De novo transcriptome assembly of the Qatari pearl oyster Pinctada imbricata radiata

De novo transcriptome assembly of the Qatari pearl oyster *Pinctada imbricata radiata*

Tim P. Bean[1], Zenaba Khatir[2], Brett P. Lyons[3], Ronny van Aerle[3], Diana Minardi[3], John P. Bignell[3], David Smyth[2,4], Bruno Welter Giraldes[2] and Alexandra Leitão[2]

Corresponding author - tim.bean@roslin.ed.ac.uk

[1]The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian, UK, EH25 9RG

[2]Environmental Science Center (ESC), Qatar University, P. O. Box: 2713 Doha, Qatar

[3]Cefas, Barrack Road, Weymouth, Dorset UK, DT4 8UB

[4]School of Ocean Science, Bangor University, Wales, UK, LL59 5AB

Abstract

The pearl oyster *Pinctada imbricata radiata* is an iconic species in Qatar, representing an integral part of the nation's cultural heritage and one of the main economic foundations upon which the nation developed. During the early part of the 20th century, nearly half of Qatar population was involved in the pearl oyster industry. However, the fishery has undergone steady decline since the 1930s, and the species is now under threat due to multiple confounding pressures. This manuscript presents the first *de novo* transcriptome of the Qatari pearl oyster assembled into 30,739 non-redundant coding sequences and with a BUSCO completeness score of 98.4%. Analysis of the transcriptome reveals the close evolutionary distance to the conspecific animal *Pinctada imbricata fucata* but also highlights differences in immune genes and the presence of distinctive transposon families, suggesting recent adaptive divergence. This data is made available for all to utilise in future studies on the species.

Prepared for submission to Marine Genomics

Working Title: *Pinctada imbricata radiata* transcriptome

Keywords: bivalve; transcriptome; pearl oyster; Pinctada; Arabian-Persian Gulf

Author input into the manuscript

TPB - conceived and designed the study, performed the field sampling and laboratory sample processing, analysed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, authored and reviewed drafts of the paper, approved the final draft.

ZK - performed the field sampling and laboratory sample processing, reviewed drafts of the paper, approved the final draft.

BPL - conceived and designed the study, performed the field sampling, reviewed drafts of the paper, approved the final draft.

RvA – Contributed to the bioinformatics analysis, reviewed drafts of the paper, approved the final draft.

DM - laboratory sample processing, reviewed drafts of the paper, approved the final draft.

DS - performed the field sampling and laboratory sample processing, reviewed drafts of the paper, approved the final draft.

BWG - performed the field sampling and laboratory sample processing, reviewed drafts of the paper, approved the final draft.

JPB - performed laboratory sample processing and assessment, approved the final draft.

AL - conceived and designed the study, performed the field sampling, laboratory sample processing, contributed reagents/materials/analysis tools, authored and reviewed drafts of the paper, approved the final draft.

Introduction

Qatar is located in the Arabian-Persian Gulf (hereafter referred to as The Gulf), a semi-enclosed sea characterized by a weak hydrodynamic flushing, high evaporation rates and low rates of discharge (Sheppard et al, 2010). Due to its naturally arid conditions, the Gulf is a challenging marine environment (Riegl et al 2012; Camp et al 2018), presenting wide variations in sea temperatures (annual range of 14-36 °C) and high salinities all year round (reaching 70 psu in the Gulf of Salwah) (Riegl et al 2012; Sheppard et al 2010). These conditions have produced several distinctive ecosystems; such as the characteristic Pearl oyster beds generated by the pearl oyster *Pinctada imbricata radiata* (Leach, 1814 and *e.g.* Smyth et al 2016). During the early part of the 20th century, nearly half of Qatar population was involved in the pearl oyster industry, which at present-day prices would have been worth an estimated $2.5 billion per annum to the nation's economy (Carter 2005). However, as Qatar has prospered and developed there has been a decline in the historical *P. i. radiata* oyster beds (Smyth et al 2016). A survey in 2014 showed that only one out of the five studied sites in Qatar could still be characterised as oyster dominant, even though all five sites previously corresponded to places of highly productive oyster fishery (Smyth et al 2016). These results highlighted the overwhelming likelihood that a combination of anthropogenic effects (such as overfishing, water quality shifts and petrochemical industry operations) have had a negative impact on the traditional Qatari pearl oyster beds (Smyth et al 2016, Al-Maslamani et al 2018).

P. i. radiata native range goes from the Indian Ocean to the Western Atlantic, including the Arabian Gulf and the Red Sea (Cunha et al. 2011). *P. i. radiata* is found as non-native species in Australasia and Japan and it is currently classified as one of the most successful invasive species in the Mediterranean Sea, likely due to the opening of the Suez Canal in 1869 (Hume, 2009), which links the Red Sea to the Mediterranean Sea. *P. i. radiata* was first reported in the Mediterranean Sea off the Egyptian coast in 1874 (Monterosato, 1878), making it one of the earliest Lessepsian invaders. Its success as an invasive species is in part due to its inherent plasticity, which has been used to cope against the substantial influence of natural and anthropogenic stressors present in the Qatar peninsula (Sheppard et al 2010; Ibrahim et al 2018). However, a recent study suggests that overall health of *P. i. radiata* beds is currently at a low ebb and not able to (Smyth et al 2016).

