Elucidating Biological Questions with Bioinformatics Tools, with a Case Study of Tube Worm Hemoglobin

by

Damien Scott Waits

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
August 1, 2015

Bioinformatics, annelid, hemoglobin, contamination, phylogenetics

Approved by

Kenneth M. Halanych, Chair, Professor of Biological Sciences
Jason E. Bond, Professor of Biological Sciences
Leslie R. Goertzen, Professor of Biological Sciences

Abstract

The goal of my Master's Thesis research was to utilize sequenced data in conjunction with new data to investigate the sulfur binding hemoglobin of the annelid family Siboglinidae. To accomplish this, multiple bioinformatic scripts were developed to help streamline the process of isolating and analyzing hemoglobin sequence data. Given the large amount of sequence data, bioinformatics pipelines are necessary to efficiently clean, sort, and analyze the information. Siboglinidae is a group of annelids living in chemosynthetic environments that has been studied extensively. Their unique symbiosis facilitated by sulfur-binding hemoglobin is shared by all members except the genus *Osedax*. Using the available and newly generated sequence data, we sought to determine if *Osedax* possessed the genetic machinery capable of producing sulfur-binding hemoglobin using a bioinformatics approach. During this study, multiple scripts were written to efficiently analyze the data. Some of these larger programs that have since been included in pipelines used to explore animal phylogeny. Herein, I describe my research on the hemoglobin of the siboglinids and the bioinformatics tools that have resulted. Chapter 1 describes my research on siboglinid hemoglobin and corresponds to a manuscript submitted to the *Journal of Molecular Evolution*. Chapter 2 reports on two bioinformatic programs that were used in a phylogenomic pipeline implemented by members of our research group.

Acknowledgments

I am grateful to my committee, Drs. Jason Bond and Les Goertzen. I am especially grateful to my advisor, Ken Halanych, for giving me the opportunity to work and learn in a biology laboratory from a computer science background. Thanks to the Halanych lab members, past and present, who have helped guide me: Kevin, Joie, and Pam for constantly getting me into the field to make sure I stuck around, Matt for sitting behind me so I always had someone to complain to, Alexis for making sure I always had a nickname or two, Amanda for teaching me lab procedures that got me through being a tech, Branson for being old, Li for messing with Branson, Nathan for ensuring the lab is never too quiet, Sammi for pinking up the lab, Victoria for that cornbread recipe, and Mike for his Asian cuisine advice. Thanks go to Scott Santos for his plentiful advice and discussions on the topics of bioinformatics and computing. A special thank you to my friends, family, funding agencies and any others who have helped me along the way. A super special thank you to Meghan, for helping me through my final year.

Table of Contents

List of Tables

Chapter 1

List of Figures

Chapter 1

Chapter 2

List of Abbreviations

cDNA	complimentary DNA
DNA	deoxyribonucleic acid
Hb	hemoglobin
H_2S	hydrogen sulfide
PCR	polymerase chain reaction
RNA	ribonucleic acid

Chapter 1. Evolution of Sulfur Binding in Hemoglobin in Siboglinidae (Annelida) with Special Reference to Bone Eating Worms, *Osedax*

1.1 Abstract

Most members of Siboglinidae (Annelida) harbor endosymbiotic bacteria that allow them to thrive in extreme environments such as hydrothermal vents, methane seeps, and whalebones. These symbioses are enabled by specialized hemoglobin (Hbs) that is able to bind hydrogen sulfide for transportation to their chemosynthetic endosymbionts. Sulfur-binding capabilities are hypothesized to be due to cysteine residues at key positions in both vascular and coelomic Hbs. Members of the genus *Osedax*, which live on whale bone, do not have chemosynthetic endosymbionts, but instead harbor heterotrophic bacteria capable of breaking down complex organic compounds. Although sulfur-binding capabilities are important in other siboglinids, we questioned whether *Osedax* retained these cysteine residues and the potential ability to bind hydrogen sulfide. To answer these questions, we used high-throughput DNA sequencing to isolate and analyze Hb sequences from 8 siboglinid lineages, including *Osedax mucofloris.* Once identified, Hb sequences from gene subfamilies A2 and B2 were translated and aligned to determine conservation of cysteine residues at previously identified key positions. Hb linker sequences were also compared to determine similarity between *Osedax* and siboglinids/sulfur-tolerant annelids. Our results found conserved cysteines within the A2 chain, but not the B2 chain, of *O. mucofloris* Hb. These residues may have been retained when *Osedax* diverged from other siboglinids. This finding suggests that Hb in *O. mucofloris* has retained some capacity to bind hydrogen

sulfide, likely due to the need to detoxify hydrogen sulfide that is abundantly produced within whalebones.

1.2 Introduction

Siboglinid annelids occur throughout the world's oceans but are best known from hydrothermal vents, cold seeps, and whalebones (Schulze and Halanych 2003; Rouse et al. 2004; Southward et al. 2005). Their dominance at these environments is largely due to symbioses with chemotrophic bacteria (Cavanaugh et al. 1981; Southward and Southward 1981; Halanych 2005; Goffredi et al. 2005; Thornhill et al. 2008). Siboglinidae is comprised of four lineages: frenulates, vestimentiferans, monoliferans, and *Osedax* (Hilário et al. 2011). Frenulates, comprising the majority of known siboglinid species, are often thread-like and found within sediments of reducing habitats (Southward 1978; Southward et al. 2005; Thornhill et al. 2008; Hilário et al. 2010). Vestimentiferans, on the other hand, are large tubeworms that are typically found in hydrothermal vents and cool seeps (McMullin et al. 2003). Monilifera is represented by a single genus (i.e., *Sclerolinum*) that shares similarities to frenulates, but can also be found on decaying organic material (Halanych et al. 2001). Finally *Osedax*, first described in 2004, are worms that colonize whalebones (Rouse et al. 2004; Glover et al. 2005).

Adult siboglinids lack a functional gut and instead rely on chemosynthetic endosymbionts to supply some or all of their energetic needs (Cavanaugh et al. 1981; Hilário et al. 2011). In this context, hydrogen sulfide (H_2S) is absorbed and transported via the blood vascular system to millions of symbiotic bacteria within the specialized organ called the trophosome (Southward 1988; Goffredi et al. 2005; Katz et al. 2011;

Bright et al. 2012). Most siboglinid endosymbionts are chemoautotrophic and generally belong to the gamma-proteobacteria (Thornhill et al. 2008; Verna et al. 2010). In contrast, *Osedax*, whose morphology is more arborescent in appearance, harbor heterotrophic Oceanospirillales endosymbionts in a root-like system that extends into the whalebone matrix (Goffredi et al. 2005) where endosymbionts utilize the complex compounds released from the bones (Rouse et al. 2004). Approximately 31 lineages of *Osedax* have been discovered (Smith et al. 2015) and phylogenetic analyses based on ribosomal genes and mitochondrial cytochrome oxidase I usually place *Osedax* as sister to a moniliferan-vestimentiferan clade (Rouse et al. 2004; Glover at al. 2005), but Glover et al. (2013) and Rouse et al. (2015) suggest a position sister to frenulate siboglinids. Despite this suggestion, recent analyses of whole mitochondrial genome data strongly favor allying *Osedax* with vestimentiferans and monoliferans (Li et al. 2015; Fig 1).

For some chemoautotroph-bearing siboglinids, H_2S uptake and transport is mediated by specialized hemoglobins (Hbs) (Numoto et al. 2005; Meunier et al. 2010). Reversible binding of H_2S to Hbs have been best studied in the vestimentiferans *Riftia pachyptila* and *Lamellibrachia luymesi,* as well as the frenulate *Oligobrachia mashikoi* (e.g., Suzuki et al. 1990; Yuasa et al. 1996; Zal et al. 1996ab; 1997). Hbs are complex structures with individual globin chains assembling into hetero-dimer subunits. Those subunits, in turn, assemble into a tetrameric functional protein, with each heme directly interacting with adjoining subunits whose size varies (Numoto et al. 2008). Vestimentiferans have one large extracellular Hb (V1 ~3500 kDa) and one small extracellular Hb (V2: ~400 kDa) in their vascular blood. Additionally, they possess one Hb (C1) in coelomic fluid that is reported to be 400 kDa (Arp and Childress 1981; Zal et al. 1996a). Whereas V1 contains 4 heme-containing globin chains (b-e) and 4 linker chains (L1-L4), V2 is composed of 6 globin chains (a-f), and C1 contains 5 globin chains (a-e). In contrast, the frenulate *Oligobrachia mashikoi* possesses a single ~400 kDA Hb composed of 24 globin chains with no linkers, comparable to the small extracellular Hbs of vestimentiferans (Yuasa et al. 1996; Numoto et al. 2005). Binding of H_2S has been hypothesized to be mediated in part by cysteine residues in the V1 chains and by disulphide bridges formed from cysteine-rich linker chains (*R. pachyptila*'s V1 chain *b* - B2 and *L. luymesi*'s V1 chain *AIII* - A2; Zal et al. 1996b, 1997). However, this only accounts for part of the binding affinity, and zinc moieties bound to amino acid residues at the interface between pairs of A2 chains may also be involved (Flores et al. 2005). With reference to *R. pachyptila*'s A2 chain, cysteines at positions 4 and 134 are common to all annelid globin chains studied and form a disulfide bridge while a free cysteine at position 75 is unique to sulfur oxidizing siboglinids (Zal et al. 1997).

Given our understanding of siboglinid phylogeny (Li et al. 2015), the bone-eating *Osedax* has likely evolved from ancestors dependent upon chemoautotrophic bacteria (Schulze and Halanych 2003; Hilario et al. 2011) at least 100 million years ago (based on fossil and molecular data; Danise and Higgs 2015). Due to its heterotrophic symbiosis, *Osedax* is apparently no longer dependent on H_2S transport or the modified blood physiology to nourish endosymbionts (Rouse et al. 2004; Goffredi et al. 2005). We assume the ability to bind H_2S carries a cost to the organism, as most Hbs lack such affinity and has been suggested to be selected against in sulfide-free habitats (Bailly et al. 2003). Based on this, we hypothesized that the *Osedax* Hb system would exhibit differences relative to other siboglinids; specifically, amino acid substitutions for carrying H_2S should be lacking in *Osedax*. To this end, we employed high-throughput DNA sequencing to generate transcriptomic data to allow examination of amino acid sequence of Hbs and linker proteins from *O. mucofloris*, three frenulates, a moniliferan, and three vestimentiferans, in addition to publically available data. Specific targets were the level of conservation among Cyt residues (especially at positions 4, 75, and 134) in Hb chains across siboglinids as well as conceptually examining how amino acid differences may influence protein-folding characteristics.

