

Supplementary Information for

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

Yuanning Li1,2\*, Michael G. Tassia1, Damien S. Waits1, Viktoria E. Bogantes1, Kyle T. David1, Kenneth M. Halanych1

Corresponding authors: Yuanning Li, Kenneth M. Halanych

Email:  [yuanning.li@yale.edu](mailto:yuanning.li@yale.edu); k[en@auburn.edu](mailto: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:**

Datasets 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. Briefly, samples were transferred to 5mL tubes appropriate for use in the Bullet Blender Storm 5 (Next Advanced). Freshly cleaned, 3.2mm stainless steel beads were added at an approximate 1:1 sample: bead volume ratio (1-5 beads). The tissue/bead mixture was then combined with 1X PBS (Hyclone) spiked with 1X Halt Protease Inhibitor cocktail (Thermo Scientific) for an approximate 2:1 liquid: solid ratio (150-600 μl). Tissue homogenization was achieved using speed setting 12 for 5 minutes. The stainless-steel beads were removed and 0.1mm glass beads added at approximately 1:1 sample:bead ratio. Further homogenization was achieved at speed 8 for 3 minutes. Homogenate was then transferred to 1.5mL microcentrifuge tubes and subjected to cup horn sonication (Amplitude 70, 10s pulse, 20s off, 7 minutes total exposure time; Qsonica) while suspended in ice water. Insoluble material was pelleted by brief centrifugation at 3000xg. The resulting supernatant was subjected to protein quantification by the Pierce BCA Protein Assay Kit (ThermoFisher-Pierce) following manufacturer’s instructions. Absorbance of reacted samples was measured at 550nm and protein quantification stemmed from a quadratic fit of the Bovine Serum Albumin (BSA) standard curve.

50 μg total protein was aliquoted from each sample and processed for in-solution trypsin digestion as previously described [(1)](https://paperpile.com/c/MnJCyS/n2oBG). 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)](https://paperpile.com/c/MnJCyS/V7qqd)) 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)](https://paperpile.com/c/MnJCyS/04ZEc) implemented in the Scaffold software (version Scaffold\_4.8.4, Proteome Software Inc., Portland, OR) [(4)](https://paperpile.com/c/MnJCyS/WeyQU). 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)](https://paperpile.com/c/MnJCyS/nKHHo). 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.**



Supplementary Information for

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

Yuanning Li1,2\*, Michael G. Tassia1, Damien S. Waits1, Viktoria E. Bogantes1, Kyle T. David1, Kenneth M. Halanych1

Corresponding authors: Yuanning Li, Kenneth M. Halanych

Email:  [yuanning.li@yale.edu](mailto:yuanning.li@yale.edu); k[en@auburn.edu](mailto:Ken@auburn.edu)

**This PDF file includes:**

Supplementary text

Figs. S1 to S7

Tables S1 to S

**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.Based on their instruction, 50 μg total protein was aliquoted from each sample and processed for in-solution trypsin digestion as previously described [(1)](https://paperpile.com/c/MnJCyS/n2oBG). 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)](https://paperpile.com/c/MnJCyS/V7qqd)) 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)](https://paperpile.com/c/MnJCyS/04ZEc) implemented in the Scaffold software (version Scaffold\_4.8.4, Proteome Software Inc., Portland, OR) [(4)](https://paperpile.com/c/MnJCyS/WeyQU). 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)](https://paperpile.com/c/MnJCyS/nKHHo). 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)](https://paperpile.com/c/MnJCyS/JoIS+Fc2s)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)](https://paperpile.com/c/MnJCyS/xPD1F)to align Hb amino acid sequences. Maximum likelihood analyses were performed in IQTree v1.5 [(9)](https://paperpile.com/c/MnJCyS/h9QK) 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)](https://paperpile.com/c/MnJCyS/eJxA) (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](https://github.com/yzl0084/Lamellibrachia-genome/tree/master/Immunity_related/Orthofinder_Archive)).

