**Supplementary Material online**

# Co-option and *de novo* gene evolution underlie molluscan shell diviersity

Felipe Aguilera[1,2], Carmel McDougall[1] and Bernard M. Degnan[1*]

[1]Centre for Marine Sciences, School of Biological Sciences, The University of Queensland, Brisbane 4072, Australia
[2]Current address: Sars International Centre for Marine Molecular Biology, University of Bergen, Thormøhlensgate 55, Bergen 5008, Norway

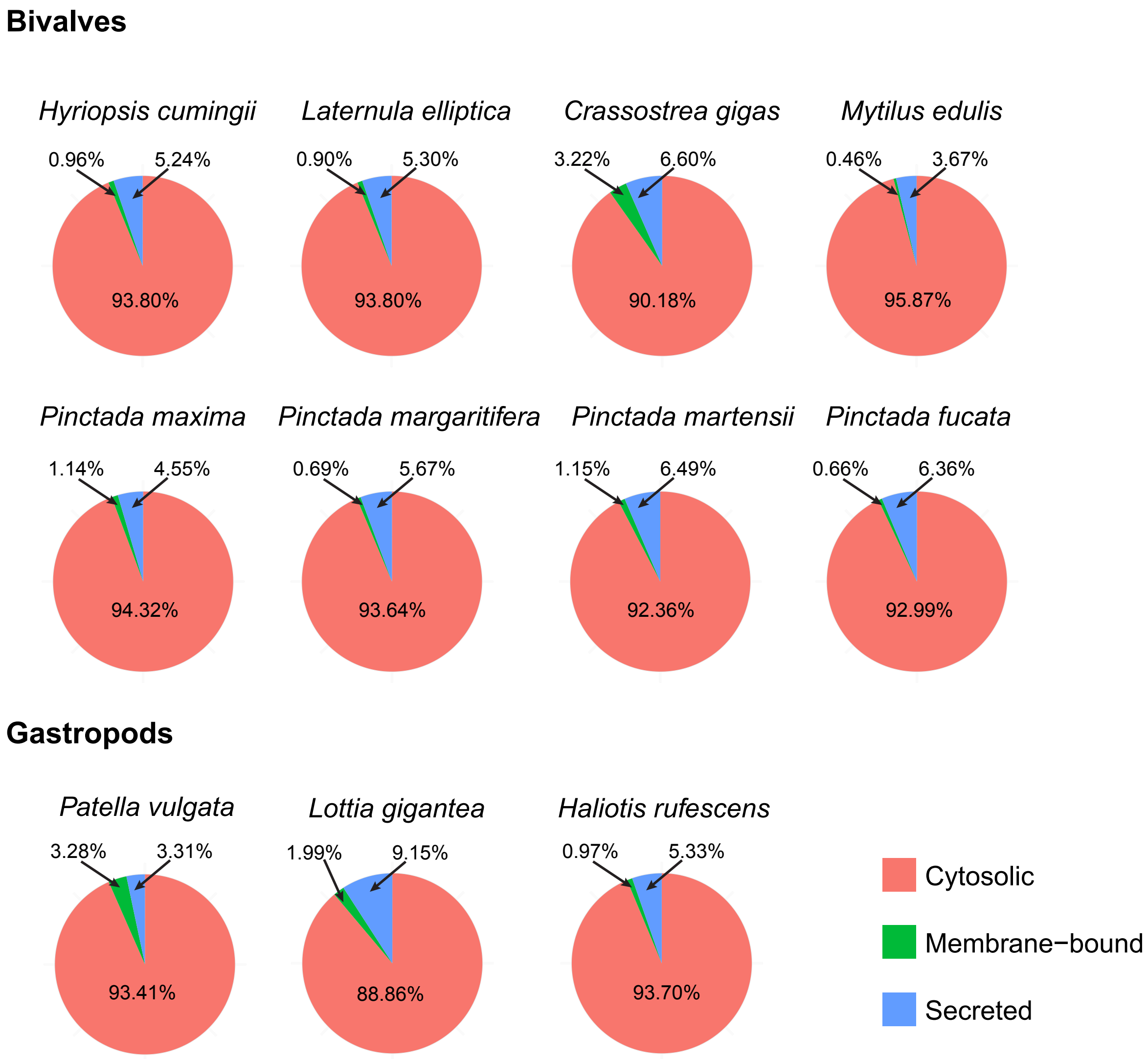


**Figure S1. Computational predictions of cytosolic, membrane-bound and secreted proteins for each species.** Pie charts represent the percentage of proteins predicted in each mantle transcriptome. Red color indicates cytosolic proteins; green color indicates membrane-bound proteins; and blue color indicates secreted proteins.

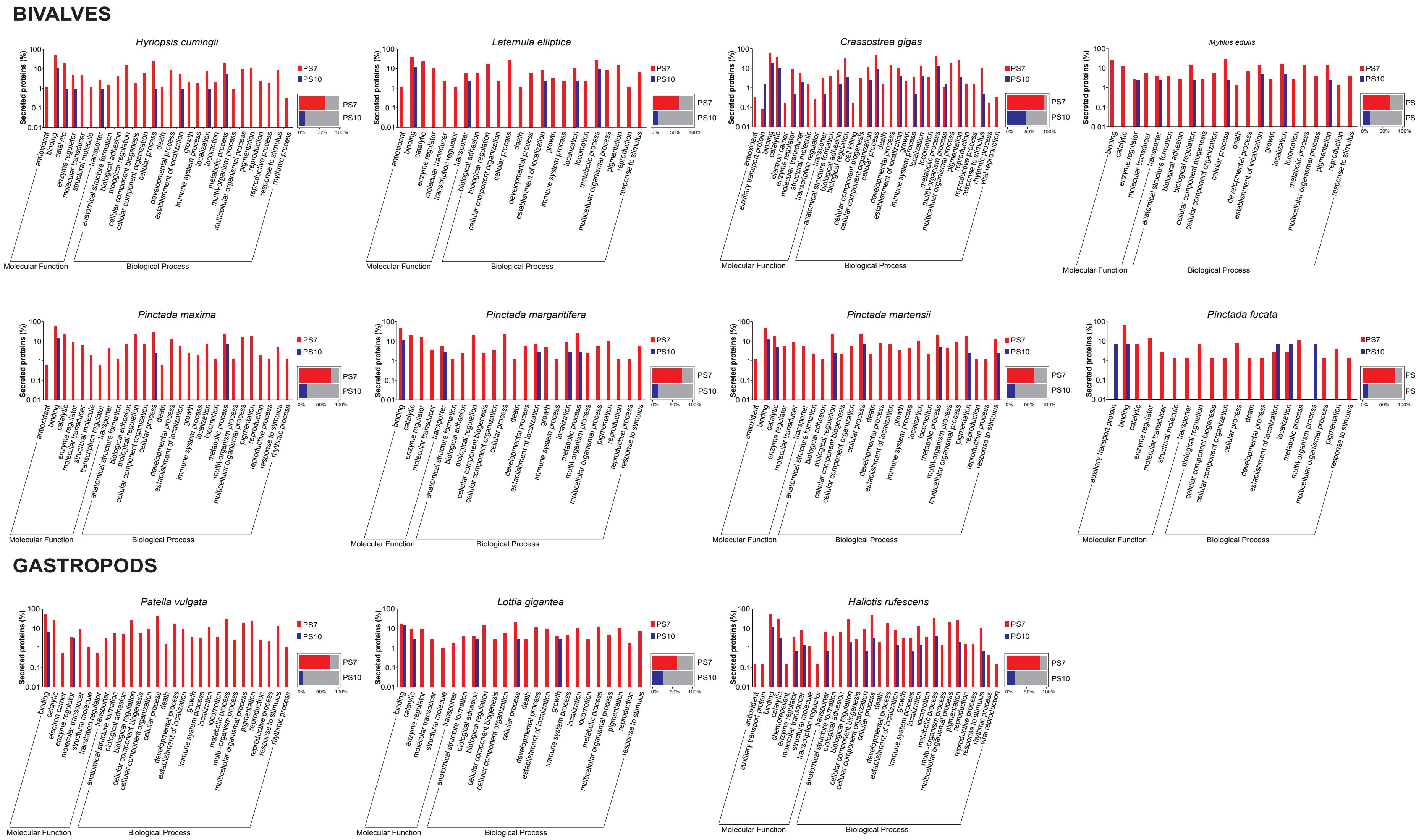

BIVALVES
Hyriopsis cumingii
Laternula elliptica
Crassostrea gigas
Mytilus edulis
Pinctada maxima
Pinctada margaritifera
Pinctada martensii
Pinctada fucata
GASTROPODS
Patella vulgata
Lottia gigantea
Haliotis rufescens
Secreted proteins (%)
100
10
1
0.1
0.01
PS7
PS10
0%
50%
100%
Molecular Function
Biological Process

**Figure S2. GO annotations of genes that arose along the stems leading to bilaterian (PS7, red) and molluscan (PS10, blue) last common ancestors.** Horizontal bars - in small squares - represent the percentage of genes that were annotated (red and blue colors) and non-annotated (gray color) in each species.