The taxonomic status of the three commercially relevant pearl producing species *Pinctada imbricata*, *Pinctada fucata*, and *Pincatada radiata* remains unresolved and has become a contentious issue in the literature (Wada and Tëmkin 2008). Recent studies based on DNA sequence analysis have shown very low levels of divergence among the three species and, in pair-wise comparison cases, the levels of divergence were comparable to conspecific individuals of other *Pinctada* species (Tëmkin

2010). Based on these data and the lack of diagnostic morphological features, the three species have been classified as subspecies of the senior synonym *P. imbricata* (*Pinctada imbricata imbricata*, *P. imbricata fucata*, and *P. imbricata radiata*) (Tëmkin 2010). This classification is followed in the present study.

Improvement in next generation sequencing technologies has led to a huge increase in the availability of genomic information, with large scale genomics and transcriptomics projects becoming commonplace. *P. i. radiata* is currently underrepresented in public databases such as NCBI (https://www.ncbi.nlm.nih.gov). In the present study we have sequenced, analysed, and publicly released a comprehensive atlas of expressed mRNA from *P. i. radiata*. This is a new resource available for detailed functional or comparative analysis of this species. In addition, as examples of the studies that can be undertaken with this data we have briefly characterised differences in genes expression between tissue types and compared genes found in *P. i. radiata* to other bivalve species.

Methods

Sampling

The Pearl oyster for this study was hand dived from a site at Al Wakra off the Qatar coast on 17/04/2017 (N 25°09.150, E 51°37.072). This is a site with moderate pollution in close proximity to Doha, the most populous city in Qatar. Live sample was transported directly to the laboratory and stored in seawater from the sampling site overnight at ambient temperature to allow depuration of the digestive tract before individual tissues (digestive gland, gill, adductor muscle, gonad, and mantle) were dissected (Figure 1). The oyster was split in two, with roughly half section stored in RNAlater at -20 °C until subsequent RNA extraction, and the reaming half section fixed in Davidson's seawater fixative for 24 hours, before being changed to 70% ethanol and then processed for formalin fixed paraffin embedded (FFPE) histological assessment using haematoxylin and eosin staining.

RNA extraction and sequencing

Total RNA was extracted from five *P. i. radiata* tissues (digestive gland, gills, adductor muscle, gonad, and mantle) with Ribozol (AMRESCO VWR, USA) using an adapted manufacturer protocol: 50 to 100 mg of individual tissues were homogenised in 1 ml of Ribozol in Lysing Matrix A FastPrep® tubes with a Fast Prep cell disrupter (1 min at 5 ms^{-1}) (MPBio, UK); samples were centrifuged at 12,000 g for 10 min at 4 °C, supernatant transferred in fresh tube and incubated at room temperature for 5 min. 200 µl of chloroform was added to each sample, vortexed for 15 sec, incubated at room temperature for 2 min, and centrifuged at 12,000 g for 15 min at 4 °C. Aqueous phase was removed and total RNA precipitated by adding 500 µl of isopropanol, followed by incubation at room temperature for 10 min,

and centrifugation 12,000 g for 10 min at 4 °C. Pellet was subsequently washed in 1 ml of 75 % ethanol, centrifuged at 7,500 g for 5 min at 4 °C, air dried for 5 minutes and dissolved in 50 µl RNase free water. Quality was checked by TapeStation (Agilent, USA) with RINs of 7, 9.3, 8.8, 9.3 and 9.2 recorded for digestive gland, gill, adductor muscle, gonad, and mantle respectively (note low value for digestive gland appeared to be due to an elevated concentration of RNA rather than low integrity). Libraries were prepared and sequenced by the University of Exeter Sequencing Service. Briefly, libraries were produced with Illumina Truseq stranded mRNA kit (Illumina, USA), QA checked by TapeStation (Agilent, USA) and Quantus Fluorometer (Promega, UK), pooled in equimolar concentrations, and sequenced as 125 bp paired-end reads on one lane of an Illumina HiSeq 2500 system in standard mode.

