Qiu Jian-Wen (Orcid ID: 0000-0002-1541-9627)

# Egg perivitelline fluid proteome of a freshwater snail (Caenogastropoda): insight into the transition from aquatic to terrestrial egg deposition

**Short title:** Egg perivitelline fluid proteome of a freshwater snail

Jack C.H. Ip[1,2,*], Huawei Mu[3,*], Yanjie Zhang[2], Horacio Heras[4,5], Jian-Wen Qiu[1,2**]

[1] HKBU Institute of Research and Continuing Education, Shenzhen, China

[2]Department of Biology, Hong Kong Baptist University, Hong Kong, China

[3]School of Life Sciences, University of Science and Technology of China, Hefei, China

[4] Instituto de Investigaciones Bioquímicas de La Plata (INIBIOLP), Universidad Nacional de La Plata (UNLP)-CONICET CCT-La Plata, La Plata, Argentina

[5]Cátedra de Química Biológica, Facultad de Ciencias Naturales y Museo, UNLP, Argentina

* These authors contributed equally to this work

** Author for correspondence

Jian-Wen Qiu

Department of Biology, Hong Kong Baptist University, Hong Kong, China

Tel: (852) 3411-7055, Fax: (852) 3411-5995

E-mail: qiujw@hkbu.edu.hk

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/rcm.8605

**Abstract**

**Rationale:** Proteins from the egg perivitelline fluid (PVF) are assumed to play critical roles in embryonic development, but for many groups of animals their identities remain unknown. Identifying egg PVF proteins is a critical step towards understanding their functions including their roles in evolutionary transition in habitats.

**Methods:** We applied proteomic and transcriptomic analysis to analyse the PVF proteome of the eggs of *Pomacea diffusa*, an out-of-water ovipositing freshwater snail in the family Ampullariidae. The PVF proteins were separated with SDS-PAGE method, and proteomic analysis was conducted using LTQ Velos Ion Trap Mass Spectrometer coupled with liquid chromatography. Comparison of PVF proteomes and evolution analyses were performed between *P. diffusa* and other ampullariids.

**Results:** In total, 32 egg PVF proteins were identified from *P. diffusa*. They were categorized as PV1-like subunits, immune responsive proteins, protein degradation, signaling and binding, transcription and translation, metabolism, oxidation-reduction and unknown function. Interestingly, the proteome includes a calcium binding protein important in forming the hard eggshell that enabled the terrestrial transition. However, it does not include PV2, a neurotoxic protein that was assumed to be present in all *Pomacea* species.

**Conclusions:** The PVF proteome data from *P. diffusa* can help us better understand the roles of reproductive proteins played during the transition from underwater to terrestrial egg deposition. Moreover, they could be useful in comparative studies of the terrestrialization in several groups of animals that occurred independently during their evolution.

**Keywords:**

Apple snail, Egg perivitelline, Positive selection, Proteomics, Terrestrial eggs

## 1. INTRODUCTION

Apple snails (Ampullariidae) are freshwater gastropods widely distributed in the wet tropics and subtropics of the Americas, South Africa and Asia.[1] They have been proposed as model organisms to study evolution due to their ancient evolutionary history, wide distribution and species diversity and various egg depositing behaviours.[2] Ampullariidae is unique among freshwater snails in that it includes members that lay gelatinous egg masses under the water surface (genera *Marisa*, *Asolene*, *Felipponea* and *Lanistes*), and terrestrial egg masses above the water surface (genera *Pomacea* and *Pila*). The remarkable terrestrial egg deposition behaviour in some of these freshwater snails probably arose due to the pressure to escape from aquatic predators such as fish.[3] However, terrestrial eggs must be able to withstand physical stressors such as desiccation and UV radiation, as well as attack by terrestrial predators.[4]

Previous studies have found that the egg perivitelline fluid (PVF) that surrounds the developing embryos is critical to the early development of *Pomacea*. The PVF is mostly composed of polysaccharides and three major glycolipoprotein complexes (perivitellins). The perivitellins were classified according to mass and density, namely perivitelline-1 (PV1), perivitelline-2 (PV2) and the protein fraction PV3.[5,6] These perivitellins serve as stores of nutrients that fuel the embryonic development and are involved in embryo protection in some *Pomacea*.[6,7] Although previous proteomic studies named the major perivitellin protein PV1 as PcOvo for *P. canaliculata*,[8] PmPV1 for *P. maculata*[9] and PsSC for *P. scalaris*,[10] to avoid confusion of the major protein in PVF, we used "PV1" throughout this study. Among two major perivitelline groups, PV1 plays several critical roles: photoprotection, antioxidant, and nutrient provision in *P. canaliculata* and *P. maculate*.[4,8,11,12], and immune response in *P. scalaris*.[13] PV2, which constitutes around 7.5% of total protein in PVF,[5] is a protein complex composed of two subunits: a membrane attack complex/perforin (MACPF) protein as the

toxic subunit, and a tachylectin-like protein as the delivery subunit, which are linked with a disulfide bridge.[14] Laboratory experiments have shown that PV2 is neurotoxic to mice in *P. canaliculata* and probably is also a neurotoxin in *P. maculata*.[7,14,15] However, the neurotoxicity was not observed in the PVF of *P. scalaris*,[7,10] suggesting that PV2-active toxins may not be present in all species of *Pomacea*. Since there are multiple copies of MACPF and tachylectin genes in *Pomacea*,[16] to avoid confusion we named the PV2 associated subunits as PV2-MACPF and PV2-tachylectin, and others as MACPF-like and tachylectin-like, respectively.

With the advance in high throughput LC-MS/MS, proteomic analyses have been conducted for two aerial egg depositors *P. canaliculate*[8] and *P. maculate,*[17] which showed that both PV1 and PV2 contain several subunits. A similar study has also been conducted in *Marisa cornuarietis*, an underwater egg depositor, which revealed 36 PVF proteins.[18] Comparing the PVF proteins of *M. cornuarietis* and the two species of *Pomacea* showed that the three species share several major PV1 perivitellins, which in *P. canaliculata* are known to nourish the embryos. However, the PVF of *M. cornuarietis* does not contain two proteins that were presumably important during the evolutionary transition from underwater to aerial egg deposition in *P. canaliculata and P. maculata*: (1) a calcium-binding protein (CaBP) which is involved in the formation of the calcareous eggshell that protects the eggs and prevents desiccation; and (2), PV2, a neurotoxin as a novel defence mechanism against potential terrestrial predators.

Phylogenetic analyses have shown that *P. canaliculata* and *P. maculata* are very closely related species located in a terminal clade among a large clade that includes most other species of *Pomacea*, and this large clade is the sister clade that includes *M. cornuarietis* and species of two other genera (*Asolene* and *Felipponea*).[2,19] A recent study based on 1,357

single-copy orthologous genes estimated that *M. cornuarietis* and the two species of *Pomacea* shared a common recent ancestor ~28.3 million years ago (Ma), whereas the divergence between *P. canaliculata* and *P. maculata* occurred much more recently at ~ 3.0 Ma.[16] Because the terminal positions of the two *Pomacea* species, and the long divergence time between the genera *Pomacea* and *Marisa*, the PVF proteins of *P. canaliculata* and *P. maculata* may not represent the diversity of PVF proteins of the genus. This inspired us to analyse the PVF proteins of *P. diffusa* (Figure 1A), a species phylogenetically located in a clade that is sister to the clade that includes *P. canaliculata* and *P. maculata*. Our aim is to provide a proteomic dataset for the PVF of *P. diffusa*, and the comparison and phylogenetic analysis against the three published PVF proteomes contribute to the understanding of the molecular mechanisms underlying the evolutionary transition from underwater to aerial egg deposition in Ampullariidae and terrestrialization of aquatic gastropods.