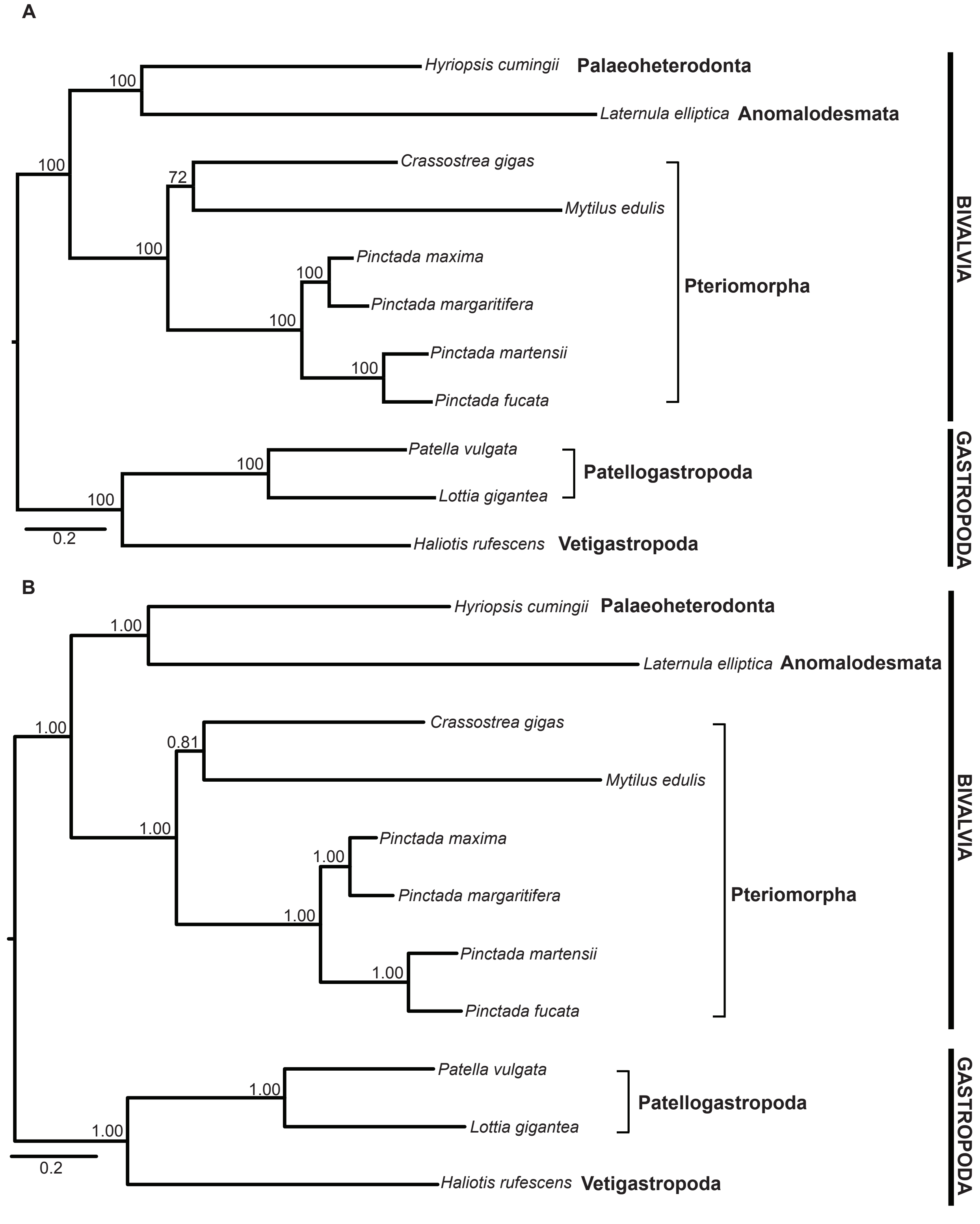


**Figure S3. Phylogenetic relationships among bivalve and gastropod lineages based on 122 gene families that encode secreted proteins. A.** Phylogenetic tree based on Maximum Likelihood is shown. The most-likely tree (log likelihood = -138,848.94) sampled in RAxML using the PROTGAMMAWAGF substitution model. ML bootstrap (BS) support values are listed at each node. **B.** A 50% majority-rule consensus Bayesian tree is shown.

Posterior probabilities (PP) are indicated at each node. The length of the matrix is 13,604 amino acids.

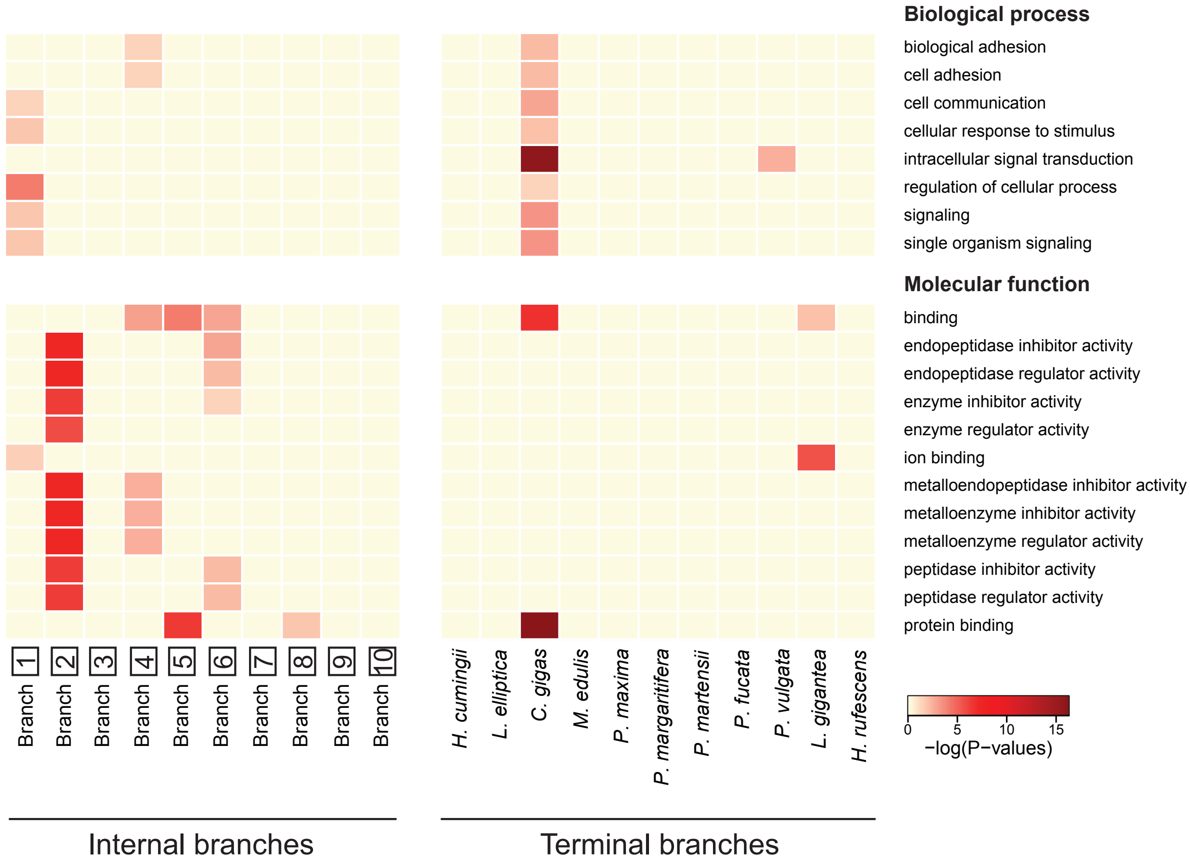


**Figure S4. GO term (biological process and molecular function) enrichment in newly gained secreted gene families across conchiferan evolution.** Only significantly enriched (P<0.05, Fisher's exact test) GO terms at least present in two phylogenetic branches are shown. GO terms (biological process and molecular function) categories of the over-represented secreted gene families at each evolutionary time point are shown at the right. For a comprehensive list of enriched GO terms across molluscan evolution, see Supplementary Table S4, Supplementary Material online.

