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

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

¹ MARBEC, Univ. Montpellier, CNRS, Ifremer, IRD, Montpellier, France

² Research Center for Oceanography, Indonesian Institute of Sciences, Jl. Pasir Putih 1, Ancol Timur, Jakarta Utara, Indonesia

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

⁴ Institut des Sciences de l'Evolution de Montpellier, Montpellier, France

⁵ SPYGEN, 73370 Le Bourget-du-Lac, France

⁶ ARC Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, Australia

* Corresponding author : Jean-Baptiste Juhel, email address : jeanbaptiste.juhel@gmail.com

Abstract :

Environmental DNA (eDNA) has the potential to provide more comprehensive biodiversity assessments, particularly for vertebrates in species-rich regions. However, this method requires the completeness of a reference database (i.e. a list of DNA sequences attached to each species), which is not currently achieved for many taxa and ecosystems. As an alternative, a range of operational taxonomic units (OTUs) can be extracted from eDNA metabarcoding. However, the extent to which the diversity of OTUs provided by a limited eDNA sampling effort can predict regional species diversity is unknown. Here, by modelling OTU accumulation curves of eDNA seawater samples across the Coral Triangle, we obtained an asymptote reaching 1531 fish OTUs, while 1611 fish species are recorded in the region. We also accurately predict ($R^2 = 0.92$) the distribution of species richness among fish families from OTU-based asymptotes. Thus, the multi-model framework of OTU accumulation curves extends the use of eDNA metabarcoding in ecology, biogeography and conservation.

Keywords : eDNA metabarcoding, sequence clustering, Operational Taxonomic Unit, diversity assessment, detectability

30 **Introduction**

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

39 In the marine realm, the detection of species occurrences is particularly challenging due to the
40 vast volume to monitor, the high diversity of habitats, the inaccessibility of some areas (e.g.
41 deep sea) and the behavior of some species (cryptobenthic or elusive) (6, 7). Environmental
42 DNA (eDNA) metabarcoding is an emerging tool that can provide more accurate and wider
43 biodiversity assessments than classical census methods particularly for rare and elusive species
44 (8, 9, 10). This non-invasive method is based on retrieving DNA naturally released by organisms
45 in their environment, amplified by polymerase chain reaction (PCR) and then sequenced to
46 ultimately identify corresponding species (11). However, inventorying and monitoring
47 biodiversity using eDNA metabarcoding requires the completeness of a reference database to
48 accurately assign each sequence to a given species (e.g. 9).

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

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

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

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

75 **Methods**

76 **Sampling area and protocol**

77 A total of 92 water samples were collected during October and November 2017 along the south
78 coast of the Bird's Head region of West Papua (500 km) across different habitats but mainly
79 coral reefs (Fig. S1). Samples were collected in DNA-free plastic bags at the surface from a
80 dinghy boat, at depths between 10 – 100m during close circuit rebreather dives, and (iii) at
81 depths between 100 - 300m using Niskin water samplers. A pressure and temperature sensor

82 was coupled to the Niskin bottle to control the sampling depth and characterize the water mass
83 via the vertical temperature profile. For each sample, 2L of seawater were filtered with sterile
84 Sterivex filter capsules (Merck© Millipore; pore size 0.22µm) using disposable sterile syringes.
85 Immediately after, the filter units were filled with lysis conservation buffer (CL1 buffer
86 SPYGEN©) and stored in 50 mL screw-cap tubes at -20°C. A contamination control protocol was
87 followed in both field and laboratory stages (21, 22). Water sample processing included the use
88 of disposable gloves and single-use filtration equipment, and the bleaching (50% bleach) of
89 Niskin water sampler.

90 **DNA extraction, amplification and high-throughput sequencing**

91 The DNA extraction and amplification were performed following the protocol of (23) including
92 12 separate PCR amplifications per sample (see Supplementary material for more details on the
93 protocol). A teleost-specific 12S mitochondrial rDNA primer (teleo, forward primer-
94 ACACCGCCCGTCACTCT, reverse primer -CTCCGGTACACTTACCATG, (21)) was used for the
95 amplification of metabarcoding sequences, generating 63 ± 3 pb (mean \pm SD) long amplicons for
96 all fish species referenced in EMBL database (European Molecular Biology Laboratory,
97 www.ebi.ac.uk, version 138, downloaded on January 2019, (24)). Eight negative extraction
98 controls and two negative PCR controls (ultrapure water) were amplified (with 12 replicates as
99 well) and sequenced in parallel to the samples to monitor possible contaminations. The teleo
100 primers were 5'-labeled with an eight-nucleotide tag unique to each PCR replicate with at least
101 three differences between any pair of tags, allowing the assignment of each sequence to the
102 corresponding sample during sequence analysis. The tags for the forward and reverse primers
103 were identical for each PCR replicate.

104 The purified PCR products were pooled in equal volumes, to achieve a theoretical sequencing
105 depth of 1,000,000 reads per sample. Library preparation and sequencing were performed at
106 Fasteris (Geneva, Switzerland). A total of five libraries were prepared using the MetaFast
107 protocol (Fasteris, [https://www.fasteris.com/dna/?q=content/metafast-protocol-amplicon-](https://www.fasteris.com/dna/?q=content/metafast-protocol-amplicon-metagenomic-analysis)
108 [metagenomic-analysis](https://www.fasteris.com/dna/?q=content/metafast-protocol-amplicon-metagenomic-analysis)), a ligation-based PCR-free library preparation. A paired-end sequencing
109 (2x125 bp) was carried out using an Illumina HiSeq 2500 sequencer on three HiSeq Rapid Flow

110 Cell v2 using the HiSeq Rapid SBS Kit v2 (Illumina, San Diego, CA, USA) following the
111 manufacturer's instructions.

112 **Sequence analyses and taxonomic assignment**

113 To evaluate the current completeness of the online database for the teleo region of the 12S
114 mitochondrial DNA, an *in silico* PCR with 3 allowed mismatches using the teleo primers
115 sequences was performed with ecoPCR (25) on the EMBL database. The generated list of
116 sequenced species was compared to the checklists of fish species present in in the Bird's Head
117 of Papua region, provided by courtesy of Kulbicki et al. 2013 (17).

