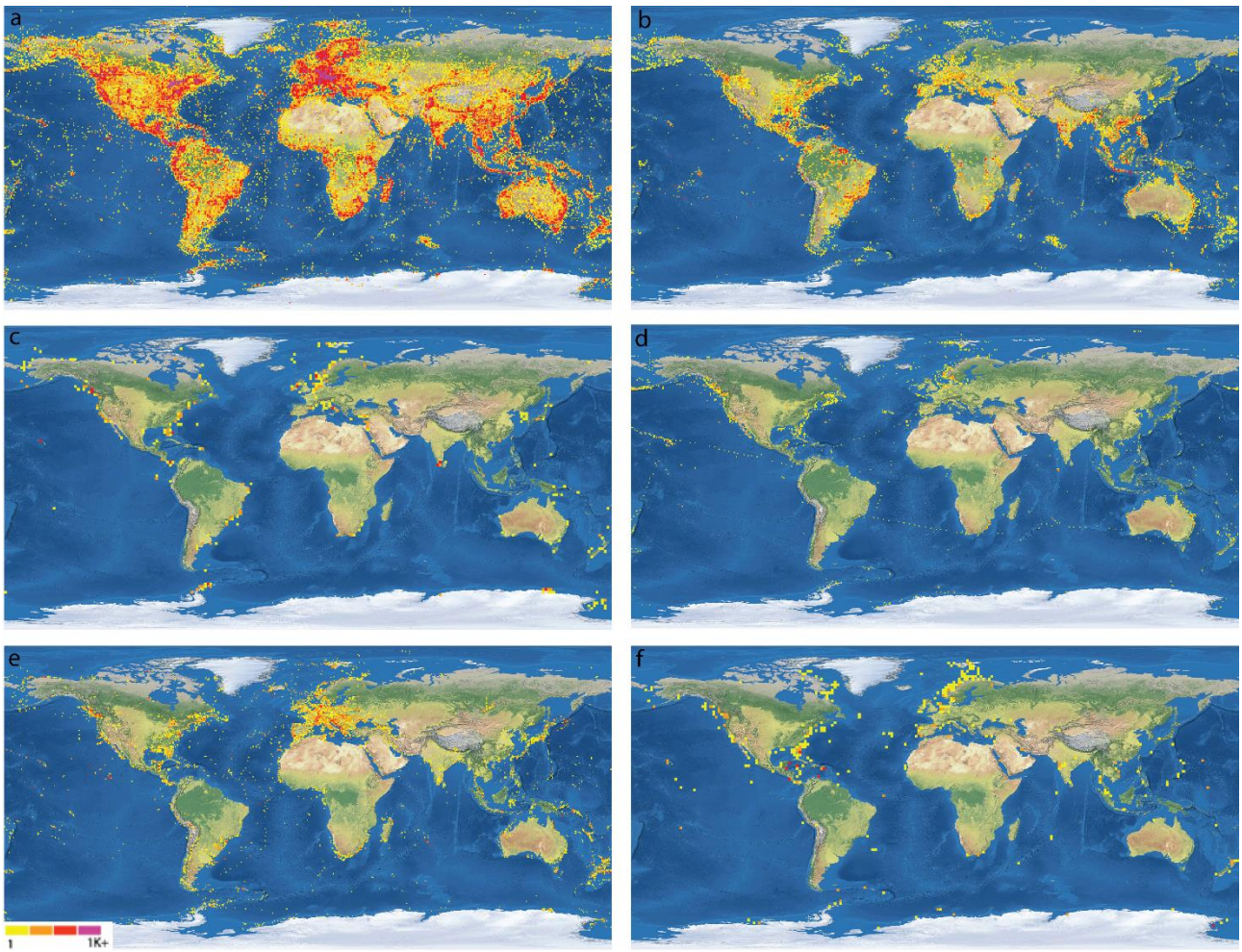


Figure 6.1 : Cartes de répartition des sites d'échantillonnage des séquences déposés dans BOLD pour l'ensemble des taxons (a), les Actinopterygii (b), les Ascidiacea (c), les Cnidaria (d), les Malacostraca (e) et les Porifera (f).

également été observée pour plusieurs taxons **(Hoareau et al. 2013 ; Oury 2022)**, avec notamment les hydraires où l'on observe un groupe génétique par île échantillonnée **(Postaire et al. 2017)**.

Cependant, au sein d'un même genre, les espèces peuvent avoir des patrons de distribution spatiale très différents. En effet, les deux autres espèces du genre *Cirripectes* échantillonnées dans les Mascareignes ont une distribution Indo-Pacifique avec des haplotypes partagées entre les océans **(Couëdel et al. 2023)**. Ces différences patrons de distribution au sein d'un même genre ont également été observées pour d'autres taxons du cryptobiome tels que les éponges **(Pasnin et al. 2020)**. Les espèces ayant des haplotypes partagés entre les océans Indien et Pacifique sont minoritaires (**Crandall et al. 2019**) avec par exemple seulement 13.5% des espèces de téléostéens récifaux à large répartition dans **Hubert et al. (2012)** et 21.5% dans **Hubert et al. (2017)**.

L'étude des différents patrons des répartitions biogéographiques des organismes récifaux permet de mieux comprendre les mécanismes à l'origine de ce phénomène (*cf*. **Gaither & Rocha 2013** pour les différentes hypothèses) **(Keyse et al. 2014)**. Dans ce contexte, l'emploi des ARMS permettrait une approche multi-taxons et standardisée, afin de définir les assemblages d'espèces et les facteurs climatiques et environnementaux qui servent de base à l'identification des modèles de biodiversité et à la définition des régions de conservation **(Bickford et al. 2007 ; Crandall et al. 2019)**. De plus, l'acquisition de données moléculaires, par exemple *via* le barcoding et le métabarcoding, couplée à la collecte de métadonnées géo-référencées complètes (ex. date de collecte, saison, profondeur, photographies, etc.), permet une réutilisation de ces données dans des contextes futurs variés. Par exemple, l'utilisation de marqueurs de métabarcoding variables, tels que le COI, fournit également une énorme quantité d'informations intra-spécifiques (intra-OTU) qui restent peu exploitées à ce jour **(Turon et al. 2020)**. Ces données COI de métabarcoding peuvent être utilisées pour étudier les variations intra-spécifiques et les caractéristiques phylogéographiques de centaines d'espèces simultanément (métaphylogéographie ; *cf.* **Turon et al. 2020**). Cependant, les comparaisons entre les régions et les études telles que dans Pearman et collaborateurs **(2020)** sont freinées par le manque de standardisations des approches moléculaires et bio-informatiques (*cf.* Chapitre 2), ainsi que l'absence de mise à disposition des jeux de données (OTU et données brutes pour réanalyse) et métadonnées associées. Toutefois, de plus en plus d'études s'inscrivent dans le principe de « science ouverte » qui a pour concept l'ouverture, la transparence, et la reproductibilité des connaissances **(Vicente-Saez & Martinez-Fuentes 2018)**. Ce principe implique la mise à disposition les pipelines et les données employées lors de la publication d'articles scientifiques, ce qui facilite la reproductibilité et la comparaison des études.

2. La structure temporelle du cryptobiome et implications pour l'évaluation de la diversité à l'aide des ARMS

Les résultats décrits dans le chapitre 4 sur le déploiement de 15 ARMS sur la pente externe d'un récif corallien de l'île de la Réunion nous ont permis d'étudier les schémas de colonisation et le remplacement temporel des communautés du cryptobiome entre trois durées d'immersion (six mois, un an et deux ans) et deux saisons de déploiement (fraîche et chaude). Les deux marqueurs moléculaires employés lors de cette étude, le 18S et le COI de résolutions taxonomiques différentes, montrent des résultats similaires en termes de structure et de composition des communautés. Alors qu'aucune variation significative du nombre d'OTU collectés par les ARMS n'a été observée entre les différents temps d'immersion et les deux saisons, la structure et la composition des communautés ont quant à elles été significativement affectées par ces deux facteurs.

En effet, ces travaux montrent que la composition des communautés du cryptobiome est significativement différente en fonction de la durée d'immersion des ARMS, en particulier pour les taxons sessiles. En effet, après 6 mois d'immersion, la communauté échantillonnée présente des diversités plus élevées en cnidaire, cirripèdes, et ascidies solitaires reflétant la capacité de ces taxons à être des colonisateurs pionniers. Après un an d'immersion, on observe une diminution de la diversité de ces premiers colonisateurs en faveur des rhodophytes, des porifères et des ascidies coloniales. Dans la fraction des organismes mobiles, on observe une forte diversité des annélides dans la communauté échantillonnée après 6 mois d'immersion, puis elle diminue avec la durée d'immersion. Cette diminution est en partie compensée par une augmentation de la diversité des arthropodes qui atteint son maximum après un an d'immersion. Les résultats observés dans les communautés mobiles et sessiles, combinés avec ceux de la diminution de la similarité entre les réplicats avec l'augmentation de la durée d'immersion, reflètent le remplacement d'un pool restreint d'espèces pionnières (cnidaires, annélides, cirripèdes et ascidies solitaires) par un ensemble plus diversifié et aléatoire d'espèces de la communauté mature (rhodophytes, porifères, arthropodes mobiles et ascidies coloniales). Ainsi, la colonisation des espèces de la communauté mature est initiée par l'établissement d'une communauté pionnière.

