Accumulation curves of environmental DNA sequences predict coastal fish diversity in the coral triangle

Jean-Baptiste Juhel, Rizkie Satriya Utama, Virginie Marquès, Indra Bayu Vimono, Hagi Yulia Sugeha, Kadarusman Kadarusman, Laurent Pouyaud, Tony Dejean, David Mouillot, Régis Hocdé

▶ **To cite this version:**

Jean-Baptiste Juhel, Rizkie Satriya Utama, Virginie Marquès, Indra Bayu Vimono, Hagi Yulia Sugeha, et al.. Accumulation curves of environmental DNA sequences predict coastal fish diversity in the coral triangle. Proceedings of the Royal Society B: Biological Sciences, 2020, 287 (1930), pp.20200248. 10.1098/rspb.2020.0248 . hal-02900375

HAL Id: hal-02900375
https://hal.science/hal-02900375v1

Submitted on 16 Jul 2020

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

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

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

Jean-Baptiste Juhel[1*], Rizkie S. Utama[2], Virginie Marques[1], Indra B. Vimono[2], Hagi Yulia Sugeha[2], Kadarusman[3], Laurent Pouyaud[4], Tony Dejean[5], David Mouillot[1,6§], Régis Hocdé[1§]

[1] UMR 9190 MARBEC, Univ Montpellier, CNRS, Ifremer, IRD, Montpellier Cedex 5, France.

[2] Lembaga Ilmu Pengetahuan Indonesia (LIPI), Pusat Penelitian Oseanografi (P2O), Ancol Timur-Jakarta, Indonesia.

[3] Politeknik Kelautan dan Perikanan Sorong, KKD BP Sumberdaya Genetik, Konservasi dan Domestikasi, Papua Barat 98411, Indonesia.

[4] Institut des Sciences de l'Evolution de Montpellier, Montpellier, France.

[5] SPYGEN, 73370 Le Bourget-du-Lac, France.

[6] ARC Centre of Excellence for Coral Reef Studies, James Cook University, Australia.

[§] Joint last authorship

***Correspondence**: Jean-Baptiste Juhel, Université de Montpellier, 163 rue Auguste Broussonnet 34095 Montpellier Cedex 5, France. email: jeanbaptiste.juhel@gmail.com

Keywords: eDNA metabarcoding, Sequence clustering, OTU, Diversity assessment, Detectability

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 not currently achieved for many taxa and ecosystems. 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). This issue is increasingly important given the accelerating human footprint on Earth. The ongoing worldwide defaunation, characterized by massive population declines, may trigger the local or even global extinction of rare, elusive and cryptic species that are still unknown or poorly documented (2, 3). Such biodiversity losses directly impact ecosystem functioning but also human health, well-being and livelihood (4, 5). This urges scientists to improve the accuracy and extend the breadth of biodiversity inventories and monitoring.

In the marine realm, the detection of species occurrences is particularly challenging due to the vast volume to monitor, the high diversity of habitats, the inaccessibility of some areas (e.g. deep sea) and the behavior of some species (cryptobenthic or elusive) (6, 7). Environmental DNA (eDNA) metabarcoding is an emerging tool that can provide more accurate and wider biodiversity assessments than classical census methods particularly for rare and elusive species (8, 9, 10). This non-invasive method is based on retrieving DNA naturally released by organisms in their environment, amplified by polymerase chain reaction (PCR) and then sequenced to ultimately identify corresponding species (11). However, inventorying and monitoring biodiversity using eDNA metabarcoding requires the completeness of a reference database to accurately assign each sequence to a given species (e.g. 9).

By now, only a minority of fish species are present in online DNA databases for mitochondrial regions targeted by metabarcoding markers, limiting the extent to which species diversity can be revealed by eDNA. This proportion of sequenced species is even lower in species-rich regions and poorly sampled habitats or taxa while the effort to complete genetic reference databases is long and costly. As an alternative, a diversity of Operational Taxonomic Units (OTUs) can be

extracted from eDNA metabarcoding through filtering and clustering techniques (12). Even if environmental genomics approaches have a long tradition of using OTU-based bioindicators (13), the extent to which the diversity of OTUs from a limited number of eDNA samples can reveal or predict the diversity of vertebrate species in a given biodiversity hotspot has not yet been investigated. This is particularly challenging for cryptobenthic fish species that are key for reef ecosystems (14) but usually missed by classical surveys (7). We thus urgently need a regional case study with a wide breadth of fish families and traits to test the potential of OTU-based assessment of biodiversity.

The Bird's Head Peninsula of West Papua (eastern Indonesia) is located in the center of the Coral Triangle which is known to host the world's richest marine biodiversity (15, 16). The current checklist of coastal fishes in the Bird's Head Peninsula identifies 1,611 species belonging to 508 genera and 112 families (15, 17) among which some are still poorly described or under severe threats (18, 19, 20). Providing a blind but accurate assessment of the level and composition of a well-known vertebrate diversity from eDNA OTUs is thus a critical step in conservation, biogeography and ecology, particularly in such biodiversity hotspots.

