

Manuscript Number:	GIGA-D-18-00377R1	
Full Title:	A haplotype-resolved draft genome of the European sardine (<i>Sardina pilchardus</i>)	
Article Type:	Data Note	
Funding Information:	Fundação para a Ciência e a Tecnologia (UID/Multi/04326/2016)	Not applicable
	Programa Operacional Mar2020 (MAR-01.04.02-FEAMP-0024)	Not applicable
	Fundação para a Ciência e a Tecnologia (22153-01/SAICT/2016)	Not applicable
	H2020 Research Infrastructures (654008)	Not applicable
Abstract:	<p>Background: 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>Findings: 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>Conclusions: 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>	
Corresponding Author:	Adelino V. M. Canário PORTUGAL	
Corresponding Author Secondary Information:		
Corresponding Author's Institution:		
Corresponding Author's Secondary Institution:		
First Author:	Bruno Louro, PhD	
First Author Secondary Information:		
Order of Authors:	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
Order of Authors Secondary Information:	
Response to Reviewers:	<p>#####</p> <p>Editor:</p> <p>#1: To put this work in context and offer sufficient validation please provide comparisons in quality and a phylogenomic tree with related sequenced fish species. This can include the other recently sequenced sardine genomes. The referees (and if published, readers) can then judge the quality, utility and context better.</p> <p>Reply 1: A section with the phylogenetic analysis and tree was added. A quality statistic metrics comparison with the other published sardine genome assembly (Machado et al, 2008) was also added (table 2) and discussed. An orthologous proteome analysis was also added.</p> <p>#2: Please include other identifiers and accessions (fishbase and NCBI taxon IDs, ORCIDs, RRIDs, sample and data accessions) for reproducibility purposes, and make sure there is sufficient methodological detail.</p> <p>Reply 2: All missing identifiers/accessions were added. The few software that don't have RRIDs are referenced and version indicated. Assembly data accession references are now clearer.</p> <p>Reviewer reports:</p> <p>Reviewer #1: In this manuscript, the authors report a draft genome for the critically important European Sardine For the most part, the approach is well thought-out and justified, but there are some key concerns involving findings and methodology.</p> <p>#3: For example, the annotation pipeline is unclear. In line 169, it is described as a custom pipeline, but there is no detail regarding how specifically MAKER is used to annotate the genome.</p> <p>Reply 3: "Custom" pipeline wasn't the best choice of words as it gave the idea of an in-house pipeline, which is not the case. Maker is an established annotation software that pipes several other programs (http://www.yandell-lab.org/software/maker.html). We now added a detailed description of the maker workflow and respective software used to annotate the genome (line 198):</p> <p>"The Maker v2.31.10 (MAKER, RRID:SCR_005309) [29] pipeline was used iteratively (five times) to annotate the SP_G consensus genome. The annotations generated in each iteration were kept in the succeeding annotation steps and in the final General Feature Format (GFF) file. During the first Maker run the already described de novo transcriptome was mapped to the genome using blastn v2.7.1 (BLASTN, RRID:SCR_001598) [30] (est2genome parameter in Maker). Moreover, the repetitive elements found with RepeatMasker were used in the Maker pipeline. This initial gene models created by Maker were then used to train Hidden Markov Model (HMM) based gene predictors. The preliminary GFF file generated by this first iteration run was used as input to train SNAP v2006-07-28 [31]. Using the scripts provided directly by Maker (maker2zff) and SNAP (fathom, forge and hmm-assembler.pl) an HMM file was created and used as input for the next Maker iteration (snaphmm option in maker configuration file). For the next iteration, the gene-finding software Augustus v3.3 (Augustus, RRID:SCR_008417) [32] was self-trained running BUSCO with the specific parameter (--long), that turn on the Augustus optimization mode for self-training. The resulted predicted species model from Augustus was included in the pipeline in the third Maker run. For the fourth iteration, GeneMark-ES v4.32 (GeneMark, RRID:SCR_011930) [33], a self-training gene prediction software, was executed and the resulting HMM file was integrated into the Maker pipeline. As further evidence for the annotation, in the last run of maker, the genome was queried using blastx v2.7.1 (BLASTX, RRID:SCR_001653) (protein2genome parameter in Maker), against the deduced proteomes of herring (GCF_000966335.1), (Clupea harengus, NCBI:txid7950, Fishbase ID:24) zebrafish (Danio rerio, NCBI:txid7955, Fishbase ID:4653) (GCF_000002035.6), blind cave fish (Astyanax mexicanus, NCBI:txid7994, Fishbase ID:2740) (GCF_000372685.2), European sardine [34] and all proteins from teleost fishes in the UniProtKB/Swiss-Prot</p>

database (UniProtKB, RRID:SCR_004426) [35]. After the five Maker runs the selected 40 777 genes models are the ab initio predictions supported by the transcriptome and proteome evidence.”

#4: There is no mention of how evidence such as transcriptome evidence from the sardine or other species was used to annotate via MAKER.

Reply 4: The use of sardine transcriptome is now better explained in the maker workflow description (see above reply 3).

Other species were used in the annotation for protein evidence, while for transcriptome evidence we used solely the transcriptome from the sardine. The transcriptome generated was quite extensive and representative of the adult stage of the sardine.

#5: In addition, it is stated in line 183 that 17,199 (65%) proteins received functional annotation. This seem like low efficiency as over a third of potential genes remain unannotated even though there is a wealth of protein sequence data available from fish genomes. This could be due to gene prediction calling a large number of false positives, but it is hard to interperet based on the brief explanation of a custom pipeline.

Reply 5: We have revised the annotation procedure using the Maker annotation pipeline. We did an extra Maker iterative run (Protein2genome) leading to an improved genome annotation with more gene models and with better AED score (median 0.16, data not shown). With this improvement we are now able to functionally annotate 95.3% of the predicted gene coding proteins. Many of the false positive predicted genes were eliminated because now we have the ab initio predictions supported by both protein and transcriptome evidences.

#6: There are other unjustified cutoffs such as the fact that only half of the genome was used for some analyses (line 190).

Reply 6: The reason for the cutoff is that the LongRanger software has a maximum input of 1000 scaffolds as reference genome. We have rewritten the sentence (line 261):

“FASTQ files were processed using the 10x Genomics LongRanger v2.2.2 pipeline [41] with a maximum input limit of one thousand scaffolds, defined as reference genome, and representing about half of the genome size (488.5 Mb).”

#7: The authors mention that the genome contains high heterozygosity, but offer no point of reference or comparison to other species so that it is demonstrated to be high.

Reply 7: A sentence addressing this point was added to the manuscript line 274: “This high SNP heterozygosity (1/206), observed solely in the comparison of the phased alleles, is higher than the average nucleotide diversity of the previously reported marine fish of wild populations: 1/390 in yellow drum [44], 1/309 in herring [45], 1/435 in coelacanth [46], 1/500 in cod [47] and 1/700 in stickleback [48].”

#8: Lastly, there is little discussion about the transcriptomes and how they were used for the genome analyses.

Reply 8: The transcriptome was used for the genome annotation and is now better described and discussed (reply 3).

In addition, the results of the UTR gene prediction based on the transcriptomes is now presented in the following added sentence (line 225): “Based on the transcriptomic evidence, 12 761 gene models were annotated with untranslated regions (UTR) features, more specifically 9 486 gene models with either 5’ or 3’ UTR and 3 275 gene models with both UTR features.”

9: This is also an awesome resource that is established by this study, but it needs more attention in the manuscript. A commented manuscript is attached.

Reply 9: The comments in the manuscript have all been dealt with and corrections made:

PDF-# 9.1: this sentence is a bit unclear. Regarding previous sentence: “A VCF file was obtained containing 2,369,617 filtered SNPs (Additional file 1), in concordance with the estimated mean distance between heterozygous SNPs in the whole genome of 197 bp, by the Supernova input report.”

Reply 9.1: This sentence was rewritten for clarity (line 270): “A VCF file was obtained containing 2 369 617 filtered SNPs (Additional file 1) resulting in a mean distance

between heterozygous phased SNPs of 206 bp. Similar results were obtained in the Supernova input report estimation (Table 1) of mean distance between heterozygous SNPs in the whole genome of 197 bp.”

PDF-# 9.2: This custom pipeline needs elaboration and description. It is a black box. For example, how was protein homology assessed? What did you use as evidence for annotation? sardine transcriptomes? proteins from other fishes?

Reply 9.2: see replies 3 and 4.

PDF-# 9.3: This seems like a low number, and should be addressed in the discussion. What about the other 35%? It seems like the ab initio gene prediction might be calling a lot of false positives. However, it is difficult to see how evidence was used in the pipeline to identify protein coding genes.

Reply 9.3: see replies 3 and 4.

PDF-# 9.4: how were they asses with TransRate?

Reply 9.4: The following was added (line 293): “...with read evidence for assembly optimization, by specifying the contigs fasta file and respective left and right edited reads to be mapped.”

PDF-# 9.5: This should be described in the above paragraph. Regarding the information: “... including 11 tissue-specific assemblies and a mult-tissue assembly.”

Reply 9.5: The sentence in the previous paragraph was edited to: “The same parameters were used for each of the 11 tissue-specific de novo assemblies.”

PDF-# 9.6: why such a low number? Regarding “Of the 170,478 transcripts contigs, 27,078 (16%) were inferred to ORF protein sequences.”

Reply 9.6: The values could be explained based on several reasons: 1) the de novo contigs may represent all types of expression products, such as non-coding RNA, by products of mRNA processing (eg, intron cleavage) or even artefacts of the de novo assembly. From our experience in transcriptome assembly, the bigger the input of RNAseq reads (553 M edited reads) the higher the number of assembled contigs (170 478) that contain non-coding products. The ORF number (27 078) is closer to the expected amount of coding expression products.

