A cause for consilience: Utilizing multiple genomic data types to resolve problematic nodes within Arthropoda and Ecdysozoa

A thesis submitted to the National University of Ireland for the Degree of
Doctor of Philosophy

NUI MAYNOOTH

Ollscoil na hÉireann Má Nuad

Presented by:
Lahcen I. Campbell
Department of Biology,
NUI Maynooth,
Maynooth,
Co. Kildare, Ireland.

February 2012

Supervisor: Dr. Davide Pisani B.Sc., Ph.D. (Bristol)
Head of Department: Professor Kay Ohlendieck, Dip.Biol., M.Sc. (Konstanz), Ph.D.

Table of Contents

For my Family and Friends

Acknowledgments

As I find myself sitting down to write these acknowledgements, a point in which I thought I would be extraordinarily lucky to ever reach, it is sorely clear to me just how many wonderful people I have been privileged to have around me over the past few years.

First off I would like to extend my deepest gratitude and thanks to my supervisor Dr. Davide Pisani. I hope you understand just how sincerely grateful I am for you providing me the opportunity to undertake something so exciting and deeply fascinating. Over the years you have been an absolute pleasure to work with, not only for your countless thoughtful and dedicated insights as a scientist, but as someone whom I have come to admire immensely.

My research would not have been possible without the funding from Science Foundation Ireland (SFI) the computational resources provided by the Irish Center for High End Computing (ICHEC) and The High Performance Computing (HPC) Facility at NUI Maynooth.

Next on the list is the lab'ers from the Bioinformatics unit, both past and present. You guys have been the best bunch of people anyone could wish to work (and play ☺) with. It would be an understatement to say I feel incredibly lucky to have got to know you all. To Therese, I would really like to thank you for your warm and helpful welcome to the Pisani lab (in the early days of no windows ☺) when I started my PhD. To the arthropod guru Dr. Omar Rota-Stabelli, I would like to thank you for your very insightful and encouraging introduction to the world of legged critters. To Dr. Stuart Longhorn, thank you for your help during those long days and nights slaving over gel trays in Dartmouth; it would have been impossible to complete on time without your help. Thanks to Brain Daley for preventing an impromptu dropout due to my utter frustration in the face of buggy scripts and uncompilable software.

I would like to sincerely thank Dr. Kevin Peterson for allowing me visit his Dartmouth lab to learn their techniques, and for your very helpful comments on past manuscripts and general accommodating nature to my many work related queries. Also thanks to Dr. Petersons past and present lab members who were all so amazingly helpful and patient during my time there.

To all my magnificent friends not only within Maynooth but also from 'the borough' and further afield, thank you all from the bottom of my heart for your unwavering support and encouragement over the years. To Carla and Leanne, thank you for always being the first ones to help me out when times were low, for providing those moments of inspiration and for your general uplifting spirit you so readily supplied. To Aoife and Sinead, thank you for the endless supply of epic banter and the countless cups of tea, without which I would have gone insane.

To my Family, thank you for being there for me, particularly my granny Kathleen whom has always had me in her thoughts and prayers. To my Parents, Rosemary and Lahcen, I am eternally indebted to the incredible encouragement and most of all belief you have always had in me. Thank you so much for the endless support you have shown me over the long time I've been in Maynooth, not only throughout my undergrad but also in the challenging times that was completing my PhD. I just want to say that I am the luckiest son in the world to have such wonderful parents.

And last but not least, thank you molecular evolution, without which none of this would have been possible ☺.

Declaration

This thesis has not been submitted in whole, or in part, to this, or any other University for any other degree and is, except where otherwise stated, the original work of the author.

Signed:________________________________

Lahcen Iouani Campbell

Abbreviations

Base Pair (bp)

Basic Local Alignment Search Tool (BLAST)

Bootstrap Support (BS)

Complementary DNA (cDNA)

Deoxyribonucleic Acid (DNA)

Expressed Sequence Tag (EST)

General Time reversible (GTR)

Horizontal Transfer (HGT)

Large Subunit (LSU)

Likelihood Ratio Test (LRT)

Long Branch Attraction (LBA)

Markov Chain Monte Carlo (MCMC)

Maximum Likelihood (ML)

Maximum Parsimony (MP)

Messenger RNA (mRNA)

MicroRNA (miRNA)

miRNA* (microRNA star)

MicroRNA-Containing Ribonucleoprotein Complex (miRNP)

MicroRNA-Containing RNA-Induced Silencing Complex (miRISC)

Neighbor Joining (NJ)

Next-Generation Sequencing (NGS)

Nucleotide (nt)

Operational Taxonomic Units (OTUs)

PAUP* (Phylogenetic analysis Using Parsimony (*and other methods)

Posterior Probability (PP)

Protein Coding Genes (PCGs)

Ribonucleic Acid (RNA)

Small Interfering RNA (siRNA)

Small Subunit (SSU)

Small Temporal RNA (stRNA)

Standard Deviation (SD)

Transfer RNA (tRNA)

Untranslated Region (UTR)

Index of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Index of Tables

Chapter 3

Chapter 5

Abstract

A major turning point in the study of metazoan evolution was the recognition of the existence of the Ecdysozoa in 1997. This is a group of eight animal phyla (Nematoda, Nematomorpha, Loricifera, Kinorhyncha, Priapulida, Tardigrada, Onychophora and Arthropoda). Ecdysozoa is the most specious clade of animals to ever exist and the relationships among its eight phyla are still heatedly debated. Similarly also the relationships among the three sub-phyla (Chelicerata, Pancrustacea and Myriapoda) within the most important ecdysozoan phylum (the Arthropoda) are still debated. Indeed, the two major problems in ecdysozoan phylogeny refer to the relationships of Myriapoda within Arthropoda, and of Tardigrada within Ecdysozoa. Difficulties in ecdysozoan relationships resides in lineages characterized by rapid, deep divergences and subsequently long periods of divergent evolution. Phylogenetic signal to resolve the relationships of these lineages is diluted, increasing the likelihood of recovery of phylogenetic artifacts.

In an attempt to resolve the relationships within Ecdysozoa, consilience of three independent phylogenetic data sets was investigated. EST and rRNA and microRNA (miRNA) data were sampled across all major ecdysozoan phyla. In particular, a major contribution of this thesis is the first time sequencing of miRNAs for all the panarthropod phyla. MicroRNAs are genome regulatory elements that recently emerged as a source of useful phylogenetic data (Sempere *et al.* 2006) because of their low homoplasy levels.

The considered data sets were analysed under phylogenetic methods and models, implemented to minimize the occurrence of phylogenetic reconstruction artifacts to understand the evolution of Ecdysozoa. Analyses of independent data types recovered well supported and corroborating evidence for the monophyly of Panarthropoda (Arthropoda, Onychophora and Tardigrada), a sister group relationships between Myriapoda and Pancrustacea within Arthropoda, and the paraphyly of Cycloneuralia (Nematoda, Nematomorpha, Loricifera, Kinorhyncha and Priapulida).

Chapter 1

Introduction

1.1 The phylum Arthropoda: The long road behind us

1.1.1 Significance of the arthropod expansion

The phylum Arthropoda is one of the most successful and diverse groups of animals to ever exist, not also in sheer numbers of species but also in terms of ecological niche filling, spanning the globe across marine, tropical, temperate, arid and polar regions. To gain perspective on species diversity within this phylum, let us consider that there is an estimated 1.9 million species of eukaryotes described presently, and that approximately half (~1.1 million) of these species are thought to be arthropods; predominately insects (Chapman, 2009).

Attempts have been made to try to estimate a precise figure for the number of arthropod species worldwide, controversial estimates initially ranged from 30 to 100 million species (Erwin, 1982; Erwin, 1988). Recently as more data have been acquired and better statistical analyses performed, revised estimates have been generated suggesting the actual figure might be much lower at between 2.5 to 4.8 million (see Hamilton, 2010; Ødegaard, 2000). Although the revised estimate of species richness now seems to be much lower than initially thought; this is not yet cause for celebration, as gaps remain in the study arthropod taxonomy with currently less than 40% of all crustacean and 20% of all insect species described and documented (Hawksworth and Kalin-Arroyo, 1995).

When one considers the enormity of the arthropod radiation, which unfolded over ~600 million years (Erwin *et al.* 2011) unsurprisingly questions arise such as, why are arthropods so successful? What mechanisms prompted arthropods to diversify on such a scale? How are all the different arthropod groups related to one another? It seems with well over a century of debate on arthropod evolution (Siebold, 1848; Snodgrass 1938; Tiegs, 1947; Manton, 1973; Friedrich and Tautz, 1995; Pisani *et al.* 2004; Rota-Stabelli *et al.* 2011), some of these questions are now becoming resolved.

Within the phylum Arthropoda, there exists a diverse array of body plans and size variation that has allowed arthropods to fill all kinds of important ecological niches. Emergence of some of the major arthropod Bauplaene (Woodger, 1945) has allowed the group to diversify to a degree unmatched anywhere else in the metazoan tree of life. One of the most abundant groups of arthropods, Crustacea, which include planktonic forms, make up the vast majority of biomass in our marine environments. These planktonic crustaceans that feed on phytoplankton and zooplankton, are responsible for injecting large amounts of nutrients into the food web and sustain many of the higher trophic level interactions in marine ecosystems. Not only are arthropods important ecologically, they appear to be pivotal to economic sustainability and growth when we look for instance at the role in crop pollination of arthropod groups like the honeybee, bumblebee and butterfly (Aizen & Harder, 2009). Data have emerged showing that the numbers of such pollinator groups are on the decline (Gross, 2008; Oldroyd, 2007), sparking much attention to a potential “pollination crisis” affecting the global production and sale of crops.

1.1.2 The demise of Coelomata and the monophyly of Arthropoda

Traditionally, studies on the interrelationships of major metazoan groups such as the bilaterian arthropods, molluscs, nematodes and chordates have been based upon morphological data. In light of the evidence taken from comparative anatomy, embryology and development, many hypotheses were proposed to describe the evolution of the Bilateria (Jenner and Schram, 1999), with the most prominent being the Coelomata hypothesis (sensu Hyman, 1940). According to this model, bilaterian relationships should be considered on the basis of a coelom (a fluid filled cavity, which grows from the mesoderm and is present only in triploblast animals). The relationships of Bilateria are then graded on the degree of its presence or absence, resulting in three defined groups: Acoelomata (Platyhelminthes and Nemertinea), Pseudocoelomata (Nematoida, Priapulids, Kinorhyncha and Rotifera), and the Coelomata (remaining Bilateria e.g. Arthropoda, Annelida, Mollusca and Vertebrata).

Morphological evidence has prompted many studies to conclude arthropods are the closely related sister phyla of other segmented protostome groups, particularly the annelid worms (sensu Hyman, 1940; Hyman, 1951). The grouping of segmented protostomes i.e. molluscs, annelids along with panarthropods (Arthropoda, Onychophora and Tardigrada) became referred to as the Articulata hypothesis (Anderson 1973; Wheeler *et al.* 1993; Schmidt-Rhaesa, 1998; Wägele and Misof, 2001; Scholtz, 2002). The first major study to refute the Coelomata grade organisation of Bilateria and in effect, the Articulata hypothesis was by Aguinaldo *et al.* (1997). In this study bilaterian evolution was investigated using 18S rRNA. The findings presented by Aguinaldo *et al.* suggested that Bilateria was in fact composed of Protostomes and Deuterostomes, with the Protostomes being further sub divided into Lophotrocozoa and Ecdysozoa. Within Ecdysozoa, the arthropods were allied to metazoan phyla, such as Nematoda, Nematomorpha, Tardigrada, Onychophora, Priapulida, Kinorhyncha and Loricifera. All groups share a common feature of exoskeletal growth through a process of repeated moulting or "ecdysis". Since its proposal, multiple studies using a range of different markers and methods e.g. complete genomes, development, EST's and Supertrees (Eernisse *et al.* 1992; Ruiz-Trillo *et al.* 2002; de Rosa *et al.* 1999; Haase *et al.* 2001; Philippe *et al.* 2005b; Dunn *et al.* 2008; Holton and Pisani 2010; Rota-Stabelli *et al.* 2011) have upheld the Ecdysozoa hypothesis to where it is now widely accepted (Kumar *et al.* 2011). We see then, that caution needs to be taken when viewing animal evolution on the basis of irreversible change of body form, as it seems clear now that secondary simplification of organisation can occur, which in some cases can lead to misinterpretation of evolutionary relationships (Philippe *et al.* 2011a).

The study of arthropod phylogeny traces its roots back to before the turn of the 20th century. Since early classical studies there has been significant headway made in the study of arthropod evolution, particularly in the last few years with important insights stemming from molecular biology, developmental biology, improved phylogenetic methods and the onset of the genomic era. Arthropods represent a very enigmatic and exciting group to study. Charles Darwin spent time refining his theory of Natural selection while studying bees and their ability to construct intricate hexagonal nectar preserving honeycombs (Darwin, 1859). Early discussions of arthropod phylogeny based on hard structures of external morphology suggested that Arthropoda evolved from a single common ancestor, uniting all the major sub groups (sub phyla) i.e. Trilobites (extinct), chelicerates, myriapods, crustaceans and insects (Snodgrass, 1938). Monophyly of Arthropoda is now widely accepted (Turbeville *et al.* 1991; Ballard *et al.* 1992; Wills *et al.* 1995; Giribet *et al.* 2001; Telford *et al.* 2008; Budd and Telford 2009; Regier *et al.* 2010; Rota-Stabelli *et al.* 2011). However, the concept of a monophyletic Arthropoda has long been contentious. In particular, Sidney Manton and others championed arthropod polyphyly in the 70's (Manton, 1973; Anderson, 1973; Anderson, 1979). Proponents of a paraphyletic or polyphyletic origin of arthropods suggested that attributes commonly used to unite Arthropoda be considered convergences due to similarities of lifestyle. Tiegs (1947) proposed a group comprising Myriapoda + Hexapoda + Onychophora (Uniramia) which was sister to a group containing Trilobita + Chelicerata + Crustacea (TCC: Cisne, 1974). According to Tiegs and Manton these groups where independently derived from annelid-like ancestors. However, phylogenetic signal within morphological data sets was weak and relied on interpretation of characters, which is always, to some extent, subjective. The utilization of molecular sequence data for phylogenetic analyses in the 1980's ushered a new era in arthropod systematics, allowing previously debated hypotheses to be tested independently, for instance, the monophyly of Arthropoda began to receive high support via the analysis of ribosomal DNA e.g. 18S rDNA (Turbeville *et al.* 1991; Wheeler *et al.* 1993; Giribet *et al.* 1996; Spears and Abele, 1997).

1.1.3 Arthropoda and the new animal Phylogeny

The Phylum Arthropoda comprises one extinct sub phylum Trilobita (which however is a now known to be a polyphyletic assemblage) and 4 extant sub phyla: Crustacea (e.g. crabs and barnacles), Hexapoda (e.g. insects and springtails), Myriapoda (e.g. millipedes and centipedes) and Chelicerata (e.g. horseshoe crabs and arachnids) see Figure 1.1. Arthropods have long been treated as a single group; Linnaeus referred to this group as Insecta in *Systema Naturae* (1758). The term "Arthropoda" was not coined until the 19th century (Siebold, 1848), deriving its name from the Greek translation of "árthros" meaning jointed, and "podós" meaning foot; which constitutes their most obvious morphological feature: the presence of jointed appendages.

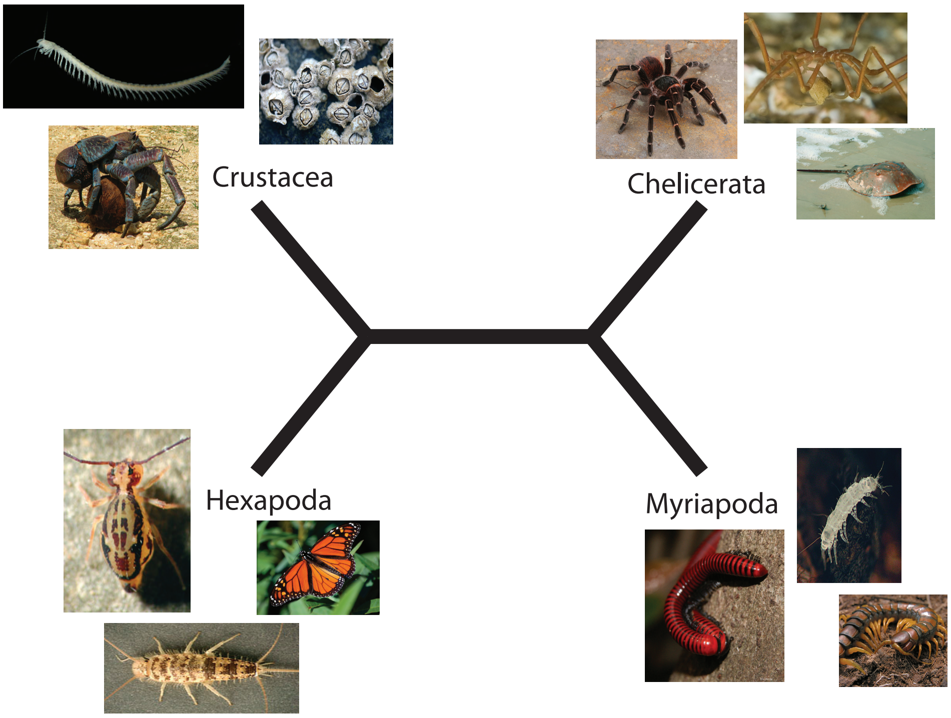


Figure 1.1: The Four Sub-Phyla of the Arthropoda. Tree shown as an unrooted network. Taxa shown represent example species for each of the four sub-phyla. Note the relationships of the Pancrustacea (Crustacea + Hexapoda) are still debated. In this figure for clarity, I represented the Crustacea as monophyletic, but a likely scenario is that the Crustacea are paraphyletic with respect to insects. Closest living sister phyla (Onychophora and Tardigrada) are not shown.

Arthropods have a number of unique synapomorphies (shared features uniting two or more taxa with their most recent common ancestor): a hard external segmented exoskeleton composed of α-chitin, intrinsic musculature between joints, segments bearing appendages with claws, a mixocoel with metanephridia, an ostiate heart, and in most cases a cephalon made up of multiple fused segments (Nielsen, 2001). Although there exists strong evidence in favour of a monophyletic origin of arthropods, relationships between the different sub phyla of Arthropoda and the arthropods closest living relatives remain the source of debate and controversy (see Figure 1.1).

Considering early traditional morphological evidence, Arthropoda is generally regarded as the closest phylum to two other ecdysozoan phyla, namely the soft-bodied Onychophora (Velvet worm) and the miniscule Tardigrada (Water bears), collectively referred to as the Panarthropoda (Nielsen, 2001). Despite some compelling evidence in support of Panarthropoda i.e. shared apomorphies of a cuticle composed of α-chitin, lateral-walking appendages on each segment, ostiate heart (absent in the miniature tardigrades) and lack of protonephridia, a consensus has yet to be reached regarding the relationships of the panarthropod phyla. Lack of phylogenetic resolution resides not only within Panarthropoda but also in relation to the remaining soft-bodied ecdysozoan phyla (Telford *et al.* 2008; Edgecombe, 2009). Accordingly, one of the main questions in panarthropod evolution is in the branching order of the three phyla, which is important to understand the processes and steps involved in the emergence of true "arthropodization", and furthermore the emergence of the ancestral ecdysozoan groundplan.

There are many morphological features, along with fossil evidence and countless phylogenetic studies that have provided compelling evidence in favour of placing onychophorans as the sister group to the Arthropoda (Ballard *et al.* 1992; Nielsen, 2001; Dunn *et al.* 2008; Budd and Telford, 2009; Braband *et al.* 2010; Mayer *et al.* 2010; Meusemann *et al.* 2010; Rota-Stabelli *et al.* 2010; Whitington and Mayer, 2011; Campbell *et al.* 2011), a clade referred to as Lobopodia (Snodgrass, 1938). However, the sister group position of Onychophora has not been universally accepted, due to numerous other analyses recovering Tardigrada as the closest living ancestral phyla to the arthropods (Giribet *et al.* 1996; Zrzavý *et al.* 1998; Edgecombe *et al.* 2000; Nielsen, 2001; Budd, 2001; Schmidt-Rhaesa 2001). The unresolved position of both Onychophora and Tardigrada in relation to arthropods leaves some open questions, not least when attempting to reconstruct the basic panarthropod groundplan. For example were ancestral panarthropods large, coelomate animals with a true blood vascular system? Questions such as this will be left open to misinterpretation and speculation if not addressed with a solid phylogenetic framework, as instances of convergencies or synapomorphies can be overlooked.

Following on from the unresolved placement of Onychophora and Tardigrada, the problem of phylogenetic and ancestral reconstruction within Panarthropoda and Ecdysozoa is further compounded by the number of recent molecular phylogenetic studies that resolve Tardigrada outside of Panarthropoda and group them within cycloneuralian ecdysozoans such as Nematoda and Priapulida (Philippe *et al.* 2005b; Sørensen *et al.* 2008; Roeding *et al.* 2007, Lartillot and Philippe, 2008; Roeding *et al.* 2009; Andrew, 2011). A tardigrade - nematode affinity raises the question; did segmentation seen in Arthropoda and Tardigrada evolve convergently, or was primitive segmentation present in the ecdysozoan ancestor subsequently lost in the remaining cycloneuralian phyla? In order to understand the origin of not only Arthropoda, but also Panarthropoda and Ecdysozoa, it is important to elucidate the true phylogenetic history of these groups in relation to their most recent common ancestors.

1.1.4 When Molecules and Morphology clash

Within the arthropods, it was once thought that the Insecta and Myriapoda sub phyla were closely related to one another (the Atelocerata, Antennata or Tracheata hypothesis) with both insects and myriapods breathing by way of a tracheal system. Support for the Atelocerata hypothesis was bolstered when it was proposed that this group should also include the Onychophora (the Uniramia Hypothesis, Tiegs 1947). Although looking back this is not unexpected as all three groups are terrestrial, breathe via tracheae, and have uniramous un-branched appendages and a single pair of antennae. This view on arthropod evolution can be seen in the context of increasing organisation, moving from a segmented annelid like ancestor, towards a lobopod like Onychophora and then eventually culminating in the "arthropodization" of Atelocerata and Crustacea plus Chelicerata (Schizoramia; sensu: Cisne, 1974) independently.

Early molecular phylogenetic analyses such as Field *et al.* (1988), and Lake (1990) set the scene for future reappraisals of some long standing established hypotheses on the evolution of arthropods and other metazoan groups (Halanych, 2004). However, early studies such as the one of Field *et al.* (1988) were hindered by poor phylogenetic reconstruction methods, sparse taxon sampling across the major metazoan lineages while also being heavily reliant upon a limited number of molecular sequences (e.g., 18S DNA, 28S DNA, Elongation Factor 1-α) (see Brusca, 2000). Still, later studies began to display overwhelming convergence of evidence for the monophyly of arthropods (Giribet *et al.* 2001; Pisani, 2004; Budd and Telford, 2009; Regier *et al.* 2010). Another interesting aspect elucidated from molecular phylogenetic analyses of Arthropoda was that the long held Atelocerata hypothesis (Hexapoda + Myriapoda) was not recovered in the vast majority of analyses. It should be noted that Atelocerata has a firm morphological basis, with shared features such as postantennal organs, Malpighian tubules, tentorial endoskeleton and a limbless intercalary segment (Klass and Kristensen, 2001, Bitsch and Bitsch, 2004). Molecular data has rejected Atelocerata, which should be viewed as a “Morphology-only” hypothesis (Edgecombe, 2010). Uniting synapomorphies of ‘Atelocerata’ should then be interpreted as convergences related to terrestrial habitats in both myriapods and hexapods (Averof and Akam, 1995) and indeed a variety of morphological apomorphies favouring a crustacean affinity for the insects have been described (mostly from the nervous system and the eye – see also Rota-Stabelli *et al.* 2011 and the below).

One of the accepted groupings within Arthropoda is that composed by crustaceans and insects. Since the onset of molecular systematics in the study of arthropod evolution, practically all analyses recover Crustacea + Insecta (Wheeler *et al.* 1993; Fredrich and Tautz, 1995; Giribet and Wheeler, 1999; Giribet *et al.* 2001; Nardi *et al.* 2003; Regier *et al.* 2005; Dunn *et al.* 2008; von Reumont *et al.* 2011; Regier *et al.* 2010) in combination with a mounting body of morphological evidence in favour of this group. The most striking morphological synapomorphy in support of this group, referred to as Pancrustacea (Zrzavý and Štys, 1997) or more specifically Tetraconata (Dohle, 2001) based on the tetrapartite crystalline cones of the ommatidia in their compound eyes (Dohle, 1997; Dohle, 2001; Harzsch, 2004). Further support for Tetraconata was found by independent studies of eye development (Harzsch and Hafner, 2006) neurogenesis (Ungerer and Scholtz, 2008) *Engrailed* expression in segmental mesoderm (Zrzavý and Štys, 1997) and mitochondrial gene order data (Boore *et al.* 1995; Boore *et al.* 1998). Although there is a sizable amount of evidence in support of Tetraconata, multiple placements of Hexapoda in relation to in-group

Crustaceans remains a particular problem, further compounded by the possibility of the joint paraphyly of both Crustacea and Hexapoda supported by some molecular data analyses (Regier and Shultz, 1997; Nardi *et al.* 2003; Cook *et al.* 2005; Giribet *et al.* 2005; Regier *et al.* 2010).

1.1.5 The question of myriapod affinity

Despite mounting evidence in support of Tetraconata, setting aside their exact relationships; a more prominent problem that has been the source of much debate and controversy over the past decade, is the position of myriapods within Arthropoda. Traditionally, arthropod groups united by the presence of a post-tritocerebral segmental appendage, which forms a jaw or 'mandible', has provided the basis to group crustaceans and hexapods together with myriapods into a clade known as Mandibulata (Nielsen, 2001). This group has long been recognised, as far back as the early work of Crampton (1921) and Snodgrass (1938); which united the three groups by way of homology of the Mandibles. Early phylogenetic studies in support of Mandibulata ranged from analyses of 18S rDNA (Giribet and Ribera, 1998) Elongation Factor-1α + RNA polymerase II (Regier and Shultz, 1997) combined histone H3 and U2 snRNA (Edgecombe *et al.* 2000) to combined nuclear and mitochondrial loci (Bourlat *et al.* 2008). At the same time, morphological evidence has mounted in support of Mandibulata, with features such as: the brain having a conserved midline neuropil, stomatogastric and labral nerves being connected to the tritocerebrum and not the deutocerebrum (Scholtz and Edgecombe, 2006) and sternal anlagen on the posterior stomodaeal region (Wolf and Scholtz, 2006). Furthermore, patterns of gene expression of the genes *Distal-less* (Scholtz *et al.* 1998) and *Dachshund* (Prpic *et al.* 2003) suggests a serial homology between the mandible, the coxal parts of the maxilla, labium, and the coxa of the legs (Edgecombe, 2004).

Taking the amassed morphological data in support of Mandibulata into consideration, it is surprising that the majority of molecular phylogenetic analyses do not support the Mandibulate affinity for myriapods. Instead of recovering Myriapoda with the tetraconatan arthropods, molecular data analyses have shown considerable support for a sister group relationship with chelicerates; in a clade referred to as Myriochelata (Pisani *et al.* 2004) or Paradoxopoda (Mallatt *et al.* 2004) this latter name referred to the seeming lack of morphological evidence supporting this clade.

Support for Myriochelata was first obtained via 18S rDNA analyses in the mid 1990's (Friedrich and Tautz, 1995; Giribet *et al.* 1996; Spears and Abele, 1997) further support was also found via mitochondrial genome analyses (Hwang *et al.* 2001; Nardi *et al.* 2003; Negrisolo *et al.* 2004; Hassanin, 2006) combined mitochondrial and nuclear genes (Pisani *et al.* 2004) combined 18S and 28S sequences (Mallatt *et al.* 2004; Mallatt and Giribet 2006; Gai *et al.* 2006) and HOX genes (Cook *et al.* 2001). Although Mandibulata is the more traditional of the two hypotheses in terms of morphology, a sparse number of uniting synapomorphies have been cited in support of Myriochelata. One proposed morphological character in support of Myriochelata derives from the developmental mechanism of neurogenesis (Dove and Stollewerk, 2003; Kandar and Stollewerk, 2004). In both insects (e.g. *Drosophila* and *Tribolium*) and some malacostracan crustaceans it has been shown that neurogenesis gives rise to both epidermal and neural cells; in contrast to chelicerates and myriapods in which there is no decision of epidermal or neural fate (see Stollewerk and Chipman, 2006 for a review) in the central neuroectoderm. A second proposed autapomorphy of Myriochelata is that both myriapods and chelicerates possess neural precursor groups as opposed to neuroblasts as seen in Tetraconata (Stollewerk and

Chipman, 2006). Although characters such as these may provide compelling evidence, polarizing the Euarthropod tree with these characters has proven difficult in the absence of similar studies in relevant outgroups, such as Onychophora or Tardigrada.

In a recent study of velvet worm development (Mayer and Whitington, 2010) it has been shown that the pattern of neurogenesis in velvet worms is more similar to that of hexapods and crustaceans than to that of myriapods and chelicerates, as Onychophora do not display post-mitotic cell clusters or segmental invaginations of the neuroectoderm. Mayer and Whittington (2010) cite another synapomorphy in favour of Myriochelata, the presence of a 'cumulus', a group of mesenchymal cells that act to initiate the breakdown of radial symmetry, leading to the dorsal split of the embryonic germ disc. The cumulus has been observed in both myriapods and chelicerates, while it has not been shown to be present within any Tetraconatan species. Considering the implications of such evidence supporting Myriochelata, one should conclude that either the uniting features of the Mandibulate head assemblage, such as the mandibles, evolved convergently in both Tetraconata and Myriapoda, or alternatively Mandibles are truly homologous representing a plesiomorphic character for Arthropoda. If we assume Myriochelata is correct, then presence of mandibles is a character that in chelicerates may have reverted from a biting mouthpart back into an ancestral biramous walking limb.

Resolving the position of myriapods within the Arthropoda when faced with the incongruences of independent analyses has proven difficult. Although the majority of molecular analyses support Myriochelata, studies that have combined all of the available evidence (Kluge, 1989) found support for Mandibulata (Zrzavý *et al.* 1998; Giribet *et al.* 2001; Giribet *et al.* 2005). However, studies that utilize sparse gene sampling, such as rDNA sequences (Zrzavý *et al.* 1998) combined rDNA, elongation factors, histone components, and mitochondrial Cytochrome C oxidase 1 (Giribet *et al.* 2001; Giribet *et al.*

2005) are becoming increasingly rare as modern phylogenetic studies tend to utilize vast amounts of genomic data; such as large scale sequencing of EST's (expressed sequence tags). Such studies have recovered highly supported topologies for many of the major metazoan clades, yet there still remains a lack of consensus regarding the placement of Myriapoda. In one of the largest phylogenomic studies conducted to date (Dunn *et al.* 2008), in which the authors analyzed a matrix of 150 genes for 77 species across the metazoa, support was found for Myriochelata with a relatively high bootstrap support of 90%. Since then, separate studies using large gene sets (>100 genes) have also recovered support for Myriochelata (Philippe *et al.* 2009; Hejnol *et al.* 2009; Pick *et al.* 2010). However all of the aforementioned phylogenomic studies suffered from poor taxonomic sampling for in-group arthropods, especially within Myriapoda and Chelicerata. Recently, the problem of myriapod affinity has been tackled by two independent phylogenomic analyses, (Regier *et al.* 2010; Rota-Stabelli *et al.* 2011) with both analyses attempting to expand the number of ingroup myriapod species in order to provide better phylogenetic signal for the placement of myriapods. Both analyses strongly supported the inclusion of Myriapoda within Mandibulata; with the study of Rota-Stabelli *et al.* (2011) supporting this hypothesis by way of three independent lines of evidence; phylogenomics, morphology and a new class of phylogenetic markers known as microRNAs. The microRNA analyses published in Rota-Stabelli *et al.* (2011) where obtained as part of the work presented in this thesis and will be presented in Chapter 4.

1.2 Molecular phylogenetics: Founding methods and modern approaches

The field of molecular phylogenetics dates back to the groundbreaking ideas of Zukerkandl and Pauling (Zukerkandl and Pauling, 1962) *"We may ask the questions where in the now living systems the greatest amount of their past history has survived and how it can be extracted"*. From a methodological point of view, modern molecular phylogenetics arose from the pragmatic mingling of ideas from the cladistic (Hennig, 1950; 1965) and the phenetic (Sokal and Sneath, 1963) schools; and championed by authors of the calibre of Joseph Felsenstein (see Felsenstein, 2004). Both the traditional cladistic and phenetic schools developed in the mid fifties, before powerful computer resources became available to phylogeneticists and before the large genomic databases we are familiar today were available. Looking back, the modern era of molecular phylogenetics has clearly come along way from these early times. Today for example we have relatively complex models of evolution that can take into account the heterogeneity of the substitution process across sites (Lartillot and Philippe, 2004) and also vastly increased computational power allowing the enormous number of calculations required for the currently widely used Maximum Likelihood and Bayesian methods. For example it is now possible to perform complete analyses of increasingly large phylogenomic data sets (Dunn *et al.* 2008; Hejnol *et al.* 2009; Campbell *et al.* 2011) under sophisticated models that are beginning to accommodate the complex processes of evolution we now know to occur. However no matter how sophisticated current models of evolution are, phylogenetic analyses are still prone to reconstruction artifacts, the most famous of which being long branch attraction – Felsenstein (1978). In this section I will discuss some of the earliest methods of modelling molecular evolution, the problems inherent in these methods and the advancements that lead to the current state of molecular phylogenetics.

1.2.1 Maximum Parsimony

Originally, pioneering algorithms developed to generate phylogenetic relationships were not focused on complex models of evolution. Models that attempt to account for inherent biases associated with many molecular data sets e.g. (among site rate variation, compositional heterogeneity and heterotachy) are recent innovations. Older, more simplistic methods such as parsimony – are based on the idea of "the minimum net amount of evolutionary change" Edwards and Cavalli-Sforza (1963; 1964). One of the first publications to use "clustering" methods for biological classification was a paper by Michener and Sokal (1957), which analysed morphological characters to classify bees. Around the same time as these early analyses, the first molecular sequence data were being generated in the form of protein sequences. Soon after sequence data began to be more commonly utilized, it was realised that molecular sequences could provide information in which to generate phylogenies. The famous paper by Zuckerlandl and Pauling (1962) is an example of such a leading innovation into the field of molecular phylogenetics in which they first proposed their hypothesis of the universal "Molecular clock" via the analysis of amino acid differences in haemoglobin sequences; which they showed changed roughly linearly with time.

The earliest computational approaches to generate phylogenetic relationships focused on methods such as parsimony (first applied to phylogeny reconstruction by Edwards and Cavalli-Sforza, 1964). The study presented by Edwards and Cavalli-Sforza (1964) focused on human gene frequency polymorphisms. This work was remarkable in the fact that they not only introduced the parsimony and likelihood method but also the use of statistical inferences to generate phylogenies more broadly. Parsimony is based on the concept of identifying the tree that minimises the number of character state transformations across all sites of the alignment, thereby finding the most "parsimonious tree". Although Edwards and Cavalli introduced parsimony, it was not used in terms of character-based phylogeny until the publication of Camin and Sokal (1965). Importantly, phylogenetic reconstruction methods that focus on the "minimum net amount of evolution" like parsimony do not consider branch lengths when selecting between alternative topologies. Due to inherent properties of parsimony, this method of phylogeny reconstruction was subsequently shown to have a number of problems. In particular Felsenstein (1978) showed parsimony to be inconsistent under certain conditions. That is when there is disproportionate rate heterogeneity in neighbouring branches of a topology, such conditions became referred to as the "Felsenstein zone" or more commonly by the manifestation of the reconstruction artifact - long branch attraction (LBA: See section1.3.3.4). As such Maximum parsimony is often criticised as being irrelevant to phylogenetics as evolution is rarely parsimonious.

1.2.2 Distance Matrix Methods

Another set of methods known as Distance matrix methods (DMM) has been in existence for a long time. In DMM, branch lengths represent expected amounts of evolution. This is a length of time, more precisely, branch lengths (BL) in phylogenetics is generally represented as the rate (ρ) of substitution multiplied by the time (the duration of the branch) hence BL = ρ * t. DMM methods calculate the distance between each pair of sequences in a multiple sequence alignment and generate a distance matrix of pairwise distances, this matrix is then used to determine the tree that reflects those distances more accurately (Felsenstein, 2004). Although the DMM do take into account the length of branches as a function of evolutionary distance, they are only simple fractions of observed amino acids differing between sequences and do not capture the reality of underlying evolutionary process. Distance matrix methods fail to fully take into account the intrinsic underlying process of evolution; processes such as biased substitutions patterns i.e. transitions vs. transversions, and unobserved multiple replacements due to high rates of substitution.

Although there has been number of different phylogenetic methods developed over the years, some of these, like Parsimony and DMM have now become essentially out-dated. Over the years improved phylogenetic reconstruction methods that allow for more accurate accounts of reality have began to replace those early methods. Specifically, Bayesian and Maximum likelihood methods, which employ more complex models of sequence evolution, have become increasingly utilized in their place. Accordingly, the work presented in this thesis features these improved methods extensively, and so I will not discuss further methods of phylogenetic reconstruction that are now viewed as being inadequate for modern phylogeny reconstruction.

1.2.3 Modelling Amino acid and protein evolution

It became apparent that the probability of one amino acid changing to anyone of the remaining nineteen amino acids was not equivalent for each pair of amino acids. Dayhoff and Eck (1968) introduced the use of empirical models of amino acid change; the first of these models was called the PAM (probability of accepted mutation) model. The first PAM substitution matrix was PAM 001. The PAM 001 model corresponds to the probability of any one of the 20 amino acids changing to any other of the 19 amino acids along a branch short enough that only 1% of the alignment positions is expected to change. A threshold of 1% allowed the assumption that the sequences were similar enough that no multiple substitutions had occurred, so for instance the likelihood of a particular mutation (e.g. F → W) being the result of the hidden substitutions (F → x → y →W) is low. Many more PAM matrices (e.g. PAM 100 and PAM 250) have been derived since then using matrix multiplication. These correspond to probabilities of changes among alternative amino acids along branches where greater evolutionary change has occurred. Currently Dayhoff matrices are no longer used for their original purpose, with Dayhoff matrices substituted by empirical derived matrices (like WAG – Whelan and Goldman, 2001) generated under a maximum likelihood framework. In any case, PAM matrices are sill sometimes used in BLAST-based database searches to access the significance of proposed matches between target and database sequences.

As additional data became available, the same methods used to derive the original PAM matrices were applied to larger datasets. Similar models of protein evolution were soon introduced, based again on empirical estimations of amino acid change. Jones, Taylor and Thornton (1992) described the empirical JTT matrix of amino acid replacement, while Whelan and Goldman (2001) then improved upon the JTT model by applying a likelihood framework to generate the WAG matrix. Models such as JTT and WAG ameliorated the assumption of Maximum Parsimony that any given site in an alignment only changes once along any single branch in a tree.

Models like the aforementioned WAG and JTT models are based on empirically derived replacement rates, and on the principle of time reversibility of the substitution process (i.e. GTR). That is, in a GTR model (mechanistic model) the probability of replacement for any amino acid is the same in both directions. Time reversibility negates the need of using a rooted topology (i.e. trees are inferred as unrooted topologies) and makes the calculation of the replacement matrix easier. As the probability of moving from amino acid J to K is the same of moving from K to J. GTR matrices are symmetrical therefore halving the number of parameters that need to be inferred.

Although mechanistic models mentioned previously like WAG and JTT improve the ability to estimate the underlying substitution process, thus improving the overall ability to correctly identify masked substitutions; a problem still remains in their use for phylogenetic reconstruction. The main problem inherent in all of the aforementioned models is that they all assume homogeneity of the replacement process. When the underlying assumptions of a given model are violated this usually results in generation of phylogenetic artifacts. Across site rate heterogeneity of the replacement process is a characteristic inherent in proteins. Amino acids are subject to heterogeneous replacement rates due to differential underlying physical properties of their amino acids, for instance globular proteins have some amino acid residues that are exposed to solution or alternatively buried in the protein core. The usual way in which among site rate heterogeneity is accounted for in homogeneous models is by way of a Gamma distribution of rates across sites (Yang, 1996). A Gamma distribution allows partial relaxation of the assumption of identical distribution of rates across sites, and has been shown to improve statistical adequacy over a uniform rate model (Yang, 1996). However most models of evolution still assume homogeneity of the replacement process, (i.e. equilibrium frequencies and rates of substitution across nucleotides and amino acids are the same across all sites) thereby promoting model violations and the occurrence of systematic errors and phylogenetic artifacts.

Attempts have been made recently to account for the problem of across site rate heterogeneity, most notably with the site heterogeneous mixture model CAT (Lartillot and Philippe, 2004). The CAT model also shares the feature of using gamma-distributed rates across sites, however CAT further relaxes the assumption of rate homogeneity across sites. This is achieved is via the clustering of columns of the alignment into a number of biochemically specific categories (*K*), each described by its own amino acid profile and equilibrium frequencies of the 20 amino acids (or 4 nucleotides). Columns of the alignment are assigned a category under which its substitutional history is to be described. The number of specific categories can be constrained to 1 (as in the standard matrix model i.e. WAG, JTT or GTR) or selected under a Dirichlet process prior on the number of equilibrium frequencies to let the value of *K* be a free parameter. CAT has been shown to be much more effective at modelling data sets that have experienced substantial degrees of substitutional saturation (Lartillot *et al.* 2007). CAT outperforms homogeneous models like WAG and the most general site homogeneous time reversible GTR model, lessening the problem of model violation and therefore generating more reliable phylogenies. The CAT model will feature extensively throughout this thesis.

1.2.4 Maximum likelihood

Maximum likelihood (ML) was first introduced back in the early 20th century by R. A. Fisher (1912; 1921; 1922). The concept of *likelihood* refers to the situation in which given some source of data *D***,** a decision must be made about an adequate explanation of the data. It wasn't until the early 1960's that ML was first applied to phylogeny by Edwards and Cavalli-Sforza (1964) when they applied ML to the analysis of gene frequency data. Although the implementation of ML for biological data had already been demonstrated, it was Joseph Felsenstein in 1981 that first showed using his pioneering "pruning algorithm" how to apply ML practically to realistic numbers of sequences.

Under a ML approach, a specific model and a hypothesis are formulated such that the model itself is not under question, but the data the model attempts to describe are. In phylogenetics, the model employed under ML assumes that sequences actually evolve according to a tree topology, and so, if point mutations or substitution events occur by chance, in principle one can calculate the probability of finding a mutation along a branch in a phylogenetic tree.

The main idea behind phylogenetic inference using ML is to determine the tree topology, branch lengths, and parameters of the evolutionary model (e.g. substitution model, base frequencies, rate variation among sites) that maximize the probability of observing the sequences under investigation. Typically the ML implementation in phylogeny reconstruction focuses around molecular sequence data such as DNA or Amino Acids (usually fixed) and a tree (part of a given "hypothesis", which is free to change). Another way to view the likelihood function is that it is the conditional probability of the data (i.e. sequence data) given a hypothesis (i.e. model of substitution with a set of parameters θ and the tree τ, including branch lengths).

$$\begin{aligned} L\ (\tau,\theta) &= \Pr\ (\mathit{Data}|\tau,\theta) \\ &= \Pr\ (\text{aligned sequences}\mid\text{tree, model of evolution}) \end{aligned}$$

One of the major advantages when using ML over other methods like Neighbour-joining (NJ) is that ML has been shown to impart robustness to systematic error and model misspecification which can affect parsimony and NJ (Hasegawa *et al.* 1991; Huelsenbeck 1995). Another advantage of ML inference is that it allows proper model selection. Currently a number of different statistical strategies exist to facilitate selection of the best fitting evolutionary model, such as information criteria, Bayesian or performance-based approaches. Probably one of the most popular methods is the likelihood ratio test (LRT). A Likelihood ratio test is a standard way of comparing the fit of two (or more) models of evolution by contrasting the maximised log-likelihoods of the null *l0* and the alternative models *l*1 (Posada and Crandall, 1998).

It has been shown that use of methods to select a best fitting evolutionary model (such as the LRT) increases the likelihood of reconstructing more accurate phylogenetic relationships (Keane *et al.* 2006). It can be said that in order to best describe the underlying evolutionary process, you should always try to avoid applying both an overly simplistic or overly parametric model of evolution. In the case of overly simplistic models it has been shown that underestimating multiple substitutions can result in statistical inconsistency during phylogeny estimation in certain situations ('Felsenstein zone') and can lead to systematic artifacts such as Long-branch attraction (Felsenstein, 1978a). Conversely, analysing small alignments (e.g. single gene data sets) with parameter rich models of evolution such as CAT (Lartillot and Philippe, 2004) and GTR can lead to overparameterisation. In such cases, finding the best tree might become impossible as all trees will have very similar likelihood (the likelihood surface will be flat) as there is not enough data to estimate all parameters in the model.

In conclusion, phylogenetic inference under a ML framework is a well-established and popular method of inference when constructing phylogenetic relationships. Specifically, ML has been shown to be largely robust to model violations and systematic errors, and so is viewed as being a substantial improvement over other less complex methods (e.g. Parsimony) (Huelsenbeck, 1995; Whelan *et al.* 2001). Indeed ML has a number of beneficial properties that promote its use as a phylogenetic reconstruction method. However, ML has not been utilized extensively in this thesis and so I will not discuss its properties any further.

1.2.5 Bayesian Inference

Bayesian inference can be viewed as being similar to likelihood methods; however, one main difference exists between the two. Bayesian inference differs by its use of a prior distribution on the entity being inferred (generally the trees) (Felsenstein, 2004). Bayesian inference has only recently become popular as a phylogenetic inference method despite its long history in statistics. This could be attributed to the effective implementation of Bayes' theorem via MCMC (Markov Chain Monte Carlo) algorithms (Rannala and Yang, 1996; Yang and Rannala, 1997; Mau *et al.* 1999; Larget and Simon, 1999). The attractiveness of Bayesian phylogenetic inference is the way in which it reflects our own "human" decision making process. In effect Bayesian inference is nothing more than a probability analysis that is updated as new information is added; thus mimicking our own rational decision making behaviour when presented with new information (Huelsenbeck and Bollback, 2001; Lemey *et al.* 2009).

Bayesian inference in phylogenetics is based upon the posterior probability of a tree. The posterior probability distribution or 'posterior' can be derived using Bayes' theorem:

$$\Pr[H|D] = \frac{\Pr[H] \text{ x } \Pr[D|H]}{\Pr[D]}$$

The posterior probability distribution (Pr[H|D]) is derived by calculating the probability of a hypothesis "H" given some data "D" (i.e. an alignment of sequences for *n* taxa). Here, the hypothesis H denotes a vector of model parameters that typically includes a topology, branch lengths and a substitution model for all alternative hypotheses (i.e. all trees possible for *n* taxa). Usually, the Prior (Pr[H]) for all trees is considered equally probable, a condition known as a vague or uninformative prior. In this equation the denominator (Pr[D]) is viewed as the normalizing constant, in which the denominator is the sum of the numerators ([Pr H and D]) over all possible hypotheses (H). This ensures that the posterior probability distribution integrates to 1, a basic requirement of a proper probability distribution. No matter how simple the model being implemented when deriving posterior probabilities, it is near impossible to calculate the denominator. To do so requires summing over all likelihood values for each hypothesis, i.e. trees; an intractable problem when viewed in terms of numbers of trees possible if *n* becomes large (Felsenstein, 1978b; Yang and Rannala, 1997; Lemey *et al.* 2009).

In a real world phylogenetic problem, calculating the posterior probability distribution analytically is impossible (Huelsenbeck and Bollback, 2001), this problem stems from the inability to estimate posteriors by drawing random samples from it (usually the posterior probability is concentrated in a small part of the parameter space). This problem can be overcome surprisingly easily by the use of MCMC (Markov chain Monte Carlo) algorithms, which allows a valid sample to be drawn from the posterior distribution (Huelsenbeck *et al.* 2001). An important property of MCMC chains is that they usually tend to converge towards an equilibrium state regardless of the starting point (i.e. a random tree) (Lemey *et al.* 2009).

The most common and flexible implementation of MCMC is via the use of the *Metropolis* algorithm, more specifically a variant referred to as *Metropolis-Hastings* algorithm (Metropolis *et al.* 1953; Hastings, 1970). The central premise of Metropolis-Hastings algorithm is to make small random changes to some current parameter value(s) then accept or reject those changes according to the appropriate probabilities. It is performed by following these steps: (1): Select a random starting state (i.e. a tree) with its associated posterior probability (θ). (2): Make a small random move by selecting a new state (θ*) from the proposal distribution. (3): Using the Metropolis-Hastings algorithm a decision is made to select or reject the new state, which is obtained by calculating the height ratio (*r*) of the posterior probabilities of the two states. There are two outcomes, either the new state is selected and becomes the starting point of the next proposal in the chain, or the current state is retained with a probability that is proportional to the height ratio (*r*) of the two states. On occasion a new state (θ*) with a lower probability than the current state (θ) is selected; which ensures the ratio of rejecting or accepting states is relative to the ratio of their posterior probabilities. In other words, the amount of time spent sampling from within a particular parameter value (i.e. a topology), is proportional to the posterior probability of that value i.e. the better the likelihood the more likely it is to be accepted.

The cycle of proposal/acceptance in a MCMC chain is repeated ad infinitum. Usually MCMC chains are not ran singularly, but more usually as a number of 'independent MCMC chains'. As chains usually start at a random state (i.e. tree) the initial posterior probability is generally quite low (*burn-in* phase) as chains sample from very different regions of tree space. As the chains progresses to regions of the posterior with high 'probability mass' we observe the likelihood increasing rapidly and the chains enter *stationarity*. Under what is known as convergence (i.e. chains in regions of tree space with similar probability distribution) of independent chains, allow you to evaluate the state of progression and cease the MCMC algorithm. The work presented in this thesis features the use of Bayesian inference extensively.

1.2.6 Posterior probability

The measure of support used in Bayesian inference is known as the Posterior probability (PP). The PP of a node within a tree is the probability that that node is correct (conditional on the model, the priors, and the data) (Huelsenbeck and Rannala, 2004). One of the benefits of PP assessment of support is that inference of PP is direct and does not require, for example repeated sampling and reanalysis, as is the case with Bootstrapping and Jack-knifing. However, mixed interpretations of support inferred from PP exist. According to some authors PP tends to be an overestimate of the real support values (Douady *et al.* 2003; Erixon *et al.* 2003). Bayesian analysis has the property that parameters are treated as random variables, and can be directly assigned probabilities thus conferring a natural way to access uncertainty in a phylogeny. This allows Bayesian inference to incorporate the use of models with greater dimensionality (Lartillot and Philippe, 2004) thereby conferring a better approximation of the true underlying evolutionary processes. This is however met with a caveat, in that it was shown that PP are more sensitive to model underspecification (Huelsenbeck and Rannala, 2004).

In conclusion BI is a powerful method of phylogenetic inference with a number of unique and intuitively positive properties; with currently a number of software implementations including *MrBayes* (Huelsenbeck and Ronquist, 2001)) and *Phylobayes* (Lartillot and Philippe, 2004). Bayesian analysis will be a widely used tool in this thesis.

Chapter 2

Considerations for phylogeny reconstruction: Data types, Phylogenetic error and Consilience

2.1 Introduction

The simplest definition of a phylogeny can be stated as follows: a phylogeny is a branching diagram depicting the genealogy or pattern of evolution for a group of operational units (typically: species, populations, single genes). Synonyms such as phylogenetic tree or evolutionary tree are usually used more commonly in place of phylogeny due to the similarity of the branching pattern to that of a tree; for instance different parts of a phylogeny are referred to accordingly (i.e. root, branch, leaf). Although one of the first evolutionary trees to appear in literature was in the publication "*Elementary geology*" by Edward Hitchcock in 1840, which depicted the relationships of plants and animals against a geological background; it wasn't until the theory of Natural selection was published in "*On the Origin of Species*" (Darwin, 1859) that popularized representing evolutionary common ancestry with the aid of a branching tree (see Figure 2.1).

When we think of a phylogeny, we will usually think of the branching tree (*topology*) leading to end points or terminal nodes (also referred to as *Operational Taxonomic*

Units; OTUs). OUT's are the focus of investigation when constructing phylogenies, and are usually extant taxa but can also be fossil taxa or individual genes. It is the branching pattern of a phylogeny that defines the relatedness of a set of OTUs, and therefore can be thought of as a hypothesis, which explains the order of evolutionary events through time (e.g. speciation, extinction and gene duplications) that we assume to have occurred.

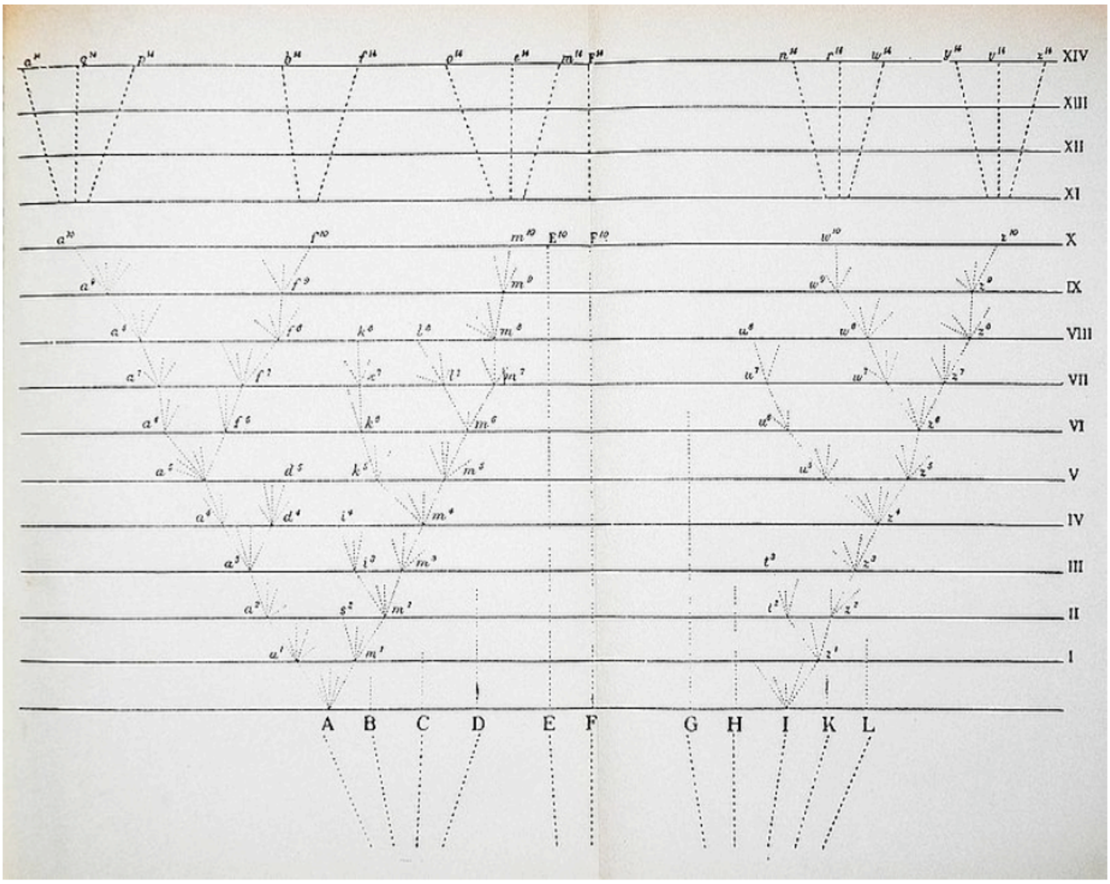


Figure 2.1: Darwin's "Tree of Life". Charles Darwin's only figure illustration from the book "On the origin of Species" (Darwin, 1859).

It is true that a phylogeny will always depict the branching pattern of its OTUs, however there is no clear outline as to what information a tree ought to convey. For instance a phylogeny may or may not display information about the phenotypes of its leaves, it may display branch lengths (*Phylogram*) or it may only display the over all branching pattern without any branch lengths (*Cladogram*). Finally the order of events of evolution may be directed (*Rooted*) i.e. indicate the direction of the evolutionary process, or directionality may not be given at all (*Unrooted*) (see Figure 2.2 for a comparison between a rooted and unrooted phylogenetic network). An unrooted network is basically a summary of the possible interconnections between OTUs; conversely a rooted network is a depiction of evolutionary history.

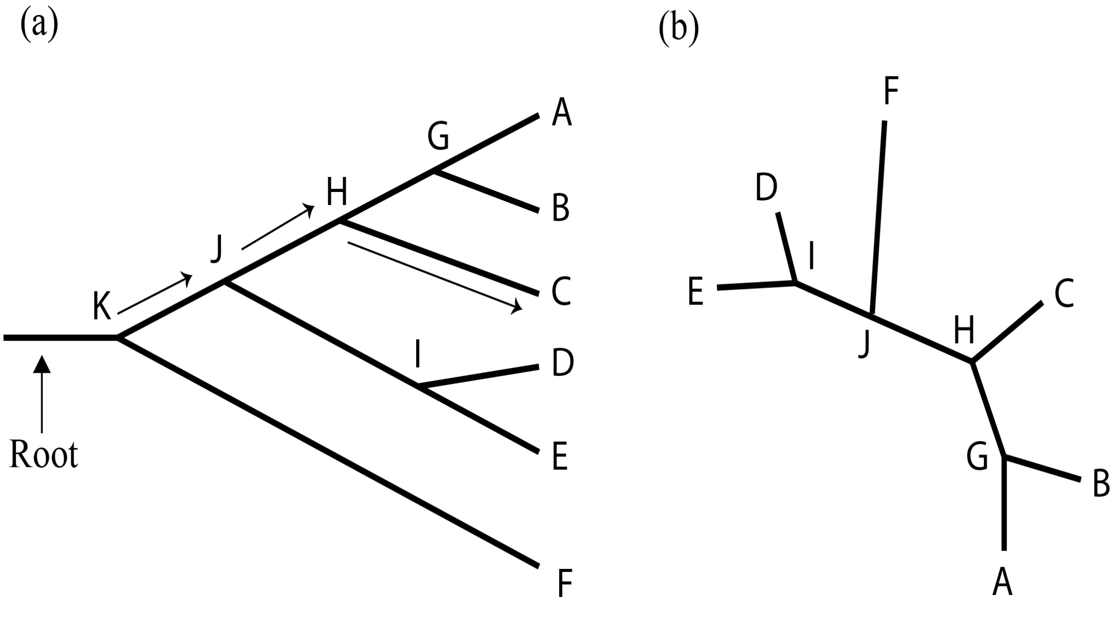


Figure 2.2: A Phylogenetic network depicted as both a rooted and unrooted cladogram. Both networks show the same topology, but the direction of evolutionary change is only evident for the rooted tree. External leaf nodes (OTUs) labelled A-F, internal ancestral nodes labelled G-K. (a) Rooted network ingroup OTUs labelled A-E, outgroup labelled F, root node labelled K. (b) Unrooted network OTUs labelled A-F, the unrooted network does not have a root node (K); therefore there is no outgroup.

The way in which we infer directionality or root a phylogeny is in terms of an *outgroup*, an outgroup is one of the OTUs that is included in the study. This outgroup has the property of being known (or it is believed anyway) to be the most distantly related to all remaining OTUs (*ingroup taxa*) then any of the ingroups are among themselves. The *root* of a given topology is positioned along the branch connecting the outgroup with the ingroup. Correct rooting of a phylogeny is crucial and should not be overlooked as this could lead to downstream biases and topological errors (see section 2.3.4: LBA). Also, it should be stated that in order to convey non-trivial information a phylogenetic tree must contain at least four species one of which can be an outgroup (Telford and Copley, 2011).

Concluding, the phylogenetic tree concept has firmly found its foothold in evolutionary thinking, effectively conveying concepts such as speciation, extinction and the over all tree like pattern of evolution we expect to be observed as a result of descent with modification from a common ancestor. However, evolution is not always tree-like. For example prokaryote evolution has both a vertical and a horizontal component (gene exchange via horizontal gene transfer or (HGT; see Ragan *et al.* 2009). Yet, this thesis is only concerned with vertical evolutionary processes. In this section I will continue by discussing some of the different data types currently used in phylogenetics, and comparing and contrasting their strengths and weaknesses. I will also discuss sources of phylogenetic biases associated with molecular sequence analyses. Finally, I will try and give insights into what we need to consider when drawing inferences from the data: is it best to combine all the evidence in search of a an hypothesis that best explain them all (i.e. "Total evidence": Kluge, 1989) or alternatively is there merit in examining multiple independent data types in the search for corroborating evidence to validate a particular hypothesis?

2.1.1 Homology and Multiple sequence alignment

Before I discuss the common data types used in modern phylogenetics, I must first introduce the concepts on which we base our assumptions of shared evolutionary history, which influence our approach to phylogenetic reconstruction. Common practice when investigating phylogenetic relationships is to begin with a set of species, with each species scored for a number of observable characters (e.g. a morphological matrix, an alignment of molecular sequence data, or combination of the two). One requirement of such an approach is identification of a set of characters (e.g., morphological traits or genes) that are known to be present in the set of OTUs through descent via common ancestry. This necessity of descent through common ancestry for any observable character intended for phylogenetic analysis introduces the idea of homology. The concept of homology has been around since the mid 19th century and forms the basis for modern phylogenetics, introduced by Owen (1843); it is traditionally defined as a “special” case of historical continuity between characters, that have descended, typically with divergence via a shared common ancestry (Patterson, 1988; Wagner, 2007; Shubin *et al.* 2009).

Classically homology was viewed in the context of shared morphological characters. For example a common instance of homology can been seen in tetrapod limb structure; with tetrapod limbs displaying stereotypical arrangement of bones regardless of necessitated function (e.g. walking, swimming, flight (Wagner, 2007)). This idea of homology between morphological characters can be extended beyond its traditional definition to molecular sequence data; two genes are homologous if they descended from the same gene in a common ancestor (regardless if they still retain the same function or the degree of similarity in the nucleotide sequence). When we consider homologous entities (homologs) it is important to distinguish between a homologous *character* (e.g., protein coding gene) and a *state* for that character (e.g. a Proline amino acid at character position six); the reason being that homology resides not in the state but in the character under examination (Fitch, 2000).

In molecular phylogenetics there exist three distinct subtypes of homology. Firstly, orthology is the relationship where two sequences diverge following a speciation event. Orthologs can then portray the "true" phylogeny of the organisms in which the orthologous genes were obtained, a property unique to orthologous sequences. A second case pertaining to homology is paralogy, where two sequences diverge following a duplication event. In this instance, paralogous sequences can diverge whilst remaining in the same organism and therefore cannot be utilized when inferring speciation. Lastly xenology is where homology arises due to interspecies transfer of genetic material (not of grave importance when investigating metazoan relationships; and so is not under consideration in this thesis). Thus when constructing phylogenetic relationships, for the reasons mentioned above, it is of central importance to know *a priori* if sequences under investigation are orthologous or paralogous (Lemey *et al.* 2009).

An essential prerequisite to phylogenetic analysis for a sequence based phylogeny is the comparison of similarity of homologous sequences. This is achieved by constructing a sequence alignment, such that homologous sites form columns in the alignment; a process commonly referred to as a multiple sequence alignment (MSA). A MSA can be thought of as a hypothesis about the homology of residues in molecular sequences. This procedure can be easy when comparing sequences with a high similarity (total number of identical residues divided by the total length of the alignment) but becomes increasingly difficult when sequences have had more time to diverge (i.e. accumulate more mutations relative to one another). There exists a large number of different MSA software programs; e.g. clustalW, Muscle and Prank (Thompson *et al.* 1994; Edgar, 2004; Löytynoja and Goldman, 2005) which have differing algorithms designed to implement a MSA each with there own strengths and weaknesses. I will not compare and contrast different MSA software packages, for a review of currently used algorithms for MSA see (Edgar and Batzoglou, 2006; Notredame, 2007). For the scope of this thesis it is enough to state that the goal, when constructing a MSA, is to identify hypotheses of homology for each residue in a set of sequences. A multiple sequence alignment thus represents a collection of "positional homologies" that are then used as inputs for phylogenetic analyses.

2.2 Data types of phylogeny reconstruction

2.2.1 Role of morphology in modern phylogenetics

Much of what we know about animal taxonomy and phylogeny today is based on classical studies of morphological data (e.g. Arthropods: Snodgrass, 1938; Chordates: Maisey, 1986; Vertebrates: Sillman, 1960). However, in the current era of comparative genomics, researchers now have vast databases of molecular sequence data available to reconstruct phylogenetic relationships from across the three domains of life (Philippe *et al.* 2004; Ciccarelli *et al.* 2006; Cox *et al.* 2008; Dunn *et al.* 2008; Hejnol *et al.* 2009; Rota-Stabelli *et al.* 2011; Campbell *et al.* 2011; Brochier-Armanet *et al.* 2011). This raises the question of morphological utility in modern phylogenetics. Indeed, some have begun to reappraise the role of morphology in today's molecular sequence era (Scotland *et al.* 2003). According to Scotland *et al.* (2003) the utility of morphology in phylogeny should be limited, because of drawbacks relating to ambiguous character definition, homology assignment and lack of useful new morphological characters. For all of these reasons they conclude: "*We view any attempt to include more morphological data in phylogeny reconstruction as inherently problematic*". This viewpoint has been meet with strong criticism by a number of researchers as they consider the reappraisal and damning of morphology based phylogenetics as unfounded (Jenner, 2004; Wiens, 2004; Smith and Turner, 2005). Cogently, Pisani *et al.* (2007) presented numerical results illustrating how the congruence / incongruence of molecular and morphological data is key to assess the likelihood that a given set of phylogenetic relationships might be correct or not.

In relation to morphological data, sequence based analyses have a number of unique beneficial properties such as: efficient data sampling enabled by next generation sequencing, automated pipelines for analyzing data, larger data sets, relatively complex substitution models; there is one limitation to the use of molecular data - the inability to incorporate fossil taxa in phylogeny reconstruction. Over the course of life on earth, it is estimated that the vast majority (~99.9%) of species to ever evolve are now extinct (Novacek and Wheeler, 1992) and so to ignore fossils is comparable to ignoring over 99% of life. It is this aspect that morphological data becomes of essential importance, as fossil taxa comprise the vast majority of all the branches on the tree of life. Morphology becomes extremely useful when elucidating phylogenies of extant taxa that are characterized by short radiations and deep divergences, such as that of the arthropods (Wheeler 1993; Budd 2001; Edgecombe 2010; Rota-Stabelli *et al.* 2011). Morphology helps to estimate the phylogenetic relationships of fossil and extant taxa by incorporating extinct species while at the same time increasing taxon sampling; potentially breaking up long branches that can occur in the absence of such fossil taxa (Donoghue *et al.* 1989; Wills *et al.* 1998).

Another important function of morphology and fossils concerns dating divergence time within a phylogeny. Although methods exists to date divergence times among living taxa using sequences and the molecular clock assumption (Zuckerkandl and Pauling, 1962, Kimura and Ohta 1971), or a relaxed molecular clock methods (e.g. Erwin *et al.* 2011) fossil data are still needed to calibrate dates of divergence, typically done by use of fossil calibration points setting an upper and/or lower bound for the emergence of a clade or species for example. It has been shown that incorrect calibration of molecular clock models due to for example incompleteness of the fossil record can lead to drastic divergence estimation errors (Rodríguez-Trelles *et al.* 2002; Blair Hedges and Kumar, 2004; Peterson *et al.* 2004; Peterson *et al.* 2008).

Not only can morphology help when reconstructing and dating phylogenies, it also allows insight into the emergence of novel bauplaene, and stem group taxa. It could be argued that proper understanding of morphological innovations that characterize early clades to be of particularly noteworthy importance. How else do we bridge the morphological gap when looking at disparate living clades such as arthropods? Some of the most famous fossil discoveries are related to stem groups of early Panarthropod ancestors, such as the iconic Cambrian predator *Anomalocaris* (Whitington and Briggs, 1985) currently thought to be an early arthropod stem group ancestor, or the armored lobopod *Hallucigenia* (Conway-Morris, 1977) now known to be the early ancestor of the extant terrestrial velvet worms. Specifically, it is this ability of fossil taxa to retain and highlight critical combinations of characters, highlighting cases of synapomorphy and plesiomorphy that can change the outcome of analyses based on hypotheses misdirected by homoplasy.

Aside from the primary role of morphology in phylogenetic reconstruction and understanding character evolution throughout the large time span of life on earth, morphology also plays an important role allowing for “reality checks”. This is important as we do not live in an age of infallible molecular phylogenies (Jenner, 2004). These reality checks allow revaluating our findings when met with conflict and ambiguity in other data types such as molecular sequence data. Independent analyses of morphology and molecules facilitate a greater understanding of the underlying processes of divergence and character evolution. Combine this with new powerful methods of (re)analysing fossils currently being pioneered, such as phase-contrast X-ray computer assisted tomography (Dunlop *et al.* 2011); which can now render exquisitely minute detail at scales never obtained before. It needs to be pointed out that one of the main goals of phylogeny reconstruction is in the understanding of the evolution of characters, typifying species, those fundamental units that natural selection acts upon. In essence, morphology is as important today as it was during the earliest investigations of life on earth, long before the appearance of molecular sequence data. The continued utility of morphological understanding in modern phylogenetics therefore cannot and should not be dismissed now or in the foreseeable future.

2.2.2 Utilization of MicroRNAs for phylogeny

In the previous section it was mentioned how much of our knowledge on metazoan interrelationships have been elucidated through our understanding of morphological similarity spanning the taxonomic hierarchy. Many of these relationships have been backed up and corroborated robustly following reanalyses conducted on molecular sequence data. Since early metazoan molecular phylogenetic analyses, most of the relationships have been investigated with a relatively few types of sequence data e.g. ribosomal, mitochondrial and nuclear protein coding genes (PCGs). Under this approach, orthologs are aligned and analysed under an array of available phylogenetic methods and evolutionary models; this has been and will continue to be the norm in molecular phylogenetics, however there are a number of problems usually encountered under this approach. Problems faced in molecular phylogenetics (across all levels of the organismal complexity) range from: differential rates of molecular evolution that can lead to systematic biases like LBA, incorrect identification of orthologs, erroneous sequence alignments, compositional bias of the replacement process, to cases in which divergence of deep nodes characterized by fast radiations result in absence or masking of genuine phylogenetic signal.

During early efforts to resolve difficult nodes within Metazoa, originally it was hoped that by simply adding greater numbers of sequences to analyses would lead to increased phylogenetic resolution; however this approach alone is now known to be insufficient (Philippe *et al.* 2011b). Essentially, the problem of low phylogenetic resolution resides in the pervasiveness of homoplasy (similarity not caused by shared ancestry but convergent evolution) in current molecular phylogenetic data types, which cannot be fully accounted for by current models of evolution commonly applied to traditional molecular sequence data. Therefore identification of data sets that minimise homoplasy as much as possible should provide the greatest hope for resolving intractable phylogenetic relationships (Sperling and Peterson, 2009), while also providing an additional data set to test hypotheses of metazoan evolution that may not be independently corroborated by both molecular and morphological data.

One of the main aims of this thesis is the utilization of novel data, characterized by low levels of homoplasy, to investigate competing hypotheses of arthropod evolution. A data type that has shown promise in fulfilling the goal of low levels of homoplasy, also with the property of having characters that arise frequently enough to record divergences across most levels of the taxonomic hierarchy is that of the recently discovered class of translational regulatory elements. These regulatory elements called microRNAs (miRNAs) are small ~22 nucleotide (nt) genomically encoded non-coding RNA genes that function as negative regulators of messenger RNA (mRNAs) expression by binding to regions of a mRNA 3' untranslated region (UTR). MicroRNAs subject a mRNA to catalytic cleavage or translational inhibition (depending on the degree of complementary nucleotide binding).

MicroRNAs were originally discovered through investigations of developmental timing in *Caenorhabditis elegans,* with the miRNA *lin-4* found to negatively regulate the protein coding gene *lin-14* (Lee *et al.* 1993). Following the initial discovery of miRNAs and their regulatory role in developmental timing, it soon also became clear that miRNAs held promise for their utilization as phylogenetic markers. In an early study into miRNA distribution across Bilateria, it was shown that the mature sequence of another early discovered miRNA, the miRNA *let-7* (Reinhart *et al.* 2000) displayed a high degree of conservation between diverse organisms like nematodes, fruitflies and Humans, while also present within every Protostome and Deuterostome investigated (Pasquinelli *et al.* 2000). Since initial investigations into miRNA distribution and conservation across Bilateria, many studies have further compounded the utility of miRNAs (Sempere *et al.* 2006; Sempere *et al.* 2007; Wheeler *et al.* 2009) as genuinely valuable independent markers for phylogenetic reconstruction. It has now become evident that miRNAs have a number of unique properties that enable them to be used in tackling phylogenetic questions; difficult questions that otherwise are yet to be fully resolved with traditional molecular sequence data (Sperling and Peterson, 2009). I must note here that the utility of miRNAs in phylogenetic reconstruction is intimately linked to their mode of biogenesis, degree of sequence conservation, mode of translational inhibition and role of regulating gene expression throughout most of an organism's life span. For details of miRNA biogenesis and mode of action please see section (4.1.1) of Chapter 4.

