A resource for sustainable management

De novo assembly and annotation of the liver transcriptome of the Atlantic chub mackerel, Scomber colias
Machado, André M.; Felício, Mónica; Fonseca, Elza; da Fonseca, Rute R.; Castro, L. Filipe C.

Published in:
Data in Brief

DOI:
10.1016/j.dib.2018.03.013

Publication date:
2018

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Data Article

A resource for sustainable management: De novo assembly and annotation of the liver transcriptome of the Atlantic chub mackerel, Scomber colias

André M. Machado\textsuperscript{a}, Mónica Felício\textsuperscript{b}, Elza Fonseca\textsuperscript{b,c}, Rute R. da Fonseca\textsuperscript{d,⁎}, L. Filipe C. Castro\textsuperscript{a,c,⁎⁎}

\textsuperscript{a} CIIMAR – Interdisciplinary Centre of Marine and Environmental Research, U. Porto – University of Porto, Porto, Portugal
\textsuperscript{b} Portuguese Institute for the Sea and Atmosphere, I.P. (IPMA), Portugal
\textsuperscript{c} Department of Biology, Faculty of Sciences, U. Porto – University of Porto, Portugal
\textsuperscript{d} The Bioinformatics Centre, Department of Biology, University of Copenhagen, Denmark

Abstract

Mackerels represent a valuable fishery worldwide. Their ample geographic distribution and capture levels make them an insightful model to address stock management strategies in the context of global changes. Yet, and despite recent impressive genome and transcriptome sequencing efforts from teleost species, available resources from the Scombridae family are comparatively scarce. Here, we generated the first high-quality de novo assembly of the liver transcriptome of the Atlantic chub mackerel (Scomber colias). Through the use of RNA-Seq Illumina technology, 111,124,228 clean reads were obtained for the liver transcriptome. De novo assembly resulted in 93,731 transcripts with an N50 of 1462 bp. This dataset provides an important insight into the context of fisheries management.

© 2018 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
### Specifications Table

<table>
<thead>
<tr>
<th>Subject area</th>
<th>Genetics and Transcriptomics</th>
</tr>
</thead>
<tbody>
<tr>
<td>More specific subject area</td>
<td>Transcriptionomics of Atlantic chub mackerel Liver</td>
</tr>
<tr>
<td>Type of data</td>
<td>Raw reads of DNA sequences</td>
</tr>
<tr>
<td>How data was acquired</td>
<td>A liver sample of Atlantic chub mackerel, <em>Scomber colias</em>, was collected for total RNA isolation. It was prepared paired-end library and sequenced by the Hiseq. 4000 system. The obtained data were subjected to <em>de novo</em> assembly and annotated using Trinotate.</td>
</tr>
<tr>
<td>Data format</td>
<td>Raw data in FASTQ, transcriptome assembly and Final transcriptome assembly in FASTA format.</td>
</tr>
<tr>
<td>Experimental factors</td>
<td>One specimen of <em>S. colias</em> was obtained from North Atlantic waters.</td>
</tr>
<tr>
<td>Experimental features</td>
<td>The <em>de novo</em> assembling of the transcriptome, decontamination, filtration and the functional annotation of Atlantic chub mackerel liver was performed.</td>
</tr>
<tr>
<td>Data source location</td>
<td>Portugal (41.501944 N 8.851667 W)</td>
</tr>
<tr>
<td>Data accessibility</td>
<td>The raw FASTQ files were deposited in the NCBI SRA database with accession number SRX3462868 (<a href="https://www.ncbi.nlm.nih.gov/sra/SRX3462868">https://www.ncbi.nlm.nih.gov/sra/SRX3462868</a>). The decontaminated transcriptome assembly was deposited in the NCBI TSA database with accession number GGCI00000000.1 (<a href="https://www.ncbi.nlm.nih.gov/nuccore/GGCI00000000.1">https://www.ncbi.nlm.nih.gov/nuccore/GGCI00000000.1</a>). The final transcriptome assembly was deposited in the figshare digital repository (<a href="https://figshare.com/s/a97f1d5b37d174d1d819">https://figshare.com/s/a97f1d5b37d174d1d819</a>).</td>
</tr>
</tbody>
</table>

### Value of the data

- This is the first high-quality *de novo* assembly of the liver transcriptome of the Atlantic chub mackerel.
- The transcriptome results presented here pave the way for developing the appropriate tools fundamental for resource management of *S. colias*.
- The on-going full genome sequence project will largely profit from this dataset at the annotation stage.

