Genomic and transcriptomic landscapes and evolutionary dynamics of molluscan glycoside hydrolase families with implications for algae-feeding biology

Jing Wanga,b,1, Hongwei Yua,1, Lijie Yaoa,1, Yuli Lia,b,⇑, Fuyun Liua, Wentao Han a, Cong Cui a, Jingjie Hua,d, Zhenmin Baoc,d, Shi Wang a,b,d,⇑

a MOE Key Laboratory of Marine Genetics and Breeding and Sars-Fang Centre, Ocean University of China, Qingdao 266003, China
b Laboratory for Marine Biology and Biotechnology, Pilot Qingdao National Laboratory for Marine Science and Technology, Qingdao 266237, China
c Laboratory for Marine Fisheries Science and Food Production Processes, Pilot Qingdao National Laboratory for Marine Science and Technology, Qingdao 266237, China
d Laboratory of Tropical Marine Germplasm Resources and Breeding Engineering, Sanya Oceanographic Institution, Ocean University of China, Sanya 572000, China

Abstract

The hydrolysis of sugar-containing compounds by glycoside hydrolases (GHs) plays essential roles in many major biological processes, but to date our systematic understanding of the functional diversity and evolution of GH families remains largely limited to a few well-studied terrestrial animals. Molluscs represent the largest marine phylum in the animal kingdom, and many of them are herbivorous that utilize algae as a main nutritional source, making them good subjects for studying the functional diversity and adaptive evolution of GH families. In the present study, we conducted genome-wide identification and functional and evolutionary analysis of all GH families across major molluscan lineages. We revealed that the remarkable expansion of the GH9, GH10, GH18 and GH20 families and the wide adoption of carbohydrate-binding modules in molluscan expanded GH families likely contributed to the efficient hydrolysis of marine algal polysaccharides and were involved in the consolidation of molluscan algae-feeding habits. Gene expression and network analysis revealed the hepatopancreas as the main organ for the prominent expression of approximately half of the GH families (well corresponding to the digestive roles of the hepatopancreas) and key or hub GHs in the coexpression gene network with potentially diverse functionalities. We also revealed the evolutionary signs of differential expansion and functional divergence of the GH family, which possibly contributed to lineage-specific adaptation. Systematic analysis of GH families at both genomic and transcriptomic levels provides important clues for understanding the functional divergence and evolution of GH gene families in molluscs in relation to their algae-feeding biology.

1. Introduction

Carbohydrates including monosaccharides, oligosaccharides and polysaccharides are the most abundant and widely distributed organic compounds on Earth. Glycoside hydrolases (GHs) are the key enzymes participating in carbohydrate metabolism, which is a prerequisite for organisms to obtain energy [1–3]. Except for some Archaea and parasitic single-cell eukaryotes, GH genes are widely present in almost all organisms [4,5]. They are widely involved in the hydrolysis of sugars and glycoconjugates and play essential roles in many important biological processes, such as energy acquisition and metabolism, degradation of cellular components, recognition and adhesion between cells and other physiological and biochemical reactions [6,7]. In recent decades, GHs have become hot research subjects in various fields of biology, involving various groups of microorganisms, plants and animals [1,8–11].

Molluscs are the largest marine phylum in the animal kingdom. Many of them are herbivorous that utilize algae as their main nutritional source, which could be a characteristic of molluscan ancestors and makes these organisms good subjects for studying GH family evolution and adaptive roles [12,13]. The polysacchar-
rides in marine algae, including cellulose, hemicellulose, chitin and laminarin, are the main carbohydrate source of herbivorous molluscs, and these compounds are substantially different from those associated with terrestrial higher plants, which include mainly cellulose and starch [14–24]. As a main source of energy, important cellular components and signaling molecules, the differences in polysaccharide types are closely related to the diet as well as morphological and physiological functions of the animals that consume these compounds. For example, most molluscs can encode cellulases to degrade and utilize cellulose [25,26], while vertebrates and most ecdysozoans are generally considered unable to encode cellulases by themselves, and their cellulose degradation depends on intestinal commensal microorganisms [27–29]. The organic matrix of the “skeleton” of ecdysozoans and molluscs is composed of chitin, and chitinase is involved in the metabolism of their “skeleton” [29,30]. While previous studies have reported multiple GH families genes of 7 molluscs (Chlamys farreri, Crassostrea gigas, Pinctada fucata, Bathymodiolus platifrons, Modiolus philippinarum, Lottia gigantea, Octopus bimaculoides) and 14 other metazoan species including the deuterostomes Homo sapiens, Danio rerio, Branchiosoma floridae, Apostichopus japonicus, the protostomes Lingula anatina, Capitella teleta, Helobdella robusta, Tribolium castaneum, Cryptotermes secundus, Drosophila melanogaster, Daphnia pulex, Caenorhabditis elegans and the non-bilaterians Nemastotella vectensis, Acropora digitifera (Table S1). Since functional domains (catalytic domains) were the most important features for identification and classification of GH family genes [3,40], we used the command “hmmscan” in the software HMMER (v3.1b2) to search for GH domains, against the Pfam-A database [41] with the e-value threshold 1e-5. The reliability of the identified GH family genes was manually checked by their functional annotations with the expected hydroxylase activity derived from functional annotation database, including the Carbohydrate active enzyme database (CAZY) [4], Uniprot database [42], the NCBI non-redundant proteins (Nr) database [43], Gene ontology (GO) [44] and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database [45], according to the sequence similarities. The genomic distributions of GH18 family genes in the seven molluscan genomes were retrieved from genomic annotation files released by previous studies (Table S1) [33–35,37,46,47]. The conserved domain structures of all molluscan GH family genes were identified using the Batch CD-Search tool in NCBI database by searching against the Pfam Database (https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi).

