Accumulation curves of environmental DNA sequences predict coastal fish diversity in the Coral Triangle

Jean-Baptiste Juhel[1], Rizkie Utama[2], Virginie Marques[1], Indra Vimono[2], Hagi Sugeha[2], kadarusman kadarusman[3], Laurent Pouyaud[4], Tony Dejean[5], David Mouillot[6], and Régis Hocdé[1]

[1]Universite Montpellier Faculte des Sciences de Montpellier
[2]Lembaga Ilmu Pengetahuan Indonesia
[3]Politeknik Kelautan dan Perikanan Sorong
[4]Institut des Sciences de l'Evolution de Montpellier
[5]SPYGEN
[6]Université Montpellier-CNRS-IFREMER

May 5, 2020

Abstract

Environmental DNA (eDNA) has the potential to provide more comprehensive biodiversity assessments particularly for vertebrates in species-rich regions. Yet, this method requires the completeness of a reference database, i.e. a list of DNA sequences attached to each species, which is never met. As an alternative, a diversity of Operational Taxonomic Units (OTUs) can be extracted from eDNA metabarcoding. However, the extent to which the diversity of OTUs provided by a limited eDNA sampling effort can predict regional species diversity is unknown. Here, by modelling OTU accumulation curves of eDNA seawater samples across the Coral Triangle, we obtained an asymptote reaching 1,531 fish OTUs while 1,611 fish species are recorded in the region. Besides, we also accurately predict ($R^2 = 0.92$) the distribution of species richness among fish families from OTU-based asymptotes. Thus, the multi-model framework of OTU accumulation curves extends the use of eDNA metabarcoding in ecology, biogeography and conservation.

Introduction

Providing accurate biodiversity assessments is a critical goal in ecology and biogeography with estimations being constantly revised for some species-rich groups ((1) Costello et al 2017). This issue is increasingly important given the accelerating human footprint on Earth. The ongoing worldwide defaunation, characterized by massive population declines, may trigger the local or even global extinction of rare, elusive and cryptic species that are still unknown or poorly documented ((2) Barlow et al 2018, (3) Lees & Pimm 2015). Such biodiversity losses directly impact ecosystem functioning but also human health, well-being and livelihood ((4) Diaz et al 2018, (5) Duffy et al 2017). This urges scientists to improve the accuracy and to extend the breadth of biodiversity inventories and monitoring.

In the marine realm, the detection of species occurrences is particularly challenging due to the vast volume to monitor, the high diversity of habitats, the inaccessibility of some areas (e.g. deep sea) and the behavior of some species (cryptobenthic or elusive) ((6) Juhel et al 2019, (7) Brandl et al 2018). Environmental DNA (eDNA) metabarcoding is an emerging tool that can provide more accurate and wider biodiversity assessments than classical census methods particularly for rare and elusive species ((8) Garlapati et al 2019, (9) Boussarie et al 2018, (10) Fukumoto et al 2015). This non-invasive method is based on retrieving DNA naturally released by organisms in their environment, amplified by polymerase chain reaction (PCR) and

then sequenced to ultimately identify corresponding species ((11) Ruppert et al 2019). However, inventorying and monitoring biodiversity using eDNA metabarcoding requires the completeness of a reference database to accurately assign each sequence to a given species (e.g. (9) Boussarie et al. 2018).

By now, only a minority of fish species are present in online DNA databases for mitochondrial regions targeted by metabarcoding markers, limiting the extent to which species diversity can be revealed by eDNA. This proportion of sequenced species is even lower in species-rich regions and poorly sampled habitats or taxa while the effort to complete genetic reference databases is long and costly. As an alternative, a diversity of Operational Taxonomic Units (OTUs) can be extracted from eDNA metabarcoding through filtering and clustering techniques ((12) Mahé et al 2014). However, the extent to which the diversity of OTUs from a limited number of eDNA samples can reveal or predict the diversity of vertebrate species in a given biodiversity hotspot has not yet been investigated. This is particularly challenging for cryptobenthic fish species that are key for reef ecosystems ((13) Brandl et al 2019) but usually missed by classical surveys ((7) Brandl et al. 2018).

The Bird's Head Peninsula of West Papua (eastern Indonesia) is located in the center of the Coral Triangle (marine area extending from Malaysia to Solomon Islands, (14) Veron et al 2009) which is known to host the world's richest marine biodiversity ((15) Allen & Erdmann 2012, (16) Mangubhai et al 2012). The current checklist of coastal fishes in the Bird's Head Peninsula identifies 1,611 species belonging to 508 genera and 112 families ((17) Kulbicki et al 2013, (15) Allen & Erdmann 2012). This exceptional level of diversity and endemism is due to a combination of factors: the complex marine currents and history of sea level changes ((18) Mora et al 2003), vicariance and dispersal at various spatio-temporal scales ((19) Hubert et al 2017), a variety of habitats from karsts to many types of coral reefs ((16) Mangubhai et al 2012), the impact of geographic isolation ((19) Hubert et al 2017), the result of plates tectonic history ((20) Leprieur et al 2016, (21) Gaboriau et al 2019) and stable environmental conditions during the quaternary ((22) Pelissier et al 2014). As a result, Indonesia is endowed with an exceptional fish diversity which is still poorly known and under severe threats ((23) Exton et al 2019, (24) Jones et al 2019, (25) Ainsworth et al 2008). Predicting the level of vertebrate diversity from eDNA OTUs is thus a critical step in conservation, biogeography and ecology, particularly in marine biodiversity hotspots.

Here, using eDNA metabarcoding from 92 seawater samples across the Bird's Head Peninsula, we (*i*) assessed the diversity of coastal fish species based on an online reference database for the teleo primers region of the 12S mitochondrial rDNA gene ((26) Valentini et al 2016), (*ii*) estimated the diversity of fish OTUs based on a custom filtering and clustering bioinformatic pipeline, and (*iii*) tested the capacity of OTU accumulation curves to predict the regional fish diversity.

Materials and Methods

Sampling area and protocol

A total of 92 water samples were collected during October and November 2017 along the south coast of the Bird's Head region of West Papua (500 km) across different habitats but mainly coral reefs (Fig. 1). Samples were collected in DNA-free plastic bags at the surface from a dinghy boat, at depths between 10 – 100m during close circuit rebreather dives, and (iii) at depths between 100 - 300m using Niskin water samplers. A pressure and temperature sensor was coupled to the Niskin bottle to control the sampling depth and characterize the water mass via the vertical temperature profile. For each sample, 2L of seawater were filtered with sterile Sterivex filter capsules (Merck© Millipore; pore size 0.22µm) using disposable sterile syringes. Immediately after, the filter units were filled with lysis conservation buffer (CL1 buffer SPYGEN©) and stored in 50 mL screw-cap tubes at -20°C. A contamination control protocol was followed in both field and laboratory stages ((27) Goldberg et al 2016; (26) Valentini et al. 2016). Water sample processing included the use of disposable gloves and single-use filtration equipment, and the bleaching (50% bleach) of Niskin water sampler.

DNA extraction, amplification and high-throughput sequencing

The DNA extraction was performed in a dedicated controlled DNA laboratory (SPYGEN,

*www.spygen.com)*equipped with separate cleanrooms with positive air pressure, UV treatment and frequent air renewal. Decontamination procedures were conducted before and after all manipulation. Each filtration capsule was agitated for 15min on S50 Shaker (Cat Ingenieurbüro) at 800 rpm. The buffer was retrieved using a 3 mL BD Disposable Syringe with Luer-Lok tips, emptied into a 50mL tube containing 33 mL of ethanol and 1.5 mL of 3M sodium acetate and, finally, stored for at least one night at -20degC. The DNA extraction and amplification were performed following the protocol of (28) Pont et al 2018 including 12 separate PCR amplifications per sample. A teleost-specific 12S mitochondrial rDNA primer (teleo, forward primer-ACACCGCCCGTCACTCT, reverse primer -CTTCCGGTACACTTACCATG, (26) Valentini et al. 2016) was used for the amplification of metabarcoding sequences. Eight negative extraction controls and two negative PCR controls (ultrapure water) were amplified (with 12 replicates as well) and sequenced in parallel to the samples to monitor possible contaminations. The teleo primers were 5'-labeled with an eight-nucleotide tag unique to each PCR replicate with at least three differences between any pair of tags, allowing the assignment of each sequence to the corresponding sample during sequence analysis. The tags for the forward and reverse primers were identical for each PCR replicate.

