A phylogenomic enquiry into Metazoan macroevolutionary dynamics

Alastair Roger Tanner

A dissertation submitted to the University of Bristol
in accordance with the requirements for award
of the degree of Doctor of Philosophy in the Faculty of Science.

University of Bristol School of Biological Sciences
Life Sciences Building
24 Tyndall Avenue BS8 1TQ

January 2018

Word count: 47,936

Abstract

The reconstruction of the Tree of Life is a central activity for understanding evolution, biodiversity and ecology in the widest temporal framework. Phylogenomics is the inference of relationships between species using inherited molecular characteristics, and is now a major methodology for tree reconstruction, contributing to the multidisciplinarity of palaeobiology. A key collaboration in palaeobiology is between palaeontologists and molecular biologists: the former provide comparative biology and geological context, the latter provide extensive molecular sequence data which can be objectively modelled in ways that traditional palaeontological data usually cannot. Since the maturation of sequencing technology molecular data have amassed, and the techniques and computational facilities to analyse such data continue to advance. For the metazoan Tree of Life, the most difficult relationships to understand are those characterised by rapid diversification, vague fossil records, conflicting phylogenetic signal, or combinations of all three. However, understanding these branches remains crucial to understanding both macroevolution and its ecological contexts, since these events are responsible for the vast biological diversity we see today. In this thesis we apply phylogenomics to investigate the evolutionary origins and ecological contexts for four recalcitrant and controversial Metazoan groups: coleoid cephalopods, chelicerates, earthworms, and the phylum Annelida. In concert with currently available data, we contribute newly sequenced species, curate new phylogenomic datasets, apply data-refinement protocols, and employ statistically supported inference methods to construct phylogenetic trees and estimate molecular divergence times. We report new hypotheses on relationships within these groups, infer divergence times for their inaugural evolutionary radiations, and develop hypotheses on the ecological dynamics at their origins. The four projects thus represent the current best insight on these questions, and outline future approaches to difficult phylogenetic questions in the palaeobiology of Metazoa.

Author declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes, and that it has not been submitted for any other academic award. The work is entirely the candidate's, except where indicated in the text. All views expressed in the dissertation are those of the author.

Alastair R Tanner
3rd of January 2018

Statement of collaboration

Chapters 2, 3 and 4 are research collaborations led by the author. Details of personal contributions are given on heading pages for these chapters. For all of these chapters, experimental design, dataset curation, bioinformatics scripting, computational experimental procedure, interpretation of results, figure design, illustration, and lead writing was carried out by Alastair R Tanner.

Chapters 1, 5 and 6 are entirely the work of Alastair R Tanner.

Alastair R Tanner
3rd of January 2018

Acknowledgements

This PhD was funded by the University of Bristol Scholarship in Teaching and Research programme, and I thank the university for their generosity, and for the wide range of educational facilities, staff and workshops that have been available to me. I am grateful for the academic guidance and enthusiasm of Jakob Vinther and Davide Pisani, without which I would have just sat in the corner writing esoteric computer script rather than applying it to anything. I am grateful to the university for granting access to the BlueCrystal computation cluster, and I thank the staff at the Advanced Computing Research Centre, in particular Callum Wright. As my teaching supervisor, I thank Wendy Gibson for her kind generosity with her time and advice. I am also grateful for the warm encouragement and reassurance offered by Sue Amesbury.

My thanks go to the University of Bristol Palaeobiology Research Group, and the people who make it supportive working environment. In particular I thank Frankie Dunn, Kelly Vargas, Dave Marshall, Fiona Walker, James Fleming, Joe O'Reilly, Holly Betts, Lucy Holloway, James Clark, Céline Petitjean, Armin Esler, Mark Puttick and Joe Keating for their erudite wisdom and incisive intellectualism.

Back in the real world, I could not have completed this work without the love and support of my friends Lucy Langley-Palmer, Tom Palmer, Owen Hastings, Chris Deas, Emily Barrett, Corin Evans-Pritchard, Charlie Albin, Milo Newman, Alex Riorden and Boz.

As ever, I offer my love and gratitude to my parents, Christine and Andrew Tanner, to whom I owe everything.

“If you wish at once to do nothing and be respectable nowadays,
the best pretext is to be at work on some profound study.”
Leslie Stephen (1832 - 1904)

“I would never die for my beliefs, because I might be wrong.”
Bertrand Russell (1872 - 1970)

Table of Contents

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

List of figures

List of tables

Thesis outline

This thesis aims to clarify some of the evolutionary history of four groups of invertebrate organisms. Its primary methodology is molecular phylogenetics, phylogenomics, and molecular clock inference. The organisms focussed upon, namely cephalopods (squid, octopuses and cuttlefish; chapter 2), chelicerates (spiders, scorpions, mites and allies; chapter 3), earthworms (terrestrial annelids; chapter 4) and the wider annelid group (chapter 5) have been chosen because these groups have represented difficult phylogenetic problems. Now that we are a decade into the genomic revolution, the increase in availability of sequence data makes molecular investigations of these palaeobiological questions feasible. These groups also offer potential insight into the narrative of invertebrate evolution over the Phanerozoic, from both an evolutionary and ecological perspective. Furthermore, understanding the evolutionary dynamics of these groups is valuable since it clarifies the nature of major ecological shifts.

The methods applied come under the now-mature area of phylogenomics, emerging from the enormous increase in availability of molecular sequence information. Concurrently, computational power has continued to develop, with high-performance facilities becoming more universally available, thus allowing research to be carried out in reasonable time frames. The statistical philosophy adopted in this thesis is Bayesian Inference, on account of the power of Bayesian statistics to both incorporate uncertainty as a component of the analysis, and express uncertainty in the results of such analyses. This leads to a more natural interpretation of results than might be possible through other inference methodologies. This is discussed in detail in papers arising from this work (but not forming chapters of this thesis); these publications are included in Appendix B.

Chapter 1 is an introductory discussion of the current state of palaeobiology from a molecular phylogenetic perspective. There is a brief discussion of the philosophical implications of palaeobiology, and the scope, limits and technical developments of molecular phylogenetics.

The evolutionary origins of modern coleoid cephalopods (squid, octopuses and cuttlefish) are the focus of chapter 2. Cephalopods provide an excellent group to introduce the methodology applied in this thesis since they have a well-understood earlier fossil-record of shelled-ancestors, including the iconic ammonites and belemnites. However, their descendants into the present day underwent a regime of shell-reduction and loss, and as such the palaeontological story told through fossils becomes less distinct. Here we use molecular methods to clarify this narrative, showing that modern coleoid cephalopods had origins coinciding with major marine ecological turnover during the Mesozoic Marine Revolution.

We move away from an exclusively marine ecology in Chapter 3, and present research into the origins and macroevolution of Chelicerata (spiders, scorpions, mites, ticks, and their allies). Recent comprehensive work has identified conflict in the phylogenetic signal for the group, which continues to be controversial. We present a hypothesis for the relationships among chelicerates which is supported by independent lines of evidence. We also infer a terrestrialisation regime, and consider the ecological context for this.

In chapter 4 we focus on the origin on earthworms, a familiar if humble organism which belies its enormous ecological impact for terrestrial systems. Molecular clocks reveal coincidence between global changes in forest makeup, the rise of seed-plants and their fungal symbionts. We suggest that change of this scale can be the result of both biotic and abiotic components, and that the diversification of earthworms, at least in part, contributed to the the demise of Carboniferous-style ever-wet coal-swamp forest ecosystems.

The annelid theme is expanded in scope in chapter 5. Gathering data to form the widest dataset on annelids assembled to date for phylogenomic study, the origin and dynamics of annelid evolution is investigated. The diversification of annelids shortly after the Cambrian has remained a controversial topic, and we here demonstrate that molecular phylogenetics may have reached a limit on their inferential power. Annelids appear to have gone through a period of such rapid evolution that resolving their relationships remains difficult, regardless of the weight of data, the data-refinement process, or the sophistication of methodology. Such a situation remains valuable in identifying limits to the power of molecular phylogenetics. After all, the fact that molecular information can reveal evolution’s path more than 500 million years ago is remarkable; and clarifying where it fails, gives us a chance to take stock of the many situations in which it has been successful.

Chapter 6 reflects and concludes on the dissertation as a whole. The successes and failures are discussed, and I consider the present and possible future directions of phylogenomic science in a palaeobiological context. I return to themes outlined in the introduction, and offer some thoughts on the state of genomic evolutionary science and its connections with science as a whole.

Chapter 1

Molecular palaeobiology and the Tree of Life

This chapter has not been published.

Abstract

A fundamental goal in evolutionary biology and palaeobiology is to understand life through a unified Tree of Life. Phylogenetics is the objective inference of evolutionary relationships, through the analysis of observational data. Phylogenies are powerful graphical tools since they express diversity, divergence, relationships, and imply a passage of time. Certain parts of the Tree of Life remain unresolved, variously through lack of palaeontological support, difficulties in analysis of data, and the reality that important evolutionary episodes are often abrupt. Four groups represent such difficulties and are outlined as the focus of this research thesis: cephalopods, chelicerates, earthworms, and annelids. These groups are timely to investigate due to advances in data resources and inferential techniques. Each group is also important since they represent enormous modern diversity, but arise from uncertain origins. Molecular data currently informs many areas of palaeobiology, but is prone to several kinds of biases. Appropriate steps can be made to minimise biased data, and the resulting refined dataset can be analysed using methods suitable for minimising systematic bias. When integrated with other independent sources of evidence, molecular palaeobiology proves a vital part of contemporary efforts to understand the history of life.

1.1 The science of history

Palaeobiology applies experimental science to investigate the history of life. Here on Earth, we find life to be characterised by broad organismal biodiversity and intricate molecular function, all united into complex layers of interconnected ecologies (Lovelock and Margulis, 1974). We also understand that the history of life on Earth spans back to nearly the inception of the planet itself (Dodd et al., 2017; Sugitani et al., 2013), and has seen cataclysmic change as well as eons of ostensible stasis. The task therefore of the palaeobiologist is to draw out the varied stories of life in this framework of billions of years, with the goal of refining our understanding of evolution and the history of life.

Palaeobiology itself has developed and diversified in recent decades, widening in scope from the more more "traditional" fields of palaeontology, geology and comparative biology, to an extended multidisciplinarity drawing from such sciences as geochemistry, geometric statistics, computer science, bioimaging, biophysics, morphometrics, molecular evolution and phylogenomics. Each approach alone might be limited in palaeontological potential, but together synthesise to provide powerful inference on the past. For example, geology and phylogenomics combine to form molecular clocks and provide a timescale beyond a literal reading of the fossil record (Donoghue and Benton, 2007; Fitch and Margoliash, 1967); geology, imaging and geometric statistics can reveal the reproductive mode of Ediacaran life (Mitchell et al., 2015); electron microscopy and camouflage ecology lead us to unexpected insight on the lives of dinosaurs (Smithwick et al., 2017); or molecular genetics and computer science can fuel important conversations on the origins of metazoans (Feuda et al., 2017; Pisani et al., 2015; Whelan et al., 2017). These kinds of synergistic collaborations now characterise palaeobiology.

Although biological enquiry and evolutionary ideas have a long history (Mayr, 1991), Charles Darwin's *On the Origin of Species* (1859) can be considered the origin of scientific evolutionary thought. Darwin's work itself exemplifies multidisciplinarity, by presenting a coherent hypothesis for "what life is", supported by a range of independent sources of evidence and lines of logic. But what Darwin's work also represents is a coming-of-age, both in terms of the scientific method and in the humanist philosophies which had developed since the European Renaissance. Evolution by natural selection stands as a fundamental paradigm shift, perhaps only really comparable with other scientific iconoclasms such as Copernican Heliocentrism, and 20th century quantum mechanics. Each of these revolutions forced a comprehensive reclassification of the universe and humanity's place in it. Heliocentrism initiated the science of unimaginable universal scale, and rendered anthropocentrism as void (Gingerich, 1973). Quantum mechanics invalidated a deterministic, clockwork cosmos, and showed that the fundamental character of reality is chance (Omnès, 1992). In a way, Darwin's work can be viewed as a biological reflection of both concepts, with evolution (and its mechanism) upheld as universal, unguided and unsentimental. As such, teleological dogma based in Christian thought (and amplified by Victorian devotion to concepts of control, progress and perfection) collapsed under the weight of evidence and humanist philosophy, which Darwin's *Origin* represents. However, we are cautioned to not be too complacent in thinking that humanism or the scientific method as being the end-state of scientific investigation, since it is becoming clear that once-sacrosanct concepts, such as Popperian falsifiability, are of diminishing service to us in the face of the complexity, and fundamental uncertainty, of reality (Stamos, 1996).

Since the *Origin*, a primary goal of evolutionary biology has been to expand on the Darwinian concept of universal ancestry by determining the relationships between lineages of organisms, and place these in a context to explain (or at least describe) the history of life. Great progress has been made in this, and today we have a broadly complete tree of eukaryotic life (perhaps especially across much of Metazoa), but the endeavour has also highlighted the limits of expressing evolution through phylogenies. We now know the branching tree-like metaphor for organismal relationships is not always appropriate, and is generally inapplicable to prokaryotic life - arguably the dominant and characteristic life-form of Earth. This lead to one of the most profound observation in evolutionary biology: all eukaryotic organisms are assemblages of prokaryotes, symbiotic on multiple interconnected levels (O'Malley, 2017; Sagan, 1967). Consequently, the most important "branching" event on the Tree of Life, at the origin of eukaryotes, turns out to not only violate but invert branching relationships, and is in fact a merger between prokaryotic lineages.

Despite that insight, for multicellular eukaryotes (and the metazoan focus of this thesis) the Tree of Life remains an ideal analogy: most evolutionary dynamics of principally isolated species is through vertical descent, and speciation almost always results in binary splits of lineages. As such, for more familiar metazoan life like whales and wasps and squid and spiders, the phylogeny is the most meaningful way to view evolutionary relationships, and the Tree of Life is an intuitive foundation from which we can understand and discuss evolutionary history. So, while evolutionary biology continues to unveil fascinating layers of complexity, the ambition to advance palaeobiology through an increasingly comprehensive Tree of Life remains a rich vein of research, and, as the single figure in Darwin's *Origin* would attest, the phylogenetic tree remains our most powerful tool to progress the science of history.

1.2 Uncertainties on the Tree of Life

Much of the power of evolutionary theory comes from its wide scope in linking casual, intuitive, everyday observations with the more intricate nature of biology. For example, humans have known for millennia how to influence crops or livestock (or humans themselves) through selective breeding. It is only recently that we have been able to quantify genetic inheritance and today we have an understanding of the deeper mechanisms, but inheritance has been human intuition for tens of millennia. Or, it is clear that nobody needs evolutionary insight to assume that, say, all dragonflies or frogs or mice or oak trees form respectively mutually exclusive groups: it is fairly safe to say that no mouse is a descendant of the common ancestor of dragonflies. But the power of evolutionary investigations has been to put these kinds of intuitions into the wider perspective, and even in the past few decades our appreciation of this has developed profoundly, and some strongly-held assumptions have been shown to be biased. To pick out some celebrated examples, it is now appreciated that insects are best understood as crustaceans (Zrzavý & Štys, 1997; Rota-Stabelli et al., 2010), or whales as artiodactyls (Graur & Higgins, 1994; Gatesy et al., 1996; Price et al., 2005; Zhou et al., 2011), and birds as dinosaurs (Ostrom 1975; Brusatte et al., 2010).

While these types of examples are gratifying (and of course remain amenable to change), they serve us by highlighting that what we celebrate as true may well turn out to be wrong. That, in itself, is of course the nature of science itself: no science can ever hope to be completely right, but can hope to be the least wrong. Given new observations, methodologies or even philosophies, science can be seen as a process of refinement, rather than an end-goal of immutable truth. For historical inference as in palaeobiology, this sentiment is given another dimension, in that regardless of the evidence, we cannot experimentally replay or replicate the past. As a judge, we can only hope to listen (without bias) to the advocates (who will be biased), and interpret their evidence to offer our best conclusion of what really happened. This thesis serves to offer new judgement on specific areas of the Tree of Life which have remained controversial or unresolved. As with all science, the closer we look and the more specific our questions become, the less sure we can be about the answers; progress requires increasingly robust enquiry. The reasons for these uncertainties, in the case of evolutionary biology, can be generalised.

Firstly, there is the question of the quality of the fossil record. As broadly discussed over the history of palaeontology, the fossil record is a sparse and fragmented reflection of past biodiversity, but when corrected or normalised for geological biases, there is a generally stable palaeontological signal (Benton et al., 2000). However, the signal shifts in its detail depending on environmental conditions, organismal morphology and ecology, and is contingent on geological processes such as burial, anagenesis, metamorphosis, subduction, uplift and erosion (Parry et al., 2017). Yet, despite its intermittent and at times misleading nature, the fossil record remains good enough for confidence in historical inferences, and it is to palaeontologists' credit that we have a relatively clear picture of the past, given the paucity of the data that we often have to work with. In the research projects in this thesis, fossils are crucial in guiding molecular divergence time estimations (see section 1.11), and as such only uncontroversial fossils are used as *a priori* knowledge. Further to this, the fossil record provides a framework to adjudge molecular inference, essentially from a parsimony philosophical standpoint (see Chapter 3). As such, the fossil record, though uncertain, acts as an important cross validation, especially for phylogenies which are rendered difficult through resolution issues.

This second issue of uncertainty is the question of resolution: whether or not a branching event, or "node", can be discerned. The greater the taxonomic complement of a phylogeny, the more difficult phylogenetic and divergence-time inference becomes. A wide but sparsely complemented example of a phylogeny, for example, would be in drawing up the relationships between our dragonflies, mice, frogs and oaks. This is relatively easy, and intuition-informed phylogenies (like the iconic, pre-Darwinian Haeckel tree) generally place these relationships correctly without making any formal scientific observations. However, drawing up a phylogeny of, say, all dragonfly species is more difficult, since dragonflies share many similarities and represent a more detailed evolutionary tree. On the other hand, comprehending the timing of evolutionary radiations requires inference on the dynamics of species divergence: bifurcation over great time-spans can determine the relationships within a group exhibiting gradual, perhaps more regular divergences. However, evolution is often a process of irregular bursts of branching events over short time-spans. It is these events which require more precise inference if we are to improve our understanding of evolutionary history. The corollary of is that, in deep-time evolutionary inference, we must contend with evolutionary dynamics at both ends of this spectrum: inferring trees characterised as having lumbering expanses of slow evolutionary dynamics, embellished with irregular episodes of baroque complexity.

Evolution is not consistent in its rate of change, and can be driven by biotic and abiotic factors (Benton, 2009). Sometimes species remain static, while at other times will experience change so rapid as to appear instantaneous, on geological timescales (Gould and Eldredge, 1993). As identified by Darwin himself, the most spectacular example of extreme evolutionary change is the Cambrian Explosion, whereby nearly all animal phyla arise within a few tens of millions of years (Erwin, 2007; Simpson, 1945). On a smaller scale, these kinds of radiations also characterise individual lineages, as species variously come across opportunity to expand or reasons to perish. Ecological niches and the incumbent organisms can be robust and long-lived, or precarious and ephemeral. As such, the expanding tree of life is characterised as both undergoing saltative pops of diversification, as well as at-times smoothly expanding diversity (Gould and Eldredge, 1977). Further to this, it should also be appreciated that while explosive radiation is an evolutionary reality, so is implosive demise: wide and long-established canopies of diversity can meet near-instantaneous death. An evocative example of coincident explosion and implosion is at the Cretaceous-Palaeogene event: the collapse of non-avian dinosaurs and explosion of birds was coincident (and likely related), probably due to the ecological changes resulting from meteorite impact (Brocklehurst et al., 2012). Examples like this illustrate that for phylogeneticists the most important nodes on evolutionary trees (perhaps defined as leading to clades of great diversity of ecological breadth) are often characterised by very short spans of time between speciation events. In light of this, it might be seen as surprising that we dispel uncertaining surrounding abrupt evolutionary events so deep in time.

1.3 Four examples of evolutionary uncertainty

This thesis focuses on four areas (see sections 1.3, 1.4, 1.5 and 1.6, and Chapters 2, 3, 4 and 5) of the metazoan Tree of Life which have remained controversial, and despite being the focus of palaeobiological research have resisted satisfactory phylogenetic investigation. They also arise as important topics to revisit at this time for a range of practical reasons. Firstly, publicly available molecular sequence data continues to accumulate and facilitate comprehensive phylogenomic analysis. Secondly, new sequence data can now be cheaply, quickly and reliably generated, independent of large financial resources, technical expertise or specialist facilities. Lastly, phylogenetic inference methodology has matured in the last decade, and advances in computational power have meant these sophisticated approaches can be applied to large datasets and produce results in sensible timescales. Further factors influencing the areas of research have been the research fields of collaborators, availability of specimens, and the fit of the work given the context of the University of Bristol Palaeobiology Research Group from 2014 to 2017.

The study therefore focuses on cephalopod molluscs (Chapter 2), terrestrial chelicerates (Chapter 3), terrestrial annelids (Chapter 4), and annelids as a whole (Chapter 5). From the perspective of this work, the common characteristics of these ostensibly disparate groups is that their origins feature rapid radiations, and (especially for cephalopods and earthworms) bear witness to major ecological turnovers that we uphold in light of our research findings. Consequently, they also share a status of being evolutionarily unresolved, and attendant to that is the uncertainty of how to treat data, particularly molecular sequence information, to tackle such evolutionary problems. Chapters 3 and 5 seek to unravel notoriously recalcitrant evolutionary relationships, and understand the reasons why these groups have stood out as difficult problems in the discipline of phylogenomics.

1.4 The origin of coleoid cephalopods

Octopuses, cuttlefish and squid are marine cephalopod molluscs showcasing bizarre and sophisticated adaptations. These characteristics include active camouflage, visual communication systems, jet locomotion, ink-discharge decoy defense mechanisms, as well as the highest intelligence of any invertebrate (Mather and Kuba, 2013a). They also have wide ecologies and adaptive morphologies, ranging from the deep-water giant squid *Architeuthis*, to shoaling oceanic squid, to reef hunters like the cuttlefish *Sepia*, and the ambush predatory octopuses (Wells and O'Dor, 1991), (see Figure 1.1 for examples of diversity in cephalopods). These animals are members of the monophyletic group of the coleoid cephalopods, distinct from their closest relatives (the two genera of *Nautilus*) by having reduced, internalised, vestigial shell structures. However, despite their ecological importance and unique adaptations, their evolutionary origins and diversification remain obscure.

Modern cephalopods present an intriguing palaeontological question. While today's diversity is dominated by predominantly shell-less coleoids, past diversity, especially prior to the mid-Mesozoic, the cephalopod shell was ubiquitous and as such left a comprehensive fossil record (Kröger et al., 2011). We have a scenario of poorly-preserved modern diversity (through shell reduction and loss) predated by a well-preserved earlier diversity, exemplified by such icons of palaeontology as ammonites and belemnites. Thus we have a significant gap in our story of the marine macroevolution, and the nature and providence of the cephalopod diversity we see today remains to be explained. When did coleoid cephalopods originate and diversify? What were the ecological contexts for this change? How does this relate to wider marine evolutionary dynamics for both invertebrates and vertebrates?

In Chapter 2 we approach these questions. Given the vagaries of the cephalopod fossil record, we tackle the question using new molecular sequence information analysed in a Bayesian framework to derive divergence time estimates for the coleoid clade. Previously, inferences on the evolution of cephalopod had relied on interpolation between fossils (Fuchs et al., 2015; Kröger, 2005; Kröger et al., 2011), while this new work provides a new view across the cephalopod crown group. We propose and support an ecological scenario for the evolutionary origin of coleoid cephalopods, upholding the narrative of coleoid cephalopods arising in competition with marine vertebrates, and thus being a facet of marine faunal turnover in the Mesozoic characterised by heightened ecological competition.

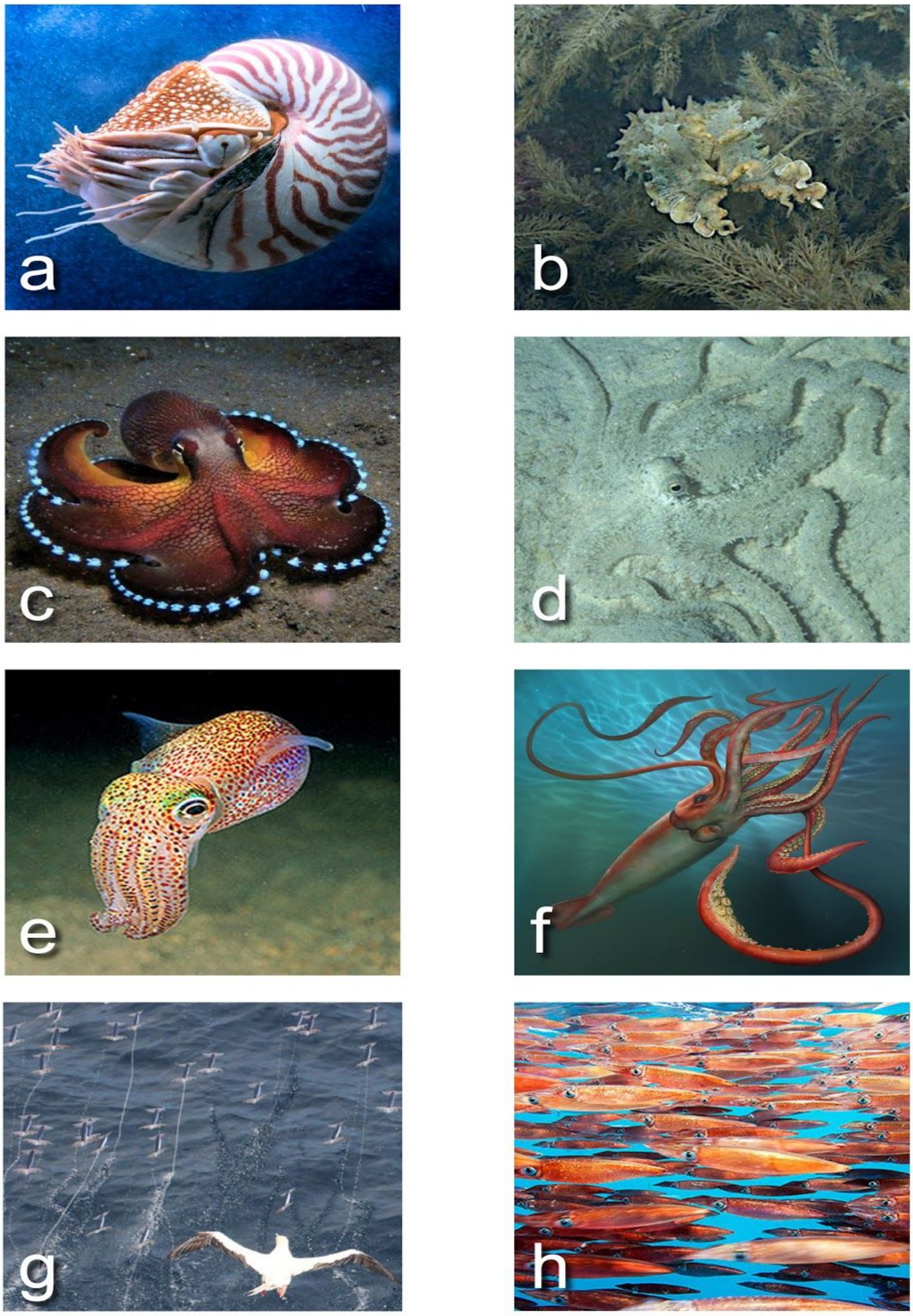


Figure 1.1. Examples of diversity in extant cephalopods. [a] The shelled, non-coleoid cephalopod *Nautilus pompilius*. [b] Cuttlefish *Sepia sepia* showing both colour and body-texture mimicry of seaweed. [c] *Octopus cyanea* showing a colourful defensive posture. [d] *Octopus vulgaris* camouflaged against sand. [e] The thumb-sized bobtail squid, *Euprymna scolopes*. [f] Artist's impression of the giant squid, *Architeuthis dux* (all photographs of *A. dux* are either dead or dying at the surface). [g] The flying squid *Todarodes pacificus* gliding above water, pursued by a gull. [h] Shoaling oceanic *Loligo pealei* near the sea surface. (All images under creative commons licence.)

1.5 The phylogeny of chelicerates

On the widest phyletic perspective, the arthropods are composed of the reciprocally monophyletic groups of Mandibulata (Pancrustacea + Myriapoda) and Chelicerata (Hejnol et al., 2009; Rota-Stabelli et al., 2011). Chelicerates include familiar animals such as spiders, scorpions and mites, the less familiar (but superficially scorpion-like) uropygids and pseudoscorpions, the enigmatic marine pycnogonid sea-spiders, and palaeontologically iconic horseshoe crabs. While sharing exoskeletal traits with mandibulates, such as ecdysis, chelicerates diverge from mandibulates in many crucial ways and as such show how evolution can take wildly different trajectories, over the widest clades (Budd and Telford, 2009). This can be seen in many phenotypic contrasts between the groups, with their mode of feeding, respiratory, locomotory, nervous and sensory systems all differing in fundamental ways. Although this work does not address evolutionary specifics, it is worth pausing to consider questions such as “why did powered flight not evolve in chelicerates?”, or “why is the use of silk webs such a prevalent predation strategy among spiders, but not in insects?”, given the context of their macroevolution.

These kinds of questions perhaps highlight the contingent nature of evolution itself, and how particular ecologies can be “locked” as a character of some groups, while others are excluded from exploiting that same niche. It also reflects not only the early divergence of chelicerates from other arthropods, but the contrast in ecologies that chelicerates have adapted to in comparison to other arthropods. Although the group have, to varying degrees, adopted lifestyles including herbivory, scavenging and detritivory, the predominant lifestyle of terrestrial chelicerates is as predators (primarily of other arthropods) and parasites (Penney, 2003). Further to this, while the group might not be represented by the vast species diversity of insects, it is clear that the chelicerates are an evolutionarily successful group, consistently featuring as significant members of ecosystems since their Palaeozoic origins (Kalmar and Currie, 2010).

In Chapters 2, 4 and 5 (and sections 1.4, 1.6, and 1.7) we discuss, identify and tackle shortcomings in evolutionary insight, in part due to paucity of the fossil record. In contrast, we do have a relatively informative chelicerate fossil record back to their Palaeozoic origins, but recovery of a palaeontologically plausible and robust phylogeny from molecular sequence information has proved difficult (Regier et al., 2010; Sharma et al., 2014). As such, reconciliation between independent lines of evidence has been elusive, and major ancestral nodes remain controversial (Giribet et al., 2002; Shultz, 1990, 2007). As identified in an exhaustive enquiry on arachnid phylogenetics (Sharma et al., 2014), chelicerate relationships remain as one of the last unresolved problem in arthropod systematics. In that work Sharma upheld some derived arachnid relationships as robust from a number of phylogenomic approaches, while highlighting ancestral (and thus more informative from a deep-time perspective) nodes as highly uncertain. While potentially frustrating, we here see the strength of multi-disciplinarity in palaeobiology, in that recourse to a single line of evidence could lead to a single result being upheld as true, but one which actually crumbles under more diverse investigation. In this way, the complexity of macroevolution is made clear, and those working to resolve deep-time macroevolutionary questions are wise to cautiously weigh the strength of each line of evidence.

Broadly, the reason for the recalcitrant nature of the chelicerate phylogeny is due to insufficient signal, or insufficient mitigation of sources of phylogenetic noise (see section 1.9 for a discussion of biases, including long-branch attraction). More specifically, we can identify problems in gene-tree discordance, which can otherwise be described as weak phylogenetic signal from the short (but critical) internodes. As common across other investigations in this thesis, the problem is explosive evolutionary divergence, in this case shortly after the start of the Phanerozoic. Exacerbating this problem, is that the Acari (mites and ticks), a major evolutionary group and one of medical interest, are parasitic. As such, these organisms' molecular sequences are subject to elevated substitution rates, under the evolutionary pressure to “out-evolve” the immune responses of their hosts (Bromham, 2009). Consequently, parasites usually exhibit very long-branched phylogenomic (and gene tree) results, misleading inference methodology through biases such as long-branch-attraction and compositional bias (see section 1.10 for further discussion on the pitfalls of phylogenetic inference). Further to this, difficulties have long been present in inference of chelicerate systematics, with morphologically and ecological nonsensical scenarios being returned as results. For example, molecular inference often struggles with the placement of horseshoe crabs, suggesting that their closest relatives are scorpions (Sharma et al., 2015). Such a topology would imply the implausible scenario of terrestrialisation and a return to marine habit for xiphosurids, which is a narrative with no support from the fossil record. Other problems are that Acari, morphologically and ecologically similar, are inferred as polyphyletic on molecular grounds (Sharma et al., 2014).

In chapter 3, we generate new data and curate phylogenomic datasets to reappraise chelicerate origins and macroevolution. Chelicerate systematics require especially sensitive measures as to retain signal while minimising noise, and we apply new methods to deal with this. We present a new phylogeny for Chelicerata, consilient with both morphological and palaeontological views on their evolution and origins.

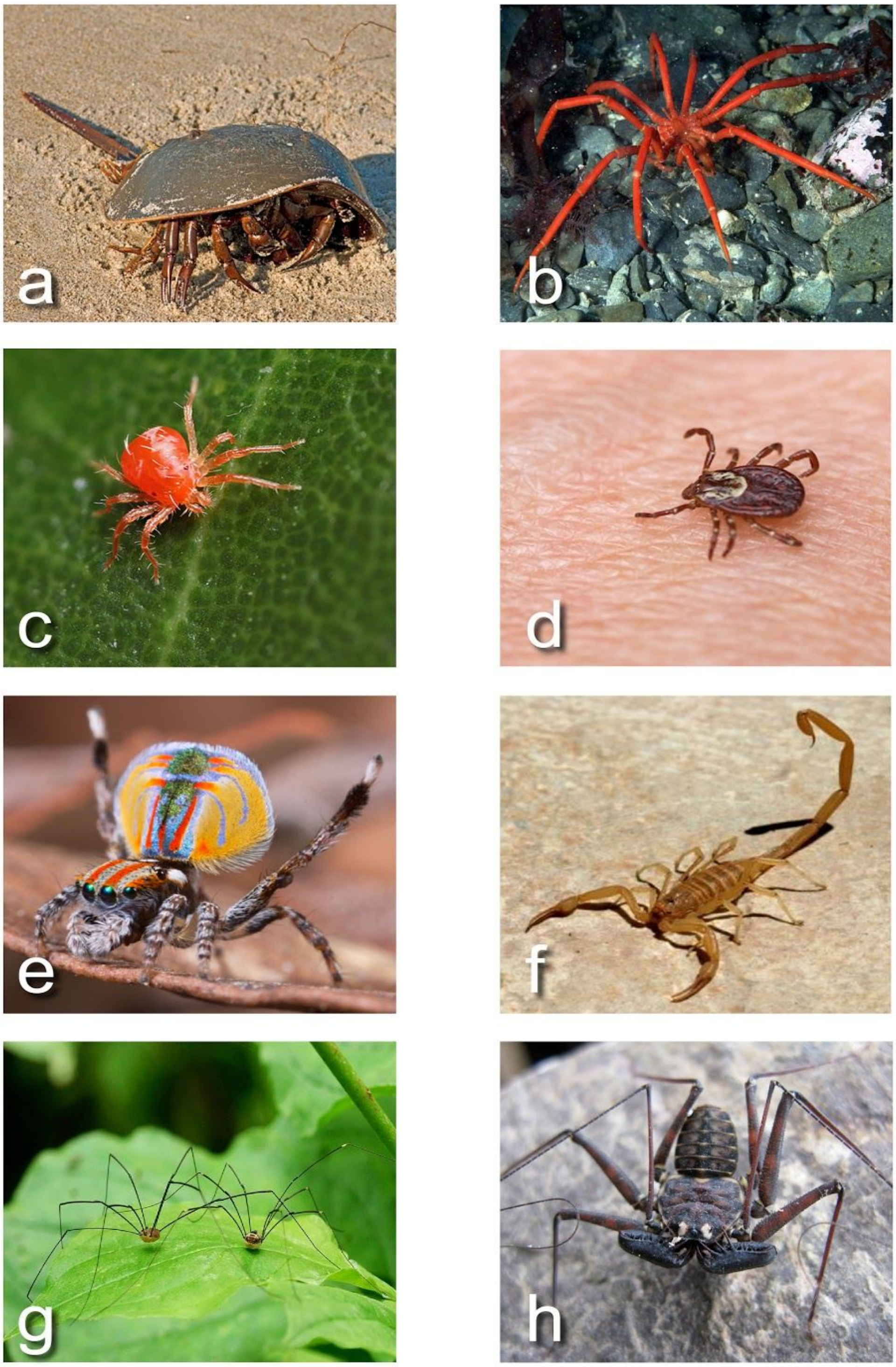


Figure 1.2. Examples of diversity in extant chelicerates. [a] The marine horseshoe crab *Limulus polyphemus* on beach during reproductive phase. [b] Pycnogonid sea-spider of *Nymphon* genus in its marine habitat. [c] The social red spider mite, *Tetranychus urticae*. [d] Courtship display of the peacock wolf spider, *Maratus volans*. [e] Haemophagic mite *Ixodes ricinus* on human skin. [f] Arizona bark scorpion *Centruroides sculpturatus*. [g] Two opiliones, the North European harvestman *Leiobunum rotundum*. [h] Amblypygid of the *Heterophrynus* genus. (All images under creative commons licence.)

1.6 The origin of earthworms

Annelids are spectacularly diverse organisms (see section 1.6 and Chapter 5), with lifestyles ranging from filter-feeding to active predation to thermal-vent symbiosis with chemotrophs (Parry et al., 2014). Most species of annelids are marine, and therefore terrestrial lineages represent special cases of adaptation, and are of major evolutionary interest. Those land-living annelids are represented by earthworms and leeches (plus a handful of other less-familiar species), and of these, earthworms are of particular interest since they qualify as ecosystem engineers (Ehrenfeld, 2010). Their presence, especially in terrestrial forest ecologies, modulates soil depth, turnover and nutrient content, leading to considerable influence on plant communities and furthermore on biodiversity and ecological resilience (Bohlen et al., 2004). But, despite their importance, earthworms and their origins cannot be traced through palaeontological means, and their history and origins are nearly completely unknown.

Earthworms have one of the poorest fossil record of all animals, being soft-bodied, and living in aerated terrestrial environments that promote decay and render the type of anoxic burial that leads to geological preservation virtually impossible. Accordingly, there is no known body fossil record for earthworms (Retallack, 2008), and important questions remain unanswered. When did annelids first become terrestrialised? Subsequently, when did earthworms arise as a major group, and can we see any correlation with macroecological events? Given the lack of fossil record for earthworms, can we suitably calibrate molecular divergence time estimations? These questions are important because terrestrialisation requires substantial evolutionary adaptation to life on land (radically different from marine evolutionary pressures), and because palaeontologists are interested in the context and impact of organisms that may well prove to become ecosystem engineers. We seek to understand if currently ecologically influential species had revolutionary effect in previous environments.

In chapter 4, we use newly acquired sequence information to expand upon the earthworm phylogeny, and infer the terrestrialisation of annelids and the origin of earthworms through molecular clocks. Responding to our results, we compare earthworm origins with inferred timings of origin of plant-fungi root symbioses, and the rise of modern plant community structures. Together with these independent lines of thought, we propose the earthworm as a contributor to global ecological turnover and significant changes to carbon sequestration towards the end of the Palaeozoic Era.

1.7 The phylogeny of annelids

Relationships among annelids represents one of the most difficult wide-perspective metazoan evolutionary problem. Annelids, classically known as “ringed worms” or “segmented worms”, are lophotrochozoans, animals characterised by a trochophore larval stage, a trait shared with with molluscs, sipunculans, brachiopods and phoronids (Edgecombe et al., 2011). The term “spiralia”, sometimes used synonymously with lophotrochozoa, refers to spiral cleavage of the blastula stage of these organisms. There is continued controversy over the phylogenetic application of cleavage-stage developmental characteristics, thus here we apply the term “lophotrochozoa” in this thesis: comprising annelids, molluscs, lophophorates and cycliophores. After molluscs and platyhelminthes, annelids are the third most species-rich lophotrochozoans, and reflecting this their ecologies range from terrestrial forests to marine abyssal trenches. Although annelids might not have the scale of species richness as seen in insects, annelids are arguably the more diverse clade. Their ecologies range from the more common heterotrophic strategies of herbivory, scavenging, predation and parasitism, to less commonly seen lifestyles for generally motile organisms, such as filter feeding, bone-boring, and symbiotic interactions with chemoautotrophs at marine hydrothermal vents (Rouse and Pleijel, 2001) (see Figure 1.3 for some examples). Given these wide lifestyles, some of the body-plan morphologies seen in annelids stray wildly from the common vermiform bauplan. For example tube-dwelling *Spirobranchus* “Christmas tree worms” feature feathery spirals of respiratory and filter-feeding structures, at a glance resembling anemone or sea-pens, while *Chaetopterus* is a pelagic drifting suspension feeder, while the abyssal *Osedax* grows “root” structures to feed on the bone of whale-falls. In the marine realm, annelids demonstrate how unexpected evolutionary niches can be exploited. While not explicitly shelled (like many molluscs), some annelids are tube making or secreting, or are rock borers, and because of this they have an appreciable fossil record back to their peri-Cambrian origins (Parry et al., 2014).

With evolutionary origins narrated by fossils, these animals provide excellent testbed for studying episodes of fast diversification. However, despite deployment of a range of palaeobiological inference methods, the early diversification of annelids remains uncertain and an active area of research (Struck et al., 2011, 2015; Weigert et al., 2014). Prior to modern evolutionary research on annelids, the worms were classified as being either “polychaete” or “clitellate”, respectively primarily marine and terrestrial, and named on account of the predominance of chaetae. Chaetae are chitinous barb- or hair-like bundles, usually taking the function of locomotion (sometimes in concert with parapodia), as fins or paddles, or for anchoring the organism within defensive burrows or tubes, and even the delivery of venom. Marine annelids tend to feature chaetae, while terrestrial annelids such as earthworms and leeches have highly reduced chaetae which serve as soil-anchoring structures, or are vestigial. However, phylogenetic investigation in the latter half of the 20th century made it clear that classification on broad chaetal characteristics was unsound (Fauchald, 1974). “Oligochaetes” can be more objectively classified as clitellates, on account of the reproductive clitellum structure, a well supported synapomorphy for the group (Martin et al., 2007). Meanwhile, under both morphological (Rouse and Fauchald, 1997) and molecular analysis (Rousset et al., 2007), the clitellates and other annelids do not form reciprocally monophyletic groups. As such, in this thesis, the term “clitellate” is used in preference over “oligochaete”, and “polychaete” is not a strict phyletic term but is generally used to mean “all annelids, except clitellates” (see further discussion in Chapters 4 and 5).

Nevertheless, clitellate monophyly is one of the few clear views of annelid relationships, and highlights the lack of resolution we have elsewhere in the clade. The reason for uncertainty is due to the topologically abrupt phylogeny: important nodes are buried in a thicket of short branches, surrounded by very long evolutionary branches. Vagaries in the literature promote different views, none of which seem well supported or backed-up by morphological inference. Chapter 5 represents a renewed attempt to untangle these relationships, using a greatly-expanded taxonomic sample, two primary methods of dataset curation, and a suite of inference methods in order to investigate annelid systematics and evolutionary origins.

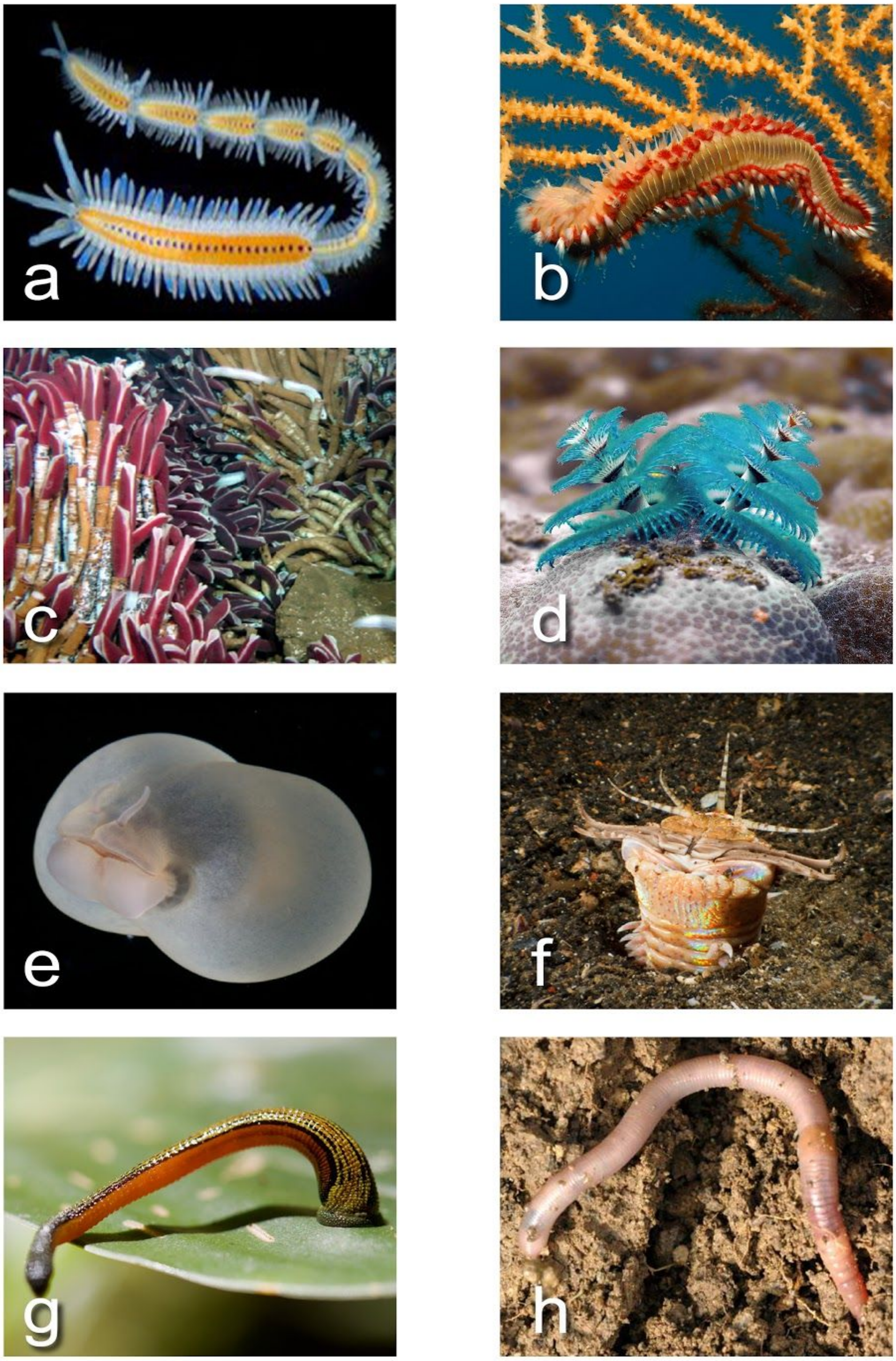


Figure 1.3. Examples of diversity in extant annelids. [a] A member of order Nereidae showing parapodia, and asexual clonal reproduction. [b] Fire bristle worm *Hermodice carunculata* on coral. [c] A colony of giant tube worms, *Riftia pachyptila,* living near a deep-sea hydrothermal vent. [d] Filter feeding appendages of the Christmas tree worm *Spirobranchus giganteus*. [e] The pig-butt worm *Chaetopterus pugaporcinus*. [f] Predatory bobbit worm *Eunice aphroditois*. [g] Terrestrial tiger leech *Haemadipsa picta.* [h] Soil-dwelling earthworm *Lumbricus terrestris.* (All images under creative commons licence.)

1.8 Molecular phylogenetics in the genomic age

Molecular phylogenetics is the science of inferring evolutionary relationships after making observations on the most fundamental heritable information of an organism: characters of its genotype. In many ways, this is merely an extension of the comparative morphological exercise of drawing up a character matrix for a group of organisms, and subjecting it to analysis. In this, observations are made on homologous phenotypical traits (for example, the morphology of skeletal components) with the resulting information arranged as a character matrix: rows of taxa and columns of observations on each *phenotypic* character. Molecular phylogeneticists carry out the same activity, using the raw information of molecular sequence information, and the tools of bioinformatics pipelines, to compile molecular matrices: rows of taxa but columns of observations on each *genotypic* character. Those observations are molecular traits (also termed “characters”, “positions” or “loci” when referring to whole genes) of linear strings of information, and as such are generally known as “sequences”, and the act of recovering the pattern of nucleotides, codons or amino acids is known as “sequencing”. Much of this information can be considered to form a nested hierarchy of information, from the lowest level of the nucleotide, to the transcribed codon, the translated amino acid, or to higher levels such as the presence or absence of a particular protein, gene or gene family. Many of these types of information are available in abundance since the “genomic revolution”.

Technical advances in the first decade of the 21st century, most notably in Next Generation Sequencing, have meant that sequence data which was previously expensive and slow is now accessible to even modestly funded researchers, and as such has spread to nearly all areas of biology. The industry of biomedical, epidemiological, agricultural and evolutionary research means that international repositories such as NCBI GenBank more than double in size annually, and as of 2017, host over two trillion nucleotides in their publicly available information (ncbi.nlm.nih.gov/genbank/statistics/).

Evolutionary biology has benefited greatly from this new wealth of information. For palaeobiology, we are now truly in the age of phylogenomics: phylogenetics using data across the whole genome (Philippe et al., 2005). In the earlier years of molecular phylogenetics, phylogeny was often inferred from the sequence content of single, easily obtainable genes. Today this technique is expanded to multi-gene datasets, bringing with it the benefit of reducing some biases (see section 1.9), and providing models of molecular evolution greater guidance. In any case, for the purpose of phylogenetics it is critical that the information analysed must be an expression of homology, just as in a traditional palaeontological character matrix: each observation must be related through vertical heredity from a common ancestral species.

For molecular data, acquisition, identification and refinement of homologous sequences requires several steps. Firstly, a particular, known gene sequence (from a known species) is used as a template for searching through other sequence information. This is now ubiquitously known as “BLASTing”: using the Basic Local Alignment Search Tool algorithm first developed in the 1990s (Altschul et al., 1990), which has now diversified into a range of search tools. BLASTing can be thought of as similar to searching an electronic document using a “find” command, except it also returns results which can vary by degree from the search query. When using the output of a BLAST operation for phylogenetics, an assumption is that the resulting sequences are similar due to common ancestry.

Naturally, this assumption may be violated, resulting in a matrix which does not express homology for a character, which then cannot drive meaningful phylogenetic inference. Sequence similarity may be due to convergent evolution, or through genetic drift. The latter of these can be assumed to be stochastic, and as such can be dealt with by adding more data: if the underlying process is random then it is unlikely that independent random processes will agree and thus conspire to present a consistent phylogenetic bias. Thus more data, hopefully with a phylogenetic signal, will eventually dilute such random noise (Philippe and Roure, 2011). Convergence is a more contentious issue, as sequence information is often functional and thus subject to selection pressures. However, the expansion of a dataset (particularly in the taxonomic dimension) as well as robust protocol for rejection of dubious data remains our best way of dealing with this kind of bias.

Related to these issues is the problem of assigning homologous relationships to sequence information, when that homology might be of a orthologous (i.e., descent from a speciation event) or a paralogous nature (Gabaldón and Koonin, 2013). Paralogy is the relationship between two sequences which originate from a duplication event within an individual, which then goes on to fixation within the species. This copying can occur at a variety of levels: the highest level is whole genome duplication (relatively common in plants, and not uncommon in animals), through to chromosome duplication, plasmid duplication (in prokaryotes), or at a lower level gene or exon duplication. Common to all of these types of duplication is that it results in two (initially identical) sets of information, that can then evolve independently. So, while those sequences do indeed share common ancestry, that ancestry does not derive from a speciation event, and the information has not experienced the same evolutionary trajectory as the species. To compare paralogues (across species, thus mistaking them for orthologues) is to violate the comparison of like-with-like, as required by orthologous heredity. Furthermore, sequence duplication often leads to exaptation or redundancy of one of the duplicates, and so each copy has its own evolutionary pressures (or lack of). Consequently the sequence content of paralogous duplicates tends to diverge (Holland et al., 2017), possibly in unusual ways (from the point of view of an evolutionary model), making phylogenetic comparison of paralogous sequences even more problematic.

In this set of investigations, the protocol to excise paralogous data from matrices is firstly to curate datasets of slowly evolving conserved genes, making it easier to reject sequences which have changed beyond a certain threshold during the BLAST stage. Secondly, especially for genes of unknown evolutionary rate, a high sequence similarity is required for it to be accepted as a BLAST hit (a positive result). Thirdly, multiple potential hits are taken, and after inference of gene trees the longer-branched or “clearly” misplaced (given uncontroversial phylogeny) sequences are removed. Finally, for larger datasets, a custom Perl script (see Appendix A) is used to automate searching through gene trees and remove long-branch data from the tree-generating matrix. While these approaches will not be perfect, and may even remove some true orthologous data, they in general lead to matrices that provide stronger phylogenetic signal, as appraised from independent lines of evidence and posterior statistical support.

In phylogenomics, the intermediate result of these kinds of processes (known as “curation”) is the supermatrix, the input data for phylogenetic inference. The supermatrix (sometimes known as a “concatenation” or “superalignment”) is merely the end-to-end joining of multiple single-gene alignment matrices to form a longer matrix. As outlined, this matrix is analogous to the palaeontological character matrix, but is typically much larger in its character count. While morphological character matrices might be tens or hundreds of characters long, a molecular supermatrix will typically be tens of thousands, and sometimes much longer. This weight of data has benefits and drawbacks. (We have mentioned some sources of bias, and we will come to some systematic biases in the next section.) But it is also the case that analysis of large datasets, when using Bayesian methodology (see section 1.10), requires considerable computational power, and even then there is a compromise between matrix size and processing time. For this work, we promote the use of matrices with as little missing data as possible, so while at times we are dealing with matrices well over a million characters long (see methods for chapter 5), curation usually leads to matrices with dimensions in the tens of thousands. We uphold this as a current optimum size regarding data quality, matrix completeness, and required computational processing time. It is also a scale which provides a suitable amount of information for evolutionary models to work well, especially seeing as some methods require data, effectively a feedback-loop, to inform a generalised model.

1.9 Phylogenetic philosophy and evolutionary models

The phylogenetic philosophy adopted in this thesis is Bayesian Inference (BI). The primary competing inference methodologies are Maximum Parsimony (hereafter just “parsimony”), and Maximum Likelihood (ML). The suitability of each of these approaches continues to be discussed (Goloboff et al., n.d.; O’Reilly et al., 2016; Puttick et al., 2017; Steel and Penny, 2000), but for the purpose of this work, the nature of the data, and the implications for how results are interpreted, a Bayesian approach can be upheld as more valid, informative and philosophically sound than either ML or parsimony.

Parsimony is the inference of relationships under the assumption that evolutionary change is rare. This assumption is not unreasonable; changes per generation, or even per speciation are small, and evolution almost never makes abrupt side-steps. Evolution is conservative. The inference philosophy under parsimony therefore is merely to present a phylogenetic tree on which the number of changes is minimised, as is the length of the tree (the sum of all branch lengths). Proponents of parsimony highlight that, since no empirical or generalised evolutionary model needs to be applied, the inference cannot be misled by model-misspecification, nor over- or under-parameterisation (Goloboff, 2003). However, in the context of this thesis, we are working with large-scale molecular data, and since the seminal work of Margaret Dayhoff (Dayhoff, 1976) it has been understood that molecular evolution can and should be modelled (see section 1.10).

Further to this, parsimony is highly prone to inferential errors, in particular long branch attraction (LBA) and compositional bias. LBA is the incorrect grouping of species due to similarity arising through large amounts of random change: large amounts of evolutionary change along the long phylogenetic branch is mistaken for similarity through homology. As a metaphor, we could infer the relationships between an imaginary population of fairly-ordered decks of cards. Two well-shuffled (but not closely related) decks are going to have more similarities with each other through chance alone, than they may have with other decks in the population. Being “well shuffled” represents large amounts of evolutionary change: the long branch on a phylogeny. Thus, if a phylogeny of the decks of cards is inferred, the two well-shuffled decks will be “attracted” to each and be inferred as closely related: an artefact merely of there being a large amount of change, rather than similarity through homology.

Sharing similarities with LBA, compositional bias is where inference is misguided simply on the similarity of, say, a G-C heavy genome, not through vertical inheritance of such a composition. The most suitable way to deal with these issues is the application of an evolutionary model, which is not a facet of parsimony. Further to this, expanded taxonomic sample “breaks” long branches (or compositional characteristics) by adding nodes in areas of the phylogeny which would otherwise be reflected as a continuous long branch.

For deriving a phylogenetic tree from molecular sequence information, Maximum Likelihood provides a statistical framework. Unlike parsimony, which commits solely to a philosophical dogma, ML employs a mathematical approach which must be guided by a model of evolutionary change. To generalise, in the case of molecular evolution, a model declares how likely certain changes are within the data, and thus not all individual changes are given the same influence over the inference. To summarise the nature of an evolutionary model, we can use a trivial morphological analogy. We might want to infer the relationships between some species of butterflies based on wing colour. Our model might declare that evolving between having blue, yellow or green wings is relatively likely, whereas evolving to become red-winged is very rare. When reconstructing the phylogeny, the model then places all red-winged butterflies in a single, monophyletic clade: it is more likely that the evolution to being red-winged happened only once (and all red-winged species are descendants of that red-winged ancestor); it is less likely that red wings evolved multiple times in independent lineages. In this way, the overall likelihood of the phylogenetic tree, given the data and the model, is maximised by the ML algorithm. For molecular data, the observations are not on high-level phenotypes (like wing-colour), but on molecular traits, such as the identity of a particular nucleotide at a particular position in a DNA sequence.

Bayesian Inference is related to ML, in that a model is applied to quantify likelihoods of observations in the data. The crucial difference is that in BI, a *prior* is a component of the statistics than guides the inference. The prior represents a probability of what is believed to be true about biological reality. This guides the process of inference, leading to a *posterior*, which is the prior modified in light of the data and the model. Thus the results of a BI analysis are expressed as *posterior probabilities* (PP) of the result, in this case the probability of a particular phylogenetic tree (or a specific branching event on the tree), given the model and the data. An issue arises in deriving these PPs, in that the “given model and data” might be enormously complex, needing to mathematically derive and define the topology, the lengths of branches on the tree, the likelihood of the data, the exchangeability frequencies of the data, and all other facets of the evolutionary hypothesis.

Considering the intractable nature of this mathematics (requiring integration across many parameters), BI instead turns to an iterative method known Markov Chain Monte Carlo (MCMC). To oversimplify, this is a trial-and-error generator. More objectively, an MCMC incrementally proposes new parameter values, tests the data and the model and (assuming the phylogenetic signal is sufficient, and the model of appropriate fit) approaches the equilibrium distribution for each parameter, and thereby the phylogeny with the highest posterior probability. This can be conceived by the Markov Chain (the series of states to be statistically tested) moving through candidate hypotheses to explain the data. At each step, a new value is proposed for each parameter, being a modification of the value in the previous step. If this new parameter has a higher likelihood than in the previous step it is retained, but if it is less likely, it is (usually) rejected. “Monte Carlo” refers to the “gambling” aspect of the algorithm. The chain is attempting to find the highest “peaks” of probability, but a local peak might not be a true global peak, which may only be found via movement across “valleys” of low probability. Thus it can “gamble”, occasionally accepting a lower likelihood move, in the speculation that this may lead to a higher likelihood area of parameter space that would otherwise remain unknown, if only “uphill” moves were accepted. The chance of a “deleterious” proposed move being accepted is proportional to how “bad” the proposal is: a modest drop in likelihood has a higher chance of being accepted, while radical drops are more likely to be rejected. Crucially though, once the MCMC algorithm reaches what it considers to be optimality, the frequency of the parameter values it visits describes the probability distribution of that value, and this is known as the stationary distribution.

The benefits of the Bayesian approach to phylogenetics is two-fold. Firstly, prior knowledge can be applied, and this is particularly crucial to molecular clocks since the known ages of fossils inform the rest of the analysis, thereby providing a scaling framework on which both the calibrated nodes and the inferred nodes can return a probability distribution. Secondly, the statistical support on nodes can more easily be interpreted, compared to methods of node support through parsimony and ML. For these latter methods, node support is usually a case of making pseudo-replicates of the dataset, and repeating the inference. This is known as “bootstrapping”, whereby the original dataset (the character matrix, or supermatrix) is chopped up (vertically, so that we have a shuffled pack of individual characters), and randomly rearranged to create multiple new datasets of the same dimensions as the original set. In doing so some of the characters might be repeated, or totally omitted. These pseudo-replicate sets are then inferred through the same method, and the differences in phylogenetic topology compared to generate a consensus tree. The issue with the bootstrap approach to node support is that it does not provide confidence values for each node. Instead, it lets us know whether any particular node is strongly driven by some of the data, or not. This in itself might be interesting, but it does not constitute a support similar to, say, a *p*-value, as bootstraps are commonly mistaken to be analogous, and thus is evolutionarily not as helpful as it first seems.

Bayesian Inference however returns posterior probabilities for whatever task it has been assigned to. In phylogenetics, this means that a given node is assigned a PP which reflects the probability of the inferred node, given the model and the data. For molecular clock inference, BI returns a probability distribution, giving an interpretable range of times for when a divergence event happened, and whether that distribution is, say, widely spread with long tails of possibility, or highly constrained to a particular date with limited tails. In general, the strengths of BI as a phylogenetic inference philosophy is that it returns evolutionarily interpretable results.

1.10 Evolutionary models

ML and BI require an evolutionary model. For molecular data inferred in a phylogenetic context, the model describes how molecular characters are likely to change through time. There are two processes for generating a model of molecular evolutionary change: empirical and generalised. For an empirical model, real-world data (in this case, aligned matrices of molecular sequences) are analysed so that the probabilities of certain types of change can be quantified. For example, it might be seen in data that an amino acid position of lycine has an elevated chance of changing through time, but only to, say, the amino acids glutamine or serine. By looking through large amounts of empirical data, these kinds of observations (probabilities of substitution) are mathematically and objectively defined. There are many empirical models now available, some are specific to certain research questions, for example epidemiological or mitochondrial sequence evolution. For the focus of this work, the most relevant model of amino acid sequence evolution is that of Le & Gascuel (LG) (2008), usually returned as the best model of fit of metazoan nuclear sequence information under an ML framework. Contrasting to this is the generalised model. This views any given matrix as an isolated case of molecular evolution, and estimates the mode of character substitution from the dataset itself. A popular example of this is the Generalised Time Reversible (GTR) model (Tavaré, 1986). This type of approach is suitable for large datasets (since the data provides the generalised model ample information to drive model parameters), which might well deviate in evolutionary “behaviour” expected by an empirical model such as LG.

Both empirical and generalised models result in an exchangeability matrix (not to be confused with gene matrix or supermatrix), which mathematically describes the likelihood of any substitution within the data. For example, for nucleotide data, the most basic model would be to enforce the assumption that the chance of a, say, nucleotide G being substituted with a C or A or T is the same (Jukes and Cantor, 1969). However, for a variety of biological and biochemical reasons, these likelihoods are not equal; the simplest case being that a purine base is more likely to be replaced by another purine (and the same for pyrimidines) (Felsenstein, 1981). This unequal exchangeability is the simplest example of modelling molecular evolution, and is the first step in building up a picture of molecular evolution based on the exchangeability matrix.

For nucleotide data the dimensions of the matrix is four by four: the likelihoods of exchange between the four nucleotides (three being actual changes, and the fourth being the stationary frequency, or the chance of remaining the same). For amino acid data, this matrix is expanded to 20 by 20, representing the 20 different amino acids used by nearly all organisms. This raises the point that amino acid sequence information is only applicable to translated parts of the genome, i.e. sequence information which codes for a polypeptide chain of amino acids, the protein product. The great majority of a eukaryote’s genome is not destined to produce a protein product, for example less than 2% of the human genome is protein coding (Claverie, 2001). However, for phylogenetic purposes, this represents a type of data-filtering, since it is easier to have confidence of homologous nature of sequence information if that sequence has an understood, or at least identified, role with some kind of cellular protein-mediated function. But, although biological function of sequence information is not a focus of phylogenomics, the way in which sequences evolve through time is.

The exchangeability matrix models the site-by-site nature of evolutionary change, and this is often augmented by modelling the rate of evolution across the matrix. This is normally done by declaring a proportion of the matrix as invariant sites (fixed for all taxa), and the rest to be of a rate of change which can be mathematically defined as a gamma distribution (Liò and Goldman, 1998). These two properties of the model, respectively known as “I” and “G”, are usually estimated from the data. Application of I and G contributes to the model by relaxing the assumption that rates of change across sites is consistent: different areas of sequences, or genes, naturally evolve at different rates, and failure to model this heterogeneity leads to biased phylogenetic inference. Reflecting this, for the type of data used in these studies, model tests through information-criterion tests nearly always support both I and G as being part of the model. Details of models and model testing are discussed further in methods sections of chapters 2, 3, 4 and 5.

As a final point, there are some sequences which are appreciated as being prone to biases, and are often omitted from phylogenetic inference. A notable example of this is mitochondrial data. In the earlier days of molecular phylogenetics, sequences from mitochondrial DNA was used, primarily because it was easier to obtain than nuclear sequences. However, today it is known that mitochondrial DNA is under unusual substitution dynamics (existing in a high-energy environment), undergoes unusual information loss (usually through information relocating to nuclear chromosomes), as well as being exposed to poorly-understood selective sweeps each generation along the maternal line. Further to this is that mitochondrial DNA is not very extensive, so phylogenomicists have turned to nuclear DNA to build their matrices.

1.11 Molecular clocks and fossil calibrations

It became clear in the earliest days of molecular biology that sequences could not only inform phylogenetics, but also be used to infer the time elapsed since speciation events (Donoghue and Benton, 2007; Kumar, 2005; Zuckerkandl and Pauling, 1962, 1965). On a phylogenetic tree, the length of any branch is proportional to the amount of evolutionary change observed along that branch. From a molecular perspective, the branch length expresses the number of substitutions inferred to have occurred between two nodes. Since the majority of molecular substitutions are neutral or nearly-neutral (Kimura, 1968), the amount of difference between homologous sequences from isolated species should be proportional to the amount of time elapsed since the species diverged: the amount of time that the species lineages have had isolated, independent evolutionary history. This generalisation is the basis of the molecular clock (Thorne et al., 1998; Thorne and Kishino, 2005).

Since its origin, molecular clock methodology has developed in its sophistication (Kumar, 2005). A significant step was the introduction of “relaxed” molecular clocks, in response to the objection that time and change are not always proportional: rates can be vary both between lineages, and through time (Bromham, 2009). More recently, molecular clocks have been integrated into Bayesian inference packages, allowing fossil observations to represent prior knowledge, and be defined as having particular probability distributions. For instance, given two fossils from phylogenetic bracketing species, an intermediate node can be assigned a prior uniform probability distribution between these two known times. In many cases, only a minimum age can be provided by a fossil, with the prior stating that there is zero probability of the node being later than this point (as the fossil by definition declares that speciation must have already occurred), but assigns a probability distribution pre-dating that fossil to reflect uncertainty on how much earlier that speciation might have been (Warnock et al., 2012). This distribution is then modified by the data and the model, to provide a posterior distribution on the age of that speciation event (which may even allow for the node post-dating the calibration, as an expression of doubt on the fossil identity and timing itself). Given this influence over molecular clock inference, careful consideration must be given to fossil calibrations (Parham et al., 2011).

Fossil calibration is usually a case of declaring some nodes on the tree to have to have occurred at, between, or (most commonly) before a particular point in time. For example, we might find a fossil with characters synapomorphic to the genus *Homo*, and be able to geologically date it to being 6 million years old. This information then becomes the prior, that we have a strong belief that *Homo* and chimpanzee ancestors speciated more than 6 million years ago: this is a minimum calibration constraint. Ascribing a maximum constraint is more difficult, and it is clearly unsound to apply a timing constraint, say, because no fossils of a particular taxon are found prior to that time (absence of evidence is not evidence of absence, of course). A maximum constraint may, although, be provided by phylogenetic bracketing with a fossil observation along another lineage. More commonly, the maximum is not a constraint but defined by a probability distribution. That distribution might promote the idea that speciation marginally pre-dated the fossil observation (an exponential distribution), or that a certain time has elapsed (a hump-shaped gamma distribution), or even a flat "uninformative" prior, that all dates before the fossil observation are possible (a uniform prior). These calibrated nodes then allow temporal rescaling: converting branch lengths from representing evolutionary change (in molecular data, the amount of substitutions between nodes), to time elapsed between speciation events.

Together with the inference of phylogeny, independent observations on macroecological conditions, and temporal inference through molecular clocks, we can derive a meaningful insight on evolution for our groups of interest. In the empirical research of Chapters 2, 3 and 4, we use molecular clocks to investigate the origin of groups which have notably poor fossil records. Thus, an auxiliary goal, particularly for Chapter 4 (earthworms), is to test inference molecular clocks which might not have ideal calibration regimes.

1.12 Descriptions and explanations of macroevolution

This thesis is an application of these tools to investigate evolution for specific groups of animals in the Phanerozoic, the aptly named era of "visible life" of the past ~540 Ma. While we now understand that life on Earth has a prokaryotic history back to at least 3.5 Ga (Djokic et al., 2017), the bewildering diversity of life that *we* can see on a human scale, to many people, is more relatable. So naturally, we seek to determine the truth of how life has expanded from the first appreciable multicellular organisms, and their visible ecologies, into the world we have today, where every conceivable niche and lifestyle seems to be ingeniously exploited. (It is worth keeping in mind that “conceivable” might be a failure of imagination; for all we know, life on Earth may be ecologically obtuse and abjectly conservative in diversity, but, given a dataset of one inhabited planet, we will have to embrace the positive (or conceited) view that life seems pretty diverse, seen through Earthling eyes.) As such it is worth investigating, describing and maybe explaining visible life.

Drivers of change over large evolutionary and temporal scales are of course exceptionally complex, and will probably forever remain impossible to firmly understand. Palaeobiology then must place itself in a position of describing the past, rather than always offering decisive explanations for the past. In their classic paper, Gould and Lewontin cautioned biologists against seeking and assigning adaptive function to every feature of an organism (Gould and Lewontin, 1979). Instead, organismal characteristics should be considered imperfect, provisional, pliant, and sometimes pointless. Scaling up that sentiment, evolution and ecology, especially on its widest perspective, should be considered composed of contingent species and processes, not one encased in a Panglossian paradigm where each and every action has a reason and a reaction. We should not assume that every ecology, past and present, has been the result of graspable interplay between organisms, or abiotic factors. Many macroecological trajectories will remain indecipherable, and suggesting cause and effect on these scales may well be impossible, even given hypothetical perfect knowledge.

Nevertheless, science will always be a process of hypothesis testing: the consensus of the present is merely science which has not been proved wrong yet. So, while the vast complexity of global ecology over enormous timescales might be essentially unknowable, there is scientific value in proposing scenarios, even if only for scientists to employ the methods of refuting such arguments. With this in mind, the strongest successes in this thesis are in offering hypotheses for the state of the past. Extending from that, to actually *explain* the past is an activity of proposing the “least wrong” answer, but, given the complexity of these topics, we promote the view that explanations are at least in part speculation to motivate scientific discussion and reaction. To say the same thing more positively, the chapters in this thesis are the best we can currently do, and, after all, speculation is the most human counterpart of an otherwise cold scientific process. If science progressed like evolution, with no recourse to hunches or intuition, no sparks of imagination, it would be a very slow process indeed.

Chapter 2

Molecular clocks indicate turnover and diversification of modern coleoid cephalopods during the Mesozoic Marine Revolution.

A version of this chapter has been published in *Proceedings of the Royal Society of London B: Biological Sciences* (Tanner et al., 2017)*,* in collaboration with Dirk Fuchs, Inger E. Winkelmann, M. Thomas P. Gilbert, M. Sabrina Pankey, Ângela M. Ribeiro, Kevin M. Kocot, Kenneth M. Halanych, Todd H. Oakley, Rute R. da Fonseca, Davide Pisani, and Jakob Vinther. The investigation was devised and developed by Alastair R Tanner, Jakob Vinther and Davide Pisani. Samples were collected and provided by JV, MTPG, IEW, MSP, AMR, KMK, KMH, THO and RRF. Sequencing was carried out at University of Copenhagen molecular biology facilities. Molecular data curation and all computational analyses were carried out by ART. Bioinformatics scripting was written and developed by ART, further details in Appendix A. Interpretation of results was carried out by ART, JV, DP and DF. The manuscript was authored by ART, JV and DP, with further input from all others on the author list. For further details, see the paper's entry in Appendix B.

Abstract

Coleoid cephalopod molluscs comprise squids, cuttlefish and octopuses, and represent nearly the entire diversity of modern cephalopods. Sophisticated adaptations such as the use of colour for camouflage and communication, jet propulsion, and the ink sac highlight the unique nature of the group. Despite these striking adaptations, there are clear parallels in ecology between coleoids and bony fishes. The coleoid fossil record is limited, however, hindering confident analysis of the tempo and pattern of their evolution. Here we use a molecular dataset (180 genes, ~36,000 amino acids) of 26 cephalopod species to explore the phylogeny and timing of cephalopod evolution. We show that crown cephalopods diverged in the Silurian-Devonian, and the crown divergence of coleoids had origins in the latest Palaeozoic. While the deep-sea vampire squids and dumbo octopuses have ancient origins extending to the Early Mesozoic Era, 242 million years ago (Ma) ± 38 million years (Myr), incirrate octopuses and the decabrachian (ten armed) coleoids diversified in the Jurassic Period. These divergence estimates are in agreement with the available fossil record, and consistent with a modern diversity of coleoid cephalopods primarily emerging in the Mesozoic Marine Revolution, a period that also witnessed the radiation of most ray-finned fish groups in addition to several other marine vertebrates. This suggests that that the origin of modern cephalopod biodiversity was contingent on ecological competition with marine vertebrates.

2.1 Introduction.

Octopus, cuttlefish and squid are the coleoid cephalopods, marine molluscs characterised by jet locomotion, sophisticated camouflage and mimicry, decoy countermeasures in the ink-sac, high intelligence, and a wide range of body sizes (Mather and Kuba, 2013b). Charismatic in these ways, and due to their importance as fishing stocks, cephalopods have been the focus of research from marine ecologists and evolutionary biologists. However, cephalopod evolutionary relationships and the timing of the origin of the modern coleoid cephalopods have remained unclear, in part due to uncertainties in their fossil record. This is in contrast to some of their extinct relatives, the ammonites and belemnites, which provide a fossil record so complete as to be used as a major tool in zoning for biostratigraphy.