## 2. EXPERIMENTAL

### 2.1 PVF protein sampling and extraction

Adults of *P. diffusa* were purchased from an aquarium shop in Hong Kong. They were reared in aerated tap water at 26 ± 1 °C and a photoperiod of 14 h light/ 10 h dark. The snails were fed with a mixed diet of lettuce, carrot and fish food daily. Two freshly deposited egg clutches (within 12 h of deposition) were collected and processed separately for proteomic analysis as two biological replicates. For each replicate, PVF was manually extracted from approximately 50 eggs using a syringe and mixed with 8 M urea.

### 2.2 SDS-PAGE and LC-MS/MS

Proteomic analysis was performed by integrating SDS-PAGE and LC-MS/MS.[17,20] In brief, each sample was centrifuged for 10 min at 12,000 *g* in a centrifuge that was set at 4 °C, the supernatant was collected, purified using the methanol/chloroform method,[21] and quantified

using a RC-DC kit (Bio-Rad, California, USA). A subsample of 100 µg of purified PVF proteins was separated using SDS-PAGE, stained with Coomassie Brilliant Blue, destained with 1% acetic acid, then cut into eight slices based on the intensity and molecular weight (Figure 1B). The gel slices were digested using sequencing grade trypsin (Promega, Madison, USA), desalted using Sep-Pak C18 cartridges (Waters, Milford, USA), dried using a SpeedVac Concentrator (Eppendorf, Hamburg, Germany), and reconstituted using 0.1% formic acid. Then the samples were analyzed using a LTQ Velos Ion Trap Mass Spectrometer (Thermo Scientific, Bremen, USA) with an 80-min LC gradient: 98% solution A (0.1% formic acid in MilliQ) (2 min), 2-45% solution B (0.1% formic acid in acetonitrile) (59 min, linear changing gradient), 80% solution B (10 min), and 98% solution A (9 min). The MS/MS was scanned at a mass range of 300 to 2000 *m*/*z*. The five most abundant multiple-charged ions with at least 2000 counts were selected for collision-induced dissociation using a normalized collision energy of 35%, activation Q of 0.25, isolation width of 2.0, and activation time of 10ms. Dynamic exclusion was applied, with a repeat count of 2, exclusion duration of 60 s and exclusion mass width of 1.5 *m*/*z*.

**2.3 PVF proteins identification**

To identify the proteins, a database that contained 27,054 protein sequences (target) with average length of 358.6 amino acids and their reversed sequences (decoy) was prepared based on published *P. diffusa* transcriptome.[22] The MS/MS files were converted from raw files to Mascot generic files by Proteome Discoverer 1.3.0.339 (Thermo Fisher Scientific, San Jose, USA), and searched against the transcriptome database using Mascot v2.3.2 (Matrix Sciences, London, UK) under the settings of 0.8 Da for peptide tolerance and MS/MS tolerance, fixed modification for carbamidomethyl, variable modification for deamidation and oxidation, and one maximum missed cleavage for trypsin. The matched peptides were filtered to retain those with ions score ≥ 42 (above the 95% confidence level), peptide length

≥ 7, false discovery rate < 0.01, detected in both biological replicates, with at least three unique peptides, and gene expression level at albumen gland > 5 TPM (transcripts per million). Protein abundance was estimated using the Exponentially Modified Protein Abundance Index (emPAI),[23] and signal peptides were predicted using SignalP4.1.[24]

The identified PVF proteins were annotated using BLASTp against NCBI non-reductant (nr) database with an *E*-value threshold of $10^{-5}$ and InterProScan 5[25] under default settings for Gene Ontology (GO). Proteins without GO annotation were further searched against COFACTOR,[26] which derives protein function based on structural comparisons and protein-protein networks, using the I-TASSER server[27] with a GO-score threshold of 0.5.[28] In addition, BLASTp search with *E*-value threshold of $10^{-5}$ was conducted to compare the PVF composition between *P. diffusa* and three other ampullariids.[8,17,18] The results were visualized as a Circos graph made using Circoletto.[29] Putative orthogroups among *P. diffusa*, *P. canaliculata*, *P. maculata* and *M. cornuarietis* were identified using OrthoMCL[30] under default settings (E-value threshold of $10^{-5}$ and sequence identity ≥ 50%).

**2.4 Phylogenetic and evolution analyses**

To investigate the evolution of CaBP and PV2 subunits, we performed BLASTp search for their homologs from all eight sequenced ampullariids, including four with PVF proteomes (i.e., *P. diffusa*, *P. canaliculata*, *P. maculata* and *M. cornuarietis*), and *Lanistes nyassanus*, *Pila ampullacea*, *Asolene platae*, and *P. scalaris*[16,18,22] with an *E*-value threshold of $10^{-5}$, and aligned length and sequence identity ≥ 30%. Phylogenetic trees were constructed from the MUSCLE aligned and manually trimmed sequences by maximum-likelihood analysis with 1000 ultrafast bootstraps using IQ-TREE.[31] The substitution models of WAG+R3 for CaBP, WAG+I+G4 for MACPF-like and WAG+G4 for tachylectin-like genes was selected by ModelFinder[32] implemented in IQ-TREE. The RNAseq data from Ip et al.[22] were mapped to

the eight transcriptomes for determination of gene expression level expressed as Transcripts Per Kilobase Million (TPM) using Salmon 0.9.1.[33]

Selective pressure was determined by the ratio (ω) of the nonsynonymous substitutions rate (dN) to the synonymous substitutions rate (dS), with purifying, neutral, or positive selection indicated by ω < 1, = 1, or > 1, respectively. Substitution rate of each branch was estimated using the codeml in PAML 4.9h[34] with the free-ratio model (model = 1). For branches with high expression in albumen gland, a likelihood ratio test (LRT) was further used to compare an alternative model (Model = 2, NSsites = 2, fix_ω= 0, ω=1), which allows sites to be under positive selection on the foreground branch, and a null model (Model = 2, NSsites = 2, fix_ω= 1, ω=1), in which sites may evolve neutrally and under purifying selection. LRT was estimated as twice the difference of log-likelihood values between nested models and compared with Chi-squared test with a degree of freedom of 1. The Bayes Empirical Bayes (BEB) approach was employed to identify sites under positive selection with a posterior probability (PP) larger than 0.95.

## 3. RESULTS AND DISCUSSION

The MS/MS analysis resulted in a total of 3,058 peptides. After filtering with the criteria described earlier, a total of 32 proteins were identified from the two biological replicates of *P. diffusa* PVF based on 342 unique peptides (Table 1; Supporting information Table S1 for detail of peptide matched). Most of the PVF proteins (31 out of 32) were successfully annotated with NCBI's nr database, among them 25 were further assigned with GO categories. However, one of the proteins was not successfully annotated and thus considered as a novel protein. The identified proteins were classified into seven functional categories (number of proteins in parentheses) based on NCBI's nr and GO annotation with manual correction: PV1 subunits (6), immune responsive proteins (13), proteins degradation (5),

signaling and binding (4), transcription and translation (1), metabolism (1), oxidation reduction (1) (Figure 1C). Among these proteins, the six PV1 subunits were most abundant with relative protein abundance ranging from 0.5% to 27.4%, and a total abundance in egg PVF of 75.2%. *P. diffusa* also invested heavily in protecting embryos against pathogenic infection, with 13 immune related proteins accounting for 17.0% of the total PVF protein abundance. Other groups of PVF proteins were less abundant, with each group contributing to no more than 5% of the total protein abundance (Figure 1D).