**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.](http://paperpile.com/b/MnJCyS/n2oBG)

2. [Li Y, Liles MR, Halanych KM (2018) Endosymbiont genomes yield clues of tubeworm success. *ISME J* 12(11):2785.](http://paperpile.com/b/MnJCyS/V7qqd)

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.](http://paperpile.com/b/MnJCyS/04ZEc)

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.](http://paperpile.com/b/MnJCyS/WeyQU)

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.](http://paperpile.com/b/MnJCyS/nKHHo)

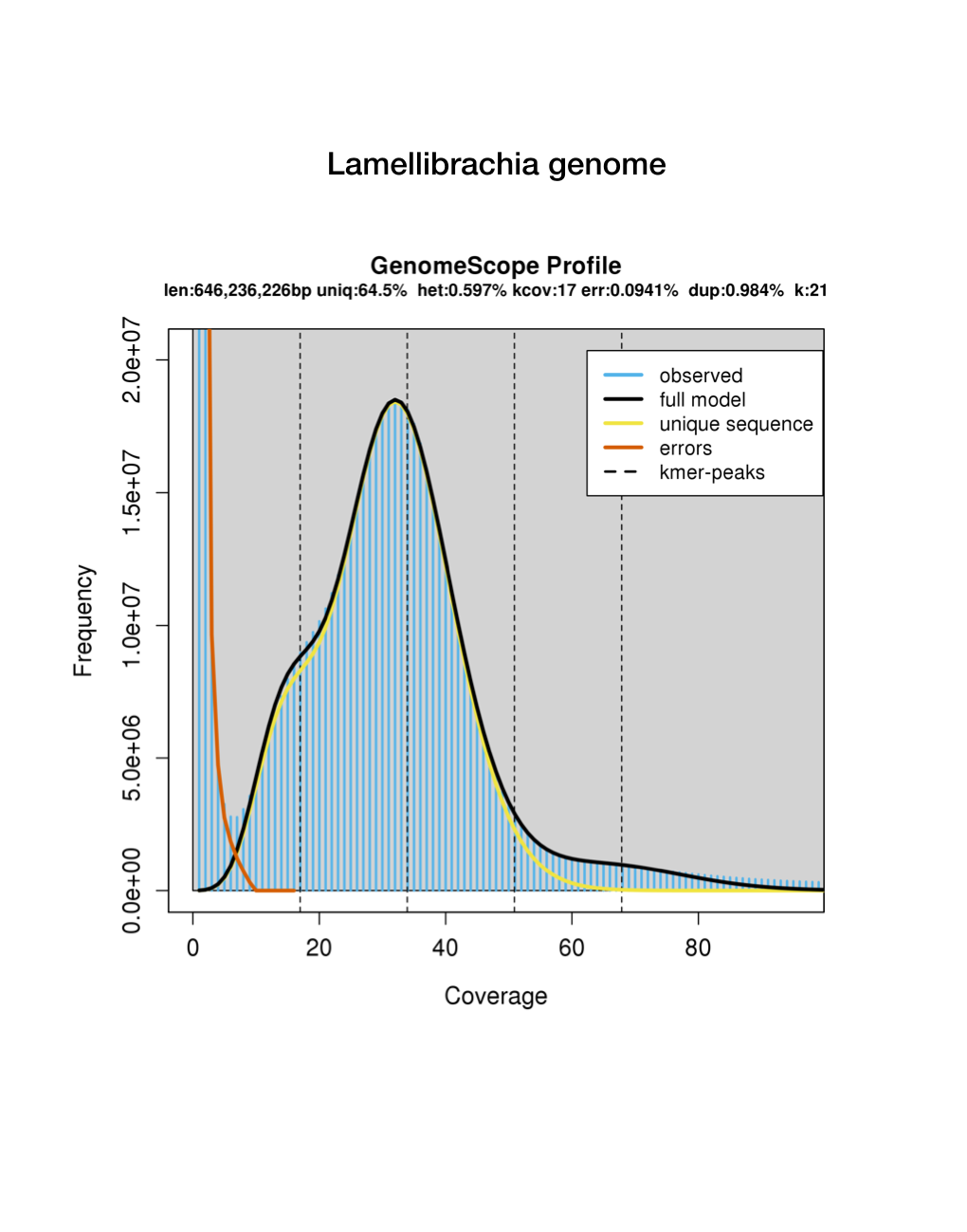
6. [Li Y, et al. (2017) Phylogenomics of tubeworms (Siboglinidae, Annelida) and comparative performance of different reconstruction methods. *Zool Scr* 46(2):200–213.](http://paperpile.com/b/MnJCyS/JoIS)

