

1 **Title**

2 *A genome assembly of the Atlantic chub mackerel (Scomber colias): a valuable teleost*
3 *fishing resource*

4 **Authors**

5 André M. Machado^{1,2}, André Gomes-dos-Santos^{1,2}, Miguel Fonseca¹, Rute R. da
6 Fonseca^{3,4}, Ana Veríssimo^{5,6}, Mónica Felício⁷, Ricardo Capela^{1,2}, Néilson Alves^{1,2},
7 Miguel Santos^{1,2}, Filipe Salvador-Caramelo^{1,2}, Marcos Domingues¹, Raquel Ruivo¹,
8 Elsa Froufe¹ and L. Filipe C. Castro^{1,2*}

9
10 **Affiliations**

- 11 1. CIIMAR – Interdisciplinary Centre of Marine and Environmental Research, U.
12 Porto – University of Porto, Porto, Portugal
13 2. Department of Biology, Faculty of Sciences, U. Porto - University of Porto,
14 Portugal
15 3. Center for Global Mountain Biodiversity, GLOBE Institute, University of
16 Copenhagen, Copenhagen, Denmark
17 4. Center for Macroecology, Evolution, and Climate, GLOBE Institute, University
18 of Copenhagen, Denmark
19 5. CIBIO - Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO
20 - Laboratório Associado, Campus de Vairão, Universidade do Porto, 4485-661
21 Vairão, Portugal
22 6. BIOPOLIS - Program in Genomics, Biodiversity and Land Planning, CIBIO,
23 Campus de Vairão, 4485-661 Vairão, Portugal
24 7. Portuguese Institute for the Sea and Atmosphere, I.P. (IPMA), Portugal

25
26 **EMAIL:**

27 AMM – andremmachado25@gmail.com

28 AGS – andreposua64@gmail.com

29 MF – mig.m.fonseca@gmail.com

30 RRdF – rute.r.da.fonseca@gmail.com

31 AV – verissimoac@gmail.com

32 MF – monicafelicio124@hotmail.com

33 RC – ricardocapela.c@gmail.com

34 NA – nalves@ciimar.up.pt
35 MS – santosmaf@gmail.com
36 FSC – filipemscaramelo@gmail.com
37 MD – markosdomingues@hotmail.com
38 RR – ruivo.raquel@gmail.com
39 EF – elsafroufe@gmail.com
40 LFCC – lfilipecastro@gmail.com
41
42 ORCID:
43 AMM – <http://orcid.org/0000-0002-6857-7581>
44 AGS – <http://orcid.org/0000-0001-9973-4861>
45 MF – <http://orcid.org/0000-0003-4598-670X>
46 RRdF – <https://orcid.org/0000-0002-2805-4698>
47 AV – <https://orcid.org/0000-0003-3396-9822>
48 MF – <https://orcid.org/0000-0003-3477-3944>
49 RC – <https://orcid.org/0000-0002-7822-4536>
50 NA – <https://orcid.org/0000-0002-4570-0647>
51 MS – <https://orcid.org/0000-0001-7347-0546>
52 FSC – <https://orcid.org/0000-0001-9345-6311>
53 MD – <https://orcid.org/0000-0002-9680-1253>
54 RR – <http://orcid.org/0000-0003-2412-8730>
55 EF – <https://orcid.org/0000-0003-0262-0791>
56 LFCC – <http://orcid.org/0000-0001-7697-386X>

57

58 * corresponding author(s): Luís Filipe Costa de Castro (filipe.castro@ciimar.up.pt)

59

60 **Abstract**

61 The Atlantic chub mackerel, *Scomber colias* Gmelin, 1789, is a medium-size pelagic
62 fish with substantial importance in the fisheries of the Atlantic Ocean and the
63 Mediterranean Sea. Over the past decade, this species has gained special relevance
64 being one of the main targets of pelagic fisheries in the NE Atlantic. Here, we
65 sequenced and annotated the first high-quality draft genome assembly of *S. colias*,
66 produced with Pacbio HiFi long reads and Illumina Paired-End short reads. The
67 estimated genome size is 814 Mb distributed into 2,028 scaffolds and 2,093 contigs
68 with an N50 length of 4,19 and 3,34 Mb, respectively. We annotated 27,675 protein-
69 coding genes and the BUSCO analyses indicated high completeness, with 97.3 % of
70 the single-copy orthologs in the Actinopterygii library profile. The present genome
71 assembly represents a valuable resource to address the biology and management of
72 this relevant fishery. Finally, this is the fourth high-quality genome assembly within
73 the Order Scombriformes and the first in the genus *Scomber*.

74

75 **Keywords:** *Scomber colias*, Atlantic chub mackerel, Scombriformes, genome,
76 fisheries management, population dynamics, endocrinology, teleosts

77

78

79

80

81

82

83

84

85

86

87

88

89

90

91

92

93 **Data description**

94 **Background and context**

95 The family Scombridae is divided into two subfamilies (Gasterochismatinae and
96 Scombrinae), with 5 genera and around 51 described species, comprising mackerels,
97 bonitos, and tunas [1]. The representative genus of the Scombridae, i.e., *Scomber*
98 includes 4 species: *S. scombrus*, *S. japonicus*, *S. australasicus*, and *S. colias*. The
99 Atlantic chub mackerel, *Scomber colias* Gmelin, 1789, (NCBI Taxonomy ID: 338315)
100 is a small coastal-pelagic fish widely distributed in the Atlantic Ocean from the Bay of
101 Biscay to South Africa (including the Canary, Madeira, Azores, and Saint Helena
102 Islands) and in the Mediterranean Sea [2]. *Scomber colias* is usually found in depths
103 up to 300 m and occupies a key position in the trophic web. This species acts as a link
104 between primary producers and top predators since it feeds mainly on zooplankton
105 and some small pelagic fish and is an essential element of the diet of larger pelagic
106 fish (e.g., tuna, swordfish, and sharks) and marine mammals (e.g., dolphins and seals)
107 [3]. Besides its ecological importance, *S. colias* also supports important commercial
108 fisheries for several countries throughout its distribution range, being an important
109 component in the diet of several local populations [1, 4]. This is probably related to its
110 nutritional value, as this mackerel is a privileged source of important and beneficial
111 fatty acids for human nutrition, particularly Docosahexaenoic acid (DHA), an omega-
112 3 fatty acid [5, 6]. Additionally, *S. colias* is used as bait for the tuna longline and
113 handline fisheries and caught in purse seine and pelagic trawl fisheries which target
114 sardine and/or anchovy [7].

115 The availability of *S. colias* makes it a sustainable marine resource [6] and a viable
116 alternative to the European's sardine (*Sardina pilchardus*), which is under fishing
117 restrictions because of population decline. Curiously, fluctuations in the abundance
118 and a northwards shift in the distribution of *S. colias*, with a likely inverse relationship
119 with sardine abundance has been recently demonstrated [8]. Due to its ecological and
120 economic importance, *S. colias* has been the focus of several recent studies focusing
121 on different aspects of its fisheries and biology [3, 8, 9]. Yet, genomic resources for
122 the species are still limited. Only a (liver) transcriptome [10], mitogenome [11], and
123 single-nucleotide polymorphism (SNP) data, obtained through RAD-seq [12], have
124 been described for the species. With the vast majority of the world's fish stocks
125 already collapsed and with climate change as additional pressure, knowledge of fish
126 genomes is becoming an invaluable tool to address conservation efforts [13, 14]. Here,

127 we report the first high-quality draft genome of *S. colias*, assembled with Illumina and
128 Pacific Biosciences (PacBio) Single-Molecule High-Fidelity (HiFi) reads. This
129 resource provides a critical platform to uncover the species' adaptive physiological
130 potential in a changing environment. Specifically, it will help understand the current
131 observed populational northward shift, postulated to be part of a more general
132 expansion of species from warmer areas [8]. Moreover, being one of the genomes
133 with higher quality within family Scombridae and the first within the *Scomber* genus,
134 the obtained information will help to improve the conservation, management, and
135 sustainable exploitation of this valuable fish resource as well as of its highly valued
136 congeners.

137



138

139 **Figure 1**

140

141 **Methods**

142 **Sampling and DNA extraction**

143 Two specimens of *S. colias* were collected at two sampling points and time frames.
144 The first specimen was collected in 2017, during the “*Programa Nacional de*
145 *Amostragem Biológica*” managed by the Instituto Português do Mar e da Atmosfera”
146 (IPMA) in North Atlantic waters (41.501944 N 8.851667 W). From this individual,
147 two tissues were collected and stored in 100% ethanol (muscle) and RNA later (liver).
148 The liver tissue was used to produce and describe the first liver transcriptome of *S.*
149 *colias* [10]. Muscle tissue was used in the present study, for genomic DNA (gDNA)
150 extraction using the DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany),
151 following the manufacturer’s instructions. The gDNA was then used for the Illumina
152 Paired-End (PE) sequencing (described below). The second specimen was caught in
153 2020 near Mira, Portugal (40.5588270 N 9.4529720 W). Immediately upon
154 harvesting, the muscle was snap-frozen in liquid nitrogen. The frozen tissue was
155 shipped to Brigham Young University DNA Sequencing Center (BYU), where gDNA
156 with high molecular weight was extracted from 1.1 g of muscle using the QIAGEN
157 Genomic-tip 20/G Kit. The quality and concentration of gDNA were assessed with
158 Qubit Fluorometric system (ThermoFisher), and the fragment size was determined
159 with a Fragment analyzer (Agilent) before loading on the Sequel II.

