

## A haplotype-resolved draft genome of the European sardine (*Sardina pilchardus*) --Manuscript Draft--

<b>Manuscript Number:</b>	GIGA-D-18-00377R2	
<b>Full Title:</b>	A haplotype-resolved draft genome of the European sardine ( <i>Sardina pilchardus</i> )	
<b>Article Type:</b>	Data Note	
<b>Funding Information:</b>	Fundação para a Ciência e a Tecnologia (UID/Multi/04326/2016)	Not applicable
	Fundação para a Ciência e a Tecnologia (22153-01/SAICT/2016)	Not applicable
	H2020 Research Infrastructures (654008)	Not applicable
	Programa Operacional Mar2020 (MAR-01.04.02-FEAMP-0024)	Not applicable
<b>Abstract:</b>	<p><b>Background:</b> The European sardine (<i>Sardina pilchardus</i> Walbaum, 1792) has a high cultural and economic importance throughout its distribution. Monitoring studies of sardine populations report an alarming decrease in stocks due to overfishing and environmental change, which has resulted in historically low captures along the Iberian Atlantic coast. Consequently, there is an urgent need to better understand the causal factors of this continuing decrease in the sardine stock. Important biological and ecological features such as levels of population diversity, structure, and migratory patterns can be addressed with the development and use of genomics resources.</p> <p><b>Findings:</b> The sardine genome of a single female individual was sequenced using Illumina HiSeq X Ten 10X Genomics linked-reads generating 113.8 Gb of data. Three draft genomes were assembled: two haploid genomes with a total size of 935 Mbp (N50 103Kb) each, and a consensus genome with a total size of 950 Mbp (N50 97Kb). The genome completeness assessment captured 84% of Actinopterygii Benchmarking Universal Single-Copy Orthologs. To obtain a more complete analysis, the transcriptomes of eleven tissues were sequenced and used to aid the functional annotation of the genome, resulting in 40 777 genes predicted. Variant calling on nearly half of the haplotype genome resulted in the identification of more than 2.3 million phased SNPs with heterozygous loci.</p> <p><b>Conclusions:</b> A draft genome was obtained with the 10X Genomics linked-reads technology, despite a high level of sequence repeats and heterozygosity that are expected genome characteristics of a wild sardine. The reference sardine genome and respective variant data are a cornerstone resource of ongoing population genomics studies to be integrated into future sardine stock assessment modelling to better manage this valuable resource.</p>	
<b>Corresponding Author:</b>	Adelino V. M. Canário PORTUGAL	
<b>Corresponding Author Secondary Information:</b>		
<b>Corresponding Author's Institution:</b>		
<b>Corresponding Author's Secondary Institution:</b>		
<b>First Author:</b>	Bruno Louro, PhD	
<b>First Author Secondary Information:</b>		
<b>Order of Authors:</b>	Bruno Louro, PhD	
	Gianluca De Moro, PhD	
	Carlos Garcia	
	Cymon J. Cox, PhD	

	Ana Veríssimo
	Stephen J. Sabatino
	António M. Santos
	Adelino V. M. Canário, PhD
<b>Order of Authors Secondary Information:</b>	
<b>Response to Reviewers:</b>	<p>Reply to reviewer reports:</p> <p>Reviewer #1: The manuscript is much improved and the annotation appears much better. All concerns appear to be addressed. One remaining question I have is how the 97 orthologs were identified to build the phylogeny. Did these come from the Orthofinder analysis? These methods to be elaborated on to show they are true orthologs. Otherwise the manuscript seems ready for publication.</p> <p>Reply Reviewer #1: The one-to-one ortholog clusters that we used to conduct the phylogenetic analyses were those assembled by Machado et al [20] to which we added the orthologous genes from our genome assembly as identified via HMMER. To make it clearer we modified the text (lines 310-315) to:          “We conducted a phylogenetic analysis of ray-finned fish (Actinopterygii) taxa based on 17 fish species. The sardine protein data set used in the phylogenetic analysis was obtained by querying the deduced proteins from our sardine genome against the one-to-one orthologous cluster dataset (106 proteins from 17 species) obtained from [20]. For the query, gene models were constructed for each protein with hmmbuild (HMMER v3.1b2) [53] using default options and the orthologous genes from the deduced sardine proteome were searched using hmmsearch (HMMER) with an e-value cutoff of 10e-3.”</p> <p>We don't describe the clustering methods used by Machado et al [20] to assemble the one-to-one ortholog clusters as we did not repeat those analyses ourselves.</p> <p>#####</p> <p>Reviewer #2: I see substantial improvements in the manuscript. One remaining issue is the quality of figures. The authors need to present the names of the species with consistency between Figures 3 and 4. Also, in Figure 4, the name of the main study species in this manuscript 'Sardine' should be consistently included in its Latin species name, <i>Sardina pilchardus</i>.</p> <p>Reply Reviewer #2: We modified and improved the quality of Figures 3 and 4. In both figures the species names are explicit in italic.</p>
<b>Additional Information:</b>	
<b>Question</b>	<b>Response</b>
Are you submitting this manuscript to a special series or article collection?	No
<b>Experimental design and statistics</b>	Yes
Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our <a href="#">Minimum Standards Reporting Checklist</a> . Information essential to interpreting the data presented should be made available in the figure legends.	
Have you included all the information	

<p>requested in your manuscript?</p>	
<p><b>Resources</b></p> <p>A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite <a href="#">Research Resource Identifiers</a> (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>?</p>	<p>Yes</p>
<p><b>Availability of data and materials</b></p> <p>All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in <a href="#">publicly available repositories</a> (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.</p> <p>Have you have met the above requirement as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>?</p>	<p>Yes</p>



[Click here to view linked References](#)

1 **A haplotype-resolved draft genome of the European sardine**

2 **(*Sardina pilchardus*)**

3 Bruno Louro<sup>1\*</sup>; Gianluca De Moro<sup>1\*</sup>; Carlos Garcia<sup>1</sup>; Cymon J. Cox<sup>1</sup>; Ana Veríssimo<sup>2</sup>;  
4 Stephen J. Sabatino<sup>2</sup>; António M. Santos<sup>2</sup>; Adelino V. M. Canário<sup>1&</sup>

5 1 CCMAR Centre of Marine Sciences, University of Algarve, Campus de Gambelas,  
6 8005-139 Faro, Portugal.

7 2 CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO,  
8 Laboratório Associado, Universidade do Porto, Vairão, Portugal

9

10 \* authors contributed equally

11 & Corresponding author: Adelino V. M. Canário, e-mail: [acanario@ualg.pt](mailto:acanario@ualg.pt)

12

13 **ORCID Details:**

14 Bruno Louro: 0000-0001-8164-581X;

15 Gianluca De Moro: 0000-0002-5542-0278;

16 Adelino V. M. Canário: 0000-0002-6244-6468

17

18

19

20

21

## 22 **Abstract**

23 **Background:** The European sardine (*Sardina pilchardus* Walbaum, 1792) has a  
24 high cultural and economic importance throughout its distribution. Monitoring studies  
25 of sardine populations report an alarming decrease in stocks due to overfishing and  
26 environmental change, which has resulted in historically low captures along the  
27 Iberian Atlantic coast. Consequently, there is an urgent need to better understand  
28 the causal factors of this continuing decrease in the sardine stock. Important  
29 biological and ecological features such as levels of population diversity, structure,  
30 and migratory patterns can be addressed with the development and use of genomics  
31 resources. **Findings:** The sardine genome of a single female individual was  
32 sequenced using Illumina HiSeq X Ten 10X Genomics linked-reads generating 113.8  
33 Gb of data. Three draft genomes were assembled: two haploid genomes with a total  
34 size of 935 Mbp (N50 103Kb) each, and a consensus genome with a total size of  
35 950 Mbp (N50 97Kb). The genome completeness assessment captured 84% of  
36 Actinopterygii Benchmarking Universal Single-Copy Orthologs. To obtain a more  
37 complete analysis, the transcriptomes of eleven tissues were sequenced and used to  
38 aid the functional annotation of the genome, resulting in 40,777 genes predicted.  
39 Variant calling on nearly half of the haplotype genome resulted in the identification of  
40 more than 2.3 million phased SNPs with heterozygous loci. **Conclusions:** A draft  
41 genome was obtained with the 10X Genomics linked-reads technology, despite a  
42 high level of sequence repeats and heterozygosity that are expected genome  
43 characteristics of a wild sardine. The reference sardine genome and respective  
44 variant data are a cornerstone resource of ongoing population genomics studies to  
45 be integrated into future sardine stock assessment modelling to better manage this  
46 valuable resource.

