manuscript

Yuanning Li

10/17/2018

# Genomic clues of deep-sea chemosynthetic symbioses from *Lamellibrachia*

## Abstract

Fourty years after discovery of chemosynthetic symbiosis in the tubeworm Riftia pachyptila, how organims

## Introduction

 **Fig 1.** *Lamellibrachia* from seep localities in Gulf of Mexico.

Considerabel debate persists concenring the evolutionary origins of siboglinids, owing to conflicting theories of its origins from fossil and molecular age estimates. Siboglinids have been claimed to be as old as 430 MYA based on fossil tubes found in the Silurian fossil vent communities (Little 2002), but molecular clock analyses suggest a much recent ( 50-126MYA) origin based on COI or 16S sequences (Little and Vrijenhoek 2003). Recently, Late Cretaceous *Osedax* fossil traces on reptile falls provides a firm calibration point for the molecular clock of the siboglinids (~100 MYA) (Danise and Higgs 2015). Moreover, a detailed chemical and morphological analyses of tubes from the Figueroa deposits were made by vestimentiferans, which greatly extends the age of this lineage than previous molecular clock analyses (Georgieva et al. 2017).

Remarkably siboglinid Hbs are able to bind oxygen and sulfide simultaneously and reversibly at two different sites. To avoid the toxic effects of sulfide while supplying it to their chemoautotrophic endosymbionts, siboglinds possesses a multihemoglobin system with three different extracellular hemoglobins (Hbs; V1, V2, and C1): two dissolved in the vascular blood, V1 and V2, and one in the coelomic fluid, C1 (Arp and Childress 1981, @zal1996multi). Siboglinid Hbs consist of four heme-containing chains (A1, A2, B1, B2). Sulfur-binding capabilities are hypothesized to be due to free cysteine residues at key positions in Hbs, especially in the A2 and B2 chains. 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 (Zal et al. 1997). Moreover, part of the H2S binding affinity has also been suggested to be mediated by the zinc moieties bound to amino acid residues at the interface between pairs of A2 chains in *Riftia* (Flores et al. 2005). ## Methods Scripts and data for the analyses are available in a git repository at <https://github.com/yzl0084/Lamellibrachia-genome>.

### Biological materials.

Adult *Lamellibrachia* *luymesi* specimens were collected from seep localities in the Mississippi Canyon at 754 m depth in Gulf of Mexico (N 28°11.58’, W 89°47.94’), using the R/V Seward Johnson and Johnson Sea Link in October 2009. All samples were frozen at 80˚C following collection.

### Genome sequencing and assembly.

Vestimentum tissue was dissected from one individual of worm, and high molecular weight genomic DNA was extracted using the DNeasy Blood & Tissue Kit (Qiagen) according to the manufacturer’s protocols. Sequencing a total of six paired-end or mate-pair genomic DNA libraries with insert sizes ranging from 180 bp to 7 kb were performed by The Genomic Services Lab at the Hudson Alpha Institute in Huntsville, Alabama on an Illumina HiSeq 2000 platform (see details in Table S1). Paired-end libraries (180 bp, 400 bp, 750 bp) were prepared using the 125 bp TrueSeq protocols, and mate-pair libraries (3-5 kbp, 5-7 kbp) were generated using the Illumina Nextera Mate Pair Library Kit followed by size selection. Moreover, a 10X sequencing library was constructed using the 10X Chromium protocol (10X genomics) at the Hudson Alpha Institute. The finished library was sequenced on an Illumina HiSeqX platform, using paired 151 bp reads with a single 8 bp index read.

Our workflow of genome assembly was shown in Fig. S1. The paired-end and 10X raw reads were checked using FastQC v0.11.5 (Andrews and others 2010) and quality filtered (Q score >30) using Trimmomatic v0.36 (Bolger, Lohse, and Usadel 2014). The estimation of genome size, level of heterozygosity and repeat contents of the Lamellibrachia genome was determined by analaysing the kmer histograms generated from the paired-end libraries using Jellyfish v2.2.3 (Marçais and Kingsford 2011) and GenomeScope (Vurture et al. 2017) (Fig. S2). The Mate-pair reads were trimmed and sorted using NxTrim v0.3.1 (O’Connell et al. 2015) which can recgonize and trim the artificial Nextera mate-pair circulation adapters. Only reads from category “mp” (true mate-pair reads) and “unkonwn” (mostly large insert size reads) were used for downstream scaffolding anlaysis. Reads from “pe” (paired-end reads) and “se” (single ends) categories were discarded.

Given that high heterozygosity of *Lamellibrachia* genome, all reads were assembled using Platanus v1.2.4 (Kajitani et al. 2014) with a kmer size of 32. Scaffolding was conducted by mapping Illumina paired-end and mate-pair reads to contigs genrated by Platanus using SSPACE v3.0 (Boetzer and Pirovano 2014). Gaps in the scaffolds were then filled with GapCloser v1.12 (Luo et al. 2012). Redundant allele scaffods were further remvoed using Redundans v0.13c with default settings (Pryszcz and Gabaldón 2016). Genome assembly quality was assessed using QUAST v3.1 (Gurevich et al. 2013). Completeness of obtained genome was assessed using BUSCO v3(Waterhouse et al. 2017) with Metazoa\_odb9 database (978 busco genes).

