Zoologica Scripta

KUNGL. VETENSKAPS-AKADEMIEN
THE ROYAL SWEDISH ACADEMY OF SCIENCES

Phylogenomics of tubeworms (Siboglinidae, Annelida) and comparative performance of different reconstruction methods

YUANNING LI, KEVIN M. KOCOT, NATHAN V. WHELAN, SCOTT R. SANTOS, DAMIEN S. WAITS, DANIEL J. THORNHILL & KENNETH M. HALANYCH

Submitted: 28 January 2016
Accepted: 18 June 2016
doi:10.1111/zsc.12201

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. —*Zoologica Scripta*, *00*: 000–000.

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 harbouring 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 data sets 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-coalescent approaches when the amount of missing data varies in a data set and when taxa susceptible to LBA are included in the analyses.

Corresponding authors: *Yuanning Li and Kenneth M. Halanych, Department of Biological Sciences & Molette Biology Laboratory for Environmental and Climate Change Studies, Auburn University, 101 Rouse Life Sciences Bldg. Auburn University, AL 36849, USA. E-mails: yzl0084@auburn.edu, ken@auburn.edu*

Yuanning Li, Department of Biological Sciences & Molette Biology Laboratory for Environmental and Climate Change Studies, Auburn University, 36830 Auburn, AL, USA. E-mail: yzl0084@auburn.edu

Kevin M. Kocot, Department of Biological Sciences & Molette Biology Laboratory for Environmental and Climate Change Studies, Auburn University, 36830 Auburn, AL, USA and Department of Biological Sciences & Alabama Museum of Natural History, The University of Alabama, 35847 Tuscaloosa, AL, USA. E-mail: kmkocot@ua.edu

Nathan V. Whelan, Scott R. Santos, Damien S. Waits, Daniel J. Thornhill, and Kenneth M. Halanych, Department of Biological Sciences & Molette Biology Laboratory for Environmental and Climate Change Studies, Auburn University, 36830 Auburn, AL, USA. E-mails: nwhelan@auburn.edu, santos@auburn.edu, dsw0002@tigermail.auburn.edu, thornhill.dan@gmail.com, ken@auburn.edu

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 *et al.* 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 regard 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* harbour 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 (i) Siboglinidae is monophyletic, (ii) the four major groups within Siboglinidae are each monophyletic, (iii) Vestimentifera is sister group to *Sclerolinum*, and (iv) 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

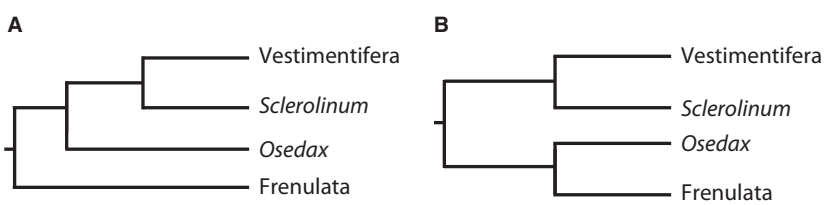

Fig. 1 Phylogenetic hypotheses from previous molecular studies. (A) Hypothesis of *Osedax* as the sister group to Vestimentifera and *Sclerolinum* clade (Rouse *et al.* 2004; Glover *et al.* 2005; Li *et al.* 2015). (B) Hypothesis of *Osedax* closely related to Frenulata (Glover *et al.* 2013; Rouse *et al.* 2015).

inference with large multilocus data sets: (i) the supermatrix (i.e. concatenation) approach and (ii) methods that use multispecies-coalescent models to resolve conflict among independently generated trees (Gatesy & Springer 2014; Edwards *et al.* 2016); methods such as *BEAST (Heled & Drummond 2010) that co-estimate gene and species trees are generally too computationally expensive for phylogenomic sized data sets. 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 supermatrix 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 analysis (BCA; Larget *et al.* 2010) with our data sets 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).

Methods

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, *et al.* 2005; *Osedax rubiplumus* Rouse, *et al.* 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 2 × 100 paired-end sequencing on an Illumina HiSeq 2000 platform (San Diego, California). cDNA for *Escarpia spicata* Jones, 1985, *Galathealinum brachiosum*

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	SRX1842875
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

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.

