Comparative genomics reveal symbiont-host evolution of deep-sea tubeworms (Siboglinidae, Annelida) and wood-boring bivalves (Xylophagaidae, Mollusca)

by

Yuanning Li

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 5, 2017

Keywords: Mitogenomics, Phylogenomics, Endosymbionts, Recruitment sources, Deep-sea, Chemosynthetic

Approved by

Kenneth M. Halanych, Chair, Professor of Biology
Scott R. Santos, Associate Professor of Biology
Jason E. Bond, Professor of Biology
Eric Peatman, Associate Professor in Fisheries

Abstract

Chemosynthetic communities are patchily distributed in deep-sea ecosystems, including hydrothermal vents, methane seeps and whale, wood, and algal falls. Unlike abyssal seafloor communities that suffer from food limitation due to lack of the input of detrital organic particles produced from the overlying water column, reducing sediments provide inorganic chemicals, such as sulfide or methane, to fuel chemoautotrophic endosymbionts, and in turn provide nutrition for hosts. Chemosynthetic habitats often support specialize communities that are attractive a wide-range of other marine species, thus facilitating adaptive radiation and evolutionary novelty. Although a wide variety of invertebrate taxa have evolved chemosynthetic symbioses independently, evolutionary relationships of hosts and symbionts, specific symbiont-host associations, dispersal potential, recruitment sources, reproductive modes of most chemosynthetic marine invertebrates are still largely unknown. Therefore, the objective of this dissertation is to explore evolutionary history and symbiont-host associations on deep-sea dominant fauna annelid Siboglinidae and wood-boring bivalve Xylophagaidae. Chapter 1 provides a brief introduction to the deep-sea chemosynthetic community with an emphasis on siboglinids and xylophgagiads, and outlines the specific aims of the dissertation. Chapter 2 presents a mitogenomic analysis to explore relationships among major lineages of Siboglinidae. Findings suggest that bone-eating *Osedax* is most closely related to the Vestimentifera+*Sclerolinum* clade, rather than Frenulata, as recently reported. My collaborators and I also identified the size variation and repeat motifs of control regions within siboglinid

mitochondrial genomes. Chapter 3 presents a subsequent analysis of siboglinid evolutionary relationship using a phylogenomic approach. Importantly, unlike previous studies, the alternative hypothesis that frenulates and *Osedax* are sister groups to one another was explicitly rejected by an approximately unbiased (AU) test. This result implies that *Osedax*, the only siboglinid lineage with heterotrophic endosymbionts, evolved from a lineage utilizing chemoautotrophic symbionts. My collaborators and I also compared the performances of different phylogenomic reconstruction methods on this empirical dataset. Although different methods showed largely congruent results, my collaborators and I found that a supermatrix method using data partitioning with site-homogenous models potentially outperformed a supermatrix method using the CAT-GTR model and multispecies-coalescence approaches when the amount of missing data varies in a dataset and when taxa susceptible to LBA are included in the analyses. Chapter 4 seeks to understand siboglinid symbiont association in hydrocarbon seeps and muddy sediments and compares them to symbiont genomes from hydrothermal vent regions using a comparative genomic approach. I found that metabolic capacity of seep-dwelling ones are largely similar to vent-living vestimentiferans. However, representative of frenulate endosymbionts, from *Galatehalinum brachiosum*, lack key enzymes associated with rTCA and can only use Calvin cycle for carbon fixation compared to vestimentiferan siboglinids. Therefore, symbionts with higher metabolic flexibility in carbon fixation are hypothesized to allow tubeworms to thrive in more reducing environments, such as seeps and vents. For my last dissertation data chapter, the host-symbiont evolution of the deep-sea wood-boring bivalve in Xylophgagaidae has been

explored using several genomic tools. Mitogenomic analysis indicates that the *Xylophaga* lineage is a paraphyleyic group within Xylophagaid. 2b-RAD SNP analysis reveals that there is no population structure identified across ~500 km, indicating that individuals from same species of xylophagaid in the study region are most likely colonized from the same gene pool. Lastly, metagenomic analysis from *Xylophaga* gill tissue shows xylophaga symbiont genome is closely related to *Teridinibacter* species isolated from their close relatives shipworm Teredinidae. Multiple genes dedicated to processing complex polysaccharides associated with wood falls were identified in partial symbiont genome, suggesting that a similar functional role in these endosymbionts from both shallow and deep wood-boring bivalves.

Acknowledgments

First, I want to thank my advisor, Dr. Ken Halanych for this guidance, support, patience, motivation and friendship during these 5 years of my graduate studies at Auburn University. Ken has been a great advisor who brought me really interested in the deep-sea chemosynthetic communities and I am especially grateful for his great advices and encouragements. I also want to thank my dissertation committee members Drs. Scott Santos, Eric Peatman and Jason Bond. I am especially grateful to Scott Santos on his advice and discussions of bioinfomatics and genomics. I thank members of Molette lab that have provided me with both academic and experiences to live and study abroad. They all have been exceptional colleagues and friends: Kevin Kocot, Pam Brannock, Justin Havird, David Branson, Joie Cannon, Nathan Whelan, Damien Waits, Matt Galaska, Mike Tassia, Viktoria Bogantes and David Weese. I am especially grateful to David Branson and Damien Waits for helping me settle down when I first arrived in this country and Kevin Kocot for tutoring me pretty much everything in bioinformatics, phylogenomics and lab skills. I have been extremely fortunate to have several friends Ning Li, Qifan Zeng, Wei Ge, Qiang Fu and Yulin Jin for their support during my stay in Auburn. Finally, I thank my family for their love, patience and support during my time as a graduate strudent.

Financial support mainly came from Chinese scholarship Council and Auburn University Biological Sciences Department.

Table Contents

List of Tables

List of Illustrations

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of Abbreviations

AU	approximately unbiased
BI	Bayesian inference
cDNA	complimentary DNA
ML	maximum likelihood
Mt	mitochondrial genome
OG	orthology group
OTU	operational taxonomic unit
RAD	restriction site-associated DNA
rDNA	ribosomal DNA
rTCA	reverse tricarboxylic acid cycle

Chapter 1. Introduction to dissertation

1.1. General introduction and background

Discovery of the Pacific deep-sea hydrothermal vents along the Galápagos hotspot in 1977 and subsequent description of abundant biofauna, set the stage for revealing symbioses between chemosynthetic bacteria and eukaryotes (Lonsdale, 1979; Vrijenhoek, 2010). Chemosynthetic communities are patchily distributed in deep-sea ecosystem, including hydrothermal vents, methane seepages and whale and wood falls, continental margins and algal detritus (Dubilier et al., 2008). Unlike abyssal seafloor communities that suffer from food limitation due to lack of the input of detrital organic particles form the water column (Smith et al., 2008), reducing sediments provide inorganic chemicals, such as sulfide or methane, to fuel chemoautotrophic endosymbionts which, in turn, provide nutrition for hosts. Chemosynthetic habitats often support specialize communities that are attractive a wide-range of other marine species, thus facilitating adaptive radiation and evolutionary novelty (Distel et al., 2000). Currently, more than seven marine invertebrate phyla have been recognized to host chemosynthetic symbionts, with hundreds of host species now described (Dubilier et al., 2008). Although a wide variety of invertebrate taxa have evolved chemosynthetic symbioses independently, several specialized organisms (e.g. Siboglinidae annelids; Vesicomyidae, Provannidae and poorly known Xylophagaidae molluscs; Alvinocarididae shrimp) dominate the biomass at deep-sea chemosynthetic habitats (Vrijenhoek, 2010). However, despite their ecological and evolutionary

importance, host and bacterial diversity and evolutionary relationships, mechanisms of symbiont-host associations, dispersal potential of most taxa are still poorly characterized.

1.1.1 Siboglinid phylogeny

The gutless tubeworm *Riftia pachyptila* discovered at the Galapagos vents were first identified as annelids. These animals were later elevated as the phyla Pogonophora and Vestimentifera due to their highly distinctive morphology (Ivanov 1963; Jones 1988), but they were later found to form a monophyletic clade within Annelida (Halanych *et al.* 2002; Southward *et al.* 2005 The *R. pachyptila* was also the first host associated with chemoautotrophic symbioses. Siboglinids are annelid worms that often dominate species in deep-sea chemosynthetic communities (e.g. hydrothermal vents, cold seeps, mud volcanoes, large organic falls; Schulze & Halanych 2003; Halanych 2005). Despite several phylogenetic studies, relationships among major siboglinid lineages lacked resolution (Black *et al.* 1997; Halanych *et al.* 1998, 2001; Rouse et al. 2004; Glover *et al.* 2005, 2013; Li *et al.* 2015). Adult siboglinids are gutless and nutritionally dependent on bacterial endosymbionts, which are typically housed in a specialized organ called the trophosome (Southward *et al.* 2005). To date, approximately 200 species have been described within 4 major siboglinid lineages: Vestimentifera, Monilifera (*Sclerolinum* Southward 1961), *Osedax* Rouse, Goffredi & Vrijenhoek, 2004, and Frenulata (Hilário *et al.* 2011). Each lineage is generally associated with a specific type of reducing habitat and group of bacterial symbionts, with vestimentiferans typically living in hydrothermal vents or

cold seeps, frenulates mainly inhabiting reducing sediments, *Sclerolinum* living on decaying organic matter (e.g. wood or rope) or in reduced sediments, and *Osedax* found on vertebrate bones (Schulze & Halanych 2003; Hilário *et al.* 2011). In regards to siboglinid habitat preference, organic-rich sediments are hypothesized to have been the ancestral habitat types and more derived taxa moved into increasingly reducing habitats such as vents or seeps (Schulze & Halanych 2003).

Prior to my dissertation work, morphological (Rouse, 2001; Schulze, 2003) and molecular (Halanych et al., 2001; Rouse et al., 2004; Rousset et al., 2004; Glover et al., 2005; Glover et al., 2013) approaches applied towards understanding siboglinid phylogeny suggest: (1) siboglinids form a monophyletic clade, (2) Vestimentifera and Frenulata are both monophyletic lineages, and (3) *Sclerolinum* is sister to Vestimentifera. Nonetheless, important aspects of siboglinid evolutionary history were still debated. A recent study (Glover et al., 2013), using combined nuclear small subunit (18S) ribosomal DNA (rDNA), mitochondrial large subunit (16S) rDNA and cytochrome oxidase subunit I (COI) data inferred *Osedax* as sister to Frenulata rather than the Vestimentifera+*Sclerolinum* clade, in contrast to previous reports (Rouse et al., 2004; Glover et al., 2005). Furthermore, habitat preference has been hypothesized to have proceeded from deep-sea muddy environments (similar to where frenulates inhabit) to more specialized reducing environments such as hydrocarbon cold seeps and hydrothermal vents (Schulze and Halanych, 2003). However, nodal support values within Vestementifera have been generally low (Halanych, 2005), obscuring our understanding of habitat shifts as well as the evolutionary

history of the group in general. Thus, important aspects of siboglinid evolution are still elusive and additional phylogenetic analyses are needed towards elucidating them.

1.1.2 Host-symbiont association of siboglinds

Siboglinids endosymbionts are passed through horizontal transmission mode (that is, acquired as free-living symbiont from surrounding environment at each generation) after the settlement of larvae. Previous phylogenetic analysis of hosts and symbionts has also shown an incongruent evolutionary history (Thornhill, et al. 2008). The horizontal symbiosis transmission of siboglinids may promote uptake and retention of bacteria from surrounding habitats and may allow them to exploit new habitats and resource (Nussbaumer et al. 2006; Lane, 2007). During infection stage, symbionts are accumulated at epidermal cells and subsequently migrate inter-and-intracellularly into a mesodermal tissue that will later develop into the trophosome (Nussbaumer, et al. 2006; Bright and Bulgheresi 2010). In the meantime, massive apoptosis of skin tissue and other non-trophosome tissue that containing symbionts have been documented in the developmental process of Rifita (Nussbaumer, et al. 2006). Overall the process is similar to acquisition of vibrio symbioses by Euprymna squid. Furthermore, symbionts are released back into the environment, enriching free-living populations, upon the death of the host (Klose, et al. 2015). Although symbiont transmission between generations is crucial for the persistence of the siboglinid-symbiont association, molecular mechanisms that underpin this specialized infection process have not been fully characterized.

Most siboglinids are generally associated with a single ribotype of γ-proteobactria (sulphur-oxidizing bacteria, but see in methanotrophic symbionts (Schmaljohann and Flugel 1987) to reduce sulfur compounds as election donors and fix CO_2 autotrophically. However, *Osedax* harbor Oceanospirillales in the posterior root region that mediates heterotrophic degradation of organic compounds from vertebrate bones. Interestingly, genomic and proteomic studies suggested that symbionts of *Riftia* and *Tevnia* (Candidatus *Endoriftiua Persephone*) are able to use reductive tricarboxylic acid cycle (rTCA) in addition to previous identified Calvin cycle for CO_2 fixation (Markert, et al. 2007; Robidart, et al. 2008; Gardebrecht, et al. 2012b). Although key enzymes RubisCO genes and ATP citrate lyase (ACL) type II gene sequences associated with these two carbon fixation cycles were identified in *Lamellibrachi*a and *Escaripia* symbionts using a traditional PCR approach (Thiel, et al. 2012), metabolic machineries based on genomic scale of seep-dwelling vestimentifernans and frenulates still needs to be further explored.

Soon after the discovery of hydrothermal vents, tubeworms were also found at cold-water hydrocarbon seeps in the Gulf of Mexico (GOM). Seep-dwelling vestimentiferans also harbor sulfide-oxidizing bacteria, and are dominant fauna in some seeps that they form a special habitat that is attractive to other marine species (Boetius 2005). Comparing to vent environment that often have a lifespan of only a few decades, seep habitats are generally much less dynamic and gas seepage caused by geological processes can last for centuries (Fisher et al., 1997). Seep vestimentiferans are often thinner, have much slower growth rates, and live longer than their vent relatives (Boetius 2005; Vrijenhoek 2010). For example, individuals of *Lamellibrachia luymesi*

are estimated to live up to 250 years found in GOM, representing one of the longest living animals on earth (Bergquist et al., 2000).

To date, only endosymbiont genomes from *Osedax* and vent-living vestimentiferans (*Rifitia*, *Tevnia* – Gardebrecht et al., 2012 and *Ridgeia* – Perez and Juniper 2016) have been sequenced and characterized. All vent-living vestimentiferan symbionts belong to the same species as "*Ca*. Endoriftia persephone", which indicates the same or close related symbiont species are able to persistent a symbioses with substantially different hydrothermal vent habitats (Gardebrecht, et al. 2012b; Perez and Juniper 2016). Moreover, the previous phylogenetic analysis suggested that three major clades were identified in symbiont bacteria associate with siboglinid hosts (besides *Osedax*), with each clade corresponding to a major siboglinid group (vent-living vestimentiferans, seep-dwelling vestimentiferans and frenulates formed their own clade). However, lack of symbiont genomic information in seep-dwelling verstimentiferans and frenulates hinders our understanding of evolutionary trends and metabolic features in siboglinid symbioses. This is particular surprising given that siboglinids are dominant fauna at seeps and frenulates are the most diverse and widely distributed lineage (Thornhill et al., 2008).

1.1.3 Deep-sea wood boring bivalve Xylophagaidae

Wood-boring (xylotrepetic) bivalves have attracted considerable interest for their unusual biology, economic impacts of destructive woods, potential role in carbon cycles and their associated endosymbionts (Distel et al., 2011). Deep-sea Xylophagaids (~100- 7500m) are

dominant fauna in organic-rich wood falls that they support specialize communities that are attractive a wide-range of other marine species, thus facilitating adaptive radiation and evolutionary novelty (Turner, 1973). Wood fall habitats are comparable with whale falls due to their ephemeral distribution in the deep-sea (Distel et al., 2000). Xylophagaids are obligate wood-boring bivalves belonging to the bivalve family Xylophagidae. They contain specialized shell that carries denticles on their anterior beak to bore a hole in the wood, and then utilize their U-shaped diverticulum to collect wood particles (Purchon 1941; Voight 2015). Similar to vent and seep living mussels such as *Bathymodiolus*, endosymbionts have been observed in their gill region (Distel and Roberts 1997). To date, approximately 60 species have been described (www.marinespecies.org) within three major xylophagaid lineages: *Xylophaga*, *Xylophalas* and *Xyloredo* (Turner 2002). However, the distribution, dispersal potential, recruitment patterns, endosymbiont community, evolutionary relationships and other biological features of this clade are still largely unknown (Voight 2007; Romano et al., 2014).

Interest in xylophagaid symbioses has been driven partly by their potential role in marine carbon cycles and a source of novel enzymes for industry (Distel et al., 2011). The ability of both xylophagaids and their close relatives – teredinids (mainly occur in shallow water) to feed on wood is thought to rely on intracellular bacterial endosymbionts contained within their gill region (Distel 2003; Distel et al., 2011, 2017). Although many wood-boring bivalves are thought to host gill-associated symbionts, only the Gammaproteobacteria (e.g. *Teredinibacter turnerae*)

found in shallow-water teredinid shipworms have been well categorized metabolically with the capacity of cellulolysis of wood and nitrogen fixation (Distel et al., 2002; O'Connor et al., 2014). In the xylophagaids, symbionts have been identified in the gills of *Xylophga atlantica* and *X. wahsingtona* (Distel and Roberts 1997), but not yet been cultured and characterized.

Wood-boring bivalves are widely distributed in the deep-sea, extending to the hadal zone, from polar to tropical regions (Stoeckle, 2006). However, most species of shipworms have only been best quantitatively sampled in the North Pacific, fewer than 10 sites has been reported in southern hemisphere which contains 60% of the world ocean (Stoeckle, 2006; Voight et al., 2009). This extreme knowledge gap results from patchy distribution at great depth, which makes quantitative sampling of wood falls very difficult. Despite a patchy deep-sea distribution, wood falls are colonized at a surprisingly short period of time by wood-boring bivalves (Turner 1973; Voight 2008; Romano et al., 2013). High density of individuals of *Xylophaga* species were recorded in wood falls after 3 month deployment in Northwest Atlantic at ~ 1800m (Turner 1973). Xylophagaid larvae are thought to disperse long distances in the water column. Their larvae may have the ability to delay metamorphosis in the absence of wood sources and to recruit rapidly following environmental cues (Gaudron 2016). Nevertheless, the mode of recruitment sources and patterns of these wood bores have not yet been characterized.

1.2 Research objectives

Previous studies addressing siboglinid phylogeny have relied primarily on several morphological characters and genes and have been unable to fully resolve evolutionary relationships among the major lineages of Siboglinidae, especially the phylogenetic position of bone-eating *Osedax* lineage.

In addition to the host phylogeny, endosymbiont genomes have only been characterized in vent communities (e.g. *Rifita*, *Ridgeia* and *Tevnia*). The symbiont-host association in cold seeps and muddy sediments remain poorly characterized. This situation is particularly surprising given that one lineage, Frenulata, generally live in deep-sea muddy sediments, and comprise the most diversity of this group. Additionally, how "free-living" bacteria migrate from the surface tissue to bacterial housing organ after settlement of siboglinid larvae is still not well understood.

Lastly, wood fall habitats are comparable with whale falls due to their ephemeral nature and distribution in the deep-sea. Wood-boring xylophagaid bivalves are dominate fauna in deep-sea wood fall habitats. Surprisingly, very little attention has been drawn in these woodborers despite their potential role in marine carbon cycles, a source of novel enzymes for industry, and extraordinary symbioses. Evolutionary relationships, recruitment sources and symbiont community of this group are still largely now known. Therefore, the research objectives of my Ph.D. dissertation work are as follows:

1. Investigate phylogeny of major siboglinid lineages using a mitogenomic approach.

2. Further investigate siboglinid phylogeny using a phylogenomic approach, as well as compare the performance of different pylogenomic reconstruction methods.
3. Investigate siboglinid symbiont evolution using a comparative genomic approach.
4. Investigate xylophgaid phylogeny, recruitment sources and symbiont community using genomic tools.

1.3 References

Black, M.B., Halanych, K.M., Maas, P.A.Y., Hoeh, W.R., Hashimoto, J., Desbruyeres, D., Lutz, R.A., Vrijenhoek, R.C., 1997. Molecular systematics of vestimentiferan tubeworms from hydrothermal vents and cold-water seeps. Marine Biology 130, 141–149.

Boetius, A., 2005. Microfauna–Macrofauna Interaction in the Seafloor: Lessons from the Tubeworm. PLOS Biology 3, e102. doi:10.1371/journal.pbio.0030102

Bright, M., Bulgheresi, S., 2010. A complex journey: transmission of microbial symbionts. Nature Reviews Microbiology 8, 218–230. doi:10.1038/nrmicro2262

Distel, D.L. and Roberts, S.J., 1997. Bacterial endosymbionts in the gills of the deep-sea wood-boring bivalves *Xylophaga atlantica* and *Xylophaga washingtona*. The Biological Bulletin, 192(2), pp.253-261.

Distel, D.L., Amin, M., Burgoyne, A., Linton, E., Mamangkey, G., Morrill, W., Nove, J., Wood, N., Yang, J., 2011. Molecular phylogeny of Pholadoidea Lamarck, 1809 supports a single origin for xylotrophy (wood feeding) and xylotrophic bacterial endosymbiosis in Bivalvia. Molecular Phylogenetics and Evolution 61, 245–254. doi:10.1016/j.ympev.2011.05.019

Distel, D.L., Baco, A.R., Chuang, E., Morrill, W., Cavanaugh, C. and Smith, C.R., 2000. Marine ecology: Do mussels take wooden steps to deep-sea vents?. Nature, 403(6771), pp.725-726.

Dubilier, N., Bergin, C. and Lott, C., 2008. Symbiotic diversity in marine animals: the art of harnessing chemosynthesis. Nature Reviews Microbiology, 6(10), pp.725-740.

Gardebrecht, A., Markert, S., Sievert, S.M., Felbeck, H., Thürmer, A., Albrecht, D., Wollherr, A., Kabisch, J., Le Bris, N., Lehmann, R., Daniel, R., Liesegang, H., Hecker, M., Schweder, T., 2012. Physiological homogeneity among the endosymbionts of *Riftia pachyptila* and *Tevnia jerichonana* revealed by proteogenomics. ISME J 6, 766–776. doi:10.1038/ismej.2011.137

Gaudron, S.M., Haga, T., Wang, H., Laming, S.R., Duperron, S., 2016. Plasticity in reproduction and nutrition in wood-boring bivalves (*Xylophaga atlantica*) from the Mid-Atlantic Ridge. Mar Biol 163, 213. doi:10.1007/s00227-016-2988-6

Glover, A.G., Källström, B., Smith, C.R., Dahlgren, T.G., 2005. World-wide whale worms? A new species of *Osedax* from the shallow north Atlantic. Proc. R. Soc. B 272, 2587–2592. doi:10.1098/rspb.2005.3275

Glover, A.G., Wiklund, H., Taboada, S., Avila, C., Cristobo, J., Smith, C.R., Kemp, K.M., Jamieson, A.J., Dahlgren, T.G., 2013. Bone-eating worms from the Antarctic: the contrasting fate of whale and wood remains on the Southern Ocean seafloor. Proc. R. Soc. B 280, 20131390. doi:10.1098/rspb.2013.1390

Halanych, K.M., 2005. Molecular phylogeny of siboglinid annelids (a.k.a. pogonophorans): a review. Hydrobiologia 535–536, 297–307. doi:10.1007/s10750-004-1437-6

Halanych, K.M., Feldman, R.A., Vrijenhoek, R.C., 2001. Molecular Evidence that *Sclerolinum brattstromi* Is Closely Related to Vestimentiferans, not to Frenulate Pogonophorans (Siboglinidae, Annelida). Biol Bull 201, 65–75.

Halanych, K.M., Lutz, R.A., Vrijenhoek, R.C., 1998. Evolutionary origins and age of vestimentiferan tube-worms. Cahiers de Biologie Marine.

Hilário, A., Capa, M., Dahlgren, T.G., Halanych, K.M., Little, C.T.S., Thornhill, D.J., Verna, C., Glover, A.G., 2011. New Perspectives on the Ecology and Evolution of Siboglinid Tubeworms. PLoS ONE 6, e16309. doi:10.1371/journal.pone.0016309

Klose, J., Aistleitner, K., Horn, M., Krenn, L., Dirsch, V., Zehl, M., Bright, M., 2016. Trophosome of the Deep-Sea Tubeworm *Riftia pachyptila* Inhibits Bacterial Growth. PLOS ONE 11, e0146446. doi:10.1371/journal.pone.0146446

Lane, C.E., 2007. Bacterial Endosymbionts: Genome Reduction in a Hot Spot. Current Biology 17, R508–R510. doi:10.1016/j.cub.2007.04.035

Lonsdale, P., 1979. A deep-sea hydrothermal site on a strike-slip fault. Nature, 281, pp.531-534.

Markert, S., Arndt, C., Felbeck, H., Becher, D., Sievert, S.M., Hügler, M., Albrecht, D., Robidart, J., Bench, S., Feldman, R.A., Hecker, M., Schweder, T., 2007. Physiological Proteomics of the Uncultured Endosymbiont of *Riftia pachyptila*. Science 315, 247–250. doi:10.1126/science.1132913

Nussbaumer, A.D., Fisher, C.R., Bright, M., 2006. Horizontal endosymbiont transmission in hydrothermal vent tubeworms. Nature 441, 345–348. doi:10.1038/nature04793

Perez, M., Juniper, S.K., 2016. Insights into Symbiont Population Structure among Three Vestimentiferan Tubeworm Host Species at Eastern Pacific Spreading Centers. Appl. Environ. Microbiol. 82, 5197–5205. doi:10.1128/AEM.00953-16

Robidart, J.C., Bench, S.R., Feldman, R.A., Novoradovsky, A., Podell, S.B., Gaasterland, T., Allen, E.E., Felbeck, H., 2008. Metabolic versatility of the *Riftia pachyptila* endosymbiont revealed through metagenomics. Environmental Microbiology 10, 727–737. doi:10.1111/j.1462-2920.2007.01496.x

Romano, C., Voight, J.R., Pérez-Portela, R. and Martin, D., 2014. Morphological and genetic diversity of the wood-boring *Xylophaga* (Mollusca, Bivalvia): new species and records from deep-sea Iberian canyons. PloS one, 9(7), p.e102887

Rouse, G.W., Wilson, N.G., Worsaae, K., Vrijenhoek, R.C., 2015. A dwarf male reversal in bone-eating worms. Curr. Biol. 25, 236–241. doi:10.1016/j.cub.2014.11.032

Rousset, V., Rouse, G.W., Siddall, M.E., Tillier, A., Pleijel, F., 2004. The phylogenetic position of Siboglinidae (Annelida) inferred from 18S rRNA, 28S rRNA and morphological data. Cladistics 20, 518–533. doi:10.1111/j.1096-0031.2004.00039.x

Schulze, A., Halanych, K.M., 2003. Siboglinid evolution shaped by habitat preference and sulfide tolerance. Hydrobiologia 496, 199–205. doi:10.1023/A:1026192715095

Smith, C.R., De Leo, F.C., Bernardino, A.F., Sweetman, A.K. and Arbizu, P.M., 2008. Abyssal food limitation, ecosystem structure and climate change. Trends in Ecology & Evolution, 23(9), pp.518-528.

Southward, E.C., Schulze, A., Gardiner, S.L., 2005. Pogonophora (Annelida): form and function, in: Bartolomaeus, T., Purschke, G. (Eds.), Morphology, Molecules, Evolution and Phylogeny in Polychaeta and Related Taxa, Developments in Hydrobiology. Springer Netherlands, pp. 227–251.

Stoeckle, M., 2006. Species richness of deep-sea wood-boring clams (subfamily Xylophagainae) from the northeast Pacific (Doctoral dissertation)

Thiel, V., Hügler, M., Blümel, M., Baumann, H.I., Gärtner, A., Schmaljohann, R., Strauss, H., Garbe-Schönberg, D., Petersen, S., Cowart, D.A., Fisher, C.R., Imhoff, J.F., 2012. Widespread Occurrence of Two Carbon Fixation Pathways in Tubeworm Endosymbionts: Lessons from Hydrothermal Vent Associated Tubeworms from the Mediterranean Sea. Front Microbiol 3. doi:10.3389/fmicb.2012.00423

Thornhill, D.J., Wiley, A.A., Campbell, A.L., Bartol, F.F., Teske, A., Halanych, K.M., 2008. Endosymbionts of *Siboglinum fiordicum* and the Phylogeny of Bacterial Endosymbionts in Siboglinidae (Annelida). Biol Bull 214, 135–144.

Turner, R.D., 1973. Wood-boring bivalves, opportunistic species in the deep sea. Science, 180(4093), pp.1377-1379.

Voight, J.R., 2015. Xylotrophic bivalves: aspects of their biology and the impacts of humans. Journal of Molluscan Studies, p.eyv008.

Vrijenhoek, R.C., 2010. Genetic diversity and connectivity of deep-sea hydrothermal vent metapopulations. Molecular ecology, 19(20), pp.4391-4411.

Chapter 2. Mitogenomics reveals phylogeny and repeated motifs in control regions in the deep-sea family Siboglinidae

2.1 Abstract

Deep-sea tubeworms in the annelid family Siboglinidae have drawn considerable interest regarding their ecology and evolutionary biology. As adults, they lack a digestive tract and rely on endosymbionts for nutrition. Moreover, they are important members of chemosynthetic environments including hydrothermal vents, cold seeps, muddy sediments, and whale bones. Evolution and diversification of siboglinids has been associated with host-symbiont relationships and reducing habitats. Despite their importance, the taxonomy and phylogenetics of this clade are debated due to conflicting results. In this study, 10 complete and 2 partial mitochondrial genomes and one transcriptome were sequenced and analyzed to address siboglinid evolution. Notably, repeated nucleotide motifs were found in control regions of these mt genomes, which may explain previous challenges of sequencing siboglinid mt genomes. Phylogenetic analyses of amino acid and nucleotide datasets were conducted in order to infer evolutionary history. Both analyses generally had strong nodal support and suggest *Osedax* is most closely related to the Vestimentifera+*Sclerolinum* clade, rather than Frenulata, as recently reported. These results imply *Osedax*, the only siboglinid lineage with heterotrophic endosymbionts, evolved from a lineage utilizing chemoautotrophic symbionts.

2.2 Introduction

Deep-sea tubeworms, Siboglinidae, are annelids that typically lack a digestive tract and instead rely on endosymbionts for nutrition. Because of their highly unusual morphology and nutritional mode with respect to other annelids, siboglinids were previously classified in the phyla Pogonophora and Vestimentifera. To date, approximately 180 species of siboglinids have been described, with four lineages in Siboglinidae being recognized: Vestimentifera, Frenulata, Monilifera (*Sclerolinum*), and *Osedax* (Fig. 1). Each group is generally associated with a specific habitat type (Schulze and Halanych, 2003; Halanych, 2005; Thornhill et al., 2008; Hilário et al., 2011), with Vestimentiferans typically living in reducing habitats (e.g. hydrothermal vents or cold seeps), frenulates on muddy sediments, *Sclerolinum* species found on decaying organic matter (e.g. woods or ropes) or mud volcanoes (Hilário et al., 2011) and *Osedax* inhabiting large vertebrate (i.e., whale) bones (Rouse et al., 2004; Glover et al., 2005). While the first three of these groups host autochemotrophic gamma proteobacteria as endosymbionts (Thornhill et al., 2008), heterotrophic Oceanospirillales are harbored by *Osedax* species (Goffredi et al., 2005).

To date, morphological (Rouse, 2001; Schulze, 2003) and molecular (Halanych et al., 2001; Rouse et al., 2004; Rousset et al., 2004; Glover et al., 2005; Glover et al., 2013) approaches applied towards understanding siboglinid phylogeny suggest: (1) siboglinids form a monophyletic clade, (2) Vestimentifera and Frenulata are both monophyletic lineages, and (3) *Sclerolinum* is sister to Vestimentifera. Nonetheless, important aspects of siboglinid evolutionary

history are still debated. A recent study (Glover et al., 2013), using combined nuclear small subunit (18S) ribosomal DNA (rDNA), mitochondrial large subunit (16S) rDNA and cytochrome oxidase subunit I (COI) data inferred *Osedax* as sister to Frenulata rather than the Vestimentifera+*Sclerolinum* clade, in contrast to previous reports (Rouse et al., 2004; Glover et al., 2005). Furthermore, habitat preference has been hypothesized to have proceeded from deep-sea muddy environments (similar to where frenulates inhabit) to more specialized reducing environments such as hydrocarbon cold seeps and hydrothermal vents (Schulze and Halanych, 2003). However, nodal support values within Vestementifera have been generally low (Halanych, 2005), obscuring our understanding of habitat shifts as well as the evolutionary history of the group in general. Thus, important aspects of siboglinid evolution are still elusive and additional phylogenetic analyses are needed towards elucidating them.

Phylogenetic analyses of mitochondrial (mt) genomes have proven useful in resolving phylogenetic relationships across a wide range of metazoans (e.g. Osigus et al., 2013; Miya et al., 2001). In Bilateria, mt genomes are circular, usually range from 14 to 17kb (but see Shao et al., 2009; Osigus et al., 2013; Boore, 1999) and typically possess 37 genes: 13 protein-coding genes (i.e., atp6, atp8, cox1–3, cob, nad1–6 and nad4l), two ribosomal RNA genes, 22 tRNA genes, and the control region (also called the unknown [UNK] region or D-loop). Within Annelida, arrangement of these genes is relatively conserved (Jennings and Halanych, 2005; Vallès and Boore, 2006; Zhong et al., 2008), but, as of March 2014, complete mitochondrial genomes were only publicly available (i.e., in GenBank) for 17 annelid species. Furthermore,

only two partial siboglinid mitochondrial genomes have so far been sequenced: *Galathealinum brachiosum* (Boore and Brown, 2000) and *Riftia pachyptila* (Jennings and Halanych, 2005). In both cases, difficulties with recovering the "control" or "unknown" region, despite considerably effort, were reported. This region of the mt genome putatively plays a role in controlling transcription and replication of mitochondrial genes (Shadel and Clayton, 1997; Boore and Brown, 2000), which may explain this situation.

Recent advances in high-throughput sequencing and bioinformatics allow novel approaches to sequencing whole mt genomes. To further explore siboglinid phylogeny, including placement of *Osedax* as either sister to Vestimentifera+*Sclerolinum* or Frenulata, and to understand the structure of siboglinid mitochondrial control region, we sequenced mt genomes from representatives of all major siboglinid lineages. These efforts included 10 complete and 2 partial mt genomes, and well as 1 transcriptome, to explore siboglinid evolutionary history.

2.3 Materials and methods

2.3.1 Specimen collection and mitochondrial genome sequencing

Specimen information is shown in Table 1. All were either stored frozen at -80°C or preserved in 80-100% non-denatured ethanol following collection. Due to a limiting amount of tissue from *Osedax mucofloris*, only RNA was extracted since (1) mitochondrial protein-coding and ribosomal RNA genes, which were used in reconstructing the phylogeny of this family, can

be recovered from whole transcriptome sequencing (Neto et al., 2000) and (2) gene order of mt genomes with the siboglinids is highly conserved (see Results and Discussion). RNA was extracted from *Osedax mucofloris* using TRIzol (Invitrogen) and purified using the RNeasy kit (Qiagen) with on-column DNase digestion. Complimentary DNA (cDNA) libraries were constructed using the SMART cDNA library construction kit (Clontech). Total genomic DNA was extracted from all other samples using the DNeasy Blood & Tissue Kit (Qiagen) according to the manufacturer's protocols.

For *Escarpia spicata*, *Seepiophila jonesi*, *Galathealinum brachiosum* and *O. mucofloris*, sequencing of genomic DNA or cDNA were performed by The Genomic Services Lab at the Hudson Alpha Institute in Huntsville, Alabama using Illumina (San Diego, California) 2x100 paired end TruSeq protocols on a Illumina HiSeq 2000 platform. For *Tevnia jerichonana*, *Oasisia alvinae*, *Ridgeia piscesae*, *Sclerolinum brattstromi*, *Siboglinum fiordicum*, *Siboglinum ekmani*, sequencing libraries were prepared from total genomic DNA using Illumina's Nextra DNA sample preparation kit and run on an Illumina Miseq sequencer with 2 × 250 paired-end reads in the Molette Laboratory, Department of Biological Sciences Department, Auburn University.

2.3.2 Mitochondrial genome assemblies and annotation

All Illumina paired-end genomic sequence data were assembled *de novo* using Ray 2.2.0 (Boisvert et al., 2010) with k-mer = 31 (value chosen based on comparing a range of k-mer

values relative to final assembly). To identify putative mt contigs, BLASTN (Altschul et al., 1997) was performed on contigs produced by Ray using the partial mt genome from *Riftia pachyptila* (GenBank Accession AY741662, Jennings and Halanych, 2005) as the query sequence. In 10 out 12 cases, the top-hitting contigs identified via BLASTN represented the entire mitochondrial genome (~15kbp; Table 2). For *R. piscesae* and *S. ekmani*, however, 2 and 3 partial contigs were recovered, respectively. In an attempt to join these partial contigs, Price 1.0.1 (Ruby et al., 2013) was employed to extend existing contigs by iteratively adding raw sequence reads to the contig ends as appropriate using default settings. For *R. piscesae*, two mt contigs increased in size by 14 bp and 89 bp, respectively. In the case of *S. ekmani*, two contigs were bridged together into a single contig. The raw paired-end reads were then remapped to each of their respective draft mt genomes using Bowtie 2-2.2.1 (Langmead and Salzberg, 2012), and visualized in Tablet 1.13.12.17 (Milne et al., 2013).

Annotation of the 13 protein-coding genes (atp6, atp8, cox1–3, cob, nad1–6 and nad4l) and two ribosomal RNAs was conducted with MITOS web server (Bernt et al., 2013) while tRNAs identified via the tRNAscan-SE web server (Lowe and Eddy, 1997; Schattner et al., 2005). Gene boundaries were examined and subsequently adjusted manually by comparison with sequenced siboglinid mt genomes (*Riftia pachyptila*, Boore and Brown, 2000 and *Galathealinum brachiosum*, Jennings and Halanych, 2005) in Artemis (Rutherford et al., 2000), with boundaries of control regions in each mt genome inferred by identifying flanking tRNA sequences. The sequences of control regions were tested for potential tandem repeats by RepeatMasker open-

4.0.3 (Smit et al., unpublished data, www.repeatmasker.org, last accessed 01/08/2014). Secondary structures of putative control regions and their thermodynamic properties were predicted using mfold web server (Zuker, 2003; mfold.rna.albany.edu).

To assemble the *Osedax* transcriptome, raw paired-end reads were first digitally normalized to a k-mer coverage of 30 using the normalize-by-median.py script (Brown et al., 2012; this step discards redundant data, thus decreasing memory usage). Remaining reads were then assembled using Trinity r2013-02-25 (Grabherr et al., 2011) with default settings. Mitochondrial protein-coding genes and ribosomal RNAs were identified by TBLASTX and BLASTN (Altschul et al., 1997), respectively (using the recovered siboglinid mt genomes above as queries). Since transcriptome data were used for *O. mucofloris*, tRNAs sequences as well as specific gene order information are lacking for this taxon. GenBank accession numbers for the above complete and partial mt genomes are provided in Table 2. Mitochondrial genome sequences (accession KJ789161-KJ789172) and individual *Osedax* genes (accession KJ806974-KJ806985) have been deposited to GenBank.

