

MolluscDB: an integrated functional and evolutionary genomics database for the hyper-diverse animal phylum Mollusca

Fuyun Liu¹, Yuli Li^{1,2,*}, Hongwei Yu¹, Lingling Zhang^{1,2}, Jingjie Hu^{1,3}, Zhenmin Bao^{1,3,4} and Shi Wang^{1,2,3,*}

¹MOE Key Laboratory of Marine Genetics and Breeding and Sars-Fang Centre, Ocean University of China, Qingdao 266003, China, ²Laboratory for Marine Biology and Biotechnology, Pilot Qingdao National Laboratory for Marine Science and Technology, Qingdao 266237, China, ³Laboratory of Tropical Marine Germplasm Resources and Breeding Engineering, Sanya Oceanographic Institution, Ocean University of China, Sanya 572000, China and ⁴Laboratory for Marine Fisheries Science and Food Production Processes, Pilot Qingdao National Laboratory for Marine Science and Technology, Qingdao 266237, China

Received August 12, 2020; Revised September 16, 2020; Editorial Decision October 01, 2020; Accepted October 03, 2020

ABSTRACT

Mollusca represents the second largest animal phylum but remains poorly explored from a genomic perspective. While the recent increase in genomic resources holds great promise for a deep understanding of molluscan biology and evolution, access and utilization of these resources still pose a challenge. Here, we present the first comprehensive molluscan genomics database, MolluscDB (<http://mgbase.qnlm.ac>), which compiles and integrates current molluscan genomic/transcriptomic resources and provides convenient tools for multi-level integrative and comparative genomic analyses. MolluscDB enables a systematic view of genomic information from various aspects, such as genome assembly statistics, genome phylogenies, fossil records, gene information, expression profiles, gene families, transcription factors, transposable elements and mitogenome organization information. Moreover, MolluscDB offers valuable customized datasets or resources, such as gene coexpression networks across various developmental stages and adult tissues/organs, core gene repertoires inferred for major molluscan lineages, and macrosynteny analysis for chromosomal evolution. MolluscDB presents an integrative and comprehensive genomics platform that will allow the molluscan community to cope with ever-growing genomic resources and will expedite new scientific discoveries for understanding molluscan biology and evolution.

INTRODUCTION

Mollusca, commonly known as shellfish, is the second largest phylum in the animal kingdom, with over 100 000 extant species. It also represents the largest marine phylum, containing ~23% of all named marine organisms (1–3). Molluscs are globally distributed and play vital roles in the structure and functioning of marine, freshwater and terrestrial ecosystems. They are among the first bilaterians to appear in fossil records and mark the extraordinary Cambrian explosion of animals ~540 million years ago (2). With tremendous diversity in morphologies, behaviours and lifestyles, they have survived several mass extinction events, which makes them well known as one of the most ancient and evolutionarily successful groups of invertebrates. Molluscs exhibit fascinating biological and evolutionary innovations, including a diversity of body plans and highly specialized structures (e.g. bivalve shells for defence and cephalopod arms for predation), adaptive life-history characters (e.g. up to 507 years life span for the bivalve *Arctica islandica* (4)) and extraordinary developmental flexibility (e.g. up to a 4.4-year egg-brooding period for the deep-sea octopus *Graneledone boreopacifica* (5)). Molluscs have been employed as excellent models for over 100 years in studies of developmental and cell biology, neurobiology, physiology, behaviour, evolution, population genetics and materials science. Moreover, many molluscs are important fishery and aquaculture species, accounting for ~22% of the total world aquaculture production (6). They therefore present an important source of food throughout the world and provide significant economic benefits to humans.

Despite their remarkable biological, evolutionary and ecological significance, molluscs have long been neglected from a genomic perspective (7,8). The rapid development of

*To whom correspondence should be addressed. Tel: +86 532 82031969; Fax: +86 532 82031969; Email: swang@ouc.edu.cn
Correspondence may also be addressed to Yuli Li. Email: liyuli@ouc.edu.cn

high-throughput sequencing technologies has pushed molluscan research into the genomics era. Decoding several molluscan genomes and transcriptomes has led to several major discoveries or breakthroughs, including heat shock protein and immune-related gene expansion for stressful intertidal zone and deep-sea adaptation (9–11), near-perfect preservation of bilaterian ancestor-like karyotypes (12,13), neural novelty evolution by extensive RNA editing (14,15), a single intercalation origin of metazoan larvae (16), and a deeply resolved molluscan phylogeny (1,17). While current molluscan genomic/transcriptomic resources have been accumulated and are rapidly increasing, the access and utilization of these scattered genomic resources pose a great challenge for the molluscan research community. There is an urgent need to establish a Mollusca genomics platform or database by integrating extensive genomic resources and developing convenient tools for comprehensive analysis of these data.

Towards this goal, we constructed the first comprehensive genomics database specifically for molluscs (named MolluscDB, <http://mgbase.qnlm.ac>) by integrating current molluscan genomic/transcriptomic resources and providing convenient tools for multi-level integrative and comparative analyses. MolluscDB enables a systematic view of genomic and transcriptomic information from various aspects and provides highly valuable, unique custom datasets or resources that are not available elsewhere. The database is compatible with computers, tablets, and mobile devices, and all data in MolluscDB can be freely accessed and downloaded.

OVERVIEW OF DATABASE STRUCTURE AND FUNCTION

MolluscDB represents the most comprehensive collection of 558 molluscan genomic/transcriptomic datasets (including 20 high-quality assembled genomes, 314 reference genome-profiled transcriptomes and 224 *de novo*-profiled transcriptomes) and 409 mitochondrial genomic resources (Figure 1, Table 1). These resources show outstandingly high taxonomy coverage of all the seven classes and ~87% of the total 53 orders (according to NCBI Taxonomy Database) in Mollusca. MolluscDB provides various genomic information, including genome assembly statistics, a genome phylogeny, fossil records, gene sequence, structure, functional annotations, expressional profiles, gene families, transcription factors and transposable elements. Convenient visualization of genomic information is compiled and integrated into a customized genome browser. MolluscDB also offers highly valuable, special-featured customized datasets or resources, including gene coexpression networks across various developmental stages and adult tissues/organs, the core gene repertoires inferred for Mollusca and descendent ancestors, and genome-by-genome macrosynteny analysis for inferring molluscan karyotype evolution. Moreover, MolluscDB provides useful and convenient tools for user-defined search of genes of interest, blast- and blat-based sequence comparison and PCR primer design. MolluscDB is implemented with the Linux operating system, using J2EE as the framework, MySQL as the back-end database and Apache Tomcat as the server.