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.* 2016) 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 data set was generated by removing any OG sampled for fewer than 13 taxa. After creation of this first data set, we found that *Spirobrachia* sp. had more missing data than other taxa

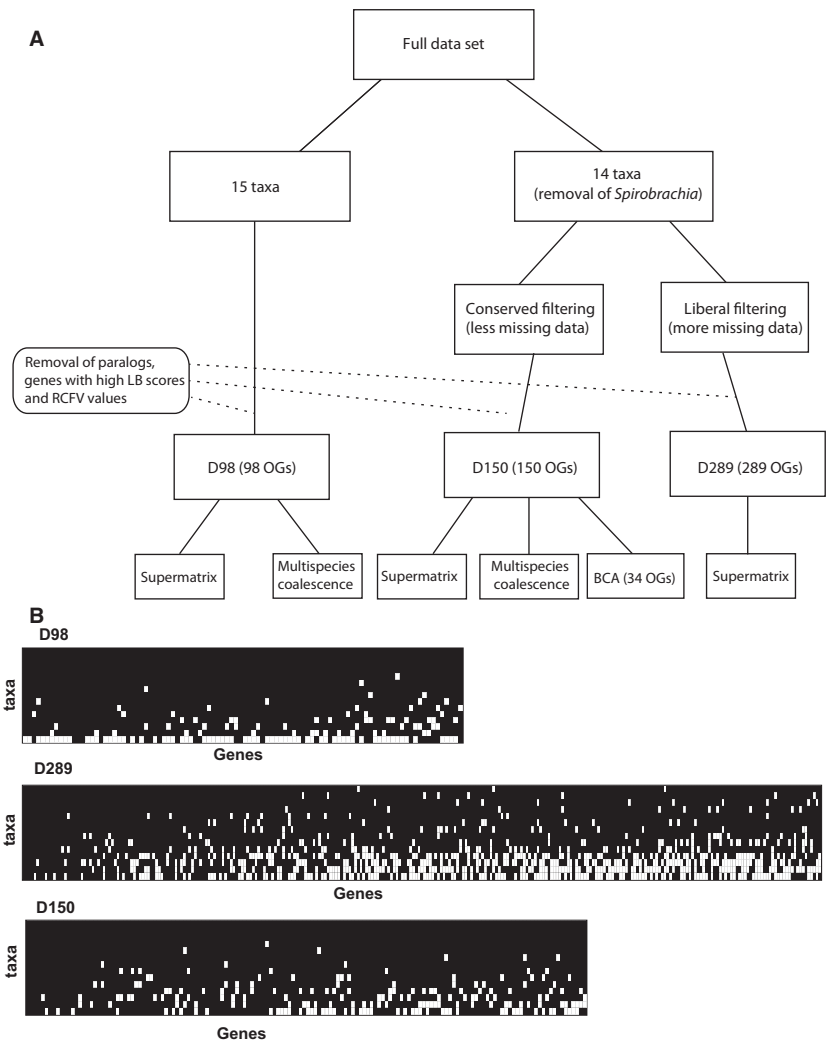

Fig. 2 (A) Chart of data filtering during data matrix construction and tree reconstruction methods employed for the D98, D150 and D298 data sets. Data statistics for each data set is shown in Table. 2. (B) Occupancy of orthologous groups in data matrices for phylogenetic analyses. Genes are ordered along 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.

Table 2 Statistics for phylogenomic data set

Data set	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

(only 24.5% of total orthologs present, Table 2), and thus, it was removed in two subsequent filtering data sets to accommodate more OGs. For these two additional filtering strategies, any gene with fewer than 10 or 12 taxa, respectively, was removed. All three data sets (D98, D150, D289 – numbers refer to numbers of OGs included; Fig. 2B) were processed by first discarding sequences that were shorter than 50 amino acid residues. 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 (Kück 2009). Alignment columns with only gaps were subsequently removed, and any OG with an alignment less than 50 amino acid residues after trimming was 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 data sets 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 (RCFV; Zhong *et al.* 2011), was calculated for each OG from each data set using BaCoCa (Fig. S1). 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 data sets.

Phylogenetic analyses

Fifteen 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, multi-species-coalescent methods and BCA. For the supermatrix approach, matrices of concatenated OGs were analysed 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 data sets.

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 and 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 approach, BCA does not make any biological assumptions about drivers of gene tree heterogeneity (Ane *et al.* 2007). Thus, BCA is not a strictly coalescent-based method. 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 the D98 and D150 data sets were estimated and nodal support was calculated with 100 fast bootstrap replicates using RAxML 8.0.23. We did not analyse data set 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 multilocus 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 data set, 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 (α = 0.1, 1, 10, 100; α = 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.