2.3.3 Southern Blot

Based on results from the mt genome assemblies (see Results and Discussion), Southern blot analyses were conducted on a subset of the examined taxa to confirm whether mitochrondrial genomes were circular, rather than linear, in nature. Genomic DNA from *R. pachyptila*, *L. luymesi*, *E. spicata* and *G. brachiosum* was submitted to Transviragen in Chapel

Hill, North Carolina for Southern blotting. Individuals of other species were too small to provide sufficient amounts of high quality DNA for Southern Blotting. For each species, restriction enzymes that would only cut the mt genome once were identified. Based on the mtDNA sequencing in hand, we used the enzymes *SbfI*, *Drd I*, *BamHI*, and *NDel* for *R. pachytila*, *L. luymesi*, *E.spicata*, and *G. brachiosum*, respectively. With these enzymes species combinations, a circular mitochondrial genome would produce a single band of roughly 15Kb, whereas a linear genome would produce two bands of which the largest would be no more than 10.3 Kb in length. Whole genomic DNA was restricted to completion and run on a 0.7% TAE agarose gel adjacent to an unrestricted sample. The agarose gel was then blotted to a nitrocellulose filter and probed with an oligonucleotide designed to hybridize to CO1. The blot was visualized using the PCR DIG Probe Synthesis Kit (Roche)

2.3.4 Phylogenetic analyses

Nineteen Operational Taxonomic Units (OTUs) were included in the phylogenetic analyses. In additional to sequence data generated here, 2 partial siboglinid mitochondrial genome and 4 outgroup species were acquired from GenBank (Table 2). *Helobdella robusta*, *Lumbricus terrestirs*, *Orbinia latreillii* and *Sipunculus nudus* were selected as outgroups based on data availability as well as current understanding of annelid evolutionary history (Struck et al., 2007, 2011; Weigert et al., 2014).

Two data sets were constructed – one being amino acid (AA) and the other being nucleotide (NUC) sequences. Nucleotide sequences were converted into amino acids using the standard invertebrate mitochondrial translation code implemented in Mega 5.2 (Tamura et al., 2011). For amino acid and nucleotide data, each gene was treated as an individual Orthology Group (OG) that was aligned in MUSCLE 3.8.31 (Edgar, 2004), followed by manual correction. All OGs were trimmed using the default settings in Gblocks 0.91b (Talavera and Castresana, 2007) to remove ambiguously aligned regions. The OGs were then concatenated into final datasets using FASconCAT (Kück and Meusemann, 2010) for phylogenetic analysis. The NUC dataset consisted of nucleotide sequences of the 13 protein-coding and the 2 ribosomal RNA genes while the AA dataset included the amino acids sequences of the 13 protein-coding genes only.

Phylogenetic relationships of siboglinids were inferred using maximum likelihood (ML) in RAxML 7.3.8 (Stamatakis, 2006) and Bayesian inference (BI) in PhyloBayes MPI 1.4f (Lartillot et al., 2009). For phylogenetic analyses, ProtTest 2.4 (Abascal et al., 2005) was performed to evaluate all evolutionary models, however, since the MtZoa evolutionary model (Rota-Stabelli et al., 2009) for amino acid data is not available on ProtTest, we evaluated tree topologies based on MtZoa and MtArt+I+G (the best-fit model according to ProtTest) separately, and MtZoa was chosen as the best-fit model because it provided better likelihood scores and less computational time.

For ML analyses, both NUC and AA datasets were partitioned by gene. Analysis of the NUC dataset was done under the GTR (general time reversible) model of substitution rate with a gamma distribution (the GTRGAMMA option) while the AA dataset was analyzed using MtZoa model with a gamma distribution using empirical base frequencies (the PROTGAMMAMTZOAF option). Topological robustness for the ML analysis was evaluated with 100 replicates of nonparametric bootstrapping. Competing phylogenetic hypotheses for both datasets were evaluated using the approximately unbiased (AU) test (Shimodaira, 2002) in CONSEL 0.20 (Shimodaira and Hasegawa, 2001). Per site likelihoods values were determined using RAxML with same evolutionary models.

For BI analyses, the CAT model (Lartillot and Philippe, 2004) was employed for both NUC and AA datasets. The CAT model in Phylobayes is a site heterogeneous model that estimates site-specific substitution rates for the 4 nucleotides or 20 amino acids in an alignment. Thus, for BI analyses, the CAT+GTR and CAT+MtZoa models were employed in analyses of the NUC and AA datasets, respectively. Five parallel chains were each run for 25,000 generations, discarding the first 5,000 generations as burn-in based on log likelihood scores for each chain once stationary was reached. A 50% majority rule consensus tree was computed from the remaining 20,000 trees from each chain, and nodal support was estimated in the post-burnin tree sample, with posterior probability values ≥ 0.95 taken as significant (Huelsenbeck and Rannala, 2004). All phylogenetic analyses were conducted on the Auburn University Molette Laboratory SkyNet server.

2.4 Results

2.4.1 Mt genome composition

Results from the high-throughput sequencing and contig assembly for the 12 mitochondrial genomes and transcriptome from *O. mucofloris* (missing the nad4l, atp8 and small subunit 12S-rDNA genes) are presented in Table 3. Two mt genomes were partial; *R. piscesae* (14,146 bp) was missing the 12S-rDNA and trnV gene and *S. ekmani* (14,838 bp) was missing trnR as well as part of atp6 and the control region. Nine of the complete mt genomes were ~15 kbp in length, varying in size from 14,779 bp (*G. brachiosum*) to 15,581 bp (*Spirobrachia* sp.). However, *S. fiordicum* has a substantially larger mt genome (19,502bp; see below). All genomes had the same gene order (Fig. 2; gene order unknown for *Osedax*).

All the mt genomes exhibited nucleotide and codon biases (Table 4). For example Vestimentifera, *Osedax*, and *Sclerolinum* were ~65% AT whereas Frenulata was ~75% AT. In all of these species, T was the most common base and G the least common base on the coding strand. Furthermore, the anti-G bias was especially pronounced in the third codon position, where G was only present at 1.63% (*G. brachiosum*) to 5.78% (*O. mucofloris*). GC-skew and AT-skew for a given strand were calculated as (G-C)/(G+C) and (A-T)/(A+T), respectively (Perna and Kocher, 1995), with negative values in skewness meaning the coding strand is enriched for T or C. In contrast, positive values infer more As and Gs. On the whole, AT-skew was slightly negative, or positive in the third codon position of vestimetiferans, and GC-skew was more negative than AT-skew (Table 4).

2.4.2 Protein-coding genes

All complete siboglinid mt genomes in this study possessed the 13 protein-coding genes, two ribosomal RNA genes, and 22 tRNA genes (Fig. 2) that are typical of bilaterian mt genomes (Boore, 1999). As in previously reported siboglinid (Boore and Brown, 2000; Jennings and Halanych, 2005) and other annelid mt genomes (e.g., Shen et al., 2009; Zhong et al., 2008; Jennings and Halanych, 2005; Boore and Brown, 1995, 2000), mitochondrial genes sequenced herein are transcribed from the same strand and gene order was conserved. The start and stop codon features of siboglinid mt genomes also showed patterns of bias (Table S1). For example, only ATG is used as an initiation codon, whereas most metazoan mt genomes use a combination of codons (i.e., ATA, ATC, GTG, GCC and GTT; Zhong et al., 2008; Boore and Brown, 2000). Most genes end with the stop codon TAA or TAG. Nevertheless, an incomplete termination codon, either a single T or TA, was observed for several protein gene sequences (Table S1).

2.4.3 Control region

Putative control regions were identified in 12 mt genomes, occurring between *trnR* and *trnH* (Table 5). Although the exact boundaries for this region are difficult to precisely define, all sequences are AT-rich and contain simple repetitive or microsatellite-like motifs (also see Table 5). Notably, lengths of putative control regions highly variable in size, ranging from 186 bp (*T. jerichonana*) to 4,737 bp (*S. fiordicum*). To further investigate size variability in the control region, raw paired-end reads were remapped to the putative control region (Table 3). In general,

presence of repetitive motifs and hairpin-like secondary structures reduced read coverage in most cases. However, *S. fiordicum* putatively showed higher than average coverage in the control region. For example, one ~400 nt sequence was mapped at >2,000X coverage compared to the 213X average across the rest of the control region. Potential secondary structure in this region had been reported in *Lumbricus terrestris* (Boore and Brown, 1995) and *Platynereis dumerilii* (Boore, 2001). However, mfold analyses failed to identify similar potential structures or even conservation of secondary structure among siboglinid mitochondrial control regions (Fig. 4).

Several of the mt genome contigs obtained from assemblies started and stopped in the control region. This could be due to the repetitive nature of the control region (repetitive elements can hinder assembly; Nagarajan and Pop, 2013) or if the molecule is linear. Either situation would offer an explanation on reported difficulties in sequencing the control region of annelids in general (see Introduction). To differentiate between these two possibilities, southern blot experiments were performed. These experiments confirmed the circular nature of the siboglinid mt genomes (*R. pachyptila* and *E. spicata*) by exhibiting a restriction pattern consistent with a circular molecule being linearized rather than a linear molecule being cut into two fragments (Fig. 5). *L. luymesi* failed to cut and *G. brachiosum* DNA degraded.

2.4.4 Phylogenetic analysis

The AA and NUC datasets contained 3,813 and 13,923 parsimony-informative characters, respectively. Both ML and BI (Fig. 3) analyses of the two concatenated datasets yielded

congruent tree topologies with high bootstrap support values (bs) or posterior probabilities (pp). In terms of higher-level relationships, both Vestimentifera and Frenulata were recovered as monophyletic clades with strong support (Fig. 3), consistent with previous molecular (Black et al., 1997; Halanych et al., 1998, 2001) and morphological analyses (Rouse, 2001; Schulze, 2003). Moreover, *Sclerolinum* was recovered sister to Vestimentifera (AA pp=1.00, bs= 100; NUC pp= 0.98, bs=100). Importantly, *Osedax* was placed sister to the *Sclerolinum*+Vestimentifera clade with moderate bootstrap support (bs=90) in analyses of the AA dataset (but pp=0.88) and stronger support (pp= 1.00, bs=90) in analyses of the NUC dataset. Although less likely than our consensus topology (Fig. 3), the hypothesis of *Osedax* as sister to Frenulata was not explicitly rejected by AU test (Table 6). Notably, vestimentiferans generally had shorter branch lengths when compared to other clades. Within Frenulata, *Siboglinum* was not monophyletic; *Spirobrachia* and *Galatheallinum* were nested as a monophyletic clade within the paraphyletic *Siboglinum* (Fig. 3).

2.5 Discussion

In this study, analyses of siboglinid mitochondrial genomes support placement of *Osedax* as sister to a vestimentiferan+*Sclerolinum* clade. Furthermore, although the overall gene order is similar to other annelids, the control regions of siboglinid mt genomes contain highly repetitive elements that generate substantial length variation among lineages.

2.5.1 Siboglinid phylogeny

Phylogenetic analyses supported traditional monophyletic clades of Vestimentifera and Frenulata. As seen in previous morphological (Rouse, 2001) and molecular analyses (Halanych et al., 2001; Rouse et al., 2004; Rousset et al., 2004; Glover et al., 2005; Glover et al., 2013), *Sclerolinum* is closely allied to Vestimentifera. Our analyses support the phylogenetic position of *Osedax* as sister to the Vestimentifera+*Sclerolinum* clade, rather than Frenulata, in agreement with previous studies based on combinations of nuclear 18S, and mitochondrial 16S or COI data (Glover et al., 2005; Rouse et al., 2004) but in contrary to the combined analysis of Glover et al. (2013). Although an Approximately Unbiased test could not reject the hypothesis of *Osedax* as sister to Frenulata, likelihood scores and nodal support values strongly support our consensus topology (Table 6). In addition, similar to Glover et al. (2013), placement of *Osedax* was variable in analyses based on a limited number of genes (i.e., 18S, 16S and COI; Hatleberg, Thornhill, Santos, and Halanych unpublished data). Although the mt genome is a single chromosome, it contains several genes with variable rates of evolution (Mueller, 2006). Our results imply that *Osedax*, which associates with heterotrophic endosymbionts and lives on whale bones, evolved from a lineage depending on chemoautotrophic symbionts that might have dwelled in deep-sea muddy sediments.

Within Vestimentifera, *Lamellibrachia* is sister to the remaining sampled vestimentiferans. Relatively low molecular diversity within vestimentiferans has been observed in all molecular

studies to date (Halanych et al., 1998, 2001; Rouse et al., 2004; Rousset et al., 2004; Glover et al., 2005; Glover et al., 2013), including this study. This limited diversity may be due to selection pressures in extreme habitats or a recent evolutionary origin of this lineage (Halanych, 2005). Within Frenulata, *Siboglinum* is not monophyletic, consistent with previous molecular and morphological studies (reviewed by Halanych, 2005), suggesting that *Siboglinum* characters – for instance, possession of a single tentacle – are symplesomorphies.

2.5.2 Genome composition and structure

Prior to this study, only 17 complete annelid mt genomes, and two partial siboglinid mt genomes, had been reported. Siboglinid mitochondrial genomes exhibit the same gene order as most annelids (Fig. 2); this gene order could be assumed to be a synapomorphy for annelids in general (Jennings and Halanych, 2005; Zhong et al., 2008). Along with this and unlike most metazoans, all genes of annelid mt genomes are transcribed on the same strand, which was hypothesized to prevent inversions that occurred on the non-transcribed strand (Boore, 1999). Additionally, only ATG is used as a start codon, whereas most metazoan mt genomes use a variety of combinations (Boore, 1999). An incomplete stop codon, either a single T or TA, is also common for many protein-genes in the examined siboglinid mt genomes. Incomplete stop codons such as T or TA may be assigned to the adjacent down-stream gene, and then modified to a complete TAA stop codon via post-transcriptional polyadenylation (Zhong et al., 2008; Boore and Brown, 2000; Yuan et al., 2012). Similar patterns are also observed in molluscs (Hrbek and

Farias, 2008; Yuan et al., 2012) and other groups (Ivey and Santos, 2007). Typically, all siboglinid mt genomes sampled to date are characterized by an anti-G bias that is relatively strong at the third codon position, where G is present at an average of only ~3.4%. The low G content may be a result of the asymmetrical replication of the mt genome (Clayton, 1982; Hrbek and Farias, 2008) or a tendency of mutational bias (Boore and Brown, 2000).

2.5.3 Control regions

The control region is often the single longest non-coding region in mt genomes and is believed to play a role in controlling mitochondrial replication and transcription (Clayton, 1991; Zhang and Hewitt, 1997; Boore, 1999). However, this region remains poorly characterized in annelids since previous efforts to sequence it in many annelid mt genomes have been unsuccessful (Boore and Brown, 2000; Jennings and Halanych, 2005; Zhong et al., 2008). We initially hypothesized that siboglinid mt genomes may be linear rather than circular for two reasons. First, the available published siboglinid mt genomes contained control regions that were difficult to amplify by long PCR primers, even though this technique has been successfully employed to amplify entire mt genomes in a variety of metazoan lineages (Yuan et al., 2012; Zhong et al., 2008; Shen et al., 2009). Secondly, a remapping-based approach identified highly repetitive motifs at the ends of the mt genome contigs similar to the telomeres of linear chromosomes (Zakian, 1995; Table 3). To test this hypothesis, Southern blot analyses were performed. Contrary to our predictions, Southern blots implied the mt genomes of siboglinid are

circular. Thus, the most reasonable conclusion is that the control region has historically been difficult to amplify and sequence due to secondary structure formation, such as hairpins, among the tandem repeat regions.

The putative length of the control region among siboglinid mt genomes shows variability among taxa, ranging from 186 bp (*T. jerichonana*) to 4,737 bp (*S. fiordicum*), which represents a nearly 25-fold difference in length. Of these cases, the unusually large control region of *S. fordicum* may be due to false extensions occurring in *de novo* assembly due to this highly repetitive region or independent transposable elements insertions from the nuclear genome, or possibly both. Notably, repeated motifs are more extensive, especially in *S. fiordicum*, compared to previously reported genomes (Boore and Brown, 2000; Jennings and Halanych, 2005; Zhong et al., 2008), and may again signal issues with artificial concatenation of repeats during assembly. Moreover, no obvious conserved regions or secondary structures (Fig. 5) have been observed in control regions across mt genomes from different siboglinid species and clades, except that all contained TA tandem repeats (Table 5). Since intronic microsatellites can affect gene transcription, mRNA splicing or export to cytoplasm (Li et al., 2004), these TA tandem repeats may have functional significance in the mt genomes of siboglinids. Interestingly, the presence of Type 2 transposons has been reported in other annelid mt genomes, which is virtually unknown among other bilaterians (Vallès et al., 2008). Although the present study represents a further step in the characterization of the mitochondrial control region in annelids, the function and reasons for variations or assembly artifacts in this region requires further study.

2.6 Acknowledgement

We thank Miquel Arnedo and two anonymous reviewers for helpful comments on the manuscript; Damien Waits and Amanda Shaver conducted RNA extraction and cDNA library preparation of *O. mucofloris*. This study was supported by awards from the U.S. National Science Foundation (NSF) to KMH, SRS, and DJT (DEB-1036537 and IOS-0843473). Yuanning Li is supported by a scholarship from the China Scholarship Council (CSC) for studying and living abroad. This is Molette Biology Laboratory contribution #35 and Auburn University Marine Biology Program contribution #126.

2.7 References

Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21, 2104–2105. doi:10.1093/bioinformatics/bti263

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25, 3389–3402. doi:10.1093/nar/25.17.3389

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., Pütz, J., Middendorf, M., Stadler, P.F., 2013. MITOS: Improved de novo metazoan mitochondrial genome annotation. Molecular Phylogenetics and Evolution, Mitogenomics and Metazoan Evolution 69, 313–319. doi:10.1016/j.ympev.2012.08.023

Black, M.B., Halanych, K.M., Maas, P. a. Y., Hoeh, W.R., Hashimoto, J., Desbruyères, D., Lutz, R.A., Vrijenhoek, R.C., 1997. Molecular systematics of vestimentiferan tubeworms from hydrothermal vents and cold-water seeps. Marine Biology 130, 141–149. doi:10.1007/s002270050233

Boisvert, S., Laviolette, F., Corbeil, J., 2010. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. J. Comput. Biol. 17, 1519–1533. doi:10.1089/cmb.2009.0238

Boore, J.L., Brown, W.M., 1995. Complete Sequence of the Mitochondrial DNA of the Annelid Worm *Lumbricus Terrestris*. Genetics 141, 305–319.

Boore, J.L., 1999. Animal mitochondrial genomes. Nucl. Acids Res. 27, 1767–1780. doi:10.1093/nar/27.8.1767

Boore, J.L., Brown, W.M., 2000. Mitochondrial Genomes of *Galathealinum*, *Helobdella*, and *Platynereis*: Sequence and Gene Arrangement Comparisons Indicate that Pogonophora Is Not a Phylum and Annelida and Arthropoda Are Not Sister Taxa. Mol Biol Evol 17, 87–106.

Brown, C.T., Howe, A., Zhang, Q., Pyrkosz, A.B., Brom, T.H., 2012. A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. arXiv:1203.4802 [q-bio].

Clayton, D.A., 1982. Replication of animal mitochondrial DNA. Cell 28, 693–705. doi:10.1016/0092-8674(82)90049-6

Clayton, D. A. 1991. Replication and transcription of vertebrate mitochondrial DNA. Annu. Rev. Cell. Biol. 7, 453–78.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl. Acids Res. 32, 1792–1797. doi:10.1093/nar/gkh340

Glover, A.G., Källström, B., Smith, C.R., Dahlgren, T.G., 2005. World-wide whale worms? A new species of *Osedax* from the shallow north Atlantic. Proc. R. Soc. B 272, 2587–2592. doi:10.1098/rspb.2005.3275

Glover, A.G., Wiklund, H., Taboada, S., Avila, C., Cristobo, J., Smith, C.R., Kemp, K.M., Jamieson, A.J., Dahlgren, T.G., 2013. Bone-eating worms from the Antarctic: the contrasting fate of whale and wood remains on the Southern Ocean seafloor. Proc. R. Soc. B 280, 20131390. doi:10.1098/rspb.2013.139

Goffredi, S.K., Orphan, V.J., Rouse, G.W., Jahnke, L., Embaye, T., Turk, K., Lee, R., Vrijenhoek, R.C., 2005. Evolutionary innovation: a bone-eating marine symbiosis. Environmental Microbiology 7, 1369–1378. doi:10.1111/j.1462-2920.2005.00824.x

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotech 29, 644–652. doi:10.1038/nbt.1883

Halanych, K.M., R.A. Lutz, and R.C. Vrijenhoek. 1998. Evolutionary origins and age of vestimentiferan tube worms. Cahiers de Biologie Marine 39: 355–358

Halanych, K.M., 2005. Molecular phylogeny of siboglinid annelids (a.k.a. pogonophorans): a review. Hydrobiologia 535-536, 297–307. doi:10.1007/s10750-004-1437-6

Halanych, K.M., Feldman, R.A., Vrijenhoek, R.C., 2001. Molecular Evidence that *Sclerolinum brattstromi*: Is Closely Related to Vestimentiferans, not to Frenulate Pogonophorans (Siboglinidae, Annelida). Biol Bull 201, 65–75.

Hilário, A., Capa, M., Dahlgren, T.G., Halanych, K.M., Little, C.T.S., Thornhill, D.J., Verna, C., Glover, A.G., 2011. New Perspectives on the Ecology and Evolution of Siboglinid Tubeworms. PLoS ONE 6, e16309. doi:10.1371/journal.pone.0016309

Hrbek, T., Farias, I.P., 2008. The complete mitochondrial genome of the pirarucu (*Arapaima gigas*, Arapaimidae, Osteoglossiformes). Genetics and Molecular Biology 31, 293–302. doi:10.1590/S1415-47572008000200024

Huelsenbeck, J., Rannala, B., 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Syst. Biol. 53, 904–913. doi:10.1080/1063515049052262.

Ivey, J.L., Santos, S.R., 2007. The complete mitochondrial genome of the Hawaiian anchialine shrimp Halocaridina rubra Holthuis, 1963 (Crustacea: Decapoda: Atyidae). Gene 394, 35–44. doi:10.1016/j.gene.2007.01.009

Jennings, R.M., Halanych, K.M., 2005. Mitochondrial Genomes of *Clymenella torquata* (Maldanidae) and *Riftia pachyptila* (Siboglinidae): Evidence for Conserved Gene Order in Annelida. Mol Biol Evol 22, 210–222. doi:10.1093/molbev/msi008

Kück, P., Meusemann, K., 2010. FASconCAT: Convenient handling of data matrices. Molecular Phylogenetics and Evolution 56, 1115–1118. doi:10.1016/j.ympev.2010.04.024

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. doi:10.1038/nmeth.1923

Lartillot, N., Lepage, T., Blanquart, S., 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25, 2286–2288. doi:10.1093/bioinformatics/btp368

Lartillot, N., Philippe, H., 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21, 1095–1109. doi:10.1093/molbev/msh112

Li, Y-C., Korol, A.B., Fahima, T., Nevo, E., 2004. Microsatellites Within Genes: Structure, Function, and Evolution. Mol Biol Evol 21, 991–1007. doi:10.1093/molbev/msh073

Lowe, T.M., Eddy, S.R., 1997. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. Nucl. Acids Res. 25, 0955–964. doi:10.1093/nar/25.5.0955

Milne, I., Stephen, G., Bayer, M., Cock, P.J.A., Pritchard, L., Cardle, L., Shaw, P.D., Marshall, D., 2013. Using Tablet for visual exploration of second-generation sequencing data. Brief. Bioinformatics 14, 193–202. doi:10.1093/bib/bbs012

Miya, M., Kawaguchi, A., Nishida, M., 2001. Mitogenomic Exploration of Higher Teleostean Phylogenies: A Case Study for Moderate-Scale Evolutionary Genomics with 38 Newly Determined Complete Mitochondrial DNA Sequences. Mol Biol Evol 18, 1993–2009.

Mueller, R. L., 2006. Evolutionary rates, divergence dates, and the performance of mitochondrial genes in Bayesian phylogenetic analysis. Syst Biol, 55.2, 289-300. doi: 10.1080/10635150500541672.

Nagarajan, N., Pop, M., 2013. Sequence assembly demystified. Nat Rev Genet 14, 157–167. doi:10.1038/nrg3367

Neto, E.D., Correa, R.G., Verjovski-Almeida, S., Briones, M.R.S., Nagai, M.A., Silva, W. da, Zago, M.A., Bordin, S., Costa, F.F., Goldman, G.H., Carvalho, A.F., Matsukuma, A., Baia, G.S., Simpson, D.H., Brunstein, A., Oliveira, P.S.L. de, Bucher, P., Jongeneel, C.V., O'Hare, M.J., Soares, F., Brentani, R.R., Reis, L.F.L., Souza, S.J. de, Simpson, A.J.G., 2000. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. PNAS 97, 3491–3496. doi:10.1073/pnas.97.7.3491

Osigus, H.-J., Eitel, M., Bernt, M., Donath, A., Schierwater, B., 2013. Mitogenomics at the base of Metazoa. Molecular Phylogenetics and Evolution 69, 339–351. doi:10.1016/j.ympev.2013.07.016

Perna, N.T., Kocher, T.D., 1995. Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. J Mol Evol 41, 353–358. doi:10.1007/BF00186547

Rota-Stabelli, O., Yang, Z., Telford, M.J., 2009. MtZoa: A general mitochondrial amino acid substitutions model for animal evolutionary studies. Molecular Phylogenetics and Evolution 52, 268–272. doi:10.1016/j.ympev.2009.01.011

Rouse, G.W., Goffredi, S.K., Vrijenhoek, R.C., 2004. *Osedax*: Bone-Eating Marine Worms with Dwarf Males. Science 305, 668–671. doi:10.1126/science.1098650

Rouse, G.W., 2001. A cladistic analysis of Siboglinidae Caullery, 1914 (Polychaeta, Annelida): formerly the phyla Pogonophora and Vestimentifera. Zoological Journal of the Linnean Society 132, 55–80. doi:10.1111/j.1096-3642.2001.tb02271.x

Rousset, V., Rouse, G.W., Siddall, M.E., Tillier, A., Pleijel, F., 2004. The phylogenetic position of Siboglinidae (Annelida) inferred from 18S rRNA, 28S rRNA and morphological data. Cladistics 20, 518–533. doi:10.1111/j.1096-0031.2004.00039.x

Ruby, J.G., Bellare, P., Derisi, J.L., 2013. PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. G3 (Bethesda) 3, 865–880. doi:10.1534/g3.113.005967

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.-A., Barrell, B., 2000. Artemis: sequence visualization and annotation. Bioinformatics 16, 944–945. doi:10.1093/bioinformatics/16.10.944

Schattner, P., Brooks, A.N., Lowe, T.M., 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucl. Acids Res. 33, W686–W689. doi:10.1093/nar/gki366

Schulze, A., 2003. Phylogeny of Vestimentifera (Siboglinidae, Annelida) inferred from morphology. Zoologica Scripta 32, 321–342. doi:10.1046/j.1463-6409.2003.00119.x

Schulze, A., Halanych, K.M., 2003. Siboglinid evolution shaped by habitat preference and sulfide tolerance. Hydrobiologia 496, 199–205. doi:10.1023/A:1026192715095

Shadel, G.S., Clayton, D.A., 1997. Mitochondrial DNA Maintenance in Vertebrates. Annual Review of Biochemistry 66, 409–435. doi:10.1146/annurev.biochem.66.1.409

Shao, R., Kirkness, E.F., Barker, S.C., 2009. The single mitochondrial chromosome typical of animals has evolved into 18 minichromosomes in the human body louse, *Pediculus humanus*. Genome Res. 19, 904–912. doi:10.1101/gr.083188.108

Shen, X., Ma, X., Ren, J., Zhao, F., 2009. A close phylogenetic relationship between Sipuncula and Annelida evidenced from the complete mitochondrial genome sequence of *Phascolosoma esculenta*. BMC Genomics 10, 136. doi:10.1186/1471-2164-10-136

Shimodaira, H., 2002. An Approximately Unbiased Test of Phylogenetic Tree Selection. Syst Biol 51, 492–508. doi:10.1080/10635150290069913

Shimodaira, H., Hasegawa, M., 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17, 1246–1247. doi:10.1093/bioinformatics/17.12.124

Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690. doi:10.1093/bioinformatics/btl446

Struck, T.H., Paul, C., Hill, N., Hartmann, S., Hösel, C., Kube, M., Lieb, B., Meyer, A., Tiedemann, R., Purschke, G., Bleidorn, C., 2011. Phylogenomic analyses unravel annelid evolution. Nature 471, 95–98. doi:10.1038/nature09864

Struck, T.H., Schult, N., Kusen, T., Hickman, E., Bleidorn, C., McHugh, D., Halanych, K.M., 2007. Annelid phylogeny and the status of Sipuncula and Echiura. BMC Evolutionary Biology 7, 57. doi:10.1186/1471-2148-7-57

Talavera, G., Castresana, J., 2007. Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. Syst Biol 56, 564–577. doi:10.1080/10635150701472164

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731–2739. doi:10.1093/molbev/msr121

Thornhill, D.J., Wiley, A.A., Campbell, A.L., Bartol, F.F., Teske, A., Halanych, K.M., 2008. Endosymbionts of *Siboglinum fiordicum* and the phylogeny of bacterial endosymbionts in Siboglinidae (Annelida). Biol. Bull. 214, 135–144.

Vallès, Y., Boore, J.L., 2006. Lophotrochozoan mitochondrial genomes. Integr. Comp. Biol. 46, 544–557. doi:10.1093/icb/icj056

Vallès, Y., Halanych, K.M., Boore, J.L., 2008. Group II Introns Break New Boundaries: Presencein a Bilaterian's Genome. PLoS ONE 3, e1488. doi:10.1371/journal.pone.0001488

Weigert, A., Helm, C., Meyer, M., Nickel, B., Arendt, D., Hausdorf, B., Santos, S.R., Halanych, K.M., Purschke, G., Bleidorn, C., Struck, T.H., 2014. Illuminating the base of the annelid tree using transcriptomics. Mol Biol Evol. 31: 1391-1401..doi:10.1093/molbev/msu080

Yuan, Y., Li, Q., Yu, H., Kong, L., 2012. The Complete Mitochondrial Genomes of Six Heterodont Bivalves (Tellinoidea and Solenoidea): Variable Gene Arrangements and Phylogenetic Implications. PLoS ONE 7, e32353. doi:10.1371/journal.pone.0032353

Zakian, V.A., 1995. Telomeres: beginning to understand the end. Science 270, 1601–1607.

Zhang, D.X., Hewitt, G.M., 1997. Insect mitochondrial control region: A review of its structure, evolution and usefulness in evolutionary studies. Biochemical Systematics and Ecology 25, 99–120. doi:10.1016/S0305-1978(96)00042-7

Zhong, M., Struck, T.H., Halanych, K.M., 2008. Phylogenetic information from three mitochondrial genomes of Terebelliformia (Annelida) worms and duplication of the methionine tRNA. Gene 416, 11–21. doi:10.1016/j.gene.2008.02.020

Zuker, M., 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31, 3406–3415. doi: 10.1093/nar/gkg595

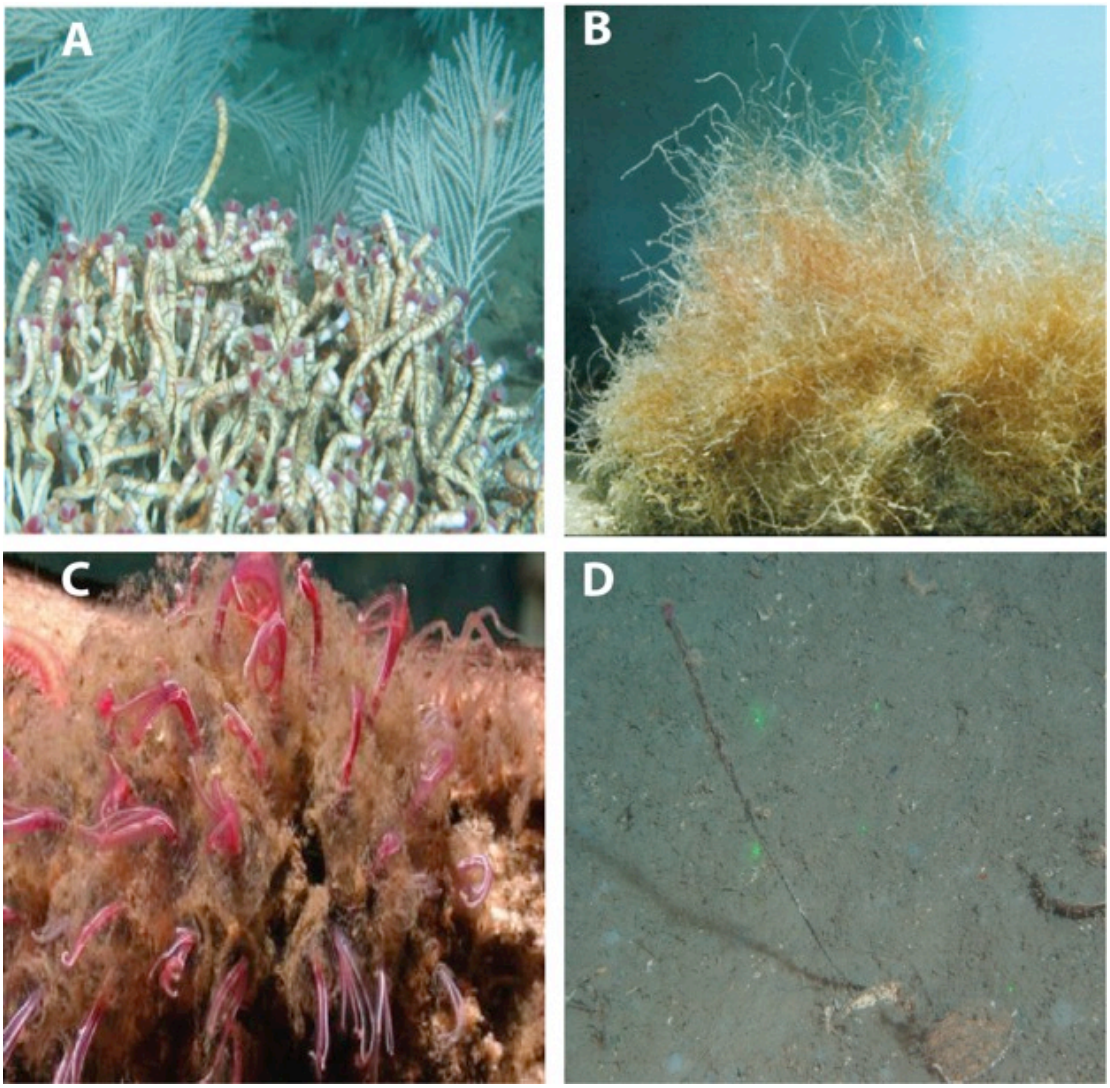


Figure 1. Major siboglinid lineages and their habitat preferences. A) *Lamellibrachia* growing near a hydrocarbon seep. B) *Sclerolinum* inhabiting decaying ropes. C) Bone-eating *Osedax* worms living on a piece of dead gray whale bone in Monterey Canyon (Image courtesy of Monterey Bay Aquarium Research Institute). D) Frenulata species growing in deep-sea muddy habitats.

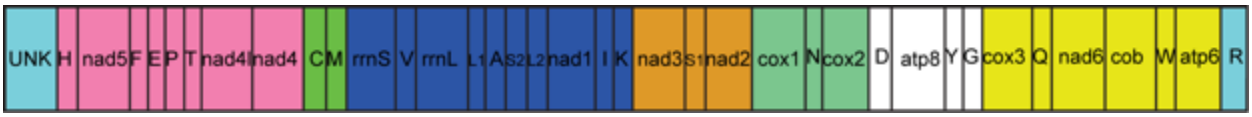


Figure 2. Gene orders of mitochondrial genomes in all Siboglinidae sampled to date. Different colors show conserved gene clusters that were previously reported (Zhong et al., 2008).

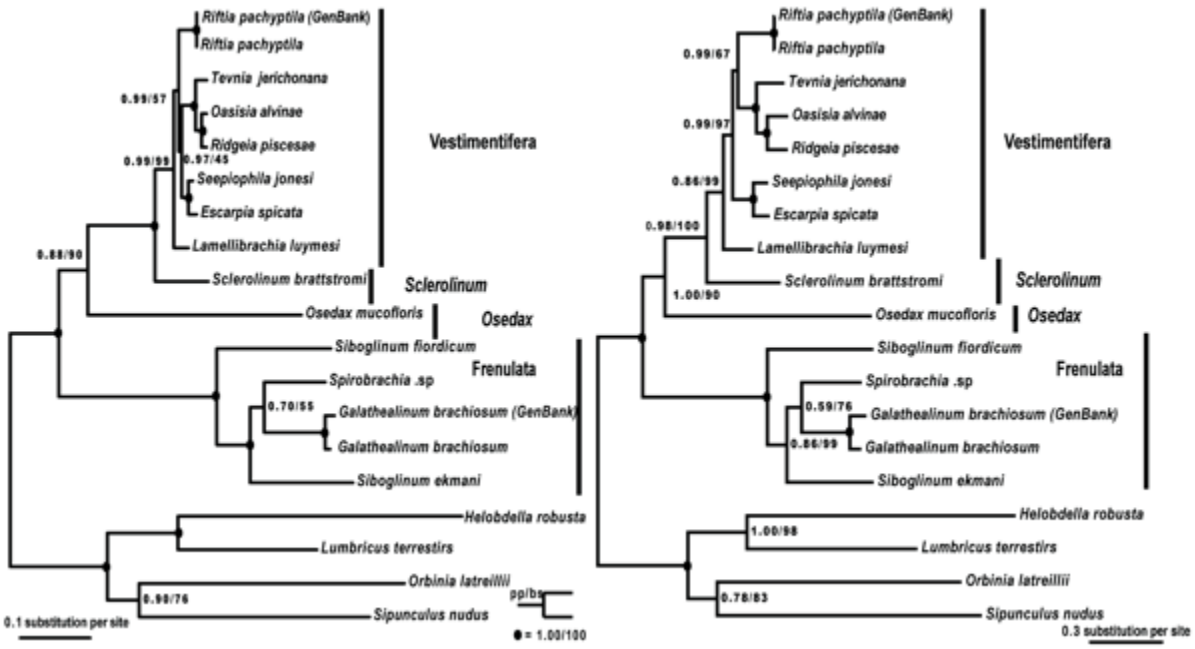


Figure 3. Phylogenetic reconstructions of Siboglinidae based on (A) concatenated amino acids of the 13 mitochondrial protein-coding genes and (B) concatenated nucleotides of the 13 mitochondrial protein-coding and 2 ribosomal RNA genes. Majority rule (50%) consensus phylograms from each of the Bayesian analyses of the concatenated data matrices are shown. Values are shown next to nodes with posterior probabilities left and ML bootstrap support values right. Filled circles indicate fully supported nodes (bs=100, pp=1.00).

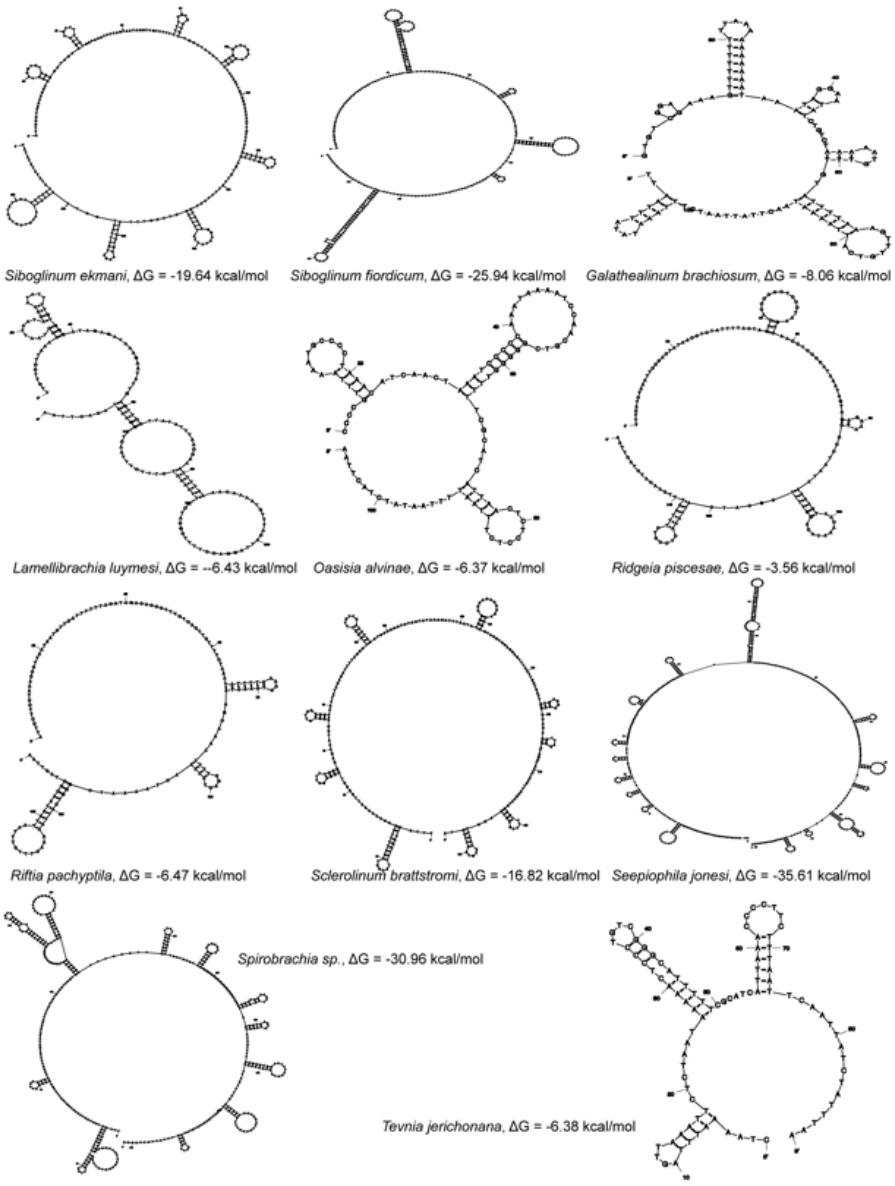


Figure 4. Putative secondary structures and their thermodynamic properties of control regions from 11 taxa.