After amplification, samples were titrated using capillary electrophoresis (QIAxcel; Qiagen GmbH, Hilden, Germany) and purified using a MinElute PCR purification kit (Qiagen GmbH, Hilden, Germany). Before sequencing, purified DNA was titrated using capillary electrophoresis. The purified PCR products were pooled in equal volumes, to achieve a theoretical sequencing depth of 1,000,000 reads per sample. Library preparation and sequencing were performed at Fasteris (Geneva, Switzerland). A total of five libraries were prepared using MetaFast protocol (Fasteris,*https://www.fasteris.com/dna/?q=content/metafast-protocol-amplicon-metagenomic-analysis*). A paired-end sequencing (2x125 bp) was carried out using an Illumina HiSeq 2500 sequencer on three HiSeq Rapid Flow Cell v2 using the HiSeq Rapid SBS Kit v2 (Illumina, San Diego, CA, USA) following the manufacturer's instructions.

Sequence analyses and taxonomic assignment

To evaluate the current completeness of the online database for the teleo region of the 12S mitochondrial DNA, an *in silico* PCR with 3 allowed mismatches using the teleo primers sequences was performed with ecoPCR ((29) Ficetola et al 2010) on the EMBL database (European Molecular Biology Laboratory,*www.ebi.ac.uk*, version 138, downloaded on January 2019, ((30) Baker et al 2000). The generated list of sequenced species was compared to the checklists of fish species present in in the Bird's Head of Papua region, provided by courtesy of (17) Kulbicki et al. 2013.

The amplified DNA sequences from the water samples were processed following two metabarcoding workflows. The first workflow used the OBITools software package ((31) Boyer et al 2016) based on direct taxonomic assignment of the sequences using the ecotag program (lower common ancestor algorithm) in EMBL database as a reference (see details in Supplementary materials).

A total of 394 species are sequenced in the Bird's Head region (24.5%, Suppl table 1). The selection of similarity thresholds for taxonomic assignment must be based on the length of the barcode and its intra taxonomic variability. We tested the resolution of the marker by running an *in silico* PCR on all fish mitochondrial DNA present in EMBL online database (downloaded the 20th of January 2019). All amplified sequences were aligned using Clustal W algorithm ((32) Larkin et al 2007) and their identity percentage calculated using Geneious R6.1.8 ((33) Kearse et al 2012). The analysis of this alignments supports the following thresholds with few false assignments at those taxonomical levels: 100-98%, 90-98%, 85-90% and 80-85% bp similarity to assign species, genus, family and order respectively. All the sequences with an assignment similarity lower than 80% were discarded from the analyses.

The second metabarcoding workflow was based on the SWARM clustering algorithm that groups multiple variants of sequences into OTUs (Operational Taxonomic Units, (12) Mahe et al 2014, see details in Supplementary materials).

The SWARM clustering workflow was used to investigate the taxa present in the samples but not revealed by the taxonomic assignment process because of gaps in the EMBL database. The number of taxa assigned

in each family was corrected to avoid taxonomical redundancy assignment. For instance, the combined assignments to the genus *Zanclus* and the species *Zanclus cornutus* were considered as one taxa as potential PCR error may have produced two different assignment levels from the same sequence. These corrected numbers of taxa were then compared to the number of OTUs from the SWARM workflow in each family to evaluate the magnitude of the diversity missed by the direct assignment method. In the SWARM workflow, a family level assignment was performed as well to remove the taxa that were not fish from nonspecific amplifications and investigate the intra family diversity.

Statistical analyses

To evaluate the number of taxa/OTUs present in the study area, a multimodel approach was implemented to fit asymptotes on the species and OTU accumulation curves. This approach considered 5 different accumulation models (Lomolino, Michaelis-Menten, Gompertz, asymptotic regression and logistic curve) and weighted them using the Akaike Information Criterion (AIC, (34) Aho et al 2014). For each curve, the accumulation model with the lowest AIC was selected. Accumulation curves and associated asymptotes were generated using the vegan R package. To estimate the sampling effort required to achieve a given proportion of asymptotes, we considered the model selected for accumulation curves. Then, we extracted the predicted number of samples producing a number of taxa/OTUs that outreached 90% and 95% of the asymptotes.

Results

High heterogeneity of fish species detection among families

A total of 299,479,007 reads (length $>$ 10 bp) were produced over the 92 eDNA samples corresponding to 14,423 unique sequences with a mean of 307 unique sequences per sample (+- 134 SD). In a conservative approach, stringent bioinformatic filters retained 9,345 unique sequences so 65% of the total. These 9,345 unique sequences were then assigned to different taxonomic levels using the following genetic similarity thresholds: 100-98% for species, 90-98% for genus, 85-90% for family and 80-85% for order. This set of thresholds retained 7,389 unique sequences resulting in 678 taxonomic assignments (Suppl. Table S2).

A total of 310 species were detected, including 211 coastal fish species present in the checklist of the Bird's Head Peninsula and 99 fish species present in other regions but absent from this checklist (Fig. 2a). Conversely, 183 sequenced fish species which are present in the Bird's Head Peninsula were not detected in our eDNA samples using our stringent filters, representing 53.6% of the sequenced species present in the checklist. Since 75.5% of fish species in the checklist of the Bird's Head Peninsula were not sequenced for the 12S rDNA, the largest part of fish species diversity remained hidden through direct assignment (Suppl. Table S1).

A total of 282 genera and 128 families of fish were detected compared to the regional checklist of 508 genera and 112 families out of which 46.1% and 72.3% are sequenced respectively (Suppl. Table S1). The number of fish species per family varied from 1 to 191 in the Bird's Head checklist (Fig. 2b), the richest family being the Gobiidae. Only 12 species of Gobiidae were detected in our 92 samples. Meanwhile, the most represented family in the eDNA samples was the Labridae with 48 species (15.5% of the species found in the samples) out of 136 in the checklist (Fig. 2b).

The percentage of fish species sequenced per family varied between 0 and 100% with a mean of 40.3% (+- 31% SD) in the Bird's Head Peninsula checklist while the percentage of detected species per family varied between 0 and 100% with a mean of 27.1% (+- 30.2% SD) in eDNA samples (Fig. 2b). These two percentages were significantly and strongly related ($p < 0.001$) with the percentage of species sequenced per family explaining 85% of variation in the percentage of detected species per family (Fig. 2c).

High but underestimated diversity of OTUs

Given that the low percentage of fish species sequenced for the 12S in the region is the main limitation to detect taxonomic diversity (Fig. 2c), we used an alternative approach based on unique clusters of genetic sequences called Operational Taxonomic Units (OTUs).

From the 331,839,591 initial reads, 183,546 OTUs were generated using the SWARM clustering algorithm. After a series of post-clustering curation processes, 972 fish OTUs were filtered among which 819 were assigned to a family (Suppl. Table S3). The number of detected OTUs varied from 1 to 54 among fish families (Fig. 3a), the richest families (>50 OTUs) being the Gobiidae, Labridae and Pomacentridae. Overall the number of OTUs was superior to the number of assigned taxa (genus and species) in 64.7% of the families found in the samples (mean $\Delta = 4 \pm 6.7$ SD, Fig. 3a). This richness difference was null in 31.4% of the families and negative in 3.9% of them (Fig. 3a). This difference was notably high in some rich families such as the Gobiidae and Pomacentridae where the number of OTUs was more than 2 times and 1.5 times higher than the number of assigned taxa, respectively. By contrast, only 7 OTUs were produced compared to 11 assigned taxa for the Scombridae so Δ = -4 units or -66.7% of this family richness.

The discrepancy between the two approaches (taxa and OTUs) was not significantly explained by the species richness of the family in the checklist ($R^2 < 0.01$, $p = 0.08$, Fig. 3b) and marginally explained, albeit non-significantly, by the percentage of sequenced fish species within each family in the checklist ($R^2 = 0.09$, $p = 0.05$, Fig. 3c).