2.2. Phylogenetic and evolutionary analysis of GH families

To construct the phylogenetic trees of GH family genes, the sequences of the longest catalytic domain (>120 amino acids) for each gene were first obtained. The multiple sequence alignments were performed using the MEGA 6 software with default parameters [48]. The Neighbor-Joining (NJ) trees were constructed using the p-distance method with 2000 bootstrap replicates. The active centre of the catalytic domain of each family was extracted based on the sequence alignment according to previous studies [7,49] and the sequence logo was generated using WebLogo (http://weblogo.berkeley.edu/) [50].

2.3. Expression analysis of the GH family genes of C. farreri and C. gigas

The transcriptome data of adult tissues/organisms and developmental stages of scallop C. farreri and oyster C. gigas used in the expression analysis were downloaded from NCBI database released by previous studies [35,37]. High-quality sequencing reads were retained by requiring 80% base with a quality value greater than 20. Then the high-quality reads were mapped to the corresponding reference genomes using the software STAR (version 2.4.1c) [51]. Reads mapped to exonic regions were counted for corresponding genes using the software featureCounts (v1.6.3) [52] to calculate the gene expression level in term of TPM (Transcripts Per Kilobase of exon model per Million mapped reads) values. The expression patterns of GH genes and GH families (the sum of the TPM values of GH genes in each family) of the C. farreri and the C. gigas were shown in heatmaps. For a given GH family, its highly expressed tissues/organisms were defined as those with TPM values higher than 1.5 folds of the average TPM of the other tissues/organisms.

2.4. Gene coexpression network analysis of the GH family genes

To ensure the reliability of network construction, the co-expression network was constructed using the expression profile of C. farreri, for which multiple replications of the main tissues/organisms are available but not in C. gigas. A total of 22,228 genes that expressed at least in one tissues/organisms of C. farreri with the TPM values higher than two were used for weighted gene co-expression network analysis (WGCNA) [53]. The coexpression network was constructed using the WGCNA package in R [54] with the parameters of ‘softPower = 15, cutHeight = 0.99 and minModuleSize = 450’. Each gene module was labeled in unique color and the unassigned genes were labeled in grey. The intramodular connectivity value (Kwithin) was used to measure a gene’s connection strength to other genes in the specified module (i.e., the hubness of a gene) [54]. The over-representation analysis of the GH family genes was performed for each module using a hypergeometric test with p-values adjusted by the Benjamini-Hochberg method for multiple-test correction [54,55]. To evaluate the expression pattern of each module among different tissues/organisms, TPM values of each gene were firstly normalized across the different tissues/organisms using the scale function in R (version 3.6.1), and the expression level of each gene cluster was represented in the median value. GO and KEGG pathway enrichment analyses of the gene modules were performed using the EnrichPath-peline with the FDR cutoff 0.05 [56,57]. Cytoscape [58] was used for visualization of the co-expression networks.
3. Results

3.1. Genomic landscape and phylogenetic analysis of molluscan GH families

We identified the full sets of GH family genes from the genomes of seven molluscs covering major molluscan lineages, in comparison with those identified from 14 additional animal species representative of the major groups in the animal kingdom. Within the molluscan group, the number of GH genes ranged from 70 to 167 (Fig. 1a; Table 1). These genes belonged to 26 different GH families (Fig. 1a; Table 1), of which 17 families were present in all seven molluscs, and only a few families were unique to certain species (Fig. S1a). Across-Metazoa comparison revealed that the four groups, Mollusca, Annelida, Ecdysozoa, and Deuterostomia, shared 17 GH families, accounting for 55% of all GH families. Compared with Annelida and Ecdysozoa, Mollusca shared more GH families with Deuterostomia (Fig. S1b), presumably a reflection of their conserved genomes [37,59]. Compared with other animal groups, a notable feature of molluscs was that both GH9 (endoglucanases) and GH10 (endo-1,4-β-xylanase) families were present and expanded in most herbivorous molluscs (Fig. 1; Table 1). The expansion of these two families was prominent in gastropods and bivalves (5–14 for GH9 and 2–13 for GH10; Fig. 1). In contrast to herbivorous molluscs, the cephapod contained only one GH9 and one GH10 family member. Compared with the shallow-water mussel M. philippinarum, the deep-sea mussel B. platifrons showed much reduced numbers of GH9 and GH10 family genes. Another remarkable feature of the molluscan GH family was the expansion of GH18 (chitinase) and GH20 (beta-hexosaminidase), which was also common in ecdyzoans (Fig. 1a). Although most GH families consist of more than one kind of enzymatic activities, approximately 77% molluscan GH families were identified as having only one enzymatic activity, including even the expanded GH9, GH10 and GH20 (Table 1; Fig. S2). For molluscan GH18, chitinase (EC 3.2.1.14) and N-acetylglucosaminidase (also known as chitosanase, EC 3.2.1.-) were identified, but only the chitinase genes were expanded in molluscs (Table 1; Fig. 2b,c). For the other GH families with two or more predicted enzymatic activities (i.e., GH2, GH13, GH30, GH31 and GH38), family members were divided into several subclades corresponding to different kinds of enzymes with different catalytic unit sequences (Table 1; Fig. S2-3). Besides this major class of carbohydrate-active enzyme, we also identified other enzymes related to polysaccharide degradation in the twenty representative metazoan species, such as alginate lyase (family PL14) and Lytic Polysaccharide Mono-Oxygenase (LPMO) (Table S2). We found that the expansion of PL14 may be a unique genomic feature belonging to gastropods, while LPMO showed remarkable expansion in molluscan genomes.