47 **Keywords:** European sardine; *Sardina*; genome; transcriptome; haplotype; SNP

48

## 49 **Data description**

### 50 **Background**

51 The European sardine (*Sardina pilchardus* Walbaum, 1792) (NCBI:txid27697,  
52 Fishbase ID:1350) (Figure 1) is a small pelagic fish occurring in temperate boundary  
53 currents of the Northeast Atlantic down to Cape Verde off the west coast of Africa,  
54 and throughout the Mediterranean to the Black Sea [1]. Two subspecies are  
55 generally recognised: *Sardina pilchardus pilchardus* occupies the north-eastern  
56 Atlantic and the North Sea whereas *S. pilchardus sardina* occupies the  
57 Mediterranean and Black seas, and the North African coasts south to Cape Verde,  
58 with a contact zone near the Strait of Gibraltar [1, 2]. As with other members of the  
59 Clupeidae family (e.g. herring, *Clupea harengus*) and allis shad (*Alosa alosa*) [3], the  
60 sardine experiences strong population fluctuations in abundance, possibly reflecting  
61 environmental fluctuations, including climate change [4, 5].

62 The sardine is of major economic and social importance throughout its range with a  
63 reported commercial catch for 2016 of 72,183 tonnes in European waters [6]. In  
64 Portugal, the sardine is an iconic and culturally revered fish and plays a central role  
65 in tourist events, such as summer festivals, throughout the country. However, recent  
66 stock assessment data strongly suggests the Iberian sardine fisheries is under  
67 threat. A recent report by the International Council for the Exploration of the Sea [6]  
68 noted a sharp decrease in the Iberian Atlantic coast sardine stock and advised that  
69 catches in 2017 should be no more than 23,000 tonnes. The sardine fishery biomass

70 has suffered from declining annual recruitment between 1978 and 2006, and more  
71 recently, it has fluctuated around historically low values indicating a high risk of  
72 collapse of the Iberian Atlantic stocks [6].

73 A number of sardine populations have been identified by morphometric methods,  
74 including as many as five populations in the north-eastern Atlantic (including the  
75 Azores), two off the Moroccan coast, and one in Senegalese waters [1, 7]. Each of  
76 these recognized sardine populations is subjected to specific climatic and oceanic  
77 conditions, mainly during larval development, which directly influence the recruitment  
78 of the sardine fisheries [4, 8, 9]. However, because of phenotypic plasticity,  
79 morphological traits are strongly influenced by environmental conditions and the  
80 underlying genetics that define those populations has proven elusive [10]. While the  
81 recognition of subspecies and localised populations might indicate significant genetic  
82 structure, the large population sizes and extensive migration of sardines are likely to  
83 increase gene flow and reduce population differences, suggesting, at its most  
84 extensive, a panmictic population with little genetic differentiation within the species'  
85 range [11].

86 It is now well established that to fully understand the genetic basis of evolutionarily  
87 and ecologically significant traits, the gene and regulatory element composition of  
88 different individuals or populations needs to be assessed [see e.g., 12, 13].  
89 Therefore, we provide a European sardine draft genome, providing the essential tool  
90 to assess the genetic structure of the sardine population(s) and for genetic studies of  
91 the life-history and ecological traits of this small pelagic fish, which will be  
92 instrumental for conservation and fisheries management.

## 93 Genome sequencing

94 Sardines were caught during commercial fishing operations in the coastal waters off  
95 Olhão, Portugal, and maintained live at the experimental fish culture facilities (EPPO)  
96 of the Portuguese Institute for the Sea and Atmosphere (IPMA), Olhão, Portugal [14].  
97 A single adult female was anesthetised with 2-phenoxyethanol (1:250 v/v), blood  
98 was collected in a heparinized syringe, and the fish euthanized by cervical section.  
99 Eleven tissues were dissected out - gill together with branchial arch, liver, spleen,  
100 ovary, midgut, white muscle, red muscle, kidney, head kidney, brain together with  
101 pituitary, and caudal fin (including skin, scales, bone and cartilage) – into RNA<sup>later</sup>  
102 (Sigma-Aldrich, USA) at room temperature followed by storage at -20 °C. Fish  
103 maintenance and sample collection were carried out in accordance with the  
104 guidelines of the European Union Council (86/609/EU) and Portuguese legislation for  
105 the use of laboratory animals from the Veterinary Medicines Directorate (DGAV), the  
106 Portuguese competent authority for the protection of animals, Ministry of Agriculture,  
107 Rural Development and Fisheries, Portugal (permit 010238 of 19/04/2016).

108 Total RNA was extracted using a total RNA purification kit (Maxwell® 16 Total RNA  
109 Purification Kit, Promega) and digested twice with DNase (DNA-free kit, Ambion,  
110 UK). The total RNA samples were kept at -80°C until shipment to the RNAseq  
111 service provider Admera Health Co. (USA) which confirmed a RIN above 8 (Qubit  
112 TapeStation) upon arrival. The mRNA library preparation was performed with  
113 NEBNext® Poly(A) mRNA Magnetic Isolation Module kit and NEBNext® Ultra™  
114 Directional RNA Library Prep kit for sequencing using Illumina HiSeq 4000 paired-  
115 end 150 bp cycle to generate about 596 million paired-end reads in total.

116 The genomic DNA (gDNA) was isolated from 20 µl of fresh blood using the DNeasy  
117 blood and tissue kit (Qiagen), followed by RNase treatment according to the



118 manufacturer's protocol. The integrity of the gDNA was confirmed using pulsed-field  
119 gel electrophoresis and showed fragment sizes largely above 50 kbp. The gDNA  
120 was stored at -20 °C before shipping to the service provider (Genome.one,  
121 Darlinghurst, Australia). Microfluidic partitioned gDNA libraries using the 10x  
122 Genomics Chromium System were made using 0.6 ng of gDNA input. Sequencing  
123 (150bp paired-end cycle) was performed in a single lane of the Illumina HiSeq X Ten  
124 instrument (Illumina, San Diego, CA, USA). Chromium library size range (580-850  
125 bp) was determined with LabChip GX Touch (PerkinElmer) and library yield (6.5-40  
126  $\mu$ M) by quantitative polymerase chain reaction.

## 127 Genome size estimation

128 A total of 759 million paired-end reads were generated representing 113.8 Gb  
129 nucleotide sequences with 76.1% bases  $\geq$  Q30. Raw reads were edited to trim 10X  
130 Genomics proprietary barcodes with a python script "filter\_10xReads.py" [15] prior to  
131 kmer counting with Jellyfish v2.2.10 (Jellyfish, RRID:SCR\_005491) [16]. Six hundred  
132 and seventy million edited reads (90.5 Gb) were used to obtain the frequency  
133 distribution of 23-mers. The histogram of the kmer counting distribution was plotted  
134 in GenomeScope v1.0.0 (GenomeScope, RRID:SCR\_017014) [17] (Figure 2) with  
135 maximum kmer coverage of 10,000 for estimation of genome size, heterozygosity  
136 and repeat content. The estimated sardine haploid genome size was 907 Mbp with a  
137 repeat content of 40.7% and a heterozygosity level of 1.43% represented in the first  
138 peak of the distribution. These high levels of heterozygosity and repeat content  
139 indicated a troublesome genome characteristic for *de novo* assembly.

## 140 *De novo* genome assembly

141 The *de novo* genome assembly was performed using the paired-end sequence  
142 reads from the partitioned library as input for the Supernova assembly algorithm  
143 v2.0.0 (7fba7b4) (Supernova assembler, RRID:SCR\_016756) (10x Genomics, San  
144 Francisco, CA, USA) [18]. Two haplotype-resolved genomes, SP\_haploid1 (ENA  
145 accession ID UOTT01000000) and SP\_haploid2 (ENA accession ID  
146 UOTU01000000), were assembled with phased scaffolds using the Supernova  
147 “mkoutput pseudohap” option. For the assembly process the Supernova run  
148 parameters for maximum reads (--maxreads) and barcode fraction (--barfrac) were  
149 set for 650M input reads and 80% of barcodes, respectively. Preliminary trials  
150 defined an optimal raw coverage of 78-fold, above the 56-fold suggested in the  
151 Supernova protocol; this reduced the problem (to some extent) of the complexity of  
152 the high repeat content (Table 1). A fraction of the 607.36 million read pairs were  
153 used after a quality control step embedded in the Supernova pipeline to remove  
154 reads that were not barcoded, not properly paired, or low-quality. Input reads had a  
155 138.5 bp mean length after proprietary 10X barcode trimming and a N50 of 612 per  
156 barcode/DNA molecule (Table 1).