1.3 Materials and Methods

1.3.1 Siboglinid sampling

Siboglinid samples were procured for transcriptome sequencing from a variety of sources (Table 1). Specifically, Christoffer Schander kindly provided *O. mucofloris* from whalebones near Bergen, Norway and *Sclerolinum contortum* from the Håkons-Mosby mud volcano off Norway. Samples of *Lamellibrachia luymesi, Escarpia spicata, Seepiophila jonesi,* and *Galathealinum brachiosum* were collected in the Gulf of Mexico using the *Johnson Sea Link* submersible aboard the *R/V Seward Johnson*. Samples of *Siboglinum fiordicum* were obtained using a small hand grab on the *R/V Aurelia* (University of Bergen) and *Siboglinum ekmani* were obtained by dredge on the *R/V Håkons-Mosby* from near Bergen, Norway. At the time of collection, all samples were morphologically identified and stored in RNALater.

1.3.2 Extraction and Sequencing

RNA extraction and cDNA preparation for high-throughput sequencing followed Kocot et al. (2011) and Li et al. (2015). Briefly, RNA was extracted using a TRIzol (Invitrogen) protocol, and then purified with the RNeasy kit (Qiagen) using an on-column 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). The double-stranded cDNA from *O. mucofloris* was sequenced on an Illumina MiSeq sequencer at Auburn University using a Nextera (Illumina) protocol, as well as an Illumina HiSeq 2000 sequencer at the Genomics Services Laboratory at the Hudson Alpha Institute for Biotechnology (Huntsville, AL, USA) using the TruSeq v3 (Illumina) protocol. cDNA for *Escarpia spicata, G. brachiosum, L. luymesi, and S. jonesi* were sent to the University of South Carolina Environmental Genomics Core Facility (Columbia, SC, USA) for Roche 454 GS-FLX sequencing. Additionally, cDNAs for *L. luymesi, S. contortum, S. ekmani, and S. fiordicum* were sequenced on an Illumina HiSeq 2000 sequencer at Hudson Alpha Institute for Biotechnology.

1.3.3 Sequence assembly

Sequencing reads were digitally normalized using the normalize-by-median script in the khmer package (https://github.com/ctb/khmer/blob/master/scripts/normalize-by-median.py) to facilitate assembly and decrease the likelihood that overrepresentation of reads would cause assembly artifacts (McDonald and Brown 2013). Transcriptome assemblies from MiSeq and 454 data were done *de novo* with the October 2012 release of

Trinity (Grabherr et al. 2011) while HiSeq 2000 data were assembled with the February 2013 release of the same software.

1.3.4 BLAST and sequence alignment

Hb and linker sequences of interest were obtained from assembled transcriptomes via BLAST (Altschul et al. 1990) by utilizing Hb and linker sequences acquired from GenBank of siboglinids as well as outgroup organisms as queries (Table 2). Specifically, an e-value cutoff of 10^{-5} was utilized in tblastn searches of nucleotide assemblies with the query protein sequences. *Arenicola marina*, a sulfur tolerant polychaete, was used as outgroup based on the availability of these sequences. Resulting BLAST hits were filtered using blast2table.pl (available from http://www.genome.ou.edu/informatics.html) with the "top" option, which reports only the best, high-scoring segment pair for each query sequence. Linker sequence hits were manually evaluated based on e-value and percent identity to determine similarity. The resulting Hb hits were translated using ESTScan version 3.0.3 (Iseli et al. 1999) and sequences aligned using MUSCLE within MEGA 5.2 (Tamura et al. 2011), The alignment was visually inspected and spuriously aligned data removed based on similarity to the alignment as a whole.

1.3.5 Gene Tree and Visualization of Data

Following alignment, A2 Hb sequences were manually trimmed of missing leading and trailing positions and Gblocks version 0.91b (Castresana 2000; Talavera and Castresana 2007) was used to trim poorly aligned positions and divergent regions with the following parameters: minimum number of sequences for a conserved position = 7, minimum number of sequences for a flank position = 7, maximum number of contiguous non-conserved positions = 8, minimum length of a block = 2, and gap positions allowed in all blocks. An appropriate amino acid substitution model for phylogenetic reconstruction was selected using Prottest version 3.4 (Darriba et al. 2011). RAxML version 7.3.8 (Stamatakis 2014) was used to infer a maximum likelihood gene tree with 100 bootstrap replicates using the PROTGAMMAWAG model, with *A. marina* serving as the outgroup. *Osedax mucofloris* Hb chain A2 was visualized as a 3D model using the RaptorX protein structure prediction server, which uses template-based tertiary structure modeling (Källberg et al. 2012).

1.4 Results

1.4.1 Sequencing results

High-throughput DNA sequencing produced 283,594 - 750,876 reads for 454, 3,027,776 reads for MiSeq, and 21,397,136 - 56,067,578 reads for HiSeq 2000 (Table 1). Contigs per assemblies were 7,209 - 12,080 for 454 data, and 17,617 - 270,658 for MiSeq and HiSeq 2000 data (Table 1).

1.4.2 BLAST results

Across the eight transcriptomes, tblastn searches returned 12 top hits (e-value cutoff of 10^{-5}) for chain A1, 17 for chain A2, 22 for chain B1, and 12 for chain B2. Upon closer inspection, the singular hit to *O. mucofloris* for chain B2 was a contig that also was returned in searches for chain A2 homologs and the B2 hit was discarded based on the higher strength of the A2 hit. These top hits were combined with data acquired from

NCBI's GenBank to generate alignments for each of the four Hb chains. After manual removal of redundant and incorrect sequences, a single contig for each chain was retained per taxon. However after inspection of the alignment, A1 sequences were not recovered for *E. spicata* and *G. brachiosum*. Additionally, the B2 sequence of *S. ekmani* has a single stop codon within the protein-coding region. This sequence was further verified via read mapping with BowTie2 (Langmead et al. 2009), and given that the sequence aligned well, we presumed it was a psuedogene.

The tblastn searches for linker sequences resulted in multiple hits for each species. The 454 libraries of *E. spicata, G. brachiosum, S. jonesi*, and *L. luymesi* had relatively few hits at 5, 6, 9, and 18 hits, respectively. Illumina libraries had higher numbers of hits, with 23 for *O. mucofloris*, 44 for *S. ekmani*, 47 for *S. fiordicum*, 75 for *L. luymesi*, and 118 for *S. brattstromi*. Upon manual inspection of each taxon's BLAST scores, all 8 transcriptomes were found to have an on-average higher score, e-value, and percent identity for hits to vestimentiferan linkers than to non-siboglinid linkers (Table 3).

1.4.3 Cysteine presence/absence

For chains A1 and B1, no free cysteine occurred at conserved amino acid positions for any taxon. For chain A2, conserved free cysteine at position 75 correlating to those found by Zal et al. (1997) were found in all taxa except *G. brachiosum* (Fig. 2). This species lacked a free cysteine between the two cysteines involved in the formation of disulfide bridges. For chain B2, one incorrect BLAST hit was recovered for *O.*

mucofloris; however, a conserved free cysteine was found for all other taxa excluding *E. spicata*, *G. brachiosum*, and *A. marina*.

1.4.4 Gene tree and 3D visualization

Final alignment of the 12 A2 chain sequences had 116 amino acid positions. Maximum-likelihood analysis of this alignment placed the *O. mucofloris* A2 sequence between the A2 sequences of frenulates and a moniliferan/vestimentiferan clade; however, frenulate sequences were recovered as paraphyletic with weak support (Fig 3). The *O. mucofloris* chain A2 sequence was recovered as sister to the monilferan/vestimentiferan chain A2 clade with moderate support (bootstrap = 73).

The 3D structure of the *O. mucofloris* Hb chain A2 protein model showed a noticeable "pocket". The disulfide chain forming cysteines and the free cysteine were positioned across from each other within this pocket (Fig 4).

1.5 Discussion

Contrary to our hypothesis, analyses presented here suggest *Osedax* has the biochemical capability of producing sulfur-binding Hbs. Specifically, *Osedax mucofloris* possesses a free cysteine at position 76 of the chain A2 of its Hbs while chain B2 does not. Additionally, 3D structure of the binding pocket (Fig 4) is consistent with the use of zinc moieties as previously described (Flores et al. 2005). Results for linker sequence comparison showed closer similarity to those of vestimentiferans than non-siboglinid taxa; however, sequence similarity based on BLAST score and percent match was on par with that of frenulates, which do not possess hexagonal bilayer Hbs with linkers and instead possess a form of ring Hb (Meunier et al. 2009). This finding was surprising since symbionts of *Osedax* spp. are not known to engage in chemosynthesis or sulfur metabolism and therefore the need for sulfur binding is unclear (Rouse et al. 2004).

One possibility is that selection for, and retention of, these residues are due to the involvement of Hbs in sulfide detoxification as part of *Osedax* life history at whale fall habitats. *Osedax mucofloris* possesses a high surface area to volume ratio in its root system, similar to the less branched root of *Lamellibrachia* where hydrogen sulfide uptake occurs (Julian et al. 1999; Huusgaard et al. 2012). Although the root epidermis of *Osedax* was suggested as an important site for nutrient uptake (Katz et al. 2010), how the mucus sheath that envelops the trunk and root structures of *Osedax mucofloris* (Higgs et al. 2011) effects chemical uptake from bones, including hydrogen sulfide, is unclear. Moreover, the exterior surface of whale bones experiences microbial sulfide production, with the potential for bone interiors to have microbial activity due to degradation of hydrophobic lipids overtime, a process that can be facilitated by *Osedax* (Treude et al. 2009). The presence of hydrogen sulfide within bones is further supported by observations of iron sulfide staining and white filamentous bacterial mats around *Osedax* boreholes (Higgs et al. 2011). These factors would indicate that *Osedax* roots are in an environment with relatively high hydrogen sulfide levels where the ability to detoxify it may be biologically advantageous. Free cysteines in Hbs are subject to negative selection in polychaetes from sulfide-free habitats (Bailly et al. 2003), further supporting that *Osedax* not only copes with hydrogen sulfide, but may use Hbs to interact with hydrogen sulfide in biologically-important ways. Whereas sulfur binding by Hbs in many siboglinids is primarily used to transport hydrogen sulfide to their endosymbionts, it also protects tissue from sulfide toxicity because it has a higher binding affinity than cytochrome-c oxidase, which is inhibited by small amounts of hydrogen sulfide (National Research Council 1979). The ability to bind sulfur could be under positive selection as part of a sulfide detoxification process (Eichinger et al. 2014).

Osedax mucofloris possesses Hb linkers with greater similarity to vestimentiferan siboglinids than to sulfide-tolerant polychaetes; a result consistent with a recent phylogeny for the group (Li et al. 2015). This could indicate that *Osedax* produces hexagonal bilayer Hbs with sites of cysteine-rich linker sequences forming disulphide bridges capable of sulfur binding. However, additional analyses are required before such conclusions can be made with confidence. In the context of the hypothesized phylogeny of Siboglinidae (Figs. 1), the presence of Hb linkers could indicate that the last common ancestor of vestimentiferan/moniliferan and *Osedax* possessed Hb that bound sulfur as well as oxygen. However, linker comparisons with frenulate transcriptomes recovered hits with sequence similarities on par with those of *Osedax mucofloris* hits, in contrast to the more robust hits to vestimentiferans, which can be explained by the taxa represented in our linker reference sequences. Currently, only vestimentiferan and moniliferan siboglinids have been shown to possess the hexagonal bilayer Hbs that self-assemble with linkers. As other annelids have large hexagonal bilayered Hbs, frenulates, possessing ring-shaped Hbs, seem to have lost the ability to produce linkers capable of creating more complex structures. Both ring and hexagonal bilayer Hbs use the same types of globins (Meunier et al. 2010), and similarities across these globin types likely confound the analyses of linker sequences presented here. Alternatively, the genetic distance between vestimentiferans/moniliferans and *Osedax* may be sufficiently high enough that linkers exhibit low homology. Quantification of the molecular mass of *Osedax* Hb would help determine whether *Osedax* Hbs are a hexagonal bilayer or a ring structure in nature.

