Genomic adaptations to chemosymbiosis in the deep-sea seep-dwelling tubeworm *Lamellibrachia luymesi* (Siboglinidae, Annelida)

Yuanning Li[1,2*], Michael G. Tassia[1], Damien S. Waits[1], Viktoria E. Bogantes[1], Kyle T. David[1], Kenneth M. Halanych[1*]

[1] Department of Biological Sciences & Molette Biology Laboratory for Environmental and Climate Change Studies, Auburn University, Auburn, AL, 36849. USA

[2] Department of Ecology and Evolutionary Biology, Yale University, 165 Prospect St, New Haven, CT 06511. USA

* Corresponding author: yuanning.li@yale.edu; ken@auburn.edu

Classification: Biological Sciences: Evolution

Abstract

Genetic mechanisms allowing organisms to maintain host-symbiont associations at the molecular level are still mostly unknown. In the case of bacterial-animal associations, most genetic studies have focused on adaptations and mechanisms of the bacterial partner. The gutless tubeworms (Siboglinidae, Annelida) are obligate hosts of chemoautotrophic endosymbionts (except for *Osedax* which houses heterotrophic Oceanospirillales). Whereas several siboglinid endosymbiont genomes have been characterized, genomes of hosts remain unexplored. Here, we present and characterize the genome of the cold-seep dwelling tubeworm *Lamellibrachia luymesi*, one of the longest-lived invertebrates. With a haploid genome size of ~688 Mb and overall completeness of ~95%, we discovered that *L. luymesi* lacks many genes essential in amino acid biosynthesis obligating them to products provided by the symbionts. In comparison, the host carries hydrogen sulfide to thiotrophic endosymbionts using hemoglobin. Interestingly, we found a large expansion of hemoglobin B1 genes many of which possess a free cysteine residue which is hypothesized to function in sulfide-binding. Moreover, sulfide-binding mediated by zinc ions is not conserved across tubeworms, suggesting the hemoglobin structure and the sulfide-binding mechanism is potentially more complex than previously thought. Our comparative analyses also suggest the Toll-like receptor pathway may be essential to host immunity and tolerance/sensitivity to symbionts and pathogens. Last, we identified several genes known to play an important role in longevity. These results help elucidate previously unknown links and potential genetic mechanisms related to the evolution of holobionts, adaptations to reducing environments, and likely extend to other chemosynthetic symbiosis.

Keywords: chemosynthetic symbiosis, cold seep, comparative genomics, nutrition mode, hemoglobins, Toll-like receptor, aging

Significance

Symbioses between bacteria and animals are ubiquitous and ecosystems (e.g., seeps, hydrothermal vents, and organic falls) driven by chemoautotrophy have received considerable attention because of the non-photosynthetic energy source. However, genomic machinery that led to evolutionary success of these chemosynthetic environments is poorly understood, especially for hosts. By characterizing the genome of the seep-dwelling tubeworm *Lamellibrachia luymesi*, we provide genetic evidence of how animals adapted to extreme environments and maintain chemosynthetic symbiosis. Host genome adaptations include loss of biosynthesis pathways, expansion of hemoglobin gene families, innate immunity mechanisms related to host-symbiont recognition, and genes related to longevity. Our findings can be extended to other taxa and shed light on the mechanisms that establish and promote symbiosis, especially in chemosynthetic systems.

Introduction

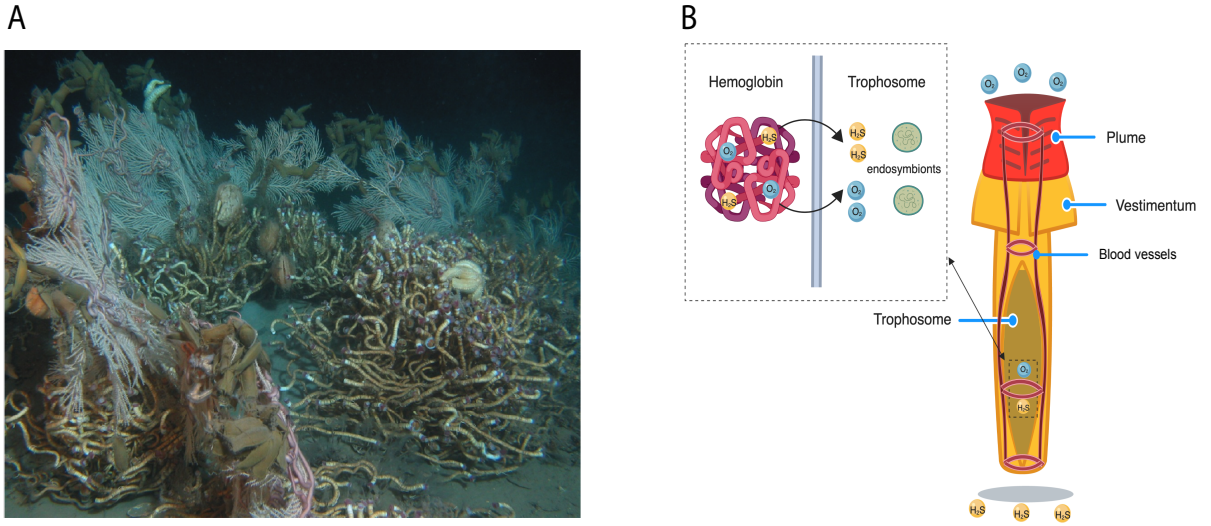

Fig. 1.*Lamellibrachia luymesi*. (A). Seep habitat in the Gulf of Mexico. (B). Diagram of adult *L. luymesi* worm to model O_2 and H_2S transport to symbionts in trophosome by hemoglobin molecules. The hemoglobin model was created with the help of Biorender (https://biorender.com/).

Recent advances in understanding the dominance of microbes on the planet has placed new emphasis on elucidating mechanisms that promote microbe-animal symbioses. Although considerable work has been undertaken on adaptations of microbial genomes to facilitate animal symbiosis (such as corals, termites, humans), examples of how animal host genomes have adapted to symbioses are limited (1). Vestimentiferan tubeworms inhabit some of Earth's most extreme environments, such as deep-sea hydrothermal vents and cold seeps, and are obligate dependents on symbiosis for survival. These animals lack a digestive tract and rely on sulfide-oxidizing bacterial symbionts for nutrition and growth. At some seeps, tubeworms, such as *Lamellibrachia luymesi* in the Gulf of Mexico, are so abundant that they transform the habitat (Fig. 1A) and thus facilitate biodiversity promoting adaptive radiations and evolutionary novelties (2). Given the obligate nature of the symbiosis between tubeworms and their gammaproteobacterial chemoautotrophic endosymbiont, one may reasonably expect adaptations in several cellular mechanisms and pathways (e.g.

nutrition, gas exchange, self-defense/self-recognition) to promote efficacy in the symbiotic relationship.

Siboglinid hosts acquire their symbionts from the surrounding environment and store them in a specialized tissue called the trophosome (3). The chemosynthetic symbionts are known to use a variety of molecules (e.g. H_2S, O_2, H_2) for final electron receptors facilitating a variety of fixation pathways (4). Primarily, vestimentiferan symbionts use both reverse TCA cycle (rTCA) and the Calvin cycle for carbon fixation providing a nutrient source for the host (4, 5). To date, metabolic studies have primarily focused on mechanisms and pathways found in symbionts and studies from the host's perspective are limited.

Another key adaptation contributing to the ability of tubeworms to thrive in chemosynthetic habitats involves hemoglobins (Hbs) that bind oxygen and sulfide simultaneously and reversibly at two different sites (6) (Fig. 1B). To avoid the toxicity of sulfide, siboglinids possess three different extracellular hemoglobins (Hbs): two dissolved in the vascular blood, V1 and V2, and one in the coelomic fluid, C1 (7, 8). Siboglinid Hbs consist of four heme-containing chains (A1, A2, B1, B2). Sulfur-binding capabilities are hypothesized to be dependent on free cysteine residues at key positions in Hbs, especially in the A2 and B2 chains (6). V1 Hb can form persulfide groups on its four linker chains (L1-L4), a mechanism that can account for the higher sulfide-binding potential of this Hb (6). However, a study suggested sulfide-binding affinity was mediated by the zinc moieties bound to amino acid residues at the interface between pairs of A2 chains in *Riftia* (9). Thus, it is still not clear which mechanism is primarily responsible for sulfide-binding in siboglinids.

In contrast to rapidly growing vent-dwelling vestimentiferans (10), seep-dwelling vestimentiferans have much slower growth rates, and are among the most long-lived non-colonial marine invertebrates (up to 250 years) (11). Immunity has important implications in aging (12), and is also a critical evolutionary driver of maintaining symbiosis (13). However, little is known about genetic mechanisms relating immunity and symbiosis. Because tubeworm endosymbionts are housed internally and their establishment process resembles infection (3), tubeworm symbiosis provides a unique

opportunity to examine evolution of immunity functions associated with host-symbiont relationships. However, Information on extremophile immunity and/or immune tolerance is lacking.

Using comparative genomics, transcriptomic and proteomic analyses on the tubeworm *Lamellibrachia luymesi*, we provide evidence for genetic pathways and novel candidate genes which may underlie the extraordinary characteristics of tubeworm symbioses. In particular, we focus on nutrition mode, hemoglobin evolution, immunity function, and longevity.

Results and Discussion

Genome features

Using Illumina paired-end, mate-pair and 10X genomic sequencing (Table S1), we the assembled genome of a single *Lamellibrachia luymesi* individual. The haploid genome assembly size is ~688 Mb (Fig. S1) with ~500X coverage and N50 values of 373 Kb (scaffolds) and 24 Kb (contigs). Although N50 lengths and assembly quality of *L. luymesi* are comparable to those of other annelids (e.g. *Capitella teleta*, *Helobdella robusta*) (Tables S2, S3), the overall genome completeness measured by BUSCO (~ 95%) is one of the highest among lophotrochozoans (Table S2). With the support of RNA-seq data from three different tissues (Table S1), we estimated *L. luymesi* genome contains 38,998 gene models. The genome also exhibits heterozygosity (0.6%) and repetitive content (36.92%) similar to other lophotrochozoans (Fig. S2, Table S4). We found 94 orthology groups (OGs) appear to have undergone a genomic expansion compared to other lophotrochozoan genomes (Tables S5).

Nutritional adaptations

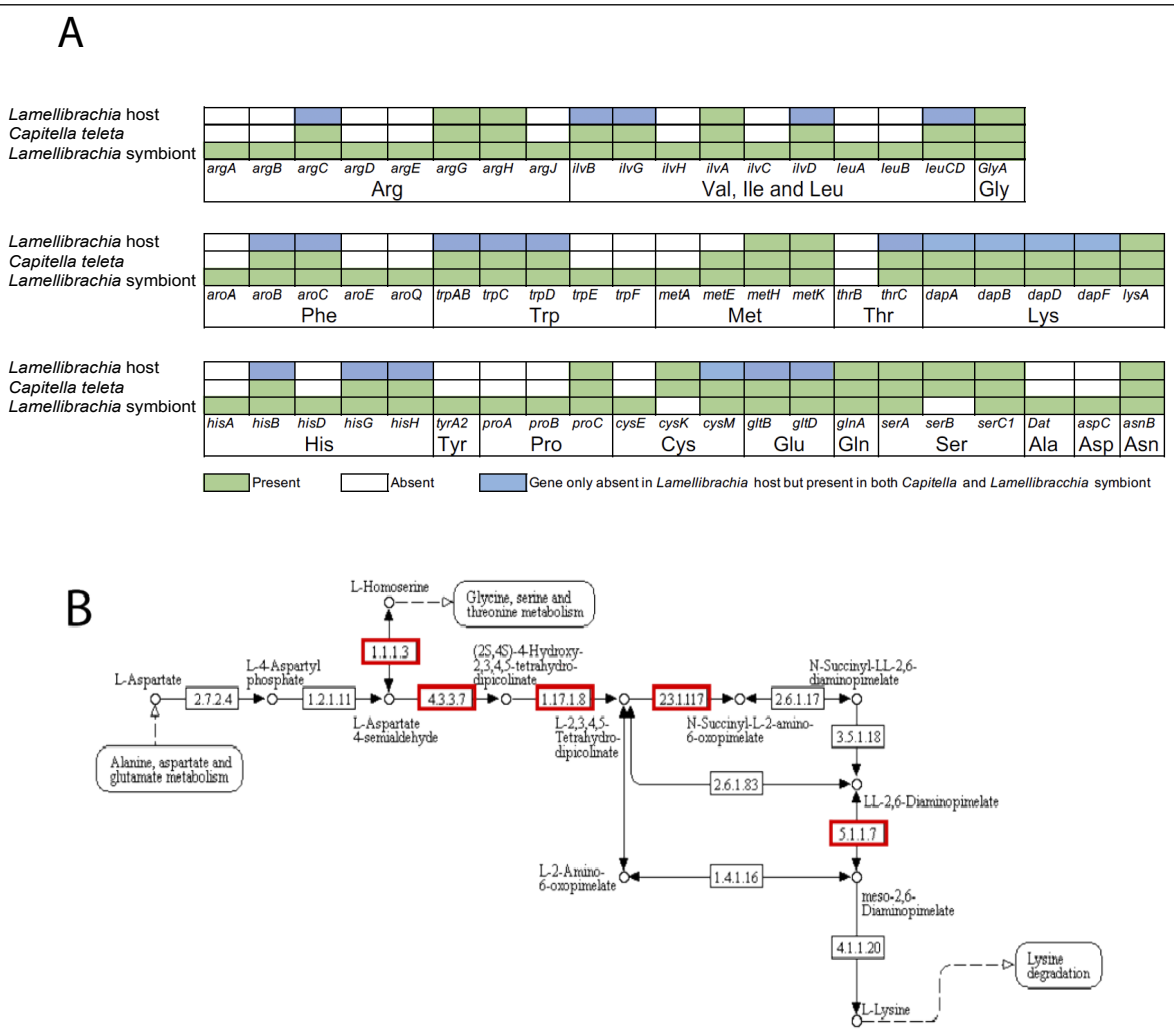

Fig. 2. *Lamellibrachia luymesi* lacks amino acid biosynthesis genes. (A) Presence (green) or absence (white boxes) of key genes associated with amino acid biosynthesis in the genomes of *Capitella teleta*, *L. luymesi* and *L. luymesi* symbionts. Blue boxes represents genes present in *C. teleta* and *L. luymesi* gammaproteobacterial symbionts but absent in *L. luymesi*. (B) Example of Lysine biosynthesis pathway. Red boxes indicate genes missing in *L. luymesi*. Figure was created with the help of KEGG webserver.