157 Further scaffolding and gap closure procedures were performed with Rails  
158 v1.2/Cobbler v0.3 pipeline script [19] to obtain the final consensus genome  
159 sequence named SP\_G (ENA accession ID GCA\_900499035.1) using the  
160 parameters anchoring sequence length (-d 100) and minimum sequence identity (-i  
161 0.95). Three scaffolding and gap closure procedures were performed iteratively with  
162 one haplotype of the initial assembly as the assembly *per se*, and previous *de novo*  
163 assemblies from Supernova v1.2.2, (315M/100% and 450M/80% reads/barcodes).  
164 By closing several gaps within scaffolds and merging other scaffolds into longer and

165 fewer scaffolds (117,259), this procedure resulted into a slightly longer genome size  
166 of 949.62 Mb, which slightly deflated the scaffold N50 length to 96.6 Kb (Table 2).  
167 The assembly metrics of the three assemblies are described in Table 2 together with  
168 a recently published Illumina paired-end assembled sardine genome (UP\_Spi) [20].  
169 The total assembly size of our genome (SP\_G) is 950 Mb and the UP\_Spi is 641 Mb  
170 (Table 2). Because the SP\_G and UP\_Spi assembly sizes are of different orders of  
171 magnitude, in addition to N50 we present NG50 values [21] for an estimated genome  
172 size of 950 Mb (Table 2). In the SP\_G assembly, 905 scaffolds (LG50) represents  
173 half of the estimated genome with an NG50 value of 96.6 Kb, in comparison to LG50  
174 of 15 422 and NG50 of 12.6 Kb in the UP\_Spi assembly. The ungapped length of the  
175 SP\_G assembly is 828 Mb. The larger gaps of the SP\_G assembly compared to the  
176 UP\_Spi can be explained by the Supernova being able to estimate gap size based  
177 on the bar codes spanning the gaps, i.e. gaps have linkage evidence through the  
178 barcodes linking reads to DNA molecules and not solely gaps based on reads pairs  
179 [22]. Such gaps are reflected in the large number of N's per 100 kb in our assemblies  
180 (Table 2). The number of scaffolds in SP\_G is 117,259 (largest 6.843 Mb) and in  
181 UP\_Spi is 44,627 (largest 0.285 Mb).

182 The genome completeness assessment was estimated with Benchmarking Universal  
183 Single-copy Orthologs (BUSCO) v3.0.1 (BUSCO, RRID:SCR\_015008) [23]. The  
184 genome was queried (options -m geno -sp zebrafish) against the "metazoa.odbg9"  
185 lineage set containing 978 orthologs from sixty-five eukaryotic organisms to assess  
186 the coverage of core eukaryotic genes, and against the "actinopterygii.odbg9" lineage  
187 set containing 4,584 orthologs from 20 different ray-finned fish species as the most  
188 taxon-specific lineage available for the sardine. Using the metazoan odbg9 database,  
189 95.4% of the genome had significant matches: 84.5% were complete genes (76.7%

190 single-copy genes and 9.8% duplicates) and 8.9% were fragmented genes. By  
191 contrast, using the actinopterygii odb9 database, 84.2% (76.0% complete genes and  
192 8.2% fragmented) had a match, with 69.3% of genes occurring as single copy and  
193 6.7% as duplicates.

194 The EMBRIC configurator service [24] was used to create a fish specific checklist  
195 (named finfish) for the submission of the sardine genome project to the European  
196 Nucleotide Archive (ENA) (European Nucleotide Archive, RRID:SCR\_006515)  
197 (project accession PRJEB27990).

## 198 Repeat Content

199 The SP\_G consensus assembly was used as a reference genome to build a *de novo*  
200 repeat library running RepeatModeler v1.0.11 (RepeatModeler, RRID:SCR\_015027)  
201 [25] with default parameters. The model obtained from RepeatModeler was used,  
202 together with Dfam\_consensus database v20171107 [26] and RepBase  
203 RepeatMasker Edition library v20170127 [27] to identify repetitive elements and low  
204 complexity sequences running RepeatMasker v4.0.7 (RepeatMasker,  
205 RRID:SCR\_012954) [28]. The analysis carried out revealed that 23.33% of the  
206 assembled genome consists of repetitive elements.

## 207 Genome annotation

208 The Maker v2.31.10 (MAKER, RRID:SCR\_005309) [29] pipeline was used iteratively  
209 (five times) to annotate the SP\_G consensus genome. The annotations generated in  
210 each iteration were kept in the succeeding annotation steps and in the final General  
211 Feature Format (GFF) file. During the first Maker run the *de novo* transcriptome was  
212 mapped to the genome using blastn v2.7.1 (BLASTN, RRID:SCR\_001598) [30]  
213 (est2genome parameter in Maker). Moreover, the repetitive elements found with

214 RepeatMasker were used in the Maker pipeline. This initial gene models created by  
215 Maker were then used to train Hidden Markov Model (HMM) based gene predictors.  
216 The preliminary GFF file generated by this first iteration run was used as input to  
217 train SNAP v2006-07-28 [31]. Using the scripts provided directly by Maker  
218 (maker2zff) and SNAP (fathom, forge and hmm-assembler.pl) an HMM file was  
219 created and used as input for the next Maker iteration (snaphmm option in maker  
220 configuration file). For the next iteration, the gene-finding software Augustus v3.3  
221 (Augustus, RRID:SCR\_008417) [32] was self-trained running BUSCO with the  
222 specific parameter (--long), that turn on the Augustus optimization mode for self-  
223 training. The resulted predicted species model from Augustus was included in the  
224 pipeline in the third Maker run. For the fourth iteration, GeneMark-ES v4.32  
225 (GeneMark, RRID:SCR\_011930) [33], a self-training gene prediction software, was  
226 executed and the resulting HMM file was integrated into the Maker pipeline. As  
227 further evidence for the annotation, in the last run of Maker, the genome was queried  
228 using blastx v2.7.1 (BLASTX, RRID:SCR\_001653) (protein2genome parameter in  
229 Maker), against the deduced proteomes of herring (GCF\_000966335.1), (*Clupea*  
230 *harengus*, NCBI:txid7950, Fishbase ID:24) zebrafish (*Danio rerio*, NCBI:txid7955,  
231 Fishbase ID:4653) (GCF\_000002035.6), blind cave fish (*Astyanax*  
232 *mexicanus*, NCBI:txid7994, Fishbase ID:2740) (GCF\_000372685.2), European  
233 sardine [20] and all proteins from teleost fishes in the UniProtKB/Swiss-Prot  
234 database (UniProtKB, RRID:SCR\_004426) [34]. After the five Maker runs the  
235 selected 40,777 genes models are the *ab initio* predictions supported by the  
236 transcriptome and proteome evidence. Based on the transcriptomic evidence,  
237 12,761 gene models were annotated with untranslated regions (UTR) features, more

238 specifically 9 486 gene models with either 5' or 3' UTR and 3,275 gene models with  
239 both UTR features.

240 InterProScan v. 5.30 (InterProScan, RRID:SCR\_005829) [35] and NCBI blastp  
241 v2.8.1 (BLASTP, RRID:SCR\_001010) [30] were used to functionally annotate the  
242 40,777 predicted protein coding genes. Thirty-three thousand five hundred and fifty-  
243 three (33,553) (82.3%) proteins were successfully annotated using blastp (e-value  
244 1e-05) against the UniProtKB/Swiss-Prot database and another 5,228 were  
245 annotated using the NCBI non-redundant protein database (nr). In addition to the  
246 above, 37,075 (90.9%) proteins were successfully annotated using InterProScan  
247 with all the InterPro v72.0 (InterPro, RRID:SCR\_006695) [36] databases: CATH-  
248 Gene3D (Gene3D, RRID:SCR\_007672), Hamap (HAMAP, RRID:SCR\_007701),  
249 PANTHER (PANTHER, RRID:SCR\_004869), Pfam (Pfam, RRID:SCR\_004726),  
250 PIRSF (PIRSF, RRID:SCR\_003352), PRINTS (PRINTS, RRID:SCR\_003412),  
251 ProDom (ProDom, RRID:SCR\_006969), ProSite Patterns (PROSITE,  
252 RRID:SCR\_003457), ProSite Profiles, SFLD (Structure-function linkage database,  
253 RRID:SCR\_001375), SMART (SMART, RRID:SCR\_005026), SUPERFAMILY  
254 (SUPERFAMILY, RRID:SCR\_007952), and TIGRFAM (JCVI TIGRFAMS,  
255 RRID:SCR\_005493). In total, 38 880 (95.3%) of the predicted proteins received a  
256 functional annotation. The annotated genome assembly is published [37] in the wiki-  
257 style annotation portal ORCAE [38] .

258 OrthoFinder v2.2.7 [39] was used to identify paralogy and orthology in our Swiss-prot  
259 annotated deduced proteome and in the deduced proteomes from herring, blind cave  
260 fish and zebrafish. The resulting orthogroups were plotted using jvenn (jVenn,  
261 RRID:SCR\_016343) [40] (Figure 3), where paralogous (two or more genes) and  
262 singletons were identified within species specific orthogroups. The deduced

263 sardine proteome has 3,413 paralogous groups containing 11406 genes, of which 31  
264 are sardine specific orthogroups. The amount of sardine singletons (9,856) can be  
265 partially due to fragmented predicted genes, but can reflect also some evolutionary  
266 divergence which requires further study to understand the biological relevance. In  
267 total, 25,560 orthogroups containing at least a single protein were identified in  
268 sardine, of which 12958 orthogroups are common to all four fish species. Within the  
269 Clupeidae, the sardine and the herring share 14,780 orthogroups with 922 family-  
270 specific orthogroups.

