

APPLICABILITY OF MITOCHONDRIAL GENOME DATA TO ANNELID
PHYLOGENY AND THE EVOLUTION OF GROUP II INTRONS

Except where reference is made to the work of others, the work described in this dissertation is my own or was done in collaboration with my advisory committee.
This dissertation does not include proprietary or classified information.

Min Zhong

Certificate of Approval:

Zhanjiang Liu
Alumni Professor
Fisheries and Allied Aquacultures

Kenneth M. Halanych, Chair
Alumni Professor
Biological Sciences

Scott R. Santos
Assistant Professor
Biological Sciences

Leslie R. Goertzen
Assistant Professor
Biological Sciences

George T. Flowers
Dean
Graduate School

APPLICABILITY OF MITOCHONDRIAL GENOME DATA TO ANNELID
PHYLOGENY AND THE EVOLUTION OF GROUP II INTRONS

Min Zhong

A Dissertation

Submitted to

the Graduate Faculty of

Auburn University

in Partial Fulfillment of the

Requirements for the

Degree of

Doctor of Philosophy

Auburn, Alabama

August 10, 2009

APPLICABILITY OF MITOCHONDRIAL GENOME DATA TO ANNELID
PHYLOGENY AND THE EVOLUTION OF GROUP II INTRONS

Min Zhong

Permission is granted to Auburn University to make copies of this dissertation at its discretion, upon request of individuals or institutions and at their expense. The author reserves all publication rights.

Signature of Author

Date of Graduation

DISSERTATION ABSTRACT

APPLICABILITY OF MITOCHONDRIAL GENOME DATA TO ANNELID
PHYLOGENY AND THE EVOLUTION OF GROUP II INTRONS

Min Zhong

Doctor of Philosophy, August 10, 2009
(M.S., China Agricultural University, 2004)
(B.S., Qufu Normal University, 2001)

150 Typed Pages

Directed by Kenneth M. Halanych

Annelida is a very diverse group of segmented worms with over 16,500 described species. They play an important role in both terrestrial and aquatic environments. Despite this, their phylogenetic and evolutionary relationships of many groups within annelids are still poorly understood. This study focused on Terebelliformia annelids, a group of tube-dwelling worms used for investigation, comprises five recognized families with ambiguous phylogeny and a myzostomid worm, whose annelid affinity has been debated in recent years. In view of the conserved composition across bilaterians and the hypothesis of conserved gene order pattern across annelids, the mitochondrial genomic data have been becoming increasingly useful for applications to resolve this issue.

The main aim of this research is to characterize the mtDNA genomes in above annelids and resolve certain relationships within annelids with mitochondrial genome to evaluate their applicability to annelid phylogeny. Additionally, two group II introns that unexpectedly discovered in the myzostomid mtDNA drew our attentions to explore the evolution of such introns as they are rarely found in bilaterian genomes.

This study showed that the mitochondrial gene arrangement pattern is evolutionarily conserved as previously hypothesized, especially for protein-coding genes. Phylogenetic analyses based on the mitochondrial genome data indicated a well-resolved phylogeny within Terebelliformia group: Pectinariidae was placed as a basal clade to all other Terebelliformia families; Ampharetidae and Alvinellidae were sister to each other; Trichobranchidae and Terebellidae were sister clade with strong support. This suggests the great potential applicability of mitochondrial genomes which could likely be applied to the phylogenetic reconstruction of other annelid clades.

Two group II introns (divergent Mintron1 and degenerated Mintron2) which are characterized here in a partial mitochondrial genome of *Endomyzostoma* sp. (Myzostomida), is the first report of multiple introns in bilaterian genomes. The study implicated that both introns belong to the mitochondrial class and they could have independent origins given the dissimilarity between their RNA structures. It offers an important basis for the future studies in regard to the evolution and function of bilaterian group II introns.

Overall, the study implies an increasingly potential applicability to explore the mitochondrial genomes of annelids in terms of the phylogenetic and evolutionary examinations.

ACKNOWLEDGEMENTS

I wish to express my deepest appreciation to Dr. Ken Halanych for his great support, assistance and encouragement throughout my PhD training and the dissertation process. Thank you for the understanding especially in the last year of my study. I am also grateful to you for contributing a lot of time, effort and patience to my writings in English as a second language. Your guidance, comprehensive knowledge and precise attitude for the science are highly invaluable for my life. I would also like to express my heartfelt gratitude to the committee members, Dr. Zhanjiang Liu for his special support for my future career, Dr. Leslie Goertzen for his valuable comments in my written and oral exams and Dr. Scott Santos for his constructive feedback and suggestions for the manuscripts, as well as Dr Fenny Dane as my outside reader. Thanks for the financial support provided by the NSF-EPSCOR and CMB program. Many thanks go to all members in KMH lab: Torsten Struck, Nerida Wilson, Heather Eccleston, Dan Thornhill, Andy Mahon, Liz Borda, Rebecca Hunter, Alexis Janosik, Joie Cannon, Kevin Kocot for their helps during the research and lab work. Special thanks are extended to Rebecca and Alexis for their kind assistances to me the many ways in the aspect of personal and life. Finally, I would like to thank my husband, Yu Xiang, for his endurance and love. It would not be easy for me to study and live in a strange country without your support. Thanks for your constant encouragement for me to achieve the goal in my life. Special thanks to my sweet little daughter, Grace Xiang. Her birth brings the family great joy!

Style manual of journal used:

Chapter 1: *Gene*

Chapter 2: *Gene*

Chapter 3: *Systematic Biology*

Chapter 4: *Molecular Biology and Evolution*

Chapter 5: *Gene*

Computer software used: Adobe Illustrator, Microsoft Excel, Microsoft Word, MacClade, DNASTAR™ Lasergene (MegAlign, GeneQuest, SeqMan) ModelTest, MrBayes, MrModelTest, PAUP, RAxML, ProtTest, mfold, Clustal X, TreeView, Se-AL, Gblocks, tRNAscan, Artemis.

TABLE OF CONTENTS

LIST OF TABLES.....	xi
LIST OF FIGURES.....	xiii
CHAPTER 1: INTRODUCTION AND BACKGROUND.....	1
1. OVERVIEW.....	2
2. ANNELID MITOCHONDRIAL GENOMES.....	3
3. TEREPELLIDORMIA ANNELIDS.....	4
4. MYZOSTOMIDA.....	5
5. GROUP II INTRONS.....	6
APPENDIX: PHYLOGENETIC METHODS.....	8
REFERENCES.....	16
CHAPTER 2: PHYLOGENETIC INFORMATION FROM THE THREE MITOCHONDRIAL GENOMES OF TEREPELLIFORMIA (ANNELIDA) WORMS AND DUPLICATION OF THE METHIONINE tRNA.....	29
ABBREVIATIONS.....	30
ABSTRACT.....	31
1. INTRODUCTION.....	32
2. MATERIALS AND METHODS.....	33
2.1 SAMPLE COLLECTION AND DNA EXTRACTION.....	33
2.2 MTDNA DATA COLLECTION.....	34
2.3 GENOMIC ASSEMBLY.....	36
2.4 PHYLOGENETIC ANALYSIS.....	37
3. RESULTS AND DISCUSSION.....	39
3.1 GENOME COMPOSITION.....	39
3.2 PROTEIN-CODING GENES.....	40
3.3 tRNAs.....	42
3.4 UNK REGION.....	43
3.5 STRUCTURE AND GENE CODE.....	43
3.6 PHYLOGENETIC ANALYSIS.....	44
ACKNOWLEDGEMENTS.....	46
REFERENCES.....	46
CHAPTER 3: PHYLOGENETIC INFERENCE OF TEREPELLIFORMIA WORMS BASED ON MITOCHONDRIAL GENOMIC DATA.....	66

ABSTRACT.....	67
1. INTRODUCTION.....	68
2. MATERIALS AND METHODS.....	71
2.1 SAMPLE COLLECTION AND DNA EXTRACTION.....	71
2.2 MTDNA DATA COLLECTION.....	72
2.3 GENOMIC ASSEMBLY.....	73
2.4 PHYLOGENETIC ANALYSES.....	74
3. RESULTS.....	76
3.1 MITOCHONDRIAL GENOMES AND GENE ORDER.....	76
3.2 PHYLOGENETIC ANALYSES.....	77
4. DISCUSSION.....	78
4.1 PHYLOGENETIC RELATIONSHIP WITH TEREBELLIFORMIA.....	78
4.2 TRNM DUPLICATION EVENT AND GENE ORDER.....	79
CONCLUSIONS.....	80
ACKNOWLEDGEMENTS.....	81
REFERENCES.....	81

CHAPTER 4: TWO GROUP II INTRONS AND THEIR EVOLUTIONARY ORIGINS
IN *ENDOMYZOSTOMA* (MYZOSTOMIDA) MITOCHONDRIAL GENOME.....95

ABSTRACT.....	96
1. INTRODUCTION.....	97
2. MATERIALS AND METHODS.....	100
2.1 SAMPLE COLLECTION AND DNA EXTRACTION.....	100
2.2 MTDNA DATA COLLECTION.....	100
2.3 GENOMIC ASSEMBLY.....	101
2.4 INTRON IDENTIFICATION.....	102
2.5 PHYLOGENETIC ANALYSIS AND SEQUENCES ALIGNMENTS.....	102
3. RESULTS.....	103
3.1 GENOME COMPOSITION AND EVOLUTION.....	103
3.2 INTRON STRUCTURES.....	104
3.3 ORF ALIGNMENT AND PHYLOGENETIC ANALYSIS.....	104
4. DISCUSSION.....	105
4.1 MINTRON1: A DIVERGENT MITOCHONDRIAL GROUP II INTRON.....	106
4.2 MINTRON2: UNDERGOING EVOLUTIONARY DEGENERATION.....	107
4.3 INTRON EVOLUTIONARY ORIGINS.....	108
ACKNOWLEDGEMENTS.....	109
LITERATURE CITED.....	109

CHAPTER 5: CONCLUSIONS.....130

1. MITOCHONDRIAL GENOMES AND PHYLOGENETIC

RELATIONSHIPS WITHIN TEREPELLIFORMIA GROUP.....	131
2. THE MITOCHONDRIAL GENOME OF A MYZOSTOMIDA WORM AND GROUP II INTRONS.....	132
REFERENCES.....	134

LIST OF TABLES

CHAPTER 1

Table 1: Parameters in evolutionary models (Model abbreviations: JC, Jukes and Cantor (1969) model; K2P, Kimura (1980) two-parameter model; TrN, TrN model with equal base frequencies, Tamura and Nei 1993; SYM, Zharkikh 1994; F81, model of Felsenstein, Felsenstein 1981); HKY85, Hasegawa-Kishino-Yano model, Hasegawa et al., 1985; K3ST, Kimura 3 substitution type model, Kimura, 1981), GTR, General time-reversible model, Lanave et al., 1984; Rodrigues et al., 1990).....	25
--	----

CHAPTER 2

Table 1: Taxonomically-inclusive PCR primers to amplify small conserved regions.....	52
Table 2: Primers for long PCR with annealing temperatures.....	53
Table 3: Sequencing Primers for sequencing the mitochondrial genomes of <i>Terebellides stroemi</i> , <i>Pista cristata</i> and <i>Eclysippe vanelli</i>	54
Table 4: Taxa used in phylogenetic analysis.....	57
Table 5: Base composition and skewness measures.....	58
Table 6: Relative synonymous codon usages (RSCU) in the 13 protein-coding genes in <i>T. stroemi</i> , <i>P. cristata</i> and <i>E. vanelli</i>	59

CHAPTER 3

Table 1: Primers used in long-run PCR amplifications.....	87
Table 2: Primers used for sequencing the mitochondrial genomes of <i>Auchenoplax crinita</i> , <i>Paravinella sulfincola</i> and <i>Pectinaria gouldii</i>	88
Table 3: Taxa used in the phylogenetic analysis.....	91

CHAPTER 4

Table 1: Long PCR primers used in amplifications.....	116
Table 2: Sequencing primers used for sequencing mitochondrial genomes.....	117
Table 3: Intron identification primers.....	119
Table 4: Group II ORFs used in the phylogenetic analysis.....	120

LIST OF FIGURES

CHAPTER 1

Figure 1: Five families of Terebelliformia worms. A. Terebellidae; B. Avinellidae (http://www.mbari.org/expeditions/ridges2005/august_11.htm); C. Pectinariidae; D. Trichobranchidae; E. Ampharetidae (Both D and E are from Rouse and Pleijel, 2001).....26

Figure 2: Microscopic views of *Endomyzostoma* sp. (Figures are from Lanterbecq et al., 2006) B: *Endomyzostoma deformatum*; D: *Endomyzostoma* sp.....27

Figure 3: Substitution rates between all bases in evolutionary models.....28

CHAPTER 2

Figure 1: Gene orders of mitochondrial genomes in Annelida. Different colors shows conserved gene clusters. Dots indicate missing regions.....60

Figure 2: Putative secondary structures of tRNA genes in 3 Terebelliformia worms. (A) 23 tRNA genes in *T. stroemi*. (B) 23 tRNA genes in *P. cristata*. (C) 20 recovered tRNA genes in *E. vanelli*.....63

Figure 3: Phylogenetic reconstructions. (A) The single combined tree represents the identical topology from both ML and Bayesian inference methods (non-partitioned and partitioned) with GTR+ Γ +I model. Nodal support values are given at branches with ML bootstrap values first and posterior probabilities of the partitioned Bayesian analysis second (non-partition values not shown). A dash indicates < 50% on trees. (B) Non-partitioned Bayesian analyses of amino acid dataset with the mixed amino acid substitution model. Posterior probabilities are shown at the nodes. (C) ML analyses of amino acid dataset using RAxML by 200 bootstrap replicates. Bootstrap values are shown at the nodes. Black bars indicate *trnM* gene duplication event. *Terebratalia transversa* (brachiopod) and *Katharina tunicata* (mollusk) were used as outgroups. Details of analyses are given in the text.....64

Figure 4: Partitioned Bayesian analyses of amino acid dataset with the mixed amino acid substitution model with partitions unlinked during the run. Posterior probabilities are shown at the nodes.....65

CHAPTER 3

Figure 1: Mitochondrial gene order of six Terebelliformia worms. Different colors shows conserved gene clusters. Dots indicate missing regions.....92

Figure 2: Phylogenetic reconstructions for 17-taxon mitochondrial datasets. (A) The mitochondrial nucleotide tree represents the identical topology from both ML and partitioned Bayesian inference methods. (B) The mitochondrial amino acid tree with the identical topology from both ML and partitioned Bayesian inference methods. Nodal support values are given at branches with posterior probabilities of the partitioned Bayesian analysis first and ML bootstrap values second. A dash indicates < 50% on trees. The representative branch lengths are from the Bayesian trees.....93

Figure 3: Character mapping on Terebelliformia ingroup trees based on mitochondrial datasets. The bar represents *trnM* duplication event. The circle represents the specific gene order of *nad5-nad4L-nad4* cluster in mitochondrial genomes. Open bar: duplication absent; Black bar: duplication present; open ellipses: *nad5- trnF-trnE-trnP-trnT-nad4L-nad4*; black ellipses: *trnE-trnT-trnP-trnY--nad4L-nad4-trnC-trnM-trnH-nad5*; shaded ellipse: *nad5- trnF-trnE-trnP-trnT-trnR-nad4L-nad4*.....94

CHAPTER 4

Figure 1: The secondary structures of 14 discovered tRNA in the mitochondrial genome of *Endomyzostoma* sp.....125

Figure 2: (A) Insertion locations of both Mintron1 and Mintron2 in the *cox1* gene. The dot line on the left end of *cox1* gene means this region was not recovered. (B) Secondary structure of Mintron1. Domain 4 contains the ORF. (C) Secondary structure of Mintron2. The stars label the conserved nucleotides in the central core as the major features of group IIA1 subclass which is the mitochondrial class. ORF, open reading frame; EBS, exon binding sequences; IBS, intron binding sequences.....127

Figure 3: (A) Intron-encoded ORF domain structure. (modified from Zimmerly et al., 2001; Fig1A). (B) Alignment of subdomain 4 region of RT domain and one conserved region in X domain. The most conserved amino acids were identified by Gblocks showing as the black bars. The representative ORFs from each major category (three mitochondrial, two chloroplast and four bacteria) were selected to align with the ORFs from both Mintron1 and *Nephtys* group II intron. The purple bars are showing the most conserved regions across all ORFs. Species abbreviations are as follows: M.m.I1, *Methanosarcina mazei*; A.v.I1, *Azotobacter vinelandii*; W.e.I2, *Wolbachia* endosymbiont of *Drosophila melanogaster*; Kl.pn.I1, *Klebsiella pneumoniae isolate YMC*. The intron category labels: Bac, Bacteria; Chl, Chloroplast; Mt, Mitochondrion. Other

information about the taxa see table 4.....128

Figure 4: Phylogenetic analyses of 105 group II intron ORFs by the MrBayes program. Taxa sampling included all mitochondrial and chloroplast group II ORFs and some representatives of bacterial classes (table 4). Dark shades are the mitochondrial group II introns including the Mintron1 and *Nephtys* ORFs (showing as bold font). Light shade is the chloroplast group II intron lineage. All other unshaded are the group II intron ORFs found in bacteria. Archaeobacterial taxa were used as outgroups.....129

CHAPTER 1

INTRODUCTION AND BACKGROUND

1. Overview

Annelida is a very diverse group of animals living in marine, freshwater and terrestrial environments. It comprises around 16,500 described species and this number is likely underestimated as annelids are the most abundant macrofauna in the unexplored deep-sea area (Grassle and Maciolek, 1992). Traditionally, Annelida was divided into Polychaeta and Clitellata, the latter including Oligochaeta and Hirudinida. However, recent phylogenetic studies have indicated a paraphyletic Polychaeta that includes several previously recognized ‘phyla’ (echiurids, sipunculids, and siboglinids – a.k.a. pogonophorans; McHugh, 1997; Boore and Staton, 2002; Bleidorn et al., 2006a; Struck et al., 2007) and the Clitellata. Annelids are typically an important food source for many economically important species and play a key role in nutrient cycling and energy flow in terrestrial and aquatic environments (Losteste and Marchese 1994; Aston, 1984). Annelids often have abundant populations with diverse life histories. Because they are environmentally important and easy-to-handle in the lab, they are a good candidate for the model of lophotrochozoan animals.

Despite their importance, we still have a poor understanding of annelid evolutionary relationships, and they continue to be debated. The overall goal of this project is to resolve certain relationships within annelids and to characterize mitochondrial genomes including their applicability to annelid phylogeny.

To this end, I focus on a well-described group, Terebelliformia annelids, comprised of five recognized families, and aim to determine their evolutionary history using mitochondrial DNA (mtDNA) data. Furthermore, I examine myzostomids, a group of worms whose annelid affiliation has been debated. Molecular evolution of myzostomid mtDNA is interesting because of the unexpected presence of two group II introns.

2. Annelid mitochondrial genomes

Gene content of the mtDNA genome is conserved across Bilateria. The mitochondrial genome contains 13 protein-coding genes, two ribosomal genes, 22 tRNA genes and one non-coding unknown region (UNK - the presumed origin of replication). Generally, the size of mitochondrial genomes is about 15-17 kB (Boore, 1999; Vallès and Boore, 2006). Whereas some taxa have highly rearranged mitochondrial genomes, gene order within annelids is hypothesized to be relatively conserved (Jennings and Halanych, 2005; Vallès and Boore, 2006; Zhong et al., 2008). Despite the utility of mitochondrial genomes for phylogenetic reconstructions of major animal taxa, only seven complete and five partial annelid mitochondrial genomes are available in GenBank (September 2008). One reason for the paucity of mtDNA genome data is that in some taxa, including annelids, the UNK region (also called the D-loop or control region) has been difficult to amplify and sequence presumably because of secondary structure and/or the present of microsatellite regions (Boore and Brown, 2000; Boore, 2001; Jennings and Halanych, 2005; Bleidorn et al., 2006a, b; Zhong et al., 2008).

3. Terebelliformia annelids

In order to accurately interpret annelid genomic evolution and molecular phylogeny, I used Terebelliformia annelids as a test group.

Terebelliformia annelids, also known as Terebellomorpha, are mainly tube-dwelling worms, found in diverse marine habitats, including the intertidal, coral reefs, the deep-sea and hydrothermal vents, and can be found in both soft and hard substrates (Fig. 1). They are sedentary in nature and have a well-developed anterior end with modified appendages for feeding and respiration. Terebelliformia worms are comprised of five ‘families’, Alvinellidae, Ampharetidae, Terebellidae, Trichobranchidae and Pectinariidae (Hessle, 1917; Holthe, 1986; Rouse and Pleijel, 2001). Even though Terebelliformia is one of the few clades in Annelida argued to be well defined by morphology, phylogenetic relationships among these five families are ambiguous and consistently debated (Rouse and Fauchald, 1997; Colgan et al., 2001; Rousset et al., 2003; Glasby et al., 2004; Rousset et al., 2007; Struck et al., 2007).

Based on morphological data, Trichobranchidae was originally treated as the subfamily Trichobranchinae within Terebellidae (Malmgren, 1866) and elevated to family rank in 1917 because two genera of Trichobranchidae (*Trichobranchus* and *Terebellides*) were suggested to be allied with Ampharetidae and Terebellidae, respectively, based on similar characteristics of their branchiae and digestive system, respectively (Hessle, 1917). Rouse and Fauchald (1997) suggested a close affinity between Terebellidae and Trichobranchiae, which was sister to the Alvinellidae/Ampharetidae/Pectinariidae clade. However, Fauchald and Rouse (1997) treated Pectinariidae together with Terebellidae and Ampharetidae as Terebellida or Terebellimorpha. Recent morphological analyses

indicated that Pectinariidae was either sister to the Ampharetidae/Alvinellidae clade (Rouse and Fauchald, 1997), or related to Terebellidae (Glasby, et al., 2004).

Nevertheless, some molecular studies and combined data indicate contradicting views of relationships within Terebelliformia. *Trichobranchus* was suggested to be associated with one subfamily of Terebellidae making Trichobranchidae paraphyletic based on five mitochondrial and nuclear genes (Colgan et al., 2001), whereas a sister relationship between Trichobranchidae and Alvinellidae was suggested based on morphology combined with molecular data (Rousset et al., 2003). Additionally, the affinity of Pectinariidae has been questioned based on recent molecular and combined data. Its sister relationship to Trichobranchidae,/Alvinellidae clade (Rousset et al., 2003), within Ampharetidae (Colgan et al., 2001), or close to all other four terebelliform lineages (Struck et al., 2007) or even other annelids (which would break up the monophyly of Terebelliformia) (Rousset et al., 2007) indicates that Pectinariidae is a family of ambiguous phylogenetic position influencing phylogenetic relationships within terebelliforms.

4. Myzostomida

Myzostomida are small parasitic marine worms, living mostly on crinoid and ophiuroid echinoderms (Grygier, 2000; Zrzavý et al., 2001; Bleidorn et al., 2007). Their body is typically soft and flattened with many radiating cirri on the thin rounded body edge and five pairs of parapodia on the ventral side (Fig. 2). They have high host-specificity as many myzostomid species are only associated with a single crinoid species (Eeckhaut et al., 1998), with a few exceptions (Lanterbecq et al., 2006). Fossils from the Ordovician suggest an ancient association between myzostomids and their crinoid hosts

(Eeckhaut et al., 1998), which may explain the highly derived body plan of myzostomids and their disputable phylogenetic position within metazoans (Zrzavý et al., 2001; Bleidorn et al., 2007). Although some studies predicted that Myzostomida are not allied with annelids (Haszprunar, 1996; Eeckhaut et al., 2000; Littlewood et al., 2001; Zrzavý et al., 2001; Giribet et al. 2004), their polychaete origins are indicated by both morphological (Eeckhaut et al., 1998; Haszprunar, 1996) and molecular studies (Bleidorn et al 2007) with robust support. The monophyletic Myzostomida group comprises two orders (Proboscidea and Pharyngidea), 12 genera and about 170 described species (Lanterbecq et al., 2006). *Endomyzostoma* sp. belongs to Endomyzostomatidae (Pharyngidea).

Originally I began work on myzostomid mtDNA to determine their phylogenetic placement. However, I discovered interesting group II introns in their mitochondrial genome. Additionally, a colleague, Christoph Bleidorn (Free University of Berlin), was also working on myzostomid mtDNA. To avoid duplication, I focused my efforts on the introns and have been working with Bleidorn on the phylogenetic issues.

5. Group II introns

Group II introns are self-splicing ribozymes that are commonly present in the genomes of bacteria, and organellar genomes of fungi, plants and animals (Lehmann and Schmidt, 2003; Rot et al., 2006; Beagley et al., 1998; Dellaporta et al., 2006; Vallès et al., 2008). The secondary structure of a typical group II intron contains 6 obvious stem-loop domains, D1 to D6. All domains radiate from a central core comprised of a few nucleotides to form the proximal helix that is required for self-splicing. There are several conserved primary sequences which are the key identifiers of group II introns including

5' and 3' splicing sites (5'-GUGYG and AY-3'), an unpaired adenosine in D6 and a number of conserved nucleotides in D5 (Knoop et al., 1994; Lambowitz and Zimmerly, 2004).

Group II introns are mobile and generally have two forms, OpenReadingFrame (ORF)-containing and ORF-less which only consists of the six domains. ORF-less introns primarily occur in organelles rather than bacteria. All ORFs in group II introns are found in D4 that involve four functional domains: reverse transcriptases (RTs) domain, X domain with maturase activities, non-conserved D domain (for DNA-binding) and En domain with endonuclease activity (Lambowitz et al., 1999; Zimmerly et al., 2001; San Filippo and Lambowitz, 2002; Dai and Zimmerly, 2002). RTs in bacterial group II introns were believed to mainly function as retroelements and play an essential role in intron homing events (Dai and Zimmerly, 2002). Intron homing is a common mobile phenomenon widely distributed in protists and other eukaryotes. Intron homing is enacted by both RNA-catalytic and RT activities. Target homing sites, both upstream and downstream of exons, are highly specific and provide an additional tool for intron classification. ORFs are also used for intron identification analyses because normally ORFs in one category are related to their specific structures of intron RNAs (Zimmerly et al., 2001; Toor et al., 2001; Dai and Zimmerly, 2002; Vallès et al., 2008). Based on phylogenetic analyses of ORFs, group II introns were divided into three major categories, mitochondrial (including fungi), chloroplast and bacterial lineages. ORFs in bacterial introns are considered basal to mitochondrial and chloroplast ORFs (Toor et al., 2001). An evolutionary model of group II introns was predicted where all currently known group II introns, including ORF-containing and ORF-less, are derivatives of bacterial

ORF-containing introns and most ORF-less introns derived from ORF-containing introns based on the relative high identity from sequence alignments among several liverwort mt-introns and land plants (Toor et al., 2001).

The first report of a group II intron in the mitochondrial genome of any bilaterian was from the annelid *Nephtys* sp. which was inferred to be derived from a recent horizontal gene transfer (Vallès et al., 2008).

Appendix: Phylogenetic Methods

Phylogenetics is a subject of studying evolutionary relationships among organisms. It can reconstruct the phylogeny through various methods. Evolution is the central idea running through the whole process to build the phylogeny. The “evolution” can be tracked and expressed using some ways which make it possible to reconstruct the original relationships among them. The most common methods involved in inferring phylogenetics or for constructing trees from morphological and genetic sequence data are distance, parsimony and likelihood. Distance is an algorithmic approach which can only yield one single tree with high speed from any given datasets. Whereas parsimony and likelihood are tree-searching approaches that can construct many trees and then performs some criteria to choose the best tree(s) with relative lower speed.

Distance

Algorithm

Whenever we want to start a phylogenetic method, the first thing we need to consider is the algorithm to make the tree. Neighbor-joining is a widely used algorithm

for tree-building (Saitou and Nei, 1987). This method infers phylogeny by using a pairwise distance matrix which can be generated from various sources, morphological data, genetic data, etc. Basically, the algorithm calculates the number of changes between pairwise sequence data as different distances to create the matrix. Then, it requires to find a pair of taxa with lowest distance value to join together and create a node, followed, find the other taxa with the lowest value to the node, so on and so forth (join the closest neighbors). Obviously, the reconstructed tree by using NJ method does not really stand for the one with the smallest branch length. However, the calculation is very fast as compared to other methods especially with large datasets. In addition, it has been proven to be comparatively accurate and statistically consistent. Therefore, with large datasets, NJ method is always first used to quickly preview the reconstructed phylogeny. This algorithm allows selection of a specific model to correct the distance matrix when it is calculated. The simple measure of the changes will be the directly observed difference between pairwise. However, it can underestimate the actual changes by ignoring multiple changes occurring at a certain site (i.e. convergent, parallel, reversal and so on). So as to minimize errors due to multiple changes, models can be chosen to correct and estimate the actual number of changes (See *Model* part in Maximum Likelihood).

The simplest distance algorithm is UPGMA (Unweighted Pair Group Method with Arithmetic mean), which assumes a constant evolution rate (the existence of molecular clock) (Sneath and Sokal 1973). After constructing the distance matrix, the nearest two groups will be clustered into a higher group level. Then, by taking average of the distances between members in one group with members in the other to decide how groups are related. Although it was once used in protein electrophoresis, this method has

been seldom used in recent studies due to the inappropriate constant evolution rate assumption and has totally been replaced by NJ (Neighbor-joining) method.

Optimality Criterion

After obtaining a tree, optimality criteria should be implemented to search or estimate the best-fit tree. Minimum evolution is an optimal criterion used in the distance approach which assumes that the best phylogenetic tree has the smallest branch length (distances) by calculating the pairwise from distance matrix. This method will compute all possible trees and then compare branch lengths for those trees. However, the branch lengths calculated under distance approach may not be evolutionary interpretable. Additionally, because distance uses the transformed pairwise matrix to build trees, it is eventually possible to lose some information from the original dataset, for instance, DNA alignment.

Maximum Parsimony

As compared to the cluster-based distance method (NJ), discrete methods which are character-based, like parsimony, maximum likelihood and Bayesian inferences operate directly on the sequence alignment by comparing characters at each site, and thus avoid loss of information.

Algorithm

The algorithm under maximum parsimony approach can also start from either a NJ tree or a random tree.

Optimality Criterion

Maximum parsimony is a non-statistical character-based criterion. Maximum parsimony tree with fewest total number of evolutionary changes or substitutions (minimize the total tree length) requires to explain the differences across taxa in a given dataset (e.g., Farris, 1970). The assumption is that all investigated taxa share a common characteristic since they all evolved from a common ancestor without considering the homoplasy which makes a character in a same state with different evolutionary history and origins caused by reversal, convergence and parallelism. The total number of evolutionary changes on a tree (branch length of this tree) is simply the sum of the change numbers at each site. Under the criterion of maximum parsimony, a tree-searching strategy will be performed to look for the best tree(s) by calculating the total number of changes for each tree.

Problems with parsimony

Parsimony doesn't guarantee finding the best tree, multiple trees may be the most parsimonious with the same branch length in some cases. So, an unresolved consensus tree can be constructed to visualize the congruity. The most parsimonious tree(s) may not be the true one. Additionally, it doesn't correct for multiple mutations within one site (homoplasy) because it doesn't imply an evolutionary model. Branch lengths cannot be accurately estimated (underestimated). It is relative slow in computing time in comparison with distance methods.

Maximum likelihood

Algorithm

The algorithm under maximum likelihood approach can also allow getting start from either a NJ tree or a random tree.

Optimality Criterion

Maximum likelihood (ML) is a powerful tree-searching criterion and very widely used in phylogenetic analyses (Cavalli-Sforza and Edwards, 1967; Felsenstein, 1981).

The principle of ML is the phylogeny with the highest likelihood is taken to be correct.

When we calculate the likelihood of a phylogeny, we score the probability of all possible ancestral states across each site for each possible topology based on the specific evolutionary model.

ML method of inference is available for both nucleic acid and protein data using PAUP for nucleic acids (Swofford, 2002) and RAxML for amino acid datasets (Stamatakis, 2006), for example. It always generates one single tree with the lowest tree length and highest likelihood score. It is most likely to be the true tree. A likelihood tree has meaningful branch lengths. If a branch length is equal to 1, meaning that one change per character on that branch.

Given the model and a possible tree, ML calculates the likelihood of observing the data:

$$L(\text{Model}) = P(\text{Data}|\text{Model}),$$

First of all, we select a specific model for a data alignment and generate all possible tree structures for each site. The state of each character within that site is plotted on the tree.

Based on the selected model, the probability of the distribution of character states is calculated to get the probability of that character, given the tree. Then, the probabilities of all of the characters is multiplied together in form of negative log-transformed and sum the probabilities for each position to get the probability of the data, given the tree. Finally, a tree is chosen with the highest likelihood (lowest negative log-transformed likelihood).

Evolutionary model

ML searches for an evolutionary tree with the highest probability of observing the data given an explicit model. Different models have variable parameters to express their evolutionary process. For instance, one-parameter model (Jukes-Cantor model, Jukes and Cantor, 1969) assumes equal nucleotide frequencies of all four bases (A, T, G, C) and constant substitution rate between all bases (Fig 3. $a=b=c=d=e=f$). In Kimura's two-parameter model (Kimura, 1980), different substitutions occur at different rates (transitions occur at rate $a=c=d=f$, transversions occur at rate $b=e$) with equal nucleotide frequencies. The F81 model assumes four unequal base frequencies with constant substitution rate (Felsenstein, 1981). Other models with different modifications of substitution rates and base frequencies are shown in Table 1.

