The shell proteome of the deep-sea barnacle Bathylasma hirsutum and the convergency in barnacle and molluscan shell proteins

Yu-Tao Xu
Shenzhen University
James Taylor
German Centre for Marine Biodiversity Research (DZMB), Universität Hamburg
Hao-Cheng Liu
Shenzhen University
Niklas Dreyer
Harvard University
Qian-Qian Cho
Shenzhen University
Yu Zhang
Shenzhen University
Shi-Feng Guo
Chinese Academy of Sciences
Saskia Brix
German Centre for Marine Biodiversity Research (DZMB), Universität Hamburg
Yue Him Wong (timwong@szu.edu.cn)
Shenzhen University

Research Article

Keywords:

Posted Date: September 26th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3287643/v1

Abstract

Background

As a group of sessile crustaceans that were being misclassified as mollusks by Carl Linnaeus, barnacles produce calcareous shell plates which, in most species, are permanently attached to the substratum. As biomineralization has independently evolved in multiple marine invertebrate taxa, a key question is how biomineralization has driven the evolution of genetic toolkits underlying shell formation. Here, we explore the shell proteome of the deep-sea acorn barnacle *Bathylasma hirsutum* (Hoek, 1883) using an integrated transcriptomic-proteomic approach and compare the properties of barnacle shell proteins with molluscan shell matrix proteins.

Results

We identified 31 *B. hirsutum* barnacle shell proteins (BSPs), including a series of key biomineralization proteins, such as carbonic anhydrase and C-type lectin. More than half of barnacle specific shell proteins (BSSPs) exhibit unknown functions. The amino acid composition of these BSSPs were biased toward A, D, E, G, S, P and Q, and were acidic and hydrophilic. Almost all BSSPs were detected with repetitive low complexity domains. Similar to molluscan shell matrix proteins, RLCDs in D-, and E-rich BSSPs constituted up to 50% amino acid of the whole protein. RLCDs in Q-rich proteins also exhibited similarity to a Q-rich abalone shell matrix protein and an insect cuticle protein.

Conclusion

From the *B. hirsutum* shell proteome, certain proteins such as carbonic anhydrase, C-type lectin, and peroxidase were implicated in shell formation or protein cross-linking across sessile invertebrate taxa. Despite the lack of sequence homology, D- and Q-rich BSSPs share similar features with molluscan shell matrix proteins in sequence redundancy, amino acid bias and thereby protein isoelectric point and hydropathy. Such convergence may reflect that similar selection pressures shape the molecular evolution of biomineralization and shell formation genes in marine invertebrates.

Introduction

In a letter to a friend, Charles Darwin wrote “I hate a Barnacle as no man ever did before, not even a Sailor in a slow-sailing ship.” [1], in order to express his frustration to classify this group of shelled sessile creatures, which nowadays are known as members of the phylum crustacea. Darwin was not the only one who felt the frustration. Carl Linnaeus in mid-1700s classified barnacles as a group of mollusks for the presence of mineralized shell plates for the protection of soft visceral mass [2]; in early-1800s, paleontologist George Cuvier noted that, beneath the hard shell, the barnacle body shares some common features with what he proposed the Articulata, a group that according to him include crustaceans and insects [3]. However, Cuvier still placed barnacles within mollusks in his famous book publication *Le règne animal* [3]. Darwin studied the life cycle of barnacle and demonstrated that barnacle nauplius larva is analogous to crustacean nauplius larva. But it still took him 8 years to conclude the previous barnacle classification was wrong and should be grouped with crustacea [4].

The presence of the calcified shell confused the big taxonomist figures when considering the classification of barnacles. From a functional perspective, barnacle shells serve the same function as bivalves or gastropod shells but is ontogenetically different. Barnacles are mainly found on the surface of any submerged objects in the marine environment, coastal rocks in the intertidal zone, floating logs, to the shell or skin of marine nektons, and even the deep-sea [5, 6]. Typical acorn barnacle shells compose of operculum plates (OPs) and wall plates (WPs). In some species, a calcified basal plate (BP) is located between the barnacle and the adhesive interface, while some species only have a chitinous membranous base [7]. The wall plates wrap around the barnacle body. The distal end of the wall plate is sealed by operculum plates, which are the only movable plates that serve as the shutter of the internal cavity. Hence, barnacle shells differ from molluscan shells in which the former typically made of multiple articulated shell plates while the latter mostly composed on one or two single shell pieces (one shell mass in gastropod and two pieces of shell in bivalves). It is generally believed that calcification of shell in both taxa is merely evolutionary convergence. But does shell biomineralization in both taxa exhibit significant similarity?

In fact, biomineralization is an extremely widespread phenomenon, even for some of the oldest living organisms on Earth, bacteria. In microorganisms, biomineralization is the chemical alteration of the environment by microbial activity resulting in the deposition of minerals [8]. Mineralization in different bacteria firstly proceeds as electrostatic interactions between metal cation and anion sites that leads to nucleation of mineral salt crystal, followed by solute deposition at the nucleation sites [9]. Unlike microorganisms, which have no real control over the mineralization process, many marine calcifiers, from unicellular forams (Foraminifera) to diverse metazoans such as echinoderms, polychaetes, and mollusks, can form complex endo- or exoskeletal structures. Marine calcifiers play an important role in the global carbon cycle, and are often the focus in paleomarine biology and paleoclimate studies [10]. Yet, the underling molecular machinery driving calcium carbonate ($CaCO_3$) biomineralization in marine calcifiers remains surprisingly poorly understood. It is believed that $CaCO_3$ biomineralization involves 1) $CaCO_3$ supersaturation at specific sites to allow simultaneous calcification and 2) reduction of crystallization inhibitors such as Mg^{2+} [11, 12]. Marine calcifiers such as Foraminifera increase intracellular CO_3^{2-} concentration by endocytosis of sea water and then transport to the mineralized site [13]. The organic component in shells, on the other hand, controls the magnesium content of the parent fluid to affect the calcification rate, crystal orientation, and crystal shape and thus the overall calcification process [11, 12].

The components of the organic shell matrix are secreted by shell-making tissue. It is this layer of organic substances that comes in closest contact with inorganic minerals during mineralization. Biochemical and micromorphological studies have revealed that that matrix plays an irreplaceable role in biologically controlled mineralization [14]. The organic shell matrix includes proteins, lipids, and polysaccharides [15] but the exact composition varies between taxa. The hard parts of matrix mainly compose of proteins and glycoproteins [16]. It was hypothesized that matrix protein can interact with crystals just like the model of protein-protein interaction or antigen-antibody interaction [17]. Here, the unit of crystal is a single “cell” while the unit of protein is an amino acid residue. As the matrix proteins fold, some residues are exposed to the external environment. The side chain of the exposed amino acid residue can then interact with the crystal and eventually forms a highly stable crystal lattice [17]. Hence, the primary structure of shell matrix proteins is critical to the shell formation function.

Shell matrix proteins have been recovered from and studied in different marine calcifying taxa, including sponges, mollusks, corals, brachiopods, and echinoderms [18–22]. From these studies, certain conserved functional proteins such as carbonic anhydrase have been recovered from the shell making tissues or direct from the soluble organic shell matrix [23]. However, many shell matrix proteins are lineage specific. For instance, in mollusks, Asprich, a family of shell matrix protein which showed heavy bias toward Aspartic acid (D) [24], is one of the most well-studied acidic shell matrix proteins that are lineage (mollusk) specific. It has been demonstrated that Asprich could induce and stabilize amorphous calcium carbonate (ACC) particles and inhibit their uncontrolled crystallization [24]. Other acidic D-rich matrix proteins have also been identified in coral skeleton organic matrices [25]. These findings beg the question whether there actually exists a biomineralization genetic toolkit that is conserved from cnidarians to higher metazoans, or the evolution of D-rich proteins in independent taxa is the result of convergency. Yet, to answer this question, broader taxonomic coverage of shell matrix protein is needed.

Barnacle shells are mostly calcitic, but the ratio of organic matrix and inorganic biomineral varies between species and the type of shell plates [26]. The growth of wall plates occurs at the junction of wall plates and basal plates, where a layer of cell containing densely packed calcium ion vacuole was observed by using TEM and the release of the cellular contents was observed in X-ray microanalysis [27]. Moreover, chitininogenous cells, cuticle-secreting cells and matrix-secreting cells were also observed at the growing edge [28]. Despite being a well-known marine calcifier, studies of barnacle shell proteins are limited. There is thus far only one study reporting the shell proteome of an intertidal acorn barnacle species *Amphibalanus amphitrite* (Darwin, 1854) [29]. Although D-rich protein was also recovered from the acetic acid shell extract, the protein sequence was highly fragmented [29]. In this study, we extracted shell proteins from *Bathylasma hirsutum* (Hoek, 1883), an acorn barnacle species collected from the mid-Atlantic semi-bathyal region ranging in depth from 600 to 1300 m. Proteomic and transcriptome analyses were performed to acquire shell proteins specifically expressed in the shell followed by analysis of the different biochemical properties of the recovered barnacle shell proteins, with specific focuses on the recovery of D-rich shell proteins and other barnacle specific shell proteins with special biochemical characteristics. The recovered barnacle shell proteins were than compared with molluscan counterparts to extract novel insights in amino acid composition, motif sequences and biochemical properties such as protein isoelectric points and hydropathy in the shell proteins from two independent marine calcifying taxa.

Materials and methods

Barnacle sample preparation

Bathylasma hirsutum specimens (Fig. 1A) were collected live in a basin at 978 m near a hydrothermal vent in the mid-Atlantic semi-bathyal region using ROV PHOCA, GEOMAR. As soon as the barnacle specimens were brought to the deck of the research vessel, they were submerged in RNAlater overnight and then kept at -80 ℃. Barnacle specimens were then transported to the lab. Five frozen *B. hirsutum* specimens were dissected. Prior to dissection, the sample was thawed at room temperature. Dissection was conducted using a sterilized razor blade. The prosoma was first separated from the wall plates and the operculum plates. For each *B. hirsutum* specimen, the cirri, maxillopods, testis were dissected. The rim and the mantle tissue underneath the operculum plate (opep), which were densely pigmented with dark purple color and was attached to the edge of the operculum plate, were also isolated.

Transcriptome sequencing, de novo assembly and functional annotation

RNAseq libraries were prepared from each total RNA sample using Illumina TruSeq RNA sample Prep Kit v2 and sequenced with the Illumina GAIIx platform at PE150 mode. The raw Illumina reads were deposited in the NCBI Short Read Archive with the accession numbers listed in Table S1.