On average, the number of OTUs underestimated the total number of coastal fish species in the Bird's Head Peninsula checklist with a mean net difference of 40.2% per family (± 38.8% SD, Fig.3d). For most families this difference was high, reaching the maximum value of 95% for the Pseudochromidae. However, this difference could also be negative with more OTUs detected than species present in the checklist as for the Dasyatidae, Leiognathidae and Orectolobidae for which this difference reached -50%. Overall, the difference was marginally but significantly explained by the species richness of the family in the regional checklist ($R^2 = 0.09$, $p = 0.04$, Fig. 3d), suggesting that the bias is not proportional to the species richness of the family with species-rich families being more underestimated by OTUs than species-poor families.

Prediction of fish species diversity from OTU accumulation curves

Since the two approaches (taxa and OTUs) underestimated the level of taxonomic diversity within fish families with a high uncertainty, we modeled accumulation curves from the diversity of species and OTUs found across our 92 samples. The modeled asymptote of the assigned species reached 429 species, a value very close to the 394 sequenced species present in the Bird's Head peninsula, but 3.7 times lower than the 1,611 species in the regional checklist (Fig. 4a). Meanwhile, the OTU accumulation curve reached an asymptote of 1,531, a value very close to the 1,611 fish species in the Bird's Head.

Applying this method to the 15 fish families which counted more than 10 OTUs and 10 species in the checklist permitted to assess the ability of eDNA-based accumulation curves to predict regional fish richness. For instance, the OTU accumulation curves for the Gobiidae, Labridae and Pomacentridae, the three richest families (51, 54 and 53 OTUs respectively), produced asymptotes and thus predictions of fish diversity much lower than those in the regional checklists with 107.5, 66.1 and 76.2 OTUs, i.e. 47.5%, 81.7% and 69.6% of the checklist richness respectively (Fig. 4b, c, d).

We then tested the ability of the assigned taxa, the OTUs and the OTU accumulation curve approaches to predict fish species richness within families of the regional checklist so the predictive power of linear or proportional relationships. The total number of assigned taxa per family in our samples was a significant but weak predictor of the number of fish species per family in the checklist ($R^2 = 0.60$, $p<0.001$, Fig. 5a) with the richness of some families being largely underestimated (e.g. 87.4% of net difference with the checklist for the Gobiidae, Fig. 5a, d). The number of OTUs per family was a better predictor of the family species richness in the checklist ($R^2 = 0.80$, $p < 0.001$) but left 20% of unexplained variation among families with still a marked underestimation (73.3% of net difference with the checklist for Gobiidae, Fig. 5b, e). Using the asymptotes of OTU accumulation curves, we obtained a high predictive accuracy of $R^2 = 0.92$ ($p < 0.001$) for the species richness within families with less bias for the Gobiidae (43.7% of net difference with the checklist) (Fig. 5c, f).

In addition, we observed that the net difference between the number of assigned taxa per family and the number of species per fish family in the checklist is not related to the number of species of the families

(Fig. 5d) suggesting an absence of systematic bias towards the underestimation of species-rich families. By contrast, the net difference between the number of OTUs per fish family and the number of species per family in the checklist significantly increased ($R^2 = 0.35$, $p = 0.02$) with the number of species per family (Fig. 5e). This bias towards the underestimation of species richness within species-rich families is nonetheless avoided when using the asymptotes of OTU accumulation curves ($p = 0.24$, Fig. 5f). Thus, asymptotes of OTU accumulation curves are most accurate and least biased eDNA-based predictors of fish species diversity within families in this marine biodiversity hotspot.

Sampling efforts necessary to achieve regional fish diversity inventory

Not only the OTU accumulation curves and their asymptotes provide diversity estimates, they also provide crucial insights into the sampling effort needed to achieve a more complete census. Here, using the asymptote on the OTU accumulation curve for all fish species (Fig. 4a), we found that our 92 cumulated samples (representing 0.2 m^3) achieved up to 63.5% of the potential fish OTU diversity in the Bird's Head Peninsula (Fig. 6). To collect 90% of this regional fish diversity, we should have filtered seawater in 735 samples so 8 times the effort of our sampling campaign, representing an aggregated sampled water volume of 1.5 m^3. This sampling effort would reach 1,883 samples (an aggregated water volume of 3.8 m^3) to collect 95% of the regional fish OTU richness (Fig. 6).

On average across fish families, our sampling effort achieved the detection of 77.1% (± 14.9 SD) of OTUs predicted by the asymptote of the accumulation curve with a variation among families ranging from 42.2% (Muraenidae) and 47.5% (Gobiidae) to 93.9% (Balistidae) (Fig. 6). The sampling effort needed to achieve 90% of the asymptotic number of OTUs in the region varied greatly among families, ranging from 37 samples for Chaetodontidae to 494 samples for Gobiidae, with a mean of 164 samples (± 123 SD). The estimated additional sampling effort to reach 95% from 90% of the OTU richness ranged from 20 more samples (Tetraodontidae) to 593 more samples (Gobiidae).

Discussion

Overcoming incompleteness of genetic reference databases

Environmental DNA metabarcoding has the potential to surpass most classical survey methods to assess biodiversity in both terrestrial and aquatic systems ((35) Deiner et al 2017). Yet, genetic reference databases are often incomplete especially for species-rich ecosystems such as the Coral Triangle, the global marine biodiversity hotspot ((14) Veron et al 2009). For instance, the current completeness of the 12S rDNA online databases for the teleo primer covers only 24.5% of fish species in the Bird's Head Peninsula. Meanwhile this cover reaches 77.3% for the COI (mitochondrial cytochrome c oxidase subunit I) but fish COI primers still perform poorly in comparison to 12S markers ((36) Collins et al 2019).

With around 28% of families, 54% of the genera and 76% of species not sequenced for the 12S rDNA teleo primers region, the largest part of fish diversity in the Bird's Head peninsula remains thus hidden through direct assignment. Additionally, sequences present in the reference online databases may have been collected from individuals not located in the region of interest. This can induce assignment errors due to biogeographical related genetic variation (e.g. (37) Wadrop et al 2016). The lack of sequencing coverage highlights the immense gap to be filled for online databases to be exhaustive, while numerous species still remain to be described ((38) Pinheiro et al 2019). This limitation prevents metabarcoding approaches from characterizing entire fish assemblages through direct species assignment. Yet, the taxa-assignment method reveals the presence of 211 fish species referenced in the checklist of coastal fishes in the Bird's Head peninsula (Fig. 2a). Conversely, 99 assigned species were absent from this checklist. These 99 detections can either be true presences extending the distribution of some species and revisiting the regional checklist or false presences due to wrong assignments or possible contaminations. For instance, the Atlantic salmon (*Salmo salar*), probably a lab kit contaminant, was found in our study and removed from the analyses (see Methods). The high number of species found in the samples but not present in the checklist of the Bird's Head region suggests that inventories of some families are still incomplete. On average 2.5 detected species per family (± 2.6 SD, Fig.2b) are missing in the checklist with a variation between 0 to 14 species (Apogonidae). This

mismatch allows to target future sampling efforts towards families and their habitats to complete the regional checklist.

As an alternative to species assignment, the use of OTUs as species proxy units is an option that has not yet been tested for vertebrates in species-rich ecosystems while currently used when the concept of species is debatable like for fungi or unicellular organisms ((39) Pawlowski et al 2018, (40) Lladó Fernández et al 2019).

Here, using a conservative and stringent bioinformatic pipeline, we show that the diversity of OTUs is a weak and biased estimator of species diversity with species-rich families being strongly underrepresented. To overcome this limitation, we propose to rely on OTU accumulation curves which provide an unbiased estimate of regional fish diversity and fish richness within families. The asymptotes underestimate the regional fish species richness but the bias is highly consistent among families (Figure 5f). We thus propose to extend this method for taxonomic inventories in poorly-sampled ecosystems like the deep sea to estimate the diversity at different taxonomic levels.

