Mar Biotechnol
DOI 10.1007/s10126-012-9481-0

ORIGINAL ARTICLE

Analysis of a deep transcriptome from the mantle tissue of *Patella vulgata* Linnaeus (Mollusca: Gastropoda: Patellidae) reveals candidate biomineralising genes

Gijsbert D. A. Werner · Patrick Gemmell · Stefanie Grosser · Rebecca Hamer · Sebastian M. Shimeld

Received: 24 April 2012 / Accepted: 17 July 2012

Abstract The gastropod *Patella vulgata* is abundant on rocky shores in Northern Europe and a significant grazer of intertidal algae. Here we report the application of Illumina sequencing to develop a transcriptome from the adult mantle tissue of *P. vulgata*. We obtained 47,237,104 paired-end reads of 51 bp, trialled de novo assembly methods and settled on the additive multiple K method followed by redundancy removal as resulting in the most comprehensive assembly. This yielded 29,489 contigs of at least 500 bp in length. We then used three methods to search for candidate genes relevant to biomineralisation: searches via BLAST and Hidden Markov Models for homologues of biomineralising genes from other molluscs, searches for predicted proteins containing tandem repeats and searches for secreted proteins that lacked a transmembrane domain. From the results of these searches we selected 15 contigs for verification by RT-PCR, of which 14 were successfully amplified and cloned. These included homologues of Pif-177/BSMP, Perlustrin, SPARC, AP24, Follistatin-like and Carbonic anhydrase, as well as three containing extensive G-X-Y repeats as found in nacrein. We selected two for further verification by in situ hybridisation, demonstrating expression in the larval shell field. We conclude that de novo assembly of Illumina data offers a cheap and rapid route to a predicted transcriptome that can be used as a resource for further biological study.

Keywords Limpet · Patella · Transcriptome · Biomineralisation · Shell

Introduction

The shells of molluscs are distinctive and are often beautiful structures that have attracted the interest of scientists from many different disciplines. Shells fossilise relatively well, providing a rich source of paleontological data. They present conceptually interesting problems to developmental biologists interested in how the structure of the shell and its pigmentation are controlled (Wilt 2005). Commercial

Electronic supplementary material The online version of this article (doi:10.1007/s10126-012-9481-0) contains supplementary material, which is available to authorized users.

G. D. A. Werner · P. Gemmell · S. Grosser · S. M. Shimeld (✉)
Department of Zoology, University of Oxford,
South Parks Road,
Oxford OX1 3PS, UK
e-mail: sebastian.shimeld@zoo.ox.ac.uk

P. Gemmell
Life Sciences Interface Doctoral Training Centre, University of Oxford,
Oxford, UK

R. Hamer
Department of Statistics, University of Oxford,
1 South Parks Road,
Oxford OX1 3TG, UK

R. Hamer
Doctoral Training Centre,
Rex Richards Building, South Parks Road,
Oxford OX1 3QU, UK

Present Address:
G. D. A. Werner
Faculty of Earth and Life Sciences, VU University Amsterdam,
De Boelelaan 1085,
1081 HV Amsterdam, The Netherlands

Present Address:
S. Grosser
Landcare Research—Manaaki Whenua,
231 Morrin Road, St Johns,
Auckland 1072, New Zealand

Published online: 03 August 2012

interests come from a biomaterials perspective, particularly in the formation of pearls by some molluscs. More recently, rising atmospheric carbon dioxide (CO_2) levels are affecting oceanic pH. The acidification this causes may compromise the ability of molluscs to build shells, attracting interest from ecologists and environmental scientists (Kleypas et al. 2006; Orr et al. 2005).

Despite this focusing of interests, we know relatively little about how mollusc shells are built. The phylum Mollusca is speciose, with estimates of around 100,000 species (Ruppert et al. 2004). While the adult shells of most of these species are well-documented by both scientists and collectors, the mechanisms underlying their construction have only been addressed in a handful of taxa. Mollusc shells are dominated by calcium carbonate ($CaCO_3$) and also include a small percentage of organic material (Marin and Luquet 2004). The organic molecules form a complex matrix upon which the various shell structures are deposited at different points in the life cycle (Wilt 2005), resulting in a layered shell structure (Fig. 1b). Though mainly containing proteins, the organic matrix also consists of other organic compounds such as glycoproteins, lipids and polysaccharides (Marin and Luquet 2004). The organic matrix is thought to regulate the deposition of different isoforms of $CaCO_3$, as well as becoming incorporated into the shell itself, and hence to control the physical properties of the shell (Veis 2003). It is thus of both scientific and commercial interest to know what proteins constitute the mollusc shell organic matrix.

The historical approach to addressing the protein constituents of shells has been biochemical: the protein component of the shell can be separated from the mineral component by physical and chemical means, and the dominant proteins extracted, sequenced and hence identified. This has been a successful approach, identifying several key constituents of some shells, for example RP-1 from *Crassostrea virginica*, calprismin and caspartin from *Pinna nobilis*, dermatopontin from *Biomphalaria glabrata* and perlustrin and perlucin from *Haliotis laevigata* (Marin et al. 2005; Donachy et al. 1992; Marxen et al. 2003; Weiss et al. 2000). More recently, this approach has been extended to include newer proteomic techniques (Marie et al. 2010).

An alternative approach is via analysis of cDNAs derived from the tissue responsible for constructing the shell. Mollusc shell formation takes places extracellularly, in the extrapallial space between the mollusc mantle and previously formed shell layers (Marin and Luquet 2004). Specifically, shell growth occurs at the mantle edges, where specialised epithelial cells secrete the organic matrix (Kniprath 1981). The analysis of cDNA derived from the mollusc mantle edge can therefore be used to build on known peptide sequences and has lead to the identification of several relevant genes, such as that encoding AP24 from *Haliotis rufescens* (Michenfelder et al. 2003).

An extension of this approach involves generating a large sequence dataset which can then be analysed bioinformatically and used as a resource for the development of cloned genes for expression and functional analyses. The first approaches in this area involved the generation of Expressed Sequence Tag (EST) datasets via Sanger sequencing (for example, Jackson et al. 2006; Fang et al. 2011) and resulted in the identification of hundreds of potential proteins relevant to shell formation (Jackson et al. 2007). A similar approach has been taken by Heyland and colleagues (2011), exploiting the extensive transcriptome information initially developed to study the neurobiology of the sea hare *Aplysia californica* (Moroz et al. 2006). More recently methods collectively known as Next-Generation sequencing have allowed larger datasets to be generated. For example, 454 pyrosequencing has been used to generate deep sequence data for the bivalve molluscs *Pinctada margaritifera* (Joubert et al. 2010), *Ruditapes philippanarum* (Milan et al. 2011) and *Laternula elliptica* (Clark et al. 2010). A key outcome of these studies is that candidate mollusc biomineralisation gene sets show low overlap between species, suggesting relatively rapid evolution of mollusc biomineralisation genes (Jackson et al. 2006; Jackson et al. 2010).

Another sequencing methodology that can be applied to transcriptome development is the Illumina platform. Compared to 454 pyrosequencing, Illumnia sequencing has the advantage of producing considerably more data per unit cost. It also has the disadvantage of generally producing shorter reads than 454 pyrosequencing, and hence, can entail greater bioinformatic challenges in assembling the data into putative transcripts. A recent report, however, has suggested these challenges can be overcome, at least for the freshwater snail *Radix balthica* (Feldmeyer et al. 2011). Here we report the application of Illumina GAII sequencing to transcriptome development for the mantle tissue of the common European limpet *Patella vulgata* (Fig. 1). We chose *P. vulgata* to study for several reasons. It is highly abundant on UK shores where limpets occupy an important ecological position. Unlike many molluscs, it deploys broadcast spawning, liberating gametes into the seawater where fertilisation, embryogenesis and larval development occur (Smith 1935). This has advantages for access to, and experimentation with, early developmental stages.