Here, using eDNA metabarcoding from 92 seawater samples across the Bird's Head Peninsula, we (*i*) assessed the diversity of coastal fish species based on an online reference database for the teleo primers region of the 12S mitochondrial rDNA gene (21), (*ii*) estimated the diversity of fish OTUs based on a custom filtering and clustering bioinformatic pipeline, and (*iii*) tested the capacity of OTU accumulation curves to predict the level and composition of regional fish diversity.

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. S1). 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 (21, 22). Water sample processing included the use of disposable gloves and single-use filtration equipment, and the bleaching (50% bleach) of Niskin water sampler.

DNA extraction, amplification and high-throughput sequencing

The DNA extraction and amplification were performed following the protocol of (23) including 12 separate PCR amplifications per sample (see Supplementary material for more details on the protocol). A teleost-specific 12S mitochondrial rDNA primer (teleo, forward primer-ACACCGCCCGTCACTCT, reverse primer -CTTCCGGTACACTTACCATG, (21)) was used for the amplification of metabarcoding sequences, generating 63 ± 3pb (mean ± SD) long amplicons for all fish species referenced in EMBL database (European Molecular Biology Laboratory, www.ebi.ac.uk, version 138, downloaded on January 2019, (24). Eight negative extraction controls and two negative PCR controls (ultrapure water) were amplified (with 12 replicates as well) and sequenced in parallel to the samples to monitor possible contaminations. The teleo primers were 5'-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.

The purified PCR products were pooled in equal volumes, to achieve a theoretical sequencing depth of 1,000,000 reads per sample. Library preparation and sequencing were performed at Fasteris (Geneva, Switzerland). A total of five libraries were prepared using the MetaFast protocol (Fasteris, https://www.fasteris.com/dna/?q=content/metafast-protocol-amplicon-metagenomic-analysis), a ligation-based PCR-free library preparation. A paired-end sequencing (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 (25) on the EMBL database. 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 Kulbicki et al. 2013 (17).

The amplified DNA sequences from the water samples were processed following two metabarcoding workflows. The first workflow used the OBITools software package (26) based on direct taxonomic assignment of the sequences using the ecotag program (lower common ancestor algorithm) in EMBL database as a reference (see details in Supplementary materials). The ecotag algorithm can sometimes wrongly assign sequences to a given species or genus despite a low-similarity percentage due to the incompleteness of reference database. We thus set the following similarity thresholds, 100-98%, 90-98%, 85-90% and 80-85% bp to assign sequences at the species, genus, family and order level, respectively. All the assignments with a similarity percentage lower than 80% were discarded from the analyses.

We evaluated the database completeness for the marker by running an *in silico* PCR on all fish mitochondrial DNA present in EMBL online database (downloaded the 20th of January 2019). A total of 394 species are sequenced in the Bird's Head region (24.5%, Suppl. table S1).

The second metabarcoding workflow was based on the SWARM clustering algorithm that groups multiple variants of sequences into OTUs (Operational Taxonomic Units (12)). Then, a post-clustering curation algorithm (LULU) was performed to curate data (see details in Supplementary material).

The SWARM clustering workflow was used to investigate the taxa present in the samples but not revealed by the taxonomic assignment process because of gaps in the EMBL database. The number of taxa assigned in each family was corrected to avoid taxonomical redundancy assignment. For instance, the combined assignments to the genus *Zanclus* and the species *Zanclus cornutus* were considered as one taxa as potential PCR error may have produced two

different assignment levels from the same sequence. These corrected numbers of taxa were then compared to the number of OTUs from the SWARM workflow in each family to evaluate the magnitude of the diversity missed by the direct assignment method. In the SWARM workflow, a family level assignment was performed as well to remove the taxa that were not fish from 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, (29)). For each curve, the accumulation model with the lowest AIC was selected. Accumulation curves and associated asymptotes were generated using the vegan R package. To estimate the sampling effort required to achieve a given proportion of asymptotes, we considered the model selected for accumulation curves. Then, we extracted the predicted number of samples producing a number of taxa/OTUs that outreached 90% and 95% of the asymptotes.

Results

High heterogeneity of fish species detection among families

A total of 299,479,007 reads were produced using the OBITools pipeline over the 92 eDNA samples corresponding to 14,423 unique sequences with a mean of 307 unique sequences per sample (± 134 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. 1a). 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. 1b), 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. 1b).

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. 1b). 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. 1c).

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. 1c), we used an alternative approach based on unique clusters of genetic sequences called Operational Taxonomic Units (OTUs).

From the 331,839,591 initial reads, 4,012 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. 2a), 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. 2a). This richness difference was null in 31.4% of the families and

195 negative in 3.9% of them (Fig. 2a). This difference was notably high in some rich families such as
196 the Gobiidae and Pomacentridae where the number of OTUs was more than 2 times and 1.5
197 times higher than the number of assigned taxa, respectively. By contrast, only 7 OTUs were
198 produced compared to 11 assigned taxa for the Scombridae so Δ = -4 units or -66.7% of this
199 family richness.

200 The discrepancy between the two approaches (taxa and OTUs) was not significantly explained
201 neither by the species richness of the family in the checklist ($R^2 < 0.01$, $p = 0.08$, Fig. 2b) nor by
202 the percentage of sequenced fish species within each family in the checklist ($R^2 = 0.09$, $p = 0.05$,
203 Fig. 2c).

