Submitted to MBE as Research Article

Mitochondrial Genomes of *Clymenella torquata* (Maldanidae) and *Riftia pachyptila* (Siboglinidae): Evidence for Conserved Gene Order in Annelida

Robert M. Jennings[*] *and Kenneth M. Halanych*[*,†]

* Biology Department, Woods Hole Oceanographic Institution, Woods Hole, MA 02543 USA

email: rjennings@whoi.edu

† Department of Biological Sciences, Auburn University, Auburn, Alabama 36849 USA

email: ken@auburn.edu

Running Title: Annelid mtDNA

Key Words: Phylogeny, gene order, mitochondria, genome, Annelida

Contact information:

Ken Halanych
Department of Biological Sciences
101 Life Science Building
Auburn University, AL 36849
Phone: (334) 844-3222
Fax: (334) 844-2333
Email: ken@auburn.edu

Abstract

Mitochondrial genomes are useful tools for inferring evolutionary history. However, many taxa are poorly represented by available data. Thus, to further understand the phylogenetic potential of complete mitochondrial genome sequence data in Annelida (segmented worms), we examined the complete mitochondrial sequence for *Clymenella torquata* (Maldanidae) and an estimated 80% of the sequence of *Riftia pachyptila* (Siboglinidae). These genomes have remarkably similar gene orders to previously published annelid genomes, suggesting that gene order is conserved across annelids. This result is interesting given the high variation seen in the closely related Mollusca and Brachiopoda. Phylogenetic analyses of DNA sequence, amino acid sequence and gene order all support the recent hypothesis that Sipuncula and Annelida are closely related. Our findings suggest that gene order data is of limited utility in annelids but that sequence data holds promise. Additionally, these genomes show AT bias (~66%) and codon usage biases, but have a typical gene complement for bilaterian mitochondrial genomes.

INTRODUCTION

Sequencing of complete mitochondrial genomes has become a useful tool for inferring animal phylogeny (e.g. Boore and Brown 1998; Lavrov, Brown, and Boore 2004; Helfenbein and Boore 2004). The haploid, non-recombining properties of animal mitochondrial DNA (mtDNA), coupled with its small size, make it a logical choice when considering phylogenetic events. Determination of the entire mitochondrial genome sequence provides several suites of characters for phylogenetic analysis; for example, DNA gene sequences (rRNA, tRNA, and protein-encoding), inferred amino acid sequences of protein-encoding genes, and the arrangement of genes in the genome. However, there is considerable disparity in taxonomic sampling. Chordata accounts for 75% of the published animal mitochondrial genomes and Arthropoda represents the next 12.5%. Thus, there is still much to learn about how mitochondria evolve in many animal lineages.

Despite the importance of Annelida (segmented worms) with over 12,000 described species and its dominance as the most abundant macrofaunal group in the deep sea (69% of the planet), only two complete annelid mitochondria have been sequenced (the nereid *Platynereis dumerilii* and the oligochaete *Lumbricus terrestris*). These genomes differ only slightly in gene order. In addition, partial genomes of the siboglinid *Galathealinum brachiosum* and the leech *Helobdella robusta* (Boore and Brown 2000), match the *L. terrestris* gene order exactly. [Note that Siboglinidae was previously referred to as Pogonophora and Vestimentifera (McHugh 1997; Rouse and Fauchald 1997; Halanych et al. 2001).] Some mtDNA genome data is available for allied Lophotrochozoan taxa; most relevant are mollusks (e.g., Hoffman, Boore, and Brown

1992; Boore and Brown 1994; Hatzoglou, Rodakis, and Lecanidou 1995; Terrett, Miles, and Thomas 1996; Wilding, Mill, and Grahame 1999; Kurabayashi and Ueshima 2000; Grande et al. 2002; Tomita et al. 2002; Serb and Lydeard 2003; Boore, Medina, and Rosenberg 2004; Dreyer and Steiner 2004; DeJong, Emery, and Adema 2004), brachiopods (Stechmann and Schlegel 1999; Noguchi et al. 2000; Helfenbein, Brown and Boore 2001), phoronids (Helfenbein and Boore 2004) and sipunculans (Boore and Staton 2002). Of these taxa, the sipunculan *Phascolopsis gouldii* is the most similar to the known annelid arrangements with 16 of the 19 sipunculan genes examined in the same order as in *L. terrestris* (but in two separate blocks). For this reason, Boore and Staton (2002) hypothesized a close relationship between annelids and sipunculans. Mollusks are notable because their mitochondrial genomes appear to have experienced numerous large-scale rearrangements and some taxa have even lost the *atp8* gene. Brachiopods also seem to have undergone numerous rearrangements. Of the three complete genomes currently available, *Laqueus rubellus* and *Terebratalia transversa* share 14 gene boundaries composed in 9 blocks; *L. rubellus* and *Terebratulina retusa* share only 8 gene boundaries in 8 separate blocks (Helfenbein, Brown, and Boore 2001).

Recent views of annelid phylogeny have moved away from the traditional view of two main groups, Clitellata (Oligochaetes and Hirudineans) and Polychaeta. Although morphological cladistic analyses have supported this hypothesis (Rouse and Fauchald 1995), multiple sources of data clearly show that the Clitellata, Echiuridae, and Siboglinidae are within the polychaete radiation (reviewed in McHugh 2000; Halanych, Dahlgren, and McHugh 2002; Halanych 2004). Such potential for morphological adaptation is not surprising given the enormous amount of diversity in annelids' body

plans, habitats, and life histories. A comprehensive molecular phylogeny of Annelida is wanting, and currently our best understanding of annelid evolutionary history comes from morphological cladistic analyses (Rouse and Fauchald 1997; Rouse and Pleijel 2001), which suggest Annelids contain three major groups, Scolecida, Aciculata, and Canalipalpata. Unfortunately, the Clitellata are not considered in these treatments.

We report here the complete mitochondrial sequence of a bamboo worm *Clymenella torquata* (Maldanidae) and an estimated 80% of the genome of the deep-sea tubeworm *Riftia pachyptila* (Siboglinidae). *Clymenella torquata* and the other members of Maldanidae are called bamboo worms because the shape of their segments gives them a bamboo-like appearance. *Clymenella torquata* is common in sandy intertidal/subtidal estuaries of the Atlantic U.S. coast, where it builds tubes from the surrounding sand and ingests sediment and the associated interstitial organisms (Mangum 1964). *Riftia pachyptila* inhabits the hydrothermal vents of the East Pacific Rise, and obtains energy from the chemosynthetic endosymbiotic bacteria in a specialized structure called the trophosome (Southward and Southward 1988). Although annelid phylogeny has not been well resolved, available molecular evidence (Halanych, unpublished) places these two annelids in very distant parts of the annelid tree. By including these two taxa, we provide representatives for all major clades outlined by Rouse and Fauchald (1997). Our goals in presenting and analyzing these new genomes are 1) to further characterize the evolution of mitochondrial genome structure among annelids and 2) to explore the potential of mitochondrial genomes in resolving annelid phylogeny.

METHODS

Organisms

Clymenella torquata and *Riftia pachyptila* were chosen to obtain better representation of annelid diversity than is currently available for mitochondrial genomes. *C. torquata* is in Maldanidae within Scolecida and *R. pachyptila* is in Siboglinidae within Canalipalpata. When combined with the available annelid genomes from GenBank (see Table 1), all of the major clades of Annelida are represented (see McHugh 2000; Rouse and Fauchald 1997; Rouse and Pleijel 2001). All of the genome of *C. torquata* and two-thirds of the *R. pachyptila* genome presented here were sequenced from total DNA extractions of a single individual of each species; the remaining *R. pachyptila* sequence reported herein came from a second individual. *C. torquata* was collected in 2002 from Hyannisport, MA (N 41°37'57.9" W 70°19'18.3"). The two *R. pachyptila* were collected in 2000 at 2500m depth near the Tica vent at 9°N on the East Pacific Rise (N 9°50'26.8", W 104°17'29.6"). All organisms were frozen at -80°C after collection.

DNA Extraction and mtDNA sequencing

Total genomic DNA was extracted from approximately 25mm^3 tissue using the DNEasy kit (Promega) according to manufacturer's protocols. Throughout this paper, gene nomenclature and abbreviations follow Boore and Brown (2000): *cox1-3* refer to cytochrome oxidase c subunits 1-3, *nad1-6* (incl. *4L*) refer to NADH dehydrogenase subunits 1-6, *atp6* and *atp8* refer to ATPase F0 subunits 6 and 8, and *cob* refers to the cytochrome oxidase b apoenzyme. tRNA genes are designated *trnX*, where *X* is the single-letter amino acid code. Contrary to Boore and Brown (2000), the large and small

ribosomal subunits are here referred to as *mLSU* (mitochondrial large subunit) and *mSSU* (mitochondrial small subunit) respectively.

