Mar Biotechnol (2014) 16:34–45
DOI 10.1007/s10126-013-9516-1

ORIGINAL ARTICLE

Biomineral Proteins from *Mytilus edulis* Mantle Tissue Transcriptome

Andy Freer · Stephen Bridgett · Jiahong Jiang · Maggie Cusack

Received: 20 November 2012 / Accepted: 5 June 2013 / Published online: 5 July 2013

Abstract The common blue mussel, *Mytilus edulis*, has a bimineralic shell composed of approximately equal proportions of the two major polymorphs of calcium carbonate: calcite and aragonite. The exquisite biological control of polymorph production is the focus of research interest in terms of understanding the details of biomineralisation and the proteins involved in the process of complex shell formation. Recent advances in ease and availability of pyrosequencing and assembly have resulted in a sharp increase in transcriptome data for invertebrate biominerals. We have applied Roche 454 pyrosequencing technology to profile the transcriptome for the mantle tissue of the bivalve *M. edulis*. A comparison was made between the results of several assembly programs: Roche Newbler assembler versions 2.3, 2.5.2 and 2.6 and MIRA 3.2.1 and 3.4.0. The Newbler and MIRA assemblies were subsequently merged using the CAP3 assembler to give a higher consensus in alignments and a more accurate estimate of the true size of the *M. edulis* transcriptome. Comparison sequence searches show that the mantle transcripts for *M. edulis* encode putative proteins exhibiting sequence similarities with previously characterised shell proteins of other species of *Mytilus*, the Bivalvia *Pinctada* and haliotid gastropods. Importantly, this enhanced transcriptome has detected several transcripts that encode proteins with sequence similarity with previously described shell biomineral proteins including Shematrins and lysine-rich matrix proteins (KRMPs) not previously found in *Mytilus*.

Keywords Transcriptome · Bivalves · Biomineralisation · *Mytilus*

Electronic supplementary material The online version of this article (doi:10.1007/s10126-013-9516-1) contains supplementary material, which is available to authorized users.

A. Freer · J. Jiang
School of Chemistry, University of Glasgow
G12 8QQ, Glasgow, UK

S. Bridgett
The GenePool, School of Biological Sciences, University of Edinburgh, EH9 3JT, Edinburgh, Scotland, UK

M. Cusack (✉)
School of Geographical and Earth Sciences, University of Glasgow, G12 8QQ, Glasgow, UK
e-mail: Maggie.Cusack@glasgow.ac.uk

Introduction

The common blue mussel, *Mytilus edulis*, is an economically important species, with more than 9,000 tonnes of farmed mussels being produced per annum in UK waters alone (Burton et al. 2001). This commercial availability provides a ready source of specimens for study. With its shell composed of an outer layer of calcite prisms and inner layer of aragonite nacre (Fig. 1), *M. edulis* is an ideal model species for the study of polymorph formation. The attractive material properties of nacre (Jackson et al. 1988; Currey et al. 2001) and a detailed working model for nacre formation (Nudelman et al. 2006; Addadi et al. 2006; Nudelman et al. 2008; Cartwright and Checa 2007) are yet more reasons for the research interest in *M. edulis* from the biomineralisation perspective. Molluscs such as *M. edulis* could also be useful for monitoring the effect of environmental changes on marine life, such as ocean acidification (Clark et al. 2010).

Despite these many advantages, progress on a detailed understanding of biomineralisation is inhibited by the paucity of protein obtained directly from the mussel, which subsequently limits the investigation of the precise function of individual proteins. Even applying lab-on-a-chip technology to investigate protein function using the more abundant extrapallial (EP) proteins (Yin et al. 2009; Ji et al. 2010) quickly runs into the problem of too little protein available

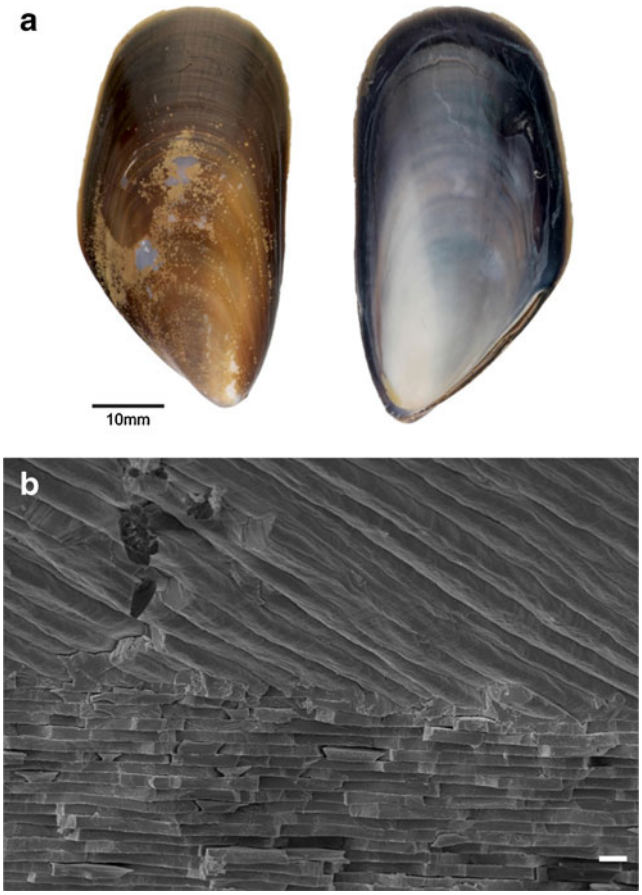

Fig. 1 Shell and constituent polymorphs of *M. edulis*. **a** Shell exterior (*left*) and interior (*right*). **b** Secondary electron image of fracture section of *M. edulis* shell showing the interface between outer calcite (*top*) and inner aragonite nacre (*bottom*). *Scale bar*=1 μm

for detailed analysis despite the small amount of protein required for this technique. Advances in the efficiency and ease of generating transcriptome data mean that the transcriptomes from the biomineralising mantle of several organisms have been analysed from a range of perspectives relevant to biomineralisation. Some examples include non-model species such as *Patella vulgata* (Werner et al. 2013), studying shell formation in the context of acidifying oceans (Clark et al. 2010) and heat stress (Truebano et al. 2010). Comparison has been made of proteins of nacre in gastropods and bivalves (Jackson et al. 2006; Jackson et al. 2010; Marie et al. 2010), of biomineral transcriptome and shell proteome in *Pinctada margaritifera* (Joubert et al. 2010) and of putative proteins encoded by nacreous and prismatic layer-producing tissues in *Pinctada fucata* (Kinoshita et al. 2011). Complementing these latest developments in genomic data acquisition is the ongoing isolation and characterisation of proteins from the shell itself, which covers a number of invertebrate species (Sarashina and Endo 2006; Zhang and Zhang 2006; Marin et al. 2008; Marie et al. 2010; Marie et al. 2011a; Marie et al. 2011b; Marie et al. 2011c; Bedouet et al. 2012; Marie et al. 2012).

In this study we generate the transcriptome from *M. edulis* mantle tissue as the source of proteins associated with biomineralisation. The assembly of this transcriptome for *M. edulis* adds to the growing sequence data base of the phylum Mollusca, which, along with current research on shell matrix and EP proteins, will allow us to continue to decipher the role and influence of particular proteins in biomineralisation.

Materials and Methods

M. edulis Specimens

Specimens of the common blue mussel, *M. edulis*, were obtained from a commercial source (Alan Beveridge, Glasgow, UK), so no specific permits were required. Extraction of total cellular RNA from the dissected mantle tissue of three specimens of locally sourced *M. edulis* was achieved using RNeasy Micro Kit (QIAGEN) according to the manufacturer's instructions. The Evrogen synthesis service was used to synthesise the cDNA using the SMART cDNA protocol (Zhu et al. 2001). This is similar to the Evrogen MINT synthesis kit in which a 3′-end CDS adapter containing an oligo(dT) sequence anneals to poly(A) stretches of RNA. A reverse transcriptase then synthesises a new first strand of cDNA, adding several non-template nucleotides at the new strand's 3′-end to incorporate a PlugOligo sequence into the cDNA. Finally, double-stranded DNA is amplified by polymerase chain reaction (PCR) using primers to the flanking CDS and PlugOligo adapters. This cDNA synthesis method requires significantly less input RNA than the random hexomer protocols and preferentially selects mRNA with polyA tails over other RNAs such as ribosomal RNA.