160

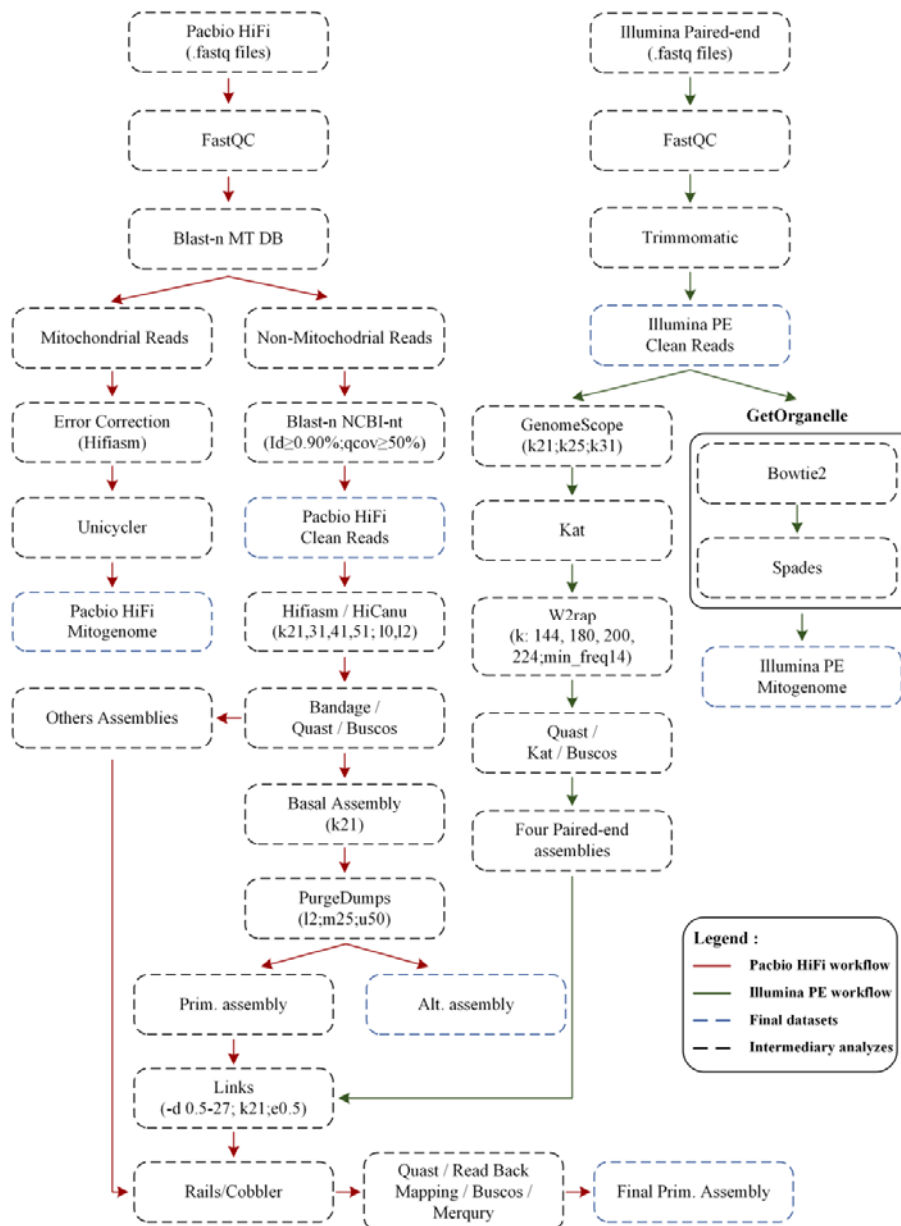
161 **DNA sequencing libraries construction and sequencing**

162 For the first DNA sample, Illumina PE library preparation and sequencing were
163 carried out by Macrogen, Inc (Seoul, Korea), using Illumina HiSeq X Ten platform
164 with 250 bp PE configuration. For the second specimen, PacBio HiFi library
165 preparation and sequencing were performed at BYU, following the manufacturer’s
166 recommendations (Pacific Biosciences) ([https://www.pacb.com/wp-](https://www.pacb.com/wp-content/uploads/Procedure-Checklist-Preparing-HiFi-SMRTbell-Libraries-using-SMRTbell-Express-Template-Prep-Kit-2.0.pdf)
167 [content/uploads/Procedure-Checklist-Preparing-HiFi-SMRTbell-Libraries-using-](https://www.pacb.com/wp-content/uploads/Procedure-Checklist-Preparing-HiFi-SMRTbell-Libraries-using-SMRTbell-Express-Template-Prep-Kit-2.0.pdf)
168 [SMRTbell-Express-Template-Prep-Kit-2.0.pdf](https://www.pacb.com/wp-content/uploads/Procedure-Checklist-Preparing-HiFi-SMRTbell-Libraries-using-SMRTbell-Express-Template-Prep-Kit-2.0.pdf)). The size-selected fraction had 15.3 kb
169 mean read length and was selected on Sage-Elf system (Sage Sciences). The
170 sequencing was conducted on 2 single-molecule, real-time (SMRT) cells using Sequel
171 II system (v.9.0), with a run time of 30h and 2.9 hours pre-extension. The circular
172 consensus analysis was performed in SMRT® Link v9.0 under the default settings
173 (check details at the Suppl. Table 1).

174

175 **Raw data quality-control, clean-up, and Genome size estimation**

176 Both short and long read datasets were assessed by FastQC (v.0.11.8)
177 (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). The Trimmomatic
178 (v.0.38) [15] software was used to filter and remove the low-quality reads as well as
179 the adaptors of the Illumina dataset (LEADING:5 TRAILING:5
180 SLIDINGWINDOW:4:20 MINLEN:50). Next, trimmed datasets were used to check
181 the overall characteristics of the *S. colias* genome (i.e., genome size, heterozygosity,
182 or unique content), through the GenomeScope2.0 [16]. Briefly, the Jellyfish (v.2.2.10)
183 [17] software was used to build the k-mer frequency distributions, and the final k-mer
184 counts (k-mer 21, 25, 31) were submitted to the GenomeScope2.0 online platform. On
185 the other hand, the HiFi reads were filtered in two ways (Fig. 2). First, the
186 mitochondrial reads were removed by blast searches (blast-n) using a pre-built
187 database of mitochondrial sequences (Database build protocol; 1 – Selection of all
188 complete mitogenomes present in NT-NCBI (nucleotide database of National Center
189 for Biotechnology Information (NCBI)); 2 – Select by taxon (Actinopterygii;
190 Taxonomy ID: 7898); 3 – Sequence length filter 15000-50000 bp; 4 – Build a
191 database with makeblastdb application of ncbi-blast+ (v.2.9.0)). Second, to filter out
192 possible sources of contamination (artefactual or biological), the HiFi reads were
193 checked by blast (blast-n) against NT-NCBI. In the end, only HiFi reads having match
194 hits with more than 90% of identity and query coverage of 50% in Actinopterygii
195 taxon (NCBI Taxonomy ID: 7898) or without match hits at all were considered for
196 further analyses (Fig. 2).



197

198 **Figure 2**

199 **Mitochondrial genome assembly**

200 Given that two specimens were used for the distinct sequencing approaches, i.e.,
 201 PacBio HiFi and Illumina PE, the whole mitochondrial genome (mtDNA) was
 202 assembled and characterized for both specimens. For specimen one, trimmed Illumina
 203 PE reads were used to assemble mtDNA in GetOrganelle (v.1.7.1) [18] (Parameters: -
 204 F animal_mt -w 121 -R 10 -k 85,95,105,115,125) (Fig. 2). For specimen two, a new
 205 pipeline was designed to produce the mtDNA assembly from the PacBio HiFi long

206 reads (Fig. 2). The PacBio HiFi mtDNA reads, previously filtered (see above), were
207 corrected using Hifiasm (v.0.13-r308) [19] (Parameters: --write-ec). Since Hifiasm is
208 not optimized to assemble circular molecules (expected for the mtDNA), the corrected
209 PacBio HiFi mtDNA reads were assembled using Unicycler (v.0.4.8.) (Parameters:
210 Defaults)[20], a software designed to assemble bacterial genomes and therefore
211 optimized for circular assemblies. Annotation and visual representation of both
212 mtDNA assemblies were produced using MitoZ (v.2.3) [21] (Parameters: --
213 genetic_code 2; --clade Chordata; --topology circular), using the PE reads for
214 coverage plotting. Furthermore, annotations were manually validated by comparison
215 with other mitochondrial genomes of the genus *Scomber* available at NCBI.