Revealing the potential and limitation of eDNA metabarcoding inventories

Fishes are the most diverse group of vertebrates on Earth with varying body sizes, environmental niches and diets. Monitoring fish assemblages in marine biodiversity hotspots like the Coral Triangle is a great challenge particularly for small, rare, cryptobenthic or elusive species. Here we show that the percentage of sequenced species is highly variable among families preventing any robust estimation of species richness. Instead Operational Taxonomic Units have the potential to reveal the presence of a broad range of fish species, i.e. from different lineages and with contrasted life-history traits. For instance, cryptobenthic families have been poorly documented and are often ignored in traditional visual censuses ((7) Brandl et al 2018) while they strongly influence ecosystem functioning ((13) Brandl et al 2019). Similarly, traditional visual censuses often miss highly mobile and elusive species such as sharks ((9) Boussarie et al 2018).

Among the 310 assigned fish species, we detected the presence of small cryptobenthic species such as *Gobiodon histrio* or*Ostorhinchus selas* , a goby and a cardinalfish with a maximum length below 40 mm, respectively. We also detected large pelagic fish such as the dogtooth tuna (*Gymnosarda unicolor*) or the thresher shark (*Alopias pelagicus*) reaching over 2 m and 4 m long, respectively. Flagship species for conservation were also present in our DNA samples such as the over-exploited Napoleon wrasse (*Cheilinus undulatus* , Endangered, IUCN redlist,*www.iucnredlist.org*), the Scalloped hammerhead shark (*Sphyrna lewini* , Endangered) and several shark species being classified as Near Threatened (NT) (*C. brevipinna* , *C. Leucas* , *C. sorrah* , *C. melanopterus* ,*T. obesus*).

Even if not assigned at species-level, OTUs can be defined as distinct entities for which their distribution and temporal variability can be assessed and monitored ((41) Cordier et al 2017). Moreover, the OTUs and their associated sequences can remain in public repositories until they are assigned to a species, subspecies or complex as databases improve ((42) Wangensteen et al 2018). However, the major caveat of using OTUs for diversity inventories is that they cannot be directly considered as species with complete certainty. Species with intra-specific genetic variability can produce two separate OTUs, overestimating species diversity. Conversely, two species phylogenetically close to each other with low genetic variability can be grouped into a single OTU, thus underestimating species diversity. The accuracy of diversity inventories using eDNA metabarcoding is thus directly based on the taxonomic resolution of the barcode used and genetic variability among families but also the number of samples.

Here we also reveal the gap of biodiversity that remains to be detected using OTU accumulation curves. The effort can be massive for some families (Fig. 6) and more ambitious eDNA sampling campaigns should be on the agenda in species-rich regions like the Coral Triangle. OTU accumulation curves can also serve to evaluate the efficiency of a sampling method (e.g. punctual filtration, transect filtration), the sampled area or the diversity of habitats that are required (e.g. depth, complexity, distance from the seafloor) and their location (e.g. proximity of reefs, hotspots) especially when targeting rare, elusive, highly mobile or cryptobenthic families of fish.

The contrasts between assigned taxa diversity, OTU diversity and OTU asymptote diversity show that the detectability varies strongly among fish families. These contrasts can be related to the ecology of the species but also to the state of the retrieved DNA fragments (intra or extracellular), their sources (e.g. gametes, larvae, feces), their release rate (different and their diffusion in the water column (limited or wide). For instance, a benthic fish species such as gobies with a small movement range would release DNA fragments through skin and feces on a small area. However, such species could release a massive number of gametes carried through the water column ((13) Brandl et al. 2019) so may appear highly detectable during breeding season. Further comparative works are urgently needed between visual, camera and eDNA metabarcoding surveys to better estimate the level of detectability of each species or family in order to provide reliable biodiversity assessments. For instance, coupling eDNA metabarcoding and video surveillance allows the detection of eighty-two fish genera from 13 orders on reefs and seagrass with only 24 genera in common ((43) Stat et al 2019). Investigating biodiversity should also consider its multiple components including functional and phylogenetic diversity that are key for reef ecosystem functioning ((44) Duffy et al 2016). Associating OTUs to species might allow to fill this gap but it will require massive sampling and sequencing efforts.

Acknowledgments

General: We especially thank the Indonesian Institute of Sciences (LIPI) for promoting our collaboration and the Sorong Polytechnic of Marine and Fisheries (Politeknik KP Sorong, West Papua) for providing the vessel Airaha 02 that we used in this campaign. We are grateful to the crew of the Aihara 02 for assisting us during the operations and SPYGEN staff for the technical support in the laboratory.

Funding: Fieldwork and laboratory activities were supported by the Lengguru 2017 Project (www.lengguru.org), conducted by the French National Research Institute for Sustainable Development (IRD), the Indonesian Institute of Sciences (LIPI) with the Research Center for Oceanography (RCO, the Politeknik KP Sorong), the University of Papua (UNIPA) with the help of the Institut Français in Indonesia (IFI) and with corporate sponsorship from the Total Foundation and TIPCO company. The sequencing was funded by the "Explorations de Monaco".

Author contributions: J.B.J, I.B.V., K., L.P., D.M. and R.H. designed research; J.B.J and R.H. design the specific research methods of data collection and the sampling strategy; J.B.J, R.S.U., K., and R.H. collected samples and data; T.D. coordinated the biomolecular analyses; J.B.J., R.S.U. and V.M. performed the bioinformatics analyses; J.B.J, R.S.U., V.M., T.D., L.P., D.M. and R.H. defined sequencing strategy, analyzed and interpreted data; J.B.J wrote the initial draft and designed the figures; J.B.J, R.S.U., V.M., I.B.V., Y.H.S., K., T.D., L.P., D.M. and R.H. wrote the paper and approved the final draft; and L.P., D.M. and R.H. acquired funding to conduct the study.

Competing interests: The authors declare no competing interests.

Data and materials availability: The sequencing run that supports the findings of this study will be available in Dryad digital repository and the metabarcoding pipelines available in GitLab will be stored in a long-term open access archive following paper acceptance.

References

1. Costello, M.J. & Chaudhary, C. (2017) Marine biodiversity, biogeography, deep-sea, and conservation. *Current Biology*,**27**: R511-R527. DOI: 10.1016/j.cub.2017.04.060.
2. Barlow, J., França, F., Gardner, T.A., Hicks, C.C., Lennox, G.D., Berenguer, E., Castello, L., Economo, E.P., Ferreira, J., Guénard, B., Gontijo Leal, C., Isaac, V., Lees, A.C., Parr, C.L., Wilson, S.K., Young, P.J. & Graham, N.A.J. (2018) The future of hyperdiverse tropical ecosystems. *Nature*, **559**: 517–526. DOI: 10.1038/s41586-018-0301-1.
3. Lees, A.C. & Pimm, S.L. (2015) Species, extinct before we know them. *Current Biology*, **5**: R177-R180. DOI: 10.1016/j.cub.2014.12.017.
4. Díaz, S., Pascual, U., Stenseke, M., Martín-lópez, B., Watson, R.T., Molnár, Z., Hill, R., Chan, K.M.A. et al (2018) Assessing nature's contributions to people. *Science*, **359**: 270-272. DOI:

10.1126/science.aap8826.
5. Duffy, J.E., Godwyn, C.M. & Cardinale, B.J. (2017) Biodiversity effects in the wild are common and as strong as key drivers of productivity. *Nature*, **0**: 1-4. DOI: 10.1038/nature23886.
6. Juhel, J.B., Vigliola, L., Wantiez, L., Letessier, T.B., Meeuwig, J.J. & Mouillot, D. (2019) Isolation and no-entry marine reserves mitigate anthropogenic impacts on grey reef shark behavior. *Scientific reports*, **9**: 2897. DOI: 10.1038/s41598-018-37145-x.
7. Brandl, S.J., Goatley, C.H.R., Bellwood, D.R. & Tornabene, L., (2018) The hidden half: ecology and evolution of cryptobenthic fishes on coral reefs. *Biological Reviews*, **93**: 1846-1873. DOI: 10.1111/brv.124233.
8. Garlapati, D., Charankumar, B., Ramu, K., Madeswaran, P. & Ramana Murthy, M.V. (2019) A review on the applications and recent advances in environmental DNA (eDNA) metagenomics. *Reviews in Environmental Science and Bio/Technology*. **18**: 389. DOI: 10.1007/s11157-019-09501-4.
9. Boussarie, G., Bakker, J., Wangensteen, O.S., Mariani, S., Bonin, L., Juhel, J.-B., Kiszka, J.J., Kulbicki, M., Manel, S., Robbins, W.D., Vigliola, L. & Mouillot, D. (2018) Environmental DNA illuminates the dark diversity of sharks. *Science Advances*, **4**: eaap9661. DOI: 10.1126/sciadv.aap9661.
10. Fukumoto, S., Ushimaru, A. & Minamoto, T. (2015) A basin-scale application of environmental DNA assessment for rare endemic species and closely related exotic species in rivers: a case study of giant salamanders in Japan. *Journal of Applied Ecology*, **52**: 358-365. DOI: 10.1111/1365-2664.12392.
11. Ruppert, K.M., Kline, R.J. & Rahman, Md S. (2019) Past, present, and future of environmental DNA (eDNA) metabarcoding: a systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation*, **17**: e00547. DOI:*10.1016/j.gecco.2019.e00547*.
12. Mahé, F., Rognes, T., Quince, C., de Vargas, C. & Dunthorn, M. (2014) Swarm: robust and fast clustering method for amplicon-based studies.*PeerJ*, **2**: e593. DOI: 10.7717/peerj.593.
13. Brandl, S.J., Rasher, D.B., Côté, I.M., Casey, J.M., Darling, E.S., Lefcheck, J.S. & Duffy, J.E. (2019) Coral reef ecosystem functioning: eight core processes and the role of biodiversity. *Frontiers in Ecology and the Environment*. **17**:445-454. DOI: 10.1002/fee.2088.
14. Veron, J.E.N., Devantier, L.M., Turak, E., Green, A.L., Kininmonth, S., Stafford-Smith, M. & Peterson, N. (2009) Delineating the Coral Triangle. *Galaxea, Journal of Coral Reef Studies*, **11**: 91-100. DOI: 10.3755/galaxea.11.91.
15. Allen, G.R. & Erdmann, M.V. (2012) Reef fishes of the East Indies. Volumes I-III. Tropical Reef Research, Perth, Australia. ISBN: 978-0-9872600-0-0. 1,292 p.
16. Mangubhai, S., Erdman, M.V., Wilson, J.R., Huffard, C.L., Ballamu, F., Hidayat, N.I., Hitipeuw, C., Lazuardi, M.E., Muhajir, Pada, D., Purba, G., Rotinsulu, C., Rumetna, L., Sumolang, K. & Wen, W. (2012) Papuan Bird's Head seascape: Emerging threats and challenges in the global center of marine biodiversity. *Marine Pollution Bulletin*,**64**: 2279-2295. DOI: 10.1016/j.marpolbul.2012.07.024.
17. Kulbicki, M., Parravicini, V., Bellwood, D.R., Arias-Gonzalez, E., Chabanet, P., Floeter, S.R., Friedlander, A., McPherson, J., Myers, R.E., Vigliola, L., Mouillot, D. (2013) Global Biogeography of Reef Fishes: A Hierarchical Quantitative Delineation of Regions. *Plos One*, **8**: e81847. DOI: 10.1371/journal.pone.0081847.
18. Mora, C., Chittaro, P.M., Sale, P.F., Kritzer, J.P. & Ludsin, S.A. (2003) Patterns and processes in reef fish diversity. *Nature*,**421**: 933-936. DOI: 10.1038/nature01393.
19. 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.
20. Leprieur, F., Descombes, P., Gaboriau, T., Cowman, P.F., Parravicini, V., Kulbicki, M., Melian, C.J., de Santana, C.N., Heine, C., Mouillot, D., Bellwood, D.R. & Pellissier, L. (2016) Plate tectonics drive tropical reef biodiversity dynamics. *Nature Communications*,**7**: 11461. DOI: 10.1038/ncomms11461.
21. Gaboriau, T., Albouy, C., Descombes, P., Mouillot, D., Pellissier, L. & Leprieur, F. (2019) Ecological constraints coupled with deep_time habitat dynamics predict the latitudinal diversity gradient in reef fishes. *Proceedings of the Royal Society B*, **286**: 20191506. DOI: 0.1098/rspb.2019.1506.
22. Pelissier, L., Leprieur, F., Parravicini, V., Cowman, P.F. & Kulbicki, M. (2014) Quaternary coral reef

refugia preserved fish diversity. *Science*, **344**: 1016-109. DOI: 10.1126/science.1249853.
23. Exton, D.A., Ahmadia, G.N., Cullen-Unsworth, L.C., Jompa, J., May, D., Rice, J., Simonin, P.W., Unsworth, R.K.F. & SmithD.J. (2019) Artisanal fish fences pose broad and unexpected threats to the tropical coastal seascape. *Nature Communications*, **10**: 2100. DOI: 10.1038/s41467-019-10051-0.
24. Jones, L.A., Mannion, P.D., Farnsworth, A., Valdes, P.J., Kelland, S.-J. & Allison, P.A. (2019) Coupling of palaeontological and neontological reef coral data improves forecasts of biodiversity responses under climatic change. *Royal Society Open Science*,**6**: 182111. DOI: 10.1098/rsos.182111.
25. Ainsworth, C.H., Pitcher, T.J. & Rotinsulu, C. (2008) Evidence of fishery depletions and shifting cognitive baselines in Eastern Indonesia. *Biological Conservation*, **141**: 848-859. DOI: 10.1016/j.biocon.2008.01.006.
26. 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., Peroux, T., Crivelli, A.J., Olivier, A., Acqueberge, M., Le Brun, M., Møller, P.R., Willerslev, E. & Dejean, T. (2016) Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding.*Molecular Ecology*, **25**: 929–942. DOI: 10.1111/mec.13428.
27. Goldberg, C.S., Turner, C.R., Deiner, K., Klymus, K.E., Thomsen, P.F., Murphy, M.A., Spear, S.F., McKee, A., Oyler-McCance, S.J., Cornman, R.S., Laramie, M.B., Mahon, A.R., Lance, R.F., Pilliod, D.S., Strickler, K.M., Waits, L.P., Fremier, A.K., Takahara, T., Herder, J.E. & Taberlet, P. (2016) Critical considerations for the application of environmental DNA methods to detect aquatic species.*Methods in Ecology and Evolution*, **7**: 1299-1307. DOI: 10.1111/2041-210X.12595.
28. Pont, D., Rocle, M., Valentini, A., Civade, R., Jean, P., Maire, A., Roset, N., Schabuss, M., Zornig, H., Dejean, T. (2018) Environmental DNA reveals quantitative patterns of fish biodiversity in large rivers despite its downstream transportation. *Scientific Reports*,**8**: 10361. DOI: 10.1038/s41598-018-28424-8.
29. Ficetola, G.T., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessière, J., Taberlet, P. & Pompanon, F. (2010) An *in silico*approach for the evaluation of DNA barcodes. *BMC Genomics*,**11**: 434. DOI: 10.1186/1471-2164-11-434.
30. Baker, W., van den Broek, Camon, E., Hingamp, P., Sterk, P., Stoesser, G. & Tuli, M.A. (2000) The EMBL nucleotide sequence database.*Nucleic Acids Research*, **28**: 19-23. DOI: 10.1093/nar/gki098.
31. Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P. & Coissac, E. (2016) OBITOOLS: a UNIX-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, **16**: 176-182. DOI: 10.1111/1755-0998.12428.
32. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettan, P.A., McWilliams, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. & Higgins, D.G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**: 2947-2948. DOI: 10.1093/bioinformatics/btm404.
33. Kearse, M., Moir, R., Wilson, A., *Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, Duran, C., Thierer, T., Ashton, B., Meintjes, P. & Drummond, A. (*2012*)* Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data*. Bioinformatics,* **28**: 1647–1649.DOI: 10.1093/bioinformatics/bts19.
34. Aho, K., Derryberry, D. & Peterson, T. (2014) Model selection for ecologists: the worldviews of AIC and BIC. *Ecology*,**95**: 631-636. DOI: 10.1890/13-1452.1.
35. Deiner, K., Bik, H.M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D.M., de Vere, N., Pfendrer, M.E. & Bernatchez, L. (2017) Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, **26**:5872-5895. DOI: 10.1111/mec.14350.
36. Collins, R.A., Bakker, J., Wangensteen, O.S., Soto, A.Z., Corrigan, L., Sims, D.W., Genner, M.J. & Mariani, S. (2019) Non-specific amplification compromises environmental DNA metabarcoding with COI.*Methods in Ecology and Evolution*, **10**: 1985-2001. DOI: 10.1111/2041-210X.13276.
37. Wadrop, E., Hobbs, J.-P., Randall, J.E., DiBattista, J.D., Rocha, L.A., Kosaki, R.K., Berumen, M.L. & Bowen, B.W. (2016) Phylogeography, population structure and evolution of coral-eating butterflyfishes (Family Chaetodontidae, genus *Chaetodon*, subgenus *Corallochaetodon*). *Journal of Biogeography*,**43**: 1116-1129. DOI: 10.1111/jbi.12680.