Pyrosequencing

The sequencing library was prepared in accordance with the Roche 454 titanium library preparation protocols (Roche 2009b) then sequenced on the 454-Flx titanium sequencing platform, and signal processing and base calling were performed, initially with the Roche gsRunProcessor version 2.3 then repeated later from the original images with the gsRunProcessor version 2.6 to obtain more reads and to improve the accuracy.

Assembly and Annotation

For assembly, we used the Roche Newbler assembler (Roche 2010), the MIRA assembler (Chevreux et al. 2004) and the CAP3 assembler (Huang and Madan 1999). For the Newbler assembly, the '-cdna' option was enabled for transcriptome assembly, and the '-vt' option was used to trim the SMART adapters from the reads. Initially, Newbler version 2.3 was used,

which produced 9,791 isotigs with a mean length of 772 bases and a total of 7 million bases. Newbler version 2.5.2 produced longer isotigs. For the MIRA 3.2.1 and 3.4.0.1 assemblies, the SMART adapters were trimmed and the options 'denovo, est, normal, 454 454_SETTINGS -CL:qc=no' were used. CAP3, with default settings, was used to merge the Newbler 2.5.2 and MIRA contigs to estimate the similarity of these two assemblies. The pre-release version of Newbler 2.5 performed best in a recent comparative experiment (Kumar and Blaxter 2010), and thus a final assembly with the latest Newbler 2.6 was performed (Table S1). Further details of the assembly commands are given in Appendix I of the "Electronic supplementary material".

Results

Pyrosequencing Summary

Using Version 2.3 of the Roche shotgun signal-processing pipeline, the Roche 454 titanium sequencing generated 385,856 reads, with an average length of 311 bases. Reads shorter than 40 bases were discarded. A total of 50 % of the bases are in reads of 407 or greater (Table 1).

Subsequent reprocessing of the sequencing images using version 2.6 generated additional reads (Table 1).

Assembly Comparisons

Since there is a limited amount of publicly available molluscan sequence data with which to validate our assemblies, we used several different assembly programs to compare assemblies, with the aim of obtaining the optimum assembly for initial annotation and future research. The reads were assembled using Newbler 2.3 (Roche 2009a), Newbler 2.5.2 (Roche 2010), MIRA 3.2.1 (Chevreux et al. 2004) and later Newbler 2.6. The statistics for the isotigs/contigs produced are given in Table S1. Initially, Newbler version 2.3 produced 9,791 isotigs with a mean length of 772 bases and a total of 7 million bases. Subsequently, a pre-released version of Newbler 2.5.2 generated 45,986 isotigs with a mean length of 518 bases and a total of 23 million bases. The assemblers were given all the reads, but those reads shorter than 20 bases were not used. Reads without significant alignment with other reads were not used in the assembly and flagged as singletons (Table S1). The MIRA assembly generated 45,966 contigs (called contigs in MIRA, rather than isotigs) with a mean length of 551 bases and a total of 25 million bases, which is similar to Newbler 2.5.2. The number of isotigs and bases generated by the Newbler 2.5.2 and MIRA assemblers are reasonable for an organism with a calculated genome C-value of 1.60 (www.genomesize.com).

Table 1 Comparison of statistics including read lengths, number of reads and number of bases for the raw 454 reads, with SMART adapters still attached, for Roche gsRunProcessor shotgun signal-processing pipeline versions 2.3 and 2.6. N50 values for versions 2.3 and 2.6, respectively, indicate that 50 % of the bases are in lengths of 407 and 398 bases or greater

Roche signal pipeline version	2.3	2.6
Minimum read length	40	40
Maximum read length	1,168	1,059
Mean read length	311.3	304.4
Standard deviation of read length	155.3	153.9
Median read length	324	315
N50 read length	407	398
Number of reads	385,856	494,391
Number of reads in N50	123,116	156,749
Number of bases in all reads	120,135,078	150,473,196
GC content of reads	34.21 %	34.02 %

The CAP3 co-assembly yielded 26,785 contigs with a mean length of 616 bases. Approximately 16,000 Newbler isotigs and MIRA contigs were not assembled at the second level of CAP3. This brief comparison of these three assemblers is echoed in the larger study of Kumar and Blaxter (2010). The GC content was consistently low for all assemblies in the range 33.3–34.3 % compared with the estimated Newbler 2.5.2 value of 42 % if all possible codons were used equally. The Newbler 2.5.2 assembly contained the largest number of assembled reads, and these Newbler 2.5.2 isotigs have been used for all subsequent annotation in this paper. The singletons probably not only contain some useful low-expression sequences but also likely contain a significant number of PCR artefacts and contaminants (Kumar and Blaxter 2010) and so have not been used in this analysis.

Data from the Newbler 2.6 assembler are presented in Table 1 and S1 for comparison. Version 2.6 generated 60,480 isotigs and 28,447,708 bases, which was consistent with the previous Newbler 2.5.2 and MIRA 3.2.1 assemblies. We expect this latest assembler to have increased accuracy as the number and percentage of aligned reads and bases have all increased.

These assembly metrics compare favourably with the assemblies obtained in other 454 mollusc sequencing projects, such as *Mytilus galloprovincialis* (Craft et al. 2010) with 8,586 contigs from 175,547 reads with Newbler 1.1 and *Laternula elliptica* (Clark et al. 2010) which had 18,290 contigs with an average size of 535 bp using Newbler with 264,289 reads. For our 454 reads, the sequencing generated an average phred (Ewing et al. 1998; Ewing and Green 1998) quality score per base of 29.6, averaged over 150,473,196 bases with a mean read length of 304 bases. For the assembled 60,000 Newbler 2.6 isotigs, the average phred quality score per base is 49.4, averaged over the 28,447,708 bases.

Gene Ontology

The annot8r script (Schmid and Blaxter 2008) was used in conjunction with the BlastX algorithm to search the EMBL UniProt database to assign subsets for Gene Ontology (GO), Kyoto Encyclopaedia of Genes and Genomes (KEGG) and Enzyme Commission (EC) annotations. A blast bitscore cutoff of 55 was used as suggested by the annot8r script prompts based on past testing by the annot8r developers. Histograms in Fig. 2 show the distributions for this gene ontology for biological processes, cellular components and molecular function based on the 1,486 isotigs (3.2 % of total) that had hits above the 55 bitscore cutoff. The majority of biological processes shared genes with the regulation of biological processes, metabolic processes and cellular processes. Molecular function annotation showed a dominance of binding functions, which is to be expected as the majority of genes expressed by the actively functioning mantle tissue could be involved in carbonate shell production. The cellular annotation showed gene sharing in the intracellular domain. This distribution is similar to that found in