From investigations into the evolution and conservation of miRNA families throughout Metazoa it has been shown that miRNAs have four unique properties that facilitate their use in phylogeny reconstruction (Sperling and Peterson, 2009; Tarver et al. 2012). These properties are as follows; (*i*) miRNA families are continuously added to genomes throughout time, (*ii*) secondary loss is rare once acquired within a genome, (*iii*) Once acquired the mature (~22 nt effecter sequence) miRNA sequence accumulates mutations very slowly, and (*iv*) There is a massively low probability of independent convergent evolution of any particular miRNA. Due to these properties, miRNAs have beneficial qualities that can overcome some of the shortcomings of traditional data types of phylogeny reconstruction.

One of the most utilized molecular markers for reconstructing deep divergence events are ribosomal genes (e.g. 16s rRNA, 18s rRNA and 28s rRNA), due to their relatively slow rate of evolution imparted by their functional constraints (Fox, 2010). Ribosomal RNA genes have been heavily used to reconstruct relationships from virtually all branches of the tree of life (Woese *et al.* 1990), with metazoan phylogeny being no exception (Giribet *et al.* 1996; Giribet and Ribera, 2000; Telford *et al.* 2003; Halanych, 2004). As the rate of evolution in ribosomal genes is slow, real phylogenetic signal is usually maintained over longer periods of time. Wheeler *et al.* (2009) however showed that compared to the rate of substitution within the mature sequence of miRNAs, slowly evolving ribosomal genes actually accumulate mutations over twice as fast. This extremely slow rate of molecular evolution in miRNA families allows the identification of homologous miRNAs that evolved independently over long periods of time with much greater certainty, thus reducing the chance of misidentification of miRNAs due to homoplasy. Furthermore, it has been shown that miRNAs were some of the most conserved genetic elements in the genome, with most miRNAs shared between both flies and higher mammals showing no substitutions to the mature sequence (Sempere *et al.* 2006).

Despite being shown how miRNAs are continuously added to genomes throughout time (Sempere *et al.* 2006; Sempere *et al.* 2007; Wheeler *et al.* 2009), instances of secondary loss of miRNA families can and do occur (Philippe *et al.* 2011a). However when regarding the phylogenetic utility of miRNAs we should consider the rate of loss compared to the rate of miRNA gain; as was demonstrated in the study of 139 miRNA families distributed throughout the Metazoa (Sperling and Peterson, 2009) in which 132 miRNA gains were found in contrast to only 7 losses; thus again highlighting the degree of phylogenetic conservation and utility of miRNAs over great evolutionary distances. The degree of homoplasy for any particular miRNA then rests on two possible factors, firstly as discussed previously cases of secondary loss or substitution within the mature nucleotide sequence, both of which inhibit our ability to trace a miRNAs true orthology, and secondly the independent evolution of the same miRNA in separate taxa. Luckily, independent evolution of the same miRNA seems highly unlikely, this low probability is due to the constraints imposed by the mode of miRNA biogenesis. The mode of miRNA biogenesis dictates that each precursor miRNA (~60-80 nt sequence containing the mature miRNA) must be able to fold with a free energy value of -20 kcal/mol into a stable hairpin loop. In addition the mature miRNA sequence must be located on one of the hairpin arms close enough to the hairpin loop so that the biogenesis machinery can process and cleave out the mature miRNA. Combining these constraints of miRNA biogenesis with the likelihood of any particular 22 nt sequence emerging by chance; estimated to be once per every 1.76×10^{13} nucleotides or once every 5,864 human-genome-sized chunks of DNA (Sperling and Peterson, 2009), makes convergent evolution of any particular miRNA to be extremely unlikely.

In summary, properties of miRNA evolution and conservation (applicability over a wide phylogenetic range, low rate of substitution to mature miRNA, and low probability of convergent evolution) allow miRNA phylogeny reconstruction to be conducted in a binary fashion; involving simply indentifying presence vs. absence of a particular miRNA within different organisms. Currently there already exists a number of studies that have embraced the use of miRNAs for phylogenetics to tackle a wide range of problematic nodes within the metazoan tree of life, e.g. sponges (Sperling *et al.* 2010), annelids (Sperling *et al.* 2009b), vertebrates (Heimberg *et al.*

2010), brachiopods (Sperling *et al.* 2011), and presented in this thesis; Arthropoda (Rota-Stabelli *et al.* 2011) and Panarthropoda (Campbell *et al.* 2011). Combining these unique properties with the ease of identification of miRNAs (Bioinformatic searching, complete genomes, Northern analysis, or next generation sequencing of small rRNA libraries) for virtually all Eumetazoa provides a Systematist with another independent and importantly homoplasy-low data set in which to test competing hypotheses of metazoan evolution.

2.2.3 Phylogenomics and Gene concatenation

The field of phylogenomics originally referred to by Eisen (1998) and O'Brien and Stanyon (1999) owes its very existence to the revolutionary change in the way in which we study genomes using genome sequencing. Since the very first genome to be sequenced was obtained nearly two decades ago via whole genome shotgun sequencing (Fleischmann *et al.* 1995) there have been major advancements in the field of DNA sequencing technology. It is now possible to sequence an organisms genome quickly and cost effectively via next-generation sequencing (NGS) technologies (Metzker, 2009), with the recent Ion torrent sequencing technology one of the latest leading innovations (Rothberg *et al.* 2011). The number of genomes available for species spanning the entire tree of life has now reached a level hardly imaginable when the very first complete genomes were sequenced. The term phylogenomics incorporates the interplay of genome wide evidence to study molecular biology and evolution; specifically phylogenomics has been utilized in investigating the mechanisms of molecular evolution and to a lesser degree for inferring phylogenetic relationships (Philippe *et al.* 2005a).

A shifting paradigm in phylogenetic investigation due to the onset of the phylogenomic era concerns the move from phylogenetic analyses of organisms using single gene sequences or limited numbers of gene sequences to data sets comprising multiple thousands of DNA or translated amino acid positions. The increase in the amount of sequence data available to study organismal relationships has largely alleviated the problem of sampling or stochastic error seen in many of the early molecular phylogenetic studies (Delsuc *et al.* 2005; Kelchner *et al.* 2007). It could be said that eliminating stochastic error is one of the major achievements of the phylogenomic approach; for instance phylogenomic analyses have lead to the confirmation of the monophyletic status of many of the higher metazoan clades such as Ecdysozoa, Lophotochozoa, Protostomia, and Deuterostomia (Telford and Copley, 2011).

There are two main ways in which phylogenomics is used. The first approach is genome structure analysis (e.g., gene order, intron location and/or presence vs. absence, protein domain structure) while the second is based on primary sequence level analysis (Philippe *et al.* 2005a). In this thesis, phylogenomic analysis of the latter type are presented, in the form of large concatenated sequence alignments also known as supermatrices, which have been applied to study the relationships of the Arthropoda and their closest relatives within Ecdysozoa.

The supermatrix approach also referred to as ‘combined analysis’ or ‘simultaneous analysis’ involves combining all systematic characters into a single large phylogenetic matrix and then analysing all the characters for all taxa simultaneously (see de Queiroz and Gatesy, 2007). This method is similar to the approach of character analysis via ‘total evidence’ (also known as ‘character congruence’) as defined by Kluge (1989) in which he advocated the combined use of all available evidence

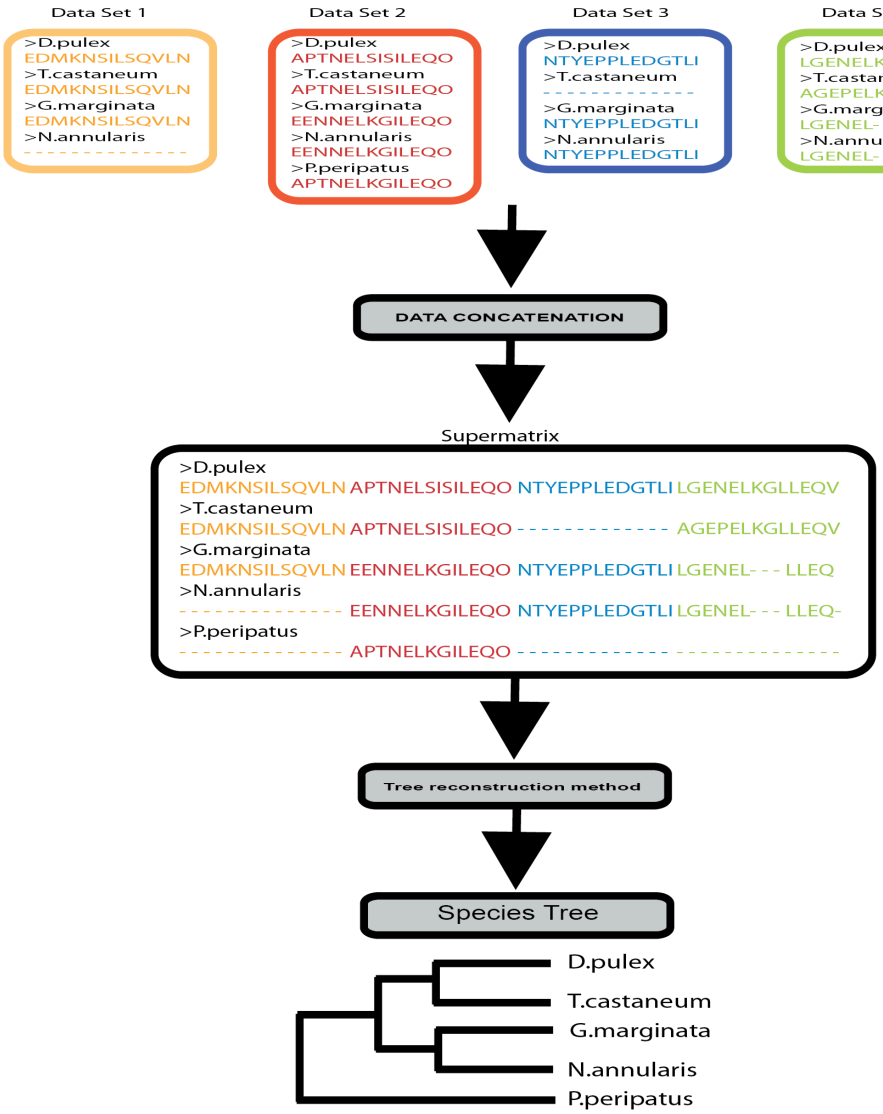


Figure 2.3: Gene concatenation pipeline. Individual data sets, which can have non-overlapping taxon sets, are joined together into a single large supermatrix containing each data set. The supermatrix is then analysed using a single tree reconstruction method, resulting in a species phylogeny.

(e.g. sequence alignments, morphological matrixes, behavioural data matrixes) into a single phylogenetic analysis. With gene concatenation, once all individual data matrixes are concatenated together, analyses are then conducted on the resulting concatenated sequence matrix under a single tree reconstruction method (see Figure 2.3). This can be seen to be a better approximation of phylogenetic relationships as it uses character evidence more fully when estimating a phylogeny (de Queiroz and

Gatesy, 2007). In other words the combined data sets enable the phylogenetic signal to assert itself more strongly over the noise (assuming there is only one phylogenetic signal in the data).

One of the advantages when using the gene concatenation approach is that there is no need to have completely overlapping sets of input taxa or sequences, i.e. the gene concatenation method allows for the presence of missing data (encoded in the form of gaps "-" for characters, or '?' for taxa); but see Sanderson *et al.* (2010). The effect of the amount of missing data has been a source of much debate, with some authors insisting that the numbers of missing characters is not as important as the quality of the numbers of characters present for any species (Wiens, 2006). However, the question still needs to be resolved, as we still do not really know whether adding large amounts of missing data can compromise phylogenetic reconstruction. Some have concluded that the effect of missing data in large phylogenomic sized data sets is limited as species for which sequence information is incomplete can be outweighed by the number of informative characters present in these phylogenomic data sets (Wiens, 2003; Philippe *et al.* 2004; Delsuc *et al.* 2005). However such conclusions were not made definitively, with others insisting this question remains to be tested thoroughly.

An important benefit of using gene concatenation resides in the ability to use probabilistic tree reconstruction methods that incorporate parameter rich mixture models (Delsuc *et al.* 2005). However one of the major limitations is that analyses are conducted using one single method of phylogeny reconstruction. Evolutionary reconstruction methods assume a treelike structure to evolutionary history and further assume the same branching history is common to all characters included in the analysis (Bull *et al.* 1993). This has implications when analysing multiple data sets (e.g. single genes) in combination, if these individual data sets violate the assumption of shared evolutionary histories (de Queiroz *et al.* 2007) as would be seen for some different gene trees.

2.3 Sources of phylogenetic error

The field of molecular phylogenetics is currently undergoing a renaissance, methodologies are continuously improving, now we have increasingly sophisticated models of evolution that for example explicitly take into consideration the biasing effect of compositional heterogeneity (Blanquart and Lartillot, 2008). Furthermore, powerful software implementations of phylogenetic inference methods like Bayesian and Maximum likelihood (Ronquist and Hulsenbeck, 2003; Stamatakis, 2006; Lartillot *et al.* 2009) allow us to analyse very large molecular data sets (e.g. Smith *et al.* 2011). Despite these advances in methodology and technology it is still clear that molecular phylogeny can be complicated by the presence of artifacts of tree reconstruction. There are two types of phylogenetic error that can affect molecular phylogenies; these are stochastic error and systematic error. Stochastic error continues to be a problem in modern phylogenetics, particularly in studies based on small numbers of genes. Stochastic (or sampling) error affects all methods of tree reconstruction, however the current “standard” of large phylogenomic sized data sets have greatly diminished its effect. Systematic error on the other hand is persistent and pervasive in current day molecular phylogenetics, and is not a source of error that can be alleviated by simply analysing huge data sets with large numbers of sampled genes and taxa. Here I will discuss some of the most prevalent sources of phylogenetic error, whilst also detailing possible methods of alleviating those errors.

2.3.1 Systematic error

Systematic error occurs when phylogenetic reconstruction methods fail to be consistent; statistically speaking, a method is said to be consistent when you move towards the correct answer as more data are considered (Philippe *et al.* 2005a). All phylogenetic methods make assumptions about the processes of evolution affecting sequences as they evolve. It is here that a method can become inconsistent when the underlying assumptions fail to describe the data, usually due to violations of the underlying model employed (Delsuc *et al.* 2005). In most cases, model violations occur and will generate different degrees of phylogenetic noise (random phylogenetic signal) that competes with the phylogenetic signal. The degree of influence of the random phylogenetic noise will depend on the strength of the true inherent underlying phylogenetic signal. Predominately in cases of ancient divergence (e.g. Arthropoda, Nematoda, Mollusca) where historical phylogenetic signal may be weak due to short radiation times masked by subsequent within lineage substitutions diluting the historical signal, can lead to cases of phylogenetic error. This also increases the evolutionary rates of convergence and site reversals. The three most predominant sources of systematic inconsistency in molecular sequence data are as follows: Compositional bias, Long Branch Attraction and Heterotachy; which will be discussed in the following sections.

2.3.2 Compositional bias and Heterotachy

Compositional bias occurs when sequences are erroneously clustered together due to non-historical similarities of the respective nucleotide or amino acid compositions, which can occur when the evolutionary models used assume homogeneity of the data (Foster, 2004; Nesnidal *et al.* 2010). Compositional heterogeneity first identified as a problem by Hasegawa *et al.* (1993) and Van Den Bussche *et al.* (1998) was thought largely to be restricted to nucleotide sequences (Loomis and Smith, 1990; Lockhart *et al.* 1992). However it has been shown that there exists a correlation between the AT/GC bias present in nucleotides and the content of AT- and GC- codons and their corresponding encoded amino acids (Foster *et al.* 1997; Foster and Hickey, 1999). Strand asymmetry, a phenomenon correlated with the origin and direction of mtDNA replication, has also been shown to be another source of compositional bias (Rota-Stabelli and Telford, 2008). However this kind of compositional bias is limited to phylogenomic analysis of mitochondrial and bacterial data sets (Rota-Stabelli and Telford, 2008), which will not be addressed in this thesis.

One of the main problems faced currently, is the analysis of phylogenomic sized data sets that include large numbers of compositionally heterogeneous sequences (Jermiin *et al.* 2004) which can be a major problem when reconstructing metazoan relationships as model assumptions can be violated substantially (Lartillot and Philippe, 2008; Rota-Stabelli and Telford, 2008). There have been a number of methods developed in an attempt to account for the biasing affects of compositional heterogeneity. For instance the use of a Log-Det transformation (Lockhart *et al.* 1994) has been said to deal affectively with compositional heterogeneity (Jermiin *et al.* 2004). One of the most popular methods for dealing with compositional heterogeneity is to employ general models of nucleotide (or amino acid) substitution that incorporate additional parameters that attempt to accommodate composition bias, such as the one introduced by Foster (2004) under a Bayesian framework.

The relatively recent development of the heterogeneous model CAT represents an important step taken to combat phylogenetic artifacts due to compositional bias, achieved by the CAT model relaxing the assumption of homogeneity among sites. The CAT model empirically assumes the existence of distinct classes of amino acids that can then be assigned into categories (based on equilibrium frequencies of the 20 amino acids) which best describes their rate of substitution (see also section 1.2.3 of preceding Chapter). Further to this a recent derivation of the CAT model has been developed, known as the CAT-BP (Blanquart and Lartillot, 2008) which has also been shown to be highly effective at accommodating composition bias between lineages by introducing ‘break points’ along the branches of a topology at which the composition is allowed to vary. Unfortunately there is a high computational burden when implementing models like the CAT-BP, with some analyses taking weeks or months to converge or sometimes not converging at all (Nesnidal *et al.* 2010).

In addition to the problems faced when analysing data sets with sequences that are compositionally heterogeneous, is the problem of Heterotachy. Heterotachy, as defined by Philippe and Lopez (2001) pertains to the variable rate at which a site in a gene sequence evolves over time. Due to functional constraints on a protein it was soon realised that the rate of substitution along a sequence was not uniformly distributed. An early attempt to take this rate change among sites into account was via the use of a gamma (Γ) distribution (Uzzell and Corbin, 1971). However the use of a gamma-distribution, which is also seen as being a “homotachous” model (Lopez *et al.* 2002) does not fully account for evolutionary processes in real data as functional constraints not only impose different rates among sites, but also may change the rate of substitution within a given site over evolutionary time, a phenomenon known as heterotachy.

The contribution of heterotachy towards generation of phylogenetic artifacts is undoubtably significant (Philippe *et al.* 2005a) so it is surprising to note that many of the current evolutionary models and phylogenetic reconstruction software implementing those models assume stationarity of the replacement rate. Accordingly, models have been developed that have somewhat addressed the problem of heterotachy, notably the covarion model (Fitch and Markowitz, 1970); in this model only a fraction of sites ("*c*" or "covarions") are allowed to accept mutations.

The covarion model, and others like it (e.g. mixture of branch length (MBL) models Kolaczkowski and Thornton, 2004; Spencer *et al.* 2005) have marginally addressed the problem faced when trying to model heterotachy. However, realistically it is very computationally expensive to do, as the number of free parameters associated with modelling independent rates of substitution for each site across all taxa generally becomes very large. In the face of this computational burden, site independent approaches have been introduced which alleviate the problem to some degree. These models use variations of the gamma model that can account for the variability of site rates over time.

Covarion-like models such as the hidden Markov model of Tuffley and Steel (1998) allows a site to be either variant or invariant, unlike the original covarion or MBL models Covarion-like has a particular beneficial property, in that it warrants the introduction of only two additional parameters (i.e. site rate switching from "on to off" and *vice-versa*). Yet it has been said that the covarion-like model of Tuffley and

Steel is limited by the assumption of rate shifts being site-independent (Zhou *et al.* 2007); a property not expected when considering sudden selective pressure changes.

2.3.3 Long Branch Attraction

The most prevalent and important source of systematic artifacts in phylogeny reconstruction is undoubtedly Long-branch attraction (LBA). LBA is a tree reconstruction error caused by undetected instances of convergent evolution (i.e. homoplasy). This results in clustering of branches regardless of the true evolutionary history. LBA can occur in cases where a species or subset of rapidly evolving species is present, or when one or more species are very evolutionarily distant from the remaining taxa (or a combination of both cases). LBA was first indentified as a problem in phylogenetics by Felsenstein (1978). Felsenstein demonstrated using a four-taxon tree that parsimony and compatibility methods could also become inconsistent (i.e. move towards a wrong answer with more certainty as more data are added) when evolutionary rates differ widely among branches. Expanding on the conditions in which parsimony can become inconsistent due to LBA, Hendy and Penny (1989) showed that not only unequal rates of evolution but differing branch lengths could also lead parsimony to fall victim to LBA. Furthermore unequal branch lengths could be caused by unequal rates or as a consequence of a non-symmetric tree. The shape of a topology and the occurrence of LBA were again demonstrated by Kim (1996), showing that even if branch lengths were equal LBA could still affect the resulting topology.

With Long Branch attraction first described under maximum parsimony, it was thought that the reduced occurrence of LBA would be achieved via the implementation of probabilistic methods like ML and Bayesian inference (Felsenstein, 1973; Yang, 1996). However, ML and Bayesian analysis are consistent if the underlying (assumed) substitution model is correct. However, ML and Bayesian analysis can also be affected by LBA, i.e. when the underlying model of evolution fits the data poorly and is widely underparameterized, therefore such a model would be a simplification of real (unknown) evolutionary processes (Philippe and Germot, 2000; Sullivan and Swofford, 2001; Inagaki *et al.* 2004). So since no real world sequence data can be expected to evolve via the oversimplified processes assumed under ML models, consistency alone does not warrant the selection of ML over parsimony. Although it has been demonstrated that ML and Bayesian inference are quite robust to violation of their assumptions (Gaut and Lewis, 1995; Sullivan and Swofford, 2001) (i.e., even when using models that do not fit the data well), ML and Bayesian analyses tend to outperform Parsimony. With this in mind it is not surprising that most phylogeneticists consider inferences made with probabilistic methods to be more robust to the effects of LBA (Bruno and Halpern, 1999; Swofford *et al.* 2001; Whelan *et al.* 2001; Philippe *et al.* 2005a).

Since the recognition of the prevalence of LBA in molecular phylogeny, many methods of reducing the artifactual effects of LBA have been developed. The most intuitive of approaches to reduce LBA is by the inclusion of additional taxa. The importance of taxon sampling for molecular phylogenetic inference was advocated early on (Lecointre *et al.* 1993) with the inclusion of additional taxa first applied to the problem of LBA by Hendy and Penny (1989). Increased taxon sampling was subsequently found to be an effective method of reducing LBA in many studies (e.g.

Swofford *et al.* 1996; Page and Holmes, 1998; Giribet and Ribera, 1998; Pollock *et al.* 2002; Dohrmann *et al.* 2006; Rota-Stabelli and Telford, 2008; Pick *et al.* 2010; Rota-Stabelli *et al.* 2010; Sperling *et al.* 2010). The importance of additional taxa when combating LBA is that additional taxa can have the effect of breaking up long branches (Hillis, 1998). Conversely, some studies have led some authors to suggest that the inclusion of additional taxa can be detrimental. However, this is only when additional "long branched" taxa are added, resulting in the exacerbation of the problem of LBA (Poe and Swofford, 1999; Rannala *et al.* 1998). The potential benefit of additional taxa has also been viewed as less important when compared to increasing sequence length (Rossenberg and Kumar, 2001); however this view has been criticized strongly in favor of taxon sampling (Pollock *et al.* 2002; Holton and Pisani, 2010) while increased taxon sampling has also been shown to benefit genomic scale studies greatly (Holton and Pisani, 2010). Finally, methodologies for identifying branches that would benefit from increased taxon sampling have been developed (Goldman, 1998; Massingham and Goldman, 2000).

In addition to inclusion of extra taxa to break up long branches, another method to alleviate LBA is to optimize outgroup selection (Rota-Stabelli and Telford, 2008). It is now understood that analyses of data sets that include divergent outgroups can artifactually attract long branched ingroup species with higher rates of evolution (Philippe and Laurent, 1998). When including phylogenetically close but genetically distant outgroup species, problems can arise and result in artifacts being generated. These artifacts can be due to problems of difficulty in sequence alignment to outgroups with accelerated substitutions, loss of signal, compositional heterogeneity, and random attraction of fast evolving ingroup species towards the root of a phylogeny (Foster and Hickey, 1999). In effect, selection of a very distantly related or highly divergent outgroups is akin to selecting a randomized, fully saturated sequence (with respect to model selection) (Wheeler *et al.* 1990). Methods for selecting optimal outgroups (Sanderson and Shaffer, 2002) representing phylogenetically closely related species have had success when applied to difficult phylogenetic problems. For example it was shown that the choice of an outgroup closely related to Arthropoda had drastic effects when recovering topologies representing the two main competing hypotheses on internal arthropod phylogeny (Rota-Stabelli and Telford, 2008). Thus robust outgroup selection can be evaluated by adding different divergent outgroups to determine if the root placement changes the resulting ingroup topology.

Aside from selecting optimal outgroup representatives; selective ingroup sampling is also a bona fide method to combat LBA. The way in which this is approached is to evaluate the evolutionary rate across the entire data set and then select taxon representatives that typify the slower more homogeneous evolutionary rate. This strategy is excellently demonstrated in the study of (Aguinaldo *et al.* 1997) in which the differential selection between a slow and fast evolving nematode species resulted in the definition of the clade Ecdysozoa, subsequently heralding the emergence of the new animal phylogeny. This method however is not always applicable, for instance when representative species with a slower evolutionary rate cannot be identified.

Lastly, a method of reducing the biasing effects of LBA is by the exclusion of character positions from the data. For example it is common practice to remove third codon positions (e.g. Regier and Shultz, 2001; Cameron *et al.* 2004; Negrisolo *et al.*

2004; Regier *et al.* 2010) from nucleotide alignments. This is because third positions are highly likely to be oversaturated as a consequence of the genetic code redundancy (Lemey *et al.* 2009). Removal of fast evolving sites has been approached using different methods. One of the most commonly used methods is based on parsimony. This approach, known as the slow-fast (SF) method (Brinkman and Philippe, 1999) identifies fast evolving sites according to *a priori* knowledge of monophyletic clades, these sites are then subsequently removed from the alignment. This method is particularly useful (Sperling *et al.* 2009a) when data sets have a limited taxon sampling, outgroup selection is not optimal or when species within the phylogeny are evolving at differential rates. The SF method is limited by analyses in which the monophyletic status of a clade is not known or is defined incorrectly. Furthermore use of methods like SF raises the question of when to stop removing sites, a subjective question that is difficult to answer, as there is no definitive cut-off. Instead monitoring the effect of progressive removal of sites and its effect on topology are required on a case-by-case basis. Alternatively the compatibility method of Pisani (2004) which uses binary character compatibility (Le Quesne, 1969) or a more recent method of Cummins and McInerney (2011) which scores sites and categorizes them according to their degree of similarity; alleviates the subjectivity and necessity of *a priori* based knowledge of topology as is inherent in the aforementioned SF method.

2.4 Congruence as a proxy for accuracy

Modern phylogenetics has access to a large array of different data types, all of which can be seen as having their own set of beneficial properties and uses for phylogeny reconstruction (Pisani *et al.* 2007). In regard to molecular sequence data, its utility no doubt stems (at least in part) from the presence of vastly greater numbers of observable characters to analyze (Scotland *et al.* 2003). It has been argued that morphological data types still have their utility and cannot be seen as being less important then molecular data (Jenner, 2004). This was most forcibly confirmed by Pisani *et al.* (2007), which demonstrated that congruence of molecules and morphology is a better proxy of phylogenetic accuracy then the congruence of alternative sequence data. In one of the papers connected to this thesis (Rota-Stabelli *et al.* 2011) congruence of molecular and morphological data was partially investigated. However, the work presented in this thesis is mostly concerned with the analysis of molecular datasets of genomic scale. More precisely, a number of different data sources were investigated, specifically data from protein coding genes (assembled from ESTs), data sets of SSU and LSU rRNA (assembled with ribosomal secondary structure), and microRNAs (which represent a new class of genomic characters which will be introduced in Chapter 4). Accuracy of our results was thus investigated through the congruence of the above mentioned data types.

From a philosophical perspective, the approach of this thesis is based on the analysis of patterns of congruence and incongruence of trees inferred from the above-mentioned data sets, an approach which can be seen as a form of "phylogenetic consilience". William Whewell introduced the concept of consilience in his work *The Philosophy of the Inductive Sciences* (1840), in which he states *"The Consilience of Inductions takes place when an induction, obtained from one class of facts, coincides with an induction obtained from another different class. Thus Consilience is a test of the truth of the Theory in which it occurs"*. Consilience (see also Wilson, 1988) is born out of human condition; it is concerned with testing the truth of a theory via the corroboration (or unification) of knowledge, thereby linking facts and fact-based theory across independent data sets to create a common groundwork of explanation. The overall opinion presented in this thesis is that in order to resolve any difficult phylogenetic problem, taking the evolution of Arthropods as a prime example, it is necessary to explain the mechanisms of evolution and the branching of a phylogeny in terms of multiple independent lines of evidence, thereby following the idea of consilience.

Chapter 3

A phylogenomic approach to resolve ecdysozoan phylogeny

3.1 Overview

3.1.1 Ecdysozoa not Coelomata

The clade known as Ecdysozoa, comprises a total of eight Phyla, these are: arthropods, onychophorans, tardigrades, priapulids, kinorhynchs, loriciferans, nematodes and nematomorphs. Ecdysozoa is the largest and most specious clade of animals to ever exist, with around ~1.5 million species described currently and a further 4.5 million living species estimated (Chapman, 2009). In a new series of annual reports (SOS: The State of Observed species; published by Arizona state university in 2011) into the diversity and cataloguing of old and newly discovered species, reports that arthropods and nematodes alone comprise in total 1,202,723 species. This large number of species, which fill a diverse array of ecological niches, is surprising when considered in terms of overall bauplan diversity. Ecdysozoan bauplaene are rather conservative, being constrained either to a segmented appendage bearing (e.g. centipedes, decapods, insects) or more worm-like (e.g. nematodes, priapulids) with an anterior circumoesophageal nerve ring and a terminal mouth (Telford *et al.* 2008). A number of morphological synapomorphies unite Ecdysozoa;

these include a lack of locomotory cilia, lack of primary larva, terminal mouth, the HRP antigen in the nervous system and a conserved mitochondrial gene order (Schmidt-Rhaesa, 1998). However, occurrence of repetitive moulting and growth cycles or 'ecdysis' is the most characteristically cited synapomorphy (Schmidt-Rhaesa, 1998), giving rise to the name Ecdysozoa.

Before the Ecdysozoa hypothesis was first proposed, the major hypothesis for the relationships of protostome phyla like arthropods and annelid worms was based on recognition of segmentation in these groups, which became known as the Articulata hypothesis (Anderson 1979; Wheeler *et al.* 1993). The Articulata hypothesis describes the pattern of emergence of increasing complexity, and posits morphological complexity in bilaterian protostomes moves from a segmented worm like ancestor with a fluid filled cavity (as seen in molluscs and annelids for example) towards a more complex segmented body with articulated appendages characteristic of arthropods. The Articulata grade of organisation was based in terms of a larger assemblage of bilaterian metazoans that possessed (to some degree) a fluid filled cavity or 'coelom'; this assemblage was referred to as the Ceolomata hypothesis (Hyman, 1940).

Under the Coelomata hypothesis, bilaterian groups with an absence of a coelom (acoelomates) like the Platyhelminthes and Nemertinea are examples of the simplest grade of coelom organisation; representing some of the earliest bilaterian groups to emerge. From these acoelomate groups, phyla such as nematodes, kinorhynchs and priapulids that possess a partial coelomic cavity (pseudocoelomates) then evolved. Finally a true fluid filled coelomic cavity developed, as is present in phyla such as annelids, molluscs, cephalopods, and arthropods (coelomates).

Despite a long history of phylogenetic study of the Metazoa, the Ecdysozoa is a relatively recent clade first proposed after the study of 18s SSU rRNA (Aguinaldo *et al.* 1997). The study of Aguinaldo *et al.* was the first study to use molecular sequence data to refute the coelomate hypothesis. In their analysis selective taxon sampling allowed them to identify phylogenetic reconstruction artifacts in previous molecular analyses (e.g. Winnepenninckx *et al.* 1995). The particular phylogenetic artifact highlighted by Aguinaldo *et al.* was the problem of Long-Branch attraction (LBA) which they showed to be prevalent in previous molecular analyses that utilized fast evolving species of nematodes (e.g. *C. elegans*). The use of such fast species resulted in the placement of Nematoda (which lack a true coelom) towards the root of the Bilateria, thus supporting the Coelomate hypothesis. However, upon use of shorter branched slowly evolving nematodes (*Trichinella* sp.) the analysis of Aguinaldo *et al.* resulted in nematodes no longer positioning near the root of Bilateria, instead moving inside a clade along with phyla like Arthropods, Kinorhynchs and Priapulids. This grouping then placed pseudocoelomate nematodes in a close relationship with other phyla that possessed a true coelom (e.g. Arthropods) thereby rejecting the Coelomata hypothesis, which posits a simple linear stepwise rise in morphological complexity.

Currently the grouping of ecdysozoan phyla has received much support from a broad range of evidence; ranging from morphology, development, phylogenomics and complete genomes and MicroRNAs (Eernisse *et al.* 1992; Schmidt-Rhaesa, 1998; de Rosa *et al.* 1999; Haase *et al.* 2001; Ruiz-Trillo *et al.* 2002; Philippe *et al.* 2005b; Sempere *et al.* 2007; Dunn *et al.* 2008; Holton and Pisani 2010; Campbell *et al.* 2011). However, there are a sizable number of publications that have supported a view of metazoan evolution in accordance with the Coelomata hypothesis.

The question of Coelomata versus a monophyletic Ecdysozoa has been a very contentious issue over the past decade, with a handful of molecular analyses recovering Coelomata over the more recently proposed Ecdysozoa hypothesis. Much of the controversy between independent molecular analyses derives from large-scale genomic wide analyses that bolster the benefit of large-scale gene sampling (Blair *et al.* 2002; Copley *et al.* 2004; Wolf *et al.* 2004; Dopazo and Dopazo, 2005; Philippe *et al.* 2005b; Rogozin *et al.* 2007). For instance, the analyses by Wolf *et al.* (2004) in light of its extremely large gene sampling (~500 genes) found support for Coelomata. This may seem convincing, however the authors themselves noted that much of the support for Coelomata in their analyses derived from phylogenetic noise i.e. LBA; the exact same problem addressed in the seminal paper of Aguinaldo *et al.* (1997). The problem inherent in many of these large scale gene sampling analyses that support Coelomata is the problem of limited taxon sampling, which when coupled with use of fast evolving species can exacerbate the systematic artifact LBA (Philippe *et al.* 2005a, Sperling *et al.* 2009a). Indeed, recently Holton and Pisani (2010) showed the potential for LBA to alter the recovery of Coelomata over Ecdysozoa when phylogenomic data sets (~1,900 genes or greater) and complete genomes were analyzed with different distantly related outgroups (fungal outgroup vs. a cnidarian outgroup).

It seems that the view of bilaterian evolution has now predominantly moved away from Coelomata, with support for Ecdysozoa reaching a turning point. Although the monophyletic status of Ecdysozoa is now generally accepted (Kumar *et al.* 2011), many controversies still remain on the interrelationships of its constituent phyla (Giribet and Ribera, 1998; Peterson and Eernisse, 2001; Mallatt *et al.* 2004; Telford *et al.* 2008). The most prominent phylogenetic questions of the Ecdysozoa regard the two major subdivisions (in terms of morphology at least) between its eight phyla; the Panarthropoda (Tardigrada, Onychophora and Arthropoda; Nielsen 2001) and the Cycloneuralia (Priapulida, Kinorhyncha, Loricifera, Nematomorpha and Nematoda; sensu Ahlrichs, 1995). Morphological support in favour of a monophyletic Panarthropoda (Nielsen, 2001), which on the face of it seems rather uncontentious, is furthermore supported by a number of molecular analyses (Zrzavý *et al.* 1998; Mallatt and Giribet, 2006; Dunn *et al.* 2008; Rota-Stabelli *et al.* 2010; Rota-Stabelli *et al.* 2011). Despite the support mentioned previously, a vast majority of molecular phylogenetic analyses support a closer relationship between tardigrades and cycloneuralian ecdysozoans (Philippe *et al.* 2005b; Roeding *et al.* 2005; Lartillot and Philippe, 2008; Sørensen *et al.* 2008; Hejnol *et al.* 2009; Roeding *et al.* 2009; Pick *et al.* 2010; Meusemann *et al.* 2010; Andrew, 2011).

These alternative hypotheses of tardigrade relationships have important consequences for our understanding of morphological evolution within Ecdysozoa. For example, if tardigrades are cycloneuralians, then the telescopic mouth cone and plated pharynx shared by tardigrades and cycloneuralians should be considered cycloneuralian apomorphies, whereas the important characteristics of segmentation and the possession of paired limbs must be homoplastic—they either evolved convergently in arthropods and tardigrades or were lost in nematodes (Edgecombe, 2010). Obviously, the opposite would be true if the tardigrades are panarthropods. Thus, accurately placing the tardigrades with respect to nematodes and arthropods is central to solving the interrelationships among the ecdysozoans and clarifying homologies within this group.

Although the rapidly growing influx of molecular data has dramatically altered our understanding of the animal tree of life, no dataset is homoplasy-free. Phylogenies derived from large, genomic- scale datasets of expressed sequence tags (ESTs) from many proteins minimize stochastic errors, yet they can exacerbate systematic errors (Jeffroy *et al.* 2006). This is because systematic errors, unlike stochastic ones, are positively misleading; the error increases with an increase in the amount of data in the analysis (Jeffroy *et al.* 2006). Although genomic-scale datasets are important for resolving difficult phylogenetic problems, suboptimal approaches to tree reconstruction, such as those using poorly fitting substitution models, can generate phylogenetic artifacts when applied to such datasets. Tools have been developed to ameliorate these problems, including comparing trees derived using differently fitting models (Rota-Stabelli *et al.* 2010; Rota-Stabelli *et al.* 2011; Philippe *et al.* 2011b), site-stripping (e.g. "slow-fast analysis": Brinkmann and Philippe, 1999; see section 2.3.3 of Chapter 2), signal dissection (Sperling *et al.* 2009a), and targeted taxon pruning (Holton and Pisani, 2010; Philippe *et al.* 2011a; Zwickl and Hillis, 2002). These tools mentioned above have been utilized in this Chapter in order to address the problem of phylogenetic affinity of Tardigrada within Ecdysozoa.

3.1.2 A closer look at the ecdysozoan phyla

Here I will discuss some of the morphological features of the major Ecdysozoan phyla and some of the evolutionary implications regarding their interrelationships. There are eight phyla that make up the Ecdysozoa (Arthropoda, Onychophora, Tardigrada, Priapulida, Nematoda, Nematomorpha, Kinorhyncha and Loricifera). The most easily recognised phylum within the Ecdysozoa must certainly be the arthropods (see Figure 3.1a); this group can be subdivided up into four main extant subphyla (crustaceans, insects, myriapods and chelicerates). The arthropods are characterised by a number of synapomorphies, most notably a hard external segmented exoskeleton (with differing degrees of tagmosis) with paired jointed appendages (Nielsen, 2001). This distinctive arthropod body plan can be broadened out more generally to include two additional ecdysozoan phyla, Tardigrada and Onychophora; which together have been grouped traditionally in a clade known as Panarthropoda (Nielsen, 2001). Morphological support in favour of Panarthropoda is conspicuous, characterized by a number of shared morphological features such as, a cuticle composed of α-chitin, paired segmentally repeated ventrolateral limbs with claws, paired leg nerves, lack of primary larvae, locomotory cilia and protonephridia. The grouping of these phyla into Panarthropoda has been further upheld by both embryological (Gabriel and Goldstein, 2007) and developmental evidence (Zantke *et al.* 2008).

Panarthropoda has received much support from fossil data obtained from a rich Cambrian fossil record; with Cambrian 'lobopod' type fossils displaying some variations of the arthropod body theme. Particular fossils have been crucial in allying Panarthropods into a monophyletic assemblage, for example the fossil taxon *Aysheaia,* was once thought to be an early annelid, but a more recent interpretation places it close to Onychophorans. Cambrian lobopods from the lower to middle Cambrian such as *Aysheaia*, and others like *Hallucigenia* and *Kerygmachela*; probably represent diverse stem groups (extinct lineages) from which the living panarthropod phyla originated and diversified (Nielsen, 2001; Budd, 2001).

The phylum Onychophora (see Figure 3.1b) more commonly known as velvet worms, comprises around 200 species all of which are terrestrial. The characteristic body plan is a worm-like cylindrical body, one pair of long anterior antennae, oral papillae and a number of segments with a pair of unjointed trunk legs ('lobopods') terminated with

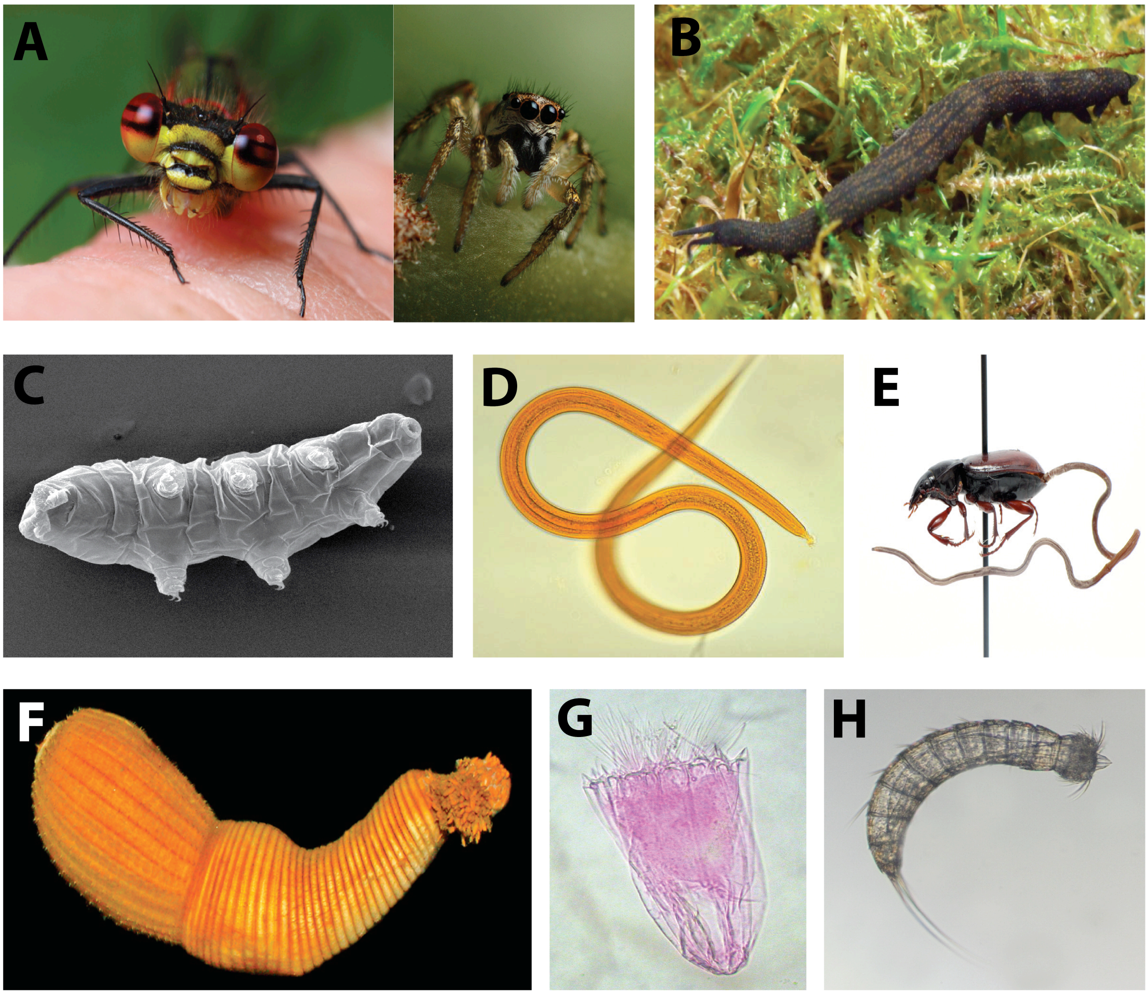


Figure 3.1: The eight phyla that comprise Ecdysozoa.

(A) Arthropoda; left side- Insect (damselfly); Right side- Arachnid (Jumping spider) – Displaying some well developed eyes and articulated legs. Images courtesy of Derek Cluskey (http://www.flickr.com/photos/degserman200/). (B) Onychophora (velvet worms) – Single pair of antennae, long worm like body with multiple pairs of lobopod legs. (C) Tardigrada (water bears), highlights the mixture of arthropod like and worm like features such as walking appendages and a terminal mouth cone. (D) Nematoda (roundworm) – displaying a transparent collagenous cuticle and terminal mouth. Image courtesy of the Tree of Life Web project (public domain). (E) Nematomorpha (Gordian worm) – Parasitic lifestyle, emerging from an arthropod host. Image courtesy of Crystal Ernest (www.crystalernst.wordpress.com). (F) Priapulida (Penis worm) – displaying an introvert with spines (scalids). Image courtesy of Herrmann, M. (2004). Macrozoobenthos communities of Svalbard. World Wide Web electronic publication. (http://www.macrozoobenthos.de). (G) Loricifera (*Spinoloricus* sp). Image courtesy of Cristina Gambi, Polytechnic University of Marche, Italy. (H) Kinorhyncha (Mud dragon) – Displaying trunk segments with locomotory spines. Specimen from kelp holdfast, Dale Fort, Wales. Collected and photographed by Ross Piper (http://scrubmuncher.wordpress.com) and identified by Martin V. Sørensen.

sclerotized terminal claws (Nielsen, 2001; Edgecombe, 2009) that gives Onychophora its name (literally translating to mean "claw-bearer"). Velvet worms can be found in warm temperate regions (e.g. Australia) but are predominantly located in tropical regions (e.g. South America) and usually inhabit environments with high humidity and dark shaded cover. They survive by predating on smaller animals like insects; catching them using a sticky slime produced in their oral papillae (modified glands). Many lower and middle Cambrian lobopods have been discovered which are believed to be early marine stem groups related to terrestrial velvet worms; like *Hallucigenia,* and the aforementioned *Aysheaia*. Evidence suggests that crown group velvet worms must have emerged terrestrially; as specialisations such as a tracheal system not being able to close fully makes them prone to desiccation in arid conditions. In addition, onychophoran nephridia are also of a structure found in many terrestrial groups (Nielsen, 2001).

The phylum Tardigrada comprises minute (~500 μm to ~1000 μm) metazoans commonly known as "water bears". Tardigrada (see Figure 3.1c) literally translating to "slow walker" a name linked to their reminiscent bear like gait, is now described to include more than one thousand species (Zhang, 2011) found ubiquitously in nature occurring in both marine and terrestrial habitats. Tardigrades are most famous for their incredible resilience to extremes of temperature (known to survive in between 150C to -272.8C or absolute zero) and radiation (surviving up to 570,000 rads – contrast to humans where 500 is a lethal dose) by way of entering into a metabolic stasis period known as cryptobiosis.

Tardigrade morphology is characteristic of a typical panarthropod Bauplan, as tardigrades share synapomorphies with onychophorans and arthropods such as paired ventrolateral legs, and an external cuticle made of α-chitin. Indeed their morphology also hints at a possible non-arthropod nature, with features more reminiscent of that of a worm-like cycloneuralian ground plan (terminal mouth, protrusible mouth cone, triradiate pharynx, and a circumesophageal brain) (Schmidt-Rhaesa, 1998; Zantke *et al.* 2008; Edgecombe, 2010). This mixture of both arthropod and cycloneuralian like morphology seen in Tardigrada hints at two possible evolutionary scenarios; either the arthropod like characters were lost in cycloneuralians or the cycloneuralian like characters were lost in the arthropods (assuming the cycloneuralian characters seen in tardigrades are homologous to that of cycloneuralians).

The remaining ecdysozoan phyla (nematodes, nematomorphs, priapulids, kinorhynchs and loriciferans) make up the lesser-known group Cycloneuralia (Ahlrichs, 1995) or Introverta (Nielsen, 2001) characterized by a "worm-like" body plan. The name Cycloneuralia derives from the collar-shaped, cicum-oral brain present in all cycloneuralian phyla; further to this is the presence of an eversible anterior end or introvert seen in most taxa (deriving the alternate name Introverta), and the shared absence of a true coelom and walking appendages (Nielsen, 2001). Many of the relationships between the separate phyla remain to be resolved (Telford *et al.* 2008) this is most likely due to the difficulty working with some phyla which are extremely small and/or hard to collect in the field. However some relationships seem to be reasonably credible. The first relationship, that has received much support is the relationship between Nematoda and Nematomorpha (see Figure 3.1d and e) in a sister group relationship known as Nematoida (Schmidt-Rhaesa, 1998). Morphologically these two groups share features like a collagenous cuticle, the reduced circular muscles in the body wall, and aflagellate sperm (Nielsen, 2001) supported further from analyses of molecular data sets including rRNA (Peterson and Eernisse, 2001; Mallatt *et al.* 2004; Mallatt and Giribet, 2006) and phylogenomics (Dunn *et al.* 2008). Another clade within Cycloneuralia known has Scalidophora (Ahlrichs, 1995) has received support from a combination of morphological and molecular analyses. Here, Scalidophora unites the three phyla Priapulida, Kinorhyncha and Loricifera (see Figure 3.1f, g and h) together on the basis of a shared possession of an introvert with scalids (spines) and the presence of two rings of retractor muscles on the introvert (Heiner and Kristensen, 2005; Telford *et al.* 2008). Importantly, molecular analyses including all three scalidophoran phyla are few and so a lack of consensus remains on their exact interrelationships.

Despite the sparse number of molecular analyses including all relevant phyla, 18S and 28S rRNA analysis and a phylogenomic analysis has placed Priapulida + Kinorhyncha in a sister group together (Garey *et al.* 2001; Mallatt and Giribet, 2006; Dunn *et al.* 2008). An expansion of the data presented in the Dunn *et al.* analysis recovered an alternative phylogeny instead grouping Kinorhyncha + Nematomorph (Hejnol *et al.* 2009). Conversely molecular phylogenetic analyses including data for Loricifera (smallest metazoan phylum known to science) have recovered a sister group relationship between Loricifera + Nematomorpha (Sørensen *et al.* 2008). Lastly, one of the other major questions regarding ecdysozoan evolution is the question of cycloneuralian monophyly versus paraphyly. Although Cycloneuralia is supported morphologically (Ahlrichs, 1995; Nielsen, 2001), some analyses have refuted this group (Zrzavý *et al.* 1998; Peterson and Eernisse 2001; Mallatt *et al.* 2004). The biological implications regarding its mono- or paraphyletic status are particularly important to the understanding of arthropod evolution and the reconstruction of the ancestral ecdysozoan ground plan. For example, monophyly of Cycloneuralia implies that characteristic features like segmentation and coeloms seen in morphologically complex protostome phyla like annelids and arthropods evolved convergently, or conversely parallel losses occurred in cycloneuralians.

In this Chapter I will present phylogenetic analyses of ESTs to investigate the relationships of Tardigrada within the Ecdysozoa, and furthermore on the interrelationships of the other Ecdysozoan phyla. The results of the analyses presented in this Chapter address the question; do the alternative hypotheses for the position of Tardigrada within Ecdysozoa (arthropod vs. nematode affinity) obtained by previous phylogenomic analyses, represent tree-reconstruction artifacts? This work has been completed under a collaborative effort, and published in the peer-reviewed journal Proceedings of the National Academy of Sciences (Campbell *et al.* 2011).

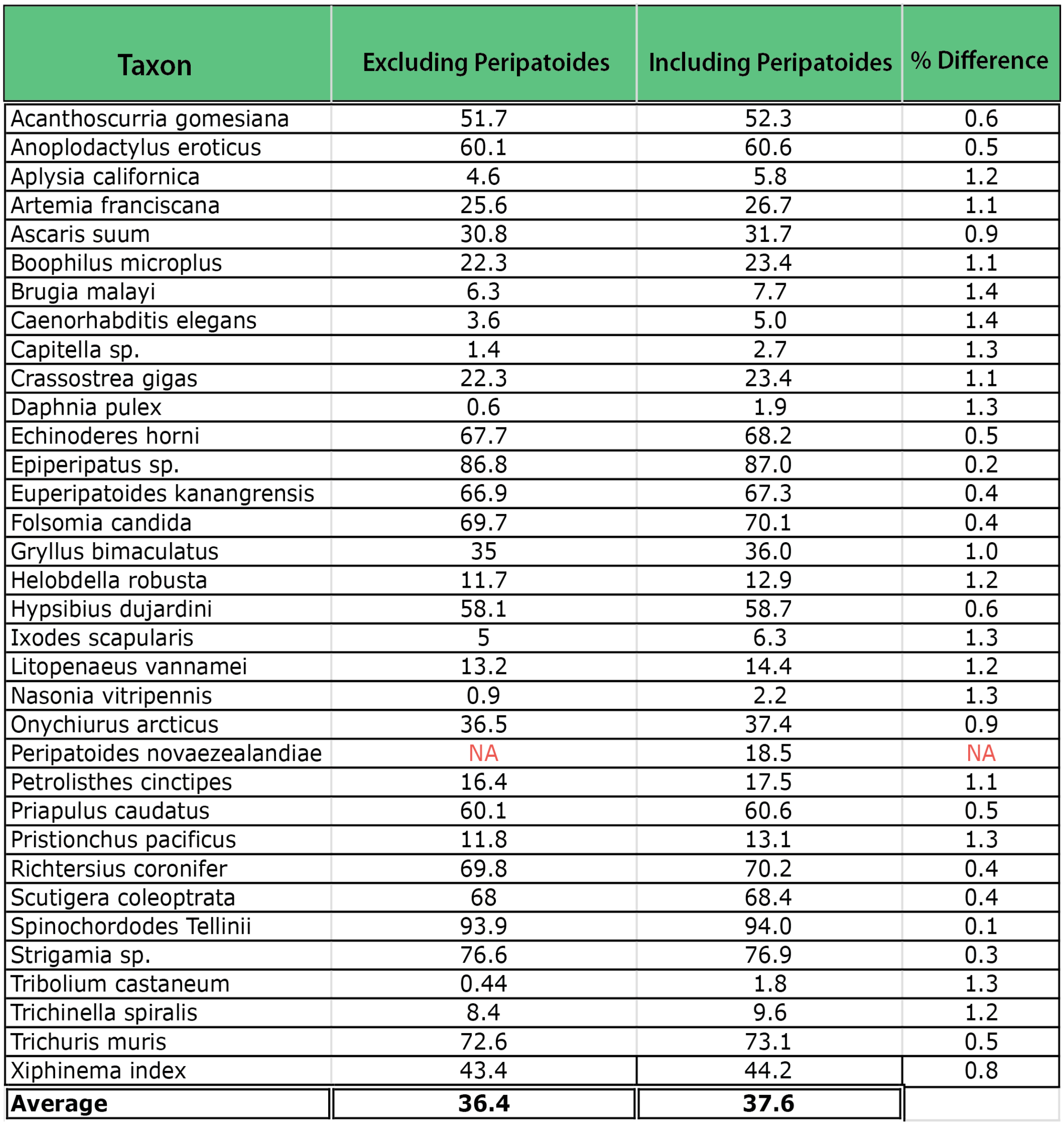
3.2 Materials and Methods

3.2.1 EST Data set assembly

For the analyses presented in this Chapter, I assembled a phylogenomic data set of 255 genes spanning 49,023 amino acid positions, for 33 ecdysozoan species by merging genes from two previously published EST data sets (Dunn *et al.* 2008; Rota-Stabelli *et al.* 2011). EST data set assembly was performed using a BLAST -based strategy, which was used to identify and eliminate redundant genes (i.e. genes present in both data sets). Single genes from (Dunn *et al.* 2008) were identified for the species

Daphnia pulex (*D. pulex* had the highest gene coverage in the Dunn data set: 99.5% total coverage). These single genes were then blasted against a local database made up of single genes (also from *D. pulex* which had a coverage of 99.4%; taken from (Rota-Stabelli *et al.* 2011). Overall, 13 orthologs from Dunn *et al.* were identified that did not have any hits in the alignment of Rota-Stabelli *et al.* Identified genes from Dunn *et al.* alignment were added to the initial 242 gene alignment of Rota-Stabelli *et al.* to generate the alignment used herein. The combined data set generated had an average of 36.4% missing data; see Table 3.1 for a list of species and their associated alignment coverage. A key difference between the alignment of Rota-Stabelli *et al.* 2011 and that used in the present study is that Rota-Stabelli *et al.* did not include any nematomorph species. With reference to (Dunn *et al.* 2008), our dataset includes 12 new taxa, including an onychophoran (*Epiperipatus sp.*) and several nematodes, including the relatively slowly evolving *Trichuris muris*. With reference to Rota-Stabelli *et al.* (2011) our dataset includes an extra onychophoran (*Epiperipatus sp.*) an additional relatively slowly evolving nematode (*T. muris*) and most importantly the nematomorph *Spinochordodes tellinii*. There is ample evidence that the Nematomorpha constitute the sister group of Nematoda within Nematoida (Schmidt-Rhaesa, 1996) and might be closely related to the Tardigrada (assuming that the latter are relatives of the Nematoda). For this study, including at the least a representative of the Nematomorpha is key, as the Nematomorpha might be useful to break the long branch leading to the Nematoda and thus help reduce LBA artifacts that could affect the position of the Tardigrada.

To include a Nematomorpha I downloaded all available 2,208 trace files from the NCBI trace archives (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi) for the nematomorph *S. tellinii*, blasted each of the genes against the genes in the alignment of (Rota-Stabelli *et al.* 2011) and identified available *Spinochordodes* orthologs of these genes. Only 37 significant blast hits were identified. The corresponding trace files were assembled into 30 contigs using Sequencher (GeneCodes) and then added to the alignment. For the 13 genes in our alignment obtained from Dunn *et al.* (2008) we did not need to identify *Spinochordodes tellinii* orthologs because the Nematomorpha were represented in the dataset of (Dunn *et al.* 2008). Unfortunately, little data on the



Taxon	Excluding Peripatoides	Including Peripatoides	% Difference
Acanthoscurria gomesiana	51.7	52.3	0.6
Anoplodactylus eroticus	60.1	60.6	0.5
Aplysia californica	4.6	5.8	1.2
Artemia franciscana	25.6	26.7	1.1
Ascaris suum	30.8	31.7	0.9
Boophilus microplus	22.3	23.4	1.1
Brugia malayi	6.3	7.7	1.4
Caenorhabditis elegans	3.6	5.0	1.4
Capitella sp.	1.4	2.7	1.3
Crassostrea gigas	22.3	23.4	1.1
Daphnia pulex	0.6	1.9	1.3
Echinoderes horni	67.7	68.2	0.5
Epiperipatus sp.	86.8	87.0	0.2
Euperipatoides kanangrensis	66.9	67.3	0.4
Folsomia candida	69.7	70.1	0.4
Gryllus bimaculatus	35	36.0	1.0
Helobdella robusta	11.7	12.9	1.2
Hypsibius dujardini	58.1	58.7	0.6
Ixodes scapularis	5	6.3	1.3
Litopenaeus vannamei	13.2	14.4	1.2
Nasonia vitripennis	0.9	2.2	1.3
Onychiurus arcticus	36.5	37.4	0.9
Peripatoides novaezealandiae	NA	18.5	NA
Petrolisthes cinctipes	16.4	17.5	1.1
Priapulus caudatus	60.1	60.6	0.5
Pristionchus pacificus	11.8	13.1	1.3
Richtersius coronifer	69.8	70.2	0.4
Scutigera coleoptrata	68	68.4	0.4
Spinochordodes Tellinii	93.9	94.0	0.1
Strigamia sp.	76.6	76.9	0.3
Tribolium castaneum	0.44	1.8	1.3
Trichinella spiralis	8.4	9.6	1.2
Trichuris muris	72.6	73.1	0.5
Xiphinema index	43.4	44.2	0.8
Average	**36.4**	**37.6**	

Table 3.1: Percent of missing data for EST datasets. Values shown for two EST datasets; excluding the onychophoran *Peripatoides novaezealandiae,* and including *Peripatoides novaezealandiae.*

Nematomorpha are available in public data repositories, and whereas the average amount of missing data in our dataset is ~36%, the amount of missing information for *Spinochordodes* is much higher (~94%; see Table 3.1). When *Spinochordodes* is not considered, the average amount of missing data in our dataset is ~34%. In the analyses presented in this Chapter (absent from original publication Campbell *et al.* 2011) a third Onychophoran (*Peripatoides novaezealandiae*) was also added to the dataset (see section 3.2.5 for details).

3.2.2 Phylogenetic analysis

All phylogenetic analyses were conducted under a Bayesian framework using PhyloBayes 3.2e (Lartillot *et al.* 2009). I first compared the fit of alternative models of evolution to our EST dataset. I then used Bayesian cross-validation (Stone, 1974), as described in the PhyloBayes manual (Lartillot *et al.* 2009) to rank the fit of alternative substitution models to the data. The models compared were WAG+Γ, GTR+Γ, CAT+Γ, and CAT-GTR+Γ.

Phylogenetic analyses of the EST dataset were performed under each model, and results were compared to evaluate whether different phylogenies were obtained when different-fitting models were used. For every PhyloBayes analysis two independent runs were executed. Convergence was tested using ‘bpcomp’ in the PhyloBayes package. Analyses were considered to have converged sufficiently when the maximum difference across bipartitions was <0.2 (see the PhyloBayes manual); thereby allowing both independent chains to sample from tree space with a similar probability distribution. For each analysis, the burn-in period was estimated independently, and trees sampled before convergence were not considered when summarizing the results of the two runs.

3.2.3 Site stripping and Signal Dissection analyses

These analyses used the slow-fast method (Brinkmann and Philippe, 1999) to estimate the rate of substitution of the sites in our alignment. First, the parsimony score of each site in our alignment was calculated for each of four groups with constrained monophyly (Pancrustacea, Chelicerata, Nematoda, and Lophotrochozoa). The rate of each site in our alignment was then estimated as the sum of its parsimony scores across all considered monophyletic groups. All parsimony analyses were performed using PAUP4b10 (Swofford, 2002). Sites in our alignment were then ranked according to their substitution rates and partitioned into classes. Alignments were generated, according to the distribution of site rates, by systematically removing (i) approximately the fastest 10% of the sites, that is, all characters with a slow-fast–estimated rate of six or more steps (total number of remaining sites, 45,292); (ii) the fastest ~20% of the sites, that is, all characters with a slow-fast estimated rate of five or more steps (total number of remaining sites, 43,316); and (iii) the fastest ~30% of the sites, that is, all characters with a slow-fast–estimated rate of three or more steps (total number of remaining sites, 37,150). However, the number of substitutions in the sites that remained after exclusion of the first 10% of characters at just five or fewer steps is already low. This implies that the proportion of fast evolving sites in our alignment is quite small. Accordingly, we did not create datasets excluding more than 30% of the fastest sites.

We also performed a signal-dissection analysis (Rota-Stabelli *et al.* 2011; Sperling *et* *al.* 2009a) to compare the signal in the slow- and fast-evolving sites. Accordingly, two datasets were generated, containing approximately 10% (3,731 sites) and 30% (11,873 sites) of the fastest sites in our alignment. The five aligned datasets that resulted, namely the three sets composed of slow-evolving sites (approximately the slowest 70%, 80%, and 90%) and the two sets of fast-evolving sites (approximately the fastest 10% and 30%), were analyzed independently using PhyloBayes 3.2e to construct trees under the best-fitting model.

3.2.4 Taxon pruning analyses

It is well known that the number and nature of the taxa used can affect phylogenetic inference, and in particular can exacerbate or reduce LBA (Aguinaldo *et al.* 1997; Philippe *et al.* 2005a; Holton and Pisani *et al.* 2010). Thus I carried out three taxon pruning experiments to evaluate the robustness of the EST results. Data sets were generated that excluded (i) the tardigrade *Richtersius coronifer* and the onychophoran *Epiperipatus sp*., which resulted in uninterrupted branches for the tardigrades and the onychophorans; (ii) the nematomorph *Spinochordodes tellinii* and the tardigrade *R. coronifer*, which resulted in uninterrupted branches leading to the nematodes and the tardigrades; and (iii) the onychophoran *Epiperipatus sp*., the tardigrade *R. coronifer*, and the nematomorph *S. tellinii*, which resulted in uninterrupted branches leading to the onychophorans, tardigrades, and nematodes. In these experiments, the retained tardigrade was always *Hypsibius dujardini* because of its greater gene coverage. All datasets were analyzed under the best fitting model.

3.2.5 Increasing taxon sampling with *Peripatoides novaezealandiae*

In a further set of analyses an additional panarthropod species, the onychophoran *P. novaezealandiae* was added to the data set, bringing the total number of onychophorans represented in the alignment to three. Increased taxon sampling has been noted to improve phylogenetic resolution (Hendy and Penny, 1989) with the potential to break long branches thus reducing LBA (Swofford *et al.* 1996; Pollock *et al.* 2002; Rota-Stabelli and Telford, 2008). The goal here was to compare the results of an increased taxon data set to that of the 33 taxon data set (Campbell *et al.* 2011; see methods section 3.2.1), in order to see what effects (if any) this additional onychophoran would make on the recovered topology. Amino acid sequences were obtained for the species *Peripatoides novaezealandiae* from a next generation sequence assembly. 658,698 contigs were assembled using the software Abyss 1.2.5 (Birol *et al.* 2009) from an initial set of 34,928,782 paired end reads sequenced in P. *novaezealandiae*. Translation of the contigs was performed using Prot4EST (Wasmuth and Blaxter, 2004) that allows translation using a reference set of ESTs. A reference set of 12,380 EST's was obtained from NCBI for the onychophoran species *Peripatopsis sedgwicki*. Upon completion, Prot4EST generated a set of 281,771 translated protein sequences for *P. novaezealandiae*. A blast based strategy was then used; translated amino acid sequences for *P. novaezealandiae* were blasted against a database of the set of non-redundant genes combined in section 3.2.1 from (Dunn *et al.* 2008; Rota-Stabelli *et al.* 2011: Hereto referred to as Campbell *et al.* 2011). In total, 210 genes were found to have a hit with orthologs from Campbell *et al.* 2011. Individual orthologs identified within Campbell *et al.* were added to the original alignment. Inclusion of *P. novaezealandiae* brought the taxon sampling for the increased taxon data set to 34, with the new alignment spanning 49,655 positions. The new alignment generated had an average of 37.69% missing data (see Table 3.1) increasing the total amount of missing data from the original alignment by 1.5%. Despite this, alignment coverage for Onychophora increased dramatically, with *P. novaezealandiae* having 18.52% missing data, compared to the species *Euperipatoides kanangrensis* and *Epiperipatus sp.,* which had a total of 66.9% and 86.8% missing data respectively.

Site stripping and signal dissection analyses were performed on this data set again, as described in section 3.2.3. Alignments were generated according to the distribution of site rates, by systematically removing (i) approximately the fastest 10% of the sites (total number of remaining sites, 45,938); (ii) the fastest ~20% of the sites (total number of remaining sites, 41,312); and (iii) the fastest ~30% of the sites, (total number of remaining sites, 37,828). Finally signal dissection analyses to compare slow and fast evolving sites were also performed. Accordingly, two further data sets were generated, containing approximately 10% (3,713) and ~30% (11,827) of the fastest sites in the alignment. These data sets were again analysed using PhyloBayes3.2e under the best fitting model.

3.2.6 Assessing support via bootstrap analysis

In all our analyses support was assessed using Posterior probabilities. However, in addition to the five analysis performed on this EST data set, I also used bootstrapping (Efron, 1979) first implemented in phylogenetics by Felsenstein (1985). Bootstrapping allows you to estimate the statistical error in situations where the underlying sampling distribution is unknown or difficult to derive. In bootstrapping the original alignment is used to generate multiple replicate data sets of the same size (pseudoreplicates) by randomly sampling alignment columns with replacement from the original alignment and then reconstructing phylogenetic trees for each. Each resulting tree is then pooled to generate a single tree by way of a majority rule consensus method. As each pseudoreplicate is analyzed independently, the proportion of times a clade is found among all bootstrap replicates is considered as the measure of robustness for the monophyly of that particular taxon subset.

It has been said that Posterior probabilities might be too permissive (Douady *et al.* 2003; Erixon *et al.* 2003) whilst it is well known that the bootstrap has the opposite problem: being too conservative (Hillis and Bull, 1993). Using both bootstrap and posterior probabilities, and their comparison I thus able to get a better feel for the real support of the relationships in the recovered trees.

To generate the bootstrap data sets I used the program SEQBOOT which is part of the software package Phylip 3.0 (Felsenstein, 2004). 100 pseudoreplicate data sets were generated from the original alignment (see section 3.2.1). Bayesian analysis was performed on each resulting pseudoreplicate data set under the best-fitting model using the software Phylobayes 3.2 (Lartillot *et al.* 2009). Bayesian analyses were performed as in section (3.2.2). Individual consensus trees were pooled, and a final bootstrap consensus tree was generated using CONSENSE (see Phylip 3.0 manual).

3.3 Results and discussion

3.3.1 Identifying the best fitting evolutionary model

In an effort to obtain a reliable phylogeny from our assembled data set, we wanted to ensure the best fitting evolutionary model was utilized to describe the data. Since the change from employing small numbers of genes (e.g. LSU and SSU rRNA, EF-1a, RNA polymerase) to large scale multi-gene, EST and phylogenomic data is now commonplace in phylogenetics, it could be fair to immediately assume an increase in phylogenetic accuracy would follow. However, a distinction must be made here; it is true that analyses of large concatenated data sets typically reduces the problem of stochastic error (Lartillot and Philippe, 2008; Dunn *et al.* 2008; Hejnol *et al.* 2009; Regier *et al.* 2010; Rota-Stabelli *et al.* 2011) this does not however translate to an improved phylogenetic accuracy. As the "true tree" is unknowable, increasing the amount of data can only increase overall phylogenetic precision. Accordingly, evolutionary models still do not encompass the entirety of the evolutionary process and so systematic error instead of being reduced in large-scale EST and phylogenomic analyses actually becomes reinforced due to the property of statistical inconsistency (obtaining the wrong result as more and more data are added) (Philippe and Delsuc, 2005). Despite current evolutionary models taking into account, for example, compositional heterogeneity (Foster, 2004) and among site rate heterogeneity (Kolaczkowski and Thornton, 2004; Tuffley and Steel, 1998) all models of evolution still make assumptions about the data. When those assumptions fail to describe the data accurately phylogenetic biases such as LBA can occur. These problems can be diminished by a number of methods, such as improvement of taxon sampling and selection of genes or sites that evolve slowly; however effective prevention of biases in the first place necessitates the use of efficient tree reconstruction methods and models of evolution that describe the data accurately.

Because the use of poorly fitting models can generate phylogenetic artifacts, I first used Bayesian cross-validation (Stone, 1974) to rank substitution models according to their fit to the alignment. The substitution models tested in the analysis were a combination of homogeneous empirical WAG+Γ and mechanistic GTR+Γ models, and the more sophisticated heterogeneous mixture models such as CAT+Γ and CAT-GTR+Γ. The results of the model selection analysis are presented in Figure 3.2, and show a regular increase in the fit of the model to the data when moving from simple

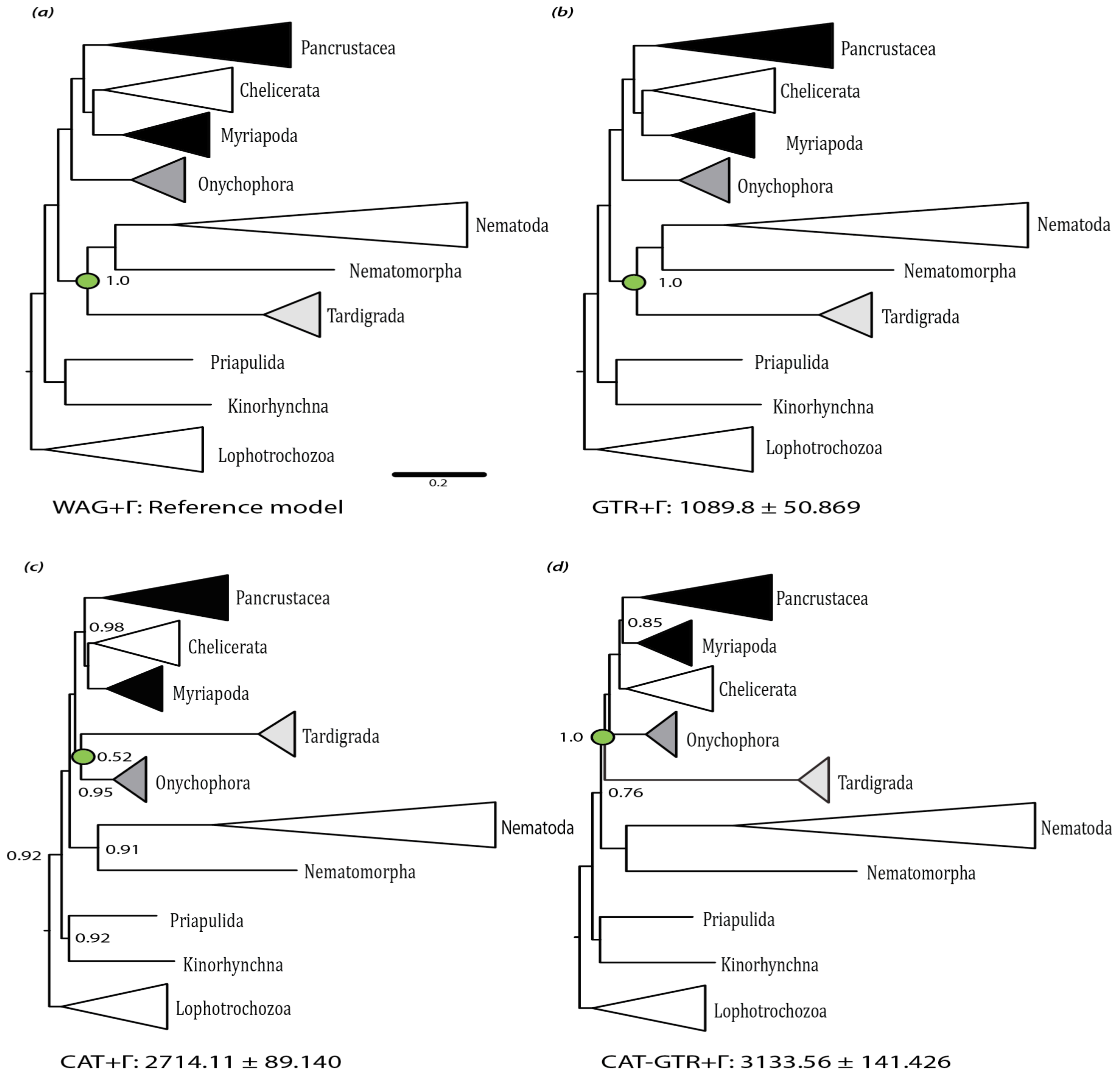


Figure 3.2: Bayesian cross-validation. Model selection test for the EST dataset, comparing models (a) WAG+Γ, (b) GTR+Γ, (c) CAT+Γ, and (d) CAT-GTR+Γ. Δ log-likelihoods and Standard deviations (SD) are shown. Positive values identify models that fit the data better than the reference model (WAG+Γ). Values at nodes are posterior probabilities (PP = 1 not shown except when they referred to the Tardigrada). Clades have been collapsed for clarity. The position of Tardigrada is indicated in green.

to more complex models, with the site-heterogeneous mixture model CAT-GTR+Γ having the best fit to our dataset. All models tested used a gamma distribution of rate variation across sites. According to the results of our model fit analysis we found it clear that the best fitting model to our data is CAT-GTR+Γ.