204 On average, the number of OTUs underestimated the total number of coastal fish species in the
205 Bird's Head Peninsula checklist with a mean net difference of 40.2% per family (± 38.8% SD,
206 Fig.2d). For most families this difference was high, reaching the maximum value of 95% for the
207 Pseudochromidae. However, this difference could also be negative with more OTUs detected
208 than species present in the checklist as for the Dasyatidae, Leiognathidae and Orectolobidae for
209 which this difference reached -50%. Overall, the difference was marginally but significantly
210 explained by the species richness of the family in the regional checklist ($R^2 = 0.09$, $p = 0.04$, Fig.
211 2d), suggesting that the bias is not proportional to the species richness of the family with
212 species-rich families being more underestimated by OTUs than species-poor families.

213 **Prediction of fish species diversity from OTU accumulation curves**

214 Since the two approaches (taxa and OTUs) underestimated the level of taxonomic diversity
215 within fish families with a high uncertainty, we modeled accumulation curves from the diversity
216 of species and OTUs found across our 92 samples. The modeled asymptote of the assigned
217 species reached 429 species, a value very close to the 394 sequenced species present in the
218 Bird's Head peninsula, but 3.7 times lower than the 1,611 species in the regional checklist (Fig.
219 3a). Meanwhile, the OTU accumulation curve reached an asymptote of 1,531 ; a value close
220 (95%) to the number of fish species (1,611) referenced in the checklist of the Bird's Head
221 Peninsula.

Applying this method to the 15 fish families which counted more than 10 OTUs and 10 species in the checklist permitted to assess the ability of eDNA-based accumulation curves to predict regional fish richness. For instance, the OTU accumulation curves for the Gobiidae, Labridae and Pomacentridae, the three richest families (51, 54 and 53 OTUs respectively), produced asymptotes and thus predictions of fish diversity much lower than those in the regional checklists with 107.5, 66.1 and 76.2 OTUs, i.e. 47.5%, 81.7% and 69.6% of the checklist richness respectively (Fig. 3b, 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. 4a) with the richness of some families being largely underestimated (e.g. 87.4% of net difference with the checklist for the Gobiidae, Fig. 4a, 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. 4b, 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. 4c, 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. 4d) 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. 4e). 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. 4f). 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. 3a), 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. 5). 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. 5).

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. 5). The sampling effort needed to achieve 90% of the asymptotic number of OTUs in the region varied greatly among families, ranging from 37 samples for Chaetodontidae to 494 samples for Gobiidae, with a mean of 164 samples (± 123 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 (30). Yet, genetic reference databases are often incomplete especially for species-rich ecosystems such as the Coral Triangle, the global marine biodiversity hotspot (14). For instance, the current completeness of the 12S rDNA online databases for the teleo primer covers only 24.5% of fish species in the Bird's Head Peninsula. Meanwhile, this cover reaches 77.3% for the COI (mitochondrial cytochrome c oxidase subunit I) but fish COI primers still perform poorly in comparison to 12S markers (31).

With around 28% of families, 54% of the genera and 76% of species not sequenced for the 12S rDNA teleo primers region, the largest part of fish diversity in the Bird's Head peninsula remains thus hidden through direct assignment. Additionally, sequences present in the reference online databases may have been collected from individuals not located in the region of interest. This can induce assignment errors due to biogeographical related genetic variation (e.g. (32)). The lack of sequencing coverage highlights the immense gap to be filled for online databases to be exhaustive, while numerous species still remain to be described (33). This limitation prevents metabarcoding approaches from characterizing entire fish assemblages through direct species assignment. Yet, the taxa-assignment method reveals the presence of 211 fish species referenced in the checklist of coastal fishes in the Bird's Head peninsula (Fig. 1a). 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 large number of species present in the samples but absent from the regional checklist suggests that inventories of some families are still incomplete. On average 2.5 detected species per family (± 2.6 SD, Fig.1b) are absent from the checklist, ranging from 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 (34, 35).

Here, using a conservative and stringent bioinformatic pipeline, we show that the diversity of OTUs is a weak and biased estimator of species diversity with species-rich families being strongly underrepresented. To overcome this limitation, we propose to rely on OTU accumulation curves which provide an unbiased estimate of regional fish diversity and fish richness within families. The asymptotes underestimate the regional fish species richness but the bias is highly consistent among families (Fig. 4f). 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) while they strongly influence ecosystem functioning (13). Similarly, traditional visual censuses often miss highly mobile and elusive species such as sharks (9).

Among the 310 assigned fish species, we detected the presence of small cryptobenthic species such as *Gobiodon histrio* or *Ostorhinchus selas*, a goby and a cardinalfish with a maximum length below 40 mm, respectively. We also detected large pelagic fish such as the dogtooth tuna (*Gymnosarda unicolor*) or the thresher shark (*Alopias pelagicus*) reaching over 2 m and 4 m long, respectively. Flagship species for conservation were also present in our DNA samples such as the over-exploited Napoleon wrasse (*Cheilinus undulatus*, Endangered, IUCN 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 (36). Moreover, the OTUs and their associated sequences can remain in public repositories until they are assigned to a species, subspecies or complex as databases improve (37). However, the major caveat of using OTUs for diversity inventories is that they cannot be directly considered as species with complete certainty. Species with intra-specific genetic variability can produce two separate OTUs, overestimating species diversity. Conversely, two species phylogenetically close to each

other with low genetic variability can be grouped into a single OTU, thus underestimating species diversity. The accuracy of diversity inventories using eDNA metabarcoding is thus directly based on the taxonomic resolution of the barcode used and genetic variability among families but also the number of samples.

