Genomics/technical resources

**De novo** assembly and annotation of the European abalone *Haliotis tuberculata* transcriptome

Ewan Harney a,⁎, Bruno Dubief a, Pierre Boudry b, Olivier Basuyaux c, Markus B. Schilhabel d, Sylvain Huchette e, Christine Paillard a, Flavia L.D. Nunes a

a Laboratoire des Sciences de l’Environnement Marin (LEMAR), UMR6539 CNRS/UBO/IRD/Ifremer, Institut Universitaire Européen de la Mer, University of Brest (UBO), Université Européenne de Bretagne (UEB), Place Nicolas Copernic, 29280, Plouzané, France
b Ifremer, Laboratoire des Sciences de l’Environnement Marin (LEMAR), UMR 6539 CNRS/UBO/IRD/Ifremer, Centre Bretagne ZI, Pointe du Diable, 29280 Plouzané, France
c SMEL (Synergie Mer Et Littoral), Centre Expérimental, 50560 Blainville-sur-Mer, France
d Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, 24105 Kiel, Germany
e France Haliotis, Kerazan Lilia, 29880 Plouguerneau, France

**A R T I C L E  I N  F O**

Article history:
Received 8 February 2016
Accepted 3 March 2016
Available online xxxx

Keywords:
RNA-Seq
Mollusca
Gastropoda
Development
Vibrio harveyi
Climate change

**A B S T R A C T**

The European abalone *Haliotis tuberculata* is a delicacy and consequently a commercially valuable gastropod species. Aquaculture production and wild populations are subjected to multiple climate-associated stressors and anthropogenic pressures, including rising sea-surface temperatures, ocean acidification and an emerging pathogenic *Vibrio* infection. Transcript expression data provides a valuable resource for understanding abalone responses to variation in the biotic and abiotic environment. To generate an extensive transcriptome, we performed next-generation sequencing of RNA on larvae exposed to temperature and pH variation and on haemolymph of adults from two wild populations after experimental infection with *Vibrio harveyi*. We obtained more than 1.5 billion raw paired-end reads, which were assembled into 328,519 contigs. Filtration and clustering produced a transcriptome of 41,099 transcripts, of which 10,626 (25.8%) were annotated with Blast2GO. A differential expression analysis comparing all samples from the two life stages identified 5690 and 10,759 transcripts with significantly higher expression in larvae and adult haemolymph respectively. This is the greatest sequencing effort yet in the *Haliotis* genus, and provides the first high-throughput transcriptomic resource for *H. tuberculata*.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Coastal marine organisms are increasingly being impacted by global climate change and anthropogenic pressures (Harley et al., 2006). The study of their responses to these stressors is vital for understanding how marine populations and species are able to acclimatise or adapt to future change and for devising management strategies for species of economic interest. The European abalone *Haliotis tuberculata* is a commercially valuable gastropod, with production based on fisheries of wild populations or hatchery-based aquaculture. As is the case with many calcifying species, rising pCO₂ and subsequent ocean acidification increase the energetic cost of calcification (Hofmann and Todgham, 2010). The concurrent effects of rising sea surface temperatures and declining pH have deleterious consequences for the development of calcifying organisms (Parker et al., 2010), with abalone being particularly sensitive to changes in pH and temperature (Byrne, 2011). In addition, the emergence of abalone disease due to *Vibrio harveyi* has been associated with the warming of coastal waters (Nicolas et al., 2002). In order to better understand how the European abalone responds to abiotic and biotic stressors, global transcript expression (RNA-Seq) was examined as part of two ongoing experiments: one that is investigating the effects of pH and temperature on larval development; and the other investigating the immune response of adults to infection with *V. harveyi*. The aim of this work was to provide a detailed transcriptomic data set for *H. tuberculata*. Rather than carrying out separate transcriptome assemblies for each life stage, Illumina sequences from both experiments were combined and assembled into a single transcriptome, allowing generation of a more complete transcriptomic resource, and direct comparison of larvae and adults. This data provides an invaluable resource for identifying transcripts involved in the response of *H. tuberculata* to global change and other anthropogenic stresses, and taken with other recent high throughput transcriptomic data for *Haliotis* species (De Wit and Palumbi, 2013; Franchini et al., 2011; Huang et al., 2014).
Comparison of molluscan transcriptomes assembled de novo from Illumina sequences using Trinity and other bioinformatic assembly tools.