PDF-# 9.7: again why such a reduction in confirmed transcripts? Regarding “Query of SwissProt (e-value cutoff of 1e-5) via blastx of total contigs resulted in 43,458 (26%) annotated transcripts.”

Reply 9.7: see reply 9.6.

PDF-# 9.8: how do we know the heterozygosity is high? What are we comparing it to?

Reply 9.8: A sentence was added to address this point (line 275): “This high SNP heterozygosity (1/206), observed solely in the comparison of the phased alleles, is higher than the average nucleotide diversity of the previously reported marine fish of wild populations: 1/390 in yellow drum [44], 1/309 in herring [45], 1/435 in coelacanth [46], 1/500 in cod [47] and 1/700 in stickleback [48].”

10: Figure 1 is also blurry and it is difficult to see the head region of the fish.

Reply 10: We have replaced figure 1.

11: Overall, more detail and justification is needed for methods and results, and the study would benefit by a comparison or the use of available data from other fish genomes. If these changes are implemented, the study would provide an excellent resource for a valuable fishery.

Reply 11: All suggested changes have been made. We thank the reviewer for valuable remarks which we greatly improved the manuscript.

#####

Reviewer #2: The authors of this manuscript report the sequencing of the Europe sardine genome and transcriptome data of selected tissues. Although the obtained resources are novel and valuable, the manuscript does not provide sufficient data to validate their reliability and utility.

#12: The 'Conclusion' part of the Abstract does not provide any conclusion from this study.

Reply 12: The abstract has been modified to address this remark: "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."

13: The epithet of the species name in the title ('Pilchardus') should not be capitalized.

Reply 13: This typo has been corrected.

14: In Abstract: 'Two haploid and a consensus draft genomes were assembled, with a total size of 935 Mbp (N50 103 Kb) and 950Mbp (N50 97 Kb), respectively.' - it is confusing to distinguish which length stats is applied to which genome assembly, in this sentence.

Reply 14: This sentence has now been rephrased: "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)."

15: In the public database NCBI Assembly, I have found two genome assemblies for this species, whose IDs are SP_G and UP_Spi. It is not clear to me which of these corresponds to the Illumina-based or the Chromium-based assembly in the manuscript. The authors need to sort out this problem and present their correspondences in a more clear-cut way.

Reply 15: We submitted our assemblies to the ENA archive project PRJEB27990, with the accession number of the three assemblies GCA_900499035.1 (consensus assembly), UOTT01000000 (haplotype1), and UOTU01000000 (haplotype2). The consensus assembly (SP_G) the reviewer accessed in the NCBI public database was synchronized automatically with ENA. The UP_Spi is a genome draft assembly submitted soon after from another study by other authors (Machado et al, 2018). All ID accessions are now clearly described in the manuscript. At the time of our manuscript submission neither the other genome (UP_Spi) nor the corresponding publication was available. Now we cite and compare the assemblies from the two studies (Table 2).

16: The composition of the two genome assemblies in NCBI Assembly differs particularly in the length of the shortest sequence (200bp vs 1000bp) which can largely affect other length-based metrics, including the N50 scaffold length. I wonder what the authors' policy behind this variable length cut-off was, and also how they describe it in the manuscript. If the authors did not have any coherent policy, they should reconsider this point and revise the manuscript and the genome assemblies in the NCBI database.

Reply 16: Only the consensus assembly of our study is present in NCBI, no filtering was performed on the contigs/scaffolds to inflate the N50. The Supernova default was 1000bp contig minimum size.

17: Also, in the genome assemblies available at NCBI Assembly, I observed a weird distribution of the lengths of 'N' tracts (stretches of undetermined bases) - they are all round numbers for SP_G, while 'N' tracts with the length of 20 is the majority. I wonder whether the authors noticed these, and think that it is worth reasoning possible causes.

Reply 17: We did noticed such behaviour from the Supernova assembler reflected in the pseudohaplotypes output assemblies. The simple explanation is Supernova is able to estimate the gap size based on barcodes spanning the gaps, i.e gaps have linkage evidence through the barcodes linking reads to DNA molecules, and not solely gaps based on reads pairs. Further detailed explanation can be found in Supernova publication (<https://www.biorxiv.org/content/early/2016/08/19/070425>) in particular at "Supplemental Note 5. Supernova gap size estimation."

We now include the "N per 100Kb" in table 2 and discuss the issue starting at line 164.

18: For completeness assessment of the genome assemblies they obtained, the authors used the eukaryote ortholog set as well as the Actinopterygii ortholog set. I wonder why the former was used, instead of the vertebrate or metazoan ortholog set. Also, in describing the numbers of orthologs retrieved by BUSCO, the authors should clearly state which category, namely, complete, fragmented, or missing.

Reply 18: We had used the eukaryote ortholog set as a substitution of the core genes

	<p>CEGMA representation. Following your recommendation, we now present the BUSCO results using the Metazoan ortholog set to represent the coverage of core genes. Following the BUSCO user guide, we also present the results of “actinopterygii” ortholog set as the most related lineage to the sardine.</p> <p>The results now include all the information requested such as complete, single copy, duplicated, fragmented and missing genes, in the following paragraph: “The genome completeness assessment was estimated with Benchmarking Universal Single-copy Orthologs (BUSCO) v3.0.1 (BUSCO, RRID:SCR_015008) [23]. The genome was queried (options -m geno -sp zebrafish) against the “metazoa.oddb9” lineage set containing 978 orthologs from sixty-five eukaryotic organisms to assess the coverage of core eukaryotic genes, and against the “actinopterygii.oddb9” lineage set containing 4584 orthologs from 20 different ray-finned fish species as the most taxon-specific lineage available for the sardine. Using the metazoan odb9 database, 95.4% of the genome had significant matches: 84.5% were complete genes (76.7% single-copy genes and 9.8% duplicates) and 8.9% were fragmented genes. By contrast, using the actinopterygii odb9 database, 84.2% (76.0% complete genes and 8.2% fragmented) had a match, with 69.3% of genes occurring as single copy and 6.7% as duplicates.”</p> <p># 19: Because Figure 2 seems to completely rely on the tool GenomeScope, the authors should cite its source at least in its legend. Reply 19: This reference is now also added in the figure 2 legend.</p> <p>We thank the reviewer for valuable remarks which we greatly improved the manuscript.</p>
Additional Information:	
Question	Response
Are you submitting this manuscript to a special series or article collection?	No
<p>Experimental design and statistics</p> <p>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends.</p> <p>Have you included all the information requested in your manuscript?</p>	Yes
<p>Resources</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 Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible.</p>	Yes

<p>Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?</p>	
<p>Availability of data and materials</p> <p>All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (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 Minimum Standards Reporting Checklist?</p>	<p>Yes</p>

[Click here to view linked References](#)

1 1 A haplotype-resolved draft genome of the European sardine (*Sardina*
2
3
4 2 *pilchardus*)

5
6
7 3 Bruno Louro^{1*}; Gianluca De Moro^{1*}; Carlos Garcia¹; Cymon J. Cox¹; Ana Veríssimo²;
8
9 4 Stephen J. Sabatino²; António M. Santos²; Adelino V. M. Canário^{1&}

10
11 5 1 CCMAR Centre of Marine Sciences, University of Algarve, Campus de Gambelas,
12
13 6 8005-139 Faro, Portugal.

14
15
16 7 2 CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO,
17
18 8 Laboratório Associado, Universidade do Porto, Vairão, Portugal

19
20
21
22 9

23
24 10 * authors contributed equally

25
26 11 & Corresponding author: Adelino V. M. Canário, e-mail: acanario@ualg.pt

27
28
29 12

30
31
32
33 13 **Abstract**

34
35
36
37 14 **Background:** The European sardine (*Sardina pilchardus* Walbaum, 1792) has a high
38
39 15 cultural and economic importance throughout its distribution. Monitoring studies of
40
41 16 sardine populations report an alarming decrease in stocks due to overfishing and
42
43 17 environmental change, which has resulted in historically low captures along the Iberian
44
45 18 Atlantic coast. Consequently, there is an urgent need to better understand the causal
46
47 19 factors of this continuing decrease in the sardine stock. Important biological and
48
49 20 ecological features such as levels of population diversity, structure, and migratory
50
51 21 patterns can be addressed with the development and use of genomics resources.