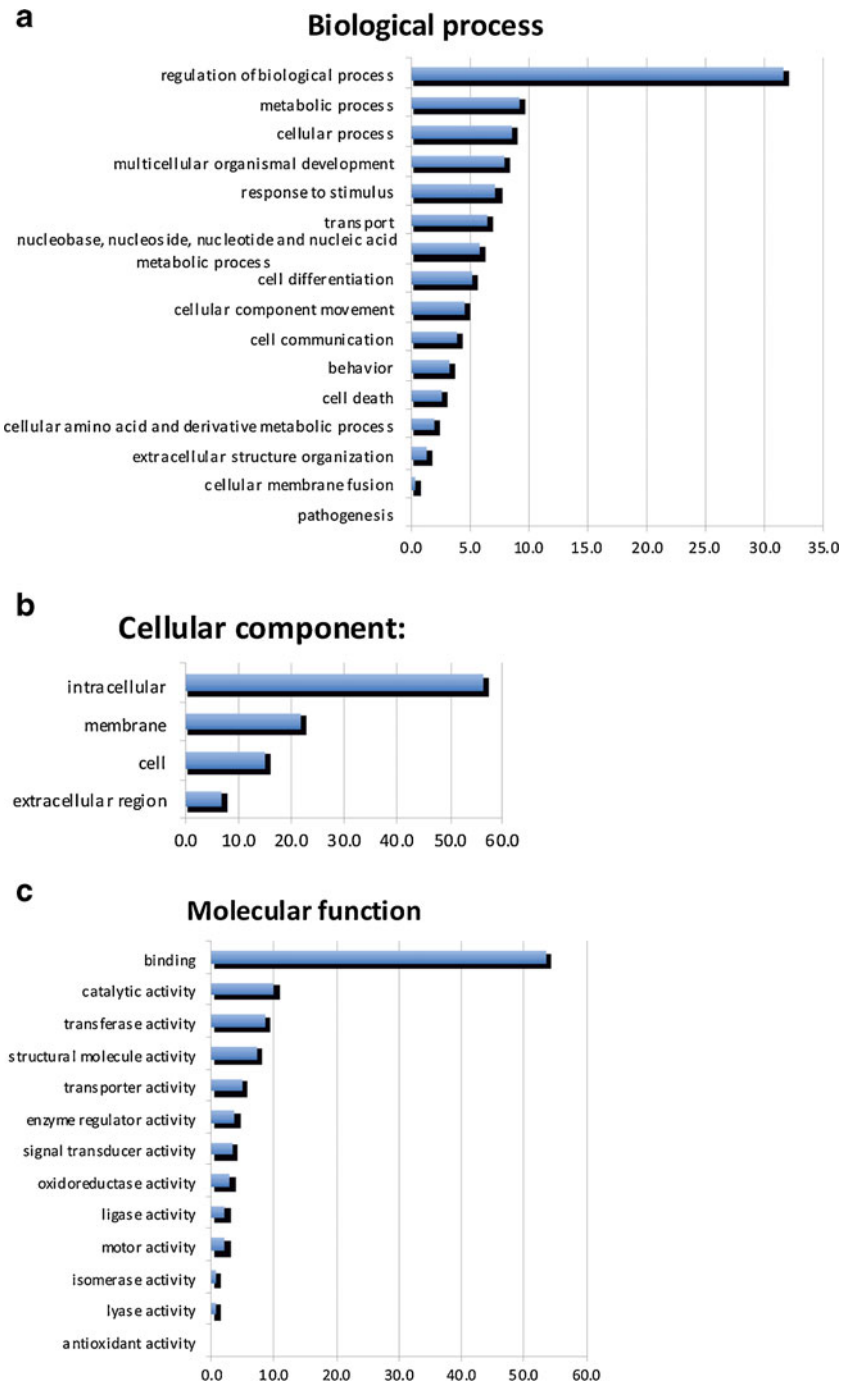

Fig. 2 GO distribution for unique sequences within *M. edulis* transcriptome. GO-slim terms are on the *y*-axis. Percentage distribution of genes shown as GO terms for **a** biological process, **b** cellular components and **c** molecular function

other marine biomineralising organisms, e.g. the bivalve, *P. margaritifera* (Joubert et al. 2010) and the bivalve *Pinctada maxima* and the gastropod *Haliotis asinina* (Jackson et al. 2010).

Exploring the Transcriptome for Biomineral Proteins

M. edulis was selected not only because of its economic importance but also because, from a biomineral perspective, it presents a bimineralic shell comprising both calcite and aragonite in almost equal proportions. This juxtaposition allows us to explore both sets of proteins involved in biomineral shell construction by exploiting this new transcriptome in relation to the growing knowledge of biomineralising proteins already characterised and now available in various data bases for both bivalves and gastropods.

The most abundant EP protein has been screened in terms of influence on carbonate polymorph formation using a novel microfluidics platform (Yin et al. 2009; Ji et al. 2010). Since the primary sequence of the most abundant EP protein is known (Hattan et al. 2001; Yin et al. 2005), this provides a good measure of the veracity of the generated transcriptome. In fact, the highest-scoring alignment of all of the proteins studied was indeed the most abundant EP protein from *M. edulis* [Q6UQ16] with a score of 486; e^{-138} with a completeness of 98 % for isotig 10720 over all 236 amino acids (aa) (Fig. S1). This score (486; e^{-138}) represents the bit score (integer value) and E-value. A higher bit score indicates a better match between the query and subject sequences. Each base that matches adds to the bit score, and each mismatch or gap reduces the bitscore. The more negative the exponent of the E-value (so the E-value approaches zero) the better the match, as the E-value is the number of matches that we expect to obtain by random chance.

Shell Matrix Proteins

Several publications discuss the evolution of disparate molluscan species within the context of shell matrix proteins (SMPs) (Jackson et al. 2006; Jackson et al. 2010; Marie et al. 2010; Marie et al. 2011a; Marie et al. 2011b; Marie et al. 2011c; Marin et al. 2008). Here we add to the debate by using this new *M. edulis* transcriptome to look firstly at the alignment of proteins from other *Mytilus* species, then between different Bivalvia, in particular proteins with repetitive low-complexity domains (RLCDs) and finally looking for alignment within the haliotid gastropods.

Mytilus Species

Using data from normalised cDNA libraries for four different bivalve species (Tanguy et al. 2008), Marie et al. (2011a) report nine novel *Mytilus* SMPs, of which three are completely new—*Mytilus* uncharacterised shell protein (MUSP)-1, MUSP-2 and MUSP-3. These three newly discovered proteins present an ideal opportunity to probe the transcriptome to see if they are found in *M. edulis*. MUSP-1 from *M. galloprovincialis* [P86853] showed 95 % identity (after the removal of the signal peptide) with a 131-aa fragment for isotig 16782 (662; $3e^{-70}$) (Fig. S2). However, both MUSP-2 [P86858] and MUSP-3 [P86859] from *Mytilus californianus* showed little sequence identity with associated poor scores. Although in this instance there was poor agreement with the *M. californianus* MUSP-2 and −3 proteins, the insoluble Ala-Gly-rich nacre-specific silk fibroin MSI60 [P86857] protein from *M. californianus* did show a 76 % alignment with isotig 06796 (216; $2e^{-57}$) across a 188-aa sequence (Fig. S3). Similarly, the *M. californianus* chitin-binding SMP [P86860] gives a 100 % sequence identity with isotig 28411 (259; e^{-70}) for a 119-aa chain (Fig. S4). Isotig 20022 (260; $9e^{-72}$) gives a 97 % sequence alignment with a 124-aa sequence from the acidic whey perlwapin-like protein from *M. galloprovincialis* [P86855], including the signal peptide (Fig. 3). This continuous sequence encompasses two of the three whey acidic protein (WAP) domains. The perlucin-like C-lectin protein from *M. galloprovincialis* [P86854] also shows a significant alignment (69 %) with isotig 14840 (78; e^{-24}) from *M. edulis* (Fig. 4). The conserved WAP domains of the Perlwapins and the C-lectin domains of the Perlucins will be discussed in detail later.

