De novo assembly and functional annotation of the transcriptome of Mimachlamys varia, a bioindicator marine bivalve

Amélia Viricel, Vanessa Becquet, Emmanuel Dubillot, Eric Pante

To cite this version:


HAL Id: hal-01856136
https://hal.archives-ouvertes.fr/hal-01856136
Submitted on 21 Aug 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
De novo assembly and functional annotation of the transcriptome of Mimachlamys varia, a bioindicator marine bivalve

Amélia Viricel\textsuperscript{a}\textsuperscript{*}, Vanessa Becquet\textsuperscript{a}, Emmanuel Dubillot\textsuperscript{a}, Eric Pante\textsuperscript{a}

Affiliation:

\textsuperscript{a}Littoral Environnement et Sociétés (LIENSs), UMR 7266, CNRS-Université de La Rochelle, 2 rue Olympe de Gouges, F-17042 La Rochelle Cedex 01, France.

Email address of each author:
amelia.viricel@gmail.com; vanessa.becquet@univ-lr.fr; emanuel.dubillot@univ-lr.fr; eric.pante@univ-lr.fr

* Corresponding author: Amélia Viricel
Email: amelia.viricel@gmail.com
Tel: +33 (0)5 46 50 76 58
Developing genomic resources for species used as bioindicators of environmental pollution facilitates identification of new biomarkers of interest. The variegated scallop *Mimachlamys varia* (Pectinidae) is a marine bivalve used to evaluate and monitor chemical contamination on the French Atlantic coast. Because natural populations of this species are commercially harvested, there is particular interest in understanding its responses to environmental pollution and pathogens. We assembled and annotated the transcriptome of *M. varia* obtained from a pool of five tissue types (gills, mantle, digestive gland, gonad, adductor muscle). In depth Illumina sequencing led to the assembly of 333,022 transcripts, covering 98% of genes conserved among eukaryotes.

**Keywords**

Pectinidae, *de novo* assembly, variegated scallop, functional annotation, RNA-seq, biomarker
Introduction

Marine bivalves are considered sentinels of environmental quality because they filter large volumes of seawater and bioaccumulate contaminants in high concentrations (Grosell and Walsh 2006). Within this group, the variegated scallop *Mimachlamys varia* (Pectinidae) has been used in ecotoxicological studies aimed at monitoring environmental contamination levels along the French Atlantic coast (Milinkovitch et al., 2015; Breitwieser et al., 2016 & 2017). These studies have detected significant physiological responses of this bivalve to chronic chemical pollution through the use of biomarkers linked to oxidative stress, mitochondrial respiration and immune system alteration. Additionally, potential long-term effects of chronic chemical contamination on *M. varia* natural populations have been investigated by comparing genetic diversity among sites along the French Atlantic coast (Breitwieser et al. submitted). Past studies on *M. varia* have focused on a few target genes for population genetic analyses and genomic resources are still lacking for this bioindicator species. Investigating gene expression would further our understanding of the responses of this bivalve to chemical pollution, and would allow identifying new biomarkers of interest for biomonitoring environmental quality.

Studies investigating transcriptomic responses to chemical contaminant exposure, performed on other marine bivalves, have revealed differential expression of genes implicated in hydrocarbons (e.g. Cai et al. 2014) and heavy metals detoxification (e.g. Meng et al. 2013). Additionally, several pectinids such as *M. varia* and *Pecten maximus* are harvested for human consumption, and there has also been a growing interest in developing genomic resources to study bivalve immune responses to pathogens (e.g. Pauletto et al. 2014, Gómez-Chiarri et al. 2015). The transcriptomes of other pectinids have been recently described (e.g. *Chlamys farreri*: Cai et al. 2014; *Chlamys nobilis*: Liu et al. 2015). The reference transcriptome of
Mimachlamys varia will i) facilitate upcoming studies of differential gene expression analyses on this bioindicator species, and ii) provide a valuable genomic resource for future comparative transcriptomic studies of pectinid bivalves.