Web user interfaces were developed based on JavaServer Pages (JSP), HTML5 and CSS3.

TAXONOMIC COVERAGE, MULTI-TYPE GENOMIC DATA AND PALEOBIOLOGICAL RECORDS

The phylum Mollusca is commonly divided into seven classes: Gastropoda, Bivalvia, Cephalopoda, Scaphopoda, Monoplacophora, Polyplacophora and Aplousobranchia. Comprehensive genomics resources offered by MolluscDB cover all seven molluscan classes. At the genome level, 20 high-quality molluscan genomes with well-annotated gene information (e.g. gene sequence, structure and function) are presented in MolluscDB (Table 2), which are derived from the Bivalvia, Gastropoda and Cephalopoda. A phylogenetic tree of the 20 molluscan genomes based on single-copy genes is shown on the MolluscDB homepage. Users can click on species names in the tree or names in the ‘Taxonomy’ module to view a brief biological introductions of each species and its genomic features or switch to frequently used modules through quick links at the bottom (Figure 2A, B).

Compared with genomic data, transcriptomic data are much more abundant and show much wider taxonomic coverage (particularly for taxa whose genomes are poorly investigated). All molluscan transcriptomic data deposited in the NCBI SRA database were searched, collected and filtered. In total, 314 reference genome-profiled transcriptomes derived from 12 species were chosen for further expression and network analysis, and 224 transcriptomes from 103 species without reference genomes (covering all seven molluscan classes) were *de novo* assembled and stored in the ‘Download’ module for free download. Users can browse detailed statistics of all the transcriptomes or download sequencing reads through related SRA links for further customized analysis in the ‘Transcriptomic Data’ module (Figure 2C).

Mitochondria, existing in almost all eukaryotic cells, are key components participating in many important biological processes. Compared with nuclear genomic data, mitogenomic data are much easier to obtain and have been an important resource for investigating molluscan phylogeny and evolution (30). We collected 409 molluscan mitochondrial genomes, covering 42 orders and seven classes. For each species, a Circos graph showing mitochondrial gene information and an associated table with detailed genomic positions are presented in the ‘Mitogenomic Data’ module (Figure 2D). Considering that some lineages in Bivalvia exhibit doubly uniparental inheritance (DUI; 31), the haplotype information for mitogenome and sex information for sequenced individual are also provided. Additionally, we also provide the sequences and annotations of each mitochondrial genome for users to download.

With an evolutionary history of ~540 million years and the possession of hardened mineralized exoskeletons, molluscs have been a well-characterized animal group with rich fossil records (32). These molluscan fossils provide crucial information for understanding molluscan phylogenetics and evolution. We collected fossil records derived from the Paleobiology Database (PBDB; (33)) for each Mollusca taxon. We organized all the searched records into a taxonomy tree presented in the ‘Paleobiological Records’ mod-

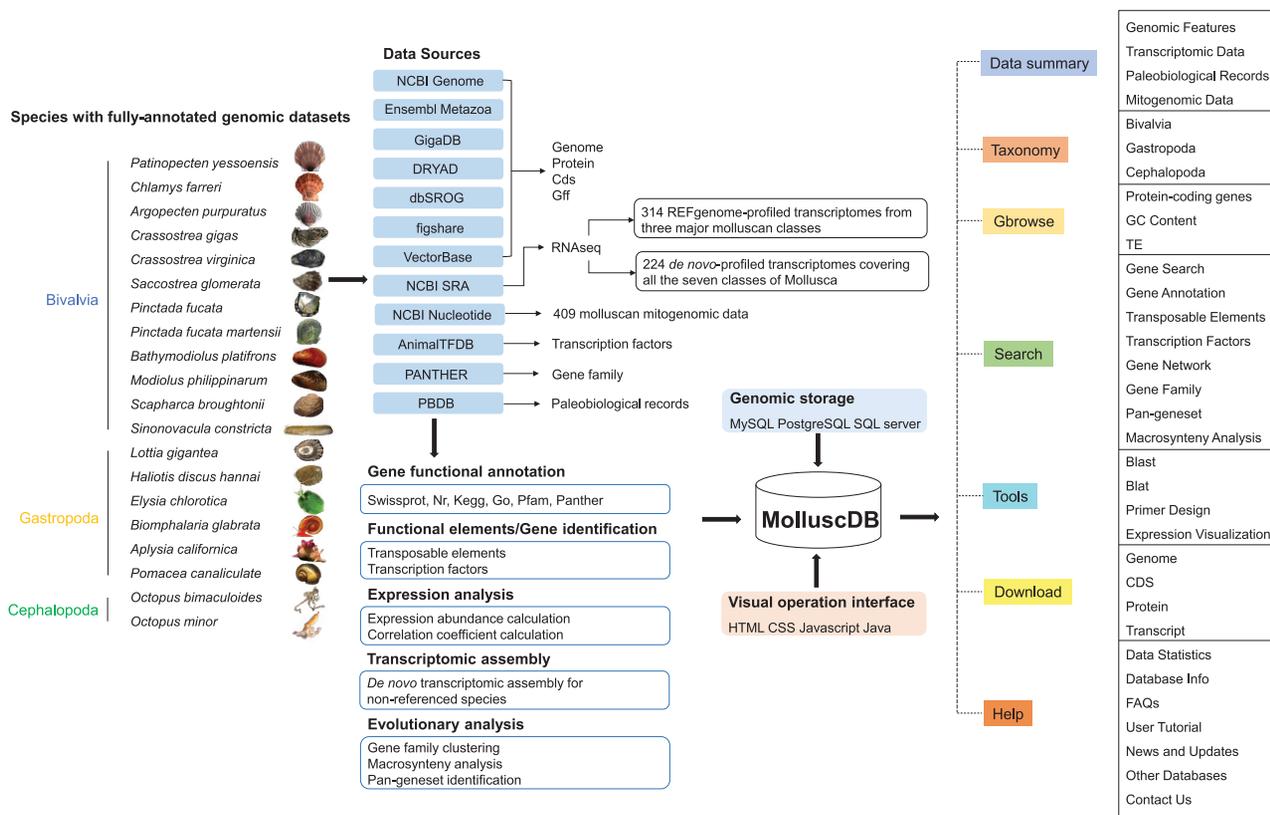


Figure 1. Overview of MolluscDB database structure and web interface features.