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

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

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

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

139 different assignment levels from the same sequence. These corrected numbers of taxa were
140 then compared to the number of OTUs from the SWARM workflow in each family to evaluate
141 the magnitude of the diversity missed by the direct assignment method. In the SWARM
142 workflow, a family level assignment was performed as well to remove the taxa that were not
143 fish from nonspecific amplifications and investigate the intra family diversity.

144 **Statistical analyses**

145 To evaluate the number of taxa/OTUs present in the study area, a multimodel approach was
146 implemented to fit asymptotes on the species and OTU accumulation curves. This approach
147 considered 5 different accumulation models (Lomolino, Michaelis-Menten, Gompertz,
148 asymptotic regression and logistic curve) and weighted them using the Akaike Information
149 Criterion (AIC, (29)). For each curve, the accumulation model with the lowest AIC was selected.
150 Accumulation curves and associated asymptotes were generated using the vegan R package. To
151 estimate the sampling effort required to achieve a given proportion of asymptotes, we
152 considered the model selected for accumulation curves. Then, we extracted the predicted
153 number of samples producing a number of taxa/OTUs that outreached 90% and 95% of the
154 asymptotes.

155 **Results**

156 **High heterogeneity of fish species detection among families**

157 A total of 299,479,007 reads were produced using the OBITools pipeline over the 92 eDNA
158 samples corresponding to 14,423 unique sequences with a mean of 307 unique sequences per
159 sample (± 134 SD). In a conservative approach, stringent bioinformatic filters retained 9,345
160 unique sequences so 65% of the total. These 9,345 unique sequences were then assigned to
161 different taxonomic levels using the following genetic similarity thresholds: 100-98% for species,
162 90-98% for genus, 85-90% for family and 80-85% for order. This set of thresholds retained 7,389
163 unique sequences resulting in 678 taxonomic assignments (Suppl. Table S2).

164 A total of 310 species were detected, including 211 coastal fish species present in the checklist
165 of the Bird's Head Peninsula and 99 fish species present in other regions but absent from this
166 checklist (Fig. 1a). Conversely, 183 sequenced fish species which are present in the Bird's Head

167 Peninsula were not detected in our eDNA samples using our stringent filters, representing
168 53.6% of the sequenced species present in the checklist. Since 75.5% of fish species in the
169 checklist of the Bird's Head Peninsula were not sequenced for the 12S rDNA, the largest part of
170 fish species diversity remained hidden through direct assignment (Suppl. Table S1).

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

178 The percentage of fish species sequenced per family varied between 0 and 100% with a mean of
179 40.3% ($\pm 31\%$ SD) in the Bird's Head Peninsula checklist while the percentage of detected
180 species per family varied between 0 and 100% with a mean of 27.1% ($\pm 30.2\%$ SD) in eDNA
181 samples (Fig. 1b). These two percentages were significantly and strongly related ($p < 0.001$) with
182 the percentage of species sequenced per family explaining 85% of variation in the percentage of
183 detected species per family (Fig. 1c).

184 **High but underestimated diversity of OTUs**

185 Given that the low percentage of fish species sequenced for the 12S in the region is the main
186 limitation to detect taxonomic diversity (Fig. 1c), we used an alternative approach based on
187 unique clusters of genetic sequences called Operational Taxonomic Units (OTUs).

188 From the 331,839,591 initial reads, 4,012 OTUs were generated using the SWARM clustering
189 algorithm. After a series of post-clustering curation processes, 972 fish OTUs were filtered
190 among which 819 were assigned to a family (Suppl. Table S3). The number of detected OTUs
191 varied from 1 to 54 among fish families (Fig. 2a), the richest families (>50 OTUs) being the
192 Gobiidae, Labridae and Pomacentridae. Overall the number of OTUs was superior to the
193 number of assigned taxa (genus and species) in 64.7% of the families found in the samples
194 (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 $\Delta = -4$ 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 ($\pm 38.8\%$ 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.

222 Applying this method to the 15 fish families which counted more than 10 OTUs and 10 species in
223 the checklist permitted to assess the ability of eDNA-based accumulation curves to predict
224 regional fish richness. For instance, the OTU accumulation curves for the Gobiidae, Labridae and
225 Pomacentridae, the three richest families (51, 54 and 53 OTUs respectively), produced
226 asymptotes and thus predictions of fish diversity much lower than those in the regional
227 checklists with 107.5, 66.1 and 76.2 OTUs, i.e. 47.5%, 81.7% and 69.6% of the checklist richness
228 respectively (Fig. 3b, c, d).

229 We then tested the ability of the assigned taxa, the OTUs and the OTU accumulation curve
230 approaches to predict fish species richness within families of the regional checklist so the
231 predictive power of linear or proportional relationships. The total number of assigned taxa per
232 family in our samples was a significant but weak predictor of the number of fish species per
233 family in the checklist ($R^2 = 0.60$, $p < 0.001$, Fig. 4a) with the richness of some families being
234 largely underestimated (e.g. 87.4% of net difference with the checklist for the Gobiidae, Fig. 4a,
235 d). The number of OTUs per family was a better predictor of the family species richness in the
236 checklist ($R^2 = 0.80$, $p < 0.001$) but left 20% of unexplained variation among families with still a
237 marked underestimation (73.3% of net difference with the checklist for Gobiidae, Fig. 4b, e).
238 Using the asymptotes of OTU accumulation curves, we obtained a high predictive accuracy of R^2
239 $= 0.92$ ($p < 0.001$) for the species richness within families with less bias for the Gobiidae (43.7%
240 of net difference with the checklist) (Fig. 4c, f).

241 In addition, we observed that the net difference between the number of assigned taxa per
242 family and the number of species per fish family in the checklist is not related to the number of
243 species of the families (Fig. 4d) suggesting an absence of systematic bias towards the
244 underestimation of species-rich families. By contrast, the net difference between the number of
245 OTUs per fish family and the number of species per family in the checklist significantly increased
246 ($R^2 = 0.35$, $p = 0.02$) with the number of species per family (Fig. 4e). This bias towards the
247 underestimation of species richness within species-rich families is nonetheless avoided when
248 using the asymptotes of OTU accumulation curves ($p = 0.24$, Fig. 4f). Thus, asymptotes of OTU
249 accumulation curves are most accurate and least biased eDNA-based predictors of fish species
250 diversity within families in this marine biodiversity hotspot.