52
53
54 22 **Findings:** The sardine genome of a single female individual was sequenced using
55
56 23 Illumina HiSeq X Ten 10X Genomics linked-reads generating 113.8 Gb of data. Three
57
58
59
60
61
62
63
64
65

24 draft genomes were assembled: two haploid genomes with a total size of 935 Mbp
25 (N50 103Kb) each, and a consensus genome with a total size of 950 Mbp (N50 97Kb).
26 The genome completeness assessment captured 84% of Actinopterygii
27 Benchmarking Universal Single-Copy Orthologs. To obtain a more complete analysis,
28 the transcriptomes of eleven tissues were sequenced and used to aid the functional
29 annotation of the genome, resulting in 40 777 genes predicted. Variant calling on
30 nearly half of the haplotype genome resulted in the identification of more than 2.3
31 million phased SNPs with heterozygous loci. **Conclusions:** A draft genome was
32 obtained with the 10X Genomics linked-reads technology, despite a high level of
33 sequence repeats and heterozygosity that are expected genome characteristics of a
34 wild sardine. The reference sardine genome and respective variant data are a
35 cornerstone resource of ongoing population genomics studies to be integrated into
36 future sardine stock assessment modelling to better manage this valuable resource.
37 **Keywords:** European sardine; *Sardina*; genome; transcriptome; haplotype; SNP

39 Data description

40 Background

41 The European sardine (*Sardina pilchardus* Walbaum, 1792) (NCBI:txid27697,
42 Fishbase ID:1350) (Figure 1) is a small pelagic fish occurring in temperate boundary
43 currents of the Northeast Atlantic down to Cape Verde off the west coast of Africa, and
44 throughout the Mediterranean to the Black Sea [1]. Two subspecies are generally
45 recognised: *Sardina pilchardus pilchardus* occupies the north-eastern Atlantic and the
46 North Sea whereas *S. pilchardus sardina* occupies the Mediterranean and Black seas,

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47 and the North African coasts south to Cape Verde, with a contact zone near the Strait
48 of Gibraltar [1, 2]. As with other members of the Clupeidae family (e.g. herring, *Clupea*
49 *harengus*, Fishbase ID:24) and allis shad (*Alosa alosa*, NCBI: txid278164, Fishbase
50 ID:101) [3], the sardine experiences strong population fluctuations in abundance,
51 possibly reflecting environmental fluctuations, including climate change [4, 5].

52 The sardine is of major economic and social importance throughout its range with a
53 reported commercial catch for 2016 of 72 183 tonnes in European waters [6]. In
54 Portugal, the sardine is an iconic and culturally revered fish and plays a central role in
55 tourist events, such as summer festivals, throughout the country. However, recent
56 stock assessment data strongly suggests the Iberian sardine fisheries is under threat.
57 A recent report by the International Council for the Exploration of the Sea [6] noted a
58 sharp decrease in the Iberian Atlantic coast sardine stock and advised that catches in
59 2017 should be no more than 23 000 tonnes. The sardine fishery biomass has suffered
60 from declining annual recruitment between 1978 and 2006, and more recently, it has
61 fluctuated around historically low values indicating a high risk of collapse of the Iberian
62 Atlantic stocks [6].

63 A number of sardine populations have been identified by morphometric methods,
64 including as many as five populations in the north-eastern Atlantic (including the
65 Azores), two off the Moroccan coast, and one in Senegalese waters [1, 7]. Each of
66 these recognized sardine populations is subjected to specific climatic and oceanic
67 conditions, mainly during larval development, which directly influence the recruitment
68 of the sardine fisheries [4, 8, 9]. However, because of phenotypic plasticity,
69 morphological traits are strongly influenced by environmental conditions and the
70 underlying genetics that define those populations has proven elusive [10]. While the
71 recognition of subspecies and localised populations might indicate significant genetic

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

72 structure, the large population sizes and extensive migration of sardines are likely to
73 increase gene flow and reduce population differences, suggesting, at its most
74 extensive, a panmictic population with little genetic differentiation within the species'
75 range [11].

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

83 Genome sequencing

84 Sardines were caught during commercial fishing operations in the coastal waters off
85 Olhão, Portugal, and maintained live at the experimental fish culture facilities (EPPO)
86 of the Portuguese Institute for the Sea and Atmosphere (IPMA), Olhão, Portugal [14].
87 A single adult female was anesthetised with 2-phenoxyethanol (1:250 v/v), blood was
88 collected in a heparinized syringe, and the fish euthanized by cervical section. Eleven
89 tissues were dissected out - gill together with branchial arch, liver, spleen, ovary,
90 midgut, white muscle, red muscle, kidney, head kidney, brain together with pituitary,
91 and caudal fin (including skin, scales, bone and cartilage) – into RNA later (Sigma-
92 Aldrich, USA) at room temperature followed by storage at -20°C . Fish maintenance
93 and sample collection were carried out in accordance with the guidelines of the
94 European Union Council (86/609/EU) and Portuguese legislation for the use of
95 laboratory animals from the Veterinary Medicines Directorate (DGAV), the Portuguese

1 96 competent authority for the protection of animals, Ministry of Agriculture, Rural
2 97 Development and Fisheries, Portugal (permit 010238 of 19/04/2016).

3
4 98 Total RNA was extracted using a total RNA purification kit (Maxwell® 16 Total RNA
5 Purification Kit, Promega) and digested twice with DNase (DNA-free kit, Ambion, UK).
6
7 99 The total RNA samples were kept at -80°C until shipment to the RNAseq service
8
9 100 provider Admera Health Co. (USA) which confirmed a RIN above 8 (Qubit TapeStation)
10 101 upon arrival. The mRNA library preparation was performed with NEBNext® Poly(A)
11 102 mRNA Magnetic Isolation Module kit and NEBNext® Ultra™ Directional RNA Library
12 103 Prep kit for sequencing using Illumina HiSeq 4000 paired-end 150 bp cycle to generate
13 104 about 596 million paired-end reads in total.
14
15 105

16 106 The genomic DNA (gDNA) was isolated from 20 µl of fresh blood using the DNeasy
17 107 blood and tissue kit (Qiagen), followed by RNase treatment according to the
18 108 manufacturer's protocol. The integrity of the gDNA was confirmed using pulsed-field
19 109 gel electrophoresis and showed fragment sizes largely above 50 kbp. The gDNA was
20 110 stored at -20 °C before shipping to the service provider (Genome.one, Darlinghurst,
21 111 Australia). Microfluidic partitioned gDNA libraries using the 10x Genomics Chromium
22 112 System were made using 0.6 ng of gDNA input. Sequencing (150bp paired-end cycle)
23 113 was performed in a single lane of the Illumina HiSeq X Ten instrument (Illumina, San
24 114 Diego, CA, USA). Chromium library size range (580-850 bp) was determined with
25 115 LabChip GX Touch (PerkinElmer) and library yield (6.5-40 µM) by quantitative
26 116 polymerase chain reaction.
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51

52 117 Genome size estimation

53 118 A total of 759 million paired-end reads were generated representing 113.8 Gb
54 119 nucleotide sequences with 76.1% bases \geq Q30. Raw reads were edited to trim 10X
55
56
57
58
59
60
61
62
63
64
65

120 Genomics proprietary barcodes with a python script “filter_10xReads.py” [15] prior to
121 kmer counting with Jellyfish v2.2.10 (Jellyfish, RRID:SCR_005491) [16]. Six hundred
122 and seventy million edited reads (90.5 Gb) were used to obtain the frequency
123 distribution of 23-mers. The histogram of the kmer counting distribution was plotted in
124 GenomeScope v1.0.0 (Genoscope, RRID:SCR_002172) [17] (Figure 2) with
125 maximum kmer coverage of 10 000 for estimation of genome size, heterozygosity and
126 repeat content. The estimated sardine haploid genome size was 907 Mbp with a
127 repeat content of 40.7% and a heterozygosity level of 1.43% represented in the first
128 peak of the distribution. These high levels of heterozygosity and repeat content
129 indicated a troublesome genome characteristic for *de novo* assembly.

130 *De novo* genome assembly

131 The *de novo* genome assembly was performed using the paired-end sequence reads
132 from the partitioned library as input for the Supernova assembly algorithm
133 v2.0.0(7fba7b4) (Supernova assembler, RRID:SCR_016756) (10x Genomics, San
134 Francisco, CA, USA) [18]. Two haplotype-resolved genomes, SP_haploid1 (ENA
135 accession ID UOTT01000000) and SP_haploid2 (ENA accession ID
136 UOTU01000000), were assembled with phased scaffolds using the Supernova
137 “mkoutput pseudohap” option. For the assembly process the Supernova run
138 parameters for maximum reads (--maxreads) and barcode fraction (--barfrac) were set
139 for 650M input reads and 80% of barcodes, respectively. Preliminary trials defined an
140 optimal raw coverage of 78-fold, above the 56-fold suggested in the Supernova
141 protocol; this reduced the problem (to some extent) of the complexity of the high repeat
142 content (Table 1). A fraction of the 607.36 million read pairs were used after a quality
143 control step embedded in the Supernova pipeline to remove reads that were not

144 barcoded, not properly paired, or low-quality. Input reads had a 138.5 bp mean length
145 after proprietary 10X barcode trimming and a N50 of 612 per barcode/DNA molecule
146 (Table 1).

147 Further scaffolding and gap closure procedures were performed with Rails
148 v1.2/Cobbler v0.3 pipeline script [19] to obtain the final consensus genome sequence
149 named SP_G (ENA accession ID GCA_900499035.1) using the parameters anchoring
150 sequence length (*-d* 100) and minimum sequence identity (*-i* 0.95). Three scaffolding
151 and gap closure procedures were performed iteratively with one haplotype of the initial
152 assembly as the assembly *per se*, and previous *de novo* assemblies from Supernova
153 v1.2.2, (315M/100% and 450M/80% reads/barcodes). By closing several gaps within
154 scaffolds and merging other scaffolds into longer and fewer scaffolds (117 259), this
155 procedure resulted into a slightly longer genome size of 949.62 Mb, which slightly
156 deflated the scaffold N50 length to 96.6 Kb (Table 2). The assembly metrics of the
157 three assemblies are described in Table 2 together with a recently published Illumina
158 paired-end assembled sardine genome (UP_Spi) [20]. The total assembly size of our
159 genome (SP_G) is 950 Mb and the UP_Spi is 641 Mb (Table 2). Because the SP_G
160 and UP_Spi assembly sizes are of different orders of magnitude, in addition to N50
161 we present NG50 values [21] for an estimated genome size of 950 Mb (Table 2). In
162 the SP_G assembly, 905 scaffolds (LG50) represents half of the estimated genome
163 with an NG50 value of 96.6 Kb, in comparison to LG50 of 15 422 and NG50 of 12.6
164 Kb in the UP_Spi assembly. The ungapped length of the SP_G assembly is 828 Mb.
165 The larger gaps of the SP_G assembly compared to the UP_Spi can be explained by
166 the Supernova being able to estimate gap size based on the bar codes spanning the
167 gaps, i.e. gaps have linkage evidence through the barcodes linking reads to DNA
168 molecules and not solely gaps based on reads pairs [22]. Such gaps are reflected in

169 the large number of N's per 100kb in our assemblies (Table 2). The number of
1 scaffolds in SP_G is 117 259 (largest 6.843 Mb) and in UP_Spi is 44 627 (largest
2 scaffolds in SP_G is 117 259 (largest 6.843 Mb) and in UP_Spi is 44 627 (largest
3 0.285 Mb).
4 171 0.285 Mb).