De novo assembly and annotation were performed on a 20 cores Dell server with 500GB RAM. All raw read data from different samples were concatenated according to their read direction and the merged read files were submitted to Trinity v2.6.2 [30] with the command "Trinity –seqType fq –max_memory 300G –trimmomatic –left ${DATADIR}/R1.fq –right ${DATADIR}/R2.fq –output ${OUTDIR}/Trintiy_${DATE} –CPU 24 –min_contig_length 200 –min_kmer_cov 2". The resulting transcriptome assembly generated from *B. hirsutum* were collapsed using the software CD-HIT-EST [31] at 95% identity. The program TransDecoder was used to extract the longest open reading frame and translate the ORFs into the coding protein sequences. Finally, CD-HIT was used to cluster protein sequences with 98% Identity. The program BUSCO (v.5.2.2) was used to access the completeness of both transcriptome assemblies [32]. The database Arthropoda_mdb10 was selected as the BUSCO background lineage. The translated protein database of *T. formosana* was annotated with the non-redundant (nr) and SwissProt protein database using BLASTX with an e-value cutoff at 1e-5. GO annotation was also conducted.

Transcript abundance estimation and differential expression analysis

The high-quality, trimmed read data from each species were mapped to the corresponding reference transcriptome assembly using the perl script Align_and_estimate_abundance.pl (in the "util" module in the Trinity package), with Salmon specified as the estimation method. The abundance of the transcript sequences was estimated for each sample using the program Salmon [33]. The transcript expression levels in each sample were normalized to Transcript Per Million Reads (TPM).

Differential expression (DE) analysis was performed using the perl script run_DE_analysis.pl (in the "Analysis" module in the Trinity package) with EdgeR [34] specified as the program to be used. For DE analysis in *B. hirsutum*, transcript expression level in the cirri, maxillopod, the tissue layer on the operculum plate, and the rim were compared. The differential expression threshold was set at 4-fold (16 times) and the FDR value set at $p < 0.001$.

Shell processing, decalcification, and protein extraction

The operculum plate (Fig. 1b) and the wall plate (Fig. 1c) of the five dissected *B. hirsutum* specimens were used to prepare the barnacle shell samples. The barnacle shell samples were cleaned with a clean toothbrush at first to remove any contamination of organisms fouled on the shell surface. The shell samples were then soaked in 10% bleach overnight to further remove organic impurities on shell surface. The shell sample was then rinsed with MilliQ water for 30 mins and cleaned again with a brush. The shells were dried and ground to fine powder.

Ten grams of shell powder from different shell parts (operculum and wall plates) were weighed and added to 100 ml of 15% acetic acid (v/v). Decalcifying shell samples were kept in the refrigerator of 4 ℃ overnight. The supernatant was retrieved as the shell acid soluble fraction and was transferred into a 15 mL centrifugal device with 3000 Dalton molecular weight cutoff filter (Ultracel, Amicon, USA) to concentrate the acid soluble shell proteins. They were centrifuged at 4000 rpm until the acid soluble fraction was

concentrated to a volume lower than 500 µL. After concentration, the shell acid soluble fraction was precipitated by performing methanol-chloroform precipitation according to standard protocol, in which methanol, chloroform, and deionized water (1:4:1:3, v/v/v/v) were added in sequence; before mixing well for further centrifugation. The acid soluble shell protein was retrieved from the interphase of the upper and lower layer, washed again with 400 µL methanol. After centrifuging and removal of supernatant, the protein pellet was air-dried, after which 500 µL of 8M urea was added to solubilize the acid soluble shell protein fraction (ASF) pellet. To extract acid insoluble proteins (AIF), 1 mL of 8M urea was first added to the acid insoluble remaining of the decalcified shell samples to extract acid insoluble proteins. After thoroughly mixing, the sample was centrifuged, and the supernatant was retrieved as the urea soluble acid insoluble fraction (AIF-U). The acid and urea insoluble remaining substances was further extracted with 3x SDS sample buffer (3% SDS, 50 mM DTT, 10% glycerol, 0.1% bromophenol blue) in a 95 ℃ water bath for 5 mins. After centrifuge, the retrieved supernatant was taken as shell proteins that are insoluble in acid but soluble in SDS (AIF-S). Protein quantification was performed using the BCA protein quantification kit (BCA protein assay kit, Sigma, USA).

Polyacrylamide gel electrophoresis and in-gel digestion

For each of the three shell protein fractions from each species, a total of 300 µg protein was applied to a 4–20% polyacrylamide protein gradient gel (ExpressPlus PAGE Gels, GenScript) loaded with broad range protein marker (Broad Multi Color Pre-Stained Protein Standard, GenScript, China). The running voltage was set at 130 V. After the dye-front reached the bottom of the gel, electrophoresis was stopped and the gel was washed with deionized water for 5 mins, then soaked in fixing solution (water: methanol: acetic acid = 5:4:1, v/v/v) and shake with orbital rotation overnight. After rehydration in deionized water, the whole gel was divided into eight size fractions according to the broad range protein marker bands and were excised separately. The excised protein band from each size fraction was washed with 50 mmol of NH_4HCO_3 solution, 50% acetonitrile/50 mmol NH_4HCO_3, and finally with 100% acetonitrile. The shrunken gel was rehydrated with 10 mM dithiothreitol (DTT) and shook at 56 ℃ for 45 mins to perform reduction to reduce the disulfide bonds in proteins. 55 mM iodoacetamide (IAA) solution was added and the samples were incubated at room temperature in the darkness for 30 mins for alkylation. The gel samples were subjected to dehydration by ascending concentration of acetonitrile. Trypsin with a cut site R/K (Sequencing Grade Modified Trypsin, PROMEGA, Germany) were applied to perform in-gel digestion at 37°C for 16 hours. 0.1% formic acid was added to stop the reaction and ultrasound was performed for 10 mins to facilitate the retrieval of digested polypeptides from the gel. The protein digest in the solution in each gel fraction was retrieved, followed by two washes of 50% acetonitrile/50 mmol NH_4HCO_3 and 100% acetonitrile to further extract digested polypeptides from the gel. All collected solution was dried by the SpeedVac vacuum centrifuge (JM50-Plus, JM, China) with temperature set at 45 ^{o}C. The dried peptide samples were then reconstituted in 1 mL of 0.1% trifluoroacetic acid solution and were then desalted using Sep-Pak (Sep-Pak C18 6 cc Vac Cartridge, Waters, USA). The desalted peptide samples were then lyophilized and submitted to LC-MS/MS for mass spectrometry analysis.

LC-MS/MS mass spectrometry

Standard shotgun proteomics techniques were used to analyze the samples on Thermo Scientific LTQ VelosTM platform (Thermo Fisher Scientific, Bremen, Germany). The LTQ mass spectrometer was interfaced to a nano-electrospray ion source coupled with a Thermo Accela LC. Protein digest (1 µg) were enriched on a trap column (Zorbax X300 SB-C18, 5×0. 3 mm, 5 µm particle size) and separated on a C18 column (Thermo Bio-Basic-18, 150×0.1mm, 300 Å pore size, 5 µm particle size) at a flow rate of 150 µL/min. with 0.1% formic acid and for solvent B was acetonitrile with 0.1% formic acid. After sample loading, the LC mobile phase is maintained at 2% of solvent B for 2 mins. The gradient started from 2 to 10% of solvent B for 2 mins and then from 10 to 32% of solvent B for 60 mins. Afterwards, the gradient quickly changed from 30 to 60% of solvent B for 6 mins and from 60 to 80% of solvent B for 2 mins. In the final stage, the mobile phase is kept at 80% of solvent B for 2 mins, and then decreased from 80 to 2% of solvent B for 2 mins. The next injection was started after equilibrating at 2% of solvent B for 14 mins. Each MS1 was followed by ten MS2 of the top ten most intense ions observed in the MS1 scan and dynamics exclusion was 60 secs. Singly charged precursor ions were excluded from MS2. The normalized collision energy was set at 30% and one microscan was acquired for each spectrum.

Protein identification by MASCOT

The mass spectrometry data from each fraction was converted to .mgf file. Mass spectra generated from acid-soluble, Urea and SDS soluble acid-insoluble fractions of the operculum and wall plates were combined and searched for matches against the customized coy-decoy translated protein database using MASCOT server v2.3.2. The coy-decoy database contained the translated protein sequences generated by *in silico* translation of the transcriptome assembly using TransDecoder v3.0.1 [35] and the decoy

reverse protein sequences generated by flip-flopping the translated protein sequences. Trypsin was specified as the digestion enzyme and two missed cleavages for peptides were allowed. The carbamidomethyl was set as fixed modification and methionine oxidation (+ 15.99492 Da) was set as a variable modification. The mass tolerant for MS1 and MS2 was set at 0.8 Da. The FDR was set at zero such that no decoy sequence would be detected in the identified list of proteins.

Gene age estimation by phylostratigraphy

The age of the recovered barnacle shell proteins was estimated using a phylostratigraphic approach [36]. Twelve phylogenetic ranks (phylostrata) were defined according to the NCBI Taxonomy database as in Aguilera et al. (2017) [37], with the first phylostratum (PS1) being at the origin of cellular life (i.e., the oldest genes), and the last phylostratum (PS12) being *B. hirsutum*.

Analysis of shell protein sequences

To search for homologous shell protein sequences, *B. hirsutum* shell proteins were blast against each other using an all-to-all BLASTP approach. All shell proteins were searched for its function via BLAST website (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and SMART website (http://smart.embl-heidelberg.de/)[38]. The hydrophobicity test of the protein was performed using GRAVY CALCULATOR (http://www.gravy-calculator.de/index.php?page=file). The isoelectric detection of proteins was performed using Protein Isoelectric Point Calculator (http://isoelectric.org/)[39]. Detection of repetitive low complexity regions (RLCDs) within protein sequences was performed using T-REKS [40]. The secreted proteins were predicted by SignalP-5.0 (https://services.healthtech.dtu.dk/service.php?SignalP-5.0)[41]. Amino acid depletion/enrichment analysis for CP52k homologs was performed using the web-tool Composition Profiler [42].

Phylogenetic analyses

RNAseq data of *Glyptelasma gigas* (Annandale, 1916) (Ggigas), *Octolasmis warwickii* Gray, 1825 (Owar) (reported in Gan et al. 2020)[43], *Chelonibia testudinaria* (Linnaeus, 1758) (Ctest) (reported in [6]), *Neolepas marisindica* (Watanabe, Chen & Chan, 2018) (Nmar), *Megabalanus ajax* (Darwin, 1854) (Majex), *Tetraclita formosana* (Hiro, 1939) (Tform) (reported in [7]) were downloaded from NCBI SRA. These RNAseq data were subjected to *de novo* assembly, collapsing by CD-HIT-EST and *in silico* translation by TransDecoder as described above. To extract homologs of the recovered *B. hirsutum* shell proteins, BLASTP was performed using the concerned barnacle shell protein sequence as the query and the translated protein of the aforementioned collection as the database, with e-value cutoff set at 1e-5. Homolog sequences from *Amphibalanus amphitrite* (Aamph) and *Pollicipes pollicipes* (Ppoll) were downloaded directly from NCBI GenBank. MUSCLE [44] was used to align amino acid sequences, and amino acid alignment further guided the alignment of coding DNA sequences in ParaAT version 1.0 [45] with the "*-g*" flag to delete gaps in the aligned sequences. To remove spuriously aligned sites, the alignments of the corresponding DNA codon sequences were further trimmed by Gblocks [46] and the *bona-fide* alignments were used to estimate phylogenetic trees using the Maximum Likelihood method by IQ-TREE (-m LG + I + G4 + F -bb 1000 -alrt 10000 -nt 4) [47]. An SOWH test for tree topologies of the Maximum Likelihood tree was performed using SOWHAT [48] with the "raxml_model" parameter set at "GTRGAMMA" (–raxml_model = GTRGAMMA).