3.2. Functional domain diversity and its association with GH18 family evolution

Most of the molluscan GH families (42%) possessed a simple structural organization with only catalytic domains (Fig. 1b). Except for the catalytic domains, 31% of molluscan GH families had non-catalytic subdomains with β-sandwich fold (GH30_C, GH59_M, GH63_N, GH81_N, NtCtMGAM_N, Gal_mutarotase_2), zinc-like fold (GH20b) or jelly-roll fold (GH2_N) structures. Some of the GH2 (28.0%), GH9 (8.2%), GH13 (9.5%), GH16 (33.3%), GH18 (34.9%), GH20 (45.3%) and GH81 (25.0%) genes of the molluscan GH family contained one or more carbohydrate-binding modules (CBMs) located at the N/C-terminus or interspersed among catalytic domains, of which the most variable domain structures were present in the GH18 family (Fig. 1b). Moreover, some GH genes also had an auxiliary domain, such as the fibronectin type 3-like (Fn3-like) domain in the GH3 family, the trefoil (P-type) domain in the GH31 family, the alpha mannosidase middle domain (Alpha-mann_mid) in the GH38 family and the solute carrier family 3 member 2N-terminus domain (SLC3A2_N) in the GH13 family (Fig. 1b). It is worth mentioning that GH3 has two kinds of catalytic domains: a triose phosphate isomerase (TIM) barrel-like domain (GH3_N) and an α/β sandwich domain (GH3_C) with distinct catalytic residues (the catalytic nucleophile/base Asp in GH3_N and the catalytic proton donor Glu in GH3_C).

Most of the GH18 genes showed expansion and resided in tandem arrays (Fig. 2a; Fig. S4), while the other expanded families (GH9, GH10 and GH20) were mostly dispersed in the genome. Based on phylogenetic analysis, GH18 family genes in the phylum of Mollusca were classified into nine separate clusters (Group A-I) (Fig. 2b). The hydrolase activities of the GH18 members are identified as chitinases (EC:3.2.1.14) except for group A, which is di-N-acetylchitobiase (Fig. 2b). For groups A-E, most genes had only catalytic domains but lacked CBMs. In contrast, genes from groups F-I mostly had an N-terminal catalytic domain followed by one or more CBMs (Fig. 2b). The conserved sequence logos showed the active centres of the molluscan GH18 catalytic domains conserved with the amino acid residues FDGLD(L/M)DWE(Y/F)P, and the most critical residues (glutamate residue E and aspartic acid residue D) in the GH18 catalytic domain were conserved among all groups except group A with LDGXXNXXEX (Fig. 2b-c). In addition, compared with group F-I, the active centres of groups B-E were less conserved for leucine and tyrosine residues (Fig. 2c). The GH18 genes in Group G were absent in the bivalves. As a group specific to Gastropoda and Cephalopoda, the active centres of Group G were mostly different from those of other CBM-containing groups (i.e., FDGLDMDWEFP instead of FDGLDLDWEYP) (Fig. 2b-c).

3.3. Comprehensive transcriptome profiling of molluscan GH families

Gene expression profiling facilitates the understanding of the function and evolution of GH families. Although many molluscan genomes have been sequenced, comprehensive transcriptome data remain largely scarce for molluscs. Here, we chose C. farreri and C. gigas, the two molluscs with the best availability of comprehensive transcriptome datasets (Table S3) [35,36], for analyzing the expression profiles of GH genes in adult tissues/organs and across various developmental stages. Molluscan GH family genes generally showed diverse expression patterns across adult tissues/organs and developmental stages (Fig. 3a-b; Fig. S5). In particular, approximately half of the GH families showed highly restricted expression in the hepatopancreas (Fig. 3a-b; Fig. S5). During development, the highly expressed GH families in the hepatopancreas started to be expressed after D-larval period and reached peak expression levels in the later umbo larval period (Fig. 3a-b). The composition of highly expressed GH families in the hepatopancreas was similar between scallop and oyster (Table 1), probably reflecting similar preferences for algae between the two molluscs. The highly expressed GH families in the hepatopancreas of C. farreri, were also previously reported as the main types of scallop polysaccharide hydrolases through mass spectrometry analysis [31].