### Transcripome assembly and analysis

Total RNA was extracted from the plume, vestimentum and trophosome tissue from the same indivdisual of *Lamellibrachia* spicemen using Trizol. RNA-seq of adult tissues from plume, vestimentum and trophosome was performed using Illumina HiSeq 2000 platform in Hudson Alpha. After quality checking and trimming of raw sequencing reads, transcripts were assembled de novo with Trinity. Transcript isoforms with high similarity (≥ 95%) were removed with CD-HIT-EST v4.7 (Li and Godzik 2006). Transcript abundance was estimated with Bowtie v2.2.9 (Langmead and Salzberg 2012) and RSEM v1.2.26 (Li and Dewey 2011) by mapping reads back to the transcriptomic assembly based on transcripts per million (TPM). A tissue specifically expressed gene was defined as a gene that had over 75% of the total transcripts in a particular tissue based on TPM (Albertin et al. 2015). GO enrichment analysis was performed based on the GO annotation using PANNZER2 (Törönen, Medlar, and Holm 2018). Statistically overrepresented GO terms of trophosome-specific genes were identifed using Lamellibrachia gene models as background with AgriGO (Du et al. 2010).

### Proteomics characterization.

Proteomics analyses of *Lamellibrachia* trophosome tissue were performed by Proteomics & Metabolomics facility at 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 (1-5beads). 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 (150-600μl). 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 (Schauer et al. 2013). 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, Thermo Scientific) equipped with a Nanospray Flex ion source (Thermo Scientific) and spectra were collected over a m/z range of 400–2000, positive mode ionization, 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 (version 3.0). 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 (Y. Li, Liles, and Halanych 2018)) 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 (Keller et al. 2002) implemented in the Scaffold software (version Scaffold\_4.8.4, Proteome Software Inc., Portland, OR) (Searle, Turner, and Nesvizhskii 2008). Peptide thresholds were set such that a low peptide FDR was achieved based on hits to the reverse database (Käll et al. 2007) or Protein Prophet (Nesvizhskii et al. 2003). 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 (Nesvizhskii et al. 2003). Proteins that contained similar peptides and could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principles of parsimony.

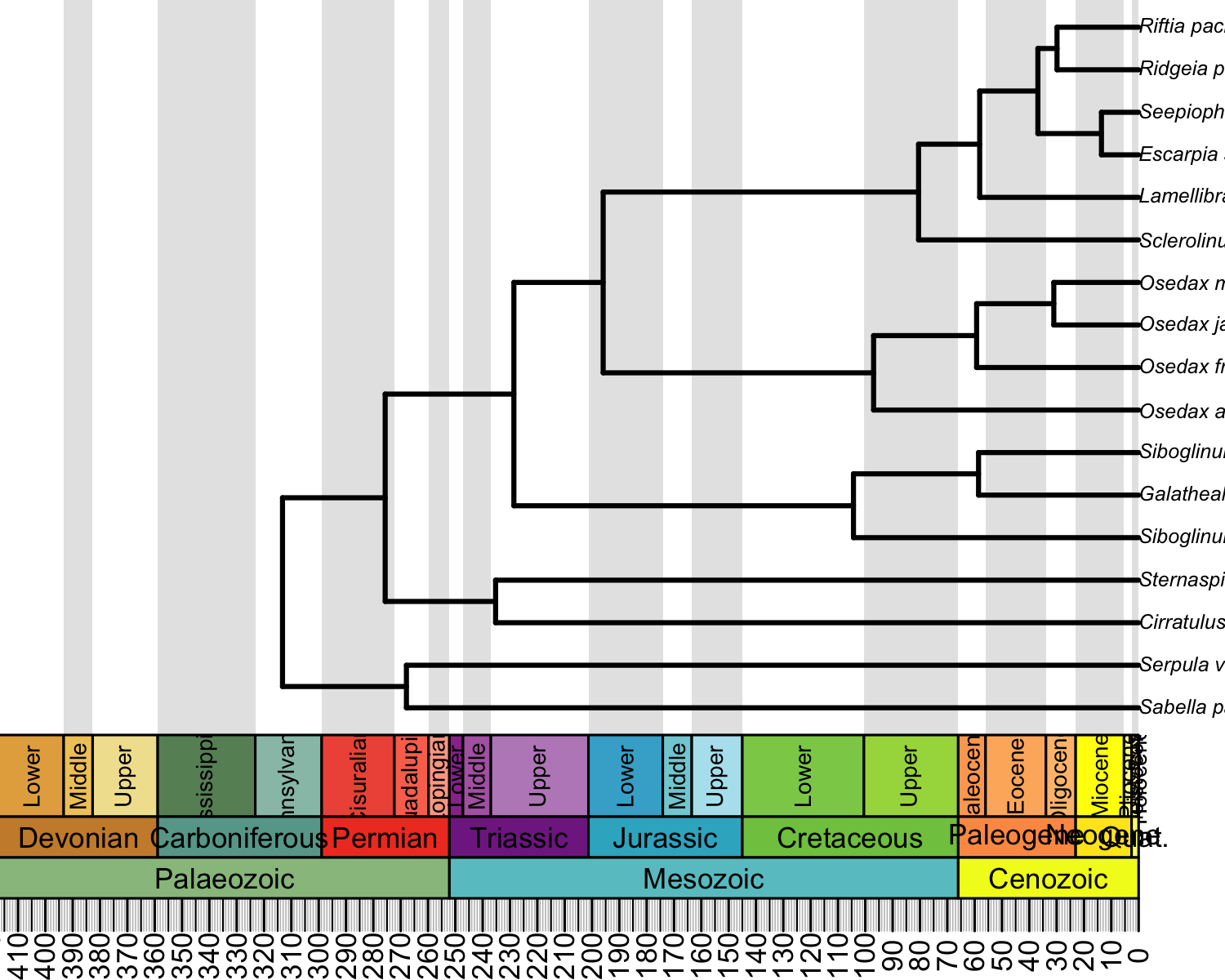
### Genome annotation.