### 1. Data

Teleostei, an infraclass of the Actinopterygii, comprise by far the most species-rich group within the vertebrates, with more than 26,000 recognized species [1]. Their fantastic variation in morphology and physiology traits is paralleled by the plethora of colonized aquatic habitats. Moreover, teleost species are a critical component of human diets providing nutrients such as proteins and lipids, including the essential “omega 3” [2,3]. Recently, a massive effort in the gathering of full genome sequences from 66 teleost species has been performed [2,3]. In addition, the implementation of international collaborative initiates with the aim to generate large-scale and comparable datasets of RNA-Seq transcriptome sequences, such as the “Transcriptomes of 1000 Fishes” (Fish-T1K; https://db.cngb.org/fisht1k/home/), is also noteworthy [4]. Mackerels from the genus *Scomber* comprise a substantial proportion of the total volume of captured fish worldwide [5], since they are highly appreciated by consumers. In Portugal, the Atlantic chub mackerel, *Scomber colias*, represents the species with the highest volume of capture recorded in 2016 (> 26,000 t) [6]. Thus, the transcriptomic dataset presented here provides important information for comparative genomics across the teleost tree of life, the definition of stock management strategies, and to investigate the biology...
and ecology of this important economic species. The provided dataset consists of raw reads of Atlantic chub mackerel, deposited in NCBI SRA database under SRX3462868 accession number. The raw reads were de novo assembled into full-length transcriptome, that after filtration, decontamination and quality control were deposited in NCBI TSA database with GGC10000000.1 accession number. In addition, we produced and annotated a filtered transcriptome assembly, final Transcriptome Assembly, derived from the previous one. Importantly, all steps of data treatment were supported by statistical analyses showed in several tables and figures.

2. Experimental design, materials and methods

2.1. Atlantic chub mackerel collection, sampling, and Illumina sequencing

One specimen of S. colias was obtained from North Atlantic waters during the “Programa Nacional de Amostragem Biológica” carried out by the Instituto Português do Mar e da Atmosfera” (IPMA) (Table 1). The liver tissue was sampled immediately upon capture, stored in RNAlater, and kept at −20 °C until RNA extraction. Total RNA (RNAt) of the liver was extracted using the RNeasy Mini Kit (Qiagen) with a pre-treatment with DNasel and subsequent elution of the extracted RNAt in nuclease-free water, according to the manufacturer’s protocol. A strand-specific library with an insert size of 250–300 bp was built after conversion of the high-quality liver RNAt to cDNA and sequenced using 150 bp paired-end reads on the Illumina HiSeq. 4000 platform by Novogene (China).

2.2. Transcriptome data processing and de novo assembly

The raw reads of liver tissue were produced by sequencing and quality filtered by Trimmomatic [7], with parameters set to “LEADING:15 TRAILING:15 SLIDING WINDOW:4:20 MINLEN:50”. The statistics of trimming reads are shown in Table 1.

Since the reference genome of S. colias is not currently available, the Trinity v2.4.0 software was used for de novo assembly [8]. We applied the software following the protocol from Hass and colleagues [9], with exception of the strand-specific data and minimum length contig parameters (SS_lib_type RF; min_contig_length 300).

Table 1
Transrate and Trinity Statistics of the original, decontaminated and final transcriptome assembly of liver transcriptome of S. colias.