Here, we analyzed *Osedax mucofloris* Hb as a first step towards determining how these proteins might function in the biology of these siboglinids bearing heterotrophic endosymbionts. Unlike most siboglinids, *Osedax* does not depend on chemotrophic endosymbionts and therefore should not require sulfur-binding Hb to support its endosymbionts. Yet sulfur-binding Hb has apparently persisted in this group of bone-eating worms. Remnants of this ability could be part of a sulfide detoxification process; an evolutionary vestige of ancient chemotrophic symbioses that has not yet been purged by mutation, selection and drift; or serve some other, yet to be discovered, functional role. Our results raise many questions about the role of sulfur binding in *Osedax*, topics that will be fruitful for future investigations.

1.6 Acknowledgements

We are indebted to the late Christopher Schander for his contribution of *Osedax mucofloris* tissue for sequencing. We would like to thank the crews of the *R/V Seward Johnson*, *R/V Aurelia*, and *R/V Håkons-Mosby* for their assistance in procuring samples. We would also like to thank Amanda Shaver and Franziska Franke for preparation of cDNA libraries for samples used in this study. Pamela M. Brannock, Kevin M. Kocot, and Nathan V. Whelan provided guidance and input on procedures used in this study and the writing of versions of this manuscript. This work was funded by National Science Foundation grants IOS-0843473 to K. M. H, S. R. S., and D. J. T. and DEB-1036537 to K. M. H. and S. R. S.

1.7 References

Altschul SF, Gish W, Miller W, et al (1990) Basic local alignment search tool. Journal of Molecular Biology 215:403–410. doi: 10.1016/S0022-2836(05)80360-2

Arp AJ, Childress JJ (1981) Blood function in the hydrothermal vent vestimentiferan tube worm. Science 213:342–344. doi: 10.1126/science.213.4505.342

Bailly X, Leroy R, Carney S, et al (2003) The loss of the hemoglobin H2S-binding function in annelids from sulfide-free habitats reveals molecular adaptation driven by Darwinian positive selection. PNAS 100:5885–5890. doi: 10.1073/pnas.1037686100

Bright M, Eichinger I, Salvini-Plawen L (2012) The Metatrochophore of a Deep-Sea Hydrothermal Vent Vestimentiferan (Polychaeta: Siboglinidae). Organisms Diversity & Evolution 13:163–88. doi:10.1007/s13127-012-0117-z.

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular biology and evolution 17:540–552.

Cavanaugh CM, Gardiner SL (1981) Prokaryotic cells in the hydrothermal vent tube worm *Riftia pachyptila* Jones: Possible chemoautotrophic symbionts. Science (New York, NY) 213:340–2. doi: 10.1126/science.213.4505.340

Danise S, Higgs ND (2015) Bone-eating *Osedax* worms lived on Mesozoic marine reptile deadfalls. Biology Letters 11:20150072–20150072. doi: 10.1098/rsbl.2015.0072

Darriba D, Taboada GL, Doallo R, Posada D (2011) ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164–1165.

Eichinger I, Schmitz-Esser S, Schmid M, et al (2014) Symbiont-driven sulfur crystal formation in a thiotrophic symbiosis from deep-sea hydrocarbon seeps. Environmental Microbiology Reports 6:364–372. doi: 10.1111/1758-2229.12149

Flores JF, Fisher CR, Carney SL, et al (2005) Sulfide binding is mediated by zinc ions discovered in the crystal structure of a hydrothermal vent tubeworm hemoglobin. Proceedings of the National Academy of Sciences of the United States of America 102:2713–2718.

Glover AG, Wiklund H, Taboada S, et al (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

Glover AG, Källström B, Smith CR, Dahlgren TG (2005) World-wide whale worms? A new species of *Osedax* from the shallow north Atlantic. Proceedings of the Royal Society of London B: Biological Sciences 272:2587–2592. doi: 10.1098/rspb.2005.3275

Goffredi SK, Orphan VJ, Rouse GW, 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 MG, Haas BJ, Yassour M, et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotech 29:644–652. doi: 10.1038/nbt.1883

Halanych KM (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 KM, Feldman RA, Vrijenhoek RC (2001) Molecular evidence that *Sclerolinum brattstromi* is closely related to vestimentiferans, not to frenulate pogonophorans (Siboglinidae, Annelida). Biol Bull 201:65–75.

Higgs ND, Glover AG, Dahlgren TG, Little CTS (2011) Bone-boring worms: Characterizing the morphology, rate, and method of bioerosion by *Osedax mucofloris* (Annelida, Siboglinidae). Biol Bull 221:307–316.

Hilário A, Capa M, Dahlgren TG, et al (2011) New perspectives on the ecology and evolution of siboglinid tubeworms. PLoS ONE 6:e16309. doi: 10.1371/journal.pone.0016309

Hilário A, Johnson SB, Cunha MR, Vrijenhoek RC (2010) High diversity of frenulates (Polychaeta: Siboglinidae) in the Gulf of Cadiz mud volcanoes: A DNA taxonomy analysis. Deep Sea Research Part I: Oceanographic Research Papers 57:143–150. doi: 10.1016/j.dsr.2009.10.004

Huusgaard RS, Vismann B, Kühl M, et al (2012) The potent respiratory system of *Osedax mucofloris* (Siboglinidae, Annelida) - A prerequisite for the origin of bone-eating *Osedax*? PLoS ONE 7:e35975. doi: 10.1371/journal.pone.0035975

Iseli C, Jongeneel CV, Bucher P (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. ISMB. pp 138–148

Julian D, Gaill F, Wood E, et al (1999) Roots as a site of hydrogen sulfide uptake in the hydrocarbon seep vestimentiferan *Lamellibrachia* sp. J Exp Biol 202:2245–2257.

Källberg M, Wang H, Wang S, et al (2012) Template-based protein structure modeling using the RaptorX web server. Nat Protocols 7:1511–1522. doi: 10.1038/nprot.2012.085

Katz S, Klepal W, Bright M (2011) The *Osedax* trophosome: Organization and ultrastructure. Biol Bull 220:128–139.

Katz S, Klepal W, Bright M (2010) The skin of *Osedax* (Siboglinidae, Annelida): An ultrastructural investigation of its epidermis. J Morphol 271:1272–1280. doi: 10.1002/jmor.10873

Kocot KM, Cannon JT, Todt C, et al (2011) Phylogenomics reveals deep molluscan relationships. Nature 477:452–456. doi: 10.1038/nature10382

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10:R25. doi: 10.1186/gb-2009-10-3-r25

Li Y, Kocot KM, Schander C, et al (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

McDonald E, Brown CT. khmer: Working with big data in bioinformatics. CoRR, abs/1303.2223, 2013.

McMullin ER, Hourdez S, Schaeffer SW, Fisher CR (2003) Phylogeny and biogeography of deep sea vestimentiferan tubeworms and their bacterial symbionts. Symbiosis 34:1–41.

Meunier C, Andersen AC, Bruneaux M, et al (2010) Structural characterization of hemoglobins from Monilifera and Frenulata tubeworms (Siboglinids): First discovery of giant hexagonal-bilayer hemoglobin in the former “Pogonophora” group. Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology 155:41–48. doi: 10.1016/j.cbpa.2009.09.010

National Research Council, Division of Medical Science, subcommittee on Hydrogen Sulfide (1979). *Hydrogen sulfide*. Baltimore: University Park Press 1979.

Numoto N, Nakagawa T, Kita A, et al (2008) Structural basis for the heterotropic and homotropic interactions of invertebrate giant hemoglobin. Biochemistry 47:11231–11238. doi: 10.1021/bi8012609

Numoto N, Nakagawa T, Kita A, et al (2005) Structure of an extracellular giant hemoglobin of the gutless beard worm *Oligobrachia mashikoi*. Proceedings of the National Academy of Sciences of the United States of America 102:14521–14526.

Okonechnikov K, Golosova O, Fursov M, Team the U (2012) Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics 28:1166–1167. doi: 10.1093/bioinformatics/bts091

Rouse GW, Wilson NG, Worsaae K, Vrijenhoek RC (2015) A dwarf male reversal in bone-eating worms. Current Biology 25:236–241. doi: 10.1016/j.cub.2014.11.032

Rouse GW, Goffredi SK, Vrijenhoek RC (2004) *Osedax*: Bone-Eating Marine Worms with Dwarf Males. Science 305:668–671. doi: 10.1126/science.1098650

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

Smith CR, Glover AG, Treude T, et al (2015) Whale-fall ecosystems: Recent insights into ecology, paleoecology, and evolution. Annual Review of Marine Science 7:571–596. doi: 10.1146/annurev-marine-010213-135144

Southward AJ, Southward EC (1981) Dissolved organic matter and the nutrition of the Pogonophora: a reassessment based on recent studies of their morphology and biology. Kieler Meeresf 5:445–453.

Southward EC, Schulze A, Gardiner SL (2005) Pogonophora (Annelida): form and function. In: Bartolomaeus T, Purschke G (eds) Morphology, Molecules, Evolution and Phylogeny in Polychaeta and Related Taxa. Springer Netherlands, pp 227–251

Southward EC (1988) Development of the gut and segmentation of newly settled stages of *Ridgeia* (Vestimentifera): implications for relationship between Vestimentifera and Pogonophora. Journal of the Marine Biological Association of the United Kingdom 68:465–487. doi: 10.1017/S0025315400043344

Southward EC (1978) A new species of *Lamellisabella* (Pogonophora) from the north Atlantic. Journal of the Marine Biological Association of the United Kingdom 58:713–718. doi: 10.1017/S0025315400041357

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

Suzuki T, Takagi T, Ohta S (1990) Primary structure of a constituent polypeptide chain (AIII) of the giant haemoglobin from the deep-sea tube worm *Lamellibrachia*. A possible H2S-binding site. http://www.biochemj.org/bj/266/bj2660221.htm.

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Systematic Biology 56:564–577.

Tamura K, Peterson D, Peterson N, et al (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 DJ, Wiley AA, Campbell AL, et al (2008) Endosymbionts of *Siboglinum fiordicum* and the phylogeny of bacterial endosymbionts in Siboglinidae (Annelida). Biol Bull 214:135–144.

Treude T, Smith C, Wenzhöfer F, et al (2009) Biogeochemistry of a deep-sea whale fall: sulfate reduction, sulfide efflux and methanogenesis. Marine Ecology Progress Series 382:1–21. doi: 10.3354/meps07972

Verna C, Ramette A, Wiklund H, Dahlgren TG, Glover AG, Gaill F, Dubilier N (2009) High symbiont diversity in the bone-eating worm *Osedax mucofloris* from shallow whale-falls in the North Atlantic. Environmental Microbiology 12:2355–70.