We show that a single lane of paired-end 51 bp Illumina GAII sequence data is sufficient to develop significant insight into the transcriptome of *P. vulgata*. We trial assembly methods and analyse the resulting assembly using bioinformatic tools. To test the quality of assembled transcripts, we target sequences homologous

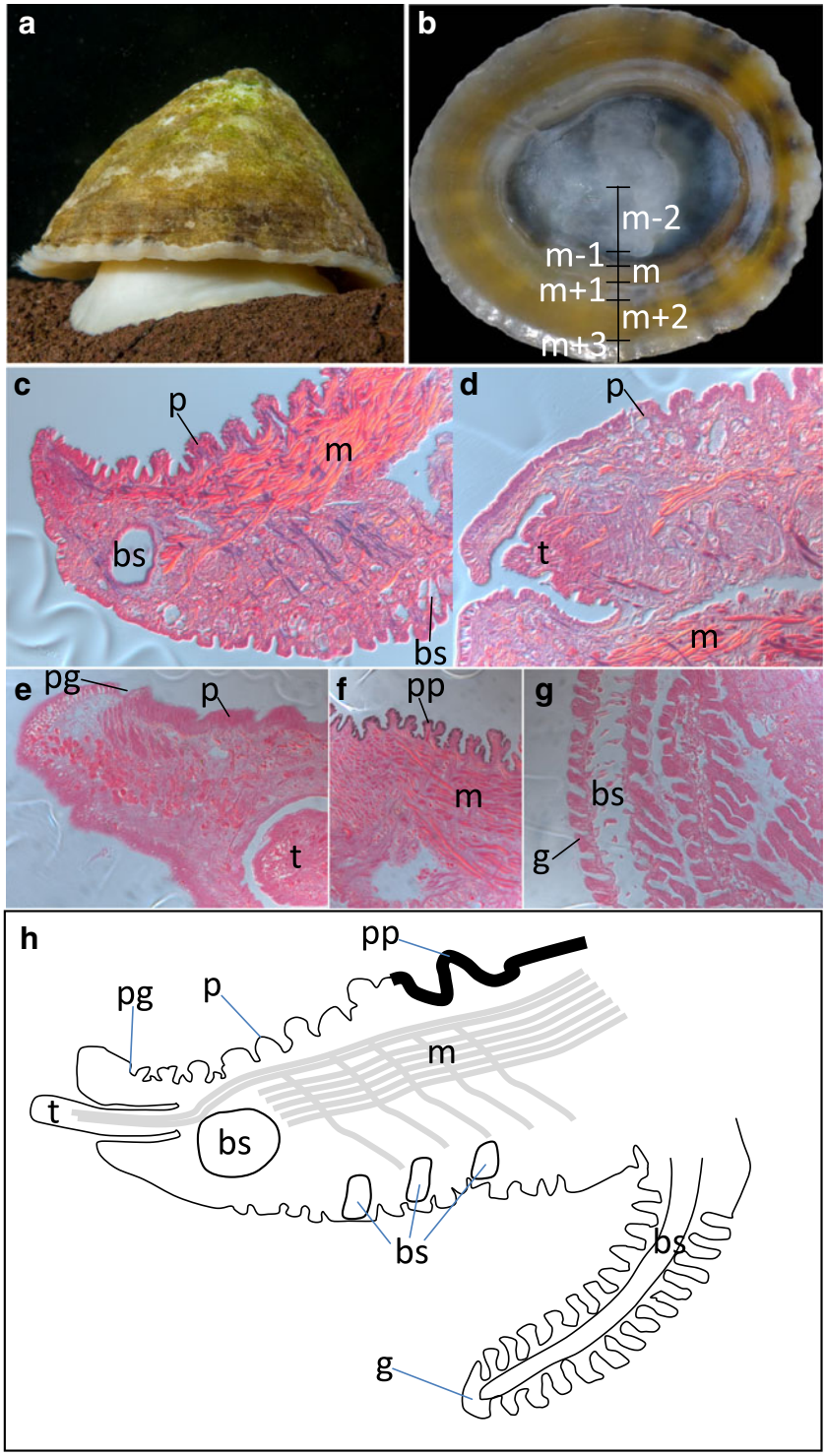

Fig. 1 Morphology and histology of the *Patella vulgata* shell and pallium. **a** Live adult *P. vulgata* showing the foot anchored to the substratum and the mantle edge projecting from underneath the shell. Image courtesy of Dr. Paul Naylor. **b** Internal view of the *P. vulgata* shell. The different zones identified by (Fuchigami and Sasaki 2005) are marked: *m* is the myostracum, and successive inner and outer layers are labelled as *m−1* and *m− 2* or *m+ 1* and *m+ 2*, respectively. **c** Section through the mantle edge showing pallium (*p*), muscle (*m*) and blood sinuses (*bs*). **d** Section through a different region of the mantle edge. A tentacle (*t*) embedded in a pit is visible. Note the distal pallium in regions adjacent to tentacles is less folded than regions without tentacles such as **c**. **e** Section through another region of mantle edge. This includes a cross section of a tentacle in its pocket. A small furrow close to the distal edge of the pallium is visible, and may be the pallial groove (*pg*). **f** Pallial surface at a more proximal position than that shown in **c**–**e**. Note the pigmentation of the pallial epithelium (*pp*) in this location. **g** Section through the pallial gills (*g*), showing their convolute structure and large internal blood sinus. **h** Schematic diagram of the mantle edge of *P. vulgata*. The pallial surface including a pigmented region is shown. Muscle fibres run proximal-distal and dorsal-ventral through the mantle, and some penetrate the tentacles. The pallial surface is highly convoluted, as is the ventral mantle surface although this is also associated with blood sinuses. A pallial gill is also shown

to known shell genes in other molluscs and amplify them by RT-PCR from mantle RNA, with a high success rate. We also examine the expression of some of these in developing embryos. This demonstrates the utility of high-throughput short-read sequencing for generating extensive sequence data from a non-model species and provides such a dataset for further investigation of *P. vulgata* biology.

Methods

Collection of adult, embryonic and larval *P. vulgata*

Adult *P. vulgata* were collected from Tinside, Plymouth, UK or from Northney Marina, Hampshire, UK. They were maintained in a recirculating sea water aquarium at approximately 10 °C. Adults are gravid between approximately

October and March. Gametes were liberated by dissection into filtered sea water. *P. vulgata* oocytes are held in a meiotic block; to release this, they were treated for 10 min with sea water containing 10 mM Tris–HCL pH 9.0 and 5 mM NH_4Cl, washed twice with sea water, then fertilised by addition of approximately 10^6 sperm per millilitre for 30 mins (Hodgson et al. 2007). Sperm were then washed off with excess sea water and the developing embryos incubated at 12 °C until they had reached the desired stage. For in situ hybridisation, they were fixed in 4 % paraformaldehyde in MOPS buffer (1 mM $MgSO_4$, 2 mM EGTA, 0.5 M NaCl, 0.1 M MOPS pH 7.5) at 4 °C for at least overnight. Embryos were subsequently washed twice with DEPC treated PBT (phosphate-buffered saline with 0.1 % Tween 20), then dehydrated through a progressively more concentrated PBT–methanol series before being washed twice with 100 % methanol and stored at −20 °C. For histology, mantles were dissected from adult *P. vulgata* and fixed and stored as for in situ hybridisation. They were then rehydrated into PBS and stained with acidified 1 % Poinceau S for 30 mins, before dehydration through graded ethanol–PBS and equilibration overnight at 4 °C in LR White medium resin (TAAB). Sections were cut on a Reichert Jung Supercut microtome at 3 or 7 μm.

RNA extraction and sequencing

A schematic summary of the workflow with summary data can be seen in Fig. 2. A single adult *P. vulgata* collected from Tinside, Plymouth, was used as a source of mRNA for sequencing. After transportation to the laboratory, the animal was immediately dissected to isolate the mantle edge (including part of the the pallium and ventral surface), and the tissue was homogenised in Tri reagent (Sigma). RNA was extracted according to manufacturer's instructions, followed by preliminary quantification and quality check by