Results and Discussion

Identification of barnacle shell proteins

To identify shell proteins, we first performed RNAseq on barnacle cirri, prosoma, opercular epithelial tissues, and rim tissues. Trinity *de novo* assembly of *B. hirsutum* Illumina RNAseq short read data generated a transcriptome assembly consisting of 174,752 transcript sequences, with a total length of 105,479,217 bp and N50 at 1,130 bp. Using the software TransDecoder with default settings, 36,237 protein coding sequences were detected, and their open reading frame (ORF) were translated into a protein sequences database with an average length of 287 amino acids. Among all the translated protein sequences, 25,690 or 70.89% were annotated by either nr or SwissProt protein databases. The complete BUSCO score (Cs) of the protein database was 86.3% (874 out of 1,013 complete BUSCO group searched) (see Table 1 for details).

Table 1
Assembly statistics, annotation status, and BUSCO result

Assembly statistics	
Total number of Unigenes	174,752
Total length of the assembly	105,479,217 bp
N50 of the assembly	1,130 bp
Annotation status	
Total protein coding transcripts	36,237
Annotated in NR	20899(57.67%)
Annotated in SwissProt	22041(60.82%)
Annotated in GO	25334(69.91%)
Annotated in at least one Database	25690(70.89%)
BUSCO result	
Total BUSCO groups searched	1,013
Complete BUSCOs (C)	874
Complete and single-copy BUSCOs (S)	767
Complete and duplicated BUSCOs (D)	107
Fragmented BUSCOs (F)	77
Missing BUSCOs (M)	62

LC-MS/MS generated 3374, 12,500, and 20,088 MS/MS spectra from *B. hirsutum* operculum plate acid soluble (OP-ASF), urea, and SDS dissolved acid insoluble fractions (OP-AIF-U and OP-AIF-S), respectively. By performing a MASCOT protein search against the transcriptome derived coy-decoy translated protein database, 274 proteins were detected in at least one of the fractions in the examined *B. hirsutum* shell plates. The proteomic data is summarized in Table 2.

Table 2
Overview of shell protein identification

Shell extract sample	Total no. of MS/MS spectra	Translated protein sequence matched*	Matched protein with signalP
WP-ASF	4641	14	9
WP-AIF-U	23314	166	37
WF-AIF-S	26503	126	45
OP-ASF	3374	12	10
OP-AIF-U	12500	104	41
OP-AIF-S	20088	119	34
*Protein identification using a coy-decoy database, with FDR set at 0.			

Following the method described in Wong et al. (2019), we assumed that true shell proteins should be those that 1) were detected in shell proteomic, 2) exhibited shell making tissue specific mRNA expression pattern, and finally 3) contained an N-terminal signal peptide sequence. Hence, among the proteins detected in the shell samples, we further screened for proteins which mRNA was up-regulated (by 16 times) or specifically expressed in the rim or the operculum mantle tissue (opep) samples. As some shell protein sequences were only partially recovered, we also checked the presence of N-terminal signal peptides but did not exclude rim/opep specific proteins which signal peptide is absent. The transcript of *B. hirsutum* 31 identified proteins (out of 274 proteins) identified

from the shell samples exhibited a rim or opep specific expression pattern. Although among these 31 proteins, three of them are lacking the N-terminal signal peptide sequence, we nonetheless recruited these three proteins as Bh-BSPs (Table 3).

Table 3
Shell protein sequences identified from *B. hirsutum* shell extract samples.

Protein_ID	Annotation	Assigned name	SignalP	SMART
Bhirs_DN101626_c1_g1_i1	uncharacterized protein LOC122393488 [Amphibalanus amphitrite]|XP_043245474.1	Bh-BSSP_DRYrich	SP(Sec/SPI)	
Bhirs_DN102860_c1_g1_i1	serine-aspartate repeat-containing protein F-like [Amphibalanus amphitrite]|XP_043226407.1	Bh-BSSP_DSGrich_iso1	SP(Sec/SPI)	
Bhirs_DN107764_c0_g1_i1	Carbonic anhydrase 6 [Amphibalanus amphitrite]|KAF0312136.1	Bh-BSP_CA_typeA	SP(Sec/SPI)	
Bhirs_DN107933_c0_g1_i1	hypothetical protein FJT64_007420 [Amphibalanus amphitrite]|KAF0294983.1	Bh-BSSP_FJT64_007420_homolog	SP(Sec/SPI)	
Bhirs_DN108902_c0_g2_i1	hypothetical protein FJT64_008003 [Amphibalanus amphitrite]|KAF0294304.1	Bh-BSSP_FJT64_008003_homolog	SP(Sec/SPI)	
Bhirs_DN109748_c0_g1_i1	hypothetical protein FJT64_008002 [Amphibalanus amphitrite]|KAF0294303.1	Bh-BSSP_FJT64_008002_homolog	SP(Sec/SPI)	
Bhirs_DN112007_c2_g1_i1	Chorion peroxidase [Amphibalanus amphitrite]|KAF0310238.1	Bh-BSP_POX	SP(Sec/SPI)	
Bhirs_DN113120_c2_g2_i2	serine-aspartate repeat-containing protein F-like [Amphibalanus amphitrite]|XP_043226407.1	Bh-BSSP_DSGrich_iso2	SP(Sec/SPI)	
Bhirs_DN113476_c6_g1_i1	Serine protease easter [Amphibalanus amphitrite]|KAF0305686.1	Bh-BSP_Serine_Protease	SP(Sec/SPI)	
Bhirs_DN115464_c0_g2_i1	collagen alpha-1(III) chain-like [Amphibalanus amphitrite]|XP_043224125.1	Bh-BSP_collagen alpha-1(III) chain-like	SP(Sec/SPI)	
Bhirs_DN115559_c2_g1_i2	Annexin B10 [Amphibalanus amphitrite]|KAF0305229.1	Bh-BSP_Annexin-B10	SP(Sec/SPI)	
Bhirs_DN115656_c2_g1_i1	lectin-like isoform X1 [Amphibalanus amphitrite]|XP_043214755.1	Bh-BSP_CLECT	SP(Sec/SPI)	
Bhirs_DN115958_c0_g1_i1	collectin-10-like isoform X1 [Branchiostoma floridae]|XP_035692282.1	Bh-BSP_Collectin-like isoform X1	SP(Sec/SPI)	
Bhirs_DN116530_c0_g1_i1	Carbonic anhydrase 3 [Amphibalanus amphitrite]|KAF0306977.1	Bh-BSP_Carbonic anhydrase	SP(Sec/SPI)	
Bhirs_DN116703_c6_g2_i1	hypothetical protein FJT64_001459 [Amphibalanus amphitrite]|KAF0287525.1	Bh-BSSP_FJT64_001459_homolog	N.D.	

Protein_ID	Annotation	Assigned name	SignalP	SMART
Bhirs_DN117275_c0_g2_i2	hypothetical protein FJT64_008005 [Amphibalanus amphitrite]|KAF0294306.1	Bh-BSSP_FJT64_008005_homolog	SP(Sec/SPI)	
Bhirs_DN117974_c1_g2_i1	Carbonic anhydrase 1 [Amphibalanus amphitrite]|KAF0287462.1	Bh-BSP_Carbonic anhydrase	SP(Sec/SPI)	
Bhirs_DN120109_c0_g2_i9	PREDICTED: uncharacterized protein LOC658528 [Tribolium castaneum]|XP_043226683.1	Bh-BSP_EGF/FOLN_domain_containing	N.D.	EGF, FOLN, TM, Zona pellucida (ZP)
Bhirs_DN177080_c1_g1_i1	no hit	Bh-BSSP_QAPrich_type2	SP(Sec/SPI)	
Bhirs_DN22363_c0_g1_i1	chitin deacetylase 1-like [Amphibalanus amphitrite]|XP_043233856.1	Bh-BSP_chitin deacetylase 1-like	SP(Sec/SPI)	
Bhirs_DN40976_c0_g1_i1	no hit	Bh-BSSP_APrich_type3	N.D.	
Bhirs_DN60548_c0_g1_i1	no hit	Bh-BSSP_GSrich_type2	SP(Sec/SPI)	
Bhirs_DN61156_c0_g3_i1	no hit	Bh-BSSP_EPrich	SP(Sec/SPI)	
Bhirs_DN64114_c0_g1_i1	no hit	Bh-BSSP_APrich_type4	SP(Sec/SPI)	
Bhirs_DN68810_c0_g1_i1	Carbonic anhydrase 2 [Amphibalanus amphitrite]|KAF0287463.1	Bh-BSP_CA_typeB	SP(Sec/SPI)	
Bhirs_DN69553_c0_g1_i1	fibropellin-1-like isoform X2 [Amphibalanus amphitrite]|XP_043214271.1	Bh-BSP_Fibropellin-like	SP(Sec/SPI)	
Bhirs_DN80141_c1_g1_i1	cuticle protein 19-like [Pollicipes pollicipes]|XP_037069594.1	Bh-BSP_Cuticle-like	SP(Sec/SPI)	
Bhirs_DN84923_c1_g31_i1	Acetylcholine receptor subunit alpha-type acr-16 [Amphibalanus amphitrite]|KAF0300376.1	Bh-BSP_AchReceptor-like	SP(Sec/SPI)	
Bhirs_DN92775_c0_g1_i5	proline-rich protein 36-like [Amphibalanus amphitrite]|XP_043236095.1	Bh-BSSP_APrich_type1	SP(Sec/SPI)	
Bhirs_DN94351_c1_g4_i1	no hit	Bh-BSSP_QAPrich_type1	SP(Sec/SPI)	
Bhirs_DN96994_c0_g1_i1	no hit	Bh-BSSP_GSrich_type1	SP(Sec/SPI)	

Functional annotations of B. hirsutum shell proteins

The annotations of these 31 *B. hirsutum* shell proteins (*Bh*-BSP) include 13 barnacle specific unknown function proteins, four carbonic anhydrase, one chitin-binding domain containing cuticle protein, peroxidase, serine protease, C-type lectin domain containing mannose binding protein, peroxidase, collagen, collectin-like protein, fibropellin-like protein, AchReceptor-like protein, and EGF/FOLN domain containing protein (Fig. 1d, note that collagen, collectin-like protein, Fibropellin-like protein, AchReceptor-like protein, and EGF/FOLN domain containing protein were categorized as “Others” in the chart). Seven *Bh*-BSPs did not match any sequences in the nr database and were considered as “novel” shell proteins. The annotation of all Bh-BSPs were summarized in Table 3.