Clymenella torquata

All mtDNA amplifications of *C. torquata* employed 1μL EXL Polymerase (Stratagene), as well as 5μL EXL buffer, 25pmol dNTPs, 200ng each primer, 1μL stabilizing solution and approximately 10ng genomic DNA per 50μL reaction. The sections *mLSU–cox1* (using primers 16Sar-L/HCO2198), *cox1–cox3* (LCO1498/COIIIr), *cox3–cob* (COIIIf/CytbR), and *cob–mLSU* (CytbF/16Sbr-H) all generated single-banded products. The *mLSU* primers are from Palumbi (1996); *cob* and *cox3* primers are from Boore and Brown (2000), and the COI primers are from Folmer et al. (1994). PCR protocols for these fragments are found in the supplementary material. Products were verified on an agarose gel, purified using the QiaQuick kit (Qiagen), eluted in 40μL water, and sheared separately in a HydroShear DNA shearer (GeneMachines) to generate random fragments of 1-2kb in length. The sticky ends were polished with the Klenow fragment, and were A-tailed using *Taq* polymerase, an excess of dATP, and incubation at 72°C for 10 min. DNA was then repurified with the QiaQuick kit, and cloned into pGEM-T Easy (Promega). Sequencing reactions were performed using Big Dye (versions 2 and 3) chemistry on an ABI 377 (Applied Biosystems). Fifteen *mLSU–cox1* clones (average coverage 5.3X), 9 *cox3-cob* clones (average 2.9X) and 7 *cob–mLSU* clones (average 2X) were sequenced in both directions using T7 and SP6 and then assembled to generate contigs. Combined, the assemblies contained ~90% of the sequence of *C. torquata*'s mt-genome. Three clones could not be entirely sequenced

using plasmid primers. To complete sequencing on these clones, 19 walking-primers were designed (see supplementary information). No clones were recovered containing the largest non-coding region (i.e. the control region or *UNK)* or the approximately 3kb surrounding it (roughly including regions of the *atp6* and *nad4L* genes, and all of *nad5, trnW, -H, -F, -E, -P,* and -*T*). This region was sequenced by amplification with flanking primers (Ctatp6f2 and Ctnad4r2) and direct sequencing using the walking primers.

Riftia pachyptila

mtDNA amplification for *R. pachyptila* was adapted from the procedure of Boore and Brown (2000). Standard primers were used to amplify short sections of *cox1* (LCO1490 and HCO2198; Folmer *et al.* 1994), and *cob* (CytbF and CytbR; Boore and Brown 2000) with *Taq* polymerase (Promega) in standard 25μL PCRs. Products were purified using the QiaQuick Gel Extraction Kit (Qiagen) and sequenced on an ABI 377 automated sequencer. These sequences were used to design *Riftia*-specific primers for long PCR. In *cox1*, the primers Rp1536 and Rp2161 were designed, and in *cob*, CytBRp. Information for all primers can be found in supplementary information.

These primers were then used to amplify long segments of the mt-genome in conjunction with the primers mentioned above: 16Sar-L and Rp1536 amplified the region spanning *mLSU–cox1*, Rp2161 and COIIIr amplified *cox1–cox3*, and COIIIf and CytBRp amplified *cox3–cob*. These long PCR reactions consisted of 5uL 10X *rTth* buffer, approximately 10ng template DNA, 25pmol dNTPs, 30pmol each primer, 0.4μL (1U) *rTth* polymerase, and 1μL of *Vent* polymerase diluted 1:100 (0.02U) per 50μL. Both polymerases are from Applied Biosystems. PCR products were verified, and when

necessary size selected, using 1% agarose gels. Single-banded products were purified and single A-overhangs added as above. A-tailed fragments were cloned into the pGEM-T Easy vector (Promega). Initial clone sequencing used the plasmid primers T7 and SP6; complete bidirectional sequencing was accomplished by primer walking, resulting in an average sequencing coverage of 7.8X.

Amplification of the *cob–mLSU* region in *Riftia*, which presumably contains UNK, was difficult. Part of this remaining region was sequenced by designing degenerate primers to *nad4* sequences obtained from the complete genomes of *Lumbricus terrestris*, *Platynereis dumerilii,* and *Katharina tunicata*. These primers (nad4f, TGR GGN TAT CAR CCN GAR CG and nad4r, GCY TCN ACR TGN GCY TTN GG) amplified a short region of *nad4,* and allowed the design of primers specific to *R. pachyptila* (Rpnad4bf and Rpnad4br). Using EXL polymerase (Stratagene), the primer combination Rpnad4bf/16Sbr-H (Palumbi 1996) amplified the region spanning *nad4–mLSU*, but the region between *cob* and *nad4*, which again was presumed to contain *UNK*, was still difficult to amplify and could not be cloned successfully after amplification. Three clones containing spliced PCR amplicons for this fragment (see Results) were partially sequenced and provided the remainder of *cob* as well as complete *trnW* and *atp6* genes. For simplicity, the *R. pachyptila* fragment will henceforth be referred to as the *R. pachyptila* genome.

Genomic Assembly

Assembled sequences were checked by BLAST (Altschul et al. 1990) searches against GenBank. Those sequences that returned strong BLAST hits to mitochondrial protein-encoding genes were translated into amino acids using the *Drosophila* mitochondrial code and aligned in CLUSTAL X (Thompson et al. 1997) with other available lophotrochozoan genome sequences (Table 1) obtained from GenBank to ensure correct identification. The full genomes were assembled by resolving ambiguous sequence reads in AutoAssembler (Applied Biosystems), checking against the amino acid alignments, and concatenating the individual alignments to make the complete genome alignment in MacClade 4.03 (Maddison and Maddison 2000).

Candidate tRNA genes were found using the tRNAScan-SE web server (http://www.genetics.wustl.edu/eddy/tRNAscan-SE); this identified all but four tRNAs in *C. torquata* and one in *R. pachyptila*. Stretches of mtDNA that did not code for protein genes and were in a similar position to tRNAs in previously published annelid genomes were scanned by eye for potential tRNA secondary structure and the presence of the anticipated anticodon sequence. The tRNA structures reported here are proposed based on the tRNAScan-SE foldings, keeping in mind the general forms suggested by Dirheimer et al. (1995). rRNA genes were identified by sequence homology with BLAST entries, and 5' and 3' ends were assumed to be directly adjacent to up- and downstream genes. The boundaries of the *C. torquata UNK* were similarly inferred from the ends of the upstream and downstream tRNAs.

Phylogenetic Analysis

Table 1 lists the taxa and their GenBank accession numbers used for phylogenetic inference. Outgroups were chosen based on knowledge of Lophotrochozoan evolutionary history (Halanych 2004). Because we hoped to develop a better understanding of the utility of mtDNA in constructing annelid phylogeny, we chose to sub-sample available lophotrochozoan mtDNA genomes for use as outgroups. For mollusks, we chose the polyplacophoran *Katharina tunicata* for its basal position, the two pulmonate gastropods *Albinaria caerulea* and *Cepaea nemoralis* because they were more easily aligned than other gastropods, and the cephalopod *Loligo bleekeri* to achieve a broader representation of mollusks. Several other molluscan genomes contained large insertions and deletions in several genes relative to annelids, greatly complicating attempts at alignment. All three available brachiopods (*Terebratalia transversa, Terebratulina retusa,* and *Laqueus rubellus*) were included in the analyses. To create the final alignment, DNA from protein-encoding genes was aligned in MacClade 4.03 using CLUSTALX alignments of the corresponding amino acids; rRNA genes were aligned manually using secondary structure as a guide, employing phylogenetic conservation diagrams obtained from the RNA database at the University of Texas's Institute for Cellular and Molecular Biology (http://www.rna.icmb.utexas.edu/topmenu.html). tRNAs, *UNK*, and non-coding DNA were not included in the alignments due to high variability (see below). This produced a single multi-partitioned alignment in MacClade 4.03, which is available at TreeBase (http://www.treebase.org) and in the supplementary information.

Two sequence-based datasets and one gene-order dataset were created. One sequence-based dataset contained nucleotide sequences from protein-encoding and rRNA

genes, and the second contained only inferred amino acid sequences. Regions that could not be unambiguously aligned, and all third codon positions were removed. The amino acids of three protein-coding genes (*atp6, atp8, nad6*) exhibited high variation, which made alignment difficult, and thus were excluded from both datasets.

All non-annelid taxa herein were treated as outgroups; however, brachiopods are drawn basally for illustrative purposes. Although mollusks, annelids, brachiopods, and sipunculids are closely related, the relationships between them are not well resolved (Halanych 2004). PAUP*4.0b10 (Swofford 2002) was used for parsimony and maximum likelihood (ML) analyses. For both datasets, gaps were treated as missing data. For the DNA dataset, maximum likelihood models and their parameters were determined with hierarchical likelihood ratio tests (hLRT's) using the program MODELTEST 3.5 (Posada and Crandall 1998). Heuristic searches in PAUP under both parsimony and ML employed random sequence addition (parsimony—100 replicates; likelihood—10 replicates) to obtain starting trees, and TBR swapping. Bootstrapping with character re-sampling was performed with 1000 replicates for parsimony and 500 replicates for ML. Decay indices (also called Bremer support, Bremer 1994) were also calculated for the parsimony trees using constraints in PAUP.