3.4. Coexpression gene network analysis of molluscan GH families

To identify coexpression modules and the key GH genes in the hydrolysis of polysaccharides and other biological processes, we further investigated the gene coexpression network of GH family genes across the main tissues/organs of C. farreri using the weighted gene coexpression network analysis (WGCNA) approach [54]. Network construction revealed 14 gene modules (M1-14) with different expression patterns among the main tissues/organs.
The top 10% genes of each module sorted by the $K_{within}$ values were regarded as the hub genes within the module, which suggested their key roles in functional regulation. A total of 110 GH genes from 25 GH families used for network construction were widely distributed in 14 gene modules and 15% of them were hub genes (Fig. 3c; Table S4). The diverse expression pattern...
Table 1
Summary of function, substrate, expression of the GH families in 7 molluscs.

| GH | Family member | Function | Enzymatic Activities | Substrate | Highly Expressed Tissue/Organs |
|----|---------------|----------|----------------------|-----------|-------------------------------|
| C. farreri (CF) | C. gigas (CG) | P. fucata (PF) | B. platifrons (BP) | M. philippinarum (MP) | L. gigantea (LG) | O. bimaculoides (OB) |
| GH1 | 6 | 1 | 4 | 5 | 10 | 7 | 1 | lactase-phlorizin hydrolase | EC:3.2.1.108; EC:3.2.1.62 | lactose, phlorizin | Hepatopancreas, Fgonad(CG) |
| GH2 | 2 | 4 | 5 | 3 | 3 | 6 | 2 | β-glucuronidase; β-mannosidase | EC:3.2.1.31; EC:3.2.1.25 | glucuronoside, mannosides | Hepatopancreas(CF), Hemolymph(CF) |
| GH3 | 2 | 2 | 4 | 4 | 5 | 21 | 2 | β-galactosidase | EC:3.2.1.21 | β-D-glucoside | Kidney(CF), Hemolymph(CG) |
| GH9 | 7 | 5 | 11 | 9 | 14 | 13 | 1 | endoglucanase | EC:3.2.1.4 | cellulose | Hepatopancreas |
| GH10 | 8 | 3 | 7 | 2 | 6 | 13 | 1 | endo-1,4-β-xylanase | EC:3.2.1.8 | xylan | Hepatopancreas, Mgonad(CG) |
| GH13 | 10 | 15 | 8 | 12 | 9 | 10 | 4 | α-amylase; α-glucosidase; 1,4-α-glucan branching enzyme | EC:3.2.1.1; EC:3.2.1.20; EC:2.4.1.18 | starch, etc | Hepatopancreas(CG) |
| GH18 | 23 | 18 | 20 | 33 | 31 | 24 | 24 | laminarinase | EC:3.2.1.39 | laminarin, etc | Hepatopancreas |
| GH20 | 18 | 19 | 13 | 18 | 9 | 11 | 7 | β-N-acetylhexosaminidase | EC:3.2.1.52 | hexosamine | Smooth_muscle(CF), Foot(CF), Hepatopancreas(CF) |
| GH26 | 1 | 0 | 1 | 0 | 0 | 4 | 0 | endo-1,4-β-mannosidase | EC:3.2.1.78 |mannans, galactomannans, glucomannanspect and other galacturonan | Hepatopancreas |
| GH30 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | polygalacturonase | EC:3.2.1.15 | / | |
| GH31 | 3 | 4 | 3 | 3 | 3 | 2 | 1 | glucosylceramidase; β-1,6-glucanase | EC:3.2.1.45; EC:3.2.1.75 | glucosylceramide, lutein, pustulan, β-glucan | Hepatopancreas |
| GH35 | 9 | 10 | 10 | 11 | 9 | 6 | 5 | β-glucosidase; maltase-glucosamylase; sucrase- \endgalactosidase | Broad Substrate (maltose, starch, etc) | | |
| GH38 | 11 | 12 | 11 | 19 | 19 | 7 | α-glucosidase; maltase-glucosamylase; sucrase- \endgalactosidase | Hepatopancreas |
| GH39 | 6 | 4 | 3 | 4 | 3 | 2 | 2 | α-mannosidase; mannosyl-oligosaccharide \alpha-1,3,1,6-α-mannosidase | | | |
| GH47 | 2 | 2 | 2 | 4 | 2 | 2 | 2 | α-mannosidase | EC:3.2.1.114 |mannoside | | |
| GH56 | 1 | 2 | 0 | 1 | 1 | 1 | 2 | hyaluronidase | EC:3.2.1.35 |hyaluronan | | |
| GH59 | 1 | 1 | 1 | 2 | 2 | 6 | 1 | galactocerebiosidase | EC:3.2.1.46 |galactosylceramidase, etc | | |
| GH63 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | mannosyl-oligosaccharide glucosidase | EC:3.2.1.106 | Specific oligosaccharide precursor | | |
| GH65 | 2 | 2 | 1 | 2 | 2 | 2 | 1 | protein-glucosylgalactosylhydroxyllysine glucosidase | EC:3.2.1.107 | Specific glucose from the disaccharide unit, trehalose | |
### Table 1 (continued)