Our genome annotation workflow was shown in Fig. S3. Gene models of *Lamellibrachia* genome were constructed following the Funannotate pipeline 1.3.0 (<https://github.com/nextgenusfs/funannotate>). Briefly, repeptive regions in the *Lamellibrachia* genome were identified usning RepeatModeler v1.0.8 (Smit and Hubley 2008) and were subsequently soft-masked using RepeatMasker v4.0.6 (Chen 2004). RNA-Seq data from different tissue were leveraged to improve the accuracy of gene prediction. RNA-Seq data were assembled *de* *novo* into transcriptomes using Trinity v2.4.0 (Haas et al. 2013) and HISAT 2.1.0 (Kim, Langmead, and Salzberg 2015) was used to algin RNA-Seq reads to the *Lamellibrachia* assembly. Transcrptome assemblies were then passed to PASA pipeline v2.3.3 (Haas et al. 2003) to identify high quality gene models. The aligned RNA-Seq data wes used to train the *ab* *initio* gene predictions using AUGUSTUS v3.3 (Stanke et al. 2006). Protein alignements from the SwissProt database to *Lamellibrachia" assembly were generated using exonerate (Slater and Birney 2005) and Trinity/PASA transcripts were aligned to the genome using Minimap2 v2.1 (H. Li 2018). The tRNA genes were identified using tRNAscan-SE v1.3.1 (Lowe and Eddy 1997). Finally, EvidenceModeler 1.1.0 (Haas et al. 2008) was used to combine all the evidences of gene prediction from protein alignemnts, transcritp alignments, and* ab\* *initio* predictions to construct high quality gene models. Finally, functional annotations of predicted gene models were analyzed using several curated databases. KEGG orthology was assinged using the KEGG Automatic Annotation server. Gene models were further annotated with domain structure and protein identity by InterProScan (Zdobnov and Apweiler 2001) and SwissProt database, respectively. Secreted proteins were predicted using SignalP (Petersen et al. 2011) and Phobius (Käll, Krogh, and Sonnhammer 2007) using InterProScan.

### Phylogenomics and Molecular clock analysis

Analysis of siboglinid phylogeny was conducted from publically available 16 transcriptomic data in conjuction with *Lamellibrachia* proteome (Table S2). Sequence assembly, annotation, homology evaluation, gene tree construction, parsing of genes trees to OGs, and supermatrix construction were conducted with Agalma (Dunn, Howison, and Zapata 2013). The final analyses presented here contained 13 siboglinids and four outgroups based on current understanding of annelid phylogeny. Maximum likelihood analyses were performed in IQTree under the best-fitting models for associated partition schemes determined by Modelfinder implemented in IQTree 1.6.3 (Nguyen et al. 2014) with ultrafast bootstrapping of 1000 replicates.

For molecular clock analysis, a relaxed molecular clock with a lognormal distribution and a Yule tree model was used in BEAST 2 v2.5.1 (Bouckaert et al. 2014). A calibration was placed on the node representing the most recent common ancestor (MRCA) of *Osedax* using a normal distribution with a mean of 100 MYA and a standard deviation of 10 following the findings of (Danise and Higgs 2015).Another calibrartion was placed on the node of MRCA of Serpulida and Sabellida using a normal distribution with a mean of 267 MYA (Sanfilippo et al. 2017). Molecular clock analyses with BEAST 2 consisted of two independent runs with 1 million MCMC generations sampled every 1000 generations. Convergence was checked and confirmed by comparing trace plots in Tracer making sure the effective sample size of each parameter was greater than 100 and that stationarity appeared to have been achieved. Log and tree fiels were combined using Logcombiner. A maximum clade credibility tree with mean heights was calculated using TreeAnnotater. The resulted time-calibrated tree was plotted using R package, phyloch, strap and OutbreakTools. Bayesian inference using a molecular clock resulted in identical branching patterns as analysis with IQTree.



### Gene family analysis

After all-to-all Diamond v1.0 (Buchfink, Xie, and Huson 2014) BLASTP searches against 22 selected lophotrochozoan proteomes (Table S3), orthology groups (OGs) were identified using Orthofinder with a dfault inflation parameter (I=1.5). Gene ontology annotation was performed using PANTHER v13.1 (Mi et al. 2016) with the PANTHER HMM scoring tool (pantherScore2.pl). Gene family expansion and contraction was estimated using CAFÉ v2.1 (De Bie et al. 2006). 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. A branch-specific *P* value was also generated for each branch/node using the Viterbi method. A family-wide *P* value less than 0.01 and a branch/node Viterbi *P* value less than 0.001 was considered as a signature of gene family expansion/contraction for a specific gene family and specific species, respectively, as suggested in manual.

### Manual annotation of gene families with potential intrest.

In addition to the automatic annotation process mentioned above, we manually annotated genes that are general intrest, such as hemoglobin gene families, genes related to amino aicd synthesize, or immunology. These genes were specifically selected *a* *priori* based on our experience after a revison of the available pulications in the field. For hmoglobin genes, datasets used for HB gene family analyses, identified homologs, and their respective accession numbers are available in Fig. S# .Hbs and linker sequences of interest were obtained from “Lamellibrahcia” genome and assembled siboglinid transcriptomes (see molecular clock analysis above) via diamond blastp (evalue cutoff 1E-5) by utilizing Hb and linker sequences acquired from GenBank of siboglinids as queries (Figure #). Sequences with best hists to target proteins were annotated for protein domain archetecture usig the Pfam databases included in InterProscan. After manual removal of redundant and incorret sequences (e.g. seuqences are too short or lack of globin domain), we used MAFFT 7.2.15 (Katoh and Standley 2013) to align HB amino acid sequences. Maximum likelihood analyses were performed in IQTree under the best-fitting models for associated partition schemes determined by Modelfinder implemented in IQTree with ultrafast bootstrapping of 1000 replicates.