The order of genes in the mitochondria was used as a third dataset for phylogenetic analysis. Although breakpoint analysis (Blanchette, Kunisawa, and Sankoff 1999) has proven useful in many cases, we prefer a newer parsimony framework (described in Boore and Staton 2002), which does not condense the data into pairwise distance measures, and allows partial genomes to be included. Briefly, 74 multistate characters were created ("upstream of gene X" and "downstream of gene X" for each of

the 37 genes), and character states were coded as “beginning of gene Y” and “end of gene Y”, for a total of 74 states (though obviously a gene cannot appear up- or downstream of itself). The matrix was then analyzed in PAUP under parsimony as previously outlined. Because the gene orders of four taxa (*P. gouldii, G. brachiosum, H. robusta, R. pachyptila*) are incompletely known, missing and ambiguous characters (52) were removed before searching for trees, leaving 22 characters. The brachiopods were again placed as the basal-most outgroup. For comparative purposes, breakpoint and inversion distances were calculated using GRAPPA 1.6 (Bader, Moret and Yan 2001).

RESULTS

Genomic Composition

The complete mitochondrial DNA (mtDNA) of *C. torquata* is 15,538 bp in length, and the *R. pachyptila* fragment is 12,016 bp long. Figure 1 shows the gene order for both genomes. The *C. torquata* genome is similar in size (i.e., about 15kb) to other lophotrochozoan mitochondrial genomes, and the portion of the *R. pachyptila* genome is of similar size to the same portions from *C. torquata* and *L. terrestris*. Table 2 shows a breakdown of nucleotide composition. Both genomes show patterns of nucleotide bias and skew[1]. The two genomes are AT-rich (~66%), and this bias is consistent across the three main gene types (those coding for proteins, tRNAs, and rRNAs). T is the most common base, and G the least. Further, the percentage of G's is markedly lower at third codon positions than even the low overall G frequency. In contrast to nucleotide bias, patterns of AT- and GC-skew are not as consistent across gene types. Skew for a given strand is calculated as (A–T)/(A+T) [or (G–C)/(G+C)] (Perna and Kocher 1995) and ranges from +1 if the coding strand has A (G) for every AT (GC) pair to –1 if the coding strand always has T (C). On the whole AT-skew is slightly negative, and GC-skew is more negative than AT-skew. In both genomes, AT-skew is most positive in 2nd codon positions, and GC-skew is most negative at 3rd codon positions.

The genome of *C. torquata* contains the standard 37 genes found in mtDNAs: 13 protein-coding genes, 2 genes for rRNAs, and 22 genes for tRNAs (Boore 1999). The *R.*

[1] Herein, " nucleotide bias" refers to unequal nucleotide frequencies (i.e., departures from 25% each) and "codon bias" to unequal frequencies of the codons that code for a single amino acid (e.g., UUA used for leucine more often than UUG). "Skew" will refer specifically to the orientation of hydrogen-bonded pairs in the molecule (e.g. whether the coding strand contains the G of a GC pair or the C).

pachyptila fragment contains 9 complete protein-coding genes (*atp8, cox1, cox2, cox3, cob, nad1, nad2, nad3, nad6)* and portions of two others (*atp6, nad4),* as well as both rRNA genes (*mLSU, mSSU)* and 16 tRNA genes (*trnA, -C, -D, -G, -I, -K, -L1, -L2, -M, -N, -Q, -S1, -S2, -V, -W, -Y);* the remaining genes (*nad4L, nad5,* and *trnE, -F, -H, -P, -R, -T)* and the *UNK* are presumably in the unsequenced portion. As seen in all other annelids to date, all genes in both genomes are encoded on a single strand.

Start and stop codon usage also shows patterns of bias. Start codons in protein-coding genes are highly biased towards ATG over ATA; ATG is observed in 12 of 13 coding genes in *C. torquata* (*nad4* uses ATA) and all 10 *R. pachyptila* coding genes for which the 5' end is known. In addition, overlap typically exists between the presumptive stop codon (TAA or TAG) and the 5' end of the next gene. In other words, some stop codon bases appear to be part of the transcript of the down stream gene (illustrated in Supplementary Information). For the purposes of annotation, the stop codon in all such cases is assumed to be incomplete (see Ojala, Montoya and Attardi 1981), and the shared bases assigned to the downstream gene.

There is considerable codon usage bias in both genomes as well, with some codons within a group being used more than an order of magnitude more frequently than others (Table 3). In codons that exhibit four-fold degeneracy, triplets ending in G tend to be the least used as expected from overall nucleotide frequencies. However, codons ending in A tend to be the most common within a codon group despite the slightly higher prevalence of T's in nucleotide frequency. In 2-fold codon groups, the use of XXG tends to be considerably less than XXA, and use of XXC is somewhat less than XXT. CCG (Pro) and CGG (Arg) were never observed in *R. pachyptila*.

Putative tRNA structures are depicted for all recovered tRNA genes in Figures 2 and 3 (*C. torquata* and *R. pachyptila*, respectively). Most possess the common cruciform structure, with an acceptor arm, anticodon arm, TΨC arm, DHU arm, and associated loop regions. In *C. torquata*, *trnS2* and -*V* have shortened TΨC stems, and *trnN* in *R. pachyptila* is missing the TΨC entirely. Additionally, *trnS1* and -*S2* in *R. pachyptila* have no DHU stems. *trnS1 and –S2* are shown without DHU stems despite the potential for some base pairing; the lack of DHU stems is a widespread feature of mitochondrial tRNA genes (Dirheimer et al. 1995). Also of interest is the single unpaired nucleotide on the 5' side of the acceptor arm of the *trnL2* gene of *R. pachyptila*, confirmed in three independent sequencing reads.

Phylogenetic analyses

A single shortest tree was recovered under parsimony for both the DNA and AA datasets. The DNA tree is shown in Figure 4a (16,680 steps, C.I. =0.549), and the AA tree in Figure 4b (12,645 steps, C.I. =0.756). Monophyly of the Annelida was recovered in both trees, as both topologies are consistent with a monophyletic Brachiopoda (100% bootstrap support in both analyses). Also in both trees, *P. gouldii* is sister to Annelida, and the two siboglinids (*R. pachyptila* and *G. brachiosum*) cluster together. There are two main differences between the trees. In the DNA tree, the oligochaete and hirudinean fall outside of the polychaetes, whereas in the AA tree they are inserted among polychaetes. The arrangement of mollusks also differs between the two trees. In the DNA tree, the mollusks are monophyletic with the polyplacophoran basal, the two gastropods together, and the cephalopod in the most derived position. In the AA tree, the cephalopod and

polyplacophoran are more closely related to the sipunculan and annelids (bootstrap support 83%) than to the gastropods.

For the ML nucleotide data, Modeltest chose the GTR+I+G model as the best fit to the data (nucleotide frequencies A=0.2557, C=0.1899, G=0.1942, T=0.3602; rates A⇔C 1.6203, A⇔G 3.4278, A⇔T 1.6946, C⇔G 2.4572, C⇔T 3.6315, G⇔T 1.000; proportion of invariable sites 0.1993, gamma shape parameter 0.8916). The single best maximum likelihood tree (-ln likelihood = 67626.63583) obtained with this model bears a strong similarity to both the DNA and AA trees (Figure 4c). Bootstrap support of 100% was found for an Annelida+Sipuncula clade with Sipuncula nested within the group as sister to the maldanid *Clymenella torquata*. Limited support (67% bootstrap support) for hirudineans and oligochaetes, the Clitellata, within polychaetes was also found.

The gene-order analysis produced 15 equally parsimonious trees of 112 steps. The strict consensus of these trees (see supplementary information) contained far less resolution than the trees derived from nucleotide or amino acid sequences with only three supported nodes. Consistent with other analyses, the two gastropods clustered together with 100% bootstrap support. Ninety-one percent support was also recovered for the node containing all annelids and *P. gouldii*. A grouping of this clade as sister to *L. rubellus* had weak support (53%). To determine if this lack of resolution was due to the parsimony method of analyzing gene order or intrinsic to the data, GRAPPA 1.6 breakpoint and inversion distances were also calculated. However, in these trees Brachiopoda and Mollusca interdigitated to a large degree (not shown). Neither algorithm can handle partial genomes; thus, *P. gouldii, G. brachiosum, H, robusta,* and *R.*

pachyptila had to be excluded from these analyses, further reducing the phylogenetic inferences that could be made. It thus appears that all of these gene order algorithms are sensitive to the disparate rates of change present in our dataset.

DISCUSSION

The present study covers all major recognized clades of annelids (Rouse and Fauchald 1997). Annelid mitochondrial gene order appears to be evolutionarily conserved. With the exception of *trnK*'s placement in *C. torquata,* and as far as could be determined for *R. pachyptila*, both genomes examined here have the same gene order as *Lumbricus terrestris* and the fragments of *Galathealinum brachiosum* and *Helobdella robusta*. *Platynereis dumerilii* differs in the placement of the *UNK* region and a few tRNAs (Figure 1). In contrast to annelids, mollusks display considerable gene order variation over a similar timescale (e.g. Dreyer and Steiner 2004). For example, even within Gastropoda and Cephalopoda large numbers of rearrangements are common (e.g. Kurabayashi and Ueshima 2000, Serb and Lydeard 2003). The three brachiopods also display very dissimilar gene orders. The origins of major taxa in these groups date back to the Cambrian (approximately 540 MYA) (Knoll and Carroll 1999). Thus, it appears that there may be a considerable difference in how annelid, mollusk, and brachiopod mitochondrial genomes evolve. This difference is interesting because of the apparent close relationship of these lophotrochozoan taxa. These results raise the possibility that gene order is highly variable in general across lophotrochozoan taxa, and that only select subgroups exhibit conserved gene orders (e.g., Annelida). If true, this situation may have considerable repercussions on how mtDNA gene order data can be used to infer evolutionary history among different animal clades.

