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

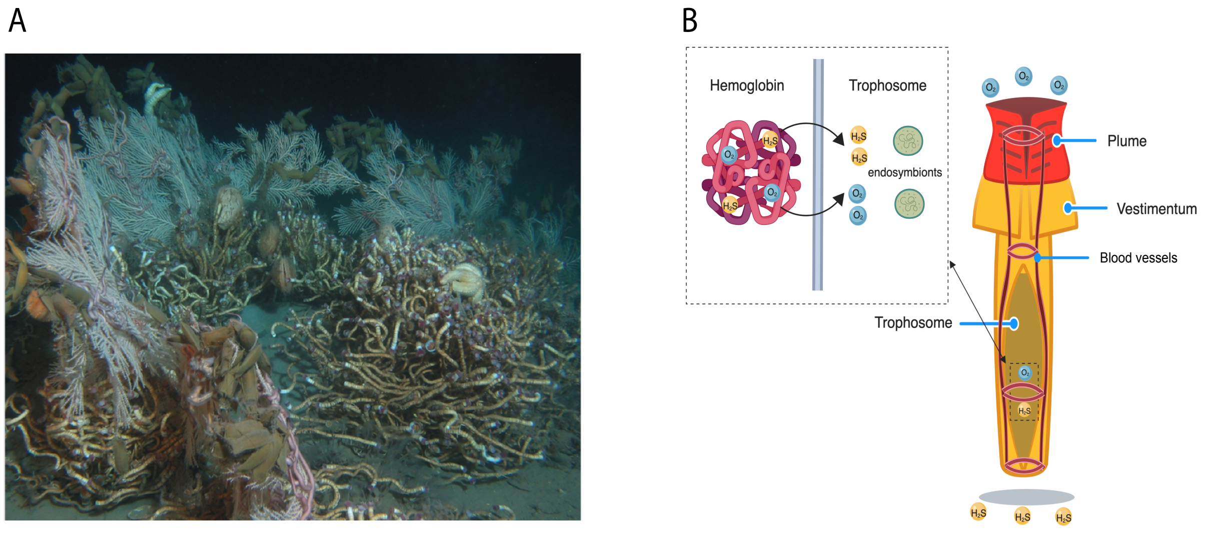


Fig. 1.*Lamellibrachia luymesi*. (A). Seep habitat in the Gulf of Mexico. (B). Diagram of adult *L. luymesi* worm to model O_2 and H_2S transport to symbionts in trophosome by hemoglobin molecules. The hemoglobin model was created with the help of Biorender (https://biorender.com/).

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.

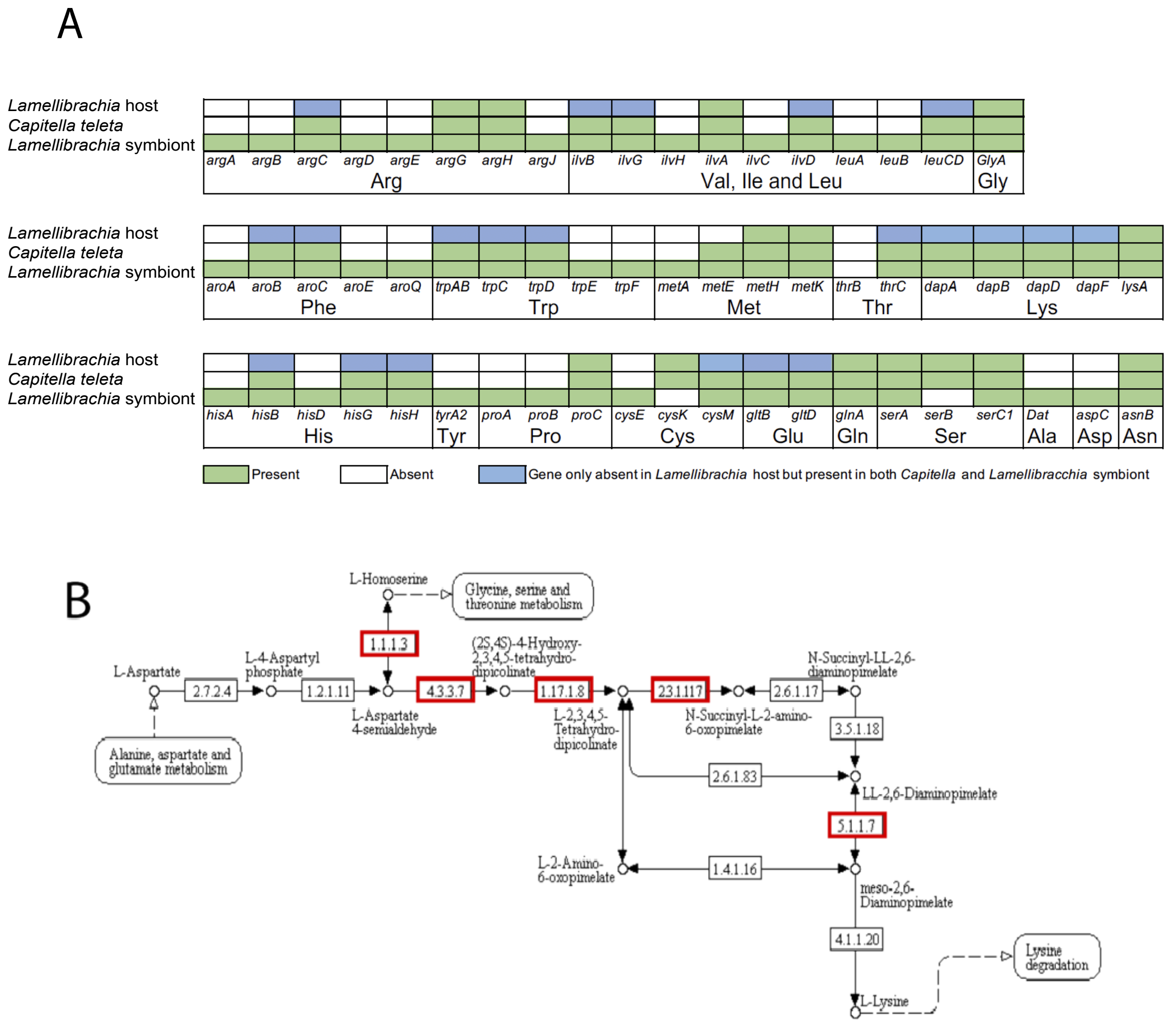
Using comparative genomics, transcriptomic and proteomic analyses on the tubeworm *Lamellibrachia luymesi*, we provide evidence for genetic pathways and novel candidate genes which may underlie the extraordinary characteristics of tubeworm symbioses. In particular, we focus on nutrition mode, hemoglobin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