Hypothesis testing

To assess the robustness of the inferred phylogenetic position of *Osedax*, an approximately unbiased (AU; Shimodaira 2002) test was used to determine whether 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).

Results

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. S1), 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 data set and the overall matrix completeness value, which considers alignment gaps as missing data, was 75.0%. For the D289 data set, on average 81.7% of orthologs were sampled per taxon, with an overall matrix completeness of 65.2%; for the D150 data set, 91.0% of the orthologs were sampled per taxon, with an overall matrix completeness of 79.5% (Table 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. S2 (D150; 14 taxa). Supermatrix analysis with data set D289 recovered an identical branching order to the tree inferred with data set D150, but with slightly higher nodal support values (Figs. 3, S2). Both data sets 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

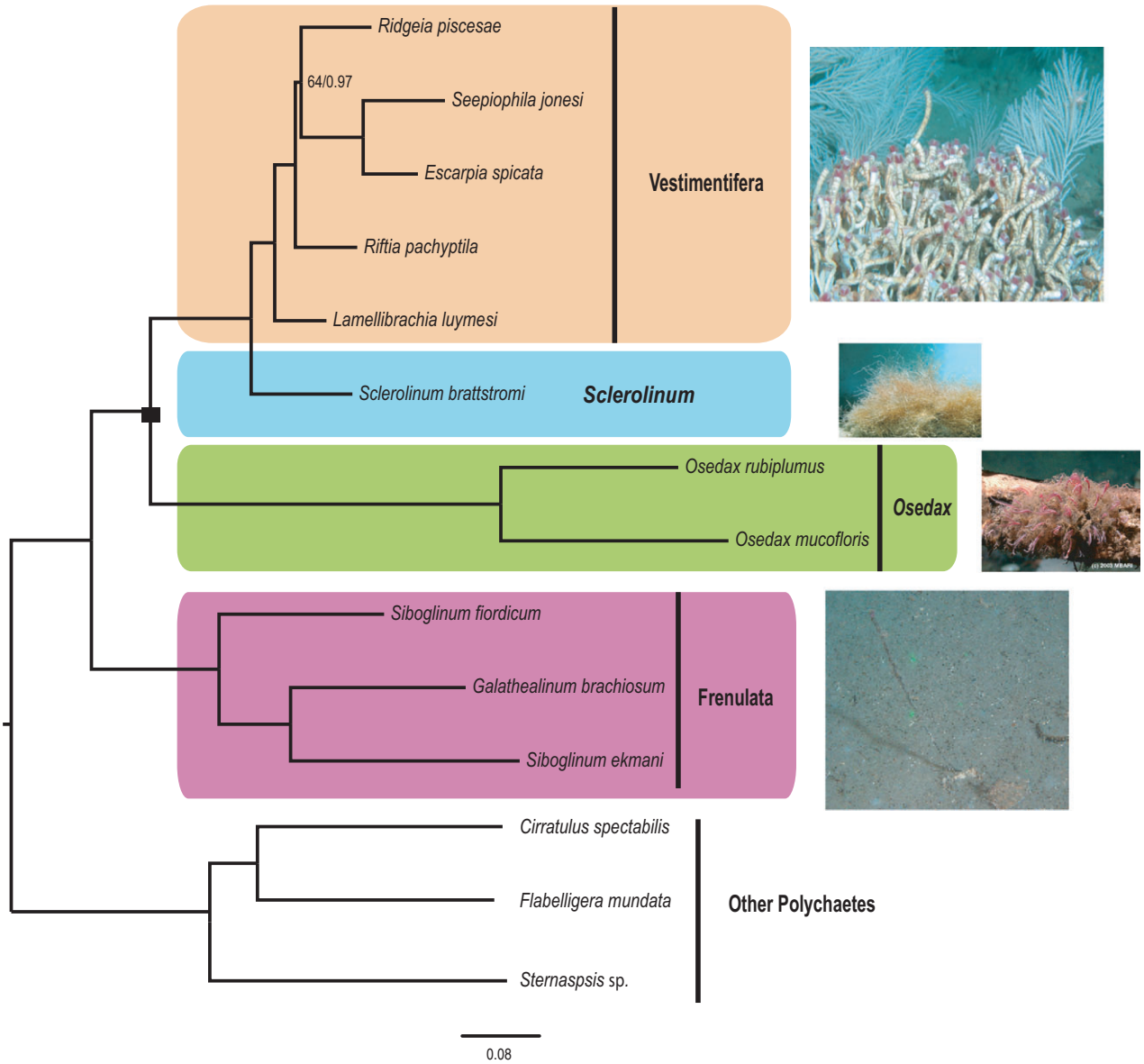

Fig. 3 Phylogenetic reconstructions of Siboglinidae based on data set D289 using supermatrix approach and a Bayesian inference approach with a CAT-GTR model. Majority-rule consensus phylogram is shown. The black square indicates the node joining the *Osedax* lineage to the vestimentiferan/*Sclerolinum*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

the sister group to Frenulata was explicitly rejected by AU tests on all three data sets (Table 3). The topology inferred from data set 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 with any other taxon, exhibited long branches in both analyses. In the BI tree inferred using CAT-GTR and data set 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 data sets in order to accommodate data sets with less missing data and more loci. Both ML and BI recovered identical topologies in data sets D289 and D150, but variability among interrelationships within Vestimentifera and Frenulata were noted in