251 **Sampling efforts necessary to achieve regional fish diversity inventory**

252 Not only the OTU accumulation curves and their asymptotes provide diversity estimates, they
253 also provide crucial insights into the sampling effort needed to achieve a more complete census.
254 Here, using the asymptote on the OTU accumulation curve for all fish species (Fig. 3a), we found
255 that our 92 cumulated samples (representing 0.2 m³) achieved up to 63.5% of the potential fish
256 OTU diversity in the Bird's Head Peninsula (Fig. 5). To collect 90% of this regional fish diversity,
257 we should have filtered seawater in 735 samples so 8 times the effort of our sampling
258 campaign, representing an aggregated sampled water volume of 1.5 m³. This sampling effort
259 would reach 1,883 samples (an aggregated water volume of 3.8 m³) to collect 95% of the
260 regional fish OTU richness (Fig. 5).

261 On average across fish families, our sampling effort achieved the detection of 77.1% (\pm 14.9 SD)
262 of OTUs predicted by the asymptote of the accumulation curve with a variation among families
263 ranging from 42.2% (Muraenidae) and 47.5% (Gobiidae) to 93.9% (Balistidae) (Fig. 5). The
264 sampling effort needed to achieve 90% of the asymptotic number of OTUs in the region varied
265 greatly among families, ranging from 37 samples for Chaetodontidae to 494 samples for
266 Gobiidae, with a mean of 164 samples (\pm 123 SD). The estimated additional sampling effort to
267 reach 95% from 90% of the OTU richness ranged from 20 more samples (Tetraodontidae) to 593
268 more samples (Gobiidae).

269 **Discussion**

270 **Overcoming incompleteness of genetic reference databases**

271 Environmental DNA metabarcoding has the potential to surpass most classical survey methods
272 to assess biodiversity in both terrestrial and aquatic systems (30). Yet, genetic reference
273 databases are often incomplete especially for species-rich ecosystems such as the Coral
274 Triangle, the global marine biodiversity hotspot (14). For instance, the current completeness of
275 the 12S rDNA online databases for the teleo primer covers only 24.5% of fish species in the
276 Bird's Head Peninsula. Meanwhile, this cover reaches 77.3% for the COI (mitochondrial
277 cytochrome c oxidase subunit I) but fish COI primers still perform poorly in comparison to 12S
278 markers (31).

279 With around 28% of families, 54% of the genera and 76% of species not sequenced for the 12S
280 rDNA teleo primers region, the largest part of fish diversity in the Bird's Head peninsula remains
281 thus hidden through direct assignment. Additionally, sequences present in the reference online
282 databases may have been collected from individuals not located in the region of interest. This
283 can induce assignment errors due to biogeographical related genetic variation (e.g. (32)). The
284 lack of sequencing coverage highlights the immense gap to be filled for online databases to be
285 exhaustive, while numerous species still remain to be described (33). This limitation prevents
286 metabarcoding approaches from characterizing entire fish assemblages through direct species
287 assignment. Yet, the taxa-assignment method reveals the presence of 211 fish species
288 referenced in the checklist of coastal fishes in the Bird's Head peninsula (Fig. 1a). Conversely, 99
289 assigned species were absent from this checklist. These 99 detections can either be true
290 presences extending the distribution of some species and revisiting the regional checklist or
291 false presences due to wrong assignments or possible contaminations. For instance, the Atlantic
292 salmon (*Salmo salar*), probably a lab kit contaminant, was found in our study and removed from
293 the analyses (see Methods). The large number of species present in the samples but absent
294 from the regional checklist suggests that inventories of some families are still incomplete. On
295 average 2.5 detected species per family (± 2.6 SD, Fig. 1b) are absent from the checklist, ranging
296 from 0 to 14 species (Apogonidae). This mismatch allows to target future sampling efforts
297 towards families and their habitats to complete the regional checklist.

298 As an alternative to species assignment, the use of OTUs as species proxy units is an option that
299 has not yet been tested for vertebrates in species-rich ecosystems while currently used when
300 the concept of species is debatable like for fungi or unicellular organisms (34, 35).

301 Here, using a conservative and stringent bioinformatic pipeline, we show that the diversity of
302 OTUs is a weak and biased estimator of species diversity with species-rich families being
303 strongly underrepresented. To overcome this limitation, we propose to rely on OTU
304 accumulation curves which provide an unbiased estimate of regional fish diversity and fish
305 richness within families. The asymptotes underestimate the regional fish species richness but
306 the bias is highly consistent among families (Fig. 4f). We thus propose to extend this method for

307 taxonomic inventories in poorly-sampled ecosystems like the deep sea to estimate the diversity
308 at different taxonomic levels.

309 **Revealing the potential and limitation of eDNA metabarcoding inventories**

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

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

329 Even if not assigned at species-level, OTUs can be defined as distinct entities for which their
330 distribution and temporal variability can be assessed and monitored (36). Moreover, the OTUs
331 and their associated sequences can remain in public repositories until they are assigned to a
332 species, subspecies or complex as databases improve (37). However, the major caveat of using
333 OTUs for diversity inventories is that they cannot be directly considered as species with
334 complete certainty. Species with intra-specific genetic variability can produce two separate
335 OTUs, overestimating species diversity. Conversely, two species phylogenetically close to each

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

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

348 The contrasts between assigned taxa diversity, OTU diversity and OTU asymptote diversity show
349 that the detectability varies strongly among fish families. These contrasts can be related to the
350 ecology of the species but also to the state of the retrieved DNA fragments (intra or
351 extracellular), their sources (e.g. gametes, larvae, feces), their release rate, their diffusion in the
352 water column (limited or wide) and their transportation (38). For instance, a benthic fish species
353 such as gobies with a small movement range would release DNA fragments through skin and
354 feces on a small area. However, such species could release a massive number of gametes
355 carried through the water column (13) so may appear highly detectable during breeding season.
356 Further comparative works are urgently needed between visual, camera and eDNA
357 metabarcoding surveys to better estimate the level of detectability of each species or family in
358 order to provide reliable biodiversity assessments. For instance, coupling eDNA metabarcoding
359 and video surveillance allows the detection of eighty-two fish genera from 13 orders on reefs
360 and seagrass with only 24 genera in common (39). Investigating biodiversity should also
361 consider its multiple components including functional and phylogenetic diversity that are key for
362 reef ecosystem functioning (40). Associating OTUs to species might allow to fill this gap but it
363 will require massive sampling and sequencing efforts.