Only 57 genes associated with amino acid biosynthesis were found in the *L. luymesi* genome, of which eight were also identified in the proteomic analysis. In contrast, the *Capitella teleta* (Capitellidae, Annelida) genome contains 90 such genes

(Fig. 2A; Supplementary Dataset 1), despite being a less complete and more fragmented genome (Table S2). These gene were not clustered together in the genomes suggesting that they were probably not missed due to random chance given the completeness of sequencing. Interestingly, the *L. luymesi* symbiont genome contains 110 genes, an essentially complete set for biosynthesis of all 20 proteinogenic amino acids and of 11 vitamins/cofactors. Genes found in C. *telata*'s genome but lacking in *L. luymesi* are involved in biosynthesis of 13 amino acids (e.g., five key enzymes in the Lysine biosynthesis pathway Fig. 2B). As amino acids are essential for protein biosynthesis in the host, the lack of many important amino acid synthesis-related genes indicate that the host depends on symbionts for amino acids and cofactors. Moreover, we found a large gene expansion of nutrient uptake ABC transport protein-coding genes in *L. luymesi* compared with other lophotrochozoans (Table S5). These findings are consistent with previous biochemical analyses which suggest that *Riftia* is also dependent on its bacterial symbiont for the biosynthesis of polyamines that are important for host metabolism and physiology (14).

Obligate bacterial symbionts often lack genes that are commonly found in other free-living bacteria, while retaining only those genes with functions essential to host needs (e.g. in sponges, (15); in termites, (16)). However, there are known cases of loss in essential gene functions in multicellular eukaryotes, but this phenomenon appears to be more frequent in bacterial symbionts (1). Interestingly, thiotrophic symbionts of the vesicomyid clam *Calyptogena magnifia* (17) and vent mussel *Bathymodiolus azoricus* (18) have been suggested provide their host with products from amino acid biosynthesis. Moreover, a recent study has suggested that the flatworm *Paracatenula* itself does not store primary energy in host cells; rather, this function is performed by its chemosynthetic symbionts (19). Although the tubeworms and bivalves under examination in the aforementioned studies live in chemosynthetic environments, the different hosts and bacteria represent disparate genomic backgrounds suggesting that modification and loss of the amino acid biosynthesis pathways may be a convergent adaptation in a variety of chemosynthetic symbioses between bacteria and animals.

In addition to the immediate release of fixed carbon and provision of amino acids by symbionts, we have found proteomic evidence of a second possible nutritional mode whereby the host directly digests symbionts, as shown by the detection of abundant host-derived digestive enzymes in trophosome tissue (Table S6). Previous observations indicated that symbionts could be digested by *Riftia* (20) but, direct evidence and mechanisms related to symbiont digestion lacked characterization. We identified 15 host proteins related to lysosomal proteases that were both highly expressed and detected as proteins in the trophosome tissue of host genome, such as Saposin and multiple copies of Cathepsin (Table S6). Lysosomes, which contain an array of digestive enzymes, are also thought to play an essential role in symbiont digestion with the chemosynthetic mussel *Bathymodilus azoricus* (18). We additionally identified 19 major proteasome components as proteins in the trophosome tissue, indicating a potential role in protein degradation of symbiont digestion (Table S6). Host lysosomal proteases and proteasome components likely facilitate degradation of symbionts and may play a role maintaining appropriate population levels of symbionts within trophosome.

We also characterized ~ 200 bacterial proteins present in the same trophosome tissue to further understand host-symbiont interactions. Key enzymatic genes, RubisCO, and ATP citrate lyase (ACL) type II associated with carbon fixation cycles, were identified in proteomic analysis from *L. luymesi* (Table S7). Our results corroborate that both rTCA and Calvin cycle, pathways for carbon fixation might be common in all vestimentiferan endosymbionts (4). Several key components related to sulfide and nitrogen metabolic pathways were identified consistent with previous analyses (4, 5).

Hemoglobin evolution

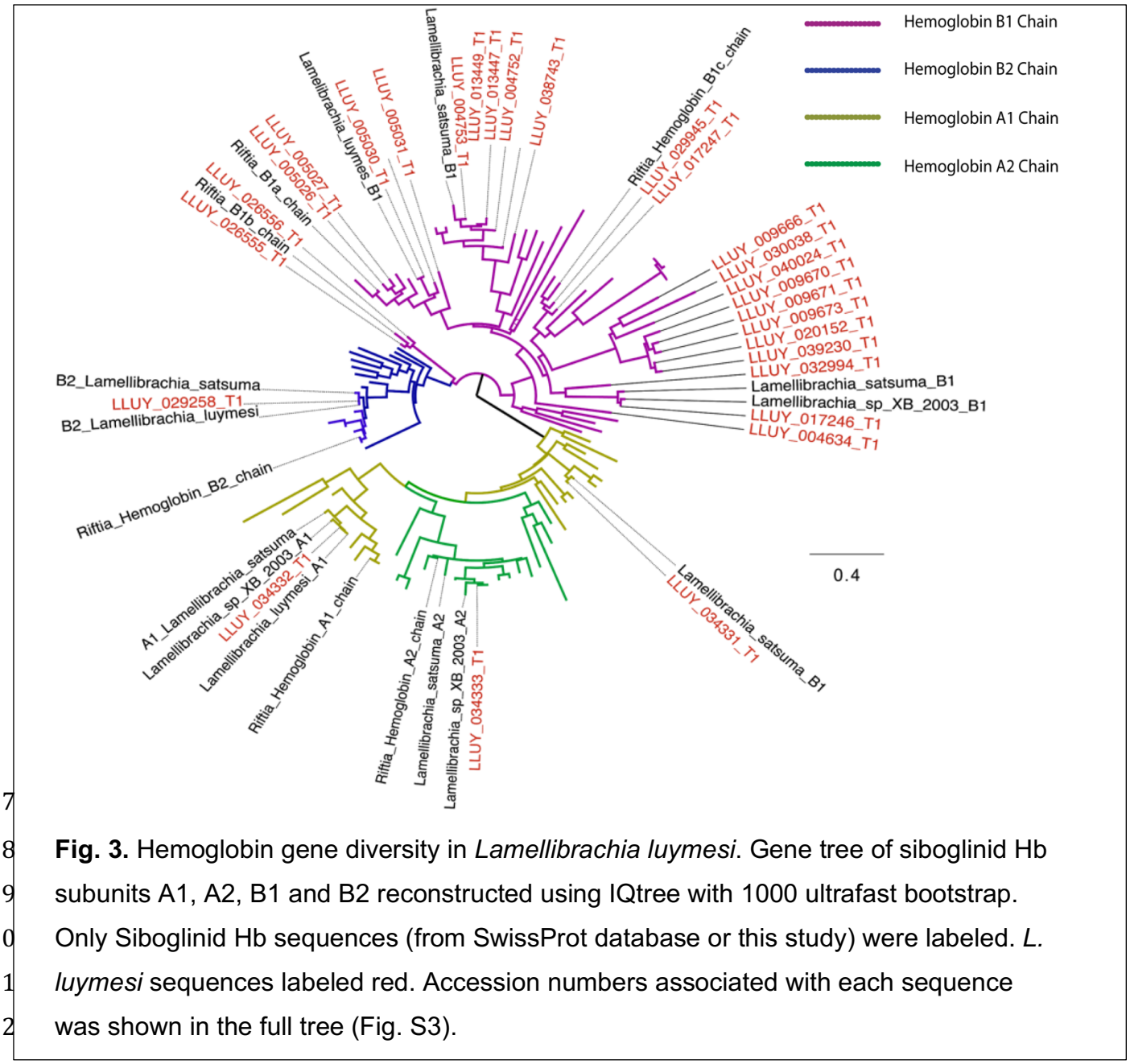

Fig. 3. Hemoglobin gene diversity in *Lamellibrachia luymesi*. Gene tree of siboglinid Hb subunits A1, A2, B1 and B2 reconstructed using IQtree with 1000 ultrafast bootstrap. Only Siboglinid Hb sequences (from SwissProt database or this study) were labeled. *L. luymesi* sequences labeled red. Accession numbers associated with each sequence was shown in the full tree (Fig. S3).

Mechanisms of Hb sulfide-binding affinity in tubeworm siboglinids are still not clear after 20 years of study. We collected all available Hb sequences from siboglinids and their close relatives and processed them through a phylogenetic framework (Figs. 3, S3). Importantly, we are be able to identify most Hbs and linkers from transcriptomic and proteomic results (Table S7). Consistent with (6, 8, 9) a single copy of A2 and B2

Hb was identified in all siboglinids which possesses a conserved-free cysteine (i.e., cysteine residues not involved in disulfide bridges) at position 77 and 67, respectively. With exception of A2 and B2 Hbs in the earthworm *Lumbricus terrestris*, homologous cysteine residues were identified in 3 annelids (*Cirratulus spectabilis*, *Sabella pacifica*, and *Sternapsis* sp.) from sulfide-free environments and *Arenicola marina* living in sulfide-rich environments (Fig. S4). These results support the hypothesis that free cysteine residues in A2 and B2 Hbs were present in all annelids and potentially involved in H_2S detoxification process (21).

Surprisingly, we found a significant expansion of B1 Hbs, 25 copies, in *L. luymesi* whereas most siboglinids and their close relatives only possess one copy indicated by previous studies (Fig. 3B), except for *Riftia pachyptila* where three B1 Hbs were identified (21). Noticeably, we found that 8 copies of *L. luymesi* B1 Hb sequences also contains a free cysteine at position 77, the same position as free cysteine in A2 Hbs. Five out the 8 copies were highly expressed in the trophosome, and one copy was identified at the protein level (Table S8).

Instead of free cysteines mediating H_2S binding, another hypothesis suggested that zinc moieties bound to amino acid residues at the interface between pairs of A2 chains influence H_2S binding (9). The Zn^{2+}-binding site contained within A2 chain is composed of three His residues (B12, B16, and G9) (9). However, none of these sites are conserved across siboglinids, or even in vestimentiferans (Fig. S5) calling into question the role of the zinc sulfide binding mechanism for H_2S transport.

Immunity function

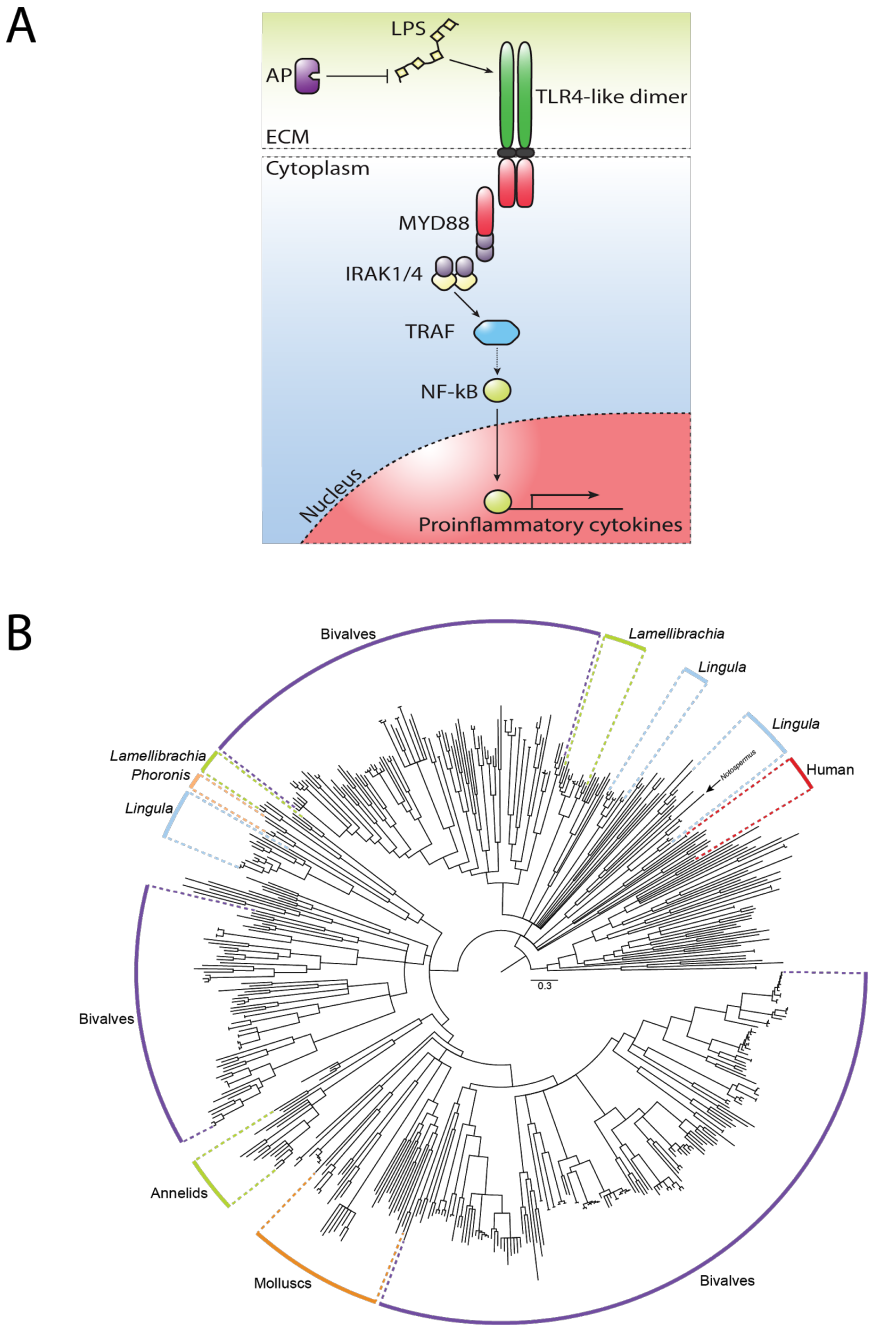

Fig. 4. Toll-Like Receptors (TLRs) in *Lamellibrachia luymesi*. (A) Putative TLR4-like pathway likely essential for immunity and response to symbionts and pathogens. AP: alkaline phosphatase; LPS: lipopolysaccharide. (B) Toll-Like Receptor gene tree from selected lophotrochozoan genomes and human reconstructed using IQtree with 1000 ultrafast bootstraps. All internal nodes possess ≥95% bootstrap support.