| GH | Family member | Highly Expressed Tissue/Organs | Substrate | Enzymatic Activities | Function | EC Number |
|----|----------------|-------------------------------|-----------|---------------------|----------|-----------|
| C. farreri | CF58153.4, CF58153.5, CF58153.7, CF58153.8.1 | Hepatopancreas, mantle, eye | mannosides | a-1,2-mannosidase | EC:3.2.1.130 | 132 |
| C. farreri | CF58153.5, CF58153.7 | Kidney, hemolymph, eye | mannosides | a-1,2-mannosidase | EC:3.2.1.130 | 116 |
| C. farreri | CF58153.6 | Hepatopancreas, mantle, eye | mannosides | a-1,2-mannosidase | EC:3.2.1.130 | 113 |
| | | | | | | |
| **Total** | | | | | | **70** |

3.5. Functional analysis of the hepatopancreas-related module M2

The highly expressed GH genes were mostly found in the gene module M2 (42 GH genes), with a p-value of 5.20e-15 (adjusted by Benjamini-Hochberg) (Fig. 5a; Table S5), indicating significant enrichment. The M2 genes were highly expressed in the hepatopancreas (Fig. 5b). Functional analysis showed that these genes were enriched in the GO terms of carbohydrate binding, hydrolase activity, trans-ferase activity and oxidoreductase activity (Fig. 5c), and the KEGG pathways of fat digestion and absorption, protein digestion and absorption, PPAR signaling pathway, vitamin digestion and absorption and starch and sucrose metabolism (Fig. 5d; Table S6), which correlated well with the digestive function of the hepatopancreas.

3.6. Expansion of the GH18 family with functional divergence

To understand molluscan GH family expansion in terms of functional divergence, we focused on the GH18 family, as it is the largest GH family expanded among molluscan species. The di-N-acetyltchitobiase genes are widely expressed in different organs, while most of the expanded chitinase genes are preferentially expressed in the hepatopancreas, mantle or eye (Fig. 6a). These chitinase genes are located in tandem arrays with variable domain structures and are widespread in different coexpression modules as hub genes, making them a good subject for studying GH family evolution and functional divergence. When focusing on the expression profile and phylogenetic relationship of GH18 family genes, we found evidence of within-cluster expression divergence of the GH18 family. For example, for the largest, four clustered GH18 chitinase genes in the C. farreri genome (CF58153.4, CF58153.5, CF58153.7, CF58153.8.1), CF58153.5 and CF58153.7 were located in the middle of the array and were specifically highly expressed in the hepatopancreas, while the remaining two flanking copies were specifically highly expressed in the mantle (Figs. 2a; 6a). A similar phenomenon was also observed in oyster, but its hepatopancreas-related copies would be evolutionarily advantageous or adaptive for their algae-feeding biology.
4. Discussion

Carbohydrates are the most abundant and widely distributed organic compounds on Earth and play a huge diversity of roles, such as energy conversion, cell structural constitution and cell-recognition processes, in living organisms [60]. GHs play essential roles in these important biological processes, primarily by polysaccharide depolymerization [6,7]. As the largest marine phylum in the animal kingdom, molluscs are mainly herbivorous, with their main nutrition source being algae (e.g., diatoms, brown
algae and green algae), which provide different types of polysaccharides, including cellulose, hemicellulose, chitin and laminarin [1,11,17–24]. Understanding the genomic distribution, expression and evolutionary characteristics of the molluscan GH family is of importance for understanding their functional divergence and adaptive roles in molluscs.

In this study, the full sets of GH family genes from seven molluscan genomes were identified and grouped into 26 different GH families based on their sequence similarity and catalytic domains supported by the CAZy database [3,4]. According to the global view of GH family distribution among 21 species representing the main metazoan groups, the GH9 and GH10 families are prominently expanded in gastropods and bivalves whereas they are contracted in cephalopod, which is possibly a reflection of adaptive evolution for the herbivory in molluscs and carnicory in cephalopod [12,13,61]. This finding is very interesting, as GH9 (cellulases) and GH10 (endo-1,4-beta-xylanase) are mainly responsible for the hydrolysis of cellulose and hemicellulose, which are important components of the algal cell wall [16,17]. Moreover, molluscs are among very few animals that have retained cellulose digestion abilities [16,17,62–64]. The expansion of these two families may indicate adaptive evolution for the degradation and utilization of cellulose and hemicellulose components in algal cell walls. Compared with the shallow-water mussel M. philippinarum, the deep-sea mussel B. platifrons shows much reduced numbers of GH9 and GH10 family genes, possibly as an adaptation to its unusual nutritional supply via chemoautotrophic endosymbionts [46].