216 **Nuclear genome assembly and assessment**

217 For whole-genome assembly, a combined approach using short and long-read
218 assemblies was applied (Fig. 2). While the long-read assemblies were mainly used to
219 produce the primary assembly, the short-read assemblies were used to scaffold and
220 improve the contiguity of the basal assembly. In summary, the short-read assemblies
221 were performed with the W2RAP pipeline (v.0.1)[22], following the authors' protocol
222 (<https://github.com/bioinfologics/w2rap>). First the Kmer analyses toolkit (KAT)
223 (v.2.4.1) [23] software (hist module) was applied to determine the ideal k-mer cut-off,
224 and then the W2RAP (Parameters: -t 30; -m 500; --min_freq 14; -d 32; --dump_all 1; -
225 k: 144, 180, 200, 224), was used to produce four assemblies (Fig. 2). To generate the
226 long-read assembly, multiple software and parameters were initially tested. The
227 PacBio HiFi reads were assembled in Hifiasm (v.0.13-r308) [19] with different
228 settings (Parameters : k=21, 25, 31, 41, 45, 51, ; l=0, 2) and HiCanu (v.2.1.1) [24]
229 (Parameters: Defaults). While Hifiasm generated 2 pseudo-haplotypes per assembly,
230 the HiCanu generated one merged assembly. To choose the “best” assembly we
231 applied a series of analyses, including Bandage (v.0.8.1) [25] and manual inspection;
232 Benchmarking Universal Single-Copy Orthologs (BUSCO) (v.5.2.2) [26] with
233 Eukaryota and Actinopterygii databases to assess the gene completeness of the
234 assemblies, and Quality Assessment Tool for Genome Assemblies (QUAST) (v.5.0.2)
235 [27] to check the general metrics of the assemblies (Fig. 2). Due to the discrepancies
236 in the length of the Hifiasm primary and alternative pseudo-haplotypes, we opted to
237 concatenate both in a single assembly. At this point, the assembly with the highest
238 complete BUSCO scores, highest contiguity (N50), and longest contig was selected

239 for further analyses. The pseudo-haplotypes were separated by purge_dups (v.1.2.5)
240 [28]. After the first round of purging and inspection by k-mer plot, produced by the
241 KAT tool, the cut-offs were manually adjusted. To assess the influence of purge_dups
242 in the genome, BUSCO (rate of deduplicates) and QUASt (N50 and genomic length
243 per pseudo-haplotype) were used. Next, to improve the contiguity/quality of the
244 assembly, the short read assemblies were used to structurally scaffold the assembly,
245 without the introduction of any new bases in the assembly similarly to [29, 30] (Fig.
246 2). The four short read assemblies were inputted to the Long Interval Nucleotide K-
247 mer Scaffolder (LINKS) (v.1.8.7) [31] (being used as long reads), and using several
248 distance values, i.e., -d 0.5,1.5,3,9,27 kb, the primary assembly was re-scaffolded
249 interactively for 5 rounds (Additional Parameters: -k 21 -e 0.5). Furthermore, the
250 scaffolded genome and the long-read assemblies, initially produced by Hifiasm and
251 HiCanu and discarded based on contiguity and completeness, were inputted to
252 Cobbler (v.0.6.1) [32] and RAILS (v.1.5.1) [32] pipeline (Parameters: Defaults). This
253 allowed gap-filling of ambiguity regions (produced by short-reads scaffolding) and
254 further re-scaffolding using long read information. To evaluate the final assembly,
255 several metrics and software were used. In addition to BUSCO and QUASt metrics,
256 read back mapping of PE read with Burrows-Wheeler Aligner (BWA) (v.0.7.17-
257 r1198) [33], long reads with Minimap2 (v.2.17) [34] and RNA-Seq with Hisat2
258 (v.2.2.0) [35, 36] were also applied. To check the consensus quality (QV) and k-mer
259 completeness we used Merqury [37] (v.1.1) (Fig. 2).

260 **Repeat masking, gene prediction, and annotation**

261 The repetitive elements of the genome were predicted and masked by RepeatMasker
262 (v.4.0.7) [38] using homologous comparisons and *ab initio* predictions. First, the *de*
263 *novo* library of repetitive elements was created with the RepeatModeler (v.2.0.1) [39].
264 Next, *ab initio* library, as well as the Dfam_consensus-20170127 [40] and RepBase-
265 20181026 [41], were used in RepeatMaker to softmask the *S. colias* genome
266 assembly.

267 The genome annotation was performed with the BRAKER2 pipeline (v.2.1.6) [42–
268 44]. Initially, the liver RNAseq reads (Accession number: SRR6367407 [10]) were
269 downloaded, mapped against the *S. colias* genome assembly using Hisat2 (v.2.2.0)
270 [35, 36] (Parameters: Defaults) and converted to bam and sorted files using the

271 SAMtools (v.1.9) [45]. Additionally, we collected 89 proteomes from NCBI RefSeq
272 [46] and Ensembl [47] databases (Suppl. Table 2). Of these, 82 species belong to the
273 Actinopterygii class (32 taxonomic orders), 81 with genome assembly at chromosome
274 level and one at scaffold level. At the date of this genome annotation, only one
275 Scombriforme genome, *Thunnus orientalis*, was annotated (at scaffold level). The
276 remaining seven proteomes were selected from well-established or phylogenetically
277 early-branching species, *Callorhinchus milii*, *Amblyraja radiata*, *Scyliorhinus*
278 *canicula*, *Lepisosteus oculatus*, *Petromyzon marinus*, *Mus musculus*, *Homo sapiens*.
279 Next, the RNAseq alignment, as well as all the above-mentioned proteomes, were
280 inputted in the BRAKER2 pipeline (Parameters: `-etpmode`; `-softmasking`; `-UTR=off`;
281 `-crf`; `-cores=30`). The final file of predictions (braker.gtf) was further filtered by
282 evidence, keeping only gene predictions with RNAseq or protein evidence (using
283 BRAKER2 auxiliary scripts; `selectSupportedSubsets.py`), converted to .gff3 format
284 (using the Augustus auxiliary scripts; `gtf2gff.pl`), and post-processed with Another
285 Gtf/Gff Analysis Toolkit (AGAT) (v.0.6.0). The post-processing stage involved the
286 correction of overlapping gene prediction coordinates and the removal of small or
287 incomplete protein-coding genes (<100 aa; without start/stop codon or both).
288 Furthermore, the proteins were extracted with AGAT tool and functional annotated
289 using InterProScan v.5.44.80 [48] and blast-p searches against RefSeq [46]
290 (Download at 15/05/2021) and SwissProt [49] (Download at 15/05/2021) databases.
291 The homology searches were performed with DIAMOND (v.2.0.11.149) [50]
292 (Parameters: `-k 1`, `-b 10`, `-e 1e-5`, `--ultra-sensitive`, `--outfmt 6`). Finally, the genome as
293 well the annotation datasets were integrated into a website using JBrowse2 [51], a
294 dynamic web platform for genome visualization and analysis that allows easy and
295 interactive exploration of the data provided. The FASTA file containing the genome
296 was indexed with SAMtools Faidx (v.1.9) [45] and added to the JBrowse component,
297 along with the annotation file sorted with “GenomeTools” (v.1.6.1) [52] and indexed
298 with SAMtools Tabix (v.1.9) [53]. In addition to the JBrowse component, ncbi-blast+
299 (v.2.12.0) [54] was integrated into the webpage allowing the blasted results from the
300 genome, mRNA, protein-coding sequences (CDS), and proteins directly from the
301 website (portugalfishomics.ccimar.up.pt/scombercolias).

302 **Phylogenomics**

303 To generate a phylogenomic analysis, the proteomes of 15 selected Actinopterygii
304 species, including the Scombriformes species *Thunnus maccoyii* and *T. orientalis*,
305 were downloaded from public databases (Suppl. Table 2). Single-copy orthologs
306 between these 15 species and *S. colias* were retrieved from the protein datasets by
307 constructing protein family clusters using OrthoFinder (v.2.4.0) [55] (Parameters: -
308 M). This resulted in a total of 392 single copy orthologous sequences that were
309 individually aligned using MUSCLE (v.3.8.31) [56] (Parameters: Defaults). Each
310 alignment was trimmed using TrimAl (v.1.2) [57] with a gap threshold of 0.5
311 (Parameters: -gt 0.5) and afterward concatenated using FASconCAT-G
312 (<https://github.com/PatrickKueck/FASconCAT-G>). Phylogenetic inferences were
313 conducted in IQ-Tree (v.1.6.12) [58] (Parameters: -bb 10000 -nt AUTO -st AA). The
314 best fitted molecular evolutionary model used in the phylogenetic analyses was
315 JTT+F+R4, which was selected by ModelFinder [59] implemented within IQ-Tree.

316 **Assessing the Nuclear Receptor and the “chemical defensible” repertoire in** 317 ***Scomber colias***

318 To collect the repertoire of the nuclear receptors (NRs) in *S. colias* tblast-n (default
319 parameters) searches were performed in the primary genome assembly. The protein
320 sequences of the DNA binding domains and ligand-binding domains of the *Homo*
321 *sapiens* NRs were collected from RefSeq [46] database and used as query
322 (NP_000466.2, NP_068804.1, NP_003241.2, XP_005257609.1, NP_001349802.1,
323 NP_068370.1, NP_599022.1, NP_009052.4, NP_001351014.1, XP_005260464.1,
324 NP_002948.1, NP_001257330.1, NP_003288.2, XP_016862607.1, NP_001273031.1,
325 NP_005645.1, NP_001278159.1, NP_004442.3, NP_000167.1, XP_005268879.1,
326 NP_004950.2, NP_201591.2). Next, the regions aligning with the *H. sapiens*
327 sequences were collected, translated to protein using the Bio.Seq module of biopython
328 (v.1.75) [60], and blasted (blast-p) against a local database containing the NRs
329 proteins of *Danio rerio* (*D. rerio* NRs Database protocol; 1 – NRs sequences and
330 classifications were retrieved from [61]; 2 – The NRs database was build using the
331 makeblastdb application of ncbi-blast+ (v.2.12.0)). For each NRs sequence in *S.*
332 *colias*, the best blast hit in the *D. rerio* database was collected. In some cases, several
333 nuclear receptors of *S. colias* matched the same receptor in *D. rerio*. In these cases,
334 the nucleotide sequences of *S. colias* were re-validated against the NT-NCBI database,
335 and all sequences matching different GeneID’s in the same organism were kept in the

336 final table of NRs. In parallel, and to assess the genome annotation performed by
337 BRAKER2, the genomic coordinates of the regions aligning to *H. sapiens* were
338 searched and identified in the annotation files.