To further understand the putative functions of the PVF proteins that were not successfully annotated using NCBI's nr database (i.e., six PV1 subunits and one novel protein), protein sequences, after removing signal peptides, were analyzed with the I-TASSER/ COFACTOR pipeline. Four PV1 (i.e., PV1-A1, PV1-A3, PV1-A4 and PV1-B1) and the novel protein were predicted with one to three GO terms using a reliable prediction GO-score (> 0.5) (Supporting information Table S2).[28] The results suggest that the PV1 subunits may function in carbohydrate metabolism with the molecular function of "ATP binding" (GO:0005524), the biological process of "aminoglycan catabolic process" (GO:0006026), "alditol metabolic process" (GO:0019400) and "primary metabolic process" (GO:0044238), suggesting a putative role of PV1 in carbohydrate metabolism (which are major PVF components used for embryogenesis).[6] The novel protein (Pdi51889_c0_g1) and its homologs were detected in the PVF proteins of the other three ampullariids, but none of them have functional annotation (Figure 2A). COFACTOR analysis predicted the molecular functions of the novel protein as "adenyl ribonucleotide binding" (GO:0032559), "purine ribonucleoside triphosphate binding" (GO:0035639), and "pyrophosphatase activity" (GO:0016462), which indicate that this protein may be involved in phosphate binding and energy metabolism during the embryonic development. These *in silico* predictions of PVF protein functions provide clues for follow-up experimental validation of the functions of

these uncharacterized proteins.[35]

BLASTp search revealed high similarities among the PVF proteomes of *P. diffusa*, *P. canaliculata*, *P. maculata* and *M. cornuarietis* (Figure 2A). Among the 32 PVF proteins in *P. diffusa*, 23 (71.9%) had at least one BLASTp hit with the corresponding PVF proteins of the other ampullariids. Moreover, these homologous proteins accounted for most (95%) of the total PVF protein abundance in *P. diffusa*. Among the homologous proteins, PV1 subunits were the major PVF proteins in all of the four species, but their abundances in *P. diffusa* (75.2%) and *M. cornuarietis* (73.2%) were substantially lower than those in *P. canaliculata* (92.2%) and *P. maculata* (87.0%). Conversely, *P. diffusa* and *M. cornuarietis* invested more heavily on immune-related proteins than *P. canaliculata* and *P. maculata*, as determined by both the number of proteins and their total abundance (about 13/17 proteins with around 17% of the total PVF protein abundance in the former against 8/9 proteins and less than 3% in the later).[18]

The PVF proteomes of the four ampullariids have a total of 139 PVF proteins, among them 109 are clustered in 31 orthogroups (i.e., genes that are present in at least two species or two paralogs within one species) and 30 singletons (genes without homolog and are present in only one species) (Figure 2B). Among the orthogroups, 13 were found in all of the four ampullariids, including five PV1 subunits, five immune responsive proteins (C1q domain-containing protein, CD109 antigen, deleted in malignant brain tumors 1, tachylectin-related protein and soma ferritin), and three other proteins (superoxide dismutase 1, apoptosis-inducing factor and novel protein). Interestingly, two orthogroups, CaBP that is involved in forming the egg shell,[16] and Kunitz/Bovine pancreatic trypsin inhibitor that inhibits the functioning of multiple digestive proteases,[12,36] were only detected in the three species of *Pomacea*. The presence of multiple immunoproteins in the PVF of the four species may

indicate their role in defense against pathogens, as has been suggested for the eggs of the pulmonate snail *Biomphalaria glabrata*.[37] The multiple homologs of immune proteins were formed by gene duplications and positive selection.[18]

The acquisition of a calcareous eggshell is one of the convergent evolutions associated with terrestrial colonization in gastropods, which provides protection of the developing embryo from physical stresses, such as desiccation and UV radiation.[4,38] Nevertheless, little is known about the involvement of proteins in forming such hard shells. In apple snails, previous studies reported the storage of calcium in albumen gland of *Pomacea* species and depositing the intracellular calcium in capsule gland (part of albumen gland) for egg formation.[39,40] The high expression of CaBP in albumen gland and its presence in both PVF and aerial eggshell of *Pomacea* indicate the crucial role of this protein in the eggshell formation.[16,18] To gain further insight into the evolution of *Pomacea* CaBP, we identified 29 homologs in the eight ampullariids. The maximum-likelihood analysis revealed two major clades (Group A and Group B; Fig. 3). We focused the analysis on Group B homologs, which comprised the subclade B of seven genes with high expression in albumen gland and their coding proteins detected in PVF. Interestingly, subclade A, containing seven genes from each species with low expression in albumen gland (TPM < 10), was sister to a clade with albumen gland specific subclade B. The positive selection analysis showed that eight branches of Group B with $\omega > 1$ (LRT: $2\Delta L=79.02$, degree of freedom=27, $p < 0.001$), indicating positive selection has acted on these genes (Fig. S1A). Although the branch-site test did not identify amino acid site with positive selection on subclade B (Table S4), the sequence alignment showed a conserved D-K/Q-DG-D/N-N calcium binding motif in subclade B,[41] and D-X-N-X-DD motif in subclade A (Fig. S2A), indicating the conservation of the putative calcium binding function in subclade B after gene duplication. The unique presence of CaBP in *Pomacea* is consistent with role of the shell in defense of terrestrial eggs

against physical damage and desiccation.[16] Similar studies in other lineages of terrestrial egg laying gastropods may unveil the mechanisms of convergent evolution of calcareous eggshell formation in Mollusca.

Remarkably, the two PV2 subunits of *P. canaliculata*, which exert neurotoxicity to mice,[14,15] were not detected in *P. diffusa* PVF (Table 1). The acquisition of neurotoxic PV2 in the terrestrial eggs of *Pomacea* may be an important biochemical defense mechanism against terrestrial predators.[15] A genome analysis of four ampullariids also indicated that the duplication and neofunctionalization contributed to the formation the neurotoxic PV2-MACPF and PV2-tachylectin in *P. canaliculata* and *P. maculata*: (1) they show tissue-specific expression in albumen gland; (2) they form clades in phylogenetic trees; and (3) their coding protein subunits are present in PVF.[16] To detect the orthologs of PV2-MACPF and PV2-tachylectin in *P. diffusa*, we integrated phylogenetic, transcriptomic and proteomic evidences. The phylogenetic analysis of two subunits were constructed using 29 MACPF homologs and 69 tachylectin homologs from the eight ampullariids. The tree topologies of MACPF-like and tachylectin-like genes agree well with that of Sun et al.[16], indicating two clustered of genes (Group A and Group B; Fig. 4A and 4B). In both genes, the extensive gene duplication events have occurred in Group B, perhaps specific the *P. canaliculata* and *P. maculata*, and possibly associated in the formation and neofunctionalization of toxic PV2-MACPF and PV2-tachylectin paralogs, with high albumen gland expression and secreted into PVF.[16] However, in Group B, none of the MACPF-like and tachylectin-like genes from *P. diffusa* were clustered with PV2 subunits and expressed in albumen gland. Therefore, the PV2-MACPF and PV2-tacykectin should be formed after the split between *P. diffusa* and *P. canaliculata*/*P. maculata*, which can explain the absence of PV2 proteins in *P. diffusa* PVF.