## Results and Discussion

### Genome features

The genome of *Lamellibrachia* was sequenced using a combination of Illumina paired-end, mate-pair and 10x techniques. Results from high-throughput sequencing and genome assembly for *Lamellibrachia* are presented in Table S3 with at least 300 fold coverage using a combination of Illumina paired-end, mate-pair and 10X genomic sequencing with a variety of insert sizes. The haploid genome assembly size is ~687 Mb, similar to the result of genome size estimation using GenomeScope (~646 Mb). The N50 value of the assembled scaffolds and contigs is 373 Kb and 24 Kb, respectively. Although N50 lengths and assembly quality of *Lamellibrachia* are comparable to those of other annelids (e.g. *Capitella* *teleta*, *Hebdella* *robusta*), the overall genome completeness measured by BUSCO (~ 95%) is one of the highest among other lophotrochozoans (Table S4), indicating the completeness of the genome assembly. With the support of RNA-seq data from different tissue, we estimated *Lamellibrachia* genome contains 38,998 gene models. The genome also exhibit similar level of heterozygosity (0.6%) and repeptitive seuqences (36.92%), compared to other lophotrochzoan genomes. The repeatComparison among three annelid genome revealed.

### Nutritional adaptations to a symbiotic life stage

Adult *Lamellirbachia* worms lacks of a digestive tract and rely soely on endosymbionts for nutrition. Noticebaly, only 57 genes asscociated with amino acid biosynthesis were found in the *Lamellibrachia* host genome, of which eight were identified as proteins (Fig S#). In contrast, the genome of the *Lamellibrachia* symbiont (110 genes) contains essentially complete gene sets for the biosynthesis of all 20 proteinogenic amino acids and of 11 vitamins and cofactors. The lack of 30 genes in “Lamellibrachia” compared to “Capitella” genome (90 genes) are involved in the biosynthesis of 13 out of 20 amino acids (e.g. five key enzymes in the Lysine biosynthesis pathway Fig. #). As amino acids are indispensable for protein biosynthesis in the host, the sparse presence of amino acid synthesis-related genes in *Lamellibrachia* genome may indicate that the host depends on its symbionts for supply with amino acids and cofactors. Moreover, we found a large gene expansion of nutrient uptake ABC transports in *Lamellibrachia* compared with other lophotrochozoans (Table #). These findings are consistent with previous chemichal analyses suggest that *Riftia* is also dependent on its bacterial symbiont for the biosynthesis of polyamines that are important for its metabolism and physiology (Minic and Hervé 2003). Intrestingly, a similar mechanism of host difficiency of amino acid synthesis has also been suggested in thiotrophic vesicomyid symbnioses of *Calyptogena* *magnifia* (Newton, Girguis, and Cavanaugh 2008) and vent musssel symbioses of *Bathymodiolus* *azoricus* (Ponnudurai et al. 2017). These results emphasize the host deficiency of amino acid syntheses and exchanges of metabolites between two partners mgiht to be a common strategy for chemosynthetic symbioses.

In addition to immediate release of fixed carbon and provide amino acids by the symbionts, there is evidence of a second possible nutritional mode - digestion of symbionts directly by the host, as indicated by the detection of abundant host-derived digestive enzymes in the trophosome tissue (Table S#). Although previous observations indicated that symbiont can be digested by *Riftia* (Bright, Keckeis, and Fisher 2000), the direct evidence and genetic mechanism related to symbiont digestion has not been characterized. We identified fifteen the host proteins related to lysosomal proteases that were highly expressed and dectected as proteins in the trophosome tissue, such as *Saposin* and multi copies of *Cathepsin* (Table S#). Lysosomes contain an array of digestive enzymes used for protein degradation in eukaryotic cells and are thought to play an important role in symbiont digestion of *B*. *azoricus* (Ponnudurai et al. 2017). We furthermore identified 19 major components of proteasome as proteins in the trophoseome tissue, indicating a potential role in protein degradation of symbiont digestion. Thus, our results suggest that lysosomal host proteases and proteasome may facilitate the degradation of symbiont proteins during symbiont digestion, and maintaining population size of symbionts within trophosome, or possiblly both.

Because the trosposme tissue of *Lamellibrachia* contain sulfur-oxidizing symbionts, we also identified ~ 200 bacterial proteins in the trophosome tissue to gain insights into the symbiotic relationship between the bacteria and their host. Key enzymatic genes, RubisCO and ATP citrate lyase (ACL) type II associated with these carbon fixation cycles, were identified as proteins in *Lamelibrachia* symbiotns (Table S#). Our resutls further corroborate the idea that the presence of rTCA in addition to CBB pathways for carbon fixation might be common in all vent-dwelling and seep-living vestimentiferan endosymbionts. We also identified several key components related sulfide and nitrogen metabolic pathways, and the results are largely consistent with preivous analyses.

### Hemoglobin evolution

![](data:application/pdf;base64,) **Fig 2.** Hemoglobin.