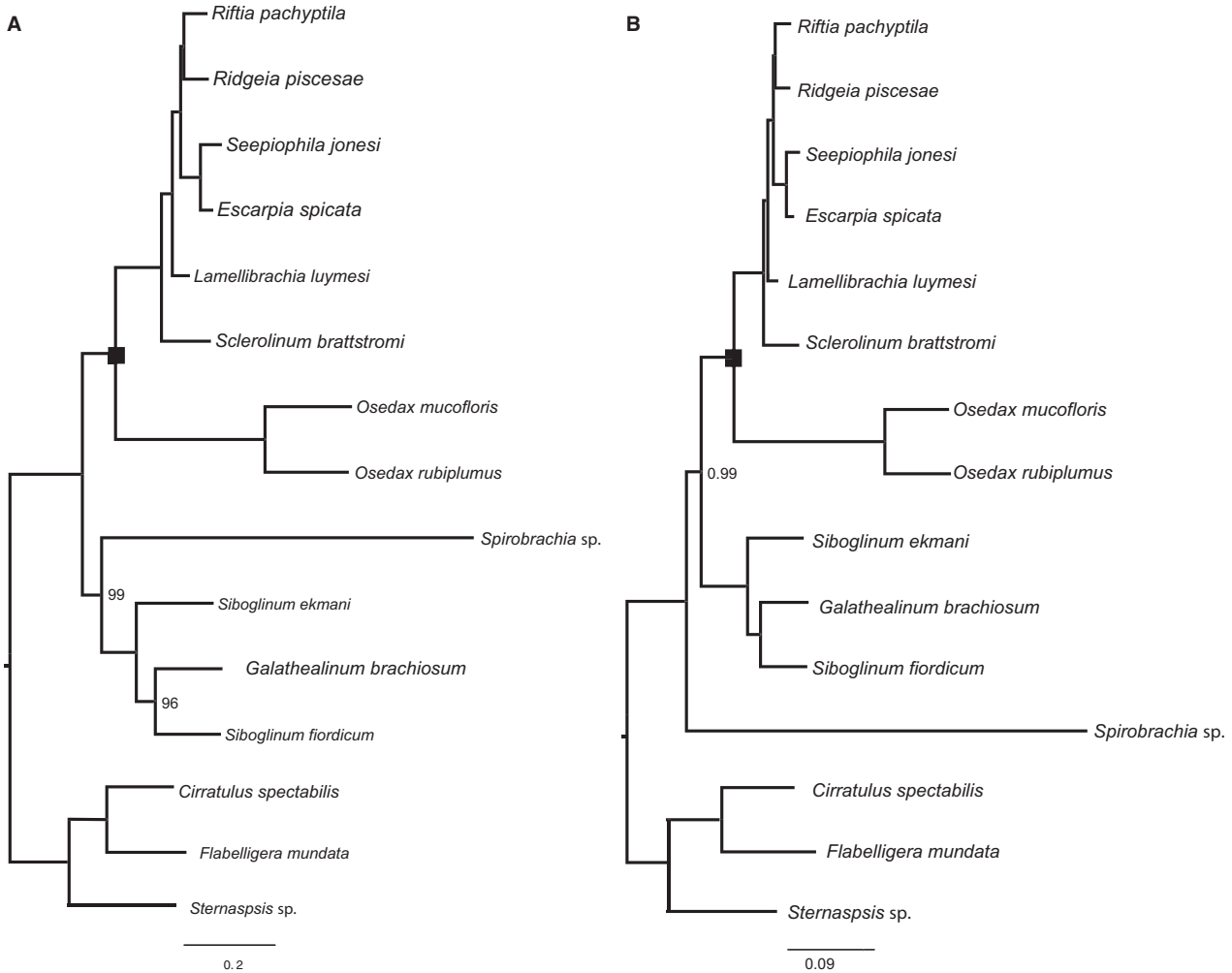

Fig. 4 Phylogenetic reconstructions of Siboglinidae from data set D98. Topologies derived from ML (A) with bootstrap support and BI using CAT+GTR model (B) with posterior probabilities. The black square indicates the node joining the *Osedax* lineage to the vestimentiferan/*Sclerolinum* clade. All nodes were supported with 100% bootstrap value or posterior probabilities of 1.0 unless otherwise noted. "Other polychaetes" form a basal polytomy but are shown here as a group for simplicity.

Table 3 AU tests of competing phylogenetic hypothesis

	D98		D150		D289	
Tree Topology	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

D98 data set (Figs. 3, 4, S2). For example, in data sets D289 and D150, *Ridgeia* was sister to *Seepiophila* and *Escarpia*, whereas it formed a clade with *Riftia* in the analysis of data set D98. This result suggests that data set size had more of an effect on relationships within Vestimentifera and Frenulata than differences between ML and BI with CAT-GTR.

Phylogenetic analysis using multispecies-coalescent approaches

Given that our supermatrix analyses showed congruent topologies between data sets D289 and D150 (Figs. 3, S2), multispecies-coalescent analyses were only performed on the smaller D150 and D98 data sets due to computational demands of some coalescent-based methods. In general,