The past 540 Ma (million years) of cephalopod evolution can be viewed as having three ecologically distinct phases. Originally shelled, sea-floor-dwelling molluscs, cephalopods are descended from superficially limpet-like ancestors in the Cambrian (Vinther et al., 2012; Yochelson et al., 1973). The protective shell later became adapted as a chambered buoyancy organ (Mutvei et al., 2007), giving rise to free-swimming forms that could then make the shift from benthic to pelagic ecologies, and by the latest Cambrian these radiated into several Ordovician lineages (Kröger, 2005). From that time and into the Mesozoic, an enormous diversity of heavily-shelled cephalopods took on pelagic ecologies. Subsequently, internalisation and reduction of the mineralised shell led to a diminishing fossil record for shelled cephalopods, which nearly entirely vanishes by the Cenozoic. It is hypothesised that this adaptive trajectory facilitated adoption of alternative ecologies in the coleoids (Boyle and Rodhouse, 2008).

Anatomical evolution is in part shaped by the ecological relationships between predator and prey species. Cephalopods take an intermediate position in food webs as both predator and prey, and in their role as prey they sustain a range of predators, especially vertebrates. Ecologically, cephalopods (and in particular oceanic squid) occupy a niche that largely overlaps with fish as active mesopredators. Considering the evolutionary narrative of cephalopods from heavily-shelled drifters to rapid hunters, the question of how and when this development took place remains unresolved. Previously, coevolution between marine predators and prey has been hypothesised from the fossil record of the Jurassic and the Cretaceous, and this ecological shift has since become known as the Mesozoic Marine Revolution (Vermeij, 1977, 1987).

Despite these views on cephalopods and their ecologies, the fossil record leaves limited insight on the providence of modern coleoid groups (Strugnell and Nishiguchi, 2007), which stands in contrast to their well documented ancestors and relatives, especially among the iconic ammonites and belemnites. The mineralised, chambered portion of the cephalopod shell (phragmocone and rostrum) has a high potential for preservation, but as the phragmocone became reduced, internalised (and in many cases lost entirely in extant coleoid lineages) so too was a clear narrative through fossils. Soft tissue fossilisation is rare, but cirrate and incirrate octopods are known from the Late Cretaceous (Cenomanian) Hâkel and Hâdjoula Lagerstätte, while cirrate forms and stem octobrachians are recorded in the Jurassic (Fuchs et al., 2015); these are known to preserve the unmineralised gladius and soft tissues. Stem group decabrachians, such as belemnites and other belemnoids are known, preserving their phragmocones and, occasionally, soft tissues (Clements et al., 2016). In contrast, the extant octopuses, cuttlefish and squid are characterised by shell-reduction and loss (Lindgren et al., 2012), and are prone to major taphonomic biases in tissue preservation (Clements et al., 2016). Consequently, clarifying evolution of coleoids from the mid-Palaeozoic to the present must rely on alternative palaeobiological approaches, such as the estimation of molecular divergence times.

The first molecular divergence times of cephalopod evolution recovered very ancient divergences for the coleoids (Strugnell et al., 2006), suggesting extensive gaps in the fossil record. However, these studies used controversial calibrations from the late Palaeozoic, such as *Shimanskya (Doguzhaeva et al., 1999)* and *Pohlsepia* (Kluessendorf and Doyle, 2000), for which the assignment to the coleoid crown group is dubious (Kröger et al., 2011). Subsequent studies attempted to estimate cephalopod divergences using calibrations from outgroups, such as bivalves and gastropods and recovered much younger divergence estimates, that were surprisingly congruent, irrespective of differences both in methodology and gene sampling (Kröger et al., 2011; Warnke et al., 2011). These independent studies recovered a divergence between the nautilids and the coleoids around the Silurian-Devonian boundary, or the earliest Devonian (~415 Ma), which is congruent with unequivocal evidence for fossil stem group coleoids (ammonoids and bactritids) (Klug et al., 2015; Kröger and Mapes, 2007) and stem group nautilids (Jerzy Dzik and Korn, n.d.) in the early Devonian. Cephalopod beaks also appear in the fossil record in the Devonian (Klug, Frey, et al., 2016). These observations suggest that the fossil record documents the origin of the crown group and that the concomitant evolution of the beak (Kröger et al., 2011) coincides with a dramatic shift in predator-prey dynamics, termed the Devonian Nekton Revolution (Klug et al., 2010). The jawed vertebrates radiated at this time, incident with a global shift in predatory style towards increased high-metabolism predation and durophagy (Bush and

Bambach, 2011), the dental and cranial adaptation for dealing with heavily-armoured prey. The coincidence of jawed vertebrates and beaked cephalopods radiating at the Silurian-Devonian boundary may thus be interpreted as a response to the changes in the predator-prey ecological landscape.

For this study we shed further light on the providence of coleoid cephalopods using molecular sequence information to infer a new phylogeny of 26 coleoid cephalopods, plus an outgroup of 30 species of molluscs and annelids. This tree then informs a molecular clock analyses, guided by fossil calibrations as suitably supplied by our robust outgrouping regime.

2.2 Methods and materials.

Data acquisition

Genomic data for *Architeuthis dux*, and transcriptomic data for *Onychoteuthis banksi* and *Sthenoteuthis banksi* were generated and assembled by the Sequencing Centre of the University of Copenhagen using the Qiagen RNeasy extraction protocol, and the TruSeq RNA Kit v2 RNA isolation, cDNA synthesis, ligation and PCR-amplification protocol. Quality control was carried out using a Bioanalyzer 2100, Agilent Technologies. Transcriptomic sequence data from *Bathypolypus arcticus, Grimpoteuthis glacialis, Lolliguncula brevis* and *Spirula spirula* were generated at the University of Bristol Life Sciences Sequencing Facility. For these species, the Trizol extraction protocol was applied, with sequencing carried out on the Illumina HiSeq platform, 100 base pair read length, paired end reading. Transcriptome assembly was carried out in Trinity version 2.0.3 (Grabherr et al., 2011) under default parameters and using Trimmomatic (default parameters, as part of the Trinity package) for quality control.

For the remainder of the taxonomic sample (see Table 2.3 for full list and accession codes), NCBI GenBank was searched for cephalopods. To root and calibrate the cephalopods, further species were selected as outgroup on account of sequence data availability and taxonomic proximity to the cephalopod ingroup. Thus the outgroup was composed of bivalves, gastropods, a scaphopod, plus four annelids (see Table 2.1, Figures 2.4, 2.5).

Matrix completeness and NCBI accessions

Species	#genes	% Completeness	AA positions	Accession
Aplysia sp.	180	100.00	36156	PRJNA253054
Hydatina sp.	180	100.00	36156	PRJNA253054
Phallomedusa	180	100.00	36156	PRJNA253054
Oxynoe sp.	179	99.44	36086	PRJNA253054
Tylodina sp.	179	99.44	35986	PRJNA253054
Philine sp.	179	99.44	35792	PRJNA253054
Sepia esculenta	176	97.78	35449	SRR1386223
Monodonta sp.	175	97.22	35406	PRJNA253054
Urosalpinx sp.	175	97.22	35099	PRJNA253054
Sepia officinalis	174	96.67	34854	SRR1325115
Solemya sp.	171	95.00	34583	PRJNA253054
Rubyspira sp.	172	95.56	34545	PRJNA253054
Hinea sp.	171	95.00	34382	PRJNA253054
Neotrigonia sp.	170	94.44	34298	PRJNA253054
Ennucula sp.	168	93.33	34172	PRJNA253054
Microhedyle sp.	163	90.56	34153	PRJNA253054
Gadila sp.	165	91.67	33487	PRJNA253054
Euspira sp.	167	92.78	33326	PRJNA253054
Granata sp.	167	92.78	33098	PRJNA253054
Dosidicus sp.	153	85.00	32709	SRR1386212
Pareledone sp.	155	86.11	32484	SRR725936
Grimpoteuthis sp.	154	85.56	32382	[Matrix on dryad]
Bathypolypus sp.	153	85.00	32097	[Matrix on dryad]
Euprymna scolopes	157	87.22	31918	SRR3472306
Lolliguncula sp.	147	81.67	31474	[Matrix on dryad]
Sepioteuthis lessoniana	160	88.89	31438	SRR1386192

Pomacea sp.	154	85.56	31129	PRJNA253054
Doryteuthis sp.	140	77.78	30847	SRR3472305
Crassostrea sp.	146	81.11	30643	PRJNA253054
Octopus vulgaris	149	82.78	30588	SRR725937
Myochama sp.	147	81.67	29805	PRJNA253054
Sthenoteuthis sp.	133	73.89	28643	[Matrix on dryad]
Architeuthis sp.	133	73.89	28241	[Matrix on dryad]
Pomatoceros lamarkii	132	71.00	27003	SRR1810802
Mytilus californianus	120	66.67	26680	PRJNA253054
Haliotis sp.	137	76.11	25388	PRJNA253054
Onychoteuthis sp.	110	61.11	25094	[Matrix on dryad]
Perotrochus sp.	94	52.22	22666	PRJNA253054
Alvinella sp.	92	50.72	22144	[Matrix on dryad]
Capitella sp.	91	49.19	21711	SRR4045567
Idiosepius paradoxus	78	43.33	17112	SRR2984343
Chiroteuthis calyx	81	45.00	15630	SRR2102319
Sepioteuthis australis	71	39.44	14835	SRR725780
Salmacina sp.	70	38.35	14635	[Matrix on dryad]
Sepia officinalis	60	33.33	14284	SRR1325115
Abdopus aculeatus	74	41.11	13398	SRR680047
Uroteuthis sp.	72	40.00	13083	DRR068682
Mytilus edulis	68	37.78	12849	PRJNA253054
Octopus cyanea	67	37.22	12314	SRR725937
Galiteuthis armata	57	31.67	11522	SRR2102359
Nautilus pompilius	36	20.00	8321	SRR108979
Vampyroteuthis infernalis	39	21.67	8218	SRR2102472
Hapalochlaena maculosa	30	16.67	6501	SRR3105559
Spirula spirula	25	13.89	6120	[Matrix on dryad]

Loliolus noctiluca	22	12.22	5177	SRR725597

Table 2.1: Data sources contributing to cephalopod supermatrix, plus accession numbers.

Phylogenomic data assembly

We compiled a supermatrix with data from 56 species (see table 2.3) for 197 genes, based and building on a dataset of genes of slow substitution rate and high evolutionary conservation (Philippe et al., 2011). This set was selected due to substitution rate consistency of the gene-sample, and since differences in read-depth of the transcriptomic data would lead to relatively sparse orthology groups. New cephalopod sequences matching those in the Philippe gene-sample were acquired through BLAST (Altschul et al., 1990) searching transcriptomic sequences, using *Aplysia californica* as the search query due to this taxon possessing full coverage for the gene dataset, and having phylogenetic proximity to the group in question. A custom Perl script (available at github.com/jairly/MoSuMa_tools/, also see Appendix A) selected sequences on the most significant expect values (e-values) among BLAST hits, taking the lowest e-values and any other significant hits within three orders of magnitude of the most significant hit. The maximum e-value threshold was set at 10^{-10}, with hits exceeding this being excluded. These selected sequences were aligned using MUSCLE (Edgar, 2004), to produce gene-alignments for each of the 197 genes (see table 2.3). Ambiguously aligned positions were removed from the gene alignments by GBlocks v0.91b (Castresana, 2000) (*b2* = 70%, *b3* = 10, *b4* = 5, *b5* = half).

Orthology assessment

These gene trees were then assessed for orthology applying for following protocol. Maximum likelihood phylogenies were inferred for each gene using PhyML (Guindon et al., 2009) version 3, modelling under LG (Le and Gascuel, 2008) and accepting the best tree of either Sub-Tree Pruning or Neighbour Joining algorithms. Sequences producing long branches were removed from the alignments, with a long branch considered to be more than 2 times the standard deviation of the average away from the average branch length for the gene in question (script available at github.com/jairly/MoSuMa_tools/). 17 genes were considered to have low orthology confidence (due to unresolved gene trees) across the taxonomic sample, and discarded, leaving 180 gene alignments. These 180 gene alignments were concatenated into a single experimental matrix using SequenceMatrix v100 (Vaidya et al., 2011), with a resulting supermatrix of 36,156 amino acid positions across 56 taxa.

To provide alternative topological inference, maximum likelihood approaches were also employed. PartitionFinder (Lanfear et al., 2012) and IQTree (von Haeseler et al., 2012; Nguyen et al., 2015a) were used to test model fit under maximum likelihood, with both returning the substitution model of Le and Gascuel (Le and Gascuel, 2008), with a gamma distribution on rates and a proportion of invariant sites as having best model fit.

Phylogenetic inference

Phylogenetic topology was inferred from the supermatrix using the Markov Chain Monte Carlo (MCMC) sampler PhyloBayes MPI v1.5a (Lartillot et al., 2013). The infinite mixture model CAT + GTR + G_4 was applied, being the most appropriate to deal with across-site heterogeneities, while minimising long-branch biases. Two independent Monte Carlo chains were run, a burn-in of 25% of the Markov chains were discarded. Convergence of these chains was considered achieved with the maximum difference (the *maxdiff* statistic under PhyloBayes) in the bipartitions of the chains < 0.1, as reported by *bpcomp* program in the PhyloBayes package. A further test of convergence was carried out using *tracecomp* (also under PhyloBayes), with effective sample sizes being > 50, and relative differences dropping below 0.1 for all parameters as compared between the independent chains. As a further test of topology, the maximum likelihood software RAxML MPI v8.1.15 (Stamatakis, 2014) was applied to the same dataset as used in Bayesian inference, applying LG + I + G_4. 1000 pseudoreplicates were run. See figure 2.1 for preliminary phylogenetic topology, which also guided divergence time inference.

Phylogenetic inference cross-validation

Cross-validation was carried out to assess model suitability, as described in PhyloBayes documentation, comparing GTR against CAT + GTR. CAT + GTR outperformed GTR, with a score of 18.87 +/- 7.99 (positive scores meaning better than reference model (GTR)). In ten replicates, CAT-GTR outperformed GTR all ten times.

Molecular divergence time inference.

Phylobayes 3.3f was used to infer molecular divergence times using CAT + GTR, soft-bounds of 0.05, and a Yule-process birth-death model. A Bayes Factor analysis of the fit of three alternative models was performed (CIR (Lepage et al., 2007), log-normal (Thorne et al., 1998), and uncorrelated gamma (Drummond et al., 2006), with CIR showing a better model-fit. Of these models, CIR was applied due to its ability to model rate change along branches and between taxa, while avoiding over-relaxation of divergence time inference. CIR also statistically performs better than other clock models under Bayesian cross-validation, see Figures 2.6, 2.7 and later in this section for details.

The topology was fixed to that inferred by PhyloBayes MPI v1.5a, the root constrained to the bifurcation between the uncontroversial monophyletic assemblages of annelids and mollusca (cephalopoda + bivalvia + gastropoda + scaphopoda); the bivalves and gastropods, plus annelids were considered a balanced outgroup (with comparable taxonomic sampling and phylogenetic crown spread). A prior was applied to the root of 565 Ma ± 10 Myr, representing the root of lophotrochozoa (Fedonkin and Waggoner, 1997). This prior was tested as being appropriate by chains being run without data, to confirm that the samples were being drawn from a distribution that includes the prior. The root age of the prior run was 552 Ma ± 8 Myr, supporting the prior as appropriate. Two independent MCMC chains were run for each model, with convergence being determined through *tracecomp*, with effective sample sizes > 50, and relative differences < 1 for all parameters as compared between independent chains. The discarded burn-in was 25% of the chain length.

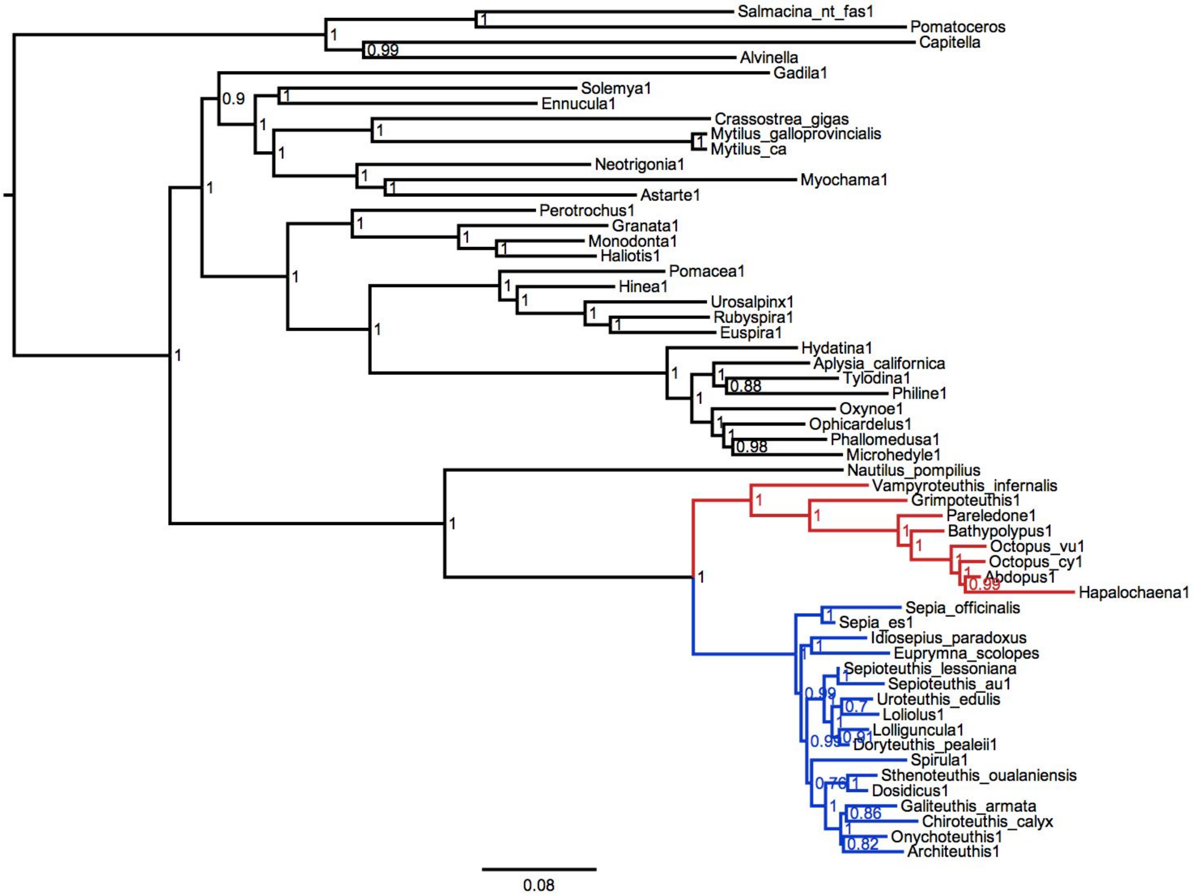


Figure 2.1: Molecular phylogeny of 26 cephalopod species, plus outgroups. 180 genes, concatenated as 36,156 aligned amino acid positions with 26% missing data, modelled under CAT + GTR + G. Numbers at nodes represent posterior probabilities. This topology is used to guide all molecular clock inferences. Black terminal branches represent outgroup taxa (see Figure 2.8 for further details). Red terminal branches represent Octobrachia, blue branches represent Decabrachia (see Figures 2.8 and 2.9 for further details and expansion on systematics).

Fossil calibrations

12 fossil calibrations were applied to the molecular clock analyses. The letters a-l in Table 2.1 below refer to the nodes labelled a-l in Figure 2.2. The root itself was constrained to 555 Ma ± 10 Myr, as evidenced by *Kimberella* of the White Sea Formation (Martin et al. 2000). *Kimberella* is here used to represent the root prior by being the oldest representative of Bilateria, and thus covering the total group under investigation. The root prior was tested as appropriate by running the analysis without data, returning a posterior distribution on the root of mean 554 Ma ± 9.2 Myr, supporting the priors as appropriate.

Node	Maximum	Minimum	Reference	Position
a: Scaphopoda + Bivalvia + Gastropoda	543 Ma	525 Ma	(Benton et al., 2015)	External
b: Scaphopoda + Bivalvia	-	532 Ma	(Nutzel and Bandel, 2000)	External
c: Bivalvia	-	485 Ma	(Parkhaev, 2008), pp. 33-69	External
d: Vetigastropoda	-	490 Ma	(Edgecombe *et al.* 2011)	External
e: Caenogastropoda + Heterobranchia	-	418 Ma	(Frýda et al., 2008), pp. 239-270	External
f: *M. edulis* + *M. californianus*	-	20 Ma	(Gosling, 2015)	External
g: Crown Decabrachia	-	68 Ma	(Klug, Schweigert, et al., 2016)	Internal
h: *Spirula*, plus descendents	-	75 Ma	(Fuchs, Iba, et al., 2013)	Internal
i: Crown Vampyromorpha	-	195 Ma	(Fuchs and Weis, 2008)	Internal
j: Crown Coleoidae	-	240 Ma	(Kröger and Mapes, 2007)	Internal
k: Crown Cephalopoda	-	408 Ma	(Kröger and Mapes, 2007)	Internal
l: Crown Mollusca	549 Ma	-	(Grant et al., 1991)	External

Table 2.2: fossil calibrations guiding molecular divergence time estimations, as shown in Figure 2.10 and 2.11. Minima have a CIR prior distribution (Lepage et al., 2006), while a uniform distribution is applied between calibrations that have a minima and maxima.

“Internal” calibrations are those placed within the coleoid cephalopod crown group. “External” calibrations are on nodes in the outgroup to coleoids.

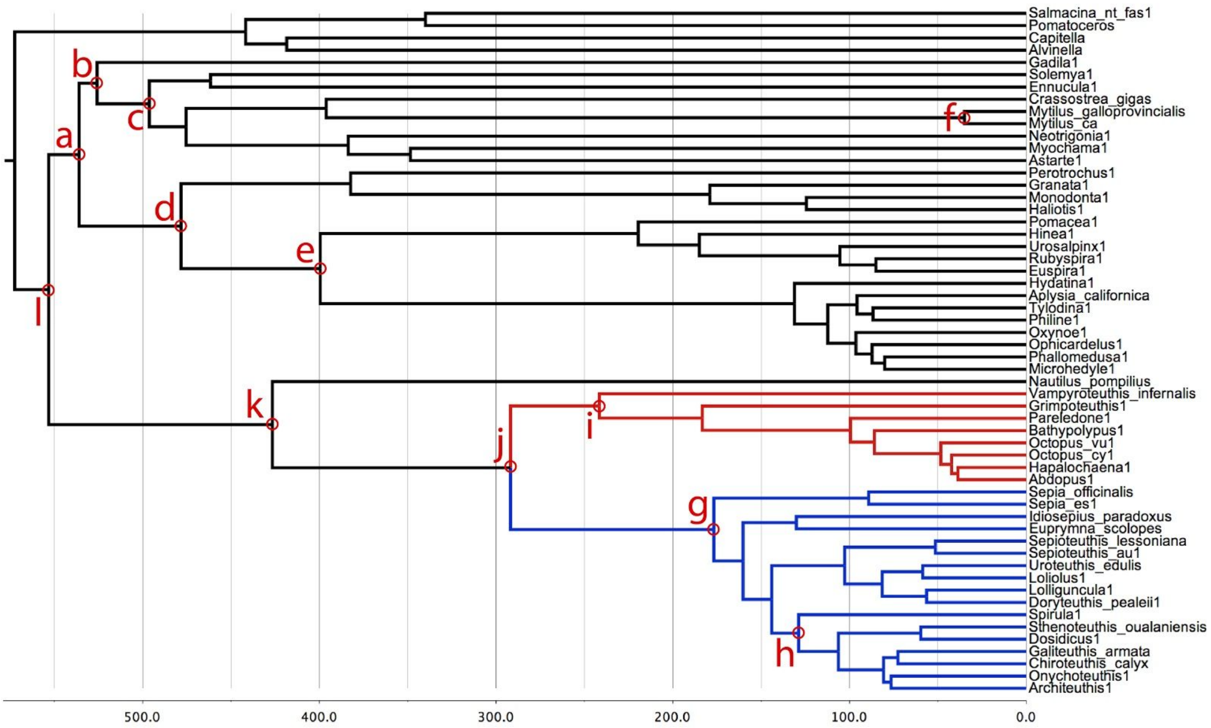


Figure 2.2: Calibration positions. 36,156 amino acid position, 56 taxa, CAT + GTR substitution model, CIR clock model, 12 calibrations (see table 2.1), root prior of 565 Ma ± 10 Myr, soft bounds of 0.05, Yule birth-death process. Outgroup constrained to annelid monophyly.

Clock model cross-validation

Cross-validation was carried out in accordance with guidelines in the PhyloBayes manual, comparing the autocorrelated CIR model against the uncorrelated UGAM model, both under the CAT-GTR substitution model. With CIR against UGAM, Bayes factor returned as 40.87 ± 10.45; a positive number is support for the fit of CIR. Out of ten replicates, all ten supported CIR.

Alternative clock models and calibration schemes

Analyses were run to assess the impact of internal versus external calibrations on divergence time estimations. See Table 2.1 for calibrations classified as “internal” and “external”. Note that “internal” refers to application of *both* internal and external calibrations.

Treatment	**Crown Cephalopoda**	**Crown Coleoidae**	**Crown Octobrachia**	**Crown Decabrachia**	**Oegopsida**	**Myopsida**	**Incirrata**
CIR externally calibrated	434	306	260	202	122	119	107
CIR internally calibrated	426	289	239	173	104	100	98
UGAM externally calibrated	405	264	210	123	74	65	101
UGAM internally calibrated	423	280	220	110	72	66	96

Table 2.3: Inferred divergence ages of key nodes under alternative clock model and calibration scheme.

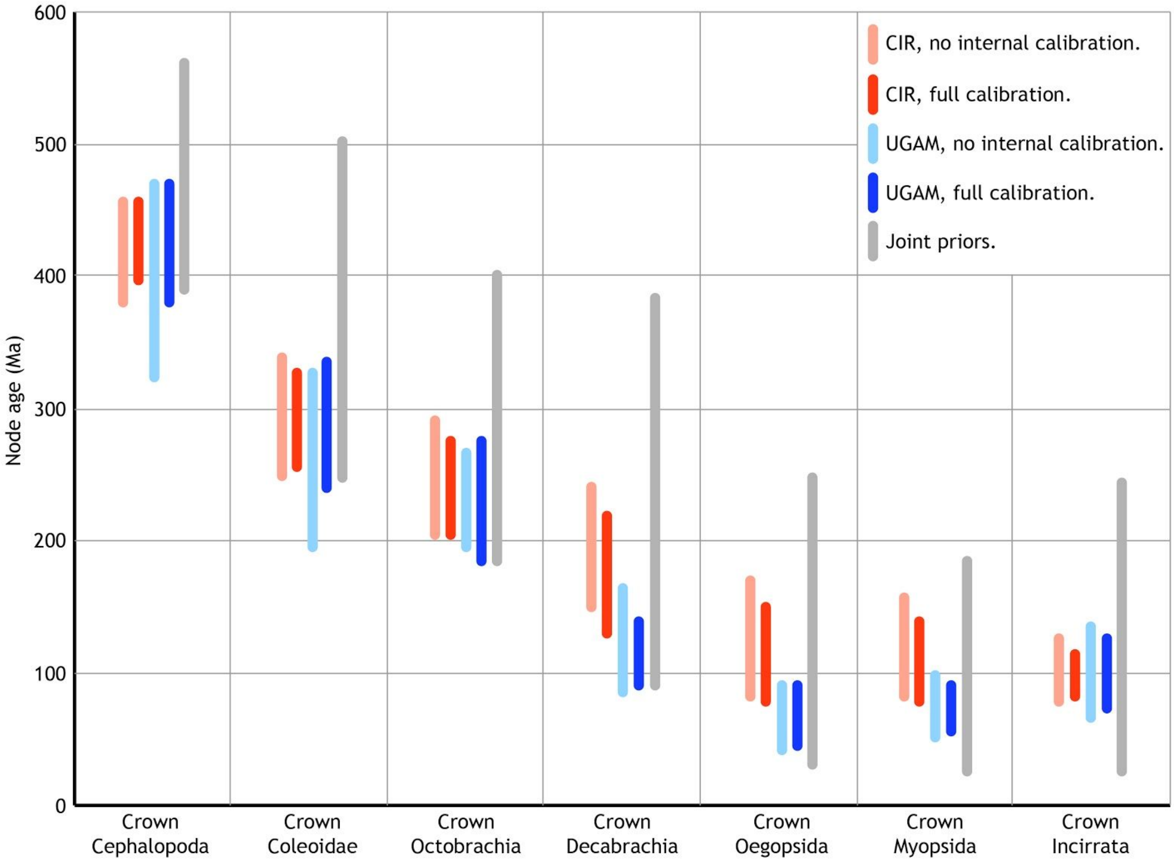


Figure 2.3: Inferred divergence ages of key nodes under alternative clock model and calibration scheme.

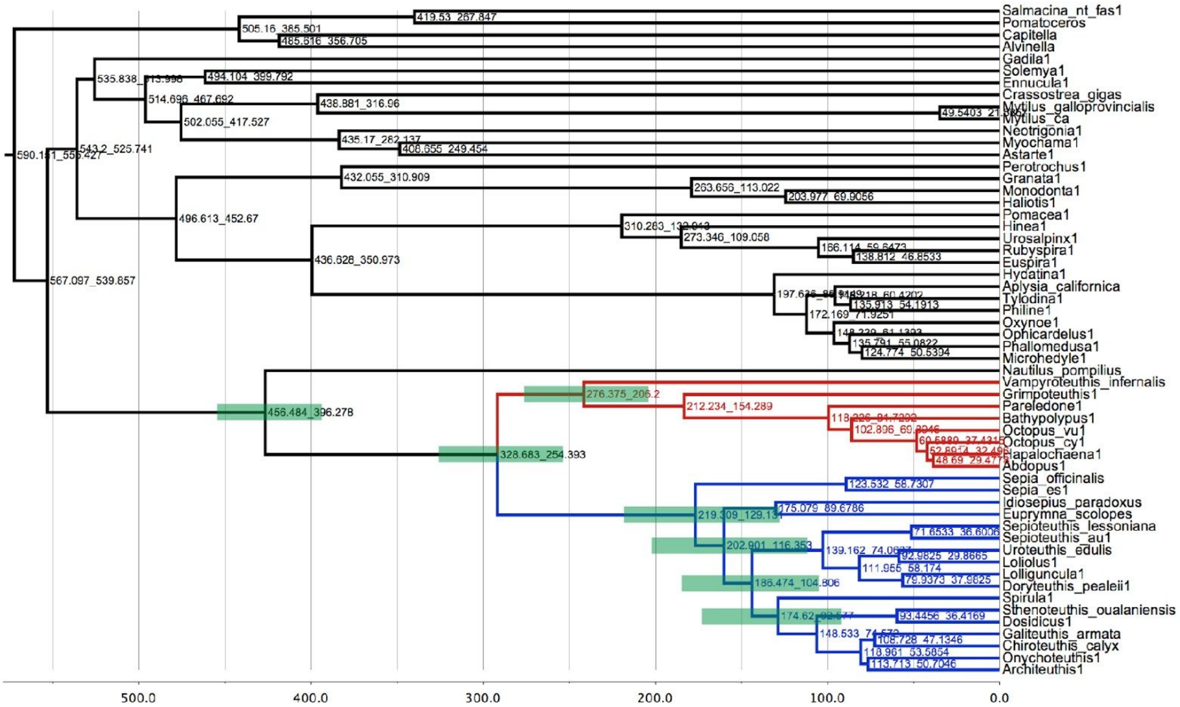


Figure 2.4: CIR clock. 36,156 amino acids, 52 taxa, CAT + GTR substitution model, CIR clock model, 12 calibrations, with annelid outgroup, root prior of 565 Ma ± 10 Myr, soft bounds of 0.05, Yule birth-death process.

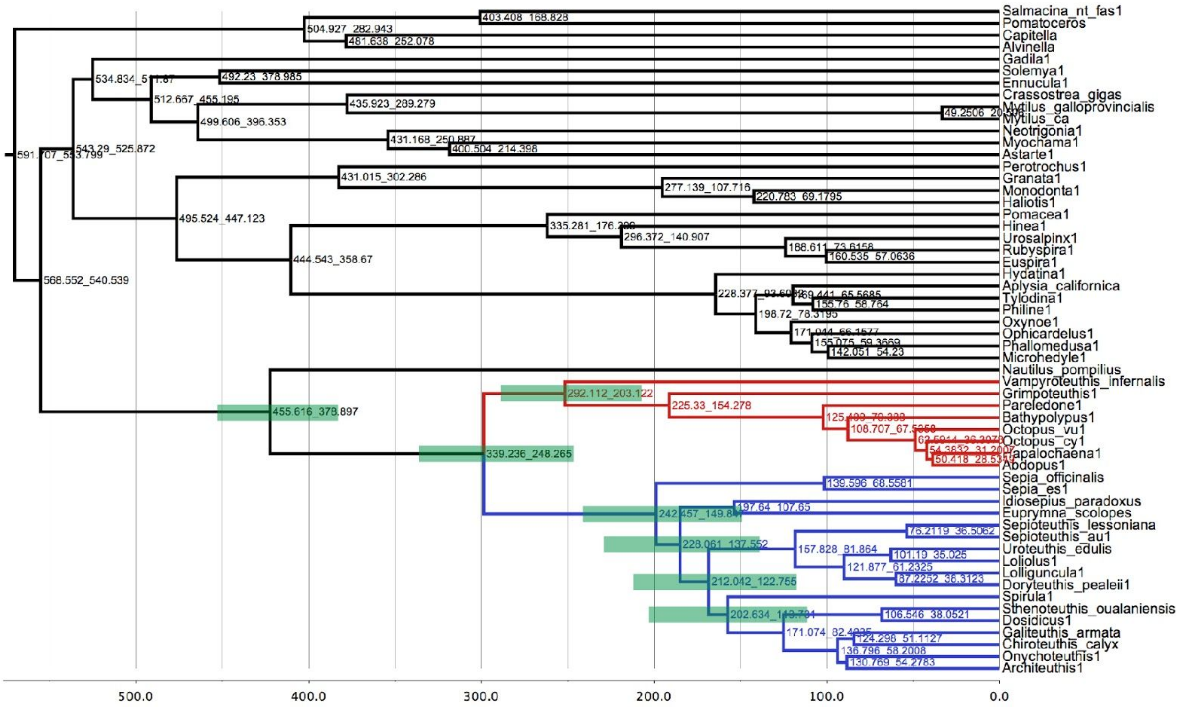


Figure 2.5: CIR clock. 36,156 amino acid position, 52 taxa, CAT + GTR substitution model, CIR clock model, 6 calibrations with none internal to Coleoidae, with annelid outgroup, root prior of 565 Ma ± 10 Myr, soft bounds of 0.05, Yule birth-death process.

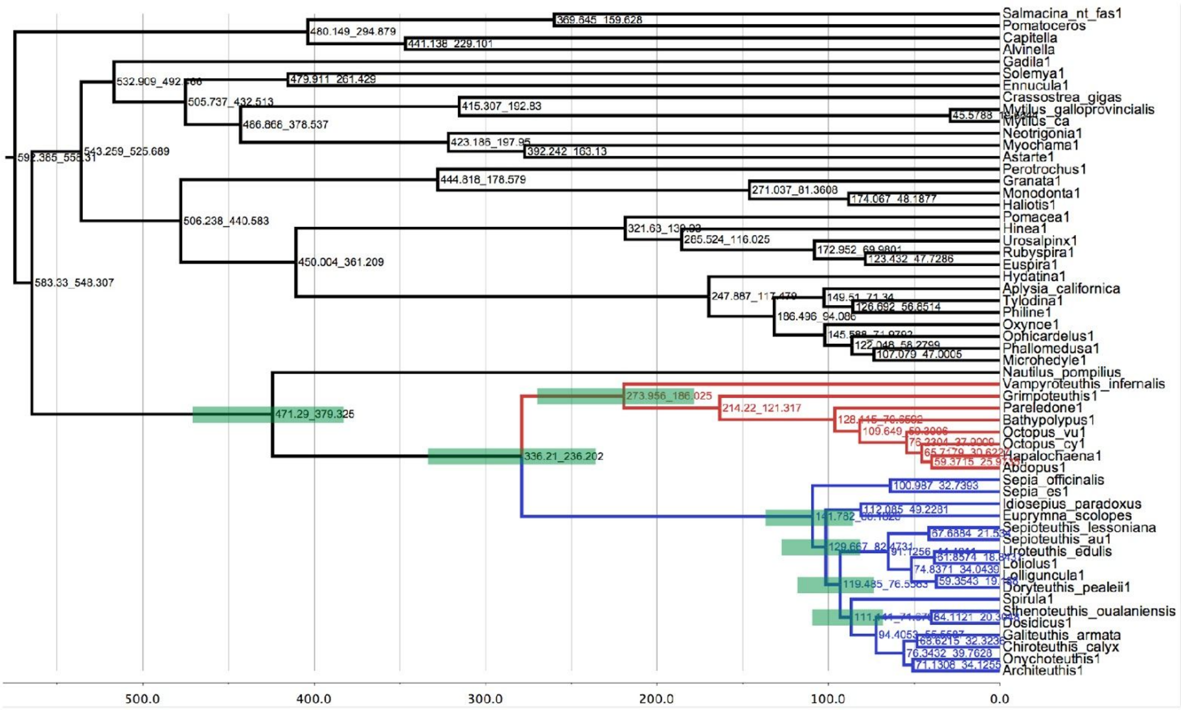


Figure 2.6: UGAM clock. 36,156 amino acids, 52 taxa, CAT + GTR substitution model, UGAM clock model, 12 calibrations, with annelid outgroup, root prior of 565 Ma ± 10 Myr, soft bounds of 0.05, Yule birth-death process.

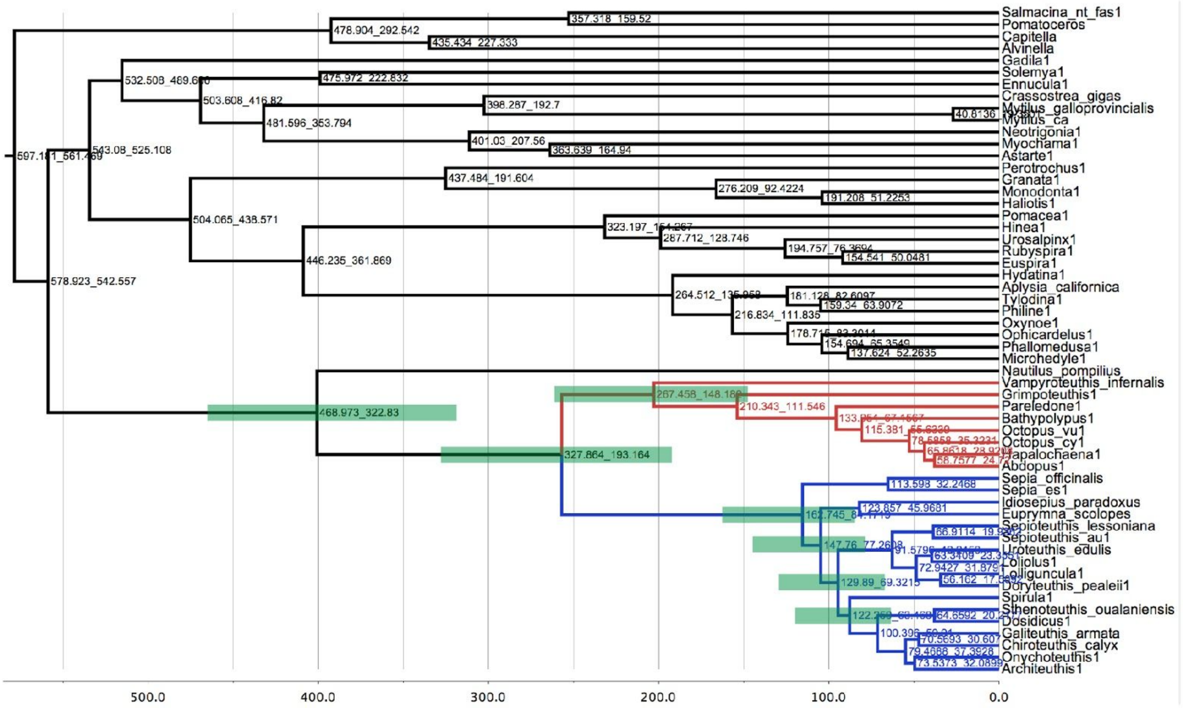


Figure 2.7: UGAM clock. 36,156 amino acids, 52 taxa, CAT + GTR substitution model, UGAM clock model, 6 calibrations with none internal to Coleoidae, with annelid outgroup, root prior of 565 Ma ± 10 Myr, soft bounds of 0.05, Yule birth-death process.

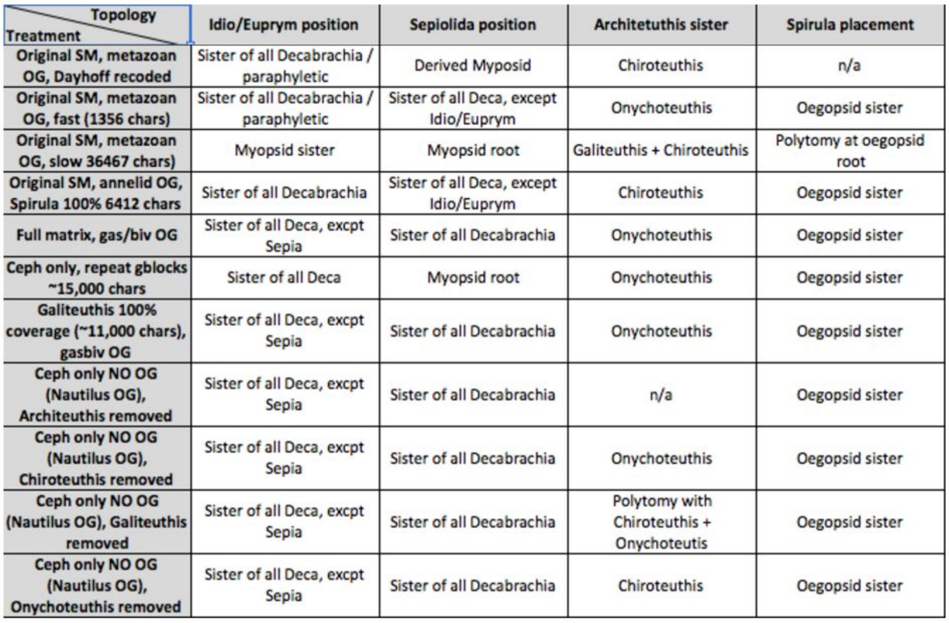
Palaeobiology Database queries

To generate diversity curves for marine vertebrates and belemnites in figure 1, occurrence data was retrieved from PBDB (pbdb.org). Queries can be repeated using the queries below, under default search options.

Query	Occurrences
Belemnitida	1565
Actinopterygii	6973
Chondrichthyes	10576
Placodermi	74
Galeaspida	26
Osteostraci	16
Thelodonti	78
Anaspida	10

Table 2.4: PBDB queries and number of records returned. Records guide diversity curves, as shown in Figure 2.11.

Alternative phylogenetic methods and key node placements.



Topology Treatment	Idio/Euprym position	Sepiolida position	Architetuthis sister	Spirula placement
Original SM, metazoan OG, Dayhoff recoded	Sister of all Decabrachia / paraphyletic	Derived Myposid	Chiroteuthis	n/a
Original SM, metazoan OG, fast (1356 chars)	Sister of all Decabrachia / paraphyletic	Sister of all Deca, except Idio/Euprym	Onychoteuthis	Oegopsid sister
Original SM, metazoan OG, slow 36467 chars)	Myopsid sister	Myopsid root	Galiteuthis + Chiroteuthis	Polytomy at oegopsid root
Original SM, annelid OG, Spirula 100% 6412 chars	Sister of all Decabrachia	Sister of all Deca, except Idio/Euprym	Chiroteuthis	Oegopsid sister
Full matrix, gas/biv OG	Sister of all Deca, excpt Sepia	Sister of all Decabrachia	Onychoteuthis	Oegopsid sister
Ceph only, repeat gblocks ~15,000 chars	Sister of all Deca	Myopsid root	Onychoteuthis	Oegopsid sister
Galiteuthis 100% coverage (~11,000 chars), gasbiv OG	Sister of all Deca, excpt Sepia	Sister of all Decabrachia	Onychoteuthis	Oegopsid sister
Ceph only NO OG (Nautilus OG), Architeuthis removed	Sister of all Deca, excpt Sepia	Sister of all Decabrachia	n/a	Oegopsid sister
Ceph only NO OG (Nautilus OG), Chiroteuthis removed	Sister of all Deca, excpt Sepia	Sister of all Decabrachia	Onychoteuthis	Oegopsid sister
Ceph only NO OG (Nautilus OG), Galiteuthis removed	Sister of all Deca, excpt Sepia	Sister of all Decabrachia	Polytomy with Chiroteuthis + Onychoteutis	Oegopsid sister
Ceph only NO OG (Nautilus OG), Onychoteuthis removed	Sister of all Deca, excpt Sepia	Sister of all Decabrachia	Chiroteuthis	Oegopsid sister

Table 2.5: alternative topologies and treatments, with key node topology inference. OG = outgroup. SM = supermatrix. Gasbiv = gastropods and bivalves. *Metazoan outgroup* refers to including outgroups to sponges at the base of metazoa.

2.3 Results

Our phylogenetic results confirm *Nautilus* as a sister group to a monophyletic coleoid cephalopods (Kocot et al., 2011; Kröger et al., 2011). In turn, coleoids comprise two monophyletic groups: Octobrachia (Vampire squids, dumbo octopuses and incirrate octopuses) and Decabrachia (cuttlefish and squid, including *Spirula*), in agreement with morphology and previous molecular studies (Lindgren, 2010; Lindgren et al., 2012; Strugnell et al., 2006) (Figure 2.8). The vampire squid *Vampyroteuthis* and the cirroctopod *Grimpoteuthis* represent cirrate octopuses, branching deep as successive sister groups to the incirrate octopuses (Figure 2.8). Within Decabrachia, we recovered a monophyletic Myopsida assemblage, along with support for Teuthoidea with the inclusion of *Spirula*, similar to previous studies (Kröger et al., 2011; Lindgren et al., 2012). However, the relationships between the orders comprising the Sepioidea (Sepiida, Idiosepiidae, Sepiolidae) are recovered as paraphyletic. Oegopsid monophyly is recovered, with *Spirula* sister to this clade, in agreement with previous studies (Lindgren et al., 2012), but the posterior probability values for many decabrachian basal nodes are generally lower than in other parts of the phylogeny. Sepioid and myopsid relationships have proven difficult to resolve (Lindgren et al., 2012), and further phylogenetic work remains to clarify these.

Molecular divergence times were estimated, from the same matrix used for phylogenetic inference, applying an autocorrelated relaxed clock model (CIR process). Analyses were first performed placing calibrations only outside of crown cephalopods, and secondly with both these external calibrations, plus calibrations of unequivocal crown cephalopods and coleoids affinity (Figures 2.9, 2.3, and Table 2.3; also see Figures 2.10 and 2.11). Alternative treatments, model applications, and comparison of the joint priors induced by our calibrations and models and the posterior divergence times supported the data as informative, and resulted in consistency in divergence time inference (Table 2.3 and figure 2.3). Notably, our molecular divergence times are congruent with previous molecular divergence estimates (Bergmann et al., 2006; Kröger et al., 2011) that used comparable calibration schemes. These studies, however, had insufficient taxonomic spread and sample required for more comprehensive investigation of the evolutionary tempo of coleoids. Furthermore, our wide sample represents crown diversity.

The fossil record of crown coleoids is scarce but still provides some points for comparison with our divergence times (Figure 2.10, Figure 2.11). The oldest unequivocal crown group coleoids appear in the latest Triassic, with belemnites representing stem group decabrachians, and phragmoteuthidids (Early Triassic or latest Permian) proposed to represent stem group Octobrachia (Fuchs, Keupp, et al., 2013). Our divergence times suggest that the coleoid crown diverged in the late Carboniferous or Permian. Further consilience is shown by stem group vampire squid (loligosepiids) fossils of the earliest Jurassic (~195 Ma) (Fuchs et al., 2015; Fuchs and Weis, 2008), consistent with our inferred origin for the group in the early Triassic. Octopus-like forms that are lacking the mantle fins and with reduced gladius appear in the latest Cretaceous (Cenomanian, 94-100 Ma) Lagerstätte of Hâkel and Hâdjoula, Lebanon (Fuchs et al., 2009). Our divergence estimate for the incirrate octopods is in the Late Cretaceous (~100 Ma). Decabrachians have a near non-existent fossil record, except for members of their stem group (e.g. belemnites) and some forms that retain remnants of the phragmocone – *Spirula* and cuttlefish. Stem group spirulids appear in the latest Cretaceous (~66-72 Ma) of West Greenland (Fuchs et al., 2012). Molecular estimates here suggest that spirulids diverged from the Oegopsids at ~128 Ma. Sepiid cuttlebones appear in the fossil record in the latest Cretaceous (~75 Ma (Fuchs et al., 2009)) and we estimate the sepiids represented in our analysis to have diverged ~88 Ma ago.

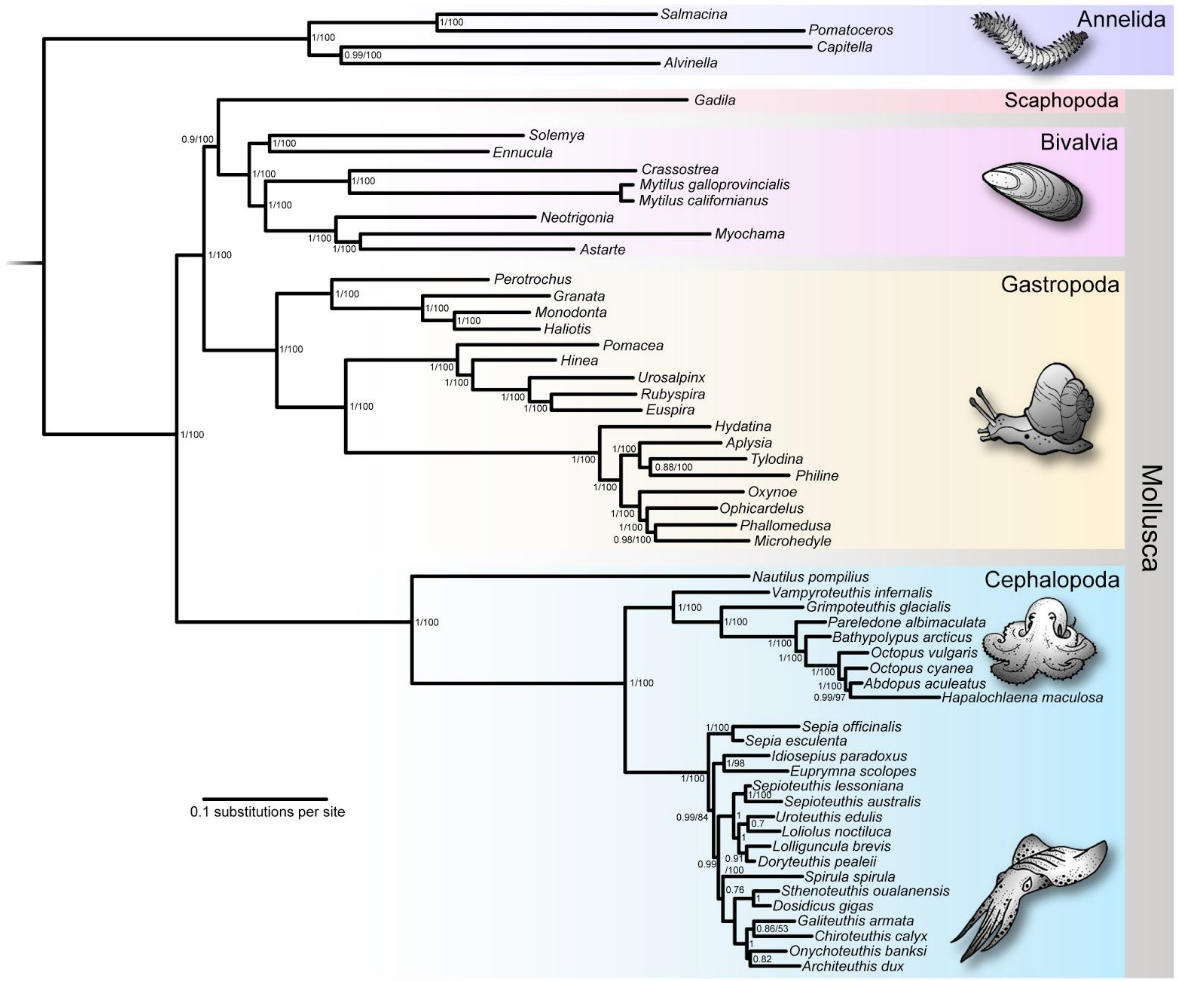


Figure 2.8: Molecular phylogeny of cephalopod, gastropod and bivalve molluscs (plus a scaphopod), with annelid outgroup. 180 genes, concatenated as 36,156 aligned amino acid positions with 26% missing data, modelled under CAT + GTR + G. Numbers at nodes denote Bayesian posterior probability / bootstrap support as returned by RAxML under the LG (Le and Gascuel, 2008) substitution model. Scale bar is expected substitutions per site.

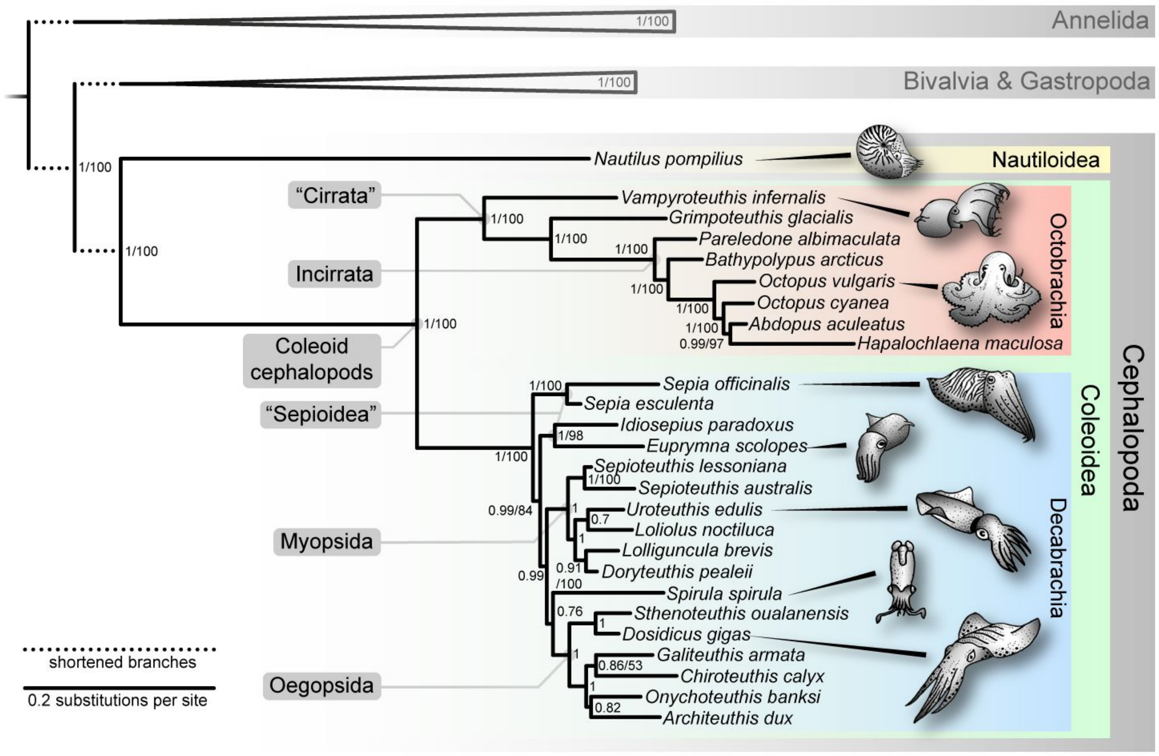


Figure 2.9: Phylogeny of 26 cephalopod species, plus outgroups (further details in Figure 2.8). 180 genes, concatenated as 36,156 aligned amino acid positions with 26% missing data, modelled under CAT + GTR + G. Numbers at nodes denote Bayesian posterior probability / bootstrap support as returned by RAxML under the LG (Le and Gascuel, 2008) substitution mode. Dotted branches at base of phylogeny are shortened for clarity, and outgroups (26 gastropods and bivalves, one scaphopod, four annelids) are collapsed for clarity, see figure 1. Scale bar is expected substitutions per site.

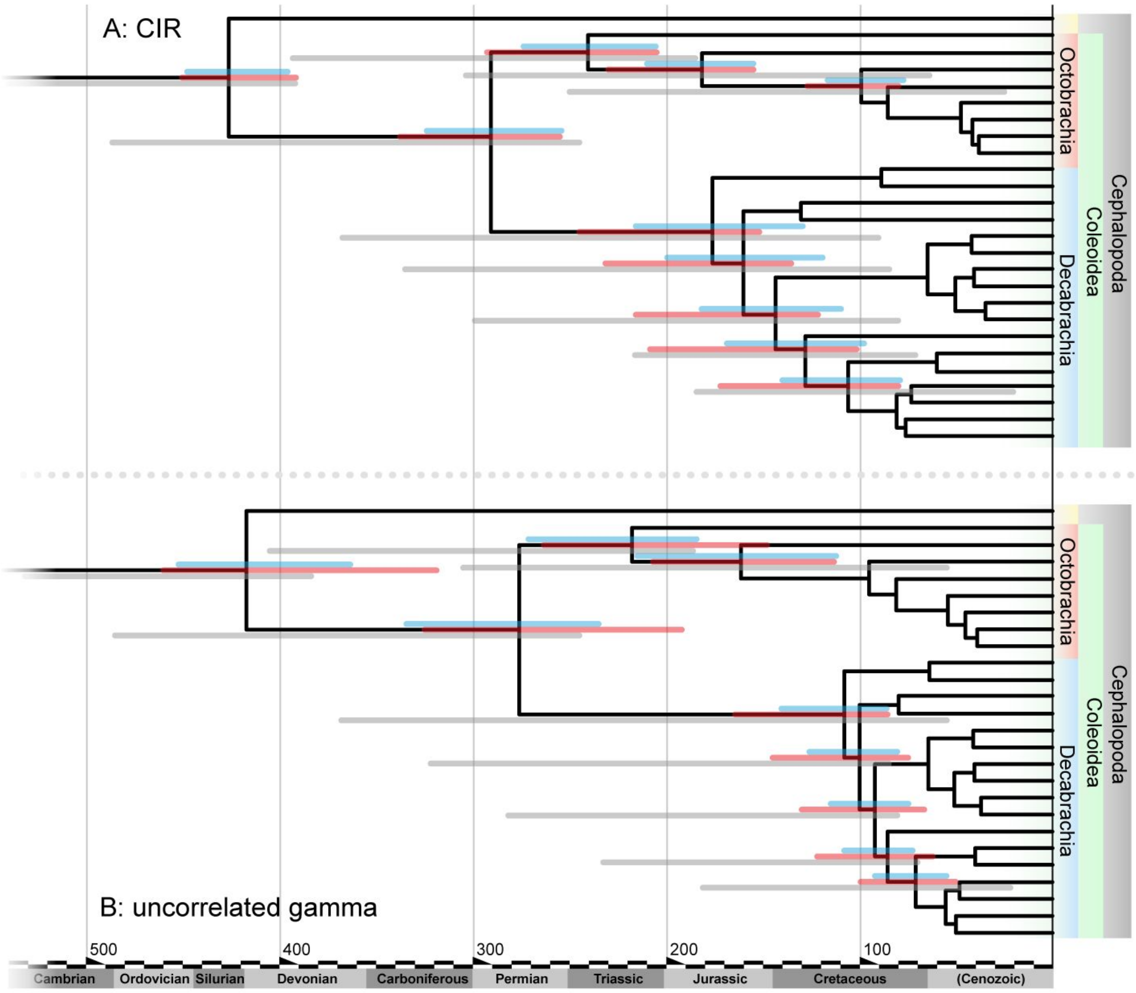


Figure 2.10: Comparison of molecular clock model and calibration scheme on confidence intervals for node timing inference. (a) applying CIR clock model, (b) applying uncorrelated gamma multipier model. Red bars at nodes are confidence intervals with only calibrations external to cephalopods applied. Blue bars are confidence intervals with the full calibration applied. Grey bars are the joint prior distribution at nodes. Not all nodes are labelled to aid clarity, for further details see figures in result section 2.3.

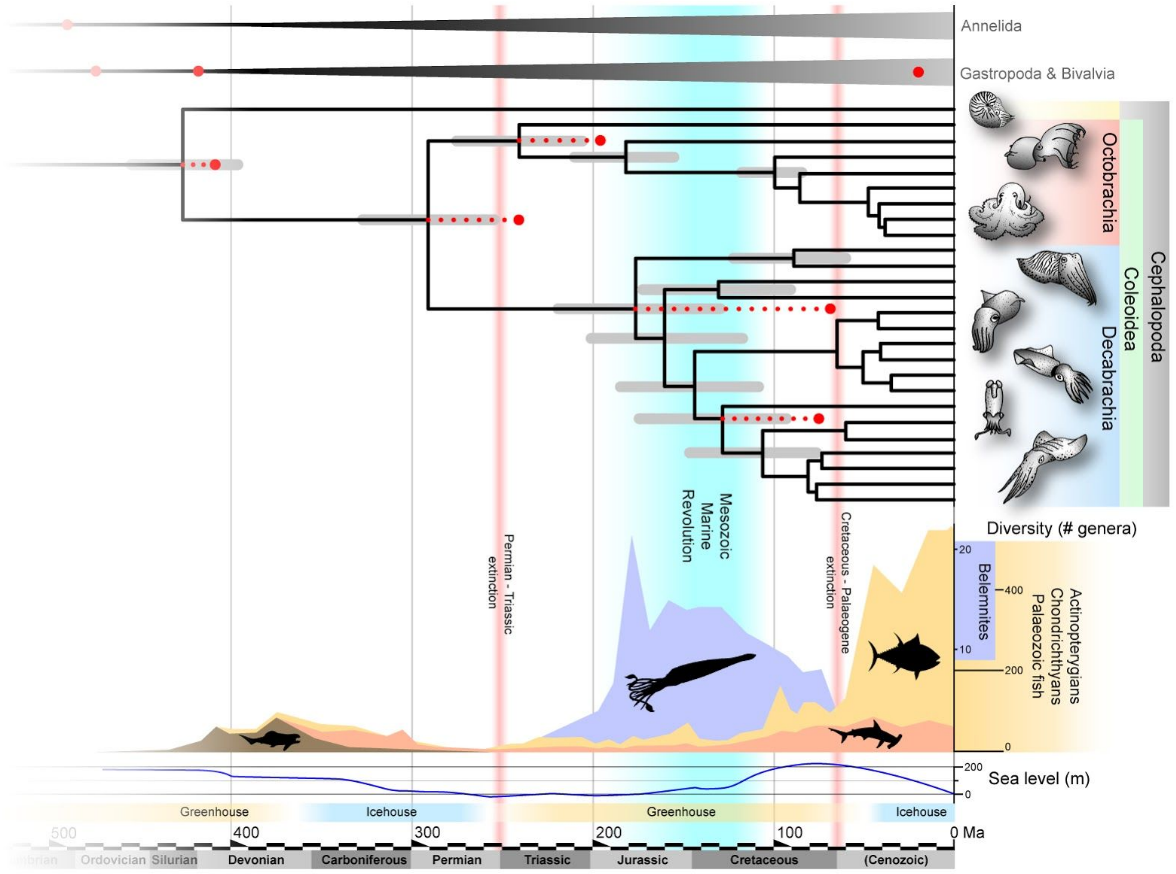


Figure 2.11: Chronogram of cephalopods, plus 26 bivalve and gastropod molluscs, one scaphopod and four annelids as outgroups and calibration nodes. 36,156 amino acid positions analysed under CAT-GTR substitution model, CIR clock model, Yule birth-death process, soft bound of 0.05, and a root prior of 565 Ma with a standard deviation of ± 10 Myr. Bars at nodes represent 95% confidence intervals (recent nodes not labelled with bars to aid clarity). Red dots indicated calibrated nodes; red dotted lines represent extent of calibration minima. Environmental conditions and sea level curve simplified from Miller *et al.* (2005) (Miller et al., 2005). Curves for belemnite, actinopterygian, chondrichthyan and Palaeozoic fish diversity are based on fossil observations on diversity, data from Palaeobiology Database (pbdb.org), see table 2.4 for query details. Red vertical lines represent major extinction events. Aqua-blue vertical bar signifies the extent of the Mesozoic Marine Revolution (Vermeij, 1977).

2.4 Discussion

Our molecular divergence estimates show that the coleoid fossil record (Fuchs et al., 2015; Schweigert and Fuchs, 2012), while sparse, ostensibly documents the appearance and radiation of key cephalopod groups. Together with the molecular clock estimates for coleoids that are lacking a fossil record, it is possible to investigate events that shaped the diversity of the group. Both decabrachians and the incirrate octopods - which together represent the bulk of modern coleoid diversity - radiate in the middle Mesozoic (Jurassic). Since this time documents an escalation - the evolution of novel predation strategies - it prompts a consideration of what anatomical changes took place in coleoids, particularly decabrachians, at this time.

In both the decabrachian and octobrachian lineages, internalisation and reduction of the skeleton evolved independently (Fuchs et al., 2015; Lindgren et al., 2012), leading to enhanced maneuverability and speed. These groups would have been in ecological competition with belemnites; stem group decabrachians (Iba et al., 2012; Schweigert and Fuchs, 2012) with an elaborate internal shell, diversifying in the Mid-Jurassic (Dera et al., 2016). Our analysis suggests that in the late Jurassic and at the onset of the Cretaceous, belemnites became marginalised and replaced by modern groups of decabrachians and finned octobrachians (Figure 2.8) (Fuchs et al., 2015). Similar patterns have been inferred from the Pacific fossil record in Japan (Iba et al., 2011), suggesting a dramatic turnover in particular ~100 Ma.

Decabrachian coleoids are nektonic predators with streamlined morphology, high metabolic rates and shoaling behaviour; adaptations in common with teleost fishes (Packard, 1972). The majority of modern teleost groups radiated during the Jurassic and Cretaceous (Near et al., 2012), concomitantly with the origin of most modern coleoids as revealed by our molecular estimates and the fossil record. The scenario in which Mesozoic ecological shifts are exhibited in teleost fish, chondrichthyans (sharks and rays), and shelled invertebrates as investigated by Vermeij (Vermeij, 1977) can be extended to cephalopods. In the face of high-metabolism, robust predators, and niche-competitors, the cephalopods may have responded in kind to these evolutionary pressures. We hypothesise that the cephalopods evolved into the forms we are familiar with today, while shelled groups fell into extinction due to the shifts in predation in this time period. The Mesozoic Marine Revolution can thus be viewed as the final stage in the shift from Palaeozoic ecologies into the modern structure of marine ecosystems, where (at least in the nektonic realm), agility superseded passive defense.

Taken together, molecular divergence times and the cephalopod fossil record are consistent with a scenario in which predator-prey arms races shaped the coleoid body plan, biodiversity and ecology. The coincidence with the evolution of jawed vertebrates and teleost fishes during the Devonian Nekton Revolution and the Mesozoic Marine Revolution, suggests that nektonic marine vertebrates have been key antagonists towards (and competitors with) cephalopods throughout most of their evolution.

Chapter 3

Chelicerata phylogenomics reconcile with morphological systematics and suggest a late-Cambrian terrestrialisation

A version of this chapter is in preparation for publication (December 2017), in collaboration with Jesus Lozano-Fernandez, Jakob Vinther, Gregory D Edgecombe and Davide Pisani. The experiment was devised and designed by ART, JLF, JV and DP. Specimens for sequencing were provided by GDE and JLF. Molecular laboratory sample preparation work was carried out by JLF. Sequencing was carried out by the University of Bristol Life Sciences genetics facility. Dataset curation, scripting, bioinformatics programming and all computational cluster analyses were carried out by ART. Bioinformatics scripting was developed by ART, further details in Appendix A. Interpretation of results was carried out by ART, DP and JLF. The manuscript was authored by ART and JLF, with further input from all authors. All figures were produced by ART.

Abstract

The interrelationships of chelicerate orders is one of the most contentious issues in the phylogeny of arthropod . While there is general consensus in grouping spiders and other book-lunged arachnids, the rest of chelicerate systematics remains highly uncertain, both from morphological and molecular perspectives. Phylogenomic inference is hampered by weak and conflicting signal arising from rapid diversification (and thus short internodes) for crucial early divergence events. Here we assemble a transcriptomic dataset with 95 species, 3 which are newly sequenced, and test alternative approaches to data refinement. When applying the model of best fit to the dataset with the least substitution saturation, the monophyly of Arachnida is recovered, as is a clade uniting Parasitiformes and Acariformes mites (Acari), and Arachnopulmonata, the latter being consistent with recent phylogenomic analyses and morphological cladistics. The phylogenetic placement of horseshoe crabs as sister group to the rest of terrestrial arachnids also reconciles with morphological and palaeontological studies, upholding our bias mitigation protocols as effective. As such, the work represents reconciliation between morphological and molecular inference of chelicerate systematics, and presents a new hypothesis for chelicerate origins and early diversification. We also infer molecular divergence time estimation to investigate evolutionary dynamics at the origin of Chelicerata, and show that a late-Cambrian terrestrialisation.

3.1 Introduction

Chelicerata is the second largest subphylum of arthropods, outnumbered in species diversity only by the insects. Chelicerata exhibits great diversity, both in terrestrial ecosystems, as represented by web-building spiders, parasite species such as haemophagic ticks, mites (some of which are eusocial), amblypygids (whip scorpions), opiliones (harvestmen) or ricinuleids (hooded tick-spiders), as well as in the marine realm, with representatives such as horseshoe crabs and the enigmatic pycnogonids (sea spiders). There is a strong consensus for a monophyletic Arthropoda from both morphological and a molecular viewpoint (Budd and Telford, 2009; Edgecombe, 2010), as well as Chelicerata (Lozano-Fernandez et al., 2016; Regier et al., 2010; Rota-Stabelli et al., 2011), which is currently favoured as the sister group to Mandibulata (Myriapoda + Pancrustacea), therefore implying deep animal origins.

Chelicerate evolutionary history spans back at least to the Cambrian, around 524 million years ago (Ma), as can be inferred both from the fossil record and molecular divergence times estimation (Dunlop, 2010; Rota-Stabelli et al., 2011). Most chelicerate lineages are predatory components of a diverse range of ecosystems, and the rock record shows that they have been dominant in both earlier Palaeozoic marine settings (Legg, 2014), and through into the Mesozoic and Cenozoic with a huge diversification of spiders and other terrestrial arachnids (Penney, 2003; Selden et al., 2009).

However, building on this through phylogenetic investigation has proved difficult. Recent molecular studies have not improved the inter-ordinal resolution within Arachnida. Since the advance in molecular biology in the 21st century, inference of phylogeny from molecular sequence information has been a powerful tool to investigate the Tree of Life, with notable advances including clear support for the once-controversial hypothesis of a paraphyletic Crustacea (on account of hexapod insects sharing a common ancestor within Crustacea) (Regier et al., 2005; von Reumont et al., 2012). But molecular phylogenetics have not proved a silver bullet for reconstruction of evolutionary history, and it has become increasingly clear that a variety of biases can mislead analyses (Philippe and Roure, 2011), and sometimes lead to strong support from wrong results (Jeffroy et al., 2006). That we continue to grapple with the phylogeny of chelicerates is a manifestation of the existence of recalcitrant nodes (Sharma et al., 2014).

Two examples of incongruence are the phylogenetic placement of horseshoe crabs (Xiphosura), and of mites and ticks (Acari). Xiphosurans have a good fossil record, showing unusual morphological stasis, and stem representatives of this lineage being traceable to the

Ordovician at around 445 Ma (Rudkin et al., 2008). With such a deep history, any inferred phylogenetic position other than branching very near the root of Chelicerata (and as sister group of terrestrial arachnids) would imply that a marine habit of this lineages is a secondary acquisition. Such a scenario is palaeontologically implausible, yet horseshoe crabs are often recovered, using molecular data, in highly derived clades, such as allied to Opiliones or palpigrades (Pepato and Klimov, 2015), ricinuleids (Sharma et al., 2014), or scorpions and spiders (von Reumont et al., 2012; Roeding et al., 2009; Sanders and Lee, 2010).

Acari has over 50,000 described species, making them the most speciose arachnid order, even outnumbering spiders. Acari are frequently parasitic, being haemophagic for at least part of their life-cycle, and exploiting nearly all metazoans as hosts, from molluscs to fish to mammals (Mans and Neitz, 2004). As such, they are of medical relevance because of their role as disease vectors. The principle synapomorphy of Acari is the presence of the gnathosoma, a specialised feeding apparatus, and this has driven some to uphold the monophyly of Acari (Garwood and Dunlop, 2014). Yet, this well-supported morphological hypothesis has not been reconciled by molecules (von Reumont et al., 2012; Roeding et al., 2009; Sanders and Lee, 2010), limiting the ability to generate robust ancestral state reconstructions to understand the terrestrialisation of the group, and the emergence of ecological innovations, such as a blood-feeding lifestyle.

Attendant to these specific examples is the general phylogenetic instability across the rest of Chelicerata, despite current data-refinement and model-application protocols (Sharma et al., 2014, 2015). While such long-running incongruence might be reason for despair, difficult phylogenetic problems, such as that of Chelicerata, can be reframed as an opportunity to understand why incongruence arises, and promote best-practice in use of data and application of models, especially for such deep-time scenarios afflicted by short internode branch lengths and high levels of sequence saturation. Together, these issues make it unsurprising that inferring phylogeny in deep time remains a difficult problem despite the welcome increases in available genetic information, the improved understanding of statistical approaches in phylogenetic inference (O’Reilly et al., 2016; Puttick et al., 2017), and the computing power necessary to process large datasets.