Yuasa HJ, Green BN, Takagi T, et al (1996) Electrospray ionization mass spectrometric composition of the 400 kDa hemoglobin from the pogonophoran Oligobrachia mashikoi and the primary structures of three major globin chains. Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology 1296:235–244. doi: 10.1016/0167-4838(96)00081-7

Zal F, Suzuki T, Kawasaki Y, et al (1997) Primary structure of the common polypeptide chain b from the multi-hemoglobin system of the hydrothermal vent tube worm *Riftia pachyptila*: An insight on the sulfide binding-site. Proteins: Structure, Function, and Bioinformatics 29:562–574.

Zal F, Lallier FH, Green BN, et al (1996a) The multi-hemoglobin system of the hydrothermal vent tube worm *Riftia pachyptila* II. Complete polypeptide chain composition investigated by maximum entropy analysis of mass spectra. J Biol Chem 271:8875–8881. doi: 10.1074/jbc.271.15.8875

Zal F, Lallier FH, Wall JS, et al (1996b) The multi-hemoglobin system of the hydrothermal vent tube worm *Riftia pachyptila* I.Reexamination of the number and masses of its constituents. J Biol Chem 271:8869–8874. doi: 10.1074/jbc.271.15.8869

Table 1. Siboglinid sample collection information.

Organism	Group	Collection Site	Sequencing Platform	Total Read Number
Escarpia spicata	Vestimentifera	N 28°11.58' W 89°47.94'	454 (Roche)	283,594
Galathealinum brachiosum	Frenulata	N 28°11.58' W 89°47.94'	454 (Roche)	456,440
Lamellibrachia luymesi	Vestimentifera	N 28°11.58' W 89°47.94'	454 (Roche)	750,876
Lamellibrachia luymesi	Vestimentifera	N 28°11.58' W 89°47.94'	HiSeq (Illumina)	50,537,812
Osedax mucofloris	*Osedax*	Artificial whale fall, near Bergen Norway	MiSeq (Illumina)	3,027,776
Osedax mucofloris	*Osedax*	Artificial whale fall, near Bergen Norway	HiSeq(Illumina)	56,067,578
Sclerolinum brattstromi	Monilifera	N 62°27.26', E 6°47.57'	HiSeq(Illumina)	44,207,372
Seepiophila jonesi	Vestimentifera	N 28°11.58' W 89°47.94'	454 (Roche)	382,144
Siboglinum ekmani	Frenulata	N 62°23.30', E 6°54.58'	HiSeq (Illumina)	21,397,136
Siboglinum fiordicum	Frenulata	N 60°16.17' E 5°05.53'	HiSeq (Illumina)	35,922,776

Table 2. GenBank accession numbers for hemoglobin and linker proteins. Novel sequences in bold.

Organism	Hb chain	Accession Number	Base Pair Length	Amino Acid Length
<u>Hemoglobin chains</u>				
Arenicola marina	A2	AJ880690	474	157
	B2	AJ880691	498	165
Escarpia spicata	**A2**	**KT166954**	**950**	**162**
	B1	**KT166953**	**316**	**106**
	B2	**KT166952**	**503**	**170**
Galathealinum brachiosum	**A2**	**KT166957**	**653**	**184**
	B1	**KT166956**	**810**	**165**
	B2	**KT166955**	**732**	**183**
Lamellibrachia luymesi	**A1**	**KT166959**	**988**	**165**
	A2	**KT166961**	**981**	**191**
	B1	**KT166960**	**664**	**216**
	B2	**KT166958**	**1255**	**169**
Lamellibrachia sp.	A1	AY273262	330	110
	A2	AY250084	210	70
	B1	AY273263	354	118
	B2	AY250085	213	71
Oasisia alvinae	A2	AY250087	228	76

	B2	AY273264	159	53
Oligobrachia mashikoi	A1	AB185392	551	156
	A2	AB185391	569	158
	B1	AB185394	851	183
Osedax mucofloris	**A1**	**KT166963**	**1024**	**148**
	A2	**KT166964**	**925**	**183**
	B1	**KT166962**	**469**	**188**
Ridgeia piscesae	B1	DQ414408	342	114
	B2	AY250083	255	85
Riftia pachyptila	A1	AJ439732	345	115
	A2	AJ439733	348	116
	B1	AJ439734	354	118
	B2	AJ439737	351	117
Sclerolinum brattstromi	**A1**	**KT166976**	**995**	**195**
	A2	**KT166977**	**1012**	**190**
	B1	**KT166978**	**898**	**192**
	B2	**KT166979**	**1121**	**196**
Seepiophila jonesi	**A1**	**KT166968**	**968**	**195**
	A2	**KT166967**	**1022**	**193**
	B1	**KT166965**	**903**	**210**
	B2	**KT166966**	**1079**	**153**
Siboglinum ekmani	**A1**	**KT166969**	**1188**	**165**
	A2	**KT166970**	**1089**	**191**

	B1	**KT166971**	**1289**	**188**
	B2	**KT166980**	**685**	**154**
Siboglinum fiordicum	**A1**	**KT166972**	**746**	**157**
	A2	**KT166973**	**635**	**146**
	B1	**KT166974**	**646**	**141**
	B2	**KT166975**	**738**	**146**
Tevnia jerichonana	A2	AY250086	264	88
Linker chains				
Alvinella pompejana	L1	CAJ00867	NA	225
	L2	CAJ00868	NA	212
	L3	CAJ00869	NA	158
Arenicola marina	L1	CAJ00866	NA	256
Lamellibrachia sp.	AV-1	P16222	NA	224
Riftia pachyptila	LX	CAJ00870	NA	141
	LY	CAJ00871	NA	182
	LZ	ABW24414	NA	120

Table 3. Averages of the BLASTX results of linker sequences from vestimentiferans and non-siboglinids to eight samples transcriptomes.

		BLAST		Percent AA
Sample	Reference	Score	E-value	Identity
Escarpia	Vestimentiferan	64.7	2.7e-07	61.3
spicata	Non-Siboglinid	37	1.0e-06	51.5
Galathealinum	Vestimentiferan	102.3	1.1e-06	35.3
brachiosum	Non-Siboglinid	22.5	2e-06	60
Lamellibrachia	Vestimentiferan	99	9.7e-07	52.5
luymesi	Non-Siboglinid	38.7	3.5e-06	50.8
Osedax	Vestimentiferan	108.8	1.3e-07	44.2
mucofloris	Non-Siboglinid	49.3	2.5e-06	41.1
Sclerolinum	Vestimentiferan	122.8	4.6e-07	50.3
brattstromi	Non-Siboglinid	42.3	7.7e-07	44.1
Seepiophila	Vestimentiferan	101.2	1.8e-07	59.4
jonesi	Non-Siboglinid	35.5	2.3e-06	53.3
Siboglinum	Vestimentiferan	76.4	2.0e-06	40.8
ekmani	Non-Siboglinid	71.6	9.1e-07	42.4
Siboglinum	Vestimentiferan	59.9	2.2e-06	41.2
fiordicum	Non-Siboglinid	48.5	1.5e-06	42.3

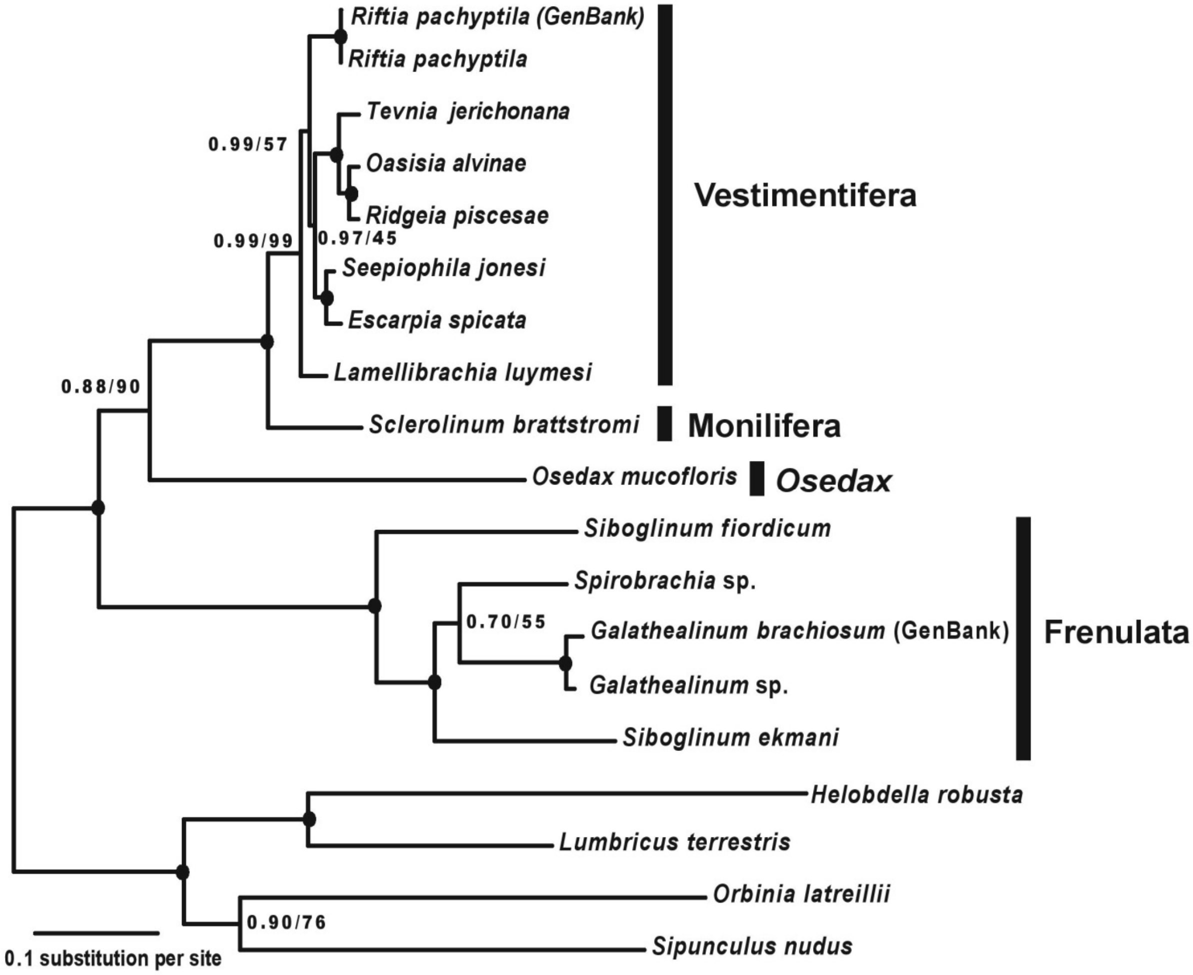

Figure 1. Current hypothesized phylogeny of Siboglinidae from Li et al. (2015).

