Liver transcriptome resources of four commercially exploited teleost species

André M. Machado1✉, Antonio Muñoz-Merida2, Elza Fonseca1,3, Ana Veríssimo2,3, Rui Pinto1, Mónica Felício4, Rute R. da Fonseca5, Elsa Froufe1✉ & L. Filipe C. Castro1,3 ✉

The generation of omic resources is central to develop adequate management strategies for species with economic value. Here, we provide high-coverage RNA-seq datasets of liver tissue (containing between 80.2 and 88.4 million of paired-end reads) from four wildtype teleost species with high commercial value: *Trachurus trachurus* (TTR; Atlantic horse mackerel), *Scomber scombrus* (SSC; Atlantic mackerel), *Trisopterus luscus* (TLU; pout), and *Micromesistius poutassou* (MPO; blue whiting). A comprehensive assembly pipeline, using de novo single and multi-kmer assembly approaches, produced 64 single high-quality liver transcriptomes – 16 per species. The final assemblies, with N50 values ranging from 2,543–3,700 bp and BUSCO (Benchmarking Universal Single-Copy Orthologs) completeness values between 81.8–86.5% of the Actinopterygii gene set, were subjected to open reading frame (ORF) prediction and functional annotation. Our study provides the first transcriptomic resources for these species and offers valuable tools to evaluate both neutral and selected genetic variation among populations, and to identify candidate genes for environmental adaptation assisting in the investigation of the effects of global changes in fisheries.

Background & Summary

Multi-data approaches using complementary techniques are essential to successfully define fish stocks and management strategies (e.g.1,2). The revolution of Next-Generation Sequencing (NGS) has created an unprecedented opportunity to contribute to each component of fisheries management (e.g. reviewed in 3), allowing to address population structure and adaptive divergence in commercially relevant teleost fish species (e.g. Atlantic cod4 and Atlantic herring5), to identify candidate genes for environmental adaptation1 or to explore the function of genes with aquaculture relevance6. Importantly, the expansion of genomic and transcriptomic datasets has been fundamental to detail the complex phylogenetic relationships of this taxon-rich clade7–9. Additionally, these resources have also been proven important in conservation strategies, where they have allowed the prediction of how species will respond to new environmental scenarios and the identification of the threats endangering species at risk (e.g.10). Yet, these formidable tools have to be applied to the vast majority of the world fisheries: 60% of which are at their maximum sustainable yield, 33% are exploited at biologically unsustainable levels and only 7.0% are considered underfished11.

Here, we generated four liver transcriptomic datasets from important fishing resources in European waters belonging to three different taxonomic families: the Atlantic horse mackerel, *Trachurus trachurus* (Linnaeus 1758), the Atlantic mackerel, *Scomber scombrus* (Linnaeus 1758) (SSC), the pout, *Trisopterus luscus* (Linnaeus 1758) and the blue whiting, *Micromesistius poutassou* (Risso 1827) (Fig. 1a). Together, these species represent an important fraction of fish captures in European waters (Fig. 1a) and are, therefore, relevant models for which the development of omic tools for research is highly desirable. Although some information regarding basic biological traits is currently available, some aspects of their biology remain poorly known, particularly population structure, nutritional requirements, reproduction, coastal recruitment processes, and studies involving NGS datasets are

1CIIMAR - Interdisciplinary Centre of Marine and Environmental Research, U. Porto – University of Porto, Porto, Portugal. 2CIBIO-InBIO, Research Network in Biodiversity and Evolutionary Biology, Universidade do Porto, Campus Agrário de Vairão, 4485-661, Vairão, Portugal. 3Department of Biology, Faculty of Sciences, U. Porto - University of Porto, Porto, Portugal. 4Portuguese Institute for the Sea and Atmosphere, I.P. (IPMA), Lisbon, Portugal. 5Center for Macroecology, Evolution and Climate, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark. ✉e-mail: andre.machado@ciimar.up.pt; filipe.castro@ciimar.up.pt
also scarce or absent. For example, in *Scomber scombrus* two RNA-seq projects are available (PRJNA272777 (low coverage RNA-seq dataset produced from a pool of tissues (muscle, liver, gonad, brain) and 454 GS-FLX Titanium System) and PRJNA305977 (RNA-seq dataset produced from white muscle tissue and with Illumina HiSeq2000 system))\textsuperscript{12,13}. For the remaining species, no RNA-seq data is currently available. To cover the above-mentioned biological aspects, the liver was selected to produce a high-quality and coverage transcriptome for each species. Typically, this tissue has a large number of expressed genes. On the other hand, the selection of a single specimen per species avoided the intraspecies variations, which associated with heterozygosity levels of marine fish species, generally higher than freshwater species, makes the construction of high-quality genomic and transcriptomic references a complex and challenging task.