38. Pinheiro, H.T., Moreau, S., Daly, M. & Rocha, L. A. (2019) Will DNA barcoding meet taxonomic needs? *Science*, **365**: 873–875. DOI: 10.1126/science.aay7174.
39. Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A., . . . , Kahlert, M. (2018). The future of biotic indices in the ecogenomic era: Integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. *Science of the Total Environment*, **637-638**: 1295-1310. DOI: 10.1016/j.scitotenv.2018.05.002.
40. Lladó Fernández, S., Větrovský, T. & Baldrian, P. (2019) The concept of operational taxonomic units revisited: genomes of bacteria that are regarded as closely related are often highly dissimilar. *Folia Microbiologica*, **64**: 19–23. DOI: 10.1007/s12223-018-0627-y.
41. Cordier, T., Esling, P., Lejzerowicz, F., Visco, J., Ouadahi, A., Martins, C., Cedhagen, T. & Pawlowski, J. (2017) Predicting the ecological quality status of marine environments from eDNA metabarcoding data using supervised machine learning.*Environmental Science & Technology*, **51**: 9118-9126. DOI: 10.1021/acs.est.7b01518.
42. Wangensteen, O., Palacín, C., Guardiola, M. & Turon, X. (2018) DNA metabarcoding of littoral hard-bottom communities: high diversity and database gaps revealed by two molecular markers. *PeerJ*,**6**: e4705. DOI: 10.7717/peerj.4705.
43. Stat, M., Jeffrey, J., DiBattista, J.D., Newman, S.J., Bunce, M. & Harvey, E.S. (2018) Combined use of eDNA metabarcoding and video surveillance for the assessment of fish biodiversity.*Conservation Biology*, **33**: 196-205. DOI: 10.1111/cobi.13183.
44. Duffy, J.E., Lelcheck, J.S., Stuart-Smith, R.D., Navarrete, S.A. & Edgar, G.J. (2016) Biodiversity enhances reef fish biomass and resistance to climate change. *Proceedings of the National Academy of Sciences*, **113**: 6230-6235. DOI: 10.1073/pnas.1524465113.

Figures and Tables

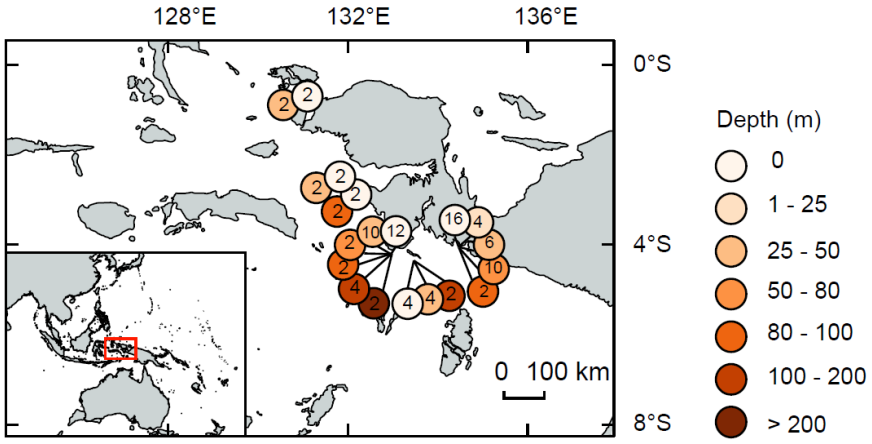

Fig. 1. Map of the Bird's Head region of West Papua showing the location of eDNA samples and their depth. The number inside the circles indicate the number of samples at the location.

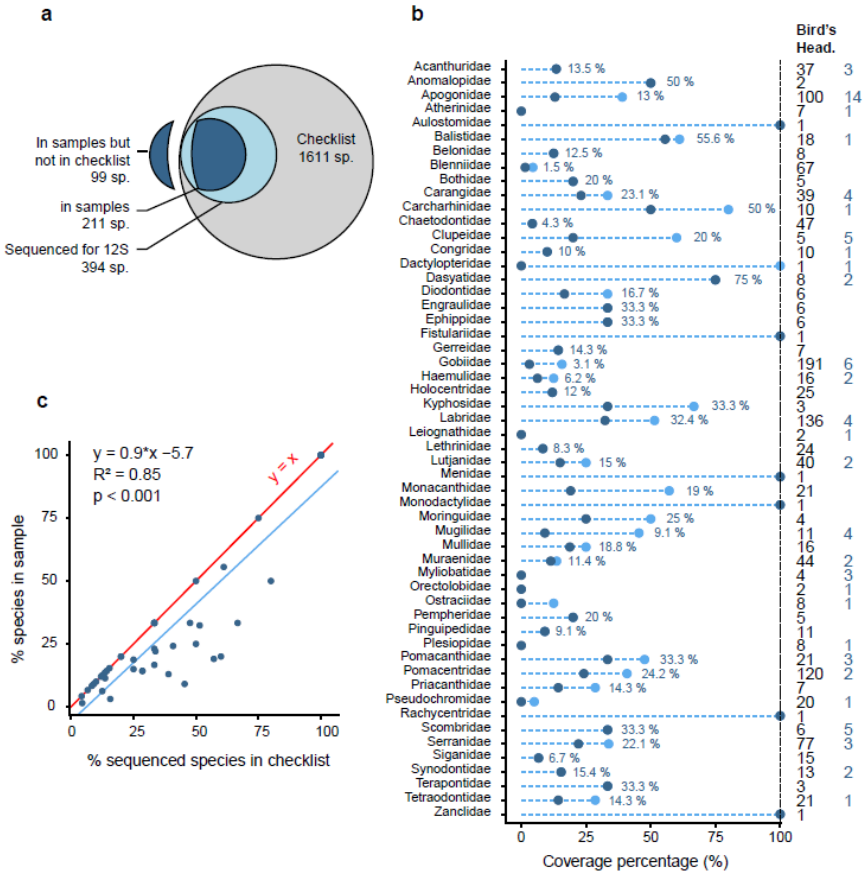

Fig. 2. Number of fish species present in the checklist of the Bird's Head region (grey), sequenced in the European Molecular Biology Laboratory database (EMBL) (light blue) and assigned in the samples (dark blue) (a) ; percentage of species present in the samples (dark blue), sequenced in EMBL (light blue) in each family of species (b) ; percentage of species found in the samples as a function of the percentage of sequenced species in EMBL (c). (b) The percentage values of the species found in the samples compared to the species present in the Bird's Head region are displayed next to the points. The number of species per family in the checklist and the number of the species detected in the samples but not present in the checklist are both on the right of the figure in black and dark blue respectively. Only the sequences assigned to species using ecotag program (similarity >98%) are used in this figure. (c) Each point corresponds to a family of fish.