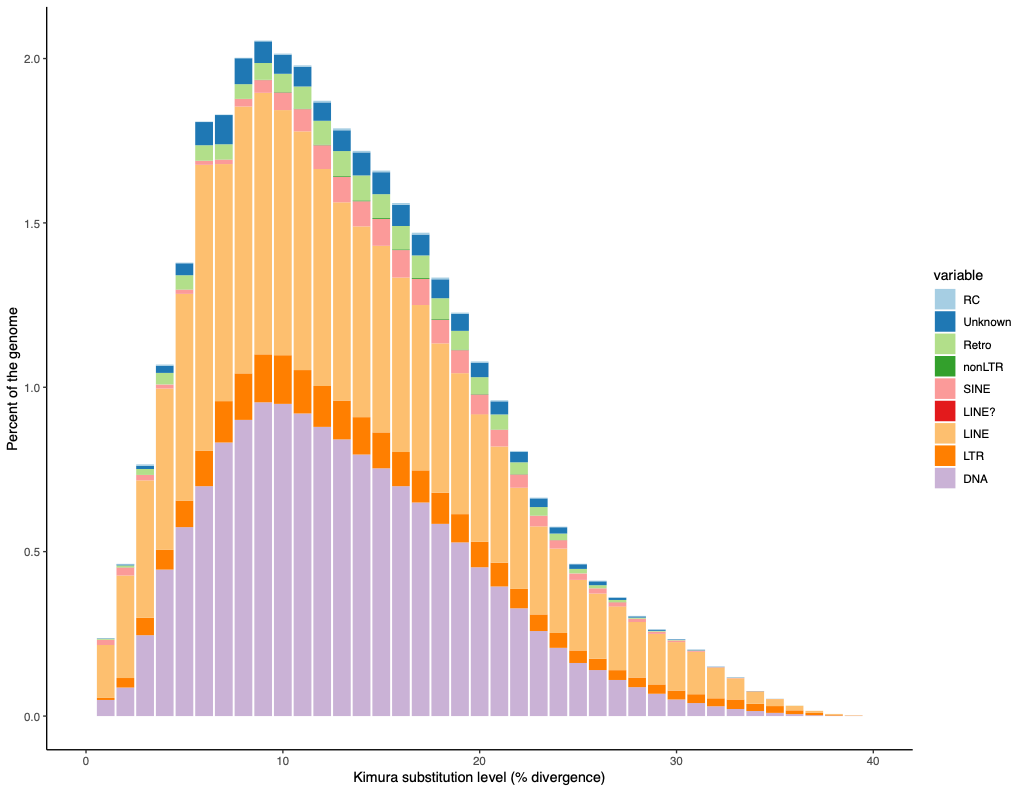
7. [Waits DS, Santos SR, Thornhill DJ, Li Y, Halanych KM (2016) Evolution of Sulfur Binding by Hemoglobin in Siboglinidae (Annelida) with Special Reference to Bone-Eating Worms, Osedax. *J Mol Evol* 82(4-5):219–229.](http://paperpile.com/b/MnJCyS/Fc2s)

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.](http://paperpile.com/b/MnJCyS/xPD1F)

