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

Amélia Viricel, Vanessa Buren Becquet, Emmanuel Dubillot, Eric Pante

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

Amélia Viricel[a]*, Vanessa Becquet[a], Emmanuel Dubillot[a], Eric Pante[a]

Affiliation :

[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

Abstract

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).

Item	Description
Investigation_type	Eukaryote
Project_name	Reference transcriptome for *Mimachlamys varia*
Lat_Lon	46.09880 ; -1.12740
Geo_loc_name	Bay of Biscay, France
Collection_date	17-October-2016
Environment	Marine water
Biome	ENVO:01000410
Feature	ENVO:01000105
env_Material	ENVO:00002150

Sequencing method	Illumina HiSeq 2500 paired-end 2x300 bp
Assembly method	Trinity v 2.4.0
Accession number of raw reads	SRP127478
Accession number of transcripts	GGGO00000000

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

(Myhre et al 2006). 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. glutathine 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).

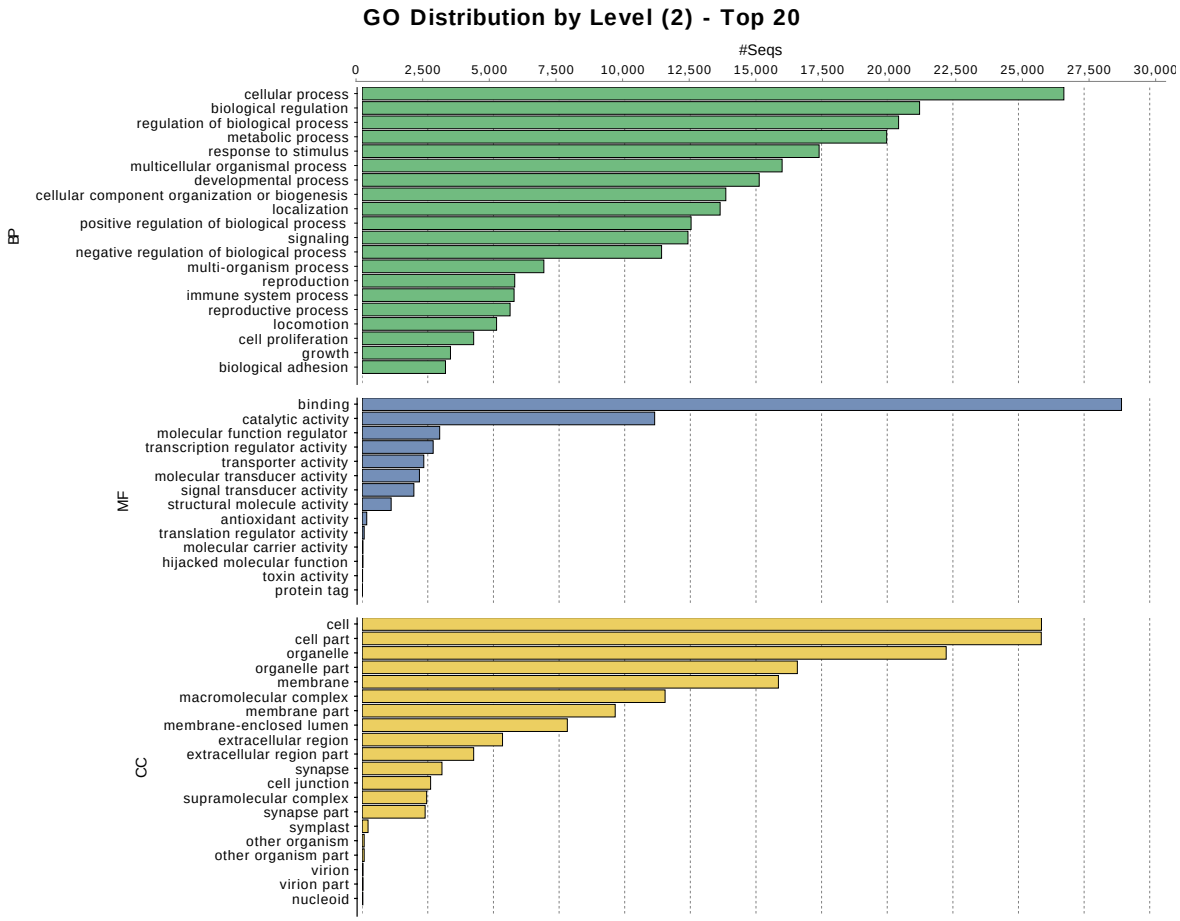

Figure 1. Gene ontology (GO) functional categories.

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-

06). 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.

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

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

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zheng Zhang J.Z., Miller, W. and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25:3389-3402.

Bolger, A.M., Lohse, M., Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114-2120

Breitwieser, M., Viricel, A., Graber, M., Murillo, L., Becquet, V., Churlaud, C., Fruitier-Arnaudin, I., Huet, V., Lacroix, C., Pante, E., Le Floch, S., Thomas-Guyon, H. 2016.