Phylogenetic Relationships

The AA parsimony and DNA likelihood phylogenetic analyses are consistent with previous findings that place Clitellata (McHugh 1997; Rota, Martin and Erséus 2001; Bleidorn, Vogt and Bartolomaeus 2003) and siboglinids (McHugh 1997; Rouse and Fauchald 1997; Kojima 1998; Halanych et al. 1998; 2001) as derived "polychaetes". Thus, the last common ancestor of "Polychaeta" and Annelida are one and the same. However bootstrap values (67% likelihood, <50% for AA parsimony) for this result were weak and Shimodaira-Hasegawa tests (Shimodaira and Hasegawa 1999), fell short of significant values (in both cases, p=0.14, 1000 replicates with RELL option). An alternative topology in the nucleotide parsimony analysis was not well supported. Clearly, considerably more taxa need to be sampled to understand the robustness of these results and placement of these groups within annelids. The groupings *R. pachyptila* + *G. brachiosum* and *H. robusta* + *L. terrestris* were highly supported in all sequence analyses in agreement with morphological expectations. An additional result consistently recovered by sequenced-based analyses was placement of the sipunculan as sister to or inside Annelida (Shimodaira-Hasegawa test p = 0.003). Boore and Staton (2002) first reported this result using many of the same mtDNA sequences used herein. Thus, although gene order may be uninformative in this case, there is high support from both DNA and amino acid sequences for an Annelida/Sipuncula clade to the exclusion of mollusks and brachiopods. Interestingly, nuclear large ribosomal subunit data also weakly supports sipunculans as the sister clade to annelids (Passamaneck and Halanych,

in prep). The likelihood tree provides the first suggestion that Sipuncula is within Annelida, but this finding requires additional verification.

In contrast to the sequence-based data sets, the gene order analysis offers little resolution. This result is to be expected with the limited observed variation in annelid gene order. Nonetheless, annelids and the sipunculan cluster together because of identical arrangement of the 11 genes between *cox1* and *cob* (inclusive) and the sequence *mSSU—trnV—mLSU*. The latter sequence appears to be somewhat conserved across lophotrochozoan clades (it is found in 10 of the 23 lophotrochozoan taxa for which data are currently available in GenBank), and potentially in other protostomes as well. Further, the subsequence *trnV–mLSU* is found in 16 of the 23 lophotrochozoan genomes, and some protostomes. In any case, based on the available data, gene order appears to be of limited utility for relationships within the annelids because of its highly conserved nature. All rearrangements seen so far are minor and found in single taxa only, although with greater taxonomic coverage potential synapomorphic gene orders may emerge. Apparently, both within annelids and between annelids and other lophotrochozoans, there is no consistent mechanism controlling the rate or types of gene order modifications. In contrast to the lack of phylogenetic signal in gene order among annelids, the resolution offered by sequence-based analyses holds promise.

Mitochondrial Genome Organization and Structure

The two genomes presented here also exhibit the pattern of post-transcriptional modification and splicing described by Ojala, Montoya, and Attardi *et al.* (1981), in

which many stop codons are incomplete in the transcript and are filled in by post-transcriptional editing machinery. This type of splicing is presumed to occur in several genes in both the *C. torquata* and the *R. pachyptila* genomes (see appendices 1 and 2). In the majority of these cases, the overlap in question contains an in-frame stop codon (TAA or TAG), but it is not presumed to be functional. Moreover, in several cases there is no in-frame stop codon at or near the end of the protein-encoding gene, making post-translational addition of a stop codon the only plausible mechanism (see supplementary Figure 2). One example is the *nad1/trnI* junction in *C. torquata*, where *nad1* presumably ends with T__, and *trnI* begins with GA, such that assigning more of the codon (TG_ or TGA) to *nad1* still does not produce a stop codon. Additionally in *C. torquata,* the last six bases of *nad4* (GGCCCT) appear to be used as the first six of *trnC*; a seven-base overlap could give *nad4* an incomplete TA_ stop codon, but the next base is a T, and therefore it is not possible to generate a full stop codon from the primary sequence.

The AT-bias seen in both genomes seems to be contributing to a strong codon bias in protein-coding genes. Although the *R. pachyptila* genome is incomplete, the absence of two GC-rich codons (CCG, encoding proline, and CGG, encoding arginine) may be linked to the low percentage of G and C. However, even given these low frequencies in the protein-coding genes as a whole, the probability of never observing CCG (Pro) in 160 proline codons given an average G content of 12% is $(0.12)^0(1-0.12)^{160} = 1.31 \times 10^{-9}$, and the probability of never seeing CGG (Arg) in 53 arginine codons is $(0.12)^0(1-0.12)^{53} = .0011$ (both assuming independence of codons). Thus, the amount of AT-bias alone does not adequately explain the lack of these two codons and suggests that some other mechanism(s) is responsible for the observed codon

bias. Cardon *et al.* (1994) discuss the paucity of CG dinucleotides in metazoan mitochondrial genomes regardless of their position in codons (i.e. positions (I,II), (II,III), and (III,I)) and overall low usage of arginine (CGN) in mitochondrial proteomes. Indeed, arginine is the least frequent of all amino acids possessing four-fold degenerate codons in both *C. torquata* and *R. pachyptila*, and is even less frequent than some two-fold degenerate amino acids. Based on the symmetrized odds-ratios (ρ_{NN}, where NN is the dinucleotide in question) of Cardon et al. (1994), *R. pachyptila* does show CG suppression (ρ_{CG}=0.5299; $0.78 \leq \rho_{NN} \leq 1.23$ is considered the normal range). Suppression of CG dinucleotides in vertebrate nuclear genomes has been linked to mutation to TG by methylation of the C followed by deaminization to T. This cannot underlie CG suppression in mtDNAs because mitochondria lack the methylation pathway, and because mtDNAs do not usually contain an excess of TG dinucleotides (*R. pachyptila* ρ_{TG}=0.83). Although no simple explanation has been found, the authors suggest that CG suppression is correlated with small genome size and "streamlined" mtDNA organization.

R. pachyptila is a large tubeworm found at Eastern Pacific hydrothermal vent fields. Early genetic analyses on this species led to speculation that hydrothermal vent animals would harbor a high GC nucleotide composition because the extra hydrogen bond, when compared to AT base pairing, would confer additional stability in the potentially high-temperature and reducing environment (Dixon, Simpson-White, and Dixon 1992). Although high GC content has been documented in thermophilic microbes (Woese et al. 1991), *R. pachyptila's* low GC content (a pervasive feature of metazoan mtDNAs in general) argues against such temperature-driven evolution in *R. pachyptila.* Possibly, the higher GC content in *R. pachyptila* postulated by Dixon and colleagues is

restricted to the nuclear genome; however, it is unclear why mitochondrial and nuclear genomes would respond in different ways to the same environmental pressure if this were true.

Genomic Amplification and Sequencing

Our difficulties in amplifying and cloning the *UNK* region of *C. torquata* and *R. pachyptila* likely stem from regulatory aspects of this region of the molecule. In *R. pachyptila,* our long PCR reactions for the region *cob–nad4* repeatedly generated 3-5 bands, even though the reactions employed two ~30mer species-specific primers. Attempts to clone the band of the expected size resulted in very low transformation efficiencies. Of three clones sequenced, each contained an apparent splice in a similar, but not exact, position just downstream of *atp6*, indicating host removal of the genes between *atp6* and *nad4* (presumably containing *trnW, UNK, trnH, nad5, trnF, -E, -P,* and *-T*). Sequencing of the 3' end of these clones provided the complete gene sequences for trnW and *atp6* but the splice prevented accurate determination of what lay farther downstream. A similar region was apparently unclonable in the sheared fragments of *C. torquata*'s mt-genome and had to be obtained by direct sequencing. Boore and Brown (2000) had similar problems when obtaining the similar region in *Platynereis dumerilii*, and suggested that the presence of signaling elements in *UNK* disrupted PCR. Our observations suggest the *UNK* region is identifiable as a foreign origin of replication and is spliced out by at least some *E. coli* cell types (in this case DH5α and JM109—both of which are $recA^{-}$) in addition to possibly interfering with PCR. Alternative strategies may need to be developed to completely sequence large numbers of complete mitochondrial

genomes in order to avoid the need to direct-sequence and primer-walk the region containing *UNK*.

CONCLUSIONS

We have expanded the phylogenetic spread of annelid taxa whose mitochondrial genomes have been sequenced. The high similarity of gene order across annelids provides sharp contrast to the variation observed in mollusks and brachiopods. In both cases, the phylogenetic utility of gene-order data may be limited. The nucleotide and amino acid data, however, produced informative trees with some measure of support. Our results are concordant with the findings of Boore and Brown (2000) and Boore and Staton (2002) on annelid relationships and the relation of Sipuncula to Annelida.

