



**HAL**  
open science

## 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

### ► To cite this version:

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. *Marine Genomics*, 2018, 41, pp.42-45. 10.1016/j.margen.2018.04.002 . hal-01856136

**HAL Id: hal-01856136**

**<https://hal.science/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.

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

3

4

5 Amélia Viricel<sup>a\*</sup>, Vanessa Becquet<sup>a</sup>, Emmanuel Dubillot<sup>a</sup>, Eric Pante<sup>a</sup>

6

7 Affiliation :

8 <sup>a</sup>*Littoral Environnement et Sociétés (LIENSs), UMR 7266, CNRS-Université de La Rochelle,*  
9 *2 rue Olympe de Gouges, F-17042 La Rochelle Cedex 01, France.*

10

11 Email address of each author :

12 *amelia.viricel@gmail.com; vanessa.becquet@univ-lr.fr; emanuel.dubillot@univ-lr.fr;*  
13 *eric.pante@univ-lr.fr*

14

15 \* Corresponding author: Amélia Viricel

16 Email: amelia.viricel@gmail.com

17 Tel: +33 (0)5 46 50 76 58

18

19

20

21

22

23

24 **Abstract**

25 Developing genomic resources for species used as bioindicators of environmental pollution  
26 facilitates identification of new biomarkers of interest. The variegated scallop *Mimachlamys*  
27 *varia* (Pectinidae) is a marine bivalve used to evaluate and monitor chemical contamination  
28 on the French Atlantic coast. Because natural populations of this species are commercially  
29 harvested, there is particular interest in understanding its responses to environmental pollution  
30 and pathogens. We assembled and annotated the transcriptome of *M. varia* obtained from a  
31 pool of five tissue types (gills, mantle, digestive gland, gonad, adductor muscle). In depth  
32 Illumina sequencing led to the assembly of 333,022 transcripts, covering 98% of genes  
33 conserved among eukaryotes.

34

35 **Keywords**

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

37

38

39

40

## 41 **Introduction**

42 Marine bivalves are considered sentinels of environmental quality because they filter large  
43 volumes of seawater and bioaccumulate contaminants in high concentrations (Grosell and  
44 Walsh 2006). Within this group, the variegated scallop *Mimachlamys varia* (Pectinidae) has  
45 been used in ecotoxicological studies aimed at monitoring environmental contamination  
46 levels along the French Atlantic coast (Milinkovitch et al., 2015; Breitwieser et al., 2016 &  
47 2017). These studies have detected significant physiological responses of this bivalve to  
48 chronic chemical pollution through the use of biomarkers linked to oxidative stress,  
49 mitochondrial respiration and immune system alteration. Additionally, potential long-term  
50 effects of chronic chemical contamination on *M. varia* natural populations have been  
51 investigated by comparing genetic diversity among sites along the French Atlantic coast  
52 (Breitwieser et al. submitted). Past studies on *M. varia* have focused on a few target genes for  
53 population genetic analyses and genomic resources are still lacking for this bioindicator  
54 species. Investigating gene expression would further our understanding of the responses of  
55 this bivalve to chemical pollution, and would allow identifying new biomarkers of interest for  
56 biomonitoring environmental quality.

57

58         Studies investigating transcriptomic responses to chemical contaminant exposure,  
59 performed on other marine bivalves, have revealed differential expression of genes implicated  
60 in hydrocarbons (e.g. Cai et al. 2014) and heavy metals detoxification (e.g. Meng et al. 2013).  
61 Additionally, several pectinids such as *M. varia* and *Pecten maximus* are harvested for human  
62 consumption, and there has also been a growing interest in developing genomic resources to  
63 study bivalve immune responses to pathogens (e.g. Pauletto et al. 2014, Gómez-Chiarri et al.  
64 2015). The transcriptomes of other pectinids have been recently described (e.g. *Chlamys*  
65 *farreri*: Cai et al. 2014; *Chlamys nobilis*: Liu et al. 2015). The reference transcriptome of

66 *Mimachlamys varia* will i) facilitate upcoming studies of differential gene expression analyses  
67 on this bioindicator species, and ii) provide a valuable genomic resource for future  
68 comparative transcriptomic studies of pectinid bivalves.