339 To identify the genes related to the “*chemical defensome*”, target genes were selected
340 based on a previous report profiling the “*chemical defensome*” of teleost species [62].
341 Next, gene names were used as queries to search the deduced *S. colias* genome
342 annotation, a simple but successful approach for well-annotated genomes such as *D.*
343 *rerio* [62]. When gene names were not retrieved from *S. colias* genome annotation
344 (i.e. *ftlh*, *gstp*, *hsph*, *maff*, *nme8*, *slc21*), further tblast-n searches were performed
345 (default parameters, except -max_hsps 1 to keep the best query-subject pair) in the
346 primary genome assembly, using *D. rerio* sequences as a query.

347 **Demography with pairwise sequentially Markovian coalescent (PSMC)**

348 To explore the variation in the demographic history of the species, the pairwise
349 sequentially Markovian coalescent (PSMC) strategy was applied [63], following the
350 authors' instructions (<https://github.com/lh3/psmc>). Briefly, the PE short reads were
351 aligned to the repeated masked genome assembly using BWA (v.0.7.17-r1198)
352 (Parameters: bwa mem) [33], and the output converted to bam and sorted using
353 SAMtools (v.1.9) [45] (Function: Sort; Parameter: Default). Next, Picard Tools
354 (v.2.19.2) (<http://broadinstitute.github.io/picard/>) was used to remove duplicate reads
355 (Function: MarkDuplicates; Parameters: Default), and SAMtools used for mapping
356 quality filtering and SNP calling (Function: mpileup; Parameters: -Q 30 -q 30 -C 50).
357 The BCFtools (v.1.9) was applied to extract consensus sequences (Function: call;
358 Parameters: -c) and the subscript vcfutils (from SAMtools) for filtering the output for
359 a minimum depth of 25, a maximum depth of 150, and a min RMS mapQ of 20
360 (Function: vcf2fq; Parameters: -d 25 -D 150 -Q 20). The resulting fastq file was
361 converted to a PSMC compatible input format using fq2psmcfa with a minimum
362 quality threshold of 20 (Parameters: -q 20). Inferences of population history were
363 performed by running PSMC for 25 iterations (Parameters: -N 15, -r 5, -p
364 4*4+13*2+4*4+6) following the recent PSCM estimations on
365 Scombriformes [64]. Furthermore, to account for uncertainties in the PSMC estimates,
366 bootstrapping of 100 replicates was performed using the splitfa script provided by the
367 PSMC authors (<https://github.com/lh3/psmc>). Finally, to scale the demographic
368 estimations, a mutation rate (μ) of 7.3×10^{-9} substitutions/site/generation was used,

369 based on the recent estimation for the Scombriformes species *Thunnus albacares* [64],
370 and a generation time for *S. colias* of 2 years [7, 65].
371

372 **Data Validation**

373 To produce the *S. colias* genome assembly two sequencing strategies were used, i.e.,
374 Illumina PE short reads and PacBio HiFi long reads. The PE dataset was used to
375 assess the genomic proprieties of the *S. colias* species and scaffold the long-read
376 assembly, while the HiFi reads were used to perform the primary genome assembly
377 and the gap closing (Fig. 2).

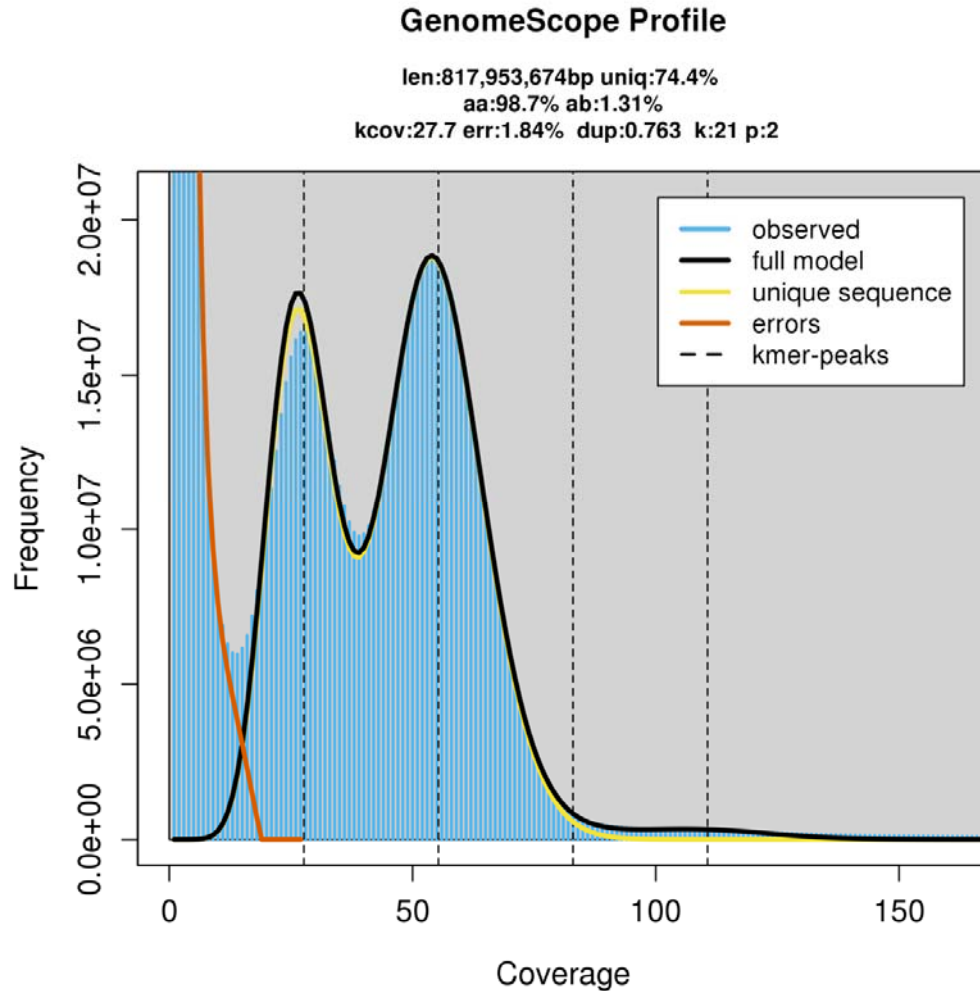
378 The Illumina sequencing yielded 149 M of PE reads and the PacBio sequencing
379 generated 1,7 M of HiFi reads (Table 1). Trimmed short reads were used to estimate
380 the genome size (817 Mb), heterozygosity rate (1.31%), and genome repeat content
381 (approximately 26%) using GenomeScope2 (Fig. 3, Suppl. Table 2). In parallel, the
382 HiFi dataset was inspected and mitochondrial reads, as well as possible sources of
383 contamination, were removed (0,31% of the initial dataset) (Table. 1).

384

385 **Table 1.** General statistics of read datasets used to perform the *Scomber colias*
386 genome assembly.

Sample	Sequencing Type	Library type	Platform	Insert size (bp)	Number of Reads (Before clean-up)	Number of Reads (After clean-up)	Application
Sco_PH	WGS	Long reads	PacBio Sequel II System	15500	1,792,104	1,786,541	Genome Assembly, Gap Closing, Assessment
Sco_PE	WGS	Short reads	HiSeq X Ten	478	149,564,893	84,738,393	Scaffold, Assessment

387



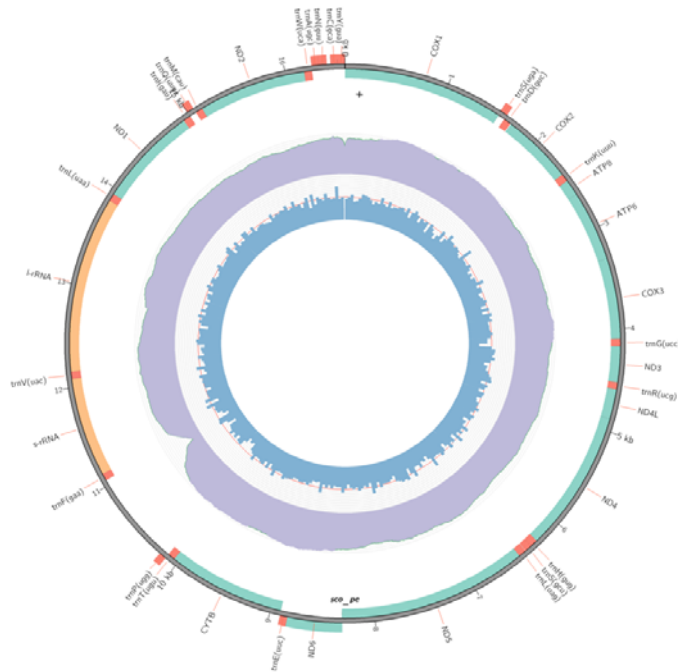
388

389

390

FIGURE 3

391 For the mtDNA assemblies, a total of 38,868 mtDNA PE reads were filtered by
392 GetOrganelle and a total of 792 mtDNA PacBio HiFi reads were filtered by blast-n
393 search. The two assemblies had the same length of 16,570 bp and differed from each
394 other 0.29% (uncorrected p -distances). Furthermore, the PE and PacBio HiFi mtDNA
395 assemblies differed from the *S. colias* mtDNA assembly available on NCBI
396 (Accession number: AB488406.1 [11]), by 0.35% and 0.40% (uncorrected p -
397 distances), respectively. The mtDNA gene content and arrangement is expected for
398 most fishes and the standard for vertebrates [66], consisting of 13 protein-coding
399 genes, 22 transfer RNA (trn), and two ribosomal RNA (rrn) (Fig. 4).