7 172 The genome completeness assessment was estimated with Benchmarking Universal
8 Single-copy Orthologs (BUSCO) v3.0.1 (BUSCO, RRID:SCR_015008) [23]. The
9 173 genome was queried (options -m geno -sp zebrafish) against the "metazoa.odbg"
10 174 lineage set containing 978 orthologs from sixty-five eukaryotic organisms to assess
11 175 the coverage of core eukaryotic genes, and against the "actinopterygii.odbg"
12 176 set containing 4584 orthologs from 20 different ray-finned fish species as the most
13 177 taxon-specific lineage available for the sardine. Using the metazoan odb9 database,
14 178 95.4% of the genome had significant matches: 84.5% were complete genes (76.7%
15 179 single-copy genes and 9.8% duplicates) and 8.9% were fragmented genes. By
16 180 contrast, using the actinopterygii odb9 database, 84.2% (76.0% complete genes and
17 181 8.2% fragmented) had a match, with 69.3% of genes occurring as single copy and
18 182 6.7% as duplicates.
19 183

20 184 The EMBRIC configurator service [24] was used to create a fish specific checklist
21 185 (named finfish) for the submission of the sardine genome project to the European
22 186 Nucleotide Archive (ENA) (European Nucleotide Archive, RRID:SCR_006515)
23 187 (project accession PRJEB27990).
24 188

25 Repeat Content

26 189 The SP_G consensus assembly was used as a reference genome to build a *de novo*
27 190 repeat library running RepeatModeler v1.0.11 (RepeatModeler, RRID:SCR_015027)
28 191 [25] with default parameters. The model obtained from RepeatModeler was used,
29 192 together with Dfam_consensus database v20171107 [26] and RepBase
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

193 RepeatMasker Edition library v20170127 [27] to identify repetitive elements and low
194 complexity sequences running RepeatMasker v4.0.7 (RepeatMasker,
195 RRID:SCR_012954) [28]. The analysis carried out revealed that 23.33% of the
196 assembled genome consists of repetitive elements.

197 Genome annotation

198 The Maker v2.31.10 (MAKER, RRID:SCR_005309) [29] pipeline was used iteratively
199 (five times) to annotate the SP_G consensus genome. The annotations generated in
200 each iteration were kept in the succeeding annotation steps and in the final General
201 Feature Format (GFF) file. During the first Maker run the *de novo* transcriptome was
202 mapped to the genome using blastn v2.7.1 (BLASTN, RRID:SCR_001598) [30]
203 (est2genome parameter in Maker). Moreover, the repetitive elements found with
204 RepeatMasker were used in the Maker pipeline. This initial gene models created by
205 Maker were then used to train Hidden Markov Model (HMM) based gene predictors.
206 The preliminary GFF file generated by this first iteration run was used as input to train
207 SNAP v2006-07-28 [31]. Using the scripts provided directly by Maker (maker2zff) and
208 SNAP (fathom, forge and hmm-assembler.pl) an HMM file was created and used as
209 input for the next Maker iteration (snaphmm option in maker configuration file). For the
210 next iteration, the gene-finding software Augustus v3.3 (Augustus,
211 RRID:SCR_008417) [32] was self-trained running BUSCO with the specific parameter
212 (--long), that turn on the Augustus optimization mode for self-training. The resulted
213 predicted species model from Augustus was included in the pipeline in the third Maker
214 run. For the fourth iteration, GeneMark-ES v4.32 (GeneMark, RRID:SCR_011930)
215 [33], a self-training gene prediction software, was executed and the resulting HMM file
216 was integrated into the Maker pipeline. As further evidence for the annotation, in the

217 last run of Maker, the genome was queried using blastx v2.7.1 (BLASTX,
1
2 218 RRID:SCR_001653) (protein2genome parameter in Maker), against the deduced
3
4
5 219 proteomes of herring (GCF_000966335.1), (*Clupea harengus*, NCBI:txid7950,
6
7 220 Fishbase ID:24) zebrafish (*Danio rerio*, NCBI:txid7955, Fishbase ID:4653)
8
9
10 221 (GCF_000002035.6), blind cave fish (*Astyanax mexicanus*, NCBI:txid7994, Fishbase
11
12 222 ID:2740) (GCF_000372685.2), European sardine [20] and all proteins from teleost
13
14 223 fishes in the UniProtKB/Swiss-Prot database (UniProtKB, RRID:SCR_004426) [34].
15
16
17 224 After the five Maker runs the selected 40 777 genes models are the *ab initio* predictions
18
19 225 supported by the transcriptome and proteome evidence. Based on the transcriptomic
20
21
22 226 evidence, 12 761 gene models were annotated with untranslated regions (UTR)
23
24 227 features, more specifically 9 486 gene models with either 5' or 3' UTR and 3 275 gene
25
26
27 228 models with both UTR features.
28
29 229 InterProScan v. 5.30 (InterProScan, RRID:SCR_005829) [35] and NCBI blastp v2.8.1
30
31
32 230 (BLASTP, RRID:SCR_001010) [30] were used to functionally annotate the 40 777
33
34 231 predicted protein coding genes. Thirty-three thousand five hundred and fifty-three (33
35
36 232 553) (82.3%) proteins were successfully annotated using blastp (e-value 1e-05)
37
38
39 233 against the UniProtKB/Swiss-Prot database and another 5 228 were annotated using
40
41 234 the NCBI non-redundant protein database (nr). In addition to the above, 37 075
42
43
44 235 (90.9%) proteins were successfully annotated using InterProScan with all the InterPro
45
46 236 v72.0 (InterPro, RRID:SCR_006695) [36] databases: CATH-Gene3D (Gene3D,
47
48
49 237 RRID:SCR_007672), Hamap (HAMAP, RRID:SCR_007701), PANTHER (PANTHER,
50
51 238 RRID:SCR_004869), Pfam (Pfam, RRID:SCR_004726), PIRSF (PIRSF,
52
53
54 239 RRID:SCR_003352), PRINTS (PRINTS, RRID:SCR_003412), ProDom (ProDom,
55
56 240 RRID:SCR_006969), ProSite Patterns (PROSITE, RRID:SCR_003457), ProSite
57
58 241 Profiles, SFLD (Structure-function linkage database, RRID:SCR_001375), SMART
59
60
61
62
63
64
65

242 (SMART, RRID:SCR_005026), SUPERFAMILY (SUPERFAMILY,
1
2 243 RRID:SCR_007952), and TIGRFAM (JCVI TIGRFAMS, RRID:SCR_005493). In total,
3
4
5 244 38 880 (95.3%) of the predicted proteins received a functional annotation. The
6
7 245 annotated genome assembly is published [37] in the wiki-style annotation portal
8
9
10 246 ORCAE [38] .
11
12 247 OrthoFinder v2.2.7 [39] was used to identify paralogy and orthology in our Swiss-prot
13
14 248 annotated deduced proteome and in the deduced proteomes from herring, blind cave
15
16
17 249 fish and zebrafish. The resulting orthogroups were plotted using jvenn (jVenn,
18
19 250 RRID:SCR_016343) [40] (Figure 3), where paralagous (two or more genes) and
20
21
22 251 singletons were identified within species specific orthogroups. The deduced
23
24 252 sardine proteome has 3 413 paralogous groups containing 11 406 genes, of which 31
25
26
27 253 are sardine specific orthogroups. The amount of sardine singletons (9 856) can be
28
29 254 partially due to fragmented predicted genes, but can reflect also some evolutionary
30
31
32 255 divergence which requires further study to understand the biological relevance. In
33
34 256 total, 25 560 orthogroups containing at least a single protein were identified in sardine,
35
36
37 257 of which 12 958 ortholgroups are common to all four fish species. Within the
38
39 258 Clupeidae, the sardine and the herring share 14 780 orthogroups with 922 family-
40
41 259 specific orthogroups.
42
43
44

45 260 Variant calling between phased alleles

46
47
48 261 FASTQ files were processed using the 10x Genomics LongRanger v2.2.2 pipeline
49
50
51 262 [41] with a maximum input limit of one thousand scaffolds, defined as reference
52
53 263 genome, and representing about half of the genome size (488.5 Mb). The
54
55
56 264 LongRanger pipeline was run with default settings, with the exception of vcmode to
57
58 265 define the Genome Analysis Toolkit (GATK) v4.0.3.0 (GATK, RRID:SCR_001876)
59
60
61
62
63
64
65