Alpha-carbonic anhydrase

We identified four alpha-carbonic anhydrase sequences from the shell proteome of *B. hirsutum*. These were unanimously assigned with a pfam α-carbonic anhydrase domain. In the maximum likelihood (ML) tree estimate, the α-CA identified from *B. hirsutum* fell into four major clusters specific to the Cirripedia (“true” barnacles Acrothoracica (burrowing barnacles), Rhizocephala (parasitic barnacles), and Thoracica (stalked and acorn barnacles); Fig. S1), suggesting that these α-CA homologs are paralogous, emerged from duplication of α-CA in the ancestor of thoracican barnacles and followed by divergence of each paralogue.

Peroxidase

The animal Heam-dependent peroxidase is a highly elaborate protein family with multiple homologs performing different functions. In the mosquito *Aedes aegypti*, peroxidase was implicated in crosslinking of chorion protein and hardening of egg capsule chorion [49, 50]. The peroxidase recovered from the shell proteome of *B. hirsutum* contains a long region unconserved region in between the N-terminal signal peptide and the C-terminal peroxidase domain, and within the region is a Proline(Pro)-rich region could be observed near the N-terminal signal peptide of all peroxidase homologs within the cluster (Fig. S2).

C-type lectin

The shell-relating C-type lectin recovered in this study formed a distinct cluster with the other homologs extracted from the transcriptome of *B. hirsutum* and also homologs from the examined stalked and acorn barnacle species (Fig. S3). C-type lectin binds to carbohydrate in the presence of calcium ion. In *M. rosa*, two C-type lectin namely BRA-2 and − 3 were suggested to be involved in healing and mineralization of the damaged basal plate [51]. But the cluster did not include BRA-2 and − 3. The result suggested 1) there may exist a specific group of C-type lectin homolog that may implicated in shell formation; 2) *M.rosa* BRA-2 and − 3 are structurally distinct and therefore distantly related to the shell-related C-type lectin homologs.

Phylostratigraphy of barnacle shell proteins

The age of the identified barnacle shell proteins was estimated using a phylostratigraphy approach [36]. Twelve phylogenetic ranks (phylostrata) were defined according to the NCBI Taxonomy database as in Aguilera et al. (2017) [37], with the first phylostratum (PS1) being at the origin of cellular life (i.e., the oldest genes), and the last phylostratum (PS12) being *B. hirsutum*. The best BLASTP hits for the BSPs and the mapping of protein to phylostrata are shown in Table S1. In total, eight Bh-BSPs can only be matched to published barnacle protein sequences and six could not match any sequences in the non-redundant database. These Bh-BSPs were classified into PS11 and PS12, which were regarded as barnacle (Cirripedia)-specific shell proteins (Fig. 1e). Other annotated Bh-BSPs were mapped to PS1 to PS8, which were proteins with conserved functions that could be traced back to the Bilateria or more ancestral PS levels (Fig. 1e). The results suggested that shell formation in barnacles likely involved substantial amounts of genetic innovations in addition to diversification of functional genes such as carbonic anhydrase that was found in multiple taxa with calcified shell structures.

B. hirsutum Barnacle Specific Shell Protein (Bh-BSSP)

As shown above, 14 out of the 31identified shell proteins of *B. hirsutum* were assigned to PS11 or PS12, indicating that these shell proteins are either novel (no match in Nr) or barnacle specific (proteins only matched barnacle sequences in Nr predicted from published barnacle genomes). We assigned protein names to these proteins with the prefix *B. hirsutum* Barnacle Specific Shell Protein, or Bh-BSSP (see Table 3, “assigned protein name” column). Some of the annotations of these Bh-BSSPs include “Serine and Aspartic acid repeat containing protein F-like”, “proline-rich protein 36-like”, “mucin-like protein”, and multiple hypothetical proteins or uncharacterized proteins (Table 3). It is noteworthy that one of the Bh-BSSPs was matched only to “collagen alpha-chain like” protein predicted from the published barnacle genome (Table 3). Despite that collagen repeats are rather conservative across the investigated taxa, this Bh-BSP did not match with any collagen alpha-chain proteins from any other taxa outside Cirripedia. Hence, we suspect that this Bh-BSP was indeed barnacle specific and the “collagen alpha-chain like” protein predicted from the published barnacle genome deposited in the nr were likely a mis-annotation.

Properties of shell proteins with unknown function(s)

The predicted hydropathy index and protein isoelectric points (pI) of Bh-BSSPs were further investigated (see Table 4 for details). Most Bh-BSSPs were acidic and hydrophilic (Fig. 2a). In mollusks, a large portion of shell matrix proteins contained low complexity repeat domains (RLCD) and were reportedly intrinsically disordered proteins. Here, we found that more than half of the Bh-BSSPs contain less than 2 cysteine (C) residues (Fig. 2a), suggesting these proteins are incapable of forming multiple intra-molecular C-C

bond. Further analysis of the amino acid composition revealed these Bh-BSSPs were low in Tyrosine (Tyr/Y), Histidine (His/H), Isoleucine (Ile/I), Phenylalanine (Phe/F), Methionine (Met/M), Tryptophan (Trp/W). Certain shell proteins shown distinctive bias (relative composition over 15%) toward certain amino acid such as Aspartic acid (Asp/D), Glycine (Gly/G), Serine (Ser/S), Proline (Pro/P), Glutamine (Gln/Q) (Fig. 2b, see relative abundance in Table S2).

Table 4
Biochemical properties of Bh-BSSPs

Sequence ID	Assigned name	GRAVY	average_pI	Mw	%_of_Cys	No._of_Cys
Bhirs_DN101626_c1_g1_i1	Bh-BSSP_DRYrich	-1.57	9.01	34585.66	4.27%	4
Bhirs_DN102860_c1_g1_i1	Bh-BSSP_DSGrich_iso1	-1.25	3.80	23350.74	4.93%	1
Bhirs_DN107933_c0_g1_i1	Bh-BSSP_FJT64_007420_homolog	-1.10	7.14	12553.83	5.66%	1
Bhirs_DN108902_c0_g2_i1	Bh-BSSP_FJT64_008003_homolog	0.30	4.96	21485.94	14.00%	5
Bhirs_DN109748_c0_g1_i1	Bh-BSSP_FJT64_008002_homolog	0.18	5.32	24690.64	11.61%	3
Bhirs_DN113120_c2_g2_i2	Bh-BSSP_DSGrich_iso2	-1.60	3.41	80310.98	1.49%	1
Bhirs_DN115464_c0_g2_i1	Bh-BSP_collagen alpha-1(III) chain-like	-0.82	7.64	41527.49	5.15%	6
Bhirs_DN116703_c6_g2_i1	Bh-BSSP_FJT64_001459_homolog	0.07	4.79	36083.09	8.12%	0
Bhirs_DN117275_c0_g2_i2	Bh-BSSP_FJT64_008005_homolog	-0.06	4.83	22752.21	12.32%	2
Bhirs_DN40976_c0_g1_i1	Bh-BSSP_APrich_type3	0.09	8.20	18053.22	6.01%	0
Bhirs_DN60548_c0_g1_i1	Bh-BSSP_GSrich_type2	-0.48	11.95	46785.24	2.69%	0
Bhirs_DN64114_c0_g1_i1	Bh-BSSP_APrich_type4	0.33	3.55	30078.11	5.68%	1
Bhirs_DN94351_c1_g4_i1	Bh-BSSP_QAPrich_type1	-0.83	6.06	63190.97	5.84%	2
Bhirs_DN61156_c0_g3_i1	Bh-BSSP_EPrich	-1.59	7.65	43888.62	2.76%	2
Bhirs_DN177080_c1_g1_i1	Bh-BSSP_QAPrich_type2	-0.67	5.57	51820.16	5.58%	3
Bhirs_DN92775_c0_g1_i5	Bh-BSSP_APrich_type1	-0.17	4.90	52818.78	6.04%	2
Bhirs_DN96994_c0_g1_i1	Bh-BSSP_GSrich_type1	-0.48	11.95	46052.49	2.74%	0

Bh-BSPs containing repetitive low complexity domain (RLCD)

During the analysis of the sequences of Bh-BSPs and Bh-BSSPs, we noticed that the sequences of certain shell proteins contain highly repetitive blocks. In mollusks, several families of shell matrix proteins were also observed with RLCDs, with unique properties and repetitive sequence blocks [37]. We therefore analyzed all Bh-BSPs and Bh-BSSPs sequences using the repeat recovering software T-REKS [40]. We found that 15 out of 31 Bh-BSP sequences were detected with RLCDs. Interestingly, these RLCDs containing Bh-BSPs or Bh-BSSPs were enriched with one or more than one of the following amino acids, namely Alanine (A), Aspartic acid (D), Glutamine (Q), Glutamic acid (E), Glysine (G), Proline (P), Serine (S), and Tyrosine (Y). C-type lectin, annexin, and one of the carbonic anhydrase paralogs were also detected with RLCD(s) outside the conserved protein domain regions. We summarized these tandem/internal repeat sequences in Table S3 and illustrate these low complexity regions in Fig. 3. Two groups of RLCD-containing Bh-BSSPs with intriguing primary structures are discussed in the following.

Bh-BSSPs with D- or E-rich RLCDs

D- and E-rich proteins have been implicated in shell formation in various taxa. In this study, two and one BSSPs were detected with D- and E-rich RLCDs, respectively. One of the D-rich barnacle specific shell proteins, *Bh*-BSSP_DSGrich, is composed of 805 amino acids with a molecular weight of 80.3 kDa (referring to the translated protein of the full-length transcript isoform Bh-BSP-DSGrich_iso2). Bh-BSSP_DSGrich was heavily biased toward Glycine (G), Serine (S), and Aspartic acid (D) (Fig. 2b) which accounted for 65.4% of all amino acid in the protein. As a result, Bh-BSP-DSGrich (Bh-BSP-DSGrich_iso2) was highly hydrophilicity (GRAVY score of -1.724) and acidic (predicted pI 3.42) (Fig. 2a and Table 4). In terms of protein structure, in addition to an N-terminal signal peptide, Bh-BSP-DSGrich contained a region made of 21 GENSKSDDSGSDSGDSDSDSDSG repeats.

Another D-rich barnacle specific shell protein, Bh-BSSP-DRYrich, was much smaller with 281 amino acids and a molecular weight of 34.5 kDa. Bh-BSP-DRYrich showed bias toward Aspartic acid (D), Arginine (R), and Tyrosine (Y), with the three amino acids accounting for 49.8% of the whole protein. In addition to an N-terminal signal peptide, Bh-BSP-DRYrich contained a region D and R rich RLCD. While the protein was also highly hydrophilic, enrichment of the basic amino acid Arginine led to a rather basic pI of the protein.