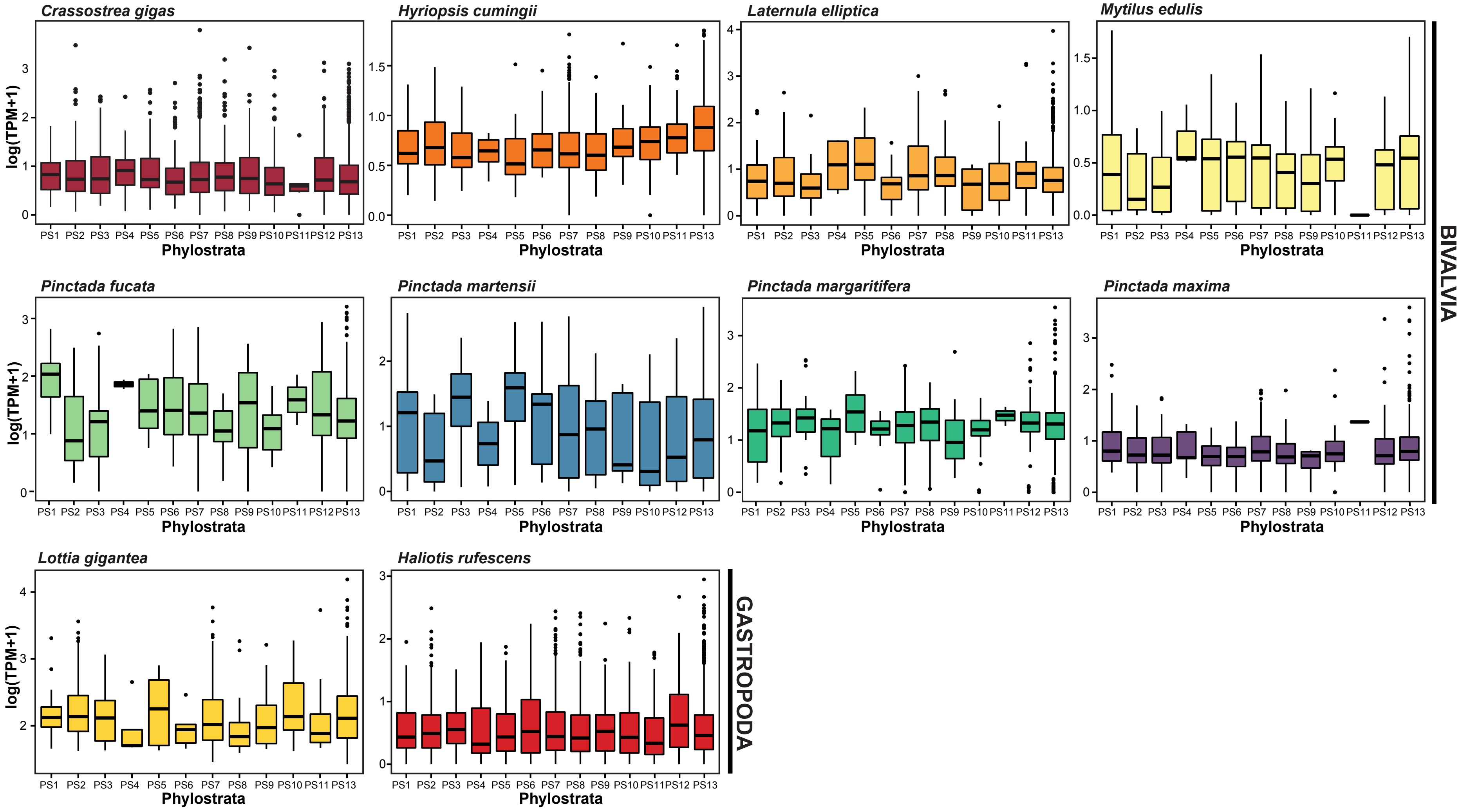


**Figure S5. Expression profiles of genes encoding secreted proteins – the mantle secretome – of eight bivalve and two gastropod species.** The boxplots show the median and interquartile ranges of the distribution of gene expression across phylostrata for bivalves and gastropods. Phylostratum (PS) is described and illustrated as per Figure 1.

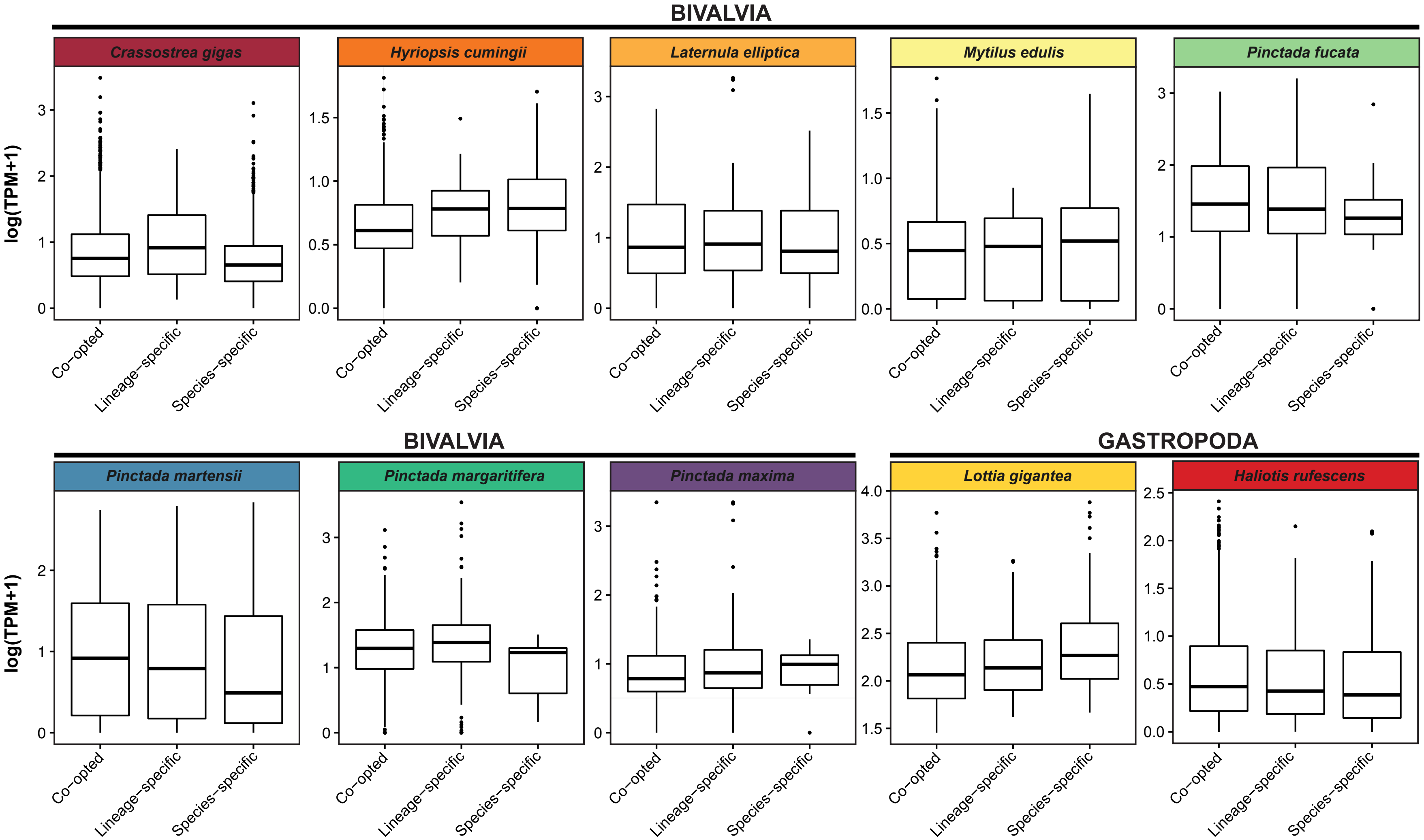


**Figure S6. Expression levels of co-opted, lineage-specific and species-specific genes encoding secreted proteins.** Boxplots display the distribution of relative gene expression of genes classified as co-opted, lineage-specific and species-specific for each eight bivalves and two gastropods. Lines in the boxplots represent the median and interquartile range. Dots correspond to outlier genes.