400

401

402

FIGURE 4

403 The primary genome assembly was produced using the filtered PacBio HiFi reads and
404 a combination of several software/settings. Following the above-mentioned criteria
405 (Material and Methods – Nuclear genome assembly and assessment) the Sco_k21
406 assembly was selected (statistics in Suppl. Table 4), both pseudo-haplotypes merged,
407 and subject to purge_dups. Although the purge_dumps generated a primary and an
408 alternative assembly, only the primary was used in the next steps. At the same time,
409 four short-read genome assemblies were performed with the W2RAP software, and
410 the contigs with more than 500bp were used as “long reads” to scaffold the primary
411 assembly (please consult statistics in the Suppl. Table 5). Importantly, during the
412 scaffolding process, only the structural information of short-read assembly was used,
413 without the inclusion of any base. Lastly, the remaining non-basal long read
414 assemblies were used to fill the gaps inserted during the scaffolding stage. The final
415 assembly (primary assembly) of *S. colias* yields a genome size of 814 Mbp,
416 distributed in 2,028 scaffolds and 2,093 contigs with an N50 length of 4,19 and 3,34
417 Mbp, respectively. On the other hand, the alternative assembly had 807 Mbp and
418 5,908 contigs with an N50 length of 0.47 Mbp (Table 2). The BUSCO analyses, at the
419 nucleotide level, in Eukaryota and Actinopterygii datasets showed high levels of
420 completeness for both primary (96.9% and 97.3% of single-copy orthologs) and

421 alternative assemblies (93.3% and 96 % single-copy orthologs) (Table 2).
 422 Consistently, Merqury determined high QV (Primary – 56.53 %; Alternative – 54.99
 423 %) and k-mer completeness (Primary – 86.11%; Alternative – 84.60%) values for both
 424 assemblies (Table 2). In the primary assembly, the k-mer analyses (via Merqury)
 425 showed a low level of k-mer duplication in the genome (color blue, green purple, and
 426 orange in Fig. 5a), indicating a high level of haplotype uniqueness (color red Fig. 5a),
 427 and a similar k-mer distribution pattern to GenomeScope2 (performed with Illumina
 428 PE reads). Additionally, we found a high rate of mappings of the Illumina, PacBio,
 429 and RNA-Seq reads, against the primary assembly of 95 %, 99.8 %, and 90.02 %,
 430 respectively. Overall, these results provide evidence of the high quality of the *S. colias*
 431 genome assembly (Table 2).

432

433 **Table 2.** Statistics of *Scomber colias* genome assembly.

434

Assembly	Alternative	Primary	
	Contigs	Contigs	Scaffolds
Number of contigs (>= 10000 bp)	5908	2093	2028
Number of contigs (>= 50000 bp)	2417	1123	1078
Number of contigs (>= 100000 bp)	1593	704	662
Number of contigs (>= 200000 bp)	1025	456	417
Number of contigs (>= 500000 bp)	421	235	209
Number of contigs (>= 1000000 bp)	123	155	138
Total length (>= 10000 bp)	807,928,680	813976802	814072661
Total length (>= 50000 bp)	721244010	781696683	782480923
Total length (>= 100000 bp)	662374873	751893146	752912084
Total length (>= 200000 bp)	580469606	716806065	718068371
Total length (>= 500000 bp)	385329197	648055626	653890381
Total length (>= 1000000 bp)	180689595	591655104	603146189
Largest contig (Mbp)	3,248	22804600	22804600
Total length (Mbp)	807,936	813,977	814,072
GC (%)	39.94	40.09	40.09
N50 (Mbp)	0,466	3,342	4,190

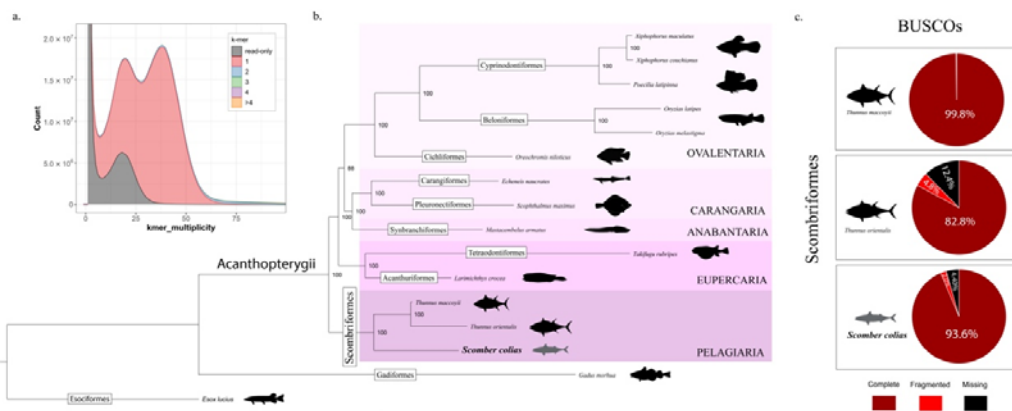
K-mer completeness (%)	84.602	86.1077
Consensus quality	56.5369	54.9969
Read back mapping PE (%)	-	95.0
Read back mapping PH (%)	-	99.8
Read back mapping Rna-Seq (%)	-	90.2
Busco statistics (Databases)	-	
Eukariota**	T:93.3%,C:90.2[S:88.6%,D:1.6%],F:3.1%,M:6.7%,n:255	T:96.9%,C:94.5%[S:92.5%,D:2.0%],F:2.4%,M:3.1%,n:255
Actinopterygii**	T:96.0%,C:94.8%[S:91.9%,D:2.9%],F:1.2%,M:4.0%,n:3640	T:97.3%,C:96.2%[S:94.9%,D:1.3%],F:1.1%,M:2.7%,n:3640

435

436 *The statistics are based on contigs/scaffolds of size ≥ 1000 bp.

437 ** (T: Total BUSCOS Found (completed + fragmented), C: Complete Buscos [S:
438 Complete and single copy, %; D: Complete and duplicated, %], F: Fragmented, %; M:
439 Missing, %; n: Number of sequences in Database)

440



441

442

FIGURE 5

443

444 The RepeatMasker software masked 29.62% of the bases in the primary genome
445 assembly. The major part of the masked regions was linked to DNA elements
446 (11.66%), long interspersed nuclear elements (4.11%), long terminal repeats (2.58%),
447 and simple repeats (2.88%). Furthermore, 8.62 % of the genome was masked and
448 annotated as Unclassified, and only a small percentage was classified as short
449 interspersed nuclear elements, Small RNA or Satellites repeats (Table 3). The genome
450 annotation process generated about 27,675 protein-coding genes and 30,999 protein-

451 coding sequences. On average, we found 9,5 exons and 1,656 bp of length per CDS
 452 (Table 4). 30,355 of the CDS had at least one blast-p hit in Swissprot and/or RefSeq
 453 databases, 27,101 were identified in the InterPro database and 21,664 of these were
 454 classified as belonging to a specific homolog superfamily (Table 5).

455

456 **Table 3.** Report of RepeatMasker software. This report contains the statistics of the
 457 repetitive elements in *Scomber colias* genome assembly.

458

Total number of sequences		2028	
Total length (bp)		814072661 bp	
GC level (%)		40.09	
Number of bases masked		241071029 bp (29.62%)	
Type	Number of elements	Length in Genome	Percentage of Genome
SINEs	16132	2679916	0.33
ALUs	0	0	0.00
MIRs	7082	1280739	0.16
LINEs	113089	33426533	4.11
LINE1	8048	4651362	0.57
LINE2	57670	14551177	1.79
L3/CR1	697	123438	0.02
LTR elements	82410	20969171	2.58
ERV_L	10	279	0.00
ERV_L-MaLRs	0	0	0.00
ERV_classI	22786	4702084	0.58
ERV_classII	11490	576448 bp	0.07
DNA elements	623126	94930706	11.66
hAT-Charlie	27952	5526534	0.68
TcMar-Tigger	169	46619	0.01
Unclassified	278199	70161089	8.62
Total interspersed repeats	-	222167415 bp	27.29
Small RNA	11380	1807250	0.22
Satellites	18552	3093792	0.38
Simple repeats	84014	23465814	2.88
Low complexity	959	200769	0.02

459

460 **Table 4.** Structural annotation report of *Scomber colias* genome assembly.