### Methods

**Animal sampling.** Adult specimens of *T. trachurus*, *S. scombrus*, *T. luscus*, and *M. poutassou* were collected to perform the RNA-seq analyses. Individuals were caught in the Northeast Atlantic Ocean, Portugal: *T. trachurus* and *S. scombrus* (40.961667 N, 9.336000 W) and *T. luscus* and *M. poutassou* (41.055000 N, 9.246667 W), under the guidelines of the "Programa Nacional de Amostragem Biológica", conducted by the Instituto Português do Mar e da Atmosfera (IPMA) (Fig. 1a; Table 1). Immediately upon capture, liver tissue from each specimen was collected and stored in RNAlater RNA Stabilization Reagent (Qiagen, Germany) during 24 h at 4 °C (Fig. 1b). The samples were later transferred to −80 °C until total RNA extraction (Fig. 1b).

**RNA extraction, library construction, and sequencing.** Total RNA was extracted from liver using the Illustra RNAspin Mini RNA Isolation Kit (GE Healthcare, UK), according to the manufacturer's instructions. The isolated RNA was treated with RNase-free DNase I to remove residual genomic DNA contamination and eluted in RNase-free water. RNA concentration was measured using a microplate spectrophotometer with Take3\textsuperscript{19} Micro-Volume Plate (BioTek, USA) (*T. trachurus* - 2816.556 ng/μl, *S. scombrus* - 2379.382 ng/μl, *T. luscus* - 1147.368 ng/μl, and *M. poutassou* - 1236.980 ng/μl). The RNA quality was verified with the measurement of the OD260/280 ratio values (1.8 to 2.0). The integrity of each RNA sample was checked by running 1 μl in a 1% agarose gel. Afterwards, the four RNA samples were used to build four strand-specific libraries, one per species, with an insert size of 250–300 bp and sequenced using 150 bp paired-end reads on the Illumina HiSeq4000 platform by Novogene (China).
Table 1. MixS descriptors of four commercial fish species used on this study.

Pre-assembly processing stage. The raw dataset for each specimen was initially inspected with the FastQC (version 0.11.8) software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Trimmomatic (version 0.38) was then used to trim, quality-filter the raw reads and remove Illumina adaptors, under the following parameters (LEADING:5 TRAILING:5 SLIDINGWINDOW:5:20 MINLEN:36) (Fig. 1c). To correct random sequencing errors introduced during the sequencing or in another stage of the pre-in silico processing, we applied a kmer-based error correction method, Rcorrector (version 1.0.3) with default settings. At this stage, all the unfixable reads were discarded. The error-corrected reads were posteriorly introduced in the Centrifuge (version 1.0.3-beta) program and taxonomically classified against the pre-compiled nucleotide database of NCBI (ftp://ftp.ncbi.nlm.nih.gov/pub/infphilo/centrifuge/data/) (version nt_2018_3_3). Importantly, all the reads not classified as belonging to Actinopterygii superclass (Taxon Id: 7898) were considered exogenous to our target species and removed from the initial datasets (Fig. 1c).