3.3.2 EST based phylogenomic analysis support Panarthropoda and Lobopodia

The results of our Bayesian analysis performed using the best fitting CAT-GTR+Γ model are shown in Figure 3.3. The majority of internal nodes have a posterior probability (PP) support value of 1. Tardigrada is recovered within Panarthropoda, sister group to Onychophora + Arthropoda, together called the Lobopodia (Snodgrass, 1938) with a PP support of 1. Within the arthropods themselves, analyses recover the chelicerate affinity of the sea spiders and are consistent with the monophyly of Mandibulata (Myriapoda + Pancrustacea) as found in recent phylogenomic based analyses (Rota-Stabelli *et al.* 2011; Regier *et al.* 2010). Our results do not support the monophyly of the Cycloneuralia, given that Nematoida (Nematoda + Nematomorpha) is recovered as the sister group of Panarthropoda, albeit with a low posterior probability (PP = 0.76) whereas Scalidophora (Priapulida + Kinorhyncha) is recovered as the sister group of all other ecdysozoans. Nematoida was recovered with PP = 1. Because Nematomorpha has the greatest amount of missing data in our EST dataset (see Table 3.1) the strong support found for Nematoida (an otherwise well-accepted clade: Nielsen, 2001; Schmidt-Rhaesa, 1996, 1998) suggests that missing data for Nematomorpha does not have a drastically negative impact on the results obtained.

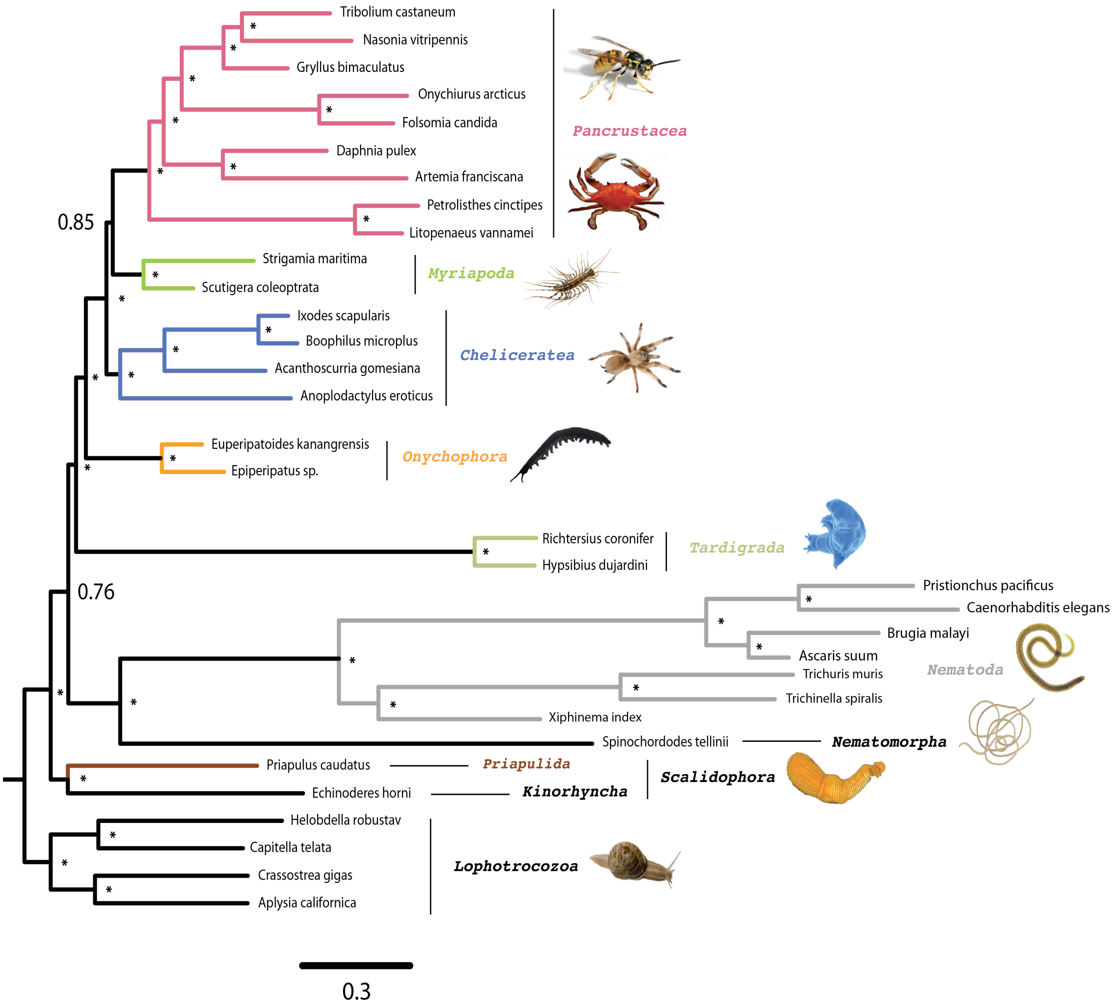


Figure 3.3: EST based phylogeny of Ecdysozoa: Supports a sister group relationship between arthropods and velvet worms inside a monophyletic Panarthropoda. Bayesian analysis of the EST data under the best-fitting CAT-GTR+Γ model supports tardigrades as the sister group of Lobopodia (Onychophora + Arthropoda) and the paraphyletic nature of Cycloneuralia, with Nematoida sister group to Panarthropoda. Support values represent posterior probabilities. Asterisks indicate a PP value of 1.0.

3.3.3 Model selection and Signal dissection reveal the artifactual nature of Tardigrada + Nematoda

To better understand the underlying phylogenetic signal present within our data set, I performed a number of analyses to test whether the grouping of Tardigrada + Nematoda obtained in previous molecular analyses (Philippe *et al.* 2005b; Roeding *et al.* 2005; Lartillot and Philippe, 2008; Sørensen *et al.* 2008; Hejnol *et al.* 2009; Roeding *et al.* 2009; Pick *et al.* 2010; Meusemann *et al.* 2010; Andrew, 2011) could

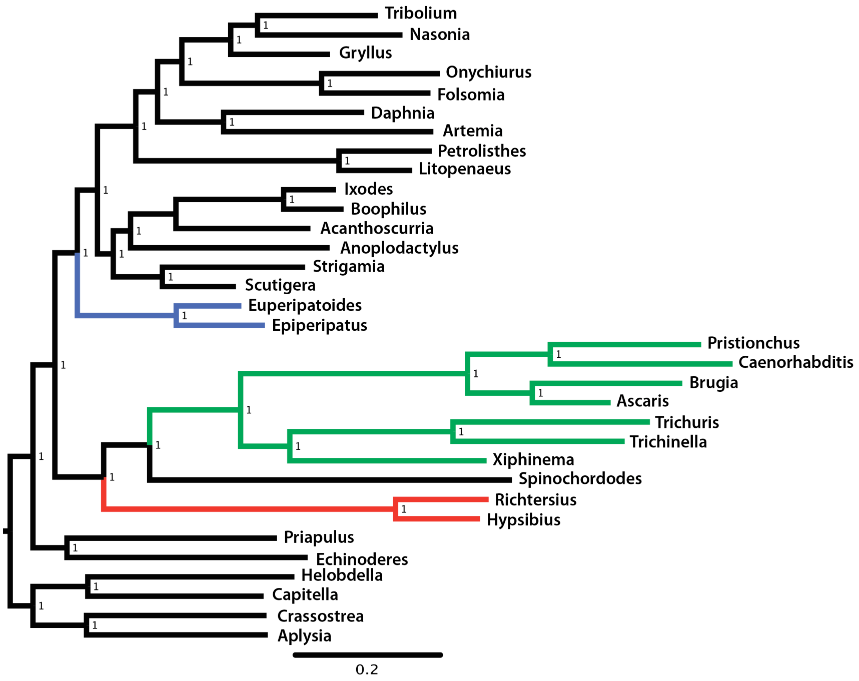


Fig 3.4.1: Bayesian analysis of our EST alignment under the poor-fitting WAG+Γ model. Topology shown is a majority rule consensus tree. Node values are posterior probabilities. Under the WAG+Γ model tardigrades cluster with the cycloneuralian group Nematoida. Clade colours: red, Tardigrada; blue, Onychophora; green, Nematoda.

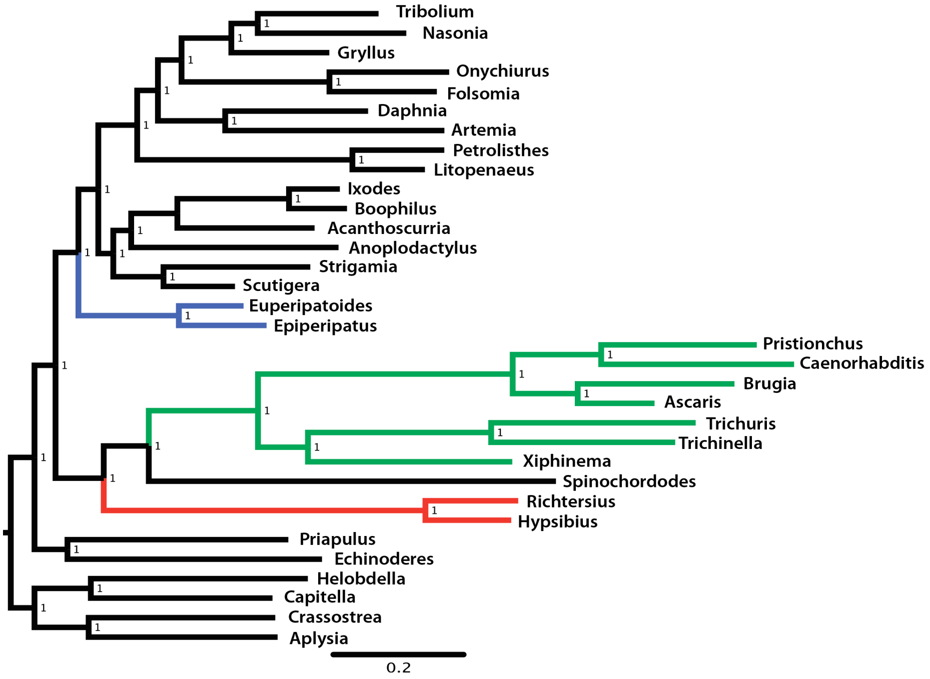


Fig 3.4.2: Bayesian analysis of our EST alignment under the poor-fitting GTR+Γ model. Topology shown is a majority rule consensus tree. Node values are posterior probabilities. Under the GTR+Γ model tardigrades cluster with the cycloneuralian group Nematoida. Clade colours: red, Tardigrada; blue, Onychophora; green, Nematoda.

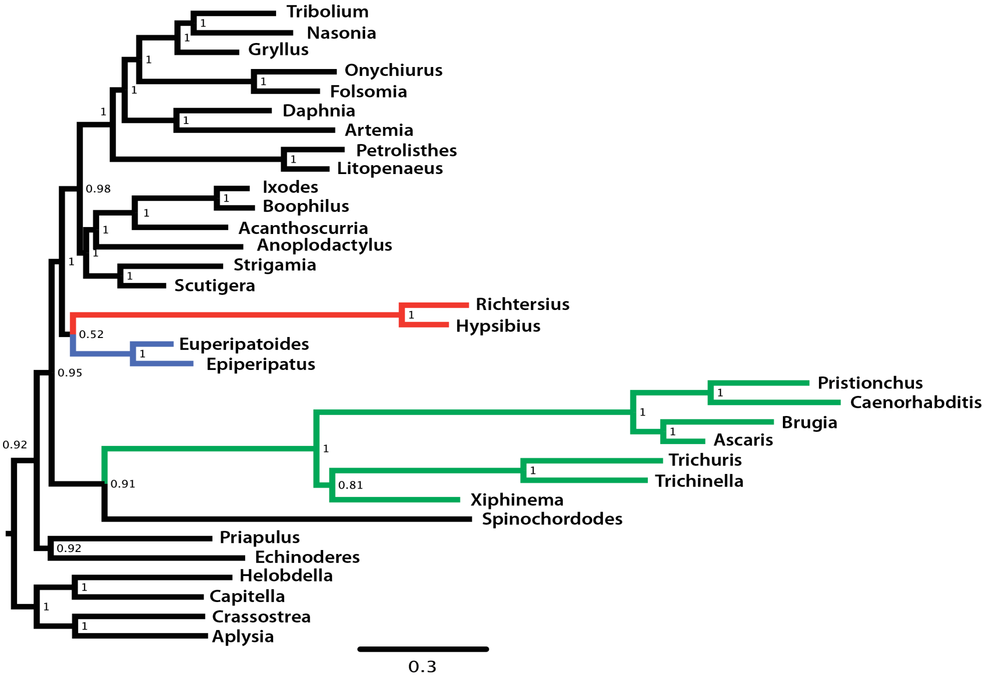


Fig 3.4.3: Bayesian analysis of our EST alignment under the better-fitting CAT+Γ model. Topology shown is a majority rule consensus tree. Node values are posterior probabilities. The CAT+Γ model supports a monophyletic Panarthropoda, with Tardigrada sister group to the Onychophora. Clade colours: red, Tardigrada; blue, Onychophora; green, Nematoda.

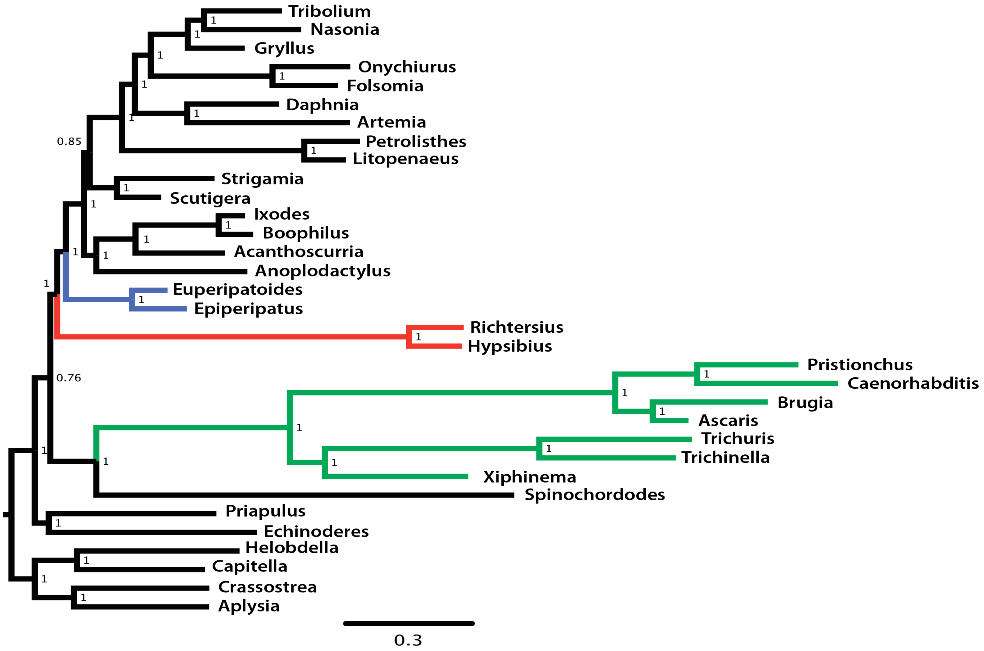
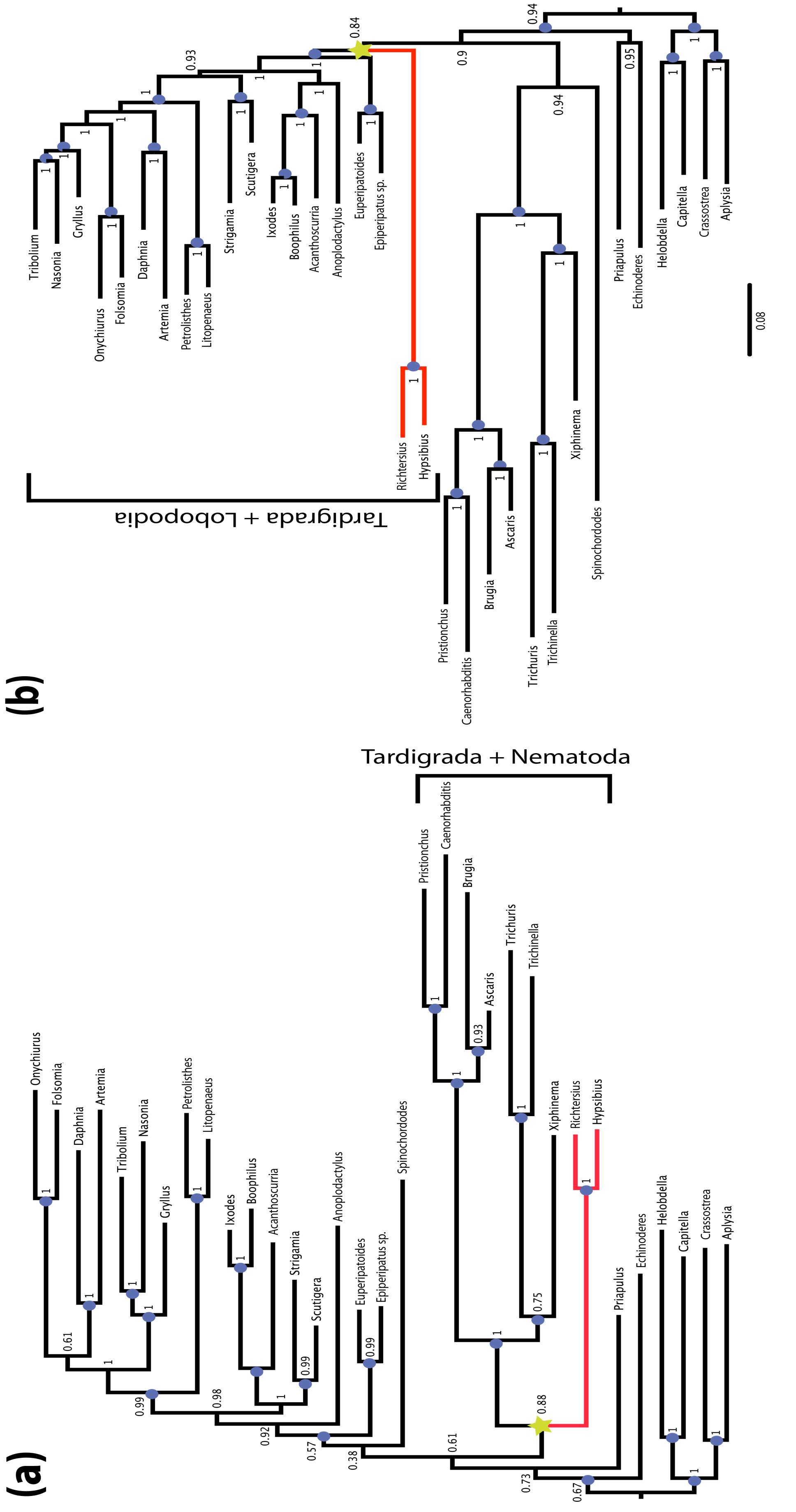


Fig 3.4.4: Bayesian analysis of our EST alignment under the best-fitting CAT-GTR+Γ model. Topology shown is a majority rule consensus tree. Node values are posterior probabilities. The CAT-GTR+Γ model supports a monophyletic Panarthropoda, with Tardigrada sister group to Onychophora + Arthropoda (Lobopodia). Clade colours: red, Tardigrada; blue, Onychophora; green, Nematoda.

be the result of systematic error. To do this, Bayesian analyses were performed on our data set under a series of four evolutionary models (WAG+Γ, GTR+Γ, CAT+Γ, CAT-GTR+Γ).

When the analyses were performed under poor-fitting models (i.e. WAG+Γ, and GTR+Γ; see Figures 3.4.1, 3.4.2 and Figure 3.2a,b) Panarthropoda was not recovered, and instead the Tardigrada were resolved as the sister group to the Nematoida (Nematoda + Nematomorpha) with full PP support of 1. In contrast, the better fitting site-heterogeneous models (CAT+Γ and CAT-GTR+Γ; see Figures 3.4.3, 3.4.4 and Figure 3.2c,d) invariably found Tardigrada as a member of monophyletic Panarthropoda.

I next performed a signal dissection analysis (Rota-Stabelli *et al.* 2010; Sperling *et al.* 2009a) based on the slow-fast technique (Brinkmann and Philippe, 1999). Sites in alignment were partitioned according to their rate of evolution, and then independently analysed (see methods section 3.2.3). We hypothesized that if the artifactual nature of Tardigrada to position sister to Nematoda was due to LBA, then the support for this grouping would be maximized in the fast-evolving sites, while conversely it would be minimised in the partitions that excluded those fast-evolving sites. Results of these the slow-fast analyses were consistent with our hypothesis, supporting Tardigrada + Nematoda in the two fast evolving partitions, whereas partitions of the slowest-evolving sites recovered monophyletic Panarthropoda. The partition containing only the fastest 10% of sites had a PP support for Tardigrada + Nematoda of 0.88 (see Figure 3.5a) while in the partition containing the fastest 30% of sites PP support for this clade decreased to 0.5 (Figure 3.6a). Results of signal dissection analyses are summarized in Table 3.2.


(a)
Tardigrada + Nematoda
Onychiurus
Folsomia
Daphnia
Artemia
Tribolium
Nasonia
Gryllus
Petrolisthes
Litopenaeus
Ixodes
Boophilus
Acanthoscurria
Strigamia
Scutigera
Anoplodactylus
Euperipatoides
Epiperipatus sp.
Spinochordodes
Pristionchus
Caenorhabditis
Brugia
Ascaris
Trichuris
Trichinella
Xiphinema
Richtersius
Hypsibius
Priapulus
Echinoderes
Helobdella
Capitella
Crassostrea
Aplysia
0.61
0.99
0.98
0.92
0.57
0.38
0.61
0.73
0.67
0.88
0.75
0.93
(b)
Tardigrada + Lobopodia
Tribolium
Nasonia
Gryllus
Onychiurus
Folsomia
Daphnia
Artemia
Petrolisthes
Litopenaeus
Strigamia
Scutigera
Ixodes
Boophilus
Acanthoscurria
Anoplodactylus
Euperipatoides
Epiperipatus sp.
Richtersius
Hypsibius
Pristionchus
Caenorhabditis
Brugia
Ascaris
Trichuris
Trichinella
Xiphinema
Spinochordodes
Priapulus
Echinoderes
Helobdella
Capitella
Crassostrea
Aplysia
0.93
0.84
0.9
0.94
0.94
0.95
0.08

←**Figure 3.5: Signal dissection of EST data. Analyses were performed under CAT-GTR+Γ.** Node values are posterior probabilities. The tardigrade branch is highlighted in red. A gold star indicates the node connecting tardigrades to the rest of the tree. (A) Analysis of the fastest 10% of the data recovers Tardigrada as the sister to Nematoda. (B) Analysis of the slowest 90% of sites recovers monophyletic Panarthropoda, with tardigrades as the sister of Lobopodia (Onychophora + Arthropoda). Monophyletic clades recovered by both the slow and fast positions are highlighted in blue.

In contrast, the analysis of the slowest evolving 90% of sites, recovered a PP support of 0.84 (Figure 3.5b) for Tardigrada + Lobopodia. An unexpected topology was recovered for the analysis of the slowest evolving 70% of sites; in this topology Tardigrada were supported with a PP support of 1.0 as the sister group to Arthropoda (Figure 3.6b) however, the Onychophora were found positioned inside the Arthropoda as the sister group to Myriapoda with a PP support of 1. To my knowledge, there are no molecular phylogenetic studies that recover the group of Onychophora sister to Myriapoda. This suggests to me that this position for Onychophora must be due to a lack of phylogenetic signal in this data set. I feel this to be likely as the remaining sites after removal of more than 30% of fastest sites in this data set have a substitution rate that is very low (2 or less substitutions across all taxa).

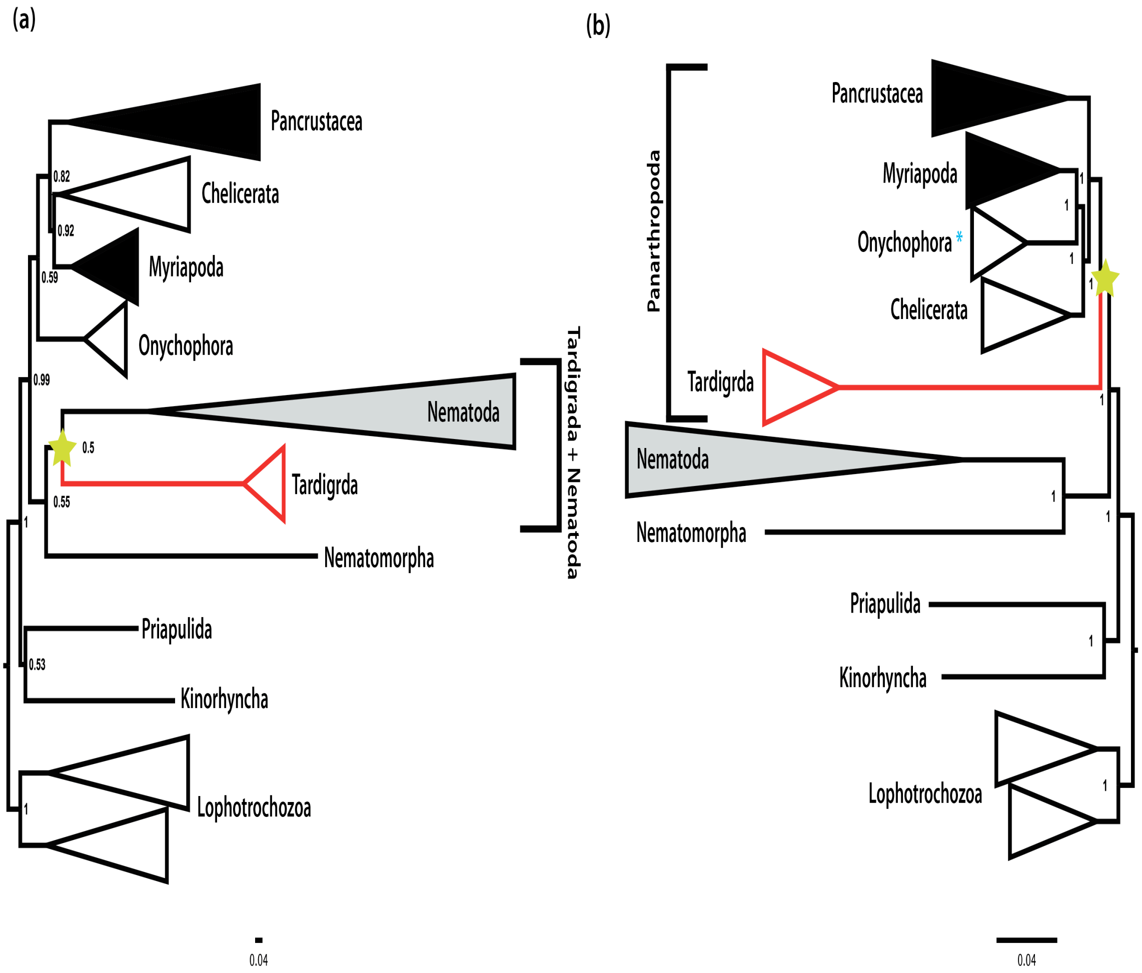
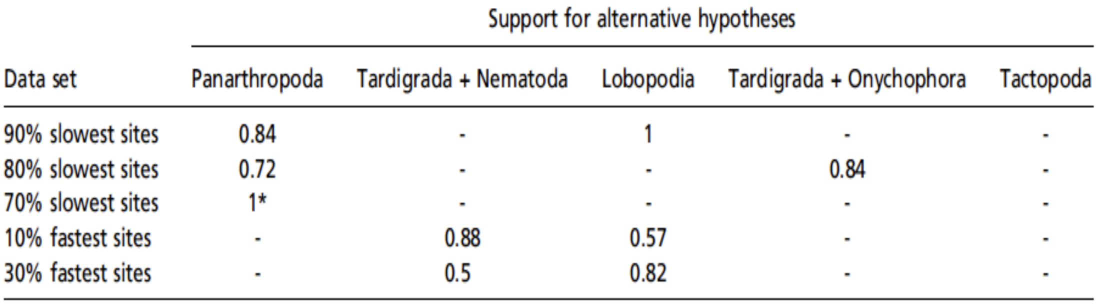


Figure 3.6: Signal dissection of EST data a larger partition of fast evolving sites. Analyses were performed under CAT-GTR+Γ. Node values are posterior probabilities. The tardigrade branch is highlighted in red, and a gold star indicates the node where tardigrades attach to the rest of the tree. (A) Analysis of the fastest 30% of sites in the alignment recovers Tardigrada as the sister to Nematoda. (B) Analysis of the slowest 70% of sites in the alignment recovers monophyletic Panarthropoda, with tardigrades as the sister of non- monophyletic Arthropoda, blue asterisks highlights artifactual position of Onychophora.



	Support for alternative hypotheses				
Data set	Panarthropoda	Tardigrada + Nematoda	Lobopodia	Tardigrada + Onychophora	Tactopoda
90% slowest sites	0.84	-	1	-	-
80% slowest sites	0.72	-	-	0.84	-
70% slowest sites	1*	-	-	-	-
10% fastest sites	-	0.88	0.57	-	-
30% fastest sites	-	0.5	0.82	-	-

Table 3.2: Summarized PP support of hypotheses of ecdysozoan evolution for all signal dissection analyses on the EST data set. *Onychophora nested inside Arthropoda. Support for Tardigrada + Nematoda is maximized by the data set including only the fastest 10% of sites. In the data set including the fastest 30% of sites, which includes some alignment positions of lower evolutionary rate, support for this group decreases substantially. On the other hand, trees derived from the data sets including only slowly evolving sites never display Tardigrada + Nematoda.

3.3.4 Taxonomic pruning and the recovery of Panarthropoda

To further test whether Tardigrada + Nematoda is an LBA artifact; a series of taxon pruning experiments was performed. These experiments were conducted by selectively removing taxa to generate uninterrupted long branches for Tardigrada, Onychophora and Nematoda (see Methods section 3.2.4). If Tardigrada + Nematoda is an LBA artifact; the results would be expected to systematically support this group (see Figure. 3.7). In summary, three different experiments designed to uncover potential sources of systematic bias in our EST alignment suggest that a nematode (or cycloneuralian) affinity for Tardigrada is most likely an LBA artifact.

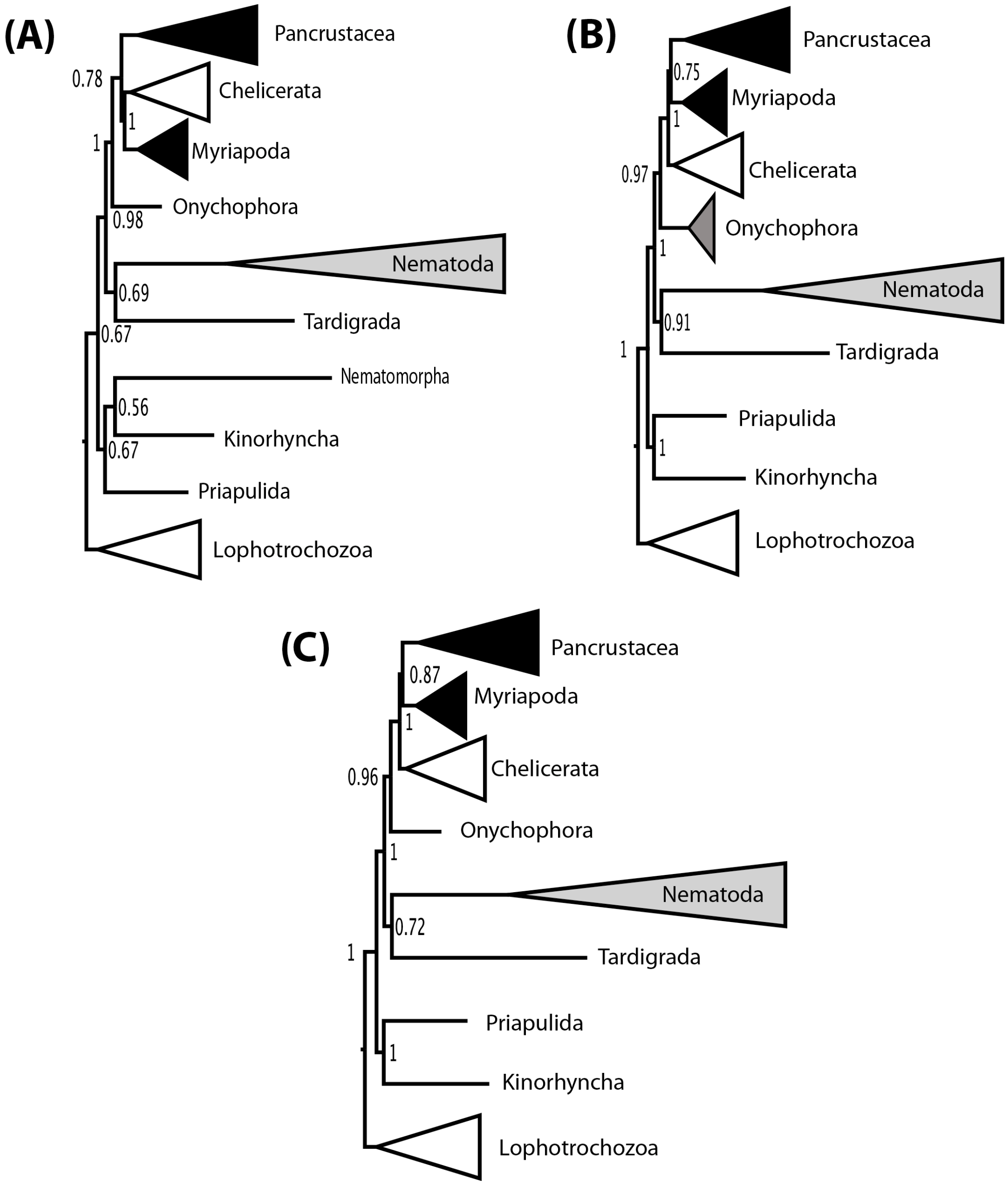


Figure 3.7: Selective taxon pruning with the aim of exacerbating LBA.

All analyses were performed under CAT-GTR+Γ. Node values are posterior probabilities. Groups with more than two taxa are collapsed for clarity. All three taxon-pruning experiments (Methods section 3.2.4) recover Tardigrada as the sister to Nematoda. (A) One onychophoran (*Epiperipatus sp.*) and one tardigrade (*Richtersius coronifer*) excluded. (B) The nematomorph (*Spinochordodes tellinii*) and one tardigrade (*R. coronifer*) excluded. (C) One onychophoran (*Epiperipatus sp.*), one tardigrade (*R. coronifer*), and the nematomorph (*S. tellinii*) excluded.

3.3.5 The effect of including *Peripatoides novaezealandiae*

Analyses of the alignment generated with *P. novaezealandiae* were conducted on the three best fitting models (GTR+Γ; CAT+Γ; CAT-GTR+Γ; see section 3.3.1 and Figure 3.2) to the exclusion of the least fitting model WAG+Γ (due to time constraints). As performed on the original alignment without the additional onychophoran species, when analyzed under the poorly fitting site homogeneous model GTR+Γ, we again observed the Tardigrada to be positioned outside Panarthropoda, instead, sister group to the Nematoida with a PP support of 0.7 (see Figure 3.8a). In addition, as the data were analysed using the better fitting site heterogeneous models CAT+Γ and CAT-GTR+Γ; PP support of 0.81 and 0.71 was recovered respectively, for Tardigrada within a monophyletic Panarthropoda (see Figure 3.8b,c). However, it must be noted here, that although support had decreased for the monophyly of Panarthropoda from full PP support of 1.0 under CAT-GTR+Γ in the original alignment to 0.71, PP support for monophyletic Mandibulata within the arthropods increased to near full PP support of 0.96. More importantly, was the recovery of Panarthropoda with Tardigrada sister group to Lobopodia with PP support of 0.99 under the less fitting model CAT+Γ model. Here, it seems that the inclusion of the single additional *P. novaezealandiae* was enough to break the sister group attraction of Onychophora with Tardigrada when analysed under CAT+Γ, a clade not recovered under the best fitting model CAT-GTR+Γ. Lastly, high support was again recovered for the paraphyletic nature of Cycloneuralia under the better fitting models CAT+Γ and CAT-GTR+Γ.

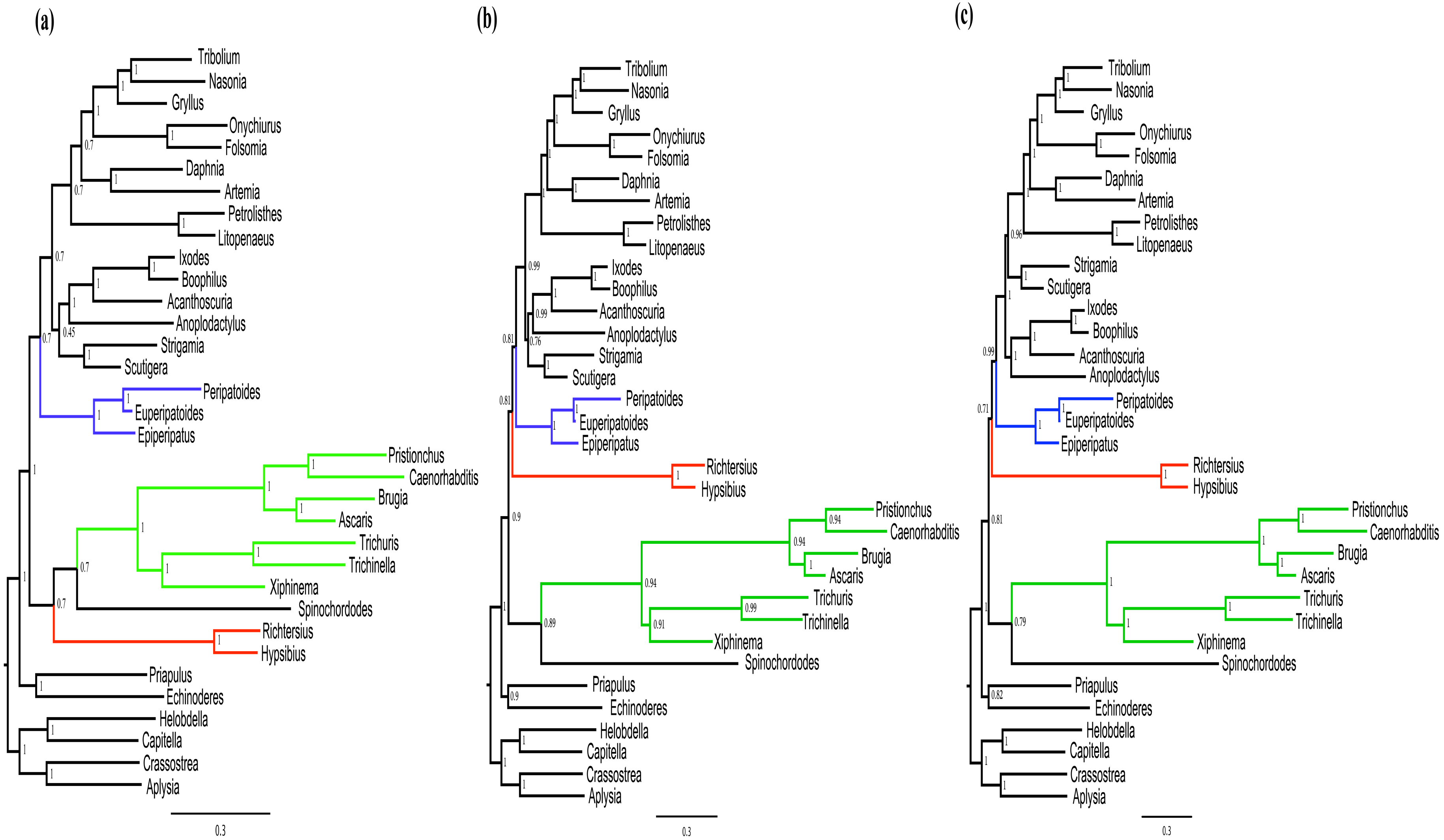
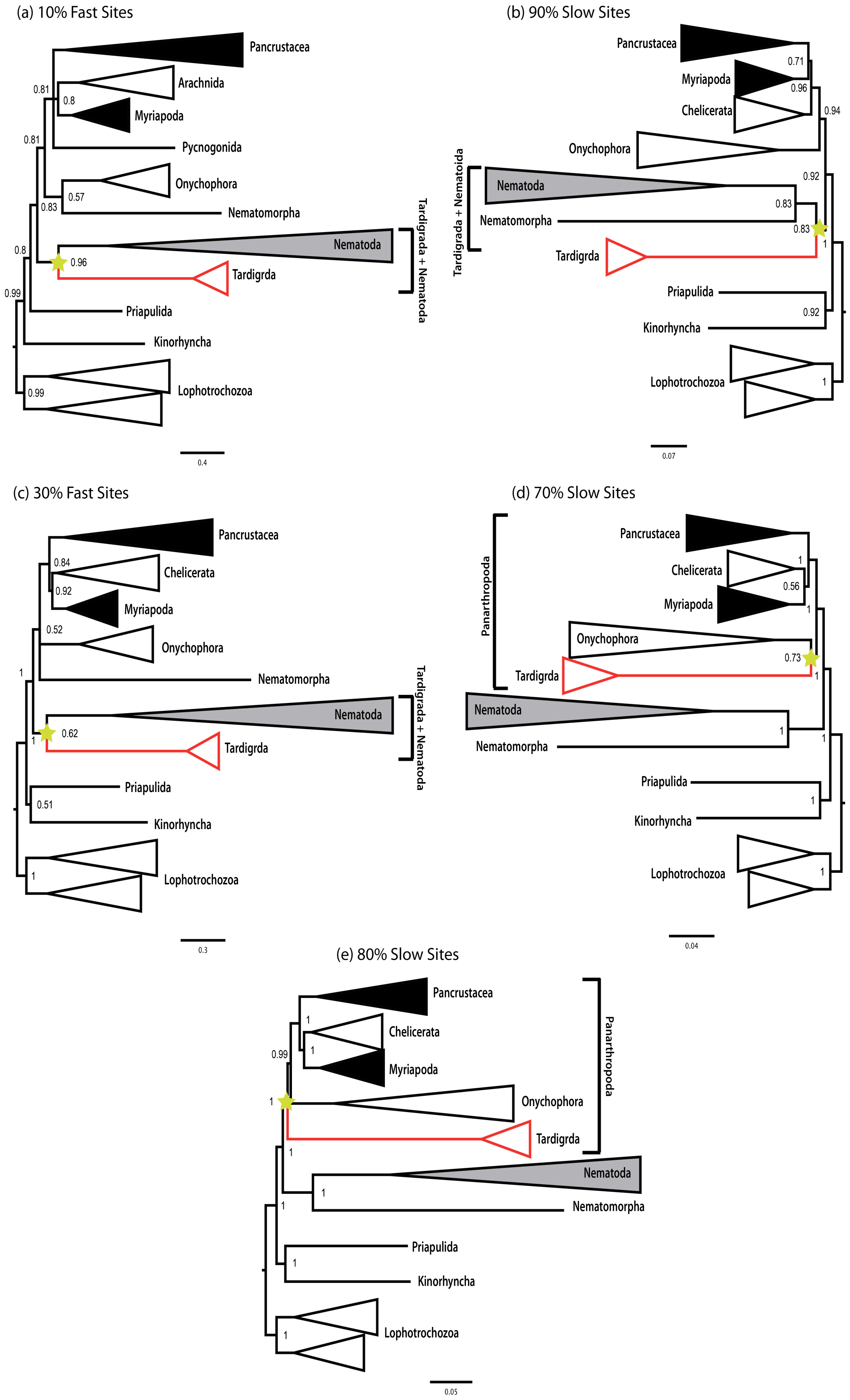


Figure 3.8: Bayesian analysis of our EST alignment including the additional onychophoran species *Peripatoides novaezealandiae*. Trees were generated under the models GTR+Γ, CAT+Γ and CAT-GTR+Γ. Topology shown is a majority rule consensus tree. Node values are posterior probabilities. (a) Analysis under the poorly fitting model GTR+Γ, supports the sister group relationship between Tardigrada + Nematoida. (b) Analysis under the better fitting CAT+Γ model, supports Tardigrda within a monophyletic Panarthropoda, sister group to Lobopodia (Onychophora + Arthropoda). (c) Analysis under the best fitting CAT-GTR+Γ model, again supports Tardigrda within a monophyletic Panarthropoda, sister group to Lobopodia (Onychophora + Arthropoda). Clade colours: red, Tardigrada; blue, Onychophora; green, Nematoda.

The data set including the additional onychophoran was subjected to the same signal dissection (Rota-Stabelli *et al.* 2010; Sperling *et al.* 2009a) analyses as performed on the original alignment, using the slow-fast technique (Brinkmann and Philippe, 1999). Sites in this alignment were again ranked and partitioned according to their rate of evolution (see section 3.2.3) and then analyzed independently. Our hypothesis for the artifactual placement of Tardigrada sister to Nematoda, as found in previous molecular analyses, if due to artifactual signal manifesting as LBA, then this grouping should also be maximized in the partitions of fast evolving sites and minimized in the partitions containing the slowly evolving sites. Results of the slow-fast analyses were again consistent with our hypothesis, fast evolving partitions of the 10% and 30% fastest support a sister group association of Tardigrada with Nematoda (see Figure 3.9a, c) with PP support of 0.96 and 0.62 respectively. Conversely, the majority of analyses for the slowly evolving site partitions recovered support for a monophyletic Panarthropoda. Not all partitions however recovered the monophyly of Panarthropoda; as in the analysis of the slowest 90% of sites (Figure 3.9b) supported with PP of 0.83 the sister group relationship of Tardigrada with a monophyletic Nematoida (Nematoda + Nematomorpha).

Here I must draw attention to new onychophoran *P. novaezealandiae, as* its branch length was quite long relative to the remaining onychophoran species, most likely due to sequencing errors possibly interpreted as autapomorphies for *P. novaezealandiae*. Accordingly, removing only 10% of the fastest sites in the alignment was not enough to reduce *P. novaezealandiae* branch length sufficiently in order to avoid a LBA artifact with Tardigrada and Nematoda. However in the analyses of the 70% and 80% slowest evolving sites, we again found high support for the monophyletic origin of

Panarthropoda; in the partition of 70% slowest sites (Figure 3.9d) we recovered a PP support of 1.0 for the sister group relationship of Arthropoda to Tardigrada + Onychophora. This topology was recovered under our second best fitting model CAT+Γ on the initial 33 taxon data set, and by some previous Mitogenomics analyses (Rota-Stabelli *et al.* 2010). In contrast, the analysis of the slowest 80% of sites, we again recovered the highly supported topology for placement of Tardigrada as the sister group to Lobopodia (Onychophora + Arthropoda) (Figure 3.9e) with a PP support value of 1.0 across all nodes, except one node connecting Onychophora to Arthropoda with had PP support of 0.99.


(a) 10% Fast Sites
Pancrustacea
Arachnida
Myriapoda
Pycnogonida
Onychophora
Nematomorpha
Nematoda
Tardigrda
Priapulida
Kinorhyncha
Lophotrochozoa
Tardigrada + Nematoda
0.81
0.8
0.81
0.57
0.83
0.8
0.96
0.99
0.99
0.4
(b) 90% Slow Sites
Pancrustacea
Myriapoda
Chelicerata
Onychophora
Nematoda
Nematomorpha
Tardigrda
Priapulida
Kinorhyncha
Lophotrochozoa
Tardigrada + Nematoida
0.71
0.96
0.94
0.92
0.83
0.83
1
0.92
1
0.07
(c) 30% Fast Sites
Pancrustacea
Chelicerata
Myriapoda
Onychophora
Nematomorpha
Nematoda
Tardigrda
Priapulida
Kinorhyncha
Lophotrochozoa
Tardigrada + Nematoda
0.84
0.92
0.52
1
1
0.62
0.51
1
0.3
(d) 70% Slow Sites
Panarthropoda
Pancrustacea
Chelicerata
Myriapoda
Onychophora
Tardigrda
Nematoda
Nematomorpha
Priapulida
Kinorhyncha
Lophotrochozoa
1
0.56
1
0.73
1
1
1
1
1
0.04
(e) 80% Slow Sites
Pancrustacea
Chelicerata
Myriapoda
Onychophora
Tardigrda
Nematoda
Nematomorpha
Priapulida
Kinorhyncha
Lophotrochozoa
Panarthropoda
1
1
0.99
1
1
1
1
1
1
0.05

←**Figure 3.9: Signal dissection of EST data including the additional onychophoran P. novaezealandiae.** Analyses were performed under the best fitting model CAT-GTR+Γ. Node values are posterior probabilities. The tardigrade branch is highlighted in red, and a gold star indicates the node where they attach to the rest of the tree. (a) Analysis of the fastest 10% of sites; recovers Tardigrada as the sister group to Nematoda. (b) Analysis of the slowest 90% of sites; recovers Tardigrada as the sister group of Nematoida. (c) Analysis of the fastest 30% of sites; recovers Tardigrada as the sister group of Nematoda. (d) Analysis of the slowest 70% of sites; recovers Tardigrada within a monophyletic Panarthropoda, sister group to the Onychophora. (e) Analysis of the slowest 80% of sites; recovers monophyletic Panarthropoda with Tardigrada sister group to Lobopodia (Onychophora + Arthropoda).

3.3.6 Bootstrap analysis of EST data set supports monophyletic Panarthropoda

In order to better understand the underlying phylogenetic signal in our EST data set, we tested how robust the highest supported topology (Figure 3.3) is under our best fitting CAT-GTR+Γ model using a 100 replicate bootstrap. Four replicates out of one hundred failed to reach convergence; however, generation of a bootstrap consensus tree including the 4 runs that failed to converge, had no effect on node support compared to the consensus tree generated with only the 96 runs that did fully converge. Consistent support for these two consensus trees was likely due to unconverged runs being affected by the unstable position of Nematomorpha.

Results of the bootstrap analysis are shown in Figure 3.10. As in the original analysis under the best fitting model CAT-GTR+Γ, Tardigrada was maintained as the sister group to Lobopodia (Onychophora + Arthropoda) within a monophyletic Panarthropoda in the majority of BS replicates; with a BS support of 66%. Although the support is low for the inclusion of Tardigrada within the panarthropods, it should be considered that, bootstrap support is well known to be over conservative, and so here a relatively low support was to be expected, given the presence of two conflicting phylogenetic signals in our data sets (as pinpointed by the signal dissection). Further to this, the relationships of Tardigrada have previously been shown to be difficult to disentangle, suggesting a general weakness of the signal with reference to this taxon. Similarly, relatively low support was found for Mandibulata within Arthropoda (BS = 64); a node that has also been shown to be affected by LBA and difficult to resolve (Regier *et al.* 2010; Rota-Stabelli *et al.* 2011) due to high levels of substitution on the short internal branch connecting Myriapoda to the

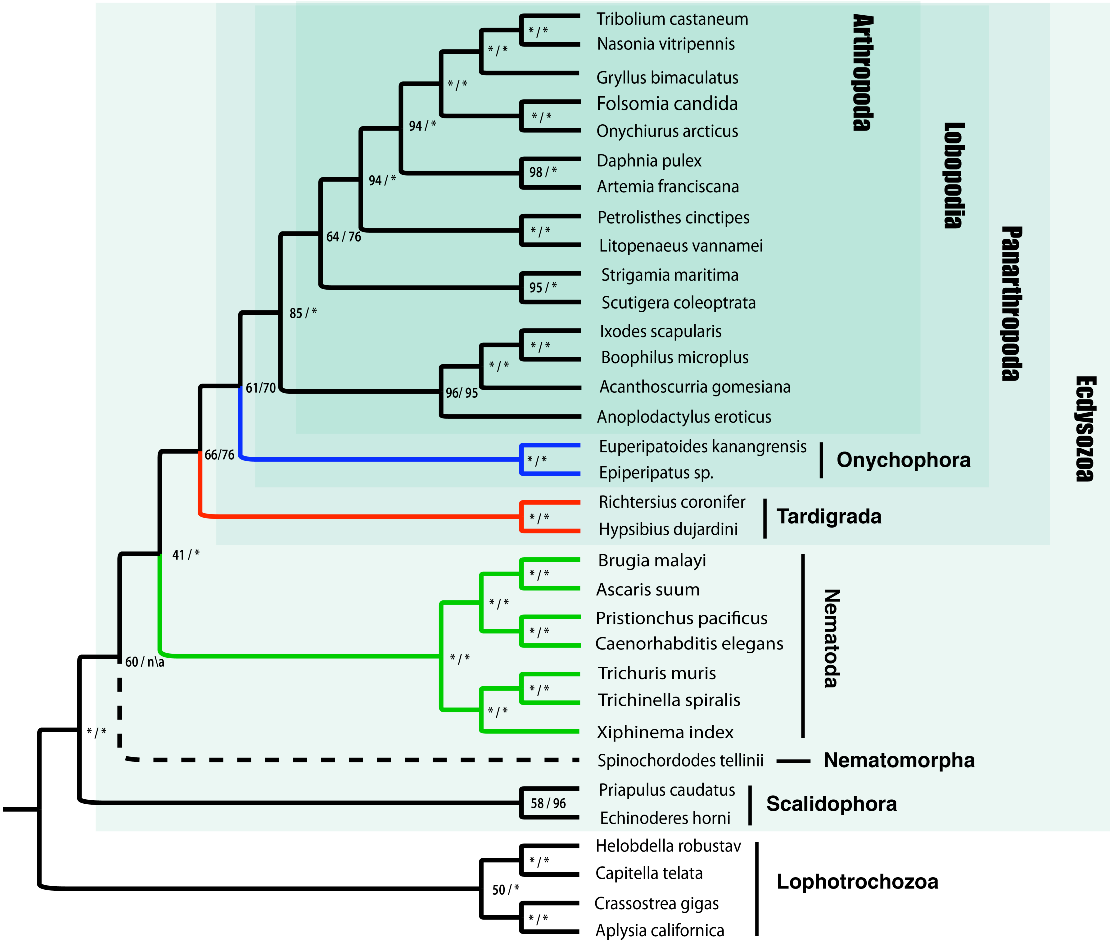


Figure 3.10: Bootstrap analysis of EST data set with 100 replicates. Analysis supports the inclusion of Tardigrada inside a monophyletic Panarthropoda sister group to Lobopodia and the unstable nature of Nematomorpha. Analysis performed using the best fitting model CAT-GTR+Γ. Node values given as the percentage out of a total of 100 replicates. Clade names given in addition to highlighted shaded regions. Support values indicated: including S. tellinii / excluding S. tellinii. Branch for S. tellinii is dashed to highlight its unstable placement. * Indicates full PP = 1.0. Onychophora branch coloured blue; Tardigrada branch coloured red; and Nematoda branch coloured green.

Pancrustacea. Also for this node our signal dissection analyses identify the presence of two conflicting signals, which can explain the relatively low support. Similar low BS support can be found for the node connecting the Nematomorpha, plausibly due to the nematomorph *S. tellinii* moving position throughout the pseudoreplicate trees because of the high amount of missing data for this taxa (93.9% missing). Due to the indication that BS support was being affected by the unstable placement of Nematomorpha, I generated a reduced consensus of the 100 bootstrap replicates after removal of *S. tellinii*. This had the effect of dramatically increasing support across the tree, particularly for Panarthropoda, Lobopodia and Nematoda + Panarthropoda. In any case, from a statistical perspective, this BS analysis provides yet further support for the inclusion of Tardigrada within the Panarthropoda (as apposed to sister to Nematoda) and the paraphyletic nature of Cycloneuralia.

3.4 Discussion

3.4.1 Systematic artifacts and the necessity of phylogenetic scrutiny

Since the advent of high throughput sequencing, molecular sequence databases now contain vast amounts of molecular sequence data; unsurprisingly ESTs are now becoming increasingly utilized to tackle a host of phylogenetic questions across the tree of life (Wolf *et al.* 2001; Bapteste *et al.* 2002; Philippe *et al.* 2004; Dunn *et al.* 2008). The use of EST data for large phylogenomic studies has a number of benefits, increasing data coverage and allowing easy expansion of taxon sampling, likely both leading to increased phylogenetic resolution (Bapteste *et al.* 2002; Rokas *et al.* 2003). Indeed use of EST data has produced many highly resolved and well-supported topologies (Philippe *et al.* 2005b; Dunn *et al.* 2008; Pick *et al.* 2010; Regier *et al.* 2010; Rota-Stabelli *et al.* 2011). However, EST based analyses do not represent the ultimate panacea to phylogeny reconstruction. It is well known that the problem of sampling or 'stochastic' error evident in many early molecular phylogenetic studies that had sparse gene sampling, is one largely alleviated by EST based phylogenomic studies (Delsuc *et al.* 2005; Philippe *et al.* 2005a; Dunn *et al.* 2008; Campbell *et al.* 2011).

Aside from the reduction in stochastic error when analysing large alignments, there is a general tendency for the increased likelihood of encountering biases (such as LBA) introduced by systematic error (Campbell *et al.* 2011; Kumar *et al.* 2011) when analysing large data sets. Systematic error is a problem of statistical inconsistency (moving towards the wrong answer as you increase the amount of data); thereby increasing the likelihood of recovering incorrect phylogenies when analysing large EST based phylogenomic data sets. EST based phylogenomic studies that utilise large numbers of genes can produce highly resolved and supported topologies, yet remain largely incongruent with one another; see (Dunn *et al.* 2008) versus (Philippe *et al.* 2009) on the phylogenetic position of the Ctenophora, or (Roeding *et al.* 2009) versus (Rota-Stabelli *et al.* 2011) on the affinity of Myriapoda within the arthropods. These incongruencies underline the ubiquitous nature of non-phylogenetic signal present in many EST datasets and suggest that analysing larger datasets is not in itself a guarantee of phylogenetic accuracy (Philippe *et al.* 2005a,b; Sperling *et al.* 2009a).

When reconstructing phylogenetic relationships, it is of utmost importance to use a model of evolution that accurately describes the data. Reconstructing difficult phylogenetic relationships requires use of a model of evolution that can describe the data, such that the model adequately employs sufficient numbers of parameters without under or over fitting the model to the data. Cases in which the model used is a poor approximation of reality, can lead to inaccurate phylogenies due to absence of key parameters and the increase of systematic biases such as LBA (Kelchner and Thomas, 2011). Furthermore, Long Branch Attraction is occasionally due to model under-fitting (Yang *et al.* 1996) particularly when inadequate taxon sampling is coupled with faster rates of substitution in one or more lineages, a situation that is more likely to mislead an analysis when the model does not include a correction for e.g. among site rate heterogeneity or compositional biases.

The focus of this study was to generate an accurate topology to describe the evolutionary relationships of Tardigrada within Ecdysozoa, while also identifying potential phylogenetic tree reconstruction artifacts that may explain why previous studies obtained conflicting hypotheses for the evolution of Tardigrada (Roeding *et al.* 2005; Meusemann *et al.* 2010; Andrew, 2011; Zrzavý *et al.* 1998; Mallatt and Giribet, 2006; Dunn *et al.* 2008; Rota-Stabelli *et al.* 2010; Rota-Stabelli *et al.* 2011). Using Bayesian crossvalidation (Stone, 1974) to rank the fit of evolutionary models to the data, I show how the choosing between different models results in the recovery of two highly supported alternate positions for Tardigrada within Ecdysozoa. Site-homogeneous models (WAG+Γ, GTR+Γ) that inadequately account for multiple hidden substitutions consistently recover (Lartillot *et al.* 2007) with full support, the sister group position of tardigrades to nematodes. Conversely site-heterogeneous mixture models (CAT+Γ, CAT-GTR+Γ) that can account for the effects of across site rate heterogeneity and compositional bias (Lartillot and Philippe, 2004) while also having a sizable improvement of fit to our data, recover the tardigrades as members of a monophyletic Panarthropoda.

Potential biases associated with taxon sampling and differential rates of substitution among the sites of our alignment were also explored. Methods employed in these analyses to uncover phylogenetic biases again strongly favour a sister group relationship of Tardigrada to Lobopodia (Onychophora + Arthropoda). One of the main problems of phylogenetic reconstruction of ancient relationships (as is the case for Ecdysozoa) is how to uncover genuine phylogenetic signal amidst the large amount of phylogenetic noise (Brinkmann and Philippe, 1999), as genuine signal can be drastically erased by millions of years of hidden multiple substitutions. According to our signal dissection analyses, which were designed to increase the signal-to-noise ratio of the data set, showed that phylogenies generated using more reliable slowly evolving sites consistently recovered Tardigrada as a member of Panarthropoda as apposed to the sister group to Nematoda. Conversely, analyses conducted on less reliable fast evolving sites (increasing the noise to signal ratio) supported the nematode affinity of tardigrades.

Furthermore, the effect of ingroup taxon sampling was investigated to uncover additional sources of systematic bias. It has been shown that the benefits of increased taxon sampling are highly advantageous in phylogeny reconstruction. Increased taxon sampling has the property of breaking up potential long branches by reducing the length of long internal nodes and in doing so reduce the incidence of systematic biases (Pollock *et al.* 2002; Zwickl and Hillis, 2002); the most common of which being LBA. The use of targeted taxonomic pruning (reducing species sampling for Onychophora, Tardigrada and the sister group to nematodes i.e. Nematomorpha) demonstrated that as specific taxa are removed thereby increasing internal branch lengths, resulted in Tardigrada being recovered outside Panarthropoda sister to Nematoda. The shifting position of Tardigrada towards the cycloneuralian Nematodes in these analyses was independent of the choice of evolutionary model, as even our best fitting CAT-GTR+Γ model recovered a nematode affinity for Tardigrada. Reanalysis including the additional onychophoran species (*P. novaezealandiae*) further demonstrated the substantial effect that taxon sampling had on the recovery of alternate hypotheses for tardigrade evolution. Support was bolstered for tardigrades as the sister group to Lobopodia, not only under our best fitting CAT-GTR+Γ model, but also under the poorer fitting CAT+Γ model, which supported a sister group relationship between Tardigrada and Onychophora (as in Rota-Stabelli *et al.* 2010 and Rota Stabelli *et al.* 2011) in analyses where *P. novaezealandiae* was not included.

3.5 Conclusion

To conclude, EST data support Tardigrada as a member of Panarthropoda. Given the pervasiveness of systematic artifacts, care must then be taken when evaluating topologies derived from large alignments, particularly when multiple highly supported competing hypotheses have been proposed. This is the case with the tardigrades, where molecular homoplasy certainly exists, as demonstrated by the fact they are recovered by previous molecular analyses in two highly discordant positions within the Ecdysozoa (Roeding *et al.* 2007; Sørensen *et al.* 2008; Hejnol *et al.* 2009; Meusemann *et al.* 2010; Andrew, 2011; Zrzavý *et al.* 1998; Mallatt and Giribet, 2006; Dunn *et al.* 2008; Rota-Stabelli *et al.* 2010; Rota-Stabelli *et al.* 2011).

Differently from most previous investigations, the results presented herein, were generated using methods designed to uncover systematic biases (LBA) in the data. I thus feel confident in concluding the affinity for tardigrades lies with the Arthropoda and Onychophora (i.e. Panarthropoda) and not with the cycloneuralian nematodes.

Chapter 4

Phylogeny reconstruction using microRNAs: Testing competing hypotheses of arthropod and panarthropod evolution

4.1 Introduction

Since the emergence and proliferation of molecular sequence data due to next generation sequencing technology advancements (Metzker, 2009) many difficult phylogenetic questions in animal evolution have been resolved (Sperling *et al.* 2009a; Regier *et al.* 2010; Rota-Stabelli *et al.* 2011; Philippe *et al.* 2011b). For instance, the arrival of the 'new animal phylogeny' and the move away from the traditional 'Coelomata' hypothesis was the result of the availability of new evidence, e.g. phylogenomics (see section 2.2.3 of Chapter 2) and developmental studies. However, despite the current availability of multiple different data types many questions still remain to be answered. There are numerous issues that can cause a phylogenetic problem to be particularly difficult (Philippe *et al.* 2005a; Philippe *et al.* 2011b), all of these could be placed under the heading of homoplasy (similarity due to convergent evolution). For example, problems arising from the analysis of molecular sequences under inadequate models (Tuffley and Steel, 1998; Lartillot and Philippe, 2004) presence of systematic biases like long-branch attraction, or use of ambiguous (i.e. evolved convergently) morphological characters (Scotland *et al.* 2003) (e.g. Atelocerata hypothesis: Klass and Kristensen 2001; Bitsch and Bitsch 2004; or Uniramia Hypothesis, Tiegs 1947), are all fundamentally caused by homoplasy. Since the true metazoan phylogeny is unknowable and given the pervasiveness of homoplasy in every type of data (Jenner, 2004), it is clear that in order to answer difficult phylogenetic questions we must look all available evidence in order to investigate pattern of congruence and incongruence among different data types (i.e. the concept of consilience). As truth is impossible to be known with certainty, convergence of alternative, independent, lines of evidence subjected to different biases, is our best proxy for phylogenetic accuracy (Wilson, 1988; Campbell *et al.* 2011).

In this Chapter I will further introduce the use of a recently emerged source of novel phylogenetic data: the genomic regulatory elements called microRNAs (miRNA). I will use these new data to tackle the unresolved competing hypotheses of evolution among the four arthropod sub-phyla (Hexapoda, Crustacea, Myriapoda and Chelicerata), and to test alternative competing hypotheses for the phylogenetic relationships of the Phylum Tardigrada within Ecdysozoa. In so doing, I will be introducing a new data type to address the problem of the evolution of the Ecdysozoa. My aim here is to identify whether miRNAs can corroborate/reject previous hypotheses derived using more traditional morphological and molecular sequence data.

Two problems will be addressed. The first is the problem of the affinity of the Myriapoda. Two major hypotheses have been proposed in the past (Mandibulata and

Myriochelata – see Nielsen, 2001 and Pisani *et al.* 2004). Of these hypotheses *Mandibulata* (Snodgrass, 1938) is the more traditional as most morphological characters are easily explained on a tree topology displaying Mandibulata. Conversely, the second hypothesis Myriochelata, can only explain a small number of morphological characters; if the possibility of convergent evolution is not considered (Dove and Stollewerk, 2003; Kandar and Stollewerk, 2004; Stollewerk and Chipman, 2006; Mayer and Whitington, 2010). Myriochelata has been recovered predominantly by molecular sequence analyses (Pisani *et al.* 2004; Mallatt, 2004; Lartillot and Philippe, 2008; Dunn *et al.* 2008; Hejnol *et al.* 2009).

Carrying on from internal arthropod phylogeny, the second problem addressed here, is the study of the relationships of the Arthropoda, the Onychophora (velvet worms) and the Tardigrada (water-bears) within the context of the Ecdysozoa. Onychophora, Tardigrada and Arthropoda, have long been recognised, on the grounds of morphology and developmental biology, as being close relatives. This group was named Panarthropoda by Nielsen (2001) and possesses features such as paired ventrolateral walking appendages and *engrailed* expression (Gabriel and Goldstein, 2007) in the posterior ectoderm of each articulated segment, representing the proposed panarthropod apomorphies. However, competing hypotheses on the placement of tardigrades have also seen them positioned outside Panarthropoda as the sister group to the phylum Nematoda. This grouping of Tardigrada + Nematoda is primarily based on a number molecular sequence analysis (Sørensen *et al.* 2008; Hejnol *et al.* 2009; Meusemann *et al.* 2010). Unfortunately, morphology alone cannot help to resolve the position of Tardigrada. This is because this phylum shares characteristics not only with the Panarthropoda, but also in with the Cycloneuralia (Nielsen, 2001) the group to which the Nematoda are generally ascribed. Putative apomorphies of the Nematoda plus Tardigrada group include a circumesophageal brain, but also a telescopic mouth cone and plated pharynx (e.g. Schmidt-Rhaesa, 1998; Campbell *et al.* 2011). In addition, even studies that agree upon the general placement of the Tardigrada within Panarthropoda disagree on the precise relationships among the three panarthropodan taxa (Arthropoda, Tardigrada and Onychophora) (see Budd, 2001; Mallatt and Giribet, 2006; Dunn *et al.* 2008). Here, I will show the results of a phylogenetic analysis on the distribution of miRNA genes present throughout both Arthropoda, and more broadly across the Panarthropoda with regard to other ecdysozoan phyla; namely Nematoda and Priapulida. The results presented here have been published in the peer-reviewed journals; *Proceedings of the Royal Society B: Biological sciences* (Rota-Stabelli *et al.* 2011) and *Proceedings of the National academy of Sciences* (Campbell *et al.* 2011).

4.1.1 MicroRNAs: Function and Biogenesis

MicroRNAs (miRNAs) are single-stranded RNAs of ~19-25 nucleotides (nt) in length that are generated from endogenous hairpin-looped transcripts (Lee *et al.* 1993; Bartel 2004), see Figure. 4.1 for a typical miRNA secondary structure. MicroRNAs were originally identified for their role in developmental timing in *Caenorhabditis elegans* (Lee *et al.* 1993), where the miRNA *lin-4* was identified as the key regulator of the gene product *lin-14* via the numerous complementary binding sites of *lin-4* present within the *lin-14* 3' untranslated region (3' UTR). Following the discovery of *lin-4* a second miRNA called *let-7* was identified as a ~22 nt regulatory RNA (Reinhart *et al.* 2000) again shown to regulate developmental timing in *C. elegans*. It was not clear at the time if these kinds of small regulatory RNAs were a peculiarity specific to nematode worms or a feature more commonly seen across the Metazoa. This however was clarified, revealing the extent of conservation of this kind of RNA regulation upon identification of the miRNA *let-7* in divergent bilaterian taxa (Pasquinelli *et al.* 2000). At this stage these small regulatory RNAs were not called miRNAs, they were instead referred to as small temporal RNAs (stRNAs) due the shared role in developmental timing. It wasn't until the cloning of sets of similar small regulatory RNAs in divergent model organisms such as humans, flies and nematodes (Lagos-Quintana *et al.* 2001; Lau *et al.* 2001; Lee and Ambros, 2001) which had similar properties to *lin-4* and *let-7* (~22 nt in length, processed from one arm of the hairpin RNA) but differed in that they were not expressed in distinct developmental stages, which prompted the introduction of the name microRNA to classify these regulatory RNAs of unknown function.