To confront these issues, we here carry out a phylogenomic investigation to resolve relationships within Chelicerata, relying on both new sequence information and robust inferential methodology. We compile a molecular matrix based on transcriptomic data from 95 species, 3 from newly sequenced organisms. Critically, we include high-quality sequence information from three horseshoe crabs, as well as a wide representation of rapidly evolving lineages, particularly members of Acariformes, Parasitiformes and Pseudoscorpiones. This expanded taxonomic sample reduces potential long branch attraction (LBA), of particular importance seeing as previously investigations (Sharma et al., 2014) suggesting that chelicerate phylogeny is prone to such bias. As such, our datasets and methodological approach represent an improvement with respect previous analyses to deal with conflicting signal and incongruence.

3.2 Methods

Data acquisition

Total RNA was extracted from individuals specimens of *Pycnogonum* sp., *Galeodes* sp. and *Damon* sp. using TriZol©. A transcriptome-wide cDNA library was generated and sequenced using two Illumina HiSeqII lanes at TrinSeq (Trinity College Dublin, Institute of Molecular Medicine, Genome Sequencing Laboratory) to an estimated coverage of <100, using 100-bp paired end reads. Row data were inspected for its quality and assembled using Abyss (Simpson et al., 2009) with k-mer of 45.

Order	Species	Source	Predicted proteins	Accession
Pycnogonida	*Endeis spinosa*	454	2930	LIBEST_025662
Pycnogonida	*Pycnogonum sp.*	Illumina	26,668	Newly sequenced
Pycnogonida	*Anoplodactylus eroticus*	454	2,595	Sharma et al. 2014
Xiphosura	*Limulus polyphemus*	Illumina	17824	SRR1145732
Xiphosura	*Tachypleus tridentatus*	Illumina	47095	SRR946952
Xiphosura	*Carcinoscorpius rotundicauda*	Illuimina	82,789	SRR1511637
Acariformes	*Tetranychus urticae*	Illumina	26106	PRJEB6152
Acariformes	*Dermatophagoides farinae*	Illumina	36507	SRR1016494
Acariformes	*Panonychus citri*	Illumina	24400	SRR341928
Acariformes	*Rhizoglyphus robini*	Illumina	72646	PRJNA213807
Acariformes	*Steganacarus magnus*	Illumina	60104	SRR4039729
Acariformes	*Hypochthonius rufulus*	Illumina	29268	SRR4039020
Acariformes	*Platynothrus peltifer*	Illumina	42123	SRR4039728
Acariformes	*Achipteria coleoptrata*	Illumina	54612	SRR4039018
Acariformes	*Hermannia gibba*	Illumina	64632	SRR4039019
Acariformes	*Nothrus palustris*	Illumina	57025	SRR4039021
Acariformes	*Tetranychus Cinnabarinus*	Illumina	23425	SRR519097

Parasitiformes	*Amblyomma americanum*	Illumina	32598	PRJNA238773
Parasitiformes	*Dermacentor andersoni*	Illumina	30501	PRJNA238802
Parasitiformes	*Ixodes scapularis*	Illumina	56503	SRR1189647
Parasitiformes	*Rhipicephalus microplus*	Illumina	13004	SRR1186998
Parasitiformes	*Hyalomma excavatum*	Illumina	54671	PRJNA311286
Parasitiformes	*Ornithodoros rostratus*	Illumina	37109	PRJNA270484
Parasitiformes	*Dermanyssus gallinae*	454	39197	SRR658515
Parasitiformes	*Varroa destructor*	Illumina	16085	SRR3927486
Parasitiformes	*Varroa jacobini*	Illumina	12342	SRR3634772
Parasitiformes	*Phytoseiulusy persimilis*	454	21218	DRR001717
Ricinulei	*Pseudocellus sp.*	Illumina	5,922	SRR1146686
Ricinulei	*Ricinoides atewa*	Illumina	14,324	SRR1145743
Ricinulei	*Cryptocellus sp. n.*	Illumina	49645	SRR1982218
Ricinulei	*Ricinoides karschii*	Illumina	87143	SRR1972991
Ricinulei	*Cryptocellus becki*	Illumina	128981	SRR1979416
Solifugae	*Eremobates sp.*	Illumina	11,765	SRR1146672
Solifugae	*Gluvia dorsalis*	454	5404	SRR1141096
Solifugae	*Galeodes* sp	Illumina	422228	Newly sequenced
Pseudoscorpiones	*Synsphyronus apimelus*	Illumina	17,820	SRR1145733
Pseudoscorpiones	*Haplochernes kraepelini*	Illumina	16376	SRR1767661
Pseudoscorpiones	*Hesperochernes sp.*	Illumina	30903	SRR1514877
Pseudoscorpiones	*Neobisium carcinoides*	Illumina	24142	Newly sequenced
Scorpiones	*Bothriurus burmeisteri*	Illumina	20574	SRR1721670
Scorpiones	*Chaerilus celebensis*	Illumina	24310	SRR1721804
Scorpiones	*Centruroides sculpturatus*	Illumina	16440	SRR1515193
Scorpiones	*Hadrurus arizonensis*	Illumina	19266	SRR1721733
Scorpiones	*Iurus dekanum*	Illumina	17619	SRR1721734
Scorpiones	*Liocheles australasiae*	Illumina	22581	SRR1721664
Scorpiones	*Superstitionia donensis*	Illumina	23916	SRR1721951
Scorpiones	*Vietbocap lao*	Illumina	20007	SRR1721740
Scorpiones	*Androctonus australis*	Illumina	19170	SRR1724216
Scorpiones	*Pandinus imperator*	Illumina	20279	SRR1721600
Scorpiones	*Scorpiops sp.*	Illumina	24941	SRR1721951
Uropygi	*Mastigoproctus giganteus*	Illumina	17674	SRR1145698
Amblypygi	*Damon variegatus*	Illumina	11823	SRR1145694
Amblypygi	*Euphrynichus bacillifer*	454	3895	SRR1141095
Amblypygi	*Damon sp.*	Illumina	24,564	Newly sequenced

Opiliones	*Larifuga capensis*	Illumina	10506	SRR1145742
Opiliones	*Metasiro americanus*	Illumina	16556	SRR618563
Opiliones	*Pachylicus acutus*	Illumina	14202	SRR1146670
Opiliones	*Protolophus singularis*	Illumina	13987	SRR1145700
Opiliones	*Trogulus martensi*	Illumina	12765	SRR1145730
Opiliones	*Hesperonemastoma modestum*	Illumina	8845	SRR1145728
Opiliones	*Siro boyerae*	Illumina	11387	SRR1145699
Araneae	*Liphistius malayanus*	Illumina	11221	SRR1145736
Araneae	*Neoscona arabesca*	Illumina	16594	SRR1145741
Araneae	*Dysdera crocota*	Illumina	30336	SRR1328258
Araneae	*Pholcus manueli*	Illumina	27519	SRR1365208
Araneae	*Oecobius cellariorum*	Illumina	30394	SRR1365089
Araneae	*Uloborus glomosus*	Illumina	15941	SRR1328334
Araneae	*Amaurobius ferox*	Illumina	19707	SRR1329250
Araneae	*Pisaurina mira*	Illumina	15940	SRR1365651
Araneae	*Microdipoena guttata*	Illumina	17704	SRR1333842
Araneae	*Sphodros rufipes*	Illumina	23297	SRR1514908
Araneae	*Antrodiaetus unicolor*	Illumina	20709	SRR1514897
Araneae	*Megahexura fulva*	Illumina	30559	SRR1514891
Araneae	*Microhexura montivaga*	Illumina	17718	SRR1514890
Araneae	*Brachythele longitarsus*	Illumina	19334	SRR1514875
Araneae	*Pionothele n. sp.*	Illumina	10538	SRR1514906
Araneae	*Paratropis sp.*	Illumina	9961	SRR1514893
Araneae	*Aphonopelma iviei*	Illumina	11968	SRR1514871
Araneae	*Hypochilus pococki*	Illumina	29268	SRR1514889
Araneae	*Ero leonina*	Illumina	33289	SRR1514886
Onychophora	*Peripatopsis capensis*	Illumina	12846	SRX451023
Onychophora	*Euperipatoides kanangrensis*	EST	3267	Dunn et al. 2008
Onychopohra	*Epiperipatus sp.*	EST	514	Roeding,F.. 2007
Crustacea	*Argulus siamensis*	Illumina	48641	SRR514120
Crustacea	*Gammarus chevreuxi*	Illumina	28997	GFCV00000000.1
Crustacea	*Calanus finmarchicus*	Illumina	126873	SRR1153469
Crustacea	*Daphnia pulex*	WGS	30611	PRJNA12756
Insecta	*Orchesella cincta*	WGS	20257	PRJNA294050
Insecta	*Drosophila melanogaster*	WGS	30443	PRJNA13812
Insecta	*Blattella germanica*	454	7302	PRJNA248247
Insecta	*Tribolium castaneum*	WGS	18534	PRJNA12540

Chilopoda	*Alipes grandidieri*	Illumina	18814	SRR619311
Chilopoda	*Scutigera coleoptrata*	Illumina	31758	SRR1158078
Symphyla	*Hanseniella sp.*	Illumina	50853	SRR3458649
Diplopoda	*Glomeris marginata*	Illumina	66936	SRR3458641

Table 3.1. Species used in this study, plus details of sequence data sources and accessions.

Orthology assignment and matrix assembly

We compiled a supermatrix with data from 95 species (see table 4.1) for 234 genes, based on and building on the gene sampling of Philippe et al. (Philippe et al., 2011). The taxonomic sample comprised 80 chelicerates, 74 of them being arachnids with 21 mites included (and 15 outgroup species) suitably covering the groups of interest, and also being the largest published chelicerate phylogenetic matrix assembled to date. New chelicerate sequences matching those in the Philippe et al. (2011) gene-sample were acquired through BLAST (Altschul et al., 1990) searching transcriptomic sequences, using *Daphnia magna* as the search query due to this taxon possessing full coverage for the gene dataset, and having phylogenetic proximity to the group in question. A custom Perl script (available at github.com/jairly/MoSuMa_tools/) isolated sequences on the most significant expect values (e-values) among BLAST hits, taking the lowest e-values and any other significant hits within three orders of magnitude of the most significant hit (in order to provide possible alternative orthologs). The maximum e-value threshold was set at 10^{-10}: sequences with e-values larger than this were rejected as candidate orthologues. These selected sequences were aligned using MUSCLE (Edgar, 2004) (applying default parameters), to produce gene-alignments for each of the 234 genes. Ambiguously aligned positions were removed from the gene alignments by GBlocks v0.91b (Castresana, 2000) (using the parameters *b2* = 70%, *b3* = 10, *b4* = 5, *b5* = half).

Gene trees were inferred for each individual gene alignment using IQTree (Nguyen et al., 2015b), applying the model of best fit as assigned by ModelFinder (Kalyaanamoorthy et al., 2017) (as part of the IQTree package). For nearly all gene trees, the model LG + I + G was best fit. The 234 gene trees were assessed for long branches using a custom Perl script (/github.com/jairly/MoSuMa_tools/blob/master/treecleaner.pl), with 8% of sequences being identified as anomalously long-branched. Sequences producing these long branches were removed from each gene matrix in order to minimise long branch bias in the supermatrix. The gene alignments, thus cleaned of ambiguous positions and long-branching sequences, were concatenated using SequenceMatrix v100 (Vaidya et al., 2011), with a resulting supermatrix of 45,939 amino acid positions across 95 taxa.

We generated an alternative molecular data set based on the same taxa we already used, with the aim of testing independently the evolutionary relationships of Chelicerata. For this purpose, we used the OMA software (Altenhoff et al. 2013), that can infer pairwise orthologs based on the whole set of transcriptomes/genomes. The default options of OMA yielded a 3,982 groups of orthologous genes, which were subsequently reduced with the most stringent option of Gblocks.

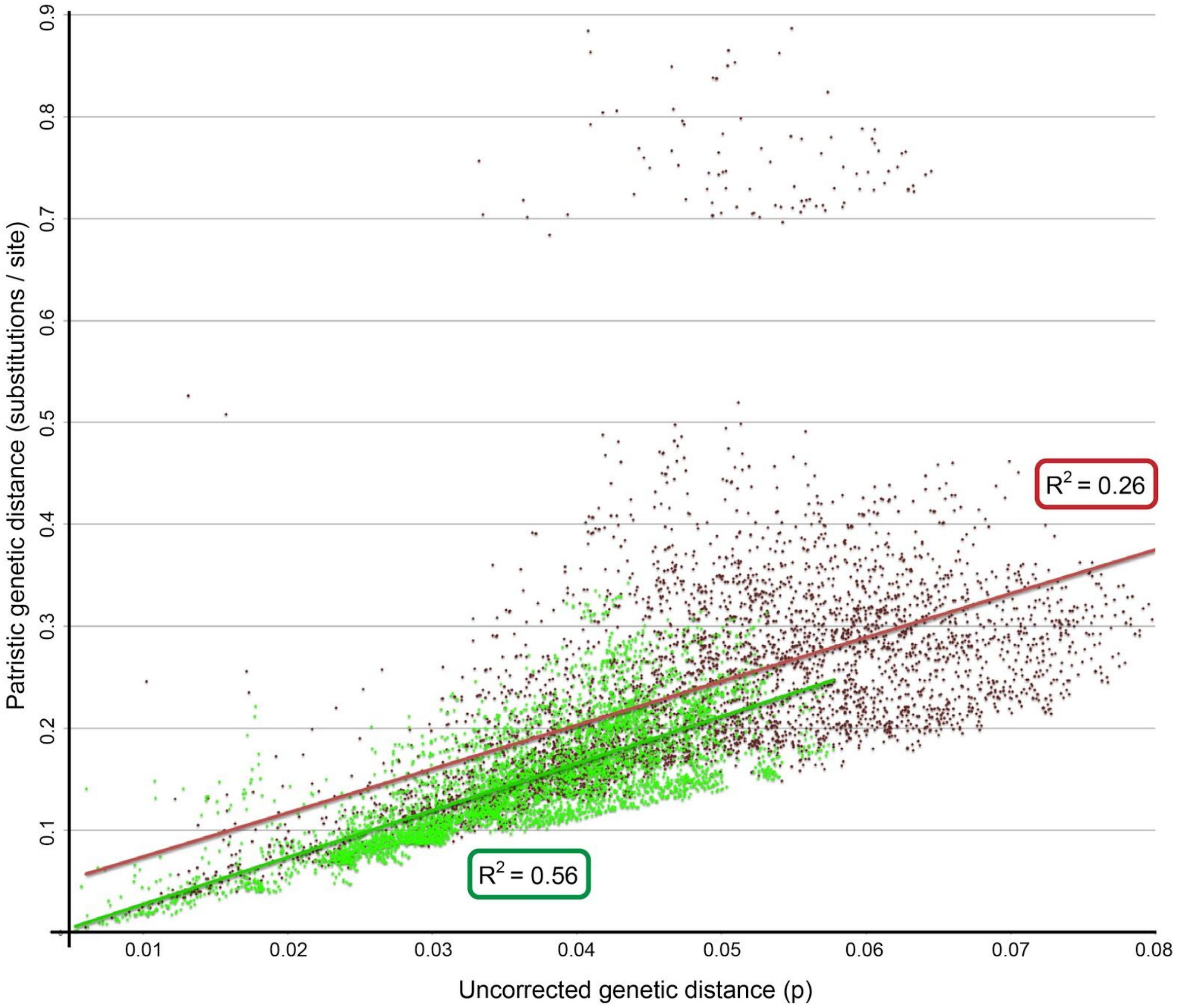


Figure 3.1. Comparison on saturation levels between primary and alternative dataset. Patristic plot of observed substitutions (y-axis) against expected substitutions (x-axis). Green points and regression line for primary dataset, red for alternative. R^2 value for the primary dataset is higher than that for the alternate, indicating that the primary dataset is exhibits less saturation, and as such more appropriate for phylogenetic inference.

Phylogenetic inference

Topological phylogenetic inferences were made using the Markov Chain Monte Carlo (MCMC) sampler PhyloBayes MPI v1.5a (Lartillot et al., 2013). The mixture model CAT + GTR + Γ_4 was applied, being the most appropriate to deal with across-site heterogeneities, while minimising long-branch biases. Cross-validation was carried out to assess model suitability, as described in PhyloBayes documentation, comparing GTR (plus a gamma distribution of rates) against CAT-GTR. CAT-GTR outperformed GTR, with a score of 21.11 +/- 8.3 (positive scores meaning better than reference model (GTR)). Out of ten cross-validation replicates, CAT-GTR outperformed GTR all ten times.

Two independent MCMC chains were run, with a burn-in of 25% of the chains discarded. These chains converged, with the maximum difference (maxdiff) in the bipartitions of the chains < 0.25, as reported by *bpcomp* program in the PhyloBayes package. A further test of convergence was carried out using *tracecomp* (also under PhyloBayes), with effective sample sizes being > 50, and relative differences dropping below 0.3 for all parameters as compared between the independent chains. For the primary experimental chains, maxdiff statistics converged to below 0.08, suggests good mixing of chain parameters and strong Bayesian convergence.

For maximum likelihood inference, the software package IQTree (Nguyen et al., 2015c) was used. As part of the software, ModelFinder (Kalyaanamoorthy et al., 2017) was used to test fit of models, through both Bayesian Information Criterion and Akaike Information Criterion. For all datasets, LG + I + G (Le and Gascuel, 2008) was shown to have best fit by both information criteria. To return node support statistics, 1000 bootstrap replicates were generated and analysed, utilising the ultrafast inference method of the IQTree package (Minh et al., 2013).

Molecular divergence time estimation

The topology inferred from Bayesian process on the primary dataset (Figure 3.2) was used to guide divergence time estimation for Chelicerata. The analysis was carried out under PhyloBayes 4.1f. We model tested through cross-validation, testing CIR (Lepage et al., 2006, 2007) against UGAM (Drummond et al., 2006). CIR applies a modified log-normal distribution which is correlated and informed by the previous branch on the phylogeny. These models represent autocorrelated and uncorrelated rates on lineages, respectively. For CIR, the previous branch’s rate of substitution can inform the estimation of subsequent branches. For UGAM, an uncorrelated gamma distribution is applied, allowing rates to shift more widely along branches. Cross-validation supported CIR as the better fitting model, 32.00 ± 9.44; a positive number is support for the fit of CIR. Out of ten replicates, all ten supported CIR. As such, CIR was applied, with CAT + GTR as the substitutions model, a Yule birth-death process, soft-bounds of 0.05 on node time estimations, and a root prior of 555 Ma ± 15 Myr, representing *Kimberella* as the earliest uncontroversial bilaterian (Martin et al., 2000). For details of fossil calibrations, see Table 3.1. Convergence was considered achieved with *tracecomp* statistics dropping below 1 for all relative difference scores, and all effective sample sizes being above 50, for all chain parameters.

Node	max	min	Reference taxon	Reference publication	Clade
Euperipatoides - Epiperipatus	-	142	Break-up of Pangea supercontinent	(Wolfe et al., 2016)	Peripatidae and Peripatopsidae
Hansiniella - Glomeris	541	426.9	*Casiogrammus ichthyeros*	(Wolfe et al., 2016)	Diplopoda
Scutigera - Alipes	541	416	*Crussolum sp.*	(Wolfe et al., 2016)	Chilopoda
Daphnia - Drosophila	559	514	*Yicaris dianensis*	(Wolfe et al., 2016)	Altocrustacea
Orchesella - Drosophila	-	405	*Rhyniella praecursor*	(Wolfe et al., 2016)	Hexapoda
Tribolium - Drosophila	-	313.7	*Westphalomerope maryvonnae*	(Wolfe et al., 2016)	Aparoglossata
Drosophila - Dysdera	559	514	*Yicaris dianensis*	(Wolfe et al., 2016)	Euarthropoda
Pycnogonum - Dysdera	559	500.5	Chasmataspis-like resting traces	(Wolfe et al., 2016)	Euchelicerata
Limulus - Tachypleura	-	236	*Tachypleus gadeai*	(Wolfe et al., 2016)	Xiphosurida
Scorpiops - Dysdera	541	432.6	*Palaeophonus loudonensis*	(Wolfe et al., 2016)	Arachnopulmonata
Pandinus - Scorpiops	-	112.6	*Protoischnurus axelrodurum*	(de Carvalho and Lourenço, 2001)	Iurida
Vietbocap - Centuroides	-	48.54	*Uintascorpio halandrasorum*	(Perry, 1995)	Buthidae + Pseudochactidae
Mastigoproctus - Amblypygi	-	319.9	*Parageralinura naufraga*	(Wolfe et al., 2016)	Pedipalpi
Tetranychus - Varroa	541	405	*Protocarus crani*	(Hirst, 1923)	Acari
Ornithodoros - Ixodes	-	98.17	*Cornupalpatum burmanicum*	(Poinar and Brown, 2003)	Ixodida
Hypochthonius - Dermatophagoides	-	382.7	*Protochthonius gilboa*	(Norton et al., 1988)	Sarcoptiformes
Siro - Pachylicus	-	405	*Eophalangium sheari*	(Wolfe et al., 2016)	Opiliones
Siro - Metasiro	-	98.17	*Palaeosiro burmanicum*	(Poinar, 2008)	Cyphopthalmi
Trogulus - Hesperonemastoma	-	298.75	*Ameticos scolos*	(Garwood and Dunlop, 2011)	Dyspnoi
Pachylicus - Larifuga	-	98.17	*Petrobunoides sharmai*	(Selden, 2016)	Laniatores
Liphistius - Dysdera	-	298.75	*Palaeothele montceauensis*	(Wolfe et al., 2016)	Araneae
Microhexura - Paratropis	-	240.5	*Rosamygale grauvogeli*	(Selden and Gall, 1992)	Avicularoidea
Aphonopelma - Paratropis	-	125	*Cretamygale chasei*	(Harvey, 2002)	Bipectina
Pholcidade - Dysdera	-	158.1	*Eoplectrurys gertschi*	(Selden and Huang, 2010)	Haplogynae
Oecobius - Neoscona	-	158.1	*Mongolarachne jurassica*	(Selden et al., 2011)	Entelegynae
Neoscona - Ero	-	129.41	*Unnamed Linyphiinae*	(Selden et al., 2011)	Araneoidea

Table 3.2. Fossil calibrations applied to molecular divergence time estimation of Chelicerata, as presented in Figure 3.5. Prior distributions between maxima and minima are uniform distributions, while minima have a CIR distribution (Lepage et al., 2006), in effect a modified log-normal, correlated between branches.

3.3 Results

Matrix and model comparison

We generated two independent molecular data for 95 taxa: the primary set being based on slow-evolving, highly conserved genes that are commonly retrieved from transcriptomic data of a range of read-depths; and an alternative set based on an orthology search across all data, and refined to be approximately the same size as the primary set. In the design of both matrices, we have prioritized minimising matrix gaps and missing data; the primary dataset is 78.1% complete, and the alternate matrix is 86.4% complete. The sets however are built from different data-selection paradigms, and as such have likely undergone different substitution processes. In order to assess these levels of saturation in our two matrices, we assessed sequence saturation through the software Patristic (Fourment and Gibbs, 2006), comparing observed against expected substitutions per site. The primary matrix was shown to suffer lower levels of saturation, as assessed by regression values, R^2 = 0.57 (primary matrix) versus R^2 = 0.26 (alternative matrix) alternative (Figure 3.1). Therefore, our main discussion is based on the results of the primary matrix, but the alternative matrix is presented to illustrate and contribute to discussion on conflict in phylogenetic signal.

Phylogenetic analyses

Our main phylogenetic analysis using the less-saturated molecular dataset and under the best fitting model is presented in Figure 3.2. Schematic cladograms showing the results using Bayesian and ML phylogenetic inference over the primary and alternative matrix are presented in Figure 3.3, and results using Dayhoff recoding are shown in Figure 3.4. Our main analysis supports the monophyly of Chelicerata, Euchelicerata and Arachnida, the latter suggesting a single terrestrialization event. In all Bayesian analyses, using both matrices and under Dayhoff recoding, and in two out of four ML phylogenies, results yielded a monophyletic Chelicerata in which Pycnogonida is sister to all other chelicerates (Euchelicerata). In the main analysis, Xiphosura appears as the sister group of Arachnida, providing molecular evidence for the monophyly of this speciose terrestrial group of chelicerates, and reconciling the phylogenetic signal found consistently in morphologically-based phylogenies. This same matrix analysed under Dayhoff recoding yielded a polytomy between Xiphosura and Acari as the earliest divergent euchelicerates, suggesting that both lineages diverged earlier than the rest, but without corroborating a sister group relationship of Xiphosura with Arachnida. The alternative data set analysed under both Bayesian and ML frameworks does not recover an alliance between the terrestrial chelicerates; and although the position of Xiphosura is unstable across analyses, it nests within the arachnids.

In Figures 3.3 and 3.4 we compare phylogenetic inferences from both datasets and different analyses. For some, highly implausible relationships are returned, allowing us confidence in rejecting these as being highly biased by conflicting phylogenetic signal. Examples of such inaccurate inference include the placement of pycnogonids in the outgroup, therefore not in support of monophyly of Chelicerata (Figures 3.3c, 3.4c), and horseshoe crabs as being derived arachnids (Figures 3.3b, c, d and 3.4c, d). These observations uphold the primary matrix and Bayesian approach as most appropriate (Figures 3.2, 3.3a).

Each of the arachnid orders, represented by three to eleven taxa, is recovered as monophyletic with maximum support in all the eight analyses performed. Interestingly, all analyses but one converge in an alliance between the Parasitiformes and Acariformes, the Acari, a result that has been elusive in most previous phylogenomic analyses (Sharma et al., 2014). The support measures in all Bayesian are very high, suggesting that phylogenetic signal is relatively strong (as is also suggested by the strong convergence between independent MCMC chains). All analyses recover Tetrapulmonata (Araneae and Pedipalpi (Uropygi and Amblypygi)) in alliance with Scorpiones (Arachnopulmonata) or to a clade composed by Scorpiones + Pseudoscorpiones. In most instances in which Arachnopulmonata is retrieved, Pseudoscorpiones are found as its sister lineage. Together, these results suggests a close relationship between pseudoscorpions and arachnopulmonates. All Bayesian analyses in both matrices yielded a relationship between Opiliones and Ricinulei, in most instances with maximum support. In contrast, three of four ML analyses support, a relationship between Ricinulei and Solifugae (sun spiders), although always with low support (PP = 55 < 77). These two latter clades, together with Xiphosura, presents the most variable topological positions.

Our improved taxonomic sample includes 10 Parasitiformes and 11 Acariformes species, and analyses converge on a sister-group relationship between the two, providing strong support for monophyly. Acariformes is composed by two major lineages, Mesostigmata and Ixodida. All analyses show a relationship between Phytoseioidea and Dermanyssoidea within Mesostigmata, and Argasidae with Ixodiadae, within Ixodida. Acariformes are composed by Trombidiformes and Sarcoptiformes. Within the latter clade, some alternative positions for an internal branch results in different definitions of Oribatida, either containing the Astigmata, or without it.

Molecular divergence time estimation

In Figure 3.5 we present a timeline for the evolution of chelicerates. The ancestral node for all arthropods, given this taxonomic sample, is recovered in the Pre-Cambrian at 563 Ma ± 5 Myr. Chelicerata are inferred to have evolutionary origins at 555 Ma ± 4 Myr. Rapid diversification then occurs between 510 and 489 Ma, with 10 orders of chelicerates becoming established between these times. By around 473 Ma, all 11 orders of chelicerates included in this analysis (and of 12 in total - palpigrades have not been sampled for this experiment) are established. Further diversification is inferred as being under more gradual tempo of evolution, in particular with Arachnopulmonata and Acari exhibiting expansion after the start of the Mesozoic.

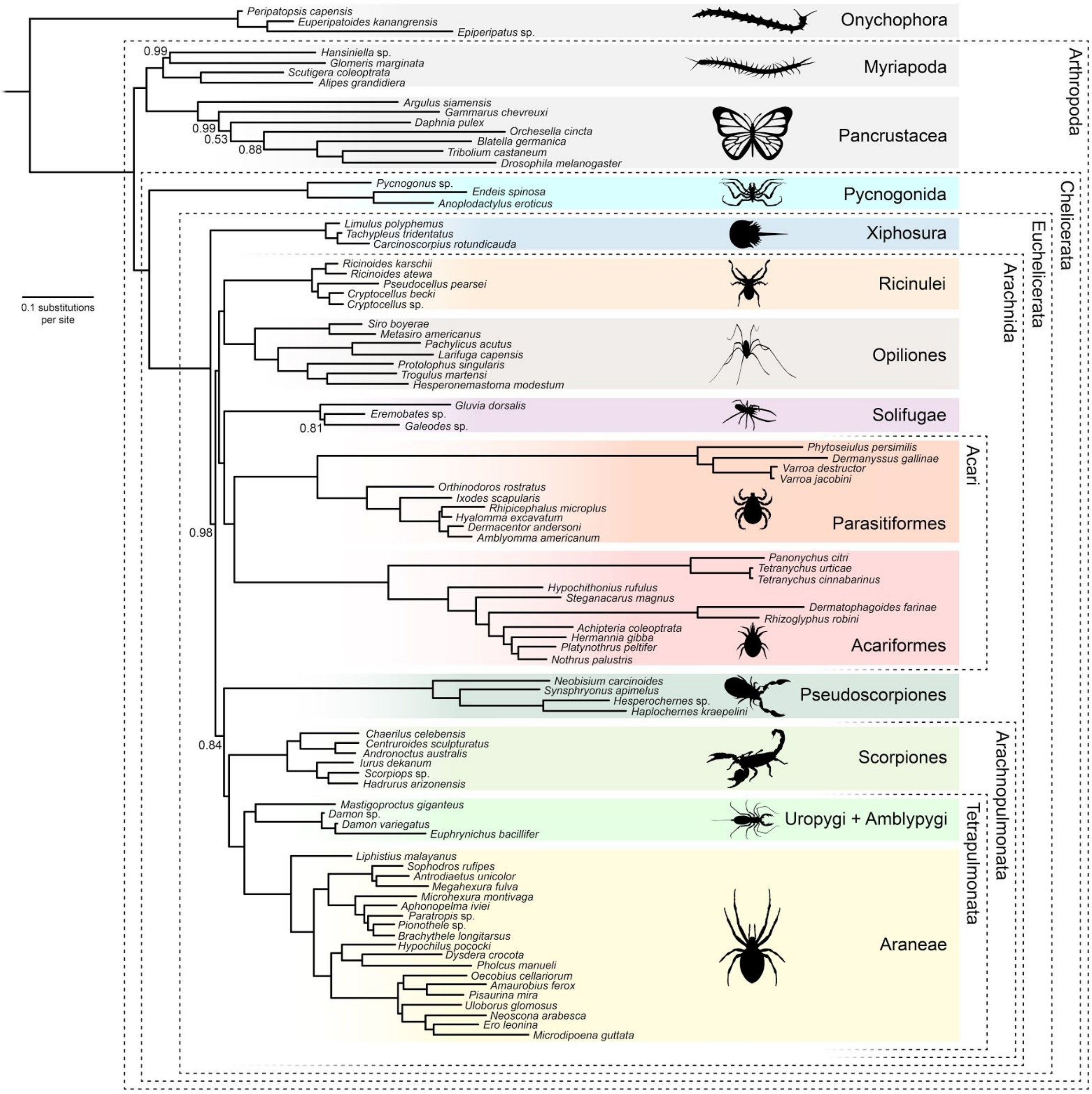


Figure 3.2. Phylogeny of Chelicerata, plus outgroups. Phylogenomic dataset of 45,939 amino across 95 taxa. Inference through PhyloBayes MPI version 1.5a, applying CAT + GTR + Γ substitution model. Numbers at nodes are posterior probabilities (PPs). All unlabelled nodes have PPs of 1. Coloured bars and silhouettes represent chelicerate orders. Dashed boxes represent further levels of monophyletic systematics recovered by the analysis. Colours used are consistent with figure 3.2, 3.3 and 3.4. All silhouettes produced by ART.

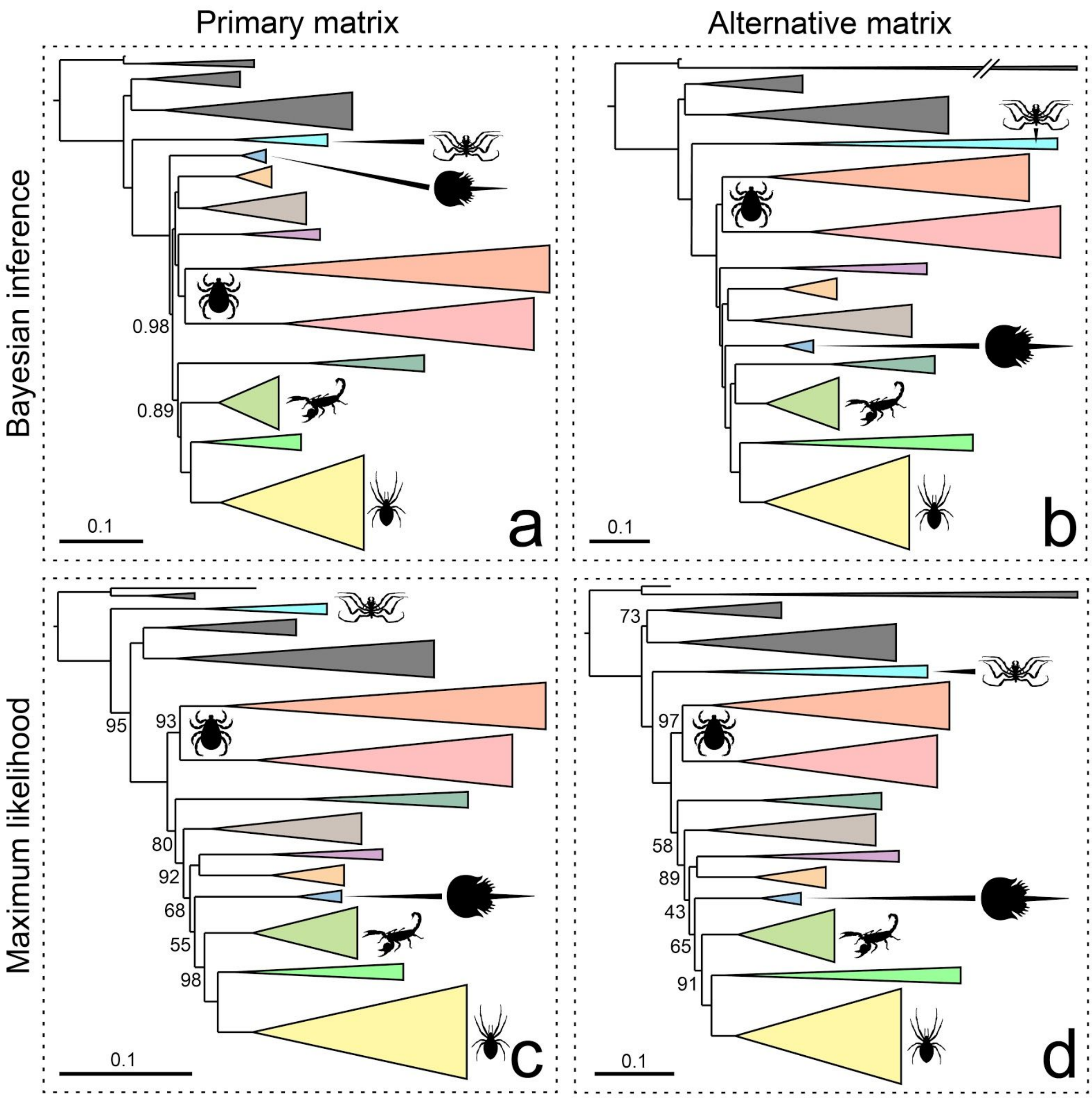


Figure 3.3. A comparison of alternative phylogenetic datasets and inference methods for Chelicerata. [a] The primary dataset inferred using PhyloBayes (CAT + GTR); node values are posterior probabilities. This figure can be seen in more detail in Figure 3.1. [b] PhyloBayes (CAT + GTR) inference of the alternative dataset, node supports are PPs. [c] Maximum likelihood (LG + I + G) inference of the primary dataset, node values are bootstrap proportions. [d] Maximum likelihood (LG + I + G) inference of the alternative dataset, node values are bootstrap proportions. All non-labelled nodes have a PP or BS of 1 or 100 respectively. All scale bars are expected substitutions per site.

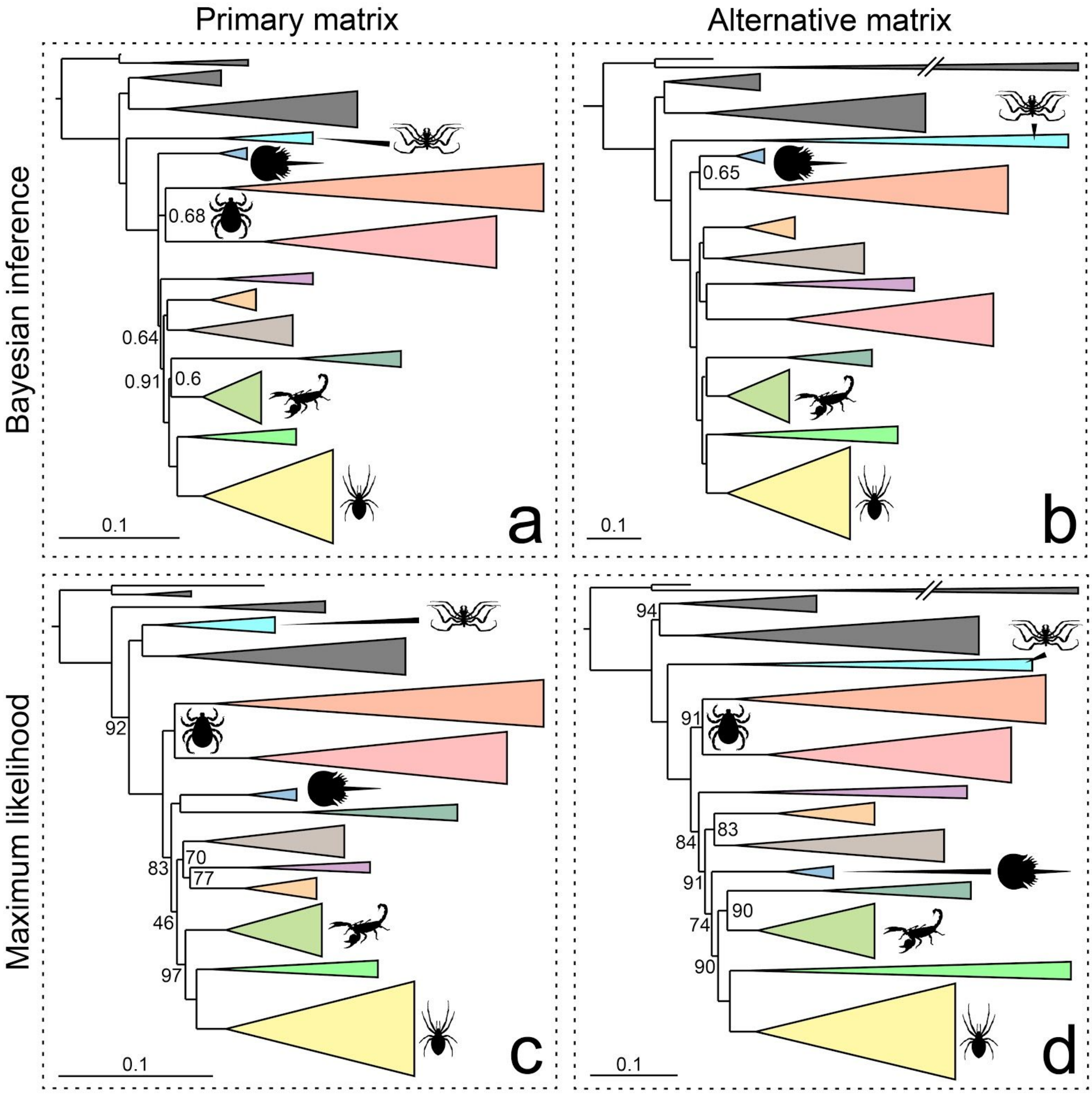


Figure 3.4. A comparison of alternative phylogenetic inference for Chelicerata, using Dayhoff-6 recoding of all datasets. [a] The primary dataset inferred using PhyloBayes (CAT + GTR); node values are posterior probabilities. [b] PhyloBayes (CAT + GTR) of the alternative dataset. [c] Maximum likelihood (LG + I + G) inference of the primary dataset, node values are bootstrap proportions. [d] Maximum likelihood (LG + I + G) inference of the alternative dataset, node values are bootstrap proportions. All unlabelled nodes have a PP or BS of 1 or 100 respectively. All scale bars are expected substitutions per site.

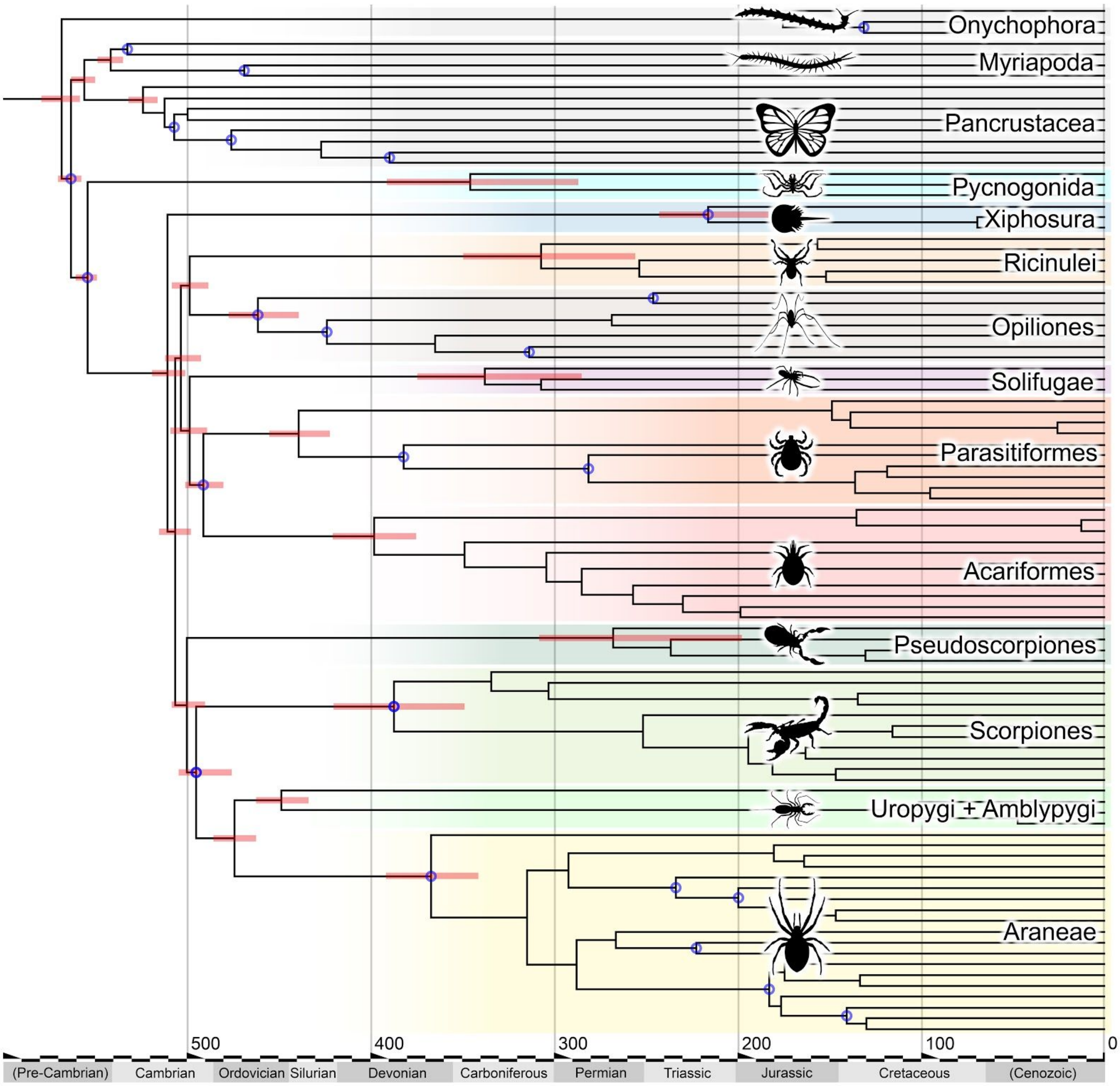


Figure 3.5. Molecular divergence time estimation for Chelicerata, plus outgroups. X-axis is Million years ago (Ma), with geological period names. Node locations are means of estimated divergence times. Red bars at nodes are 95% confidence intervals (nodes labelled to ordinal-level for clarity). Blue circles represent calibrated nodes (see Table 3.2 for full details). Silhouettes produced by ART.

3.5 Discussion

The results strongly favour a monophyletic Chelicerata (Figure 3.2), in reciprocal sisterhood to Mandibulata (Myriapoda + Panarthropoda). Pycgnogonids are descended from the earliest diversification within Chelicerata, and represent the sister clade to Euchelicerata. Xiphosurans are descended from the earliest branching event within Euchelicerata, and are the sister clade to all other members of Arachnida. With the exception of Palpigradi (the “micro whip scorpions”), we have included members of all arachnid orders, and all of these are returned as monophyletic, in agreement with morphological studies and comparative systematics (Edgecombe, 2010). Although two out of eight analyses returned Xiphosura as the sister group to the terrestrial Arachnida, in our main analysis and using the same matrix under Dayhoff recoding (in a polytomy with Acari), we advocate this sisterhood relationship as the most congruent hypothesis, given that fact that has been found in the most optimal matrix and is supported by morphological and palaeontological evidence. Placing Xiphosura in sisterhood to Arachnida is consistent with other observations, not least that a more derived position would suggest terrestrialisation and a return to marine environment for horseshoe crabs. Such a hypothesis is not supported by the fossil record, and as such a permanently marine habit for Xiphosura is the most parsimonious conclusion to support. The six analyses contravening this result can thus be seen as further evidence for deep conflicts in phylogenetic signal, as previously contended by others (Sharma et al., 2014). Here we have increased the taxonomic sample in effort to break long branches, in particular of more slowly evolving members of Acari. Considering the strength of the Bayesian inference, in concert with its reconciliation of morphological systematics, we suggest this deeper representation protocol has increased the signal for Arachnida. That arachnids are all terrestrial and share a common ancestor suggests that only a single terrestrialisation event took place. We investigate this further with the molecular divergence time estimation (Figure 3.5), and discuss this later in this section.

Mites (Acariformes) and ticks (Parasitiformes) have been traditionally grouped under the name Acari, in which the most conspicuous character exclusively shared between both lineages is the presence of the gnathosoma, a morphological unit at the front of the body composed of the chelicerae, mouth lips and pedipalps (Alberti et al., 2011). Our results support a monophyletic Acari. We did not find support some of the previous associations made between their constituent lineages, such as Acariformes allied with Opiliones (“Opilioacariform”) or with Solifugae (“Poecilophysidea”), or Acari as a whole allied with Ricinulei (“Acaromorpha”). All our analyses recovered Tetrapulmonata, and in most instances Arachnopulmonata, suggesting an association between spiders and their closest relatives (Pedipalpi) with scorpions. Therefore, our results give support to a single origin of book lungs in arachnid evolution, a hypothesis underpinned by detailed structural similarity between scorpion and tetrapulmonate book lungs (Scholtz and Kamenz, 2006). Interestingly, in all Bayesian analyses we recover Opiliones allied with Ricinulei, a clade that previously has not been proposed.

Our molecular clock experiment is guided by extensive fossil calibrations (see Table 3.2), and provides a new insight on early chelicerate evolution (Figure 3.5), given the wide taxonomic sample and appropriate outgrouping. The ancestral pycnogonid divergence is inferred to have happened prior to the Cambrian, around 555 Ma. This date suggests cryptic evolution of the euchelicerate stem group from that time into the Cambrian, due to a paucity of fossils from these very early and pre-Phanerozoic times. Subsequently, divergence from the horseshoe crab ancestral species is estimated to have occurred around the mid-Cambrian. Crown-Arachnida is inferred to have origin ~ 503 Ma ± 10 Myr. This node subtends all terrestrial chelicerates, so it is reasonable to suggest that chelicerates become terrestrial at this time, rather than later (which would require multiple terrestrialisation events across crown Arachnida). Interestingly, the origin of hexapoda is inferred to be at a similar time from this analysis.

From an ecological context, it has been suggested that appreciably complex terrestrial ecosystems may have existed up to 1 Gy (Clarke et al., 2011). As such, it suggests that the very first terrestrial chelicerates (and hexapods) were early adopters of terrestrial environments, as other recent molecular investigations have suggested (Lozano-Fernandez et al., 2016). If it is the case that these groups of arthropods were on land so early, it would lead us to speculate that the animals may have been early grazers on littoral bacterial mats, or perhaps as predators on the already-terrestrialised pancrustaceans (Clarke et al., 2011). Naturally, these speculated ecologies represent habitats highly unfavourable to fossilisation, being high-energy environments characterised by erosion (rather than deposition), and so anaerobic burial as is typically required for fossilisation would have been generally impossible (Parry et al., 2017). It is unsurprising that palaeontological insight is thus limited, and inference of the molecular kind as used here becomes more important as an investigative tool.

Moving into the present, there is clear contrast in the evolutionary tempo after the explosive radiations of the Cambrian and Ordovician. More gradual cladogenesis characterises the later Phanerozoic macroevolutionary dynamics of chelicerates, as is seen in the step-wise acquisition of sub-ordinal clades. While further taxonomic sample would prove more revealing about this, it seems clear from this molecular clock study that there is a contrast in dynamics between the early Palaeozoic, and later times, being rapid then gradual, respectively.

3.6 Conclusions

We here show that an effective approach to tackle the difficult phylogeny of chelicerates is both the expansion of taxonomy breadth, and the use of slow-evolving genes analysed under a Bayesian paradigm. Although we found a consistent phylogenetic position of several arachnid orders across datasets and methods, still some lineages are recalcitrant to robust and consistent results. Valuable future work would be to inspect the nature of gene evolution, in particular to see if orthologous assignment is being violated by usual molecular evolution, as we are cautioned to expect (Holland et al., 2017). While we show some success in reconciling molecules with morphology, a major cause for incongruence is likely due scant signal from rapid diversification during the early Palaeozoic linked to the terrestrialisation of Arachnida (Garwood and Dunlop, 2014; Rota-Stabelli et al., 2013). The monophyly of Acari we recover here encourages further work on this group to understand the origin of parasitic ecologies, and may prove valuable to epidemiological researchers. From a palaeobiological perspective, we advocate the sequencing of more taxa to approach the chelicerate phylogeny, and that sophisticated models under a Bayesian framework are statistically shown to prove most profitable and objective when viewing phylogenetic problems of this kind.

Chapter 4

Earthworm origins coincide with global ecosystem turnover and the retreat of Palaeozoic coal-forming forests.

A version of this chapter is in preparation for publication (December 2017), as a research collaboration with Luke Parry, Alexander J Hetherington, Christoffer Bugge Harder, Samuel W James, Elena Kupriyanova, Yanan Sun, Jakob Vinther. The investigation was devised and developed by ART, LP and JV. Specimens were provided for sequencing by SWJ. Molecular laboratory work was carried out by LP. Sequencing was carried out in University of Bristol Life Sciences genetics facility. Data handling, dataset curation, bioinformatics code-writing, experimental procedure, computer cluster management, and all computational analyses were carried out by ART. Bioinformatics scripting developed by ART, further details in Appendix A. Interpretation of results was carried out by ART, JV, DP, LP and AH. All figures were produced by ART. The manuscript was authored by ART, AH, CH and JV with further input from all authors.

Abstract

Earthworms are powerful ecosystem engineers in modern soils, affecting soil depth, carbon cycles, plant-fungi symbioses, and biodiversity. Earthworms are descended from marine annelid worms, however the near non-existent earthworm fossil record obscures palaeontological insight on their terrestrial origins and diversification. Here we use a molecular dataset to investigate the timing of diversification of earthworms. We recover a shift away from ancestral marine habitats in the Devonian (~380 million years ago (Ma)), and the origin of terrestrial crown group earthworms over the Carboniferous-Permian boundary (~298 Ma). This timing coincides with the collapse of continental carbon burial rates, with fundamental changes in the composition of forests flora, and with diversification of symbiotic ectomycorrhizal fungi. As such, earthworm diversification coincides with the demise of everwet coal-forming forests, and the rise of gymnosperm forests adapted to drier environments. Earthworms were thus significant factors in the rise to global dominance of gymnosperm and angiosperm forests from the Permian through to the present day.

4.1 Introduction

Earthworms have long been appreciated as crucial components of terrestrial ecosystems. In his final book, Charles Darwin (1881) compiled observations from a lifetime of fascination with earthworms, laying the foundations for investigations into their impact on soil structure and turnover, on chemical, biological and decompositional processes, and their place in terrestrial food webs (Lavelle, 2011). While the role of earthworms in enriching and aerating soils is well known to agriculture and horticulture, it is now recognised that they greatly influence terrestrial ecosystems, especially forests (Lavelle et al., 1997; Lehmann et al., 2011; Milleret et al., 2009). On a wider context, earthworms influence global biospheric chemical cycles, in particular the carbon (Pollierer et al., 2007) and nitrogen cycles (Bohlen et al., 2004; Drake and Horn, 2007). Furthermore, symbioses between plant roots and fungi are promoted by the action of worms, transporting fungal spores (Hutchinson and Kamel, 1956) and promoting air movement around roots (Dexter, 1978). With this wide repertoire of ecological roles earthworms are key ecosystem engineers, and yet our understanding of earthworm evolution and their origins has remained poor.

The earthworm fossil record is nearly non-existent, making it impossible to test historical correlation of ecosystem change against the evolutionary history of earthworms. Not only are these animals devoid of biomineralised anatomy, but also live in terrestrial, aerated habitats which seldom allow for preservation, and even then, soils are rare in the rock record (Briggs and Kear, 1993). Consequently, earthworms have among the lowest preservation potentials of all animals; it comes as little surprise that the unequivocal clitellate body fossil record features nothing beyond leech cocoons from the Lower Triassic onwards (Bomfleur et al., 2012; Manum et al., 1991a), while trace fossils attributed to earthworms are known from the Lower Triassic (Retallack, 1997). Comparative biology is also hindered, due to the morphological simplicity of earthworms.

Considering these limitations, we here apply a molecular approach to investigate the hypothesis that modern terrestrial soils and ecosystems emerged from the late Palaeozoic following the colonisation of lands by earthworms. To guide this molecular approach, a robust fossil calibration scheme is critical since the earthworm clade itself cannot be explicitly calibrated. Such a scheme was provided through the use of marine annelids, gastropods and molluscs as outgroups. Our results indicate that earthworms colonised land 298 Ma ± ~30 Myr and their initial radiation into the major families was completed by ~230 Ma and the end-Permian mass extinction, corroborating biogeographic evidence for a Pangean distribution of major crassiclittelate groups (Anderson et al., 2017). Our earthworm divergence time estimates directly and consiliently correlate with the emergence of modern terrestrial ecosystems, supporting the view that the activity of earthworms engineered modern terrestrial ecosystems. We discuss this overlap between our inference on earthworm origins and the emergence of important terrestrial ecological characteristics such as the collapse of Carboniferous-style forests, the expansion of seed-plant diversity, and the rise of root-fungi symbioses.

4.2 Experimental procedures

For all newly acquired specimens (see table 4.3), material was stored in RNAlater and total RNA was extracted using Trizol. mRNA was then purified using NEXTflex™ Poly(A) Beads. cDNA libraries were prepared using the NEXTflex™ Rapid Illumina Directional RNA-Seq Library Prep Kit using the NEXTflex™ RNA-Seq Barcodes to allow for multiplexing.

Phylogenetic data assembly

We compiled a supermatrix with data from 58 species (see table 3.1) for 197 genes, based on the gene loci of Philippe et al. (Philippe et al., 2011). The taxonomic focus was on earthworms, with 20 species representing the group. This sample covers the major families of earthworms (namely Megascolecidae, Glossoscolecidae, Moniligastridae, Lumbricidae, Eudrilidae, Hormogastridae, Sparganophilidae and Ocnerodrilidae) and represents global biogeographical coverage. As such this represents crown-group coverage, and is the widest sample coverage used for divergence time estimation for earthworms, and the second largest for general phylogenetic inference, behind the earthworm origins and Phanerozoic distribution study of James et al. (2017). New gene sequences were acquired through BLAST (Altschul et al., 1990) searching transcriptomic sequences, using *Helobdella robusta* as the search query on account of appropriate phylogenetic proximity to the group under investigation. A custom Perl script (available at github.com/jairly/MoSuMa_tools/) selected sequences on the most significant expect values (e-values) among BLAST hits, taking the lowest e-values and any other significant hits within three orders of magnitude of the most significant hit. The maximum e-value threshold was set at 10^{-10}, with hits exceeding this being excluded. These selected sequences were aligned using MUSCLE (Edgar, 2004) (default parameters), to produce gene-alignments for each of the 197 genes.

Data completeness and accessions

Taxa (as named in matrices)	% Completeness	AA positions	Location
Alloderahylae	**99.4**	**40204**	**[matrix on dryad]**
Arenicola	**97.3**	**39354**	**SRR2005653**
Astarte	**86.9**	**35145**	**PRJNA253054**
Criodrilus	**100.0**	**40430**	**[matrix on dryad]**
Diplocardia	**100.0**	**40430**	**[matrix on dryad]**
Drawida nelamburensis	**98.3**	**39743**	**[matrix on dryad]**
***Drawida* sp.**	**99.0**	**40040**	**[matrix on dryad]**
***Drilocrius* sp.**	**100.0**	**40430**	**[matrix on dryad]**
Eisenia fetida	**89.5**	**36168**	**SRR3004261**
Eisenoides	**100.0**	**40430**	**[matrix on dryad]**
Enchytraeus crypticus	**91.1**	**36842**	**[matrix on dryad]**
Ennucula	**99.0**	**40009**	**PRJNA253054**
Eudrilus	**100.0**	**40430**	**[matrix on dryad]**
Euspira	**96.3**	**38922**	**PRJNA253054**
Geogenia benhami	**100.0**	**40430**	**[matrix on dryad]**
Glossoscolex	**100.0**	**40430**	**[matrix on dryad]**
Granata	**99.8**	**40364**	**PRJNA253054**
Haementeria depressa	**40.2**	**16239**	**CN807912 [*]**
Helobdella robusta	**84.4**	**34133**	**SAMN02769625**
Hinea	**97.9**	**39568**	**SRX644676**
Hirudo medicinalis	**76.5**	**30915**	**SRR799268**
Hormogaster	**100.0**	**40430**	**SRR786599**
Hydatina	**100.0**	**40430**	**PRJNA253054**
Hydroides	**99.6**	**40277**	**[matrix on dryad]**

Lumbricus rubellus	**88.9**	**35935**	**SRX1453289 [*]**
Lumbricus terrestris	**100.0**	**40430**	**SRX2559187**
Marphysa bellii	**94.4**	**38182**	**SRR1232833**
Metavermilia	**100.0**	**40430**	**[matrix on dryad]**
Microhedyle	**94.5**	**38188**	**SRX644682**
Monodonta	**99.4**	**40198**	**PRJNA253054**
Myochama	**86.3**	**34909**	**PRJNA253054**
Neotrigonia	**96.2**	**38906**	**PRJNA253054**
Ocnerodrilidae	**100.0**	**40430**	**[matrix on dryad]**
Oxynoe	**99.8**	**40362**	**PRJNA253054**
Perotrochus	**72.2**	**29186**	**PRJNA253054**
Phallomedusa	**99.9**	**40396**	**PRJNA253054**
Pomacea	**92.2**	**37275**	**SRX2538617**
Pristina leidyi	**91.5**	**36990**	**SRX110479**
Protula	**98.9**	**39998**	**[matrix on dryad]**
Pseudopolydora vexillosa	**79.6**	**32178**	**SRR125361, SRR125360**
Rhinodrilus priolli	**99.4**	**40204**	**[matrix on dryad]**
Rubyspira	**97.2**	**39286**	**SRX644700**
Sabella	**98.0**	**39627**	**SRR1232634**
Sabellastarte	**99.5**	**40229**	**[matrix on dryad]**
Scolelepis squamata	**100.0**	**40430**	**SRR1222145**
Serpula	**97.6**	**39448**	**[matrix on dryad]**
Solemya	**98.7**	**39910**	**SRX091478**
Sparganophilus	**100.0**	**40430**	**[matrix on dryad]**
Tubifex tubifex	**89.4**	**36161**	**EY453399 [*]**
Tylodina	**99.6**	**40267**	**PRJNA253054**

Allolobophora chlorotica	**100.0**	**40430**	**SRR1324778, SRR13247**
Aporrectodea icterica	**100.0**	**40430**	**SRR1324787**
Eisenia andrei	**100.0**	**40430**	**DRR023799**
Hormogaster elisae	**99.6**	**40287**	**SRR786599**
Paralvinella sulfincola	**97.9**	**39587**	**SRX055402**
Platynereis dumerilii	**92.7**	**37464**	**ERR700601, ERR700609**
Spirobranchus lamarcki	**96.8**	**39143**	**SRR516531**
Urechis unicinctus	**96.8**	**39131**	**SRR1057936**
TOTAL	**88%**	**n/a**	**n/a**

Table 4.1: details of species used in study, number of orthologous genes recovered, completeness of data, and accessions. Accessions marked with an asterisk [*] are from multiple expressed sequence tag records - please search on NCBI for full details.

Gene matrices were then assessed for orthology in a two step process. Maximum likelihood phylogenies were inferred for each gene using PhyML (Guindon et al., 2009) version 3, modelling under LG (Le and Gascuel, 2008) and accepting the best tree of either SPR or NJ. First, sequences producing long branches were removed from the alignments, with a long branch considered to be more than 1.8 standard deviations away from the average branch length for the gene in question (script available at github.com/jairly/MoSuMa_tools/). Then trees were assessed by eye, and taxa inferred within an anomalous class were excluded from the dataset. No gene matrices failed this test. Ambiguously aligned positions were removed from the gene alignments by GBlocks v0.91b (Castresana, 2002) (*b2* = 70%, *b3* = 10, *b4* = 5, *b5* = half). The output of the GBlocks was concatenated using SequenceMatrix v100 (Vaidya et al., 2011), with a resulting supermatrix of 40,430 amino acid positions across 58 taxa.

To provide alternative topological inference, maximum likelihood approaches were also employed. PartitionFinder (Lanfear et al., 2012) and IQTree (von Haeseler et al., 2012) were used to test model fit under maximum likelihood, with both returning the substitution model of Le and Gascuel (Le and Gascuel, 2008), with a gamma distribution of rates and a proportion of invariant sites as having best model fit. The maximum likelihood topology is shown in figure 4.1.

Phylogenetic inference

The superalignment was analysed using the Markov chain Monte Carlo (MCMC) sampler PhyloBayes MPI v1.5a (Lartillot et al., 2013). The mixture model CAT + GTR + Γ_4 was applied, being the most appropriate to deal with across-site heterogeneities, while minimising long-branch biases (Bayes factor 62.2 ± 39.2, positive result in support of CAT + GTR over GTR alone). Model selection was carried out in two steps. Firstly, PartitionFinder(Lanfear et al., 2012) and IQTree(von Haeseler et al., 2012) tested empirical model fit, both returning LG(Le and Gascuel, 2008), with a gamma distribution of rates and a proportion of invariant sites as having best fit. Secondly, LG was tested against CAT+GTR by cross-validation under PhyloBayes (IQTree or PartitionFinder cannot test CAT+GTR). Bayesian cross validation scores strongly support CAT+GTR as the better fitting model (mean score = 79.47 ± 17.9879, positive score is support for CAT+GTR. Out of ten replicates, CAT+GTR outperforms both GTR+Γ_4 and LG+Γ_4 ten times).

Two independent Monte Carlo chains were run, with 25% of the Markov chains discarded as burn-in. Convergence was considered achieved with the maximum difference in the bipartitions of the chains < 0.3, as reported by *bpcomp* program in the PhyloBayes package. A further test of convergence was carried out using *tracecomp* (also under PhyloBayes), with effective sample sizes being > 50, and relative differences dropping below 0.3 for all parameters as compared between the independent chains.

To provide alternative topology, IQTree and was applied to return maximum likelihood phylogeny. LG + I + Γ_4 was selected as model of best fit by ModelFinder (Kalyaanamoorthy et al., 2017), with 1000 bootstrap replicates to generate node confidence statistics. Output trees of all analyses were viewed in FigTree version 1.4.2 (available at tree.bio.ed.ac.uk). The output consensus tree from IQTree returned the same topology as the main Bayesian analysis, see Figure 4.6.

In order to test if long-branches proximal to the earthworm clade were causing topological instability, further Bayesian analyses were run with leeches removed from the matrix. Leeches are mostly parasitic species, known to be under selection pressures leading to phylogenetic long branches. The topology across the remaining taxonomic sample remained stable, suggesting that although the leeches exhibit long branches on the tree, that this was not biasing other results, see Figure 4.2 and 4.6.

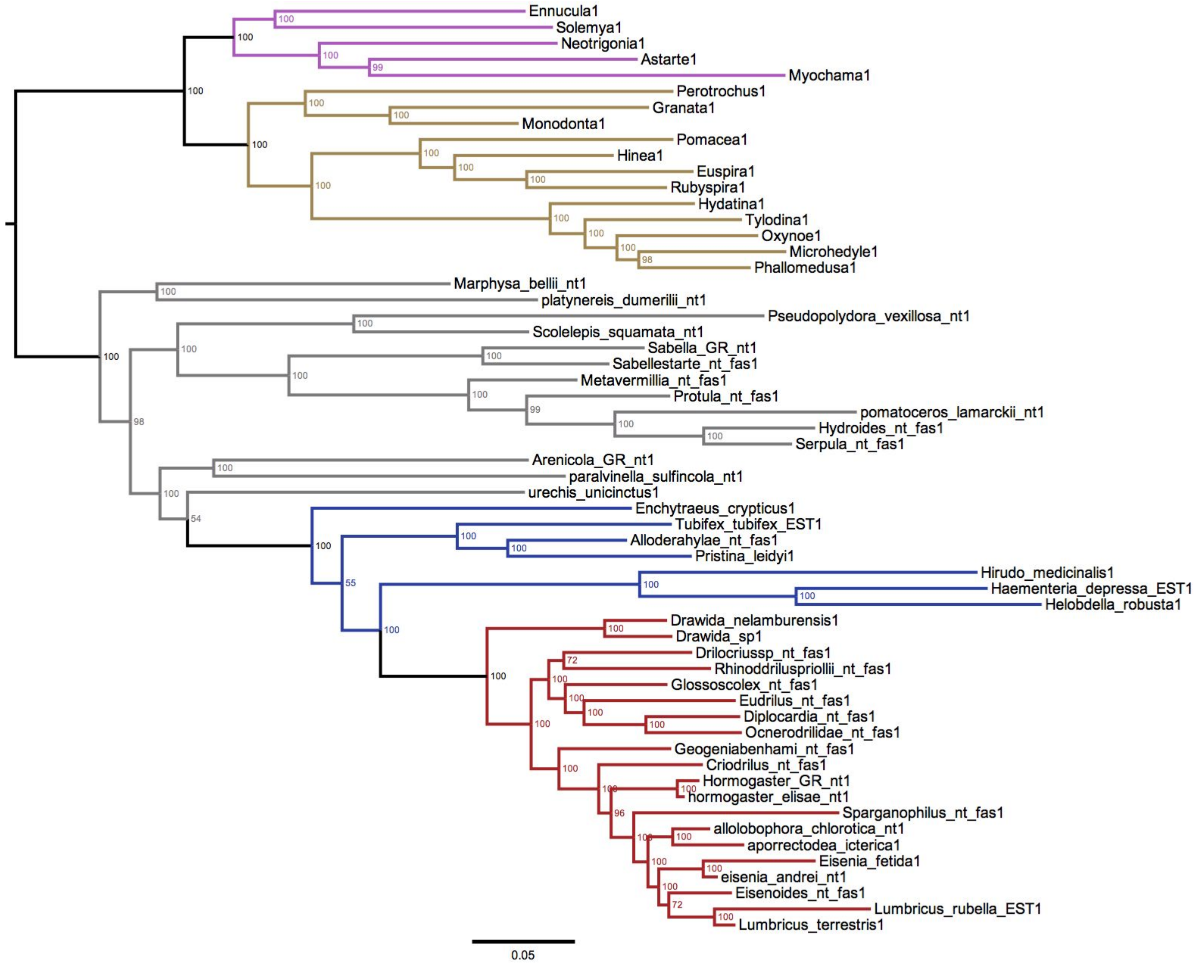


Figure 4.1. Maximum likelihood phylogeny of annelids, plus gastropod and bivalve outgroup. LG + I + Γ, 1000 bootstrap replicates, run under IQtree 1.4.3. Numbers at nodes represent bootstrap support values, branch colours follow colouring regime used in main paper figures. Pink: bivalve molluscs, yellow: gastropod molluscs, grey: non-clitellate annelids, blue: non-earthworm clitellates, red: earthworms. See figure 4.6 for full details. Scale bar is number of expected substitutions per site.

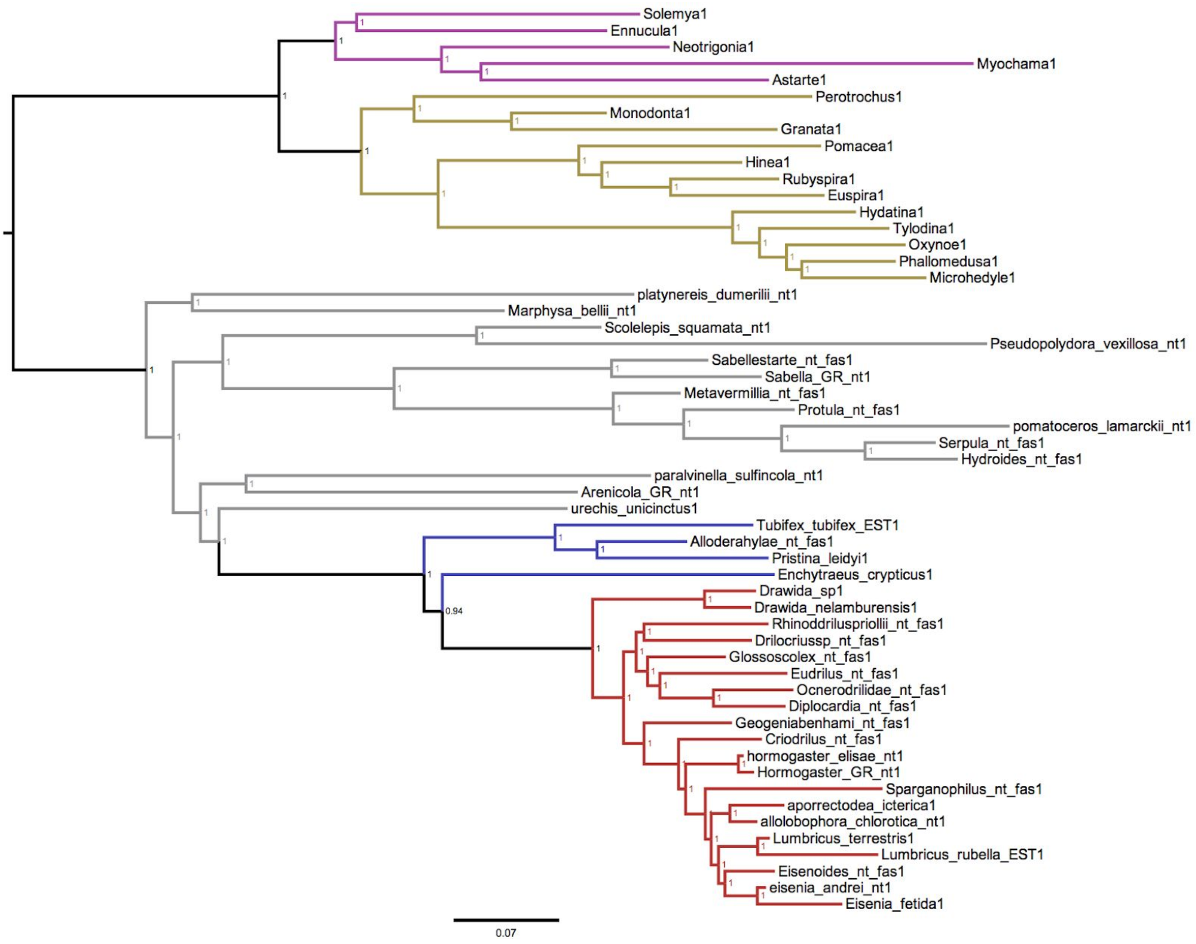


Figure 4.2. Bayesian analysis across annelids with gastropod and bivalve outgroup, minus leechs to assess potential long-branch bias artefacts. Two MCMC chains run under CAT + GTR, convergence assessed from *tracecomp* and *bpcomp* statistics: ESS scores > 50, relative differences < 0.3, and maximum bipartition differences < 0.1. Branch colours follow colouring regime used in main paper figures. Pink: bivalve molluscs, yellow: gastropod molluscs, grey: non-clitellate annelids, blue: non-earthworm clitellates, red: earthworms. See figure 4.6 for full details. Scale bar is number of expected substitutions per site.

Molecular divergence time estimation.

Phylobayes 3.3f was used to infer the molecular clock using CAT + GTR, as supported as the best fitting substitution through cross-validation (62.2 ± 39.2, positive result in support of CAT + GTR over GTR alone: CAT + GTR outperforms GTR on ten out of ten replicates), soft-bounds on the calibration intervals (Yang and Rannala, 2006) of 0.05, root prior of 555 Ma ± 10 Myr, and a Yule-process birth-death model. Cross-validation returned support for CIR over UGAM (1.91 ± 1.22, positive result in support of CIR over UGAM: CIR outperforms UGAM in seven out of ten replicates). Root prior standard deviations were tested at 5, 10, 20 and exponential distribution, see Figure 4.3. All priors were shown to be appropriate except exponential distribution, which would have influenced posterior inference through unrealistic spread of prior values. A root prior standard deviation of 10 was chosen to reasonably represent the uncertainty in *Kimberella* timing, the fossil use as root prior. This prior was tested as being appropriate by MCMC chains being run without data, to confirm that the samples were being drawn from a distribution that includes the prior. The root age of the prior run was 552 Ma ± 8 Myr, supporting the prior as appropriate (Figure 4.3). The topology was fixed to that inferred by PhyloBayes MPI v1.5a (Figure 4.5), the root constrained to the bifurcation between the uncontroversial reciprocal monophylogenetic assemblages of Annelida and Mollusca; the bivalves and gastropods were considered a balanced outgroup with comparable taxonomic sampling and phylogenetic crown spread.