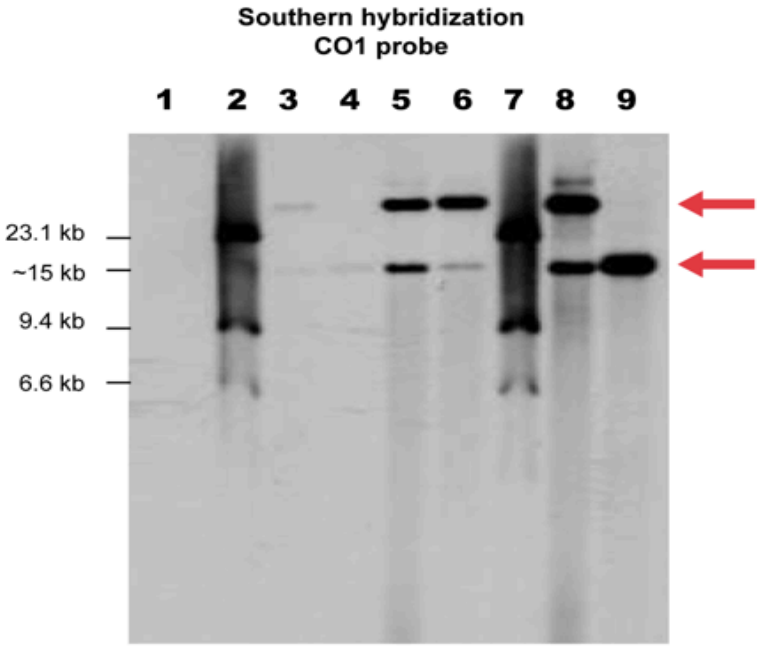


Figure 5. Southern hybridization of the mitochondrial DNA of *R. pachyptila, L. luymesi*, and *E. spicata*. (Lane 1) 1kb Ladder; (lanes 2, 7) DIG Marker (Roche); (lanes 3,5,8) undigested mitochondrial DNA extracted from *R. pachyptila, L. luymesi* and *E. spicata*, respectively. The presence of the ~15kb bank in the undigested samples indicated that DNA contained nicked or fragmented DNA; (lanes 4, 6, 9) digested mitochondrial DNA extracted from from *R. pachyptila, L. luymesi* and *E. spicata*, respectively. Restriction enzyme digestion resulted in detection of only the 15 kb band in *R. pachyptila* and *E. spicata*, suggesting that the mitochondrial genome is circular and cutting at a single site shifts the product to linear DNA. The presence of both the 15 and > 25 kb bands in digested sample *L. luymesi* suggested that the restriction enzyme used did not cut the DNA

Table 1. Specimen data for sequenced taxa.

Spieces	Clade	Specimen Collection		
		location	depth (m)	GPS coordinates
Riftia pachyptila	Vestimentifera	East Pacific Rise	~2522	N 9°50.89′ W 104°17.49′
Tevnia jerichonana	Vestimentifera	East Pacific Rise	~2537	N 9°47.13′ W 104°16.13′
Oasisia alvinae	Vestimentifera	East Pacific Rise	~2630	N 9°48.12′ W 103°56.12′
Ridgeia piscesae	Vestimentifera	Hulk, Canada	2190	N 47°56.95′ W 104°16.87′
Seepiophila jonesi	Vestimentifera	Mississippi Canyon, U.S.	754	N 28°11.58', W 89°47.94'
Escarpia spicata	Vestimentifera	Mississippi Canyon, U.S.	754	N 28°11.58' W 89°47.94'
Lamellibrachia luymesi	Vestimentifera	Mississippi Canyon, U.S.	754	N 28°11.58' W 89°47.94'
Sclerolinum brattstromi	Monilifera	Storfjorden Fjord, Norway	660	N 62°27.26', E 6°47.57'
Siboglinum fiordicum	Frenulata	Skoge Inlet, Norway	36	N 60°16.17' E 5°05.53'
Spirobrachia sp.	Frenulata	Aleutian Trench, U.S.	4890	N 57°27.39' W 148°00.01'
Galathealinum brachiosum	Frenulata	Mississippi Canyon, U.S.	754	N 28°11.58' W 89°47.94'
Siboglinum ekmani	Frenulata	Storfjorden Fjord, Norway	515	N 62°23.30', E 6°54.58'
Osedax mucofloris	*Osedax*	Near Bergen, Norway	N/A	on artificial whale fall

Table 2. Taxa used in phylogenetic analysis.

Species	Clade	Mt Genome Size	GenBank Number
Riftia pachyptila	Siboglinidae, Vestimentifera	14,987 complete	KJ789166
Riftia pachyptila (Genbank)	Siboglinidae, Vestimentifera	12,016 partial	AY741662
Tevnia jerichonana	Siboglinidae, Vestimentifera	14,891 complete	KJ789172
Oasisia alvinae	Siboglinidae, Vestimentifera	14,849 complete	KJ789164
Ridgeia piscesae	Siboglinidae, Vestimentifera	14,146 partial	KJ789165
Seepiophila jonesi	Siboglinidae, Vestimentifera	15,092 complete	KJ789168
Escarpia spicata	Siboglinidae, Vestimentifera	15,445 complete	KJ789161
Lamellibrachia luymesi	Siboglinidae, Vestimentifera	14,991 complete	KJ789163
Sclerolinum brattstromi	Siboglinidae, *Sclerolinum*	15,383 complete	KJ789167
Osedax mucofloris	Siboglinidae, *Osedax*	N/A	*
Siboglinum fiordicum	Siboglinidae, Frenulata	19,502 complete	KJ789170
Spirobrachia sp.	Siboglinidae, Frenulata	15,581 complete	KJ789171
Galathealinum brachiosum	Siboglinidae, Frenulata	14,779 complete	KJ789162
Galathealinum brachiosum (Genbank)	Siboglinidae, Frenulata	7,568 partial	AF178679
Siboglinum ekmani	Siboglinidae, Frenulata	14,838 partial	KJ789169
Helobdella robusta	Hirudinea, Glossiphoniidae	7,553 partial	AF178680
Lumbricus terrestirs	Oligochaeta, Lumbricidae	14,998 complete	NC_001673
Orbinia latreillii	Scolecida, Orbiniidae	15,558 complete	NC_007933
Sipunculus nudus	Polychaeta, Sipunculidae	15,502 complete	NC_011826

* GenBank Numbers of 11 protein-coding and 1 ribosomal RNA genes of O. mucofloris are

KJ806974, KJ806975, KJ806976, KJ806977, KJ806978, KJ806979, KJ806980, KJ806981,

Table 3. Sequencing information..

Species	Number of mt contig(s) recovered	Total reads of mt contig(s)	Average Sequencing Depth of Control Region (X)	Average Sequencing Depth of Control Region (X)
Vestimentifera				
Riftia pachyptila	1	4,632	45	71
Tevnia jerichonana	1	1,484	5	22
Oasisia alvinae	1	2,802	9	39
Ridgeia piscesae	2	1,883	22	29
Seepiophila jonesi	1	32,811	98	220
Escarpia spicata	1	35,127	133	232
Lamellibrachia luymesi	1	4,193	24	67
Sclerolinum				
Sclerolinum brattstromi	1	1,504	21	23
Osedax				
Osedax mucofloris	12	120,480	N/A	802
Frenulata				
Siboglinum fiordicum	1	11,297	213	90
Spirobrachia sp.	1	8,706	103.51	44.62
Galathealinum brachiosum	1	38,282	103	261
Siboglinum ekmani	3	1,969	35	36

Table 4 Structural features of control region.

Species	Size (bp) of Control region	(A+T)%A + ()()()	Proportion of the mt genome (%)	Repeat motifs
Vestimentifera				
Riftia pachyptila	303	81.19	2.02	$(TA)_n$
Tevnia jerichonana	186	81.72	1.25	$(TA)_n$
Oasisia alvinae	147	66.67	0.99	$(TA)_n$
Ridgeia piscesae	309	67.31	2.18	$(TA)_n$
Seepiophila jonesi	381	76.12	2.52	$(TA)_n$
Escarpia spicata	741	82.05	4.8	$(TA)_n$; $(TATATG)_n$
Lamellibrachia luymesi	309	80.91	2.06	$(TA)_n$
Sclerolinum				
Sclerolinum brattstromi	654	63.61	4.25	$(TA)_n$
Osedax				
Osedax mucofloris	N/A	N/A	N/A	N/A
Frenulata				
Siboglinum fiordicum	4737	79.23	24.29	$(TA)_n$; $(CACA)_n$; $(CATA)_n$; $(TATATG)_n$; $(CA)_n$; $(CATATA)_n$; AT rich
Spirobrachia sp.	975	71.90	6.26	$(TA)_n$; $(GA)_n$; AT rich
Galathealinum brachiosum	240	87.92	1.62	$(TA)_n$; $(A)_n$; AT rich
Siboglinum ekmani	645	75.97	4.35	$(TA)_n$; AT rich

Table 5. AU tests of competing phylogenetic hypothesis.

Tree Topology	AA Dataset		NUC Dataset	
	Log-likelihood	AU test (*P*-value)	Log-likelihood	AU test (*P*-value)
Osedax + Vestimentifera/*Sclerolinum*	-58555.90	0.882	-151,727.36	0.914
Osedax + Frenulata	-58570.38	0.118	-151,750.80	0.083

Chapter 3. Phylogenomics of tubeworms and comparative performance of supermatrix versus multispecies-coalescent and Bayesian-Concordance approaches

3.1 Abstract

Deep-sea tubeworms (Annelida, Siboglinidae) represent dominant species in deep-sea chemosynthetic communities (e.g. hydrothermal vents and cold methane seeps) and occur in muddy sediments and organic falls. Siboglinids lack a functional digestive tract as adults, and they rely on endosymbiotic bacteria for energy, making them of evolutionary and physiological interest. Despite their importance, inferred evolutionary history of this group has been inconsistent among studies based on different molecular markers. In particular, placement of bone-eating *Osedax* worms has been unclear in part because of their distinctive biology, including harboring heterotrophic bacteria as endosymbionts, displaying extreme sexual dimorphism, and exhibiting a distinct body plan. Here, we reconstructed siboglinid evolutionary history using 12 newly sequenced transcriptomes. We parsed data into three datasets that accommodated varying levels of missing data, and we evaluate effects of missing data on phylogenomic inference. Additionally, several multispecies-coalescent approaches and Bayesian Concordance Analysis (BCA) were employed to allow for a comparison of results to a supermatrix approach. Every analysis conducted herein strongly supported *Osedax* being most closely related to the Vestimentifera and *Sclerolinum* clade, rather than Frenulata, as previously reported. Importantly, unlike previous studies, the alternative hypothesis that frenulates and

Osedax are sister groups to one another was explicitly rejected by an approximately unbiased (AU) test. Furthermore, although different methods showed largely congruent results, we found that a supermatrix method using data partitioning with site-homogenous models potentially outperformed a supermatrix method using the CAT-GTR model and multispecies-coalescence approaches when the amount of missing data varies in a dataset and when taxa susceptible to LBA are included in the analyses.

3.2 Introduction

Siboglinids are annelid worms that can be the dominant species in deep-sea chemosynthetic communities (e.g. hydrothermal vents, cold seeps, mud volcanoes, large organic falls; Schulze & Halanych 2003; Halanych 2005). Despite several phylogenetic studies, relationships among major siboglinid lineages lack resolution (Black *et al.* 1997; Halanych *et al.* 1998, 2001; Glover *et al.* 2005, 2013; Li *et al.* 2015). These animals were formerly recognized as the phyla Pogonophora and Vestimentifera due to their highly distinctive morphology (Ivanov 1963; Jones 1988), but they were later found to form a monophyletic clade within Annelida (Halanych *et al.* 2002; Southward *et al.* 2005). Adult siboglinids are gutless and nutritionally dependent on bacterial endosymbionts, which are typically housed in a specialized organ called the trophosome (Southward *et al.* 2005). To date, approximately 200 species have been described within 4 major siboglinid lineages: Vestimentifera, Monilifera (*Sclerolinum* Southward 1961), *Osedax* Rouse, Goffredi & Vrijenhoek, 2004, and Frenulata (Hilário *et al.* 2011). Each lineage is

generally associated with a specific type of reducing habitat and group of bacterial symbionts, with vestimentiferans typically living in hydrothermal vents or cold seeps, frenulates mainly inhabiting reducing sediments, *Sclerolinum* living on decaying organic matter (e.g. wood or rope) or in reduced sediments, and *Osedax* found on vertebrate bones (Schulze & Halanych 2003; Hilário *et al.* 2011). In regards to siboglinid habitat preference, organic-rich sediments are hypothesized to have been the ancestral habitat types and more derived taxa moved into increasingly reducing habitats such as vents or seeps (Schulze & Halanych 2003).

Endosymbionts of siboglinids are passed through horizontal transmission mechanisms that promote uptake and retention of bacteria from surrounding habitats and may allow them to exploit new habitats and resources (Nussbaumer *et al.* 2006; Lane 2007). Siboglinids are generally dominated by a single ribotype of chemosynthetic endosymbiont (Southward 1982; Thornhill *et al.* 2008, but see Chao *et al.* 2007; Vrijenhoek *et al.* 2007). Whereas most siboglinids use chemoautotrophic gammaproteobacteria hosted in the trophosome (Thornhill *et al.* 2008), *Osedax* harbor Oceanospirillales in a root-like system that facilitates heterotrophic degradation of large organic compounds from vertebrate bones (Goffredi *et al.* 2005). Unlike other lineages of Siboglinidae, most bone-eating *Osedax* species exhibit extreme male dwarfism (Rouse *et al.* 2004, 2015).

To date, most morphological (Rouse 2001; Schulze 2003) and molecular (Black *et al.* 1997; Halanych *et al.* 2001; Rouse *et al.* 2004; Rousset *et al.* 2004; Glover *et al.* 2005, 2013; Li *et al.* 2015) phylogenetic studies indicate that: (1) Siboglinidae is monophyletic, (2) the four

major groups within Siboglinidae are each monophyletic, (3) Vestimentifera is sister group to *Sclerolinum*, and 4) Frenulata is sister group to all other siboglinids. However, aspects of siboglinid phylogeny are still debated, especially the placement of *Osedax*. In contrast to previous molecular and morphological phylogenetic studies (Rouse *et al.* 2004; Glover *et al.* 2005) that inferred *Osedax* as closely related to the Vestimentifera and *Sclerolinum* clade (Fig. 1A), recent molecular phylogenetic studies using five nuclear and mitochondrial loci reported *Osedax* as the sister group to Frenulata (Glover *et al.* 2013; Rouse *et al.* 2015; Fig. 1B). Additionally, a recent study using whole mitochondrial genomes supported the original hypothesis that *Osedax* is the sister group to the Vestimentifera/*Sclerolinum* clade, but explicit hypothesis testing could not reject the alternative hypothesis of *Osedax* as the sister group to Frenulata (Li *et al.* 2015). Given that mitochondrial genomes represent a single locus and that mitochondrial-based trees occasionally are inaccurate due to introgression, saturation, or selection (Funk & Omland 2003), phylogenetic analyses based on multiple nuclear loci are desirable for elucidating evolutionary history of siboglinids.

The ability to utilize genome scale data for phylogenetic analyses, or "phylogenomics," has significantly improved our understanding of metazoan evolution (Delsuc *et al.* 2005; Matus *et al.* 2006; Dunn *et al.* 2008; Kocot *et al.* 2011; Bond *et al.* 2014; Misof *et al.* 2014; Weigert *et al.* 2014; Whelan *et al.* 2015). Currently, two different systematic approaches are primarily used for phylogenetic inference with large multilocus datasets: (1) the supermatrix (i.e. concatenation) approach and (2) methods that use multispecies-coalescence models to resolve conflict among

independently generated trees (Gatesy & Springer 2014; Edwards *et al.* 2015); methods such as *BEAST (Heled & Drummond 2010) that co-estimate gene and species trees are generally too computationally expensive for phylogenomic sized datasets. However, performance of the supermatrix approach relative to coalescent-based estimation is still debated (Gatesy & Springer 2013; Oliver 2013; Wu *et al.* 2013; Zhong *et al.* 2013, 2014; Springer & Gatesy 2015). The supermatix approach assumes that phylogenetic signal from genes that do not share the species phylogeny will be overwhelmed by the signal from the majority of genes whose genealogy mirrors that of the species evolutionary history (Lanier & Knowles 2012). In contrast, multispecies-coalescent approaches can account for gene tree heterogeneity (Rannala & Yang 2003) by taking incomplete lineage sorting into account. Most multispecies-coalescent approaches (e.g. STAR; Liu *et al.* 2009, MP-EST; Liu *et al.* 2010, NJst; Liu & Yu 2011, and ASTRAL; Mirarab *et al.* 2014) resolve gene tree conflict by estimating species trees from individual gene trees (i.e. gene trees are the required input for multispecies-coalescent methods).

To further explore siboglinid phylogeny, including testing the placement of *Osedax* as the sister group to a clade of Vestimentifera and *Sclerolinum* or to Frenulata (Fig. 1), we sequenced 12 transcriptomes including representatives from all major siboglinid lineages and 3 outgroups. We also evaluated the relative performance of supermatrix approaches employing maximum likelihood and Bayesian inference, multispecies-coalescent methods, and the Bayesian Concordance Analaysis (BCA; Larget *et al.* 2010) with our datasets to understand how these

different approaches performed on inferring evolutionary events that occurred presumably 60-126 millions of years ago (Little & Vrijenhoek 2003; Hilário *et al.* 2011).

3.3 Methods

3.3.1 Taxon sampling, sequencing and assembling

Specimen information is given in Tables 1 and S1. Upon collection, all specimens were either stored at -80°C or preserved in RNAlater (Life Technologies Inc.). RNA extraction and cDNA preparation for high-throughput sequencing followed Kocot *et al*. (2011) and Whelan *et al*. (2015), Briefly, total RNA was extracted using TRIzol (Invitrogen) and purified using the RNeasy kit (Qiagen) with on-column DNase digestion. Next, single strand cDNA libraries were reverse transcribed using the SMART cDNA Library Construction kit (Clontech) followed by double-stranded cDNA synthesis using the Advantage 2 PCR system (Clontech). Illumina sequencing library preparation and sequencing of *Osedax mucofloris* Glover, Kallstrom, Smith & Dahlgren, 2005; *Osedax rubiplumus* Rouse, Goffredi & Vrijenhoek, 2004; *Lamellibrachia luymesi* van der Land & Nørrevang, 1975; *Sclerolinum brattstromi* Webb, 1964; *Siboglinum fiordicum* Webb, 1963; *Siboglinum ekmani* Jägersten, 1956; *Sternaspsis* sp. Otto, 1821*; Flabelligera mundata* Gravier, 1906; and *Cirratulus spectabilis* Kinberg, 1866 were performed by The Genomic Services Lab at the Hudson Alpha Institute in Huntsville, Alabama using 2x100 paired-end sequencing on an Illumina HiSeq 2000 platform (San Diego, California). cDNA for *Escarpia spicata* Jones, 1985*, Galathealinum brachiosum* Ivanov, 1961, *L. luymesi* and

Seepiophila jonesi Gardiner, McMullin & Fisher, 2001 were sent to the University of South Carolina Environmental Genomics Core Facility (Columbia, SC, USA) for Roche 454 GS-FLX sequencing. Additionally, transcriptome data were obtained from the NCBI SRA database (Table 1).

Prior to assembly, Illumina paired-end transcriptome sequence data were digitally normalized to a k-mer coverage of 30 using *normalize-by-median.py* (Brown *et al.* 2012). Remaining reads were then assembled using Trinity r2013-02-25 (Grabherr *et al.* 2011) with default settings. Raw 454 data were assembled using Newbler v2.5.3 (Margulies *et al.* 2005) with -*cdna* and -*large* parameters.

3.3.2 Orthology determination, filtering and data matrix assembly

A brief schematic of the phylogenomic pipeline for this study is shown in Fig. 2A. Putative orthologous groups (OGs) were retrieved from each transcriptome following bioinformatics pipelines of Kocot *et al*. (2011) and Whelan *et al*. (2015). Briefly, each assembled transcriptome was scanned for open reading frames and translated using TransDecoder (Grabherr *et al.* 2011). Initial orthology determination was performed with HaMStR local v13 (Ebersberger *et al.* 2009), and the "Lophotrochozoa-Kocot" core ortholog set, which consisted of 2,046 nuclear genes (Kocot et al. *in press*) using *Capitella teleta* as the reference taxon.

Because missing data can mislead phylogenetic reconstruction (Lemmon *et al.* 2009), three filtering strategies were used to evaluate how missing data may affect phylogenomic performance (Fig. 2B). First, a dataset was generated by removing any OG sampled for fewer than 13 taxa. After creation of this first dataset, we found that *Spirobrachia* sp. had more missing data than other taxa (only 24.5% of total orthologs present, Table 2) and thus it was removed in two subsequent filtering datasets to accommodate more OGs. For these two additional filtering strategies, any gene with fewer than 10 or 12 taxa, respectively, was removed. All three datasets (D98, D150, D289 – numbers refer to numbers of OGs included; Fig. 2B) were processed by first discarding sequences that were shorter than 50 bp. Sequences of each OG were then aligned using MAFFT (Katoh *et al.* 2002) with the "*-auto*" and "*-localpair*" parameters and 1,000 maximum iterations. Uninformative and ambiguously aligned positions were trimmed with Aliscore (Misof & Misof 2009) and Alicut (Kück 2009). Alignment columns with only gaps were subsequently removed, and any OG with an alignment less than 50 bp after trimming were discarded. For each OG, a custom javascript, *AlignmentCompare.java* was used to remove any sequence that did not overlap other sequences by at least 20 amino acids. After these filtering steps, any OG that had fewer than the minimum taxa thresholds of the three filtering strategies (see above) was removed. FastTreeMP (Price *et al.* 2010) with the "*-slow*" and "*-gamma*" parameters was then employed to generate single-gene trees for each OG to screen for suspected paralogs that were then trimmed from the data matrix using PhyloTreePruner (Kocot *et al.* 2013) with a minimum bootstrap support value of 95%. All scripts used for initial orthology

determination, except PhyloTreePruner, can be found at https://github.com/kmkocot/basal_metazoan_phylogenomics_scripts_01-2015.

To further identify potential causes of systematic error, TreSpEx (Struck 2014) and BaCoCa (Kuck & Struck 2014) were employed to examine and parameterize tree-based information to filter potential sources of systematic error from the three datasets generated under different minimum taxon values. To do this, ProtTest 2.4 (Abascal *et al.* 2005) was used to select the best-fitting protein evolutionary model for each OG, and then individual gene trees were inferred using RAxML 8.0.23 (Stamatakis 2014) with 100 fast bootstrap replicates. Next, possible paralogs and exogenous contamination missed by HaMStR and PhyloTreePruner were further filtered using the tree- and blast-based method of TreSpEx. For this method, we used gene trees generated by RAxML and the *Capitella teleta* and *Helobdella robusta* blast databases packaged with TreSpEx. Both "certain" (high-confident paralogs) and "uncertain" (potential paralogs) sequences, as identified by TreSpEx were removed. Standard deviation of LB scores, a metric designed to quantify a gene's potential for causing long-branch attraction (LBA; Struck 2014), was also calculated with TreSpEx. Amino acid compositional heterogeneity for each gene, as measured by relative composition frequency variability (RCVF; Zhong *et al.* 2011), was calculated for each OG from each dataset using BaCoCa (Fig. 6). Both genes with high RCFV values and standard deviation of LB scores can cause systematic error in phylogenetic inference. Therefore, genes with outlier values for both of these metrics were identified based on density

plots generated in R (R Core Development Team, 2015). Outliers were subsequently removed from all three datasets.

3.3.3 Phylogenetic analyses

15 siboglinid taxa were included in phylogenomic analyses. *Sternaspsis* sp.*, F. mundata* and *C. spectabilis* were selected as outgroups based on data availability and current understanding of annelid phylogeny (Struck *et al*. 2011; Weigert *et al*. 2014). Three major approaches were used to reconstruct phylogenetic relationships: supermatrix, multispecies-coalescent methods, and BCA. For the supermatrix approach, matrices of concatenated OGs were analyzed using both maximum likelihood (ML) in RAxML and Bayesian inference (BI) in PhyloBayes 1.5a (Lartillot *et al.* 2009). Prior to ML analyses, PartitionFinderV1.1.1 (Lanfear *et al.* 2012, 2014) was used to evaluate best-fit partition schemes and associated best-fit amino acid substitution models for each partition using 20% relaxed clustering (Lanfear *et al.* 2014). Each ML analyses employed best-fit models and partitions indicated by PartitionFinder and a gamma distribution to model rate heterogeneity. Nodal support for ML analyses was evaluated with 100 fast bootstrap replicates. For BI, the CAT+GTR +Γ model (Lartillot & Philippe 2004) was employed because it accounts for site-specific heterogeneity in the substitution process. PhyloBayes analyses were run with four parallel chains for 10,000-20,000 generations, depending on the datasets. Burn-in of 20% was determined with trace plots as viewed in Tracer (Rambaut *et al.* 2014; available from http://tree.bio.ed.ac.uk/software/tracer/). Chains were

considered to have reached convergence when the maxdiff statistic among chains was below 0.3 (as measured by bpcomp) and Effective Sample Size > 50 for each parameter (as measured by tracecomp). A 50% majority rule consensus tree was computed with bpcomp, and nodal support was estimated by posterior probability (Huelsenbeck & Rannala 2004).

Four multispecies-coalescent approaches (i.e. STAR, MP-EST, NJst, ASTRAL) were also used for phylogenetic inference. Differences in these methods are briefly summarized here. STAR estimates a species tree from average ranks of coalescent units from each rooted gene tree (Liu *et al.* 2009). MP-EST estimates a species tree from a set of rooted individual gene trees by maximizing a pseudo-likelihood function of triplets (Liu *et al.* 2010). In contrast to the former approaches, NJst can incorporate unrooted gene trees to infer a species tree. The NJst method estimates the species tree using neighbor-joining trees built from a distance matrix in which the distance is defined as the internode distance between two species (Liu & Yu 2011). Similarly, ASTRAL can also estimate the species tree from unrooted gene trees by minimizing the quartet distance between gene trees and the species tree (Mirarab *et al.* 2014). Unlike multispecies-coalescent, BCA does not make any biological assumptions about drivers of gene-tree heterogeneity (Ane *et al.* 2007). Thus, BCA is not strictly coalescent-based methods. We also employed BUCKy, a phylogenetic program for BCA that summarizes the proportion of sampled loci that support each clade by revising posterior distributions from every individual gene trees (Barrow *et al.* 2014; Liu *et al.* 2015). However, this method has not been widely used in deep-

level phylogeny because it requires that all taxa must be present in the gene tree for every locus (i.e. no missing data is permitted).

As input for these multispecies-coalescent approaches, individual gene trees from D98 and D150 datasets were estimated and nodal support was calculated with 100 fast bootstrap replicates using RAxML 8.0.23. We did not analyze dataset D289 with multispecies-coalescent approaches because of computational demands and preliminary analyses suggested similar results to those of analyses with D150. The best-fitting evolutionary model for each gene was evaluated in ProtTest and best-fit models were determined with Bayesian information criteria. STAR, MP-EST and NJst were conducted on the Species Tree Analysis Web server (STRAW; Shaw *et al.* 2013) with 100 multiloci bootstraps. A species tree was also estimated using ASTRAL with default parameters and 100 bootstrap replicates. OGs that included all taxa were used to estimate the primary concordance tree (34 OGs from D150 dataset, without *Spirobrachia*) using BUCKy 1.4.3. BUCKy required posterior distributions of individual gene trees, and these were estimated using MrBayes 3.2.2 (Ronquist & Huelsenbeck 2003). MrBayes analyses of the 34 OGs comprised two independent runs, with four coupled chains that were run for 2,000,000 generations. The first 10% of generations were discarded as burn-in based on trace plots. BUCKy 1.4.3 was run using four Markov Chain Monte Carlo chains for 1 million generations with four different priors ($\alpha = 0.1, 1, 10, 100$; $\alpha = 0$ indicates all gene trees possess the same topology; $\alpha = \infty$ indicates topology of each gene tree is completely incongruent), discarding the first 10% generations as burn-in.

3.3.4 Hypothesis testing

To assess the robustness of the inferred phylogenetic position of *Osedax*, an approximately unbiased (AU; Shimodaira 2002) test was used to determine if any *a priori* hypothesis of phylogenetic position *of Osedax* could be rejected (Fig. 1). Per site log-likelihoods for trees were calculated in RAxML and AU test were employed in CONSEL 0.20 (Shimodaira & Hasegawa 2001).

3.4 Results

3.4.1 Data matrix assembly

Initial orthology filtering of assembled transcriptomic data, followed by additional paralog screening and removal of genes that may cause systematic error using TreSpEx and BaCoCa (Fig. 6), resulted in 98 OGs for D98, 150 OGs for D150, and 289 OGs for D289. On average, 90.1% of orthologs were sampled per taxon in D98 dataset and the overall matrix completeness value, which considers alignment gaps as missing data, was 75.0%. For the D289 dataset, an of average 81.7% of orthologs were sampled per taxon, with an overall matrix completeness of 65.2%; for the D150 dataset, 91.0% of the orthologs were sampled per taxon, with an overall matrix completeness of 79.5% (Table 2).

3.4.2 Phylogenetic analysis using the supermatrix approach

Resulting tree topologies from all supermatrix analyses are shown in Fig. 3 (D289; 14 taxa), Fig. 4 (D98; 15 taxa), and Supplementary Fig. 7 (D150; 14 taxa). Supermatrix analysis with dataset D289 recovered an identical branching order to the tree inferred with dataset D150, but with slightly higher nodal support values (Figs. 3, S2). Both datasets recovered strong support for *Osedax* as closely related to Vestimentifera/*Sclerolinum* rather than Frenulata (bs=100; pp=1.00). Importantly, the hypothesis of *Osedax* as the sister group to Frenulata was explicitly rejected by AU tests on all three datasets (Table 3). The topology inferred from dataset D98 also supported *Osedax* as the sister group with the Vestimentifera/*Sclerolinum* clade (bs=100; pp=1.00). *Spirobrachia,* which had the most missing data (Table 2) and highest LB score compared to any other taxon, exhibited long branches in both analyses. In the BI tree inferred using CAT-GTR and dataset D98, *Spirobrachia* was placed unexpectedly as sister to all other Siboglinidae (Fig. 4B), whereas it was sister to the other frenulates in the ML analysis (Fig. 4A). *Spirobrachia* was not included in the other two datasets in order to accommodate datasets with less missing data and more loci. Both ML and BI recovered identical topologies in datasets D289 and D150, but variability among interrelationships within Vestimentifera and Frenulata were noted in D98 dataset (Figs. 3, 4, S2). For example, in datasets D289 and D150, *Ridgeia* was sister to *Seepiophila* and *Escarpia*, whereas it formed a clade with *Riftia* in the analysis of dataset D98. This result suggests that dataset size had more of an effect on relationships within Vestimentifera and Frenulata than differences between ML and BI with CAT-GTR.

3.4.3 Phylogenetic analysis using multispecies-coalescent approaches

Given that our supermatrix analyses showed congruent topologies between dataset D289 and D150 (Figs. 3, S2), multispecies-coalescent analyses were only performed on the smaller D150 and D98 datasets due to computational demands of some coalescent-based methods. In general, topologies derived from STAR, NJst, MP-EST and ASTRAL were largely in agreement with trees generated by the supermatrix approach, although some variations in branching patterns were observed (Figs. 5, S3, S4). For example, consistent with the analysis of D98, *Riftia* was placed as sister to *Ridgeia* in all multispecies-coalescent approaches (albeit with low nodal support values), but they were not sister taxa in the D289 and D150 supermatrix analyses. Importantly, all multispecies-coalescent analyses inferred *Osedax* as the sister group to Vestimentifera/*Sclerolinum* with 100% multiloci bootstrap support. For both the D150 and D98 datasets, species trees derived from STAR, NJst, and MP-EST exhibited the same tree topology (Figs. 5A, S3). *Siboglinum ekmani* was placed sister to other Frenulata from all D98 multispecies- coalescent analyses, whereas *S. fiordicum* was sister to other frenulates based on ASTRAL in both datasets (Figs. 5B, S4) and in supermatrix analyses. Similar to the BI analysis of D98 using the CAT-GTR model, *Spirobrachia* was placed sister to all other siboglinids in the multispecies-coalescent analyses.

3.4.4 Bayesian Concordance Analysis

The BCA tree (Fig. 10) derived from the reduced D150 dataset that only included the 34 OGs with every taxon present also exhibited a similar overall topology to other analyses in that a sister relationship between *Osedax* and Vestimentifera/*Sclerolinum* was recovered, albeit with moderate support (CF = 0.42; a CF = 0.5 indicates 50% of individual gene trees support this clade). Two differences were recovered between relationships estimated using BCA and the supermatrix approach. Within the Vestimentifera clade, placement of *Riftia* and *Ridgeia* were different compared to the supermatrix approach and multispecies coalescent, but these branches were weakly supported (CF = 0.27), and the lower CFs indicate the high level of gene tree discordance. Similar to topologies derived from multispecies-coalescent analyses, *S. ekmani* was placed sister to other frenulates (CF = 0.61), instead of *S. fiordicum* as inferred from supermatrix analyses.

3.5 Discussion

3.5.1 Siboglinid phylogeny

Different analyses have yielded conflicting hypotheses regarding the phylogenetic position of *Osedax* (Rouse *et al.* 2004, 2015; Glover *et al.* 2005, 2013; Li *et al.* 2015). Our results are consistent with previous molecular phylogenetic studies based on combinations of nuclear *18S* rDNA, mitochondrial *16S* rDNA, and *COI* (Rouse *et al.* 2004; Glover *et al.* 2005), indicating that *Osedax* is the sister group to the Vestimentiferan/*Sclerolinum* clade. A recent

mitogenomic analysis (Li *et al.* 2015) yielded the same topology as this study, but the nodal support for *Osedax* with the Vestimentifera/*Sclerolinum* clade was relatively low. Furthermore, Li *et al.* (2015) failed to reject the alternative placement of *Osedax* as the sister group to Frenulata with AU hypothesis tests. The lack of statistical support for the placement of *Osedax* in Li *et al.* (2015) and previous molecular studies with a limited number of loci (Rouse *et al*. 2004, 2015; Glover *et al*. 2005, 2013) could be explained as stochastic effects from a small number of loci (Delsuc *et al.* 2006) or saturation of the mitochondrial genes. Moreover, given that the entire siboglinid family can be traced back to a Late Mesozoic-Cenozoic origin (Little & Vrijenhoek 2003; Danise & Higgs 2015), utilizing only several mitochondrial loci and/or nuclear ribosomal loci may result in analyses with too little signal for resolving evolutionary relationships of major groups within siboglinids.

Both supermatrix and multispecies-coalescent analyses robustly supported placement of bone-eating *Osedax* as the sister group to a Vestimentifera plus *Sclerolinum* clade in all three datasets. More importantly, contrary to mitogenomic analyses (Li *et al.* 2015), our hypothesis testing strongly rejected the hypothesis of *Osedax* as the sister group to Frenulata (Table 3). Our results imply that bone-eating *Osedax*, the only lineage of siboglinids utilizing heterotrophic endosymbionts, is most likely derived from a lineage relying on chemoautotrophic bacteria that lived in deep-sea muddy sediments. Given that the association between non-*Osedax* siboglinids and chemoautotrophic bacteria is an obligate symbiosis, understanding the evolutionary

transition from a chemoautotrophic endosymbiont to a heterotrophic one in *Osedax* is of interest as the switch likely involved several changes in host physiology.

The monophyly of Frenulata was strongly supported in the supermatrix analyses of the D150 and D289 datasets (Figs. S2, 3), but not in dataset D98 because the tree inferred with CAT-GTR placed *Spirobrachia* sister to all other siboglinids. This placement of *Spirobrachia* was also recovered by all multispecies-coalescent approaches; *Spirobrachia* was not included in BCA because of a high level of missing data. As seen in previous analyses (Halanych *et al.* 2001; Li *et al* . 2015), our results also strongly supported *Lamellibrachia* sister to other vestimentiferans (Figs. 3-5, S2-S5). *Lamellibrachia* and *Escarpia* mainly inhabit seeps, whereas more derived vestimentiferans (e.g. *Riftia*, *Ridgeia*) live in association with vents, which is consistent with the hypothesis that habitat preferences of vestimentiferans have proceeded from less to more reducing sediments (Schulze & Halanych 2003).

3.5.2 Performance of supermatrix versus multispecies-coalescent approaches

Large phylogenomic datasets potentially contain genes with conflicting signal – for example due to incomplete lineage sorting, introgression, and paralogs – that can confound phylogenomic analyses (Smith *et al.* 2015). Additionally, given the recent debate between supermatrix and multispecies-coalescent approaches (Gatesy & Springer 2014; Edwards *et al.* 2015), we wished to explore the performance of these approaches on a phylogenomic dataset of manageable size.

In our analyses, relationships among the four major siboglinid lineages were largely consistent across approaches. Although variability among interrelationships within Vestimentifera and Frenulata were noted above, some of these conflicts were likely due to differences in dataset size. Notably, conflicting results were obtained from supermatrix and multispecies-coalescent methods with dataset D98. BI analysis using CAT models have been widely used for phylogenomic analyses because of its purported superiority in handling LBA (Delsuc *et al.* 2008; Philippe *et al.* 2009; Philippe *et al.* 2011). Yet *Spirobrachia* was unexpectedly placed sister to all other siboglinids in BI with dataset D98 (Fig. 4B), the same result as multispecies-coalescent based analyses (Figs. 5, S4) of the D98 dataset. In contrast, ML analyses of the D98 supermatrix supported a monophyletic Frenulata as previously reported in molecular and morphological studies (Rouse 2001; Li *et al.* 2015). ML analysis using data partitioning with site-homogeneous models is a common alternative approach to site-heterogeneous models for handling substitutional heterogeneity in large datasets (Lanfear *et al*. 2012). Several synapomorphies support the monophyly of frenulates including the presence of a cuticular and ventral ciliated band in the forepart region (Ivanov & Petrunkevitch 1955; Hilário *et al.* 2011). Given that sequences resulting from sample contamination (e.g. endosymbionts) have also likely been removed with TreSpEx, misplacement of *Spirobrachia* was most likely a result of this taxon having a large amount of missing data and consequently the highest LB score of any taxon rather than a paraphyletic group of frenulates. Thus, placement of *Spirobrachia*

sister to all other siboglinids seems unlikely (Rouse 2001; Halanych *et al.* 2001). As such, BI with CAT-GTR and multispecies-coalescence analyses are both potentially more susceptible to error when at least one taxon has large amounts of missing data compared to ML with data partitioning and site-homogeneous models. This conclusion implies that BI with the CAT-GTR model, as well as multispecies-coalescent analyses, likely produced trees not representative of siboglinid phylogeny.