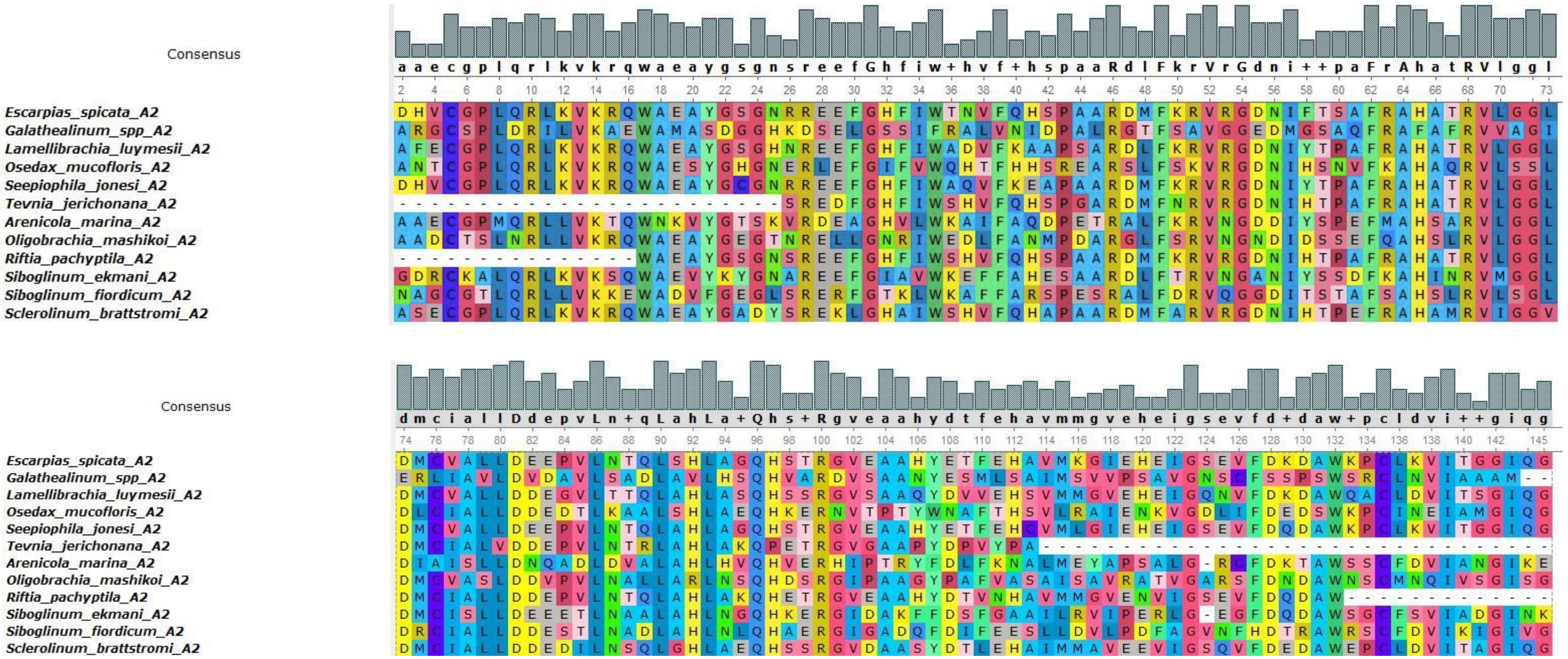
Consensus
a a e c g p l q r l k v k r q w a e a y g s g n s r e e f G h f i w + h v f + h s p a a R d l F k r V r G d n i + + p a F r A h a t R V l g g l
2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 73
Escarpias_spicata_A2 DHVCGPLQRLKVKRQWAEAYGSGNRREEFGHFIWTNVFQHSPAARDMFKRVRGDNIFTSAFRAHATRVLGGL
Galathealinum_spp_A2 ARGCSPLDRILVKAEWAMASDGGHKDSELGSSIFRALVNIDPALRGTFSAVGGEDMGSAQFRAFAFRVVAGI
Lamellibrachia_luymesii_A2 AFECGPLQRLKVKRQWAEAYGSGHNREEFGHFIWADVFKAAPSARDLFKRVRGDNIYTPAFRAHATRVLGGL
Osedax_mucofloris_A2 ANTCGPLQRLKVKRQWAESYGHGNERLEFGIFVWQHTFHHSREARSLFSKVRGDNIHSNVFKAHAQRVLSSL
Seepiophila_jonesi_A2 DHVCGPLQRLKVKRQWAEAYGCGNRREEFGHFIWAQVFKEAPAARDMFKRVRGDNIYTPAFRAHATRVLGGL
Tevnia_jerichonana_A2 ------------------------SREDFGHFIWSHVFQHSPGARDMFNRVRGDNIHTPAFRAHATRVLGGL
Arenicola_marina_A2 AAECGPMQRLLVKTQWNKVYGTSKVRDEAGHVLWKAIFAQDPETRALFKRVNGDDIYSPEFMAHSARVLGGL
Oligobrachia_mashikoi_A2 AADCTSLNRLLVKRQWAEAYGEGTNRELLGNRIWEDLFANMPDARGLFSRVNGNDIDSSEFQAHSLRVLGGL
Riftia_pachyptila_A2 ---------------WAEAYGSGNSREEFGHFIWSHVFQHSPAARDMFKRVRGDNIHTPAFRAHATRVLGGL
Siboglinum_ekmani_A2 GDRCKALQRLKVKSQWAEVYKYGNAREEFGIAVWKEFFAHESAARDLFTRVNGANIYSSDFKAHINRVMGGL
Siboglinum_fiordicum_A2 NAGCGTLQRLLVKKEWADVFGEGLSRERFGTKLWKAFFARSPESRALFDRVQGGDITSTAFSAHSLRVLSGL
Sclerolinum_brattstromi_A2 ASECGPLQRLKVKRQWAEAYGADYSREKLGHAIWSHVFQHAPAARDMFARVRGDNIHTPEFRAHAMRVIGGV
Consensus
d m c i a l l D d e p v L n + q L a h L a + Q h s + R g v e a a h y d t f e h a v m m g v e h e i g s e v f d + d a w + p c l d v i + + g i q g
74 76 78 80 82 84 86 88 90 92 94 96 98 100 102 104 106 108 110 112 114 116 118 120 122 124 126 128 130 132 134 136 138 140 142 145
Escarpias_spicata_A2 DMCVALLDEEPVLNTQLSHLAGQHSTRGVEAAHYETFEHAVMKGIEHEIGSEVFDKDAWKPCLKVITGGIQG
Galathealinum_spp_A2 ERLIAVLDVDAVLSADLAVLHSQHVARDVSAANYESMLSAIMSVVPSAVGNSCFSSPSWSRCLNVIAAAM--
Lamellibrachia_luymesii_A2 DMCVALLDDEGVLTTQLAHLASQHSSRGVSAAQYDVVEHSVMMGVEHEIGQNVFDKDAWQACLDVITSGIQG
Osedax_mucofloris_A2 DLCIALLDDEDTLKAALSHLAEQHKERNVTPTYWNAFTHSVLRAIENKVGDLIFDEDSWKPCINEIAMGIQG
Seepiophila_jonesi_A2 DMCVALLDEEPVLNTQLAHLAGQHSTRGVEAAHYETFEHCVMLGIEHEIGSEVFDQDAWKPCLKVITGGIQG
Tevnia_jerichonana_A2 DMCIALVDDEPVLNTRLAHLAKQPETRGVGAAPYDPVYPA--------------------------------
Arenicola_marina_A2 DIAISLLDNQADLDVALAHLHVQHVERHIPTRYFDLFKNALMEYAPSALG-RCFDKTAWSSCFDVIANGIKE
Oligobrachia_mashikoi_A2 DMCVASLDDVPVLNALLARLNSQHDSRGIPAAGYPAFVASAISAVRATVGARSFDNDAWNSCMNQIVSGISG
Riftia_pachyptila_A2 DMCIALLDDEPVLNTQLAHLAKQHETRGVEAAHYDTVNHAVMMGVENVIGSEVFDQDAW-------------
Siboglinum_ekmani_A2 DMCISLLDEEETLNAALAHLNGQHKERGIDAKFFDSFGAAILRVIPERLG-EGFDQDAWSGCFSVIADGINK
Siboglinum_fiordicum_A2 DRCIALLDDESTLNADLAHLNLQHAERGIGADQFDIFEESLLDVLPDFAGVNFHDTRAWRSCFDVIKIGIVG
Sclerolinum_brattstromi_A2 DMCIALLDDEDILNSQLGHLAEQHSSRGVDAASYDTLEHAIMMAVEEVIGSQVFDEDAWEPCLDVITAGIQG

Figure 2. Amino acid alignment of chain A2 for siboglinids. Alignment was generated in MEGA 5.2 using MUSCLE and visualized using UniPro UGENE (Okonechnikov et al. 2012). Bars at the top of the alignment show percentage of conserved amino acid for that position. Conserved free cysteine at position 76 shown in purple.