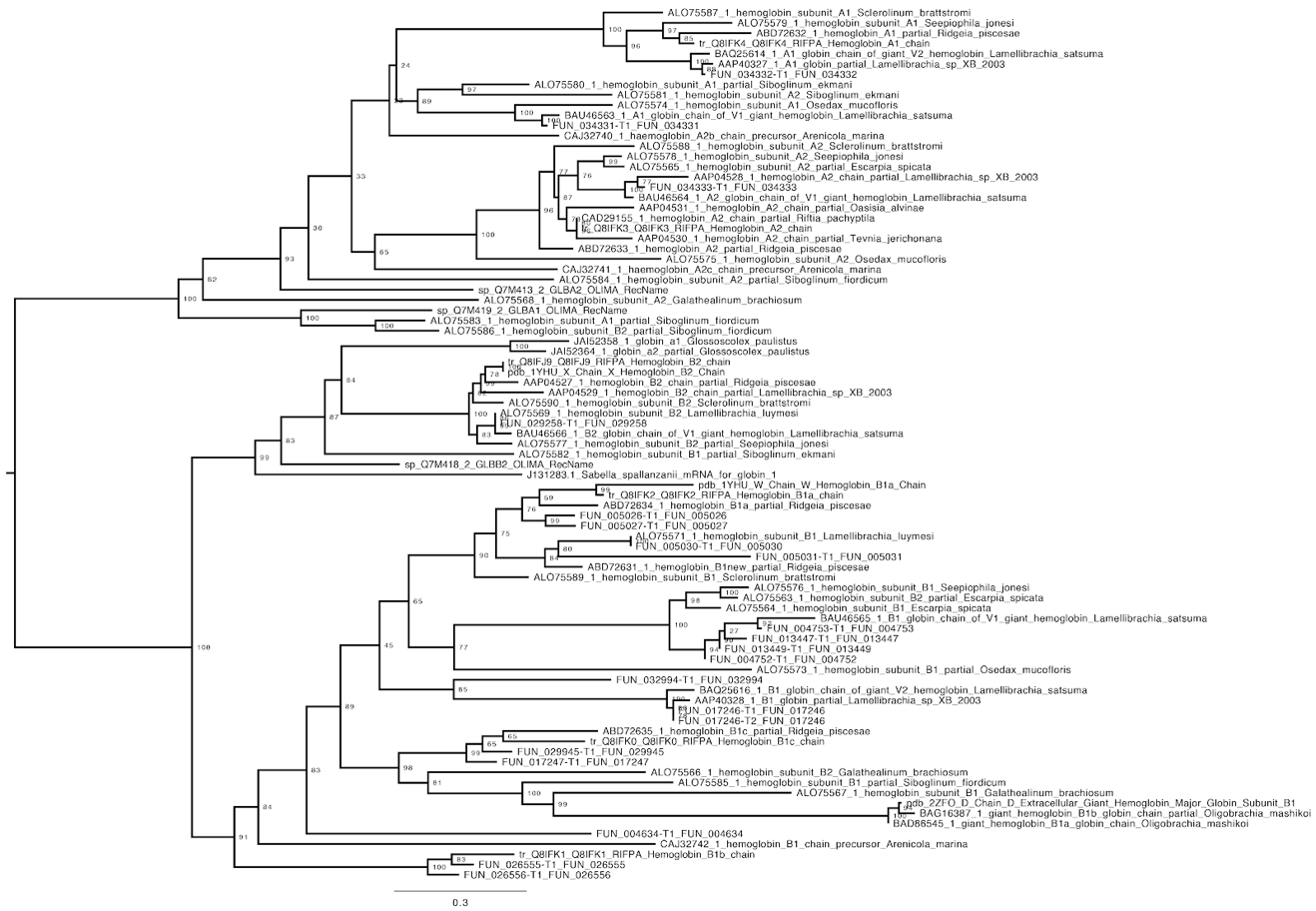
9. [Chernomor O, von Haeseler A, Minh BQ (2016) Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Syst Biol* 65(6):997–1008.](http://paperpile.com/b/MnJCyS/h9QK)

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.](http://paperpile.com/b/MnJCyS/eJxA)