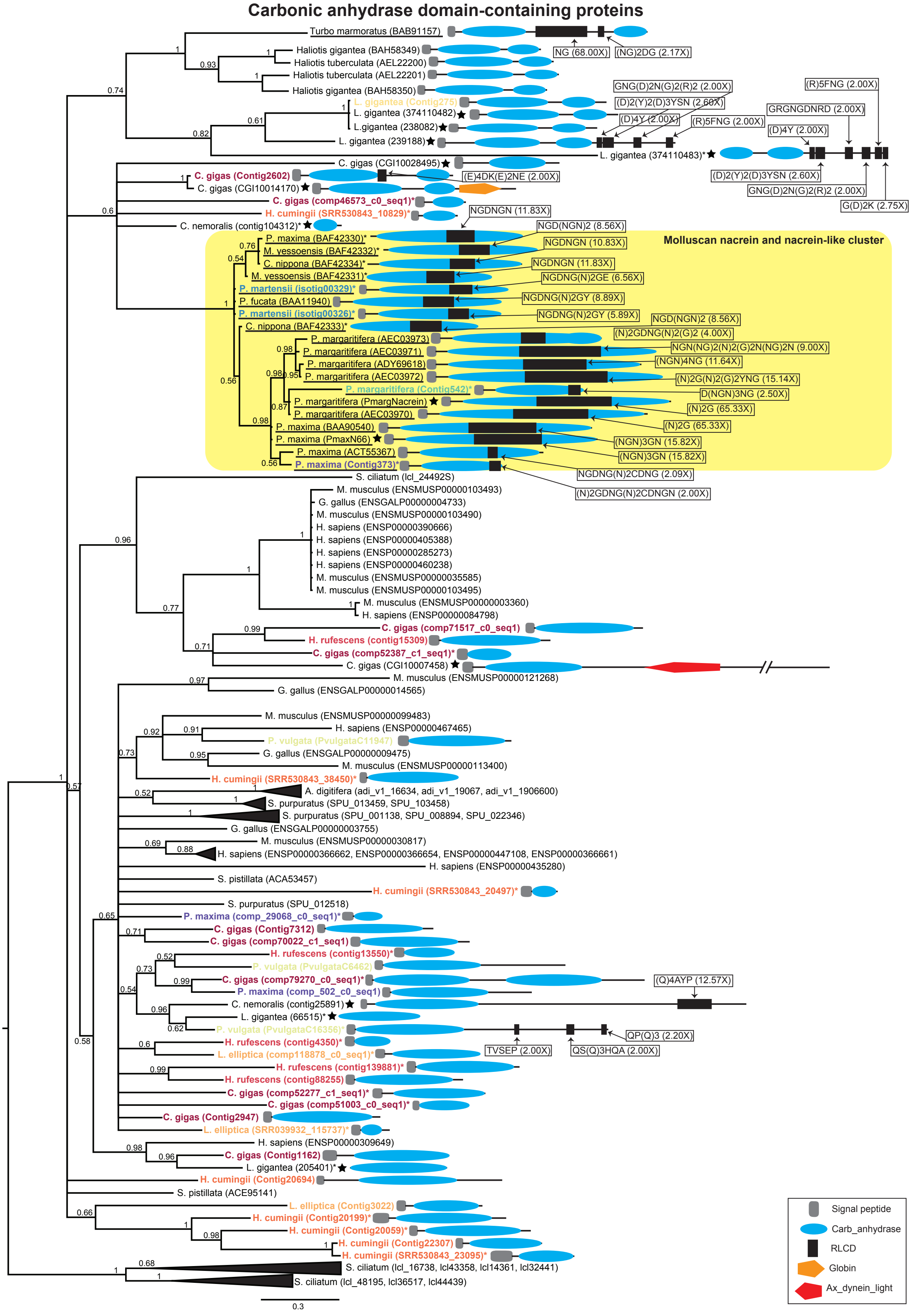
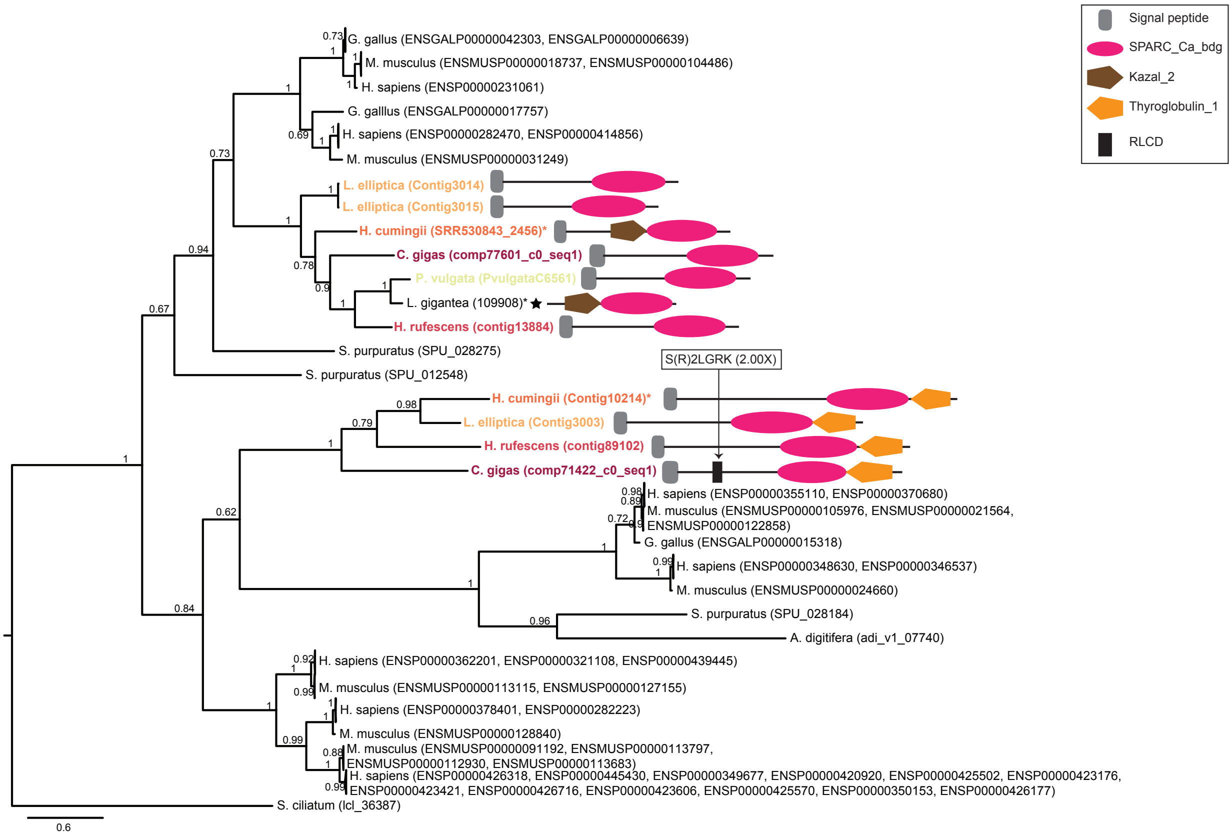

Carbonic anhydrase domain-containing proteins
Turbo marmoratus (BAB91157)
NG (68.00X)
(NG)2DG (2.17X)
1
Haliotis gigantea (BAH58349)
Haliotis tuberculata (AEL22200)
0.93
1
Haliotis tuberculata (AEL22201)
Haliotis gigantea (BAH58350)
0.74
L. gigantea (Contig275)
L. gigantea (374110482)
L.gigantea (238082)
L. gigantea (239188)
L. gigantea (374110483)*
1
0.61
0.82
GNG(D)2N(G)2(R)2 (2.00X)
(D)2(Y)2(D)3YSN (2.60X)
(D)4Y (2.00X)
(R)5FNG (2.00X)
(R)5FNG (2.00X)
GRGNGDNRD (2.00X)
(D)4Y (2.00X)
(D)2(Y)2(D)3YSN (2.60X)
GNG(D)2N(G)2(R)2 (2.00X)
G(D)2K (2.75X)
C. gigas (CGI10028495)
1
C. gigas (Contig2602)
C. gigas (CGI10014170)
(E)4DK(E)2NE (2.00X)
C. gigas (comp46573_c0_seq1)*
H. cumingii (SRR530843_10829)*
0.6
C. nemoralis (contig104312)*
NGDNGN (11.83X)
NGD(NGN)2 (8.56X)
Molluscan nacrein and nacrein-like cluster
P. maxima (BAF42330)*
M. yessoensis (BAF42332)*
C. nippona (BAF42334)*
M. yessoensis (BAF42331)*
P. martensii (isotig00329)*
P. fucata (BAA11940)
P. martensii (isotig00326)*
C. nippona (BAF42333)*
0.76
0.54
1
NGDNGN (10.83X)
NGDNGN (11.83X)
NGDNG(N)2GE (6.56X)
NGDNG(N)2GY (8.89X)
NGDNG(N)2GY (5.89X)
NGD(NGN)2 (8.56X)
(N)2GDNG(N)2G(G)2 (4.00X)
P. margaritifera (AEC03973)
P. margaritifera (AEC03971)
P. margaritifera (ADY69618)
P. margaritifera (AEC03972)
P. margaritifera (Contig542)*
P. margaritifera (PmargNacrein)
P. margaritifera (AEC03970)
P. maxima (BAA90540)
P. maxima (PmaxN66)
P. maxima (ACT55367)
P. maxima (Contig373)*
0.98
0.98
0.95
0.56
0.87
0.98
0.56
NGN(NG)2(N)2(G)2N(NG)2N (9.00X)
(NGN)4NG (11.64X)
(N)2G(N)2(G)2YNG (15.14X)
D(NGN)3NG (2.50X)
(N)2G (65.33X)
(N)2G (65.33X)
(NGN)3GN (15.82X)
(NGN)3GN (15.82X)
NGDNG(N)2CDNG (2.09X)
(N)2GDNG(N)2CDNGN (2.00X)
S. ciliatum (lcl_24492S)
M. musculus (ENSMUSP00000103493)
G. gallus (ENSGALP00000004733)
M. musculus (ENSMUSP00000103490)
H. sapiens (ENSP00000390666)
H. sapiens (ENSP00000405388)
H. sapiens (ENSP00000285273)
H. sapiens (ENSP00000460238)
M. musculus (ENSMUSP00000035585)
M. musculus (ENSMUSP00000103495)
M. musculus (ENSMUSP00000003360)
H. sapiens (ENSP00000084798)
0.96
1
1
1
0.77
C. gigas (comp71517_c0_seq1)
H. rufescens (contig15309)
C. gigas (comp52387_c1_seq1)*
C. gigas (CGI10007458)
0.99
0.71
M. musculus (ENSMUSP00000121268)
G. gallus (ENSGALP00000014565)
0.97
M. musculus (ENSMUSP00000099483)
H. sapiens (ENSP00000467465)
P. vulgata (PvulgataC11947)
G. gallus (ENSGALP00000009475)
M. musculus (ENSMUSP00000113400)
H. cumingii (SRR530843_38450)*
0.92
0.91
0.95
0.73
1
0.57
A. digitifera (adi_v1_16634, adi_v1_19067, adi_v1_1906600)
S. purpuratus (SPU_013459, SPU_103458)
S. purpuratus (SPU_001138, SPU_008894, SPU_022346)
0.52
1
1
1
G. gallus (ENSGALP00000003755)
M. musculus (ENSMUSP00000030817)
H. sapiens (ENSP00000366662, ENSP00000366654, ENSP00000447108, ENSP00000366661)
0.69
0.88
H. sapiens (ENSP00000435280)
S. pistillata (ACA53457)
H. cumingii (SRR530843_20497)*
S. purpuratus (SPU_012518)
P. maxima (comp_29068_c0_seq1)*
0.65
C. gigas (Contig7312)
C. gigas (comp70022_c1_seq1)
0.71
H. rufescens (contig13550)*
P. vulgata (PvulgataC6462)
C. gigas (comp79270_c0_seq1)*
P. maxima (comp_502_c0_seq1)
0.52
0.73
0.99
0.54
(Q)4AYP (12.57X)
C. nemoralis (contig25891)
L. gigantea (66515)*
P. vulgata (PvulgataC16356)*
0.96
0.62
TVSEP (2.00X)
QS(Q)3HQA (2.00X)
QP(Q)3 (2.20X)
H. rufescens (contig4350)*
L. elliptica (comp118878_c0_seq1)*
0.6
0.58
H. rufescens (contig139881)*
H. rufescens (contig88255)
0.99
C. gigas (comp52277_c1_seq1)*
C. gigas (comp51003_c0_seq1)*
C. gigas (Contig2947)
L. elliptica (SRR039932_115737)*
H. sapiens (ENSP00000309649)
C. gigas (Contig1162)
L. gigantea (205401)*
0.98
0.96
H. cumingii (Contig20694)
S. pistillata (ACE95141)
L. elliptica (Contig3022)
H. cumingii (Contig20199)*
H. cumingii (Contig20059)*
H. cumingii (Contig22307)
H. cumingii (SRR530843_23095)*
0.66
1
0.98
1
S. ciliatum (lcl_16738, lcl43358, lcl14361, lcl32441)
S. ciliatum (lcl_48195, lcl36517, lcl44439)
1
0.68
1
0.3
Signal peptide
Carb_anhydrase
RLCD
Globin
Ax_dynein_light