Immune interactions between hosts and symbionts is a key evolutionary driver that has potential implications in aging (12). The genetic machinery and functionality of the immune system in chemosynthetic symbioses have not been extensively characterized. Toll-Like Receptor (TLRs) provides a core cellular and molecular interface between invading pathogens and recognition of host-microbial symbiosis (13) (Fig. 4A). Consistent with previous analyses (Luo et al. 2018), we found that TLR gene families experienced expansion within lophotrochozoan lineages (Fig. 4B; Table S9). Within *L. luymesi*, 33 unique TLR proteins were identified compared to 5 in *Capitella telata*, suggesting TLR genes have additional functions in tubeworms.

A substantial subset of TLR sequences recovered form *L. luymesi* best identify as TLR4 by primary sequence identity and domain structures. In mammals, TLR4 recognizes and binds lipopolysaccharide (LPS; a major cell-membrane component of Gram-negative bacteria which include tubeworm symbionts). LPS-bound TLR4 then initiates a signal-transduction pathway that activates NF-kB, a transcription factor that promotes the expression of pro-inflammatory cytokines (22) (Fig. 4). *Lamellibrachia luymesi* encodes seven TLR4-like proteins, which is in contrast to the one sequence found in other annelid genomes suggesting potential for increased sensitivity to Gram-negative bacteria in *L. luymesi*. Interestingly, we also found genomic expansions of tumor necrosis factor receptors (TNFRs) and TNFR-associated factors (TRAFs) (Table S5) which play vital roles in activation and the downstream responses of NF-kB, further supporting a specialized/expanded role for TLR4-like signaling. Whereas some other components of the innate immunity (e.g. RIG-1-like receptor signaling pathway which recognizes virus-derived nucleotide present in the cytoplasm) showed no indication of gene expansion, the NLRP gene family (which plays a key role in an innate immunity recognition of infectious pathogens and regulates inflammatory caspases) and Sushi domain-containing genes (potential recognition and adhesion between hosts and symbionts, (18) showed expansion relative to other lophotrochozoans. (Table S9).

The initial physical encounter between tubeworms and symbionts occurs in an extracellular mucus secreted by pyriform glands of newly settled larvae (3). Within these mucus matrices, symbionts can attach to the host using extracellular components

secreted from symbionts, such as LPS. The symbiont's colonization process induces massive apoptosis of host skin tissue as symbionts travel from host epidermal cells into trophosome (3). Recognition of lipopolysaccharide (LPS) by TLR4 can result in the induction of signaling cascades that lead to activation of NF-kB and the production of proinflammatory cytokines (13). Although the mechanism by which host distinguishes between symbionts and pathogens in most symbioses is still not clear, alkaline phosphatase has been shown to be involved in the maintenance of homeostasis of commensal bacteria in the squids, mouse, and zebrafish (23). The commensal bacterially-derived LPS signaling via TLR4 yields an upregulation of intestinal alkaline phosphatase and prevents inflammatory responses to resident microbiota. Importantly, we also identified 8 copies of alkaline phosphatase, whereas only one copy was found in each of the *Capitella teleta* and *Helobdella robusta* genomes, further supporting a potential mechanism of tolerating Gram-negative bacteria and facilitating symbiotic colonization. Thus, although further analysis is warranted, a TLR4-like signaling pathway may be central for host immunity and in distinguishing between symbionts and pathogens (Fig. 4A).

Aging

Seep-living vestimentiferans are long lived, and in addition to innate immunity, our analyses of gene family expansion highlighted families that may play a direct role in aging. We found expansion of interleukin 6 receptors (IL6R) which are the key component of the main signaling pathway implicated in aging (24). Superoxide dismutases (SODs) have important function role in cells to protect against oxidative damage induced by metabolism and are implicated in aging and redox balancing. We found genomic expansions of CuZn-superoxide dismutase (SOD1) genes and Mn-superoxide dismutase (SOD2) in *L. luymesi*'s genome compared to other lophotrochozoans (Fig. S6). Most lophotrochozoan genomes contain one or two copies of SOD1 and SOD2, but *L. luymesi* has 5 copies of each gene (Fig. S6). Three of 5 SOD2 genes were recovered in transcriptomic and proteomic data (Table S6). Previous studies suggested that overexpression of SOD1 or SOD2 could significantly extend lifespan in mammals, fruit flies and *C. elegans* (25) and SOD gene product may help

316 symbionts overcome host cellular immune responses (26). Consistent with previous
317 studies, we also be able to identify symbionts' SOD gene as proteins in proteomic
318 analysis. Thus, SODs from both bacteria and tubeworms may play a central role for
319 overcoming oxidative damage and essential for extreme longevity for seep-living
320 vestimentiferans.

321 **Conclusion**

322 We characterize of the genome for the deep-sea seep-living tubeworm
323 *Lamellibrachia luymesi*. This report provides critical insight that hosts, like their bacterial
324 partners, may lose essential genomic components when their life-history strategy relies
325 on symbiotic interactions. Analyses show that *Lamellibrachia luymesi* has lost key
326 genes for amino acid biosynthesis making it necessarily dependent on endosymbionts.
327 Additionally, expansions have occurred in a number of gene families (e.g., TLRs, SODs,
328 Hemoglobins) that have been implicated in bacterial symbiosis. Evolutionarily,
329 increasing the number of paralogs provides opportunity for neofunctionalization or
330 subfunctionalization allowing more refined gene-gene interactions to promote symbiotic
331 efficacy. This balance of gene family expansion and gene loss may be a hallmark of
332 how genomic machinery adapts and develops interdependence across of variety of
333 bacterial-animal symbioses.

334 **Methods.**

335 **Organismal collection**

336 *Lamellibrachia luymesi* was collected from seeps in the Mississippi Canyon in the
337 Gulf of Mexico (N 28°11.58', W 89°47.94' 754m depth), using the *R/V Seward Johnson*
338 and *Johnson Sea Link* in October 2009. Samples were frozen at 80˚C following
339 recovery.

340 **Genome sequencing and assembly**

341 Using vestimentum tissue of one individual, high molecular weight genomic DNA
342 was extracted using the DNeasy Blood & Tissue Kit (Qiagen). Four TruSeq paired-end
343 and two Nextra mate-pair genomic DNA libraries were generated and sequenced by

The Genomic Services Lab at the Hudson Alpha Institute for Biotechnology in Huntsville, Alabama on an Illumina HiSeq platform (Table S1). Additionally, Hudson Alpha constructed and sequenced a Chromium 10X sequencing library (10X genomics) on an Illumina HiSeqX platform.

Our genome assembly workflow is shown in Fig. S7. Paired-end and 10X raw reads were checked with FastQC v0.11.5 (27) and quality filtered (Q score >30) with Trimmomatic v0.36 (28). Genome size, level of heterozygosity, and repeat content were determined using kmer histograms generated from the paired-end libraries in Jellyfish v2.2.3 (29) and GenomeScope (30) (Fig. S1). Mate-pair reads were trimmed and sorted using NxTrim v0.3.1 (31), and only “mp” (true mate-pair reads) and “unknown” (mostly large insert size reads) reads were used for downstream scaffolding analysis.

Given high heterozygosity in non-model species, all reads were assembled using Platanus v1.2.4 (32) with a kmer size of 32. Scaffolding was conducted by mapping PE and MP reads to Platanus contigs using SSPACE v3.0 (33). Gaps in scaffolds were filled with GapCloser v1.12 (34) and redundant allele scaffolds were removed using Redundans v0.13c (default settings; (35). Genome assembly quality was assessed with QUAST v3.1 (36) and genome completeness with BUSCO v3 (37) using the Metazoa_odb9 database (978 Busco genes). We also assemble the genome using 10X data in Supernova 1.2.0 (Weisenfeld et al. 2017), but the genome quality and completeness was inferior to the Platanus assembly (Fig. S7) and there for ignored.

Transcriptome assembly and analysis

Total RNA was extracted via Trizol (Thermo Fisher Scientific) from the plume, vestimentum and trunk/trophosome tissue of the same *L. luymesi*. RNA-Seq was carried out by Hudson Alpha on using Illumina HiSeq platform. After quality checking and trimming of raw sequencing reads, transcripts were assembled in Trinity v2.4.0 (38). Transcript isoforms with high similarity (≥ 95%) were removed with CD-HIT-EST v4.7 (39). Transcripts were verified and abundance estimated by read mapping with Bowtie v2.2.9 (40) and RSEM v1.2.26 (41).

Genome annotation

Gene models were constructed following the Funannotate pipeline 1.3.0 (https://github.com/nextgenusfs/funannotate; Fig. S8) using information from the genome assembly, transcriptome assembly, and SwissProt/Uniprot. For genome data, repetitive regions were identified using RepeatModeler v1.0.8 (43) and soft-masked using RepeatMasker v4.0.6 (44). For each transposable element (TE) superfamily, relative ages of different copies were estimated by calculating Kimura distances assuming that most of the mutations are neutral using repeatLandscape.R (https://github.com/dunnlab/genome_annotation/blob/master/repeatLandscape.R). RNA-Seq data combined into a single *de novo* assembly with Trinity and a spliced alignment indexed against the genome assembly with HISAT 2.1.0 (45). The PASA pipeline v2.3.3 (46) was used to identify high-quality gene models that were used to train the *ab initio* gene predictor in AUGUSTUS v3.3 (47) and GenMark. Additionally, SwissProt protein data was aligned to the genome assembly using Exonerate (48) and *L. luymesi* transcripts aligned using Minimap2 v2.1 (49). tRNA genes were identified with tRNAscan-SE v1.3.1 (50). Finally, EvidenceModeler 1.1.0 (51) was used to combine all evidence of gene prediction from protein alignments, transcript alignments, and *ab initio* predictions to construct high-quality consensus gene models. Functional annotations of predicted gene models were performed using curated databases: KEGG orthology was assigned using the KEGG Automatic Annotation server (52), domain structure by InterProScan (53), and protein identity with the SwissProt database. Secreted proteins were predicted using SignalP (54) and Phobius (55) in InterProScan.

Proteomics characterization

Proteomic analysis was performed by Proteomics & Metabolomics Facility at Colorado State University. Briefly, trunk/trophosome tissue was cleaned and homogenizated. Protein in resulting supernatant was quantified by the Pierce BCA Protein Assay Kit (ThermoFisher-Pierce). Absorbance was measured at 550nm and using a Bovine Serum Albumin (BSA) control. 50 µg total protein was processed for in-solution trypsin digestion (56). Tandem mass spectra were extracted, charge state deconvoluted and deisotoped by ProteoWizard MsConvert (version 3.0). Spectra searched against gene models of *L. luymesi* host (herein) and symbiont genomes ((4)) using Mascot (Matrix Science, London, UK; version 2.6.0) with a fragment ion mass

tolerance of 0.80 Da and a parent ion tolerance of 20 PPM. Search results assessed with probabilistic protein identification algorithms (57) implemented in the Scaffold software v. 4.8.4, (Proteome Software Inc., Portland, OR; (58). Protein identifications required >99.0% probability (with Protein Prophet algorithm; (59) and presence of ≥1 identified peptide. Proteins that contained similar peptides and could not be differentiated based on MS/MS analysis alone were grouped (SI methods).

Gene family analysis

Following all-to-all Diamond v1.0 (60) BLASTP searches against 22 selected lophotrochozoan proteomes (Table S3), orthology groups (OGs) were identified using Orthofinder with a default inflation parameter (I=1.5). Gene ontology annotation used PANTHER v13.1 (61) with the PANTHER HMM scoring tool (pantherScore2.pl). Gene family expansion and contraction was estimated using CAFÉ v2.1 (62). For each gene family, CAFÉ generated a family-wide P value, with a significant *P* value indicating a possible gene-family expansion or contraction event. Significantly expanded gene families ($p < 0.05$) were then identified by InterProscan.

Manual annotation of gene families

In addition to the annotation pipeline mentioned above, we manually annotated genes of interest herein: hemoglobin gene families, genes related to amino acid synthesize, immunity, and longevity. These gene families were specifically selected *a priori* based on our experience and review of available publications in the field. See *SI methods* for detailed procedure.

Data Availability

Raw reads, assembled genome sequences and annotation are accessible from NCBI under BioProject numbers PRJNA516467, Sequence Read Archive accession numbers SRR851910-SPR851919 and Whole Genome Shotgun project numbers SDWI00000000. The genome annotations, proteomic results, scripts and data for the

analyses are available from the Github Repository at https://github.com/yzl0084/Lamellibrachia-genome.

Acknowledgments

This study was supported by awards from the U.S. National Science Foundation (NSF) (DEB-1036537 and IOS-0843473 to KMH, Scott Santos and DanThornhill). YL was supported by the China Scholarship Council (CSC). We thank Chris Little, Maggie Georgieva, Luke Parry, and Jason Flores for the helpful discussions. We thank Zack and Ian Gilman for help with revising the manuscript. We thank Jon Palmer helped troubleshoot the Funannoate pipeline. We thank Kitty Brown for help with proteomic data interpretation. Bioinformatic analyses were conducted on the Auburn University Molette Laboratory SkyNet server, Auburn University Hopper HPC system, and the Alabama Supercomputer Authority. This is Molette Biology Laboratory contribution ### and Auburn University Marine Biology Program contribution ###.

Author Contributions

YL and KMH designed research; YL, MGT, DSW, VEB, KTD and KMH performed research and data analysis; YL, MGT and KMH wrote the paper. All authors contributed to revise the paper.