Bh-BSSP_EPrich is another barnacle specific shell protein which showed heavy bias toward the acidic amino acids glutamic acid (E) and proline (P). This Bh-BSSP was characterized by the presence of a region containing 30 KPEP repeats, such that the protein was heavily biased toward E and P and the two amino acids accounted for 36.4% of the whole protein. But the protein also contained 17.5% of lysine (K) and Arginine (R) so that the predicted pI of Bh-BSSP_EPrich was 7.65. In addition, there was a region of four AEARAAE repeats after the N-terminal signal peptide.

Bh-BSSPs with Q-rich RLCDs

We recovered two types of Q-rich BSSPs from the shell of *B. hirsutum*. Both types of Q-rich BSSPs were also rich in A and P, with the three amino acid accounting for more than 60.6% for Bh-BSSP_QAPrich_type1 and 62.6% for Bh-BSSP_QAPrich_type2. Bh- Bh-BSSP_QAPrich_type1 composed of 484 amino acids and contains four different RLCDs, with 14 AQQQ repeats closest to the N-terminal, followed by two regions of Q(T/A)Q repeats, 10 (Q/H)PAPVA repeats, and finished with seven PSGPAAGT repeats at the C-terminal; Bh-BSSP_Qrich_type2 composed of 582 amino acids and contains four different RLCDs, with 18 QPV-QH- repeats closest to the N-terminal, followed by a stretch of successive Q residues, four QLQQ repeats and six APATY repeats at the C-terminal portion (Fig. 3).

Exceptional bias of tandem repeat sequences toward proline, alanine, glycine, and serine

Intriguingly, Q- and D-rich barnacle shell proteins exhibited exceptional bias towards

proline, alanine, glycine, and serine. For instance, the QAPrich proteins also contained up to 31.6% P and 34.9% A; the D-rich protein Bh-BSSP_DSGrich also contained 29% S and 14.1% G; Bh-BSSP_EPrich contained 20.6% P. In addition to Q- and acidic D-rich proteins, more barnacle shell proteins were also reported with RLCDs with bias towards A, G, P, and S. For instance, two types of Bh-BSSP_APrich were detected. Bh-BSSP_APrich_type1 contained 26 successive PAVE tandem repeating blocks, followed six PTEAPAGP repeats; Bh-BSSP_APrich_type2, which was recovered with the C-terminal sequence, contained seven PGTAVN repeats; Bh-BSSP_Prich, which was annotated as homolog of “Proline-rich 36-like” from *A. amphitrite* (see Table 3), was highly biased toward P and contained the P-rich QPGPASAEP repeats in the N-terminal sequence, another GPASGTSS repeats in the C-terminal sequence and a section of successive A-stretch in the middle section; the “Collagen-like” protein Bh-BSSP_collagen_like, also contain four successive repetitive FGGPRGGPAD-Q blocks (see Table S3).

Acidic barnacle specific shell proteins hint to convergent evolution of genes involved with biomineralization

D-rich proteins such as Asprich have been detected as shell matrix proteins in multiple molluscan taxa [24]. The implication of molluscan Asprich family proteins has been widely discussed. It was believed that the Aspartic acid domains in Asprich are responsible for calcium ion binding. *In vitro* experimental results indicated that Asprich induces and transiently stabilizes the deposition of amorphous calcium carbonate (ACC). It was then suggested that the function of Asprich was to induce and stabilize ACC particles and inhibits their uncontrolled crystallization [52]. In term of protein structure, in addition to a D-rich domain, Asprich

family proteins also contain Calsequestrin-like region and a domain with DEAD repeats that was believed to have Mg-binding capabilities [24]. In *B. hirsutum*, two different D-rich proteins were recovered with D-rich domain. However, based on nr BLASTP results, these D-rich shell proteins were barnacle specific and therefore showed no homology to molluscan Asprich family protein.

While the functions of these acidic barnacle shell specific proteins remain to be explored, the presence of D-rich proteins suggested convergent evolution of shell matrix protein sequences from mollusks and barnacles. We further analyzed the primary structure, amino acid composition, predicted pI and hydropathy of Asprich of *Atrina rigida*, acidic shell matrix proteins from *Isognomon perna* and *Pinctada fucata*, the Pif protein from *Pinctada margaritifera*, and finally compared their biochemical characters to barnacle D-rich shell specific proteins. In terms of amino acid composition, all examined D-rich proteins showed heavy bias towards D. Yet, barnacle D-rich shell specific proteins and Pif contain 18 to 21% of D while the Asprich and acidic SMP contained more than 50% of D (Fig. 4a). Interestingly, the barnacle DRY-rich and Pif also shown enrichment in basic R (21.3%) and K (12.9%). Another interesting point to note was that the examined acidic SMPs and the barnacle DSGrich protein were also enriched in S and G. When the amino acid composition of these six D-rich proteins were analyzed by principal component analysis (PCoA), the barnacle DRY-rich protein showed obvious distinction from other examined D-rich proteins (Fig. 4b). The difference in amino acid composition was also reflected on the estimated biochemical properties of the protein, in which DRY-rich protein was the only basic D-rich protein (Fig. 4c). On the other hand, barnacle DSG-rich shell proteins showed similar amino acid content bias as in the three molluscan D-rich proteins, leading to highly acidic pI and hydrophilic properties in these proteins (Fig. 4c, see details in Table 3).

Previous studies have suggested that molluscan D-rich shell proteins might be important for Calcium or Magnesium ion binding during shell formation [24, 52, 53]. The presence of D stretch in Asprich likely functions to form a high-affinity Calcium ion capturing domain that inhibits the spontaneous crystallization of calcium carbonate in the mantle-shell growth compartment and thereby regulating shell formation. In all six examined D-rich proteins, the RLCDs were all enriched in D and covered more than half or even up to 90% of the proteins (Fig. 4a & d). Yet, detailed examination of the amino acid sequence of D-rich RLCD in each of the proteins revealed similarities and distinctive features in different proteins. The RLCDs the Asprich and the acidic SMPs were characterized by the presence of a highly redundant D-stretch; Bh-BSSP_DSGrich did not show D-stretch structure but was characterized the presence of large among of S and G in the RLCD. Interestingly, in both Bh-BSSP_DSGrich and Pf_SMP, multiple DSG sequence motifs could be observed distributing throughout the RLCD (highlighted in black boxes in Fig. 4d). DSG motif could also be detected in smaller number in Ip_SMP and Ar_Asprich (Fig. 4d). On the other hand, the RLCD in Pif and Bh-BSSP_DRY-rich both contained sequence motif constituted by the acidic D and the basic K or R in alternative order. Examples included DDRR, DDK(R)K, RRDD, KKDD (highlighted in purple boxes in Fig. 4d). Since Pif has been demonstrated to facilitate crystal nucleation during the crystallization of calcium carbonate in mollusk [54], we suspect that the alternative acidic/basic motif structure (DDKK/KKDD) could be a crucial motif in initiating the nucleation process.

While further functional studies will be needed to elucidate the reason for S and G presence in D-rich RLCD, we anticipate that the presence of these two amino acids may also relate to calcification. Serine is a residue of potential phosphorylation sites. It has been suggested that at least part of the shell proteome was phosphorylated, because phosphorylation of serine could result in negative charge of the amino acid residue, thereby facilitating Calcium binding during the shell formation process [55]. On the other hand, glycine is the smallest amino acid with no side chain group. Because of its structural simplicity, glycine spacer in protein sequences has been suggested to influence the exposure of functional motifs and protein self-assembling propensity [56]. While the reasons behind the co-incidental bias of S and G remained unclear, the occurrence of DSG motif in barnacle DSGrich shell protein and the two molluscan shell matrix protein, and the co-occurrence of acid/basic alternative motif structures in barnacle DRYrich protein and Pif suggests that there is substantial sequence convergence in the shell proteins in barnacles and bivalves.

Possible function of Q-rich shell protein

Compared to acidic shell matrix proteins, Q-rich proteins have rarely been considered as the major molecules involved with shell formation. While the roles of the two Q-rich Bh-BSSPs remain to be explored, several matrix protein surveys of different taxa did recover Q-rich proteins from biominerals. For instance, in the crayfish *Procambarus clarkia*, a protein recovered from the gastroliths, and later termed gastrolith matrix protein (GAMP), was found to contain a Q-rich RLCD [57]; in mollusks, a Q-rich protein has been recovered from a proteomic analysis of the organic matrix of the abalone *Haliotis asinina* [58]; in the adult Colorado potato beetles *Leptinotarsa decemlineata*, proline glutamine rich domain was found in the cuticle protein Ld-CP1v1 [59] and as such Q-rich domain was suggested to be involved in mediating crystallization of $CaCO_3$ on chitinous substrate. BLASTP against nr suggested these two

types of Q-rich shell proteins are barnacle specific, the Q-biased amino acid composition in the two Q-rich barnacle specific shell proteins, GAMP, Ld-CP1v1 and the abalone Q-rich shell matrix protein (SMP) could be again the result of convergent evolution. To facilitate visual comparison, the distribution, and the sequences of Q-rich RLCDs in GAMP, Ld-CP1v1, and abalone Q-rich matrix protein are illustrated in Fig. 5. Unlike D-rich RLCDs, no Q-rich RLCDs cover more than 50% of the examined full-length protein sequences (Fig. 5a). In terms of primary structure, GAMP RLCDs contain multiple AQE motif while other proteins are characterized by the presence of continuous Q-stretch. Ld-CP1v1 was characterized by the presence of PQPQYX repeats (where X refers to amino acid residue other than P, Q, Y), which is also distinct from the two types of barnacle QAP-rich proteins and the abalone Q-rich SMP (Fig. 5b).

While further functional studies and *in vitro* or *in vivo* data will be needed to elucidate the ultimate role of Q-rich barnacle shell proteins in biomineralization processes, we speculate that these proteins could be involved in stabilizing calcium carbonate crystal in chitinous cuticle, since barnacle shell plates are known to contain chitinous cuticle [60]. Interestingly, the PYQR repeats in the N-terminal of barnacle shell C-type lectin (see Fig. 3) were rather similar to the PQYX repeats in Ld-CP1v1. Further in-depth analyses must investigate the function of the shell C-type lectin in barnacle shell, and particularly the PYQR repeats, in biomineralization processes. Nevertheless, our results are the first report of Q-rich proteins in barnacle shells and adds to the repertoire of Q-rich RLCD possibly involved with biomineralization in marine invertebrates.

Conclusions

We identified 31 barnacle shell protein sequences from the deep-sea acorn barnacle *Bathylasma hirsutum*. Among these BSPs, 14 were found to be barnacle specific shell proteins (BSSPs). Among the annotated BSPs, carbonic anhydrase, C-type lectin and perioxidase were implicated in biomineralization or egg capsule hardening in other shell forming taxa. The BSSPs were found to be biased toward D, E, Q, A, P, G, and S. Almost half of the detected BSPs contained RLCDs. Interestingly, we observed no obvious homology between D-rich BSSPs, Asprich or molluscan shell matrix proteins, despite these proteins all being enriched in D. Similarly, Q-rich BSSPs also exhibited a strikingly similar Q-biased amino acid composition with the bee cuticle protein Ld-CP1v1 and an abalone shell matrix protein. The convergent evolution of amino acid compositions suggest, in two independent taxa, that biomineralization has independently exerted strong selective pressures on the molecular evolution of shell matrix protein genes.

