


# Genomic Insights into the Origin and Evolution of Molluscan Red-Bloodedness in the Blood Clam *Tegillarca granosa*

Yongbo Bao <sup>1,†</sup> Qifan Zeng<sup>2,†</sup> Jing Wang<sup>2,†</sup> Zelei Zhang<sup>1,3,†</sup> Yang Zhang<sup>4,†</sup> Sufang Wang<sup>1,†</sup> Nai-Kei Wong<sup>4,5,†</sup> Wenbin Yuan,<sup>1,3</sup> Yiyi Huang,<sup>1,3</sup> Weifeng Zhang,<sup>1,3</sup> Jing Liu,<sup>2</sup> Liyuan Lv,<sup>1</sup> Qinggang Xue,<sup>1</sup> Shanjie Zha,<sup>1</sup> Zhilan Peng,<sup>1</sup> Hanhan Yao,<sup>1</sup> Zhenmin Bao,<sup>2</sup> Shi Wang<sup>2,\*</sup> and Zhihua Lin<sup>1,\*</sup>

<sup>1</sup>Key Laboratory of Aquatic Germplasm Resource of Zhejiang, College of Biological & Environmental Sciences, Zhejiang Wanli University, Ningbo 315100, China

<sup>2</sup>Sars-Fang Centre & MOE Key Laboratory of Marine Genetics and Breeding, Ocean University of China and National Laboratory for Marine Science and Technology (LMBB & LMFSFP), Qingdao 266000, China

<sup>3</sup>School of Marine Sciences, Ningbo University, Ningbo 315010, China

<sup>4</sup>CAS Key Laboratory of Tropical Marine Bio-resources and Ecology and Guangdong Provincial Key Laboratory of Applied Marine Biology, South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou 510301, China

<sup>5</sup>The Eighth Affiliated Hospital of Sun Yat-Sen University, Shenzhen 518000, China

<sup>†</sup>These authors contributed equally to this work

\*Correspondence authors: E-mails: zhihua9988@126.com; swang@ouc.edu.cn

Associate editor Amanda Larracuent

## Abstract

Blood clams differ from their molluscan kins by exhibiting a unique red-blood (RB) phenotype; however, the genetic basis and biochemical machinery subserving this evolutionary innovation remain unclear. As a fundamental step toward resolving this mystery, we presented the first chromosome-level genome and comprehensive transcriptomes of the blood clam *Tegillarca granosa* for an integrated genomic, evolutionary, and functional analyses of clam RB phenotype. We identified blood clam-specific and expanded gene families, as well as gene pathways that are of RB relevant. Clam-specific RB-related hemoglobins (Hbs) showed close phylogenetic relationships with myoglobins (Mbs) of blood clam and other molluscs without the RB phenotype, indicating that clam-specific Hbs were likely evolutionarily derived from the Mb lineage. Strikingly, similar to vertebrate Hbs, blood clam Hbs were present in a form of gene cluster. Despite the convergent evolution of Hb clusters in blood clam and vertebrates, their Hb clusters may have originated from a single ancestral Mb-like gene as evidenced by gene phylogeny and synteny analysis. A full suite of enzyme-encoding genes for heme synthesis was identified in blood clam, with prominent expression in hemolymph and resembling those in vertebrates, suggesting a convergence of both RB-related Hb and heme functions in vertebrates and blood clam. RNA interference experiments confirmed the functional roles of Hbs and key enzyme of heme synthesis in the maintenance of clam RB phenotype. The high-quality genome assembly and comprehensive transcriptomes presented herein serve new genomic resources for the super-diverse phylum Mollusca, and provide deep insights into the origin and evolution of invertebrate RB.

**Key words:** blood clam, red-bloodedness, genome sequencing, hemoglobin evolution, heme biosynthesis.

## Introduction

The vivid red color of erythrocytes characteristically arises from the heme group of hemoglobin (*Hb*). Virtually all vertebrates rely on red blood (RB) cells and *Hb* to sustain oxygen transport in a circulatory system. In contrast, red-bloodedness rarely occurs in invertebrates, which have been found only in limited species among six invertebrate phyla, namely: Phoronida, Annelida, Nemertina, Echiuroidea, Mollusca, and Echinodermata (Mangum 2010). Other than oxygen transport, RB cells are also functionally vital to ATP and S-nitrosothiol release (Diesen et al. 2008; Wan et al. 2008), nitric oxide synthesis (Kleinbongard et al. 2006), hydrogen sulfide production (Benavides et al. 2007), and immune defense

(Jiang et al. 2007). Compared with species which have no respiratory protein, owning *Hb* in RB cells can dramatically increase blood oxygen (O<sub>2</sub>) concentrations and permit many other physiological functions (Storz 2018). Although the structure, function, and evolution of *Hb* have been a subject of intensive studies in vertebrates, the origin, evolutionary dynamics, and functional significance of invertebrate *Hb* remain poorly understood.

In the majority of invertebrates, hemolymph usually appears slightly bluish with hemocyanin as respiratory protein or colorless when there is no respiratory pigment. Within the super-diverse phylum Mollusca, the bivalve family Arcidae is one of few animal groups that possess RB-related *Hbs* as

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

Open Access

respiratory proteins (Terwilliger 1998; Mangum 2010; Bao et al. 2011). The blood clam *Tegillarca granosa* is a marine bivalve that belongs to the Arcidae family and is of commercial importance globally. This peculiar bivalve is thus named due to the unusual presence of *Hb* in its hemocytes. Typically, blood clams reside in sulphuretted mudflat or muddy sand, where oxygen availability becomes periodically limited. Genetic and functional diversity of *Hbs* has been reported in Arcidae species (Suzuki et al. 1989a; Titchen et al. 1991; Gambacurta et al. 2000; Mangum 2010; Bao et al. 2011; Ronda et al. 2013). To date, homodimeric, heterodimeric, tetrameric, and polymeric *Hbs* have been found in blood clams (Nicol and O’Gower 1967; Chiancone et al. 1981; Furuta and Kajita 1983; Suzuki et al. 1989b; Suzuki and Arita 1995; Bao et al. 2011). Consequently, blood clams constitute an excellent model for elucidating the complex evolutionary history of invertebrate *Hbs*.

The origin and evolution of RB-related *Hbs* are of great interest, in part, because of their probable roles in vertebrate adaptation to different environments. During vertebrate evolution, gene duplication played a vital role in *Hb* gene evolution and functional divergence (Storz 2016). For instance, Gadiformes fish possess expanded numbers of *Hb* genes driven by gene duplications to adapt the variable environment in shallower water (Baalsrud et al. 2017). Tandem duplications also enable subfunctionalization of *Hb* paralogs in birds (Grispo et al. 2012). Recently, the evolutionary origin and functions of human *Hb*’s heterotetrameric architecture have been elucidated, with a few substitutions driving the evolution of heterotetramer *Hb* from its ancestor dimeric precursor (Pillai et al. 2020). In comparison, in blood clams, the major *Hb* component *HbII* is formed by two polypeptide chains encoded as *HbIIA* and *HbIIB* to form a heterotetramer similar to human *Hb*. The minor *Hb* component, *HbI*, is composed of a third type of chain responsible for forming a homodimer (Chiancone et al. 1981; Furuta and Kajita 1983; Terwilliger et al. 1988; Bao et al. 2011; Ronda et al. 2013). The homo dimer in blood clam is significantly different in structure and type of *Hbs* from human. Despite these structural insights, the functional significance and evolution of *Hbs* in molluscs remain insufficiently clarified in comparison with vertebrates.

To better understand how RB-related *Hbs* arose and evolved in Mollusca, we sequenced and assembled the blood clam *T. granosa* genome at chromosomal level, and generated comprehensive transcriptomic resources. Through comparative genomic, transcriptomic analyses, and functional assays, we investigated the genomic determinants of clam RB phenotype and particularly gained novel insights into the origin and evolution of clam RB-related *Hbs*. The high-quality blood clam genome and comprehensive transcriptomes lay the foundation for deep understanding of the innovative evolution of molluscan RB.

## Results and Discussion

### Genome Sequencing and Characterization of *T. granosa*

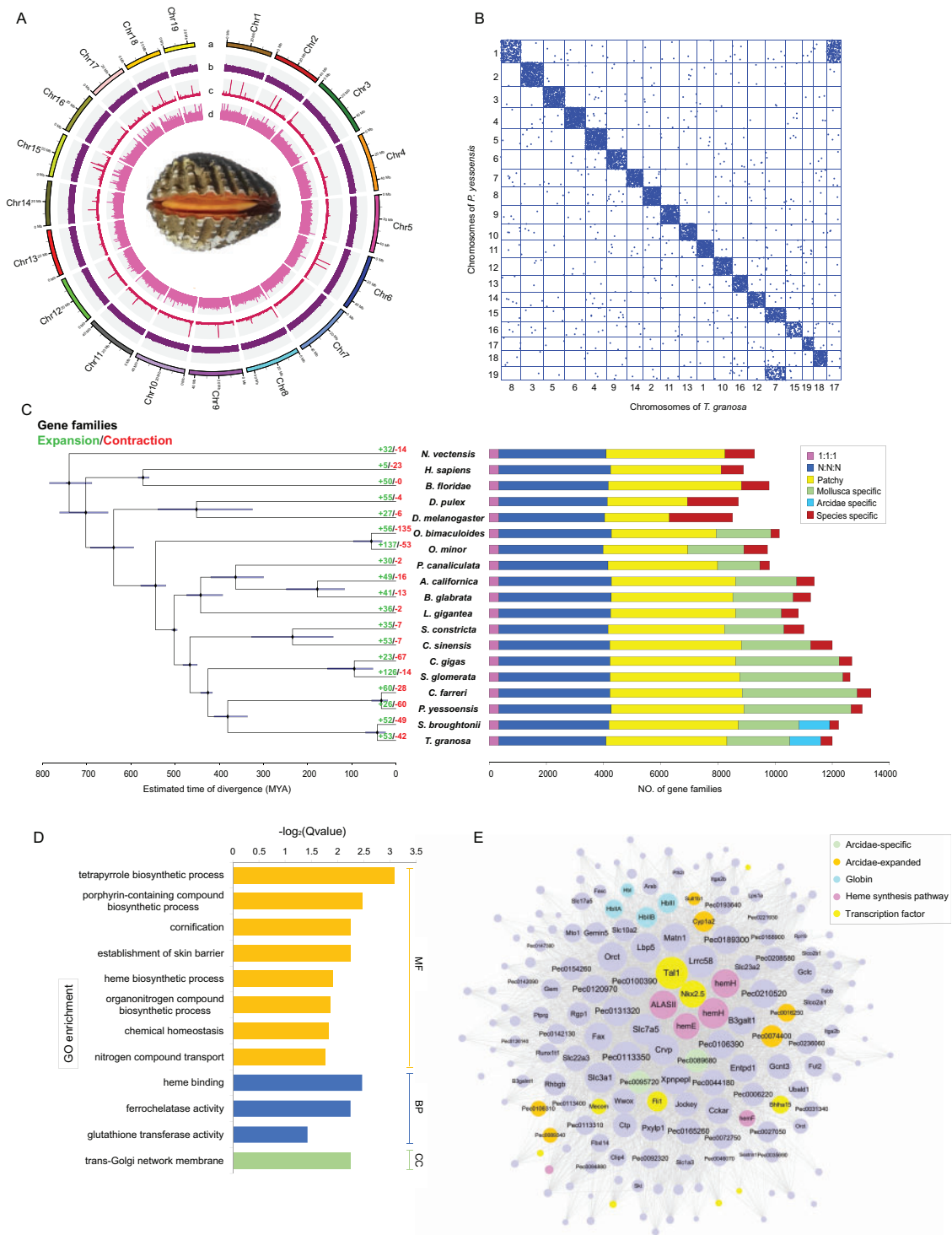
To obtain a high-quality blood clam reference genome, we conducted the deep genome sequencing of a single individual