<table>
<thead>
<tr>
<th>Species</th>
<th>Raw reads (millions)</th>
<th>Assembler</th>
<th>Total contigs</th>
<th>Filtered and/or non-redundant contigs</th>
<th>Mean contig length (bp)*</th>
<th>N50 (bp)*</th>
<th>Number of annot. contigs (percent)*</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haliotis tuberculata</td>
<td>1517</td>
<td>Trinity</td>
<td>328,519</td>
<td>41,099</td>
<td>1033</td>
<td>1544</td>
<td>10,626 (25.9)</td>
<td>This study</td>
</tr>
<tr>
<td>H. laevigata</td>
<td>38</td>
<td>Trinity</td>
<td>222,172</td>
<td>97,420</td>
<td>–</td>
<td>1313</td>
<td>20,702 (21.3)</td>
<td>Shiel et al., 2015</td>
</tr>
<tr>
<td>H. midae</td>
<td>25</td>
<td>CLC</td>
<td>22,761</td>
<td>–</td>
<td>260</td>
<td>356</td>
<td>3841 (16.9)</td>
<td>Franchini et al., 2011</td>
</tr>
<tr>
<td>H. rufescens</td>
<td>355.68</td>
<td>CLC</td>
<td>162,928</td>
<td>–</td>
<td>522</td>
<td>653</td>
<td>48,004 (29.5)</td>
<td>De Wit and Palumbi, 2013</td>
</tr>
<tr>
<td>Anadara trapezia</td>
<td>27</td>
<td>Trinity</td>
<td>75,024</td>
<td>63,778</td>
<td>505</td>
<td>505</td>
<td>29,913 (38.7)</td>
<td>Prentis and Pavasovic, 2014</td>
</tr>
<tr>
<td>Clione pyramidalis</td>
<td>400</td>
<td>Trinity</td>
<td>45,739</td>
<td>30,800</td>
<td>618</td>
<td>852</td>
<td>9280 (20.3)</td>
<td>Maas et al., 2015</td>
</tr>
<tr>
<td>Conus tribblei</td>
<td>33.54</td>
<td>Trinity</td>
<td>163,513</td>
<td>–</td>
<td>513</td>
<td>614</td>
<td>21,609 (12.9)</td>
<td>Barghi et al., 2014</td>
</tr>
<tr>
<td>Mytilus galloprovincialis</td>
<td>393</td>
<td>Trinity</td>
<td>–</td>
<td>151,320</td>
<td>570</td>
<td>–</td>
<td>50,998 (33.7)</td>
<td>Moreira et al., 2015</td>
</tr>
<tr>
<td>Nacella pulchra</td>
<td>42.08</td>
<td>Trinity</td>
<td>90,674</td>
<td>–</td>
<td>–</td>
<td>413</td>
<td>20,922 (21.3)</td>
<td>Chu et al., 2014</td>
</tr>
<tr>
<td>Patinopecten yessoensis</td>
<td>55.88</td>
<td>Trinity</td>
<td>135,963</td>
<td>86,521</td>
<td>733</td>
<td>1266</td>
<td>28,228 (20.8)</td>
<td>Sun et al., 2015</td>
</tr>
<tr>
<td>Retzia clavigera</td>
<td>110.77</td>
<td>Trinity</td>
<td>197,324</td>
<td>151,684</td>
<td>499</td>
<td>582</td>
<td>28,948 (14.7)</td>
<td>Ho et al., 2014</td>
</tr>
<tr>
<td>Trionida diomedia</td>
<td>133.1</td>
<td>Trinity</td>
<td>185,546</td>
<td>123,154</td>
<td>74</td>
<td>1353</td>
<td>18,246 (14.8)</td>
<td>Senatore et al., 2015</td>
</tr>
<tr>
<td>Cragonsea gigas</td>
<td>2204</td>
<td>Oases</td>
<td>–</td>
<td>55,651</td>
<td>1659</td>
<td>–</td>
<td>44,912 (80.7)</td>
<td>Riviere et al., 2015</td>
</tr>
<tr>
<td>C. virginica</td>
<td>52.86</td>
<td>Oases</td>
<td>48,562</td>
<td>–</td>
<td>874</td>
<td>11,237</td>
<td>26,689 (19.6)</td>
<td>Zhang et al., 2014</td>
</tr>
<tr>
<td>Elliptio complanata</td>
<td>361.03</td>
<td>CLC</td>
<td>136,000</td>
<td>–</td>
<td>–</td>
<td>982</td>
<td>33,882 (36.4)</td>
<td>Cormann et al., 2014</td>
</tr>
<tr>
<td>Pinctada fascata</td>
<td>26.51</td>
<td>SOAP</td>
<td>723,797</td>
<td>92,964</td>
<td>434</td>
<td>–</td>
<td>33,882 (36.4)</td>
<td>Huang et al., 2013</td>
</tr>
</tbody>
</table>