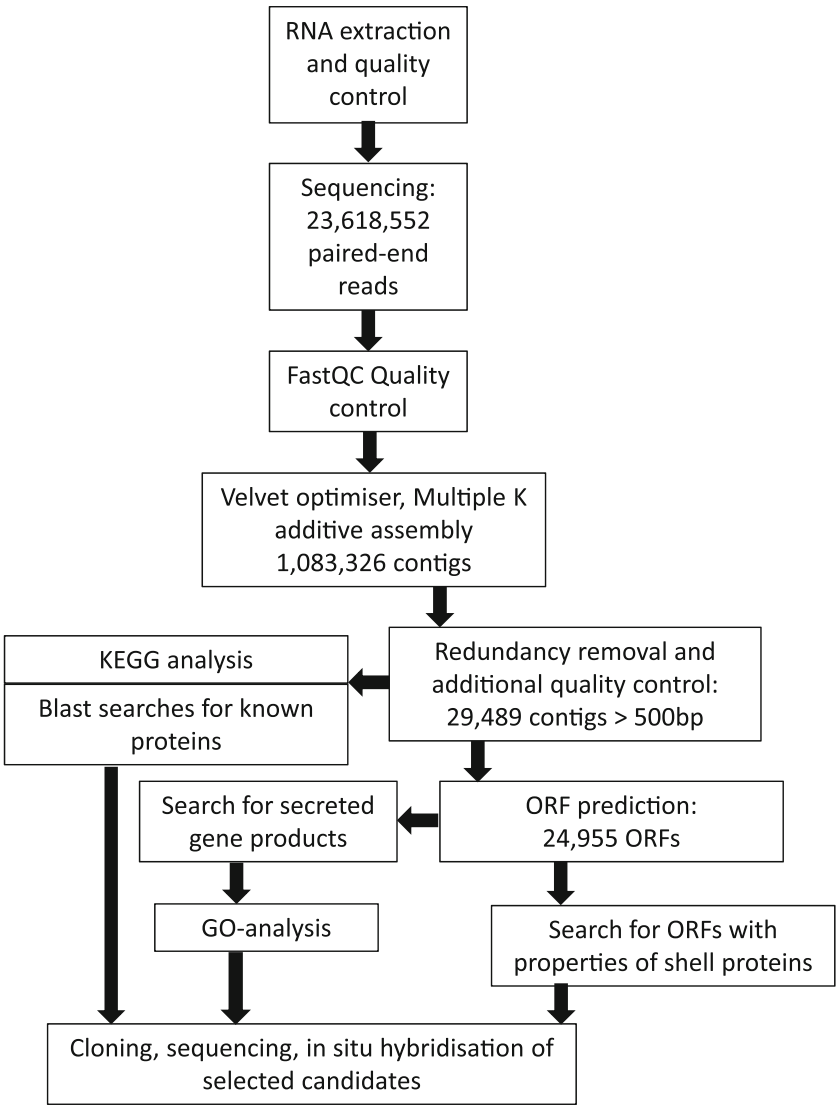

Fig. 2 Flow diagram depicting the assembly and analysis process, including numbers of sequences/contigs at key stages

spectrophotometry then further analysis via Bioanalyser (Agilent technologies). Sequencing was conducted by the Wellcome Trust Centre for Human Genetics sequencing service (Oxford, UK) using the Illumina GAII system. This generated 47,237,104 reads in total, comprising 23,618,552 paired-end reads of 51 bp from a library of average insert size 300 bp. We used FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/) to assess the quality of the dataset. A bias in GC content over the first 12 bases of the reads was revealed, consistent with that previously reported for Illumina data and thought to derive from biased reverse transcription hexamer primer binding (Hansen et al. 2010).

Assembly and initial bioinformatic analyses

We used the Velvet assembler (Zerbino and Birney 2008). We first ran assemblies for different kmer sizes, 21, 23, 25, 27 and 29, adopting the VelvetOptimiser script to optimise each assembly (http://bioinformatics.net.au/software.velvetoptimiser.shtml). Output statistics for these analyses can be seen in Table 1. We then employed the additive multiple k method of assembly combination (Surget-Groba and Montoya-Burgos 2010), followed by removal of redundant contigs by CD-hit (Li and Godzik 2006) and TGICL (Pertea et al. 2003) and finally removal of contigs shorter than 500 bp. This value was chosen as a compromise between the loss of data incurred when removing short contigs versus the utility of including only contigs of sufficient length to allow generation of a probe for in situ hybridisation. This left 29,652 contigs. Since the Illumina sample preparation and sequencing were carried out in parallel with human samples and cross contamination is a known possibility, we next asked whether any contigs had unusually high matches to human sequence. We used BLASTN to search all raw reads against the human genome. One hundred twenty-seven thousand three hundred twenty-nine (0.27 %) had significant matches. Significant matches could derive from genuine *P. vulgata* reads from genes highly conserved with human genes or from reads including simple sequence repeats. However, it could also indicate a low level of humans sequence contamination. Accordingly, we searched the 29,652 contigs against human genomic and EST data using BLASTN. This identified 163 contigs with $e<10-3$. This is a low threshold, so again, some of these will be very highly conserved coding sequences; however, some matched at a very high level suggesting contamination. To exclude any possibility of these affecting subsequent analyses, all 163 were removed from the dataset, leaving 29,489 contigs, available from the corresponding author on request. Since repetitive sequences might also generate significant matches between *P. vulgata* and human sequences, and identifying repeats as found in nacrein was one of our aims (see below), we also tested whether any of the excluded 163 contigs contained repetitive sequence. None had nacrein-like repeats.

Table 1 Summary of assembly statistics deriving from different kmer sizes implemented in Velvet with VelvetOptimiser

kmer size	Total Contigs	Contigs >100 bp	Contigs >1 kb	N50
21	233,903	87,674	4,359	313
23	273,587	98,817	3,972	275
25	248,645	98,108	3,695	279
27	169,693	80,809	4,278	379
29	157,498	79,559	3,745	367

Identification of transcripts relevant to shell construction

KEGG analysis was undertaken using the KAAS server (Moriya et al. 2007). The longest open-reading frame (ORF) for each contig was predicted using a perl script and ORFs of less than 50 amino acids were removed. This yielded 24,955 ORFs that were taken forward for further analysis. Gene ontology (GO) analysis was performed using Blast2GO (Conesa et al. 2005). We employed three methods to identify transcripts encoding proteins potentially relevant to shell construction. First we searched for genes with homology to known shell proteins derived from other species using TBLASTN. A list of these can be found in the Supplementary Table S1. We also used Hidden Markov Models (HMMs) based on Pfam domains for protein domains previously identified in known shell proteins. These were implemented in HMMER (Finn et al. 2011) with Pfam domains, PF01607 chitin binding, PF00194 carbonic anhydrase, PF00095 whey acidic protein, PF00092 von-Willebrand factor type-A, PF01391 collagen, PF00059 lectin C, PF00219 insulin growth factor binding protein and PF00779 BTK. Second we searched for tandem repeats using XSTREAM v1.73 (Newman and Cooper 2007). Third we searched for potential secreted proteins using SignalP (Bendtsen et al. 2004) and TargetP (Emanuelsson et al. 2000), excluding sequences also predicted by TMHMM (Emanuelsson et al. 2007) to encode a transmembrane domain.

For molecular phylogenetic analyses, we generated amino acid alignments using MAFFT (Katoh and Toh 2008) or ClustalX. Trees were built using the maximum likelihood method implemented in MEGA 5 (Tamura et al. 2011) with the WAG model and a gamma distribution with invariant sites and four discrete Gamma categories. Protein domain figures were created using DOG 2.0 (Ren et al. 2009). We also used the same datasets to construct Bayesian tress, using MrBayes v3.2.1 (Ronquist and Huelsenbeck 2003) with one million generations, discarding the first 250,00 trees when compiling summary trees and statistics.

RT-PCR and cloning of selected transcripts and in situ hybridisation

Primers used for RT-PCR are summarised in Table 2. cDNA was synthesised from *P. vulgata* mantle total RNA using Superscript III reverse transcriptase (Stratagene) according to manufacturer's instructions. Amplified products were cloned and sequenced to confirm their identity. In all cases, the sequence matched the predicted transcript with at least 99 % similarity. Accession numbers for these can be seen in Table 3. In situ hybridisation to embryos was conducted as described (Shimeld et al. 2010).

Results and discussion

Sequence generation, assembly and initial analysis

We generated 47,237,104 paired-end reads of 51 bp. Assembly optimisation, followed by redundancy removal, surveillance for contamination and removal of sequences <500 bp resulted in 29,489 contigs, of which 24,955 had an ORF of at least 50 amino acids. KEGG analysis of this entire dataset showed good coverage of core metabolic pathways, including 21 components of the citrate (TCA) cycle and 23 of glycolysis (data not shown). The assembly also reliably recovered some very long transcripts, for example Contig352 is 15,563 bp long, encoding an ORF of 5,141 amino acids that aligns throughout its length with the FAT-4 protocadherin (data not shown).