of *T. granosa*. In total, 101.96 Gb and 131.15 Gb sequencing data were obtained from the PacBio and Illumina platforms respectively, representing  $\sim 287\times$  genome coverage (Supplementary table S1). Based on k-mer analysis, the genome heterozygosity of *T. granosa* was estimated to be 1.3% (Supplementary fig. S1). Following de novo assembly and polishing, a final genome assembly of 812.61 Mb was generated with a contig N50 of 599.92 Kb. The assembly size is comparable to those estimated by k-mer (812.32 Mb) and flow cytometry (828 Mb) analyses (Supplementary fig. S1 and S2). With the aid of 113 Gb Hi-C sequencing data for scaffolding, 97.13% of the obtained contigs were successfully anchored into 19 chromosomes, consistent with the haploid karyotype of *T. granosa* (Afiati 1999; Lu et al. 2008) and *Scapharca broughtonii* (Bai et al. 2019). The accuracy and completeness of the genome assembly was assessed by short reads alignment (fig. 1A). The mapping rate and genome coverage reached 96.82% and 99.73%, respectively. Compared with the reference genome assembly, only 26,437 (0.003%) homozygous single-nucleotide polymorphisms and insertions and deletions (Indels) were identified in the sequencing reads, indicating a low error rate in the assembly (Supplementary fig. S3). In addition, 978 metazoan Benchmarking Universal Single-copy Orthologs (BUSCOs) were present in this genome assembly, wherein 912 were complete (93.3%), 13 partial (1.3%), and 53 missing (5.4%) (Supplementary table S2), confirming the high degree of quality and completeness of the genome assembly.

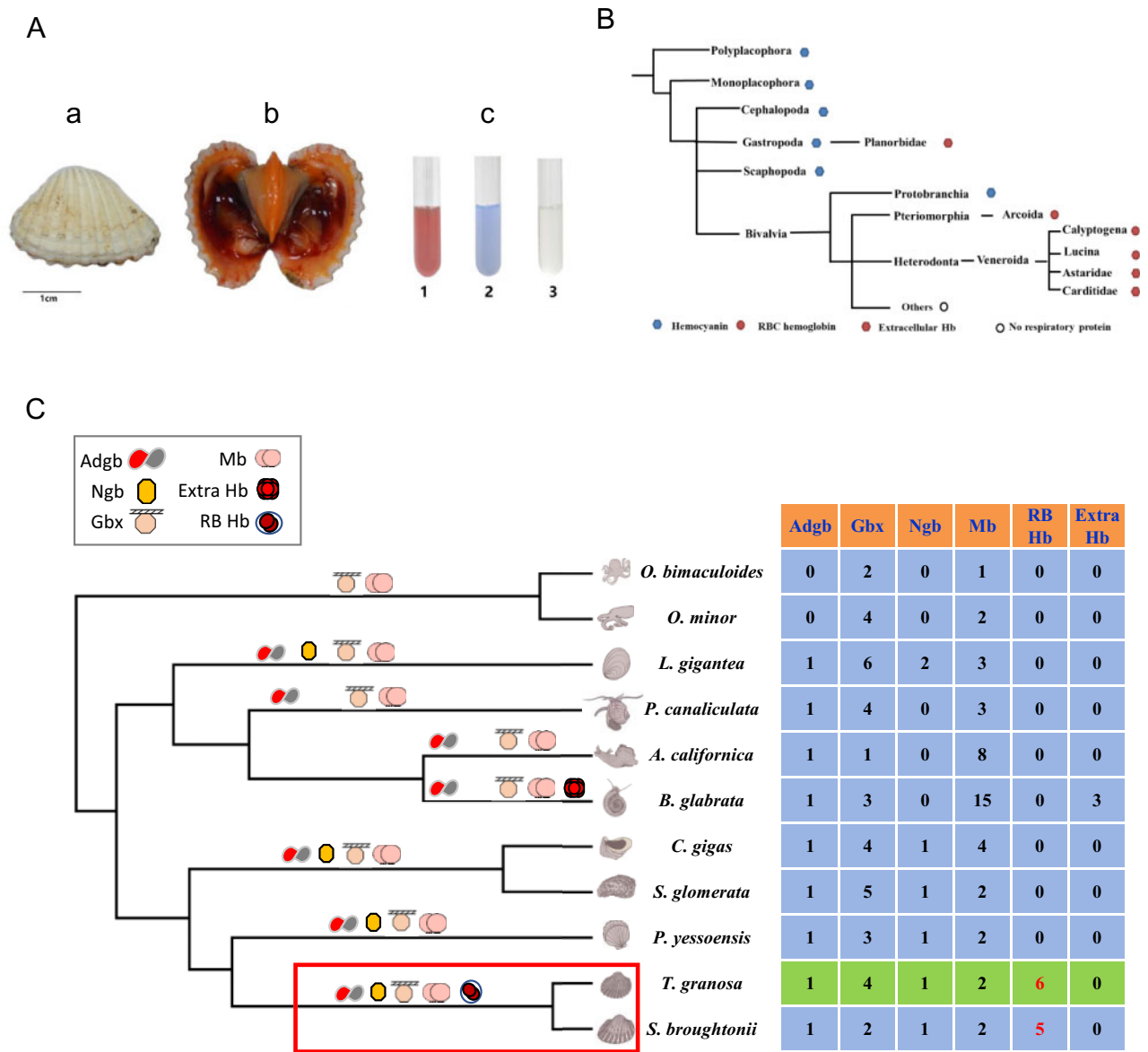
We identified 24,398 protein-coding genes in the *T. granosa* genome, of which 81.18% were annotated based on known proteins in public databases. Repeat content accounted for 53.56% of the assembled genome, which were dominated by DNA transposons (26.72%) (Supplementary fig. S4). Chromosomal macrosynteny analysis revealed high karyotype conservation (despite existence of few chromosome rearrangements) between *T. granosa* and the scallop *Patinopekten yessoensis* (fig. 1B), the latter possessing a highly conserved 19-chromosome karyotype similar to that of bilaterian ancestors (Wang et al. 2017b; Simakov et al. 2020). Phylogenomic analysis suggested that the Arcidae family diverged from its closest Pectinidae family approximately 383 million years ago, whereas the speciation of two blood clams *T. granosa* and *S. broughtonii* occurred in the mid-Eocene epoch, about 44.8 million years ago (fig. 1C).

### Identification of RB-Related Gene Families and Pathways

To investigate genomic components contributing to clam RB phenotype, we first conducted comparative genomic analysis. Through genome comparison with 18 other molluscan and animal groups spanning the animal kingdom, we identified 1,079 and 74 Arcidae-specific and expanded gene families respectively. Notably, the *Hb* gene family exhibited both Arcidae-specific and gene family expansion, supporting *Hbs* as crucial components for clam RB phenotype. Next, to enable comparative transcriptomic analysis, we generated extensive transcriptomic resources of *T. granosa* for hemolymph and other major organs/tissues. Transcriptome



**FIG. 1.** (A): Global genome landscape of *T. granosa*. From outer to inner circles: the 19 chromosomes at the Mb scale (a), GC content (b), depth of coverage of Illumina reads (c), depth of coverage of PacBio reads (d). (B): Oxford dot plot of orthologous genes between *T. granosa* and *P. yessoensis*. The horizontal axis represents the chromosomes of *T. granosa*, and the vertical axis represents the chromosomes of *P. yessoensis*. (C): Phylogenetic tree and number of shared orthologs among *T. granosa* and other animal species. Numbers of gene families undergoing rapid expansion and contraction for each lineage are shown in red and green, respectively. (D): Significantly enriched GO terms of the HREGs unique to blood clam, including tetrapyrrole biosynthesis, heme binding, and ferrochelatase activity regulation, which involved in the conventional RB-related functions. MF: molecular function; BP: biological process; CC: cellular component. (E): Gene co-expression network of the hemolymph-related module. The top 200 genes with the highest intramodular connectivity are chosen for network display. Gene names or IDs of top 100 genes are noted. Node size represents the intramodular connectivity of a given gene.