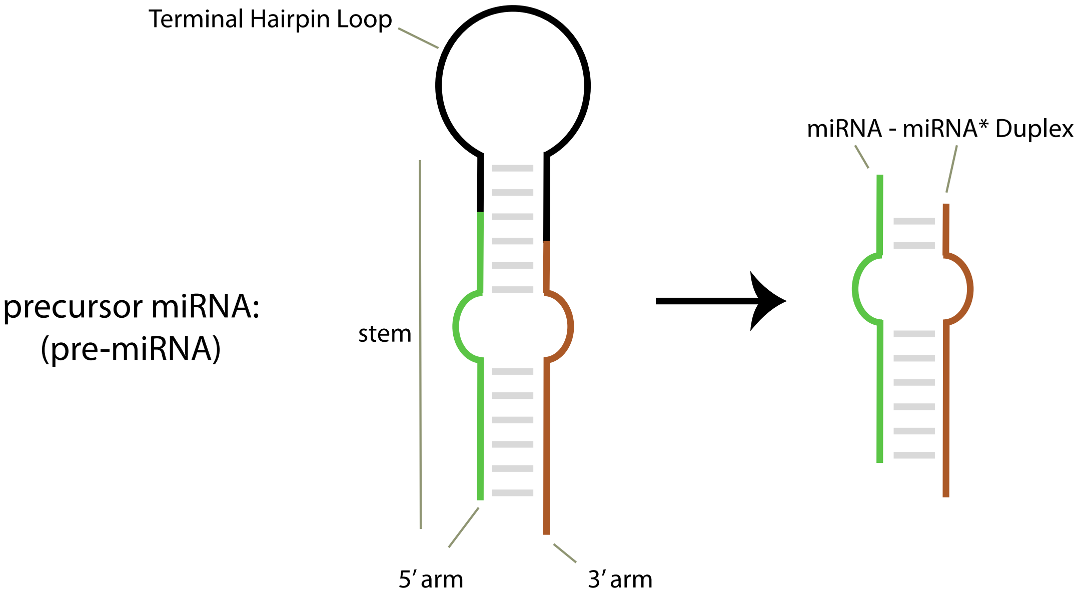


Figure 4.1: Typical microRNA secondary structure. Watson-Crick base paring shown between two arms of the pre-miRNA (green and brown) with a central miss-match bulge.

Since their discovery, miRNAs have now been shown to be crucial regulators in a multitude of different physiological processes such as developmental timing, neuronal patterning, cell proliferation, apoptosis, tissue differentiation and cell signalling (Bartel, 2004). MicroRNAs function as posttranscriptional repressors of their target genes when bound to the specific sites present in the 3' UTR of the target messenger RNA (mRNA) (Berezikov 2011). In Metazoa, miRNA mediated silencing of mRNAs is usually achieved by imperfect base paring to the 3' UTR, thereby blocking the access of the target mRNA to the translational machinery (Lee *et al.* 1993). However depending on the degree of base complementarity, metazoan miRNAs can also direct catalytic cleavage (Bartel, 2009; Brodersen and Voinnet, 2009). Individual miRNAs may regulate up to hundreds of different loci, and it has been estimated that a majority of human genes are potential miRNA targets (Lim *et al.* 2005; Lewis *et al.* 2005). The diversity of the different physiological processes that miRNAs coordinate is evident, yet miRNA regulation follows a single strict pathway of biogenesis.

Before moving forward to discussing the stages of miRNA biogenesis, I must clarify that miRNAs are biogenically defined. In short, a nucleotidic sequence represents a miRNA gene if it produces an RNA with a secondary hairpin structure that is identifiable by the miRNA biogenesis machinery, which will transform it into a functional miRNA effecter complex (Kim, 2005). This statement is crucial as only these types of genes can be recognised and processed correctly into the miRNA effecter complex, therefore enabling translational repression.

MicroRNAs are genomically encoded non-protein coding genes located in various regions within a genome. Early studies showed that many identified miRNA genes were located in distinct intergenic regions or with an antisense orientation to annotated genes, indicating that those miRNAs derive from independent transcription units (Lagos-Quintana *et al.* 2001; Lau *et al.* 2001; Lee and Ambros, 2001). However it is now known that a sizeable minority of miRNAs are located within intron regions of protein coding genes (PCGs) usually in the same sense orientation suggesting that those miRNAs are co-transcribed with their associated PCG (Rodriguez *et al.* 2004). In addition it is now understood that many miRNAs are also located in close proximity to one another, arranged and transcribed in a pattern suggesting that transcription occurs via a multi-cistronic primary transcript. As for all PCGs transcription of a miRNA is primarily mediated by RNA polymerase II (Lee *et al.* 2004). However, occasionally miRNAs have been observed to be transcribed by RNA polymerase III (Lee *et al.* 2004; Borchert *et al.* 2006).

The stages of miRNA biogenesis (described below; but see Figure. 4.2) begins with the transcription of a miRNA locus by RNA polymerase II, resulting in a long primary transcript or primary miRNA (pri-miRNA; Lee *et al.* 2002) that is usually several kilo bases long containing local hairpin structures. Primary miRNAs are also capped and polyadenylated in typical Pol II fashion. Pri-miRNAs fold into characteristic hairpin-like structures, providing the basis of recognition by the RNase III enzyme complex *Drosha*. Pri-miRNA transcription is followed by processing or 'cleavage' by *Drosha* to liberate a shorter ~60-70 nt stem loop intermediate known as a precursor miRNA (pre-miRNA; Kim, 2005). Evidence suggests that the tertiary structure of the pri-miRNA allows the recognition by *Drosha* to cleave out the pre-miRNA and subsequently a downstream functioning miRNA.

Following the nuclear processing by *Drosha* the pre-miRNA is then exported from the nucleus into the cytoplasm by a nuclear pore complex mediated by the nuclear transport receptors *exportin-5* (Kim, 2005). After the pre-miRNA enters into the cytoplasm it is then acted upon by another RNase III enzyme called *Dicer* (Bartel, 2004). Dicer cleaves the pre-miRNA by loping off the terminal base pairs and the stem loop; this generates a shorter ~20-24 nt mature miRNA duplex. This Duplex contains the mature miRNA, which is associated with its reverse complement sequence known together as the miRNA:miRNA* duplex ("miRNA-miRNA Star"). The pre-miRNA duplex cleaved by Dicer contains a staggered cut typical of RNase III endonucleases, with the base of the pre-miRNA stem loop characterized by a 5' phosphate and a 2 nt 3' overhanging tail end (Filipowicz, 2008).

It has been shown that the RNase III enzyme Dicer is associated with a number of different proteins which function not in the catalytic cleavage of pre-miRNAs but miRNA stability and effecter complex formation (Kim, 2005). One of the most important Dicer associated protein families is the Argonaute family, with the Argonaute protein *Ago2* shown to function as the 'slicer' enzyme that cleaves target mRNA (Song *et al.* 2004). The role of *Dicer* in miRNA biogenesis is conserved across all animals (also in plant miRNA biogenesis) however recently it has been demonstrated for the first time that the miRNA miR-451 (present in mammals) is generated independent of *Dicer* activity, instead relying upon the endonuclease activity of Ago2 (Cheloufi *et al.* 2010; Bossé and Simard, 2010).

Mature miRNAs are then incorporated into the effecter complex known as 'miRNP' (miRNA-containing ribonucleoprotein complex) or miRISC (miRNA-containing RNA-induced silencing complex). Once a miRNA:miRNA* duplex is formed by *Dicer*, strand selection occurs resulting in a single arm of the duplex being incorporated into the miRISC complex. Usually the duplex does not persist long in the cell, as one strand of the duplex (usually the miRNA*) will degrade

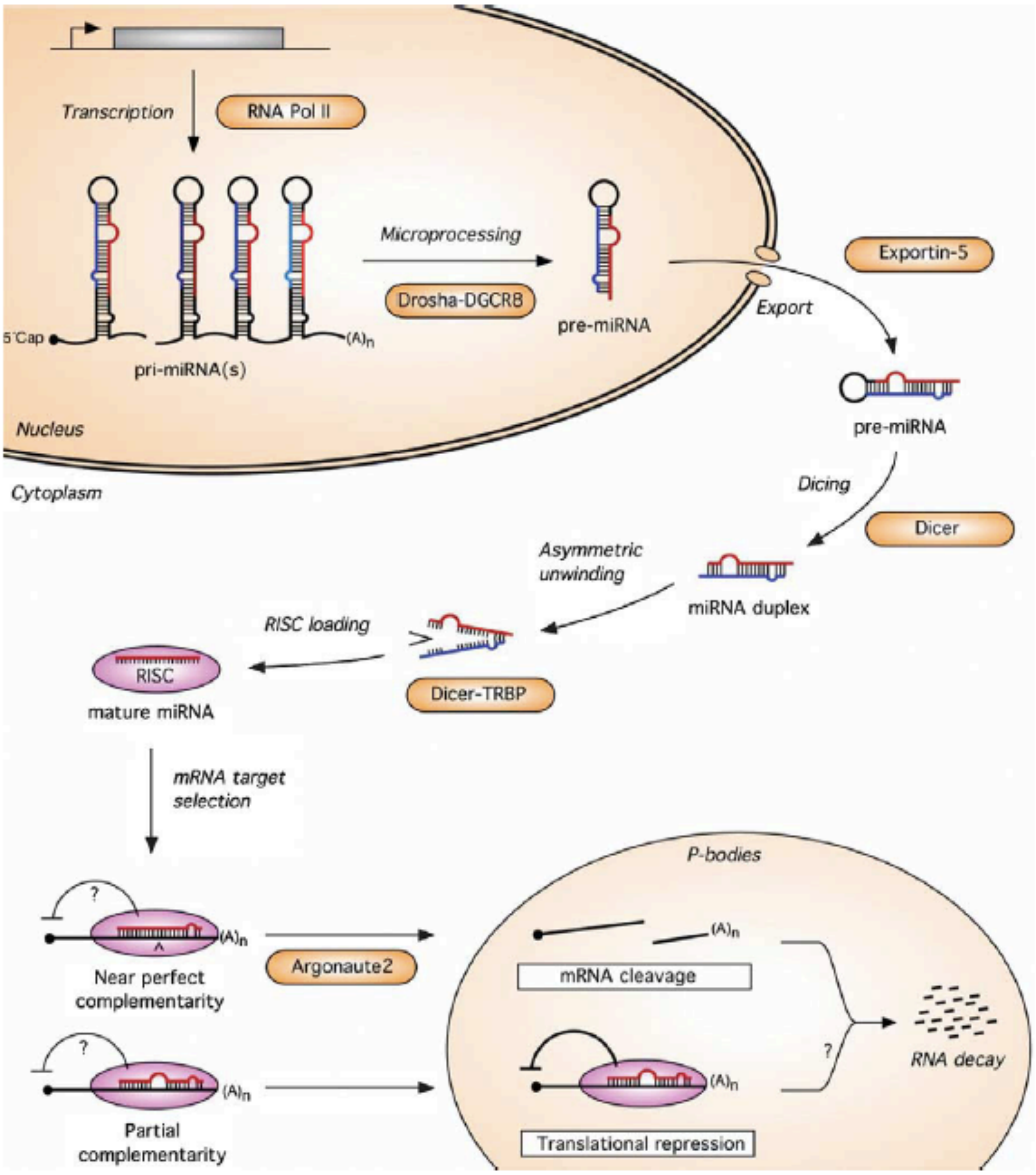


Figure 4.2: Typical metazoan miRNA biogenesis pathway. Figure taken from (Wienholds and Plasterk, 2005). See text for the details of the stages of miRNA biogenesis.

(Filipowicz, 2008) whereas the other will be selected as the mature miRNA. MicroRNA Strand selection is not fully understood, but mounting evidence indicates that selection of a mature miRNA strand resides in the relative stability of the two ends of the miRNA:miRNA* duplex (Bartel *et al.* 2004; Kim, 2005). The strand selected for the miRISC complex is usually the one whose 5' end is less tightly paired (for example, G:U pair vs. G:C pair; Khvorova *et al.* 2003; Schwarz *et al.* 2003). Interestingly, the miRNA* sequence which is essentially a palindrome of the mature miRNA has been shown on occasion to be incorporated into the miRISC complex also, with studies showing new miRNA loci can be generated via antisense transcription of existing miRNA genes (Berezikov, 2011).

Once a mature miRNA is loaded into the miRISC complex, it can then be used to locate its target sequence(s). In Metazoa, miRNAs target mRNA transcripts by imperfect base-pairing to multiple sites within the target 3' UTR regions. Specifically, there are two regions of a mature miRNA sequence that are crucial for effective target binding, these are the "seed" region (Lewis *et al.* 2003) located usually in nt positions 2-8 on the 5' arm, while the second region also shown to be important to target binding is located in nt positions 13-16 in the 3' arm. Targeting of a miRNA to locations within a 3' UTR was first observed in the earliest discovered miRNAs, *lin-4* and *let-7*. However it is now currently known that miRNA target sites, although usually located in 3' UTRs, can also be found in other locations such as open reading frames (ORFs) as seen in *Drosophila* (Stark *et al.* 2007) and vertebrates (Forman *et al.* 2008); but this seems to be the exception rather than the rule (Filipowicz *et al.* 2008).

The degree of complementarity of Watson-Crick base paring between a miRNA and its target, specifically in regions such as the 'seed' and 3' compensatory region, are fundamental to conferring regulation. Perfect complementarity between the seed region of a miRNA and its target have been shown to be sufficient to confer target regulation (Brennecke *et al.* 2005) however, compensatory pairing between the 3' miRNA region (nt 13-16) is required when mismatches occur in the 5' seed region. The importance of the Watson-Crick base pairing in these two regions has been highlighted, with the seed and 3' compensatory regions shown to be the most conserved nt positions in a miRNA (Wheeler *et al.* 2009); with the number of substitutions in these regions considerably lower compared to the remaining nt positions. In addition, documented instances of a modified seed region or 'seed shifts' have been made, seed region starting positions can be modified by moving them in a 3' or 5' direction usually by insertion of 1 – 2 nt. Importantly though, seed shifts are conserved evolutionary events (Wheeler *et al.* 2009) highlighting the importance of seed regions in mRNA targeting. Lastly, the mode of miRNA regulation has been correlated to the degree of base-pair complementarity. Plant miRNAs in contrast to animal miRNAs usually repress their targets by binding with near perfect complementarity, thereby inducing target cleavage; whereas this mode of target cleavage is rarely observed in animals (Filipowicz *et al.* 2008). It seems now that contrary to general belief, it is not the degree of base-complementarity *per se,* in animal miRNA targeting, but the presence of central base-pair mismatches in the miRNA-target interaction (Bordersen and Voinnnet, 2009). This prediction is consistent with structural models that suggest that the RNase active site in the miRISC complex is located ~ 10 nt from the beginning of the miRNA (Song *et al.* 2004); therefore located between the 5' seed or 3' compensatory regions that have been shown to be crucial for miRNA target specificity in animals.

4.1.2 MicroRNAs in phylogeny reconstruction

Ever since their discovery, miRNAs have been scrutinized for their properties of gene regulation in a wide variety of physiological roles (Bartel, 2004). However it is only recently that miRNAs have been seen as promising genomic markers for phylogeny reconstruction. Today it is now widely accepted that the addition of new miRNAs to genomes has been commonplace since the emergence of the Metazoa (Hertel *et al.* 2006; Sperling *et al.* 2010), but specifically the expansion of miRNA repertoires to eumetazoan (Grimson *et al.* 2008; Peterson *et al.* 2009) and bilaterian genomes (Christodoulou *et al.* 2010) has been more dramatic. The continual addition of miRNAs to genomes was apparent in early studies (Sempere *et al.* 2006) and since then has been confirmed in numerous investigations (Hertel *et al.* 2006; Wheeler *et al.* 2009; Sperling *et al.* 2009b; Sperling *et al.* 2010; Heimberg, 2010; Philippe *et al.* 2011a). The upshot of this is in terms of phylogeny reconstruction is that nearly every metazoan clade thus far investigated can be characterized by at least one new miRNA family acquisition, making these characters extremely useful for resolving phylogenetic relationships. However it is important to note that the rate of acquisition of families is not constant between taxa, with different lineages experiencing different rates of acquisition (Tarver *et al.* 2012). Furthermore, it is important to point out that contra to some initial claims, miRNA data sets, despite being homoplasy low, are not homoplasy free (e.g. Philippe *et al.* 2011a). Yet, the rate of acquisition of new miRNA families substantially outweighs the rate of losses (Campbell *et al.* 2011).

The major benefit for using miRNAs to recreate phylogenies is in regards to their mode of biogenesis, and the fact that the recognition of a miRNA for processing by the biogenesis machinery (i.e. RNase III enzymes *Drosha* & *Dicer*) relies upon the miRNA stem loop structure and not the primary miRNA sequence. This greatly increases the utility of miRNAs for phylogeny as it negates the need for a researcher to know any particular miRNA sequence prior to sequencing and analysis. Studies investigating the expansion and conservation of miRNAs and miRNA families throughout different groups of animals have led to the realization of a number of evolutionary characteristics endowing miRNAs with a level of phylogenetic utility rivalling the most commonly used phylogenetic data. MicroRNAs have four characteristics that make them exceptional phylogenetic candidates to resolve conflicting hypotheses of evolution or even provide fresh hypotheses previously overlooked: (*i*) miRNA families are continuously added to genomes throughout time, (*ii*) secondary loss of a miRNA is rare once acquired within a genome, (*iii*) Once acquired the mature miRNA sequence accumulates mutations very slowly, and (*iv*) there is a massively low probability of independent convergent evolution of any particular miRNA in separate lineages. Due to the aforementioned properties of miRNA evolution (also see section 2.2.2 of Chapter 2) miRNAs are endowed with the potential ability to overcome pitfalls of using traditional data types of phylogeny reconstruction (Sperling and Peterson, 2009), see Figure. 4.3 for a pipeline of the implementation of miRNAs in phylogenetic analysis.

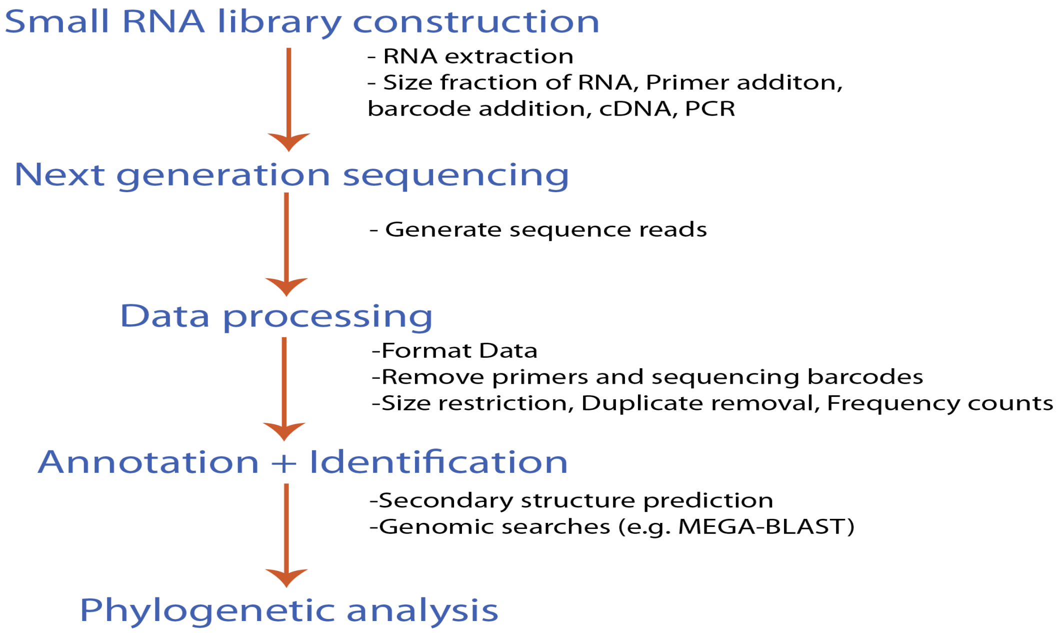

Small RNA library construction
- RNA extraction
- Size fraction of RNA, Primer additon, barcode addition, cDNA, PCR
Next generation sequencing
- Generate sequence reads
Data processing
-Format Data
-Remove primers and sequencing barcodes
-Size restriction, Duplicate removal, Frequency counts
Annotation + Identification
-Secondary structure prediction
-Genomic searches (e.g. MEGA-BLAST)
Phylogenetic analysis

Figure 4.3: Flowchart pipeline of the typical stages involved in implementing miRNAs in phylogeny reconstruction.

Apart from continual lineage specific miRNA expansion, it has been shown that once a miRNA gene emerges and is incorporated into a specific lineage gene regulatory network, it is rarely secondarily lost in the descendent lineages (Sempere *et al.* 2007, Heimberg *et al.* 2008, Wheeler *et al.* 2009). However, lineage specific loss has been observed; with the absence of specific miRNA families (27 losses from 36 families) recently shown in the Acoela flatworms. This instance of major miRNA loss in the Acoel *Symsagittifera roscoffensis* can be met with a caveat, in that large-scale secondary simplification of this species (Philippe *et al.* 2011a) was not enough to completely lose all derived miRNAs. The acoel flatworm *Symsagittifera roscoffensis* and the species *Xenoturbella* were both found to posses the miRNA *miR*-103, a deuterostome specific miRNA suggesting the placement of Acoela resides within Deuterostomia (Philippe *et al.* 2011a), in contrast to previously posited competing hypotheses which placed Acoels as either a group or grade of basal bilaterians or associated them with the Platyhelminthes (Baguna and Riutort, 2004). This deuterostome placement of *Xenoturbella* and the Acoel worms might seem counterintuitive, but it has been confirmed by the analyses of nuclear protein coding genes and mitogenomic datasets (Bourlat *et al.* 2006).

Many studies have shown how miRNAs can regulate up to hundreds of different genes (Lim *et al.* 2005; Lewis *et al.* 2005) and so because miRNAs regulate so many different transcripts they must be able to retain the ability to interact with all transcripts 3' UTRs. This necessity of sequence conservation therefore makes it difficult to lose a miRNA or change its primary sequence. Loss of miRNAs can occur as previously mentioned, however evidence suggests that loss of miRNAs is more likely for species that have undergone significant secondary morphological simplification such as seen in the case of Acoela. This suggests that the mosaic pattern of miRNA loss is related to the reduced number of gene targets for those miRNAs (Sperling and Peterson, 2009). It is here that a critical distinction must be made when concerned with secondary loss of miRNAs: we must be able to clarify between genuine lineage specific secondary loss in contrast to the apparent loss as a consequence of searching for example an incomplete genome or small RNA library.

By far the most effective way to detect miRNAs for a species is by using small RNA sequencing, such as NGS technologies like Illumina and 454. However depending on the developmental stage and or tissue sampled not all miRNAs may be expressed and thus identified; this is usually not a problem as the depth of sequencing with NGS technologies should ensure that even the most lowly expressed miRNAs are detected (Berezikov *et al.* 2006). Problems can be introduced however when searching (using BLAST) previously identified miRNAs against a known genome to identify orthologues. Tarver *et al.* (2012) have shown that the level of genome coverage is a major factor when identifying miRNA orthologues, with high coverage genomes (~7x) missing on average 5.16 miRNA families in contrast to low coverage genomes (~2x) missing on average 26 families. Thus presumed instances of secondary loss may in fact be false negatives i.e. failing to detect a miRNA due to the incomplete nature of some genomes.

There is now mounting evidence on the different ways new miRNAs and miRNA families can arise, with the ease of RNA to form into a stable fold-back structure, indicating novel miRNA genes may actually be more likely to arise than a protein-coding gene. MicroRNAs that contain significant sequence homology to each other in the mature region are grouped into families (Ambros *et al.* 2003), with new miRNAs and miRNA families arising via a number of evolutionary processes. One of the major sources of novel miRNAs is via gene duplication; these events are then usually followed by sub- and neo-functionalization (Ruby *et al.* 2007) of the acquired miRNA. Yet this only increases the dimension of an existing family. Interestingly, as many miRNAs are located within intronic regions (Rodriguez *et al.* 2004) it is not surprising that miRNAs can arise via the acquisition a miRNA like hairpin in a intron sequence, a term called 'intronic exaptation' (Campo-Paysaa *et al.* 2011). In addition, new miRNAs have also been shown to arise via de novo acquisition, or acquisition of miRNA function from an antisense transcript of an existing miRNA (miRNA*).

Because miRNA families arise independently, they can be treated as a discrete set of characters. In other words their presence versus absence in a taxon can be coded in the same manner as other discrete characters such as morphological characters. Thus miRNA phylogeny reconstruction is essentially performed via binary analysis i.e. the presence (1) vs. absence (0). Groups of taxa containing the greatest number of orthologous miRNAs can be inferred to be more closely related to one another then they are to groups of taxa with a smaller subset of orthologous miRNAs; (e.g. Human+Mouse will contain bilaterian, Deuterostome and mammalian specific miRNA genes, whilst Human+Nematoda will also share bilaterian specific miRNAs but Humans will not posses any protostome or ecdysozoan specific miRNAs present in Nematoda). Novel miRNA acquisitions represent the gain of a *de novo* trans-acting gene classes (Tarver *et al.* 2012). Here, the outgroup state can be determined with a high level of certainty (i.e. absence) coupled with rarity of secondary loss (special care needs to be taken to avoid false negatives), high conservation to the mature sequence and improbability of convergent evolution; with such properties demonstrated experimentally across the entire metazoan tree of life. MicroRNAs are thus excellent candidates for delineating the position of the root in a phylogenetic tree, which is vital to understanding the emergence and evolution of groups of species.

Many of the current competing hypotheses of animal evolution are not a disagreement of topology *per se* but instead a problem of root placement; for example the placement of the root in Arthropoda when changed can result in recovery of both of the main competing hypotheses of the four main arthropod classes (Rota-Stabelli and Telford, 2008); with the application of miRNA data resulting in the unambiguous support of the Mandibulata hypothesis of arthropod evolution (Rota-Stabelli *et al.* 2011). Considering the properties mentioned above, miRNAs are an invaluable new phylogenetic marker for the goal of resolving some of the most intractable phylogenetic problems, across all levels of the animal hierarchy from species to phylum. Moreover, the continual reduction of NGS costs of sequencing small RNA libraries makes miRNAs a cost effective tool. In addition to cost, the ease of analysing miRNA data sets that are vastly smaller compared to large scale multi-gene analyses as seen in phylogenomics further promotes the increased use in the future of phylogenetic studies.

4.1.3 Validating MicroRNAs

Before conducting a phylogenetic analysis using miRNAs, sequence data (e.g. small RNA libraries) or miRNA orthologue sequences indentified from genomic searches (e.g. BLAST) must first be annotated in order to indentify and validate candidate miRNAs. Initially miRNA discovery relied upon conventional Sanger sequencing of size restricted (~22 nt) RNAs, but with the introduction of NGS technologies the task of small RNA sequencing has been greatly simplified. MicroRNA annotation can be achieved using a set of guidelines based on the secondary structure and mode of biogenesis; these guidelines have been summarized (Ambros *et al.* 2003) providing a high level of scrutiny to miRNA discovery. Importantly, miRNAs and other regulatory RNAs (e.g. small interfering RNAs (siRNA)) can not be distinguished based on their functions; this is due to the differential preference of some miRNAs to act upon their target transcripts by repressing their translation while some miRNAs along with siRNAs direct cleavage of their target transcripts (Bantounas *et al.* 2004; Filipowicz *et al.* 2008).

The characteristic features seen in actual miRNAs relate to the endogenous transcripts found in local hairpin structures, which ordinarily are processed such that a single mature miRNA sequence accumulates from only one arm of the hairpin precursor molecule (pre-miRNA). Moreover, if indeed it is a bona fide miRNA then it will also have the characteristic processing sites consistent with *Drosha* and *Dicer* biogenesis (Berezikov, 2011) i.e. staggered cleavage; producing phased small RNA reads with the most abundant RNA reads corresponding to the mature ~22nt miRNA sequence. Given the desire to distinguish between small RNAs like miRNAs and siRNA, miRNAs are identified from other small RNAs by their mode of biogenesis, which is intimately linked to a miRNAs secondary fold back structure as previously stated. The identification and annotation of a given miRNA is based on the following criteria, which can be categorized under two distinct headings; expression and biogenesis (Ambros *et al.* 2003).

According to expression criteria, true identification of a miRNA should include the following criteria: **A**) Detection of a distinct ~22 nt RNA transcript by hybridization to a size fractioned RNA, usually achieved by the northern blotting method, **B**)

Identification of the ~22 nt sequence in a cDNA library made from size fractioned RNA, with these sequences precisely matching the genomic sequence of the organism they were cloned from. In addition to expression criteria, validation of a bona fide miRNA must reside in two or more of the following biogenesis criteria: **C**) Predicting the potential fold back hairpin structure which contains the mature ~22 nt miRNA within one of the hairpin arms, this hairpin must be the folding alternative with the lowest free energy value (~ -20 kcal/mol) and this fold back structure must contain at least 16 base pairs (bp) derived from the ~22nt mature miRNA whilst also not containing any large internal loops or bulges (particularly asymmetric bulges); **D**) Phylogenetic conservation of the mature miRNA sequence and its associated fold back precursor (i.e. pre-miRNA), with the same minimal base pairing requirements as seen in criterion C; but need not meet the lowest free energy folding alternative; and **E**) Detection of increased pre-miRNA accumulation in organism with reduced Dicer function; however, reduced Dicer function criterion alone is not strictly a characteristic of miRNA biogenesis, as Dicer is known to cleave dsRNA to generate siRNAs (Bantounas *et al.* 2004).

Correct annotation of a miRNA relying on a single criterion based on either expression or biogenesis is not sufficient (Ambros *et al.* 2003). Ideally, the identification of a bona fide miRNA would meet all criteria but in practice variations are possible; with the very minimum criterion requirements being the expression of a ~22 nt form and the presence of a hairpin precursor need to be verified to classify a small RNA as a miRNA (Berezikov *et al.* 2006).

4.2 Materials and Methods

In this section I will describe the materials and methods used in order to generate small RNA libraries that we sequenced to identify miRNA genes for our selected taxa. The protocols described apply to the taxa presented in the two publications Rota-Stabelli et al (2011) and Campbell et al (2011). The protocols described in this section will differ in respect to selective species and the next generation sequencing method applied to sequence the small RNA library of that species. Next generation sequencing technologies used to generate the data presented in this thesis are *454* Life Sciences (Bradford, CT, USA) and Illumina (Yale sequencing center).

4.2.1 RNA extraction

I used standard RNA extraction methods that were outlined according to the Invitrogen^{TM} TRIzol® Reagent protocol Catalogue No (15596-018). For all species in which we present miRNA data in this thesis the same RNA extraction protocol was performed.

Depending on the size of the specimen, an initial tissue homogenization step was applied in order to fully breakdown the tissue before applying the TRIzol® reagent protocol. This was performed using Liquid nitrogen (LN_2) and a pestle and mortar, which resulted in snap frozen tissue that was then ground down using a pestle and mortar. The resulting tissue was further homogenized using TRIzol® reagent, using 1ml of TRIzol®per 5-100 mg of tissue. Homogenized tissue was then incubated for 5 minutes at 15 to 30°C to permit complete disassociation of nucleoprotein complexes. Then added 0.2 ml of chloroform per 1 ml of TRIzol® reagent used in the initial homogenization. Tubes were then capped and shaken vigorously followed by an incubation period of 3 minutes at 15 to 30°C. Tissue samples were then centrifuged at 12,000 x g for 15 minutes at 4°C. Following the centrifugation the mixture separated into three distinct phases, the lower red phenol-chloroform phase, an interphase, and an upper aqueous phase; this phase contained the RNA that was then transferred into a fresh Oakridge tube. In order to precipitate the RNA we added in some isopropyl alcohol, in the amount of 0.5 ml of isopropyl alcohol per every 1ml of TRIzol® used initially. Samples were then capped, mixed and incubated for 10 minutes at 15 to 30°C followed by a centrifugation cycle at 12,000 x g for 10 minutes at 4°C. This last step resulted in a final RNA pellet formed at the side and bottom of the Oakridge tube. The TRIzol® protocol next calls for an RNA wash in 75% ethanol, however this step was removed, as it would have resulted in loss of miRNAs from the RNA pellet. The pellet was allowed to dry by pouring off the remaining isopropyl alcohol and subjecting the pellet to a final centrifugation at 7,500 x g for 5 minutes at 4°C. The aqueous phase was then pipeted off, making sure not to disturb the pellet, which was then allowed to air dry for 10-15 minutes at 15 to 30°C. Finally the RNA pellet was resuspended in 200-500 μl of RNase free water; depending on the size of the pellet.

4.2.2 Small RNA library generation

RNA libraries generated for this thesis differed in the next generation sequencing technology used to produce the reads for each library. The initial sequencing method selected to generate our small RNA library reads was the large-scale parallel Pyrosequencing system developed by 454 Life Sciences. In order to generate our reads using the 454 platform required a lengthy protocol with steps in which RNA is size fractioned, 3' and 5' adaptors, complementary DNA (cDNA) synthesis, vector cloning and unique barcode identifiers which are ligated in order to facilitate amplification and identification of species reads from a pooled sample of species sequences. However, some species small RNA libraries were sequenced using the Illumina platform. Sequencing using the Illumina platform did not require the same lengthy protocol; instead it just required a sample of extracted RNA (preformed as detailed in previous section) to be sent for sequencing.

4.2.2.1 454 small RNA library protocol

This protocol is purpose made to generate a miRNA library for sequencing and identification of novel miRNAs; for fully detailed protocol see Appendix 1 of Appendices. Small RNA libraries were constructed as described (Wheeler *et al.* 2001). Small RNAs were isolated with fluorescein-labeled DNA oligonucleotides equivalent to 21 and 27 nucleotides (nt) in molecular weight were combined with 200- 500 mg of total RNA and electrophoresed on a 15% urea-polyacrylamide gel. Following the 3' linker ligation, 31 and 43 nt fluorescein markers were combined with the ligated RNA just before electrophoresis; these were used to guide the excision of the 3' ligated RNAs (between 35 and 41 nt in size). Following the 5' linker ligation a 51 nt fluorescein marker was used in the same manner. The gel was then excised above the marker to include the 5' and 3' ligated RNAs (between 52 and 58 nt in size). Small RNA cDNA was then generated by way of reverse transcription of the 3' and 5' linker-ligated small RNAs. PCR amplification of the small RNA cDNA was performed next, under the following temperature conditions: an initial denaturation at 96°C for 1min; 33 cycles at 96°C (10 sec), 50°C (1min), and 72°C (15 sec); a final extension time of 5min; and then held indefinitely at 10°C. The PCR primers included a unique 4 nt barcode so that the source of the sequence could be identified after sequencing; and the 454 primers. The resultant PCR amplicons were then electrophoresed through a 3% agarose gel. After running out the gel, product bands that approximately migrated the same distance as the 100 nt ladder band were excised, gel extracted (Qiagen QIAquick Gel Extraction Kit; Qiagen, CA, USA). DNA concentrations were measured using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA). Libraries with different barcodes were pooled for a total of 100 ng and submitted to *454* Life Sciences (Branford, CT, USA) and the Yale Center for Genomics and Proteomics Sequencing Facility for sequencing.

4.2.3 Identifying microRNAs using miRMiner

The miRNA analyses presented in this thesis were conducted using software called miRMiner (Wheeler *et al.* 2009) that allows discovery of known and novel miRNAs in newly sequenced taxa, and identifies conserved miRNA complements from all taxa considered. The program miRMiner was designed to be implemented using sequence reads generated by *454* Pyrosequencing; therefore data generated by other sequencing methods (in this thesis Illumina was the alternate sequencing method) needed to be pre-processed in order to be compatible with miRMiner. The pre-processing of Illumina data was achieved by implementing a pipeline (developed specifically and coded in PERL and python) that mimicked the sequence read processing properties of miRMiner.

In order to identify miRNA reads, group them according to primary sequence, and finally identify novel or conserved miRNAs, miRMiner first processes raw *454* sequence reads in the following manner. Firstly the 5' and 3' primers and taxon specific *454* barcode identifiers are removed by applying a 21nt cut off on either end of each sequence read. Resulting reads are then organised by taxon according to the barcode identifiers removed, while also enforcing a 17 – 25 nt cut-off limit. All reads that do not fall within the 17 – 25 nt cut off and/or do not have a matching 3' and 5' barcode are removed from the data. Within each species, duplicate reads are eliminated, and the number of duplicates annotated as the reads frequency count. In each resulting non-redundant set, reads that are identical to reads with a higher frequency count when ignoring differences on the 5' and 3' end and allowing a one gap or mismatch are grouped. Finally, sequences from each sub group with the highest frequency count are selected as representative sequences for further analysis, all remaining reads are not considered.

Our pipeline developed to mimic miRMiner, processes Illumina SOLiD sequence reads according to the same criteria implemented in miRMiner. Sequence data generated by Illumina was output as a set of reads, one set per species. Our pipeline was developed to deal with one species per run. The first stage of the pipeline converts raw paired-end SOLiD FastQ data into the format FastA. FastA is the input format for downstream processing in the pipeline and miRMiner. From the formatted sequence file, the SOLiD sequencing primer is then searched and removed from the 3' end (SOLiD primer used: CTGCTGTACGGCCAAGGCG). Primer removal is performed by searching each read for the full-length primer sequence and all possible sub words (441 words, including all palindromes) for that sequence. Reads are then grouped by removing all duplicate reads, again annotating the number of duplicates as that reads frequency count. If required, reads could then be reduced to contain only sequences that had a length of between 20 – 25 nt; this limit differed from the cut off implemented by miRMiner. In each non-redundant set of reads, reads that are identical in length but differ only in the last 3 nt on the 3' end were grouped together, with the number of reads annotated as the new frequency count. The pipeline culminates by formatting the sequence read fastA headers to be compatible with miRMiner.

Annotation of known miRNAs was achieved by identifying homologous mature and miRNA* (miRNA star) sequences in miRBase (Griffith-Jones *et al.* 2007) release version 15, by way of a stand-alone BLAST search. The resulting list of candidate identities were then filtered according to three criteria on an ungapped global alignment of the read and the hit sequence, beginning at the 5' end: (i) sequence lengths must match within 2 nt; (ii) positions 2-7 of the seed sequence must be identical; (iii) the remainder of the alignment may contain up to only 3 mismatches. Sequence reads matched to a known miRNA or miRNA* sequence within the above criteria were annotated and removed from the data set. Reads identified to not be of miRNA origin were found by comparison with NCBI's nucleotide database (nt) using Standalone MEGABLAST (version 2.2.17). Reads matching a non-miRNA RNA (rRNA, tRNAs etc.) molecule with percent identity >95% were removed, and the remaining sequence reads were then investigated for phylogenetic conservation. Reads from all species were combined, and those that "matched" a read with a higher frequency count were grouped. Matches were determined using the three criteria used to identify known miRNAs given above (similar length, seed sequence identity, and non-seed sequence similarity). Reads conserved across multiple taxa were grouped; and ranked by the frequency count of the most frequently occurring sequence. Reads not conserved across multiple taxa were divided by taxon and ranked by frequency count. This completed the automated analysis by miRMiner, resulting in a list of conserved reads across all taxa and lists of unique reads for each taxon.

4.3 Results

4.3.1 Results of Library generation

4.3.1.1 Testing success of library generation via amplicon gel electrophoresis.

The results presented in this section correspond to the final stages of the 454 RNA library generation protocol (Protocol in Appendix 1: Day 7 – Step 8). It should be noted, results served the purpose of indicating whether the PCR products generated was the correct ~103 bp sized DNA fragment length following initial size fractionation (of the extracted RNA), addition of 5' and 3' linker-ligated primers, generation of RNA cDNA and the final PCR product. Product bands for individual species seen to be just above the 100bp DNA ladder band indicated that the library generation protocol was performed successfully; see Figure 4.4 for a schematic representation of the expected result.

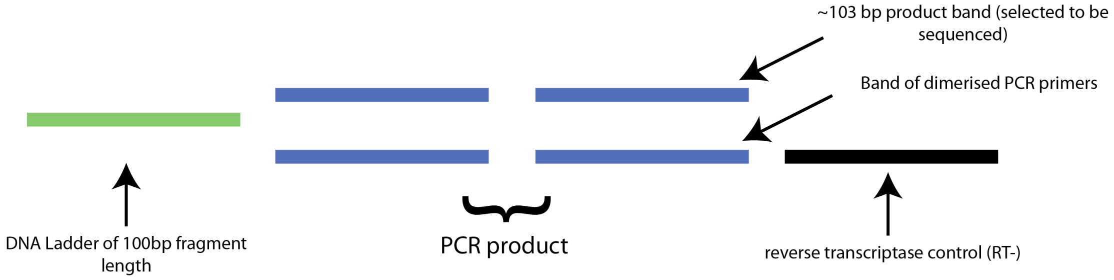

~103 bp product band (selected to be sequenced)
Band of dimerised PCR primers
DNA Ladder of 100bp fragment length
PCR product
reverse transcriptase control (RT-)

Figure 4.4: Schematic view of ideal result of 3% agarose gel electrophoresis. Performed to isolate DNA fragments corresponding to amplified miRNA genes. Bars represent individual lanes of the gel, with colour-coded bars representing different DNA fragments types.

The gels shown (Figures 4.5.1 to 4.5.7) here all yielded sequences of the length required for the sequence to contain potential *bona fide* miRNA sequences. Resulting product bands for these gel runs were all processed according to the protocol (section 4.2.2.1) and selected for sequencing.

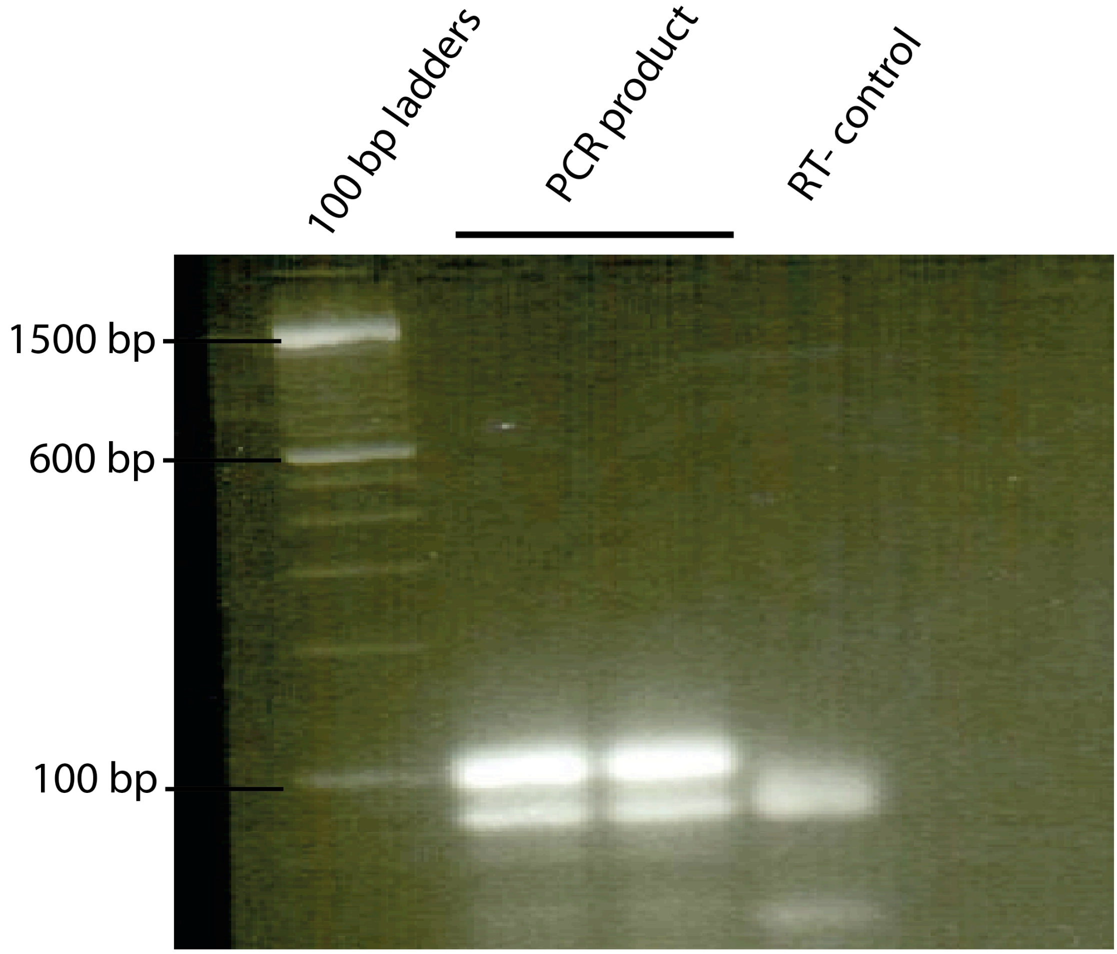


Figure 4.5.1: 3% agarose gel electrophoresis of *Astacoidea sp.* (crayfish). Four lanes shown from left to right, from left most to right lanes corresponding to the DNA ladder, 2 lanes of PCR product/dimmerized primers and lastly a reverse transcriptase control lane.

Gel electrophoresis for the crayfish *Astacoidea sp*. (see Figure 4.5.1) was performed in triplicate. This was due to the first two gel runs yielding no visible ~103 bp product band, indicating that the PCR amplification for that species did not work. The third PCR amplification yielded a clear band just above the 100bp DNA ladder.

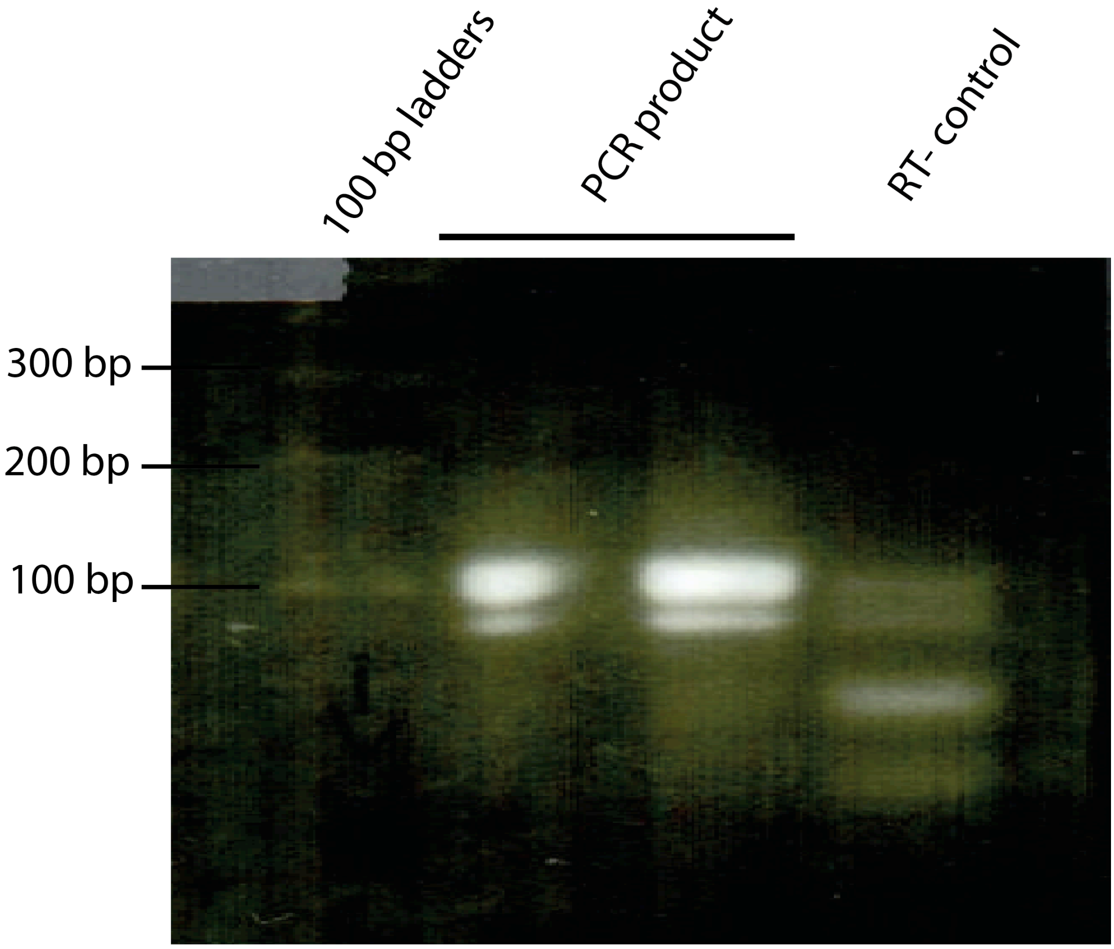


Figure 4.5.2: 3% agarose gel electrophoresis of *Glomeris marginata* (Pill millipede). Four lanes shown from left to right, from left most to right lanes corresponding to the DNA ladder, 2 lanes of PCR product/dimmerized primers and lastly a reverse transcriptase control lane.

Gel electrophoresis for the pill millipede *Glomeris marginata* (see Figure 4.5.2) was ran as stated in the protocol of section 4.2.2.1, however the gel did not run smoothly and resulted in a band separation that was not straight. The bands seen for *G. marginata* were thick, which we believe to be the result of two individual bands migrating in close proximity to one another. Nonetheless each band was seen to have migrated above the 100bp ladder band indicating that those bands contained sequences of the correct DNA fragment length.

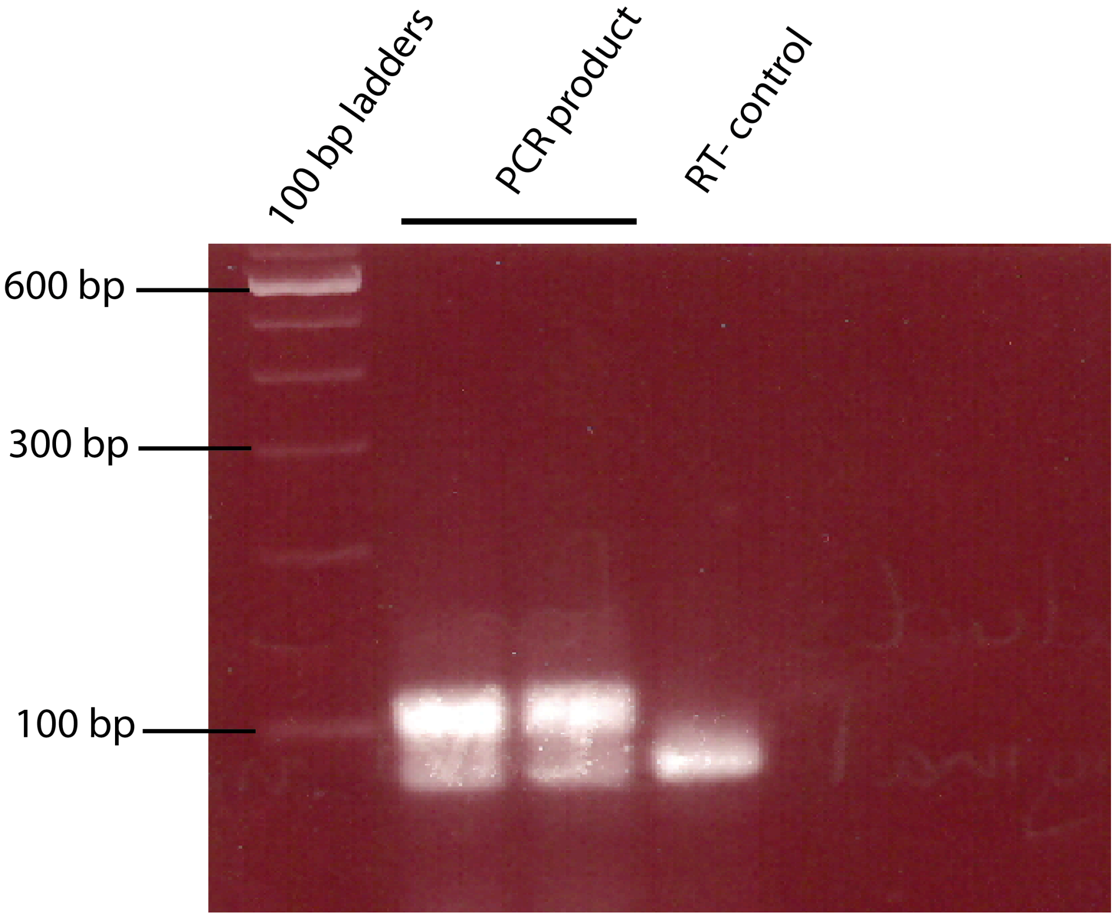


Figure 4.5.3: 3% agarose gel electrophoresis for *Limulus polyphemus* (Horseshoe crab). Four lanes shown from left to right, from left most to right lanes correspond to the DNA ladder, 2 lanes of PCR product/dimmerized primers and lastly a reverse transcriptase control lane.

RNA library generation performed on the chelicerate *Limulus polyphemus* (see Figure 4.5.3) resulted in a clear-cut and satisfactory migration of ~103 bp product bands above the 100bp DNA ladder band.

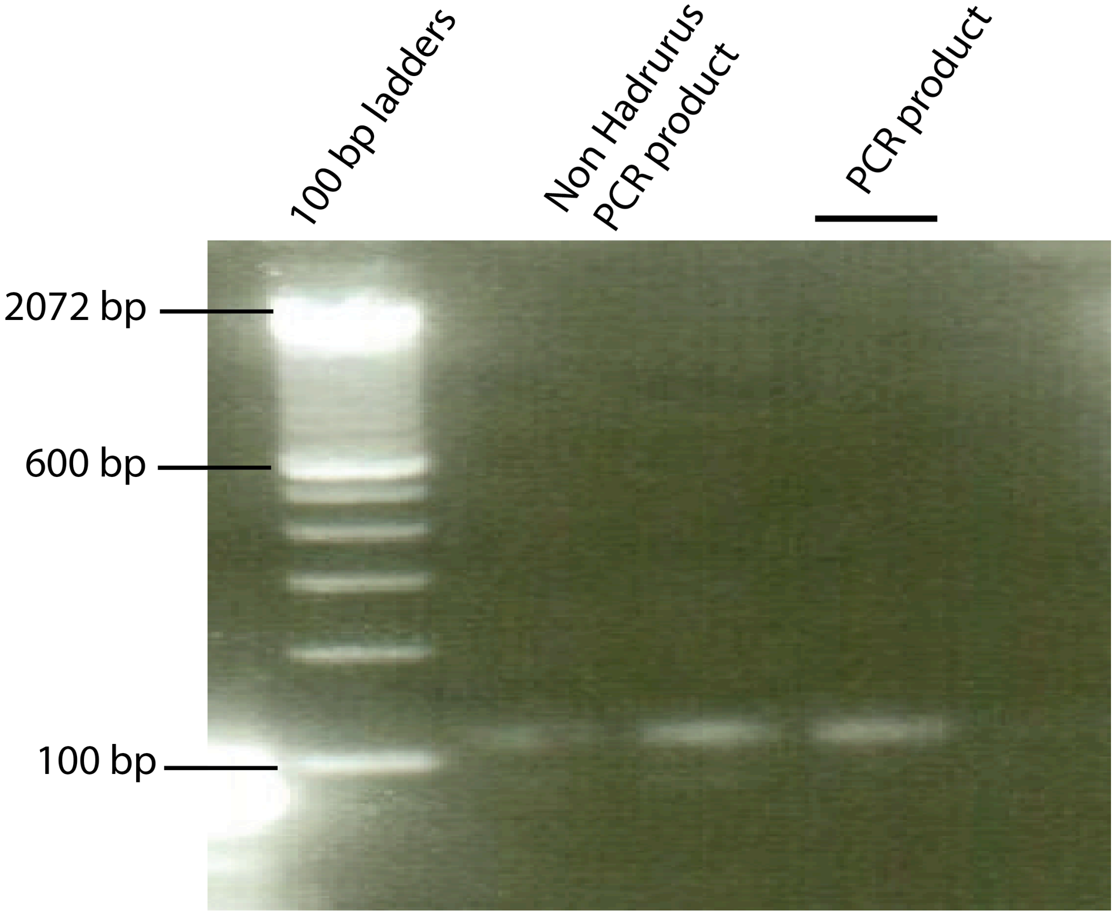


Figure 4.5.4: 3% agarose gel electrophoresis for *Hadrurus sp.* (Scorpion). Four lanes shown from left to right, from left most to right lanes correspond to the DNA ladder, 2 lanes of PCR product for a non Hadrurus species (not considered here) and lastly the PCR product for Hadrurus sp. The reverse transcriptase control lane is not visible.

Gel electrophoresis for the scorpion *Hadrurus sp.* (see Figure 4.5.4) was performed as set out in the protocol (section 4.2.2.1), and resulted in a clearly defined ~103bp product band above the 100bp ladder band.

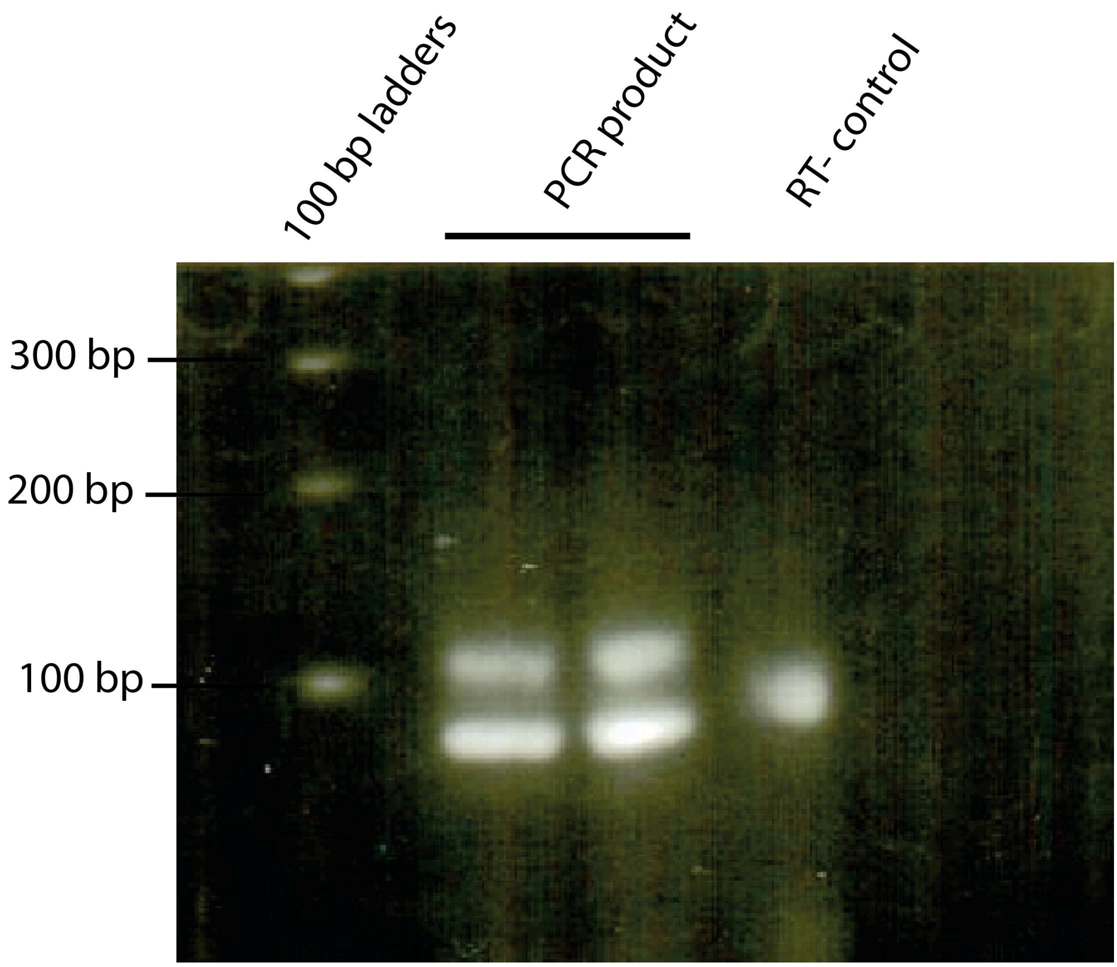


Figure 4.5.5: 3% agarose gel electrophoresis for *Scutigera coleoptrata* (House centipede). Four lanes shown from left to right, from left most to right lanes correspond to the DNA ladder, 2 lanes of PCR product/dimmerized primers and lastly a reverse transcriptase control lane.

Gel electrophoresis for the centipede *Scutigera coleoptrata* (see Figure 4.5.5) resulted in clearly defined bands above the 100bp ladder band. These bands had good migratory separation were seen to have a clear-cut size fraction of around ~103bp.

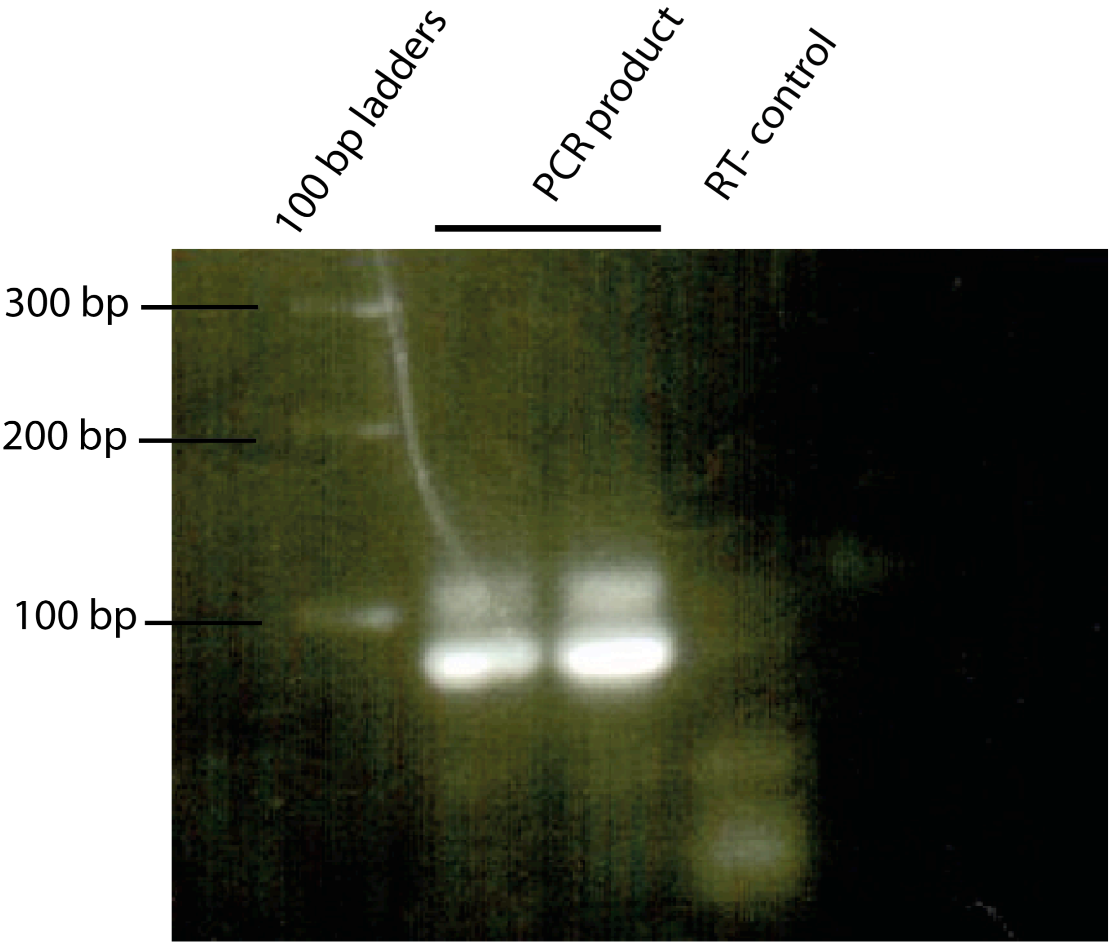


Figure 4.5.6: 3% agarose gel electrophoresis for *Acanthoscurria chacoana* (Tarantula). Four lanes shown from left to right, from the left most to right lane corresponds to: DNA ladder, 2 lanes of PCR product/dimmerized primers and lastly a reverse transcriptase control lane.

Gel electrophoresis for the spider *Acanthoscurria chacoana* (see Figure 4.5.6) resulted in migration of faint product bands with good separation above the 100 bp DNA ladder band.

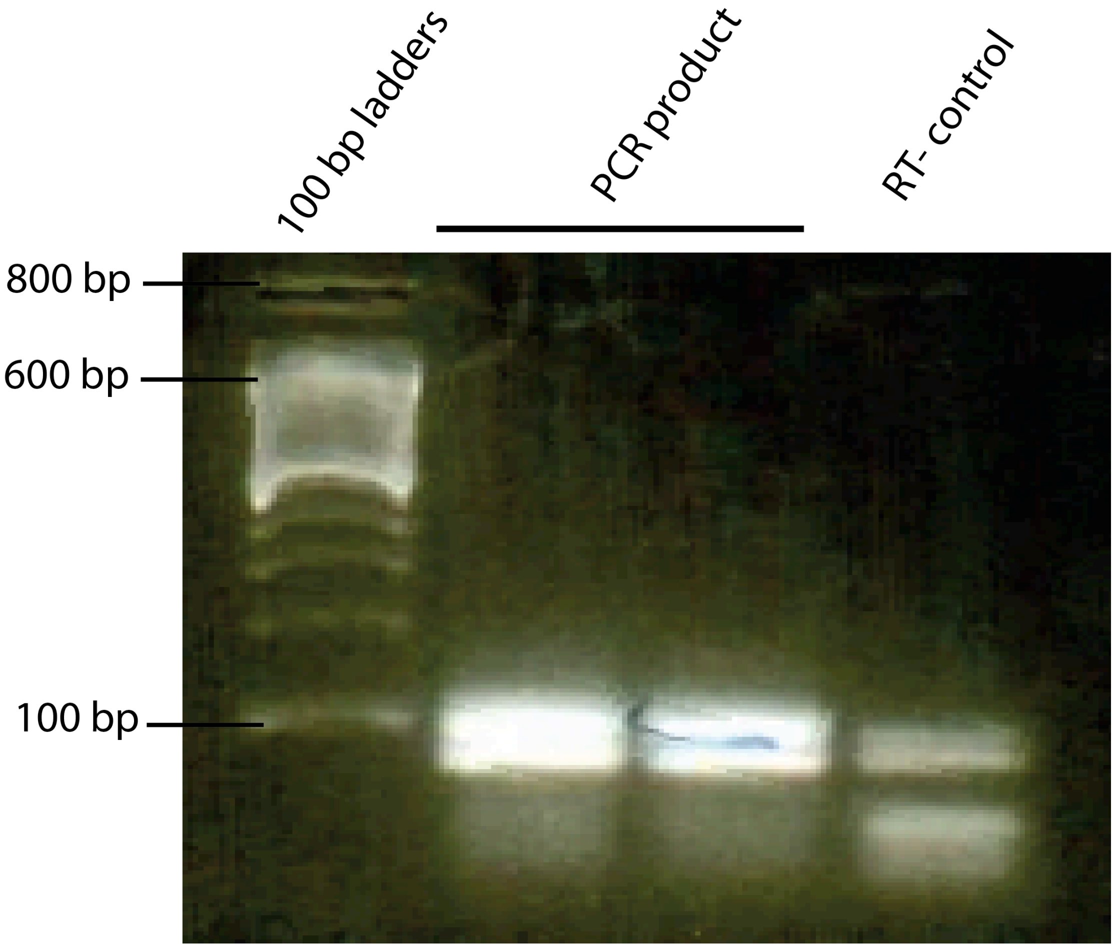


Figure 4.5.7: 3% agarose gel electrophoresis for Thermobia domestica (Firebrat). Four lanes shown from left to right, from left most to right lanes correspond to the DNA ladder, 2 lanes of PCR product/dimmerized primers and lastly a reverse transcriptase control lane.

Gel electrophoresis for the firebrat *Thermobia domestica* (see Figure 4.5.7) was performed as stated in the protocol (section 4.2.2.1). Band migration and separation was not satisfactory due to a short migration period, and so was allowed time to migrate further. A longer migration time resulted in a clearer defined band separation with the product band located above the 100bp DNA ladder.

4.3.2 Arthropoda miRNA distribution supports Mandibulata not Myriochelata.

MicroRNAs are an ideal source of data to tackle the issue of competing phylogenetic hypotheses, previously unresolved by conventional data; therefore providing an additional independent source of data (Sperling and Peterson, 2009). To tackle the problem of the competing Mandibulata and Myriochelata (see section 1.1.5 of Chapter 1) phylogenetic hypotheses, miRNA complements for key arthropod taxa were explored using a combination of genomic searches in addition to small RNA libraries which were sequenced and analysed for their miRNA reads. Consideration of the number of beneficial phylogenetic properties of miRNAs (Tarver *et al.* 2012; but see section 2.2.2 and 4.1.2), properties such as continual addition to genomes through time, high conservation of their primary sequence (~22 nt) allows miRNAs to be readily identifiable between descendant taxa of interest. Ease of identifying conserved and novel miRNAs between taxa coupled with the apparent rarity of secondary loss and low probability of convergent evolution of any miRNA ensures these regulatory elements are an invaluable class of phylogenetic characters.

From the analyses of our miRNA libraries we found a variety of miRNA that can be used to characterise arthropod groups. One miRNA, *miR*-965, had previously been found only in Pancrustacea and had been shown to be absent from the genome of the chelicerate *Ixodes scapularis* (Wheeler *et al.* 2009). Importantly, we found reads of the mature *miR*-965 in the small RNA libraries of both myriapods (*Glomeris marginata* and *Scutigera coleoptata*), and also in the genome of the centipede *S. maritima* (see Figure 4.6). Screening our miRNA libraries also showed that in addition to being absent from the genomic sequence of the tick (*I. scapularis*), *miR*-965 could not be detected in the xiphosuran *Limulus polyphemus* nor in the arachnid *Acanthoscurria chacoana*. This distribution supports miR-965 (see Figure 4.6 and

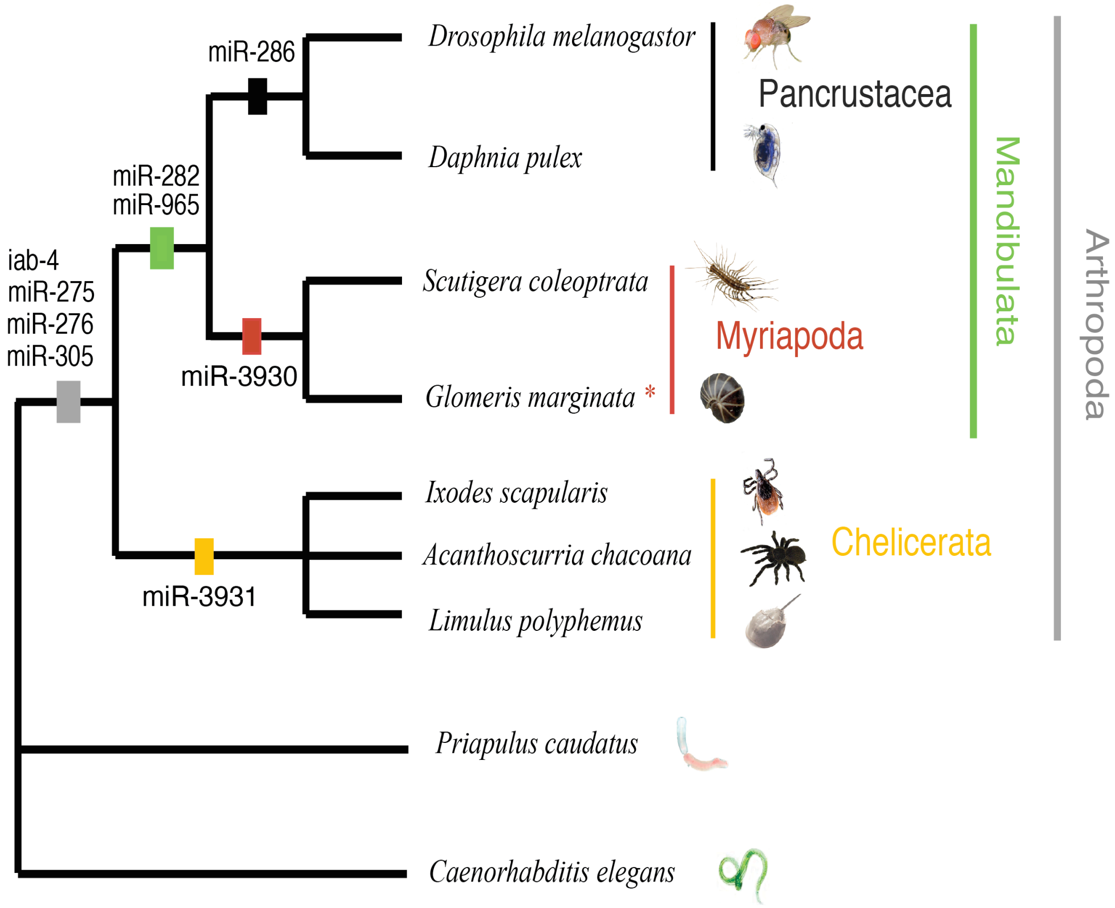


Figure 4.6: Phylogenetic distribution of miRNAs supports Mandibulata. The monophyly of Mandibulata is supported by the presence of miR-965 and miR-282, also discovered in the genome of the centipede *Strigamia maritima*, and in the small RNA libraries of the millipede *Glomeris marginata* and the house centipede *Scutigera coleoptrata*. miR-965 and miR-282 are not known from any chelicerate or non-arthropod. N.B. miR-282 was not found in the small RNA library of *Glomeris**. In addition a novel chelicerate miRNA (miR-3931) is present only in chelicerates, but in none of the mandibulates considered. A novel myriapod miRNA (miR-3930) is found only in myriapods. Major clades highlighted with colour coded nodes (miRNA gains) and bars (delineate clades).

4.7) as a putative genomic apomorphy (a rare genomic change) of the Mandibulata. This same distribution is true of a second miRNA *miR*-282 that we have found only in insects, crustaceans and the centipedes *Strigamia* and *Scutigera*. *miR*-282 was not found in the *Glomeris* small RNA library (Figure 4.6 and Figure 4.7). Absence of *miR*-282 is most likely a result of the low expression of *miR*-282 across all

Mandibulata sampled here, while also the total number of reads and sequencing depth was relatively low in the *Glomeris* miRNA library (* - see Figure 4.6).