364 **Acknowledgments**

365 **General:** We thank the Indonesian Institute of Sciences (LIPI) for promoting our collaboration
366 and the Sorong Polytechnic of Marine and Fisheries (Politeknik KP Sorong, West Papua) for
367 providing the vessel Airaha 02 that we used in this campaign. We thank the crew of the Aihara
368 02 for assisting us during the operations and the SPYGEN staff for the technical support in the
369 laboratory.

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

376 **References**

- 377 1. Costello, M.J. & Chaudhary, C. (2017) Marine biodiversity, biogeography, deep-sea, and
378 conservation. *Current Biology*, **27**: R511-R527. DOI: 10.1016/j.cub.2017.04.060.
- 379 2. Barlow, J. et al. (2018) The future of hyperdiverse tropical ecosystems. *Nature*, **559**: 517–
380 526. DOI: 10.1038/s41586-018-0301-1.
- 381 3. Lees, A.C. & Pimm, S.L. (2015) Species, extinct before we know them. *Current Biology*, **5**:
382 R177-R180. DOI: 10.1016/j.cub.2014.12.017.
- 383 4. Díaz, S. et al. (2018) Assessing nature’s contributions to people. *Science*, **359**: 270-272.
384 DOI: 10.1126/science.aap8826.
- 385 5. Duffy, J.E., Godwyn, C.M. & Cardinale, B.J. (2017) Biodiversity effects in the wild are
386 common and as strong as key drivers of productivity. *Nature*, **0**: 1-4. DOI:
387 10.1038/nature23886.
- 388 6. Juhel, J.B., Vigliola, L., Wantiez, L., Letessier, T.B., Meeuwig, J.J. & Mouillot, D. (2019)
389 Isolation and no-entry marine reserves mitigate anthropogenic impacts on grey reef
390 shark behavior. *Scientific reports*, **9**: 2897. DOI: 10.1038/s41598-018-37145-x.
- 391 7. Brandl, S.J., Goatley, C.H.R., Bellwood, D.R. & Tornabene, L., (2018) The hidden half:
392 ecology and evolution of cryptobenthic fishes on coral reefs. *Biol Rev*, **93**: 1846-1873.
393 DOI: 10.1111/brv.124233.