Partial alingment and gene tree of siboglinid HBs is shown in Figure #.Tthe expression of Hb subunit A1, A2, B1, B2 and linker L1-L4 were among the highest in the trophosome comared to the plume and vestimentum, and we were able to identify most of them in the mass specmetry. Consistent with previous studies (Zal et al. 1996), a single copy of A2 and B2 HB was identifed in all siboglinids with a conserved-free cystine (i.e., cysteine residues not involved in dissulfide brideges) at position 77 and 67, respectively. With the exception of A2 and B2 HB globins in *Lumbricus* *terrestris*, Cystine residues in same positions are also identified in *Cirratulus* *spectabilis*, *Sabella* *pacifica* and *Sternapsis* sp. from sulfide-free and *Arenicola* *marina* living in sulfide-rich environments (Fig. 3). These results supports the hypothesis that free cysteine residues in A2 and B2 globins have been conserved before the radiation of the Siboglinidae and potentially involve in H2S detoxification process (Bailly et al. 2002).

Surprisingly, we found a large expansion of B1 globins in *Lamellibrachia* genome with 25 copies. *Riftia* might also contain more copies since three B1 globins were identified using a set of degenerate primers by PCR in the previous study (Bailly et al. 2002). Noticablly, we found that eight copies of “Lamellibrachia” B1 seqeuences also contains a free cysteinine at position 77, same position as free cysteine in A2 globins. Five out of eight copies were highly expressed in the trophosome and one copy was identified at the protein level. Unlike A2 and B2 globins, it was long thought that B1 globins can only bind O2 and lack of the ability of sulfid binding. Although the presence of free cystine residues is not proof of sulfide binding also occurs in B1 globins, it is at least possible that the hemoglobin structure and sulfide binding function is more complex that previous thought due to unexpected large expansion and highly expression of “Lamellibrachia” B1 sequences.

Instead of free cysteinies, another hypothesis suggested that the H2S binding affinity has been mediated by the zinc moieties bound to amino acid residues at the interface between pairs of A2 chains in *Riftia* (Flores et al. 2005). Figure S# shows an alginment of amino acid sequences of A2 globins from Siboglinidae and other annelids. Amino acids involved in the binding to of zinc ions in *Riftia* A2 globin pairs are highlighted. Each A2 chain contains one complete Zn2+-binding site and shares two others with adjoining A2 subunits. The Zn2+-binding site contained within A2 chain is composed of three His residues (B12, B16, and G9) (Flores et al. 2005). However, none of these sites are conserved across siboglinids, even in vestimentiferans (Fig. S#). For example, B16 His site is replaced by a varietly of amino acids in seep-living vestimentiferans. Thus, it is less likely that the zinc sulfide binding mechanism is a conserved functional in siboglinids, and might only be a derived function only for vent-living vestimentiferans.

### Immunity function and aging

The immune function between hosts and symbionts is a key evolutionry driver that is under strong selective pressure and has potential . However, the genetic machienary and funtionality of immune system in most chemosynthetic symbioses have not been extensively characterized. Our detailed analysis revealed several genes related to immunne function found duplications in *Lamellibrachia* compared with other lophotrochozoans (Table S#). Toll-like receptor (TLRs) provides vital cellular and molecular defense against invading pathogens and recognition of host-microbial symbiosis (Chu and Mazmanian 2013). Consistent with previous analysis (Luo et al. 2018), we found that TLR genes were generally expanded in most lophotrochozoans, especially in bivalves (Fig. S#). Noticebaly, we found several lineage-specific expansion of putative TLR genes in *Lamellibrachia* genome with 31 copies, whereas only five copies were found in *Capitella* genome, suggesting TLR genes in *Lamellibrachia* may play a more relevant role for symbiosis.

The initial physical encounter between tubeworms and symbionts occurs in an extracellular mucous secreted by pyriform glands by newly settled larvae (Nussbaumer, Fisher, and Bright 2006). In mucus matrices, symbionts can attach to the host using extracellular components secreted from symbionts, such as lipopolysaccharide - a major component of the outer membrane of *Riftia* symbionts. TLR4 orthologs have been hypothesized to recognize lipopolysaccharide in several studies. We found a genomic expansion of TLR4 in *Lamellibrachia* (7 copies) compared with *Capitella* (1 copy) from KEGG mapping. Moreover, the infection process finishes simultaneously with massive apoptosis of skin tissue during symbionts travel from host epidermal cells into trophosome (Nussbaumer, Fisher, and Bright 2006). This is consistent with that recognition of lipopolysaccharide (LPS) by TLR4 can result in indcution of sinalling cascades that lead to activation of NF-kb and the produciton ot proinflamatory cytokines (Chu and Mazmanian 2013). Intrestingly, we also found genomic expansions of Tumor necrosis factor receport (TRFR) and TNF recepotr associated factor (TNAF) (Table S#), both play key roles in NF-kb pathway. Although it is still not clear how the hosts distinguish between symbionts and pathogens in most symbioses, alkaline phosphatase has been shown to be involved in the maintenance of hemeostasis in the squids, mouse and zebrafish. The commensal bactieria-dirived LPS signaling via TLR4-MYD88 leads to upreregulation of intestinal alkaline phophatase and prevents infalmmatory responses to the resident microbiota. Importantly, we also identified eight copies of alkaline phosphatase in *Lamellibrachia*, whereas only one copy was found in *Capitella*, suggesting a potential mechanism of in response to symbionts and pathogens, thus facilating colonization by the endosymbionts.

In addition of TLRs, we also found evidence of genomic expansion of NLRP family, which play a key role in the inflammasome, an innate immune system protein complex that recognizes infectious pathogens and regulates the activation of inflammatory caspases. Moreover, We found large expansions of several members of Sushi domain proteins that could be involved recognition and adhesion between hosts and symbionts.