De novo transcriptome assembly stage. To build the transcriptome, we opted by the de novo assembly strategy using a multi-kmer approach. Thus, to generate the liver transcriptomes of TTR, SSC, TLU, and MPO we used four assemblers – Trinity (version 2.8.4) rnaSPAdes mode of SPAdes (version 3.11.1) SOAPdenovo-trans (version 1.03) and IDBA-trans (version 1.1.3) (Fig. 1c). The first assembly was carried out by Trinity software with a fixed Kmer of 25 and a strand-specific data parameter on (SS_lib_type RF). The RnaSPAdes tool was then applied with the default parameters (kmer of 73, following the strategy used in the original publication (read length/2 – 1) in the remaining assemblers, we used a multi-kmer approach that required both the mean insert size (IS) and standard deviation (SD) values of the raw dataset. To calculate these values, we used the transcriptome generated by Trinity as a reference, the Bowtie2 (version 2.3.5) to map the clean raw reads, and finally CollectInsertSizeMetrics function of Picard tools (version 2.19.2) to estimate the insert size and standard deviation values. The IDBA-trans assemblies were performed with the SD and IS, previously calculated, the kmer values of 25, 31, 41, 51, 61, 71, 81 and the–ss-fr parameter on. On the other hand, SOAPdenovo-trans used the kmer values of 25, 31, 41, 51, 61, 71, 81 with (-L 200; -F YES) parameters. To remove the gaps inserted during the SOAPdenovo-trans assemblies, the GapCloser (version 1.2) module of SOAPdenovo-trans software was used, with default settings.

Post-assembly processing stage. In the post-assembly processing stage, all assemblies were processed with the SeqKit (version 0.10.1) toolkit. This tool removed all contigs with less than 200 nucleotides, and concatenated all assemblies, per species, in one multi-assembly file. Furthermore, we also conducted several searches against the nucleotide NCBI (nt-NCBI) (downloaded on 30/03/2019) and UniVec (downloaded on 02/04/2019) databases to identify and remove biological contaminations, vectors or adapters not identified in the previous stages. These searches were done via blastn (version 2.9.0) against the nt-NCBI database with the parameters (-evalue 1e-5; -max_target_seqs 1; -perc_identity 90; -max_hsps 1; and minimum alignment length of 50bp), and against UniVec database with the settings (-reward 1; -penalty -5; -gopenpen 3; -gapextend 3; -dist yes; -soft_masking true; -evalue 700; -searchsp 17500000000000). For the nt-NCBI searches, all contigs with the best match hits out of the Actinopterygii taxon were considered contaminations and removed from the transcriptome assemblies. The remaining transcripts, without any match hit or with match hits in Actinopterygii taxon were kept into the transcriptomes and used for further analyses. Regarding the UniVec database, all transcripts with a match hit were considered exogenous and removed from the dataset. To decrease the redundancy and complexity within the decontaminated transcriptomes, we clustered highly similar nucleotide sequences with the CD-HIT-EST (version 4.7) software, with the following settings (-c 0.95; -g 1; -M 60000; -T 30) (Fig. 1c). Essentially, the
software clusters and compares nucleotide sequences, keeping the longest sequence per cluster above a certain similarity threshold, in our case 95% of similarity. All steps of the transcriptome assembly and post-processing stage (from the single kmer assemblies build until the concatenation, decontamination and the clean-up of redundancy) were further inspected to guarantee the accuracy of our approach. Thus, we used the Trinity and Transrate (1.0.3) for primary statistics and the Benchmarking Universal Single-Copy Orthologs (BUSCO version 3.0.2) – with four lineage-specific profile libraries (Metazoa, Eukaryota, Vertebrates and Actinopterygii) – to evaluate the gene completeness of each assembly. In addition, the rate of reads back mapping to the transcriptome (RBMT) was also calculated for all the assemblies after the decontamination step (Fig. 1c). The RBMT was performed with Bowtie2 (version 2.3.5) tool.