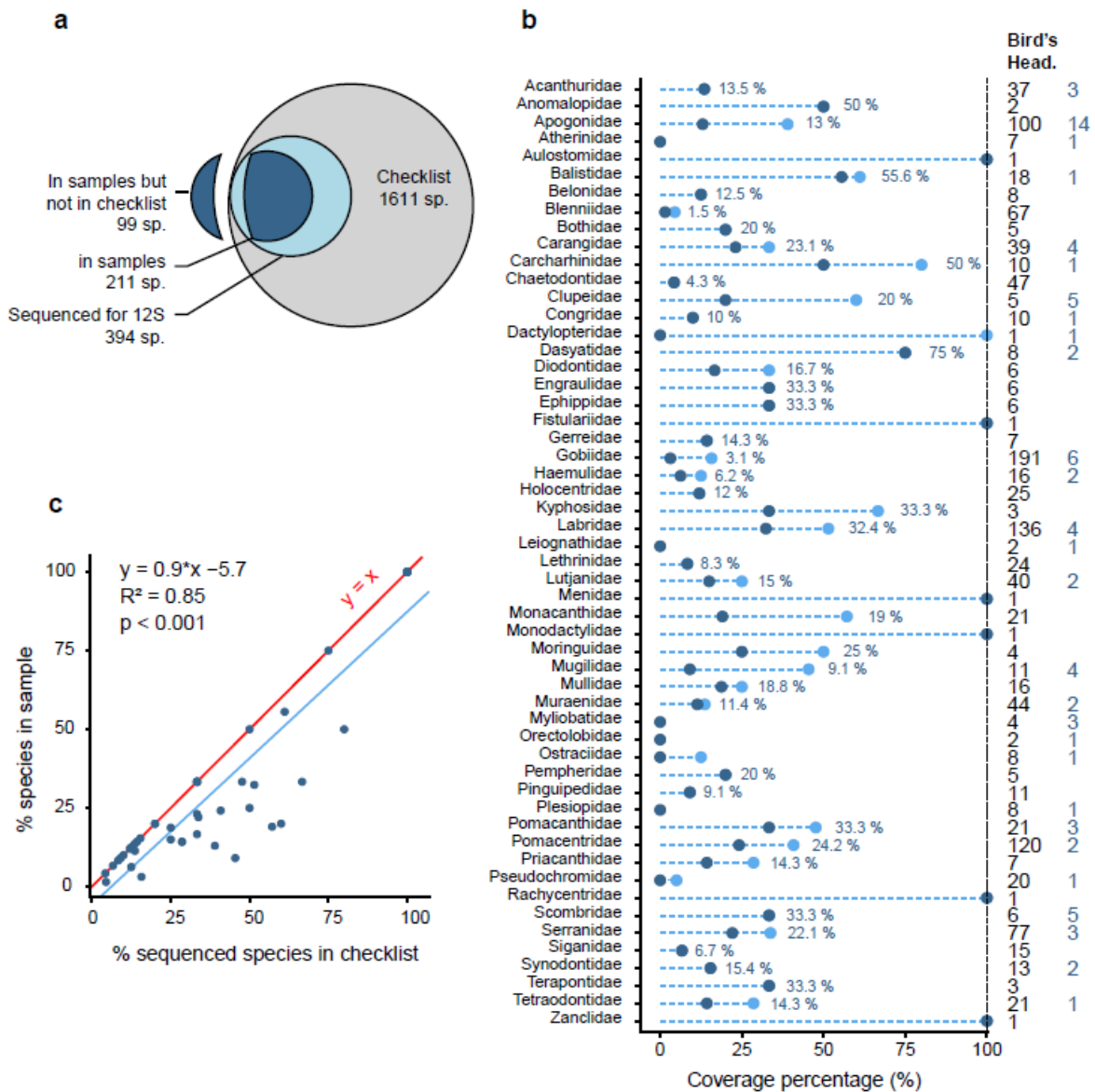
- 394 8. Garlapati, D., Charankumar, B., Ramu, K., Madeswaran, P. & Ramana Murthy, M.V.
395 (2019) A review on the applications and recent advances in environmental DNA (eDNA)
396 metagenomics. *Rev Environ Sci Bio*. **18**: 389. DOI: 10.1007/s11157-019-09501-4.
- 397 9. Boussarie, G. et al. (2018) Environmental DNA illuminates the dark diversity of sharks. *Sci*
398 *Adv*, **4**: eaap9661. DOI: 10.1126/sciadv.aap9661.
- 399 10. Fukumoto, S., Ushimaru, A. & Minamoto, T. (2015) A basin-scale application of
400 environmental DNA assessment for rare endemic species and closely related exotic
401 species in rivers: a case study of giant salamanders in Japan. *J Appl Ecol*, **52**: 358-365.
402 DOI: 10.1111/1365-2664.12392.
- 403 11. Ruppert, K.M., Kline, R.J. & Rahman, Md S. (2019) Past, present, and future of
404 environmental DNA (eDNA) metabarcoding: a systematic review in methods, monitoring,
405 and applications of global eDNA. *Glob Ecol Conserv*, **17**: e00547. DOI:
406 10.1016/j.gecco.2019.e00547.
- 407 12. Mahé, F., Rognes, T., Quince, C., de Vargas, C. & Dunthorn, M. (2014) Swarm: robust and
408 fast clustering method for amplicon-based studies. *PeerJ*, **2**: e593. DOI:
409 10.7717/peerj.593.
- 410 13. Cordier, T (2019) Multi-marker eDNA metabarcoding survey to assess the environmental
411 impact of three offshore gas platforms in the North Adriatic Sea (Italy). *Mar. Environ.*
412 *Res.*, **146**: 24-34. DOI: 10.1016/j.marenvres.2018.12.009.
- 413 14. Brandl, S.J., Rasher, D.B., Côté, I.M., Casey, J.M., Darling, E.S., Lefcheck, J.S. & Duffy, J.E.
414 (2019) Coral reef ecosystem functioning: eight core processes and the role of
415 biodiversity. *Front Ecol Environ*. **17**:445-454. DOI: 10.1002/fee.2088.
- 416 15. Veron, J.E.N., Devantier, L.M., Turak, E., Green, A.L., Kininmonth, S., Stafford-Smith, M. &
417 Peterson, N. (2009) Delineating the Coral Triangle. *Galaxea, JCRS*, **11**: 91-100. DOI:
418 10.3755/galaxea.11.91.
- 419 16. Allen, G.R. & Erdmann, M.V. (2012) Reef fishes of the East Indies. Volumes I-III. Tropical
420 Reef Research, Perth, Australia. ISBN: 978-0-9872600-0-0. 1,292 p.
- 421 17. Kulbicki, M. et al. (2013) Global Biogeography of Reef Fishes: A Hierarchical Quantitative
422 Delineation of Regions. *Plos One*, **8**: e81847. DOI: 10.1371/journal.pone.0081847.
- 423 18. Exton, D.A. et al. (2019) Artisanal fish fences pose broad and unexpected threats to the
424 tropical coastal seascape. *Nat Commun*, **10**: 2100. DOI: 10.1038/s41467-019-10051-0.
- 425 19. Jones, L.A., Mannion, P.D., Farnsworth, A., Valdes, P.J., Kelland, S.-J. & Allison, P.A.
426 (2019) Coupling of palaeontological and neontological reef coral data improves forecasts
427 of biodiversity responses under climatic change. *Roy Soc Open Sci*, **6**: 182111. DOI:
428 10.1098/rsos.182111.
- 429 20. Ainsworth, C.H., Pitcher, T.J. & Rotinsulu, C. (2008) Evidence of fishery depletions and
430 shifting cognitive baselines in Eastern Indonesia. *Biol Conserv*, **141**: 848-859. DOI:
431 10.1016/j.biocon.2008.01.006.

- 432 21. Valentini, A. et al. (2016) Next-generation monitoring of aquatic biodiversity using
433 environmental DNA metabarcoding. *Mol Ecol*, **25**: 929–942. DOI: 10.1111/mec.13428.
- 434 22. Goldberg, C.S., Turner, C.R., Deiner, K., Klymus, K.E., Thomsen, P.F., Murphy, M.A., Spear,
435 S.F., McKee, A., Oyler-McCance, S.J., Cornman, R.S., Laramie, M.B., Mahon, A.R., Lance,
436 R.F., Pilliod, D.S., Strickler, K.M., Waits, L.P., Fremier, A.K., Takahara, T., Herder, J.E. &
437 Taberlet, P. (2016) Critical considerations for the application of environmental DNA
438 methods to detect aquatic species. *Methods Ecol Evol*, **7**: 1299-1307. DOI: 10.1111/2041-
439 210X.12595.
- 440 23. Pont, D. et al. (2018) Environmental DNA reveals quantitative patterns of fish
441 biodiversity in large rivers despite its downstream transportation. *Sci Rep*, **8**: 10361. DOI:
442 10.1038/s41598-018-28424-8.
- 443 24. Baker, W., van den Broek, Camon, E., Hingamp, P., Sterk, P., Stoesser, G. & Tuli, M.A.
444 (2000) The EMBL nucleotide sequence database. *Nucleic Acids Res*, **28**: 19-23. DOI:
445 10.1093/nar/gki098.
- 446 25. Ficetola, G.T., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessièrè, J., Taberlet, P. &
447 Pompanon, F. (2010) An *in silico* approach for the evaluation of DNA barcodes. *BMC*
448 *Genomics*, **11**: 434. DOI: 10.1186/1471-2164-11-434.
- 449 26. Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P. & Coissac, E. (2016) OBITOOLS: a
450 UNIX-inspired software package for DNA metabarcoding. *Mol Ecol Res*, **16**: 176-182. DOI:
451 10.1111/1755-0998.12428.
- 452 27. Larkin, M.A. et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**: 2947-
453 2948. DOI: 10.1093/bioinformatics/btm404.
- 454 28. Kearse, M. et al. (2012) Geneious basic: an integrated and extendable desktop software
455 platform for the organization and analysis of sequence data. *Bioinformatics*, **28**: 1647–
456 1649. DOI: 10.1093/bioinformatics/bts19.
- 457 29. Aho, K., Derryberry, D. & Peterson, T. (2014) Model selection for ecologists: the
458 worldviews of AIC and BIC. *Ecology*, **95**: 631-636. DOI: 10.1890/13-1452.1.
- 459 30. Deiner, K. et al. (2017) Environmental DNA metabarcoding: Transforming how we survey
460 animal and plant communities. *Mol Ecol*, **26**:5872-5895. DOI: 10.1111/mec.14350.
- 461 31. Collins, R.A., Bakker, J., Wangenstein, O.S., Soto, A.Z., Corrigan, L., Sims, D.W., Genner,
462 M.J. & Mariani, S. (2019) Non-specific amplification compromises environmental DNA
463 metabarcoding with COI. *Methods Ecol Evol*, **10**: 1985-2001. DOI: 10.1111/2041-
464 210X.13276.
- 465 32. Wadrop, E., Hobbs, J.-P., Randall, J.E., DiBattista, J.D., Rocha, L.A., Kosaki, R.K., Berumen,
466 M.L. & Bowen, B.W. (2016) Phylogeography, population structure and evolution of coral-
467 eating butterflyfishes (Family Chaetodontidae, genus *Chaetodon*, subgenus
468 *Corallochaetodon*). *J Biogeogr*, **43**: 1116-1129. DOI: 10.1111/jbi.12680.
- 469 33. Pinheiro, H.T., Moreau, S., Daly, M. & Rocha, L. A. (2019) Will DNA barcoding meet
470 taxonomic needs? *Science*, **365**: 873–875. DOI: 10.1126/science.aay7174.

