REP RTS

I

C

L D F H L K D

H
I

B A
C

speci c shell pro. ein[11]. Recen l , high-. hro ghp . . ran crip. omic and pro. eomic anal e ha e been c cce-i el ed . o charac. eri e shell ma. ri pro. eome (SMPE) in *Haliotis asinina*[8], *Pinctada margaritifera*[12], *Pinctada maxima*[12], *Crassostrea gigas*[13], *Lottia gigantea*[14], *Cepaea nemoralis*[6], *Mytilus coruscus*[15], *Pinctada. fucata*[10], *Magellania venosa*[2], *Mytilus galloprovincialis*[16], *Mya truncata*[17], and *C. gigas*, *Mytilus edulis*, and *Pecten maximus*[18]. Ho e er, onl . o global SMPE compari on ha e been performed in *C. nemoralis* and *M. venosa*, re ealing lo imilari ie. E en in bi al e, . o di eren. model ha e been propo ed for calcif in

Gene name	Categories	Expression in mantle[a]	Best matched gene ID	BLAST best hit to NCBI accession (species)	Shell layer[b]	Domains[c]
ACCBP1	CaCO3	N	CGI_10024902	EKC41060 (*C. gigas*)	.	NCLBD
ACCBP2	CaCO3	N	CGI_10024903	EKC41058 (*C. gigas*)	.	NCLBD
BMSP	Chi. in	N	CGI_10009194	BAK86420 (*M. galloprovincialis*)	P,N	VWA;CBD
CaLP	CaCO3	N	CGI_10011294	P41041 (*Pneumocystis carinii*)	P,N	CaBEF
CaM	CaCO3	N	CGI_10011293	EKC20234 (*C. gigas*)	P,N	CaBEF
Cg r1	O. here	S	CGI_10007793	EKC29813 (*C. gigas*)	.	CBS
CgT r2	O. here	S	CGI_10011913	EKC18549 (C. gigas)	p	CBS
Chi. in e n. hase1	Chi. in	H	CGI_10009438	AAY86556 (*Atrina rigida*)	P,N	MHD
Chi. in e n. hase2	Chi. in	N	CGI_10012656	BAF73720 (*P. fucata*)	P,N	MHD
chi obiase	Chi. in	S	CGI_10007857	H2A0L6 (*P. margaritifera*)	P,N	Gl co_20
Chi. o. rioidase1	Chi. in	S	CGI_10024867	AFO53261 (*Hyriopsis cumingii*)	P,N	Gl co_20
Chi. o. rioidase2	Chi. in	S	CGI_10026605	CAI96027 (*C. gigas*)	P,N	Gl co_20
Clp3	Chi. in	S	CGI_10026599	H2A0L5 (*P. margaritifera*)	P	Gl co_18
CopAmO	ECM	H	CGI_10026457	EKC31553 (*C. gigas*)	P	Copper amine o idase domain
EGF-ZP1	ECM	S	CGI_10017543	P86785 (*C. gigas*)	.	EGF;ZP
EGF-ZP2	ECM	S	CGI_10017544	EKC41439 (*C. gigas*)	.	EGF;ZP
EGF-ZP3	ECM	H	CGI_10017545	P86954 (*C. gigas*)	.	EGF;ZP
Fibronec. in1	ECM	H	CGI_10016964	EKC41462 (*C. gigas*)	P	bronec. in . pe III
Fibronec. in2	ECM	H	CGI_10016965	EKC41461 (*C. gigas*)	P	bronec. in . pe III
Pero idase1	O. here	S	CGI_10023200	EKC34657 (*C. gigas*)	.	CaBS
Pero idase2	O. here	S	CGI_10010240	EKC26108 (*C. gigas*)	.	CaBS
PFMG9	O. here	H	CGI_10010153	ADC52432 (*P. fucata*)	.	KAZAL_FS
Pif-like1	Chi. in	N	CGI_10014497	AKV63183 (*P. fucata*)	P,N	VWA;CBD
Pif-like2	Chi. in	N	CGI_10017473	BAK86420 (*M. galloprovincialis*)	P,N	VWA;CBD
SPARC	ECM	N	CGI_10005088	AND99565 (*P. fucata*)		N-. erminal acidic domain

Table 1. Identification and characterization of SMPs from *C. gigas*. [a]repre en . he gene e pre ion le el in man le compared i h o her organ: S = pecial, H = high, N = no pecial or high. [b]repre en . he hell la er . ha ere de ec ed . o con ain . he con i en pro ein : p = perio rac m, P = pri ma ic la er, N = nacreo la er; 'P, N' mean . ha . he SMP ere fo nd in bo h . he la er ; . – repre en . nkno n. [c]Abbre ia ion: CBS = copper-binding i e ; MHD = m o in head domain; VWA = on Willebrand fac or (WF) pe A domain; CBD = chi in-binding domain; CaBS = calci m-binding i e; EGF = Epidermal gro h fac or; ZP = ona pell cida; CaBEF = Ca2 + -binding EF hand domain; NCLBD = ne ro ran mi er-ga ed ion-channel ligand-binding domain ; KAZAL_FS = Ka al pe erine pro ea e inhibi or and folli a in-like domain ; Gl co_18 = famil 18 Gl co l h drola e ; Gl co_20 = famil 20 Gl co l h drola e .

I Since . he r . SMP gene Nacrein a cloned, man SMP ha e been cloned and charac eri ed. Be ide . he hell' perior mechanical and remarkable biocompa ibili proper ie , . he high commercial al e of pearl ha made pearl o er one of . he be . died biominerali a ion model . Here, e ed . he e SMP mainl from *P. fucata* . o iden if . he homolog e in *C. gigas* and anal ed . he . r c ral domain nderl ing biominerali a ion. A . o al of 58 SMP eq ence ere ed . o perform BLAST (S pplemen ar Table S3) and 25 pro ein eq ence ere anno a ed and charac eri ed (Table 1). F r hermore, . he e SMP ere cla i ed in o fo r ca egorie ba ed on f nc ion and domain : cr alli a ion of $CaCO_3$, chi in rela ed pro ein , ECM rela ed pro ein and o her pro ein .

Crystallization of $CaCO_3$: ACCBP, CaLP, Nacrein. Amorpho calci m carbona e-binding pro ein (ACCBP) i a member of . he ace lcholine-binding pro ein famil . ha a r . i ola ed from . he e rapallial id of *P. fucata*[31]. I ha been repor ed . ha ACCBP inhibi . nde ired cr al gro h and pla a ke role in forming . he e ceedingl orderl micro r c re of nacre[31]. ACCBP con ain an ID region near . he N- ermin (ACCN) . ha ha been implica ed in reg la ion of $CaCO_3$ precipi a ion[32]. BLASTp earch iden i ed nine gene i h eq ence imilari . o ACCBP (E- al e < e-21, core > 100). Ph logene ic anal i of he e nine gene formed . o clade (Fig. 2), of hich ACCBP, ACCBP-like, CGI_10024902, and CGI_10024903 ere cl ered in o one clade. i gge . ha CGI_10024902 and CGI_10024903 di erged from a common ance ral gene of ACCBP. e e . o gene are loca ed on . he ame ca old146, gge ing an in ragenic d plica ion. Bo h pro ein con ain . o ne ro ran mi er-ga ed ion-channel ligand-binding domain (NCLBD), a ell a ACCBP and ACCBP-like.