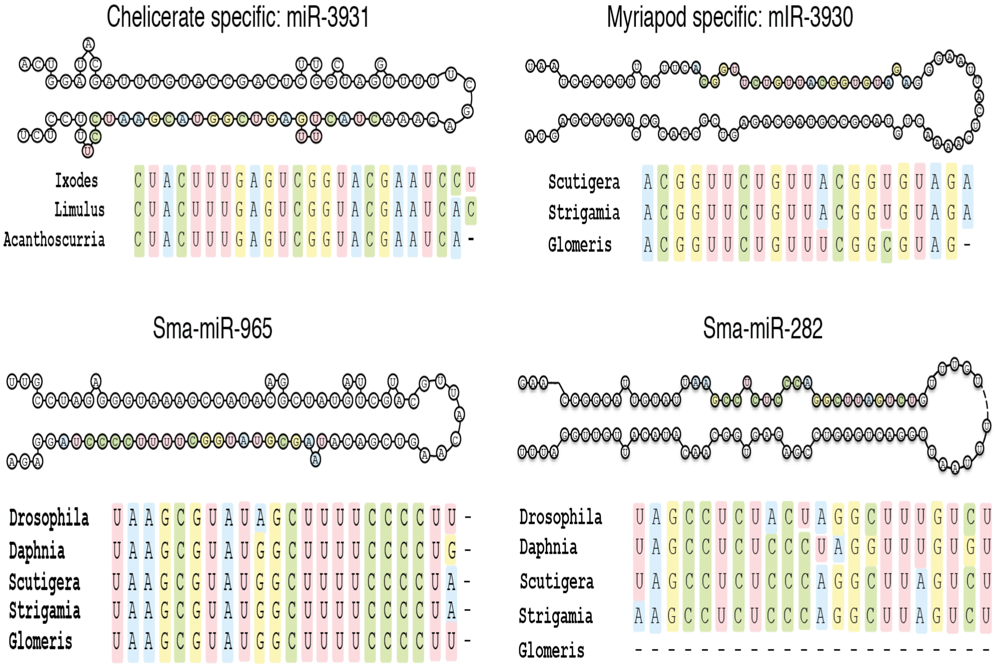

Chelicerate specific: miR-3931
Ixodes CUACUUUGAGUCGGUACGAAUCCU
Limulus CUACUUUGAGUCGGUACGAAUCAC
Acanthoscurria CUACUUUGAGUCGGUACGAAUCA-
Myriapod specific: mIR-3930
Scutigera ACGGUUCUGUUACGGUGUAGA
Strigamia ACGGUUCUGUUACGGUGUAGA
Glomeris ACGGUUCUGUUUCGGCGUAG-
Sma-miR-965
Drosophila UAAGCGUAUAGCUUUUCCCCUU-
Daphnia UAAGCGUAUGGCUUUUCCCCUG-
Scutigera UAAGCGUAUGGCUUUUCCCCUA-
Strigamia UAAGCGUAUGGCUUUUCCCCUA-
Glomeris UAAGCGUAUGGCUUUUCCCCUU-
Sma-miR-282
Drosophila UAGCCUCUACUAGGCUUUGUCU
Daphnia UAGCCUCUCCCUAGGUUUGUGU
Scutigera UAGCCUCUCCCAGGCUUAGUCU
Strigamia AAGCCUCUCCCAGGCUUAGUCU
Glomeris ----------------------

Figure 4.7: Stem-loop structures of Mandibulata, Chelicerata and Myriapoda specific miRNAs. Fold-back RNAs shown are precursor miRNAs (pre-miRNA). The folded miRNAs miR-965 and miR-282 shown are taken from stem-loop hairpins for *Strigamia maritima* (Sma). Shaded regions in each pre-miRNA highlight the mature miRNA.

In addition, upon screening the *L. polyphemus* and *A. chacoana* small-RNA libraries, we identified a novel chelicerate miRNA (*miR*-3931) that is not present in the Mandibulata, but is present in the genome of the tick *I. scapularis* (see Figure 4.6 and Figure 4.7), and we thus suggest this miRNA to be a new genomic apomorphy for the Euchelicerata (Xiphosura and Arachnida). We have also identified a novel myriapod-specific miRNA (*miR*-3930) in the small-RNA libraries of *G. marginata* and *S. coleoptrata*, and in the genome of *S. maritima*, but not in the libraries or genomes of any other non-myriapod taxon analysed. Genome sequences for the myriapod *S.*

maritima were obtained from the sequenced cDNA library of *S. maritima*, provided freely by the Baylor college of Medicine Human Genome Sequencing Center (http://www.hgsc.bcm.tmc.edu/collaborations/insects/dros_modencode/GAsm/centepede/). Further Myriapod-specific molecular synapomorphies have recently been described (Janssen and Budd, 2010). Results presented in this section have been published in the peer-reviewed journal Philosophical Transactions of the Royal Society B (Rota-Stabelli *et al.* 2011).

4.3.3 MicroRNAs suggest velvet worms as the arthropod sister group within a monophyletic Panarthropoda

In a second miRNA phylogenetic analysis, to investigate the complement of miRNAs that appeared in the two closest living panarthropod (Nielsen, 2001) sister phyla to Arthropoda, small RNA libraries were sequenced for the tardigrade *Paramacrobiotus cf. richtersi* and the onychophoran *Peripatoides novaezelandiae*. MicroRNA complements were obtained from sequenced small RNA libraries for the tardigrade and onychophoran and analysed in conjunction with previously identified myriapod and chelicerate specific miRNA data, described in the previous section (Rota-Stabelli *et al.* 2011). MicroRNA complements for the ecdysozoan phyla Nematoda, Priapulida and the arthropod species *Drosophila melanogaster* and *Daphnia pulex* were obtained from an on line miRNA database miRBase. According to the analysis of Rota-Stabelli *et al.* (2011), the four arthropod specific miRNAs described (*iab*-4, *miR*-275, *miR*-276, *miR*-305) have previously never been identified in any other non-arthropod ecdysozoans. Here, the aim was to investigate whether or not the panarthropod phyla Tardigrada and Onychophora shared any miRNAs that were previously only identified within Arthropoda. Furthermore, previous molecular sequence analyses (Philippe *et al.* 2005b; Sørensen *et al.* 2008; Roeding *et al.* 2007, Lartillot and Philippe, 2008; Roeding *et al.* 2009; Andrew, 2011) that positioned Tardigrada to lie outside Panarthropoda and as sister group to Nematoda were scrutinized by investigating whether or not miRNA complements could be identified to be unique to just Tardigrada and Nematoda.

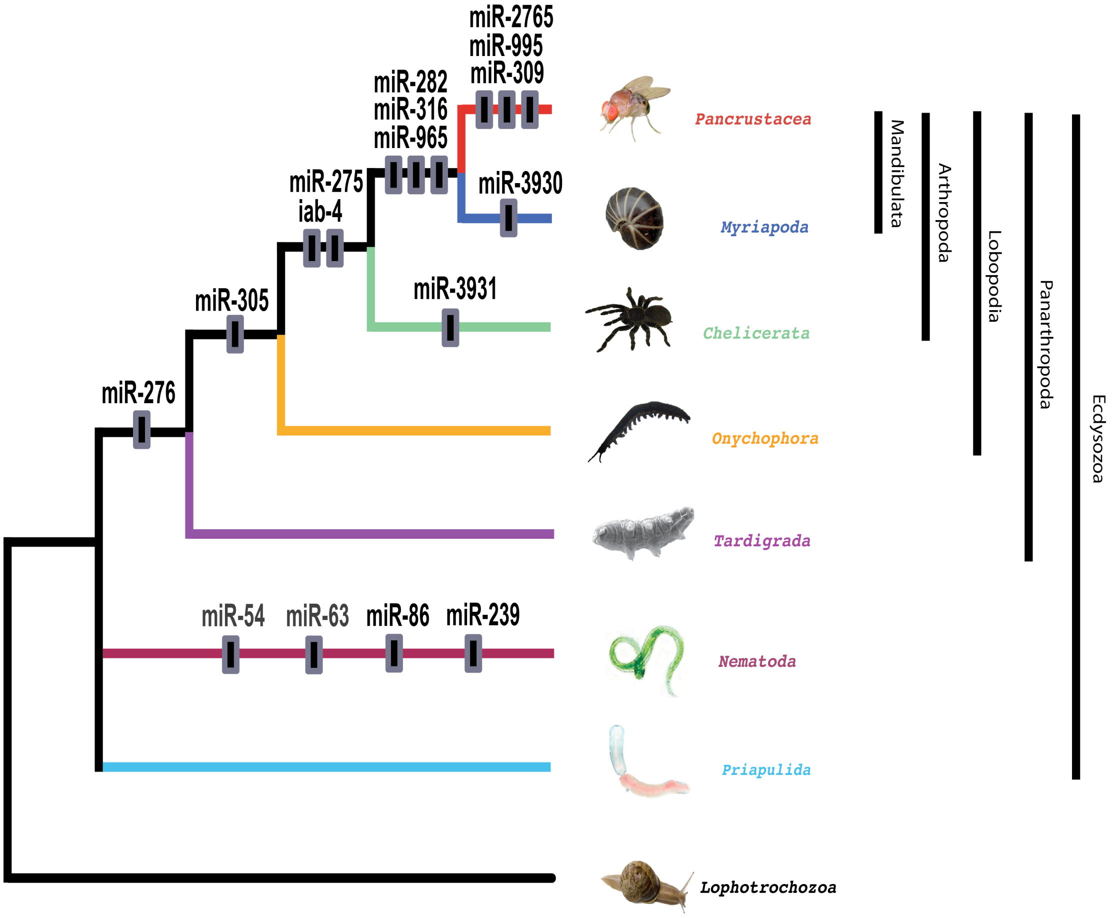


Figure 4.8: MicroRNA distribution supports a sister group relationship between velvet worms and Arthropoda within a monophyletic Panarthropoda. Single grey/black bars represent miRNA gains. Clades are colour coded for clarity, higher level clades depicted with black vertical bars.

There are four miRNAs that are conserved between the nematode genera *Caenorhabditis* and *Pristionchus* (de Wit *et al.* 2009): *miR*-54, -63, -86, and -239

(Figure 4.8). From the analysis of our tardigrade small RNA library, we could not detect any nematode specific miRNAs shared also in Tardigrada. Similarly, no potential miRNAs were found shared exclusively between the tardigrade and onychophoran. Instead, in both the tardigrade and onychophoran libraries we found a single miRNA, *miR*-276, that formerly had been identified only in arthropods (Rota-Stabelli *et al.* 2011). Furthermore, in the onychophoran library, but not in the tardigrade library, we found a second miRNA, *miR*-305, which is also considered arthropod specific (Figure 4.8).

According to the results of the miRNA distribution found within Panarthropoda a number of hypotheses can be made. The miRNA *miR*-276 was found to be present within only all three panarthropod phyla; therefore I infer that the gain of *miR*-276 represents a single genomic apomorphy supporting the monophyly of Panarthropoda (Tardigrada + Lobopodia). Further to this, the miRNA *miR*-305, found to be present only within Onychophora and Arthropoda, suggests that this miRNA gain represents a genomic apomorphy supporting Lobopodia (Onychophora + Arthropoda). Lastly, building upon the previous analysis of arthropod specific miRNAs (Rota-Stabelli *et al.* 2011) I hereto hypothesize that the miRNA gains of *iab*-4 and *miR*-275 are apomorphies of Arthropoda. Results presented in this section have been published in the peer-reviewed journal Proceedings of the National Academy of Sciences (Campbell *et al.* 2011).

4.4 Discussion

4.4.1 Phylogenetic uncertainty and the need for microRNAs

Looking back, the early days of phylogeny reconstruction were concerned with analyses of small data sets of morphological or molecular characters, using simplistic methods of phylogenetic inference. It is clear now that modern phylogenetics has come a long way; to the stage were it is commonplace for analyses to be performed on expansive data sets (Regier *et al.* 2008; Hejnol *et al.* 2009; Holton and Pisani, 2010; Rota-Stabelli *et al.* 2011; Smith *et al.* 2011) under increasingly sophisticated models of evolution (Tuffley and Steel, 1998; Lartillot and Philippe, 2004; Foster, 2004; Blanquart and Lartillot, 2008) (not to mention the vastly improved computational resources). This being said, in spite of the vast amounts of available data and improved methods of inference, numerous open questions remain in modern day systematics. One of the most obvious problems existent in the study of animal evolution regards peculiar and obscure phyla, for example Placozoa, Rotifera, Acanthocephala, and Chaetognatha; yet to be reliably resolved within the metazoan tree of life (Telford, 2006). Much of the complication can be attributed to homoplasy; for instance obscure or strange morphology (as seen in *Trichoplax*) complicated by sparse numbers of useful morphological characters; to cases in which molecular data is exceedingly rapidly evolving making some groups phylogenetically unstable and predisposed to reconstruction artifacts such as LBA (Friedrich and Tautz, 1995; Hwang *et al.* 2001; Pisani *et al.* 2004; Lartillot and Philippe, 2008; Andrew, 2011). This would certainly be the case in Tardigrada, as it has been shown that this group suffers particularly from lack of phylogenetic resolution brought about by rapid molecular evolution (see molecular phylogenetic studies of Campbell *et al.* 2011 vs.

Meusemann *et al.* 2010) and difficulty in interpreting its morphology (i.e. having mixtures of panarthropod and cycloneuralian features; see Telford *et al.* 2008; Edgecombe, 2009; Campbell *et al.* 2011).

Inherent biases in obscure or fast evolving taxa are not the sole reason why so many open questions remain in systematics. One of the biggest hurdles to improving phylogenetic resolution resides in the methods used to investigate those relationships, methods based on models of evolution that currently are not able to (fully) account for the real underlying evolutionary processes encountered in everyday phylogenetic data sets (Lartillot and Philippe, 2004; Foster, 2004; Kelchner, 2007). This notion is made even more apparent by the use of large scale data sets including hundreds of genes; as these data sets can introduce considerable non-phylogenetic signal due to model violations brought about by systematic error (Deulsc *et al.* 2005; Philippe *et al.* 2005a; Nesnidal *et al.* 2010). Furthermore, inclusion of greater amounts of data does not diminish the overall magnitude of the problem of model violation, as it is now well known that although increasing the amount of data in a phylogenetic analysis can overcome problems of sampling error (i.e. stochastic error), the degree to which model violations occur (systematic bias) has been shown to increase according to the amount of data added (Delsuc *et al.* 2005; Kelchner *et al.* 2011).

Many of the metazoan relationships have been proposed on grounds of morphology, with many of these groups now corroborated by molecular data; while others are refuted, such as Coelomata (Hyman, 1951), to Arthropoda and the Articulata hypothesis (Aguinaldo *et al.* 1997; Ruiz-Trillo *et al.* 2002; Philippe *et al.* 2005b; Sempere *et al.* 2007; Dunn *et al.* 2008; Holton and Pisani 2010). However, as already stated many areas of the metazoan tree lack substantial phylogenetic resolution, with persistent competing hypotheses of evolution recovered with high statistical support.

Lack of phylogenetic resolution for some of the major nodes on the tree of life likely stems from a handful of causes, for instance problems of differential rates of molecular evolution or long internodes caused by a recent origin of the crown group, and fast, deep radiations (Philippe and Laurent, 2005; Pisani *et al.* 2011). The build up of phylogenetic evidence over the years has made it apparent that the aforementioned problems typify particular metazoan lineages more than others; prime examples being the dual phylogenetic affinity for groups like Myriapoda (Mandibulata vs. Myriochelata) and Tardigrada (Panarthropoda vs. Cycloneuralia). Ultimately, the problems related to resolving evolutionary relationships like those seen in Arthropods and Ecdysozoa, resides in homoplasy – (similarity in different species brought about by convergent evolution) and how readily we can identify and deal with phylogenetic problems hindered by high levels of homoplasy.

It has been said before that the best way in which to deal with the problems faced in modern day phylogenetics is to use a data source that minimizes homoplasy (Sperling and Peterson, 2009; Campbell *et al.* 2011). One such data set, with low levels of homoplasy is that of the recently discovered class of regulatory elements i.e. miRNAs (Lee *et al.* 1993). However, being homoplasy low is not the only prerequisite for a particular data type to resolve the intractable nodes in the metazoan tree. In addition, for miRNAs to be truly useful, they need to arise quickly enough as to characterize the divergences in question. Providentially miRNAs meet the above criteria, as they not only have beneficial properties that make them excellent phylogenetic markers (see section 4.1.2; Sperling and Peterson, 2009; Campbell *et al.* 2011, Tarver *et al.* 2012), but they have been shown to arise rapidly enough to characterize most of the major metazoan lineages (Hertel *et al.* 2006; Sempere et al 2006; Heimberg *et al.*

2009; Wheeler *et al.* 2009; Sperling *et al.* 2010; Campo-Paysaa *et al.* 2011; but see Pisani *et al.* 2011).

Since their relatively recent discovery, miRNAs have been scrutinized for the role in regulating the expression of genes (Bartel, 2004) but only over the last decade or so have they received notable attention for their phylogenetic utility. In the time since, miRNA complements have been investigated for some of the more problematic nodes of the animal tree of life, for example Annelida (Sperling *et al.* 2009b), Brachiopods (Sperling *et al.* 2011), Deuterostomia (Campo-Paysaa, 2011), Vertebrata (Heimberg *et al.* 2010), Porifera (Sperling *et al.* 2010) and the obscure Acoelomorpha (Philippe *et al.* 2011a). The success of miRNAs in phylogeny reconstruction is already clear, with some of the seemingly obvious phylogenetic relationships not finding support, leading to critical reappraisals of important evolutionary groups (e.g. Acoelomorpha and *Xenoturbella* now considered deuterostomes and not early bilaterians; Philippe *et al.* 2011a) to other cases in which support is recovered to corroborate longstanding classically studied phylogenetic relationships. Accordingly, in this thesis, miRNAs have been further demonstrated as a truly useful phylogenetic data type in which to investigate groups of related species, as for the first time, complements of miRNAs have been described throughout the four major groups of Arthropods (Rota-Stabelli *et al.* 2011) and also the arthropod sister phyla: Onychophora and Tardigrada (Campbell *et al.* 2011).

Complications introduced by homoplasy, from instances of high level sequence saturation and LBA, to cases in which morphology can not provide adequate polarizing characters in which to resolve the position of a group, should, I think encourage the phylogenetic community to seek phylogenetic precession elsewhere. Indeed, this has already begun, as there has been a large increase in the number of studies published presenting miRNA complements for metazoan groups. In essence, miRNAs provide a way in which to contribute to modern phylogenetic study, by introducing additional sources of phylogenetic data in which to test alternate competing hypotheses of evolutionary relationships. Although miRNAs do not provide the ultimate "fix all" solution to phylogeny reconstruction, as they do not represent a data set completely free of homoplasy (i.e. cases of secondary loss due to simplification; see Philippe *et al.* 2011a) they do represent an excellent platform in which to reappraise already established hypotheses based on traditional phylogenetic data types.

4.4.2 Evaluating the strength of miRNA evidence: A case in Arthropoda

In this chapter, I have already described the different strengths (i.e. continual addition, absence of convergent evolution, minimal loss and high conservation) and weakness (difficulty in demonstrating true absence) associated with the use of miRNAs for phylogeny reconstruction. However, I would like to draw attention to how a particular hypothesis, one corroborated by miRNA distribution, can be evaluated on the basis of additional evidence from new sources of genomic data.

The first complete genome of a chelicerate species; the spider mite *Tetranychus urticae* was recently published (Grbic *et al.* 2011). This genome provided additional genomic data with which to test the robustness of miRNA results in support of the mandibulate affinity of the myriapod arthropods and the monophyletic status of Chelicerata. According to the distribution of arthropod miRNAs, the division of the four main sub phyla is characterized by miRNA complements shared exclusively between Myriapoda (autapomorphy of *miR*-3930) and Pancrustacea: together referred to as Mandibulata (*miR*-282, *miR*-965); and Chelicerata (autapomorphy of *miR*-3931). No doubt the best way to ensure one obtains the clearest picture of miRNA distribution to polarize a group of taxa is by deep sequencing of small RNA libraries for all taxa concerned; however in the absence of these libraries the next best option is to mine complete genomes of related species.

The recently published genome for *T. urticae* is the first fully complete and annotated genome for Chelicerata. This provides an excellent resource to facilitate the testing of the distribution of arthropod clade specific miRNAs, thereby allowing me to more precisely infer their validity. *T. urticae* is a plant pest and represents a particularly rapidly adaptive species, with one of the highest incidences of resistance to pesticides (Grbic *et al.* 2011). In addition to this species being highly adaptive to pesticides, *T. urticae* is particularly interesting as it has so far been shown to have the smallest known arthropod genome, estimated at 90Mb (Grbic *et al.* 2011). This is in stark contrast to its closest sequenced relative the acariform tick species *Ixodes scapularis* (with an uncompleted genome estimated at 2,100Mb). Presence of the proposed chelicerate specific miRNA *miR*-3931 in the reduced genome of *T. urticae* would provide additional strong evidence in favour of *miR*-3931 being a true genomic apomorphy of Chelicerata, but not a certainty, as complete genomes and small RNA libraries are still absent for many of the chelicerate lineages.

In order to test the validity of the arthropod relationships supported by the distribution of clade specific miRNAs, I blasted the mature sequences (most conserved region of a miRNA) of the panarthropod, arthropod, mandibulate, myriapod and chelicerate specific miRNAs against the complete genome of *T. urticae.* The only non-chelicerate specific miRNA for which we got significant blast hits were the two non-arthropod specific miRNAs, i.e. *miR*-276 and *miR*-305 that are present in Panarthropoda and Lobopodia respectively. Absence of the two arthropod specific miRNAs (*miR*-257, *iab*-4) is surprising, as these miRNAs have been recovered in all arthropod species analyzed in this thesis; while also being present in the closely related chelicerate species *I. scapularis.* None of the Mandibulate miRNAs were found in T. urticae, including the myriapod (*miR*-3930) and pancrustacean (*mir*-286) specific. The lack of Mandibulate specific miRNAs provides additional support in favour of the Mandibulata hypothesis, with miR-282 and miR-956 being true genomic apomorphies of mandibulates.

Differently, and as expected, the chelicerate specific miRNA (*miR*-3931) from *I. scapularis* was found to be present in the genome of *T. urticae*. The T. urticae homolog of *miR*-3931 had a near complete identity (mismatch of 1 nucleotide at the 3' end). This sequence was then extracted with 100 bp flanking regions, and was subjected to RNA folding using the online folding software mFold (Zuker, 2003). It was then confirmed that the *T. urticae* sequence found to hit *miR*-3931, produced a canonical miRNA hairpin structure, with a minimal free folding energy value of -19.10 Kcal/mol; see Figure 4.9. The fact that the blast hit sequence for *T. urticae* folds into a *bona fide* miRNA structure is convincing evidence to support the presence of miR-3931 in the genome of *T. urticae*. Presence of this miRNA allows me to infer with greater certainty that miR-3931 is a valid genomic apomorphy for Chelicerata.

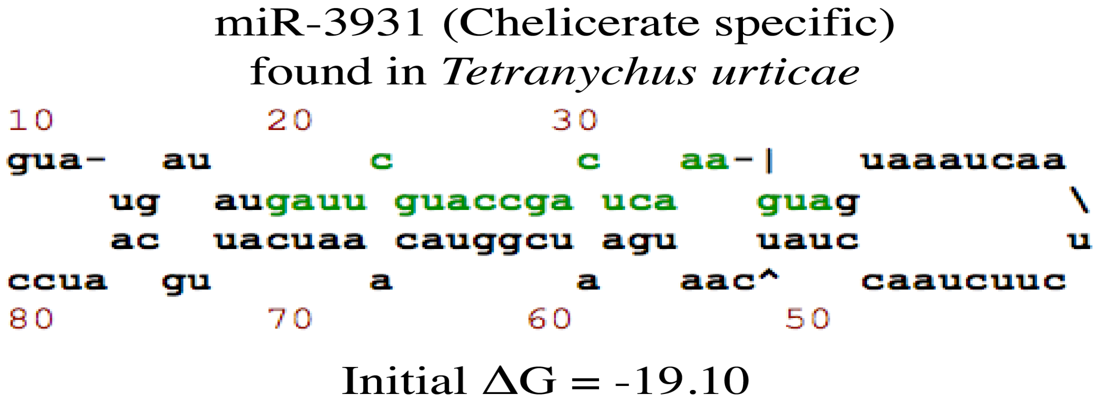

miR-3931 (Chelicerate specific)
found in *Tetranychus urticae*
10 20 30
gua- au c c aa-| uaaaucaa
ug augauu guaccga uca guag \
ac uacuaa cauggcu agu uauc u
ccua gu a a aac^ caaucuuc
80 70 60 50
Initial ΔG = -19.10

Figure 4.9: pre-miRNA structure for chelicerate specific miR-3931 found in *T. urticae*. Free folding energy for *T. urticae* miR-3931 was found to be below the threshold of -20 kcal/mol (Ambros *et al.* 2003).

The investigation of conserved, arthropod clade specific miRNAs against the first fully sequenced and annotated chelicerate genome*,* bolsters results of miRNA distributions in support of Mandibulata and the monophyly of Chelicerata (Rota-Stabelli *et al.* 2011). The absence of arthropod specific miRNAs (*miR*-275 and *iab*-4) is an unexpected result, but goes to further demonstrate that secondary loss of miRNAs can and does occur; particularly in species that are seen as rapidly evolving or that have significant genome size reductions (Sperling and Peterson, 2009; Philippe *et al.* 2011a). Conversely, the presence of the proposed chelicerate specific miRNA (miR-3931) is cogent evidence to support the true monophyletic status of Chelicerata, as *miR*-3931 is retained in a genome that had significant amounts of genome reduction in the course of its evolution.

4.5 Conclusion

Considering the evidence presented in this Chapter, a number of statements can be made about the evolution of one of the oldest and most diverse group of animals to ever exist. In conclusion, the results of investigations into the distribution of highly conserved, and tightly controlled genome regulatory elements or miRNAs, supports many of the classically defined phylogenetic hypotheses for arthropods and their two closest relatives, the onychophorans and tardigrades.

From the analyses of the miRNAs within the arthropods themselves, I have found evidence to support Mandibulata, a classical hypotheses grouping Pancrustacea and Myriapoda; with all groups possessing biting mouthparts or mandibles (Nielsen, 2001). In light of miRNA corroboration, previous analyses of traditional molecular sequence data in support of Chelicerata + Myriapoda (Friedrich and Tautz, 1995; Hwang *et al.* 2001; Cook *et al.* 2001; Pisani *et al.* 2004; Mallatt and Giribet 2006), must on the face of mounting evidence be due to phylogenetic reconstruction artifacts (Rota-Stabelli *et al.* 2011), made even more likely by the sparsity of morphological evidence for Myriochelata.

Lastly, the investigation of miRNAs in the closest living relatives to Arthropoda, supports the monophyletic status of Panarthropoda to include Tardigrada as the earliest branching phylum sister group to a clade composed of Onychophora plus Arthropoda (Lobopodia). Reflecting on previous molecular support for the inclusion of Tardigrada within the Cycloneuralia sister to nematodes (Sørensen *et al.* 2008; Roeding *et al.* 2009; Meusemann *et al.* 2010; Andrew, 2011) I must here infer that the most likely explanation for this grouping is again down to a case of artifactual LBA, between the fast evolving tardigrades and nematodes (Campbell *et al.* 2011).

Contrary to substantial evidence in support of alternate placements of myriapods and tardigrades within Arthropoda and Ecdysozoa, considering the strength of miRNA evidence due to their unique and beneficial properties for phylogeny reconstruction; I must conclude with confidence: that Arthropoda is composed of Mandibulata sister to group to chelicerates, with Panarthropoda composed of Tardigrada sister group to Lobopodia. The work presented here, I feel, will provide substantial phylogenetic resolution to questions of evolutionary relations that currently are still hotly debated despite the long history of phylogenetic investigations into these fascinating animals.

Chapter 5

“Classical” molecular data & the within-Ecdysozoa phylogeny

5.1 Overview

The study of the evolutionary relationships between the major metazoan groups, from the level of Phyla down to the level of genus and species, has traditionally relied heavily upon large matrices of morphological characters (Pisani *et al.* 2007) analysed using simplistic phylogenetic reconstruction methods such as Maximum Parsimony (MP). Previous studies focusing on these characters and methods often yielded highly unresolved phylogenies or incompatible sets of relationships, possibly due to widespread problems with character coding and homology assessment (e.g. Scotland *et al.* 2003). From this point of view molecular data is generally viewed as being less ambiguous than morphological data, even though homology assessment is not straightforward also with reference to molecular data. However, as pointed out by Scotland *et al.* (2003) certainly molecular data has the advantage of providing a greater number of observable characters. Furthermore, molecular data can be subjected to better phylogenetic analyses as model development for morphological data in a likelihood or Bayesian framework has lagged behind, with the Lewis model (Lewis, 2001) which is equivalent to a Jukes and Cantor model for nucleotidic data (Jukes and Cantor 1969) being still the only available option.

Phylogenetic analyses utilizing molecular sequence data started to emerge during the eighties. Following Carl Woese seminal work on the tree of life (Woese *et al.* 1990) the most commonly used molecule in early phylogenetic studies became the Small Subunit rRNA (SSU rRNA – that in Metazoa is the 18S rRNA). Indeed, the earliest molecular phylogeny of the Metazoa was also based on the study of an 18S SSU rRNA data set (Field *et al.* 1988).

Since then also the 28S rRNA (Large Subunit – LSU) has been widely used often in combination with the 18S rRNA. The legacy of these studies is that the SSU rRNA is the taxonomically better sampled gene in the NCBI database. Indeed, one can say that the use of ribosomal sequences, primarily 18S (SSU) and 28S (LSU) rRNA, typify the "classical" period in metazoan molecular phylogenetics. SSU and LSU rRNA have indeed many interesting features for the study of animal phylogenetics, as they can be applied over large evolutionary distances (Field *et al.* 1988; Philippe and Germot, 2000; Peterson and Eernisse, 2001). In addition, because they have a stem-loop based three-dimensional structure they contain regions that evolve at very different rates. Thus, careful site selection from the same rRNA alignment allows investigation of problems at different phylogenetic depth (Mallatt and Giribet, 2006).

Despite the SSU being the best sampled RNA molecule in NCBI today, many early classical molecular phylogenetic studies suffered from sparse taxonomic sampling (Giribet *et al.* 1996; Garey *et al.* 1996; Moon and Kim, 1996; Garey *et al.* 1999). In addition, because of a lack of adequate methods and models, they often failed to account for unequal rates of nucleotide substitution in the different taxa (Ballard *et al.* 1992; Winnepenninckx *et al.* 1995; Aguinaldo *et al.* 1997; Felsenstein, 2004) as well as compositional biases leading to phylogenetic artifacts and low statistical support for important nodes (Hillis *et al.* 1993).

More recently, studies of animal evolution that use rRNA had a much wider taxon sampling (Spears and Abele, 1997; Zrzavý *et al.* 1998; Giribet and Ribera, 2000; Peterson and Eernisse, 2001; Mallatt and Giribet, 2006) while also taking measures to counteract the detrimental effects of including species that have high substitutional saturation, unequal rates of substitution and compositionally biased sequences.

An obvious success of rRNA data in animal phylogenetics was the recovery (by Aguinaldo *et al.* 1997) of Ecdysozoa. This study refuted both the Coelomata (Hyman, 1951) and the Articulata (Anderson, 1973) hypothesis, which were at that time considered fairly well supported clades (but see Eernisse *et al.* 1992 for a different opinion). Ecdysozoa has since received much support from subsequent analyses of other rRNA data sets (Giribet and Ribera, 1998; Giribet *et al.* 2000; Peterson *et al.* 2001; Mallatt *et al.* 2004; Mallatt and Giribet, 2006; Telford *et al.* 2008), large scale phylogenomic analyses (Philippe *et al.* 2005b; Hejnol *et al.* 2009; Holton and Pisani, 2010; Rota-Stabelli *et al.* 2011; Campbell *et al.* 2011) and other molecular data sources, for e.g. mitochondria (Bourlat *et al.* 1999; Rota-Stabelli *et al.* 2010).

Morphological analyses in support of ecdysozoan relationships (Eernisse *et al.* 1992; Schmidt-Rhaesa, 1998) generally recognise the subdivision of Ecdysozoa into Panarthropoda (Arthropoda, Onychophora, Tardigrada; sensu Nielsen, 2001) and Cycloneuralia (Nematoida, Priapulida, Kinorhyncha, Loricifera; sensu Ahlrichs, 1995); However, despite the increase of molecular phylogenetic studies of Ecdysozoa, competing hypotheses for the relationships among the ecdysozoan constituent phyla remain to be resolved; both from a molecular and morphological point of view (Telford *et al.* 2008; Campbell *et al.* 2011).

Within the Ecdysozoa, for e.g. the group Nematoida (Nematoda + Nematomorpha) which has strong morphological support (Schmidt-Rhaesa, 1996) has been recovered as a monophyletic group using rRNA (Zrzavý *et al.* 1998; Garey *et al.* 2001; Giribet *et al.* 2000; Mallatt *et al.* 2004; and Mallatt and Giribet*,* 2006); whilst the studies of Giribet *et al.* (2000) and Peterson and Eernisse (2001) which utilized the same data type did not recover any support for monophyly of Nematoida. Similarly, rRNAs have yet to resolve the interrelationships of the panarthropod phyla. Although it is now generally accepted that Onychophora are the most likely sister group to Arthropoda (Edgecombe, 2010), early analyses supported the inclusion of Onychophora within the Arthropoda (Ballard *et al.* 1992). Further to this, subsequent analyses recovered multiple competing hypotheses for the placement of Tardigrada within Ecdysozoa. The earliest analyses of 18S rRNA supported a sister group relationship between water bears and arthropods (Garey *et al.* 1996; Giribet *et al.* 1996; Garey *et al.* 1999) while later analyses suggested either a sister group relationship with Onychophora (Garey *et al.* 2001; Mallatt *et al.* 2004; Mallatt and Giribet, 2006), Lobopodia (Giribet *et al.* 2000; Garey *et al.* 2001) or alternatively a placement of Tardigrada within the Cycloneuralia as the sister group to Nematoda (Giribet and Ribera, 1998; Giribet and Wheeler, 1999; Park *et al.* 2006; Sørensen *et al.* 2008).

The remaining ecdysozoan phyla, which make up the Scalidophora (Priapulida, Kinorhyncha, and Loricifera; sensu Schmidt-Rhaesa, 1998) are equally problematic to resolve. The main issue for these phyla is their lack of sufficient taxon sampling in previously published phylogenetic analyses (Halanych, 2004). Support has been recovered for a number of competing phylogenetic hypotheses regarding Scalidophora; ranging from: Priapulida + Kinorhyncha (Garey *et al.* 2001; Mallatt and Giribet, 2006; Campbell *et al.* 2011), Kinorhyncha + Nematomorpha (Hejnol *et al.* 2009) or Priapulida + Kinorhyncha + Nematoda (Dunn *et al.* 2008). However, these analyses all lack sufficient taxon sampling to be able to posit scalidophoran relationships with any substantial level of phylogenetic precision. This is particularly true of the Loricifera. Loricifera is one of the most recently discovered metazoan phyla (Kristensen, 1983) and so it is one of the least explored in terms of its evolutionary relationships (Park *et al.* 2006; Sørensen *et al.* 2008). To compound the lack of data and lack of inclusion in previous analyses, the study of Park *et al.* (2006) failed to find any significant support for placement of Loricifera within Ecdysozoa. This leaves us with only one relevant study, that of Sørensen *et al.* (2008) which included two loriciferan species and tentatively supported a relationship of this phylum with the Nematomorpha.

In this Chapter, given the abundance of SSU and LSU sequences in NCBI, I will investigate the evolutionary relationships of Ecdysozoa using these classic molecular markers. The aims of the analyses presented here are two. The first is to evaluate the extent to which the results we obtained using ESTs and miRNA are confirmed by the classic ribosomal markers. The second is to exploit the good taxonomic sampling available for this marker in order to attempt drafting a complete (i.e. including representative of all phyla) ecdysozoan phylogeny.

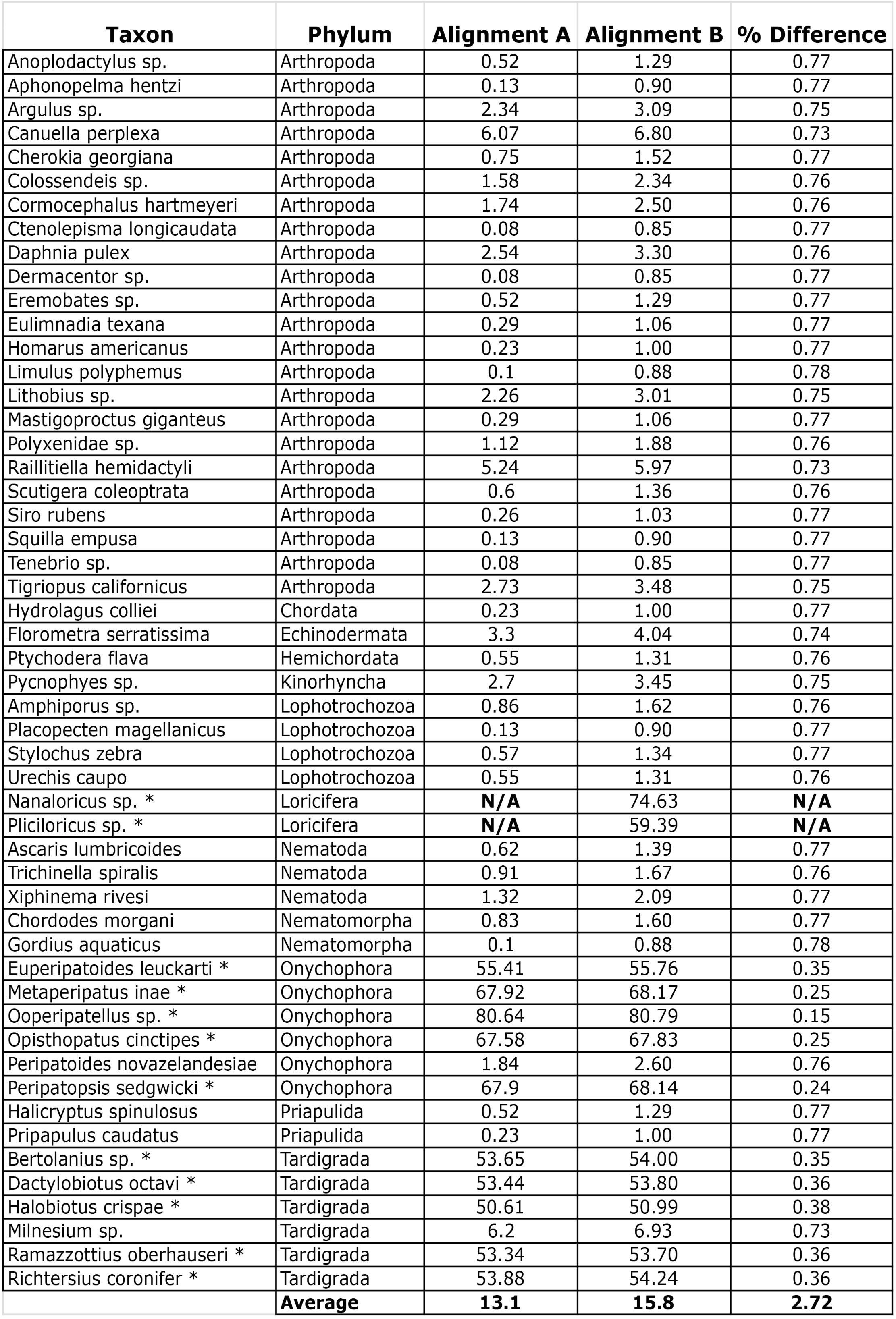
5.2 Materials and Methods

5.2.1 Alignment assembly

Two alignments were generated for the analyses presented in this Chapter. The main base-alignment (referred to here as alignment **A**) comprises a 50 taxon nuclear SSU and LSU rRNA gene dataset based on the alignment of Mallatt and Giribet (2006), and the second (referred to here as alignment **B**) is based on the same 50 taxon data set, but includes additional sequences for two species of Loricifera (*Nanaloricus*. sp. and *Pliciloricus* sp.). The starting alignment of Mallatt and Giribet was chosen for its quality (it was originally aligned using ribosomal secondary structure information, and only positions from easily aligned conserved regions were retained). However, differently from Mallatt and Giribet (2006) possible pseudogenes (e.g. for *Hanseniella* and *Sphaerotheriidae* – c.f. Mallatt and Giribet 2006) were not considered. The original alignment properties were kept, with the alignment length retained at 3,853 nucleotides as in the original Mallatt and Giribet dataset. The taxon sampling of Mallatt and Giribet was altered by deleting some of the taxa they used (41 ingroup arthropods) while also adding in complete or nearly complete SSU and partial LSU sequences for five tardigrades, five onychophorans and two loriciferans. The new sequences were obtained by blasting the NCBI database with *Peripatoides* and *Milnesium* SSU and LSU sequences. In addition, in order to reduce the computational burden, I reduced the sampling of over represented taxa such as Pancrustacea in the arthropods, keeping mostly moderately evolving taxa, also removing the two very fast evolving nematodes *Meloidogyne* and *Caenorhabditis* and the most fastest evolving of the Onychophora - *Peripatus* sp. The resulting rRNA alignments averaged 13.1% and 15.8% missing data for alignment **A** and **B** respectively (see Table 5.1), and comprised in total 23 Arthropoda, 6 Onychophora, 6 Tardigrada, 3 Nematoda, 2 Nematomorpha, 5 Scalidophora and 7 outgroups to the Ecdysozoa.

5.2.2 Phylogenetic analysis

All phylogenetic analyses were conducted under a Bayesian framework using PhyloBayes 3.2e (Lartillot *et al.* 2009). We first compared the fit of alternative models of evolution to our rRNA dataset excluding species for Loricifera. This was performed using Bayesian cross-validation (Stone, 1974) to rank the fit of alternative substitution models to the data. The models compared in this analysis were GTR+Γ, CAT+Γ, and CAT-GTR+Γ. Phylogenetic analyses of the rRNA datasets were performed under each model, and results were compared to evaluate whether different phylogenies were obtained when different-fitting models were used. Full details of how phylogenetic analyses were performed for these analyses see section 4.2.2 of Chapter 4.



Taxon	Phylum	Alignment A	Alignment B	% Difference
Anoplodactylus sp.	Arthropoda	0.52	1.29	0.77
Aphonopelma hentzi	Arthropoda	0.13	0.90	0.77
Argulus sp.	Arthropoda	2.34	3.09	0.75
Canuella perplexa	Arthropoda	6.07	6.80	0.73
Cherokia georgiana	Arthropoda	0.75	1.52	0.77
Colossendeis sp.	Arthropoda	1.58	2.34	0.76
Cormocephalus hartmeyeri	Arthropoda	1.74	2.50	0.76
Ctenolepisma longicaudata	Arthropoda	0.08	0.85	0.77
Daphnia pulex	Arthropoda	2.54	3.30	0.76
Dermacentor sp.	Arthropoda	0.08	0.85	0.77
Eremobates sp.	Arthropoda	0.52	1.29	0.77
Eulimnadia texana	Arthropoda	0.29	1.06	0.77
Homarus americanus	Arthropoda	0.23	1.00	0.77
Limulus polyphemus	Arthropoda	0.1	0.88	0.78
Lithobius sp.	Arthropoda	2.26	3.01	0.75
Mastigoproctus giganteus	Arthropoda	0.29	1.06	0.77
Polyxenidae sp.	Arthropoda	1.12	1.88	0.76
Raillitiella hemidactyli	Arthropoda	5.24	5.97	0.73
Scutigera coleoptrata	Arthropoda	0.6	1.36	0.76
Siro rubens	Arthropoda	0.26	1.03	0.77
Squilla empusa	Arthropoda	0.13	0.90	0.77
Tenebrio sp.	Arthropoda	0.08	0.85	0.77
Tigriopus californicus	Arthropoda	2.73	3.48	0.75
Hydrolagus colliei	Chordata	0.23	1.00	0.77
Florometra serratissima	Echinodermata	3.3	4.04	0.74
Ptychodera flava	Hemichordata	0.55	1.31	0.76
Pycnophyes sp.	Kinorhyncha	2.7	3.45	0.75
Amphiporus sp.	Lophotrochozoa	0.86	1.62	0.76
Placopecten magellanicus	Lophotrochozoa	0.13	0.90	0.77
Stylochus zebra	Lophotrochozoa	0.57	1.34	0.77
Urechis caupo	Lophotrochozoa	0.55	1.31	0.76
Nanaloricus sp. *	Loricifera	**N/A**	74.63	**N/A**
Pliciloricus sp. *	Loricifera	**N/A**	59.39	**N/A**
Ascaris lumbricoides	Nematoda	0.62	1.39	0.77
Trichinella spiralis	Nematoda	0.91	1.67	0.76
Xiphinema rivesi	Nematoda	1.32	2.09	0.77
Chordodes morgani	Nematomorpha	0.83	1.60	0.77
Gordius aquaticus	Nematomorpha	0.1	0.88	0.78
Euperipatoides leuckarti *	Onychophora	55.41	55.76	0.35
Metaperipatus inae *	Onychophora	67.92	68.17	0.25
Ooperipatellus sp. *	Onychophora	80.64	80.79	0.15
Opisthopatus cinctipes *	Onychophora	67.58	67.83	0.25
Peripatoides novazelandesiae	Onychophora	1.84	2.60	0.76
Peripatopsis sedgwicki *	Onychophora	67.9	68.14	0.24
Halicryptus spinulosus	Priapulida	0.52	1.29	0.77
Pripapulus caudatus	Priapulida	0.23	1.00	0.77
Bertolanius sp. *	Tardigrada	53.65	54.00	0.35
Dactylobiotus octavi *	Tardigrada	53.44	53.80	0.36
Halobiotus crispae *	Tardigrada	50.61	50.99	0.38
Milnesium sp.	Tardigrada	6.2	6.93	0.73
Ramazzottius oberhauseri *	Tardigrada	53.34	53.70	0.36
Richtersius coronifer *	Tardigrada	53.88	54.24	0.36
	Average	**13.1**	**15.8**	**2.72**

Table 5.1: Percent of missing data for all 52 ecdysozoan taxa. Asterisks (*) indicate species absent from Mallatt and Giribet (2006) that were added as part of these analyses (see section 5.2.1).

5.2.3 Generating site stripping and signal dissection data sets

Site stripping analyses were performed using the Slow-fast method of (Brinkmann and Philippe, 1999) to estimate the rate of substitution of the sites in both alignments **A** and **B**. Parsimony scores for every site in each of the alignments were calculated for groups in our taxon set had constrained monophyly: these groups are as follows; alignment **A**: - (Pancrustacea, Myriapoda, Chelicerata, Tardigrada, Onychophora, Nematoda), alignment **B**: - (Pancrustacea, Myriapoda, Chelicerata, Tardigrada, Onychophora, Nematoda, Scalidophora). The rate of each site, in both alignments, was then independently estimated as the sum of their parsimony scores across all considered monophyletic groups. All parsimony analyses were performed using PAUP4b10 (Swofford, 2002). Sites in both alignments were then ranked according to their substitution rate and partitioned into classes. Both alignments had near identical rate distributions (max parsimony steps of 14 vs. 11) but differed in the number of sites found to be in a particular rate class, this is due to the alignment **B** including Loricifera (3,883 nucleotides) which had a larger fraction of fast evolving sites with reference to the remaining species we retained from Mallatt and Giribet (2006).

Partitioned alignments were then generated according to the distribution of site rates, by systematically removing (i) approximately the fastest 5% of the sites, that is, all characters with a slow-fast–estimated rate of five or more steps (total number of sites remaining in alignment **A**: 3,580; alignment **B**: 3,619; (ii) approximately the fastest 10% of the sites, that is, all characters with a slow-fast–estimated rate of four or more steps (total number of sites remaining in alignment **A**: 3,439; alignment **B**: 3,449); (iii) approximately the fastest 15% of the sites, that is, all characters with a slow-fast–estimated rate of three or more steps (total number of sites remaining in alignment **A**: 3,220; alignment **B**: 3,224). No additional data sets were created after removal of the fastest 15% of sites, as the rate of the remaining sites was extremely low; at a rate of 2 parsimony steps or less. However, a signal dissection analysis (Sperling *et al.* 2009a) was also performed, this was to examine the phylogenetic signal present in the data set containing only the fastest 10% of sites (414 nucleotides). This data set was then independently analysed in conjunction with the slow-fast partitioned data sets excluding the fastest 5%, 10% and 15% of sites.

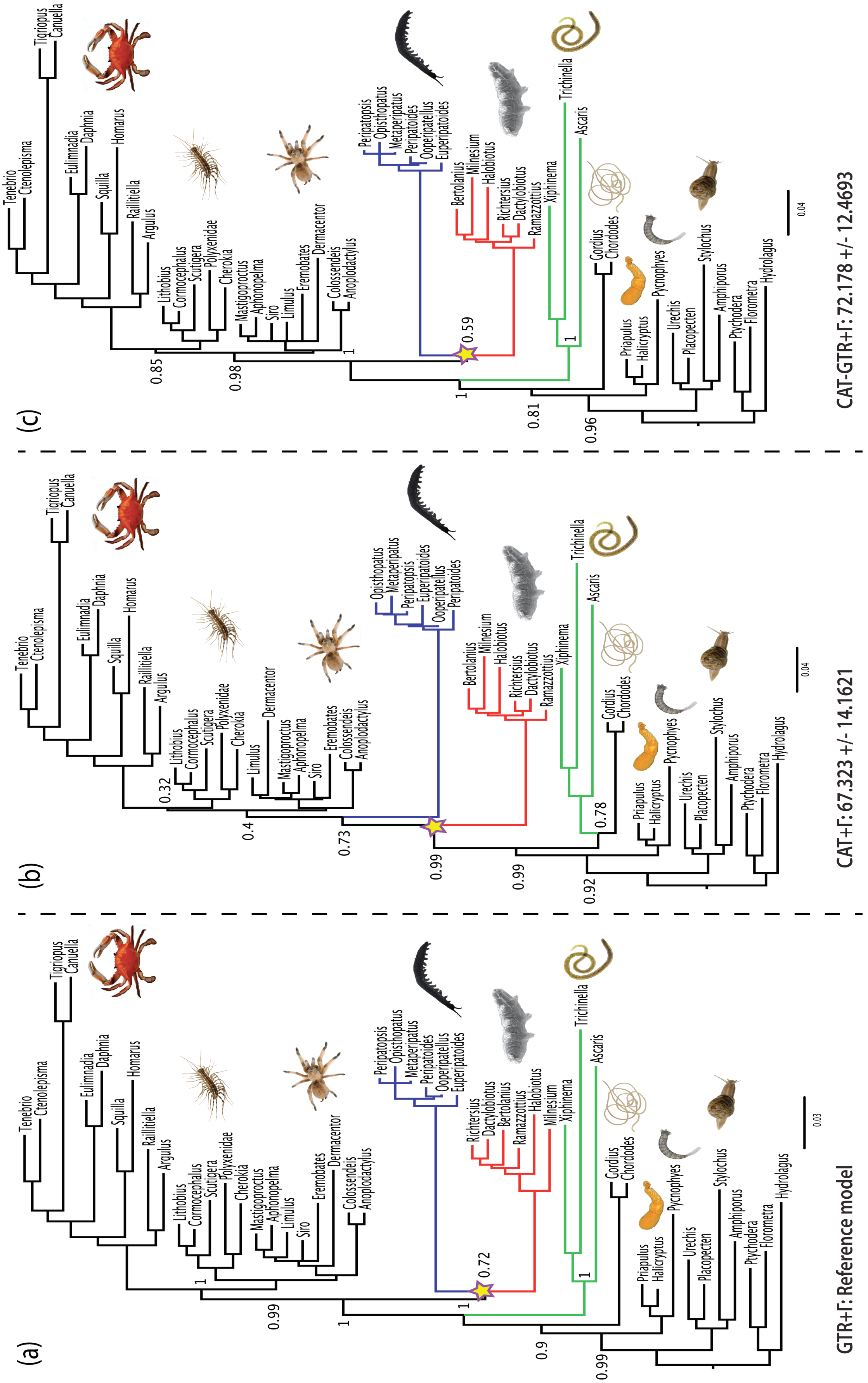
5.2.4 Taxon pruning analyses

In the same manner as for the EST analyses, the rRNA data sets were analyzed in order to examine the effect of taxon sampling on the recovery of alternate phylogenetic hypotheses within Ecdysozoa. It is well known that the number and nature of the taxa used can affect phylogenetic inference and, in particular, can exacerbate or reduce LBA (Aguinaldo *et al.* 1997; Philippe *et al.* 2005b; Holton and Pisani *et al.* 2010). Therefore I carried out two taxon-pruning experiments to evaluate the robustness of the RNA results. In the first, all slowly evolving ecdysozoan outgroups were excluded: the nematomorphs *Chordodes morgani* and *Gordius aquaticus*, and the scalidophorans *Halicryptus spinulosus*, *Priapulus caudatus* and *Pycnophyes sp*. This left the Nematoda as the sole, long branched outgroup. In the second experiment the onychophorans *Euperipatoides leuckarti, Metaperipatus inae, Opisthopatus cinctipes, Peripatopsis sedgwicki, Ooperipatellus* sp. and the tardigrades *Ramazzottius oberhauseri, Richtersius coronifer, Dactylobius octavi, Halobiotus crispae, Bertolanius* sp. were excluded. This left both the Onychophora and the Tardigrada represented by a single uninterrupted branch. Taxon pruning experiments were performed on alignment **A** solely, this was because the branch leading to Loricifera was the longest within my data set, indicating this group were the most unstable within our data set; therefore the species *Nanaloricus*. sp. and *Pliciloricus* sp. were excluded to avoid unwanted LBA artifacts.

5.3 Results

5.3.1 Deep divergences require site-heterogeneous models

Prior to conducting phylogenetic analysis of the RNA data sets, I ranked the fit of alternate substitution models to the data; with the aim of avoiding encountering systematic errors and the generation of tree biased by phylogenetic reconstruction artifacts. I first performed Bayesian crossvalidation (Stone, 1974) to rank substitution models according to their fit to the alignment. The substitution models tested in these analyses were the mechanistic GTR+Γ model, and the more complex heterogeneous mixture models CAT+Γ, and CAT-GTR+Γ. The results of the crossvalidation analyses are presented in Figure 5.1, in which they show the GTR+Γ (Figure 5.1a) model fits the dataset significantly less well than either the site-heterogeneous CAT+Γ (Figure 5.1b) or CAT-GTR+Γ (Figure 5.1c) model. It is apparent that the model GTR+Γ fits the data least, however results of these analyses do not clearly indicate which of site-heterogeneous models fits the data better. Despite CAT-GTR+Γ having a marginally better fit to the data, it is difficult to discriminate statistically between the two site-heterogeneous models for this data set, thus preventing me from drawing precise phylogenetic conclusions on the base of model fit alone. In addition to the three models discussed above, I also expanded model selection to include the Q-Matrix mixture model (QMM); this model employs multiple Q-Matrices each with



GTR+Γ: Reference model

CAT+Γ: 67.323 +/- 14.1621

CAT-GTR+Γ: 72.178 +/- 12.4693

🡐 **Figure 5.1: Bayesian crossvalidation and Phylogenetic analysis support monophyletic Panarthropoda and the paraphyletic nature of Cycloneuralia.** Compared models: GTR+Γ, CAT+Γ and CAT-GTR+Γ. (a) Tree obtained under GTR+Γ (i.e. under the reference model). (b) Tree obtained under CAT+Γ. (c) Tree obtained under CAT-GTR+Γ. The affinity of Tardigrada to Onychophora is model dependent. Δ log-likelihoods and standard deviations are shown for each model. The reference model is GTR+Γ and positive Δ-likelihood values identify a model with better fit than the reference model. Clades coloured for clarity, Onychophora (blue), Tardigrada (red), Nematoda (green); Node connecting Tardigrada highlighted by gold star.

their own distinct set of exchange rates and equilibrium frequencies. The Δ-likelihood value obtained in the crossvalidation analysis for the QMM model was exactly equal to that of CAT-GTR+Γ. Analyses under QMM are considerably more computationally expensive as QMM in essence uses multiple GTR matrixes; therefore this model was not considered in subsequent analyses.

5.3.2 rRNA supports the inclusion of Tardigrada within Panarthropoda and the paraphyletic nature of Cycloneuralia

Results of analyses under all considered models support Panarthropoda ((Posterior probability (PP) = 0.72, 0.99 and 1.0 for GTR, CAT and CAT-GTR respectively; see Figure 5.1)), while also supporting the paraphyletic origin of Cycloneuralia. However, exact topological relationships of the Tardigrada and Onychophora were model dependent. More precisely, the CAT model supports Tardigrada as the sister group of Lobopodia (PP=0.73; Figure 5.1b) while GTR and CAT-GTR support a sister group relationship between Onychophora and Tardigrada (PP = 1.0 and 0.59 for GTR and CAT-GTR; Figure 5.1a,c). Similar topological disagreement between the CAT model and the models GTR and CAT-GTR were observed in regard to the mono- vs paraphyletic nature of Nematoida (Nematoda + Nematomorpha); CAT was the only model found to support the monophyletic origin of Nematoida (PP=0.78). In a change of support, some agreement between models was obtained as all models corroborated the paraphyletic origin of Cycloneuralia, with GTR, CAT and CAT-GTR supporting this topology with a PP = 0.99, 0.92, 0.96 respectively. Finally, the topological relationships for the remaining ecdysozoan taxa are in broad agreement with one another across all models, with one exception; GTR was the only model found to recover Myriapoda as the sister group to Chelicerata (PP = 0.99; i.e. Myriochelata hypothesis) in contrast to CAT and CAT-GTR which both supported the monophyletic origin of Mandibulata (PP = 0.32 and 0.85).

5.3.3 Methods to uncover phylogenetic biases further support artifactual nature of Tardigrada plus Nematoda

In the analysis I present in this section, my aim is to understand the potential for phylogenetic artifacts as a result of model misspecification, presence of over saturated sites and the effect of reduced taxon sampling on the recovery of the different phylogenetic hypotheses supported in previous molecular studies of Ecdysozoa. Similarly to the rationale of experimental design presented in the EST Chapter of this thesis, I hypothesised that the suspected artifactual nature of Tardigrada sister group to Nematoda, obtained in previous analyses (e.g. Sørensen *et al.* 2008) being the result of LBA, should find support for this grouping to be highest in the data sets containing the largest proportion of fast evolving sites. Correspondingly, the opposite trend should be expected, in that support for the inclusion of Tardigrada within Panarthropoda would be maximised in the data sets excluding those fast evolving sites.

Results obtained from the series of slow-fast analyses (Brinkmann and Philippe, 1999) of the site rate partitioned data sets (generated for alignment A, see section 5.2.1) are again consistent with my hypothesis, in that all analyses conducted on the slowest evolving site partitions uniformly recover a monophyletic origin of Panarthropoda, while none of these analyses support the inclusion of Tardigrada within Cycloneuralia. Results of these analyses are summarized in Figure 5.2 and Table 5.2. Unsurprisingly, the most evident finding from these analyses is that no

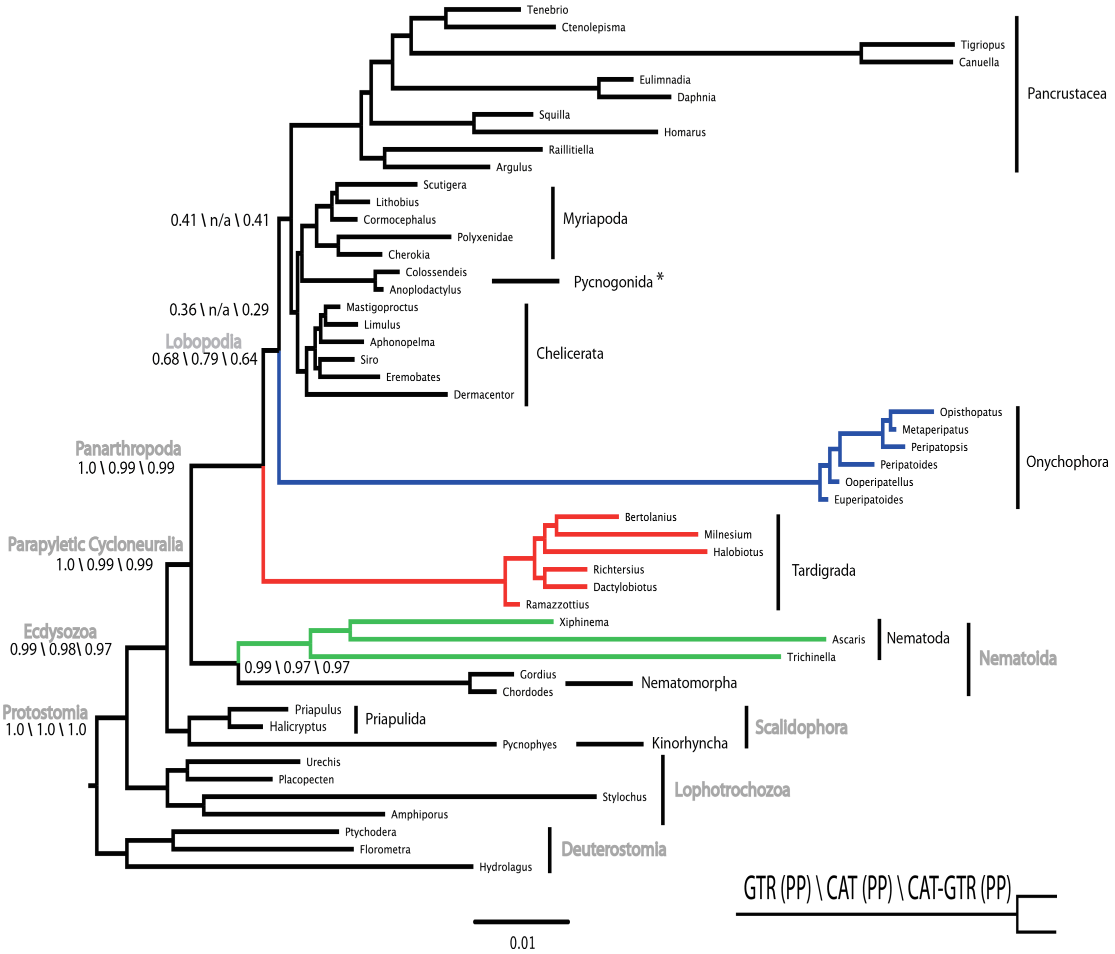
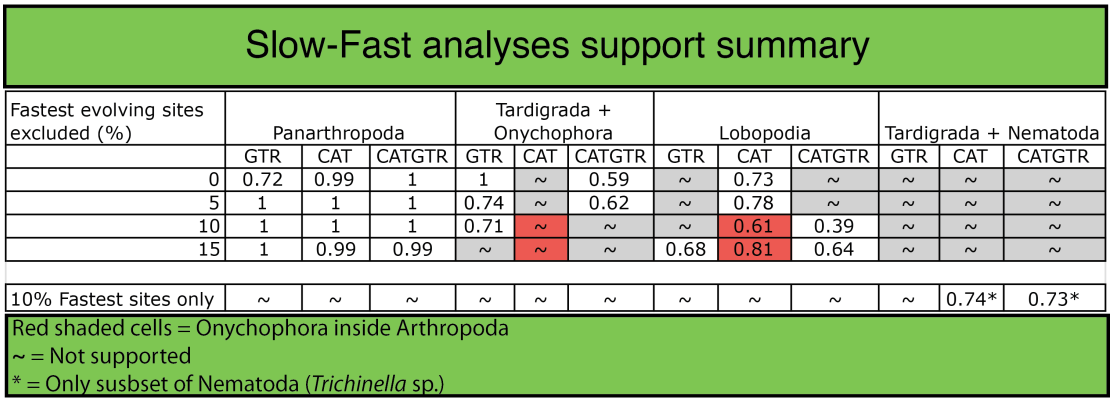


Figure 5.2: Summary of 18s + 28s rRNA site stripping analyses with 15% cut-off. Analyses under all three (GTR+ Γ, CAT+ Γ, CAT-GTR+ Γ) considered models support the inclusion of Tardigrada within a monophyletic Panarthropoda under all considered site-stripping cut-offs (5%, 10% and 15% of the fastest sites – see Methods and Table. 5.2). Elimination of the fastest 15% of sites for the GTR+Γ and CAT-GTR+Γ results in trees converging on Lobopodia. Support for Lobopodia peak in CAT+ Γ analyses when the fastest 15% of the sites are excluded (see Table 5.2) but at cut-offs of 10 and 15% Onychophora is found to nest within Arthropoda. Support values shown are Posterior probabilities (PP), (n/a) = not supported. Data set used in these analyses was alignment A (see Methods). * Indicates artifactual position of Pycnogonida sister group to the myriapods.

matter the degree to which fast evolving sites are removed (5%, 10% or 15%) support is never diminished for the monophyletic origin of Panarthropoda; in fact, we observe an overall increase in support for Panarthropoda across all models (GTR, CAT, CAT-GTR); particularly across all slow-fast data sets analysed under GTR (PP = 1.0). In addition, when 15% of the fastest sites were removed, I observed a switch in topology for both GTR and CAT-GTR in regards to the monophyly of Nematoida, to where all models (including CAT) now supported this group with near full PP support (see Figure 5.2).

Interestingly in the signal dissection analysis of the fastest evolving 10% of sites in the alignment (see Table. 5.2) all considered models (GTR, CAT, CAT-GTR) obtained weak PP support for topologies that were biologically implausible, with groups like Panarthropoda, and Lobopodia never recovered. Instead, spurious groups of taxa were recovered across all models, for instance Onychophora was found as the sister group to a clade composed of arthropods, tardigrades, nematodes and nematomorphs. Minor support was obtained for a Tardigrada – Nematode affinity, however the longest branched nematode (*Trichinella* sp.) was the only nematode to be recovered in such a position. The results of these signal dissection analyses clearly indicate that the data set containing only the 10% of fastest site is one that contains a high noise-to-signal ratio; accordingly these results have little phylogenetic utility on deciphering the relationships of Ecdysozoa. Signal dissection, albeit irrelevant here when trying to better understand the phylogenetic relationships among the Ecdysozoa, did confirm that the exclusion of the 10% fastest evolving sites from our data set could not have caused a loss of important phylogenetic signal.



Slow-Fast analyses support summary

Fastest evolving sites excluded (%)	Panarthropoda			Tardigrada + Onychophora			Lobopodia			Tardigrada + Nematoda		
	GTR	CAT	CATGTR	GTR	CAT	CATGTR	GTR	CAT	CATGTR	GTR	CAT	CATGTR
0	0.72	0.99	1	1	~	0.59	~	0.73	~	~	~	~
5	1	1	1	0.74	~	0.62	~	0.78	~	~	~	~
10	1	1	1	0.71	🟥 ~	~	~	🟥 0.61	0.39	~	~	~
15	1	0.99	0.99	~	🟥 ~	~	0.68	🟥 0.81	0.64	~	~	~
10% Fastest sites only	~	~	~	~	~	~	~	~	~	~	0.74*	0.73*

Red shaded cells = Onychophora inside Arthropoda
~ = Not supported
* = Only susbset of Nematoda (*Trichinella* sp.)

Table 5.2: Support summary for all Slow-fast analyses performed on alignment A. In all of the SF analyses, nearly full support was recovered for the monophyletic origin of Panarthropoda, yet support for the relationships of Tardigrada and Onychophora were model dependent. The details of support values are given at the bottom of the table.

In a final test, taxon-pruning experiments were conducted to evaluate further the robustness of my RNA results. This was done by selectively removing taxa (see section 5.2.4) to generate uninterrupted long branches for Tardigrada, Onychophora, and Nematoda. Results of these analyses are presented in Figure 5.3; and show that the affect of removing specific taxa to exacerbate LBA had no effect on the position of Tardigrada. One apparent trend observed in these taxon pruning experiments was the overall loss of support for Lobopodia, with some analyses supporting unlikely affinities for Onychophora (see Figure 5.3c). Accordingly, interpretation of these experiments suggest that this new rRNA data set, when analyzed using complex models of evolution like CAT+Γ and CAT-GTR+Γ (which fit the data markedly better), or alternatively even poor fitting site-homogeneous models (GTR+Γ) is robust against the recovery of artifactual topologies brought about by LBA.

Concluding, it is clear that adequate phylogenetic signal is present within this data set, phylogenetic signal that undoubtedly supports the panarthropodan affinities of the Tardigrada. However, according to the results of SSU/LSU rRNA analyses presented

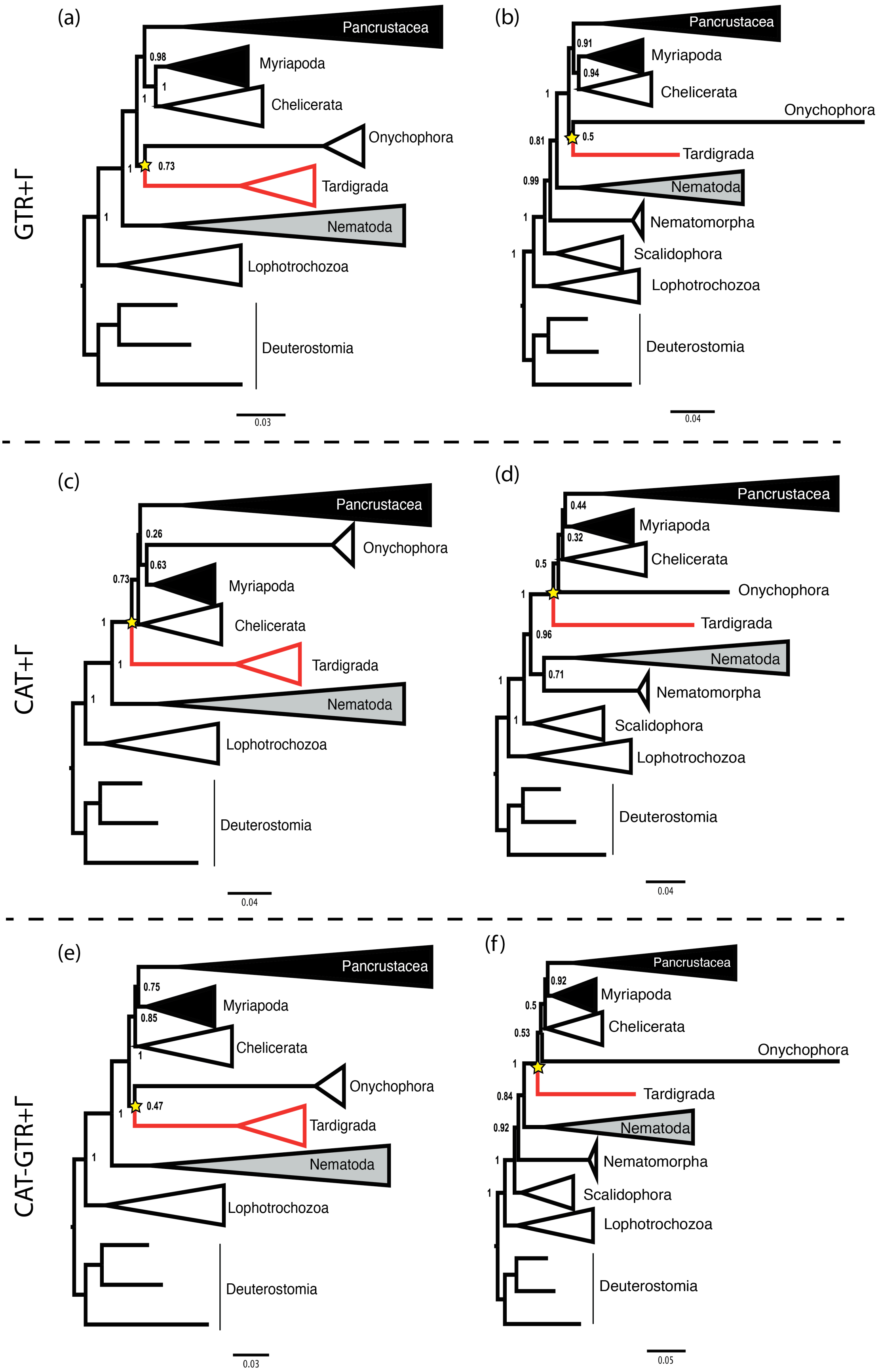


Figure 5.3: Taxon pruning analyses designed to exacerbate LBA.
Two taxon pruning experiments performed under GTR+Γ, CAT+Γ, CAT-GTR+Γ; designed to increase the affect of LBA, by removing (i) all the slowly evolving ecdysozoan outgroups leaving only Nematoda (**a, c, e**), (ii) All but one species from both Onychophora and Tardigrada (**b, d, f**). Taxon prunning experiments support the Panarthropod affinity of Tardigrada and show this position is unaffected by LBA.

here (which are model dependent) validation of Lobopodia warrants further investigation as this group is only partially supported in these analyses.