### 271 Variant calling between phased alleles

272 FASTQ files were processed using the 10x Genomics LongRanger v2.2.2 pipeline  
273 [41] with a maximum input limit of one thousand scaffolds, defined as reference  
274 genome, and representing about half of the genome size (488.5 Mb). The  
275 LongRanger pipeline was run with default settings, with the exception of vcmode  
276 to define the Genome Analysis Toolkit (GATK) v4.0.3.0 (GATK,  
277 RRID:SCR\_001876) [42] as the variant caller and the somatic parameters. The  
278 longest phase block was 2.86 Mb and the N50 phase block was 0.476 Mb.

279 Single nucleotide polymorphisms (SNP's) were furthered filtered to obtain only  
280 phased and heterozygous SNP's between the two alleles with a coverage higher  
281 than 10-fold using VCFtools v0.1.16 (VCFtools, RRID:SCR\_001235). A VCF file was  
282 obtained containing 2,369,617 filtered SNPs (Additional file 1) resulting in a mean  
283 distance between heterozygous phased SNPs of 206 bp. Similar results were  
284 obtained in the Supernova input report estimation (Table 1) of mean distance  
285 between heterozygous SNPs in the whole genome of 197 bp. This high SNP  
286 heterozygosity (1/206), observed solely in the comparison of the phased alleles, is

287 higher than the average nucleotide diversity of the previously reported marine fish of  
288 wild populations: 1/390 in yellow drum [43], 1/309 in herring [44], 1/435 in coelacanth  
289 [45], 1/500 in cod [46] and 1/700 in stickleback [47].

## 290 *De novo* transcriptome assembly

291 The 596 million paired-end raw transcriptomic reads were edited for contamination  
292 (e.g. adapters) using TrimGalore v0.4.5 wrapper tool (TrimGalore,  
293 RRID:SCR\_016946) [15], low-quality base trimming with Cutadapt v1.15 (cutadapt,  
294 RRID:SCR\_011841) [48] and the output overall quality reports of the edited reads  
295 with FastQC v0.11.5 (FastQC, RRID:SCR\_014583) [49].

296 The 553 million edited paired-end reads were *de novo* assembled as a multi-tissue  
297 assembly using Trinity v2.5.1 (Trinity, RRID:SCR\_013048) [50] with a minimum  
298 contig length of 200 bp, 50x coverage read depth normalization, and RF strand-  
299 specific read orientation. The same parameters were used for each of the 11 tissue  
300 specific *de novo* assemblies. The genome and transcriptome assemblies were  
301 conducted on the Portuguese National Distributed Computing Infrastructure [49].

302 The twelve *de novo* transcriptome assemblies (Table 3) were each quality assessed  
303 using TransRate v1.0.3 [51] with read evidence for assembly optimization, by  
304 specifying the contigs fasta file and respective left and right edited reads to be  
305 mapped. The multi-tissue assembly with all reads resulted in an assembled  
306 transcriptome of 170,478 transcript contigs following the TransRate step. Functional  
307 annotation was performed using the Trinotate v3.1.1 pipeline [24] and integrated into  
308 a SQLite database. All annotations were based on the best deduced open reading  
309 frame (ORF) obtained with the Transdecoder v1.03 [51]. Of the 170 478 transcripts  
310 contigs, 27,078 (16%) were inferred to ORF protein sequences. Query of the



311 UniProtKB/Swiss-Prot (e-value cutoff of 1e-5) database via blastx v2.7.1 of total  
312 contigs resulted in 43 458 (26%) annotated transcripts. The ORFs were queried  
313 against UniProtKB/Swiss-Prot (e-value cutoff of 1e-5) via blastp v2.7.1 and PFAM  
314 using hmmscan (HMMER v3.1b2) (Hmmer, RRID:SCR\_005305) [52] resulting in  
315 19,705 (73% of ORF) and 16 538 (61% of ORF) UniProtKB/Swiss-Prot and PFAM  
316 annotated contigs respectively. The full annotation report with further functional  
317 annotation, such as signal peptides, transmembrane regions, eggnoG, Kyoto  
318 Encyclopedia of Genes and Genomes (KEGG) (KEGG, RRID:SCR\_012773), and  
319 Gene Ontology annotation (Gene Ontology, RRID:SCR\_002811) are listed in tabular  
320 format in Additional file 2.

## 321 **Ray-finned fish phylogeny**

322 We conducted a phylogenetic analysis of ray-finned fish (Actinopterygii) taxa based  
323 on 17 fish species. The sardine protein data set used in the phylogenetic analysis  
324 was obtained by querying the deduced proteins from our sardine genome against the  
325 one-to-one orthologous cluster dataset (106 proteins from 17 species) obtained from  
326 [20].

327 For the query, gene models were constructed for each protein with hmmbuild  
328 (HMMER v3.1b2) [53] using default options and the orthologous genes from the  
329 deduced sardine proteome were searched using hmmsearch (HMMER) with an e-  
330 value cutoff of 10e-3. The best protein hits, as indicated by the bitscores, were  
331 aligned to the original protein sequence alignments using hmmsalign (HMMER) with  
332 default options. Gapped and poorly aligned sites were identified by Gblocks v0.91b  
333 (Gblocks, RRID:SCR\_015945) [54] using default options and removed using p4  
334 v1.3.0 [55]. Protein alignment statistics were calculated, and the proteins

335 concatenated into a single alignment using novel scripts in p4. Of the 106 fish  
336 proteins alignments, 97 contained sites which were considered correctly aligned by  
337 the Gblocks analysis; statistics for these alignments are presented in Table S1  
338 (Additional file 3). The concatenated sequence alignment of the 97 proteins  
339 contained 14,515 sites without gaps of which 7,391 were constant, 7,123 variable,  
340 and 3,879 parsimony informative.

341 The best-fitting empirical protein model of the concatenated data was evaluated  
342 using ModelFinder [56] in IQ-TREE v1.6.7.1 [57]. The best-fitting empirical  
343 substitution model was estimated to be the JTT model [58] with a discrete gamma-  
344 distribution of among-site rate variation (4 categories) and empirical composition  
345 frequencies (typical notation: JTT+ $\Gamma_4$ +F).

346 Optimal maximum likelihood tree searches (100 replicates) and bootstrap analyses  
347 (300 replicates) were conducted using RAxML v8.2.12 (RAxML, RRID:SCR\_006086)  
348 [59] with the best-fitting model. The optimal maximum likelihood tree (-ln likelihood:  
349 146565.6438) is presented in Figure 4 with bootstrap support values given at nodes,  
350 and is rooted to the outgroups *Petromyzon marinus* (lamprey) and *Latimeria*  
351 *chalumnae* (coelacanth).

352

## 353 **Conclusion**

354 Despite the sardine genome having a high level of repeats and heterozygosity,  
355 factors which pose a challenge to *de novo* genome assembly, a more than adequate  
356 draft genome was obtained with the 10X Genomics linked-reads (Chromium)  
357 technology. The Chromium technology's ability to tag and cluster the reads to  
358 individual DNA molecules has proven advantages for scaffolding, just as long reads

359 technologies such as Nanopore and Pacific Biosciences, but with high coverage and  
360 low error rates. The advantage of linked-reads for *de novo* genomic assemblies is  
361 evident in comparison to typical short read data, especially in the case of wild  
362 species with highly heterozygous genomes, where the latter often result in many  
363 uncaptured genomic regions and with a lower scaffolding yield due to repeated  
364 content.

365 The high degree of heterozygosity identified here in the sardine genome illustrates I  
366 future problems for monitoring sardine populations using low-resolution genetic data.  
367 However, the phased SNPs obtained in this study can be used to initiate the  
368 development of a SNP genetic panel for population monitoring, with SNPs  
369 representative of haplotype blocks, allowing insights into the patterns of linkage  
370 disequilibrium and the structure of haplotype blocks across populations.

371 The genomic and transcriptomic resources reported here are important tools for  
372 future studies to understand sardine response at the levels of physiology, population  
373 genetics and ecology of the causal factors responsible for the recruitment and  
374 collapse of the sardine stock in Iberian Atlantic coast. Besides the commercial  
375 interest, the sardine plays a crucial role at a key trophic level by bridging energy from  
376 the primary producers to the top predators in the marine ecosystem. Therefore,  
377 disruption of the sardine population equilibrium is likely to reverberate throughout the  
378 food chain via a trophic cascade. Consequently, these genomic and genetic  
379 resources are the prerequisites needed to develop tools to monitor the population  
380 status of the sardine and thereby provide an important bio-monitoring system for the  
381 health of the marine environment.