**Figure S7. Phylogenetic analysis of carbonic anhydrase domain-containing proteins.** Bayesian phylogenetic tree was obtained under the LG+G substitution model. The tree is rooted using the calcareous sponge sequences as outgroup. Statistical support for each node is indicated as posterior probabilities after 20,000,000 generations. Domain architectures of molluscan secreted proteins are shown on the right of each sequence, and domain nomenclature is shown in the rectangle. *Nacrein* and *nacrein-like* proteins are in bold/underline. Molluscan nacreins are highlighted in the yellow box. Proteins previously extracted from shells and/or known to be involved in shell formation are indicated by a black star. For each RLCD, the consensus repeat sequence is shown, followed by the copy number (in brackets).

SPARC domain-containing proteins


G. gallus (ENSGALP00000042303, ENSGALP00000006639)
M. musculus (ENSMUSP00000018737, ENSMUSP00000104486)
H. sapiens (ENSP00000231061)
G. galllus (ENSGALP00000017757)
H. sapiens (ENSP00000282470, ENSP00000414856)
M. musculus (ENSMUSP00000031249)
L. elliptica (Contig3014)
L. elliptica (Contig3015)
H. cumingii (SRR530843_2456)*
C. gigas (comp77601_c0_seq1)
P. vulgata (PvulgataC6561)
L. gigantea (109908)* ★
H. rufescens (contig13884)
S. purpuratus (SPU_028275)
S. purpuratus (SPU_012548)
S(R)2LGRK (2.00X)
H. cumingii (Contig10214)*
L. elliptica (Contig3003)
H. rufescens (contig89102)
C. gigas (comp71422_c0_seq1)
H. sapiens (ENSP00000355110, ENSP00000370680)
M. musculus (ENSMUSP00000105976, ENSMUSP00000021564, ENSMUSP00000122858)
G. gallus (ENSGALP00000015318)
H. sapiens (ENSP00000348630, ENSP00000346537)
M. musculus (ENSMUSP00000024660)
S. purpuratus (SPU_028184)
A. digitifera (adi_v1_07740)
H. sapiens (ENSP00000362201, ENSP00000321108, ENSP00000439445)
M. musculus (ENSMUSP00000113115, ENSMUSP00000127155)
H. sapiens (ENSP00000378401, ENSP00000282223)
M. musculus (ENSMUSP00000128840)
M. musculus (ENSMUSP00000091192, ENSMUSP00000113797, ENSMUSP00000112930, ENSMUSP00000113683)
H. sapiens (ENSP00000426318, ENSP00000445430, ENSP00000349677, ENSP00000420920, ENSP00000425502, ENSP00000423176, ENSP00000423421, ENSP00000426716, ENSP00000423606, ENSP00000425570, ENSP00000350153, ENSP00000426177)
S. ciliatum (lcl_36387)
0.73
1
0.69
0.94
0.78
0.9
0.67
0.98
0.79
0.62
0.89
0.72
0.99
0.96
0.84
0.92
0.88
0.6
Signal peptide
SPARC_Ca_bdg
Kazal_2
Thyroglobulin_1
RLCD

Zona pellucida-like domain-containing proteins

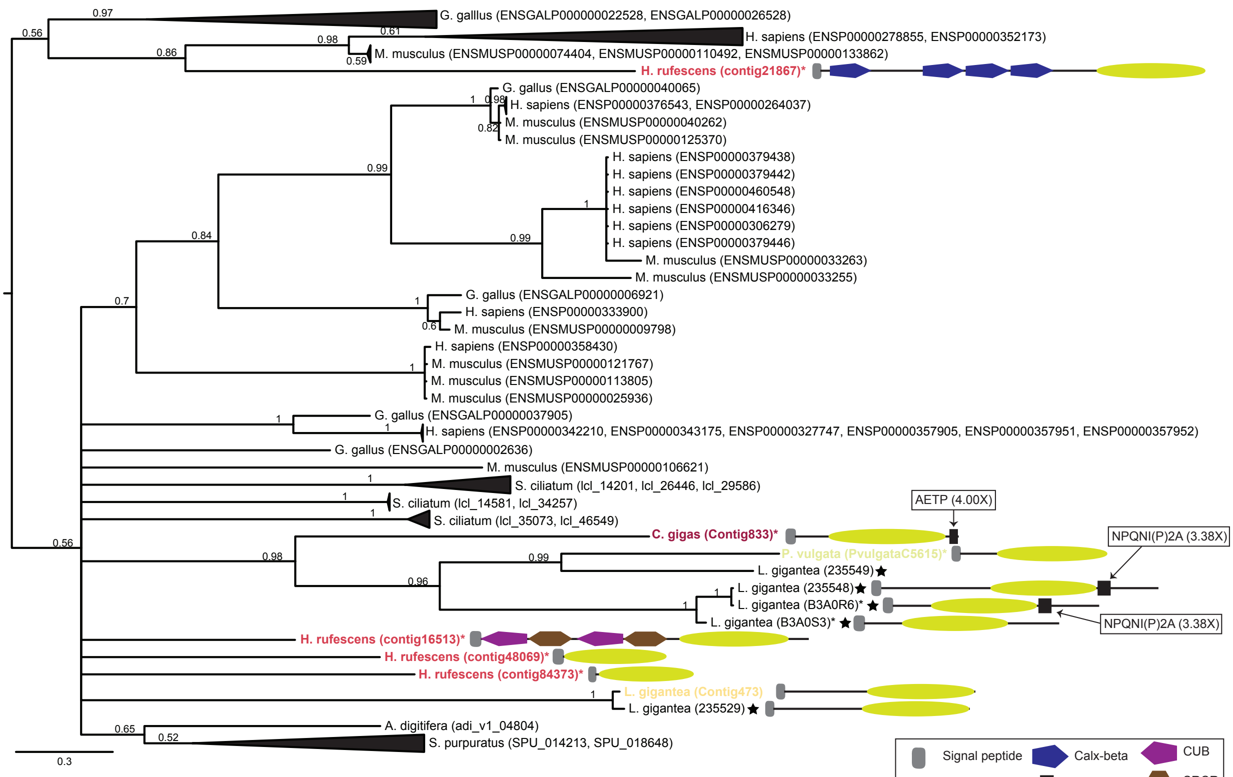

G. galllus (ENSGALP000000022528, ENSGALP00000026528)
H. sapiens (ENSP00000278855, ENSP00000352173)
M. musculus (ENSMUSP00000074404, ENSMUSP00000110492, ENSMUSP00000133862)
H. rufescens (contig21867)*
G. gallus (ENSGALP00000040065)
H. sapiens (ENSP00000376543, ENSP00000264037)
M. musculus (ENSMUSP00000040262)
M. musculus (ENSMUSP00000125370)
H. sapiens (ENSP00000379438)
H. sapiens (ENSP00000379442)
H. sapiens (ENSP00000460548)
H. sapiens (ENSP00000416346)
H. sapiens (ENSP00000306279)
H. sapiens (ENSP00000379446)
M. musculus (ENSMUSP00000033263)
M. musculus (ENSMUSP00000033255)
G. gallus (ENSGALP00000006921)
H. sapiens (ENSP00000333900)
M. musculus (ENSMUSP00000009798)
H. sapiens (ENSP00000358430)
M. musculus (ENSMUSP00000121767)
M. musculus (ENSMUSP00000113805)
M. musculus (ENSMUSP00000025936)
G. gallus (ENSGALP00000037905)
H. sapiens (ENSP00000342210, ENSP00000343175, ENSP0000327747, ENSP00000357905, ENSP00000357951, ENSP00000357952)
G. gallus (ENSGALP00000002636)
M. musculus (ENSMUSP00000106621)
S. ciliatum (lcl_14201, lcl_26446, lcl_29586)
S. ciliatum (lcl_14581, lcl_34257)
S. ciliatum (lcl_35073, lcl_46549)
AETP (4.00X)
C. gigas (Contig833)*
NPQNI(P)2A (3.38X)
P. vulgata (PvulgataC5615)*
L. gigantea (235549) ★
L. gigantea (235548) ★
L. gigantea (B3A0R6)* ★
L. gigantea (B3A0S3)* ★
NPQNI(P)2A (3.38X)
H. rufescens (contig16513)*
H. rufescens (contig48069)*
H. rufescens (contig84373)*
L. gigantea (Contig473)
L. gigantea (235529) ★
A. digitifera (adi_v1_04804)
S. purpuratus (SPU_014213, SPU_018648)
0.97
0.56
0.98
0.61
0.86
0.59
0.82
0.99
0.84
0.7
0.6
0.96
0.65
0.52
0.3
Signal peptide
Calx-beta
CUB
Zona_pellucida
RLCD
SRCR