Bioinformatics

Sequences were trimmed by trim_galore (version 0.4.0) for paired sequences, but with –retain_unparied and –fastqc arguments included, and then used for all subsequent analysis. To produce the transcriptome, all sequences were pooled and normalised with bbnorm, including the pre-filter parameter, before being assembled by Trinity (version 2.8.4) with default parameters except for inclusion of the --no_normalize_reads parameter (Haas et al 2013). The transcriptome was run against the NCBI nr protein database (07/09/2018) using the blastx feature in diamond (version 0.9.22) including the –sensitive parameter an e-value cut off of 0.001 and arguments to increase speed on a high memory server (--index-chunks 1 and --block-size 10). Results visualised in MEGAN (version 6.5.8). Reads identified as metazoan were used for all further analysis, while non-metazoan reads were discarded. The transcriptome (isoform sequences) was loaded into Blast2GO with associated Blast results for annotation. The gene expression matrix was calculated by RSEM using the dispersion index of 0.4, which was deemed most appropriate for no replicate reads. Differential expression was calculated by edgeR using two different methods. A standard matrix was used for tissue to tissue comparison, whereas for Gene Set Enrichment Analysis (GSEA), each tissue was analysed in comparison to all other tissues as replicates. Differentially expressed genes (FDR > 0.01, fold change >2) were used for GSEA analysis within Blast2GO software. Tissues were furthermore analysed for uniquely expressed genes; any genes in any tissue with FPKM > 0.25 was selected as expressed, and visualised within the package VennDiagram (version 1.6.2.0). Transdecoder (version 5.5.0) was used to identify probable open reading frames and redundancy removed with cd-hit (Version 4.8.1, Li et al 2006, Fu et al 2012) using an identity threshold of 0.9 resulting in 30739 non-redundant coding sequences (Supplementary Figure 1). These sequences were run against metazoan databases with benchmarking single-copy orthologs software (BUSCO, version 3.0.2) to check the completeness of the transcriptome (Waterhouse et al., 2017).

Orthology analysis

Orthologous gene groups were assigned to *P. i. radiata* and five other species using OrthoFinder software (Emms et al., 2015 and 2018). Briefly, total non-redundant protein sequence files for *P. i. fucata* (Takeuchi et al., 2016), *Crassostrea gigas* (Zhang et al., 2012), *C. virginica* (https://www.ncbi.nlm.nih.gov), *M. yesso* (Wang et al., 2017) and *Octopus bimaculoides* (Albertin et al., 2015) were downloaded from publically available databases. These sequences were run through Orthofinder (version 2.3.3), using default parameters, alongside the 30739 non-redundant *P. i. radiata* protein sequences output from Transdecoder (see above and Supplementary data 1).

Results

Gross morphology and histopathology

The oyster chosen for sequencing had height of 53 mm (anterior to posterior), a width of 17 mm (maximum distance from left to right valve) and a total wet-weight of 62 g. No pearls or notable morphologies were observed. Pathology samples were examined histologically as in previous studies (Ward *et al*., 2006; Hines *et al*., 2007). The oyster individual sequenced was a female, with a developed gonad. One instance of an unknown trematode with granuloma was observed. No other notable pathologies were observed.

Sequencing and transcriptome Assembly

After trimming, over 45 Gb of data were available for further analysis (Table 1). The transcriptome assembled into 179,599 contigs with max, min, and average lengths of 16,371 bp, 201 bp, and 1,119 bp respectively (Table 2). N50 and N90 were 2,013 bp and 430 bp respectively (Table 2). Trinity assigned these transcripts into 24,676 gene clusters (Table 2). Of the transcripts, 70,114 mapped to a sequence record associated with a cellular organism in the NCBI nr reference database (https://www.ncbi.nlm.nih.gov), of these there were sequences for 68,930 Metazoa; 60,285 Protostomia; 58,378 Lophotrochozoa; 55,935 Mollusca and 49,835 Bivalvia. On a species level, most of the bivalve sequences mapped to rock oysters, for which two species have whole genome sequences available within the NCBI nr reference database, with just under 2,000 reads mapping to *Pinctada* species. Twenty reads mapped to the common molluscan parasite phylum Platyhelminthes. After removal of redundant ORFs, Transcoder and cd-hit returned 30,739 non-redundant, expressed sequences, or hypothetical proteins. BUSCO checks on the overall completeness of both redundant and non-redundant set of coding sequences found 957 complete BUSCOs and 5 fragmented BUSCOs in both sequence sets, with an overall completeness score of 98.4%.

Gene expression and gene set enrichment analysis

Gene expression was analysed quantitatively and qualitatively. Expression of 13,657 genes was shared by all tissues. Each tissue had uniquely expressed genes with the digestive gland having the most unique transcripts 730, compared to 471 in the mantle, 401 in the gill, 145 in the gonad, and just 80 in the adductor muscle. The pattern of differentially expressed transcripts was distinct, with the digestive gland having 3,972 differentially expressed transcripts (compared to all other tissues), adductor muscle 3,649, gill 2,208, gonad 1,835, and mantle 1,466. Gene Set Enrichment Analysis (GSEA) was performed on these differentially expressed genes from each of the tissue types (assessed against all other tissues). Table 3 shows how each tissue type was enriched for several gene ontology (GO) terms which relate to the function of that particular tissue (for example contractile fiber in the adductor muscle). Digestive gland, which had the highest number of differentially and uniquely expressed genes, also showed the highest number of enrichment terms. Functions, including endopeptidase and peptidase inhibitor activity, were enriched in differentially expressed genes, suggesting function associated with the process of digestion.