BCA is similar to multispecies-coalescent approaches in that it does not assume loci share the same underlying topology, but unlike other methods, it reports proportions of genes supporting inferred relationships. However, BUCKy requires that all taxa must be present in the posterior distribution of trees for every locus. In this transcriptome-based study, only 34 OGs had full taxon representation and could be used to estimate the primary concordance tree (Fig. 10). Although topologies derived from both methods were largely congruent, variation occurred in the placement of *Riftia* and *Ridgeia*, a node with low concordance (CF = 0.266). Given that performance of most phylogenetic methods can be dramatically improved by increasing the number of genes (Liu *et al.* 2015), this conflict should not be surprising, especially as only a small number of OGs could were suitable for analysis with BUCKy.

In conclusion, the three contrasting phylogenetic approaches used in this study produced largely congruent results, especially for datasets D289 and D150. In contrast to previous studies, we failed to recover an *Osedax*/Frenulata sister relationship with any datasets across analytical methods. Explicit hypothesis testing with AU tests also significantly rejected *Osedax* as the sister

group to Frenulata. Moreover, a significant discrepancy was found in dataset D98 in terms of the placement of *Spriobrachia*. Given that placement of *Spirobrachia* sister to all other siboglinids is not consistent with other sources of data (Rouse 2001; Li *et al.* 2015) and that *Spirobrachia* shares putative morphological synapomorphies with Frenulata, the supermatrix approach with ML using data partitioning with site-homogenous models appears to have outperformed both the supermatrix method with CAT-GTR and multispecies-coalescent approaches. In particular, methods that recovered *Spirobrachia* sister to all other siboglinids appears to be susceptible to error associated with missing data. The well-supported phylogenetic hypotheses generated here should serve as a foundation for future studies on siboglinid evolution including the evolution of different obligate symbioses, adaptation, and colonization to different reducing habitats.

3.6 Acknowledgement

This study was supported by awards from the U.S. National Science Foundation (NSF) (DEB-1036537 to KMH and SRS; IOS-0843473 to KMH, SRS and DJT; and DBI-1306538 to KMK). Yuanning Li is supported by a scholarship from the China Scholarship Council (CSC) for studying and living abroad. All phylogenetic analyses were conducted on the Auburn University Molette Laboratory SkyNet server and Auburn University CASIC HPC system. This is Molette Biology Laboratory contribution #51 and Auburn University Marine Biology Program contribution #143.

3.7 References

Abascal, F., Zardoya, R., & Posada, D. (2005). Prottest: Selection of best-fit models of protein evolution. *Bioinformatics*, 21, 2104-2105.10.1093/bioinformatics/bti263

Ane, C., Larget, B., Baum, D. A., Smith, S. D., & Rokas, A. (2007). Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution*, 24, 412-426.Doi 10.1093/molbev/msl170

Barrow, L. N., Ralicki, H. F., Emme, S. A., & Lemmon, E. M. (2014). Species tree estimation of North American chorus frogs (Hylidae: *Pseudacris*) with parallel tagged amplicon sequencing. *Molecular Phylogenetics and Evolution*, 75, 78-90.Doi 10.1016/j.ympev.2014.02.007

Black, M. B., Halanych, K. M., Maas, P. A. Y., Hoeh, W. R., Hashimoto, J., Desbruyeres, D., et al. (1997). Molecular systematics of vestimentiferan tubeworms from hydrothermal vents and cold-water seeps. *Marine Biology*, 130, 141-149.Doi 10.1007/S002270050233

Bond, J. E., Garrison, N. L., Hamilton, C. A., Godwin, R. L., Hedin, M., & Agnarsson, I. (2014). Phylogenomics resolves a spider backbone phylogeny and rejects a prevailing paradigm for orb web evolution. *Current Biology*, 24, 1765-1771.10.1016/j.cub.2014.06.034

Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B., & Brom, T. H. (2012). A reference-free algorithm for computational normalization of shotgun sequencing data. *arXiv:1203.4802 [q-bio]*

Chao, L. S. L., Davis, R. E., & Moyer, C. L. (2007). Characterization of bacterial community structure in vestimentiferan tubeworm *Ridgeia piscesae* trophosomes. *Marine Ecology and Evolutionary Perspective*, 28, 72-85.Doi 10.1111/j.1439-0485.2007.00151.x

Danise, S., & Higgs, N. D. (2015). Bone-eating *Osedax* worms lived on mesozoic marine reptile deadfalls. *Biology Letter*, 11, 20150072.Doi 10.1098/rsbl.2015.0072

Delsuc, F., Brinkmann, H., Chourrout, D., & Philippe, H. (2006). Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature*, 439, 965-968.Doi 10.1038/nature04336

Delsuc, F., Brinkmann, H., & Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, 6, 361-375.Doi 10.1038/nrg1603

Delsuc, F., Tsagkogeorga, G., Lartillot, N., & Philippe, H. (2008). Additional molecular support for the new chordate phylogeny. *Genesis (New York, N.Y.: 2000)*, 46, 592-604.Doi 10.1002/dvg.20450

Dunn, C. W., Hejnol, A., Matus, D. Q., Pang, K., Browne, W. E., Smith, S. A., et al. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, 452, 745-749.Doi 10.1038/nature06614

Ebersberger, I., Strauss, S., & von Haeseler, A. (2009). Hamstr: Profile hidden markov model based search for orthologs in ests. *BMC Evolutioanry Biology*, 9, 157.Doi 10.1186/1471-2148-9-157

Edwards, S. V., Xi, Z., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., et al. (2015). Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Molecular Phylogenetics and Evolution*. Doi 10.1016/j.ympev.2015.10.027

Funk, D. J., & Omland, K. E. (2003). Species-level paraphyly and polyphyly: Frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annual Review of Ecology Evolution and Systematics*, 34, 397-423.Doi 10.1146/annurev.ecolsys.34.011802.132421

Gatesy, J., & Springer, M. S. (2013). Concatenation versus coalescence versus "concatalescence". *Proceedings of the National Academy of Sciences of the United States of America*, 110, E1179.Doi 10.1073/pnas.1221121110

Gatesy, J., & Springer, M. S. (2014). Phylogenetic analysis at deep timescales: Unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Molecular Phylogenetics and Evolution*, 80, 231-266.Doi 10.1016/j.ympev.2014.08.013

Glover, A. G., Kallstrom, B., Smith, C. R., & Dahlgren, T. G. (2005). World-wide whale worms? A new species of *Osedax* from the shallow north atlantic. *Proceedings of the Royal Society B: Biological Sciences,* 272, 2587-2592.Doi 10.1098/rspb.2005.3275

Glover, A. G., Wiklund, H., Taboada, S., Avila, C., Cristobo, J., Smith, C. R., et al. (2013). Bone-eating worms from the Antarctic: The contrasting fate of whale and wood remains

on the southern ocean seafloor. *Proceedings of the Royal Society B: Biological Sciences*, 280, 20131390.Doi 10.1098/rspb.2013.1390

Goffredi, S. K., Orphan, V. J., Rouse, G. W., Jahnke, L., Embaye, T., Turk, K., et al. (2005). Evolutionary innovation: A bone-eating marine symbiosis. *Environmental Microbiology*, 7, 1369-1378.Doi 10.1111/j.1462-2920.2005.00824.x

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from rna-seq data without a reference genome. *Nature Biotechnology*, 29, 644-652.Doi 10.1038/nbt.1883

Halanych, K. M. (2005). Molecular phylogeny of siboglinid annelids (a.K.A. Pogonophorans): A review. *Hydrobiologia*, 535, 297-307.Doi 10.1007/S10750-004-1437-6

Halanych, K. M., Dahlgren, T. G., & McHugh, D. (2002). Unsegmented annelids? Possible origins of four lophotrochozoan worm taxa. *Integrative and Comparative Biology*, 42, 678-684.Doi 10.1093/icb/42.3.678

Halanych, K. M., Feldman, R. A., & Vrijenhoek, R. C. (2001). Molecular evidence that *Sclerolinum brattstromi* is closely related to vestimentiferans, not to frenulate pogonophorans (siboglinidae, annelida). *Biological Bulletin*, 201, 65-75

Halanych, K. M., Lutz, R. A., & Vrijenhoek, R. C. (1998). Evolutionary origins and age of vestimentiferan tube-worms. *Cahiers de Biologie Marine*, 39, 355-358

Heled, J., & Drummond, A. J. (2010). Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, 27, 570-580.Doi 10.1093/molbev/msp274

Hilário, A., Capa, M., Dahlgren, T. G., Halanych, K. M., Little, C. T., Thornhill, D. J., et al. (2011). New perspectives on the ecology and evolution of siboglinid tubeworms. *PLoS One*, 6, e16309.Doi 10.1371/journal.pone.0016309

Huelsenbeck, J., & Rannala, B. (2004). Frequentist properties of bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Systematic Biology*, 53, 904-913.Doi 10.1080/10635150490522629

Ivanov, A. V. (1963). *Pogonophora*: Academic Press.

Ivanov, A. V., & Petrunkevitch, A. (1955). On external digestion in pogonophora. *Systematic Zoology*, 4, 174.Doi 10.2307/2411670

Jones, M. L. (1988). The vestimentifera, their biology, systematic and evolutionary patterns. *Oceanologica Acta, Special issue*

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). Mafft: A novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research*, 30, 3059-3066

Kocot, K. M., Cannon, J. T., Todt, C., Citarella, M. R., Kohn, A. B., Meyer, A., et al. (2011). Phylogenomics reveals deep molluscan relationships. *Nature*, 477, 452-456.Doi 10.1038/nature10382

Kocot, K. M., Citarella, M. R., Moroz, L. L., & Halanych, K. M. (2013). Phylotreepruner: A phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. *Evolutionary Bioinformatics Online*, 9, 429-435.Doi 10.4137/EBO.S12813

Kocot, K. M., Struck, T. H., Merkel, J., Waits, D. S., Todt, C., Brannock, P. M., Weese D. A., et al. (2016). Phylogenomics of Lophotrochozoa with consideration of systematic error. *Systematic Biology.*

Kück, P. (2009). Alicut: A perlscript which cuts aliscore identified rss. *Department of Bioinformatics, Zoologisches Forschungsmuseum A. Koenig (ZFMK), Bonn, Germany, version*, 2

Kuck, P., & Struck, T. H. (2014). BaCoCa--a heuristic software tool for the parallel assessment of sequence biases in hundreds of gene and taxon partitions. *Molecular Phylogenetics and Evolution*, 70, 94-98.Doi 10.1016/j.ympev.2013.09.011

Lane, C. E. (2007). Bacterial endosymbionts: Genome reduction in a hot spot. *Current Biology*, 17, R508-510.Doi 10.1016/j.cub.2007.04.035

Lanfear, R., Calcott, B., Ho, S. Y., & Guindon, S. (2012). Partitionfinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evoltion*, 29, 1695-1701.Doi 10.1093/molbev/mss020

Lanfear, R., Calcott, B., Kainer, D., Mayer, C., & Stamatakis, A. (2014). Selecting optimal partitioning schemes for phylogenomic datasets. *BMC Evolutionary Biology*, 14, 82.Doi 10.1186/1471-2148-14-82

Lanier, H. C., & Knowles, L. L. (2012). Is recombination a problem for species-tree analyses? *Systematic Biology*, 61, 691-701.Doi 10.1093/sysbio/syr128

Larget, B. R., Kotha, S. K., Dewey, C. N., & Ane, C. (2010). BUCKy: Gene tree/species tree reconciliation with bayesian concordance analysis. *Bioinformatics*, 26, 2910-2911.Doi 10.1093/bioinformatics/btq539

Lartillot, N., Lepage, T., & Blanquart, S. (2009). PhyloBayes 3: A bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, 25, 2286-2288.Doi 10.1093/bioinformatics/btp368

Lartillot, N., & Philippe, H. (2004). A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21, 1095-1109.Doi 10.1093/molbev/msh112

Lemmon, A. R., Brown, J. M., Stanger-Hall, K., & Lemmon, E. M. (2009). The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and bayesian inference. *Systematic Biology*, 58, 130-145.10.1093/sysbio/syp017

Li, Y., Kocot, K. M., Schander, C., Santos, S. R., Thornhill, D. J., & Halanych, K. M. (2015). Mitogenomics reveals phylogeny and repeated motifs in control regions of the deep-sea

family Siboglinidae (Annelida). *Molecular Phylogenetics and Evolution*, 85, 221-229.Doi 10.1016/j.ympev.2015.02.008

Little, C. T. S., & Vrijenhoek, R. C. (2003). Are hydrothermal vent animals living fossils? *Trends in Ecology & Evolution*, 18, 582-588.Doi 10.1016/j.tree.2003.08.009

Liu, L., Wu, S. Y., & Yu, L. L. (2015). Coalescent methods for estimating species trees from phylogenomic data. *Journal of Systematics and Evolution*, 53, 380-390.Doi 10.1111/jse.12160

Liu, L., & Yu, L. (2011). Estimating species trees from unrooted gene trees. *Systematic Biology*, 60, 661-667.Doi 10.1093/sysbio/syr027

Liu, L., Yu, L., & Edwards, S. V. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, 10, 302.Doi 10.1186/1471-2148-10-302

Liu, L., Yu, L., Pearl, D. K., & Edwards, S. V. (2009). Estimating species phylogenies using coalescence times among sequences. *Systematic Biology*, 58, 468-477.Doi 10.1093/sysbio/syp031

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437, 376-380.Doi 10.1038/nature03959

Matus, D. Q., Copley, R. R., Dunn, C. W., Hejnol, A., Eccleston, H., Halanych, K. M., et al. (2006). Broad taxon and gene sampling indicate that chaetognaths are protostomes. *Current Biology*, 16, R575-576.Doi 10.1016/j.cub.2006.07.017

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., & Warnow, T. (2014). Astral: Genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30, i541-548.Doi 10.1093/bioinformatics/btu462

Misof, B., Liu, S., Meusemann, K., Peters, R. S., Donath, A., Mayer, C., et al. (2014). Phylogenomics resolves the timing and pattern of insect evolution. *Science*, 346, 763-767.Doi 10.1126/science.1257570

Misof, B., & Misof, K. (2009). A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: A more objective means of data exclusion. *Systematic Biology*, 58, 21-34.Doi 10.1093/sysbio/syp006

Nussbaumer, A. D., Fisher, C. R., & Bright, M. (2006). Horizontal endosymbiont transmission in hydrothermal vent tubeworms. *Nature*, 441, 345-348.Doi 10.1038/nature04793

Oliver, J. C. (2013). Microevolutionary processes generate phylogenomic discordance at ancient divergences. *Evolution*, 67, 1823-1830.Doi 10.1111/evo.12047

Philippe, H., Brinkmann, H., Copley, R. R., Moroz, L. L., Nakano, H., Poustka, A. J., et al. (2011). Acoelomorph flatworms are deuterostomes related to *Xenoturbella*. *Nature*, 470, 255-258.Doi 10.1038/nature09676

Philippe, H., Derelle, R., Lopez, P., Pick, K., Borchiellini, C., Boury-Esnault, N., et al. (2009). Phylogenomics revives traditional views on deep animal relationships. *Current Biology*, 19, 706-712.Doi 10.1016/j.cub.2009.02.052

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One*, 5, e9490.Doi 10.1371/journal.pone.0009490

R Development Core Team. (2015). R: A language and environment for statistical computing. Vienna, Austria; 2014. *URL* http://www. *R-project. Org.*

Rambaut, A., Suchard M, A., Xie, D., & Drummond A, J. (2014). Tracer v1.6.

Rannala, B., & Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164(4), 1645-1656.

Ronquist, F., & Huelsenbeck, J. P. (2003). Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19, 1572-1574

Rouse, G. W. (2001). A cladistic analysis of Siboglinidae caullery, 1914 (Polychaeta, Annelida): Formerly the phyla Pogonophora and Vestimentifera. *Zoological Journal of the Linnean Society*, 132, 55-80.Doi 10.1111/J.1096-3642.2001.Tb02271.X

Rouse, G. W., Goffredi, S. K., & Vrijenhoek, R. C. (2004). *Osedax*: Bone-eating marine worms with dwarf males. *Science*, 305, 668-671.Doi 10.1126/science.1098650

Rouse, G. W., Wilson, N. G., Worsaae, K., & Vrijenhoek, R. C. (2015). A dwarf male reversal in bone-eating worms. *Current Biology*, 25, 236-241.Doi 10.1016/j.cub.2014.11.032

Rousset, V., Rouse, G. W., Siddall, M. E., Tillier, A., & Pleijel, F. (2004). The phylogenetic position of Siboglinidae (Annelida) inferred from 18s rRNA, 28s rRNA and morphological data. *Cladistics*, 20, 518-533.Doi 10.1111/J.1096-0031.2004.00039.X

Schulze, A. (2003). Phylogeny of vestimentifera (siboglinidae, annelida) inferred from morphology. *Zoologica Scripta*, 32, 321-342.Doi 10.1046/J.1463-6409.2003.00119.X

Schulze, A., & Halanych, K. M. (2003). Siboglinid evolution shaped by habitat preference and sulfide tolerance. *Hydrobiologia*, 496, 199-205.Doi 10.1023/A:1026192715095

Shaw, T. I., Ruan, Z., Glenn, T. C., & Liu, L. (2013). Straw: Species tree analysis web server. *Nucleic Acids Research*, 41, W238-241.Doi 10.1093/nar/gkt377

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, 51, 492-508.Doi 10.1080/10635150290069913

Shimodaira, H., & Hasegawa, M. (2001). Consel: For assessing the confidence of phylogenetic tree selection. *Bioinformatics*, 17, 1246-1247.10.1093/bioinformatics/17.12.1246

Smith, S. A., Moore, M. J., Brown, J. W., & Yang, Y. (2015). Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology*, 15, 150. Doi 10.1186/s12862-015-0423-0

Southward, E. C. (1982). Bacterial symbionts in pogonophora. *Journal of the Marine Biological Association of the United Kingdom*, 62, 889-906.Doi 10.1017/S0025315400070417

Southward, E. C., Schulze, A., & Gardiner, S. L. 2005. Pogonophora (Annelida): Form and function. In T. Bartolomaeus & G. Purschke (Eds) *Morphology, molecules, evolution and phylogeny in polychaeta and related taxa* pp. 227-251): Springer Netherlands.

Springer, M. S., & Gatesy, J. (2015). The gene tree delusion. *Molecular Phylogenetics and Evolution*, 94, 1-33.Doi 10.1016/j.ympev.2015.07.018

Stamatakis, A. (2014). Raxml version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30, 1312-1313.Doi 10.1093/bioinformatics/btu033

Struck, T. H. (2014). TreSpEx-detection of misleading signal in phylogenetic reconstructions based on tree information. *Evolution Bioinformatics Online*, 10, 51-67.10.4137/EBO.S14239

Struck, T. H., Paul, C., Hill, N., Hartmann, S., Hösel, C., Kube, M., Lieb, B., Meyer, A., Tiedemann, R., Purschke, G. & Bleidorn, C. (2011). Phylogenomic analyses unravel annelid evolution. *Nature,* 471(7336), 95-98.

Thornhill, D. J., Wiley, A. A., Campbell, A. L., Bartol, F. F., Teske, A., & Halanych, K. M. (2008). Endosymbionts of siboglinum fiordicum and the phylogeny of bacterial endosymbionts in Siboglinidae (Annelida). *Biological Bulletin*, 214, 135-144

Vrijenhoek, R. C., Duhaime, M., & Jones, W. J. (2007). Subtype variation among bacterial endosymbionts of tubeworms (Annelida : Siboglinidae) from the Gulf of California. *Biological Bulletin*, 212, 180-184

Weigert, A., Helm, C., Meyer, M., Nickel, B., Arendt, D., Hausdorf, B., et al. (2014). Illuminating the base of the annelid tree using transcriptomics. *Molecular Biology and Evolution*, 31, 1391-1401.Doi 10.1093/molbev/msu080

Whelan, N. V., Kocot, K. M., Moroz, L. L., & Halanych, K. M. (2015). Error, signal, and the placement of Ctenophora sister to all other animals. *Proceedings of the National Academy of Sciences of the United States of America*, 112, 5773-5778.Doi 10.1073/pnas.1503453112

Wu, S. Y., Song, S., Liu, L., & Edwards, S. V. (2013). Reply to Gatesy and Springer: The multispecies coalescent model can effectively handle recombination and gene tree heterogeneity. *Proceedings of the National Academy of Sciences of the United States of America*, 110, E1180-E1180.Doi 10.1073/pnas.1300129110

Zhong, B., Liu, L., & Penny, D. (2014). The multispecies coalescent model and land plant origins: A reply to Springer and Gatesy. *Trends Plant Science*, 19, 270-272.Doi 10.1016/j.tplants.2014.02.011

Zhong, B., Liu, L., Yan, Z., & Penny, D. (2013). Origin of land plants using the multispecies coalescent model. *Trends Plant Science*, 18, 492-495.Doi 10.1016/j.tplants.2013.04.009

Zhong, M., Hansen, B., Nesnidal, M., Golombek, A., Halanych, K. M., & Struck, T. H. (2011). Detecting the symplesiomorphy trap: A multigene phylogenetic analysis of terebelliform annelids. *BMC Evolutioanry Biology*, 11, 369.Doi 10.1186/1471-2148-11-369

Fig. 1

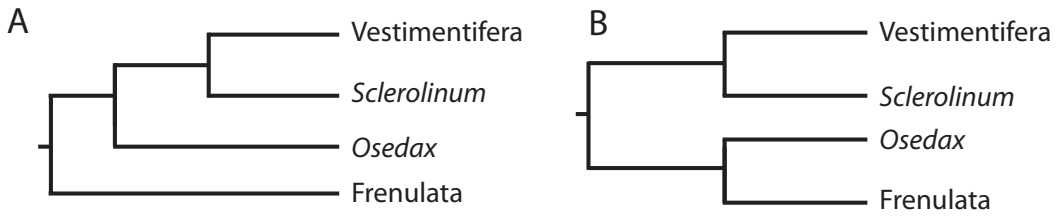
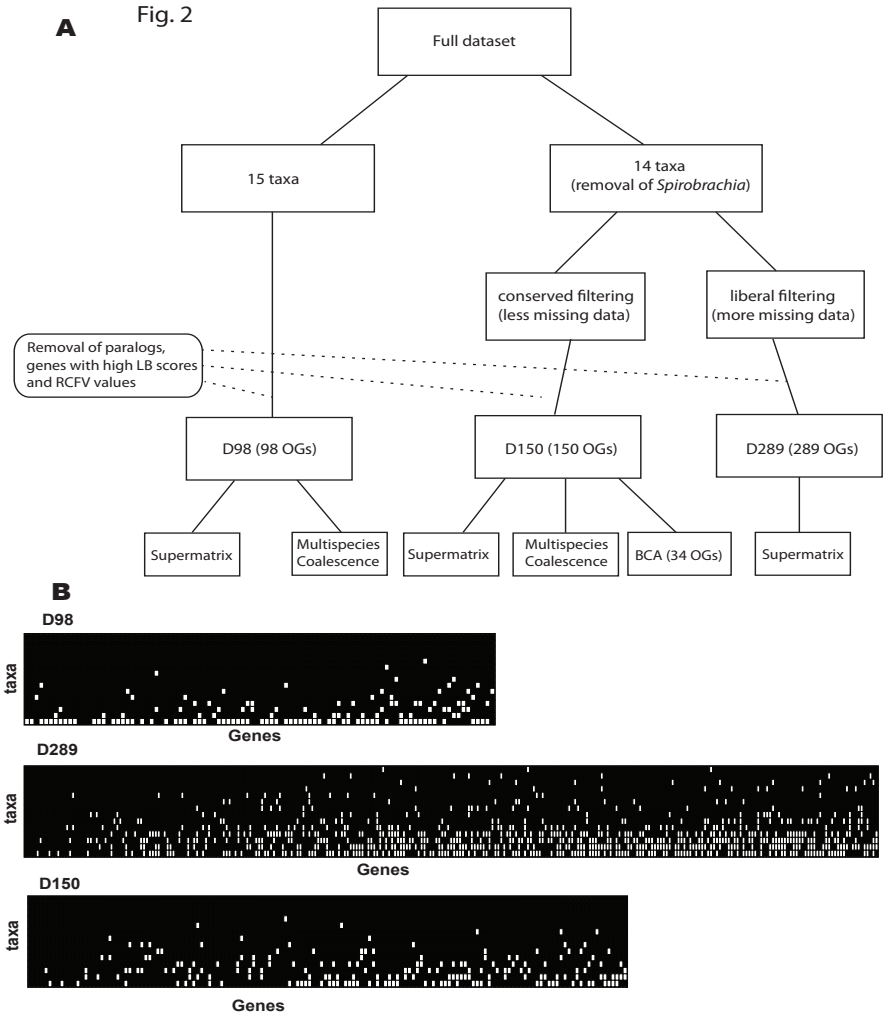


Fig. 1. Phylogenetic hypotheses from previous molecular studies. (A) Hypothesis of *Osedax* as the sister group to Vestimentifera and *Sclerolinum* clade (Rouse *et al.* 2004; Glover *et al.* 2005; Li *et al*. 2015). (B) Hypothesis of *Osedax* close related to Frenulata (Glover *et al*. 2011; Rouse *et al*. 2015).

Fig. 2



c for the D98, D150 and D298 datasets. Data statistics for each dataset is shown in Table. 2. (B) Occupancy of orthologous groups in data matrices for phylogenetic analyses. Genes are ordered along the X-axis and taxa are ordered along the Y-axis. For any given gene fragment, black squares represent sampled sequence data, and white squares represent missing data.

Fig. 3

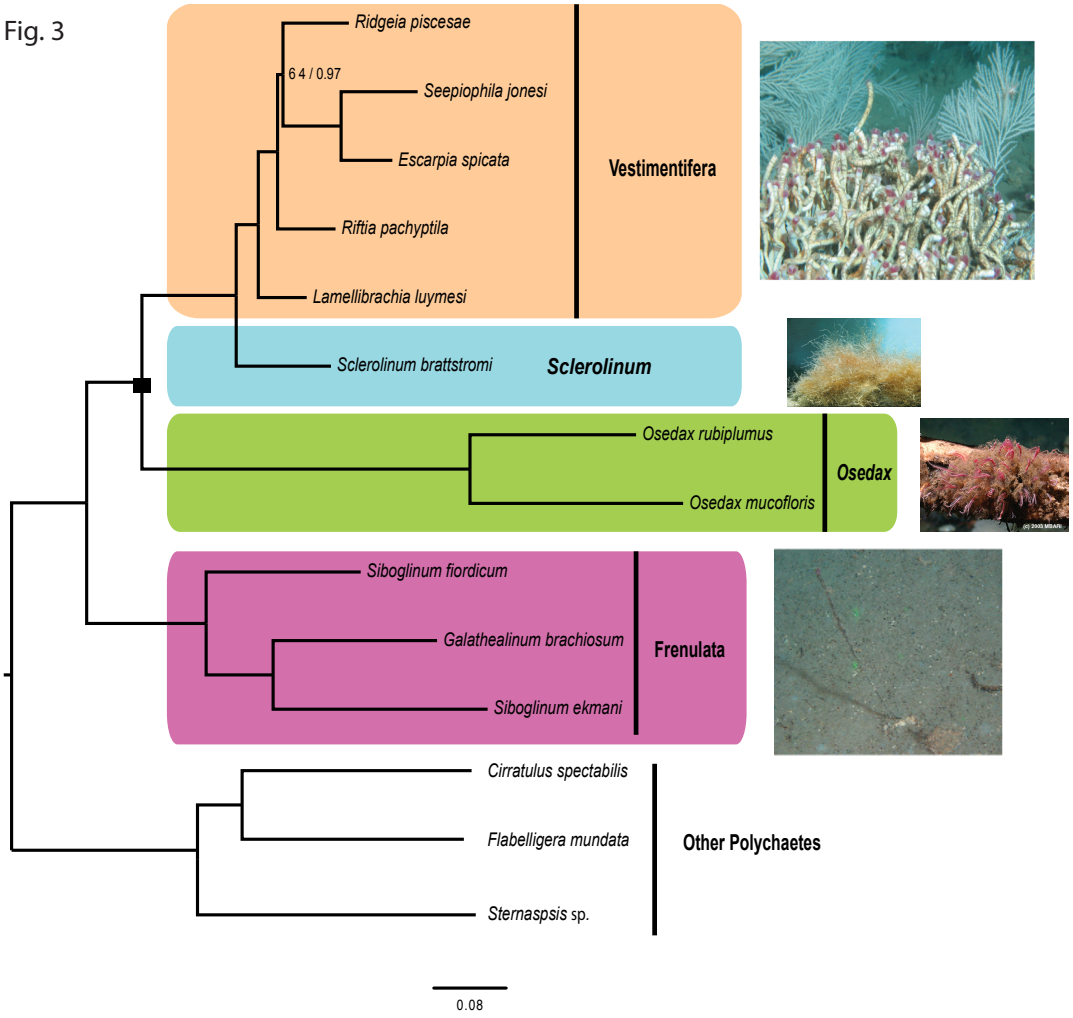


Fig. 3. Phylogenetic reconstructions of Siboglinidae based on dataset D289 using supermatrix approach and a Bayesian inference approach with a CAT-GTR model. Majority rule consensus phylogram is shown. Values shown next to nodes are posterior probabilities on the left and ML bootstrap support values on the right. The black square indicates the node joining the *Osedax* lineage to the vestimentiferan/*Sclerolinum* clade. All nodes were supported with 100% bootstrap value or posterior probabilities of 1.0 unless otherwise noted.

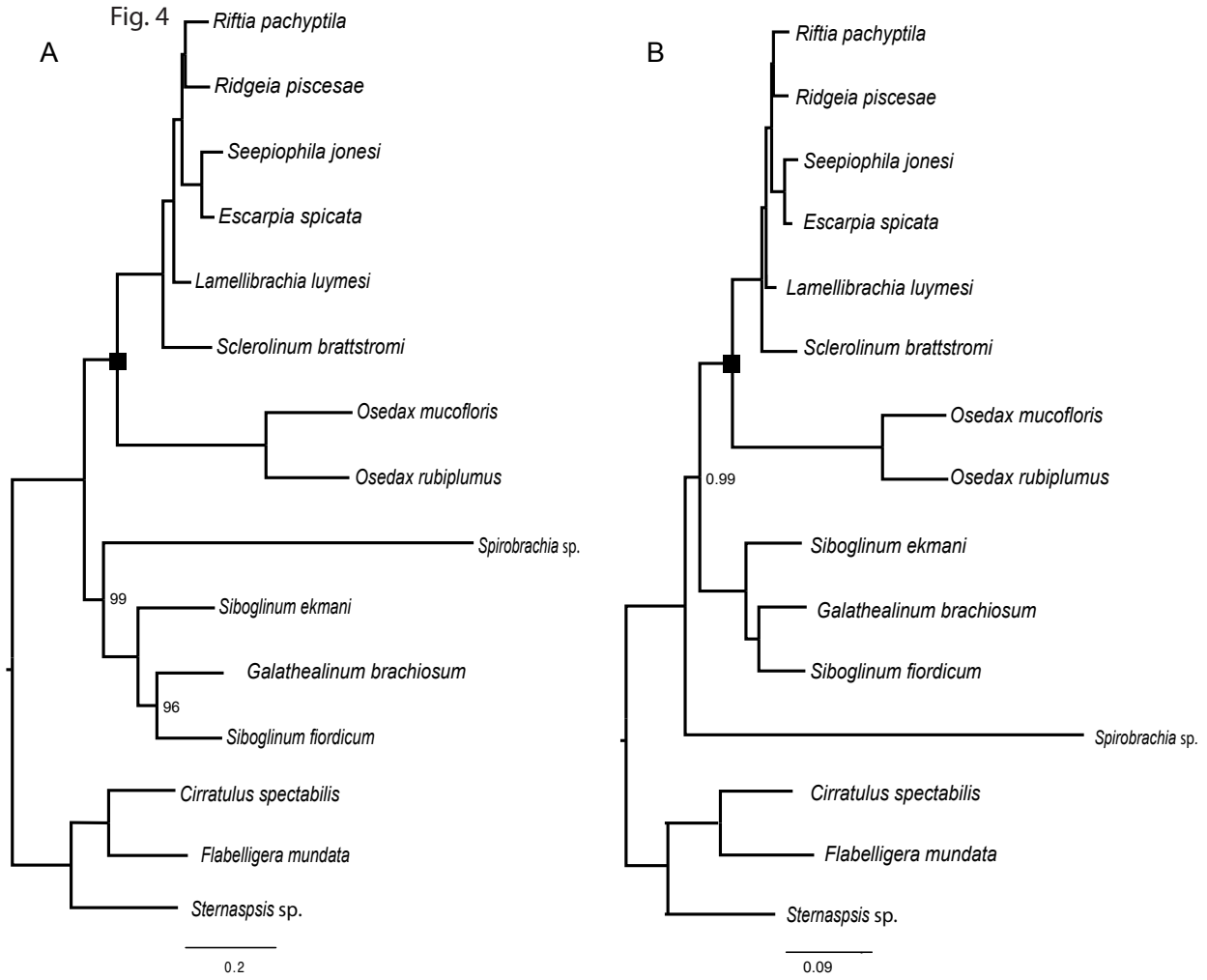


Fig. 4. Phylogenetic reconstructions of Siboglinidae from dataset D98. Topologies derived from ML (A) with bootstrap support and BI using CAT+GTR model (B) with posterior probabilities. The black square indicates the node joining the *Osedax* lineage to the vestimentiferan/*Sclerolinum* clade. All nodes were supported with 100% bootstrap value or posterior probabilities of 1.0 unless otherwise noted.

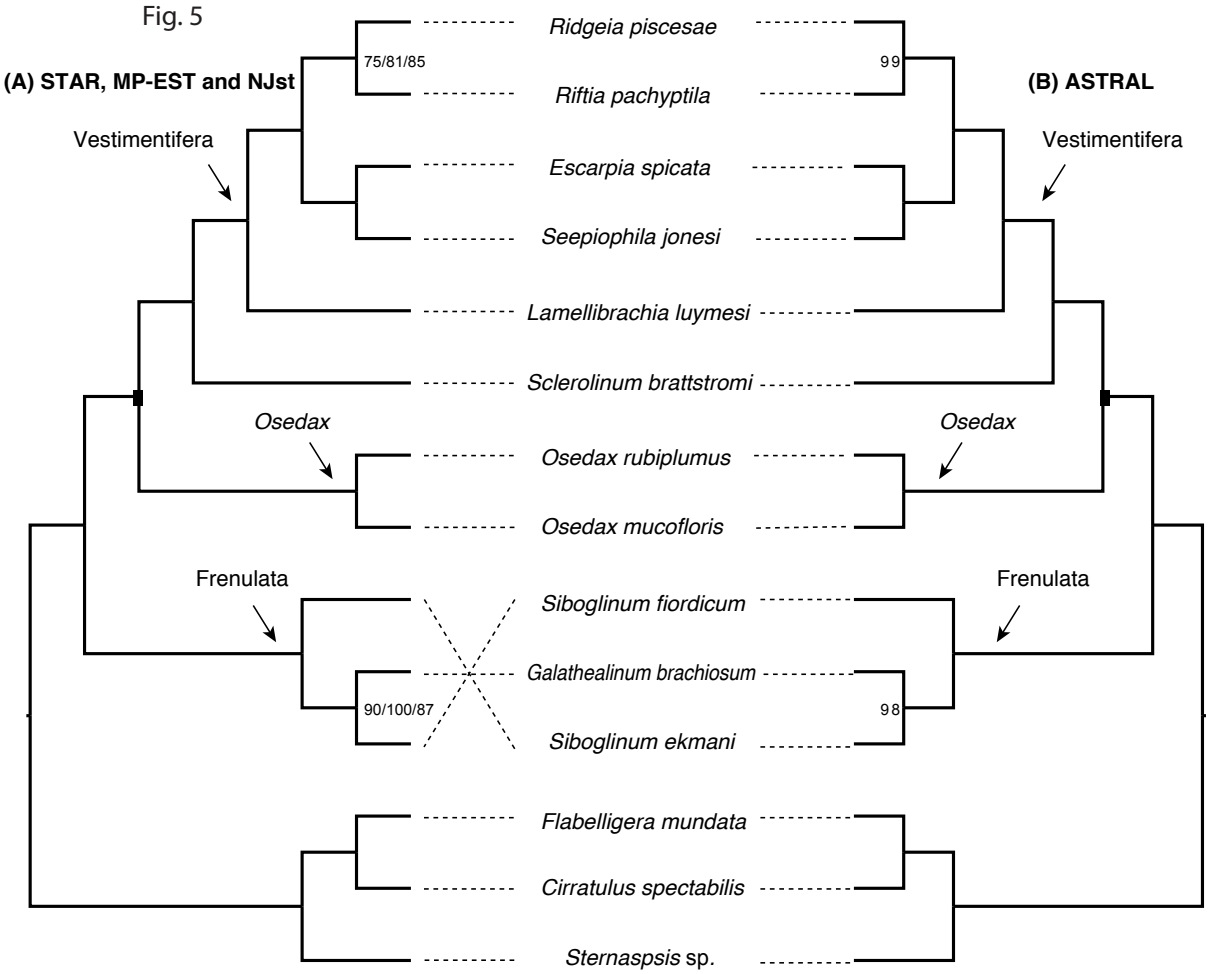


Fig. 5. Species trees inferred from (A) based on STAR, MP-EST, NJst and (B) ASTRAL from dataset D150. The black square indicates the node joining the *Osedax* lineage to the vestimentiferan/*Sclerolinum* clade. Nodal support values (A) left: STAR; middle: MP-EST; right: NJst (B) ASTRAL indicate bootstrap proportion based upon 100 multilocus bootstraps

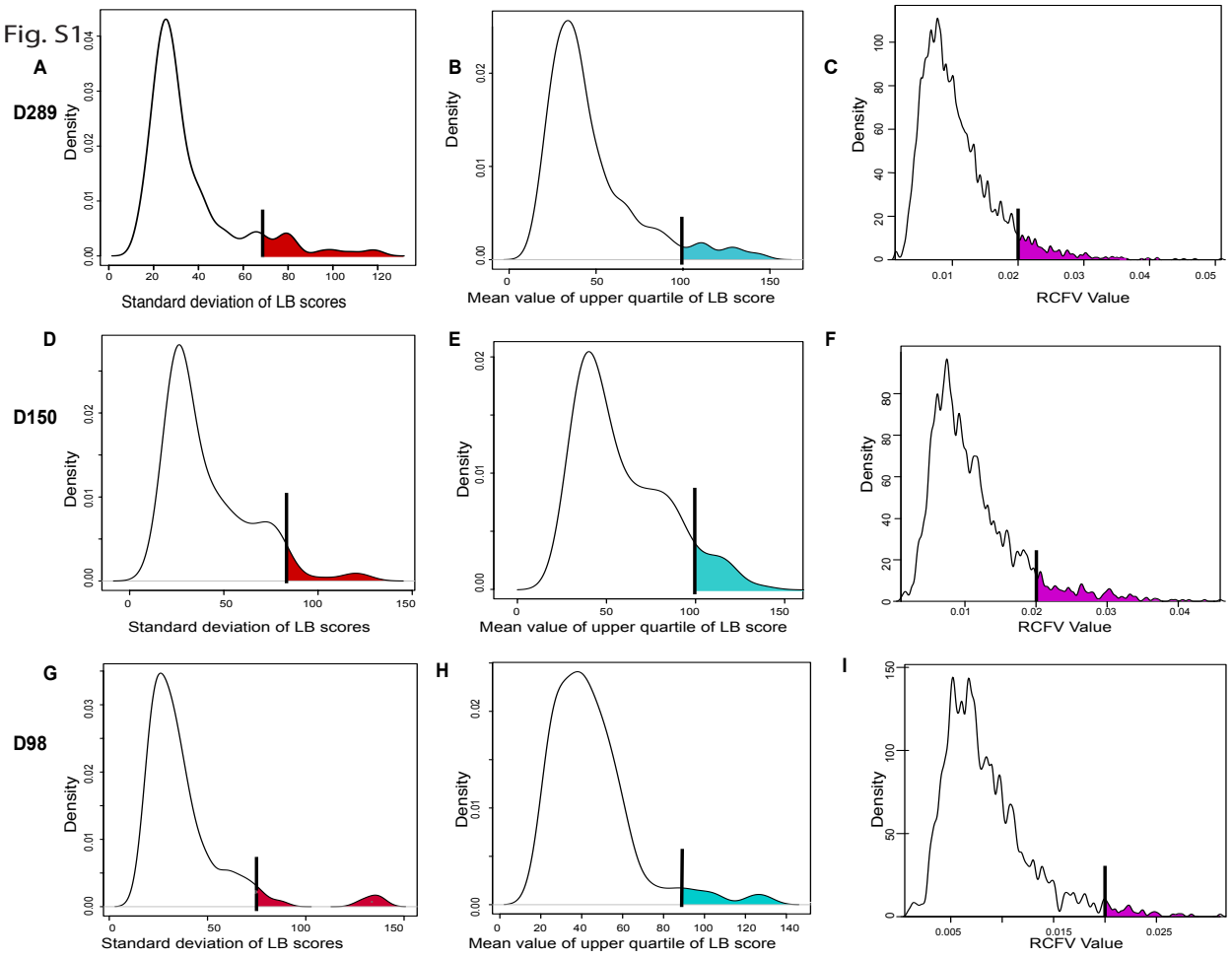


Fig. 6. Density plots of (A, D, G) standard deviation of LB scores for OGs, (B, E, H) average upper quartile LB score for each OG, and (C, F, I) RCFV values for each OG from D289, D150 and D98 datasets, respectively. Shaded areas include taxa or genes that were considered to have "high" LB scores or RCFV values and were trimmed for subsequent analysis.