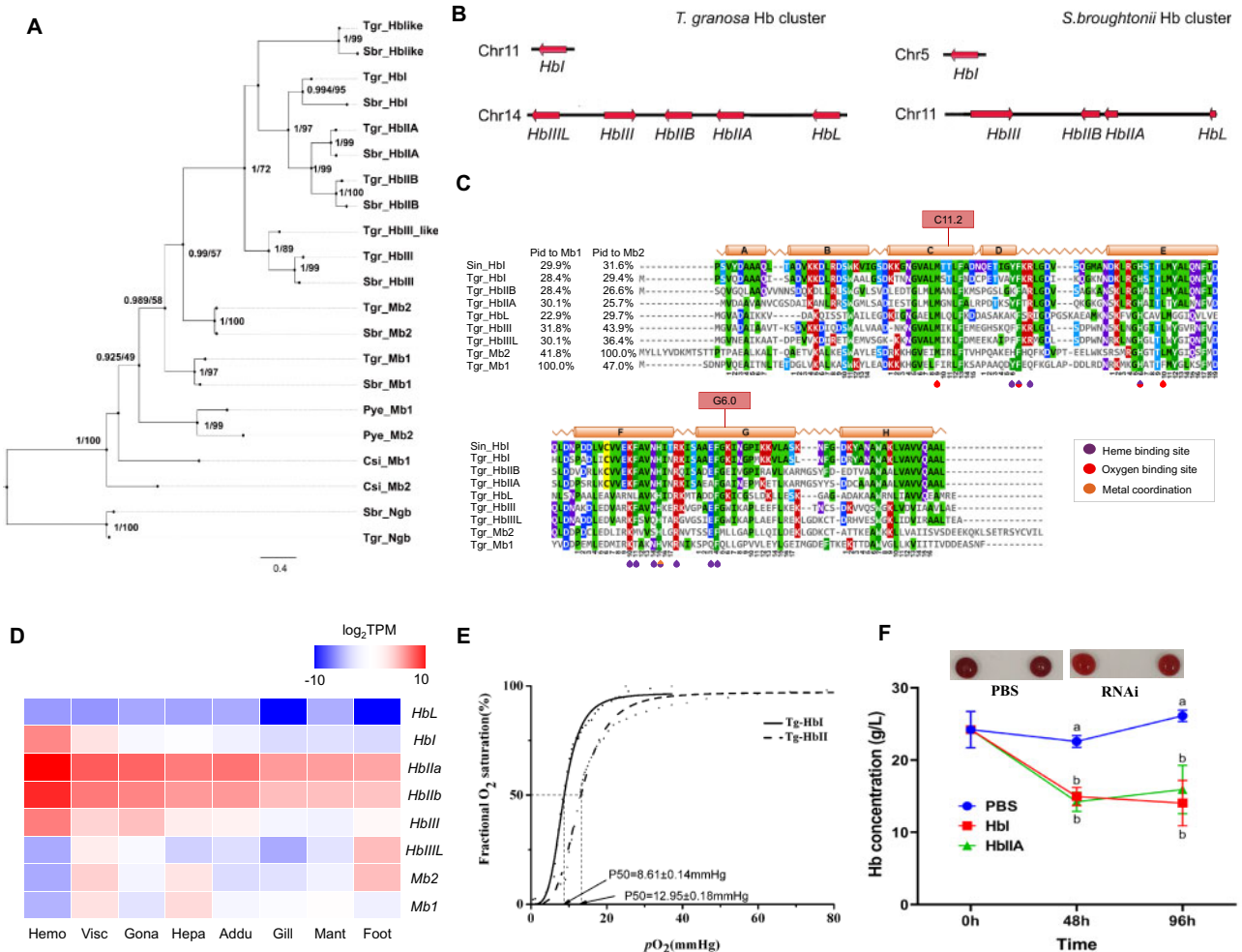


**Fig. 2.** Morphology of blood clam *T. granosa* and diversity of respiratory proteins in *Mollusca*. (A): (a) External morphology of blood clam *T. granosa*. (b) A blood clam with an open shell. (c) The blood traits of three molluscs: 1. *T. granosa* (red, hemoglobin); 2. *Haliotis discus hannai* (blue, hemocyanin), 3. *S. constricta* (colorless, unknown). (B): Diversity and distribution of respiratory proteins in molluscs. (C): Classification and distribution of globin family in molluscs. Abbreviations for globins: androglobin (Adgb), neuroglobin (Ngb), globin X (Gbx), myoglobin (Mb), Red blood cell hemoglobin (RB Hb), extracellular hemoglobin (Extra Hb).

comparison across bivalve molluscs revealed 148 genes that showed highly restricted expression in the hemolymph of blood clam but not in other bivalve molluscs (such as scallop, oyster, and razor clam) without the RB phenotype. Functional enrichment of these genes revealed conventional RB-related functions, including tetrapyrrole biosynthesis, heme binding, and ferrochelatase activity regulation (fig. 1D).

Furthermore, we set out to identify the RB-related gene module by adopting the weighted gene co-expression network analysis (WGCNA) approach. The gene module AM1, which showed high expression in the hemolymph of blood clam and was enriched with the hemolymph-restricted

expressed genes (HREGs), was identified as the RB-related gene module (Supplementary table S4). Notably, Hbs *HbI*, *HbIIA*, *HbIIB*, *HbIII*, and heme biosynthesis-related genes including *ALASII*, *hemH*, and *hemE* were enriched in the AM1 module, suggesting their pivotal roles in the maintenance of clam RB function and physiology (fig. 1E). We then put a focus on transcription factors (TFs), which led to the identification of *Tal1*, *Nkx2.5*, and *Fli1* with high network connectivity in the AM1 module (fig. 1E). Interestingly, these TFs have known crucial roles in the regulation of erythropoiesis in vertebrates (Porcher et al. 1996; Starck et al. 2010; Caprioli et al. 2011), suggesting that these TFs may also be important regulators with similar functional roles in the blood clam.



**FIG. 3.** (A): Maximum likelihood phylogram of two blood clam *Hbs* and bivalve *Mbs*. Numbers at the nodes correspond to support from the aBayes test and 10,000 pseudoreplicates of the ultrafast bootstrap procedure. The tree was rooted using blood clam *Ngbs* as outgroup sequences. (B): Structure of *Hb* gene clusters of *T. granosa*, *S. broughtonii* and *P. yessoensis*. The direction of arrows indicates that of gene transcription. (C): Alignment of *T. granosa* *Hb* and *Mb* sequences. The percentages of identity of *Hb* to *Mb1* and *Mb2* is showed ahead of the alignment. The globin alpha-helical structure is illustrated according to the *HbI* of blood clam *S. inaequivalvis*. The two intron positions shared by all the genes are shown on top of the alignment. Functional residues related to metal coordination, heme and oxygen binding are indicated at the bottom. (D): Expression patterns of *Hb* and *Mb* genes among the 8 different tissues in *T. granosa*. (E): Oxygen binding curves of *T. granosa* *HbI* (black diamond, solid line) and *HbII* (red circle, dashed line).  $Hb-O_2$  affinity was indexed by P50 values (i.e., the  $PO_2$  in mmHg at which *Hb* becomes half-saturated). (F): A faded blood color was observed after interference of *Hb* genes by RNAi. The *Hb* concentrations were verified with decreased levels. Vertical bars are reported as means  $\pm$  standard error. Significant differences between group PBS and group interference (*HbI* and *HbIIA*) are indicated with the different alphabet for  $P < 0.05$ .

**Globin Gene Family and Clam Hb Evolution**

Globins are a superfamily of heme-containing globular proteins that are widely present in animal kingdom (Wajcman et al. 2009). In contrast to most other molluscan groups, blood clams are unusual in their possession of red-bloodedness phenotype and *Hbs* as respiratory proteins (fig. 2A and B). To gain a better understanding of clam *Hb* evolution and its relationships with other types of globins in Mollusca, we first investigated globin gene diversity in the genomes of blood clams as well as in other molluscan groups. In blood clams, five known globin types were identified, including *Hb*, myoglobin (*Mb*), globin X (*GbX*), neuroglobin (*Ngb*), and androglobin (*Adgb*). Remarkably, globin type distribution varies greatly among molluscs. Of these, *Mb* and

*GbX* are present in all ten assayed molluscs, while *Adgb* was found lost in the octopus and *Ngb* lost in the octopus and most gastropods. Extracellular *Hb* is present in the freshwater snail *Biomphalaria glabrata*, while RB cell *Hb* (RB *Hb*) is found only in the blood clams *T. granosa* and *S. broughtonii* (fig. 2C).

Thus far, three *T. granosa* *Hb* genes including *HbI*, *HbIIA*, and *HbIIB* have been isolated and previously characterized by our group (Bao et al. 2011,2013,2016). In the present study, three additional *Hb* genes (*HbIII*, *HbIII\_Like*, and *Hb\_Like*) were found in blood clamgenomes (fig. 3A–D). The high-quality genome of *T. granosa* enables to obtain the complete picture of genomic organization of *Hb* genes. Notably, five of the six *Hb* genes (*HbL*, *HbIIA*, *HbIIB*, *HbIII*, and *HbIII\_Like*) co-localized together (on the Chr14), featuring a vertebrate-like

Hb cluster. A similar Hb cluster in *S. broughtonii* was also observed with conserved microsynteny (fig. 3B). It suggested that the evolutionary origin of a variety of RB-related Hb genes in blood clams may be driven by tandem gene duplication events.

It has been postulated that extracellular Hb of the snail *B. glabrata* evolved from Mb by a single gene duplication event (Lieb et al. 2006). We were thus motivated to investigate whether blood clam Hb followed a similar evolutionary path. Supporting this view, phylogenetic analysis suggested that blood clam Hbs were more closely related to Mb than other globins (fig. 3A). Notably, the clade of HbIII and HbIII\_Like showed shorter branch lengths and more sequence similarity with Mb2 than other Hbs (fig. 3A and C). HbIII\_Like even showed Mb-characteristic transcriptional preference in muscle-dominant foot organ but not in hemolymph (fig. 3D), suggesting HbIII\_Like likely resemble the ancestral Hb gene under initial transitional state from Mb to Hb.

Transcriptional profiles of *T. granosa* revealed that HbI, HbIIA, HbIIB, and HbIII distributed mainly in hemolymph and thus likely to be the major functional Hbs (fig. 3D). Protein domain analysis suggested that blood clam Hbs well preserved typical functional residues that are crucial for heme and oxygen binding (fig. 3B). Our previous study (Bao et al. 2016), together with the additional biochemical assays of *S. broughtonii* presented in this study (Supplementary fig. S5) supported that homodimeric HbI and heterotetrameric HbII (composed of HbIIA and HbIIB) should be general features in blood clams. Intriguingly, compared with HbII, HbI of *T. granosa* showed much higher restricted expression in hemolymph and also stronger oxygen affinity, peroxidase and antibacterial activity (fig. 3E; Supplementary fig. S5), indicating the potential functional divergence of these Hbs. To further examine the role of Hb genes in red-bloodedness, we conducted RNA interference (RNAi) experiments of HbI and HbII genes of *T. granosa* and observed faded blood color with significantly decreased Hb concentrations after 96-h gene knockdown (fig. 3F; Supplementary fig. S6), for the first time, functionally demonstrating that Hb genes are indeed responsible for hemolymph coloration in blood clam.