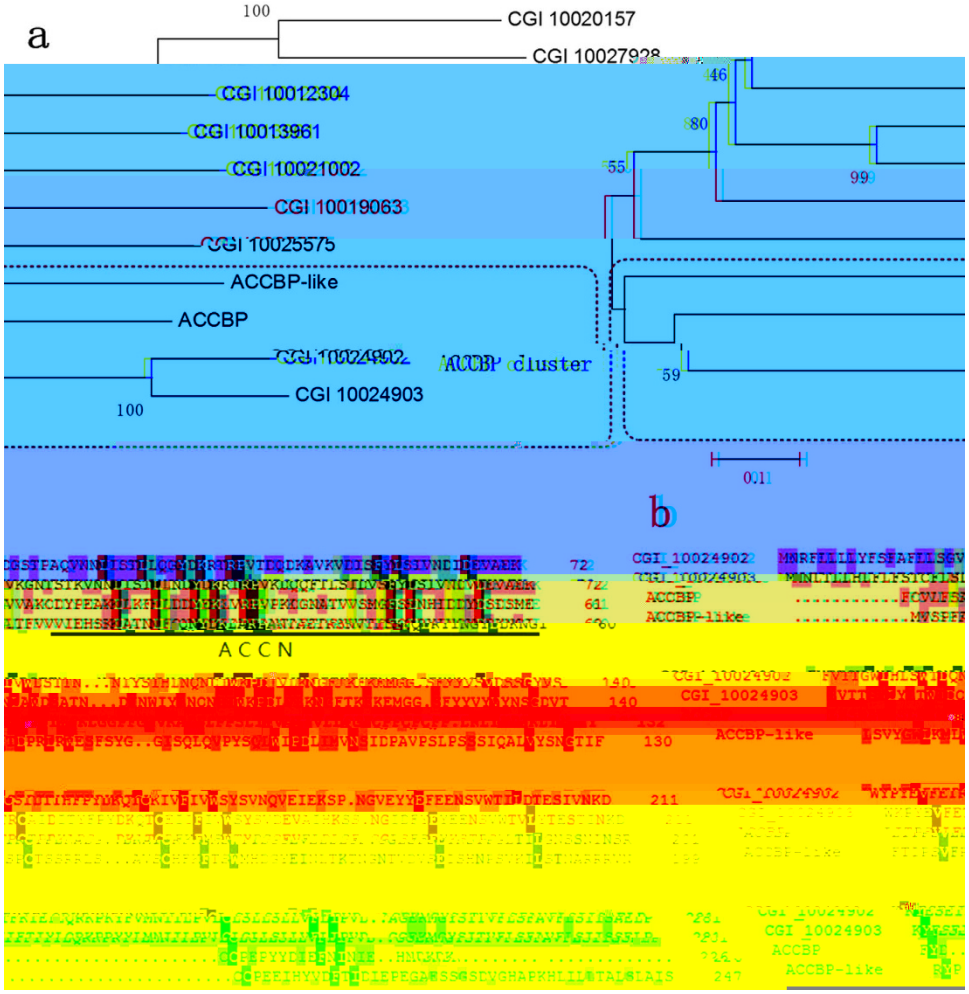

Figure 2. (**a**). Molec lar ph logene ic ree of he nine ACCBP hi s in *C. gigas* and o ACCBP in P. f ca a. A ph logene ic ree as inferred from he amino acid seq ences sing he neighbor-joining me hod. Boo s rap al es from 1000 rials are indica ed a each branch node. e scale bar indica es 0.1 amino acid replacemen s per si e. b. Comparison of ne ro ransmi er-ga ed ion-channel ligand binding domains (NCLBD) of ACCBP. e ACCN seq ence is nderlined.

Seq ence alignmen sho ed ha he conser ed resid es are di eren from hose resid es be een ACCBP and he nAChR famil (NCLBD-con aining pro eins in *D. melanogaster*)[32], especiall in he ACCN region (Fig. 2). ere are signi can l more conser ed resid es among ACCBP, ACCBP-like and he o pro eins in ACCN seq ence, hich are ass med o pla a role in minerali a ion ac i i .

Calmod lin-like pro ein (CaLP) is a m l if nc ional calci m sensor ha belongs o a ne member of he CaM s perfamil [33,34]. In bi al e, CaLP con ains o Ca^{2+}-binding EF hand domains, each of hich con ains a pair of EF-hand mo if. Imm nos aining re ealed ha CaLP as locali ed in he organic la er sand iched be een nacre (aragoni e) and he prisma ic la er (calci e) and he prisma ic la er in *P. fucata*[35]. ese res l s gges ed ha CaLP migh be in ol ed in he gro h of nacre la er and prisma ic la er. We ha e iden i ed hree p a i e genes encoding CaLP sing BLASTn search (E- al e < E-05, score > 50). BLASTp search iden i ed 26 gene models (E- al e < E-05 and score > 100), hich incl ded he aforemen ioned hree gene models in he bes si hi s. Using he bes si pro ein seq ences in BLASTp, he ph logene ic ree sho ed ha CGI_10011294 and CaLP cons i e a single clade, as ell as CGI_10011293 and CaM (S pplemen ar Fig S2). erefore, e anno- a ed CGI_10011294 as CaLP, and CGI_10011293 as CaM.

Nacrein is he r. iden i ed moll scan organic ma ri componen [36]. Nacrein e presses hro gho he en ire man le epi heli m and i f nc ions in he prod c ion of bo h prisma ic and nacreo s la er[37,38]. Nacrein con ains o f nc ional domains, a CA and a NG-repea domain i h a repea ed seq ence rich in Asn and Gl [36]. Nacrein-rela ed pro eins ha e been fo nd in some congeneric species (*P. maxima* and *P. margaritifera*) and also one gas ropod, *Turbo marmoratus*[39]. e ha e a similar primar s r c re as nacrein in *P. fucata*[11]. No nacrein-rela ed genes co ld be iden i ed sing BLASTp search i h E- al e < E-21 and score > 100. Ho e er, fo r nacrein-rela ed genes ere iden i ed i h E- al e < E-05 and score > 80. Seq ence alignmen sho ed ha he fo r hi s ha e no NG-repea domain and corresponding ac i e si e[39]. Ph logene ic ree of he aforemen ioned pro ein seq ences sho ed o di eren cl s ers of nacrein gro p and hi s gro p (S pplemen ar Fig. S3). ese