Correlation Between *Pinctada* Bivalvia

Looking for similarities between the sequences derived from the new *M. edulis* transcriptome and the pearl oyster species *Pinctada*, we tentatively report a GN (glycine–asparagine) repeat domain for the alignment of isotig 48548 with Nacrein from *P. fucata* [Q27908] (Miyamoto et al. 2005) across a 61-aa sequence with 65 % identity (Fig. S5a). Further evidence for the existence of GN repeats in the *Mytilus* clade comes from a 64 % alignment identity along the complete GN repeat sequence of N16.1 matrix protein from *P. fucata* [Q9TVT2] (Samata et al. 1999) with the same isotig 48548 (39; $5e^{-4}$) (Fig. S5b). A longer GN repeat sequence was also picked up in the N66 matrix protein from *P. maxima* [Q9NL38] (Kono et al. 2000) where isotig 26479 (110; $8e^{-25}$) defines a 144-aa sequence with 50 % identity (Fig. S5c). Many of these repeat sequences obtained from *M. edulis* transcriptome were only obtained by switching off the BlastX low-complexity (SEG) filter control, which would normally have discarded these repeat sequences as garbage. Since many biomineral proteins contain very long sequence repeats, normally referred to as RLCDs, these would have been missed otherwise. With the filter on, isotig 36106 shows a (42; $3e^{-4}$) match and 57 % alignment with a 57-aa sequence found in Nacrein. There is also an analogous alignment with this isotig

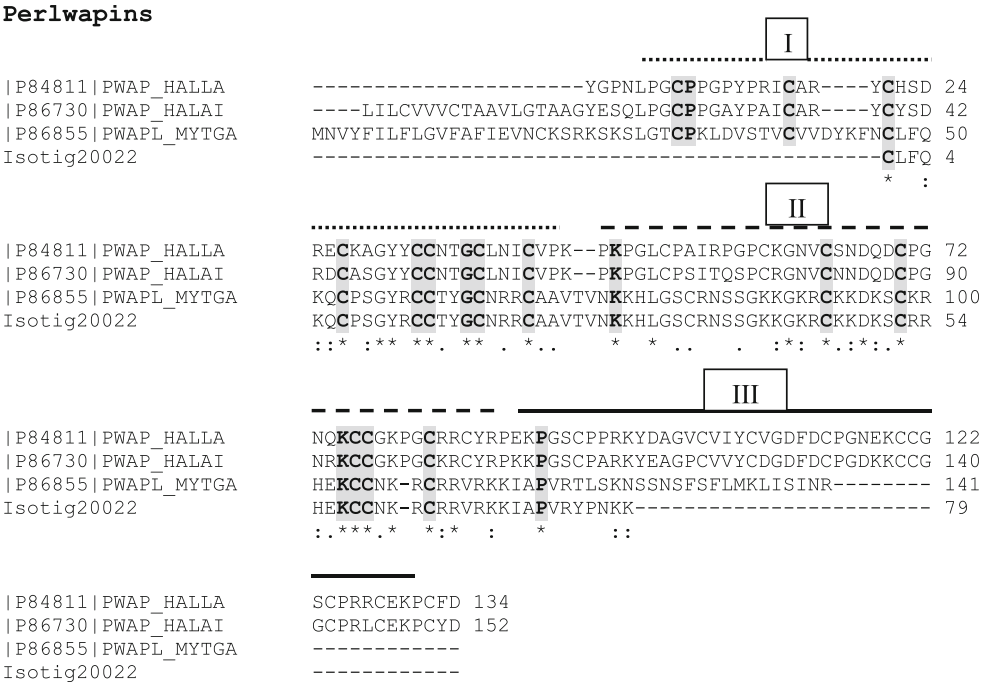

Fig. 3 Comparison of protein sequence from *M. edulis* transcriptome with sequence from Perlwapins. Comparison of protein sequence of isotig 20022 from *M. edulis* transcriptome with perlwapin sequence from *M. galloprovincialis* (*MYTGA*) and the gastropods, *Haliotis laevigata* (*HALLA*) and *H. asinina* (*HALAI*). The three WAP domains are indicated by *broken*, *dashed* and *bold lines above* the appropriate sequences. In this figure and in subsequent figures, including supplementary figures, we have used bold type and grey shading to highlight conserved amino acids across sequences and isotigs

for N45 nacrein-like protein [C7BCT8] from *P. maxima* (Yu et al. 2011) and the nacrein-like protein from *M. californianus* [P86856] (The *M. californianus* nacrein-like protein does not have a GN repeat). In these three alignments, the aa sequence GSLTTPPC is conserved (Fig. S5d–f). Isotig 46928 confirms this identity (41; $4e^{-4}$), which forms part of the second subdivision of the carbonic anhydrase catalytic domain (Miyamoto et al. 1996).

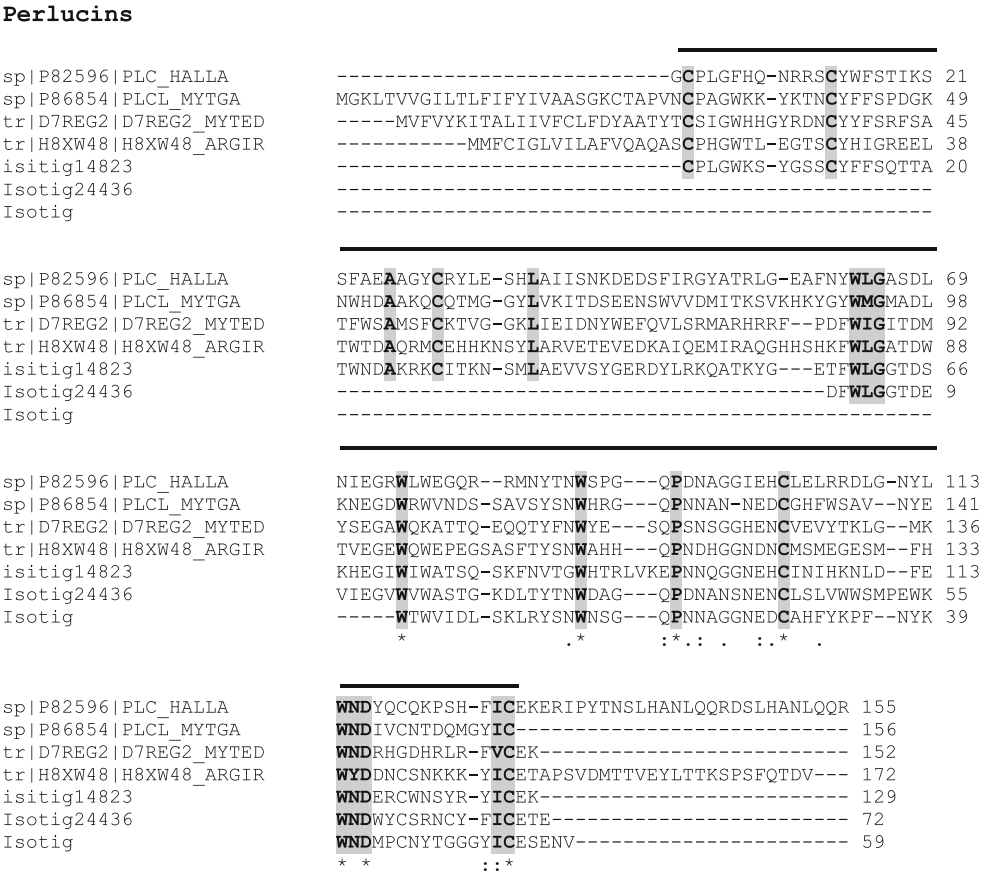

Fig. 4 Comparison of protein sequence from *M. edulis* transcriptome with sequence from Perlucins. Comparison of protein sequence of isotig 14840, 14823 and 24436 from *M. galloprovincialis* (*MYTGA*) and the gastropod *H. laevigata* (*HALLA*). Also included are the C-type lectin domains from *M. edulis* (*MYTED*) and the scallop *A. irradians* (ARGIR). The C-type lectin domain is shown by a *bold line above* the sequence

Repeat Low Complexity Domains—the Shematrins and Lysine-Rich Matrix Proteins

Jackson et al. (2010) highlight the relative abundance of proteins with repetitive low-complexity domains in both the bivalve *P. maxima* and the gastropod *H. asinina*. In particular, the silk fibroin domains of the Shematrins are absent from the *H. asinina* gene products and show divergent evolution within the three species of *Pinctada* studied. Since Shematrins are thought to be important in prism formation (Yano et al. 2006), we explored the new transcriptome for these Gly-Tyr-rich domains with the characteristic RKKKY, RRKKY, RRRKY, IRRKK and PRKKY C-terminal signature. Several Shematrins from both *P. fucata* and *P. maxima* were used (Fig. 5a, b; Fig. S6a–c). Isotig 17419 was dominant for all the Shematrins input, being aligned to most G_nY (n=2,3) repeat domains, typically shown for the Shematrin-like protein 2 from *P. maxima* [P86950] (Jackson et al. 2010) showing a 62 % identity for a 139-aa sequence (Fig. S6a). Interestingly, a common motif for this isotig shows $(GGGY\ GGYGI)_n$ where n varies from 2 to 4 and concurs with the usual G_nY repeat being followed by a hydrophobic amino acid. Of course, to define these as Shematrins, we looked at isotigs with lower value scores to find the characteristic C-terminal signature described earlier. All showed a close alignment with several isotigs as shown in Fig. 5b and Fig. S6c.