Here we also reveal the gap of biodiversity that remains to be detected using OTU accumulation curves. The effort can be massive for some families (Fig. 5) and more ambitious eDNA sampling campaigns should be on the agenda in species-rich regions like the Coral Triangle. OTU accumulation curves can also serve to evaluate the efficiency of a sampling method (e.g. punctual filtration, transect filtration), the sampled area or the diversity of habitats that are required (e.g. depth, complexity, distance from the seafloor) and their location (e.g. proximity of reefs, hotspots) especially when targeting rare, elusive, highly mobile or cryptobenthic families of fish.

The contrasts between assigned taxa diversity, OTU diversity and OTU asymptote diversity show that the detectability varies strongly among fish families. These contrasts can be related to the ecology of the species but also to the state of the retrieved DNA fragments (intra or extracellular), their sources (e.g. gametes, larvae, feces), their release rate, their diffusion in the water column (limited or wide) and their transportation (38). 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) 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 (39). Investigating biodiversity should also consider its multiple components including functional and phylogenetic diversity that are key for reef ecosystem functioning (40). Associating OTUs to species might allow to fill this gap but it will require massive sampling and sequencing efforts.

Acknowledgments

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

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

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. et al. (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. 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. *Biol Rev*, **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. *Rev Environ Sci Bio*. **18**: 389. DOI: 10.1007/s11157-019-09501-4.

9. Boussarie, G. et al. (2018) Environmental DNA illuminates the dark diversity of sharks. *Sci Adv*, **4**: eaap9661. DOI: 10.1126/sciadv.aap9661.

10. Fukumoto, S., Ushimaru, A. & Minamoto, T. (2015) A basin-scale application of environmental DNA assessment for rare endemic species and closely related exotic species in rivers: a case study of giant salamanders in Japan. *J Appl Ecol*, **52**: 358-365. DOI: 10.1111/1365-2664.12392.

11. Ruppert, 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. *Glob Ecol Conserv*, **17**: e00547. DOI: 10.1016/j.gecco.2019.e00547.

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

13. Cordier, T (2019) Multi-marker eDNA metabarcoding survey to assess the environmental impact of three offshore gas platforms in the North Adriatic Sea (Italy). *Mar. Environ. Res*., **146**: 24-34. DOI: 10.1016/j.marenvres.2018.12.009.

14. Brandl, 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. *Front Ecol Environ*. **17**:445-454. DOI: 10.1002/fee.2088.

15. 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, JCRS*, **11**: 91-100. DOI: 10.3755/galaxea.11.91.

16. 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.

17. Kulbicki, M. et al. (2013) Global Biogeography of Reef Fishes: A Hierarchical Quantitative Delineation of Regions. *Plos One*, **8**: e81847. DOI: 10.1371/journal.pone.0081847.

18. Exton, D.A. et al. (2019) Artisanal fish fences pose broad and unexpected threats to the tropical coastal seascape. *Nat Commun*, **10**: 2100. DOI: 10.1038/s41467-019-10051-0.

19. Jones, 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. *Roy Soc Open Sci*, **6**: 182111. DOI: 10.1098/rsos.182111.

20. Ainsworth, C.H., Pitcher, T.J. & Rotinsulu, C. (2008) Evidence of fishery depletions and shifting cognitive baselines in Eastern Indonesia. *Biol Conserv*, **141**: 848-859. DOI: 10.1016/j.biocon.2008.01.006.

21. Valentini, A. et al. (2016) Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Mol Ecol*, **25**: 929–942. DOI: 10.1111/mec.13428.

22. Goldberg, 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 Ecol Evol*, **7**: 1299-1307. DOI: 10.1111/2041-210X.12595.

23. Pont, D. et al. (2018) Environmental DNA reveals quantitative patterns of fish biodiversity in large rivers despite its downstream transportation. *Sci Rep*, **8**: 10361. DOI: 10.1038/s41598-018-28424-8.

24. Baker, W., van den Broek, Camon, E., Hingamp, P., Sterk, P., Stoesser, G. & Tuli, M.A. (2000) The EMBL nucleotide sequence database. *Nucleic Acids Res*, **28**: 19-23. DOI: 10.1093/nar/gki098.

25. 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.

26. Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P. & Coissac, E. (2016) OBITOOLS: a UNIX-inspired software package for DNA metabarcoding. *Mol Ecol Res*, **16**: 176-182. DOI: 10.1111/1755-0998.12428.

27. Larkin, M.A. et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**: 2947-2948. DOI: 10.1093/bioinformatics/btm404.

28. Kearse, M. et al. (2012) Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics,* **28**: 1647–1649. DOI: 10.1093/bioinformatics/bts19.