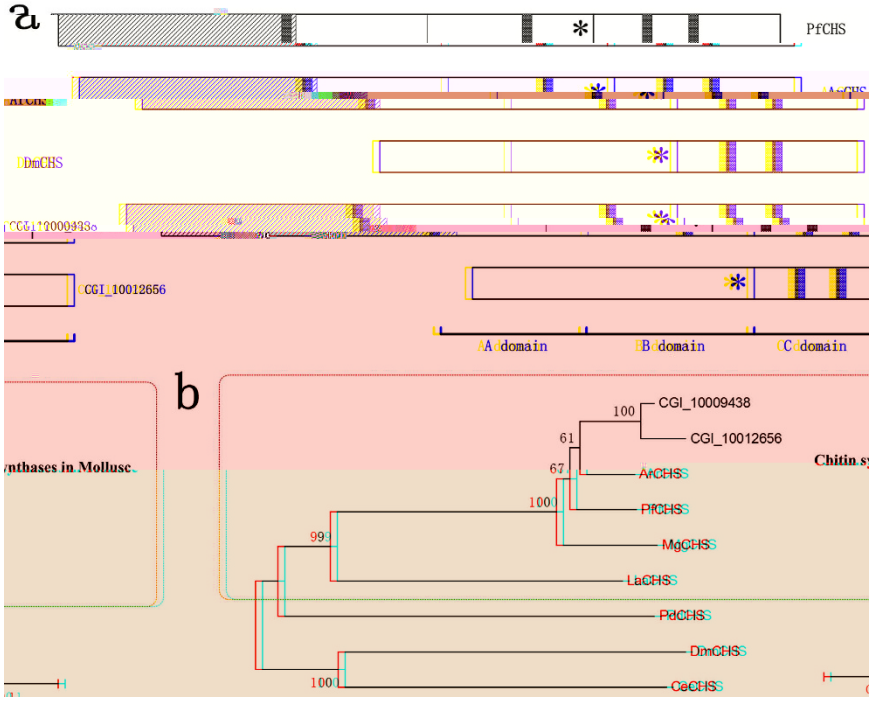

Figure 3. Diagrams of domain structures and phylogenetic tree analyses. (**a**). Diagramma ic repre en a ion of he domain r c re of chi in n ha e in di eren pecie . e m o in head domain i repre en ed b a la hed bo . Predic ed coiled-coil region are par ial haded. A. erik repre en he po i ion of he chi in n ha e ac i e i e (QRRRW), hich i ed o align he diagram. e hree domain in he e pro ein are deno ed A, B, and C. PfCHS, ArCHS and DmCHS are chi in n ha e from *Pinctada fucata*, *Atrina rigida* and *Drosophila melanogaster*[66]. (**b**). A ph logene ic ree ho ing he e ol ionar rela ionship of he predic ed chi in n ha e in *C. gigas* i h kno n chi in n ha e. e ree a con r c ed ing ma im m likelihood me hod. Seq ence in he QRRRW ca al ic domain ere aligned o genera e he ree. N mber on he node indica e boo. rap al e. ArCHS (AAY86556.1), PfCHS (BAF73720.1), MgCHS (ABQ08059.1), LaCHS (AHX26699.1), PdCHS (AHX26708.1), DmCHS (AAG09735.1) and CeCHS (NP_493682.2) are chi in n ha e from *A. rigida*, *P. fucata*, *Mytilus galloprovincialis*, *Leptochiton asellus*, *Platynereis dumerilii*, *Drosophila melanogaster*, *Caenorhabditis elegans*.

fo r hi are no peci call or highl e pre ed in man le. Taken oge her, e arg e ha he fo r hi a CA, b no nacrein, gge ing he NG-repea domain in CA pro ein a acq ired independen l in he lineage of bi al e and ga ropod.

e hree SMP ha e been repor ed o par icipa e in concen ra ion of Ca^{2+} and CO_3^{2-}, cr al n clea ion, and inhibi ion of cr al gro h.

Chitin Related Proteins: Chitin synthase, Clp. Chi in e i ing in bo h nacreo and prisma ic organic ma ri i ho gh o pla impor an role in biominerali a ion[40]. Chi in n ha e are e pre ed in man le edge, con rib ing o he forma ion of he frame ork for hell calci ca ion[29]. Chi in n ha e charac eri ed o far compri e hree domain, A, B, and C[29]. I i no e or h ha he chi in n ha e iden i ed in Moll ca hare a pecial fea re-a m o in domain in he N- ermin [29,41]. We ha e iden i ed o predic ed gene model (CGI_10009438, CGI_10012656) ha are homologo o chi in n ha e ing BLASTn earch (E- al e < E-63, core > 250). e o predic ed gene model ere loca ed o di eren ca old. e encode o di eren predic ed pro ein ha ere al o iden i ed b BLASTp i h E- al e = 0, core > 1800.

We aligned he amino acid eq ence of chi in n ha e and di co ered ha he e o pro ein incl ded all domain (a m o in domain and A, B, C domain) in *PfCHY* (Fig. 3). COILS anal i of CGI_10009438 ho ed a rong po en ial for coiled coil forma ion a fo r po i ion i h *PfCHS*. Speci call , hree po i ion ho ed a rong po en ial and one po i ion ho ed a fairl eak po en ial. COILS anal i of CGI_10012656 ho ed rong po en ial for coiled coil forma ion a he ame o rela i e po i ion i h DmCHS. e ph logene ic ree rongl con rmed ha he p a i e chi in n ha e from *C.gigas* belong o he chi in n ha e famil (Fig. 3). e moll k chi in n ha e gro p incl ding PfCHS, ArCHS, MgCHS i epara ed from o her chi in n ha e gro p b forming an independen cl er. i po i ion of CgCHS in he ph logene ic ree i con i en i h he e ol ionar po i ion in Moll ca. e e da a gge ed ha he chi in n ha e of moll can pecie gained he coiled-coil eq ence in he N- erminal region d ring e ol ion.

Chi ina e-like pro ein (Clp) ere . de ec ed from *P. margaritifera* and *P. maxima*. Clp ran crip ere locali ed in he man le edge peci call implica ed in he biominerali a ion of he prism[12]. Clp3 in *P. margaritifera* incl de in a region of GH18_chi ina e-like, hich can h drol e chi in. U ing he Clp3 eq ence,

14 gene models are identified by BLASTp with E-value < E-39, Score > 150. Of these, three hits CGI_10026599, CGI_10024867, CGI_10026605 are described to specially expressed in mantle of *C.gigas*[13], which strongly suggested that these gene models are related with shell formation. These three hits contain GH18 domain, showing hydrolysis of chitin activity.

Both of chitinase and chitinase play key roles in construction and reconstruction of chitin framework, just like chitin-silk broin gel proteins-acidic macromolecules model.

ECM related Proteins: Pif and BMSP, EGF-ZP, SPARC. Pif is an acidic matrix protein which regulates nacre formation. It was rst identified in the pearl oyster *P. fucata*. The Pif gene encoded a precursor protein, which was post-translationally cleaved to produce Pif 97 and Pif 80, respectively. Pif 97 has two conserved domains, a VWA domain for protein-protein interaction and chitin-binding domain. Pif 80 has aragonite-binding activity[42]. Sequence analysis revealed that bivalve mollusk shell protein (BMSP) is a Pif homologous preproprotein in *M. galloprovincialis*. BMSP consisted of a signal peptide and two proteins, BMSP 120 and BMSP 100, respectively[43]. In addition, other Pif homologues from bivalves (*P. margaritifera, P. maxima*, and *Pteria penguin*) and gastropods (*L. gigantea*) have been identified, suggesting a common ancestral gene during evolution[43].