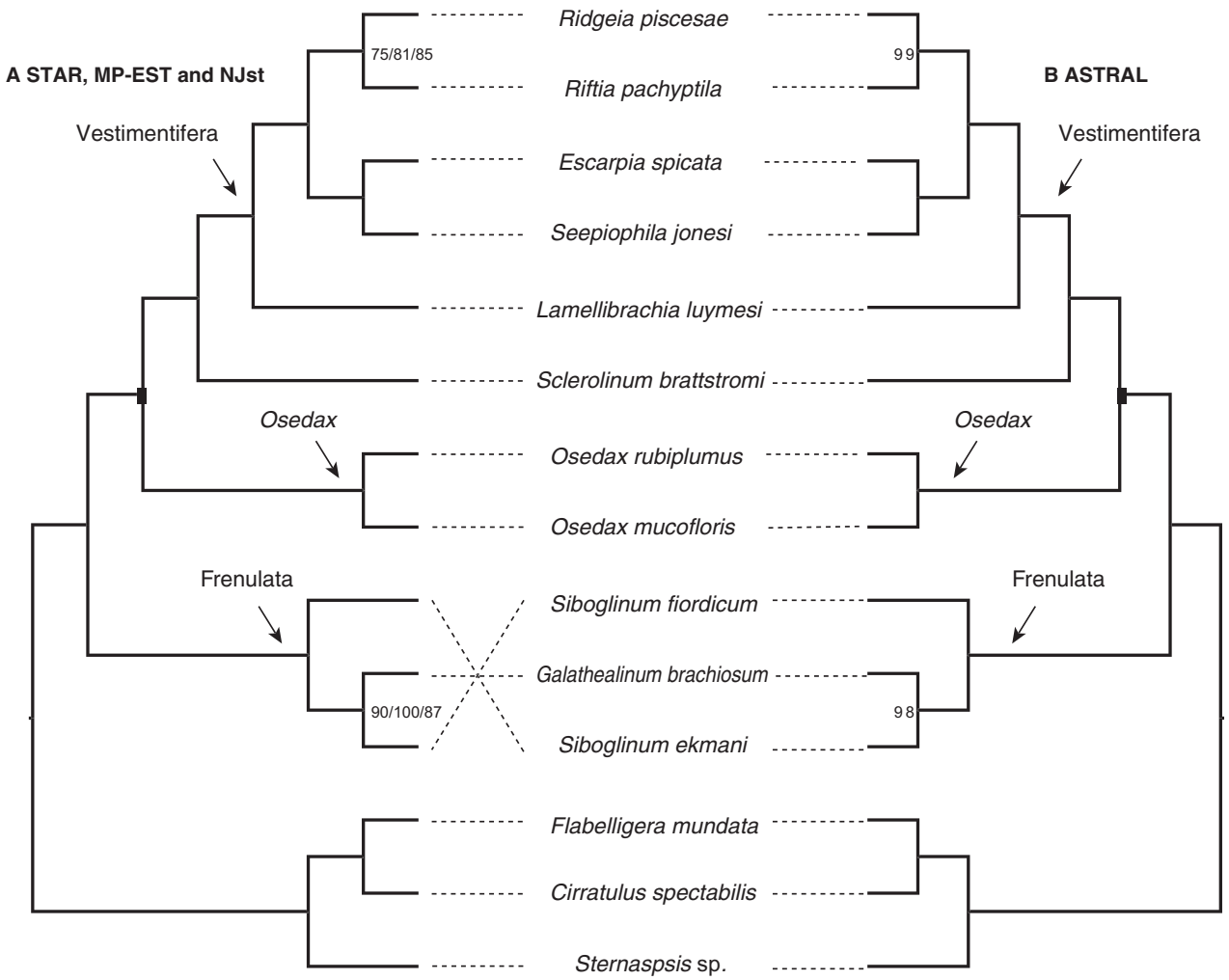

Fig. 5 Species trees inferred from (A) based on STAR, MP-EST, NJst and (B) ASTRAL from data set D150. The black square indicates the node joining the *Osedax* lineage to the vestimentiferan/*Sclerolinum* clade. Nodal support values (A) left: STAR; middle: MP-EST; right: NJst (B) ASTRAL indicate bootstrap proportion based upon 100 multilocus bootstraps. "Other polychaetes" form a basal polytomy but are shown here as a group for simplicity.

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% multilocus bootstrap support. For both the D150 and D98 data sets, species trees derived from STAR, NJst and MP-EST exhibited the same tree topology (Figs. 5A, S3). *Siboglinum ekmani* was placed as sister to other Frenulata from all D98 multispecies-coalescent analyses, whereas *S. fiordicum* was sister to other frenulates based on ASTRAL in both data sets (Figs. 5B, S4) and in supermatrix analyses. Similar to the BI analysis of D98 using the CAT-GTR model, *Spirobrachia* was placed as sister to all other siboglinids in the multispecies-coalescent analyses.

Bayesian concordance analysis

The BCA tree (Fig. S5) derived from the reduced D150 data set 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* was different compared with 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 as sister to other frenulates (CF = 0.61), instead of *S. fiordicum* as inferred from supermatrix analyses.

Discussion

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 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 data sets. 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 data sets (Figs. S2, 3), but not in data set D98 because the tree inferred with CAT-GTR placed *Spirobrachia as* 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).

Performance of supermatrix vs. multispecies-coalescent approaches

Large phylogenomic data sets 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.* 2016), we wished to explore the performance of these approaches on a phylogenomic data set 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 was noted above, some of these conflicts were likely due to differences in data set size. Notably, conflicting results were obtained from supermatrix and multispecies-coalescent methods with data set D98. BI analysis using CAT models has been widely used for phylogenomic analyses because of its purported superiority in handling LBA (Delsuc *et al.* 2008; Philippe *et al.* 2009, 2011). Yet *Spirobrachia* was unexpectedly placed as sister to all other siboglinids in BI with data set D98 (Fig. 4B), the same result as multispecies-coalescent-based analyses (Figs. 5, S4) of the D98 data set. 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 data sets (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-coalescent analyses are both potentially more susceptible to error when at least one taxon has large amounts of missing data compared with 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. S5). 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 data sets D289 and D150. In contrast to previous studies, we failed to recover an *Osedax*/Frenulata sister relationship with any data sets 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 data set D98 in terms of the placement of *Spirobrachia*. 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 appear 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.

Acknowledgements

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.

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.

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.

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.

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.

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.

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

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

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

Delsuc, F., Tsagkogeorga, G., Lartillot, N. & Philippe, H. (2008). Additional molecular support for the new chordate phylogeny. *Genesis*, *46*, 592–604.

Dunn, C. W., Hejnol, A., Matus, D. Q., Pang, K., Browne, W. E., Smith, S. A., Seaver, E., Rouse, G. W., Obst, M., Edgecombe, G. D., Sørensen, M. V., Haddock, S. H., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R. M., Wheeler, W. C., Martindale, M. Q. & Giribet, G. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, *452*, 745–749.