## 382 **Availability of the supporting data**

383 Raw data, assembled transcriptomes, and assembled genomes are available at the  
384 European Bioinformatics Institute ENA archive with the project accession  
385 PRJEB27990. The annotated genome assembly is published in the wiki-style  
386 annotation portal ORCAE [37]. Supporting data and materials are available in the  
387 *GigaScience* GigaDB database [60].

## 388 **Abbreviation**

389 gDNA: genomic DNA; BUSCO: Benchmarking Universal Single-copy Orthologs;  
390 GFF: General Feature Format; HMM: Hidden Markov Model; KEGG: Kyoto  
391 Encyclopedia of Genes and Genomes

## 392 **Acknowledgements**

393 This research was supported by national funds from FCT - Foundation for Science  
394 and Technology through project UID/Multi/04326/2016 and by FCT and FEDER  
395 under projects 22153-01/SAICT/2016 (to INCD), ALG-01-0145-FEDER-022121 and  
396 ALG-01-0145-FEDER-022231; and co-funds from MAR2020 operational programme  
397 of the European Maritime and Fisheries Fund (project SARDINOMICS MAR-  
398 01.04.02-FEAMP-0024). The EMBRIC configurator service received funding from the  
399 European Union's Horizon 2020 research and innovation programme under grant  
400 agreement No 654008. The authors acknowledge Pedro Guerreiro for providing the  
401 sardine samples.

402

## 403 References

- 404 1. Parrish RH, Serra R and Grant WS. The monotypic sardines, *Sardina* and  
405 *Sardinops* - Their taxonomy, distribution, stock structure, and zoogeography.  
406 Can J Fish Aquat Sci. 1989;46 11:2019-36. doi:10.1139/f89-251.
- 407 2. Silva A. Morphometric variation among sardine (*Sardina pilchardus*)  
408 populations from the northeastern Atlantic and the western Mediterranean.  
409 ICES J Mar Sci. 2003;60 6:1352-60. doi:10.1016/S1054-3139(03)00141-3.
- 410 3. Lavoue S, Miya M, Saitoh K, Ishiguro NB and Nishida M. Phylogenetic  
411 relationships among anchovies, sardines, herrings and their relatives  
412 (Clupeiformes), inferred from whole mitogenome sequences. Mol Phylogenet  
413 Evol. 2007;43 3:1096-105. doi:10.1016/j.ympev.2006.09.018.
- 414 4. Santos AMP, Borges MDF and Groom S. Sardine and horse mackerel  
415 recruitment and upwelling off Portugal. ICES J Mar Sci. 2001;58 3:589-96.  
416 doi:10.1006/jmsc.2001.1060.
- 417 5. Checkley Jr. DM, Asch RG and Rykaczewski RR. Climate, anchovy, and  
418 sardine. Annual Review of Marine Science. 2017;9 1:469-93.  
419 doi:10.1146/annurev-marine-122414-033819.
- 420 6. ICES. *Report of the Working Group on Southern Horse Mackerel, Anchovy*  
421 *and Sardine (WGHANSA), 24–29 June 2017, Bilbao, Spain. CM*  
422 *2017/ACOM:17, 640 p. 2017.*
- 423 7. Atarhouch T, Ruber L, Gonzalez EG, Albert EM, Rami M, Dakkak A, et al.  
424 Signature of an early genetic bottleneck in a population of Moroccan sardines  
425 (*Sardina pilchardus*). Mol Phylogenet Evol. 2006;39 2:373-83.  
426 doi:10.1016/j.ympev.2005.08.003.

- 427 8. Santos MB, Gonzalez-Quiros R, Riveiro I, Cabanas JM, Porteiro C and Pierce  
428 GJ. Cycles, trends, and residual variation in the Iberian sardine (*Sardina*  
429 *pilchardus*) recruitment series and their relationship with the environment.  
430 ICES J Mar Sci. 2012;69 5:739-50. doi:10.1093/icesjms/fsr186.
- 431 9. Leitao F, Alms V and Erzini K. A multi-model approach to evaluate the role of  
432 environmental variability and fishing pressure in sardine fisheries. J Mar Syst.  
433 2014;139:128-38. doi:10.1016/j.jmarsys.2014.05.013.
- 434 10. Tinti F, Di Nunno C, Guarniero I, Talenti M, Tommasini S, Fabbri E, et al.  
435 Mitochondrial DNA sequence variation suggests the lack of genetic  
436 heterogeneity in the Adriatic and Ionian stocks of *Sardina pilchardus*. Mar  
437 Biotechnol (NY). 2002;4 2:163-72. doi:10.1007/s10126-002-0003-3.
- 438 11. Jemaa S, Bacha M, Khalaf G, Dessailly D, Rabhi K and Amara R. What can  
439 otolith shape analysis tell us about population structure of the European  
440 sardine, *Sardina pilchardus*, from Atlantic and Mediterranean waters? J Sea  
441 Res. 2015;96:11-7. doi:10.1016/j.seares.2014.11.002.
- 442 12. Boehm JT, Waldman J, Robinson JD and Hickerson MJ. Population genomics  
443 reveals seahorses (*Hippocampus erectus*) of the western mid-Atlantic coast to  
444 be residents rather than vagrants. PLoS One. 2015;10 1:e0116219.  
445 doi:10.1371/journal.pone.0116219.
- 446 13. Hendricks S, Anderson EC, Antao T, Bernatchez L, Forester BR, Garner B, et  
447 al. Recent advances in conservation and population genomics data analysis.  
448 Evol Appl. 2018;11 8:1197-211. doi:10.1111/eva.12659.
- 449 14. Marcalo A, Guerreiro PM, Bentes L, Rangel M, Monteiro P, Oliveira F, et al.  
450 Effects of different slipping methods on the mortality of sardine, *Sardina*  
451 *pilchardus*, after purse-seine capture off the Portuguese Southern coast

- 452 (Algarve). PLoS One. 2018;13 5:e0195433.  
453 doi:10.1371/journal.pone.0195433.
- 454 15. Krueger F: "Trim galore" A wrapper tool around Cutadapt and FastQC to  
455 consistently apply quality and adapter trimming to FastQ files.  
456 [https://www.bioinformatics.babraham.ac.uk/projects/trim\\_galore/](https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) (2015).  
457 Accessed 9/24/2018.
- 458 16. Marcais G and Kingsford C. A fast, lock-free approach for efficient parallel  
459 counting of occurrences of k-mers. Bioinformatics. 2011;27 6:764-70.  
460 doi:10.1093/bioinformatics/btr011.
- 461 17. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski  
462 J, et al. GenomeScope: fast reference-free genome profiling from short reads.  
463 Bioinformatics. 2017;33 14:2202-4. doi:10.1093/bioinformatics/btx153.
- 464 18. Weisenfeld NI, Kumar V, Shah P, Church DM and Jaffe DB. Direct  
465 determination of diploid genome sequences. Genome Res. 2017;27 5:757-67.  
466 doi:10.1101/gr.214874.116.
- 467 19. Warren RL. RAILS and Cobbler: Scaffolding and automated finishing of draft  
468 genomes using long DNA sequences. JOSS. 2016;1 7:116.  
469 doi:10.21105/joss.00116.
- 470 20. Machado A, Tørresen O, Kabeya N, Couto A, Petersen B, Felício M, et al.  
471 "Out of the Can": A draft genome assembly, liver transcriptome, and  
472 nutrigenomics of the European sardine, *Sardina pilchardus*. Genes. 2018;9  
473 10:485. doi:10.3390/genes9100485.
- 474 21. Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, et al. Assemblathon 1:  
475 a competitive assessment of de novo short read assembly methods. Genome  
476 Res. 2011;21 12:2224-41. doi:10.1101/gr.126599.111.