References

1. Moran NA (2007) Symbiosis as an adaptive process and source of phenotypic complexity. *Proc Natl Acad Sci U S A* 104 Suppl 1:8627–8633.
2. Boetius A (2005) Microfauna-macrofauna interaction in the seafloor: lessons from the tubeworm. *PLoS Biol* 3(3):e102.
3. Nussbaumer AD, Fisher CR, Bright M (2006) Horizontal endosymbiont transmission in hydrothermal vent tubeworms. *Nature* 441(7091):345.
4. Li Y, Liles MR, Halanych KM (2018) Endosymbiont genomes yield clues of tubeworm success. *ISME J* 12(11):2785.

458 5. Markert S, et al. (2007) Physiological proteomics of the uncultured endosymbiont of
459 *Riftia pachyptila. Science* 315(5809):247–250.

460 6. Zal F, et al. (1997) Primary structure of the common polypeptide chain b from the
461 multi-hemoglobin system of the hydrothermal vent tube worm Riftia pachyptila: An
462 insight on the sulfide binding-site. *Proteins: Struct Funct Bioinf* 29(4):562–574.

463 7. Arp AJ, Childress JJ (1981) Blood function in the hydrothermal vent vestimentiferan
464 tube worm. *Science* 213(4505):342–344.

465 8. Zal F, Lallier FH, Green BN, Vinogradov SN, Toulmond A (1996) The multi-
466 hemoglobin system of the hydrothermal vent tube worm *Riftia pachyptila* II.
467 Complete polypeptide chain composition investigated by maximum entropy analysis
468 of mass spectra. *J Biol Chem* 271(15):8875–8881.

469 9. Flores JF, et al. (2005) Sulfide binding is mediated by zinc ions discovered in the
470 crystal structure of a hydrothermal vent tubeworm hemoglobin. *Proceedings of the*
471 *National Academy of Sciences* 102(8):2713–2718.

472 10. Lutz RA, et al. (1994) Rapid growth at deep-sea vents. *Nature* 371(6499):663.

473 11. Bergquist DC, Williams FM, Fisher CR (2000) Longevity record for deep-sea
474 invertebrate. *Nature* 403(6769):499.

475 12. Quesada V, et al. (2018) Giant tortoise genomes provide insights into longevity and
476 age-related disease. *Nature ecology & evolution*:1.

477 13. Chu H, Mazmanian SK (2013) Innate immune recognition of the microbiota
478 promotes host-microbial symbiosis. *Nat Immunol* 14(7):668.

479 14. Minic Z, Hervé G (2003) Arginine metabolism in the deep sea tube worm *Riftia*
480 *pachyptila* and its bacterial endosymbiont. *J Biol Chem*.

481 15. Tian R-M, et al. (2017) Genome Reduction and Microbe-Host Interactions Drive
482 Adaptation of a Sulfur-Oxidizing Bacterium Associated with a Cold Seep Sponge.
483 *mSystems* 2(2). doi:10.1128/mSystems.00184-16.

484 16. Tokuda G, et al. (2013) Maintenance of essential amino acid synthesis pathways in
485 the Blattabacterium cuenoti symbiont of a wood-feeding cockroach. *Biol Lett*
486 9(3):20121153.

487 17. Newton ILG, Girguis PR, Cavanaugh CM (2008) Comparative genomics of
488 vesicomyid clam (Bivalvia: Mollusca) chemosynthetic symbionts. *BMC Genomics*
489 9(1):585.

490 18. Ponnudurai R, et al. (2017) Metabolic and physiological interdependencies in the
491 *Bathymodiolus azoricus* symbiosis. *ISME J* 11(2):463.

492 19. Jäckle O, et al. (2019) Chemosynthetic symbiont with a drastically reduced genome

493 serves as primary energy storage in the marine flatworm *Paracatenula*. *Proc Natl*
494 *Acad Sci U S A*. doi:10.1073/pnas.1818995116.

495 20. Bright M, Keckeis H, Fisher CR (2000) An autoradiographic examination of carbon
496 fixation, transfer and utilization in the *Riftia pachyptila* symbiosis. *Mar Biol*
497 136(4):621–632.

498 21. Bailly X, et al. (2002) Evolution of the sulfide-binding function within the globin
499 multigenic family of the deep-sea hydrothermal vent tubeworm *Riftia pachyptila*. *Mol*
500 *Biol Evol* 19(9):1421–1433.

501 22. Park BS, Lee J-O (2013) Recognition of lipopolysaccharide pattern by TLR4
502 complexes. *Exp Mol Med* 45(12):e66.

503 23. Bates JM, Akerlund J, Mittge E, Guillemin K (2007) Intestinal alkaline phosphatase
504 detoxifies lipopolysaccharide and prevents inflammation in zebrafish in response to
505 the gut microbiota. *Cell Host Microbe* 2(6):371–382.

506 24. Maggio M, Guralnik JM, Longo DL, Ferrucci L (2006) Interleukin-6 in aging and
507 chronic disease: a magnificent pathway. *J Gerontol A Biol Sci Med Sci* 61(6):575–
508 584.

509 25. Melov S, et al. (2000) Extension of life-span with superoxide dismutase/catalase
510 mimetics. *Science* 289(5484):1567–1569.

511 26. Bright M, Bulgheresi S (2010) A complex journey: transmission of microbial
512 symbionts. *Nat Rev Microbiol* 8(3):218–230.

513 27. Andrews S, Others (2010) FastQC: a quality control tool for high throughput
514 sequence data.

515 28. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina
516 sequence data. *Bioinformatics* 30(15):2114–2120.

517 29. Marçais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel
518 counting of occurrences of k-mers. *Bioinformatics* 27(6):764–770.

519 30. Vurture GW, et al. (2017) GenomeScope: fast reference-free genome profiling from
520 short reads. *Bioinformatics* 33(14):2202–2204.

521 31. O'Connell J, et al. (2015) NxTrim: optimized trimming of Illumina mate pair reads.
522 *Bioinformatics* 31(12):2035–2037.

523 32. Kajitani R, et al. (2014) Efficient de novo assembly of highly heterozygous genomes
524 from whole-genome shotgun short reads. *Genome Res*:gr–170720.

525 33. Boetzer M, Pirovano W (2014) SSPACE-LongRead: scaffolding bacterial draft
526 genomes using long read sequence information. *BMC Bioinformatics* 15(1):211.

527 34. Luo R, et al. (2012) SOAPdenovo2: an empirically improved memory-efficient short-

read de novo assembler. *Gigascience* 1(1):18.

35. Pryszcz LP, Gabaldón T (2016) Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res* 44(12):e113–e113.

36. Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29(8):1072–1075.

37. Waterhouse RM, et al. (2017) BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol* 35(3):543–548.

38. Haas BJ, et al. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* 8(8):1494.

39. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22(13):1658–1659.

40. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357.

41. Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12(1):323.

42. Albertin CB, et al. (2015) The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature* 524(7564):220.

43. Smit AFA, Hubley R (2008) RepeatModeler Open-1.0. *Available fom http://www repeatmasker org*.

44. Chen N (2004) Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* 5(1):4–10.

45. Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 12(4):357.

46. Haas BJ, et al. (2003) Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* 31(19):5654–5666.

47. Stanke M, et al. (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* 34(suppl_2):W435–W439.

48. Slater GSC, Birney E (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6(1):31.

49. Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 1:7.

50. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25(5):955.

51. Haas BJ, et al. (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* 9(1):1.

52. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 35(suppl_2):W182–W185.

53. Zdobnov EM, Apweiler R (2001) InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17(9):847–848.

54. Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* 8(10):785.

55. Käll L, Krogh A, Sonnhammer ELL (2007) Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. *Nucleic Acids Res* 35(suppl_2):W429–W432.

56. Schauer KL, Freund DM, Prenni JE, Curthoys NP (2013) Proteomic profiling and pathway analysis of the response of rat renal proximal convoluted tubules to metabolic acidosis. *American Journal of Physiology-Renal Physiology* 305(5):F628–F640.

57. Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* 74(20):5383–5392.

58. Searle BC, Turner M, Nesvizhskii AI (2008) Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. *J Proteome Res* 7(1):245–253.

59. Nesvizhskii AI, Keller A, Kolker E, Aebersold R (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* 75(17):4646–4658.

60. Buchfink B, Xie C, Huson DH (2014) Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12(1):59.

61. Mi H, et al. (2016) PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res* 45(D1):D183–D189.

62. 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.

PNAS
www.pnas.org

Supplementary Information for

Genomic adaptations to chemosymbiosis in the deep-sea seep-dwelling tubeworm *Lamellibrachia luymesi* (Siboglinidae, Annelida)

Yuanning Li[1,2,*], Michael G. Tassia[1], Damien S. Waits[1], Viktoria E. Bogantes[1], Kyle T. David[1], Kenneth M. Halanych[1]

Corresponding authors: Yuanning Li, Kenneth M. Halanych
Email: yuanning.li@yale.edu; ken@auburn.edu

This PDF file includes:

Supplementary text
Figs. S1 to S8
Tables S1 to S10

Other supplementary materials for this manuscript include the following:

Dataset S1

Supplementary Information Text

SI Methods

Proteomics characterization.

Proteomic analysis of *Lamellibrachia luymesi* trunk/trophosome tissue was performed by Proteomics & Metabolomics Facility at Colorado State University. Here we restate the protocol provided by Colorado State University. 50 µg total protein was aliquoted from each sample and processed for in-solution trypsin digestion as previously described (1). A total of 0.5µg of peptides were then purified and concentrated using an on-line enrichment column (Waters Symmetry Trap C18 100Å, 5µm, 180 µm ID x 20mm column). Subsequent chromatographic separation was performed on a reverse phase nanospray column (Waters, Peptide BEH C18; 1.7µm, 75 µm ID x 150mm column, 45°C) using a 90 minute gradient: 5%-30% buffer B over 85 minutes followed by 30%-45%B over 5 minutes (0.1% formic acid in ACN) at a flow rate of 350 nanoliters/min. Peptides were eluted directly into the mass spectrometer (Orbitrap Velos Pro, Thermo Scientific) equipped with a Nanospray Flex ion source (Thermo Scientific) and spectra were collected over a m/z range of 400–2000 under positive mode ionization. Ions with charge state +2 or +3 were accepted for MS/MS using a dynamic exclusion limit of 2 MS/MS spectra of a given m/z value for 30 s (exclusion duration of 90 s). The instrument was operated in FT mode for MS detection (resolution of 60,000) and ion trap mode for MS/MS detection with a normalized collision energy set to 35%. Compound lists of the resulting spectra were generated using Xcalibur 3.0 software (Thermo Scientific) with a S/N threshold of 1.5 and 1 scan/group.

Tandem mass spectra were extracted, charge state deconvoluted and deisotoped by ProteoWizard MsConvert v3.0. Spectra from all samples were searched using Mascot (Matrix Science, London, UK; version 2.6.0) against gene models of *Lamellibrachia* host and symbiont genomes (derived from (2)) assuming the digestion enzyme trypsin. Mascot was searched with a fragment ion mass tolerance of 0.80 Da and a parent ion tolerance of 20 PPM. Oxidation of methionine and carbamidomethyl of cysteine were specified in Mascot as variable modifications. Search results from all samples were imported and combined using the probabilistic protein identification algorithms (3) implemented in the Scaffold software (version Scaffold_4.8.4, Proteome Software Inc., Portland, OR) (4). Protein identifications were accepted if they could be established at greater than 99.0% probability and contained at least 1 identified peptide. Protein probabilities were assigned by the Protein Prophet algorithm (5). Proteins that contained similar peptides and could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principles of parsimony.

Manual annotation of gene families with potential interest.

We manually annotated some gene families of interest including hemoglobin gene families, genes related to amino acid synthesize, immunity function and longevity. Hbs and linker sequences (Fig. S4) of interest were obtained from *L. luymesi*genome and assembled siboglinid transcriptomes derived from previous studies (6, 7)via diamond BLASTP (evalue cutoff 1E-5) by using bait Hb and linker bait sequences acquired from *Riftia*Hbs (downloaded from SwissProt Database). Sequences with best hits to target proteins were annotated for protein domain architecture using the Pfam databases included in InterProscan. After manual removal of redundant and incorrect sequences (e.g., sequences are too short or lack of globin domain), we used MAFFT 7.2.15 (8)to align Hb amino acid sequences. Maximum likelihood analyses were performed in IQTree v1.5 (9) under the best-fitting models for associated partition schemes

determined by Modelfinder implemented in IQTree with ultrafast bootstrapping of 1000 replicates.

Discovery of SODs and immunity-related genes largely follows the same workflow as used for Hbs. For immunity genes, targeted genes were additionally processed through the Extract_Homologs2 script used in (10) (available at https://github.com/mtassia/Homolog_identification). We examined major signaling components of the TLR signaling pathway, as well as RLRs, NFkB-associated proteins and interferon-regulatory factors. We only included identification of TLR and RIGs signaling components in the manuscript as other immunity related genes did not clearly reveal any evolutionary patterns of interest across lophotrochozoans (Table S9). Importantly, the Extract_Homologs2 script identifies unique protein sequences within an amino acid dataset that fall within user-defined domain architecture criteria (Table S10). Due to this stringency, the pipeline only identifies the complement of unique proteins for any target family encoded in a genome. Full amino acid sequences for TLRs were placed in a phylogenetic context using the bioinformatic workflow delineated above for Hbs.

Searches for genes related to amino acids synthesis from *Lamellibrachia*, *Lamellibrachia* symbionts, and *Capitella teleta* genomes were performed by using the KEGG2 KAAS genome annotation web server and then visualized by the KEGG Mapper Reconstruct Pathway. A BLASTp search of protein sequences from the genome annotation were queried against the Swiss-Prot database was used to search and supplement for proteins that were missing in the visualized KEGG pathway.
The results of gene alignments, tree files mentioned above were available at the Github Repository (https://github.com/yzl0084/Lamellibrachia-genome).