* Contig stats and annotation reported either from total or non-redundant contigs, depending on cited example.

2. Data description

2.1. Sampling and sequencing

Biological samples used for generating the H. tuberculata transcriptome were collected as part of two ongoing studies in our group. Combining the sequences from two life stages (larvae and adults) and two tissues (whole organism and haemolymph) enables the assembly of a more complete transcriptome which reflects a broader range of functions. Here we present the bioinformatic analysis used to assemble the de novo transcriptome, and provide a differential expression analysis of whole larvae versus adult haemolymph. Detailed differential expression analysis dealing with the experimental treatments performed for each study will be examined in two companion papers.

H. tuberculata larvae were produced following controlled fertilizations carried out at a commercial hatchery (France Haliotis, Plouguerneau, Finistère, France). Larvae were transferred to experimental facilities at Ifremer (Plouzané, Finistère, France) and reared in one of the facilities at Ifremer (Plouzané, Finistère, France) and reared in one of the following four treatments (control = 20 °C pH 8.1; acidified = 20 °C pH 7.9; warmed = 22 °C pH 8.1; interaction = 22 °C pH 7.9), RNA samples deriving from approximately 10,000 pooled larvae were collected by filtration from two replicated experimental tanks for each treatment at 14, 24 and 38 h post-fertilisation. These 24 samples were flash frozen in liquid nitrogen and stored at −80 °C until RNA extraction. Adult haemolymph RNA was collected during successive infection experiments on abalone from two different natural populations in Brittany (Molène and Saint-Malo, average size of individuals = 74 ± 15 mm) with the bacteria V. harveyi (strain ORM4), carried out during the spawning period. Haemolymph was sampled 72 h post-infection, following the first exposure for abalone from Molène and after the third exposure for abalone from Saint-Malo. Uninfected control abalones were sampled at the same time point as infected individuals for both populations. Three replicates were sampled per treatment for a total of 12 samples. Two millilitres of haemolymph was withdrawn with a 5 mL syringe from each abalone, and spun at 200 rcf for 10 min. The serum was removed, 1 mL of TRIzol reagent (Life Technologies) was added to the pellet, and then the samples were frozen and stored at −80 °C until RNA extraction.

All samples were homogenised by bead beating and total RNA was extracted using TRIzol reagent (Life Technologies) according to the manufacturer’s instructions. RNA quality and concentration were determined using an Agilent 2100 RNA Bioanalyser (Agilent, Santa Clara, CA, USA) and a NanoDrop 8000 Spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA) respectively. Library preparation and sequencing of larval samples were carried out at the Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, Germany, with Illumina HiSeq (101 bp paired-end, non-strand-specific), with 8 samples multiplexed per lane (two samples were resequenced later with 126 bp paired-end reads). Library preparation and sequencing of adult haemolymph RNA samples were performed by the GenEpool, University of Edinburgh, Scotland, on Illumina HiSeq (100 bp paired-end, strand-specific). For adult haemolymph RNA, two runs were performed with lanes of 12 multiplexed samples. Raw sequence data is available from the NCBI sequence read archive (SRA) under the accession SRA 303338.