Surprisingly, the positive selection analysis in Group B identified 4 branches in MACPF-like (LRT: $2\Delta L=104.26$, degree of freedom=33, $p < 0.001$) and 14 branches in tachylectin-like (LRT: $2\Delta L=90.88$, degree of freedom=47, $p < 0.001$) with $\omega > 1$, indicating wide spread positive selection on these genes (Figs. S1B and S1C).The branch-site test further suggested the PV2-MACPF branch (Pca_1306_1.39 and Pma_3499_0.31) under episodic bursts of selection after putative gene duplication with six amino acid sites under positive selection ($PP > 0.95$; Table S4), whilst 3 sites are located in the MACPF domain predicted with PROSITE (Fig. S3).[42] Studies of venom proteins showed that positive selection sites might be adaptive to predation and defense because most of the adaptive sites are located on the functional domains.[43,44] Therefore, positive selection might have been involved in the divergent evolution of MACPF-like and tachylectin-like genes among ampullariids, and the PV2-MACPF in *Pomacea*, driving the neofunctionalization of toxic function in the PV2 complex.

In summary, our study has identified 32 proteins in the PVF of *P. diffusa* eggs. Comparison with published data showed that several proteins in the PV1 complex dominate the PVF proteomes of all ampullariids. Nevertheless, *P. diffusa* and *M. cornuarietis* invest more on immune proteins and protease inhibitors. The three *Pomacea* species have a unique CaBP required for the formation of the calcareous shell of terrestrial eggs, and a Kunitz/Bovine pancreatic trypsin inhibitor that, at least in *P. canaliculata* and *P. maculata* inhibit multiple digestive proteases of mice.[7,36] In addition, duplication and positive selection of the CaBP and PV2 subunits were likely responsible for the aerial egg deposition and the evolution of PV2 toxic protein as an anti-predation mechanism in *P. canaliculata* and *P. maculata*.[16] Overall, the results of this study will facilitate further research on the evolution of PVF proteins during the transition from underwater to terrestrial egg deposition in Ampullariidae, in particular the Old World lineage (e.g., the underwater ovipositor *Lanistes*

and terrestrial ovipositor *Pila*). The data can also be used in comparative genomic studies aiming to understand the mechanisms of terrestralization in Gastropoda.[45-48]

## DATA ACCESSIBILITY

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD012269.

## CONFLICT OF INTEREST

The authors declare no competing financial interest.

## ACKNOWLEDGEMENTS

This study was supported by Shenzhen Science and Technology Innovation Committee (JCYJ20170307161326613) and General Research Fund of Hong Kong (HKBU 12301415).

## REFERENCES

1. Hayes KA, Burks RL, Castro-Vazquez A, et al. Insights from an integrated view of the biology of apple snails (Caenogastropoda: Ampullariidae). *Malacologia* 2015;58:245–302.
2. Hayes KA, Cowie RH, Jørgensen A, Schultheiß R, Albrecht C, Thiengo SC. Molluscan models in evolutionary biology: apple snails (Gastropoda: Ampullariidae) as a system for addressing fundamental questions. *Am. Malacol. Bull*. 2009;27:47–58.
3. Turner, R.L. Effects of submergence on embryonic survival and developmental rate of the Florida apple snail, *Pomacea paludosa*: implications for egg predation and marsh management. *Fla. Sci*. 1998;118-129.
4. Heras H, Dreon M, Ituarte S, Pollero R. Egg carotenoproteins in neotropical Ampullariidae (Gastropoda: Arquitaenioglossa). *Comp. Biochem. Physiol. C Toxicol.*

*Pharmacol*. 2007;146:158–167.

5. Garin CF, Heras H, Pollero RJ. Lipoproteins of the egg perivitelline fluid of *Pomacea canaliculata* snails (Mollusca: Gastropoda). *J. Exp. Zool*. 1996;276:307–314.
6. Heras H, Garin CF, Pollero RJ. Biochemical composition and energy sources during embryo development and in early juveniles of the snail *Pomacea canaliculata* (Mollusca: Gastropoda). *J. Exp. Zool.* 1998;280:375–383.
7. Giglio ML, Ituarte S, Pasquevich MY, Heras H. The eggs of the apple snail *Pomacea maculata* are defended by indigestible polysaccharides and toxic proteins. *Can. J. Zool*. 2016;94:777–785.
8. Sun J, Zhang H, Wang H, et al. First proteome of the egg perivitelline fluid of a freshwater gastropod with aerial oviposition. *J. Proteome Res.* 2012;11:4240–4248.
9. Pasquevich M, Dreon M, Heras H. The major egg reserve protein from the invasive apple snail *Pomacea maculata* is a complex carotenoprotein related to those of *Pomacea canaliculata* and *Pomacea scalaris*. *Comp. Biochem. Physiol. B Biochem. Mol. Biol*. 2014;169:63–71.
10. Ituarte S, Dreon M, Ceolín M, Heras H. Isolation and characterization of a novel perivitellin from the eggs of *Pomacea scalaris* (Mollusca, Ampullariidae). *Mol. Reprd. Dev.* 2008;75:1441-1448.
11. Dreon M.S, Schinella G, Heras H, Pollero RJ. Antioxidant defense system in the apple snail eggs, the role of ovorubin. *Arch. Biochem. Biophys*. 2004;422:1-8.
12. Dreon MS, Ituarte S, Heras H. The role of the proteinase inhibitor ovorubin in apple snail eggs resembles plant embryo defense against predation. *PLoS One* 2010;5:e15059.
13. Ituarte S, Brola TR, Fernández PE, Mu H, Qiu JW, Heras H, Dreon MS. A lectin of a non-invasive apple snail as an egg defense against predation alters the rat gut morphophysiology. *PloS One*. 2018;13: e0198361.
14. Dreon MS, Frassa MV, Ceolín M, et al. Novel animal defenses against predation: a snail

egg neurotoxin combining lectin and pore-forming chains that resembles plant defense and bacteria attack toxins. *PLoS One* 2013;8:e63782.

15. Heras H, Frassa MV, Fernandez PE, Galosi CM, Gimeno EJ, Dreon MS. First egg protein with a neurotoxic effect on mice. *Toxicon* 2008;52:481–488.
16. Sun J, Mu H, Ip, JCH, et al. Signatures of divergence, invasiveness, and terrestrialization revealed by four apple snail genomes. *Mol. Biol. Evol.* 2019;36:1507-1520.
17. Mu H, Sun J, Heras H, Chu KH, Qiu JW. An integrated proteomic and transcriptomic analysis of perivitelline fluid proteins in a freshwater gastropod laying aerial eggs. *J. Proteomics* 2017;155:22–30.
18. Ip JCH, Mu H, Zhang Y, et al. Understanding the transition from water to land: Insights from multi-omic analyses of the perivitelline fluid of apple snail eggs. *J. Proteomics*. 2019;194:79-88.
19. Yang Q, Liu S, He C, Cowie RH, Yu X, Hayes KA. Invisible apple snail invasions: importance of continued vigilance and rigorous taxonomic assessments. *Pest Manag. Sci.* 2018;75:1277-1286.
20. Zhang Y, Mu H, Lau SC, Zhang Z, Qiu JW. Sperm proteome of *Mytilus galloprovincialis*: Insights into the evolution of fertilization proteins in marine mussels. *Proteomics* 2015;15:4175–4179.
21. Friedman DB. Quantitative proteomics for two-dimensional gels using difference gel electrophoresis. *Methods Mol. Biol.* 2007:367;219–239
22. Ip JCH, Mu H, Chen Q, et al. AmpuBase: a transcriptome database for eight species of apple snails (Gastropoda: Ampullariidae). *BMC Genomics* 2018;19:179.
23. Ishihama Y, Oda Y, Tabata T, et al. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell. Proteomics* 2005;4:1265–1272.
24. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal

peptides from transmembrane regions. *Nat. Methods*. 2011;8:785–786.