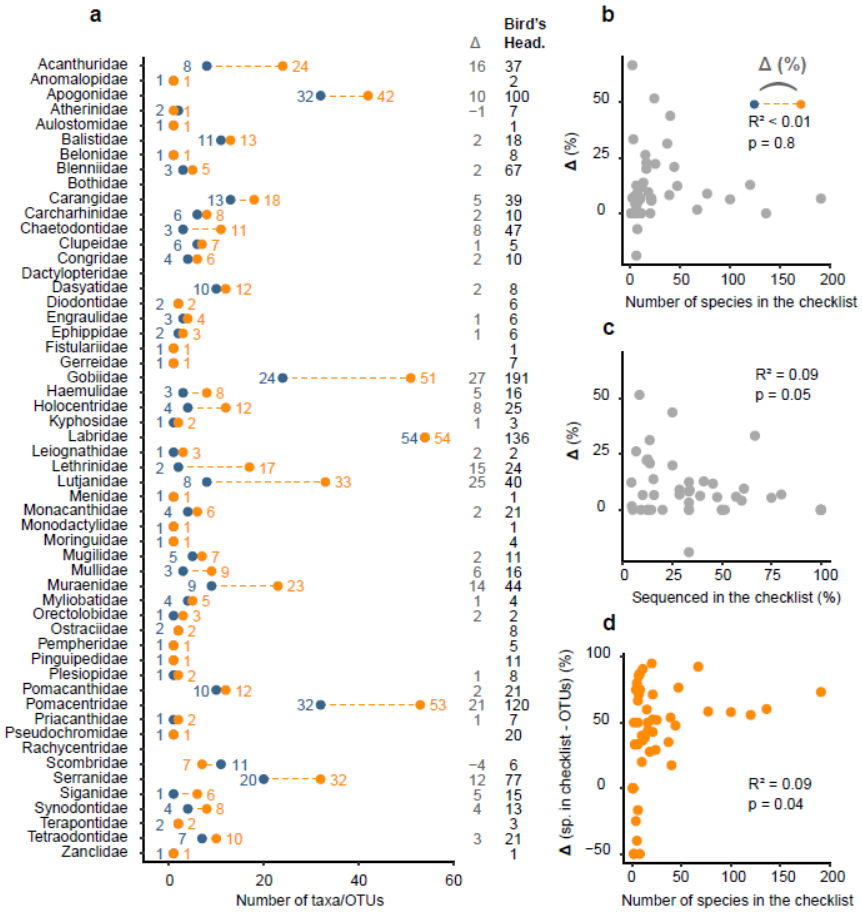

Fig. 3. Number of taxa assigned by the OBITools workflow (blue) and number of OTUs generated by the SWARM workflow (orange) in the families of fish (a) ; distribution of the differences between the two workflows as a function of family richness (b) and as a function of family sequencing coverage (c); distribution of the differences between OTUs and the number of taxa (species and genus) in the checklist as a function of family richness (d). (a)The difference of taxa/OTUs between the two methods (noted Δ) and the number of species in the checklist of the Bird's Head region are on the right of the figure in grey and black respectively. For the OBITools workflow, only the sequences assigned to species and genus using ecotag program (similarity $> 98\%$ and $> 90\%$ respectively) are used in this figure. For the SWARM workflow, only the OTUs curated by LULU and assigned to family (similarity $> 85\%$) are used in this figure.

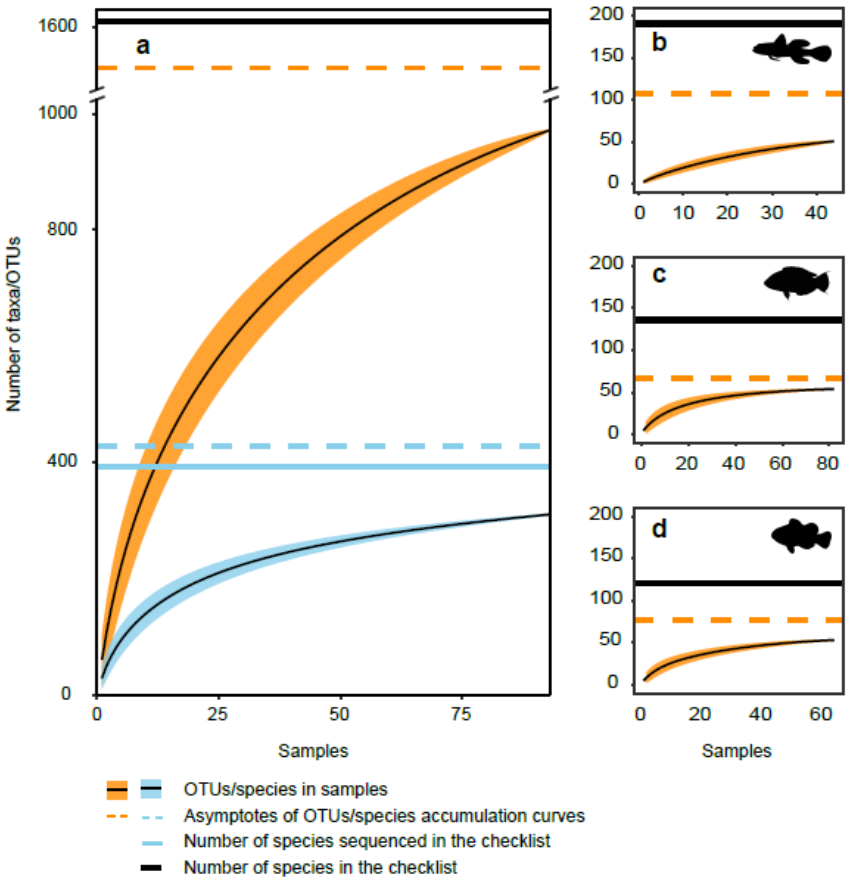

Fig. 4. Accumulation curves of species assigned (blue) and the OTUs (orange) obtained in the whole sampling (a) and within the three most diverse families: Gobiidae (b), Labridae (c) and Pomacentridae (d). The detection of species and OTUs was randomized 100 times and the results were used to generate the standard deviation confidence intervals. The asymptotes were modeled by a multi-model approach weighted by the Akaike Information Criterion (AIC). Fish silhouettes are from phylopic.org (Kent Sorgon & Lily hughes)

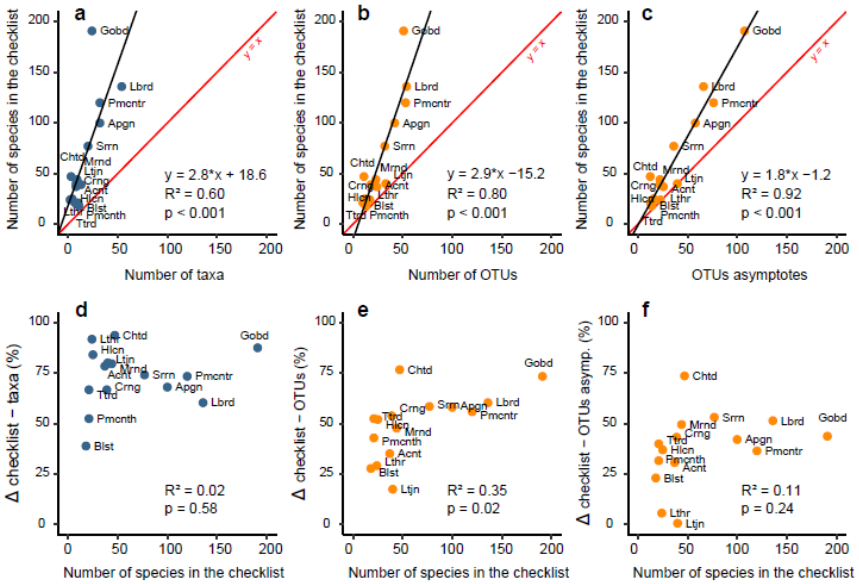

Fig. 5. Linear regression of the diversity of the most diverse families as a function of the number taxa assigned (a), the number of OTUs (b), the asymptotes of the OTUs accumulation curves (c) ; and differences between the number of taxa assigned (d), the number of OTUs (e), the asymptotes of OTUs accumulation curves (f) and the number of species in the checklist as a function of the number of species in the checklist. Only the families with a number of OTU and a number of species in the checklist [?] 10 are presented to provide accurate estimations.