Acknowledgements

We thank Tim Shank for *R. pachyptila* tissues. Early drafts of this paper were greatly improved by comments from Lauren Mullineaux, Stace Beaulieu, and Lara Gulmann; we also thank the two anonymous reviewers for their thorough and helpful comments. Support by CICOR to RJM is gratefully acknowledged. This work was support by the National Science Foundation grants (DEB-0075618 and EAR-0120646) to KMH. This work is WHOI contribution number 11244.

Literature Cited

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. **215**:403-410.

Bader, D. A., Moret, B. M. E. and Yan, M. 2001. A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. J. Comput. Biol. **8:** 483-491.

Blanchette, M., T. Kunisawa, and D. Sankoff. 1999. Gene order breakpoint evidence and animal mitochondrial phylogeny. J. Mol. Evol. **49**:193-203.

Bleidorn, C., L. Vogt, and T. Bartolomaeus. 2003. A contribution to sedentary polychaete phylogeny using 18S rRNA sequence data. J Zool. Syst. Evol. Res. **41**:186-195.

Boore, J. L. 1999. Animal mitochondrial genomes. Nucl. Acid. Res. **27**:1767-1780.

Boore, J. L., and W. M. Brown. 2000. Mitochondrial genomes of *Galathealinum, Helobdella,* and *Platynereis*: Sequence and gene rearrangement comparisons indicate the Pogonophora is not a phylum and Annelida and Arthropoda are not sister taxa. Mol. Biol. Evol. **17**:87-106.

Boore, J. L., and W. M. Brown. 1998. Big trees from little genomes: mitochondrial gene order as a phylogenetic tool. Curr. Opin. Genet. Dev. **8**:668-674.

Boore, J. L., and W. M. Brown. 1994. Complete DNA sequence of the mitochondrial genome of the black chiton, *Katharina tunicata*. Genetics **138**:423-443.

Boore, J. L., and J. L. Staton. 2002. The mitochondrial genome of the Sipunculid *Phascolopsis gouldii* supports its association with Annelida rather than Mollusca. Mol. Biol. Evol. **19**:127-137.

Boore, J. L., M. Medina, and L.A. Rosenberg. 2004. Complete sequences of the highly

rearranged molluscan mitochondrial genomes of the scaphopod *Graptacme eborea* and the bivalve *Mytilus edulis*. Mol Biol Evol **21:** 1492-1503.

Bremer, K. 1994. Branch support and tree stability. Cladistics **10**:295-304.

Cardon, L. R., C. Burge, D. A. Clayton and S. Karlin. 1994. Pervasive CpG suppression in animal mitochondrial genomes. Proc. Nat. Acad. Sci. USA **91:** 3799-3803.

DeJong, R. J., A. M. Emery, and C. M. Adema. 2004. The mitochondrial genome of *Biomphalaria glabrata* (Gastropoda, Basommatophora), intermediate host of *Schistosoma mansoni*. J. Parasitol. **in press**.

Dirheimer, G., G. Keith, P. Dumas and E. Westhof. 1995. Primary, secondary, and tertiary structures of tRNAs. Pp. 93-126 *in* D. Söll and U. RajBhandary, eds. tRNA: Structure, Biosynthesis, and Function. ASM Press, Washington, DC.

Dixon, D. R., R. Simpson-White, and L. R. J. Dixon. 1992. Evidence for thermal stability of ribosomal DNA sequences in hydrothermal-vent organisms. J. Mar. Biol. Assoc. U.K. **72**:519-527.

Dreyer, H., and G. Steiner. 2004. The complete sequence and gene organization of the mitochondrial genome of the gadilid scaphopod *Siphonondentalium lobatum* (Mollusca). Mol. Phylogenet. Evol. **31**:605-617.

Folmer, O., M. Black, W. Hoeh, R. Lutz, and R. Vrijenhoek. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Mol. Mar. Biol. Biotech. **3**:294-299.

Grande, C., J. Templado, J. L. Cervera, and R. Zardoya. 2002. The complete mitochondrial genome of the nudibranch *Roboastra europaea* (Mollusca: Gastropoda) supports the monophyly of opisthobranchs. Mol. Biol. Evol. **19**:1672-1685.

Halanych, K. M. 2004. The new view of animal phylogeny. Ann. Rev. Ecol. Evol. Syst. **35**:229-256.

Halanych, K. M., T. G. Dahlgren, and D. McHugh. 2002. Unsegmented annelids? Possible origins of four lophotrochozoan worm taxa. Integrat. Compar. Biol. **42**:678-684.

Halanych, K. M., R. A. Feldman, and R. C. Vrijenhoek. 2001. Molecular evidence that *Sclerolinum brattstromi* is closely related to vestimentiferans, not to frenulate pogonophorans (Siboglinidae, Annelida). Biol. Bull. **201**:65-75.

Halanych, K. M., R. A. Lutz, and R. C. Vrijenhoek. 1998. Evolutionary origins and age of vestimentiferan tube-worms. Cah. Biol. Mar. **39**:355-358.

Hatzoglou, E., G. C. Rodakis, and R. Lecanidou. 1995. Complete sequence and gene organization of the mitochondrial genome of the land snail *Albinaria caerulea*. Genetics **140**:1353-1366.

Helfenbein, K. G., and J. L. Boore. 2004. The mitochondrial genome of *Phoronis architecta* - Comparisons demonstrate that phoronids are lophotrochozoan protostomes. Mol. Biol. Evol. **21**:153-157.

Helfenbein, K. G., W. M. Brown, and J. L. Boore. 2001. The complete mitochondrial genome of the articulate brachiopod *Terebratalia transversa*. Mol. Biol. Evol. **18**:1734-1744.

Hoffmann, R. J., J. L. Boore, and W. M. Brown. 1992. A novel mitochondrial genome organization for the blue mussel, *Mytilus edulis*. Genetics **131**:397-412.

Knoll, A., and S. B. Carroll. 1999. Early animal evolution: Emerging views from comparative biology and geology. Science **284**:2129-2137.

Kojima, S. 1998. Paraphyletic status of Polychaeta suggested by phylogenetic analysis based on the amino acid sequences of elongation factor-1-alpha. Mol. Phylogenet. Evol. **9**:255-

261.

Kurabayashi, A., and R. Ueshima. 2000. Complete sequence of the mitochondrial DNA of the primitive opisthobranch gastropod *Pupa strigosa*: systematic implication of the genome organization. Mol. Biol. Evol. **17**:266-277.

Lavrov, D. V., W. M. Brown, and J. L. Boore. 2004. Phylogenetic position of the Pentastomida and (pan)crustacean relationships. Proc. R. Soc. Lond. B. **271**:537-544.

Maddison, D. R., and W. P. Maddison. 2000. MacClade. Sinauer Associates, Inc., Sunderland, MA.

Mangum, C. P. 1964. Studies on speciation in Maldanid polychaetes of the North American Atlantic Coast. II. Distribution and competitive interaction of five sympatric species. Limnol. Oceanogr. **9:** 12-26.

McHugh, D. 1997. Molecular evidence that echiurans and pogonophorans are derived annelids. Proc. Natl. Acad. Sci. USA **94**:8006-8009.

McHugh, D. 2000. Molecular Phylogeny of the Annelida. Can. J. Zool. **78**:1873-1884.

Noguchi, Y., Endo, K., Tajima, F. and Ueshima, R. 2000. The mitochondrial genome of the brachiopod *Laqueus rubellus*. Genetics **155:** 245-259.

Ojala, D., J. Montoya, and G. Attardi. 1981. tRNA punctuation model of RNA processing in human mitochondria. Nature **290**:470-474.

Palumbi, S. R. 1996. Nucleic acids II: The polymerase chain reaction. Pp. 205-248 *in* D. M. Hillis, C. Mortiz, and B. K. Mable, eds. Molecular Systematics. Sinauer Associates, Inc., Sunderland, Massachusetts.

Perna, N. T. and Kocher, T. D. 1995. Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. J Mol. Evol. **41:** 353-358.

Posada, D., and K. A. Crandall. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics **14**:817-818.

Rota, E., P. Martin, and C. Erséus. 2001. Soil-dwelling polychaetes: enigmatic as ever? Some hints on their phylogenetic relationship as suggested by a maximum parsimony analysis of 18S rRNA gene sequences. Contri. Zool. **70**:127-138.

Rouse, G. W., and K. Fauchald. 1997. Cladistics and polychaetes. Zool. Scripta **26**:139-204.

Rouse, G. W., and K. Fauchald. 1995. The articulation of annelids. Zool. Scripta **24**:269-301.

Rouse, G. W., and F. Pleijel. 2001. Polychaetes. Oxford University Press, New York.

Serb, J. M., and C. Lydeard. 2003. Complete mtDNA sequence of the North American freshwater mussel, *Lampsilis ornata* (Unionidae): An examination of the evolution and phylogenetic utility of mitochondrial genome organization in Bivalvia (Mollusca). Mol. Biol. Evol. **20**:1854-1866.

Shimodaira, H., and M. Hasegawa. 1999. Multiple comparisons of Log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. **16**:1114-1116.