Ebersberger, I., Strauss, S. & von Haeseler, A. (2009). Hamstr: profile hidden markov model based search for orthologs in ests. *BMC Evolutioanry Biology*, *9*, 157.

Edwards, S. V., Xi, Z., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., Zhong, B., Wu, S., Lemmon, E. M., Lemmon, A. R., Leaché, A. D., Liu, L. & Davis, C. C. (2016). Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. *Molecular Phylogenetics and Evolution*, *94*, 447–462.

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.

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.

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.

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.

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. *Proceedings of the Royal Society B: Biological Sciences*, *280*, 20131390.

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.

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. *Nature Biotechnology*, *29*, 644–652.

Halanych, K. M. (2005). Molecular phylogeny of siboglinid annelids (a.K.A. Pogonophorans): a review. *Hydrobiologia*, *535*, 297–307.

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.

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., Dahlgren, T. G. & McHugh, D. (2002). Unsegmented annelids? Possible origins of four lophotrochozoan worm taxa. *Integrative and Comparative Biology*, *42*, 678–684.

Heled, J. & Drummond, A. J. (2010). Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, *27*, 570–580.

Hilário, A., Capa, M., Dahlgren, T. G., Halanych, K. M., Little, C. T., Thornhill, D. J., Verna, C. & Glover, A. G. (2011). New perspectives on the ecology and evolution of siboglinid tubeworms. *PLoS ONE*, *6*, e16309.

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.

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

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

Jones, M. L. (1988). The vestimentifera, their biology, systematic and evolutionary patterns. *Oceanologica Acta*, *Special issue*, *8*, 69–82.

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., Santos, S. R., Schander, C., Moroz, L. L., Lieb, B. & Halanych, K. M. (2011). Phylogenomics reveals deep molluscan relationships. *Nature*, *477*, 452–456.

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.

Kocot, K. M., Struck, T. H., Merkel, J., Waits, D. S., Todt, C., Brannock, P. M., Weese, D. A., Cannon, J. T., Moroz, L. L. & Halanych, K. M. (2016). Phylogenomics of Lophotrochozoa with consideration of systematic error. *Systematic Biology*, Epub ahead of print.

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.

Lane, C. E. (2007). Bacterial endosymbionts: genome reduction in a hot spot. *Current Biology*, *17*, R508–R510.

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.

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

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

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.

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.

Lartillot, N., Lepage, T. & Blanquart, S. (2009). PhyloBayes 3: a bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, *25*, 2286–2288.

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.

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.

Little, C. T. S. & Vrijenhoek, R. C. (2003). Are hydrothermal vent animals living fossils? *Trends in Ecology & Evolution*, *18*, 582–588.

Liu, L. & Yu, L. (2011). Estimating species trees from unrooted gene trees. *Systematic Biology*, *60*, 661–667.

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

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.

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.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L., Jarvie, T. P., Jirage, K. B., Kim, J. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F. & Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*, 376–380.

Matus, D. Q., Copley, R. R., Dunn, C. W., Hejnol, A., Eccleston, H., Halanych, K. M., Martindale, M. Q. & Telford, M. J. (2006). Broad taxon and gene sampling indicate that chaetognaths are protostomes. *Current Biology*, *16*, R575–R576.

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–i548.

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.