The small (~10 kDa) lysine-rich matrix proteins are characterised by a short lysine–tryptophan domain which follows immediately after the signal peptide. KRMPs have a Gly-Tyr-rich pre C-terminal domain involved in protein cross-linking, which can vary in length usually between ten and 40 residues in length and, in an analogous manner to the Shematrins, a short RKYKY, RPKKY, RRKY C-terminal motif. There was no precise match for the characteristic lysine tryptophan-rich lead domain. However, this threw up an interesting pseudo-match with many of the KRMPs where, although the tryptophans aligned with those from several *Pinctada* species, the lysines were out of sync. Often, the

Fig. 5 Comparison of protein sequence from *M. edulis* transcriptome with sequence from Shematrins. **a** Comparison of protein sequence of isotig 17419 from *M. edulis* transcriptome with the glycine-rich repeat domain of the Shematrin sequence from the bivalve *P. fucata* (*PINFU*) and **b** comparison of protein sequence of isotig 16784 with the C-terminal signature of Shematrin sequence from *P. maxima* (*PINMA*)

Shematrins

(a) Gly-rich repeat domain

```
|B6CHA9|PINFU        GNIATGSISSSVSGNIPYGGVLGIGGYGIGLG-GYGGYGLGG-YGGYGLG 48
Isotig17419          --------------NAEYGGYG--GGFGRGFGRGYGRYGIGGGYGGYGIG 34
                                  *  ***    **:* *:* *** **:** *****:*

|B6CHA9|PINFU        G-YGGYGLGG-YGGYGLGG-YGGYFPSYGSSLYGVSQSYPFGNAVFSGQA 95
Isotig17419          GGYGGYGIGGGYGGYGIGGGYGG----YGKGKVVVVKGYGGYGGGYGGLG 80
                     * *****:** *****:** ***    **..   * :.*   .. :.* .

|B6CHA9|PINFU        SGAGVPLFGSYN--FGGVGVGYPGGYYGGGGLIGGGGIIGGGGGVIGGGG 143
Isotig17419          GGYGG--YGGYGGGYGGLGGGY-GGYGGYGGGFGG---IGGGYGGYGGG- 123
                     .* *   :*.*.  :**:* ** *** * ** :**   **** *  ***

|B6CHA9|PINFU        VTVIRKKKY 152
Isotig17419          --------- 
```

(b) C-terminal signature

```
|P86950|SLP2_PINMA      MKPFISLASLIVLIASASAGGDDDYGKYGYGSYGPGIGGIGGGGGGIVIGGGGGGIGGGI 60
Isotig16784            ------------------------------------------------------------

|P86950|SLP2_PINMA      GGGIGGGIGGGGLIGGGGLIGGFGPGSVSGSVNQFGGVRTRAFGLGGTSPAVRGAAQGAA 120
Isotig16784            ------------------------------------------------------------

|P86950|SLP2_PINMA      TLSALGVASGRPSRVSGVSVGTGGGRALVSGSATPIGGYGIGIPYGVYGGGYGGYGGGYG 180
Isotig16784            ------------------------------------------------GYGGMGG-YG 9
                                                                         **** ** **

|P86950|SLP2_PINMA      GYGLGYGGYGGGYGGYGYGGYGYGPDLATFQGHTFGNLATGSINSLTGGYQIPYGGILSL 240
Isotig16784            GMG-GYGGMGG-YG--GMGGFG-G--------------------MSG------------ 31
                       * * **** ** **  * **:* *                    ::*

|P86950|SLP2_PINMA      YGPYGGYGGGYGGYGGGYGGYGGGYGIGIGSGGLGYGGYGYGGYGLGGSTLTGVSQSGPF 300
Isotig16784            ---YGGMMGGYGGIGGGHGGYWPG---------------------------------- 52
                          ***  ***** ***:***  *

|P86950|SLP2_PINMA      GTASMYGQAYGAGVPLFGTTYFGGVNVGSPYGIYGGGYPIGGIGGGAGPIGGGGIVIGGG 360
Isotig16784            --SSKYWLTYGRG---------------------------------------------- 63
                         :* *  :** *

|P86950|SLP2_PINMA      VGGIGGGGIGGIGGGGIGGGGIIGGGPIIRKKKY 394
Isotig16784            ----------------------IGGGSWIRPKK- 74
                                            ****. ** **
```

number of lysines was the same, but out of register with the *Pinctada* sequence for KRMP4 [B5KFE2] (Fig. 6a and Fig. S7a). We did find evidence for the GGY domains, especially for KRMP7 from *P. maxima* [P86960] (Jackson et al. 2010), with a 75 % identity for a 64-aa sequence (isotig 46876 (96; $5e^{-21}$)) (Fig. 6b). Similarly, the GGY domain in *P. margaritifera* KRMP11 [A7X133] is clearly seen in isotig 17419 (98; e^{-21}) with 60 % identity over this 90-aa sequence domain (Fig. S7b). As with the Shematrins, we looked for the C-terminal signature and could establish a closer alignment with the KRMPs than with the Shematrins shown in Fig. 6c and Fig. S7c.

Similar to the Shematrins is the acidic poly Gly shell framework proteins MSI31 [H3JZ93] (Sudo et al. 1997) and Prismalin-14 [Q6F4C6] (Suzuki et al. 2004) from *P. fucata*, where again isotig 17419 aligns with many of the G_nY repeat domains (Fig. S8a, b). Unlike the Shematrins, MSI31 has an interesting Glu-Asp-rich repeat domain dominated by EDXESE sequence repeat, where X is threonine or methionine. Isotig 17419 shows a 58 % alignment with this repeat domain (Fig. S8c).

One of the largest contiguous sequences found in our analysis is for isotig 09923 (400; e^{-112}) and a 636-aa sequence for poly Ala Gly MSI60 from *P. fucata* [G9MD31] (Sudo et al. 1997) which gives a 52 % identity and in particular a good alignment for the poly alanine repeats (Fig. S9). Looking further afield, we found an interesting alignment with the Japanese scallop *Mizuhopecten* (*Patinopecten*) *yessoensis* for the highly acidic protein MSP1 [Q95YF6] (Sarashina and Endo 2001). Isotig 09923 (197; $9e^{-51}$) gives a 47 % alignment (accepting positives) over a 541-aa sequence (Fig. S10). What makes it interesting is that many of the serine repeats have been replaced by alanine repeats and only 16 of the acidic 107 Asp residues have remained, which would make this an almost neutral protein. Interestingly, the alignment of the glycine residues is almost in complete register between MSP1 and isotig 09923 (Fig. S10). This pseudo-alignment between MSP1, MSI60 and isotig 09923 where one protein

KRMPs

(a) Lys-rich domain

```
|B5KFE3|PINMG        MRYAVLLAVVLLLGAFTAEGYWHPPLNICKWKLWKCLKWCAPWDWRCRKR 50
Isotig48548          ------------------KWKWR-----CKWK-WK---WRWKWKWKWRWR 23
                                        :  *:     **** **   *   *.*: * *

|B5KFE3|PINMG        CYWRYWWCLKRYGGGYGGYDYGDGGGYGGGGYGGGGYGGGGYGGGGYGGG 100
Isotig48548          WIWKWKW---------------------------------------------- 30
                       *:: *

|B5KFE3|PINMG        YDGGYDGGYDGGYGGGYGGGYGGGYGGGYGGGYGGGYHHRPKKY 144
Isotig48548          --------------------------------------------
```

(b) *P.maxima* Gly-rich domain

```
|P86960|GRP3_PINMA      MRYAVLLAVVLLLGAFTAEAFSPPPFHICKWKLWKCLKWCAPWDWKCRRKCFWKYWWCLK 60
Isotig46876          ------------------------------------------------------------

|P86960|GRP3_PINMA      KFGGHYGGYGYGDDGYGGGGYGG-GGYGGGYGGGYGGGYGGGYGDVGYGGGYSGGYSGGY 119
Isotig46876          --GGQ-GGFGGGLGGYGGG-YGGKGGYGGGYGGGYGGGYGGGYGG-GLGGGFGGGF-GGF 54
                       **: **:* * .***** *** ********************. * ***:.**: **:

|P86960|GRP3_PINMA      SGGYGSYGHRRKY 132
Isotig46876          GGGKGGYG----- 62
                     .** *.**
```