References

1. Schauer KL, Freund DM, Prenni JE, Curthoys NP (2013) Proteomic profiling and pathway analysis of the response of rat renal proximal convoluted tubules to metabolic acidosis. *American Journal of Physiology-Renal Physiology* 305(5):F628–F640.

2. Li Y, Liles MR, Halanych KM (2018) Endosymbiont genomes yield clues of tubeworm success. *ISME J* 12(11):2785.

3. Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* 74(20):5383–5392.

4. Searle BC, Turner M, Nesvizhskii AI (2008) Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. *J Proteome Res* 7(1):245–253.

5. Nesvizhskii AI, Keller A, Kolker E, Aebersold R (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* 75(17):4646–4658.

6. Li Y, et al. (2017) Phylogenomics of tubeworms (Siboglinidae, Annelida) and comparative performance of different reconstruction methods. *Zool Scr* 46(2):200–213.

7. Waits DS, Santos SR, Thornhill DJ, Li Y, Halanych KM (2016) Evolution of Sulfur Binding by Hemoglobin in Siboglinidae (Annelida) with Special Reference to Bone-Eating Worms, Osedax. *J Mol Evol* 82(4-5):219–229.

8. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30(4):772–780.

9. Chernomor O, von Haeseler A, Minh BQ (2016) Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Syst Biol* 65(6):997–1008.

10. Tassia MG, Whelan NV, Halanych KM (2017) Toll-like receptor pathway evolution in deuterostomes. *Proceedings of the National Academy of Sciences* 114(27):7055–7060.

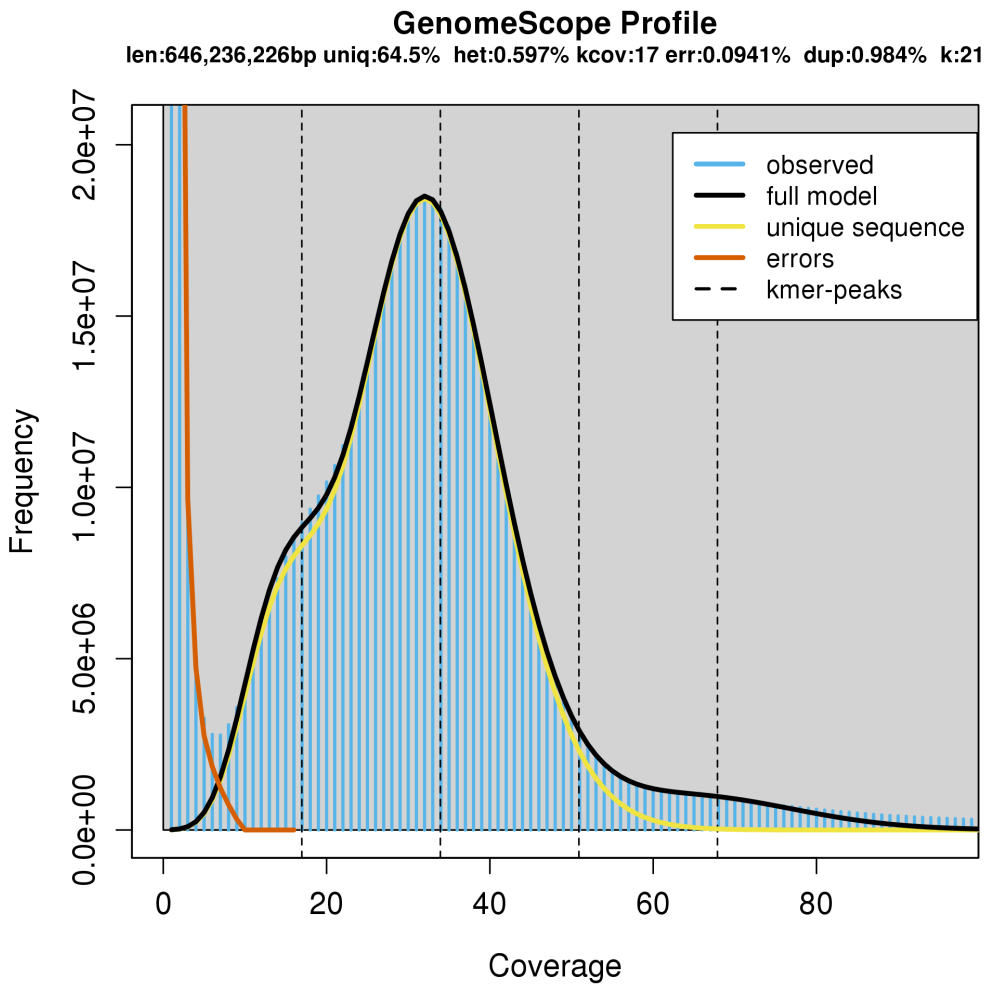

Figure S1. Estimation of genome size, repetitive content and level of heterozygosity from 100 million Illumina paired-end reads.

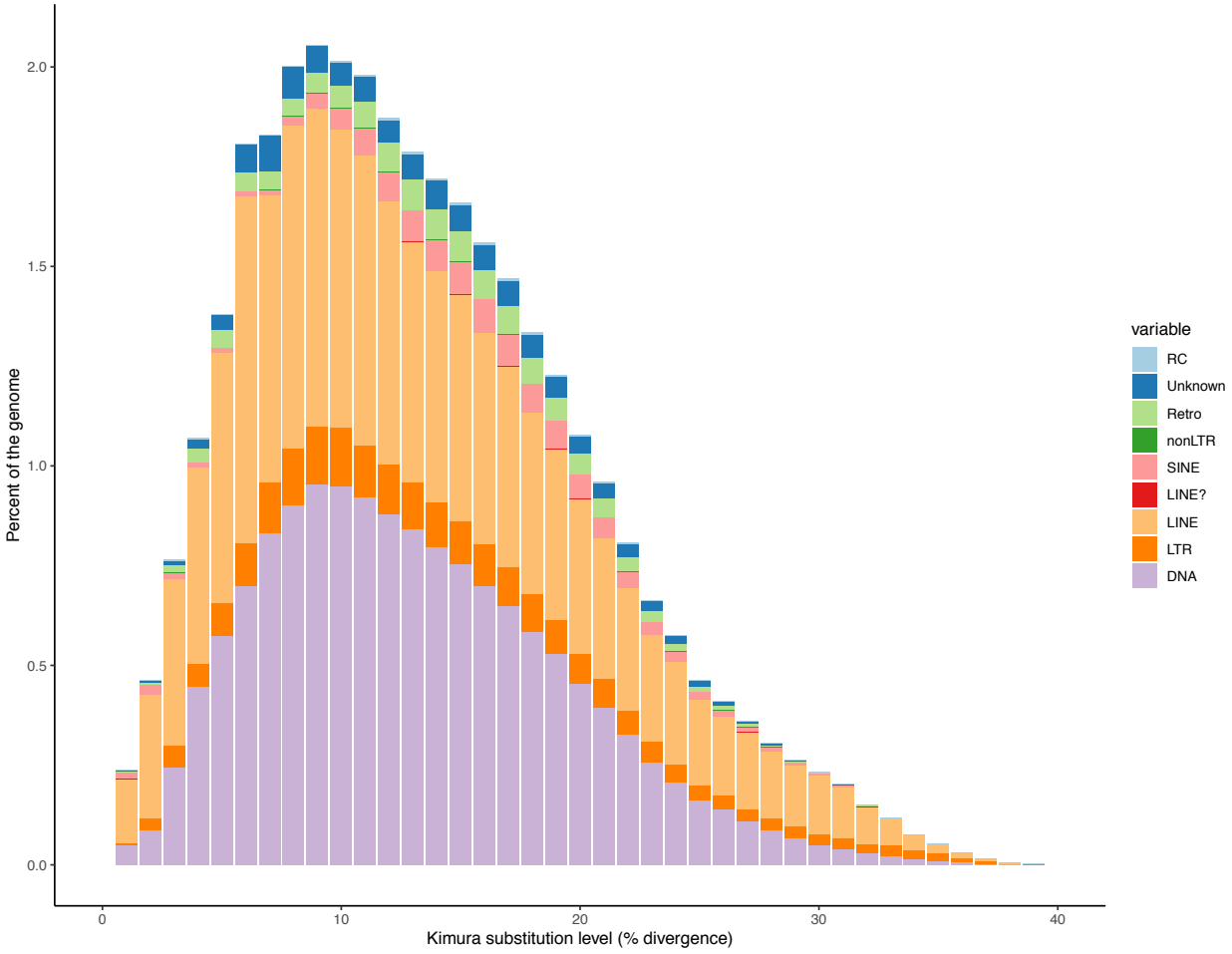

Figure S2. History from major superfamilies of transposable elements in the *Lamellibrachia* genome**.** Kiruma distances are arranged from value 0 representing recent TE copies to 40 for ancient TE insertions.

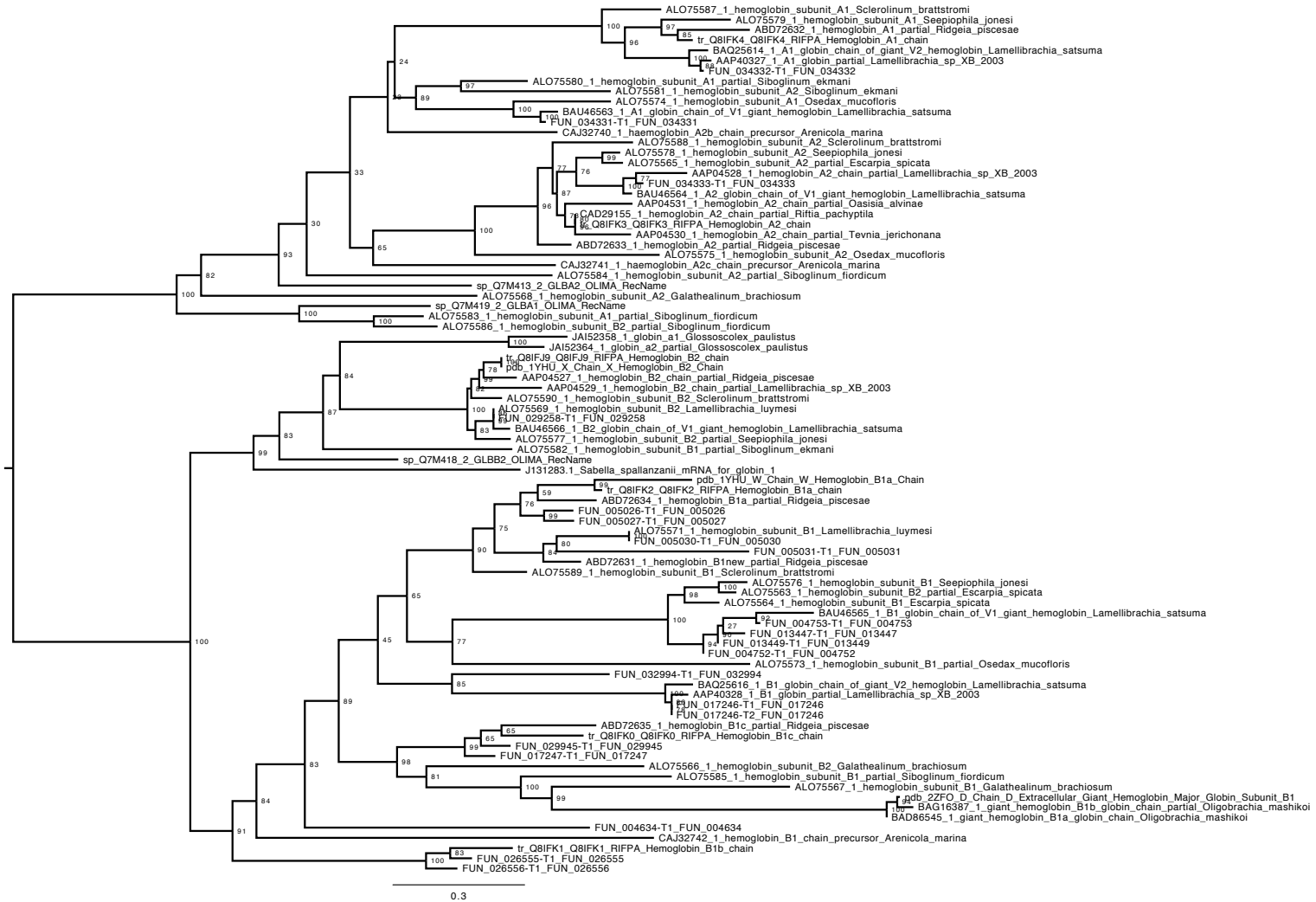

Figure S3. Siboglinid hemoglobin maximum-likelihood tree reconstructed with IQtree with midpoint rooting with 1000 ultrafast bootstraps using LG model. Bootstrap support values are shown at the relevant node. GenBank accession numbers are listed on the terminal nodes.

