

1 **Large scale gene duplication affected the European eel (*Anguilla***  
2 ***anguilla*) after the 3R teleost duplication**

3  
4 Christoffer Rozenfeld<sup>1\*</sup>, Jose Blanca<sup>2\*</sup>, Victor Gallego<sup>1</sup>, Víctor García-Carpintero<sup>2</sup>, Juan Germán  
5 Herranz-Jusdado<sup>1</sup>, Luz Pérez<sup>1</sup>, Juan F. Asturiano<sup>1▲</sup>, Joaquín Cañizares<sup>2†</sup>, David S.  
6 Peñaranda<sup>1†</sup>

7  
8 1 Grupo de Acuicultura y Biodiversidad. Instituto de Ciencia y Tecnología Animal. Universitat  
9 Politècnica de València. Camino de Vera s/n, 46022 Valencia, Spain

10 2 Instituto de Conservación y Mejora de la Agrodiversidad Valenciana, Universitat Politècnica  
11 de València, Camino de Vera 14, 46022, Valencia, Spain.

12  
13 Authors marked with \* or † contributed equally to this work, respectively.

14  
15  
16  
17 Running title: Evidence of European eel large scale gene duplication

18  
19 Keywords: European eel, PHYLOG, 4dTv, whole genome duplication

20  
21  
22  
23  
24  
25 ▲ Corresponding author:

26  
27 Dr. Juan F. Asturiano  
28 Grupo de Acuicultura y Biodiversidad  
29 Instituto de Ciencia y Tecnología Animal  
30 Universitat Politècnica de València  
31 Camino de Vera s/n 46022 Valencia (Spain)  
32 E-mail: [jfastu@dca.upv.es](mailto:jfastu@dca.upv.es)

33  
34

35

36 **Abstract**

37 Genomic scale duplication of genes generates raw genetic material, which may facilitate new  
38 adaptations for the organism. Previous studies on eels have reported specific gene duplications,  
39 however a species-specific large-scale gene duplication has never before been proposed. In  
40 this study, we have assembled a *de novo* European eel transcriptome and the data show more  
41 than a thousand gene duplications that happened, according to a 4dTv analysis, after the  
42 teleost specific 3R whole genome duplication (WGD). The European eel has a complex and  
43 peculiar life cycle, which involves extensive migration, drastic habitat changes and  
44 metamorphoses, all of which could have been facilitated by the genes derived from this large-  
45 scale gene duplication.

46 Of the paralogs created, those with a lower genetic distance are mostly found in tandem  
47 repeats, indicating that they are young segmental duplications. The older eel paralogs showed a  
48 different pattern, with more extensive synteny suggesting that a Whole Genome Duplication  
49 (WGD) event may have happened in the eel lineage. Furthermore, an enrichment analysis of  
50 eel specific paralogs further revealed GO-terms typically enriched after a WGD. Thus, this  
51 study, to the best of our knowledge, is the first to present evidence indicating an Anguillidae  
52 family specific large-scale gene duplication, which may include a 4R WGD.

53

54

55

56

57

58

59

60

61

## 62 **Introduction**

63 Large-scale gene duplications can originate from one single event, like a whole genome  
64 duplication (WGD; Ohno, 1970) or from multiple smaller segmental duplication events (SDs; Gu  
65 et al. 2002). Any of these duplication events may contribute to species radiation, since both  
66 provide raw material for new genetic variation (Cañestro et al. 2013; Gu et al. 2002; Ohno  
67 1970). It has been suggested that early in the vertebrate lineage two WGDs (1R and 2R)  
68 happened, resulting in species radiation and evolution of new traits (Cañestro et al. 2013; Dehal  
69 and Boore 2005; Gu et al. 2002; Ohno 1970). In teleosts, there is strong evidences to support  
70 an additional WGD, called the 3<sup>rd</sup> teleost specific WGD (3R), which occurred in the base of the  
71 teleost lineage, between 350 and 320 million years ago (MYA; Aparicio et al. 2002; Christoffels  
72 et al. 2004; Howe et al. 2013; Vandepoele et al. 2004; Jaillon et al. 2004; Kasahara et al. 2007;  
73 Meyer and Peer 2005; Schartl et al. 2013). Previous studies have proposed that this extra 3R  
74 WGD is one of the possible causes of the massive species radiation observed in teleosts  
75 (Hoegg et al. 2004; Santini et al. 2009). In addition to 3R, multiple genus or species specific  
76 WGDs have been documented in teleosts, e.g. in salmonids (order Salmoniformes; Allendorf  
77 and Thorgaard, 1984; Johnson et al. 1987), sturgeons (order Acipenseriform; Ludwig et al.  
78 2001), common carp (*Cyprinus carpio*; Larhammar and Risinger, 1994), goldfish (*Carassius*  
79 *auratus*; Ohno, 1970), suckers (family Catostomidae; Uyeno and Smith, 1972), and loaches  
80 (*Botia macracantha* and *Botia modesta*; Ferris and Whitt 1977).

81 As mentioned previously, other mechanisms to WGDs can create large-scale gene duplications.  
82 Several species have shown a high occurrence of relatively recent segmental duplications (SD),  
83 often found in tandem, with segments spanning from a few hundred base pairs to several genes  
84 e.g. in yeast (Llorente et al. 2000), daphnia (Colbourne et al. 2011), humans (Bailey et al. 2002;  
85 Gu et al. 2002; Vallente Samonte and Eichler 2016) and teleosts (Blomme et al. 2006; David et  
86 al. 2003; Jaillon et al. 2004; Lu et al. 2012; Rondeau et al. 2014). It is quite common for one of

87 the copies of these SDs to get lost over time, possibly due to genetic drift or purifying selection.  
88 As a consequence, the genetic distance between two copies often tends to be quite small  
89 (Ohno, 1970). This process is known as the continuous mode hypothesis (Gu et al. 2002). In  
90 some cases however, these SDs have been conserved in high frequency at particular times,  
91 e.g. in yeast (Llorente et al. 2000), common carp (David et al. 2003) and humans (Asrar et al.  
92 2013; Bailey et al. 2002; Gu et al. 2002; Hafeez et al. 2016). Some mechanisms, which could be  
93 facilitating this conservation include the processes of subfunctionalization, neofunctionalization  
94 or dosage selection (for review see Zhang, 2003). Furthermore, these processes have also  
95 been associated with the adaptation to new environments (Colbourne et al. 2011; Tautz and  
96 Domazet-lošo 2011).

97 The elopomorpha cohort, is one of the most basal teleost groups (Greenwood et al. 1966; Inoue  
98 et al. 2004). Elopomorphas are believed to originally be a marine species however, the 19  
99 species of Anguillidae family, broke away from the ancestral trait and adapted a catadromous  
100 life style, migrating from their feeding grounds in freshwater rivers and lakes to their marine  
101 spawning grounds (Inoue et al. 2010; Munk et al. 2010; Schmidt 1923; Tsukamoto Katsumi,  
102 Nakai Izumi 1998). It is likely that species of the Anguillidae family originally performed relatively  
103 short reproductive migrations however, due to continental drift (Inoue et al. 2010; Tsukamoto et  
104 al. 2002) or changes in oceanic currents these migrations have since become vastly extensive  
105 (Jacobsen et al. 2014), with a total migrating distance of >6.000 km in the case of the European  
106 eel (Righton et al. 2016).

107 Several previous studies have revealed a high occurrence of duplicated genes in eels (Dufour et  
108 al. 2005; Henkel et al. 2012; Lafont et al. 2016; Maugars and Dufour 2015; Morini et al. 2015;  
109 Pasqualini et al. 2009; Pasquier et al. 2012; Rozenfeld et al. 2016; Morini et al. 2017). E.g.  
110 Lafont et al. (2016) found two paralog genes of *ift140*, *tleo2*, *nme4*, *xpo6*, and *unkl*, in the *gper*  
111 genomic regions of the eel. Only one copy of these genes has been observed in other teleosts.  
112 These results led Lafont et al. (2016) to hypothesize i) that the whole region containing *gper*

113 could have been duplicated in *Anguilla* eels, and maybe also in other teleosts, and ii) that the  
114 retention of duplicated genes may be higher in eels than in other teleosts.

115 For the present study, we assembled a *de novo* European eel transcriptome from Illumina RNA  
116 sequencing data. In order to study species-specific duplications and the timings of the events  
117 that created them, we ran phylogenetic reconstructions and calculated fourfold synonymous  
118 third-codon transversion (4dTv) distances. This analysis was performed on our transcriptome,  
119 and on multiple other fish transcriptomes and genomes. Our analysis revealed a high  
120 accumulation of duplicated genes in eel compared to other teleost species (which do not have a  
121 confirmed 4R duplication event in their lineage). Many of these duplications are restricted to the  
122 eel lineage, and the 4dTv analyses suggested that these duplications happened much later than  
123 the 3R WGD shared by all teleosts. We will discuss in more depth if these eel specific  
124 duplications are the result of several SDs or one eel-specific 4R WGD. To our knowledge, this is  
125 the first published evidence of a large-scale lineage specific duplication in the elopomorpha  
126 cohort.

127

## 128 **Results**

### 129 *Transcriptome assemblies and genomes*

130 To assemble a *de novo* European eel transcriptome, we performed high quality RNA extractions  
131 from forebrain, pituitary, and testis samples, of one eel, following the protocol described by  
132 Peña-Llopis and Brugarolas (2013). The RNA was then quality tested on the Bio-Rad  
133 Bioanalyser, which yielded average RIN values of 8.90. From this RNA, in total 181 million  
134 Illumina reads, with a length of 101 bp, were produced. These reads were assembled by using  
135 the Trinity assembler after a digital normalization step that left 75 million representative reads.  
136 The transcriptomes of Northern pike (*Esox lucius*), elephantnose fish (*Gnathonemus petersii*)  
137 and silver arowana (*Osteoglossum bicirrhosum*) were also assembled by Trinity using Illumina  
138 reads from the Phylofish database (Pasquier et al. 2016). The resulting unigenes were clustered

139 by using a transitive clustering approach to create sets of very similar transcripts. The number of  
 140 unigenes (henceforth referred to as transcripts) assembled ranged from 68489 to 78610 (table  
 141 1) and the number of transcript clusters from 49154 to 55667 (henceforth referred to as genes;  
 142 table 2).

Table 1 Size and quality of included transcriptomes from: European eel (*Anguilla Anguilla*), Northern Pike (*Esox Lucius*), elephantnose fish (*Gnathonemus petersi*), and silver arowana (*Osteoglossum bicirrhosum*).