We have identified six putative BMSP genes by BLASTp search (E-value < E-05 and score > 100), but no putative BMSP gene hits using BLASTn search (E-value < E-05, score > 50). Of these, the gene CGI_10009194 is homologous to BMSP with E-value = 0, score = 1415. Another two genes, CGI_10014497 and CGI_10017473, consist of the VWA domains and a CBD. We have identified ten putative Pif genes by BLASTp (E-value < E-05 and score > 100) using Pif177, but none using BLASTn search (E-value < E-05, score > 50). In addition to the former three genes (CGI_10014497, CGI_10017473, CGI_10009194), one more gene (CGI_10006697) with both VWA domains and a CBD is considered to be homologue of Pif, other have neither VWA domain no CBD. In phylogenetic analysis, the CGI 10009194 and BMSP formed a separate clade, suggesting that CGI 10009194 should be annotated as BMSP. The CGI 10014497 and CGI 10017473 clustered with Pif genes from *L. gigantea* with a low condence value. CGI 10006697 formed a single clade with others and showed distant evolution, not being annotated as Pif (Fig. 4). Phylogenetic analysis of VWA domain showed that the former three VWA domains in CGI 10009194 formed a single clade consisten l with the former three in BMSP that were evolved from BMSP-4 and CGI 10009194-4 (Fig. 4). These data indicated that BMSP and CGI 10009194 were likely evolved from an ancestral with four VWA domains. The phylogenetic tree showed that VWA domains in CGI 10014497 and CGI 10017473 have evolved from CGI 10009194-4 and Pif177, being annotated as *Pif-like1* and *Pif-like2*. Schematic representation of Pif showed the common VWA domains, CBDs, chitin-binding like domains and various C-terminal aragonite-binding sequences (Fig. 4).

Epidermal growth factor (EGF) domain-containing SMPs were rst identified from *C. gigas*[44] and named as Cgigas-IMSP-2. These were subsequently discovered in *P. maxima, P. margaritifera,* and *L. gigantea*[12,14]. Generally, the SMP consists of both EGF-like domain and one ona pellucida (ZP) domain. The presence of both domains in one protein is uncommon[12]. EGF-like domains, which are characteristic conserved cysteines linked in a characteristic pattern of disulfide bonds and two small β-sheets, are among the smallest but most idely distributed models found in extracellular proteins. EGF-like domains occur in a variety of proteins associated with diverse biological functions such as cell adhesion, signaling, and Ca^{2+}-binding[45]. The ZP domains are present in a range of extracellular lamentous matrix proteins. The ZP domain is characteristically conserved cysteine residues, which are involved in protein polymerization. The specic function of the EGF- and ZP-containing SMPs in calcied shell biomineralization and ali alidation. Three EGF-ZP genes were identified by BLASTp using the known IMSP-2 genes of *C. gigas* with (E-value < E-48, score > 100). CGI_10017543, CGI_10017544, and CGI_10017545 are located on the same scaffold 120. In combination with their relative level high degree of sequence identity (37.53%), suggesting that these originated from a gene duplication event. A sequence alignment of these EGF-ZPs illustrated a strong conservation of each domain (signal peptide, EGF, ZP), suggesting a fundamental role in biomineralization (Fig. S4).

SPARC (secreted protein, acidic, rich in cysteine), also known as BM-40 or osteonectin, is a major noncollagenous matrix protein of bone and a common mineralization-related protein of other ebrates and molluscs[45]. The primary structure of SPARC is characterized by the presence of three functional domains: the N-terminal acidic domain I; the follistatin-like domain II with 10 conserved cysteine residues; and the C-terminal domain III, which is involved in interactions with collagen molecules[46]. By searching the *C. gigas* genome, we could only identify a single SPARC gene (CGI_10005088). Besides, the *C. gigas* genome contains one SPARC gene as observed in most triploblastic organisms. The SPARC protein consists of a region Kazal type serine protease inhibitors, follistatin-like domains (KAZAL_FS), and a region extracellular Ca^{2+} binding domain (SPARC_EC). KAZAL_FS can inhibit serine proteases and play an important role in tissue specic regulation. SPARC_EC functions to regulate cell-matrix interactions and binds to proteins such, acidic, rld a ro kni1.60000038(P)84-6..199999861(es a)0000038(

and o ida ion of *o*-diphenol . I is h s also clas i ed as a phenolo idase. T ro ina e is ell kno n for i s ke biological role in melanin bios n he is ia ransforma ion of ro ine o L-DOPA. T ro ina e f nc ions in pigmen a ion and inna e imm ni [47]. In addi ion, o her prod c s of he melanin pa h a par icipa e in c icle sclero i a ion in insec s[49].

In Moll sca, ro ina e ha e been s gges ed in pigmen a ion and biomineraliza ion of seashell. Cephalopod ro ina es are e pre ed in he ink sac, s gges ing an impor an role in melanin prod c ion[50]. In *P. fucata*, hree ro ina e genes ha e been charac eri ed, P 1 and P 2 are s gges ed o f nc ion in pris ma ic forma ion and OT47 is propo ed o in ence he perios rac m forma ion[51,52]. In *C. gigas, Cgtyr1* as cloned and propo ed o peci call f nc ion in he ini ial phase of he lar al shell biogenesis[23]. *CgTyr2* as also cloned and sho ed high le els of e pression in man le edges. I has been s gges ed o pla a role in he forma ion of perios rac m/pigmen a ion[24]. ese repor s s rongl s gges ha ro ina es pla di erse roles in s ages hen seashell is cons r c ed, pigmen ed, and co ered i h he perios rac m.

T en si ro ina e genes ere iden i ed from *C. gigas* genome. T o ro ina e genes CGI_10007793 and CGI_10011913 ere fo nd o be iden ical o he repor ed *Cgtyr1* and *CgTyr2,* respec i el (Fig. 5). e ro ina e gene famil can be f r her clas i ed in o hree pes: secre ed form i h signal pep ides (T pe A), c o olic form (T pe B) and member-bo nd form (T pe C). According o SignalP 4.0, and TMHMM Ser er 2.0, here are si T peA ro ina e genes, 15 T pe B ro ina e genes and e T pe C ro ina e genes. e ph logene ic ree of 26 ro ina es sho ed ha here are eigh pairs of d plica ion genes. Among hem, onl o pairs of CGI_10021076 T pe C and CGI_10021075 T pe A, CGI_10009319 T pe A and CGI_10009318 T pe B are loca ed in he same sca old separa el , belonging o in ragenic d plica ion. e ph logene ic ree sho ed ha he cl s ers of ro i