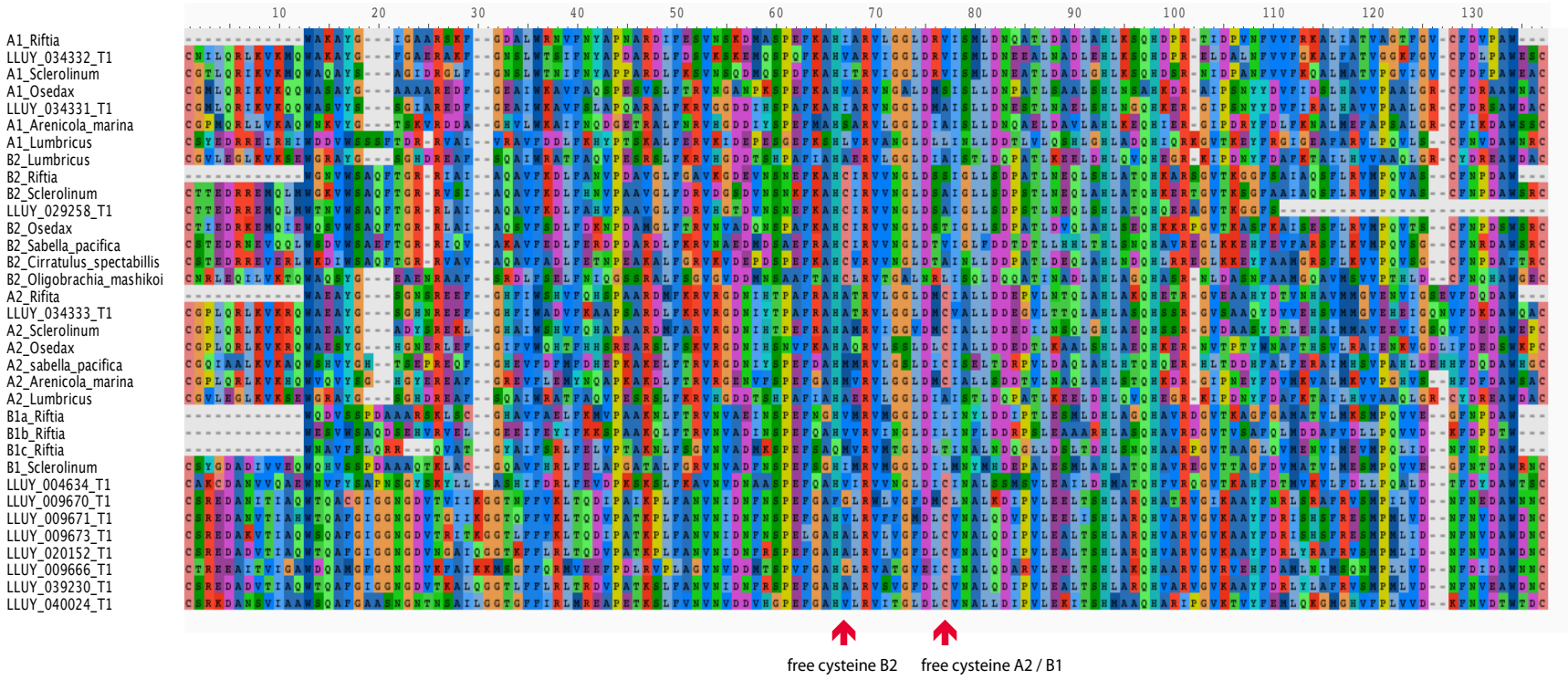

Figure S4. Hemoglobin gene diversity in *Lamellibrachia luymesi*. (Partial alignment of sampled siboglinid Hb subunit A1, A2, B1, B2 sequences. Red arrows indicate positions contain free cysteines or cysteine residues in HB B2 chains, and B1/A2 chains, respectively.

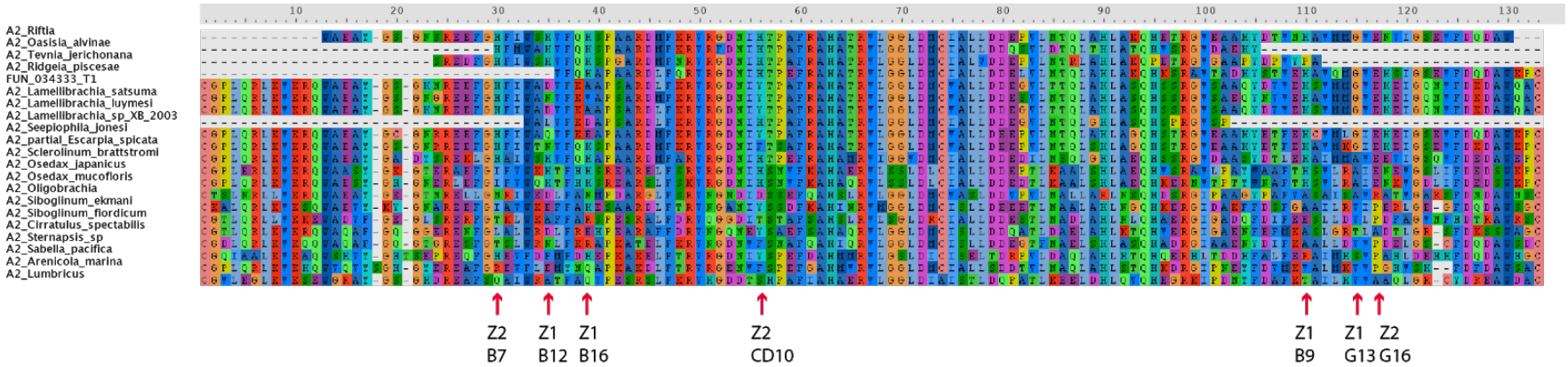

Figure S5. Partial alignment of sampled siboglinid HB subunit A2 sequences. Red arrows indicate amino acid residues at the interface between pairs of A2 chains with zinc moieties for H2S binding.

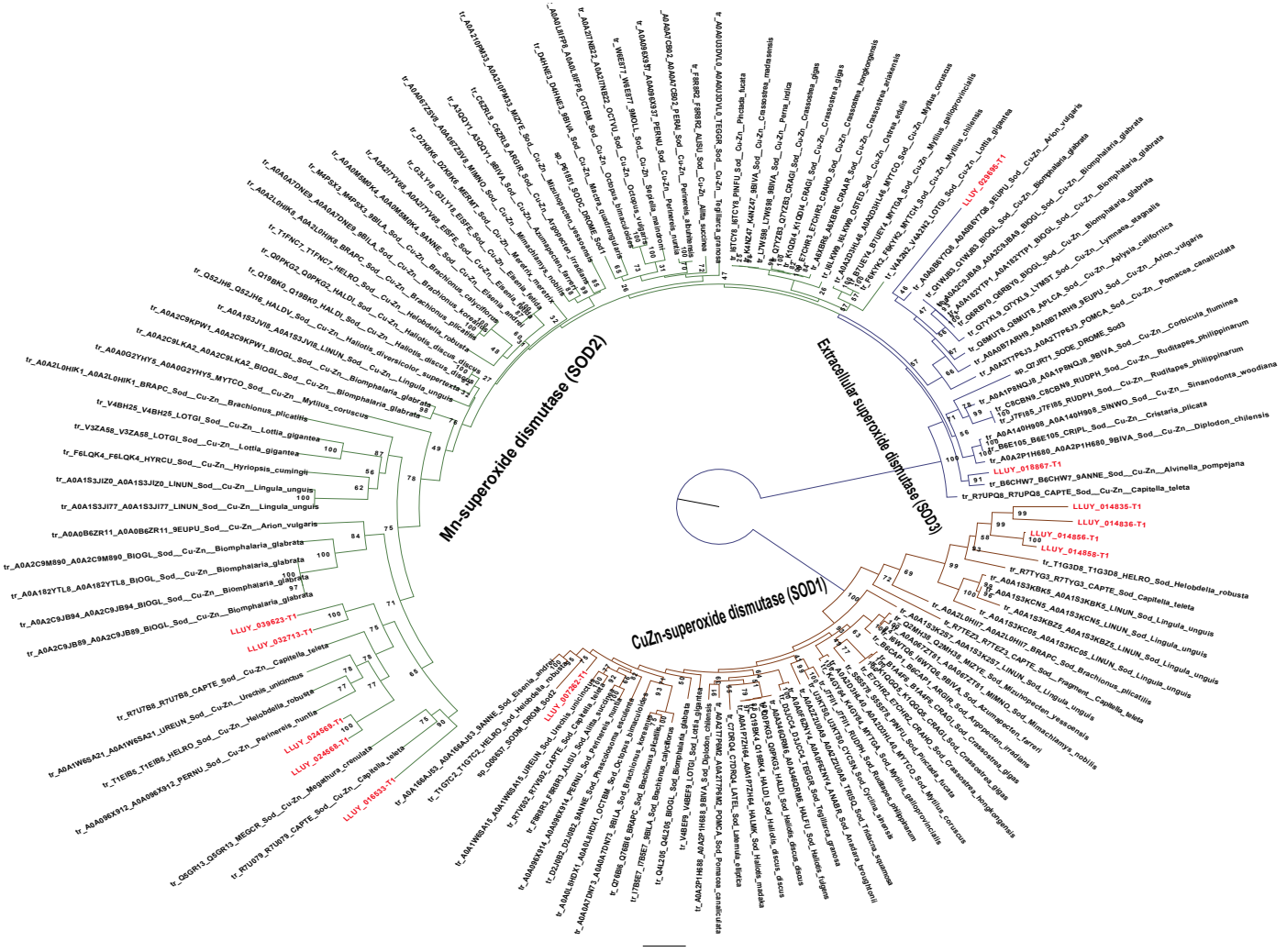

Figure S6. Lophotrochozoan SOD maximum-likelihood tree reconstructed with IQtree with midpoint rooting with 1000 ultrafast bootstraps using LG model. Bootstrap support values are shown at the relevant node. GenBank accession numbers are listed on the terminal nodes. GenBank accession numbers are next to the tip names.

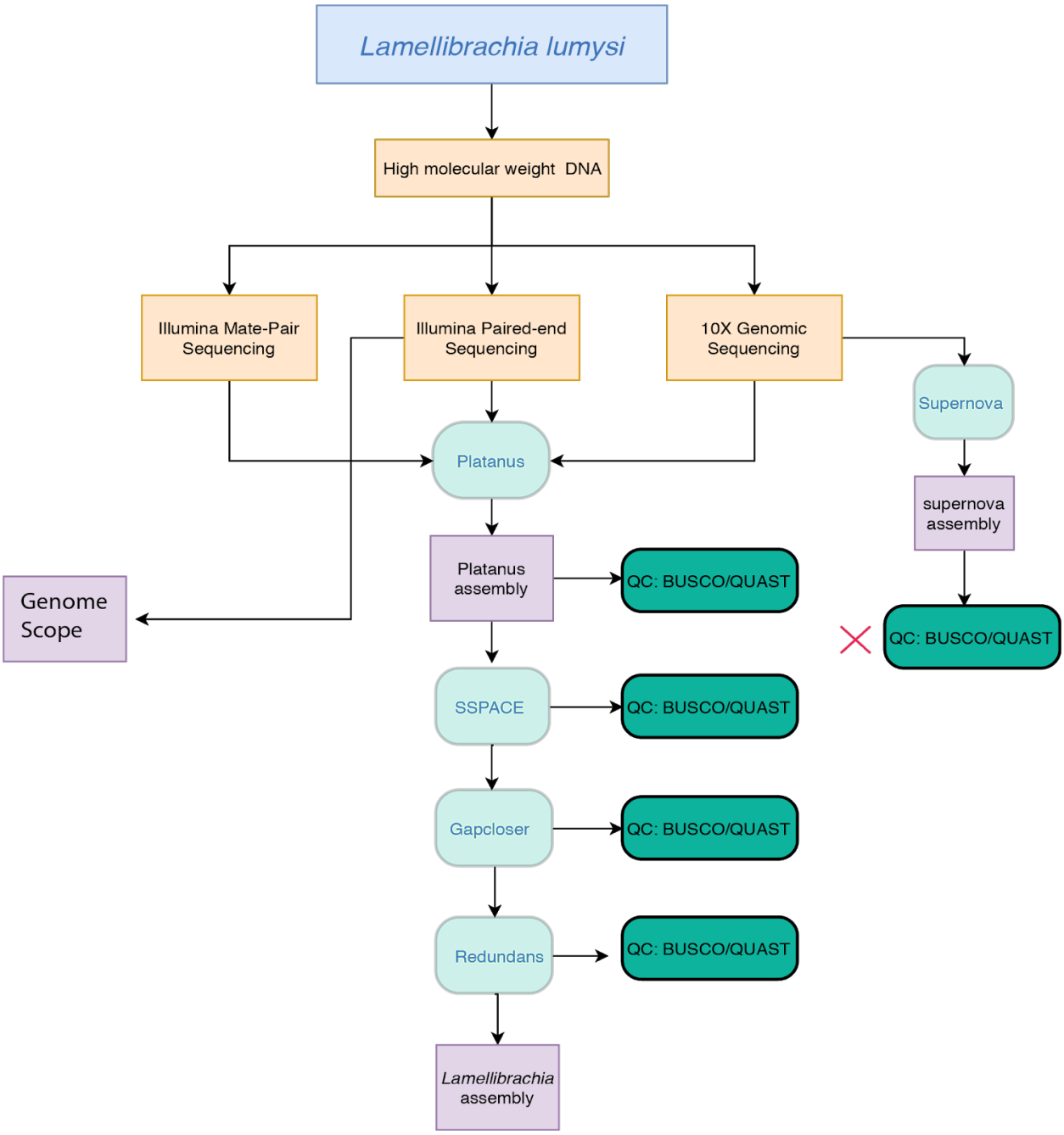

Figure S7. Workflow of *Lamellibrachia luymesi* genome assembly. 10X genomics assembly alone provide worse assembly and failed QC compared to Platanus indicated by red X.