Table 1. Summary of MolluscDB data composition

Data	Statistics
Class /order/species	3/46/123
Protein-coding genes	563 593
Transcriptomic data/expression profiles	538
Mitogenomic data	409
Taxonomic categories with paleobiological records	241
Types of functional annotation database	6
Swissprot/Nr/Go/Kegg/Pfam/Panther annotation	347 623/508 505/277 773/165 238/411 647/455 626
Transposable elements/associated genes	72 640 596/522 372
Gene families/associated genes	29 151/513 684
Groups of Pan-geneset	38
Core gene families	122 434
Dispensable gene families	169 392
Core genes	513 684
Unclustered genes	49 909
Transcription factors/TF families	26 441/71
Co-expressed gene networks	18
Synten gene pairs	363 152

ule (Figure 2E) and linked the fossil record of each species with its relevant genomic/transcriptomic data. In total, 241 taxa distributed in seven Mollusca classes were labelled and linked with fossil records. Clicking on a labelled taxon name links to the external PBDB database and provides related paleobiology information, such as the morphology, dating and collection locations of fossils.

GENE ANNOTATION, TRANSPOSABLE ELEMENTS AND TRANSCRIPTION FACTORS

Functional annotation by homology comparison against public databases is crucial for understanding the possible functions of protein-coding genes. To comprehensively annotate 563,593 molluscan protein-coding genes, the ‘Gene Annotation’ module was set up, which compiles and integrates functional annotation information from six mainstream databases (Figure 3A), including NR (34), SwissProt (35), KEGG (36,37), GO (38,39), Pfam (40) and Panther (41). In total, 504 210 genes were annotated with at least one type of annotation. The detailed annotation information can be accessed by searching gene IDs for accurate matching or key words in annotation descriptions for fuzzy matching. Links to the ‘Gene Search’ and ‘Gbrowse’ modules in MolluscDB and external annotation databases are related to each gene ID or annotation ID, respectively. Download options are also provided for user-defined downloading of annotation information for selected genes or all the protein-coding genes of selected species.

Transposable elements (TEs) are major components of eukaryotic genomes, with significant impacts on genome evolution, function and disease (42). To ensure the consistency of TE identification across various genome datasets, we developed a uniform pipeline to re-annotate all 20 molluscan genomes for TE identification by referring to previously published methods (29,10). Specifically, in MolluscDB, all annotated TEs were correlated with protein-coding genes for conveniently exploring relationships be-

Table 2. Summary of 20 high-quality molluscan genome assemblies

Taxonomy	Species	Genome_size (Mb)	Number of protein-coding genes	Contig N50 (Kb)	Scaffold N50 (Kb)	GC_content (%)	Repeat_rate (%)	References/Resources	
Bivalvia	<i>Patinopecten yessoensis</i>	988	24 738	38	804	36.52	27.85	(13)	
	<i>Chlamys farreri</i>	780	28 602	22	602	35.49	27.73	(11)	
	<i>Argopecten purpuratus</i>	725	26 256	80	1 020	35.40	32.04	(18)	
	<i>Crassostrea gigas</i>	559	28 072	19	401	33.44	34.71	(9)	
	<i>Crassostrea virginica</i>	685	34 596	1 971	75 944	34.83	39.69	(19)	
	<i>Saccostrea glomerata</i>	788	29 738	40	804	33.31	45.39	(20)	
	<i>Pinctada fucata</i>	1024	31 477	21	167	35.03	43.35	(21)	
	<i>Pinctada fucata martensii</i>	991	30 815	21	324	35.32	48.01	(22)	
	<i>Bathymodiolus platifrons</i>	1660	33 584	13	343	34.17	47.25	(13)	
	<i>Modiolus philippinarum</i>	2630	36 549	20	100	33.96	59.66	(13)	
	<i>Scapharca broughtonii</i>	885	24 045	1798	4500	33.70	46.41	(23)	
	<i>Sinonovacula constricta</i>	1 332	26 273	679	57 990	35.45	36.65	(24)	
	Cephalopoda	<i>Octopus bimaculoides</i>	2 372	33 609	5	470	36.04	50.43	(14)
		<i>Octopus minor</i>	5 090	30 010	197	3020	36.34	75.62	(25)
Gastropoda	<i>Lottia gigantea</i>	360	23 818	96	1870	33.28	23.73	(12)	
	<i>Haliotis discus hamai</i>	1 865	29 449	14	211	40.51	36.07	(26)	
	<i>Elysia chlorotica</i>	558	24 980	29	422	37.65	29.25	(27)	
	<i>Biomphalaria glabrata</i>	916	25 550	19	48	35.99	43.79	(28)	
	<i>Aphysia californica</i>	927	19 944	10	917	40.35	39.70	NCBI Genome (AplCal3.0)	
	<i>Pomacea canaliculata</i>	440	21 533	1073	31 530	40.62	20.72	(29)	

tween TEs and potential target genes (Figure 3B), with associations between 72 640 596 TEs and 522 372 genes characterized. Users can search for a certain genomic interval, TE subfamily type or gene ID to obtain and download full annotation information. Gbrowse links are also provided for visualization of each TE and its related gene.

Transcription factors (TFs), functioning as ‘master regulators’ and ‘selector genes’, exert control over biological processes that regulate growth, development and response to the external environment (43,44). We identified TF genes and classified them into gene families according to the AnimalTFDB database (version 3.0; (45)). In total, 26 441 TF genes were obtained from the 20 molluscan genomes and then classified into 71 gene families. In the ‘Transcription Factors’ module, users can search for the TF family of a species, a class or even all classes by family name to obtain TF gene family member information (Figure 3C). Links are also provided in the ‘Gene Search’ module for TF genes, and download options are provided for TFs of interest to users.

GENOME BROWSING AND GENE SEARCHING

Basic genomic features and annotated functional elements for 20 high-quality molluscan genomes in MolluscDB are visualized using a customized ‘Gbrowse’ module (46). Users can quickly browse any selected genomic region through the genome browser and obtain a convenient view of related genomic annotations, including GC content, sequence and structure of protein-coding genes, and types of transposable elements (Figure 3D). Clicking on any element embedded in the browser will display detailed information in a new page. Users are also allowed to create custom tracks by uploading genomic files with prescribed forms.

The ‘Gene Search’ module, which is cross-linked to other modules through the gene ID, ingrates basic gene information from multiple aspects for the whole gene sets of 20 molluscan genomes. Users can search for specific genes through three types of key words, namely, genomic region, gene ID and gene name. The search results contain the

downloadable content of gene location, gene size, transcription direction, gene structure, functional annotations and genomic/CDS/protein sequence (Figure 3E). Links to Gbrowse and functional databases are also provided in this module for deep gene mining.