De surcroît, les variations saisonnières influencent également les communautés récoltées et plus fortement les communautés sessiles que les communautés mobiles. À La Réunion, la saison fraîche montre des concentrations plus élevées de chlorophylle et de matière organique particulaire. Les ARMS déployés pendant cette saison présentent (à la récolte) des diversités plus élevées des taxons suspensivores tels que les porifères les cnidaires et les bryozoaires. Inversement,

les rhodophytes présentaient une diversité plus importante pendant la saison chaude, saison ayant les plus fortes valeurs d'irradiation quotidienne, ainsi que de diversité des mollusques qui s'en nourrissent **(Morton & Blackmore 2009 ; Larkin et al. 2017)**. Par ailleurs, seules les ascidies semblent être influencées plus fortement par la saison de récolte que la saison de pose. Elles ont montré une diversité plus importance dans les ARMS récoltés en saison fraîche. Les ascidies ont des patrons de recrutement spécifiques à chaque espèce, avec des variations selon les saisons, l'orientation et la position sur le substrat **(Shenkar et al. 2008)**. Néanmoins, l'augmentation de leur diversité concorde avec le pic d'enrichissement en chlorophylle, **POC** (Carbone Organique Particulaire) et $\delta^{13}C$ des eaux récifales **(Kolasinski et al. 2011)** et des conditions d'hydrodynamique plus fortes **(Naim 1993 ; Chazottes et al. 2008)**, qui favorisent la dispersion des larves et aussi l'advection de la nourriture (**POM** [Matière Organique Particulaire], etc.) pour les suspensivores. Dans l'ensemble, les variations saisonnières des communautés récifales comprennent des interactions complexes de facteurs environnementaux tels que la température de l'eau de mer, l'irradiation, les précipitations et la disponibilité des nutriments. Des études ciblées à chaque taxon sont nécessaires pour mieux comprendre comment ils réagissent aux variations saisonnières des différents paramètres environnementaux.

Nos travaux sont les premiers à évaluer l'effet de la saison de déploiement et de récolte des ARMS. Il est fort probable ces variations se reflètent également dans les résultats obtenus par d'autres méthodes d'échantillonnages, telles que l'analyse de plaques de recrutement artificielles (*Artificial Sampling Units* (**ASU**) ; **Cahill et al. 2018 ; Monroy-Velázquez et al. 2020**) ou le suivi des communautés d'épifaune des coraux branchus **(Pisapia et al. 2020)**. L'étude du recouvrement des organismes sessiles des plaques de ces mêmes ARMS est réalisée dans le cadre de l'étude doctorale de Baptiste Frattini et montre des résultats cohérents avec ces observations (*com. pers.* B. Frattini). Ainsi, à la lumière de ces résultats, le temps d'immersion, et non seulement la saison de déploiement mais aussi de récolte, doivent être pris en compte lors de la conception du plan d'échantillonnage et de l'analyse des résultats pour permettre la comparaison des études. Il est essentiel de déployer et de récupérer les ARMS, et autres structures, pendant la même période de l'année/saison. La durée d'immersion pourrait être réduite à un an au lieu des deux préconisés, du moins dans les milieux chauds, car elle ne semble pas avoir d'effet sur le nombre d'espèces récoltées. Cela permettrait d'obtenir des informations plus fréquentes et de détecter des changements plus rapides au sein des communautés. Toutefois, cette étude reste pionnière dans le domaine et les résultats peuvent différer des schémas à plus long terme ou d'autres récifs. De ce fait, des études supplémentaires sont nécessaires afin de différencier les effets temporels des effets aléatoires, et

ainsi mettre en évidence les effets liés aux perturbations environnementales et aux changements sur le long terme.

Les communautés marines sont généralement dominées par quelques taxons, la majeure partie de la diversité étant représentée par des espèces rares dont la présence peut varier dans l'espace et le temps **(Logares et al. 2014 ; Lindh et al. 2017)**. Nos résultats sont en accord avec cette observation, puisque seule une petite fraction de la communauté (environ 10%) semble être stable dans le temps, comme c'était également le cas en Mer Rouge **(Carvalho et al. 2019 ; Villalobos et al. 2022)**. Pour mieux comprendre le rôle de ces organismes rares dans les écosystèmes et leur stabilité, il est nécessaire de déterminer s'ils sont réellement rares, périodiquement rares (leur abondance fluctue périodiquement) ou que l'effort d'échantillonnage était insuffisant **(Bouchet et al. 2002)**. Plusieurs taxons marins, tels que les algues rouges **(Ateweberhan et al. 2006)**, les coraux **(Guest et al. 2012 ; Massé 2014)**, les mollusques **(Aranda et al. 2003 ; Palant & Fishelson 2013 ; Chetoui et al. 2019)**, les échinodermes **(Muthiga & Jaccarini 2005)** et les poissons **(Takemura et al. 2004 ; Abesamis & Russ 2010)** montrent un cycle de reproduction annuel, synchroniser avec les variations saisonnières. Cela peut influencer les communautés échantillonnées à une certaine saison, avec par exemple, la présence d'œufs ou de larves benthiques d'organismes dont la phase adulte est démersale et ne serait donc pas détectée dans les ARMS **(Benkendorff & Davis 2004)**. Par ailleurs, pour les ARMS, l'écartement d'environ 1 cm entre les plaques peut conditionner la taille maximale ou la phase de développement des organismes présents. Ainsi, par exemple, seuls les juvéniles de certaines espèces peuvent être collectés. C'est le cas d'un juvénile de *Gymnothorax margaritophorus* récolté dans un des ARMS déployé à La Réunion, cette espèce pouvant atteindre une taille totale adulte de 70 cm (**Myers (1991)** dans Fishbase **(2023)**). Dans le cas de *C. castaneus*, les spécimens retrouvés dans les ARMS étaient en moyenne plus petits que ceux trouvés sur les platiers récifaux peu profonds (37 mm contre 71 mm ; *com.pers.* H. Brugemmann). C'est pourquoi, les suivis biologiques réalisés à un moment donné sont généralement inadéquats pour décrire la biodiversité totale ou pour étudier les changements de la diversité dans le temps **(Logares et al. 2014 ; Lindh et al. 2017)**. Les réplicats temporels permettent d'échantillonner une plus grande biodiversité totale (gamma) et permettent de mesurer les changements temporels de la diversité (bêta). Ces résultats soulignent à la fois l'importance de la collecte d'échantillons temporels avec les métadonnées adéquates (**Jarman et al. 2018**), et l'importance du métabarcoding multi-marqueurs pour la surveillance à long terme des écosystèmes marins.

3. Les limites de l'approche métabarcoding et les perspectives d'amélioration

3.1. Les bases de référence

Ces dernières années, le métabarcoding s'est montré être un outil fondamental pour évaluer et suivre la diversité des communautés biologiques **(Plaisance et al. 2011 ; Leray & Knowlton 2016 ; Taberlet et al. 2018)**. Toutefois, son efficacité est dépendante de la qualité et de la complétude des bases de données de séquence de référence **(Gold et al. 2021)**. La grande diversité des taxons présents dans les récifs coralliens complexifie l'exhaustivité de ces bases de référence, et leur incomplétude entraîne une proportion élevée de taxons non assignés et des assignements erronés **(Ransome et al. 2017 ; Gaither et al. 2022 ; Mugnai et al. 2023)**. Ainsi, malgré l'ajout de séquences locales, seule une petite fraction du cryptobiome a pu être identifiée dans notre étude. Ces résultats soulignent la nécessité de poursuivre les efforts pour compléter les bases de référence. De ce fait, l'incomplétude des bases de référence reste une des principales limites lors de l'assignement **(Wang et al. 2018 ; Keck et al. 2023)**. Néanmoins, cette limite n'est pas définitive et apparaît plus comme un contretemps. En effet, la pérennité des données sous forme d'OTU rend possible l'exécution de nouveaux assignements dans le futur, lorsque les bases de référence auront gagné en exhaustivité. Ainsi, il est d'autant plus important de collecter des métadonnées adéquates et de mettre à disposition les données (brutes en *reads* et traitées en OTU), par exemple en les publiant sous forme de *data paper* pour informer la communauté scientifique de l'existence d'un tel jeu de données **(Chavan & Penev 2011 ; Jarman et al. 2018 ; Shea et al. 2023)**.

L'inclusion de séquences locales de qualité aux bases de référence en amont des études de métabarcoding améliore la quantité et la qualité des assignements **(Mugnai et al. 2023)**. Ainsi, cette thèse contribue à documenter le cryptobiome des Mascareignes à l'aide : (1) des photographies et des métadonnées complètes rattachées aux spécimens, (2) de spécimens déposés en collection, dont notamment les 39 spécimens de *Cirripectes* dans les collections de poissons du MNHN, et (3) de trois marqueurs moléculaires (18S, COI et 16S). La production de référentiels barcode contenant un maximum de métadonnées **(Hebert et al. 2003 ; Ratnasingham & Hebert 2007 ; Rimet et al. 2021)** permet de créer des jeux de données fiables et facilite la vérification des séquences **(Ratnasingham & Hebert 2007 ; Chakrabarty et al. 2013 ; Joly et al. 2014 ; Rulik et al. 2017 ; Rimet et al. 2021 ; Berba & Matias 2022)**. En effet, comme relaté dans le chapitre 5, certaines séquences du genre *Cirripectes* étaient mal identifiées dans la base de données BOLD, mais les métadonnées fournies (coordonnées et photographies), en plus des méthodes de délinéation moléculaire, nous ont permis de corriger leur identification et de réaliser une étude de phylogéographie du genre **(Couëdel et al. 2023)**. De plus, l'inclusion de différents marqueurs moléculaires dans les référentiels

barcode permet de compléter les bases de référence pour les différents marqueurs, mais permet également d'ancrer les arbres produits à partir des OTU de métabarcoding en reliant les OTU assignés des différents marqueurs les uns aux autres et au spécimen de référence. L'ajout de nos séquences locales aux bases de référence a amélioré l'assignement des OTU retrouvés par métabarcoding pour les deux marqueurs moléculaires employés (18S et COI) et a contribué à abonder les bases de référence **(Mugnai et al. 2023)**. Toutefois, cela n'assure pas pour autant que les séquences de référence soient bien identifiées. Par exemple, sans notre étude de phylogéographie des *Cirripectes*, l'identification associée aux séquences publiques de spécimens collectés sur la zone d'étude aurait été erronée. De ce fait, nous aurions assigné les OTU de *Cirripectes castaneus* à *C. stigmaticus*. Dans notre cas, cette mauvaise identification n'aurait pas eu d'impact sur les résultats, car les analyses écologiques ont été réalisées au niveau du phylum. En revanche, cela peut avoir de réelles conséquences pour des études où le rang taxonomique d'intérêt est celui de l'espèce, comme dans les analyses de contenus stomacaux ou des études avec des objectifs de conservation, où différencier des espèces endémiques de celles à large répartition, comme dans le cas de nos *Cirripectes,* est primordial.