29. Aho, K., Derryberry, D. & Peterson, T. (2014) Model selection for ecologists: the worldviews of AIC and BIC. *Ecology*, **95**: 631-636. DOI: 10.1890/13-1452.1.

30. Deiner, K. et al. (2017) Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Mol Ecol*, **26**:5872-5895. DOI: 10.1111/mec.14350.

31. Collins, 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 Ecol Evol*, **10**: 1985-2001. DOI: 10.1111/2041-210X.13276.

32. 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*). *J Biogeogr*, **43**: 1116-1129. DOI: 10.1111/jbi.12680.

33. 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.

34. Pawlowski, J. et al. (2018). The future of biotic indices in the ecogenomic era: Integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. *Sci Total Environ*, **637-638**: 1295-1310. DOI: 10.1016/j.scitotenv.2018.05.002.

35. Lladó 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 Microbiol*, **64**: 19–23. DOI: 10.1007/s12223-018-0627-y.

36. Cordier, T., Esling, P., Lejzerowicz, F., Visco, J., Ouadahi, A., Martins, C., Cedhagen, T. & Pawlowski, J. (2017) Predicting the ecological quality status of marine environments from eDNA metabarcoding data using supervised machine learning. *Environ Sci Technol*, **51**: 9118-9126. DOI: 10.1021/acs.est.7b01518.

37. Wangensteen, O., Palacín, C., Guardiola, M. & Turon, X. (2018) DNA metabarcoding of littoral hard-bottom communities: high diversity and database gaps revealed by two molecular markers. *PeerJ*, **6**: e4705. DOI: 10.7717/peerj.4705.

38. Harrison, J.B., Sunday, J.M. & Rogers S.M. (2019) Predicting the fate of eDNA in the environment and implications of studying biodiversity. *Proc R Soc B*, **286**: 20191409. DOI: 10.1098/rspb.2019.1409.

39. Stat, M., Jeffrey, J., DiBattista, 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. *Conserv Biol*, **33**: 196-205. DOI: 10.1111/cobi.13183.

40. 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. *P Natl Acad Sci*, **113**: 6230-6235. DOI: 10.1073/pnas.1524465113.

41. Juhel J-B et al. (2020) Data from: Accumulation curves of environmental DNA sequences predict coastal fish diversity in the coral triangle. Dryad Digital Repository. (doi: 10.5061/dryad.t1g1jwt05)