69

## 70 **Data description**

### 71 *Sampling, RNA extraction and Illumina sequencing*

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

84

<b>Item</b>	<b>Description</b>
Investigation_type	Eukaryote
Project_name	Reference transcriptome for <i>Mimachlamys varia</i>
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	GGGO000000000

---

85

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

87

88 *De novo transcriptome assembly and quality control*

89 Sequencing produced 64,291,972 raw reads that were trimmed and quality filtered using

90 Trimmomatic v. 0.36 (parameters : HEADCROP:10 LEADING:15 TRAILING:15

91 SLIDINGWINDOW:4:15 MINLEN:100) including adapter removal (ILLUMINACLIP:

92 2:30:10) (Table 2). Reads were then filtered using Deconseq v. 0.4.3 to remove potential

93 transcripts from human, bacteria, viruses, archae and microalgae that could be present in

94 scallop tissues as environmental or laboratory contaminants. In addition to the Deconseq

95 transcript databases, we included the SILVA (complete database release 128, Quast et al.

96 2013), MarREF and MarDB (Klemetsen et al. 2017) databases, and 4 microalgae genomes

97 (Accession nb NW\_011934117.1, NW\_005202428.1, NC\_011669.1 and NC\_012064.1) in

98 this decontamination step. SILVA is a comprehensive database of ribosomal RNA (rRNA)

99 sequences (including Bacteria, Archaea and Eukarya), and thus allowed removal of

100 potentially remaining endogenous and microbial rRNAs from our transcriptome data. The

101 MarREF and MarDB databases comprise genome sequences from marine prokaryotes that

102 could have been present in our sample, particularly in the gut content. In total, 178,425 reads

103 (0.29% of all quality-filtered reads) were excluded using Deconseq (search parameters: 90%

104 minimum identity, 50% minimum coverage). Read quality was assessed using FastQC for

105 raw reads and after each quality control step (Trimmomatic and Deconseq).

106 The *de novo* assembly was achieved using Trinity v.2.4.0 (Grabherr et al. 2011) with default

107 parameters and *in-silico* normalization. Transdecoder was used to identify putative coding

108 regions within transcripts (ORFs  $\geq$  100 AA long, homology to known proteins determined  
109 using blastp and pfam searches following <http://transdecoder.github.io>). Four metrics were  
110 used to assess assembly quality, following recommendations from Honaas et al. (2016). First,  
111 the proportion of quality-filtered reads mapping back to the assembly was high (93.3%).  
112 Second, the N50 based on the longest isoform per gene was 1,378 bp. Third, 98.0% and  
113 97.9% genes that are conserved and widely expressed in eukaryotes and metazoans,  
114 respectively, were recovered in our Transdecoder candidate ORFs, as determined using  
115 BUSCO v. 3 (Simão et al 2015 ; Waterhouse et al 2017). Finally, the total number of  
116 transcripts ( $n = 333,022$ ) and genes (*sensu* Trinity;  $n = 180,900$ ) is consistent with other  
117 pectinid bivalve assemblies (e.g. *Mizuhopecten yessoensis*: Meng et al. 2013; *Chlamys*  
118 *nobilis*: Liu et al. 2015). Other assembly statistics are described in Table 2.