OrthoFinder

In brief, 20,870 orthogroups were assigned to the six different species and included 83.5 % of the total number of genes (Supplementary Data 2). Only 1.6 % of the orthogroups were species-specific. As expected, the all-gene phylogeny grouped the *Pinctada* species together and the *Crassostrea* species together, with the scallop *M. yesso* completing the bivalve clade and the cephalopod mollusc *O. bimaculoides* being most distance to all other species. The two *Crassostrea* species and the *Pinctada* species each had around 16,000 orthogroups and each genus shared similar numbers of orthogroup overlaps. As a lone and more distant species, *O. bimaculoides* only had around 9,000 orthogroups but shared the majority of these with all other species sequenced. Orthogroups containing at least five genes in *P. i. radiata* and zero genes from *P. i. fucata* (Supplementary data 3) were analysed in Blast2GO. These 24 gene groups included functional groups such as transposable elements, transcription factors and immune system receptors (data not shown).

Discussion

All metrics suggest that the overall quality of *P. i. radiata* transcriptome produced in the present study is very high, with a similar number of non-redundant expressed sequences close to the highly related species *P. i. fucata* (Du et al., 2017), which has had its whole genome sequenced, and with benchmarking single-copy ortholog (BUSCO) analysis giving a score of 98.4% completeness. The N50 score of 2,013 bp, which is within the region of most complete bivalve transcriptomes (*e.g.* Ryu et al

2019, Viricel 2018, Patnaik et al 2016). This sequence data is now available online via public databases with relevant details available in Table 4.

Orthogroups, representative of groups of homologous genes, are a useful way of inferring and comparing functional biology of multiple species, and also identifying shared genes with which multigene phylogenies can be drawn (Figure 2A) (Emms et al 2015). In order to assess the comparative differences between *P. i. radiata* and other molluscs, non-redundant protein databases of four other bivalves (*P. i. fucata , C. virginica, C. gigas and M. yesso*) and one cephalopod mollusc (*O. bimaculoides*) were analysed with OrthoFinder (Emms et al 2015) (Figure 2). In total circa 20,000 orthogroups were identified across the five different species, 83.5 % of which spanned across more than one species. As expected*, P. i. radiata* groups closely to and shares a high proportion of its orthogroups with *P. i. fucata*. Interestingly, the multi-gene phylogeny inferred within OrthoFinder suggests the distance between *P. i. radiata* and *P. i. fucata* is similar to that between *C. gigas* and *C. virginica*. In addition, the two *Pinctada* species compared in this study share roughly the same number of orthogroups as the two *Crassostrea* species (14,801 vs 15,308 respectively). Together, these data suggest a similar level of phylogenetic relationship between the *Pinctada*, rather than a conspecific relationship. However, it should be noted that less than 0.6 % of the *Pinctada* genes reside in species specific orthogroups, compared to over 1 % of the *Crassostrea* genes, suggesting that the two *Crassostrea* species have several more divergent orthogroups, in addition to the large number of shared groups. The cephalopod (*O. bimaculoides*) only had genes assigned to around 9000 orthogroups, but it shared the majority of these with all other species sequenced. This finding includes the orthogroups which are present across mollusca, and likely includes genes with many essential functions, rather than those evolved for lineage specific functions. It will be interesting to continue repeating this analysis with more bivalve transcriptomes and genomes as they continue to become available and identify the groups of genes specific to each class of mollusc, and to further study those genes which allowed such successful adaptive radiation of the molluscs (Seed 1983). In order to elucidate some of the functional differences between *P. i. radiata* and *P. i. fucata* genomes, orthogroups which contained at least five genes from *P. i. radiata* and none from *P. i. fucata* were studied in more detail. Among this set of 24 orthogroups were genes with homology to transposable elements, transcription factors and innate immune signalling. The function of these genes suggests they have evolved in relation to specific pressures, which may underlie some of the more recent lineage specific adaptations. In general, the relationship between *P. i. radiata* and *P. i. fucata* has proven to be challenging to resolve both from a morphological and genetic point of view (Tëmkin 2010), with the current taxonomic sub-species designation being somewhat of a compromise. The analysis presented in this study, however, suggests that the current designation could someday be re-visited with a thorough genome-wide analysis.

In order to demonstrate the potential of this transcriptome for study of functional properties of the differentially expressed transcripts from each organ, gene set enrichment analysis (GSEA) was utilised (Table 3). Most organs were enriched for categories highly symbolic of the overall function, for example the adductor muscle was enriched for myosin complex, contractile fibers, myofibril, sarcomere and actin cytoskeleton, all of which are associated with muscle contraction. The mantle, perhaps the most bivalve-specific tissue, was enriched for chitin binding and metabolism, glucosamine containing processes, aminoglycan and amino sugar metabolic processes. These findings point towards the key function of shell formation, with chitin metabolism being previously identified as a basic component of nacre in *P. fucata martensii* (Du et al., 2017) and the amino glycan and amino sugar pathways previously identified as enriched protein components in *C. gigas* shell (Wang et al 2013), likely to be involved in formation of complex matrices. Otherwise, the digestive gland was enriched for categories associated with peptidase regulation, the gonad, enriched for categories including nucleoplasm, biosynthesis and protein assembly suggesting active biosynthetic processes, such as gonadongenesis. Enrichment categories in the gill suggested the process of post-translational modification via dephosphorylation, which may, for example, play a key role in regulation of ion-transport across the membrane of the gill (Lucena et al., 2017).