266 [42] as the variant caller and the somatic parameters. The longest phase block was
1
2 267 2.86 Mb and the N50 phase block was 0.476 Mb.
3
4
5 268 Single nucleotide polymorphisms (SNP's) were furthered filtered to obtain only
6
7 269 phased and heterozygous SNP's between the two alleles with a coverage higher than
8
9
10 270 10-fold using VCFtools v0.1.16 (VCFtools, RRID:SCR_001235). A VCF file was
11
12 271 obtained containing 2 369 617 filtered SNPs (Additional file 1) resulting in a mean
13
14 272 distance between heterozygous phased SNPs of 206 bp. Similar results were obtained
15
16
17 273 in the Supernova input report estimation (Table 1) of mean distance between
18
19 274 heterozygous SNPs in the whole genome of 197 bp. This high SNP heterozygosity
20
21
22 275 (1/206), observed solely in the comparison of the phased alleles, is higher than the
23
24 276 average nucleotide diversity of the previously reported marine fish of wild populations:
25
26
27 277 1/390 in yellow drum [43], 1/309 in herring [44], 1/435 in coelacanth [45], 1/500 in cod
28
29 278 [46] and 1/700 in stickleback [47].
30

33 279 *De novo* transcriptome assembly

34
35
36 280 The 596 million paired-end raw transcriptomic reads were edited for contamination
37
38
39 281 (e.g. adapters) using TrimGalore v0.4.5 wrapper tool (TrimGalore,
40
41 282 RRID:SCR_016946) [15], low-quality base trimming with Cutadapt v1.15 (cutadapt,
42
43 283 RRID:SCR_011841) [48] and the output overall quality reports of the edited reads with
44
45
46 284 FastQC v0.11.5 (FastQC, RRID:SCR_014583) [49].
47

48 285 The 553 million edited paired-end reads were *de novo* assembled as a multi-tissue
49
50
51 286 assembly using Trinity v2.5.1 (Trinity, RRID:SCR_013048) [50] with a minimum contig
52
53 287 length of 200 bp, 50x coverage read depth normalization, and RF strand-specific read
54
55
56 288 orientation. The same parameters were used for each of the 11 tissue specific *de novo*
57
58
59
60
61
62
63
64
65

289 assemblies. The genome and transcriptome assemblies were conducted on the
1
2 290 Portuguese National Distributed Computing Infrastructure [49].
3
4
5 291 The twelve *de novo* transcriptome assemblies (Table 3) were each quality assessed
6
7 292 using TransRate v1.0.3 [51] with read evidence for assembly optimization, by
8
9 293 specifying the contigs fasta file and respective left and right edited reads to be mapped.
10
11
12 294 The multi-tissue assembly with all reads resulted in an assembled transcriptome of
13
14 295 170 478 transcript contigs following the TransRate step. Functional annotation was
15
16 296 performed using the Trinotate v3.1.1 pipeline [24] and integrated into a SQLite
17
18 297 database. All annotations were based on the best deduced open reading frame (ORF)
19
20 298 obtained with the Transdecoder v1.03 [51]. Of the 170 478 transcripts contigs, 27 078
21
22 299 (16%) were inferred to ORF protein sequences. Query of the UniProtKB/Swiss-Prot
23
24 300 (e-value cutoff of 1e-5) database via blastx v2.7.1 of total contigs resulted in 43 458
25
26 301 (26%) annotated transcripts. The ORFs were queried against UniProtKB/Swiss-Prot
27
28 302 (e-value cutoff of 1e-5) via blastp v2.7.1 and PFAM using hmmscan (HMMER v3.1b2)
29
30 303 (Hmmer, RRID:SCR_005305) [52] resulting in 19 705 (73% of ORF) and 16 538 (61%
31
32 304 of ORF) UniProtKB/Swiss-Prot and PFAM annotated contigs respectively. The full
33
34 305 annotation report with further functional annotation, such as signal peptides,
35
36 306 transmembrane regions, eggnoG, Kyoto Encyclopedia of Genes and Genomes
37
38 307 (KEGG) (KEGG, RRID:SCR_012773), and Gene Ontology annotation (Gene
39
40 308 Ontology, RRID:SCR_002811) are listed in tabular format in Additional file 2.
41
42
43
44
45
46
47
48
49

50 309 **Ray-finned fish phylogeny**

51
52
53 310 We conducted a phylogenetic analysis of ray-finned fish (Actinopterygii) taxa based
54
55 311 on 97 genes obtained from the newly constructed proteome.
56
57
58
59
60
61
62
63
64
65

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

312 Sequence alignments for 106 proteins from 17 fish species were obtained from [20].
313 Gene models for each protein were constructed with hmmbuild (HMMER v3.1b2) [53]
314 using default options and orthologous genes from the new proteome searched for
315 using hmmsearch (HMMER) with an e-value cutoff of 10e-3. Best protein hits from
316 the new genome according to the bitscores were aligned to the original protein
317 sequence alignments using hmalign (HMMER) with default options. Gapped and
318 poorly aligned sites were identified by Gblocks v0.91b (Gblocks, RRID:SCR_015945)
319 [54] using default options and removed using p4 v1.3.0 [55]. Protein alignment
320 statistics were calculated, and the proteins concatenated into a single alignment using
321 novel scripts in p4. Of the 106 fish proteins alignments, 97 contained sites which were
322 considered correctly aligned by the Gblocks analysis; statistics for these alignments
323 are presented in Table S1 (Additional file 3). The concatenated sequence alignment
324 of the 97 proteins contained 14 515 sites without gaps of which 7 391 were constant,
325 7 123 variable, and 3 879 parsimony informative.

326 The best-fitting empirical protein model of the concatenated data was evaluated using
327 ModelFinder [56] in IQ-TREE v1.6.7.1 [57]. The best-fitting empirical substitution
328 model was estimated to be the JTT model [58] with a discrete gamma-distribution of
329 among-site rate variation (4 categories) and empirical composition frequencies (typical
330 notation: JTT+ Γ_4 +F).

331 Optimal maximum likelihood tree searches (100 replicates) and bootstrap analyses
332 (300 replicates) were conducted using RAxML v8.2.12 (RAxML, RRID:SCR_006086)
333 [59] with the best-fitting model. The optimal maximum likelihood tree (-ln likelihood:
334 146565.6438) is presented in Figure 4 with bootstrap support values given at nodes,
335 and is rooted to the outgroups *Petromyzon marinus* (lamprey) and *Latimeria*
336 *chalumnae* (coelacanth).

337

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

338 **Conclusion**

339 Despite the sardine genome having a high level of repeats and heterozygosity, factors
340 which pose a challenge to *de novo* genome assembly, a more than adequate draft
341 genome was obtained with the 10X Genomics linked-reads (Chromium) technology.
342 The Chromium technology's ability to tag and cluster the reads to individual DNA
343 molecules has proven advantages for scaffolding, just as long reads technologies such
344 as Nanopore and Pacific Biosciences, but with high coverage and low error rates. The
345 advantage of linked-reads for *de novo* genomic assemblies is evident in comparison
346 to typical short read data, especially in the case of wild species with highly
347 heterozygous genomes, where the latter often result in many uncaptured genomic
348 regions and with a lower scaffolding yield due to repeated content.

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

355 The genomic and transcriptomic resources reported here are important tools for future
356 studies to understand sardine response at the levels of physiology, population
357 genetics and ecology of the causal factors responsible for the recruitment and collapse
358 of the sardine stock in Iberian Atlantic coast. Besides the commercial interest, the
359 sardine plays a crucial role at a key trophic level by bridging energy from the primary
360 producers to the top predators in the marine ecosystem. Therefore, disruption of the

361 sardine population equilibrium is likely to reverberate throughout the food chain via a
1
2 362 trophic cascade. Consequently, these genomic and genetic resources are the
3
4
5 363 prerequisites needed to develop tools to monitor the population status of the sardine
6
7 364 and thereby provide an important bio-monitoring system for the health of the marine
8
9
10 365 environment.

14 366 **Availability of the supporting data**

17 367 Raw data, assembled transcriptomes, and assembled genomes are available at the
18
19
20 368 European Bioinformatics Institute ENA archive with the project accession
21
22 369 PRJEB27990. The annotated genome assembly is published in the wiki-style
23
24
25 370 annotation portal ORCAE [37].

29 371 **Acknowledgements**

32 372 This research was supported by national funds from FCT - Foundation for Science
33
34
35 373 and Technology through project UID/Multi/04326/2016 and by FCT and FEDER under
36
37 374 projects 22153-01/SAICT/2016 (to INCD), ALG-01-0145-FEDER-022121 and ALG-
38
39
40 375 01-0145-FEDER-022231; and co-funds from MAR2020 operational programme of the
41
42 376 European Maritime and Fisheries Fund (project SARDINOMICS MAR-01.04.02-
43
44
45 377 FEAMP-0024). The EMBRIC configurator service received funding from the European
46
47 378 Union's Horizon 2020 research and innovation programme under grant agreement No
48
49
50 379 654008. The authors acknowledge Pedro Guerreiro for providing the sardine samples.