The most commonly used model for DNA data is the general time-reversible (GTR) model which combines all of the above models and assumes six different substitution rates with unequal base frequencies as special cases (Lanave et al., 1984; Barry and Hartigan, 1987; Rodriguez et al., 1990). This model can be time-reversible. Taking evolutionary rate heterogeneity among the lineages of particular genes into account, GTR+G+I is often modified from GTR with the discrete Gamma distribution (G) and addition of invariable sites (I). Gamma distribution is used to correct the substitution

rate across sites by estimating rates across all sites. Gamma distribution has two parameters, a (shape parameter) and b (scale parameter). As a increases, higher frequency of variable sites have equal substitution rate (equal-rate model). As a decreases, stronger among-site variation in the data. Gamma distribution models the variant distribution among sites with different substitution rates. With this parameter, rate heterogeneity can be corrected.

Furthermore, because almost all genes have invariable sites across lineages, the proportion of invariable positions may change with an enlarged time scale or different lineages, which is usually considered as a plus parameter.

A specific model can be selected by Modeltest program (Posada and Crandall, 1998) for nucleotide dataset or ProtTest for amino acid dataset (Abascal et al., 2005).

Advantages and disadvantages

Because ML requires computation of likelihood for all the candidate topologies of each site, it evaluates all possible tree topologies and calculates all their likelihoods. It provides more confidence than other methods. Additionally, ML is a model-based method which tends to be robust. The big problem of ML is that it is time consuming and computationally intensive. It is virtually impossible to compute all the probabilities when taxa number is large. Additionally, results are based on the model chosen. A bad model can lead to a tree with less reliability.

Tree-searching strategy under optimality criteria

Under the selected optimality criterion, the optimal tree(s) need(s) to be evaluated among all searched trees by tree-searching methods. NJ tree is often treated as a starting

point for tree-searching. The methods include exhaustive search which uses exact algorithm (with branch-and-bound or branch-addition algorithm) and heuristic strategy (with stepwise addition, branch swapping or star decomposition method; Hendy and Penny, 1982; Swofford, 1996). Exhaustive search evaluates every tree and guarantees finding the optimal tree. It can only conduct small datasets. Heuristic search can be used to search trees with a larger dataset, however, it can't guarantee that the best tree is found and can result in more than one equally parsimonious tree. So a consensus tree needs to be created to mix all possible relationships in one unresolved tree. The reason that heuristic search can't guarantee the best tree is because it is essentially based on a hill-climbing algorithm in which a single initial tree is built then rearrangements are performed to improve the tree and keep the tree better than the previous one.

Bootstrapping

Once we obtain the best tree, its topology with the order of branches may not be very reliable. So it is very important to perform a statistical estimation for a reliability test. Bootstrapping is a way to estimate the reliability of the tree by generating a pseudoreplicate datasets (alignment) and creating trees based on those datasets using same tree-construction methods (Bradley et al., 1996). The randomly resampling method is used to sample the original alignment so as to produce a new character alignment with same size as the original one. The resampling allows replacement meaning that the same site can occur in new alignment more than once or disappear. In phylogenetic analyses, nonparametric bootstrapping is very commonly used. The frequency that a given branch can be found during bootstrapping will be written as a proportion number on each branch. These numbers can measure the reliability of individual branches. Since the larger the

number, the higher the reliability. These frequency numbers can examine how often an individual branch appears in resampled pseudoalignment and evaluate their reliability, the larger the number, the higher the reliability for each branch. Both parsimony and likelihood method need bootstrapping after a tree is obtained.

Bayesian Inference

The Bayesian approach is relative new and has a strong connection with likelihood method (Geyer, 1991; Hueslenbeck and Ronquist, 2001). The only difference is that it calculates the posterior probability which is proportional to the likelihood multiplied by the prior probability (Yang and Rannala, 1997; Larget and Simon, 1999). The product of these two quantities over all possible parameter values is integrated to score the posterior probability for the tree.

In Bayesian approach, the Markov chain Monte Carlo (MCMC) algorithm is implemented. MCMC can approximate the probability distribution over a vast range. MCMC works by forming a conceptual chain (Larget and Simon, 1999). In each step, it proposed a new location as the next link of the chain. Then the posterior-probability density of that new location is calculated. If the score is higher than the one in current location of the chain, the new location will be the next link of the chain so as to push the chain moving forward. If the score is lower than the current one, the chain will keep its present location and not move on until it finds a new one with higher score. After repeating the same procedure a million times, a long chain forms and would always stay in locations with the highest posterior probability. Then, it will combine all trees to calculate the probability of the appearance for each tree, the highest probability, the best tree. During phylogenetics, MCMC forms a long chain by estimating the posterior

probability through different trees and evolutionary models.

References

- Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: Selection of best-fit models of protein evolution. *Bioinformatics*. 21(9), 2104-2105.
- Aston, R.J., 1984. The culture of *Branchiura sowerbyi* (Oligochaeta: Tubificidae) using cellulose substrate. *Aquaculture* 40(1), 89–94.
- Barry, D., Hartigan, J.A., 1987, Asynchronous distance between homologous DNA sequences. *Biometrics*. 43, 261–276.
- Beagley, C.T., Okimoto, R., Wolstenholme, D.R., 1998. The mitochondrial genome of the sea anemone *Metridium senile* (Cnidaria): introns, a paucity of tRNA genes, and a near-standard genetic code. *Genetics* 148, 1091–1108.
- Bleidorn, C., Podsiadlowski, L., Bartolomaeus, T., 2006a. The complete mitochondrial genome of the orbiniid polychaete *Orbinia latreillii* (Annelida, Orbiniidae)--A novel gene order for Annelida and implications for annelid phylogeny. *Gene* 370, 96–103.
- Bleidorn, C., Kruse, I., Albrecht, S., Bartolomaeus, T., 2006b. Mitochondrial sequence data expose the putative cosmopolitan polychaete *Scoloplos armiger* (Annelida, Orbiniidae) as a species complex. *BMC Evol. Biol.* 6, 47.
- Bleidorn, C., Eeckhaut, I., Podsiadlowski, L., Schult, N., McHugh, D., Halanych, K.M., Milinkovitch, M.C., Tiedemann, R., 2007. Mitochondrial genome and nuclear sequence data support Myzostomida as part of the annelid radiation. *Mol. Biol. Evol.* 24(8): 1690–1701.

- Boore, J.L., 1999. Animal mitochondrial genomes. *Nucleic Acids Res.* 27, 1767–1780.
- Boore, J.L., 2001. Complete mitochondrial genome sequence of the polychaete annelid *Platynereis dumerilii*. *Mol. Biol. Evol.* 18, 1412–1416,
- Boore, J.L., Brown, W.M., 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., Staton, J.L., 2002. The mitochondrial genome of the Sipunculid *Phascolopsis gouldii* supports its association with Annelida rather than Mollusca. *Mol. Biol. Evol.* 19, 127–137.
- Bradley, E., Halloran, E., Holmes, S., 1996. Bootstrap confidence levels for phylogenetic trees. *PNAS.* 93, 23.
- Cavalli-Sforza, L.L., Edwards, A.W.F., 1967. Phylogenetic analysis: Models and estimation procedures. *Evolution* 32: 550-570 and *Am. J. Hum. Genet.* 19:233–257.
- Clark, B.F.C., 2006. The crystal structure of tRNA. *J. Biosciences.* 31, 453–457.
- Colgan, D.J., Hutchings, P.A., Brown, S., 2001. Phylogenetic relationships within the Terebellomorpha. *J. Mar. Biol. Ass. U.K.* 81, 3806/1–9.
- Dai, L., Zimmerly, S., 2002. Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic Acids Res.* 30, 1091–1102.
- Dellaporta, S.L., Xu, A., Sagasser, S., Jakob, W., Moreno, M.A., et al., 2006. Mitochondrial genome of *Trichoplax adhaerens* supports placozoa as the basal

- lower metazoan phylum. *Proc. Natl. Acad. Sci. U S A.* 103, 8751–8756.
- Eeckhaut, I., 1998. *Mycomyzostoma caldicola* gen. et sp. n., the first extant myzostomid infesting crinoid stalks, with a nomenclatural appendix by M. J. Grygier. *Spec. Div.* 3:89–103.
- Eeckhaut, I., McHugh, D., Mardulyn, P., Tiedemann, R., Monteyne, D., Jangoux, M., Milinkovitch, M.C., 2000. Myzostomida: a link between trochozoans and flatworms? *Proc. R. Soc. Lond. B.* 267, 1383–1392.
- Farris, J.S., 1970. Methods for computing Wagner trees. *Syst. Zool.* 34, 21–34.
- Fauchald, K., Rouse, G.W., 1997. Polychaete systematics: Past and present. *Zool. Scr.* 26, 71–138.
- Felsenstein, J., 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* 17, 368–376.
- Geyer, C.J., 1991 Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium of the Interface* (ed. E.M. Keramidas), pp. 156–163. Interface Foundation, Fairfax Station, VA.
- Giribet, G., Soerensen, M.V., Funch, P., Kristensen, R.M., Sterrer, W., 2004. Investigations into the phylogenetic position of Micrognathozoa using four molecular loci. *Cladistics* 20, 1–13.
- Glasby, C.J., Hutchings, P.A., Hall, K., 2004. Assessment of monophyly and taxon affinities within the polychaete clade Terebelliformia (Terebellida). *J. Mar. Biol. Ass. U.K.* 84, 961–971.
- Grassle, J.F., Maciolek, N.J., 1992. Deep-sea species richness: regional and local diversity estimates from quantitative bottom samples. *Am. Nat.* 139, 313–341.

- Grygier M.J., 2000. Myzostomida. In: Beesley PL, Ross GJB, Glasby CJ, editors. Polychaetes and Allies: The Southern Synthesis. Fauna of Australia, Vol 4A Polychaeta, Myzostomida, Pogonophora, Echiura, Sipuncula. Melbourne: CSIRO Publishing. 297–330.
- Hasegawa, M., Kishino, H., Yano, T., 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 21, 160–174.
- Haszprunar G., 1996. The Mollusca: coelomate turbellarians or mesenchymate annelids? In: Taylor JD, editor. *Origin and Evolutionary radiation of the Mollusca*. Oxford: University Press. 3–28.
- Hendy, M.D., Penny, D., 1982. Branch and bound algorithms to determine minimal evolutionary trees. *Mathematical Bioscience* 59:277–290.
- Hessle, C., 1917. Zur Kenntnis der terebellomorphen polychaeten. *Zoologiska Bidrag från Uppsala*. 5, 39–25.
- Holthe, T., 1986. Evolution, systematic and distribution of the Polychaeta Terebellomorpha, with a catalogue of the taxa and a bibliography. *Gunneria*. 55, 1–236.
- Huelsenbeck, J.P., Ronquist, F., 2001. MrBayes: Bayesian inference in phylogenetic trees. *Bioinformatics*, 17, 754–755.
- Jennings, R.M., Halanych, K.M., 2005. Mitochondrial genomes of *Clymenella torquata* (Maldanidae) and *Rifta pachyprila* (Siboglinidae): Evidence for conserved gene order in Annelida. *Mol. Biol. Evol.* 22, 210–222.
- Jukes, T., Cantor, C., 1969. Evolution of protein molecules. In *Mammalian protein metabolism* (ed. H. Munro), pp. 21–132. New York: Academic Press.

- Kimura, M., 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120.
- Kimura, M., 1981. Estimation of evolutionary distances between homologous nucleotide sequences. *PNAS UAS.* 78, 454–458.
- Knoop, V., Kloska, S., Brennicke, A., 1994. On the identification of group II introns in nucleotide sequence data. *J. Mol. Biol.* 242, 389–396.
- Lambowitz, A.M., Zimmerly, S., 2004. Mobile group II introns. *Annu. Rev. Genet.* 38: 1–35.
- Lambowitz, A.M., Caprara, M.G., Zimmerly, S., Perlman, P.S., 1999. Group I and group II ribozymes as RNPs: Clues to the past and guides to the future. In *The RNA world*, 2nd ed., pp. 451–485. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Lanave, C., Preparata, G., Saccone, C., Serio, G., 1984. A new method for calculating evolutionary substitution rates. *J. Mol. Evol.* 20, 86–93.
- Larget, B., Simon, D.L., 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* 16, 750–759.
- Lehmann, K., Schmidt, U., 2003. Group II introns: structure and catalytic versatility of large natural ribozymes. *Crit. Rev. Biochem. Mol. Biol.* 38, 249–303.
- Lanterbecq, D., Rouse, G.W., Milinkovitch, M.C., Eeckhaut, I., 2006. Molecular phylogenetic analyses indicate multiple independent emergences of parasitism in Myzostomida (Protostomia). *Syst. Biol.* 55, 208–227.
- Littlewood, D.T.J., Olson, P.D., Telford, M.J., Herniou, E.A., Riutort, M., 2001. Elongation Factor 1- alpha sequences alone do not assist in resolving the position

- of the Acoela within Metazoa. *Mol. Biol. Evol.* 18, 437–442.
- Loteste, A., Marchese, M., 1994. Ammonium excretion by *Paranadrilus descolei* Gavrilov, 1955 and *Limnodrilus hoffmeisteri* Claparède, 1862 (Oligochaeta: Tubificidae) and their role in nitrogen delivery from sediment. *Pol. Arch. Hydrobiol.* 41(2), 189–194.
- Malmgren, A. J., 1866. Nordiska Hafs-Annulater. Öfversigt af Kongl. Vetenskaps-Akademiens Förhandlingar, 22, 355–410.
- McHugh, D., 1997. Molecular evidence that echiurans and pogonophorans are derived annelids. *Proc. Natl. Acad. Sci. U S A.* 94, 8006–8009.
- Palumbi, S.R., 1996. Nucleic acid II: the polymerase chain reaction. Pp. 205–247 in *Molecular Systematics*, Hillis, D.M., Moritz, C., Mable, B.K. eds. Sinauer Associates, Sunderland, MA.
- Perna, N.T., Kocher, T.D., 1995. Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. *J. Mol. Evol.* 41, 535–558.
- Posada, D., Crandall, K.A. 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics* 14, 817–818.
- Rambaut, A., 1996. The Use of Temporally Sampled DNA Sequences in Phylogenetic Analysis. PhD Thesis. Oxford, UK: Oxford University.
- Rouse, G.W., Fauchald, K., 1997. Cladistics and polychaetes. *Zool. Scri.* 26, 139–204.
- Rouse, G.W., Pleijel, F., 2001. Polychaetes. Oxford: Oxford University Press. Pp. 235–250.
- Rousset, V., Rouse, G.W., Féral, J.P., Desbruyères, D., Pleijel, F., 2003. Molecular and morphological evidence of Alvinellidae relationships (Terebelliformia, Polychaeta,

- Annelida). *Zool. Scri.* 32, 185–197.
- Rousset, V., Pleijel, F., Rouse, G.W., Erséus, C., Siddall, M. E., 2007. A molecular phylogeny of annelids. *Cladistics* 23, 41–63.
- Rot, C., Goldfarb, I., Ilan, M., Huchon, D., 2006. Putative cross-kingdom horizontal gene transfer in sponge (Porifera) mitochondria. *BMC Evol. Biol.* 6, 71.
- Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M-A., Barrell, B., 2000. Artemis: sequence visualisation and annotation. *Bioinformatics* 16 (10), 944–945.
- Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4 (4), 406-425.
- San Filippo, J., Lambowitz, A.M., 2002. Characterization of the C-terminal DNA-binding/DNA endonuclease region of a group II intron-encoded protein. *J. Mol. Biol.* 324, 933–951.
- Sneath, P.H.A., Sokal, R.R., 1973. *Numerical Taxonomy*. Freeman, San Francisco, CA.
- Stamatakis A., 2006. RaxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics.* 22 (21), 2688-2690.
- Struck, T.H., Schult, N., Kusen, T., Hickman, E., Bleidorn, C., McHugh, D., Halanych, K.M., 2007. Annelida phylogeny and the status of Sipuncula and Echiura. *BMC Evol. Biol.* 7, 57.
- Swofford, D.L., 2002. PAUP*. *Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Version 4. Sinauer Associates, Sinderland, Massachusetts.
- Swofford, D.L., Olsen, G.J., Waddell, P.J., Hillis, D.M., 1996. Phylogenetic inference. Pp. 407–514 in *Molecular systematics* (D. M. Hillis, C. Moritz, and B. K. Mable,

- eds.). Sinauer Associates, Inc., Publishers, Sunderland, Massachusetts.
- Toor N., Hausner G., Zimmerly S., 2001. Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. *RNA*. 7, 1142–1152.
- Vallès, Y., Boore, J.L., 2006. Lophotrochozoan mitochondrial genomes. *Int. Comp. Biol.* 46, 544–557.
- Vallès, Y., Halanych, K.M., Boore, J.L., 2008. Group II Introns Break New Boundaries: Presence in a Bilaterian's Genome. *PLoS ONE*. 3(1), e1488.
- Yang, Z., Rannala, B., 1997. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. *Mol. Biol. Evol.* 14, 717–724.
- Zhong, M., Struck, T.H., Halanych, K.M., 2008. Phylogenetic information from three mitochondrial genomes of Terebelliformia (Annelida) worms and duplication of the methionine tRNA. *Gene* 416, 11–21.
- Zimmerly, S., Hausner, G. Wu, X., 2001. Phylogenetic relationships among group II intron ORFs. *Nucleic Acids Res.* 29, 1238–1250.
- Zrzavy, J., Hypsa, V., Tietz, D., 2001. Myzostomida are not annelids: molecular and morphological support for a clade of animals with anterior sperm flagella. *Cladistics* 17, 1–29.

Table 1. Parameters in evolutionary models (Model abbreviations: JC, Jukes and Cantor (1969) model; K2P, Kimura (1980) two-parameter model; TrN, TrN model with equal base frequencies, Tamura and Nei 1993; SYM, Zharkikh 1994; F81, model of Felsenstein, Felsenstein 1981); HKY85, Hasegawa-Kishino-Yano model, Hasegawa et al., 1985; K3ST, Kimura 3 substitution type model, Kimura, 1981), GTR, General time-reversible model, Lanave et al., 1984; Rodrigues et al., 1990)

Model	Substitution Rate	Base frequencies
JC	$a=b=c=d=e=f$	equal
K2P	$a=c=d=f, b=e$	equal
TrN	$a=c=d=f, b, e$	equal
SYM	a, b, c, d, e, f	equal
K3ST	$a=c=d=f, b, e$	equal
F81	$a=b=c=d=e$	unequal
HKY85	$a=c=d=f, b=e$	unequal
GTR	a, b, c, d, e, f	unequal

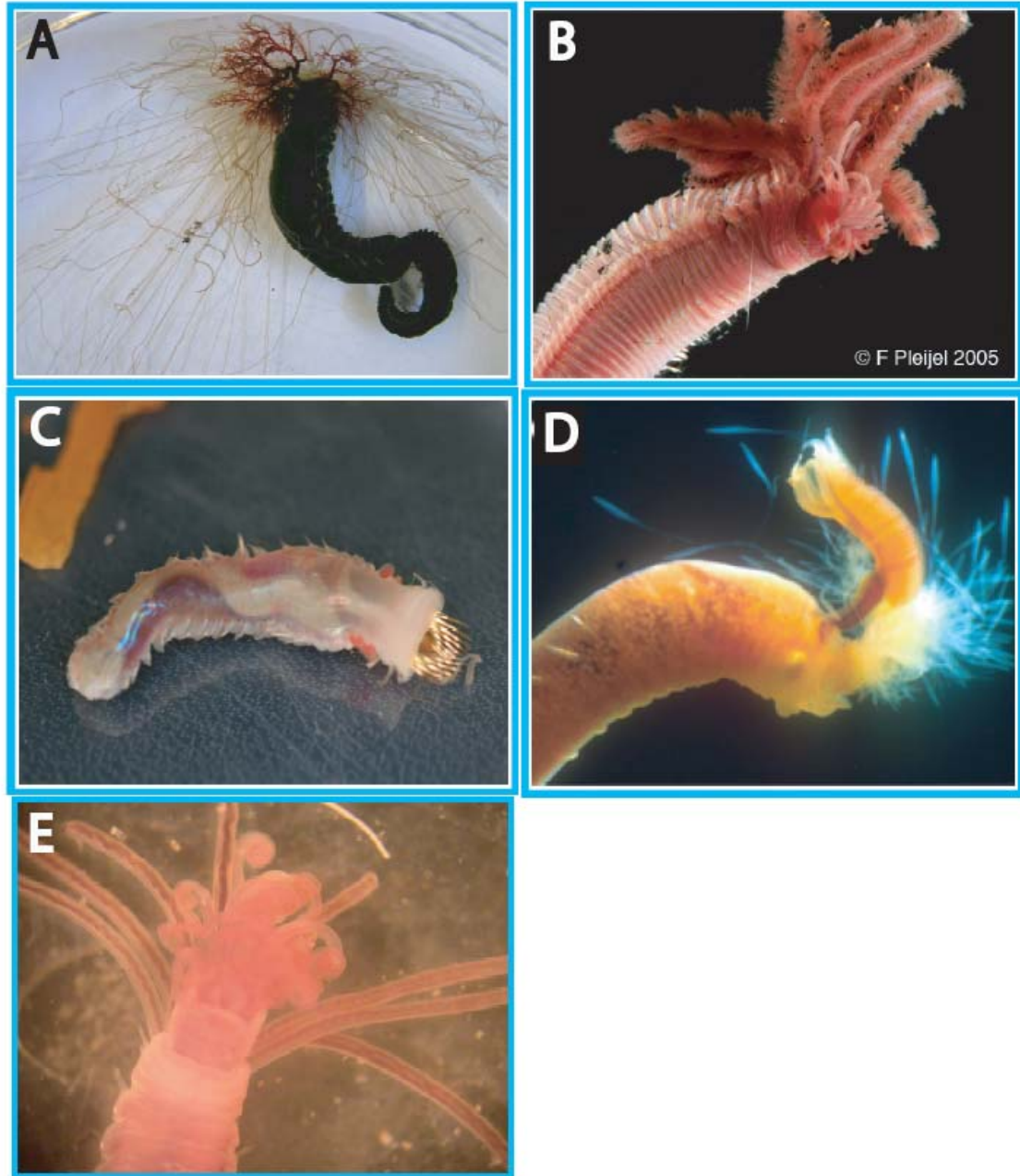


Figure 1. Five families of Terebelliformia worms. A. Terebellidea; B. Avinellidae (http://www.mbari.org/expeditions/ridges2005/august_11.htm); C. Pectinariidae; D. Trichobranchidae; E. Ampharetidae (Both D and E are from Rouse and Pleijel, 2001).

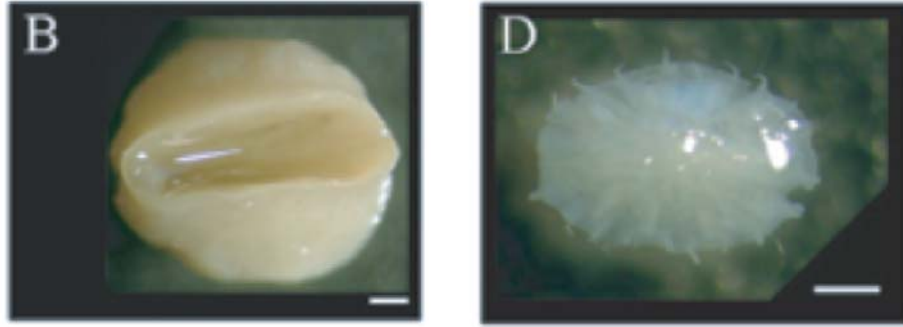


Figure 2. Microscopic views of *Endomyzostoma* sp. (Figures are from Lanterbecq et al., 2006) B: *Endomyzostoma deformatior*; D: *Endomyzostoma* sp..

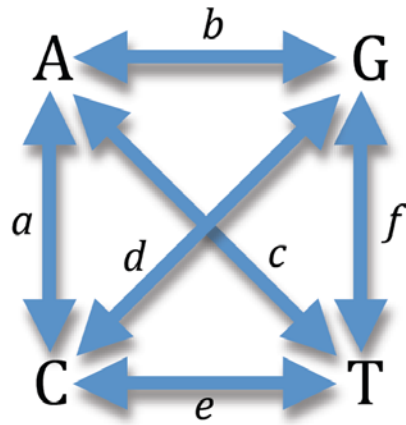


Figure 3. Substitution rates between all bases in evolutionary models.

CHAPTER 2

PHYLOGENETIC INFORMATION FROM THE THREE MITOCHONDRIAL GENOMES OF TEREPELLIFORMIA (ANNELIDA) WORMS AND DUPLICATION OF THE METHIONINE tRNA

Abbreviations

A - adenine; AIC - Akaike information criterion; *atp6* and *8* – ATP synthase subunits 6 and 8; bp – base pair; °C – degrees Celsius; C - cytosine; *cob* – cytochrome b apoenzyme; *coxI-3* – cytochrome c oxidase subunits 1-3; dNTP – deoxy- nucleotide triphosphate; G - guanine; kb – kilobases; ML – maximum likelihood; *mLSU* – mitochondrial large subunit ribosomal gene; *mSSU* – mitochondrial small subunit ribosomal gene; mtDNA - mitochondrial deoxyribonucleic acid; N - nucleotide; *nad1-6* and *4L* – NADH dehydrogenase subunits 1-6 and 4L; NADH – nicotinamide adenine dinucleotide; Nc – effective number of codons; PCR – polymerase chain reaction; RSCU - relative synonymous codon usage; T – thymine; TBR - tree-bisection-reconnection; tRNA – transfer ribonucleic acid; *trnL1* and *trnL2* – tRNA leucine 1 and 2; *trnM* - tRNA methionine; *trnR* – tRNA arginine; *trnS1* and *trnS2* – tRNA serine 1 and 2; *trnW* – tRNA tryptophane; UNK – unknown region.

Abstract

Mitochondrial genomes have been useful for inferring animal phylogeny across a wide range of clades, however they are still poorly sampled in some animal taxa, limiting our knowledge of mtDNA evolution. For example, despite being one of the most diverse animal phyla, only 5 complete annelid mitochondrial genomes have been published. To address this paucity of information, we obtained complete mitochondrial genomic sequences from *Pista cristata* (Terebellidae) and *Terebellides stroemi* (Trichobranchidae) as well as one nearly complete mitochondrial genome from *Eclysippe vanelli* (Ampharetidae). These taxa are within Terebelliformia (Annelida), which include spaghetti worms, icecream cone worms and their relatives. In contrast to the 37 genes found in most bilaterian metazoans, we recover 38 genes in the mitochondrial genomes of *T. stroemi* and *P. cristata* due to the presence of a second methionine tRNA (*trnM*). Interestingly, the two *trnMs* are located next to each other and are possibly a synapomorphy of these two taxa. The *E. vanelli* partial mitochondrial genome lacks this additional *trnM* at the same position, but it may be present in the region not sampled. Compared to other annelids, gene orders of these three mitochondrial genomes are generally conserved except for the *atp6-mSSU* region. Phylogenetic analyses reveal that mtDNA data strongly supports a Trichobranchidae/Terebellidae clade.

1. Introduction

Composition of the mtDNA genome is conserved across Bilateria containing 13 protein-coding genes, two ribosomal genes, 22 tRNA genes and one non-coding UNK region (unknown region), the presumed origin of replication. Generally, the size of mitochondrial genomes is about 15-17kb with homologous genes across Bilateria (Boore, 1999; Vallès and Boore, 2006). Whereas some taxa have highly rearranged mitochondrial genomes, gene order within annelids is hypothesized to be relatively conserved (Jennings and Halanych, 2005; Vallès and Boore, 2006). Despite the utility of mitochondrial genomes for phylogenetic reconstructions of major animal taxa, only five complete and five partial annelid mitochondrial genomes are available in GenBank as of October 2007. One reason for this bias in sampling is that in some taxa, including annelids, the UNK region (also called the D-loop or control region) has been difficult to amplify and sequence presumably because of secondary structure and/or microsatellite regions (Boore and Brown, 2000; Boore, 2001; Jennings and Halanych, 2005; Bleidorn et al., 2006a, b).

Annelida is a very diverse group of animals comprising around 16,500 described species including previously recognized ‘phyla’ (echiurids, sipunculids, and siboglinids – a.k.a. pogonophorans; McHugh, 1997; Boore and Staton, 2002; Bleidorn et al. 2006a; Struck et al., 2007). Even though relationships among annelid groups are not well understood, some higher-level annelid clades have been recovered with both morphology (Rouse and Fauchald, 1997) and molecular tools (Rousset et al., 2006; Struck et al.,

2007). One such group, Terebelliformia, comprises five ‘families’ (including Alvinellidae, Ampharetidae, Terebellidae, Trichobranchidae and Pectinariidae), which are typically sedentary benthic tube dwellers (Rouse and Pleijel, 2001). Phylogenetic position of Terebelliformia within annelids and relationships among those five ‘families’ remain debatable, but terebelliform monophyly has been recovered (Colgan et al., 2001; Struck et al., 2007) with Trichobranchidae treated as either a ‘subfamily’ of Terebellidae (e.g. Rouse and Pleijel, 2001), or a recognized ‘family’ (McHugh, 1995; Fauchald and Rouse, 1997; Glasby et al., 2004).

To further understand the phylogenetic utility of mitochondrial genome data within annelids, we examined mitochondrial genomes of three Terebelliformia worms (*Eclysippe vanelli*--Ampharetidae, *Pista cristata*--Terebellidae and *Terebellides stroemi*--Trichobranchidae). Gene composition and arrangement, tRNA folding patterns, and nucleotide patterns were examined. A phylogenetic assessment was performed to assess these Terebelliformia taxa relative to other annelid mitochondrial genome data available in GenBank.

2. Materials and Methods

2.1 Sample collection and DNA extraction

Pista cristata (Terebellidae) was collected in central Californian waters (36°23.0N, 121°57.9W) using the *R/V Pt. Sur*; *Terebellides stroemi* (Trichobranchidae) was obtained off southern New England (40° 46.1N, 71°156.2W) using the *R/V Oceanus*, and *Eclysippe vanelli* (Ampharetidae) was collected near Trosøm Norway (63°30.8N, 10°25.0E) using the *R/V Håkon Mosby*. All organisms were frozen at -80°C after

collection. Total genomic DNA extractions employed the DNeasy® Tissue Kit (Qiagen) according to the manufacture's instructions.

2.2 mtDNA data collection

Gene nomenclature and abbreviations in this paper follow Jennings and Halanych (2005).

Terebellides stroemi

The genome was amplified in three overlapping sections. First, conserved regions of *mLSU*, *cox1*, *cob* genes were amplified using taxonomically-inclusive primers (Table 1). After purification with QIAquick PCR purification kit (Qiagen), and sequencing of these products using a CEQ8000 (Beckmann), three pairs of species-specific primers for long PCRs were generated: Ter-16S-F/Ter-CO1-R, Ter-CO1-F/Ter-Cytb-R and Ter-Cytb-F/Ter-16S-R (Table 2). Long PCRs were performed on Eppendorf Mastercycler (Eppendorf) using Takara LA-Taq or the Eppendorf® TripleMaster PCR System. 50 µl long PCR reactions using Takara LA-Taq were set up as follows: 5µl 10× buffer, 8µl dNTP (2mM), 5µl MgCl₂ (25mM), 2µl of each long PCR specific primers (10µM each), 0.5µl Takara LA-Taq (5U/µl), 2µl DNA template and 25.5µl sterilized distilled water. The protocol for HotStart long PCRs was 94 °C for 3min; then polymerase was added by pipette (while samples were in the 94°C block) followed by 35 cycles with 94°C for 30 sec, 52°C for 30s, and 70°C for 12min; final extension at 72°C for 10min and hold at 4°C. Both *mLSU-cox1* and *cox1-cob* regions were 4-5 kb in size. For *cob-mLSU* fraction we used Eppendorf® TripleMaster PCR System with the 50µl reactions containing 5µl Tuning buffer with 25mM Mg²⁺, 12.5µl dNTP (2mM), 3µl of each long PCR specific

primers (10 μ M each), 0.4 μ l TripleMaster (5U/ μ l), 2 μ l DNA template and 24.1 μ l sterilized distilled water. The cycling profile was as follows: 93 °C for 3min; then 10 cycles with 93°C for 15sec, 50°C for 30sec, and 68°C for 4min 30sec; followed by 28 cycles with 93°C for 15sec, 50°C for 30sec, and 68°C for 4min 30sec (20sec increase each subsequent cycle); final extension at 68°C for 15min and hold at 4°C. The *cob-mLSU* fragment was approximately 7.5 kb in size.

All three long fragments were purified with QiaQuick Gel Extraction Kit (Qiagen) and cloned into the pGEM-T Easy vector (Promega). Clones were verified by PCR, grown overnight, and plasmids were isolated using the QIAprep[®] Spin Miniprep Kit (Qiagen). Purified plasmids were digested by *EcoRI* to check the insert size.