Figure legends

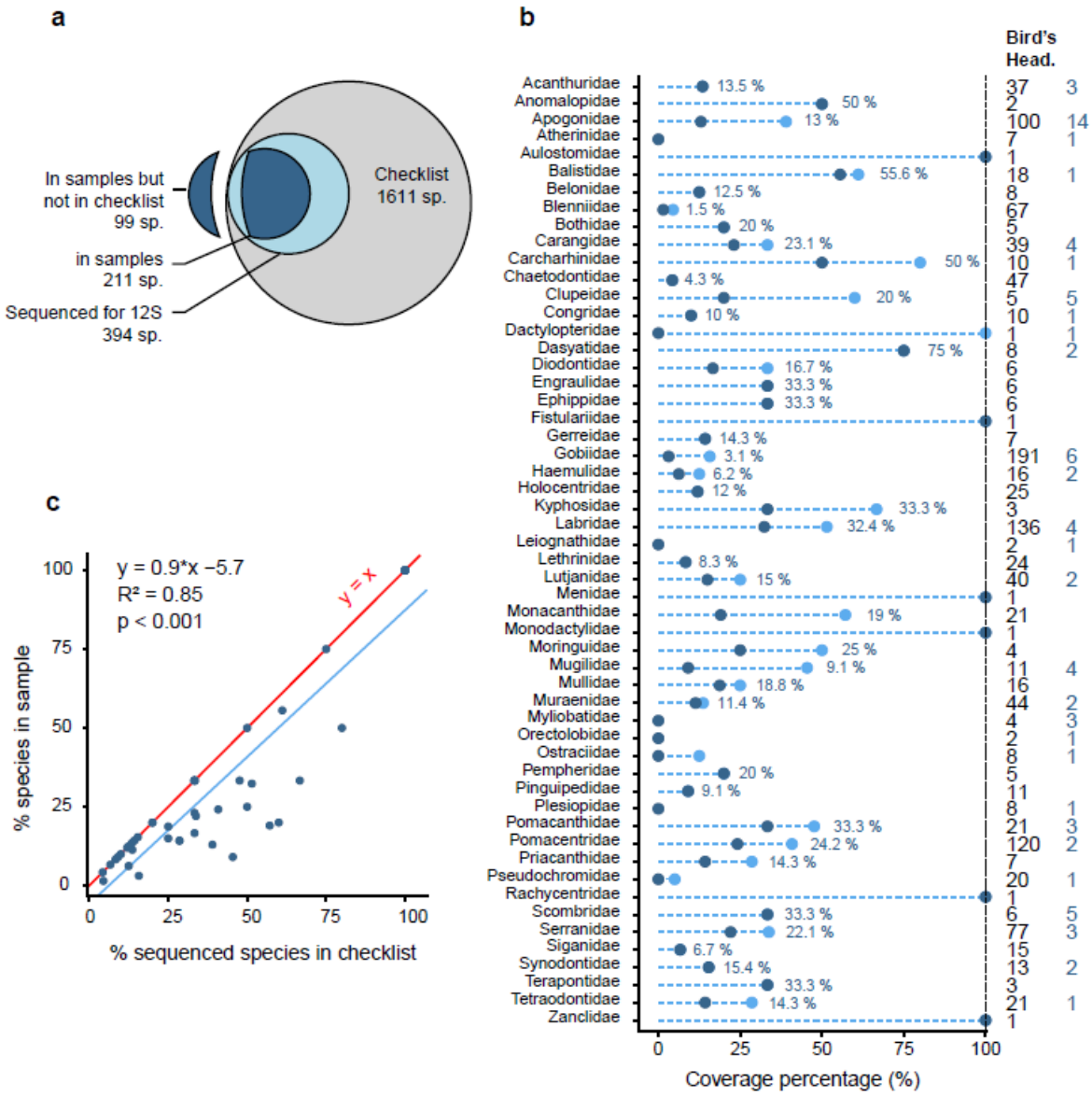


Fig. 1. Number of fish species present in the checklist of the Bird's Head region (grey), sequenced in the European Molecular Biology Laboratory database (EMBL) (light blue) and detected in the eDNA samples (dark blue) (a) ; percentage of species detected in the samples (dark blue), sequenced in EMBL (light blue) in each family of species (b) ; percentage of species detected in the samples as a function of the percentage of sequenced species in EMBL (c). (b) The percentages of the species detected in the eDNA samples compared 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 species detected in the samples but not present in the checklist are both on the right of the figure in black and dark blue, respectively. Only the sequences assigned to species using ecotag program (similarity >98%) are used in this figure. (c) Each point corresponds to a fish family.

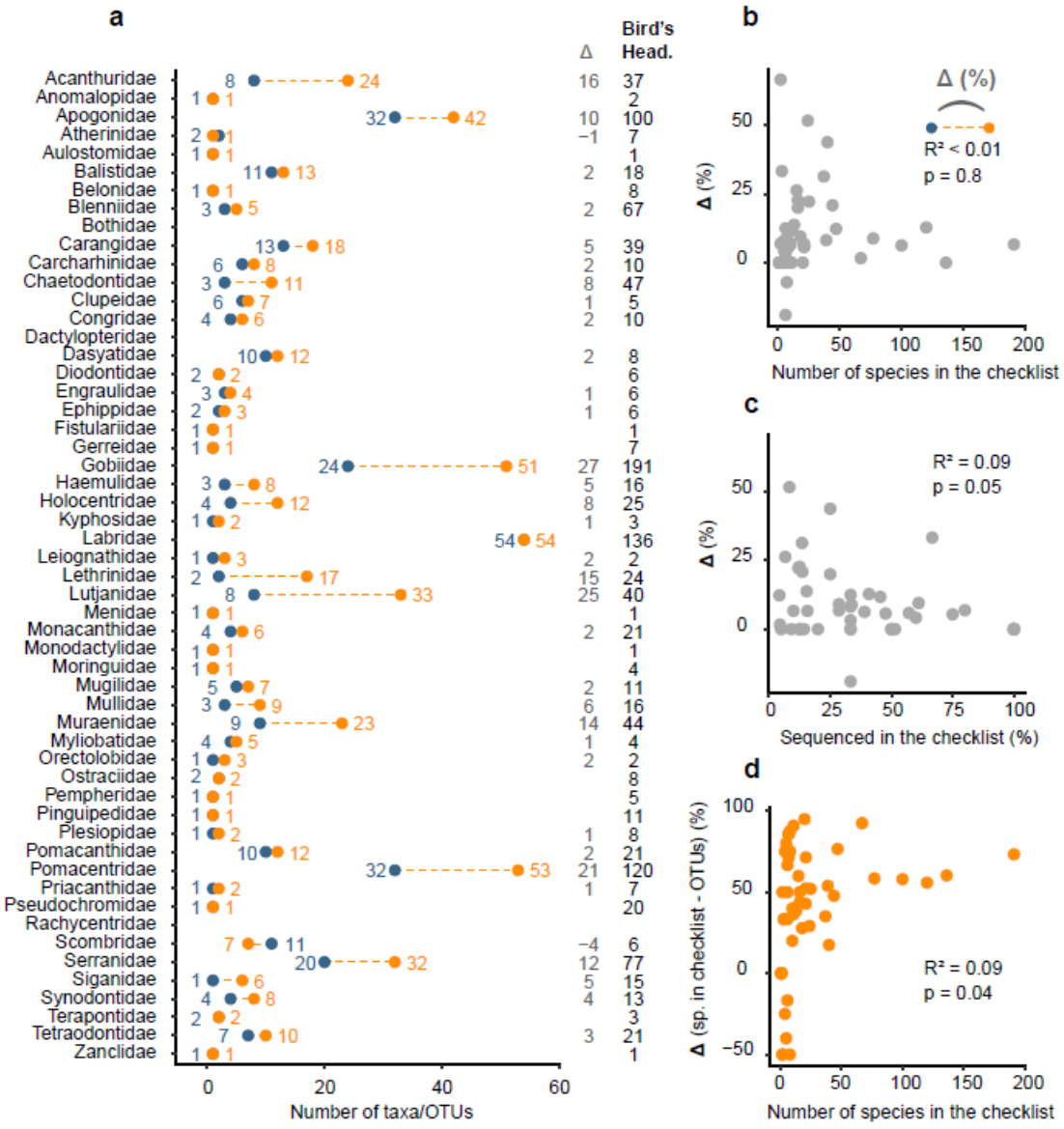


Fig. 2. Number of taxa assigned by the OBITools workflow (blue) and number of OTUs generated by the SWARM workflow (orange) in the different fish families (a) ; distribution of the differences between the two workflows as a function of family richness (b) and as a