<table>
<thead>
<tr>
<th>Trimming &amp; Assembly</th>
<th>Liver tissue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw sequencing reads</td>
<td>121,656,039</td>
</tr>
<tr>
<td>Reads used in assembly</td>
<td>111,124,228</td>
</tr>
<tr>
<td>Percentage of reads submitted to assembly</td>
<td>91.34%</td>
</tr>
<tr>
<td>Assembly Versions</td>
<td>Original transcriptome assembly</td>
</tr>
<tr>
<td>Number of “genes”</td>
<td>72,618</td>
</tr>
<tr>
<td>Number of transcripts</td>
<td>114,174</td>
</tr>
<tr>
<td>n50 transcript length (bp)</td>
<td>1299</td>
</tr>
<tr>
<td>Median transcript length (bp)</td>
<td>544</td>
</tr>
<tr>
<td>Mean transcript length (bp)</td>
<td>899</td>
</tr>
<tr>
<td>Smallest Contig</td>
<td>301</td>
</tr>
<tr>
<td>Largest Contig</td>
<td>14,405</td>
</tr>
<tr>
<td>Number of Contigs over 1k nn</td>
<td>30,346</td>
</tr>
<tr>
<td>Number of Contigs over 10k nn</td>
<td>28</td>
</tr>
<tr>
<td>GC %</td>
<td>45.33</td>
</tr>
<tr>
<td>Total Assembled bases</td>
<td>102,639,256</td>
</tr>
<tr>
<td>RMBT %</td>
<td>–</td>
</tr>
</tbody>
</table>
To check the raw assembled contigs for contamination, the assembled transcriptome was queried in the MCSC decontamination pipeline [10] with the following parameters: LVL = 5; TAX-O_LVL = superclass; WHITE_NAME = Actinopterygii, and all contigs with a match to Actinopterygii sequences of Uniref90 database were kept. The remaining contigs with a match to other taxa or without hits at all, were re-blasted against the nucleotide database (NT) of NCBI, with an E-value cut-off of 1e-5. Again, contigs having top hits outside of Actinopterygii taxa were excluded, while the contigs without hits at all were retained. To check and remove vector sequences, adapter and linkers, not previously identified, we also filtered the assembled transcripts against the UniVec database. Any sequences of our assembly with a strong match against UniVec database (1/1,000,000 chance of a random match for queries of 350 Kb, terminal match score ≥ 24, internal match score ≥ 30) were removed.

To decrease the isoform redundancy of the clean assembly, the tr2aacsds pipeline, from the Evidential – Gene package (http://arthropods.eugenes.org/EvidentialGene/), was used. This pipeline reduces redundancy by selecting the ‘optimal’ set of assembled transcripts, based on coding potential. For each filtration stage of the assembly, the Trinity and Transrate [11] statistics are shown in the Table 1. Furthermore, and despite the redundancy removal and decontamination clean-up steps in the initial stages, the rate of reads mapped back to transcripts (RMBT) as well the distribution length of final assembled sequences was calculated and plotted as a measure of assembly quality (Table 1; Fig. 1).

To assess the completeness of our transcriptome, in terms of gene content, the Benchmarking Universal Single-Copy Orthologs tool (BUSCO) was used [12]. The statistics of complete BUSCO hits against the eukaryota and metazoa lineage-specific profile libraries are provided in Table 2.

---

**Fig. 1.** Length distribution of the final transcriptome contigs. The x-axis represents the length, and the y-axis represents the number of Contigs.

**Table 2**

<table>
<thead>
<tr>
<th>BUSCO Statistics</th>
<th>Metazoa DB (%)</th>
<th>Eukaryota DB (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
<td>82.90</td>
<td>83.20</td>
</tr>
<tr>
<td>Single</td>
<td>73.60</td>
<td>71.60</td>
</tr>
<tr>
<td>Multi</td>
<td>9.30</td>
<td>11.60</td>
</tr>
<tr>
<td>Fragment</td>
<td>13.00</td>
<td>13.20</td>
</tr>
<tr>
<td>Missing</td>
<td>4.10</td>
<td>3.60</td>
</tr>
</tbody>
</table>
2.3. Functional annotation

The functional annotation was performed in two independent steps. Firstly, the final transcriptome assembly was queried against the non-redundant (NR) database of NCBI, through the blastx tool of DIAMOND v0.8.36 software [13] and applying an E-value cut-off of 1e-5. The top 30 species with the blastx best hits are provided in Table 3. To facilitate the visualization, only the top 15 best blast hits are plotted in the species distribution in Fig. 2A. Moreover, and to strengthen the blast analysis, the E-value and sequence similarity distributions were also tabulated and displayed in Fig. 2B and C and in Tables 4 and 5.