Using EZ::TN5[™] Insertion Kit (EPICENTRE), primer regions were inserted according to the manufacturer's instruction for subsequent sequencing. Unfortunately this kit was biased with many more transposons inserted (per kb) into the 3kb vector rather than the cloned mtDNA region, thus primer walking was employed to complete the fragments. Information for all sequencing primers can be found in Table 3.

Pista cristata and *Eclysippe vanelli*

Genomes for *P. cristata* and *E. vanelli* were amplified in four segments. Fragments of *mLSU*, *cox1*, *cob* and *nad4* genes were amplified by taxonomically inclusive primers (Table 1) and sequenced. For each species, four pairs of specific long-PCR primers (Table 2) were designed to amplify long fragments: *mLSU-cox1*, *cox1-cob*, *cob-nad4* and *nad4-mLSU*. Long PCRs employed Takara LA-Taq as described above. The protocols of long PCR and purification were similar to those of *T. stroemi* except for

annealing temperatures (see Table 2). All fragments were cloned into pGEM-T Easy vector (Promega) and clones screened as above. Fragments of *mLSU-cox1* and *cox1-cob* in both species were sequenced by EZ::TN5™ Insertion Kit (EPICENTRE) (with insertion bias again) and several internal primers. *cob-nad4* and *nad4-mLSU* fragments were sequenced by primer walking.

The *nad4-mLSU* fragment of *Echysippe*, apparently containing the UNK region, was difficult to amplify. Thus, new degenerate primers were designed to amplify regions of *nad5* and *mSSU* (Table 1) based on alignment of published annelid mitochondrial genomes. Subsequent long PCR primers (Nad5-Ev-698F/ Nad5-Ev-820R and 12S-Ev-longR) were produced. This approach allowed recovery of the *mSSU-mLSU* (primers 12S-Ann-256F/16SbrH) and the *nad4-nad5* region (primers Nad4-Ev-startF/ Nad5-Ev-820R). Resulting fragments were cloned, but the *nad5-mSSU* fragment could not be amplified despite several attempts. Based on the available information, *E. vanelli*'s missing region presumably includes part of *nad5* and *mSSU* genes, the UNK region, *trnK*, and *trnR*.

2.3 Genomic Assembly

Sequences were edited and aligned using DNASTAR™ Lasergene programs SeqMan and MegAlign (Burland, 2000). Protein-coding genes and ribosomal RNA genes were identified by Artemis (Rutherford et al., 2000) and Blast (Altschul et al., 1990). All tRNA genes were identified using tRNAscan-SE web server (<http://lowelab.ucsc.edu/tRNAscan-SE/>; Lowe and Eddy, 1997) under default settings and source = “mito/chloroplast”, or by eye based on their potential secondary structures and anticodon sequences. Boundaries of UNK regions in *P. cristata* and *T. stroemi* were

inferred by identifying flanking tRNA sequence. Secondary structure of UNK region was examined with the “mfold” online server (Zuker et al., 1999) and DNASTAR™ Lasergene program GENEQUEST (Burland, 2000).

2.4 Phylogenetic analysis

Analyses included all available annelid mitochondrial genomes with about 50% coverage or greater (Table 4). The alignment of Jennings and Halanych (2005) was employed, to which *Orbinia latreillii*, *Scoloplos cf. armiger*, *Urechis caupo*, and the three terebelliform mitochondrial genomes were added (Table 4). Because we are interested in relationships within annelids and given results of recent mitochondrial genome analyses (Jennings and Halanych, 2005; Bleidorn et al., 2006a), a representative brachiopod and a mollusk were used as outgroups. Two datasets were created for phylogenetic analyses. In the nucleotide dataset, the 2 rRNA genes and all protein-coding genes (except for *atp6*, *atp8* and *nad6* genes which exhibit high variability) were included. We used Clustal X (Thompson et al., 1997) under default setting to realign rRNA genes. MacClade4.08 (Maddison and Maddison, 2002) was used to exclude most regions that contain insertions/deletions and all third codon positions in protein coding genes. Gblocks 0.91b (Castresana, 2000) was used to identify ambiguous aligned regions in rRNA genes (*mLSU* and *mSSU*) that were excluded from analyses. The amino acid dataset was created from the translated (invertebrate mitochondrial code) aligned nucleotide dataset with exclusion of rRNA genes by MacClade4.08 and Se-AI v2.0a11 (Rambaut, 1996). All alignments are available at TreeBase (accession no SN3857 and SN3858).

For the nucleotide dataset, phylogenetic analyses employed both maximum likelihood and Bayesian-inference approaches. Maximum-likelihood analyses were

performed in PAUP4.0b10 (Swofford et al., 2002) with a GTR+ Γ +I model as determined by MODELTEST v3.7 based on the Akaike information criterion (AIC) (Posada and Crandall, 1998). Heuristic searches were run with random-taxon addition (10 replicates) using Tree-Bisection-Reconnection (TBR) swapping. All model parameters used fixed values as determined by MODELTEST v3.7. Bootstrap analysis employed 1,000 iterations using heuristic searches with 10 random taxa addition. Bayesian inference analyses in MRBAYES version 3.1.2 (Huelsenbeck and Ronquist, 2001) used 5,000,000 generations with 2 runs of chains (3 heated and 1 cold), and sampling every 100 generations. Both non-partitioned and partitioned Bayesian analyses were employed. For the non-partitioned analyses, we use prior distributions according to the GTR+ Γ +I model selected under the AIC in MrModeltest (Nylander, 2004). The partitioned analysis used unlinked GTR+ Γ +I models for which parameters were individually estimated for each gene partition (note *nad3* and *nad4* partitions used no proportion of invariant sites resulting in a GTR+ Γ model). Resulting -ln likelihood scores were graphed using X-Y scatter plots to identify the “burn-in” point at which all estimated parameters reached stationarity (burnin = 2,500).

For the amino acid dataset, a non-partitioned ML analyses was run in addition to non-partitioned and partitioned Bayesian analyses. For ML analyses, model selection was performed in ProtTest (Abascal et al., 2005) and MtArt + Γ +I+F model was chosen as the best one under the AIC. As there is no MtArt model used in RAxML, we chose the next best model, MtREV+ Γ +I+F, available in RAxML. A maximum likelihood search was implemented by 200 bootstrap replicates using RAxML 7.0.0 (Stamatakis, 2006) with MtREV+ Γ +I+F model by the “PROTGAMMAIMTREV” option. The non-partitioned

Bayesian analysis was conducted with the mixed amino acid substitution model option plus a Γ distribution and a proportion of invariant sites in MRBAYES v3.1.2 with 5,000,000 generations sampled every 100 generations (burnin = 2,500). In the mixed model option, a specific model is not specified a priori, but each model is chosen during the runs based on its posterior probability. We also employed a partitioned amino acid analyses in MrBayes in which, the mixed amino acid substitution model option plus a Γ distribution and a proportion of invariant sites was assigned to each partition individually and unlinked during the run. All other settings remained the same (burnin = 2,500).

3. Results and Discussion

3.1 Genome Composition

The complete mtDNA of *T. stroemi* and *P. cristata* are 15,755 bp and 15,894 bp in length respectively. Both contain 38 genes including 13 protein-coding genes, two rRNA genes and 23 tRNAs. Both genomes contain 2 copies of the methionine tRNA gene. The partially sequenced mitochondrial genome of *E. vanelli* is 13,749 bp in length. However, regions of *nad5* and *mSSU* genes, the UNK region and 2 tRNA genes (*trnK* and *trnR*) were not recovered for *E. vanelli*. Fig. 1 shows gene orders of these taxa as well as other annelids taxa available in GenBank. As in other annelids (Boore and Brown, 2000; Boore and Staton, 2002; Jennings and Halanych, 2005; Bleidorn et al., 2006a, b;), all Terebelliformia mitochondrial genes are transcribed from the same strand. *Terebellides stroemi* and *P. cristata* have identical gene arrangements, and are unique in containing two methionine tRNAs in tandem. *Eclysippe vanelli* has a different gene order and presumably, like other bilaterians, only one *trnM*.

All three genomes are AT-rich (>66%) with T being the most common base (Table 5). The low percentage of G in the third codon positions is also notable in both *P. cristata* (3.1%) and *T. stroemi* (4.5%), but in *E. vanelli* percentage of C is the lowest (see Table 5). Skewness, a measure of the bias of the base composition, is calculated as $(A-T)/(A+T)$ and $(G-C)/(G+C)$ for an individual strand (Perna and Kocher, 1995). Table 5 lists the AT-skew and GC-skew in both protein-coding genes and whole genome sequences for the terebelliform genomes. Whereas both *T. stroemi* and *P. cristata* exhibit GC-skews more negative than AT-skews, *E. vanelli* has positive GC-skews. Negative values in skewness mean the coding strand has more Ts or Cs, respectively. In contrast, positive values indicate more As or Gs.

3.2 Protein-coding genes

Terebelliform genomes contain all 13 protein-coding genes typically found in metazoan mtDNA (Boore, 1999). Effective number of codons used in a gene (N_c) can be used to quantify departure from equal codon usage of synonymous codons, that is codon usage bias (Wright, 1990) (Table 6). N_c is reported as values from 20 to 61, where 20 represents extreme bias of one codon exclusively used for each amino acid and 61 represents equal probability of alternative synonymous codons. N_c values, which are similar in all three terebelliform genomes (42.5 for *T. stroemi*, 40.1 for *P. cristata* and 40.8 for *E. vanelli*), show greater codon usage bias than other annelid mitochondrial genomes (Bleidorn et al., 2006a). In all terebelliforms, third position As or Ts (NNA or NNU) are more used than third position Gs or Cs (NNG or NNC), respectively, for 2-fold degenerated codons (Table 6). For *T. stroemi* and *P. cristata*, the difference between the usage of NNA and NNG is more pronounced than for NNU versus NNC, but in *E. vanelli*

the situation is reversed. For 4-fold degenerated codons, NNC is more frequent than NNG in *T. stroemi* and *P. cristata*, but both are less common than NNU and NNA with the latter being used more often (Table 6). In *E. vanelli* however, NNG is more common than NNC and NNU than NNA. Thus, the bias in codon usage reflects the AT richness as well as the distribution of base frequencies at the third positions except for *P. cristata*, which shows a higher frequency of T than A confirming that codon usage bias and base frequency bias are tightly linked. Furthermore, the average relative synonymous codon usage (RSCU) of the tRNA's found in the mitochondrial genomes of either *T. stroemi* or *P. cristata* is 1.41 or 1.31, respectively. In *E. vanelli* on the other hand, the RSCU for these tRNA's is only 0.85. Correspondingly, the patterns in the 4-fold and 2-fold degenerated codons are the opposite to the ones in *T. stroemi* and *P. cristata* indicating a possible relative preference of nuclear tRNA's in *E. vanelli*.

Other codon features include ATG, the start codon in all protein-coding genes with the exception of *cox1* in *E. vanelli* which is initiated by the codon ATA. Our observations are consistent with previous reports in annelids where ATG is the most common initiation codon with the occasional use of ATA, ATC, GTG, GCC and GTT in some genes (Boore and Brown, 2000; Jennings and Halanych, 2005; Bleidorn et al., 2006a).

An incomplete termination codon, a single T, is used for many NADH genes (*nad4* and *nad1* genes in *T. stroemi*; *nad5*, *nad1* and *atp8* genes in *P. cristata*; *cox1*, *cob*, *nad5*, *nad4*, *nad3* and *nad2* genes in *E. vanelli*). This incomplete codon is presumably completed to a TAA stop codon via a polyadenylation process (Ojala et al., 1981). All other protein-coding genes in *T. stroemi* and *P. cristata* contain the complete TAA

termination codon. In *E. vanelli*, TAG is used as a termination codon in three genes (*cox3*, *cob* and *atp6*). The *cob* gene in *E. vanelli* is ~20 bp shorter than in other annelids, yielding 19 intergenic nucleotides between *cob* and the adjacent *trnW* gene.

Overlapping genes have been reported in compact genomes (Kurabayashi and Ueshima, 2000; Firth and Brown, 2006). In addition to the one overlap mentioned above, *nad4L* (which has a complete termination codon TAA) has a 7 bp overlap with the downstream *nad4* genes in all three genomes.

3.3 tRNAs

Secondary structures of all tRNAs are depicted in Fig. 2. Both *T. stroemi* and *P. cristata* have 23 tRNA genes respectively in their mitochondrial genomes. Only 20 tRNA genes were found in the partial *E. vanelli* mtDNA.

We found two *trnMs* adjacent to each other in mtDNAs of both *T. stroemi* and *P. cristata*. Unlike 2 *trnLs* and 2 *trnSs*, the *trnMs* have the same anticodon triplet (AUG) coding for methionine. For the sake of clarity, we have called the 5' upstream copy of the translated strand *trnM1* and the 3' downstream copy *trnM2*. The TΨC stems of both *trnM2s* with only 2 or 3 matching bases are much shorter than the ones of *trnM1*. The shorter TΨC arm occurs in other Annelida taxa, including *Clymenella torquata*, *Eclysippe vanelli*, *Platynereis dumerilii*, *Lumbricus terrestris*, *Riftia pachyptila* and *Urechis caupo* (Boore and Brown, 1995, 2000; Boore et al., 1999; Jennings and Halanych, 2005). Additionally, both *trnM2s* possess identical 4 base-pair DHU stems, same as *trnM1* of *P. cristata*, which is conserved in *trnM* across the sampled Annelida taxa and found in other bilaterians (e.g., the mollusk *Katharina tunicata*, and the cephalochordate *Branchiostoma*

floridae; Boore and Brown, 1994, 1995, 2000; Boore et al., 1999; Boore and Staton, 2002; Jennings and Halanych, 2005). In terms of both nucleotides and structure, *trnM2* is much more similar to other annelid *trnMs* including *E. vanelli*.

In *T. stroemi*, *trnR* lack potential for folding a DHU stem (D arm) which is thought to act as a recognition site for aminoacyl-tRNA synthetase (Clark, 2006). The lack of a DHU stem has been reported in several known mtDNAs of annelids taxa (Boore, 2001; Jennings and Halanych, 2005).

3.4 UNK region

The UNK region of *T. stroemi*, 802 bp, and *P. cristata*, 844 bp, is located between *trnR* and *trnH*. Both of them are AT-rich (*T.stroemi*: 72.5%; *P. cristata*: 78.4%) and possess microsatellite-like sequences. About 50 uncontinuous TATA-repeats mainly occur within 200 bp from *trnR* in *T. stroemi* and 350 bp in *P. cristata*. We have examined this region for secondary structure using both the “mfold” online server (Zuker et al., 1999) and DNASTAR™ Lasergene program GENEQUEST (Burland, 2000). Both approaches reported several hairpin structures within the UNK region. However, results between the two programs differed in size and location of secondary structures. Thus, secondary structure, in addition to nucleotide repeats, likely hindered our ability to recover the *E. vanelli* UNK region.

3.5 Structure and Gene order

Gene orders of *P. cristata* and *T. stroemi* are identical to those of Clitellata, Siboglinidae and Maldinidae, except for the second *trnM* reported above and the position of *trnK* in maldanids (Fig. 1). Therefore, this gene order is most likely symplesiomorphic

for annelids. When genomes are aligned, a few clusters with conserved gene boundaries can be identified (see Fig. 1). The observation of conserved gene order for several protein-coding genes in association with rearrangements for tRNA genes and UNK region across Annelida worms has been previously noted (Jennings and Halanych, 2005; Vallès and Boore, 2006; Bleidorn et al., 2006a). In contrast, *E. vanelli* is expected to show variation in the arrangement of several tRNA genes and the UNK region compared to these taxa (Fig. 1). Furthermore, *nad4L* and *nad4* are placed before *nad5* in *E. vanelli*, suggesting the *atp6-mSSU* region across annelids is more variable than other regions. In general, gene order is conserved across sampled annelids with the exception of *Urechis caupo* (the echiurid) and *Phascolopsis gouldii* (the sipunculid). Whereas the first part from *cox1* to *cob* of the incomplete sipunculid genome is conserved, the remaining gene order shows rearrangements compared to other annelids (Vallès and Boore, 2006).

3.6 Phylogenetic Analyses

Maximum likelihood and Bayesian inference (both non-partitioned and partitioned) of nucleotide data yielded identical topologies which are slightly different from that of the non-partitioned Bayesian inference and RAxML amino acid analyses (Fig. 3). The partitioned Bayesian amino acid analysis was poorly resolved (Fig. 4) and resulted in topology that was consistent with nucleotide trees. The siboglinids, echiurid and sipunculid fell inside Annelida with significant posterior probabilities in nucleotide and amino acid data. This finding is consistent with previous results from nuclear markers (e.g., McHugh, 1997; Bleidorn et al., 2003), mitochondrial DNA (Bleidorn et al., 2006a) and combined molecular data (Struck et al., 2007). Placement of these taxa within Annelida confirms that segmentation has been highly modified or lost within annelid

lineages (Halanych et al., 2002; Struck et al., 2007).

The close relationship between the echiurid and the aciculate *Platynereis*, found in the nucleotide trees here and in Bleidorn et al.'s (2006a) mtDNA analyses, has not been recovered in ML amino acid tree and combined nuclear gene trees (Rousset et al., 2006; Struck et al., 2007), which placed echiurans as sister to Capitellidae. The sister relationship between Siboglinidae and Clitellata in both nucleotide and amino acid trees agrees with mitochondrial data analyses (Jennings and Halanych, 2005), but contrasts with other molecular markers (Rousset et al. 2003, Struck et al., 2007). Comparing three trees in figure 3, the positions of Nereididae, Echiura, Sipuncula and Orbiniidae are quite different and with low nodal support indicating that their relationships are not fully resolved. Orbiniidae is placed at the base of the MrBayes amino acid tree (Fig 3B) disagreeing with suggestions to place Orbiniidae closer to Nereididae (Struck et al., 2007). In both nucleotide trees (Fig 3A) and ML amino acid tree (Fig 3C), the sipunculid *Phascolopsis* is close to orbiniids (88% of ML, 100% of Bayesian; 90% of amino acid ML), which is consistent with mitochondrial genomic data (Bleidorn et al., 2006a).

The three terebelliform taxa cluster with *Clymenella torquata* (Maldanidae, Scolecida) (69% of nucleotide ML bootstrap, 100% of posterior probability for both nucleotide and the non-partitioned amino acid tree, and 91% in the amino acid ML bootstrap), albeit given that a more comprehensive taxon sampling is necessary to delineate the sister group of Terebelliformia. Within Terebelliformia, a Trichobranchidae/Terebellidae clade is strongly supported by the mitochondrial genomic data as hypothesized based on morphology (Rouse and Fauchald, 1997). The unique presence of two methionine-tRNAs appears to be an additional synapomorphy. In

contrast, previous morphological and molecular analyses (Rousset et al., 2003; Glasby et al., 2004; Struck et al., 2007) place other taxa (e.g. alvinellids) between terebellids and trichobranchids. However, in general the taxon sampling of mitochondrial genomes is still limited, and more taxa, especially from Terebellidae (see Colgan et al., 2001), need to be sampled to confirm such results.

Acknowledgements

We are grateful to Les Goertzen and Scott Santos for their helpful comments on this manuscript. This work was supported by National Science Foundation (NSF) WormNet grant (EAR-0120646) to K. M. H., and in part, by the Alabama Commission on Higher Education Graduate Research Scholar's Program through the Auburn University Cellular and Molecular Biosciences Program and NSF EPS-0447675. Contribution #35 to the AU Marine Biology Program.

References

- Abascal, F., Zardoya, R., Posada, D. 2005. ProtTest: Selection of best-fit models of protein evolution. *Bioinformatics*. 21(9), 2104-2105.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.M., Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.
- Bleidorn, C., Vogt, L., Bartolomaeus, T. 2003. New insights into polychaete phylogeny (Annelida) inferred from 18S rDNA sequences. *Mol. Phylogenet. Evol.* 29, 279–288.

- Bleidorn, C., Podsiadlowski, L., Bartolomaeus, T. 2006a. The complete mitochondrial genome of the orbiniid polychaete *Orbinia latreillii* (Annelida, Orbiniidae)--A novel gene order for Annelida and implications for annelid phylogeny. *Gene* 370, 96–103.
- Bleidorn, C., Kruse, I., Albrecht, S., Bartolomaeus, T. 2006b. Mitochondrial sequence data expose the putative cosmopolitan polychaete *Scoloplos armiger* (Annelida, Orbiniidae) as a species complex. *BMC Evol. Biol.* 6, 47.
- Boore, J.L. 1999. Animal mitochondrial genomes. *Nucleic Acids Res.* 27, 1767–1780.
- Boore, J.L. 2001. Complete mitochondrial genome sequence of the polychaete annelid *Platynereis dumerilii*. *Mol. Biol. Evol.* 18, 1412-1416,
- Boore, J.L., Brown, W.M. 1994. Complete DNA sequence of the mitochondrial genome of the black chiton, *Katharina tunicata*. *Genetics* 138, 423–443.
- Boore, J.L., Brown, W.M. 1995. Complete sequence of the mitochondrial DNA of the annelid worm *Lumbricus terrestris*. *Genetics* 141, 305-319.
- Boore, J.L., Brown, W.M. 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., Daehler, L.L., Brown, W.M. 1999. Complete sequence, gene arrangement, and genetic code of mitochondrial DNA of the Cephalochordate *Branchiostoma floridae* (Amphioxus). *Mol. Biol. Evol.* 16, 410–418.
- Boore, J.L., Staton, J.L. 2002. The mitochondrial genome of the Sipunculid *Phascolopsis*

- gouldii* supports its association with Annelida rather than Mollusca. *Mol. Biol. Evol.* 19, 127–137.
- Burland, T.G. 2000. DNASTAR's lasergene sequence analysis software. *Methods Mol. Biol.* 132, 71–91.
- Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17, 540–552.
- Clark, B.F.C. 2006. The crystal structure of tRNA. *J. Biosciences.* 31, 453–457.
- Colgan, D.J., Hutchings, P.A. Brown, S. 2001. Phylogenetic relationships within the Terebellomorpha. *J. Mar. Biol. Ass. U.K.* 81, 3806/1-9.
- Fauchald, K., Rouse, G.W. 1997. Polychaete systematics: Past and present. *Zool. Scr.* 26, 71–138.
- Firth, A.E., Brown, C.M. 2006. Detecting overlapping coding sequences in virus genomes. *BMC Bioinform.* 16, 75–80.
- Folmer, O., Black, M., Hoeh, W., Lutz, R., Vrijenhoek, R. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotech.* 3, 294-299.
- Glasby, C.J., Hutchings, P.A., Hall, K. 2004. Assessment of monophyly and taxon affinities within the polychaete clade Terebelliformia (Terebellida). *J. Mar. Biol. Ass. U.K.* 84, 961–971.
- Halanych, K.M., Dahlgren, T. G., McHugh, D. 2002. Unsegmented Annelids? Possible origins of four Lophotrochozoan worm taxa. *Int. Comp. Biol.* 42, 678–684.
- Huelsenbeck, J.P., Ronquist F. 2001. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* 17, 754–755.

- Jennings, R.M., Halanych, K.M. 2005. Mitochondrial genomes of *Clymenella torquata* (Maldanidae) and *Rifta pachyprila* (Siboglinidae): Evidence for conserved gene order in Annelida. *Mol. Biol. Evol.* 22, 210–222.
- Kurabayashi, A., Ueshima, R. 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.
- Lowe, T.M., Eddy, S.R. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–964.
- Maddison, D.R., Maddison, W.P. 2002. MacClade 4: Analysis of Phylogeny and Character Evolution, version 4.0. Sunderland, MA: Sinauer Associates.
- McHugh, D. 1995. Phylogenetic analysis of the Amphitritinae (Polychaeta: Terebellidae). *Zool. J. Linn. Soc.* 114, 405–429.
- McHugh, D. 1997. Molecular evidence that echiurans and pogonophorans are derived annelids. *Proc. Natl. Acad. Sci. U S A.* 94, 8006–8009.
- Nylander, J.A.A. 2004. MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre, Uppsala University.
- Ojala, D., Montoya, J., Attardi, G. 1981. tRNA punctuation model of RNA processing in human mitochondria. *Nature* 290, 470–474.
- Palumbi, S.R., 1996. Nucleic acid II: the polymerase chain reaction. Pp. 205–247 in *Molecular Systematics*, Hillis, D.M., Moritz, C., Mable, B.K. eds. Sinauer Associates, Sunderland, MA.
- Perna, N.T., Kocher, T.D. 1995. Patterns of nucleotide composition at fourfold

- degenerate sites of animal mitochondrial genomes. *J. Mol. Evol.* 41, 535-358.
- Posada, D., Crandall, K.A. 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics* 14, 817–818.
- Rambaut, A. 1996. The Use of Temporally Sampled DNA Sequences in Phylogenetic Analysis. PhD Thesis. Oxford, UK: Oxford University.
- Rouse, G.W., Fauchald, K. 1997. Cladistics and polychaetes. *Zool. Scri.* 26, 139–204.
- Rouse, G.W., Pleijel, F. 2001. Polychaetes. Oxford: Oxford University Press. Pp. 235–250.
- Rousset, V., Rouse, G.W., Féral, J.P., Desbruyères, D., Pleijel, F. 2003. Molecular and morphological evidence of Alvinellidae relationships (Terebelliformia, Polychaeta, Annelida). *Zool. Scri.* 32, 185–197.
- Rousset, V., Pleijel, F., Rouse, G.W., Erséus, C., Siddall, M. E. 2006. A molecular phylogeny of annelids. *Cladistics* 22, 1–23.
- Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M-A., Barrell, B. 2000. Artemis: sequence visualisation and annotation. *Bioinformatics* 16 (10), 944–945.
- Stamatakis A. 2006. RaxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics.* 22 (21), 2688-2690.
- Struck, T.H. Schult, N., Kusen, T., Hickman, E., Bleidorn, C., McHugh, D., Halanych, K.M. 2007. Annelida phylogeny and the origins of Sipuncula and Echiura. *BMC Evol. Biol.* 7, 57.
- Swofford, D.L. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other

Methods). Version 4. Sinauer Associates, Sinderland, Massachusetts.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins. D.G. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 24, 4876–4882.

Vallès, Y., Boore, J.L. 2006. Lophotrochozoan mitochondrial genomes. *Int. Comp. Biol.* 46, 544–557.

Wright, F. 1990. The effective number of codons used in a gene. *Gene* 87, 23–29.

Zuker, M., Mathews, D.H., Turner, D.H. 1999. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide in RNA biochemistry and biotechnology. Barciszewski, J. and Clark, B.F.C. eds., NATO ASI series, Kluwer Academic Publishers.

Table 1. Taxonomically-inclusive PCR primers to amplify small conserved regions.

Primer name	Sequence	Reference
<i>mSSU</i>		
12S-Ann-256F	5'—AWW TYY GTG CCA GCW RCC GC—3'	Reported here
12S-Ann-800R	5'—TCW TGT TAC GAC TTA YCT C—3'	Reported here
<i>mLSU</i>		
16SarL	5'—CGC CTG TTT ATC AAA AAC AT—3'	Palumbi (1996)
16SbrH	5'—CCG GTC TGA ACT CAG ATC ACG T—3'	Palumbi (1996)
<i>cob</i>		
Cytb 424F	5'—GGW TAY GTW YTW CCW TGR GGW CAR AT—3'	Boore and Brown (2000)
Cytb 876R	5'—GCR TAW GCR AAW ARR AAR TAY CAY TCW GG—3'	Boore and Brown (2000)
<i>cox1</i>		
LCO1490	5'—GGT CAA CAA ATC ATA AAG ATA TTG G—3'	Folmer et al. (1994)
HCO2198	5'—TAA ACT TCA GGG TGA CCA AAA AAT CA—3'	Folmer et al. (1994)
<i>nad4</i>		
Nad4f	' —TGR GGN TAT CAR CCN GAR CG—3'	Jennings and Halanych (2005)
Nad4r	5' —GCY TCN ACR TGN GCY TTN GG—3'	Jennings and Halanych (2005)
<i>nad5</i>		
Nad5-Ann-482F	5'—ACN AAY CGW ATY GGR GA—3'	Reported here
Nad5-Ann-937R	5'—GCY TTA AAT ADH GCR TGD GT—3'	Reported here

Table 2. Primers for long PCR with annealing temperatures.

Fragments	Primer name	Sequence	Annealing Temp.
<i>T. stroemi</i>			
<i>mLSU-cox1</i>	Ter-16S-F	5'—TGG GCT TGT ATG AAC GGA TAA ACG AAG GC—3'	52
	Ter-CO1-R	5'—AGC TAA GTG AAG TGA GAA AAT AGC AAG GTC—3'	52
<i>cox1-cob</i>	Ter-CO1-F	5'—TAT TGG AGG TTT CGG TAA TTG ATT AAT CCC-3'	52
	Ter-Cytb-R	5'—CAG GTT TGA TAT GGA TTG GTG TGA CTA ATG—3'	52
<i>cob-mLSU</i>	Ter-Cytb-F	5'—AAT AAT CCT CTT GGA ATT AGT AGT ACA TCC—3'	50
	Ter-16S-R	5'—TTT GTG GAG GGT ATT TAC TCC ATA GCC G—3'	50
<i>P. cristata</i>			
<i>mLSU-cox1</i>	Pis-16S-F	5'—CCT GAC CGT GCT AAG GTA GCG TGA TAA TTC—3'	52
	Pis-CO1-R	5'—CGT AGT CCT TTT CAT CGT ATA TTA GCT ACG G—3'	52
<i>cox1-cob</i>	Pis-CO1-F	5'—GCC TTA CTT CTT CTA CTC AGT TCA GCT G—3'	52
	Pis-Cytb-R	5'—CTA GAA TAT TAG GTC AAA ATA TAA CAA CGG—3'	52
<i>cob-nad4</i>	Pis-Cytb-F	5'—AGG AGT TAG GTC AAA TAC AGA CAA AAT CCC—3'	55
	Nad4-Pc-290R	5'—GAG AAG CAA ATC GTA ATA AAC CAT AGC TGC—3'	55
<i>nad4-mLSU</i>	Nad4-Pc-startF	5'—TCT GGC TAA CCA AAG CCC ATG TCG AAG C—3'	50
	Pis-16S-R	5'—TAG ATT CAG GTC TCT TCA CCT GAA GGC C—3'	50
<i>E. vanelli</i>			
<i>mLSU-cox1</i>	Ecl-16S-F	5'—CTG ACT GTG CTA AGG TAG CGT GAT AAT TCG—3'	52
	Ecl-CO1-R	5'—CCG TAA CAT GGA TTG ACC AAA CAA AAA GCC—3'	52
<i>cox1-cob</i>	Ecl-CO1-F	5'—TTG TAT GAA TAC AAT TGT GAC AGC TCA TGC—3'	52
	Ecl-Cytb-R	5'—ATC AAT TCT CTG GAT CCC CTA ACA CAA CAG—3'	52
<i>cob-nad4</i>	Ecl-Cytb-F	5'—ACA TCA AAC TGG GTC TAG TAA TCC AAT TGG—3'	52
	Nad4-Ev-startR	5'—CCC AAC TAA CAA AGG CAA AGA TGC ACT TGA—3'	52
<i>nad4-nad5</i>	Nad4-Ev-startF	5'—TCA AGT GCA TCT TTG CCT TTG TTA GTT GGG—3'	52
	Nad5-Ev-820R	5'—ATC ACC CCA AGT TGA CTC AAA GTT GAT AAA G—3'	52
<i>mSSU-mLSU</i>	12S-Ann-256F	5'—AWW TYY GTG CCA GCW RCC GC—3'	Degenerate 45-60
	16SbrH	5'—CCG GTC TGA ACT CAG ATC ACG T—3'	Degenerate 45-60

Table 3. Sequencing Primers for sequencing the mitochondrial genomes of *Terebellides stroemi*, *Pista cristata* and *Eclysippe vanelli*.