This species has demonstrated an incredible ability to survive a range of challenging conditions, but it appears that is now reaching the limit of this inherent flexibility. The existence of this high-quality reference transcriptome will now allow for transcriptomic studies into the ability of *P. i. radiata* to survive challenging conditions.

Acknowledgments

This publication was made possible by the NPRP award [NPRP9-394-1-090 “The Pearl Oyster: from national icon to guardian of Qatar's marine environment”] from the Qatar National Research Fund (a member of Qatar Foundation). The findings herein reflect the work, and are solely the responsibility, of the authors. In addition, the authors would like to thank Prof. Eduarda Santos for advice regarding sequencing requirements and Mr. Mark Chatting and Mr. Reyniel Gasang for their support in the collection of the samples.

References

Alagarswami, K., & Victor, A. C. C. (1976). Salinity tolerance and rate of filtration of the pearl oyster Pinctada fucata. Journal of the Marine Biological Association of India, 18 (1), 149-158.

Albertin, C.B., Simakov, O., Mitros, T., Wang, Z.Y., Pungor, J.R., Edsinger-Gonzales, E., Brenner, S., Ragsdale, C.W. and Rokhsar, D.S.,(2015). The octopus genome and the evolution of cephalopod neural and morphological novelties. Nature, 524 (7564), p.220.

Al-Maslamani, I., Smyth, D., Giraldes, B., Chatting, M., Al-Mohannadi, M. and Le Vay, L. (2018). Decline in oyster populations in traditional fishing grounds; is habitat damage by static fishing gear a contributory factor in ecosystem degradation?. Journal of Sea Research, 140, 40-51.

Almatar, S.M., Carpenter, K.E., Jackson, R., Alhazeem, S.H., Alsaffar, A.H., Ghaffar, A.R.A. and Carpenter, C. (1993). Observations on the pearl oyster fishery of Kuwait. Journal of Shellfish research, 12 (1), 35-40.

Camp, E.F., Schoepf, V., Mumby, P.J., Hardtke, L.A., Rodolfo-Metalpa, R., Smith, D.J. and Suggett, D.J. (2018). The future of coral reefs subject to rapid climate change: lessons from natural extreme environments. Frontiers in Marine Science, 5:4. doi: 10.3389/fmars.2018.00004

Carter, R., 2005. The history and prehistory of pearling in the Persian Gulf. J. Econ. Soc. Hist. Orient. 48 (2), 139-209.

Cunha, R.L., Blanc, F., Bonhomme, F. and Arnaud-Haond, S. (2011). Evolutionary patterns in pearl oysters of the genus Pinctada (Bivalvia: Pteriidae). Marine Biotechnology, 13 (2), 181-192.

de Mora, S., Tolosa, I., Fowler, S. W., Villeneuve, J-P., Cassi, R., Cattini, C. (2010). Distribution of petroleum hydrocarbons and organochlorinated contaminants in marine biota and coastal sediments from the ROPME sea area during 2005. Marine Pollution Bulletin. 60, 2323-2349. https://doi.org/10.1016/j.marpolbul.2010.09.021

Emms, D.M. and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome biology, 16:157. DOI 10.1186/s13059-015-0721-2

Emms, D.M. and Kelly, S. (2018). OrthoFinder2: fast and accurate phylogenomic orthology analysis from gene sequences. *BioRxiv*, p.466201. doi: https://doi.org/10.1101/466201

Freije, A. M. (2015). Heavy metal, trace element and petroleum hydrocarbon pollution in the Arabian Gulf. Journal of the Association of Arab Universities for Basic and Applied Sciences, 17(1), 90-100. http://dx.doi.org/10.1016/j.jaubas.2014.02.001

Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics, 28 (23), 3150-3152.

Galil, B., Zenetos, A., (2002). A sea change – Exotics in the Eastern Mediterranean Sea. In: Leppäkoski E., Gollasch S., Olenin S. (eds) Invasive Aquatic Species of Europe. Distribution, Impacts and Management. Springer, Dordrecht.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M. and MacManes, M.D. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols, 8(8), p.1494.

Hines, A., Oladiran, G.S., Bignell, J.P., Stentiford, G.D., Viant, M.R. (2007). Direct Sampling of Organisms from the Field and Knowledge of their Phenotype: Key Recommendations for Environmental Metabolomics. Environmental Science and Technology, 41 (9), 3375-3381.

Hulme, P.E. ed., 2009. *Handbook of alien species in Europe* (Vol. 569). Dordrecht, The Netherlands: Springer.

Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658-1659.

Lucena, M.N., Garçon, D.P., Fontes, C.F.L., McNamara, J.C. and Leone, F.A. (2017). Polyamines regulate phosphorylation–dephosphorylation kinetics in a crustacean gill (Na+, K+)-ATPase. *Molecular and cellular biochemistry*, *429*(1-2), 187-198.

Monterosato T.A. (1878). Enumerazione e sinonimia delle conchiglie Mediterranee. Giornale di scienze naturali ed economiche di Palermo, 13: 61-113.