2.2. De novo transcriptome assembly

A total of 849,307,722 and 668,045,828 paired-end reads were generated for larvae and adult haemolymph respectively. Adapters and low quality bases were trimmed using the Trimmomatic tool, with a 4 bp sliding window, cutting when the average quality score dropped below Q = 20 (Bolger et al., 2014), leaving 790,015,978 paired-end reads for larvae (mean length = 99.6 bp) and 544,556,316 paired-end reads for adult haemolymph (mean length = 88.9 bp). Separate in silico normalisation of kmers was carried out for larvae and adult haemolymph (Haas et al., 2013), and the transcriptome was assembled from the combined normalised reads using Trinity ver. 2.0.6 (Grabherr et al., 2011) with default parameters. Initial assembly produced 514,943 contigs of between 224 and 32,662 bp length. Removal of short contigs (< 300 bp) reduced this to 328,519 contigs with an average length of 825 bp and an N50 of 1103 bp. This Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GEAU00000000. The version described in this paper is the first version, GEAU001000000. The relatively large number of contigs was in part a result of sequencing effort, which was an order of magnitude greater than in many comparable molluscan de novo transcriptome assemblies (Table 1). Therefore to maximise confidence in our transcriptome, we carried out a number of filtrations. Initially, lowly expressed contigs were removed by applying a FPKM cut-off of 1. This filtration produced 79,374 contigs, with an average length of 1022 bp and an N50 of 1509 bp. Contigs with high similarity, which likely represent different spliced isoforms or duplicated genes were grouped by Trinity into 41,099 non-redundant ‘unigenes’ (hereafter referred to as transcripts), with an average length of 1015 bp and an N50 of 1529 bp. This filtered transcriptome is used in subsequent
stages of annotation and analysis, and is available in the supplementary material (Appendix A).

2.3. Functional annotation

Coding sequence detection was carried out with TransDecoder (Haas et al., 2013). Among the 41,099 transcripts in the filtered transcriptome were 12,587 (30.6%) that contained at least one open reading frame (ORF) of \( \geq 100 \) amino acids length. Translated ORFs were used as Blastp queries against the NR and Swiss-Prot databases, and non-translated nucleotide sequences from all transcripts were used as Blastx queries against the same databases. All blasts were carried out using an e-value cutoff of \( 1e^{-5} \).

A total of 10,626 transcripts (25.85%) were annotated following Blast analysis: 10,603 transcripts were associated with blast hits in the NR database, and 9,347 transcripts were associated with blast hits in the Swiss-Prot database. Longer transcripts were more likely to contain an ORF and be associated with blast hits (Fig. 1A); however, our total of 10,626 annotated transcripts is comparable in terms of percentage annotation (25.9%) with other Illumina transcriptome assemblies of non-model marine molluscs without a reference genome (Table 1). When blast results from the NR database were given precedence over those from the Swiss-Prot database, the majority of hits among the 10,626 annotated transcripts derived from mollusc species, with four species accounting for over 63% of results: Lottia gigantea (25.22%), Crassostrea gigas (17.46%), Aplysia californica (13.96%), and

Fig. 1. A) Sequence length distribution of 41,099 filtered transcripts showing those with Open Reading Frames (ORFs) and Blast hits in the NR database; B) species distribution of blast hits among the 10,626 annotated transcripts. Results from NR take precedence over Swiss-Prot; C) Gene Ontology (GO) classification of 7380 H. tuberculata Blast2GO annotated transcripts split by category. Results from Swiss-Prot take precedence over NR.
Biomphalaria glabrata (6.55%). Furthermore, the congeners H. discus and H. diversicolor were also among the top 10 most represented species, accounting for 1.56% and 1.46% of hits respectively (Fig. 1B). To estimate the influence of fragmentation on the transcriptome, non-duplicated accession numbers among blast annotated transcripts were counted (with NR results again given priority over Swiss-Prot results). A total of 9378 different accession numbers were present among the 10,626 annotated transcripts, suggesting that the large majority (88.25%) of annotated transcripts had unique identities.