Fragments	Primer name	Sequence
<i>T. stroemi</i>		
<i>mLSU-cox1</i>	Nad1-Ts-753F	5'—TAA CAT CCT AGT TAT AAG CC—3'
	Nad3-Ts-endR	5'—TAG CTG TCA CCA GAG TTC CC—3'
	Nad2-Ts132F	5'—TCA TTT ATC CCT TTA CTA GC—3'
	Nad2-Ts786R	5'—TTA CTA TAG CAG AAA TGA C—3'
<i>cox1-cob</i>	CO1-Ts1500R	5'—TAT TAT GGA AGT CTA AGG G—3'
	Cytb-Ts550R	5'—ATG TAG AC AAA GAA ACG G—3'
	COIII-Ts653F	5'—TAT ACG AAC TAT GTC ACA CC—3'
	Cytb-Ts27R	5'—ATT GAT GAT TTT CAT TAG TGG—3'
<i>cob-mLSU</i>	tRNAArg-TsF	5'—CGG CTC AAA CTT CAA GGG TG —3'
	Nad5-Ts-startR	5'—CAA CAG ATG AGA ATA GTA TCC C—3'
	12S-Ts-startF	5'—TCA CAG CCT ATG TAT TGC CG—3'
	12S-Ts-midR	5'—AAC GCG GAA GGT ATG TAG CC—3'
<i>P. cristata</i>		
<i>mLSU-cox1</i>	16S-Pc400F	5'—AGC TAC CTC GGG GAT AAC AG —3'
	16S-Pc-270F	5'—ATA AAG ATG GCA GGT GTG AAG C—3'
	tRNA-Pc-befNad1R	5'—AAA CAA TAA TCT GTC AAC C—3'
	Nad3-Pc207F	5'—GAA ATT GCT CTC CTA ATA CC—3'
	Nad2-Pc31R	5'—ATA ATA CCA GAT AGT AAG G—3'
	Nad2-Pc307F	5'—CAA CAG TTA TAT CTG CTC—3'
<i>cox1-cob</i>	COIII-Pc407F	5'—TTT AGC TTC AGG CGT TAC CG —3'
	COIII-Pc-654R	5'—CGA AGC ATG GAA TTG TGT AG—3'
	Nad6-Pc106F	5'—CAA CAG CAG GAT TTA TTG GG—3'
	Cytb-Pc357R	5'—CCT CCT AAA CGT CAA CTA T—3'
<i>cob-nad5</i>	Cytb-Pc-900F	5'—TTG CAA TAT TTG CTG CTA TCC —3'
	Nad5-Pc-405R	5'—TAC CTG CAC CAA TGT ATT TAG CG —3'
	Atp6-Pc-172F	5'—TAT ACA TAG TCA AGT ATC GCG—3'
	Nad5-Pc-startR	5'—AAA TCT GAA ACC ACA ACT CAG—3'
	Nad5-Pc-startF	5'—AAT ACT GAG TTG TGG TTT CAG—3'

Fragments	Primer name	Sequence
	Atp6-Pc-690F	5'—TCT ATG CAG ATG ACC ATA CTC—3'
	Atp6-Pc-5F	5'—TCT AAA TTC TGA TGC TGA CTG—3'
	UNK-Pc-endR	5'—GGT AAA CCC TAT TAA CAA GGT G—3'
	Nad5-Pc-350R	5'—CTG CAC CAA GTG ATT TAG CG—3'
<i>nad5-nad4</i>	Nad5-Pc-300R	5'—ACA AAG TAT TAT ATC TTG C —3'
	Nad5-Pc-1616F	5'—CAG AAA TAA GAT CTG CTG AAG G—3'
	Nad4-Pc-900F	5'—ATA TCA TTA TGT GCT GCA CC —3'
	Nad4L-Pc-405R	5'—GAT ATA AGT AGA TGT TTT CGT TG—3'
<i>nad4-mLSU</i>	Nad4-Pc-695F	5'—ATA TAG CAG CTC CTC CTT CCC TC—3'
	Nad4-Pc-1350F	5'—TAA CTG ACT AAA CTA ACC CGG —3'
	Nad4-Pc-990R	5'—TAA ACT TCG GGT ATG AGT TG—3'
	16S-Pc-96R	5'—TTA AGT CCT TTA CAG TAC TAA G—3'
	16S-Pc-240R	5'—GTG AAA CAT AAG ACG GTG GG—3'
	12S-Pc-640R	5'—ATC TCT TCT TTC TCA TAG GC—3'
	12S-Pc-180R	5'—TAA TCG AAT CTA GGT GTC CC—3'
	CystRNA—Pc-10R	5'—ATT TCA ACT ATG AGA CCG GG—3'
	Nad4-Pc-262R	5'—TCG TAA TAA ACC ATA GCT GCC—3'
<i>E. vanelli</i>		
<i>mLSU -cox1</i>	Nad1-Ev498F	5'—ATA ACA CTA ATT CTT TTA AGG C —3'
	Nad1-Ev854F	5'—GTC TTT TCC TCG AAT GCG—3'
	Nad2-Ev486R	5'—AGT CTG ATT TAT CCC ACC—3'
	Nad2-Ev938R	5'—ACT TAT TGC TAA CCT GCT C—3'
<i>cox1-cob</i>	COII-Ev331F	5'—TCT GAT GTT TGT GAT CTT GC —3'
	COIII-Ev40R	5'—ACC ATG AAA TCA CGA TAC C —3'
<i>cob-nad5</i>	Cytb-Ev-1042F	5'—TTA TGA CTG GTT AGG ACA GG —3'
	Nad5-Ev-200R	5'—ACT ATA GAT GTC ACC AAC CC—3'
	Nad5-Ev-582R	5'—GCT CTT TTT GTT ATA GCT GC —3'
	Nad5-Ev-384F	5'—GGG GTG AGA TGG TTT AGG TTT AAC—3'
	Nad4-Ev-1300R	5'—TAT TCA ACT ATA CAG CCA TAA ACC—3'
	Atp6-Ev-350F	5'—TTG GTT TTA CTA TTT GGG CGA G—3'
	tRNA-Ev-33R	5'—AAA TAC CCG TTC AAC CCG GC—3'
	Atp6-Ev-116F	5'—GGT TTT GAG TTG TTC CGA GG—3'

Fragments	Primer name	Sequence
	tRNA-Ev-30F	5'—AAC ATT GCG CCG GGT TGA ACG—3'
	Nad4-Ev-947R	5'—TAC TAT TAA ACC TAA AGC CC—3'
	Nad4-Ev-942F	5'—TAG GTT TAA TAG TAG CCC ACG G—3'
	Nad4L-Ev-132F	5'—AAC CAG GTT AGT AAC CCT GCG—3'
	Nad4-Ev-415R	5'—ATA AAA CCT AGC CTG CAA CCG—3'
	Nad4-Ev-1049R	5'—CAC CAA ATT CTC ATA CAC GG—3'
<i>mMSU-mLSU</i>	12S-Ecl-midF	5'—GTT TGG TTC TTG GTA TGG —3'
	16S-Ev-mid3R	5'—GCG CTA ACA ATA AGA AGG C—3'
	16S-Ecl-midR	5'—AAT AGA GAC AGC ATA ACC —3'

Table 4. Taxa used in phylogenetic analysis.

Species	Clade	Nucleotides	GenBank Number
<i>Terebellides stroemi</i>	Annelida, “Canalipalpata”, Trichobranchidae	15,755 complete	EU236701
<i>Pista cristata</i>	Annelida, “Canalipalpata”, Terebellidae	15,894 complete	EU239688
<i>Echysippe vanelli</i>	Annelida, “Canalipalpata”, Ampharetidae	13,749 partial	EU239687
<i>Clymenella torquata</i>	Annelida, “Scolecida”, Maldanidae	15,538 complete	AY741661
<i>Riftia pachyptila</i>	Annelida, “Canalipalpata”, Siboglinidae	12,016 partial	AY741662
<i>Galathealinum brachiosum</i>	Annelida, “Canalipalpata”, Siboglinidae	7,576 partial	AF178679
<i>Platynereis dumerilii</i>	Annelida, “Aciculata”, Nereididae	15,619 complete	NC_000931
<i>Lumbricus terrestris</i>	Annelida, “Oligochaeta”, Lumbricidae	14,998 complete	NC_001673
<i>Helobdella robusta</i>	Annelida, Hirudinea, Glossiphoniidae	7,553 partial	AF178680
<i>Orbinia latreillii</i>	Annelida, “Scolecida”, Orbiniidae	15,558 complete	AY961084
<i>Scoloplos cf. armiger</i>	Annelida, “Scolecida”, Orbiniidae	12,042 partial	DQ517436
<i>Phascolopsis gouldii</i>	Annelida, Sipuncula	7,470 partial	AF374337
<i>Urechis caupo</i>	Annelida, Echiura	15,113 complete	AY619711
<i>Katharina tunicata</i>	Mollusca, Polyplacophora	15,532 complete	NC_001636
<i>Terebratalia transversa</i>	Brachiopoda, Articulata	14,291 complete	NC_003086

Table 5. Base composition and skewness measures.

	<i>T. stroemi</i>			<i>P. cristata</i>			<i>E. vanelli</i>		
	Whole genome	Protein-coding genes	3 rd position	Whole genome	Protein-coding genes	3 rd position	Whole sequences	Protein-coding genes	3 rd position
A	31.0%	29.5%	40.4%	29.6%	28.0%	37.6%	27.2%	24.8%	29.8%
T	36.1%	37.3%	38.0%	38.5%	39.3%	44.1%	41.9%	43.7%	50.0%
G	12.8%	12.2%	4.5%	12.2%	11.6%	3.1%	19.3%	19.5%	15.9%
C	20.1%	21.0%	17.1%	19.7%	21.0%	15.2%	11.6%	12%	4.4%
A+T	67.1%	66.8%	78.4%	68.1%	67.3%	81.7%	69.1%	68.2%	79.8%
AT-skew	-0.075	-0.118	0.03	-0.13	-0.168	-0.079	-0.213	-0.277	-0.253
GC-skew	-0.221	-0.226	-0.585	-0.235	-0.288	-0.662	0.249	0.238	0.567

Table 6. Relative synonymous codon usages (RSCU) in the 13 protein-coding genes in *T. stroemi*, *P. cristata* and *E. vanelli*.

AA	Codon	<i>T. stroemi</i>	<i>P. cristata</i>	<i>E. vanelli</i>	AA	Codon	<i>T. stroemi</i>	<i>P. cristata</i>	<i>E. vanelli</i>							
Ala	GCA	82	1.276	104	1.541	57	1.163	Lys	AAA	93	1.86	76	1.949	59	1.457	
	GCC	57	0.887	52	0.77	4	0.082		AAG	7	0.14	2	0.051	22	0.543	
	GCG	3	0.047	6	0.089	19	0.388		Met	AUA	194	1.709	151	1.787	184	1.579
	GCU	115	1.79	108	1.6	116	2.367			AUG	33	0.291	18	0.213	49	0.421
Arg	CGA	41	2.563	42	2.625	22	1.222	Phe	UUC	91	0.565	51	0.289	10	0.054	
	CGC	9	0.563	3	0.188	2	0.111		UUU	231	1.435	302	1.711	361	1.946	
	CGG	3	0.188	3	0.188	17	0.944	Pro	CCA	84	1.768	67	1.396	35	0.979	
Asn	CGU	11	0.688	16	1	31	1.722		CCC	13	0.274	32	0.667	1	0.028	
	AAC	48	0.756	52	0.689	24	0.39		CCG	8	0.168	2	0.042	18	0.503	
Asp	AAU	79	1.244	99	1.311	99	1.61	CCU	85	1.789	91	1.896	89	2.49		
	GAC	30	0.882	21	0.627	13	0.371	Ser	AGA	39	0.782	49	1.071	78	1.651	
GAU	38	1.118	46	1.373	57	1.629	AGC		19	0.381	10	0.219	6	0.127		
Cys	UGC	14	1	10	0.5	10	0.328		AGG	9	0.18	2	0.044	45	0.952	
	UGU	14	1	30	1.5	51	1.672	AGU	18	0.361	17	0.372	47	0.995		
Gln	CAA	64	1.753	66	1.886	37	1.194	UCA	122	2.446	118	2.579	59	1.249		
	CAG	9	0.247	4	0.114	25	0.806	UCC	44	0.882	33	0.721	6	0.127		
Glu	GAA	59	1.71	65	1.857	35	0.875	UCG	3	0.06	2	0.044	14	0.296		
	GAG	10	0.29	5	0.143	45	1.125	UCU	145	2.907	135	2.951	123	2.603		
Gly	GGA	83	1.747	97	2.132	59	0.904	Ter(.)	UAA/U	13	2	13	2	10	1.538	
	GGC	30	0.632	28	0.615	18	0.276	UAG	0	0	0	0	3	0.462		
	GGG	32	0.674	20	0.439	90	1.379	Thr	ACA	91	1.562	98	1.587	49	1.441	
	GGU	45	0.947	37	0.813	94	1.44		ACC	38	0.652	59	0.955	8	0.235	
His	CAC	38	0.835	25	0.568	11	0.268	ACG	5	0.086	3	0.049	12	0.353		
	CAU	53	1.165	63	1.432	71	1.731	ACU	99	1.699	87	1.409	67	1.971		
Ile	AUC	106	0.607	78	0.449	13	0.099	Trp	UGA	92	1.804	80	1.616	41	0.812	
	AUU	243	1.393	269	1.55	250	1.901		UGG	10	0.196	19	0.384	60	1.188	
Leu	CUA	147	1.586	114	1.136	30	0.35	Tyr	UAC	55	0.873	36	0.61	21	0.288	
	CUC	23	0.248	60	0.598	2	0.023		UAU	71	1.127	82	1.39	125	1.712	
	CUG	12	0.129	3	0.03	10	0.117	Val	GUA	66	1.833	42	1.292	75	0.92	
	CUU	122	1.316	193	1.923	75	0.875		GUC	22	0.611	15	0.462	14	0.172	
	UUA	236	2.547	216	2.153	285	3.327		GUG	7	0.194	9	0.277	54	0.663	
	UUG	16	0.173	16	0.159	112	1.307		GUU	49	1.361	64	1.969	183	2.245	
2fold	NNA	100.4	1.767	87.6	1.819	71.2	1.183	4fold	NNA	74.5	1.792	75.0	1.762	49.5	1.105	
	NNG	13.8	0.233	9.6	0.181	40.2	0.817		NNC	28.2	0.603	31.5	0.610	7.8	0.151	
	NNC	54.6	0.788	39.0	0.533	14.6	0.257		NNG	9.7	0.226	7.2	0.181	35.0	0.705	
	NNU	104.1	1.212	127.3	1.467	144.9	1.743		NNU	67.3	1.379	67.2	1.448	96.7	2.039	

AA, amino acid; Ter(.), Terminator codons

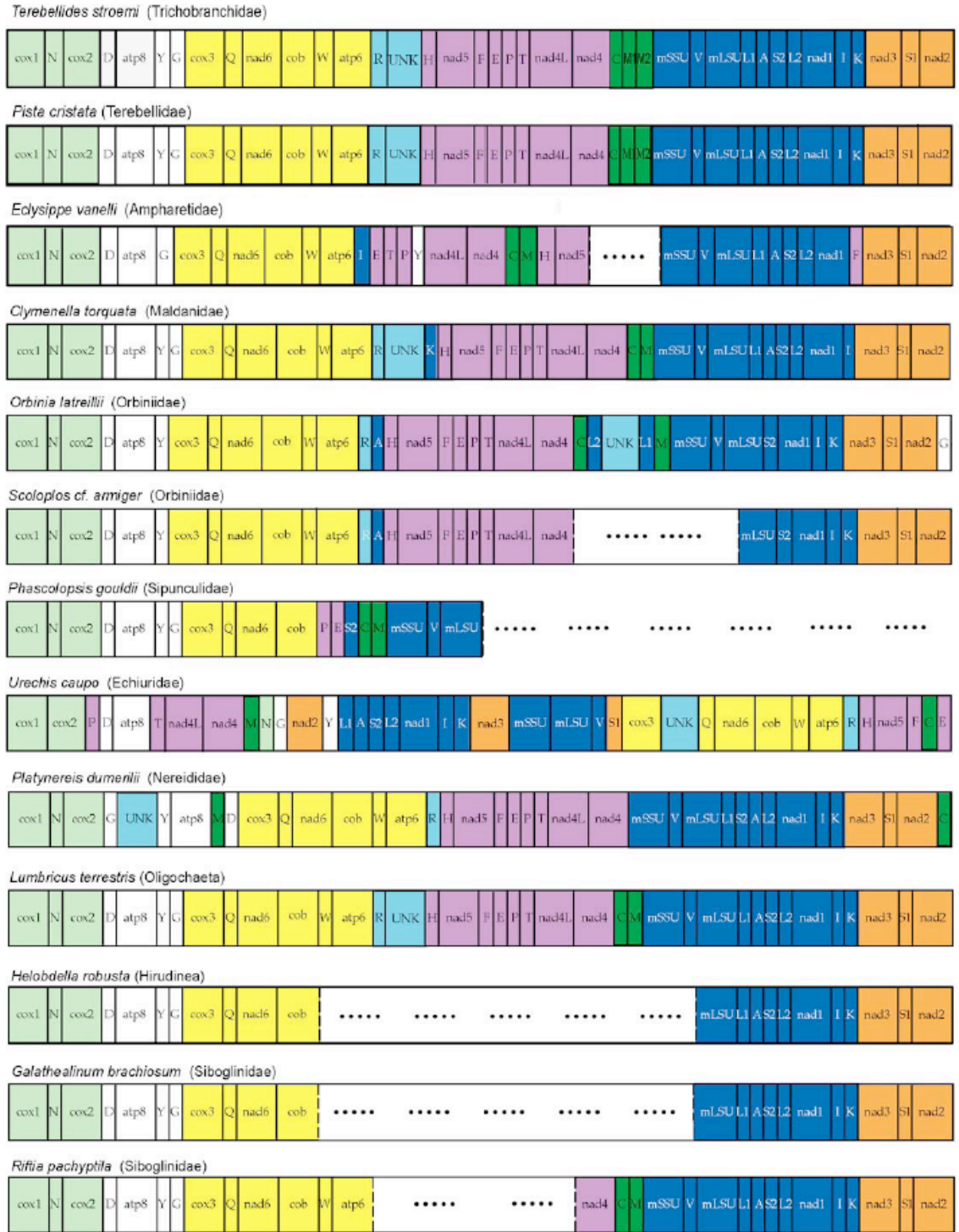
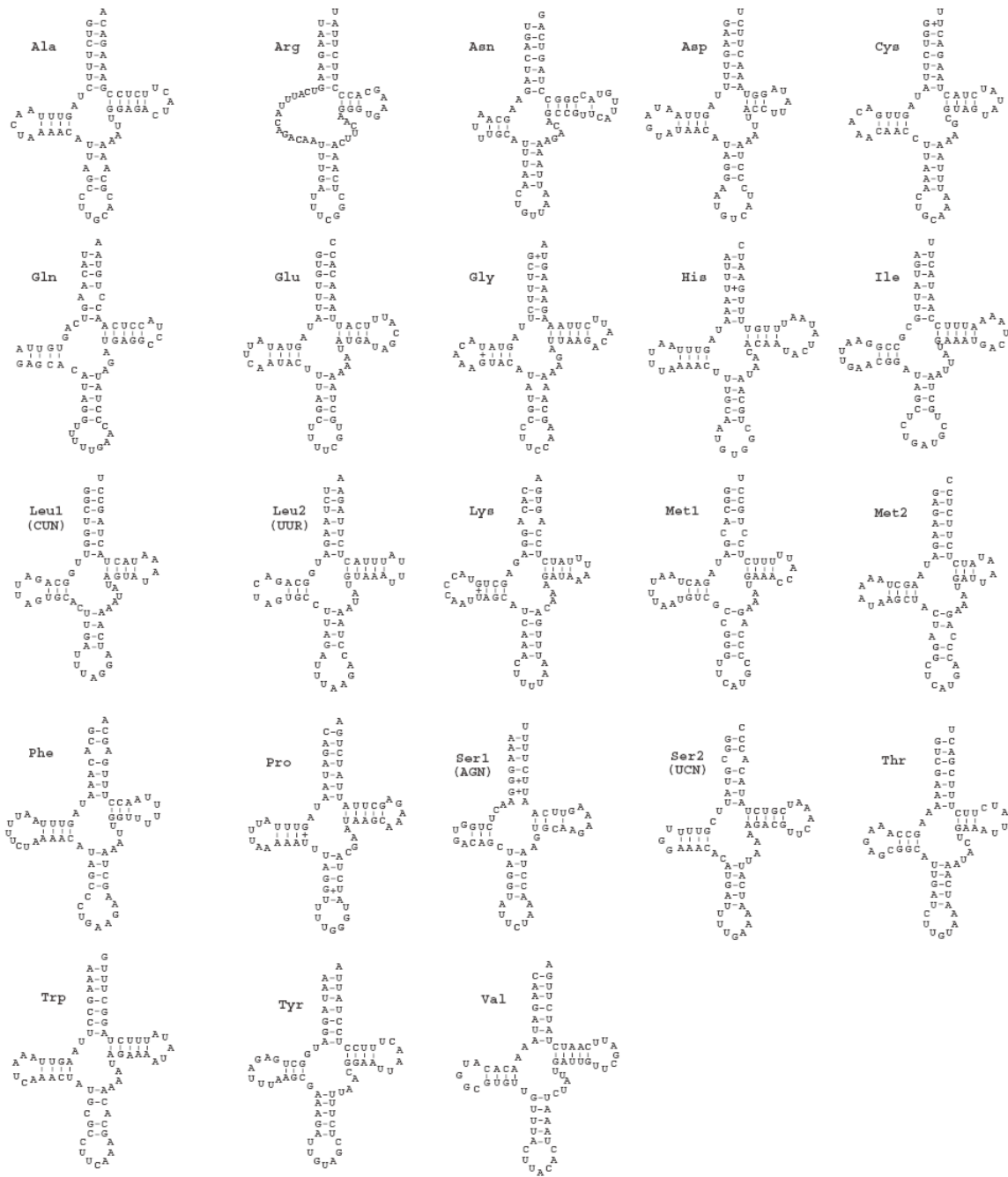
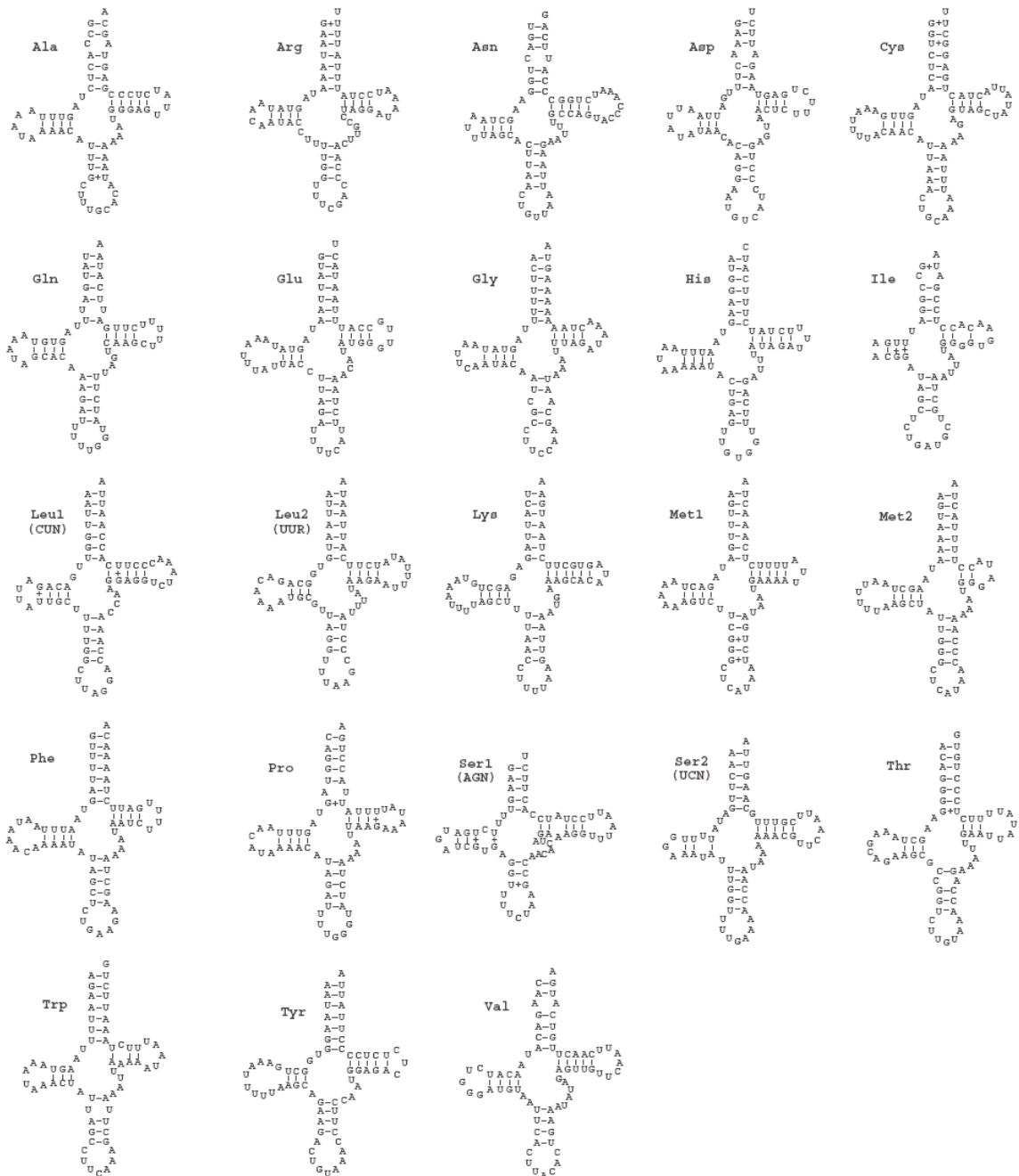


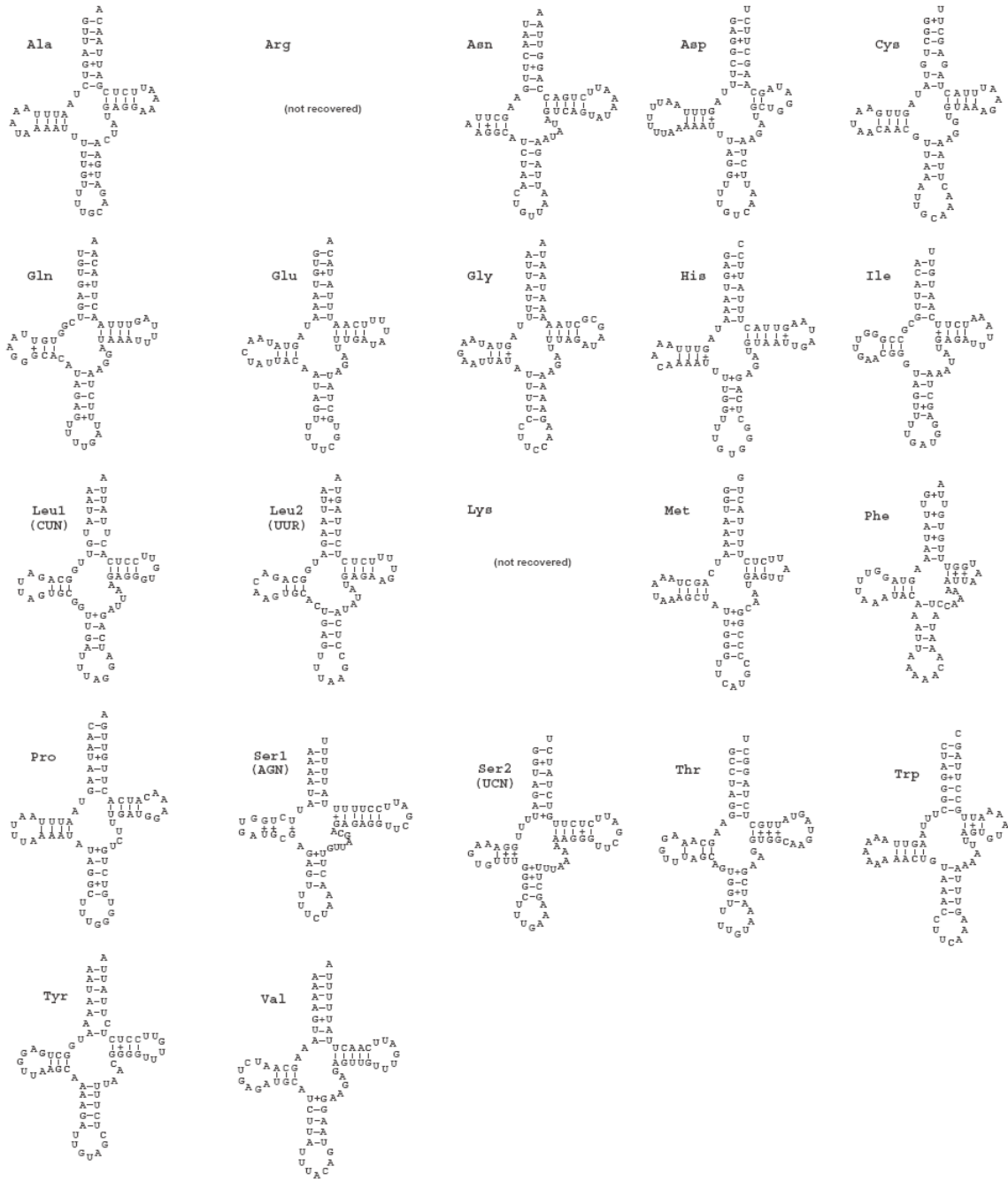
Figure 1. Gene orders of mitochondrial genomes in Annelida. Different colors shows conserved gene clusters. Dots indicate missing regions.



A.



B.



C.

Figure 2. Putative secondary structures of tRNA genes in 3 Terebelliformia worms. (A) 23 tRNA genes in *T. stroemi*. (B) 23 tRNA genes in *P. cristata*. (C) 20 recovered tRNA genes in *E. vanelli*.

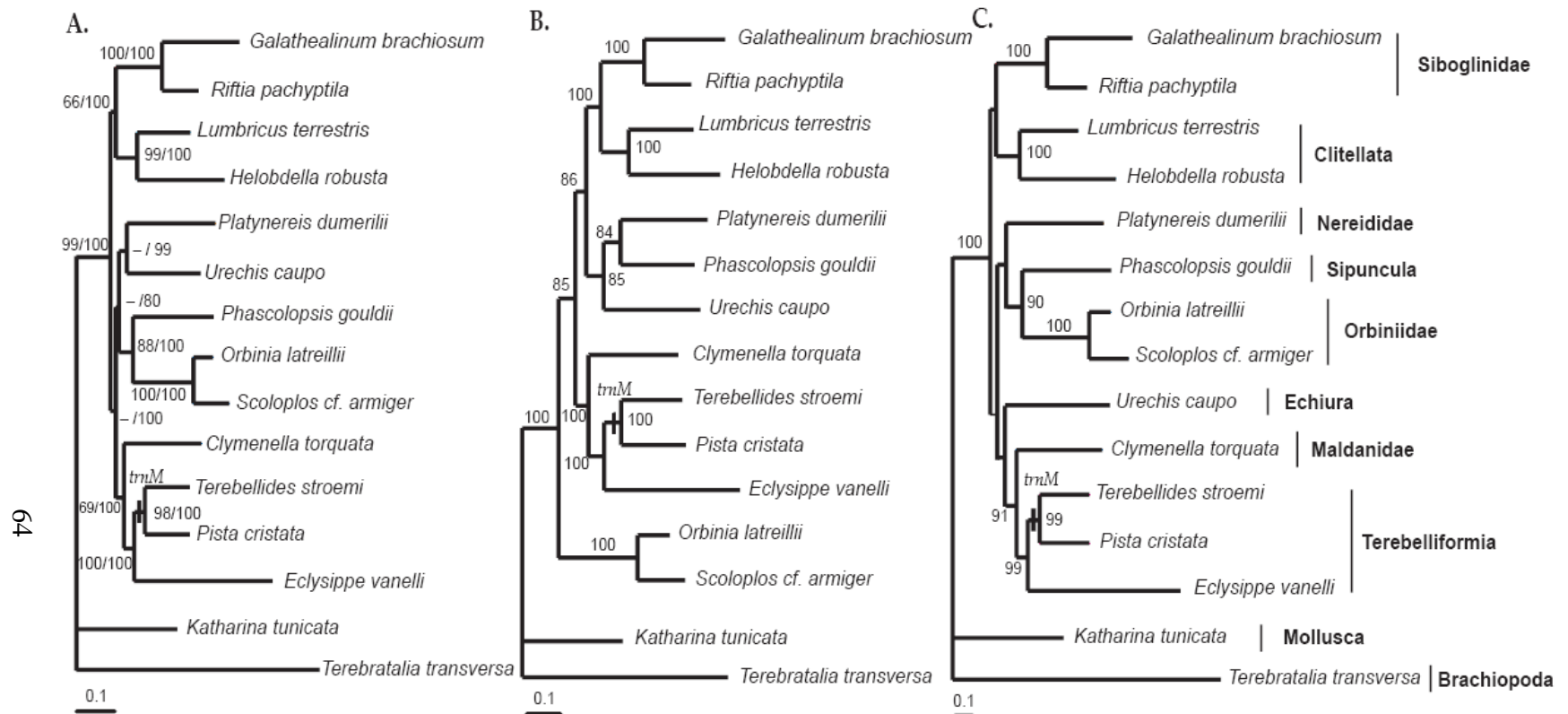


Figure 3. Phylogenetic reconstructions. (A) The single combined tree represents the identical topology from both ML and Bayesian inference methods (non-partitioned and partitioned) with GTR+ Γ +I model. Nodal support values are given at branches with ML bootstrap values first and posterior probabilities of the partitioned Bayesian analysis second (non-partition values not shown). A dash indicates < 50% on trees. (B) Non-partitioned Bayesian analyses of amino acid dataset with the mixed amino acid substitution model. Posterior probabilities are shown at the nodes. (C) ML analyses of amino acid dataset using RAxML by 200 bootstrap replicates. Bootstrap values are shown at the nodes. Black bars indicate *trnM* gene duplication event. *Terebratalia transversa* (brachiopod) and *Katharina tunicata* (mollusk) were used as outgroups. Details of analyses are given in the text.