- 477 22. Weisenfeld NI, Kumar V, Shah P, Church DM and Jaffe DB. Direct  
478 determination of diploid genome sequences. bioRxiv. 2016:070425.  
479 doi:10.1101/070425.
- 480 23. Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov  
481 G, et al. BUSCO applications from quality assessments to gene prediction  
482 and phylogenomics. Mol Biol Evol. 2017;35 3:543-8.  
483 doi:10.1093/molbev/msx319.
- 484 24. Bryant DM, Johnson K, DiTommaso T, Tickle T, Couger MB, Payzin-Dogru D,  
485 et al. A tissue-mapped axolotl *de novo* transcriptome enables identification of  
486 limb regeneration factors. Cell Rep. 2017;18 3:762-76.  
487 doi:10.1016/j.celrep.2016.12.063.
- 488 25. Smit A and Hubley R: RepeatModeler Open-1.0. <http://www.repeatmasker.org>  
489 (2008). Accessed 9/24/2018.
- 490 26. Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, et al. The Dfam  
491 database of repetitive DNA families. Nucleic Acids Res. 2016;44 D1:D81-9.  
492 doi:10.1093/nar/gkv1272.
- 493 27. Bao W, Kojima KK and Kohany O. Repbase Update, a database of repetitive  
494 elements in eukaryotic genomes. Mob DNA. 2015;6 1:11.  
495 doi:10.1186/s13100-015-0041-9.
- 496 28. Smit A, Hubley R and Green P: 2013–2015. RepeatMasker Open-4.0.  
497 <http://www.repeatmasker.org> (2013).
- 498 29. Holt C and Yandell M. MAKER2: an annotation pipeline and genome-  
499 database management tool for second-generation genome projects. BMC  
500 Bioinformatics. 2011;12 1:491. doi:10.1186/1471-2105-12-491.



- 501 30. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al.  
502 BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421-  
503 doi:10.1186/1471-2105-10-421.
- 504 31. Korf I. Gene finding in novel genomes. *BMC Bioinformatics*. 2004;5 1:59.  
505 doi:10.1186/1471-2105-5-59.
- 506 32. Stanke M and Waack S. Gene prediction with a hidden Markov model and a  
507 new intron submodel. *Bioinformatics*. 2003;19 suppl\_2:ii215-ii25.  
508 doi:10.1093/bioinformatics/btg1080.
- 509 33. Lomsadze A, Burns PD and Borodovsky M. Integration of mapped RNA-Seq  
510 reads into automatic training of eukaryotic gene finding algorithm. *Nucleic  
511 Acids Res*. 2014;42 15:e119. doi:10.1093/nar/gku557.
- 512 34. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, et al.  
513 UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*. 2004;32  
514 Database issue:D115-9. doi:10.1093/nar/gkh131.
- 515 35. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan  
516 5: genome-scale protein function classification. *Bioinformatics*. 2014;30  
517 9:1236-40. doi:10.1093/bioinformatics/btu031.
- 518 36. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, et al.  
519 InterPro in 2017—beyond protein family and domain annotations. *Nucleic  
520 Acids Res*. 2016;45 D1:D190. doi:10.1093/nar/gkw1107.
- 521 37. Sardine Genome Annotation Portal.  
522 <https://bioinformatics.psb.ugent.be/orcae/overview/Spil>. Accessed 9/24/2018.
- 523 38. Sterck L, Billiau K, Abeel T, Rouze P and Van de Peer Y. ORCAE: online  
524 resource for community annotation of eukaryotes. *Nat Methods*. 2012;9  
525 11:1041. doi:10.1038/nmeth.2242.

- 526 39. Emms DM and Kelly S. OrthoFinder: solving fundamental biases in whole  
527 genome comparisons dramatically improves orthogroup inference accuracy.  
528 Genome Biol. 2015;16 1:157. doi:10.1186/s13059-015-0721-2.
- 529 40. Bardou P, Mariette J, Escudie F, Djemiel C and Klopp C. jvenn: an interactive  
530 Venn diagram viewer. BMC Bioinformatics. 2014;15 1:293. doi:10.1186/1471-  
531 2105-15-293.
- 532 41. Zheng GX, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, et al.  
533 Haplotyping germline and cancer genomes with high-throughput linked-read  
534 sequencing. Nat Biotechnol. 2016;34 3:303-11. doi:10.1038/nbt.3432.
- 535 42. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et  
536 al. The Genome Analysis Toolkit: A MapReduce framework for analyzing  
537 next-generation DNA sequencing data. Genome Res. 2010;20 9:1297-303.  
538 doi:10.1101/gr.107524.110.
- 539 43. Han Z, Li W, Zhu W, Sun S, Ye K, Xie Y, et al. Near-complete genome  
540 assembly and annotation of the yellow drum (*Nibea albiflora*) provide insights  
541 into population and evolutionary characteristics of this species. Ecology and  
542 Evolution. 2019;9 1:568-75. doi:doi:10.1002/ece3.4778.
- 543 44. Barrio AM, Lamichhaney S, Fan GY, Rafati N, Pettersson M, Zhang H, et al.  
544 The genetic basis for ecological adaptation of the Atlantic herring revealed by  
545 genome sequencing. Elife. 2016;5:e12081. doi:10.7554/eLife.12081.
- 546 45. Amemiya CT, Alföldi J, Lee AP, Fan SH, Philippe H, MacCallum I, et al. The  
547 African coelacanth genome provides insights into tetrapod evolution. Nature.  
548 2013;496 7445:311-6. doi:10.1038/nature12027.

- 549 46. Star B, Nederbragt AJ, Jentoft S, Grimholt U, Malmstrom M, Gregers TF, et  
550 al. The genome sequence of Atlantic cod reveals a unique immune system.  
551 Nature. 2011;477 7363:207-10. doi:10.1038/nature10342.
- 552 47. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, et al.  
553 The genomic basis of adaptive evolution in threespine sticklebacks. Nature.  
554 2012;484 7392:55-61. doi:10.1038/nature10944.
- 555 48. Martin M. Cutadapt removes adapter sequences from high-throughput  
556 sequencing reads. EMBnet journal. 2011;17 1:10-2. doi:10.14806/ej.17.1.200.
- 557 49. Andrews S: FastQC: a quality control tool for high throughput sequence data.  
558 <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (2010). Accessed  
559 9/24/2018.
- 560 50. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et  
561 al. De novo transcript sequence reconstruction from RNA-seq using the Trinity  
562 platform for reference generation and analysis. Nat Protoc. 2013;8 8:1494-  
563 512. doi:10.1038/nprot.2013.084.
- 564 51. Smith-Unna R, Boursnell C, Patro R, Hibberd JM and Kelly S. TransRate:  
565 reference-free quality assessment of de novo transcriptome assemblies.  
566 Genome Res. 2016;26 8:1134-44. doi:10.1101/gr.196469.115.
- 567 52. Finn RD, Clements J and Eddy SR. HMMER web server: interactive  
568 sequence similarity searching. Nucleic Acids Res. 2011;39 Web Server  
569 issue:W29-37. doi:10.1093/nar/gkr367.
- 570 53. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998;14 9:755-63.
- 571 54. Castresana J. Selection of conserved blocks from multiple alignments for their  
572 use in phylogenetic analysis. Mol Biol Evol. 2000;17 4:540-52. doi:DOI  
573 10.1093/oxfordjournals.molbev.a026334.

- 574 55. Foster PG. Modeling compositional heterogeneity. *Syst Biol.* 2004;53 3:485-  
575 95.
- 576 56. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A and Jeremiin LS.  
577 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat*  
578 *Methods.* 2017;14 6:587-9. doi:10.1038/nmeth.4285.
- 579 57. Nguyen LT, Schmidt HA, von Haeseler A and Minh BQ. IQ-TREE: a fast and  
580 effective stochastic algorithm for estimating maximum-likelihood phylogenies.  
581 *Mol Biol Evol.* 2015;32 1:268-74. doi:10.1093/molbev/msu300.
- 582 58. Jones DT, Taylor WR and Thornton JM. The rapid generation of mutation  
583 data matrices from protein sequences. *Comput Appl Biosci.* 1992;8 3:275-82.
- 584 59. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-  
585 analysis of large phylogenies. *Bioinformatics.* 2014;30 9:1312-3.  
586 doi:10.1093/bioinformatics/btu033.
- 587 60. Louro B, Moro GD, Garcia C, Cox CJ, Veríssimo A, Sabatino SJ, Santos AM,  
588 Canário AVM (2019): Supporting data for "A haplotype-resolved draft genome  
589 of the European sardine (*Sardina pilchardus*)" GigaScience Database.  
590 <http://dx.doi.org/10.5524/100596>  
591  
592

## 593 Figure legends

594 Figure 1. The European sardine, *Sardina pilchardus* (photo credit ©[Eduardo Soares](#),  
595 [IPMA](#))

596

597 Figure 2. The histogram of the 23-mer depth distribution was plotted in  
598 GenomeScope [17] to estimate genome size (907Mb), repeat content (40.7%) and  
599 heterozygosity level (1.43%). Two kmer coverage peaks are observed at 28X and  
600 50X.

601

602 Figure 3. Optimal maximum likelihood tree (-ln likelihood: 146565.6438) under a  
603 best-fitting JTT+ $\Gamma_4$ +F substitution model of 97 concatenated proteins. Maximum  
604 likelihood bootstrap support values are given below or to the right of nodes. Scale  
605 bar represents mean numbers of substitutions per site. The Actinopterygii ingroup  
606 was rooted to two outgroup taxa, namely *Petromyzon marinus* (lamprey) and  
607 *Latimeria chalumnae* (coelacanth) (not shown).