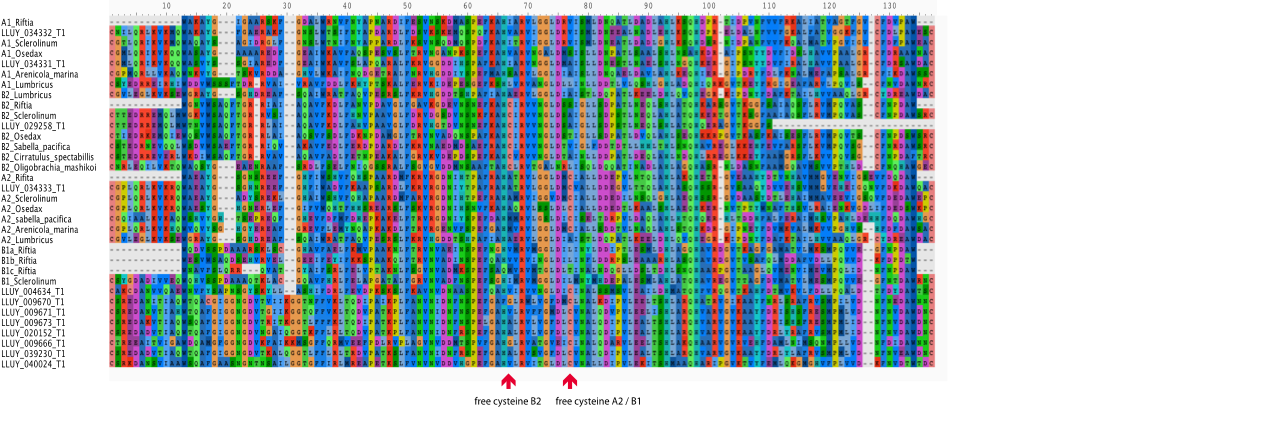
**Figure S1.** Estimation of genome size, repetitive content and level of heterozygosity from 100 million Illumina paired-end reads.



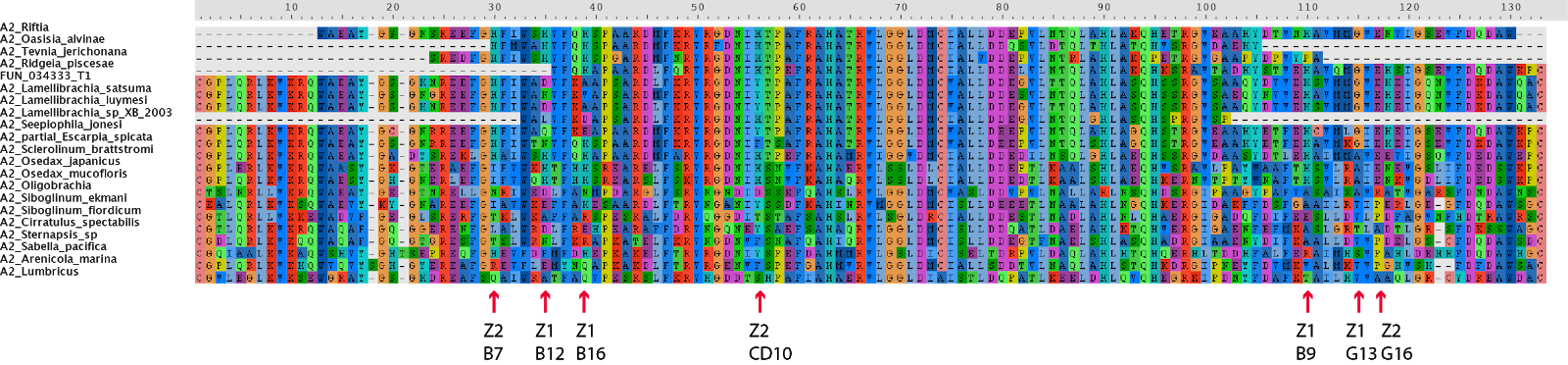
**Figure S2.** History from major superfamilies of transposable elements in the *Lamellibrachia* genome**.** Kiruma distances are arranged from value 0 representing recent TE copies to 40 for ancient TE insertions.