Species	N.º Reads	Q30	Transcripts
European eel	181322106	0.994	77247
Northern Pike	553710218	0.989	68489
Elephantnose fish	498451616	0.993	74642
Silver arowana	490649254	0.992	78610

143  
 144

Table 2 Quantities of included genes per included species: European eel (*Anguilla anguilla*), Zebrafish (*Danio rerio*), Northern pike (*Esox Lucius*), Elephantnose fish (*Gnathonemus petersi*), Spotted gar (*Lepisosteus oculatus*), Silver arowana (*Osteoglossum bicirrhosum*), Atlantic salmon (*Salmo salar*), Fugu (*Takifugu rubripes*), and Platyfish (*Xiphophorus maculatus*).

Species	Transcripts	Genes	Representative transcripts	Representative transcripts with predicted protein	Gene family transcripts	% of genes assigned to a gene family
European eel	77247	54879	54845	27696	25862	93.38
Zebrafish	58274	32189	32189	25790	22703	88.03
Northern pike	68489	49154	49154	23843	21696	90.99
Elephantnose fish	74642	50455	50455	24857	22036	88.65
Spotted gar	22483	18341	18341	18341	17872	97.44
Silver arowana	78610	55667	55667	24938	21604	86.63
Atlantic salmon	109584	55104	55104	48593	42625	87.72
Fugu	47841	18523	18523	18523	17698	95.55
Platyfish	20454	20379	20379	20379	19807	97.19

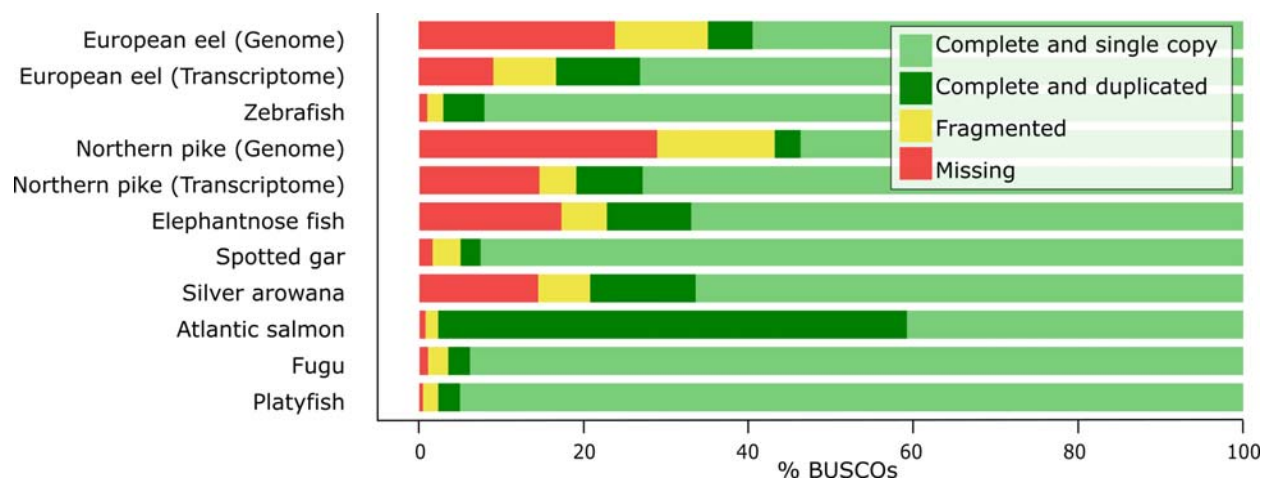
145  
 146 The genomes of zebrafish (*Danio rerio*), northern pike, spotted gar (*Lepisosteus oculatus*), fugu  
 147 (*Takifugu rubripes*), and platyfish (*Xiphophorus maculatus*) were obtained from the ENSEMBL

148 database, the Atlantic salmon (*Salmo salar*) genome was downloaded from NCBI and the  
149 published eel genome was downloaded from the ZF-genomics web site (Henkel et al. 2012).

150

### 151 *Genome and transcriptome quality assessment*

152 In order to test the completeness of the transcriptomes and genomes we ran a BUSCO analysis  
153 in which we looked for a set of single-copy orthologues, typically found in fish genomes (Simão  
154 et al. 2015; Fig. 1). In general, genomes were more complete than transcriptomes according to  
155 the BUSCO assessment, and were thus preferred. However, in the cases of the pike and eel,  
156 the transcriptomes outperformed the genomes (Fig. 1), and these transcriptomes were therefore  
157 used for further analysis.



158

Figure 1

BUSCO (Benchmarking set of Universal Single-Copy Orthologues) result for every genome and transcriptome, one per row. The sequence of a BUSCO gene can be found complete or fragmented in each genome and it can be found once (single copy), more than once (duplicated) or not found (missing). Included genomes: European eel (*Anguilla anguilla*), zebrafish (*Danio rerio*), northern pike (*Esox lucius*), spotted gar (*Lepisosteus oculatus*), fugu (*Takifugu rubripes*), platyfish (*Xiphophorus maculatus*) and Atlantic salmon (*Salmo salar*). Included transcriptomes: European eel, northern pike, elephantnose fish (*Gnathonemus petersii*) and silver arowana (*Osteoglossum bicirrhosum*).

159

160 In order to further test the completeness of the eel and pike transcriptomes, we mapped the eel  
161 and pike RNA-seq reads to the transcriptome assembly using BWA-MEM (Li and Durbin 2010)

162 and to the genome using the software HISAT2 (Pertea et al. 2016). The percentages of reads  
163 that mapped concordantly against the genome and the transcriptome were 65.8 and 91.9%  
164 respectively for eel, and 44.6 and 85.8% for pike. Likewise, previous published eel RNA-  
165 sequencing experiments were also mapped to the eel genome and transcriptome. In this case,  
166 52.2% (Coppe et al. 2010), 57.9% (Burgerhout et al. 2016), and 66.18% (Ager-Wick et al. 2013)  
167 reads mapped concordantly against the eel genome whereas 84.3% (Coppe et al. 2010), 69.5%  
168 (Burgerhout et al. 2016), and 87.32 % (Ager-Wick et al. 2013) mapped against the  
169 transcriptome.

170

#### 171 *Gene families*

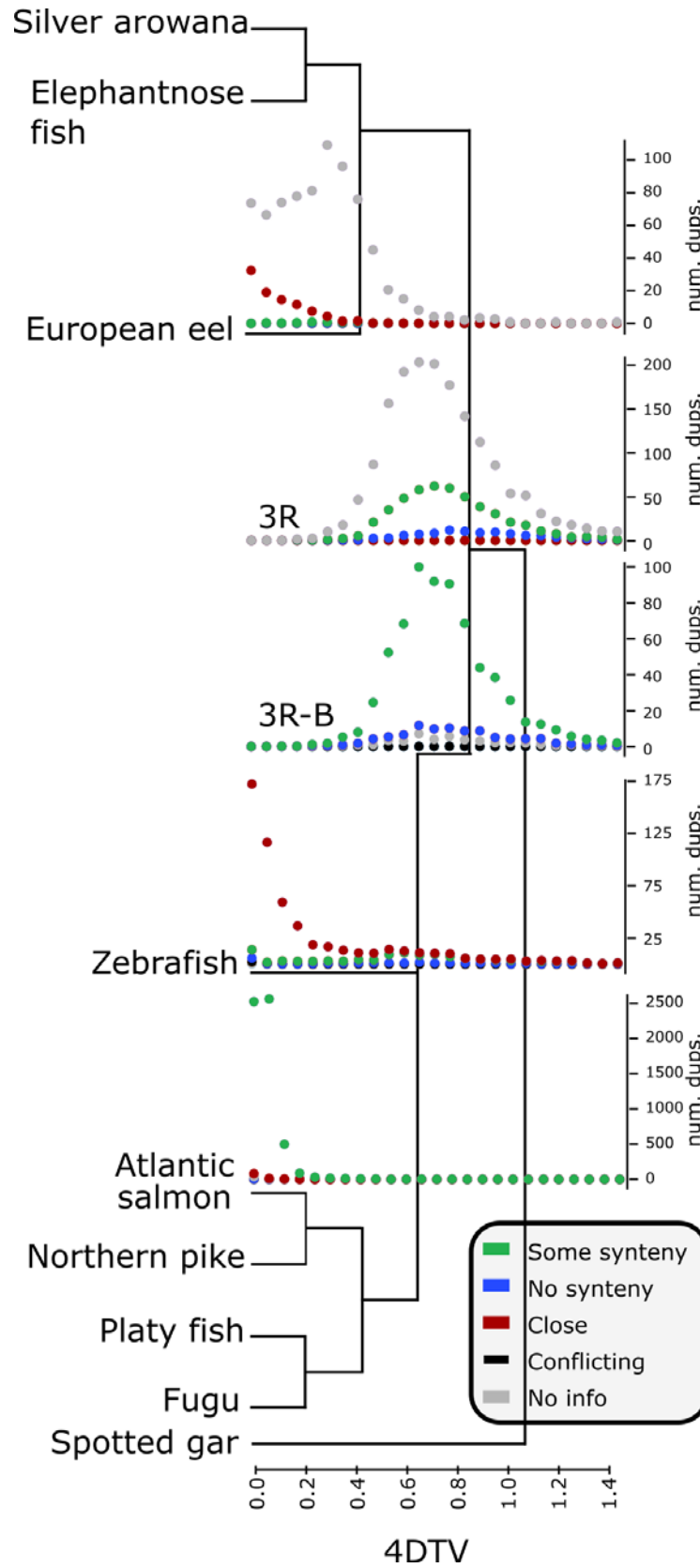
172 One representative transcript for each gene and species was selected; the longest one for  
173 genomes and the most expressed one for transcriptomes. The OrthoMCL web service (Li et al.  
174 2003) assigned gene families to the genes. The percentage of genes assigned to a family  
175 ranged from 88.0% (zebrafish) to 97.4% (spotted gar; table 2). Overall, 17003 gene families  
176 were covered, from which 13823 protein and codon alignments were built. These families  
177 contained between 2 and 161 genes, with 9 genes per family being the mode (suppl. Fig. 1).

178

#### 179 *Phylogenetic reconstruction and duplication dating*

180 PHYLOG (Boussau et al. 2013) was run 10 independent times using 8,000 protein alignments  
181 chosen at random. Overall, PHYLOG created trees for 10,352 gene families and, based on the  
182 tree topology, it labelled the branches in which gene duplication events had happened. All 10  
183 runs produced a species tree that matched the species tree topology created by phylobayes  
184 (Lartillot et al. 2009) with a CAT-GTR model built with the concatenation of 100 protein  
185 alignments and a Neighbour-Joining tree built with a 4dTv distance matrix (Fig. 2).