Declarations

Ethics approval and consent to participate.

Not applicable.

Consent for publication

Not applicable.

Competing of interests

The authors declare they have competing interests.

Fundings

This research was jointly supported by the Innovation Team Project of Universities in Guangdong Province (no. 2020KCXTD023) and Natural Science Foundation of China General Project (award no. 42276104) awarded to Y.H.W; Scientific and Technical Innovation Council of Shenzhen Government (grant no. jcyj20210324093412035) award to Y.Z.; the National Natural Science Foundation of China General Project (award no. 52071332), the Department of Science and Technology of Guangdong Province (grant no. 2019QN01H430) and the Science and Technology Innovation Commission of Shenzhen (grant no. JCYJ20180507182239617) awarded to S.F.G. The sampling was supported by the German Science Foundation under grant number MErMet17-05 to S.B.

Availability of data and materials

All data presented in the manuscript have been stored with the article or its online supplementary materials. New RNA-seq datasets have been uploaded to National Centre for Biotechnological Information (NCBI) Sequence Read Archive (SRA) under the following bioproject accession number PRJNA681321.

Authors' contributions The study was conceived by Y.H.W. and S.B. and designed by Y.T.X., Y.H.W., Y.Z., S.F.G. and S.B. Barnacle samples were collected by J.T. and S.B. Y.T.X., H.C.L., Q.Q.C., N.D., Y.Z. and Y.H.W. performed experiments and analyzed the data. Y.T.X., H.C.L., N.D. and Y.H.W. wrote the first draft. All authors contributed to and agreed upon the final version of the manuscript.

References

1. **Letter from Darwin to William D. Fox, 24 October 1852. Source: (http://www.darwinproject.ac.uk)**.
2. Linnaeus C: **Systema naturae per regna tria naturae, Secundum Classes, Ordines, Genera, Species, Cum Characteribus, Differentiis, Synonymis, Locis (in Latin)**. *Laurentius Salvius, Stockholm* 1758.
3. Cuvier G: **Le règne animal distribué d'après son organisation: Pour servir de base à l'histoire naturelle des animaux et d'introduction à l'anatomie comparée**, vol. 1. Cambridge: Cambridge University Press; 2012.
4. Darwin C: **A monograph on the sub-class Cirripedia, with figures of all the species**, vol. 1. London: Ray society; 1851.
5. Chan B, Høeg J: **Diversity of lifestyles, sexual systems and larval development patterns in sessile crustaceans**. 2015, **2**:14-34.
6. Chan B, Wong Y, Robinson N, Lin J-C, Yu S-P, Dreyer N, Cheng IJ, Høeg J, Zardus J: **Five hundred million years to mobility: directed locomotion and its ecological function in a turtle barnacle**. *Proceedings of the Royal Society B: Biological Sciences* 2021, **288**.
7. Lin H-C, Wong T, Chan B: **Histology and transcriptomic analyses of barnacles with different base materials and habitats shed lights on the duplication and chemical diversification of barnacle cement proteins**. *BMC Genomics* 2021, **22**.
8. Stocks-Fischer S, Galinat J, Bang S: **Microbiological Precipitation of CaCO3**. *Soil Biology and Biochemistry* 1999, **31**:1563-1571.
9. Beveridge T, Fyfe W: **Metal Fixation by Bacterial Cell Walls**. *Canadian Journal of Earth Sciences* 2011, **22**:1893-1898.
10. Goldstein S, Gupta B: **Modern Foraminifera**. In.; 2003: 37-55.
11. Takeuchi T, Sarashina I, Iijima M, Endo K: **In vitro regulation of CaCO3 crystal polymorphism by the highly acidic molluscan shell protein Aspein**. *FEBS letters* 2008, **582**:591-596.
12. Bentov S, Erez J: **Impact of biomineralization processes on the Mg content of foraminiferal shells: A biological perspective**. *Geochemistry Geophysics Geosystems* 2006, **7**.
13. Bentov S, Brownlee C, Erez J: **The role of seawater endocytosis in the biomineralization process in calcareous foraminifera**. *Proceedings of the National Academy of Sciences of the United States of America* 2009, **106**:21500-21504.
14. Nagasawa H: **Mollusk shell structures and their formation mechanism1**. *Canadian Journal of Zoology* 2013, **91**.
15. Agbaje O, Thomas D, Dominguez J, McLnerney B, Kosnik M, Jacob D: **Biomacromolecules in bivalve shells with crossed lamellar architecture**. *Journal of Materials Science* 2019, **54**.
16. Anderson J: **: Bone and Tooth: Proceedings of the First European Symposium Held at Somerville College, Oxford, April 1963 . H. J. J. Blackwood**. *American Anthropologist* 2009, **67**:517-518.
17. Addadi L, Weiner S, Geva M: **On how proteins interact with crystals and their effect on crystal formation**. *Zeitschrift für Kardiologie* 2001, **90 Suppl 3**:92-98.
18. Germer J, Mann K, Wörheide G, Jackson D: **The Skeleton Forming Proteome of an Early Branching Metazoan: A Molecular Survey of the Biomineralization Components Employed by the Coralline Sponge Vaceletia Sp**. *PLoS ONE* 2015, **10**.
19. Karakostis K, Zanella-Cléon I, Immel F, Guichard N, Dru P, Lepage T, Plasseraud L, Matranga V, Marin F: **A minimal molecular toolkit for mineral deposition? Biochemistry and proteomics of the test matrix of adult specimens of the sea urchin Paracentrotus lividus**. *Journal of Proteomics* 2016, **136**.
20. Luo Y-J, Takeuchi T, Koyanagi R, Yamada L, Kanda M, Khalturina M, Fujie M, Yamasaki S-i, Endo K, Satoh N: **The Lingula genome provides insights into brachiopod evolution and the origin of phosphate biomineralization**. *Nature Communications* 2015, **6**:9301.

21. Marie B, Zanella-Cléon I, Guichard N, Becchi M, Marin F: **Novel Proteins from the Calcifying Shell Matrix of the Pacific Oyster Crassostrea gigas**. *Marine biotechnology (New York, NY)* 2011, **13**:1159-1168.
22. Drake JL, Mass T, Haramaty L, Zelzion E, Bhattacharya D, Falkowski PG: **Proteomic analysis of skeletal organic matrix from the stony coral Stylophora pistillata**. *Proc Natl Acad Sci U S A* 2013, **110**(10):3788-3793.
23. Jackson D, Macis L, Reitner J, Degnan B, Wörheide G: **Sponge Paleogenomics Reveals an Ancient Role for Carbonic Anhydrase in Skeletogenesis**. *Science (New York, NY)* 2007, **316**:1893-1895.
24. Gotliv B-A, Kessler N, Sumerel J, Morse D, Tuross N, Addadi L, Weiner S: **Asprich: A Novel Aspartic Acid-Rich Protein Family from the Prismatic Shell Matrix of the Bivalve Atrina rigida**. *Chembiochem : a European journal of chemical biology* 2005, **6**:304-314.
25. Ramos-Silva P, Kaandorp J, Huisman L, Marie B, Zanella-Cléon I, Guichard N, Miller D, Marin F: **The Skeletal Proteome of the Coral Acropora millepora: The Evolution of Calcification by Co-Option and Domain Shuffling**. *Molecular biology and evolution* 2013, **30**.
26. Bourget E: **Barnacle shells: Composition, structure and growth**. In.; 2018: 267-285.
27. Nousek N: **Shell formation and calcium transport in the barnacle Chthamalus fragilis**. *Tissue & cell* 1984, **16**:433-442.
28. Bubel A: **An ultrastructural study of the mantle of the barnacle, Elminius modestus Darwin in relation to shell formation**. *Journal of Experimental Marine Biology and Ecology - J EXP MAR BIOL ECOL* 1975, **20**:287-324.
29. Zhang G, He L, Wong T, Xu Y, Zhang Y, Qian P-Y: **Chemical Component and Proteomic Study of the Amphibalanus (= Balanus) amphitrite Shell**. *PloS one* 2015, **10**:e0133866.
30. Haas B, Papanicolaou A, Yassour M, Grabherr M, Blood P, Bowden J, Couger M, Eccles D, Li B, Lieber M *et al*: **De novo transcript sequence reconstruction from RNA-Seq using the Trinity platform for reference generation and analysis**. *Nature protocols* 2013, **8**:1494-1512.
31. Li W, Godzik A: **Cd-Hit: a Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences**. *Bioinformatics (Oxford, England)* 2006, **22**:1658-1659.
32. Simão F, Waterhouse R, Ioannidis P, Zdobnov E: **BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs**. *Bioinformatics (Oxford, England)* 2015, **31**.
33. Patro R, Duggal G, Love M, Irizarry R, Kingsford C: **Salmon provides fast and bias-aware quantification of transcript expression**. *Nature Methods* 2017, **14**.
34. Robinson M, McCarthy D, Smyth G: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics* 2010, **26**:139-140.
35. [Haas, BJ. https://github.com/TransDecoder/TransDecoder]
36. Domazet-Loso T, Brajković J, Tautz D: **A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages**. *Trends Genet* 2007, **23**(11):533-539.
37. Aguilera F, McDougall C, Degnan B: **Co-Option and De Novo Gene Evolution Underlie Molluscan Shell Diversity**. *Molecular biology and evolution* 2017, **34**.
38. Letunic I, Khedkar S, Bork P: **SMART: recent updates, new developments and status in 2020**. *Nucleic Acids Research* 2020, **49**(D1):D458-D460.
39. Kozlowski LP: **IPC – Isoelectric Point Calculator**. *Biology Direct* 2016, **11**(1):55.
40. Jorda J, Kajava AV: **T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm**. *Bioinformatics* 2009, **25**(20):2632-2638.
41. Armenteros J, Tsirigos K, Sønderby C, Nordahl Petersen T, Winther O, Brunak S, Heijne G, Nielsen H: **SignalP 5.0 improves signal peptide predictions using deep neural networks**. *Nature Biotechnology* 2019, **37**.
42. Vacic V, Uversky VN, Dunker AK, Lonardi S: **Composition Profiler: a tool for discovery and visualization of amino acid composition differences**. *BMC Bioinformatics* 2007, **8**:211.
43. Gan Z, Yuan J, Liu X, Dong D, Li F, Li X: **Comparative transcriptomic analysis of deep- and shallow-water barnacle species (Cirripedia, Poecilasmatidae) provides insights into deep-sea adaptation of sessile crustaceans**. *BMC Genomics* 2020, **21**.
44. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res* 2004, **32**(5):1792-1797.