Open reading frame prediction and transcriptome annotation. The open reading frames (ORFs) were predicted using the Transdecoder software (version 5.3.0) (https://transdecoder.github.io/) (Fig. 1c). This pipeline is mainly subdivided into three stages. In the first stage, the software pre-predicted the longest ORF per transcript with a cut off length of 100 aminoacids. In the second stage, to find homology and protein evidence, two databases were screened – blast-p (version 2.9.0) with cut-off evalue of 1e-5 against UniProtKB/Swiss-Prot database (downloaded on 12/04/2019) and hmmscan of hmmer2 package (version 2.4i) to find protein profiles against PFAM database (downloaded on 12/04/2019) respectively. In the last stage, all the information collected from both databases, together with the pre-predicted ORFs were used to perform the final prediction of the amino acid sequence. Afterwards, all transcripts codifying for a protein, per species, were used to carry out a functional annotation step with the Sma3s (version 2.1) tool (Fig. 1c). Functional annotation was assigned applying consecutive filters to a blast record (performed against the Uniref90 database (downloaded on 2019-02)) based on both similarity and query coverage. Functional domains were also identified clustering all significant blast hits, and their annotations were retrieved only when their frequency was higher than the frequency of appearance in the reference database following the hypergeometric distribution. Annotation types retrieved consist of GO terms, EC codes from ENZYME repository and Uniprot keywords and pathways. Gene name is only associated in cases of a blast hit greater than 75% of identity and 75% of query coverage.

Data Records
The data generated in this study is subdivided in three main categories: the raw reads, the transcriptome assemblies, and the functional annotation. The decontaminated raw reads, for each species, were deposited in the NCBI Sequence Read Archive – SRP216187. Furthermore, the non-redundant transcriptome assemblies were submitted in fasta format, to the NCBI Transcriptome Shotgun Assembly Sequence Database, under the accession number, GHSR00000000 (TTR), GHRT00000000 (SSC), GHRZ00000000 (TLU), GHRY00000000 (MPO). The remaining transcriptome assemblies (fasta format), the open reading frames, as well the annotation files per species were uploaded to figshare online repository. In detail, it is possible to consult in figshare, the initial individual transcriptome assemblies of the four species produced with four different assemblers, the multi assembly and the final assembly, per species, in fasta format. In addition, also the predicted open reading frames (.pep file), as well as the remaining outputs of TransDecoder software (.bed,.gff3 and.cds files) and the functional annotation files are available, per species.