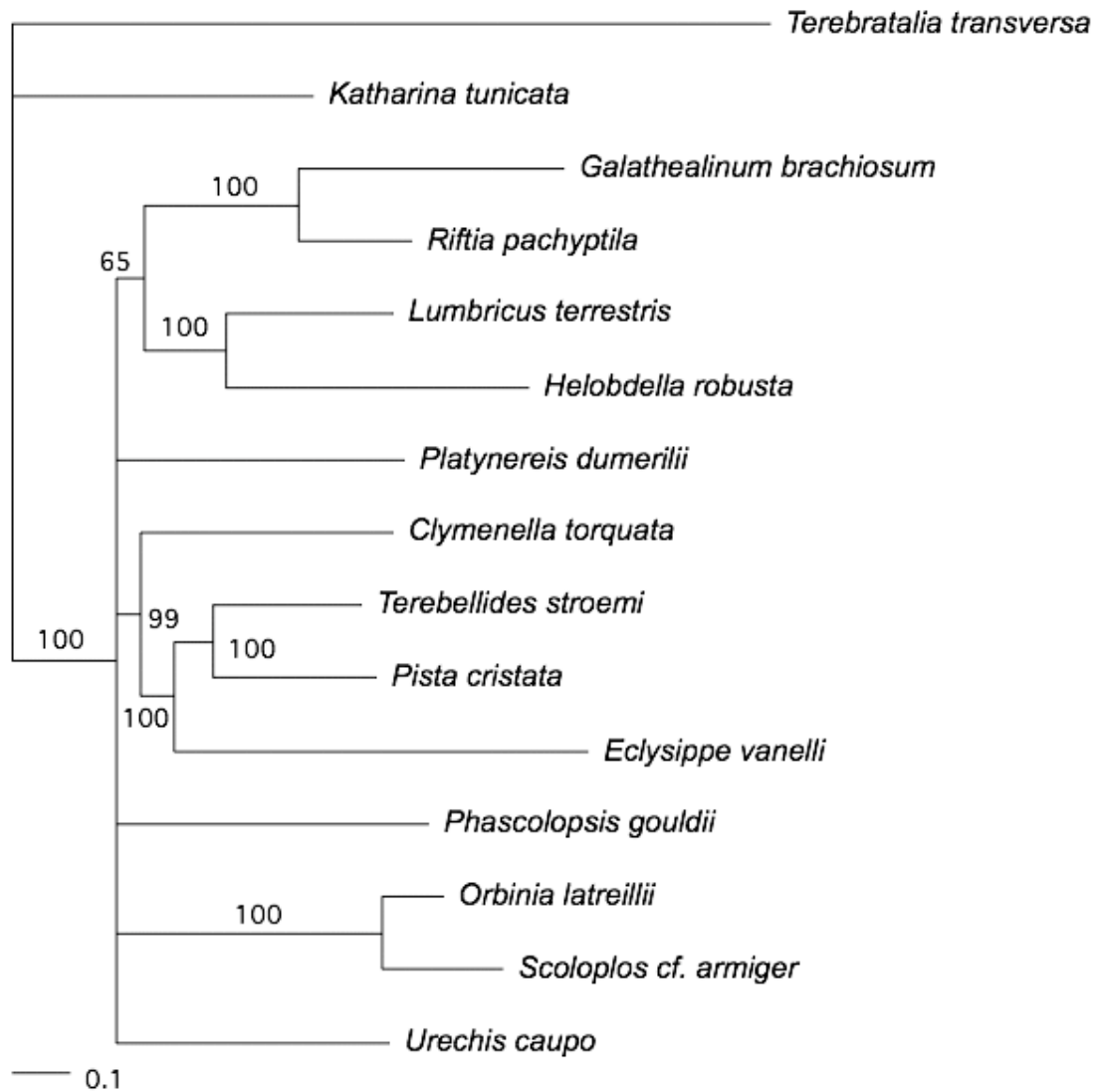


Figure 4. Partitioned Bayesian analyses of amino acid dataset with the mixed amino acid substitution model with partitions unlinked during the run. Posterior probabilities are shown at the nodes.

CHAPTER 3

PHYLOGENETIC INFERENCE OF TEREPELLIFORMIA WORMS BASED ON MITOCHONDRIAL GENOMIC DATA

Abstract

Terebelliformia worms are mainly tube-dwelling annelids comprising of five ‘families’, Alvinellidae, Ampharetidae, Terebellidae, Trichobranchidae and Pectinariidae. Even though Terebelliformia is one of the few, more-inclusive clades of Annelida well-defined by morphology, relationships among terebelliform lineages are not resolved. Here we addressed phylogenetic relationships among five Terebelliformia families using mitochondrial genomic datasets. Monophyly of Terebelliformia was strongly supported. Both nucleotide and amino acid datasets supported a sister relationship between Ampharetidae and Alvinellidae and the basal placement of Pectinariidae. Trichobranchidae and Terebellidae were sisters to each other with high nodal values. Furthermore, we mapped mitochondrial gene order characters (*trnM* duplication and gene order of *nad5-nad4L-nad4* cluster) on the topology to understand mtDNA genome evolution. *trnM* duplication was presumed to be a primitive trait of terebelliforms. Ampharetidae and Alvinellidae lineages have secondarily lost the duplication event. Based on the mapping of the *nad5-nad4L-nad4* gene order, ampharetids were indicated to be a more divergent clade.

1. Introduction

Terebelliformia worms are mainly tube-dwelling annelids, found in diverse marine habitats, including intertidal, deep-sea and even hydrothermal vents. Terebellidae is a species-rich family containing over 400 described species (Hutchings, 2000), living in marine environments worldwide. Terebelliformia worms comprise of five ‘families’, Alvinellidae, Ampharetidae, Terebellidae, Trichobranchidae and Pectinariidae (Hessle, 1917; Holthe, 1986; Rouse and Pleijel, 2001).

These annelid clades have a long and intertwined history. Terebellidae was erected as a family in 1850, although the first terebellid worm, *Terebella lapidaria*, was described in 1776 (Müller, 1776; Grube, 1850). Trichobranchidae, only containing a few genera (*Octobranchus*, *Terebellides*, *Trichobranchus* and *Artacamella*), was originally treated as a subfamily Trichobranchinae within Terebellidae (Malmgren, 1866) and suggested to be family rank by Hessle in 1917. Ampharetidae, which typically live in fragile tubes as deposit feeders, is another species-rich family from relative diverse habitats. Many ampharetids have been found in deep sea with some taxa occurred in hydrothermal vent areas and fresh water (Solis-Weiss, 1993; Desbruyères and Laubier, 1991; Holthe, 1986). Ampharetidae was referred to as a family in 1866 by Malmgren (1866). Pectinariidae, first recognized as a separate taxon by Quatrefages in 1865, was once named as Amphictenidae and recurred in a ruling of the International Commission on Zoological Nomenclature in 1982 (Rouse and Pleijel, 2001). Pectinariids can be

generally found in muddy or sandy shallow waters with filter/suspension feeding during burrowing and digging activity (Watson, 1928). Alvinellidae was originally erected based on the description of the “Pompeii worm”, *Alvinella pompejana*, Desbruyères and Laubier 1980. Alvinellids were originally placed as a subfamily within Ampharetidae, and only later described as a separate family (Desbruyères and Laubier 1986).

Phylogenetic relationship among these five families is ambiguous and consistently debated by recent researchers (Rouse and Fauchald, 1997; Colgan et al., 2001; Rousset et al., 2003; Glasby et al., 2004; Rousset et al., 2007; Struck et al., 2007; Zhong et al., 2008). For example, Trichobranchidae was originally treated as a subfamily, Trichobranchinae, within Terebellidae (Malmgren, 1866), but later elevated to family rank (Hessle, 1917) because two genera of Trichobranchidae (*Trichobranthus* and *Terebellides*) were thought not to be closely related. *Trichobranthus* was placed close to Ampharetidae based on similar branchiae, and *Terebellides* was placed close to Terebellidae based on the digestive system. Placement of Trichobranchidae has been debated by subsequent authors (Day, 1967; Fauchald, 1977; McHugh, 1995; Fauchald and Rouse, 1997; Rouse and Fauchald, 1997; Rouse and Pleijel, 2001; Colgan et al., 2001). Rouse and Fauchald (1997) referred a close affinity between Terebellidae and Trichobranchidae and, based on morphological characters, placed this group as sister to an Alvinellidae/Ampharetidae/Pectinariidae clade. In contrast, combined morphological and molecular data suggest that Trichobranchidae is sister to Alvinellidae and not Terebellidae (Rousset et al., 2003). Although the close relationship between alvinellids and ampharetids has been suspected based on morphology (Rouse and Fauchald, 1997; Glasby et al., 2004), it has not been supported by molecular data (e.g., Rousset et al.,

2003; Struck et al., 2007). In contrast, some have assumed Ampharetidae as closely related to Terebellidae and Trichobranchidae (Hessle, 1917; Fauchald and Rouse, 1997). However, Pectinariidae, together with Terebellidae and Ampharetidae, was proposed to be Terebellida or Terebellimorpha (Fauchald and Rouse, 1997). Recent morphological analyses indicated that Pectinariidae as either sister to Ampharetidae and Alvinellidae clade (Rouse and Fauchald, 1997), or related to Terebellidae (Glasby, et al., 2004). Nonetheless, some molecular studies and combined data indicated contradictory views for terebelliform relationships. Based on the five combined genes, *Trichobranchus* was suggested to be associated with one subfamily of Terebellidae making Trichobranchidae paraphyletic (Colgan et al., 2001). Whereas a sister relationship between Trichobranchidae and Alvinellidae was suggested based on combined molecular and morphological data (Rousset et al., 2003). Furthermore, using complete mitochondrial genomes, the terebellid *Pista cristata* and the trichobranchid *Terebellides stroemi* were recovered together (Zhong et al., 2008). This study also reported additional copy of the methionine tRNA in these taxa, and such duplications are not common bilaterian mitochondrial genomes. Elucidating the affinity of Pectinariidae has been particularly problematic as it has been recovered as sister to a Trichobranchidae./Alvinellidae clade (Rousset et al., 2003), within Ampharetidae (Colgan et al., 2001), basal to all other terebelliform lineages (Struck et al., 2007), or even other annelids outside of Terebelliformia (Rousset et al., 2007).

As more and more bilaterian mitochondrial genomes are sequenced, they become useful for phylogenetic reconstructions of major groups. Mitochondrial genomes are generally 15-17kb in size and contain conserved homologous genes across bilaterian

animals (Boore, 1999; Vallès and Boore, 2006). Apart from highly variable non-coding UNK region (unknown region), typical bilaterian mitochondrial genomes are comprised of 37 genes with 13 protein-coding, two ribosomal and 22 tRNAs. The gene arrangement of mitochondrial genomes within annelids is presumably conserved with some exceptions (Jennings and Halanych, 2005; Vallès and Boore, 2006; Zhong et al., 2008).

To elucidating the ambiguous annelid phylogeny so as to provide a powerful framework to further studies, the goal of the present study is to further examine relationships of Terebelliformia by examining a mitochondrial genome data. Partial mitochondrial genomes of three Terebelliformia worms (*Auchenoplax crinita* – Ampharetidae, *Paralvinella sulfincola* – Alvinellidae and *Pectinaria gouldii* -- Pectinariidae) are reported here. Both nucleotide and amino acid datasets were established to reconstruct the evolutionary history within Terebelliformia.

2. Materials and Methods

2.1 Sample collection and DNA extraction

Auchenoplax crinita (Ampharetidae) was collected south of Cap Code, Massachusetts, USA, *Pectinaria gouldii* (Pectinariidae) was obtained in Egypt Lane, Fairhaven, Massachusetts, USA, and *Paralvinella sulfincola* (Alvinellidae) was collected in Juan de Fuca Ridge system on the Visions 05 cruise in Oct. 2005 (and kindly provided by Dr. Peter Girguis from Harvard University). Each organism is a representative of every above family. All organisms were stored at -80°C after collection. Total genomic DNA extractions employed the DNeasy® Tissue Kit (Qiagen) according to the manufacture's instructions.

2.2 mtDNA data collection

Gene nomenclature and abbreviations in this paper follow Jennings and Halanych (2005).

Auchenoplax crinita and *Paralvinella sulfincola*

The genomes were amplified in four overlapping segments. Conserved regions of *mLSU*, *cox1*, *cob* and *nad5* genes were amplified using taxonomically-inclusive primers (Zhong et al., 2008) and purified by QIAquick PCR purification kit (Qiagen). PCR products were sequenced on a CEQ8000 (Beckmann) and four pairs of species-specific primers for long PCRs were designed (Table 1). HotStart long PCRs were performed on Eppendorf Mastercycler (Eppendorf) using Takara LA- Taq™ (Takara) by following protocol: 94 °C for 3min; then addition of polymerase followed by 35 cycles with 94°C for 30 sec, 52°C for 30 sec, and 70°C for 8 min; final extension at 72°C for 10 min and hold at 6°C. 50 µl long PCR reactions using Takara LA-Taq were set up as follows: 5µl 10×buffer, 8µl dNTP (2mM), 5µl MgCl₂ (25mM), 2µl of each long PCR specific primers (10µM each), 0.5µl Takara LA-Taq (5U/µl), 2µl DNA template and 25.5µl sterilized distilled water. Both *mLSU-cox1* and *cox1-cob* regions were about 4 kb in size. Both *cob-nad5* fraction of *A. crinita* and *nad5- mLSU* of *P. sulfincola* were obtained about 6-7 kb in size. The unobtained *nad5-mLSU* of *A. crinita* and *cob-nad5* of *P. sulfincola* containing the UNK region, were difficult to amplify despite many tries (see Boore and Brown, 2000; Jennings and Halanych 2005).

All long fragments were purified with QiaQuick Gel Extraction Kit (Qiagen) and sequenced directly or cloned into the pGEM-T Easy vector (Promega). Clones were

verified by PCR and plasmids were isolated using the QIAprep[®] Spin Miniprep Kit (Qiagen). Purified plasmids were digested by *EcoRI* to check the insert size and sequenced by primer walking.

Pectinaria gouldii

The genome of *P. gouldii* was amplified in six segments. Conserved fragments of *mLSU*, *cox1*, *nad5* and *nad4* genes were amplified by taxonomically inclusive primers (Zhong et al., 2008) and sequenced. Two new primers were designed based on *cox3* conserved region (see Table 1). Specific long-PCR primers (Table 1) were then designed to amplify long fragments: *mLSU-cox1*, *cox1-cox3*, *nad5-nad4* and *nad4-mLSU*. The universal *cob* reverse primer “Cytb 876R” was used to pair with *cox3* forward one to amplify *cox3-cob* fragment. After sequence it, a specific *cob* forward long-PCR primer was designed to pair with *nad5* reverse primer. Long PCRs employed Takara LA-Taq as described above. The protocols of long PCR and purification were similar with some modifications of annealing temperatures based on different pairs of primers (Table 1). All long fragments were sequenced directly by primers walking. The unobtained *cob-nad5* fragment was hard to obtain presumably due to the inclusion of the UNK region. Information for all sequencing primers can be found in the table 2.

2.3 Genomic Assembly

Sequences were edited and aligned using DNASTAR[™] Lasergene programs SeqMan and MegAlign (Burland, 2000). Protein-coding genes and ribosomal RNA genes were identified by BLAST (Altschul et al., 1990). tRNA genes were identified using tRNAscan-SE web server (<http://lowelab.ucsc.edu/tRNAscan-SE/>; Lowe and Eddy, 1997)

under default settings and source = “mito/chloroplast”, or by hand based on their potential secondary structures and anticodon sequences.

2.4 Phylogenetic analyses

Seventeen available annelid mitochondrial genomes with 50% coverage or greater were used for phylogenetic analyses. The alignment of Zhong et al. (2008) was employed with the additions of *Nephtys sp. 'San Juan Island'*, *Pectinaria gouldii*, *Paravinella sulfincola* and *Auchenoplax crinita* (Table 3). Because we are interested in the relationships within Terebelliformia, we deleted the mitochondrial data of *Katharina* (Mollusca) and *Terebratalia* (Branchiopoda) from the Zhong et al. data set and used all other annelids as outgroups.

Phylogenetic analyses were based on nucleotide and amino acid datasets. The nucleotide dataset included all protein-coding genes (except for *atp6*, *atp8* and *nad6* genes which exhibit high variability) and the two rRNA genes (*mLSU* and *mSSU*). Clustal X (Thompson et al., 1997) under default setting was used to realign rRNA genes. MacClade4.08 (Maddison and Maddison, 2002) was used to exclude the most regions that contain insertions/deletions and all third codon positions in protein coding genes. Gblocks 0.91b (Castresana, 2000) was used to identify ambiguous aligned regions in rRNA genes that were excluded from analyses. Using MacClade4.08 (Maddison and Maddison, 2002) and Se-A1 v2.0a11 (Rambaut, 1996), the amino acid dataset was created from the aligned nucleotide dataset translated with the *Drosophila* mitochondrial code, and with rRNA genes excluded.

Both maximum likelihood and Bayesian-inference approaches were employed on all datasets. For the nucleotide dataset, Maximum-likelihood analyses were performed in PAUP4.0b10 (Swofford et al., 2002) with a GTR+ Γ +I model as determined by MODELTEST v3.7 based on the Akaike information criterion (AIC) (Posada and Crandall, 1998). Heuristic searches were run with random-taxon addition (10 replicates) using Tree-Bisection-Reconnection (TBR) swapping. All model parameters used fixed values as determined by MODELTEST v3.7. Bootstrap analysis employed 1,000 iterations using heuristic searches with 10 random taxa addition. Partitioned Bayesian inference analyses in MRBAYES version 3.1.2 (Huelsenbeck and Ronquist, 2001) used 5,000,000 generations with 2 runs of chains (3 heated and 1 cold), and sampling every 100 generations. Unlinked GTR+ Γ +I models were selected under the AIC in MrModeltest (Nylander, 2004) for which parameters were separately estimated for each gene partition with the exception of GTR+I models for both 12S and 16S genes; GTR+ Γ model for *cox3* gene and HKY+ Γ model for *nad2*, *nad4L* and *nad5* genes. Resulting -ln likelihood scores were graphed using X-Y scatter plots to identify the “burn-in” point at which all estimated parameters reached stationarity (burnin = 10,000). For the nucleotide dataset of the combined matrix, similar ML and partitioned Bayesian analyses were performed in PAUP and MrBayes separately after model-selections.

For the amino acid dataset, a non-partitioned ML analyses was run in addition to partitioned Bayesian analyses. For ML analyses, model selection was performed in ProtTest (Abascal et al., 2005) and MtArt + Γ +F model was chosen as the best one under the AIC. As there is no MtArt model used in RAxML, we chose the next best model, MtREV+ Γ +F, available in RAxML. A maximum likelihood search was implemented by

200 bootstrap replicates using RAxML web-server (<http://phylobench.vital-it.ch/raxml-bb/>; Stamatakis, 2006) with MtREV+ Γ +F model. In partitioned Bayesian analysis, the mixed amino acid substitution model option plus a Γ distribution and a proportion of invariant sites was assigned to each partition individually and unlinked in MRBAYES v3.1.2 with 2,000,000 generations sampled every 500 generations (burnin = 10,000). In the mixed model option, a specific model is not specified a priori, but each model is chosen during the runs based on its posterior probability.

3. Results

3.1 Mitochondrial genomes and gene order

The three genomes recovered generally conform to the conserved annelid mtDNA gene order (Jennings and Halanych, 1995), but for all recover the UNK region proved difficult and can't be discovered here. The partial mtDNA of *Aucheoplax crinita* is 13,759 bp in length, of *Paralvinella sulfincola* is 13,640 bp and of *Pectinaria gouldii* is 13,438 bp. *A. crinita* contains 34 recovered genes including 13 protein-coding genes, two rRNA genes and 19 tRNAs. There are 33 genes found in mtDNA of *P. sulfincola* with 12 protein-coding genes, two rRNA genes and 19 tRNAs, 34 genes in the one of *P. gouldii* with 12 protein-coding genes, two rRNA genes and 20 tRNAs. Interestingly, the partial genome of *P. gouldii* possesses the identical gene order as both the trichobranchid *T. stroemi* and the terebellid *P. cristata*, including adjacent duplicate methionine tRNA gene (see Zhong et al., 2008). Fig. 1 shows gene orders of all six available terebelliform taxa. All mitochondrial genes are transcribed from the same strand, similar to other annelids (Boore and Brown, 2000; Boore and Staton, 2002; Jennings and Halanych, 2005;

Bleidorn et al., 2006a, 2006b; Zhong et al., 2008). Gene order of *A. crinita* is identical to another published ampharetid, *E. vanelli*, (Zhong et al., 2008) except for the *trnF* between *nad1* and *nad3* genes in *E. vanelli*.

3.2 Phylogenetic analyses

The mitochondrial nucleotide dataset comprised 17 taxa with 6287 unambiguous characters. Both maximum likelihood and partitioned Bayesian analyses inferred an identical topology (Fig. 2A). The amino acid dataset from 10 mitochondrial protein-coding genes (with the exception of *atp6*, *atp8* and *nad6* genes) consisted 17 taxa with 3149 unambiguous characters. Maximum likelihood (RAxML) and Bayesian inference of amino acid data resulted in the same topology (Fig. 2B), which was identical to the nucleotide topology within Terebelliformia (including strong nodal support), but differed in outgroup relationships. Monophyly of Terebelliformia is also well supported by both datasets (BS: 100 for DNA dataset /97 for AA-dataset; PP: 1.00 for both). The sister relationship between Trichobranchidae and Terebellidae was strongly suggested by mitochondrial datasets (BS: 95 for DNA/85 for AA; PP: 1.00 for both). Two Ampharetidae taxa were close to Alvinellidae with highest values from both analyses for both datasets (BS: 100; PP: 1.00). The mitochondrial datasets also showed that the Trichobranchidae/Terebellidae clade was sister to Ampharetidae/Ampharetidae. However, it didn't have a good nodal support value by the ML analysis for DNA dataset (BS: 69). Pectinariidae was shown to be a basal clade in Terebelliformia (BS: 100 for DNA/97 for AA; PP: 1.00 for both). Outgroup topologies differed in the placement of *Urechis caupo*

(*Echiura*), *Platynereis dumerilii* (Nereididae) and *Nephtys sp.* (Nephtyidae), suggesting it still need more informative datasets to make clear their relationships.

4. Discussion

4.1 Phylogenetic relationship within Terebelliformia

Monophyly of Terebelliformia was recovered in both topologies, agreeing with some morphological and molecular studies (Rouse and Fauchald, 1997; Colgan et al., 2001; Rousset et al., 2004; Rousset et al., 2007). Furthermore, Zhong et al. (2008) hypothesis that *Clymenella torquata* might be a sister group to Terebelliformia is supported herein by both mitochondrial datasets (Fig. 2).

In our analyses, Pectinariidae was placed as the most basal Terebelliformia lineage in both nucleotide and amino acid topologies with strong nodal supports (BS: 100 for both mtDNA and AA; PP: 1.00 for mtDNA/97 for AA). This placement is congruent with previous molecular study (Struck et al., 2007) of annelid phylogeny which showed Pectinariidae basal. Both trees provided strong support for the close affinity between Ampharetidae and Alvinellidae, which contradicted views proposing *Paralvinella* (Alvinellidae) as close to Terebellidae (Colgan et al., 2001; Rousset et al., 2007), or to Trichobranchidae (Rousset et al., 2003). However, this relationship was congruent with the previous morphological studies which described the similar features, such as buccal tentacles attached to a dorsal curtain and retractable into the mouth, hood-like upper lip regions and so on (Rouse and Fauchald, 1997; Glasby et al., 2004).

A sister relationship between Trichobranchidae and Terebellidae was strongly supported by both datasets (Fig. 2) and the identical mitochondrial gene order (Zhong et

al., 2008). This was in agreement with the family-based morphological cladistic analyses (Rouse and Fauchald, 1997). In contrast, other morphological and molecular analyses of Terebelliformia did not find a closer relationship of Terebellidae and Trichobranchidae (Rousset et al., 2003; Galsby et al., 2004; Struck et al., 2007). Additionally, all datasets used here haven't included the genus *Trichobranchus*, which was speculated to be distant to *Terebellides* used in our analyses (Hessle, 1917; Colgan et al., 2001). We would consequently suggest the further investigations with larger sampling are still in need, especially the coverage of *Trichobranchus* genus.

4.2 *trnM* duplication event and gene order mapping

Character mapping on the phylogenetic tree is essential to understand the evolutionary history of some traits (Felsenstein, 1985). Parsimony is the most universal method used to map characters implemented in some programs, such as MacClade and Mesquite (Maddison and Maddison, 2002; 2004). Two adjacent *trnM* genes were discovered in pectinariid *P. gouldii*, as well as terebellid and trichobranchid worms (Zhong et al., 2008), but not in the alvinellid and ampharetids. The unique duplication event was speculated based on their similar secondary structure, identical anticodons and the same sequences in stems. Their functions are still under explorations. To examine the evolutionary history of *trnM* duplication trait, we mapped it on the Terebelliformia tree by hand as well as the gene order of *nad5-nad4L-nad4* cluster characters (Fig. 3). As Pectinariidae was highly supported as a basal clade of Terebelliformia, *trnM* duplication was noted to be a symplesiomorphy of terebelliforms instead of synapomorphy for Trichobranchidae and Terebellidae

(Zhong et al., 2008). Both Ampharetidae and Alvinellidae have secondarily lost the duplication event and one *trnM*. In term of the *nad5-nad4L-nad4* gene cluster, both ampharetids have different gene arrangement from all other published annelids (except for the echiurid *Urechis caupo*) (Fig. 1). Their *nad5* were placed to 3' end of *nad4L* and *nad4* genes with *trnF-trnE-trnP-trnT* genes translocated. Comparing with Alvinellid *Paralvinella sulfincola* that only has variable tRNA gene order with the invariable protein-coding gene order, ampharetids have more evolutionarily divergent.

The well-resolved phylogenetic trees of terebelliforms in our study sill provide some insight into the potential capability of mtDNA for annelid phylogenetic reconstruction. However, the combined datasets including mitochondrial genomes and nuclear genes are still highly suggested for continued research investigations as well as the large sampling size for each family.

Conclusions

Our data addressed well-resolved phylogenetic relationships within Terebelliformia worms based on mitochondrial genome data. (1) Pectinariidae was placed as a basal clade to all other Terebelliformia families. (2) Ampharetidae and Alvinellidae were sister to each other. (3) Trichobranchidae and Terebellidae were sister clade with strong support by both phylogenetic analyses and mitochondrial gene order, but require further attention with large samplings. (4) *trnM* duplication event was inferred to be a symplesiomorphy of terebelliforms. Both Ampharetidae and Alvinellidae have secondarily lost the duplication.

Acknowledgements

We are grateful to Dr Peter Girguis (Harvard University) for kindly providing *Paralvinella sulfincola* samples. This work was supported by the USA National Science Foundation (NSF) WormNet grant (EAR-0120646) to K.M.H. and by the Alabama Commission on Higher Education Graduate Research Scholar's Program through the Auburn University Cellular and Molecular Biosciences Program and NSF EPS-0447675.

References

- Abascal, F., R. Zardoya, D. and Posada. 2005. ProtTest: Selection of best-fit models of protein evolution. *Bioinformatics*. 21:2104–2105.
- Altschul, S. F., W. Gish, W. Miller, E. M. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.
- Bleidorn, C., L. Podsiadlowski, and T. Bartolomaeus. 2006a. The complete mitochondrial genome of the orbiniid polychaete *Orbinia latreillii* (Annelida, Orbiniidae)--A novel gene order for Annelida and implications for annelid phylogeny. *Gene*. 370:96–103.
- Bleidorn, C., I. Kruse, S. Albrecht, and T. Bartolomaeus. 2006b. Mitochondrial sequence data expose the putative cosmopolitan polychaete *Scoloplos armiger* (Annelida, Orbiniidae) as a species complex. *BMC Evol. Biol.* 6:47.
- Boore, J. L. 1999. Animal mitochondrial genomes. *Nucleic Acids 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 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.
- Burland, T. G. 2000. DNASTAR's lasergene sequence analysis software. *Methods Mol. Biol.* 132:71–91.
- Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17:540–552.
- Colgan, D. J., P. A., Hutchings, and S. Brown. 2001. Phylogenetic relationships within the Terebellomorpha. *J. Mar. Biol. Assoc. U.K.* 81:765–773.
- Day, J. H. 1967. A monograph on the Polychaeta of South Africa. Part2. Sedentaria. London: British Museum. Natural History.
- Desbruyères, D. and L. Laubier. 1980. *Alvinella pompejana* g.en sp. nov., Ampharetidae aberrant des sources hydrothermales de la ride Est-Pacifique. *Ocean. Acta*, 3:267–274.
- Desbruyères, D, and L. Laubier. 1986. Les Alviellidae, une famille nouvelle d'annélides polychètes inféodées aux sources hydrothermales sous-marines: systématique, biologie et écologie. *Can. J. Zoo.* 64:2227–2245.
- Desbruyères, D. and L. Laubier. 1991. Systematics, phylogeny, ecology and distribution of the Alvinellidae (Polychaeta) from deep-sea hydrothermal vents. *Ophelia. Proceedings of the 2nd International Polychaeta Conference, Supplement.* 5:31–45.

- Fauchald K. 1977. The polychaete worms. Definitions and keys to the orders, families and genera. Natural History Museum of Los Angeles City Science Series. 28:1–88.
- Fauchald, K., and G. W. Rouse. 1997. Polychaete systematics: Past and present. *Zool. Scri.* 26:71–138.
- Felsenstein J. 1985. Phylogenies and the comparative method. *American Naturalist.* 125:1–15.
- Glasby, C. J., P. A. Hutchings, and K. Hall. 2004. Assessment of monophyly and taxon affinities within the polychaete clade Terebelliformia (Terebellida). *J. Mar. Biol. Assoc. U.K.* 84:961–971.
- Grube, A. E. 1850. Die Familien der Anneliden. *Arch. Naturgesch.*, 16:249–364.
- Hessle, C. 1917. Zur Kenntnis der terebellomorphen polychaeten. *Zoologiska Bidrag från Uppsala.* 5:39–25.
- Holthe, T. 1986. Evolution, systematic and distribution of the Polychaeta Terebellomorpha, with a catalogue of the taxa and a bibliography. *Gunneria.* 55:1–236.
- Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics.* 17:754–755.
- Hutchings, P. A. 2000. Family Ampharetidae-Trichobranchidae. In: Beesly, PL, Ross, GJ, Glasby, CJ (Eds.), *Polychaetes and Allies: The Southern Synthesis. Fauna of Australia. Vol. 4A Polychaeta, Myzostomida, Pogonophora, Echiura, Sipuncula.* CSIRO Publishing, Melbourne, 203–235.

- Jennings, R. M., and K. M. Halanych. 2005. Mitochondrial genomes of *Clymenella torquata* (Maldanidae) and *Rifta pachyprila* (Siboglinidae): Evidence for conserved gene order in Annelida. *Mol. Biol. Evol.* 22:210–222.
- Lowe, T. M., and S. R. Eddy. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25:955–964.
- Maddison, D. R., and W. P. Maddison. 2002. *MacClade 4: Analysis of Phylogeny and Character Evolution*, version 4.0. Sunderland, MA: Sinauer Associates.
- Maddison, W. P., and D. R. Maddison. 2004. *Mesquite: a modular system for evolutionary analysis*, version 1.01.
- Malmgren, A. J. 1866. Nordiska Hafs—Annulater. K. Svenska Vetenskopsakadamien. *Ofversigt of Forhandlingar*, Stockholm, 22:355–410.
- McHugh, D. 1995. Phylogenetic analysis of the Amphitritinae (Polychaeta: Terebellidae). *Zool. J. Linn. Soc.*, 114:405–429.
- Müller, O. F. 1776. *Zoologica Danica Prodrumus, seu Animalium Daniae et Norvegiae indigenarum characteres, nomina et synonyma imprimis popularium*. Copenhagen Hallageriis. xxxii. 274. pp.
- Nylander, J. A. A. 2004. *MrModeltest v2*. Program distributed by the author. Evolutionary Biology Centre, Uppsala University.
- Posada, D., and K. A. Crandall. 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics* 14:817–818.
- Rambaut, A. 1996. *The Use of Temporally Sampled DNA Sequences in Phylogenetic Analysis*. PhD Thesis. Oxford, UK: Oxford University.