A
B
L-Homoserine
Glycine, serine and threonine metabolism
1.1.1.3
(2S,4S)-4-Hydroxy-2,3,4,5-tetrahydro-dipicolinate
N-Succinyl-LL-2,6-diaminopimelate
L-Aspartate
L-4-Aspartyl phosphate
2.7.2.4
1.2.1.11
4.3.3.7
1.17.1.8
2.3.1.117
2.6.1.17
L-Aspartate 4-semialdehyde
L-2,3,4,5-Tetrahydro-dipicolinate
N-Succinyl-L-2-amino-6-oxopimelate
3.5.1.18
Alanine, aspartate and glutamate metabolism
2.6.1.83
LL-2,6-Diaminopimelate
5.1.1.7
L-2-Amino-6-oxopimelate
1.4.1.16
meso-2,6-Diaminopimelate
4.1.1.20
Lysine degradation
L-Lysine

	Arg								Val, Ile and Leu									Gly
	argA	*argB*	*argC*	*argD*	*argE*	*argG*	*argH*	*argJ*	*ilvB*	*ilvG*	*ilvH*	*ilvA*	*ilvC*	*ilvD*	*leuA*	*leuB*	*leuCD*	*GlyA*
Lamellibrachia host	□	□	🟦	□	□	🟩	🟩	□	🟦	🟦	□	🟩	□	🟦	□	□	🟦	🟩
Capitella teleta	□	□	🟩	□	□	🟩	🟩	□	🟩	🟩	□	🟩	□	🟩	□	□	🟩	🟩
Lamellibrachia symbiont	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩

	Phe					Trp					Met				Thr		Lys				
	aroA	*aroB*	*aroC*	*aroE*	*aroQ*	*trpAB*	*trpC*	*trpD*	*trpE*	*trpF*	*metA*	*metE*	*metH*	*metK*	*thrB*	*thrC*	*dapA*	*dapB*	*dapD*	*dapF*	*lysA*
Lamellibrachia host	□	🟦	🟦	□	□	🟦	🟦	🟦	□	□	□	□	🟩	🟩	□	🟦	🟦	🟦	🟦	🟦	🟩
Capitella teleta	□	🟩	🟩	□	□	🟩	🟩	🟩	□	□	□	🟩	🟩	🟩	□	🟩	🟩	🟩	🟩	🟩	🟩
Lamellibrachia symbiont	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	□	🟩	🟩	🟩	🟩	🟩	🟩

	His					Tyr	Pro			Cys			Glu		Gln	Ser			Ala	Asp	Asn
	hisA	*hisB*	*hisD*	*hisG*	*hisH*	*tyrA2*	*proA*	*proB*	*proC*	*cysE*	*cysK*	*cysM*	*gltB*	*gltD*	*glnA*	*serA*	*serB*	*serC1*	*Dat*	*aspC*	*asnB*
Lamellibrachia host	□	🟦	□	🟦	🟦	□	□	□	🟩	□	🟩	🟦	🟦	🟦	🟩	🟩	🟩	🟩	□	□	🟩
Capitella teleta	□	🟩	□	🟩	🟩	□	□	□	🟩	□	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	□	□	🟩
Lamellibrachia symbiont	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	□	🟩	🟩	🟩	🟩	🟩	□	🟩	🟩	🟩	🟩