## Figure 2

Species cladogram generated by PHYLOGENETIC ANALYSIS (PHYLOGENETIC ANALYSIS) for the species included in this study: European eel (*Anguilla anguilla*), zebrafish (*Danio rerio*), northern pike (*Esox lucius*), spotted gar (*Lepisosteus oculatus*), fugu (*Takifugu rubripes*), platyfish (*Xiphophorus maculatus*), Atlantic salmon (*Salmo salar*), elephantnose fish (*Gnathonemus petersii*) and silver arowana (*Osteoglossum bicirrhosum*). PHYLOGENETIC ANALYSIS also determined the duplication events and for each of these events the 4dTv and the synteny type found around the gene was determined. Only the 4dTv distributions for the branches with most duplications are represented over the corresponding cladogram branch, for the distributions for all branches refer to supplementary figure 2. The synteny types are the following: close, the copies originated by the duplication are close in the genome; some synteny, some genes close to the one duplicated are also found to be duplicated close by; no synteny, there are no paralogs for others genes found close to the paralog copies created by the duplication; no information, the duplicated genes are located in small scaffolds with not enough genes close by; conflicting syntenies, different synteny classification found in the genomes of the different species affected by the duplication)

186

187 For each duplication found in each gene family tree, the 4dTv distance between the genes was  
188 calculated, and by grouping them according to the species tree branch in which it happened, the  
189 distribution of the 4dTvs for each lineage was built. Each 4dTv distribution was further divided  
190 according to the synteny type found in the region where the paralogs of each gene family were  
191 located (Fig. 2 and suppl. Fig. 2). The duplications were thus labeled according to the genomic  
192 region where the resulting paralogs were found. In some cases the paralog pairs were found  
193 close to each other (labelled as close), denoting a tandem SD, in other cases they were in  
194 syntenic regions where paralogs from other gene families were also located (labelled as “some  
195 synteny”), possibly denoting a WGD and, finally, in some other cases, there were not enough  
196 close genes (labelled as “no info”) in the genome assembly or conflicting evidence was found  
197 (labelled as “conflicting syntenies”).

198 PHYLOGENETIC ANALYSIS assigned 4,308 duplications to the basal teleost branch, after the split of the spotted  
199 gar (Fig. 2 and Fig. 3), with a 4dTv mode of 0.8. Of the paralogs created by these duplications  
200 63.1% were located in regions with some synteny, 2.4% were close to each other, and 32.3%  
201 had no synteny. These percentages are calculated without taking into account the duplications  
202 where no information regarding the physical location of the genes could be established. The  
203 following branch (directly following the split of the eel, arowana and elephantnose fish) was

204 assigned 1,525 duplications and showed very similar distributions with an overall 4dTv mode of  
 205 0.75. The eel specific branch was assigned 1460 duplications of which 16.5, 75.8, and 7.2%  
 206 were labelled as some synteny, close and without synteny, respectively. Notably, most of the  
 207 eel specific duplications lacked sufficient physical genomic location information. The  
 208 duplications that generated close genes in tandem within the eel genome clearly showed a  
 209 different distribution to the ones located in syntenic regions and the ones with no information.  
 210 The tandem ones tended to be more recent according to their 4dTv (fig. 2) while the syntenic  
 211 ones, and most of the duplications without sufficient genomic location information, showed a  
 212 4dTv mode of 0.4. In both the salmon and zebrafish specific branches most duplications  
 213 seemed quite recent, according to their 4dTv values, with 8,712 and 1,452 branch specific  
 214 duplications, respectively. In the case of the salmon, most duplications (80.6%) were  
 215 characterized by paralogs located in syntenic regions whereas most zebrafish paralogues  
 216 (54.6%) were close tandem SDs (Fig. 2 and Fig. 3).

217

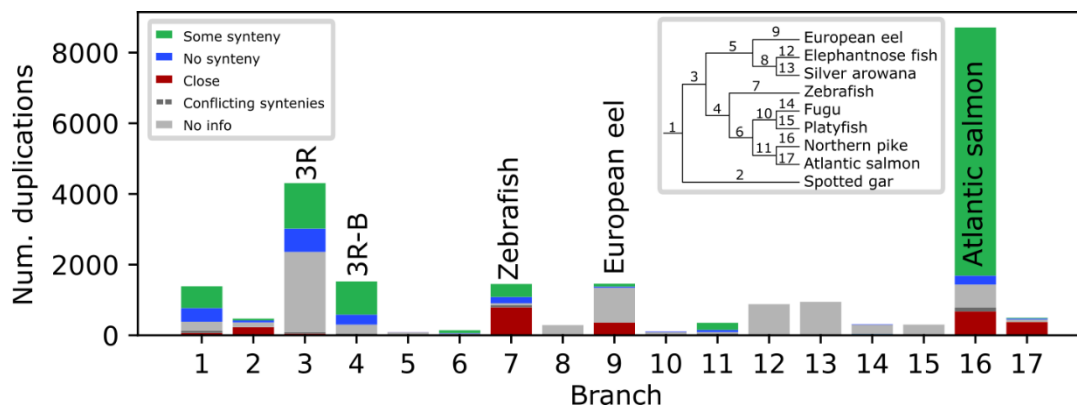


Figure 3

Barplot of the number of duplications PHYLDG assigned to each branch of the species tree. Bars are numbered according to the cladogram in the upper righthand corner. 3R indicates the branch where the 3R teleost-specific whole genome duplication is hypothesized to have happened. 3R-B indicates the basal branch of the remaining teleosts after the split of the elopomorphas and osteoglossomorphas. Each bar is subdivided into the synteny types described in figure 2.

218

219 In order to investigate the timing of the main eel duplication event in greater depth, we  
220 compared the 4dTv distribution found for eel paralogs with the 4dTv distribution built for the eel  
221 orthologs with elephantnose fish, and arowana. The results showed a 4dTv maximum of 0.4 for  
222 the main eel paralog peak, and 0.5 for the peaks corresponding to the speciation event that  
223 separated elephantnose fish and arowana from eel (Fig. 4).

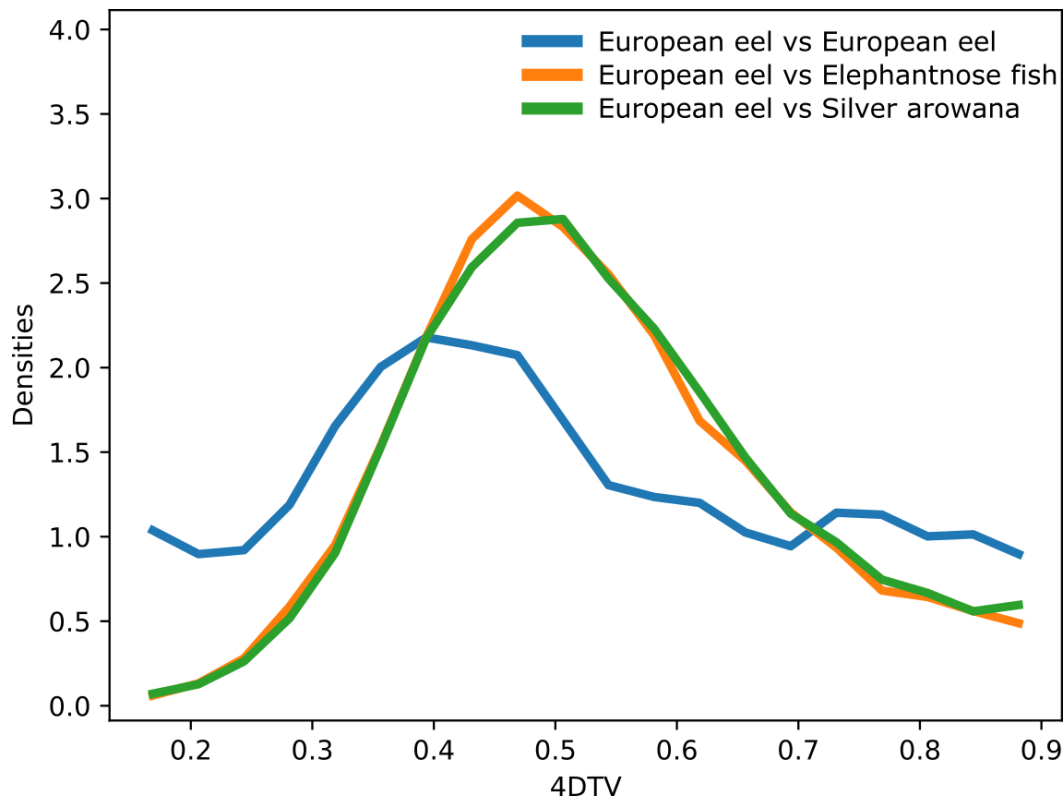


Figure 4  
4dTv distribution of European eel (*Anguilla anguilla*) paralogs (blue), European eel and elephantnose fish (*Gnathonemus petersii*) orthologs (green), and European eel and silver arowana (*Osteoglossum bicirrhosum*) orthologs (orange). This analysis was carried out in the gene families without taking into account the PHYLOGEN result.

224

#### 225 *Investigation of functional category enrichment*

226 To investigate if some functional categories were overrepresented in the eel duplications, an  
227 enrichment test was carried out. GO terms were assigned to 9570 gene families by comparing  
228 them to the annotated EggNOG gene families (Huerta-Cepas et al. 2016). The GO annotation  
229 can be found in suppl. table 1. From these terms we performed an enrichment analysis with the

230 topGO R library (Alexa and Rahnenfuhrer 2016).

231 The resulting enriched GO-terms, are presented in the suppl. table 2. In most cases these terms  
232 are involved either in signal transduction (GTPase, MAPK, sphingosine-1-phosphate),  
233 morphological alterations (convergent extension involved in axis elongation and gastrulation,  
234 heart development, and pronephric glomerulus development, pigmentation), or forebrain  
235 development.

236 Additionally, KEGG terms were assigned to 16466 eel genes using BlastKOALA (Kanehisa et  
237 al. 2016) and the terms related to the genes involved in the eel duplication were mapped onto  
238 the KEGG pathways using the KEGG Mapper tool. A fisher test, corrected for multiple testing  
239 using False Discovery Rate, was used to look for enriched KEGG pathways in the eel branch  
240 (Suppl. Table 3). Most of the KEGG pathways found to be enriched were related to immune  
241 system, nervous system, oocyte, apoptosis, cell adhesion, amino acid metabolism, glycan  
242 biosynthesis, and signal transduction. Several key genes related to the immune system were  
243 also duplicated, including: Cytohesin-associated scaffolding protein (*casp*) c-jun N-terminal  
244 kinase (*jnk*) and vav proto-oncogene (*vav*), involved in the lymphocyte differentiation and  
245 activation, the interleukine and T-cell receptors (*tcr's*), major histone compatibility complex  
246 (*mch*) I, *mch* II, and several cytokines including *tgf*-beta (transforming growth factor beta) and B-  
247 cell activating factor (*baff*). There were also numerous duplications related to apoptosis,  
248 including activator protein 1 (*ap-1*), c-fos proto-oncogene (*c-fos*), *casp*, serine protease (*omi*),  
249 mitochondrial Rho GTPase (*miro*), Death executioner (*bcl-2*), Caspase Dredd, X-linked inhibitor  
250 of apoptosis protein (*xiap*), *casp8*, and Baculoviral inhibitor-of-apoptosis repeat (*bir*). In the  
251 nervous system tyrosine hydroxylase (*th*) and monoamine oxidase, as well as some receptors  
252 (*gaba-a*, *gaba-b*, *ampa*, *mglur*, *crfr1*, and *girk*) were duplicated, as well as several genes  
253 involved in synaptic exocytosis (*rab3a*, *munc-18*, syntaxin, *vgat*), and Golf and Transductin,  
254 which are involved in olfaction and phototransduction. In the oocyte more affected genes were  
255 found, including the progesterone receptor (*pgr*), early mitotic inhibitor 2 (*emi2*), aurora-B

256 phosphatase (*glc7*), cullin F-box containing complex (*scf-skp*), insulin-like growth factor 1 (*igf1*),  
257 and cytoplasmic polyadenylation element binding protein (*cpeb*).