function of family sequencing coverage (c); distribution of the differences between OTUs and the number of 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.

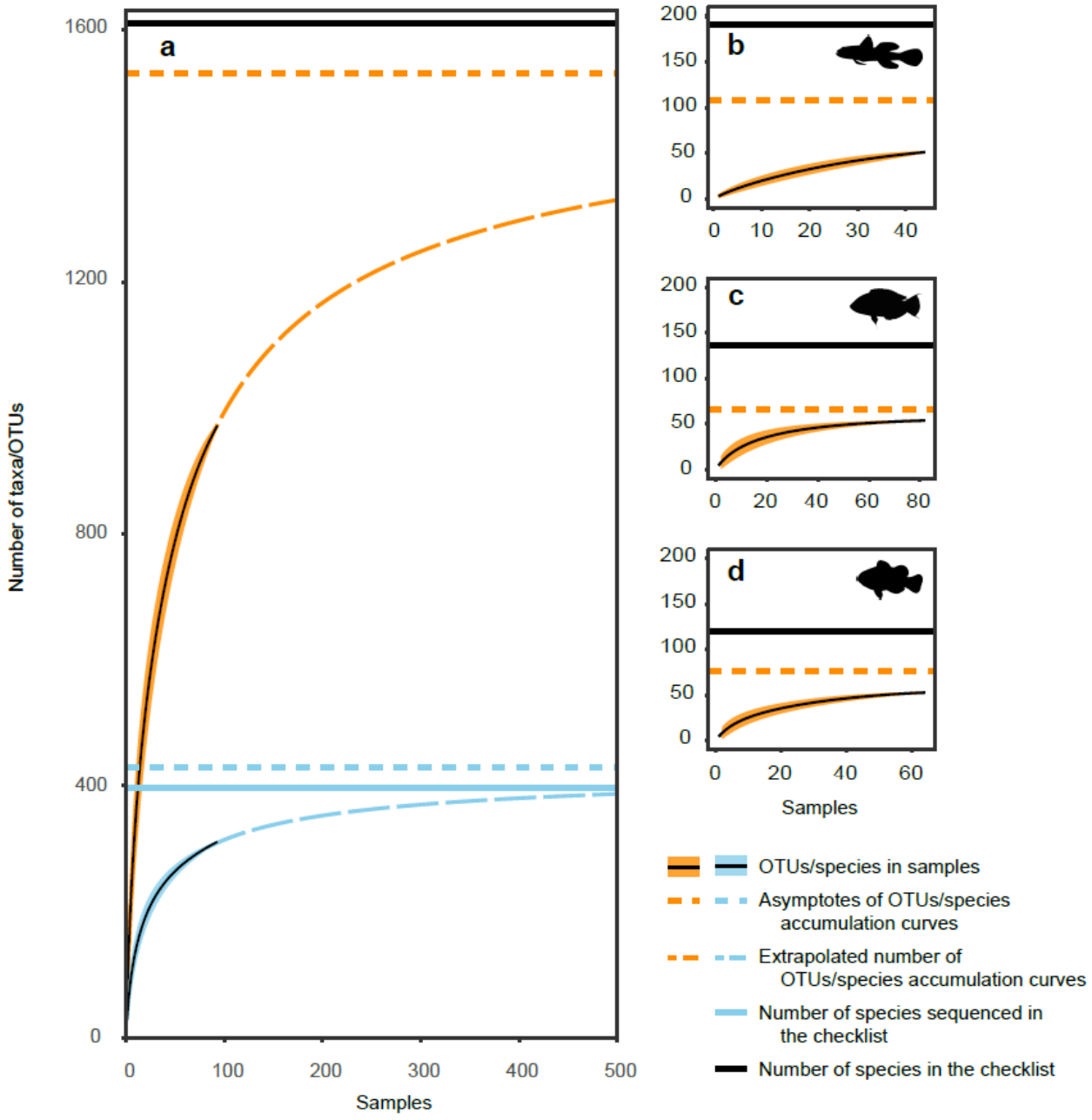


Fig. 3. Accumulation curves of species assigned (blue) and the OTUs (orange) obtained in the whole sampling (a) and within the three most diverse families: Gobiidae (b), Labridae (c) and Pomacentridae (d). The detection of species and OTUs was randomized 100 times and the results were used to generate the confidence intervals. The asymptotes were modeled by a multi-model approach weighted by the Akaike Information Criterion (AIC). Fish silhouettes are from phylopic.org (Kent Sorgon & Lily hughes)