Par ailleurs, certains auteurs proposent d'assigner les OTU à partir d'un sous-jeu de données régional des bases de référence publiques pour limiter les erreurs d'assignement **(Gold et al. 2021 ; Mugnai et al. 2023)**. Ici, cette approche n'aurait pas empêché les mauvais assignements entre *C. castaneus*, *C. stigmaticus* et *C. randalli*. De plus, les résultats présentés dans le chapitre 5 montrent que l'espèce *C. castaneus* possède des haplotypes partagés entre des zones géographiques éloignées telles que La Réunion et la Nouvelle-Calédonie **(Couëdel et al. 2023)**. Ainsi, l'emploi d'un sous-jeu de données aurait, au contraire, pu réduire le nombre de séquences correctement assignées. C'est pourquoi il peut être préférable, et c'est le choix qui a été fait lors de ce projet, de travailler à partir de bases de données nettoyées, mais non restreintes géographiquement, et d'appliquer un pipeline d'assignement strict (*cf.* Chapitre 2 ; **Hleap et al. 2021 ; Bourret et al. 2023**).

3.2. Les méthodes d'assignement

La méthode d'assignement la plus communément utilisée est celle du plus proche voisin. Cependant cette approche est fortement impactée par l'incomplétude des bases de données (*cf.* Chapitre 2 ; **Hleap et al. 2021**). Pour contourner ce biais, il est recommandé d'utiliser la méthode du plus proche ancêtre commun **(Hleap et al. 2021)**, qui assigne à l'entité taxonomique partagée par l'ensemble des séquences (*cf.* Chapitre 2 ; **Huson et al. 2007 ; Kahlke & Ralph 2019**). Cette seconde

approche est, quant à elle, impactée par les erreurs d'identification des séquences de référence et fournit des assignements avec une résolution taxonomique plus faible que celle du plus proche voisin **(Huson et al. 2007 ; Bourret et al. 2023)**. En outre, la stratégie d'assignement avec plusieurs niveaux taxonomiques a été récemment plébiscitée **(Hleap et al. 2021)**. D'une façon générale, aucune méthode d'assignement n'est optimale si elle est appliquée seule. De ce fait, pour augmenter la probabilité d'obtenir des assignements corrects, une combinaison des différentes méthodes d'assignement a été employée ici. À notre connaissance, cette étude est la première à employer jusqu'à quatre niveaux d'assignement différents (pour le COI ; trois pour le 18S), en combinant deux méthodes d'assignement et trois seuils de similarité.

L'une des meilleures façon d'évaluer les aspects méthodologiques du métabarcoding est d'appliquer l'ensemble du processus expérimental à des communautés artificielles de composition connue, également appelées communautés fictives (*mock community* en anglais **Cristescu 2014 ; Leray & Knowlton 2017 ; Braukmann et al. 2019**). La création de communautés fictives permet de connaître (1) le nombre de spécimens, (2) leur biomasse, (3) leur taxonomie et (4) leur séquence barcode **(O'Rourke et al. 2020 ; Iwaszkiewicz-Eggebrecht et al. 2023)**. Elles peuvent être utilisée pour (1) évaluer les erreurs dans les données de séquence obtenues en fin de pipeline **(Gohl et al. 2016 ; Wei et al. 2021)**, (2) optimiser les paramètres de filtrage **(Bokulich et al. 2013 ; González et al. 2023)** et (3) optimiser les différentes méthodes d'assignement **(Bokulich et al. 2018)**. De ce fait, nous aurions dû idéalement travailler avec une telle communauté pour évaluer les gains produits par l'approche bio-informatique développée dans ces travaux. La communauté fictive doit ressembler étroitement à la communauté attendue pour être effective et permettre l'optimisation du pipeline **(Hleap et al. 2021)**. Cependant, les communautés du cryptobiome récifal sont extrêmement diversifiées avec plus de 30 phylums retrouvés **(Leray & Knowlton 2015 ; Pearman et al. 2016 ; Casey et al. 2021 ; Ip et al. 2022)**, ce qui complexifie la création d'une telle communauté. Leray et Knowlton **(2017)** ont créé une communauté fictive à partir d'organismes collectés avec les ARMS composé 34 organismes incluant de 6 phylums, or la communauté obtenue après séquençage était composé des 34 OTU des organismes ciblés et de 86 OTU d'organismes non ciblés (N=120 OTU). Parmi les 86 OTU non ciblés, 31 OTU appartenaient à la faune locale de la zone échantillonnée. Les organismes non ciblés étaient présents sur les pattes des arthropodes et / ou dans les contenus stomacaux des organismes carnivores **(Leray & Knowlton 2017)**. Cet exemple et les contraintes nécessaires au développement d'une telle communauté (organisme identifié et exempte de contamination, séquence connue, temps et matériel nécessaire), nous ont conduit à ne pas en constituer une. Néanmoins, notre étude approfondie du genre *Cirripectes* et des espèces présentes

dans les Mascareignes confirme la validité du pipeline bio-informatique employé lors de ces travaux. L'étude sur *Cirripectes* apporte plusieurs éléments à l'étude métabarcoding :

(1) L'intérêt d'employer des OTU regroupés à un seuil de similarité de *clustering* de 99 % au lieu de 97 % (*cf.* Chapitre 2). En effet, le jeu de données des OTU regroupées au seuil de 99 % a permis la détection de 2 OTU de *Cirripectes castaneus*, alors que dans le jeu de données des OTU regroupés à 97 % aucun OTU n'a été assigné au genre *Cirripectes*.

(2) La nécessité d'utiliser un seuil de similarité plus restrictif pour l'assignement des OTU avec le COI. En effet, généralement, le seuil le plus restrictif employé pour l'assignement des OTU en COI par la méthode du plus proche voisin est de 97 %. Cependant, les résultats présentés dans le chapitre 5 montrent que chez le genre *Cirripectes*, *C. chelomatus* a une variabilité inter-spécifique minimale de 2,6 %. Cette variabilité étant inférieure à 3 %, il est probable que les OTU correspondant à cette espèce ne soient pas correctement assignés (ici, assignés à *C. filamentosus*). De plus, l'analyse de la variabilité intra-spécifique montre que pour 14 des 15 espèces de *Cirripectes* (pour lesquelles l'information est disponible), la variabilité intra-spécifique moyenne est inférieure à 1 %. De ce fait, un assignement avec la méthode du plus proche voisin avec un seuil de similarité de 99 % permet d'assigner correctement au moins 14 espèces de *Cirripectes* sur 15. *C. jenningsi* montre une variation intra-spécifique moyenne de 1,4 %. Ainsi, l'utilisation d'un seuil de similarité à 98% permettrait d'assigner correctement l'ensemble des *Cirripectes* inclus dans l'analyse.

(3) La pertinence d'employer plusieurs méthodes d'assignement pour limiter les erreurs d'assignement. En effet, le point précédent (2) montre qu'avec des seuils de similarités stricts (99 %, voire 98 %), la méthode du plus proche voisin assigne correctement les espèces de *Cirripectes* mais que le risque d'introduire des assignements erronés augmente avec la diminution du seuil de similarité (97 %). Dans ce cas de figure, la méthode du plus proche ancêtre commun, avec un seuil de 97 %, aurait assigné les OTU correctement à l'espèce pour 14 des 15 espèces, et les OTU de la 15^{e} espèce auraient été assignés au genre. De ce fait, l'ensemble des OTU aurait été assigné correctement, même si cela implique une diminution de la résolution taxonomique dans certains cas. C'est pourquoi combiner la méthode du plus proche voisin avec des seuils élevés et la méthode du plus proche ancêtre commun avec des seuils moins stricts permet d'avoir des assignements plus fiables tout en limitant la perte de résolution taxonomique.

(4) La combinaison des différentes méthodes d'assignement permet de diminuer les biais introduits par l'incomplétude des bases de référence, mais ne les résout pas. En effet, on a

observé que pour 10 des 17 espèces de *Cirripectes*, la variabilité inter-spécifique était supérieure à 5 %. Or, dans notre cas, le seuil de similarité le plus bas employé était de 95 %. De ce fait, si la base de référence n'inclut pas de référence pour ces 10 espèces, les OTU correspondants ne pourront pas être assignés, même à des rangs taxonomiques supérieurs.

Toutefois, cet exemple a ses limites car il repose sur un cas idéal où les taxons sont bien représentés dans les bases de référence et où les espèces sont bien délimitées, et de plus sur des vertébrés qui peuvent ne pas être représentatifs d'un point de vue divergence inter-spécifique et variabilité.

4. Coûts cachés du métabarcoding

Plusieurs études ont évalué le coût des différentes techniques de séquençage dans le cas de la création d'un référentiel barcode ou/et d'un suivi de la biodiversité **(Stein et al. 2014 ; Meier et al. 2016 ; Wang et al. 2018)**. Cependant, les coûts humains post-séquençage sont rarement pris en compte et seul le temps nécessaire à la collecte jusqu'à la préparation du séquençage est comptabilisé **(Fu et al. 2021)**.