EXPRESSION PROFILES AND GENE NETWORKS

In addition to the basic gene information of sequence, structure, and function, MolluscDB also provides gene expression profiles in various developmental stages or major adult tissues/organs. We retrieved 314 reference genome-profiled RNA-Seq datasets belonging to 12 molluscan species from the NCBI SRA database to calculate gene expression profiles in the ‘Expression Visualization’ module (Figure 4A) based on a uniform processing pipeline (16). Users need to input gene IDs of specific species to view expression profiles in selected developmental stages or adult tissues/organs. The expression profiles are presented in the format of a heatmap or transcript per million [TPM] value table, which can be switched by clicking on the ‘Display Heatmap/TPM’ button.

Co-expressed genes, reflecting possible relationships in expression regulation and important for elucidating gene interactions, can be displayed in the format of a gene co-expression network according to the similarity of gene expression patterns (47). Co-expressed gene networks for 12 species were constructed based on Pearson’s correlation coefficient (PCC) values between pairs of genes (Figure 4B) and visualized by using JavaScript Cytoscape.js (48). In total, we filtered and acquired 61,500 highly correlated co-expressed gene pairs. For a given query gene, displayed as a red dot, we show the network of the top 20 target genes with the highest correlation values, displayed in black dots. Users can click on any co-expressed gene in the network to view its co-expression network. In addition, a summary table of all co-expressed genes (also with links to the ‘Gene Search’ module) and corresponding functional annotations are provided below the network.

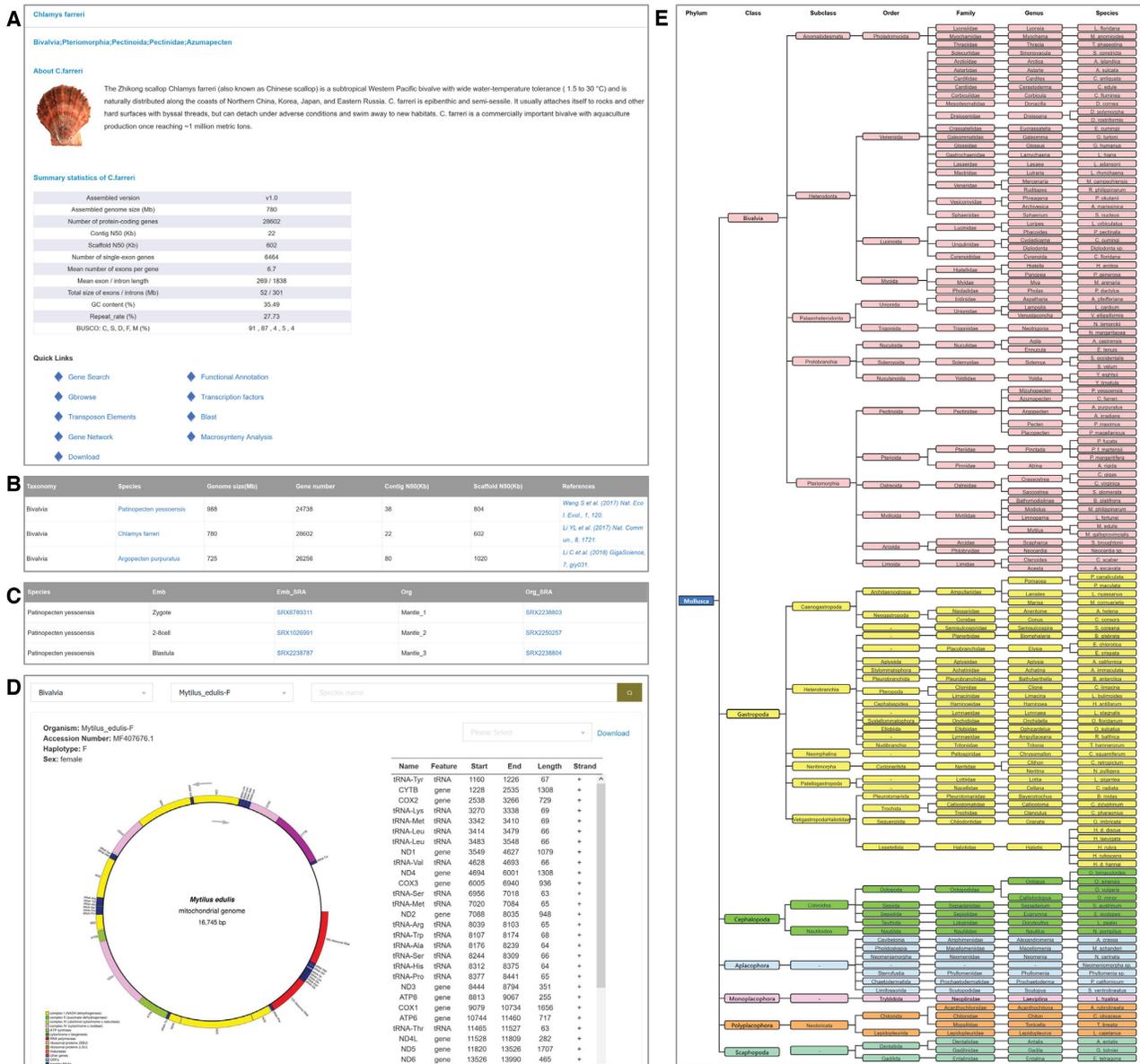


Figure 2. Screenshots for (A) overview of species information, (B) summary of genome assembly, (C) summary of transcriptomic data, (D) overview of mitogenomic information and (E) summary of paleobiological records.

GENE FAMILY, PAN-GENE SET AND MACROSYNTENY ANALYSIS

Identification and comparison of gene families are critical for understanding evolution and adaptation (49). Previous studies illustrated that the expansion of specific gene families is characteristic of molluscan genomes, which possibly corresponds to molluscan evolutionary success in terms of ecological adaptation and morphological diversity (9–11,14). In the ‘Gene Family’ module, we clustered and annotated gene families of 20 molluscan genomes based on OrthoMCL software (v2.0.9; (50)) and the Panther database (40), which resulted in 29 151 gene clusters containing 513 684 genes (Figure 3F). Users can search key words in annotation descriptions to obtain gene families of interest. Clicking on the number of each cluster will display

the genes with information on species, Panther annotation ID and description. To enable comparative analysis of gene families with other model organisms (e.g. fruit fly, mouse and zebrafish), Panther IDs in clustered gene families of molluscs were externally linked to the Panther database.