5.3.4 Maintaining support for Panarthropoda and the weak phylogenetic signal for the placement of Loricifera

Following from initial phylogenetic analyses using rRNA, I wanted investigate the phylogenetic placement of Loricifera within the Ecdysozoa. Currently, there are scant numbers of molecular phylogenetic studies that deal with the placement of Loricifera (Park *et al.* 2006; Sørensen *et al.* 2008). In the most recent analyses including data for this phylum, there is some evidence to suggest the placement of Loricifera resides with the parasitic horsehair worms (Nematomorpha) (Sørensen *et al.* 2008). However, this position disagrees with established morphological support in favour of a monophyletic sister group relationship between Nematoda and Nematomorpha (Nematoida; Schmidt-Rhaesa, 1998; Nielsen, 2001). In addition to this unlikely position for Loricifera, Sørensen *et al.* (2008) also recover a sister group position of Tardigrada + Nematoda.

Preliminary analyses of the rRNA data set that includes full and partial SSU (18S) sequences for two species of Loricifera was carried out using the models (GTR, CAT, and CAT-GTR) on the full length alignment (see section 5.2.1). The results of these analyses are shown in Figure 5.4a,b; which support the sister group relationship between Loricifera and Onychophora, a position recovered across all evolutionary models considered (PP = 0.64, 0.53, 0.55 for GTR, CAT and CAT-GTR respectively). However, this position is highly likely to be artifactual, as neither morphological nor molecular sequence data has been found previously to support such a relationship.

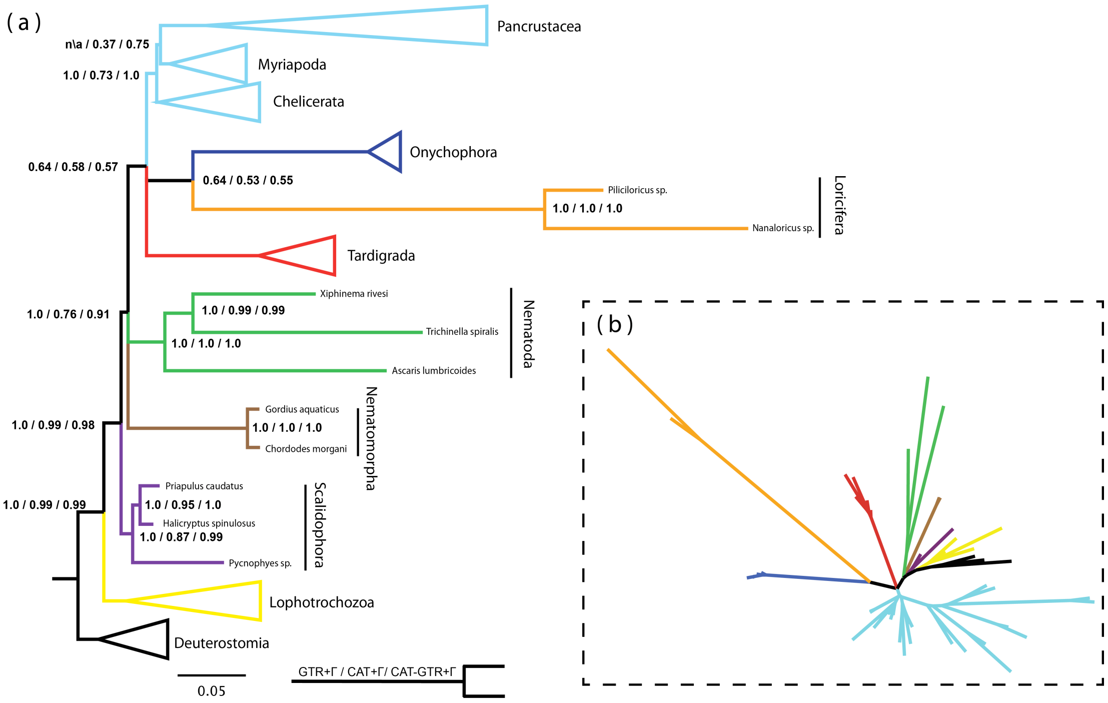
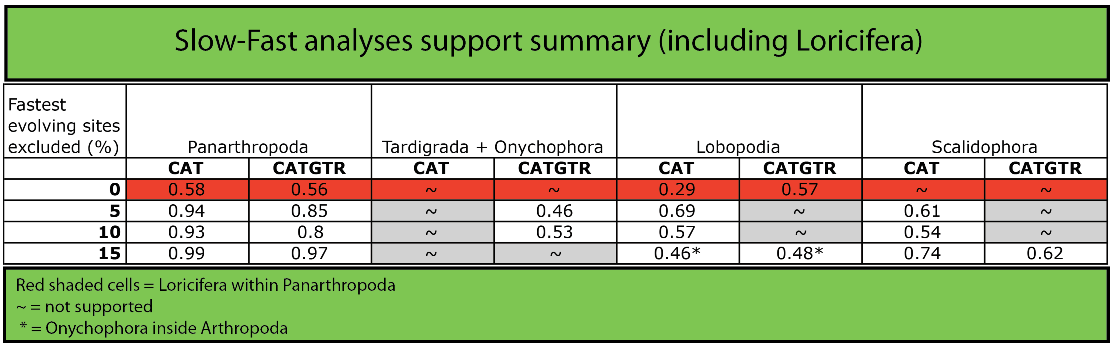


Figure 5.4: Unresolved position of Loricifera due to weak phylogenetic signal. Phylogenetic analysis of rRNA data set including two species for Loricifera. Analyses performed under all models (GTR+Γ, CAT+Γ, CAT-GTR+Γ) on the full alignment, recover the same artifactual position of Loricifera as the sister group to Onychophora. (a) Consensus topology of all three considered models supports Loricifera as sister to Onychophora. (b) Radial tree of same topology highlighting the extremely long branch for Loricifera.

Furthermore, the amount of missing data within this data set is greatest for these two phyla, averaging 57.2% for Onychophora and 67.0% for Loricifera. The long branch connecting the Loricifera to the rest of the tree is obvious from Fig. 5.4b, and confirms the unlikely nature of this result. The large amount of missing data in these species could have caused problem with ancestral character state optimisation (under ML and Bayesian analyses gapped sites are inferred as those maximising the likelihood for the considered site). In any case it seems likely that genuine phylogenetic signal is low for these two groups, which would increase the ratio between noise and phylogenetic signal. Unsurprisingly, inclusion of data for Loricifera had the effect of dramatically reducing the support for a monophyletic origin of Panarthropoda, likely due to the unstable placement of Loricifera.



Slow-Fast analyses support summary (including Loricifera)

Fastest evolving sites excluded (%)	Panarthropoda		Tardigrada + Onychophora		Lobopodia		Scalidophora	
	CAT	CATGTR	CAT	CATGTR	CAT	CATGTR	CAT	CATGTR
0	0.58	0.56	~	~	0.29	0.57	~	~
5	0.94	0.85	~	0.46	0.69	~	0.61	~
10	0.93	0.8	~	0.53	0.57	~	0.54	~
15	0.99	0.97	~	~	0.46*	0.48*	0.74	0.62

Red shaded cells = Loricifera within Panarthropoda
~ = not supported
* = Onychophora inside Arthropoda

Table 5.3: Support summary for all Slow-fast analyses performed on alignment B. Results of the SF analyses highlight unstable nature of Loricifera, as multiple placements are recovered; either within Panarthropoda, sister group to Nematomorpha, or within a monophyletic Scalidophora. Support was however maintained for the monophyly of Panarthropoda, but again the relationships of Tardigrada and Onychophora were model dependent. The details of support values are given at the bottom of the table.

In a final attempt to uncover genuine phylogenetic signal, I performed another series of site-stripping experiments using the slow-fast technique (Brinkmann and Philippe, 1999). To do this I progressively removed sites from the alignment, resulting in three additional data sets excluding the fastest evolving 5%, 10% and 15% of sites from the alignment (section 5.2.3) and then analysed these using the two best fitting models (CAT+Γ, CAT-GTR+Γ) identified from the crossvalidation analysis (section 5.3.1). Analyses performed on these more rate homogeneous data sets, with the analysis of the slowest 95% of sites shown in figure 5.5 (also summarized in Table 5.3) resulted in a number of topological changes compared to results found from the original full-length data set. Most notably was the now lack of support for a sister group position of Loricifera and Onychophora, Loricifera was instead recovered as the earliest branching phylum within a monophyletic Scalidophora, a position recovered under both models (CAT+Γ: PP = 0.74 and CAT-GTR+Γ: PP = 0.62) when analyses were performed on the most site-homogeneous data set (15% cut-off). This position was also found under CAT+Γ for the less stringent cut-offs (5% and 10%), however

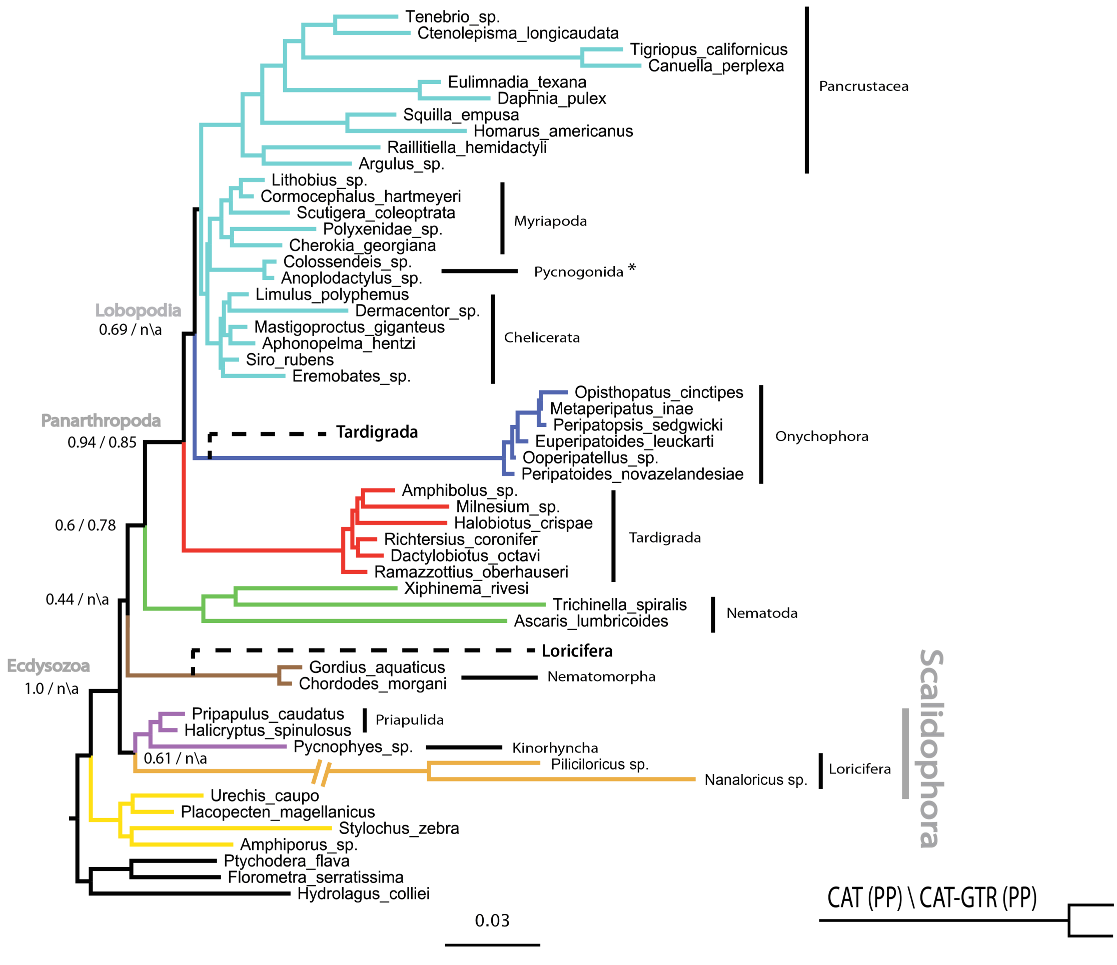


Figure 5.5: Site stripping analysis including sequences for Loricifera. Topology generated by analyzing the dataset (alignment **B**) with 5% of the fastest sites removed. Analyses performed under both CAT+Γ and CAT-GTR+Γ. Moderate PP support is recovered for the inclusion of Loricifera within a monophyletic Scalidophora (under CAT-Γ), while CAT-GTR-Γ supports Loricifera + Nematomorpha with weak PP support. Alternate position of Loricifera and Tardigrada recovered under CAT-GTR+Γ indicated by dashed branches. Asterisks indicate artifactual position for Pycnogonida.

support was reduced in both cases (see Figure 5.5); while CAT-GTR+Γ weakly supported either a branching position of Loricifera between the remaining scalidophorans and the other Ecdysozoan taxa, or alternatively as the sister group to Nematomorpha (weak PP of 0.29; see Figure 5.5). As a consequence of the recovery of Loricifera within Cycloneuralia, high support was again recovered for the monophyly of Panarthropoda, peaking under both CAT+Γ (PP = 0.99) and CAT-GTR+Γ (PP = 0.97) when analysing the data set with 15% of the fastest sites removed (see Table 5.3). However, within Panarthropoda, the recovery of Lobopodia versus a sister group relationship of Tardigrada + Onychophora was again model dependent, with CAT+Γ weakly supporting Lobopodia for all three site-stripped data sets (See Table 5.3). Conversely, support for Lobopodia was only recovered under CAT-GTR+Γ for the most site-homogeneous data set (15%), yet CAT-GTR+Γ and CAT+Γ both recovered Onychophora to be within a partially unresolved Arthropoda under this most stringent of the site-stripped data sets.

5.4 Concluding remarks

Since the earliest days of molecular phylogenetics, the phylogenetic utility of rRNA molecules has been recognised (e.g. Woese *et al.* 1990). With reference to Metazoa, the work of Aguinaldo et al (1997) was the first to introduce the now well-accepted ‘new animal phylogeny’. Following from this classic study, many of the now well-accepted hypotheses of relationships among the major metazoan groups were proposed from the analyses of rRNA data e.g. Arthropoda: (Giribet and Ribera, 1998); Annelida: (Rousset *et al.* 2004).

With reference to the Ecdysozoa, different studies based on rRNA data obtained multiple well supported competing hypotheses e.g. Giribet *et al.* (1996) and Garey *et al.* (1999) versus Garey *et al.* (2001) and Mallatt and Giribet (2006) with reference to the position of the Tardigrada. Accordingly, given also that only rRNA data are available for all phyla within Ecdysozoa, I attempted to establish a reliable rRNA-based ecdysozoan phylogeny. To do so I modified the well curated 18S + 28S rRNA dataset of Mallatt and Giribet (2006) to which I added sequences for underrepresented lineages (Tardigrada, Onychophora, Loricifera) while also removing some of the most rapidly evolving and over represented taxa. This dataset was subjected to model selection, taxon pruning and site-stripping experiments, and allowed generation of yet another independently acquired set of phylogenies to describe the evolution of Ecdysozoa. It has been noted, and widely discussed in this thesis, that one of the best proxies for phylogenetic accuracy is the congruence of independent data sets (Pisani *et al.* 2007; Campbell *et al.* 2011). In relation to the work presented in the preceding Chapters of this thesis, here, a further line of evidence to test the ecdysozoan phylogeny has been presented.

The results of the rRNA analyses found further support for clades found by our EST and miRNA analyses e.g. Panarthropoda (Nielsen, 2001). Cycloneuralia (Ahlrichs, 1995) is supported by our rRNA analyses and this is also congruent with the results of our miRNA and EST analyses. Considering previous morphological support, in the recent publication of Telford *et al.* (2008), the authors were in favour of the paraphyletic origin of Cycloneuralia. Scalidophora is supported as a monophyletic group in these rRNA analyses. According to this study the Loricifera might also be true scalidophorans, a position highly supported by morphology (Nielsen, 2012; but see Nielsen, 2001 for ref). This result is interesting because it has not been previously obtained from the analyses of rRNA or other types of molecular data. However, there is morphological evidence that could support it as the Loricifera share with the other scalidophorans the possession of an introvert with scalids and the presence of two rings of retractor muscles on the introvert (Heiner and Kristensen, 2005; Telford *et al.* 2008). An association of the Loricifera to the scalidophorans is thus expected (morphologically speaking) but needs further confirmation as support was low in these molecular analyses, additionally the high amount of missing data in the loriciferan sequences is potentially problematic (see above).

Within Panarthropoda, our rRNA analyses provide further support for a sister group relationship of Tardigrada + Lobopodia (Onychophora + Arthropoda). This result, which contradict previous finding by Mallatt and Giribet (2006) is in agreement with my miRNA and EST analyses. Dissimilarly to the results of my EST data sets, the rRNA analyses did not find any robust evidence that could possibly support a sister group relationship between Tardigrada and Nematoda. This finding further increases the likelihood that previous molecular support for Tardigrada + Nematoda (Giribet and Ribera, 1998; Giribet and Wheeler, 1999; Park *et al.* 2006; Philippe *et al.* 2005b; Roeding *et al.* 2005; Lartillot and Philippe, 2008; Sørensen *et al.* 2008; Pick *et al.* 2010; Andrew, 2011) could have been caused by uncorrected systematic biases. Substantial support was also recovered in favour of a monophyletic Nematoida (Nematoda + Nematomorpha) as analyses under the two best fitting models both supported this topology. This was not unexpected, as many previous studies also supported this group (Peterson and Eernisse, 2001; Mallatt *et al.* 2004; Mallatt and Giribet, 2006; Dunn *et al.* 2008); in addition to the strong morphological support for Nematoida (Schmidt-Rhaesa, 1998; Nielsen, 2001).

In conclusion, the phylogeny of the Ecdysozoa has received much attention since the onset of the molecular era (Field *et al.* 1988; Giribet *et al.* 1996; Aguinaldo *et al.* 1997; Telford *et al.* 2003; Philippe *et al.* 2005b; Dunn *et al.* 2008; Hejnol *et al.* 2009; Rota-Stabelli *et al.* 2010; Campbell *et al.* 2011) despite this, a consensus has yet to be reached on the exact topological relationships of its constituent phyla (Telford *et al.* 2008; Edgecombe, 2009; Campbell *et al.* 2011; Nielsen, 2012). Following on from the results presented in this Chapter, the most credible hypothesis for the evolutionary relationships among the Ecdysozoa are reported in Figure 5.2. Ecdysozoa can be partitioned into a monophyletic Panarthropoda (Tardigrada + Lobopodia) the sister group of which is represented by the Nematoida. The name "Ambulavermia" is proposed for this, currently unnamed group. Finally, the sister group of the ambulavermians is represented by the Scalidophora to which the Loricifera also seem to belong.

Chapter 6

Discussion and Perspectives

> *"No naturalist can avoid being fascinated by the diversity of the animal kingdom, and by the sometimes quite bizarre specializations that have made it possible for the innumerable species to inhabit almost all conceivable ecological niches"*
>
> *-Claus Nielsen*

6.1 Making sense of Cryptic divergences with phylogenomics

There is no doubt that we live in a world that has seen tremendous transformation over its extensive geological history; yet as a species, humans have been absent for the vast majority of this time. Our relatively momentary existence is in stark contrast to the immense age of the deepest branches of the animal tree of life, some of which have flourished for well over 700 million years (Peterson *et al.* 2008; Erwin *et al.* 2011). The notion of expansive geological history coupled with an ever-increasing diversity of animal life is even more profound when we consider for instance current estimates of extant species numbers compared to those that are long extinct. Ecdysozoa alone comprises ~1.5 million species (Chapman, 2009) yet despite being one of the most specious groups of animals to exist today, pales in insignificance when all living species are estimated to only represent a meager fraction (~0.1%) of the total number of species that ever existed (Raup, 1981; Novacek and Wheeler, 1992).

Since the first molecular phylogeny was published in support of Ecdysozoa (Aguinaldo *et al.* 1997), It is now generally accepted that Ecdysozoa is monophyletic

(Kumar *et al.* 2011). This group is generally assumed to comprise two distinct clades, Panarthropoda with segmental bodies with limbs and paired ganglia (the monophyly of which was confirmed in this thesis) and Cycloneuralia without limbs and with a collar shaped brain (Nielsen, 2001; Edgecombe *et al.* 2011) the monophyly of which was rejected in this thesis. We have shown that despite the "cycloneuralians" and the Panarthropoda are for the most part morphologically well delineated, the positions of the tardigrades has long been unstable in both morphological and molecular analyses (Edgecombe, 2009). In Chapter 3, I presented phylogenomic analyses of a 255 gene (49,023 amino acids) concatenated alignment, to investigate the major Ecdysozoan relationships, paying particular heed to the affinity of Tardigrada. Results of these analyses support the inclusion of Tardigrada within Panarthropoda, but they also demonstrate the unstable nature of Tardigrada in phylogenomic analyses, highlighting the importance of taxon sampling, and the presence of conflicting phylogenetic signal for Tardigrada. In addition, these results rejected the monophyly of the cycloneuralians. This is important, given that we found Tardigrada to be monophyletic, as it concurs with support of the plesiomorphic nature of the cycloneuralian morphological characters present in tardigrades. That is, paraphyly of cycloneuralians suggests that the characters shared by the Scalidophora, the Nematoida and Tardigrada represent retained plesiomorphies that presumably characterised the last common ecdysozoan ancestor. With reference to taxonomic sampling, inclusion of a representative species for Nematomorpha was of particular importance, as the sister group relationship between Nematoda and Nematomorpha (to form Nematoida) seems unquestionable, supported by the majority of morphological and molecular analyses (Nielsen, 2001; Kristensen, 2003; Mallatt and Giribet, 2006; Dunn *et al.* 2008).

Its clear that in recent years we have witnessed a marked move from small scale studies of limited numbers of genes, commonplace in 'classical' molecular phylogeny (e.g. SSU and LSU rRNA) towards large scale analyses of greater numbers of genes characteristic of EST (i.e. phylogenomics) based studies. Yet there exist limitations to the phylogenomic approach. For instance the study of Hejnol *et al.* (2009) generated an encompassing data set of 1,487 genes for 97 taxa, however this study and others like it require extremely powerful computational resources, not readily available within the phylogenetic community. Furthermore, one of the major limitations to phylogenomic scale analysis under the supermatrix paradigm is the use of evolutionary models that are required to describe the evolution of multiple genes that have undergone radically different evolutionary trajectories (de Queiroz *et al.* 2007; Philippe *et al.* 2011b; Philippe and Roure, 2011).

In much of this thesis I have highlighted problems in current models of evolution and their propensity to misinterpret or their failure to detect multiple substitutions, leading to what we collectively refer to as "non-phylogenetic signal" (Philippe *et al.* 2011b). In such cases, evident in the majority of phylogenetic analyses of ancient groups of taxa such as tardigrades and the arthropod sub phyla (e.g. Myriapoda and Chelicerata (Pisani *et al.* 2004; Pick *et al.* 2010; Rota-Stabelli *et al.* 2011; Campbell *et al.* 2011)) there is an inherent absence of natural phylogenetic signal to the point where sophisticated models of evolution often fail to unambiguously solve these problematic nodes. Lack of genuine phylogenetic signal and occurrence of systematic bias was forcefully addressed in Chapter 3, where I showed how the use of alternate models of evolution resulted in generation of radically different tree topologies (Figure 3.2). Failure of current models to capture inherent evolutionary process, for example the misconception of particular models to assume homogeneity of the replacement process, is one of the major hurdles of current evolutionary models used in the phylogenomic study of organismal relationships (Philippe and Roure, 2011).

The inability to fully account for unequal rates of evolution in current models and the difficulty in resolving deep nodes characterized by rapid divergence and multiple hidden substitutions was demonstrated in Chapter 3, and again in Chapter 5 were I presented a classical molecular phylogeny for Ecdysozoa using SSU/LSU rRNA. In both Chapters I performed site-striping and signal dissection analyses with the aim of generating data sets with more homogeneous rates of evolution, and then compare and contrast results of their analysis against results generated from heterogeneous-fast evolving site alignments. The artifactual nature of Tardigrada was clearly shown in Chapter 3, as analyses generated under the more rate homogeneous data sets (therefore less likely to have diluted phylogenetic signal) compared to those of the faster evolving data sets, unequivocally demonstrated the recovery of two highly supported but conflicting tree topologies. According to my results in Chapter 3, there is clearly need for caution when investigating problematic nodes like those of Tardigrada and Myriapoda when using a phylogenomic approach. Nodes such as these are characterized by short internal branches, rapid divergences, and high rate of substitution in extant lineages making them prone to errors of tree reconstruction introduced by systematic bias (i.e. LBA) and problems of taxon sampling leading to the recovery of highly supported yet equally contradictory phylogenies.

Dissimilarly to the non-phylogenetic signal shown to be pervasive for Tardigrada in Chapter 3, analyses of rRNA showed a seemingly clear-cut phylogenetic signal for their placement within a monophyletic Panarthropoda. Drawing conclusions on the contrasting phylogenetic signal strength of the rRNA data compared to that of the EST data set I suspect is due to various factors. Firstly, experimental design in Chapter 5 focused on generating a robust phylogeny for Ecdysozoa by utilizing sequence data for all ecdysozoan phyla. Secondly, taxonomic sampling within focal groups such as Onychophora and Tardigrada was considerably improved in our analyses compared to previous rRNA data sets (e.g. Mallatt and Giribet, 2006), promoting substantial reduction of stem branches for both groups. And lastly, the sequence alignment was based on proportions of sites taken from the most conserved ribosomal regions. Combining thorough taxon sampling with a compact data set of highly conserved SSU/LSU regions, then performing analyses with sophisticated models of sequence evolution should as it has already been noted (Lartillot and Philippe, 2008; Pisani *et al.* 2011; Philippe *et al.* 2011b) dramatically improve the ratio of phylogenetic signal to noise, and lead to well resolved and supported taxon placement.

In Chapter 3, and elsewhere throughout this thesis I have reiterated that certain phylogenetic relationships are heavily dependent upon the methods used e.g. model fitting, signal dissection (Sperling *et al.* 2009a; Pisani *et al.* 2009), and the importance of targeted taxon sampling (Rota-Stabelli and Telford, 2008). But how does one ascertain satisfactory confidence in the results obtained from different methods of analysis when faced with multiple conflicting and highly supported hypotheses? In accordance with the overwhelming trend I have witnessed from the analyses presented in this thesis, the most promising way to achieve robust confidence in a particular phylogenetic hypotheses is by critical dissection of the underlying phylogenetic signal(s) present in the data.

I must point out that it is not sufficient enough to merely present phylogenies obtained under the most 'optimal' methodological settings or via the criterion of it's the 'best fitting' model, as although the model may be suitable enough it may not be the best available. While the aforementioned properties provide initial phylogenetic confidence, to achieve a high level of confidence you must adhere to comparing and evaluating phylogenies obtained over different methods, to best identify cases of systematic or stochastic error. Comparing results over different methods, if found to be consistent, can indicate whether or not the resulting phylogeny is robust. For instance it can be useful to compare phylogenies generated under conditions that minimise potential sources of error against those that are generated under settings that maximise sources of phylogenetic error. Comparisons of trees generated under such different methodological settings was shown in both Chapters 3 and 5, for instance trees generated under different models (e.g. Figure 3.2), selective taxon pruning (3.7, 5.3) or in data sets generated to increase the level of rate homogeneity (e.g. Figure 3.5, 5.2). Comparisons of trees generated under these different analytical settings provided the opportunity to indentify what affects these settings had the recovery of alternate topologies. The presence of conflicting phylogenetic signals and non-phylogenetic signal has been demonstrated throughout this thesis, however, experimental approaches based on taxon sampling and signal dissections allowed distinguishing the most robust signal, one likely to represent the real phylogeny i.e. monophyletic Panarthropoda.

A major focal point in this thesis is the first use of miRNA evidence to polarise the phylogenetic placement the major Arthropod sub-phyla, and arthropod sister phyla

Onychophora and Tardigrada. In the preceding paragraphs, and elsewhere throughout this thesis I have discussed the problems inherent in, and the limitations of, analysis of large phylogenomic data sets, which I advocate are problems related to homoplasy. Chapter 4 sees a move away from use of classical mainstream molecular and morphological data types to investigate animal evolution, towards use of a relatively novel source of phylogenetic data (miRNAs) recently shown to be invaluable for testing alternate hypotheses of evolution (Pisani *et al.* 2011; Philippe *et al.* 2011a; Tarver *et al.* 2012). Accordingly, one of the major goals of this work is to test the alternate, conflicting hypotheses of within-ecdysozoan evolution by utilizing inherent properties of miRNA evolution (discussed at length in sections 2.2.2 & 4.1.2), properties that make them a homoplasy-low source of phylogenetic data (Sperling and Peterson, 2009; Tarver *et al.* 2012).

In Chapter 4, I presented two separate miRNA analyses that were performed in order to resolve competing hypotheses of evolution for Myriapoda (Mandibulata vs. Myriochelata) and the panarthropod phyla Onychophora and Tardigrada (mono- vs. paraphyletic Panarthropoda). According to results of investigations into shared miRNA complements for all considered taxa, I have recovered unequivocal support for some long held traditional hypotheses, these being monophyletic Mandibulata (see Figure 4.6; Snodgrass, 1938) and Panarthropoda comprised of tardigrades as the sister group to Onychophora plus Arthropoda (see Figure 4.8: Lobopodia; Snodgrass, 1938).

In conclusion, I advocate the thorough and detailed investigation of phylogenetic signal, when faced with resolving difficult problematic nodes characterized by high levels of homoplasy and the recovery of highly supported yet conflicting hypotheses. This can be achieved by indentifying potential sources of change in phylogenetic signal and the support of alternate topologies, by examining factors (for example taxon sampling shown as a key factor throughout this thesis) that can potentially lead to occurrences of stochastic/systematic error (e.g. unequal rates of evolution in both sites and taxa). There are many publications that follow this principle of detailed phylogenetic scrutiny (Lartillot and Philippe, 2008; Sperling *et al.* 2009a; Pisani *et al.* 2011; Rota-Stabelli *et al.* 2011) and indeed these already have provided some well-supported and robust phylogenies. I would also like to reiterate here, following on from what has already been advocated throughout this thesis, is the crucial importance of evaluating the robustness of a particular tree (hypothesis) with corroboration of multiple lines of independent evidence. This is the principle of consilience (see Wilson, 1998) and it is one in which I adhered to in this thesis. Accordingly, phylogenetic results supported in this thesis are those supported by the multiple lines of evidence used (ESTs, miRNAs, and rRNA). These lines of evidence provided robust evidence to support previously proposed relationships within Ecdysozoa, specifically monophyly of Panarthropoda comprised of Tardigrada + Lobopodia, and the sister group relationship of Myriapoda to Pancrustacea; robust evidence provided from the corroboration of not only phylogenomics (Chapter 3), classical rRNA (Chapter 5) and the recently emerged miRNAs (Chapter 4).

6.2 Resurrecting ancestral bauplaene within Ecdysozoa based on current evidence

In the past 15 years since the proposal of Ecdysozoa (Aguinaldo *et al.* 1997) the debate over whether or not Ecdysozoa is monophyletic (contra to Articulata; discussed in section 3.1.1) has largely been put to rest from analyses of molecular data

(Edgecombe *et al.* 2011; Kumar *et al.* 2011). Some ecdysozoan apomorphies have been evident since the inception of the group, relating to moulting of the external cuticle, mediated in all ecdysozoan phyla by ecdysteroids (Garey *et al.* 2001; but see Pilato *et al.* 2005 for a different opinion); while all phyla further lack locomotory cilia (Nielsen, 2001). These characteristics are some of the more striking features of Ecdysozoa, but to provide the most robust reconstruction of the ecdysozoan ancestor it is crucial to understand whether or not the worm like phyla comprising Cycloneuralia are a monophyletic or paraphyletic assemblage.

According to the majority of well-supported analyses presented in this thesis, from phylogenomics (Figure 3.3) and rRNA (Figure 5.2, 5.5), the paraphyletic origin of "Cycloneuralia" made up of Nematoida (Nematoda + Nematomorpha) sister to Panarthropoda (Nielsen, 2001), and Scalidophora (Priapulida, Kinorhyncha and Loricifera; sensu Schmidt-Rhaesa, 1996) as the sister group of nematoids plus Arthropoda is strongly supported. miRNAs are mute about this issue, but what is certain is that no miRNA characterising a monophyletic Cycloneuralia were found. Cycloneuralia has traditionally been regarded as a monophyletic group on the grounds of morphology (Ahlrichs, 1995) with all members sharing possession of collar-shaped circumesophageal brain (Nielsen, 2001). Similarly to the findings presented in this thesis, the recovery of paraphyletic Cycloneuralia is also recovered in some previous rRNA analyses (Garey *et al.* 2001 and Mallatt and Giribet, 2006). Importantly the analyses of Garey *et al.* (2001) and Mallatt and Giribet (2006) did not include data for Loricifera; which upon inclusion in analyses presented in Chapter 5 further where found to be member of the Scalidophora within the context of a paraphyletic "Cycloneuralia". In contrast to the paraphyletic origin supported in this thesis and by rRNA analyses mentioned above, the recent phylogenomic study of Dunn *et al.* (2008)

supported a monophyletic origin of Cycloneuralia. This is an interesting contradiction, as the phylogenomic analyses presented in Chapter 3 in support of paraphyly of Cycloneuralia are based on a sub sampling of genes from Dunn *et al.* (2008) while also having a larger taxon sampling for Tardigrada and Nematoda.

Accordingly, I conjecture that monophyletic Cycloneuralia in Dunn *et al.* (2008) might have been a tree reconstruction artifact perhaps resulting from low taxonomic sampling or the inclusion of fast evolving genes (many fast evolving genes from Dunn *et al.* 2008 where not included here). The overwhelming support provided in this thesis for the paraphyletic origin of the cycloneuralians, suggests that the ecdysozoan ancestor was cycloneuralian-like, with a collar shaped brain. Additionally, this organism could have possessed an introvert, as this is characteristic of all the Scalidophoran taxa. Considering the paraphyletic nature of Cycloneuralia, with Nematoida as the sister group to Panarthropoda, I suggest the name "Ambulavermia" for the still unnamed Nematoida plus Panarthropoda clade, a name that literally translates to "walking worm". With reference to the last common ecdysozoan ancestor, it has been suggested based on evidence from living and fossil ecdysozoans, that the predicted ancestral (plesiomorphic) characters of the Ecdysozoa are remarkably similar to those of extant Priapulida (Webster *et al.* 2006). Specifically ancestral characters such as an annulated, worm-like body, with a terminal mouth, proboscis, direct development, of macrofaunal body size, growth via ecdysis and finally a collar-shaped circumesophageal brain (Schmidt-Rhaesa, 1998; Budd, 2001). This depiction of the ecdysozoan ancestor is parsimonious when we consider the likely derived small size of the meiofaunal phyla Kinorhyncha and Loricifera in addition to the fact that priapulids are the only ecdysozoan phylum to have radial embryonic cleavage (Aguinaldo *et al.* 1997). In any case it is noted that Priapulids retain the largest proportion of plesiomorphic characters for Ecdysozoa compared to all other Introverta (Webster *et al.* 2006). It will be interesting to evaluate, once the priapulid genome is released, whether the priapulid worms are living fossils.

In a recent 2012 edition of the book "Animal evolution: Interrelationships of the living phyla", by Claus Nielsen, paraphyly of Cycloneuralia is not supported; instead the monophyletic origin is supported by the shared morphological feature of Nematoida + Scalidophora having a collar shaped brain with anterior and posterior rings of soma (neuron terminal cell body) separated by a ring of neuropile. According to our analyses, these results presented by Nielsen (2012) should be rejected. Within the Scalidophora, Nielsen suggests that the large priapulid worms are sister group to Kinorhyncha + Loricifera. This position for Priapulida is in disagreement with the basal branching position for Loricifera supported by the rRNA analyses in Chapter 5, and with previous rRNA analyses (Mallatt and Giribet, 2006; Sørensen *et al.* 2008) and phylogenomic analyses (Dunn *et al.* 2008; Hejnol *et al.* 2009) which support a sister group relationship of Priapulida + Kinorhyncha. However taxon sampling within Scalidophora in our and other studies is inadequate, leaving some doubts on the correct relationships among the Scalidophora. Indeed, the lack of resolution both within and between different molecular and morphological analyses has left the Scalidophora essentially as an unresolved trichotomy (Nielsen, 2012; but reference Nielsen, 2001). This trichotomy in Scalidophora calls for closer examination, with the potential of resolving these phyla residing in molecular analyses conducted with a much richer taxon sampling for Scalidophora; and eventually sequenced miRNA complements. This approach, one that I have advocated in this thesis is crucial to resolve problematic groups like Scalidophora; but is one that is met with a caveat, in that it is well known that small meiofaunal animals are difficult to obtain in the field, e.g. Loricifera being found in permanently anoxic conditions (Danovaro *et al.* 2010).

6.3 The nature of Panarthropoda and the rise of Lobopodia

Paraphyly of the Cycloneuralia, in relation to the last common ancestor of Ecdysozoa implies that this animal was an annulated, proboscis-bearing worm like organism with a collar-shaped brain. This has important implications for the evolution of the Panarthropoda. It is not surprising to note that the panarthropods thus represent the morphologically most divergent assemblage within Ecdysozoa, which must have evolved from a worm-like ancestor with a collar-shaped brain. Evolution from such an ancestor is supported by analysis of Eriksson and Budd (2000) in which they suggested that the onychophoran brain evolved from a circumesophageal ring by extending dorsal portions of the collar-shaped brain.

With respect to morphology, tardigrades have a melange of arthropod and cycloneuralian characters, suggesting that either the arthropod-like characters were lost in the cycloneuralians, or conversely the cycloneuralian-like characters were lost in the arthropods. According to the analyses presented in this thesis, which I consider to be robust corroborating evidence to support Tardigrada + Lobopodia, the arthropod-like features of tardigrades, such as the paired ventrolateral appendages with segmental leg nerves and *Engrailed* expression in the posterior ectoderm of each segment (Gabriel and Goldstein, 2007; Edgecombe, 2009) appear to be panarthropod apomorphies that are not present in Cycloneuralia.

I would like to draw attention to the small level of uncertainty for the placement of tardigrades, which stems from some analyses presented in this thesis supporting a tardigrade + onychophoran clade. This uncertainty can be diminished when we consider the results supporting this relationship were dependent on choice of model in both EST and rRNA (Figure 3.4.3; Figure 5.1) analyses, while also being reliant on taxon sampling (Figure 5.3) and so for the most part support was obtained from analyses that might have exacerbated phylogenetic artifacts. Overall, results in this thesis favour a clade composed of Tardigrada + Lobopodia, a finding bolstered greatly by the distribution of homoplasy-low miRNAs in Panarthropoda (Figure 4.8) providing accountability for the uniquely shared features of Onychophora + Arthropoda; features like an open hemocoelic circulatory system, dorsal heart with segmented ostia, nephridia forming from segmented mesoderm, without having to posit loss their secondary loss in Tardigrada due to miniaturization. Although Tardigrada + Onychophora has been recovered in previous molecular analyses of rRNA (Garey *et al.* 2001; Mallatt *et al.* 2004; Mallatt and Giribet, 2006) and multi-gene data sets (Rota-Stabelli *et al.* 2010) there are yet no commonly accepted morphological synapomorphies linking these taxa.

Contra to the large amount of molecular support for the monophyly of Onychophora + Arthropoda presented herein, previous morphological studies have suggested a sister group relationship of Tardigrada + Arthropoda. In these studies, this group was supported by shared features such as sclerotized cuticle, reduced numbers of nephridia (Wills *et al.* 1998) and segmental ganglia in the nerve cord – contrast to the unganglionated nerve cord in Onychophora (Whittington and Mayer, 2011). Interpreting these features in the face of significant support for Tardigrada + Lobopodia indicates that either convergent gain of segmental ganglia occurred in tardigrades and arthropods, or onychophorans developed a secondarily unsegmented nerve cord. The analyses presented in this thesis never found support for a sister group relationship of Tardigrada + Arthropoda ("Tactopoda"; sensu Budd, 2001), thus I fully reject this relationship in favour of Tardigrada + Lobopodia.

The findings presented in this thesis suggest that characters shared by tardigrades and cycloneuralians, such as a terminal mouth, protrusible mouth cone, triradiate pharynx, and a circumesophageal brain (Zantke *et al.* 2008; Edgecombe, 2010; Schmidt-Rhaesa, 1998) are most likely ecdysozoan plesiomorphies. This hypothesis is also consistent with the fossil record of arthropods, in that taxa in the arthropod stem group such as armoured lobopodians and anomalocaridids, show a melange of arthropod-like and cycloneuralian-like features, the latter (e.g. radially arranged mouthparts) then lost in the arthropod crown group (Edgecombe, 2010, Budd, 2001). Furthermore, my results suggest that paired limbs and a shared mode of segment patterning (Gabriel and Goldstein, 2007) are apomorphic for Panarthropoda. Regardless of the exact interrelationships of the three Panarthropod phyla, I have presented robust evidence throughout this thesis to support the monophyly of Panarthropoda. Carrying on from analyses of genomic data sets (ESTs, miRNAs, rRNA) morphology further provides unavoidable support for their monophyly, as all lineages have uniquely derived synapomorphies such as lateral walking appendages, segmented mesoderm, ventral nerve cords and a tripartite brain.

6.4 Potential role of miRNAs in the emergence of arthropod Bauplaene

In this thesis I have discussed the properties of, and presented the results of miRNA complements within arthropods and their close relatives to elucidate their phylogenetic relationships. However, the most prominent utilization of miRNAs is in the study of developmental regulation, prompting many to investigate their possible role of developmental canalization throughout metazoan evolution. Canalization refers to the process by which phenotypes are stabilized within species (Hornstein and Shomron, 2006). It is well known that arthropods are an incredibly diverse and specious group, but the degree of morphological disparity is one that is sometimes overlooked. One of the major questions in arthropod evolution regards the evolutionary developmental processes that led to what we refer to as "Arthropodization" and the endowment in arthropods of phenomenal environmental adaptability and diverse solutions to survival.

It was long thought that the rise in morphological complexity was one that was tightly correlated to that of an organism's repertoire of protein coding genes, but upon sequencing of complete genomes for model organisms like *C. elegans,* this was soon falsified, with morphologically simple roundworms having roughly the same number of PCGs as morphologically complex organisms (e.g. *Homo sapien*). So what other factors contributed to rising morphological complexity? Apart from the role of gene regulation, miRNAs are now beginning to be recognised for their dual role of developmental canalization over evolutionary time (Wu *et al.* 2009). miRNAs are crucial in gene regulatory networks, working in conjunction with typical regulatory network elements such as transcription factors. Yet, miRNAs also have specific attributes that allow them to not only regulate transcription, but also to reduce the overall 'genetic noise' in gene regulatory networks imparted by the stochasticity of transcription factors in the translational process (Hornstein and Shomron, 2006).

MicroRNAs are being continuously added to, and conserved within genomes throughout evolutionary time, a fact that is largely unique with respect to transcription factors. When considering metazoan development, it is important to note that all metazoan transcription factors are present and conserved throughout all Metazoa (Wheeler *et al.* 2009). This is in contrast to miRNAs, which have been shown to be largely lineage specific, for instance the Bilateria and Deuterostomia had a massive burst of miRNA expansion compared to that of early branching metazoans (Sempere *et al.* 2007; Campo-Paysaa *et al.* 2011). Similarly to expansion of miRNAs through time, it is now recognised that morphological complexity is intimately linked to expansion of novel cell types (Valentine *et al.* 1994) with miRNAs known to play a key role in cell regulation and differentiation (Ambros, 2004). This suggests that miRNAs must be intimately tied to the evolution of novel cell types and therefore morphological complexity (Heimberg *et al.* 2008; Wheeler *et al.* 2009).

Although questions regarding the appearance of lineage specific miRNAs and their correlation with the rise in morphological complexity are ones outside the scope of this thesis, I would like to briefly consider the emergence of arthropod specific miRNAs (Figure 4.8) and their potential role in the emergence of the many diverse arthropod Bauplaene. From results of analyses into panarthropod miRNAs, I showed that there are two miRNAs (*miR*-275 and *iab*-4) conserved throughout Arthropoda. *Iab*-4 as an interesting example, as this miRNA has been shown to be intimately linked to the regulation of developmentally important HOX transcription factors (*Abd*-A, *Ubx* and *Antp*; Enright *et al.* 2004; Miura *et al.* 2012). In the recent study of Miura *et al.* (2012), the miRNA *iab*-4 is shown to have an incredibly high conservation of its seed region throughout the ~400 MYA evolutionary period since the last common ancestor of *Drosophila* and *Daphnia*. Interestingly, the number of target sites for *iab*-4 in the HOX genes *Abd*-A, *Ubx* and *Antp* varied considerably across the Arthropoda (see Miura *et al.* 2012; Table 1).

Given that HOX genes are particularly important in the developmental process, differential expression of these genes brought about by acquisitions or changes in the number of *iab*-4 target sites across Arthropoda, may have to some degree driven changes in morphological evolution. Evidence then at least suggests that the emergence of arthropod specific *iab*-4 could have been pivotal in the canalization of developmental segmentation, and might have played a role in the evolution of the complex appendages (e.g. walking legs) observed in Arthropoda but not found in Onychophora and Tardigrada which have much simpler walking appendages. This conclusion is of course speculative and well outside the scope of this work, but at least hints at the possible role of novel arthropod specific miRNAs (already known to be key players in canalization) and the emergence of and construction of the most successful of all animal body plans, that of the arthropods.

6.5 Closing remarks

The arthropods and to a lesser extent their closest living relatives the Onychophora and the Tardigrada, are an excellent example of the emergence of a group of animals that have come to dominate animal diversity. There is no doubt that the arthropods alone represent the phylum with the greatest number of living species, yet these species only represent the surviving branches of a long history of diverse extinct forms. Notwithstanding the evidence presented in this thesis in relation to the interrelationships of Arthropoda and the remaining panarthropods, it is clear that the diversity of species leading to the panarthropod groups began deep in geological history. Recent estimates on the emergence of stem lineages leading to extant forms of

Arthropods and their relatives Onychophora dates the timing of origin to be between ~593 and 534 million years ago (Erwin *et al.* 2011). A period commonly referred to as the 'Cambrian explosion' falls within this interval, and is an era of animal evolution marked by explosive and abrupt appearances of taxa and subsequent rapid diversification. Consequently, much of the evolutionary information has been eroded by millions of years of mutational saturation, exquisitely highlighted in Tardigrada with their long branches. Yet despite this, in the past 20 years or so, the availability of ever larger molecular data sets to analyse using phylogenetics, and the development of sophisticated models of evolution has lead, in recent years, to significant insights into the evolution of this diverse group of animals.

In chapter 2 (page 35) of the book '*Arthropod fossils and phylogeny'* (Edgecombe, 1998) which was published relatively recently, a list of six "principle issues in arthropod evolution" were outlined to promote further investigations in elucidating the evolution of arthropods and their relatives. I will not recall all of these, as some are not relatable to the work presented in this thesis. Outstanding questions at the time: (1) *Whether the crustaceans and tracheates* (Hexapoda + Myriapoda) *form a clade (Mandibulata);* (2) *Where the Onychophora and Tardigrada lie with respect to the tracheates and the rest of the euarthropods in general;* (3) *Whether the euarthropods arose once from a single soft-bodied ancestor that was itself an arthropod, or whether two or more events occurred;* and (4) *Whether the tracheates are monophyletic, or the myriapods branched off lower in the phylogeny.*

The work presented in this thesis (summarized in Figure 6.1) I feel has significantly improved the overall resolution of arthropod/panarthropod evolution, specifically addressing these questions with corroborating evidence of phylogenomics, miRNA distributions and classical rRNA molecular data. According to my results, Mandibulata is a true clade composed of Pancrustacea + Myriapoda, sister group to Chelicerata; all arising within a single monophyletic origin of Arthropoda, or '*Euarthropoda'* from within Panarthropoda "Arthropoda" in Edgecombe (1998) terminology. Further more, according to all data types considered in the analyses here presented, supports the monophyly of Panarthropoda, with arthropods being the sister group to Onychophora, and Tardigrada sister group to Lobopodia.

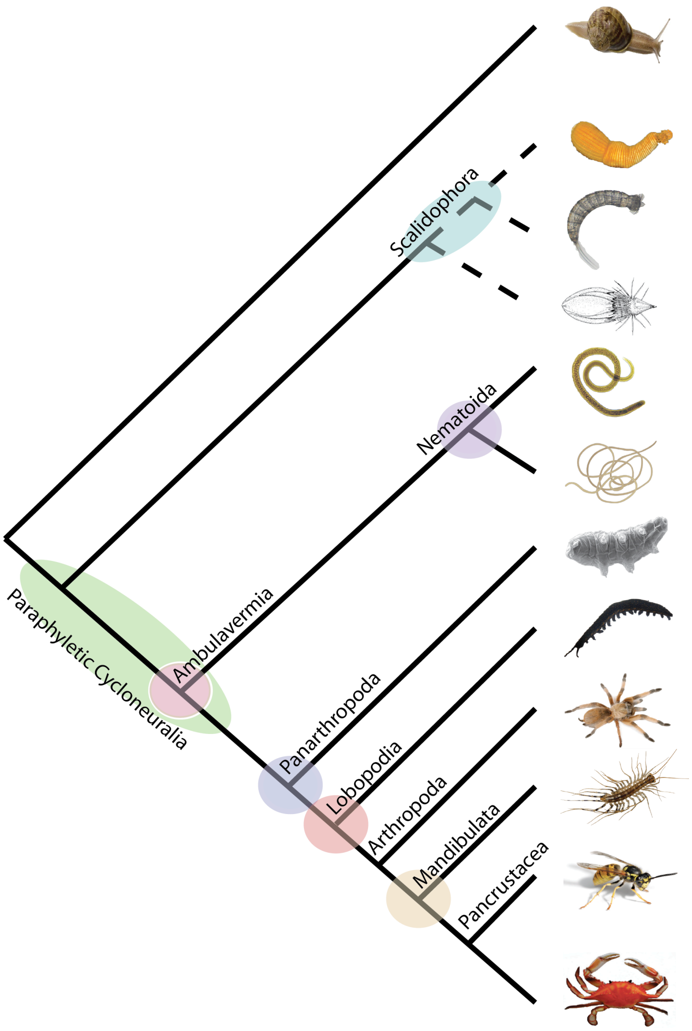


Figure 6.1: Summary of major hypotheses addressed in this thesis. Ecdysozoa composed of a paraphyletic Cycloneuralia (green oval) and monophyletic Panarthropoda (blue circle). Scalidophora (light blue circle) is sister to a clade (here named "Ambulavermia" (pink circle)) of Nematoida (purple circle) plus Panarthropoda. Within the Panarthropoda, Tardigrada is sister group to Lobopodia (red circle), while Arthropoda is made up of Chelicerata sister group to Mandibulata (yellow circle).

Lastly, and in addition to the points presented by Edgecombe (1998) the results here presented suggest that the Cycloneuralia are a paraphyletic assemblage, with Panarthropods evolving from a cycloneuralian, worm-like ancestor. However, the hypotheses of cycloneuralian monophyly vs. paraphyly are ones that necessitate further investigation; particularly from increased taxon sampling of crucial scalidophoran phyla such as Kinorhyncha and Loricifera. I would like to state here that the findings related to monophyly of Mandibulata, Lobopodia and Panarthropoda should be considered robust, specifically in light of the experiments performed throughout this thesis to uncover instances of systematic bias (e.g. Tardigrada + Nematoda, Myriochelata) in addition to striking morphological synapomorphies and the recovery of clade specific miRNAs characterized by low levels of homoplasy.

I would like to conclude by saying, I sincerely hope the phylogenetic methods and hypotheses presented in this thesis, will improve the understanding of arthropod, ecdysozoan and animal evolution more broadly.

Chapter 7

Bibliography

AGUINALDO, A. M. A., TURBEVILLE, J. M., LINFORD, L. S., RIVERA, M. C., GAREY, J. R., RAFF, R. A. & LAKE, J. A. (1997). Evidence for a clade of nematodes, arthropods and other moulting animals. Nature, 387, 489-493.

AHLRICHS, W. (1995). Ultrastruktur und Phylogenie von Seison nebaliae (Grube 1859) und Seison annulatus (Claus 1876): Hypothesen zu phylogenetischen Verwandtschaftsverhaltnissen innerhalb der Bilateria, Cuvillier.

AMBROS, V., BARTEL, B., BARTEL, D. P., BURGE, C. B., CARRINGTON, J. C., CHEN, X., DREYFUSS, G., EDDY, S. R., GRIFFITHS-JONES, S. & MARSHALL, M. (2003). A uniform system for microRNA annotation. Rna, 9, 277-279.

ANDERSON, D. (1979). Embryos, fate maps, and the phylogeny of arthropods. Arthropod Phylogeny. Van Nostrand Reinhold, New York, 59-106.

ANDERSON, D. T. (1973). Embryology and phylogeny in annelids and arthropods, Pergamon Press Oxford.

ANDREW, D. R. (2011). A New View of Insect-Crustacean Relationships II. Inferences from Expressed Sequence Tags and Comparisons with Neural Cladistics. Arthropod Structure & Development.

AVEROF, M. & AKAM, M. (1995). Insect-crustacean relationships: insights from comparative developmental and molecular studies. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 347, 293.

BAGUNA, J. & RIUTORT, M. (2004). The dawn of bilaterian animals: the case of acoelomorph flatworms. Bioessays, 26, 1046-1057.

BALLARD, J., OLSEN, G. J., FAITH, D. P., ODGERS, W. A., ROWELL, D. M. & ATKINSON, P. W. (1992). Evidence from 12S ribosomal RNA sequences that onychophorans are modified arthropods. Science, 258, 1345.

BANTOUNAS, I., PHYLACTOU, L. & UNEY, J. (2004). RNA interference and the use of small interfering RNA to study gene function in mammalian systems. Journal of molecular endocrinology, 33, 545-557.

BAPTESTE, E., BRINKMANN, H., LEE, J. A., MOORE, D. V., SENSEN, C. W., GORDON, P., DURUFLÉ, L., GAASTERLAND, T., LOPEZ, P. & MÜLLER, M. (2002). The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. Proceedings of the National Academy of Sciences, 99, 1414.

BARTEL, D. P. (2004). MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. Cell, 116, 281-297.

BARTEL, D. P. (2009). MicroRNAs: target recognition and regulatory functions. Cell, 136, 215-233.

BEREZIKOV, E. (2011). Evolution of microRNA diversity and regulation in animals. Nature Reviews Genetics, 12, 846-860.

BEREZIKOV, E., CUPPEN, E. & PLASTERK, R. H. A. (2006). Approaches to microRNA discovery. Nature genetics, 38, S2-S7.

BITSCH, C. & BITSCH, J. (2004). Phylogenetic relationships of basal hexapods among the mandibulate arthropods: a cladistic analysis based on comparative morphological characters. Zoologica Scripta, 33, 511-550.

BLAIR HEDGES, S. & KUMAR, S. (2004). Precision of molecular time estimates. TRENDS in Genetics, 20, 242-247.

BLAIR, J., IKEO, K., GOJOBORI, T. & HEDGES, S. B. (2002). The evolutionary position of nematodes. BMC Evolutionary Biology, 2, 7.

BLANQUART, S. & LARTILLOT, N. (2008). A site-and time-heterogeneous model of amino acid replacement. Molecular biology and evolution, 25, 842.

BOORE, J. L., COLLINS, T. M., STANTON, D., DAEHLER, L. L. & BROWN, W. M. (1995). Deducing the pattern of arthropod phylogeny from mitochondrial-DNA rearrangements. Nature, 376, 163-165.

BOORE, J. L., LAVROV, D. V. & BROWN, W. M. (1998). Gene translocation links insects and crustaceans. Nature, 392, 667-668.

BORCHERT, G. M., LANIER, W. & DAVIDSON, B. L. (2006). RNA polymerase III transcribes human microRNAs. Nature structural & molecular biology, 13, 1097-1101.

BOSSÉ, G. D. & SIMARD, M. J. (2010). A new twist in the microRNA pathway: Not Dicer but Argonaute is required for a microRNA production. Cell Research, 20, 735-737.

BOURLAT, S. J., JULIUSDOTTIR, T., LOWE, C. J., FREEMAN, R., ARONOWICZ, J., KIRSCHNER, M., LANDER, E. S., THORNDYKE, M., NAKANO, H. & KOHN, A. B. (2006). Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. Nature, 444, 85-88.

BOURLAT, S. J., NIELSEN, C., ECONOMOU, A. D. & TELFORD, M. J. (2008). Testing the new animal phylogeny: a phylum level molecular analysis of the animal kingdom. Molecular Phylogenetics and Evolution, 49, 23-31.

BRABAND, A., CAMERON, S. L., PODSIADLOWSKI, L., DANIELS, S. R. & MAYER, G. (2010). The mitochondrial genome of the onychophoran Opisthopatus cinctipes (Peripatopsidae) reflects the ancestral mitochondrial gene arrangement of Panarthropoda and Ecdysozoa. Molecular Phylogenetics and Evolution, 57, 285-292.

BRENNECKE, J., STARK, A., RUSSELL, R. B. & COHEN, S. M. (2005). Principles of microRNA target recognition. PLoS biology, 3, e85.

BRINKMANN, H. & PHILIPPE, H. (1999). Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Molecular biology and evolution, 16, 817.

BROCHIER-ARMANET, C., FORTERRE, P. & GRIBALDO, S. (2011). Phylogeny and evolution of the Archaea: one hundred genomes later. Current opinion in microbiology.

BRODERSEN, P. & VOINNET, O. (2009). Revisiting the principles of microRNA target recognition and mode of action. Nature Reviews Molecular Cell Biology, 10, 141-148.

BRUNO, W. J. & HALPERN, A. L. (1999). Topological bias and inconsistency of maximum likelihood using wrong models. Molecular biology and evolution, 16, 564-566.

BRUSCA, R. C. (2000). Unravelling the history of arthropod biodiversification. Annals of the Missouri Botanical Garden, 13-25.

BUDD, G. E. (2001). Tardigrades as Stem-Group Arthropods: The Evidence from the Cambrian Fauna. Zoologischer Anzeiger-A Journal of Comparative Zoology, 240, 265-279.

BUDD, G. E. & TELFORD, M. J. (2009). The origin and evolution of arthropods. Nature, 457, 812-817.

BULL, J., HUELSENBECK, J. P., CUNNINGHAM, C. W., SWOFFORD, D. L. & WADDELL, P. J. (1993). Partitioning and combining data in phylogenetic analysis. Systematic Biology, 42, 384-397.

CAMERON, S. L., MILLER, K. B., D'HAESE, C. A., WHITING, M. F. & BARKER, S. C. (2004). Mitochondrial genome data alone are not enough to unambiguously resolve the relationships of Entognatha, Insecta and Crustacea sensu lato (Arthropoda). Cladistics, 20, 534-557.

CAMIN, J. & SOKAL, R. (1965). A method for deducing branching sequences in

phylogeny. Evolution, 19, 311-326.

CAMPBELL, L. I., ROTA-STABELLI, O., EDGECOMBE, G. D., MARCHIORO, T., LONGHORN, S. J., TELFORD, M. J., PHILIPPE, H., REBECCHI, L., PETERSON, K. J. & PISANI, D. (2011). MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda. Proceedings of the National Academy of Sciences, 108, 15920-15924.

CAMPO PAYSAA, F., SÉMON, M., CAMERON, R. A., PETERSON, K. J. & SCHUBERT, M. (2011). microRNA complements in deuterostomes: origin and evolution of microRNAs. Evolution & Development, 13, 15-27.

CHAPMAN, A. D., SERVICES, A. B. I. & STUDY, A. B. R. (2009). Numbers of living species in Australia and the world. 2nd ed. Australian Biological Resources Study, Canberra.

CHELOUFI, S., DOS SANTOS, C. O., CHONG, M. M. W. & HANNON, G. J. (2010). A dicer-independent miRNA biogenesis pathway that requires Ago catalysis. Nature, 465, 584-589.

CHRISTODOULOU, F., RAIBLE, F., TOMER, R., SIMAKOV, O., TRACHANA, K., KLAUS, S., SNYMAN, H., HANNON, G. J., BORK, P. & ARENDT, D. (2010). Ancient animal microRNAs and the evolution of tissue identity. Nature, 463, 1084-1088.

CICCARELLI, F. D., DOERKS, T., VON MERING, C., CREEVEY, C. J., SNEL, B. & BORK, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. Science, 311, 1283.

CISNE, J. L. (1974). Trilobites and the origin of arthropods. Science, 186, 13-18.

CONWAY MORRIS, S. (1977). A new metazoan from the Cambrian Burgess Shale of British Columbia. Palaeontology, 20, 623-40.

COOK, C. E., SMITH, M. L., TELFORD, M. J., BASTIANELLO, A. & AKAM, M. (2001). Hox genes and the phylogeny of the arthropods. Current Biology, 11, 759-763.

COPLEY, R. R., ALOY, P., RUSSELL, R. B. & TELFORD, M. J. (2004). Systematic searches for molecular synapomorphies in model metazoan genomes give some support for Ecdysozoa after accounting for the idiosyncrasies of Caenorhabditis elegans. Evolution & Development, 6, 164-169.

COX, C. J., FOSTER, P. G., HIRT, R. P., HARRIS, S. R. & EMBLEY, T. M. (2008). The archaebacterial origin of eukaryotes. Proceedings of the National Academy of Sciences, 105, 20356.

CRAMPTON, G. (1921). The Phylogenetic Origin of the Mandibles of Insects and Their Arthropodan Relatives: A Contribution to the Study of the Evolution of the Arthropoda. Journal of the New York Entomological Society, 29, 63-100.

CUMMINS, C. A. & MCINERNEY, J. O. (2011). A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. Systematic Biology, 60, 833-844.

DANOVARO, R., DELL'ANNO, A., PUSCEDDU, A., GAMBI, C., HEINER, I. & KRISTENSEN, R. M. (2010). The first metazoa living in permanently anoxic conditions. BMC biology, 8, 30.

DARWIN, C. (1859). On the origin of the species by natural selection. London, John Murray.

DAYHOFF, M. O. & ECK, R. V. (1968). Atlas of protein sequence and structure 1967-(1968). National Biomedical Research Foundation, Silver Spring, Maryland.

DE QUEIROZ, A. & GATESY, J. (2007). The supermatrix approach to Systematics. Trends in Ecology & Evolution, 22, 34-41.

DE ROSA, R., GRENIER, J. K., ANDREEVA, T., COOK, C. E., ADOUTTE, A., AKAM, M., CARROLL, S. B. & BALAVOINE, G. (1999). Hox genes in brachiopods and priapulids and protostome evolution. Nature, 399, 772-776.

DE WIT, E., LINSEN, S. E. V., CUPPEN, E. & BEREZIKOV, E. (2009). Repertoire and evolution of miRNA genes in four divergent nematode species. Genome research, 19, 2064-2074.

DELSUC, F., BRINKMANN, H. & PHILIPPE, H. (2005). Phylogenomics and the reconstruction of the tree of life. Nature reviews. Genetics, 6, 361.

DOHLE, W. (1997). Are the insects more closely related to the crustaceans than to the myriapods? ENTOMOLOGICA SCANDINAVICA SUPPLEMENTUM, 7-16.

DOHLE, W. Year. Are the insects terrestrial crustaceans? A discussion of some new facts and arguments and the proposal of the proper name 'Tetraconata' for the monophyletic unit Crustacea+ Hexapoda. In, (2001). Société entomologique de France, 85-103.

DOHRMANN, M., VOIGT, O., ERPENBECK, D. & WORHEIDE, G. (2006). Non-monophyly of most supraspecific taxa of calcareous sponges (Porifera, Calcarea) revealed by increased taxon sampling and partitioned Bayesian analysis of ribosomal DNA. Molecular Phylogenetics and Evolution, 40, 830-843.

DONOGHUE, M. J., DOYLE, J. A., GAUTHIER, J., KLUGE, A. G. & ROWE, T. (1989). The importance of fossils in phylogeny reconstruction. Annual Review of Ecology and Systematics, 20, 431-460.

DOPAZO, H. & DOPAZO, J. (2005). Genome-scale evidence of the nematode-arthropod clade. Genome Biology, 6, R41.

DOUADY, C. J., DELSUC, F., BOUCHER, Y., DOOLITTLE, W. F. & DOUZERY, E. J. P. (2003). Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Molecular biology and evolution, 20, 248.

DOVE, H. & STOLLEWERK, A. (2003). Comparative analysis of neurogenesis in the myriapod Glomeris marginata (Diplopoda) suggests more similarities to chelicerates than to insects. Development, 130, 2161.

DUNLOP, J. A., WIRTH, S., PENNEY, D., MCNEIL, A., BRADLEY, R. S., WITHERS, P. J. & PREZIOSI, R. F. (2011). A minute fossil phoretic mite recovered by phase-contrast X-ray computed tomography. Biology Letters.

DUNN, C. W., HEJNOL, A., MATUS, D. Q., PANG, K., BROWNE, W. E., SMITH, S. A., SEAVER, E., ROUSE, G. W., OBST, M. & EDGECOMBE, G. D. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. Nature, 452, 745-749.

EDGAR, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research, 32, 1792.

EDGAR, R. C. & BATZOGLOU, S. (2006). Multiple sequence alignment. Current opinion in structural biology, 16, 368-373.

EDGECOMBE, G. D. (1998). Arthropod fossils and phylogeny, Columbia Univ Pr.

EDGECOMBE, G. D. (2004). Morphological data, extant Myriapoda, and the myriapod stem-group. Contributions to Zoology, 73, 207-252.

EDGECOMBE, G. D. (2009). Palaeontological and molecular evidence linking arthropods, onychophorans, and other Ecdysozoa. Evolution: Education and Outreach, 2, 178-190.

EDGECOMBE, G. D. (2010). Arthropod phylogeny: an overview from the perspectives of morphology, molecular data and the fossil record. Arthropod Structure & Development, 39, 74-87.

EDGECOMBE, G. D., WILSON, G. D. F., COLGAN, D. J., GRAY, M. R. & CASSIS, G. (2000). Arthropod cladistics: combined analysis of histone H3 and U2 snRNA sequences and morphology. Cladistics, 16, 155-203.

EDWARDS, A. W. F. & CAVALLI-SFORZA, L. L. (1963). The reconstruction of evolution. Annals of Human genetics, 27, 105-106.

EDWARDS, A. W. F. & CAVALLI-SFORZA, L. L. (1964). Reconstruction of evolutionary trees. Phenetic and Phylogenetic classification, ed. V.H. Haywood and J. McNeill, Systematics Association Publ. No. 6, London, 67-76.

EERNISSE, D. J., ALBERT, J. S. & ANDERSON, F. E. (1992). Annelida and Arthropoda are not sister taxa: a phylogenetic analysis of spiralian metazoan morphology. Systematic Biology, 41, 305.

EFRON, B. (1979). Bootstrap methods: another look at the jackknife. The annals of Statistics, 7, 1-26.

EISEN, J. A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome research, 8, 163.

ENRIGHT, A. J., JOHN, B., GAUL, U., TUSCHL, T., SANDER, C. & MARKS, D. S. (2004). MicroRNA targets in Drosophila. Genome Biology, 5, 1-1.

ERIXON, P., SVENNBLAD, B., BRITTON, T. & OXELMAN, B. (2003). Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. Systematic Biology, 52, 665.

ERWIN, D. H., LAFLAMME, M., TWEEDT, S. M., SPERLING, E. A., PISANI, D. & PETERSON, K. J. (2011). The Cambrian conundrum: Early divergence and later ecological success in the early history of animals. Science, 334, 1091-1097.

ERWIN, T. L. (1982). Tropical forests: their richness in Coleoptera and other arthropod species. Coleopterists Bulletin, 36, 74,Äì75.

ERWIN, T. L. (1988). The tropical forest canopy: the heart of biotic diversity. in E. O. Wilson and F. M. Peter, eds. Biodiversity, National Academy, Washington, DC., 123 - 129.

FELSENSTEIN, J. (1973). Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Systematic Zoology, 240-249.

FELSENSTEIN, J. (1978a). Cases in which parsimony or compatibility methods will be positively misleading. Systematic Biology, 27, 401.

FELSENSTEIN, J. (1978b). The number of evolutionary trees. Systematic Biology, 27, 27.

FELSENSTEIN, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of molecular evolution, 17, 368-376.

FELSENSTEIN, J. (1985). Phylogenies and the comparative method. American Naturalist, 1-15.

FELSENSTEIN, J. (2004). Inferring phylogenies. Sunderland, Massachusetts: Sinauer Associates.

FIELD, K. G., OLSEN, G. J., LANE, D. J., GIOVANNONI, S. J., GHISELIN, M. T., RAFF, E. C., PACE, N. R. & RAFF, R. A. (1988). Molecular phylogeny of the animal kingdom. Science, 239, 748.

FILIPOWICZ, W., BHATTACHARYYA, S. N. & SONENBERG, N. (2008). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nature Reviews Genetics, 9, 102-114.

FISHER, R. A. (1912). On an Absolute Criterion for Fitting Frequency Curves.

FISHER, R. A. (1921). On the" Probable Error" of a Coefficient of Correlation Deduced from a Small Sample. Metron, 1, 3-32.

FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 222, 309-368.

FITCH, W. M. (2000). Homology: a personal view on some of the problems. TRENDS in Genetics, 16, 227-231.

FLEISCHMANN, R. D., ADAMS, M. D., WHITE, O., CLAYTON, R. A., KIRKNESS, E. F., KERLAVAGE, A. R., BULT, C. J., TOMB, J. F., DOUGHERTY, B. A. & MERRICK, J. M. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science, 269, 496.

FORMAN, J. J., LEGESSE-MILLER, A. & COLLER, H. A. (2008). A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. Proceedings of the National Academy of Sciences, 105, 14879.

FOSTER, P. G. (2004). Modelling compositional heterogeneity. Systematic Biology, 53, 485.

FOSTER, P. G. & HICKEY, D. A. (1999). Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. Journal of molecular evolution, 48, 284-290.

FOSTER, P. G., JERMIIN, L. S. & HICKEY, D. A. (1997). Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. Journal of molecular evolution, 44, 282-288.

FOX, G. E. (2010). Origin and Evolution of the Ribosome. Cold Spring Harbor Perspectives in Biology, 2.

FRIEDRICH, M. & TAUTZ, D. (1995). Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods. Nature, 376, 165-167.

GABRIEL, W. N. & GOLDSTEIN, B. (2007). Segmental expression of Pax3/7 and Engrailed homologs in tardigrade development. Development Genes and Evolution, 217, 421-433.

GAI, Y. H., SONG, D. X., SUN, H. Y. & ZHOU, K. Y. (2006). Myriapod monophyly and relationships among myriapod classes based on nearly complete 28S and 18S rDNA sequences. Zoological Science, 23, 1101-1108.

GAREY, D. R., NELSON, L. Y. & MACKEY, J. L. (1999). Tardigrade phylogeny: congruency of morphological and molecular evidence. Zool. Anz., 238, 205 - 210.

GAREY, J. R. (2001). Ecdysozoa: the relationship between Cycloneuralia and Panarthropoda. Zoologischer Anzeiger-A Journal of Comparative Zoology, 240, 321-330.

GAREY, J. R., KROTEC, M., NELSON, D. R. & BROOKS, J. (1996). Molecular analysis supports a tardigrade-arthropod association. Invertebrate Biology, 79-88.

GAUT, B. S. & LEWIS, P. O. (1995). Success of maximum likelihood phylogeny inference in the four-taxon case. Molecular biology and evolution, 12, 152.

GIRIBET, G., CARRANZA, S., BAGUNA, J., RIUTORT, M. & RIBERA, C. (1996). First molecular evidence for the existence of a Tardigrada+ Arthropoda clade. Molecular biology and evolution, 13, 76.

GIRIBET, G., EDGECOMBE, G. D. & WHEELER, W. C. (2001). Arthropod phylogeny based on eight molecular loci and morphology. Nature, 413, 157-161.

GIRIBET, G. & RIBERA, C. (1998). The Position of Arthropods in the Animal Kingdom: A Search for a Reliable Outgroup for Internal Arthropod Phylogeny. Molecular Phylogenetics and Evolution, 9, 481-488.

GIRIBET, G. & RIBERA, C. (2000). A review of arthropod phylogeny: new data based on ribosomal DNA sequences and direct character optimization. Cladistics, 16, 204-231.

GIRIBET, G., RICHTER, S., GREGORY, D. E. & WHEELER, W. C. (2005). The position of crustaceans within Arthropoda-Evidence from nine molecular loci and morphology. In: Koenemann S, Jenner R (eds) Crustacea and arthropod relationships. , 307-352.

GIRIBET, G. & WHEELER, W. C. (1999). The Position of Arthropods in the Animal Kingdom: Ecdysozoa, Islands, Trees, and the" Parsimony Ratchet". Molecular Phylogenetics and Evolution, 13, 619-623.

GOLDMAN, N. (1998). Phylogenetic information and experimental design in molecular systematics. Proceedings of the Royal Society of London. Series B: Biological Sciences, 265, 1779.

GRBI , M., VAN LEEUWEN, T., CLARK, R. M., ROMBAUTS, S., ROUZÉ, P., GRBI , V., OSBORNE, E. J., DERMAUW, W., NGOC, P. C. T. & ORTEGO, F. (2011). The genome of Tetranychus urticae reveals herbivorous pest adaptations. Nature, 479, 487-492.

GRIMSON, A., SRIVASTAVA, M., FAHEY, B., WOODCROFT, B. J., CHIANG, H. R., KING, N., DEGNAN, B. M., ROKHSAR, D. S. & BARTEL, D. P. (2008). Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature, 455, 1193-1197.

HAASE, A., STERN, M., W,Ä∞CHTLER, K. & BICKER, G. (2001). A tissue-specific marker of Ecdysozoa. Development Genes and Evolution, 211, 428-433.