25. Jones P, Binns D, Chang HY, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 2014;30:1236-1240.
26. Zhang C, Freddolino PL, Zhang Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic Acids Res*. 2017;45:W291-W299.
27. Yang J, Zhang Y. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res.* 2015;43:W174-W181
28. Roy A, Xu D, Poisson J, Zhang Y. A protocol for computer-based protein structure and function prediction. *J. Vis. Exp.* 2011;*JoVE*:e3259.
29. Darzentas N. Circoletto: visualizing sequence similarity with Circos. *Bioinformatics* 2010;26:2620–2621.
30. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 2003;13:2178-2189.
31. 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.* 2014;32:268-274.
32. Kalyaanamoorthy S, Minh BQ, Wong TK, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods*. 2017;14: 587.
33. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 2017;14:417–419.
34. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol*. 2007;24:1586–1591.
35. Zhang C, Wei X, Omenn GS, Zhang,Y. Structure and Protein Interaction-based Gene Ontology Annotations Reveal Likely Functions of Uncharacterized Proteins on Human Chromosome 17. *J. Proteome Res*. 2018;17: 4186-4196.

36. Ituarte S, Brola TR, Dreon MS, Sun J, Qiu JW, Heras H. Non-digestible proteins and protease inhibitors: Implications for defense of the colored eggs of freshwater apple snails. *Can. J. Zool.* 2019;97:558-566.

37. Hathaway JJ, Adema C. M, Stout BA, Mobarak CD, Loker ES. Identification of protein components of egg masses indicates parental investment in immunoprotection of offspring by *Biomphalaria glabrata* (Gastropoda, Mollusca). *Dev. Comp. Immunol*. 2010;34:425–435.

38. Bigatti G, Giraud-Billoud M, Vega IA, Penchaszadeh PE, Castro-Vazquez A. The calcareous egg capsule of the Patagonian neogastropod *Odontocymbiola magellanica*: morphology, secretion and mineralogy. *J. Molluscan Stud*. 2010;76:279-288.

39. Catalán N, Fernández S, Winik B. Oviductal structure and provision of egg envelops in the apple snail *Pomacea canaliculata* (Gastropoda, Prosobranchia, Ampullariidae). *Biocell*. 2002;26:91-100.

40. Meenakshi V, Blackwelder P, Watabe N. Studies on the formation of calcified egg-capsules of ampullarid snails. *Calcif. Tissue Res.* 1974;16:283-291.

41. Rigden DJ, Galperin MY. The DxDxDG motif for calcium binding: multiple structural contexts and implications for evolution. *J Mol. Biol.* 2004;343:971-984.

42. Sigrist CJ, Cerutti L, De Castro E, et al. PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res*. 2009;38:D161-D166.

43. Sunagar K, Johnson WE, O'Brien SJ, Vasconcelos V, Antunes A. Evolution of CRISPs associated with toxicoferan-reptilian venom and mammalian reproduction. *Mol. Biol. Evol.* 2012;29:1807-1822.

44. Jouiaei M, Sunagar K, Gross AF, et al. Evolution of an ancient venom: recognition of a novel family of cnidarian toxins and the common evolutionary origin of sodium and potassium neurotoxins in sea anemone. *Mol. Biol. Evol*. 2015;32:1598-1610.

45. Kameda Y, Kato M. Terrestrial invasion of pomatiopsid gastropods in the heavy-snow

region of the Japanese Archipelago. *BMC Evol. Biol*. 2011;11:118.

46. Kano Y, Chiba S, Kase T. Major adaptive radiation in neritopsine gastropods estimated from 28S rRNA sequences and fossil records. *Pro R Soc Lond. B: Biol. Sci*. 2002;269:2457-2465.

47. Dayrat B, Conrad M, Balayan S, et al. Phylogenetic relationships and evolution of pulmonate gastropods (Mollusca): new insights from increased taxon sampling. *Mol. Phylogenet. Evol*. 2011;59:425-437.

48. Vermeij GJ, Dudley R. Why are there so few evolutionary transitions between aquatic and terrestrial ecosystems? *Biol. J. Linn. Soc*. 2000;70:541-554.

**TABLE 1.** List of *Pomacea diffusa* PVF proteins showing unigene ID, protein name, number of peptides and average percentage of the two replicates. Transcriptomic expression levels of the corresponding genes in albumen gland (AG) and other tissues (OT) are also shown as transcripts per million (TPM).

| Unigene ID | Description[a] | Number of peptides | Average %[b] | AG_TPM[c] | OT_TPM[c] |
|---|---|---|---|---|---|
| **Perivitellin ovorubin (PV1)** | | | | | |
| Pdi37843_c0_g1 | Perivitellin ovorubin-A2 (S) | 239 | 27.36±2.25 | 41637.8 | 19.1 |
| Pdi37785_c0_g1 | Perivitellin ovorubin-A4 (S) | 379 | 13.81±2.49 | 69997.6 | 34.5 |
| Pdi67619_c0_g1 | Perivitellin ovorubin-A1 (S) | 289 | 20.49±0.61 | 39804.9 | 20.8 |
| Pdi33234_c1_g2 | Perivitellin ovorubin-B3a (S) | 15 | 0.54±0.52 | 1080.4 | 0.4 |
| Pdi14773_c0_g1 | Perivitellin ovorubin-A3 (S) | 277 | 11.71±3.69 | 59971.0 | 28.9 |
| Pdi118877_c0_g1 | Perivitellin ovorubin-B1 (S) | 164 | 1.26±0.54 | 39661.5 | 18.8 |
| **Immune response** | | | | | |
| Pdi101720_c0_g3 | C1q domain-containing protein (S) | 466 | 4.78±1.23 | 34962.5 | 16.4 |
| Pdi101708_c0_g1 | C1q domain-containing protein (S) | 18 | 1.57±1.50 | 33042.0 | 11.3 |
| Pdi5997_c0_g1 | CD109 antigen-like | 24 | 0.19±0.14 | 136.2 | 46.6 |
| Pdi35904_c0_g1 | CD109 antigen-like | 12 | 0.83±0.22 | 167.2 | 53.7 |
| Pdi104588_c0_g1 | CD109 antigen-like | 15 | 0.22±0.10 | 113.4 | 43.0 |
| Pdi36100_c1_g1 | CD109 antigen-like (S) | 46 | 0.26±0.12 | 88.7 | 508.1 |
| Pdi85127_c0_g1 | C-type lectin (S) | 13 | 0.79±0.75 | 6425.2 | 4.1 |
| Pdi35677_c2_g3 | C-type lectin (S) | 13 | 0.72±0.19 | 311.7 | 0.0 |
| Pdi36174_c0_g1 | Deleted in malignant brain tumors 1 protein-like (S) | 550 | 1.05±0.06 | 10479.3 | 5.7 |
| Pdi26423_c1_g1 | Deleted in malignant brain tumors 1 protein-like (S) | 11 | 0.19±0.01 | 395.2 | 1.8 |
| Pdi26788_c0_g1 | Soma ferritin | 33 | 3.07±0.79 | 5169.0 | 3130.3 |
| Pdi67699_c0_g1 | Tachylectin-related protein (S) | 57 | 2.18±1.53 | 8434.1 | 9.7 |
| Pdi19678_c1_g1 | Tachylectin-related protein (S) | 57 | 1.14±0.44 | 4234.5 | 1.7 |
| **Protein degradation** | | | | | |
| Pdi27770_c0_g3 | Aldehyde dehydrogenase family 3 member B1-like | 8 | 0.10±0.06 | 65.7 | 37.7 |
| Pdi35145_c2_g1 | Kunitz-like protease inhibitor (S) | 20 | 0.89±0.10 | 9256.2 | 5.0 |
| Pdi35260_c0_g1 | Leukocyte elastase inhibitor-like (S) | 107 | 1.41±0.70 | 6452.0 | 715.2 |
| Pdi27557_c0_g1 | ubiquitin-60S ribosomal L40 | 6 | 0.31±0.08 | 2814.5 | 1611.0 |
| Pdi14151_c0_g1 | WAP four-disulfide core domain protein 3-like (S) | 20 | 2.04±0.13 | 2954.4 | 2.7 |
| **Signaling and binding** | | | | | |
| Pdi35891_c1_g1 | 16 kDa calcium-binding protein-like | 7 | 0.24±0.06 | 220.3 | 39.5 |
| Pdi30433_c0_g2 | calcium-binding protein, partial (S) | 16 | 0.71±0.86 | 38457.9 | 18.4 |
| Pdi19486_c0_g1 | calcium-binding protein, partial | 8 | 0.29±0.07 | 2683.0 | 1.0 |
| Pdi37550_c0_g2 | Hypothetical protein (S) | 7 | 0.02±0.01 | 24.8 | 11.1 |
| **Transcription and translation** | | | | | |
| Pdi25317_c0_g1 | Apoptosis-inducing factor 3-like | 132 | 0.57±0.22 | 199.7 | 25.9 |
| **Metabolism** | | | | | |
| Pdi35984_c0_g2 | Acylcarnitine hydrolase-like | 4 | 0.07±0.07 | 10.6 | 11.7 |
| **Oxidation reduction** | | | | | |
| Pdi30870_c0_g1 | Superoxide dismutase [Cu-Zn] (S) | 4 | 0.03±0.04 | 26.6 | 13.4 |
| **Unknown** | | | | | |
| Pdi51889_c0_g1 | Novel protein (S) | 41 | 1.18±0.26 | 3373.1 | 2.6 |