Southward, A. J. and E. C. Southward. 1988. Pogonophora: Tube-worms dependent on endosymbiotic bacteria. Anim. Plant Sci. **1:** 203-207.

Stechmann, A., and M. Schlegel. 1999. Analysis of the complete mitochondrial DNA sequence of the brachiopod *Terebratulina retusa* places Brachiopoda within the protostomes. Proc. Roy. Soc. Lond. Ser. B. **266**:2043.

Swofford, D. L. 2002. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer Associates, Inc., Sunderland, MA.

Terrett, J. A., S. Miles, and R. H. Thomas. 1996. Complete DNA sequence of the mitochondrial genome of *Cepaea nemoralis* (Gastropoda: Pulmonata). J. Mol. Evol.

42:160-168.

Thompson, J., T. Gibson, F. Plewniak, F. Jeanmougin, and D. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucl. Acid. Res. **25**:4876-4882.

Tomita, K., Yokobori, S. I., Oshima, T., Ueda, T. and Watanabe, K. 2002. The cephalopod *Loligo bleekeri* mitochondrial genome: multiplied noncoding regions and transposition of tRNA genes. J Mol. Evol. **54:** 486-500.

Wilding, C. S., P. J. Mill, and J. Grahame. 1999. Partial sequence of the mitochondrial genome of *Littorina saxatilis*: Relevance to gastropod phylogenetics. J. Mol. Evol. **48**:0348-0359.

Woese, C.R., L. Achenbach, P. Rouviere and L. Mandelco. 1991. Archaeal phylogeny: reexamination of the phylogenetic position of *Archaeoglobus fulgidus* in light of certain composition-induced artifacts. Syst. Appl. Microbiol. **14:**364-371.

Table 1. Taxa used in phylogenetic analyses

Species	Clade	Nucleotides	GenBank Accession
Clymenella torquata	Annelida, Scolecida, Maldanidae	15,538 complete	AY741661
Riftia pachyptila	Annelida, Canalipalpata, Siboglinidae	12,016 partial	AY741662
Galathealinum brachiosum	Annelida, Canalipalpata, Siboglinidae	7,576 partial	AF178679
Platynereis dumerilii	Annelida, Aciculata, Nereididae	15,619 complete	NC_000931
Lumbricus terrestris	Annelida, Oligochaeta, Lumbricidae	14,998 complete	NC_001673
Helobdella robusta	Annelida, Hirudinea, Glossiphoniidae	7,553 partial	AF178680
Phascolopsis gouldii	Sipuncula	7,470 partial	AF374337
Katharina tunicata	Mollusca, Polyplacophora	15,532 complete	NC_001636
Loligo bleekeri	Mollusca, Cephalopoda	17,211 complete	NC_002507
Albinaria caerulea	Mollusca, Gastropoda	14,130 complete	NC_001761
Cepaea nemoralis	Mollusca, Gastropoda	14,100 complete	NC_001816
Terebratulina retusa	Brachiopoda, Articulata	15,451 complete	NC_000941
Laqueus rubellus	Brachiopoda, Articulata	14,017 complete	NC_002322
Terebratalia transversa	Brachiopoda, Articulata	14,291 complete	NC_003086

Table 2. Base Composition, Bias, and Skew

A. *C. torquata*

	Protein Coding				rRNA	tRNA	Whole Genome
	All Positions	1st Positions	2nd Positions	3rd Positions			
A	31.17%	32.19%	18.19%	43.11%	38.99%	36.17%	32.96%
T	35.03%	27.45%	43.81%	33.83%	29.49%	31.74%	34.28%
A+T	66.20%	59.64%	62.00%	76.94%	68.48%	67.91%	67.24%
C	20.66%	19.97%	24.59%	17.45%	17.49%	15.73%	19.46%
G	13.14%	20.40%	13.41%	5.61%	13.99%	16.36%	13.30%
AT-skew	-0.06	0.08	-0.41	0.12	0.14	0.07	-0.02
GC-skew	-0.22	0.01	-0.29	-0.51	-0.11	0.02	-0.19
base pairs	11146	3716	3715	3715	2116	1424	15538

B. *R. pachyptila*

	Protein Coding				rRNA	tRNA	Whole Genome
	All Positions	1st Positions	2nd Positions	3rd Positions			
A	29.55%	28.54%	18.40%	41.79%	38.02%	34.14%	31.49%
T	36.50%	29.73%	43.55%	36.23%	28.47%	32.02%	34.70%
A+T	66.05%	58.27%	61.95%	78.02%	66.50%	66.16%	66.19%
C	21.87%	21.86%	24.92%	18.81%	19.43%	16.59%	20.96%
G	12.07%	19.88%	13.14%	3.18%	14.07%	17.16%	12.84%
AT-skew	-0.11	-0.02	-0.41	0.07	0.14	0.03	-0.05
GC-skew	-0.29	-0.05	-0.31	-0.71	-0.17	0.03	-0.24
base pairs	8799	2933	2930	2928	2146	1037	12016

Table 3. Codon usage.[1]

Codon	AA	*C. torquata* N	*C. torquata* %	*R. pachyptila* N	*R. pachyptila* %
UUU	Phe (F)	211	75	143	61
UUC	Phe	70	25	92	39
		281		**235**	
UUA	Leu (L)	218	95	191	97
UUG	Leu	12	5	5	3
		230		**196**	
CUU	Leu (L)	108	33	116	40
CUC	Leu	51	16	45	15
CUA	Leu	150	46	127	43
CUG	Leu	14	4	5	2
		323		**293**	
AUU	Ile (I)	238	73	196	77
AUC	Ile	86	27	59	23
		324		**255**	
AUA	Met (M)	224	90	138	84
AUG	Met	25	10	26	16
		249		**164**	
GUU	Val (V)	46	29	34	26
GUC	Val	21	13	19	15
GUA	Val	83	52	76	58
GUG	Val	9	6	2	2
		159		**131**	
UAU	Tyr (Y)	76	65	66	74
UAC	Tyr	41	35	23	26
		117		**89**	
UAA	Ter (.)	7	78	3	75
UAG	Ter	2	22	1	25
		9		**4**	
CAU	His (H)	53	64	49	69
CAC	His	30	36	22	31
		83		**71**	
CAA	Gln (Q)	71	93	56	97
CAG	Gln	5	7	2	3
		76		**58**	
AAU	Asn (N)	80	54	72	67
AAC	Asn	67	46	35	33
		147		**107**	
UCU	Ser (S)	70	33	83	38
UCC	Ser	37	18	50	23
UCA	Ser	99	47	83	38
UCG	Ser	3	1	3	1
		209		**219**	
CCU	Pro (P)	70	39	79	49
CCC	Pro	20	11	34	21
CCA	Pro	77	43	47	29
CCG	Pro	12	7	0	0
		179		**160**	
ACU	Thr (T)	64	26	75	41
ACC	Thr	42	17	41	22
ACA	Thr	132	54	65	36
ACG	Thr	5	2	2	1
		243		**183**	
GCU	Ala (A)	87	35	69	36
GCC	Ala	58	23	51	27
GCA	Ala	96	39	68	36
GCG	Ala	6	2	2	1
		247		**190**	
UGU	Cys (C)	15	11	18	15
UGC	Cys	16	12	13	11
UGA	Trp (W)	80	61	85	73
UGG	Trp	20	15	1	1
		131		**117**	
CGU	Arg (R)	13	22	10	19
CGC	Arg	3	5	5	9
CGA	Arg	38	66	38	72
CGG	Arg	4	7	0	0
		58		**53**	
AGU	Ser (S)	15	15	6	9
AGC	Ser	12	12	6	9
AGA	Ser	64	62	53	78
AGG	Ser	12	12	3	4
		103		**68**	
GGU	Gly (G)	36	19	30	19
GGC	Gly	37	19	19	12

AAA	Lys (K)	79	95	62	95	GGA	Gly	68	35	101	63
AAG	Lys	4	5	3	5	GGG	Gly	52	27	10	6
		83		**65**				**193**		**160**	
GAU	Asp (D)	33	52	28	53						
GAC	Asp	30	48	25	47						
		63		**53**							
GAA	Glu (E)	57	80	59	97						
GAG	Glu	14	20	2	3						
		71		**61**							
							TOTAL	**3578**		**2932**	

[1]Stop codons are only listed if complete.

AA=Amino Acid, N=number of occurrences in all protein-encoding genes observed.

FIG. 1 – Gene orders of annelid and the sipunculan mitochondrial genomes. Abbreviations are as explained in the text. Genomes have been arbitrarily linearized at *cox1* after Boore and Brown (2000). Dashed lines with ellipses in *Riftia, Galathealinum, Helobdella,* and *Phascolopsis* indicate unsequenced regions whose gene order is unknown. Shaded boxes highlight different sets of gene orders conserved among the taxa shown.

FIG. 2 – *Clymenella torquata* assumed tRNA structure diagrams. tRNAs are designated by their single-letter abbreviations.

FIG. 3 – *Riftia pachyptila* assumed tRNA structure diagrams. tRNAs are designated by their single-letter abbreviations.