(c) Gly-rich domain and C-terminal signature

```
|B5KFE3|PINMG        MRYAVLLAVVLLLGAFTAEGYWHPPLNICKWKLWKCLKWCAPWDWRCRKR 50
Isotig16784          --------------------------------------------------
Isotig01803          --------------------------------------------------

|B5KFE3|PINMG        CYWRYWWCLKRYGGGYGGYDYGDGGGYGG-GGYGGGGYGGGGYGGG-GYG 98
Isotig16784          ----------GYGG-MGG--YGGMGGYGGMGGYGG----MGGFGGMSGYG 33
Isotig01803          -----GRGRGGFGGDRGG--RGGFGGGRGGGGYGGD---RGGRGGG-GRG 39
                           :**  **   *. **  * *****     ** **  * *

|B5KFE3|PINMG        GGYDGGYDGGYDGGYGGGYGGGYGGGYGGGYGGGYHHRPKKY 144
Isotig16784          G-MMGGY-GGIGGGHGGYWPG--SSKYWLTYGRGIGGGSWIRPKK- 74
Isotig01803          G---GGFGGGRDRGGGRDRPG------------GDRGD--RRPKPY 6
                     *   **: ** . * *    *            *  *.   ***
```


Fig. 6 Comparison of protein sequence from *M. edulis* transcriptome with sequence from lysine-rich matrix proteins (KRMPs). **a** Comparison of protein sequence of isotig 48548 from *M. edulis* transcriptome with KRMP sequence from the bivalve *P. margaritifera* (*PINMA*), **b** comparison of protein sequence of isotig 46876 with glycine-rich domain of KRMP from *P. maxima* (*PINMA*) and **c** comparison of protein sequence of isotig 16784 and 01803 with glycine-rich domain and C-terminal signature of KRMP from *P. margaritifera* (*PINMG*)

almost "ghosts" the other may allude to the premise that they originally shared a common locus.

Another highly acidic protein, Aspein, from both *P. maxima* [G9MBW9] (Isowa et al. 2012) and *P. fucata* [Q76K52] (Tsukamoto et al. 2004) shows remarkable sequence identity of 71 and 63 % for a 107- and 109-aa sequence respectively, with the same isotig 05617 (145; $3e^{-57}$ and 113; e^{-40}, respectively) where the *M. edulis* sequence occasionally interrupts the long Asp repeats usually with a MRERRN mutation (Fig. S11).

M. edulis and Gastropods

Having looked at the sequence correlation between *Mytilus* and *Pinctada* bivalves, we turned to see if there were any conserved domains between the *M. edulis* transcriptome and the haliotid gastropods. Marie et al. (2010) have already reported a strong sequence similarity between the *M. galloprovincialis* Perlwapin [P86855] and Perlucin [P86854] shell matrix proteins. Figure 3 shows the alignment of the Perlwapins from *H. asinina* and *H. laevigata* and *M. edulis* with isotig 20022 (49; $8e^{-7}$). Figure S12 shows the same alignment, but with an additional isotig 29469 (39; $6e^{-4}$). Both figures show good alignment for the WAP protease inhibitor-like domains with highly conserved cysteines. Further protease inhibition was identified by a Kunitz-like type II [P86733] domain found with isotigs 11877 (54; $2e^{-8}$) and 38255 (61; $6e^{-11}$) (Fig. S13).

An analogous situation to the Perlwapins arises for the Perlucins where again three isotigs define the majority of the sequence for *H. laevigata* [P82596]—isotig 14823 (94; $2e^{-20}$), isotig 24436 (72; $7e^{-14}$) and isotig 14840 (78; e^{-24}). The alignments, along with that for *M. galloprovincialis*, are shown in Fig. 4. Perlucins exhibit high sequence homology to the C-type lectin family of calcium-dependent carbohydrate-binding proteins. Figure 4 includes the C-type lectin domain from the scallop *Argopecten irradians* (AiCTL5) [H8XW48] (Mu et al. 2012) and the putative C-lectin domain from *M. edulis* [D7REG2] (Espinosa et al. 2010), both of which are indicative of this type of domain.

The glutamine-rich protein from *H. asinina* [P86727] which is thought to be an intermediary in protein aggregation (Barton et al. 2007) is denoted in a number of isotigs, the strongest being isotig 14179 (45; $9e^{-6}$) which gives a relatively conserved alignment of the glutamine residues over a 94-aa sequence (Fig. S14).

As with the bivalves, the Gly-Ala-rich protein from *H. asinina* [P86732] shows a tentative alignment with isotig 09923 (190; $9e^{-49}$), having 42 % alignment over a 527-aa sequence (Fig. S15a). Isotig 09923 was also dominant in MSI60 and, to a lesser extent, MSPI. Interestingly, isotig 26749 (118; $4e^{-27}$) gives a 62 % alignment with the GN repeat C-terminal domain (Fig. S15b).

Discussion

The objective of this research was to generate a mantle transcriptome for *M. edulis* in order to search for characteristic biomineral protein signatures. It is important to stress that in this paper we highlight only the strongest sequence agreements for brevity. The gene ontology data (Fig. 2) show the expected dominance of binding and catalytic activity in terms of molecular function, and the high quality of the phred scores allows confidence in the analysis of the assembled transcriptome. The depth of reads afforded by 454 pyrosequencing has allowed us to discover several SMPs that previously could not be detected in other *Mytilus* data bases. In so doing, it potentially fills some of the gaps between proteins found in the shell matrix of the pearl oyster *Pinctada* and that of *Mytilus*—in particular, matching of identities for the Shematrins and KRMPs and also evidence for the GN repeats from the nacrein suite of proteins. Although it is not possible to determine which specific Shematrin or KRMP has been identified, the characteristic protein profile for each of these sets of proteins has been unmistakably extracted from this new *M. edulis* transcriptome. For example, isotig 17419 is heavily aligned to the G_nY domains for both of these classes of protein, although not exclusively. Many isotigs also show good alignment and sequence identity with the G_nY domains of the Shematrins and KRMPs, although isotig 17419 is the most prevalent. Isotig 17419 shows two distinct regions for MSI31 (Fig. S8) in which good identity occurs at the beginning and end of the sequence with poor alignment in the middle. The C-terminal R_nK_mY signature for both Shematrins and KRMPs was found to involve several isotigs and may indicate diversity in this signature for *M. edulis*. Specifically, the lysine tryptophan-rich N-terminal domain, which defines the KRMPs, could be identified through a number of isotigs, although the sequence showed partial mismatch mainly for the lysines being out of sync with those of the *Pinctada* sequence.

With two separate isotigs (06796 and 09923) defining substantial sequence identity in the silk fibroin framework MSI60 protein from both *M. californianus* and *P. fucata* adds weight to the idea that MSI60 may be ubiquitous to nacre-forming organisms. Although we have demonstrated (for the few proteins shown here) a strong correlation among other species of *Mytilus*, we were surprised that the MUSP-2 and −3 proteins showed poor sequence identity with the isotigs of the transcriptome. Other proteins where consistent identity matching was found (MSI60, chitin-binding MSP, Perlwapins and Perlucins) are well characterised in terms of their role in mineral shell formation and regulation, whereas the MUSP proteins have yet to be characterised. Perhaps once their role in shell formation is more clearly resolved, it may be easier to explain this.

The high level of similarity between Perlwapin and WAP protease inhibitor-like domains is consistent with previous observations of acellular protease inhibition in biominerals including abalone (Marie et al. 2010) and pearl oysters (Bedouet et al. 2007; Liu et al. 2007). Marie et al. (2010) suggest that these inhibitors may play a role in protecting against proteolytic degradation and may also function to remodel the shell matrix. Interestingly, in siliceous biominerals, silicatein-α with six conserved cysteine residues functions to polymerise silica while having similarity with the cysteine-protease cathepsin-L (Cha et al. 1999).