Technical Validation
Raw datasets and pre-assembly processing quality control. The sequencing process generated a total of 88.4 M in TTR, 87.8 M in SSC, 80.2 M in TLU and 84 M in MPO paired-end raw reads. All raw datasets were initially scrutinized by the FastQC tool, trimmed with Trimmomatic, error-corrected with Rcorrector and decontaminated with Centrifuge software. The percentage of removed reads per step in the pre-assembly processing stage can be visualized in Table 2. Overall, ~86.7 M in TTR, 87.1 M in SSC, 79.2 M in TLU and 82.8 M in MPO reads had Phred scores higher than 20 and were retained for the transcriptome assembly stage (Table 2 and Fig. 2a–d).
Transcriptome assembly metrics. The *de novo* transcriptome assembly was performed using multiple software including Trinity, rnaSPAdes, SOAPdenovo-trans and IDBA-trans. This strategy has been previously applied with success in transcriptome projects of multiple organisms, without a reference genome. The first assembly obtained with Trinity tool was used to map the raw decontaminated reads and to calculate the IS and SD for each species. In all species, the IS and SD values ranged between 267–280 and 59–61 (Table 2). The initial multi-assembly approach generated 64 transcriptome assemblies, 16 for each species. The statistic metrics for each assembly, such as N50 transcript length, mean transcript length, percentage of GC, longest transcript length, were calculated and presented in Table 2.
| Basic statistics | T. trachurus | S. scombrus | T. luscus | M. poutassou |
|-----------------|-------------|-------------|-----------|--------------|
| Number of transcripts | 111866 | 97811 | 150334 | 167124 |
| Longest transcript | 29599 | 44312 | 32751 | 26373 |
| n bases | 223085715 | 237628172 | 247718527 | 288345963 |
| Mean transcript length (bp) | 1994.22 | 2429.46 | 1647.79 | 1725.34 |
| Number of transcripts over 1k nt | 70814 | 68168 | 81638 | 92951 |
| Number of transcripts over 10k nt | 567 | 1050 | 379 | 528 |
| N90 transcript length (bp) | 957 | 1212 | 739 | 776 |
| N70 transcript length (bp) | 1994 | 2477 | 1648 | 1739 |
| N50 transcript length (bp) | 2991 | 3700 | 2543 | 2699 |
| N30 transcript length (bp) | 4370 | 5249 | 3718 | 3926 |
| N10 transcript length (bp) | 7131 | 8290 | 6125 | 6431 |
| Percentage of GC (%) | 0.49 | 0.46 | 0.55 | 0.53 |
| RBMT (%) | 93.78 | 94.63 | 91.45 | 91.52 |
| **BUSCO analysis (%)** | — | — | — | — |
| BUSCO Complete (Single + Duplicated)* | 99.01978583997842 | 97.69974484497681 | 99.0198478885058185 | 99.019888889038242 |
| BUSCO Single* | 36.63348732953061 | 33.33327233643187 | 23.76281230122829 | 23.76281228542829 |
| BUSCO Duplicated* | 62.38629951044782 | 64.36647250854494 | 75.25703557935536 | 75.2570767068535412 |
| BUSCO Fragmented* | 0.6611217546606 | 0.999123596502 | 0.000051507445 | 0.330331460408 |
| BUSCO Missing* | 0.3311284715151 | 1.321339551817 | 0.99910268813370 | 0.660826501350 |
| Total Buscos Found* | 99.67989891538449 | 98.68986789458183 | 99.01989893128630 | 99.34991893505865 |
| **Annotation** | — | — | — | — |
| Transcripts with ORF | 111866 | 97811 | 150334 | 167124 |
| Transcripts annotated with Gene Name | 87847 | 77369 | 93067 | 102433 |
| Transcripts annotated with GO terms | 88269 | 78116 | 93274 | 104086 |
| Transcripts annotated with ENZYME | 32485 | 29241 | 34832 | 39244 |
| Transcripts annotated with PATHWAY | 8653 | 7462 | 8712 | 9760 |
| Final number of transcripts annotated | 90428 | 79911 | 95110 | 106354 |

| Number of transcripts annotated with ORF | 111866 | 97811 | 150334 | 167124 |
| Percentage of GC (%) | 0.49 | 0.46 | 0.55 | 0.53 |
| RBMT (%) | 93.78 | 94.63 | 91.45 | 91.52 |
| **BUSCO analysis (%)** | — | — | — | — |
| BUSCO Complete (Single + Duplicated)* | 99.01978583997842 | 97.69974484497681 | 99.0198478885058185 | 99.019888889038242 |
| BUSCO Single* | 36.63348732953061 | 33.33327233643187 | 23.76281230122829 | 23.76281228542829 |
| BUSCO Duplicated* | 62.38629951044782 | 64.36647250854494 | 75.25703557935536 | 75.2570767068535412 |
| BUSCO Fragmented* | 0.6611217546606 | 0.999123596502 | 0.000051507445 | 0.330331460408 |
| BUSCO Missing* | 0.3311284715151 | 1.321339551817 | 0.99910268813370 | 0.660826501350 |
| Total Buscos Found* | 99.67989891538449 | 98.68986789458183 | 99.01989893128630 | 99.34991893505865 |