Test of timings and priors as robust and appropriate were carried out by altering the standard deviation on the root prior. Standard deviation values of 5, 10, 20 Ma, plus an exponential standard deviation (the root prior given a distribution extending from 0 to the age of the distribution, ie a highly uninformative prior), were applied and the posterior intervals compared (Figure 4.3). These supported a root prior with standard deviation of ± 10 Myr as most appropriate, applied to constrain the primary analysis (Figure 4.5).

Ten fossil calibration points were applied to the analysis (see Table 4.2). Two independent MCMC chains were run for each model, with convergence being determined by the root age of each run agreeing to within 1 Ma, with the first 25% of states being discarded as burn-in, and with *tracecomp* statistics between the two chains returning effective sample sizes over 50, and relative differences between parameters below 0.3.

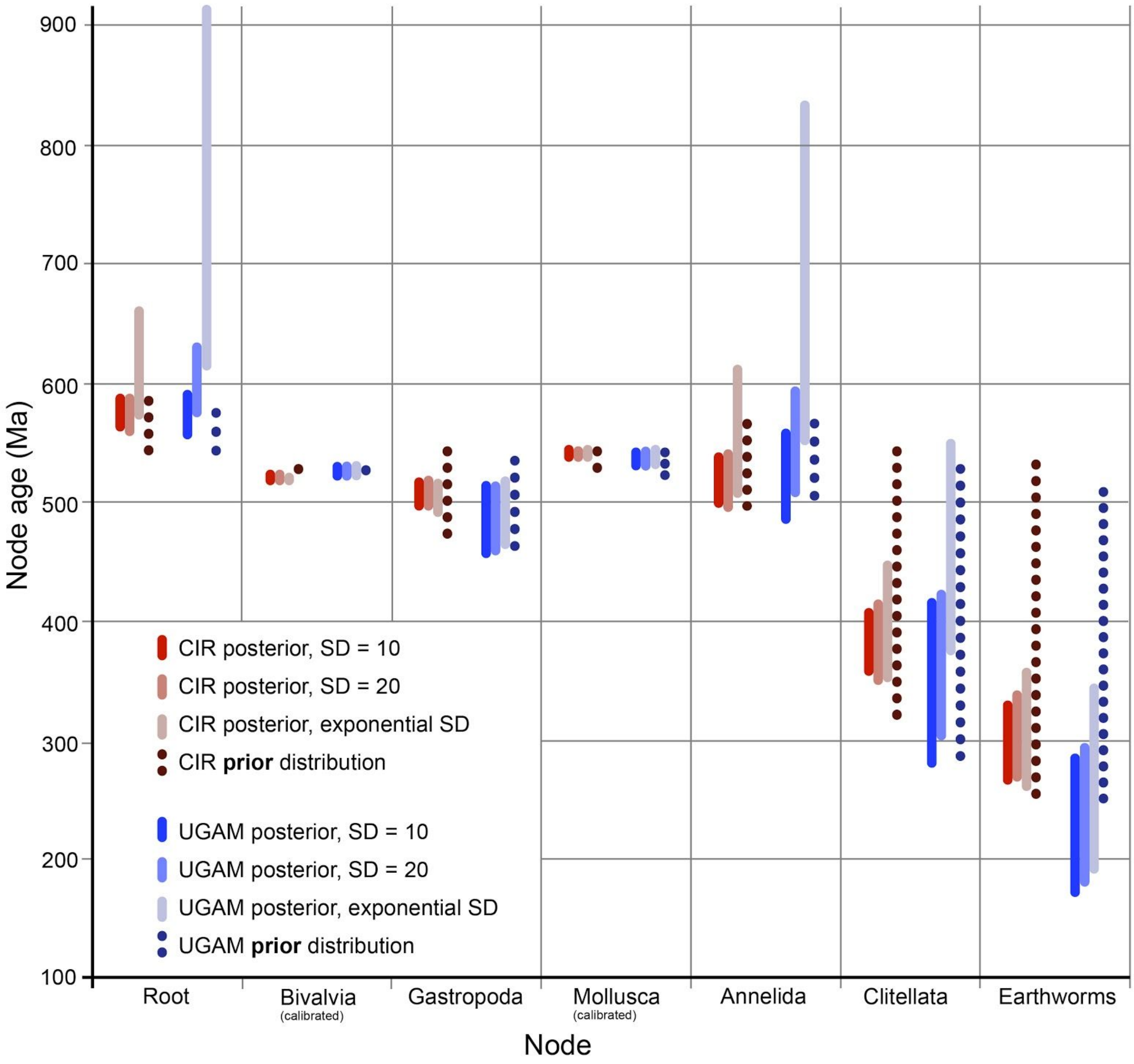


Figure 4.3. Comparison of CIR (Lepage et al., 2007) vs UGAM (Drummond et al., 2006) clock models on timing of node inference, and comparisons of root prior standard deviation constraint on timing of node inference. Red bars denote CIR node age distributions; blue bars denote UGAM node age distributions, see legend for further details.

Fossil calibrations

Node on fig S4	Node	Max.	Min.	Reference
a	Gastropoda + Bivalvia	543 Ma	525 Ma	(Benton et al., 2015)
b	Bivalvia	-	485 Ma	(Parkhaev, 2008), pp. 33-69
c	Vetigastropoda	-	490 Ma	Edgecombe *et al.* 2011
d	Caenogastropoda + Heterobranchia	-	418 Ma	(Frýda et al., 2008), pp. 239-270
e	*Serpula* + *Hydroides*	-	66 Ma	Ippolitov *et al.* 2014(Ippolitov et al., 2014)
f	*Serpula* + *Metavermilia*	-	205 Ma	(Ippolitov et al., 2014)
g	*Serpula* + *Spirobranchus*	-	190 Ma	(Ippolitov et al., 2014)
h	*Serpula* + *Sabellastarte*	-	254 Ma	(Sanfilippo et al., 2017)
i	*Platynereis dumerilii* + *Marphysa bellii* (crown group Errantia)	-	476.5 Ma	(Hints and Eriksson, 2007)
j	*Hirudo medicinalis* + *Glossoscolex*	-	200 Ma	(Manum et al., 1991b)

Table 4.2. Fossil calibrations applied to molecular clock analyses, with source references.

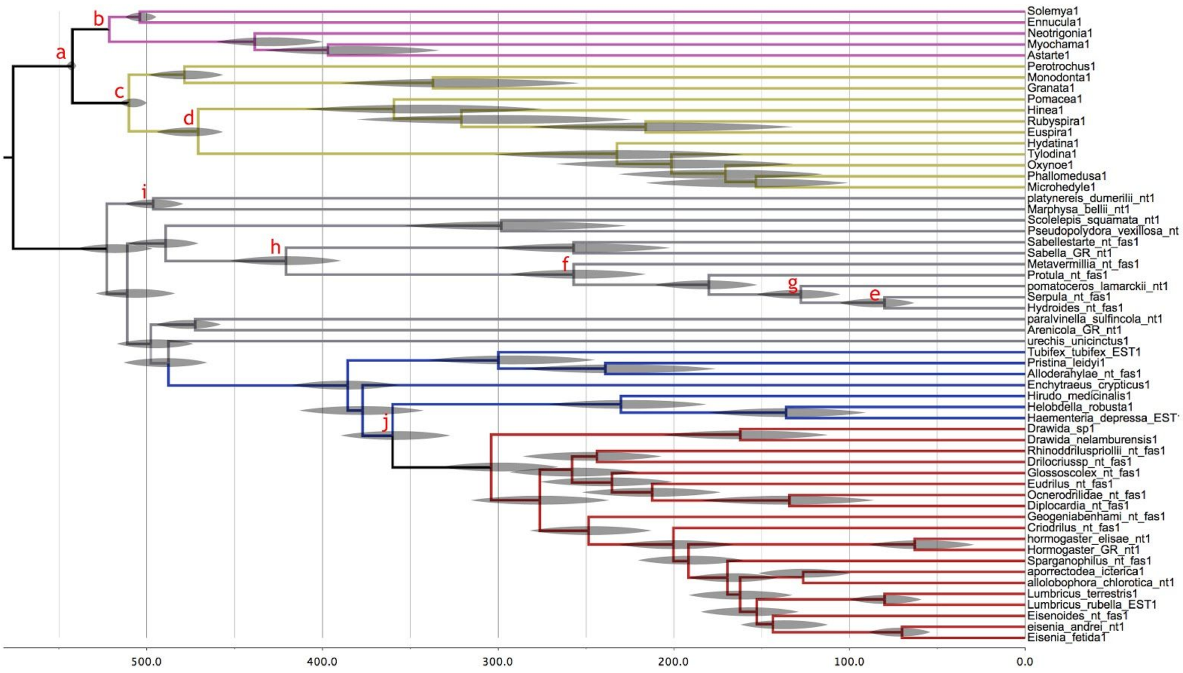


Figure 4.4. Node positions of fossil calibrations, as defined in table 4.2. Letters a-j denote calibrated nodes. Elipses at nodes represent 95% confidence intervals for node ages, taper representing lower probability at the extremities of these distributions. Scale is Ma, millions of years before present.

Ancestral state reconstruction

The R package Ancestral Character Estimation was used to investigate the likelihood of ancestral conditions for states at nodes on the Bayesian molecular clock tree. The character states of marine (red in Figure 4.5), freshwater (blue), and terrestrial (green) were inferred through marginal likelihood.

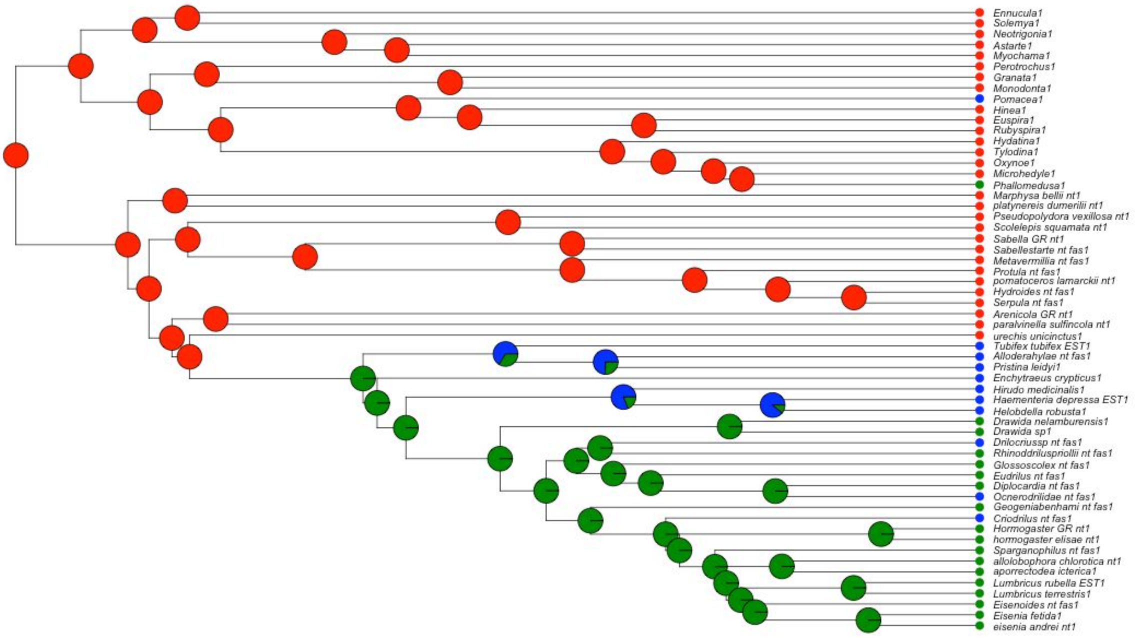


Figure 4.5. Ancestral state reconstruction of marine (red), terrestrial (green) and freshwater (blue) across annelids plus outgroups, with a focus on earthworms. Colours at tips represent habitat of extant taxonomy used in this study. Pie charts at nodes represent probabilities of ancestral states - completely red nodes suggest very high likelihood of marine condition, but arithmetic limitations have results in probabilities of 1.

PBDB diversity over time curve data retrieval

Searches were carried out on The Palaeobiology DataBase, (https://paleobiodb.org/classic/displayDownloadGenerator), on 16/03/2017. Searches were made at a genera level, with time bins at epoch resolution.

Seed plant curve used the search terms "Alethopteridae, Alismaceae, Araceae, Arberiopsida, Archaefructaceae, Arecaceae, Bennettitopsida, Cordaitaceae, Cycadopsida, Cymodoceaceae, Cyperaceae, Dicotyledonae, Ginkgoopsida, Gramineae, Magnoliopsida, Medullosaceae, Musaceae, Najadaceae, Palmae, Peltaspermopsida, Pinopsida, Posidoniaceae, Smilacaceae, Sparganiaceae, Sphenopteridae, Triuridaceae, Voltziopsida, Zannichelliaceae, Zingiberaceae, Zosteraceae".

Spore plant curve applied "Calamitaceae, Cladoxylopsida, Equisetopsida, Filicopsida, Lepidodendraceae, Lycopsida, Polypodiopsida, Progymnospermopsida, Psilophytopsida, Pteridopsida, Rhyniaceae, Ruppiaceae, Sigillariaceae, Zosterophyllopsida, Lycopodiophyta, Lycophytina".

Classification	Search term	Records
Seed plant	Alethopteridae	214
Seed plant	Alismaceae	34
Seed plant	Araceae	51
Seed plant	Arberiopsida	459
Seed plant	Archaefructaceae	2
Seed plant	Arecaceae	140
Seed plant	Bennettitopsida	2442
Seed plant	Cordaitaceae	20
Seed plant	Cycadopsida	1763
Seed plant	Cymodoceaceae	14
Seed plant	Cyperaceae	223
Seed plant	Dicotyledonae	5803
Seed plant	Ginkgoopsida	4636
Seed plant	Gramineae	94
Seed plant	Magnoliopsida	20519
Seed plant	Medullosaceae	517
Seed plant	Musaceae	3
Seed plant	Najadaceae	209
Seed plant	Palmae	98
Seed plant	Peltaspermopsida	1370
Seed plant	Pinopsida	8936
Seed plant	Posidoniaceae	21
Seed plant	Smilacaceae	17
Seed plant	Sparganiaceae	200
Seed plant	Sphenopteridae	836
Seed plant	Triuridaceae	4
Seed plant	Voltziopsida	85
Seed plant	Zannichelliaceae	6
Seed plant	Zingiberaceae	68
Seed plant	Zosteraceae	45
Spore plant	Calamitaceae	1096
Spore plant	Cladoxylopsida	108

Spore plant	Equisetopsida	1953
Spore plant	Filicopsida	10938
Spore plant	Lepidodendraceae	551
Spore plant	Lycopsida	2095
Spore plant	Polypodiopsida	10938
Spore plant	Progymnospermopsida	18
Spore plant	Psilophytopsida	106
Spore plant	Pteridopsida	10938
Spore plant	Rhyniaceae	8
Spore plant	Ruppiaceae	6
Spore plant	Sigillariaceae	135
Spore plant	Zosterophyllopsida	41
Spore plant	Lycopodiophyta	59
Spore plant	Lycophytina	96

Table 4.3: Details of records returned by the Palaebiology Database for each search term included in the classifications of “seed plant” and “spore plant”, as used to generate diversity curves on figure 4.8.

4.3 Results

Earthworms (Crassiclitelata) emerge as a monophyletic group within a monophyletic Clitellata (figure 4.6). Within Clitellata, leeches and Naididae (typified by the sludge worm *Tubifex*) emerge as sequential sister groups of the earthworms. Earthworm systematics is still controversial, and some of the relationships that we recovered disagree with previous studies such as James and Davidson (James and Davidson, 2012), yet recognised groups such as Lumbricidae are recovered. However, investigation of earthworm origins requires crown-group representation such that important ancestral divergences can be dated, which the taxonomic sample of our dataset satisfies, particularly by the inclusion of *Drawida* and *Enchytraeus* to demark the root of earthworms, as well as a wide crown-group representation (Figure 4.6).

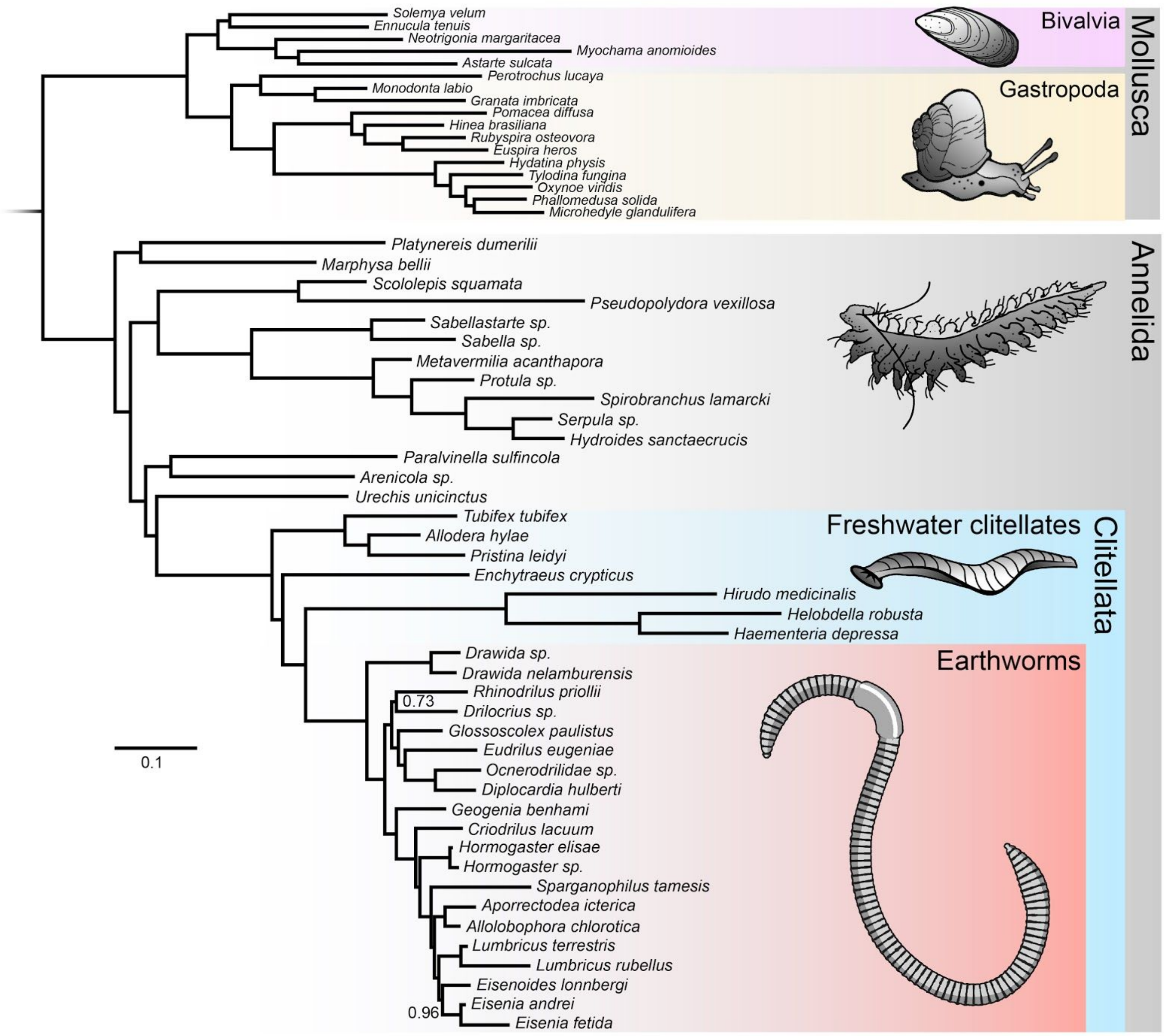


Figure 4.6. A phylogeny of Annelida (plus molluscan outgroup) inferred from molecular sequence data. Concatenated matrix of 197 genes, with matrix dimensions of 58 taxa and 40,430 amino acid positions. Topology inferred through PhyloBayes MPI version 1.6j, applying CAT + GTR + G. All nodes have a posterior probability of 1, except two marked nodes. Scale bar is expected substitutions per site. Illustrations by ART.

Our molecular divergence time estimates indicate an origin of annelids between 495 and 539 Ma (Figure 4.7), with the highest probability density placed over the mid- to early Cambrian. Molecular clocks were employed applying the auto-correlated CIR clock and uncorrelated gamma clock, with cross validation supporting CIR as model of best fit. The common ancestor of Clitellata arose between 350 and 410 Ma, while the earthworm crown is estimated to have radiated at or just before the Carboniferous-Permian (Figure 4.6, figure 4.7). The major earthworm families Lumbricidae, Glossoscolecidae, Moniligastridae and Microchaetidae emerged between 238 and 298 Ma. Subsequent cladogenesis among the sampled earthworms took place throughout the Mesozoic in particular. The earthworm sistergroup of leeches (Hirudinea), have estimated origins around 225 Ma.

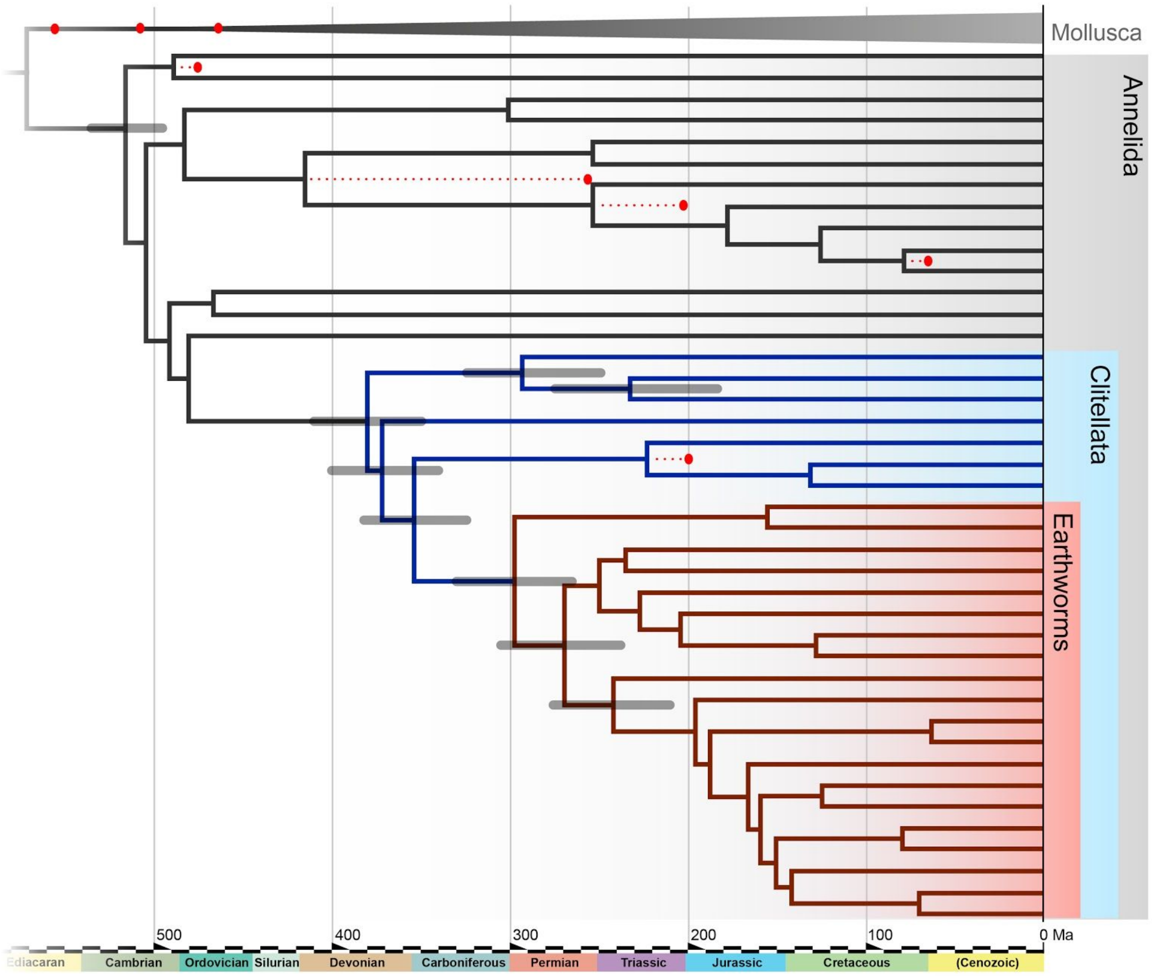


Figure 4.7. Divergence time estimations of earthworms and other clitellates, calibrated by annelids and molluscs. Matrix dimensions of 58 taxa and 40,430 amino acid positions. Timings inferred through PhyloBayes version 4.1c, applying a CAT + GTR substitution model, CIR clock model, soft-bounds of 0.05, Yule birth-death prior, root prior of 555 Ma ± 10 Myr representing *Kimberella*. Grey bars at nodes are 95% confidence intervals; not all nodes are labelled to aid clarity (see Figure 4.4 for further details). Red circles are calibration positions, see Table 4.2. Topology was constrained to that presented in Figure 4.6.

4.4 Discussion

Earthworms, plant communities and soils are closely linked (Lemtiri et al., 2014; Wurst, 2010), yet our results show that earthworms originated into already diverse terrestrial ecosystems rather than in concert with other major terrestrial groups earlier in the Palaeozoic. These finding are are consistent with the very vague fossil record for earthworms. Notably, trace fossils of earthworm activity are not common, and entirely absent prior to the Triassic. This may be due to a combination of factors, principally that soil preservation and subsequent sampling is uncommon. That trace fossils are absent in the Palaeozoic is consistent with a Carbo-Permian origin of earthworms, as upheld by our analysis. Our phylogenetic inference recovers already well-supported clades, with earthworms nested as monophyletic within clitellates, clitellates themselves as a clade within annelids, which itself is in sisterhood to molluscs (given this taxonomic sample). This robust results is suitable for further inference and interpretation, an in particular molecular divergence estimates and ensuing reflection on macroecological conditions.

Although the origin of earthworms occurred in the context of an already-mature terrestrial ecology, our investigation places their origins at a time characterised by comprehensive environmental and ecosystem turnover. Phanerozoic carbon burial peaked in the late Carboniferous before diminishing rapidly into the Permian (Nelsen et al., 2016) (Figure 4.7, Figure 4.8), a shift also recorded in atmospheric CO_2 levels (Montañez et al., 2007). This same time period was further characterised by climate oscillations between glacial and interglacial periods similar to our present climate (DiMichele et al., 2001), and with associated changes in both temperature and sea level (Montañez et al., 2007). Between the Carboniferous and Permian, land plants underwent a major mass extinction due to loss of tropical wetlands forests across Euramerica (Cascales-Miñana and Cleal, 2014). Coal forests which dominated the ever-wet intervals of the Carboniferous period receded and the seed plants of seasonally dry biomes of the Carboniferous rose to global dominance (DiMichele et al., 2001; Falcon-Lang, 2015; Montañez et al., 2007). These shifts led to a rapid diversification of seed plants, continued diversification of seed ferns such as glossopterids, as well as the expansion of the conifers, cycads and ginkgoales (DiMichele et al., 2001; Falcon-Lang, 2015; Montañez et al., 2007). Seed plants therefore underwent a major diversification in the Permian, a time when spore plants were declining in diversity (Niklas et al., 1983) (Figure 4.8). This marked the start of seed plant dominated forests. Our results therefore place the origin of earthworms at a point of critical environmental and ecosystem turnover. As such, an examination of the terrestrial context for the origin of earthworms and their roles as ecosystem engineers becomes pertinent.

Coal swamp forests dominated equatorial regions during wet glacial phases of the Carboniferous, and led to the burial of vast volumes of peat, drawing down CO_2, later to become coal (Herendeen, 2015; Nelsen et al., 2016). Carboniferous organic carbon burial rates peaked in the Phanerozoic, but abruptly declined over the Carboniferous-Permian boundary, coinciding with the collapse of the coal swamp forests as the dominant tropical forest ecotype. Post-dating the Permian (the predicted origin of earthworms), the mode and scale of carbon burial never returns to Carbo-Permian levels (Berner, 2003). Why organic carbon never increased to a similar scale after the Carboniferous period, despite tectonic and climatic conditions cycling through similar scenarios conducive to mass carbon burial, has yet to be understood (Floudas et al., 2012; Robinson, 1990). Our results prompt us to advocate earthworms as playing an important role in attenuating post-Palaeozoic cycles of terrestrial carbon burial.

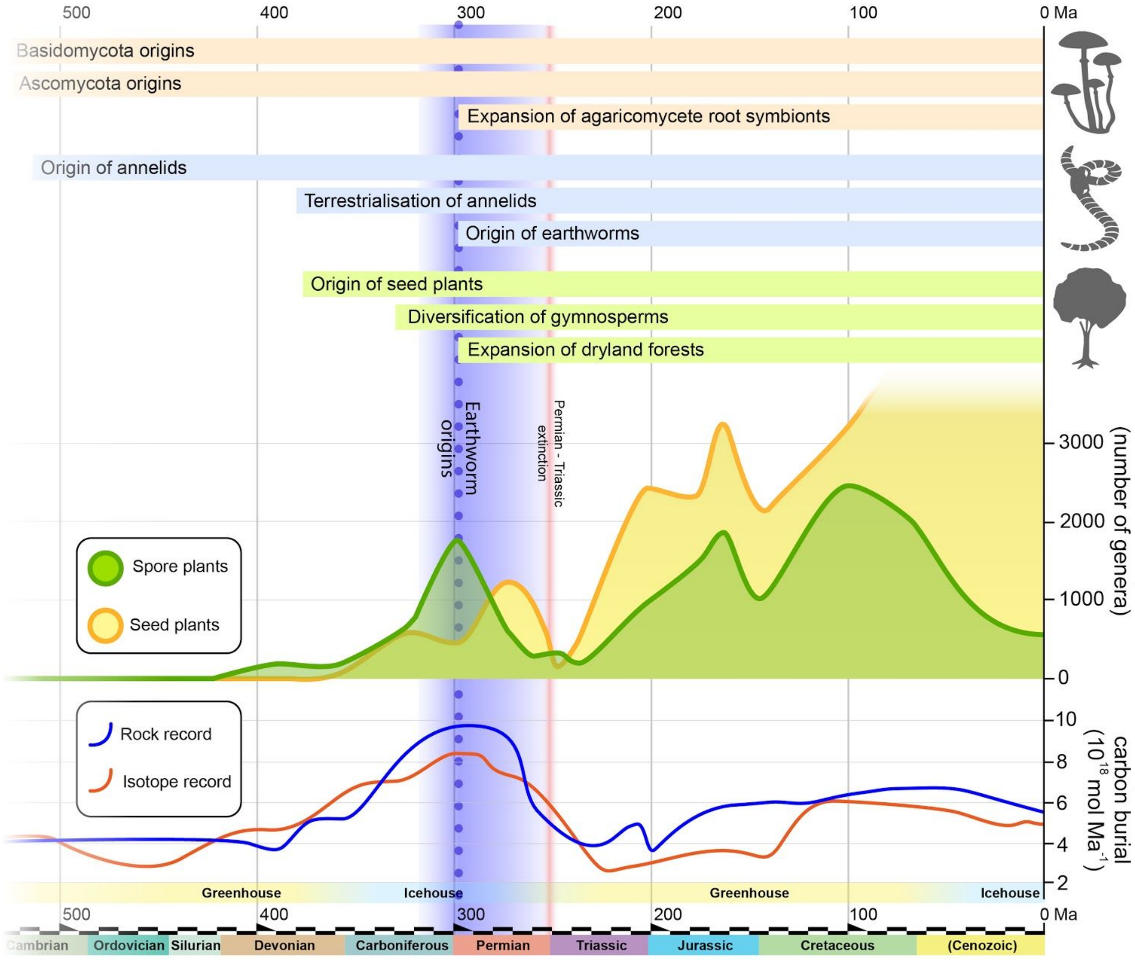


Figure 4.8. A timeline of major evolutionary transitions and the evolution of key fungi and terrestrial plants, in the context of annelid divergence estimates as presented in figure 4.7. Blue shaded vertical bar represents estimated time of earthworm origins. Timings of fungal transitions are adapted from Floudas et al (Floudas et al., 2012). Major events in plant evolution from Gibling and Davies (Gibling and Davies, 2012). Plant diversity curve recovered from Palaeobiology Database (pbdb.org), see table 4.3 for search criteria. Organic carbon burial curves adapted from Berner (Berner, 2003). Illustrations by ART.

Earthworms have a central role in accelerating organic material breakdown and the release of carbon to atmospheric CO_2 (Lubbers et al., 2013). The origin of earthworms would therefore have increased the rate of organic decay, decreasing the potential for organic carbon burial. Carbon burial is also highest in areas with limited decay such as anoxic and waterlogged soils. Earthworms may have helped decrease the amount of waterlogged soils by increasing soil drainage (Edwards et al., 1990). It is notable that the non-linear decline in carbon burial rates (Berner, 2003) (Figure 4.8) suggest underlying positive-feedback processes, as satisfied by our earthworm hypotheses. Underpinning this idea further, is that diversification of earthworms decelerates carbon storage in peatlands, and, reciprocally, the drainage of peatlands promoted earthworm invasion (Wu et al., 2017). We suggest that a similar mechanism but on vast geographical scale, as supported by our molecular clock results, contributed to the collapse in carbon burial rates at the Carboniferous-Permian transition.

As the coal swamp forests were receding, the warmer, drier climate of the Permian allowed the rise to dominance and subsequent taxonomic radiation of seed plants (DiMichele et al., 2001; Falcon-Lang, 2015; Montañez et al., 2016). This expansion of dryland forests (Poulsen et al., 2007) saw a taxonomic diversification of gymnosperms including the rise in diversity of conifers e.g. *Walchia* and *Ernestiodendron* (DiMichele et al., 2001; Falcon-Lang, 2015; Montañez et al., 2007), the continued diversification of seed ferns such as glossopterids, as well as the evolution of the cycads and ginkgoales (DiMichele et al., 2001; Falcon-Lang, 2015). Such shifts can be the consequence of both abiotic and biotic factors, including the action of organisms that function as ecosystem engineers. Extant earthworms are key drivers of plant community change (Bohlen et al., 2004; Gundale, 2002; Hale et al., 2006; Nuzzo et al., 2009), and can furthermore be seen to directly increase plant productivity (van Groenigen et al., 2014). We predict earthworms played a central role in this transition in land plant macroevolution: the diversification of seed plants was due to the direct ecological engineering roles earthworms have on plants communities. Our findings indicate that this role for earthworms extends back to their Palaeozoic origins, and that they were key in the increased diversity of land plants (Figure 4.3).

Seed plants have historically been more adept than spore plants in exploiting fungal symbioses and our timing for earthworm origins coincides not only with gymnosperm taxonomic radiation, but also with the earliest known gymnosperm ectomycorrhizal (ECM) host families Pinaceae/Gnetaceae (Hibbett and Matheny, 2009), which have a minimum age of 270 Ma (Wang, 2004). Earthworm burrowing behaviour is known to fundamentally alter soil environments by removing the litter layer and thereby decreasing carbon storage (Bohlen et al., 2004), facilitating transport and accessibility of nitrogen and phosphorous (Milleret et al., 2009), transporting fungal spores along their paths (Hutchinson and Kamel, 1956), and enhancing the nitrogen uptake of mycorrhizal host plants (Yang et al., 2015).

Thus, their appearance also likely had major implications for the below-ground biotic interactions in Carbo-Permian soils.

The arbuscular mycorrhiza (AM)-forming Glomeromycota has a fossil record extending to the early Devonian (Taylor et al., 1994), and it is widely accepted that their coevolution with plant hosts was an important factor for the radiation and diversification of early land plants (Brundrett, 2002; Wang et al., 2010). However, invasive earthworms in North America negatively affected AM-communities by impairing their colonisation of hosts; reducing the diversity of AM host plants; and increasing phosphorus availability, thus reducing host dependence on AM services (Hale et al., 2006; Lawrence et al., 2003; Paudel et al., 2016). The most likely mechanism behind this is that aseptate hyphae are sensitive to the mechanical damage caused by burrowing behaviour (Gundale, 2002), meaning that large parts of a widely-spread fungal network can be destroyed by leakage at just one point. In contrast to AM fungi, filamentous hyphae of ECM species are all septate (i.e., split into cell compartments separated by septa that can be closed if the neighbouring cell is damaged), which makes their networks more robust to mechanical damage by burrowing. A plausible hypothesis consistent with these results is that the origin of earthworms and their burrowing behaviour created a soil environment that was more conducive to the evolution of gymnosperms and their ectomycorrhizal symbionts at the expense of the more ancient arbuscular mycorrhiza species and their hosts.

Symbioses are correlated with expansions of diversity (Herre et al., 1999) and consequently we suggest that the action of earthworms as distributors of ECM fungal spores qualifies earthworms as drivers of the expanded plant diversity ECM promotes. Our molecular clock results strongly support this hypothesis, and our general conclusion: that the narrative of terrestrial ecosystem macroevolution over the Carbo-Permian transition was driven, at least in part, by the emergence of earthworms.

4.5 Conclusions

The collapse of the Carboniferous forests can be viewed as one of the most important terrestrial events of the Phanerozoic, and undoubtedly is the result of many interacting biotic and abiotic factors. We promote the view that key ecosystem engineers, in this case earthworms, have major roles in ecological change at a global level. Further support for this hypothesis comes from the observation that over the ensuing 250 Ma through to the present, carbon burial and terrestrial forest ecosystems are not seen to return to Carbo-Permian conditions, despite abiotic drivers returning to their Carbo-Permian states at various times. Thus, biotic drivers must have been significant in influencing terrestrial ecosystem turnover, and in preventing a return to earlier ecological community structure. With the timing of earthworm diversification overlapping this ecological shift, and in light of the importance of earthworms to plant communities, we conclude that the demise of the Carboniferous forests, and the accompanying collapse of carbon burial rate and expansion of seed plant diversity, was at least in part driven by the origin and diversification of earthworms.

Chapter 5

The origin of annelids

This chapter has not been published. Sections of the introduction are adapted from the published review article, The Origin of Annelids (2014), Luke Parry, Alastair R Tanner, and Jakob Vinther, *Palaeobiology doi: 10.1111/pala.12129.* Phylogenetic work is a collaboration with Luke Parry, Samuel W James, Elena Kupriyanova, Yanan Sun, Katrine Woorsae, Jakob Vinther and Davide Pisani. The investigation was devised and developed by ART, LP and JV. Specimens were provided for sequencing by SWJ, LP, JV, KW and EK. Molecular laboratory work was carried out by LP. Sequencing was carried out in University of Bristol Life Sciences genetics facility. Data handling, dataset curation, bioinformatics code-writing, experimental procedure, computer cluster management, and all computational analyses were carried out by ART. Bioinformatics scripting developed by ART, further details in Appendix A. Interpretation of results was carried out by ART, JV, DP and LP. All figures were produced by ART.

Abstract

Annelids, the ringed worms, make up one of the largest and most ecologically diverse animal phyla. Annelids have important and distinct ecological roles, being predators, detritivores, filter-feeders, terrestrial bioturbators, and colonial-hydrothermal symbionts. Annelid phylogeny has represented one of the most difficult to understand among Metazoa, characterised by discord between morphological and molecular systematics. Despite utilising a range of sources of molecular data, phylogenetic inferences have not approached a consensus on annelid evolutionary relationships, and further interpretation of their early history has remained obscure. Here we apply a suite of inference methods and dataset assembly protocols to approach the annelid phylogeny. We confirm that phylogenetic signal is highly conflicted, and that rapid divergence and speciation early in their evolutionary history has probably contributed to this lack of resolution. We suggest further methods to explore the nature of the phylogenetic conflict, and how potentials for minimising biases, and discuss how such difficult phylogenetic problems can allow us to reflect on evolutionary inference, and contribute to phylogenomic methodology in general.

5.1 Introduction

Annelids, the segmented worms, are a diverse phylum adopting ecologies from terrestrial forests to oceanic abyssal plains. Their adaptations and lifestyles are exceptionally varied, with ecologies including filter feeding, ambush predation, parasitism, pelagic detritivory, whale-bone-scavenging, hydrothermal vent colonialism, as well as the familiar soil-dwelling habit of earthworm (Rouse and Pleijel, 2001). Beyond segmentation, most annelids are characterised by having bristle-like chaetae, as well as serially-iterated lateral outgrowths of the body, termed parapodia (Rouse and Pleijel, 2001). In many groups this vermiform body plan is highly specialised, resulting in bizarre derivations in morphology. While the breadth of annelidan diversity, speciosity, and ecology has led to the group being the focus of palaeontological and phylogenetic research, their evolutionary origins and systematics remain controversial.

At a broad classification level, Annelida is a phylum within the clade Lophotrochozoa: the annelids plus the phyla Mollusca, Sipuncula, Brachiopoda and Phoronida (Halanych et al., 1995; Edgecombe et al., 2011). Annelids are generally supported as monophyletic (Rousset et al., 2007; Weigert et al., 2014), although there is ongoing discussion and research as to whether sipunculans represent their own phylum, or should be considered to be annelids (Struck et al., 2011). Traditionally, annelids were classified as two morphologically distinct groups, the clitellates (also "oligochaetaes") and polychaetes (Rouse and Fauchald, 1997; Weigert and Bleidorn, 2016), and this view led some to suggest reciprocal monophyly between the two (Fauchald, 1974). Systematic and molecular phylogenetic work has now rejected this view, and it is strongly supported that while clitellates are monophyletic, they share a common ancestor within polychaetes, thus rendering the latter paraphyletic (Struck et al., 2011; Weigert et al., 2014).

Clitellates are predominantly terrestrial annelids, with the most familiar members of the group being leeches and earthworms (see Chapter 4), plus two minor and obscure clades, acanthobdellans and the branchiobdellans. Morphologically, clitellates are united in having the "clitellum" reproductive structure (Erséus, 2005), lacking parapodia and nuchal organs, and having highly reduced chaetae (thus "oligochaete"). While the clitellates make up a fraction of annelid diversity, they are of critical importance to terrestrial ecosystems, and are of evolutionary interest since they represent a terrestrialisation process within Annelida (see Chapter 4 for further discussion).

Clitellates aside, the bulk of annelid diversity is marine (Glasby and Timm, 2008). Morphological taxonomy upholds three major clades: two that possess palps, the Aciculata and Canalipalpata (sometimes known together as “Palpata”), and a third that lacks palps and other head appendages, the Scolecida (Rouse and Fauchald, 1997). Aciculates are named such on account of their aciculae: internal chaetae acting as pseudo-skeletal support for parapodia. Other morphological features, such as dorsal and ventral cirri, ventral sensory palps, compound chaetae, prostomial palps, and multiple prostomial antennae lead to strong phyletic support (Rouse and Pleijel, 2001). The orders Phyllodocida and Eunicida are upheld as members of Aciculata since they both possess jaws; a sister group relationship of the two orders seems likely (Rouse and Fauchald, 1997; Struck et al., 2011; Weigert et al., 2014; Zrzavý et al., 2009). Canalipalpates, in contrast, feature grooved, ciliated palps that are usually derived from the prostomium (Rouse and Pleijel, 2001), although they have been secondarily lost in several taxa (Orrhage, 2001).

On an intermediate perspective, legacy classification of many annelids as being either members of Errantia or Sedentaria has seen support from molecular approaches (Weigert and Bleidorn, 2016). However, these terms should not be seen as objective characteristics, since not all Sedentaria are sedentary, not are Errantia all errant (meaning mobile). Errantia and Sedentaria, though, have emerged as reciprocally monophyletic on molecular grounds, although placement of other annelids (such as sipunculans, amphinomids, chaetopterids) through molecular approaches remains uncertain (Struck et al., 2015; Weigert et al., 2014). Phylogenetic and deep-history inference on the group is hindered by this lack of resolution, since while placement is uncertain, analyses seem consistent in placing them on less derived branches of the annelidan tree (Weigert and Bleidorn, 2016) (see Figure 5.1).

Annelids are further represented by interstitial species, previously known as “Archiannelids” on suggestion the group represent an ancestral state for annelids (Hermans, 1969). The interstitium is defined as the ecology of living between grains of sand, or other small fragments of the environment. Consequently, these animals are extremely small, and often have simple morphologies, sometimes resembling larval forms (Struck et al., 2002). Phylogenetic investigation of the interstitial annelids did not support their monophyly, and instead suggested that perhaps two major groups of annelids independently adopted an interstitial ecology, either through progenesis (sexual maturation of larval forms) or anagenic miniaturisation.

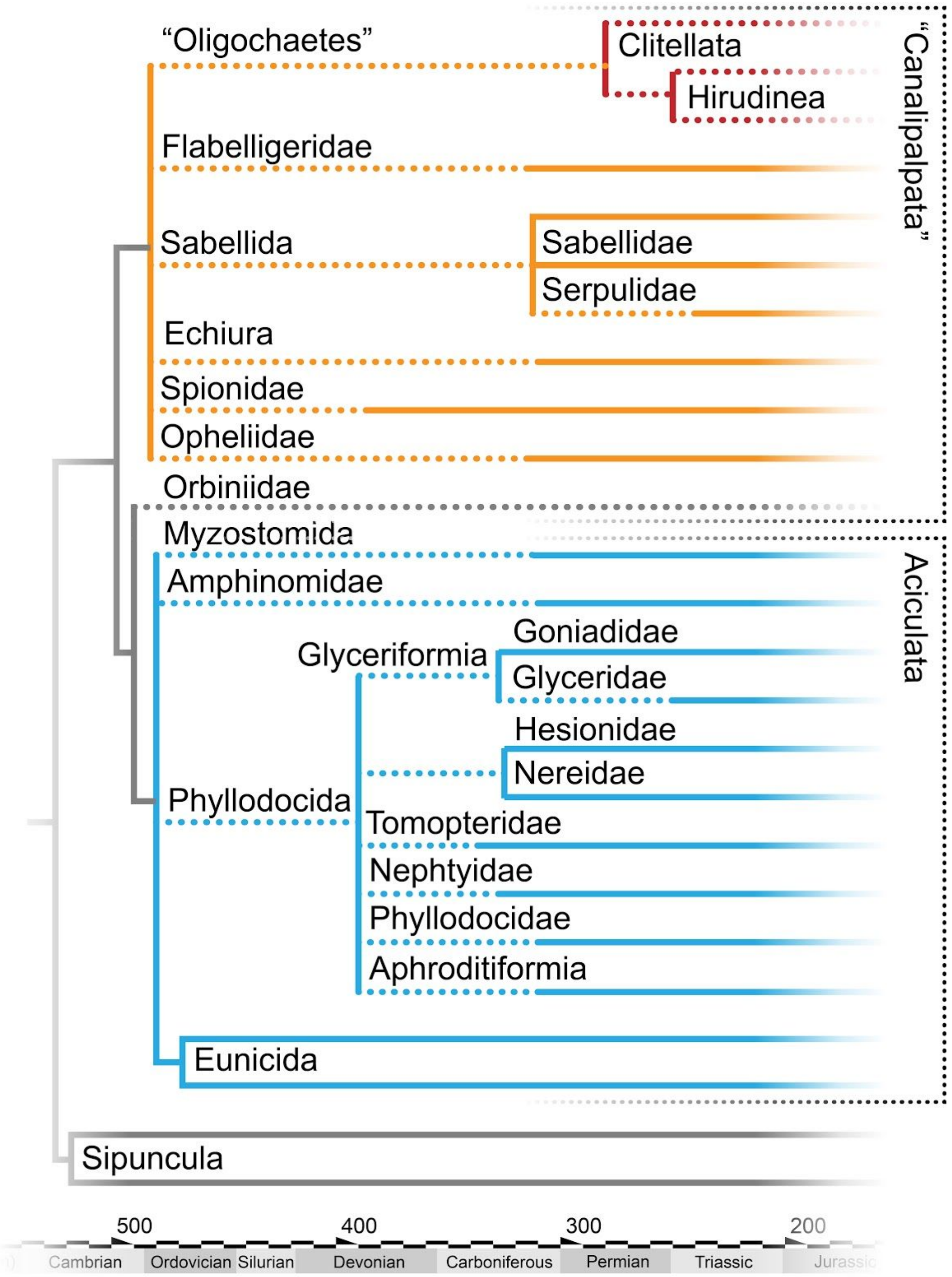


Figure 5.1. Phylogeny of annelid families. Dotted lines are inferred lineages, while solid lines represent known fossil records. Divergence times represent means of node timings, as adapted from Edgecombe et al. (2011) and Erwin et al. (2011), and are not intended to present known diversification times of groups. The project aims to resolve these polytomies.

What is highlighted, especially when this latter example of the interstitial annelids is included, is that phylogenetic studies based on molecular sequence data have not robustly recovered the higher polychaete taxa recognized by morphologists (Rousset et al., 2007). Furthermore, there is a lack of congruence between the earlier molecular studies, which employed data from a diversity of sources including protein-coding genes and nuclear genes (Rousset et al., 2007), mitochondrial genes and gene order (Mwinyi et al., 2009), miRNAs (Sperling et al., 2009), ESTs (Struck et al., 2011), and combined morphological and molecular analyses (Zrzavý et al., 2009). Despite more sophisticated studies involving expanded taxonomic sample (Struck et al., 2015; Weigert et al., 2014) annelid phylogeny has remained difficult to understand, or to uphold consistently from a variety of independent approaches.

As seen elsewhere in this thesis (see Chapter 3 in particular), incongruence is possibly due to (proximally) conflict in signal, and (ultimately) due to closely spaced divergence events. Compounding this, is that molecular evolutionary process across the annelids is under a wide range of selection pressures, perhaps especially for colonial, parasitic and interstitial species (Struck et al., 2015). Therefore, the difficulties in inferring annelid phylogeny may be both due to explosive radiation, and that extreme ecological diversity has led to molecular sequence evolution that violates the consistency that an evolutionary model expects. Here, we approach the problem through expansion of taxonomic sample to richly represent the crown-group diversity of Annelida, while assembling a range of molecular matrices based on both slow-evolving genes, and a genome-wide search for orthologous sequences. Combining 36 newly sequenced annelid species with 168 previously-sequenced species, plus an appropriately balanced outgroup of 44 molluscs, brachiopods and nemerteans, our taxonomic sample totals 248 species, the most comprehensive annelid phylogenetic dataset assembled to date. We apply strict data-refinement protocols to minimise systematic and inherent biases, and also apply models best suited to dealing with across-site heterogeneity, under a Bayesian paradigm so as to return interpretable results through posterior probabilities.

5.2 Materials and Methods

Data acquisition and assembly

36 previously-unsequenced species of annelids were acquired (see Table 5.1 for details). Specimens were stored in RNAlater, and total RNA was extracted using Trizol process. mRNA was purified using NEXTflex™ Poly(A) Beads. cDNA libraries were prepared using the NEXTflex™ Rapid Illumina Directional RNA-Seq Library Prep Kit using NEXTflex™ RNA-Seq Barcodes to allow for multiplexing.

168 already sequenced, publicly available species of annelids were recovered from NCBI GenBank (see Table 5.1 for full details). Sequence reads from 44 further species of molluscs, brachiopods and nemerteans were also recovered from GenBank, representing an extensive yet proximal outgroup to the annelid study species. This information varied in type,

being mostly transcriptomic, with a few species represented by genomic or expressed sequence tag (EST) data. In total, sequence read information from 248 species was gathered. Genomes were acquired already assembled by the respective projects generating the data. EST data was pooled into single continuous FASTA files, one for each species represented by EST data.

For transcriptomic read information, sequence assembly was consistently applied using the following protocol (see Figure 5.2 for graphical outline of data handling and matrix assembly). The RNA-seq de-novo transcriptome assembly software Trinity (Grabherr et al., 2011) was used. Read quality was assessed and edited using the standard parameters of Trimmomatic (Bolger et al., 2014), as part of the Trinity suite of software. The resulting assemblies were then cleaned of redundant repeated sequences using CD-hit (Huang et al., 2010), resulting in FASTA files of only unique sequences, effectively the transcriptome of each organism. These were then translated from nucleotide to amino acid sequences, using TransDecoder (Haas and Papanicolaou, 2016), to convert the transcriptome into the proteome of each organism. These refined sets of information are suitable for BLAST (Altschul et al., 1990) operations since the removal of redundancy accelerates computation, and for this project expression levels are not relevant.

Supermatrix compilation

To generate a range of datasets and provide alternative approaches to phylogenetic inference, we followed two compilation pathways (see Figure 5.2). Process one, denoted by blue on figures 5.2 and 5.3, and resulting in supermatrices 1A, 1B and 1C, is described in the rest of this paragraph. A set of 254 genes as BLAST targets, with these genes known to be highly conserved and slowly evolving (Philippe et al., 2011), suitable (Lozano-Fernandez et al., 2016) for investigating relationships on the timescales of the evolution of annelids; at least to the Cambrian (Parry et al., 2014). A custom Perl script (/github.com/jairly/MoSuMa_tools/blast_all.pl) was used to BLAST these 254 genes again all taxa, but only accepting results with an e-value smaller than 10^{-1}, while accepting any other hits within three orders of magnitude of the smallest e-value hit. The resulting BLAST hits were then filtered, so that genes found in less than 50% of taxa, and taxa returning less than 50% of genes were excluded (resulting in 244 taxa and 233 genes). These FASTA files, one per gene containing the top hits per species, were then aligned using MUSCLE (Edgar, 2004) under default parameters. Since gene alignments often contain poor-quality or ambiguously aligned regions (especially at the ends of gene sequences), this information was removed using Gblocks (Castresana, 2002). This was automated using a shell script, applying the parameters (-t=p, -b2=[half of the taxa count], -b3=20, -b4=2, -b5=h). Each of these aligned, trimmed matrices was then phylogenetically inferred using IQtree (Nguyen et al., 2015d), to producing individual gene-phylogenies for each matrix. Using a custom Perl script of the MoSuMa tools suite (/github.com/jairly/MoSuMa_tools/treecleaner.pl), these trees were inspected for long branches, with a branch considered to be “long” if being more than 1.8 standard deviations longer than the mean branch length across the whole tree. Sequences generating long branches were then deleted from the gene matrix producing that tree. Trees were also assessed by eye, and duplicate sequences (due to passing e-value criteria earlier in the process) were removed, judged by placement (nearly all were in sisterhood on the phylogeny, thus the choice was not relevant, or it was producing a long branch and was removed automatically). These gene matrices, with most long-branch-generating sequence information removed, were then concatenated using SequenceMatrix (Vaidya et al., 2011), resulting in a supermatrix of 41,293 amino acids across 233 taxa. Preliminary phylogenetic inference was carried out on this set to identify rogue taxa (those with significant instability in placement, or implausibility in placement suggesting mislabelling or other upstream errors). These taxa were removed, refining the matrix to 41,293 amino acids across 213 taxa, supermatrix 1A. Two further datasets were made, on account of known issues for parasites in phylogenies, and also with the known difficulties in reliable phylogenetic reconstruction of interstitial annelids (Struck et al., 2015). Supermatrix 1B (188 taxa) is the same as 1A but with parasites removed, and supermatrix 1C (159 taxa) is the same as 1A but with both parasites and interstitial taxa removed.

A full orthology reciprocal BLAST method was applied to find all orthologous groups for all taxa. This process, and the resultant supermatrices, is denoted in red on figures 5.2 and 5.3. The computational burden of a comprehensive reciprocal BLAST search of 244 taxa against each other was not feasible. Also, full taxonomic reciprocal BLASTing is not an appropriate use of computational resources, since the genes recovered when BLASTing between just and handful of taxa and the whole taxonomic sample will be very similar. Considering this, we chose ten species from both the annelid group and the outgroup to reciprocally BLAST, on account of crown-group coverage, and the quality of the data. The species selected were *Crassostrea virginica, Salmacina* sp*., Terebratalia transversa, Pareurythoe* sp.*, Phascolopsis gouldii, Nereimyra* sp.*, Pherusa flabellata, Owenia fusiformis* and *Phyllochaetopterus* sp.. The software OMA (Roth et al., 2008) was applied, recovering 82,355 orthologous groups for the ten species. This set was then refined by excluding all genes with less than 50% taxonomic coverage, resulting in a set of 3,880 genes. A single instance of each of these genes was used as BLAST target for the remaining taxa that were not part of the OMA reciprocal search. A custom Perl script (/github.com/jairly/MoSuMa_tools/blast_all.pl) was used to automate BLASTing, taking only the top hit of each search. 3,880 gene matrices were then produced by aligning each orthologous gene using MUSCLE. Each of these aligned gene matrices was then phylogenetically inferred using IQtree to producing individual gene-phylogenies for each matrix. Using a custom Perl script of the MoSuMa tools suite (/github.com/jairly/MoSuMa_tools/treecleaner.pl), these trees were inspected for long branches, with a branch considered to be “long” if being more than 1.8 standard deviations longer than the mean branch length across the whole tree. Sequences generating long branches were then deleted from the gene matrix producing that tree. The resulting gene matrices were concatenated using FASconCAT (Kück and Meusemann, 2010), resulting in a supermatrix of 1,855,381 amino acids across 244 taxa. Four smaller supermatrices were then produced from this extensive matrix: one matrix with 100% completeness, one matrix approximately the same size as matrix 1B, one matrix twice the size of matrix 1B, and one ten times the size of 1B (see Figure 5.2), named matrices 2A, 2B, 2C and 2D respectively. (On preliminary analysis, matrix 2D was seen as being computationally intractable due to size, and was thus rejected from all downstream analysis.) These matrices were made by using increasingly stringent Gblocks parameters, in order to retain conserved sequence blocks. Matrix 1B (188 taxa) was used as a guide taxon size due to parasite sequence data being significantly difficult to deal with in a phylogenetic framework, and because other data was considered valuable to the investigation. Thus, matrix 1B was considered a suitable intermediate matrix size, and most appropriate for approaching the evolutionary questions at hand.

Species	Family	Phylum	# of 243 gene hits	Accession
Allodera hylae	Clitellata	Annelida	217	Newly sequenced
Allolobophora chlorotica	Clitellata	Annelida	216	SRR1324778
Alvinella pompejana	Terebelliformia	Annelida	219	EST on NCBI
Amynthas koreanus	Clitellata	Annelida	63	EST on NCBI
Antillesoma antillarum	Phascolosomatidae	Sipuncula	194	SRR1646260
Apharyngtus punicus	Dinophilidae	Annelida	203	SRR2014574
Aplysia californica	Gastropoda	Mollusca	214	SRR016568
Aporrectodea icterica	Clitellata	Annelida	213	SRR1324787
Arabella sp	Oenonidae	Annelida	206	SRR2040141
Arenicola marina	Terebelliformia	Annelida	193	SRR2005653
Arenicola marina	Terebelliformia	Annelida	68	EST on NCBI
Aricidea	Paraonidae	Annelida	221	SRR1219647
Aspidosiphon parvulus	Aspidosiphonidae	Sipuncula	173	SRR1646391
Astarte	Bivalvia	Mollusca	193	SRX687759
Baseodiscus unicolor	Anopla	Nemertea	218	SRR1505175
Boccardia proboscidea	Spionidae	Annelida	220	SRR2057014
Bonellia	Echiura	Annelida	215	SRR016568
Bonellia viridis	Echiura	Annelida	216	SRR2017645
Brada	Cirratuliformia	Annelida	224	Newly sequenced
Branchiomma	Sabellidae	Annelida	219	Newly sequenced
Capitella telata	Capitellidae	Annelida	232	genome.jgi.doe.gov
Cephalothrix hongkongiensis	Palaeonemertea	Nemertea	214	SRR618505
Cephalothrix linearis	Palaeonemertea	Nemertea	217	SRR1273789
Cerebratulus marginatus	Anopla	Nemertea	216	SRR618507
Cerebratulus sp	Anopla	Nemertea	175	SRR1797867
Chaetopterus	Chaetopteridae	Annelida	125	SRX755856
Chaetopterus variopedatus	Chaetopteridae	Annelida	47	SRR1219647
Cirratulus	Cirratuliformia	Annelida	221	EST on NCBI
Cirratulus specabilis	Cirratuliformia	Annelida	178	SRX2848072
Claudrilus	Clitellata	Annelida	215	SRR2017810
Crassostrea gigas	Bivalvia	Mollusca	215	SRR1925760
Crassostrea virginica	Bivalvia	Mollusca	222	SRX3467356
Criodrilus	Clitellata	Annelida	177	Newly sequenced
Dinophilus taeniatus	Dinophilidae	Annelida	213	SRR2018886
Dinophilus gyrociliatus	Dinophilidae	Annelida	189	SRR2040285
Diopatra cuprea	Eulabidognatha	Annelida	221	SRR2131612

Diplocardia	Clitellata	Annelida	220	Newly sequenced
Dorivillea sp	Dorvilleidae	Annelida	219	SRR2040378
Drawida nelamburensis	Clitellata	Annelida	56	Newly sequenced
Drawida sp	Clitellata	Annelida	224	Newly sequenced
Echiniscus	Tardigrada	Arthropoda	221	SRR1224604
Eisenia andrei	Clitellata	Annelida	171	SRR1224604
Eisenia fetida	Clitellata	Annelida	218	DRR023799
Eisenoides	Clitellata	Annelida	213	Newly sequenced
Enchytraeus crypticus	Clitellata	Annelida	205	SRR2014693
Endomyzostoma scotia	Myzostomidae	Annelida	223	SRR2005728
Ennucula	Bivalvia	Mollusca	228	SRR331123
Eudrilus	Clitellata	Annelida	79	Newly sequenced
Eulalia clavigera	Phyllodocida	Annelida	185	EST on NCBI
Eunice	Eulabidognatha	Annelida	214	SRX1038837
Eunice pennata	Eulabidognatha	Annelida	183	SRR2040479
Eunice torquata	Eulabidognatha	Annelida	201	SRR2005375
Euspira	Gastropoda	Mollusca	180	DRX031588
Exallopus sp	Dorvilleidae	Annelida	177	SRR2040480
Fauveliopsis	Cirratuliformia	Annelida	214	SRR2017643
Ficopomatus	Serpulidae	Annelida	187	Newly sequenced
Flabelliderma ockeri	Cirratuliformia	Annelida	80	SRR2005668
Flabelligera affinis	Cirratuliformia	Annelida	228	EST on NCBI
Flabelligera mundata	Cirratuliformia	Annelida	230	SRR3574613
Gadila	Scaphopoda	Mollusca	215	SRR2040285
Galeolaria	Serpulidae	Annelida	222	Newly sequenced
Gattyana	Phyllodocida	Annelida	223	SRR016568
Geogeniabenhami	Clitellata	Annelida	220	Newly sequenced
Glossoscolex	Clitellata	Annelida	223	Newly sequenced
Glottidia pyramidata	Linguliformia	Brachiopoda	201	SRR1611555
Glycera dibranchiata	Phyllodocida	Annelida	157	SRR1611557
Glycera tridactyla	Phyllodocida	Annelida	221	SRR1237833
Golfingia vulgaris	Golfingiidae	Sipuncula	224	SRR1797875
Goniada	Phyllodocida	Annelida	199	Newly sequenced
Granata	Gastropoda	Mollusca	77	EST on NCBI
Haementeria depressa	Clitellata	Annelida	151	EST on NCBI
Haliotis	Gastropoda	Mollusca	220	SRR2131255
Harmothoe extenuata	Phyllodocida	Annelida	223	SRR1237766

Harmothoe	Phyllodocida	Annelida	224	SRR2005364
Helobdella robusta	Clitellata	Annelida	228	genome.jgi.doe.gov/
Hemithiris psittacea	Rhynchonellida	Brachiopoda	131	SRR1611556
Hermodice carunculata	Amphinomida	Annelida	208	SRR651044
Hirudo medicinalis	Clitellata	Annelida	220	SRX256698
Hormogaster elisae	Clitellata	Annelida	220	SRR786599
Hormogaster samnitica	Clitellata	Annelida	226	SRR618446
Hubrechtella ijimai	Palaeonemertea	Nemertea	221	SRR1505100
Hydatina	Gastropoda	Mollusca	215	SRR1505113
Hydroides sanctaecrucis	Serpulidae	Mollusca	225	SRX1509335
Laevipelina hyalina	Monoplacophora	Mollusca	216	EST on NCBI
Lamellibranchia luymesi	Bivalvia	Mollusca	227	SRX1782977
Laqueus californicus	Rhynchonellida	Brachiopoda	221	SRR1611557
Leptochiton	Aculifera	Mollusca	218	SRR1611558
Lindrilus	Protodrilidae	Annelida	220	SRR1224604
Liothyrella uva	Rhynchonellida	Brachiopoda	220	SRR3205210
Lobatocerum sp	Incertae sedis	Annelida?	199	SRR2131397
Lumbricus rubella	Clitellata	Annelida	179	EST on NCBI
Lumbricus terrestris	Clitellata	Annelida	218	SRX316160
Lumbrinereis zonata	Lumbrineridae	Annelida	63	EST on NCBI
Lysarete	Lumbrineridae	Annelida	212	Newly sequenced
Macandrevia cranium	Rhynchonellida	Brachiopoda	69	SRR1611130
Macrochaeta clavicornis	Cirratuliformia	Annelida	123	SRR1221445
Magelona berkeleyi	Magelonidae	Annelida	149	SRR1257638
Magelona	Magelonidae	Annelida	219	SRR1224604
Magelona johnstoni	Magelonidae	Annelida	215	SRR1222290
Magelona pitelkai	Magelonidae	Annelida	216	SRR2015609
Malacobdella grossa	Enopla	Nemertea	219	SRR1611560
Malacoceros fulginosus	Serpulidae	Annelida	92	EST on NCBI
Marphysa bellii	Eulabidognatha	Annelida	198	SRR1232833
Megadrilus nsp	Protodrilidae	Annelida	224	SRR2020581
Megalomma vesiculosum	Sabellidae	Annelida	210	SRR1231830
Meganerilla	Nerilidae	Annelida	215	SRR1222288
Meiodrilus	Protodrilidae	Annelida	210	SRR2005873
Mesochaetopterus	Chaetopteridae	Annelida	163	SRR1925760
Mesochaetopterus minutus	Chaetopteridae	Annelida	218	SRR2005708
Mesonerilla fagei	Nerilidae	Annelida	208	SRR2014581

Metavermillia acanthaphora	Serpulidae	Annelida	221	Newly sequenced
Microhedyle	Gastropoda	Mollusca	177	EST on NCBI
Monodonta	Gastropoda	Mollusca	220	SRR1505119
Myochama	Bivalvia	Mollusca	199	SRR1224604
Mytilus californianus	Bivalvia	Mollusca	218	SRX565220
Myzostoma cirriferum	Myzostomidae	Annelida	117	EST on NCBI
Myzostoma cirriferum	Myzostomidae	Annelida	208	SRR1237872
Myzostoma seymourcollegium	Myzostomidae	Annelida	213	SRR2005822
Myzostomida sp	Myzostomidae	Annelida	203	SRR2008168
Nainereis	Orbiniidae	Annelida	216	SRR016568
Naineris dendritica	Orbiniidae	Annelida	217	SRR2017044
Neosabellaria cementarium	Sabellariidae	Annelida	193	SRR2017810
Neotrigonia	Bivalvia	Mollusca	215	SRR2005708
Nephasoma pellucidum	Golfingiidae	Sipuncula	214	SRR1646439
Nephtys caeca	Phyllodocida	Annelida	217	SRR1232795
Nereimyra	Phyllodocida	Annelida	221	Newly sequenced
NewMyzostome	Myzostomidae	Annelida	205	SRR016568
Ninoe nigrens	Eulabidognatha	Annelida	211	SRR2040484
Novocrania anomala	Craniiformia	Brachiopoda	151	SRR1611564
Ocnerodrilidae	Clitellata	Annelida	218	Newly sequenced
Ophelina	Opheliidae	Annelida	224	Newly sequenced
Ophicardelus	Gastropoda	Mollusca	217	SRR1611132
Ophryotrocha globopalpata	Dorvilleidae	Annelida	197	SRR1926090
Osedax frankpressi	Siboglinidae	Annelida	159	SRR2005641
Osedax mucofloris	Siboglinidae	Annelida	217	SRR1232833
Osedax rubiplumus	Siboglinidae	Annelida	216	SRR1611132
Owenia fusiformis	Oweniidae	Annelida	223	SRR1222288
Owenia	Oweniidae	Annelida	215	SRR2005873
Oxynoe	Gastropoda	Mollusca	215	SRX755857
Paralvinella sulfincola	Terebelliformia	Annelida	192	SRR1646442
Paramphinome jeffreysii	Amphinomida	Annelida	213	SRR1257731
Paranemertes peregrina	Enopla	Nemertea	223	SRR1611562
Parborlasia corrugatus	Anopla	Nemertea	28	SRR1611132
Pareurythoe	Amphinomida	Annelida	171	SRR1926090
Pectinaria koreni	Terebelliformia	Annelida	223	SRR1325083
Perionyx excavatus	Clitellata	Annelida	79	EST on NCBI

Perotrochus	Gastropoda	Mollusca	112	SRR1646442
Phallomedusa	Gastropoda	Mollusca	222	SRR1222216
Pharyngocirrus tridentiger	Protodrilida	Annelida	218	SRR2016714
Phascolion cryptum	Phascolionidae	Sipuncula	161	SRR1646440
Phascolopsis gouldii	Sipunculidae	Sipuncula	210	SRR1654498
Phascolopsis	Sipunculidae	Sipuncula	203	SRX755857
Phascolosoma granulatum	Phascolosomatidae	Sipuncula	114	SRR1231565
Phascolosoma perlucens	Phascolosomatidae	Sipuncula	171	SRR1646442
Pherusa flabellata	Cirratuliformia	Annelida	221	Newly sequenced
Philine	Gastropoda	Mollusca	202	SRR2124792
Phoronis australis	Phoronidae	Phoronida	214	SRR2018856
Phoronis psammophila	Phoronidae	Phoronida	223	SRR1611565
Phoronis vancouverensis	Phoronidae	Phoronida	224	SRR1611566
Phoronopsis harmeri	Phoronidae	Phoronida	224	SRR2131255
Phyllochaetopterus sp	Chaetopterida	Annelida	218	SRR1257898
Phyllodoce	Phyllodocida	Annelida	184	SRR2016923
Phyllodoce medipapillata	Phyllodocida	Annelida	222	SRR2005653
Phylo foetida	Orbiniidae	Annelida	204	SRR1222216
Platynereis dumerilii	Phyllodocida	Annelida	210	SRR1324778
Poecilobdella javanica	Clitellata	Annelida	221	SRR5429897
Polygordius	Polygordiidae	Annelida	213	SRR1231565
Polygordius lacteus	Polygordiidae	Annelida	226	SRR2014676
Polygordius nsp	Polygordiidae	Annelida	213	SRR2005365
Polygordius sp1	Polygordiidae	Annelida	225	SRR2124758
Polygordius sp2	Polygordiidae	Annelida	221	SRR2124789
Pomacea	Gastropoda	Mollusca	165	SRR2005708
Pomatoceros lamarckii	Serpulidae	Annelida	212	SRR516531
Praxiellaaffinis	Terebelliformia	Annelida	219	Newly sequenced
Praxiella p	Terebelliformia	Annelida	223	Newly sequenced
Prionospio	Spionidae	Annelida	199	SRR2017831
Prionospio	Spionidae	Annelida	223	Newly sequenced
Prionospio sp	Spionidae	Annelida	227	Newly sequenced
Pristina leidyi	Clitellata	Annelida	185	SRR125360
Protodorvillea kefersteini	Dorvilleidae	Annelida	195	SRR2014681
Protodorvillea	Dorvilleidae	Annelida	223	Newly sequenced
Protodriloides chaetifer	Protodrilidae	Annelida	218	SRR2016233
Protodriloides	Protodrilidae	Annelida	215	SRX1023293

Protodriloides symbioticus	Protodrilidae	Annelida	217	SRR2124792
Protodrilus adhaerens	Protodrilidae	Annelida	208	SRR2014684
Protodrilus	Protodrilidae	Annelida	226	SRR2564755
Protula	Serpulidae	Annelida	216	Newly sequenced
Pseudopolydora vexillosa	Spionidae	Annelda	168	SRR125360
Rhinoddriluspriollii	Clitellata	Annelida	220	Newly sequenced
Ridgeia piscesae	Siboglinidae	Annelida	199	JGI database§
Riftia pachyptila	Siboglinidae	Annelida	208	SRR346549
Rubyspira	Gastropoda	Mollusca	197	SRR1505136
Sabella pavonina	Sabellidae	Annelida	214	SRR2005708
Sabellaria alveolata	Sabellariidae	Annelida	128	SRR1232634
Sabellestarte sp.	Sabellidae	Annelida	218	Newly sequenced
Saccocirrus burchelli	Protodrilidae	Annelida	220	SRR2014689
Saccocirrus papillocerchs	Protodrilidae	Annelida	226	SRS931954
Saccocirrus	Protodrilidae	Annelida	221	SRR125360
Salmacina	Serpulidae	Annelida	213	Newly sequenced
Scalibregma	Scalibregmatidae	Annelida	224	Newly sequenced
Sclerolinum fiordicum	Siboglinidae	Annelida	214	SRR1646442
Scolelepis squamata	Spionidae	Annelida	222	SRR1222145
Scoloplos armiger	Orbiniidae	Annelida	86	SRR1221444
Serpula	Serpulidae	Annelida	213	Newly sequenced
Siboglinum ekmani	Siboglinidae	Annelida	173	SRR2017643
Siphonosoma cumanense	Sipunculidae	Sipuncula	121	SRR1646441
Sipunculus	Sipunculidae	Sipuncula	197	SRR619011
Sipunculus nudus	Sipunculidae	Sipuncula	89	EST on NCBI
Solemya	Bivalvia	Mollusca	221	SRX2354100
Sparganophilus	Clitellata	Annelida	219	Newly sequenced
Spinther	Spintheridae	Annelida	222	SRR2005641
Spinther sp	Spintheridae	Annelida	218	SRR016568
Spiochaetopterus sp	Chaetopterida	Annelida	215	SRR1224605
Spirobrachia sp	Siboglinidae	Annelida	219	SRR1224604
Spirobranchus	Serpulidae	Annelida	214	Newly sequenced
Sternapsis affinis	Cirratuliformia	Annelida	218	SRR2017800
Sternapsis sp	Cirratuliformia	Annelida	213	SRR2005708
Streblospio benedicti	Spionidae	Annelida	186	SRR626652
Stygiocapitella	Stygocapitellidae	Annelida	207	SRR0226109
Syllis sp	Phyllodocida	Annelida	96	SRR1224604

Terebratalia transversa	Rhynchonellida	Brachiopoda	221	SRR2564755
Thelepus nsp	Terebelliformia	Annelida	221	SRR2017640
Thoracophelia mucronata	Opheliidae	Annelida	218	SRR2017631
Tomopteris helgolandica	Phyllodocida	Annelida	212	SRR1237767
Trilobodrilus axi	Dinophilidae	Annelida	216	SRR2014693
Tubifex tubifex	Clitellata	Annelida	188	EST on NCBI
Tubulanus polymorphus	Palaeonemertea	Nemertea	221	SRR1273849
Tylodina	Gastropoda	Mollusca	217	SRR0226110
Urechis unicinctus	Echiura	Annelida	217	SRR2564755
Urosalpinx	Gastropoda	Mollusca	221	SRR1224604
Vermiliopsis	Serpulidae	Annelida	220	Newly sequenced

Table 5.1. Species used in this study, with accession numbers and number of genes recovered from datasets built from pre-selected slow-evolving gene dataset (onward to make matrices 2A, 2B and 2C, see Figures 5.2 and 5.3).