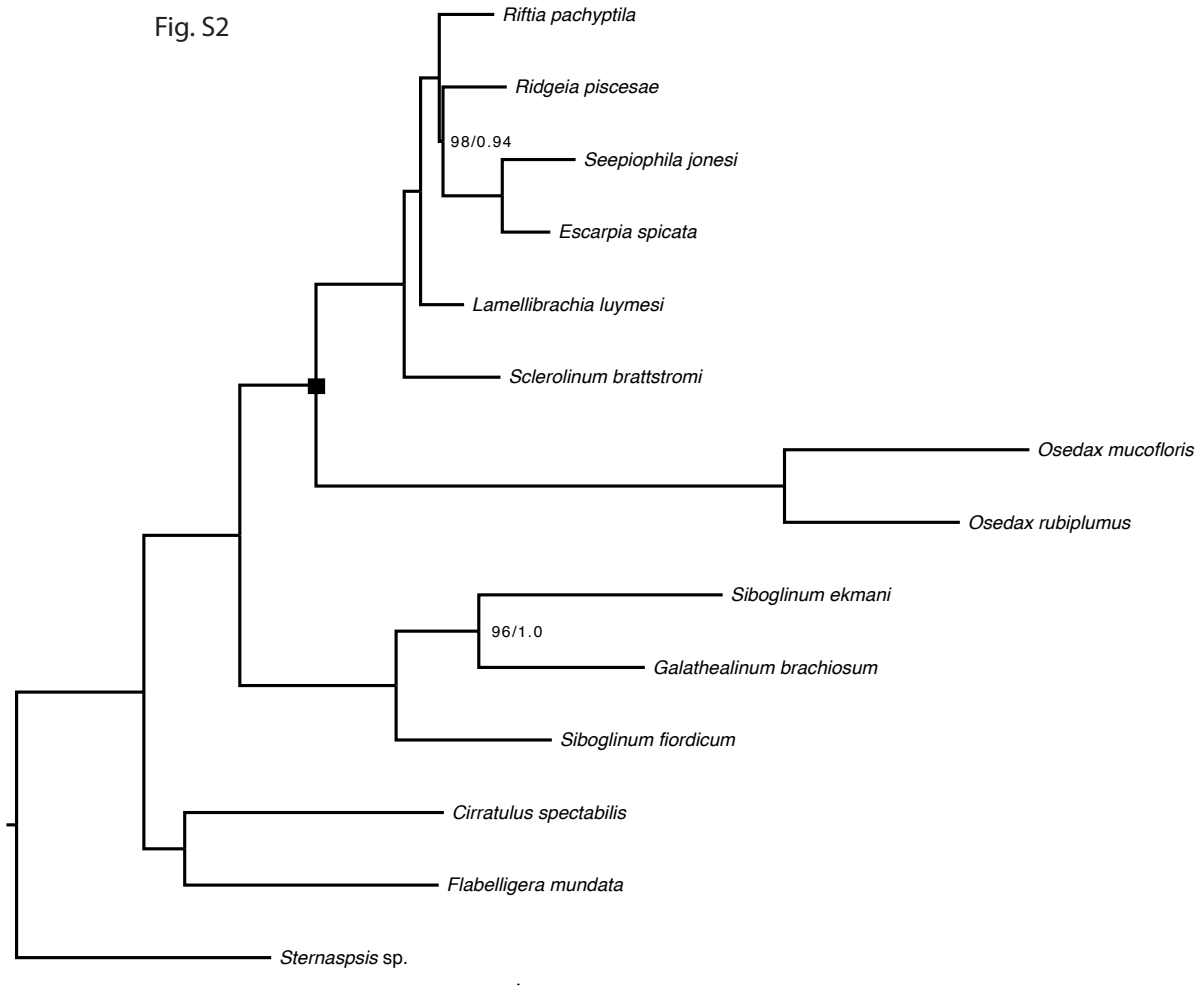


Fig. 7. Phylogenetic reconstructions of Siboglinidae inferred from D150 dataset. BI majority rule (50%) consensus phylogram (using the CAT+GTR model) of the concatenated data matrix is shown. Values are shown next to nodes with posterior probabilities to the left and ML bootstrap support values to the right. All nodes were supported with 100% bootstrap value or posterior probabilities of 1.0 unless otherwise noted

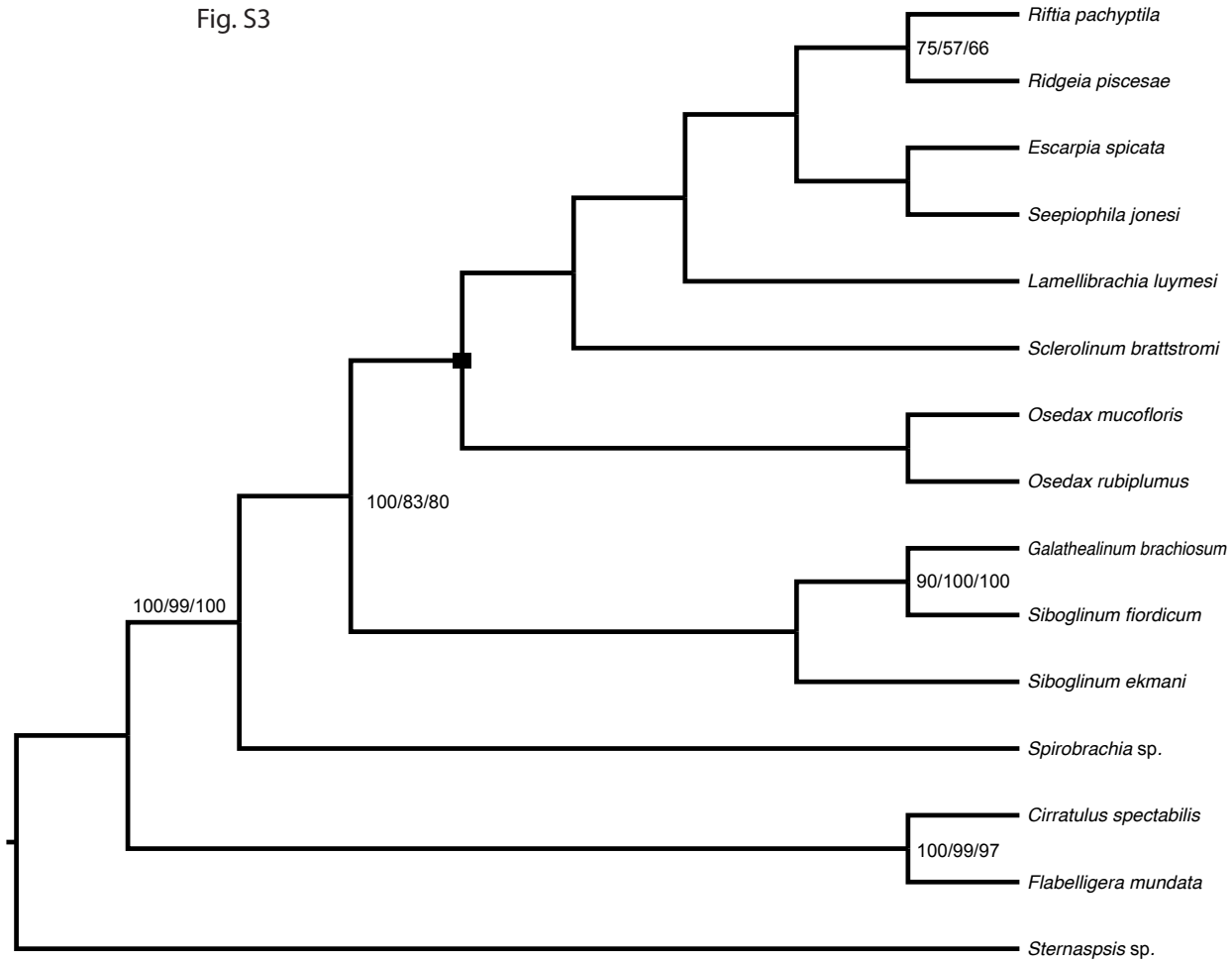


Fig. 8. Species tree inferred from D98 based on STAR, MP-EST and NJst. The black square indicates the node joining the *Osedax* lineage to the vestimentiferan/*Sclerolinum* clade. Nodal support values (left: STAR; middle: MP-EST; right: NJst) indicate bootstrap proportion based upon 100 multilocus bootstraps.

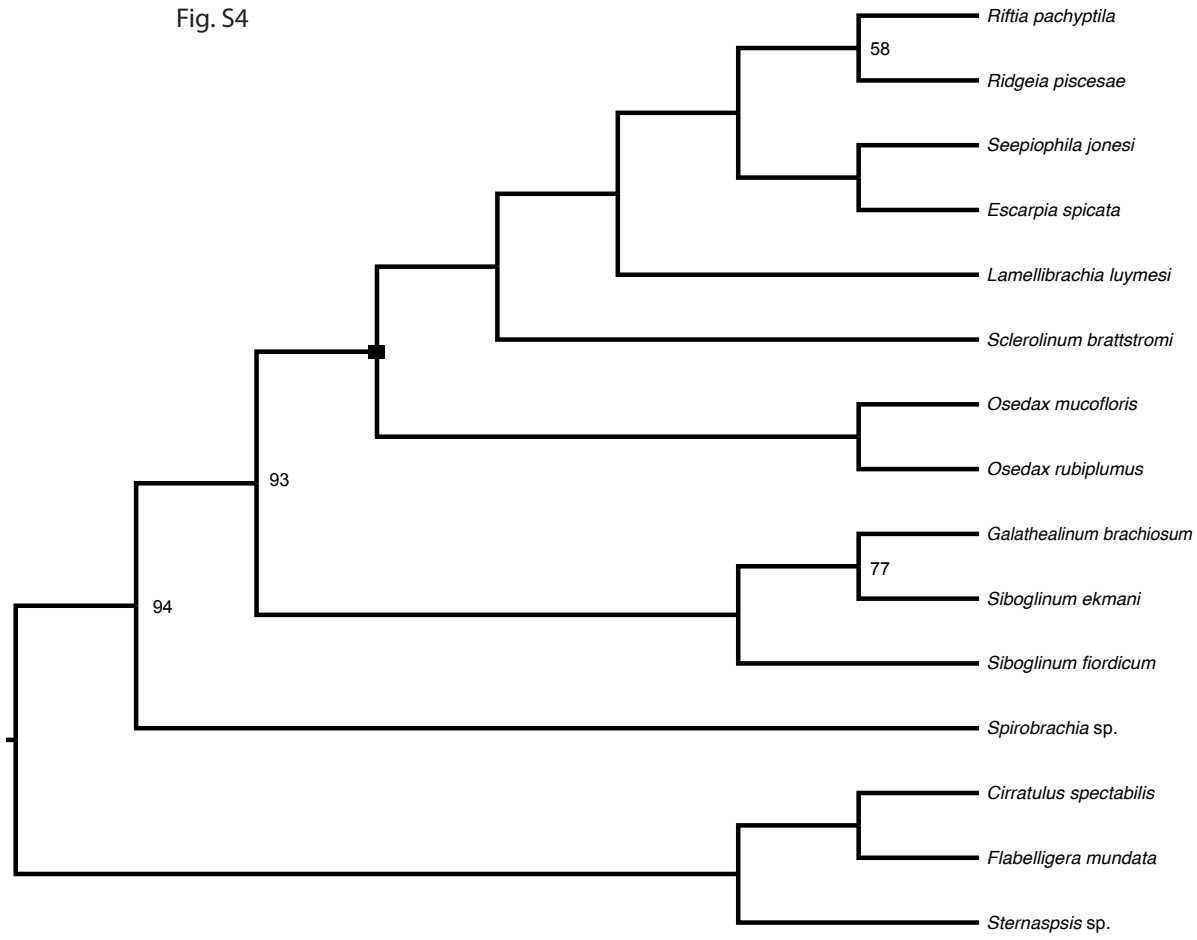


Fig. 9. Species tree inferred from ASTRAL using the D98 database. The black square indicates the node joining the *Osedax* lineage to the vestimentiferan/*Sclerolinum* clade. Nodal support values indicate bootstrap proportion based upon 100 multilocus bootstraps.

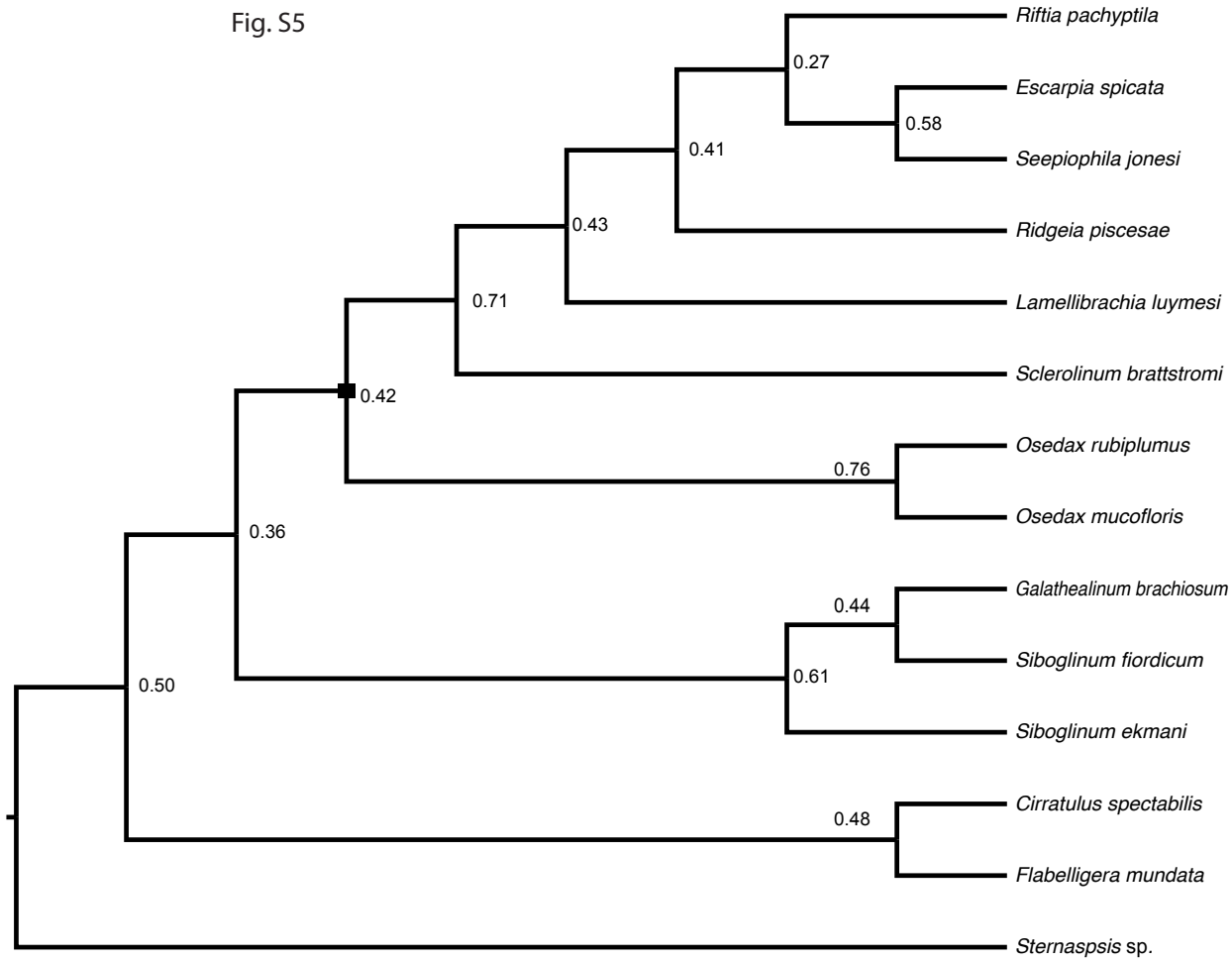


Fig. 10. Primary concordance tree reconstructed using BUCKy with 34 OGs derived from D150 dataset. The black square indicates the node joining the *Osedax* lineage to the vestimentiferan/*Sclerolinum* clade. Node values represent concordance percentages.

Table 1. Taxon sampling and source of data used in phylogenomic analyses.

Taxon	Clade	Data	Reads	Source	Accession #s
Riftia pachyptila	Siboglinidae - Vestimentifera	454	1,333,110	NCBI SRA	SRR346550
Riftia pachyptila	Siboglinidae - Vestimentifera	454	623,927	NCBI SRA	SRR346549
Escarpia spicata	Siboglinidae - Vestimentifera	454	283,594	This study	SRR3554587
Lamellibrachia luymesi	Siboglinidae - Vestimentifera	Illumina	50,537,812	This study	SRR3556248
Lamellibrachia luymesi	Siboglinidae - Vestimentifera	454	760,876	This study	SRR3556245
Ridgeia piscesae	Siboglinidae - Vestimentifera	454	1,092,906	NCBI SRA	SRR346554
Ridgeia piscesae	Siboglinidae - Vestimentifera	Sanger	515	NCBI EST	EV802484 - EV802997, EV823675
Seepiophila jonesi	Siboglinidae - Vestimentifera	454	382,144	This study	SRR3554599
Sclerolinum brattstromi	Siboglinidae - *Sclerolinum*	Illumina	44,207,372	This study	SRR3560108
Osedax mucofloris	Siboglinidae - *Osedax*	Illumina	56,067,578	This study	SRR3574511
Osedax rubiplumus	Siboglinidae - *Osedax*	Illumina	50,339,804	This study	SRR3574382
Spirobrachia sp.	Siboglinidae - Frenulata	Illumina	46,610,870	This study	SRR3571603
Siboglinum fiordicum	Siboglinidae - Frenulata	Illumina	35,922,776	This study	SRR3560206
Siboglinum ekmani	Siboglinidae - Frenulata	Illumina	63,511,320	This study	SRR3560562
Galathealinum sp.	Siboglinidae - Frenulata	454	456,440	This study	XXX
Sternaspsis sp.	Sternaspidae	Illumina	54,186,104	This study	SRR3574594
Flabelligera mundata	Flabelligeridae	Illumina	66,330,138	This study	SRR3574613
Cirratulus spectabilis	Cirratulidae	Illumina	57,767,330	This study	SRR3574861

Table 2. Statistics for phylogenomic dataset

Dataset	Taxa	HaMStR OGs	TreSpex OGs	LB scores and RCFV values	Sites	Gene occupancy%	Missing data (Including gaps) %
D98	15	244	128	98	31,276	90.1	25.0
D150	14	265	171	150	48,125	91.p	21.5
D289	14	715	301	289	103,421	81.7	34.8

Table 3

AU tests of competing phylogenetic hypothesis.

Tree Topology	D98		D150		D289	
	Log-likelihood	AU test	Log-likelihood	AU test	Log-likelihood	AU test
Osedax + Vestimentifera/*Sclerolinum*	-342993.31	1.00	-542050.62	1.00	-1021627.38	1.00
Osedax + Frenulata	-343600.84	3e-60	-542961.36	5e-41	-1022921.31	3e-18

Table 4

Specimen data for sequenced taxa.

Spieces	Clade	Specimen Collection		
		location	depth (m)	GPS coordinates
Seepiophila jonesi	Vestimentifera	Mississippi Canyon, U.S.	754	28°11.58' N, 89° 47.94' W
Escarpia spicata	Vestimentifera	Mississippi Canyon, U.S.	754	28°11.58' N, 89°47.94' W
Lamellibrachia luymesi	Vestimentifera	Mississippi Canyon, U.S.	754	28°11.58' N, 89°47.94' W
Sclerolinum brattstromi	Monilifera	Storfjorden Fjord, Norway	660	62°27.26' N, 6°47.57' E
Siboglinum fiordicum	Frenulata	Skoge Inlet, Norway	36	60°16.17' N, 5°05.53' E
Spirobrachia sp.	Frenulata	Aleutian Trench, U.S.	4890	57°27.39' N, 148°00.01' W
Galathealinum sp.	Frenulata	Mississippi Canyon, U.S.	754	28°11.58' N, 89°47.94' W
Siboglinum ekmani	Frenulata	Storfjorden Fjord, Norway	515	62°23.30' N, 6°54.58' E
Osedax mucofloris	*Osedax*	Near Bergen, Norway	N/A	on artificial whale fall
Osedax *rubiplumus*	*Osedax*	Flanders Bay, Antarctic	700	65°05.99’ S, 63°09.94’ W
Flabelligera mundata	Sternaspidae	Trinity Peninsula, Antarctic	87	63°13.74' S, 58°45.33' W
Sternaspsis sp.	Flabelligeridae	Eagle Island, Antarctic	335	63°40.00 ' S, 19°57.34' W
Cirratulus spectabilis	Cirratulidae	Cattle Point, U.S.	N/A	48° 27.18' N, 122°57.76' W

Chapter 4. Comparative genomics of seep-dwelling tubeworm (Siboglinidae: Annelida) endosymbionts

4.1 Abstract

The evolution of gutless siboglinids, which are important members of vents, seeps, muddy sediments, and whale bone communities, has been hypothesized to be driven by preference for reducing habitats and their dependence on endosymbionts. However, genomes from only a few vent-dwelling vestimentiferan and bone-eating *Osedax* endosymiont genomes have been sequenced and characterized. Here we focus on the genomes of gamma proteobacteria endosymbionts from vestimentiferan and frenulate siboglinids. Vestimentiferans tend to grow to relatively large sizes (up to 2m long and 5cm in diameter) whereas frenulates are typically more diminutive (10cm long but 0.5cm in diameter). To understand differences in these holobiont systems, we sequenced 3 vestimentiferan and 1 frenulate symbiont genomes collected at hydrocarbon seeps in Gulf of Mexico and compared them to endosymbiont genomes from hydrothermal vent regions. Similar to vent-living vestimentiferans symbionts, all sampled endosymbionts from seep-dwelling siboglinids are also able to use rTCA cycle in addition to Calven-Benson cycle for carbon fixation. However, representative of frenulates, the *Galathealinum* symbionts lack key enzymes associated with rTCA and can only use Calvin cycle for carbon fixation. Thus, we hypothesize that symbionts with higher metabolic flexibility in carbon fixation may allow tubeworms to thrive in more reducing environments, such as seeps and vents. In addition, we show that metabolisms of sulfur, nitrogen are largely conserved across

all siboglinid chemoautotrophic symbionts. Surprisingly, we find that the ability to use hydrogen as an additional energy source is probably also widespread in cold seeps than previous recognized, especially for siboglinid symbionts. Lastly, we take a comparative approach to systematically characterize the molecular mechanisms related to the process of infection, including motility guided by chemotaxis, secretion systems, type IV pili and genes potentially related to toxin and immunity. These results suggest that there are previous unrecognized links among siboglind symbionts from different deep-sea chemosynthetic environments and shed light on understanding of evolutionary trends of siboglinid host-symbiont evolution.

4.2 Introduction

Adult siboglinids are gutless and nutritionally dependent on bacterial symbionts, which are typically housed within bacteriocytes of a specialized organ called the trophosome (Southward, 1982), or in the root system in the case of *Osedax* (Rouse et al. 2004). As currently recognized, more than 200 species of siboglinids have been described within four major lineages (Li *et al.*, 2015; Li *et al.*, 2016) (Fig. 1). Vestimentifera, the best-studied lineage, includes tubeworms from hydrothermal vents (e.g., *Riftia*, *Tevnia* and *Ridgeia*) and hydrocarbon seeps (e.g., *Lamellibrachia*, *Escarpia* and *Seepiophila*). Other lineages include Monilifera (*Sclerolinum*), which live on decaying organic matter, and the bone-eating *Osedax*. Last are Frenulata, comprising the most species-rich (141 species) and ecologically diverse siboglinid lineage, mainly living in reducing sediments. In regards to siboglinid habitat preference, frenulates, sister

to all other siboglinids, mainly inhabit muddy sediments which contain lower levels of sulfide whereas the more derived taxa tend to live in increasingly reducing habitats such as seeps or vents (Schulze & Halanych, 2003).

Most siboglinids are generally associated with a single ribotype of sulphur-oxidizing γ-proteobactria (but see in methanotrophic symbionts in (Schmaljohann & Flugel, 1987)) that reduce sulfur compounds as electron donors and fix CO_2 autotrophically. However *Osedax* harbors heterotrophic Oceanospirillales that aid digestion of vertebrate bones. Siboglinid endosymbionts are obtained each generation through horizontal transmission of free-living symbionts from surrounding environment presumably after settlement of larvae (Harmer *et al.*, 2008). Previous phylogenetic analysis has also shown different lineages of siboglinds host specific lineages of symbionts (Thornhill *et al.*, 2008). The horizontal symbiosis transmission of siboglinids is similar to the *Vibrio*-squid system (Nyholm & Mcfall-Ngai, 2004) in that an infection-like process is used. Potential symbionts migrate across the epidermis into mesodermal tissue that will later develop into the trophosome (Nussbaumer *et al.*, 2006) (Bright & Bulgheresi, 2010). Upon death of the host, symbionts are released back into the environment enriching free-living populations (Klose *et al.*, 2015). Although symbiont transmission between generations is crucial for the persistence of the siboglinid-symbiont association, mechanisms underpinning this specialized infection process have not been fully characterized.

To date, only *Osedax* and vent-living vestimentiferan symbiont genomes (*Rifitia*, *Tevnia* – (Gardebrecht, 2012 #792) and *Ridgeia* – (Perez & Juniper, 2016) have been sequenced and

characterized. These recent genomic and proteomic studies suggested that vestimentiferan symbionts (Candidatus *Endoriftiua persephone*) are able to use reductive tricarboxylic acid cycle (rTCA) in addition to previous identified Calvin-Benson-Bassham (CBB) cycle for CO_2 fixation (Markert *et al.*, 2007; Robidart *et al.*, 2008; Gardebrecht *et al.*, 2012). Moreover, key enzymatic genes, RubisCO and ATP citrate lyase (ACL) type II associated with these carbon fixation cycles, were identified in *Lamellibrachi*a and *Escarpia* symbionts using a PCR approach (Thiel *et al.*, 2012). Given this limited data, how metabolic machineries differ between the endosymbionts of vent and seep-dwelling vestimentiferans, or between symbionts of vestimentiferans and their diminuative cousins, the frenulates, is not well understood.

To further explore host-symbiont associations in siboglinid tubeworms and how the endosymbiont mechanisms vary between siboglinid lineages, we sequenced and assembled four symbiont genomes from three seep-dwelling vestimentiferan (*Lamellibrachia luymesi*, *Seepiophila jonesii* and *Escarpia spicata*) and one frenulate species (*Galathealinum brachiosum*) collected from the same seep locality in the Gulf of Mexico. We compare these new endosymbiont genomes to previously published data from vent vestimentiferans to enhance our understanding of genomic structures and evolutionary trends in siboglinid symbiosis.

4.3 Materials and Methods

4.3.1 Sampling collection, DNA extraction and sequencing

Tubeworm specimens were collected from seep localities of Mississippi Canyon at 754 m depth in Gulf of Mexico (N 28°11.58', W 89°47.94'), during a *R/V Seward Johnson* and *Johnson Sea Link* in October 2010. All were frozen at 80˚C following collection. Total genomic DNA was extracted from each worm's trophosomal tissue using the DNeasy Blood & Tissue Kit (Qiagen) according to the manufacturer's protocols. Sequencing of genomic DNA was performed by The Genomic Services Lab at the Hudson Alpha Institute in Huntsville, Alabama using Illumina (San Diego, California) 2 X 100 paired-end TruSeq protocols on an Illumina HiSeq 2000 platform.

4.3.2 Genome assembly, completeness assessment

Resulting metagenomic data were assembled *de novo* with several different assemblers, Ray 2.2.0 (Boisvert *et al.*, 2012) with k-mer = 31 (value chosen based on comparing a range of k-mer values relative to final assembly) , hierarchical genome assembly (HGA) methodology (Al-Okaily, 2016) and MetaPlatanus 1.03 (Kajitani *et al.*, 2014). To identify putative symbiont contigs, BLASTN or TBLASTX (Altschul *et al.*, 1997) was performed on contigs produced by Ray using the *Riftia* symbiont genome (GenBank Accession: NZ_AFOC00000000.1, (Gardebrecht, 2012) as the query sequence. Symbiont assemblies from different assemblers were evaluated using Quast (Gurevich *et al.*, 2013) based on size N50, number of contigs, GC content and number of recovered genes and operons.

Completeness of the obtained bacteria genomes was assessed using CheckM (Parks *et al.*, 2015). Completeness of genomes was estimated via presence of 111 essential single-copy genes proposed by (Dupont *et al.*, 2012) using hmmer3 (Finn *et al.*, 2015), requiring 70% length match for each Hidden Markov models (HMMs) of these genes, with the trusted cutoff as the minimum score (Fig. 6). An additional estimate of completeness for siboglinid symbionts was compared 577 lineage-specific single copy marker sets (Parks *et al.*, 2015) from 112 γ-proteobacteria genomes (Table 2).

To ensure the relative symbiont purity in the assembly, sequences of 16S ribosomal RNA gene were extracted using BLASTN and subsequently blasted against the 16S rRNA gene database (http://greegenes.lbl.gov). To test if there was evidence of multiple symbiont ribotypes within single hosts. Raw paired-end reads were remapped to each of their respective 16S rDNA contigs using Bowtie2 (Langmead & Salzberg, 2012), and then visualized in Tablet (Milne *et al.*, 2013). Gammaproteobacteria taxa used in the phylogenetic analysis were based on previous analysis of Thornhill et al. (2008) and downloaded from TREEBASE (http://www.Treebase.org). Sequences of each OG were then aligned using MAFFT (Katoh *et al.*, 2002) with the '-auto' and '-localpair' parameters and 1000 maximum iterations. Maximum likelihood analyses were performed in RAxML under GTRGAMAMA model with rapid bootstrapping of 100 replicates (Fig. 7).

4.3.3 Genome annotation and pathway analysis

Extracted endosymbiont genomes were uploaded to the RAST server (http://rast.nmpdr.org/) (Overbeek *et al.*, 2014) for annotation. Metabolic pathway analysis of newly sequenced and publically available genomes was performed by using KEGG2 KAAS genome annotation web server (Moriya *et al.*, 2007) and then visualized by KEGG Mapper Reconstruct Pathway (http://www.genome.jp/kegg/tool/map_pathway.html). Functional genes or pathway holes were filled by using BLASTP to search for proteins that were missing in the visualized KEGG pathway or RAST annotations.

4.3.4 Comparative genomic and phylogenetic analysis

Four seep-dwelling and three previously sequenced vent-living siboglinid endosymbiont genome sequences were subjected to BLASTN comparisons and then visualized using BRIG (Alikhan *et al.*, 2011) (Fig. 2A). Whole draft genomes were also aligned with the progressive Mauve algorithm with default settings (Darling *et al.*, 2010) (Fig. 2B). To roughly characterize the protein composition of each siboglinid symbiont genome, annotated genes were assigned to RAST subsystem category and then plotted for comparison (Fig. 8).

Analysis of co-phylogeny was conducted from six siboglinids and corresponding endosymbionts; *Tevnia* was not included here due to the lack of a host transcriptome. A phylogenomic analysis, using highly conserved orthologous genes, compared the newly

sequenced four genomes with other publically available siboglinid symbiont genomes. 1-to-1 orthologous genes (OGs) were generated using ProteinOrtho (Lechner *et al.*, 2011) with the default parameters. Resulting OG clusters were then filtered from each bacterial genome following modified bioinformatics pipelines of (Whelan *et al.*, 2015; Li *et al.*, 2016). Briefly, all sequences shorter than 100 amino acid residues were discarded. Sequences of each OG were then aligned using MAFFT (Katoh *et al.*, 2002) with the '-auto' and '-localpair' parameters and 1000 maximum iterations. Uninformative and ambiguously aligned positions were trimmed with Aliscore (Misof & Misof, 2009) and Alicut (P. Kück, 2009). Alignment columns with only gaps were subsequently removed, and any OG with an alignment less than 100 amino-acid residues in length after trimming was discarded. For each OG, a custom java program, *AlignmentCompare.java*, was used to remove any sequence that did not overlap other sequences by at least 20 amino acids. For the host dataset, supermatrix dataset from selected taxa was derived from Li et al. (2016), yielding 289 OGs. For both data sets, OGs were concatenated into a single alignment using FASconCAT (Patrick Kück & Meusemann, 2010). Maximum likelihood analyses were performed in RAxML under the best-fitting models for associated partition schemes determined by PartitionFinder with rapid bootstrapping of 100 replicates (Fig. 7).

4.4 Results and discussion

4.4.1 General genomic features

The general genomic features of all siboglinid symbiont assemblies are listed in Table 1, 2. Reported completeness estimation based on the 106 essential bacterial single-copy and subsequent 577 lineage-specific core genes suggested that a > 95% overall completeness of newly-sequenced symbiont genomes (Fig. 6 and Table 2, respectively). The *Escarpia* and *Galathealinum* symbiont assembly (22 and 19 contigs, respectively) contained significantly fewer and longer contigs in comparison with other sequenced siboglinid symbionts (Table 1). Although *Seepiophila* (337 contigs), *Lamellibrachia* (345 contigs) symbiont assemblies possessed relatively more contigs than *Escarpia* and *Galathealinum* genome, the sequencing depths (~ 100X) were similar or even better than other assemblies (Table 1). Notably, the GC-content of seep-dwelling vestimentiferan symbionts (~ 54%) was slightly lower than vent-living ones (~ 58%). The GC-content of heterotrophic *Osedax* bacteria was ~ 43%, whereas the *Galathealinum* symboint genome was even lower with a content of 38.9%. The GC content of bacterial genomes has been thought to associate with genome size, oxygen and nitrogen exposure and habitats. The base composition in GC-rich bacterial genomes might reflect stronger selective forces than AT-rich genomes (Bohlin *et al.*, 2010). This correspondence in siboglinid symbionts might reflect that host-symbiont specificity in major siboglinid lineages (Schulze & Halanych, 2003), an adaptation of symbiont genomes to more-reducing environments, or possibly both. Consistent with other siboglinid assemblies, we were unable to close and circularize these symbiont genomes even with such high sequencing coverage due to the presence of repetitive regions (see Perez and Juniper, 2016). Either situation would offer an explanation on reported

difficulties in assembling siboglinid symbiont genomes in general.

The blast and pairwise comparisons of whole genome alignments revealed relatively strong homology across siboglinid symbiont genomes (Fig 2A, 2B) in line with their degree of relatedness (e.g., seep-dwelling vestimentiferan symbionts show greatest homology and the *Galathealinum* symbiont showed the least). In contrast, clusters of functional groups based on RAST subsystem showed largely physiological homogeneity among all sequenced siboglinid symbionts (Fig. 8).

4.4.2 Endosymbiont purity, phylogenetic affiliation, and cophylogenetic analysis

Consistent with results from previous analyses that most siboglinid species only host a single ribotype of bacterial symbiont (Southward, 1982), only one good hit (evalue cutoff: 1e-5) of the 16S sequence was recovered from each symbiont assembly using BLASTN against the GREENGENES database (DeSantis *et al.*, 2006), and no nucleotide sequence difference was observed via subsequent remapping of raw reads. Both results strongly supported that only one bacterial endosymbiont phylotype was present in any of the hosts examined in this study.

The dataset for phylogenetic analysis using 16S rDNA consisted of 1,458 nucleotide positions. All four endosymbionts belonged to sulfur-oxidizing bacteria within gammaproteobactria (Fig. 7). Although four seep-dwelling siboglinid endosymbionts were collected from basically the same seep locality in GOM, they harbored different ribotypes of endosymbionts. *Escarpia, Seepiophila* and *Lamellibrachia* symbiont sequences were clustered as

a well-supported clade with other seep-dwelling vestimentiferans (bs = 96), whereas *Galathealinum* was most close related with other frenulate species with a strong nodal support (bs = 100). Within seep-dwelling vestimentiferans endosymbionts, three monophyletic groups were identified in previous studies (McMullin *et al.*, 2003). *Lamellibrachia* and *Seepiophila* shared the identical 16S rRNA gene sequences and clustered within "group 2" endosymbionts. However, *Escarpia* endosymbiont was highly different than *Lamellibrachia* and *Seepiophila* symbionts (differing by 26 bp) and formed its own clade instead of affiliating with previous sequenced *Escarpia* symbionts (Fig. 7).

Coevolution between host and their symbiont was not expected to arise in siboglinids because symbionts are acquired through horizontal transmission mode. Here, we used phylogenomic analyses to assess the topological patterns between host and symbiont trees. Two datasets were generated here: the host dataset contained 289 orthologous groups (OGs) which derived from the study of Li et al. (2016); the symbiont dataset comprised of 552 highly conserved OGs after several steps of filtering. Consistent with previous results (Thornhill et al., 2008; Thial et al., 2012), cophylogenetic analyses (Fig. 3) revealed incongruence phylogeny between the hosts and symbionts. Specifically, although Galthealinum symbiont was sampled from basically the same locality from other seep-dwelling vestimentiferans, it was placed as the most basal lineage in both host and symbiont trees. This is consistent with previous analysis that major sibolginid lineages generally associate with a different bacterial clade regardless of

inhibited habitats of the hosts (Thornhill *et al.*, 2008).

4.4.3 Metabolic Pathways

Carbon fixation

All sequenced siboglinid symbiont genomes contained core components of the Calvin-Benson-Bassham (CBB) cycle for carbon fixation (Fig. 4). Similar to previous sequenced *Ca*. *Endoriftia persephone* species (Markert *et al.*, 2007; Robidart *et al.*, 2008; Gardebrecht *et al.*, 2012), all components of rTCA cycle were also found in *Lamellibrachia*, *Escarpia* and *Seepiophila* symbiont genomes. Most of the enzymes in the TCA/rTCA pathways are shared, with the exception of three key enzymes that allow the cycle to run in reverse: ATP citrate lyase (*aclAB*), 2-oxoglutarate:ferredoxin oxidoreductase *(korB*), and fumarate reductase (*sdhAC*) (Hugler *et al.*, 2005). Unlike vestimentiferan symbionts, the three key enzymes specifically associated with rTCA cycle were completely missing in the *Galathealinum* symbiont genome (Fig. 4). More importantly, these three key gene sets were not genomically clustered together as identified in all vestimentiferan symbiont genomes so lack of these genes in *Galathealinum* symbiont genome was unlikely due to incompleteness of genome assembly. In addition, multiple copies of shared enzymes that function in both pathways were identified in all sequenced vestimentiferan symbiont genomes since they have been proposed to employ separate sets of genes for the oxidative and reverse direction of TCA cycle, whereas only one copy of these genes was discovered in the *Galathealinum* symbiont (e.g. pyruvate:ferredoxin oxidoreductase

(PorAG): 4 - 5 copies in vestimentiferan symbionts, only 1 copy in *Galathelainum* symbiont) (Table S1).-The presence of a rTCA cycle and multiple copies of associated genes involved in oxidation may account for the ability of vestimentiferans to fix far more energy, resulting in greater body size and larger population size, than their Frenulate cousins.