Pour le barcoding, plusieurs études ont souligné un coût effectif par échantillon plus faible en séquençage NGS qu'en séquençage SANGER **(Meyer & Kircher 2010 ; Hinsinger et al. 2015 ; Shokralla et al. 2015 ; Meier et al. 2016 ; Kerdrel et al. 2020)**. Cependant, Srivathsan et collaborateurs **(2019)** suggèrent la rentabilité des approches NGS seulement partir de 10 000 spécimens séquencés. En outre, le séquençage NGS nécessite de consacrer plus de temps à la reconstitution des séquences. En effet, lors de cette thèse, j'ai été amenée à reconstituer des séquences à partir des deux types de séquençage et ai estimé le temps moyen de reconstitution d'une séquence SANGER à environ 3 min (quasiment toujours < 5 min) contre 7 min pour une séquence multiplexée en NGS. Ce temps pouvait aller jusqu'à 30 min lorsque la qualité des *reads* était moins bonne. Ainsi, si on extrapole aux 200 séquences de COI produites au cours de ces travaux, le temps estimé en SANGER est de 10 h contre supérieur à 23 h pour des séquences en NGS.

De même pour les analyses métabarcoding, le temps consacré aux analyses bio-informatiques est rarement discuté **(Fu et al. 2021)** alors qu'il représente une partie considérable du temps total nécessaire pour acquérir la donnée analysable (ex. liste d'OTU assignées). C'est pourquoi, il est doit être pris en compte au même titre que le temps nécessaire à l'échantillonnage lors de la comparaison des coûts. En effet, il faut prendre en compte le temps de développement du pipeline, qui sera le même pour 10 ou 100 échantillons. Il faut également ajouter le temps nécessaire pour apprendre à utiliser correctement, comparer et évaluer les différentes méthodes. Par ailleurs, après

automatisation du pipeline, il est toujours nécessaire d'y consacrer du temps avec la maintenance, la mise à jour des logiciels et des formats de données. On peut avoir les meilleures données et le meilleur pipeline, si les deux ne sont pas compatibles, ils seront inutilisables en l'état et nécessiteront d'y investir du temps pour les adapter. Rétrospectivement, j'ai tenté d'évaluer la proportion du temps que j'ai passé sur les différentes grandes parties de mon doctorat (Figure 6.2). Mon apprentissage au langage bash et le développement du pipeline d'analyse métabarcoding correspondent à un quart du temps total que j'ai investi dans ce doctorat. Cependant, la figure 6.2 ne reflète pas le temps investi par l'ensemble des personnes impliquées dans l'acquisition des données utilisées lors de ce doctorat et sous-estime notamment le temps d'échantillonnage. En effet, la gestion du programme, le temps de construction et de mise en place des ARMS n'est pas pris en compte, et le tri d'organismes après récolte a été réalisé collégialement par 7, voire 8 personnes. Toutefois, ces étapes auraient également été nécessaires pour des analyses plus traditionnelles. Pour finir, le temps de laboratoire est sous-estimé car une partie des PCR ont été externalisées (laboratoire SSM à Paris, 3 personnes) pour pallier au retard engendré par la COVID-19.

Garijo et collaborateurs **(2013)** ont tenté de quantifier le temps nécessaire pour reproduire un pipeline d'analyse de résultats déjà publiés. Leur étude a montré qu'il faudrait 280 h, soit 8 semaines de travail de 35 h, à un bio-informaticien débutant (chercheur possédant une expertise fondamentale en bio-informatique) pour reproduire les différentes étapes d'analyse. Bien que cette étude ait été publiée en 2013 hors du champ de métabarcoding (biomédicale), les problématiques rencontrées restent les mêmes.

Pour finir, j'ai tenté de lister les différentes étapes qui prennent ou peuvent prendre du temps (extrêmement long pour certaines) dans les analyses bio-informatiques (Figure 6.3). À titre d'expérience personnelle, ouvrir le logiciel Arlequin 3.5 (pour le calcul de la diversité des nucléotides et densité de Tajima ; **Excoffier & Lischer 2010**) sur mon PC a été un vrai parcours du combattant. En effet, le logiciel n'a jamais voulu s'ouvrir sur mon PC Windows 10. Après deux jours de recherches, j'ai trouvé qu'il fallait (1) passer par une machine virtuelle Linux (VirtualBox sous Linux Ubuntu 18) ; (2) installer un émulateur Windows sous cette machine virtuelle (wine), (3) monter un disque dur virtuel pour relier la machine virtuelle au PC et (4) ouvrir Arlequin en ligne de commande *via* la machine virtuelle. Une fois le logiciel ouvert, il a ensuite fallu comprendre le format des données acceptées unique à ce logiciel. Je conseille d'ailleurs les tutoriels Youtube de @DrJBanta.

Ainsi, la mise à disposition des pipelines bio-informatiques et d'un organigramme illustrant le pipeline d'analyse permet une meilleure reproductibilité et dans de meilleurs délais. C'est pourquoi,

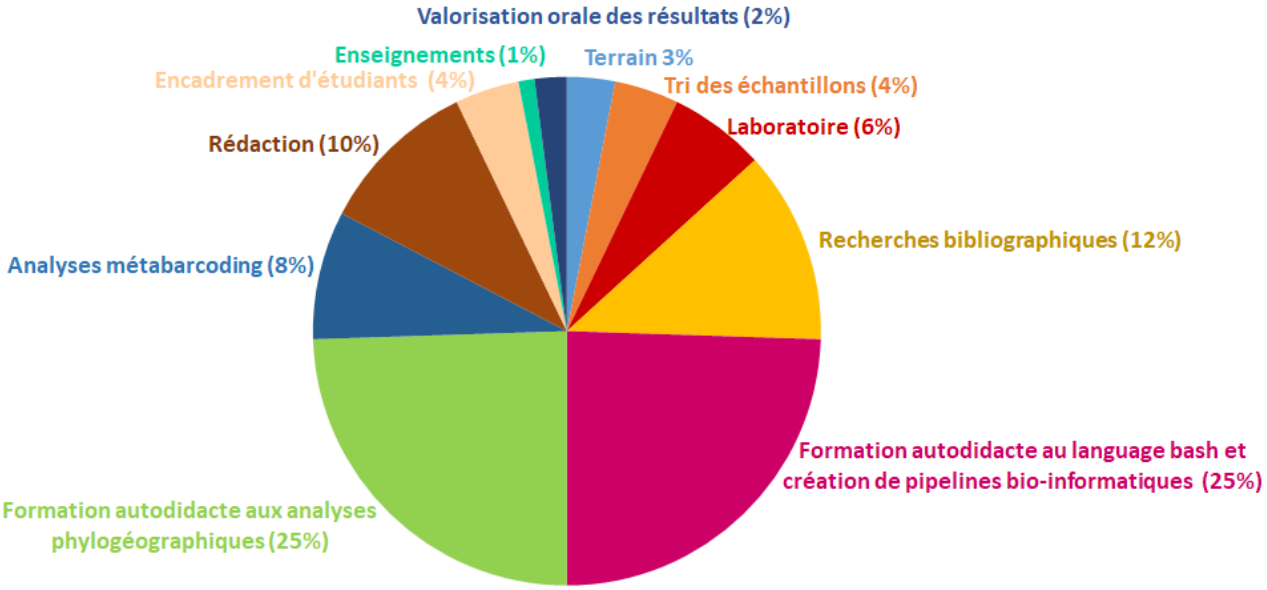


Figure 6.2 : Diagramme circulaire représentant le temps alloué à chaque grande partie de ces travaux de thèse.

lors de la publication de l'article "*New insights into the diversity of cryptobenthic* Cirripectes *blennies in the Mascarene Archipelago sampled using Autonomous Reef Monitoring Structures (ARMS)*", présenté dans le chapitre 5, les données, les scripts et l'organigramme des analyses (Figure 6.4) ont été mis à disposition sur Github :