52 380

381 **References**

- 1
2
3 382 1. Parrish RH, Serra R and Grant WS. The monotypic sardines, *Sardina* and
4
5 383 *Sardinops* - Their taxonomy, distribution, stock structure, and zoogeography.
6
7
8 384 Can J Fish Aquat Sci. 1989;46 11:2019-36. doi:10.1139/f89-251.
- 9
10 385 2. Silva A. Morphometric variation among sardine (*Sardina pilchardus*)
11
12
13 386 populations from the northeastern Atlantic and the western Mediterranean.
14
15 387 ICES J Mar Sci. 2003;60 6:1352-60. doi:10.1016/S1054-3139(03)00141-3.
- 16
17
18 388 3. Lavoue S, Miya M, Saitoh K, Ishiguro NB and Nishida M. Phylogenetic
19
20 389 relationships among anchovies, sardines, herrings and their relatives
21
22 390 (Clupeiformes), inferred from whole mitogenome sequences. Mol Phylogenet
23
24 391 Evol. 2007;43 3:1096-105. doi:10.1016/j.ympev.2006.09.018.
- 25
26
27 392 4. Santos AMP, Borges MDF and Groom S. Sardine and horse mackerel
28
29
30 393 recruitment and upwelling off Portugal. ICES J Mar Sci. 2001;58 3:589-96.
31
32 394 doi:10.1006/jmsc.2001.1060.
- 33
34
35 395 5. Checkley Jr. DM, Asch RG and Rykaczewski RR. Climate, anchovy, and
36
37 396 sardine. Annual Review of Marine Science. 2017;9 1:469-93.
38
39
40 397 doi:10.1146/annurev-marine-122414-033819.
- 41
42 398 6. ICES. *Report of the Working Group on Southern Horse Mackerel, Anchovy*
43
44 399 *and Sardine (WGHANSA), 24–29 June 2017, Bilbao, Spain. CM*
45
46 400 *2017/ACOM:17, 640 p.* 2017.
- 47
48
49 401 7. Atarhouch T, Ruber L, Gonzalez EG, Albert EM, Rami M, Dakkak A, et al.
50
51
52 402 Signature of an early genetic bottleneck in a population of Moroccan sardines
53
54 403 (*Sardina pilchardus*). Mol Phylogenet Evol. 2006;39 2:373-83.
55
56 404 doi:10.1016/j.ympev.2005.08.003.
- 57
58
59
60
61
62
63
64
65

- 405 8. Santos MB, Gonzalez-Quiros R, Riveiro I, Cabanas JM, Porteiro C and Pierce
1
2 406 GJ. Cycles, trends, and residual variation in the Iberian sardine (*Sardina*
3
4
5 407 *pilchardus*) recruitment series and their relationship with the environment.
6
7 408 ICES J Mar Sci. 2012;69 5:739-50. doi:10.1093/icesjms/fsr186.
8
9
10 409 9. Leitao F, Alms V and Erzini K. A multi-model approach to evaluate the role of
11
12 410 environmental variability and fishing pressure in sardine fisheries. J Mar Syst.
13
14 411 2014;139:128-38. doi:10.1016/j.jmarsys.2014.05.013.
15
16
17 412 10. Tinti F, Di Nunno C, Guarniero I, Talenti M, Tommasini S, Fabbri E, et al.
18
19 413 Mitochondrial DNA sequence variation suggests the lack of genetic
20
21 414 heterogeneity in the Adriatic and Ionian stocks of *Sardina pilchardus*. Mar
22
23
24 415 Biotechnol (NY). 2002;4 2:163-72. doi:10.1007/s10126-002-0003-3.
25
26
27 416 11. Jemaa S, Bacha M, Khalaf G, Dessailly D, Rabhi K and Amara R. What can
28
29 417 otolith shape analysis tell us about population structure of the European
30
31 418 sardine, *Sardina pilchardus*, from Atlantic and Mediterranean waters? J Sea
32
33
34 419 Res. 2015;96:11-7. doi:10.1016/j.seares.2014.11.002.
35
36
37 420 12. Boehm JT, Waldman J, Robinson JD and Hickerson MJ. Population genomics
38
39 421 reveals seahorses (*Hippocampus erectus*) of the western mid-Atlantic coast to
40
41 422 be residents rather than vagrants. PLoS One. 2015;10 1:e0116219.
42
43
44 423 doi:10.1371/journal.pone.0116219.
45
46 424 13. Hendricks S, Anderson EC, Antao T, Bernatchez L, Forester BR, Garner B, et
47
48
49 425 al. Recent advances in conservation and population genomics data analysis.
50
51 426 Evol Appl. 2018;11 8:1197-211. doi:10.1111/eva.12659.
52
53
54 427 14. Marcalo A, Guerreiro PM, Bentes L, Rangel M, Monteiro P, Oliveira F, et al.
55
56 428 Effects of different slipping methods on the mortality of sardine, *Sardina*
57
58 429 *pilchardus*, after purse-seine capture off the Portuguese Southern coast
59
60
61
62
63
64
65

430 (Algarve). PLoS One. 2018;13 5:e0195433.
1
2 431 doi:10.1371/journal.pone.0195433.
3
4
5 432 15. Krueger F: "Trim galore" A wrapper tool around Cutadapt and FastQC to
6
7 433 consistently apply quality and adapter trimming to FastQ files.
8
9
10 434 https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (2015).
11
12 435 Accessed 9/24/2018.
13
14 436 16. Marcais G and Kingsford C. A fast, lock-free approach for efficient parallel
15
16 437 counting of occurrences of k-mers. Bioinformatics. 2011;27 6:764-70.
18
19 438 doi:10.1093/bioinformatics/btr011.
20
21
22 439 17. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski
23
24 440 J, et al. GenomeScope: fast reference-free genome profiling from short reads.
25
26 441 Bioinformatics. 2017;33 14:2202-4. doi:10.1093/bioinformatics/btx153.
27
28
29 442 18. Weisenfeld NI, Kumar V, Shah P, Church DM and Jaffe DB. Direct
30
31 443 determination of diploid genome sequences. Genome Res. 2017;27 5:757-67.
32
33 444 doi:10.1101/gr.214874.116.
34
35
36 445 19. Warren RL. RAILS and Cobbler: Scaffolding and automated finishing of draft
37
38 446 genomes using long DNA sequences. JOSS. 2016;1 7:116.
39
40 447 doi:10.21105/joss.00116.
41
42
43 448 20. Machado A, Tørresen O, Kabeya N, Couto A, Petersen B, Felício M, et al.
44
45 449 "Out of the Can": A draft genome assembly, liver transcriptome, and
46
47 450 nutrigenomics of the European sardine, *Sardina pilchardus*. Genes. 2018;9
48
49 451 10:485. doi:10.3390/genes9100485.
50
51
52
53 452 21. Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, et al. Assemblathon 1:
54
55 453 a competitive assessment of de novo short read assembly methods. Genome
56
57 454 Res. 2011;21 12:2224-41. doi:10.1101/gr.126599.111.
58
59
60
61
62
63
64
65

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
- 455 22. Weisenfeld NI, Kumar V, Shah P, Church DM and Jaffe DB. Direct
456 determination of diploid genome sequences. bioRxiv. 2016:070425.
457 doi:10.1101/070425.
- 458 23. Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov
459 G, et al. BUSCO applications from quality assessments to gene prediction
460 and phylogenomics. Mol Biol Evol. 2017;35 3:543-8.
461 doi:10.1093/molbev/msx319.
- 462 24. Bryant DM, Johnson K, DiTommaso T, Tickle T, Couger MB, Payzin-Dogru D,
463 et al. A tissue-mapped axolotl *de novo* transcriptome enables identification of
464 limb regeneration factors. Cell Rep. 2017;18 3:762-76.
465 doi:10.1016/j.celrep.2016.12.063.
- 466 25. Smit A and Hubley R: RepeatModeler Open-1.0. <http://www.repeatmasker.org>
467 (2008). Accessed 9/24/2018.
- 468 26. Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, et al. The Dfam
469 database of repetitive DNA families. Nucleic Acids Res. 2016;44 D1:D81-9.
470 doi:10.1093/nar/gkv1272.
- 471 27. Bao W, Kojima KK and Kohany O. Repbase Update, a database of repetitive
472 elements in eukaryotic genomes. Mob DNA. 2015;6 1:11.
473 doi:10.1186/s13100-015-0041-9.
- 474 28. Smit A, Hubley R and Green P: 2013–2015. RepeatMasker Open-4.0.
475 <http://www.repeatmasker.org> (2013).
- 476 29. Holt C and Yandell M. MAKER2: an annotation pipeline and genome-
477 database management tool for second-generation genome projects. BMC
478 Bioinformatics. 2011;12 1:491. doi:10.1186/1471-2105-12-491.

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
- 479 30. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al.
480 BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421-.
481 doi:10.1186/1471-2105-10-421.
- 482 31. Korf I. Gene finding in novel genomes. *BMC Bioinformatics*. 2004;5 1:59.
483 doi:10.1186/1471-2105-5-59.
- 484 32. Stanke M and Waack S. Gene prediction with a hidden Markov model and a
485 new intron submodel. *Bioinformatics*. 2003;19 suppl_2:ii215-ii25.
486 doi:10.1093/bioinformatics/btg1080.
- 487 33. Lomsadze A, Burns PD and Borodovsky M. Integration of mapped RNA-Seq
488 reads into automatic training of eukaryotic gene finding algorithm. *Nucleic
489 Acids Res*. 2014;42 15:e119. doi:10.1093/nar/gku557.
- 490 34. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, et al.
491 UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*. 2004;32
492 Database issue:D115-9. doi:10.1093/nar/gkh131.
- 493 35. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan
494 5: genome-scale protein function classification. *Bioinformatics*. 2014;30
495 9:1236-40. doi:10.1093/bioinformatics/btu031.
- 496 36. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, et al.
497 InterPro in 2017—beyond protein family and domain annotations. *Nucleic
498 Acids Res*. 2016;45 D1:D190. doi:10.1093/nar/gkw1107.
- 499 37. Sardine Genome Annotation Portal.
500 <https://bioinformatics.psb.ugent.be/orcae/overview/Spil>. Accessed 9/24/2018.
- 501 38. Sterck L, Billiau K, Abeel T, Rouze P and Van de Peer Y. ORCAE: online
502 resource for community annotation of eukaryotes. *Nat Methods*. 2012;9
503 11:1041. doi:10.1038/nmeth.2242.