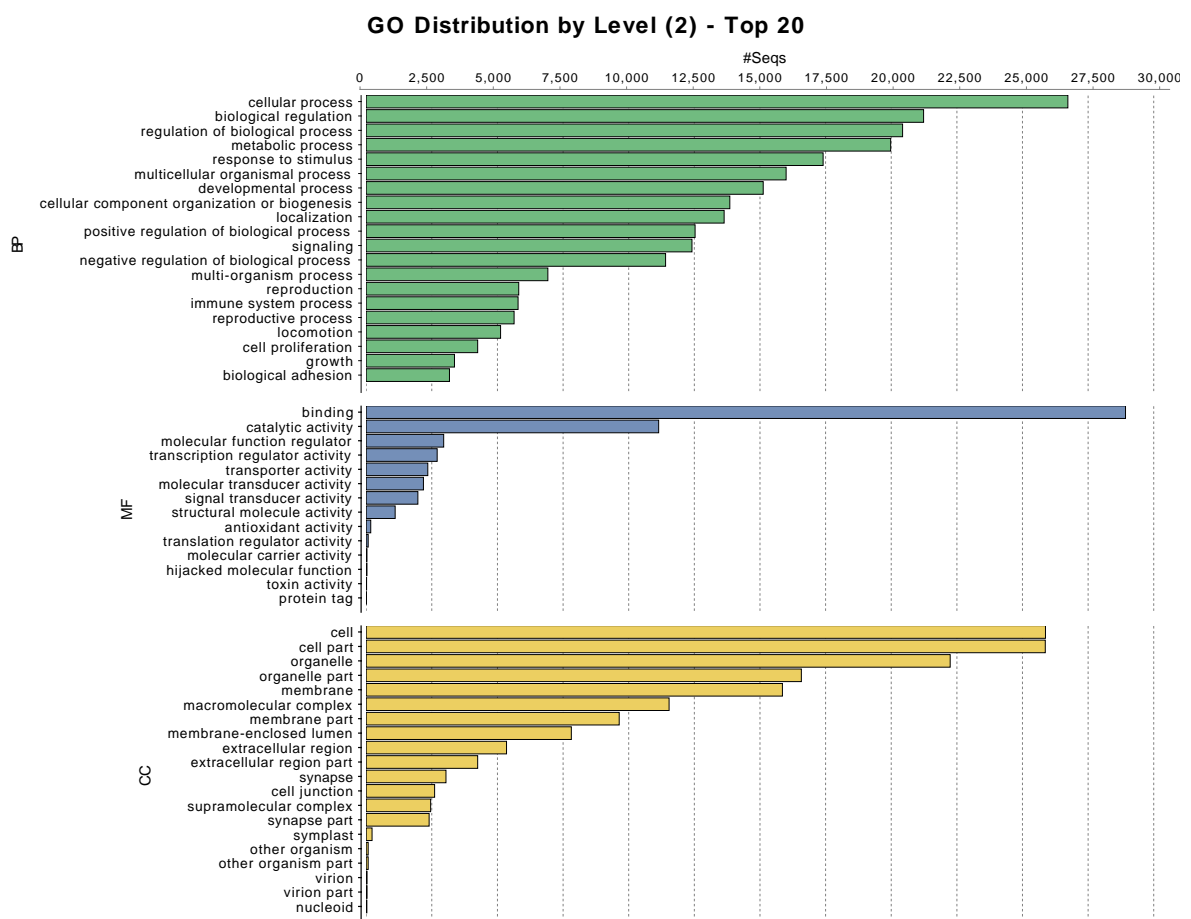
119 Prior to annotation, Transdecoder candidate ORFs were blasted (blastp, evaluate  $1e-5$  cutoff)  
120 against a custom database of 369,629 sequences of non-eukaryotes (viruses, bacteria, archaea)  
121 assembled from the curated UniProtKB/Swiss-Prot database (2018-01-30). All sequences  
122 returning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

123

#### 124 *Transcriptome annotation*

125 Transcript annotation was performed using Blast2GO PRO v. 5.0.8 based on 78,784  
126 Transdecoder candidate ORFs (Supplementary files 1, 2 and 3). Blast searches were  
127 performed using blastp (Altschul et al. 1997) against the UniProtKB/ Swiss-Prot database  
128 (2017-06-06), using an e-value threshold of  $1e-3$  and retaining 20 blast hits per query. We  
129 chose to use a rather lax e-value since we were comparing data from a non-model organism to  
130 the curated database. Mapping and annotation were performed in Blast2GO PRO v. 5.0.8 with  
131 default settings. The InterProScan pipeline (Finn et al. 2017) was run and GO terms were  
132 merged to the Blast2GO annotation. ANNEX annotation augmentation was performed

133 (Myhre et al 2006). A total of 36,278 peptide sequences (46%) were annotated (Figure 1).  
 134 Among those, we detected enzymes involved in immune response (phenol oxidases such as  
 135 genes belonging to the Laccase family), oxidative stress response (e.g. glutathione  
 136 peroxidase) and toxin biotransformation (e.g. glutathione S-transferase), which are commonly  
 137 used as biomarkers in ecotoxicological studies on invertebrates (e.g. Valavanidis et al. 2006;  
 138 Breitwieser et al. 2016; Luna-Acosta et al. 2017).  
 139



140  
 141 **Figure 1.** Gene ontology (GO) functional categories.

142  
 143 *Matches to other Pteriomorphia bivalves*

144 79.5% of the 78,784 Transdecoder candidate ORFs blasted to a custom Pteriomorphia (a  
 145 subclass of Bivalvia including scallops, oysters and mussels) TrEMBL database (2018-02-



146 06). Among these, 79.5% of peptides best hits corresponded to the Yesso scallop  
 147 (*Mizuhopecten yessoensis*), which reference genome was recently sequenced (Wang et al  
 148 2017). Best hits to *Crassostrea gigas* represented 9.9%. Other best hits (including Pectinidae,  
 149 Ostreidae, Mytilidae, Pteriidae and Arcidae) represented 0.2% of peptide sequences.  
 150  
 151  
 152

---

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

---

153  
 154 **Table 2.** Assembly statistics. These statistics are based on the longest isoform per gene. The  
 155 terminology corresponds to assembly with Trinity. BUSCO codes indicate the percentage of  
 156 widely expressed genes that were recovered completely (C) (for single-copy (S) and  
 157 duplicated (D) genes), that were only partially recovered (F for “Fragmented”), or that were  
 158 missing (M). The total number of orthologous groups of genes (n) that was searched in  
 159 BUSCO is also indicated.