461

Structural Annotation	Number
Number of genes	27675
Number of mrnas	30999

Number of cdss	30999
Number of exons	295102
Number of introns	264103
Number of exon in cds	295102
Number of intron in cds	264103
Number of intron in exon	264103
Number of intron in intron	235209
Number gene overlapping	71
Number of single exon gene	2036
Number of single exon mrna	2105
mean mrnas per gene	1.1
mean cdss per mrna	1.0
mean exons per mrna	9.5
mean introns per mrna	8.5
mean exons per cds	9.5
mean introns in cdss per mrna	8.5
mean introns in exons per mrna	8.5
mean introns in introns per mrna	7.6
Total gene length	269856447
Total mrna length	310580471
Total cds length	51346678
Total exon length	51346678
Total intron length	259233793
Total intron length per cds	259233793
Total intron length per exon	259233793
Total intron length per intron	35919947
mean gene length	9750
mean mrna length	10019
mean cds length	1656
mean exon length	173
mean intron length	981
mean intron in exon length	981
mean intron in intron length	152
Longest gene	242447
Longest mrna	242447
Longest cds	98436
Longest exon	14939
Longest intron	76003
Longest cds piece	14939
Shortest gene	303
Shortest mrna	144
Shortest cds	18
Shortest exon	3
Shortest intron	30

463 **Table 5.** Functional annotation report of *S. colias* genome assembly.

464

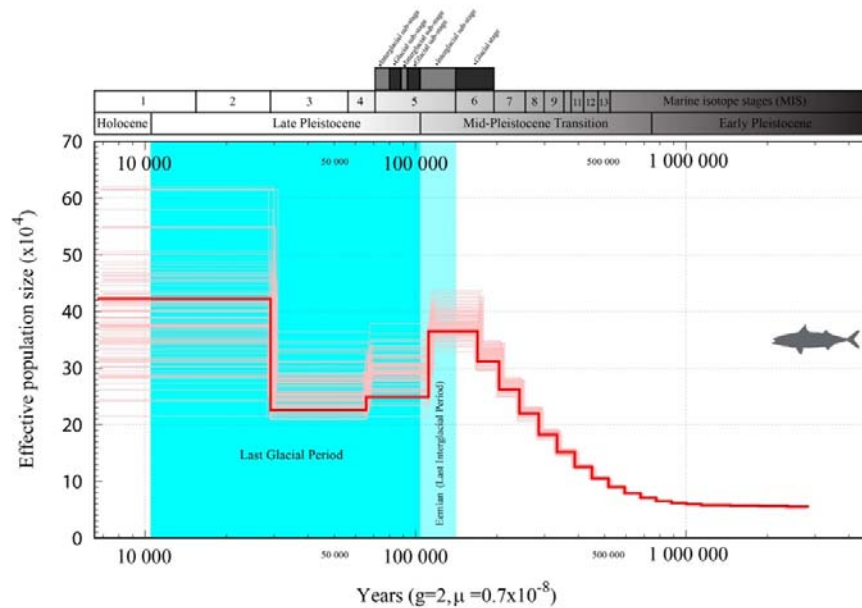
Functional Annotation	Number
Swissprot / RefSeq	30355
InterPro	27101
CDD	12832
Coils	7705
GO	18643
Gene3D	22209
Hamap	463
KEGG	1402
MetaCyc	1140
MobiDBlite	16765
PIRSF	1755
PRINTS	7143
Pfam	25708
ProSitePatterns	8082
PrositeProfiles	16229
Reactome	7376
SFDL	114
SMART	14906
SUPERFAMILY	21664
TIGRFAM	1427

465

466 To validate the protein-coding sequences we performed phylogenetic (via
467 OrthoFinder) and BUSCO analyses (using Actinopterygii library profile) (Fig. 4 b, c).
468 Of the 16 Actinopterygii proteins datasets imputed to OrthoFinder, 98.3% were
469 assigned to 29,066 orthogroups, with 12,334 orthogroups present in all species (Suppl.
470 Table 6). Furthermore, a total of 392 single-copy orthologues were retrieved by
471 OrthoFinder and used for the phylogenomic analysis. Alignment, trimming and
472 concatenation of all single-copy orthologues, resulted in a final 120,886 aa long
473 supermatrix alignment that was used for the phylogenomic inference in IQ-Tree. The
474 resulting Maximum Likelihood phylogenetic tree has maximum support for almost all
475 nodes (Fig. 5b). The phylogeny recovered the reciprocal monophyletic
476 Acanthopterygii groups Pelagiaria, Eupercaria, Anabantaria, Carangaria, and
477 Ovalentaria, with Pelagiaria being the basal clade and represented by the three
478 Scombrifomes, including *S. colias* (Fig. 5b). These results are in accordance with the
479 most recent phylogenomic study of ray-finned fishes [67], as well as the Ensembl
480 Compara Species Tree (<https://m.ensembl.org/info/about/speciestree.html>). The

481 BUSCOs analysis showed the *S. colias* proteome with 93.6% of the groups complete,
482 2% fragmented, and 4.4% missing (Fig. 5c). In comparison, *T. maccoyii* had 99,8 %
483 BUSCO groups complete while *T. orientalis* had only 82,8 %. These results are
484 expected, since the *T. maccoyii* genome assembly, part of the Vertebrate Genome
485 Project [68], was built at chromosome level with multiple technologies (including 46x
486 PacBio data, 46x 10X Genomics Chromium data, BioNano data, and Arima Hi-C
487 data) and several manual curation steps [69]. On the other hand, both *T. orientalis* [70]
488 and *S. colias* were built at scaffold level using only short and long read information.
489 We further explored the quality of the annotation by investigating the repertoire of the
490 nuclear receptor (NR) superfamily in the *S. colias* assembly. NRs are critical
491 molecular physiology components, with vital roles in animal physiology and
492 disruption [71]. Moreover, their exact NR gene complement in vertebrate lineages has
493 been shown to vary [61]. We were able to deduce the existence of 76 NRs in the *S.*
494 *colias* genome (Suppl. Table 7), in line with the repertoire described for other teleost
495 species [72]. Among the retrieved NRs we found those that are key components of the
496 “chemical defensome”—an ensemble of related and unrelated genes that protect
497 organisms against chemical stressors, and thus critical under anthropogenic chemical
498 build-up and climate change scenarios —such as the xenobiotic-inducible pregnane X
499 receptor (*pxr*, *nr1i2*) [62, 73]. Subsequent analysis, using gene names, further
500 suggested the presence of gene annotations for the vast majority of the reported
501 members of the teleost “chemical defensome” in *S. colias*, similarly to that described
502 for *D. rerio* [62]. Additional blast searches were performed for a reduced set of genes
503 (*fthl*, *gstp*, *hsph*, *maff*, *nme8*, and *slc21*), uncovering possible homologs for this gene
504 subset, except for a single member of the GST family (*gstp*) (Suppl. Table 8).
505 We additionally validated our dataset by examining the present population structure of
506 the species, since the genome may also provide clues regarding its past demographic
507 history [63]. One popular method to produce these inferences is the pairwise
508 sequentially Markovian coalescent (PSMC) model, here implemented to the *S. colias*
509 final genome assembly. Since PSMC requires an estimation of the genome-wide
510 mutation rate, and since this has never been produced for *S. colias*, we used the
511 recently estimated genome-wide mutation rate of the yellowfin tuna, *T. albacares* of
512 7.3×10^{-9} mutations/site/generation [64]. The results suggest a past population
513 expansion between 160,000 – 115,000 years ago with maximum effective population
514 size (N_e) of 36,000, during the end of the Mid-Pleistocene Transition, corresponding

515 to the Eemian (i.e., the last interglacial period) and the transition between the Marine
516 Isotope Stages (MIS) 6 to 5, (Fig. 6). This population expansion is followed by an
517 apparent decrease in the N_e to around 25.000, at the beginning of the Late Pleistocene,
518 corresponding to the entering of the Last Glacial Period. These results, suggesting the
519 influence of the climatic changes from the Pleistocene glaciation cycles on the N_e , are
520 following other recent studies on Scombriformes, e.g., Pacific Sierra mackerel,
521 *Scomberomorus sierra* [74], and the Indo-Pacific Yellowfin Tuna *T. albacares* [64],
522 as well as in other pelagic marine species, e.g., the killer whale [75].
523



524
525
526
527

FIGURE 6

528 **Reuse potential**

529 This study reports the first genome assembly of Atlantic chub mackerel. *Scomber*
530 *colias* is a valuable marine resource, with a high impact on the fisheries of several
531 countries on the west coast of the Atlantic Ocean and/or the Mediterranean Sea.
532 Ecologically, this species establishes an important link between primary producers
533 and top predators of the marine trophic web. Despite the ecological and economic
534 importance of *S. colias*, few genomic resources are available for this species. Thus,
535 this genome is timely and is expected to contribute to the effective conservation,
536 management, and sustainable exploitation of *S. colias* species in the Anthropocene.
537 Additionally, this genome will be a key tool to decipher biological features of *S.*
538 *colias*, such as population dynamics, physiology, or endocrinology.

539

540 **Data Records**

541 The raw datasets of PacBio HiFi and Illumina sequencing were deposited in the NCBI
542 Sequences Read Archive under the Bioproject: PRJNA769550. Additionally, both
543 primary and alternative pseudo-haplotype assemblies were submitted to NCBI
544 GenBank (Accession number: JAJDFG000000000 and JAJDFH000000000). The
545 mitochondrial genome assemblies and annotations were submitted to GenBank
546 (Accession number: OK501306 and OK501307). The four W2RAP assemblies as well
547 as the genome annotation files were uploaded to figshare online repository
548 (<https://doi.org/10.6084/m9.figshare.17025506.v1>). The genome and annotation
549 datasets also can be interactively explored at the website -
550 <http://portugalfishomics.ciimar.up.pt/app/scombercolias/>.