Data description

Sampling, RNA extraction and Illumina sequencing

An adult male variegated scallop (shell length: 46 mm, shell height: 40 mm) was collected in the sublittoral zone of Angoulins, France (Table 1) at low tide in October 2016. Total RNA was isolated from five distinct tissues collected from this individual (digestive gland, mantle, gills, adductor muscle and gonads), using 50 mg of each tissue type. RNA extractions were performed using the Nucleospin RNA Set for Nucleozol kit (Macherey-Nagel). After determining RNA concentration using a Nanodrop 2000 spectrophotometer (Thermo Scientific), the five RNA extractions were pooled in equal amounts (4 µg of RNA per tissue type). The quality of the RNA pool was assessed on an Agilent Bioanalyzer before poly(A) enrichment and normalized random primed cDNA library preparation. The library was sequenced using an Illumina HiSeq 2500 with a modified protocol producing long paired-end reads (2 x 300 bp). Sample and sequencing information is given in Table 1 following MIxS standard descriptors (Yilmaz et al. 2011).

<table>
<thead>
<tr>
<th>Item</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Investigation_type</td>
<td>Eukaryote</td>
</tr>
<tr>
<td>Project_name</td>
<td>Reference transcriptome for Mimachlamys varia</td>
</tr>
<tr>
<td>Lat_Lon</td>
<td>46.09880 ; -1.12740</td>
</tr>
<tr>
<td>Geo_loc_name</td>
<td>Bay of Biscay, France</td>
</tr>
<tr>
<td>Collection_date</td>
<td>17-October-2016</td>
</tr>
<tr>
<td>Environment</td>
<td>Marine water</td>
</tr>
<tr>
<td>Biome</td>
<td>ENVO:01000410</td>
</tr>
<tr>
<td>Feature</td>
<td>ENVO:01000105</td>
</tr>
<tr>
<td>env_Material</td>
<td>ENVO:00002150</td>
</tr>
<tr>
<td><strong>Sequencing method</strong></td>
<td>Illumina HiSeq 2500 paired-end 2x300 bp</td>
</tr>
<tr>
<td>-----------------------</td>
<td>----------------------------------------</td>
</tr>
<tr>
<td><strong>Assembly method</strong></td>
<td>Trinity v 2.4.0</td>
</tr>
<tr>
<td><strong>Accession number of raw reads</strong></td>
<td>SRP127478</td>
</tr>
<tr>
<td><strong>Accession number of transcripts</strong></td>
<td>GGGO00000000</td>
</tr>
</tbody>
</table>

**Table 1.** Data description following MIxS standards.

**De novo transcriptome assembly and quality control**

Sequencing produced 64,291,972 raw reads that were trimmed and quality filtered using Trimmomatic v. 0.36 (parameters: HEADCROP:10 LEADING:15 TRAILING:15 SLIDINGWINDOW:4:15 MINLEN:100) including adapter removal (ILLUMINACLIP: 2:30:10) (Table 2). Reads were then filtered using Deconseq v. 0.4.3 to remove potential transcripts from human, bacteria, viruses, archae and microalgae that could be present in scallop tissues as environmental or laboratory contaminants. In addition to the Deconseq transcript databases, we included the SILVA (complete database release 128, Quast et al. 2013), MarREF and MarDB (Klemetsen et al. 2017) databases, and 4 microalgae genomes (Accession nb NW_011934117.1, NW_005202428.1, NC_011669.1 and NC_012064.1) in this decontamination step. SILVA is a comprehensive database of ribosomal RNA (rRNA) sequences (including Bacteria, Archaea and Eukarya), and thus allowed removal of potentially remaining endogenous and microbial rRNAs from our transcriptome data. The MarREF and MarDB databases comprise genome sequences from marine prokaryotes that could have been present in our sample, particularly in the gut content. In total, 178,425 reads (0.29% of all quality-filtered reads) were excluded using Deconseq (search parameters: 90% minimum identity, 50% minimum coverage). Read quality was assessed using FastQC for raw reads and after each quality control step (Trimmomatic and Deconseq).