### Phylogeny and molecular clock of tubeworm

Two additional *Osedax* taxa were added in the study compared to the previous siboglinid phylogenomic analyses (Li et al. 2017) (Table #). The final supermatrix dataset contains 191 genes single-copy orthologs. Bayesian inference with a relaxed molecular clock (Fig. 2) recovered the same topology as ML analysis in IQTree with strong nodal support (Fig. #). Both analyses strongly support *Osedax* is most closely related to the Vestimentifera + *Sclerolinum* clade and Frenulata is the early diverging group, as recently reported (Li et al. 2015). Within Vestimentifera, *Lamellibrachia* is sister to the remaining sampled vestimentiferans.

Molecular clock analyses based on phylogenomic dataset suggest modern siboglinid diversity originated in Mesozoic (223MYA ± 80 MY), conflicting with previous hypotheses indicate a Late Mesozoic or Cenozoic approximately 50-126 MYA. The previous analyses are solely based on estimation of nucleotide divergence of COI sequences on limited taxa sampling (mainly vestimentiferans). However, mitochondrial genes of vestiemntiferans may have experienced a “slow-down” in rate of nucleotide subsitution relative to other siboglinid lineages (Li et al. 2015). Moreover, molecular clock analysis suggest a yonger split of Vestimentifernas during the Cenozoic (60 MYA). Recent analyses suggested that Jurassic tubes from Figueroa deposits is likely to have been made by vestimentiferans (Georgieva et al. 2017), but this hypothesisis is not supported in our analysis. Thus, we provide the molecular clock of the siboglinid phylogenetic tree, placing a common siboglinid ancestor as far back as the Jurassic. Siboglinid vestimentifernas might exploit and adapt to vents and seeps during Cenozoic (<65 MYA). This adds to the growing evidence that the Cenozoic was a key period for the radiation of most dominant invertebrate taxa now occupying in deep-sea chomosynthetic communities (Vrijenhoek 2013).

## Conclusion

## Ackowledgements

## Author contribution

## Supplemental Information

Albertin, Caroline B, Oleg Simakov, Therese Mitros, Z Yan Wang, Judit R Pungor, Eric Edsinger-Gonzales, Sydney Brenner, Clifton W Ragsdale, and Daniel S Rokhsar. 2015. “The Octopus Genome and the Evolution of Cephalopod Neural and Morphological Novelties.” *Nature* 524 (7564): 220.

Andrews, Simon, and others. 2010. “FastQC: A Quality Control Tool for High Throughput Sequence Data.”

Arp, Alissa J, and James J Childress. 1981. “Blood Function in the Hydrothermal Vent Vestimentiferan Tube Worm.” *Science* 213 (4505): 342–44.

Bailly, Xavier, Didier Jollivet, Stephano Vanin, Jean Deutsch, Franck Zal, François Lallier, and André Toulmond. 2002. “Evolution of the Sulfide-Binding Function Within the Globin Multigenic Family of the Deep-Sea Hydrothermal Vent Tubeworm Riftia Pachyptila.” *Molecular Biology and Evolution* 19 (9): 1421–33.

Boetzer, Marten, and Walter Pirovano. 2014. “SSPACE-Longread: Scaffolding Bacterial Draft Genomes Using Long Read Sequence Information.” *BMC Bioinformatics* 15 (1): 211.

Bolger, Anthony M, Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics* 30 (15): 2114–20.

Bouckaert, Remco, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A Suchard, Andrew Rambaut, and Alexei J Drummond. 2014. “BEAST 2: A Software Platform for Bayesian Evolutionary Analysis.” *PLoS Computational Biology* 10 (4): e1003537.

Bright, M, H Keckeis, and CR Fisher. 2000. “An Autoradiographic Examination of Carbon Fixation, Transfer and Utilization in the Riftia Pachyptila Symbiosis.” *Marine Biology* 136 (4): 621–32.

Buchfink, Benjamin, Chao Xie, and Daniel H Huson. 2014. “Fast and Sensitive Protein Alignment Using Diamond.” *Nature Methods* 12 (1): 59.

Chen, Nansheng. 2004. “Using Repeatmasker to Identify Repetitive Elements in Genomic Sequences.” *Current Protocols in Bioinformatics* 5 (1): 4–10.

Chu, Hiutung, and Sarkis K Mazmanian. 2013. “Innate Immune Recognition of the Microbiota Promotes Host-Microbial Symbiosis.” *Nature Immunology* 14 (7): 668.

Danise, Silvia, and Nicholas D Higgs. 2015. “Bone-Eating Osedax Worms Lived on Mesozoic Marine Reptile Deadfalls.” *Biology Letters* 11 (4): 20150072.

De Bie, Tijl, Nello Cristianini, Jeffery P Demuth, and Matthew W Hahn. 2006. “CAFE: A Computational Tool for the Study of Gene Family Evolution.” *Bioinformatics* 22 (10): 1269–71.

Du, Zhou, Xin Zhou, Yi Ling, Zhenhai Zhang, and Zhen Su. 2010. “AgriGO: A Go Analysis Toolkit for the Agricultural Community.” *Nucleic Acids Research* 38 (suppl\_2): W64–W70.

Dunn, Casey W, Mark Howison, and Felipe Zapata. 2013. “Agalma: An Automated Phylogenomics Workflow.” *BMC Bioinformatics* 14 (1): 330.

Flores, Jason F, Charles R Fisher, Susan L Carney, Brian N Green, John K Freytag, Stephen W Schaeffer, and William E Royer. 2005. “C.” *Proceedings of the National Academy of Sciences* 102 (8): 2713–8.