Candidate gene identification by homology search

Homology searches identified candidate homologues of seven proteins identified in other mollusc species as involved in shell construction (Table 3). AP24 was first identified in the abalone *H. rufescens* (Michenfelder et al. 2003) and has also been described in the sea hare *A. californica* (Heyland et al. 2011). *P. vulgata* C20395 contains a 143-amino acid ORF with homology to AP24 (Fig. S1). Carbonic anhydrase (CA) activity is associated with biomineralisation in molluscs (Freeman and Wilbur 1948), and CA domains have been identified in some shell proteins such as nacrein from the oyster *Pinctada fucata* (Miyamoto et al. 1996; Miyamoto et al. 2005). We searched for genes containing homologous domains and identified two *P. vulgata* contigs, C11947 and C16356, encoding ORFs of 300 and 698 amino acids, respectively (Table 3; Fig. S2). *Nacrein* also encodes an extensive Gly-Xaa-Asn repeat (where Xaa=Asp, Asn, or Glu); however, neither C11947 or C16356 encodes a similar repeat, although C16356 does contain low complexity sequence towards the carboxy end of the predicted ORF. LustrinA was identified from *H. rufescens* (Shen et al. 1997) and contains a Wey Acidic Protein (WAP) domain found in a number of extracellular protease inhibitors. A BLAST search of the *P. vulgata* nucleotide contig set identified C21697 as a significant hit, which encodes an ORF of 123 amino acids (Fig. S3). This ORF, however, lacks homology to LustrinA. Further examination of the sequence showed an additional ORF of 240 amino acids but lacking an initiation codon, which encoded three

Table 2 PCR primers used in this study

Contig	Annotation	Forward primer	Reverse primer
C20395	AP24	CCGATGTGGAATGATGGTTA	CGGCTCTAAAATTCTTGGGTA
C11947	Carbonic anhydrase	ATCAATGGTCGCGTTTATCC	GCCATTGCTTGTGTTGTTTG
C16356	Carbonic anhydrase	TGCGGCCCTTATTTTACTTG	CTGAACGGGTCGGAAGTTAT
C21697	LustrinA	AGACCTGTGCGCTCTGTATGT	GCACGACAGCATTTCTGATT
C21206	Perlustrin	TTTAACAGTTTGGAGGGGATTT	TCGTTTTATATAGATCCGGGTTTT
C4677	Pif-177/BSMP	ATGCCGCAGAACTAGAAGGA	TCTGGGTGATTTGTTGACCA
C14403	Dermatopontin	GTGACCGTGACGACAAAACA	CATGTGATCACCAACGCATC
C8761	Nacrein-like	ATCTTCTCCGTGTCGTCCAG	GGAACCCCGATCCTAACAAT
C17157	Nacrein-like	ATCACGAGGACCACAAGGAC	GTTCGTTGAAGTTGGGCATT
C10800	Nacrein-like	CAGGATTCCCAGGACAGAAA	CACGTGATCCCTTCAAACCT
C6561	SPARC	CTGGTGTGAGATGGCACTGT	CGATGATGACGTTGATGAGG
C6756	Periotrophin	AATGGCACAATGATCGTCAC	ACAGTATTCGCTGGACATGG
C21942	SLRP	TTTCTAAAGGCTTTCGGGTGT	CAATCTGCGAGGAGAACCAT
C11723	Follistatin	CTGGAGCACCAAGACAGACA	GACATGACGCCTTACAACGA
C22093	Follistatin-like	GCGCCTTTGATCTTATACC	CGGCATCTTTAACCTCGTTC

Table 3 Candidate contigs for which sequence was confirmed by RT-PCR and cloning; of the primer pairs shown in Table 1, only one (for C11723: Follistatin) did not amplify as predicted

Contig	Accession number	Identified by	Annotation	Length (bp)	ORF length	Reference
C20395	HE962372	Homology1	AP24	614	143	Michenfelder et al. (2003)
C11947	HE962373	Homology1	Carbonic anhydrase	1,102	300	Miyamoto et al. (1996)
C16356	HE962374	Homology1	Carbonic anhydrase	2,224	698	Miyamoto et al. (1996)
C21697	HE962375	Homology1	LustrinA	908	123 (240)	Shen et al. (1997)
C21206	HE962376	Homology1	Perlustrin	586	109	Weiss et al. (2000)
C4677	HE962377	Homology1	Pif-177/BSMP	1,955	564	Suzuki et al. (2011)
C14403	HE962381	Homology1	Dermatopontin	531	160	Marxen et al. (2003)
C8761	HE962378	Repeat2	Nacrein-like	4,804	1,452	Miyamoto et al. (2005)
C17157	HE962379	Repeat2	Nacrein-like	4,271	1,324	Miyamoto et al. (2005)
C10800	HE962380	Repeat2	Nacrein-like	2,576	814	Miyamoto et al. (2005)
C6561	HE962382	Secreted3	SPARC	2,264	267	Bradshaw (2009)
C6756	HE962383	Secreted3	Periotrophin	1,246	393	Tellam et al. (1999)
C21942	HE962384	Secreted3	SLRP	608	131	Kalamajski and Oldberg (2010)
C22093	HE962385	Secreted3	Follistatin-like	2,072	336	Bragdon et al. (2011)

Annotation is based on BLAST. 1Contig is homologous to a known shell protein from another mollusc species. 2Contig contains repeat region detected by XSTREAM. 3Contig detected as secreted but lacking a transmembrane domain. References relate to studies implicating homologous genes and/or proteins in biomineralisation. Two ORF sizes are given for LustrinA as discussed in the text

WAP domains (Fig. S3). Perlustrin was first identified in *H. laevigata* (Weiss et al. 2000), and homologous genes have been reported from other molluscs including *P. margaritifera* (Joubert et al. 2010). Perlustrins have homology to vertebrate insulin growth factor binding protein (IGFBP). We identified a *P. vulgata* contig, C21206, with homology to perlustrin, including an IGFBP domain (Fig. S4). The ORF encoded by this contig is 109 amino acids long, a bit larger than that encoded by *Haliotis discus* at 71 amino acids and the sequence reported from *H. laevigata* at 84 amino acids. Pif-177 was identified from the oyster *P. fucata* (Suzuki et al. 2009), and the protein it produces is cleaved to yield two fragments, Pif-97 (encoding a VWA domain and chitin-binding domain) and Pif-80 (encoding an aragonite binding domain). A related protein, BSMP, has also been reported from another bivalve, *Mytilus galloprovincialis*, and encodes multiple VWA domains, a chitin binding domain and a calcium carbonate binding domain (Suzuki et al. 2011). *P. vulgata* contig C4677 encodes a 564 amino acid ORF with homology to both proteins (Fig. S5). It includes a VWA domain and three chitin binding domains. Dermatopontin was extracted from the shell of the fresh water snail *B. glabrata* (Marxen et al. 2003; Sarashina et al. 2006) and is part of a family of extracellular matrix proteins of widespread phylogenetic distribution. *P. vulgata* C14403 encodes a 160-amino acid predicted ORF with homology to *B. glabrata* dermatopontin (Fig. S6).

Candidate gene identification by detection of sequence repeats

Several proteins known to form part of the protein matrix of mollusc shells contain highly repetitive sections, most notably the nacrein proteins as discussed above. The repeated sequence in nacrein is similar to that of the collagens and has been hypothesised to act as a regular structure for the nucleation of organised calcium carbonate deposition. Nacrein was first identified in *P. fucata* (Miyamoto et al. 1996; Miyamoto et al. 2005) and also contains a divergent CA domain. Our searches for CA-domain containing sequences as described above did not identify a nacrein candidate. Hence, we searched for sequences encoding repeated motifs. We did not identify any contigs specifically matching the Gly-Xaa-Asn repeat of nacrein; however, contigs C8761, C17157 and C10800 (Table 3) all contained extensive Gly-Xaa-Yaa repeats. Figure 3 shows the predicted ORF of C8761, including 338 repeat units. This ORF also includes a short sequence with limited similarity to the CA domain (Fig. 4); this was not detected by our initial BLAST searches but was weakly detected by HMM searches. Analysis of this sequence via the SMART domain annotation server also suggests the presence of a COLF1 (fibrillar collagen C-terminal) domain at the carboxy terminal, overlapping with the putative CA domain, as well as highlighting some similarity to collagens elsewhere in the sequence (Fig. 4). C17157 encodes an ORF of 1,324