In the second step, the nucleotide sequences of the final transcriptome were submitted to the Trinotate v3.0.1 (http://trinotate.github.io). Into Trinotate pipeline were used several annotation software such as Transdecoder (http://transdecoder.github.io), Hmmer v.3.1b1 [14], PFAM [15], TMHMM v.2.0c [16], GOseq [17] and eggNOG v.3.0 [18] to perform the functional annotation of transcriptome. Open reading frames (ORFs) were predicted using the Transdecoder. Obtained ORFs were blasted using the blastp tool of DIAMOND v0.8.36 software [13] against the NCBI NR, Uniref90, and SwissProt databases with an E-value cut-off of 1e-5. To avoid statistical gene overrepresentation, when more than one isoform per ‘gene’ remained after filtering with the tr2aacd pipeline, only the top representative of each ‘gene’ was selected for further analysis. To belong to the final subset of

<table>
<thead>
<tr>
<th>Specie</th>
<th>Taxon ID</th>
<th>Number of blastx Hits</th>
<th>Percentage of blastx Hits (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lates calcarifer</td>
<td>8187</td>
<td>8,503</td>
<td>28.20</td>
</tr>
<tr>
<td>Larimichthys crocea</td>
<td>215358</td>
<td>4,696</td>
<td>15.58</td>
</tr>
<tr>
<td>Stegastes partitus</td>
<td>144197</td>
<td>2,683</td>
<td>8.90</td>
</tr>
<tr>
<td>Acanthochromis polycanthus</td>
<td>80966</td>
<td>2,597</td>
<td>8.61</td>
</tr>
<tr>
<td>Labrus bergylta</td>
<td>56723</td>
<td>1,447</td>
<td>4.80</td>
</tr>
<tr>
<td>Paralichthys oliveaceus</td>
<td>8255</td>
<td>1,346</td>
<td>4.46</td>
</tr>
<tr>
<td>Monopterus albus</td>
<td>43700</td>
<td>1,270</td>
<td>4.21</td>
</tr>
<tr>
<td>Oreoichromis niloticus</td>
<td>8128</td>
<td>822</td>
<td>2.73</td>
</tr>
<tr>
<td>Nototenia coriceps</td>
<td>8208</td>
<td>551</td>
<td>1.83</td>
</tr>
<tr>
<td>Neolamprologus brichardi</td>
<td>32507</td>
<td>431</td>
<td>1.43</td>
</tr>
<tr>
<td>Haplochromis burtoni</td>
<td>8153</td>
<td>375</td>
<td>1.24</td>
</tr>
<tr>
<td>Dicentrarchus labrax</td>
<td>13489</td>
<td>322</td>
<td>1.07</td>
</tr>
<tr>
<td>Cyprinodon variegatus</td>
<td>28743</td>
<td>290</td>
<td>0.96</td>
</tr>
<tr>
<td>Maylandia zebra</td>
<td>106582</td>
<td>262</td>
<td>0.87</td>
</tr>
<tr>
<td>Hippocampus comes</td>
<td>109280</td>
<td>256</td>
<td>0.85</td>
</tr>
<tr>
<td>Fundulus heteroclitus</td>
<td>8078</td>
<td>255</td>
<td>0.85</td>
</tr>
<tr>
<td>Cygnoglossus semilaevis</td>
<td>244447</td>
<td>220</td>
<td>0.73</td>
</tr>
<tr>
<td>Kryptolebias marmoratus</td>
<td>37003</td>
<td>205</td>
<td>0.68</td>
</tr>
<tr>
<td>Austrofundulus limnaeus</td>
<td>52670</td>
<td>203</td>
<td>0.67</td>
</tr>
<tr>
<td>Orzias latipes</td>
<td>8090</td>
<td>199</td>
<td>0.66</td>
</tr>
<tr>
<td>Nothobranchius furzeri</td>
<td>105023</td>
<td>175</td>
<td>0.58</td>
</tr>
<tr>
<td>Poecilia latipinna</td>
<td>48699</td>
<td>164</td>
<td>0.54</td>
</tr>
<tr>
<td>Boleophthalmus pectinrostris</td>
<td>150288</td>
<td>152</td>
<td>0.50</td>
</tr>
<tr>
<td>Tetraodon nigroviridis</td>
<td>99883</td>
<td>151</td>
<td>0.50</td>
</tr>
<tr>
<td>Pandamia nyereeri</td>
<td>303518</td>
<td>144</td>
<td>0.48</td>
</tr>
<tr>
<td>Oncorhynchus mykiss</td>
<td>8022</td>
<td>142</td>
<td>0.47</td>
</tr>
<tr>
<td>Poecilia mexicana</td>
<td>48701</td>
<td>142</td>
<td>0.47</td>
</tr>
<tr>
<td>Takifugu rubripes</td>
<td>31033</td>
<td>139</td>
<td>0.46</td>
</tr>
<tr>
<td>Cyprinus carpio</td>
<td>7962</td>
<td>133</td>
<td>0.44</td>
</tr>
<tr>
<td>Oplegnathus fasciatus</td>
<td>163134</td>
<td>120</td>
<td>0.40</td>
</tr>
</tbody>
</table>
Fig. 2. Blastx analysis of Scomber colias final transcriptome assembly. (A) Homologous gene-species distribution (B) E-value distribution, (C) Similarity distribution.