45. Zhang Z, Xiao J, Wu J, Zhang H, Liu G, Wang X, Dai L: **ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments**. *Biochem Biophys Res Commun* 2012, **419**(4):779-781.
46. Castresana J: **Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis**. *Mol Biol Evol* 2000, **17**(4):540-552.
47. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ: **IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies**. *Mol Biol Evol* 2015, **32**(1):268-274.
48. Church S, Ryan J, Dunn C: **Automation and evaluation of the SOWH test with SOWHAT**. *Systematic biology* 2015, **64**.
49. Li J, Li J: **Major chorion proteins and their crosslinking during chorion hardening in Aedes aegypti mosquitoes**. *Insect biochemistry and molecular biology* 2007, **36**:954-964.
50. Li J, Hodgeman B, Christensen B: **Involvement of peroxidase in chorion hardening in Aedes aegypti**. *Insect Biochemistry and Molecular Biology* 1996, **26**:309-317.
51. Kamiya H, Jimbo M, Yako H, Muramoto K, Nakamura O, Kado R, Watanabe T: **Participation of the C-type hemolymph lectin in mineralization of the acorn barnacle Megabalanus rosa**. *Marine Biology* 2002, **140**:1235-1240.
52. Politi Y, Mahamid J, Goldberg H, Weiner S, Addadi L: **Asprich mollusk shell protein: In vitro experiments aimed at elucidating function in CaCO3 crystallization**. *Crystengcomm* 2007, **9**.
53. Endo H, Takagi Y, Ozaki N, Kogure T, Watanabe T: **A Crustacean Ca2+-binding protein with a glutamate-rich sequence promotes CaCO3 crystallization**. *The Biochemical journal* 2004, **384**:159-167.
54. Suzuki M, Saruwatari K, Kogure T, Yamamoto Y, Nishimura T, Kato T, Nagasawa H: **An Acidic Matrix Protein, Pif, Is a Key Macromolecule for Nacre Formation**. *Science (New York, NY)* 2009, **325**:1388-1390.
55. Du J, Xu G, Liu C, Zhang R: **The role of phosphorylation and dephosphorylation of shell matrix proteins in shell formation: an in vivo and in vitro study**. *CrystEngComm* 2018, **20**.
56. Taraballi F, Natalello A, Campione M, Villa O, Doglia SM, Paleari A, Gelain F. **Glycine-spacers influence functional motifs exposure and self-assembling propensity of functionalized substrates tailored for neural stem cell cultures**. *Frontiers in Neuroengineering* 2010, **3**:1161.
57. Naoaki T, Katsuaki I, Yasuaki T, Toshiki W, Hiromichi N: **Cloning and Expression of a cDNA Encoding an Insoluble Matrix Protein in the Gastroliths of a Crayfish**. *Zoological Science* 1999, **16**(4):619-628.
58. Marie B, Marie A, Jackson D, Dubost L, Degnan B, Milet C, Marin F: **Proteomic analysis of the organic matrix of the abalone Haliotis asinia calcified shell**. *Proteome science* 2010, **8**:54.
59. Zhang J: **Characterization of cuticular chitin-binding proteins of Leptinotarsa decemlineata (Say) and post-ecdysial transcript levels at different developmental stages**. *Insect molecular biology* 2010, **19**:517-525.
60. Fernández M, Vergara I, Oyarzun A, Arias J, Rodriguez R, Wiff J, Fuenzalida V, Arias J: **Extracellular Matrix Molecules Involved in Barnacle Shell Mineralization**. *Materials Research Society Symposium - Proceedings* 2002, **724**:3-9.

Figures

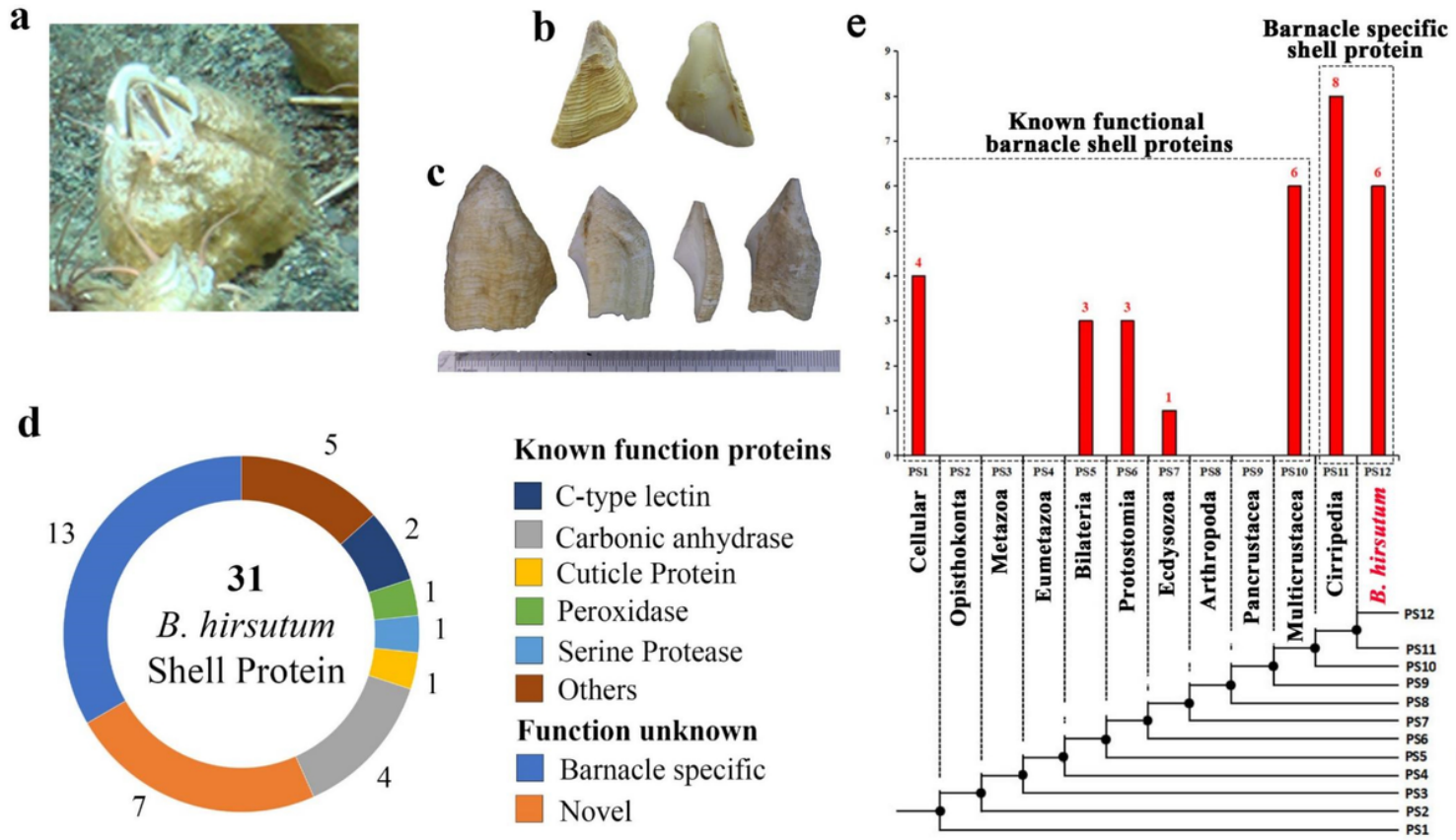

Figure 1

***Bathylasma hirsutum* shell plates and the recovered shell proteome. a**. a specimen of *B. hirsutum* inhabiting near a 973m deep hydrothermal vent. The image was a screenshot of a video recorded by ROV PHOCA. **b**. shell plates dissected from an adult *B. hirsutum* specimen. Upper panel: the operculum plates; lower panel: the wall plates. **c**. Phylostratography of the identified shell proteome of *B. hirsutum*. Detail annotation results were listed in Table S1. **d**. *B. hirsutum* shell proteome composition. Note that barnacle specific proteins referred to proteins which only matched to published barnacle genome databases in the nr protein database (as of Oct 2022) by BLASTP (evalue cutoff 1e-5).

Figure 2

Biochemical properties of barnacle specific shell proteins (Bh-BSSPs) in *B. hirsutum* shell proteome. a. a pI vs GRAVY chart of all Bh-BSSPs. Each dot represented one BSSPs and the size of the dot represent number of cysteine in the protein. **b**. Amino acid composition heatmap of all Bh-BSSPs. The rows (BSSPs) and columns (amino acids) were clustered using k-mean cluster method.