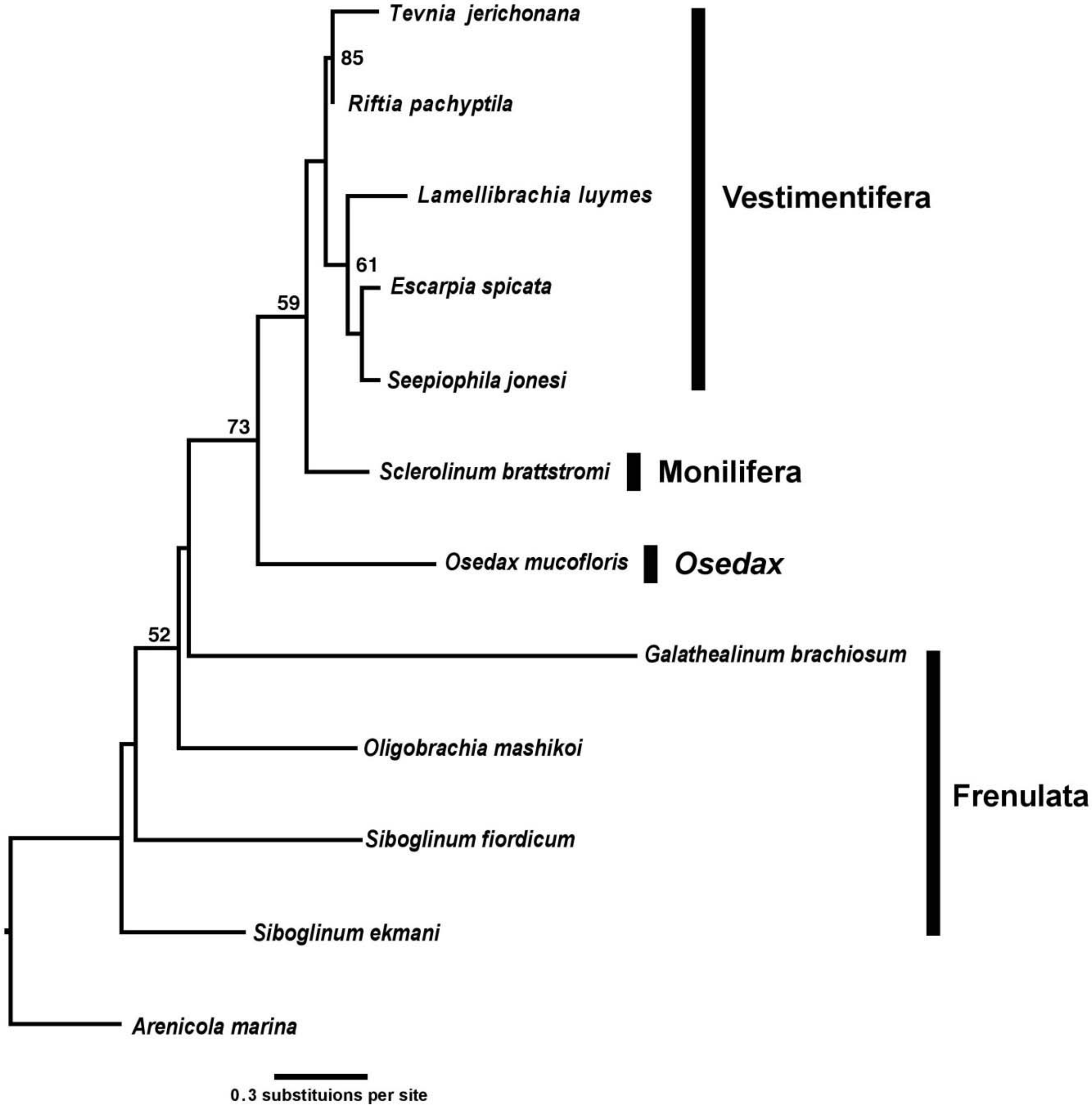

Figure 3. Hemoglobin chain A2 gene maximum likelihood tree reconstructed with RAxML using the PROTGAMMAWAG model. The optimal topology had a –ln Likelihood of -1995.806816. Bootstrap support values greater than %50 are shown at the relevant node.

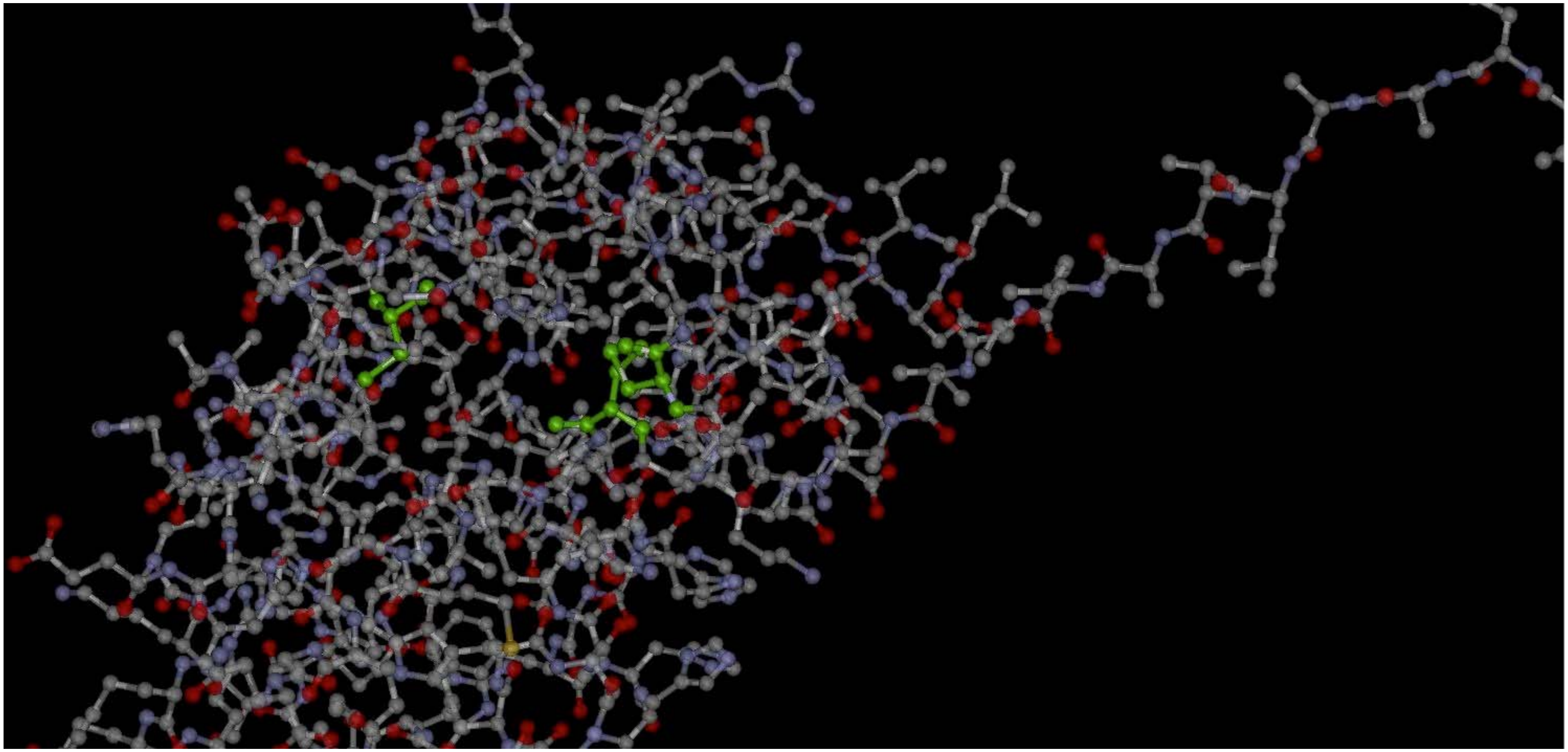

Figure 4. 3D visualization of *Osedax mucofloris* A2 chain protein structure using the RaptorX protein structure prediction server. Cysteine residues are highlighted in green.

Chapter 2. Filling the Cracks: Two Programs for Phylogenetic Pipelines

2.1 Introduction

With the prevalence of high throughput sequencing data, researchers are faced with the problem of handling massive amounts of data in an efficient way. Many tools have been developed for use in genomic projects that lead from raw reads to final phylogenetic topology, such as Trinity for *de novo* transcriptome assembly (Grabherr et al. 2011), Ray for *de novo* genome assembly (Boisvert et al. 2012), Transdecoder for gene translation (Haal et al. 2013), HaMStR for ortholog prediction (Ebersberger et al 2009), a suite of sequence aligners (Edgar 2003; Katoh and Standly 2013; Sivers and Higgins 2014), and multiple programs for topology predictions such as RAxML (Stamatakis 2006). Programs like these are workhorses in pipelines that facilitate fast, efficient execution of procedures that are common to most phylogenomic projects (Dunn et al. 2008; Kocot et al 2011). The cracks between these programs must be filled based on intentions of the researcher and the scope of the project at hand. Trimming of missing data, contamination screening, and other small tasks are executed at the discretion of the researcher and the quality of the data.

Our research group has developed a phylogenomic pipeline (Fig. 1) for inference of various phylogenetic questions (Cannon et al. 2014, Kocot et al. submitted). I have developed two programs for inclusion in this pipeline with specifications given by researchers. The first program, ContamScreen, is a tool that removes sequence contamination from transcriptome or genome assemblies. The second program, Alignment_Compare, is an alignment trimmer that removes sequences that do not align with every other sequence in the dataset by at least 20 amino acids. Both these programs can be accessed at github.com/DamienWaits.

2.2 Details

ContamScreen is a tool designed to detect and isolate sequence contamination from assembled transcriptomes or genomes based on reference datasets. Contamination may occur from several sources or lab errors. Prey items remaining in the stomach or parasites that cannot be removed due to the necessity of using whole smaller organisms for extraction may be causes of contamination. Human error during any point in sample preparation also may introduce contamination before sequencing even occurs. Even sequencing facilities may have errors even though they try to ensure that their data remains free of contamination. A recent study found that out of 202 viral and bacterial metagenomes, 145 contained human contamination with some having up to 64% of reads belonging to humans (Schmeider and Edwards 2011). Bleed-through is also a possible on Illumina platforms where reads can be misattributed within a lane on the flowcell as sequencing occurs (pers. obs.).

Due to these types of complications, the following program was written to help filter contamination of sequence data from organisms that are not the sample of interest to help ensure correct phylogenetic inference. The program BLAST is utilized to determine similarity of target sequences to good vs. contaminant reference sequences, which are then output to corresponding files (Altschul et al. 1990). Two files are supplied by the user and renamed as good.fasta and contaminated.fasta. These files are given headers corresponding to their identity and merged. This file is then used to generate a BLAST database. The target assembly is searched using the BLAST database and a BLAST report is generated in a table format. This table is parsed line-by-line and all hits for each contig are compared pairwise to determine the best hit. The header for the best hit is used to determine whether the contig in question is “good” or “contam”.

2.2 ContamScreen

This program was written in bash. Explanations of the script are given prior to commands and are denoted by a “#” symbol. The expected input is a target transcriptome or genome in a fasta file, a file containing sequences belonging to closely related species of your target samples (parameter 1), and a file containing sequences belonging to possible contaminants such as parasites or prey items (parameter 2), how many orders of magnitude a “contaminant” hit must be greater than a “good” hit to be “contaminant” (parameter 3), and how many orders of magnitude a “good” hit must be greater than a “contaminant” hit to be “good” (parameter 4).

```
#This program screens a given assembly or assembly for contamination based on two
#reference databases supplied by the user as fasta files, parameter 1 and parameter 2.
#Parameter 1 should include sequences that are similar to those of the targeted assembly.
#Parameter 2 should include sequences of known or suspected contaminants. These files
#are merged and used to generate a BLAST database which is run on the target assembly.
#The best hit for each contig in the assembly is chosen based on e-value score comparison.
#E-values are compared based on order of magnitude and the sensitivity of this comparison
i#s decided by the user. Parameter 3 should be an integer and corresponds to how much
#better a “contam” hit must be than the best “good” hit to be considered contaminated.
#Example: If parameter 3 is input as 5 and the best “good” hit for Contig1 is 1e-10 and the
#best “contam” hit is 1e-16, Contig1 is returned as “contaminated”. If the best “contam”
#hit were 1e-13, Contig1 would be returned as “suspect” for manual evaluation. Parameter
#4 is similar except it is how much better a “good” hit must be than a “contam” hit. Three
#files are generated as output: TAXON-good.fasta, TAXON-contam.fasta, TAXON-
#suspect.fasta. BLAST output files can be found in the misc_intermediate_files directory.

#This program was developed by Damien S Waits and Kevin M Kocot. Version July 3rd,
#2015

#!/bin/bash
#Creates a directory where intermediate output files are stored.
mkdir "misc_intermediate_files"

#Appends “good” to headers in the good file and writes “contam”
#to headers in the contaminatant file.
sed -i 's/^>/>good\ /' $1 > good.fasta
sed -i 's/^>/>contam\ /' $2 > contaminated.fasta