551

552

553 **Declarations**

554 **List of abbreviations**

555 AGAT: Another Gtf/Gff Analysis Toolkit; BUSCO: Benchmarking Universal Single-
556 Copy Ortholog; BWA: Burrows-Wheeler Aligner; CDS: Coding sequences; DHA:
557 Docosahexaenoic acid; Gbp: gigabase pair(s); gDNA: Genomic DNA; KAT: Kmer
558 analyses toolkit; LINKS: Long Interval Nucleotide K-mer Scaffold; Mbp: megabase
559 pair(s); mtDNA: Mitochondrial genome; NCBI: National Center for Biotechnology
560 Information; NRs: Nuclear receptors; NT-NCBI: Nucleotide database of NCBI;
561 PacBio: Pacific Biosciences; PE: Paired-end; PSMC: Pairwise sequentially Markovian
562 coalescent; QUASt: Quality Assessment Tool for Genome Assemblies; QV:
563 consensus quality value; SMRT: Single Molecule, Real-Time; SNP: Single-nucleotide
564 polymorphism.

565 **Ethical approval**

566 This work has been approved by the CIIMAR ethical committee and by CIIMAR
567 Managing Animal Welfare Body (ORBEA) according to the European Union
568 Directive 2010/63/EU.

569 **Funding**

570 This research was funded by COMPETE 2020, Portugal 2020, and the European
571 Union through the ERDF, grant number 031342, and by FCT through national funds
572 (PTDC/CTA-AMB/31342/2017), and is part of the CIIMAR-lead initiative Portugal-
573 *Fishomics*. FCT - Foundation for Science and Technology supported AMM.
574 (DFA/BD/8069/2020), AGS. (SFRH/BD/137935/2018), AV (DL57/2016), NA
575 (DFA/BD/6218/2020). RRdF. thanks, the Villum Foundation for its funding of the
576 Center for Macroecology, Evolution, and Climate (DNRF96).

577 **Competing interests**

578 The authors declare that they have no competing interests

579 **Author contributions**

580 LFCC designed and conceived this work; MF, RC, and NA collected the samples;
581 AMM, AGS, EF, LFCC wrote the manuscript; AMM, AGS., MF, FC, MS, MD, RdF,
582 RR, AV, and LFCC coordinated and carried out the bioinformatics analyses. All
583 authors read, revised, and approved the final manuscript.

584 **Acknowledgments**

585 Not applicable.

586

587 **References**

- 588 1. Collette BB, Nauer CE. Scombrids of the world. An Annotated and Illustrated
589 Catalogue of Tunas, Mackerels, Bonitos and Related Species Known to Date. FAO
590 Species Catalogue. 1983;2:2–137. <http://www.fao.org/3/ac478e/ac478e00.htm>.
591 Accessed 20 Jan 2020.
- 592 2. Hernández JJC, Ortega ATS, Castro Hernandez JJ, Santana Ortega AT. Synopsis of
593 biological data on the chub mackerel (*Scomber japonicus* Houttuyn, 1782). FAO Fish
594 Synopsis. 2000;157:1–77. [https://agris.fao.org/agris-](https://agris.fao.org/agris-search/search.do?recordID=XF2000393177)
595 [search/search.do?recordID=XF2000393177](https://agris.fao.org/agris-search/search.do?recordID=XF2000393177). Accessed 20 Feb 2020.
- 596 3. Velasco EM, del Arbol J, Baro J, Sobrino I. Age and growth of the Spanish chub
597 mackerel *Scomber colias* off southern Spain: a comparison between samples from the
598 NE Atlantic and the SW Mediterranean. *Rev Biol Mar Oceanogr*. 2011;46:27–34.
- 599 4. Gamito R, Pita C, Teixeira C, Costa MJ, Cabral HN. Trends in landings and
600 vulnerability to climate change in different fleet components in the Portuguese coast.
601 *Fish Res*. 2016;181:93–101.
- 602 5. Karakoltsidis PA, Zotos A, Constantinides SM. Composition of the commercially
603 important mediterranean finfish, crustaceans, and molluscs. *J Food Compos Anal*.
604 1995;8:258–73.
- 605 6. Ferreira I, Gomes-Bispo A, Lourenço H, Matos J, Afonso C, Cardoso C, et al. The
606 chemical composition and lipid profile of the chub mackerel (*Scomber colias*) show a
607 strong seasonal dependence: Contribution to a nutritional evaluation. *Biochimie*.
608 2020;178:181–9.
- 609 7. Carvalho N, Perrotta RG, Isidro E. Age, growth and maturity in the chub mackerel
610 (*Scomber japonicus* Houttuyn, 1782) from the Azores. *Arquipélago Ciências*
611 *Biológicas e Mar*. 2002;19:93–9.
- 612 8. Martins MM, Skagen D, Marques V, Zwolinski J, Silva A. Changes in the
613 abundance and spatial distribution of the Atlantic chub mackerel (*Scomber colias*) in
614 the pelagic ecosystem and fisheries off Portugal. *Sci Mar*. 2013;77:551–63.
- 615 9. Vasconcelos J, Afonso-Dias M, Faria G. Atlantic chub mackerel (*Scomber colias*)
616 spawning season, size and age at first maturity in Madeira waters. *Arquipélago Life*
617 *Mar Sci*. 2012;29:43–51.
- 618 10. Machado AM, Felício M, Fonseca E, da Fonseca RR, Castro LFC. A resource for
619 sustainable management: De novo assembly and annotation of the liver transcriptome
620 of the Atlantic chub mackerel, *Scomber colias*. *Data Br*. 2018;18:276–84.

- 621 11. Catanese G, Manchado M, Infante C. Evolutionary relatedness of mackerels of the
622 genus *Scomber* based on complete mitochondrial genomes: Strong support to the
623 recognition of Atlantic *Scomber colias* and Pacific *Scomber japonicus* as distinct
624 species. *Gene*. 2010;452:35–43.
- 625 12. Rodríguez-Ezpeleta N, Bradbury IR, Mendibil I, Álvarez P, Cotano U, Irigoien X.
626 Population structure of Atlantic mackerel inferred from RAD-seq-derived SNP
627 markers: effects of sequence clustering parameters and hierarchical SNP selection.
628 *Mol Ecol Resour*. 2016;16:991–1001.
- 629 13. Ravi V, Venkatesh B. The divergent genomes of teleosts. *Annu Rev Anim Biosci*.
630 2018;6:47–68.
- 631 14. Formenti G, Theissinger K, Fernandes C, Bista I, Bombarely A, Bleidorn C, et al.
632 The era of reference genomes in conservation genomics. *Trends Ecol Evol*.
633 2022;accepted.
- 634 15. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina
635 sequence data. *Bioinformatics*. 2014;30:2114–20.
- 636 16. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot
637 for reference-free profiling of polyploid genomes. *Nat Commun*. 2020;11:1–10.
- 638 17. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting
639 of occurrences of k-mers. *Bioinformatics*. 2011;27:764–70.
- 640 18. Jin JJ, Yu W Bin, Yang JB, Song Y, Depamphilis CW, Yi TS, et al. GetOrganelle:
641 A fast and versatile toolkit for accurate de novo assembly of organelle genomes.
642 *Genome Biol*. 2020;21:241.
- 643 19. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo
644 assembly using phased assembly graphs with hifiasm. *Nat Methods*. 2021;18:170–5.
- 645 20. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome
646 assemblies from short and long sequencing reads. *PLOS Comput Biol*.
647 2017;13:e1005595.
- 648 21. Meng G, Li Y, Yang C, Liu S. MitoZ: A toolkit for animal mitochondrial genome
649 assembly, annotation and visualization. *Nucleic Acids Res*. 2019;47:63.
- 650 22. Clavijo BJ, Garcia Accinelli G, Wright J, Heavens D, Barr K, Yanes L, et al.
651 W2RAP: A pipeline for high quality, robust assemblies of large complex genomes
652 from short read data. *bioRxiv*. 2017;:110999. doi:10.1101/110999.
- 653 23. Mapleson D, Accinelli GG, Kettleborough G, Wright J, Clavijo BJ. KAT: A K-
654 mer analysis toolkit to quality control NGS datasets and genome assemblies.

- 655 Bioinformatics. 2017;33:574–6.
- 656 24. Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, et al. HiCanu:
657 Accurate assembly of segmental duplications, satellites, and allelic variants from high-
658 fidelity long reads. *Genome Res.* 2020;30:1291–305.
- 659 25. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de
660 novo genome assemblies. *Bioinformatics.* 2015;31:3350–2.
- 661 26. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update:
662 Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic
663 Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol.*
664 2021;38:4647–54.
- 665 27. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUILT: Quality assessment tool
666 for genome assemblies. *Bioinformatics.* 2013;29:1072–5.
- 667 28. Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and
668 removing haplotypic duplication in primary genome assemblies. *Bioinformatics.*
669 2020;36:2896–8.
- 670 29. Jones S, Taylor G, Chan S, Warren R, Hammond S, Bilobram S, et al. The
671 Genome of the Beluga Whale (*Delphinapterus leucas*). *Genes (Basel).* 2017;8:378.
- 672 30. Taylor GA, Kirk H, Coombe L, Jackman SD, Chu J, Tse K, et al. The Genome of
673 the North American Brown Bear or Grizzly: *Ursus arctos ssp. horribilis*. *Genes*
674 (Basel). 2018;9:598.
- 675 31. Warren RL, Yang C, Vandervalk BP, Behsaz B, Lagman A, Jones SJM, et al.
676 LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads.
677 *Gigascience.* 2015;4:35.
- 678 32. Warren RL. RAILS and Cobbler: Scaffolding and automated finishing of draft
679 genomes using long DNA sequences. *J Open Source Softw.* 2016;1:116.
- 680 33. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler
681 transform. *Bioinformatics.* 2010;26:589–95.
- 682 34. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.*
683 2018;34:3094–100.
- 684 35. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome
685 alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol.*
686 2019;37:907–15.
- 687 36. Kim D, Langmead B, Salzberg SL. HISAT: A fast spliced aligner with low
688 memory requirements. *Nat Methods.* 2015;12:357–60.