- Rouse, G. W., and F. Pleijel. 2001. Polychaetes. Oxford: Oxford University Press. Pp. 235–250.
- Rouse, G. W., and K. Fauchald. 1997. Cladistics and polychaetes. *Zool. Scri.* 26:139–204.
- Rousset, V., G. W. Rouse, J. P. Féral, D. Desbruyères, and F. Pleijel. 2003. Molecular and morphological evidence of Alvinellidae relationships (Terebelliformia, Polychaeta, Annelida). *Zool. Scri.* 32:185–197.
- Rousset, V., G. W. Rouse, M. E. Siddall, A. Tillier, and F. Pleijel. 2004. The phylogenetic position of Siboglinidae (Annelida) inferred from 18S rRNA, 28S rRNA and morphological data. *Cladistics.* 20:518–533.
- Rousset, V., F. Pleijel, G. W. Rouse, C. Erséus, and M. E. Siddall. 2007. A molecular phylogeny of annelids. *Cladistics.* 23:41–63.
- Solis-Weiss, V. 1993. *Grassleia hydrothermalis*, a new genus and species of Ampharetidae (Annelida: Polychaeta) from the hydrothermal vents off the Oregon coast (U.S.A) at Gorda Ridge. *Proceedings of the Biological Society of Washington.* 106:661–665.
- Stamatakis, A. 2006. RaxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics.* 22:2688–2690.
- Struck, T. H., N. Schult, T. Kusen, E. Hickman, C. Bleidorn, D. McHugh, and K. M. Halanych. 2007. Annelida phylogeny and the origins of Sipuncula and Echiura. *BMC Evol. Biol.* 7:57.

- Swofford, D. L. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sinderland, Massachusetts.
- Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 24:4876–4882.
- Watson, A. T. 1928. Observations on the habits and life-history of *Pectinaria (Lagis) Koreni*, Mgr. *Proceedings and Transactions of the Liverpool Biological Society.* 42:25–60.
- Vallès, Y., and J. L. Boore. 2006. Lophotrochozoan mitochondrial genomes. *Int. Comp. Biol.* 46:544–557.
- Zhong M., T. H. Struck, and K. M. Halanych. 2008. Phylogenetic information from three mitochondrial genomes of Terebelliformia (Annelida) worms and duplication of the methionine tRNA. *Gene.* 416:11–21.

TABLE 1: Primers used in long-run PCR amplifications.

Fragments	Primer name	Sequence	Annealing Temp.
<i>A. crinita</i>			
<i>mLSU-cox1</i>	16S-Ac-longF	5'—GCG GTA TCC TGA CTG TGC TAA GGT AGC GTG—3'	53
	CO1-Ac-longR	5'—AGA ATG AGC AAT ATT ACT AGA TAA GGG CGG—3'	53
<i>cox1-cob</i>	CO1-Ac-longF	5'—TTA ATT CGT GTT GAG CTT GGA CAG CCA GGC—3'	53
	Cytb-Ac-longR	5'—AGA GGG GTT ACT AGT GGA TTT GCT AAT AGC C—3'	53
<i>cob-nad5</i>	Cytb-Ac-longF	5'—ATT TTC TTG TGC CTT TTA TTA TGG TAG CAC—3'	54
	Nad5-Ac-789R	5'—ATC GCA CTC TAC CAA AGC TGA AAA CGC AGC—3'	54
<i>nad5-mLSU</i>	Nad5-Ac-645F	5'—CGC TAG TTC ATT CGT CTA CAC TTG TAA CAG C—3'	--
	16S-Ac-longR	5'—TGT CCC ACA CTT ACA TTC AGG TAT TTT CAC C—3'	--
<i>P. sulfincola</i>			
<i>mLSU-cox1</i>	16S-PS-longF	5'—GTT CTA ACT GCA TAT CAA GGC AAA ACC AGC C—3'	58
	CO1-PS-longR	5'—GCC TGC GTG TGC CAT GTT TCC TGC TAG TGG—3'	58
<i>cox1-cob</i>	CO1-PS-longF1	5'—AAT CTA CCC ACC ACT AGC AGG AAA CAT GGC—3'	55
	cob-PS-longR	5'—AAT TGA TCA GTT TTC TGG TTC TCC CAG GGC—3'	55
<i>cob-nad5</i>	cob-PS-longF	5'—ACA ACC CTC TTG GTA TTA ACT CAG ACT CCG—3'	--
	Nad5-PS-longR	5'—AGT GTG GAT GAG TGG ACT AGT GCA GAG ACG—3'	--
<i>nad5-mLSU</i>	Nad5-PS-longF	5'—GAC TTC CAG CAG CAA TGG CAG CAC CTA CGC—3'	58
	16S-PS-longR	5'—ATC AGT TGT GCT TGT GTG GCT GGT TTT GCC—3'	58
<i>P. gouldii</i>			
<i>mLSU-cox1</i>	16S-PecG-longF	5'—AAA TCA TAG GAC AAG AAG ACC CCG TAG AGC—3'	58
	CO1-PecG-longR	5'—AAA AAT AGC AAG ATC CAC GGA AGG GCC TGC—3'	58
<i>cox1-cox3</i>	CO1-PecG-longF	5'—CTT AAT TCG TGT AGA ACT TGG TCA ACC AGG C—3'	52
	CO3-Ann-724R	5'—ACR TCS ACA AAR TGT CAR TAY CA—3'	52
<i>cox3-cob</i>	CO3-PG-long627F	5'—GAT TCC ACG GGC TTC ATG TTC TAA TTG GC—3'	52
	Cytb 876R	5'—GCR TAW GCR AAW ARR AAR TAY CAY—3'	52
<i>cob-nad5</i>	Cob-PG-midF	5'—TTG CTT CGA AAC CTT CAT GCT AAC GGA GC—3'	--
	Nad5-PG-675R	5'—AAT CGG ACT AAA AGA AAC ACT CCT GCC G—3'	--
<i>nad5-nad4</i>	Nad5-PG-453F	5'—TTA TAC TTC TTT TAT CTA TTG GAT GAG CCC—3'	53
	Nad4-PecG-longR	5'—ATT GGG AGG AGT TAA AAG AGT AAA GGA TTG C—3'	53
<i>nad4-mLSU</i>	Nad4-PecG-longF	5'—TTT AAC TCC TCC CAA TCC CTA TCT TTT AGC—3'	58
	16S-PecG-longR	5'—GTT TAG GTT AGG CGG GAT GCC TTA TTG CTC—3'	58

Table 2: Primers used for sequencing the mitochondrial genomes of *Auchenoplax crinita*, *Paravinella sulfincola* and *Pectinaria gouldii*.

Fragments	Primer name	Sequence	
<i>A. crinita</i>			
<i>mLSU -cox1</i>	CO1-Auc1-startR	5'—GTC GCA TAT AAC CAA CGC—3'	
	16S-Auc1-560F	5'—AAT CCT ACA TGA GCT GAG—3'	
	16S-Auc1-midF	5'—AAT CCT ACA TGA GCT GAG—3'	
	Nad2-Auc1-705R	5'—CTA CTT GTC TAG ATC TTC C—3'	
	Nd2-Auc-189R	5'—CAA CTC AAA TAG CTA GCC—3'	
	Nd2-Auc-508R	5'—TAA TAC TAC TTG TCT AGA TC—3'	
	Nad1-Auc-37F	5'—TTT ATG CGC TAT ACT AGC—3'	
	Nad3-Auc1-startR	5'—AGA AAT AAC CAT TAA CGC C—3'	
	Nad1-Auc1-561F	5'—TGT TTC GGT TTT AGC TGA G—3'	
	Nad1-Auc1-521R	5'—GAT AAT CAA ACT TGC CG—3'	
	Nad2-Auc-599R	5'—AAT TAT TGA CCC TGC TAC—3'	
	Nad2-Auc-588F	5'—AAT TTC TAG TGT TAG AGC—3'	
	Nad1-Auc-startF	5'—TGG CAG ACT AGT GCG TTG G—3'	
	Nad2-A1-5R	5'—AGG TTT TCC ATG TTA ATG C—3'	
	Nd1-A1-500F	5'—TGT TTC GGT TTT AGC TGA G—3'	
	<i>cox1-cob</i>	CO1-Auc1-591F	5'—TTC GTT ACC AGT ACT GGC—3'
		Cytb-Auc1-417R	5'—CCA TAA TAA ATT CCA CGC—3'
		Cob-Auc1-mid2R	5'—ACA TAC ACG CTG AGT AAC—3'
		CO3-Auc1-middR	5'—CTA CTG GCG GTC ATA CAC—3'
Nad6-Auc1-156R		5'—TTC ATG ACG TTG ATT AGG—3'	
CO1-Auc1-1004F		5'—AAA ATA GTT TAT GAT CCT GC—3'	
Atp8-Auc1-startR1		5'—AAT AGG AGA CAA ATG AGG C—3'	
CO1-Auc-911R		5'—TTG ACA AAA TAA TCC CAG—3'	
Nad6-Auc-startR		5'—AAT CAA AAC CCT ATG TGC—3'	
Nad6-Auc-45F		5'—TAG TTT TAG TAT TAC TTG G—3'	
Atp8-A1-endR		5'—TAA TAG GAG ACA AAT GAG GC—3'	
CO3-A1-startR		5'—TAC TGT CAC ACC AGA AGC—3'	
<i>cob-nad5</i>		Cob-Auc1-942F	5'—AGC AGA AGC GTA GAA GAC —3'

Fragments	Primer name	Sequence
	Nd5-Auc1-436R	5'—AAC TAT GAT TAT CCC AGC—3'
	Atp6-Auc1-70F	5'—AGT TTC TGT AAT AGG AGT G—3'
	Cob-Auc1-endsR	5'—AAA GTA CAC ATA AAG CTC C—3'
	Nad4-Auc-812R	5'—ACA AAC TTC CAT GTT GAG—3'
	Nad5-Auc-middR	5'—GAA GTT ATT ATC CCA GGC—3'
	Nad5-Auc-middF	5'—TTA ACT TAT ATG GCG TAT G—3'
	Nad4-Auc1-672R	5'—TAG ACC CAG CAA CAG GAG C—3'
	tRNAX-Auc-befNd5R	5'—AAC ATC GAA ACC ATG GGC—3'
	Atp6-Auc-220F	5'—TAT GTT TAC TCA GTC ACG—3'
	Atp6-Auc-235R	5'—ACG CTG ATA AAA CAC TTC—3'
	Nad4-A1-312R	5'—CTA AAA TTA TAG ACC CAG C—3'
	Atp6-A1-530F	5'—TAA GAG CAG GGC ATA TTG CG—3'
<i>P. sulfincola</i>		
<i>mLSU-cox1</i>	16S-PS-midF	5'—CTA CCT GCT ACA GTT CTC C —3'
	16S-PS-endF	5'—AAA ATA TGC CCT ACT AGG CTC—3'
	tRNAX-PS-R	5'—TTA TGA AGA CGA CTT TGA GG—3'
	Nad2-PS-235R	5'—GGA AAC CAT TGA TGA CAC GG—3'
	Nad1-PS-endF	5'—TGA ACG GAT AGC TCT GAT GC—3'
	Nad3-PS-211R	5'—TGT GGC TGA GTG CTT AGA CG—3'
	Nad1-PS-308R	5'—GTT GAT ATT GGA TAA TTG GC—3'
<i>cox1-cob</i>	CO1-PS-834F	5'—AGA CAC ACG AGC CTA CTT TAC CGC —3'
	Cob-PS-202R	5'—TGA GAA GGC TAG GTC TAC ATT TGG—3'
	CO1-PS-1500F	5'—CAC CAG CAT TCC ACT CAG GAG CCG—3'
	Nad6-PS-108R	5'—GTG TTA GAG TCG ATA GTC TGC—3'
	CO2-PS-497F	5'—AGC AGA CGT AAT TCA TTC ATG AGC C—3'
	CO3-PS-333R	5'—TTA AGA GTG GTA CTG CAA ATG GGT C—3'
<i>nad5-mLSU</i>	12S-PS-endR	5'—ATG CAG TGA CAT GGT GGC TTG CTG CGG —3'
	Nad5-PS-1160F	5'—GTG CCG CAG CTG CCT ACT CAA CAC G —3'
	12S-PS-startR	5'—TTC AGT GTA AGT GAG TGG CAT CAC C—3'
	tRNAX-PS-R2	5'—TCT ATT TGG ACA TTT CGT TAA ACC G—3'
	Nad4-PS-200F	5'—ATT CAT CAT TGT AGT CTC TTT CGC—3'
	Nad4-PS-531F	5'—CGT TGC CGC ATC TTT ACC GC—3'

Fragments	Primer name	Sequence
	Nad4-PS-1176R	5'—ATA AGA TTA AAG AGT ATG TAC C—3'
<i>P. gouldii</i>		
<i>mLSU -cox1</i>	16S-PG-720F	5'—GCC CAG CTA ATT GGC AGA CTA GTG C —3'
	CO1-PG-startR	5'—GAG ATA ACG TGG CAC AAA CCA AAG C—3'
	Nad1-PG-261F	5'—TCT GCG CTT TAA CCT TAG CCC TGC—3'
	Nad2-PG-495R	5'—CCC GTA TTT GGG TTT GAT TTA GC—3'
	Nad1-PG-656F	5'—ATT GAA TAT AGA AGT GGC AGC—3'
	Nad2-PG-2R	5'—TTT ATA GGG GCA GTC ATA GG—3'
	Nad2-PG-321F	5'—GTA CAA TTT CAA TCA ACC AG—3'
<i>cox1-cox3</i>	CO1-PG-900R	5'—CAT GGT TGC GGC AGT AAA ATA GGC —3'
	CO3-PG-204R	5'—CCT AAA AAA GTT CCC TCA CG —3'
	CO2-PG-591R	5'—GCA TTG ACC ATA GAA TAC TCC G—3'
	CO2-PG-startR	5'—TCT TGG AAG GAT AGT TGG GCT C—3'
	CO1-PG-1257F	5'—TTC TTC CCT CAA CAC TTC CTC GG—3'
	CO1-PG-1134R	5'—AGA GGA AAC CAG TGG TTA AAT GC—3'
<i>cox3-cob</i>	Cob-PG-midF	5'—TTG CTT CGA AAC CTT CAT GCT AAC GGA GC —
	Nad6-PG-273F	5'—CTA ATT GCA CTC CTT TTG TCA GGC—3'
	CO3-PG-L627F	5'—GAT TCC ACG GGC TTC ATG TTC TAA TTG GC —3'
<i>nad5-nad4</i>	Nad4-PG-186F	5'—TCT TAA CCT TAT GAA TTT CCA GCC—3'
	Nad4-PG-516R	5'—TTT ATG AGG ATC GAG TAA CTT GC—3'
	Nad4L-PG-75F	5'—ATG GCG ACC CTG ATT ATT CAA CG—3'
	Nd5-PG-1005R	5'—GTT TGC AAT TAA AAG GCA GGA C—3'
	Nd5-PG-1065F	5'—GCC GGG TTC TAC TCT AAA GAC C—3'
	Nd5-PG-1587F	5'—CAA GGT TGA AAC GAA ATA GCC GG—3'
	Nad4-PG-453R	5'—CTA TAA GGA GTG GTA GTG AAG C—3'
<i>nad4-mLSU</i>	Nad4-PG-846F	5'—CTT ATT GCT TAT TCC TCA GTT GGT C —3'
	Nad4-PG-1026F	5'—GGG TTA CTC ATT AAT ACT TTC CAG—3'
	12S-PG-startR	5'—AAT TTT TCA ATA GTT TAC TCA TGG AG —3'
	tRNAX-PG-F	5'—CAT GAC CCA AAA GTG GAA ACA ATC C—3'
	12S-PG-111F	5'—TGT CCT CTA ACT CGA TAA TCC ACG—3'
	12S-PG-midR	5'—ATT CCT GAT GCT ATG CTT TGT GGC—3'
	Nad4-PG-519R	5'—TGG CTG GAA AGT ATA TGA GTA ACC—3'

Table 3: Taxa used in the phylogenetic analysis.

Species	Clade	Nucleotides	GenBank Number
<i>Auchenoplax crinita</i>	Annelida, "Canalipalata", Ampharetidae	13,759 partial	FJ976041
<i>Pectinaria gouldii</i>	Annelida, "Canalipalata", Pectinariidae	13,438 partial	FJ976040
<i>Paralvinella sulfincola</i>	Annelida, "Canalipalata", Alvinellidae	13,640 partial	FJ976042
<i>Terebellides stroemi</i>	Annelida, "Canalipalata", Trichobanchidae	15,755 complete	EU236701
<i>Pista cristata</i>	Annelida, "Canalipalata", Terebellidae	15,894 complete	EU239688
<i>Eclysippe vanelli</i>	Annelida, "Canalipalata", Ampharetidae	13,749 partial	EU239687
<i>Clymenella torquata</i>	Annelida, "Scolecida", Maldanidae	15,538 complete	AY741661
<i>Riftia pachyptila</i>	Annelida, "Canalipalata", Siboglinidae	12,016 partial	AY741662
<i>Galathealinum brachiosum</i>	Annelida, "Canalipalata", Siboglinidae	7,576 partial	AF178679
<i>Platynereis dumerilii</i>	Annelida, "Aciculata", Nereididae	15,619 complete	NC_000931
<i>Lumbricus terrestris</i>	Annelida, "Oligochaeta", Lumbricidae	14,998 complete	NC_001673
<i>Helobdella robusta</i>	Annelida, Hirudinea, Glossiphoniidae	7,553 partial	AF178680
<i>Orbinia latreillii</i>	Annelida, "Scolecida", Orbiniidae	15,558 complete	AY961084
<i>Scoloplos cf. armiger</i>	Annelida, "Scolecida", Orbiniidae	12,042 partial	DQ517436
<i>Phascolopsis gouldii</i>	Annelida, Sipuncula	7,470 partial	AF374337
<i>Urechis caupo</i>	Annelida, Echiura	15,113 complete	AY619711
<i>Nephtys sp.</i>	Annelida, "Aciculata", Nephtyidae	17,217 complete	EU293739

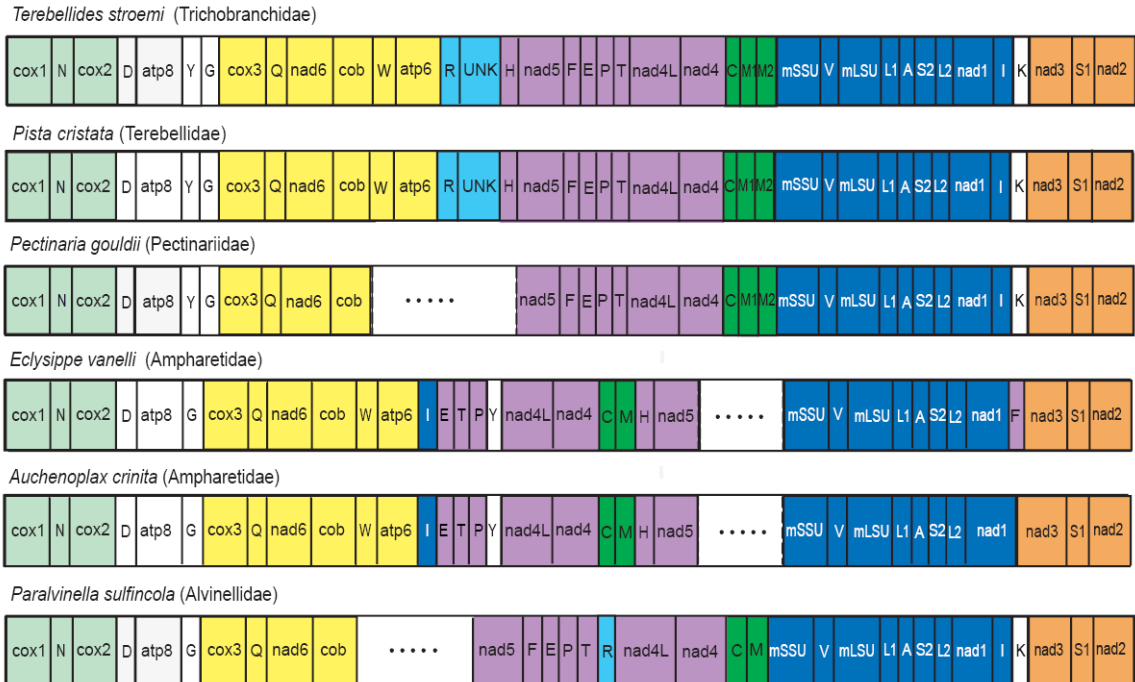


Figure 1: Mitochondrial gene order of six Terebelliformia worms. Different colors shows conserved gene clusters. Dots indicate missing regions.

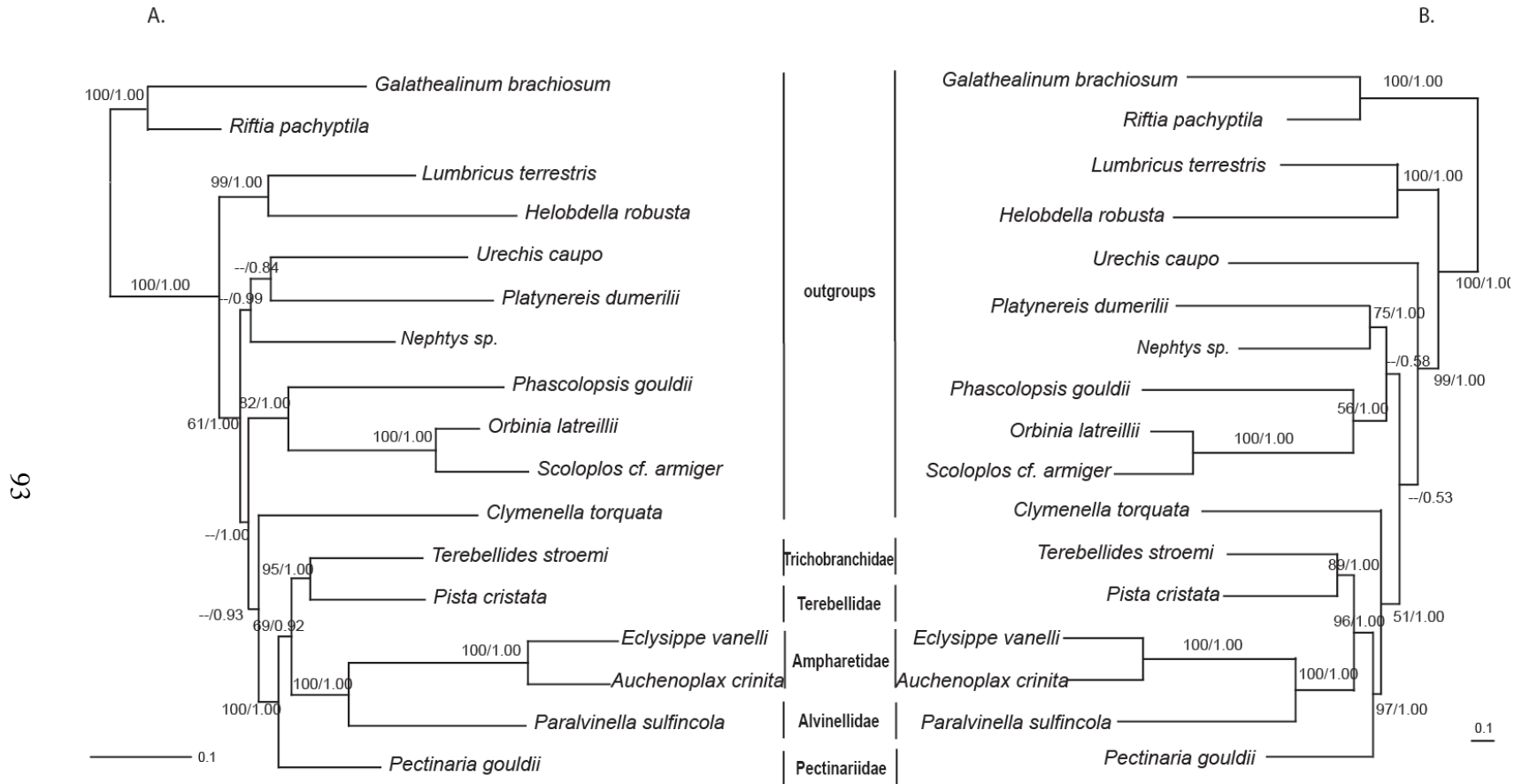


Figure 2: Phylogenetic reconstructions for 17-taxon mitochondrial datasets. (A) The mitochondrial nucleotide tree represents the identical topology from both ML and partitioned Bayesian inference methods. (B) The mitochondrial amino acid tree with the identical topology from both ML and partitioned Bayesian inference methods. Nodal support values are given at branches with posterior probabilities of the partitioned Bayesian analysis first and ML bootstrap values second. A dash indicates < 50% on trees. The representative branch lengths are from the Bayesian trees.

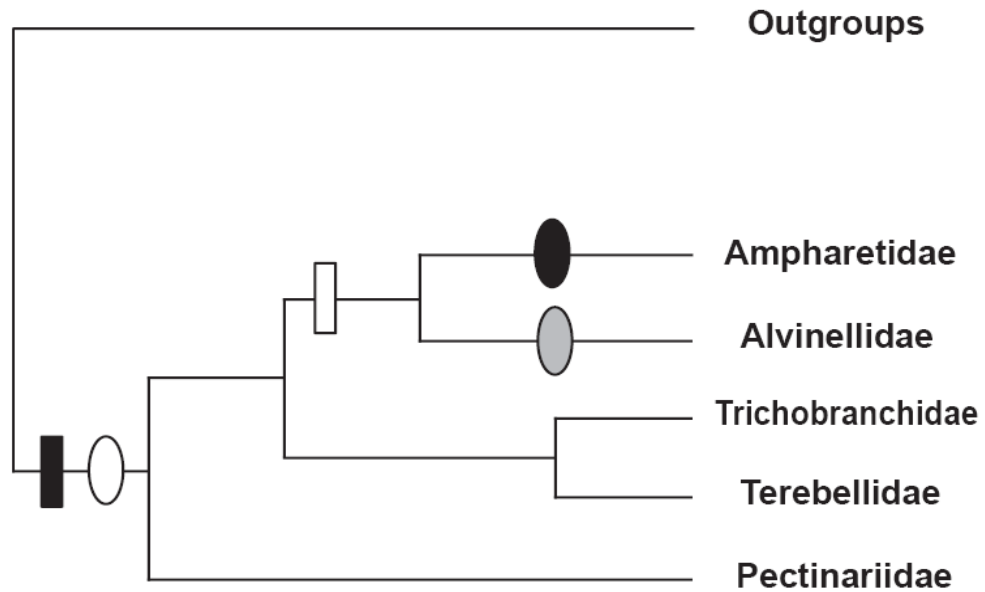


Figure 3: Character mapping on Terebelliformia ingroup trees based on mitochondrial datasets. The bar represents *trnM* duplication event. The circle represents the specific gene order of *nad5-nad4L-nad4* cluster in mitochondrial genomes. Open bar: duplication absent; Black bar: duplication present; open ellipses: *nad5-trnF-trnE-trnP-trnT-nad4L-nad4*; black ellipses: *trnE-trnT-trnP-trnY--nad4L-nad4-trnC-trnM-trnH-nad5*; shaded ellipse: *nad5-trnF-trnE-trnP-trnT-trnR-nad4L-nad4*.

CHAPTER 4

TWO GROUP II INTRONS AND THEIR EVOLUTIONARY ORIGINS IN *ENDOMYZOSTOMA* (MYZOSTOMIDA) MITOCHONDRIAL GENOME

Abstract

Group II introns are self-splicing ribozymes or retroelements thought to be evolutionary precursors to spliceosomal introns. They consist of a distinctive secondary structure and have been described from eukaryotic organelles and bacteria. Interestingly, they are lacking from genomes of bilaterian animals, with the exception of a species annelid worm, *Nephtys* sp. Here, we report two group II introns identified in the mitochondrial genome of the annelid-like myzostomid worm, *Endomyzostoma*. Similar to the *Nephtys* intron, the myzostomid introns, designated as Mintron1 and Mintron2, are found in the *coxI* gene. Mintron1 has a 1424 bp ORF in the loop area of domain 4 which includes only the RT and X domains that are assumed to assist the self-splicing efficiency *in vivo*. Based on both the ORF nucleotide sequence and RNA structure, we hypothesize Mintron1 is a divergent group II intron. In contrast Mintron2 is a short (384 bp) degenerate intron that lacks an ORF which cannot perform retrohoming mobility, but can perform RNA-mediated splicing *in vitro*. Mintron1 lacks of both D and En domains in the ORF which play a key role in retrohoming process, but it may still maintain its capability for retrohoming by insertion into replication forks. Both introns are in the mitochondrial class of group II introns as suggested by RNA folding structure (for Mintron2) and phylogenetic analysis (for Mintron1). They likely are products of independent horizontal transfer events based on the close relationship to *Nephtys*' intron suggested by the ORF phylogeny.

1. Introduction

Group II introns are self-splicing ribozymes that are found in the genomes of bacteria as well as organelle genomes of fungi, plants and animals (Beagley et al. 1998; Lehmann and Schmidt 2003; Rot et al. 2006; Dellaporta et al. 2006; Vallès et al. 2008) and are hypothesized to be the ancestor of eukaryotic spliceosomal introns (e.g., Cavalier-Smith, 1991). Secondary structure of a typical group II intron consists of 6 stem-loop domains, D1 to D6, with all domains radiating from a central core to form the proximal helix that facilitates self-splicing. Several conserved nucleotide sequences are diagnostic of group II introns including 5' and 3' splicing sites (5'-GUGYG and AY-3'), an unpaired adenosine in D6 and a number of conserved nucleotides in D5 (Knoop et al. 1994; Lambowitz and Zimmerly 2004; Lang et al. 2007).

Group II introns are mobile and generally have two forms, a larger ORF-containing form and an ORF-less form that only consists of the six domains. ORF-less introns primarily occur in organelles rather than bacteria (Dai and Zimmerly 2002). All ORFs in group II introns are found in D4 and involve four functional domains: reverse transcriptases (RTs) domains, a X domain with maturase activities, a non-conserved D domain (for DNA-binding) and an En domain with endonuclease activity (Lambowitz et al. 1999; Zimmerly et al. 2001; San Filippo and Lambowitz 2002; Dai and Zimmerly 2002). ORF proteins mainly have two functions: 1) to facilitate the RNA-mediated intron splicing process through both RT and X activities, and 2) to act as an element for

retrohoming which requires all four domains cooperating with intron RNA (reviewed by Robart and Zimmerly 2005). Intron retrohoming is a common mobile phenomenon in group II introns by inserting into specific target sites with the assistance of ORF proteins that widely distributed in protists and eukaryotes. But introns lacking an ORF will only perform the RNA-mediated splicing activity *in vitro* with the requirement of extreme unphysiological conditions (high concentrations of magnesium and high temperature) and cannot home (Michel and Ferat 1995). But *in vivo*, the splicing activity relies on the aid of proteins coded by the ORF (Michel and Ferat 1995; Bonen and Vogel 2001).

Retrohoming usually occurs at very specific insertion site sequences, but insertions into non-specific site are also known (Cousineau et al. 2000). ORFs are associated with their specific intron RNA structures allowing their sequences to be for categorizing and identifying group II introns (Zimmerly et al. 2001; Toor et al. 2001; Dai and Zimmerly 2002). Phylogenetic analyses of the various ORFs can divide group II introns into several major classes: six bacterial classes (A, B, C, D, E, F), one mitochondrial class (with uniform group IIA1 RNA structures) and two chloroplast classes (IIB1 and IIB2 in structures) (Toor et al. 2001; Zimmerly et al. 2001; Toro et al. 2002; Simon et al. 2008). Bacterial introns are considered basal to mitochondrial and chloroplast forms, and all currently known group II introns have been hypothesized to descend from bacterial ORF-containing group II introns (Toor et al. 2001).

Eukaryotic mitochondrial genomes, which can harbor Group II introns, have variable gene compliments across organismal lineages. In contrast, bilaterian mitochondrial genomes are remarkably conserved with a typical size of 15-17kb and a 37 gene compliment consisting of 13 protein-coding genes, two ribosomal genes, 22 tRNA

genes and one non-coding region referred to as the UNK or D-loop (Boore 1999; Vallès and Boore 2006; Zhong et al. 2008). To date, there have been over one thousand bilaterian mitochondrial genomes published. Of these, only the mitochondrial genome of the annelid *Nephtys* sp. possesses a group II intron which was the first report of group II intron in animal genomes (Vallès et al. 2008).

Myzostomida are small parasitic marine worms, living mostly on as ectocommensals or endoparasites of echinoderms (Grygier 2000). Their body is typically soft and flattened with many radiating cirri on the thin rounded body edge and five pairs of parapodia on the ventral side. With limited exception (Lanterbecq et al. 2006), all have high host-specificity, as a myzostomid species usually associates with a single crinoid species (Eeckhaut 1998). Fossils from the Ordovician suggest an ancient association between myzostomids and their crinoid hosts (Eeckhaut 1998), which may explain the highly derived body plan of myzostomids and their disputable phylogenetic position within metazoans (Zrzavý et al. 2001; Bleidorn et al. 2007). Although some studies question their placement with annelids (e.g., Eeckhaut et al. 2000; Zrzavý et al. 2001), their affinity with the group is indicated by morphological data (e.g., Haszprunar 1996; Rouse and Fauchald 1997) and a recent molecular study (Bleidorn et al. 2007) provides strong additional support.

While determining the mitochondrial genome of a marine worm of *Endomyzostoma* sp. (Myzostomida), we discovered two group II introns. Interestingly, the mtDNA genome of *Myzostoma seymourcollegiorum* (Bleidorn et al. 2007) does not have introns. Below we describe these introns and discuss implications for both bilaterian mitochondrial genome and group II intron evolution. This is the first report of multiple

introns in bilaterian mitochondrial genomes. The association between two introns as well as their evolutionary origins will be the focus in our study.