**Figure S8. Phylogenetic analyses of SPARC domain-containing proteins and zona pellucida-like domain-containing proteins.** Bayesian phylogenetic trees were obtained under the WAG+G and LG+I+G substitution models, respectively. The secreted protein acidic and rich in cysteine Ca binding domain-containing protein tree is rooted using the calcareous sponge sequence as outgroup, while the Zona pellucida domain-containing protein tree is rooted using the midpoint-rooted option. Statistical support for each node is indicated as posterior probabilities after 1,000,000 and 6,000,000 generations, respectively. Domain architectures of molluscan secreted proteins are shown on the right of each sequence, and domain nomenclature is shown in the rectangles. Proteins previously extracted from shells and/or known to be involved in shell formation are indicated by a black star. For each RLCD, the consensus repeat sequence is shown, followed by the copy number (in brackets).

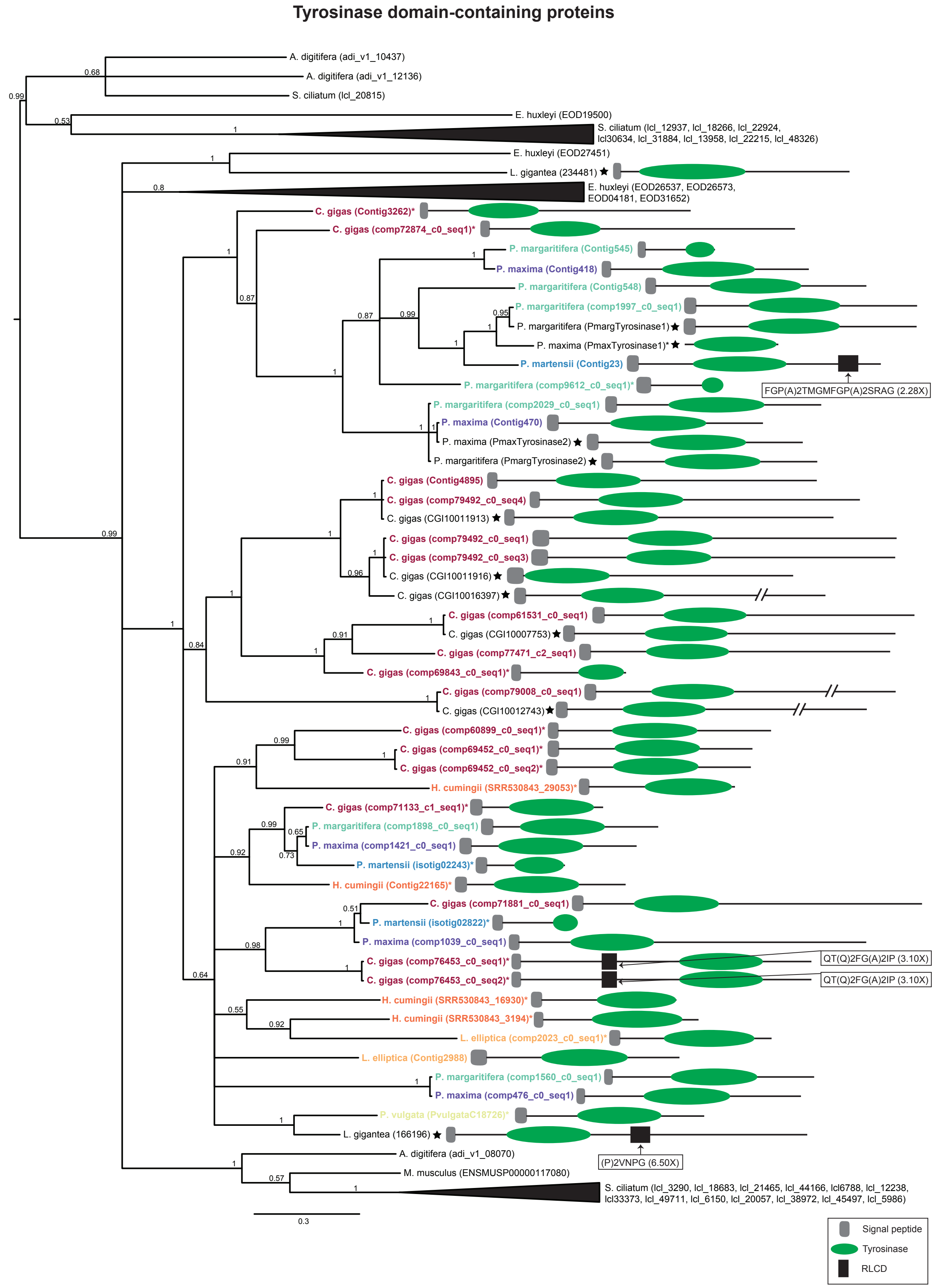

Tyrosinase domain-containing proteins
A. digitifera (adi_v1_10437)
A. digitifera (adi_v1_12136)
S. ciliatum (lcl_20815)
E. huxleyi (EOD19500)
S. ciliatum (lcl_12937, lcl_18266, lcl_22924, lcl30634, lcl_31884, lcl_13958, lcl_22215, lcl_48326)
E. huxleyi (EOD27451)
L. gigantea (234481) ★
E. huxley (EOD26537, EOD26573, EOD04181, EOD31652)
C. gigas (Contig3262)*
C. gigas (comp72874_c0_seq1)*
P. margaritifera (Contig545)
P. maxima (Contig418)
P. margaritifera (Contig548)
P. margaritifera (comp1997_c0_seq1)
P. margaritifera (PmargTyrosinase1) ★
P. maxima (PmaxTyrosinase1)* ★
P. martensii (Contig23)
FGP(A)2TMGMFGP(A)2SRAG (2.28X)
P. margaritifera (comp9612_c0_seq1)*
P. margaritifera (comp2029_c0_seq1)
P. maxima (Contig470)
P. maxima (PmaxTyrosinase2) ★
P. margaritifera (PmargTyrosinase2) ★
C. gigas (Contig4895)
C. gigas (comp79492_c0_seq4)
C. gigas (CGI10011913) ★
C. gigas (comp79492_c0_seq1)
C. gigas (comp79492_c0_seq3)
C. gigas (CGI10011916) ★
C. gigas (CGI10016397) ★
C. gigas (comp61531_c0_seq1)
C. gigas (CGI10007753) ★
C. gigas (comp77471_c2_seq1)
C. gigas (comp69843_c0_seq1)*
C. gigas (comp79008_c0_seq1)
C. gigas (CGI10012743) ★
C. gigas (comp60899_c0_seq1)*
C. gigas (comp69452_c0_seq1)*
C. gigas (comp69452_c0_seq2)*
H. cumingii (SRR530843_29053)*
C. gigas (comp71133_c1_seq1)*
P. margaritifera (comp1898_c0_seq1)
P. maxima (comp1421_c0_seq1)
P. martensii (isotig02243)*
H. cumingii (Contig22165)*
C. gigas (comp71881_c0_seq1)
P. martensii (isotig02822)*
P. maxima (comp1039_c0_seq1)
C. gigas (comp76453_c0_seq1)*
C. gigas (comp76453_c0_seq2)*
QT(Q)2FG(A)2IP (3.10X)
QT(Q)2FG(A)2IP (3.10X)
H. cumingii (SRR530843_16930)*
H. cumingii (SRR530843_3194)*
L. elliptica (comp2023_c0_seq1)*
L. elliptica (Contig2988)
P. margaritifera (comp1560_c0_seq1)
P. maxima (comp476_c0_seq1)
P. vulgata (PvulgataC18726)*
L. gigantea (166196) ★
(P)2VNPG (6.50X)
A. digitifera (adi_v1_08070)
M. musculus (ENSMUSP00000117080)
S. ciliatum (lcl_3290, lcl_18683, lcl_21465, lcl_44166, lcl6788, lcl_12238, lcl33373, lcl_49711, lcl_6150, lcl_20057, lcl_38972, lcl_45497, lcl_5986)
0.99
0.68
0.53
1
1
0.8
0.99
1
0.87
1
0.87
0.99
0.95
1
1
1
1
1
1
1
0.96
1
1
1
0.91
1
0.84
1
0.99
1
0.91
0.99
0.65
0.73
0.92
0.51
1
0.98
1
0.64
0.55
0.92
1
1
1
0.57
1
0.3
Signal peptide
Tyrosinase
RLCD

**Figure S9. Phylogenetic analysis of tyrosinase domain-containing proteins.** Bayesian phylogenetic tree was obtained under the LG+I+G substitution model. The tree is rooted

using the midpoint-rooted option. Statistical support for each node is indicated as posterior probabilities after 5,000,000 generations. Domain architectures of molluscan secreted proteins are shown on the right of each sequence, and domain nomenclature is shown in the rectangle. Proteins previously extracted from shells and/or known to be involved in shell formation are indicated by a black star. For each RLCD, the consensus repeat sequence is shown, followed by the copy number (in brackets).