### Evolution of Hb Genes in Molluscs and Vertebrates

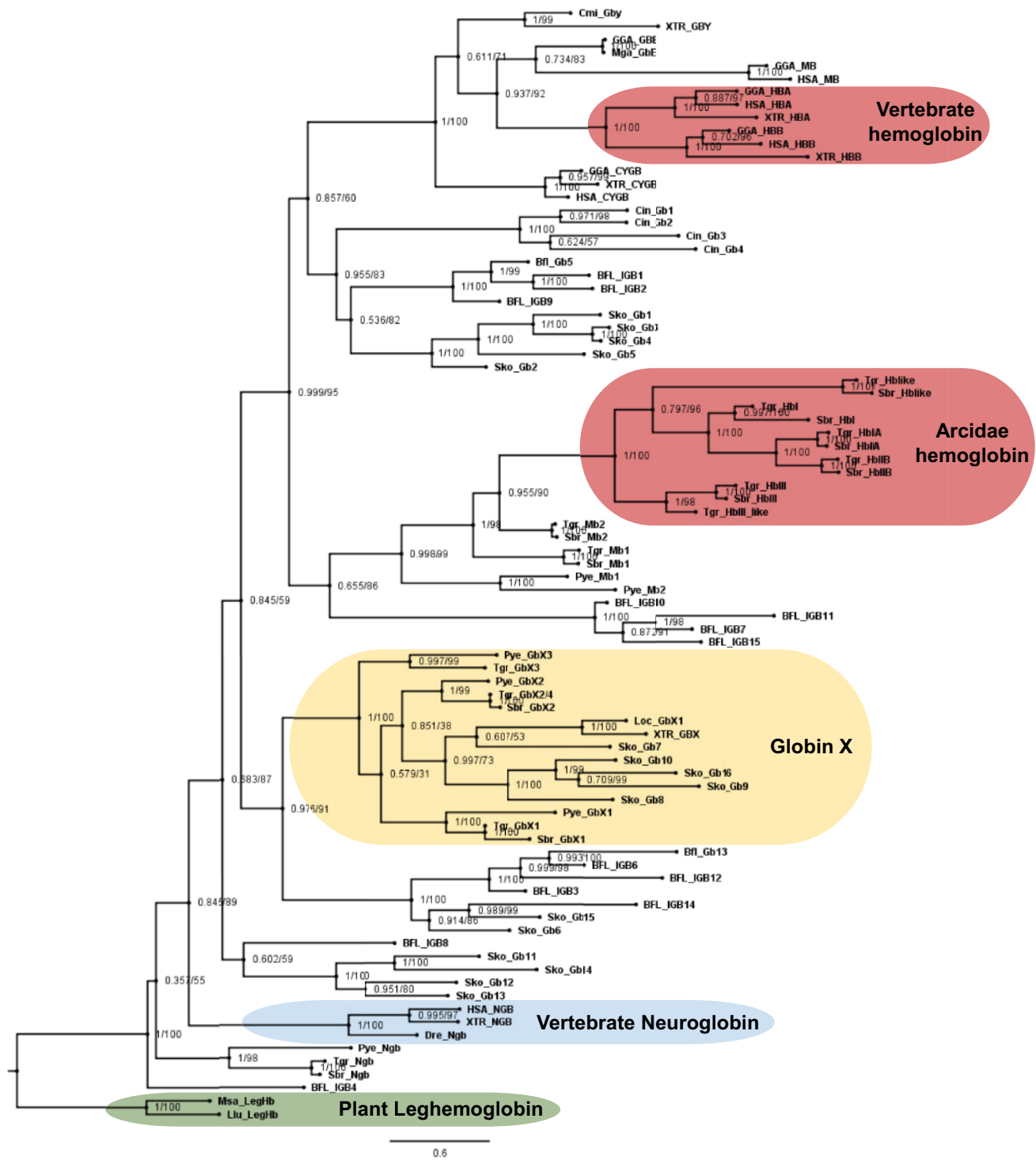
Arcidae and mammalian Hbs have similar structures and functions (Royer et al. 1990; Supplementary figs. S7 and S8), and had been previously regarded as a classic case of convergent evolution (Kuriyan et al. 1991). Decker et al. (2014) previously reported that Mb showed higher oxygen affinity than Hb in the vesicomyid clam *Laubiericoncha chuni*. Similar phenomena have also been found in vertebrates and some other invertebrates with both globin types, indicating a functional convergence of vertebrate and invertebrate Hb evolution that can facilitate oxygen transfer from Hb to Mb and oxygen storage in muscle (Wittenberg 1970; Decker et al. 2014). Despite the convergent evolution of blood clam and vertebrate Hb genes, it remains unclear whether the progenitors of their Hb gene lineages are evolutionarily independent or not. The deuterostome globins have been clustered into four major clades that diverged before the split of protostomes and

deuterostomes in previous studies, including *Ngb*, *GbX*, the vertebrate-specific globins, and globin X-like globins (*GbXL*) that are structurally and phylogenetically similar to *GbX* (Hoffmann et al. 2012; Hoogewijs et al. 2012; Prothmann et al. 2020). Consistent with previous studies, our phylogenetic analysis recovered the monophyly of the groups defined by *GbX*, *GbXL*, and vertebrate-specific globins, placing vertebrate *Ngb* and their putative orthologs from mollusks at the deepest diverging lineage. The molluscan *GbX* genes formed separate monophyletic groups with the corresponding globins from vertebrates and invertebrate deuterostomes, indicating that their evolutionary origin is prior to the split of deuterostomes and protostomes. Intriguingly, the blood clam Hb and Mb genes were grouped into a single large clade that also included vertebrate-specific Hb and Mb genes and their deuterostome homologs (fig. 4) suggesting the intimate evolutionary relationship of molluscan and vertebrate Hbs and Mbs.

We further evaluated the chromosomal evolution of Hb gene clusters in blood clams and vertebrates based on macrosynteny and microsynteny analyses. Through macrosynteny analysis, we found that despite the existence of extensive genomic rearrangements, Hb-cluster adjacent regions showed recognizable syntenic similarities between blood clams and vertebrates (fig. 5A). Interestingly, such correspondence is more prominent to vertebrate  $\beta$ -cluster rather than  $\alpha$ -cluster. This is consistent with previous finding of higher conserved synteny of  $\beta$ -cluster than  $\alpha$ -cluster (Ebner et al. 2010). The syntenic relationships of Hb-flanking genes were detailly checked among blood clams, chicken, and human. To our surprise, the Hb cluster in blood clams shared remarkable conserved microsynteny with the *HB $\beta$*  clusters of chicken and human (fig. 5B). Among 17 conserved flanking genes, six have been recently identified as ancient genes in the ancestral chordate linkage group F (CLG-F) (Simakov et al. 2020). Since the Hb-bearing Chr14 of *T. granosa* exhibited notable correspondence with CLG-F (Supplementary fig. S9) and the *HB $\beta$*  clusters in human and chicken were both indicated to be mainly descended from CLG-F (Simakov et al. 2020), it can be speculated that there might exist an ancient Mb-like globin gene in the bilaterian ancestor which gave rise to the convergent evolution of Hbs in vertebrates and blood clams. The Mb genes seem to be a preferential substrate for the evolution of Hbs, likely because of its oxygen-sensing and binding ability, an ancestral function that can probably be traced back to microorganisms (Hou et al. 2000).

### Heme Synthesis in Relation with Clam RB

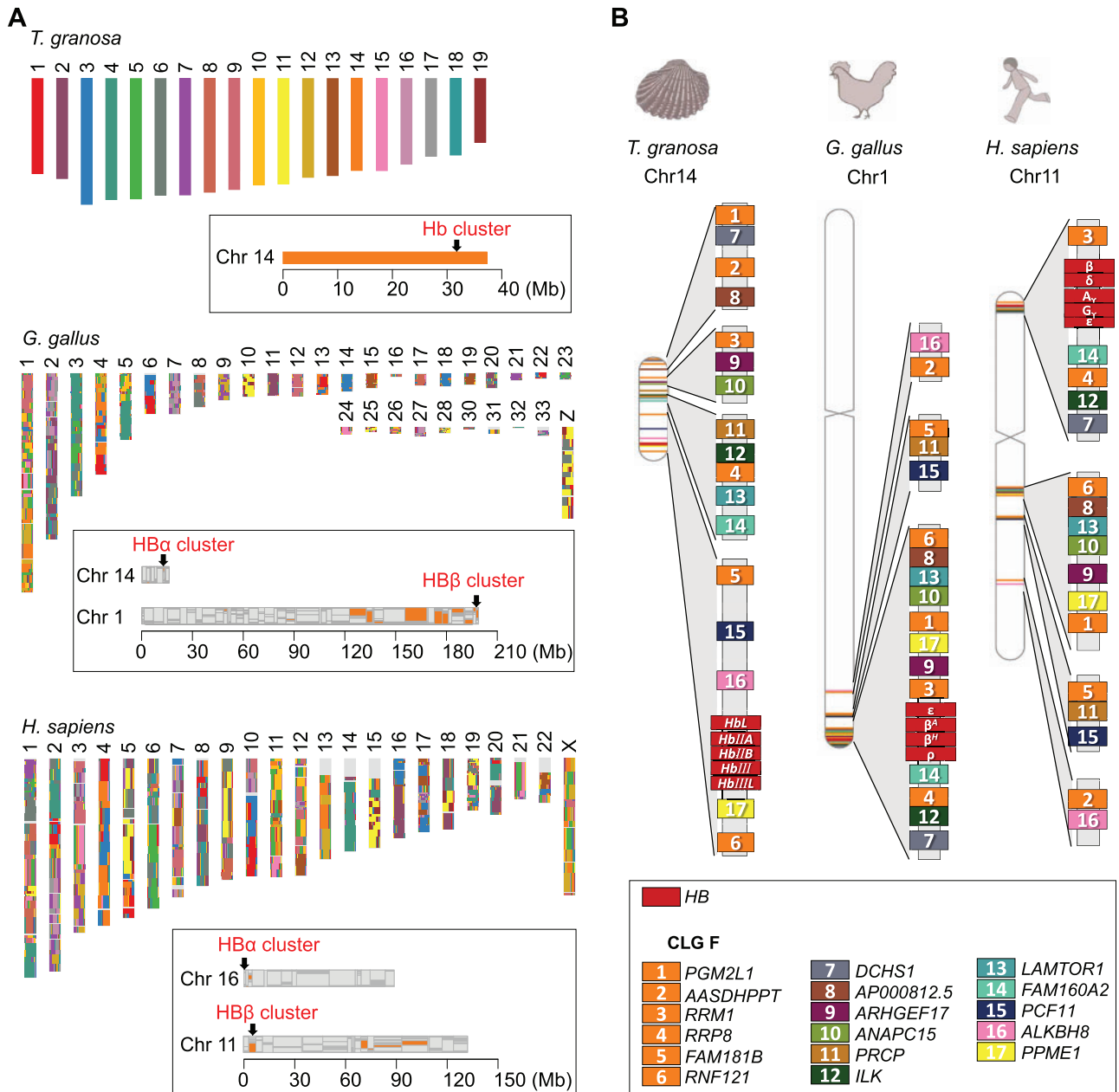
Heme commonly occurs in biological systems as an iron-porphyrin complex, most notably as an essential component of Hb, the characteristic red pigment in blood. Moreover, heme is also integral to a variety of biologically important hemoproteins such as other globins, cytochromes, catalase, heme peroxidases, and endothelial nitric oxide synthase (Paoli et al. 2002). Heme shuttles electrons between proteins as in mitochondrial respiration or electron transport, and stores O<sub>2</sub> in globins (Poulos 2014). As in the human heme synthesis pathway, all key enzymes and genes were identified in the



**Fig. 4.** Maximum likelihood phylogenetic reconstruction reveals possible orthologous relationships among globin genes from representative mollusks and deuterostome taxa. Numbers on the nodes represent support from the aBayes test and 10,000 pseudoreplicates of the ultrafast bootstrap procedure. The tree was rooted using plant leghemoglobins as outgroup sequences. Abbreviations of species, *Callorhynchus milii* (Cmi), *Xenopus tropicalis* (Xtr), *Gallus gallus* (Gga), *Meleagris gallopavo* (Mga), *H. sapiens* (Hsa), *Ciona intestinalis* (Cin), *B. floridae* (Bfl), *Lepisosteus oculatus* (Loc), *Saccoglossus kowalevskii* (Sko), *T. granosa* (Tgr), *S. broughtonii* (Sbr), *P. yessoensis* (Pye), *Medicago sativa* (Msa), *Lupinus luteus* (Llu).