Patnaik, B.B., Wang, T.H., Kang, S.W., Hwang, H.J., Park, S.Y., Park, E.B., Chung, J.M., Song, D.K., Kim, C., Kim, S. and Lee, J.S. (2016). Sequencing, de novo assembly, and annotation of the transcriptome of the endangered freshwater pearl bivalve, Cristaria plicata, provides novel insights into functional genes and marker discovery. *PLoS One*, *11*(2), p.e0148622.

Rabaoui, L., Tlig-Zouari, S., Cosentino, A., Hassine, O. (2009). Associated fauna of the fan shell *Pinna nobilis* (Mollusca: Bivalvia) in the northern and eastern Tunisian coasts. Scientia Marina, 73(1), 129-141.

Riegl, B.M. and Purkis, S.J. (2012). Coral reefs of the Gulf: adaptation to climatic extremes in the world's hottest sea. In Coral reefs of the Gulf (pp. 1-4). Springer, Dordrecht.

Ryu, T., Woo, S. and Lee, N. (2019). The first reference transcriptome assembly of the stalked barnacle, Neolepas marisindica, from the Onnuri Vent Field on the Central Indian Ridge. *Marine Genomics*.

Seed, R., (1983). Structural organization, adaptive radiation, and classification of molluscs. In Metabolic Biochemistry and Molecular Biomechanics (pp. 1-54). Academic Press.

Sheppard, C., Al-Husiani, M., Al-Jamali, F., Al-Yamani, F., Baldwin, R., et al., (2010). The Gulf: a young sea in decline. Marine Pollution Bulletin, 60, 13–38.

Smyth, D., Al-Maslamani, I., Chatting, M., Giraldes, B. (2016). Benthic surveys of the historic pearl oyster beds of Qatar reveal a dramatic ecological change. Marine Pollution Bulletin, 113(1-2), 147–155.

Takeuchi, T., Kawashima, T., Koyanagi, R., Gyoja, F., Tanaka, M., Ikuta, T., Shoguchi, E., Fujiwara, M., Shinzato, C., Hisata, K. and Fujie, M. (2012). Draft genome of the pearl oyster Pinctada fucata: a platform for understanding bivalve biology. DNA research, 19(2), 117-130.

Takeuchi, T., Koyanagi, R., Gyoja, F., Kanda, M., Hisata, K., Fujie, M., Goto, H., Yamasaki, S., Nagai, K., Morino, Y. and Miyamoto, H. (2016). Bivalve-specific gene expansion in the pearl oyster genome: implications of adaptation to a sessile lifestyle. Zoological letters, 2(1), p.3.

Tëmkin I. (2010). Molecular phylogeny of pearl oysters and their relatives (Mollusca, Bivalvia, Pterioidea). BMC Evolutionary Biology 10: 342.

Tlig-Zouari, S., Rabaoui, L., Irathni, I., Ben Hassine, O. (2009). Distribution, habitat and population densities of the invasive species *Pinctada radiata* (Molluca : Bivalvia) along the Northern and Eastern coasts of Tunisia. Cahiers de Biologie Marine 50(2): 131-142.

Verbruggen, B., Bickley, L.K., Santos, E.M., Tyler, C.R., Stentiford, G.D., Bateman, K.S. and van Aerle, R., (2015). De novo assembly of the Carcinus maenas transcriptome and characterization of innate immune system pathways. BMC genomics, 16(1), p.458.

Viricel, A., Becquet, V., Dubillot, E. and Pante, E. (2018). De novo assembly and functional annotation of the transcriptome of Mimachlamys varia, a bioindicator marine bivalve. *Marine genomics*, *41*, 42-45.

Wada KT, Tëmkin I. (2008). Taxonomy and phylogeny. In The Pearl oyster. Edited by: Southgate PC, Lucas JS. Amsterdam, Boston, Heidelberg, London, New York, Oxford, Paris, San Diego, San Francisco, Singapore, Sydney, Tokyo: Elsevier:37-75.

Wang, X., Li, L., Zhu, Y., Du, Y., Song, X., Chen, Y., Huang, R., Que, H., Fang, X. and Zhang, G. (2013). Oyster shell proteins originate from multiple organs and their probable transport pathway to the shell formation front. PloS one, 8(6), p.e66522.

Wang, S., Zhang, J., Jiao, W., Li, J., Xun, X., Sun, Y., Guo, X., Huan, P., Dong, B., Zhang, L. and Hu, X. (2017). Scallop genome provides insights into evolution of bilaterian karyotype and development. Nature ecology & evolution, 1(5), p.0120.

Ward, D.G., Wei, W., Cheng, Y., Billingham, L.J., Martin, A., Johnson, P.J., Lyons, B.P., Feist, S.W., Stentiford, G.D. (2006). Environmental Science and Technology, 40 (12), 4031-4036.

Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V. and Zdobnov, E.M. (2017). BUSCO applications from quality assessments to gene prediction and phylogenomics. Molecular biology and evolution, 35 (3), 543-548.

Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., Yang, P., Zhang, L., Wang, X., Qi, H. and Xiong, Z. (2012). The oyster genome reveals stress adaptation and complexity of shell formation. Nature, 490 (7418), 49-54. https://doi.org/10.1038/nature11413

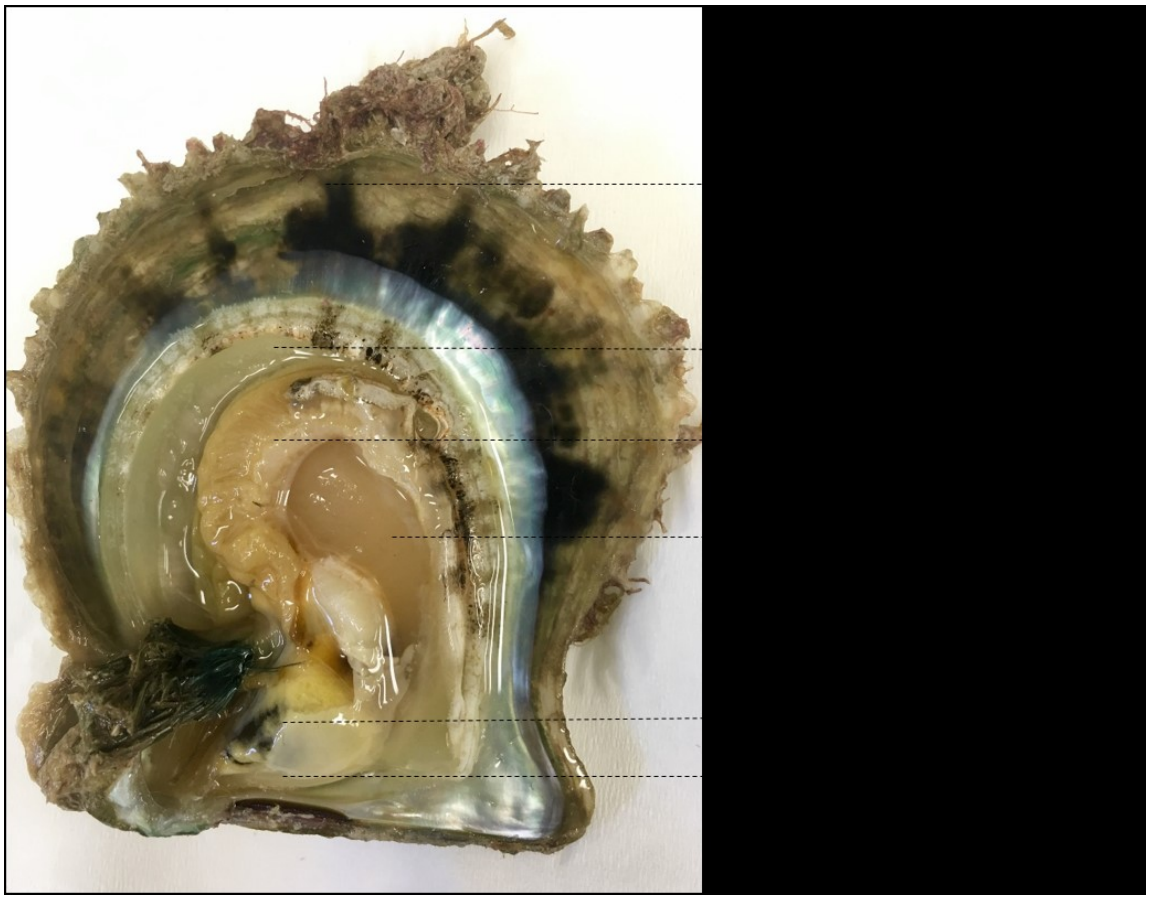

Figure 1. Left valve of the pearl oyster, *Pinctada imbricata radiata*, with tissues utilised in transcriptomic analysis identified.