Nevertheless, the number of transcripts presented here is higher than the number of genes commonly reported in marine mollusc genomes (e.g. 21,013 for A. californica, 24,676 for L. gigantea and 32,250 for C. gigas), and the majority remain unannotated. Sequencing of expressed transcripts is expected to identify many novel contigs that do not align to known gene models: up to 30–40% in organisms with reference genomes (Wang et al., 2009). Unidentified transcripts may derive from non-coding RNA (Eddy, 2001), which can represent a large part of the transcriptome of many organisms, and play an important functional role in the regulation of processes such as gene expression (Guttman and Rinn, 2012). A recent study on transcript expression in C. gigas by Riviere et al. (2015) demonstrates that de novo transcriptome assembly will often produce many novel contigs with unknown functions; in that case, over 56,000 contigs were identified including 11,817 contigs that did not match any known region of the reference genome or RefSeq protein records.

Finally, Blast2GO (Götz et al., 2008) was used to retrieve Gene Ontology (GO) terms among blast-annotated transcripts. For this analysis, blast results from the Swiss-Prot database took precedence over those that do not align to known gene models.
from the NR database, as they often include more functional information. As part of the annotation procedure, the InterProScan plugin of Blast2GO was used to search for functional domains (Jones et al., 2014). This step was applied to all 41,099 transcripts, revealing 16,305 transcripts that contained at least one InterPro domain (including 8265 that had not been annotated following blast analysis). For Blast2GO annotation, default parameters were applied, and subsequent GO terms were merged with InterProScan results; furthermore, the annex function of Blast2GO was used to augment GO term annotation among the transcripts. A total of 46,336 GO terms were annotated to 7380 transcripts (17.96% of total transcripts). The number of transcripts with second level GO terms is provided for each of the three GO categories of biological process, molecular function and cellular component (Fig. 1C).

2.4. Comparison of GO terms in larvae and adults

Broad differences in function between larvae and adult haemolymph were visualised by comparing relative contributions of second level GO terms in contigs differentially expressed between the life stages. Differential expression analysis was carried out in R using the package edgeR (McCarthy et al., 2012), and considered normalised expression values from all adult haemolymph libraries (across both populations and both treatments) against normalised expression from all larval libraries (across all stages and all treatments). Using all the libraries increased statistical replication for each life stage, and thus confidence in the identification of differentially expressed transcripts. Of the 41,099 transcripts considered, 16,449 were significantly differentially expressed between adult haemolymph and larvae (Benjamini–Hochberg FDR < 0.01, log-fold change > 5). Specifically, 5690 transcripts were significantly more expressed in larvae, and 10,759 were significantly more expressed in adult haemolymph. Among these transcripts, 1185 and 3537 were blast-annotated in larvae and adult haemolymph respectively, of which 566 and 2416 also had Blast2go annotation. Full 8265 that had not been annotated following blast analysis). For

Acknowledgements

This work was supported by a grant from the Regional Council of Brittany, from the European Funds (ERDF) and supported by LabexMER (ANR-10-LABX-19) and co-funded by a grant from the French Government under the programme “Investissements d’Avenir”. Pierrick Le Soucru and Philippe Minier provided technical aide during larval experiments at Ifremer. Fabien Riera, Gaëlle Richard, Naidja Lainsy and Stéphane Pénatiny assisted with experimental infections of adult abalones. Ewann Corre at Abims (Analysis bioinformatics for marine science, Roscoff Marine Station) provided invaluable advice on Trinity usage.

Appendix A. Supplementary data

Supplemental data to this article can be found online at http://dx.doi.org/10.1016/j.margen.2016.03.002.

References