[a] Proteins were annotated using NCBI nr database. [b] The mean and standard deviation of emPAI value of two or three biological replicates. [c] The expression level (TPM) of transcripts in albumen gland (AG) and other tissues (OT). "S" indicates the protein with a predicted signal peptide.

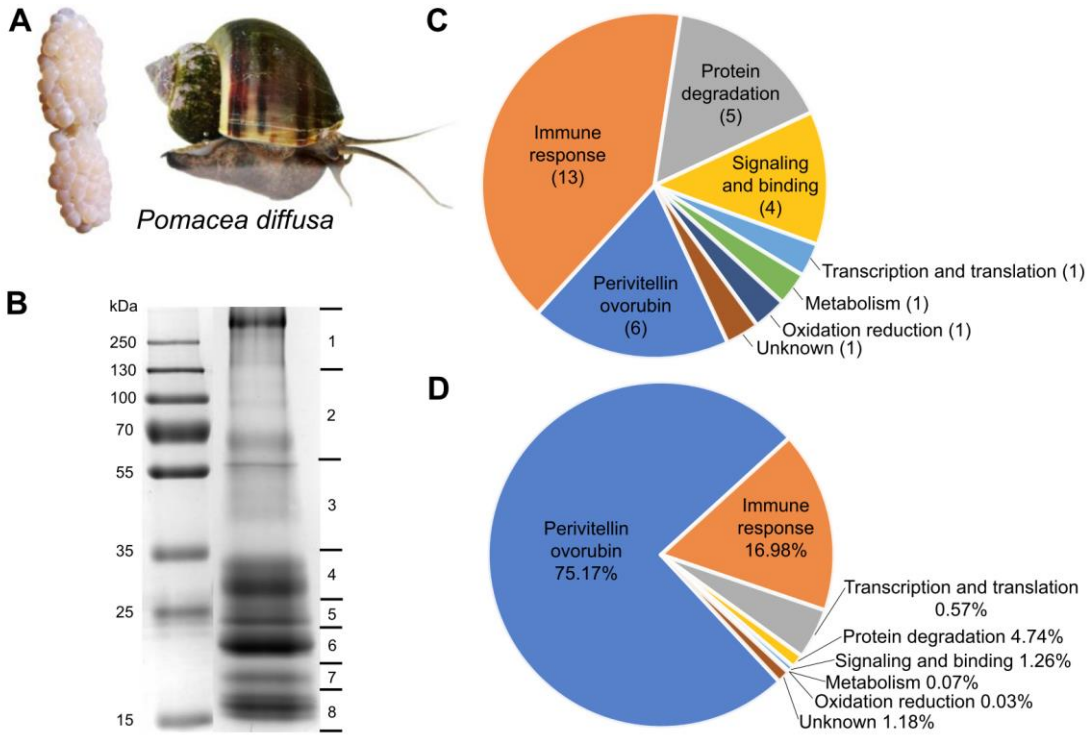



**FIGURE 1.** *Pomacea diffusa*. (A) A female (4.2 cm shell length) and its egg clutch (4.6 cm in length). (B) A SDS-PAGE gel image showing a molecular weight ladder (kDa, left) and two replicate eight gel slices of PVF for LC-MS/MS analysis (right). (C) Functional classification of PVF proteins. The number of proteins in each class is stated in parentheses. (D) Functional classification of PVF proteins by protein abundances based on emPAI values.