2. Materials and Methods

2.1 Sample collection and DNA extraction

Endomyzostoma sp. was collected in the Bransfield Straight region of the Antarctic Peninsula (S 63°40.145', W 61°10.047') by Blake trawl using the *R/V Lawrence M. Gould*. Tissue was initially frozen at -80°C then dissected in ethanol after DNA collections. Total genomic DNA extractions employed the DNeasy® Tissue Kit (Qiagen) according to the manufacture's instructions.

2.2 MtDNA data collection

The mitochondrial genome of *Endomyzostoma* sp. was amplified in four overlapping fragments. Taxonomically-inclusive primers (Zhong et al. 2008) were first used to amplify conserved regions of *mLSU*, *cox1*, *cob* and *nad5* genes. PCR products were then purified using QIAquick PCR purification kit (Qiagen) and sequenced using a Beckman CEQ8000. Four pairs of specific long-PCR primers (table 1) were subsequently designed to amplify fragments spanning *mLSU-cox1*, *cox1-cob*, *cob-nad5* and *nad5-mLSU*. Long PCRs were employed on an Eppendorf Mastercycler (Eppendorf) using Takara LA-Taq PCR System. 50 µl long PCR reactions were set up including 5µl 10×buffer, 8µl dNTP (2mM), 5µl MgCl₂ (25mM), 2µl of each long PCR specific primers (10µM each), 0.5µl Takara LA-Taq (5U/µl), 2µl DNA template and 25.5µl sterilized distilled water. The long PCR protocol was 94 °C for 3min, followed by 35 cycles with

94°C for 30 sec, 53 or 54°C for 30s (annealing temperatures for different pairs of primers are in table 1), and 70°C for 12min; final extension at 72°C for 10 min and hold at 4°C. The *cox1-cob* fragment was approximately 8 kb, *cob-nad5* was 2 kb, and *nad5-mLSU* was about 4.5kb in size. These three fragments were purified using QiaQuick Gel Extraction Kit (Qiagen) and then cloned into the pGEM-T Easy vector (Promega). Positive clones were screened by PCRs and plasmids were isolated by QIAprep[®] Spin Miniprep Kit (Qiagen). Size of the inserted plasmid was confirmed by *EcoRI* digestion. Primer walking was employed to sequence inserts of plasmids. All sequencing primers were listed in the table 2.

The *mLSU-cox1* fragment repeatedly proved difficult to be amplified probably due to the effects of the potential UNK region that it contained (see Boore and Brown 2000; Jennings and Halanych 2005). Thus, we were not able to obtain the complete sequence of the mtDNA genome. Presumably, *nad1*, *nad3*, *nad2* and several tRNA genes are located in this currently uncharacterized region (see below). The GenBank accession number for the partial mitochondrial genome is FJ975144. Gene nomenclature and abbreviations used here follow Jennings and Halanych (2005).

2.3 Genomic Assembly

Sequences were assembled and edited using SeqMan and MegAlign (Burland 2000) in the DNASTAR[™] Lasergene suite. BLAST (Altschul et al. 1990) was used to identify protein-coding genes and ribosomal RNA genes. Boundaries of tRNA sequences were inferred by identifying flanking protein-coding and rRNA genes. The tRNA genes were identified using the tRNAscan-SE web server (<http://lowelab.ucsc.edu/tRNAscan->

SE/; Lowe and Eddy 1997) under default settings with the sequence source set to “mito/chloroplast”, or drawn by hand based on their potential secondary structures and anticodon sequences.

2.4 Intron identification

Secondary structures of introns were examined with the “mfold” online server (Zuker 2003) and refined by eye based on comparisons with the typical structures of group II introns (Toor et al. 2001). To confirm introns existed in the mitochondrial genome of *Endomyzostoma* sp., two pairs of primers spanning the intron-exon boundary were employed for PCR amplifications. For intron1 (here after called Mintron1), Intron-Myz-635R was designed to anneal within the intron1 sequence and paired with the *coxI* primer LCO1490 (Folmer et al. 1994). A 1.4 kb fragment was obtained using the Eppendorf High Fidelity PCR system and genomic DNA as template. A similar test was performed for the second intron (here after Mintron2) by using the intron specific primer, CO1-Myz-Intron2F and a *coxI* gene degenerated primer, CO1-Ann-1300R from the 3' region of *coxI*. In this case, a 500 bp fragment was generated. Sequencing results of these two fragments confirm sequences obtained from clone fragments, thus verifying that Mintron1 & 2 are not recombinant products or PCR artifacts. All the primers used for intron identification are shown in table 3.

2.5 Phylogenetic analyses and sequence alignments

For phylogenetic analyses, 105 ORF amino acid sequences for group II introns were collected from bacteria, mitochondrion and chloroplast based on the organism list in the Zimmerly database (<http://www.fp.ucalgary.ca/group2introns/>; Dai et al. 2003). The

dataset included all ORFs from published mitochondrial and chloroplast group II introns as well as 46 bacterial group II introns (table 4). An Archaeobacteria ORF (M.a.II-1) was used as the outgroup to polarize the resultant tree. Sequences were aligned by MAFFT version 6 (Kato and Toh 2008) and examined by Gblocks 0.91b (Castresana 2000). MacClade 4.08 (Maddison and Maddison 2002) was used to exclude ambiguously aligned regions. The employed alignment is available in TreeBASE with the accession number XXXXXX (www.TreeBASE.org). Bayesian analyses of amino acid sequences were performed in MrBayes version 3.1.2 (Huelsenbeck and Ronquist 2001) with the mixed amino acid substitution model. Two sets of four chains were run simultaneously for 2,000,000 generations sampled every 100 generations. Burnin (2,500 generations) was determined by assessing stationarity in $-\ln$ likelihood values.

3. Results

3.1 Genome Composition and Evolution

The partial mitochondrial genome of *Endomyzostoma* sp. is 13,190 bp in length and contains 26 genes including 10 protein-coding genes, two rRNA genes and 14 tRNAs. The region from *mLSU* to *cox1* gene was not recovered. All recovered genes are transcribed from the same strand. The tRNA secondary structures are presented in figure 1. The gene orders of *Endomyzostoma* sp. and *Myzostoma seymourcollegiorum* mtDNA have the same protein-coding and rRNA gene order (Bleidorn et al. 2007; Bleidorn et al. revised). tRNAs have relocated in the fragment from *atp6* to *mSSU*, the region Zhong et al. (2008) noticed was less conserved. The partial mitochondrial gene arrangements in two myzostomids share the same protein-coding gene arrangement pattern as all published

annelids other than the sipunculid and echiurid (Vallès and Boore 2006; Zhong et al. 2008; Bleidorn et al. revised).

3.2 Intron structures

Two group II introns, Mintron 1 and Mintron 2, were identified in the *cox1* gene by their potential secondary structures and ORF sequences (fig. 2A). The total length of Mintron1 is 1424bp, with an ORF 1056bp in size whereas Mintron2 is only 384bp in length and lacks an ORF. The secondary structures contain six helical domains radiating from a central core. Conserved GUGYG and AY nucleotides at 5' and 3' end, respectively, were identified in Mintron1 (fig. 2B). However, one nucleotide difference was located at the 5' end (GCGCG) of the Mintron2 sequences (fig. 2C). Two Exon Binding Sequences (EBSs) were located in D1 for both Mintron1 and Mintron2. These sites can bind specifically with intron binding sequences (IBSs) in 5' end exon sequences during splicing (fig. 2B and C) (Lambowitz et al. 1999; Bonen and Vogel 2001).

3.3 ORF alignment and phylogenetic analyses

Group II intron ORFs typically have several domains with a few conserved regions (fig. 3A) (Zimmerly et al. 2001; San Filippo and Lambowitz 2002; Robart and Zimmerly 2005). The alignment of 105 ORF amino acid sequences revealed that both Mintron1 and the *Nephtys* intron have the RT and X domains while lacking both D and En domains. The amino acid sequences of both ORFs were highly divergent. From the alignment, the Mintron1 ORF shows sub-domain 2 and 4 of the RT domain and the X domain are the most conserved sub-domains present (fig. 3B).

The ORF phylogenetic analysis revealed that most group II introns found in mitochondria fall into one main clade (fig. 4). Mintron1 is inferred to be sister to the *Nephtys*' although nodal support as judged by posterior probabilities is weak (fig. 4).

4. Discussion

Animal mitochondrial genomes usually lack any introns, with the exception of some Cnidaria and the Placozoa (Lavrov 2007). It has been noticed that bilaterian mitochondrial genomes provide a population genetic environment which encourages elimination of mutationally disadvantageous noncoding DNA (Lynch 2007). Discovery of two group II introns within the mitochondrial genome of the myzostomid, *Endomyzostoma* sp., along with their previous discovery in *Nephtys* (Vallès et al. 2008), contrast sharply with the otherwise highly conserved and intron-free mitochondrial genomes found in the bilaterian. This suggests that some animal lineages may be more susceptible to retroelement “attack”. Additionally, one of the most basal groups of metazoans, Placozoa, also contains group II introns (Dellaporta et al. 2008), as do some other annelids (KMH unpublished data). To date, all of these metazoan mitochondrial group II introns have been discovered in the *cox1* gene. When aligned, the insertion site of the *Nephtys* intron and Mintron1 are only 24 bp apart, and Mintron 2 is further downstream in the 3' direction. Presumably, the sequence motif recognized by these group II introns for insertion is found in higher prevalence in this region of the *cox1* gene than other mtDNA genes. In view of the conserved regions of host *cox1* genes for both Mintron1 and Mintron2, our result is consistent with the previous prediction that

organellar group II introns often locate in conserved housekeeping genes (Dai and Zimmerly 2002; Robart and Zimmerly 2005).

4.1 Mintron1: a divergent mitochondrial group II intron

Although Mintron1 is an ORF-containing intron, the amino acid sequence of the ORF is considerably different from other investigated taxa. Only a limited number of amino acid residues found in the RT 0-7 subdomains and one region in X domain are conserved relative to other group II ORFs. Given that both RT (for reverse transcriptase activity) and X (maturase activity for splicing) domains are required for mobility *in vivo* (Matsuura et al. 1997; Wank et al. 1999; Bonen and Vogel 2001; Cui et al. 2004;), Mintron1 can presumably accomplish reverse splicing in the genome. Generally, only single-stranded DNA will be reverse transcribed as the lacking D (C-terminal DNA-binding domain) and En (endonuclease activity) domains interfere with the recognition of the second-strand DNA cleavage (Guo et al. 1997; Lambowitz and Zimmerly 2004). Mintron1, however, probably does not totally lose mobility since they may insert into the replication fork to complete the double strand DNA replication for retrotransposition without the second-strand cleavage (Ichiyanaqi et al. 2002; Zhong and Lambowitz 2003).

Secondary structure of Mintron1 is typical of group II introns with six domains and one ORF in the D4 domain. Comparing secondary structures in each sub-class, we don't find any similarity in Mintron1 to any previously identified category (Toor et al. 2001). However, the ORF phylogenetic tree suggests it belongs to the mitochondrial class with close relationship to the intron found in *Nephtys* (fig. 4). Therefore, we infer

Mintron1 to be a divergent mitochondrial group II intron. Alternatively, Mintron1 could fold in other configurations.

4.2 Mintron2: undergoing evolutionary degeneration

Compared to the typical mitochondrial class group IIA1 structure (Toor et al. 2001), the secondary structure of Mintron2 has several conserved nucleotides in the central core. These are key A1 features, for instance, CGGA between D2 and D3, only one A locating between D3 and D4, and GGA between D4 and D5 (fig. 2C). Mintron2 has apparently lost its mobility and is speculated to be undergoing degeneration as it lacks any ORFs and has one substitution in the critical 5' region. Nevertheless, in order for the *coxI* gene that harbors the insertion to be functional, Mintron2 must be spliced out prior to translation (Dai and Zimmerly 2002; Robart and Zimmerly 2005). Thus, because Mintron2 lacks the ORF encoded proteins, it can't perform splicing *in vivo* only by the RNA structure and presumably needs host accessory proteins to increase splicing efficiency allowing the *coxI* gene, which is critical to oxidative phosphorylation, to function (Jenkins and Barkan 2001; Dai and Zimmerly 2002). Alternatively, Mintron 2 could use the reverse transcriptase and maturase activities encoded by Mintron1's ORF for splicing.

The presence of substitutions at the 5' end of Mintron2 challenges our knowledge of the actual mechanism of self splicing. The 5' splice sites play a critical role in initiating group II intron self-splicing events by interaction with a bulged adenosine in D6 domain to form a lariat (Bonen and Vogel 2001). Whereas such substitutions would be expected to cause lose of splicing function, the 5' sites retain splicing proficiency

suggesting the mechanism is not as conserved as previously hypothesized (Michel and Ferat 1995; Bonen and Vogel 2001). Such substitutional changes are consistent with the “Retroelement Ancestor Hypothesis” which posits that most ORF-less introns are derived from ORF-containing elements (Toor et al. 2001; Robart and Zimmerly 2005).

4.3 Intron evolutionary origins

Phylogenetic analysis shows the ORF-containing Mintron1 belongs to the main mitochondrial class (fig. 4). The similarity of secondary structure of the Mintron2 to the group IIA1 class also suggests its mitochondrial origin. Although Mintron2 has lost retrohoming capability as a transposable element, it can presumably still perform splicing *in vivo* under aid of recruited proteins and despite substitutions at the 5' end. Toor et al.'s (2001) retroelement ancestor hypothesis posits that major RNA structural forms of group II introns coevolve with the intron-encoded proteins leading to most ORF-less introns evolving from ORF-containing ones (see also Hausner et al. 2005). However, the difference in secondary structure of Mintron 1 and 2 seems to suggest the two introns originated from different lineages.

Interestingly, Mintron 1 is recovered as sister to the *Nephtys* intron, but based on the ORF alignment dataset, limited similarity was found among Mintron1, the *Nephtys* intron and *Trichoplax* ORF domains, suggesting that an independent origin of introns in these three metazoans. The close phylogenetic placement of Mintron1 to the *Nephtys*' intron, however, may indicate a common ancestor that separately infected both organelles by a horizontal transfer (as discussed in Vallès et al. 2008). In these organisms, introns occur in roughly the same 3' region of *cox1* perhaps indicating a common insertion site

by the group II introns. The discovery of multiple groups II introns restricted to one lineage of bilaterians, annelids, implicates the bigger potential to accept retroelements in its mitochondrial genomes and may provide insight into their origins and functional contributions to the hosts.

Acknowledgements

We thank T. Struck, Y. Vallés, and J. Boore for constructive comments with the mtDNA genome work, S. Zimmerly for the helpful suggestions with the group II intron analyses. This work was made possible with support from the National Science Foundation (EAR-0120646) to KMH. This is contribution #XX to the AU Marine Biology Program.

Literature Cited

- Altschul SF, Gish W, Miller W, Myers EM, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol.* 215:403–410.
- Beagley CT, Okimoto R, Wolstenholme DR. 1998. The mitochondrial genome of the sea anemone *Metridium senile* (Cnidaria): introns, a paucity of tRNA genes, and a near-standard genetic code. *Genetics.* 148:1091–1108.
- Bleidorn C, Eeckhaut I, Podsiadlowski L, Schult N, McHugh D, Halanych KM, Milinkovitch MC, Tiedemann R. 2007. Mitochondrial genome and nuclear sequence data support Myzostomida as part of the annelid radiation. *Mol Biol Evol.* 24(8):1690–1701.

- Bleidorn C, Podsiadlowski L, Zhong M, Eeckhaut I, Hartmann S, Halanych KM, Tiedemann R. Revised. On the phylogenetic position of Myzostomida: can 77 genes get it wrong? *BMC Evol Biol*.
- Bonen L, Vogel J. 2001. The ins and outs of group II introns. *Trends Genet*. 17: 322–331.
- Boore JL. 1999. Animal mitochondrial genomes. *Nucleic Acids Res*. 27:1767–1780.
- Boore JL, Brown MW. 2000. Mitochondrial genomes of *Galathealinum*, *Helobdella*, and *Platynereis*: sequence and gene arrangement comparisons indicate that Pogonophora is not a phylum and Annelida and Arthropoda are not sister taxa. *Mol Biol Evol*. 17:87–106.
- Burland TG. 2000. DNASTAR's lasergene sequence analysis software. *Methods Mol Biol*. 132:71–91.
- Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*. 17:540–552.
- Cavalier-Smith T. 1991. Intron phylogeny: a new hypothesis. *Trends Genet*. 7:145–148.
- Cousineau B, Lawrence S, Smith D, Belfort M. 2000. Retrotransposition of a bacterial group II intron. *Nature*. 404:1018–1021.
- Cui X, Mastsuura M, Wang Q, Ma H, Lambowitz AM. 2004. A group II intron-encoded maturase functions preferentially in *cis* and requires both the reverse transcriptase and X domains to promote RNA splicing. *J Mol Biol*. 340 (2):211–231.
- Curcio MJ, Belfort M. 1996. Retrohoming: cDNA-mediated mobility of group II introns requires a catalytic RNA. *Cell*. 84:9–12.
- Dai L, Toor N, Olson R, Keeping A, Zimmerly S. 2003. Database for mobile group II introns. *Nucleic Acids Res*. 31:424–426.

- Dai L, Zimmerly S. 2002. Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic Acids Res.* 30:1091–1102.
- Dellaporta SL, Xu A, Sagasser S, Jakob W, Moreno MA, Buss LW, Schierwater B. 2006. Mitochondrial genome of *Trichoplax adhaerens* supports placozoa as the basal lower metazoan phylum. *Proc Natl Acad Sci USA.* 103:8751–8756.
- Eeckhaut I. 1998. *Mycomyzostoma caldicola* gen. et sp. n., the first extant myzostomid infesting crinoid stalks, with a nomenclatural appendix by M. J. Grygier. *Spec Div.* 3:89–103.
- Eeckhaut I, McHugh D, Mardulyn P, Tiedemann R, Monteyne D, Jangoux M, Milinkovitch MC. 2000. Myzostomida: a link between trochozoans and flatworms? *Proc R Soc Lond B.* 267:1383–1392.
- Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol Mar Biol Biotech.* 3:294–299.
- Guo H, Zimmerly S, Perlman PS, Lambowitz AM. 1997. Group II intron endonucleases use both RNA and protein subunits for recognition of specific sequences in double-stranded DNA. *EMBO J.* 16:6835–6848.
- Grygier MJ. 2000. Myzostomida. In: Beesley PL, Ross GJB, Glasby CJ, editors. *Polychaetes and Allies: The Southern Synthesis. Fauna of Australia, Vol 4A Polychaeta, Myzostomida, Pogonophora, Echiura, Sipuncula.* Melbourne: CSIRO Publishing. 297–330.

- Haszprunar G. 1996. The Mollusca: coelomate turbellarians or mesenchymate annelids?
In: Taylor JD, editor. Origin and Evolutionary radiation of the Mollusca. Oxford:
University Press. 3–28.
- Hausner G, Olson R, Simon D, Johnson I, Sanders ER, Karol KG, McCourt RM,
Zimmerly S. 2005. Origin and evolution of the chloroplast trnK (matK) intron: a
model for evolution of group II intron RNA structures. *Mol Biol Evol.* 23(2):380–
391.
- Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogeny.
Bioinformatics. 17:754–755.
- Ichiyanagi K, Beauregard A, Lawrence S, Smith D, Cousineau B, Belfort M. 2002.
Retrotransposition of the Ll.LtrB group II intron proceeds predominantly via
reverse splicing into DNA targets. *Mol Microbiol.* 46:1259–1272.
- Jenkins BD, Barkan A. 2001. Recruitment of a peptidyl-tRNA hydrolase as a facilitator
of group II intron splicing in chloroplasts. *EMBO J.* 20:872–879.
- Jennings RM, Halanych KM. 2005. Mitochondrial genomes of *Clymenella torquata*
(Maldanidae) and *Riftia pachyptila* (Siboglinidae). Evidence for conserved gene
order in Annelida. *Mol Biol Evol.* 22:210–222.
- Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence
alignment program. *Brief Bioinform.* 9:286–298.
- Knoop V, Kloska S, Brennicke A. 1994. On the identification of group II introns in
nucleotide sequence data. *J Mol Biol.* 242:389–396.
- Lambowitz AM, Caprara MG, Zimmerly S, Perlman PS. 1999. Group I and group II
ribozymes as RNPs: clues to the past and guides to the future. In Gesteland RF,

- Cech TR and Atkins JF (eds), *The RNA World*, 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 451–484.
- Lambowitz AM, Zimmerly S. 2004. Mobile group II introns. *Annu Rev Genet.* 38: 1-35.
- Lang BF, Laforest MJ, Burger G. 2007. Mitochondrial introns: a critical view. *Trends Genet.* 23 (3):119–126.
- Lanterbecq D, Rouse GW, Milinkovitch MC, Eeckhaut I. 2006. Molecular phylogenetic analyses indicate multiple independent emergences of parasitism in Myzostomida (Protostomia). *Syst. Biol.* 55:208–227.
- Lavrov D. 2007. Key transitions in animal evolution: a mitochondrial DNA perspective. *Integr. Com. Biol.* 47:734–743.
- Lehmann K, Schmidt U. 2003. Group II introns: structure and catalytic versatility of large natural ribozymes. *Crit Rev Biochem Mol Biol.* 38:249–303.
- Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25:955–964.
- Lynch M. 2007. *The origins of genome architecture*. Sunderland MA: Sinauer Associates.
- Maddison DR, Maddison WP. 2002. *MacClade 4: Analysis of Phylogeny and Character Evolution*, version 4.0. Sunderland, MA: Sinauer Associates.
- Matsuura M, Saldanha R, Ma HW, Wank H, Yang J, Mohr G, Cavanagh S, Dunny GM, Belfort M, Lambowitz AM. 1997. A bacterial group II intron encoding reverse transcriptase, maturase, and DNA endonuclease activities: biochemical demonstration of maturase activity and insertion of new genetic information within the intron. *Genes Dev.* 11:2910–2924.

- Michel F, Ferat JL. 1995. Structure and activities of group II introns. *Annu Rev Biochem.* 64:435–461.
- Robart AR, Zimmerly S. 2005. Group II intron retroelements: function and diversity. *Cytogenetic Genome Res.* 1100:589–597.
- Rot C, Goldfarb I, Ilan M, Huchon D. 2006. Putative cross-kingdom horizontal gene transfer in sponge (Porifera) mitochondria. *BMC Evol Biol.* 6:71.
- Rouse GW, Fauchald K. 1997. Cladistics and polychaetes. *Zool Scr.* 26:139–204.
- San Filippo J, Lambowitz AM. 2002. Characterization of the C-terminal DNA-binding/DNA endonuclease region of a group II intron-encoded protein. *J Mol Biol.* 324:933–951.
- Simon DM, Clarke NAC, McNeil BA, Johnson I, Pantuso D, Dai L, Chai D, Zimmerly S. 2008. Group II introns in Eubacteria and Archaea: ORF-less introns and new varieties. *RNA.* 14:1704–1713.
- Toor N, Hausner G, Zimmerly S. 2001. Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. *RNA.* 7:1142–1152.
- Toro N, Molina-Sánchez MD, Fernández-López M. 2002. Identification and characterization of bacterial class E group II introns. *Gene.* 299:245–250.
- Vallès Y, Boore JL. 2006. Lophotrochozoan mitochondrial genomes. *Int Comp Biol.* 46:544–557.
- Vallès Y, Halanych KM, Boore JL. 2008. Group II Introns Break New Boundaries: Presence in a Bilaterian's Genome. *PLoS ONE.* 3(1): e1488.

- Wank H, San Filippo J, Singh RN, Matsuura M, Lambowitz AM. 1999. A reverse transcriptase/maturase promotes splicing by binding at its own coding segment in a group II intron RNA. *Mol Cell.* 4:239–250.
- Zimmerly S, Hausner G, Wu X. 2001. Phylogenetic relationships among group II intron ORFs. *Nucleic Acids Res.* 29:1238–1250.
- Zhong J, Lambowitz AM. 2003. Group II intron mobility using nascent strands at DNA replication forks to prime reverse transcription. *EMBO J.* 22:4555–4565.
- Zhong M, Struck TH, Halanych KM. 2008. Phylogenetic information from three mitochondrial genomes of Terebelliformia (Annelida) worms and duplication of the methionine tRNA. *Gene.* 416:11–21.
- Zrzavy J, Hypsa V, Tietz D. 2001. Myzostomida are not annelids: molecular and morphological support for a clade of animals with anterior sperm flagella. *Cladistics* 17:1–29.
- Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31:3406–3415.

Table 1. Long PCR primers used in amplifications.

Fragments	Primer name	Sequence	Annealing Temp.
<i>mLSU-cox1</i>	16S-Myz-longF	5'—AGG GTT AAT TAA ATC AAT CCT ACT AGT AAG AC—3'	
	CO1-Myz-longR	5'—TAT TCG TTC TAA TTT CAA TGC TGT TGA TCG—3'	
<i>cox1-cob</i>	CO1-Myz-longF	5'—ATT TTT TCC TTA CAT TTA GCT GGG GCT AGG-3'	53
	Cytb-Myz-longR	5'—TGT TTA ACT CCT AAA GGG TTT GAT GAC CCG C—3'	53
<i>cob-nad5</i>	Cytb-Myz-longF	5'—TCC TCA TTA ATA AAA ATC CCG TTC CAC CCG—3'	54
	Nad5-Myz-618R	5'—TAC TAG TGC AGA AAC GGG TGT AGG TGC TGC—3'	54
<i>nad5-mLSU</i>	Nad5-Myz-615F	5'—GTA CAC TCA TCA ACA TTA GTA ACA GCA GGC—3'	54
	16S-Myz-longR	5'—CTT TAG AAA AAT AAA CCT GTT ATC CCT GTG G—3'	54

Table 2. Sequencing primers used for sequencing mitochondrial genomes.

Fragments	Primer name	Sequence	
<i>coxI-cob</i>	CO1-Myzo-720F	5'—GCA TAC AAA AAT ATC TTT CAA GGT G—3'	
	CO1-Myzo-636F	5'—CAT CAT TTT TTG ACC CAT CG—3'	
	Cytb-Myzo-211R	5'—TGA ATT AAC CGA ATC ACC C—3'	
	Intron-Myz-713F	5'—AGG AAG CTA TAT TAA CGA C—3'	
	Intron-Myz-635R	5'—AGG TAA TTT ATG CCT GTG—3'	
	Intron-Myz-201R	5'—TTT TAG GTG CTA TTT AGC—3'	
	Cob-Myzo-241R	5'—TTT GAG TAC AGG TGG TAG—3'	
	Cob-Myzo-161R	5—AAG TTT ATT GGT GCT GG—3'	
	CO3-Myz-453R	5'—GAG TGG CCT ATA ATG CCT G—3'	
	CO1-Myz-1068F	5'—ACT CAT CAC TAG ATA TAC C—3'	
	Intron-Myz-954F	5'—AAT GAT AAT GTT AAA ATA CG—3'	
	CO3-Myz-startR	5'—TAA ATG GTA TGG TTG TCG G—3'	
	CO3-M-378F	5'—AGA TTG GAA TAG TAT GGC—3'	
	CO1-M-918F	5'—AGC CTA TTT CAC GTC TGC G—3'	
	Nad6-Myzo-164R	5'—AGC TGT TGA AAT AGC TAA GC—3'	
	Intron-Myzo-470F	5'—AAC TAC AAA ATT AAC CGC—3'	
	Intron-Myzo-1163F	5'—ATA ATT TGA TAG GGC CAC C—3'	
	CO1-Myz-1389F	5'—TTA TCA GGA ATA CCA CGA CG—3'	
	CO1-Myz-Intron2F	5'—AAT ATC ACG TAC AAT TCG G—3'	
	CO2-Myz-422R	5'—TTC AGT CGG TAA TAT GTA TG—3'	
	tRNAD-Myz-R	5'—TAT TGG TGC AAG TTG AGG C—3'	
	<i>cob-nad5</i>	Nad5-Myz-411R	5'—TTA TAC CTG CAG ATA GTG—3'
		Nad5-Myz-261R	5'—GTA TTT GAT TTT GCT CAT C—3'
		Cob-Myz-1020F	5'—TAA TAT AGC TGT GCT ACC—3'
Cob-Myz-840R		5'—GAT TGA TCG AAG CAT GGC—3'	
Cob-Myz-endsR		5'—CAT CAA AAT CTA GGC TAC C—3'	
Atp6-Myz-midR		5'—GAA ATG TAT AGT GAT AAC—3'	
Atp6-Myz-midF		5'—TTT TCT TTG TGC GTT AGG—3'	
<i>Nad5-mLSU</i>	Nad5-Myz-1015F	5'—TAT TTA TCA GGG TTC TAT TC—3'	
	16S-Myz-59R	5'—ATT ATG CTA CCT TAG TAC G —3'	
	12S-Myz-midF	5'—GTT TCT AGC CTA TAA GGC —3'	

Fragments	Primer name	Sequence
	12S-Myz-mid2F	5'—ACC AAC CTT ACA CAT TTC CG—3'
	12S-Myz-endR	5'—TTA TAT TTG CCG AAT TCC—3'
	Nad5-Myz-1179F	5'—TAA TAA TAT CAA CAC TAA TGA C—3'
	12S-Myz-statR	5'—TAT TTA TTT TTC CAA GGT TG—3'
	Nd5-Myz-1611F	5'—ATT ACA AAA GAG AAT CCC AAT C—3'
	12S-Myz-startR2	5'—ATA ATA GAG ATA GAT CAC ATC—3'
	Nad4L-M-251F	5'—ATC ATA ACT CGA ACA TAC GG—3'
	Nad4-M-915R	5'—CTT GAT GTA AGC CCA TGG GC—3'
	Nad4-M1-824R	5'—ATG GGT AAT TGA TGA GTA TGC G—3'

Table 3. Intron identification primers.

Introns	Primer name	Sequence
Intron1	LCO1490 (Folmer et al.,1994)	5'—GGT CAA CAA ATC ATA AAG ATA TTG—3'
	Intron-Myz-635R	5'—AGG TAA TTT ATG CCT GTG—3'
Intron2	CO1-Myz-Intron2F	5'—AAT ATC ACG TAC AAT TCG G—3'
	CO1-Ann-1300R	5'—TCC GGG TAR TCW GAR TAT CGT CGW GG—3'

Table 4. Group II ORFs used in the phylogenetic analysis.