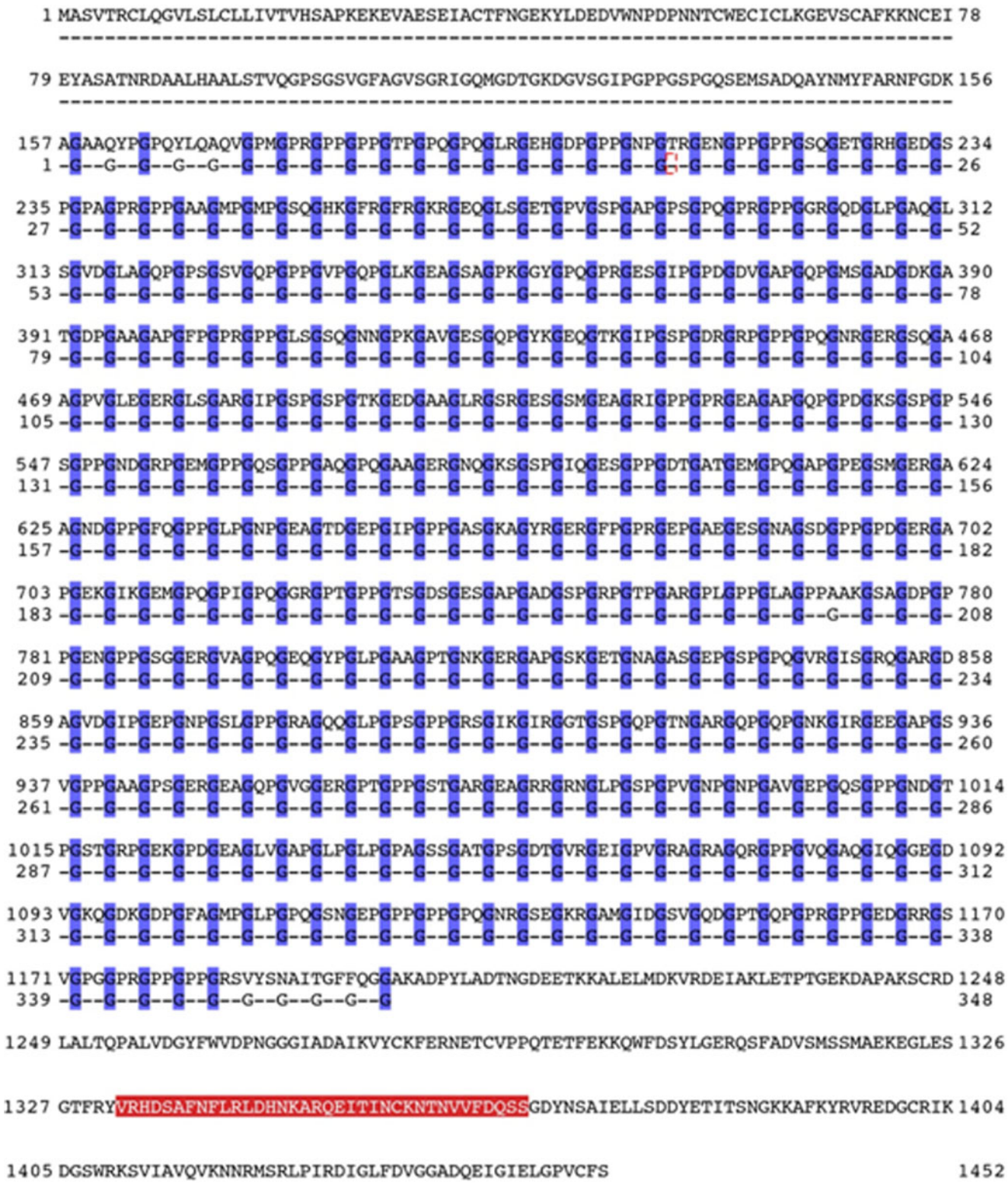

Fig. 3 Predicted ORF of C8761 (*top sequence*) aligned to the Gly-Xaa-Yaa repeat (*bottom sequence*). The short sequence blocked towards the carboxy terminal has similarity to the carbonic anhydrase domain

amino acids which includes 333 Gly-Xaa-Yaa repeats and, as for C8761, a COLF1 domain at the carboxy terminal (Fig. 4; Fig. S7). C10800 encodes an ORF of 814 amino acids which includes 148 Gly-Xaa-Yaa repeats

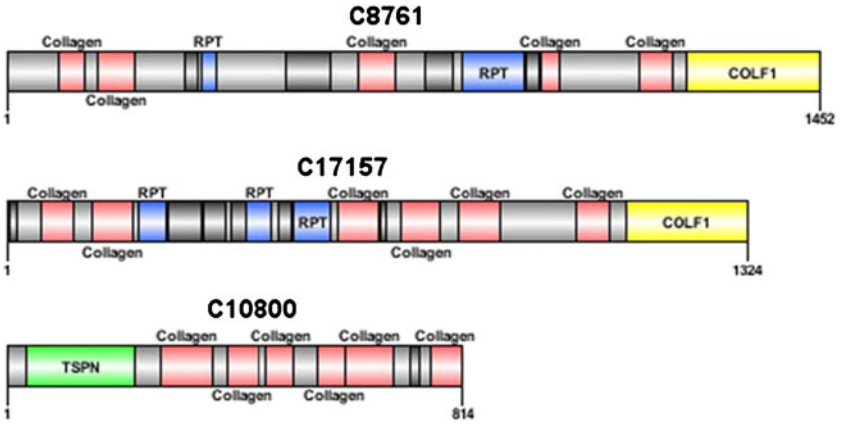

Fig. 4 Domain diagram of the three *P.vulgata* contigs encoding Gly-Xaa-Yaa similar to those in nacrein. Protein domains indicated are: PFam Collagen Domain (marked *Collagen*), repeat sequences (*RPT*), Fibrillar Collagen C-terminal Domain (*COLF1*) and the Thrombospondin N-terminal-like Domain (*TSPN*). Low complexity regions are indicated in *black*. *Numbers* indicate the length of the protein in amino acid residues. Domains were identified using SMART (Letunic et al. 2004)

and a TSPN (thrombospondin N terminal-like) domain at the N terminal (Fig. 4; Fig. S8).

We hence conclude that C8761 and C17157 are probably originally derived from collagen genes since they share both similar repeat structure and the COLF1 domain at the carboxy terminal. The weak match in C8761 to a CA domain is intriguing, and may suggest that the classical nacrein domain structure, with a Gly-Xaa-Yaa repeat and CA domain, evolved from an ancestral collagen gene. C10800 has a different domain structure, though this could still have evolved from an ancestral collagen gene. All three contigs are candidates for genes involved in providing simple repeated amino acid sequences for regulation of ordered calcium carbonate crystal deposition.

Candidate gene identification by detection of secreted proteins

In order to minimise the number of false positives, we adopted a stringent approach to identifying secreted proteins. SignalP and TargetP identified 344 and 1,698 ORFs, respectively, that were predicted to be secreted. We only considered those predicted by both methods and then excluded any predicted by TMHMM to include a transmembrane domain. This left 166 contigs, which we name the secretome and which were then annotated using Blast2GO (Table S2). Comparison of GO annotation between the secretome and the whole predicted proteome revealed

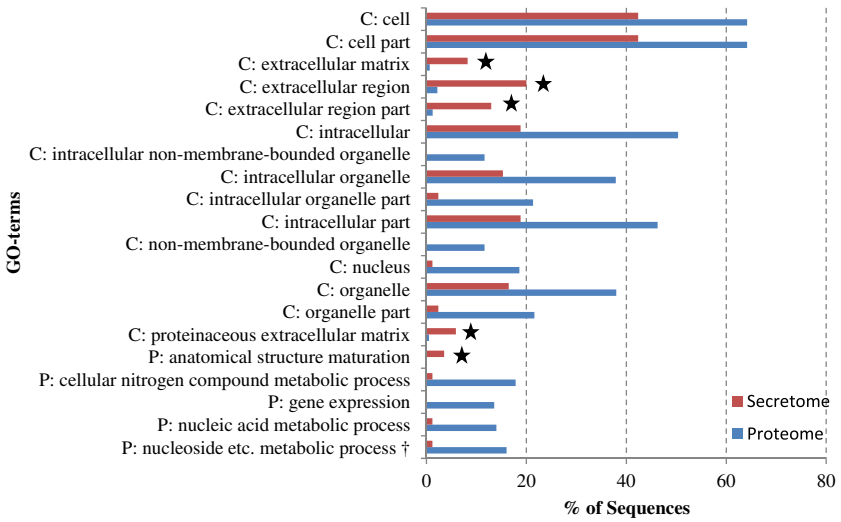

Fig. 5 GO-term distribution in the *P. vulgata* secretome (*top bar* for each GO term) compared to the entire predicted proteome (*bottom bar*). All GO terms that were under or overrepresented in the secretome are reported. A *letter* indicates the domain of a particular term (*C* cellular component, *P* biological process; none were found to be differentially distributed in the domain molecular function). Under or overrepresentation was statistically significant at the $\alpha<0.01$ level in all cases (Fisher's exact test with FDR applied). *Bars* indicate the percentage of sequences assigned to a particular GO term relative to the total number of sequences in that set. GO terms that are overrepresented in the secretome are indicated with a *star*. *Dagger* the full name of this GO term is "nucleobase, nucleoside, nucleotide and nucleic acid metabolic process"