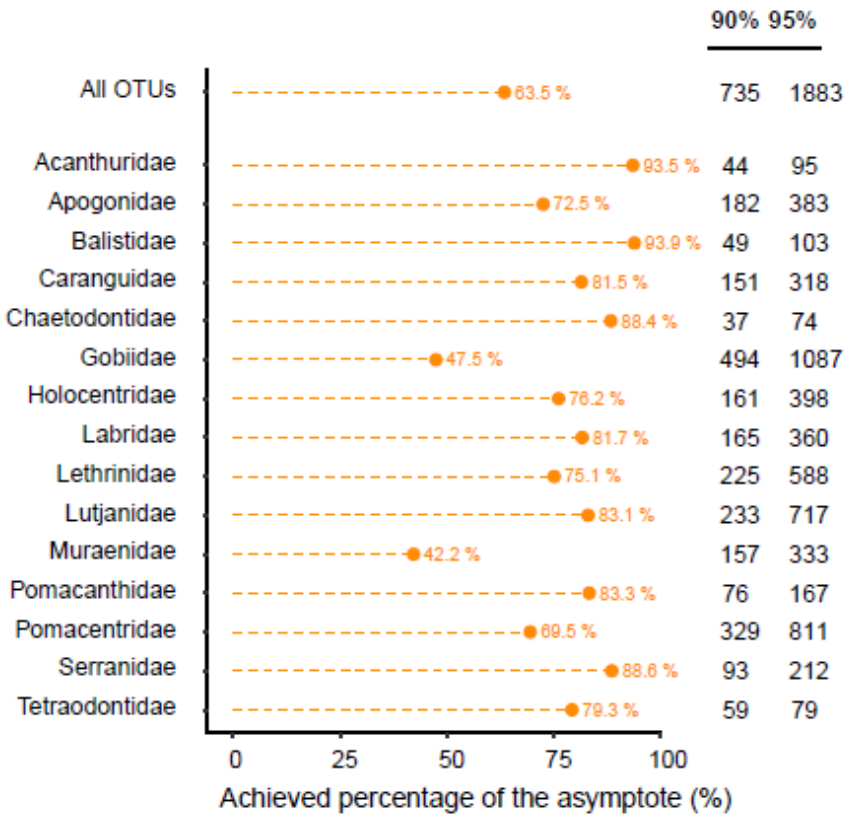

Fig. 6. Percentage of the OTUs diversity covered by the current sampling effort (N = 92) in the families of fish (orange) and the estimated sampling effort required to achieve both 90% and 95% of the diversity. Only the families with a number of OTU and a number of species in the checklist [?] 10 are presented to provide accurate estimations.

Supplementary materials

OBITools bioinformatics workflow

Sequences were aligned using illuminapairedend with a minimum quality score of 40. Then, sequences were demultiplexed using ngsfilter and obisplit. Sequences with ambiguities were removed using obigrep, and filtered using obiclean with a threshold of 0.05 before the taxonomical assignation using ecotag. Further bioinformatic filters were applied to remove PCR or sequencing related errors and non-specific amplifications: i) removal of amplicons with less than 10 reads per PCR, ii) removal of the non-specific amplifications (non-fish orders), iii) removal of the amplicons which size was not comprised in the range of the targeted sequence (50 - 75 bp) and iv) cross-sample contamination cleaning removing amplicons with less than 1/1000 reads per PCR run (i.e. tag jumps, (S1) Schnell et al 2015) and in less than 2 PCR runs from the same sample ((S2) Ficetola et al 2015). An additional check was performed to remove species from contamination using DNA extraction and PCR controls. For instance, 34 to 230,685 reads in 7 sites passed through the bioinformatics filters and were assigned with a similarity of 98.4% and 100% to the Atlantic salmon (*Salmo salar*), a species absent from Indonesia but which DNA was reported as a contaminant in some marine eDNA studies (e.g. (S3) Thomsen et al 2016).

SWARM clustering workflow

Sequences were merged using vsearch. Then, we used cutadapt for demultiplexing and primer trimming

and vsearch to remove sequences containing ambiguities. SWARM was run with a minimum distance of 1 mismatch to make clusters. Once OTUs are generated, the most abundant sequence within each cluster was used for taxonomic assignment using ecotag. The same bioinformatics filters previously described were applied on the OTUs to remove PCR related errors. Then, a post-clustering curation algorithm (LULU, (S4) Frøslev et al 2017) was performed to curate data. This algorithm uses sequence similarity and co-occurrence patterns to detect and remove erroneous OTUs produced by the clustering algorithm. Following author's recommendation, we set the thresholds at 84% sequence similarity and 95% of co-occurrence to identify errors.

1. Schnell, I.B., Bohmann, K. & Gilbert, T.P. (2015) Tag jumps illuminated – reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources*,**15**: 1289–1303. DOI: 10.1111/1755-0998.12402.
2. Ficetola G. F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., Gielly, L., Lopes, C. M., Boyer, F., Pompanon, F., Rayé, G. & Taberlet, P. (2015) Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources*, **15**: 543-556. DOI: 10.1111/1755-0998.12338.
3. Thomsen, P.F., Møller, P.R., Sigsgaard, E.E., Knudsen,S.W., Jørgensen, O.A. & Willerslev, E. (2016) Environmental DNA from seawater samples correlate with trawl catches of subarctic, deepwater fishes.*PLoS One,* **11**: e0165252. DOI: 10.1371/journal.pone.016525.
4. Frøslev, T.G., Kjøller, R., Bruun, H.H., Ejrnæs, R., Brunbjerg, A.K., Pietroni, C. & Hansen, A.J. (2017) Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates.*Nature Communications*, **8**: 1188. DOI: 10.1038/s41467-017-01312-x.

Table S1. Proportion of the fish taxa for which the 12S mitochondrial rDNA is sequenced and referenced in the EMBL database. The checklists of fishes from Indonesia were retrieved from Fishbase (*www.fishbase.de*) and the checklists of fishes from the Bird's Head Peninsula was obtained from Kulbicki et al. 2013.

Habitat	**Taxa level**	**Number in the checklist**	**Sequenced for 12S in EMBL da**
Indonesia	**Indonesia**	**Indonesia**	**Indonesia**
Marine	Family	246	61.8
	Genus	1070	38.7
	Species	3592	18.2
Deep water	Family	109	28.4
	Genus	212	19.8
	Species	316	14.9
Pelagic	Family	22	50
	Genus	44	34.1
	Species	109	18.3
Reef	Family	115	68.7
	Genus	549	44.4
	Species	2068	21.1
Bird's Head Peninsula	**Bird's Head Peninsula**	**Bird's Head Peninsula**	**Bird's Head Peninsula**
Coastal	Family	112	72.3
	Genus	508	46.1
	Species	1611	24.5

Table S2. Summary of the bioinformatic filters applied on the sequences and the associated number of taxa assignments using the ecotag program of the OBITools package.

Bioinformatic filter	Number of reads	Number of sequences	Number of taxa assignations
Sequences with < 10 reads discarded	299,479,007	14,423	765

Bioinformatic filter	Number of reads	Number of sequences	Number of taxa assignations
Assignation to fish taxa only	226,630,600	10,521	719
Sequence length between 50 & 75 bp	226,391,708	10,408	719
PCR errors removed	214,864,059	9,345	714
Taxa assignation (order>80%)	181,034,672	7,389	678

Table S3. Summary of the bioinformatic filters applied on the OTUs using clustering SWARM algorithm and post-clustering LULU curation.

Bioinformatic filter	Number of reads	Number of OTUs
No filter	331,839,591	183,546
OTUs with < 10 reads discarded	331,427,418	4,012
Assignments to non-fish taxa removed	252,276,717	2,737
OTUs length between 50 & 75 bp	251,635,415	2,643
PCR errors removed	228,944,832	1,252
LULU algorithm curation	228,569,519	972
Assignment to families (similarity > 85%)	174,357,646	819