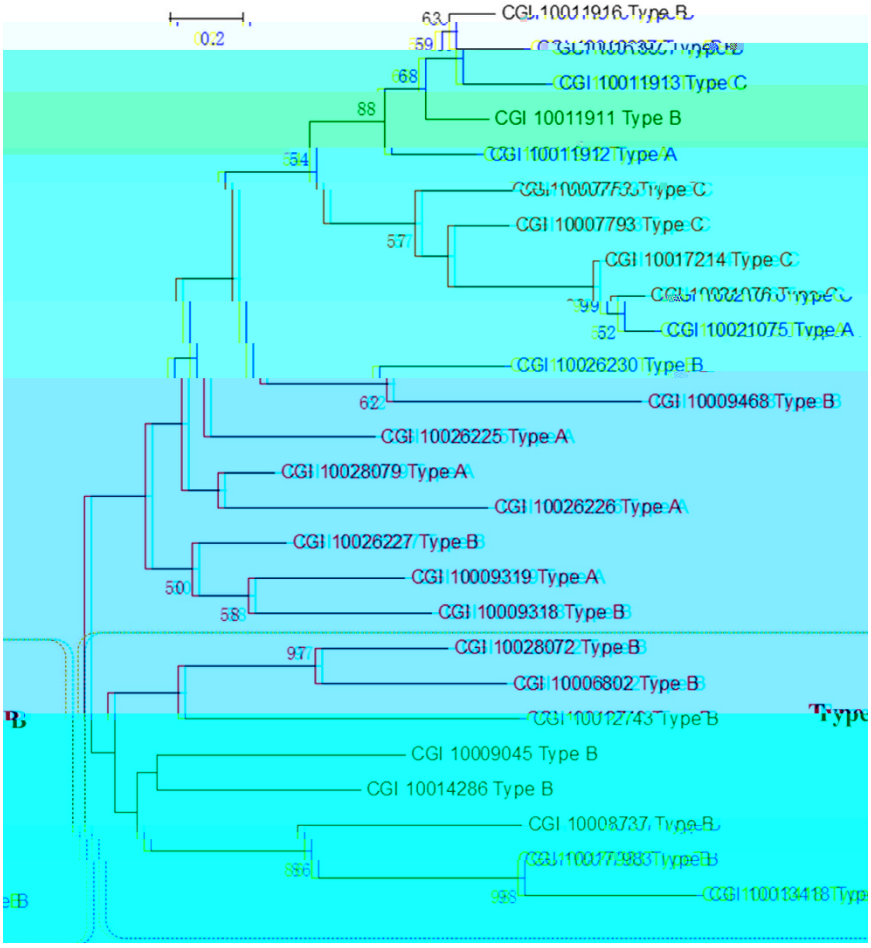

Figure 5. Tyrosinase genes diversity in *C. gigas*. e ph logene ic ree of ro inases as cons r c ed b he ma im m likelihood me hod. N mbers on he node indica e boo s rap al es. e ro inase gene are named b acces ion n mber and pe.

T pe B, CGI_10011913 T pe C, CGI_10011911 T pe B and CGI_10011912 T pe A are cl s ered, all of hich are loca ed in Sca old 43702. Abo e all, i s gges s ha he ro inases ha e e ol ed hro gh bo h in ergenic d plica ion and in ragenic d plica ion.

Pero idases are iron pro eins ha ca al e he o ida ion of man aroma ic amines and phenols b h drogen pero ide. Inac i a ion e perimen s pro ide s rong e idence ha he DOPA reac ion in he man le of *Lymnaea stagnalis* is ca al ed b pero idase, s gges ing ha pero idase is in ol ed in he q inone-anning of peri- os rac m pro eins[53]. Pero idase is e cl si el e pressed in he ink gland of *Sepia officinalis*, likel in ol ed in melanin bios n hesis[54]. T pical pero idase is charac eri ed b a niq e s r c ral domain ha con ains o his- idines (pro imal and dis al his idines) and one calci m-binding si e, hich as s gges ed o f nc ion in main- aining he pro ein s r c re in he heme en ironmen[54,55]. Pero idase has been re rie ed from he shell ma ri of *P. margaritifera* and *L. gigantea*. I as proposed o be in ol ed in biomineral-h drogel forma ion ia pro ein ma ri frame ork assembl [12]. We iden i ed 26 pero idase gene b BLASTp sing he kno n pero idase gene of *P. margaritifera* i h (E- al e < E-23, s core > 100). e pero idase gene ha e gone hro gh an e pansion as sho n for ro inases in *C. gigas*. Nine o of he 26 pero idase ere speciall or highl e pressed in man le of *C. gigas*. e o be hi s (CGI_10023200 and CGI_10010240) are speci call e pressed in man le[13]. All nine pero idases consis of o his idines and one calci m-binding si e. Ph logene ic anal sis sho ed ha hese per- o idases formed o cl s ers ha co ld be iden i ed as melanin bios n hesis gro p and he shell forma ion gro p (Fig. 6). Pero idases from *Drosophila melanogaster*, *Bombyx mori* and *Sepia officinalis* ha e been implica ed in melanin s n hesis form melanin pol mer[54].

We no ed in o r bioinforma ics anal ses ha in man cases, BLASTn ga e no hi s, ho e er, BLASTp gi e hi s (Table S2). is s gges ed ha genes encoding SMPs ha e ndergone more aria ion han SMPs pro ein seq ences. Un il no , li era re on biominerali a ion is ca ered i h hese iden i ed SMPs. Ob io sl , SMPs do more han pro iding a frame ork for cr s alli a ion. e co ld be in ol ed in in erac ions i h o her mac- romolec lar componen s of he ma ri and cells, gi ing s b le feedbacks be een he shell and he calcif ing man le epi heli m.

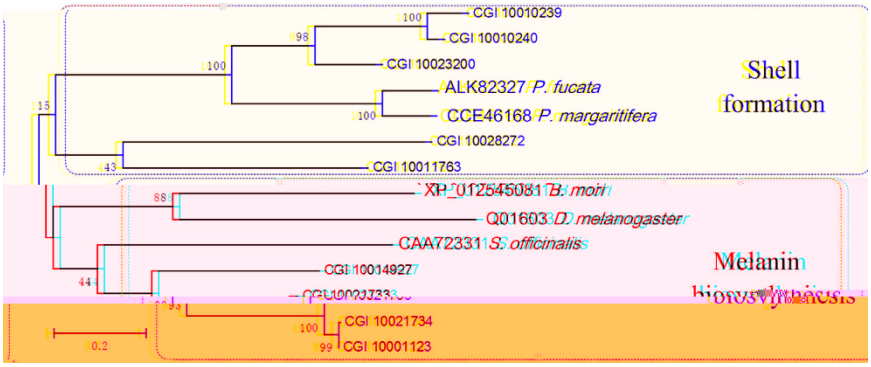

Figure 6. Peroxidase genes diversity in *C. gigas*. e ph logene ic ree of pero idase as cons r c ed b he ma im m likelihood me hod. N mbc enga