https://github.com/Mcouedel/Couedel_etal2023_Cirripectes

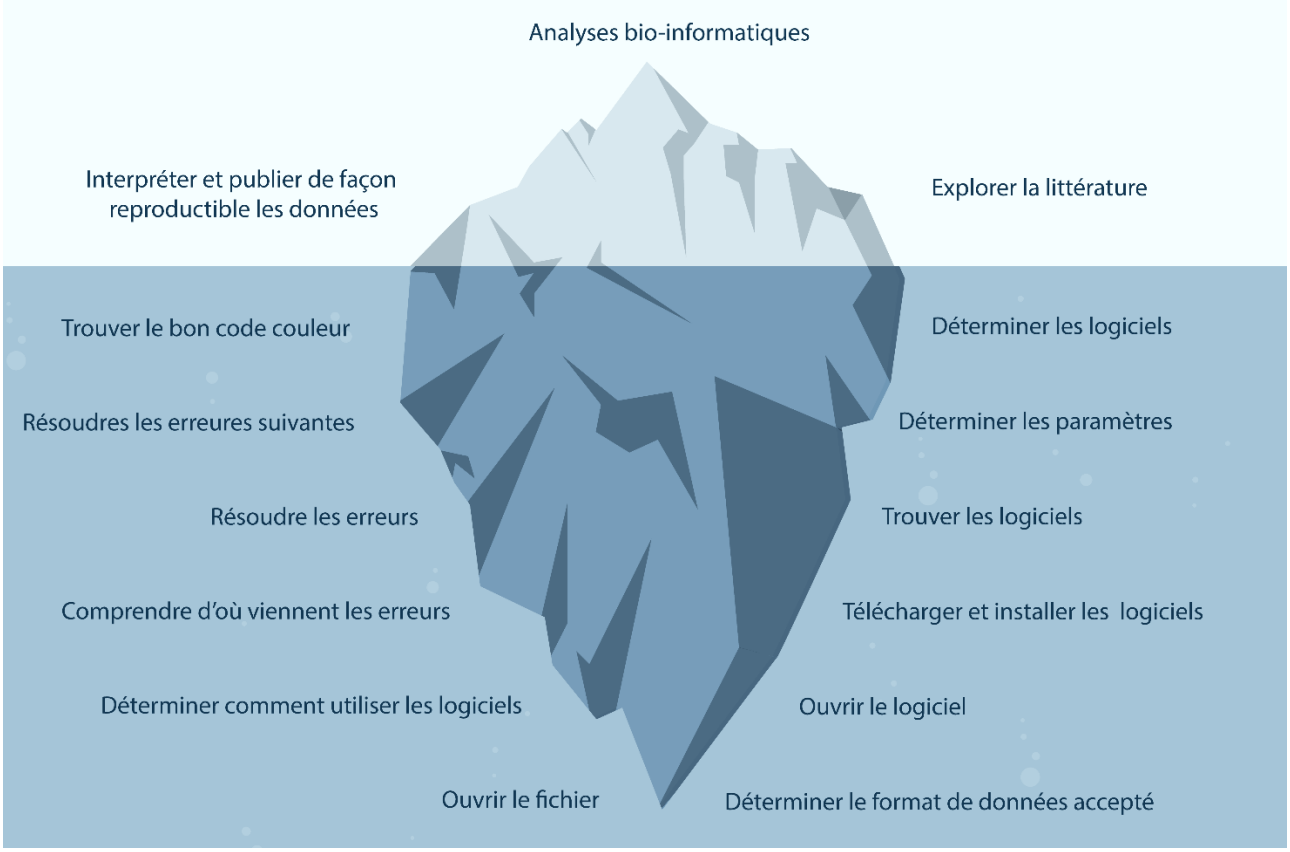


Figure 6.3 : Infographie représentant les différentes étapes qui peuvent poser problème et ainsi prendre du temps dans les analyses bio-informatiques.

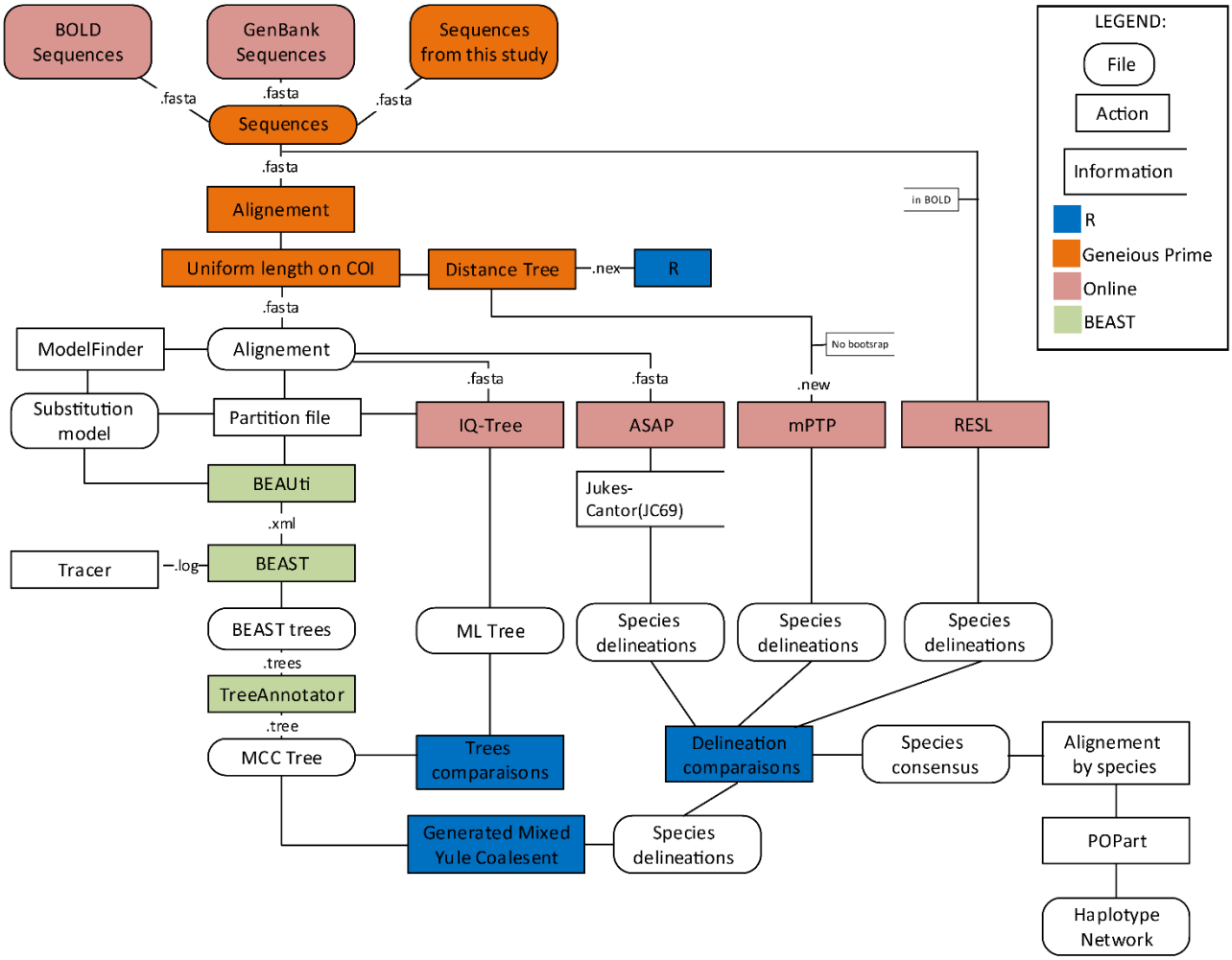


Figure 6.4 : Organigramme illustrant le pipeline d'analyse employé pour l'étude de la phylogéographie des espèces du genre*Cirripectes*.

5. Conclusion

L'ensemble des résultats obtenus durant cette thèse apporte des nouveaux éléments à la connaissance du cryptobiome récifal et contribue à optimiser la méthode des ARMS et du métabarcoding pour évaluer et suivre les écosystèmes coralliens. En effet, ces travaux s'intègrent dans la première étude à coupler les ARMS aux métabarcoding pour étudier le cryptobiome récifal dans l'océan Indien. Les taxons retrouvés correspondent aux taxons du cryptobiome observés dans d'autres régions du monde et reflètent la grande diversité taxonomique observée dans les récifs coralliens. En revanche, la composition des communautés est significativement différente en fonction de la durée d'immersion des ARMS, de la saison de déploiement et de récolte, en particulier pour les taxons sessiles. En effet, la colonisation des ARMS est accomplie dans un premier temps par des espèces pionnières telles que des ascidies solitaires, des cirripèdes et des cnidaires, puis dans un second temps par des espèces de la communauté mature comme des arthropodes mobiles, des ascidies coloniales, des porifères et des rhodophytes. À cela s'ajoute les variations saisonnières. En effet, les bryozoaires, les cnidaires et les porifères ont montré des diversités plus élevées dans les ARMS déployés pendant la saison fraîche, qui concordent avec les concentrations plus élevées de chlorophylle et de matière organique particulaire. Inversement, les rhodophytes présentaient une diversité plus importante pendant la saison chaude.

Par ailleurs, les taxons composant le cryptobiome récifal des Mascareignes sont éloignés génétiquement des espèces référencées dans les bases de données publiques. Il est donc nécessaire de poursuivre son exploration et son référencement. À l'instar des travaux réalisés durant cette thèse avec la production de séquences pour différents marqueurs moléculaires, de nombreuses équipes s'attellent à combler les lacunes dans les connaissances taxonomiques et moléculaires de la biodiversité récifale. Idéalement pour les Mascareignes, l'ensemble des taxons devrait être investigué, néanmoins au vu de l'investissement que cela représente, il semble avantageux de commencer par les arthropodes pour augmenter rapidement le nombre d'identifications par métabarcoding. En outre, l'acquisition de nouvelles connaissances moléculaires pour les Mascareignes permettrait d'évaluer les différents patrons des répartitions biogéographiques des organismes récifaux afin de mieux comprendre les mécanismes à l'origine de cette répartition. Dans ce contexte, la grande diversité échantillonnée par ARMS permettrait une approche multi-taxa et standardisée, améliorable au fil du temps et de l'acquisition du référentiel.

En outre, bien que l'emploi des ARMS connaisse un rapide essor ces dernières années, plusieurs aspects dans leur utilisation manquent de standardisation et limitent les comparaisons entre études. Le temps d'immersion et la saison de déploiement varient entre les études, et nos

travaux ont montré un effet significatif de ces deux facteurs sur les communautés benthiques échantillonnées. Dès lors, nous suggérons de déployer et de récupérer les ARMS pendant la même période de l'année/saison.

Pour finir, ces travaux fournissent un pipeline bio-informatique adapté aux métazoaires des récifs coralliens dans l'objectif de travailler vers la standardisation du traitement des *reads* en sortie de séquenceur. Fournir un nouveau pipeline peut paraître paradoxal vis-à-vis de cet objectif de standardisation. En effet, au premier abord, l'utilisation d'un pipeline déjà publié peut s'avérer plus pertinent. Toutefois, je n'ai pas trouvé de pipeline où tous les paramètres utilisés étaient argumentés et disponibles pour la reproductibilité des analyses. Ainsi, cette thèse s'attache à fournir un pipeline bio-informatique documenté pour les futures analyses de métabarcoding.

6. Références du chapitre 6

Al-Rshaidat MMD., Snider A., Rosebraugh S., Devine AM., Devine TD., Plaisance L., Knowlton N., Leray M. (2016) Deep COI sequencing of standardized benthic samples unveils overlooked diversity of Jordanian coral reefs in the northern Red Sea - Genome. Génome 59:724–737.

Aranda DA., Cárdenas EB., Martínez I., Zárate AZ., Brulé T. (2003) A Review of the Reproductive Patterns Of Gastropod Mollusks from Mexico. BULLETIN OF MARINE SCIENCE 73.