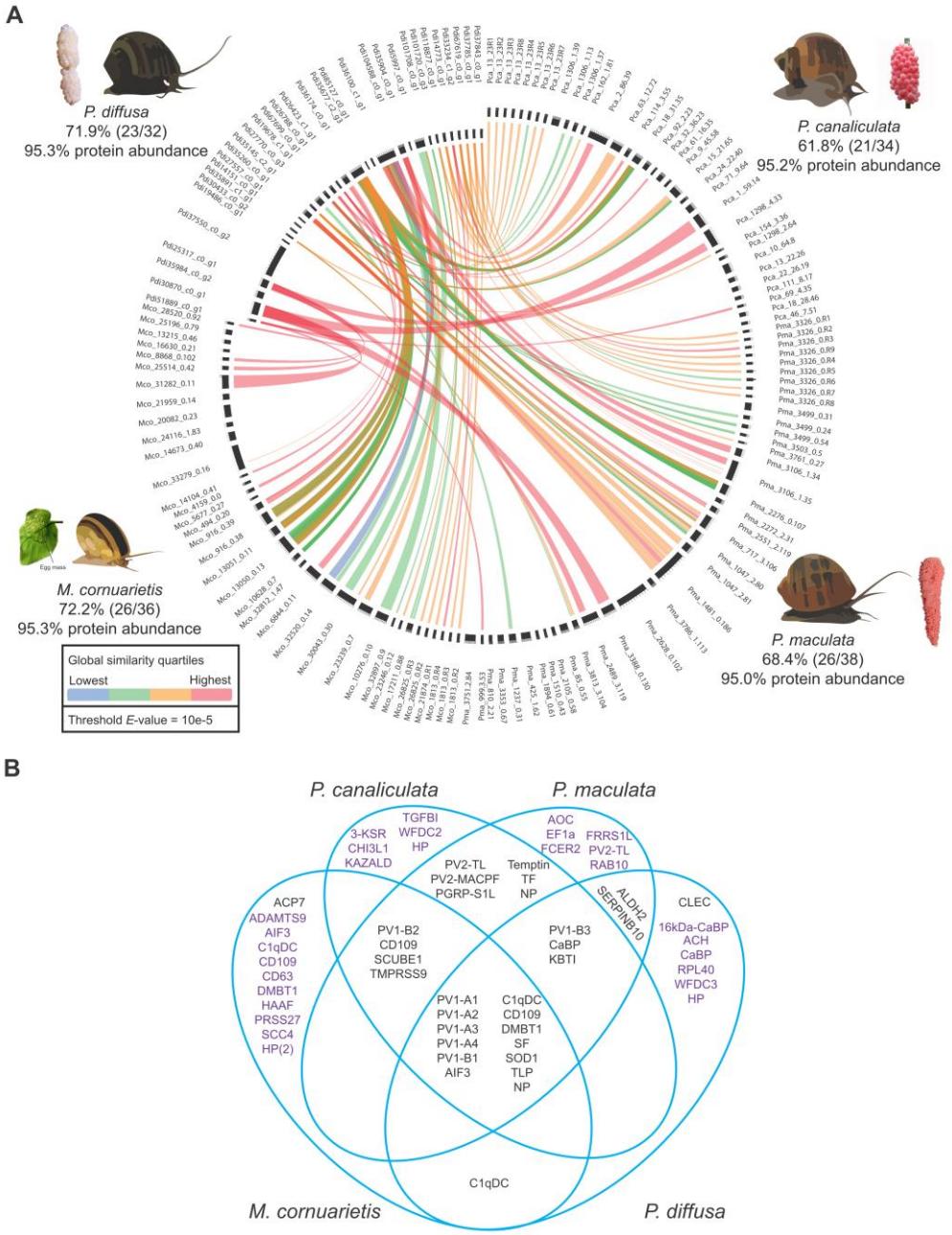



**FIGURE 2.** Comparison of PVF proteins in four ampullariids. (A) BLASTp results among the four ampullariids. Semi-transparent ribbons represent the quartile of similarity between proteins and black colour boxes represent the gene lengths. The number and percentage of PVF proteins with BLAST hit are shown with their corresponding protein abundance. Egg masses with characteristic color are shown in *M. cornuarietis* (colourless), *P. diffusa* (vivid pink), *P. canaliculata* (reddish) and *P. maculata* (reddish). (B) A Venn diagram of PVF orthologous families among the four species (details of the protein abbreviation and number of genes in orthologous group are included in Supporting information Table S3). Protein names in black are orthologous groups with two or more homologs, whilst those in purple are unique protein without any homolog.

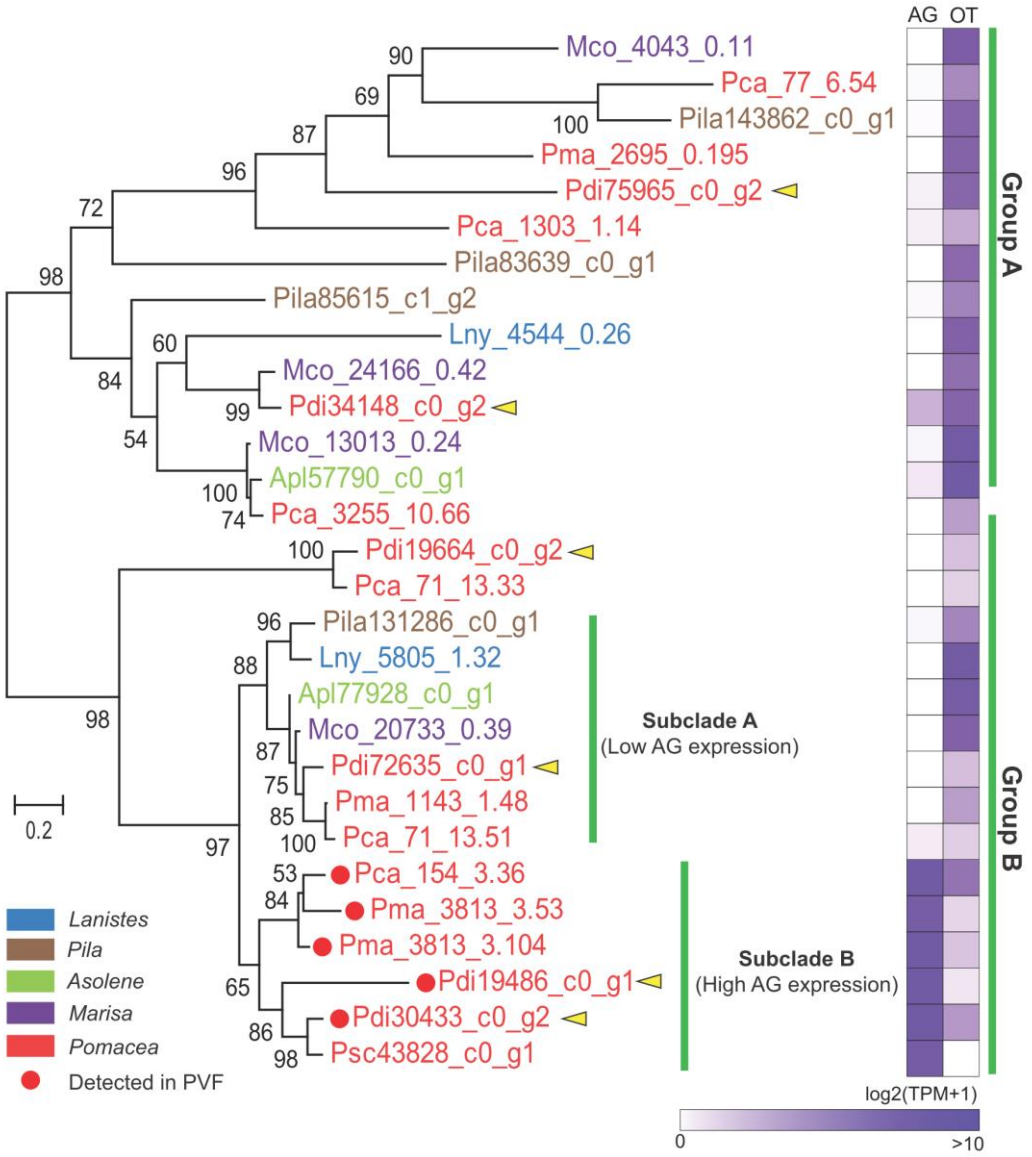
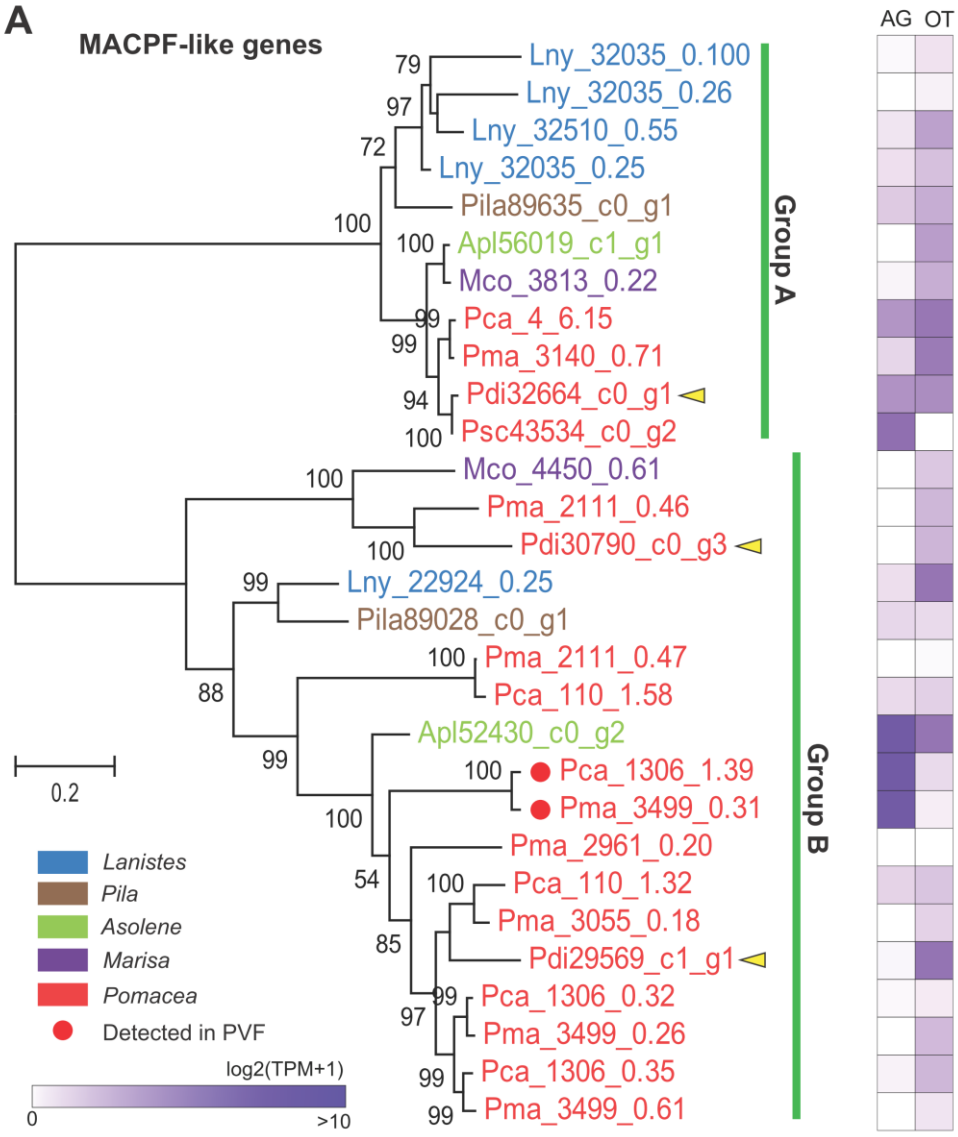