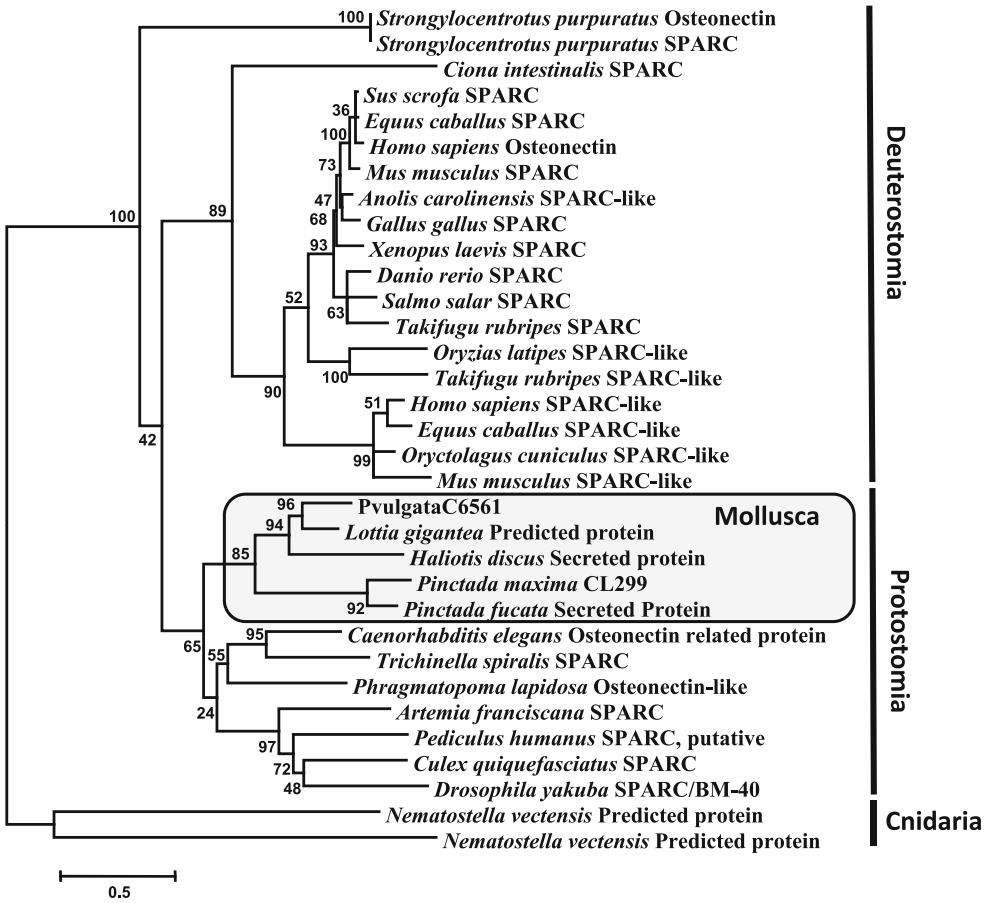

Fig. 6 Molecular phylogenetic analysis of SPARC amino acid sequences. *P. vulgata* contig 6561 is embedded in a group of mollusc SPARC orthologues (*boxed*). *Numbers adjacent to nodes* are percentage bootstrap support values. The *scale bar* represents number of substitutions per site. The tree is rooted with sequences from the sea anemone *Nematostella vectensis*, the earliest-diverging animal lineage represented in the dataset. Full details of accession numbers can be found in Supplementary Table S3. A Bayesian analysis also showing monophyly of mollusc SPARC sequences can be seen in Fig. S12

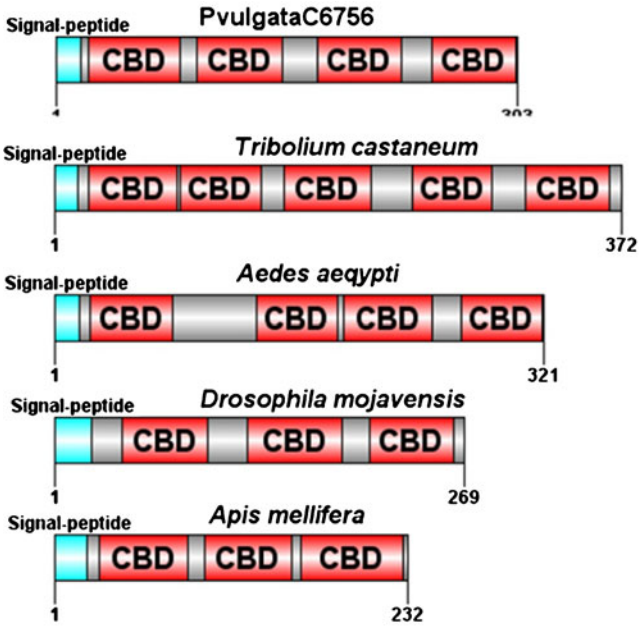

Fig. 7 Domain structure of *P. vulgata* C6756 and a selection of peritrophic matrix proteins. Numbers indicate length of the protein in amino acid residues. All have a signal-peptide followed by multiple chitin-binding domains (CBD). Chitin binding domains were identified through SMART (Letunic et al. 2004). Accession numbers are: *Tribolium castaneum*; ACY95486. *Aedes aegypti*; XP_001655685. *Drosophila mojavensis*; XP_002007741. *Apis mellifera*; NP_001165850

several terms significantly overrepresented in the secretome, the majority of which are associated with extracellular components (Fig. 5).

To further investigate these data, we selected five candidates for additional analysis on the basis of their annotation with GO terms associated with biomineralisation: contigs C6561, C6756, C11723, C21942 and C22093. C6561 encodes a SPARC homologue (Fig. S9). SPARC is also known as osteonectin in humans and is the most abundant non-collagenous organic matrix protein in vertebrate bone (Delany and Hankenson 2009), where it functions in the assembly and organisation of collagen fibrils (Bradshaw 2009). Sequences of mollusc SPARC homologues have been reported before in bivalves (Clark et al. 2010; Joubert et al. 2010) and additional mollusc sequences have been deposited in GenBank (Fig. S9). To further investigate the evolutionary relationships of SPARC sequences, we conducted a molecular phylogenetic analysis (Fig. 6). This showed a well-supported clade of molluscs SPARC sequences related to vertebrate SPARC, and we hence conclude C6561 is the *P. vulgata* orthologue of SPARC.

C6756 (Fig. S10) encodes a 393 amino acid ORF with similarity to the periotrophic matrix (PM) protein family (Tellam et al. 1999). These encode at least one chitin binding domain (CBD), usually repeated multiple times (Fig. 7).