In conclusion, this is another data set to add to the already expanding data that are now available to try and piece together the evolutionary story of biomineralisation. In this study, we can start to see tentative links between proteins found in other clades which have previously not been found. Perhaps the direction that now needs to be taken, now that sequencing technology continues at a pace, is in the careful application of bioinformatics with an essential hands-on approach where researchers use their savvy to derive the best from multiple data sets.

Acknowledgments AF, JJ and MC gratefully acknowledge funding from BBSRC grant number BB/E025110/1. 454 library preparation and sequencing was carried out by Anna Montazam and Denis Cleven at The GenePool sequencing facility, University of Edinburgh. We thank Ben Elsworth and Sujai Kumar of GenePool sequencing facility, University of Edinburgh, for setting up the InterproScan pipeline and sge_blast.

Appendix I

The main commands used for assembly were:

Newbler assembly:

(a) Newbler 2.3:

```
runAssembly -cdna -vt SMARTadapters.fna -o assembly_newbler23 454reads.sff
```

(b) Newbler 2.5.2:

```
runAssembly -cdna -vt SMARTadapters.fna -urt -o assembly_newbler252_urt 454reads.sff
```

(c) Newbler 2.6: same command as used for Newbler 2.5.2 above.

Note: The '-cdna' option enables the Newbler transcriptome assembly to construct the possible isoforms and isogroups from the reads. The '-vt' option locates and trims adapter sequences given in the 'SMARTadapters.fna' file from the reads before using the read in the assembly. The '-urt' ('use read tips') is an option introduced in Newbler version 2.5 which uses low-coverage read tips to extend the contigs, yielding a larger assembly although less accurate in the low-coverage read-tip regions (For a more detailed explanation of these options, see the Roche Newbler Assembler manual, 2011).

(2) MIRA assemblies:

MIRA does not trim adapters from reads, so a custom Perl script was used to locate and trim the SMART adapters from the reads. This script used the Roche 'sffinfo' program to extract the fasta sequences from the .sff file. Then a local 'blastn' executable (part of from NCBI's blast+) was used to locate the position of any SMART adapters in the reads (using option for search word-size of 4). The Roche 'sfffile' program was then used to adjust the left and right trim positions within the sff file for each read. The trimmed reads were then extracted using the MIRA third-party tool 'sff_extract':

```
sff_extract -s mira_in.454.fasta -q mira_in.454.fasta.qual -x mira_traceinfo_in.454.xml 454reads_SMARTtrimmed.sff
```

Then run the MIRA assembly, redirecting the output to log files:

```
mira -project=mira - -job=denovo,est,normal,454 454_SETTINGS -CL:qc=no> log_asm.out 2> log_asm.err
```

Note: The 'est' option is for transcriptome assembly; the '-CL:qc=no' is to disable quality-clipping of reads as the reads clipping points have already been set to trim both low quality and the SMART adapters. This MIRA assembly command is based on the (MIRA online manual, 2010). The same command was used for MIRA version 3.2.1 and for the subsequent MIRA version 3.4.0.1.

(3) CAP3:

All FASTA sequences for the Isotigs from Newbler 2.5, and the contigs from MIRA 3.4.1, were concatenated into one file, and quality scores into another file:

```
cat Newbler252_Isotigs.fna Mira_v321_unpadded.fasta> newbler252_mira321.fasta
```

```
cat Newbler252_Isotigs.qual Mira_v321_unpadded.qual> newbler252_mira321.fasta.qual
```

(CAP3 expects the quality scores be in file with ".qual" appended to the reads fasta filename)

Then CAP3 was run with default settings:

```
cap3 newbler252_mira321.fasta> cap3.log
```

A script was used to count CAP3 contigs that contain Newbler+MIRA contigs, only Newbler, and only MIRA contigs.

(4) Assembly metrics and extracting singletons:

For each assembly, a Perl script was used to summarise the statistics for the assembled contigs. Another script was used to extract the unassembled (singleton) reads from each assembly.

References

Addadi L, Joester D, Nudelman F, Weiner S (2006) Mollusk shell formation: a source of new concepts for understanding biomineralization processes. Chem-Eur J 12:981–987

Barton S, Jacak R, Khare SD, Ding F, Dokholyan NV (2007) The length dependence of the PolyQ-mediated protein aggregation. J Biol Chem 282:25487–25492

Bedouet L, Duplat D, Marie A, Dubost L, Berland S, Rousseau M, Milet C, Lopez E (2007) Heterogeneity of proteinase inhibitors in the water-soluble organic matrix from the oyster nacre. Mar Biotechnol 9:437–449

Bedouet L, Marie A, Berland S, Marie B, Auzoux-Bordenave S, Marin F, Milet C (2012) Proteomic strategy for identifying mollusc shell proteins using mild chemical degradation and trypsin digestion of insoluble organic shell matrix: a pilot study on *Haliotis tuberculata*. Mar Biotechnol 14:446–458

Burton CA, MacMillan JT, Learmouth MM (2001) Shellfish ranching in the UK. Hydrobiologia 465:1–5

Cartwright JHE, Checa AG (2007) The dynamics of nacre self-assembly. J Royal Soc Interface 4:491–504

Cha JN, Shimizu K, Zhou Y, Christiansen SC, Chmelka BF, Stucky GD, Morse DE (1999) Silicatein filaments and subunits from a marine sponge direct the polymerization of silica and silicones in vitro. Proc Natl Acad Sci 96:361–365

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WEG, Wetter T, Suhai S (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res 14:1147–1159

Clark MS, Thorne MAS, Vieira FA, Cardoso JCR, Power DM, Peck LS (2010) Insights into shell deposition in the Antarctic bivalve *Laternula elliptica*: gene discovery in the mantle transcriptome using 454 pyrosequencing. BMC Genomics 11

Craft JA, Gilbert JA, Temperton B, Dempsey KE, Ashelford K, Tiwari B, Hutchinson TH, Chipman JK (2010) Pyrosequencing of *Mytilus galloprovincialis* cDNAs: tissue-specific expression patterns. PLoS One 5:e8875

Currey JD, Zioupos P, Davies P, Casinos A (2001) Mechanical properties of nacre and highly mineralized bone. Proc R Soc London, Ser B 268:107–111

Espinosa EP, Perrigault M, Allam B (2010) Identification and molecular characterization of a mucosal lectin (MeML) from the blue mussel *Mytilus edulis* and its potential role in particle capture. Comp Biochem Physiol A-Mol Integr Physiol 156:495–501

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8:186–194

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8:175–185

Hattan SJ, Laue TM, Chasteen ND (2001) Purification and characterization of a novel calcium-binding protein from the extrapallial fluid of the mollusc, *Mytilus edulis*. J Biol Chem 276:4461–4468

Huang XQ, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9:868–877

Isowa Y, Sarashina I, Setiamarga DHE, Endo K (2012) A comparative study of the shell matrix protein aspein in Pterioid bivalves. J Mol Evol 75:11–18

Jackson AP, Vincent JFV, Turner RM (1988) The mechanical design of nacre. Proc R Soc London, Ser B 234:415

Jackson DJ, McDougall C, Green K, Simpson F, Worheide G, Degnan BM (2006) A rapidly evolving secretome builds and patterns a sea shell. BMC Biol 4

Jackson DJ, McDougall C, Woodcroft B, Moase P, Rose RA, Kube M, Reinhardt R, Rokhsar DS, Montagnani C, Joubert C, Piquemal D, Degnan BM (2010) Parallel evolution of nacre building gene sets in molluscs. Mol Biol Evol 27:591–608

Ji BZ, Cusack M, Freer A, Dobson PS, Gadegaard N, Yin HB (2010) Control of crystal polymorph in microfluidics using molluscan 28 kDa Ca^{2+}-binding protein. Integr Biol 2:528–535

Joubert C, Piquemal D, Marie B, Manchon L, Pierrat F, Zanella-Cleon I, Cochennec-Laureau N, Gueguen Y, Montagnani C (2010) Transcriptome and proteome analysis of *Pinctada margaritifera* calcifying mantle and shell: focus on biomineralization. BMC Genomics 11