**Post-assembly processing and annotation verification.** At this stage, the 16 assemblies per species were concatenated and decontaminated, resulting in four multi-assembly transcriptomes – Online-only Table 4. During the decontamination phase, all blast-n results were manually inspected and the threshold of minimum alignment length of 50 bp and taxonomic superclass Actinopterygii, were specifically selected due to two factors: the considerable number of plausible biological contamination hits with more 50 bp of nucleotide alignment length (e.g. Lathesia calliflora (Taxon ID: 149440)) and the relatively high number of Actinopterygii species with genome sequenced and annotated on nt-NCBI database (at least 43 species), respectively. Although, some transcripts have been discarded at this stage (e.g. novel sequences not present in this 43 Actinopterygii species), we have ensured the application of sensitive blast parameters. Thus, instead of focusing the analyses on the removal of all sequences with a match hit, we performed first an effort on the identification of the hits, via homology, against the possible sources of contamination and only after excluded if confirmed. Apart from a few transcripts that have been removed, this approach increased the confidence levels of each dataset. Posteriorly, four multi-assemblies were subject to a redundancy removal step to decrease the complexity and to remove the overlapping transcripts above 95% of sequence similarity. This method has been highly used to remove redundancy in several datasets and organisms [50-53], namely in the build of new transcriptomic references. In addition, length, or transcript number over 1k or 10k nucleotides can be visualized in Online-only Tables 1–4.

---

### Table 3. Transrate, Busco and RBMT statistics of the final liver transcriptome assemblies of *T. trachurus, S. scombrus, T. luscus, M. poutassou.* **Euk**: Dataset with 303 genes of Eukaryota library profile. **Met**: Dataset with 978 genes of Metazoa library profile. **Ver**: Dataset with 2586 genes of Vertebrata library profile. **Actino**: Dataset with 4584 genes of Actinopterygii library profile.
this strategy still decreased the natural heterozygosity of the organisms in the assemblies. It should be noted that although naturally present in the organisms, the heterozygosity continues to be a crucial parameter to consider during the generation of transcriptomic and genomic datasets. In some cases, where heterozygosity is not contemplated in the bioinformatic approach, the results change completely and their interpretation can be affected. Using this approach, the total number of transcripts in each dataset was substantially reduced, TTR – 2769441 to 414729, SSC – 2728965 to 377586, TLU – 3203445 to 548983, and MPO – 3675167 to 602418, allowing a better overview and understanding of the datasets (Online-only Table 5). After this step, we implemented another quality control strategy, the RBMT. This method, together with the already established metrics, general statistics, and gene completeness, allowed us to verify the raw read content of the transcriptomes. Importantly, this metric showed that even with the redundancy removal step, the RBMT rate in the non-redundant transcriptomes has kept very high values, TTR – 96.78, SSC – 97.16, TLU – 95.3, MPO – 95.51% (Online-only Table 5).

In the process of ORF prediction, the TransDecoder software identified TTR – 111866, SSC – 97811, TLU – 150334, MPO – 167124 transcripts with an assigned ORF. Importantly, the entire set of transcripts codifying for a protein was collected from the non-redundant transcriptomes and placed in files classified as final transcriptome assemblies.

The basic metrics, BUSCO analyses and RBMT rates for TTR, SSC, TLU, and MPO final transcriptome assemblies are available in Table 3. Notwithstanding, the final transcriptome statistics demonstrate the power of the assembly and processing strategy chosen, with N50 values ranging from 2543 to 3700 bp, BUSCO values between 81.8–86.5% in Actinopterygii gene set, and raw reads rate mapping ranging from 91.45 to 94.63% (Table 3). Interestingly, the BUSCO analyses in Vertebrata and Actinopterygii gene sets still showed a decrease in the percentage of fragmented genes (less than half) and missing genes (slightly), in the final transcriptome assemblies when compared with the initial individual assemblies, for all three species.

In the end, the final transcriptomes were functionally annotated using the Sma3s software. A high number of the transcripts were annotated and most of them including the gene name which suggests a remarkable quality of the assemblies. Annotation distribution across the different species is very similar, keeping a logical proportion based on the total number of transcripts. All the annotations stats, including the gene homology, the most probable gene name, the GO terms, Kegg Pathways and EC numbers for Enzymes can be consulted in Table 3.

**Code availability**

All the software programs used in the bioinformatic workflow (transcriptome assembly, pre and post-assembly processing stages and transcriptome annotation) are presented in the Methods section. All programs and databases have the versions, download dates, and parameters described. Software programs with no parameters associated were used with the default settings.