GH18 (chitinase) and GH20 (beta-N-acetylhexosaminidase) families were also expanded in molluscs. The beta-N-acetylhexosaminidase is an exo-type hydrolytic enzyme involved in chitin degradation by cleaving chitin oligosaccharides into the N-acetylglucosamine monomers, which has synergistic effect with GH18 family on chitin hydrolysis [65–67]. Chitin is not only the main polysaccharide from molluscan herbivore diets but also the main component of molluscan shell organic scaffolds [12,68,69], which may suggest the adaptive evolution of the GH18 family for the digestion of food-source chitin and the calcification of shells. Although diverse enzyme functions have been identified in GH9,
GH10, GH18 and GH20 according to the CAZy database, only one kind of enzyme genes that closely related to the hydrolysis of algal polysaccharides have been identified and expanded in each of the expanded GH family, which suggested the adaptive evolution of molluscs for algae feeding.

The characteristic gene structure of the GH family is its catalytic domain, which is often a $(\beta/\alpha)_8$ TIM barrel structure with conserved catalytic residues that shape the active pocket. The non-catalytic subdomains with $\beta$-sandwich fold, zincin-like fold or jelly-roll fold structures beside the catalytic units are involved in the modification of the pocket-shaped catalytic site and are supposed to be critical for the stabilization of enzyme-carbohydrate interactions [70–72]. The catalytic domains are responsible for the hydrolysis reaction of polysaccharides [1,26]. However, the insoluble feature of polysaccharides makes the GHs inefficient when targeting glycosidic bonds. To overcome these problems, some GH family genes often possess one or more CBMs, which can specifically bind to polysaccharide substrates and efficiently promote adsorption of the enzyme onto insoluble polysaccharides [73]. Approximately 20% members of all molluscan GH family genes contain one or more CBMs, which is higher than the rough ratio (7%) of the CAZy database entries containing one or more CBMs [4]. It is interesting that most of the CBM-containing families (GH9, GH10, GH18, GH20) that participate in cellulose, hemicellulose and chitin digestion are expanded in molluscs [30,74–76]. Especially in the GH18 family, 35% of copies possessed up to five CBMs located at the N/C-terminus or interspersed among catalytic domains, which can promote the association of the enzyme with the substrate. The location of CBMs has no effect on their carbohydrate-binding activity [77,78]. Approximately 8% of GH9 members in mollusks possessed CBMs, while no CBM domain was identified in cellulases of the termite C. secundus. Termites are one of the insects that can degrade lignocellulose efficiently with the aid of intestinal microbiota (e.g. bacteria, fungi and flagellates) [79,80]. The cellulase-acting enzymes from flagellates usually have no CBMs, which show activity towards water-soluble cellulose and hardly any activity towards insoluble cellulose. In contrast, cellulases from the symbiotic bacteria of termites usually contain CBMs and show remarkable activities towards insoluble cellulose [81–83]. The wide adoption of CBMs in molluscan-expanded GH families is therefore evolutionarily advantageous and facilitates the functional achievement of GH families in the digestion of diverse marine algal polysaccharides. In view of the important roles of intestinal microbiota in the energy metabolism, the symbiotic microbiota in molluscs is probably involved in the hydrolyzation of polysaccharides from marine algae. However, the exact roles of symbiotic microbiota in the complex metabolic process in Mollusca are still poorly understood, which deserves in-depth analysis for more comprehensive understanding on the evolution of molluscan feeding habits.

Moreover, some GH genes also had an auxiliary domain, such as the Fn3-like domain in the GH3 family, the trefoil domain in GH31 family, the Alpha-mann_mid domain in the GH38 family and the SLC3A2_N domain in the GH13 family. The Fn3-like domain in GH3 has been suggested to stabilize the incomplete catalytic domain and can aid the hydrolysis of cellulose by modifying its surface [84,85]. The SLC3A2_N domain in GH13 can mediate the transmembrane transport of all essential amino acids as a compo-
nent of heteromeric amino acid transport (HAT) [86]. The duplicated trefoil domain in the GH31 family is probably involved in protein–protein or carbohydrate moieties interactions [87]. The alpha-mann_mid domains are suggested to be involved in enzyme folding and substrate recognition [88]. Above all, different kinds of auxiliary domains are suggested to have important contributions to substrate binding and hydrolysis efficiency.

To understand molluscan GH family expansion for functional divergence, we focused on the GH18 family (chitinase), which is known to have complex and diverse biological functions and participate in molluscan shell formation, algal digestion, immunity, and early embryonic development [68,69,89–91]. GH18 family genes are widely distributed in representatives of all kingdoms. Extensive studies on the domain architecture, functional divergence and structure-function relationship have been carried out in mammals, insects, plants, fungi, bacteria, archaea, nematodes and viruses [92–94]. However, little is known about the characteristics and evolution of the molluscan GH18 family. The expansion of GH18 genes was previously found in insects [95,96]. In this study, molluscan GH18 family genes showed more prominent expansions than those of insects. Most of the expanded GH18 genes reside in tandem arrays, possibly resulting from recent gene duplications. Based on the phylogenetic analysis of molluscan GH18 family genes, species-specific members are dominant in several molluscan species, possibly indicating their recent intraspecific expansion. Functional analysis of active residues in molluscan GH18 catalytic domains showed their good preservation of key residues [e.g., glutamate residue (E) and aspartic acid (D)] that are crucial for the retention of enzymatic activity [97,98], indicating that the functionality of the expanded GH18 copies is possibly ensured by purifying selection.