Table 4
E-value distribution of blastx hits of final transcriptome assembly against NR database.

<table>
<thead>
<tr>
<th>E-values Ranges</th>
<th>Number of blastx Hits</th>
<th>Percentage of blastx Hits (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0 - 1e-100$</td>
<td>10492</td>
<td>34.80</td>
</tr>
<tr>
<td>$1e-100 - 1e-60$</td>
<td>7057</td>
<td>23.41</td>
</tr>
<tr>
<td>$1e-60 - 1e-45$</td>
<td>4366</td>
<td>14.48</td>
</tr>
<tr>
<td>$1e-45 - 1e-30$</td>
<td>3667</td>
<td>12.16</td>
</tr>
<tr>
<td>$1e-30 - 1e-15$</td>
<td>2899</td>
<td>9.62</td>
</tr>
<tr>
<td>$1e-15 - 1e-5$</td>
<td>1667</td>
<td>5.53</td>
</tr>
</tbody>
</table>
unigenes, a sequence would have to obey the following criteria (in this order): 1) codify an ORF, 2) display a blastx or blastp hit in at least one of the 3 chosen databases (NR, Swissprot, Uniref90), 3) represent the longest ORF per 'gene'. The annotation statistics of final transcriptome assembly and final transcriptome subset can be consulted in Table 6. Additionally, the Clusters of Orthologous Groups (COG) screening was performed using the eggNOG database, integrated within the Trinotate pipeline. The COG distribution is available in Fig. 3.
Acknowledgments

We acknowledge the North Portugal Regional Operational Program (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF) that supported this research through the Northern Regional Operational Program (NORTE 2020) through the European Regional Development Fund (ERDF), under the Framework of the Structured Program of R&D&I INNOVMAR - Innovation and Sustainability in the Management and Exploitation of Marine Resources (Reference NORTE-01-0145-FEDER-000035), within the Research Line ECO-SERVICES. EF was funded by the FCT (SFRH/BD/100262/2014). This dataset is part of the CIIMAR-lead initiative Portugal-Fishomics. The biological material used here was collected under the project Programa Nacional de Amostragem Biológica - IPMA.

Transparency document. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2018.03.013.

References