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
- 504 39. Emms DM and Kelly S. OrthoFinder: solving fundamental biases in whole
505 genome comparisons dramatically improves orthogroup inference accuracy.
506 Genome Biol. 2015;16 1:157. doi:10.1186/s13059-015-0721-2.
- 507 40. Bardou P, Mariette J, Escudie F, Djemiel C and Klopp C. jvenn: an interactive
508 Venn diagram viewer. BMC Bioinformatics. 2014;15 1:293. doi:10.1186/1471-
509 2105-15-293.
- 510 41. Zheng GX, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, et al.
511 Haplotyping germline and cancer genomes with high-throughput linked-read
512 sequencing. Nat Biotechnol. 2016;34 3:303-11. doi:10.1038/nbt.3432.
- 513 42. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et
514 al. The Genome Analysis Toolkit: A MapReduce framework for analyzing
515 next-generation DNA sequencing data. Genome Res. 2010;20 9:1297-303.
516 doi:10.1101/gr.107524.110.
- 517 43. Han Z, Li W, Zhu W, Sun S, Ye K, Xie Y, et al. Near-complete genome
518 assembly and annotation of the yellow drum (*Nibea albiflora*) provide insights
519 into population and evolutionary characteristics of this species. Ecology and
520 Evolution. 2019;9 1:568-75. doi:doi:10.1002/ece3.4778.
- 521 44. Barrio AM, Lamichhaney S, Fan GY, Rafati N, Pettersson M, Zhang H, et al.
522 The genetic basis for ecological adaptation of the Atlantic herring revealed by
523 genome sequencing. Elife. 2016;5:e12081. doi:10.7554/eLife.12081.
- 524 45. Amemiya CT, Alföldi J, Lee AP, Fan SH, Philippe H, MacCallum I, et al. The
525 African coelacanth genome provides insights into tetrapod evolution. Nature.
526 2013;496 7445:311-6. doi:10.1038/nature12027.

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
- 527 46. Star B, Nederbragt AJ, Jentoft S, Grimholt U, Malmstrom M, Gregers TF, et
528 al. The genome sequence of Atlantic cod reveals a unique immune system.
529 Nature. 2011;477 7363:207-10. doi:10.1038/nature10342.
- 530 47. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, et al.
531 The genomic basis of adaptive evolution in threespine sticklebacks. Nature.
532 2012;484 7392:55-61. doi:10.1038/nature10944.
- 533 48. Martin M. Cutadapt removes adapter sequences from high-throughput
534 sequencing reads. EMBnet journal. 2011;17 1:10-2. doi:10.14806/ej.17.1.200.
- 535 49. Andrews S: FastQC: a quality control tool for high throughput sequence data.
536 <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (2010). Accessed
537 9/24/2018.
- 538 50. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et
539 al. De novo transcript sequence reconstruction from RNA-seq using the Trinity
540 platform for reference generation and analysis. Nat Protoc. 2013;8 8:1494-
541 512. doi:10.1038/nprot.2013.084.
- 542 51. Smith-Unna R, Boursnell C, Patro R, Hibberd JM and Kelly S. TransRate:
543 reference-free quality assessment of de novo transcriptome assemblies.
544 Genome Res. 2016;26 8:1134-44. doi:10.1101/gr.196469.115.
- 545 52. Finn RD, Clements J and Eddy SR. HMMER web server: interactive
546 sequence similarity searching. Nucleic Acids Res. 2011;39 Web Server
547 issue:W29-37. doi:10.1093/nar/gkr367.
- 548 53. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998;14 9:755-63.
- 549 54. Castresana J. Selection of conserved blocks from multiple alignments for their
550 use in phylogenetic analysis. Mol Biol Evol. 2000;17 4:540-52. doi:DOI
551 10.1093/oxfordjournals.molbev.a026334.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

552 55. Foster PG. Modeling compositional heterogeneity. *Syst Biol.* 2004;53 3:485-
553 95.

554 56. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A and Jermini LS.
555 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat*
556 *Methods.* 2017;14 6:587-9. doi:10.1038/nmeth.4285.

557 57. Nguyen LT, Schmidt HA, von Haeseler A and Minh BQ. IQ-TREE: a fast and
558 effective stochastic algorithm for estimating maximum-likelihood phylogenies.
559 *Mol Biol Evol.* 2015;32 1:268-74. doi:10.1093/molbev/msu300.

560 58. Jones DT, Taylor WR and Thornton JM. The rapid generation of mutation
561 data matrices from protein sequences. *Comput Appl Biosci.* 1992;8 3:275-82.

562 59. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-
563 analysis of large phylogenies. *Bioinformatics.* 2014;30 9:1312-3.
564 doi:10.1093/bioinformatics/btu033.

Figure legends

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

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

Figure 3. Optimal maximum likelihood tree (-ln likelihood: 146565.6438) under a best-fitting JTT+ Γ_4 +F substitution model of 97 concatenated proteins. Maximum likelihood bootstrap support values are given below or to the right of nodes. Scale bar represents mean numbers of substitutions per site. The Actinopterygii ingroup was rooted to two outgroup taxa, namely *Petromyzon marinus* (lamprey) and *Latimeria chalumnae* (coelacanth) (not shown).

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

1 588 **Additional files**

2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

589 **Additional file 1.** Heterozygous SNPs identified in the phased haploid blocks listed
590 in a VCF file format.

591
592 **Additional file 2.** Annotation of all tissues transcriptome assembly in a tabular
593 format.

594
595 **Additional file 3.** Sequence alignment statistics of the 97 proteins concatenated for
596 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

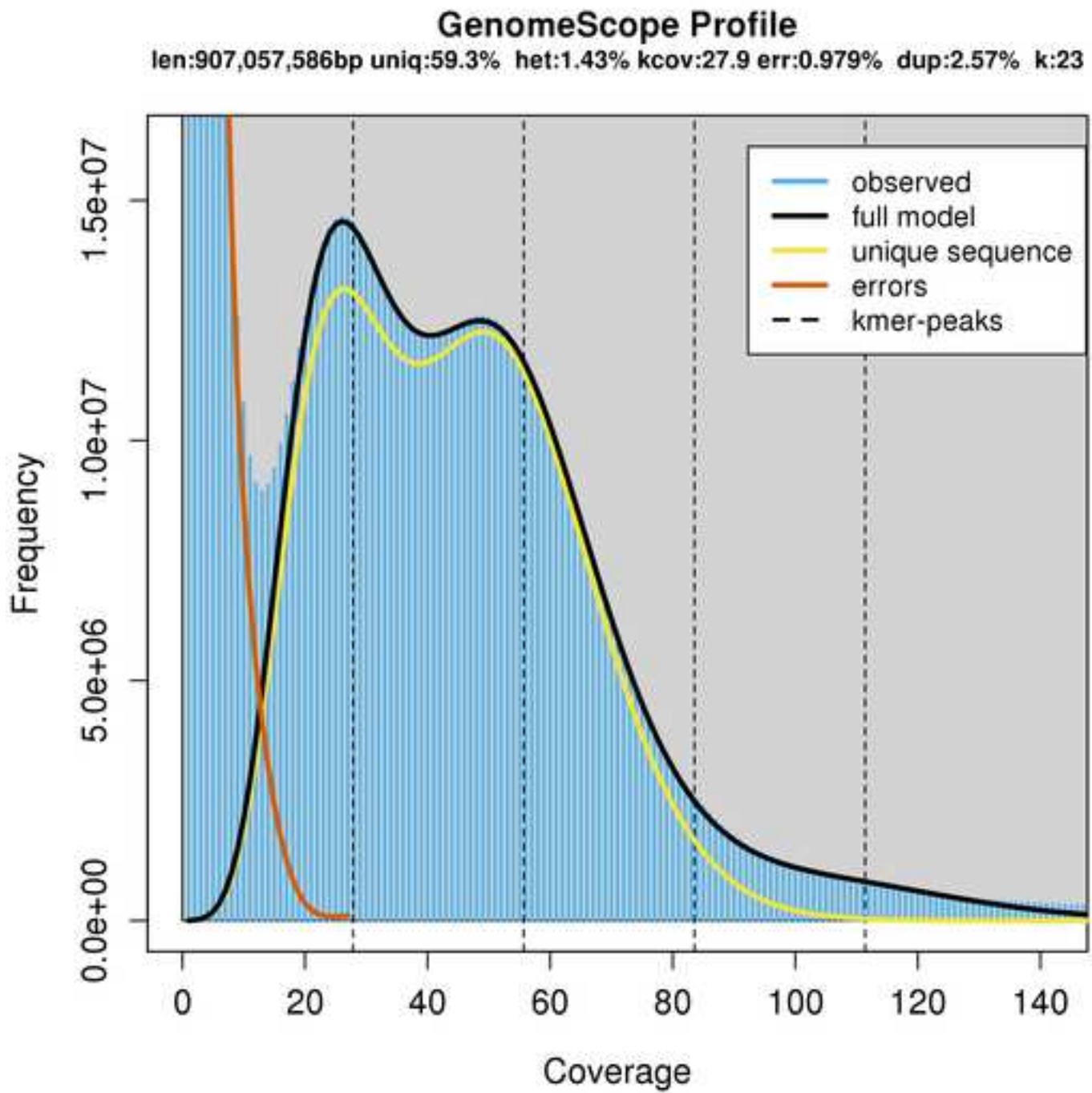
Table 2. Descriptive metrics of sardine genome assemblies. SP_haploid1/ SP_haploid2: haploids genomes ([UOTT01000000](#) and [UOTU01000000](#)). SP_G: consensus genome (NCBI representative genome assembly, GCA_900499035.1). UP_Spi: Illumina paired-end assembled genome from [20] (GCA_003604335.1). Values for scaffolds equal or larger than 1Kb, 10Kb and 100 Kb are presented in separated rows.