Gene expression profiling showed that the molluscan GH family genes are generally widely expressed in a variety of adult tissues/organisms as well as at different developmental stages, which may be related to the extensive involvement of GH in various biological processes [6,7]. In particular, approximately half of GH families showed highly restricted expression in the hepatopancreas, which was related to the important roles of GH families in the digestion of polysaccharides. During the developmental process, the highly expressed GH families in the hepatopancreas started to be expressed after D-larval period and reached peak expression levels in the later umbo larval period. This corresponds well to the developmental formation of the molluscan hepatopancreas and the nutritional supply change as molluscs consumed their egg-stored energy and then started to obtain energy by ingesting algae in the surrounding aquatic environment [99]. According to the coexpression network, GH genes from 25 GH families were widely distributed in 14 gene modules and 15% of them were hub genes, which suggested their key roles in multiple biological processes. The GH gene-enriched module was the hepatopancreas-specific module. The wide-presence of diverse highly expressed GH families (14 GH families) in the hepatopancreas suggests a complex polysaccharide-hydrolase system in scallop, which coincides with the largely unique substrate specificity of polysaccharide-degrading enzymes and the diverse marine algal polysaccharides consumed by scallop [1,9,14]. Moreover, genes from the same GH family are widely distributed in different gene modules, which suggests their potential functional diversity. Taking GH18 genes
as an example, in addition to the highly expressed GH18 genes in the hepatopancreas, which are related to the digestion of chitin-containing food, four tandem-arranged genes are specifically expressed in the eyes of scallop (also the hub genes of eye-related module M10) and are therefore potentially involved in eye function. Chitin has been identified in the compound eyes of arthropods, where it is considered as a part of the visual system [100]. This may suggest that the expansion of eye-related GH18 in the scallop could be responsible for maintaining eye function in the scallop by chitin hydrolysis. The within-cluster expression divergence of the GH18 family was found both in scallop and oyster. The expression features and gene location of cluster members suggested that both hepatopancreas-related and mantle-related GH18 copies are already present in the common ancestor of scallop and oyster, and recent expansion of hepatopancreas-related copies independently occurred after the divergence between scallop and oyster. The higher expansion of hepatopancreas-related copies in oyster than scallop is likely a reflection of their differential efficiency of algae digestion based on the previous observation of the higher filtration efficiency of oyster over scallop especially at low algal concentrations [101].

5. Conclusions

We conducted the most comprehensive genome-wide identification, expression and evolutionary analysis of GH families to date for the largest marine phylum Mollusca. We revealed that compared with other animal groups, the remarkable expansion of GH9, GH10, GH18 and GH20 families in molluscs plays prominent roles in consolidating their algae-feeding habits. The wide adoption of CBMs in molluscan-expanded GH families likely contributes to their efficient hydrolysis of marine algal polysaccharides. Gene expression and network analysis revealed that the hepatopancreas was the main organ for the prominent expression of approximately half of GH families, which corresponds well to the herbivorous diets of molluscan species and the digestive roles of the hepatopancreas. We also revealed the evolutionary signs of differential expansion and functional divergence of the GH18 family for possi-
by contributing to lineage-specific adaptation. Taken together, through systematic analysis of GH families at both genomic and transcriptomic levels, our findings provide important clues for understanding the functional diversity and evolution of molluscan GH gene families and their association with molluscan algae-feeding biology.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We acknowledge the grant support from National Key Research and Development Program of China (2018YFC0310802), National Natural Science Foundation of China (U1706203 and 31702330), Fundamental Research Funds for the Central Universities (201762001 and 201841001), and Taishan Scholar Project Fund of Shandong Province of China.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.09.021.