- 471 34. Pawlowski, J. et al. (2018). The future of biotic indices in the ecogenomic era: Integrating
472 (e)DNA metabarcoding in biological assessment of aquatic ecosystems. *Sci Total Environ*,
473 **637-638**: 1295-1310. DOI: 10.1016/j.scitotenv.2018.05.002.
- 474 35. Lladó Fernández, S., Větrovský, T. & Baldrian, P. (2019) The concept of operational
475 taxonomic units revisited: genomes of bacteria that are regarded as closely related are
476 often highly dissimilar. *Folia Microbiol*, **64**: 19–23. DOI: 10.1007/s12223-018-0627-y.
- 477 36. Cordier, T., Esling, P., Lejzerowicz, F., Visco, J., Ouadahi, A., Martins, C., Cedhagen, T. &
478 Pawlowski, J. (2017) Predicting the ecological quality status of marine environments
479 from eDNA metabarcoding data using supervised machine learning. *Environ Sci Technol*,
480 **51**: 9118-9126. DOI: 10.1021/acs.est.7b01518.
- 481 37. Wangensteen, O., Palacín, C., Guardiola, M. & Turon, X. (2018) DNA metabarcoding of
482 littoral hard-bottom communities: high diversity and database gaps revealed by two
483 molecular markers. *PeerJ*, **6**: e4705. DOI: 10.7717/peerj.4705.
- 484 38. Harrison, J.B., Sunday, J.M. & Rogers S.M. (2019) Predicting the fate of eDNA in the
485 environment and implications of studying biodiversity. *Proc R Soc B*, **286**: 20191409. DOI:
486 10.1098/rspb.2019.1409.
- 487 39. Stat, M., Jeffrey, J., DiBattista, J.D., Newman, S.J., Bunce, M. & Harvey, E.S. (2018)
488 Combined use of eDNA metabarcoding and video surveillance for the assessment of fish
489 biodiversity. *Conserv Biol*, **33**: 196-205. DOI: 10.1111/cobi.13183.
- 490 40. Duffy, J.E., Lelcheck, J.S., Stuart-Smith, R.D., Navarrete, S.A. & Edgar, G.J. (2016)
491 Biodiversity enhances reef fish biomass and resistance to climate change. *P Natl Acad*
492 *Sci*, **113**: 6230-6235. DOI: 10.1073/pnas.1524465113.
- 493 41. Juhel J-B et al. (2020) Data from: Accumulation curves of environmental DNA sequences
494 predict coastal fish diversity in the coral triangle. Dryad Digital Repository. (doi:
495 10.5061/dryad.t1g1jw05)
- 496