FIG. 4 – Phylogenetic reconstructions. A. The single best DNA sequence parsimony tree (protein-coding genes and rRNA; see text for details). B. The single best amino-acid parsimony tree. Numbers above branches are bootstrap percentages out of 1000 replicates (percentages below 50 not shown). Numbers below branches are Bremer support values (decay indices). C. DNA sequence maximum likelihood tree (model chosen via Modeltest 3.5). Numbers nearest the node indicate bootstrap percentage out of 500 replicates (percentages below 50 not shown).

Clymenella
(Polychaeta)

Riftia
(Polychaeta)

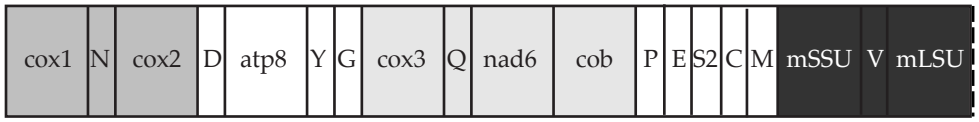
Galathealinum
(Polychaeta)

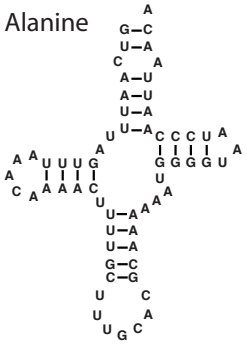
Lumbricus
(Oligochaeta)

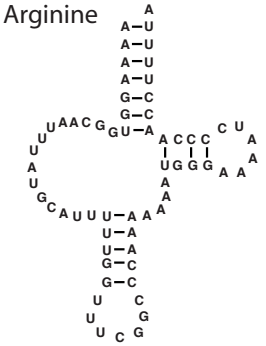
Helobdella
(Hirudinea)

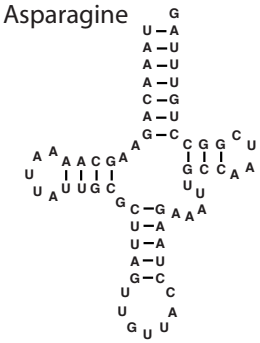
Platynereis
(Polychaeta)

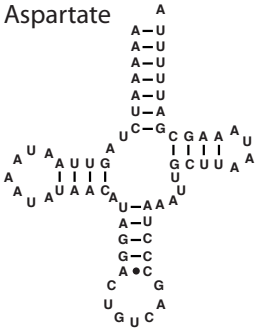
Phascolopsis
(Sipuncula)

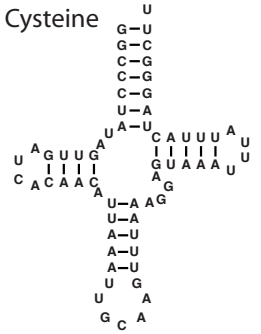


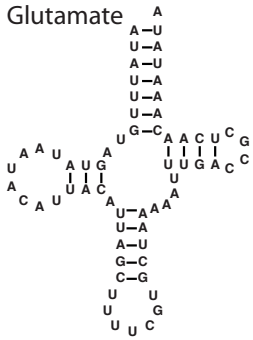

Alanine

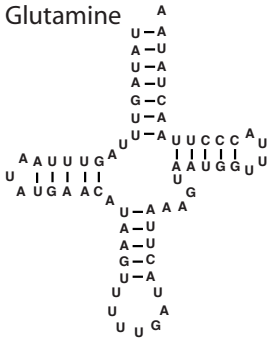

Arginine

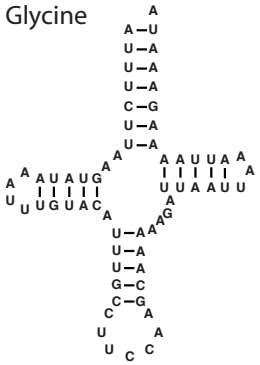

Asparagine

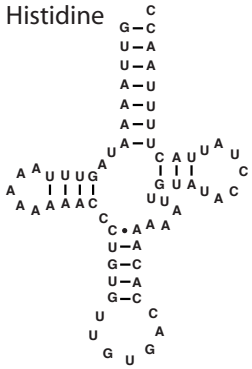

Aspartate

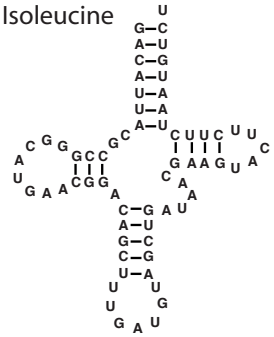

Cysteine

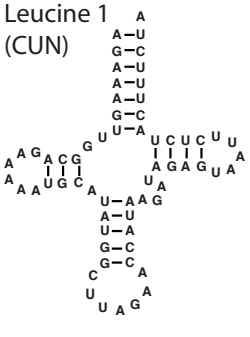

Glutamate

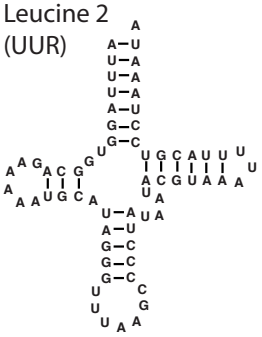

Glutamine

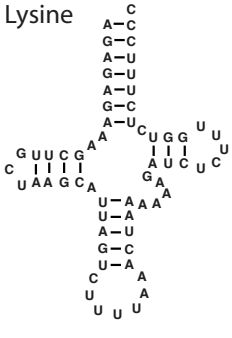

Glycine

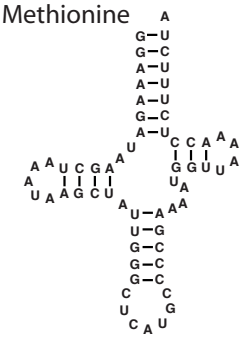

Histidine

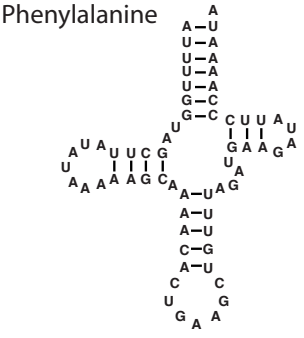

Isoleucine

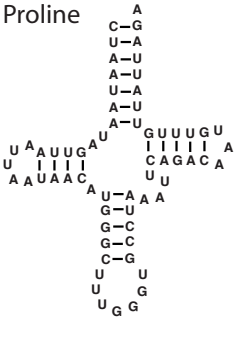

Leucine 1
(CUN)

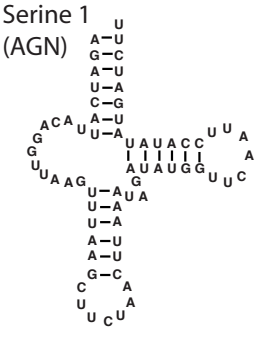

Leucine 2
(UUR)

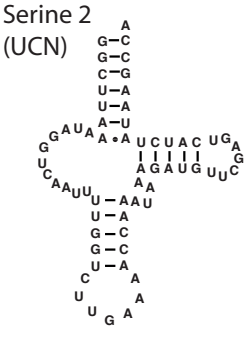

Lysine

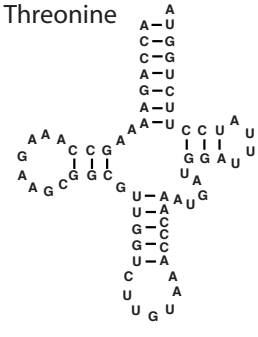

Methionine

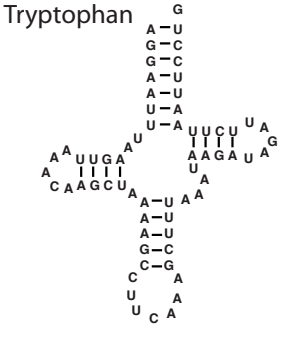

Phenylalanine

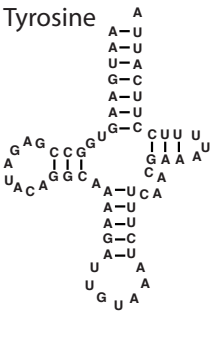

Proline

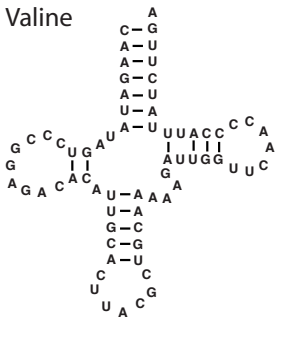

Serine 1
(AGN)


Serine 2
(UCN)

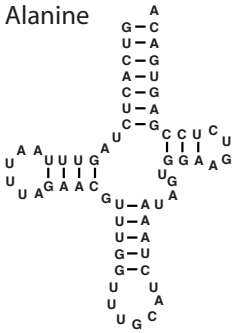

Threonine


Tryptophan


Tyrosine


Valine

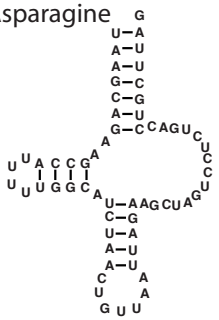
Alanine



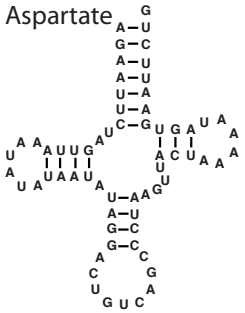
Arginine

(not recovered)