The *de novo* assembly was achieved using Trinity v.2.4.0 (Grabherr et al. 2011) with default parameters and *in-silico* normalization. Transdecoder was used to identify putative coding
regions within transcripts (ORFs ≥ 100 AA long, homology to known proteins determined using blastp and pfam searches following http://transdecoder.github.io). Four metrics were used to assess assembly quality, following recommendations from Honaas et al. (2016). First, the proportion of quality-filtered reads mapping back to the assembly was high (93.3%). Second, the N50 based on the longest isoform per gene was 1,378 bp. Third, 98.0% and 97.9% genes that are conserved and widely expressed in eukaryotes and metazoans, respectively, were recovered in our Transdecoder candidate ORFs, as determined using BUSCO v. 3 (Simão et al 2015; Waterhouse et al 2017). Finally, the total number of transcripts ($n = 333,022$) and genes (sensu Trinity; $n = 180,900$) is consistent with other pectinid bivalve assemblies (e.g. Mizuhopecten yessoensis: Meng et al. 2013; Chlamys nobilis: Liu et al. 2015). Other assembly statistics are described in Table 2.

Prior to annotation, Transdecoder candidate ORFs were blasted (blastp, evalue 1e-5 cutoff) against a custom database of 369,629 sequences of non-eukaryotes (viruses, bacteria, archaea) assembled from the curated UniProtKB/Swiss-Prot database (2018-01-30). All sequences returning a significant match to these taxa were excluded.

Transcriptome annotation

Transcript annotation was performed using Blast2GO PRO v. 5.0.8 based on 78,784 Transdecoder candidate ORFs (Supplementary files 1, 2 and 3). Blast searches were performed using blastp (Altschul et al. 1997) against the UniProtKB/Swiss-Prot database (2017-06-06), using an e-value threshold of 1e-3 and retaining 20 blast hits per query. We chose to use a rather lax e-value since we were comparing data from a non-model organism to the curated database. Mapping and annotation were performed in Blast2GO PRO v. 5.0.8 with default settings. The InterProScan pipeline (Finn et al. 2017) was run and GO terms were merged to the Blast2GO annotation. ANNEX annotation augmentation was performed
A total of 36,278 peptide sequences (46%) were annotated (Figure 1). Among those, we detected enzymes involved in immune response (phenol oxidases such as genes belonging to the Laccase family), oxidative stress response (e.g. glutathione peroxidase) and toxin biotransformation (e.g. glutathione S-transferase), which are commonly used as biomarkers in ecotoxicological studies on invertebrates (e.g. Valavanidis et al. 2006; Breitwieser et al. 2016; Luna-Acosta et al. 2017).