```

```
#Writes both good.fasta and contaminated.fasta to an all.fasta
#and generates a blast database since e-values are not
#informative across multiple blast searches.
cat good.fasta contaminated.fasta > all.fasta
makeblastdb -in all.fasta -dbtype nucl -title ALL -out ALL

#Performs the following actions on all files that have the
#suffix .fa until “done” is read.
for FILENAME in *.fa
do

#Removes line breaks from taxon.fa
#using nentferner.pl distributed with HaMStR
nentferner.pl -in=$FILENAME -out=$FILENAME".nent"
rm $FILENAME
mv $FILENAME".nent" $FILENAME

#Removes unneeded header text that will make blast output
#unnecessarily long.
sed -i '/^>/ s/ .\+//g' $FILENAME
sed -i '/^>/ s/;.\+//g' $FILENAME

#Deletes sequences shorter than 100 amino acids in length.
grep -B 1 "[^>].\{100,\}" $FILENAME > $FILENAME".tmp"
#Deletes blank lines.
sed -i '/^$/d' $FILENAME".tmp"
#Deletes original file.
rm -rf $FILENAME
#Restores original file.
mv $FILENAME".tmp" $FILENAME

#Sets variable "taxon" equal to the part of the file name before
#the extension.
taxon=`echo $FILENAME | cut --delimiter=. --fields=1`

#Performs blast search on assembly using the ALL database that
#includes both good and contam sequences, and formats output to
#a table.
blastn -db ALL -query $FILENAME -num_descriptions 10 -
num_alignments 10 -num_threads 6 > $taxon"_vs_all.txt"
blast2table2.pl -format 10 -evalue 0.0001 $taxon"_vs_all.txt" >
$taxon"_vs_all.table"

#Formats e-values to allow for comparison.
sed 's/[1-9]e-0*//' $taxon"_vs_all.table" > temp.table
sed -i 's/0.0/0/' temp.table

```

```
#Loops through all lines in the table format of the blast
#output.
while read line;
do

#Returns the sixth column of the table.
thisLine=`echo $line | awk '{print $6}'`

#The following nested if statements compares the e-values of
#hits to the same contig in the targeted assembly. Contaminant
#or true sequences are determined based on the e-value of the
#hit. If the hit is within a threshold determined by the end
#user, it is instead output to a file of suspect sequences.
if [ "$lastLine" == "" ]
then
    lastLine=$thisLine
    currentValue=`echo $line | awk '{print $2}'`
    currentState=`echo $line | awk '{print $12}'`
    suspect=false
    if [ "$currentValue" -eq "0" ]
    then
        perfect=$currentState
        perfectLine=$thisLine
    fi
elif [ "$thisLine" == "$lastLine" ]
then
    newValue=`echo $line | awk '{print $2}'`
    newState=`echo $line | awk '{print $12}'`
    if [ "$newValue" -eq "0" ]
    then
        perfect=$newState
        perfectLine=$thisLine
    fi
    if [ "$currentState" == "contam" ] && [ "$newState" ==
"good" ]
    then
        if [ "$newValue" -gt "$((currentValue+$3))" ]
        then
            currentState=`echo $line | awk '{print $12}'`
            currentValue=$newValue
            suspect=false
        elif [ "$newValue" -gt "$((currentValue-$4))" ]
        then
            suspect=true
        fi
    elif [ "$currentState" == "good" ] && [ "$newState" ==
"contam" ]
```

```
    then
        if [ "$newValue" -gt "$((currentValue+$4))" ]
        then
            currentState=`echo $line | awk '{print $12}'`
            currentValue=$newValue
            suspect=false
        elif [ "$newValue" -gt "$((currentValue-$3))" ]
        then
            suspect=true
        fi
    else
        if [ "$newValue" -gt "$currentValue" ]
        then
            currentValue=$newValue
            currentState=`echo $line | awk '{print $12}'`
        fi
    fi
else
    if [ "$perfect" != "" ]
    then
        echo "$perfectLine" >> $perfect"_headers.txt"
    elif [ "$suspect" ]
    then
        echo "$lastLine" >> suspect_headers.txt
    elif [ "$currentState" == "good" ]
    then
        echo "$lastLine" >> good_headers.txt
    elif [ "$currentState" == "contam" ]
    then
        echo "$lastLine" >> contam_headers.txt
    fi
    lastLine=$thisLine
    currentValue=`echo $line | awk '{print $2}'`
    currentState=`echo $line | awk '{print $12}'`
    suspect=false
fi

done <  $taxon"_vs_all.table"

#Writes out the name of each contig to either good or contam
#headers files depending on the comparisons from the above if
#statement.
if [ "$perfect" != "" ]
then
        echo "$perfectLine" >> $perfect"_headers.txt"
elif  $suspect
then
```

```
        echo "$lastLine" >> suspect_headers.txt
elif [ $currentState == "good" ]
then
        echo "$lastLine" >> good_headers.txt
elif [ $currentState == "contam" ]
then
        echo "$lastLine" >> contam_headers.txt
fi

#Removes redundant sequence headers.
sort good_headers.txt | uniq > good_headers_sorted.txt
sort suspect_headers.txt | uniq > suspect_headers_sorted.txt
sort contam_headers.txt | uniq > contam_headers_sorted.txt

#Extracts the sequences that belong to the headers in the above
#files.
select_contigs.pl -n good_headers_sorted.txt $FILENAME
$taxon"-good.fasta"
select_contigs.pl -n contam_headers_sorted.txt $FILENAME
$taxon"-contam.fasta"
select_contigs.pl -n suspect_headers_sorted.txt $FILENAME
$taxon"-suspect.fasta"

#Clean-up of garbage files.
rm -f good_headers.txt
rm -f contam_headers.txt
rm -f suspect_headers.txt
rm -f temp.table
mv $taxon"_vs_all.txt" ./misc_intermediate_files/
mv $taxon"_vs_all.table" ./misc_intermediate_files/
done
```

Testing of this program has shown almost perfect results. Four transcriptomes were seeded with artificial “contaminant” data in the form of sequences from a transcriptome outside the phylum of the target , and ran through the contamination screening program, with the original transcriptomes used as the “good” database, and the seeded data used as the “contaminant” database. This resulted in 99.99% accurate detection of “good” sequences, and 99.98% accurate detection of “contaminant” sequences.

2.4 Alignment_Compare

Sequence contamination is a possibility in almost all datasets; however some pipelines require very specific procedures that are not necessarily shared across other pipelines. In the pipeline our research group developed, core orthologous sequences are mined from the transcriptome or genome and aligned. Previously this alignment was manually trimmed of sequences that did not align with all other sequences by at least 20 amino acids. To streamline this process, the following Java program was developed. The expected input is an alignment in a fasta or simple text format. The program compares positions in each line of the alignment and removes poorly aligned sequences by determining which sequence does not align with the most sequences by at least 20 amino acids. Explanations of code are indicated by "//" symbols before the block of code. Continuation of a line of code is denoted by an indentation.

```
/*This program parses through an alignment and flags each
sequences that does not align with any other sequences by 20
amino acids. Upon finding a non-aligning sequence, it counts the
number of sequences that do not overlap with it by at least 20.
Upon parsing through the entire alignment, the sequences with
the highest number of counts, is deleted. The program then
cycles back through the alignment and recounts. It continues to
do this until all sequences align with all other sequences by at
least 20 amino acids. Ties are broken by deleting the shortest
sequence in the tie.

This program was developed by Damien S. Waits with
specifications and design input from Kevin M. Kocot. Version
July 3rd, 2015. */

import java.io.*;
import java.util.ArrayList;

public class AlignmentCompare {

public static void main(String[] args) throws IOException {

boolean notFinished = true;
String line1;
String line2;
String writeLine;
int overlaps;
int badLine;
int i;
int j;

//Sanity check to determine if the correct parameters were
input.
if (args[0] == null) {
System.out.print("Please give an input file.");
System.exit(0);
}
//Runs until each line overlaps with every other by at least 20.
while (notFinished) {
File alignmentFilename = new File(args[0]);
File tempFilename = new File("myTempFile.txt");

//Reader for iterating through lines to be compared.
BufferedReader AlignmentReader1 = new BufferedReader(new
     FileReader(alignmentFilename));
//Reader for going through remaining lines to compare to Reader1
//lines.
BufferedReader AlignmentReader2 = null;
BufferedWriter AlignmentWriter = new
BufferedWriter(new FileWriter(tempFilename));
line1 = null;
line2 = null;
writeLine = null;
overlaps = 0;
badLine = 0;
i = 0;
j = 0;

//ArrayList of ArrayList that holds the line numbers that don't
//overlaps for each line.
ArrayList<ArrayList<Integer>> noOverlaps = new
     ArrayList<ArrayList<Integer>>(0);
//Stores the number of amino acids in each line.
ArrayList<Integer> lineLengths = new ArrayList<Integer>(0);
int lineToDelete;

//While loop for comparisons of 1 line to each other line.
while ((line1 = AlignmentReader1.readLine()) != null) {
//ArrayList to keep track of the line numbers that don't overlap
//will with the line currently being looked at.
```

```
ArrayList<Integer> currentNoOverlaps = new
     ArrayList<Integer>(0);
AlignmentReader2 = new BufferedReader(new
     FileReader(alignmentFilename));
//If the current line is a header, skip it.
if (line1.charAt(0) == '>') {
line1 = AlignmentReader1.readLine();
i++;
}
//Line 1 shouldn't be compared to line 1. Move down to the line
//after the line that Reader1 just read.
j = i;
for (int inc = 0; inc < i; inc++) {
@SuppressWarnings("unused")
String temp =  AlignmentReader2.readLine();
}
//Here we use Reader2 to start reading in all lines after
//Reader1's current line for comparison.
while ((line2 = AlignmentReader2.readLine()) != null)
{
//Skip headers
if (line2.charAt(0) == '>') {
line2 = AlignmentReader2.readLine();
j++;
}
overlaps = 0;
//Counting overlaps.
for (int k = 0; k < line1.length() && k < line2.length(); k++) {
//If both chars at position k aren't gaps, increment.
if (checkForGaps(line1, k) == -1 && checkForGaps(line2, k) ==
     -1) {
overlaps++;
}
}
//If we don't have more than 20 overlaps, add this line to our
//list.
if (overlaps <= 20) {
currentNoOverlaps.add(j);
}
j++;
}
//Reader2 has iterated through the whole file. Close reader2.
AlignmentReader2.close();
//Get the length of this line in amino acids.
lineLengths.add(findLength(line1, i));
//Add this line's results to the ArrayList of ArrayLists.
noOverlaps.add(currentNoOverlaps);
```

```
i++;
}
//Reader1 has iterated through the whole file. Close
reader1.  AlignmentReader1.close();
//To determine which line is the worst and should be deleted, we
//count the number of lines that don't overlap with it.
int[] overlapCounts;
overlapCounts = new int[noOverlaps.size()];

for (int l = 0; l < noOverlaps.size(); l++) {
for (int m = 0; m < noOverlaps.get(l).size(); m++) {
badLine = noOverlaps.get(l).get(m);
if (badLine != 0) {
badLine = (badLine + 1) / 2;
//Since comparisons don't go backwards(1 was recorded as not
//overlapping with 2, but not vice versa) we need to account for
//this.
overlapCounts[badLine - 1]++;
overlapCounts[l]++;
}
}
}
//Use findLargestAndSmallest to determine which line has the
//most lines that do not overlap with it.
lineToDelete = findLargestAndSmallest(overlapCounts,
     lineLengths);

//Stop when all lines overlap.
if (lineToDelete == 0) {
notFinished = false;
tempFilename.delete();
break;
}
//Start readering in file for writing.
AlignmentReader1 = new BufferedReader(new
     FileReader(alignmentFilename));

// Find the lineToDelete and don't write it.
i = 1;
while ((writeLine = AlignmentReader1.readLine()) != null) {
//Doubling lineToDelete to account for headers.
if ((i != (lineToDelete * 2)) && ((i + 1) != (lineToDelete *
     2))) {
AlignmentWriter.write(writeLine);
AlignmentWriter.newLine();
}
i++;
```

```
}

AlignmentReader1.close();
AlignmentWriter.close();
tempFilename.renameTo(alignmentFilename);
}
}
//Method to check if the current char is a gap.
public static int checkForGaps(String line, int inc) {
String badChars = "-X?";
int check = badChars.indexOf(line.charAt(inc));
return check;
}

//Determines which line has the most non-overlapping lines and
//returns it.
public static int findLargestAndSmallest(int[] integers,
     ArrayList<Integer> lengths) {
boolean conflict = false;
int largestNonOverlaps = 0;
int smallestLength = 0;
int position = 0;
//If two lines have the same number of non-overlapping lines,
//this ArrayList holds their positions.
ArrayList<Integer> choices = new ArrayList<Integer>(1);

//Checking which has the most non-overlaps.
for (int i = 0; i < integers.length; i++) {
if (integers[i] > largestNonOverlaps) {
conflict = false;
choices.clear();
choices.add(i + 1);
largestNonOverlaps = integers[i];
//Save which line has the most non-overlaps.
position = i + 1;
//If equal to the current line with most non-overlaps, add to
//choices without clearing it.
} else if (integers[i] == largestNonOverlaps) {
conflict = true;
choices.add(i + 1);
}
}
//If there are no non-overlaps, we're done.
if (largestNonOverlaps == 0)
return 0;
//If two lines have the same amount of non-overlapping lines,
//find the one with the smallest number of amino acids.
```

```
if (conflict) {
smallestLength = lengths.get(choices.get(0));
position = choices.get(0);
for (int j = 1; j < choices.size(); j++) {
if (lengths.get((choices.get(j))-1) < smallestLength) {
smallestLength = lengths.get((choices.get(j))- 1);
position = choices.get(j);
}
}
}
//Returns the line number of line with the most non
//overlaps/smallest length of amino acids.
return position;
}

//Finds the length in amino acids of a line.
public static int findLength(String line, int position) {
int length = 0;
for (int i = 0; i < line.length(); i++) {
if (checkForGaps(line, i) == -1)
length++;
}
return length;
}
}
```