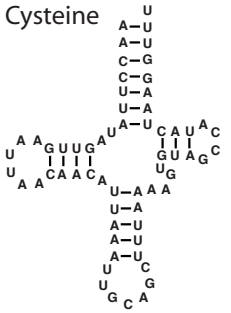
Asparagine



Aspartate



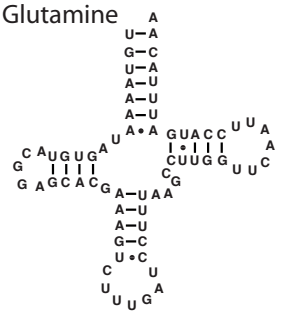
Cysteine



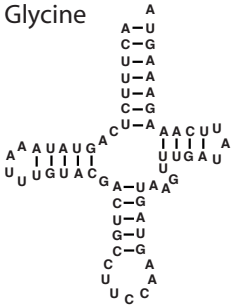
Glutamate

(not recovered)

Glutamine



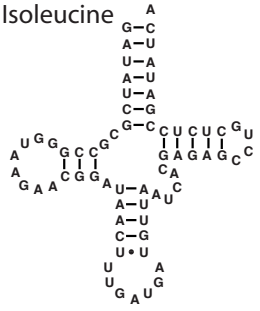
Glycine



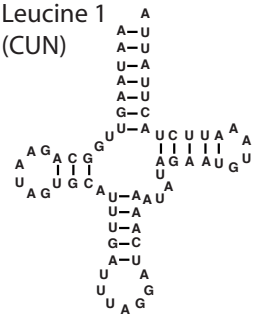
Histidine

(not recovered)

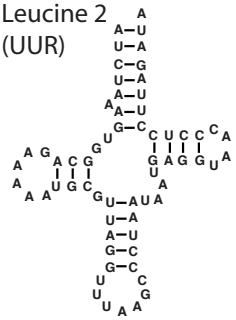
Isoleucine



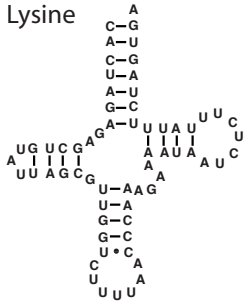
Leucine 1 (CUN)



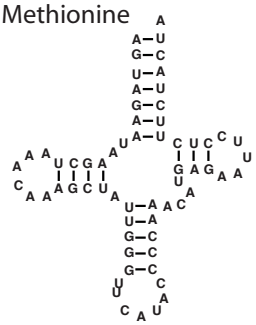
Leucine 2 (UUR)



Lysine



Methionine



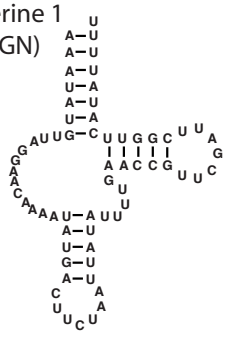
Phenylalanine

(not recovered)

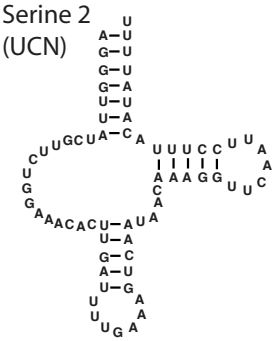
Proline

(not recovered)

Serine 1 (AGN)



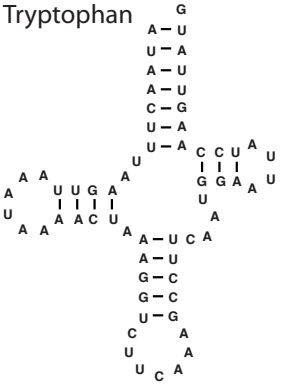
Serine 2 (UCN)



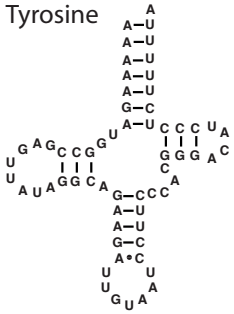
Threonine

(not recovered)

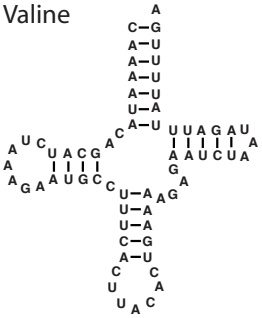
Tryptophan

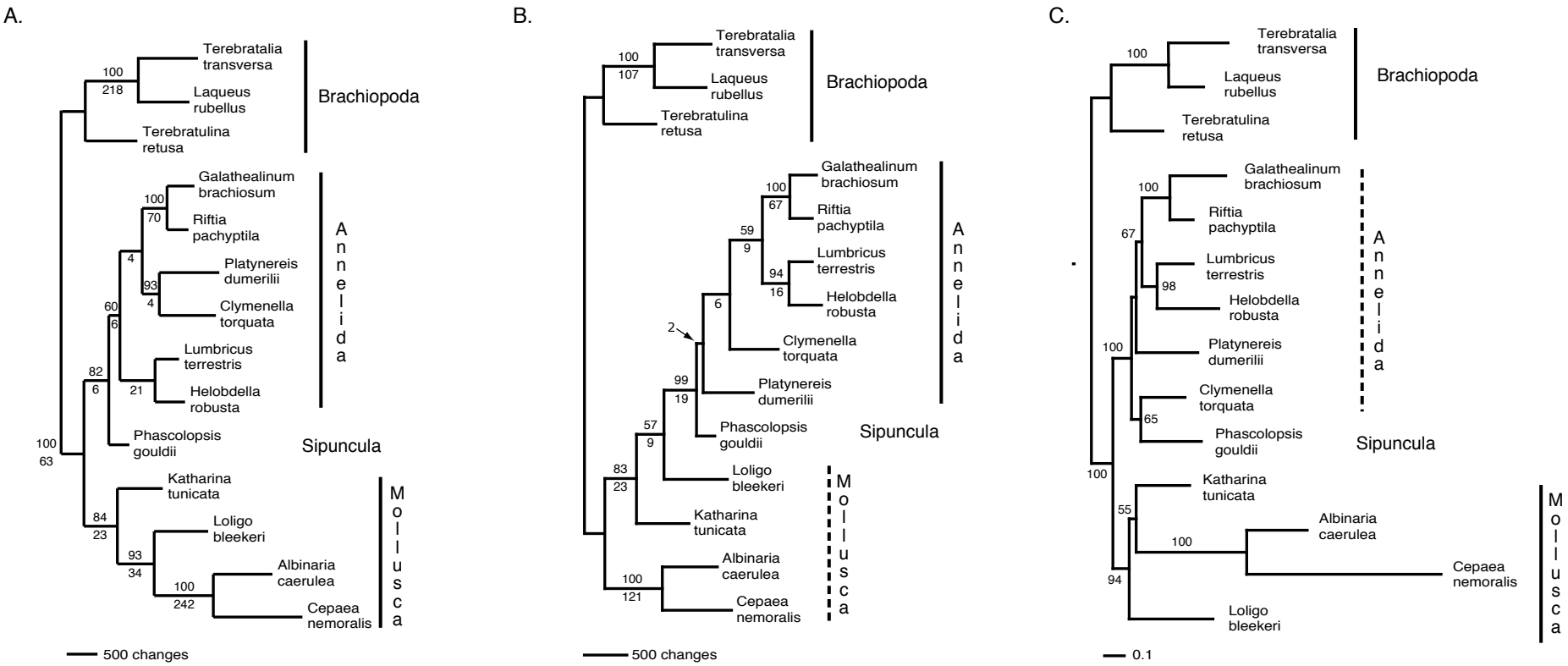


Tyrosine



Valine




A.
Terebratalia transversa
Laqueus rubellus
Terebratulina retusa
Brachiopoda
Galathealinum brachiosum
Riftia pachyptila
Platynereis dumerilii
Clymenella torquata
Lumbricus terrestris
Helobdella robusta
Annelida
Phascolopsis gouldii
Sipuncula
Katharina tunicata
Loligo bleekeri
Albinaria caerulea
Cepaea nemoralis
Mollusca
500 changes
B.
Terebratalia transversa
Laqueus rubellus
Terebratulina retusa
Brachiopoda
Galathealinum brachiosum
Riftia pachyptila
Lumbricus terrestris
Helobdella robusta
Clymenella torquata
Platynereis dumerilii
Annelida
Phascolopsis gouldii
Sipuncula
Loligo bleekeri
Katharina tunicata
Albinaria caerulea
Cepaea nemoralis
Mollusca
500 changes
C.
Terebratalia transversa
Laqueus rubellus
Terebratulina retusa
Brachiopoda
Galathealinum brachiosum
Riftia pachyptila
Lumbricus terrestris
Helobdella robusta
Platynereis dumerilii
Clymenella torquata
Annelida
Phascolopsis gouldii
Sipuncula
Katharina tunicata
Albinaria caerulea
Cepaea nemoralis
Loligo bleekeri
Mollusca
0.1