**FIGURE 3.** Phylogenetic tree of calcium binding proteins in eight ampullariids with numbers on nodes showing bootstrap values (>50%). Expression levels in log2 scale are presented on right side of the gene ID, which are colored by genus. *P. diffusa* sequences are indicated by yellow arrows. AG, albumen gland; OT, Other tissues (pooled tissues of digestive gland, foot, mantle). Lny: *L. nyassanus*, Pila: *Pila ampullacea*, Apl: *A. platae* Mco: *M. cornuarietis*, Pdi: *P. diffusa*, Psc: *P. scalaris*, Pca: *P. canaliculata*, Pma: *P. maculata*.



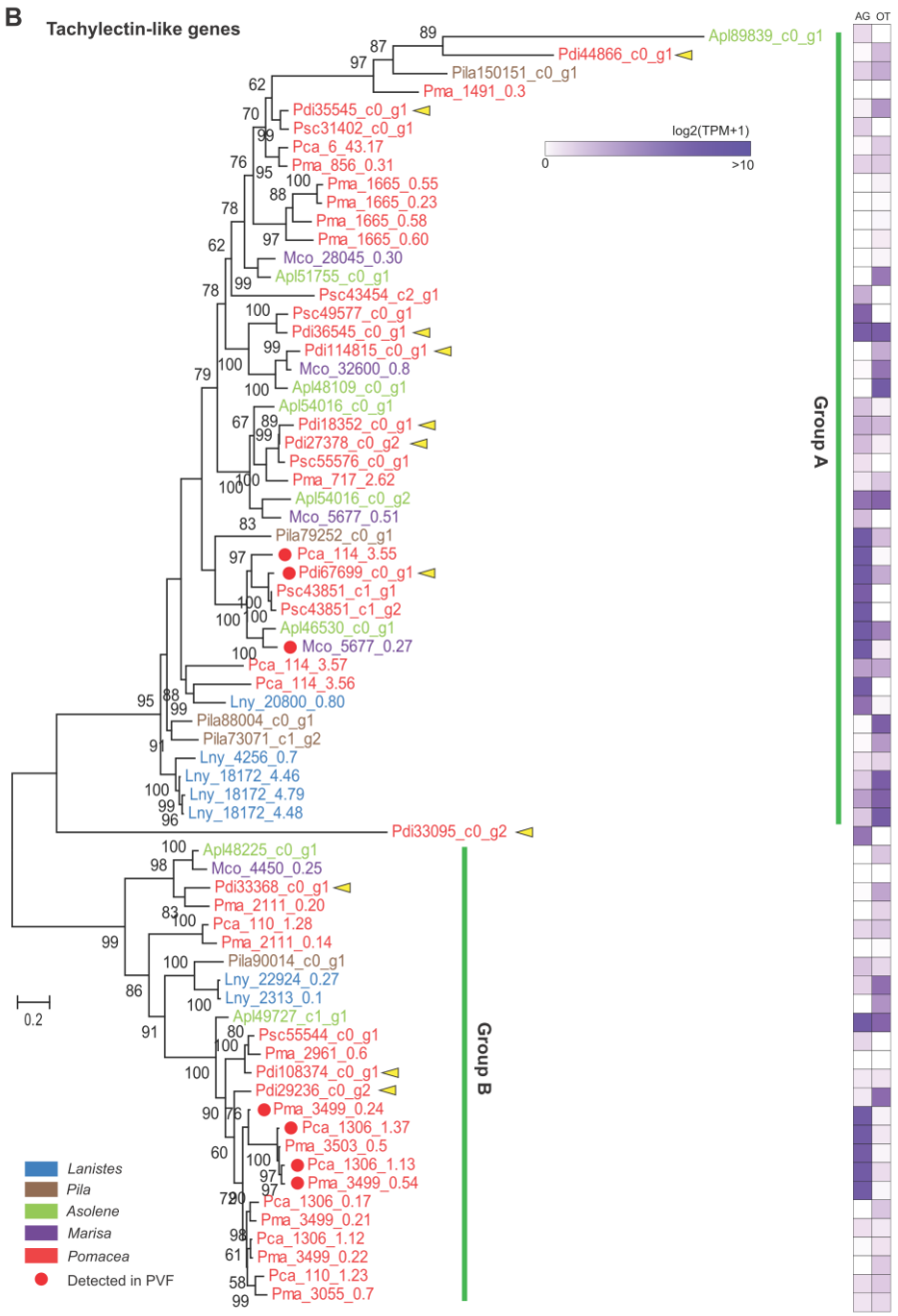



**FIGURE 4.** Phylogeny and expression of homologues of (A) MACPF-like and (B) tachylectin-like genes in eight ampullariids. Numbers on nodes indicate the bootstrap values (>50%). Gene expression levels are in log2. Yellow arrows indicate *P. diffusa* sequences. AG, albumen gland; OT, Other tissues (pooled tissues of digestive gland, foot, mantle). Lny: *L. nyassanus*, Pila: *Pila ampullacea*, Apl: *A. platae* Mco: *M. cornuarietis*, Pdi: *P. diffusa*, Psc: *P. scalaris*, Pca: *P. canaliculata*, Pma: *P. maculata*.