As suggested by Thiel et al., 2012, the presence of rTCA in addition to CBB pathways for carbon might be common in all vent/seep-living vestimentiferan endosymbionts. Presently available data suggest that CBB cycle is the predominant carbon fixation pathway in both seeps and vents, whereas thiotrophs using rTCA cycle are thought to be found only in symbionts from vent environments (Hugler & Sievert, 2011). However, endosymbionts of seep-dwelling vestimentiferans and some other polychaetes (e.g. *Branchinotogluma sandersi*) also relied on rTCA cycle for carbon fixation, thus providing additional support for widespread occurrence and importance of rTCA cycle in seeps (Portail *et al.*, 2016).

The rTCA cycle is significantly less energy demanding than CBB cycle and generally considered the most energy-efficient CO_2 fixation pathway (Berg, 2011). However, several enzyme associated with rTCA cycle are highly oxygen sensitive compared to CBB cycle, therefore limit rTCA cycle only found in anaerobic and microanaerobic bacteria that occur in certain anoxic environments (e.g. vents)(Erb, 2011). The possession of rTCA cycle in vestimentiferan symbionts may be important for symbionts' heterotrophic living stage since they must travel across the oxic-anoxic interface to access both reducing compounds from anoxic habitats as energy sources and the oxygen from the water column for aerobic metabolism

(Cavanaugh *et al.*, 2006). Gardebrecht et al., 2012 also reported that spot volume in enzymes associated with rTCA was higher in *Tevnia* than *Riftia* symbionts owing to higher oxygen levels in the vent fluid surrounding *Riftia*. Moreover, symbionts with depleted energy sources because of low sulfur contents may switch to using the rTCA cycle, which would allow for a high metabolic flexibility and thus facilitate adaptation in different environmental conditions (Markert *et al.*, 2007; Hugler & Sievert, 2011)

Although, the *Galathealinum* symbiont genome contains all components required for CBB cycle similar to vestimentiferans, key enzymes associated with rTCA cycle were lacking. Frenulates are sister to other siboglinid lineages and mainly found in anoxic reducing muddy sediments with a relatively lower sulfur level than vents and seeps (Hilario *et al.*, 2011). Although frenulate species *Galathealinum brachiosum* was collected in the muddy sediment nearby the seep locality along with other seep-dwelling vestimentiferans, it harbored endosymbionts belong to a different bacterial clade (see above). Along with these lines, we speculate that host siboglinids uptake and switch to different lineages of bacterial that have a higher metabolic flexibility of carbon fixation (e.g. rTCA) in surroundings which would allow them to successfully exploit and thrive in more reducing habitats, such as vents and seeps.

Sulfur, nitrogen and hydrogen metabolism

Given reducing habitats in which chemoautotrophic symbioses occur, sulfur-oxidizing symbiont bacteria oxidize reduced inorganic sulfur compounds to provide energy needed for hosts (Cavanaugh *et al.*, 2006). Consistent with *Riftia* symbionts (details of the sulfide oxidation pathway are given in (Markert *et al.*, 2011)), all sequenced tubeworm symbiont genomes also possessed a reverse sulfate reduction pathway for sulfide oxidation (APS reductase pathway), involving the enzymes APS reductase (*AprA*/*AprB*), ATP sulfurylase (*DsrA*/*DsrB*) and ATP sulfurylase (*SopT*) (Fig. 4). Furthermore, components of thiosulfate-oxidizing Sox enzyme system (*SoxABXYZ*) were also identified, whereas SoxCD homolog was missing in all siboglinid symbionts. Given that genomic capacities of sulfur metabolic pathways are largely conserved across all siboglinid symbionts, we suggest that seep-dwelling vestimentiferan and frenulate symbionts also oxidize sulfide to sulfate via the APS reductase pathway (Fig. 4). Moreover, since sulfide, instead of thiosulfate, is inferred to be the preferred energy source for vent-living vestimentiferan symbionts (Markert *et al.*, 2011) (Gardebrecht *et al.*, 2012), the presence of Sox system in symbiont genomes suggests that thiosulfate oxidation might serve as an alternative sulfur metabolic pathway while sulfide is at minimum levels. SoxCD homolog is absent in most sulfur globule-forming organisms, such as *Allochromatium vinosum* (Weissgerber *et al.*, 2011). Among these organisms, the Sox complex is involved in the oxidation of thiosulfate to sulfate and then store sulfur globules as intermediates (Welte *et al.*, 2009). Since sulfur globules have been isolated in *Riftia* symbionts, formation of intracellular sulfur globules for sulfur storage might be common across siboglinids (Fig. 4).

Nitrate is extremely abundant in deep-sea hydrothermal vents and cold seeps (Bowles & Joye, 2011) (Lee & Childress, 1994). All four seep-dwelling siboglinid symbiont genomes were found to encode the entire set of enzymes required for using nitrate as a terminal electron acceptor during anaerobic respiration (nitrate respiration) and nitrate can also be reduced to ammonia and then incorporated into amino acids through the sequential action of glutamine synthase (GS) and glutamate synthase (GOGAT) enzymes for biosynthesis and growth (ammonia assimilation)(Hentschel & Felbeck, 1993). Previous analysis suggested that majority of nitrate is only reduced to ammonia instead of nitrogen gas for nitrate respiration in *Riftia* symbionts (Girguis *et al.*, 2000). Genes required to convert nitrite to nitrous oxide were found in all symbiont genomes (nitrite reductase—*nirK* and nitrite oxidoreductase—*norCB*). Unexpectedly, two types of nitrate reductases (membrane-bound respiratory nitrate reductase: *NarGHIJ*; periplasmic dissimilatory nitrate reductase: *NapABC*) that can catalyze reduction of nitrate to nitrite and their associated electron carriers were found in *Riftia*, *Tevnia*, *Lamellibrachia* and *Seepiophila* symbionts, whereas only Nap operon was found in *Escarpia* and *Galathealinum* symbiont genomes (Fig. 4). Comprising both Nap and Nar or single Nap enzyme systems have been documented in other organisms. For example, analysis of multiple species within the bacteria genus *Shewanella* (widespread in diverse environments, such as freshwater, hydrothermal vents, deep-sea sediments, etc.) revealed that at least five species within the bacteria genus *Shewanella* have both Nap and Nar systems, while other species only contain Nap

system (Chen & Wang, 2015). The Nap system used in *Riftia* symbiont has been proposed to compensate *NarGHIJ* and perform nitrate respiration in environments in which the amount of nitrate is extremely low (Gardebrecht *et al.*, 2012). Indeed, many Epsilonproteobacteria species only contain Nap are capable of anaerobic nitrate respiration (Kern & Simon, 2009). However, respiration via Nap system is thought to typically generate less energy than Nar system (Sparacino-Watkins *et al.*, 2014), although diverse physiological functions have been suggested for the bacterial Nap systems, such as the dissipation of excess reducing equivalents for redox balancing, denitrification, the adaptation to anaerobic growth, and in compensating nitrate respiration in nitrate-limited environments (Gavira *et al.*, 2002). Along with these lines, nitrate metabolism in siboglinid symbionts are more complex than we previously recognized and might be selectively modified to adapt to each bacterium's physiological function or associated with symbiont's different life stages, although further analysis is warranted to explicitly investigate these hypotheses.

Many symbionts associated with animals living in hydrothermal vents can also use hydrogen to power primary production, such as symbionts of *Bathymodiolus* mussels and shrimp *Rimicaris exoculata* (Petersen *et al.*, 2011; Anantharaman *et al.*, 2013). The key gene for hydrogen oxidation, *hupL,* has been identified in vent-living vestimentiferan symbiont genomes. A *hupL* gene fragment has also been amplified in *Lamellibrachia anaximandri* found at sedimented hydrothermal vent sites (Thiel et al., 2012). However, molecular hydrogen cycling in seep-dwelling organisms has not been extensively studied before. Previous analysis failed to

amplify *hupL* gene from any cold seep *Batyhmodiolus* mussels collected from GOM, and the authors have suggested that was because of multiple habitat-specific gene loss events occurred in the cold seep environment (Petersen *et al.*, 2011). Interestingly, we were be able to identify the core enzyme sets required for hydrogen oxidation, including genes related to structural assembly (*HypABCDEF* operon), hydrogen uptake (*HupS*, *HupL*), oxidation and energy production (membrane-bound group 1 Ni, Fe hydrogenase) in all seep-dwelling vetimentiferan and *Galathealinum* symbiont genomes in this study. Thus, the presence of molecular hydrogen metabolism pathway is not restricted to the vent symbioses, but rather might be more widespread in cold seep symbioses than previously recognized. Besides *Osedax*, the ability to use hydrogen as an additional energy source is probably common in siboglinid symbionts (Fig. 4).

4.4.4 Chemotaxis and motility

Motility and chemotaxis are essential for many pathogenic or symbiotic species during the free-living stage to allow colonization and infection of a host (Wadhams & Armitage, 2004). For example, chemotaxis can guide *Vibrio anguillarium* to the surface tissue of fish and *Vibrio fischeri* to the squid light organs (Nyholm & Mcfall-Ngai, 2004). Although previous Cand. E *Persephone* metagenomic analysis showed remarkable dedication to chemoreception and motility (Robidart *et al.*, 2008), detailed molecular pathway has not been well characterized. A spectacular range of methy-accepting chemotaxis proteins (MCPs) and their associated chemotaxis protein homologs (Che) were identified in vent-living (Gardebrecht *et al.*, 2012;

Perez & Juniper, 2016) and seep-dwelling siboglinid symbiont genomes (Table S1). Most environmental signals are sensed by receptors (MCPs) and MCPs subsequently regulate the activity of the histidine protein kinase CheA through the linker protein CheW (Zusman *et al.*, 2007). Although chemosensory pathways can usually be identified by a specific CheA (Zusman *et al.*, 2007), several conserved chemosensory co-located gene clusters were identified in most siboglinid symbiont genomes (Fig. 5), although these gene clusters were not found in *Lamellibrachia* and *Seepiophila* symbiont genomes probably due to highly fragmented genome assembly. Although the actual function is still unknown, the gene arrangement of cheI cluster in tubeworm symbionts is highly similar to che3 system in *M. xanthus* and che 2 system in *Rhizobium leguminosarum* (Miller *et al.*, 2007) which has a minor effect on bacterial motility. The cheII system identified in siboglinid symbiont genomes is highly similar to bivalve symbiont *Solemya velum*, both contain pilGHIJ-cheAW gene clusters, which are known to control twitching motility in other bacteria (Dmytrenko *et al.*, 2014). For cheIII chemosensory system, chemotaxis protein homologs are co-located with flagellar biosynthesis gene clusters and flagellar motility proteins (MotA and MotB). As shown in *E. coli*, motA and motB are required for the generation of rotation force during the flagellar movement (Stolz & Berg, 1991). Thus, the cheIII system identified here might have a function in controlling flagellar motility guided by chemotaxis for siboglinid symboints. Noticeable, all sequenced symbiont genomes contained the full complement of genes that were indispensable for a functional flagellum and type IV pili (T4P), plus a number of accessories or duplicated components. Although a flagellum was not

detected in *Riftia* symbionts while in the trophosome, it might be present in the free-living life stage and genes associated with flagellar synthesis are no longer expressed after infection (Millikan *et al.*, 1999) (Robidart *et al.*, 2008). Altogether, motility of siboglinid bacteria was probably mediated by the flagellar motility and extension, retraction of type IV pili T4P guided by chemotaxis.

4.4.5 Host infection

Although the symbiotic association between tubeworms and symbionts are essential for them to thrive in deep-sea reducing habitats, molecular mechanisms that underpin the infection process are still largely unknown. For horizontally transmitted symbionts, the infection process can be divided into two steps: colonization of host cells and then travel to the symbiont housing organ (Bright & Bulgheresi, 2010). Siboglinids symbionts infect the skin of the larvae during post-settlement stage and then travel through host epidermal cells into a mesodermal tissue that will later develop into the trophosome, while the infection process finishes simultaneously with massive apoptosis of skin tissue in the juvenile stage (Nussbaumer *et al.*, 2006). Below we discuss several important mechanisms in this host infection process.

Adhesion: To establish the initial adhesion of bacteria to host surfaces, siboglinid symbionts might reach their respective hosts through flagelllar and T4P motility guided by chemotaxis (Bright and Bulgheresi; also see above section). The initial physical encounter between host and symbionts occurs in mucous extracellular substance secreted by newly settled

larvae (Nussbaumer *et al.*, 2006). In mucus matrices, extracellular components from symbionts, such as lipopolysaccharide (LPS) and O-antigen, can mediate direct physical contact between symbionts and their hosts (Bright & Bulgheresi, 2010; Powell *et al.*, 2016). T4P might also be essential for adhesion and biofilm formation through its twitching motility, which are thought to play a role in pathogenesis (Pizarro-Cerda & Cossart, 2006). In addition to T4P, several adhesion-related proteins such as ankyrin like protein, fironectin type III domain (previous identified in (Gardebrecht *et al.*, 2012)) and tetratrico peptide repeat domain (only found in vent-living vestimentiferan symbionts) were identified here (Table S1), which might be also involved mediating host-symbiont adhesion.

Secretion system and virulence gene homologs: Unlike many symbionts (e.g. rhizobia, enteropathogenic *Escherichia coli*) used proteins secreted by a type III secretion system (T3SS) to avoid phagocytosis and facilitate bacterial invasion to host cells (Costa *et al.*, 2015), T3SS was lacking in siboglinid symbionts and instead possessed all the conserved core components of a type II secretion system (T2SS). Hemolysin and chitinase exported by the T2SS have been shown to be important for virulence in many pathogens and beneficial microbes (e.g. *Aeromonas veronii* in the leech gut (Maltz, 2011 #816) and *Burkholderia rhizoxinica* in the fungus (Moebius *et al.*, 2014). A hemolysin III (hlyIII) gene encoding a hemolysin (Gardebrecht *et al.*, 2012; Goffredi *et al.*, 2014) and a gene for a chitinase were identified in all siboglinid symbionts which may enable the bacteria to permeablilize the host cells and inter- and intracellularly migrate to newly developed trophosome tissue (Bright & Bulgheresi, 2010) (Gardebrecht *et al.*, 2012).

Although symbionts are mostly dependent on the production of specific toxins, extracellular proteolytic enzymes are also thought to play key roles in host colonization (Duarte *et al.*, 2016). We found genes encoding a chitinase (see above) and collagenase in siboglinid symbiont genomes, which might assist the symbionts travel into host tissues and facilitate diffusion of toxins (Duarte *et al.*, 2016). Furthermore, unlike symbionts of *Bathymodiolus* mussels, which contain abundant toxin-related genes (Sayavedra *et al.*, 2015), only a few RTX (repeats in toxins) genes were identified in all siboglinid symbiont genomes, which might also involve in host-symbiont interactions (Sayavedra *et al.*, 2015). In accordance with *Riftia* and *Tevnia* symbiont (Gardebrecht *et al.*, 2012), the predicted proteins belonging to the virulence resistance category are involved in the production of bacteriocin Colicin V.

Oxidative stress: In addition to physical barriers, traveling symbionts have to overcome the host cellular immune responses (Bright & Bulgheresi, 2010). In accordance with the previous description of oxidative stress response in *Riftia* (Gardebrecht *et al.*, 2012), superoxide dismutase (FeSOD) and alkylhydroperoxide reductase (AhpC) were identified in all vestimentiferan symbiont genomes in this study. However, FeSOD was lacking in the *Galathealinum* symbiont genome. Furthermore, a rubrerythrin (Rbr) was detected in all siboglinid symbiont genomes, which has been suggested to play a role in protection against hydrogen peroxide-mediated oxidative stress and damage (Konig *et al.*, 2016). In addition, multiple gene copies of Cytochrome c peroxidase and Thiol peroxidase were also identified in siboglinid sulfur-oxidizing symbionts.

4.6 References

Al-Okaily, A. A. (2016). Hga: De novo genome assembly method for bacterial genomes using high coverage short sequencing reads. *BMC Genomics*, 17, 193.10.1186/s12864-016-2515-7

Alikhan, N. F., Petty, N. K., Ben Zakour, N. L., & Beatson, S. A. (2011). Blast ring image generator (brig): Simple prokaryote genome comparisons. *BMC Genomics*, 12, 402.10.1186/1471-2164-12-402

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped blast and psi-blast: A new generation of protein database search programs. *Nucleic Acids Research*, 25, 3389-3402.10.1093/nar/25.17.3389

Anantharaman, K., Breier, J. A., Sheik, C. S., & Dick, G. J. (2013). Evidence for hydrogen oxidation and metabolic plasticity in widespread deep-sea sulfur-oxidizing bacteria. *Proc Natl Acad Sci U S A*, 110, 330-335.10.1073/pnas.1215340110

Berg, I. A. (2011). Ecological aspects of the distribution of different autotrophic co2 fixation pathways. *Applied and Environmental Microbiology*, 77, 1925-1936.10.1128/Aem.02473-10

Bohlin, J., Snipen, L., Hardy, S. P., Kristoffersen, A. B., Lagesen, K., Donsvik, T., et al. (2010). Analysis of intra-genomic gc content homogeneity within prokaryotes. *BMC Genomics*, 11.Artn 464

10.1186/1471-2164-11-464

Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F., & Corbeil, J. (2012). Ray meta: Scalable de novo metagenome assembly and profiling. *Genome Biol*, 13, R122.10.1186/gb-2012-13-12-r122

Bowles, M., & Joye, S. (2011). High rates of denitrification and nitrate removal in cold seep sediments. *The ISME Journal*, 5, 565-567.10.1038/ismej.2010.134

Bright, M., & Bulgheresi, S. (2010). A complex journey: Transmission of microbial symbionts. *Nature reviews. Microbiology*, 8, 218-230.10.1038/nrmicro2262

Cavanaugh, C. M., McKiness, Z. P., Newton, I. L., & Stewart, F. J. 2006. Marine chemosynthetic symbioses*The prokaryotes* pp. 475-507): Springer.

Chen, Y., & Wang, F. (2015). Insights on nitrate respiration by shewanella. *Frontiers in Marine Science*, 1.10.3389/fmars.2014.00080

Costa, T. R., Felisberto-Rodrigues, C., Meir, A., Prevost, M. S., Redzej, A., Trokter, M., et al. (2015). Secretion systems in gram-negative bacteria: Structural and mechanistic insights. *Nature reviews. Microbiology*, 13, 343-359.10.1038/nrmicro3456

Darling, A. E., Mau, B., & Perna, N. T. (2010). Progressivemauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS One*, 5.ARTN e1114710.1371/journal.pone.0011147

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., et al. (2006). Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. *Applied and Environmental Microbiology*, 72, 5069-5072.10.1128/Aem.03006-05

Dmytrenko, O., Russell, S. L., Loo, W. T., Fontanez, K. M., Liao, L., Roeselers, G., et al. (2014). The genome of the intracellular bacterium of the coastal bivalve, solemya velum: A blueprint for thriving in and out of symbiosis. *BMC Genomics*, 15, 924.10.1186/1471-2164-15-924

Duarte, A. S., Correia, A., & Esteves, A. C. (2016). Bacterial collagenases - a review. *Critical Reviews in Microbiology*, 42, 106-126.10.3109/1040841x.2014.904270

Dupont, C. L., Rusch, D. B., Yooseph, S., Lombardo, M. J., Richter, R. A., Valas, R., et al. (2012). Genomic insights to sar86, an abundant and uncultivated marine bacterial lineage. *The ISME Journal*, 6, 1186-1199.10.1038/ismej.2011.189

Erb, T. J. (2011). Carboxylases in natural and synthetic microbial pathways. *Applied and Environmental Microbiology*, 77, 8466-8477.10.1128/Aem.05702-11

Finn, R. D., Clements, J., Arndt, W., Miller, B. L., Wheeler, T. J., Schreiber, F., et al. (2015). Hmmer web server: 2015 update. *Nucleic Acids Research*, 43, W30-38.10.1093/nar/gkv397

Gardebrecht, A., Markert, S., Sievert, S. M., Felbeck, H., Thurmer, A., Albrecht, D., et al. (2012). Physiological homogeneity among the endosymbionts of riftia pachyptila and

tevnia jerichonana revealed by proteogenomics. *The ISME Journal*, 6, 766-776.10.1038/ismej.2011.137

Gavira, M., Roldan, M. D., Castillo, F., & Moreno-Vivian, C. (2002). Regulation of nap gene expression and periplasmic nitrate reductase activity in the phototrophic bacterium rhodobacter sphaeroides dsm158. *Journal of Bacteriology*, 184, 1693-1702.10.1128/JB.184.6.1693-1702.2002

Girguis, P. R., Lee, R. W., Desaulniers, N., Childress, J. J., Pospesel, M., Felbeck, H., et al. (2000). Fate of nitrate acquired by the tubewormriftia pachyptila. *Applied and Environmental Microbiology*, 66, 2783-2790.10.1128/AEM.66.7.2783-2790.2000

Goffredi, S. K., Yi, H., Zhang, Q., Klann, J. E., Struve, I. A., Vrijenhoek, R. C., et al. (2014). Genomic versatility and functional variation between two dominant heterotrophic symbionts of deep-sea osedax worms. *The ISME Journal*, 8, 908-924.10.1038/ismej.2013.201

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). Quast: Quality assessment tool for genome assemblies. *Bioinformatics*, 29, 1072-1075.10.1093/bioinformatics/btt086

Harmer, T. L., Rotjan, R. D., Nussbaumer, A. D., Bright, M., Ng, A. W., DeChaine, E. G., et al. (2008). Free-living tube worm endosymbionts found at deep-sea vents. *Applied and Environmental Microbiology*, 74, 3895-3898.10.1128/Aem.02470-07

Hentschel, U., & Felbeck, H. (1993). Nitrate respiration in the hydrothermal vent tubeworm riftia-pachyptila. *Nature*, 366, 338-340.DOI 10.1038/366338a0

Hilario, A., Capa, M., Dahlgren, T. G., Halanych, K. M., Little, C. T., Thornhill, D. J., et al. (2011). New perspectives on the ecology and evolution of siboglinid tubeworms. *PLoS One*, 6, e16309.10.1371/journal.pone.0016309

Hugler, M., & Sievert, S. M. (2011). Beyond the calvin cycle: Autotrophic carbon fixation in the ocean. *Annual Review of Marine Science, Vol 3*, 3, 261-289.10.1146/annurev-marine-120709-142712

Hugler, M., Wirsen, C. O., Fuchs, G., Taylor, C. D., & Sievert, S. M. (2005). Evidence for autotrophic co2 fixation via the reductive tricarboxylic acid cycle by members of the epsilon subdivision of proteobacteria. *Journal of Bacteriology*, 187, 3020-3027.10.1128/JB.187.9.3020-3027.2005

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., et al. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res*, 24, 1384-1395.10.1101/gr.170720.113

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). Mafft: A novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research*, 30, 3059-3066

Kern, M., & Simon, J. (2009). Electron transport chains and bioenergetics of respiratory nitrogen metabolism in wolinella succinogenes and other epsilonproteobacteria. *Biochimica Et Biophysica Acta-Bioenergetics*, 1787, 646-656.10.1016/j.bbabio.2008.12.010

Klose, J., Polz, M. F., Wagner, M., Schimak, M. P., Gollner, S., & Bright, M. (2015). Endosymbionts escape dead hydrothermal vent tubeworms to enrich the free-living population. *Proc Natl Acad Sci U S A*, 112, 11300-11305.10.1073/pnas.1501160112

Konig, S., Gros, O., Heiden, S. E., Hinzke, T., Thurmer, A., Poehlein, A., et al. (2016). Nitrogen fixation in a chemoautotrophic lucinid symbiosis. *Nat Microbiol*, 2, 16193.10.1038/nmicrobiol.2016.193

Kück, P. (2009). Alicut: A perlscript which cuts aliscore identified rss. *Department of Bioinformatics, Zoologisches Forschungsmuseum A. Koenig (ZFMK), Bonn, Germany, version*, 2

Kück, P., & Meusemann, K. (2010). Fasconcat: Convenient handling of data matrices. *Molecular Phylogenetics and Evolution*, 56, 1115-1118.10.1016/j.ympev.2010.04.024

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9, 357-359.10.1038/nmeth.1923

Lechner, M., Findeiss, S., Steiner, L., Marz, M., Stadler, P. F., & Prohaska, S. J. (2011). Proteinortho: Detection of (co-)orthologs in large-scale analysis. *Bmc Bioinformatics*, 12, 124.10.1186/1471-2105-12-124

Lee, R. W., & Childress, J. J. (1994). Assimilation of inorganic nitrogen by marine-invertebrates and their chemoautotrophic and methanotrophic symbionts. *Applied and Environmental Microbiology*, 60, 1852-1858

Li, Y., Kocot, K. M., Schander, C., Santos, S. R., Thornhill, D. J., & Halanych, K. M. (2015). Mitogenomics reveals phylogeny and repeated motifs in control regions of the deep-sea family siboglinidae (annelida). *Mol Phylogenet Evol*, 85, 221-229.10.1016/j.ympev.2015.02.008

Li, Y., Kocot, K. M., Whelan, N. V., & Halanych, K. M. (2016). Phylogenomics of tubeworms (annelida: Siboglinidae) facilitates comparative performance of supermatrix versus species tree phylogenetic approaches. *Integr Comp Biol*, 56, E126-E126

Markert, S., Arndt, C., Felbeck, H., Becher, D., Sievert, S. M., Hugler, M., et al. (2007). Physiological proteomics of the uncultured endosymbiont of riftia pachyptila. *Science*, 315, 247-250.10.1126/science.1132913

Markert, S., Gardebrecht, A., Felbeck, H., Sievert, S. M., Klose, J., Becher, D., et al. (2011). Status quo in physiological proteomics of the uncultured riftia pachyptila endosymbiont. *Proteomics*, 11, 3106-3117.10.1002/pmic.201100059

McMullin, E. R., Hourdez, S., Schaeffer, S. W., & Fisher, C. R. (2003). Phylogeny and biogeography of deep sea vestimentiferan tubeworms and their bacterial symbionts. *Symbiosis*, 34, 1-41

Miller, L. D., Yost, C. K., Hynes, M. F., & Alexandre, G. (2007). The major chemotaxis gene cluster of rhizobium leguminosarum bv. Viciae is essential for competitive nodulation. *Mol Microbiol*, 63, 348-362.10.1111/j.1365-2958.2006.05515.x

Millikan, D. S., Felbeck, H., & Stein, J. L. (1999). Identification and characterization of a flagellin gene from the endosymbiont of the hydrothermal vent tubeworm riftia pachyptila. *Appl Environ Microbiol*, 65, 3129-3133

Milne, I., Stephen, G., Bayer, M., Cock, P. J. A., Pritchard, L., Cardle, L., et al. (2013). Using tablet for visual exploration of second-generation sequencing data. *Briefings in bioinformatics*, 14, 193-202.10.1093/bib/bbs012

Misof, B., & Misof, K. (2009). A monte carlo approach successfully identifies randomness in multiple sequence alignments: A more objective means of data exclusion. *Syst Biol*, 58, 21-34.10.1093/sysbio/syp006

Moebius, N., Uzum, Z., Dijksterhuis, J., Lackner, G., & Hertweck, C. (2014). Active invasion of bacteria into living fungal cells. *Elife*, 3, e03007.10.7554/eLife.03007

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., & Kanehisa, M. (2007). Kaas: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research*, 35, W182-185.10.1093/nar/gkm321

Nussbaumer, A. D., Fisher, C. R., & Bright, M. (2006). Horizontal endosymbiont transmission in hydrothermal vent tubeworms. *Nature*, 441, 345-348.10.1038/nature04793

Nyholm, S. V., & Mcfall-Ngai, M. J. (2004). The winnowing: Establishing the squid-vibrio symbiosis. *Nature Reviews Microbiology*, 2, 632-642.10.1038/nrmicro957

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., et al. (2014). The seed and the rapid annotation of microbial genomes using subsystems technology (rast). *Nucleic Acids Research*, 42, D206-214.10.1093/nar/gkt1226

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). Checkm: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*, 25, 1043-1055.10.1101/gr.186072.114

Perez, M., & Juniper, S. K. (2016). Insights into symbiont population structure among three vestimentiferan tubeworm host species at eastern pacific spreading centers. *Appl Environ Microbiol*, 82, 5197-5205.10.1128/AEM.00953-16

Petersen, J. M., Zielinski, F. U., Pape, T., Seifert, R., Moraru, C., Amann, R., et al. (2011). Hydrogen is an energy source for hydrothermal vent symbioses. *Nature*, 476, 176-180.10.1038/nature10325

Pizarro-Cerda, J., & Cossart, P. (2006). Bacterial adhesion and entry into host cells. *Cell*, 124, 715-727.10.1016/j.cell.2006.02.012

Portail, M., Olu, K., Dubois, S. F., Escobar-Briones, E., Gelinas, Y., Menot, L., et al. (2016). Food-web complexity in guaymas basin hydrothermal vents and cold seeps. *PLoS One*, 11.ARTN e016226310.1371/journal.pone.0162263

Powell, J. E., Leonard, S. P., Kwong, W. K., Engel, P., & Moran, N. A. (2016). Genome-wide screen identifies host colonization determinants in a bacterial gut symbiont. *Proc Natl Acad Sci U S A*, 113, 13887-13892.10.1073/pnas.1610856113

Robidart, J. C., Bench, S. R., Feldman, R. A., Novoradovsky, A., Podell, S. B., Gaasterland, T., et al. (2008). Metabolic versatility of the riftia pachyptila endosymbiont revealed through metagenomics. *Environmental Microbiology*, 10, 727-737.10.1111/j.1462-2920.2007.01496.x

Sayavedra, L., Kleiner, M., Ponnudurai, R., Wetzel, S., Pelletier, E., Barbe, V., et al. (2015). Abundant toxin-related genes in the genomes of beneficial symbionts from deep-sea hydrothermal vent mussels. *Elife*, 4, e07966.10.7554/eLife.07966

Schmaljohann, R., & Flugel, H. J. (1987). Methane-oxidizing bacteria in pogonophora. *Sarsia*, 72, 91-99

Schulze, A., & Halanych, K. M. (2003). Siboglinid evolution shaped by habitat preference and sulfide tolerance. *Hydrobiologia*, 496, 199-205.Doi 10.1023/A:1026192715095

Southward, E. C. (1982). Bacterial symbionts in pogonophora. *Journal of the Marine Biological Association of the United Kingdom*, 62, 889-906

Sparacino-Watkins, C., Stolz, J. F., & Basu, P. (2014). Nitrate and periplasmic nitrate reductases. *Chemical Society Reviews*, 43, 676-706.10.1039/c3cs60249d

Stolz, B., & Berg, H. C. (1991). Evidence for interactions between mota and motb, torque-generating elements of the flagellar motor of escherichia-coli. *Journal of Bacteriology*, 173, 7033-7037.10.1128/jb.173.21.7033-7037.1991

Thiel, V., Hugler, M., Blumel, M., Baumann, H. I., Gartner, A., Schmaljohann, R., et al. (2012). Widespread occurrence of two carbon fixation pathways in tubeworm endosymbionts: Lessons from hydrothermal vent associated tubeworms from the mediterranean sea. *Front Microbiol*, 3, 423.10.3389/fmicb.2012.00423

Thornhill, D. J., Wiley, A. A., Campbell, A. L., Bartol, F. F., Teske, A., & Halanych, K. M. (2008). Endosymbionts of siboglinum fiordicum and the phylogeny of bacterial endosymbionts in siboglinidae (annelida). *Biol Bull*, 214, 135-144.10.2307/25066670

Wadhams, G. H., & Armitage, J. P. (2004). Making sense of it all: Bacterial chemotaxis. *Nat Rev Mol Cell Biol*, 5, 1024-1037.10.1038/nrm1524

Weissgerber, T., Zigann, R., Bruce, D., Chang, Y. J., Detter, J. C., Han, C., et al. (2011). Complete genome sequence of allochromatium vinosum dsm 180(t). *Standards in Genomic Sciences*, 5, 311-330.10.4056/sigs.2335270

Welte, C., Hafner, S., Kratzer, C., Quentmeier, A., Friedrich, C. G., & Dahl, C. (2009). Interaction between sox proteins of two physiologically distinct bacteria and a new protein involved in thiosulfate oxidation. *FEBS Lett*, 583, 1281-1286.10.1016/j.febslet.2009.03.020

Whelan, N. V., Kocot, K. M., Moroz, L. L., & Halanych, K. M. (2015). Error, signal, and the placement of ctenophora sister to all other animals. *Proc Natl Acad Sci U S A*, 112, 5773-5778.10.1073/pnas.1503453112

Zusman, D. R., Scott, A. E., Yang, Z., & Kirby, J. R. (2007). Chemosensory pathways, motility and development in myxococcus xanthus. *Nature reviews. Microbiology*, 5, 862-872.10.1038/nrmicro1770

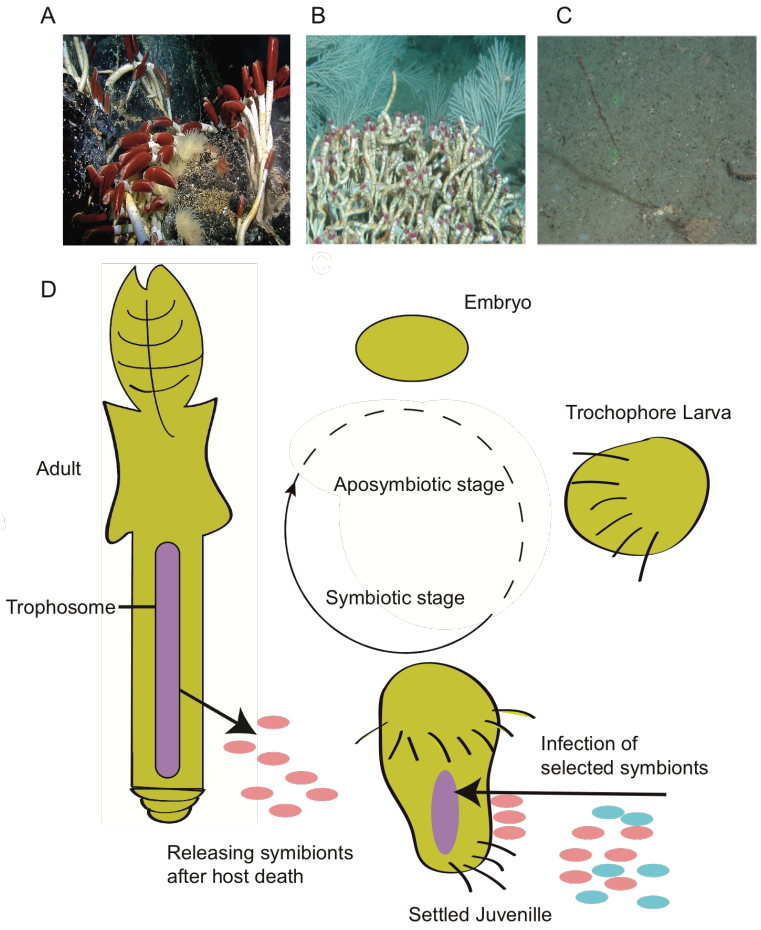


Figure 1 | Major siboglinid lineages and life cycles associated with horizontally transmitted symbionts. A, Giant tubeworm *Riftia* growing in hydrothermal vent. B, *Lamellibrachia* growing near a hydrocarbon seep. C, Frenulata species growing in deep-sea muddy habitats. D, The different life stages of siboglinids accociated with horizontally transmitted symbionts. The embryo and larval stage are aposymbiotic. Symbiotns infect the settled larval skin, and then migrate to mesoderm that later will develop into trophosome. Environmental bacteria are shown in purple

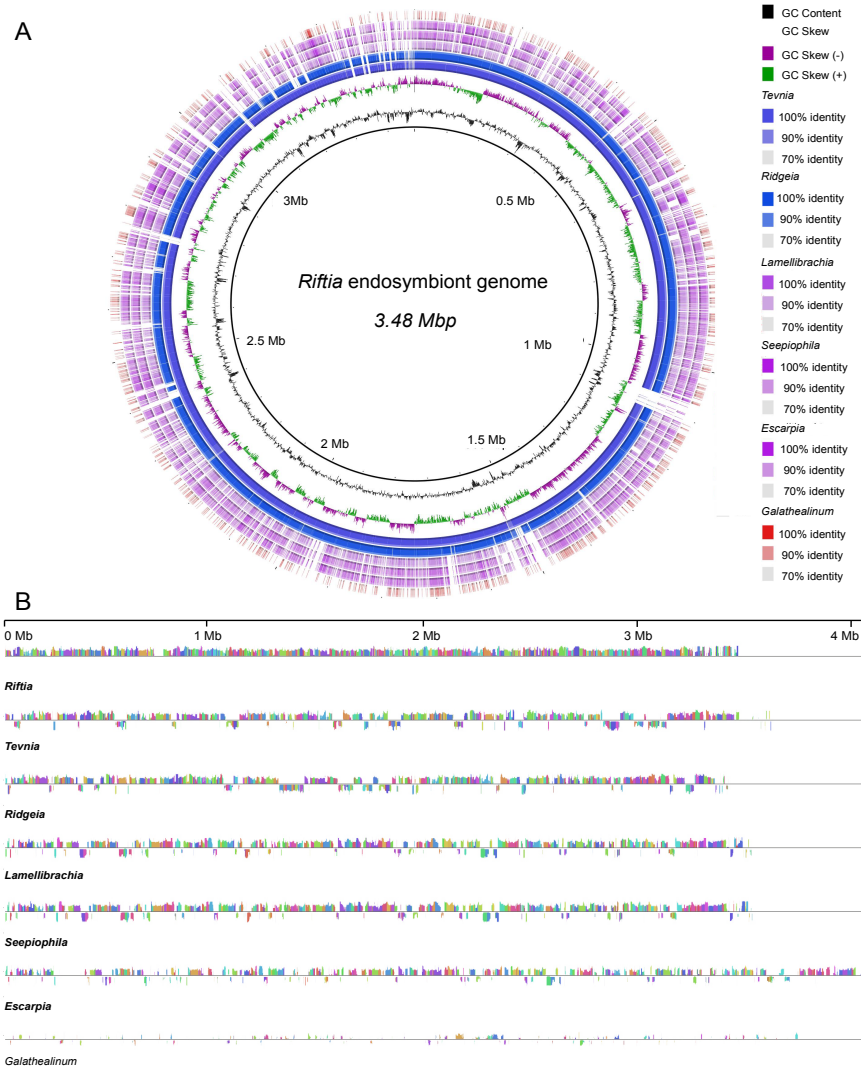


Figure 2 | Whole genome comparisons of sulfur-oxidizing symbionts from siboglinids. A, Genomic map of siboglinid symbiont genomes. The inner circle represents the reference sequence, *Riftia* symbiont. B, Synteny across sequenced symbiont genomes. The plot shows an alignment created using progressiveMauve after each draft genome was ordered against the reference genome of *Riftia* symbionts.