C E C In *C. gigas*, a SMPE consi ing of 53 pro eins as recen l p blished[18]. e SMPE as cons r c ed from pep ide fragmen s in shell, hich can be mapped o he genome. In addi ion, a SMPE in *C. gigas* consi ing of 259 pro eins as also anal ed, hich as cons r c ed b Zhang and colleag es sing he same me hod. Man ho se-keeping pro eins, s ch as elonga ion fac or 1α and ribosomal pro eins, ere fo nd in hese 259 SMPs, hich are signi can l more han SMPs iden i ed from o her moll scan shell pro eome o da e. Generall ho se-keeping pro eins sho ld no be fo nd in shell ma ri pro eome and he do no pla special role in biominerali a ion. Gi en ha ho sekeeping genes are generall e pressed a rela i el cons an le els in mos non-pa hological si a ions, e iden i ed genes ha are specif- icall or highl e pressed in man le o minimi e he in erference from ho sekeeping pro eins. An in egra ed SMPE consi ing of 76 SMPs as cons r c ed. e re ined SMPE as cons r c ed from he in ersec ion of o da abases:1) 259 pro eins isola ed from o s er shells; 2) 492 genes ha ere speci call or highl e pressed in man le, of hich, he highl e pressed genes ere de ned as ha ing RPFM al es of a leas 5 and a leas e imes of o her organ a erage[13]. In rinsicall n r c red/disordered (ID) domains ere predic ed b IUPRED[56] and XSTREAM[57]. IUPRED as sed o recogni e disordered regions from he amino acid seq nces of SMPs based on he es ima ed pair ise energ con en . XSTREAM as sed o de ec pro eins i h andem-arranged repea ni s in he defa l se ings.

C Based on BLASTp, he global similari comparison of *C. gigas* SMPE as performed agains 443 SMPs deri ed from nine o her biocalcif ing me a oans. ese incl ded 53 *M. coruscus* pro eins[15], 75 *P. fucata* pro eins[10], 45 *P. margaritifera* pro eins[12], 26 *P. maxima* pro eins[12], 39 *L. gigantea* pro eins[14], 14 *H. asinina* pro eins[8], 59 *C. nemoralis* pro eins[6], 67 *M. truncata*[17], and 65 *M. venosa*[2] pro eins. All pro ein seq nces are alida ed b mapping o he ranscrip omes, ESTs in biocalcif ing organs or genome assemblies. e e- al e hreshold as se o 1e-06. ese comparisons ere made sing blas + [58]: blas p -q er XX.fa -db XX.change.fas a -o XX.blp -o fm 6 -e al e 10e-6 -n m_ hreads 10. e *.blp les genera ed b blas + ere modi ed sing c s om Perl scrip s and hen passed o Circos in order o genera e an ideogram[59]. e *.blp le i h de ails of he BLASTp res l s is pro ided (S pplemen ar Table S4).

I C e SMP searches in he *C. gigas* gene models (o s er_gene_ 1) and ind ced pro ein seq nces (o s er_pep. ides_ 1) ere performed sing O s erbase[13]. Shell forma ion complemen ar DNAs (cDNAs) in *C.gigas* ere BLASTn and BLASTp searched. Iden i es of he hi gene models associa ed i h he original cDNAs ere con rmed b seq nce alignmen . Similarl , SMPs iden i ed from o her moll scan species ere BLAST searched, and he ob ained gene models ere reciprocall BLASTp searched agains he NCBI nonred ndan (nr) da abase o con rm he bes -hi seq nce.

C C e conser ed s r c ral domains ere e amined sing he SMART[60] and In erproScan[61]. e amino acid seq nces ere aligned sing MEGA5[62] or DNAMAN (L non Biosо). For ph logene ic anal sis, poorl aligned posi ions ere checked and remo ed man all , he ph - logene ic ree i h boo s rap al e as cons r c ed sing MEGA5. Pro ein seq nces for he ph logene ic anal sis ere re rie ed from GenBank, S iss-Pro , or he O s erbase. In he case of secre or pro eins or pep ides, he pres- ence of a signal pep ide as predic ed b SignalP 4.0[63]. COILS as sed o de ec coiled coil forma ion[64]. To iden- if membrane pro eins, he TMHMM Ser er 2.0 predic ion algori hm as sed for ransmembrane helices[65].

C We compared o di eren SMPEs of *C. gigas* consi ing of 76 SMPs and 53 SMPs and cho se he la er one o perform a broad le el comparison b bioinforma ic anal sis. ese SMPE ere charac eri ed b ha ing a high propor ion of ID pro eins, especiall RLCD pro eins. We sed a marine SMPE in *C. gigas* o perform a broad comparison agains 443 SMPs from nine o her species sing BLASTp. O r da a con rms he earlier ndings ha he SMPs similari depends no onl on he e ol ionar dis ance b is in enced b mineralog of shell,

parallel e ol . ion, adap. a. ion . o . he en ironmen. e. c. e highl con.er. ed pro. eins . ro.inase and chi. o. rio-.idase are iden. i ed in bi al e, and . he rela. i el con.er. ed pro. eins i. h domains of CA, VWA, CBD, IG-like and LaG are iden. i ed from all . en .pecies. 25 genes encoding SMPs ere anno. a. ed and charac. eri ed . ha. are chi. in rela. ed or ECM rela. ed pro. eins in ol ed in cr .. alli a. ion of $CaCO_3$. ese con.er. ed SMPs and . ni er.al domains enrich . he molec lar kno ledge of .hell forma. ion mechani.m in *C. gigas*, . rging for a re ned .hell forma. ion model incl ding bo. h chi. in and ECM-rela. ed pro. eins.