In an effort to define and characterize the pan-gene set for Mollusca at different phylogenetic levels, we set up the ‘Pan-geneset’ module, which provides information on core gene sets that are common to all species at a certain molluscan phylogenetic level and potentially dispensable gene sets that show presence/absence variations across species at the same phylogenetic level. Based on the gene family clustering results described above, we identified core/dispensable gene sets at 38 molluscan phylogenetic levels in the 20-mollusc phylogenetic tree (Figure 5A). To enable a view of the distri-

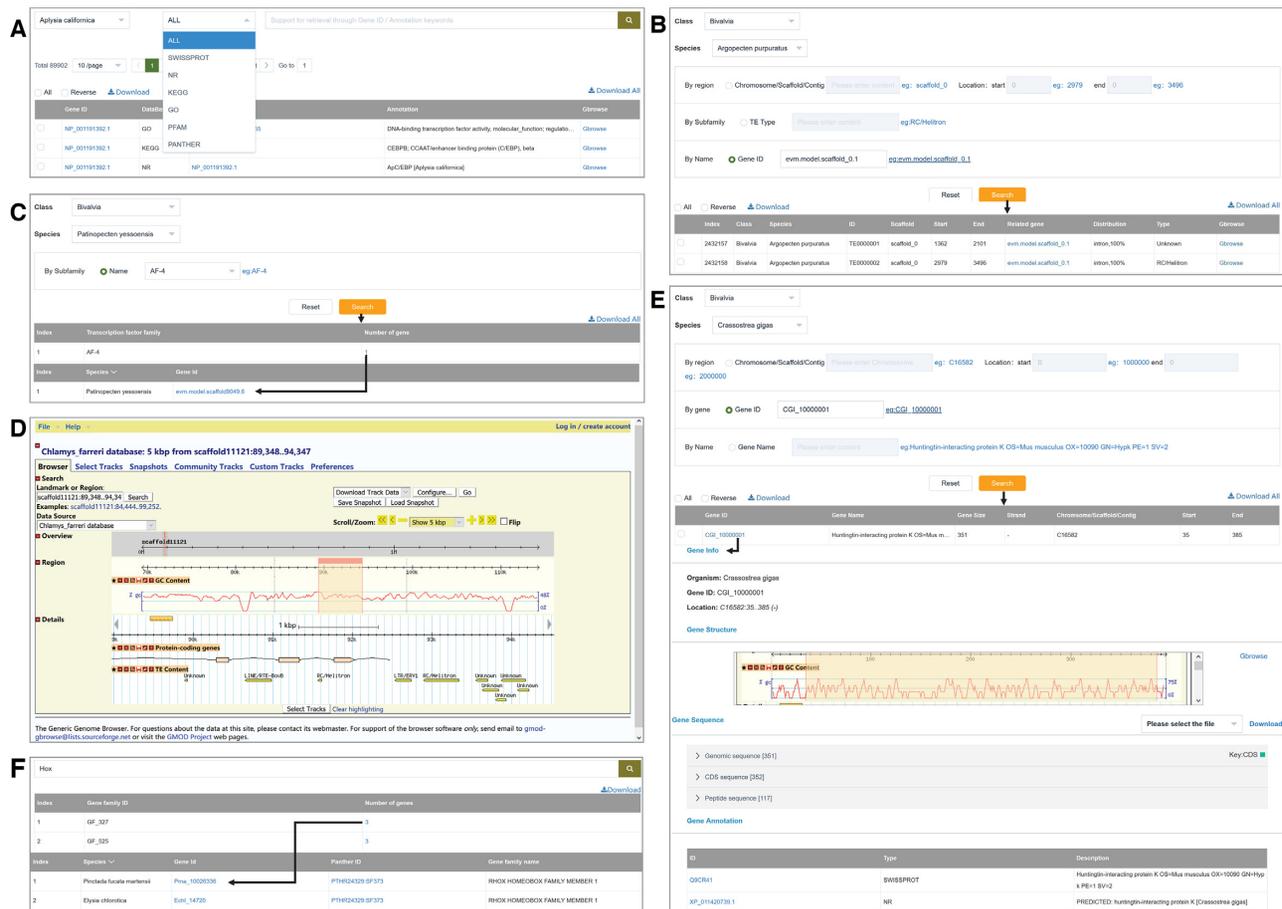


Figure 3. Screenshots for (A) gene annotation, (B) transposable elements, (C) transcription factors, (D) Gbrowse, (E) gene search and (F) gene family.

tribution of core/dispensable gene sets in individual genomes, we classify and visualize all protein-coding genes of each species according to their commonness at certain phylogenetic levels (i.e. phylum/class/order/family/genus/species). By clicking on certain bar graphs, the user can download the gene IDs of corresponding gene sets.

Macrosynteny analysis enables deep phylogenetic comparisons and an understanding of karyotype evolution by investigating conserved linkages between orthologous genes that are independent of intra-chromosomal rearrangements (12). Our previous macrosynteny analysis of 19 scallop chromosomes revealed that scallops may have a karyotype close to that of the bilaterian ancestor (13). Consistently a recent study supported the 19 presumed ancestral linkage groups (ALGs) of the bilaterian ancestor (51). To comprehensively investigate the evolution of molluscan karyotypes, we analysed the macrosyntenic relationships of 20 molluscan genomes with ancestral linkage groups represented by three conserved genomes (*Patinopecten yessoensis*, *Branchiostoma floridae* and *Nematostella vectensis*) by adopting the approach described by Simakov *et al.* (12) and Wang *et al.* (13). In this module (Figure 5B), users can view and compare the conservation level among 20 molluscan genomes according to different referred ALGs or focus on particular species by clicking on the dot plot to investigate detailed synteny relationships in an enlarged view. The download

option is provided for users to obtain macrosynteny dot plots, homologous gene pairs and related gene sequences.

CONVENIENT ONLINE TOOLS

MolluscDB also provides users with several convenient online tools. Using the 'primer design' tool, users can choose a genomic region or directly input a sequence to design primers for PCR experiments. Users can use 'Blast' or 'Blat' to search for targeted genes by entering user-supplied sequences that are aligned against the genome, CDSs, protein sequences or *de novo* assembled transcripts.