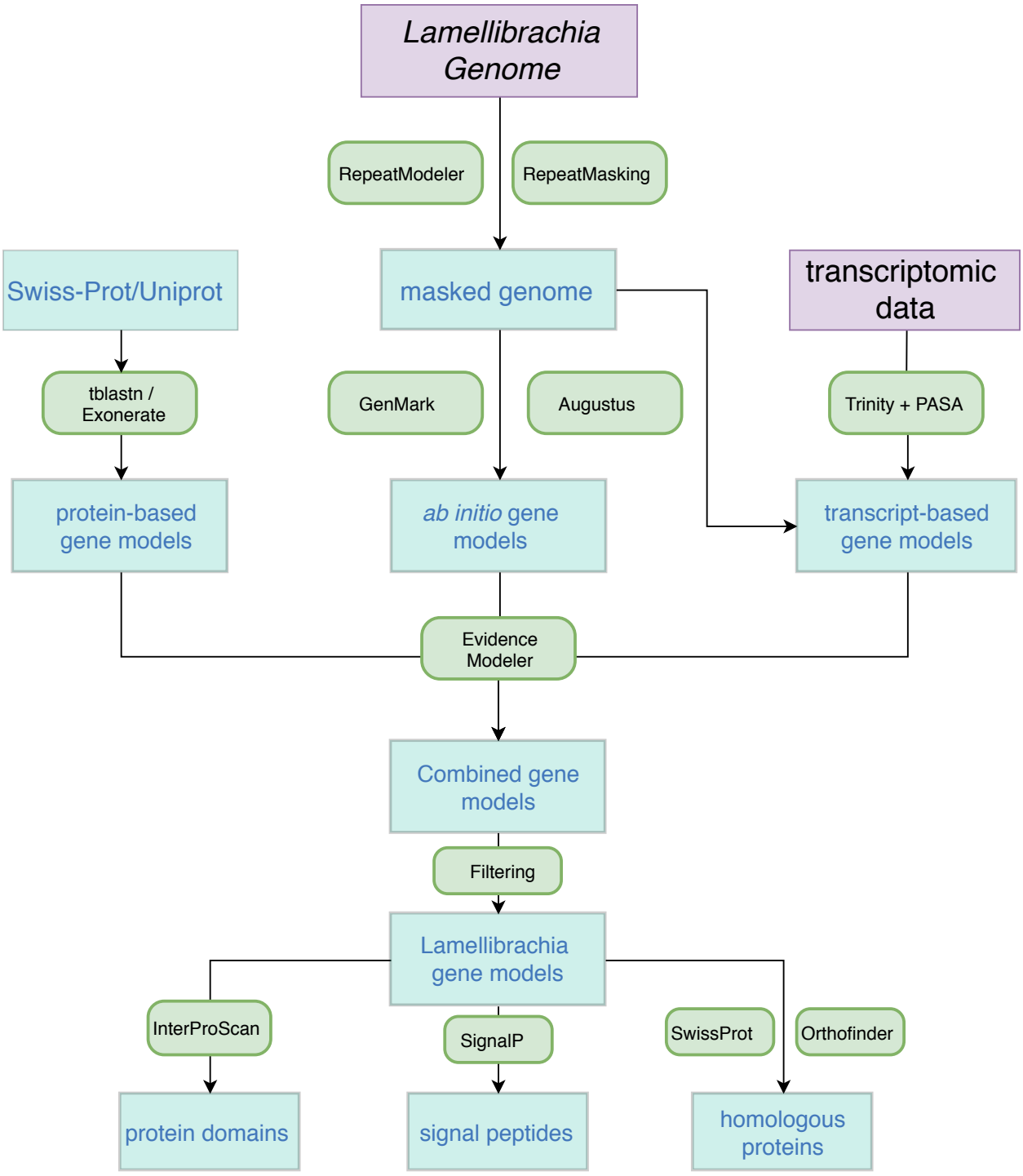

Figure S8. Workflow of *Lamellibrachia* genome annotation pepline using Funannotate.

Table S1. Sequencing information of *Lamellibrachia* luymesi genome.

Tissue	Data Type	Sequencing Chemistry	Total Read Number	Lab Accession	Accession Number	Coverage (X)
Vestimentum	Genomics	10X Genomics	648,546,716	KH-4260-0006	SRR8519110	141.48
Vestimentum	Genomics	180 bp Paired-end	530,601,282	SL84794	SRR8519115	96.43
Vestimentum	Genomics	180 bp Paired-end	318,356,186	SL115013	SRR8519114	57.86
Vestimentum	Genomics	400 bp Paired-end	237,879,494	SL84795	SRR8519113	43.12
Vestimentum	Genomics	750 bp Paired-end	118,008,914	SL84796	SRR8519112	21.40
Vestimentum	Genomics	3-5 kbp Mate-pair	344,803,888	SL85812	SRR8519119	60.77
Vestimentum	Genomics	5-7 kbp Mate-pair	352,639,094	SL85813	SRR8519118	64.04
Plume	Transcriptome	Paired-end	58,660,044	SL85796	SRR8519117	
Trophosome	Transcriptome	Paired-end	75,640,660	SL85798	SRR8519111	
Vestimentum	Transcriptome	Paired-end	50,537,812	SL85797	SRR8519116	

Table S2.
Genome assembly and BUSCO statistics of *Lamellibrachia luymesi* compared to other lophotrochozoan genomes.

Taxon	Species Name	# contigs	Total length	Largest contig	GC (%)	N50	BUSCO (%)				
							Complete	Single	Duplicate	Fragment	Missing
Annelida	*Lamellibrachia lumysi*	11,871	687,711,696	2,117,112	40.16	372,990	95.80	93.00	2.80	2.90	1.30
	Capitella teleta	20,803	333,283,208	1,620,044	40.36	188,402	92.30	87.60	4.70	1.10	6.60
	Hydroides elegans	188,407	1,026,046,400	244,066	35.43	17,725	79.90	53.00	26.90	8.80	11.30
	Helobdella robusta	1,991	235,376,169	13,640,604	32.82	3,060,193	85.30	83.80	1.50	3.70	11.00
Phoronida	*Phoronis australis*	3,983	498,443,662	4,871,659	39.34	655,058	91.90	89.40	2.50	1.30	6.80
Nemertea	*Notospermus geniculatus*	11,108	858,599,399	1,576,180	42.85	239,235	91.90	89.40	2.50	1.30	6.80
Mollusca	*Crassostrea virginica*	10	684,723,884	104,168,038	34.83	75,944,018	90.70	88.10	2.60	0.80	8.50
Mollusca	*Crassostrea gigas*	7,658	557,717,710	1,964,558	33.42	402,213	90.80	85.70	5.10	0.90	8.30
Mollusca	*Bathymodiolus platifrons*	65,662	1,658,191,953	2,790,175	34.17	345,477	89.30	88.00	1.30	2.20	8.50
Mollusca	*Mytilus galloprovincialis*	1,002,334	1,500,149,602	67,529	31.77	3,239	91.20	63.00	28.20	0.90	7.90
Mollusca	*Octopus bimaculoides*	151,674	2,338,188,782	4,064,693	36.06	485,615	85.80	85.30	0.50	3.40	10.80
Mollusca	*Modiolus philippinarum*	74,573	2,629,556,424	715,382	33.96	100,386	85.20	82.70	2.50	4.90	9.90
Mollusca	*Mizuhopecten yessoensis*	82,658	987,568,220	7,498,238	36.52	827,226	89.80	87.80	2.00	1.20	9.00
Mollusca	*Lottia gigantea*	4,469	359,505,668	9,386,848	33.28	1,870,055	91.30	90.20	1.10	0.90	7.80
Brachiopoda	*Lingula anatina*	2,677	406,282,338	2,166,018	36.42	460,090	90.20	70.20	20.00	0.90	8.90
Rotifer	*Aplysia californica*	4,331	927,296,314	6,102,535	40.35	917,541	88.30	87.80	0.50	1.60	10.10

Table S3.
Proteomics and genome assemblies used in comparative analyses.

Taxon	Species	Genome source	RefSeq assembly accession
Annelida	*Lamellibrachia luymsi*	This study	SDWI00000000
	Capitella teleta	NCBI	GCA_000328365.1
	Helobdella robusta	NCBI	GCA_000326865.1
Mollusca	*Lottia gigantea*	NCBI	GCA_000327385.1
	Octopus bimaculoides	NCBI	GCA_001194135.1
	Chlamys farreri	NCBI	
	Bathymodiolus platifrons	NCBI	GCA_002080005.1
	Biomphalaria glabrata	NCBI	
	Mizuhopecten yessoensis	NCBI	GCA_002113885.2
	Modiolus philippinarum	NCBI	GCA_000457365.1
	Patinopecten yessoensis	NCBI	
	Crassostrea gigas	NCBI	GCF_000297895.1
	Crassostrea virginica	NCBI	GCA_002022765.4
Nemertea	*Notospermus geniculatus*	NCBI	GCA_002633025.1
Phoronida	*Phoronis australis*	NCBI	GCA_002633005.1
Brachiopoda	*Lingula anatina*	NCBI	GCA_001039355.2
Flatworm	*Schistosoma mansoni*	NCBI	GCA_000237925.2
	Schmidtea mediterranea	NCBI	GCA_002600895.1
	Macrostomum lignano	NCBI	GCA_002269645.1
	Echinococcus multilocularis	NCBI	GCA_000469725.3
Rotifera	*Aplysia californica*	NCBI	GCA_000002075.2
Ecdysozoa	*Diphania pulex*	NCBI	GCA_000187875.1
	Drosophila melanogaster	NCBI	GCA_000001215.4

Table S4.
Repetitive element contained in the *Lamellibrachia luymesi* genome

	Subclass	Number of elements	length occupied (bp)	percentage of sequence
SINEs		56,783	8,181,892	1.19
LINEs		307,497	96,235,311	13.99
	LINE1	1,333	337,473	0.05
	LINE2	51,791	18,267,908	2.66
	L3/CR1	83,528	33,202,610	4.83
LTR elements		70,108	17,360,126	2.52
DNA elements		442,010	101,825,130	14.81
	hAT-Charlie	6,246	1,710,664	0.25
	TcMar-Tigger	4,078	1,496,397	0.22
Unclassified		95,876	17,185,149	2.5
Tot. interspersed repeats			240,787,608	35.01
Small RNA		101	27,655	0
Satellites		813	376,172	0.05
Simple repeats		211,949	13,781,625	2
Low complexity		8,125	682,495	0.1

Table S5.
PANTHER gene family annotation of gene families that are under expansion or contraction as identified by CAFE.

Orthology Group	Number of gain and loss	PANTHER gene family	PANTHER annotation
OG0000040	+23*	PTHR44025	LOW-DENSITY LIPOPROTEIN RECEPTOR-RELATED PROTEIN
OG0000044	+16*	PTHR12011	G-PROTEIN COUPLED RECEPTOR
OG0000059	+25*	PTHR11177	CHITINASE
OG0000062	+20*	PTHR28576	PIGGYBAC TRANSPOSABLE ELEMENT-DERIVED PROTEIN
OG0000075	+20*	PTHR19325	COMPLEMENT COMPONENT-RELATED SUSHI DOMAIN-CONTAINING
OG0000091	+27*	PTHR11119	XANTHINE-URACIL / VITAMIN C PERMEASE FAMILY MEMBER
OG0000118	+30*	PTHR10877	POLYCYSTIN-RELATED
OG0000124	+9*	PTHR13802	MUCIN 4-RELATED
OG0000128	+23*	PTHR10131	TNF RECEPTOR ASSOCIATED FACTOR
OG0000137	+14*	PTHR11908	XANTHINE DEHYDROGENASE
OG0000146	+18*	PTHR11709	MULTI-COPPER OXIDASE
OG0000150	+20*	PTHR14647	GALACTOSE-3-O-SULFOTRANSFERASE
OG0000155	+13*	PTHR23033	BETA1,3-GALACTOSYLTRANSFERASE
OG0000171	+17*	PTHR10283	SOLUTE CARRIER FAMILY 13 MEMBER
OG0000180	+16*	PTHR24039	FIBRILLIN
OG0000183	+21*	PTHR14453	PARP/ZINC FINGER CCCH TYPE DOMAIN CONTAINING PROTEIN
OG0000184	+35*	PTHR24221	ABC TRANSPORTER
OG0000198	+27*	PTHR24033	FAMILY NOT NAMED
OG0000219	+24*	PTHR10579	CALCIUM-ACTIVATED CHLORIDE CHANNEL REGULATOR
OG0000238	+26*	PTHR11039	NEBULIN
OG0000242	+11*	PTHR34415	FAMILY NOT NAMED
OG0000247	+15*	PTHR14453	PARP/ZINC FINGER CCCH TYPE DOMAIN CONTAINING PROTEIN
OG0000250	+29*	PTHR13800	TRANSIENT RECEPTOR POTENTIAL CATION CHANNEL, SUBFAMILY M, MEMBER 6
OG0000262	+16*	PTHR43645	UPF0214 PROTEIN YFEW
OG0000268	+13*	PTHR43998	FILAMIN
OG0000277	+13*	PTHR44131	CUB DOMAIN-CONTAINING PROTEIN
OG0000287	+18*	PTHR44097	PROTEIN SERRATE
OG0000290	+7*	PTHR10166	VOLTAGE-DEPENDENT CALCIUM CHANNEL SUBUNIT ALPHA-2/DELTA-RELATED
OG0000293	+26*	PTHR22605	AAA+ ATPASE, CORE DOMAIN-CONTAINING PROTEIN
OG0000304	+12*	PTHR23097	TUMOR NECROSIS FACTOR RECEPTOR SUPERFAMILY MEMBER
OG0000305	+17*	PTHR23232	KRAB DOMAIN C2H2 ZINC FINGER
OG0000307	+10*	PTHR18966	IONOTROPIC GLUTAMATE RECEPTOR
OG0000318	+11*	PTHR12622	DELTEX-RELATED
OG0000322	+17*	PTHR23024	MEMBER OF 'GDXG' FAMILY OF LIPOLYTIC ENZYMES
OG0000324	+8*	PTHR11106	GANGLIOSIDE INDUCED DIFFERENTIATION ASSOCIATED PROTEIN 2-RELATED