258

## 259 **Discussion**

260 The observed data has shown, for the first time to our knowledge, that in the eel lineage more  
261 than a thousand gene families have genes that are the result of a large-scale gene duplication  
262 that happened after the teleost specific 3R WGD. Only Atlantic salmon, due to its salmonid-  
263 specific 4R WGD, showed a higher amount of conserved duplicates, whereas the number of  
264 duplicated gene families found in eel and zebrafish were similar. Furthermore, the paralog 4dTv  
265 distribution shows a unique pattern in the case of the eel branch when compared to the other  
266 teleosts.

267 The duplications assigned by the phylogenetic analysis to the eel specific branch, after the split  
268 of arowana and elephantnose fish, showed a distribution of 4dTv distances younger than those  
269 corresponding to the 3R duplication and older than the salmonid duplications. This result was  
270 replicated in an independent analysis, not based on phylogenetic tree topologies. In this case  
271 the distributions of the 4dTv distances (Fig. 4) were calculated between the eel genes found in  
272 the gene families (eel paralogs) and between the eel sequences and the arowana and  
273 elephantnose fish sequences (orthologs). Both the phylogenetic topologies and the 4dTv  
274 distances corroborate the hypothesis that the main eel duplication happened after the teleost  
275 specific 3R duplication event (320-350 MYA; Vandepoele et al. 2004 and Christoffels et al.  
276 2004), and after the arowana and elephantnose fish split, but before the salmonid duplication  
277 (80-100 MYA; Macqueen et al. 2014). Given the 4dTv distribution found, we could assume that  
278 this duplication event is shared by all members of the *Anguilla* genus, as they are estimated to  
279 first appear 20-50 MYA (Minegishi et al. 2005). To be more precise about the timing of the main  
280 eel duplication event it would be advisable to study other elopomorpha transcriptomes or  
281 genomes to analyze if they share the same duplication.

282 Not all the PHYLOGENETIC duplication assignments to the basal branches of the species tree are as  
283 trustworthy as the lineage specific ones. According to PHYLOGENETIC the duplications usually  
284 associated with the 3R WGD would be split into two events, one would correspond to the 3R  
285 branch and the other with the one that gave rise to zebrafish, salmon and fugu (indicated as 3R-  
286 B; Fig. 2). If PHYLOGENETIC is right and there were really two genome duplications the 4dTv  
287 distributions of the events should be different. These distributions are quite similar, although not  
288 completely identical. Nevertheless, duplications this old are almost completely saturated in  
289 terms of 4dTv, decreasing the accuracy of the measurement. Alternatively, the PHYLOGENETIC split  
290 of the 3R duplication could be explained as an artifact. Only 3 species make up the daughter  
291 clade of the 3R branch at one side of the split and some gene families will only include  
292 representatives from the species of the other daughter clade. This could make the phylogenetic  
293 based assignment of the gene duplication to a particular branch more error prone. Thus, some  
294 gene family duplications created by 3R might end up wrongly assigned by PHYLOGENETIC to 3R-B.  
295 This might have happened even with some zebrafish lineage specific duplications. The 4dTv  
296 distribution for the zebrafish branch shows a bump that overlaps with the 3R duplication. That  
297 might be due to a real old duplication in the zebrafish lineage, but it is also possible that it is a  
298 PHYLOGENETIC artifact. The zebrafish genome is among the most complete genomes, with the most  
299 well-supported protein coding gene annotation, used in this study. Therefore, it might be the  
300 case that some duplicated genes could be found only in this species making the PHYLOGENETIC  
301 assignment more difficult.

302 In previous studies, data consistent with an eel specific duplication event has been reported,  
303 although it has never been interpreted in this way. Recently, the transcriptome of several  
304 teleosts was sequenced (Pasquier et al. 2016), and the European eel was the species with the  
305 highest number of contigs, except for species with a documented 4R WGD in their lineage. In  
306 the additional data included by Inoue et al. (2015), in their analysis of the gene loss process that  
307 followed the teleost 3R WGD, both the eel and zebrafish are the species with the highest

308 percentage of duplications (36.6 and 31.9%, respectively). Thus, these two studies are in  
309 concordance with our analysis.

310 Previously, several lineage specific gene duplications have been found and studied in eel, in  
311 studies focusing on particular genes. For instance, Morini et al. (2015) found that the leptin  
312 receptors were duplicated in the eel. However, their phylogeny did not include other basal  
313 teleosts and was compatible with both the hypothesis that those genes were duplicated in an  
314 eel specific duplication event or in the teleost 3R WGD. Similarly, Lafont et al. (2016) found  
315 several species-specific duplicated genes in the eel genome, but they attributed it to less  
316 extensive loss of genes in the eel lineage after the 3R duplication. Furthermore, several other  
317 analyses of single genes have reached the same conclusion (Dufour et al. 2005; Maugars and  
318 Dufour 2015; Pasqualini et al. 2009; Pasquier et al. 2012; Morini et al. 2017; Henkel et al. 2012).  
319 In general, these studies were based on tree topologies that did not include other basal teleost  
320 species and often did not report the genetic distances.

321 The eel duplication was also not detected in the analysis of the published eel genome (Henkel  
322 et al. 2012). This genome was quite fragmented and, according to our BUSCO analysis, it was  
323 more incomplete than our transcriptome, thus including less duplications (Fig. 1). Moreover, no  
324 global 4dTv or Ks distribution was calculated, but a remarkably high number of Hox genes (73)  
325 were found, thus these results are also compatible with our current analysis.

326 In the proposed species phylogeny (Fig. 2) the eel is located in a basal position as found in  
327 other previous phylogenies. The location of the osteoglossomorphes order, represented by the  
328 elephantnose fish and arowana is, however, in disagreement with previous phylogenies based  
329 on elopomorpha mitochondrial genomes (Inoue et al. 2003) and on the nuclear arowana  
330 genome (Austin et al. 2015). In these phylogenies arowana is the basal node to the teleosts,  
331 whereas it is grouped with the eel in ours. In the transcriptome based phylogeny proposed by  
332 the PhyloFish project (Pasquier et al. 2016) the topology is reversed and the Anguilliformes  
333 appear as the basal teleosts whereas the Osteoglossiformes appear to have split more recently.



334 The branch that separated eel and arowana, in the arowana genome paper, was one of the  
335 shortest and had a lower posterior probability than the others; this could be the reason why  
336 these different phylogenies disagree. This disparity in the results is unlikely to be due to the  
337 phylogenetic methods used. In this study we have used the neighbor joining, maximum  
338 likelihood and bayesian approaches, and all of them agree that arowana and eel form two sister  
339 clades. The difference might be due to the extra species that we are including, the elephantnose  
340 fish. To be more confident of the topology of these events in the base of the teleosts, it would be  
341 advisable to include other basal species.

342 Two different mechanisms can create genomic scale duplications: WGDs or many small-scale  
343 SDs, a process referred to as the continuous mode hypothesis (Gu et al. 2002). The latter  
344 process has been observed in many species, including yeast (Llorente et al. 2000), fruit flies  
345 (Zhou et al. 2008), water fleas (Colbourne et al. 2011), humans (Bailey et al. 2002; Gu et al.  
346 2002; Vallente Samonte and Eichler 2016), several plant species (Cui et al. 2006) and teleosts  
347 (Blomme et al. 2006; David et al. 2003; Jaillon et al. 2004; Lu et al. 2012; Rondeau et al. 2014).  
348 Usually, most of these SDs are lost soon after they are generated. Therefore, in the genome,  
349 many young paralog pairs, and few old, can be found. In a 4dTv distribution this pattern would  
350 be detected as a peak of 4dTv values with a mode close to 0. Moreover, SDs are usually the  
351 result of tandem duplications, therefore very similar paralogs located in tandem are likely to be  
352 the result of this process. Patterns compatible with this scenario were seen in several species in  
353 our analysis, including: zebrafish, elephantnose fish, arowana, pike, fugu, spotted gar, salmon  
354 and eel (suppl. Fig. 2). The high amount of SDs that we detected in zebrafish has been  
355 previously documented (Blomme et al. 2006; Lu et al. 2012). In some other species it has been  
356 shown that these SDs could be retained at specific points in time, possibly during specific  
357 evolutionary events e.g. in yeast (Llorente et al. 2000), common carp (David et al. 2003) and  
358 humans (Bailey et al. 2002; Hughes et al. 2001). These events have been linked to the

359 adaptation of a species to a new environment (Chain et al. 2014; Colbourne et al. 2011; Tautz  
360 and Domazet-lošo 2011).

361 In eel, the 4dTv distribution pattern found is quite distinct as it reflects two modes; one of  
362 younger and one of older duplications. Of these duplications, the older ones are clearly older  
363 than those found, for example, in zebrafish or salmon. Furthermore, the genomic surroundings  
364 of the paralogs of the younger duplications are quite different from those of the older  
365 duplications. The younger duplications, which have a lower 4dTv, tend to be located close  
366 together in the genome and are likely to have been generated recently by tandem SDs, whereas  
367 the older duplications are not usually found in tandem. A WGD should have left behind blocks of  
368 syntenic regions similar to those detected in our analysis for the 3R teleost and the salmon  
369 WGD. In the case of the eel, an increase in these syntenic blocks was also detected in the older  
370 duplications. However most genomic regions are very fragmented in the genome assembly and  
371 thus, we lack physical genomic location information for many genes. However, from the  
372 evidence available we can hypothesize that most of the older duplications are likely to be the  
373 result of a WGD which occurred in the eel lineage, and that an analysis with the latest eel  
374 genome assembly published (Jansen et al. 2017), but not available without restrictions, would  
375 detect more syntenic regions. In other words, the numerous duplications found in eel are likely  
376 to have been generated by a WGD followed by many SDs, a pattern which has also been  
377 observed in primates (Gu et al. 2002), and common carp (David et al. 2003).