FUTURE DIRECTIONS

Currently, high-quality genomes are largely biased to the bivalves, gastropods, and cephalopods, but the situation is expected to quickly change as the rapid increase of genomic resources would eventually cover all molluscan lineages. In the future, we will continuously update MolluscDB as new molluscan genomes and omics data become available and will add more annotation and functionalities to the database, such as the incorporation of multiomics data (e.g. epigenome, proteome, metabolome, phenome and microbiome), developmental transcriptome age-based analysis for evo-devo research (16), molecular marker resources

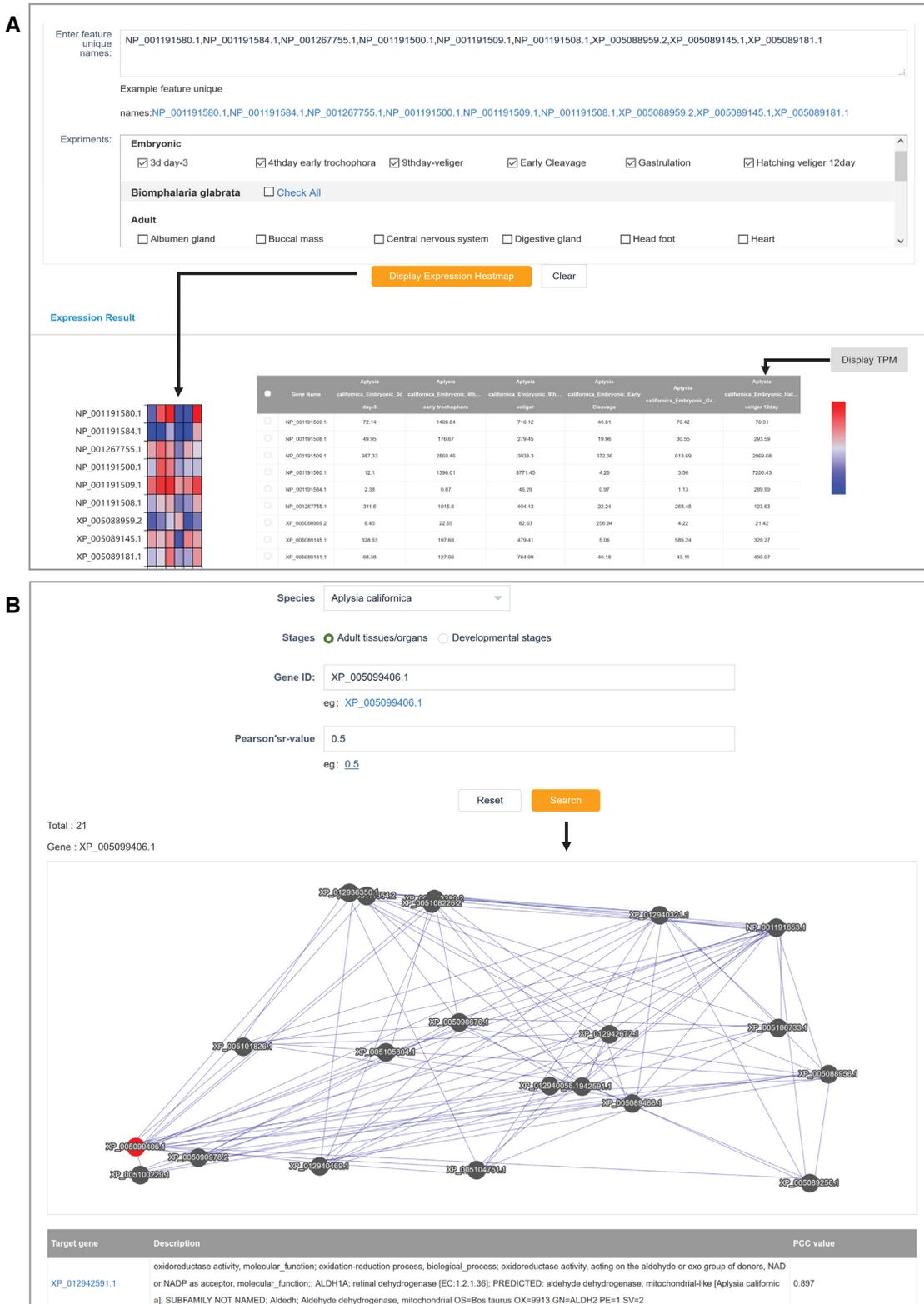


Figure 4. Screenshots for (A) expression visualization and (B) gene coexpression network.

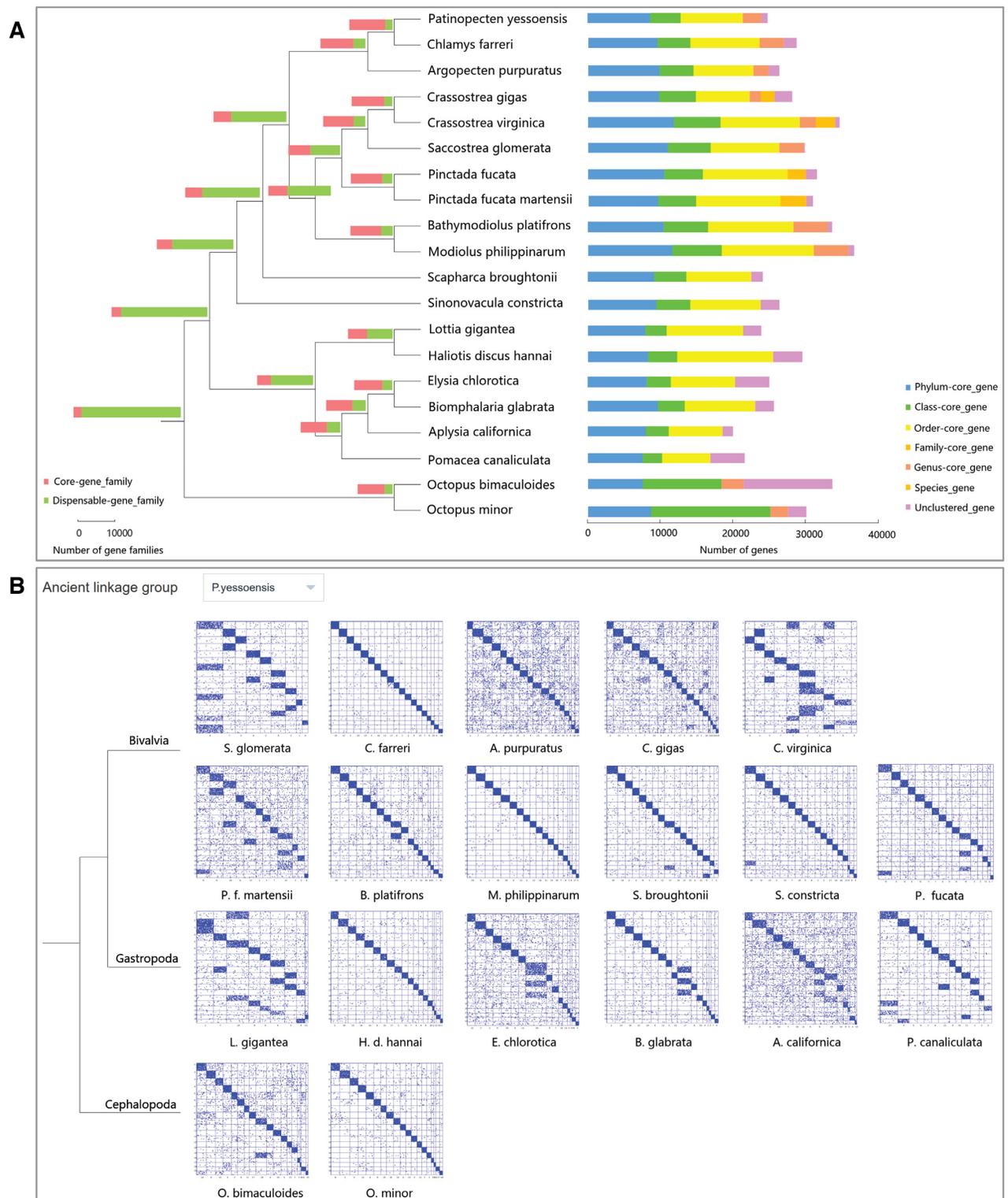


Figure 5. Screenshots for specially customized modules for (A) Pan-geneset analysis and (B) macrosynteny analysis.