Benkendorff K., Davis A r. (2004) Gastropod egg mass deposition on a temperate, wave-exposed coastline in New South Wales, Australia: implications for intertidal conservation. Aquatic Conservation: Marine and Freshwater Ecosystems 14:263–280. DOI: 10.1002/aqc.604

Berba CMP., Matias AMA. (2022) State of biodiversity documentation in the Philippines: Metadata gaps, taxonomic biases, and spatial biases in the DNA barcode data of animal and plant taxa in the context of species occurrence data. PeerJ 10:e13146. DOI: 10.7717/peerj.13146

Bickford D., Lohman DJ., Sodhi NS., Ng PKL., Meier R., Winker K., Ingram KK., Das I. (2007) Cryptic species as a window on diversity and conservation. Trends in Ecology & Evolution 22:148–155. DOI: 10.1016/j.tree.2006.11.004

Bokulich NA., Kaehler BD., Rideout JR., Dillon M., Bolyen E., Knight R., Huttley GA., Caporaso JG. (2018) Optimizing taxonomic classification of marker gene amplicon sequences. PeerJ Inc. DOI: 10.7287/peerj.preprints.3208v2

Bokulich NA., Subramanian S., Faith JJ., Gevers D., Gordon JI., Knight R., Mills DA., Caporaso JG. (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nature Methods 10:57–59. DOI: 10.1038/nmeth.2276

Bouchet P., Lozouet P., Maestrati P., Heros V. (2002) Assessing the magnitude of species richness in tropical marine environments: exceptionally high numbers of molluscs at a New Caledonia site: Molluscan species richness in a tropical environment. Biological Journal of the Linnean Society 75:421–436. DOI: 10.1046/j.1095-8312.2002.00052.x

Bourret A., Nozères C., Parent E., Parent GJ. (2023) Maximizing the reliability and the number of species assignments in metabarcoding studies using a curated regional library and a public repository. Metabarcoding and Metagenomics 7:e98539. DOI: 10.3897/mbmg.7.98539

Braukmann TWA., Ivanova NV., Prosser SWJ., Elbrecht V., Steinke D., Ratnasingham S., de Waard JR., Sones JE., Zakharov EV., Hebert PDN. (2019) Metabarcoding a diverse arthropod mock community. Molecular Ecology Resources 19:711–727. DOI: 10.1111/1755-0998.13008

Cahill AE., Pearman JK., Borja A., Carugati L., Carvalho S., Danovaro R., Dashfield S., David R., Féral J-P., Olenin S., Šiaulys A., Somerfield PJ., Trayanova A., Uyarra MC., Chenuil A. (2018) A comparative analysis of metabarcoding and morphology-based identification of benthic communities across different regional seas. Ecology and Evolution 8:8908–8920. DOI: https://doi.org/10.1002/ece3.4283

Carvalho S., Aylagas E., Villalobos R., Kattan Y., Berumen M., Pearman JK. (2019) Beyond the visual: using metabarcoding to characterize the hidden reef cryptobiome. Proceedings of the Royal Society B: Biological Sciences 286:20182697. DOI: 10.1098/rspb.2018.2697

Casey JM., Ransome E., Collins AG., Mahardini A., Kurniasih EM., Sembiring A., Schiettekatte NMD., Cahyani NKD., Wahyu Anggoro A., Moore M., Uehling A., Belcaid M., Barber PH., Geller JB., Meyer CP. (2021) DNA metabarcoding marker choice skews perception of marine eukaryotic biodiversity. Environmental DNA 3:1229–1246. DOI: 10.1002/edn3.245

Chakrabarty P., Warren M., Page LM., Baldwin CC. (2013) GenSeq: An updated nomenclature and ranking for genetic sequences from type and non-type sources. ZooKeys 346:29–41. DOI: 10.3897/zookeys.346.5753

Chavan V., Penev L. (2011) The data paper: a mechanism to incentivize data publishing in biodiversity science. BMC Bioinformatics 12:S2. DOI: 10.1186/1471-2105-12-S15-S2

Chazottes V., Reijmer JJG., Cordier E. (2008) Sediment characteristics in reef areas influenced by eutrophication-related alterations of benthic communities and bioerosion processes. Marine Geology 250:114–127. DOI: 10.1016/j.margeo.2008.01.002

Chetoui I., Telahigue K., Bejaoui S., Rabeh I., Ghribi F., Denis F., ElCafsi M. (2019) Annual reproductive cycle and condition index of *Mactra corallina* (Mollusca: Bivalvia) from the north coast of Tunisia. Invertebrate Reproduction & Development 63:40–50. DOI: 10.1080/07924259.2018.1529636

Couëdel M., Dettai A., Guillaume MMM., Bruggemann F., Bureau S., Frattini B., Verde Ferreira A., Azie J-L., Bruggemann JH. (2023) New insights into the diversity of cryptobenthic *Cirripectes* blennies in the Mascarene Archipelago sampled using Autonomous Reef Monitoring Structures (ARMS). Ecology and Evolution 13:e9850. DOI: 10.1002/ece3.9850

Crandall ED., Riginos C., Bird CE., Liggins L., Treml E., Beger M., Barber PH., Connolly SR., Cowman PF., DiBattista JD., Eble JA., Magnuson SF., Horne JB., Kochzius M., Lessios HA., Liu SYV., Ludt WB., Madduppa H., Pandolfi JM., Toonen RJ., Network CM of the D of the I-P., Gaither MR. (2019) The molecular biogeography of the Indo-Pacific: Testing hypotheses with multispecies genetic patterns. Global Ecology and Biogeography 28:943–960. DOI: 10.1111/geb.12905

Cristescu ME. (2014) From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. Trends in Ecology & Evolution 29:566–571. DOI: 10.1016/j.tree.2014.08.001

Enochs IC. (2012) Motile cryptofauna associated with live and dead coral substrates: implications for coral mortality and framework erosion. Marine Biology 159:709–722. DOI: 10.1007/s00227-011-1848-7

Enochs IC., Manzello DP. (2012) Species richness of motile cryptofauna across a gradient of reef framework erosion. Coral Reefs 31:653–661. DOI: 10.1007/s00338-012-0886-z

Excoffier L., Lischer HEL. (2010) Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources 10:564–567.

Froese R., Pauly D. (2023) FishBase. World Wild Web electronic publication. https://www.fishbase.se (accessed 25 April 2023)

Fu M., Hemery L., Sather N. (2021) Cost Efficiency of Environmental DNA as Compared to Conventional Methods for Biodiversity Monitoring Purposes at Marine Energy Sites. Pacific Northwest National Laboratory, Richland, Washington 99354.

Gaither MR., DiBattista JD., Leray M., von der Heyden S. (2022) Metabarcoding the marine environment: from single species to biogeographic patterns. Environmental DNA 4:3–8. DOI: 10.1002/edn3.270

Gaither MR., Rocha LA. (2013) Origins of species richness in the Indo-Malay-Philippine biodiversity hotspot: evidence for the centre of overlap hypothesis. Journal of Biogeography 40:1638–1648. DOI: 10.1111/jbi.12126

Garijo D., Kinnings S., Xie L., Xie L., Zhang Y., Bourne PE., Gil Y. (2013) Quantifying Reproducibility in Computational Biology: The Case of the Tuberculosis Drugome. PLOS ONE 8:e80278. DOI: 10.1371/journal.pone.0080278

Gohl DM., Vangay P., Garbe J., MacLean A., Hauge A., Becker A., Gould TJ., Clayton JB., Johnson TJ., Hunter R., Knights D., Beckman KB. (2016) Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. Nature Biotechnology 34:942–949. DOI: 10.1038/nbt.3601

Gold Z., Curd EE., Goodwin KD., Choi ES., Frable BW., Thompson AR., Walker Jr. HJ., Burton RS., Kacev D., Martz LD., Barber PH. (2021) Improving metabarcoding taxonomic assignment: A case study of fishes in a large marine ecosystem. Molecular Ecology Resources 21:2546–2564. DOI: 10.1111/1755-0998.13450

González A., Dubut V., Corse E., Mekdad R., Dechatre T., Castet U., Hebert R., Meglécz E. (2023) VTAM: A robust pipeline for validating metabarcoding data using controls. Computational and Structural Biotechnology Journal 21:1151–1156. DOI: 10.1016/j.csbj.2023.01.034

Guest J., Goh B., Chou L. (2012) Reproductive seasonality of the reef building coral *Platygyra pini* on Singapore's reefs. The Raffles bulletin of zoology Supplement:123–131.