378 In this study, several gene function analyses were carried out to study overrepresented  
379 functions among the eel specific duplications. These overrepresentations could be linked to  
380 several adaptations that have taken place throughout eel evolution, e.g. the inclusion of a  
381 leptocephali larvae stage to their life history (Inoue et al. 2004), the adaptation to a catadromous  
382 lifecycle (Inoue et al. 2010), and the adaptation to withhold maturation until after the extensive  
383 reproductive migration (Righton et al. 2016; van Ginneken et al. 2005). Other mechanisms  
384 which perhaps influence the conservation of paralogs are: dosage selection (Glasauer and

385 Neuhauss 2014) and segregation avoidance (Hahn 2009). It has been suggested that these  
386 mechanisms conserve duplicated genes related to specific biological processes, such as  
387 development, signaling, ion transport, metabolism and neuronal function after WGDS (Berthelot  
388 et al. 2014; Blomme et al. 2006; Brunet et al. 2006; Kassahn et al. 2009).

389 Specifically, 54 GO-terms were found to be enriched among the eel specific duplications (suppl.  
390 table 2). Interestingly, several of the GO-terms found to be enriched among the eel specific  
391 duplications form part of some of the aforementioned processes, including; development, ion  
392 transport, signaling, neuronal function, and metabolism. The high number of enriched GO-  
393 terms, which are part of processes that are often conserved after WGDs, suggests that the  
394 duplication event here described is a WGD. It is likely that they have been conserved due to the  
395 mechanisms regulating gene conservation after WGD rather than due to specific necessities of  
396 the *Anguilla* species. Other GO terms that were found to be duplicated in eel are not usually  
397 found in other WGDs, for example “pigmentation”. As the eel has incorporated several  
398 pigmentation alterations into its lifecycle, it is likely that the genes associated with this GO-term  
399 are conserved due to new functions acquired in the *Anguilla* species. Most of the pigmentation  
400 changes undergone by the eel are linked to the transition between the marine and freshwater  
401 environment, therefore the duplication of these genes might have generated the necessary raw  
402 genetic material for adaptation to the catadromous lifecycle.

403 Furthermore, 54 KEGG pathways were also found to be enriched among the eel specific  
404 duplications (suppl. table 3). As in the case of the GO-terms, several of these pathways are  
405 involved in signaling, metabolism and neuronal function. Additionally, olfactory transduction and  
406 several pathways involved in immune response e.g. Tuberculosis, Th1 and Th2 cell  
407 differentiation, Bacterial invasion of epithelial cells, Th17 cell differentiation, and others were  
408 found to be enriched. Lu et al. (2012) also found that immune response pathways and olfactory  
409 receptor activity were enriched among the recent segmental duplications found in zebrafish.  
410 Several studies have also found immune response genes to be enriched among other recent

411 SDs (Conrad and Antonarakis 2007; Kasahara et al. 2007; She et al. 2008; Stein et al. 2007;  
412 Wang et al. 2012). Thus this enrichment in immune response genes could be linked to eel SDs.  
413 However, in these studies, the recent duplications were found to be mostly in the components  
414 interacting with pathogens, possibly to contribute to the response against different pathogens,  
415 as opposed to the components downstream of the receptors, which make up most of the  
416 enriched pathways found in our study. Also, among the eel specific duplications were the  
417 progesterin receptors which have recently been characterized in eel (Morini et al. 2017).  
418 Progesterins are known as maturation-inducing steroids promoting sperm maturation and  
419 spermiation (for review see Scott et al. 2010), and the two paralogs do show differential  
420 expression during maturation (Morini et al. 2017).

421 The most significantly enriched pathway found in the eel duplications is the dopaminergic  
422 synapse pathway. Dopamine (DA) is an essential neurotransmitter in vertebrates, with several  
423 functions (Davila et al. 2003; Hsia et al. 1999). DA has been proven to be important in teleosts,  
424 where DA has been found to have an inhibitory role on the gonadotropic activity of the pituitaries  
425 (Dufour et al. 2005; Peter et al. 1986). In the case of the eel in particular, DA inhibition  
426 completely arrests puberty before their oceanic migration (Vidal et al. 2004), indicating that DA  
427 has a much stronger inhibitor effect in eel compared to most other teleosts (Dufour 1988; Vidal  
428 et al. 2004). This suggests that the duplicated genes involved in the dopaminergic synapse  
429 pathway may have been conserved during the adaptation to block maturation until after the  
430 extensive reproductive migration. Among the duplicated genes assigned to the dopaminergic  
431 synapse pathway, we found tyrosine hydroxylase (TH). TH is the rate limiting enzyme of the DA  
432 biosynthesis (Nagatsu et al. 1964), and is therefore often used as an indicator of DA tone in eel  
433 (Davila et al. 2003; Weltzien et al. 2015). As genes are rarely conserved without a specific  
434 function or necessity, the presence of two TH genes in the eel encourages suspicion of potential  
435 differential expression or function between the two, which may prove important for the regulation  
436 of the DA induced inhibition of puberty observed in pre-migration eels.

437 In conclusion, the data presented strongly suggest that a vast amount of genes have been  
438 duplicated specifically in the eel lineage. Furthermore, the synteny, 4dTv, and enrichment  
439 analyses suggest that these genes derive both from a WGD as well as continuously created  
440 SDs, and that they are related to the eel specific physiology. To our knowledge this is thus the  
441 first evidence published suggesting a possible eel lineage specific 4R WGD.

442

## 443 **Materials and methods**

### 444 *Fish husbandry*

445 Ten immature farm eel males (mean body weight  $96.7 \pm 3.6$  g $\pm$ SEM) supplied by Valenciana de  
446 Acuicultura S.A. (Puzol, Valencia, Spain) were transported to the Aquaculture Laboratory at the  
447 Universitat Politècnica de València, Spain. The fish were kept in a 200-L tank, equipped with  
448 individual recirculation systems, a temperature control system (with heaters and, coolers), and  
449 aeration. The fish were gradually acclimatized to sea water (final salinity  $37 \pm 0.3\%$ ), over the  
450 course of two weeks. The temperature, oxygen level and pH of rearing were 20 °C, 7-8 mg/L  
451 and ~ 8.2, respectively. The tank was covered to maintain, as much as possible, a constant  
452 dark photoperiod and the fish were starved throughout the holding period. After acclimation, the  
453 fish were sacrificed in order to collect samples of forebrain, pituitary, and testis tissues.

454

### 455 *Human and Animal Rights*

456 This study was carried out in strict accordance with the recommendations given in the Guide for  
457 the Care and Use of Laboratory Animals of the Spanish Royal Decree 53/2013 regarding the  
458 protection of animals used for scientific purposes (BOE 2013), and in accordance with the  
459 European Union regulations concerning the protection of experimental animals (Dir  
460 86/609/EEC). The protocol was approved by the Experimental Animal Ethics Committee from  
461 the Universitat Politècnica de València (UPV) and final permission was given by the local

462 government (Generalitat Valenciana, Permit Number: 2014/VSC/PEA/00147). The fish were  
463 sacrificed using anesthesia and all efforts were made to minimize suffering.

464

#### 465 *RNA extraction and sequencing*

466 High quality RNA was extracted from forebrain, pituitary, and testis samples following the  
467 protocol developed by Peña-Llopis and Brugarolas (2013). Quantity and quality were tested on  
468 a Bio-Rad Bioanalyser (Bio-Rad Laboratories, Hercules, CA, USA), selecting the samples with  
469 RIN values and amounts higher than  $>8$   $>3$   $\mu\text{g}$  of total RNA, respectively. Total RNA samples  
470 were shipped to the company MacroGen Korea (Seoul, South Korea). Then, a mRNA  
471 purification was carried out using Sera-mag Magnetic Oligo (dT) Beads, followed by buffer  
472 fragmentation. Reverse transcription was followed by PCR amplification to prepare the samples  
473 for sequencing. The strand information was kept in an Illumina HiSeq-2000 sequencer (Illumina,  
474 San Diego, USA). Resulting raw sequences are available at the NCBI Sequence Read Archive  
475 (SRA) as stated in the section titled "Data accessibility".

476

#### 477 *Transcriptome assemblies and genomes*

478 The software FastQC (Andrews 2010) was used to assess the quality of the raw reads  
479 generated by MacroGen. Thereafter, trimmomatic (Bolger et al. 2014) was used to trim the  
480 reads, eliminating known adaptor sequences, and low quality regions. Finally, trimmed reads  
481 shorter than 50 bp were filtered out. Eel reads were digitally normalized before assembly by  
482 khmer software (Crusoe et al. 2015) using a k-mer length of 25 and a coverage of 100. Further,  
483 The RNA-Seq raw reads for pike, arowana and elephantnose fish were downloaded from the  
484 PhyloFish project (Pasquier et al. 2016). All transcriptomes were then assembled using Trinity  
485 software (Haas et al. 2013), with the read orientation and sense (in the eel case) into account.  
486 The transcripts assembled were filtered according to their complexity (with a DUST score  
487 threshold of 7 and a DUST window of 64), length (with a minimum length of 500 bp), and level

488 of expression (with a TPM threshold of 1). After assembly, the CDSs and proteins were  
489 annotated using the Trinotate functional annotation pipeline (Haas et al. 2013).

490 Transcripts that share k-mers are clustered by Trinity, however, these transcripts might  
491 correspond to different transcript forms of the same gene or to closely related genes from a  
492 gene family. We split these transcripts into genes by running a transitive clustering based on a  
493 blast search. In this clustering transcripts which shared at least 100 bp with a minimum identity  
494 of 97% were considered to be isoforms of the same gene. Thus, some Trinity clusters were split  
495 into several genes. For each gene, the most expressed transcript, according to Salmon (Patro  
496 et al. 2017), was chosen as its representative.

497 The available eel genome was downloaded from the ZF-Genomics web site (Henkel et al.  
498 2012). The salmon genome assembled by the International Cooperation to Sequence the  
499 Atlantic Salmon Genome was downloaded from NCBI (Lien et al. 2016). The genomes of  
500 zebrafish (Howe et al. 2013), fugu (Kai et al. 2011), spotted gar (Braasch et al. 2016), and  
501 platyfish (Schartl et al. 2013) were downloaded from ENSEMBL (release 87). The pike genome  
502 (Rondeau et al. 2014) was downloaded from the Northern Pike Genome web site (Genbank  
503 accession GCA\_000721915.1). For each gene in the genomes, the longest transcript was  
504 chosen as the representative.

505

#### 506 *Genome and transcriptome quality assessment*

507 In order to check the quality of the transcriptomes and genomes we looked for the BUSCO  
508 conserved gene set in them (Simão et al. 2015). BUSCOs are conserved proteins, and are  
509 expected to be found in any complete genome or transcriptome. Therefore, the number of  
510 present, missing, or fragmented BUSCOs can be used as a quality control of a genome or  
511 transcriptome assembly. For this assessment the Actinopterygii (*odb9*) gene set, which consists  
512 of 4584 single-copy genes that are present in at least 90% of Actinopterygii species was used.