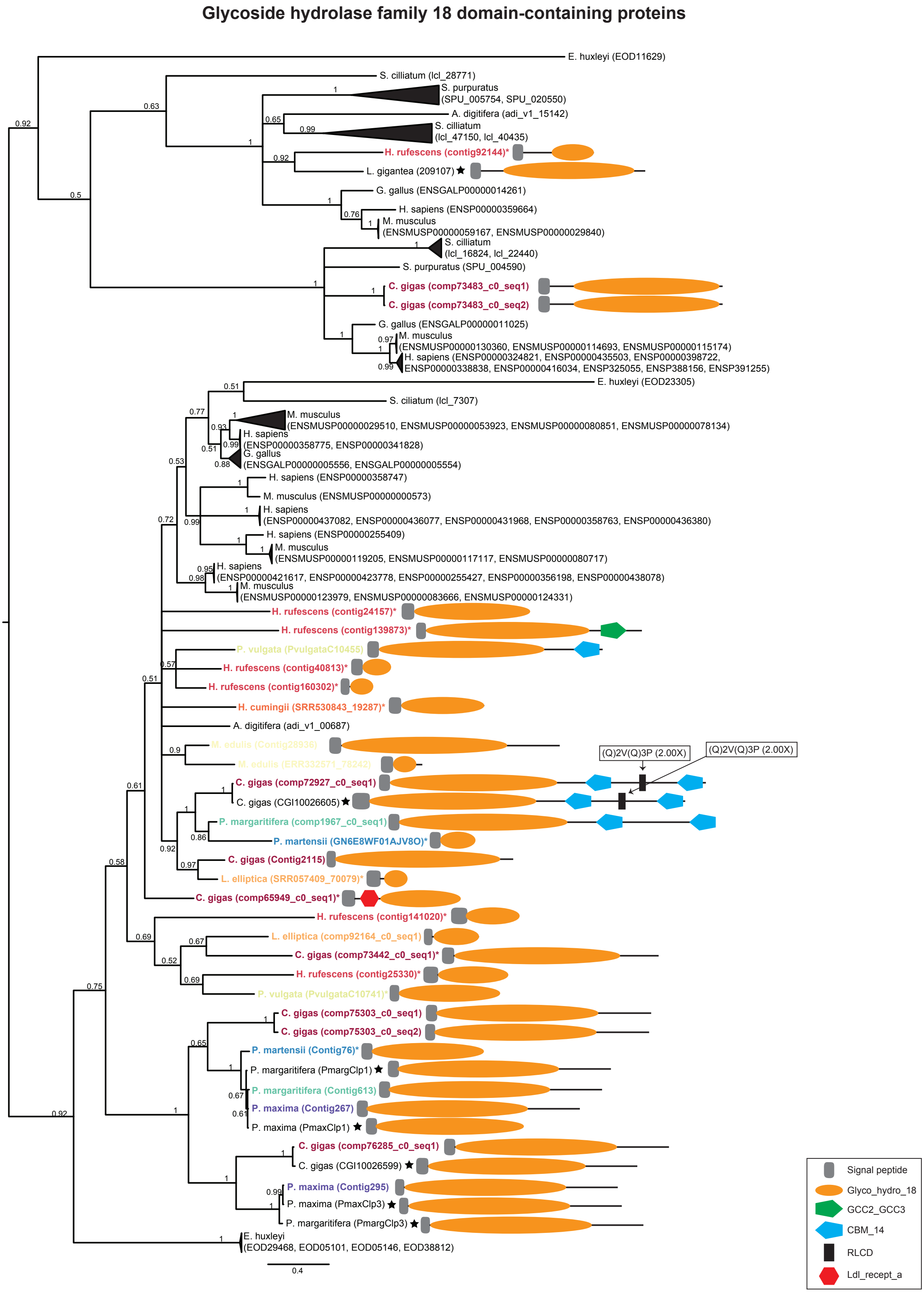

Glycoside hydrolase family 18 domain-containing proteins
E. huxleyi (EOD11629)
S. cilliatum (lcl_28771)
S. purpuratus (SPU_005754, SPU_020550)
A. digitifera (adi_v1_15142)
S. cilliatum (lcl_47150, lcl_40435)
H. rufescens (contig92144)*
L. gigantea (209107) ★
G. gallus (ENSGALP00000014261)
H. sapiens (ENSP00000359664)
M. musculus (ENSMUSP00000059167, ENSMUSP00000029840)
S. cilliatum (lcl_16824, lcl_22440)
S. purpuratus (SPU_004590)
C. gigas (comp73483_c0_seq1)
C. gigas (comp73483_c0_seq2)
G. gallus (ENSGALP00000011025)
M. musculus (ENSMUSP00000130360, ENSMUSP00000114693, ENSMUSP00000115174)
H. sapiens (ENSP00000324821, ENSP00000435503, ENSP00000398722, ENSP00000338838, ENSP00000416034, ENSP325055, ENSP388156, ENSP391255)
E. huxleyi (EOD23305)
S. ciliatum (lcl_7307)
M. musculus (ENSMUSP00000029510, ENSMUSP00000053923, ENSMUSP00000080851, ENSMUSP00000078134)
H. sapiens (ENSP00000358775, ENSP00000341828)
G. gallus (ENSGALP00000005556, ENSGALP00000005554)
H. sapiens (ENSP00000358747)
M. musculus (ENSMUSP00000000573)
H. sapiens (ENSP00000437082, ENSP00000436077, ENSP00000431968, ENSP00000358763, ENSP00000436380)
H. sapiens (ENSP00000255409)
M. musculus (ENSMUSP00000119205, ENSMUSP00000117117, ENSMUSP00000080717)
H. sapiens (ENSP00000421617, ENSP00000423778, ENSP00000255427, ENSP00000356198, ENSP00000438078)
M. musculus (ENSMUSP00000123979, ENSMUSP00000083666, ENSMUSP00000124331)
H. rufescens (contig24157)*
H. rufescens (contig139873)*
P. vulgata (PvulgataC10455)
H. rufescens (contig40813)*
H. rufescens (contig160302)*
H. cumingii (SRR530843_19287)*
A. digitifera (adi_v1_00687)
M. edulis (Contig28936)
M. edulis (ERR332571_78242)
(Q)2V(Q)3P (2.00X)
(Q)2V(Q)3P (2.00X)
C. gigas (comp72927_c0_seq1)
C. gigas (CGI10026605) ★
P. margaritifera (comp1967_c0_seq1)
P. martensii (GN6E8WF01AJV8O)*
C. gigas (Contig2115)
L. elliptica (SRR057409_70079)*
C. gigas (comp65949_c0_seq1)*
H. rufescens (contig141020)*
L. elliptica (comp92164_c0_seq1)
C. gigas (comp73442_c0_seq1)*
H. rufescens (contig25330)*
P. vulgata (PvulgataC10741)*
C. gigas (comp75303_c0_seq1)
C. gigas (comp75303_c0_seq2)
P. martensii (Contig76)*
P. margaritifera (PmargClp1) ★
P. margaritifera (Contig613)
P. maxima (Contig267)
P. maxima (PmaxClp1) ★
C. gigas (comp76285_c0_seq1)
C. gigas (CGI10026599) ★
P. maxima (Contig295)
P. maxima (PmaxClp3) ★
P. margaritifera (PmargClp3) ★
E. huxleyi (EOD29468, EOD05101, EOD05146, EOD38812)
0.4
Signal peptide
Glyco_hydro_18
GCC2_GCC3
CBM_14
RLCD
Ldl_recept_a

**Figure S10. Phylogenetic analysis of glycoside hydrolase family 18 domain-containing proteins.** Bayesian phylogenetic tree was obtained under the LG+G

substitution model. The tree is rooted using the midpoint-rooted option. Statistical support for each node is indicated as posterior probabilities after 4,000,000 generations. Domain architectures of molluscan secreted proteins are shown on the right of each sequence, and domain nomenclature is shown in the rectangle. Proteins previously extracted from shells and/or known to be involved in shell formation are indicated by a black star. For each RLCD, the consensus repeat sequence is shown, followed by the copy number (in brackets).