References

[1] Henrissat B, Davies GJ. Glycoside hydrolases and glycosyltransferases. Families, modules, and implications for genomics. Plant Physiol. 2000;124:1515–9.
[2] Stutz AE, Wrodnigg TM. Inmino sugars and glycosyl hydrolases: historical context, current aspects, emerging trends. Adv Carbohydr Chem Biochem 2011;66:187–298.
[3] Henrissat B. A classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem J 1991;280:109–16.
[4] Lombard V, Goloncza Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZY) in 2013. Nucl Acids Res 2014;42:D490–5.
[5] Naumann DG. Hierarchical classification of glycoside hydrolases. Biochemistry 1995;34:5835–9.
[6] Coutinho PM, Henrissat B. Life with no sugars? J Mol Microbiol Biotechnol 1999;1:307–8.
[7] Davies GJ, Henrissat B. Structures and mechanisms of glycosyl hydrolases. Structure 1995:3:853–9.
[8] Berlemont R, Martinez AC. Glycoside hydrolases across environmental microbial communities. PLoS Comput Biol 2016;12:e1005300.
[9] Cones J, Raich L, Rovira C. Modeling catalytic reaction mechanisms in glycoside hydrolases. Curr Opin Chem Biol 2019;53:183–91.
[10] Nakamura T, Fahmi M, Tanaka J, Seki K, Kubota Y, et al. Genome-wide analysis of whole human glycoside hydrolases by data-driven analysis in Silico. Int J Mol Sci 2019;20:6290.
[11] Ojima T. Marine enzymes for biocatalysis: polysaccharide-degrading enzymes from heriborvous marine invertebrates. Philadelphia: Woodhead Publishing; 2013. p. 333–71.
[12] Steenek RS, Watling L. Feeding capabilities and limitation of herbivorous molluscs: a functional group approach. Mar Biol 1982;68:299–319.
[13] Haszpznaru G, Wanninger A, Molinsius. Curr Biol 2012;22:510–4.
[14] Ahmadi A, Moghadamtousi S, Abubakar S, Zandi K. Antiviral potential of chitin and its derivatives. Crit Rev Food Sci Nutr 2018;58:1664–80.
[15] Yang Z, Zhang L, Hu J, Wang J, Bao Z, et al. The evo-devo of molluscs: insights into bilaterian evolution from three silurian genomes. Nature 2013;493:526–31.
[16] Albertin CB, Simakov O, Mitrovs T, Wang Z, Pungor J, et al. The octopus genome and the evolution of cephalopod neural and morphological novelties. Nature 2015;524:220–4.
[17] Zhang G, Fang X, Guo X, Li L, Luo R, et al. The oyster genome reveals stress adaptation and complexity of shell formation. Nature 2012;490:49–54.
[18] Wang S, Zhang J, Jiao W, Li J, Xun X, et al. Scallop genome provides insights into evolution of bilaterian karyotype and development. Nat Ecol Evol 2017;1:0120.
[19] Li Y, Sun X, Hu X, Xun X, Zhang J, et al. Scallop genome reveals molecular adaptations to semi-sessile life and neurotrophs. Nature Commun 2017;8:1721.
[20] Takeuchi T. Molluscan genomics: implications for biology and aquaculture. Curr Mol Bio Rep 2017;3:175–85.
[21] Yang Z, Zhang L, Hu J, Wang J, Bao Z, et al. The de-vo of molluscs: insights from a genomic perspective. Evol Dev 2020. e12356.
[22] Henrissat B, Davies G. Structural and sequence-based classification of glycoside hydrolases. Curr Opin Struct Biol 1997;7:637–44.
[23] Punta M, Cogoll P, Eberhardt RY, Mistry J, Tate J, et al. The Pfam protein families database. Nucl Acids Res 2012;40:290–301.
[24] The UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 47:D200–7.
[25] O’Leary N, Wright M, Brister J, Ciufo S, Haddad D, et al. Insights into the cell wall. Eukaryot Cell 2009;8:1038–50.
[26] Stutz AE, Wrodnigg TM. Imino sugars and glycosyl hydrolases: historical context and current aspects. Front Physiol 2017;8:231.
[27] Davison A, Blaxter M. Ancient origin of glycosyl hydrolase family 9 cellulase from green algae. Proc Biol Sci 2008;178:209–15.
[28] Suzuki K, Ojima T, Bao Z, Wang S, et al. Molecular cloning and characterization of two novel cellulase genes from the mollusc Ampullaria crossean. J Comp Physiol B 2015;2015:825203.
[29] Ahmadi A, Moghadamtousi S, Abubakar S, Zandi K. Antiviral potential of chitin and its derivatives. Crit Rev Food Sci Nutr 2018;58:1664–80.
[30] Yang Z, Zhang L, Hu J, Wang J, Bao Z, et al. The evo-de of molluscs: insights from a genomic perspective. Evol Dev 2020. e12356.
[31] Henrissat B, Davies G. Structural and sequence-based classification of glycoside hydrolases. Curr Opin Struct Biol 1997;7:637–44.
[32] Punta M, Cogoll P, Eberhardt RY, Mistry J, Tate J, et al. The Pfam protein families database. Nucl Acids Res 2012;40:290–301.
[33] O’Leary N, Wright M, Brister J, Ciufo S, Haddad D, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucl Acids Res 2016;44:D733–45.
[34] The Gene Ontology Consortium. Ashburner M, Ball C, Blake J, Botstein D, et al. Gene ontology: tool for the unification of biology. Nat Genet (2000) 25: 25-29.
[35] The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res (2019) 47: D330-D338.
[36] Sun J, Zhang Y, Xu T, Zhang Y, Mu H, et al. Adaption to deep-sea chemosynthetic environments as revealed by mussels genomes. Nat Ecol Evol 2017;1:0121.
[37] Takeuchi T, Koyanagi R, Gyoda F, Kanda M, Hisata K, et al. Bivalve-specific gene expansion in the pearl oyster genome: implications of adaptation to sessile lifestyle. Zool Lett 2016;2:3.
[38] Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGAB: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 2013;30:2725–9.
[39] Liu S, Li H, Yang Y, Yang R, Yang W, et al. Genome-wide identification of chitinase and chitin deacetylase gene families in the oriental fruit fly, Bactrocera dorsalis (Hendel). Comp Biochem Physiol Part D 2018;27:13–22.
[40] Guo R, Smyth GK, Shi W, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 2014;30:923–5.
[41] Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis, statistical applications in genetics and molecular biology. Stat Appl Genet Mol Biol 2005;4:e17.