497 **Figure legends**

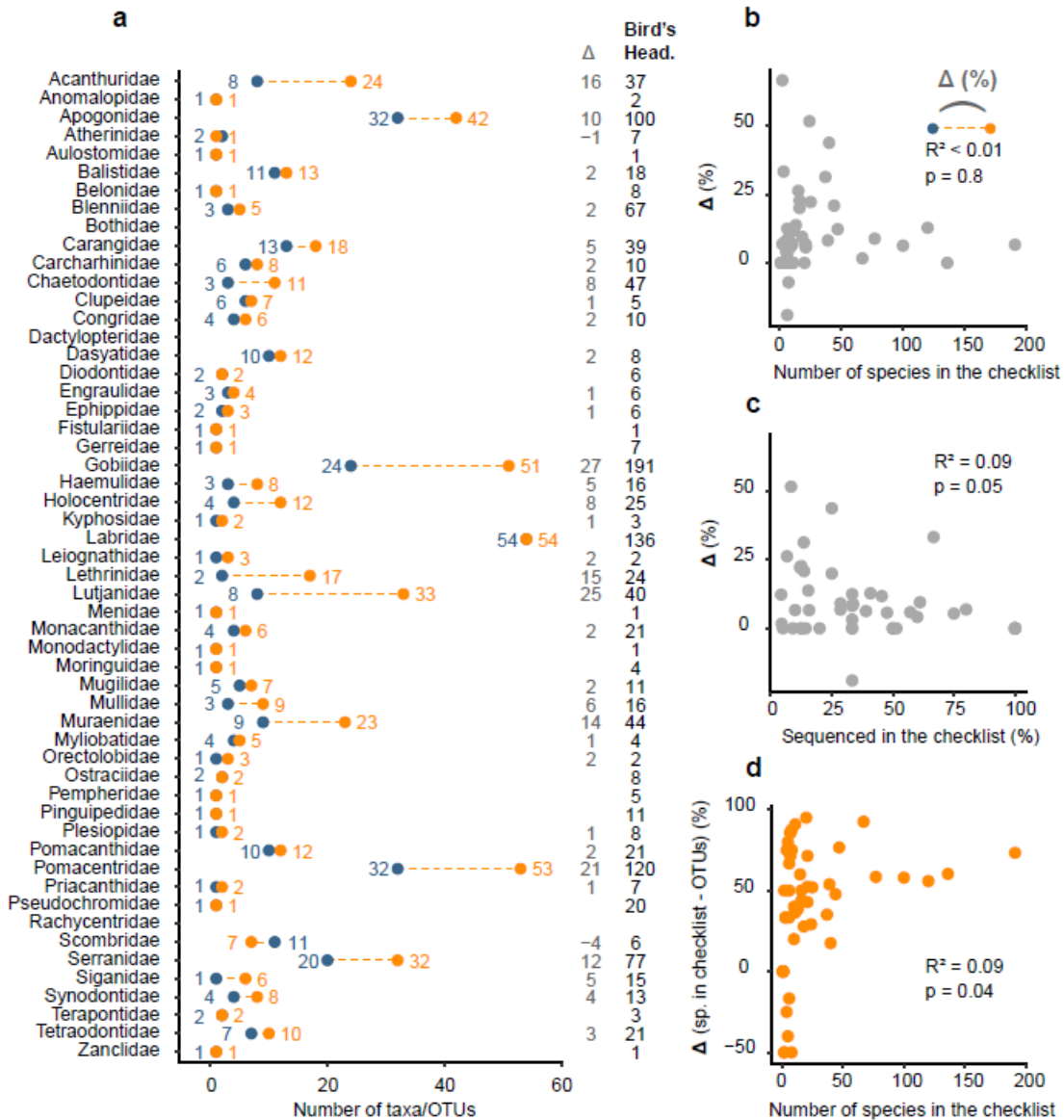
498



499

500 **Fig. 1. Number of fish species present in the checklist of the Bird's Head region (grey),**
 501 **sequenced in the European Molecular Biology Laboratory database (EMBL) (light blue) and**
 502 **detected in the eDNA samples (dark blue) (a) ; percentage of species detected in the samples**
 503 **(dark blue), sequenced in EMBL (light blue) in each family of species (b) ; percentage of**
 504 **species detected in the samples as a function of the percentage of sequenced species in EMBL**
 505 **(c). (b) The percentages of the species detected in the eDNA samples compared to the species**
 506 **present in the Bird's Head region are displayed next to the points. The number of species per**

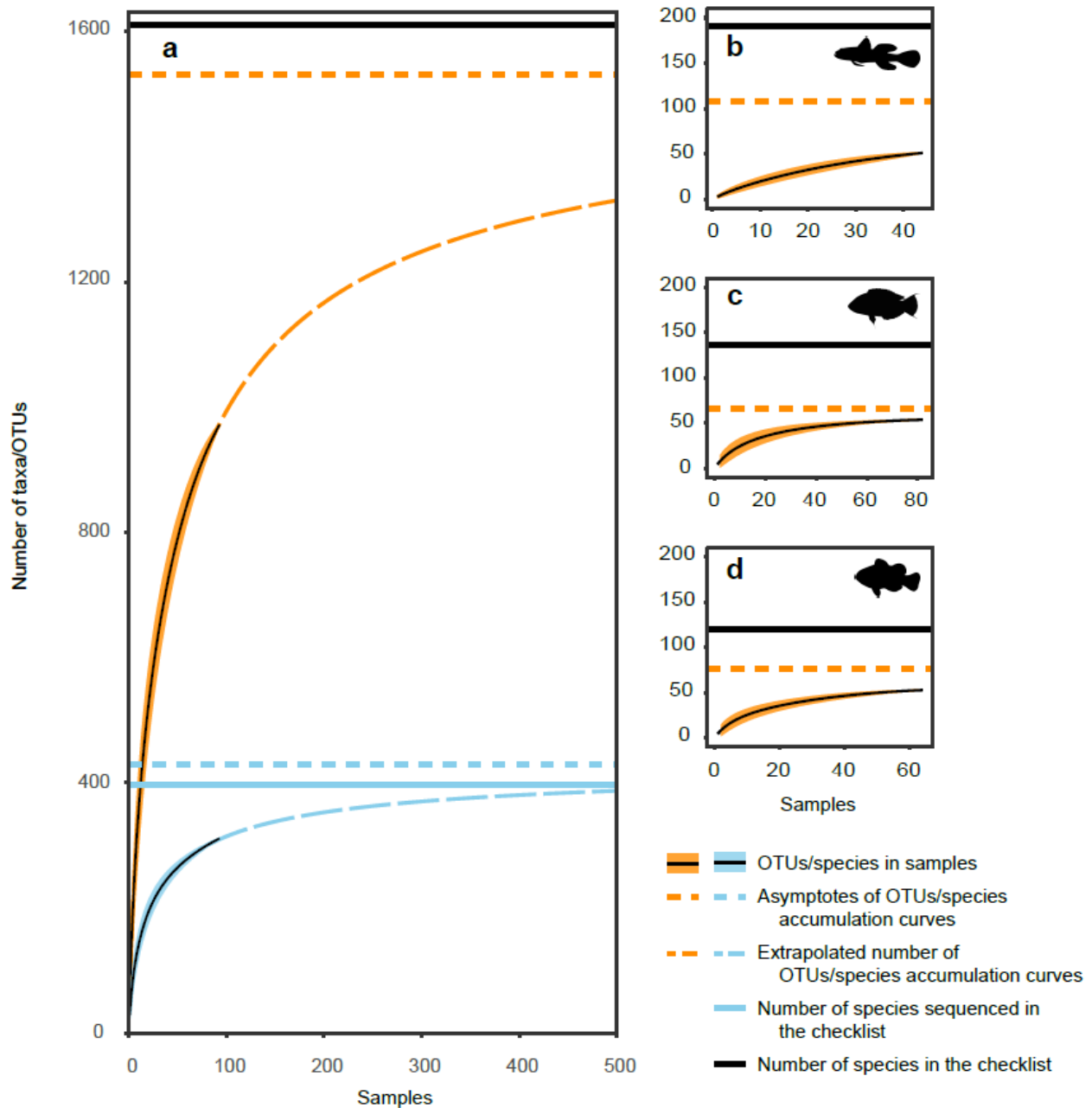
507 family in the checklist and the number of species detected in the samples but not present in the
 508 checklist are both on the right of the figure in black and dark blue, respectively. Only the
 509 sequences assigned to species using ecotag program (similarity >98%) are used in this figure. (c)
 510 Each point corresponds to a fish family.
 511