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

Hinsinger DD., Debruyne R., Thomas M., Denys GPJ., Mennesson M., Utage J., Dettai A. (2015) Fishing for barcodes in the Torrent: from COI to complete mitogenomes on NGS platforms. DNA Barcodes 3:170–186. DOI: 10.1515/dna-2015-0019

Hleap JS., Littlefair JE., Steinke D., Hebert PDN., Cristescu ME. (2021) Assessment of current taxonomic assignment strategies for metabarcoding eukaryotes. Molecular Ecology Resources 21:2190–2203. DOI: 10.1111/1755-0998.13407

Hoareau TB., Boissin E., Paulay G., Bruggemann JH. (2013) The Southwestern Indian Ocean as a potential marine evolutionary hotspot: perspectives from comparative phylogeography of reef brittle-stars. Journal of Biogeography 40:2167–2179. DOI: 10.1111/jbi.12155

Hubert N., Dettai A., Pruvost P., Cruaud C., Kulbicki M., Myers R., Borsa P. (2017) Geography and life history traits account for the accumulation of cryptic diversity among Indo-West Pacific coral reef fishes. Marine Ecology Progress Series 583:179–193. DOI: 10.3354/meps12316

Hubert N., Meyer CP., Bruggemann HJ., Guérin F., Komeno RJL., Espiau B., Causse R., Williams JT., Planes S. (2012) Cryptic Diversity in Indo-Pacific Coral-Reef Fishes Revealed by DNA-Barcoding Provides New Support to the Centre-of-Overlap Hypothesis. PLOS ONE 7:e28987. DOI: 10.1371/journal.pone.0028987

Huson DH., Auch AF., Qi J., Schuster SC. (2007) MEGAN analysis of metagenomic data. Genome Research 17:377–386. DOI: 10.1101/gr.5969107

Ip YCA., Chang JJM., Oh RM., Quek ZBR., Chan YKS., Bauman AG., Huang D. (2022) Seq' and ARMS shall find: DNA (meta)barcoding of Autonomous Reef Monitoring Structures across the tree of life uncovers hidden cryptobiome of tropical urban coral reefs. Molecular Ecology n/a. DOI: 10.1111/mec.16568

Iwaszkiewicz-Eggebrecht E., Granqvist E., Buczek M., Prus M., Kudlicka J., Roslin T., Tack AJM., Andersson AF., Miraldo A., Ronquist F., Łukasik P. (2023) Optimizing insect metabarcoding using replicated mock communities. Methods in Ecology and Evolution 14:1130–1146. DOI: 10.1111/2041-210X.14073

Jarman S., Berry O., Bunce M. (2018) The value of environmental DNA biobanking for long-term biomonitoring. Nature Ecology & Evolution 2. DOI: 10.1038/s41559-018-0614-3

Joly S., Davies TJ., Archambault A., Bruneau A., Derry A., Kembel SW., Peres-Neto P., Vamosi J., Wheeler TA. (2014) Ecology in the age of DNA barcoding: the resource, the promise and the challenges ahead. Molecular Ecology Resources 14:221–232. DOI: 10.1111/1755-0998.12173

Kahlke T., Ralph PJ. (2019) BASTA – Taxonomic classification of sequences and sequence bins using last common ancestor estimations. Methods in Ecology and Evolution 10:100–103. DOI: 10.1111/2041-210X.13095

Keck F., Couton M., Altermatt F. (2023) Navigating the seven challenges of taxonomic reference databases in metabarcoding analyses. Molecular Ecology Resources 23:742–755. DOI: 10.1111/1755-0998.13746

Kerdrel GA de., Andersen JC., Kennedy SR., Gillespie R., Krehenwinkel H. (2020) Rapid and cost-effective generation of single specimen multilocus barcoding data from whole arthropod

communities by multiple levels of multiplexing. Scientific Reports 10:1–12. DOI: 10.1038/s41598-019-54927-z

Keyse J., Crandall ED., Toonen RJ., Meyer CP., Treml EA., Riginos C. (2014) The scope of published population genetic data for Indo-Pacific marine fauna and future research opportunities in the region. Bulletin of Marine Science 90:47–78. DOI: 10.5343/bms.2012.1107

Kolasinski J., Rogers K., Cuet P., Barry B., Frouin P. (2011) Sources of particulate organic matter at the ecosystem scale: a stable isotope and trace element study in a tropical coral reef. Marine Ecology Progress Series 443:77–93. DOI: 10.3354/meps09416

Larkin M., Smith S., Willan R., Davis T. (2017) Diel and seasonal variation in heterobranch sea slug assemblages within an embayment in temperate eastern Australia. Marine Biodiversity. DOI: 10.1007/s12526-017-0700-9

Leray M., Knowlton N. (2015) DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. Proceedings of the National Academy of Sciences 112:2076–2081. DOI: 10.1073/pnas.1424997112

Leray M., Knowlton N. (2017) Random sampling causes the low reproducibility of rare eukaryotic OTUs in Illumina COI metabarcoding. PeerJ 5:e3006. DOI: 10.7717/peerj.3006

Leray M., Knowlton N. (2016) Visualizing Patterns of Marine Eukaryotic Diversity from Metabarcoding Data Using QIIME. In: *Marine Genomics*. Methods in Molecular Biology, Bourlat SJ (ed) Springer New York, New York, NY, p 219–235 DOI: 10.1007/978-1-4939-3774-5_15

Lindh MV., Sjöstedt J., Ekstam B., Casini M., Lundin D., Hugerth LW., Hu YOO., Andersson AF., Andersson A., Legrand C., Pinhassi J. (2017) Metapopulation theory identifies biogeographical patterns among core and satellite marine bacteria scaling from tens to thousands of kilometers. Environmental Microbiology 19:1222–1236. DOI: 10.1111/1462-2920.13650

Logares R., Audic S., Bass D., Bittner L., Boutte C., Christen R., Claverie J-M., Decelle J., Dolan JR., Dunthorn M., Edvardsen B., Gobet A., Kooistra WHCF., Mahé F., Not F., Ogata H., Pawlowski J., Pernice MC., Romac S., Shalchian-Tabrizi K., Simon N., Stoeck T., Santini S., Siano R., Wincker P., Zingone A., Richards TA., de Vargas C., Massana R. (2014) Patterns of Rare and Abundant Marine Microbial Eukaryotes. Current Biology 24:813–821. DOI: 10.1016/j.cub.2014.02.050

Massé L. (2014) Comparaison de la reproduction sexuée et du recrutement des coraux scléractiniaires entre un récif tropical (La Réunion) et subtropical (Afrique du Sud) du sud-ouest de l'océan Indien. These de doctorat, La Réunion

Meier R., Wong W., Srivathsan A., Foo M. (2016) $1 DNA barcodes for reconstructing complex phenomes and finding rare species in specimen-rich samples. Cladistics 32:100–110. DOI: 10.1111/cla.12115

Meyer M., Kircher M. (2010) Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. Cold Spring Harbor protocols 2010:pdb.prot5448. DOI: 10.1101/pdb.prot5448

Monchamp M-E., Taranu Z., Garner R., Re T., Morissette O., Iversen L., Fugère V., Littlefair J., Barbosa da Costa N., Desforges J., Schacht J., Derry A., Cooke S., Barrett R., Walsh D., Ragoussis J., Albert M., Cristescu M., Gregory-Eaves I. (2023) Prioritizing taxa for genetic reference database development to advance inland water conservation. Biological Conservation 280. DOI: 10.1016/j.biocon.2023.109963

Monroy-Velázquez LV., Rodríguez-Martínez RE., Blanchon P., Alvarez F. (2020) The use of artificial substrate units to improve inventories of cryptic crustacean species on Caribbean coral reefs. PeerJ 8:e10389. DOI: 10.7717/peerj.10389

Morton B., Blackmore G. (2009) Seasonal variations in the density of and corallivory by *Drupella rugosa* and *Cronia margariticola* (Caenogastropoda: Muricidae) from the coastal waters of Hong Kong: 'plagues' or 'aggregations'? Journal of the Marine Biological Association of the United Kingdom 89:147–159. DOI: 10.1017/S002531540800218X

Mugnai F., Costantini F., Chenuil A., Leduc M., Ortega JMG., Meglécz E. (2023) Be positive: customized reference databases and new, local barcodes balance false taxonomic assignments in metabarcoding studies. PeerJ 11:e14616. DOI: 10.7717/peerj.14616

Muthiga NA., Jaccarini V. (2005) Effects of seasonality and population density on the reproduction of the Indo-Pacific echinoid *Echinometra mathaei* in Kenyan coral reef lagoons. Marine Biology 146:445–453. DOI: 10.1007/s00227-004-1449-9

Muths D., Tessier E., Bourjea J. (2015) Genetic structure of the reef grouper *Epinephelus merra* in the West Indian Ocean appears congruent with biogeographic and oceanographic boundaries. Marine Ecology 36:447–461. DOI: 10.1111/maec.12153

Myers RF. (1991) Micronesian reef fishes., 2nd Edition. Coral Graphics, Guam, USA., 298 p.

Naim O. (1993) Seasonal responses of a fringing reef community to eutrophication (Reunion Island, Western Indian Ocean). Marine Ecology Progress Series 99:137–151. DOI: 10.3354/meps099137

Nichols PK., Timmers M., Marko PB. (2021) Hide 'n seq: Direct versus indirect metabarcoding of coral reef cryptic communities. Environmental DNA 4:93–107. DOI: 10.1002/edn3.203

O'Rourke DR., Bokulich NA., Jusino MA., MacManes MD., Foster JT. (2020) A total crapshoot? Evaluating bioinformatic decisions in animal diet metabarcoding analyses. Ecology and Evolution 10:9721–9739. DOI: 10.1002/ece3.6594

Oury N. (2022) De la délimitation des espèces à la diversité intra-coloniale : apport de la génomique chez les coraux du genre *Pocillopora* dans l'Indo-Pacifique. phdthesis, Université de la Réunion

Palant B., Fishelson L. (2013) *Littorina punctata* (gmelin) and *Pittorina neritoides* (l.), (Mollusca, Gastropoda) from israel: ecology and annual cycle of genital system. Israel Journal of Zoology.