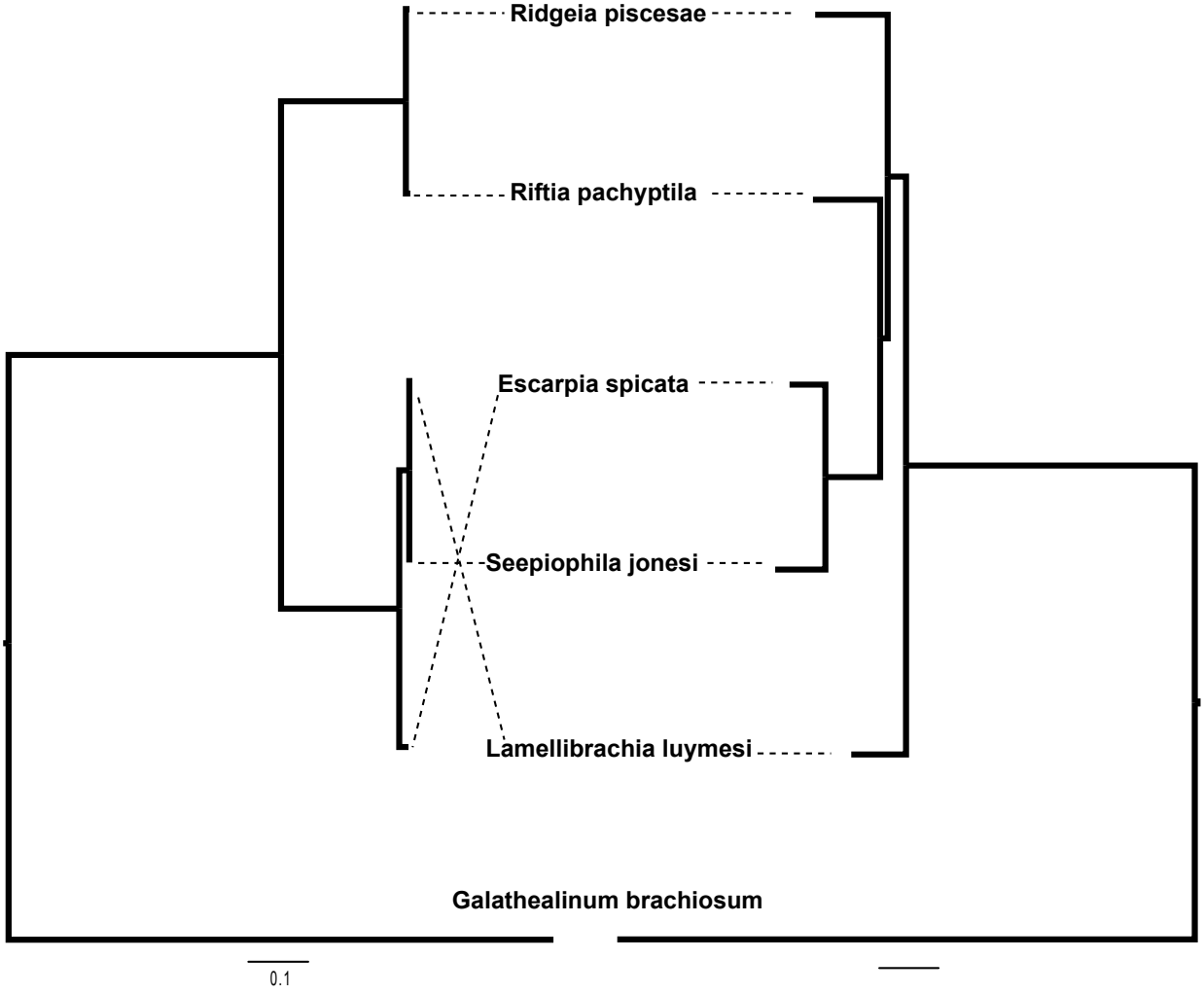


Figure 3 | Cophylogeny of siboglinid speies hosts and bacterial symbionts. All nodes are 100 bootstrp values.

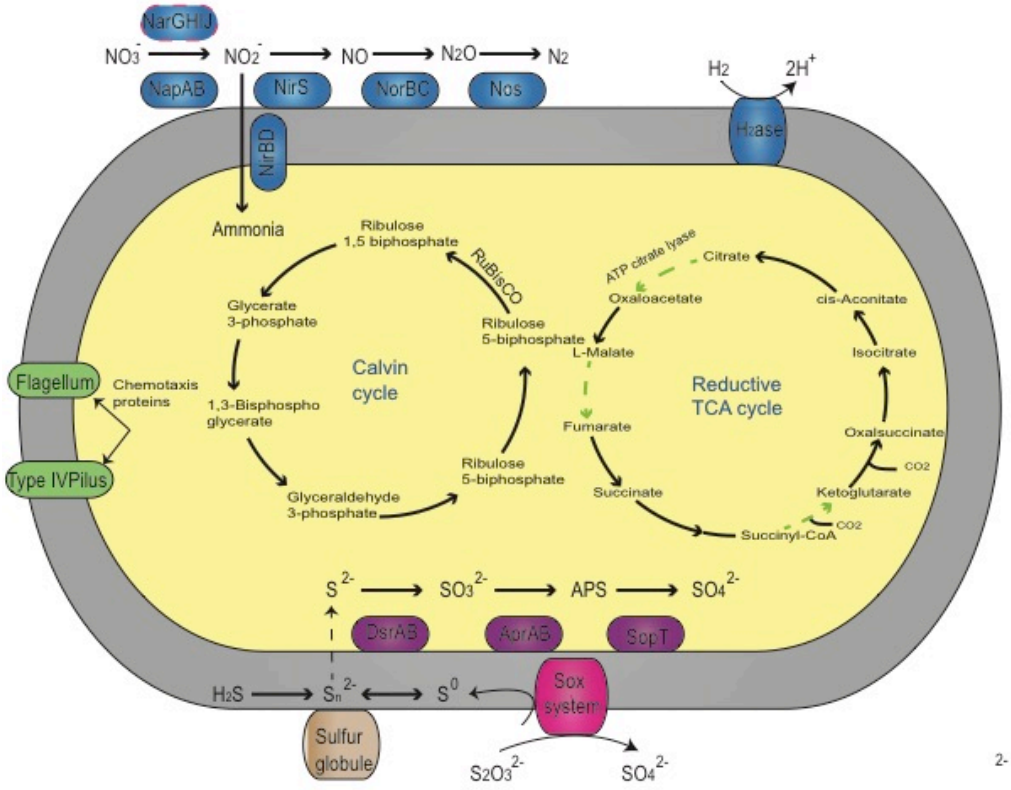


Figure 4 | Overview of the major cellular features and central metabolism in the deep-sea sulfur-oxidizing siboglinid symbiotns. Pathways for which no predictable enzymes were found in frenulate species *Galatehalinum* and seep-dwelling *Escarpia* symbiont genome are shown in green and red dash lines, respectively. Numbers of transport machineries are shown for both strains. The KEGG database was used for the reconstruction of metabolic pathways. H2ase, hydrogenase; Nap, periplasmic nitrate reductase; Nir, cytochrome nitrite reductase; Nor, nitric oxide reductase; Nos, nitrous oxide reductase.

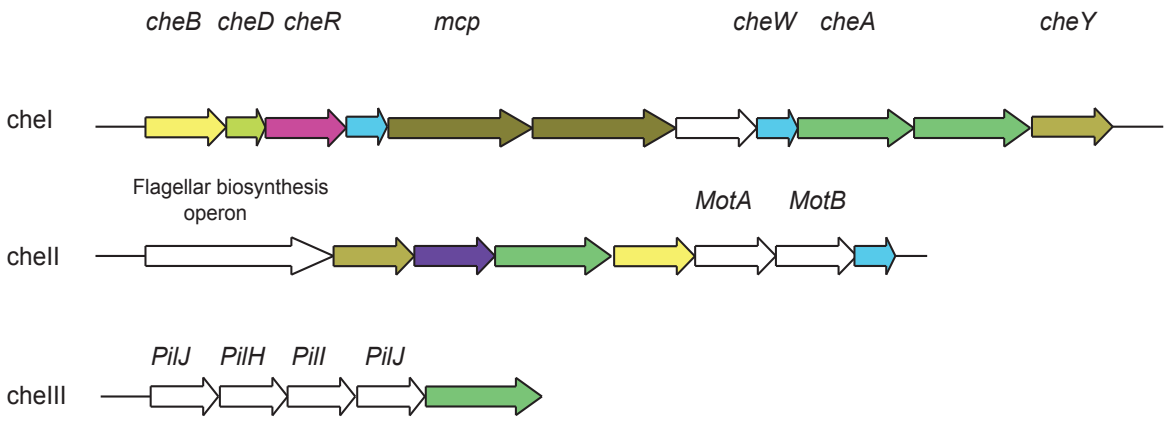


Figure 5 | Multiple chemosensory gene clusters in siboglind symbionts. Chemotaxis pathways in were identified by computer searches (BLAST) for genes that encode CheA homologues and other chemotaxis proteins.

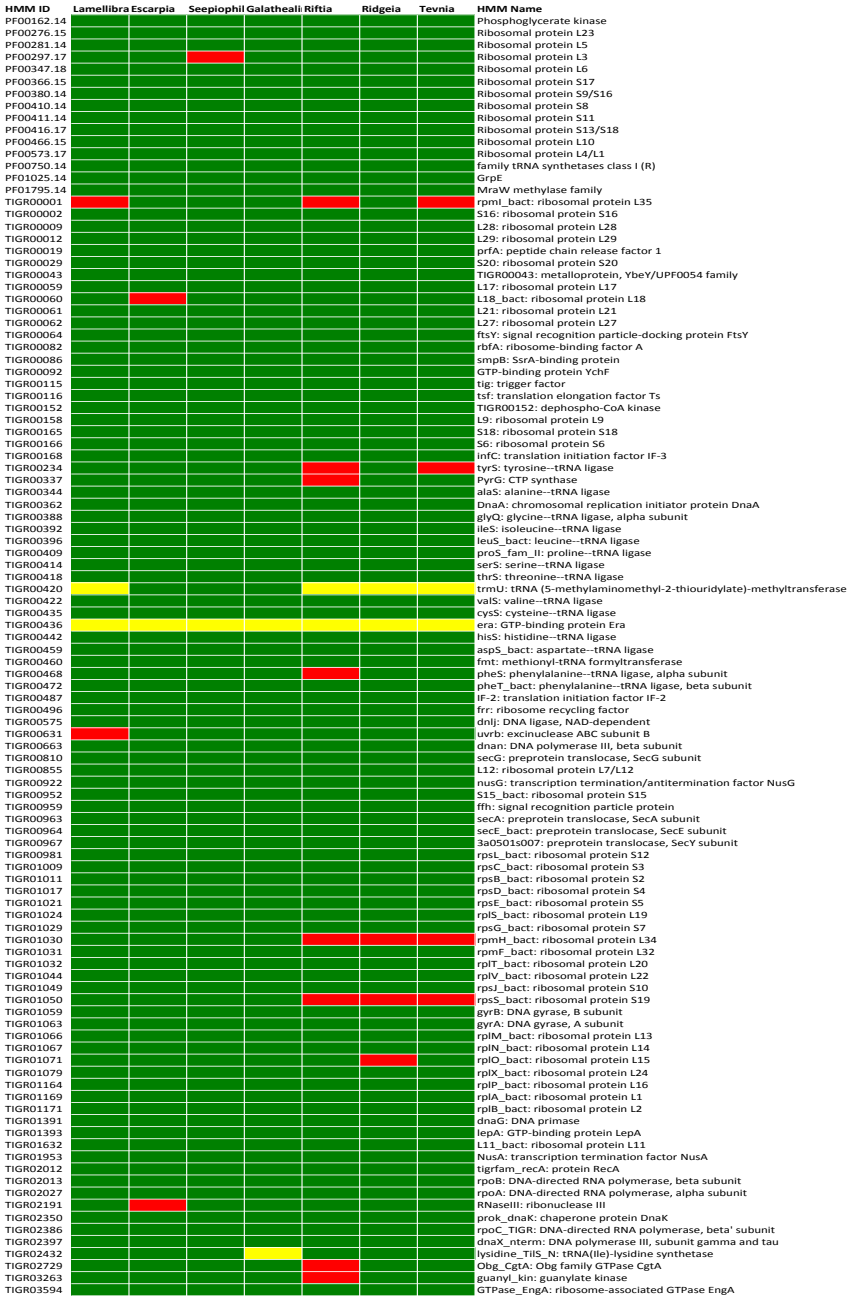


Figure 6. Completeness check based on 106 bacterial universal single-copy genes.

Figure 7. Phylogenetic analysis based on 16S rRNA genes of siboglinid symbionts.

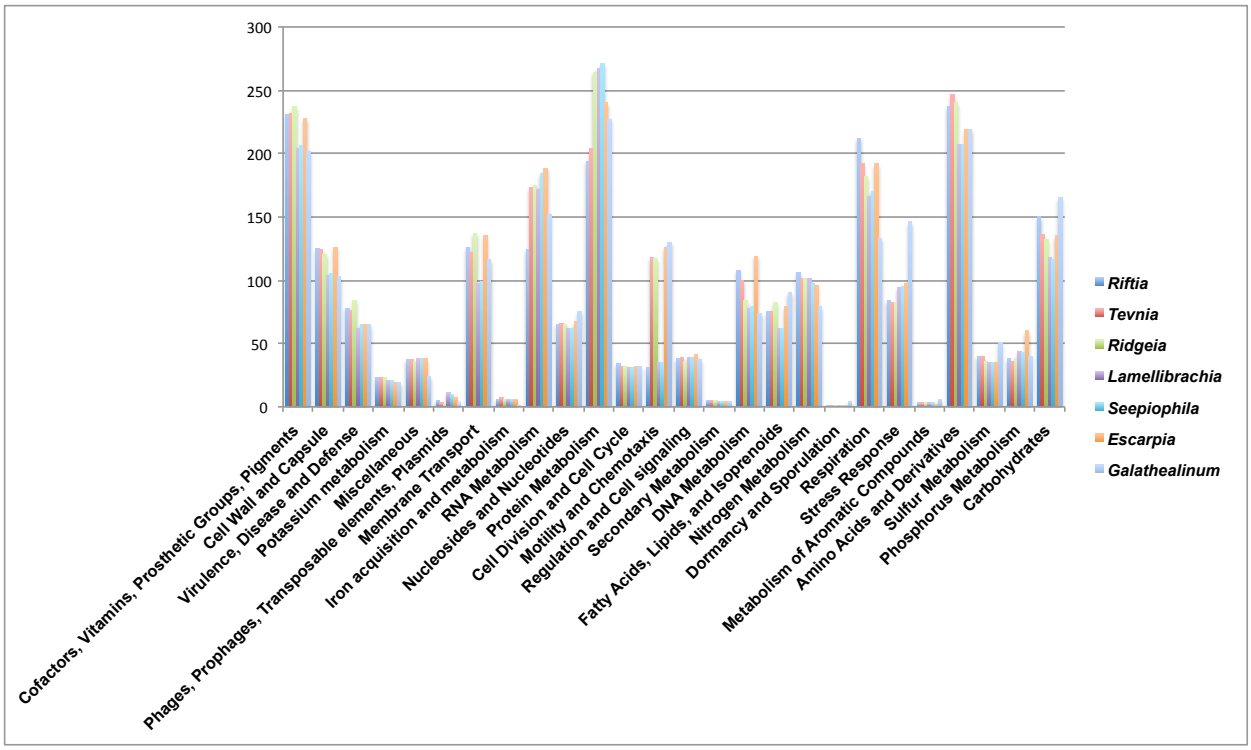


Figure 8. Gene ontology comparison of siboglinid symbiont genomes.

Table 1 Overview of the siboglinid symbiont assemblies.

	Vent-living vestimentiferans			Seep-dwelling vestimentifernas			Frenulata
Features	*Riftia*	*Ridgeia*	*Tevnia*	*Escarpia*	*Lamlllibrachia*	*Seepiophila*	*Galathealinum*
Genome size (Mb)	3.48	3.44	3.64	4.01	3.54	3.55	3.77
N50 (Kb)	29.6	83.9	92.7	313.6	20.5	20.5	486.2
Coverage (folds)	25	180	15	121	120	108	75.6
% G+C	58.8	58.9	58.2	54.2	54.3	54.3	38.9
No. of contigs	197	97	184	23	345	337	19
No. of predicted genes	3341	3188	3566	3698	3552	3542	3497
No. of RNA	45	50	47	48	48	48	43

Table 2

Completeness check

Bin ID	Marker sets	Present	Duplicates	Missing	Completeness%	Contamination%
Riftia	577	491	7	80	87.81	1.74
Ridgeia	577	564	4	10	98.41	0.7
Tevnia	577	540	5	32	93.85	1.05
Seepiophila	577	562	5	15	97.64	1.74
Lamellibrachia	577	551	6	21	96.19	2.09
Escarpia	577	566	4	7	99.18	0.7
Galathealinum	544	535	5	4	99.44	1.23

Chapter 5. Using genomic tools to understand host-symbiont evolution of wood-boring bivalves (Mollusca, Xylophagaidae)

5.1 Abstract

Wood falls occur widely in the deep-sea. They support specialized communities for limited periods of time, contribute fundamentally to biodiversity, and contain evolutionary novelties. Wood-boring Xylophagidae species dominate deep-sea wood fall communities and play a key role in facilitating wood decomposition. Despite evolutionary and ecological importance of these obligate wood bores, very little is known of their distribution, dispersal potential, and recruitment sources. For this study, wood landers were deployed at two depths (1500 and 3000m), 250 km apart, in the NE Pacific and SW Atlantic basins. Because so little is known about xylophagaid bivalves, this chapter sought to address 3 basic evolutionary issues relating to the group. 1) Phylogenetic analysis based on complete mitochondrial genome suggests that *Xylophaga* is a paraphyletic clade, and xylophagaid evolutionary relationships do not corresponding to isolation by depth in the deep-sea. 2) Connectivity and recruitment of two xylophagaid species from two landers in NE Pacific and two landers in SW Atlantic where evaluated by 2b-RAD sequencing approach. Our findings show that there is no population structure identified across 500 km spacing in each species, suggesting that individuals from same xylophagaid species living in bathyal site (~1500m) are most likely from the same gene pool in both basins. 3) Lastly, metagenomic analysis from *Xylophaga* gill tissue shows their symbiont genome is closely related

to *Teridinibacter* species isolated from the closely related shallow water shipworm family Teredinidae. Multiple genes dedicated to processing complex polysaccharides associated with wood falls were identified in partial symbiont genome, which potentially indicating a similar functional role in these endosymbionts from both shallow and deep wood-boring bivalves.

5.2 Introduction

Deep-sea xylophagaids are dominant fauna in organic-rich wood falls that support diverse, yet specialize, communities that contain adaptive radiations and evolutionary novelties (Turner, 1973). Wood-fall habitats are comparable with whale falls due to their ephemeral distribution in the deep-sea (Distel et al., 2000). Xylophagids are obligate wood-boring bivalves belonging to the bivalve family Xylophaginae. They contain specialized shells that carries denticles on their anterior beak to bore a hole in the wood, and then utilize their U-shaped diverticulum to collect wood particles (Purchon 1941; Voight 2015) (Fig. 1). Similar to vent and seep living mussels such as *Bathymodiolus*, endosymbionts have been observed in their gill region (Distel and Roberts 1997). To date, approximately 60 species have been described (www.marinespecies.org) within three genera: *Xylophaga*, *Xylophalas* and *Xyloredo* (Turner 2002). Despite their great ecological, evolutionary and potential industrial importance, the distribution, dispersal potential, recruitment patterns, endosymbiont community, evolutionary relationships and other biological features of this clade are still largely unknown (Voight 2007; Romano et al., 2014).

Wood-boring bivalves are thought to be widely distributed in the deep-sea, extending from shallow depths to the hadal zone, and from polar to tropical regions (Stoeckle, 2006). However, most species of shipworms have only been best quantitatively sampled in the North Pacific, fewer than 10 sampling localities that include xylophagids have been reported in southern hemisphere which contains 60% of the world ocean (Stoeckle, 2006; Voight et al., 2009). Despite a patchy deep-sea distribution, wood falls are colonized at a surprisingly high speed by wood-boring bivalves (Turner 1973; Voight 2008; Romano et al., 2013). High densities of individuals of *Xylophaga* species were recorded in wood falls after 3 month deployment in Northwest Atlantic at ~ 1800m (Turner 1973). The current hypothesis of xylophagaid dispersal is that their larvae are abundant in the water column, transported by bottom currents and then colonize at where wood falls are present over long distances. They were thought have the ability to delay metamorphosis in the absence of wood sources and to recruit rapidly following environmental cues (Gaudron 2016). Nevertheless, the mode of recruitment sources and patterns of these wood bores have not yet been characterized.

Interest in xylophagaid symbioses has been driven partly by their potential role in marine carbon cycles and a source of novel enzymes for industry (Distel et al., 2011). The ability of both xylophagaids and their close relatives,teredinids, to feed on wood is thought to rely on intracellular bacterial endosymbionts contained within their gill region (Distel 2003; Distel et al., 2011, 2017). Although many wood-boring bivalves are thought to host gill-associated symbionts, only the Gammaproteobacteria (e.g. *Teredinibacter turnerae*) found in shallow-water teredinid

shipworms have been well categorized metabolically with the capacity of cellulolysis of wood and nitrogen fixation (Distel et al., 2002; O'Connor et al., 2014). In xylophagaids, symbionts have been identified in the gills of *Xylophga atlantica* and *X. wahsingtona* (Distel and Roberts 1997), but not yet been cultured and characterized.

To further explore evolution of these unusual deep-sea xylophagaid shipworms, my colleagues and I collected specimens from artificial Bone and Wood landers (BOWLS) in Northeast (NE) Pacific and Southwest (SW) Atlantic (Fig. 2) to undertook (i) phylogenetic analyses of 13 mitochondrial genomes to explore evolutionary relationships, (ii) 2b-RAD sequencing of two xylophagaid species to evaluate their recruitment sources and patterns, (iii) metagenomics to better characterize their gill endosymbiont community. Here we described some of these efforts.5.3 Materials and Methods

5.3.1 Experimental design

To collect wood-boring bivalves from wood-fall habitats in the deep sea, my colleague Drs. Craig Smith (University of Hawaii) and Paulo Sumida (University of São Paulo) designed free-vehicle landers to deploy standardized wood substrates at the deep-sea floor (Fig. 2C). Six replicate experimental landers were deployed in a stratified design at two seafloor depths (~ 1500 m and ~3000 m) on the NE Pacific margin (Fig. 2A). A parallel set of experiments, with a similar experimental design of six landers, was conducted with Brazilian collaborators, on the

SW Atlantic Margin (Fig. 2B). All the landers were recovered after 15 months or 19 months deployment, from NE Pacific and SW Atlantic, respectively.

5.3.2 Specimen collection and mitochondrial genome sequencing

Specimen information used for mitogenomic analysis is shown in Table 1. All were preserved in 95% non-denatured ethanol following collection. DNA was extracted using the DNeasy Blood & Tissue Kit (Qiagen) following manufacture's protocols. Sequencing of genomic DNA was performed by The Genomic Services Lab at the Hudson Alpha Institute in Huntsville, Alabama on an Illumina HiSeq 2500 platform (San Diego, California) using 2 x 150 paired-end v4 chemistry. Paired-end reads were assembled de novo using IDBA-UD (Peng et al., 2012) with default settings. Contigs of interest where identified by using blast with previously published bivalve mtDNA genomes against the assembled genomic data. Annotation of the 13 protein-coding genes, 2 ribosomal RNAs and tRNAs was conducted initially with MITOS web server (Bernt et al., 2013), followed by manual genome annotation of start and stop positions of each gene using Artemis (Carver et al., 2008).

Fourteen Operational Taxonomic Units (OTUs) were included in phylogenetic analyses. *Arctica islandica* was selected as outgroup based on data availability of this poorly studied group. Two data sets were constructed – one being amino acid (AA) and the other being nucleotide (NUC) sequences. Nucleotide sequences were converted into amino acids using the standard invertebrate mitochondrial translation code implemented in Mega 5.2 (Tamura et al.,

2011). For amino acid and nucleotide data, each gene was treated as an individual Orthology Group (OG) that was aligned in MUSCLE 3.8.31 (Edgar, 2004), followed by manual correction. All OGs were trimmed using the default settings in Gblocks 0.91b (Talavera and Castresana, 2007) to remove ambiguously aligned regions. OGs were then concatenated into final datasets for phylogenetic analyses using FASconCAT (Kück and Meusemann, 2010). The NUC dataset consisted of nucleotide sequences of the 13 protein-coding and the 2 ribosomal RNA genes while the AA dataset included the amino acids sequences of the 13 protein-coding genes only. Prior to ML analyses, PartitionFinderV2 (Lanfear et al., 2016) was used to evaluate best-fit partition schemes and associated best-fit substitution models for both datasets. Topological robustness for the ML analysis was evaluated with 100 replicates of fast-bootstrapping.

5.3.3 Specimen collection and 2b-RAD sequencing

To evaluate recruitment patterns of wood-boring bivalves, 58 individuals of *X. oregona* and 28 individuals of *Xyloredo sp1* were collected from two landers each in NE Pacific and SW Atlantic, respectively (Table 3). Due to logistical issues, other collected Xylophaga species in this study were not included either because of small sample size or only colonized in only one lander. DNA was extracted using the DNeasy Blood & Tissue Kit (Qiagen) following manufacture's protocols and 2b-RAD library preparation for high-throughput sequencing followed Wang et al., 2012 with the restriction enzyme *AlfI*. A 1/16 genome reduction scheme was used on samples to target roughly 2,000 SNPs based on estimated genome size (best guess

was based on *Platyodon cancellatus* (C-value = 1.40) (Hinegarder, 1973) and pooling strategy. Sequencing of dual barcoded 2b-RAD libraries were performed by The Genomic Services Lab at the Hudson Alpha Institute in Huntsville, Alabama on an Illumina HiSeq 2500 platform (San Diego, California) using 50 bp single-end v4 chemistry.

The non-reference SNP calling pipeline 2bRAD2.0 (Eli Meyer's script: http://people.oregonstate.edu/~meyere/2bRAD_analysis2.0.html) was used for SNP discovery and genotyping for samples from each species. Briefly, data was filtered to only keep RAD tags with a minimum coverage of 25X. For all SNPs, loci scored as homozygotes were chosen with a maximum variance of 1%, those scored as heterozygotes had a minimum of 25% variance, and others were excluded. Individuals or loci with more than 30% of missing data were also eliminated. Details of the number of loci retained are provided in Table 4. Loci potentially under balancing or divergent selection were removed using the BAYESCAN v.2.1 (Foll and Gaggiotti 2008). Loci with a posterior probability over 0.95 were considered as outliers (100,000 generations).

A Bayesian clustering method – STRUCTURE (Falush et al., 2003) was used to infer fine-scale population structure of individuals from each species. Structure can infer the best value of *K*, the number of putative populations from the dataset. With STRUCTURE, I used 10,000 burn-in generations followed by 50,000 additional MCMC repetitions with 5 replicates at each potential *K* (1-10). An admixture model was also applied with correlated allele frequencies. The optimal *K* will be chosen using Structure Harvester (Earl and VonHoldt 2012). Resulting data

was summarized with CLUMPP (Jakobsson and Rosenberg 2007) and visualized using DISTRUCT (Rosenberg 2004). We then performed a Discriminant Analysis of Principal Components (DAPC) in the R package *adegenet* (Jombart et al., 2010), without prior information on population structures. The optimal number of populations was evaluated using the BIC criteria. Hierarchical F_{st} test and genetic diversity across all the SNP loci (e.g. level of heterozygosity) was calculated in R package *adegenet*.

5.3.4 Metagenomics of bacterial community of gill region

Cross-section tissue samples were excised from gill tissue of two specimens of *X. oregona* 1500m in NE Pacific, respectively. Genomic DNA was extracted using the DNeasy Blood & Tissue Kit (Qiagen) following manufacture's protocols. Sequencing of metagenomic libraries were performed by The Genomic Services Lab at the Hudson Alpha Institute in Huntsville, Alabama on an Illumina HiSeq 2500 platform (San Diego, California) using 125 bp paired-end v4 chemistry. Illumina paired-end fastq reads were trimmed using Sickle (Joshi and Fass 2011) with default parameters, and then merged as FASTA files using the Perl script "fq2fq" in IDBA_ud package (Peng et al., 2012). Merged FASTA files were assembled using IDBA_ud with standard parameters.

5.4 Results and Discussion

5.4.1 Mitogenomics

Results from high-throughput sequencing and genome assembly for all the 13 mitochondrial genomes from 7 xylophagaid species (5 *Xylophaga* species, 2 *Xyloredo* speceis) are presented in Table 2. All xylophagaid mt genomes are ~16 kbp in length, varying in size from 14,450 bp (*X. hetorosiphoni*) to 18,624 bp (*X. washingtona*) (Table 2). Interestingly, *X. heterosiphoni* and *X. zirenbergi* collected from ~3,000m exhibited smaller mitochondrial genome size with slightly higher in GC content than other Xylophaga (Table 2). Moreover, species collected from SW Atlantic generally showed a lower GC content of mt genomes than ones from NE Pacific (Table 2). For each mitochondrial genome sequenced herein, the genome was composed of 36 genes. Similar to many other molluscs (Boore, 1999), the ATP8 gene was missing in all xylophagaid mt genomes. Substantially gene rearrangements were observed in xylophagaid mt genomes (Fig. 4), which was also common in other mollusk lineages (e.g. *Mytilus*).

The AA and NUC datasets contained 3,483 and 12,962 position after trimming using Gblocks, respectively. The ML analyses of the two concatenated datasets yielded congruent tree topologies with high bootstrap support values (Fig. 3). In terms of higher-level relationships, *Xyloredo* lineage was recovered as monophyletic clade, and it was nested within *Xylophaga* with strong support (100/94), indicating *Xylophaga* is paraphyletic. Among xylophagaids, partially calcareous tubes surround the siphons only in the genus *Xyloredo* (Haga and Kase 2008),

whereas periostracal cones cover the siphons that does not become calcified were identified in a few *Xylophaga* species (Voight 2015). However, morphological characters used to distinguish between *Xylophaga* and *Xyloredo* lineage was usually vague (personal communication with Janet Voight), making taxonomy work of this group problematic. For example, calcareous tubes surround the siphons are thought only in *Xyloredo*, although periostracal cones have also been found in several *Xylophaga* species (Voight 2015). Within *Xylophaga*, shallow water *X. washingtiona* was close related to *X. oregona* collected in 1500m at NE Pacific, and this clade was sister to *Xyloredo* species collected in SW Atlantic. *Xylophaga zirenbergi* was sister to a clade comprising *Xylophaga* species collected in 1500m at SW Atlantic.

Isolation by depth, rather than distance is thought to be the most important isolation barrier in the deep-sea. For example, the bivalve *Deminucula atacellana* had more divergence within an ocean basin at different depths than across thousands of kilometers at the same depth (Zardus et al., 2006). Such pattern has also been identified in other organisms and habitats (e.g. brittle stars on seamounts – Cho and Shank 2010; corals in the Pacific ocean – Miller et al., 2011; *Lamellibrachia* tubeworms in Gulf of Mexico - Cowart et al., 2014). As hypothesized by Zardus et al., 2006, isolation by distance was most occurred across different ocean basins. However, isolation model of depths and distances are only partly supported in xylophagaid evolutionary relationships in this study. Xylophagaid species were not overlapped across different depth zones and ocean basins (Fig. 3). However, although *Xylophaga heterosiphoni* and *Xylophaga zirenbergi* were collected from the same lander in 3000m at NE Pacific, they were not sister to

each other (Fig. 3). *Xylophaga zirenbergi* from NE Pacific is even most close related to *Xylophaga* species from SW Atlantic at bathyal zone with strong nodal support (BS = 100), rather than sister to any species collected from NE Pacific. In summary, although more comprehensive sampling is needed, the phylogenetic analysis suggest that genus *Xylophaga* is a paraphyletic clade, and evolutionary relationships of Xylophagaids shows a slightly mosaic pattern instead of corresponding to isolation by depth or distance models.

5.4.2 2b-RAD sequencing to evaluate xylphagaid recruitment sources.

Following quality filtering, SNP calling and removing loci that potentially under selection, 1,910 and 2,100 loci were recovered for X. oregona and Xyloredo sp1 for the subsequent analysis, respectively. Under calculations of Delta K from STRUCTURE HAVESTER for the filtered datasets, K of 1 had the highest maximum-likelihood scores for both species (Fig. 6). DAPC analysis yielded the same interpretation of number of genetic clusters. Although specimens were collected across ~ 500 kilometers at the same depth, both results strongly suggested a well-mixing population with no population structures were observed in both species. The level of expected against observed heterozygosity was further analyzed and plotted in Fig 5. The results suggested a significant higher value of expected heterozygosity from individuals in both species (X. oregona: Bartlett's K-squared = 1308.1, df = 1, p-value < 2.2e-16; Xyloredo sp1: Bartlett's K-squared = 32.841, df = 1, p-value = 1e-08). It was generally assumed that populations with reduced heterozygosity are more inbred than related populations with greater heterozygosity (Markert et al., 2004). Typically, a lower observed heterozygosity than expected

values could be attributed the discrepancy to forces such as inbreeding (Castric et al., 2002). If heterozygosity is higher than expected, which might suspect an isolate-breaking effect (individuals are colonized the wood fall from different cohorts).

The successful colonization of an ephemeral and patchily distributed wood fall habitats largely depends on life-history traits (Tyler et al. 2009). Although the reproductive and dispersal mode of xylophgaid species are still poorly known so far, some species are brooders although Haga and Kase (2013) suggested that some or all *Xylophaga* species could instead be carrying dwarf males. Our results here support the hypothesis brought by Tuner 1973, xylophagaid larvae might be abundant in the deep-sea, transported by bottom currents and guided by an ability to detect wood over considerable distances (500 km in this case at both basins), potentially with the capacity to delay metamorphosis and to recruit rapidly on the wood falls. In summary, our results strongly suggest that the recruitment sources of same xylophagiad species at bathyal sites (~1500m) are mostly like from the same gene pool. Their larvae might have extraordinary power for dispersal as an adaptation to ephemeral wood fall habitats in the deep-sea.

5.4.3 Metagenomics to characterize *Xylophaga* gill symbiont communities.

To better characterize the gill endosymbiont community, we sequenced metagenomes from the gills of *X. oregona*, yielded a total of 142 million reads after quality filtering. Of these, 10.59% were derived form host; the remainder was bacterial (Fig. 7A). The majority of bacterial contigs (24%) originated form a genome closely related to *Teredinibacter turnerae* (Fig. 7B).

Teredinibacter turnerae is a cellulolytic, dinitrogen-fixing bacterium associated within the gills of shallow-water shipworm teredinids (Distel et al., 2002). The 16s rRNA gene was isolated from *Xylophaga* metagenomic dataset, and phylogenetic analysis place this dominant symbiont ribotype within a clade with other previous sequenced teredinid symbionts (Fig. 8). This clade is placed within the family Alteromonadaceae, a group known as obligate aerobic heterotrophs (Distel et al., 2017). The remaining contigs might represent other bacteria associated with bivalves' gill surface area due to low representative.

Although we were not be able to isolate the complete endosymbiont genome from metagenomic data due to potential lack of sequencing coverage, ~30 genes involving in processing complex polysaccharides were identified. These results are consistent with previous sequenced *T. turnerae* genome, suggesting that *Xylophaga* symbiont might also produce enzymes that may assist host in degrading carbohydrate components from wood falls (e.g. cellulose, hemicellulose) (Yang et al., 2009). Moreover, teredinid symbiont is thought to have ability to fix nitrogen to supplement the host's nitrogen deficient diet of wood (Alison et al., 2000; Yang et al., 2009). However, only one nitrogen fixation gene (nifA) was identified in the symbiont genome. The lack of nitrogen fixation pathway in *Xylophga* symbiont might be due to incomplete metagenome assembly.

5.5 References

Purchon (1941) On the Biology and relationships of the Lamellibrach *Xylophaga dorsalis*

(Turton). Journal of the Marine Biological Association of the United Kingdom 25: 1–39.

Distel DL, Roberts SJ (1997) Bacterial endosymbionts in the gills of the deep-sea wood-boring bivalves *Xylophaga atlantica* and *Xylophaga washingtona*. Biological Bulletin (Woods Hole) 192: 253–261.

Voight JR (2007) Experimental deep-sea deployments reveal diverse Northeast Pacific wood-boring bivalves of Xylophagainae (Myoida: Pholadidae). J Molluscan Stud 73: 377–391.

Stoeckle, M., 2006. Species richness of deep-sea wood-boring clams (subfamily Xylophagainae) from the northeast Pacific, University of Victoria (Electronic Theses and Dissertations), pp. 188.

Voight, J.R., 2009. Diversity and reproduction of near-shore vs offshore wood-boring bivalves (Pholadidae: Xylophagainae) of the deep eastern Pacific ocean, with three new species. Journal of Molluscan Studies 75(2), 167-174.

Romano C, Voight JR, Pérez-Portela R, Martin D. Morphological and Genetic Diversity of the Wood-Boring *Xylophaga* (Mollusca, Bivalvia): New Species and Records from Deep-Sea Iberian Canyons. PLoS ONE. 2014;9:e102887.

Haga T, Kase T. Progenetic dwarf males in the deep-sea wood-boring genus *Xylophaga* (Bivalvia: Pholadoidea). J Molluscan Stud. 2013;79:90–4.

HAGA, T. & KASE, T. 2008. Redescription of the deep-sea wood borer *Neoxylophaga teramachii* Taki & Habe, 1950 and its assignment to the genus Xyloredo (Bivalvia: Myoida: Pholadoidea) with comments on fossil Pholadoidea. Veliger, 50: 107–119.

Peng, Yu, Henry CM Leung, Siu-Ming Yiu, and Francis YL Chin. "IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth." Bioinformatics 28, no. 11 (2012): 1420-1428.

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., Pütz, J., Middendorf, M. and Stadler, P.F., 2013. MITOS: Improved de novo metazoan mitochondrial genome annotation. Molecular phylogenetics and evolution, 69(2), pp.313-319.

Carver, T., Berriman, M., Tivey, A., Patel, C., Böhme, U., Barrell, B.G., Parkhill, J. and Rajandream, M.A., 2008. Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. Bioinformatics, 24(23), pp.2672-2676.

Lanfear, R., Frandsen, P.B., Wright, A.M., Senfeld, T. and Calcott, B., 2016. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Molecular Biology and Evolution, p.msw260.

Hinegardner, R. (1974a). Cellular DNA content of the Mollusca. Comparative Biochemistry and Physiology 47A: 447-460.

Lischer HEL, Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. Bioinformatics, 28, 298–299.

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genetics, 11, 94.

Peng Y, Leung HC, Yiu SM, Chin FY (2012) IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28:1420–1428

Joshi NA, Fass JN (2011) Sickle: A sliding-window, adaptive, quality-based trimming tool for FASTQ files (version 1.33).

Cowart, D.A., Fisher, C.R., Halanych, K.M., & Schaeffer, S.W. (2014). Depth-dependent gene flow in Gulf of Mexico cold seep *Lamellibrachia* tubeworms (Annelida, Siboglinidae). Hydrobiologia, 736, 139-154.

Cho, W., Shank, T.M. 2010. Incongruent patterns of genetic connectivity among four ophiuroid species with differing coral host specificity on North Atlantic seamounts. Marine Ecology. 31 (Suppl. 1), 121- 143.

Zardus, J.D., Etter, R.J., Chase, M.R., Rex, M.A., Boyle, E.E., 2006. Bathymetric and geographic population structure in the pan-Atlantic deep-sea bivalve *Deminucula atacellana* (Schenck, 1939). Molecular Ecology 15, 639-651.

Voight JR (2008) Deep-sea wood-boring bivalves of *Xylophaga* (Myoida: Pholadidae) on the continental shelf: a new species described. J Mar Biol Assoc UK 88:1459–1464

Markert JA, Grant PR, Grant BR, Keller LF, Coombs JL, Petren K. Neutral locus heterozygosity, inbreeding, and survival in Darwin's ground finches (*Geospiza fortis* and *G. scandens*). Heredity. 2004;92:306–15.