1. Sim i..., . & Wilb r, . M. *Biomineralization* (El.e ier, 2012).
2. Jac .on, D. J. *et al.* e *Magellania venosa* biominerali ing pro. eome: a indo in. o brachiopod .hell e ol . ion. *Genome biology and evolution* **7,** 1349 1362 (2015).
3. noll, A. H. Biominerali a. ion and e ol .ionar his. or . *Reviews in mineralogy and geochemistry* **54,** 329 356 (2003).
4. M. oco. , ., Ag ilera, F., McDo gall, C., J.Jac .on, D. & M.Degnan, B. Sea .hell di er.i and rapidl e ol ing .ecre. ome.: insigh. . in. o he e ol . ion of biominerali a. ion. *Frontiers in Zoology* **13,** 1 (2016).
5. Marie, B., Le_ o , N., Zanella-Cleon, I., Becchi, M. & Marin, F. Molec lar e ol . ion of moll .c .hell pro. eins: insigh. . from pro. eomic anal .is of . he edible m .. el *Mytilus*. *J Mol Evol* **72,** 531 46 (2011).
6. Mann, . & Jac .on, D. J. Charac. eri a. ion of . he pigmen. ed .hell-forming pro. eome of . he common gro e .nail *Cepaea nemoralis*. *BMC genomics* **15,** 1 (2014).
7. Ag ilera, F., McDo gall, C. & Degnan, B. M. Co-op. ion and de no o gene e ol . ion nderlie moll .can .hell di er.i . *Molecular Biology and Evolution* doi, h. .p.://doi.org/10.1093/molbe /m. 294 (2016).
8. Marie, B. *et al.* Pro. eomic anal .is of . he organic ma ri. of . he abalone *Haliotis asinina* calci ed .hell. *Proteome Sci* **8,** 54 (2010).
9. Marin, F. & L q e , G. Moll .can .hell pro. ein. *Comptes Rendus Palevol* **3,** 469 492 (2004).
10. Li , C. *et al.* In-dep. h pro. eomic anal .is of .hell ma ri. pro. eins of *Pinctada fucata*. *Scientific reports* **5** (2015).
11. Mi amo. o, H. *et al.* e di er.i of .hell ma ri. pro. eins: Genome- ide in .. iga. ion of . he pearl o .. er, *Pinctada fucata*. *Zoological science* **30,** 801 816 (2013).
12. Marie, B. *et al.* Di eren. .ecre. or reper. oires con. rol . he biominerali a. ion proce..es of pri.m and nacre depo.i. ion of . he pearl o .. er .hell. *Proceedings of the National Academy of Sciences* **109,** 20986 20991 (2012).
13. Zhang, G. *et al.* e o .. er genome re eals .. re.. adap. a. ion and comple_i. of .hell forma. ion. *Nature* **490,** 49 54 (2012).
14. Marie, B. *et al.* e .hell forming pro. eome of *Lottia gigantea* re eals bo. h deep con.er. a. ion. and lineage .peci c no el ie.. *FEBS Journal* **280,** 214 232 (2013).
15. Liao, Z. *et al.* In-dep. h pro. eomic anal .is of nacre, pri.m, and m o.. rac m of *Mytilus* .hell. *Journal of proteomics* **122,** 26 40 (2015).
16. Gao, P. *et al.* La er-b -La er Pro. eomic Anal .is of *Mytilus galloprovincialis* Shell. *PLOS ONE* **10** (2015).
17. Ari alagan, J. *et al.* Shell ma ri. pro. eins of . he clam, *Mya truncata*:_ ole. be ond .hell forma. ion . hro gh pro. eomic .. d . *Marine Genomics* **27,** 69 74 (2016).
18. Ari alagan, J. *et al.* Insigh. . from . he .hell pro. eome: biominerali a. ion . o adap. a. ion. *Molecular Biology and Evolution* **34,** 66 77 (2016).
19. F r ha.hi, T., Sch ar inger, C., Mi .i , I., Smr , M. & Beran, A. Moll .can .hell e ol . ion i. h re ie of .hell calci ca. ion h po. he.i.. *Comparative biochemistry and physiology Part B: Biochemistry and molecular biology* **154,** 351 371 (2009).
20. John.. one, M. *et al.* Cell lar orche.. ra. ed biominerali a. ion of cr .. alline compo.i. e. on implan. .. rface.. b . he ea.. ern o .. er, *Crassostrea virginica* (Gmelin, 1791). *Journal of Experimental Marine Biology and Ecology* **463,** 8 16 (2015).
21. Mo n , A. S., Wheeler, A. P., Parad ar,_ . P. & Snider, D. Hemoc . e-media. ed .hell minerali a. ion in . he ea.. ern o .. er. *Science* **304,** 297 (2004).
22. Y , H., Zhao, X. & Li, Q. Genome- ide iden. i ca. ion and charac. eri a. ion of long in. ergenic noncoding_ NAs and . heir po. en. ial a..ocia. ion i. h lar. al de elopmen. in . he Paci c o .. er. *Scientific reports* **6** (2016).
23. H an, P., Li , G., Wang, H. & Li , B. Iden. i ca. ion of a . ro.inase gene po. en. iall in ol ed in earl lar. al .hell biogene.i. of . he Paci c o .. er *Crassostrea gigas*. *Development genes and evolution* **223,** 389 394 (2013).
24. Y , X. *et al.* Molec lar cloning and di eren. ial e_pre..ion in . i.. e. of a . ro.inase gene in . he Paci c o .. er *Crassostrea gigas*. *Molecular biology reports* **41,** 5403 5411 (2014).
25. E an., J. S. Aragoni. e-a..ocia. ed biominerali a. ion pro. eins are di.ordered and con. ain in. erac. i e mo. if.. *Bioinformatics* **28,** 3182 3185 (2012).
26. Jac .on, D. *et al.* Parallel e ol . ion of nacre b ilding gene .e. . in moll .c.. *Molecular biology and evolution* **27,** 591 608 (2010).
27. Feng, D., Li, Q., Y , H., Zhao, X. & ong, L. Compara. i e Trans-crip. ome Anal .is of . he Pacific O .. er *Crassostrea gigas* Charac. eri ed b Shell Color.: Iden. i ca. ion of Gene. ic Ba.e. Po. en. iall In ol ed in Pigmen. a. ion. *PloS one* **10** (2015).
28. Ag ilera, F., McDo gall, C. & Degnan, B. M. E ol . ion of . he . ro.inase gene famil in bi al e moll .c.: independen. e_pan..ion of . he man. le gene reper. oire. *Acta biomaterialia* **10,** 3855 3865 (2014).
29. S . i, M., Sa da, S. & Naga.a a, H. Iden. i ca. ion of chi. in in . he pri.ma. ic la er of . he .hell and a chi. in .. n ha.e gene from . he Japane.e pearl o .. er, *Pinctada fucata*. *Bioscience, biotechnology, and biochemistry* **71,** 1735 1744 (2007).
30. Wei.., I. M., Sch ni er, V., Eichner, N. & S mper, M. e chi. in .. n ha.e in ol ed in marine bi al e moll .. .hell forma. ion con. ain. a m o.in domain. *FEBS letters* **580,** 1846 1852 (2006).
31. Ma, Z. *et al.* A no el e_trapallial id pro. ein con. rol. . he morpholog of nacre lamellae in . he pearl o .. er, *Pinctada fucata*. *Journal of Biological Chemistry* **282,** 23253 23263 (2007).
32. Amo., F. F., Ndao, M. & E an., J. S. E idence of minerali a. ion ac. i i. and .. pramolec lar a..embl b . he N- erminal .eq ence of ACCBP, a biominerali a. ion pro. ein . ha. i. homologo .. o . he ace. lcholine binding pro. ein famil . *Biomacromolecules* **10,** 3298 3305 (2009).
33. Li, S. *et al.* Cloning and e_pre..ion of a pi o. al calci m me. abolism reg la. or: calmod lin in ol ed in .hell forma. ion from pearl o .. er (*Pinctada fucata*). *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* **138,** 235 243 (2004).
34. Li, S., Xie, L., Ma, Z. & Zhang,_ . cDNA cloning and charac. eri a. ion of a no el calmod lin, li e pro. ein from pearl o .. er *Pinctada fucata*. *Febs Journal* **272,** 4899 4910 (2005).
35. Yan, Z. *et al.* Biominerali a. ion: f nc. ion. of calmod lin-li e pro. ein in . he .hell forma. ion of pearl o .. er. *Biochimica et Biophysica Acta (BBA)-General Subjects* **1770,** 1338 1344 (2007).
36. Mi amo. o, H. *et al.* A carbonic anh dra.e from . he nacreo .. la er in o .. er pearl.. *Proceedings of the National Academy of Sciences* **93,** 9657 9660 (1996).
37. Mi amo. o, H., Mi o.hi, F. & ohno, J. e carbonic anh dra.e domain pro. ein nacrein i. e_pre..ed in . he epi. helial cell. of . he man. le and ac.. a. a nega. i e reg la. or in calci ca. ion in . he moll .c *Pinctada fucata*. *Zoological science* **22,** 311 315 (2005).
38. Ta e chi, T. & Endo, . Bipha.ic and d all coordina. ed e_pre..ion of . he gene. encoding major .hell ma ri. pro. eins in . he pearl o .. er *Pinctada fucata*. *Marine biotechnology* **8,** 52 61 (2006).
39. Nori i, M. & Sama a, T. Dis. rib . ion and f nc. ion of . he nacrein-rela. ed pro. eins inferred from .. r c. ral anal .is. *Marine Biotechnology* **10,** 234 241 (2008).