160

161 **Data availability**

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

167

## 168 **Acknowledgments**

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

174

## 175 **Supplementary data**

176 Supplementary data to this article can be found at xxxx.

177

## 178 **References**

- 179 Altschul, S.F., Madden, T.L., Schäffer, A.A., Zheng Zhang J.Z., Miller, W. and Lipman, D.J.  
180 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search  
181 programs. *Nucleic Acids Research* 25:3389-3402.
- 182 Bolger, A.M., Lohse, M., Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina  
183 sequence data. *Bioinformatics* 30, 2114-2120
- 184 Breitwieser, M., Viricel, A., Graber, M., Murillo, L., Becquet, V., Churlaud, C., Fruitier-  
185 Arnaudin, I., Huet, V., Lacroix, C., Pante, E., Le Floch, S., Thomas-Guyon, H. 2016.

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

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

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

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

198 Finn, R.D., Attwood, T.K., Babbitt, P.C., Bateman, A., Bork, P., Bridge, A.J., Chang, H-Y.,  
199 Dosztányi, Z., El-Gebali, S., Fraser, M., Gough, J., Haft, D., Holliday, G.L., Huang, H.,  
200 Huang, X., Letunic, I., Lopez, R., Lu, S., Marchler-Bauer, A., Mi, H., Mistry, J., Natale,  
201 D.A., Necci, M., Nuka, G., Orengo, C.A., Park, Y., Pesseat, S., Piovesan, D., Potter, S.C.,  
202 Rawlings, N.D., Redaschi, N., Richardson, L., Rivoire, C., Sangrador-Vegas, A., Sigrist,  
203 C., Sillitoe, I., Smithers, B., Squizzato, S., Sutton, G., Thanki, N., Thomas, P.D., Tosatto,  
204 S.C.E., Wu, C.H., Xenarios, I., Yeh, L-S., Young, S-Y, Mitchell, A.L. 2017. InterPro in  
205 2017—beyond protein family and domain annotations. Nucleic Acids Research 45, D190–  
206 D199

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

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

212 Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N.,  
213 Regev, A. 2011. Full-length transcriptome assembly from RNA-se data without a reference  
214 genome. *Nat Biotechnol.* 29, 644-652

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

217 Honaas, L.A., Wafula, E.K., Wickett, N.J., Der, J.P., Zhang, Y., Edger, P.P., Altman, N.S.,  
218 Pires, J.C., Leebens-Mack, J.H., dePamphilis C.W. 2016. Selecting superior de novo  
219 transcriptome assemblies : lessons learned by leveraging the best plant genome. *PLoS*  
220 *ONE* 11(1), e0146062

221 Klemetsen, T., Raknes, I.A., Fu, J., Agafonov, A., Balasundaram, S.V., Tartari, G.,  
222 Robertsen, E., Willassen, N.P. 2017. The MAR databases: development and  
223 implementation of databases specific for marine metagenomics. *Nucl. Acids. Res.*, gkx1036

224 Liu, H., Zheng, H., Zhang, H., Deng, L., Liu, W., Wang, S., Meng, F., Wang, Y., Guo, Z., Li,  
225 S., Zhang, G. 2015. A de novo transcriptome of the noble scallop, *Chlamys nobilis*,  
226 focusing on mining transcripts for carotenoid-based coloration. *BMC Genomics* 16:44

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

229 Meng, X-L., Liu, M., Jiang, K-y., Wang, B-j., Tian, X., Sun, S-j., Luo, Z-y., Qiu, C-w.,  
230 Wang, L. 2013. De novo characterization of Japanese scallop *Mizuhopecten yessoensis*  
231 transcriptome and analysis of its gene expression following cadmium exposure. *PLoS*  
232 *ONE*, 8, e64485

233 Milinkovitch, T., Bustamante, P., Huet, V., Reigner, A., Churlaud, C., Thomas-Guyon, H.  
234 2015. *In situ* evaluation of oxidative stress and immunological parameters as  
235 ecotoxicological biomarkers in a novel sentinel species (*Mimachlamys varia*). *Aquatic*  
236 *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., Scoullou, 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.