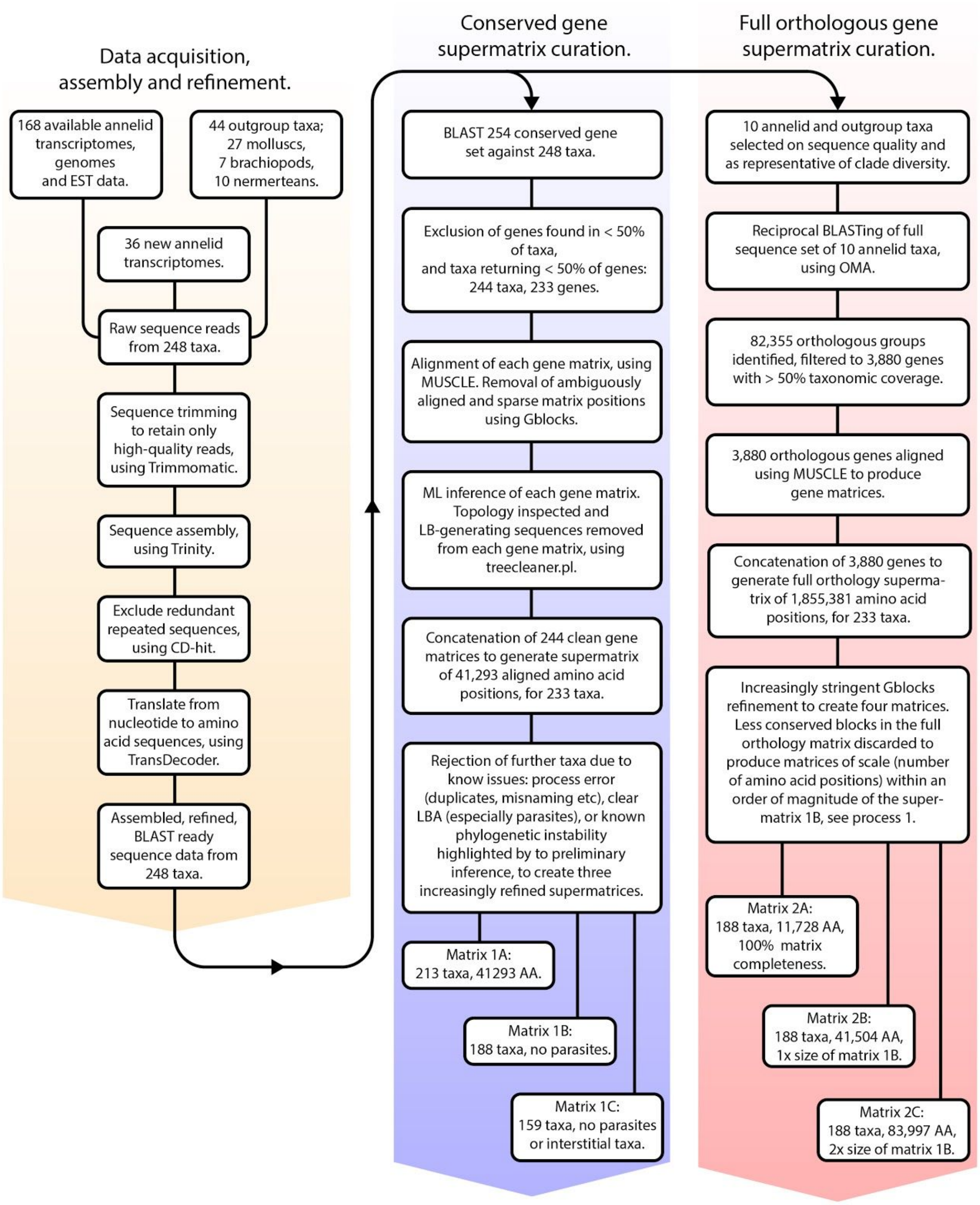

Data acquisition, assembly and refinement.
168 available annelid transcriptomes, genomes and EST data.
44 outgroup taxa; 27 molluscs, 7 brachiopods, 10 nermerteans.
36 new annelid transcriptomes.
Raw sequence reads from 248 taxa.
Sequence trimming to retain only high-quality reads, using Trimmomatic.
Sequence assembly, using Trinity.
Exclude redundant repeated sequences, using CD-hit.
Translate from nucleotide to amino acid sequences, using TransDecoder.
Assembled, refined, BLAST ready sequence data from 248 taxa.
Conserved gene supermatrix curation.
BLAST 254 conserved gene set against 248 taxa.
Exclusion of genes found in < 50% of taxa, and taxa returning < 50% of genes: 244 taxa, 233 genes.
Alignment of each gene matrix, using MUSCLE. Removal of ambiguously aligned and sparse matrix positions using Gblocks.
ML inference of each gene matrix. Topology inspected and LB-generating sequences removed from each gene matrix, using treecleaner.pl.
Concatenation of 244 clean gene matrices to generate supermatrix of 41,293 aligned amino acid positions, for 233 taxa.
Rejection of further taxa due to know issues: process error (duplicates, misnaming etc), clear LBA (especially parasites), or known phylogenetic instability highlighted by to preliminary inference, to create three increasingly refined supermatrices.
Matrix 1A: 213 taxa, 41293 AA.
Matrix 1B: 188 taxa, no parasites.
Matrix 1C: 159 taxa, no parasites or interstitial taxa.
Full orthologous gene supermatrix curation.
10 annelid and outgroup taxa selected on sequence quality and as representative of clade diversity.
Reciprocal BLASTing of full sequence set of 10 annelid taxa, using OMA.
82,355 orthologous groups identified, filtered to 3,880 genes with > 50% taxonomic coverage.
3,880 orthologous genes aligned using MUSCLE to produce gene matrices.
Concatenation of 3,880 genes to generate full orthology supermatrix of 1,855,381 amino acid positions, for 233 taxa.
Increasingly stringent Gblocks refinement to create four matrices. Less conserved blocks in the full orthology matrix discarded to produce matrices of scale (number of amino acid positions) within an order of magnitude of the supermatrix 1B, see process 1.
Matrix 2A: 188 taxa, 11,728 AA, 100% matrix completeness.
Matrix 2B: 188 taxa, 41,504 AA, 1x size of matrix 1B.
Matrix 2C: 188 taxa, 83,997 AA, 2x size of matrix 1B.

Figure 5.2. Schematic flow-chart for phylogenomic matrix compilation and refinement. Yellow represents initial data acquisition and BLAST operations. Blue represents generation of supermatrices based on pre-selected gene sample. Red represents process of supermatrices based on orthology searches of all sequence data, and the steps to refine this to make datasets suitable for phylogenetic inference.

Phylogenetic inference

Three primary inference methods were applied to the resulting supermatrices (see figures 5.1 and 5.2). Firstly, the site-heterogenous infinite-mixture model (CAT) was applied, guided by exchangeability frequencies inferred from the dataset (GTR) (Lartillot et al., 2013). In order to test the fit of the model, cross-validation comparing GTR against CAT-GTR was carried out. Ten replicate learning and test datasets were assembled and analysed. Statistical summary of these showed that CAT-GTR outperformed GTR in all ten replicates (Bayes factor 48.8 ± 21.0, positive result in support of CAT-GTR over GTR alone). Due to computational constraints, cross-validation was only carried out on matrix 1B, and the assumption was made (especially on past experience on a wide variety of datasets as CAT-GTR being best fit (Lozano-Fernandez et al., 2016; Tanner et al., 2017)) that CAT-GTR will nearly always fit better than GTR alone, for this type of dataset. For each matrix analysis, two independent chains were run. Under the PhyloBayes package, two methods were employed to assess convergence. Firstly *bpcomp* was used to compare the bipartitions of the trees visited by each independent chain, after a burn-in of 25% of each chain. For the matrixes recoded to Dayhoff-6 amino acid properties, all datasets converged with *maxdiff* < 0.2, and *tracecomp* statistics showing effective sample sizes > 50, and relative differences < 1. For the non-recoded matrices, convergence was not achieved: *bpcomp* remained at 1, indicating that the independent chains had converged to non-overlapping local optima for parameters, and therefore could not return consistent topology. However, *tracecomp* statistics returned sample sizes > 50, log likelihood scores dropped below 1, while other statistics, notably *statent* remained > 1. Convergence is discussed in more detail in the results and discussion sections of this chapter.

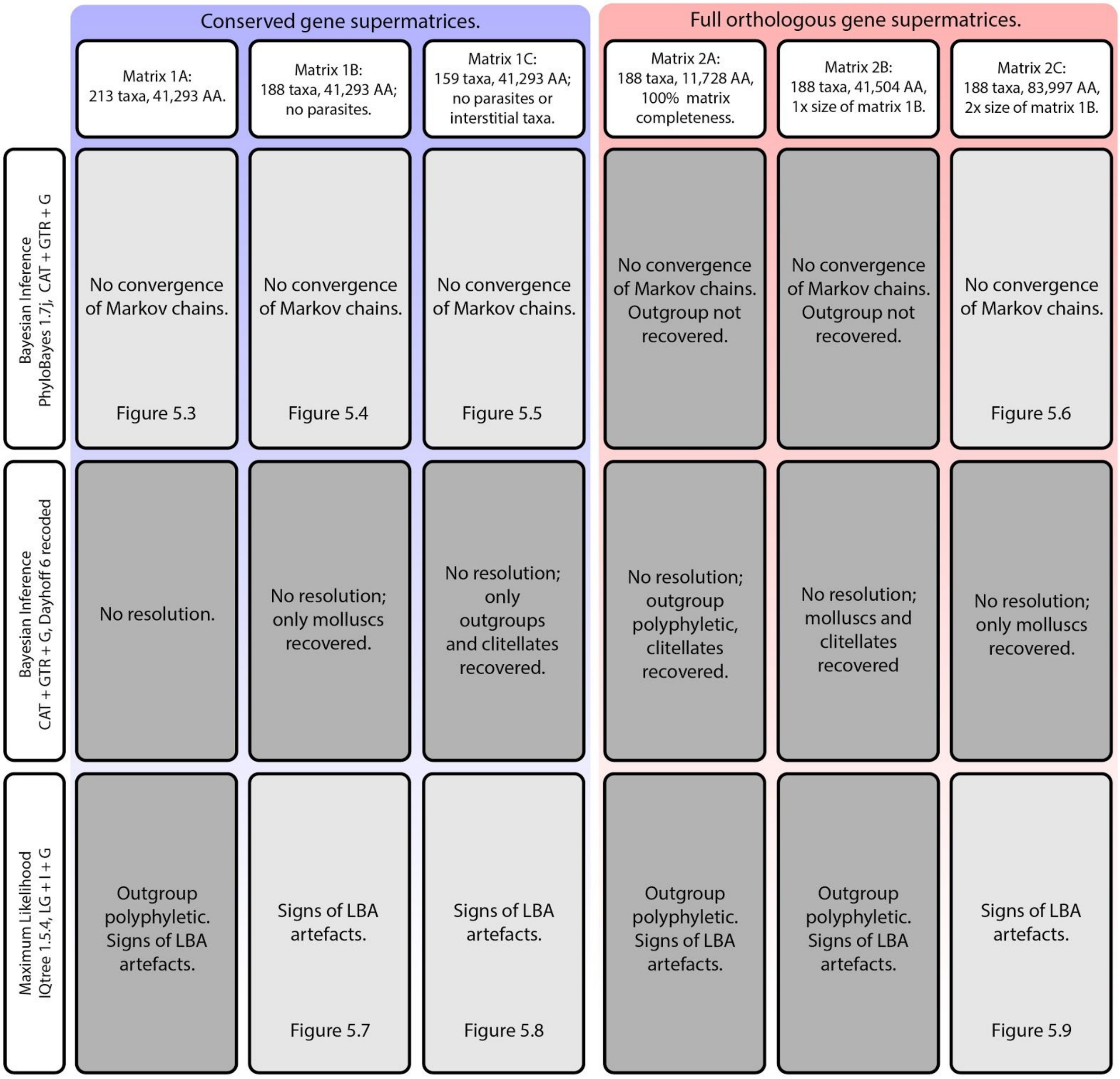
The second approach was to recode each of the seven matrices on account of the physiochemical properties of amino acid postions, using six categories: dayhoff6. This approach intends to lower the resolution of the information in the matrix (the alphabet of possible states reduces from 20 to 6), but as such mitigate compositional and branch-attraction artefacts (Feuda et al., 2017). These recoded matrices were then analysed with the same model as above: CAT-GTR, as applied in the PhyloBayes softwares.

Thirdly, Maximum Likelihood (ML) was applied to investigate node support and alternative hypotheses, in light of the Bayesian methods. A topology was inferred from the seven experimental matrices, using IQtree (Nguyen et al., 2015d). As part of the IQtree software suite, model testing was carried out using ModelFinder (Kalyaanamoorthy et al., 2017). For all matrices except 2D, the best fitting model was LG + I + G4: exchangeability matrix of Le and Gascuel (Le and Gascuel, 2008), with a proportion of invariant sites (I), and a gamma distribution of rates under four categories (G4). Matrix 2D had best fit with the model LG + I + F + G; the same as above, except with amino acid frequencies being normalised to account for differences from the frequencies expected by LG. For all analyses, 1,000 bootstrap replicates matrices were inferred to generate node support statistics on the majority-rule consensus tree for each matrix.

5.3 Results

Convergence of Bayesian MCMC chains was assessed through bipartition comparison and chain trace comparison, using *bpcomp* and *tracecomp* respectively, as part of the PhyloBayes package. For all analysis, maxdiff statistics remained equalling 1, indicating that convergence was not reached, and no consistent topology was returned under any conditions. Trace statistics showed that multiple parameters did not approach their stationary distribution, again indicating poor performance of MCMC chains. Overall, no topology can be upheld as being credible (see Figure 5.2). It is notable that for two of the Bayesian analyses, even the outgroup was not recovered, which indicates that the datasets or inference methodologies have very little resolution power.

For the ML analyses, some failed to recover outgroups meaningfully, and these analyses were rejected on the grounds that if the outgroup cannot be recovered, then ingroup relationships are certainly not credibly recovered. For all topologies, LBA seems a dominant characteristic of the phylogenies, with clustering of species at buried nodes, that otherwise have no support from other analyses, or are in conflict with published phylogenies (Struck et al., 2015; Weigert et al., 2014).



	Conserved gene supermatrices.			Full orthologous gene supermatrices.		
	Matrix 1A: 213 taxa, 41,293 AA.	Matrix 1B: 188 taxa, 41,293 AA; no parasites.	Matrix 1C: 159 taxa, 41,293 AA; no parasites or interstitial taxa.	Matrix 2A: 188 taxa, 11,728 AA, 100% matrix completeness.	Matrix 2B: 188 taxa, 41,504 AA, 1x size of matrix 1B.	Matrix 2C: 188 taxa, 83,997 AA, 2x size of matrix 1B.
Bayesian Inference PhyloBayes 1.7j, CAT + GTR + G	No convergence of Markov chains. Figure 5.3	No convergence of Markov chains. Figure 5.4	No convergence of Markov chains. Figure 5.5	No convergence of Markov chains. Outgroup not recovered.	No convergence of Markov chains. Outgroup not recovered.	No convergence of Markov chains. Figure 5.6
Bayesian Inference CAT + GTR + G, Dayhoff 6 recoded	No resolution.	No resolution; only molluscs recovered.	No resolution; only outgroups and clitellates recovered.	No resolution; outgroup polyphyletic, clitellates recovered.	No resolution; molluscs and clitellates recovered	No resolution; only molluscs recovered.
Maximum Likelihood IQtree 1.5.4, LG + I + G	Outgroup polyphyletic. Signs of LBA artefacts.	Signs of LBA artefacts. Figure 5.7	Signs of LBA artefacts. Figure 5.8	Outgroup polyphyletic. Signs of LBA artefacts.	Outgroup polyphyletic. Signs of LBA artefacts.	Signs of LBA artefacts. Figure 5.9

Figure 5.3. Matrices made and inference methodologies applied. Grey boxes give generalised result of the inference on that dataset. Dark grey boxes represent results that were rejected due to major failures in inference. Light grey boxes are inferences that, while still invalid, have been presented in Figures 5.4 - 5.10.

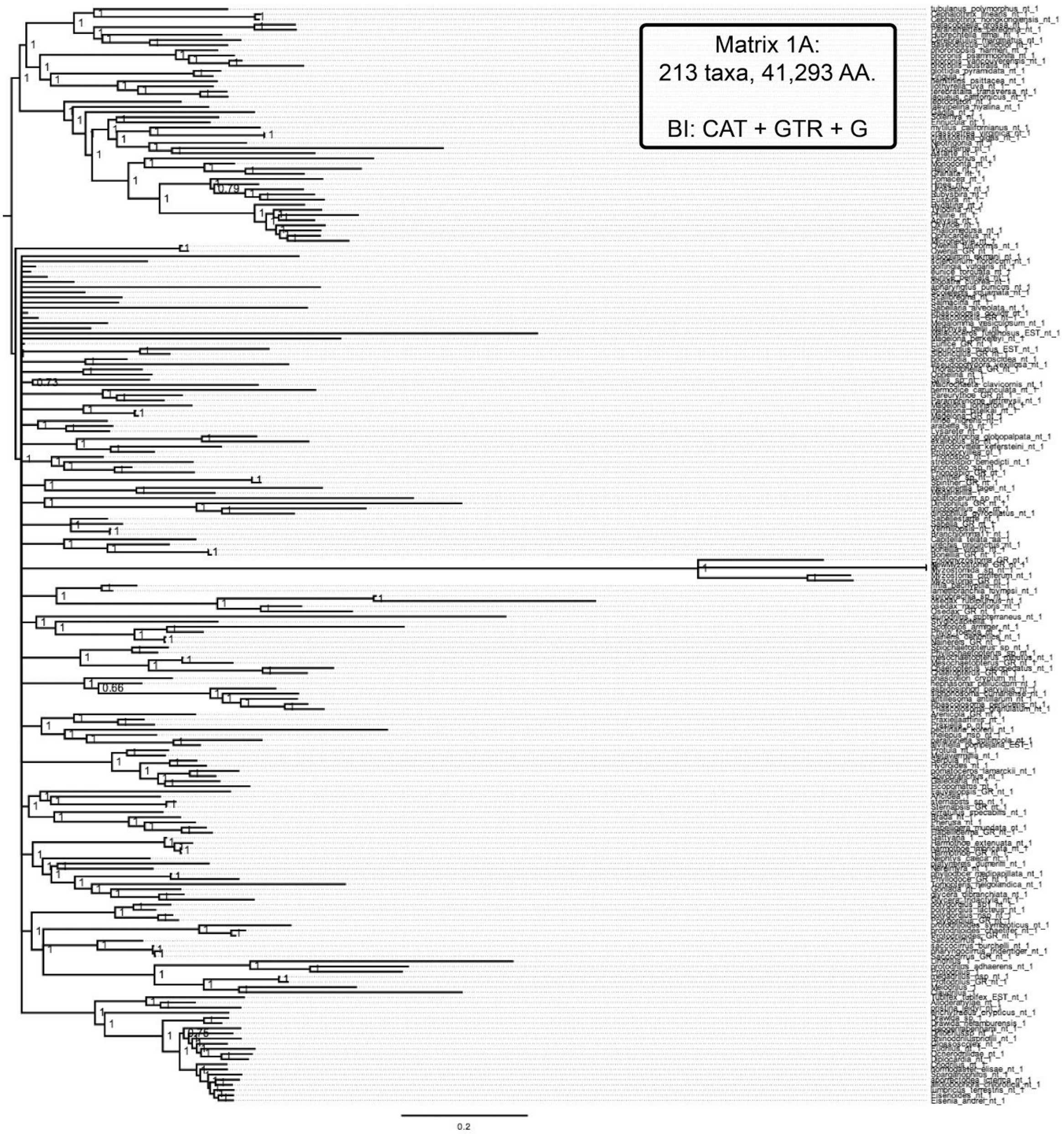


Figure 5.4. Unconverged topology from Bayesian inference of Matrix 1A. 213 annelid and outgroup taxa, 41,293 AA. Topology inferred through PhyloBayes version 1.7j, applying Cat + GTR + G.

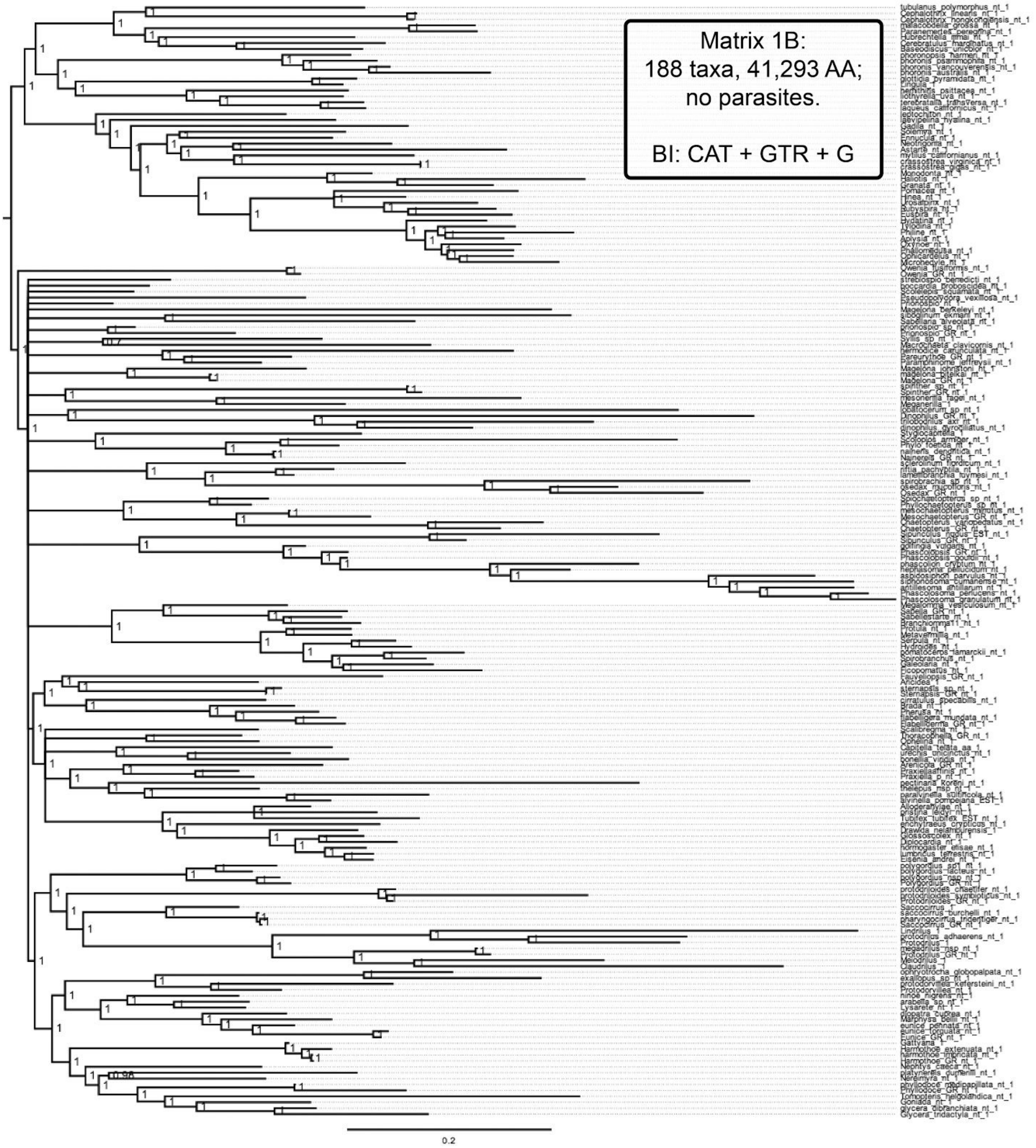


Figure 5.5. Unconverged topology from Bayesian inference of Matrix 1B. 213 annelid and outgroup taxa, 41,293 AA. Topology inferred through PhyloBayes version 1.7j, applying CAT + GTR + G. 188 annelid and outgroup taxa, 41,293 AA. Topology inferred through PhyloBayes version 1.7j, applying Cat + GTR + G.

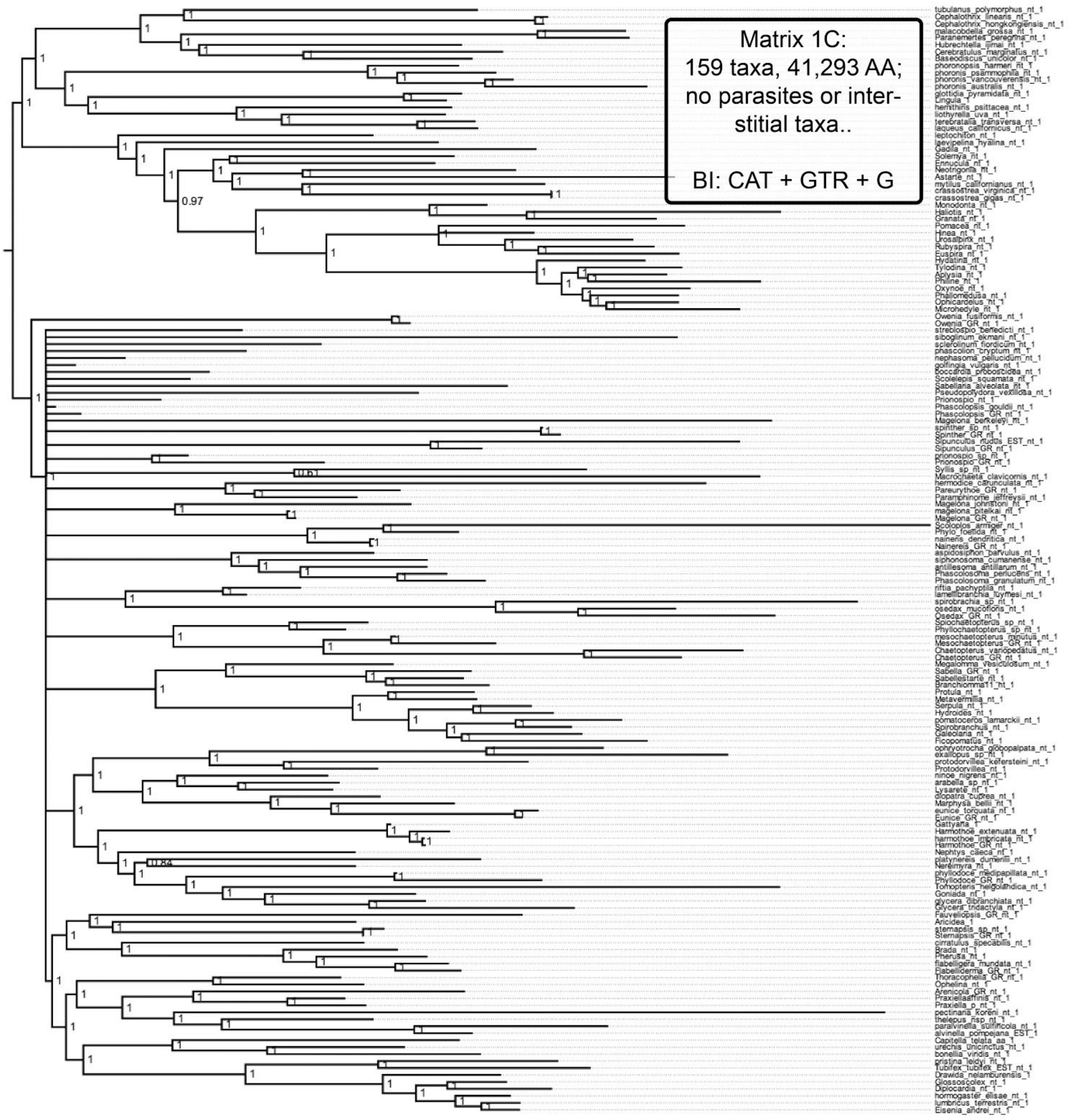


Figure 5.6. Unconverged topology from Bayesian inference of Matrix 1C. 159 annelid and outgroup taxa, 41,293 AA. Topology inferred through PhyloBayes version 1.7j, applying CAT + GTR + G.

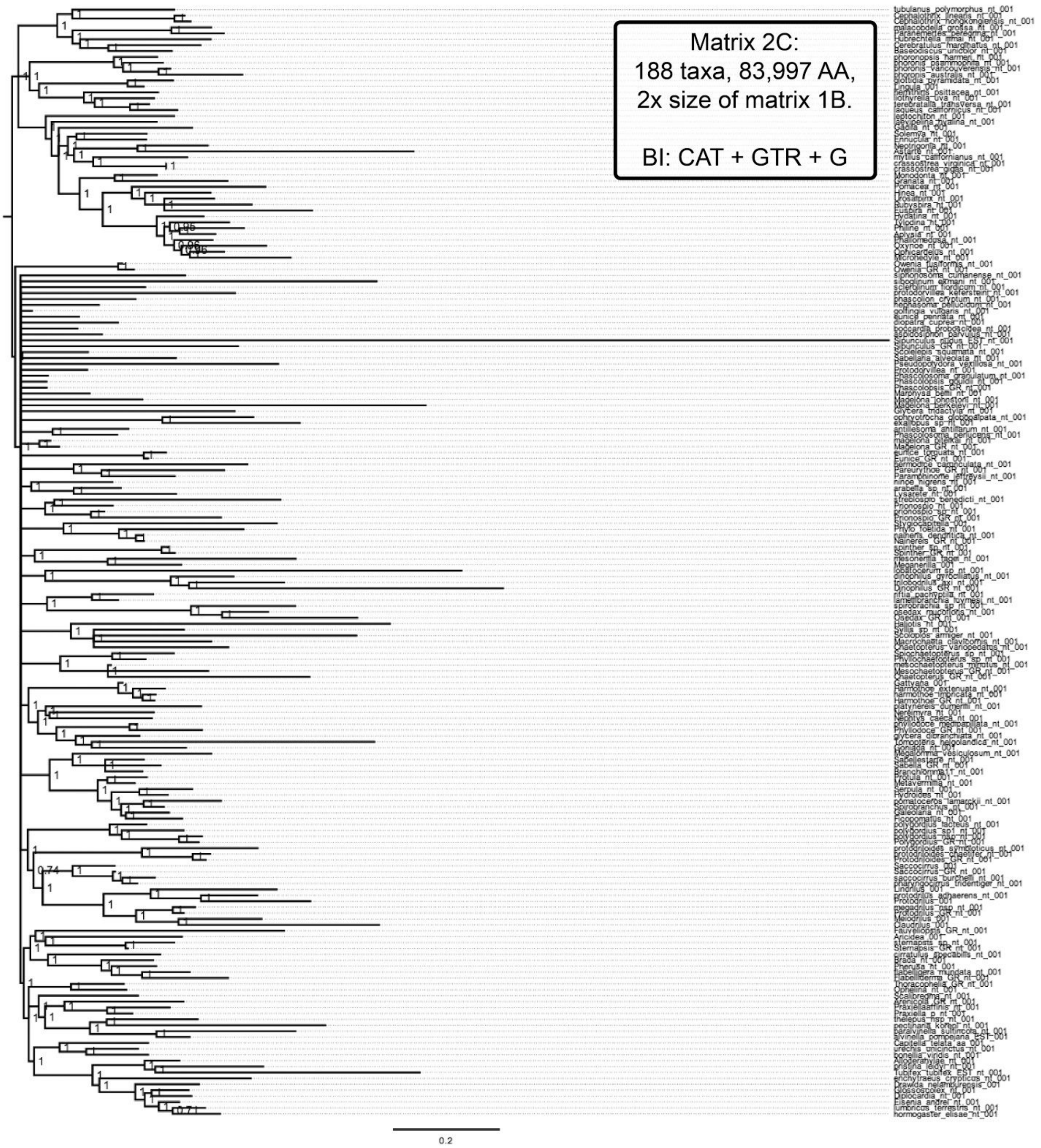


Figure 5.7. Unconverged topology from Bayesian inference of Matrix 2C. 188 annelid and outgroup taxa, 83,997 AA. Topology inferred through PhyloBayes version 1.7j, applying CAT + GTR + G.

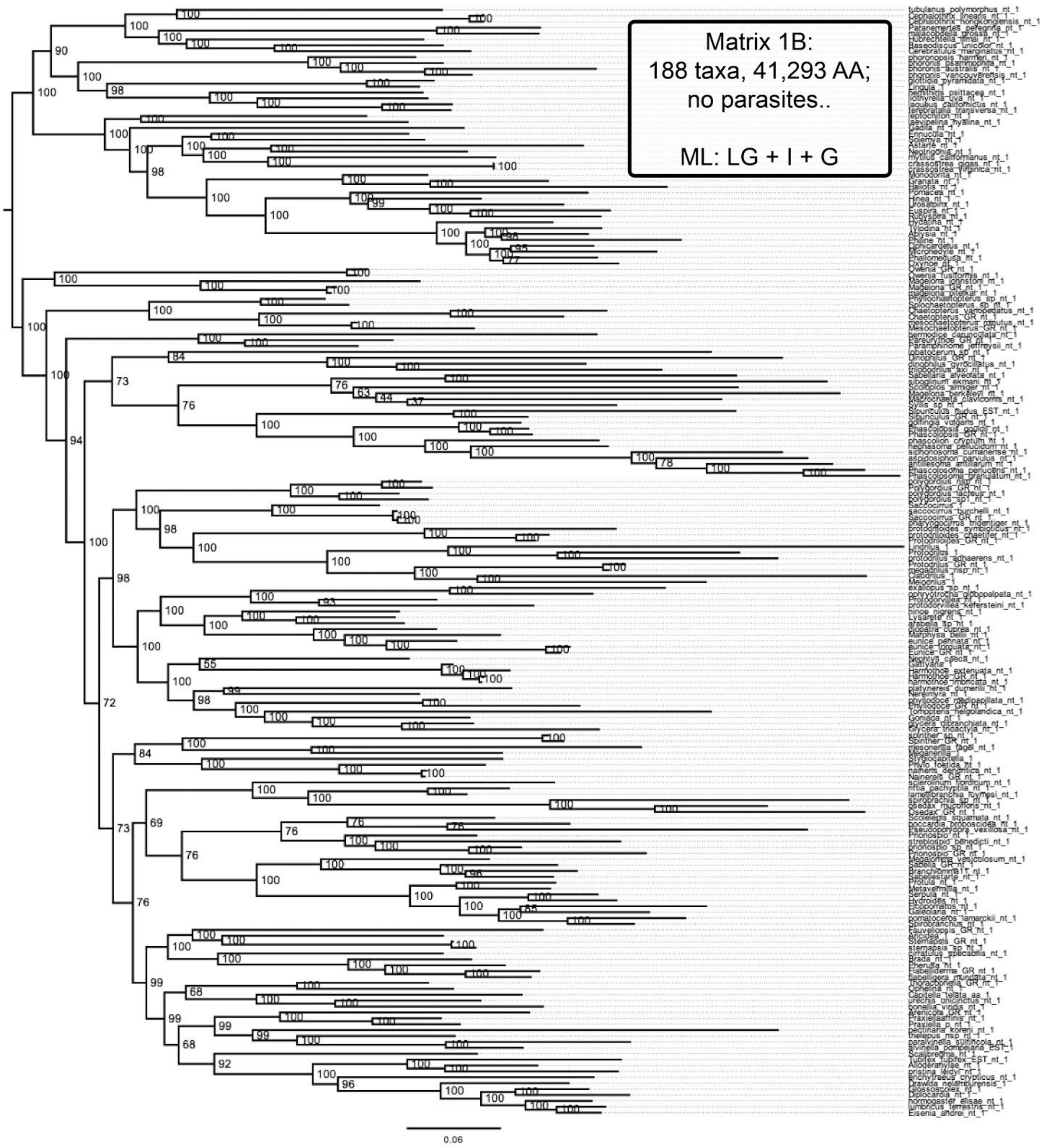


Figure 5.8. Maximum likelihood inference of Matrix 1B. 188 taxa, 41,293 AA, IQTree applying LG + I + G, 1000 bootstrap replicates.

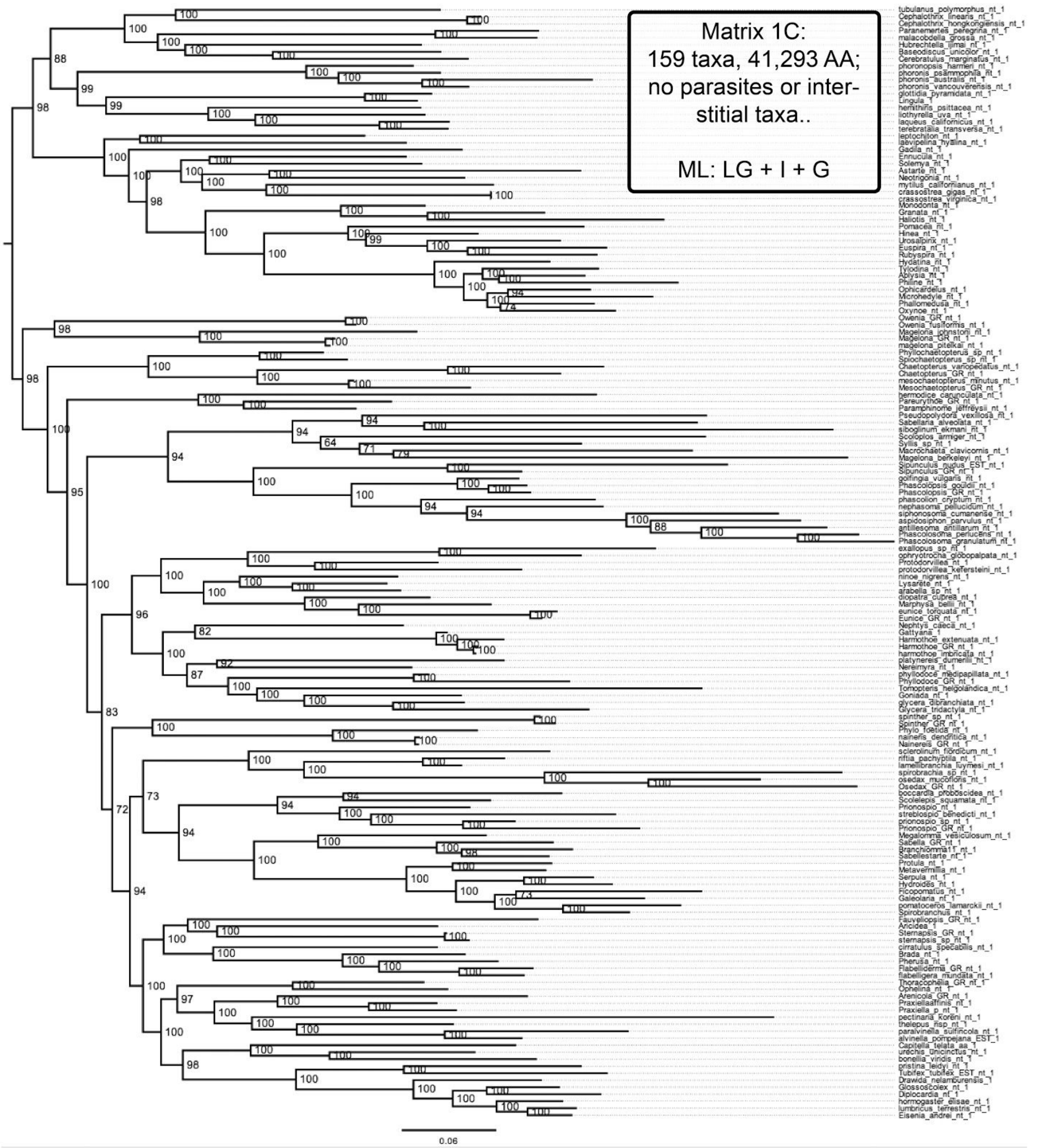


Figure 5.9. Maximum likelihood inference of Matrix 1C. 159 taxa, 41,293 AA, IQTree applying LG + I + G, 1000 bootstrap replicates.

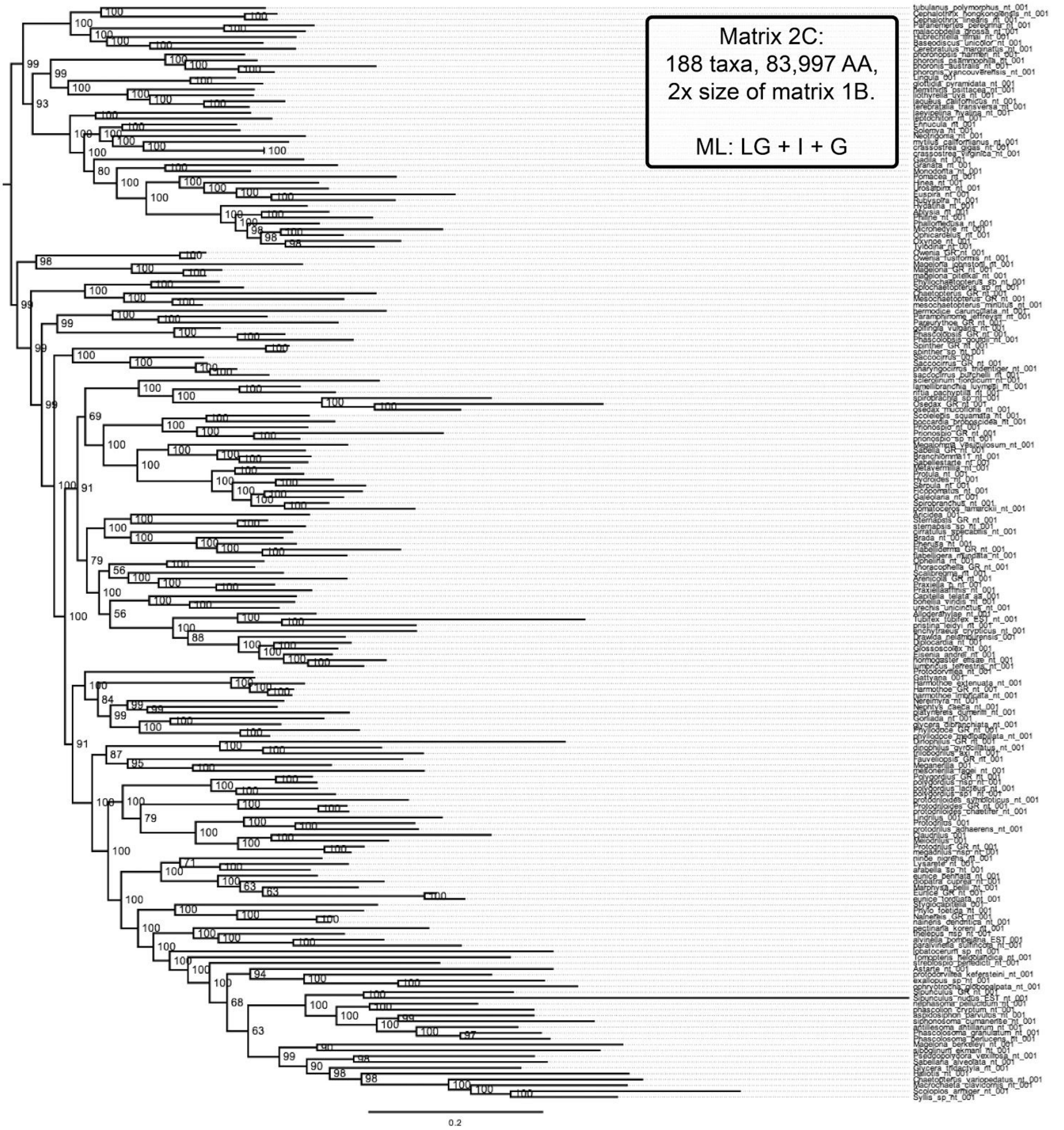


Figure 5.10. Maximum likelihood inference of Matrix 2C. 188 taxa, 41,293 AA, IQTree applying LG + I + G, 1000 bootstrap replicates.

5.4 Discussion

None of the inferential methods or datasets return consistent phylogenetic topology for annelids. Bayesian inference (of any of the datasets, or their Dayhoff-recoded counterparts) leads to Markov chains that cannot converge. This indicates that stationary distributions of parameters cannot be estimated given the data and the model, and is likely due to very weak signal, or a signal that is overwhelmed by the non-phylogenetic signal. This is compounded by the complexity of the dataset: with such a large number of taxa, inference seems to be beyond reasonable computationally tractable, even given powerful computing clusters. Without parameter values reaching stationarity, no phylogenetic topology can be supported as credible, through consensus reached by independent chains. As such, no meaningful discussion can be made on macroevolutionary dynamics, nor can any downstream analysis be carried out, such a divergence time inference. Molecular clocks rely on very robust phylogeny, both so that we can be certain of suitable signal from the molecular matrix to drive clock inference, but also so that calibrations can be placed on nodes that are strongly supported by independent inference of topology.

Bipartition comparisons between two independent chains do return posterior topologies that at first glance seem fairly well resolved, but all important nodes in the annelid clade are shown to be supported with posterior probabilities of 0.5. This can be highly misleading, and as such all figure (5.3, 5.4, 5.5 and 5.6) of Bayesian inference have nodes which have PP support below 0.6 are collapsed. If PP statistics are built from two equally-sized lists of trees, nodes with PP ≈ 0.5 should raise concerns regarding the interpretation of such support. It does not follow that, for both chains, they settled on the probability of that node existing being 0.5, but that the node has a posterior probability of 1 in one of the chains, and doesn’t exist at all (PP = 0) in the other chain. As such, the node should only be shown as a collapsed polytomy, not a support of PP = 0.5 (as evidenced by bipartition statistics equalling 1). Such an example heeds us to be wary of PPs, since, say, a PP of 0.75 could be expressing two different things: between two chains it could be an average of 0.5 and 1. Or it could be that both chains consider that node to have a PP of 0.75 (or anything in between these two examples). This highlights the importance of convergence, and the invalidity of any Bayesian inference that has not displayed appropriate statistical concord between independent chains. As an aside, extra chains were run to test whether *any* convergence was arrived at, and (informally) check whether we simply hadn’t had “poor luck” in MCMC getting stuck at local optima (which admittedly should not be the case given appropriate model and data). Running eight independent chains did not return further hints of convergence: all 28 interchain comparisons showed no convergence. Such a test, while informal, suggests a highly complex dataset, and that signal refinement approaches have not been effective enough in this case.

Efforts to minimise composition and saturation bias, as carried out through Dayhoff recoding of amino acids to six amino acid categories, does lead to Markov chain convergence. However, these inferences are only certain about there being uncertainty: all Dayhoff recoded phylogenies are characterised by extensive polytomies. As such, none of these phylogenies are presented here.

Maximum likelihood always returns resolved topologies, and bootstrap node-support that is based on information distribution in the dataset, not a probabilistic determination of inference-certainty. As such, ML should always been treated with caution, and here we provide strong examples of how misleading bootstraps can be. Each ML inference returns worrying high support on nodes (see Figures 5.8, 5.9 and 5.10), yet the Bayesian counterparts cannot agree with there being a solution for such data, let alone uphold nodes as being statistically supported. The results here exemplify these kinds of spurious results.

The likelihood in this case is that the signal to noise ratio is so poor that dataset refinement and application of the most currently suitable model is not sufficient to recover anything other than spurious or unresolved phylogenies (Townsend, 2007; Townsend et al., 2012). As such, analysis could me made on the datasets in order to assess signal strength. A first approach could be carried out using the methods of Yang (Yang, 1998) or Goldman (Goldman, 1998). These methods optimise experimental and inference parameters, using Markov chain methodology, and may well offer ways in which to identify, refine or otherwise modify datasets to maximise phylogenetic suitability and power. However, such methods are highly computationally demanding, and even modest sized molecular matrices represent nearly intractable problems (Townsend et al., 2012). Therefore, were such an approach to be applied, further matrix curation should be carried out to better tackle the problem of annelids.

A potential weak step in the curation methods used here is that paralogy is not suitably identified and excised. Orthology, here, has been assessed on strength of e-value statistics of BLAST results, and through removal of long-branch generating sequences (in pre-concatenation gene matrices) using the custom script treecleaner.pl (github.com/jairly/MoSuMa_tools). Considering the size of the taxonomic sample, any "by eye" assessment of gene trees not only suffers from being a subjective activity which no prior belief can guide, but of being too time-consuming a task for anybody to reasonably accomplish with any useful accuracy. Application of more stringent e-value threshold might be suggested, but this will still allow paralogous sequences into the final dataset, and because of our high (but potentially misled) confidence of homologous status these could lead to even stronger support for incorrect phylogeny. In any case, it remains difficult to ascertain whether signal is in conflict due to inclusion of paralogues, or due to a more general lack of information for the sequences included in a matrix.

It has been recognised that phylogenetic datasets can have unexpected instabilities, in that taxa may take up different phylogenetic positions after ostensibly minor changes to the dataset or inference protocol, particularly if the signal to noise ratio is poor (Cueto and Matsen, 2011). Methods to identify rogue taxa (those with highly unstable phylogenetic position through analysis) have failed to reasonably identify problems within this dataset. For example, the software Roguenarok (Aberer et al., 2011, 2013) (github.com/aberer/RogueNaRok) does not, in this case, appear to function as intended. The software seems to return almost any taxon as rogue, and if those sets of taxa are removed, and reanalysed with Roguenarok, it simply returns more and more rogue taxa. Those returned as rogue also seem highly unstable themselves - minor changes to the dataset (for example removing some taxa or using a subsection of the dataset) leads to radical changes in what Roguenarok considers rogue. This can be considered as further evidence that the dataset is of insufficient power to return sensible phylogeny, although it is accepted that a more systematic approach to Roguenarok could provide better detail on notably unstable taxa. For example, a matrix could be built up, starting with a very small number of taxa, say 10 or 20, then added to until instability becomes dominant. However, careful experimental procedure would be required since the number of potential approaches is astronomical, and inference of a large number of phylogenies is intractable. Thus, a justified regime to build up the matrix would have to be applied, lest the experiment become computationally unrealistic, or the results too difficult to interpret with any confidence. We also appreciate that there are alternative methods for identification of rogue taxa, and promote a diverse approach to identification of rogues.

Matrix simplification may prove a profitable approach, but would again be reliant on carefully planned experimental procedure. As mentioned, the taxonomic spread could be incrementally built up until thresholds of instability are reached. Another method may be to create highly complete dataset, but with refined taxonomic spread, through the creation of chimeric or consensus taxa. For example, species which, from independent lines of evidence, are strongly believed to form clades (such as clitellates, or outgroups) could then be merged into consensus sequences to represent that clade. As such, data from many species (as in this project) could be distilled down to a much smaller complement of taxa, allowing more tractable computation. A major criticism would naturally be in questioning the strength of the independent confidence in those clades existing (and, for annelids and other controversial groups), that those clades are the very ones that require inference in the first place. Further to this, it is clear that information loss through merging of taxa to make consensus sequences could represent an unacceptable modification to the data, essentially smearing out detail of sequence content. Nevertheless, this approach might be useful for guiding just where problems lie, if not in returning a fully meaningful and upholdable topology.

The development of Bayesian inference software that can be highly customised, such as RevBayes (Höhna et al., 2016), may well prove to set a new paradigm for phylogenetics, and could be the key to sensible analysis of recalcitrant evolutionary datasets. Through RevBayes, Bayesian parameters can be specified that other inference, including PhyloBayes, cannot. It seems that, for this dataset, the complexity of the data leads to Markov chains that cannot properly identify or describe parameter optima, and thus the phylogeny inference as a whole fails. This may be due to improper "mixing" behaviour of the Markov chains, that is, the ability of a chain to suitable explore parameter-space and settle on a stationary distribution of the parameter. In the parameter "landscape" metaphor, this can be thought of as either making new parameter propositions that are too large, thus missing global optima because they are never visited; or too small, so that local optima become locked, or too slow to approach anyway due to each new proposal being too conservative (Huelsenbeck et al., 2002). Refined control of chain behaviour may offer a solution to this problem, although it is possibly true that such computational issues would vanish, given a meaningful dataset. Ascertaining whether such a dataset could be compiled should be a priority over application of exotic methodologies.

A greater degree of control over the inference model and data partitioning, as offered by RevBayes, may also be profitable. Careful sectioning of the dataset, perhaps based on evolutionary tempo and mode of particular genes or positions. It is also suggested that lineage- or clade-specific models might well be applicable in RevBayes, but of course this necessitates at least some kind of prior knowledge of the phylogeny itself. The method could be criticised as having an element of circularity, unless strong independent evidence could be presented for some characteristics of the topology. For the time-being, these kinds of methods are still in early development, and we are yet to see evolutionary scenarios which can (only) be dealt with through such complex dataset handling and inference. In all cases, these ideas offer exciting opportunities for evolutionary biologists dealing with impenetrable phylogeny, as has been the case here with annelids.

A final point to consider is on the nature of polytomies. In phylogenetic inference, a polytomy is usually regarded as a methodological failure. However, it may be the case that some evolutionary events can only be described as a polytomy. Speciation normally results in two isolated lineages, but it is not impossible for three or more branches to originate from a single node; it merely means that genetic isolation was not complete before new branches were inaugurated (as is reflected in incomplete lineage sorting). This kind of phylogenetic feature though is likely to confound our inference methodologies, as not only can a polytomy be “accepted” as a possibility (either by Bayesian or ML approaches), but also the presence of a single “real” polytomy on a tree may have knock-on effects elsewhere in the inference, causing other nodes to be difficult to infer. As such, we are cautioned to remember that a topology is only a graphical hypothesis given the data and the model, not a result that rigidly lays out the nature of evolution. Trees are highly useful for us to conceptualise evolution, but they are not perfect, and some evolutionary narratives may be impossible to meaningfully relate through the tree metaphor.

5.5 Conclusion

The evolutionary origins of annelids continues to be one of the most difficult problems in the metazoan Tree of Life. We have here shown that approaches using either pre-selected genes of known conservation and slow evolutionary rate do not seem able to return a supported topology. Our expanded dataset derived from from transcriptome-wide orthology searches also did not provide informative datasets. Our methodologies of BI, ML, and amino acid recoding also fail to return consistent phylogenetic topology. We urge further research to focus on new methods to deal with the annelid phylogeny, in signal-refinement, taxonomic selection, data-partitioning, model selection, and statistical examination of Markov chains in order to provide meaningful Bayesian results. The annelids remain a highly valuable Metazoan group, phylogenetically, firstly for the practical reason that our knowledge on their early evolution remain vague. Secondly, and perhaps more importantly, they represent an ideal set of organisms on which to develop and refine our methodological approaches; work on annelids will contribute to best-practice on deep-time phylogenetic inference, and will allow us to reconsider other groups, altogether helping up to understand evolution seen through phylogeny.

Chapter 6

The future of phylogenomics

This chapter has not been published.

Abstract

The contributions to molecular palaeobiology made by the research in this thesis are outlined. From an empirical point of view we summarise the progress made through this research. We also highlight the publicly-available sequence information and coding arising from these projects, and its potential for use in dealing with further palaeobiological questions. The strengths and weaknesses of current methodologies are discussed, as well as present trends in terms of software, approaches, and data-handling. We speculate on future phylogenetic methods, and how its data and methods may change in light of expanding information, and how data is handled in an increasingly connected world.

6.1 Advances made through this thesis

The research projects presented here are hoped first of all to be contributions to knowledge on the evolutionary relationships and histories of the organisms focused upon. A crucial success of the research has been to demonstrate empirical phylogenomic investigation of evolutionary dynamics, especially in Chapters 2, 3 and 4, using a wealth of current information, and the best current methods. Further to this, we have integrated palaeontological evidence to support timelines of evolutionary dynamics, and where possible relate these to ecology on its widest scale. On a more practical note, this work relied on the transcriptomic sequencing of new species, and the now-publicly-available data generated from this can be seen as another contribution to molecular biology. These new data significantly expand repositories of information for the cephalopods, chelicerates and especially annelids, and are now integrated into the ever-growing banks of sequence information that we have of life on Earth. It is hoped these data will prove valuable to future phylogenetic, genomic, and palaeobiological investigations. The work here also makes publicly available a pipeline for the assembly of transcriptomes, the curation of phylogenomic datasets, and the preparation of such data for downstream analysis; these scripts are open-source and can be modified and updated by anyone.

To summarise the four empirical research project covered in this thesis, the following can be said. We provide evidence that modern coleoid cephalopods are a consequence of Mesozoic marine turnover, when ecologies shifted towards more kinetic lifestyles where retaining a shell was untenable in the face of both competition for niches, and prey advantage (Chapter 2). In Chapter 3, we show that a meaningful phylogeny of chelicerates can be recovered when data and methods are refined, and this allows us to place the chelicerate invasion of land around the end of the Cambrian. The origin of earthworm diversity coincides with forest ecosystem turnover, and considering their current ecosystem engineer roles, we uphold our molecular work as evidence for earthworms as (at least in part) instrumental in shifting Palaeozoic terrestrial ecologies (Chapter 4). Finally, we see that phylogenetic conflict can overwhelm some current methods, and the approach of expansion of taxonomic representation can lead to further problems, rooted in the complexity of the data and phylogeny (Chapter 5). Together, these projects show how careful approach to phylogenomics can result in meaningful results, which can then drive interpretations of macroevolutionary dynamics (Chapters 2, 3 and 4). Chapter 5 provides an example of an evolutionary scenario which cannot currently be resolved using the methodologies developed and applied in this thesis, and we thus suggest reasons for this failure and potential approaches to deal with the phylogeny of Annelida.

We have covered how, in the genomic age, molecular palaeobiology has significant investigatory power, shedding light on areas that are often obscured by paucity of traditional palaeontological evidence. However, this work has also highlighted that much work is still to be done to understand metazoan macroevolution, especially at moments of great ecological change. With growing synergism between disciplines, and the ongoing analytical advances, consilient views will emerge in both describing and explaining the past. Considering the crucial role of fossils in guiding molecular clock analyses (Pisani and Liu, 2015), we hope that palaeontologists continue their valuable work in sampling and describing fossils. Fossils are ascribed systematic position through comparative biology, and temporal position through the various methods of dating strata, and if these activities are done as well as possible, then molecular workers cannot ask for more. However, palaeontologists may well ask for more from their molecular colleagues. Firstly, it is important when inferring molecular clocks that proper calibration prior distributions are applied, and understanding in this areas now seems mature (Foster and Ho, 2017; Parham et al., 2011; Warnock et al., 2012; Wolfe et al., 2016). By understanding and adhering to these best practices, hard-found fossils can be as useful as possible in the, arguably, more labile arena of molecular inference. Secondly, we should also hope to integrate timed phylogenies of extant organisms (as this thesis does) with information from long-dead species. After all, palaeontologists and molecular biologist both hope to explain life through the phylogeny, and further uniting these two should prove a powerful technique. Of course, a major stumbling block is that while molecular data can be reasonably well modelled, the same cannot so easily be said of morphological data.

6.2 Further methodologies and models

Techniques such as Total Evidence Dating (TED) (O’Reilly et al., 2015; Ronquist, Klopfstein, et al., 2012) extends the fossil guidance of time-scale inference from the “standard” inter-branch node calibration scheme to incorporating fossils as tips, or “terminal nodes”. While uptake is possibly impeded by philosophical and data-handling conflict between palaeontologists and molecular biologists, the technique will certainly prove profitable for some areas of evolutionary inference. Challenges remain to be addressed in TED include the lack of morphological evolutionary models, how to deal with non-random presence or absence of fossil characteristics, and best practice for the accomodation of uncertainty in fossil ages (O’Reilly et al., 2015). However, for some areas the use of TED is proving valuable, with the evolution of penguins representing an exemplary group for which understanding would otherwise be more vague, without the technique (Gavryushkina et al., 2017). Other palaeontological areas will continue to be addressed by TED, and if the method becomes well established it may well prove to be a suitable approach to the widest phylogeny of Annelida, given a strongly supported topology. Beyond that, it is even possible that such investigations may have the potential to contribute to the currently vague comprehension of the relationships between the genotype and the phenotype, a topic that continues to rise in profile with the rapid growth of gene-editing technology (Singh et al., 2017), synthetic biology (Davies, 2017) and evolutionary-developmental biology (Schwab and Moczek, 2017), “evo-devo”.

Evo-devo is concerned with the question of how developmental processes of an organism are related to its evolutionary history. In the *Origin*, Darwin used embryological characteristics as a line of evolutionary evidence, and as such carried out a simple type of evo-devo. Today evo-devo has become a sophisticated school of molecular biology. Like phylogenomics, evo-devo has advanced apace since the genomic revolution, and the field is becoming increasingly dynamic as biologists become proficient with sequencing technologies and computational biological techniques. Links are starting to be confidently made between palaeobiology and evo-devo, and integration of understanding evo-devo in a deep-time context will lead to important advances (Haug and Haug, 2017). For the work presented here, perhaps in future we will be able to reconcile developmental biology with phylogenomics to provide further insight on the history or cephalopods, chelicerata or annelids?

In terms of application of molecular evolutionary models, it feels that currently we have a stable repertoire. Popular models such as LG, CAT and GTR are statistically supported as best fit, especially for wide-group and deep-time inference of phylogeny, as palaeobiology often interested in. However, it is hoped that the way in which these models are applied is refined for future investigation. Firstly, data partitioning is sometimes appropriate, so that a separate model can be applied to each section of sequence information, which may be under a particular evolutionary regime. With the advance of fully-scale phylogenomics (rather than inference based on a handful of genes), partitioning by gene has been neglected, due to the complexity of having hundreds or thousands of genes. Perhaps, though, it is time to reconsider a new partitioning approach, given that matrix-wide modelling can remain unresolved (see Chapter 5). With our wealth of information, is it time to test partitioning based on characteristics of data in the matrix, for example on chemicophysical properties of amino acids, tertiary structure traits (such as buried index, helix and turn propensity), or even entropy levels of sequences? With the rapid advance of cryo-electron-microscopy (Fernandez-Leiro and Scheres, 2016; Kowal et al., 2018), it future years it might even be worth assessing three-dimensional morphological characteristics of protein-products for phylogenetic ends, representing in a way a bridge between the genotype and the phenotype.

Secondly, it may be appropriate to apply clade-specific models, thus dealing with the mode of molecular evolution as it changes across a phylogeny. Such approaches could expand on clade-specific modelling approaches (Roure and Phillipe 2011), building on implementations such as in PROCOV (Wang et al., 2009). Of course, this requires at least some prior understanding of the phylogeny itself, and so a type of co-estimation would probably be required, inferring nodes and then modifying the fit of the model for subtending nodes. This kind of approach may be possible with the highly customisable software RevBayes (Höhna et al., 2016), a successor to MrBayes (Ronquist, Teslenko, et al., 2012). In this, not only can the model be specified precisely, but it will also be possible to integrate these kinds of lineage-specific modelling methodologies. Uptake of RevBayes is increasing, and so scripts to apply such approaches will become more publicly available. Potentially, the approach will be highly beneficial for palaeobiology as we deal with more difficult evolutionary problems. Further to that, RevBayes also represents model specification at its most fundamental, and (for those that use it) will mean the end to “black box” inference softwares, leading to biologists with a true understanding of their method.

For molecular clocks, a potentially profitable route may be in partitioning data by gamma category, and applying an independent clock for each partition. Through this, it may be possible to return better-constrained confidence intervals for node ages, which naturally through relaxed clock methods can be wide. This type of approach can be applied in the software package BEAST (Baele et al., 2017), and may also be constructed using the RevBayes language. However, the challenge will be to suitably validate results, and avoid pitfalls in presenting inappropriately precise node-age estimates. Gamma partitioned clock methodology would, as ever, require the careful application of fossil calibrations schemes. Nevertheless, such an approach appears to have potential, and would benefit from an empirical proof-of-concept to encourage more workers to develop such methods.

Another, speculative, future potential for molecular palaeobiology is in machine-learning and AI. The amount of sequence data we currently have should lead us to return to the question of how it should be handled. Currently, and as in these projects, datasets have been curated and analysed manually, with a few automated steps via scripting and coding. Naturally, this is long-winded, fairly error-prone, and by the time the work is done curating a set, new data is available to be added. (To illustrate this, the matrix curated for Chapter 2 (assembled 2014-2015), if made today, would contain more than double the number of species.) A solution to these kinds of issues may be in automation and real-time interrogation of new additions to sequence repositories, as they happen. Using such techniques, it could be possible to generate, curate and refine molecular datasets in a nearly fully-automated way. A basic version of such a system could be a conceivable modification to some of the scripts written for this thesis, only requiring code that reliably integrates with online servers, which monitors the species, family or phyla of information that has been uploaded. Potentially, such a program could present molecular palaeobiologists with a permanently-up-to-date phylogenomic matrix for their organisms of interest, removing the most onerous step of phylogenomics. A further conceptual advance might be to apply machine-learning algorithms to search for patterns in the data which may prove phylogenetically informative, but which otherwise would not be picked up by current methods. It is possible to foresee datasets which integrate composition, synteny, epigenetics, and other as-yet unthought of characteristics of sequence information as fuel for phylogenetic analysis. So, we might ask palaeobiologists to consider whether (or when) they should be extending their collaborations to include machine-learning and AI computer scientists.

6.3 In conclusion

In the introduction we introduced the nature of phylogenetics and phylogenomics, and stated our goals as being in the expansion of knowledge on the macroevolutionary dynamics of four groups of organisms. Some of those empirical goals have been met, and where resolution has proved difficult, it is hoped that suggestions for progress influence future research direction. But, together as a whole, the thesis contributes to our understanding of Metazoan evolutionary biology, and has shown how past ecologies can be investigated using molecular methodology. Future sequence information, methodologies and computational approaches have great promise for phylogenomics and palaeobiology, as we continue to refine our picture of the narrative of evolutionary history.

6.4 Epilogue - thoughts from (and for) a PhD candidate

The experience of carrying out and completing PhD will vary widely, from person to person, subject to subject, culture to culture. Perhaps, though, it is commonly experienced as being difficult, since it requires the development of a range of skills that are not (and cannot be) taught or studied. In research, a PhD candidate has to accept stepping into the unknown, has to manage a range of personal relationships, has to learn resilience to criticism, and must come to terms with being in a peculiar sub-culture, populated exclusively by ultra-high-achievers. I provide a few points, that I wish I had known when I started, in the hope that they can help others (both students and supervisors). These points are mostly in the context of molecular biology and bioinformatics, but they might be applicable more generally.

- Play to your strengths. It might take a couple of years to identify these, but once you do, develop them. Don’t try to be good at everything. Have pride in the things you are good at. Be honest about the things which are not your strengths. If anyone thinks you are weak for admitting weakness, that is their problem.
- Through your PhD you are in effect learning a new language. It takes time, but constant immersion in academic literature eventually leads to comprehension. Being able to use your new glossary in spoken communication leads to fluency. Try to find a friendly colleague to look up and discuss unfamiliar language with. Don’t be afraid, even late in your PhD, to ask “what do you understand by the word _____?”, even of words you feel you understand; your comprehension will change as your vocabulary expands.
- Ensure your supervisors identify the most important papers, both recently and in earlier development of their field. Supervisors might assume you have this knowledge already - there is no shame in not knowing the literature comprehensively, even late in your PhD. Some people just don’t have a mind for remembering papers. That is fine, it might not be your strength. The literature, especially these days, is effectively infinite - do not let the endlessness of publications get you down. Much of it is bollocks veiled in esoteric language. If, even after years of immersion in your field, you cannot grasp a paper, it is a failure of the authors, not you. Ignore and move on.
- Ensure your supervisors provide comprehensive detail on research groups working in similar areas. There is no reason why you should have insight on these groups - you should be briefed on who is active in the field, what their research style is, perhaps even what their personality is. Be wary of your supervisor’s opinions - they are not impartial or unbiased.
- Start coding as soon as possible, but question the advice from your supervisor. Your supervisor may know the field, but will likely be methodologically obsolete (even if they learned only five years ago). Today learning to code is far easier than even the very recent past, because of hubs like stack-overflow, github and other coding sites and forums. Do not learn perl for bioinformatics, or at least not exclusively - you may find yourself isolated from other bioinformaticians, and you will not maximise your transferable skills. In short, learn python and shell. (Be wary of what I am saying, I might be obsolete too.)
- Have two prongs to your research.

  - Have a “mundane but achievable” goal of repeating well-established methods on new data. Aim for this to be your first paper and first chapter, and should have clear goals of what data to use and methods to apply from the outset. Your secondary goal will be to understand the limitations of those methods.
  - Have a “risky but high-reward” goal of building on established methods to confront limitations, but only consider this once you have mastered the current method repertoire. Ask your supervisors about already-tried methods in the literature - many of those experiments might be quite old, and you can’t be expected to know of little-cited (because it doesn’t work, or simply wasn’t taken up by the community) methods from decades ago. Merely tweaking an existing method may well prove very fruitful; you have to rewrite the rulebook.
- Don’t worry if you can’t come up with a new method, it is difficult, requires a certain type of creativity, and might not be your strength. That’s fine. Don’t worry if a methodology fails. Keep in mind the mountains of unpublished negative results out there. What is published is a misrepresentation of research; it is just the “good bits” that people are happy for others to see, a bit like a facebook profile. Sadly, there is no forum or channel for crucial negative results to be disseminated.
- Be brave when speaking to your supervisor. They might seem like they know everything, but they don’t. You are witnessing survivorship-bias - those with (or who develop) self-confidence make it to principal investigator positions. PIs and faculty are the result of selection, primarily on the trait of competitiveness. Your current self-doubt is entirely reasonable, and crucial for robust science. If anything, your supervisor’s self-confidence is less reasonable than your self-doubt; if good science is the goal, they need you more than you need them. It is fine to have impostor syndrome - in fact, if you don’t, you should be worried.
- Always have sympathy for your supervisor. They are under enormous pressure, with responsibilities to carry out research; to understand the current state of their field; to teach hordes of students who may be generationally alien to them; to have people-management skills in orchestrating a range of postgraduates and other colleagues; to respond to management about progress and future plans; to perpetually apply for grants with looming deadlines; to be resilient to rejected grant applications; to carry out significant pro-bono work in reviewing articles, organising conferences and engaging with their research community; and to generally comply with the “publish or perish” existence of academia, when they would probably prefer things to be different. They can’t be qualified and skilled in all these roles, and will inevitably lack competence or natural ability in some of them. That is fine, we all have strengths and weaknesses.

References

Aberer, A.J., Krompaß, D. and Stamatakis, A. (2011), “RogueNaRok: An efficient and exact algorithm for rogue taxon identification”, *Heidelberg Institute for Theoretical Studies: Exelixis-RRDR-2011--10*, available at: http://sco.h-its.org/exelixis/pubs/Exelixis-RRDR-2011-10.pdf.

Aberer, A.J., Krompass, D. and Stamatakis, A. (2013), “Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice”, *Systematic Biology*, Vol. 62 No. 1, pp. 162–166.

Alberti, G., Heethoff, M., Norton, R.A., Schmelzle, S., Seniczak, A. and Seniczak, S. (2011), “Fine structure of the gnathosoma of Archegozetes longisetus Aoki (Acari: Oribatida, Trhypochthoniidae)”, *Journal of Morphology*, Wiley Subscription Services, Inc., A Wiley Company, Vol. 272 No. 9, pp. 1025–1079.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990), “Basic local alignment search tool”, *Journal of Molecular Biology*, Vol. 215 No. 3, pp. 403–410.

Anderson, F.E., Williams, B.W., Horn, K.M., Erséus, C., Halanych, K.M., Santos, S.R. and James, S.W. (2017), “Phylogenomic analyses of Crassiclitellata support major Northern and Southern Hemisphere clades and a Pangaean origin for earthworms”, *BMC Evolutionary Biology*, Vol. 17 No. 1, p. 123.

Baele, G., Lemey, P., Rambaut, A. and Suchard, M.A. (2017), “Adaptive MCMC in Bayesian phylogenetics: an application to analyzing partitioned data in BEAST”, *Bioinformatics* , Vol. 33 No. 12, pp. 1798–1805.

Benton, M.J. (2009), “The Red Queen and the Court Jester: species diversity and the role of biotic and abiotic factors through time”, *Science*, Vol. 323 No. 5915, pp. 728–732.

Benton, M.J., Donoghue, P. and Asher, R.J. (2015), “Constraints on the timescale of animal evolutionary history”, *Palaeontologia Electronica* , palaeo-electronica.org.

Benton, M.J., Wills, M.A. and Hitchin, R. (2000), “Quality of the fossil record through time”, *Nature*, Vol. 403 No. 6769, pp. 534–537.

Bergmann, S., Lieb, B., Ruth, P. and Markl, J. (2006), “The hemocyanin from a living fossil, the cephalopod Nautilus pompilius: protein structure, gene organization, and evolution”, *Journal of Molecular Evolution*, Vol. 62 No. 3, pp. 362–374.

Berner, R.A. (2003), “The long-term carbon cycle, fossil fuels and atmospheric composition”, *Nature*, Vol. 426 No. 6964, pp. 323–326.

Bohlen, P.J., Scheu, S., Hale, C.M., McLean, M.A., Migge, S., Groffman, P.M. and Parkinson, D. (2004), “Non-native invasive earthworms as agents of change in northern temperate forests”, *Frontiers in Ecology and the Environment*, Wiley Online Library, Vol. 2 No. 8, pp. 427–435.

Bolger, A.M., Lohse, M. and Usadel, B. (2014), “Trimmomatic: a flexible trimmer for Illumina sequence data”, *Bioinformatics* , Vol. 30 No. 15, pp. 2114–2120.

Bomfleur, B., Kerp, H., Taylor, T.N., Moestrup, Ø. and Taylor, E.L. (2012), “Triassic leech cocoon from Antarctica contains fossil bell animal”, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 109 No. 51, pp. 20971–20974.

Boyle, P. and Rodhouse, P. (2008), *Cephalopods: Ecology and Fisheries*, Wiley.

Briggs, D.E.G. and Kear, A.J. (1993), “Decay and preservation of polychaetes: taphonomic thresholds in soft-bodied organisms”, *Paleobiology*, Cambridge Univ Press, Vol. 19 No. 01, pp. 107–135.