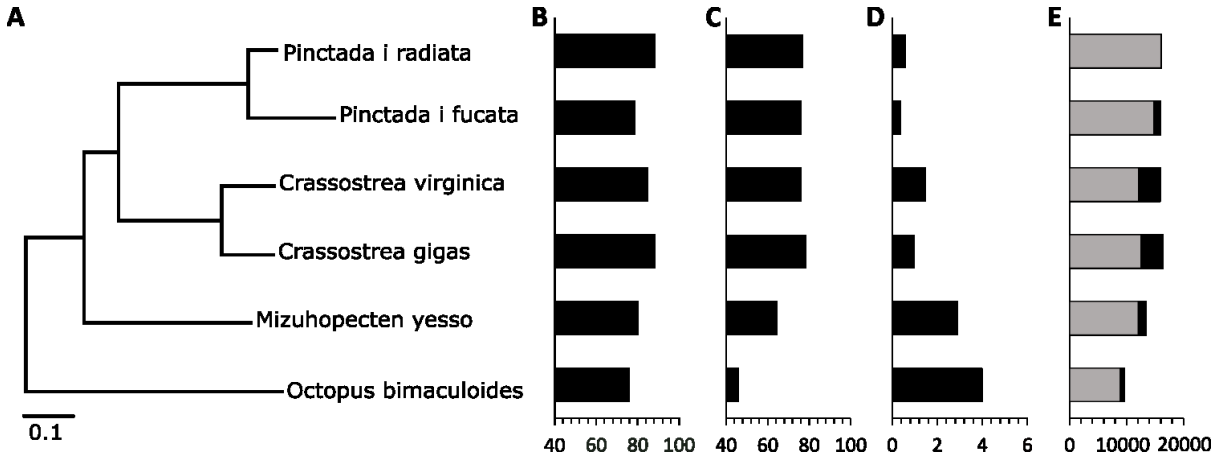

Figure 2. Results of phylogenomic orthology screens using OrthoFinder. A) Phylogeny of species as defined by all genes. B) Percentage of genes from each species in orthogroups. C) Percentage of orthogroups containing each species. D) Percentage of genes in species-specific orthogroups. E) Total number of orthogroups (black) which included genes from this species, and number of those shared with *Pinctada i radiata* (grey).

392 Table 1. Number and length of reads pre and post trimming via trimmomatic.

Tissue	Number of read pairs	Average length	Number of trimmed pairs	Average length	
				Forward	Reverse
Digestive gland	35,856,234	2x125	35,770,112	114.4	111.6
Gill	38,123,232	2x125	37,997,395	116.9	112.9
Adductor muscle	43,089,935	2x125	43,009,382	116.2	112.7
Gonad	39,136,760	2x125	39,044,852	115.5	112
Mantle	44,816,064	2x125	44,737,749	114.9	111.3

393

394

395 Table 2. Transcriptome statistics

Descriptive Statistic	Summary
Number of transcripts	179,599
Number of genes*	24,676
Total length (bp)	201,029,654
Shortest transcript length (bp)	201
Mean transcript length (bp)	1,119.30
Longest Transcript length (bp)	16,371
N50 (bp)	2,013

*gene cluster as identified with Trinity assembler (see methods)

396

397

398

399

400 Table 3. Top five (or all) categories for gene set enrichment analysis from each tissue.

Tissue	GO ID	GO Name	GO Category	Nominal p-val	FDR q-val
Digestive Gland	GO:0004866	endopeptidase inhibitor activity	Molecular Function	0	0
	GO:0061135	endopeptidase regulator activity	Molecular Function	0	0
	GO:0061134	peptidase regulator activity	Molecular Function	0	0
	GO:0004857	enzyme inhibitor activity	Molecular Function	0	0
	GO:0004867	serine-type endopeptidase inhibitor activity	Molecular Function	0	0
Gill	GO:0004721	phosphoprotein phosphatase activity	Molecular Function	5.941E-03	7.582E-02
	GO:0004725	protein tyrosine phosphatase activity	Molecular Function	7.905E-03	2.096E-01
	GO:0016311	dephosphorylation	Biological Process	7.937E-03	1.103E-01
	GO:0006570	tyrosine metabolic process	Biological Process	1.504E-02	8.915E-02
	GO:0006470	protein dephosphorylation	Biological Process	1.590E-02	7.120E-02
Adductor muscle	GO:0015629	actin cytoskeleton	Cellular Component	0	0
	GO:0016459	myosin complex	Cellular Component	0	0
	GO:0043292	contractile fiber	Cellular Component	0	0
	GO:0030016	myofibril	Cellular Component	0	0
	GO:0030017	sarcomere	Cellular Component	0	0
Gonad	GO:0034622	cellular protein-containing complex assembly	Biological Process	0.000E+00	3.651E-02
	GO:0005654	nucleoplasm	Cellular Component	7.937E-03	1.067E-01
	GO:0016053	organic acid biosynthetic process	Biological Process	1.235E-02	1.510E-01
	GO:0046394	carboxylic acid biosynthetic process	Biological Process	1.594E-02	1.517E-01
Mantle	GO:1901071	glucosamine-containing compound metabolic process	Biological Process	0	0
	GO:0006030	chitin metabolic process	Biological Process	0	0
	GO:0008061	chitin binding	Molecular Function	0	0
	GO:0006040	amino sugar metabolic process	Biological Process	0	0
	GO:0006022	aminoglycan metabolic process	Biological Process	0	0

401

402 Table 4. MIS specifications of the *P. i. radiata* transcriptome.

Item	Description
Investigation_type	Eukaryote
Project_name	Reference transcriptome of *Pinctada imbricata radiata*
Organism	*Pinctada imbricata radiata*
Classification	Metazoa (kingdom); Mollusca (phylum); Bivalvia (class); Pteriida (order); Pteriidae (family); Pinctada (genus)
Lat_lon	25°09.150 N 51°37.072 E
Geo_loc_name	Al Wakrah, Qatar
Collection_date	17/04/2018
Collector	Alexandra Leitão
Environment (biome)	marine benthic biome (ENVO:01000024)
Environment (feature)	sand (ENVO:01000017)
Environment (material)	sea water (ENVO:00002149)
Env_package	Water
Seq_meth	Illumina
Transcriptome_platform	HiSeq 2500
Assembly_method	Trinity v2.8.4
Submitted_to_INSDC	Bioproject ID: PRJDB8463 Biosample ID: SAMD00178207-SAMD00178211 Short read archive ID: DRA008674 Accession: ICPG01000001-ICPG01068930

403