****

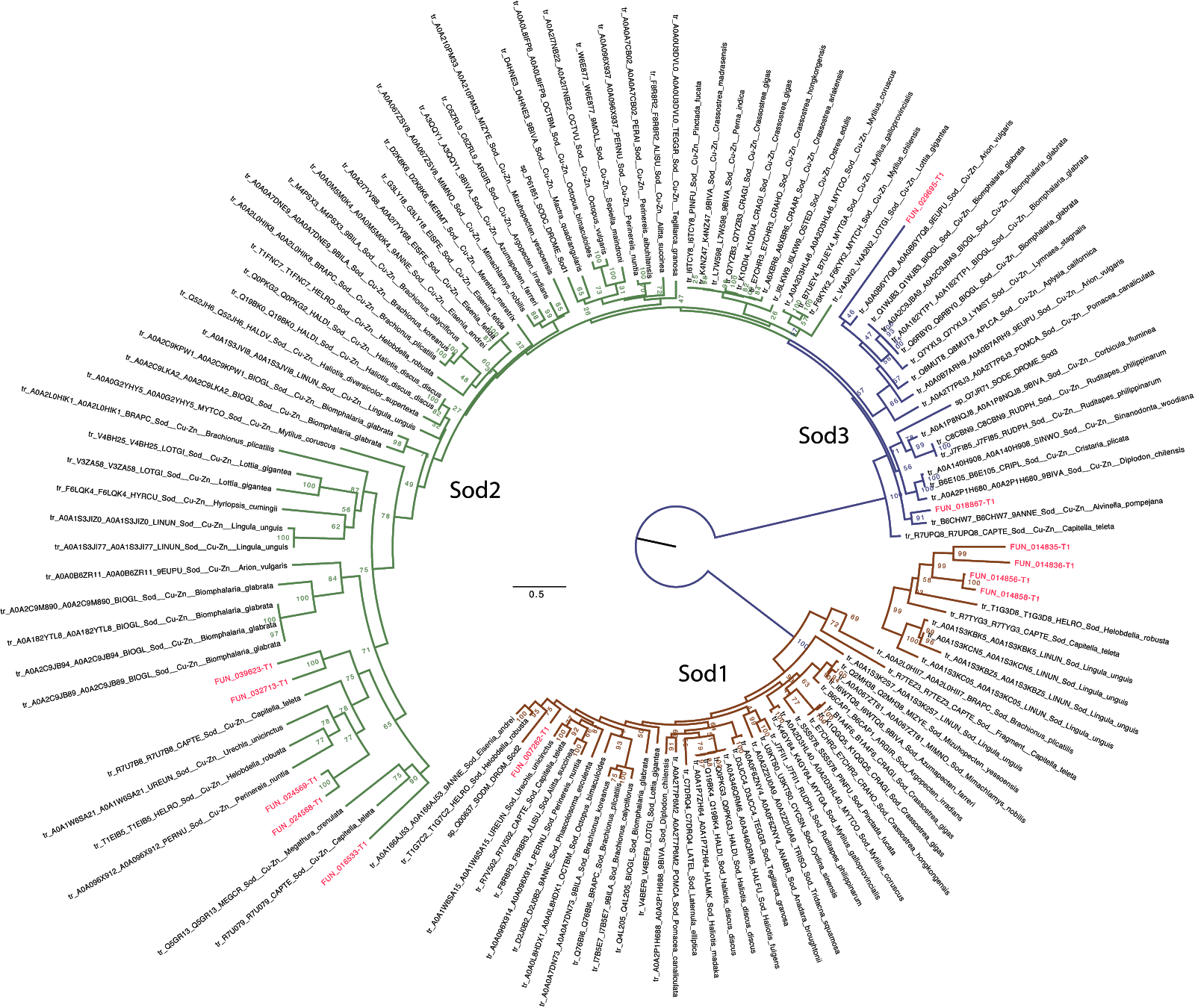
**Figure S3.** Siboglinid hemoglobin maximum-likelihood tree reconstructed with IQtree with midpoint rooting with 1000 ultrafast bootstraps using LG model. Bootstrap support values are shown at the relevant node. GenBank accession numbers are listed on the terminal nodes.