513 As an additional comparison between the transcriptome and genomes of pike and eel, the RNA-

514 seq reads were mapped both to the genome and transcriptome assemblies using the softwares  
515 HISAT2 (Pertea et al. 2016) and BWA-MEM (Li and Durbin 2010), respectively.

516

### 517 *Gene families*

518 Genes were clustered into gene families by the OrthoMCL web service (Li et al. 2003). For each  
519 gene family a multiple protein alignment was built. To avoid transcriptome assembly artifacts  
520 proteins longer than 1,500 amino acids, transcripts with a DUST score higher than 7 or  
521 sequences with more than 40% of gaps in the alignment were filtered out. The software Clustal  
522 Omega (Sievers et al. 2011) carried out the protein multiple alignment and trimAl (Capella-  
523 Gutiérrez et al. 2009) removed the regions with too many gaps or those difficult to align. The  
524 protein alignment was used as a template to build the codon alignment by aligning the transcript  
525 sequences against the corresponding protein using the protein2dna exonerate algorithm (Slater  
526 and Birney 2005).

527

### 528 *Phylogenetic reconstruction and duplication dating*

529 The resulting protein alignments were used by PHYLOG (Boussau et al. 2012) software to  
530 generate a species tree as well as a family tree corresponding to each alignment. Due to the  
531 high memory requirements of PHYLOG not all gene families could be run in the same analysis  
532 so 10 analyses were carried out, choosing 8000 protein alignments at random for each. Once all  
533 runs were finished, we checked that the species tree topology of all the 10 species trees,  
534 matched exactly. . PHYLOG uses a maximum likelihood approach to simultaneously co-  
535 estimate the species and gene family trees from all individual alignments. From the topology of  
536 the gene family trees, it is capable of inferring when the duplications in each family happened.  
537 Alternatively, the species phylogeny was also reconstructed using a bayesian approach by  
538 using PhyloBayes MPI version 1.7 (Lartillot et al. 2009). From the gene families that had one  
539 gene for each species, 100 were chosen at random to create a concatenated alignment of



540 43566 aminoacids. The model used was CAT-GTR and three independent MCMC chains were  
541 run for 39872, 56328, and 39285 iterations.

542 Finally, a neighbor joining tree based on the fourfold synonymous third-codon transversion  
543 distances (4dTv) was also calculated (Tang et al. 2008). Between any pair of sequences the  
544 number of transversions found in the third base of the codon was divided by the number of four-  
545 fold degenerated codons. A correction to the 4dTv was applied:  $\ln(1 - 2 * \text{distance}) / -2$ . For  
546 each pair of species a 4dTv distance was calculated. 4dTvs were calculated between the  
547 sequences of those species found in each gene family codon alignment. The distribution of  
548 those 4dTvs was fitted with a log normal mixture model using the scikit-learn Gaussian Mixture  
549 class. The number of components required was one for all the species pairs, except for those  
550 where a bimodal distribution was found due to a recent speciation event. The distance between  
551 any two species was the mode of the fitted model. The neighbor joining tree was built using the  
552 BioPython Tree Construction class. This process is implemented in the `fit_fdtv_distributions`  
553 module found in the Python scripts (suppl. Material 1).

554 The 4dTv was calculated for each duplication tagged by PHYLDOG within any gene family. A  
555 duplication event defines a subtree in the gene family tree, and this subtree defines two child  
556 branches, so the 4dTv calculated for that event was the mean of the 4dTv between all  
557 combinations of sequences found between those branches. These calculations are  
558 implemented by the functions `calculate_4dTv` and `calculate_mean_fdtv_for_tree` found in the  
559 Python scripts (suppl. Material 1).

560

### 561 *Synteny*

562 Furthermore, the kind of event that created each duplication was characterized by analyzing the  
563 conserved synteny between the paralogs created by that duplication within a particular genome.  
564 A duplication may derive from a SD that could have occurred in tandem or not, or from a WGD,  
565 among others. Tandem SDs would create paralogs found close to each other in the genome,

566 whereas the paralogs created by a WGD would be far away, but surrounded by similar genes in  
567 each of the duplicated regions. We also have to consider that several phylogenetically close  
568 species can be affected by the same older duplication event. Therefore, these traces of the  
569 duplication event could be found in the genomes of those different species and, if no other  
570 genomic rearrangement happened since, these traces should match each other and convey the  
571 same information. With this in mind, we categorized duplications as one of 4 classes: i) the  
572 paralog genes that were found close to each other in the genome, within a 50 gene distance  
573 were labelled as close, ii) the paralogs which were found in syntenic regions where 2 or more  
574 paralogous from other gene families were located within a 50 genes distance, not necessarily in  
575 the same colinear order, were labelled as “some synteny”, iii) the cases in which fewer than 2  
576 gene families could be identified within a 50 gene distance from both of the paralogous genes  
577 were labelled as “no info”, and iv) the cases in which conflicting evidence was found in the  
578 genomes of the different species affected by the duplication were labelled as “conflicting  
579 syntenies”. This labelling of the duplications was carried out by the Python function  
580 `determine_if_pair_is_close_or_syntenic` and the Python class `GenomeLocator`, found in the  
581 scripts (suppl. material 1). The location of each gene in a genome was obtained by performing a  
582 BLAST search with its representative transcript against the genome.

583

#### 584 *Investigation of functional category enrichment*

585 The EggNOG database has GO annotations for each of its gene families (Huerta-Cepas et al.  
586 2016). To match our gene families with those from the EggNOG database, the protein sequence  
587 with least gaps for each of our families was selected and a HMMER search (Finn et al. 2011)  
588 was carried out against the EggNOG position weight matrices with an e-value threshold of  
589 0.0001. The GO annotation of the best EggNOG hit in this search was transferred to our family.  
590 The enrichment analysis was carried out using the fisher statistic and the weight algorithm of the  
591 topGO library (Alexa and Rahnenfuhrer 2016) from the bioconductor project. The R script

592 go\_enrichment\_analysis found in the scripts (suppl. Material 1) implements this analysis. Eel  
593 transcripts were annotated using the BlastKOALA KEGG service (Kanehisa et al. 2016) and a  
594 fisher exact test was carried out, using the scipy implementation, to look for overrepresented  
595 KEGG pathways in the eel duplications.

596

#### 597 **Disclosure declaration**

598 The authors declare no potential conflicts of interest with respect to the authorship, research,  
599 and/or publication of this article.

600

#### 601 **Data accessibility**

602 The raw RNA-sequencing reads from brain, pituitary, and testis samples from European eel  
603 (*Anguilla anguilla*) have been deposited at GenBank (<http://www.ncbi.nlm.nih.gov/genbank>)  
604 under accession no. XX. Deposition and acquiring of accession number was not finished at the  
605 time of first submission but will be settled within a few days.

606

#### 607 **Acknowledgements**

608 This study received funding from the project REPRO-TEMP (AGL2013-41646-R) funded by the  
609 Spanish Ministry of Economy and Competitiveness, and from the European Union's Horizon  
610 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No  
611 642893 (IMPRESS). V. Gallego has a postdoc grant from the UPV (PAID-10-16).

612

#### 613 **References**

614 Ager-Wick E, Dirks RP, Burgerhout E, Nourizadeh-Lillabadi R, de Wijze DL, Spaink HP, van den  
615 Thillart GEEJM, Tsukamoto K, Dufour S, Weltzien FA, et al. 2013. The pituitary gland of  
616 the European eel reveals massive expression of genes involved in the melanocortin

- 617 system. *PLoS One* **8**: 1–12.
- 618 Alexa A, Rahnenfuhrer J. 2016. topGO: Enrichment analysis for gene ontology. R package  
619 version 2.29.0.
- 620 Allendorf F, Thorgaard G. 1984. Tetraploidy and the evolution of salmonid fishes. In  
621 *Evolutionary Genetics of Fishes*, pp. 1–53.
- 622 Andrews S. 2010. FastQC: A quality control tool for high throughput sequence data.  
623 <http://www.bioinformatics.babraham.ac.uk/projects>.
- 624 Aparicio S, Chapman J, Stupka E, Putnam N, Chia J, Dehal P, Christoffels A, Rash S, Hoon S,  
625 Smit A, et al. 2002. Whole-Genome shotgun assembly and analysis of the genome of *fugu*  
626 *rubripes*. *Science* **297**: 1301–1310.
- 627 Asrar Z, Haq F, Abbasi AA. 2013. Molecular Phylogenetics and Evolution Fourfold paralogy  
628 regions on human HOX-bearing chromosomes: Role of ancient segmental duplications in  
629 the evolution of vertebrate genome. *Mol Phylogenet Evol* **66**: 737–747.
- 630 Austin CM, Tan MH, Croft LJ, Hammer MP, Gan HM. 2015. Whole genome sequencing of the  
631 asian arowana of ray-finned fishes. *Genome Biol Evol* **7**: 2885–2895.
- 632 Bailey JA, Gu Z, Clark RA, Reinert K, Samonte R V, Schwartz S, Adams MD, Myers EW, Li PW,  
633 Eichler EE. 2002. Recent segmental duplications in the human genome. *Science* **297**:  
634 1003–1007.
- 635 Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noël B, Bento P, Da Silva C,  
636 Labadie K, Alberti A, et al. 2014. The rainbow trout genome provides novel insights into  
637 evolution after whole-genome duplication in vertebrates. *Nat Commun* **5**: 3657.
- 638 Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y. 2006. The gain and  
639 loss of genes during 600 million years of vertebrate evolution. *Genome Biol* **7**: 1–12.
- 640 Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence  
641 data. *Bioinformatics* **30**: 2114–2120.
- 642 Boussau B, Szöll GJ, Duret L, Gouy M, Tannier E, Daubin V, Lyon U De, Lyon U. 2012.