🟩 Present □ Absent 🟦 Gene only absent in *Lamellibrachia* host but present in both *Capitella* and *Lamellibracchia* symbiont

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

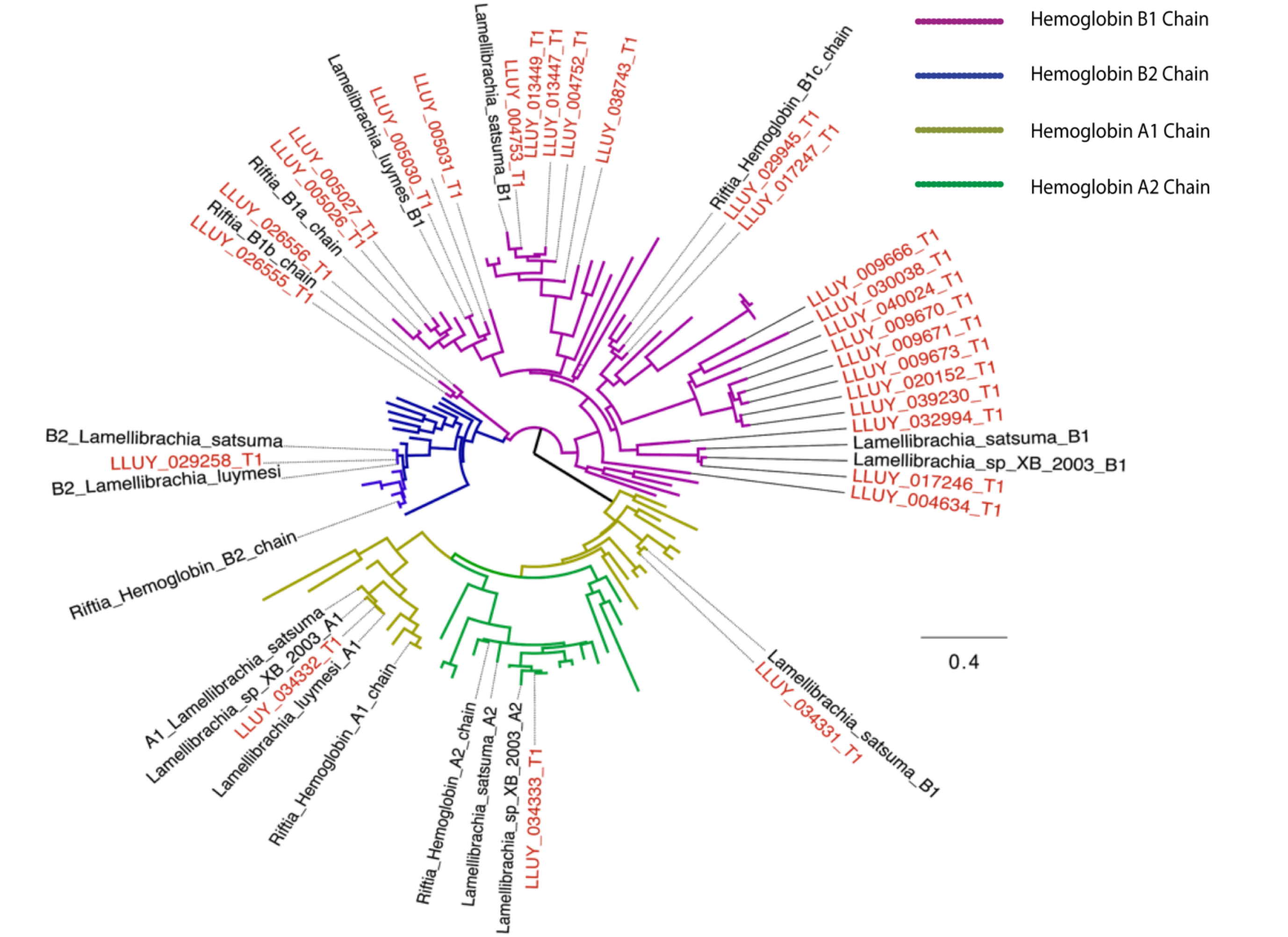


Fig. 3. Hemoglobin gene diversity in *Lamellibrachia luymesi*. Gene tree of siboglinid Hb subunits A1, A2, B1 and B2 reconstructed using IQtree with 1000 ultrafast bootstrap. Only Siboglinid Hb sequences (from SwissProt database or this study) were labeled. *L. luymesi* sequences labeled red. Accession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