40. F r ha hi, T. *et al.* P rol is GC/MS and I pec. ro cop in chi. in anal is of moll can shell. *Bioscience, biotechnology, and biochemistry* **73,** 93 103 (2009).
41. Wei e, I. M. & Sch ni er, V. e dis. rib . ion of chi. in in lar al shell of he bi al e moll *Mytilus galloprovincialis*. *Journal of structural biology* **153,** 264 277 (2006).
42. S i, M. *et al.* An acidic ma ri pro ein, Pif, is a e macromolec le for nacre forma ion. *Science* **325,** 1388 1390 (2009).
43. S i, M. *et al.* Iden i ca ion and Charac eri a ion of a Calci m Carbona e Binding Pro ein, Bl e M el Shell Pro ein (BMSP), from he Nacreo La er. *Chembiochem* **12,** 2478 2487 (2011).
44. Marie, B., Zanella-Cl on, I., G ichard, N., Becchi, M. & Marin, F. No el pro eins from he calcif ing hell ma ri of he Paci c o er *Crassostrea gigas*. *Marine biotechnology* **13,** 1159 1168 (2011).
45. Ma rer, P. & Hohene er, E. S. r c ral and f nc. ional a pec of calci m binding in e racell lar ma ri pro eins. *Matrix biology* **15,** 569 580 (1997).
46. Sa a i, T., Hohene er, E., G hring, W. & Timpl, . Cr al r c re and mapping b i e direc ed m agenei of he collagen binding epi ope of an ac i a ed form of BM 40/SPA C/o eonec in. *The EMBO Journal* **17,** 1625 1634 (1998).
47. Garc a Borr n, J. C. & Solano, F. Molec lar Ana om of T rosinase and i ela ed Pro eins: Be ond he His idine Bo nd Me al Ca al ic Cen er. *Pigment Cell Research* **15,** 162 173 (2002).
48. Ag ilera, F., McDo gall, C. & Degnan, B. M. Origin, e ol ion and clas i ca ion of pe-3 copper pro eins: lineage-pecif c gene e pan ions and lo e acro s he Me a oa. *BMC evolutionary biology* **13,** 96 (2013).
49. Ander en, S. O. Ins c c ic lar clero i a ion: a re ie . *Insect biochemistry and molecular biology* **40,** 166 178 (2010).
50. Narao a, T. *et al.* P ri ca ion, charac eri a ion and molec lar cloning of ro inase from he cephalopod moll , *Illex argentinus*. *European Journal of Biochemistry* **270,** 4026 4038 (2003).
51. Nagai, ., Yano, M., Morimo o, . & Mi amo o, H. T ro inase locali a ion in moll c hell. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* **146,** 207 214 (2007).
52. Zhang, C., Xie, L., H ang, J., Chen, L. & Zhang, . A no el p a i e ro inase in ol ed in perio rac m forma ion from he pearl o er (*Pinctada fucata*). *Biochemical and biophysical research communications* **342,** 632 639 (2006).
53. Timmerman, L. P. S. die on hell forma ion in moll c. *Netherlands Journal of Zoology* **19,** 413 523 (1968).
54. Ge aldo, I., Aniello, F., Branno, M. & Pal mbo, A. Molec lar cloning of a pero idase m NA peci call e pre ed in he in gland of *Sepia officinalis*. *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression* **1353,** 111 117 (1997).
55. Shiro, Y., rono, M. & Mori hima, I. Pre ence of endogeno calci m ion and i f nc. ional and r c ral reg la ion in hor eradi h pero idase. *Journal of Biological Chemistry* **261,** 9382 9390 (1986).
56. Do n i, Z., Ci m , V., Tompa, P. & Simon, I. IUPred: eb er er for he predic ion of in rin icall n r c red region of pro eins ba ed on e ima ed energ con en . *Bioinformatics* **21,** 3433 3434 (2005).
57. Ne man, A. M. & Cooper, J. B. XST EAM: A prac ical algori hm for iden i ca ion and archi ec re modeling of andem repea in pro ein eq ence. *BMC Bioinformatics* **8,** 382 (2007).
58. Camacho, C. *et al.* BLAST+: archi ec re and applica ion. *BMC bioinformatics* **10,** 1 (2009).
59. r in i, M. *et al.* Circo : an informa ion ae he ic for compara i e genomic. *Genome research* **19,** 1639 1645 (2009).
60. Le nic, I., Doer , T. & Bor , P. SMA T: recen pda e, ne de elopmen and a a in 2015. *Nucleic Acids Research* **43,** 257 260 (2015).
61. Q e illon, E. *et al.* In erProScan: pro ein domain iden i er. *Nucleic Acids Research* **33,** 116 120 (2005).
62. Tam ra, . *et al.* MEGA5: molec lar e ol ionar gene ic anal is ing ma im m li elihood, e ol ionar di ance, and ma im m par imon me hod. *Molecular biology and evolution* **28,** 2731 2739 (2011).
63. Pe er en, T. N., Br na , S., on Heijne, G. & Niel en, H. SignalP 4.0: dis crimina ing ignal pep ide from ran membrane region. *Nature methods* **8,** 785 786 (2011).
64. L pa , A., Van D e, M. & S oc , J. Predic ing coiled coil from pro ein eq ence. *Science* **252,** 1162 1164 (1991).
65. M ller, S., Croning, M. D. & Ap eiler, . E al a ion of me hod for he predic ion of membrane panning region. *Bioinformatics (Oxford, England)* **17,** 646 653 (2001).
66. Tellam, . L., V ocolo, T., John on, S. E., Jarme , J. & Pear on, . D. Ins c chi in n ha e. *European Journal of Biochemistry* **267,** 6025 6043 (2000).

A

This d a pp or ed b he gran from Na ional Na ral Science Fo nda ion of China (31372524), Shandong Seed Projec , Shandong Pro ince (2016ZDJS06A06), and Qingdao Na ional Labora or for Marine Science and Technolog (2015ASKJ02).

A C

D.D.F. and S.H.D. anal ed he da a and ro e he paper. Q.L., H.Y. and L.F.K. concei ed and designed he d .

A I

Supplementary information accompanie hi paper a h p:// .na re.com/rep

Competing Interests: e a hor declare no compe ing nancial in ere .

How to cite this article: Feng, D. *et al.* Iden i ca ion of con er ed pro ein from di er e hell ma ri pro eome in *Crassostrea gigas:* charac eri a ion of gene ic ba e reg la ing hell forma ion. *Sci. Rep.* **7**, 45754; doi: 10.1038/ rep45754 (2017).