Brocklehurst, N., Upchurch, P., Mannion, P.D. and O’Connor, J. (2012), “The completeness of the fossil record of mesozoic birds: implications for early avian evolution”, *PloS One*, Vol. 7 No. 6, p. e39056.

Bromham, L. (2009), “Why do species vary in their rate of molecular evolution?”, *Biology Letters*, Vol. 5 No. 3, pp. 401–404.

Brundrett, M.C. (2002), “Coevolution of roots and mycorrhizas of land plants”, *The New Phytologist*, Blackwell Science Ltd, Vol. 154 No. 2, pp. 275–304.

Brusatte, S.L., Nesbitt, S.J., Irmis, R.B., Butler, R.J., Benton, M.J. and Norell, M.A. (2010), “The origin and early radiation of dinosaurs”, *Earth-Science Reviews*, Vol. 101 No. 1, pp. 68–100.

Budd, G.E. and Telford, M.J. (2009), “The origin and evolution of arthropods”, *Nature*, Vol. 457 No. 7231, pp. 812–817.

Bush, A.M. and Bambach, R.K. (2011), “Paleoecologic Megatrends in Marine Metazoa”, *Annual Review of Earth and Planetary Sciences*, Vol. 39 No. 1, pp. 241–269.

de Carvalho, M. da G.P. and Lourenço, W.R. (2001), “A new family of fossil scorpions from the Early Cretaceous of Brazil”, *Comptes Rendus de l’Académie Des Sciences - Series IIA - Earth and Planetary Science*, Vol. 332 No. 11, pp. 711–716.

Cascales-Miñana, B. and Cleal, C.J. (2014), “The plant fossil record reflects just two great extinction events”, *Terra Nova*, Vol. 26 No. 3, pp. 195–200.

Castresana, J. (2000), “Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis”, *Molecular Biology and Evolution*, Vol. 17 No. 4, pp. 540–552.

Castresana, J. (2002), “Gblocks, v. 0.91 b. Online version”.

Clarke, J.T., Warnock, R. and Donoghue, P.C.J. (2011), “Establishing a time-scale for plant evolution”, *The New Phytologist*, Wiley Online Library, Vol. 192 No. 1, pp. 266–301.

Claverie, J.M. (2001), “Gene number. What if there are only 30,000 human genes?”, *Science*, Vol. 291 No. 5507, pp. 1255–1257.

Clements, T., Colleary, C., De Baets, K. and Vinther, J. (2016), “Buoyancy mechanisms limit preservation of coleoid cephalopod soft tissues in Mesozoic Lagerstätten”, *Palaeontology*, doi.org/10.1111/pala.12267.

Condon, D., Zhu, M., Bowring, S., Wang, W., Yang, A. and Jin, Y. (2005), “U-Pb ages from the neoproterozoic Doushantuo Formation, China”, *Science*, Vol. 308 No. 5718, pp. 95–98.

Cueto, M.A. and Matsen, F.A. (2011), “Polyhedral geometry of phylogenetic rogue taxa”, *Bulletin of Mathematical Biology*, Vol. 73 No. 6, pp. 1202–1226.

Darwin, Charles and of Edinburgh., R.C. of P. (n.d.). *The Formation of Vegetable Mould, through the Action of Worms : With Observations on Their Habits*, London: J. Murray.

Davies, J. (2017), “Using synthetic biology to explore principles of development”, *Development* , Vol. 144 No. 7, pp. 1146–1158.

Dayhoff, M.O. (1976), “The origin and evolution of protein superfamilies”, *Federation Proceedings*, Vol. 35 No. 10, pp. 2132–2138.

Dera, G., Toumoulin, A. and De Baets, K. (2016), “Diversity and morphological evolution of Jurassic belemnites from South Germany”, *Palaeogeography, Palaeoclimatology, Palaeoecology*, Vol. 457, pp. 80–97.

Dexter, A.R. (1978), “Tunnelling in soil by earthworms”, *Soil Biology & Biochemistry*, Pergamon-Elsevier Science Ltd, the Boulevard, Langford Lane, Kidlington, Oxford, England OX5 1GB, Vol. 10 No. 5, pp. 447–449.

DiMICHELE, W.A., Mamay, S.H., Chaney, D.S., Hook, R.W. and Nelson, W.J. (2001), “An Early Permian flora with Late Permian and Mesozoic affinities from North-Central Texas”, *Journal of Paleontology*, Vol. 75 No. 2, pp. 449–460.

DiMichele, W.A., Stein, W.E. and Bateman, R.M. (2001), “Ecological sorting of vascular plant classes during the Paleozoic evolutionary radiation”, *Evolutionary Paleoecology. Columbia University Press, New York*, pp. 285–335.

Djokic, T., Van Kranendonk, M.J., Campbell, K.A., Walter, M.R. and Ward, C.R. (2017), “Earliest signs of life on land preserved in ca. 3.5 Ga hot spring deposits”, *Nature Communications*, Vol. 8, p. 15263.

Dodd, M.S., Papineau, D., Grenne, T., Slack, J.F., Rittner, M., Pirajno, F., O’Neil, J., et al. (2017), “Evidence for early life in Earth’s oldest hydrothermal vent precipitates”, *Nature*, Vol. 543 No. 7643, pp. 60–64.

Doguzhaeva, L.A., Mapes, R.H. and Mutvei, H. (1999), “A Late Carboniferous Spirulid Coleoid from ahe Southern Mid-Continent (USA)”, in Olóriz, F. and Rodríguez-Tovar, F.J. (Eds.), *Advancing Research on Living and Fossil Cephalopods*, Springer US, pp.

47–57.

Donoghue, P.C.J. and Benton, M.J. (2007), “Rocks and clocks: calibrating the Tree of Life using fossils and molecules”, *Trends in Ecology & Evolution*, Vol. 22 No. 8, pp. 424–431.

Drake, H.L. and Horn, M.A. (2007), “As the worm turns: the earthworm gut as a transient habitat for soil microbial biomes”, *Annual Review of Microbiology*, Vol. 61, pp. 169–189.

Drummond, A.J., Ho, S.Y.W., Phillips, M.J. and Rambaut, A. (2006), “Relaxed phylogenetics and dating with confidence”, *PLoS Biology*, Vol. 4 No. 5, p. e88.

Dunlop, J.A. (2010), “Geological history and phylogeny of Chelicerata”, *Arthropod Structure & Development*, Vol. 39 No. 2-3, pp. 124–142.

Edgar, R.C. (2004), “MUSCLE: multiple sequence alignment with high accuracy and high throughput”, *Nucleic Acids Research*, Vol. 32 No. 5, pp. 1792–1797.

Edgecombe, G.D. (2010), “Arthropod phylogeny: an overview from the perspectives of morphology, molecular data and the fossil record”, *Arthropod Structure & Development*, Vol. 39 No. 2-3, pp. 74–87.

Edgecombe, G.D., Giribet, G., Dunn, C.W., Hejnol, A., Kristensen, R.M., Neves, R.C., Rouse, G.W., et al. (2011), “Higher-level metazoan relationships: recent progress and remaining questions”, *Organisms, Diversity & Evolution*, Springer-Verlag, Vol. 11 No. 2, pp. 151–172.

Edwards, W.M., Shipitalo, M.J., Owens, L.B. and Norton, L.D. (1990), “Effect of Lumbricus terrestris L. burrows on hydrology of continuous no-till corn fields”, *Geoderma*, Vol. 46 No. 1, pp. 73–84.

Ehrenfeld, J.G. (2010), “Ecosystem consequences of biological invasions”, *Annual Review of Ecology, Evolution, and Systematics*, Annual Reviews, Vol. 41, pp. 59–80.

Erséus, C. (2005), “Phylogeny of oligochaetous Clitellata”, *Hydrobiologia*, Kluwer Academic Publishers, Vol. 535-536 No. 1, pp. 357–372.

Erwin, D.H. (2007), “Disparity: morphological pattern and developmental context”, *Palaeontology*, Wiley Online Library, Vol. 50 No. 1, pp. 57–73.

Falcon-Lang, H.J. (2015/3), “A calamitalean forest preserved in growth position in the Pennsylvanian coal measures of South Wales: Implications for palaeoecology, ontogeny and taphonomy”, *Review of Palaeobotany and Palynology*, Vol. 214, pp. 51–67.

Fauchald, K. (1974), “Polychaete Phylogeny: A Problem in Protostome Evolution”, *Systematic Biology*, Vol. 23 No. 4, pp. 493–506.

Fedonkin, M.A. and Waggoner, B.M. (1997), “The Late Precambrian fossil Kimberella is a mollusc-like bilaterian organism”, *Nature*, Nature Publishing Group, Vol. 388 No. 6645, pp. 868–871.

Felsenstein, J. (1981), “Evolutionary trees from DNA sequences: a maximum likelihood approach”, *Journal of Molecular Evolution*, Vol. 17 No. 6, pp. 368–376.

Feuda, R., Dohrmann, M., Pett, W., Philippe, H., Rota-Stabelli, O., Lartillot, N., Wörheide, G., et al. (2017), “Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals”, *Current Biology: CB*, doi.org/10.1016/j.cub.2017.11.008.

Fitch, W.M. and Margoliash, E. (1967), “Construction of phylogenetic trees”, *Science*, Vol. 155 No. 3760, pp. 279–284.

Floudas, D., Binder, M., Riley, R., Barry, K., Blanchette, R.A., Henrissat, B., Martínez, A.T., et al. (2012), “The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes”, *Science*, Vol. 336 No. 6089, pp. 1715–1719.

Foster, C.S.P. and Ho, S.Y.W. (2017), “Strategies for Partitioning Clock Models in Phylogenomic Dating: Application to the Angiosperm Evolutionary Timescale”, *Genome Biology and Evolution*, Vol. 9 No. 10, pp. 2752–2763.

Fourment, M. and Gibbs, M.J. (2006), “PATRISTIC: a program for calculating patristic distances and graphically comparing the components of genetic change”, *BMC Evolutionary Biology*, Vol. 6, p. 1.

Frýda, J., Nützel, A. and Wagner, P.J. (2008), “Paleozoic gastropoda”, *Phylogeny and Evolution of the Mollusca*, fzp.czu.cz, available at: http://fzp.czu.cz/~fryda/pdfs/189.pdf.

Fuchs, D., Bracchi, G. and Weis, R. (2009), “New Octopods (Cephalopoda: Coleoidea) from the Late Cretaceous (upper Cenomanian) of Hâkel and Hâdjoula, Lebanon”, *Palaeontology*, Vol. 52 No. 1, pp. 65–81.

Fuchs, D., Iba, Y., Ifrim, C., Nishimura, T., Kennedy, W.J., Keupp, H., Stinnesbeck, W., et al. (2013), “Longibelus gen. nov., a new Cretaceous coleoid genus linking Belemnoidea and early Decabrachia”, *Palaeontology*, Vol. 56 No. 5, pp. 1081–1106.

Fuchs, D., Iba, Y., Tischlinger, H., Keupp, H. and Klug, C. (2015), “The locomotion system of Mesozoic Coleoidea (Cephalopoda) and its phylogenetic significance”, *Lethaia*, available at:https://doi.org/10.1111/let.12155.

Fuchs, D., Keupp, H. and Schweigert, G. (2013), “First record of a complete arm crown of the Early Jurassic coleoid Loligosepia (Cephalopoda)”, *Paläontologische Zeitschrift*, Springer Berlin Heidelberg, Vol. 87 No. 3, pp. 431–435.

Fuchs, D., Keupp, H., Trask, P. and Tanabe, K. (2012), “Taxonomy, morphology and phylogeny of Late Cretaceous spirulid coleoids (Cephalopoda) from Greenland and Canada”, *Palaeontology*, Blackwell Publishing Ltd, Vol. 55 No. 2, pp. 285–303.

Fuchs, D. and Weis, R. (2008), “Taxonomy, morphology and phylogeny of Lower Jurassic loligosepiid coleoids (Cephalopoda)”, *Neues Jahrbuch Fur Geologie Und Palaeontologie - Abhandlungen*, Vol. 249 No. 1, pp. 93–112.

Gabaldón, T. and Koonin, E.V. (2013), “Functional and evolutionary implications of gene orthology”, *Nature Reviews. Genetics*, Vol. 14 No. 5, pp. 360–366.

Garwood, R.J. and Dunlop, J. (2014), “Three-dimensional reconstruction and the phylogeny of extinct chelicerate orders”, *PeerJ*, Vol. 2, p. e641.

Garwood, R.J. and Dunlop, J.A. (2011), “Morphology and systematics of anthracomartidae (Arachnida: Trigonotarbida)”, *Palaeontology*, Blackwell Publishing Ltd, Vol. 54 No. 1, pp. 145–161.

Gavryushkina, A., Heath, T.A., Ksepka, D.T., Stadler, T., Welch, D. and Drummond, A.J. (2017), “Bayesian Total-Evidence Dating Reveals the Recent Crown Radiation of

Penguins”, *Systematic Biology*, Vol. 66 No. 1, pp. 57–73.

Gibling, M.R. and Davies, N.S. (2012), “Palaeozoic landscapes shaped by plant evolution”, *Nature Geoscience*, Vol. 5 No. 2, pp. 99–105.

Gingerich, O. (1973), “From Copernicus to Kepler: Heliocentrism as Model and as Reality”, *Proceedings of the American Philosophical Society*, American Philosophical Society, Vol. 117 No. 6, pp. 513–522.

Giribet, G., Edgecombe, G.D., Wheeler, W.C. and Babbitt, C. (2002), “Phylogeny and systematic position of Opiliones: a combined analysis of chelicerate relationships using morphological and molecular data”, *Cladistics: The International Journal of the Willi Hennig Society*, Vol. 18 No. 1, pp. 5–70.

Glasby, C.J. and Timm, T. (2008), “Global diversity of polychaetes (Polychaeta; Annelida) in freshwater”, *Hydrobiologia*, Springer Netherlands, Vol. 595 No. 1, pp. 107–115.

Goldman, N. (1998), “Phylogenetic information and experimental design in molecular systematics”, *Proceedings. Biological Sciences / The Royal Society*, Vol. 265 No. 1407, pp. 1779–1786.

Goloboff, P.A. (2003), “Parsimony, likelihood, and simplicity”, *Cladistics: The International Journal of the Willi Hennig Society*, Blackwell Publishing Ltd, Vol. 19 No. 2, pp. 91–103.

Goloboff, P.A., Torres, A. and Arias, J.S. “Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology”, *Cladistics: The International Journal of the Willi Hennig Society*, available at:https://doi.org/10.1111/cla.12205.

Gosling, E. (2015), “Phylogeny and evolution of bivalve molluscs”, *Marine Bivalve Molluscs*, John Wiley & Sons, Ltd, pp. 1–11.

Gould, S. and Eldredge, N. (1993), “Punctuated equilibrium comes of age”, *Nature*, Springer, Vol. 366 No. 6452, pp. 223–227.

Gould, S.J. and Eldredge, N. (1977), “Punctuated equilibria: the tempo and mode of evolution reconsidered”, *Paleobiology*, Cambridge University Press, Vol. 3 No. 2, pp.

115–151.

Gould, S.J. and Lewontin, R.C. (1979), “The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme”, *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character. Royal Society* , Vol. 205 No. 1161, pp. 581–598.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., et al. (2011), “Full-length transcriptome assembly from RNA-Seq data without a reference genome”, *Nature Biotechnology*, Vol. 29 No. 7, pp. 644–652.

Grant, S.W., Knoll, A.H. and Germs, G.J. (1991), “Probable calcified metaphytes in the latest Proterozoic Nama Group, Namibia: origin, diagenesis, and implications”, *Journal of Paleontology*, Vol. 65 No. 1, pp. 1–18.

van Groenigen, J.W., Lubbers, I.M., Vos, H.M.J., Brown, G.G., De Deyn, G.B. and van Groenigen, K.J. (2014), “Earthworms increase plant production: a meta-analysis”, *Scientific Reports*, Vol. 4, p. 6365.

Guindon, S., Dufayard, J.F., Hordijk, W., Lefort, V. and Gascuel, O. (2009), “PhyML: fast and accurate phylogeny reconstruction by maximum likelihood”, *Infection Genetics and Evolution*, Vol. 9, pp. 384–385.

Gundale, M.J. (2002), “Influence of Exotic Earthworms on the Soil Organic Horizon and the Rare Fern Botrychium mormo”, *Conservation Biology: The Journal of the Society for Conservation Biology*, Blackwell Science Inc, Vol. 16 No. 6, pp. 1555–1561.

Haas, B.J. and Papanicolaou, A. (2016), “TransDecoder (find coding regions within transcripts)”.

von Haeseler, A., Minh, B.Q., Nguyen, L.T. and Schmidt, H.A. (2012), “IQ-TREE version 0.9. 3 Efficient phylogenetic tree reconstruction and ultrafast bootstrap approximation”.

Halanych, K.M., Bacheller, J.D., Aguinaldo, A.M., Liva, S.M., Hillis, D.M. and Lake, J.A. (1995), “Evidence from 18S ribosomal DNA that the lophophorates are protostome animals”, *Science*, Vol. 267 No. 5204, pp. 1641–1643.

Hale, C.M., Frelich, L.E. and Reich, P.B. (2006), “Changes in hardwood forest understory plant communities in response to European earthworm invasions”, *Ecology*, Vol. 87 No. 7, pp. 1637–1649.

Harvey, M.S. (2002), “The neglected cousins: what do we know about the smaller arachnid orders?”, *The Journal of Arachnology*, BioOne, Vol. 30 No. 2, pp. 357–372.

Haug, C. and Haug, J.T. (2017), “Methods and Practices in Paleo-Evo-Devo”, in de la Rosa, L.N. and Müller, G. (Eds.), *Evolutionary Developmental Biology*, Springer International Publishing, pp. 1–14.

Hejnol, A., Obst, M., Stamatakis, A., Ott, M., Rouse, G.W., Edgecombe, G.D., Martinez, P., et al. (2009), “Assessing the root of bilaterian animals with scalable phylogenomic methods”, *Proceedings. Biological Sciences / The Royal Society*, Vol. 276 No. 1677, pp. 4261–4270.

Herendeen, E.P.S. (2015), “Wetland-Dryland Vegetational Dynamics in the Pennsylvanian Ice Age Tropics”, *International Journal of Plant Sciences*, University of Chicago PressChicago, IL, available at:https://doi.org/10.1086/675235.

Hermans, C.O. (1969), “The Systematic Position of the Archiannelida”, *Systematic Biology*, Oxford University Press, Vol. 18 No. 1, pp. 85–102.

Herre, E.A., Knowlton, N., Mueller, U.G. and Rehner, S.A. (1999), “The evolution of mutualisms: exploring the paths between conflict and cooperation”, *Trends in Ecology & Evolution*, Vol. 14 No. 2, pp. 49–53.

Hibbett, D.S. and Matheny, P.B. (2009), “The relative ages of ectomycorrhizal mushrooms and their plant hosts estimated using Bayesian relaxed molecular clock analyses”, *BMC Biology*, Vol. 7, p. 13.

Hints, O. and Eriksson, M.E. (2007), “Diversification and biogeography of scolecodont-bearing polychaetes in the Ordovician”, *Palaeogeography, Palaeoclimatology, Palaeoecology*, Vol. 245 No. 1–2, pp. 95–114.

Hirst, S. (1923), “XLVI.—On some Arachnid remains from the Old Red Sandstone (Rhynie

Chert Bed, Aberdeenshire)”, *Annals and Magazine of Natural History*, Taylor & Francis, Vol. 12 No. 70, pp. 455–474.

Höhna, S., Landis, M.J., Heath, T.A., Boussau, B., Lartillot, N., Moore, B.R., Huelsenbeck, J.P., et al. (2016), “RevBayes: A flexible framework for Bayesian inference of phylogeny”, *Systematic Biology*, Vol. 64, pp. 726–736.

Holland, P.W.H., Marlétaz, F., Maeso, I., Dunwell, T.L. and Paps, J. (2017), “New genes from old: asymmetric divergence of gene duplicates and the evolution of development”, *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, Vol. 372 No. 1713, available at:https://doi.org/10.1098/rstb.2015.0480.

Huang, Y., Niu, B., Gao, Y., Fu, L. and Li, W. (2010), “CD-HIT Suite: a web server for clustering and comparing biological sequences”, *Bioinformatics* , Vol. 26 No. 5, pp. 680–682.

Huelsenbeck, J.P., Larget, B., Miller, R.E. and Ronquist, F. (2002), “Potential applications and pitfalls of Bayesian inference of phylogeny”, *Systematic Biology*, Vol. 51 No. 5, pp. 673–688.

Hutchinson, S.A. and Kamel, M. (1956), “The effect of earthworms on the dispersal of soil fungi”, *Journal of Soil Science*, Blackwell Publishing Ltd, Vol. 7 No. 2, pp. 213–218.

Iba, Y., Mutterlose, J., Tanabe, K., Sano, S.-I., Misaki, A. and Terabe, K. (2011), “Belemnite extinction and the origin of modern cephalopods 35 Ma prior to the Cretaceous−Paleogene event”, *Geology*, Vol. 39 No. 5, pp. 483–486.

Iba, Y., Sano, S.-I., Mutterlose, J. and Kondo, Y. (2012), “Belemnites originated in the Triassic—A new look at an old group”, *Geology*, Vol. 40 No. 10, pp. 911–914.

Ippolitov, A.P., Vinn, O., Kupriyanova, E.K. and Jäger, M. (2014), “Written in stone: history of serpulid polychaetes through time”, *Memoirs of Museum Victoria*, Vol. 71.

James, S.W. and Davidson, S.K. (2012), “Molecular phylogeny of earthworms (Annelida : Crassiclitellata) based on 28S, 18S and 16S gene sequences”, *Invertebrate Systematics*, Vol. 26 No. 2, pp. 213–229.

Jeffroy, O., Brinkmann, H., Delsuc, F. and Philippe, H. (2006), “Phylogenomics: the beginning of incongruence?”, *Trends in Genetics: TIG*, Vol. 22 No. 4, pp. 225–231.

Jerzy Dzik, W. and Korn, D. (n.d.). “Devonian ancestors of Nautilus”, *Paläontologische Zeitschrift*, Springer-Verlag, Vol. 66 No. 1-2, pp. 81–98.

Jukes, T.H. and Cantor, C.R. (1969), “Chapter 24 - Evolution of Protein Molecules”, in Munro, H.N. (Ed.), *Mammalian Protein Metabolism*, Academic Press, pp. 21–132.

Kalmar, A. and Currie, D.J. (2010), “The completeness of the continental fossil record and its impact on patterns of diversification”, *Paleobiology*, The Paleontological Society, Vol. 36 No. 1, pp. 51–60.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A. and Jermiin, L.S. (2017), “ModelFinder: fast model selection for accurate phylogenetic estimates”, *Nature Methods*, Vol. 14 No. 6, pp. 587–589.

Kimura, M. (1968), “Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles”, *Genetical Research*, Vol. 11 No. 3, pp. 247–269.

Kluessendorf, J. and Doyle, P. (2000), “Pohlsepia mazonensis, an early ‘octopus’ from the carboniferous of Illinois, USA”, *Palaeontology*, Vol. 43, pp. 919–926.

Klug, C., Frey, L., Korn, D., Jattiot, R. and Rücklin, M. (2016), “The oldest Gondwanan cephalopod mandibles (Hangenberg Black Shale, Late Devonian) and the mid-Palaeozoic rise of jaws”, *Palaeontology*, Wiley Online Library, available at: http://onlinelibrary.wiley.com/doi/10.1111/pala.12248/full.

Klug, C., Korn, D., De Baets, K., Kruta, I. and Mapes, R.H. (2015), *Ammonoid Paleobiology: From Macroevolution to Paleogeography*, Springer.

Klug, C., Kröger, B., Kiessling, W., Mullins, G.L., Servais, T., Frýda, J., Korn, D., et al. (2010), “The Devonian nekton revolution”, *Lethaia*, Blackwell Publishing Ltd, Vol. 43 No. 4, pp. 465–477.

Klug, C., Schweigert, G., Fuchs, D., Kruta, I. and Tischlinger, H. (2016), “Adaptations to squid-style high-speed swimming in Jurassic belemnitids", *Biology Letters*, Vol. 12 No. 1.

Kocot, K.M., Cannon, J.T., Todt, C., Citarella, M.R., Kohn, A.B., Meyer, A., Santos, S.R., et al. (2011), "Phylogenomics reveals deep molluscan relationships", *Nature*, Vol. 477 No. 7365, pp. 452–456.

Kowal, J., Biyani, N., Chami, M., Scherer, S., Rzepiela, A.J., Baumgartner, P., Upadhyay, V., et al. (2018), "High-Resolution Cryoelectron Microscopy Structure of the Cyclic Nucleotide-Modulated Potassium Channel MloK1 in a Lipid Bilayer", *Structure*, Vol. 26 No. 1, pp. 20–27.e3.

Kröger, B. (2005), "Adaptive evolution in Paleozoic coiled cephalopods", *Paleobiology*, Cambridge Univ Press, Vol. 31 No. 02, pp. 253–268.

Kröger, B. and Mapes, R.H. (2007), "On the origin of bactritoids (Cephalopoda)", *Paläontologische Zeitschrift*, Springer, Vol. 81 No. 3, pp. 316–327.

Kröger, B., Vinther, J. and Fuchs, D. (2011), "Cephalopod origin and evolution: a congruent picture emerging from fossils, development and molecules", *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, Vol. 33 No. 8, pp. 602–613.

Kück, P. and Meusemann, K. (2010), "FASconCAT: Convenient handling of data matrices", *Molecular Phylogenetics and Evolution*, Vol. 56 No. 3, pp. 1115–1118.

Kumar, S. (2005), "Molecular clocks: four decades of evolution", *Nature Reviews. Genetics*, Vol. 6 No. 8, pp. 654–662.

Lanfear, R., Calcott, B., Ho, S.Y.W. and Guindon, S. (2012), "Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses", *Molecular Biology and Evolution*, Vol. 29 No. 6, pp. 1695–1701.

Lartillot, N., Rodrigue, N., Stubbs, D. and Richer, J. (2013), "PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment", *Systematic Biology*, Vol. 62 No. 4, pp. 611–615.

Lavelle, P. (2011), "Earthworms as Ecosystem Engineers", in Gliński, J., Horabik, J. and Lipiec, J. (Eds.), *Encyclopedia of Agrophysics*, Springer Netherlands, pp. 233–235.

Lavelle, P., Bignell, D., Lepage, M., Wolters, W., Roger, P., Ineson, P., Heal, O.W., et al.

(1997), “Soil function in a changing world: the role of invertebrate ecosystem engineers”, *European Journal of Soil Biology*, Elsevier, Vol. 33 No. 4, pp. 159–193.

Lawrence, B., Fisk, M.C., Fahey, T.J. and Suárez, E.R. (2003), “Influence of non-native earthworms on mycorrhizal colonization of sugar maple (Acer saccharum)”, *The New Phytologist*, Blackwell Science Ltd, Vol. 157 No. 1, pp. 145–153.

Legg, D.A. (2014), “Sanctacaris uncata: the oldest chelicerate (Arthropoda)”, *Die Naturwissenschaften*, Vol. 101 No. 12, pp. 1065–1073.

Lehmann, J., Rillig, M.C., Thies, J., Masiello, C.A., Hockaday, W.C. and Crowley, D. (2011/9), “Biochar effects on soil biota – A review”, *Soil Biology & Biochemistry*, Vol. 43 No. 9, pp. 1812–1836.

Lemtiri, A., Colinet, G., Alabi, T., Cluzeau, D., Zirbes, L., Haubruge, É. and Francis, F. (2014), “Impacts of earthworms on soil components and dynamics. A review”, *Biotechnologie, Agronomie, Société et Environnement*, Les Presses Agronomiques de Gembloux, Vol. 18 No. 1, p. 121.

Lepage, T., Bryant, D., Philippe, H. and Lartillot, N. (2007), “A general comparison of relaxed molecular clock models”, *Molecular Biology and Evolution*, Vol. 24 No. 12, pp. 2669–2680.

Lepage, T., Lawi, S., Tupper, P. and Bryant, D. (2006), “Continuous and tractable models for the variation of evolutionary rates”, *Mathematical Biosciences*, Vol. 199 No. 2, pp. 216–233.

Le, S.Q. and Gascuel, O. (2008), “An improved general amino acid replacement matrix”, *Molecular Biology and Evolution*, Vol. 25 No. 7, pp. 1307–1320.

Lindgren, A.R. (2010), “Molecular inference of phylogenetic relationships among Decapodiformes (Mollusca: Cephalopoda) with special focus on the squid Order Oegopsida”, *Molecular Phylogenetics and Evolution*, Vol. 56 No. 1, pp. 77–90.

Lindgren, A.R., Pankey, M.S., Hochberg, F.G. and Oakley, T.H. (2012), “A multi-gene phylogeny of Cephalopoda supports convergent morphological evolution in association with multiple habitat shifts in the marine environment”, *BMC Evolutionary Biology*, Vol. 12 No. 1, p. 129.

Liò, P. and Goldman, N. (1998), “Models of molecular evolution and phylogeny”, *Genome Research*, Vol. 8 No. 12, pp. 1233–1244.

Lovelock, J.E. and Margulis, L. (1974), “Atmospheric homeostasis by and for the biosphere: the gaia hypothesis”, *Tell’Us*, Blackwell Publishing Ltd, Vol. 26 No. 1-2, pp. 2–10.

Lozano-Fernandez, J., Carton, R., Tanner, A.R., Puttick, M.N., Blaxter, M., Vinther, J., Olesen, J., et al. (2016), “A molecular palaeobiological exploration of arthropod terrestrialization”, *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, Vol. 371 No. 1699, doi.org/10.1098/rstb.2015.0133.

Lubbers, I.M., van Groenigen, K.J., Fonte, S.J., Six, J., Brussaard, L. and van Groenigen, J.W. (2013), “Greenhouse-gas emissions from soils increased by earthworms”, *Nature Climate Change*, Nature Research, Vol. 3 No. 3, pp. 187–194.

Mans, B.J. and Neitz, A.W.H. (2004), “Adaptation of ticks to a blood-feeding environment: evolution from a functional perspective”, *Insect Biochemistry and Molecular Biology*, Vol. 34 No. 1, pp. 1–17.

Manum, S.B., Bose, M.N. and Sawyer, R.T. (1991), “Clitellate cocoons in freshwater deposits since the Triassic”, *Zoologica Scripta*, Blackwell Publishing Ltd, Vol. 20 No. 4, pp. 347–366.

Martin, M.W., Grazhdankin, D.V., Bowring, S.A., Evans, D.A., Fedonkin, M.A. and Kirschvink, J.L. (2000), “Age of Neoproterozoic bilatarian body and trace fossils, White Sea, Russia: implications for metazoan evolution”, *Science*, Vol. 288 No. 5467, pp. 841–845.

Martin, P., Martinez-Ansemil, E., Pinder, A., Timm, T. and Wetzel, M.J. (2007), “Global diversity of oligochaetous clitellates (‘Oligochaeta’; Clitellata) in freshwater”, *Hydrobiologia*, Springer Netherlands, Vol. 595 No. 1, pp. 117–127.

Mather, J.A. and Kuba, M.J. (2013a), “The cephalopod specialties: complex nervous system, learning, and cognition”, *Canadian Journal of Zoology*, NRC Research Press, Vol. 91 No. 6, pp. 431–449.

Mather, J.A. and Kuba, M.J. (2013b), “The cephalopod specialties: complex nervous system, learning, and cognition 1”, *Canadian Journal of Zoology*, NRC Research Press, available at: http://www.nrcresearchpress.com/doi/abs/10.1139/cjz-2013-0009.

Mayr, E. (1991), *One Long Argument: Charles Darwin and the Genesis of Modern Evolution Thought*, Harvard University Press.

Milleret, R., Le Bayon, R.-C. and Gobat, J.-M. (2009), “Root, mycorrhiza and earthworm interactions: their effects on soil structuring processes, plant and soil nutrient concentration and plant biomass”, *Plant and Soil*, Springer Netherlands, Vol. 316 No. 1-2, pp. 1–12.

Miller, K.G., Kominz, M.A., Browning, J.V., Wright, J.D., Mountain, G.S., Katz, M.E., Sugarman, P.J., et al. (2005), “The Phanerozoic record of global sea-level change”, *Science*, Vol. 310 No. 5752, pp. 1293–1298.

Minh, B.Q., Nguyen, M.A.T. and von Haeseler, A. (2013), “Ultrafast approximation for phylogenetic bootstrap”, *Molecular Biology and Evolution*, Vol. 30 No. 5, pp. 1188–1195.

Mitchell, E.G., Kenchington, C.G., Liu, A.G., Matthews, J.J. and Butterfield, N.J. (2015), “Reconstructing the reproductive mode of an Ediacaran macro-organism”, *Nature*, Vol. 524 No. 7565, pp. 343–346.

Montañez, I.P., McElwain, J.C., Poulsen, C.J., White, J.D., DiMichele, W.A., Wilson, J.P., Griggs, G., et al. (2016), “Climate, pCO2 and terrestrial carbon cycle linkages during late Palaeozoic glacial–interglacial cycles”, *Nature Geoscience*, Vol. 9 No. 11, pp. 824–828.

Montañez, I.P., Tabor, N.J., Niemeier, D., Dimichele, W.A., Frank, T.D., Fielding, C.R., Isbell, J.L., et al. (2007), “CO2-forced climate and vegetation instability during Late Paleozoic deglaciation”, *Science*, Vol. 315 No. 5808, pp. 87–91.

Mutvei, H., Zhang, Y.-B. and Dunca, E. (2007), “Late Cambrian plectronocerid nautiloids and their role in cephalopod evolution”, *Palaeontology*, Blackwell Publishing Ltd, Vol. 50 No.

6, pp. 1327–1333.

Mwinyi, A., Meyer, A., Bleidorn, C., Lieb, B., Bartolomaeus, T. and Podsiadlowski, L. (2009), “Mitochondrial genome sequence and gene order of Sipunculus nudus give additional support for an inclusion of Sipuncula into Annelida”, *BMC Genomics*, Vol. 10, p. 27.

Near, T.J., Eytan, R.I., Dornburg, A., Kuhn, K.L., Moore, J.A., Davis, M.P., Wainwright, P.C., et al. (2012), “Resolution of ray-finned fish phylogeny and timing of diversification”, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 109 No. 34, pp. 13698–13703.

Nelsen, M.P., DiMichele, W.A., Peters, S.E. and Boyce, C.K. (2016), “Delayed fungal evolution did not cause the Paleozoic peak in coal production”, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 113 No. 9, pp. 2442–2447.

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A. and Minh, B.Q. (2015), “IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies”, *Molecular Biology and Evolution*, Vol. 32 No. 1, pp. 268–274.

Niklas, K.J., Tiffney, B.H. and Knoll, A.H. (1983), “Patterns in vascular land plant diversification”, *Nature*, Nature Publishing Group, Vol. 303 No. 5918, pp. 614–616.

Norton, R.A., Bonamo, P.M., Grierson, J.D. and Shear, W.A. (1988), “Oribatid mite fossils from a terrestrial Devonian deposit near Gilboa, New York”, *Journal of Paleontology*, Cambridge University Press, Vol. 62 No. 2, pp. 259–269.

Nutzel, A. and Bandel, K. (2000), “Goniasmidae and Orthonemidae: two new families of the Palaeozoic Caenogastropoda (Mollusca, Gastropoda)”, *Neues Jahrbuch Fur*.

Nuzzo, V.A., Maerz, J.C. and Blossey, B. (2009), “Earthworm invasion as the driving force behind plant invasion and community change in northeastern North American forests”, *Conservation Biology: The Journal of the Society for Conservation Biology*, Vol. 23 No. 4, pp. 966–974.

O’Malley, M.A. (2017), “From endosymbiosis to holobionts: Evaluating a conceptual legacy”, *Journal of Theoretical Biology*, Vol. 434 No. Supplement C, pp. 34–41.

Omnès, R. (1992), “Consistent interpretations of quantum mechanics”, *Reviews of Modern Physics*, APS, Vol. 64 No. 2, p. 339.

O’Reilly, J.E., Puttick, M.N., Parry, L., Tanner, A.R., Tarver, J.E., Fleming, J., Pisani, D., et al. (2016), “Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data”, *Biology Letters*, Vol. 12 No. 4, available at:https://doi.org/10.1098/rsbl.2016.0081.

O’Reilly, J.E., dos Reis, M. and Donoghue, P.C.J. (2015), “Dating Tips for Divergence-Time Estimation”, *Trends in Genetics: TIG*, Vol. 31 No. 11, pp. 637–650.

Orrhage, L. (2001), “On the anatomy of the central nervous system and the morphological value of the anterior end appendages of Ampharetidae, Pectinariidae and Terebellidae (Polychaeta)”, *Acta Zoologica* , Blackwell Science Ltd, Vol. 82 No. 1, pp. 57–71.

Packard, A. (1972), “Cephalopods and fish: the limits of convergence”, *Biological Reviews of the Cambridge Philosophical Society*, Blackwell Publishing Ltd, Vol. 47 No. 2, pp. 241–307.

Parham, J.F., Donoghue, P.C.J., Bell, C.J., Calway, T.D., Head, J.J., Holroyd, P.A., Inoue, J.G., et al. (2011), “Best practices for justifying fossil calibrations”, *Systematic Biology*, Oxford University Press, Vol. 61 No. 2, pp. 346–359.

Parkhaev, P.Y. (2008), “The Early Cambrian Radiation of Mollusca”, in Ponder, W. (Ed.), *Phylogeny and Evolution of the Mollusca*, University of California Press, pp. 33–69.

Parry, L.A., Smithwick, F., Nordén, K.K., Saitta, E.T., Lozano-Fernandez, J., Tanner, A.R., Caron, J.-B., et al. (2017), “Soft-Bodied Fossils Are Not Simply Rotten Carcasses - Toward a Holistic Understanding of Exceptional Fossil Preservation: Exceptional Fossil Preservation Is Complex and Involves the Interplay of Numerous Biological and Geological Processes”, *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, available at:https://doi.org/10.1002/bies.201700167.

Parry, L., Tanner, A. and Vinther, J. (2014), “The origin of annelids”, *Palaeontology*, available at:https://doi.org/10.1111/pala.12129.

Paudel, S., Wilson, G.W.T., MacDonald, B., Longcore, T. and Loss, S.R. (2016), “Predicting spatial extent of invasive earthworms on an oceanic island”, *Diversity & Distributions*, Vol. 22 No. 10, pp. 1013–1023.

Penney, D. (2003), “Does the fossil record of spiders track that of their principal prey, the insects?”, *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, Royal Society of Edinburgh Scotland Foundation, Vol. 94 No. 3, pp. 275–281.

Pepato, A.R. and Klimov, P.B. (2015), “Origin and higher-level diversification of acariform mites--evidence from nuclear ribosomal genes, extensive taxon sampling, and secondary structure alignment”, *BMC Evolutionary Biology*, BioMed Central, Vol. 15 No. 1, p. 178.

Perry, M.L. (1995), “Preliminary Description of a New Fossil Scorpion from the Middle Eocene, Green River Formation, Rio Blanco County, Colorado”, archives.datapages.com, available at: http://archives.datapages.com/data/grand-junction-geo-soc/data/013/013001/131_gjgs-sp0130131.htm.

Philippe, H., Brinkmann, H., Copley, R.R., Moroz, L.L., Nakano, H., Poustka, A.J., Wallberg, A., et al. (2011), “Acoelomorph flatworms are deuterostomes related to Xenoturbella”, *Nature*, Vol. 470 No. 7333, pp. 255–258.

Philippe, H., Delsuc, F., Brinkmann, H. and Lartillot, N. (2005), “Phylogenomics”, *Annual Review of Ecology, Evolution, and Systematics*, Annual Reviews, Vol. 36, pp. 541–562.

Philippe, H. and Roure, B. (2011), “Difficult phylogenetic questions: more data, maybe; better methods, certainly”, *BMC Biology*, Vol. 9 No. 1, p. 91.

Pisani, D. and Liu, A.G. (2015), “Animal Evolution: Only Rocks Can Set the Clock”, *Current Biology: CB*, Vol. 25 No. 22, pp. R1079–81.

Pisani, D., Pett, W., Dohrmann, M., Feuda, R., Rota-Stabelli, O., Philippe, H., Lartillot, N., et al. (2015), “Genomic data do not support comb jellies as the sister group to all other animals”, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 112 No. 50, pp. 15402–15407.

Poinar, G. (2008), “Palaeosiro burmanicum n. gen., n. sp., a fossil Cyphophthalmi (Arachnida: Opiliones: Sironidae) in Early Cretaceous Burmese amber”, *Advances in Arachnology and Developmental Biology. Papers Dedicated to Prof. Dr. Bozidar Curcic. Vienna, Belgrade, Sofia: Faculty of Life Sciences, University of Vienna, and Serbian Academy of Sciences and Arts*, pp. 267–274.

Poinar, G., Jr and Brown, A.E. (2003), “A new genus of hard ticks in Cretaceous Burmese amber (Acari: Ixodida: Ixodidae)”, *Systematic Parasitology*, Vol. 54 No. 3, pp. 199–205.

Pollierer, M.M., Langel, R., Körner, C., Maraun, M. and Scheu, S. (2007), “The underestimated importance of belowground carbon input for forest soil animal food webs”, *Ecology Letters*, Vol. 10 No. 8, pp. 729–736.

Poulsen, C.J., Pollard, D., Montañez, I.P. and Rowley, D. (2007), “Late Paleozoic tropical climate response to Gondwanan deglaciation”, *Geology*, Vol. 35 No. 9, pp. 771–774.

Price, S.A., Bininda-Emonds, O.R.P. and Gittleman, J.L. (2005), “A complete phylogeny of the whales, dolphins and even-toed hoofed mammals (Cetartiodactyla)”, *Biological Reviews of the Cambridge Philosophical Society*, Vol. 80 No. 3, pp. 445–473.

Puttick, M.N., O’Reilly, J.E., Tanner, A.R., Fleming, J.F., Clark, J., Holloway, L., Lozano-Fernandez, J., et al. (2017), “Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data”, *Proceedings. Biological Sciences / The Royal Society*, Vol. 284 No. 1846, available at:https://doi.org/10.1098/rspb.2016.2290.

Regier, J.C., Shultz, J.W. and Kambic, R.E. (2005), “Pancrustacean phylogeny: hexapods are terrestrial crustaceans and maxillopods are not monophyletic”, *Proceedings. Biological Sciences / The Royal Society*, Vol. 272 No. 1561, pp. 395–401.

Regier, J.C., Shultz, J.W., Zwick, A., Hussey, A., Ball, B., Wetzer, R., Martin, J.W., et al.

(2010), “Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences”, *Nature*, Vol. 463 No. 7284, pp. 1079–1083.

Retallack, G.J. (1997), “Palaeosols in the upper Narrabeen group of New South Wales as evidence of Early Triassic palaeoenvironments without exact modern analogues*”, *Australian Journal of Earth Sciences*, Taylor & Francis, Vol. 44 No. 2, pp. 185–201.

Retallack, G.J. (2008), *Soils of the Past: An Introduction to Paleopedology*, John Wiley & Sons.

von Reumont, B.M., Jenner, R.A., Wills, M.A., Dell’ampio, E., Pass, G., Ebersberger, I., Meyer, B., et al. (2012), “Pancrustacean phylogeny in the light of new phylogenomic data: support for Remipedia as the possible sister group of Hexapoda”, *Molecular Biology and Evolution*, Vol. 29 No. 3, pp. 1031–1045.

Robinson, J.M. (1990), “Lignin, land plants, and fungi: Biological evolution affecting Phanerozoic oxygen balance”, *Geology*, Vol. 18 No. 7, pp. 607–610.

Roeding, F., Borner, J., Kube, M., Klages, S., Reinhardt, R. and Burmester, T. (2009), “A 454 sequencing approach for large scale phylogenomic analysis of the common emperor scorpion (Pandinus imperator)”, *Molecular Phylogenetics and Evolution*, Vol. 53 No. 3, pp. 826–834.

Ronquist, F., Klopfstein, S., Vilhelmsen, L., Schulmeister, S., Murray, D.L. and Rasnitsyn, A.P. (2012), “A total-evidence approach to dating with fossils, applied to the early radiation of the hymenoptera”, *Systematic Biology*, Vol. 61 No. 6, pp. 973–999.

Ronquist, F., Teslenko, M., Mark, P. van der, Ayres, D.L., Darling, A., Höhna, S., Larget, B., et al. (2012), “MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space”, *Systematic Biology*, Vol. 61 No. 3, pp. 539–542.

Rota-Stabelli, O., Campbell, L., Brinkmann, H., Edgecombe, G.D., Longhorn, S.J., Peterson, K.J., Pisani, D., et al. (2011), “A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata”, *Proceedings. Biological Sciences / The Royal Society*, Vol. 278 No. 1703, pp. 298–306.

Rota-Stabelli, O., Daley, A.C. and Pisani, D. (2013), “Molecular timetrees reveal a Cambrian colonization of land and a new scenario for ecdysozoan evolution”, *Current Biology: CB*, Vol. 23 No. 5, pp. 392–398.

Rota-Stabelli, O., Kayal, E., Gleeson, D., Daub, J., Boore, J.L., Telford, M.J., Pisani, D., et al. (2010), “Ecdysozoan mitogenomics: evidence for a common origin of the legged invertebrates, the Panarthropoda”, *Genome Biology and Evolution*, Vol. 2, pp. 425–440.

Roth, A.C.J., Gonnet, G.H. and Dessimoz, C. (2008), “Algorithm of OMA for large-scale orthology inference”, *BMC Bioinformatics*, Vol. 9, p. 518.

Rouse, G. and Pleijel, F. (2001), *Polychaetes*, OUP Oxford.

Rouse, G.W. and Fauchald, K. (1997), “Cladistics and polychaetes”, *Zoologica Scripta*, Blackwell Publishing Ltd, Vol. 26 No. 2, pp. 139–204.

Rousset, V., Pleijel, F., Rouse, G.W., Erséus, C. and Siddall, M.E. (2007), “A molecular phylogeny of annelids”, *Cladistics: The International Journal of the Willi Hennig Society*, Blackwell Publishing Ltd, Vol. 23 No. 1, pp. 41–63.

Rudkin, D.M., Young, G.A. and Nowlan, G.S. (2008), “The oldest horseshoe crab: a xiphosurid from the Late Ordovician konservat-lagerstatten deposits, Manitoba, Canada”, *Palaeontology*, Blackwell Publishing Ltd, Vol. 51 No. 1, pp. 1–9.

Sagan, L. (1967), “On the origin of mitosing cells”, *Journal of Theoretical Biology*, Vol. 14 No. 3, pp. 225–IN6.

Sanders, K.L. and Lee, M.S.Y. (2010), “Arthropod molecular divergence times and the Cambrian origin of pentastomids”, *Systematics and Biodiversity*, Taylor & Francis, Vol. 8 No. 1, pp. 63–74.

Sanfilippo, R., Rosso, A., Reitano, A. and Insacco, G. (2017), “First record of sabellid and serpulid polychaetes from the Permian of Sicily”, available at: https://www.app.pan.pl/archive/published/app62/app002882016.pdf.

Scholtz, G. and Kamenz, C. (2006), “The book lungs of Scorpiones and Tetrapulmonata (Chelicerata, Arachnida): Evidence for homology and a single terrestrialisation event of a common arachnid ancestor”, *Zoology* , Vol. 109 No. 1, pp. 2–13.

Schwab, D.B. and Moczek, A.P. (2017), “Evo-Devo and Niche Construction”, in de la Rosa, L.N. and Müller, G. (Eds.), *Evolutionary Developmental Biology*, Springer International Publishing, pp. 1–14.

Schweigert, G. and Fuchs, D. (2012), “First record of a true coleoid cephalopod from the Germanic Triassic (Ladinian)”, *Neues Jahrbuch Fur Geologie Und Palaeontologie - Abhandlungen*, Vol. 266 No. 1, pp. 19–30.

Selden, P.A. (2016), “Land Animals, Origins of”, in Kliman, R.M. (Ed.), *Encyclopedia of Evolutionary Biology*, Academic Press, Oxford, pp. 288–295.

Selden, P.A., Anderson, H.M. and Anderson, J.M. (2009), “A Review of the Fossil Record of Spiders (Araneae) with Special Reference to Africa, and Description of a New Specimen from the Triassic Molteno Formation of South Africa”, *African Invertebrates*, KwaZulu-Natal Museum, Vol. 50 No. 1, pp. 105–116.

Selden, P.A. and Gall, J.-C. (1992), “A Triassic mygalomorph spider from the northern Vosges, France”, The Palaeontological Association, available at: https://kuscholarworks.ku.edu/handle/1808/8350..

Selden, P.A. and Huang, D. (2010), “The oldest haplogyne spider (Araneae: Plectreuridae), from the Middle Jurassic of China”, *Die Naturwissenschaften*, Vol. 97 No. 5, pp. 449–459.

Selden, P.A., Shih, C. and Ren, D. (2011), “A golden orb-weaver spider (Araneae: Nephilidae: Nephila) from the Middle Jurassic of China”, *Biology Letters*, Vol. 7 No. 5, pp. 775–778.

Sharma, P.P., Fernández, R., Esposito, L.A., González-Santillán, E. and Monod, L. (2015), “Phylogenomic resolution of scorpions reveals multilevel discordance with morphological phylogenetic signal”, *Proceedings. Biological Sciences / The Royal Society*, Vol. 282 No. 1804, p. 20142953.

Sharma, P.P., Kaluziak, S.T., Pérez-Porro, A.R., González, V.L., Hormiga, G., Wheeler, W.C. and Giribet, G. (2014), “Phylogenomic interrogation of arachnida reveals systemic conflicts in phylogenetic signal”, *Molecular Biology and Evolution*, Vol. 31 No. 11, pp. 2963–2984.

Shultz, J.W. (1990), “Evolutionary morphology and phylogeny of Araghnida”, *Cladistics: The International Journal of the Willi Hennig Society*, Wiley Online Library, Vol. 6 No. 1, pp. 1–38.

Shultz, J.W. (2007), “A phylogenetic analysis of the arachnid orders based on morphological characters”, *Zoological Journal of the Linnean Society*, Oxford University Press, Vol. 150 No. 2, pp. 221–265.

Simpson, G.G. (1945), “Tempo and mode in evolution”, *Transactions of the New York Academy of Sciences*, Vol. 8, pp. 45–60.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M. and Birol, I. (2009), “ABySS: a parallel assembler for short read sequence data”, *Genome Research*, Vol. 19 No. 6, pp. 1117–1123.

Singh, V., Braddick, D. and Dhar, P.K. (2017), “Exploring the potential of genome editing CRISPR-Cas9 technology”, *Gene*, Vol. 599, pp. 1–18.

Smithwick, F.M., Nicholls, R., Cuthill, I.C. and Vinther, J. (2017), “Countershading and Stripes in the Theropod Dinosaur Sinosauropteryx Reveal Heterogeneous Habitats in the Early Cretaceous Jehol Biota”, *Current Biology: CB*, Vol. 27 No. 21, pp. 3337–3343.e2.

Sperling, E.A., Vinther, J., Moy, V.N., Wheeler, B.M., Sémon, M., Briggs, D.E.G. and Peterson, K.J. (2009), “MicroRNAs resolve an apparent conflict between annelid systematics and their fossil record”, *Proceedings. Biological Sciences / The Royal Society*, Vol. 276 No. 1677, pp. 4315–4322.

Stamatakis, A. (2014), “RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies”, *Bioinformatics* , Vol. 30 No. 9, pp. 1312–1313.

Stamos, D.N. (1996), “Popper, falsifiability, and evolutionary biology”, *Biology & Philosophy*, Kluwer Academic Publishers, Vol. 11 No. 2, pp. 161–191.

Steel, M. and Penny, D. (2000), “Parsimony, likelihood, and the role of models in molecular phylogenetics”, *Molecular Biology and Evolution*, Vol. 17 No. 6, pp. 839–850.

Struck, T.H., Golombek, A., Weigert, A., Franke, F.A., Westheide, W., Purschke, G., Bleidorn, C., et al. (2015), “The evolution of annelids reveals two adaptive routes to the interstitial realm”, *Current Biology: CB*, Vol. 25 No. 15, pp. 1993–1999.

Struck, T.H., Paul, C., Hill, N., Hartmann, S., Hösel, C., Kube, M., Lieb, B., et al. (2011), “Phylogenomic analyses unravel annelid evolution”, *Nature*, Vol. 471 No. 7336, pp. 95–98.

Struck, T.H., Westheide, W. and Purschke, G. (2002), “Progenesis in Eunicida (‘Polychaeta,’ Annelida)—separate evolutionary events? Evidence from molecular data”, *Molecular Phylogenetics and Evolution*, Vol. 25 No. 1, pp. 190–199.

Strugnell, J., Jackson, J., Drummond, A.J. and Cooper, A. (2006), “Divergence time estimates for major cephalopod groups: evidence from multiple genes”, *Cladistics: The International Journal of the Willi Hennig Society*, Vol. 22 No. 1, pp. 89–96.

Strugnell, J. and Nishiguchi, M.K. (2007), “Molecular phylogeny of coleoid cephalopods (Mollusca: Cephalopoda) inferred from three mitochondrial and six nuclear loci: a comparison of alignment, implied alignment and analysis methods”, *The Journal of Molluscan Studies*, Vol. 73 No. 4, pp. 399–410.

Sugitani, K., Mimura, K., Nagaoka, T., Lepot, K. and Takeuchi, M. (2013), “Microfossil assemblage from the 3400Ma Strelley Pool Formation in the Pilbara Craton, Western Australia: Results form a new locality”, *Precambrian Research*, Vol. 226 No. Supplement C, pp. 59–74.

Tanner, A.R., Fuchs, D., Winkelmann, I.E., Gilbert, M.T.P., Pankey, M.S., Ribeiro, Â.M., Kocot, K.M., et al. (2017), “Molecular clocks indicate turnover and diversification of modern coleoid cephalopods during the Mesozoic Marine Revolution”, *Proceedings. Biological Sciences / The Royal Society*, Vol. 284 No. 1850, available at:https://doi.org/10.1098/rspb.2016.2818.

Tavaré, S. (1986), “Some probabilistic and statistical problems in the analysis of DNA sequences”, *Lectures on Mathematics in the Life Sciences*, Vol. 17 No. 2, pp. 57–86.

Taylor, T.N., Remy, W. and Hass, H. (1994), “Allomyces in the Devonian”, *Nature*, Nature Publishing Group, Vol. 367 No. 6464, pp. 601–601.

Thorne, J.L. and Kishino, H. (2005), “Estimation of Divergence Times from Molecular Sequence Data”, *Statistical Methods in Molecular Evolution*, Springer New York, pp. 233–256.

Thorne, J.L., Kishino, H. and Painter, I.S. (1998), “Estimating the rate of evolution of the rate of molecular evolution”, *Molecular Biology and Evolution*, Vol. 15 No. 12, pp. 1647–1657.

Townsend, J.P. (2007), “Profiling phylogenetic informativeness”, *Systematic Biology*, Vol. 56 No. 2, pp. 222–231.

Townsend, J.P., Su, Z. and Tekle, Y.I. (2012), “Phylogenetic signal and noise: predicting the power of a data set to resolve phylogeny”, *Systematic Biology*, Vol. 61 No. 5, pp. 835–849.

Vaidya, G., Lohman, D.J. and Meier, R. (2011), “SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information”, *Cladistics: The International Journal of the Willi Hennig Society*, Wiley Online Library, Vol. 27 No. 2, pp. 171–180.

Vermeij, G.J. (1977), “The Mesozoic marine revolution: evidence from snails, predators and grazers”, *Paleobiology*, Cambridge University Press, Vol. 3 No. 03, pp. 245–258.

Vermeij, G.J. (1987), *Evolution and Escalation: An Ecological History of Life*, Princeton University Press.

Vinther, J., Sperling, E.A., Briggs, D.E.G. and Peterson, K.J. (2012), “A molecular palaeobiological hypothesis for the origin of aplacophoran molluscs and their derivation from chiton-like ancestors”, *Proceedings of the Royal Society B-Biological Sciences*, Vol. 279 No. 1732, pp. 1259–1268.

Wang, B., Yeun, L.H., Xue, J.-Y., Liu, Y., Ané, J.-M. and Qiu, Y.-L. (2010), “Presence of three mycorrhizal genes in the common ancestor of land plants suggests a key role of mycorrhizas in the colonization of land by plants”, *The New Phytologist*, Vol. 186 No. 2, pp. 514–525.

Wang, Z.-Q. (2004), “A new Permian gnetalean cone as fossil evidence for supporting current molecular phylogeny”, *Annals of Botany*, Vol. 94 No. 2, pp. 281–288.

Warnke, K.M., Meyer, A., Ebner, B. and Lieb, B. (2011), “Assessing divergence time of Spirulida and Sepiida (Cephalopoda) based on hemocyanin sequences”, *Molecular Phylogenetics and Evolution*, Vol. 58 No. 2, pp. 390–394.

Warnock, R.C.M., Yang, Z. and Donoghue, P.C.J. (2012), “Exploring uncertainty in the calibration of the molecular clock”, *Biology Letters*, Vol. 8 No. 1, pp. 156–159.

Weigert, A. and Bleidorn, C. (2016), “Current status of annelid phylogeny”, *Organisms, Diversity & Evolution*, Springer Berlin Heidelberg, Vol. 16 No. 2, pp. 345–362.

Weigert, A., Helm, C., Meyer, M., Nickel, B., Arendt, D., Hausdorf, B., Santos, S.R., et al. (2014), “Illuminating the base of the annelid tree using transcriptomics”, *Molecular Biology and Evolution*, p. 80.

Wells, M.J. and O’Dor, R.K. (1991), “Jet Propulsion and the Evolution of the Cephalopods”, *Bulletin of Marine Science*, Vol. 49 No. 1-2, pp. 419–432.

Whelan, N.V., Kocot, K.M., Moroz, T.P., Mukherjee, K., Williams, P., Paulay, G., Moroz, L.L., et al. (2017), “Ctenophore relationships and their placement as the sister group to all other animals”, *Nature Ecology & Evolution*, Vol. 1 No. 11, pp. 1737–1746.

Wolfe, J.M., Daley, A.C., Legg, D.A. and Edgecombe, G.D. (2016), “Fossil calibrations for the arthropod Tree of Life”, *Earth-Science Reviews*, Vol. 160 No. Supplement C, pp. 43–110.

Wurst, S. (2010), “Effects of earthworms on above- and belowground herbivores”, *Applied Soil Ecology: A Section of Agriculture, Ecosystems & Environment*, Vol. 45 No. 3, pp. 123–130.

Wu, X., Cao, R., Wei, X., Xi, X., Shi, P., Eisenhauer, N. and Sun, S. (2017), “Soil drainage facilitates earthworm invasion and subsequent carbon loss from peatland soil”, *The Journal of Applied Ecology*, available at:https://doi.org/10.1111/1365-2664.12894.

Yang, N., Schützenmeister, K., Grubert, D., Jungkunst, H.F., Gansert, D., Scheu, S., Polle, A., et al. (2015), “Impacts of earthworms on nitrogen acquisition from leaf litter by arbuscular mycorrhizal ash and ectomycorrhizal beech trees”, *Environmental and*

Experimental Botany, Vol. 120, pp. 1–7.

Yang, Z. (1998), “On the best evolutionary rate for phylogenetic analysis”, *Systematic Biology*, Vol. 47 No. 1, pp. 125–133.

Yang, Z. and Rannala, B. (2006), “Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds”, *Molecular Biology and Evolution*, Vol. 23 No. 1, pp. 212–226.

Yochelson, E.L., Flower, R.H. and Webers, G.F. (1973), “The bearing of the new Late Cambrian monoplacophoran genus Knightoconus upon the origin of the Cephalopoda”, *Lethaia*, Blackwell Publishing Ltd, Vol. 6 No. 3, pp. 275–309.

Zhou, X., Xu, S., Yang, Y., Zhou, K. and Yang, G. (2011), “Phylogenomic analyses and improved resolution of Cetartiodactyla”, *Molecular Phylogenetics and Evolution*, Vol. 61 No. 2, pp. 255–264.

Zrzavý, J., Říha, P., Piálek, L. and Janouskovec, J. (2009), “Phylogeny of Annelida (Lophotrochozoa): total-evidence analysis of morphology and six genes”, *BMC Evolutionary Biology*, Vol. 9, p. 189.

Zuckerkandl, E. and Pauling, L. (1962), “Molecular evolution”, *Horizons in Biochemistry and Biophysics*, Academic Press, pp. 189–225.

Zuckerkandl, E. and Pauling, L. (1965), “Molecules as documents of evolutionary history”, *Journal of Theoretical Biology*, Vol. 8 No. 2, pp. 357–366.