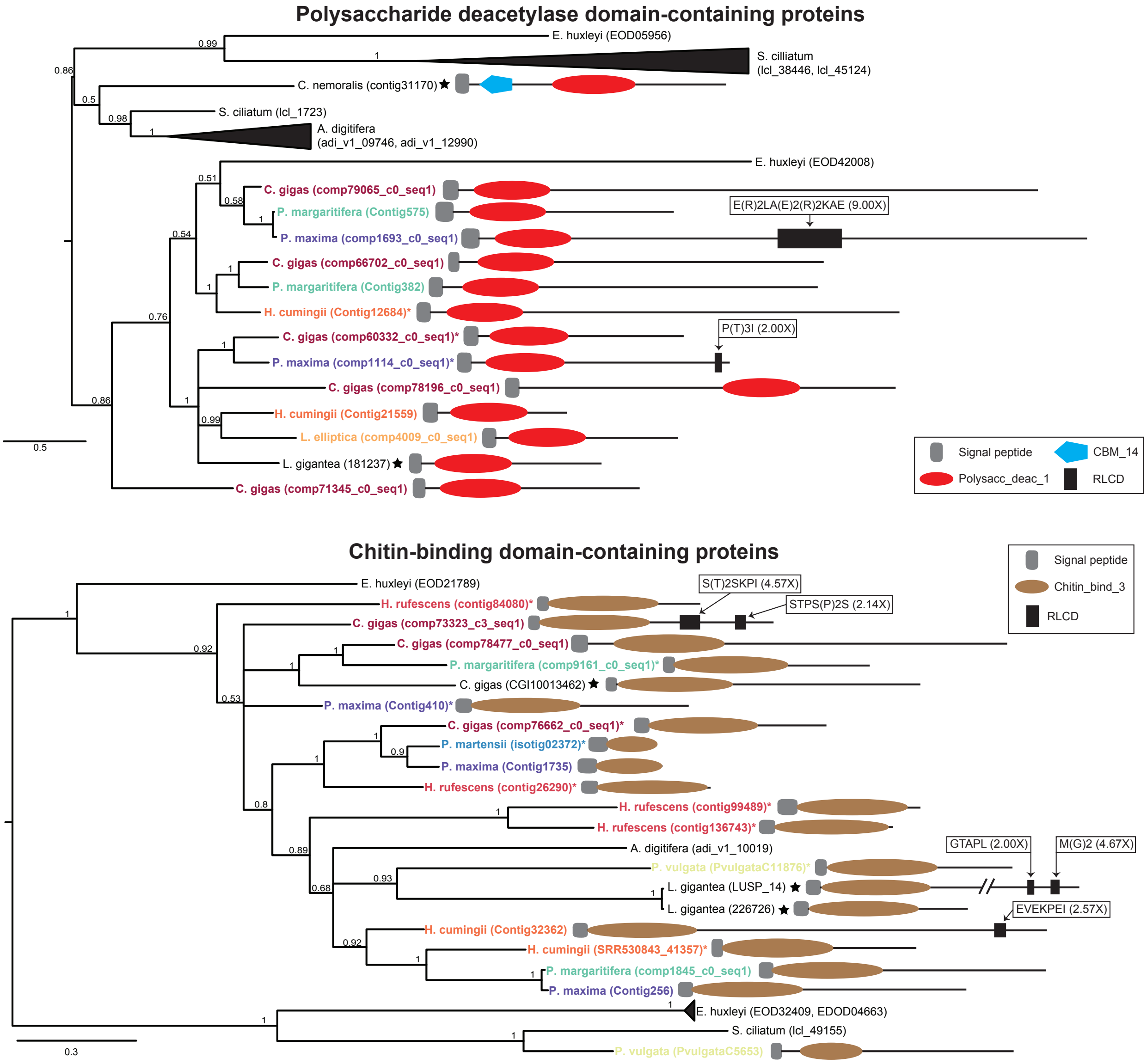

Polysaccharide deacetylase domain-containing proteins
E. huxleyi (EOD05956)
S. cilliatum (lcl_38446, lcl_45124)
C. nemoralis (contig31170) ★
S. ciliatum (lcl_1723)
A. digitifera (adi_v1_09746, adi_v1_12990)
E. huxleyi (EOD42008)
C. gigas (comp79065_c0_seq1)
P. margaritifera (Contig575)
E(R)2LA(E)2(R)2KAE (9.00X)
P. maxima (comp1693_c0_seq1)
C. gigas (comp66702_c0_seq1)
P. margaritifera (Contig382)
H. cumingii (Contig12684)*
C. gigas (comp60332_c0_seq1)*
P(T)3I (2.00X)
P. maxima (comp1114_c0_seq1)*
C. gigas (comp78196_c0_seq1)
H. cumingii (Contig21559)
L. elliptica (comp4009_c0_seq1)
L. gigantea (181237) ★
C. gigas (comp71345_c0_seq1)
0.99
1
0.86
0.5
0.98
1
0.51
0.58
1
0.54
1
1
0.76
1
1
0.86
0.99
0.5
Signal peptide
CBM_14
Polysacc_deac_1
RLCD
Chitin-binding domain-containing proteins
Signal peptide
Chitin_bind_3
RLCD
E. huxleyi (EOD21789)
S(T)2SKPI (4.57X)
H. rufescens (contig84080)*
STPS(P)2S (2.14X)
C. gigas (comp73323_c3_seq1)
C. gigas (comp78477_c0_seq1)
P. margaritifera (comp9161_c0_seq1)*
C. gigas (CGI10013462) ★
P. maxima (Contig410)*
C. gigas (comp76662_c0_seq1)*
P. martensii (isotig02372)*
P. maxima (Contig1735)
H. rufescens (contig26290)*
H. rufescens (contig99489)*
H. rufescens (contig136743)*
A. digitifera (adi_v1_10019)
GTAPL (2.00X)
M(G)2 (4.67X)
P. vulgata (PvulgataC11876)*
L. gigantea (LUSP_14) ★
L. gigantea (226726) ★
EVEKPEI (2.57X)
H. cumingii (Contig32362)
H. cumingii (SRR530843_41357)*
P. margaritifera (comp1845_c0_seq1)
P. maxima (Contig256)
E. huxleyi (EOD32409, EDOD04663)
S. ciliatum (lcl_49155)
P. vulgata (PvulgataC5653)
1
0.92
1
1
0.53
1
0.9
1
0.8
1
0.89
0.93
0.68
1
0.92
1
1
1
1
1
0.3

**Figure S11. Phylogenetic analyses of polysaccharide deacetylase domain-containing proteins and chitin-binding domain-containing proteins.** Bayesian phylogenetic trees were obtained under the LG+I+G and WAG+G substitution models, respectively. Both trees were rooted using the midpoint-rooted option. Statistical support for each node is indicated as posterior probabilities after 2,000,000 and 1,000,000 generations,

respectively. Domain architectures of molluscan secreted proteins are shown on the right of each sequence, and domain nomenclature is shown in the rectangles. Proteins previously extracted from shells and/or known to be involved in shell formation are indicated by a black star. For each RLCD, the consensus repeat sequence is shown, followed by the copy number (in brackets).

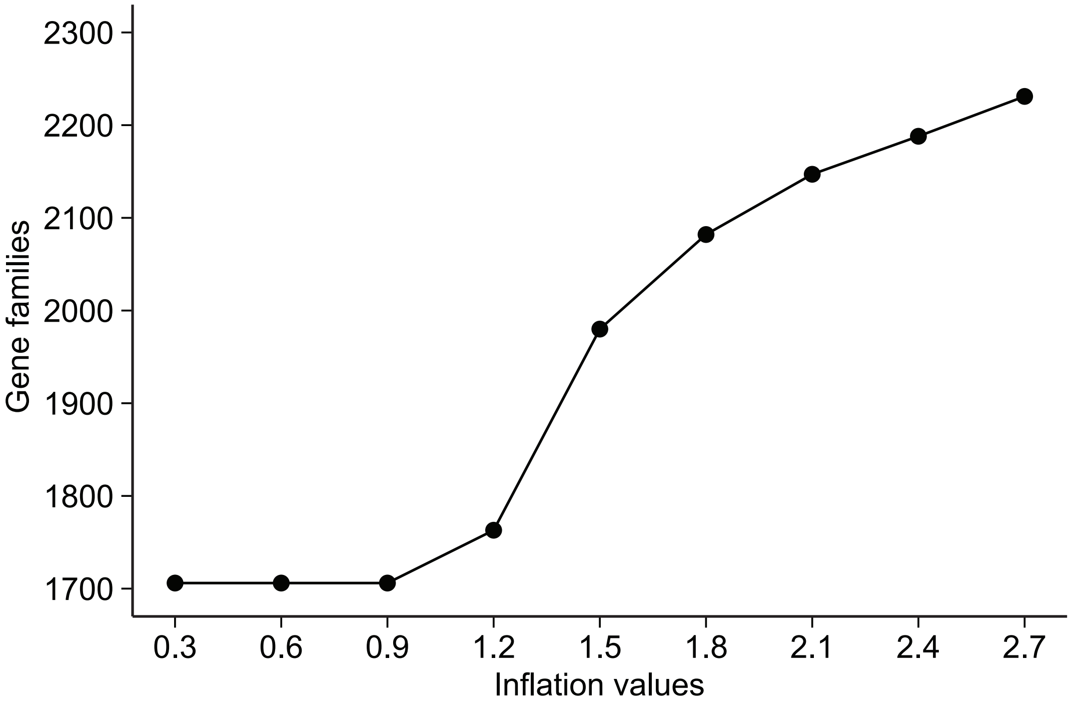


**Figure S12. Changes in the prediction of gene families using different MCL inflation values.** The effect of varying the MCL inflation value on the number of gene families are shown.