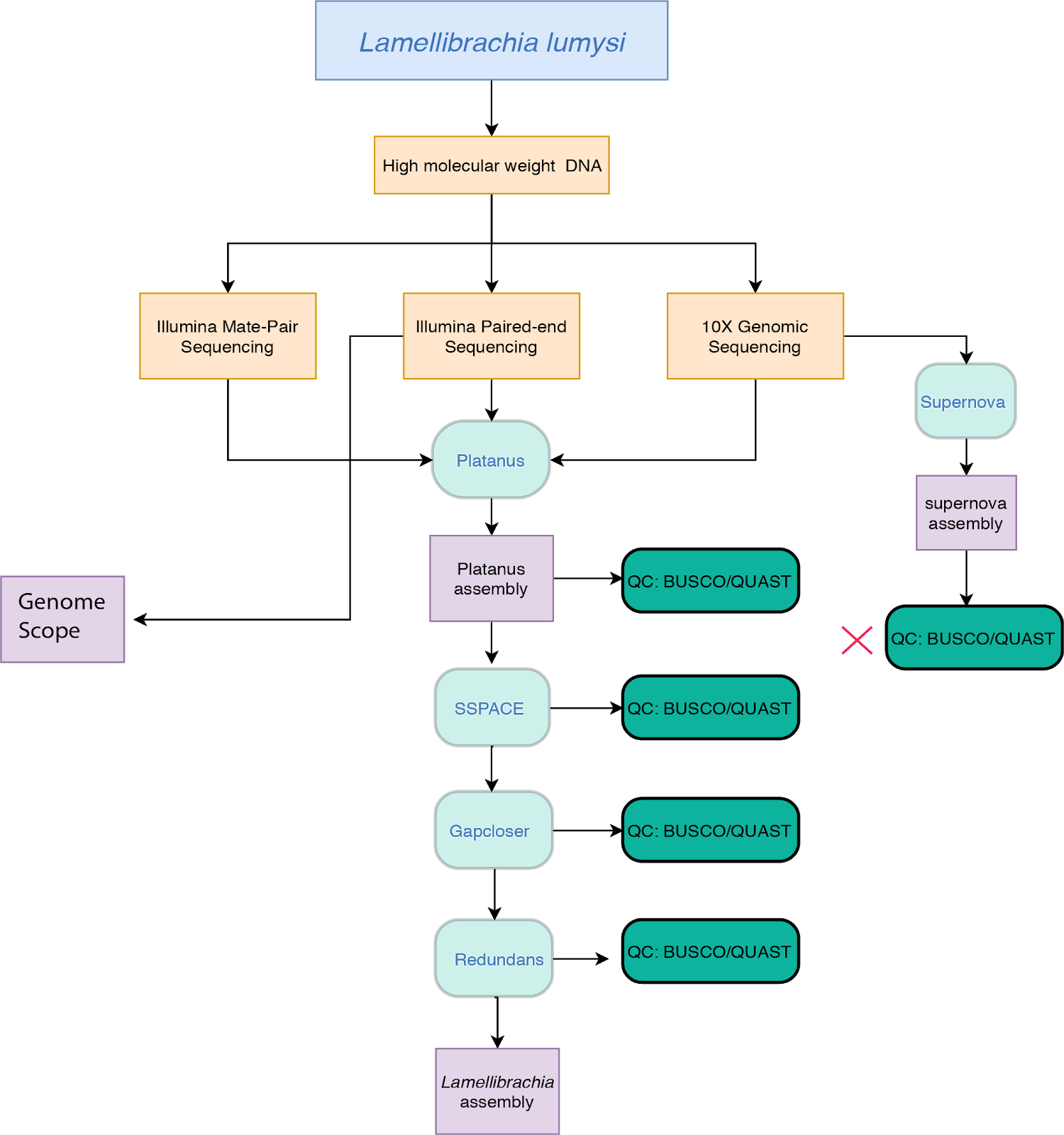
**Figure S4.** Hemoglobin gene diversity in *Lamellibrachia luymesi*. (Partial alignment of sampled siboglinid Hb subunit A1, A2, B1, B2 sequences. Red arrows indicate positions contain free cysteines or cysteine residues in HB B2 chains, and B1/A2 chains, respectively.

****

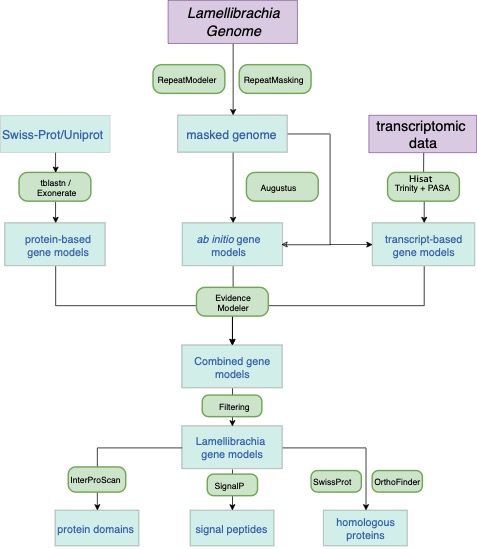
**Figure S5.** Partial alignment of sampled siboglinid HB subunit A2 sequences. Red arrows indicate amino acid residues at the interface between pairs of A2 chains with zinc moieties for H2S binding.