HALANYCH, K. M. (2004). The new view of animal phylogeny. Annual Review of Ecology, Evolution, and Systematics, 229-256.

HAMILTON, A. J., BASSET, Y., BENKE, K. K., GRIMBACHER, P. S., MILLER, S. E., NOVOTY, V., SAMUELSON, G. A., STORK, N. E., WEIBLEN, G. D. & YEN, J. D. L. (2010). Quantifying uncertainty in estimation of tropical arthropod species richness. The American Naturalist, 176, 90-95.

HARZSCH, S. (2004). Phylogenetic comparison of serotonin immunoreactive neurons in representatives of the Chilopoda, Diplopoda, and Chelicerata: Implications for arthropod relationships. Journal of Morphology, 259, 198-213.

HARZSCH, S. & HAFNER, G. (2006). Evolution of eye development in arthropods: Phylogenetic aspects. Arthropod Structure & Development, 35, 319-340.

HASEGAWA, M., HASHIMOTO, T., ADACHI, J., IWABE, N. & MIYATA, T. (1993). Early branchings in the evolution of eukaryotes: ancient divergence of entamoeba that lacks mitochondria revealed by protein sequence data. Journal of molecular evolution, 36, 380-388.

HASEGAWA, M., KISHINO, H. & SAITOU, N. (1991). On the maximum likelihood method in molecular phylogenetics. Journal of molecular evolution, 32, 443-445.

HASSANIN, A. (2006). Phylogeny of Arthropoda inferred from mitochondrial sequences: strategies for limiting the misleading effects of multiple changes in pattern and rates of substitution. Molecular Phylogenetics and Evolution, 38, 100-116.

HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57, 97.

HAWKSWORTH, D. & KALIN-ARROYO, M. (1995). Magnitude and distribution of biodiversity. Global biodiversity assessment, 107-191.

HEIMBERG, A. M. (2010). microRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrate. Proceedings of the National Academy of Sciences, 107, 19379.

HEJNOL, A., OBST, M., STAMATAKIS, A., OTT, M., ROUSE, G. W., EDGECOMBE, G. D., MARTINEZ, P., BAGUÑÀ, J., BAILLY, X. & JONDELIUS, U. (2009). Assessing the root of bilaterian animals with scalable phylogenomic methods. Proceedings of the Royal Society B: Biological Sciences, 276, 4261.

HENDY, M. D. & PENNY, D. (1989). A framework for the quantitative study of evolutionary trees. Systematic Biology, 38, 297-309.

HENNIG, W. (1950). Grundzüge einer theorie der phylogenetischen systematik. Deutscher Zentralverlag, Berlin.

HENNIG, W. (1965). Phylogenetic Systematics. Annual Review of Entomology, 10, 97-116.

HERTEL, J., LINDEMEYER, M., MISSAL, K., FRIED, C., TANZER, A., FLAMM, C., HOFACKER, I. & STADLER, P. (2006). The expansion of the metazoan microRNA repertoire. BMC genomics, 7, 25.

HILLIS, D. M. (1998). Taxonomic sampling, phylogenetic accuracy, and investigator bias. Systematic Biology, 47, 3-8.

HILLIS, D. M. & BULL, J. J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Systematic Biology, 42, 182.

HOLTON, T. A. & PISANI, D. (2010). Deep genomic-scale analyses of the metazoa reject Coelomata: evidence from single-and multigene families analyzed under a supertree and supermatrix paradigm. Genome Biology and Evolution, 2, 310.

HORNSTEIN, E. & SHOMRON, N. (2006). Canalization of development by microRNAs. Nature genetics, 38, S20-S24.

HUELSENBECK, J. P. (1995). The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbour joining. Molecular biology and evolution, 12, 843-849.

HUELSENBECK, J. P. & BOLLBACK, J. P. (2001). Empirical and hierarchical Bayesian estimation of ancestral states. Systematic Biology, 50, 351.

HUELSENBECK, J. P. & RANNALA, B. (2004). Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Systematic Biology, 53, 904.

HUELSENBECK, J. P., RONQUIST, F., NIELSEN, R. & BOLLBACK, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. Science, 294, 2310.

HWANG, U. W., FRIEDRICH, M., TAUTZ, D., PARK, C. J. & KIM, W. (2001). Mitochondrial protein phylogeny joins myriapods with chelicerates. Nature, 413, 154-157.

HYMAN, L. 1940. The invertebrates: Protozoa through Ctenophora. New York and London: McGraw-Hill.

HYMAN, L. H. (1951). The Invertebrates: The Acoelomate Bilateria. Platyhelminthes and Rhynchocoela, McGraw-Hill.

INAGAKI, Y., SUSKO, E., FAST, N. M. & ROGER, A. J. (2004). Covarion shifts cause a long-branch attraction artifact that unites Microsporidia and Archaebacteria in EF-1 phylogenies. Molecular biology and evolution, 21, 1340.

JEFFROY, O., BRINKMANN, H., DELSUC, F. & PHILIPPE, H. (2006). Phylogenomics: the beginning of incongruence? TRENDS in Genetics, 22, 225-231.

JENNER, R. A. (2004). Accepting partnership by submission? Morphological phylogenetics in a molecular millennium. Systematic Biology, 53, 333.

JENNER, R. A. & SCHRAM, F. R. (1999). The grand game of metazoan phylogeny: rules and strategies. Biological Reviews, 74, 121-142.

JERMIIN, L. S., HO, S. Y. W., ABABNEH, F., ROBINSON, J. & LARKUM, A. W. D. (2004). The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Systematic Biology, 53, 638.

JONES, D. T., TAYLOR, W. R. & THORNTON, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. Computer applications in the biosciences: CABIOS, 8, 275-282.

JUKES, T. H. & CANTOR, C. R. (1969). Evolution of protein molecules. IN MUNRO, H. N.

(Ed.) Mammalian Protein Metabolism. New York, Academic Press.

KADNER, D. & STOLLEWERK, A. (2004). Neurogenesis in the chilopod Lithobius forficatus suggests more similarities to chelicerates than to insects. Development Genes and Evolution, 214, 367-379.

KEANE, T., CREEVEY, C., PENTONY, M., NAUGHTON, T. & MCLNERNEY, J. (2006). Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evolutionary Biology, 6, 29.

KELCHNER, S. A. & THOMAS, M. A. (2007). Model use in phylogenetics: nine key questions. Trends in Ecology & Evolution, 22, 87-94.

KHVOROVA, A., REYNOLDS, A. & JAYASENA, S. D. (2003). Functional siRNAs and miRNAs exhibit strand bias. Cell, 115, 209-216.

KIM, J. (1996). General inconsistency conditions for maximum parsimony: Effects of branch lengths and increasing numbers of taxa. Systematic Biology, 45, 363.

KIM, V. N. (2005). MicroRNA biogenesis: coordinated cropping and dicing. Nature Reviews Molecular Cell Biology, 6, 376-385.

KIMURA, M. & OHTA, T. (1971). On the rate of molecular evolution. Journal of molecular evolution, 1, 1-17.

KLASS, K. D. & KRISTENSEN, N. P. Year. The ground plan and affinities of hexapods: recent progress and open problems. In, (2001). Société entomologique de France, 265-298.

KLUGE, A. G. (1989). A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). Systematic Biology, 38, 7.

KOLACZKOWSKI, B. & THORNTON, J. W. (2008). A mixed branch length model of heterotachy improves phylogenetic accuracy. Molecular biology and evolution, 25, 1054.

KRISTENSEN, R. M. (2003). Comparative morphology: Do the ultrastructural investigations of Loricifera and Tardigrada support the clade Ecdysozoa? The New Panorama of Animal evolution, 467-477.

KUMAR, S., FILIPSKI, A. J., BATTISTUZZI, F. U., POND, S. L. K. & TAMURA, K. (2011). Statistics and Truth in Phylogenomics. Molecular biology and evolution.

LAGOS-QUINTANA, M., RAUHUT, R., LENDECKEL, W. & TUSCHL, T. (2001). Identification of novel genes coding for small expressed RNAs. Science, 294, 853.

LAKE, J. A. (1990). Origin of the Metazoa. Proceedings of the National Academy of Sciences, 87, 763-766.

LARGET, B. & SIMON, D. L. (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Molecular biology and evolution, 16, 750-759.

LARTILLOT, N., BRINKMANN, H. & PHILIPPE, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evolutionary Biology, 7, S4.

LARTILLOT, N., LEPAGE, T. & BLANQUART, S. (2009). PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics, 25, 2286.

LARTILLOT, N. & PHILIPPE, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Molecular biology and evolution, 21, 1095.

LARTILLOT, N. & PHILIPPE, H. (2008). Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Philosophical Transactions of the Royal Society B: Biological Sciences, 363, 1463.

LAU, N. C., LIM, L. P., WEINSTEIN, E. G. & BARTEL, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science, 294, 858.

LECOINTRE, G., PHILIPPE, H., VAN LE, H. & LE GUYADER, H. (1993). Species sampling has a major impact on phylogenetic inference. Mol. Phylogenet. Evol, 2, 205-224.

LEE, R. C. & AMBROS, V. (2001). An extensive class of small RNAs in Caenorhabditis elegans. Science, 294, 862.

LEE, R. C., FEINBAUM, R. L. & AMBROS, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell, 75, 843-854.

LEE, Y., KIM, M., HAN, J., YEOM, K. H., LEE, S., BAEK, S. H. & KIM, V. N. (2004). MicroRNA genes are transcribed by RNA polymerase II. The EMBO journal, 23, 4051-4060.

LEMEY, P., SALEMI, M. & VANDAMME, A. M. (2009). The Phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing, Cambridge University Press Cambridge.

LEWIS, B. P., BURGE, C. B. & BARTEL, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell, 120, 15-20.

LEWIS, B. P., SHIH, I., JONES-RHOADES, M. W., BARTEL, D. P. & BURGE, C. B. (2003). Prediction of mammalian microRNA targets. Cell, 115, 787-798.

LEWIS, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. Systematic Biology, 50, 913.

LIM, L. P., LAU, N. C., GARRETT-ENGELE, P., GRIMSON, A., SCHELTER, J. M., CASTLE, J., BARTEL, D. P., LINSLEY, P. S. & JOHNSON, J. M. (2005). Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature, 433, 769-773.

LOCKHART, P., HOWE, C., BRYANT, D., BEANLAND, T. & LARKUM, A. (1992). Substitutional bias confounds inference of cyanelle origins from sequence data. Journal of molecular evolution, 34, 153-162.

LOCKHART, P. J., STEEL, M. A., HENDY, M. D. & PENNY, D. (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. Molecular biology and evolution, 11, 605-612.

LOOMIS, W. F. & SMITH, D. W. (1990). Molecular phylogeny of Dictyostelium discoideum by protein sequence comparison. Proceedings of the National Academy of Sciences, 87, 9093.

LOPEZ, P., CASANE, D. & PHILIPPE, H. (2002). Heterotachy, an important process of protein evolution. Molecular biology and evolution, 19, 1.

LÖYTYNOJA, A. & GOLDMAN, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. Proceedings of the National Academy of Sciences of the United States of America, 102, 10557.

MAISEY, J. G. (1986). Heads and tails: a chordate phylogeny. Cladistics, 2, 201-256.

MALLATT, J. & GIRIBET, G. (2006). Further use of nearly complete 28S and 18S rRNA genes to classify Ecdysozoa: 37 more arthropods and a kinorhynch. Molecular Phylogenetics and Evolution, 40, 772-794.

MALLATT, J. M., GAREY, J. R. & SHULTZ, J. W. (2004). Ecdysozoan phylogeny and Bayesian inference: first use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin. Molecular Phylogenetics and Evolution, 31, 178-191.

MANTON, S. M. (1973). Arthropod phylogeny-a modern synthesis. Journal of Zoology, 171, 111-130.

MASSINGHAM, T. & GOLDMAN, N. (2000). EDIBLE: experimental design and information calculations in phylogenetics. Bioinformatics, 16, 294.

MAU, B., NEWTON, M. & LARGET, B. (1999). Bayesian phylogenetic inference via Markov chain Monte carlo methods. Biometrics, 55, 1-12.

MAYER, G., WHITINGTON, P., SUNNUCKS, P. & PFLÜGER, H. J. (2010). A revision of brain composition in Onychophora (velvet worms) suggests that the tritocerebrum evolved in arthropods. BMC Evolutionary Biology, 10, 255.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. & TELLER, E. (1953). Equation of state calculations by fast computing machines. The journal of chemical physics, 21, 1087.

METZKER, M. L. (2009). Sequencing technologies - the next generation. Nature Reviews Genetics, 11, 31-46.

MEUSEMANN, K., VON REUMONT, B. M., SIMON, S., ROEDING, F., STRAUSS, S., KÜCK, P., EBERSBERGER, I., WALZL, M., PASS, G. & BREUERS, S. (2010). A phylogenomic approach to resolve the arthropod tree of life. Molecular biology and evolution, 27, 2451.

MICHENER, C. D. & SOKAL, R. R. (1957). A quantitative approach to a problem in classification. Evolution, 130-162.

MIURA, S., NOZAWA, M. & NEI, M. (2011). Evolutionary changes of the target sites of two microRNAs encoded in the Hox gene cluster of Drosophila and other insect species. Genome Biology and Evolution, 3, 129.

MOON, S. Y. E. O. & KIM, W. (1996). Phylogenetic position of the Tardigrada based on the 18S ribosomal RNA gene sequences. Zoological Journal of the Linnean Society, 116, 61-69.

NARDI, F., SPINSANTI, G., BOORE, J. L., CARAPELLI, A., DALLAI, R. & FRATI, F. (2003). Hexapod origins: monophyletic or paraphyletic? Science, 299, 1887.

NEGRISOLO, E., MINELLI, A. & VALLE, G. (2004). The mitochondrial genome of the house centipede Scutigera and the monophyly versus paraphyly of myriapods. Molecular biology and evolution, 21, 770.

NESNIDAL, M. P., HELMKAMPF, M., BRUCHHAUS, I. & HAUSDORF, B. (2010). Compositional heterogeneity and phylogenomic inference of metazoan relationships. Molecular biology and evolution, 27, 2095.

NIELSEN, C. (2001). Animal evolution: interrelationships of the living phyla. 2nd Edition, Oxford University Press, USA.

NOTREDAME, C. (2007). Recent evolutions of multiple sequence alignment algorithms. PLoS Computational Biology, 3, e123.

NOVACEK, M. & WHEELER, Q. (1992). Extinct taxa: accounting for 99.999...% of the earths biota. Extinction and phylogeny (MJ Novacek and QD Wheeler, eds.). Columbia University Press, New York, 1-16.

O'BRIEN, S. J. & STANYON, R. (1999). Phylogenomics. Ancestral primate viewed. Nature, 402, 365.

ØDEGAARD, F. (2000). How many species of arthropods? Erwin's estimate revised. Biological Journal of the Linnean Society, 71, 583-597.

OLDROYD, B. P. (2007). What's killing American honey bees? PLoS biology, 5, e168.

OWEN, R. (1843). Lectures on the comparative anatomy and physiology of the invertebrate animals. Longman, Brown, Green and Longman, London.

PARK, J. K., RHO, H. S., KRISTENSEN, R. M., KIM, W. & GIRIBET, G. (2006). First molecular data on the phylum Loricifera-an investigation into the phylogeny of Ecdysozoa with emphasis on the positions of Loricifera and Priapulida. Zoological Science, 23, 943-954.

PASQUINELLI, A. E., REINHART, B. J., SLACK, F., MARTINDALE, M. Q., KURODAK, M. I. & MALLER, B. (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature, 408, 86-89.

PATTERSON, C. (1988). Homology in classical and molecular biology. Molecular biology and evolution, 5, 603.

PETERSON, K. J., COTTON, J. A., GEHLING, J. G. & PISANI, D. (2008). The Ediacaran emergence of bilaterians: congruence between the genetic and the geological fossil records. Philosophical Transactions of the Royal Society B: Biological Sciences, 363, 1435.

PETERSON, K. J. & EERNISSE, D. J. (2001). Animal phylogeny and the ancestry of bilaterians: inferences from morphology and 18S rDNA gene sequences. Evolution & Development, 3, 170-205.

PETERSON, K. J., LYONS, J. B., NOWAK, K. S., TAKACS, C. M., WARGO, M. J. & MCPEEK, M. A. (2004). Estimating metazoan divergence times with a molecular clock. Proceedings of the National Academy of Sciences of the United States of America, 101, 6536.

PHILIPPE, H., BRINKMANN, H., COPLEY, R. R., MOROZ, L. L., NAKANO, H., POUSTKA, A. J., WALLBERG, A., PETERSON, K. J. & TELFORD, M. J. (2011a). Acoelomorph flatworms are deuterostomes related to Xenoturbella. Nature, 470, 255-258.

PHILIPPE, H., BRINKMANN, H., LAVROV, D. V., LITTLEWOOD, D. T. J., MANUEL, M., WÀÜRHEIDE, G. & BAURAIN, D. (2011b). Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS biology, 9, e1000602.

PHILIPPE, H., DELSUC, F., BRINKMANN, H. & LARTILLOT, N. (2005a). Phylogenomics. Annual Review of Ecology, Evolution, and Systematics, 541-562.

PHILIPPE, H., LARTILLOT, N. & BRINKMANN, H. (2005b). Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Molecular biology and evolution, 22, 1246.

PHILIPPE, H., DERELLE, R., LOPEZ, P., PICK, K., BORCHIELLINI, C., BOURY-ESNAULT, N., VACELET, J., RENARD, E., HOULISTON, E. & QUÉINNEC, E. (2009). Phylogenomics revives traditional views on deep animal relationships. Current Biology, 19, 706-712.

PHILIPPE, H. & GERMOT, A. (2000). Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution. Molecular biology and evolution, 17, 830.

PHILIPPE, H. & LAURENT, J. (1998). How good are deep phylogenetic trees? Current opinion in genetics & development, 8, 616-623.

PHILIPPE, H. & LOPEZ, P. (2001). On the conservation of protein sequences in evolution. Trends Biochem. Sci, 26, 414-416.

PHILIPPE, H. & ROURE, B. (2011). Difficult phylogenetic questions: more data, maybe; better methods, certainly. BMC biology, 9, 91.

PHILIPPE, H., SNELL, E. A., BAPTESTE, E., LOPEZ, P., HOLLAND, P. W. H. & CASANE, D. (2004). Phylogenomics of eukaryotes: impact of missing data on large alignments. Molecular biology and evolution, 21, 1740.

PICK, K., PHILIPPE, H., SCHREIBER, F., ERPENBECK, D., JACKSON, D., WREDE, P., WIENS, M., ALIE, A., MORGENSTERN, B. & MANUEL, M. (2010). Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Molecular biology and evolution, 27, 1983.

PILATO, G., BINDA, M. G., BIONDI, O., D'URSO, V., LISI, O., MARLETTA, A., MAUGERI, S., NOBILE, V., RAPPAZZO, G. & SABELLA, G. (2005). The clade Ecdysozoa, perplexities and questions. Zoologischer Anzeiger-A Journal of Comparative Zoology, 244, 43-50.

PISANI, D. (2004). Identifying and removing fast-evolving sites using compatibility analysis: an example from the Arthropoda. Systematic Biology, 53, 978.

PISANI, D., BENTON, M. J. & WILKINSON, M. (2007). Congruence of morphological and molecular phylogenies. Acta Biotheoretica, 55, 269-281.

PISANI, D., FEUDA, R., PETERSON, K. J. & SMITH, A. B. (2011). Resolving phylogenetic signal from noise when divergence is rapid: A new look at the old problem of echinoderm class relationships. Molecular Phylogenetics and Evolution.

POE, S. & SWOFFORD, D. L. (1999). Taxon sampling revisited. Nature, 398, 299-300.

POLLOCK, D. D., ZWICKL, D. J., MCGUIRE, J. A. & HILLIS, D. M. (2002). Increased taxon sampling is advantageous for phylogenetic inference. Systematic Biology, 51, 664.

POSADA, D. & CRANDALL, K. A. (1998). Modeltest: testing the model of DNA substitution. Bioinformatics, 14, 817-818.

PRPIC, N. M. & TAUTZ, D. (2003). The expression of the proximodistal axis patterning genes Distal-less and dachshund in the appendages of Glomeris marginata (Myriapoda: Diplopoda) suggests a special role of these genes in patterning the head appendages. Developmental biology, 260, 97-112.

QUESNE, W. J. L. (1969). A method of selection of characters in numerical taxonomy. Systematic Biology, 18, 201.

RAGAN, M. A., MCINERNEY, J. O. & LAKE, J. A. (2009). The network of life: genome beginnings and evolution. Philosophical Transactions of the Royal Society B: Biological Sciences, 364, 2169.

RANNALA, B., HUELSENBECK, J. P., YANG, Z. & NIELSEN, R. (1998). Taxon sampling and the accuracy of large phylogenies. Systematic Biology, 47, 702-710.

RANNALA, B. & YANG, Z. (1996). Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. Journal of molecular evolution, 43, 304-311.

RAUP, D. M. (1981). Extinction: bad genes or bad luck? Acta geológica hispánica, 16, 25-33.

REGIER, J. C. & SHULTZ, J. W. (1997). Molecular phylogeny of the major arthropod groups indicates polyphyly of crustaceans and a new hypothesis for the origin of hexapods. Molecular biology and evolution, 14, 902.

REGIER, J. C. & SHULTZ, J. W. (2001). Elongation factor-2: a useful gene for arthropod phylogenetics. Molecular Phylogenetics and Evolution, 20, 136-148.

REGIER, J. C., SHULTZ, J. W. & KAMBIC, R. E. (2005). Pancrustacean phylogeny: hexapods are terrestrial crustaceans and maxillopods are not monophyletic. Proceedings of the Royal Society B: Biological Sciences, 272, 395.

REGIER, J. C., SHULTZ, J. W., ZWICK, A., HUSSEY, A., BALL, B., WETZER, R., MARTIN, J. W. & CUNNINGHAM, C. W. (2010). Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature, 463, 1079-1083.

REINHART, B. J., SLACK, F. J., BASSON, M., PASQUINELLI, A. E., BETTINGER, J. C., ROUGVIE, A. E., HORVITZ, H. R. & RUVKUN, G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature, 403, 901-906.

RODRIGUEZ, A., GRIFFITHS-JONES, S., ASHURST, J. L. & BRADLEY, A. (2004). Identification of mammalian microRNA host genes and transcription units. Genome research, 14, 1902-1910.

RODRÌGUEZ-TRELLES, F., TARRÌO, R. & AYALA, F. J. (2002). A methodological bias toward overestimation of molecular evolutionary time scales. Proceedings of the National Academy of Sciences, 99, 8112.

ROEDING, F., BORNER, J., KUBE, M., KLAGES, S., REINHARDT, R. & BURMESTER, T. (2009). A 454 sequencing approach for large scale phylogenomic analysis of the common emperor scorpion (Pandinus imperator). Molecular Phylogenetics and Evolution, 53, 826-834.

ROEDING, F., HAGNER-HOLLER, S., RUHBERG, H., EBERSBERGER, I., VON HAESELER, A., KUBE, M., REINHARDT, R. & BURMESTER, T. (2007). EST sequencing of Onychophora and phylogenomic analysis of Metazoa. Molecular Phylogenetics and Evolution, 45, 942-951.

ROGOZIN, I. B., WOLF, Y. I., CARMEL, L. & KOONIN, E. V. (2007). Analysis of rare amino acid replacements supports the Coelomata clade. Molecular biology and evolution, 24, 2594-2597.

RONQUIST, F. & HUELSENBECK, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics, 19, 1572.

ROSENBERG, M. S. & KUMAR, S. (2001). Incomplete taxon sampling is not a problem for phylogenetic inference. Proceedings of the National Academy of Sciences, 98, 10751.

ROTA-STABELLI, O., CAMPBELL, L., BRINKMANN, H., EDGECOMBE, G. D., LONGHORN, S. J., PETERSON, K. J., PISANI, D., PHILIPPE, H. & TELFORD, M. J. (2011). A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. Proceedings of the Royal Society B: Biological Sciences, 278, 298.

ROTA-STABELLI, O., KAYAL, E., GLEESON, D., DAUB, J., BOORE, J. L., TELFORD, M. J., PISANI, D., BLAXTER, M. & LAVROV, D. V. (2010). Ecdysozoan mitogenomics: evidence for a common origin of the legged invertebrates, the Panarthropoda. Genome Biology and Evolution, 2, 425.

ROTA-STABELLI, O. & TELFORD, M. J. (2008). A multi criterion approach for the selection of optimal outgroups in phylogeny: recovering some support for Mandibulata over Myriochelata using mitogenomics. Molecular Phylogenetics and Evolution, 48, 103-111.

ROTHBERG, J. M., HINZ, W., REARICK, T. M., SCHULTZ, J., MILESKI, W., DAVEY, M., LEAMON, J. H., JOHNSON, K., MILGREW, M. J. & EDWARDS, M. 2011. An integrated semiconductor device enabling non-optical genome sequencing. Nature, 475, 348-352.

RUBY, J. G., STARK, A., JOHNSTON, W. K., KELLIS, M., BARTEL, D. P. & LAI, E. C. (2007). Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. Genome research, 17, 1850-1864.

RUIZ-TRILLO, I., PAPS, J., LOUKOTA, M., RIBERA, C., JONDELIUS, U., BAGUNA, J. & RIUTORT, M. (2002). A phylogenetic analysis of myosin heavy chain type II sequences corroborates that Acoela and Nemertodermatida are basal bilaterians. Proceedings of the National Academy of Sciences, 99, 11246.

SANDERSON, M., MCMAHON, M. & STEEL, M. (2010). Phylogenomics with incomplete taxon coverage: the limits to inference. BMC Evolutionary Biology, 10, 155.

SANDERSON, M. J. & SHAFFER, H. B. (2002). Troubleshooting molecular phylogenetic analyses. Annual Review of Ecology and Systematics, 49-72.

SCHMIDT RHAESA, A., BARTOLOMAEUS, T., LEMBURG, C., EHLERS, U. & GAREY, J. R. (1998). The position of the Arthropoda in the phylogenetic system. Journal of Morphology, 238, 263-285.

SCHMIDT-RHAESA, A. (1996). The nervous system of Nectonema munidae and Gordius aquaticus, with implications for the ground pattern of the Nematomorpha. Zoomorphology, 116, 133-142.

SCHMIDT-RHAESA, A. (2001). Tardigrades - Are They Really Miniaturized Dwarfs? Zoologischer Anzeiger-A Journal of Comparative Zoology, 240, 549-555.

SCHOLTZ, G. (2002). The Articulata hypothesis-or what is a segment? Organisms Diversity & Evolution, 2, 197-215.

SCHOLTZ, G. & EDGECOMBE, G. D. (2006). The evolution of arthropod heads: reconciling morphological, developmental and palaeontological evidence. Development Genes and Evolution, 216, 395-415.

SCHOLTZ, G., MITTMANN, B. & GERBERDING, M. (1998). The pattern of Distal-less expression in the mouthparts of crustaceans, myriapods and insects: new evidence for a gnathobasic mandible and the common origin of Mandibulata. The International journal of developmental biology, 42, 801.

SCHWARZ, D. S., HUTV¬∑GNER, G., DU, T., XU, Z., ARONIN, N. & ZAMORE, P. D. (2003). Asymmetry in the assembly of the RNAi enzyme complex. Cell, 115, 199-208.

SCOTLAND, R. W., OLMSTEAD, R. G. & BENNETT, J. R. (2003). Phylogeny reconstruction: the role of morphology. Systematic Biology, 52, 539-548.

SEMPERE, L. F., COLE, C. N., MCPEEK, M. A. & PETERSON, K. J. (2006). The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. JOURNAL OF EXPERIMENTAL ZOOLOGY PART B MOLECULAR AND DEVELOPMENTAL EVOLUTION, 306, 575.

SEMPERE, L. F., MARTINEZ, P., COLE, C., BAGUÑÀ, J. & PETERSON, K. J. (2007). Phylogenetic distribution of microRNAs supports the basal position of acoel flatworms and the polyphyly of Platyhelminthes. Evolution & Development, 9, 409-415.

SHUBIN, N., TABIN, C. & CARROLL, S. (2009). Deep homology and the origins of evolutionary novelty. Nature, 457, 818-823.

SIEBOLD, C. T. W. V. & H, S. (1848). Lehrbuch der vergleichenden Anatomie der wirbellosen Tiere. Veit, Berlin.

SILLMAN, L. R. (1960). The origin of the vertebrates. Journal of Palaeontology, 540-544.

SMITH, N. D. & TURNER, A. H. (2005). Morphology's role in phylogeny reconstruction: perspectives from palaeontology. Systematic Biology, 54, 166.

SMITH, S. A., WILSON, N. G., GOETZ, F. E., FEEHERY, C., ANDRADE, S. C. S., ROUSE, G. W., GIRIBET, G. & DUNN, C. W. (2011). Resolving the evolutionary relationships of molluscs with phylogenomic tools. Nature, 480, 364 – 367.

SNODGRASS, R. E. (1938). Evolution of the Annelida, Onychophora and Arthropoda, The Smithsonian institution.

SOKAL, R. R. & SNEATH, P. H. A. (1963). Principles of numerical taxonomy, Freeman San Francisco.

SONG, J. J., SMITH, S. K., HANNON, G. J. & JOSHUA-TOR, L. (2004). Crystal structure of Argonaute and its implications for RISC slicer activity. Science, 305, 1434.

SØRENSEN, M. V., HEBSGAARD, M. B., HEINER, I., GLENNER, H., WILLERSLEV, E. & KRISTENSEN, R. M. (2008). New data from an enigmatic phylum: evidence from molecular sequence data supports a sister group relationship between Loricifera and Nematomorpha. Journal of Zoological Systematics and Evolutionary Research, 46, 231-239.

SPEARS, T. & ABELE, L. (1997). Crustacean phylogeny inferred from 18S rDNA. In R. A. Fortey and R.H. Thomas (eds.), Arthropod relationships. Chapman & Hall, London.

SPENCER, M., SUSKO, E. & ROGER, A. J. (2005). Likelihood, parsimony, and heterogeneous evolution. Molecular biology and evolution, 22, 1161.

SPERLING, E., ROBINSON, J., PISANI, D. & PETERSON, K. (2010). Where's the glass? Biomarkers, molecular clocks, and microRNAs suggest a 200 Myr missing Precambrian fossil record of siliceous sponge spicules. Geobiology, 8, 24-36.

SPERLING, E. A. & PETERSON, K. J. (2009). microRNAs and metazoan phylogeny: big trees from little genes. Animal Evolution: Genomes, Fossils, and Trees, 157.

SPERLING, E. A., PETERSON, K. J. & PISANI, D. (2009a). Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa. Molecular biology and evolution, 26, 2261-2274.

SPERLING, E. A., VINTHER, J., MOY, V. N., WHEELER, B. M., SÉMON, M., BRIGGS, D. E. G. & PETERSON, K. J. (2009b). MicroRNAs resolve an apparent conflict between annelid systematics and their fossil record. Proceedings of the Royal Society B: Biological Sciences, 276, 4315.

SPERLING, E. A., PISANI, D. & PETERSON, K. J. (2011). Molecular paleobiological insights into the origin of the Brachiopoda. Evolution & Development, 13, 290-303.

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

STARK, A., LIN, M. F., KHERADPOUR, P., PEDERSEN, J. S., PARTS, L., CARLSON, J. W., CROSBY, M. A., RASMUSSEN, M. D., ROY, S. & DEORAS, A. N. (2007). Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature, 450, 219-232.

STOLLEWERK, A. & CHIPMAN, A. D. (2006). Neurogenesis in myriapods and chelicerates and its importance for understanding arthropod relationships. Integrative and comparative biology, 46, 195.

STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society. Series B (Methodological), 111-147.

SULLIVAN, J. & SWOFFORD, D. L. (2001). Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? Systematic Biology, 50, 723-729.

SWOFFORD, D. (1998). PAUP 4.0: phylogenetic analysis using parsimony, Smithsonian Institution.

SWOFFORD, D. L., OLSEN, G. J., WADDELL, P. J. & HILLIS, D. M. (1996). Phylogenetic inference.

SWOFFORD, D. L., WADDELL, P. J., HUELSENBECK, J. P., FOSTER, P. G., LEWIS, P. O. & ROGERS, J. S. (2001). Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Systematic Biology, 50, 525-539.

TARVER, J., E., SPERLING, E. A., NAILOR, A., HEIMBERG, A. M., ROBINSON, J. M., PISANI, D., DONOGHUE, P. C. J. & PETERSON, K. J. (2012). The use of miRNAs in Phylogenetics as Rare Genomic Characters. Personal Communication. In preparation.

TELFORD, M. (2006). Animal phylogeny. Current Biology, 16, R981-R985.

TELFORD, M. J., BOURLAT, S. J., ECONOMOU, A., PAPILLON, D. & ROTA-STABELLI, O. (2008). The evolution of the Ecdysozoa. Philosophical Transactions of the Royal Society B: Biological Sciences, 363, 1529.

TELFORD, M. J. & COPLEY, R. R. (2011). Improving animal phylogenies with genomic data. Trends in Genetics, 27, 186-195.

TELFORD, M. J., LOCKYER, A. E., CARTWRIGHT-FINCH, C. & LITTLEWOOD, D. T. J. (2003). Combined large and small subunit ribosomal RNA phylogenies support a basal position of the acoelomorph flatworms. Proceedings of the Royal Society of London. Series B: Biological Sciences, 270, 1077.

THOMPSON, J. D., HIGGINS, D. G. & GIBSON, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic acids research, 22, 4673.

TIEGS, O. (1947). The development and affinities of the Pauropoda, based on a study of Pauropus silvaticus. Quarterly Journal of Microscopical Science, 3, 275.

TUFFLEY, C. & STEEL, M. (1998). Modeling the covarion hypothesis of nucleotide substitution. Mathematical Biosciences, 147, 63-91.

TURBEVILLE, J., PFEIFER, D., FIELD, K. & RAFF, R. (1991). The phylogenetic status of arthropods, as inferred from 18S rRNA sequences. Molecular biology and evolution, 8, 669-702.

UNGERER, P. & SCHOLTZ, G. (2008). Filling the gap between identified neuroblasts and neurons in crustaceans adds new support for Tetraconata. Proceedings of the Royal Society B: Biological Sciences, 275, 369.

UZZELL, T. & CORBIN, K. W. (1971). Fitting discrete probability distributions to evolutionary events. Science, 172, 1089.

VALENTINE, J. W., COLLINS, A. G. & MEYER, C. P. (1994). Morphological complexity increase in metazoans. Paleobiology, 131-142.

VAN DEN BUSSCHE, R. A., BAKER, R. J., HUELSENBECK, J. P. & HILLIS, D. M. (1998). Base Compositional Bias and Phylogenetic Analyses: A Test of the "flying DNA" hypothesis. Molecular Phylogenetics and Evolution, 10, 408-416.

VON REUMONT, B. M., JENNER, R. A., WILLS, M. A., DELL'AMPIO, E., PASS, G., EBERSBERGER, I., MEYER, B., KOENEMANN, S., ILIFFE, T. M. & STAMATAKIS, A. (2011). Pancrustacean phylogeny in the light of new phylogenomic data: support for Remipedia as the possible sister group of Hexapoda. Molecular biology and evolution.

WAGELE, J. & MISOF, B. (2001). On quality of evidence in phylogeny reconstruction: a reply to Zrzavýs defence of the Ecdysozoa hypothesis. J. Zool. Syst. Evol. Res, 39, 165-176.

WAGNER, G. P. (2007). The developmental genetics of homology. Nature Reviews Genetics, 8, 473-479.

WASMUTH, J. & BLAXTER, M. (2004). prot4EST: translating expressed sequence tags from neglected genomes. BMC Bioinformatics, 5, 187.

WEBSTER, B. L., COPLEY, R. R., JENNER, R. A., MACKENZIE DODDS, J. A., BOURLAT, S. J., ROTA STABELLI, O., LITTLEWOOD, D. & TELFORD, M. J. (2006). Mitogenomics and phylogenomics reveal priapulid worms as extant models of the ancestral Ecdysozoan. Evolution & Development, 8, 502-510.

WHEELER, B. M., HEIMBERG, A. M., MOY, V. N., SPERLING, E. A., HOLSTEIN, T. W., HEBER, S. & PETERSON, K. J. (2009). The deep evolution of metazoan microRNAs. Evol Dev, 11, 50-68.

WHEELER, W. C. (1990). Nucleic acid sequence phylogeny and random outgroups. Cladistics, 6, 363-367.

WHEELER, W. C., CARTWRIGHT, P. & HAYASHI, C. Y. (1993). Arthropod phylogeny: a combined approach. Cladistics, 9, 1-39.

WHELAN, S. & GOLDMAN, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Molecular biology and evolution, 18, 691-699.

WHITINGTON, P. M. & MAYER, G. (2011). The origins of the arthropod nervous system: insights from the Onychophora. Arthropod Structure & Development.

WHITTINGTON, H. B. & BRIGGS, D. E. G. (1985). The largest Cambrian animal, Anomalocaris, Burgess Shale, British Columbia. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 569-609.

WIENS, J. J. (2003). Missing data, incomplete taxa, and phylogenetic accuracy. Systematic Biology, 52, 528.

WIENS, J. J. (2004). The role of morphological data in phylogeny reconstruction. Systematic Biology, 53, 653.

WIENS, J. J. (2006). Missing data and the design of phylogenetic analyses. Journal of Biomedical Informatics, 39, 34-42.

WILLS, M., BRIGGS, D., FORTEY, R. & WILKINSON, M. (1995). The significance of fossils in understanding arthropod evolution. VERHANDLUNGEN-DEUTSCHEN ZOOLOGISCHEN GESELLSCHAFT, 88, 203-216.

WILLS, M. A., BRIGGS, D. E. G., FORTEY, R. A., WILKINSON, M. & SNEATH, P. H. A. (1998). An arthropod phylogeny based on fossil and recent taxa. Arthropod fossils and phylogeny, 33-105.

WILSON, E. O. (1988). Consilience: The Unity of Knowledge. Alfred A. Knopf, New York, 332.

WINNEPENNINCKX, B., BACKELJAU, T., MACKEY, L. Y., BROOKS, J. M., DE WACHTER, R., KUMAR, S. & GAREY, J. R. (1995). 18S rRNA data indicate that Aschelminthes are polyphyletic in origin and consist of at least three distinct clades. Molecular biology and evolution, 12, 1132-1137.

WOESE, C. R., KANDLER, O. & WHEELIS, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proceedings of the National Academy of Sciences, 87, 4576.

WOLF, Y., ROGOZIN, I., GRISHIN, N., TATUSOV, R. & KOONIN, E. (2001). Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evolutionary Biology, 1, 8.

WOLF, Y. I., ROGOZIN, I. B. & KOONIN, E. V. (2004). Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. Genome research, 14, 29-36.

WOLFF, C. & SCHOLTZ, G. (2006). Cell lineage analysis of the mandibular segment of the amphipod Orchestia cavimana reveals that the crustacean paragnaths are sternal outgrowths and not limbs. Frontiers in Zoology, 3, 19.

WOODGER, J. (1945). On biological transformations. Essays on Growth and Form, 95-120.

YANG, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. Trends in Ecology & Evolution, 11, 367-372.

YANG, Z. & RANNALA, B. (1997). Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Molecular biology and evolution, 14, 717.

ZANTKE, J., WOLFF, C. & SCHOLTZ, G. (2008). Three-dimensional reconstruction of the central nervous system of Macrobiotus hufelandi (Eutardigrada, Parachela): implications for the phylogenetic position of Tardigrada. Zoomorphology, 127, 21-36.

ZHANG, Z. H. I. Q. (2011). Animal biodiversity: An introduction to higher-level classification and taxonomic richness. Zootaxa, 3148, 7-12.

ZHOU, Y., RODRIGUE, N., LARTILLOT, N. & PHILIPPE, H. (2007). Evaluation of the models handling heterotachy in phylogenetic inference. BMC Evolutionary Biology, 7, 206.

ZRZAV, J., MIHULKA, S., KEPKA, P., BEZD K, A. & TIETZ, D. (1998). Phylogeny of the Metazoa based on morphological and 18S ribosomal DNA evidence. Cladistics, 14, 249-285.

ZRZAVÝ, J. & ŠTYS, P. (1997). The basic body plan of arthropods: insights from evolutionary morphology and developmental biology. Journal of Evolutionary Biology, 10, 353-367.

ZUCKERKANDL, E. & PAULING, L. (1962). Molecular disease, evolution and genetic heterogeneity. Horizons in biochemistry, 189-225.

ZUKER, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic acids research, 31, 3406-3415.

ZWICKL, D. J. & HILLIS, D. M. (2002). Increased taxon sampling greatly reduces phylogenetic error. Systematic Biology, 51, 588.

Appendices

Appendix 1

This protocol is purpose made to generate a miRNA library for sequencing and identification of novel miRNAs.

Materials:

RNA isolated at a minimum concentration of 1.3 mg/ml in 150 µl (need a minimum of 200 µg in 150 µl)

Trizol Reagent - Invitrogen Catalogue no. (15596-018)

Dry ice and liquid nitrogen

Large mortar and pestle

50 ml polypropylene copolymer

Chloroform

Isopropanol alcohol

Ethanol alcohol

RNase Free Water (DEPC treated or otherwise prepared)

National Diagnostics RNA gel reagents:

Gel concentrate EC-830

Gel diluent EC-840

Gel buffer EC-835

1X TBE Running Buffer (diluted from 10X stock) - National Diagnostics Catalogue no. EC-860)

2X sample loading buffer 8M Urea, 0.5 mM EDTA, Bromo Blue (dry chemicals)

18, 28, 40, and 50 nucleotide fitc labeled markers (Integrated DNA Tech.)

1 mg/ml glycogen

3' Ligation:

2 µl 5x Ligation Buffer

2 µl 100 mM App 17.91x

1 µl T4 RNA Ligase

28 and 40 fitc nucleotide markers

5’ Ligation:

2 µl 5x Ligation Buffer

2 µl 200 µM 17.93R

1 µl 4 mM ATP

1 µl T4 RNA Ligase

50 fitc nucleotide marker

cDNA:

1 µl 100 µM 15.22

10 µl dH20

6 µl 5X first strand buffer

7 µl 10X dNTP’s

3 µl 100 mM DTT

1 µl Superscript III reverse transcriptase

1 µl RNase H

Library Amplification:

10 µl 10X PCR Buffer

10 µl 10X dNTPs (1X = 0.2 mM of each dNTP)

1 µl 100 µM Barcoded forward primer

1 µl 100 µM Barcoded reverse primer

72 µl dH20

1 µl of Taq Polymerase

100 % Ethanol

0.3 M NaCl

RNase Free Water

10 bp or 100bp DNA ladder

Phenol

Chloroform

pGEM T-easy vector (Promega #A1360)

SOC or LB broth

LB +Amp/Xgal/IPTG bacterial agar plates

Magnificent broth (MacConnell Research Catalogue no. MR2001)

Qiagen miniprep kit

Qiagen QIAquick Gel Extraction Kit (Qiagen, CA, USA)

Siliconized eppendorf tubes (1.5 ml)

Ethidium Bromide

Razor Blades

Agarose

TAE solution

Equipment:

BioRad minigel apparatus. Catalogue no. (165-8003)

Casting Tray (including clamps, 10 well comb, short plates, 1.0 mm spacer plates)

Power Source capable of running at a constant 2 Watts

Hot Block capable of reaching 80°C

PCR Thermocycler

Rotator in 4°C environment (i.e. a cold room)

Ultraviolet box for gel visualization

Camera for gel pictures (GEL DOC)

Microcentrifuge

NanoDrop ND-1000 spectrophotometer

Solutions:

LB Amp/Xgal/IPTG

0.3 M NaCl

2X Urea Loading Buffer

Ligation Buffer (For 5 ml-aliquot into 1ml)

250 mM Hepes pH 8.3	1.25 ml of 1M Hepes pH 8.3
50 mM MgCl2	250 µl of 1M MgCl2
16.5 mM DTT	82.5 µl of 1M DTT (made in dH2O)
50 µg/ml BSA	5 µl of 50 µg/ml BSA
41.5% glycerol	2.075 ml of 100% glycerol

Protocol Procedure:

DAY 1 - (Size fractioning of RNA)

1. Pour 15% 1.5 mm denaturing polyacrylamide gel using bio-rad mini gel apparatus. Use 10 well comb.
   Volumes for 2 Gels:

   9.0 ml Concentrate

   4.5 ml Diluent

   1.5 ml Buffer

   150 µl 10 % APS (ammonium persulfate made in water)

   7.5 µl Temed
2. Prepare an aliquot of total RNA (200-500 µg) with an equal volume of 8 M + 0.5 mM EDTA loading dye in a total volume of 300 µl or less (this means that you need a minimum of 150 µl of RNA at 1.3 mg/ml to start each library).
3. Add 1 µl of each 28 fitc and 18 fitc nucleotide molecular markers (10 µM stocks) per lane going to load (i.e. add 10µl of loading into 10 lanes)
4. Heat samples to 80 °C for 5 min.
5. Flush wells using a pipette to push out the dense urea.
6. Fill Chamber with 500 ml 1X TBE made in RNase free water.
7. Load sample into flushed wells using as many wells as possible (~30µl/well).
8. Run gel at 2 watts for 1-2 hours until the lower dye-front is approx 1 cm from the bottom.

9. Remove gel, open glass plates, and wrap gel in plastic wrap (keeping note of the orientation).
10. Take a picture of the gel (using UV gel doc system) to document placement of the markers.
11. Then, over a UV light box, draw a rectangle around the area including the two molecular markers.
12. Use the rectangle as a guide where to cut the gel.
    a. Using a new blade cut along the rectangle and place the small gel piece into a clean (pre-weighed) eppendorf tube.
    b. Weigh the eppendorf tube + gel piece and calculate the weight of the gel piece alone.
13. Crush the gel pieces then and add 3 times the volume (of the gel piece) of 0.3 M NaCl.
14. Let rotate at 4 °C overnight.
15. Stain the remaining gel with Ethidium Bromide for approximately 10 minutes.

DAY 2 – (Precipitation)

1. Remove samples from 4 °C.
2. Transfer as much as possible of the liquid portion (containing NaCl and RNA) to a clean eppendorf tube.
3. Spin briefly to pellet small pieces of acrylamide gel and again transfer the supernatant to a clean eppendorf tube.
4. Add 2 times the volume of 100% ethanol to the supernatant.
5. Add 1 µg/ml glycogen (using ~1µl of 1mg/ml stock stored in an eppendorf at -20).
6. Mix by inverting 2-3 times.
7. Store at -20 °C minimum of overnight.

DAY 3 (3' linker ligation)

1. Remove samples from -20 °C freezer.
2. Spin tubes at 13,000 x g for 30 min at 4 °C.
3. Remove supernatant and allow pellet to air dry for approximately 10 minutes in a fume hood.
4. Resuspend pellets (of the same organism) in a total of 10 µl RNase free water (i.e. if you have multiple tubes repeat this serially with same 10 µl).
5. Set up 3' adaptor ligation reaction (all reagents stored at -20 °C)

   2 µl 5x Ligation Buffer

   2 µl 100 mM App 17.91x

   1 µl T4 RNA Ligase

   5 µl purified small RNAs (from the 10 µl resuspension step 4)

   Store remaining 5µl of small RNAs at -20°C

6. Let incubate at 15 - 30°C for 2 hours.
7. During 2 hour incubation prepare 15% denaturing polyacrylamide gel.

8. Stop reaction with 15 µl 2X Urea Loading Dye (8 M Urea 0.5 M EDTA).
9. Add 2 µl each of 40 fitc and 28 fitc nucleotide molecular marker.
10. Heat samples for 5 min at 80 °C.
11. Load samples in 2-4 lanes (use more than one lane to prevent overloading and to dilute the salt in the reaction).
12. Run gel at 2 watts until good separation between the BB and XC dyes (~3 inches).
13. Take picture of gel as before and mark a rectangle within each lane above the 28 nt marker (don't include) and above 40 nt marker (include). Cut out the fragment.
14. Place gel pieces in eppendorf tube and elute overnight (follow **Day 1**, step 12).

DAY 4 – (Precipitation)

15. Precipitate RNA with glycogen (follow **Day 2**). Store at -20°C.

DAY 5 – (5' linker ligation)

1. Remove samples from -20°C freezer.
2. Spin tubes at 13,000 x g for 30 min at 4 °C.
3. Remove supernatant and allow pellet to air dry for approximately 10 minutes in a fume hood.
4. Resuspend pellets (of the same organism) in a total of 10 µl RNase free water.
5. Set up 5' adaptor ligation reaction (All reagents stored at -20 °C)

   2 µl 5x Ligation Buffer

   2 µl 200 µM 17.93R

   1 µl 4 mM ATP

   1 µl T4 RNA Ligase

   5 µl small RNAs (from the 10 µl resuspension in step 4)

6. Allow reaction to sit at room temperature (15-30°C) for 6 hours.
7. During 6 hour incubation pour 15 % polyacrylamide gels with 10 well comb.
8. Stop reaction with 13 µl 2X Urea loading dye.
9. Add 2 µl of 50 nt fitc molecular marker.
10. Heat to 80 °C for 5 minutes.
11. Remove comb and flush wells thoroughly.
12. Load sample into 2-4 wells.
13. Run gel at 2 Watts for 1-2 hours.
14. Cut out gel pieces above the 50 nt marker and elute overnight at 4°C with 0.3 M NaCl.

DAY 6 – (Precipitation)

1. Precipitate RNA with 2X volume of ethanol and 1µg/ml glycogen overnight (same as **Day 2**).

DAY 7 – (cDNA synthesis)

1. Remove samples from -20°C.
2. Spin tubes at 13,000 x g for 30 min at 4 °C.
3. Remove supernatant and allow pellet to air dry for approx 10 minutes in a fume hood.
4. Resuspend pellets (of the same organism) in a total of 10 µl RNase free water.
5. Set up RT-PCR of small RNAs with Adaptors to synthesize cDNA (all reagents stored at -20°C).

   5 µl of ligated RNAs

   1 µl 100 µM 15.22

   10 µl dH20

   HEAT to 80°C for 2 min

   SPIN down to cool

   Add

   6 µl 5X first strand buffer

   7 µl 10X dNTP's

   3 µl 100 mM DTT

   HEAT to 48°C for 2 min

   REMOVE 3µl to a new tube (for –RT control)

   Add 1 µl Superscript III reverse transcriptase (**NOT** to –RT control tube)

   HEAT to 48°C for 1 hour.

   Add 1 µl RNase H (to + and – controls)

   HEAT to 37°C for 30 minutes.

   Store at -20°C or continue with amplification.

PCR amplification:

6. Set up 100 µl reactions for + and – reverse transcriptase (RT) samples.

7. Combine the following (all reagents stored at -20°C).
   5 µl of cDNA

   10 µl 10X PCR Buffer

   10 µl 10X dNTPs (1X = 0.2 mM of each dNTP)

   1 µl 100 µM 17.92 (or barcoded primer A)

   1 µl 100 µM 17.93D (or barcoded primer B)

   72 µl dH20

   HEAT to 96 °C for 5 min

   or add 1 µl of Taq and use continue with program 454amp on thermocycler if using barcodes.

   REDUCE heat to 80°C

   Add 1 µl of Taq Polymerase

   Barcoded PCR conditions:

   Let cycle 33 times

   96°C 1 min

   96°C 10 sec

   50°C 1 min

   72°C 15 sec

   17.92/17.93 PCR conditions:

   After add Taq at 80°C let cycle 25 times

   94°C 30 sec

   50°C 30 sec

   72°C 30 sec

8. Check reaction by running 5µl (+ equal volume of sample buffer) on an acrylamide gel with a 10 bp ladder and staining with SYBR gold (or run on a 3% agarose gel with a 100bp ladder if using barcodes)

   a. Should see a smear around 100 nt (this is the ligated RNA) and some sharper bands of primers

9. If a smear is visible then continue, otherwise the library didn't work and you must start over from the beginning.
10. Gel purify positive band using Qiagen QIAquick Gel Extraction Kit.
11. Ligate into vector overnight.
    a. To ligate into pGEM combine 3µl of ppt product, one frozen aliquot of vector + buffer (5 µl of 2X buffer and 1 µl of vector), and 1µl of enzyme. Incubate at 16°C overnight.

Day 8 – (Vector ligation)

12. Combine ligation (may want to ethanol ppt ligation first) with 5 µl electocompetent cells.
    a. To ethanol ppt combine 2.5X volume of 100 % ethanol and 0.1X volume of sodium acetate. Let sit at -20°C for 1 hr
    b. Spin for 20 minutes at 14,000 x g at 4°C.
    c. Remove the supernatant then add 70 µl of 75% ethanol.
    d. Spin for a further 5 minutes then remove the supernatant.
    e. Allow pellet to air dry, then resuspend in 5 µl of RNase free water.
13. Electroporate at 1.8.
14. Immediately add 500 µl LB and transfer everything (500 µl LB + 10 µl of vector+product+E. coli) to 15 ml snap-cap tube.
15. Let incubate in shaker at 37°C for 45 min
16. Meanwhile pre warm 4 LB amp/Xgal/IPTG plates to 37°C
17. After 45 minutes streak 4 plates with ~200 µl of cells each and let incubate at 37°C overnight.

DAY 9 – (Colony Picking)

1. Pick one colony (using sterile toothpicks) per 3 ml of LB in a snap-cap 15 ml tube or, if you have lots of colonies, pick enough to fill a 96-1ml plate (for automated mini-prep).
2. Shake colonies overnight at 37°C (if using 96 well plate use special incubator).

DAY 10

1. Miniprep each sample.
2. NanoDrop spec some samples to check for approximate concentration.
3. Combine 500 ng DNA and 3.2 pMoles T7 or SP6 primer in 20 µl total volume with water for each sample.
4. Send for sequencing.

Publications

PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B
doi:10.1098/rspb.2010.0590
Published online

A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata

Omar Rota-Stabelli[1,2,†], Lahcen Campbell[2,†], Henner Brinkmann[3], Gregory D. Edgecombe[4], Stuart J. Longhorn[2], Kevin J. Peterson[5], Davide Pisani[2,*], Hervé Philippe[3,*] and Maximilian J. Telford[1,*]

[1]*Department of Genetics, Evolution and Environment, University College London, Darwin Building,*
Gower Street, London WC1E 6BT, UK
[2]*Department of Biology, The National University of Ireland, Maynooth, Maynooth, Co. Kildare, Ireland*
[3]*Centre Robert-Cedergren, Département de Biochimie, Université de Montréal, Succursale Centre-Ville,*
Montréal, Quebec, Canada H3C3J7
[4]*Department of Palaeontology, Natural History Museum, Cromwell Road, London SW7 5BD, UK*
[5]*Department of Biology, Dartmouth College, Hanover, NH, USA*

While a unique origin of the euarthropods is well established, relationships between the four euarthropod classes—chelicerates, myriapods, crustaceans and hexapods—are less clear. Unsolved questions include the position of myriapods, the monophyletic origin of chelicerates, and the validity of the close relationship of euarthropods to tardigrades and onychophorans. Morphology predicts that myriapods, insects and crustaceans form a monophyletic group, the Mandibulata, which has been contradicted by many molecular studies that support an alternative Myriochelata hypothesis (Myriapoda plus Chelicerata). Because of the conflicting insights from published molecular datasets, evidence from nuclear-coding genes needs corroboration from independent data to define the relationships among major nodes in the euarthropod tree. Here, we address this issue by analysing two independent molecular datasets: a phylogenomic dataset of 198 protein-coding genes including new sequences for myriapods, and novel microRNA complements sampled from all major arthropod lineages. Our phylogenomic analyses strongly support Mandibulata, and show that Myriochelata is a tree-reconstruction artefact caused by saturation and long-branch attraction. The analysis of the microRNA dataset corroborates the Mandibulata, showing that the microRNAs miR-965 and miR-282 are present and expressed in all mandibulate species sampled, but not in the chelicerates. Mandibulata is further supported by the phylogenetic analysis of a comprehensive morphological dataset covering living and fossil arthropods, and including recently proposed, putative apomorphies of Myriochelata. Our phylogenomic analyses also provide strong support for the inclusion of pycnogonids in a monophyletic Chelicerata, a paraphyletic Cycloneuralia, and a common origin of Arthropoda (tardigrades, onychophorans and arthropods), suggesting that previous phylogenies grouping tardigrades and nematodes may also have been subject to tree-reconstruction artefacts.

Keywords: arthropod; phylogeny; Mandibulata; microRNA

1. INTRODUCTION

With over 1 million living species described and a rich 520 Myr fossil record, arthropods are the most species-rich clade of animals on Earth, accounting for nearly 80 per cent of animal biodiversity [1]. Four main euarthropod sub-phyla are recognized: Hexapoda (including insects); Crustacea (lobsters, water fleas and others); Myriapoda (e.g. millipedes and centipedes); and Chelicerata (including arachnids, horseshoe crabs and possibly sea spiders). After many years of debate, a consensus has emerged that these four classes (or sub-phyla) form a monophyletic group called the Euarthropoda [2,3]. The relationships between the four euarthropod groups remain disputed, however, as is the validity of their close relationship to tardigrades (water bears) and onychophorans (velvet worms) in a more inclusive clade called Arthropoda (named Panarthropoda by Nielsen [4]).

Within the Euarthropoda, the main point of disagreement concerns the position of the myriapods, which were long thought to be most closely related to the hexapods [5]. Myriapods and hexapods notably share a distinctive head composed of five segments distinguished by their unique appendages—the antennal, intercalary

[*] Authors for correspondence (davide.pisani@nuim.ie; herve.philippe@umontreal.ca; m.telford@ucl.ac.uk).
[†] These authors contributed equally to this work.

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2010.0590 or via http://rspb.royalsocietypublishing.org.

Received 19 March 2010
Accepted 21 July 2010

(appendage-less), mandibular, and usually two pairs of maxillae (the second being the insect labium). Molecular data, however, have shown crustaceans, which differ in having a second antennal rather than an intercalary segment, to be the closest sister group of hexapods in a clade named Pancrustacea or Tetraconata [6,7]. When compared with chelicerates, the detailed similarities of the arrangement of head segments and associated appendages in Pancrustacea and myriapods strongly support their sister group relationship within a wider clade that has been named the Mandibulata in recognition of the similarity of their biting mouthparts (see the electronic supplementary material). Considering the complex shared features of myriapod and pancrustacean head morphology, it is surprising that the majority of published molecular phylogenetic analyses do not support the Mandibulata, instead placing the myriapods as the sister group of the chelicerates in an assemblage that has been named the Myriochelata or Paradoxopoda [8,9]. Molecular support for Myriochelata was initially obtained using large and small subunit rRNAs [10] and later Hox genes [8], mitochondrial protein-coding sequences [11] and combined datasets of both nuclear and mitochondrial genes [9]. Myriochelata was also supported by several phylogenomic analyses [12–15]. However, recently, a dataset of 62 nuclear protein-coding genes found support for Mandibulata [16]. Regier *et al.* [16] did not identify the factors underpinning the difference between their new results and those of previously published phylogenies that supported Myriochelata. Consequently, and in light of the varying results from these molecular samples, the Mandibulata versus Myriochelata controversy remains an open question.

Uncertainty in deep arthropod phylogeny has recently been reinforced as Mayer & Whitington [17] proposed various putative synapomorphies of the Myriochelata, including a revised character polarity for the well-studied neuro-developmental pattern [18], and the mechanism of dorsoventral patterning. Here, debate surrounds the ancestral conditions, specifically whether nervous tissue forms from immigration of single or clusters of cells, and whether or not the neuroectoderm invaginates in each developing segment.

In a similar conflict between molecules and morphology, arthropods share features including segmentation and appendages with tardigrades and onychophorans [1], yet a close relationship between these three phyla has not been clearly supported by molecular analyses. A close relationship between onychophorans and euarthropods is widely accepted, but affinities of tardigrades are less clear, to the extent that they have been linked with nematodes in several phylogenomic studies [13–15]. Recently, a mitogenomic study of the Ecdysozoa supported a monophyletic origin of these three groups, although support is model-dependent [19].

There are two explanations for the discrepancies between different molecular datasets and between molecules and morphology. First, morphology may mislead—mandibles might have evolved independently in pancrustaceans and myriapods or been lost in chelicerates; similarly, segmentation and legs may have appeared separately in arthropods, onychophorans and tardigrades. The second explanation is that some molecular data may be affected by errors—either stochastic (unlikely with phylogenomic scale datasets) or systematic such as compositional bias or long-branch attraction (LBA) [20–22]. The possibility of systematic error is suggested by some datasets being equivocal regarding myriapod [7,9,19,23,24] or tardigrade affinities [12,19].

To resolve the phylogenetic relationships of the arthropods and their ecdysozoan outgroups, we present analyses of three independent datasets. The first is a phylogenomic dataset of 198 protein-coding genes, which includes new data from the pivotal myriapods. The second is a novel set of arthropod microRNAs (miRNAs), small non-coding regulatory genes implicated in the control of cellular differentiation and homeostasis. The third is a comprehensive dataset of 393 morphological characters, including the recently proposed morphological homologies of Myriochelata [17] and recent gene expression data [25] alongside new and traditional characters supporting the Mandibulata.

In addition, we have explored the nature of the conflict between molecular datasets supporting alternative arthropod phylogenies by assaying the potential effects of systematic error on our phylogenomic dataset using an experimental approach coupling targeted taxon-sampling, the use of alternative models of molecular evolution, and the analyses of subsets of slowly evolving sites extracted from our full dataset.

2. MATERIAL AND METHODS

Detailed description of methods used to generate novel expressed sequence tags and, miRNA datasets, to assemble and align sets of orthologous genes, and for phylogenetic analyses of phylogenomic and morphological datasets, are available in the electronic supplementary material.

3. RESULTS

(a) *Phylogenomic analyses support Mandibulata*

To elucidate the phylogenetic position of myriapods and the discrepancy between recent analyses [12,16], we first analysed a phylogenomic dataset of 198 genes (corresponding to 40 100 reliably aligned amino acid positions) from 30 taxa (see figure 1). The dataset contains new sequences from the centipede *Strigamia maritima*. Bayesian analysis using the CAT + Γ model in the software package PHYLOBAYES [26] supports monophyly of Mandibulata with a posterior probability (PP) of 0.92 and a non-parametric bootstrap support (BS) value of 79 per cent. A Bayesian analysis using an even larger sampling of 59 taxa and the mixed CAT-general time reversible (GTR) + Γ model corroborates these findings (see the electronic supplementary material, figures S1 and S2). Furthermore, our analysis supports the monophyly of Chelicerata (Pycnogonida plus Arachnida), a close relationship between Branchiopoda and Hexapoda, monophyly of Arthropoda (Eurthropoda, Tardigrada and Onychophora), and a paraphyletic origin of the Cycloneuralia (Nematoda more closely related to Arthropoda than to Scalidophora). These relationships are further addressed in §3*e*.

(b) *Myriochelata is the result of a LBA artefact*

Our results are in accordance with those of Regier *et al.* [16], but in contradiction of other phylogenomic studies

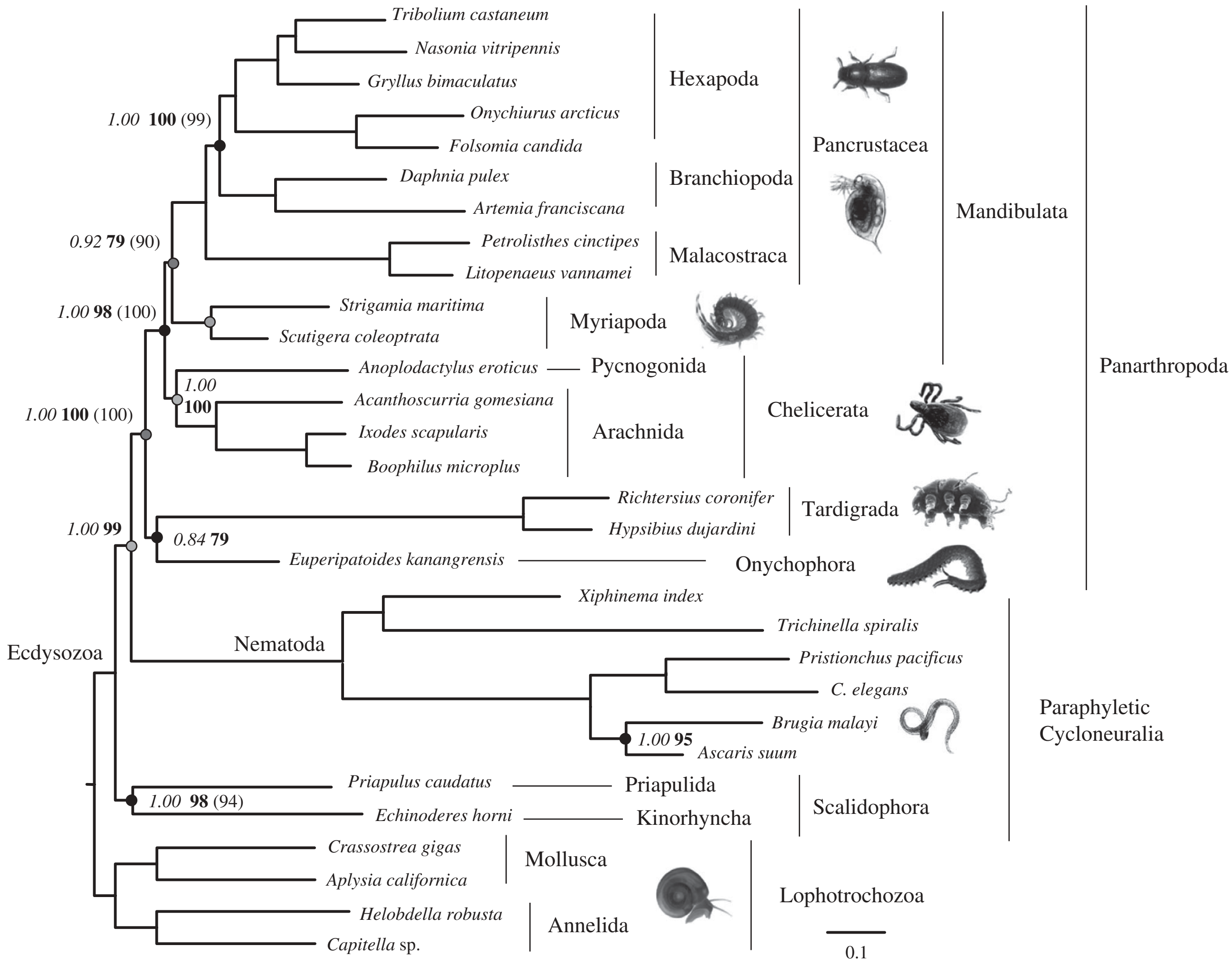


Figure 1. Phylogenomic analyses support Mandibulata, Arthropoda, Chelicerata and paraphyletic Cycloneuralia. Bayesian analyses using the CAT + Γ model. Values at nodes correspond to posterior probabilities (PP) (in italics) and bootstrap support (BS) from 100 pseudo-replicates (in bold); values in brackets are the BS for the same dataset reanalysed without the long branched Nematoda and Tardigrada lineages. Analyses support a monophyletic group of Mandibulata (Myriapoda, Hexapoda and Crustacea), a monophyletic group of Arthropoda (Eurthropoda, Tardigrada and Onychophora), monophyly of Chelicerata (Pycnogonida plus Euchelicerata) and a paraphyletic origin of the Cycloneuralia (Nematoda sister group of the Arthropoda). Where not shown, support values correspond to a PP of 1.00 and BS of 100 per cent. Images have been modified from http://commons.wikimedia.org.

[12,13,15]. We therefore explored whether systematic errors, in particular LBA, could have caused the discrepancy between our results and those of studies supporting Myriochelata. In this context, one notable aspect of the tree in figure 1 is the different branch lengths seen in various taxonomic groups. Pancrustacea have long branches in comparison to Myriapoda and Chelicerata, suggesting that in previous studies the fast evolving Pancrustacea could have been attracted towards the distant outgroup, resulting in the clustering of slowly evolving Myriapoda and Chelicerata owing to LBA. Because systematic errors, particularly LBA, become more apparent when the substitution model is unable to handle multiple substitutions correctly [14], we first asked how models such as Whelan and Goldman (WAG) + F + Γ and GTR + Γ—which assume homogeneity of the substitution process—fit our data. We find that WAG + F + Γ and GTR + Γ fit the data significantly less well than the heterogeneous CAT + Γ model (see the electronic supplementary material), and that this reduced fit is matched by reduction in support for Mandibulata over Myriochelata (figure 2*a* and electronic supplementary material, figure S3*a*).

We next explored the possible effects of LBA using a strategy of different taxon sampling. Logically, if Myriochelata is the result of an LBA artefact, exaggerating this source of error by using long-branched or evolutionarily distant outgroups will result in more support for this artefactual clade. Conversely, the use of the shortest branched outgroups should reduce the effects of LBA and result in lower support for Myriochelata. Both of these predictions are supported; when we used either the most phylogenetically distant outgroup (Lophotrochozoa, figure 2*b* and electronic supplementary material, figure S3*b*) or the fastest evolving ecdysozoan outgroup (Nematoda, figure 2*c* and electronic supplementary material, S3*c*), support decreases for Mandibulata and the artefactual group of slow evolving Myriapoda and Chelicerata (Myriochelata, in grey) increases. Equally, removal of these distant outgroups

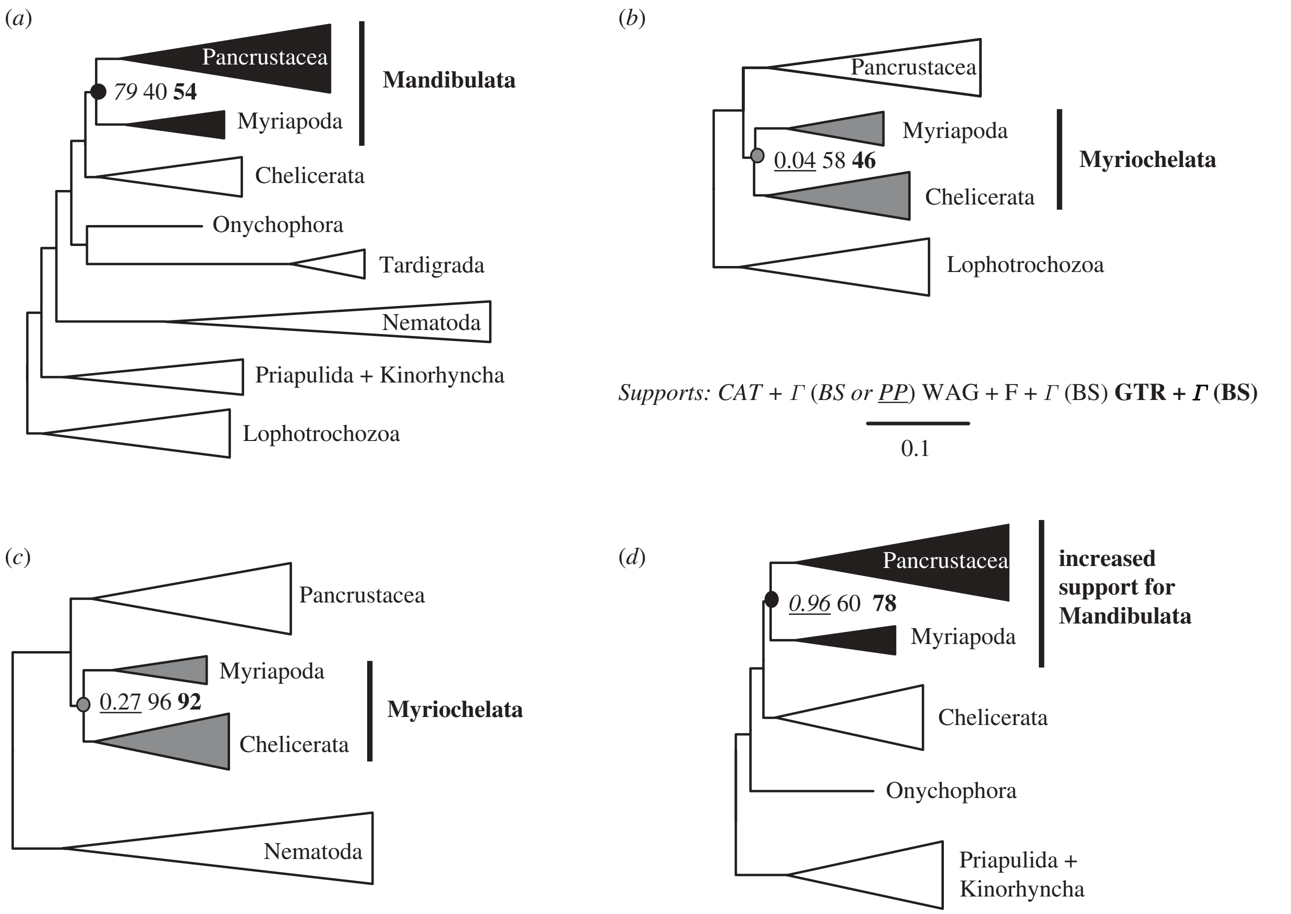


Figure 2. Taxon sampling and the artefactual nature of Myriochelata. Phylogenetic analyses of our 198 gene dataset using different taxon samples and both Bayesian and maximum likelihood inference. (*a*) Use of the less well fitting WAG + F + Γ and GTR + Γ homogeneous models results in lower support for Mandibulata (black node and lineages) compared to the best fitting CAT + Γ model (figure 1). The tree depicted is from the Bayesian CAT + Γ analyses. (*b*) Phylogenetically distant Lophotrochozoa and (*c*) fast evolving Nematoda outgroups exert an LBA with the fast evolving Tetraconata lineage, thereby regrouping slow evolving Myriapoda and Chelicerata (Myriochelata) (*d*) When using slowly evolving and phylogenetically close ecdysozoan outgroups, the support for Mandibulata increases. Trees *b*, *c* and *d* are the WAG + F + Γ maximum likelihood trees. Note that support for Mandibulata is high regardless of which outgroup is used when the dataset is analysed using best fitting model CAT + Γ, but significantly varies when using the less well fitting WAG + F + Γ and GTR + Γ models. Values at nodes are PPs from the Bayesian analyses using CAT + Γ model (PP in italics) BS from 100 replicates using the WAG + F + Γ (BS plain text) and GTR + Γ (BS in bold text) models. When not shown, the support is PP 1.00 and BS 100 per cent. Lineages have been collapsed for clarity with the length of triangles equal to the longest terminal branch in the collapsed lineage and stems are equal to the original length. Original trees with full support values are indicated in the electronic supplementary material, figure S3.

and their replacement with shorter branched taxa (e.g. Onychophora and Priapulida [27]) results in increased support for Mandibulata over Myriochelata (figure 2*d* and electronic supplementary material, figure S3*d*). We also performed a bootstrap analysis (under CAT + Γ) excluding the fast evolving nematodes and tardigrades, which found 90 per cent support for Mandibulata. Notably, both Lophotrochozoa and Nematoda contain species with divergent amino acid composition (see the electronic supplementary material, table S1), supporting our inference that they represent less suitable outgroups [19].