**Figure 1.** Gene ontology (GO) functional categories.

**Matches to other Pteriomorphia bivalves**

79.5% of the 78,784 Transdecoder candidate ORFs blasted to a custom Pteriomorphia (a subclass of Bivalvia including scallops, oysters and mussels) TrEMBL database (2018-02-
Among these, 79.5% of peptides best hits corresponded to the Yesso scallop (*Mizuhopecten yessoensis*), which reference genome was recently sequenced (Wang et al. 2017). Best hits to *Crassostrea gigas* represented 9.9%. Other best hits (including Pectinidae, Ostreidae, Mytilidae, Pteriidae and Arcidae) represented 0.2% of peptide sequences.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw reads</td>
<td>64,291,972</td>
</tr>
<tr>
<td>Quality-filtered reads</td>
<td>62,039,219</td>
</tr>
<tr>
<td>Total assembled bases</td>
<td>148,351,501</td>
</tr>
<tr>
<td>% reads mapping back to</td>
<td>93.3</td>
</tr>
<tr>
<td>assembly</td>
<td></td>
</tr>
<tr>
<td>Number of transcripts</td>
<td>333,022</td>
</tr>
<tr>
<td>Number of genes</td>
<td>180,900</td>
</tr>
<tr>
<td>Median contig length</td>
<td>445</td>
</tr>
<tr>
<td>Average contig length</td>
<td>820</td>
</tr>
<tr>
<td>contig N50 (in bp, based</td>
<td>1,378</td>
</tr>
<tr>
<td>on the longest isoform)</td>
<td></td>
</tr>
<tr>
<td>Transdecoder peptides</td>
<td>78,784</td>
</tr>
<tr>
<td>BUSCO Eukaryote</td>
<td>C: 98.0% [S: 43.9%, D: 54.1%],</td>
</tr>
<tr>
<td></td>
<td>F: 1.3%, M: 0.7%, n: 303</td>
</tr>
<tr>
<td>BUSCO Metazoa</td>
<td>C: 97.9% [S: 41.9%, D: 56.0%],</td>
</tr>
<tr>
<td></td>
<td>F: 1.0%, M: 1.1%, n: 978</td>
</tr>
</tbody>
</table>

**Table 2.** Assembly statistics. These statistics are based on the longest isoform per gene. The terminology corresponds to assembly with Trinity. BUSCO codes indicate the percentage of widely expressed genes that were recovered completely (C) (for single-copy (S) and duplicated (D) genes), that were only partially recovered (F for “Fragmented”), or that were missing (M). The total number of orthologous groups of genes (n) that was searched in BUSCO is also indicated.

**Data availability**
Raw reads are available through the NBCI Sequence Read Archive (SRP127478). The Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GGGO00000000. The version described in this paper is the first version, GGGO01000000. Both data sources are linked to the NCBI BioSample and BioProject numbers SAMN08235964 and PRJNA427371, respectively.

Acknowledgments

This work was supported by the contrat de plan Etat-Région (CPER/FEDER) ECONAT (RPC DYPOMAR). We would like to thank Nathalie Imbert and Denis Fichet for coordinating axe 2 of DYPOMAR and Benoit Simon-Bouhet for additional funding. RNA extractions were prepared at the Molecular Core Facility at the University of La Rochelle. We thank two anonymous reviewers for their constructive comments on the manuscript.

Supplementary data

Supplementary data to this article can be found at xxxx.

References


Short-term and long-term biological effects of chronic chemical contamination on natural populations of a marine bivalve. PLoS ONE 11, e0150184


Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N.,

Regev, A. 2011. Full-length transcriptome assembly from RNA-se data without a reference
genome. Nat Biotechnol. 29, 644-652


Oceanography 19, 126-133


transcriptome assemblies: lessons learned by leveraging the best plant genome. PLoS
ONE 11(1), e0146062

Klemetsen, T., Raknes, I.A., Fu, J., Agafonov, A., Balasundaram, S.V., Tartari, G.,

Robertsen, E., Willassen, N.P. 2017. The MAR databases: development and
implementation of databases specific for marine metagenomics. Nucl. Acids. Res, gkx1036


S., Zhang, G. 2015. A de novo transcriptome of the noble scallop, Chlamys nobilis,
focusing on mining transcripts for carotenoid-based coloration. BMC Genomics 16:44

Luna-Acosta, A., Breitwieser, M., Renault, T., Thomas-Guyon, H. 2017. Recent findings on
phenoloxidases in bivalves. Marine Pollution Bulletin 122, 5-16


Wang, L. 2013. De novo characterization of Japanese scallop Mizuhopecten yessoensis
transcriptome and analysis of its gene expression following cadmium exposure. PLoS
ONE, 8, e64485

Milinkovitch, T., Bustamante, P., Huet, V., Reigner, A., Churlaud, C., Thomas-Guyon, H.

2015. In situ evaluation of oxidative stress and immunological parameters as
ecotoxicological biomarkers in a novel sentinel species (Mimachlamys varia). Aquatic
Toxicology 161, 170-175