Georgieva, Magdalena N, Crispin TS Little, Jonathan S Watson, Mark A Sephton, Alexander D Ball, and Adrian G Glover. 2017. “Identification of Fossil Worm Tubes from Phanerozoic Hydrothermal Vents and Cold Seeps.” *Journal of Systematic Palaeontology*, 1–43.

Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. “QUAST: Quality Assessment Tool for Genome Assemblies.” *Bioinformatics* 29 (8): 1072–5.

Haas, Brian J, Arthur L Delcher, Stephen M Mount, Jennifer R Wortman, Smith JrRoger K, Linda I Hannick, Rama Maiti, et al. 2003. “Improving the Arabidopsis Genome Annotation Using Maximal Transcript Alignment Assemblies.” *Nucleic Acids Research* 31 (19): 5654–66.

Haas, Brian J, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, et al. 2013. “De Novo Transcript Sequence Reconstruction from Rna-Seq Using the Trinity Platform for Reference Generation and Analysis.” *Nature Protocols* 8 (8): 1494.

Haas, Brian J, Steven L Salzberg, Wei Zhu, Mihaela Pertea, Jonathan E Allen, Joshua Orvis, Owen White, C Robin Buell, and Jennifer R Wortman. 2008. “Automated Eukaryotic Gene Structure Annotation Using Evidencemodeler and the Program to Assemble Spliced Alignments.” *Genome Biology* 9 (1): 1.

Kajitani, Rei, Kouta Toshimoto, Hideki Noguchi, Atsushi Toyoda, Yoshitoshi Ogura, Miki Okuno, Mitsuru Yabana, et al. 2014. “Efficient de Novo Assembly of Highly Heterozygous Genomes from Whole-Genome Shotgun Short Reads.” *Genome Research*, gr–170720.

Katoh, Kazutaka, and Daron M Standley. 2013. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” *Molecular Biology and Evolution* 30 (4): 772–80.

Käll, Lukas, Anders Krogh, and Erik LL Sonnhammer. 2007. “Advantages of Combined Transmembrane Topology and Signal Peptide Prediction—the Phobius Web Server.” *Nucleic Acids Research* 35 (suppl\_2): W429–W432.

Käll, Lukas, John D Storey, Michael J MacCoss, and William Stafford Noble. 2007. “Assigning Significance to Peptides Identified by Tandem Mass Spectrometry Using Decoy Databases.” *Journal of Proteome Research* 7 (01): 29–34.

Keller, Andrew, Alexey I Nesvizhskii, Eugene Kolker, and Ruedi Aebersold. 2002. “Empirical Statistical Model to Estimate the Accuracy of Peptide Identifications Made by Ms/Ms and Database Search.” *Analytical Chemistry* 74 (20): 5383–92.

Kim, Daehwan, Ben Langmead, and Steven L Salzberg. 2015. “HISAT: A Fast Spliced Aligner with Low Memory Requirements.” *Nature Methods* 12 (4): 357.

Langmead, Ben, and Steven L Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357.

Li, Bo, and Colin N Dewey. 2011. “RSEM: Accurate Transcript Quantification from Rna-Seq Data with or Without a Reference Genome.” *BMC Bioinformatics* 12 (1): 323.

Li, Heng. 2018. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” *Bioinformatics* 1: 7.

Li, Weizhong, and Adam Godzik. 2006. “Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences.” *Bioinformatics* 22 (13): 1658–9.

Li, Yuanning, Kevin M Kocot, Christoffer Schander, Scott R Santos, Daniel J Thornhill, and Kenneth M Halanych. 2015. “Mitogenomics Reveals Phylogeny and Repeated Motifs in Control Regions of the Deep-Sea Family Siboglinidae (Annelida).” *Molecular Phylogenetics and Evolution* 85: 221–29.

Li, Yuanning, Kevin M Kocot, Nathan V Whelan, Scott R Santos, Damien S Waits, Daniel J Thornhill, and Kenneth M Halanych. 2017. “Phylogenomics of Tubeworms (Siboglinidae, Annelida) and Comparative Performance of Different Reconstruction Methods.” *Zoologica Scripta* 46 (2): 200–213.

Li, Yuanning, Mark R Liles, and Kenneth M Halanych. 2018. “Endosymbiont Genomes Yield Clues of Tubeworm Success.” *The ISME Journal* 12 (11): 2785.

Little, Crispin TS. 2002. “The Fossil Record of Hydrothermal Vent Communities.” *Cahiers de Biologie Marine* 43 (3/4): 313–16.

Little, Crispin TS, and Robert C Vrijenhoek. 2003. “Are Hydrothermal Vent Animals Living Fossils?” *Trends in Ecology & Evolution* 18 (11): 582–88.

Lowe, Todd M, and Sean R Eddy. 1997. “TRNAscan-Se: A Program for Improved Detection of Transfer Rna Genes in Genomic Sequence.” *Nucleic Acids Research* 25 (5): 955.

Luo, Ruibang, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, et al. 2012. “SOAPdenovo2: An Empirically Improved Memory-Efficient Short-Read de Novo Assembler.” *Gigascience* 1 (1): 18.

Luo, Yi-Jyun, Miyuki Kanda, Ryo Koyanagi, Kanako Hisata, Tadashi Akiyama, Hirotaka Sakamoto, Tatsuya Sakamoto, and Noriyuki Satoh. 2018. “Nemertean and Phoronid Genomes Reveal Lophotrochozoan Evolution and the Origin of Bilaterian Heads.” *Nature Ecology & Evolution* 2 (1): 141.