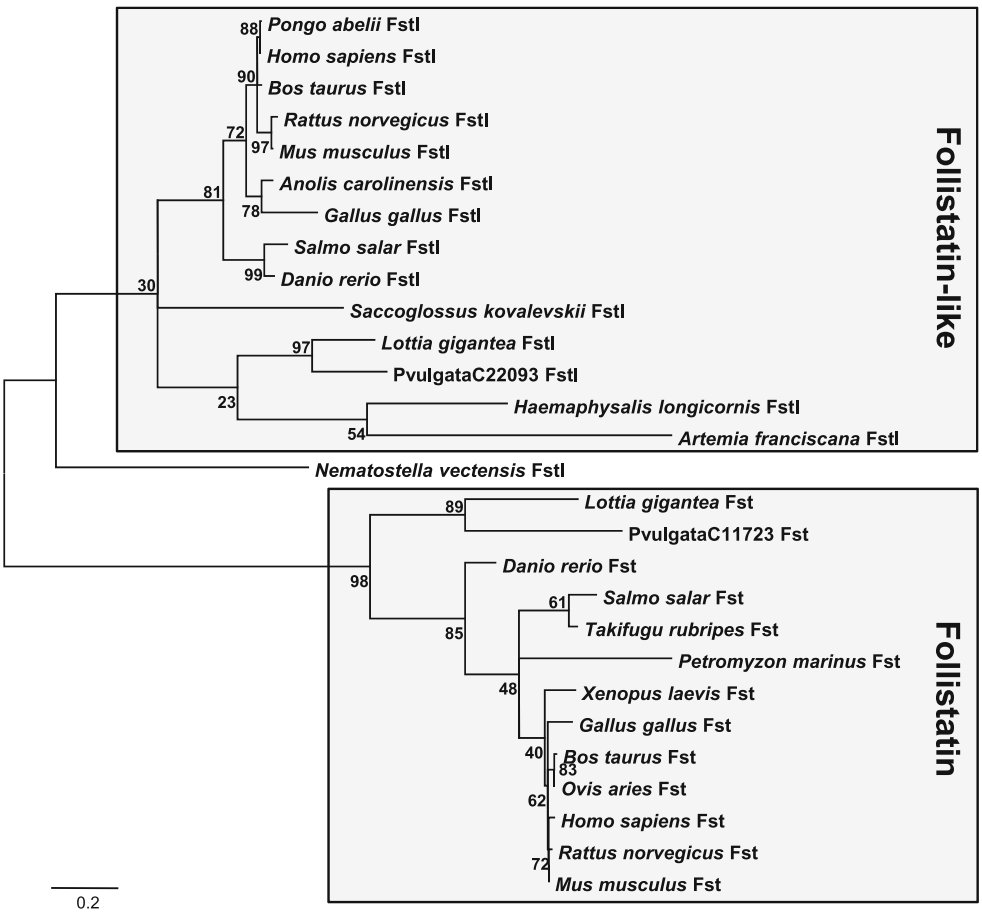

Fig. 8 Molecular phylogenetic analysis of Follistatin (Fst) and Follistatin-like (Fstl) protein sequences. C11723 and C22093 are embedded in the groups of Fst and Fstl proteins, respectively (*boxed*). Both sequences cluster with other mollusc sequences showing similarity to Fst/Fstl, and in the case of Fstl also with other protostomes. *Numbers adjacent to nodes* are percentage bootstrap support values. The *scale bar* represents number of substitutions per site. We were not able to identify a more distantly related homologue to use as an effective outgroup, hence the phylogeny is depicted as mid-point rooted to show the Fst and Fstl gene groups. The *Nematostella* gene named Fstl remains outside the Follistatin-like box; as this is without an outgroup, we cannot formally determine its position. Full details of accession numbers can be found in Supplementary Table S3. A Bayesian analysis also showing monophyly of Fst and Fstl sequences can be seen in Fig. S12

PM proteins are thought to cross link chitin fibrils in the insect gut (Tellam et al. 1999); and since chitin is part of the organic component of mollusc shells, it may be fulfilling the same role here. A protein with a single chitin binding domain has been reported from *H. asinina* (Jackson et al. 2010), however this lacked a signal peptide, hence *P. vulgata* C6756 is to our knowledge the only full PM protein reported to date from a mollusc.

C21942 and C9544 (Fig. S11) encode proteins with similarity to decorin, aspirin and biglycan, all members of the class I small leucine-rich repeat proteoglycans (SLRPs). SLRPs are secreted matrix proteins with multiple leucine-rich repeats (LRRs) and are involved in collagen fibrillogensis and its subsequent mineralisation (Kalamajski and Oldberg 2010). Like other SLRPs, C21942 encodes a signal peptide; however, it only encodes two LRRs, whereas other SLRPs may encode many more. It is hence possible C21942 is truncated at the carboxy terminal. C9544 encodes an ORF of 604 amino acids with a signal peptide and nine LRRs.

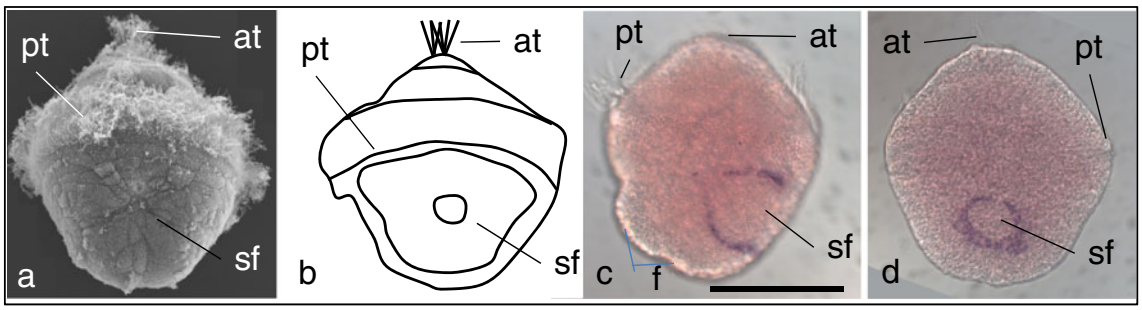

Fig. 9 **a** Scanning electron micrograph of a dorsal view of a 21-h trochophore. The apical tuft (*at*) is shown, as is the prototroch (*pt*) and shell field (*sf*). At this stage, the shell field has started to invaginate. Image courtesy of Dr. Helen Thompson. **b** Schematic diagram of a dorsal view of a 21-h trochophore, modelled on **a**. **c** Expression of C6561 (SPARC) in a ring around the shell field. This larva is slightly rotated onto its right side, such that the ventrally sited foot (*f*) is also visible. **d** Expression of C6757 (PM), also in a ring around the shell field. The *scale bar* on **b** is 100 μm and applies to all images

C11723 and C22093 encode proteins with similarity to respectively Follistatin (Fst) and Follistatin-like (Fstl). Follistatin is a well-known antagonist of the Bone Morphogenetic Protein 2/4 (BMP2/4), a class of proteins that regulates a large range of developmental processes, including vertebrate bone and skeleton (re)generation (Bragdon et al. 2011). Although the mollusc BMP2/4-homologue *dpp-BMP2/4* has been studied quite extensively (Iijima et al. 2008; Kin et al. 2009; Shimizu et al. 2011; Nederbragt et al. 2002; Koop et al. 2007), neither Fst nor Fstl has previously been described in molluscs. However, other sequences similar to Fst and Fstl have been deposited in GenBank, and phylogenetic analyses confirm the expected evolutionary relationship of C11723 and C22093 to these and other Fst and Fstl sequences (Fig. 8)

Evaluating contig integrity by RT-PCR

Since the assembly of short-read data is a relatively new technique, we sort to evaluate assembly predictions by directly amplifying and cloning predicted sequences. We designed primers to all 15 contigs shown in Table 2 and attempted to amplify them via RT-PCR from *P. vulgata* mantle RNA. Fourteen out of 15 were successfully amplified, producing bands of the correct predicted size (data not shown), and were then cloned and sequenced to confirm their identity. In all the cases, the sequence matched the predicted transcript with at least 99 % similarity. This demonstrates that, at least for these 14 contigs, the assembly process has accurately captured the mRNA sequence. We conclude that our sequencing and assembly process is an effective predictor of genuine mRNA sequence data and hence can provide a rapid and reliable route to cloned genes for further analyses.

Expression analysis of selected genes

Our searches discussed above were designed to identify genes likely to be involved in shell construction. However, the transcriptome was by necessity developed from a composite tissue, including the pallium, and also other tissues such as muscle, tentacle and pallial gill (Fig. 1c–g). Thus the genes in the transcriptome are not necessarily expressed by cells involved in shell construction. We hence sought to directly test if the cloned genes were genuinely expressed in cells related to shell construction. Gene expression protocols have yet to be developed for adult *P. vulgata* tissue; however, they have been established for embryos and larvae (Shimeld et al. 2010), including the stages where the shell field and shell gland are prominent (Fig. 9a, b).

We used the clones derived for C6561 (SPARC) and C6756 (PM) to generate sense and antisense probes for in situ hybridisation on *P. vulgata* larvae. The results showed staining around the developing shell field, with both genes expressed in a ring (Fig. 9b, c), reminiscent of other shell-related genes in *P. vulgata* (Nederbragt et al. 2002). This shows that, at least for the genes and stage tested, expression is confined to tissues involved in shell construction. We conclude that our analyses have identified genes that in *P. vulgata* larvae are expressed in cells involved in building the shell. This implies our original transcriptome should provide a source of other genes involved in this process.

Conclusions

We have shown that a single lane of Illumina GAII short read sequence data can provide extensive insight into the transcriptome of a non-model species. Analysis of these data has identified a number of genes encoding proteins likely to be involved in shell construction in *P. vulgata*, and these were amplified from *P. vulgata* mRNA with a high success rate, illustrating the general effectiveness of the assembly process. In two cases, we have also demonstrated specific expression in the developing shell gland of early *P. vulgata* trochophore embryos. These data provide an extensive resource for future investigations into the biology of *P. vulgata*.

Acknowledgements We thank The Elizabeth Hannah Jenkinson fund for financial support for the sequencing, and the High-Throughput Genomics Group at the Wellcome Trust Centre for Human Genetics in Oxford for conducting the Illumina GAII sequencing. We also thank Dr. Paul Naylor for Fig. 1a, and Dr. Helen Thompson for Fig. 9a. SG was supported by the EU Lifelong Learning Programme. GDAW was supported by the Huygens Scholarship Programme. PG was supported by an EPSRC studentship.