608

609 Figure 4. Venn diagram representing paralogous and orthologous groups  
610 between sardine, blind cave fish, zebrafish, and herring obtained with OrthoFinder  
611 and plotted with Jvenn [40]. Orthogroups of singleton genes are showed in  
612 parenthesis.

613

## 614 Additional files

615 **Additional file 1.** Heterozygous SNPs identified in the phased haploid blocks listed

616 in a VCF file format.

617

618 **Additional file 2.** Annotation of all tissues transcriptome assembly in a tabular

619 format.

620

621 **Additional file 3.** Sequence alignment statistics of the 97 proteins concatenated for

622 the phylogenetics analyses.

Table 1. Descriptive metrics, estimated by Supernova, of the input sequence data for the *de novo* genome assembly.

Number of paired reads used	607.36 M
Mean read length after trimming	138.50 bp
Median insert size	345 bp
Weighted mean DNA molecule size	46.41 Kb
N50 reads per barcode	612
Raw coverage	78.35 X
Effective read coverage	52.91 X
Mean distance between heterozygous SNPs	197 bp

Table 2. Descriptive metrics of sardine genome assemblies. SP\_haploid1/ SP\_haploid2: haploids genomes ([UOTT01000000](#) and [UOTU01000000](#)). SP\_G: consensus genome (NCBI representative genome assembly, GCA\_900499035.1). UP\_Spi: Illumina paired-end assembled genome from [20] (GCA\_003604335.1). Values for scaffolds equal or larger than 1Kb, 10Kb and 100 Kb are presented in separated rows.

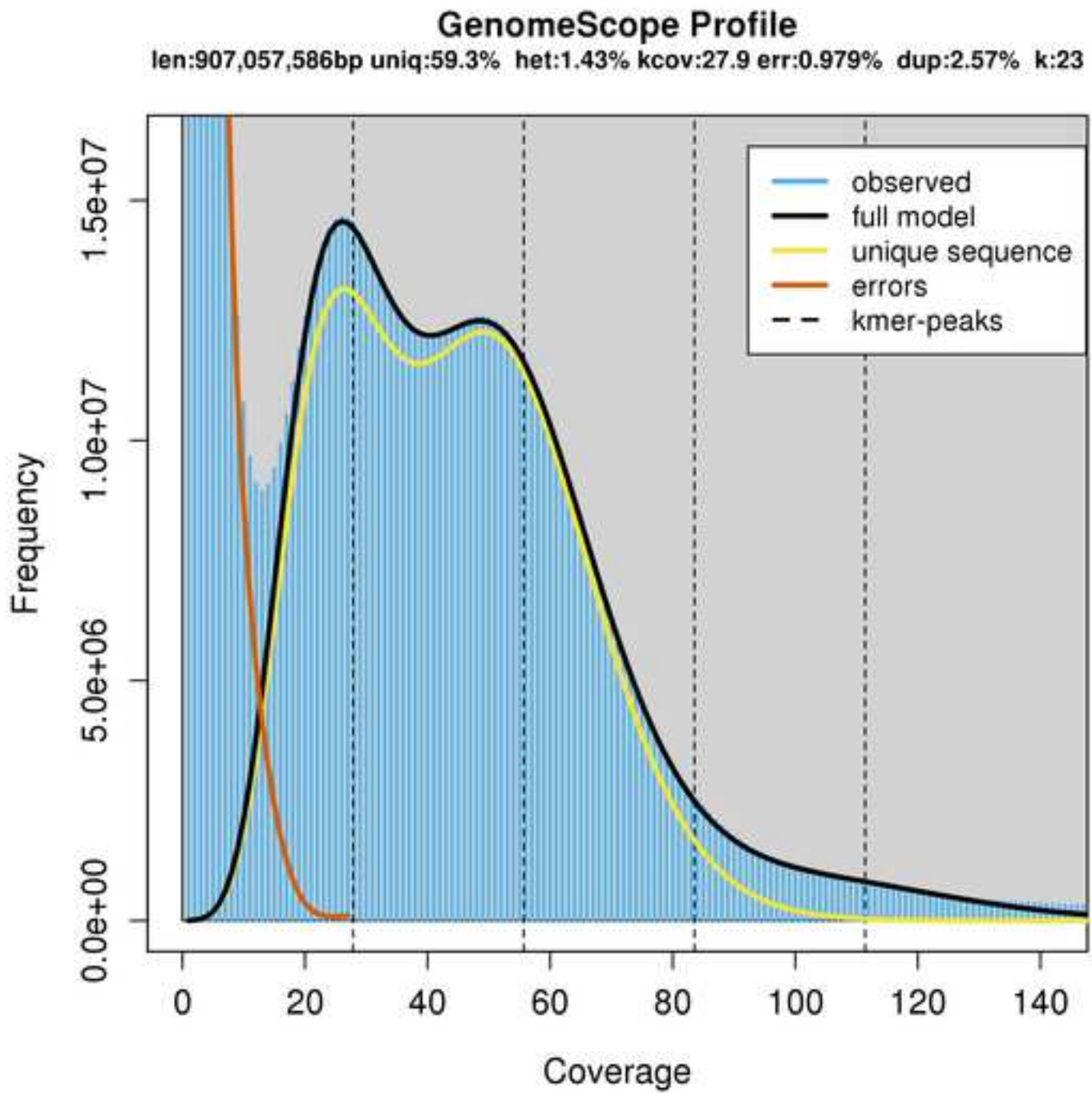
<b>Scaffolds</b>	<b>Spil_haploid1</b>	<b>Spil_haploid2</b>	<b>SP_G</b>	<b>UP_Spi</b>
<b>Largest</b>	6.835 Mb	6.850 Mb	6.843 Mb	0.285 Mb
<b>Number</b>				
>=100Kb	874	872	890	309
>= 10Kb	8 301	8 298	8 760	18 863
>= 1Kb ( <b>total</b> )	117 698	117 698	117 259	44 627
<b>L50 / N50</b>				
>=100Kb	135 / 906.0 Kb	134 / 925.2 Kb	137 / 899.1 Kb	130 / 122.5 Kb
>= 10Kb	242 / 572.7 Kb	242 / 568.2 Kb	254 / 552.2 Kb	4 594 / 32.9 Kb
>= 1Kb ( <b>total</b> )	859 / 102.9 Kb	860 / 102.7 Kb	903 / 96.6 Kb	6 797 / 25.6 Kb
<b>LG50/NG50</b>	<b>935 / 87.7 Kb</b>	<b>939 / 87.1 Kb</b>	<b>905 / 96.6 Kb</b>	<b>15 422 / 12.6 Kb</b>
<b>Assembly size</b>				
>=100Kb	469.371 Mb	468.838 Mb	473.550 Mb	39.274 Mb
>= 10Kb	622.165 Mb	621.688 Mb	636.491 Mb	513.719 Mb
>= 1Kb ( <b>total</b> )	935.548 Mb	935.082 Mb	949.618 Mb	641.169 Mb
<b>GC content</b>	43.9 %	43.9 %	43.9 %	44.5 %
<b>N's per 100 Kb</b>	12 955	12 961	12 834	169