- 643           Genome-scale coestimation of species and gene trees. *Life Sci* **23**: 323–330.
- 644   Brunet FG, Crollius HR, Paris M, Aury JM, Gibert P, Jaillon O, Laudet V, Robinson-Rechavi M.  
645           2006. Gene loss and evolutionary rates following whole-genome duplication in teleost  
646           fishes. *Mol Biol Evol* **23**: 1808–1816.
- 647   Braasch I, Gehrke AR, Smith JJ, Kawasaki K, Manousaki T, Pasquier J, Amores A, Desvignes  
648           T, Batzel P, Catchen J, et al. 2016. Corrigendum: The spotted gar genome illuminates  
649           vertebrate evolution and facilitates human-teleost comparisons. *Nat Genet* **48**: 700–700.
- 650   Burgerhout E, Minegishi Y, Brittijn SA, de Wijze DL, Henkel C V., Jansen HJ, Spaink HP, Dirks  
651           RP, van den Thillart GEEJM. 2016. Changes in ovarian gene expression profiles and  
652           plasma hormone levels in maturing European eel (*Anguilla anguilla*); Biomarkers for  
653           broodstock selection. *Gen Comp Endocrinol* **225**: 185–196.
- 654   Cañestro C, Albalat R, Irimia M, Garcia-Fernández J. 2013. Impact of gene gains, losses and  
655           duplication modes on the origin and diversification of vertebrates. *Semin Cell Dev Biol* **24**:  
656           83–94.
- 657   Capella-gutiérrez S, Silla-martínez JM, Gabaldón T. 2009. trimAl: a tool for automated  
658           alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972–1973.
- 659   Chain FJJ, Feulner PGD, Panchal M, Eizaguirre C, Samonte IE, Kalbe M, Lenz TL, Stoll M,  
660           Bornberg-Bauer E, Milinski M, et al. 2014. Extensive copy-number variation of young  
661           genes across stickleback populations. *PLoS Genet* **10**: e1004830.
- 662   Christoffels A, Koh EGL, Chia JM, Brenner S, Aparicio S, Venkatesh B. 2004. Fugu genome  
663           analysis provides evidence for a whole-genome duplication early during the evolution of  
664           ray-finned fishes. *Mol Biol Evol* **21**: 1146–1151.
- 665   Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A,  
666           Arnold GJ, Basu MK, et al. 2011. The ecoresponsive genome of *daphnia pulex*. *Science*  
667           **331**: 555–562.
- 668   Conrad B, Antonarakis SE. 2007. Gene Duplication: A drive for phenotypic diversity and cause

- 669 of human disease. *annurev.genom* **8**: 17–35.
- 670 Coppe A, Pujolar JM, Maes GE, Larsen PF, Hansen MM, Bernatchez L, Zane L, Bortoluzzi S.  
671 2010. Sequencing, de novo annotation and analysis of the first *Anguilla anguilla*  
672 transcriptome: EeelBase opens new perspectives for the study of the critically endangered  
673 european eel. *BMC Genomics* **11**: 635.
- 674 Crusoe MR, Alameldin HF, Awad S, Boucher E, Caldwell A, Cartwright R, Charbonneau A,  
675 Constantinides B, Edverson G, Fay S, et al. 2015. The khmer software package: enabling  
676 efficient nucleotide sequence analysis. *F1000Research* **4**: 900.
- 677 Cui L, Wall PK, Leebens-mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE,  
678 Arumuganathan K, Barakat A, et al. 2006. Widespread genome duplications throughout the  
679 history of flowering plants. *Genome Res* **814**: 738–749.
- 680 David L, Blum S, Feldman MW, Lavi U, Hillel J. 2003. Recent duplication of the common carp  
681 (*Cyprinus carpio* L.) Genome as revealed by analyses of microsatellite loci. *Mol Biol Evol*  
682 **20**: 1425–1434.
- 683 Davila NG, Blakemore LJ, Trombley PQ, Nestor G, Blakemore LJ, Trombley PQ. 2003.  
684 Dopamine modulates synaptic transmission between rat olfactory bulb neurons in culture. *J*  
685 *Neurophysiol* **90**: 395–404.
- 686 Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate.  
687 *PLoS Biol* **3**: 1700–1708.
- 688 Dufour S. 1988. Stimulation of gonadotropin release and of ovarian development, by the  
689 administration of a gonadoliberin agonist and of dopamine antagonists, in female silver eel  
690 pretreated with estradiol. *Gen Comp Endocrinol* **70**: 20–30.
- 691 Dufour S, Weltzien F, Sebert M-E, Le Belle N, Vidal B, Vernier P, Pasqualini C. 2005.  
692 Dopaminergic inhibition of reproduction in teleost fishes: ecophysiological and evolutionary  
693 implications. *Ann N Y Acad Sci* **1040**: 9–21.
- 694 Ferris SD, Whitt GS. 1977. Duplicate gene expression in diploid and tetraploid loaches

- 695 (Cypriniformes, Cobitidae). *Biochem Genet* **15**: 1097–1112.
- 696 Finn RD, Clements J, Eddy SR. 2011. HMMER web server: Interactive sequence similarity  
697 searching. *Nucleic Acids Res* **39**: 29–37.
- 698 Glasauer SMK, Neuhauss SCF. 2014. Whole-genome duplication in teleost fishes and its  
699 evolutionary consequences. *Mol Genet Genomics* **289**: 1045–1060.
- 700 Greenwood PH, Rosen DE, Weitsman SH MG. 1966. *Phyletic studies of teleostean fishes, with*  
701 *a provisional classification of living forms*. bull am Mus nat.
- 702 Gu X, Wang Y, Gu J. 2002. Age distribution of human gene families shows significant roles of  
703 both large- and small-scale duplications in vertebrate evolution. *Nat Genet* **31**: 205–209.
- 704 Hafeez M, Shabbir M, Altaf F, Abbasi AA. 2016. Phylogenomic analysis reveals ancient  
705 segmental duplications in the human genome. *Mol Phylogenet Evol* **94**: 95–100.
- 706 Hahn MW. 2009. Distinguishing among evolutionary models for the maintenance of gene  
707 duplicates. *J Hered* **100**: 605–617.
- 708 Henkel C V., Burgerhout E, de Wijze DL, Dirks RP, Minegishi Y, Jansen HJ, Spaink HP, Dufour  
709 S, Weltzien FA, Tsukamoto K, et al. 2012. Primitive duplicate hox clusters in the european  
710 eel's genome. *PLoS One* **7**.
- 711 Hoegg S, Brinkmann H, Taylor JS, Meyer A. 2004. Phylogenetic timing of the fish-specific  
712 genome duplication correlates with the diversification of teleost fish. *J Mol Evol* **59**: 190–  
713 203.
- 714 Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JEJE, Humphray S,  
715 McLaren K, Matthews L, et al. 2013. The zebrafish reference genome sequence and its  
716 relationship to the human genome. *Nature* **496**: 498–503.
- 717 Hsia AY, Vincent J, Lledo P, National C, Recherche D, Fessard IA. 1999. Dopamine depresses  
718 synaptic inputs into the olfactory bulb. *J Physiol* **82**: 1082–1085.
- 719 Huerta-cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR,  
720 Sunagawa S, Kuhn M, et al. 2016. eggNOG 4.5: a hierarchical orthology framework with

- 721 improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic*  
722 *Acids Res* **44**: 286–293.
- 723 Hughes AL, Silva J, Friedman R. 2001. Ancient genome duplications did not structure the  
724 human hox -bearing chromosomes. *Genome Res* **11**: 771–780.
- 725 Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D,  
726 Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-seq  
727 using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**: 1494–1512.
- 728 Inoue JG, Miya M, Miller MJ, Sado T, Hanel R, Hatooka K, Aoyama J, Minegishi Y, Nishida M,  
729 Tsukamoto K. 2010. Deep-ocean origin of the freshwater eels. *Biol Lett* **6**: 363–366.
- 730 Inoue JG, Miya M, Tsukamoto K, Nishida M. 2003. Basal actinopterygian relationships: A  
731 mitogenomic perspective on the phylogeny of the “ancient fish.” *Mol Phylogenet Evol* **26**:  
732 110–120.
- 733 Inoue JG, Miya M, Tsukamoto K, Nishida M. 2004. Mitogenomic evidence for the monophyly of  
734 elopomorph fishes (Teleostei) and the evolutionary origin of the leptocephalus larva. *Mol*  
735 *Phylogenet Evol* **32**: 274–286.
- 736 Jacobsen MW, Pujolar JM, Gilbert MTP, Moreno-Mayar J V, Bernatchez L, Als TD, Lobon-  
737 Cervia J, Hansen MM. 2014. Speciation and demographic history of Atlantic eels (*Anguilla*  
738 *anguilla* and *A. rostrata*) revealed by mitogenome sequencing. *Heredity (Edinb)* **113**: 1–11.
- 739 Jaillon O, Aury J, Brunet F, Petit J-L, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C,  
740 Ozouf-costaz C, Bernot A, et al. 2004. Genome duplication in the teleost fish *Tetraodon*  
741 *nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431**: 946–957.
- 742 Jansen HJ, Liem M, Jong-raadsen SA, Dufour S, Swinkels W, Koelewijn A, Palstra AP, Pelster  
743 B, Herman P, Thillart GE Van Den, et al. 2017. Rapid de novo assembly of the European  
744 eel genome from nanopore sequencing reads. *Sci Rep* **7**: 7213.
- 745 Johnson KR, Wright JE, May B. 1987. Linkage relationships reflecting ancestral tetraploidy in  
746 salmonid fish. *Genetics* **116**: 579–591.



- 747 Kai W, Kikuchi K, Tohari S, Chew AK, Tay A, Fujiwara A, Hosoya S, Suetake H, Naruse K,  
748 Brenner S, et al. 2011. Integration of the genetic map and genome assembly of evolution in  
749 teleosts and mammals. *Genome Biol Evol* **3**: 424–442.
- 750 Kanehisa M, Sato Y, Morishima K. 2016. BlastKOALA and GhostKOALA: KEGG tools for  
751 functional characterization of genome and metagenome sequences. *J Mol Biol* **428**: 726–  
752 731.
- 753 Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, Yamada T, Nagayasu Y, Doi K,  
754 Kasai Y, et al. 2007. The medaka draft genome and insights into vertebrate genome  
755 evolution. *Nature* **447**: 714–719.
- 756 Kassahn KS, Dang VT, Wilkins SJ, Kassahn KS, Dang VT, Wilkins SJ, Perkins AC, Ragan MA.  
757 2009. Evolution of gene function and regulatory control after whole-genome duplication□:  
758 Comparative analyses in vertebrates. *Genome Res* **19**: 1404–1418.
- 759 Lafont AG, Rousseau K, Tomkiewicz J, Dufour S. 2016. Three nuclear and two membrane  
760 estrogen receptors in basal teleosts, *Anguilla* sp.: Identification, evolutionary history and  
761 differential expression regulation. *Gen Comp Endocrinol* **235**: 177–191.
- 762 Larhammar D, Risinger C. 1994. Molecular genetic aspects of tetraploidy in the common carp  
763 *Cyprinus carpio*. *Mol Phylogenet Evol* **3**: 59–68.
- 764 Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: A Bayesian software package for  
765 phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**: 2286–2288.
- 766 Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform.  
767 *Bioinformatics* **26**: 589–595.
- 768 Li L, Stoeckert CJJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic  
769 genomes. *Genome Res* **13**: 2178–2189.
- 770 Lien S, Koop BF, Sandve SR, Miller JR, Matthew P, Leong JS, Minkley DR, Zimin A, Grammes  
771 F, Grove H, et al. 2016. The Atlantic salmon genome provides insights into rediploidization.  
772 *Nature* **533**: 200–205.