Marçais, Guillaume, and Carl Kingsford. 2011. “A Fast, Lock-Free Approach for Efficient Parallel Counting of Occurrences of K-Mers.” *Bioinformatics* 27 (6): 764–70.

Mi, Huaiyu, Xiaosong Huang, Anushya Muruganujan, Haiming Tang, Caitlin Mills, Diane Kang, and Paul D Thomas. 2016. “PANTHER Version 11: Expanded Annotation Data from Gene Ontology and Reactome Pathways, and Data Analysis Tool Enhancements.” *Nucleic Acids Research* 45 (D1): D183–D189.

Minic, Zoran, and Guy Hervé. 2003. “Arginine Metabolism in the Deep Sea Tube Worm Riftia Pachyptila and Its Bacterial Endosymbiont.” *Journal of Biological Chemistry*.

Nesvizhskii, Alexey I, Andrew Keller, Eugene Kolker, and Ruedi Aebersold. 2003. “A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry.” *Analytical Chemistry* 75 (17): 4646–58.

Newton, Irene LG, Peter R Girguis, and Colleen M Cavanaugh. 2008. “Comparative Genomics of Vesicomyid Clam (Bivalvia: Mollusca) Chemosynthetic Symbionts.” *BMC Genomics* 9 (1): 585.

Nguyen, Lam-Tung, Heiko A Schmidt, Arndt von Haeseler, and Bui Quang Minh. 2014. “IQ-Tree: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.” *Molecular Biology and Evolution* 32 (1): 268–74.

Nussbaumer, Andrea D, Charles R Fisher, and Monika Bright. 2006. “Horizontal Endosymbiont Transmission in Hydrothermal Vent Tubeworms.” *Nature* 441 (7091): 345.

O’Connell, Jared, Ole Schulz-Trieglaff, Emma Carlson, Matthew M Hims, Niall A Gormley, and Anthony J Cox. 2015. “NxTrim: Optimized Trimming of Illumina Mate Pair Reads.” *Bioinformatics* 31 (12): 2035–7.

Petersen, Thomas Nordahl, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. 2011. “SignalP 4.0: Discriminating Signal Peptides from Transmembrane Regions.” *Nature Methods* 8 (10): 785.

Ponnudurai, Ruby, Manuel Kleiner, Lizbeth Sayavedra, Jillian M Petersen, Martin Moche, Andreas Otto, Dörte Becher, et al. 2017. “Metabolic and Physiological Interdependencies in the Bathymodiolus Azoricus Symbiosis.” *The ISME Journal* 11 (2): 463.

Pryszcz, Leszek P, and Toni Gabaldón. 2016. “Redundans: An Assembly Pipeline for Highly Heterozygous Genomes.” *Nucleic Acids Research* 44 (12): e113–e113.

Sanfilippo, Rossana, Antonietta Rosso, Agatino Reitano, and Gianni Insacco. 2017. “First Record of Sabellid and Serpulid Polychaetes from the Permian of Sicily.” *Acta Palaeontologica Polonica* 62 (1): 25–38.

Schauer, Kevin L, Dana M Freund, Jessica E Prenni, and Norman P Curthoys. 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.

Searle, Brian C, Mark Turner, and Alexey I Nesvizhskii. 2008. “Improving Sensitivity by Probabilistically Combining Results from Multiple Ms/Ms Search Methodologies.” *The Journal of Proteome Research* 7 (1): 245–53.

Slater, Guy St C, and Ewan Birney. 2005. “Automated Generation of Heuristics for Biological Sequence Comparison.” *BMC Bioinformatics* 6 (1): 31.

Smit, AFA, and R Hubley. 2008. “RepeatModeler Open-1.0.” *Available Fom Http://Www. Repeatmasker. Org*.

Stanke, Mario, Oliver Keller, Irfan Gunduz, Alec Hayes, Stephan Waack, and Burkhard Morgenstern. 2006. “AUGUSTUS: Ab Initio Prediction of Alternative Transcripts.” *Nucleic Acids Research* 34 (suppl\_2): W435–W439.

Törönen, Petri, Alan Medlar, and Liisa Holm. 2018. “PANNZER2: A Rapid Functional Annotation Web Server.” *Nucleic Acids Research*.

Vrijenhoek, Robert C. 2013. “On the Instability and Evolutionary Age of Deep-Sea Chemosynthetic Communities.” *Deep Sea Research Part II: Topical Studies in Oceanography* 92: 189–200.

Vurture, Gregory W, Fritz J Sedlazeck, Maria Nattestad, Charles J Underwood, Han Fang, James Gurtowski, and Michael C Schatz. 2017. “GenomeScope: Fast Reference-Free Genome Profiling from Short Reads.” *Bioinformatics* 33 (14): 2202–4.

Waterhouse, Robert M, Mathieu Seppey, Felipe A Simão, Mosè Manni, Panagiotis Ioannidis, Guennadi Klioutchnikov, Evgenia V Kriventseva, and Evgeny M Zdobnov. 2017. “BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics.” *Molecular Biology and Evolution* 35 (3): 543–48.

Zal, Franck, François H Lallier, Brian N Green, Serge N Vinogradov, and André Toulmond. 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.” *Journal of Biological Chemistry* 271 (15): 8875–81.

Zal, F, T Suzuki, Y Kawasaki, JJ Childress, FH Lallier, and A Toulmond. 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: Structure, Function, and Bioinformatics* 29 (4): 562–74.

Zdobnov, Evgeni M, and Rolf Apweiler. 2001. “InterProScan–an Integration Platform for the Signature-Recognition Methods in Interpro.” *Bioinformatics* 17 (9): 847–48.