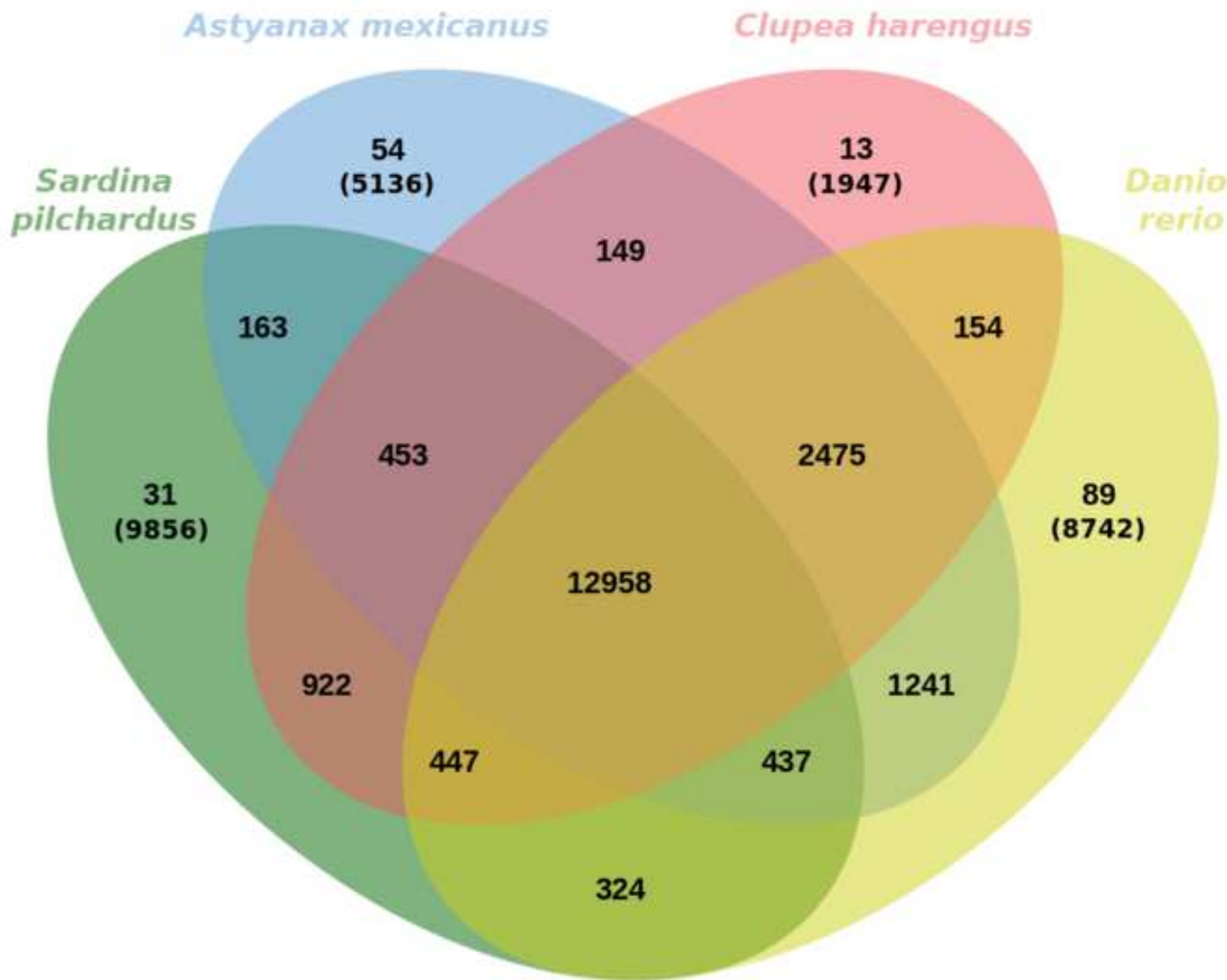


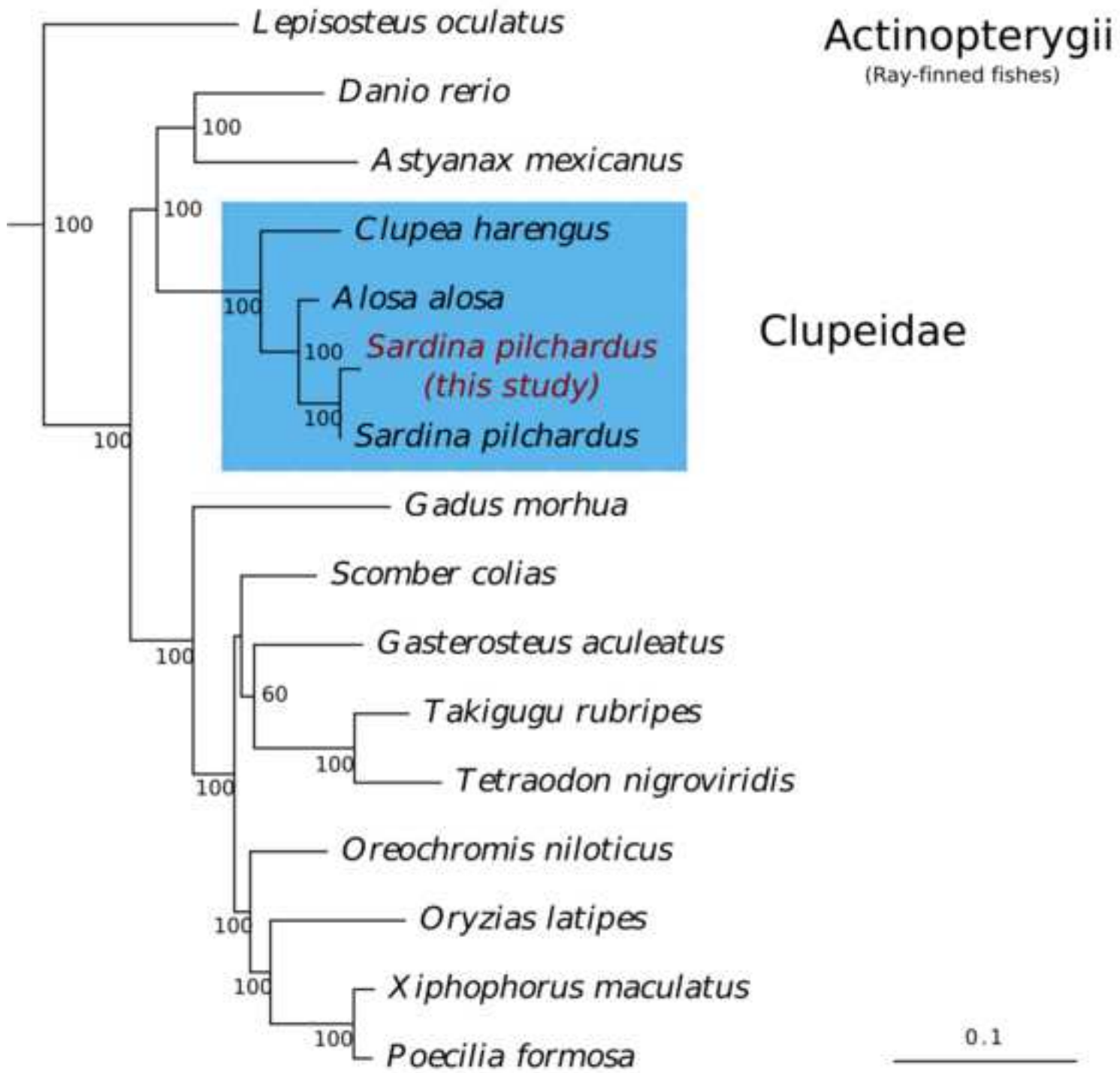
Table 3 – Summary statistics of transcriptome data for the eleven tissues.

Tissue	Paired raw reads	Contigs	CDS deduced	SwissProt annotated	Accession number
Gill/Branchial Arch	29 783 994	62 526	29.3%	38.6%	ERS2629269
Liver	33 479 471	53 104	29.7%	40.1%	ERS2629273
Spleen	25 634 530	66 419	31.6%	40.4%	ERS2629276
Ovary	22 241 327	42 521	38.1%	42.5%	ERS2629270
Midgut	28 016 117	75 782	31.0%	39.5%	ERS2629274
White Muscle	24 409 160	49 266	35.4%	44.8%	ERS2629277
Red Muscle	30 653 774	55 873	30.3%	42.1%	ERS2629275
Kidney	27 861 879	59 495	30.8%	37.3%	ERS2629272
Head Kidney	25 280 960	65 888	32.2%	38.4%	ERS2629271
Brain/Pituitary	24 467 352	75 620	24.5%	37.1%	ERS2629267
Caudal Fin (Skin/Cartilage/Bone)	26 342 097	64 832	23.9%	38.0%	ERS2629268
All Tissues	298 170 661	170 478	15.9%	25.5%	ERS2629362





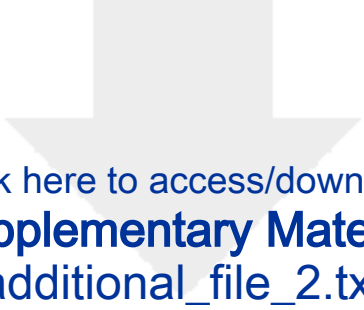




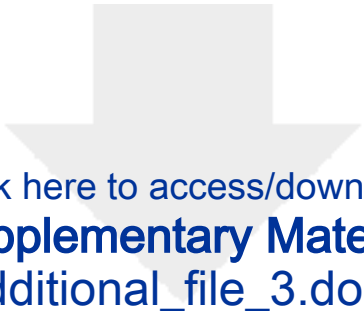


Click here to access/download  
**Supplementary Material**  
Spil\_SNP\_phased\_COV10\_nbc.vcf






Click here to access/download  
**Supplementary Material**  
additional\_file\_2.txt



Click here to access/download  
**Supplementary Material**  
additional\_file\_3.docx





April 05, 2019

Dear Editor,

Please find the revised manuscript "**A haplotype-resolved draft genome of the European sardine (*Sardina pilchardus*)**" by Louro *et al.* for publication in GigaScience as a Data Note article.

We have followed the reviewers' suggestions and made the required changes and corrections which are detailed in separate file.

We take the opportunity to thank the Editor and reviewers for their detailed comments which greatly helped to improve the manuscript.

We hope that the manuscript can now be accepted in GigaScience.

Yours sincerely,

Adelino Canário