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

Breitwieser, M., Viricel, A., Churlaud, C., Guillot, B., Martin, E., Stenger, P-L., Huet, V., Fontanaud, A., Thomas-Guyon, H. 2017. First data on three bivalve species exposed to an intra-harbour polymetallic contamination (La Rochelle, France). Comparative Biochemistry and Physiology, Part C 199, 28-37

Breitwieser, M., Becquet, V., Thomas-Guyon, H., Pillet, V., Sauriau, P-G., Graber, M., Viricel, A. (submitted) Genetic evidence for a single population of variegated scallop *Mimachlamys varia* across a biogeographic break on the French coastline. Submitted to Journal of Molluscan Studies

Cai, Y., Pan, L., Hu, F., Jin, Q., Liu, T. 2014. Deep sequencing-based transcriptome profiling analysis of *Chlamys farreri* exposed to benzo[a]pyrene. Gene 551,261-270

Finn, R.D., Attwood, T.K., Babbitt, P.C., Bateman, A., Bork, P., Bridge, A.J., Chang, H-Y., Dosztányi, Z., El-Gebali, S., Fraser, M., Gough, J., Haft, D., Holliday, G.L., Huang, H., Huang, X., Letunic, I., Lopez, R., Lu, S., Marchler-Bauer, A., Mi, H., Mistry, J., Natale, D.A., Necci, M., Nuka, G., Orengo, C.A., Park, Y., Pesseat, S., Piovesan, D., Potter, S.C., Rawlings, N.D., Redaschi, N., Richardson, L., Rivoire, C., Sangrador-Vegas, A., Sigrist, C., Sillitoe, I., Smithers, B., Squizzato, S., Sutton, G., Thanki, N., Thomas, P.D., Tosatto, S.C.E., Wu, C.H., Xenarios, I., Yeh, L-S., Young, S-Y, Mitchell, A.L. 2017. InterPro in 2017—beyond protein family and domain annotations. Nucleic Acids Research 45, D190–D199

Gómez-Chiarri, M., Guo, X., Tanguy, A., He, Y., Proestou, D. 2015. The use of –omic tools in the study of disease processes in marine bivalve mollusks. Journal of Invertebrate Pathology 131, 137-154

Grabherr, M.G., Haas, B.J., Yassour M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Rauchowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A.,

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

Grosell, M., Walsh, P.J. 2006. Sentinel species and animal models of human health. Oceanography 19, 126-133

Honaas, L.A., Wafula, E.K., Wickett, N.J., Der, J.P., Zhang, Y., Edger, P.P., Altman, N.S., Pires, J.C., Leebens-Mack, J.H., dePamphilis C.W. 2016. Selecting superior de novo 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

Liu, H., Zheng, H., Zhang, H., Deng, L., Liu, W., Wang, S., Meng, F., Wang, Y., Guo, Z., Li, 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

Meng, X-L., Liu, M., Jiang, K-y., Wang, B-j., Tian, X., Sun, S-j., Luo, Z-y., Qiu, C-w., 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

237 Myhre, S., Tveit, H., Mollestad, T., Laegreid, A. 2006. Additional gene ontology structure for
238 improved biological reasoning. Bioinformatics 22, 2020-7
239 Pauletto, M., Milan, M., Moreira, R., Novoa, B., Figueras, A., Babbucci, M., Patarnello, T.,
240 Bargelloni, L. 2014. Deep transcriptome sequencing of *Pecten maximus* hemocytes : A
241 genomic resource for bivalve immunology. Fish & Shellfish Immunology 37, 154-165
242 Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner,
243 F.O. 2013. The SILVA ribosomal RNA gene database project: improved data processing
244 and web-based tools. Nucl. Acids. Res 41(D1), D590-D596
245 Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M. 2015.
246 BUSCO: assessing genome assembly and annotation completeness with single-copy
247 orthologs. Bioinformatics 31, 3210-3212
248 Valavanidis, A., Vlahogianni, T., Dassenakis, M., Scoullos, M. 2006. Molecular biomarkers
249 of oxidative stress in aquatic organisms in relation to toxic environmental pollutants.
250 Ecotoxicology and Environmental Safety 64, 178-189
251 Wang, S., Zhang, J., Jiao, W., Li, J., Xun, X., Sun, Y., Guo, X., Huan, P., Dong, B., Zhang,
252 L., et al. 2017. Scallop genome provides insights into evolution of bilaterian karyotype and
253 development. Nature Ecology & Evolution 1, 0120.
254 Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G.,
255 Kriventseva, E.V., Zdobnov, E.M. 2017. BUSCO applications from quality assessments to
256 gene prediction and phylogenomics. Molecular Biology and Evolution, doi:
257 10.1093/molbev/msx319
258 Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J.R., Amaral-Zettler, L., Gilbert, J.A.,
259 Karsch-Mizrachi, I., Johnston, A., Cochrane, G., et al. 2011. Minimum information about a
260 marker gene sequence (mimarks) and minimum information about any (x) sequence (mixs)
261 specifications. Nature Biotechnology 29, 415–420.