References

Bendtsen JD, Nielsen H, Von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340:783–95

Bradshaw AD (2009) The role of SPARC in extracellular matrix assembly. J Cell Commun Signal 3:239–46

Bragdon B, Moseychuk O, Saldanha S, King D, Julian J, Nohe A (2011) Bone morphogenetic proteins: a critical review. Cell Signal 23:609–20

Clark MS, Thorne MA, Vieira FA, Cardoso JC, 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 11:362

Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–6

Delany AM, Hankenson KD (2009) Thrombospondin-2 and SPARC/osteonectin are critical regulators of bone remodeling. J Cell Commun Signal 3:227–38

Donachy JE, Drake B, Sikes CS (1992) Sequence and atomic-force microscopy analysis of a matrix protein from the shell of the oyster *Crassostrea virginica*. Mar Biol 114:423–428

Emanuelsson O, Brunak S, Von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2:953–71

Emanuelsson O, Nielsen H, Brunak S, Von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300:1005–16

Fang D, Xu G, Hu Y, Pan C, Xie L, Zhang R (2011) Identification of genes directly involved in shell formation and their functions in pearl oyster, Pinctada fucata. PLoS One 6:e21860

Feldmeyer B, Wheat CW, Krezdorn N, Rotter B, Pfenninger M (2011) Short read Illumina data for the de novo assembly of a non-model snail species transcriptome (Radix balthica, Basommatophora, Pulmonata), and a comparison of assembler performance. BMC Genomics 12:317

Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39:W29–37

Freeman JA, Wilbur KM (1948) Carbonic anhydrase in molluscs. Biol Bull 94:55–9

Fuchigami T, Sasaki T (2005) The shell structure of the recent Patellogastropoda (Mollusca: Gastropda). Palaeontological Research 9:143–168

Hansen KD, Brenner SE, Dudoit S (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res 38:e131

Heyland A, Vue Z, Voolstra CR, Medina M, Moroz LL (2011) Developmental transcriptome of Aplysia californica. J Exp Zool B Mol Dev Evol 316B:113–34

Hodgson AN, Le Quesne WJF, Hawkins SJ, Bishop JDD (2007) Factors affecting fertilization success in two species of patellid limpet (Mollusca: Gastropoda) and development of fertilization kinetics models. Mar Biol 150:415–426

Iijima M, Takeuchi T, Sarashina I, Endo K (2008) Expression patterns of engrailed and dpp in the gastropod Lymnaea stagnalis. Dev Genes Evol 218:237–51

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 4:40

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

Jackson DJ, Worheide G, Degnan BM (2007) Dynamic expression of ancient and novel molluscan shell genes during ecological transitions. BMC Evol Biol 7:160

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 11:613

Kalamajski S, Oldberg A (2010) The role of small leucine-rich proteoglycans in collagen fibrillogenesis. Matrix Biol 29:248–53

Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform 9:286–98

Kin K, Kakoi S, Wada H (2009) A novel role for dpp in the shaping of bivalve shells revealed in a conserved molluscan developmental program. Dev Biol 329:152–66

Kleypas J, Feely R, Fabry V, Langdon C, Sabine C, Robbins L (2006) Imapcts of ocean acidification on coral reefs and other marine calcifiers. Report of the workshop sponsored by NSF, NOAA and USGS

Kniprath E (1981) Ontogeny of the molluscan shell field—a review. Zoologica Scripta 10:61–79

Koop D, Richards GS, Wanninger A, Gunter HM, Degnan BM (2007) The role of MAPK signaling in patterning and establishing axial symmetry in the gastropod Haliotis asinina. Dev Biol 311:200–12

Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, Bork P (2004) SMART 4.0: towards genomic data integration. Nucleic Acids Res 32:D142–4

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–9

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 8:54

Marin F, Amons R, Guichard N, Stigter M, Hecker A, Luquet G, Layrolle P, Alcaraz G, Riondet C, Westbroek P (2005) Caspartin and calprismin, two proteins of the shell calcitic prisms of the Mediterranean fan mussel Pinna nobilis. J Biol Chem 280:33895–908

Marin F, Luquet G (2004) Molluscan shell proteins. CR Palevol 3:469–492

Marxen JC, Nimtz M, Becker W, Mann K (2003) The major soluble 19.6 kDa protein of the organic shell matrix of the freshwater snail Biomphalaria glabrata is an N-glycosylated dermatopontin. Biochim Biophys Acta 1650:92–8

Michenfelder M, Fu G, Lawrence C, Weaver JC, Wustman BA, Taranto L, Evans JS, Morse DE (2003) Characterization of two molluscan crystal-modulating biomineralization proteins and identification of putative mineral binding domains. Biopolymers 70:522–33

Milan M, Coppe A, Reinhardt R, Cancela LM, Leite RB, Saavedra C, Ciofi C, Chelazzi G, Patarnello T, Bortoluzzi S, Bargelloni L (2011) Transcriptome sequencing and microarray development for the Manila clam, Ruditapes philippinarum: genomic tools for environmental monitoring. BMC Genomics 12:234

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 USA 93:9657–60

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. Zoolog Sci 22:311–5

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 35:W182–5

Moroz LL, Edwards JR, Puthanveettil SV, Kohn AB, Ha T, Heyland A, Knudsen B, Sahni A, Yu F, Liu L, Jezzini S, Lovell P, Iannucculli W, Chen M, Nguyen T, Sheng H, Shaw R, Kalachikov S, Panchin YV, Farmerie W, Russo JJ, Ju J, Kandel ER (2006) Neuronal transcriptome of Aplysia: neuronal compartments and circuitry. Cell 127:1453–67

Nederbragt AJ, Van Loon AE, Dictus WJ (2002) Expression of Patella vulgata orthologs of engrailed and dpp-BMP2/4 in adjacent domains during molluscan shell development suggests a conserved compartment boundary mechanism. Dev Biol 246:341–55

Newman AM, Cooper JB (2007) XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. BMC Bioinforma 8:382

Orr JC, Fabry VJ, Aumont O, Bopp L, Doney SC, Feely RA, Gnanadesikan A, Gruber N, Ishida A, Joos F, Key RM, Lindsay K, Maier-Reimer E, Matear R, Monfray P, Mouchet A, Najjar RG, Plattner GK, Rodgers KB, Sabine CL, Sarmiento JL, Schlitzer R, Slater RD, Totterdell IJ, Weirig MF, Yamanaka Y, Yool A (2005) Anthropogenic ocean acidification over the twenty-first century and its impact on calcifying organisms. Nature 437:681–6

Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J, Quackenbush J (2003)

TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19:651–2

Ren J, Wen LP, Gao XJ, Jin CJ, Xue Y, Yao XB (2009) DOG 1.0: illustrator of protein domain structures. Cell Research 19:271–273

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–4

Ruppert EE, Fox RS, Barnes RD (2004) Invertebrate zoology. Thomson learning Inc., Belmont

Sarashina I, Yamaguchi H, Haga T, Iijima M, Chiba S, Endo K (2006) Molecular evolution and functionally important structures of molluscan Dermatopontin: implications for the origins of molluscan shell matrix proteins. J Mol Evol 62:307–18

Shen X, Belcher AM, Hansma PK, Stucky GD, Morse DE (1997) Molecular cloning and characterization of lustrin A, a matrix protein from shell and pearl nacre of Haliotis rufescens. J Biol Chem 272:32472–81

Shimeld SM, Boyle MJ, Brunet T, Luke GN, Seaver E (2010) Clustered Fox genes in molluscs and annelids and the evolution of mesoderm. Dev Biol 340:234–248

Shimizu K, Sarashina I, Kagi H, Endo K (2011) Possible functions of Dpp in gastropod shell formation and shell coiling. Dev Genes Evol 221:59–68

Smith FGW (1935) The development of Patella vulgata. Phil Trans Royal Soc B 225:95–125

Surget-Groba Y, Montoya-Burgos JI (2010) Optimization of de novo transcriptome assembly from next-generation sequencing data. Genome Res 20:1432–40

Suzuki M, Iwashima A, Tsutsui N, Ohira T, Kogure T, Nagasawa H (2011) Identification and characterisation of a calcium carbonate-binding protein, blue mussel shell protein (BMSP), from the nacreous layer. ChemBioChem 16:2478–2487

Suzuki M, Saruwatari K, Kogure T, Yamamoto Y, Nishimura T, Kato T, Nagasawa H (2009) An acidic matrix protein, Pif, is a key macromolecule for nacre formation. Science 325:1388–90

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–9

Tellam RL, Wijffels G, Willadsen P (1999) Peritrophic matrix proteins. Insect Biochem Mol Biol 29:87–101

Veis A (2003) Mineralization in organic matrix frameworks. Biomineralization 54:249–289

Weiss IM, Kaufmann S, Mann K, Fritz M (2000) Purification and characterization of perlucin and perlustrin, two new proteins from the shell of the mollusc Haliotis laevigata. Biochem Biophys Res Commun 267:17–21

Wilt FH (2005) Developmental biology meets materials science: morphogenesis of biomineralized structures. Dev Biol 280:15–25

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–9