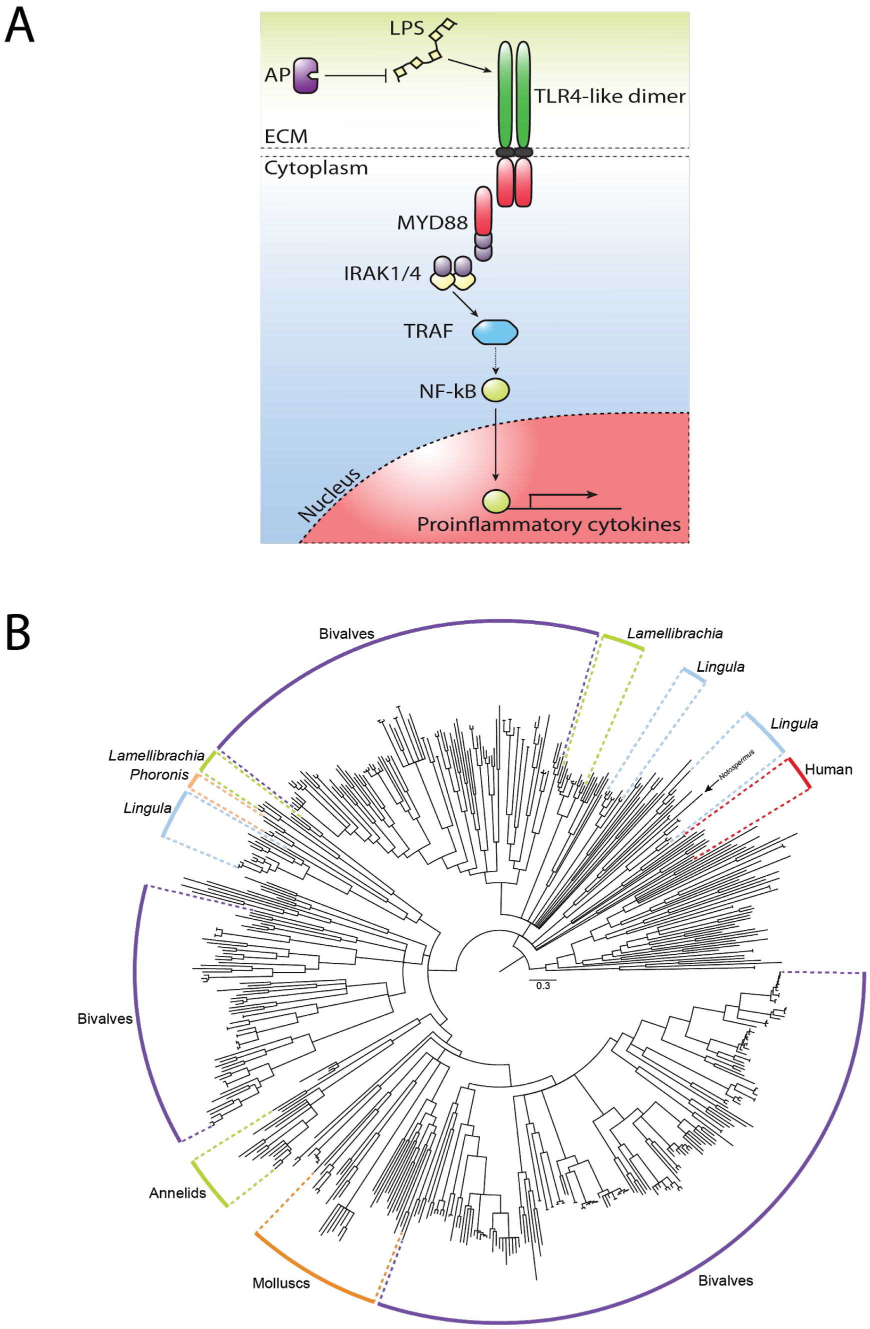


Fig. 4. Toll-Like Receptors (TLRs) in *Lamellibrachia luymesi*. (A) Putative TLR4-like pathway likely essential for immunity and response to symbionts and pathogens. AP: alkaline phosphatase; LPS: lipopolysaccharide. (B) Toll-Like Receptor gene tree from selected lophotrochozoan genomes and human reconstructed 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 symbionts overcome host cellular immune responses (26). Consistent with previous studies, we also be able to identify symbionts' SOD gene as proteins in proteomic analysis. Thus, SODs from both bacteria and tubeworms may play a central role for overcoming oxidative damage and essential for extreme longevity for seep-living vestimentiferans.

Conclusion

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

Methods.

Organismal collection

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

Genome sequencing and assembly

Using vestimentum tissue of one individual, high molecular weight genomic DNA was extracted using the DNeasy Blood & Tissue Kit (Qiagen). Four TruSeq paired-end 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.