*T. granosa* genome (fig. 6A and B). These heme synthesis-related genes showed more preferential expression in the hemolymph of blood clam than other bivalve molluscs without RB phenotype (fig. 6C). Of all these genes, aminolevulinic acid synthase (ALAS) is of particular importance, as it

catalyzes the first and rate-limiting step in the process of heme synthesis. In contrast to most non-red-blood bivalves that have only one ALAS gene (fig. 6A), blood clam has two ALAS genes that likely originated via gene duplication (fig. 6B) and may ensure the production of heme in RB. Two ALAS



**FIG. 5.** (A): Macro-synteny of the 19 *T. granosa* chromosomes to contemporary chicken and human genomes. The conserved syntenic blocks are shown by the local fraction of genes from each *T. granosa* chromosomes. The *Hb $\beta$*  clusters in chicken and human revealed a higher level of conservation compared with *Hb $\alpha$* . (B): Micro-synteny of conserved *Hb* cluster flanking genes among blood clam, chicken, and human. Genes in the ancestral chordate linkage group F (CLG-F) was indicated in orange.

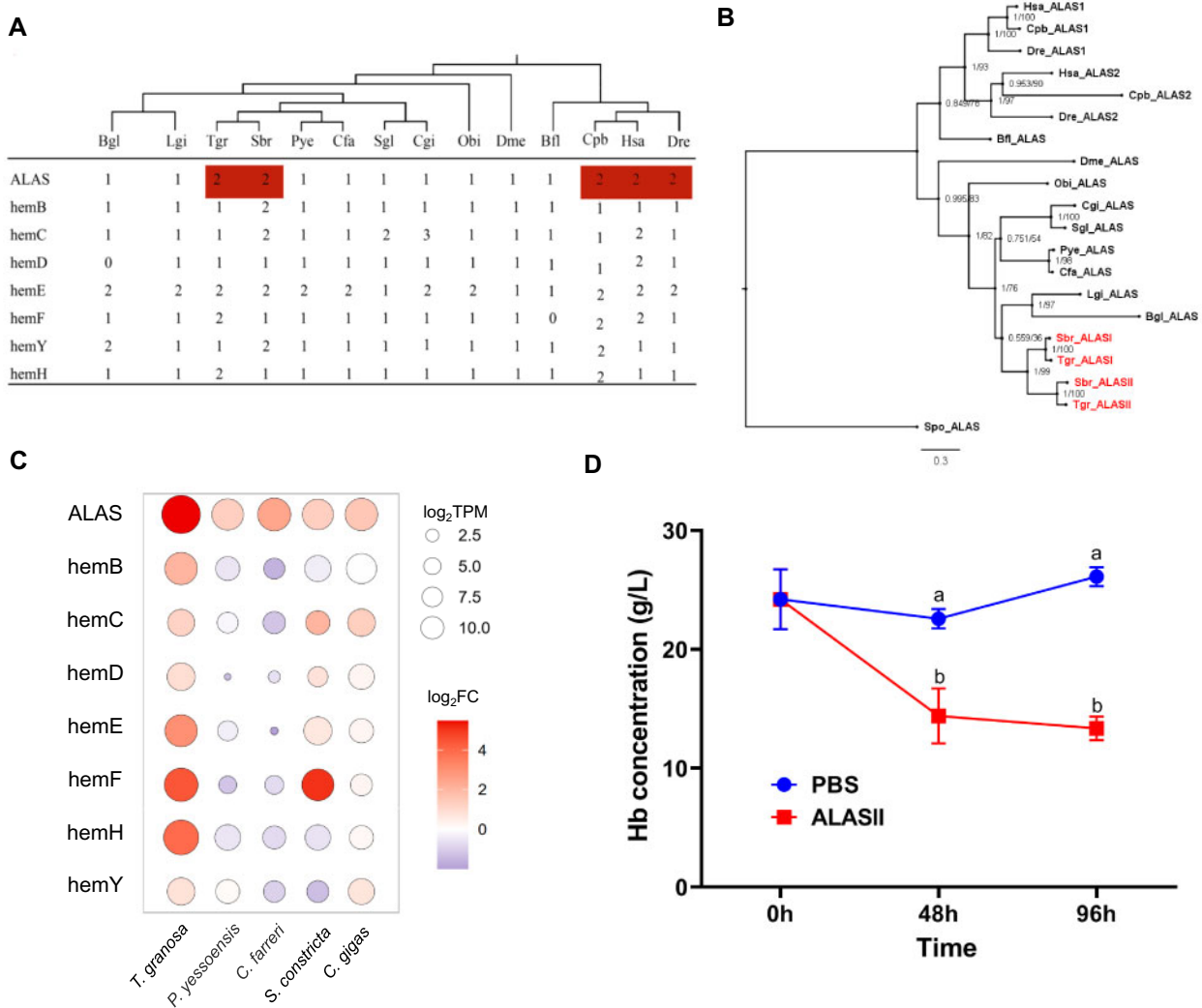
genes have also been identified in vertebrate RB cells (fig. 6A and B), with human *ALAS1* ubiquitously expressed throughout the body and *ALAS2* solely expressed in RB precursor cells (Ajioka et al. 2006). The erythroid-specific *ALAS* is typically absent in molluscs and cephalochordate amphioxus, which have no RB-related Hb (Ebner et al. 2010). Notably, *ALASII* showed the highest expressional specificity in the hemolymph of *T. granosa* over other bivalves without red-blood phenotype (fig. 6C), possibly illustrating functional specificity of *ALASII* in RB heme synthesis, reminiscent of the function of *ALAS2* in human. The vital role of *ALASII* in the RB phenotype of *T. granosa* was also supported by RNAi-based gene

knockdown experiment, where significantly decreased Hb concentrations were observed after double-stranded RNA (dsRNA) injection (fig. 6D; Supplementary fig. S6).

## Conclusion

Our study provides the first high-quality chromosome-level reference genome and comprehensive transcriptomes for the blood clam *T. granosa*. Through an integrated genomic, evolutionary, and functional analyses, we revealed that blood clam *Hbs* were crucial components for the red-bloodedness phenotype, as well as a convergence of both RB-related *Hb*





**FIG. 6.** (A): Numbers of enzyme-encoding genes within the heme biosynthetic pathway in the blood clams *T. granosa* (Tgr) and *S. broughtonii* (Sbr) and representative mollusks and deuterostomes invertebrates and vertebrates. Bgl, *B. glabrata*; Lgi, *L. gigantea*; Pye, *P. yessoensis*; Cfa, *C. farreri*; Sgl, *Saccostrea glomerata*; Cgi, *C. gigas*; Obi, *O. bimaculoides*; Dme, *Drosophila melanogaster*; Bfl, *Branchiostoma floridae*; Gga, *Gallus gallus*; Hsa, *Homo sapiens*; Dre, *Danio rerio*; Cpb, *Chrysemys picta bellii*. (B): Maximum likelihood phylogenetic tree of ALAS genes. Numbers above the nodes correspond to support from the aBayes test and 10,000 pseudoreplicates of the ultrafast bootstrap procedure. The tree was rooted using the ALAS of yeast *Schizosaccharomyces pombe* (Spo) as outgroup sequence (UniProtKB accession O14092). (C): Relative expression levels of ALAS and other heme biosynthetic key genes in the hemolymph compared with other tissues/organs in *T. granosa* and four other bivalves. Note that ALASII exhibits the strongest hemolymph-specific expression in *T. granosa* and is selected for illustration. (D): Hb concentrations in hemolymph of *T. granosa* after interference of ALASII by RNAi. Vertical bars are reported as means  $\pm$  standard error. Statistical significance is indicated with different letters.

and heme functions in vertebrates and blood clams. These permit critical glimpses into the origin and evolution of red-bloodedness in blood clams as an intricate process driven by gene duplication and functional divergence of *Hb* genes. Our study lays a fundamental foundation for deep understanding of the origin and evolutionary dynamics of invertebrate RB phenotype, with important implications for RB evolution in vertebrates.

## Materials and Methods

### Samples Preparation and Sequencing

A healthy individual *T. granosa* was collected from a brood stock at the Genetic Breeding Research Center of Zhejiang

Wanli University, China, for genome sequencing. Genomic DNA was extracted from the muscle by the phenol–chloroform extraction method (Harlow and Lane 1988). A paired-end Illumina library with an insert size of 350 bp was prepared with the Illumina Genomic DNA sample preparation kit, and sequenced on an Illumina HiSeq X Ten system. Genomic DNA from the same sample was used to construct a library for PacBio sequencing. Size-selection was performed to enrich DNA fragments longer than 10 kb for sequencing on a PacBio Sequel Single-molecule Real-time (SMRT) platform. The adductor muscle of a blood clam from the same breeding family was collected for Hi-C library construction by following a procedure as described previously (Burton et al. 2013). Briefly, tissue specimen was fixed with 1% formaldehyde.

Obtained genomic DNA was cross-linked, digested by the restriction enzyme *Mbol*, labeled on biotinylated residues, and end repaired. The library was subsequently sequenced on an Illumina NovaSeq platform.

Eight adult tissues/organs (including hemocytes, gonad, foot, mantle, visceral mass, gill, hepatopancreas, and adductor muscle) were collected from three adult individuals. Artificial fertilization and larval culture were performed according to laboratory protocols described in previous studies (Zhang et al. 2012). All samples were stored at  $-80^{\circ}\text{C}$  after being flash-frozen in liquid nitrogen. Total mRNA was extracted with TRIzol reagent (OMEGA, USA) according to the manufacturer's instructions. All of the 24 libraries were sequenced on an Illumina HiSeq X Ten system. Raw reads were first filtered by removing those containing undetermined bases ("N") or excessive numbers of low-quality positions ( $>10$  positions with quality scores  $<10$ ). Then, high-quality reads were mapped onto the *T. granosa* genome by using Tophat (v2.0.9) with the parameters of "`-p 10 -N 3 -read-edit-dist 3 -m 1 -r 0 -coverage-search -microexon-search`". The expression levels of all genes were normalized by calculating the transcripts per kilobase of exon model per million mapped reads.

### Genome Survey, Assembly, and Scaffolding

Genome size, heterozygosity, and repeat content of *T. granosa* were estimated by using k-mer analysis. Briefly, Illumina short reads were first trimmed to remove adaptors and reads with  $>10\%$  ambiguous or  $>20\%$  low-quality bases by using Trimmomatic (Bolger et al. 2014). Distribution of 17-mer frequency was estimated by using clean reads by Jellyfish (Marçais and Kingsford 2011). Genome size was estimated according to the following formula: genome size = k-mer number/peak depth (Varshney et al. 2012). The gills of *T. granosa* and *Crassostrea gigas* were cut into pieces. The DNA was stained by Propidium Iodide (PI). The genome size of *T. granosa* was estimated using *C. gigas* as reference by the flow cytometry BD FACSAria II.