Misof, B., Liu, S., Meusemann, K., Peters, R. S., Donath, A., Mayer, C., Frandsen, P. B., Ware, J., Flouri, T., Beutel, R. G., Niehuis, O., Petersen, M., Izquierdo-Carrasco, F., Wappler, T., Rust, J., Aberer, A. J., Aspöck, U., Aspöck, H., Bartel, D., Blanke, A., Berger, S., Böhm, A., Buckley, T. R., Calcott, B., Chen, J., Friedrich, F., Fukui, M., Fujita, M., Greve, C., Grobe, P., Gu, S., Huang, Y., Jermiin, L. S., Kawahara, A. Y., Krogmann, L., Kubiak, M., Lanfear, R., Letsch, H., Li, Y., Li, Z., Li, J., Lu, H., Machida, R., Mashimo, Y., Kapli, P., McKenna, D. D., Meng, G., Nakagaki, Y., Navarrete-Heredia, J. L., Ott, M., Ou, Y., Pass, G., Podsiadlowski, L., Pohl, H., von Reumont, B. M., Schütte, K., Sekiya, K., Shimizu, S., Slipinski, A., Stamatakis, A., Song, W., Su, X., Szucsich, N. U., Tan, M., Tan, X., Tang, M., Tang, J., Timelthaler, G., Tomizuka, S., Trautwein, M., Tong, X., Uchifune, T., Walzl, M. G., Wiegmann, B. M., Wilbrandt, J., Wipfler, B., Wong, T. K., Wu, Q., Wu, G., Xie, Y., Yang, S., Yang, Q., Yeates, D. K., Yoshizawa, K., Zhang, Q., Zhang, R., Zhang, W., Zhang, Y., Zhao, J., Zhou, C., Zhou, L., Ziesmann, T., Zou, S., Li, Y., Xu, X., Zhang, Y., Yang, H., Wang, J., Wang, J., Kjer, K. M. & Zhou, X. (2014). Phylogenomics resolves the timing and pattern of insect evolution. *Science*, *346*, 763–767.

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

Oliver, J. C. (2013). Microevolutionary processes generate phylogenomic discordance at ancient divergences. *Evolution*, *67*, 1823–1830.

Philippe, H., Derelle, R., Lopez, P., Pick, K., Borchiellini, C., Boury-Esnault, N., Vacelet, J., Renard, E., Houliston, E., Quéinnec, E., Da Silva, C., Wincker, P., Le Guyader, H., Leys, S., Jackson, D. J., Schreiber, F., Erpenbeck, D., Morgenstern, B., Wörheide, G. & Manuel, M. (2009). Phylogenomics revives traditional views on deep animal relationships. *Current Biology*, *19*, 706–712.

Philippe, H., Brinkmann, H., Copley, R. R., Moroz, L. L., Nakano, H., Poustka, A. J., Wallberg, A., Peterson, K. J. & Telford, M. J. (2011). Acoelomorph flatworms are deuterostomes related to *Xenoturbella*. *Nature*, *470*, 255–258.

Price, M. N., Dehal, P. S. & Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE*, *5*, e9490.

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*, 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.

Rouse, G. W., Goffredi, S. K. & Vrijenhoek, R. C. (2004). *Osedax*: bone-eating marine worms with dwarf males. *Science*, *305*, 668–671.

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.

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.

Schulze, A. (2003). Phylogeny of vestimentifera (siboglinidae, annelida) inferred from morphology. *Zoologica Scripta*, *32*, 321–342.

Schulze, A. & Halanych, K. M. (2003). Siboglinid evolution shaped by habitat preference and sulfide tolerance. *Hydrobiologia*, *496*, 199–205.

Shaw, T. I., Ruan, Z., Glenn, T. C. & Liu, L. (2013). Straw: species tree analysis web server. *Nucleic Acids Research*, *41*, W238–W241.

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, *51*, 492–508.

Shimodaira, H. & Hasegawa, M. (2001). Consel: for assessing the confidence of phylogenetic tree selection. *Bioinformatics*, *17*, 1246–1247.

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.

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

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). Netherlands: Springer.

Springer, M. S. & Gatesy, J. (2015). The gene tree delusion. *Molecular Phylogenetics and Evolution*, *94*, 1–33.

Stamatakis, A. (2014). Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*, 1312–1313.

Struck, T. H. (2014). TreSpEx-detection of misleading signal in phylogenetic reconstructions based on tree information. *Evolution Bioinformatics Online*, *10*, 51–67.

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.

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., Santos, S. R., Halanych, K. M., Purschke, G., Bleidorn, C. & Struck, T. H. (2014). Illuminating the base of the annelid tree using transcriptomics. *Molecular Biology and Evolution*, *31*, 1391–1401.

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.

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.

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.

Zhong, B., Liu, L., Yan, Z. & Penny, D. (2013). Origin of land plants using the multispecies coalescent model. *Trends Plant Science*, *18*, 492–495.

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.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. 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.

Fig. S2. Phylogenetic reconstructions of Siboglinidae inferred from D150 dataset.

Fig. S3. Species tree inferred from D98 based on STAR, MP-EST and NJst.

Fig. S4. Species tree inferred from ASTRAL using the D98 database.

Fig. S5. Primary concordance tree reconstructed using BUCKy with 34 OGs derived from D150 dataset.

Table S1. Specimen data for sequenced taxa.