- 773 Llorente B, Malpertuy A, Neuvéglise C, De Montigny J, Aigle M, Artiguenave F, Blandin G,  
774 Bolotin-Fukuhara M, Bon E, Brottier P, et al. 2000. Genomic exploration of the  
775 hemiascomycetous Yeasts: 18. comparative analysis of chromosome maps and synteny  
776 with *Saccharomyces cerevisiae*. *FEBS Lett* **487**: 101–112.
- 777 Lu J, Peatman E, Tang H, Lewis J, Liu Z. 2012. Profiling of gene duplication patterns of  
778 sequenced teleost genomes: evidence for rapid lineage-specific genome expansion  
779 mediated by recent tandem duplications. *BMC Genomics* **13**: 246.
- 780 Ludwig A, Belfiore NM, Pitra C, Svirsky V, Jenneckens I. 2001. Genome duplication events and  
781 functional reduction of ploidy levels in sturgeon (*Acipenser*, *Huso* and *Scaphirhynchus*).  
782 *Genetics* **158**: 1203–1215.
- 783 Macqueen DJ, Johnston IA, Macqueen DJ. 2014. A well-constrained estimate for the timing of  
784 the salmonid whole genome duplication reveals major decoupling from species  
785 diversification. *Proc R Soc B* **281**: 20132881.
- 786 Maugars G, Dufour S. 2015. Demonstration of the Coexistence of Duplicated LH Receptors in  
787 Teleosts, and Their Origin in Ancestral Actinopterygians. *PLoS One* **10**: e0135184.
- 788 Meyer A, Peer Y Van De. 2005. From 2R to 3R: evidence for a fish-specific genome duplication  
789 (FSGD). *BioEssays* **27**: 937–945.
- 790 Minegishi Y, Aoyama J, Inoue JG, Miya M. 2005. Molecular phylogeny and evolution of the  
791 freshwater eels genus *Anguilla* based on the whole mitochondrial genome sequences. *Mol*  
792 *Phylogenet Evol* **34**: 134–146.
- 793 Morini M, Pasquier J, Dirks R, Van Den Thillart G, Tomkiewicz J, Rousseau K, Dufour S, Lafont  
794 AG. 2015. Duplicated leptin receptors in two species of eel bring new insights into the  
795 evolution of the leptin system in vertebrates. *PLoS One* **10**: 1–31.
- 796 Morini M, Peñaranda DS, Vélchez MC, Nourizadeh-Lillabadi R, Lafont AG, Dufour S, Asturiano  
797 JF, Weltzien FA, Pérez L. 2017. Nuclear and membrane progesterin receptors in the  
798 European eel: Characterization and expression in vivo through spermatogenesis. *Comp*

- 799 *Biochem Physiol -Part A Mol Integr Physiol* **207**: 79–92.
- 800 Munk P, Hansen MM, Maes GE, Nielsen TG, Castonguay M, Riemann L, Sparholt H, Als TD,  
801 Aarestrup K, Andersen NG, et al. 2010. Oceanic fronts in the Sargasso Sea control the  
802 early life and drift of Atlantic eels. *Proc Biol Sci* **277**: 3593–3599.
- 803 Nagatsu T, Levitt M, Udenfriend S. 1964. Tyrosine Hydroxylase: The initial step in  
804 norepinephrine biosynthesis. *J Biol Chem* **239**: 2910–2917.
- 805 Ohno S. 1970. *Evolution by Gene Duplication*. Springer-Verlag, New York.
- 806 Pasqualini C, Weltzien FA, Vidal B, Baloche S, Rouget C, Gilles N, Servent D, Vernier P, Dufour  
807 S. 2009. Two distinct dopamine D2 receptor genes in the European eel: Molecular  
808 characterization, tissue-specific transcription, and regulation by sex steroids. *Endocrinology*  
809 **150**: 1377–1392.
- 810 Pasquier J, Cabau C, Nguyen T, Jouanno E, Severac D, Braasch I, Journot L, Pontarotti P,  
811 Klopp C, Postlethwait JH, et al. 2016. Gene evolution and gene expression after whole  
812 genome duplication in fish: the PhyloFish database. *BMC Genomics* **17**: 368.
- 813 Pasquier J, Lafont AG, Jeng SR, Morini M, Dirks R, van den Thillart G, Tomkiewicz J, Tostivint  
814 H, Chang CF, Rousseau K, et al. 2012. Multiple kisspeptin receptors in early osteichthyans  
815 provide new insights into the evolution of this receptor family. *PLoS One* **7**: e48931.
- 816 Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides accurate, fast,  
817 and bias-aware transcript expression estimates using dual-phase inference. *bioRxiv* **14**:  
818 21592.
- 819 Peña-Llopis S, Brugarolas J. 2013. Simultaneous isolation of high-quality DNA, RNA, miRNA  
820 and proteins from tissues for genomic applications. *Nat Protoc* **8**: 2240–55.
- 821 Perteu M, Kim D, Perteu GM, Leek JT, Salzberg SL. 2016. Transcript-level expression analysis  
822 of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc* **11**: 1650–1667.
- 823 Peter RE, Chang JP, Nahorniak RJ, Sokolowska OM, Shih SH, Billard R. 1986. Interaction of  
824 catecholamin and GnRH in regulation of gonadotropin secretion in teleost fish. *Recent*

- 825 *Prog Horm Res* **42**: 513–548.
- 826 Righton D, Westerberg H, Feunteun E, Okland F, Gargan P, Amilhat E, Metcalfe J, Lobon-  
827 Cervia J, Sjo berg N, Simon J, et al. 2016. Empirical observations of the spawning  
828 migration of European eels: The long and dangerous road to the Sargasso Sea. *Sci Adv* **2**:  
829 e1501694.
- 830 Rondeau EB, Minkley DR, Leong JS, Messmer AM, Jantzen JR, Von Schalburg KR, Lemon C,  
831 Bird NH, Koop BF. 2014. The genome and linkage map of the northern pike (*Esox lucius*):  
832 Conserved synteny revealed between the salmonid sister group and the neoteleostei.  
833 *PLoS One* **9**: e102089.
- 834 Rozenfeld C, Butts IAE, Tomkiewicz J, Zambonino-Infante J-L, Mazurais D. 2016. Abundance of  
835 specific mRNA transcripts impacts hatching success in European eel, *Anguilla anguilla* L.  
836 *Comp Biochem Physiol A Mol Integr Physiol* **191**: 59–65.
- 837 Santini F, Harmon LJ, Carnevale G, Alfaro ME. 2009. Did genome duplication drive the origin of  
838 teleosts? A comparative study of diversification in ray-finned fishes. *BMC Evol Biol* **9**: 194.
- 839 Schartl M, Walter RB, Shen Y, Garcia T, Catchen J, Amores A, Braasch I, Chalopin D, Volf J-N,  
840 Lesch K-P, et al. 2013. The genome of the platyfish, *Xiphophorus maculatus*, provides  
841 insights into evolutionary adaptation and several complex traits. *Nat Genet* **45**: 567–572.
- 842 Schmidt J. 1923. The breeding places of the eel. *Philos Trans R Soc B Biol Sci* **211**: 179–208.
- 843 Scott a P, Sumpter JP, Stacey N. 2010. The role of the maturation-inducing steroid, 17,20beta-  
844 dihydroxypregn-4-en-3-one, in male fishes: a review. *J Fish Biol* **76**: 183–224.
- 845 She X, Cheng Z, Zöllner S, Church DM, Eichler EE. 2008. Mouse segmental duplication and  
846 copy number variation. *Nat Genet* **40**: 909–14.
- 847 Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M,  
848 Söding J, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence  
849 alignments using Clustal Omega. *Mol Syst Biol* **7**: 539.
- 850 Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. 2015. BUSCO:

- 851           Assessing genome assembly and annotation completeness with single-copy orthologs.  
852           *Bioinformatics* **31**: 3210–3212.
- 853 Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence  
854           comparison. *BMC Bioinformatics* **6**: 31.
- 855 Stein C, Caccamo M, Laird G, Leptin M. 2007. Conservation and divergence of gene families  
856           encoding components of innate immune response systems in zebrafish. *Genome Biol* **8**:  
857           R251.
- 858 Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH. 2008. Unraveling ancient  
859           hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res* **18**: 1944–1954.
- 860 Tautz D, Domazet-lošo T. 2011. The evolutionary origin of orphan genes. *Nat Publ Gr* **12**: 692–  
861           702.
- 862 Tsukamoto K, Aoyama J, Miller MJ. 2002. Migration, speciation, and the evolution of diadromy  
863           in anguillid eels. *Can J Fish Aquat Sci* **59**: 1989–1998.
- 864 Tsukamoto Katsumi, Nakai Izumi TW-. V. 1998. Do all freshwater eels migrate? *Nature* **396**:  
865           635–636.
- 866 Uyeno T, Smith GR. 1972. Tetraploid origin of the karyotype of catostomid fishes. *Science* **175**:  
867           644–646.
- 868 Vallente Samonte R, Eichler EE. 2016. Segmental duplications and the evolution of the primate  
869           genome. *Nat Rev Genet* **3**: 65–72.
- 870 van Ginneken V, Antonissen E, Müller UK, Booms R, Eding E, Verreth J, van den Thillart G.  
871           2005. Eel migration to the Sargasso: remarkably high swimming efficiency and low energy  
872           costs. *J Exp Biol* **208**: 1329–35.
- 873 Vandepoele K, De Vos W, Taylor JS, Meyer A, Van de Peer Y. 2004. Major events in the  
874           genome evolution of vertebrates: paranome age and size differ considerably between ray-  
875           finned fishes and land vertebrates. *Proc Natl Acad Sci U S A* **101**: 1638–1643.
- 876 Vidal B, Pasqualini C, Le Belle N, Holland MCH, Sbaihi M, Vernier P, Zohar Y, Dufour S. 2004.

877 Dopamine inhibits luteinizing hormone synthesis and release in the juvenile European eel:  
878 a neuroendocrine lock for the onset of puberty. *Biol Reprod* **71**: 1491–500.

879 Wang J, Li J, Zhang X, Sun X. 2012. Transcriptome analysis reveals the time of the fourth round  
880 of genome duplication in common carp (*Cyprinus carpio*). *BMC Genet* **13**: 96.

881 Weltzien F, Pasqualini C, Se M, Vidal B, Belle N Le, Kah O, Vernier P, Dufour S. 2015.  
882 Androgen-dependent stimulation of brain dopaminergic systems in the female European  
883 eel (*Anguilla anguilla*). *Recent Prog Horm Res* **146**: 2964–2973.

884 Zhang J. 2003. Evolution by gene duplication: An update. *Trends Ecol Evol* **18**: 292–298.

885 Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W. 2008. On  
886 the origin of new genes in *Drosophila*. *Genome Res* **18**: 1446–1455.

887