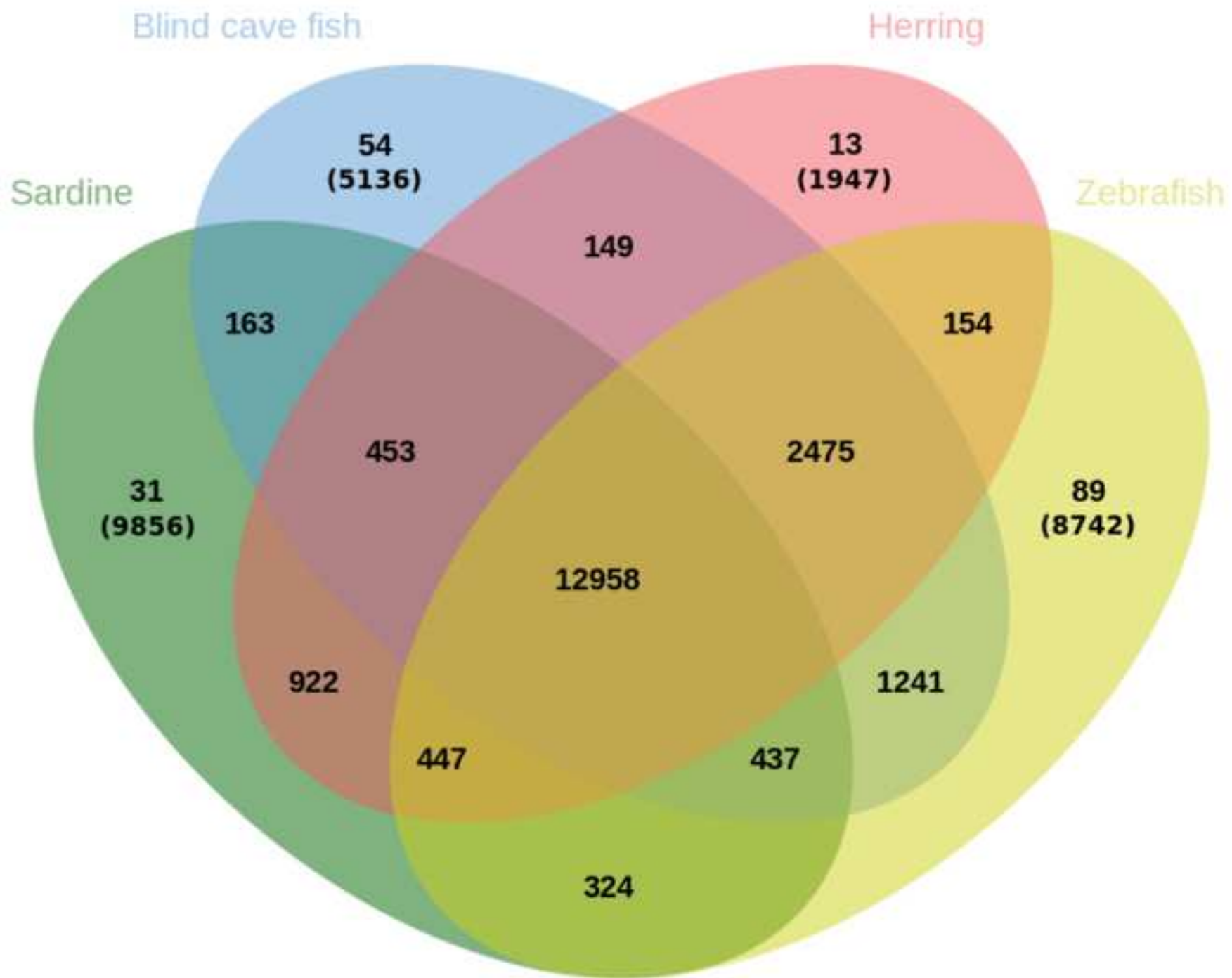
Scaffolds	Spil_haploid1	Spil_haploid2	SP_G	UP_Spi
Largest	6.835 Mb	6.850 Mb	6.843 Mb	0.285 Mb
Number				
>=100Kb	874	872	890	309
>= 10Kb	8 301	8 298	8 760	18 863
>= 1Kb (total)	117 698	117 698	117 259	44 627
L50 / N50				
>=100Kb	135 / 906.0 Kb	134 / 925.2 Kb	137 / 899.1 Kb	130 / 122.5 Kb
>= 10Kb	242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

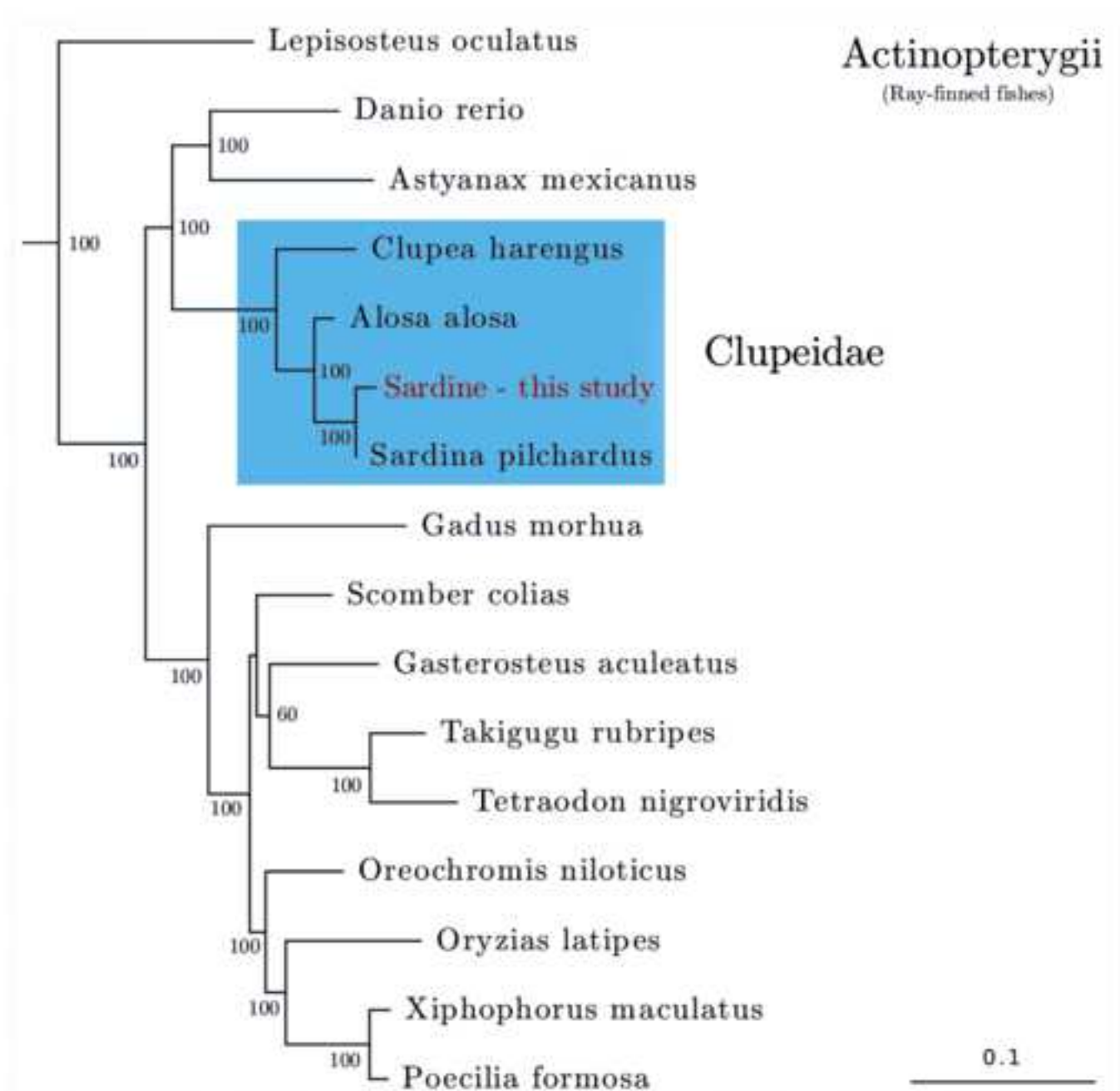
Table 3 – Summary statistics of transcriptome data for the eleven tissues.

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





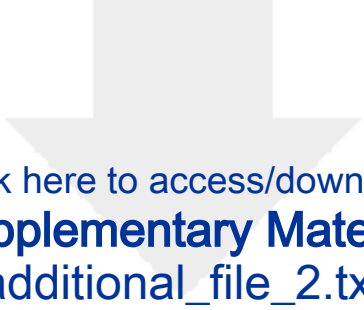




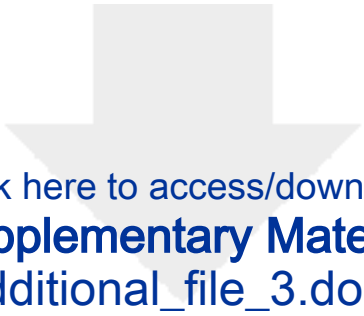


Click here to access/download
Supplementary Material
Spil_SNP_phased_COV10_nbc.vcf






Click here to access/download
Supplementary Material
additional_file_2.txt



Click here to access/download
Supplementary Material
additional_file_3.docx



March 11, 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