Using our phylogenomic dataset, we have shown that conditions which reduce LBA result in the highest support for Mandibulata, whereas conditions that increase LBA result in increased support for Myriochelata, implying the artefactual nature of the latter. We replicated these findings using the set of 150 genes of Dunn *et al.* [12], hereafter 'Dunn'. Reanalysis of a dataset using their original taxon sampling (of 16 ecdysozoans) resulted in strong support for Myriochelata (figure 3*a* and electronic supplementary material, figure S4*a*) in accordance with their original analysis. To test if the difference between our phylogeny (which supports Mandibulata) and that of Dunn (which favoured Myriochelata) is owing to taxonomic sampling we expanded their taxonomic representation to include all of our 30 taxa. Under these conditions, modest support for Mandibulata is obtained using the CAT + Γ model while support for Myriochelata decreased under WAG + F + Γ and GTR + Γ (figure 3*b* and electronic supplementary material, figure S4*b*). However, when we remove fast evolving outgroups the support for Mandibulata increases significantly (figure 3*c* and electronic supplementary material, figure S4*c*). Removal of fast evolving characters (see the electronic supplementary material, figure S5*a*) also

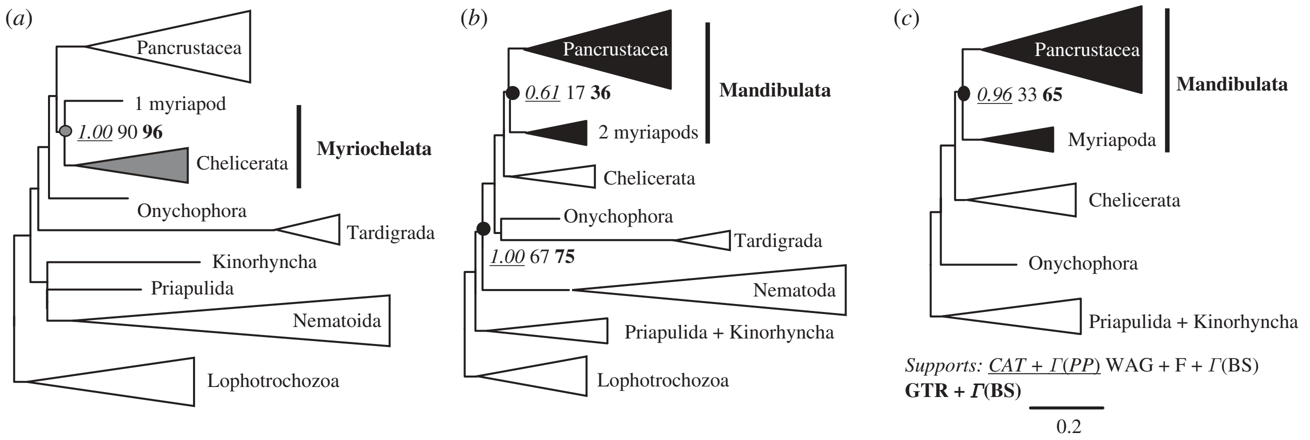


Figure 3. Support for Mandibulata from the gene set of Dunn *et al.* [12]. Bayesian and maximum likelihood analyses of the dataset of Dunn *et al.* [12]. (*a*) Using their original set of genes and taxa, Myriochelata is recovered with high support. (*b*) Using our taxon sampling (with the key addition of additional myriapod data) support for Myriochelata decreases and limited support for Mandibulata is recovered. (*c*) Support for Mandibulata increases when fast evolving or distant outgroups are excluded. Tree topologies correspond to the whole dataset Bayesian CAT + Γ trees. Values at nodes are PPs from the Bayesian analyses using CAT + Γ model (in italics and underscored) BS from 100 replicates using the WAG + F + Γ (plain text) and GTR + Γ (bold text) models. When not shown, the support is PP 1.00 and BS 100 per cent. Lineages have been collapsed for clarity with the length of triangles equal to the longest terminal branch in the collapsed lineage and stems are equal to the original length. Original trees with full support values are shown in the electronic supplementary material, figure S4.

results in support for Mandibulata instead of Myriochelata. Notably, even with identical taxonomic sampling our 198 gene set provides more support for Mandibulata than do the 150 genes of Dunn *et al.* (compare figures 2*c* and 3*c*). The difference may be partly explained by our dataset being larger and more complete (40 100 positions, 69% complete versus 18 829 positions, 61% complete), but also by the lower substitutional saturation of our genes (see the electronic supplementary material, figure S5*b*).

(c) *miRNAs corroborate Mandibulata, Euchelicerata and Myriapoda*

A useful way to test between the competing Mandibulata and Myriochelata phylogenetic hypotheses is to use an independent data source. We therefore explored the miRNA complements of key arthropod taxa using a combination of genomic sequence searches coupled with the generation and analysis of multiple small-RNA libraries. Novel miRNAs appear to have accumulated in animal genomes through time, and, although short, they show a level of sequence conservation exceeding that of ribosomal DNA [28], making it relatively easy to identify these novel miRNAs in descendant taxa. The apparent rarity of loss of miRNAs within evolutionary lineages coupled with the low likelihood of convergent evolution [29] makes miRNAs a valuable class of rare genomic characters in phylogenetics.

One miRNA, miR-965, had previously been found only in Pancrustacea and had been shown to be absent from the genome of the chelicerate *Ixodes scapularis* [28]. Importantly, we found reads of the mature miR-965 in the small RNA libraries of both myriapods (*Glomeris marginata* and *Scutigera coleoptata*), and also in the genome of the centipede *S. maritima* (figure 4). Screening our miRNA libraries also showed that in addition to being absent from the genomic sequence of the tick (*I. scapularis*), miR-965 could not be detected in the xiphosuran *Limulus polyphemus* or in the arachnid *Acanthoscurria chacoana*. Consequently, this distribution supports miR-965 as a genomic apomorphy (a rare genomic change) of the Mandibulata (figure 4). This same distribution is true of a second miRNA miR-282 that we have found only in insects, crustaceans and the centipedes *Strigamia* and *Scutigera*. miR-282 was not found in the *Glomeris* small RNA library and this may be because miR-282 is expressed at low levels in all Mandibulata sampled and the total number of reads and sequencing depth was relatively low in the *Glomeris* miRNA library.

In addition, upon screening the *L. polyphemus* and *A. chacoana* small-RNA libraries, we identified a novel chelicerate miRNA (Arthropod-Novel-1) that is not present in the Mandibulata, but is present in the genome of the tick *I. scapularis* (figure 4), and we thus suggest this miRNA to be a new genomic apomorphy for the Euchelicerata (Xiphosura and Arachnida). We have also identified a novel myriapod-specific miRNA (Arthropod-Novel-2) in the small-RNA libraries of *G. marginata* and *S. coleoptrata*, and in the genome of *S. maritima*, but not in the libraries or genomes of any other non-myriapod taxon analysed (figure 4). Further Myriapod-specific molecular synapomorphies have recently been described [30].

(d) *Updated morphological analyses support Mandibulata*

We assembled a large matrix of morphological data, which provides a third independent line of evidence in support of Mandibulata. While a number of possible morphological apomorphies of Myriochelata have recently been identified [17], inclusion of these characters in a cladistic analysis of 393 morphological characters still results in overall support for Mandibulata (Bremer support = 5)

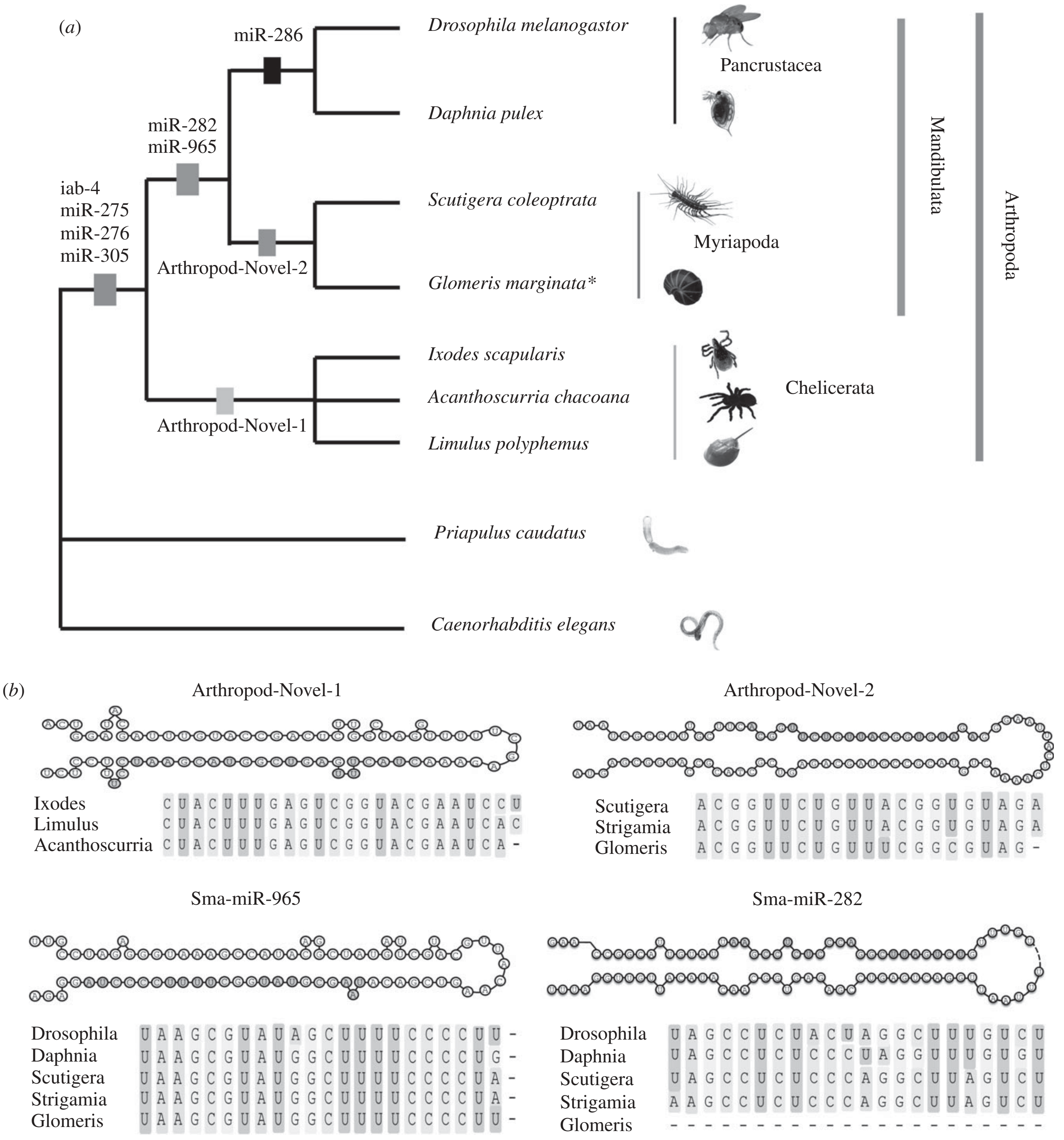

(a)
miR-286
miR-282
miR-965
iab-4
miR-275
miR-276
miR-305
Arthropod-Novel-2
Arthropod-Novel-1
Drosophila melanogastor
Daphnia pulex
Scutigera coleoptrata
Glomeris marginata*
Ixodes scapularis
Acanthoscurria chacoana
Limulus polyphemus
Priapulus caudatus
Caenorhabditis elegans
Pancrustacea
Myriapoda
Chelicerata
Mandibulata
Arthropoda
(b)
Arthropod-Novel-1
Ixodes C U A C U U U G A G U C G G U A C G A A U C C U
Limulus C U A C U U U G A G U C G G U A C G A A U C A C
Acanthoscurria C U A C U U U G A G U C G G U A C G A A U C A -
Arthropod-Novel-2
Scutigera A C G G U U C U G U U A C G G U G U A G A
Strigamia A C G G U U C U G U U A C G G U G U A G A
Glomeris A C G G U U C U G U U U C G G C G U A G -
Sma-miR-965
Drosophila U A A G C G U A U A G C U U U U C C C C U U -
Daphnia U A A G C G U A U G G C U U U U C C C C U G -
Scutigera U A A G C G U A U G G C U U U U C C C C U A -
Strigamia U A A G C G U A U G G C U U U U C C C C U A -
Glomeris U A A G C G U A U G G C U U U U C C C C U U -
Sma-miR-282
Drosophila U A G C C U C U A C U A G G C U U U G U C U
Daphnia U A G C C U C U C C C U A G G U U U G U G U
Scutigera U A G C C U C U C C C A G G C U U A G U C U
Strigamia A A G C C U C U C C C A G G C U U A G U C U
Glomeris -

Figure 4. miRNAs corroborate the monophyly of Mandibulata. (*a*) The monophyly of Mandibulata is supported by the presence of miR-965 and miR-282, also discovered in the genome of the centipede *Strigamia maritima*, and in the small RNA libraries of the millipede *Glomeris marginata* and the house centipede *Scutigera coleoptrata*. miR-965 and miR-282 are not known from any chelicerate or non-arthropod. N.B. miR-282 was not found in the small RNA library of *Glomeris*. (*b*) In addition a novel chelicerate miRNA (Arthropod-Novel-1) is present only in chelicerates, but in none of the mandibulates considered, and a novel myriapod miRNA (Arthropod-Novel-2) is found only in myriapods. Shaded residues highlight the mature miRNA sequence within the folded pre-miRNAs.

rather than Myriochelata, with or without the inclusion of fossil taxa (see figure 5 and electronic supplementary material). The Palaeozoic fossil taxa *Tanazios*, *Martinssonia*, and Trilobita (*Olenoides*) are resolved progressively more stemward relative to the mandibulate crown group. Although support values for the deep nodes in the mandibulate stem- and crown groups are weak when the fossils are included (Bremer values mostly 1 and jackknife frequencies mostly less than 50%), support for the mandibulate crown-group is increased when the analysis is confined to extant taxa because support is concentrated at a single node rather than broken up at series of nodes along the stem lineage.

Morphological support for Mandibulata includes complex similarities of head structure [31] and specifically of their mandibles, arrangements of midline neuropils in the brain, correspondences in cell numbers and specialized cell types in the ommatidia, similar sternal buds in the stomodeal region, and specific arrangements of serotonin-reactive neurons in the nerve cord (see the

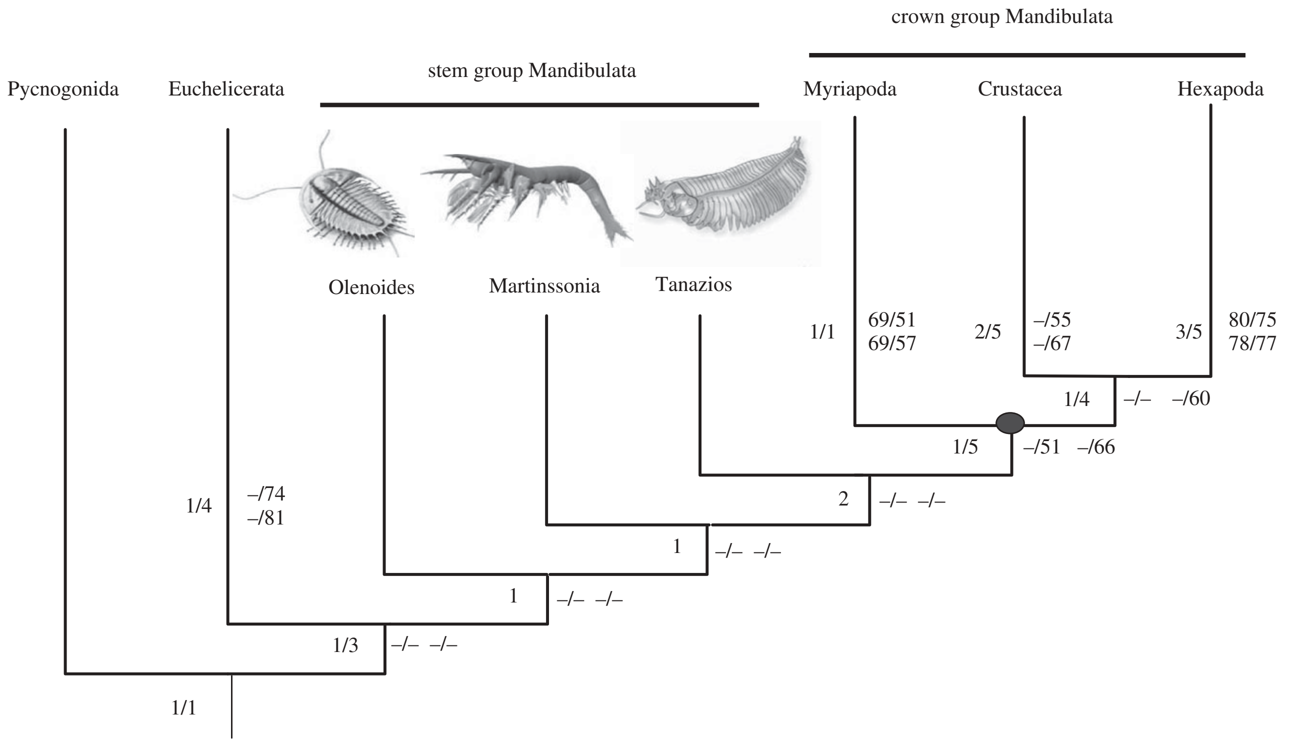


Figure 5. Morphology supports monophyly of crown Mandibulata. Summary cladogram of crown group euarthropod relationships based on morphological data (393 characters listed in the electronic supplementary material). Clades shown here are a strict consensus of shortest cladograms computed by TNT and PAUP*. Numbers to left of branches are Bremer support values; for extant taxa, values for analyses with (left) and without fossils (right) are separated by a slash. Numbers to right of branches are bootstrap (top) and jackknife (bottom) frequencies (indicated by a dash if less than 50%); values for analysis with and without fossils are separated by a slash. The fossils *Tanazios*, *Martinssonia*, and trilobites (*Olenoides*) are resolved progressively more stemward relative to the mandibulate crown group.

electronic supplementary material for a detailed compilation of morphological and developmental genetic characters).

(e) *Phylogenomic analyses support monophyletic Arthropoda, Chelicerata and Paraphyletic Cycloneuralia*

Most of our phylogenomic analyses support the monophyly of Arthropoda (euarthropods, tardigrades, onychophorans), either using our gene sampling (figure 1) or that of Dunn (figure 3*b*). The position of tardigrades is more unstable, varying from being sister to the onychophorans (figure 1 using CAT + Γ model) to being sister to a group of arthropods plus onychophorans (see the electronic supplementary material, figure S2 using the CAT + GTR model). Whereas the CAT + Γ model supports Arthropoda consistently, site-homogeneous WAG + F + Γ and GTR + Γ models tend to group tardigrades with nematodes (dotted arrows in the electronic supplementary material, figures S3 and S4). Our interpretation is that site-homogeneous models, which fit our data less well than the CAT model (see §2), are unable to overcome the effect of systematic errors responsible for the grouping of fast evolving nematodes and tardigrades.

All our phylogenomic analyses support a monophyletic origin of the chelicerates in which pycnogonids are sister to a group of arachnids plus Xiphosura. This finding is significant in light of recent debates over the position of the Pycnogonida, which some studies find to be the sister group to all other arthropods, a hypothesis known as Cormogonida [23,32,33]. The possibility that systematic/stochastic errors were affecting the affinity of pycnogonids in previous studies is highlighted by their position being parameter-dependent in other studies [16,24,34].

Finally, all our phylogenomic analyses support a paraphyletic origin of the Cycloneuralia, with the Scalidophora (priapulids and kinorhynchs) sister to a group of nematodes plus arthropods. This is in accordance with ribosomal markers [23], but in contrast to previous phylogenomic studies [12,13], which instead supported monophyly of Cycloneuralia (Nematodoida + Scalidophora). Notably, when updating the gene selection of Dunn *et al.* [12] to our larger taxon sampling, a paraphyletic origin of the Cycloneuralia is recovered. Ultimately, the relationships of Nematodoida, Scalidophora and Arthropoda remain uncertain.

4. DISCUSSION

Arguably the strongest evidence of phylogenetic accuracy is the congruence of independent lines of evidence supporting the same tree topology [22,35]. In order to test current hypotheses of arthropod evolution, we have analysed three independent lines of evidence: a phylogenomic dataset of 198 genes, a new miRNA dataset and a large morphological dataset. All three datasets unambiguously support the monophyly of Mandibulata.

We have examined the possibility that previous molecular phylogenies supporting Myriochelata might have been affected by systematic error and the robustness of the result from our phylogenomic dataset is supported by experiments designed to reduce the effects of systematic errors. Increased taxon sampling, exclusion of outgroups with the longest branches, removal of the fastest evolving positions and the use of better evolutionary models systematically increase support for Mandibulata over Myriochelata.

The presence of miR-965 and miR-282 in Pancrustacea and in two groups of Myriapoda also represents compelling evidence in support of Mandibulata. These two miRNA are absent from both arachnids and horseshoe crabs as well as from all other Ecdysozoans for which the miRNA complement is known (nematodes and priapulids worms). As it is implausible for this miRNA to have been independently acquired in the different mandibulate lineages [29], we conclude that it constitutes a rare genomic change supporting Mandibulata. In light of congruence of these novel miRNA autapomorphies with other lines of evidence presented here (phylogenomics and morphology) and with the complementary findings of Regier *et al.* [16], we conclude that the most tenable position of the Myriapoda is as the sister group of the Pancrustacea within a monophyletic Mandibulata.

Our phylogenomic analyses suggest that studies which have grouped tardigrades with nematodes may have been similarly affected by LBA. When analysed using the CAT model, which has been shown to help in overcoming systematic errors [14], both our dataset and that of Dunn *et al.* [12] group Tardigrada with Euarthropoda and Onychophora in a monophyletic Arthropoda clade. Tardigrada are a sister group of the Onychophora in these trees, a topology which finds no support from a morphological point of view, but is in accordance with mitochondrial markers [36]. Furthermore, if the paraphyletic nature of the Cycloneuralia is correct, as supported by our phylogenomic analyses, this would suggest that the ancestral Ecdysozoa was cycloneuralian-like, possessing a circumpharyngeal brain and an introvert [37].

The Mandibulata, which includes insects, is by far the largest clade of animals on Earth, but the origin of this successful bodyplan in terms of the evolution of its development remains obscure. The picture from palaeontology is, however, somewhat clearer. Cambrian fossils that have been identified as a grade of stem-group mandibulates [38] indicate a crustacean-like *habitus* for basal members of the Mandibulata and may shed light on how the mandible common to these groups evolved. The limb on the third cephalic segment (the mandible homologue) in Cambrian stem-group mandibulates such as *Martinssonia* displays a stronger development of a movable, setose process at the limb base (‘proximal endite’; [39]) than that on the adjacent limbs [40]. The more elaborated proximal endite used for food manipulation is viewed as a precursor to the fully differentiated coxal chewing surface in the mandibulate crown group [40]. Further studies of fossils and embryos in the light of what we suggest is a reliable phylogeny of arthropod classes should clarify the evolution of the mandibulate bodyplan [41], and consequently how anatomical novelties may have promoted their hugely successful radiation.

We thank Michael Akam and Ariel Chipman for the *Strigamia maritima* cDNA library, Baylor College of Medicine Human Genome Sequencing Center for access to the *Strigamia* genome data and Alexandros Stamatakis for a RAxML version implementing the GTR model for amino acids. D.P. and L.C. are supported by a Science Foundation Ireland RFP grant to D.P. (08/RFP/EOB1595), and by a Science Foundation Ireland Short Term Travel Fellowship to D.P. O.R.S. was supported by the Zoonet Marie Curie Training Network to M.J.T. and by a Science Foundation Ireland FRP grant (08/RFP/EOB1595) to D.P. S.J.L. is supported by an IRCSET fellowship. K.J.P. is supported by the National Science Foundation. H.P. gratefully acknowledges the financial support provided by NSERC, the Canadian Research Chair Program and the Université de Montréal, and the Réseau Québecois de Calcul de Haute Performance for computational resources.

REFERENCES

1 Telford, M. J., Bourlat, S. J., Economou, A. D., Papillon, D. & Rota-Strabelli, O. 2008 The evolution of the Ecdysozoa. *Phil. Trans. R. Soc. B* **363**, 1529–1537. (doi:10.1098/rstb.2007.2243)

2 Edgecombe, G. D. 2010 Arthropod phylogeny: an overview from the perspectives of morphology, molecular data and the fossil record. *Arthropod Struct. Dev.* **39**, 74–87. (doi:10.1016/j.asd.2009.10.002)

3 Budd, G. E. & Telford, M. J. 2009 The origin and evolution of arthropods. *Nature* **417**, 812–817. (doi:10.1038/nature07890)

4 Nielsen, C. 1995 *Animal evolution. Interrelationships of the living phyla*, 2nd edn. Oxford, UK: Oxford University Press.

5 Telford, M. J. & Thomas, R. H. 1995 Demise of the Atelocerata? *Nature* **376**, 123–124. (doi:10.1038/376123a0)

6 Boore, J. L., Lavrov, D. V. & Brown, W. M. 1998 Gene translocation links insects and crustaceans. *Nature* **392**, 667–668. (doi:10.1038/33577)

7 Dohle, W. 2001 Are the insects terrestrial crustaceans? A discussion of some new facts and arguments and the proposal of a proper name ‘Tetraconata’ for the monophyletic unit Crustacea + Hexapoda. *Ann. Soc. Entomol. France* **37**, 85–103.

8 Cook, C. E., Smith, M. L., Telford, M. J., Bastianello, A. & Akam, M. 2001 Hox genes and the phylogeny of the arthropods. *Curr. Biol.* **11**, 759–763. (doi:10.1016/S0960-9822(01)00222-6)

9 Pisani, D., Poling, L. L., Lyons-Weiler, M. & Hedges, S. B. 2004 The colonization of land by animals: molecular phylogeny and divergence times among arthropods. *BMC Biol.* **2**, 1. (doi:10.1186/1741-7007-2-1)

10 Friedrich, M. & Tautz, D. 1995 rDNA phylogeny of the major extant arthropod classes and the evolution of myriapods. *Nature* **376**, 165–167. (doi:10.1038/376165a0)

11 Hwang, U.-W., Friedrich, M., Tautz, D., Park, C. J. & Kim, W. 2001 Mitochondrial protein phylogeny joins myriapods with chelicerates. *Nature* **413**, 154–157. (doi:10.1038/35093090)

12 Dunn, C. W. *et al.* 2008 Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **452**, 745–749. (doi:10.1038/nature06614)

13 Hejnol, A. *et al.* 2009 Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc. R. Soc. B* **276**, 4261–4270. (doi:10.1098/rspb.2009.0896)

14 Lartillot, N. & Philippe, H. 2008 Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. *Phil. Trans. R. Soc. B* **363**, 1463–1472. (doi:10.1098/rstb.2007.2236)

15 Roeding, F., Borner, J., Kube, M., Klages, S., Reinhardt, R. & Burmester, T. 2009 A 454 sequencing approach for

large scale phylogenomic analysis of the common emperor scorpion (*Pandinus imperator*). *Mol. Phylogenet. Evol.* **53**, 826–834. (doi:10.1016/j.ympev.2009.08.014)

16 Regier, J. C., Shultz, J. W., Zwick, A., Hussey, A., Ball, B., Wetzer, R., Martin, J. W. & Cunningham, C. W. 2010 Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature* **463**, 1079–1083. (doi:10.1038/nature08742)

17 Mayer, G. & Whitington, P. M. 2009 Velvet worm development links myriapods with chelicerates. *Proc. R. Soc. B* **276**, 3571–3579. (doi:10.1098/rspb.2009.0950)

18 Stollewerk, A. & Chipman, A. D. 2006 Neurogenesis in myriapods and chelicerates and its importance for understanding arthropod relationships. *Integr. Comp. Biol.* **46**, 195–206. (doi:10.1093/icb/icj020)

19 Rota-Stabelli, O. & Telford, M. J. 2008 A multi criterion approach for the selection of optimal outgroups in phylogeny: recovering some support for Mandibulata over Myriochelata using mitogenomics. *Mol. Phylogenet. Evol.* **48**, 103–111. (doi:10.1016/j.ympev.2008.03.033)

20 Brinkmann, H. & Phillipe, H. 1999 Archaea sister group of bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. *Mol. Biol. Evol.* **16**, 817–825.

21 Philippe, H., Delsuc, F., Brinkmann, H. & Lartillot, N. 2005 Phylogenomics. *Annu. Rev. Ecol. Evol. Syst.* **36**, 541–562. (doi:10.1146/annurev.ecolsys.35.112202.130205)

22 Philippe, H., Lartillot, N. & Brinkmann, H. 2005 Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Mol. Biol. Evol.* **22**, 1246–1253. (doi:10.1093/molbev/msi111)

23 Mallatt, J. & Giribet, G. 2006 Further use of nearly complete 28S and 18S rRNA genes to classify Ecdysozoa: 37 more arthropods and a kinorhynch. *Mol. Phylogenet. Evol.* **40**, 772–794. (doi:10.1016/j.ympev.2006.04.021)

24 Regier, J. C. *et al.* 2008 Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. *Syst. Biol.* **57**, 920–938. (doi:10.1080/10635150802570791)

25 Janssen, R., Eriksson, B. J., Budd, G. E., Akam, M. & Prpic, N.-M. 2010 Gene expression patterns in an onychophoran reveal that regionalization predates limb segmentation in pan-arthropods. *Evol. Dev.* **12**, 363–372. (doi:10.1111/j.1525-142X.2010.00423.x)

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

27 Webster, B. L., Copley, R. R., Jenner, R. A., Mackenzie-Dodds, J. A., Bourlat, S. J., Rota-Stabelli, O., Littlewood, D. T. J. & Telford, M. J. 2006 Mitogenomics and phylogenomics reveal priapulid worms as extant models of the ancestral Ecdysozoan. *Evol. Dev.* **8**, 502–510. (doi:10.1111/j.1525-142X.2006.00123.x)

28 Wheeler, B. M., Heimberg, A. M., Moy, V. N., Sperling, E. A., Holstein, T. W., Heber, S. & Peterson, K. J. 2009 The deep evolution of metazoan microRNAs. *Evol. Dev.* **11**, 50–68. (doi:10.1111/j.1525-142x.2008.00302.x)

29 Sperling, E. A. & Peterson, K. J. 2009 MicroRNAs and metazoan phylogeny: big trees from little genes. In *Animal evolution: genomes, fossils, and trees* (eds M. J. Telford & D. T. J. Littlewood), pp. 157–170. Oxford, UK: Oxford University Press.

30 Janssen, R. & Budd, G. E. 2010 Gene expression suggests conserved aspects of Hox gene regulation in arthropods and provides additional support for monophyletic Myriapoda. *EvoDevo* (doi:10.1186/2041-9139-1-4)

31 Scholtz, G. & Edgecombe, G. D. 2006 The evolution of arthropod heads: reconciling morphological, developmental and palaeontological evidence. *Dev. Genes Evol.* **216**, 395–415. (doi:10.1007/s00427-006-0085-4)

32 Giribet, G., Edgecombe, G. D. & Wheeler, W. C. 2001 Arthropod phylogeny based on eight molecular loci and morphology. *Nature* **413**, 157–161. (doi:10.1038/35093097)

33 Maxmen, A., Browne, W. E., Martindale, M. Q. & Giribet, G. 2005 Neuroanatomy of sea spiders implies an appendicular origin of the protocerebral segment. *Nature* **437**, 1144–1148. (doi:10.1038/nature03984)

34 Podsiadlowski, L. & Braband, A. 2006 The mitochondrial genome of the sea spider *Nymphon gracile* (Arthropoda: Pycnogonida). *BMC Genom.* **7**, 284. (doi:10.1186/1471-2164-7-284)

35 Pisani, D., Benton, M. J. & Wilkinson, M. 2007 Congruence of morphological and molecular phylogenies. *Acta Biotheoretica* **55**, 269–281. (doi:10.1007/s10441-007-9015-8)

36 Rota-Stabelli, O., Kayal, E., Gleeson, D., Daub, J., Boore, J., Pisani, D., Blantor, M. & Lavrov, D. V. 2010 Ecdysozoan mitogenomics: evidence for a common origin of the legged invertebrates, the Panarthropoda. *Genom. Biol. Evol.* **2**, 425–440. (doi:10.1093/gbe/evq030)

37 Garey, J. R. 2001 Ecdysozoa: the relationship between Cycloneuralia and Panarthropoda. *Zool. Anz.* **240**, 321–330. (doi:10.1078/0044-5231-00039)

38 Richter, S. & Wirkner, S. 2004 Kontroversen in der phylogenetische systematik der Euarthropoda. In *Kontroversen in der phylogenetischen systematik der Metazoa.* (eds S. Richter & W. Sudhaus), pp. 73–102. Berlin, Germany: Sitzungs-Berichte der Gesellschaft Naturforschender Freunde zu Berlin.

39 Waloszek, D., Maas, A., Chen, J. & Stein, M. 2007 Evolution of cephalic feeding structures and the phylogeny of Arthropoda. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **254**, 273–287. (doi:10.1016/j.palaeo.2007.03.027)

40 Zhang, X. G., Siveter, D. J., Waloszek, D. & Maas, A. 2007 An epipodite-bearing crown-group crustacean from the Lower Cambrian. *Nature* **449**, 595–598. (doi:10.1038/nature06138)

41 Telford, M. J. & Budd, G. E. 2003 The place of phylogeny and cladistics in Evo-Devo research. *Int. J. Dev. Biol.* **47**, 479–490.

MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda

Lahcen I. Campbell[a,1], Omar Rota-Stabelli[a,1,2], Gregory D. Edgecombe[b], Trevor Marchioro[c], Stuart J. Longhorn[a], Maximilian J. Telford[d], Hervé Philippe[e], Lorena Rebecchi[c], Kevin J. Peterson[f,3], and Davide Pisani[a,3]

[a]Department of Biology, The National University of Ireland, Maynooth, Kildare, Ireland; [b]Department of Palaeontology, The Natural History Museum, London SW7 5BD, United Kingdom; [c]Dipartimento di Biologia, Università di Modena e Reggio Emilia, 41125 Modena, Italy; [d]Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, United Kingdom; [e]Centre Robert-Cedergren, Département de Biochimie, Université de Montréal, Montréal, QC, Canada H3C3J7; and [f]Department of Biology, Dartmouth College, Hanover, NH 03755

Edited* by James W. Valentine, University of California, Berkeley, CA, and approved August 11, 2011 (received for review April 7, 2011)

Morphological data traditionally group Tardigrada (water bears), Onychophora (velvet worms), and Arthropoda (e.g., spiders, insects, and their allies) into a monophyletic group of invertebrates with walking appendages known as the Panarthropoda. However, molecular data generally do not support the inclusion of tardigrades within the Panarthropoda, but instead place them closer to Nematoda (roundworms). Here we present results from the analyses of two independent genomic datasets, expressed sequence tags (ESTs) and microRNAs (miRNAs), which congruently resolve the phylogenetic relationships of Tardigrada. Our EST analyses, based on 49,023 amino acid sites from 255 proteins, significantly support a monophyletic Panarthropoda including Tardigrada and suggest a sister group relationship between Arthropoda and Onychophora. Using careful experimental manipulations—comparisons of model fit, signal dissection, and taxonomic pruning—we show that support for a Tardigrada + Nematoda group derives from the phylogenetic artifact of long-branch attraction. Our small RNA libraries fully support our EST results; no miRNAs were found to link Tardigrada and Nematoda, whereas all panarthropods were found to share one unique miRNA (miR-276). In addition, Onychophora and Arthropoda were found to share a second miRNA (miR-305). Our study confirms the monophyly of the legged ecdysozoans, shows that past support for a Tardigrada + Nematoda group was due to long-branch attraction, and suggests that the velvet worms are the sister group to the arthropods.

Ecdysozoa | cycloneuralia | Lobopodia | Tactopoda

Ecdysozoa (1) is the clade of molting invertebrates that includes two of the ecologically most important and evolutionarily most successful animal phyla—the arthropods and the nematodes—as well as several other, less diversified taxa, including the tardigrades (water bears), the onychophorans (velvet worms), and the priapulids (penis worms). Although the monophyly of Ecdysozoa is now well established (2, 3), the phylogenetic relationships within this group have proven difficult to resolve (4–7). Morphological and embryological evidence suggests a close affinity among Arthropoda, Onychophora, and Tardigrada (the Panarthropoda) (8, 9), although the interrelationships among these three taxa are uncertain. Despite the concordance between these morphological studies and a few molecular analyses (10–14), most molecular studies instead support a close relationship between the water bears and the cycloneuralian ecdysozoans (nematodes, priapulids, and their close relatives), particularly the nematodes (2, 15–22). These alternative hypotheses of tardigrade relationships have important consequences for our understanding of morphological evolution within Ecdysozoa. For example, if tardigrades are cycloneuralians, then the telescopic mouth cone and plated pharynx shared by tardigrades and cycloneuralians should be considered cycloneuralian apomorphies, whereas the important characteristics of segmentation and the possession of paired limbs must be homoplastic—they either evolved convergently in arthropods and tardigrades or were lost in nematodes (23). Obviously, the opposite would be true if the tardigrades are panarthropods. Thus, accurately placing the tardigrades with respect to nematodes and arthropods is central to solving the interrelationships among the ecdysozoans and clarifying homologies within this group.

Although the rapidly growing influx of molecular data has radically altered our understanding of the animal tree of life, no dataset is homoplasy-free. Phylogenies derived from large, genomic-scale datasets of expressed sequence tags (ESTs) from many proteins minimize stochastic errors; however, they can exacerbate systematic errors (24), such as the well-known long-branch attraction (LBA) artifact (25). This is because systematic errors, unlike stochastic ones, are positively misleading; the error increases with an increase in the amount of data in the analysis (24). Although genomic-scale datasets are important for resolving difficult phylogenetic problems, suboptimal approaches to tree reconstruction, such as those using poorly fitting substitution models, can generate phylogenetic artifacts when applied to such datasets. Tools have been developed to ameliorate these problems, including comparing trees derived using differently fitting models (13, 14, 26), site-stripping (e.g., "slow-fast" analyses; ref. 27), signal dissection (28), and targeted taxon pruning (3, 26, 29). These tools have recently been applied to address, for example, the position of the Myriapoda (centipedes and their relatives) within Arthropoda (12, 14, 20, 30) and the position of the Ctenophora (comb jellies) among the nonbilaterian animals (12, 26, 31, 32).

Given the inherent difficulties and potential biases associated with the analyses of genome-scale datasets, the use of a single type of data might not be sufficient to solve a particularly difficult phylogenetic problem (33). We have contended that consilience (34)—the congruence of multiple lines of evidence—is a particularly cogent indicator of phylogenetic accuracy (14, 35, 36). A

Author contributions: K.J.P., and D.P. designed research; L.I.C., O.R.-S., S.J.L., K.J.P., and D.P. performed research; K.J.P. and D.P. contributed new reagents/analytic tools; G.D.E. contributed morphological and paleontological expertise as needed; T.M. cultured, classified, and selected animals (Tardigrada); L.I.C., O.R.-S., L.R., K.J.P., and D.P. analyzed data; and L.I.C., O.R.-S., G.D.E., T.M., M.J.T., H.P., L.R., K.J.P., and D.P. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

[1]L.I.C. and O.R.-S. contributed equally to this work.

[2]Present address: Research and Innovation Center, Fondazione Edmund Mach, San Michele all'Adige, Italy.

[3]To whom correspondence may be addressed. E-mail: kevin.j.peterson@dartmouth.edu or davide.pisani@nuim.ie.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1105499108/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1105499108

class of molecules whose utility for phylogenetic reconstruction has recently been recognized is the microRNAs (miRNAs), genomically encoded nonprotein coding RNAs of approximately 22 nucleotides in length that are found in many eukaryotes, including the metazoans (37, 38). MiRNAs are important post-transcriptional regulators (39), but it is their use as phylogenetic markers that is of interest here. MiRNAs have four properties that make them reliable indicators of phylogenetic relationships: (*i*) New miRNA families are continually added through time to evolving metazoan genomes; (*ii*) once a new miRNA is acquired, its mature sequence accumulates mutations only very slowly; (*iii*) the rate of miRNA acquisition outweighs the rate of miRNA losses in most metazoan taxa; and (*iv*) there is a low probability of convergent evolution of an miRNA gene (38, 40). Indeed, the use of miRNAs has already provided important insights into the interrelationships among annelids (41), sponges (42), arthropods (14) vertebrates (43), and brachiopods (44), and has helped place enigmatic taxa, such as acoel flatworms, into the animal tree of life (36).

In the present study, we investigated the phylogenetic relationships of the Tardigrada within Ecdysozoa by studying the consilience of two independent genomic datasets, ESTs and miRNAs. We first present our EST results and use these to ask whether alternative hypotheses of tardigrade relationships (arthropod vs. nematode affinity), as found in previous phylogenomic analyses, could be tree-reconstruction artifacts. We then assembled the miRNAs complements of a tardigrade and an onychophoran, and compare these with the miRNA complements of all other known metazoans. Finally, we compare the results of our EST and miRNA analyses to evaluate the extent to which these genomic markers corroborate or, alternatively, disagree with each other. These lines of evidence support the monophyly of Panarthropoda including Tardigrada. We show that support from previous studies for a nematode+tardigrade group is the result of an LBA artifact, and provide evidence that Onychophora is the sister group of Arthropoda. These results imply that panarthropod limbs and segmentation are homologous, and that characters shared by tardigrades, nematodes, and other cycloneuralians are ecdysozoan plesiomorphies.

Results

EST-Based Phylogenomic Analyses Support Panarthropoda and Lobopodia. To address the phylogenetic position of tardigrades, we assembled a dataset of 255 genes (49,023 reliably aligned amino acid positions) from all of the ecdysozoan phyla except the Loricifera. Because the use of poor-fitting models can cause the recovery of artifactual phylogenies, we first used Bayesian cross-validation (45) to rank substitution models according to their fit to our alignment. Results of our cross-validation analysis (Fig. S1) show a regular increase in the fit of the model to the data when moving from simple to more complex models, with the site-heterogeneous mixture model CAT-GTR+Γ having the best fit to our dataset. (All models tested used a gamma distribution to account for rate variation among sites.) Results of the Bayesian analyses performed using the CAT-GTR+Γ model are shown in Fig. 1*A*. The majority of internal nodes have a posterior probability (PP) = 1. Tardigrada is recovered within Panarthropoda as the sister group of Onychophora + Arthropoda, together called the Lobopodia (46), with PP = 1. Within Arthropoda, our analyses confirm the chelicerate affinity of the sea spiders and are consistent with the monophyly of Mandibulata (Myriapoda + Pancrustacea) (14, 30).

Our results do not support the monophyly of the Cycloneuralia, given that Nematoida (Nematoda + Nematomorpha) is recovered as the sister group of Panarthropoda, albeit with a low posterior probability (PP = 0.76), whereas Scalidophora

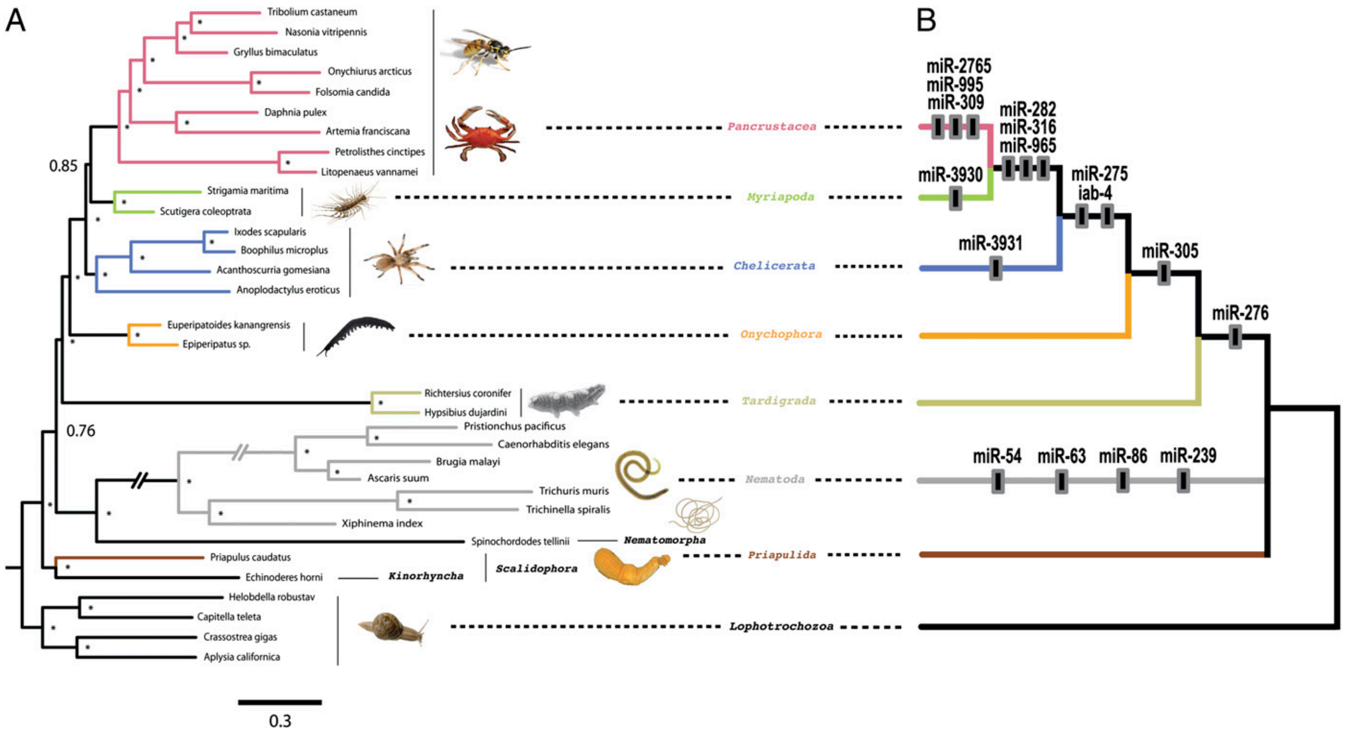


Fig. 1. Phylogenomics and miRNAs suggest velvets worm are the sister group to the arthropods within a monophyletic Panarthropoda. (*A*) Phylogenetic tree derived using Bayesian analysis of the EST data under the best-fitting CAT-GTR+Γ model supports tardigrades as the sister group of Lobopodia (Onychophora + Arthropoda). Support values represent posterior probabilities. Asterisks indicate a PP value of 1.0. Note that for Nematoda alone, the branch lengths are not shown to scale. (*B*) MiRNA distribution is consistent with the results obtained from the phylogenomic analysis. Single gray/black rectangles represent a miRNA gain. Clades are color-coded to highlight congruence between ESTs and miRNAs (see text for more details).

(Priapulida + Kinorhyncha) is recovered as the sister group of all other ecdysozoans. Nematoida was recovered with PP = 1. Because Nematomorpha is the taxon with the greatest amount of missing data in our EST dataset (Table S1), the strong support found for Nematoida (an otherwise well-accepted clade) suggests that missing data for Nematomorpha do not have a negative impact on our results.

Model Selection, Signal Dissection, and Targeted Taxonomic Pruning Highlight the Artifactual Nature of Tardigrada + Nematoda. To better understand the nature of the signal in our EST dataset, we performed three experiments to test whether the Tardigrada + Nematoda group recovered in previous analyses (2, 15–22) could result from a systematic error. First, Bayesian analyses were performed under a series of alternative models (Figs. S1 and S2). When the data were analyzed under poor-fitting site-homogenous models (i.e., WAG+Γ and GTR+Γ) (Fig. 2*A* and Figs. S1 *A* and *B* and S2 *A* and *B*), Panarthropoda was not recovered, and instead Tardigrada was found as the sister group of Nematoida (PP = 1 with both models). In contrast, analyses using the better-fitting site-heterogeneous CAT+Γ and CAT-GTR+Γ invariably identified Tardigrada as a member of Panarthropoda (Fig. 1*A* and Figs. S1 *C* and *D* and S2 *C* and *D*).

We next performed a signal-dissection analysis (13, 28), based on the slow-fast technique (27). We partitioned sites into subsets according to their rate of evolution, and independently analyzed these partitions. We hypothesized that if Tardigrada + Nematoda were an LBA artifact, then support for this group would be favored by the partitions of fast-evolving sites, whereas it would be minimized in partitions that exclude these sites (*Methods*). Consistent with our hypothesis, analyses of the fast-evolving sites show Nematoda + Tardigrada with PP = 0.88, whereas analyses of the slow-evolving sites show Tardigrada + Lobopodia with PP = 0.84 (Fig. 2 *B* and *C*, Fig. S3, and Table S2).

To further test whether Tardigrada + Nematoda is an LBA artifact, we performed a series of taxon pruning experiments. We selectively removed taxa to generate uninterrupted long-branches for Tardigrada, Onychophora, and Nematoda (*Methods*). As expected if Tardigrada + Nematoda is an LBA artifact, the results systematically support this group (Fig. 2*D* and Fig. S4).

In summary, three different experiments designed to uncover potential sources of systematic bias in our EST alignment suggest that a nematode (or cycloneuralian) affinity for Tardigrada is most likely an LBA artifact.

MiRNAs Corroborate the EST-Based Phylogenomic Analyses, and Confirm the Monophyly of Panarthropoda and Lobopodia. Our second dataset derives from the newly sequenced small RNA complements of the tardigrade *Paramacrobiotus* cf. *richtersi* and the onychophoran *Peripatoides novaezelandiae*, and characterization of their respective miRNA complements. Rota-Stabelli et al. (14) identified four miRNAs that characterize arthropods and had not yet been found in other ecdysozoans: miR-275, -276, -305, and -iab-4. There are also four miRNAs that are conserved between the nematode genera *Caenorhabditis* and *Pristionchus* (47): miR-54, -63, -86, and -239 (Fig. 1*B*). Consistent with our EST results, we did not find any nematode miRNAs in our tardigrade small-RNA library. Similarly, we did not find any potential miRNAs shared exclusively between the tardigrade and the onychophoran. Instead, in both the tardigrade and onychophoran libraries we found a single miRNA, miR-276, that formerly had been identified only in arthropods (14). In addition, in the onychophoran library, but not in the tardigrade library, we found a second miRNA, miR-305, which is also considered arthropod-specific (Fig. 1*B*). Based on these discoveries, we hypothesize that miR-276 is an apomorphy of Panarthropoda (Tardigrada + Lobopodia) and miR-305 is an apomorphy of Lobopodia (Onychophora + Arthropoda). Finally, our results suggest that miR-275 and miR-iab-4 are apomorphies of Arthropoda (Fig. 1*B*).

Discussion

Given the pervasiveness of systematic artifacts, care must be taken when evaluating topologies derived from large alignments, especially when well-supported competing hypotheses have been proposed. In the case of the tardigrades, molecular homoplasy certainly exists, as demonstrated by the fact some molecular studies support a nematode affinity of tardigrades, whereas others support an arthropod affinity. With respect to morphology, tardigrades have a melange of arthropod and cycloneuralian characters, suggesting that either the arthropod-like characters were lost in cycloneuralians or cycloneuralian-like characters

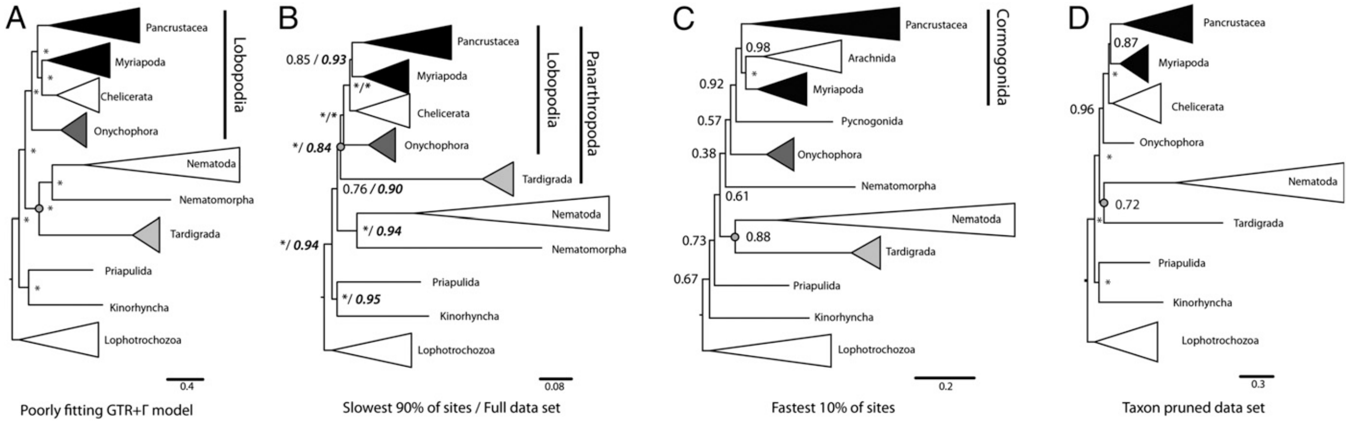


Fig. 2. Model selection, signal dissection and taxon pruning experiments show LBA explains previous support for a tardigrade/nematode clade. As in Fig. 1, these are trees from the EST data; node values represent posterior probabilities, and asterisks indicate a PP of 1.0. The node where the Tardigrada join the tree is identified by a circle. Clades have been collapsed for clarity. (*A*) Tardigrades are recovered as the sister group of Nematoida under the poorly fitting GTR+Γ model of sequence evolution (for Δ-likelihoods and SDs; Fig. S1 and *Methods*). (*B*) Tree recovered from the analysis of the slowest-evolving 90% of the sites in our dataset (Fig. S3*A*). The PP values are reported in italics, whereas support values obtained from the analysis of the complete dataset are in roman type (Fig. 1*A*). (*C*) Topology recovered from the 10% fastest evolving sites in our dataset, under CAT-GTR+Γ. The fast-evolving sites support Tardigrada as the sister group of Nematoda. (*D*) Phylogeny generated under a reduced-taxon set (one onychophoran, one tardigrade, and no nematomorph) designed to exacerbate LBA artifact.

were lost in arthropods (assuming that cycloneuralian and tardigrade characters are homologous). Consilience between our EST and miRNA analyses, as well as the experiments performed to identify LBA artifacts, congruently suggest that the closest affinity of tardigrades is with the Arthropoda and the Onychophora (i.e., Panarthropoda), not with the cycloneuralian ecdysozoans (nematodes). These results supersede our previous mitogenomic analyses (13), which could not reject a nematode affinity of Tardigrada because of the extremely high evolutionary rate of nematode mitochondrial genomes. The arthropod-like features of tardigrades, such as the paired ventrolateral appendages with segmental leg nerves and *Engrailed* expression in the posterior ectoderm of each segment (23, 48), appear to be panarthropod apomorphies that are not present in Cycloneuralia.

The position of tardigrades within the panarthropods is less certain. Overall, our results favor a sister group relationship between the Tardigrada and the Lobopodia. This relationship is favored because our EST and miRNA data both suggest a sister group relationship between onychophorans and arthropods and account for the uniquely shared features of onychophorans and arthropods (e.g., an open, hemocoelic circulatory system, a dorsal heart with segmental ostia, nephridia forming from segmented mesoderm), without the need to force their secondary loss in tardigrades as the result of miniaturization. Nonetheless, arthropods and tardigrades do share segmental ganglia in the nerve cord, in contrast to the unganglionated nerve cord in onychophorans (49), in which the commissures are not in segmental register. Our best tree, however, implies either convergent gain of segmental ganglia in tardigrades and arthropods or a secondarily unsegmented nerve cord in onychophorans, given that tardigrades share no miRNAs with arthropods to the exclusion of onychophorans and were not recovered as sister taxa in any of our EST analyses (Figs. 1 and 2 and Figs. S1 and S2). Analyses performed using the CAT+Γ model, similar to previous mitogenomic analyses (13), still pointed toward a Tardigrada + Onychophora group within Panarthropoda (Fig. S2*C*). CAT+Γ is not the overall best-fitting model for our dataset, however. When the overall best-fitting model (CAT-GTR+Γ) is used, our dataset support Lobopodia (Fig. 1), whereas mitogenomic data are known to be not very reliable markers for resolving deep divergences. In addition, no morphological evidence has been shown to support such a grouping, and no miRNA has been found to be shared exclusively between these two taxa. We conclude that by fully rejecting "Arthropoda + Tardigrada" (i.e., Tactopoda: ref. 50), which was never recovered in our analyses, and by favoring Lobopodia over Onychophora + Tardigrada, our results significantly reduce uncertainty regarding the placement of Tardigrada within Panarthropoda.

Our findings suggest that characters shared by tardigrades and cycloneuralians, such as a terminal mouth, protrusible mouth cone, triradiate pharynx, and a circumesophageal brain (9, 23, 51), are most likely ecdysozoan plesiomorphies. This is consistent with the fact that in our proposed phylogeny (Fig. 1*A*), even if the Tardigrada are excluded, the remaining cycloneuralian taxa do not form a monophyletic group (14). Instead, they are arranged as a paraphyletic grade at the base of Ecdysozoa (Fig. 1*A*). This hypothesis is also consistent with the fossil record of arthropods, in that taxa in the arthropod stem group, such as armoured lobopodians and anomalocaridids, show a melange of arthropod-like and cycloneuralian-like features, the latter (e.g., radially arranged mouthparts) then lost in the arthropod crown group (23, 50). Our phylogeny suggests that paired limbs and a shared mode of segment patterning (48) are apomorphic for Panarthropoda. Thus Tardigrades, as a living taxon with a mixture of cycloneuralian and arthropod characters, are placed center stage in our pursuit of understanding of the mechanisms underlying the construction of the most successful of all animal body plans, that of the arthropods.

Methods

EST Dataset Assembly. We assembled a 255-gene phylogenomic dataset of 49,023 amino acid positions from 33 ecdysozoan species by merging genes from two previous EST datasets (12, 14) (available on request). By merging these two datasets, we were able to improve taxonomic sampling with reference to (14) and particularly to (12). In addition, we were able to investigate the effect of including genes unique to (12) to the initial gene sets that we analyzed in (14) to address the problem of the relationships within Arthropoda. Improving taxonomic sampling is a key to alleviating LBA, and by merging the two datasets we were able to add data for one nematomorph, a second onychophoran, and an additional, relatively slowly evolving nematode. More details on dataset assembly, taxonomic sampling, and ortholog identification are provided in *SI Methods*. The average amount of missing data in our superalignment is ~36% (Table S1).

MiRNA Library Generation. Specimens of a velvet worm *Peripatoides novaezealandiae* were obtained commercially and identified by S.J.L.. A small-RNA library was constructed according to established protocols (38) and sequenced at 454 Life Sciences. The total RNA preparation of the tardigrade *Paramacrobiotus* cf. *richtersi* (~4,400 pooled individuals) was sequenced using Illumina technology at the Yale Center for Genome Analysis. Tardigrades were cultured by L.R. and T.M. and stored in RNAlater. MiRNA data for the arthropod subclasses Myriapoda and Chelicerata were obtained from previously described miRNA complements (14), and those for *Drosophila melanogaster*, *Daphnia pulex*, *Priapulus caudatus*, and *Caenorhabditis elegans* were obtained from miRBase (52). Sequences from the tardigrade and onychophoran small-RNA libraries were processed using PERL scripts written by L.I.C. and D.P. (available on request) and analyzed using miRMiner as described previously (14, 38).

Phylogenetic Analyses. All phylogenetic analyses were conducted under a Bayesian framework using PhyloBayes 3.2e (53). We first compared the fit of alternative models of evolution to our EST dataset. We used Bayesian cross-validation (45), as described in the PhyloBayes manual (53), to rank the fit of alternative substitution models to the data. The models compared were WAG+Γ, GTR+Γ, CAT+Γ, and CAT-GTR+Γ.

Phylogenetic analyses of the EST dataset were performed under each model, and results were compared to evaluate whether different phylogenies were obtained when different-fitting models were used. For every PhyloBayes analysis, two independent runs were executed. Convergence was tested using "bpcomp" in the PhyloBayes package. Analyses were considered to have converged when the maximum difference across bipartitions was <0.2 (see the PhyloBayes manual). For each analysis, the burn-in period was estimated independently, and trees sampled before convergence were not considered when summarizing the results of the two runs.

Site Stripping and Signal Dissection Analyses. These analyses used the slow-fast method (27) to estimate the rate of substitution of the sites in our alignment. First, the parsimony score of each site in our alignment was calculated for each of four groups with constrained monophyly (Pancrustacea, Chelicerata, Nematoda, and Lophotrochozoa). The rate of each site in our alignment was then estimated as the sum of its parsimony scores across all considered monophyletic groups. All parsimony analyses were performed using PAUP4b10 (54). Sites in our alignment were then ranked according to their substitution rates and partitioned into classes. Alignments were generated, according to the distribution of site rates, by systematically removing (*i*) approximately the fastest 10% of the sites, that is, all characters with a slow-fast–estimated rate of six or more steps (total number of remaining sites, 45,292); (*ii*) the fastest ~20% of the sites, that is, all characters with a slow-fast estimated rate of five or more steps (total number of remaining sites, 43,316); and (*iii*) the fastest ~30% of the sites, that is, all characters with a slow-fast–estimated rate of three or more steps (total number of remaining sites, 37,150). However, the number of substitutions in the sites that remained after exclusion of the first 10% of characters at just five or fewer steps is already low. This implies that the proportion of fast-evolving sites in our alignment is quite small. Accordingly, we did not create datasets excluding more than 30% of the fastest sites.

We also performed a signal-dissection analysis (14, 28) to compare the signal in the slow- and fast-evolving sites. Accordingly, two datasets were generated, containing approximately 10% (3,731 sites) and 30% (11,873 sites) of the fastest sites in our alignment. The five aligned datasets that resulted, namely the three sets composed of slow-evolving sites (approximately the slowest 70%, 80%, and 90%) and the two sets of fast-evolving sites (approximately the fastest 10% and 30%), were analyzed independently using PhyloBayes 3.2e to construct trees under the best-fitting model (i.e., the site-heterogeneous mixture model CAT-GTR+Γ).

Taxonomic Pruning Experiment. It is well known that the number and nature of the taxa used can affect phylogenetic inference and, in particular, can exacerbate or reduce LBA (2, 3). Thus, we carried out three taxon pruning experiments to evaluate the robustness of our EST results. We generated datasets that excluded (*i*) the tardigrade *Richtersius coronifer* and the onychophoran *Epiperipatus* sp., which resulted in uninterrupted branches for the tardigrades and the onychophorans; (*ii*) the nematomorph *Spinochordodes tellinii* and the tardigrade *R. coronifer*, which resulted in uninterrupted branches leading to the nematodes and the tardigrades; and (*iii*) the onychophoran *Epiperipatus* sp., the tardigrade *R. coronifer*, and the nematomorph *S. tellinii*, which resulted in uninterrupted branches leading to the onychophorans, tardigrades, and nematodes. In these experiments, the retained tardigrade was always *Hypsibius dujardini* because of its greater gene coverage. All of these datasets were analyzed under CAT-GTR+Γ.

ACKNOWLEDGMENTS. We thank A. Heimberg, E. Sperling, and V. Moy for technical assistance. K.J.P. is supported by the National Science Foundation and National Aeronautic and Space Administration (NASA) Astrobiology Institute. D.P., L.I.C., and O.R.-S. are supported by a Science Foundation Ireland Research Frontiers Programme grant (08/RFP/EOB1595, to D.P.). O.R.-S. and S.J.L. are also supported by individual Irish Research Council for Science Engineering and Technology, Empower Postdoctoral Fellowships. L.R. and T. M. are supported by the MoDNA project, financed by Fondazione Cassa di Risparmio di Modena and the University of Modena and Reggio Emilia. All analyses were performed using the computational infrastructures provided by the Science Foundation Ireland/Higher Education Authority Irish Centre for High-End Computing (ICHEC) and the National University of Ireland, Maynooth High-Performance Computing Facility.

1. Aguinaldo AM, et al. (1997) Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* 387:489–493.
2. Philippe H, Lartillot N, Brinkmann H (2005) Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Mol Biol Evol* 22:1246–1253.
3. Holton TA, Pisani D (2010) Deep genomic-scale analyses of the metazoa reject Coelomata: Evidence from single- and multigene families analyzed under a supertree and supermatrix paradigm. *Genome Biol Evol* 2:310–324.
4. Giribet G, Ribera C (1998) The position of arthropods in the animal kingdom: A search for a reliable outgroup for internal arthropod phylogeny. *Mol Phylogenet Evol* 9: 481–488.
5. Peterson KJ, Eernisse DJ (2001) Animal phylogeny and the ancestry of bilaterians: inferences from morphology and 18S rDNA gene sequences. *Evol Dev* 3:170–205.
6. Mallatt JM, Garey JR, Shultz JW (2004) Ecdysozoan phylogeny and Bayesian inference: First use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin. *Mol Phylogenet Evol* 31:178–191.
7. Telford MJ, Bourlat SJ, Economou A, Papillon D, Rota-Stabelli O (2008) The evolution of the Ecdysozoa. *Philos Trans R Soc Lond B Biol Sci* 363:1529–1537.
8. Nielsen C (2001) *Animal Evolution: Interrelationships of the Living Phyla* (Oxford Univ Press, Oxford), 2nd Ed.
9. Zantke J, Wolff C, Scholtz G (2008) Three-dimensional reconstruction of the central nervous system of *Macrobiotus hufelandi* (Eutardigrada, Parachela): Implications for the phylogenetic position of Tardigrada. *Zoomorphology* 127:21–36.
10. Zrzavy J, Mihulka S, Kepka P, Bezdek A, Tietz D (1998) Phylogeny of the Metazoa based on morphological and 18S ribosomal DNA evidence. *Cladistics* 14:249–285.
11. Mallatt J, Giribet G (2006) Further use of nearly complete 28S and 18S rRNA genes to classify Ecdysozoa: 37 more arthropods and a kinorhynch. *Mol Phylogenet Evol* 40: 772–794.
12. Dunn CW, et al. (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452:745–749.
13. Rota-Stabelli O, et al. (2010) Ecdysozoan mitogenomics: Evidence for a common origin of the legged invertebrates, the Panarthropoda. *Genome Biol Evol* 2:425–440.
14. Rota-Stabelli O, et al. (2011) A congruent solution to arthropod phylogeny: Phylogenomics, microRNAs and morphology support monophyletic Mandibulata. *Proc Biol Sci* 278:298–306.
15. Roeding F, et al. (2007) EST sequencing of Onychophora and phylogenomic analysis of Metazoa. *Mol Phylogenet Evol* 45:942–951.
16. Lartillot N, Philippe H (2008) Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. *Philos Trans R Soc Lond B Biol Sci* 363:1463–1472.
17. Hejnol A, et al. (2009) Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc Biol Sci* 276:4261–4270.
18. Roeding F, et al. (2009) A 454 sequencing approach for large-scale phylogenomic analysis of the common emperor scorpion (*Pandinus imperator*). *Mol Phylogenet Evol* 53:826–834.
19. Pick KS, et al. (2010) Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. *Mol Biol Evol* 27:1983–1987.
20. Meusemann K, et al. (2010) A phylogenomic approach to resolve the arthropod tree of life. *Mol Biol Evol* 27:2451–2464.
21. Sørensen MV, et al. (2008) New data from an enigmatic phylum: Evidence from molecular sequence data supports a sister-group relationship between Loricifera and Nematomorpha. *J Zoological Syst Evol Res* 46:231–239.
22. Andrew DR (2011) A new view of insect–crustacean relationships, II: Inferences from expressed sequence tags and comparisons with neural cladistics. *Arthropod Struct Dev* 40:289–302.
23. Edgecombe GD (2010) Arthropod phylogeny: An overview from the perspectives of morphology, molecular data and the fossil record. *Arthropod Struct Dev* 39:74–87.
24. Jeffroy O, Brinkmann H, Delsuc F, Philippe H (2006) Phylogenomics: The beginning of incongruence? *Trends Genet* 22:225–231.
25. Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool* 27:401–410.
26. Philippe H, et al. (2011) Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biol* 9:e1000602.
27. Brinkmann H, Philippe H (1999) Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. *Mol Biol Evol* 16:817–825.
28. Sperling EA, Peterson KJ, Pisani D (2009) Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa. *Mol Biol Evol* 26:2261–2274.
29. Zwickl DJ, Hillis DM (2002) Increased taxon sampling greatly reduces phylogenetic error. *Syst Biol* 51:588–598.
30. Regier JC, et al. (2010) Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature* 463:1079–1083.
31. Schierwater B, et al. (2009) Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis. *PLoS Biol* 7:e20.
32. Philippe H, et al. (2009) Phylogenomics revives traditional views on deep animal relationships. *Curr Biol* 19:706–712.
33. Philippe H, Delsuc F (2005) Phylogenomics. *Annu Rev Ecol Evol Syst* 36:541–562.
34. Wilson EO (1998) *Consilience: The Unity of Knowledge* (Alfred A. Knopf, New York), p 332.
35. Pisani D, Benton MJ, Wilkinson M (2007) Congruence of morphological and molecular phylogenies. *Acta Biotheor* 55:269–281.
36. Philippe H, et al. (2011) Acoelomorph flatworms are deuterostomes related to Xenoturbella. *Nature* 470:255–258.
37. Grimson A, et al. (2008) Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature* 455:1193–1197.
38. Wheeler BM, et al. (2009) The deep evolution of metazoan microRNAs. *Evol Dev* 11: 50–68.
39. Bartel DP (2009) MicroRNAs: Target recognition and regulatory functions. *Cell* 136: 215–233.
40. Sperling EA, Peterson KJ (2009) MicroRNAs and metazoan phylogeny: Big trees from little genes. *Animal Evolution: Genomes, Fossils, and Trees*, eds Telford MJ, Littlewood DTJ (Oxford Univ Press, Oxford), pp 157–210.
41. Sperling EA, et al. (2009) MicroRNAs resolve an apparent conflict between annelid systematics and their fossil record. *Proc Biol Sci* 276:4315–4322.
42. Sperling EA, Robinson JM, Pisani D, Peterson KJ (2010) Where's the glass? Biomarkers, molecular clocks, and microRNAs suggest a 200-Myr missing Precambrian fossil record of siliceous sponge spicules. *Geobiology* 8:24–36.
43. Heimberg AM, Cowper-Sal-lari R, Sémon M, Donoghue PC, Peterson KJ (2010) microRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrate. *Proc Natl Acad Sci USA* 107:19379–19383.
44. Sperling EA, Pisani D, Peterson KJ (2011) Molecular paleobiological insights into the origin of the Brachiopoda. *Evol Dev* 13:290–303.
45. Stone M (1974) Cross-validatory choice and assessment of statistical predictions. *J R Stat Soc Series B Stat Methodol* 36:111–147.
46. Snodgrass RE (1938) Evolution of the Annelida, Onychophora, and Arthropoda. *Smithsonian Miscellaneous Collections* 97:1–159.
47. de Wit E, Linsen SEV, Cuppen E, Berezikov E (2009) Repertoire and evolution of miRNA genes in four divergent nematode species. *Genome Res* 19:2064–2074.
48. Gabriel WN, Goldstein B (2007) Segmental expression of Pax3/7 and engrailed homologs in tardigrade development. *Dev Genes Evol* 217:421–433.
49. Whitington PM, Mayer G (2011) The origins of the arthropod nervous system: Insights from the Onychophora. *Arthropod Struct Dev* 40:193–209.
50. Budd GE (2001) Tardigrades as "stem-group arthropods": The evidence from the Cambrian fauna. *Zool Anz* 240:265–279.
51. Schmidt-Rhaesa A (1998) The position of the Arthropoda in the phylogenetic system. *J Morphol* 238:263–285.
52. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ (2006) miRBase: MicroRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* 34(Database issue):D140–D144.
53. Lartillot N, Lepage T, Blanquart S (2009) PhyloBayes 3: A Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25:2286–2288.
54. Swofford DL (2002) *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4.0 beta 10* (Sinauer Associates, Sunderland, MA).