Contig assembly of *T. granosa* was conducted by using CANU (version 1.4). In the correction step, longer seed reads were selected with the parameters "`genomeSize = 800m`" and "`corOutCoverage = 80`", while overlapping raw reads were detected by the mhap overlacer with the option "`corMhapSensitivity = normal`". Reads error correction was performed with the parameter "`correctedErrorRate = 0.065`". Unsupported bases and hairpin adaptors were trimmed with default parameters to acquire the longest supported range. Draft contigs were assembled by using the longest 80 coverage-trimmed reads to mitigate the error rate and maintain separation of haplotypes. The assembly was subsequently polished by Arrow (SMRT Link version 5.1.0) and Pilon (Walker et al. 2014) with PacBio long reads and Illumina short reads, respectively. Redundancy of heterozygous contigs was assessed and removed by using the Purge Haplotigs pipeline (Roach et al. 2018). Integrity and accuracy of the genome assembly were evaluated at the single-base level. Illumina short-insert library reads were mapped onto the contigs by using Burrows-Wheeler Aligner (BWA) and genetic variants

were called out by SAMtools (version 1.10). Genome completeness was also assessed by using BUSCOs (version 3) analysis (Waterhouse et al. 2018).

Hi-C sequencing was performed for chromosome-level scaffolding. Hi-C sequencing reads were first truncated at junctions and aligned to polished contigs by using BWA (version 0.7.17) with default parameters. Only uniquely aligned reads with a mapping quality of  $>20$  were further processed. After filtering invalid interactions in HiC-Pro (version 2.8.0) (Servant et al. 2015), valid read pairs were utilized to evaluate the interaction strength among whole genome contigs. Lachesis (version 2e27abb) was used to cluster and anchor contigs into 19 pseudochromosomes by using an agglomerative hierarchical clustering method (Burton et al. 2013).

### Genome Annotation

Repetitive sequences in the genome assembly were identified and masked prior to gene annotation. RepeatScout (version 1.0.5) and Repeat Modeler version 1.0.11 (<http://www.repeat-masker.org/>, last accessed June 18, 2020) were used for de novo identification of repeat families. Full-length long terminal repeat (LTR) retrotransposons were identified by using LTR-finder (version 1.0.7) (Xu and Wang 2007) with the parameters "`-E -C`". Tandem Repeats Finder (version 4.09) (Benson 1999) was used to screen for tandem repeats with the parameters "`match = 2, mismatching penalty = 7, indel penalty = 7, match probability = 80, indel probability = 10, minimum alignment score = 50, maximum period size = 500`". Predicted repetitive sequences along with the RepBase database (Bao et al. 2015) were used for homology-based searches by using Repeatmasker (version 4.0.9) with the parameters "`-a -nolow -no_is -norna -parallel 32 -small -xsmall -poly -e ncbi -pvalue 0.0001`" (Chen 2004).

Protein-coding genes were annotated by incorporating de novo prediction, homology-based searches, and transcriptome-assisted methods. Protein sequences of Yesso scallop (*P. yessoensis*), freshwater snail (*B. glabrata*), Pacific oyster (*C. gigas*), eastern oyster (*Crassostrea virginica*), and California sea hare (*Aplysia californica*) were downloaded from NCBI and aligned onto the genome assembly by using TBLASTN with the parameters "`-evalue 1e-5`". Gene structures were predicted with GeMoMa (version 1.6.4). Illumina RNA-seq reads of the nine tissues were aligned to the genome assembly using Tophat (version 2.1.1) (Trapnell et al. 2009). Cufflinks (version 2.1.0) (Trapnell et al. 2010) was used to generate gene models with the parameter "`-multi-read-correct`". In addition, the clean RNA-seq data from all samples were pooled and assembled by using Trinity (version 2.0.2) (Grabherr et al. 2011), followed by the alignment of assembled sequences against the genome by using Program to Assemble Spliced Alignment (Haas et al. 2008). The resultant effective alignments were clustered based on genome mapping locations and assembled into gene structures. De novo gene prediction packages Augustus (version 3.3.2) (Stanke et al. 2006) and Genscan (version 3.1) (Burge and Karlin 1997) were used to predict genes with repeat-masked genome sequences by default settings. All evidence thus generated from the gene model was integrated by using Maker (version 2.31.10)

(Cantarel et al. 2008). Next, functional annotation was performed by homology comparisons of the predicted protein sequences against public databases including KEGG, SwissProt, TrEMBL, TF, KOG, and NCBI-NR by using BLASTP with an *E*-value threshold of  $1e^{-5}$ . InterProScan (version 4.8) (Jones et al. 2014) was also used to identify motifs and domains by searching through the Pfam, InterPro, and Gene Ontology (GO) databases.

Noncoding RNA genes including miRNAs, rRNAs, snRNAs, and tRNAs were annotated in the *T. granosa* genome. tRNAs were predicted by tRNAscan-SE 1.4 (Lowe and Chan 2016) with parameters for eukaryotes. Screens for miRNAs and snRNAs were done by using INFERNAL 1.1.2 against the Rfam database (version 14.1) (Kalvari et al. 2018) with default parameters.

### Genome Phylogeny and Gene Family Analysis

Full protein sets of 18 species were retrieved from NCBI database for gene family analysis, including blood clam (*S. broughtonii*), scallops (*P. yessoensis*, *Chlamys farreri*), oysters (*C. gigas*, *Saccostrea glomerata*), Venus clam (*Cyclina sinensis*), razor clam (*Sinonovacula constricta*), snails (*B. glabrata*, *Pomacea canaliculata*), California sea hare (*A. californica*), owl limpet (*Lottia gigantea*), octopuses (*Octopus bimaculoides*, *Octopus minor*), fruit fly (*Drosophila melanogaster*), water flea (*Daphnia pulex*), human (*Homo sapiens*), Florida lancelet (*Branchiostoma floridae*), and starlet sea anemone (*Nematostella vectensis*). The longest protein sequence was selected as representative, when a gene possesses multiple splicing isoforms. Gene family clusters from all the 19 species were assigned by using Orthofinder (version 2.3.3) (Li et al. 2003).

Phylogeny of *T. granosa* was inferred by using the shared single-copy orthologs. Amino acid sequences of these genes were first aligned by Mafft (version 7.221) with “E-INS-I” iterative refinement (Edgar 2004). The resultant alignments were subsequently concatenated and used for maximum-likelihood (ML) phylogenetic inference with IQ-TREE v1.6.12 (Nguyen et al. 2015). An optimal substitution model was automatically selected, whose robustness was assessed by using the bootstrap method with 1,000 replicates. MCMCtree tool in PAML package (Yang 2007) was used to estimate the dates of divergence calibrated by four reference divergence times obtained from TimeTree database (Kumar et al. 2017). CAFE (version 3) (De Bie et al. 2006) was used to analyze expansion and contraction of gene families with separated birth ( $\lambda$ ) and death ( $\mu$ ) rates under a *P*value threshold of 0.01.

The *Hb* protein structures of *T. granosa* were predicted by PHYRE2 (Kelley et al. 2015). We aligned the amino acid sequences and protein structures of *Hb* genes of *T. granosa* and *H. sapiens* using Mafft (version 7.221) and TM-align (Zhang and Skolnick 2005), respectively.

### Phylogenetic Analysis and Classification of Molluscan Globins

Globin genes were identified in the *T. granosa* genome by using HHsearch with an *E*-value threshold of  $1e^{-5}$  against a globin hmm file (PF00042) from pfam database, and were

further confirmed by comparing to the Conserved Domains Database (<http://www.ncbi.nlm.nih.gov/cdd>, last accessed September 26, 2020) and SMART (<http://smart.embl-heidelberg.de/>, last accessed September 26, 2020). Genes were then classified based on BLAST results, molecular phylogeny and manual inspection of conserved residues. The same approach was also applied to identify globin genes in *S. broughtonii*, *P. yessoensis*, *C. gigas*, *C. virginica*, *B. glabrata*, *P. canaliculata*, *A. californica*, *L. gigantea*, *O. bimaculoides*, and *O. minor*. Validated sequences were aligned by CLUSTAL 2.1 and used to construct an updated shellfish-specific globin hmm file. A second-round of searches was performed with the updated hmm file. Globin sequences from representative deuterostomes were also included in the analysis, including amphioxus *B. floridae*, acorn worm *Saccoglossus kowalevskii*, sea squirt *Ciona intestinalis*, Australian ghostshark *Callorhynchus milii*, spotted gar *Lepisosteus oculatus*, frog *Xenopus tropicalis*, chicken *Gallus gallus*, and human *H. sapiens*. Plant leghemoglobin from *Medicago sativa* and *Lupinus luteus* were used as an outgroup (Supplementary table S5). Phylogenetic relationships of globin amino acid sequences were estimated using maximum likelihood and Bayesian analyses. ML analyses were run using IQ-Tree (v1.6.12). Branch supports were evaluated with the Shimodaira–Hasegawa approximate likelihood-ratio test from Anisimova et al. (2011) and the aBayes tests of 10,000 pseudoreplicates of the ultrafast bootstrap procedure (Hoang et al. 2018). ModelFinder from IQ-Tree was used to select the best-fitting model (LG + R4).

### Synteny Analysis

Conservation of gene macro-synteny among *T. granosa*, *S. broughtonii* and *P. yessoensis* was presented in the form of Oxford dot plot. Each dot in the dot-plot comparison represents an orthologous gene pair derived from the same gene family. A macro-synteny conservation index was calculated as measure of conservation (Simakov et al. 2013; Wang et al. 2017b). To further check the distribution of shared ancestral chromosomal segments among blood clam and vertebrates. The local ancestry across chromosomes of *T. granosa*, *G. gallus*, and *H. sapiens* was identified based on orthologous gene families according to Simakov et al.’s method (2020). Briefly, the corresponding chromosomal ancestry was determined by the fraction of genes in a window of approximate 20 genes. The discontinuity was measured using a sliding window method to subdivide chromosomes to relatively homogeneous ancestry.