(e.g. SNPs and microsatellites) for genomic breeding (52) and new machine learning-based tools for deep mining of multi-omics data (53) for understanding molluscan biology and evolution.

ACKNOWLEDGEMENTS

We wish to thank all researchers who have generated invaluable molluscan genomic resources that are gathered in the MolluscDB database. We thank Biomarker Technologies Corporation and Wuhan Goolgene Technology Co., Ltd. for assisting in MolluscDB construction. We also thank the Center for High Performance Computing and System Simulation (Qingdao Pilot National Laboratory for Marine Science and Technology) for the support of hardware resources and network services.

FUNDING

National Key Research and Development Program of China [2018YFC0310802]; National Natural Science Foundation of China [31871499, 31702330]; Major basic research projects of Shandong Natural Science Foundation [ZR2018ZA0748]; Fundamental Research Funds for the Central Universities [201841001, 202064008]; Taishan Scholar Project Fund of Shandong Province of China. *Conflict of interest statement.* None declared.

REFERENCES

- Kocot, K.M., Cannon, J.T., Todt, C., Citarella, M.R., Kohn, A.B., Meyer, A., Santos, S.R., Schander, C., Moroz, L.L., Lieb, B. *et al.* (2011) Phylogenomics reveals deep molluscan relationships. *Nature*, **477**, 452–456.
- Wanninger, A. and Wollesen, T. (2019) The evolution of molluscs. *Biol. Rev.*, **94**, 102–115.
- Wang, X. and Wang, Y. (2019) Editorial: molecular physiology in molluscs. *Front. Physiol.*, **10**, 1131.
- Butler, P.G., Wanamaker, A.D., Scourse, J.D., Richardson, C.A. and Reynolds, D.J. (2013) Variability of marine climate on the North Icelandic Shelf in a 1357-year proxy archive based on growth increments in the bivalve *Arctica islandica*. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **373**, 141–151.
- Robison, B., Seibel, B. and Drazen, J. (2014) Deep-sea octopus (*Graneledone boreopacifica*) conducts the longest-known egg-brooding period of any animal. *PLoS One*, **9**, e103437.
- FAO. (2019) In: *FAO yearbook. Fishery and Aquaculture Statistics 2017/FAO annuaire*.
- Gomes-dos-Santos, A., Lopes-Lima, M., Castro, L.F.C. and Froufe, E. (2020) Molluscan genomics: the road so far and the way forward. *Hydrobiologia*, **847**, 1705–1726.
- Yang, Z., Zhang, L., Hu, J., Wang, J., Bao, Z. and Wang, S. (2020) The evo-devo of molluscs: insights from a genomic perspective. *Evol. Dev.*, e12336, doi:10.1111/ede.12336.
- Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., Yang, P., Zhang, L., Wang, X., Qi, H. *et al.* (2012) The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*, **490**, 49–54.
- Sun, J., Zhang, Y., Xu, T., Zhang, Y., Mu, H.W., Zhang, Y.J., Lan, Y., Fields, C.J., Hui, J.H.L., Zhang, W.P. *et al.* (2017) Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. *Nat. Ecol. Evol.*, **1**, 121.
- Li, Y., Sun, X., Hu, X., Xun, X., Zhang, J., Guo, X., Jiao, W., Zhang, L., Liu, W., Wang, J. *et al.* (2017) Scallop genome reveals molecular adaptations to semi-sessile life and neurotoxins. *Nat. Commun.*, **8**, 1721.
- Simakov, O., Marletaz, F., Cho, S.J., Edsinger-Gonzales, E., Havlak, P., Hellsten, U., Kuo, D.H., Larsson, T., Lv, J., Arendt, D. *et al.* (2013) Insights into bilaterian evolution from three spiralian genomes. *Nature*, **493**, 526–531.
- Wang, S., Zhang, J., Jiao, W., Li, J., Xun, X., Sun, Y., Guo, X., Huan, P., Dong, B., Zhang, L. *et al.* (2017) Scallop genome provides insights into evolution of bilaterian karyotype and development. *Nat. Ecol. Evol.*, **1**, 120.
- Albertin, C.B., Simakov, O., Mitros, T., Wang, Z.Y., Pungor, J.R., Edsinger-Gonzales, E., Brenner, S., Ragsdale, C.W. and Rokhsar, D.S. (2015) The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature*, **524**, 220–224.
- Liscovitch-Brauer, N., Alon, S., Porath, H.T., Elstein, B., Unger, R., Ziv, T., Admon, A., Levanon, E.Y., Rosenthal, J.J.C. and Eisenberg, E. (2017) Trade-off between transcriptome plasticity and genome evolution in cephalopods. *Cell*, **169**, 191–202.
- Wang, J., Zhang, L., Lian, S., Qin, Z., Zhu, X., Dai, X., Huang, Z., Ke, C., Zhou, Z., Wei, J. *et al.* (2020) Evolutionary transcriptomics of metazoan biphasic life cycle supports a single intercalation origin of metazoan larvae. *Nat. Eco. Evol.*, **4**, 725–736.
- Smith, S.A., Wilson, N.G., Goetz, F.E., Feehery, C., Andrade, S.C., Rouse, G.W., Giribet, G. and Dunn, C.W. (2011) Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature*, **480**, 364–367.
- Li, C., Liu, X., Liu, B., Ma, B., Liu, F., Liu, G., Shi, Q. and Wang, C. (2018) Draft genome of the Peruvian scallop *Argopecten purpuratus*. *GigaScience*, **7**, g1031.
- Gómez-Chiarri, M., Warren, W.C., Guo, X. and Proestou, D. (2015) Developing tools for the study of molluscan immunity: the sequencing of the genome of the eastern oyster, *Crassostrea virginica*. *Fish Shellfish Immun.*, **46**, 2–4.
- Powell, D., Subramanian, S., Suwansa-Ard, S., Zhao, M., O'Connor, W., Raftos, D. and Elizur, A. (2018) The genome of the oyster *Saccostrea* offers insight into the environmental resilience of bivalves. *DNA Res.*, **25**, 655–665.
- Takeuchi, T., Koyanagi, R., Gyoja, F., Kanda, M., Hisata, K., Fujie, M., Goto, H., Yamasaki, S., Nagai, K., Morino, Y. *et al.* (2016) Bivalve-specific gene expansion in the pearl oyster genome: implications of adaptation to a sessile lifestyle. *Zool. Lett.*, **2**, 3.
- Du, X., Fan, G., Jiao, Y., Zhang, H., Guo, X., Huang, R., Zheng, Z., Bian, C., Deng, Y., Wang, Q. *et al.* (2017) The pearl oyster *Pinctadafucata martensii* genome and multi-omic analyses provide insights into biomineralization. *GigaScience*, **6**, 1–12.
- Bai, C., Xin, L., Rosani, U., Wu, B., Wang, Q., Duan, X.K., Liu, Z. and Wang, C. (2019) Chromosomal-level assembly of the blood clam, *Scapharca (Anadara) broughtonii*, using long sequence reads and Hi-C. *GigaScience*, **8**, g1067.
- Dong, Y., Zeng, Q., Ren, J., Yao, H., Lv, L., He, L., Ruan, W., Xue, Q., Bao, Z., Wang, S. *et al.* (2020) The chromosome-level genome assembly and comprehensive transcriptomics of the razor clam (*Sinonovacula constricta*). *Front. Genet.*, **11**, 664.
- Kim, B.M., Kang, S., Ahn, D.H., Jung, S.H., Rhee, H., Yoo, J.S., Lee, J.E., Lee, S., Han, Y.H., Ryu, K.B. *et al.* (2018) The genome of common long-arm octopus *Octopus minor*. *GigaScience*, **7**, g1119.
- Nam, B.H., Kwak, W., Kim, Y.O., Kim, D.G., Kong, H.J., Kim, W.J., Kang, J.H., Park, J.Y., An, C.M., Moon, J.Y. *et al.* (2017) Genome sequence of pacific abalone (*Haliotis discus hannai*): the first draft genome in family Haliotidae. *GigaScience*, **6**, doi:10.1093/gigascience/gix014.
- Cai, H., Li, Q., Fang, X., Li, J., Curtis, N.E., Altenburger, A., Shibata, T., Feng, M., Maeda, T., Schwartz, J.A. *et al.* (2019) A draft genome assembly of the solar-powered sea slug *Elysia chlorotica*. *Sci. Data*, **6**, 190022.
- Adema, C.M., Hillier, L.W., Jones, C.S., Loker, E.S., Knight, M., Minx, P., Oliveira, G., Raghavan, N., Shedlock, A., do Amaral, L.R. *et al.* (2017) Whole genome analysis of a schistosomiasis-transmitting freshwater snail. *Nat. Commun.*, **8**, 15451.
- Liu, C., Zhang, Y., Ren, Y., Wang, H., Li, S., Jiang, F., Yin, L., Qiao, X., Zhang, G., Qian, W. *et al.* (2018) The genome of the golden apple snail *Pomacea canaliculata* provides insight into stress tolerance and invasive adaptation. *GigaScience*, **7**, g1011.
- Simmons, W.B. and Boore, J.L. (2005) In: *Molluscan Evolutionary Genomics*. United States.
- Breton, S., Beaupré, H.D., Stewart, D.T., Hoeh, W.R. and Blier, P.U. (2007) The unusual system of doubly uniparental inheritance of mtDNA: isn't one enough? *Trends Genet.*, **23**, 465–474.
- Parkhaev, Y.P. (2017) Origin and the early evolution of the phylum Mollusca. *Paleontol. J.*, **51**, 663–686.