512
 513 **Fig. 2.** Number of taxa assigned by the OBITools workflow (blue) and number of OTUs
 514 generated by the SWARM workflow (orange) in the different fish families (a) ; distribution of
 515 the differences between the two workflows as a function of family richness (b) and as a

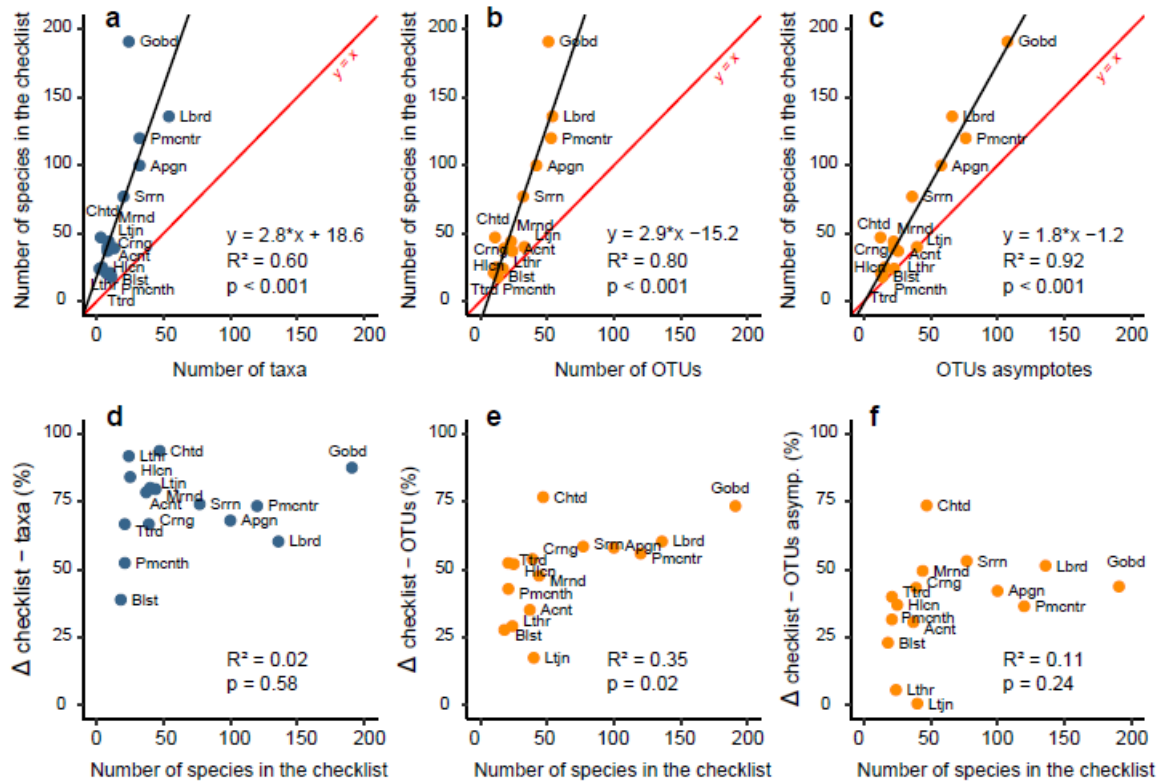
516 **function of family sequencing coverage (c); distribution of the differences between OTUs and**
517 **the number of taxa (species and genus) in the checklist as a function of family richness (d).**

518 (a)The difference of taxa/OTUs between the two methods (noted Δ) and the number of species
519 in the checklist of the Bird's Head region are on the right of the figure in grey and black,
520 respectively. For the OBITools workflow, only the sequences assigned to species and genus
521 using ecotag program (similarity > 98% and > 90% respectively) are used in this figure. For the
522 SWARM workflow, only the OTUs curated by LULU and assigned to family (similarity > 85%) are
523 used in this figure.

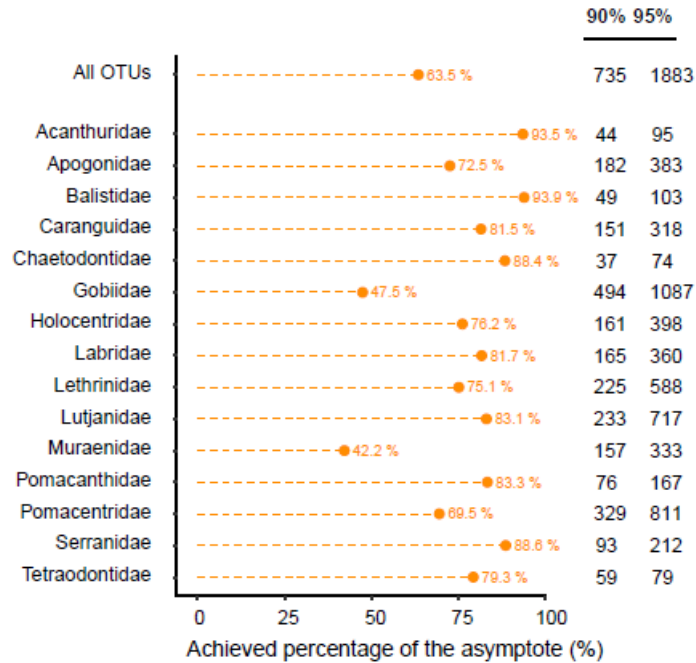


524

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



531
 532 **Fig. 4.** Linear regression of the diversity of the most diverse families as a function of the
 533 number taxa assigned (a), the number of OTUs (b), the asymptotes of the OTUs accumulation
 534 curves (c); and differences between the number of taxa assigned (d), the number of OTUs (e),
 535 the asymptotes of OTUs accumulation curves (f) and the number of species in the checklist as
 536 a function of the number of species in the checklist. Only the families with a number of OTU
 537 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.**