****

**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 luymesi* | 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](https://www.ncbi.nlm.nih.gov/assembly/1062421) |
|  | *Biomphalaria glabrata* | NCBI |  |
|  | *Mizuhopecten yessoensis* | NCBI | [GCA\_002113885.2](https://www.ncbi.nlm.nih.gov/assembly/1131331) |
|  | *Modiolus philippinarum* | NCBI | GCA\_000457365.1 |
|  | *Patinopecten yessoensis* | NCBI |  |
|  | *Crassostrea gigas* | NCBI | GCF\_000297895.1 |
|  | *Crassostrea virginica* | NCBI | [GCA\_002022765.4](https://www.ncbi.nlm.nih.gov/assembly/1188741) |
| **Nemertea** | *Notospermus geniculatus* | NCBI | [GCA\_002633025.1](https://www.ncbi.nlm.nih.gov/assembly/1369591) |
| **Phoronida** | *Phoronis australis* | NCBI | [GCA\_002633005.1](https://www.ncbi.nlm.nih.gov/assembly/1369581) |
| **Brachiopoda** | *Lingula anatina* | NCBI | [GCA\_001039355.2](https://www.ncbi.nlm.nih.gov/assembly/1536921) |
| **Flatworm** | *Schistosoma mansoni* | NCBI | [GCA\_000237925.2](https://www.ncbi.nlm.nih.gov/assembly/557858) |
|  | *Schmidtea mediterranea* | NCBI | GCA\_002600895.1 |
|  | *Macrostomum lignano* | NCBI | GCA\_002269645.1 |
|  | *Echinococcus multilocularis* | NCBI | [GCA\_000469725.3](https://www.ncbi.nlm.nih.gov/assembly/594471) |
| **Rotifera** | *Aplysia californica* | NCBI | [GCA\_000002075.2](https://www.ncbi.nlm.nih.gov/assembly/683478) |
| **Ecdysozoa** | *Diphania pulex* | NCBI | [GCA\_000187875.1](https://www.ncbi.nlm.nih.gov/assembly/244278) |
|  | *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** |
| **Poteosoma** | 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](https://www.genome.jp/dbget-bin/www_bget?ko:K03038) |
|  | 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]](https://www.genome.jp/dbget-bin/www_bget?ec:3.4.14.1) |
|  | LLUY\_007735-T1, LLUY\_007737-T1, LLUY\_036642-T1, LLUY\_036643-T1 | [cathepsin B [EC:3.4.22.1]](https://www.genome.jp/dbget-bin/www_bget?ec:3.4.22.1) |
|  |  |
|  | LLUY\_009810-T1, LLUY\_026908-T1 | [cathepsin L [EC:3.4.22.15]](https://www.genome.jp/dbget-bin/www_bget?ec:3.4.22.15) |
|  | LLUY\_012982-T1 | [legumain [EC:3.4.22.34]](https://www.genome.jp/dbget-bin/www_bget?ec:3.4.22.34) |
|  | LLUY\_007693-T1 | [cathepsin F [EC:3.4.22.41]](https://www.genome.jp/dbget-bin/www_bget?ec:3.4.22.41) |
|  | LLUY\_016715-T1 | [cathepsin D [EC:3.4.23.5]](https://www.genome.jp/dbget-bin/www_bget?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]](https://www.genome.jp/dbget-bin/www_bget?ec:3.4.18.1) |
|  | LLUY\_002319-T1 | [lysosomal alpha-mannosidase [EC:3.2.1.24]](https://www.genome.jp/dbget-bin/www_bget?ec:3.2.1.24) |
|  | LLUY\_004297-T1 | [lysosomal alpha-glucosidase [EC:3.2.1.20]](https://www.genome.jp/dbget-bin/www_bget?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]](https://www.genome.jp/dbget-bin/www_bget?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; <br>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 |