33. Peters, S.E. and McClennen, M. (2016) The Paleobiology Database application programming interface. *Paleobiology*, **42**, 1–7.
34. O’Leary, N.A., Wright, M.W., Brister, J.R., Ciuffo, S., McVeigh, D.H.R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
35. Consortium, T.U.P. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
36. Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.
37. Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. and Tanabe, M. (2019) New approach for understanding genome variations in KEGG. *Nucleic Acids Res.*, **47**, D590–D595.
38. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
39. Consortium, T.G.O. (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.
40. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A. *et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.
41. Mi, H., Muruganujan, A., Ebert, D., Huang, X. and Thomas, P.D. (2019) PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.*, **47**, D419–D426.
42. Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvak, Z., Levin, H.L., Macfarlan, T.S., Mager, D.L. and Feschotte, C. (2018) Ten things you should know about transposable elements. *Genome Biol.*, **19**, 199.
43. Hsia, C.C. and McGinnis, W. (2003). Evolution of transcription factor function. *Curr. Opin. Genet. Dev.*, **13**, 199–206.
44. Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R. and Weirauch, M.T. (2018) The human transcription factors. *Cell*, **172**, 650–665.
45. Hu, H., Miao, Y.R., Jia, L.H., Yu, Q.Y., Zhang, Q. and Guo, A.Y. (2019) AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Res.*, **47**, D33–D38.
46. Stein, L.D., Mungall, C., Shu, S.Q., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
47. Stuart, J.M., Segal, E., Koller, D. and Kim, S.K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.
48. Franz, M., Lopes, C.T., Huck, G., Dong, Y., Sumer, O. and Bader, G.D. (2016). Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics*, **32**, 309–311.
49. Raghupathy, N. and Durand, D. (2009) Gene cluster statistics with gene families. *Mol. Biol. Evol.*, **26**, 957–968.
50. Li, L., Stoeckert, C.J. Jr and Roos, D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.
51. Simakov, O., Marletaz, F., Yue, J.X., O’Connell, B., Jenkins, J., Brandt, A., Calef, R., Tung, C.H., Huang, T.K., Schmutz, J. *et al.* (2020) Deeply conserved synteny resolves early events in vertebrate evolution. *Nat. Ecol. Evol.*, **4**, 820–830.
52. Houston, R.D., Bean, T.P., Macqueen, D.J., Gundappa, M.K., Jin, Y.H., Jenkins, T.L., Selly, S.L.C., Martin, S.A.M., Stevens, J.R., Santos, E.M. *et al.* (2020) Harnessing genomics to fast-track genetic improvement in aquaculture. *Nat. Rev. Genet.*, **21**, 389–409.
53. Eraslan, G., Avsec, Ž., Gagneur, J. and Theis, F.J. (2019) Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.*, **20**, 389–403.