Pasnin O., Voigt O., Wörheide G., Murillo Rincón AP., von der Heyden S. (2020) Indo-Pacific Phylogeography of the Lemon Sponge *Leucetta chagosensis*. Diversity 12:466. DOI: 10.3390/d12120466

Pearman JK., Anlauf H., Irigoien X., Carvalho S. (2016) Please mind the gap – Visual census and cryptic biodiversity assessment at central Red Sea coral reefs. Marine Environmental Research 118:20–30. DOI: 10.1016/j.marenvres.2016.04.011

Pearman JK., Chust G., Aylagas E., Villarino E., Watson JR., Chenuil A., Borja A., Cahill AE., Carugati L., Danovaro R., David R., Irigoien X., Mendibil I., Moncheva S., Rodríguez-Ezpeleta N., Uyarra MC., Carvalho S. (2020) Pan-regional marine benthic cryptobiome biodiversity patterns revealed by metabarcoding Autonomous Reef Monitoring Structures. Molecular Ecology n/a. DOI: 10.1111/mec.15692

Pearman JK., Leray M., Villalobos R., Machida RJ., Berumen ML., Knowlton N., Carvalho S. (2018) Cross-shelf investigation of coral reef cryptic benthic organisms reveals diversity patterns of the hidden majority. Scientific Reports 8:1–17. DOI: 10.1038/s41598-018-26332-5

Pisapia C., Stella J., Silbiger NJ., Carpenter R. (2020) Epifaunal invertebrate assemblages associated with branching Pocilloporids in Moorea, French Polynesia. PeerJ 8:e9364. DOI: 10.7717/peerj.9364

Plaisance L., Caley MJ., Brainard RE., Knowlton N. (2011) The Diversity of Coral Reefs: What Are We Missing? PLOS ONE 6:e25026. DOI: 10.1371/journal.pone.0025026

Postaire B., Gélin P., Bruggemann JH., Magalon H. (2017) One species for one island? Unexpected diversity and weak connectivity in a widely distributed tropical hydrozoan. Heredity 118:385–394. DOI: 10.1038/hdy.2016.126

Ransome E., Geller JB., Timmers M., Leray M., Mahardini A., Sembiring A., Collins AG., Meyer CP. (2017) The importance of standardization for biodiversity comparisons: A case study using autonomous reef monitoring structures (ARMS) and metabarcoding to measure cryptic diversity on Mo'orea coral reefs, French Polynesia. PLOS ONE 12:e0175066. DOI: 10.1371/journal.pone.0175066

Ratnasingham S., Hebert PDN. (2007) BOLD: The Barcode of Life Data System. Molecular Ecology Notes. DOI: 10.1111/j.1471-8286.2006.01678.x

Rimet F., Aylagas E., Borja Á., Bouchez A., Canino A., Chauvin C., Chonova T., Ciampor Jr F., Costa FO., Ferrari BJD., Gastineau R., Goulon C., Gugger M., Holzmann M., Jahn R., Kahlert M., Kusber W-H., Laplace-Treyture C., Leese F., Leliaert F., Mann DG., Marchand F., Méléder V., Pawlowski J., Rasconi S., Rivera S., Rougerie R., Schweizer M., Trobajo R., Vasselon V., Vivien R., Weigand A., Witkowski A., Zimmermann J., Ekrem T. (2021) Metadata standards and practical guidelines for specimen and DNA curation when building barcode reference libraries for aquatic life. Metabarcoding and Metagenomics 5:e58056. DOI: 10.3897/mbmg.5.58056

Rulik B., Eberle J., von der Mark L., Thormann J., Jung M., Köhler F., Apfel W., Weigel A., Kopetz A., Köhler J., Fritzlar F., Hartmann M., Hadulla K., Schmidt J., Hörren T., Krebs D., Theves F., Eulitz U., Skale A., Rohwedder D., Kleeberg A., Astrin JJ., Geiger MF., Wägele JW., Grobe P., Ahrens D. (2017) Using taxonomic consistency with semi-automated data pre-processing for high quality DNA barcodes. Methods in Ecology and Evolution 8:1878–1887. DOI: 10.1111/2041-210X.12824

Shea MM., Kuppermann J., Rogers MP., Smith DS., Edwards P., Boehm AB. (2023) Systematic review of marine environmental DNA metabarcoding studies: toward best practices for data usability and accessibility. PeerJ 11:e14993. DOI: 10.7717/peerj.14993

Shenkar N., Bronstein O., Loya Y. (2008) Population dynamics of a coral reef ascidian in a deteriorating environment. Marine Ecology Progress Series 367:163–171. DOI: 10.3354/meps07579

Shokralla S., Porter TM., Gibson JF., Dobosz R., Janzen DH., Hallwachs W., Golding GB., Hajibabaei M. (2015) Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. Scientific Reports 5:9687. DOI: 10.1038/srep09687

Srivathsan A., Hartop E., Puniamoorthy J., Lee WT., Kutty SN., Kurina O., Meier R. (2019) Rapid, large-scale species discovery in hyperdiverse taxa using 1D MinION sequencing. BMC Biology 17:96. DOI: 10.1186/s12915-019-0706-9

Stein ED., Martinez MC., Stiles S., Miller PE., Zakharov EV. (2014) Is DNA Barcoding Actually Cheaper and Faster than Traditional Morphological Methods: Results from a Survey of Freshwater Bioassessment Efforts in the United States? PLOS ONE 9:e95525. DOI: 10.1371/journal.pone.0095525

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

Takemura A., Rahman MdS., Nakamura S., Park YJ., Takano K. (2004) Lunar cycles and reproductive activity in reef fishes with particular attention to rabbitfishes. Fish and Fisheries 5:317–328. DOI: 10.1111/j.1467-2679.2004.00164.x

Turon X., Antich A., Palacín C., Præbel K., Wangensteen OS. (2020) From metabarcoding to metaphylogeography: separating the wheat from the chaff. Ecological Applications 30:e02036. DOI: 10.1002/eap.2036

Vicente-Saez R., Martinez-Fuentes C. (2018) Open Science now: A systematic literature review for an integrated definition. Journal of Business Research 88:428–436. DOI: 10.1016/j.jbusres.2017.12.043

Villalobos R., Aylagas E., Pearman J., Cúrdia J., Lozano-Cortés D., Coker D., Jones B., Berumen M., Carvalho S. (2022) Inter-annual variability patterns of reef cryptobiota in the central Red Sea across a shelf gradient. Scientific Reports 12. DOI: 10.1038/s41598-022-21304-2

Wang WY., Srivathsan A., Foo M., Yamane SK., Meier R. (2018) Sorting specimen-rich invertebrate samples with cost-effective NGS barcodes: Validating a reverse workflow for specimen processing. Molecular ecology resources 18:490–501.

Wei Z-G., Zhang X-D., Cao M., Liu F., Qian Y., Zhang S-W. (2021) Comparison of Methods for Picking the Operational Taxonomic Units From Amplicon Sequences. Frontiers in Microbiology 12.

Résumé

Dans le contexte du changement global, la surveillance de la biodiversité est essentielle pour détecter les facteurs de perturbation et comprendre les réponses des communautés. Les récifs coralliens font partie des écosystèmes les plus riches et diversifiés de la planète, mais également des plus menacés. Pourtant, les organismes, souvent de petite taille, vivant cachés dans les anfractuosités du récif et constituant le cryptobiome, restent très peu étudiés alors qu'ils représentent la majeure partie de la diversité récifale et un maillon fondamental du réseau trophique. Si l'évaluation de la diversité est essentielle pour une gestion efficace des écosystèmes, les méthodes traditionnelles de taxonomie, basées sur la morphologie, sont peu adaptées au cryptobiome en étant chronophages, nécessitant des connaissances spécialisées et difficilement comparables entre sites. Ces petits taxons présentent des défis supplémentaires en étant plus difficiles à trouver et à identifier. Avec le métabarcoding d'ADN, les récents progrès des techniques de séquençage à haut-débit et de la bio-informatique offrent une alternative aux méthodes traditionnelles.

Cette thèse porte sur la diversité et l'écologie du cryptobiome récifal des Mascareignes (La Réunion et Rodrigues), collecté à l'aide de structures d'échantillonnage standardisées, les ARMS, et étudié par l'approche métabarcoding. Dans un premier temps, un pipeline bio-informatique adapté aux analyses métabarcoding de métazoaires a été élaboré pour évaluer les communautés échantillonnées. Les interprétations écologiques découlant de l'approche par métabarcoding sont très dépendantes des références moléculaires disponibles. C'est pourquoi, dans un second temps, ces travaux se sont attachés à initier un référentiel moléculaire pour documenter le cryptobiome des Mascareignes, en produisant des séquences pour trois marqueurs moléculaires, le 18S, le COI et le 16S, ainsi que le mitogénome complet pour l'ensemble des téléostéens échantillonnés. Dans un troisième temps, ces travaux ont évalué la diversité du cryptobiome et sa cinétique de colonisation à travers les saisons chaude et fraîche à La Réunion. Les taxons retrouvés correspondent aux taxons du cryptobiome observé dans d'autres régions du monde et reflète la grande diversité taxonomique observée dans les récifs coralliens. Ces travaux sont les premiers mettre en évidence des différences significatives dans la composition des communautés en fonction de la durée d'immersion des ARMS, de la saison de déploiement et de récolte. Ces facteurs influencent en particulier les taxons sessiles tels que les ascidies, les cnidaires, les porifères et les rhodophytes. Dès lors, ces travaux démontrent la nécessité de considérer ces variations dans l'analyse des résultats et leurs comparaisons avec d'autres études, ainsi que leurs implications dans l'utilisation des ARMS en tant qu'outil de suivi. Dans un dernier temps, la méthode des ARMS couplée aux méthodes d'identification moléculaire pour évaluer la diversité et la distribution des espèces a été examinée. Un regard particulier a été porté sur les patrons de distribution des téléostéens cryptobenthiques du genre *Cirripectes* trouvés dans les Mascareignes. En effet, de récentes études ont découvert l'endémisme de certaines espèces du genre à des zones géographiques restreintes, malgré la distribution généralement large des espèces de blennies tropicales et subtropicales. Parmi les trois espèces collectées, deux ont montré une distribution Indo-Pacifique et la troisième un endémisme aux Mascareignes.

Les travaux de cette thèse posent une base indispensable aux connaissances sur la diversité du cryptobiome récifal des Mascareignes. Ils apportent une meilleure compréhension des processus de colonisation des communautés cryptobenthiques et permettront d'améliorer l'utilisation des ARMS dans les récifs coralliens et l'emploi de méthodes de taxonomie moléculaire dans le Sud-Ouest de l'océan Indien.

Mots clés : cryptobiome, récifs coralliens, métabarcoding, écologie moléculaire, Sud-Ouest de l'océan Indien, diversité