OG0000337	+24*	PTHR10887	DNA2/NAM7 HELICASE FAMILY
OG0000342	+25*	PTHR16897	CARNOSINE N-METHYLTRANSFERASE
OG0000347	+11*	PTHR10796	PATCHED-RELATED
OG0000357	+10*	PTHR23302	TRANSMEMBRANE CHANNEL-RELATED
OG0000365	+7*	PTHR12042	LACTOSYLCERAMIDE 4-ALPHA-GALACTOSYLTRANSFERASE ALPHA- 1,4-GALACTOSYLTRANSFERASE
OG0000373	+23*	PTHR31649	LD46221P-RELATED
OG0000378	+39*	PTHR13715	RYANODINE RECEPTOR AND IP3 RECEPTOR
OG0000388	+16*	PTHR15600	SACSIN
OG0000398	+13*	PTHR11697	GENERAL TRANSCRIPTION FACTOR 2-RELATED ZINC FINGER PROTEIN
OG0000423	+8*	PTHR44014	FAMILY NOT NAMED
OG0000432	+14*	PTHR14454	GRB2-ASSOCIATED AND REGULATOR OF MAPK PROTEIN
OG0000460	+10*	PTHR43905	CONTACTIN
OG0000465	+11*	PTHR23130	FERRIC-CHELATE REDUCTASE
OG0000468	+15*	PTHR16897	CARNOSINE N-METHYLTRANSFERASE
OG0000477	+10*	PTHR11616	SODIUM/CHLORIDE DEPENDENT TRANSPORTER
OG0000491	+20*	PTHR11046	OLIGORIBONUCLEASE, MITOCHONDRIAL
OG0000500	+12*	PTHR24106	CASPASE RECRUITMENT DOMAIN-CONTAINING PROTEIN 8/NACHT, LRR AND PYD DOMAINS-CONTAINING PROTEIN
OG0000524	+50*	PTHR19277	PENTRAXIN
OG0000555	+9*	PTHR31009	S-ADENOSYL-L-METHIONINE:CARBOXYL METHYLTRANSFERASE FAMILY PROTEIN
OG0000562	+8*	PTHR15698	PHYTANOYL-COA HYDROXYLASE-INTERACTING PROTEIN
OG0000570	+27*	PTHR10697	MAMMALIAN EPENDYMIN-RELATED PROTEIN 1
OG0000578	+21*	PTHR33748	FAMILY NOT NAMED
OG0000586	+14*	PTHR34153	SI:CH211-262H13.3
OG0000592	+41*	PTHR19325	COMPLEMENT COMPONENT-RELATED SUSHI DOMAIN-CONTAINING
OG0000613	+11*	PTHR44854	FIBROCYSTIN-
OG0000617	+20*	PTHR19325	COMPLEMENT COMPONENT-RELATED SUSHI DOMAIN-CONTAINING
OG0000640	+28*	PTHR23145	NUCLEOSOMAL BINDING PROTEIN 1
OG0000656	+9*	PTHR19297	GLYCOSYLTRANSFERASE 14 FAMILY MEMBER
OG0000687	+13*	PTHR24243	G-PROTEIN COUPLED RECEPTOR
OG0000715	+10*	PTHR10796	PATCHED-RELATED
OG0000728	+22*	PTHR45240	ZINC METALLOPROTEINASE NAS
OG0000789	+14*	PTHR10773	DNA-DIRECTED RNA POLYMERASES I, II, AND III SUBUNIT RPABC2
OG0000849	+37*	PTHR12673	FACIOGENITAL DYSPLASIA PROTEIN
OG0000862	+23*	PTHR11514	MYC
OG0000917	+10*	PTHR31569	ZINC FINGER SWIM DOMAIN-CONTAINING PROTEIN
OG0000927	+21*	PTHR44252	CARBONYL REDUCTASE NADPH
OG0000938	+17*	PTHR11360	MONOCARBOXYLATE TRANSPORTER
OG0000943	+39*	PTHR13954	IRE1-RELATED
OG0000987	+10*	PTHR24280	CYTOCHROME P450 20A1

OG0001012	+19*	PTHR16897	CARNOSINE N-METHYLTRANSFERASE
OG0001096	+11*	PTHR11799	PARAOXONASE
OG0001166	+10*	PTHR16897	HISTAMINE N-METHYLTRANSFERASE
OG0001243	+12*	PTHR11903	PROSTAGLANDIN G/H SYNTHASE
OG0001293	+52*	PTHR23194	PYGOPUS
OG0001377	+10*	PTHR31513	GLYCINE-RICH PROTEIN
OG0001383	+9*	PTHR20956	HEH2P
OG0001422	+41*	PTHR23259	RIDDLE
OG0001533	+24*	PTHR37558	FAMILY NOT NAMED
OG0001594	+10*	PTHR37445	FAMILY NOT NAMED
OG0001671	+22*	PTHR10199	THROMBOSPONDIN
OG0001730	+7*	PTHR24106	NACHT, LRR AND PYD DOMAINS-CONTAINING PROTEIN
OG0001853	+6*	PTHR13627	FUKUTIN RELATED PROTEIN
OG0001862	+10*	PTHR11697	GENERAL TRANSCRIPTION FACTOR 2-RELATED ZINC FINGER PROTEIN
OG0001954	+10*	PTHR15031	CARTILAGE INTERMEDIATE LAYER PROTEIN CLIP
OG0002014	+17*	PTHR12419	OTU DOMAIN CONTAINING PROTEIN
OG0002286	+6*	NONE	NA
OG0002380	+6*	NONE	NA
OG0002381	+22*	PTHR11046	OLIGORIBONUCLEASE, MITOCHONDRIAL
OG0002455	+11*	PTHR35558	FAMILY NOT NAMED
OG0000077	-10*	PTHR14918	PROTEIN SZT2

Table S6.

Key genes of host genes identified as proteins from proteomic analysis.

Function System	Feature ID	Function
Proteosome	LLUY_028786-T1	PSMA6; 20S proteasome subunit alpha 1 [EC:3.4.25.1]
	LLUY_012964-T1	PSMA2; 20S proteasome subunit alpha 2 [EC:3.4.25.1]
	LLUY_008627-T1	PSMA4; 20S proteasome subunit alpha 3 [EC:3.4.25.1]
	LLUY_027290-T1	PSMA7; 20S proteasome subunit alpha 4 [EC:3.4.25.1]
	LLUY_003537-T1	PSMA5; 20S proteasome subunit alpha 5 [EC:3.4.25.1]
	LLUY_040090-T1	PSMA1; 20S proteasome subunit alpha 6 [EC:3.4.25.1]
	LLUY_012936-T1	PSMA3; 20S proteasome subunit alpha 7 [EC:3.4.25.1]
	LLUY_023805-T1	PSMB3; 20S proteasome subunit beta 1
	LLUY_029937-T1	PSMB3; 20S proteasome subunit beta 3
	LLUY_002855-T1	PSMB2; 20S proteasome subunit beta 4
	LLUY_015520-T1	PSMB5; 20S proteasome subunit beta 5
	LLUY_022140-T1	PSMB1; 20S proteasome subunit beta 6
	LLUY_022807-T1	PSMB4; 20S proteasome subunit beta 7
	LLUY_003098-T1	SMD2, RPN1; 26S proteasome regulatory subunit N1
	LLUY_026382-T1	PSMD7, RPN8; 26S proteasome regulatory subunit N8
	LLUY_017796-T1	EIF3E, INT6; translation initiation factor 3 subunit E
	LLUY_014232-T1	HSPA1s; heat shock 70kDa protein 1/2/6/8
	LLUY_037621-T1	HSP90A, htpG; molecular chaperone HtpG
	LLUY_000099-T1	HSP90A, htpG; molecular chaperone HtpG
	LLUY_000099-T1	RAD23, HR23; UV excision repair protein RAD23
Lysosome	LLUY_033571-T1	cathepsin C [EC:3.4.14.1]
	LLUY_007735-T1, LLUY_007737-T1, LLUY_036642-T1, LLUY_036643-T1	cathepsin B [EC:3.4.22.1]
	LLUY_009810-T1, LLUY_026908-T1	cathepsin L [EC:3.4.22.15]
	LLUY_012982-T1	legumain [EC:3.4.22.34]
	LLUY_007693-T1	cathepsin F [EC:3.4.22.41]
	LLUY_016715-T1	cathepsin D [EC:3.4.23.5]
	LLUY_023349-T2, LLUY_023349-T1	clathrin heavy chain
	LLUY_017989-T1	lysosomal-associated membrane protein 1/2
	LLUY_024617-T1	cathepsin X [EC:3.4.18.1]
	LLUY_002319-T1	lysosomal alpha-mannosidase [EC:3.2.1.24]
	LLUY_004297-T1	lysosomal alpha-glucosidase [EC:3.2.1.20]
	LLUY_005359-T1	saposin
	LLUY_016480-T1	lysosome membrane protein 2
	LLUY_003890-T1	cathepsin A (carboxypeptidase C) [EC:3.4.16.5]
Longevity regulating pathway	LLUY_007262-T1, LLUY_014836-T1, LLUY_014856-T1	SOD2; superoxide dismutase, Fe-Mn family [EC:1.15.1.1]
	LLUY_018867-T1	SOD1; superoxide dismutase, Cu-Zn family [EC:1.15.1.1]

Table S7.

Key genes of symbiont genes identified as proteins from proteomic analysis.

Function System	Feature ID	Length (bp)	Function
rTCA Cycle	Lamellibrachia_symbiont.peg.150	1524	Fumarate hydratase class I, aerobic (EC 4.2.1.2)
	Lamellibrachia_symbiont.peg.2185	1911	2-oxoglutarate oxidoreductase, alpha subunit (EC 1.2.7.3)
	Lamellibrachia_symbiont.peg.2186	954	2-oxoglutarate oxidoreductase, beta subunit (EC 1.2.7.3)
	Lamellibrachia_symbiont.peg.2943	987	Malate dehydrogenase (EC 1.1.1.37)
	Lamellibrachia_symbiont.peg.986	1161	ATP citrate lyase beta chain (EC 4.3.1.8)
	Lamellibrachia_symbiont.peg.987	870	ATP citrate lyase alpha chain (EC 4.3.1.8)
	Lamellibrachia_symbiont.peg.2924	1167	2-oxoglutarate oxidoreductase, alpha subunit (EC 1.2.7.3)
	Lamellibrachia_symbiont.peg.2925	999	2-oxoglutarate oxidoreductase, beta subunit (EC 1.2.7.3)
	Lamellibrachia_symbiont.peg.2926	570	2-oxoglutarate oxidoreductase, gamma subunit (EC 1.2.7.3)
Calvin Cycle	Lamellibrachia_symbiont.peg.2754	1389	Ribulose bisphosphate carboxylase (EC 4.1.1.39)
	Lamellibrachia_symbiont.peg.2757	804	Rubisco activation protein CbbQ
	Lamellibrachia_symbiont.peg.2758	2256	Rubisco activation protein CbbO
Sulfer Oxidataion	Lamellibrachia_symbiont.peg.303	942	Dissimilatory sulfite reductase, beta subunit (EC 1.8.99.3)
Nitrogen Mitobolism	Lamellibrachia_symbiont.peg.724	1602	Respiratory nitrate reductase beta chain (EC 1.7.99.4Denitrifying reductase gene clusters; Nitrate and nitrite ammonification
	Lamellibrachia_symbiont.peg.725	3762	Respiratory nitrate reductase alpha chain (EC 1.7.99.4)
Adhesion-related proteins	Lamellibrachia_symbiont.peg.2820	978	Ankyrin
	Lamellibrachia_symbiont.peg.2856	3813	Fibronectin type III domain protein
Oxidative stress	Lamellibrachia_symbiont.peg.1649	582	Superoxide dismutase [Fe] (EC 1.15.1.1) (FeSOD)
	Lamellibrachia_symbiont.peg.2936	489	Rubrerythrin

Table S8.
Lamellibrachia luymesi Hb sequences identified that are highly expressed in the trophosome tissue or as proteins from proteomic data.

HBs	*Lamellibrachia luymsei* Hbs	Differential expressed	Mass spectrum
A1 Chain	LLUY_034331-T1	*	*
	LLUY_034332-T1	*	*
A2 Chain	LLUY_034333-T1	*	*
B2 Chain	LLUY_029258-T1	x	*
B1 Chain	LLUY_004752-T1	*	*
	LLUY_004753-T1	*	x
	LLUY_005026-T1	x	x
	LLUY_005027-T1	x	x
	LLUY_005030-T1	*	x
	LLUY_005031-T1	x	x
	LLUY_009666-T1	x	x
	LLUY_009670-T1	x	x
	LLUY_009671-T1	*	x
	LLUY_009673-T1	x	x
	LLUY_013447-T1	*	*
	LLUY_013449-T1	*	x
	LLUY_017246-T1	*	*
	LLUY_017247-T1	*	*
	LLUY_020152-T1	*	x
	LLUY_026555-T1	x	x
	LLUY_026556-T1	*	*
	LLUY_029945-T1	x	x
	LLUY_032994-T1	x	x
	LLUY_038743-T1	x	x
	LLUY_039230-T1	x	x
	LLUY_040024-T1	x	x
	LLUY_004634-T1	*	x
Linkers	LLUY_002344-T1	*	*
	LLUY_024479-T1	*	x
	LLUY_026441-T1	*	*
	LLUY_026443-T1	*	*

*: Positive; x: negative.
Sequences with putative free-cystenine were colored as red.

Table S9.
Number of unique TLR proteins encoded in lophotrochozoan genomes.

Taxon	Species	Number of TLR identified	Number of RLR identified
Annelida	*Lamellibrachia luymesi*	33	2
	Capitella telata	5	3
	Helobdella robusta	4	3
Mollusca	*Bathymodiolus platifrons*	61	7
	Crassostrea gigas	61	11
	Modiolus philippinarum	90	2
	Mizuhopecten yessoensis	30	4
	Octopus bimaculoides	5	5
	Patinopecten yessoensis	22	4
	Lottia gigantea	7	3
	Crassostrea virginica	109	8
Nemertea	*Notospermus geniculatus*	5	4
Phoronida	*Phoronis australis*	23	3
Brachiopoda	*Lingula anatina*	46	13
Flatworm	*Biomphalaria glabrata*	17	6
Rotifera	*Aplysia californica*	15	2
Vertebrata	*Homo sapiens*	11	3

Table S10.
Domain requirements for identifying components of TLR pathway.

Protein	Domain Requirements
TLR/TOLL	TIR+LRR(≥3)
MYD88	TIR+DEATH
SARM1	TIR+SAM(2)
DDX58	CARD(2)+Helicase_ATP_binding+Helicase_C_Terminal+Rig1_Regulatory_Domain
DHX58	Helicase_ATP_binding+Helicase_C_Terminal+Rig1_Regulatory_Domain
IFIH1	CARD(2)+Helicase_ATP_binding+Helicase_C_Terminal+Rig1_Regulatory_Domain
TRAF	Zn_Finger+MATH/TRAF or Zn_Finger+WD40_repeats(≥3)
NLR	(x)+NACHT+LRR(≥3)
IKK	Kinase
IKB	ANK(≥3)
NFKB	RHD+ANK(≥3)+DEATH
NEMO	NEMO/Coiled_coil
IRF	Interferon_regulatroy_factor_DNA_binding_domain