Received: 21 November 2019; Accepted: 9 June 2020;
Published online: 07 July 2020

**References**

1. Begg, G. A. & Waldman, J. R. An holistic approach to fish stock identification. In Fisheries Research 43, 35–44 (Elsevier, 1999).
2. Cadrin, S. X., Kerr, L. A. & Mariani, S. Stock Identification Methods: An Overview. In Stock Identification Methods: Applications in Fishery Science: Second Edition 1–5 (Elsevier Inc., 2013).
3. Kelley, J. L., Brown, A. P., Therkildsen, N. O. & Foote, A. D. The life aquatic: advances in marine vertebrate genomics. Nat. Rev. Genet. 17, 523–534 (2016).
4. LIMBORG, M. T. et al. Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (Clupea harengus). Mol. Ecol. 21, 3686–3703 (2012).
5. Lamichhaney, S. et al. Population-scale sequence scaling reveals genetic differentiation due to local adaptation in Atlantic herring. Proc. Natl. Acad. Sci. U. S. A. 109, 19345–50 (2012).
6. Machado, A. et al. Out of the Can: A Draft Genome Assembly, Liver Transcriptome, and Nutrigenomics of the European Sardine, Sardina pilchardus. Genes (Basel) 9, 485 (2018).
7. Ravi, V. & Venkatesh, B. The divergent genomes of teleosts. Annu. Rev. Anim. Biosci. 6, 47–68 (2018).
8. Hughes, L. C. et al. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. Proc. Natl. Acad. Sci. U. S. A. 115, 6249–6254 (2018).
9. Sun, Y. et al. Fish-T1K (Transcriptomes of 1,000 Fishes) Project: large-scale transcriptome data for fish evolution studies. Gigascience 5, 18 (2016).
10. Gonzalez, E. G. et al. Population Proteomics of the European Hake (Merluccius merluccius). J. Proteome Res. 9, 6392–6404 (2010).
11. FAO. State of world fisheries and aquaculture (SOFIA) - Meeting the sustainable development goals. Food and Agriculture Organization of the United Nations (2018).
12. Alvarez, P. et al. Genomic Resources Notes Accepted 1 June 2015–31 July 2015. Mol. Ecol. Resour. 15, 1510–1512 (2015).
13. Ciezarek, A. G. et al. Substitutions in the glycogenin-1 gene are associated with the evolution of endothermy in sharks and tunas. Genome Biol. Evol. 8, 3011–3021 (2016).
14. Bolger, A. M., Lohse, M. & Usadel, B. Trimmmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30, 2144–2142 (2014).
15. Song, L. & Florea, L. Ccrocotor: efficient and accurate error correction for Illumina RNA-seq reads. Gigascience 4, 48 (2015).
16. Kim, D., Song, L., Jeffrey, S., Byung-soo, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 26, 1721–1729 (2016).
17. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652 (2011).
18. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nat. Protoc. 8, 28 (2013).
19. Bushmanova, E., Antipov, D., Lapidas, A. & Pribeliski, A. D. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. Gigascience 8 (2019).
20. Xie, Y. et al. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics 30, 1660–1666 (2014).
21. Peng, Y. et al. IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. Bioinformatics 29, 1326–1334 (2013).
Acknowledgements
This work was financed by the Project The Sea and the Shore, Architecture and Marine Biology: The Impact of Sea Life on the Built Environment [PTDC/ART-DAQ/29537/2017] from FCT/MCTES through National Funds (PIDDAC) and co-financing from the European Regional Development Fund (FEDER) [POCI-01-0145-FEDER-022184]. R.R.F. thanks the Danish National Research Foundation for its funding of the Center for Macroecology, Evolution, and Climate.

Author contributions
L.F.C.C. supervised the study. A.M.M., L.F.C.C. and R.R.D.F. designed and conceived the work. M.F. collected the samples. E.F. performed the RNA extraction. A.M.M., A.M.-M., A.V. and R.P. conducted the transcriptome analyses. A.M.M., L.F.C.C., R.R.D.F., A.V. and E.F. wrote the first version of the manuscript. All authors revised and contribute to the final version of the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to A.M.M. or L.F.C.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