Appendix A: original shell scripting and code

Research Chapters 2, 3, 4 and 5 include dataset compilation and refinement methodologies which refer to the *MoSuMa Tools* (*Mo*lecular *Su*per *Ma*trix) pipeline, available at github.com/jairly/MoSuMa_tools. All of these scripts and code were authored by Alastair R Tanner. Some of the major scripts are given in full below. Further details, a complete manual, and other auxiliary scripts are available on github.

```
##########################################################################
# extract_blast_2015.pl, Al Tanner, March 2015 rev Dec 2015
# counts the number of high hits (the number of evalue = 0.0 hits,
# or the highest hit plus any other hits within 3 orders of magnitude)
# then extracts that many results from the blast output file,
# and places them in a fasta formatted output file.
##########################################################################

use strict;
use warnings;
use Data::Dumper;
use List::Util qw(max);

my @files = <blast_out*>;
my $counter = 0;
my $logfile = "extract_blast.log";
my $hash_size = 0;

if (! $ARGV[1]) {
    print "  USAGE: perl extract_blast_2015.pl [e-value cut-off] [taxon name]\n";
    die "EXAMPLE: perl extract_blast_2015.pl -10 Gallus (only evalues smaller than e-10 will be accepted, faster headers 
will be named \">Gallus\")\n";
}

###################
# set cut off value
$ARGV[0] =~ s/-//;
my $cut_off = $ARGV[0];

################
# set taxon name
my $taxon_name = $ARGV[1];

open (LOGFILE, ">$logfile") || die "Problem opening logfile...\n";

`mkdir selected_hits/`; # make the output folder

#########################################################################
# loop over blast_out files and recovers the e-values from the hits table
foreach my $infile (@files) {
    $counter++;
    print LOGFILE "=== Reading file number $counter :: $infile ===" . "\n";
    my %seqs;
    my $evalue_zero = 0;
    my $bool = 0;
    open (IN, "<$infile") || die "Problem opening $infile";
    while (<IN>) {
         chomp;
         if ($_ =~ /^Sequences producing/) { # the table starts with this string
             $bool = 1;
         }
         if (($_ =~ /^>/) && ($bool == 1)) { # the table ends when a line starts with ">"
             $bool = 0;
         }
         if (($_ =~ /^\w/) && ($bool == 1)) {
             my ($name, $bits, $evalue) = split (/[ ]+/, $_); # THIS MIGHT NOT WORK IF THERE ARE SPACES IN THE TITLE
             $seqs{$name} = $evalue;
         }
    }

    my %edited_seqs;
    my @keys = keys (%seqs);

###################################################################
# this loop standardises 0.0 to 999 and removes crap around evalue.
# and zero pads to three figures (makes comparisons easier).
    foreach my $key (@keys) {
         if ($seqs{$key} =~  /^0.0$/) {
             $edited_seqs{$key} = 999;
             $evalue_zero = 1;
             next;
         }
         if (! ($seqs{$key} =~ /e-/)) {
             next;
         }
         if ($seqs{$key} =~ /e-/) {
             my $local_var = $seqs{$key};
             $local_var =~ s/[\d]*e-//;
             $local_var = sprintf("%03d", $local_var);
             print LOGFILE "E-value exponent = $local_var \n";
             $edited_seqs{$key} = $local_var;
             next;
         }
    }
# here we need to add reference to arvg, and have subroutine for
# taking the top value and taking the 3 top values.
    ########################################################
    # find highest value, and remove evalues not within 10^3
    if (! %edited_seqs) {
         print LOGFILE "No acceptably high hits: all e-values really rubbish (not expressed as exponents).\n";
         print LOGFILE "=== File number $counter done ===" . "\n\n";
         next;
    }
    else {
         my $max = max(values(%edited_seqs));
         if ($max < $cut_off) {
             %edited_seqs = (); # remove all hits, if they are are not over cut-off
             print LOGFILE "No acceptably high hits: nothing smaller than cut_off (e-$cut_off)\n";
         }
         print LOGFILE "Best hit = $max (cut-off = e-$cut_off) \n";
         foreach my $key (keys %edited_seqs) {
             if ($max - ($edited_seqs{$key}) > 3) { # delete any values more than 3
                   delete ($edited_seqs{$key});       # orders of magnitude beyond high hit
             }
         }
    }
    $hash_size = keys %edited_seqs; # how many hits did you keep? (often 1)
    print LOGFILE "$hash_size highest hit(s).\n";
    if ($evalue_zero == 1) {
         print LOGFILE "These are 0.0 e-value hits.\n";
    }

    if ($hash_size > 0) { # if the hits have already been counted, extract stuff
         &extract ($infile, $hash_size, $taxon_name);
    }

    close IN;
    print LOGFILE "=== File number $counter done ===" . "\n\n";
}

`mv extract_blast.log ..`;
print "===\nextract_blast_2015.pl: $counter files processed. Details in $logfile\n";
print "extract_blast_2015.pl: Selected sequences saved to selected_hits/*.sel\n===\n";
print LOGFILE "=== $counter files processed. Done. ===" . "\n\n";
close LOGFILE;

exit;

###############
# SUBROUTINES #
###############

sub extract { # puts the best hits into a fasta output file
    my $seq_count = 0;
    my $name = ">" . "$_[2]";
    my $selected_sequence;
    my %name_n_seq = ();
    my $fasta_bool = 0;
    my $extract_bool = 0;
    open (IN, "<$_[0]") || die "Cannot find $_[0]...\n";
    while (<IN>) {
         chomp;
        if ($seq_count > $_[1]) { # only take as many records as there are high hits
            next;
        }
         if ((/^>/) || (/^ Score =/) && ($extract_bool == 1)) {
             $seq_count ++;
             $extract_bool = 0;
             $fasta_bool = 0;
             $selected_sequence = "";
         }
         if ((/^>/) && ($extract_bool == 0)) {
             $fasta_bool = 1;
         }
         if ((/^ Score =/) && ($fasta_bool == 1)) {
             $extract_bool = 1;
         }
        if ((/^Sbjct/) && ($extract_bool == 1)) {
             my $extracted_sequence = $_;
             $extracted_sequence =~ s/^Sbjct: [0-9]+//; # remove sbjct and hit line numbers
            $extracted_sequence =~ s/^\s+//; # get rid of opening space, if there is one
             my @extract_bits = split (/[ ]+/, $extracted_sequence); # avoid ending number
            $selected_sequence = $selected_sequence . $extract_bits[0];
             my $zero_padded_seq_count = sprintf("%03d", $seq_count);
             $name_n_seq{$name . $zero_padded_seq_count} = $selected_sequence;
         }
    }
    close IN;
    $_[0] =~ s/^blast_out_//;
    $_[0] .= ".sel";
    open (OUTFILE, ">$_[0]") || die "Cannot open $_[0] \n\n";
    for (sort keys %name_n_seq) {
         print OUTFILE "$_\n$name_n_seq{$_}\n";
    }
    close OUTFILE;
    # clean up by putting selected files in the output folder
    `mv *.sel selected_hits/`;
}
```

```
###################################################################
# blast_all.pl, a Perl script of the automation of BLAST operations
# taking a sequence target file and blasting it against sequence information fasta files
# Modified March 2015 by Al T. Change to deal with nucleotide or protein blast approaches.
###################################################################

# usage: perl script_name name_file_with_sequences_to_blast* basename_of_blast_database (from formatdb)
# *note that the file of sequences to blast has to be in a simple format where in the same line there is sequence name
followed by sequence itself (the two are separated by one or more spaces)

use strict; use warnings;

if (! $ARGV[2]) {
    die "blast_all_2015.pl: USAGE: perl blast_all_2015.pl [sequences to blast] [blast database] [database sequence format
\(either -aa or -nt\)]\n                 EXAMPLE: perl blast_all_2015.pl sequences.txt Gallo.fas -aa\n";
}

if ($ARGV[2] !~ m/-nt|-NT|-aa|-AA/) {
    die "blast_all_2015.pl: USAGE: perl blast_all_2015.pl [sequences to blast] [blast database] [database sequence format
\(either -aa or -nt\)]\n                 EXAMPLE: perl blast_all_2015.pl sequences.txt Gallo.fas -aa\n";
}

my $infile = $ARGV[0];
my $database_base = $ARGV[1];
my %sequences;
my $seq_count = 1;
open (IN, "<$infile") || die "blast_all_2015.pl: I cannot find \"$infile\"\n";

while (<IN>)
{
    print "blast_all_2015.pl: Storing sequence $seq_count\n";
    my $line = $_;
    chomp $line;
    my ($key, $value) = split (/ /, $line);
    $sequences{$key} = $value;
    $seq_count++;
}

my @sequencenames = keys (%sequences);
my $blast_count = 0;
my $sequences_in_database = scalar(@sequencenames);
print "=====\nblast_all_2015.pl: Sequences in database = $sequences_in_database \n=====\n";

# set which blast search version to use
my $database_format = "";
if ($ARGV[2] =~ "-aa|-AA") { # if amino acids, use blastp
    $database_format = "blastp";
}
if ($ARGV[2] =~ "-nt|-NT") { # if nucleotides, use tblastn
```

```
    $database_format = "tblastn";
}

foreach my $key (@sequencenames)
{
    $blast_count++;
    print "blast_all_2015.pl: Blasting sequence $blast_count\n";
    open (OUT, ">infile") || die "blast_all_2015.pl: there is a problem opening infile.\n";
    print OUT ">" . $key . "\n";
    print OUT $sequences{$key} . "\n";
    close OUT;

    system ("blastall -p $database_format -d ${database_base} -i infile -o out");
    `mv out blast_out_${key}`;
    `rm infile`;
}

close IN;

print "=====\nblast_all_2015.pl: Finished, $blast_count blast operations completed.\n";
my @keys2 = keys(%sequences); # check that a random file looks the right length
my $random_key = $keys2[rand(@keys2)];
if (`grep -c "" blast_out_$random_key` < 12) {
    print "blast_all_2015.pl: Are you sure you told me to use the right blast format? \($ARGV[2]\)\n";
    print "blast_all_2015.pl: (I examined a random file \(blast_out_$random_key\) and it doesn't look right...)\n"
}

# clean up by putting things in a folder
`mkdir blast_out`;
`mv blast_out_* blast_out/`;

print "====\n";

exit;
```

```
#!/bin/bash
###########################################################################
# matrix_compiler.pl
# A script for automating the MoSuMa tools pipeline,
# Al Tanner November 2016
# USAGE: run the script from a folder containing fasta files,
# and nothing else.
# USAGE: sh matrix_compiler.sh [absolute path to blast target file]
###########################################################################

if [ $# -eq 0 ]
  then
    echo "matrix_compiler.sh :: USAGE"
    echo "Please provide the absolute path to your blast target file. for example:"
    echo "sh matrix_compiler.sh /home/john/project1/blast_targets"
    exit 1
fi

# make a folder for each file, named after each file and move each into it
echo "Making folders for these files to go in..."
find . -not -path '*/\.*' -type f -not -name '.' -exec sh -c 'mkdir "${1%.*}" ; mv "$1" "${1%.*}" ' _ {} \;

# clean the .fas in each folder
starting_folder=$PWD;
for folder in */; do
    cd $folder;
    for file in *; do
         perl ~/mosuma_dev/fasta_clean_2015.pl $file $file.clean; done;
    cd $starting_folder;
done;

# format the clean file in each folder
for folder in */; do
    cd $folder;
    for i in *.clean; do
         formatdb -i $i -p T;
    done;
    cd $starting_folder;
done;

# blast
for folder in */; do
    cd $folder;
    for file in *.clean; do
         perl ~/git/MoSuMa_tools/blast_all_2015.pl $1 $file -aa; done;
    cd $starting_folder;
done;

# select the top hits
for folder in */; do
    cd $folder/blast_out/;
    perl ~/git/MoSuMa_tools/extract_blast_2015.pl -10 $folder;
    cd $starting_folder;
done;

# make an output folder named after the date, one layer up
out_folder=`date +"%d%h%y_%H.%M.%S"`;
mkdir $starting_folder/../$out_folder/;

# put the top selected hit into a file named after the blast target
for folder in */; do
    cd $folder/blast_out/selected_hits/;
    for file in *; do
         head -2 $file >> $starting_folder/../$out_folder/$file; done;
    cd $starting_folder;
done;

# rename .sel to .fas in the gene folder
find $starting_folder/../$out_folder -name '*.sel' -exec sh -c 'mv "$0" "${0%.sel}.fas"' {} \;

# remove slashes that might have cropped up in the files
sed -i "s/[/]/_/g" $starting_folder/../$out_folder/*;

# align all of the selected hits to make gene matrices
for file in $starting_folder/../$out_folder/*; do
    muscle -in $file -out $file.ali;
done;

# move the aligned MUSCLE output matrices to their own folder
ali=_aligned;
mkdir $starting_folder/../$out_folder$ali;
mv $starting_folder/../$out_folder/*.ali $starting_folder/../$out_folder$ali/;

# rename .fas.ali to .ali in the aligned folder
find $starting_folder/../$out_folder$ali -name '*.fas.ali' -exec sh -c 'mv "$0" "${0%.fas.ali}.ali"' {} \;

# ummmmmmm that should be it
echo "=== matrix_compiler.sh ==="
echo "All done. Aligned gene matrices are in $out_folder$ali."
echo "=========================="
exit;
```

```
####################################################################
# TREECLEANER.pl by Al Tanner, July 2014. Latest mod: 26 May 2015
# Examines NEWICK trees for long branches
# Please report bugs on github/jairly/MoSuMa_tools         Thanks :)
####################################################################

use strict;
use warnings;
use Bio::TreeIO;
use Text::Balanced 'extract_bracketed';
use List::Util qw(sum);
```

```
# use Statistics::Basic qw(:all); # this might cause problems to some people

if (! $ARGV[1]) {
    print "=== treecleaner.pl: USAGE: perl treecleaner.pl [tree file in Newick format] [threshold branch length]\n";
    print "=== EXAMPLE (3 standard deviations of mean length as threshold): perl treecleaner.pl fibro.tree 3\n";
    print "=== The higher the threshold branch length, the fewer branches will be identified as LONG.\n";
    print "=== To automatically modify the phylip file corresponding to the tree, add this after the threshold.\n";
    die "=== EXAMPLE: perl treecleaner.pl fibro.tree 3 fibro.phy\n";
}

if ($ARGV[2]) { # examine matrix file to modify
    if (! -e $ARGV[2]) {
         die "Phylip file \"$ARGV[2]\" doesn't exist here.\n";
    }
    if (`grep "^>" $ARGV[2]`) {
         die "File \"$ARGV[2]\" looks like a fasta file. I can only modify phylip files.\n";
    }
}

my $threshold = $ARGV[1];
my $phylip_to_modify = $ARGV[2];

# open file
open (TREEFILE, "<$ARGV[0]") || die "treecleaner.pl: Cannot find $ARGV[0] [$!]\n";
my $newick = <TREEFILE>;
if (`grep -o ";" $ARGV[0] | wc -l` != 1) { # checks for tree formatting.
    print "Tree file \"$ARGV[0]\" doesn't seem to have the correct number of semi-colons.\n";
    die "Please check the tree is in Newick format.\n";
}
if (`grep -o "(" $ARGV[0] | wc -l` != `grep -o ")" $ARGV[0] | wc -l`) {
    print "There are an unequal number of close and open brackets in $ARGV[0].\n";
    die "Please check the tree is in Newick format.\n";
}
close (TREEFILE);                          # end format checks

# clean up newick and generate warnings
chomp $newick;
my $space_warning = 0;
my $pipe_warning = 0;
my $plus_warning = 0;
if ($newick =~ m/\s+/) {
    $newick =~ s/\s//g;
    $space_warning = 1; # space warning
}
if ($newick =~ m/\|/) {
    $newick =~ s/\|//g;
    $pipe_warning = 1;  # pipe warning
}
if ($newick =~ m/\+/) {
    $newick =~ s/\+//g;
    $plus_warning = 1;  # plus warning
}

# save a clean version of the newick string (for grepping later)
my $clean_newick_temporary = $ARGV[0] . ".cln";
open (OUT, ">$clean_newick_temporary") || die "Problem making temporary clean newick file...\n";
print OUT "$newick";
close OUT;

# isolate branch lengths and taxa names
my @branch_lengths = $newick =~ /\d+[\.?\d+]*/g;        # match number, with or without decimal point
my @terminal_branches = $newick =~ /\w+:\d+[\.?\d+]*/g; # match [string] ":" [number, with or without decimal point]
for my $index (reverse 0..$#terminal_branches) {        # clean out bootstrap supports that have been
    if ( $terminal_branches[$index] =~ /^\d/ ) {        # mistaken for taxa names.
        splice(@terminal_branches, $index, 1, ());
    }
}

push (my @clade_search_input, $newick);
my $clade_count = () = $newick =~ /\)/g;                # counts occurence of "(" in tree string.
my $total_clade_count = $clade_count + (scalar (@terminal_branches));

print "\ntreecleaner.pl =======================================================\n\n";
print "Terminal clade (taxa) count\t\t" . @terminal_branches . "\n";
print "Multiple-member clade count\t\t$clade_count\n";
print "Total clade count\t\t\t$total_clade_count\n";
```

```
# uncomment next line for verbose output
#print "\nTERMINAL CLADES (" . @terminal_branches . ")\t\tBRANCH LENGTHS\n";
my $taxa;
my $length;
my %taxa_length;
foreach (@terminal_branches) {
    ($taxa, $length) = split (/:/,$_);
    $taxa_length{$taxa} = $length;
}

my @keys = keys{%taxa_length};
# uncomment this loop for verbose output
#foreach my $key (@keys) {
#    printf ("%-25s\t%-20s\n", $key, $taxa_length{$key});
#}

# multiple clades search regex
my $clade_search_regex = qr/
    (                   # start of bracket 1
    \(                  # match an opening bracket
        (?:
        [^\(\)]++       # one or more brackets, non backtracking
            |
           (?1)         # recurse to bracket 1
        )*
    \)                  # match a closing bracket
    )                   # end of bracket 1
    /x;

$" = "\n\t";

my @multi_clades;
while (@clade_search_input) {
    my $string = shift @clade_search_input;
    my @groups = $string =~ m/$clade_search_regex/g;
    push (@multi_clades, @groups) if @groups;
    unshift @clade_search_input, map { s/^\(//; s/\)$//; $_ } @groups;
}

# uncomment next line for verbose output header
#print "\nMULTIPLE MEMBER CLADES (" . $clade_count . ")\n"; # displays readable clade members

my %multi_clades_and_lengths;
foreach (@multi_clades) {
    my $multi_clade_members = $_;
    $multi_clade_members =~ s/\(+//;
    $multi_clade_members =~ s/:.*?,/ + /g;   # replace stuff between : and , with +
    $multi_clade_members =~ s/\(//g;         # remove other brackets
    $multi_clade_members =~ s/:.*?\)$//;     # remove closing bracket
    s/\(/\\\(/g;                             # replace open bracket with ACTUAL backslash open bracket
    s/\)/\\\)/g;                             # replace close bracked with ACTUAL backslash close bracket
#    my $multi_clade_length = `grep -E -o "$_.*?[,|)]" $clean_newick_temporary`;
#    $multi_clade_length =~ s/^.*://g;        # remove stuff at start
#    $multi_clade_length =~ s/.$//g;          # remove last char, usually ","
#    $multi_clades_and_lengths{$multi_clade_members} = $multi_clade_length;
}

# The four lines above are commented because for some reason grep can run out of
# memory and it wont work... not sure about this. Clearly I'm being stupid.

# uncomment the following line for full multiple member clade commentary output
#print "$_ $multi_clades_and_lengths{$_}\n" for (keys %multi_clades_and_lengths);

my $sum_branch_lengths;
my @lengths_difference;
my %clade_branch_length;

# tree statistics
my $branch_count = scalar @branch_lengths;
print "Branch count\t\t\t\t$branch_count\n";
foreach (@branch_lengths) {
    $sum_branch_lengths += $_;
}
my $mean_branch_length = $sum_branch_lengths / $branch_count;
print "Mean branch length\t\t\t$mean_branch_length\n";
```

```
# generate standard deviation
@lengths_difference = @branch_lengths;
foreach (@lengths_difference) {
    $_ = ($_ - $mean_branch_length);
    $_ *= $_;
}
my $differences_summed = sum (@lengths_difference);
my $lengths_standard_deviation = sqrt ($differences_summed / $branch_count);
print "Branch length standard deviation\t$lengths_standard_deviation\n";
print "Threshold length = $threshold standard deviations more than the mean.\n";
my $actual_threshold = ($mean_branch_length + ($threshold * $lengths_standard_deviation));
print "                \(= $actual_threshold\)\n";

# look for long terminal branches
my $terminal_long_branch_count = 0;
foreach my $branch_length (@branch_lengths) {
    if ($branch_length > $actual_threshold) {
         $terminal_long_branch_count++;
    }
}

# initiate hash of taxa to remove
my @taxa_to_remove = "";

# print terminal long branches
if ($terminal_long_branch_count > 0) {
    print "\n----- Long branched taxa or clades in $ARGV[0] -----\n";
    my @keys = sort { $taxa_length{$b} <=> $taxa_length{$a} } keys(%taxa_length);
    my @vals = @taxa_length{@keys};
    my $counter1 = 0;
    for (my $i=0; $i < $terminal_long_branch_count; $i++) {
         print "\t\($keys[$counter1]\)\n";
         push (@taxa_to_remove, $keys[$counter1]); # add to the list of taxa to remove
         $counter1++;
    }
}

# print internal long branches
# if there are more than half the entire tree in a long branch clade,
# the search has picked up the wrong end of the branch, and should ignore
# the content of the multi clade hash.
my $half_taxa_count = (scalar (@terminal_branches) / 2);
my $internal_long_branch_count = 0;
my $majority_clade_bool = 0;
for (keys %multi_clades_and_lengths) {
    chomp;
    if (! $multi_clades_and_lengths{$_}) { # skip empty values (clade of whole tree will
         next;                              # have an empty length value)
    }
    if ($multi_clades_and_lengths{$_} > $actual_threshold) {
         print "\t\($_\)\n";
         s/[ \+ ]/\+/g;
         my @internal_long_branch_clades_to_add = split ('\+', $_);
         @internal_long_branch_clades_to_add = grep /\S/, @internal_long_branch_clades_to_add;
         foreach my $clade_members_to_add (@internal_long_branch_clades_to_add) {
             chomp;
             s/\s//g;
             s/\+//g;
             my $array_size = scalar (@internal_long_branch_clades_to_add);
             if ($half_taxa_count > $array_size) {
                   push (@taxa_to_remove, $clade_members_to_add);
             }
             else {
                   $majority_clade_bool = 1;
             }
         }
         $internal_long_branch_count++;
    }
}
if ($majority_clade_bool == 1) {
    print "      ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^\n";
    print "----- This clade contains more than half of the taxa representation -----\n";
    print "----- Long branches leading to major clades are ignored - these taxa will not be removed from matrix -----\n";
}

@taxa_to_remove = grep /\S/, @taxa_to_remove; # clean up array of empty lines
```

```
my $total_long_branch_count = $terminal_long_branch_count + $internal_long_branch_count;
my $number_of_taxa_to_remove = scalar (@taxa_to_remove);
print "\n----- $number_of_taxa_to_remove taxa associated with long branches -----\n";
print "\t@taxa_to_remove\n";

# report any warnings generated
if ($pipe_warning == 1) {
    print "\nWARNING: $ARGV[0] contained the symbol \"\|\", this symbol was removed before parsing.\n";
}
if ($space_warning == 1) {
    print "\nWARNING: $ARGV[0] contained spaces. Spaces were removed before parsing.\n";
}
if ($plus_warning == 1) {
    print "\nWARNING: $ARGV[0] contained the symbol \"+\". Plusses were removed before parsing.\n";
}

# create new phylip file with taxa removed and metadata updated for correct taxa count.
if ($ARGV[2]) {
    `cp $phylip_to_modify $phylip_to_modify.edited`;
    foreach my $taxa_to_be_removed_from_phylip (@taxa_to_remove) {
         `grep -v "^$taxa_to_be_removed_from_phylip " $phylip_to_modify.edited > temp && mv temp
$phylip_to_modify.edited`;
    }
    my $old_phylip_taxa_count = `grep -o -m 1 "[0-9].* " $phylip_to_modify.edited`;
    chomp $old_phylip_taxa_count;
    my $new_phylip_taxa_count = (`grep -c "" $phylip_to_modify.edited` - 1);
    chomp $new_phylip_taxa_count; # perl inline: update phylip metadata taxa count vvvvv
    `perl -p -i -e "s/$old_phylip_taxa_count/$new_phylip_taxa_count /" $phylip_to_modify.edited`;
    print "\n$number_of_taxa_to_remove long branching taxa removed from $ARGV[2], saved to $phylip_to_modify.edited\n";
}

print "\nDone ================================================================\n\n";

# remove temporary newick file
`rm $clean_newick_temporary`;

exit;
```

```
#!/bin/bash
#
#############################################
# ortho2blasttarget.sh # Al Tanner Jan 2017 #
#############################################
#
# Looks through fasta files of orthologous groups of sequences,
# and extracts just the fasta files with more than a given number
# of sequences in them. Produces an output file with the single longest
# hit from each of those orthologous fastas, ready for blast operations.
#
# USAGE: bash ortho2blasttarget.sh [minimum number of seqs per orthologous group]
# EXAMPLE: bash ortho2blasttarget.sh 20 fa (will extract all orthogroups with 20 or more seqs in, looking through files
suffixed with .fa)
#

minimum_seqs=$1
fasta_suffix=$2
# if there is no argument, quit
if [[ $# -ne 2 ]] ; then
    echo 'ortho2blasttarget.sh: please include the minimum number of sequences per fasta file :)'
    echo 'ortho2blasttarget.sh: please include the suffix of your fasta files :)'
    echo 'example: bash ortho2blastarget.sh 10 fa (will place files with 10 or more seqs in a folder called "10seqs",
looking through files suffixed .fa)'
    exit 0
fi
# quit if output directory already exists...
if [ -d "$1seqs" ]; then
    echo "An output folder called $1seqs already exists here. Better not overwrite that... exiting."
    exit 0
```

```
fi
echo "Making an output folder called $1seqs."
mkdir $1seqs
# if the file contains more > fasta headers than $1, put it in a folder.
total_fasta_files=`ls -l *$2 | wc -l`
echo "There are $total_fasta_files fasta files here."
echo "Looking for files with a minimum of $1 seqs and copying them into folder $1seqs."
for file in *.fa; do
    seqs=`grep -c ">" $file`;
    if [ $seqs -ge $1 ]; then
         `cp $file $1seqs/`;
    fi
done;
# move into selected seqs folder
echo "Moving into folder $1seqs."
cd $1seqs/
number_of_selected_files=`ls *.fa -l | wc -l`
echo "$number_of_selected_files files had a minimum of $1 sequences."
# convert to phylip so seq is all on one line
echo "Converting $number_of_selected_files files to phylip."
perl ~/git/MoSuMa_tools/fasta2phylip.pl .fa .phy
# remove phylip header crap / shorten names
echo 'Shortening headers.'
for file in *.phy; do
    awk '{print $(NF-1) " " $NF}' $file > $file.shortnames;
done
# order by longest to shortest line
echo "Putting longest sequence to top of file."
for file in *.shortnames; do
    awk '{ print length($0) " " $0; }' $file | sort -r -n | cut -d ' ' -f 2- > $file.ordered;
done
# extract the top line, the longest hit
echo 'Taking just the longest hit.'
for file in *.ordered; do
    head -1 $file > $file.onehit;
done
# remove phylip headers
echo 'Cleaning up redundant phylip headers.'
for file in *.onehit; do
    cut -d " " -f2- $file > $file.noname;
done
# remove filename crap
echo 'Cleaning filenames.'
find . -name '*.noname' -exec sh -c 'mv "$0" "${0%.phy.shortnames.ordered.onehit.noname}seq"' {} \;
# rename phylip header with OGnumber, now the filename
echo 'Renaming output.'
for file in *seq; do
    perl -p -i -e "s/^/$file /" $file;
done
# remove "seq" from files
for file in *seq; do
    perl -p -i -e "s/seq//g" $file;
done
# concatenate into a single file
echo 'Concatenating into single file.'
cat *seq > blast_targets_$1seqs
# cleanup
echo 'Cleaning up temporary files.'
rm *.phy*
rm *seq
echo "Done. Blast targets are in the folder $1seqs."
echo "These are also in a single file called blast_targets$1seqs, in folder $1seqs."
exit;
```

Appendix B: papers arising from, or related to, work in this thesis.

The following papers are provided in their published format.

Parry, Luke, **Alastair R. Tanner**, and Jakob Vinther. 2014. "The Origin of Annelids." *Palaeontology*, September. https://doi.org/10.1111/pala.12129.

O'Reilly, Joseph E., Mark N. Puttick, Luke Parry, **Alastair R. Tanner**, James E. Tarver, James Fleming, Davide Pisani, and Philip C. J. Donoghue. 2016. "Bayesian Methods Outperform Parsimony but at the Expense of Precision in the Estimation of Phylogeny from Discrete Morphological Data." *Biology Letters* 12 (4). https://doi.org/10.1098/rsbl.2016.0081.

Puttick, Mark N., Joseph E. O'Reilly, **Alastair R. Tanner**, James F. Fleming, James Clark, Lucy Holloway, Jesus Lozano-Fernandez, et al. 2017. "Uncertain-Tree: Discriminating among Competing Approaches to the Phylogenetic Analysis of Phenotype Data." *Proceedings. Biological Sciences / The Royal Society* 284 (1846). https://doi.org/10.1098/rspb.2016.2290.

Lozano-Fernandez, Jesus, Robert Carton, **Alastair R. Tanner**, Mark N. Puttick, Mark Blaxter, Jakob Vinther, Jørgen Olesen, Gonzalo Giribet, Gregory D. Edgecombe, and Davide Pisani. 2016. "A Molecular Palaeobiological Exploration of Arthropod Terrestrialization." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 371 (1699). https://doi.org/10.1098/rstb.2015.0133.

Alastair R. Tanner, Dirk Fuchs, Inger E. Winkelmann, M. Thomas P. Gilbert, M. Sabrina Pankey, Ângela M. Ribeiro, Kevin M. Kocot, et al. 2017. "Molecular Clocks Indicate Turnover and Diversification of Modern Coleoid Cephalopods during the Mesozoic Marine Revolution." *Proceedings. Biological Sciences / The Royal Society* 284 (1850). https://doi.org/10.1098/rspb.2016.2818.

Puttick, Mark N., Joseph E. O'Reilly, Derek Oakley, **Alastair R. Tanner**, James F. Fleming, James Clark, Lucy Holloway, et al. 2017. "Parsimony and Maximum-Likelihood Phylogenetic Analyses of Morphology Do Not Generally Integrate Uncertainty in Inferring Evolutionary History: A Response to Brown et Al." In *Proc. R. Soc. B*, 284:20171636. The Royal Society.

Parry, Luke A., Fiann Smithwick, Klara K. Nordén, Evan T. Saitta, Jesus Lozano-Fernandez, **Alastair R. Tanner**, Jean-Bernard Caron, Gregory D. Edgecombe, Derek E. G. Briggs, and Jakob Vinther. 2017. "Soft-Bodied Fossils Are Not Simply Rotten Carcasses - Toward a Holistic Understanding of Exceptional Fossil Preservation: Exceptional Fossil Preservation Is Complex and Involves the Interplay of Numerous Biological and Geological Processes." *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, November. https://doi.org/10.1002/bies.201700167.

Palaeontology The Palaeontological Association www.palass.org

[Palaeontology, 2014, pp. 1–13]

FRONTIERS IN PALAEONTOLOGY

THE ORIGIN OF ANNELIDS

by LUKE PARRY[1,3], ALASTAIR TANNER[2] *and* JAKOB VINTHER[1,2]*

[1]School of Earth Sciences, University of Bristol, Wills Memorial Building, Queen's Road, Bristol, Avon, BS8 1RJ, UK; e-mail: lp13932@bristol.ac.uk
[2]School of Biological Sciences, University of Bristol, Life Sciences Building, 24 Tyndall Avenue, Bristol, UK; e-mails: at9362@bristol.ac.uk, jakob.vinther@bristol.ac.uk
[3]Department of Earth Sciences, Natural History Museum, Cromwell Road, London, SW7 5BD, UK
*Corresponding author

Typescript received 28 May 2014; accepted in revised form 15 August 2014

Abstract: Annelids are a phylum of segmented bilaterian animals that have become important components of ecosystems spanning terrestrial realms to the deep sea. Annelids are remarkably diverse, possessing high taxonomic diversity and exceptional morphological disparity, and have evolved numerous feeding strategies and ecologies. Their interrelationships and evolution have been the source of much controversy over the past century with the composition of the annelid crown group, the relationship of major groups and the body plan of the ancestral annelid having undergone major recent revisions. There is a convincing body of molecular evidence that polychaetes form a paraphyletic grade and that clitellates are derived polychaetes. The earliest stem group annelids from Cambrian Lagerstätten are errant, epibenthic polychaetes, confirming that biramous parapodia, head appendages and diverse, simple chaetae are primitive for annelids. Current evidence from molecular clocks and the fossil record suggest that crown group annelids are a Late Cambrian – Ordovician radiation, with clitellates radiating in the Late Palaeozoic. Their body fossil record is largely confined to deposits showing exceptional preservation and is punctuated by the acquisition of hard parts in major groups. The discovery of an Ordovician fossil with soft tissues has shown that machaeridians are in fact a clade of crown polychaetes. They were in existence for more than 200 million years and possess unique calcitic dorsal armour, allowing their mode of life and phylogeny to be interpreted in the context of the annelid body plan. We identify a novel clade of machaeridians, the Cuniculepadida, which exhibit a series of adaptations for burrowing.

Key words: Annelida, Clitellata, Polychaeta, Cuniculepadida, Machaeridia.

ANNELIDS belong to the clade Lophotrochozoa (Halanych *et al.* 1995), a group that includes animals that produce trochophore larvae (including Mollusca and Sipuncula) as well as Brachiopoda and Phoronida (Edgecombe *et al.* 2011). The position of Annelida within this Lophotrochozoa is still under substantial debate from both molecular and morphological points of view.

Annelids have traditionally been split into two morphologically distinct groups, the clitellates and polychaetes, that were considered reciprocally monophyletic in morphological studies (Rouse and Fauchald 1997).

Polychaetes are primarily marine, although a small proportion are known from freshwater (Glasby and Timm 2008), and possess abundant chaetae and lateral outgrowths of the body called parapodia (Rouse and Pleijel 2001). These animals have evolved a great diversity of feeding strategies including (but not limited to) predation, parasitism, suspension feeding and detritus feeding, and can be tube dwelling, epibenthic, burrowing or pelagic in habit (Rouse and Pleijel 2001). The most comprehensive suprafamilial taxonomy of polychaetes based on morphology splits them into three clades: two that possess palps, the Aciculata and Canalipalpata, and a third that lacks palps and other head appendages, the Scolecida (Rouse and Fauchald 1997).

Palps function as either sensory or feeding structures, with both sharing an identical pattern of innervation, suggesting they are homologous structures (Orrhage 1993). As the scolecids are an artificial group united by absences (Rouse and Pleijel 2001), it is reasonable to argue that the ancestral crown annelid possessed palps, and palp homologues have indeed been identified by neuroanatomical observations in several scolecid families (Orrhage and Müller 2005).

Aciculates are the best supported clade based on morphology and are defined by the presence of dorsal and ventral cirri, ventral sensory palps, aciculae, compound chaetae and multiple prostomial antennae (Rouse and Pleijel 2001). The palps of all aciculates are derived from the prostomium and are sensory (Rouse and Pleijel 2001). A close relationship between the Phyllodocida and

doi: 10.1111/pala.12129

Eunicida, the only polychaetes with jaws, is suggested by both morphological and molecular evidence (Rouse and Fauchald 1997; Zrzavy *et al.* 2009; Struck *et al.* 2011; Weigert *et al.* 2014), and a sister group relationship of the two orders is likely.

Canalipalpates are defined by the presence of grooved, ciliated palps that are situated most commonly between the peristomium and prostomium or are derived from the prostomium (Rouse and Pleijel 2001) while they have been secondarily lost in several taxa (Orrhage 2001).

Molecular studies have challenged the old view of reciprocally monophyletic clitellates and polychaetes, consistently placing clitellates as derived within polychaetes, thus rendering the latter paraphyletic (McHugh 1997; Struck *et al.* 2007). Many nonsegmented taxa, including Echiura, Pogonophora and Sipuncula previously thought to be distinct at the phylum level, have also been subsumed into the annelids, suggesting that segmentation is a labile character within Annelida (Struck *et al.* 2007).

However, the base of the annelid tree remains poorly resolved, with different taxa forming the earliest divergences depending on taxon sampling, analysis and data choice. Typically, phylogenetic studies based on molecular sequence data have not recovered the higher polychaete taxa recognized by morphologists, which has been attributed to short branches that are the consequence of rapid early diversification (Rousset *et al.* 2007). There is a lack of congruence between the earlier studies of molecular data, which employed data from a diversity of sources including protein-coding genes and nuclear genes (Rousset *et al.* 2007), mitochondrial genes and gene order (Mwinyi *et al.* 2009; Shen *et al.* 2009); miRNAs (Sperling *et al.* 2009), ESTs (Struck *et al.* 2011); and combined morphological and molecular analyses (Zrzavy *et al.* 2009). Nonetheless, as more data have become available and statistically better fitting models in a Bayesian and maximum likelihood framework have been employed, a topology is emerging, which to some extent is congruent with a morphological scenario.

Struck *et al.* (2011) recovered a topology that is broadly congruent with morphological hypotheses of polychaete relationships. In their analysis, aciculates plus Orbiniidae form a monophyletic clade while canalipalpates form a paraphyletic grade to clitellates, but the taxa that form the base of the tree are surprising and suggest an ancestral annelid morphology that is at odds with the fossil record (Eibye-Jacobsen and Vinther 2012). However, these basal relationships are unstable, as highlighted by more recent studies of a similar data set with paralogous sequences pruned (Struck 2013) and additional taxa (Weigert *et al.* 2014), which resulted in profoundly different polychaete families forming the most basal diverging taxa.

Hypotheses of relationships based on rare genomic changes have yielded results that are in conflict with phylogenies from sequence data. Rare genomic changes are thought to be reliable signatures in phylogenetics (Rokas and Holland 2000) as they are conserved over geological time, are analytically simple and, in the case of miRNAs, have the potential to be homoplasy free (Tarver *et al.* 2013). Sipunculans are commonly recovered as in-group annelids in studies of sequence data (Struck *et al.* 2011). This has led to the interpretation that sipunculans are derived annelids that have lost segmentation and parapodia. Mwinyi *et al.* (2009) found Annelida and Sipuncula to be sister taxa based on mitochondrial gene order, but in contrast found Sipuncula nested within Annelida when analysing mitochondrial protein-coding genes. Sperling *et al.* (2009) employed a miRNA data set to address deep annelid phylogeny and similarly found a sister group relationship between Annelida and Sipuncula and a more derived position of Chaetopteridae than suggested by sequence data (Struck *et al.* 2011). The earliest crown group sipunculans are known from the Early Cambrian Chengjiang biota (Huang *et al.* 2004), which are roughly coeval with stem group polychaete fossils from the Sirius Passet biota of North Greenland (Conway Morris and Peel 2008; Vinther *et al.* 2011). Sipunculans are possibly also known from the middle Cambrian 'thin' Stephen Formation of British Columbia (Caron *et al.* 2010). An in-group position of sipunculans within annelids suggests that annelids radiated in the Early Cambrian (Weigert *et al.* 2014), whereas preliminary molecular clock analyses (Edgecombe *et al.* 2011; Erwin *et al.* 2011) and current interpretations of the annelid body fossil record (Eibye-Jacobsen 2004) suggest annelids underwent a Late Cambrian – Ordovician radiation.

Recent morphological investigations of embryological characters of the sipunculan pharyngeal apparatus (Tzetlin and Purschke 2006) and nervous system (Wanninger *et al.* 2009) suggest that sipunculans may indeed be primitively segmented. The phylogenetic position of sipunculans, with respect to annelids, may therefore hold the key to the order of the acquisition of parapodia and segmentation within this branch of the Lophotrochozoa.

The phylogenetic position of myzostomids within or outside of annelids has proved to be similarly problematic. Myzostomids are small, obligate parasites of crinoids that have a fossil record extending to at least the Carboniferous (Welch 1976). They share a suite of characters with annelids, more specifically with aciculates (Rouse and Pleijel 2001), including segmentation, chaetae, aciculae and parapodial cirri. Myzostomids are frequently recovered either as nonannelids (Eeckhaut *et al.* 2000) or in the basal portion of the annelid tree on a very long branch (Struck *et al.* 2011) in analyses of sequence data. Helm *et al.* (2012) found that myzostomids are derived within the annelids based on their complement of miRNAs. Weigert *et al.* (2014) recovered the first tree from sequence data in which myzostomids are recovered as aciculates, lending further weight to this long-held morphological hypothesis (Rouse and Pleijel 2001).

The stable results from recent molecular phylogenies suggest that canalipalpates plus clitellates form a monophyletic clade with the majority of scolecids (such as arenicolids and opheliids) distributed polyphyletically (Fig. 1). Aciculates plus orbiniids form the sister group of the mainly sedentary and infaunal annelids, with miRNA data suggesting that sipunculans are the sister group of all annelids. While chaetopterids are commonly recovered as forming deep branches of the annelid tree (Struck *et al.* 2007, 2011), morphological data strongly suggest that they are closely related to canalipalpates given that they possess uncini and grooved palps that are morphologically similar to those of spionids (Eibye-Jacobsen and Vinther 2012).

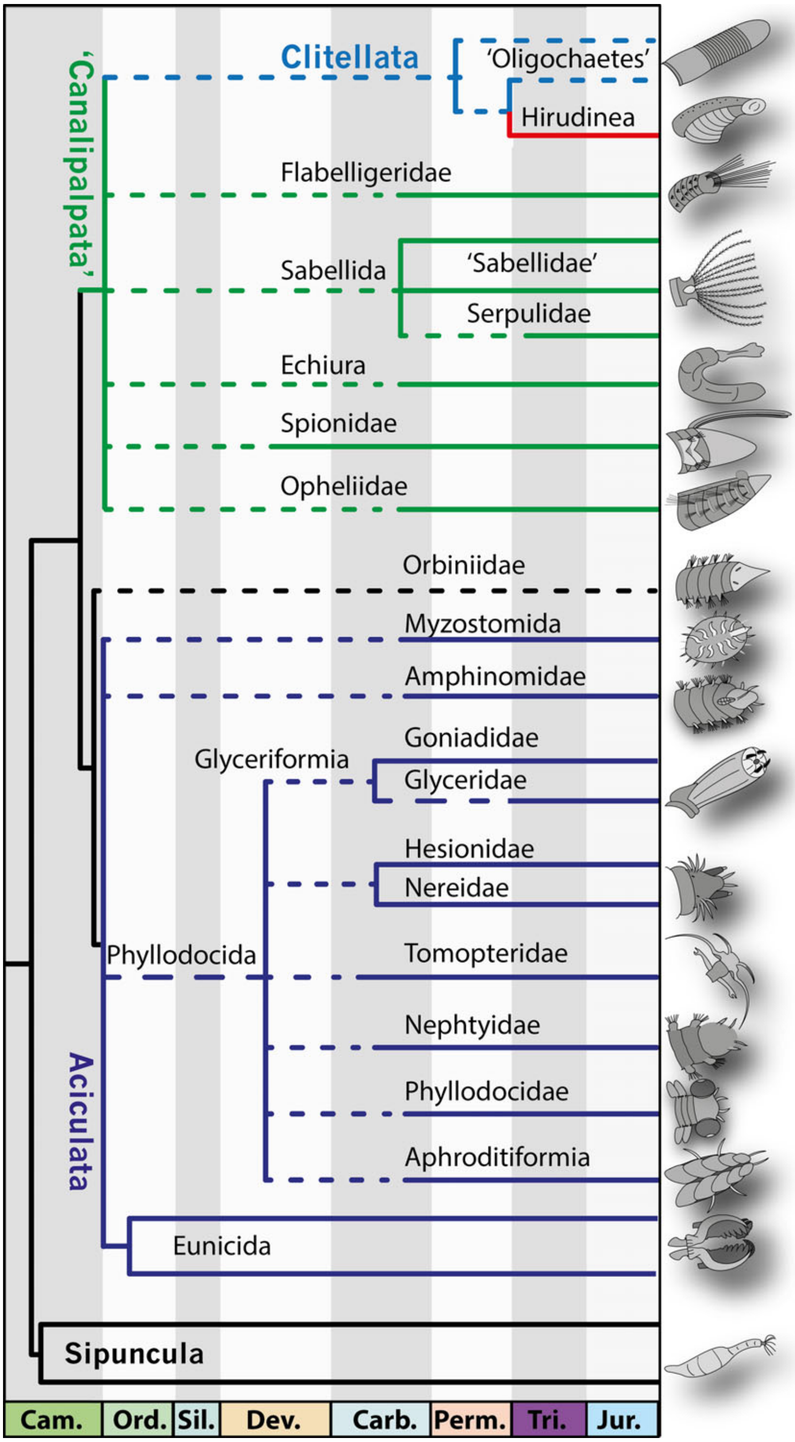


FIG. 1. Phylogeny of extant annelids summarizing common results from analyses of sequence data and rare genomic changes. Polytomies represent areas of the tree that are poorly resolved or uncertain. Only polychaete clades with a fossil record are shown (except Orbiniidae), with fossil ranges based on the published literature (see Table 1). The divergence times within crown clitellates are based on the molecular clock estimates of Edgecombe *et al.* (2011) and Erwin *et al.* (2011) with the first appearance of leech cocoons constraining Hirudinea. The age of Eunicida is discussed in text.

Terrestrial annelids are represented by earthworms and leeches, characterized by the nonsegmented epidermal reproductive structure known as the clitellum (Brinkhurst 1992) and unique spermatozoan specializations such as the intercalation of mitochondria between the nucleus and the axoneme, the presence of an acrosomal tube and the presence of a prominent central sheath in the axoneme (Ferraguti and Erséus 1999). Aquatic clitellates make up a minor part of their diversity, but include the leech-like ectoparasitic branchiobdellids and acanthobdellids, and the semi-infaunal Naididae (formerly Tubificidae – see Erséus *et al.* 2008). Apart from these, clitellates are primarily a terrestrial phenomenon and adaptations for life on land clearly distinguish them from the polychaetes. These adaptations include the absence of parapodia and head appendages, a distinct neuroanatomy, and specialized reproductive and developmental characteristics (Purschke *et al.* 2000). Postfertilization, the clitellum secretes and deposits a cocoon in which all further development takes place: juvenile clitellates have no larval stages. These synapomorphies firmly support clitellate monophyly (Erséus 2005) and molecular studies concur (Struck *et al.* 2011; Weigert *et al.* 2014). Nevertheless, the placement of clitellates within polychaetes is uncertain, not least because (as a consequence of their physiology and habitat) clitellates have no documented fossil record beyond Triassic leech cocoons (Manum *et al.* 1991; Bomfleur *et al.* 2012). Presently, phylogenomics provide no further congruence on clitellate position within annelids, with recent studies placing a clade of echiurids + capitellids (Struck *et al.* 2011) or terrebellids (Weigert *et al.* 2014) as their sister group.

CAMBRIAN STEM GROUP ANNELIDS AND THE ANCESTRAL ANNELID

Historically, the treatment of clitellates and polychaetes as monophyletic groups resulted in two polarized views of the origin and evolution of the annelid body plan. Clark (1964) suggested that the ancestral annelid was an infaunal organism with segmentation arising from a need to compartmentalize the coelom to assist peristaltic burrowing. Westheide (1997), in contrast, argued that the development of parapodia drove internal segmentation to satisfy the need for blood supply. The new phylogenetic hypothesis of annelids in which clitellates are derived polychaetes clearly demonstrates that a body plan with abundant and well-developed chaetae, parapodia and palps is the primitive condition for annelids (Struck *et al.* 2011; Eibye-Jacobsen and Vinther 2012). However, rogue taxa that often resolve at the base of the annelid tree hinder our efforts to reconstruct the ancestral annelid body plan based on their phylogeny alone (Eibye-Jacobsen and Vinther 2012), and evidence from molecular phylogenies, the early fossil record and the morphology of living organisms all need to be taken into account.

Annelids have classically been considered to possess a muscle apparatus composed of an outer layer of circular muscle and inner bands of longitudinal muscle, which greatly influenced the evolutionary hypothesis of Clark (1964). The use of F-actin staining and confocal microscopy has allowed polychaete muscle anatomy to be studied in detail (Tzetlin and Filippova 2005; Zhadan *et al.* 2014). Such investigations have revealed that enclosing circular muscles are absent in many polychaete families and longitudinal muscles are instead antagonized by other muscle groups, such as those of the parapodia (Tzetlin and Filippova 2005). This observation led Tzetlin and Filippova (2005) to the interesting hypothesis that circular muscles may be primitively absent in annelids, a question that may become more tractable once annelid phylogeny is sufficiently well resolved.

Capillary chaetae are present in almost all groups of polychaetes (Merz and Woodin 2006), and observations from the fossil record suggest that they are the most primitive chaetal type (Eibye-Jacobsen 2004). The presence of ultrastructurally identical chaetae in brachiopods (Lüter 2000) as well as in chitons (Leise and Cloney 1982) suggests that chaetae have deeper origins than the annelid total group. Capillary chaetae have a diversity of functions in modern polychaetes including protection (Westheide 1997) and locomotion (Merz and Woodin 2006), while more complex chaetal forms have evolved to perform more specialized functions such as anchoring in tubes with uncini (Merz and Woodin 2006).

The oldest polychaete whole-body fossils are known from the Early Cambrian Sirius Passet biota of North Greenland (Fig. 2A; Conway Morris and Peel 2008; Vinther *et al.* 2011). Despite the similar age of the Chengjiang biota of China, this deposit has yielded no convincing polychaete body fossils (Conway Morris and Peel 2008). However, the geographically close and taphonomically similar but younger Emu Bay shale has yielded a single polychaete specimen (Greg Edgecombe, pers. comm. 2013), making it unclear whether the absence of polychaetes in the Chengjiang is the result of biogeography, low abundance, taphonomic bias or true absence.

The best-known Cambrian polychaetes are *Canadia* (Fig. 3B) and *Burgessochaeta* (Fig. 2C) of the Burgess Shale of British Columbia. The relationships of these fossils to modern polychaete groups have been the subject of much debate (Conway Morris 1979; Eibye-Jacobsen 2004). The majority of Cambrian polychaetes have been interpreted as errant and epibenthic, including forms from Sirius Passet, but unlike living aciculates, lack characters such as dorsal and ventral cirri, compound chaetae, antennae, jaws and aciculae (Eibye-Jacobsen

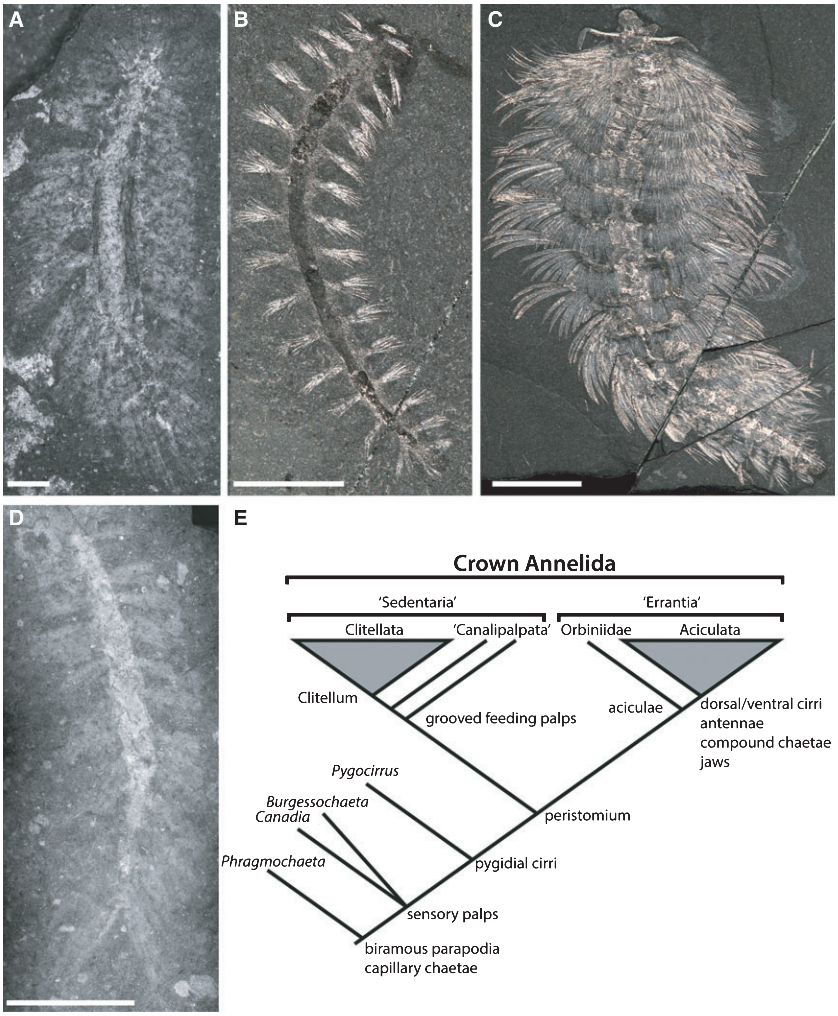


FIG. 2. Cambrian stem group annelids. A, complete specimen of *Phragmochaeta canicularis* from Sirius Passet, showing anterior chaetae and longitudinal muscle bands; scale bar represents 1.5 mm (Geological Museum of Copenhagen, MGUH 30888). B, *Burgessochaeta setigera* (Royal Ontario Museum 61042) from the Burgess Shale; scale bar represents 5 mm (image courtesy of Jean-Bernard Caron). C, *Canadia spinosa* (ROM 56972) from the Burgess Shale showing sensory palps and paleae (enlarged, scale-shaped chaetae); scale bar represents 5 mm (image courtesy of Jean-Bernard Caron). D, posterior fragment of *Pygocirrus butyricampum* (MGUH 29288) from Sirius Passet showing pygidial cirri; scale bar represents 5 mm. E, tentative phylogeny of Cambrian stem annelids showing the acquisition of key characters on the annelid stem.

2004; Conway Morris and Peel 2008; Vinther *et al.* 2011). *Burgessochaeta* has been interpreted as infaunal (Briggs *et al.* 1994), although the evidence for this is weak. Stronger evidence for an infaunal mode of life is observed in *Peronochaeta* due to the presence of backward-facing 'hooks' (Merz and Woodin 2006). Aciculae are one of the most robust types of chaetae and have a high preservation potential and therefore should have been preserved by the taphonomic regime of the Burgess Shale (Butterfield 2003). Aciculae are well known from younger Lagerstätte, such as the Carboniferous Mazon Creek (Thompson 1979) and Bear Gulch (Schram 1979), and the Cretaceous Haqel and Hadjoula, Lebanon (Bracchi and Alessandrello 2005). The absence of such a diagnostic character from Cambrian polychaetes has led to the interpretation that they are members of the annelid stem group (Budd and Jensen 2000; Eibye-Jacobsen 2004). The morphology of these animals is therefore informative of the morphology of the ancestral annelid and the sequence of character acquisition on the annelid stem (Fig. 2E; Vinther *et al.* 2011; Eibye-Jacobsen and Vinther 2012).

Palps are considered a basal character within annelids, with palps only secondarily absent in some taxa (Eibye-Jacobsen 2004) and are split into two morphological varieties. The palps of aciculates are sensory in function and derived from the prostomium, whereas the palps of the canalipalpates are used in feeding, arising dorsally, either from the junction between the prostomium and peristomium or from the prostomium (Rouse and Pleijel 2001). Despite these key differences, both types of palp share the same pattern of innervation with the brain (Orrhage 1993). The palps of Cambrian annelids are most likely prostomial in origin, are contractile and lacked a ciliated groove (Eibye-Jacobsen 2004). This morphology strongly suggests that primitive palps were sensory and not feeding structures (Eibye-Jacobsen 2004; Eibye-Jacobsen and Vinther 2012).

Pygocirrus (Fig. 2D) is the only Cambrian fossil known to possess pygidial cirri, a possible apomorphy shared by living polychaetes (Eibye-Jacobsen 2004), suggesting it is the most crownward of the Cambrian taxa (Fig. 2E). Both biramous and uniramous parapodia are known from Cambrian polychaetes (Conway Morris 1979). It is possible that

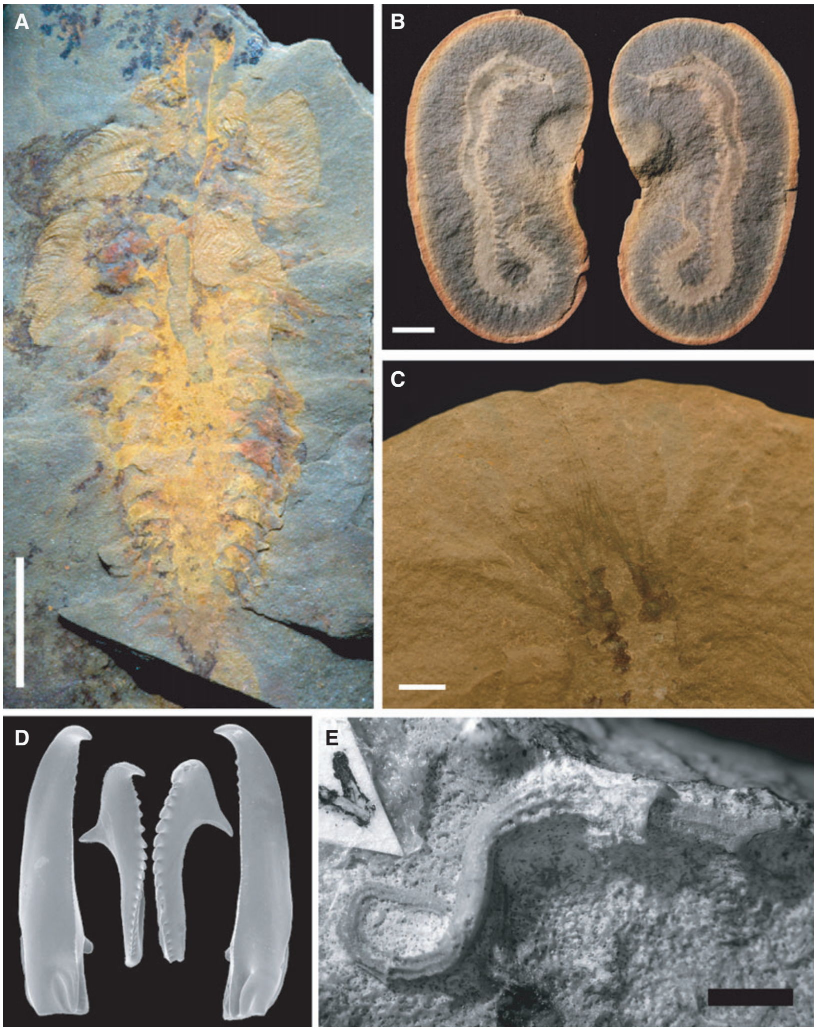


FIG. 3. Annelid body fossils, scolecodonts and calcareous tubes. A, the machaeridian *Plumulites bengtsoni* showing preservation of soft tissue; scale bar represents 5 mm (image courtesy of Peter Van Roy). B, *Esconites zelus* (ROM 47973), Upper Carboniferous, Mazon Creek; scale bar represents 1 cm (image courtesy of David Rudkin). C, anterior portion of *Mazopherusa prinosi* (University of Bristol BRSUG 29380), Upper Carboniferous, Mazon Creek, showing the cephalic cage characteristic of flabelligerids; scale bar represents 2 mm. D, *Kingnites diamondi* (Department of Geology, Lund University LO11518t, LO11512t, LO11513t, LO11521t from left to right), an Ordovician eunicidan scolecodont of the family Paulinitidae. Elements have been rescaled to form a complete apparatus; image courtesy of Mats Eriksson. E, *Vermiliopsis negevensis* (University of Tartu, Natural History Museum TUG 1372-2), a serpulid worm tube from the Middle Jurassic of Israel; scale bar represents 2 mm (image courtesy of Olev Vinn).

some Cambrian taxa possessed a simple eversible proboscis (Eibye-Jacobsen 2004), and living polychaetes both jawed and jawless display diverse feeding habits (Fauchald and Jumars 1979), rendering the feeding strategies of the Cambrian taxa uncertain.

In conclusion, the ancestral annelid was likely an errant polychaete, possessing biramous parapodia with simple chaetae, paired dorsal and ventral longitudinal muscle bands, prostomial sensory palps and possibly an eversible pharynx. Grooved palps, hooks and uncini are later innovations within sedentary polychaetes, and aciculae, compound chaetae, parapodial cirri and antennae are novel features of aciculates (Vinther *et al.* 2011).

THE CROWN ANNELID FOSSIL RECORD

There are few molecular clock estimates of the date of the origin of the annelid crown group and of major groups within the annelids. Recent studies have used a restricted taxon sample and, due to uncertainties of the internal relationships within annelids, it is uncertain whether such analyses have properly bracketed the annelid crown node. Such analyses recover crown group ages ranging from mid-Cambrian (Edgecombe *et al.* 2011) to Early Ordovician (Erwin *et al.* 2011). There is agreement that clitellates radiated during the Late Palaeozoic, which is in part corroborated by the presence of leech cocoons from the Triassic (Manum *et al.* 1991; Bomfleur *et al.* 2012) and the separation from polychaetes by a long branch (Struck *et al.* 2011). The first-known appearances of major polychaete clades based on well-studied fossils are listed in Table 1. Many of the scolecodont-bearing eunicidan families are omitted from this list as the relationship between scolecodont families and living families remains obscure (but see Bracchi and Alessandrello (2005) for a discussion of the fossil record of these groups).

As annelids are primitively entirely soft-bodied and decay rapidly (Briggs and Kear 1993), their whole-body

TABLE 1. First appearances of major polychaete groups shown in Figure 1 with ages based on the published literature, with a first appearance of Eunicidae also given.

Clade	Taxon	Age	Reference
Amphinomidae	*Paleocampa Anthrax*	Pennsylvanian	Thompson (1979)
Aphroditiformia	*Dryptoscolex matthiesae, Faustoscolex gemmatus, Histriocola deliculata*	Pennsylvanian	Thompson (1979)
Echiura	*Coprinoscolex ellogimus*	Pennsylvanian	Jones and Thompson (1977)
Eunicidae	*Esconites zelus*	Pennsylvanian	Thompson and Johnson (1977)
Flabelligeridae	*Mazopherusa prinosi*	Pennsylvanian	Hay (2002)
Glyceridae	'*Glycera*' (scolecodont)	Permian	Nakrem *et al.* (2001)
Goniadidae	*Carbosesostris megaliphagon*	Mississippian	Schram (1979)
Hesionidae	*Rutellifrons wolfforum*	Pennsylvanian	Thompson (1979)
Myzostomida	Galls on crinoid stems	Carboniferous	Welch (1976)
Nephtyidae	*Astreptoscolex anasilosus, Didontogaster cordylina*	Mississippian	Schram (1979)
Nereidae	*Fossundecima konecniorum*	Pennsylvanian	Fitzhugh *et al.* (1997)
Opheliidae	Undescribed	Pennsylvanian	Thompson (1979)
Pectinariidae	Unnamed	Santonian	Vinn and Luque (2013)
Phyllodocidae	*Levisettius campylonectus*	Pennsylvanian	Thompson (1979)
Sabellidae	*Glomerula*	Late Carboniferous	Ippolitov *et al.* (in press)
Serpulidae	*Filograna*	Triassic	Ippolitov *et al.* (in press)
Spionidae	Unnamed	Mid-Devonian	Cameron (1967)
Tomopteridae	*Eotomopteris aldridgei*	Mississippian	Briggs and Clarkson (1987)

fossil record is confined to Lagerstätten and is punctuated by the acquisition of hard parts in the conventional fossil record. Canalipalpates did not acquire calcareous tubes until the Carboniferous (Ippolitov *et al.* in press), and consequently, their body fossil record is otherwise restricted to four occurrences, a flabelligerid (Hay 2002) and echiuran (Jones and Thompson 1977) from the Pennsylvanian Mazon Creek (Fig. 3C), a possible Devonian spionid (Cameron 1967), an undescribed opheliid (Thompson 1979) and a dubious Triassic arenicolid (Horwood 1912; Rouse and Pleijel 2001). The spionid is supposedly preserved in pyrite in an agglutinated tube within a boring of a bivalve shell. While this mode of life is consistent with modern spionids, this specimen displays puzzling three-dimensional preservation in pyrite, is poorly figured and demands further investigation.

Such first appearances are likely to postdate true canalipalpate origins and are more likely a consequence of taphonomic bias. The internode, separating canalipalpates from the aciculates, is remarkably short (Struck *et al.* 2011), which suggests a radiation of the major canalipalpate clades soon after these two lineages split. Many other groups of filter-feeding organisms radiated during the Great Ordovician Biodiversification Event (Servais *et al.* 2009), and it is possible that the filter-feeding canalipalpates formed part of this adaptive radiation.

The best evidence for the timing of the origins of major groups within the canalipalpates comes from the fossil record of tube-building polychaetes. Modern polychaetes use diverse materials to build tubes including mud/mucous (e.g. most Sabellidae), chitin (e.g. Chaetopteridae) and calcium carbonate (e.g. Serpulidae; Ippolitov *et al.* in press). Calcareous tubes have the highest preservation potential and are built by polychaetes belonging to the Serpulidae, Sabellidae and Cirratulidae (Vinn and Mutvei 2009; Ippolitov *et al.* in press). Calcareous tube building is restricted to a single genus in both the cirratulids and the sabellids, with serpulids forming the majority of calcareous polychaetes (Ippolitov *et al.* in press). Molecular evidence suggests that serpulids are nested within sabellids (Kupriyanova and Rouse 2008), which is corroborated by the fossil record. While many Palaeozoic tubes and the Ediacaran genus *Cloudina* have been interpreted as serpulids, recent reassessments of these fossils have found them to show morphological differences that distinguish them from polychaetes (Vinn and Mutvei 2009; Vinn and Zaton 2012). Sabellids are known as fossils from the Late Carboniferous (Ippolitov *et al.* in press) while serpulids first appear in the Mid-Triassic (Vinn and Mutvei 2009). Calcareous cirratulids are far younger by comparison and first appear in the Oligocene (Ippolitov *et al.* in press). The history of agglutinated tube builders is less certain due to their lower preservation potential. Pectinariids have a fossil record that extends to the Cretaceous (Vinn and Luque 2013), suggesting that a diversity of fossilizable tube-building strategies originated in the Mesozoic. The origins of the Siboglinidae are more controversial, with possible fossils dating back to the Silurian

(Little *et al.* 1998), while molecular clock estimates suggest that they are Mesozoic or younger (Hilário *et al.* 2011). However, this question should be treated in a relaxed molecular clock framework, using fossil calibration.

Aciculates are best represented in the fossil record by scolecodonts, the jaws of the Eunicida and Phyllodocida (e.g. Fig. 3D). The overwhelming majority of scolecodonts are from the order Eunicida, which possess a multi-element jaw apparatus composed of ventral mandibles and dorsal maxillae that is commonly asymmetrical (Szaniawski 1996). Eunicidan scolecodonts originated in the latest Cambrian and radiated in the Ordovician (Hints and Eriksson 2007) and are consequently the oldest polychaete fossils assignable to the crown group. Six distinct types of jaw apparatus have been identified within the Eunicida, with two such types known exclusively from the fossil record (Paxton 2009). The earliest representatives are placognath and ctenognath type apparatuses, the former being wholly extinct (Paxton 2009) lying outside the eunicidan crown group. Ctenognath apparatuses are known from living Dorvilleidae, and this jaw architecture has been suggested as the primitive condition from which the jaws of the other eunicidan groups are derived (Paxton 2009). The labidognath apparatuses characteristic of modern Eunicidae and Onuphidae first occur in the Mid-Ordovician (Eriksson *et al.* 2013) placing the only constraint on the origin of the eunicidan crown group, given that 'Ctenognatha' is paraphyletic.

Although parataxonomy has plagued early studies of scolecodonts (Eriksson and Bergman 1998), recent studies have combined modern and fossil data to study the evolution of jaws in deep time, increasingly using an apparatus-based approach (Whittle *et al.* 2008; Paxton 2009; Paxton and Eriksson 2012).

The oldest unequivocal phyllodocidan body fossils are from the Middle Devonian of Canada (Farrell and Briggs 2007). It is possible that polychaetes bearing aciculae and noneunicidan jaw apparatuses from the Early Devonian Hunsrück Slate are phyllodocidans, but diagnostic characters such as head appendages and tentacular cirri are absent and have most likely been lost due to decay (Briggs and Bartels 2010). *Kenostrychus* from the Silurian of Herefordshire may also be a phyllodocidan, but a cladistic analysis by Sutton *et al.* (2001) was inconclusive. Many phyllodocidan families first appear in the Carboniferous (Fig. 1) in the Mazon Creek biota (Thompson 1979), Bear Gulch limestone (Schram 1979) and Granton Shrimp Bed (Briggs and Clarkson 1987). This clustering of first appearance dates in the Carboniferous is probably attributable to favourable taphonomic conditions that allowed polychaetes to be well preserved and easily identified, rather than a true Carboniferous radiation of phyllodocidans. Jawed phyllodocidans such as Gonididae, Nereidae and Nephtyidae are known from the Mazon Creek (Thompson 1979; Fitzhugh *et al.* 1997) and Bear Gulch (Schram 1979), but isolated phyllodocidan scolecodonts are not known until the Permian, possibly due to a lack of intensive sampling (Nakrem *et al.* 2001).

The scolecodont record is not a perfect chronicle of jawed polychaete evolution. The preservation potential of different jaws is contingent on composition, which varies greatly between families (Colbath 1988). Glyceridae were found to have jaws with the highest preservation potential by Colbath (1988) whereas the jaws of Nereidae, Onuphidae, Lumbrineridae and Eunicidae were found to be unlikely to persist into the fossil record. Such findings are congruent with the complete absence of nereid scolecodonts from the fossil record (Szaniawski 1974) and the discovery that glycerid jaws contain high concentrations of melanin (Moses *et al.* 2006), an organic pigment that has recently been demonstrated to have a high preservation potential (Vinther *et al.* 2008*a*; Glass *et al.* 2012). The first appearance of Glyceridae as scolecodonts in the Late Permian fossil record (Nakrem *et al.* 2001) may therefore closely approximate their true first appearance despite the first appearance of their sister group, the Goniadidae, in the Carboniferous (Thompson 1979).

Machaeridians are a group with an unique, calcitic dorsal armour that are known from the Ordovician – Mid-Permian (Högström and Taylor 2001). They were for a long time enigmatic, being attributed variously to barnacles, echinoderms, molluscs and annelids. Their calcitic skeleton is composed of a longitudinal series of two or four-shell plates that are ornamented by concentric rugae and growth lines (Högström *et al.* 2009; Vinther and Briggs 2009). Machaeridians have long been recognized as a monophyletic group based upon the growth and structure of their shell plates (Adrain 1992), but their higher-level affinity was unknown until the discovery of an articulated plumulitid with parapodia and chaetae from the Ordovician of Morocco (Fig. 3A; Vinther *et al.* 2008*b*). This discovery firmly established that machaeridians are an extinct clade of annelids and thus has allowed these fossils to be interpreted in the context of a segmented body plan (Vinther and Briggs 2009).

Machaeridians have been classified into three families that are differentiated by the arrangement and articulation of their shell plates, the plumulitids, turrelipadids and lepidocoleids. Soft anatomy has only been observed in plumulitids (Vinther *et al.* 2008*b*) and lepidocoleids (Högström *et al.* 2009) with both families demonstrating a segmental body plan, reinforcing the case for machaeridian monophyly and their placement within the annelids. The shell plates appear to attach to elaborate parapodial cirri akin to elytra of modern scale worms with inner and outer shell plates attaching to alternating segments (Vinther *et al.* 2008*b*). This potentially affiliates machaeridians with

Phyllodocida, either within or on the stem of living aphroditaceans (Vinther *et al.* 2008*b*), as the latter are generally characterized by forms with dorsal cirri of alternating morphology by segment with some variability in alternation pattern (Glasby *et al.* 2008). However, no machaeridian has been demonstrated to possess the defining characters of aciculate polychaetes, such as aciculae, dorsal and ventral cirri or sensory palps (Vinther and Rudkin 2010), so their exact phylogenetic placement remains an open question until new soft-bodied specimens are discovered.

Plumulitid shell plates are thin and lack muscle scars, and have been interpreted as passive armour (Vinther and Briggs 2009). It is likely that plumulitids were epibenthic, using their parapodia to crawl across the seafloor much like extant scale worms (Vinther and Briggs 2009). The two most anterior segments of plumulitids lack outer shell plates and possess inner plates that are morphologically distinct from the plates of other segments (Vinther and Rudkin 2010).

Turrilepadids and lepidocoleids both display lateral displacement of the outer shell plates so that their skeleton encloses their entire body (Vinther and Briggs 2009) and asymmetric left/right morphology of the inner shell plates. Unlike the plumulitids, their shell plates display evidence of muscle attachment and thickening, indicating that their shell plates were actively employed during burrowing (Vinther and Briggs 2009).

Turrilepadids are square in transverse cross-section with inner plates bending through 90 degrees at their midline (Fig. 4B). Articulated turrilepadids are rare, but like articulated plumulitids lack outer shell plates on the most anterior segments (Adrain *et al.* 1991).

Lepidocoleids display the body plan most divergent from that of plumulitids. Their most distinguishing

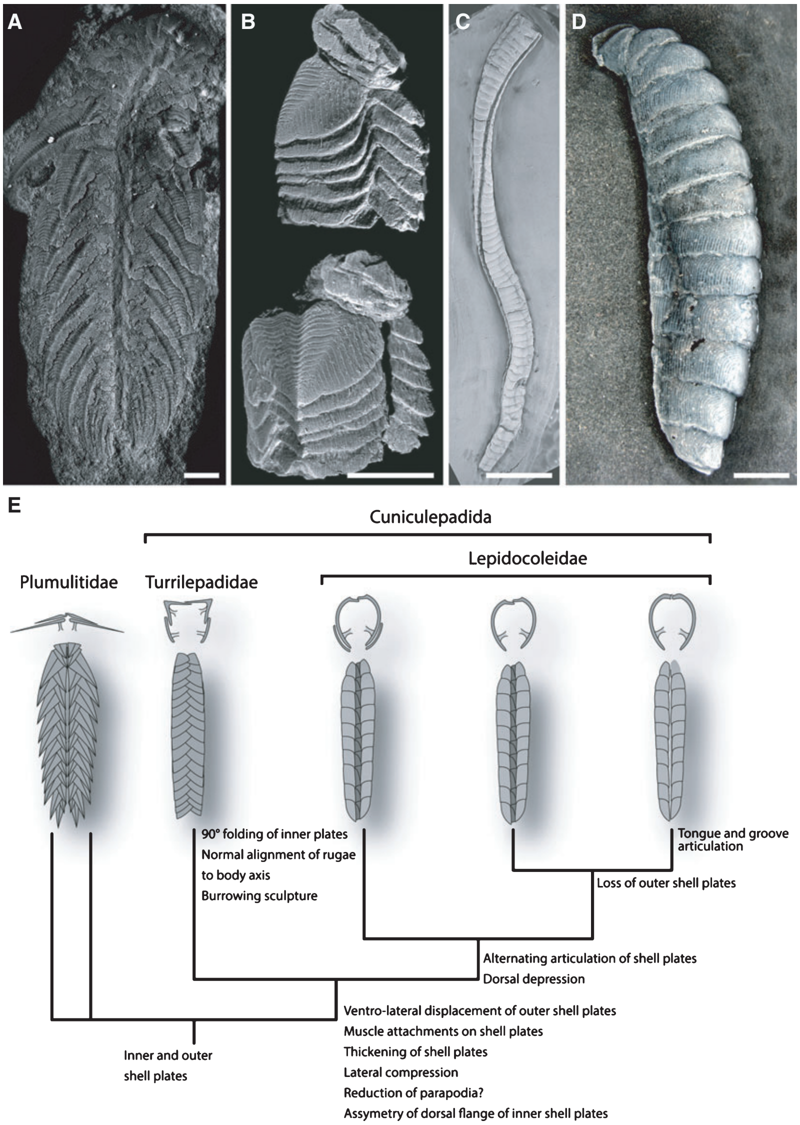

A
B
C
D
E
Cuniculepadida
Lepidocoleidae
Plumulitidae
Turrilepadidae
90° folding of inner plates
Normal alignment of rugae to body axis
Burrowing sculpture
Tongue and groove articulation
Loss of outer shell plates
Alternating articulation of shell plates
Dorsal depression
Ventro-lateral displacement of outer shell plates
Muscle attachments on shell plates
Thickening of shell plates
Lateral compression
Reduction of parapodia?
Assymetry of dorsal flange of inner shell plates
Inner and outer shell plates

FIG. 4. Articulated machaeridian scleritomes and their phylogeny. A, *Plumulites richorum,* Middle Devonian, South Australia (Yale Peabody Museum IP227508, latex cast of National Museum of Victoria P54265); scale bar represents 2 mm. B, *Turrilepas wrightiana,* Silurian, Gotland, Sweden (Naturhistoriska Riksmuseet 3852), incomplete specimen in different views; scale bar represents 5 mm. C, the tetraseriate lepidocoleid *Lepidocoleus ketleyanus*, Wenlock, Dudley, UK (Birmingham University 2804); scale bar represents 1 cm (image courtesy of Liam Herringshaw). D, *Lepidocoleus sarlei*, Middle Silurian, Wenlock, Rochester Shale, New York, USA (YPM IP227508); scale bar represents 2 mm. E, phylogeny showing character evolution within machaeridians based upon the scenario of Vinther and Briggs (2009) and the cladistic analysis of Vinther *et al.* (2008*a*, *b*).

character is the dorsal depression where their inner shell plates articulate (Högström *et al.* 2009), the inner shell plates either directly abutting each other or alternating along the midline (Högström and Taylor 2001). Where alternating articulation is present, there are two known types of articulation, either with a tongue and groove hinge (Adrain 1992) or by continuous overlap of the shell plates (Högström and Taylor 2001). Lepidocoleids commonly possess only the inner shell plates (Högström 1997) with the outer shell plates presumably secondarily lost (Vinther and Briggs 2009), leaving every other segment plateless, which was subsequently confirmed by Högström *et al.* (2009). In lepidocoleids that retain both sets of plates as well as in some that lack them, the inner plates articulate by overlapping (Dzik 1986; Högström 1997) as in turrilepadids, suggesting that alternating articulation of shell preceded loss of the outer shell plates and hinged articulation (Fig. 4). CT scanning has also revealed the reduction of parapodia within lepidocoleids (Högström *et al.* 2009), which is a common trend among polychaetes that have transitioned to an infaunal lifestyle.

Vinther and Briggs (2009) presented a phylogenetic hypothesis for machaeridians in which plumulitids are the most primitive and the characters that define the turrilepadids and lepidocoleids are adaptations to infaunality, which is presented as a cladogram in Figure 4. A more primitive position of plumulititids is corroborated by the presence of parapodia in this family, a character that is either known or can be inferred to be absent in the other two families (Högström *et al.* 2009; Vinther and Briggs 2009), and has been indicated in a cladistic analysis by Vinther *et al.* (2008*a*, *b*). The oldest unequivocal record of machaeridians is of plumulitids, which also suggests their more primitive status (but see Herringshaw and Raine 2007). Turrilepadids and lepidocoleids are united by lateral compression of the shell plates, which enclose the body, and lateral displacement of the outer shell plates. We here name this clade Cuniculepadida for their inferred burrowing mode of life. Turrilepadids further modified this body plan by utilizing the concentric rugae characteristic of machaeridian shell plates as a burrowing sculpture (Vinther and Briggs 2009) by aligning them near normal to the body axis (Seilacher 1984).

CONCLUSION AND FUTURE DIRECTIONS

Building on recent advances within the last decade in both palaeontological discoveries and molecular phylogenetics, a novel scenario for the origin of annelids has developed. Machaeridians demonstrate a unique extinct body plan within the phylum and convergent trends towards infaunality and biomineralization in both extinct and living polychaetes. Future studies should strive towards obtaining a more stable phylogeny of annelids and expanding the fossil record of annelids in the Ordovician in particular.

Acknowledgements. We thank Jean-Bernard Caron, Peter Van Roy, David Rudkin, Mats Eriksson, Olev Vinn and Liam Herringshaw for the images used in Figures 2B–C, 3A–B, D–E and 4C, respectively. We also thank Danny Eibye-Jacobsen and Mats Eriksson, and the Editor for their helpful comments that greatly improved the quality of this manuscript.

Editor. Andrew Smith

REFERENCES

ADRAIN, J. 1992. Machaeridian Classification. *Alcheringa*, **16**, 15–32.

—— CHATTERTON, B. and COCKS, L. R. M. 1991. A new species of machaeridian from the Silurian of Podolia, USSR, with a review of the Turrilepadidae. *Palaeontology*, **34**, 637–651.

BOMFLEUR, B., KERP, H., TAYLOR, T., MOESTRUP, O. and TAYLOR, E. 2012. Triassic leech cocoon from Antarctica contains fossil bell animal. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 20971–20974.

BRACCHI, G. and ALESSANDRELLO, A. 2005. Paleodiversity of the free-living polychaetes (Annelida, Polychaeta) and description of new taxa from the Upper Cretaceous Lagerstätten of Haqel, Hadjula and Al-Namoura (Lebanon). *Società Italiana di Scienze Naturali e Museo Civico di Storia Naturale di Milano*, **32**, 1–64.

BRIGGS, D. E. and BARTELS, C. 2010. Annelids from the Lower Devonian Hunsrueck Slate (lower Emsian, Rhenish Massif, Germany). *Palaeontology*, **53**, 215–232.

—— and CLARKSON, E. N. 1987. The first tomopterid, a polychaete from the Carboniferous of Scotland. *Lethaia*, **20**, 257–262.

—— and KEAR, A. J. 1993. Decay and preservation of polychaetes; taphonomic thresholds in soft-bodied organisms. *Paleobiology*, **19**, 107–135.

—— ERWIN, D. H., COLLIER, F. J. and CLARK, C. 1994. *The fossils of the Burgess Shale*. Smithsonian Institution Press, Washington, DC, 238 pp.

BRINKHURST, R. O. 1992. Evolutionary relationships within the Clitellata. *Soil Biology and Biochemistry*, **24**, 1201–1205.

BUDD, G. E. and JENSEN, S. 2000. A critical reappraisal of the fossil record of the bilaterian phyla. *Biological Reviews*, **75**, 253–295.

BUTTERFIELD, N. J. 2003. Exceptional fossil preservation and the Cambrian explosion. *Integrative and Comparative Biology*, **43**, 166–177.

CAMERON, B. 1967. Fossilization of an ancient (Devonian) soft-bodied worm. *Science*, **155**, 1246–1248.

CARON, J.-B., GAINES, R. R., MÁNGANO, M. G., STRENG, M. and DALEY, A. C. 2010. A new Burgess

Shale–type assemblage from the 'thin' Stephen Formation of the southern Canadian Rockies. *Geology*, **38**, 811–814.

CLARK, R. 1964. *Dynamics in metazoan evolution: the origin of the coelom and segments*. Clarendon Press, Oxford, 313 pp.

COLBATH, G. K. 1988. Taphonomy of Recent polychaete jaws from Florida and Belize. *Micropaleontology*, **34**, 83–89.

CONWAY MORRIS, S. 1979. Middle Cambrian polychaetes from the Burgess shale of British Columbia. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, **285**, 227–274.

—— and PEEL, J. 2008. The earliest annelids: Lower Cambrian polychaetes from the Sirius Passet Lagerstatte, Peary Land, North Greenland. *Acta Palaeontologica Polonica*, **53**, 135–146.

DZIK, J. 1986. Turrilepadida and other Machaeridia. *Problematic Fossil Taxa*, **5**, 116–134.

EDGECOMBE, G., GIRIBET, G., DUNN, C., HEJNOL, A., KRISTENSEN, R., NEVES, R., ROUSE, G., WORSAAE, K. and SORENSEN, M. 2011. Higher-level metazoan relationships: recent progress and remaining questions. *Organisms Diversity and Evolution*, **11**, 151–172.

EECKHAUT, I., McHUGH, D., MARDULYN, P., TIEDEMANN, R., MONTEYNE, D., JANGOUX, M. and MILINKOVITCH, M. 2000. Myzostomida: a link between trochozoans and flatworms? *Proceedings of the Royal Society of London, Series B: Biological Sciences*, **267**, 1383–1392.

EIBYE-JACOBSEN, D. 2004. A reevaluation of *Wiwaxia* and the polychaetes of the Burgess Shale. *Lethaia*, **37**, 317–335.

—— and VINTHER, J. 2012. Reconstructing the ancestral annelid. *Journal of Zoological Systematics and Evolutionary Research*, **50**, 85–87.

ERIKSSON, M. and BERGMAN, C. F. 1998. Scolecodont systematics exemplified by the polychaete *Hadoprion cervicornis* (Hinde, 1879). *Journal of Paleontology*, **72**, 477–485.

—— HINTS, O., PAXTON, H. and TONAROVÁ, P. 2013. Ordovician and Silurian polychaete diversity and biogeography. *Geological Society of London, Memoirs*, **38**, 265–272.

ERSÉUS, C. 2005. Phylogeny of oligochaetous Clitellata. *Hydrobiologia*, **535**, 357–372.

—— WETZEL, M. J. and GUSTAVSSON, L. 2008. ICZN rules–a farewell to Tubificidae (Annelida, Clitellata). *Zootaxa*, **1744**, 66–68.

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

FARRELL, U. and BRIGGS, D. 2007. A pyritized polychaete from the Devonian of Ontario. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, **274**, 499–504.

FAUCHALD, K. and JUMARS, P. A. 1979. The diet of worms: a study of polychaete feeding guilds. *Oceanography and Marine Biology – An Annual Review*, **73**, 193–284.

FERRAGUTI, M. and ERSÉUS, C. 1999. Sperm types and their use for a phylogenetic analysis of aquatic clitellates. *Hydrobiologia*, **402**, 225–237.

FITZHUGH, K., SROKA, S., KRUTY, M., HENDERSON, A. and HAY, A. 1997. Polychaete worms. 64–83. *In* SHABICA, C. W. and HAY, A. (eds). *Richardson's guide to the fossil fauna of Mazon creek*. Northeastern Illinois University, Chicago, 308 pp.

GLASBY, C. and TIMM, T. 2008. Global diversity of polychaetes (Polychaeta; Annelida) in freshwater. *Hydrobiologia*, **595**, 107–115.

—— GLASBY, S. and PLEIJEL, F. 2008. Worms by number. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, **275**, 2071–2076.

GLASS, K., ITO, S., WILBY, P., SOTA, T., NAKAMURA, A., BOWERS, C., VINTHER, J., DUTTA, S., SUMMONS, R., BRIGGS, D., WAKAMATSU, K. and SIMON, J. 2012. Direct chemical evidence for eumelanin pigment from the Jurassic period. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 10218–10223.

HALANYCH, K. M., BACHELLER, J. D., AGUINALDO, A., LIVA, S. M., HILLIS, D. M. and LAKE, J. A. 1995. Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. *Science, AAAS Weekly Paper Edition*, **267**, 1641–1642.

HAY, A. A. 2002. Flabelligerida from the Francis Creek shale of Illinois. *Journal of Paleontology*, **76**, 764–766.

HELM, C., BERNHART, S., SIEDERDISSEN, C., NICKEL, B. and BLEIDORN, C. 2012. Deep sequencing of small RNAs confirms an annelid affinity of Myzostomida. *Molecular Phylogenetics and Evolution*, **64**, 198–203.

HERRINGSHAW, L. G. and RAINE, R. J. 2007. The earliest turrilepadid: a machaeridian from the Lower Ordovician of the Northwest Highlands. *Scottish Journal of Geology*, **43**, 97–100.

HILÁRIO, A., CAPA, M., DAHLGREN, T., HALANYCH, K., LITTLE, C., THORNHILL, D., VERNA, C. and GLOVER, A. 2011. New perspectives on the ecology and evolution of siboglinid tubeworms. *PLoS One*, **6**, doi:10.1371/journal.pone.0016309.

HINTS, O. and ERIKSSON, M. 2007. Diversification and biogeography of scolecodont-bearing polychaetes in the Ordovician. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **245**, 95–114.