### Transcriptome Analysis and Gene co-Expression Network Construction

RNA-seq data from different tissues of *P. yessoensis*, *C. farreri*, *C. gigas*, and *S. constricta* were downloaded from the NCBI (Supplementary table S6). The gene expression levels were calculated following the protocol mentioned above. The HREGs of each species were defined by the software package RNentropy (Zambelli et al. 2018), with the uniform criteria of the corrected global sample specificity test  $P < 0.01$  by the Benjamini–Hochberg method, local sample specificity test

$P < 0.01$  and  $\log_2(\text{fold change}) > 1.5$ . The HREGs that unique to blood clam (i.e. these were not identified as HREGs in the other bivalve molluscs without RB phenotype) were extracted and used for GO enrichment analysis by EnrichPipeline (Chen et al. 2010). All 18,549 genes that expressed in eight tissues/organs of blood clam were used for the gene co-expression network construction by the R package WGCNA (Langfelder and Horvath 2008), with the parameters of  $\text{softPower} = 12$ , minimum module size = 300 and cutting height = 0.99. These were assigned into ten gene modules (AM1-10) that were labeled in different colors. The intramodular connectivity (Kwithin) represents the importance of a gene in one module, which measures the connection strength of a gene to others in the specified module (Langfelder and Horvath 2008). To identify the hemolymph-related modules, the HREGs of blood clam were used to perform the overrepresentation analysis for each module using a hypergeometric test, with  $P$  values adjusted by the Benjamini–Hochberg method for multiple-test correction (Langfelder and Horvath 2008). Cytoscape (Version 3.7.1) was used for visualization of co-expression networks (Shannon et al. 2003).

### Oxygen Binding, Peroxidase Activity, and Antibacterial Activity Assays for Hbs

Individual components and whole *Hb* samples from blood clams *T. granosa* and *S. broughtonii* were stripped of organic phosphate by passing *Hb* solutions through a Sephacryl S100 HR (GE Healthcare; HiPrep 16/60) column equilibrated with 50 mM HEPES buffer (pH 7.2). Purified Hbs were diluted with 0.1 M HEPES buffer (pH 7.2) and were stored at  $-80^\circ\text{C}$  prior to the measurement of  $\text{O}_2$ -equilibrium curves. We measured  $\text{O}_2$ -equilibria of purified *Hb* solutions under optimized conditions ( $13^\circ\text{C}$ , pH 7.2,  $50\ \mu\text{M}$  heme) by using our independently developed instruments, which combine functionalities of a spectrophotometer, oxygen analyzer, and gas mixing system instrument. These three devices operating with absorption at 560 nm (where oxy and deoxy Hbs differ in absorbance) and 570 nm (where oxy and deoxy Hbs are the same in absorbance) allowed monitoring of stepwise changes in mixtures of oxygen and nitrogen prepared by Wösthoff gas mixing pumps. We estimated the values of  $P_{50}$  by fitting experimental  $\text{O}_2$  saturation data to the Hill equation  $Y = PO_2^n / (P_{50}^n + PO_2^n)$  by means of nonlinear regression (where  $Y$  = fractional  $\text{O}_2$  saturation;  $n$  = cooperativity coefficient). Model-fitting was based on a series of equilibration steps between 5% and 95% oxygenation.

Peroxidase activity of *Hb* was measured as described in previous studies (Wang et al. 2017a). Detection time was set at 0.5 min. Minimum inhibitory concentration (MIC) value of *Hb* was determined by using the broth microdilution method according to the National Committee for Clinical Laboratory Standards (NCCLS; 1993). Following inoculation by streaking, bacterial strains to be tested were cultured overnight on Luria-Bertani agar and individual colonies were picked from the plates. Following liquid culture in Mueller-Hinton broth, a bacterial suspension corresponding to an absorbance of 0.5 on a McFarland calibration standard was prepared and further diluted 1,000 times with Mueller-

Hinton broth, followed by seeding into wells ( $20\ \mu\text{L}$  per well) in at least triplicates of a 96-well microtiter plate. Hbs were serially diluted in 50 mM PBS buffer (pH 7.0), and distributed (in a  $100\text{-}\mu\text{L}$  volume) into wells containing the bacterial suspension. Results were compared with a bacteria alone control, a *Hb* alone control and a medium alone control. Plates were incubated for 16–20 h at  $35^\circ\text{C}$  without shaking. Minimum concentrations for Hbs at which no turbidity was detected were regarded as the MICs.

### RNAi and Hb Concentration Measurement

To verify the function of *Hb* genes and *ALASII* gene in *Hb* synthesis, *T. granosa Hbl*, *HbIIA*, *ALASII* genes were knocked down in vivo via dsRNA-mediated RNAi. Healthy blood clams averaging about 30 mm in shell length were obtained from a local seafood market and acclimatized in flat-bottomed glass tanks with aerating for seven days. The temperature of seawater was  $21 \pm 0.5^\circ\text{C}$  and salinity was 24‰ during the experiments. They were fed with chlorella (*Chlorella vulgaris*) powder and changing seawater daily. For the RNAi experiment, the clams were divided into five groups (30 clams in each group). For each individual,  $20\ \mu\text{L}$  ( $20\ \mu\text{M}$ ) of three genes (*Hbl*, *HbIIA*, *ALASII*) specific dsRNA was injected into the adductor muscle (Supplementary table S7). Uninjected clams and 30 clams injected with  $20\ \mu\text{L}$  PBS were used as the time 0 (at 0 h) and control groups, respectively.

Blood samples were collected from four randomly selected clams per group at 48 h and 96 h respectively, and were subsequently centrifuged  $1000 \times g$  for 5 min. The supernatant was discarded and the remaining was immediately stored  $-80^\circ\text{C}$  until RNA isolation. The *Hb* concentration of fresh blood was tested by TU-1810PC UV-spectrophotometer (Purkinje, China) with *Hb* test kit (Real-tech, China). Total RNA was extracted from hemocytes using the EASY spin Plus tissues/cells RNA extraction kit (Aidlab, China) according to manufacturer protocols. The RNA quality and quantity were controlled via NanoReady micro-volume spectrophotometers (LifeReal, China). The PrimeScript™ RT reagent kit (Takara, Japan) was used to purify total RNA and synthesize first strand cDNA. Expression of *Hbl*, *HbIIA*, and *ALASII* in hemocytes at each time point were measured with the LightCycler® 480 II SYBR Green Quantitative Real-Time PCR System (Roche, Switzerland). The specific primers were used amplified for qRT-PCR listed in Supplementary Table S7. It was carried out in a  $10\ \mu\text{L}$  reaction volume containing  $1.0\ \mu\text{L}$  cDNA,  $5.0\ \mu\text{L}$  of  $2 \times$  SYBR Green Master Mix,  $0.5\ \mu\text{L}$  of each primer ( $10\ \text{pmol}/\mu\text{L}$ ), and  $3.0\ \mu\text{L}$  of PCR grade water. The Real-time PCR cycling process was conducted as follows: one cycle of  $95^\circ\text{C}$  for 10 min, amplification for 40 cycles ( $95^\circ\text{C}$ , 10 s;  $60^\circ\text{C}$ , 60 s). All experiments were performed in triplicates by using clam 18S mRNA as an internal control. Specificity of primers was confirmed by the dissociation curve of the amplification products at the end of each PCR. The relative expression level was determined by the Livak ( $2^{-\Delta\Delta\text{CT}}$ ) method. Data were processed with SPSS 24.0 statistical software, and significance among groups was determined by t-test.

## Data Availability

*T. granosa* genome assembly is publicly available at NCBI under the accession number JABXW000000000. Raw reads data from PacBio, Illumina and Hi-C sequencing for genome assembly were deposited at the NCBI Sequence Read Archive (SRA) under the accession numbers of SRR10685239-SRR10685249, SRR10662764 and SRR10662389 under BioProject PRJNA593692. Transcriptomic sequencing raw data files are available at the NCBI SRA under the accession numbers of SRR10713949-SRR10713972 of BioProject PRJNA594182.

## Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

## Acknowledgements

We are deeply grateful to all lab members and collaborators who provided assistance or valuable advice at all stages of this study. We would like to acknowledge grant support from the National Science Foundation of China (31672678 and U1706203), the Key Natural Science Foundation of Zhejiang (LZ20C190001), the National Key Research and Development Program of China (2018YFD0901404 and 2018YFD0901405), Modern Agro-industry Technology Research System (CARS-49), the Demonstration Project for Innovative Development of Marine Economy (NBHY-2017-54), Grant of Sanya Yazhou Bay Science and Technology City (SKJC-KJ-2019KY01), and the Taishan Scholar Project Fund of Shandong Province of China (to S. W.).

## References

- Afiati N. 1999. Cytoplasmic granules in the red blood cells and the karyotype of rounded ecomorph of *Anadara granosa* (L.) (Bivalvia: arcaidae) from Central Java. *Indonesia. Majalah Ilmu Kelautan* 14:51–59.
- Ajioka RS, Phillips JD, Kushner JP. 2006. Biosynthesis of heme in mammals. *Biochim Biophys Acta* 1763(7):723–736.
- Anisimova M, Gil M, Dufayard JF, Dessimoz C, Gascuel O. 2011. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Syst Biol* 60(5):685–699.
- Baalsrud HT, Voje KL, Tørresen OK, Solbakken MH, Matschiner M, Malmstrøm M, Hanel R, Salzburger W, Jakobsen KS, Jentoft S. 2017. Evolution of hemoglobin genes in codfishes influenced by ocean depth. *Sci Rep* 7(1):7956.
- Bai CM, Xin LS, Rosani U, Wu B, Wang QC, Duan XK, Liu ZH, Wang CM. 2019. Chromosomal-level assembly of the blood clam, *Scapharca (Anadara) broughtonii*, using long sequence reads and Hi-C. *GigaScience* 8(7):giz067.
- Bao W, Kojima KK, Kohany O. 2015. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* 6:11.
- Bao YB, Wang JJ, Li CH, Li PF, Wang SF, Lin ZH. 2016. A preliminary study on the antibacterial mechanism of *Tegillarca granosa* hemoglobin by derived peptides and peroxidase activity. *Fish Shellfish Immunol* 51:9–16.
- Bao YB, Wang Q, Lin ZH. 2011. Hemoglobin of the bloody clam *Tegillarca granosa* (Tg-Hb1) is involved in the immune response against bacterial infection. *Fish Shellfish Immunol* 31(4):517–523.
- Bao YB, Wang Q, Guo XM, Lin ZH. 2013. Structure and immune expression analysis of hemoglobin genes from the blood clam *Tegillarca granosa*. *Genet Mol Res* 12(3):3110–3123.
- Benavides GA, Squadrito GL, Mills RW, Patel HD, Isbell TS, Patel RP, Darley-Usmar VM, Doeller JE, Kraus DW. 2007. Hydrogen sulfide mediates the vasoactivity of garlic. *Proc Natl Acad Sci* 104(46):17977–17982.
- Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27(2):573–580.
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120.
- Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 268(1):78–94.
- Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol* 31(12):1119–1125.
- Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell M. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* 18(1):188–196.
- Caprioli A, Koyano-Nakagawa N, Iacovino M, Shi X, Ferdous A, Harvey RP, Olson EN, Kyba M, Garry DJ. 2011. Nkx2-5 represses Gata1 gene expression and modulates the cellular fate of cardiac progenitors during embryogenesis. *Circulation* 123(15):1633–1641.
- Chen N. 2004. Using repeat masker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* 5:4.10.11–14.10.14.
- Chen S, Yang P, Jiang F, Wei Y, Ma Z, Kang L. 2010. De novo analysis of transcriptome dynamics in the migratory locust during the development of phase traits. *PLoS One* 5(12):e15633.
- Chiancone E, Vecchini P, Verzili D, Ascoli F, Antonini E. 1981. Dimeric and tetrameric hemoglobins from the mollusc *Scapharca inaequivalvis*: structural and functional properties. *J Mol Biol* 152(3):577–592.
- De Bie T, Cristianini N, Demuth JP, Hahn MW. 2006. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22(10):1269–1271.
- Decker C, Zorn N, Potier N, Leize-Wagner E, Lallier F, Olu K, Andersen A. 2014. Globin's structure and function in Vesicomid bivalves from the Gulf of Guinea cold seeps as an adaptation to life in reduced sediments. *Physiol Biochem Zool* 87(6):855–869.
- Diesen DL, Hess DT, Stamler JS. 2008. Hypoxic vasodilation by red blood cells: evidence for an S-nitrosothiol-based signal. *Circ Res* 103(5):545–553.
- Ebner B, Panopoulou G, Vinogradov SN, Kiger L, Marden MC, Burmester T, Hankeln T. 2010. The globin gene family of the cephalochordate amphioxus: implications for chordate globin evolution. *BMC Evol Biol* 10(1):370.
- Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.
- Furuta H, Kajita A. 1983. Dimeric hemoglobin of the bivalve mollusc *Anadara broughtonii*: complete amino acid sequence of the globin chain. *Biochemistry* 22(4):917–922.
- Gambacurta A, Basili P, Ascoli F. 2000. *Scapharca inaequivalvis* A and B miniglobin genes: promoter activity of the 5' flanking regions and in vivo transcription. *Gene* 255(1):75–81.
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat Biotechnol* 29(7):644–652.
- Grispo MT, Natarajan C, Projecto-Garcia J, Moriyama H, Weber RE, Storz JF. 2012. Gene duplication and the evolution of hemoglobin isoform differentiation in birds. *J Biol Chem* 287(45):37647–37658.
- Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVIDENCEModeler and the program to assemble spliced alignments. *Genome Biol* 9(1):R7.
- Harlow E, Lane D. 1988. A laboratory manual. New York: Cold Spring Harbor Laboratory. p. 579.

- Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. 2018. UFBBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol*35(2):518–522.
- Hoffmann FG, Opazo JC, Hoogewijs D, Hankeln T, Ebner B, Vinogradov SN, Bailly X, Storz JF. 2012. Evolution of the globin gene family in deuterostomes: lineage-specific patterns of diversification and attrition. *Mol Biol Evol*29(7):1735–1745.
- Hoogewijs D, Ebner B, Germani F, Hoffmann FG, Fabrizio A, Moens L, Burmester T, Dewilde S, Storz JF, Vinogradov SN, et al. 2012. Androglobin: a chimeric globin in metazoans that is preferentially expressed in mammalian testes. *Mol Biol Evol*29(4):1105–1114.
- Hou S, Larsen R, Boudko D, Riley C, Karatan E, Zimmer M, Ordal G, Alam M. 2000. Myoglobin-like aerotaxis transducers in Archaea and Bacteria. *Nature*403(6769):540–544.
- Jiang N, Tan NS, Ho B, Ding JL. 2007. Respiratory protein-generated reactive oxygen species as an antimicrobial strategy. *Nat Immunol*8(10):1114–1122.
- Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics*30(9):1236–1240.
- Kalvari I, Nawrocki EP, Argasinska J, Quinones-Olvera N, Finn RD, Bateman A, Petrov AI. 2018. Non-coding RNA analysis using the Rfam database. *Curr Protoc Bioinformatics*62(1):e51.
- Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc*10(6):845–858.
- Kleinbongard P, Schulz R, Rassaf T, Lauer T, Dejam A, Jax T, Kumara I, Gharini P, Kabanova S, Özüyman B, et al. 2006. Red blood cells express a functional endothelial nitric oxide synthase. *Blood*107(7):2943–2951.
- Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol*34(7):1812–1819.
- Kuriyan J, Krishna T, Wong L, Guenther B, Pahler A, Williams C, Model P. 1991. Convergent evolution of similar function in two structurally divergent enzymes. *Nature*352(6331):172–174.
- Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* 9:559.
- Li L, Stoekert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*13(9):2178–2189.
- Lieb B, Dimitrova K, Kang H, Braun S, Gebauer W, Martin A, Hanelt B, Saenz S, Adema C, Markl J. 2006. Red blood with blue-blood ancestry: intriguing structure of a snail hemoglobin. *Proc Natl Acad Sci*103(32):12011–12016.
- Lowe TM, Chan PP. 2016. tRNAscan-SE on-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res*44(W1):W54–W57.
- Lu R, Lin Z, Zhang Y, Chai X, Dong Y, Xiao G, Zhang J, Fang J, Hu L. 2008. Comparison on the karyotypes of *Scapharca subcrenata*, *Tegillarca granosa* and *Estellarca olivacea* (In Chinese). *J Shanghai Fish Univ*17:625–629.
- Mangum CP. 2010. Comprehensive physiology. Invertebrate blood oxygen carriers. Hoboken (NJ): John Wiley & Sons, Inc.
- Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*27(6):764–770.
- Nicol PI, O’Gower AK. 1967. Haemoglobin variation in *Anadara trapezia*. *Nature*216(5116):684.
- Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*32(1):268–274.
- Paoli M, Marles-Wright J, Smith A. 2002. Structure–function relationships in heme-proteins. *DNA Cell Biol*21(4):271–280.
- Pillai AS, Chandler SA, Liu Y, Signore AV, Cortez-Romero CR, Benesch JLP, Laganowsky A, Storz JF, Hochberg GKA, Thornton JW. 2020. Origin of complexity in haemoglobin evolution. *Nature*581(7809):480–485.
- Porcher C, Swat W, Rockwell K, Fujiwara Y, Alt FW, Orkin SH. 1996. The T cell leukemia oncoprotein SCL/tal-1 is essential for development of all hematopoietic lineages. *Cell*86(1):47–57.
- Poulos TL. 2014. Heme enzyme structure and function. *Chem Rev*114(7):3919–3962.
- Prothmann A, Hoffmann FG, Opazo JC, Herbener P, Storz JF, Burmester T, Hankeln T. 2020. The globin gene family in Arthropods: evolution and functional diversity. *Front Genet*11:858.
- Roach MJ, Schmidt SA, Borneman AR. 2018. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*19(1):460.
- Ronda L, Bettati S, Henry ER, Kashav T, Sanders JM, Royer WE, Mozzarelli A. 2013. Tertiary and quaternary allostery in tetrameric hemoglobin from *Scapharca inaequivalvis*. *Biochemistry*52(12):2108–2117.
- Royer W, Hendrickson W, Chiancone E. 1990. Structural transitions upon ligand binding in a cooperative dimeric hemoglobin. *Science*249(4968):518–521.
- Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, Heard E, Dekker J, Barillot E. 2015. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol*16(1):259.
- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*13(11):2498–2504.
- 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(6):820–830.
- 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(7433):526–531.
- Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res*34(Web Server):W435–W439.
- Starck J, Weiss-Gayet M, Gonnet C, Guyot B, Vicat J-M, Morlé F. 2010. Inducible Fli-1 gene deletion in adult mice modifies several myeloid lineage commitment decisions and accelerates proliferation arrest and terminal erythrocytic differentiation. *Blood*116(23):4795–4805.
- Storz JF. 2016. Gene duplication and evolutionary innovations in hemoglobin-oxygen transport. *Physiology (Bethesda)*31(3):223–232.
- Storz JF. 2018. Hemoglobin: insights into protein structure, function, and evolution. Oxford: Oxford University Press.
- Suzuki T, Arita T. 1995. Two-domain hemoglobin from the blood clam, *Barbatia lima*. The cDNA-derived Amino Acid Sequence. *J Protein Chem*14(7):499–502.
- Suzuki T, Shiba M, Furukohri T, Kobayashi M. 1989a. Hemoglobins from the two closely related clams *Barbatia lima* and *Barbatia virescens*. Comparison of their subunit structures and N-terminal sequence of the unusual two-domain chain. *Zool J Linn Soc*6:269–281.
- Suzuki T, Takagi T, Ohta S. 1989b. Amino acid sequence of the dimeric hemoglobin (Hb I) from the deep-sea cold-seep clam *Calyptogena soyoae* and the phylogenetic relationship with other molluscan globins. *Biochim Biophys Acta*999(3):254–259.
- Terwilliger NB. 1998. Functional adaptations of oxygen-transport proteins. *J Exp Biol*201(Pt 8):1085–1098.
- Terwilliger NB, Terwilliger RC, Meyhöfer E, Morse MP. 1988. Bivalve hemocyanins-A Comparison with other molluscan hemocyanins. *Comp Biochem Physiol B*89(1):189–195.
- Titchen DA, Glenn WK, Nassif N, Thompson AR, Thompson EOP. 1991. A minor globin gene of the bivalve mollusc *Anadara trapezia*. *Biochim Biophys Acta*1089(1):61–67.
- Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*25(9):1105–1111.
- Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*28(5):511–515.

- Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MT, Azam S, Fan G, Whaley AM, et al. 2012. Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat Biotechnol*30(1):83–89.
- Wajcman H, Kiger L, Marden MC. 2009. Structure and function evolution in the superfamily of globins. *Cr Biol*332(2–3):273–282.
- Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*. 9(11):e112963.
- Wan J, Ristenpart WD, Stone HA. 2008. Dynamics of shear-induced ATP release from red blood cells. *Proc Natl Acad Sci*105(43):16432–16437.
- Wang S, Yu X, Lin Z, Zhang S, Xue L, Xue Q, Bao Y. 2017a. Hemoglobins likely function as peroxidase in blood clam *Tegillarca granosa* hemocytes. *J Immunol Res*2017:1–10.
- Wang S, Zhang JB, Jiao WQ, Li J, Xun XG, Sun Y, Guo XM, Huan P, Dong B, Zhang LL, et al. 2017b. Scallop genome provides insights into evolution of bilateral karyotype and development. *Nat Ecol Evol*1(5):1–12.
- Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol*35(3):543–548.
- Wittenberg JB. 1970. Myoglobin-facilitated oxygen diffusion: role of myoglobin in oxygen entry into muscle. *Physiol Rev*50:560–601.
- Xu Z, Wang H. 2007. LTR\_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res*35(Web Server):W265–W268.
- Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*24(8):1586–1591.
- Zambelli F, Mastropasqua F, Picardi E, D’Erchia AM, Pesole G, Pavesi G. 2018. RNentropy: an entropy-based tool for the detection of significant variation of gene expression across multiple RNA-Seq experiments. *Nucleic Acids Res*46(8):e46.
- Zhang GF, Fang XD, Guo XM, Li L, Luo RB, Xu F, Yang PC, Zhang LL, Wang XT, Qi HG, et al. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*490(7418):49–54.
- Zhang Y, Skolnick J. 2005. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*33(7):2302–2309.