- 689 37. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: Reference-free quality,
690 completeness, and phasing assessment for genome assemblies. *Genome Biol.*
691 2020;21:245.
- 692 38. Chen N. Using Repeat Masker to identify repetitive elements in genomic
693 sequences. *Curr Protoc Bioinforma.* 2004;5:4–10.
- 694 39. Smit AFA, Hubley R. RepeatModeler Open-1.0. <http://www.repeatmasker.org>.
- 695 40. Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, et al. The Dfam
696 database of repetitive DNA families. *Nucleic Acids Res.* 2016;44:D81–9.
- 697 41. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements
698 in eukaryotic genomes. *Mob DNA.* 2015;6:1–6.
- 699 42. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1:
700 Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and
701 AUGUSTUS. *Bioinformatics.* 2015;32:767–9.
- 702 43. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-genome annotation with
703 BRAKER. *Methods Mol Biol.* 2019;1962:65–95.
- 704 44. Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic
705 eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a
706 protein database. *NAR Genomics Bioinforma.* 2021;3:lqaa108.
- 707 45. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
708 Alignment/Map format and SAMtools. *Bioinformatics.* 2009;25:2078–9.
- 709 46. O’Leary NA, Wright MW, Brister JR, Ciuffo S, Haddad D, McVeigh R, et al.
710 Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion,
711 and functional annotation. *Nucleic Acids Res.* 2016;44:D733–45.
- 712 47. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al.
713 Ensembl 2020. *Nucleic Acids Res.* 2020;48:D682–8.
- 714 48. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5:
715 Genome-scale protein function classification. *Bioinformatics.* 2014;30:1236–40.
- 716 49. Bateman A, Martin MJ, Orchard S, Magrane M, Agivetova R, Ahmad S, et al.
717 UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.*
718 2021;49:D480–9.
- 719 50. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using
720 DIAMOND. *Nature Methods.* 2014;12:59–60.
- 721 51. Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, et al. JBrowse:
722 A dynamic web platform for genome visualization and analysis. *Genome Biol.*

- 723 2016;17:66.
- 724 52. Gremme G, Steinbiss S, Kurtz S. Genome tools: A comprehensive software library
725 for efficient processing of structured genome annotations. *IEEE/ACM Trans Comput*
726 *Biol Bioinforma.* 2013;10:645–56.
- 727 53. Li H. Tabix: Fast retrieval of sequence features from generic TAB-delimited files.
728 *Bioinformatics.* 2011;27:718–9.
- 729 54. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA
730 sequences. *J Comput Biol.* 2000;7:203–14.
- 731 55. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome
732 comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.*
733 2015;16:157.
- 734 56. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high
735 throughput. *Nucleic Acids Res.* 2004;32:1792–7.
- 736 57. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T, Capella-Gutiérrez S, Silla-
737 Martínez JM, Gabaldón T. trimAl: A tool for automated alignment trimming in large-
738 scale phylogenetic analyses. *Bioinformatics.* 2009;25:1972–3.
- 739 58. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and
740 Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.
741 *Mol Biol Evol.* 2015;32:268–74.
- 742 59. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermini LS.
743 ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat Methods.*
744 2017;14:587–9.
- 745 60. Chapman B, Chang J. Biopython. *ACM SIGBIO Newsl.* 2000;20:15–9.
- 746 61. Fonseca E, Machado AM, Vilas-Arrondo N, Gomes-dos-Santos A, Veríssimo A,
747 Esteves P, et al. Cartilaginous fishes offer unique insights into the evolution of the
748 nuclear receptor gene repertoire in gnathostomes. *Gen Comp Endocrinol.*
749 2020;295:113527.
- 750 62. Eide M, Zhang X, Karlsen OA, Goldstone J V., Stegeman J, Jonassen I, et al. The
751 chemical defensome of five model teleost fish. *Sci Rep.* 2021;11:1–13.
- 752 63. Li H, Durbin R. Inference of human population history from individual whole-
753 genome sequences. *Nature.* 2011;475:493–6.
- 754 64. Barth JMI, Damerau M, Matschiner M, Jentoft S, Hanel R. Genomic
755 Differentiation and Demographic Histories of Atlantic and Indo-Pacific Yellowfin
756 Tuna (*Thunnus albacares*) Populations. *Genome Biol Evol.* 2017;9:1084–98.

- 757 65. Martins MM. Growth variability in Atlantic mackerel (*Scomber scombrus*) and
758 Spanish mackerel (*Scomber japonicus*) off Portugal. *ICES J Mar Sci.* 2007;64:1785–
759 90.
- 760 66. Satoh TP, Miya M, Mabuchi K, Nishida M. Structure and variation of the
761 mitochondrial genome of fishes. *BMC Genomics.* 2016;17:719.
- 762 67. Hughes LC, Ortí G, Huang Y, Sun Y, Baldwin CC, Thompson AW, et al.
763 Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on
764 transcriptomic and genomic data. *Proc Natl Acad Sci.* 2018;115:6249–54.
- 765 68. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards
766 complete and error-free genome assemblies of all vertebrate species. *Nature.*
767 2021;592:737–46.
- 768 69. Howe K, Chow W, Collins J, Pelan S, Pointon DL, Sims Y, et al. Significantly
769 improving the quality of genome assemblies through curation. *GigaScience.*
770 2021;10:1–9.
- 771 70. Suda A, Nishiki I, Iwasaki Y, Matsuura A, Akita T, Suzuki N, et al. Improvement
772 of the Pacific bluefin tuna (*Thunnus orientalis*) reference genome and development of
773 male-specific DNA markers. *Sci Rep.* 2019;9:1–12.
- 774 71. Santos MM, Ruivo R, Capitão A, Fonseca E, Castro LFC. Identifying the gaps:
775 Resources and perspectives on the use of nuclear receptor based-assays to improve
776 hazard assessment of emerging contaminants. *J Hazard Mater.* 2018;358:508–11.
- 777 72. Bertrand S, Thisse B, Tavares R, Sachs L, Chaumot A, Bardet PL, et al.
778 Unexpected novel relational links uncovered by extensive developmental profiling of
779 nuclear receptor expression. *PLoS Genet.* 2007;3:2085–100.
- 780 73. Eide M, Rydbeck H, Tørresen OK, Lille-Langøy R, Puntervoll P, Goldstone J V.,
781 et al. Independent losses of a xenobiotic receptor across teleost evolution. *Sci Rep.*
782 2018;8:1–13.
- 783 74. López MD, Alcocer MU, Jaimes PD. Phylogeography and historical demography
784 of the Pacific Sierra mackerel (*Scomberomorus sierra*) in the Eastern Pacific. *BMC*
785 *Genet.* 2010;11:34.
- 786 75. Moura AE, Van Rensburg CJ, Pilot M, Tehrani A, Best PB, Thornton M, et al.
787 Killer whale nuclear genome and mtDNA reveal widespread population bottleneck
788 during the last glacial maximum. *Mol Biol Evol.* 2014;31:1121–31.
- 789
790

791 **Figure legends**

792 **Figure 1:** Photograph of Atlantic chub mackerel, *Scomber colias* (Specimen caught in
793 2020 and used to do the Pacbio HiFi genome assembly).

794 **Figure 2:** Bioinformatics workflow used to perform the genome assembly of *Scomber*
795 *Colias* species.

796 **Figure 3:** Genomescope2 plot with K-mer spectra content and fitted models of the
797 *Scomber colias* Illumina PE dataset.

798 **Figure 4:** Circular mitochondrial genome assembly of *Scomber colias*, obtained from
799 the Illumina PE dataset (equal to that obtained from the PacBio HiFi long reads
800 assembly). From the center to the outmost features: the GC content distribution;
801 sequencing depth distribution of aligned Paired-End reads; and gene elements (i.e,
802 PCGs, rRNA genes, tRNA genes)

803 **Figure 5:** Validation of the genome assembly and annotation process. a. K-mer
804 analyses of the *Scomber colias* genome assembly (Merqury). b. Maximum Likelihood
805 phylogenetic tree based on the concatenated alignments of amino acid sequences of
806 392 single-copy orthologs retrieved by OrthoFinder. Bootstrap values are shown next
807 to the nodes. c. BUSCOs scores were obtained from searching the proteomes of the
808 three Scombriformes species with genome annotation available, against the
809 actinopterygii_odb10 (n:3640) lineage.

810 **Figure 6:** Pairwise sequentially Markovian coalescent (PSMC) estimates from the
811 *Scomber colias* genome assembly. Estimations were obtained using a generation time
812 of 2 years and genome-wide mutation rate of 7.3×10^{-9} mutations/site/generation.
813 Effective population size (N_e) is presented in the left vertical axis and changes
814 estimated since the present over the last 3 myr in the horizontal axis.

815