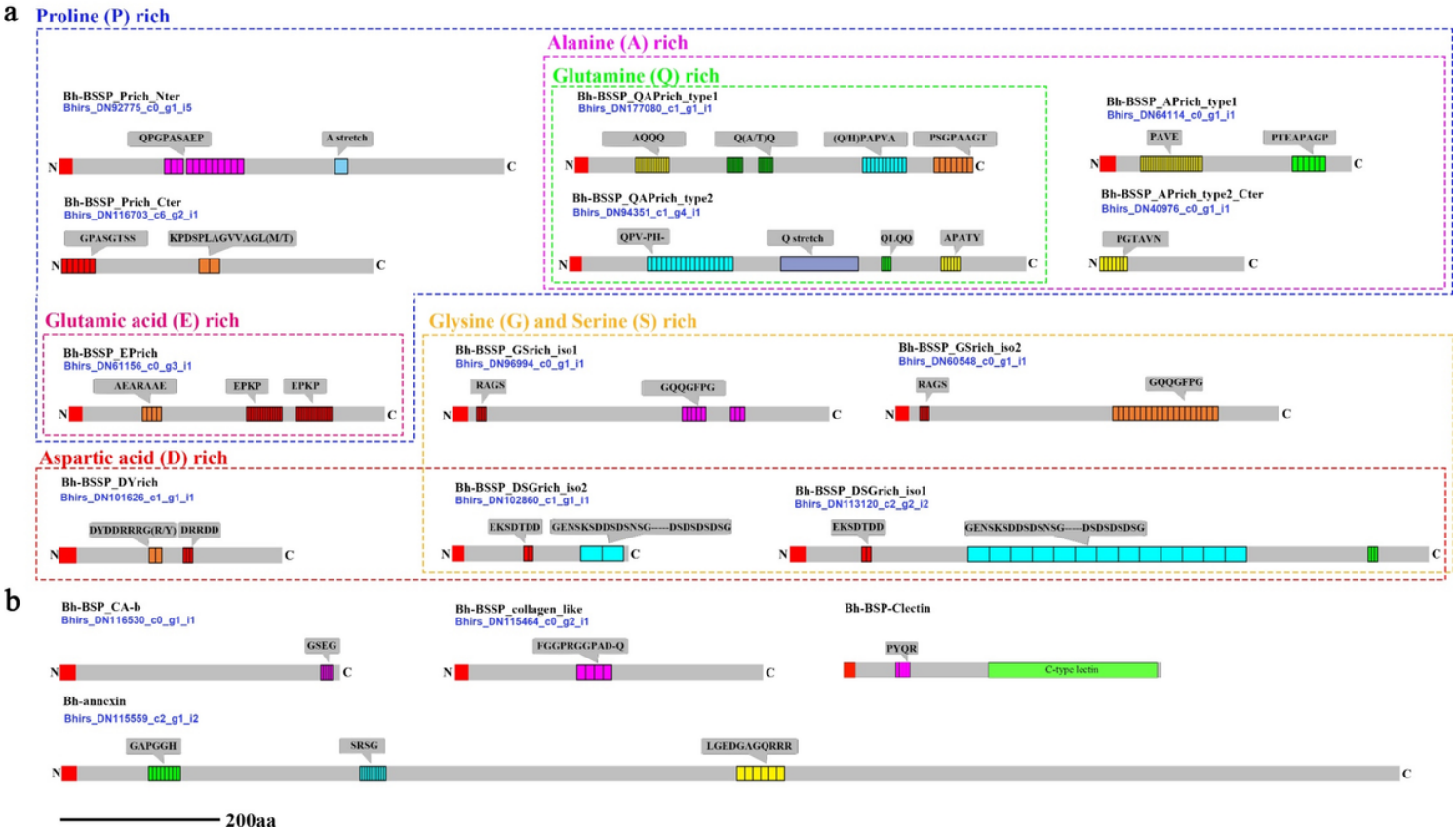

Figure 3

Protein domain schematics illustrating repetitive low complexity domains (RLCDs) in *B. hirsutum* shell proteins. a. Schematics of RLCDs in barnacle specific shell proteins. BSSPs were highlighted by dotted boxes in different colors, with each color representing bias toward a particular amino acid. **b**. Schematics of RLCDs in annotated barnacle shell proteins.

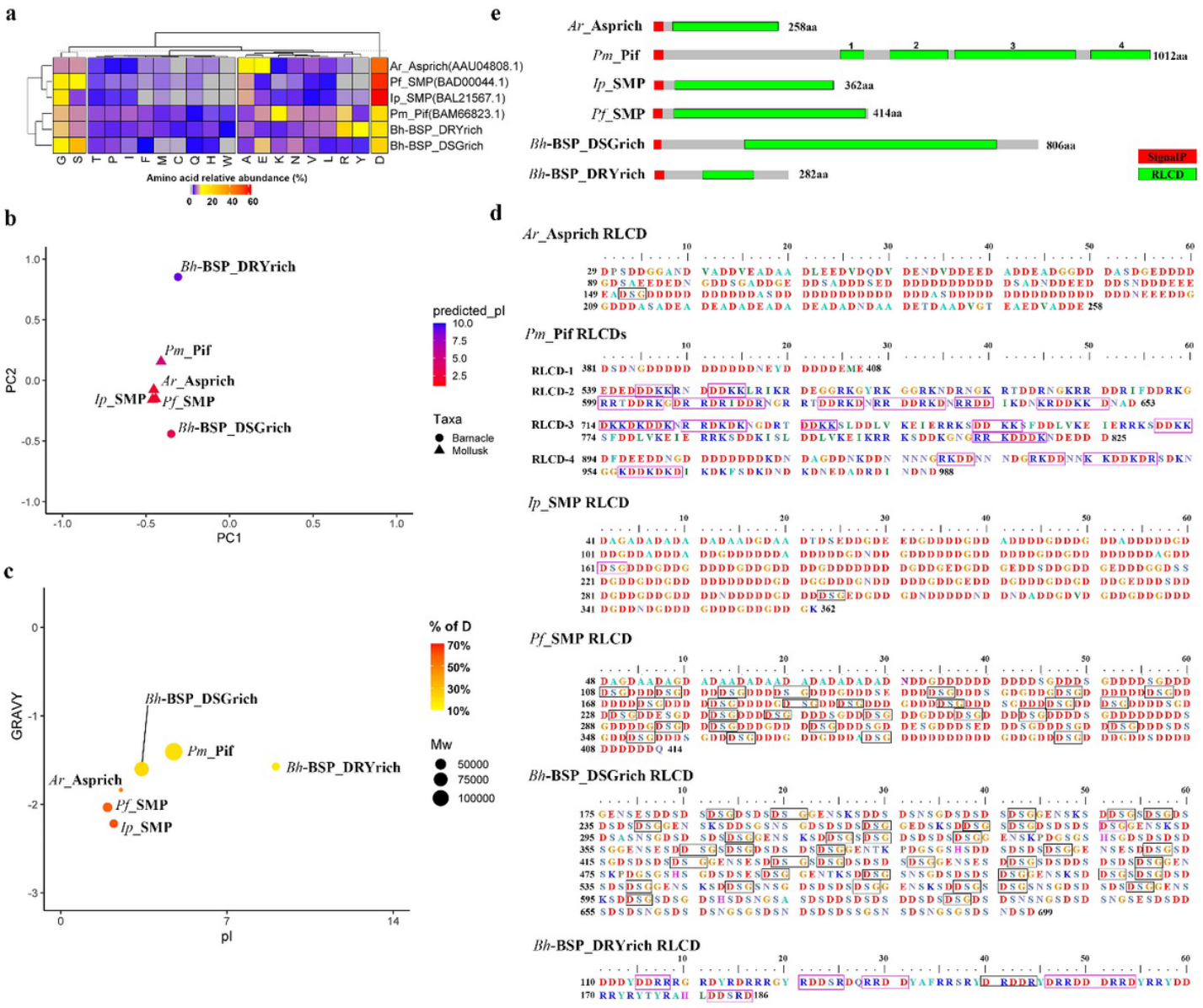

Figure 4

Comparison of molluscan acidic shell matrix proteins with *B. hirsutum* D-rich shell proteins. a. The amino acid composition of *Atrina rigida* Asprich (*Ar*_Asprich, AAU04808.1), acidic shell matrix protein from *Isognomon perna* (*Ip*_SMP, BAL21567.1) and *Pinctada fucata* (*Pf*_SMP, BAD00044.1), *Pinctada margaritifera* Pif protein (*Pm*_Pif, BAM66823.1), *Bh*-BSSP_DRYrich and *Bh*-BSSP_DSGrich. Enrichment amino acids were indicated with *. **b**. 2D Principal component analysis (PCoA) of the amino acid composition of the six D-rich proteins, with dot color illustrating the percentage of D in each protein. **c**. pI-GRAVY plot of the six D-rich proteins, with the size of each dot illustrating the predicted molecular weight of the proteins. **d**. Schematic of RLCDs in the six proteins. e. Sequences of RLCDs in each protein. DSG motifs were highlighted by black boxes and the alternative acidic/basic amino acid motif (DDKK/KKDD/DDRR) were highlighted by purple boxes.

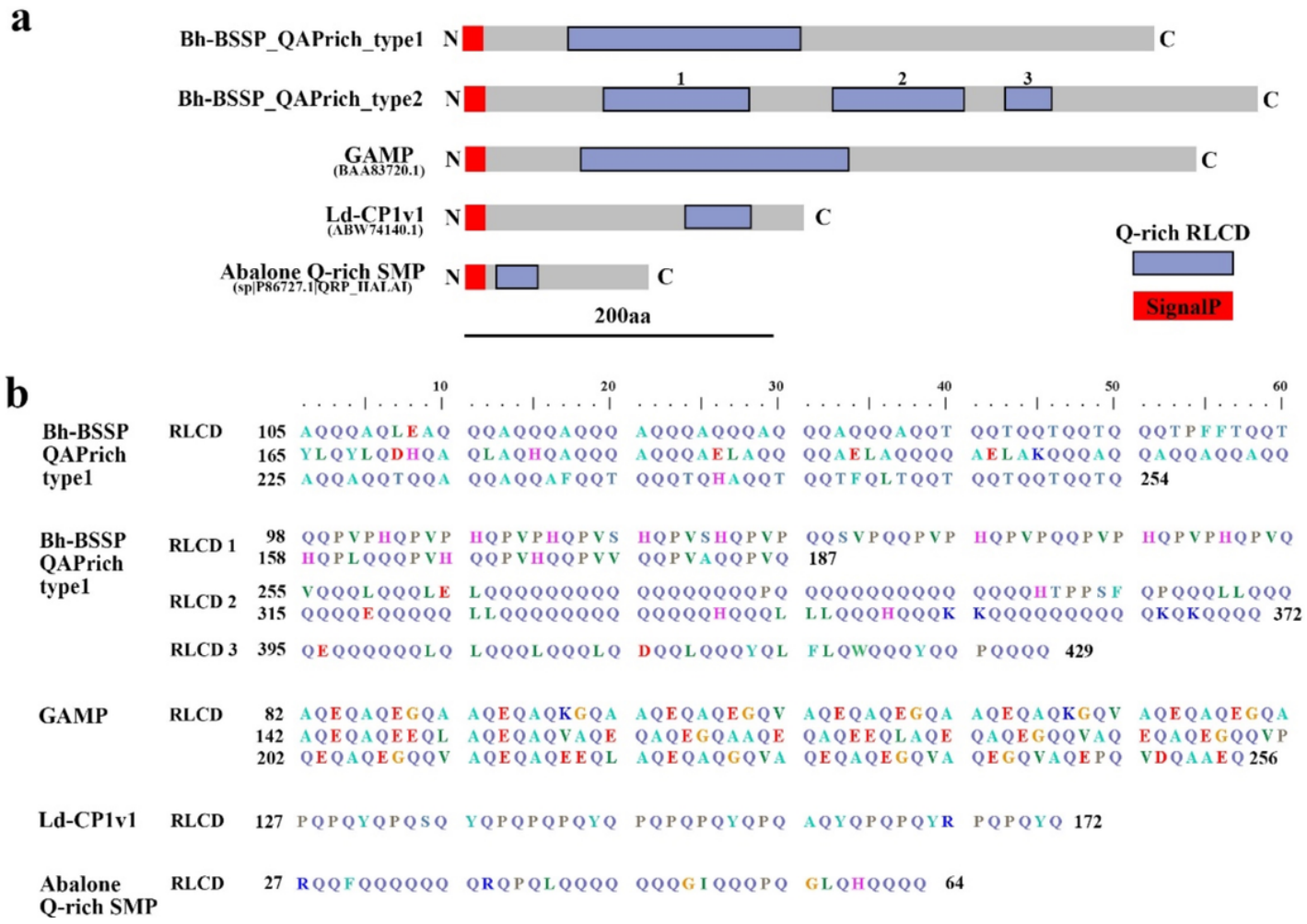

Figure 5

A comparison of known and *B. hirsutum* Q-rich proteins. The five examined proteins are *Procambarus clarkia* GAMP (BAA83720.1), *Haliotis asinina* (abalone) shell matrix protein (sp|P86727.1|QRP_HALAI), *Leptinotarsa decemlineata* cuticular protein Ld-CP1v1 (ABW74140.1). **a**. Schematic of RLCDs in the six proteins. **b**. Sequences of RLCDs in each protein.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- BhshellproteinBMCgenomics230718JTSBSupplementary.docx