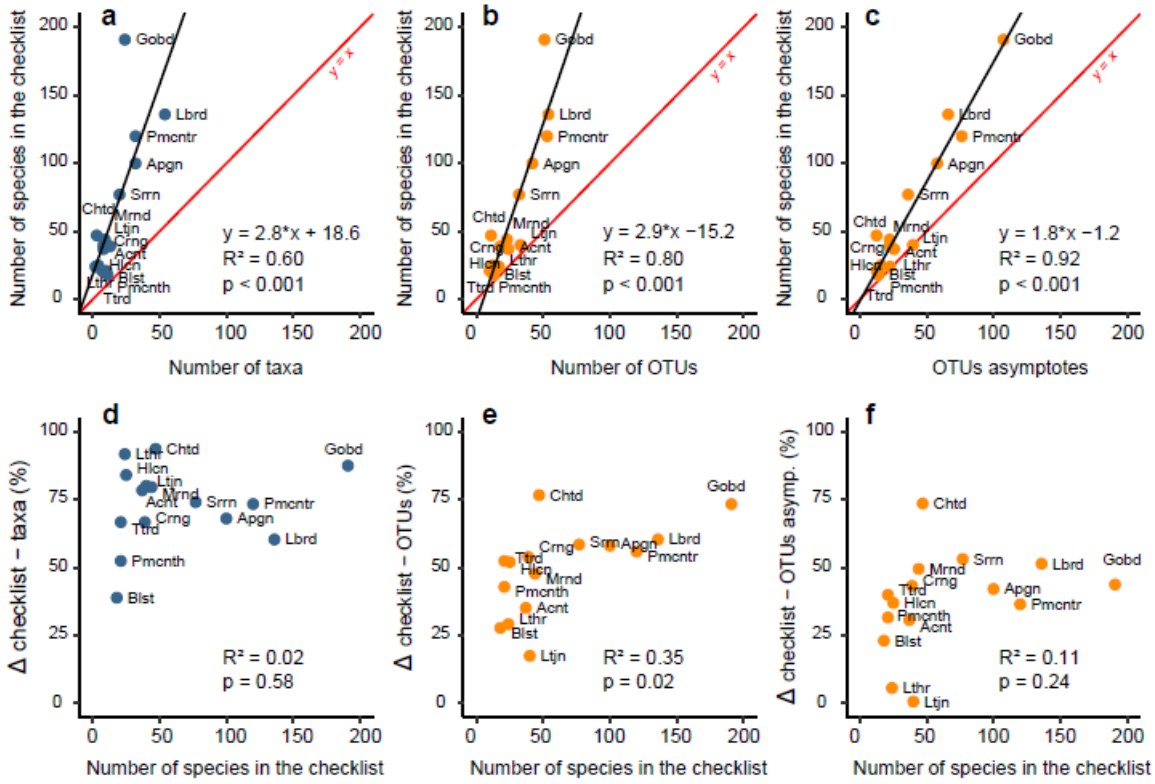
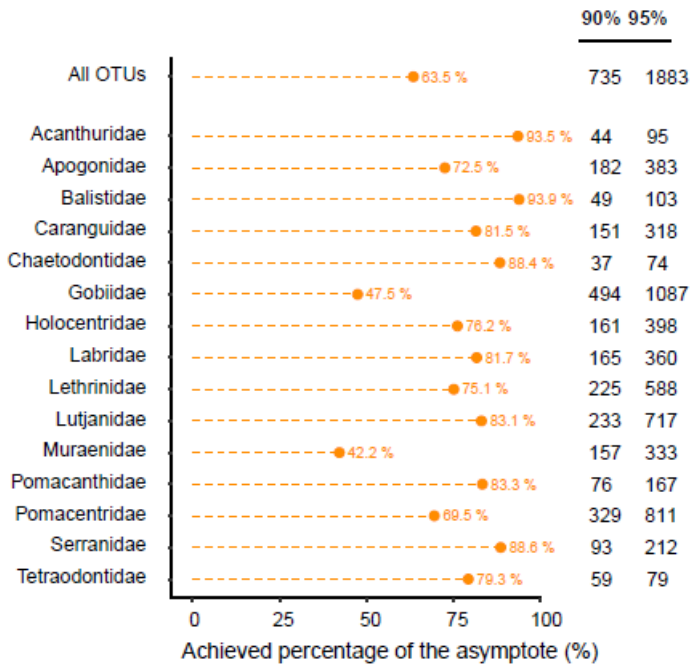


Fig. 4. Linear regression of the diversity of the most diverse families as a function of the number taxa assigned (a), the number of OTUs (b), the asymptotes of the OTUs accumulation curves (c); and differences between the number of taxa assigned (d), the number of OTUs (e), the asymptotes of OTUs accumulation 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.



538

539 **Fig. 5. Percentage of the OTUs diversity covered by the current sampling effort (N = 92) in the**
540 **families of fish (orange) and the estimated sampling effort required to achieve both 90% and**
541 **95% of the diversity.** Only the families with a number of OTU and a number of species in the
542 checklist ≥ 10 are presented to provide accurate estimations.