Kinoshita S, Wang N, Inoue H, Maeyama K, Okamoto K, Nagai K, Kondo H, Hirono I, Asakawa S, Watabe S (2011) Deep sequencing of ESTs from nacreous and prismatic layer producing tissues and a screen for novel shell formation-related genes in the pearl oyster. PLoS One 6:e21238

Kono M, Hayashi N, Samata T (2000) Molecular Mechanism of the Nacreous Layer Formation in Pinctada maxima. Biochem Biophys Res Commun 269:213–218

Kumar S, Blaxter ML (2010) Comparing de novo assemblers for 454 transcriptome data. BMC Genomics 11

Liu HL, Liu SF, Ge YJ, Liu J, Wang XY, Xie LP, Zhang RQ, Wang Z (2007) Identification and characterization of a biomineralization related gene PFMG1 highly expressed in the mantle of *Pinctada fucata*. Biochemistry 46:844–851

Marie B, Marie A, Jackson DJ, Dubost L, Degnan BM, Milet C, Marin F (2010) Proteomic analysis of the organic matrix of the abalone *Haliotis asinina* calcified shell. Proteome Sci 8

Marie B, Le Roy N, Zanella-Cleon I, Becchi M, Marin F (2011a) Molecular evolution of mollusc shell proteins: insights from proteomic analysis of the edible mussel *Mytilus*. J Mol Evol 72:531–546

Marie B, Trinkler N, Zanella-Cleon I, Guichard N, Becchi M, Paillard C, Marin F (2011b) Proteomic identification of novel proteins from the calcifying shell matrix of the Manila clam *Venerupis philippinarum*. Mar Biotechnol 13:955–962

Marie B, Zanella-Cleon I, Guichard N, Becchi M, Marin F (2011c) Novel proteins from the calcifying shell matrix of the Pacific oyster *Crassostrea gigas*. Mar Biotechnol 13:1159–1168

Marie B, Joubert C, Belliard C, Tayale A, Zanella-Cleon I, Marin F, Gueguen Y, Montagnani C (2012) Characterization of MRNP34, a novel methionine-rich nacre protein from the pearl oysters. Amino Acids 42:2009–2017

Marin F, Luquet G, Marie B, Medakovic D (2008) Molluscan shell proteins: primary structure, origin, and evolution. Curr Top Dev Biol 80:209–276

Miyamoto H, Miyashita T, Okushima M, Nakano S, Morita T, Matsushiro A (1996) A carbonic anhydrase from the nacreous layer in oyster pearls. Proc Natl Acad Sci U S A 93:9657–9660

Miyamoto H, Miyoshi F, Kohno J (2005) The carbonic anhydrase domain protein nacrein is expressed in the epithelial cells of the mantle and acts as a negative regulator in calcification in the mollusc *Pinctada fucata*. Zool Sci 22:311–315

Mu C, Song X, Zhao J, Wang L, Qiu L, Zhang H, Zhou Z, Wang M, Song L, Wang C (2012) A scallop C-type lectin from *Argopecten irradians* (AiCTL5) with activities of lipopolysaccharide binding and Gram-negative bacteria agglutination. Fish & Shellfish Immunol 32:716–723

Nudelman F, Gotliv BA, Addadi L, Weiner S (2006) Mollusk shell formation: mapping the distribution of organic matrix components underlying a single aragonitic tablet in nacre. J Struct Biol 153:176–187

Nudelman F, Shimoni E, Klein E, Rousseau M, Bourrat X, Lopez E, Addadi L, Weiner S (2008) Forming nacreous layer of the shells of the bivalves *Atrina rigida* and *Pinctada margaritifera*: an environmental- and cryo-scanning electron microscopy study. J Struct Biol 162:290–300

Roche (2009a) Genome Sequencer FLX System Software Manual, version 2.3 Part C: GS De Novo Assembler – GS Reference Mapper – SFF Tools, 454. User Manual. Life Sciences Corp. A Roche Company, Brandford

Roche (2009b) GS FLX titanium general library preparation method manual. Method Manual. Roche Diagnostics GmbH, Mannheim

Roche (2010) 454 Sequencing System Software Manual, v 2.5.3: Part C – GS De Novo Assembler, GS Reference Mapper, SFF Tools, 454. Life Sciences Corp, A Roche Company, Branford

Samata T, Hayashi N, Kono M, Hasegawa K, Horita C, Akera S (1999) A new matrix protein family related to the nacreous layer formation of Pinctada fucata. FEBS Lett 462:225–229

Sarashina I, Endo K (2001) The complete primary structure of molluscan shell protein 1 (MSP-1), an acidic glycoprotein in the shell matrix of the scallop *Patinopecten yessoensis*. Mar Biotechnol 3:362–369

Sarashina I, Endo K (2006) Skeletal matrix proteins of invertebrate animals: comparative analysis of their amino acid sequences. Paleontol Res 10:311–336

Schmid R, Blaxter ML (2008) annot8r: GO, EC and KEGG annotation of EST datasets. BMC Bioinformatics 9

Sudo S, Fujikawa T, Nagakura T, Ohkubo T, Sakaguchi K, Tanaka M, Nakashima K, Takahashi T (1997) Structures of mollusc shell framework proteins. Nature 387:563–564

Suzuki M, Murayama E, Inoue H, Ozaki N, Tohse H, Kogure T, Nagasawa H (2004) Characterization of Prismalin-14, a novel matrix protein from the prismatic layer of the Japanese pearl oyster (*Pinctada fucata*). Biochem J 382:205–213

Tanguy A, Bierne N, Saavedra C, Pina B, Bachere E, Kube M, Bazin E, Bonhomme F, Boudry P, Boulo V, Boutet I, Cancela L, Dossat C, Favrel P, Huvet A, Jarque S, Jollivet D, Klages S, Lapegue S, Leite R, Moal J, Moraga D, Reinhardt R, Samain J-F, Zouros E, Canario A (2008) Increasing genomic information in bivalves through new EST collections in four species: development of new genetic markers for environmental studies and genome evolution. Gene 408:27–36

Truebano M, Burns G, Thorne MAS, Hillyard G, Peck LS, Skibinski DOF, Clark MS (2010) Transcriptional response to heat stress in the Antarctic bivalve *Latemula elliptica*. J Exp Mar Biol Ecol 391:65–72

Tsukamoto D, Sarashina I, Endo K (2004) Structure and expression of an unusually acidic matrix protein of pearl oyster shells. Biochem Biophys Res Commun 320:1175–1180

Werner GDA, Gemmell P, Grosser S, Hamer R, Shimeld SM (2013) Analysis of a deep transcriptome from the mantle tissue of *Patella vulgata* Linnaeus (Mollusca: Gastropoda: Patellidae) reveals candidate biomineralising genes. Mar Biotechnol 15:230–243

Yano M, Nagai K, Morimoto K, Miyamoto H (2006) Shematrin: a family of glycine-rich structural proteins in the shell of the pearl oyster *Pinctada fucata*. Comp Biochem Physiol B 144:254–262

Yin Y, Huang D, Paine ML, Reinhold VN, Chasteen ND (2005) Characterization of the primary structure of the EP protein from the extrapallial fluid of the mollusc, *Mytilus edulis*. Abstr Pap Am Chem Soc 229:U228–U229

Yin HB, Ji BZ, Dobson PS, Mosbahi K, Glidle A, Gadegaard N, Freer A, Cooper JM, Cusack M (2009) Screening of biomineralization using microfluidics. Anal Chem 81:473–478

Yu DH, Wang Y, Tang R (2011) Cloning and characterization of nacre-related genes in silver-lip pearl oyster *Pinctada maxima*. Shanghai Hai Yang Da Xue Xue Bao 20:121–128

Zhang C, Zhang R (2006) Matrix proteins in the outer shell of molluscs. Mar Biotechnol 8:572–586

Zhu YY, Machleder EM, Chenchik A, Li R, Siebert PD (2001) Reverse transcriptase template switching: a SMART (TM) approach for full-length cDNA library construction. Biotechniques 30:892–897