Intron Class	Organism Class	Species	Host gene	Intron			
mitochondrial	Annelids	<i>Nephtys sp.</i>	<i>cox1</i>				
	Myzostomids	<i>Endomyzostoma sp.</i>	<i>cox1</i>	Mintron1			
	Placozoans	<i>Trichoplax ashaerens</i>	<i>cox1</i>	OFR677			
	Green Plant		<i>Arabidopsis thaliana</i>	<i>nad1</i>	I4		
			<i>Glycine max</i>	<i>nad1</i>	I4		
			<i>Oenothera berteriana</i>	<i>nad1</i>	I4		
			<i>Vicia faba</i>	<i>nad1</i>	I4		
			<i>Zea mays</i>	<i>nad1</i>	I4		
			Liverwort		<i>Marchantia polymorpha</i>	<i>atpA</i>	I1
	<i>Marchantia polymorpha</i>	<i>atpA</i>			I2		
	<i>Marchantia polymorpha</i>	<i>Atp9</i>			I1		
	<i>Marchantia polymorpha</i>	<i>cob1</i>			I3		
	<i>Marchantia polymorpha</i>	<i>cox1</i>			I1		
	<i>Marchantia polymorpha</i>	<i>cox1</i>			I2		
	<i>Marchantia polymorpha</i>	<i>cox2</i>			I2		
	<i>Marchantia polymorpha</i>	<i>SSU</i>			I1		
	Algae				<i>Chara vulgaris</i>	<i>nad3</i>	I2
					<i>Mesostigma viride</i>	<i>cox2</i>	I1
			<i>Porphyra purpurea</i>	<i>LSU rDNA</i>	I1		
			<i>Porphyra purpurea</i>	<i>LSU rDNA</i>	I2		
			<i>Pavlova lutheri</i>	<i>cox1</i>	I1		
			<i>Pylaiella littoralis</i>	<i>LSU rDNA</i>	I1		
			<i>Pylaiella littoralis</i>	<i>LSU rDNA</i>	I2		
			<i>Pylaiella littoralis</i>	<i>cox1</i>	I1		
			<i>Pylaiella littoralis</i>	<i>cox1</i>	I2		
			<i>Pylaiella littoralis</i>	<i>cox1</i>	I3		
			<i>Thalassiosira pseudoana</i>	<i>cox1</i>	I1		
			<i>Rhodomonas salina</i>	<i>cox1</i>	I1		

Intron Class	Organism Class	Species	Host gene	Intron
		<i>Rhodomonas salina</i>	<i>cox1</i>	I2
	Fungus	<i>Allomyces macrogynus</i>	<i>cox1</i>	I3
		<i>Rhizphydium sp. 136</i>	<i>cox1</i>	I11
		<i>Neurospora crassa</i>	<i>cox1</i>	I1
		<i>Podospora anserina</i>	<i>cox1</i>	I1
		<i>Podospora anserine</i>	<i>cox1</i>	I4
		<i>Podospora anserina</i>	<i>nad5</i>	I4
		<i>Podospora comata</i>	<i>cox1</i>	I1
		<i>Podospora curvicolla</i>	<i>nad5</i>	I1
		<i>Venturia inaequalis</i>	<i>cob1</i>	I1
	Yeast	<i>Candida parapsilosis</i>	<i>cox1</i>	I1
		<i>Candida stellata</i>	<i>LSU</i>	I1
		<i>Kluyveromyces lactis</i>	<i>cox1</i>	I1
		<i>Saccharomyces cerevisiae</i>	<i>cox1</i>	I1
		<i>Saccharomyces cerevisiae</i>	<i>cox1</i>	I2
		<i>Schizosaccaromyces pombe EF2</i>	<i>cob1</i>	I1
		<i>Schizosaccaromyces pombe</i>	<i>cox1</i>	I1
		<i>Schizosaccaromyces pombe</i>	<i>cox2</i>	I2
		<i>Schizosaccaromyces octosporus</i>	<i>cox2</i>	I1
	Ichthyospora	<i>Amoebidium parasiticum</i>	<i>cox1</i>	I6
Chloroplast	Plant	<i>Nicotiana tabacum</i>	<i>trnK</i>	I1
		<i>Marchantia polymorpha</i>	<i>trnK</i>	I1
	Green Algae	<i>Scenedesmus obliquus</i>	<i>petD</i>	I1
		<i>Oocystacea sp.</i>	<i>petD</i>	I1
		<i>Bryopsis maxima</i>	<i>rbcL</i>	I1
		<i>Pyrenomonas salina</i>	<i>cpn60</i>	I1
	Euglenoid	<i>Euglena gracilis</i>	<i>psbC</i>	I4
		<i>Euglena gracilis</i>	<i>psbD</i>	I8
		<i>Euglena deces</i>	<i>psbC</i>	I4

Intron Class	Organism Class	Species	Host gene	Intron
		<i>Euglena myxocylindracea</i>	<i>psbC</i>	I4
		<i>Euglena viridis</i>	<i>psbC</i>	I4
		<i>Lepocinclis buetschli</i>	<i>psbC</i>	I4
Bacteria	Eubacteria	<i>Alkaliphilus metalliredigenes (Al.me.I4)</i>	None	
		<i>Sinorhizobium terangae strain ORS22 (Sr.t.I1)</i>	<i>ISRm10-1</i>	
		<i>Uncultured marine bacterium (UMB.I1)</i>	Unknown	
		<i>Bradyrhizobium japonicum (B.j.I2-1)</i>	Putative transposase	
		<i>Chlorobium phaeobacteroides DSM 266 (Ch.ph.I2-1)</i>	none	
		<i>Frankia sp. (Fr.sp.I1)</i>	Transposase	
		<i>Uncultured bacterium (UB.I1)</i>	none	
		<i>Clostridium beijerincki NCIM 0852 (Cl.be.I2)</i>	none	
		<i>Wolbachia endosymbiont of Drosophila melanogaster (W.e.I2)</i>	Hypothetical protein	
		<i>Mycobacterium vanbaalenii (My.va.I1)</i>	none	
		<i>Burkholderia fungorum Bcep_271 (B.f.I1)</i>	none	
		<i>Sodalis glossinidius (So.gl.I1)</i>	none	
		<i>Pseudomonas putida (P.p.I1)</i>	none	
		<i>Syntrophobacter fumaroxidans (Sy.fu.I1)</i>	none	
		<i>Solibacter usitatus Ellin6076 (So.us.I3-1)</i>	none	
		<i>Klebsiella pneumoniae isolate YMC (Kl.pn.I1)</i>	<i>Metallo-beta-lactamase VIM-2</i>	
		<i>Oceanobacillus theyensis (O.i.I1)</i>	none	
		<i>Bacillus sp. NRRL (B.sp.I1)</i>	none	
		<i>Clostridium acetobutylicum (C.a.I1)</i>	none	
		<i>Syntrophomonas wolfei subsp. wolfei str. Goettingen (Sy.wo.I2-1)</i>	none	
		<i>Desulfotalea psychrophila LSv54 (D.p.I1)</i>	none	
		<i>Bacillus anthracis (B.a.I2)</i>	<i>pX01-24/ORFX</i>	

Intron Class	Organism Class	Species	Host gene	Intron
		<i>Clostridium perfringens</i> (Cl.pe.II)	none	
		<i>Bacillus megaterium</i> (B.me.II)	Conserved region	
		<i>Bacillus thuringiensis</i> (B.th.II)	Conserved	
			hypothetical protein	
		<i>Enterococcus faecalis</i> (E.f.I4)	FtsK/SpoIIIE family	
			protein	
		<i>Crocospaera watsonii</i> WH 8501 (C.w.I7)	none	
		<i>Nostoc punctiforme</i> Npun_200 (N.p.II)	none	
		<i>Trichodesmium erythraeum</i> IMS101(Tr.e.I4)	Putative RP	
			ribonucleotide	
			reductase	
		<i>Onion yellows phytoplasma</i> (OYPII)	Putative helicase	
			DnaB	
		<i>Shigella flexneri</i> (S.f.II)	IS629-like ORF	
		<i>Azotobacter vinelandii</i> (A.v.II)	groEL	
		<i>Alcanivorax borkumensis</i> SK2 (Al.bo.II)	none	
		<i>Xylella fastidiosa</i> (X.f.II)	DNA	
			methyltransferase	
		<i>Bacteroides thetaiotaomicron</i> (B.t.II)	TraG like protein	
		<i>Shewanella baltica</i> OS155 (Sh.ba.II)	none	
Archaeobacteria		<i>Methanococcoides burtonii</i> DSM 6242	RNA-directed DNA	
		(Mc.b.II-1)	polymerase	
		<i>Methanosarcina acetivorans</i> (M.a.II-1)	Conserved region	
			(species)	
		<i>Methanosarcina acetivorans</i> (M.a.II-2)	ORF of M.a.I3	
		<i>Methanosarcina mazei</i> (M.m.II)	Transposase	
		Uncultured archaeon GZfos26G2 (UA.II)	none	
		Uncultured archaeon GZfos28G7 (UA.I5)	none	
		<i>Methanospirillum hungatei</i> (Me.hu.I2)	none	

Intron Class	Organism Class	Species	Host gene	Intron
		<i>Uncultured archaeon GZfos14B8 (UA.14)</i>	none	
		<i>Methanosarcina acetivorans (M.a.15-1)</i>	ORF of M.a.I1-3	
		<i>Methanosarcina acetivorans (M.a.15-3)</i>	ORF of M.a.I1-2	

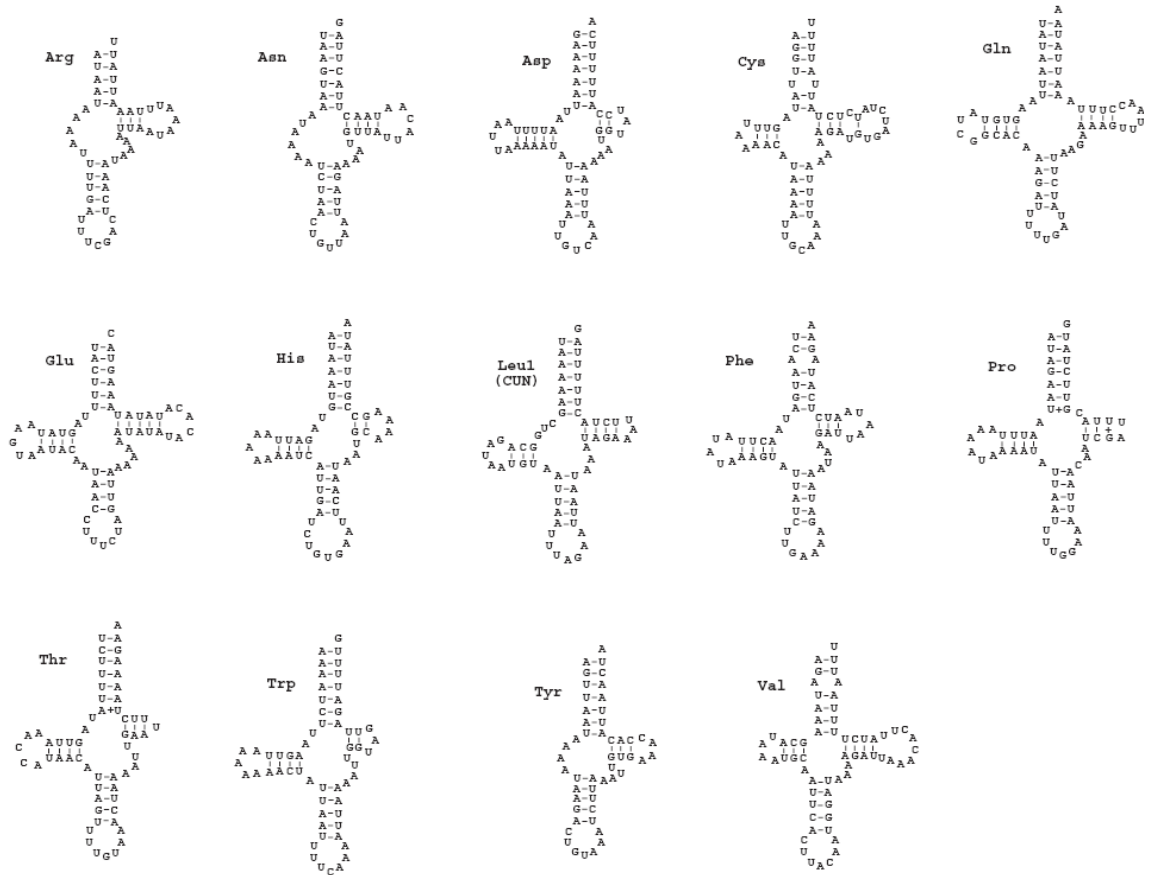
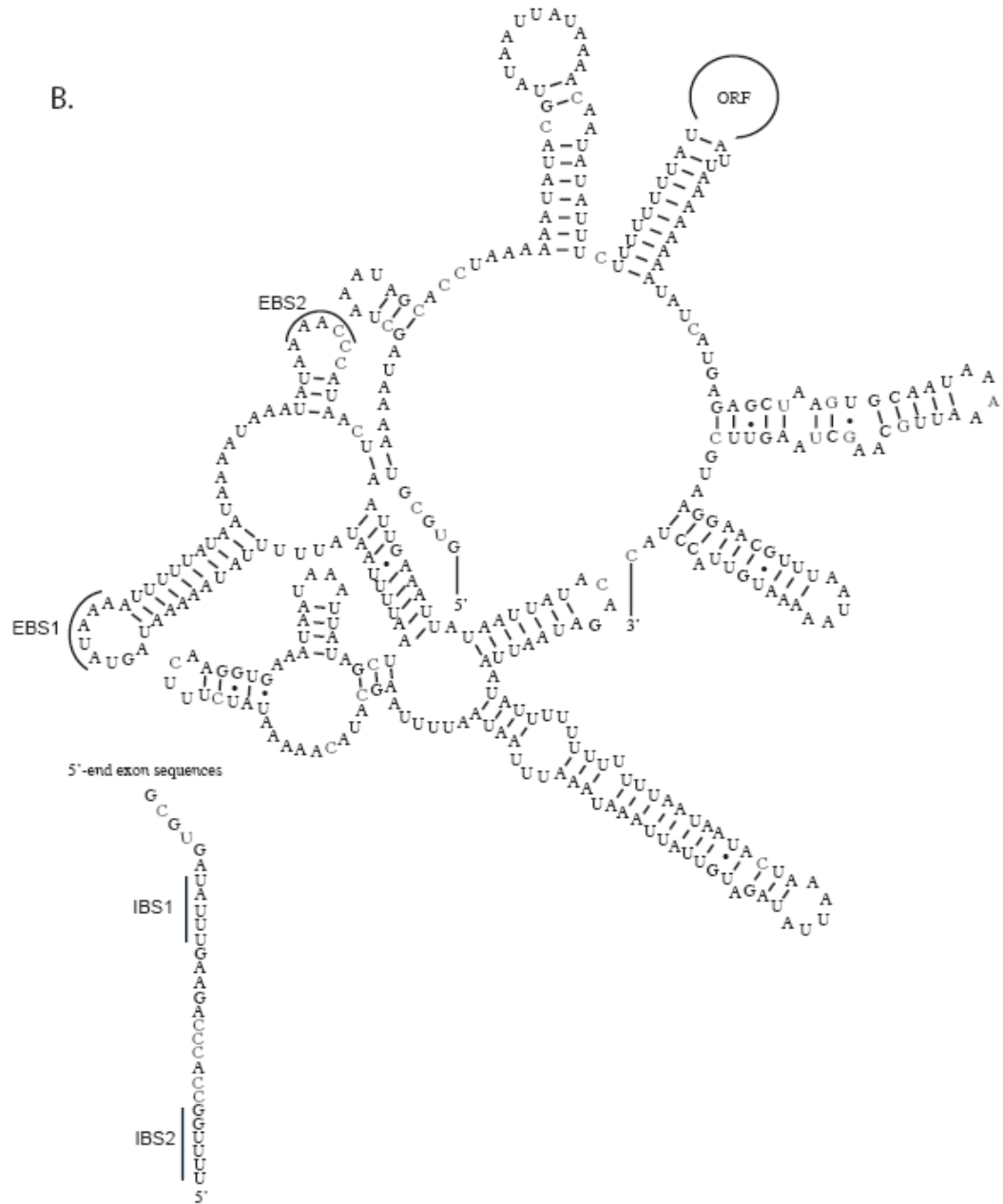


Figure 1. The secondary structures of 14 discovered tRNA in the mitochondrial genome of *Endomyzostoma* sp..

A.

cox1(609bp)	Mintron1 (1424bp)	cox1 (405bp)	Mintron2 (384bp)	cox1 (408bp)
-------------	-------------------	--------------	------------------	--------------

B.



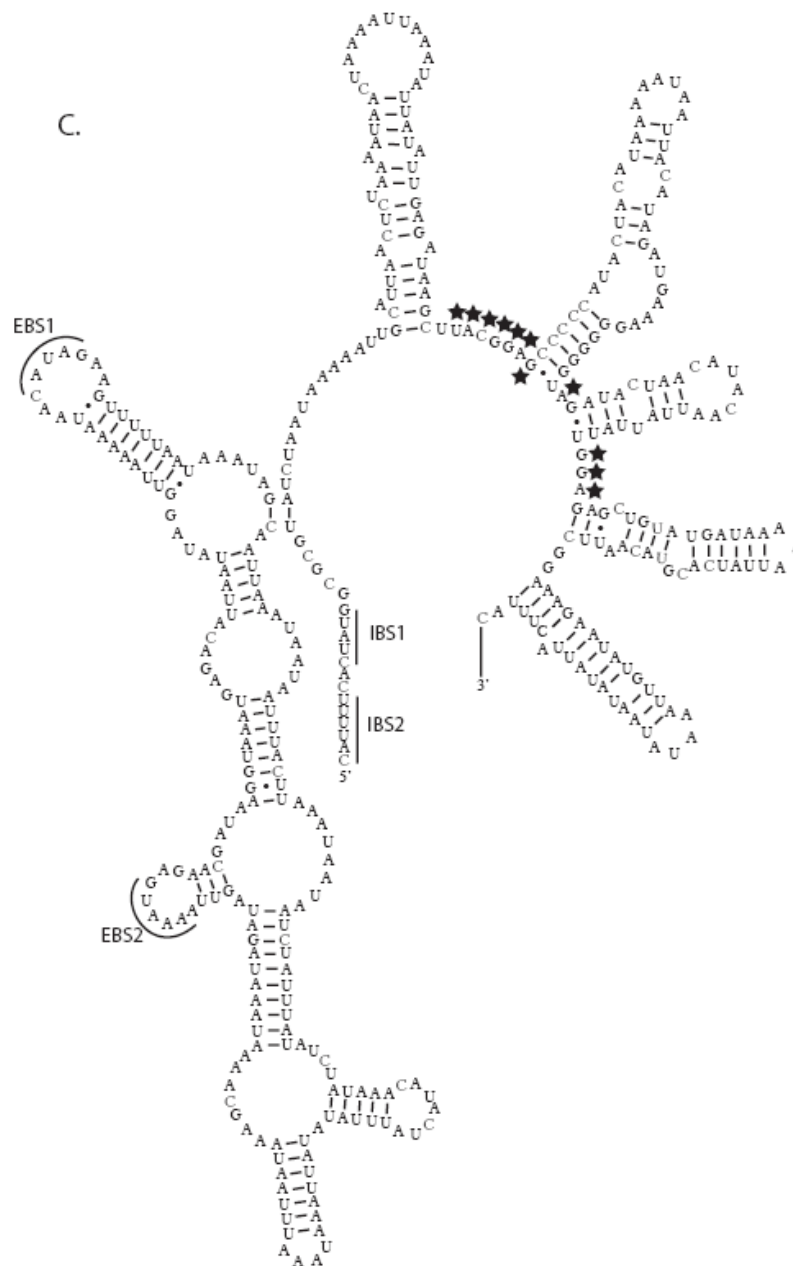


Figure 2. (A) Insertion locations of both Mintron1 and Mintron2 in the *cox1* gene. The dot line on the left end of *cox1* gene means this region was not recovered. (B) Secondary structure of Mintron1. Domain 4 contains the ORF. (C) Secondary structure of Mintron2. The stars label the conserved nucleotides in the central core as the major features of group IIA1 subclass which is the mitochondrial class. ORF, open reading frame; EBS, exon binding sequences; IBS, intron binding sequences.

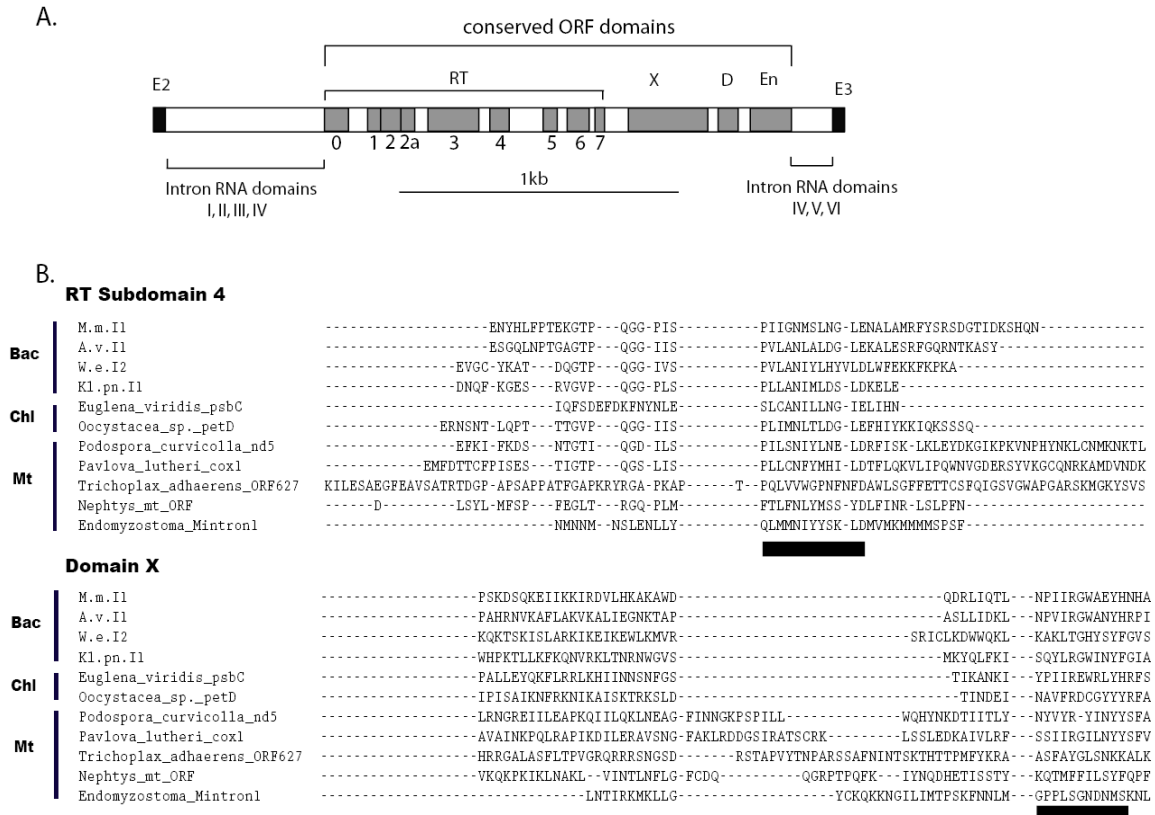


Figure 3. (A) Intron-encoded ORF domain structure. (modified from Zimmerly et al., 2001; Fig1A). (B) Alignment of subdomain 4 region of RT domain and one conserved region in X domain. The most conserved amino acids were identified by Gblocks showing as the black bars. The representative ORFs from each major category (three mitochondrial, two chloroplast and four bacteria) were selected to align with the ORFs from both Mintron1 and *Nephtys* group II intron. The purple bars are showing the most conserved regions across all ORFs. Species abbreviations are as follows: M.m.I1, *Methanosarcina mazei*; A.v.I1, *Azotobacter vinelandii*; W.e.I2, *Wolbachia* endosymbiont of *Drosophila melanogaster*; Kl.pn.I1, *Klebsiella pneumoniae isolate YMC*. The intron category labels: Bac, Bacteria; Chl, Chloroplast; Mt, Mitochondrion. Other information about the taxa see table 4.

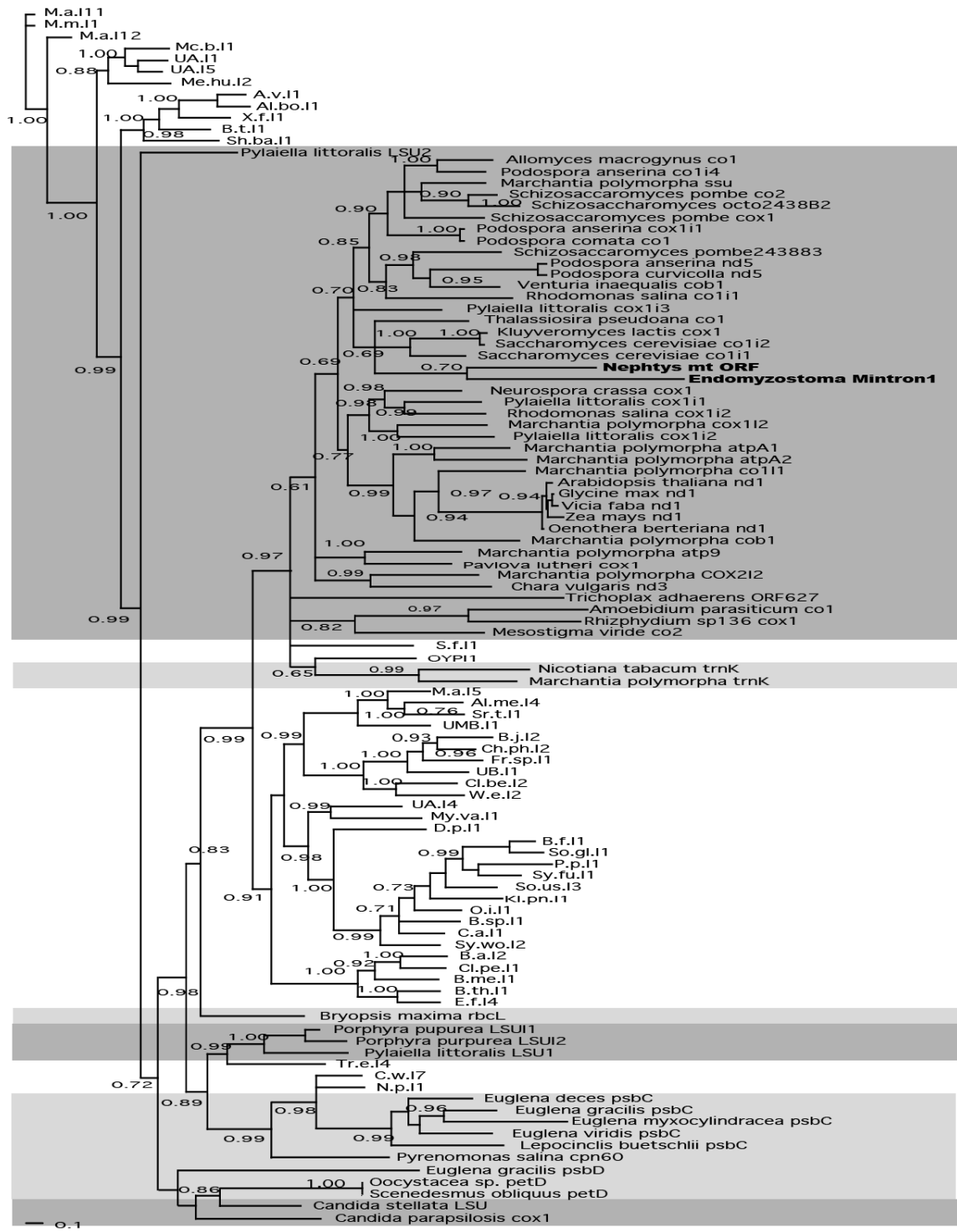


Figure 4. Phylogenetic analyses of 105 group II intron ORFs by the MrBayes program. Taxa sampling included all mitochondrial and chloroplast group II ORFs and some representatives of bacterial classes (table 4). Dark shades are the mitochondrial group II introns including the Mintron1 and *Nephlys* ORFs (showing as bold font). Light shade is the chloroplast group II intron lineage. All other unshaded are the group II intron ORFs found in bacteria. Archaeobacterial taxa were used as outgroups.

CHAPTER 5

CONCLUSIONS

1. Mitochondrial genomes and phylogenetic relationships within terebelliformia group

Over the past few decades, phylogenetic relationships within annelids have been a disputed topic with a growing interest from many researchers (e.g. Rouse and Fauchald, 1997; Bleidorn et al., 2003; McHugh, 2005; Rousset et al., 2006; Struck et al., 2007). Differences in data used for analyses, such as morphology or various molecular loci, can cause such phylogenetic discrepancies. In order to resolve this issue, current tendency is to use multiple-genes, and even genomic datasets (e.g, mitochondrial genome or Expressed Sequence Tag data) to develop robust phylogenetic hypotheses (Bleidorn et al., 2006; Zhong et al., 2008). Our study proposed a phylogenetic reconstruction among five families of Terebelliformia annelids by using data from the mitochondrial genome.

Mitochondrial genomes were sequenced from six worms, representative of five recognized Terebelliformia families. For two of species examined, UNK regions of the mitochondrial genomes were fully described. By comparison, the complete UNK regions in both *Terebellides stroemi* (Trichobranchidae) and *Pista cristata* (Terebellidae) are AT rich and consist microsatellite-like sequences which could form complicated secondary structures to affect the ability of polymerases during amplification or sequencing (see also Boore and Brown, 2000; Jennings and Halanych, 2005; Bleidorn et al., 2006; Zhong et al., 2008).

The data recovered for Terebelliformia mitochondrial gene order fits previous hypotheses of conserved gene order among protein-coding genes in annelida (Jennings and Halanych, 2005; Vallès and Boore, 2006; Bleidorn et al., 2006). Although rearrangements are observed for tRNA genes and UNK regions., mitochondrial gene order, within annelids, offers limited phylogenetic signal. Exceptions to this observation are the presence of a *trnM* duplication present in a terebellid, trichobranchid and pectinariid as well as a variable protein-coding gene order in *atp6-mSSU* region only found in ampharetids.

Based on the mtDNA data, maximum likelihood and Bayesian inference approaches were used to reconstruct the tree topolog for both mitochondrial nucleotide and amino acid data. Both resultant trees have an identical topology within Terebelliformia groups with high nodal support values, suggesting well-resolved relationships among the five families based on mtDNA data. The results herein not only provide a phylogenetic hypothesis for Terebelliformia groups, but also indicate that mitochondrial genomes could be a useful tool for understanding phylogenetic relationships within annelids, especially when combined with other sources of data, such as nuclear markers. Moreover, as our efficiency of sequencing mtDNA genomes continues to improve, additional sampling will be essential to more thoroughly elucidate annelid phylogenetic relationships.

2. The mitochondrial genome of a Myzostomida worm and group II introns

Group II introns are self-splicing ribozymes that have only recently been reported in bilaterian animals (Vallès et al., 2008, in a *Nephtys* annelid). My studies discovered

two group II introns in a partial mitochondrial genome of an *Endomyzostoma* (Myzostomida) worm. The protein-coding and rRNA gene order in mtDNA of *Endomyzostom* sp. matches another Myzostomida worm, *Myzostoma seymourcollegiorum*, which lacks such introns (Bleidorn et al., 2007). And both share the same protein-coding gene arrangement pattern as the majority published annelids (Vallès and Boore, 2006; Zhong et al., 2008), further supporting the hypothesis of an annelid origin of myzostomids.

This is the first study to describe multiple group II introns found in a bilaterian genome. Two group II introns, named as Mintron1 and Mintron2, were identified in *cox1* gene. The similarity of the insertion site between Mintron1 and *Nephtys*' intron implies the susceptible region in annelid *cox1* genes for the group II intron recognition. My work focused on characterizing the secondary structures of the introns, which holds implications for self-splicing and retrohoming ability. Mintron2 lacks of an Open Reading Frame (ORF) and therefore presumably lacks retrohoming mobility. It could perform splicing ability *in vivo* by recruitment of some host proteins to keep the activity of the *cox1* gene. Based on its contracted secondary structure and it has lost retrohoming capability, I predicted Mintron2 is undergoing evolutionary degeneration. Mintron1, however, possesses an ORF in D4 domain with only RT and X domains, inferring that it can presumably accomplish reverse splicing *in vivo*. But needs to insert into the replication fork to assist itself synthesize the double strand DNA for retrotransposition due to lacking of both D and En domains in the ORF (Zhong and Lambowitz, 2003). Despite there are several conserved regions found in the ORF, phylogenetic analyses

showed Mintron1 is a divergent group II intron related to the other group II intron in *Nephtys*.

In conclusion, the study indicated it is unquestionable to be very informative for the family level phylogenetic analyses using mitochondrial genome data. Future work would be focused phylogenetic reconstruction based on the combination datasets with nuclear markers as well as the large sampling of each Terebelliformia family. In terms of group II introns, the continuous work into their evolutionary and functional significance in bilaterian genomes is needed.

References

- Bleidorn, C., Vogt, L., Bartolomaeus, T., 2003. New insights into polychaete phylogeny (Annelida) inferred from 18S rDNA sequences. *Mol. Phylogenet. Evol.* 29, 279–288.
- Bleidorn, C., Podsiadlowski, L., Bartolomaeus, T., 2006. The complete mitochondrial genome of the orbiniid polychaete *Orbinia latreillii* (Annelida, Orbiniidae)--A novel gene order for Annelida and implications for annelid phylogeny. *Gene* 370, 96–103.
- Bleidorn, C., Eeckhaut, I., Podsiadlowski, L., Schult, N., McHugh, D., Halanych, K.M., Milinkovitch, M.C., Tiedemann, R., 2007. Mitochondrial genome and nuclear sequence data support Myzostomida as part of the annelid radiation. *Mol Biol Evol.* 24(8):1690–1701.
- Boore, J.L., Brown, W.M., 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., Staton, J.L., 2002. The mitochondrial genome of the Sipunculid *Phascolopsis gouldii* supports its association with Annelida rather than Mollusca. Mol. Biol. Evol. 19, 127–137.
- Felsenstein, J., 1985. Phylogenies and the comparative method. American Naturalist. 125:1–15.
- Jennings, R.M., Halanych, K.M., 2005. Mitochondrial genomes of *Clymenella torquata* (Maldanidae) and *Rifta pachyprila* (Siboglinidae): Evidence for conserved gene order in Annelida. Mol. Biol. Evol. 22, 210–222.
- Rouse, G.W., Fauchald, K., 1997. Cladistics and polychaetes. Zool. Scri. 26, 139–204.
- Rousset, V., Pleijel, F., Rouse, G.W., Erséus, C., Siddall, M. E., 2006. A molecular phylogeny of annelids. Cladistics 22, 1–23.
- Vallès, Y., Boore, J.L., 2006. Lophotrochozoan mitochondrial genomes. Int. Comp. Biol. 46:544–557.
- Vallès, Y., Halanych, K.M., Boore, J.L., 2008. Group II Introns Break New Boundaries: Presence in a Bilaterian's Genome. PLoS ONE. 3(1): e1488.
- Zhong, J., Lambowitz, A.M., 2003. Group II intron mobility using nascent strands at DNA replication forks to prime reverse transcription. EMBO J. 22:4555–4565.
- Zhong, M., Struck, T.H., Halanych, K.M., 2008. Phylogenetic information from three mitochondrial genomes of Terebelliformia (Annelida) worms and duplication of the methionine tRNA. Gene 416:11–21.