Castric, V., Bernatchez, L., Belkhir, K., Bonhomme, F., 2002. Heterozygote deficiencies in small lacustrine populations of brook charr Salvelinus Fontinalis Mitchill (Pisces, Salmonidae): a test of alternative hypotheses. Heredity 89, 27–35. doi:10.1038/sj.hdy.6800089

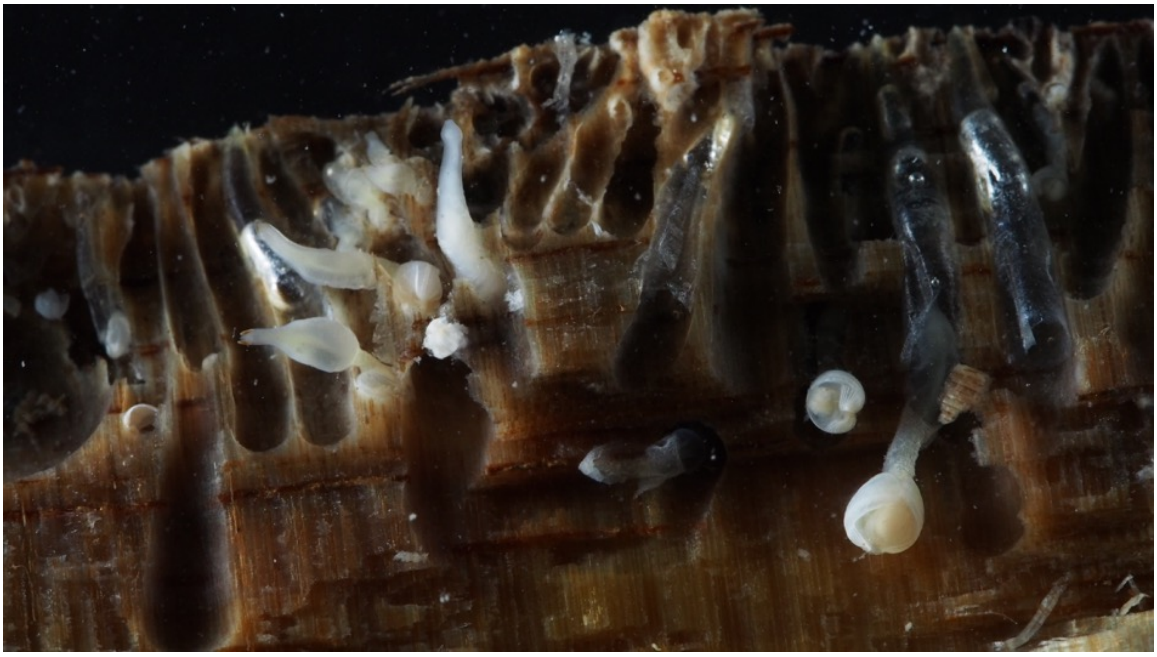


Figure 1. *Xyloredo* individuals colonized at wood falls fro SW Atlantic Basin (Figure is provided with my collaborator Dr. Angelo Bernardino).

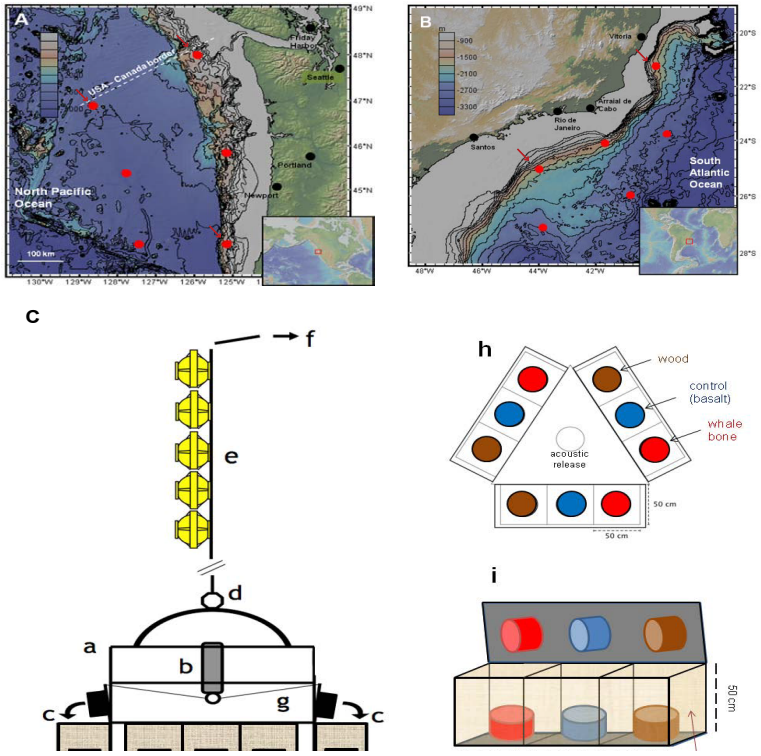


Figure 2. Sampling locality and experimental design in this study. (A) Locality of deployment landers in NE Pacific; (B) Sampling locality in SW Atlantic. On each margin, six replicate landers (red circles) were deployed and collected with approximately ~ 250 km spacing; three BOWL landers at ~1500m and tree landers at ~3000m depths. Red arrows indicated xylophagaid samples were found in these landers for this study. Figures are provided by my collaborator Drs. Craig Smith and Angelo Bernardino.

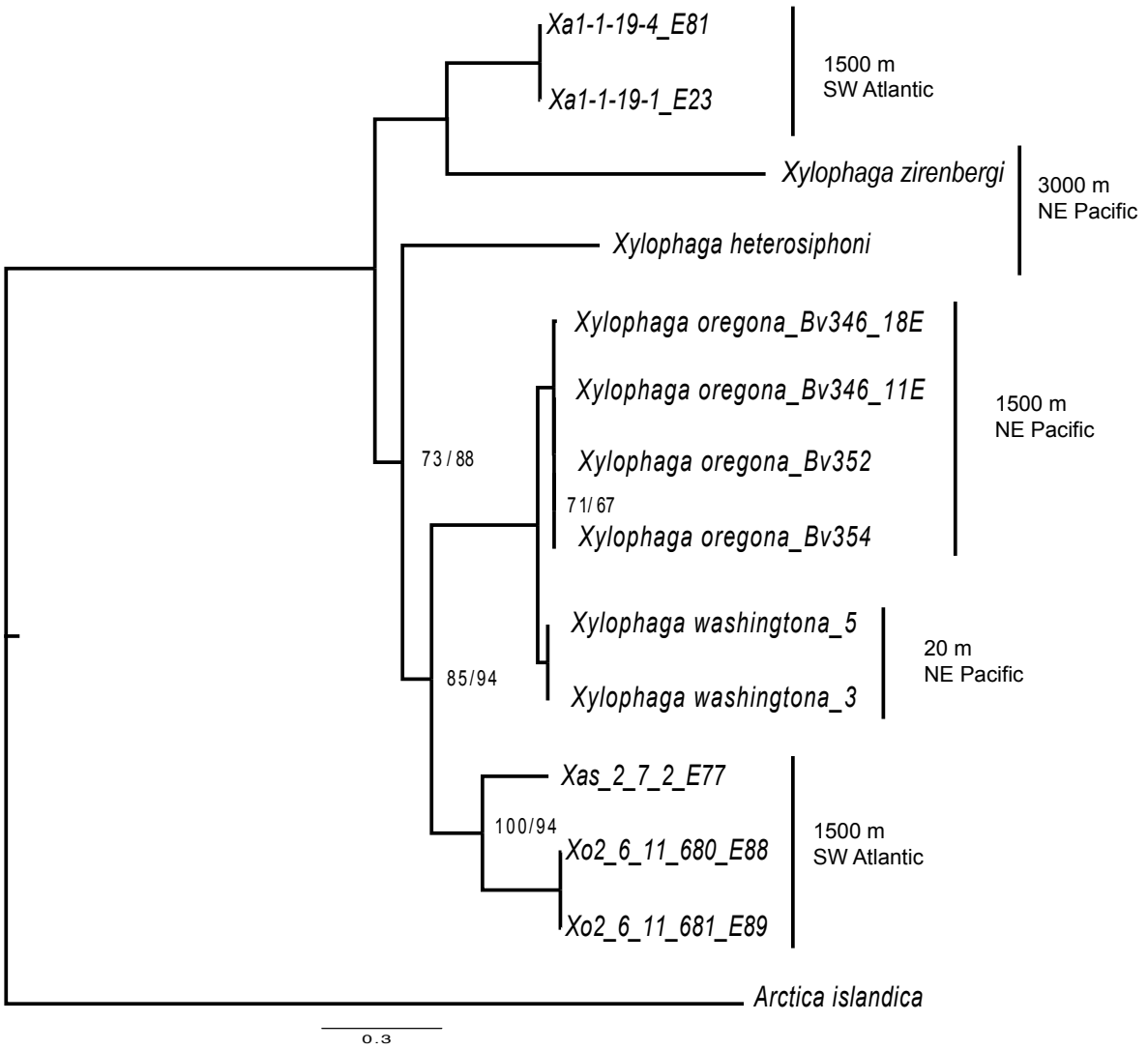


Figure 3. Phylogenetic tree based complete mitochondrial genomes.

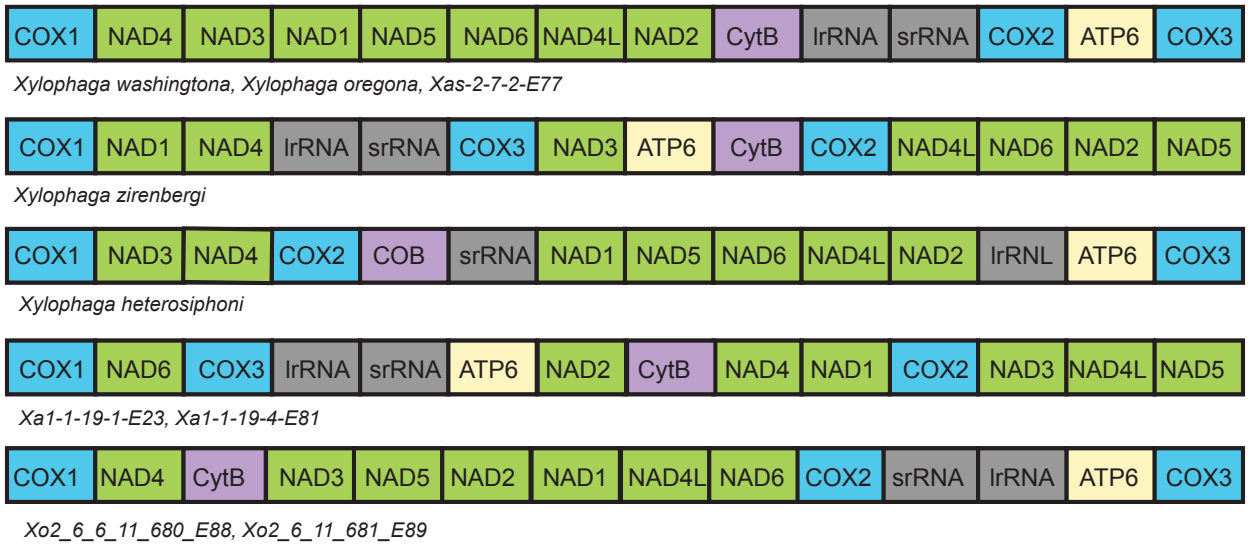

COX1 NAD4 NAD3 NAD1 NAD5 NAD6 NAD4L NAD2 CytB lrRNA srRNA COX2 ATP6 COX3
Xylophaga washingtona, Xylophaga oregona, Xas-2-7-2-E77
COX1 NAD1 NAD4 lrRNA srRNA COX3 NAD3 ATP6 CytB COX2 NAD4L NAD6 NAD2 NAD5
Xylophaga zirenbergi
COX1 NAD3 NAD4 COX2 COB srRNA NAD1 NAD5 NAD6 NAD4L NAD2 lrRNL ATP6 COX3
Xylophaga heterosiphoni
COX1 NAD6 COX3 lrRNA srRNA ATP6 NAD2 CytB NAD4 NAD1 COX2 NAD3 NAD4L NAD5
Xa1-1-19-1-E23, Xa1-1-19-4-E81
COX1 NAD4 CytB NAD3 NAD5 NAD2 NAD1 NAD4L NAD6 COX2 srRNA lrRNA ATP6 COX3
Xo2_6_6_11_680_E88, Xo2_6_11_681_E89

Figure 4. Gene arrangement of sequenced xylophagaid mt genomes

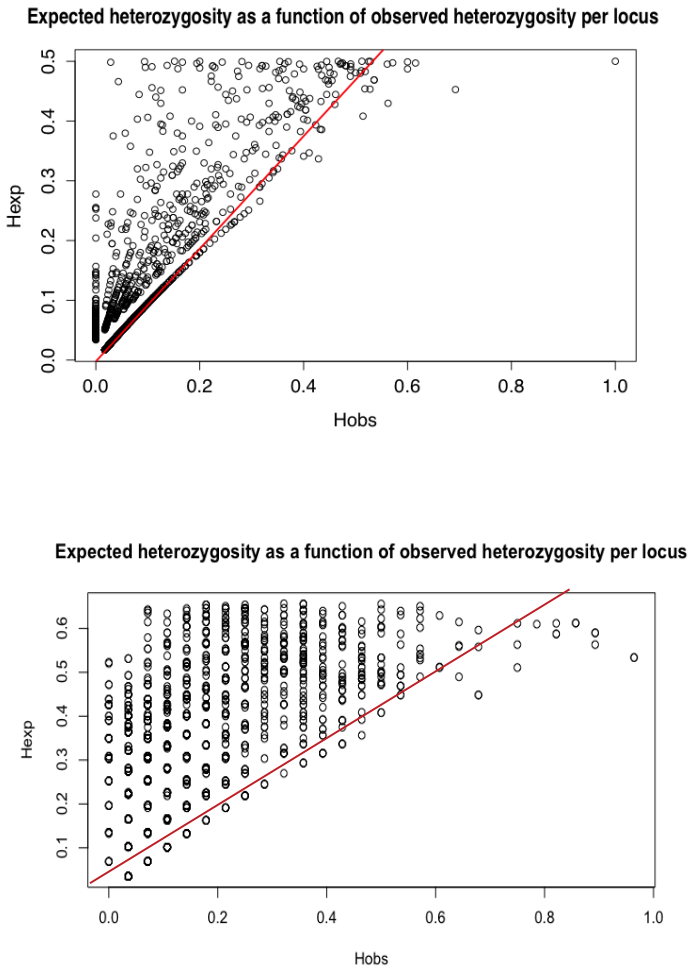


Figure 5. Values of expected heterozygosity was plotted against observed values across all SNP loci from (A) *Xylophaga oregona* and (B) *Xyloredo sp1*. The red line indicated same value of expected and observed heterozygosity.

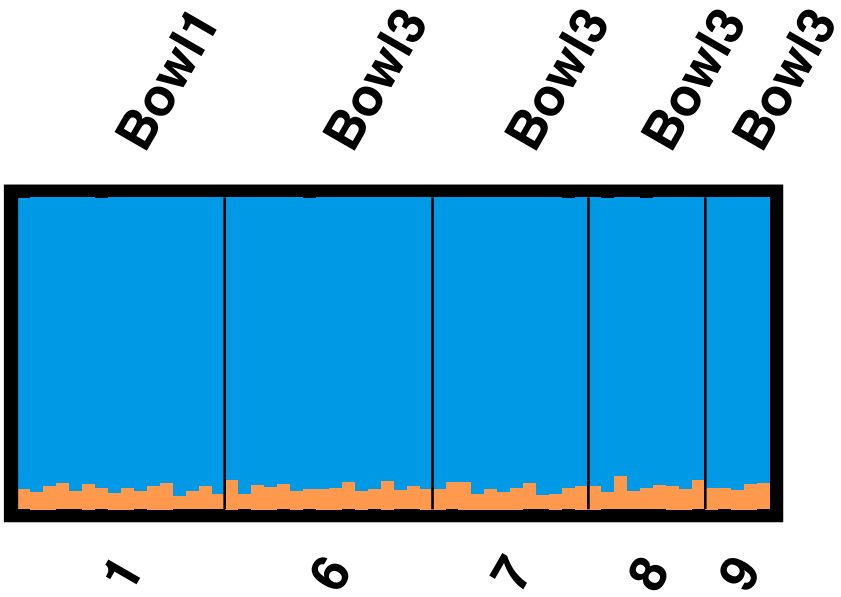


Figure 6. Patterns of population structure for *Xylophaga oregona* based on SNP data analyzed in STRUCTURE and visualized in DISTRUCT testing for the true number of populat

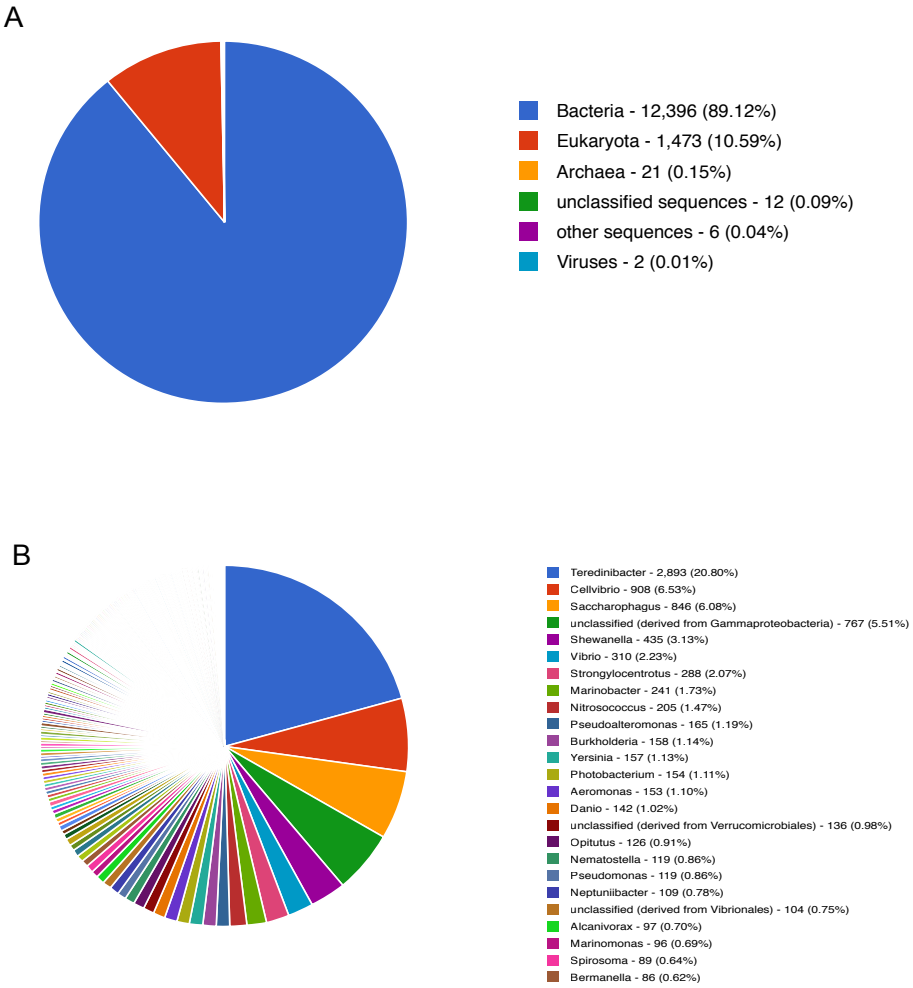


Figure 7. Taxonomic hits distribution from metagenomic assembly from MG-RAST. (A) Domain level and (B) Genus level.

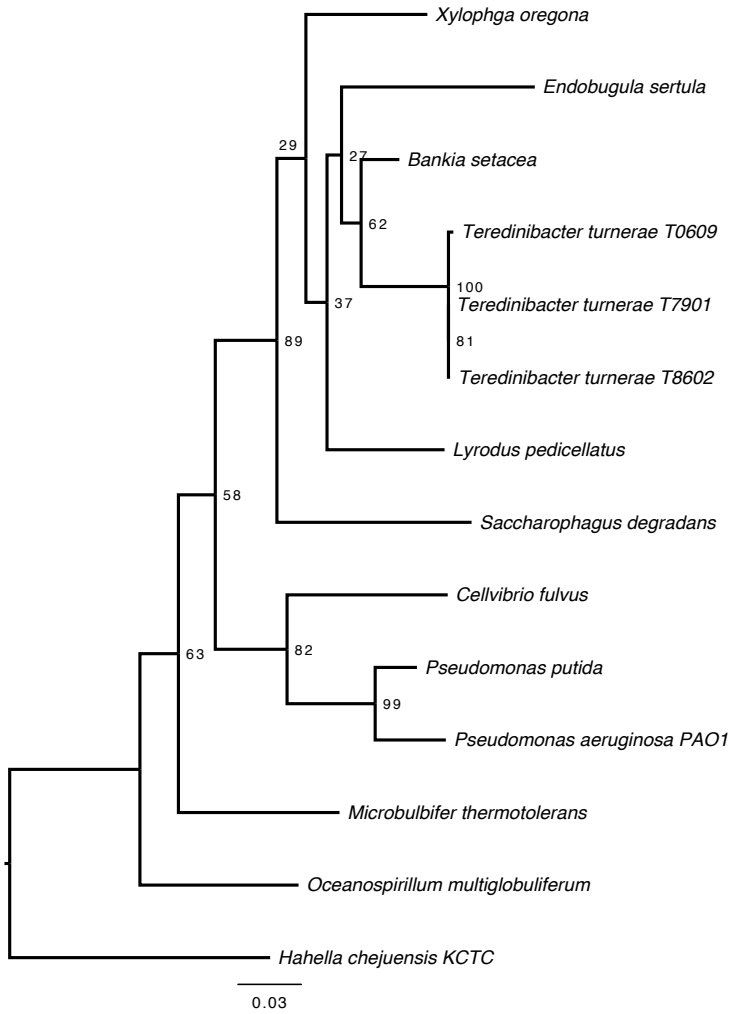


Figure 8. Maximum likelihood analysis based on 16s gene.

Table 1

Specimen data for sequenced taxa.

OTU	Clade	Specimen Collection		
		location	depth (m)	GPS coordinates
Xylophaga washingtona_3	*Xylophaga*	Friday Harbor Dock	~20	N 9°50.89′ W 104°17.49′
Xylophaga washingtona_5	*Xylophaga*	Friday Harbor Dock	~20	N 9°47.13′ W 104°16.13′
Xylophaga oregona_Bv346_18E	*Xylophaga*	BOWL6 (NE Pacific)	1605	N 43°54.52′ W 125°10.29′
Xylophaga oregona_Bv346_11E	*Xylophaga*	BOWL6 (NE Pacific)	1605	N 43°54.52′ W 125°10.29′
Xylophaga oregona_Bv352	*Xylophaga*	BOWL2 (NE Pacific)	1596	N 47°57.46′ W 126°02.19′
Xylophaga oregona_Bv354	*Xylophaga*	BOWL2 (NE Pacific)	1596	N 47°57.46′ W 126°02.19′
Xylophaga heterosiphoni	*Xylophaga*	BOWL3 (NE Pacific)	~3000	N 47°16.20′ W 127°35.57′
Xylophage zirenbergi	*Xylophaga*	BOWL3 (NE Pacific)	~3000	N 47°16.20′ W 127°35.57′
Xylophaga sp1_E81	*Xylophaga*	Santos Basin (SP)	1508m	S 25o53.64’ W 45o02.09'
Xylophaga sp1_E23	*Xylophaga*	Santos Basin (SP)	1508m	S 25o53.64’ W 45o02.09'
XYLOREDO	*Xylophaga*	Espirito Santo Basin (ES)	1491m	S 21o27.01’ W 39o53.79'
Xyloredo sp.E88	*Xyloredo*	Santos Basin (SP)	3358m	S 28o01.71’ W 43o31.78'
Xyloredo sp.E89	*Xyloredo*	Santos Basin (SP)	3358m	S 28o01.71’ W 43o31.78'

Table 2 Xylophagiad mt genome stats.

OTU	Clade	Mt genome statistics		
		Genome Size (bp)	GC%	Sequencing Coverage
Xylophaga washingtona_3	*Xylophaga*	18,624	40.03	N 9°50.89′ W 104°17.49′
Xylophaga washingtona_5	*Xylophaga*	18,599	40.02	N 9°47.13′ W 104°16.13′
Xylophaga oregona_Bv346_18E	*Xylophaga*	17,883	40.19	N 43°54.52′ W 125°10.29′
Xylophaga oregona_Bv346_11E	*Xylophaga*	18,178	40.15	N 43°54.52′ W 125°10.29′
Xylophaga oregona_Bv352	*Xylophaga*	18,477	40.43	N 47°57.46′ W 126°02.19′
Xylophaga oregona_Bv354	*Xylophaga*	18,477	40.43	N 47°57.46′ W 126°02.19′
Xylophaga heterosiphoni	*Xylophaga*	14,450	43.82	N 47°16.20′ W 127°35.57′
Xylophage zirenbergi	*Xylophaga*	15,083	42.9	N 47°16.20′ W 127°35.57′
Xylophaga sp1_E81	*Xylophaga*	16,921	38.3	S 25o53.64’ W 45o02.09'
Xylophaga sp1_E23	*Xylophaga*	16,461	38.2	S 25o53.64’ W 45o02.09'
Xyloredo sp1	*Xylophaga*	17,322	37.68	S 21o27.01’ W 39o53.79'
Xyloredo sp2_E88	*Xyloredo*	16,156	32.12	S 28o01.71’ W 43o31.78'
Xyloredo sp2_E89	*Xyloredo*	16,415	32.4	S 28o01.71’ W 43o31.78'

Table 3

Specimen information for 2b-RAD sequencing.

Species	Clade	Specimen Collection		
		location	Numeber of loci	Number of specimen
Xylophaga oregona		BOWL6 (NE Pacific)	1,910	16
Xylophaga		BOWL2 (NE Pacific)		44
Xyloredo sp1	*Xyloredo*	Santos Basin (SP)	2,100	10
		Espirito Santo Basin (ES)		18

Chapter 6. Conclusions and Future Directions

6.1 Siboglinid mitogenomics

My first chapter (Li et al., 2015 (Chapter 2)) has demonstrated that using whole mitochondrial genome sequences are useful for resolving siboglinid phylogeny. Within siboglinids, bone-eating *Osedax* is most closely related to a clade comprising Vestimentifera and *Sclerolinum* with storng nodal support, not Frenulata as recently reported (Glover et al., 2013). However, AU test could not reject the alternative hypothesis.

In terms of mitochondrial genome evolution, gene orders and sizes in siboglinds are largely similar to other annelids. Interestingly, the putative length of the control region among siboglinid mt genomes shows variability among taxa, ranging from 186 bp (*Tevnia jerichonana*) to 4,737 bp (*Siboglinium fiordicum*), which represents a nearly 25-fold difference in length. Moreover, no obvious conserved regions or secondary structures have been observed in control regions across mt genomes from different siboglinid species and clades, except that all contained TA tandem repeats. Since intronic microsatellites can affect gene transcription, mRNA splicing or export to cytoplasm these TA tandem repeats may have functional significance in the mt genomes of siboglinids.

6.2 Siboglinid phylogenomics

Because of the AU test could not reject the alternative hypothesis of the phylogenetic position of *Osedax*, we sought to apply a phylogenomic approach towards resolving the evolutionary history of siboglinids (Li et al., 2016). In particular, placement of the bone-eating *Osedax* worms has been unclear mainly due to their distinctive biology, including harboring heterotrophic bacteria as endosymbionts, displaying marked sexual dimorphism, and exhibiting a distinct body-plan. Every analysis conducted herein strongly supported *Osedax* being most closely related to the Vestimentifera+*Sclerolinum* clade, rather than the Frenulata. More improtanly, Explicit hypothesis testing with AU tests also significantly rejected *Osedax* as the sister group to Frenulata. Together with the results from Chapter 1, our analyses indicates that *Osedax*, the only siboglinid lineage with heterotrophic endosymbionts, evolved from a lineage utilizing chemoautotrophic symbionts.

Additionally, several supermatrix and species-tree approaches were also conducted in order to evaluate the effects of gene incongruence when inferring phylogeny from large concatenated datasets. Our results suggest that the supermatrix approach with ML using data partitioning with site-homogenous models appears to have outperformed both the supermatrix method with CAT-GTR and multispecies-coalescent approaches.

Therefore, the results conducted in Chapter 1 and 2 provide a well-supported phylogenetic hypotheses generated here should serve as a foundation for future studies on siboglinid evolution

including the evolution of different obligate symbioses, adaptation, and colonization to different reducing habitats.

6.3 Siboglinid symbiont evolution

Given the success of the well-resolved host phylogeny within siboglind, we sought to employ a comparative genomics of their associated symbionts. Importantly, most symbiont genomes have been characterized are only limited to hydrothermal vent localities. Given this limited data, how metabolic machineries differ between the endosymbionts of vent and seep-dwelling vestimentiferans, or between symbionts of vestimentiferans and their diminuative cousins, the frenulates, is not well understood.

Our results strongly suggest that all sampled endosymbionts from seep-dwelling siboglinids are also able to use rTCA cycle in addition to Calven-Benson cycle for carbon fixation. However, representative of frenulates, the *Galathealinum* symbionts lack key enzymes associated with rTCA and can only use Calvin cycle for carbon fixation. Thus, we hypothesize that symbionts with higher metabolic flexibility in carbon fixation may allow tubeworms to thrive in more reducing environments, such as seeps and vents. In addition, we take a comparative approach to systematically characterize the molecular mechanisms related to the process of infection, including motility guided by chemotaxis, secretion systems, type IV pili and genes potentially related to toxin and immunity. These results suggest that there are previous unrecognized links among siboglind symbionts from different deep-sea chemosynthetic

environments and shed light on understanding of evolutionary trends of siboglinid host-symbiont evolution.

6.4 Symbiont-host evolution of the deep-sea wood-boring Xylophagaidae

Chapter 5 of my dissertation was trying to address the mitochondrial genome evolution, recruitment sources, and symbioses of wood-boring Mollusca family Xylophagaidae. Wood fall habitats are comparable with whale falls due to their ephemeral distribution in the deep-sea (Distel et al., 2000). As in Chapter 2, a mitogenomic analysis was performed to explore evolutionary relationships with xylophagaids. Our analysis *Xylophaga* is a paraphyletic clade, and xylophagaid evolutionary relationships do not corresponding to isolation by depth in the deep-sea. Connectivity and recruitment of two xylophagaid species from two landers in NE Pacific and two landers in SW Atlantic where evaluated by 2b-RAD sequencing approach. Our findings show that there is no population structure identified across 500 km spacing in each species, suggesting that individuals from same xylophagaid species living in bathyal site (~1500m) are most likely from the same gene pool in both basins. Lastly, metagenomic analysis from *Xylophaga* gill tissue shows their symbiont genome is closely related to *Teridinibacter* species isolated from the closely related shallow water shipworm family Teredinidae. A gene encoding cellulase is identified in the *Xylophaga* symbiont genome, which potentially indicating a similar functional role in these endosymbionts from both shallow and deep wood-boring bivalves.

6.5. Future Directions

Although phylogenetic relationships within Siboglinidae are much better understood now, some important aspects still remain unanswered. In particular, the molecular mechanisms of interaction between microbes and hosts have not been well characterized. Moreover, innate immunity and apoptosis are also critical for organisms with host-symbiont interactions and being held in such extreme environments. Evolution of innate immunity across invertebrate taxa as little is know, even though it is far more complex in some invertebrates than traditionally recognized (Halanych and Kocot 2014, Tassia et al. submitted). The vestimentiferan tubeworm *Lamellibrachia lumysi* distributed along the Gulf of Mexico are one of the most common recognized and extensively studied seep tubeworms (Fig. 1). Similar to other siboglinid tubeworms, typically they lack a digestive tract and rely on sulfide-oxidizing bacterial symbionts for their nutrition. Hosts acquire their endosymbionts from the surrounding environment and store them in a specialized worm tissue called the trophosome. Interestingly, seep vestimentiferans have a much slower growth rate and have greater longevity than their vent relatives. For example, individuals of *L. luymesi* are estimated to live up to 300 years.

Given that recent advances in high-throughput sequencing and bioinformatics allow novel approaches to sequencing whole nuclear genomes more efficiently, high-quality annotated genomes from siboglindis will undoubtedly help allow us to gain insights into the adaptation of

deep-sea tubeworms to abiotic stresses in extreme chemosynthetic environments as well as characterize the general mechanisms of host-symbiont interactions. Although we don't have complete siboglinid genomes yet, in the course of my dissertation research, my collaborators and I have generated a large amount of transcriptome from several siboglnids. Moreover, although sequences from multiple paired-end and mate-pair libraries with different insertion size were already generated for this project, the quality of raw assembly was relative low (N50 = 43,518 bp). Thus, we employed a 10X genomics (http://www.10xgenomics.com) approach (e.g., Zheng et al. 2016) with an approximately 50X coverage of the genome. This approach yields long scaffolds inexpensively.

Whale-falls represent one of the most extraordinary and poorly sample habitats (Rouse et al., 2004; Glover et al., 2005). The sulfide and lipid from decomposing whale-bones is though to serve as an stepping-stones for several taxa to that of vents and seep habitats (Distel, 2000). The recent discovery of bone-eating *Osedax*, a novel siboglinid lineage is associated with heterotrophic endosymbiont rather than chemoautotrophic symbionts as they used in their close relatives. However, despite their worldwide distribution, *Osedax* symbiont genome has been only characterized in only one species at NE Pacific (Goffredi et al., 2012). We found that *Osedax rubiplumus* collected in Antarctic shared the same symbiont species as they found in NE Pacific. In the future, I intend to sample more *Osedax* symbionts from different ocean basins (e.g.

Atlantic) and depth zones (abyssal and bathyal zones) to fully understand their extraordinary symbioses.

In terms of mitochondrial genome evolution, although over 65,000 genomes have been sequenced to date (records from GenBank), majority of them are from insects or vertebrates. Mitochondrial genome evolution in most of marine invertebrate groups is still poorly known. For example, my collaborators and I found that multiple introns were present in introns within the *cox1*, *nad1* and *nad4* genes of deep-sea Ampharetid *Decemunciger* worms. This is the greatest number of introns observed in annelid mtDNA genomes, and possibly in bilaterians (Bernardino et al., 2017). The sequence of the introns within *cox1* is similar to Group II introns previously identified, suggesting that introns in the mitochondrial genome of annelids may be more widespread then realized. Moreover, we just recently found that a novel codon usage in mitochondrial genome of *Cephalodiscus* species. More interestingly, this alternative codon usage was first reported (Bessho et al., 1992) in flatworms (flatworm and echinoderms shared the same mitochondrial codon usage). The existence of this particular genetic code was disputed in Telford et al. (2000) but they didn't rule out that possibility and it was subsequently confirmed in two nematode species (Jacob et al., 2009). Therefore, the animal mitochondrial genome evolution is more interesting and sophisticated that we previously recognized.

Ark clams or ark shells (order Arcoida Gray, 1854) are a well-known, economically important group of bivalves. They are amongst the oldest extant bivalve lineages, dating to the lower Ordovician (~450 Mya; Morton et al., 1998). Today, species of Arcoida are globally distributed, predominantly in the tropical shallow waters and warm temperate seas, containing approximately 570 species (Huber, 2010), and have their maximum species richness in the Indo-West Pacific. Several species have significant economic value. Arcoida encompasses two superfamilies: Arcoidea (including Arcidae, Cucullaeidae, Noetiidae, Glycymerididae and Parallelodontidae), and Limopsoidea, (including Philobryidae and Limopsidae). Although Arcoida is monophyletic (e.g., Steiner and Hammer, 2000; Giribet and Wheeler, 2002; Matsumoto, 2003; Bieler et al., 2014), its internal relationships remain controversial and a conundrum for bivalve systematics (Feng et al., 2015; Combosch and Giribet, 2016). Despite several studies (citations), including two recent studies employing multi-locus datasets with a broad Arcoida sampling, relationships among major families remain controversial. Mitogenomics have proven useful in resolving phylogenetic relationships across a wide range of metazoans (e.g. Osigus et al., 2013; Miya et al., 2001; Li et al., 2015). At present, only six complete mt genomes of ark shells were available (Liu et al., 2013; Sun et al., 2015a, 2015b, 2015c; Sun et al., 2016), but ark shell mitochondrial genomes appear unique in that sizes of some mitogenomes are 2-3 times the size of other bilaterians (typically 15-17kb), i.e. 46,985 bp for *S. broughtonii*, 46,713 bp for *S. kagoshimensis*, 31,589 bp for *T. granosa*, 34,147 bp for A. vellicata, 28,470 bp *P. pilula* and 19,614 bp for *T. kiyoni*. Much of this increase size is the result

of non-coding regions which have the potential to influence mitochondrial energetics and replication. Moreover, their genomes show distinct gene arrangement patterns, namely unique rearrangements involving the tRNA genes (Liu et al., 2013; Sun et al., 2015a, 2015b, 2015c; Sun et al., 2016). To further explore Arcoida phylogeny, and to understand unusual large and variable sizes of the mitochondrial genomes, I intend to sequence mt genomes from representatives of all major Arcoida lineages as well as nuclear genomes from 3 commercially important taxa during my postdoctoral work. This information will serve as a comparative framework for understanding genomic evolution of arc shells. Because arc shells are commercially important, the genomic work will initially focus on comparison of gene systems relevant to aquaculture. Specifically, we will explore gene pathways that relate to shell deposition (Kocot et al. 2016) and growth as well as immunity (Tassia et al. 2016).

6.6 References

Combosch, D.J., Giribet, G., 2016. Clarifying phylogenetic relationships and the evolutionary history of the bivalve order Arcida (Mollusca: Bivalvia: Pteriomorphia). *Mol. Phylogenet. Evol*. 94, 298–312.

Feng, Y., Li, Q., Kong, L., 2015. Molecular phylogeny of Arcoidea with emphasis on Arcidae species (Bivalvia: Pteriomorphia) along the coast of China: challenges to current classification of arcoids. *Mol. Phylogenet. Evol*. 85, 189–196.

Goffredi, S.K., Yi, H., Zhang, Q., Klann, J.E., Struve, I.A., Vrijenhoek, R.C., Brown, C.T., 2014. Genomic versatility and functional variation between two dominant heterotrophic symbionts of deep-sea Osedax worms. ISME J 8, 908–924. doi:10.1038/ismej.2013.201

Halanych, K.M. and Kocot, K.M., 2014. Repurposed transcriptomic data facilitate discovery of innate immunity toll-like receptor (TLR) genes across lophotrochozoa. *The Biological Bulletin*, *227*(2), pp.201-209.

Huber, M (2010). Compendium of bivalves. *ConchBooks*: Hackenheim, Germany.

Kocot, K.M., Aguilera, F., McDougall, C., Jackson, D.J. and Degnan, B.M., 2016. Sea shell diversity and rapidly evolving secretomes: insights into the evolution of biomineralization. *Frontiers in Zoology*, 13(1), p.1.

Li, Y., Kocot, K.M., Schander, C., Santos, S.R., Thornhill, D.J., Halanych, K.M., n.d. Mitogenomics reveals phylogeny and repeated motifs in control regions of the deep-sea family Siboglinidae (Annelida). Molecular Phylogenetics and Evolution. doi:10.1016/j.ympev.2015.02.008

Li, Y., Kocot, K.M., Whelan, N.V., Santos, S.R., Waits, D.S., Thornhill, D.J., Halanych, K.M., 2016. Phylogenomics of tubeworms (Siboglinidae, Annelida) and comparative performance of different reconstruction methods. Zool Scr n/a-n/a. doi:10.1111/zsc.12201

Sun, S. E., Kong, L., Yu, H., Li, Q., 2015b. The complete mitochondrial DNA of *Tegillarca granosa* and comparative mitogenomic analyses of three Arcidae species. Gene, 557, 61–70.

Sun, S. E., Kong, L., Yu, H., Li, Q., 2015c. Complete mitochondrial genome of *Anadara vellicata* (Bivalvia: Arcidae): A unique gene order and large atypical non-coding region. Comp. Biochem. *Physiol. D: Genomics Proteomics*. 16, 73–82.

Sun, S. E., Li, Q., Kong, L., Yu, H., 2016. Complete mitochondrial genomes of *Trisidos kiyoni* and *Potiarca pilula*: Varied mitochondrial genome size and highly rearranged gene order in Arcidae. *Scientific Report*. 6, 33794.

Sun, S.E., Kong, L.F., Yu, H., Li, Q., 2015a. The complete mitochondrial genome of *Scapharca kagoshimensis* (Bivalvia: Arcidae). *Mitochondrial DNA* 26, 957-958.