Testing with the alignment-trimming program showed that the program worked as intended and trimmed out sequences that do not overlap with every other sequence by at least 20 amino acids, with sequence length breaking ties of equally poor aligning sequences.

2.5 Conclusion

There are limitations to ContamScreen. These include lengthy runtime and the necessity for prior knowledge of contaminants. Execution of the contamination script can take up to 4 days, depending on how large of a dataset is used. This runtime was seen during testing when one transcriptome for *Saccoglossus mereschkowskii* was seeded with another small transcriptome, *Osedax mucofloris* for a final size of 163 megabytes and each transcriptome was used as the databases (156 megabyte “good” database and 9.7 “contam” database). Optimally, a reference genome for the suspected contaminant and the sample of interest would be used, however housekeeping genes from closely related species should be sufficient to detect the presence of contamination. This leads to a trade-off: more robust reference databases will allow for more accurate detection of contamination but will also increase runtime. The program also relies on the end-user to supply appropriate reference and contaminant databases. Unknown contamination will most likely go undetected. Therefore, a proper working knowledge of the sequenced organism is needed. Prey items and parasites should be known. If sequence bleed-through is suspected or human error possibly caused contamination, databases should include sequences from humans and samples from adjacent sequencing lanes.

High throughput sequencing platforms have opened up a world of new opportunities for researchers. However, with the prevalence of this data, more and more tools are required for analysis. Without these tools, data for phylogenetic inference would take a prohibitively large amount of time to prepare. Programs described above help to alleviate problems associated with errors in sequencing and allow for more efficient uses of time due to automation of alignment trimming.

2.6 References

Altschul SF, Gish W, Miller W, et al (1990) Basic local alignment search tool. Journal of Molecular Biology 215:403–410. doi: 10.1016/S0022-2836(05)80360-2

Boisvert S, Raymond F, Godzaridis É, et al (2012) Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biology 13:R122. doi: 10.1186/gb-2012-13-12-r122

Cannon JT, Kocot KM, Waits DS, et al (2014) Phylogenomic Resolution of the Hemichordate and Echinoderm Clade. Current Biology 24:2827–2832. doi: 10.1016/j.cub.2014.10.016

Dunn CW, Hejnol A, Matus DQ, 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, Haeseler A von (2009) HaMStR: Profile hidden markov model based search for orthologs in ESTs. BMC Evolutionary Biology 9:157. doi: 10.1186/1471-2148-9-157

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research 32:1792–1797.

Grabherr MG, Haas BJ, Yassour M, et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotech 29:644–652. doi: 10.1038/nbt.1883

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution 30:772–780.

Kocot KM, Cannon JT, Todt C, et al (2011) Phylogenomics reveals deep molluscan relationships. Nature 477:452–456. doi: 10.1038/nature10382

Kocot KM (2013) A combined approach toward resolving the phylogeny of Mollusca. (Doctoral Dissertation) Auburn University, Auburn, AL. Retrieved from http://etd.auburn.edu/handle/10415/3581

Schmieder R, Edwards R (2011) Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets. PLoS One. doi: 10.1371/journal.pone.0017288

Sievers F, Higgins DG (2014) Clustal Omega, accurate alignment of very large numbers of sequences. In: Multiple sequence alignment methods. Springer, pp 105–116

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcosts.

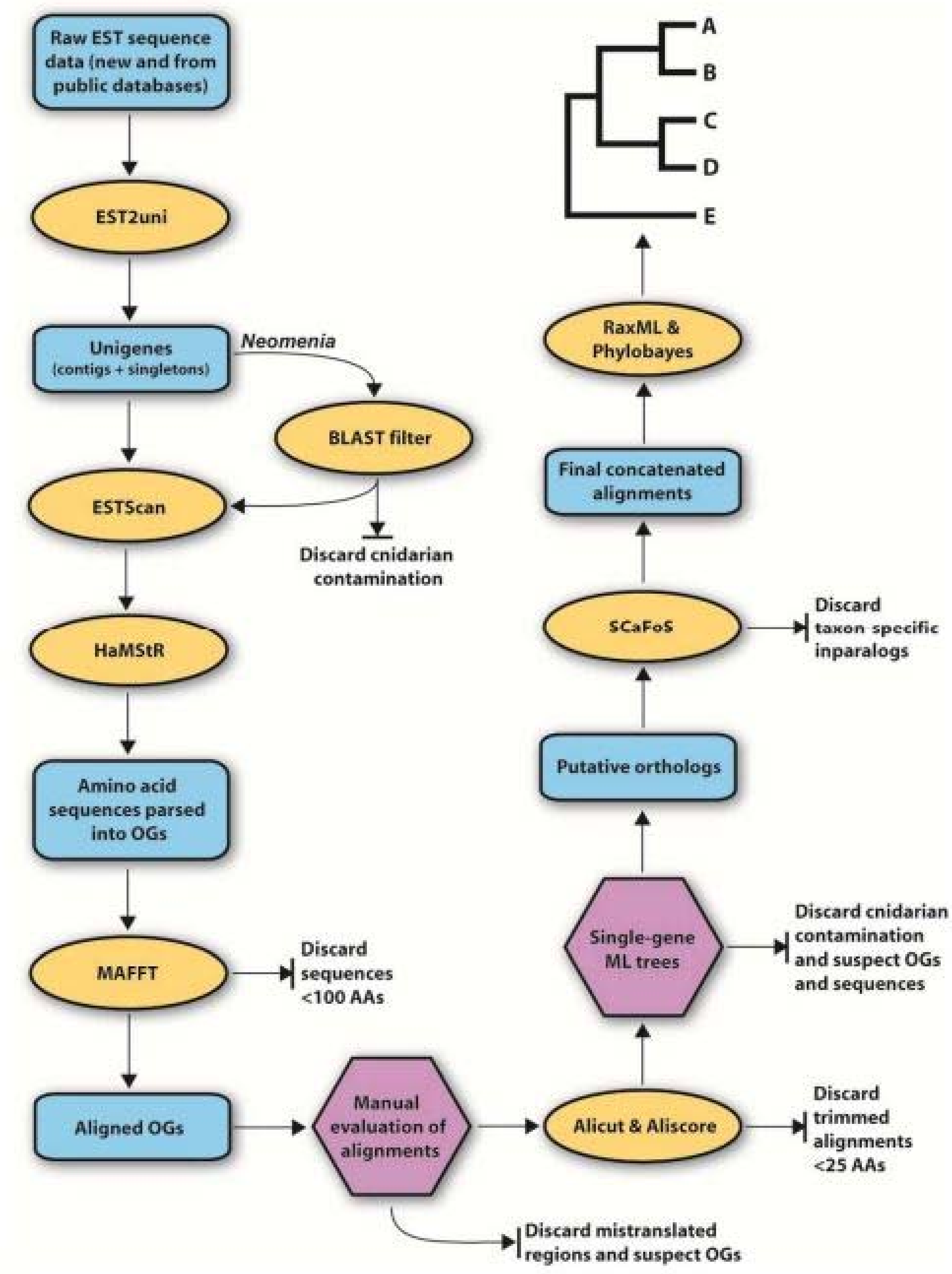
Raw EST sequence data (new and from public databases)
EST2uni
Unigenes (contigs + singletons)
Neomenia
BLAST filter
ESTScan
Discard cnidarian contamination
HaMStR
Amino acid sequences parsed into OGs
MAFFT
Discard sequences <100 AAs
Aligned OGs
Manual evaluation of alignments
Discard mistranslated regions and suspect OGs
Alicut & Aliscore
Discard trimmed alignments <25 AAs
Single-gene ML trees
Discard cnidarian contamination and suspect OGs and sequences
Putative orthologs
SCaFoS
Discard taxon specific inparalogs
Final concatenated alignments
RaxML & Phylobayes
A
B
C
D
E

Figure 2. Workflow of the phylogenomic pipeline designed by our research group. ContamScreen takes the place of “BLAST filtering”, and Alignment_Score replaces “Manual evaluation of alignments”. (Kocot 2013)