HÖGSTRÖM, A. E. 1997. Machaeridians from the upper Wenlock (Silurian) of Gotland. *Palaeontology*, **40**, 817–831.

—— and TAYLOR, W. 2001. The machaeridian *Lepidocoleus sarlei* Clarke, 1896, from the Rochester Shale (Silurian) of New York State. *Palaeontology*, **44**, 113–130.

—— BRIGGS, D. E. and BARTELS, C. 2009. A pyritized lepidocoleid machaeridian (Annelida) from the Lower Devonian Hunsrück Slate, Germany. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, **276**, 1981–1986.

HORWOOD, A. 1912. II. – On Archarenicola Rhætica, sp. nov. *Geological Magazine (Decade V)*, **9**, 395–399.

HUANG, D. Y., CHEN, J. Y., VANNIER, J. and SALINAS, J. I. S. 2004. Early Cambrian sipunculan worms from southwest China. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, **271**, 1671–1676.

IPPOLITOV, A., VINN, O., KUPRIYANOVA, E. and JÄGER, M. in press. Written in stone: history of serpulid polychaetes through time. *Memoirs of the Museum of Victoria*.

JONES, D. and THOMPSON, I. 1977. Echiura from the Pennsylvanian Essex fauna of northern Illinois. *Lethaia*, **10**, 317–325.

KUPRIYANOVA, E. and ROUSE, G. W. 2008. Yet another example of paraphyly in Annelida: molecular evidence that Sabellidae contains Serpulidae. *Molecular Phylogenetics and Evolution*, **46**, 1174–1181.

LEISE, E. M. and CLONEY, R. A. 1982. Chiton integument: ultrastructure of the sensory hairs of *Mopalia muscosa* (Mollusca: Polyplacophora). *Cell and Tissue Research*, **223**, 43–59.

LITTLE, C., HERRINGTON, R., MASLENNIKOV, V. and ZAYKOV, V. 1998. The fossil record of hydrothermal vent communities. *Geological Society of London, Special Publications*, **148**, 259–270.

LÜTER, C. 2000. Ultrastructure of larval and adult setae of Brachiopoda. *Zoologischer Anzeiger*, **239**, 75–90.

MANUM, S., BOSE, M. and SAWYER, R. 1991. Clitellate cocoons in fresh-water deposits since the Triassic. *Zoologica Scripta*, **20**, 347–366.

McHUGH, D. 1997. Molecular evidence that echiurans and pogonophorans are derived annelids. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 8006–8009.

MERZ, R. A. and WOODIN, S. A. 2006. Polychaete chaetae: function, fossils, and phylogeny. *Integrative and comparative biology*, **46**, 481–496.

MOSES, D., MATTONI, M., SLACK, N., WAITE, J. and ZOK, F. 2006. Role of melanin in mechanical properties of Glycera jaws. *Acta Biomaterialia*, **2**, 521–530.

MWINYI, A., MEYER, A., BLEIDORN, C., LIEB, B., BARTOLOMAEUS, T. and PODSIADLOWSKI, L. 2009. Mitochondrial genome sequence and gene order of *Sipunculus nudus* give additional support for an inclusion of Sipuncula into Annelida. *BMC Genomics*, **10**, 27.

NAKREM, H., SZANIAWSKI, H. and MORK, A. 2001. Permian–Triassic scolecodonts and conodonts from the Svalis Dome, central Barents Sea, Norway. *Acta Palaeontologica Polonica*, **46**, 69–86.

ORRHAGE, L. 1993. On the microanatomy of the cephalic nervous system of Nereidae (Polychaeta), with a preliminary discussion of some earlier theories on the segmentation of the polychaete brain. *Acta Zoologica*, **74**, 145–172.

—— 2001. On the anatomy of the central nervous system and the morphological value of the anterior end appendages of Ampharetidae, Pectinariidae and Terebellidae (Polychaeta). *Acta Zoologica*, **82**, 57–71.

—— and MÜLLER, M. C. 2005. Morphology of the nervous system of Polychaeta (Annelida). *Hydrobiologia*, **535–536**, 79–111.

PAXTON, H. 2009. Phylogeny of Eunicida (Annelida) based on morphology of jaws. *Zoosymposia*, **2**, 241–264.

—— and ERIKSSON, M. 2012. Ghosts from the past – ancestral features reflected in the jaw ontogeny of the polychaetous annelids *Marphysa fauchaldi* (Eunicidae) and *Diopatra aciculata* (Onuphidae). *GFF*, **134**, 309–316.

PURSCHKE, G., HESSLING, R. and WESTHEIDE, W. 2000. The phylogenetic position of the Clitellata and the Echiura – on the problematic assessment of absent characters. *Journal of Zoological Systematics and Evolutionary Research*, **38**, 165–173.

ROKAS, A. and HOLLAND, P. 2000. Rave genomic changes as a tool for phylogenetics. *Trends in Ecology and Evolution*, **15**, 454–459.

ROUSE, G. W. and FAUCHALD, K. 1997. Cladistics and polychaetes. *Zoologica Scripta*, **26**, 139–204.

—— and PLEIJEL, F. 2001. *Polychaetes*. Oxford University Press, New York, 354 pp.

ROUSSET, V., PLEIJEL, F., ROUSE, G. W., ERSEUS, C. and SIDDALL, M. E. 2007. A molecular phylogeny of annelids. *Cladistics*, **23**, 41–63.

SCHRAM, F. R. 1979. Worms of the Mississippian Bear Gulch Limestone of Central Montana, USA. *Transactions of the San Diego Society of Natural History*, **19**, 107–120.

SEILACHER, A. 1984. Constructional morphology of bivalves: evolutionary pathways in primary versus secondary soft-bottom dwellers. *Palaeontology*, **27**, 207–237.

SERVAIS, T., HARPER, D. A., MUNNECKE, A., OWEN, A. W. and SHEEHAN, P. M. 2009. Understanding the Great Ordovician Biodiversification Event (GOBE): influences of paleogeography, paleoclimate, or paleoecology. *GSA Today*, **19**, 4–10.

SHEN, X., MA, X., REN, J. and ZHAO, F. 2009. A close phylogenetic relationship between Sipuncula and Annelida evidenced from the complete mitochondrial genome sequence of *Phascolosoma esculenta*. *BMC Genomics*, **10**, 136.

SPERLING, E. A., VINTHER, J., MOY, V. M., WHEELER, B. M., SEMON, M., BRIGGS, D. E. G. and PETERSON, K. J. 2009. MicroRNAs resolve an apparent conflict between annelid systematics and their fossil record. *Integrative and Comparative Biology*, **276**, 4315–4322.

STRUCK, T. 2013. The impact of paralogy on phylogenomic studies – a case study on Annelid relationships. *PLoS One*, **8**, doi:10.1371/journal.pone.0062892.

—— SCHULT, N., KUSEN, T., HICKMAN, E., BLEIDORN, C., MCHUGH, D. and HALANYCH, K. 2007. Annelid phylogeny and the status of Sipuncula and Echiura. *BMC Evolutionary Biology*, **7**, 57.

—— PAUL, C., HILL, N., HARTMANN, S., HOSEL, C., KUBE, M., LIEB, B., MEYER, A., TIEDEMANN, R., PURSCHKE, G. and BLEIDORN, C. 2011. Phylogenomic analyses unravel annelid evolution. *Nature*, **471**, 95–113.

SUTTON, M. D., BRIGGS, D. E., SIVETER, D. J. and SIVETER, D. J. 2001. A three-dimensionally preserved fossil polychaete worm from the Silurian of Herefordshire, England. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, **268**, 2355–2363.

SZANIAWSKI, H. 1974. Some Mesozoic scolecodonts congeneric with recent forms. *Acta Palaeontologica Polonica*, **19**, 179–199.

—— 1996. Chapter 12. Scolecodonts, 337–354. *In* JASONIUS, J. and MCGREGOR, D. C. (eds). *Palynology: principles and applications*, *Volume 1*. AASP Foundation, Dallas, TX, 1330 pp.

TARVER, J., SPERLING, E., NAILOR, A., HEIMBERG, A., ROBINSON, J., KING, B., PISANI, D., DONOGHUE, P. and PETERSON, K. 2013. miRNAs: small genes with big potential in metazoan phylogenetics. *Molecular Biology and Evolution*, **30**, 2369–2382.

THOMPSON, I. 1979. Errant polychaetes (Annelida) from the Pennsylvanian Essex fauna of northern Illinois. *Palaeontographica Abteilung A*, **163**, 169–199.

—— and JOHNSON, R. G. 1977. New fossil polychaete from Essex, Illinois. *Fieldiana*, **33**, 471–487.

TZETLIN, A. and FILIPPOVA, A. 2005. Muscular system in polychaetes (Annelida). *Hydrobiologia*, **535**, 113–126.

—— and PURSCHKE, G. 2006. Fine structure of the pharyngeal apparatus of the pelagosphera larva in *Phascolosoma agassizii* (Sipuncula) and its phylogenetic significance. *Zoomorphology*, **125**, 109–117.

VINN, O. and LUQUE, J. 2013. First record of a pectinariid-like (Polychaeta, Annelida) agglutinated worm tube from the Late Cretaceous of Colombia. *Cretaceous Research*, **41**, 107–110.

—— and MUTVEI, H. 2009. Calcareous tubeworms of the Phanerozoic. *Estonian Journal of Earth Sciences*, **58**, 286–296.

—— and ZATON, M. 2012. Inconsistencies in proposed annelid affinities of early biomineralized organism *Cloudina* (Ediacaran): structural and ontogenetic evidences. *Carnets de Géologie, Brest, Article*, **3**, 39–47.

VINTHER, J. and BRIGGS, D. G. 2009. Machaeridian locomotion. *Lethaia*, **42**, 357–364.

—— and RUDKIN, D. 2010. The first articulated specimen of *Plumulites canadensis* (Woodward 1889) from the upper Ordovician of Ontario, with a review of the anterior region of Plumulitidae (Annelida: Machaeridia). *Palaeontology*, **53**, 327–334.

—— BRIGGS, D. G., PRUM, R. O. and SARANATHAN, V. 2008*a*. The colour of fossil feathers. *Biology Letters*, **4**, 522–525.

—— VAN ROY, P. and BRIGGS, D. 2008*b*. Machaeridians are Palaeozoic armoured annelids. *Nature*, **451**, 185–188.

—— EIBYE-JACOBSEN, D. and HARPER, D. A. 2011. An Early Cambrian stem polychaete with pygidial cirri. *Biology Letters*, **7**, 929–932.

WANNINGER, A., KRISTOF, A. and BRINKMANN, N. 2009. Sipunculans and segmentation. *Communicative and Integrative Biology*, **2**, 56–59.

WEIGERT, A., HELM, C., MEYER, M., NICKEL, B., ARENDT, D., HAUSDORF, B., SANTOS, S. R., HALANYCH, K. M., PURSCHKE, G. and BLEIDORN, C. 2014. Illuminating the base of the annelid tree using transcriptomics. *Molecular Biology and Evolution*, **31**, 1391–1401.

WELCH, J. R. 1976. Phosphannulus on Paleozoic crinoid stems. *Journal of Paleontology*, **50**, 218–225.

WESTHEIDE, W. 1997. The direction of evolution within the Polychaeta. *Journal of Natural History*, **31**, 1–15.

WHITTLE, R., GABBOTT, S., ALDRIDGE, R. and THERON, J. 2008. Late Ordovician (Hirnantian) scolecodont clusters from the Soom Shale Lagerstatte, South Africa. *Journal of Micropalaeontology*, **27**, 147–159.

ZHADAN, A., VORTSEPNEVA, E. and TZETLIN, A. 2014. Three-dimensional reconstruction of the musculature of *Cossura pygodactylata* Jones, 1956 (Annelida: Cossuridae). *Zoologischer Anzeiger*, **253**, 181–191.

ZRZAVY, J., RIHA, P., PIALEK, L. and JANOUSKOVEC, J. 2009. Phylogeny of Annelida (Lophotrochozoa): total-evidence analysis of morphology and six genes. *BMC Evolutionary Biology*, **9**, 1–14.

BIOLOGY
LETTERS

rsbl.royalsocietypublishing.org

Research

Cite this article: O'Reilly JE, Puttick MN, Parry L, Tanner AR, Tarver JE, Fleming J, Pisani D, Donoghue PCJ. 2016 Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. *Biol. Lett.* **12**: 20160081.
http://dx.doi.org/10.1098/rsbl.2016.0081

Received: 28 January 2016
Accepted: 21 March 2016

Subject Areas:
palaeontology, taxonomy and systematics, evolution

Keywords:
parsimony, Bayesian, likelihood, phylogenetics, morphology

Authors for correspondence:
Davide Pisani
e-mail: davide.pisani@bristol.ac.uk
Philip C. J. Donoghue
e-mail: phil.donoghue@bristol.ac.uk

[†] These authors contributed equally to this study.

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsbl.2016.0081 or via http://rsbl.royalsocietypublishing.org.

THE ROYAL SOCIETY
PUBLISHING

Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data

Joseph E. O'Reilly[1,†], Mark N. Puttick[1,†], Luke Parry[1], Alastair R. Tanner[1,2], James E. Tarver[1], James Fleming[1], Davide Pisani[1,2] and Philip C. J. Donoghue[1]

[1] School of Earth Sciences, and [2] School of Biological Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol BS8 1TQ, UK

JEO, 0000-0001-9775-253X; DP, 0000-0003-0949-6682; PCJD, 0000-0003-3116-7463

Different analytical methods can yield competing interpretations of evolutionary history and, currently, there is no definitive method for phylogenetic reconstruction using morphological data. Parsimony has been the primary method for analysing morphological data, but there has been a resurgence of interest in the likelihood-based Mk-model. Here, we test the performance of the Bayesian implementation of the Mk-model relative to both equal and implied-weight implementations of parsimony. Using simulated morphological data, we demonstrate that the Mk-model outperforms equal-weights parsimony in terms of topological accuracy, and implied-weights performs the most poorly. However, the Mk-model produces phylogenies that have less resolution than parsimony methods. This difference in the accuracy and precision of parsimony and Bayesian approaches to topology estimation needs to be considered when selecting a method for phylogeny reconstruction.

1. Introduction

Morphology once provided the only means of inferring evolutionary trees, but it was effectively rendered obsolete by molecular sequence data and the development of sophisticated molecular evolutionary models for phylogenetic analysis [1]. However, with the recognition that fossil species are integral to correctly inferring patterns of character evolution and changes in diversity, as well as in establishing evolutionary timescales, morphological data are enjoying a phylogenetic renaissance [2], allowing fossil species to be assigned to their correct branches in the Tree of Life. Methods for phylogenetic analysis of morphological data remain underdeveloped and though likelihood models are available that may more accurately accommodate the vagaries of morphological datasets [3], including high rates of heterogeneity and a preponderance of missing data [4], parsimony remains the method of choice, principally perhaps as a consequence of tradition. Indeed, a recent simulation-based study by Wright & Hillis [5] demonstrated that a Bayesian implementation of Lewis's Mk-model [3] strongly outperforms parsimony, especially when rates of character change are high, or when relatively few characters are analysed. The conclusions drawn by Wright & Hillis [5] were based on data effectively simulated using the Mk-model, potentially biasing the test in favour of the Mk-model. Furthermore, they did not consider whether the simulated data exhibited realistic levels of homoplasy, analysed unrealistically large simulated datasets, and evaluated only the relative performance of

Table 1. The differences in median and the 95th percentile range of Robinson–Foulds values between the Mk and both parsimony models are greater in the full dataset compared with the realistic homoplasy subsets. mk, Bayesian Mk model; ew, equal-weights parsimony; iw, implied weights parsimony and its attendant *K* values.

	100 characters	100 characters CI	350 characters	350 characters CI	1000 characters	1000 characters CI
mk	45 (29–64)	40.5 (28.2–62.5)	20 (10–51)	19.5 (10.2–57.3)	19.5 (10.2–57.3)	11 (5–27.8)
ew	61 (31–98)	53 (29–91.8)	27 (12–70)	28 (12–74.8)	28 (12–74.8)	16 (6.2–43.7)
iw k2	89 (39–119)	77 (38.2–117.7)	36 (18–76)	36 (17.2–81.3)	36 (17.2–81.3)	19.5 (10–35.7)
iw k3	76 (38–112)	69 (36.4–108)	32 (16–69)	34 (15.2–70)	34 (15.2–70)	18 (9.2–35.7)
iw k5	68 (36–104)	61 (32.2–102)	30 (14–66)	31.5 (15.2–68)	31.5 (15.2–68)	18 (9–34)
iw k10	63 (34–100)	55.5 (32–98)	28 (13–68)	30 (15.2–69.7)	30 (15.2–69.7)	16 (8–34)
iw k20	64 (34–100)	53 (33–97.8)	28 (14–68)	30 (13.2–71.7)	30 (13.2–71.7)	17 (8–39.3)
iw k200	65 (34–100)	55 (32.2–97.7)	28 (14–72)	30.5 (15–76)	30.5 (15–76)	18 (8–44)

equal-weights parsimony when morphological data are now commonly analysed under implied-weights parsimony [6].

In an attempt to evaluate the relative performance of likelihood and parsimony methods for the phylogenetic analysis of discrete character morphological data, we simulated datasets of 100, 350 and 1000 discrete morphological characters using a modified HKY85 model, discriminating datasets that failed to meet expected levels of homoplasy. We evaluated the relative performance of equal-weights parsimony, implied-weights parsimony and model-based methods of phylogenetic analysis in terms of their ability to recover the tree used to simulate the data. We found that the Mk-model performs best in the analysis of all simulated datasets, largely because the Bayesian consensus trees are poorly resolved. Equal-weights parsimony exhibits lower levels of accuracy but this is combined with higher resolution. Implied-weights parsimony performed most poorly of all the methods considered.

2. Material and methods

To simulate binary morphological data, we used the HKY + $\Gamma_{\text{continuous}}$ model to generate nucleotide data which we translated into purines (0) and pyrimidines (1)—R/Y coding. The recoded HKY-model possesses an uneven equilibrium distribution of state frequencies, resulting in structurally realistic morphological matrices while facilitating violation of assumptions of the Mk-model; thus, our data are not biased in favour of either method of phylogenetic inference. Initial tests were performed to determine values for the model parameters which produce binary data with empirically observed levels of homoplasy [7]. Following [5], data were simulated using the lissamphibian tree presented in [8], yielding datasets of 100, 350 and 1000 characters; most real morphological datasets contain in the order of 100 characters, but we included 350 and 1000 character matrices to investigate the effect of scaling and for ease of comparison to [5]. In total, 100 unique underlying substitution rates were drawn from a U(0.1,10) distribution, facilitating rates spanning two orders of magnitude. For each substitution rate, 10 unique matrices were produced, modelling among-character rate heterogeneity as gamma distributed uniquely within each matrix.

Matrices were analysed with the Mk + Γ model using default priors in MrBayes v. 3.2 [9], and both standard and implied-weights parsimony in TNT [10]. The Mk-model is more suitable for our simulated data than the Mkv-model as we did not strip invariant sites from the final matrices. Majority-rule consensus trees were produced for each method. For implied-weights parsimony, we used a range of *K*-values: 2, 3, 5, 10, 20 and 200. As the underlying substitution rate is varied, the per-matrix level of homoplasy may violate the empirically observed range; to produce the most empirically justified morphological matrices, we implemented an empirically derived minimum consistency index (CI) cut-off of 0.26 [7] for each simulated dataset and repeated analyses for these treated matrices (electronic supplementary material, figure S1). This cut-off reduced the size of the datasets to 128 (100 characters), 149 (350 characters) and 126 (1000 characters) matrices. In-depth description of the initial parameter value tests and further details of matrix generation are presented in the electronic supplementary material.

The accuracy of topologies estimated by the different reconstruction techniques was assessed using the Robinson–Foulds distance [11] from the generator tree. We also explored the relationship between resolution of output trees, measured by the number of nodes per tree.

3. Results

The Mk-model achieved the highest levels of accuracy across all datasets. Median Robinson–Foulds distances are lower for the Mk-model compared with both equal-weights and implied-weights parsimony (table 1 and figure 1), and for all approaches, accuracy of topology reconstruction increases with increasing dataset size. Furthermore, equal-weights parsimony out-performs implied-weights parsimony for all datasets and values of *K*, but this is less pronounced for the 1000 character dataset (table 1). For convenience, all further results for implied weights are for $K = 2$.

The same relative performance of the phylogenetic reconstruction methods is seen when considering only those datasets exhibiting realistic levels of homoplasy. The median Robinson–Foulds distance for the Mk-model is still lowest for each dataset, but the median and range of Robinson–Foulds distances for equal and implied-weights parsimony are closer to the distribution seen from the Mk-model (table 1 and figure 1). Additionally, for a given dataset, there is a similar Robinson–Foulds distance regardless of the reconstruction method employed (electronic supplementary material, figure S2). Unless otherwise stated, all subsequent results are from the subset of datasets exhibiting realistic levels of homoplasy.

The higher accuracy (lower Robinson–Foulds values) of the Mk-model against other methods for 100 and 350 characters is due to trees being less resolved (figure 2). The density of Robinson–Foulds distance is lower for the Mk compared with equal weights, which itself is lower than implied weights, but both equal and implied weights achieve higher levels of

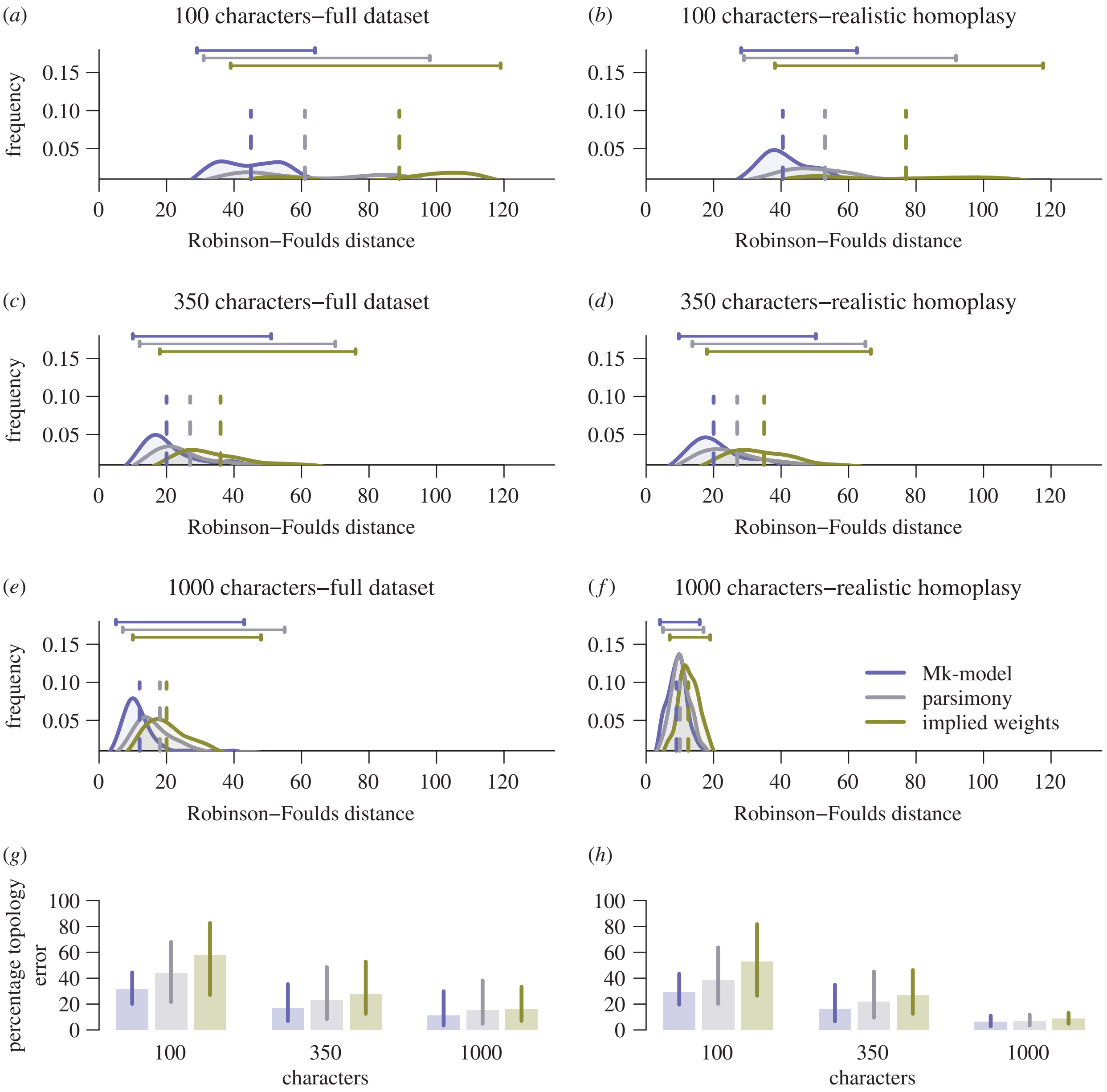


Figure 1. Mk tree reconstructions (blue) outperform equal-weights parsimony (grey) and implied-weights parsimony (green) for 100, 350 and 1000 characters (*a*,*c*,*e*,*g*), and these differences remain in the subset of the simulated data matrices that exhibit realistic levels of homoplasy (*b*,*d*,*f*,*h*). Bars above the plots mark the 95th percentile range for each method, and dashed vertical lines show the median values. Percentage topology error (*g*,*h*) is the Robinson–Foulds value of the reconstructed tree compared with the worst possible value, as shown in [5].

precision (number of nodes reconstructed). These differences are negligible in the 1000 character datasets (figure 2).

There is a significant overlap in the set of nodes correctly recovered across methods, when mapped against the reference phylogeny (figure 2; electronic supplementary material, figure S3). In particular, for all methods there is a trend for nodes closer to the root to be more accurately estimated in small datasets, but this relationship decreases as the number of characters increases (electronic supplementary material, table S2 and figures S2, S4, S5). The percentage of times a node from the reference tree was accurately reconstructed showed a strong correlation for 100 and 350 characters, but decreases with 1000 characters (electronic supplementary material, table S2).

4. Discussion

Only minor differences are seen in the accuracy of phylogenetic topology reconstruction between the Bayesian implementation of the Mk-model and parsimony methods. Our findings both support and contradict elements of the results of Wright & Hillis [5] in that we can corroborate their observation, that the Mk-model outperforms equal-weights parsimony in accuracy, but the Mk-model achieves this at the expense of precision. Unexpectedly, implied-weights parsimony is less effective than either equal-weights parsimony or the Mk-model, in datasets with small numbers of characters. Implied-weights parsimony outperforms equal-weights parsimony only in the analyses of unrealistically large datasets. These results challenge the increasingly common view that implied-weighting better accommodates homoplasy than does equal-weights parsimony [6], and this result is true for a range of *K*-values (table 1).

In comparison with the other approaches, equal-weights parsimony analyses of the datasets exhibiting realistic levels of homoplasy and large number of characters yield a set of trees with a longer tailed distribution of Robinson–Foulds distances. In large part, this reflects estimation of a small quantity of trees markedly different from the generating tree (figure 1).

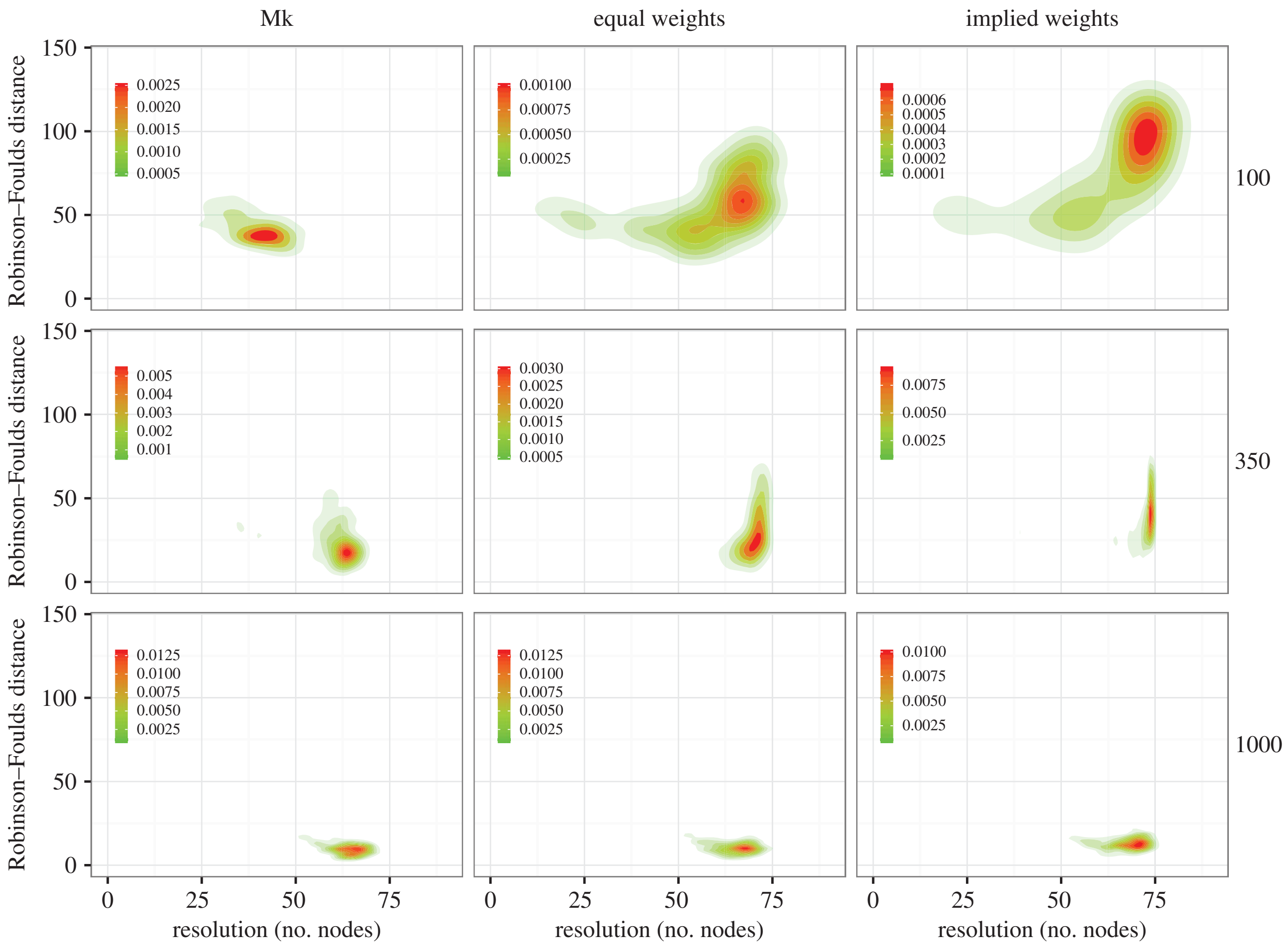


Figure 2. The Mk model exhibits higher accuracy with lower precision than parsimony methods; these results are less clear as more characters are added. Contour plots of Robinson–Foulds distances against the number of resolved nodes in each tree; the contours represent the density of the distribution of trees.

Inaccuracy in topological estimation is more prevalent towards the tips in all analyses, with the inclusion of more characters reducing the intensity of this phenomenon. For this effect to be completely removed, it would require the analysis of well over 1000 empirically justifiable characters, a number that is rarely achieved for morphological datasets. The accuracy of node reconstruction is correlated significantly between all three techniques, demonstrating that most nodes in the tree that were difficult to resolve for one method were difficult to resolve for all. This phenomenon is observed across all character quantities and suggests a general difficulty in accurately estimating topology given the same data.

Our results can be interpreted to advocate use of the Mk-model over parsimony methods in the analysis of discrete morphological data. Parsimony methods produce precision without the accuracy achieved by the Mk-model and precision without accuracy is a poor basis for any science. We anticipate that the implementation of the Mk-model within a maximum-likelihood framework will exhibit levels of accuracy and precision more comparable to the parsimony methods, simply because it estimates a single, fully resolved topology. Integration over parameters while producing an acceptable level of accuracy is a quality of Bayesian inference, and our Mk-model results are probably dependent on a Bayesian implementation. While comparative phylogenetic methods often require fully resolved trees, these may be accommodated through analyses using the posterior sample of trees estimated using the Mk-model. Therefore, the prior requirement of a fully resolved tree need not necessarily lead to a preference for parsimony over the Mk-model.

In comparison to parsimony methods, the Mk-model has undergone little development since its conception [12,13], while attempts to improve the performance of parsimony methods, like implied-weights parsimony [3], have not led to increased accuracy (table 1). Thus, model-based phylogenetics can be expected to offer more opportunity for development, e.g. through relaxing the assumption of symmetrically distributed stationary distribution of character states [12,13] and improvement in the accuracy of phylogeny estimation from discrete character data. We suggest, however, that more focus should be invested in assessing whether the data are sufficiently informative to discriminate between competing phylogenetic hypotheses.

5. Conclusion

Phylogenies produced using likelihood models are more accurate than parsimony approaches, but have lower precision. Likelihood models offer greater scope for development in attempting to achieve greater accuracy but, in the interim, we suggest that phylogeneticists should consider the aims of their analyses when choosing the appropriate method.

Data accessibility. Data used in this manuscript are archived at http://dx.doi.org/10.5061/dryad.10qf3 and as the electronic supplementary material.

Authors' contributions. All authors made substantial contributions to (i) conception and design, or acquisition of data, or analysis and interpretation of data and (ii) drafting the article or revising it critically for important intellectual content; gave final approval of the version to be published, and agreed to be accountable for all aspects

of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Competing interests. We have no competing interests.

Funding. This research was funded by NERC (NE/L501554/1 to J.E.O'R. and L.P.; NE/K500823/1 to M.N.P.; NE/L002434/1 to J.F.; NE/N003438/1 to P.C.J.D.), BBSRC (BB/N000919/1 to P.C.J.D.), the University of Bristol (STaR scholarship to A.R.T.), Royal Society Wolfson Research Merit Award (P.C.J.D.) and the John Templeton Foundation (43915 to D.P.).

Acknowledgements. We thank the other members of the Bristol Palaeobiology research group for discussion, as well as April Wright and Peter Wagner for the constructive reviews that helped shape the published version of our study.

References

1. Scotland RW, Olmstead RG, Bennett JR. 2003 Phylogeny reconstruction: the role of morphology. *Syst. Biol.* **52**, 539–548. (doi:10.1080/10635150390223613)
2. Lee MSY, Palci A. 2015 Morphological phylogenetics in the genomic age. *Curr. Biol.* **25**, R922–R929. (doi:10.1016/j.cub.2015.07.009)
3. Lewis PO. 2001 A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* **50**, 913–925. (doi:10.1080/106351501753462876)
4. Wagner PJ. 2012 Modelling rate distributions using character compatibility: implications for morphological evolution among fossil invertebrates. *Biol. Lett.* **8**, 143–146. (doi:10.1098/rsbl.2011.0523)
5. Wright AM, Hillis DM. 2014 Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. *PLoS ONE* **9**, e109210. (doi:10.1371/journal.pone.0109210)
6. Goloboff PA, Carpenter JM, Arias JS, Rafael D, Esquivel M. 2008 Weighting against homoplasy improves phylogenetic analysis of morphological data sets. *Cladistics* **24**, 758–773. (doi:10.1111/j.1096-0031.2008.00209.x)
7. Sanderson MJ, Donoghue MJ. 1989 Patterns of variation in levels of homoplasy. *Evolution* **43**, 1781–1795. (doi:10.2307/2409392)
8. Pyron RA. 2011 Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. *Syst. Biol.* **60**, 466–481. (doi:10.1093/sysbio/syr047)
9. Ronquist F *et al.* 2012 MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542. (doi:10.1093/sysbio/sys029)
10. Goloboff P, Farris S, Nixon K. 2000 *TNT* (*Tree analysis using New Technology*). Tucumán, Argentina: published by the authors.
11. Robinson DR, Foulds LR. 1981 Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147. (doi:10.1016/0025-5564(81)90043-2)
12. Klopfstein S, Vilhelmsen L, Ronquist F. 2015 A nonstationary Markov model detects directional evolution in hymenopteran morphology. *Syst. Biol.* **64**, 1089–1103. (doi:10.1093/sysbio/syv052)
13. Wright AM, Lloyd GT, Hillis D. 2015 Modeling character change heterogeneity in phylogenetic analyses of morphology through the use of priors. *Syst. Biol.* syv122. (doi:10.1093/sysbio/syv122)

PROCEEDINGS B

rspb.royalsocietypublishing.org

Research

Cite this article: Puttick MN *et al.* 2017 Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data. *Proc. R. Soc. B* **284**: 20162290. http://dx.doi.org/10.1098/rspb.2016.2290

Received: 19 October 2016
Accepted: 5 December 2016

Subject Category:
Palaeobiology

Subject Areas:
evolution, palaeontology, taxonomy and systematics

Keywords:
phylogeny, Bayesian, parsimony, cladistics, morphology, palaeontology

Authors for correspondence:
Davide Pisani
e-mail: davide.pisani@bristol.ac.uk
Philip C. J. Donoghue
e-mail: phil.donoghue@bristol.ac.uk

Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data

Mark N. Puttick[1,3,†], Joseph E. O'Reilly[1,†], Alastair R. Tanner[2], James F. Fleming[1], James Clark[1], Lucy Holloway[1], Jesus Lozano-Fernandez[1,2], Luke A. Parry[1], James E. Tarver[1], Davide Pisani[1,2] and Philip C. J. Donoghue[1]

[1]School of Earth Sciences, and [2]School of Biological Sciences, University of Bristol, Life Sciences Building, 24 Tyndall Avenue, Bristol BS8 1TQ, UK
[3]Department of Life Sciences, Natural History Museum, Cromwell Road, London SW7 5BD, UK

MNP, 0000-0002-1011-3442; JEO, 0000-0001-9775-253X; JC, 0000-0003-2896-1631; LH, 0000-0003-1603-2296; JL-F, 0000-0003-3597-1221; LAP, 0000-0002-3910-0346; PCJD, 0000-0003-3116-7463

Morphological data provide the only means of classifying the majority of life's history, but the choice between competing phylogenetic methods for the analysis of morphology is unclear. Traditionally, parsimony methods have been favoured but recent studies have shown that these approaches are less accurate than the Bayesian implementation of the Mk model. Here we expand on these findings in several ways: we assess the impact of tree shape and maximum-likelihood estimation using the Mk model, as well as analysing data composed of both binary and multistate characters. We find that all methods struggle to correctly resolve deep clades within asymmetric trees, and when analysing small character matrices. The Bayesian Mk model is the most accurate method for estimating topology, but with lower resolution than other methods. Equal weights parsimony is more accurate than implied weights parsimony, and maximum-likelihood estimation using the Mk model is the least accurate method. We conclude that the Bayesian implementation of the Mk model should be the default method for phylogenetic estimation from phenotype datasets, and we explore the implications of our simulations in reanalysing several empirical morphological character matrices. A consequence of our finding is that high levels of resolution or the ability to classify species or groups with much confidence should not be expected when using small datasets. It is now necessary to depart from the traditional parsimony paradigms of constructing character matrices, towards datasets constructed explicitly for Bayesian methods.

1. Introduction

The fossil record affords the only direct insight into evolutionary history of life on the Earth, but the incomplete preservation and temporal distribution of fossils has long prompted biologists to seek alternative perspectives, such as molecular phylogenies of living species, eschewing palaeontological evidence altogether [1]. However, there is increasing acceptance that analyses of historical diversity cannot be made without phylogenies that incorporate fossil species [2,3] and calibrating molecular phylogenies to time cannot be achieved effectively without recourse to the fossil record [4]. Integrating fossil and living species has become the grand challenge and there has been a modest proliferation of phylogenetic approaches to the analysis of phenotypic data. While conventional parsimony remains the most widely employed method, alternative parsimony [5] and probabilistic [6] models have been developed to better accommodate heterogeneity in

[†]These authors contributed equally to this study.

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3653186.

THE ROYAL SOCIETY
PUBLISHING

the rate of evolution among characters and across phylogeny. Unfortunately, these competing methods invariably yield disparate phylogenetic hypotheses among which it is difficult to discriminate as the true tree is never known for empirical data.

A number of studies have attempted to establish the efficacy of competing phylogenetic methods using data simulated from known trees [7–9], finding that the probabilistic Mkv model outperforms parsimony methods, among which, conventional equal-weights parsimony (EW-Parsimony) performs best. However, these studies were potentially biased by their experimental design: (i) two of the studies employed a generating tree that was unresolved and, therefore, biased against parsimony methods which recover resolved trees; (ii) these studies did not discriminate between the impact of the probabilistic model and its implementation in a Bayesian framework; (iii) based on single empirical trees, the impact of tree symmetry, which is known to confound phylogeny estimation [10], was not explored; and (iv) only binary characters were considered, whereas empirical datasets are commonly a mixture of binary and multistate characters. Therefore, we compare the performance of EW-Parsimony, implied-weights parsimony (IW-Parsimony), maximum-likelihood and Bayesian implementations of the Mk model, based on datasets with different numbers of characters, comprising binary and multistate characters and simulated on a fully balanced and a maximally imbalanced phylogenetic tree. We find that Bayesian inference outperforms all other methods, while EW-Parsimony performs better than IW-Parsimony, and maximum likelihood performs worst of all. We apply these competing phylogenetic methods to empirical morphological datasets of similar sizes to our simulated datasets and explore the efficacy of the ensuing phylogenetic hypotheses in the light of the conclusions derived from our simulation-based study.

2. Material and methods

(a) Simulation of morphological matrices

We simulated data on two 32-taxon generating trees at the extremes of tree symmetry: one fully asymmetrical and one fully symmetrical (see electronic supplementary material, figure S1). For each tree, we simulated matrices of three sizes: 100, 350 and 1000 characters. We generated matrices using the HKY + Γ Continuous model of molecular substitution, with $\kappa = 2$, the shape (set equal to rate) of the gamma distribution and underlying substitution rate for each replicate sampled from independent and identically distributed exponential distributions with a mean of 1, and character state stationary frequencies fixed as $\pi = [0.2,0.2,0.3,0.3]$. We used a fixed and uneven stationary distribution of nucleotide frequencies to ensure our simulation model did not collapse into the Mk model, as this would bias the analysis in favour of Mk model-based approaches. We simulated 1000 replicate matrices with unique substitution parameters for each tree and each character number, resulting in a total of 6000 matrices. We set two types of character within each matrix, binary and multistate, and we simulated a proportion of 55 binary : 45 multistate characters, based on the mean ratio found in a survey of empirical morphological data matrices [11]. We established binary characters by converting data simulated under the HKY model to R/Y coding (i.e. 0/1): morphological multistate characters were simulated by converting DNA bases to integers.

To ensure that our simulated data are realistic, we generated each set of 1000 unique replicate matrices such that the among-matrix distribution of homoplasy approximated the distribution of empirical homoplasy, characterized by the consistency index (CI), reported by Sanderson & Donoghue [12]. To approximate this distribution of homoplasy, we placed the Sanderson and Donoghue data into quantized bins of CI spanning 0.05, between the empirical bounds of 0.26 and 1.0, and simulated matrices until we matched this expected density per bin (electronic supplementary material, figure S2).

The code used to simulate these data is available in the electronic supplementary material.

(b) Phylogenetic analysis

We analysed the simulated matrices with EW-Parsimony, IW-Parsimony ($k = 2$) and the Mk model [6] under both maximum-likelihood and Bayesian implementations. EW-Parsimony and IW-Parsimony estimation of topology was performed in TNT [13]. We used the Mk + Γ model for maximum-likelihood estimation of topology in RAxML v. 7.2 [14], and Bayesian estimation of topology in MrBayes v. 3.2 [15]. As the approximate likelihood calculation of RAxML may be distant from the true likelihood [16], we conducted a sensitivity test by re-analysing a subset of our data with the likelihood implementation of the Mk model in IQ-tree [17]; both methods gave effectively identical results, indicating results from the likelihood Mkv model are not software specific.

The Mkv model is inappropriate due to the lack of acquisition bias in the simulated data. For maximum-likelihood and Bayesian analyses, we applied the discretized gamma distribution model to account for between-character rate heterogeneity. For Bayesian analyses, the posterior distribution was sampled 1 million times by four chains using the Metropolis-coupled Markov-chain Monte Carlo algorithm with every 100th sample stored, resulting in 10 000 samples; two independent runs were performed for each replicate and the two resulting posterior samples were combined after qualitative assessment of convergence. For parity, we characterized the result of all phylogenetic methods as the majority-rule consensus of resultant tree samples. We did not employ bootstrap methods to measure support for parsimony and likelihood analyses because phenotypic data does not meet the assumption that phylogenetic signal is distributed randomly among characters.

We used the Robinson–Foulds metric [18] to compare the similarity of estimated topologies against their respective generating tree. We also noted the per-node resolution, and the variation of node accuracy across the topology.

(c) Empirical analyses

We analysed four published palaeontological phenotype character matrices that encompass a range of character numbers and a diverse sample of taxa from the Tree of Life [19–22]. We resolved any ambiguities in character coding to their most derived state for each matrix to make analyses compatible across the different phylogenetic methods, facilitating comparison of results. We analysed each matrix by applying the same settings used to analyse our simulated matrices: EW-Parsimony, IW-Parsimony, as well as Bayesian and maximum-likelihood implementations of the Mk model. Empirical morphological matrices are rarely constructed to contain invariant or parsimony uninformative characters. Therefore, the Mkv extension of the Mk model, which uses conditional likelihood to correct for such acquisition biases, is more appropriate than the Mk model for analysis of these empirical data matrices [6].

3. Results

(a) Simulated data

Accuracy is higher for trees inferred from data simulated on a symmetrical topology compared with trees

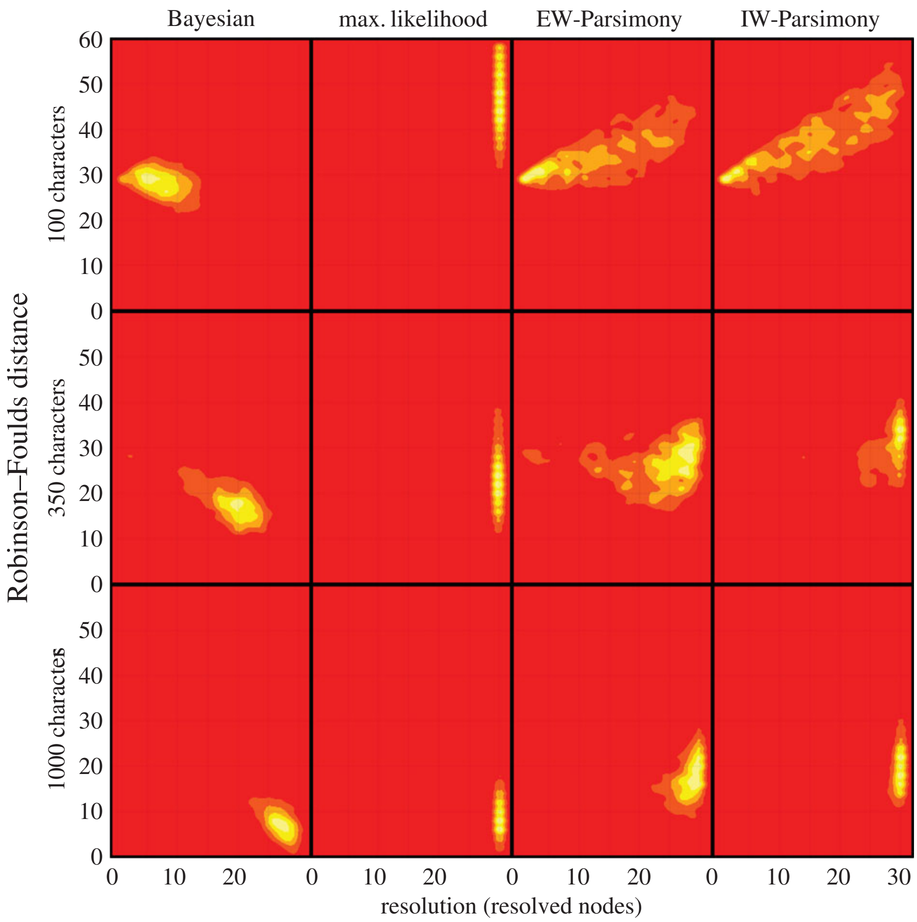


Figure 1. Contour plots of Robinson–Foulds distance against phylogenetic resolution, indicating the higher accuracy of Bayesian implementations against all other methods with data generated on the asymmetrical phylogeny. The spectrum of red to yellow, reflect lower to higher density of trees. As the number of characters increases all methods converge on the correct phylogeny, although Bayesian phylogenies are generally the least resolved. The other methods achieve higher resolution but at a cost of lower accuracy. Data generated on the symmetrical phylogeny shows similar patterns but with much less variance and higher accuracy for all iterations; this lack of variance means point estimates cannot be shown as density estimates. (Online version in colour.)

estimated from data simulated on the asymmetrical topology (cf. figures 2 and 3). Bayesian consensus phylogenies are generally the least well-resolved (figure 1). All methods estimated topologies with greater accuracy as the number of analysed characters increased (figures 2 and 3; electronic supplementary material, table S5–S7). All methods, apart from maximum likelihood, produced phylogenies with greater resolution with higher numbers of characters (figure 1).

For all implementations and dataset sizes, the Bayesian implementation of the Mk model achieves higher accuracy compared with other methods (table 1; figures 1–3). The two parsimony methods achieved the next highest levels of accuracy, EW-Parsimony achieving greater accuracy than IW-Parsimony. Maximum likelihood was the least accurate method for topology reconstruction for both the symmetrical and asymmetrical phylogenies (table 1). The relative accuracy of these phylogenetic methods remains the same across all dataset sizes and the two simulation topologies (table 1; figures 1–3).

Nodes closer to the tips are significantly more accurately reconstructed in the asymmetrical phylogenies across all dataset sizes (table 2 and figure 2; electronic supplementary material, figure S8). In the symmetrical trees, there was no significant correlation between distance from the tips and the accuracy of node reconstruction, except in the maximum-likelihood analysis of 100 characters (figure 2 and table 2).

(b) Empirical phylogenies

Patterns of resolution achieved from the simulated datasets are similar for the empirical datasets. The Bayesian implementation of the Mk model estimates the least resolved phylogenies and maximum likelihood produces fully resolved trees (full trees are shown electronic supplementary material, figure S9–S15).

Kulindroplax, from the Sutton *et al.* [22] dataset, is supported as a crown-mollusc based on maximum likelihood, EW-Parsimony and IW-Parsimony (figure 4*a*–*d*). The results of the IW-Parsimony analysis are most similar to the original results [22], with *Kulindroplax* resolved as a crown-aplacophoran; maximum-likelihood analysis of the dataset resolved *Kulindroplax* as the stem-aplacophoran. The result of the Bayesian analysis of the dataset is largely unresolved, and *Kulindroplax* is not discriminated as a member of any clade within molluscs or even as a member of total-group Mollusca.

The anthophyte hypothesis (non-monophyletic gymnosperms sister to seed ferns plus angiosperms) recovered by Hilton & Bateman [19] is supported by our EW-Parsimony and maximum-likelihood analyses of their dataset which recovered a paraphyletic seed ferns plus Gnetophyta as sister to angiosperms (figure 4*f*,*g*); the results of Bayesian and IW-Parsimony analyses of the same dataset contradict the anthophyte hypothesis (figure 4*e*,*h*). The Bayesian analysis produced a non-monophyletic gymnosperms with the relationships between them and seed ferns unresolved with the exception of

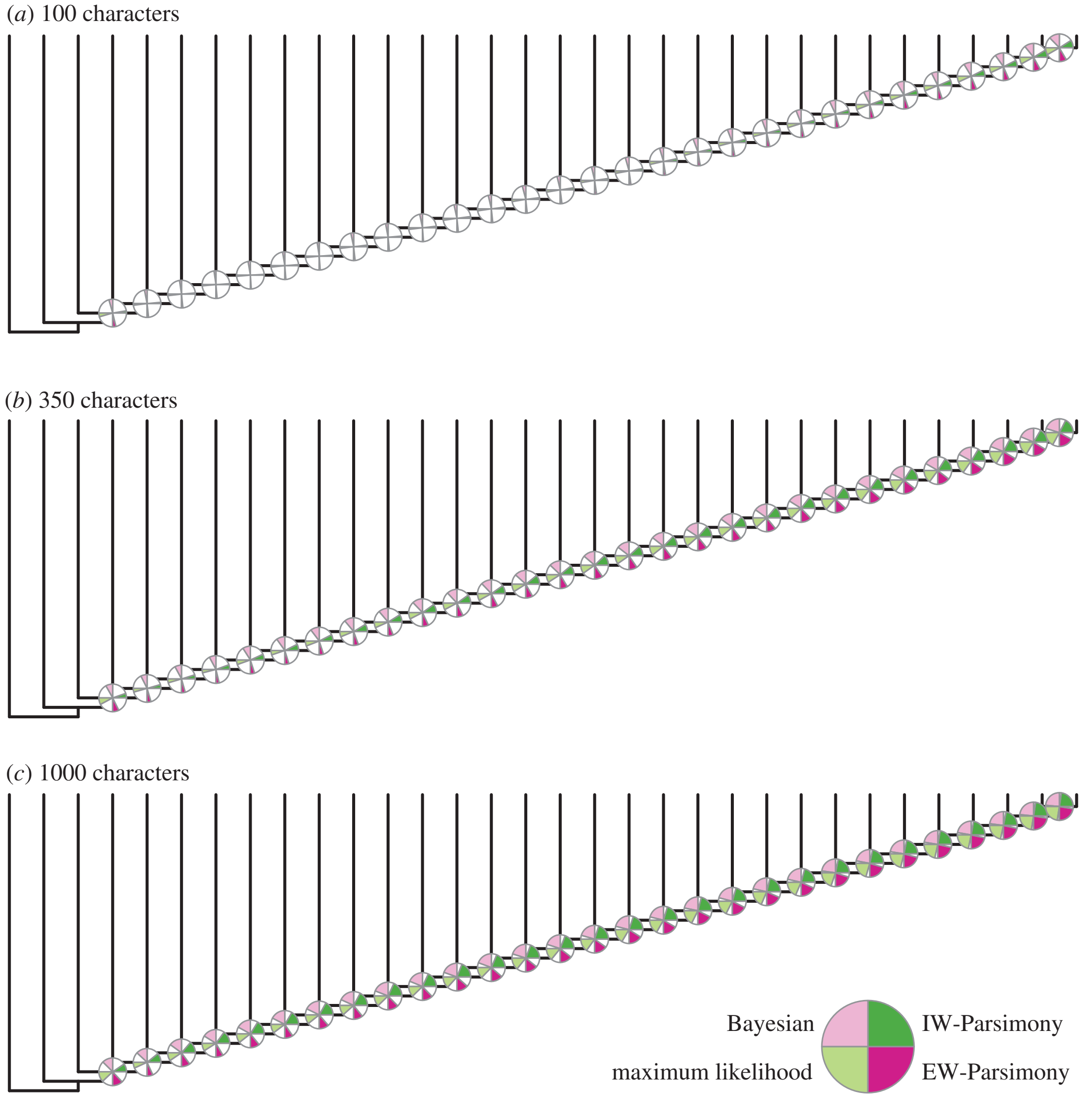


Figure 2. Accuracy of nodes is higher for those closer to the tips in the asymmetrical trees. The percentage of times a node was accurately reconstructed is shown as a proportion of a quarter of a circle in anticlockwise order for Bayesian, maximum likelihood, EW-Parsimony and IW-Parsimony at each node. Accuracy of reconstructions is significantly lower in the 100 character dataset (*a*), and increases in the 350 character (*b*) and 1000 character datasets (*c*). (Online version in colour.)

Bennettitales which resolved as a gnetophyte, and *Caytonia* as sister to the angiosperms.

Analyses of the Luo *et al.* [20] dataset yielded congruent results with the original study, with the placement of *Haramiyavia* outside of crown-Mammalia and multituberculates, although some haramiyids are resolved as crown mammals in the IW-Parsimony analysis (figure 5*a*–*d*).

Nyasasaurus is recovered as a member of Dinosauria in the maximum likelihood, EW-Parsimony and IW-Parsimony analyses of the dataset from Nesbitt *et al.* [21] (figure 5*e*–*h*). The Bayesian analysis recovers *Nyasasaurus* in a polytomy with the two major clades of dinosaurs, corroborating the conclusion of Nesbitt *et al.* [21] that, given the data, its precise phylogenetic position is uncertain.

4. Discussion

(a) Simulations indicate that the Bayesian implementation of the Mk model outperforms all other methods and implementations

Previous simulation-based analyses that have attempted to evaluate the performance of likelihood and parsimony-based phylogenetic methods for analysing phenotypic data have found that the probabilistic model performs best [7,8]. However, these studies were biased against parsimony because they employed an unresolved generating tree that is problematic as parsimony methods will attempt to recover a fully resolved tree from the simulated data yielding a non-zero RF distance from the generating tree, even if the two trees are effectively compatible. Furthermore, since previous simulation studies considered the Mk model only within a Bayesian framework, they did not distinguish between the impact of the probabilistic model of character evolution and the statistical framework in which it was implemented.

Our analyses control for these shortcomings of previous simulation studies and show consistently that the Bayesian implementation of the Mk model performs best. In line with previous simulations [8], we found that EW-Parsimony performs better than IW-Parsimony. There is overlap between model performance shown by the distribution of Robinson–Foulds distances (table 1), but there is reason to have different degrees of confidence in the models; only the Bayesian implementation produces a relatively small distribution of tree performance compared with the large tails signifying worse performance in the two parsimony methods (table 1). We also found that the Bayesian implementation of the Mk model outperforms the

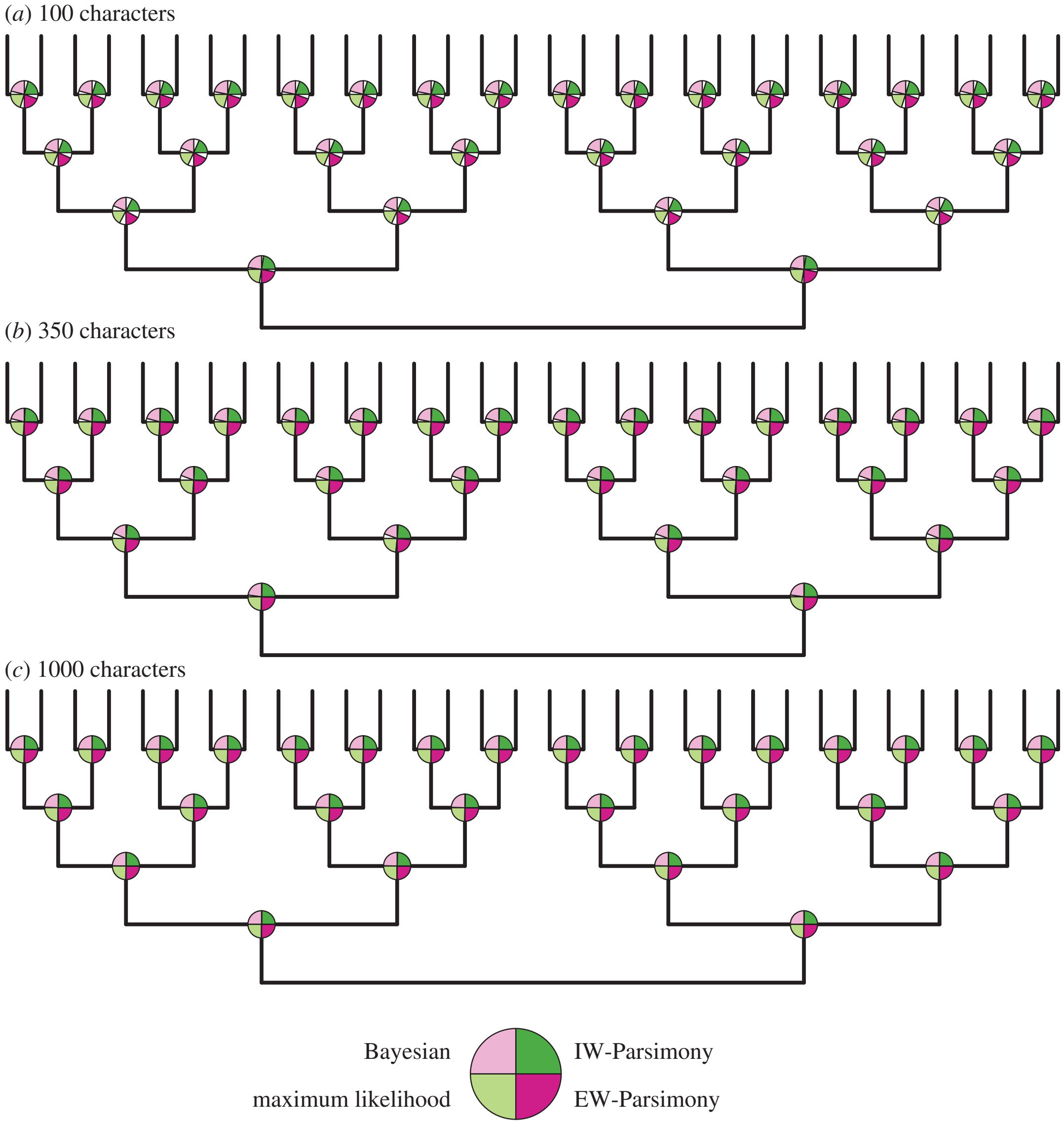


Figure 3. Accuracy of nodes is high for all nodes in the symmetrical phylogeny. The percentage of times a node was accurately reconstructed is shown as a proportion of a quarter of a circle in anticlockwise order for Bayesian, maximum likelihood, EW-Parsimony and IW-Parsimony at each node. Accuracy of reconstructions is high in each dataset size, but there is a non-significant increase in accuracy as dataset size increases (*a–c*). (Online version in colour.)

Table 1. Bayesian approaches produce the most accurate trees for all character sets. Mean and range (in brackets) of Robinson–Foulds distances are lower for topologies estimated using Bayesian methods for both the symmetrical and asymmetrical generating tree. Maximum likelihood is the generally the most inaccurate method for the symmetrical generating tree, and implied weights parsimony performs worst for the asymmetrical generating tree.

	equal weights parsimony	implied weights parsimony	maximum likelihood	Bayesian
asymmetrical generating phylogeny				
100	34.89 (22–56)	37.85 (22–56)	45.84 (20–58)	28.1 (18–39)
350	26.57 (11–51)	29.2 (12–51)	26.49 (6–58)	19.21 (7–35)
1000	17.82 (3–40)	19.16 (2–33)	11.94 (0–58)	9.34 (0–31)
symmetrical generating phylogeny				
100	8.08 (0–33)	9.29 (0–29)	10.1 (0–58)	7.51 (0–29)
350	1.33 (0–28)	1.43 (0–28)	1.8 (0–52)	1.2 (0–28)
1000	0.32 (0–26)	0.31 (0–26)	0.51 (0–52)	0.31 (0–26)

maximum-likelihood implementation, indicating that it is not merely the probabilistic transition model that outperforms parsimony methods, but the implementation of the Mk model within a Bayesian statistical framework. Indeed, the maximum-likelihood implementation of the Mk model was the worst-performing method, worse even than IW-Parsimony. In part, the poor performance of the maximum-likelihood-Mk method is because we did not capture phylogenetic uncertainty

Table 2. p-Values from Spearman's rank correlation between the percentage of nodes being accurately reconstructed and their distance from the root. Nodes closer to the tips are significantly more likely to be accurately reconstructed in asymmetrical trees but this is not generally true for symmetrical phylogenies.

	asymmetrical tree	symmetrical tree
MB 100	<0.001	0.09919
maximum likelihood 100	<0.001	0.027295
EW 100	<0.001	0.106712
IW 100	<0.001	0.092736
MB 350	<0.001	0.638242
maximum likelihood 350	<0.001	0.057809
EW 350	<0.001	0.19683
IW 350	<0.001	0.148108
MB 1000	<0.001	0.256976
maximum likelihood 1000	<0.001	0.085987
EW 1000	<0.001	0.179186
IW 1000	<0.001	0.287058

associated with this phylogenetic method. This is normally achieved in analyses of molecular datasets through bootstrapping methods, but these are inappropriate for the analysis of phenotypic data as the basic methodological assumption, that the phylogenetic signal is randomly distributed across sites (characters), is not true for morphological data.

However, irrespective of the phylogenetic method used, dataset size correlated positively with both phylogenetic accuracy and resolution, diminishing differences in the relative performance of the competing phylogenetic methods. All phylogenetic methods also performed best when attempting to recover a symmetrical target tree; all methods found recovery of asymmetrical trees challenging and phylogenetic accuracy diminished from tip to root. The impact of tree topology is of particular concern since empirical phylogenetic trees are invariably asymmetric [23], and trees of fossil species are infamous for their asymmetry [24,25]. However, there is a broad spectrum of tree symmetry, with fully symmetric and fully asymmetric trees representing end-members. Palaeontological trees with the dimensions used in our simulations are typically far from the fully asymmetric pectinate-generating tree we employed (Ic = ∼0.4 for 32 species) [25]. Furthermore, the asymmetry of many palaeontological trees is often a representational artefact of attempting to summarize character evolution, or an analytic artefact of analysing the relationships among diverse clades based on representative species or higher taxa [26]. Thus, the challenge of recovering trees of extinct taxa may not be as great as a simplistic interpretation of our results might suggest.

(b) Analyses of empirical data bear out conclusions based on simulations

Maximum-likelihood, IW-Parsimony and EW-Parsimony methods of the simulated datasets commonly identify a single optimal tree, but the differences between the optimal trees derived from these methods provides no confidence that any one of the inferred topologies is accurate with reference to the placement of a taxon of interest. This view is corroborated by our reanalysis of empirical datasets which recovered poorly resolved trees using the Bayesian implementation of the Mk model, and in a number of instances, indicate that the conclusions drawn in the corresponding original studies are not supported by the data.

In an extreme example, our re-analyses of the dataset published by Sutton *et al.* [22], which attempted to demonstrate a crown-aplacophoran mollusc affinity for *Kulindroplax*, yielded disparate hypotheses of affinity. EW-Parsimony and IW-Parsimony recovered the published result, while maximum likelihood recovered *Kulindroplax* as a stem-aplacophoran, and Bayesian could not discriminate *Kulindroplax* as a total-group mollusc (figure 4*a*). This poor resolution is unlikely to be a result of poor fossil evidence but, rather, the lack of discriminatory power in the small character matrix. Among the analyses of the dataset from Hilton & Bateman [19], we recovered some of the principal competing topologies that have featured in debate over the affinity of seed plants in past decades. However, the Bayesian analysis of the dataset recovered a topology that is largely unresolved in terms of the relationships among key clades. This suggests that the available data are insufficient to discriminate among the competing hypotheses, and this long-standing debate is largely an artefact of the false resolution of parsimony methods.

Bayesian analyses need not overturn the results from previous analyses based on deterministic phylogenetic methods like EW-Parsimony, IW-Parsimony and maximum likelihood. A phylogenetic position for haramiyids, outside crown-Mammalia, is corroborated by our Bayesian analysis of the dataset from Luo *et al.* [20]—in contrast with the crown-Mammalia affinity recovered for some haramiyids through IW-Parsimony analysis of the same data (figure 5*d*). Similarly, *Nyasasaurus* was posited as the earliest dinosaur, and this conclusion is supported by the Bayesian analyses (figure 5*e*) although this is not supported by EW-Parsimony, IW-Parsimony and maximum-likelihood analyses (figure 5*f*–*h*). However, the Bayesian analysis is more robust in expressing the phylogenetic ambiguity identified by the original authors [19], as *Nyasasaurus* falls in a polytomy alongside the two major clades of dinosaurs.

Some of the differences between methods may simply reflect the dimensions of the dataset. The two datasets that cannot resolve relationships under Bayesian inference and exhibit significant topological discordance among phylogenetic methods [19,22] are both comparatively small (34 taxa, 48 characters and 48 taxa, 82 characters). These both fall within the scope of simulated datasets that yield low resolution from the Bayesian method and, from other phylogenetic methods, high resolution but low accuracy (figure 1). The two empirical datasets that yield trees with greater congruence from the different phylogenetic methods, are both larger: Luo (114 taxa, 497 characters) and Nesbitt (82 taxa, 413 characters). The size of these matrices is comparable with our simulation results in which we see marked increases in topological accuracy and agreement between methods (figure 1, between 350 and 1000 characters).

(c) Implications for phylogenetic analysis of phenotypic data

The results of our simulation studies indicate that the cadre of phylogenetic hypotheses generated from phenotypic data

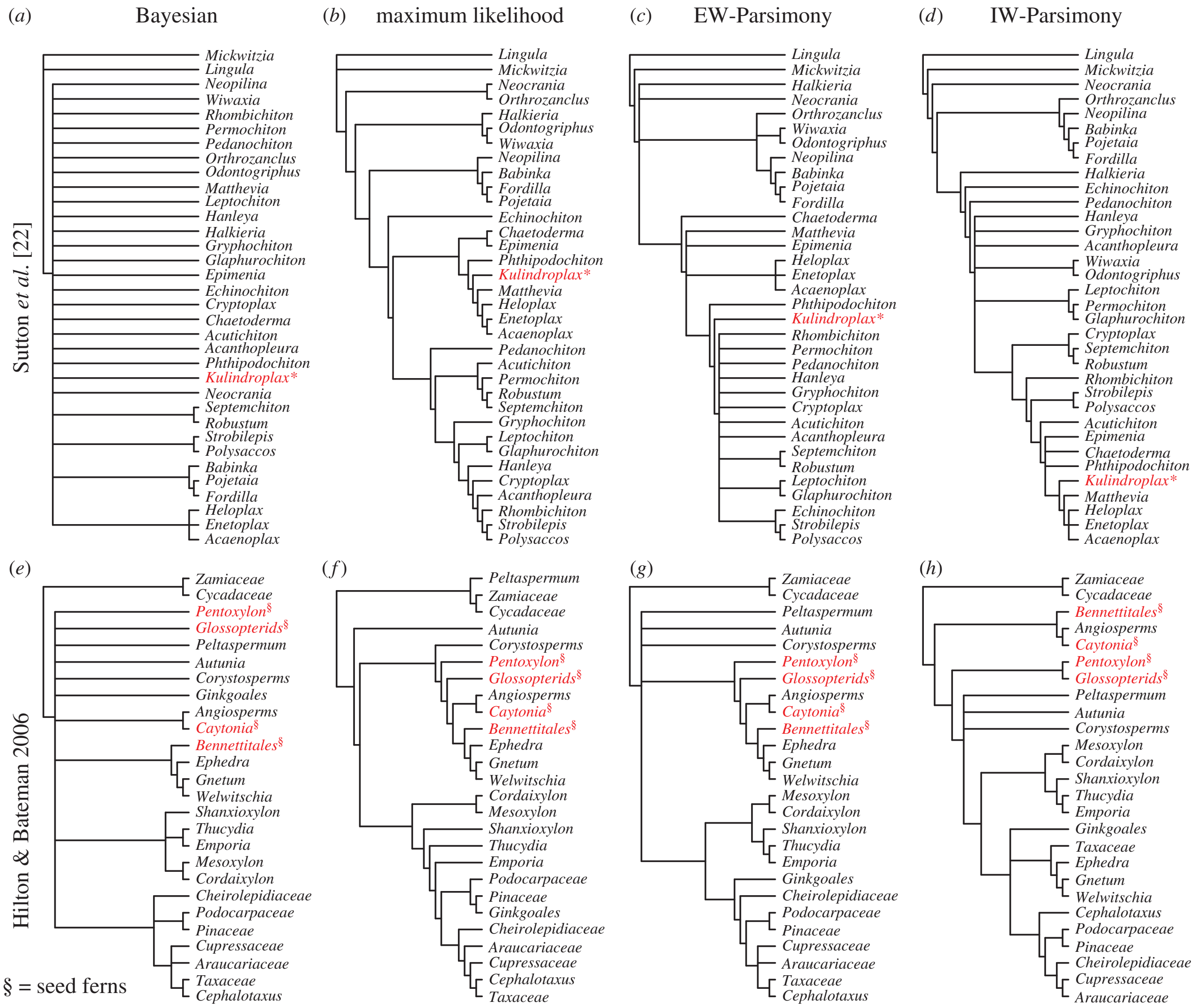


Figure 4. Alternative phylogenetic reconstruction methods alter our understanding of evolution with empirical matrices. However, the relationship of fossil seed ferns from Hilton & Bateman [19] is changed according to implementation (*a–d*), although *Caytonia* remains as sister to angiosperms in all analyses. Alternative analyses change the taxonomic affinity of *Kulindroplax* from Sutton *et al*. [22] (*e–h*). (Online version in colour.)

using parsimony methods require reassessment using the Bayesian implementation of the Mk model. It is likely that many evolutionary interpretations are contingent on precise but inaccurate phylogenetic hypotheses. In this undertaking, it is important that the implications of our simulation studies are considered in the design of phylogenetic studies.

Firstly, phylogenies of fossils tend towards strong asymmetries [25] and, like all phylogenetic methods, Bayesian inference struggles with the recovery of deep nodes within asymmetric trees. Therefore, it is important that outgroups are sampled extensively, ensuring that contentious in-group relationships are closer to the tips, where topological accuracy is highest. Further, in-group lineages should be sampled in a manner that does not accentuate tree asymmetry.

Secondly, phylogenetic accuracy and resolution correlates positively with the relative dimensions of the dataset. Accordingly, phylogenetic resolution or certainty should not be expected from cladistic analyses of small morphological datasets (i.e. those around 100 characters or fewer), particularly if they include fossils. There are finite limits to the number of available phylogenetically informative characters [27] and, for well-studied clades, it may be perceived that these phylogenetically informative characters have already been found. However, it is important to note that the concept of phylogenetic informativeness is different within a likelihood versus a parsimony framework: in parsimony characters that undergo few changes are prized in favour of homoplastic characters. Under the likelihood model, branch length, informed by the number of character changes, contributes to topology estimation. Thus, traditionally 'bad' phylogenetic characters (those exhibiting homoplasy) may find utility in expanding the dimensions of phenotypic character matrices as long as homoplasy falls within the limits that the model can accommodate. In a Bayesian framework, this can be tested using posterior predictive tests of model adequacy (e.g. [28]).

Finally, we may need to alter our expectations to anticipate less well-resolved but more accurate phylogenetic hypotheses, which will both constrain and guide research. Greater resolution may be found by generating matrices suited to likelihood- rather than parsimony-based phylogenetic methods. However, we must also come to terms with the prospect that for some groups of organisms, or their fossil remains, there may be insufficient data. As such, their evolutionary relationships might not therefore be resolvable using morphological data alone and, if they are fossils, their evolutionary significance may never be realized. Nevertheless, resolving phylogenies is not the end game for evolutionary biology.

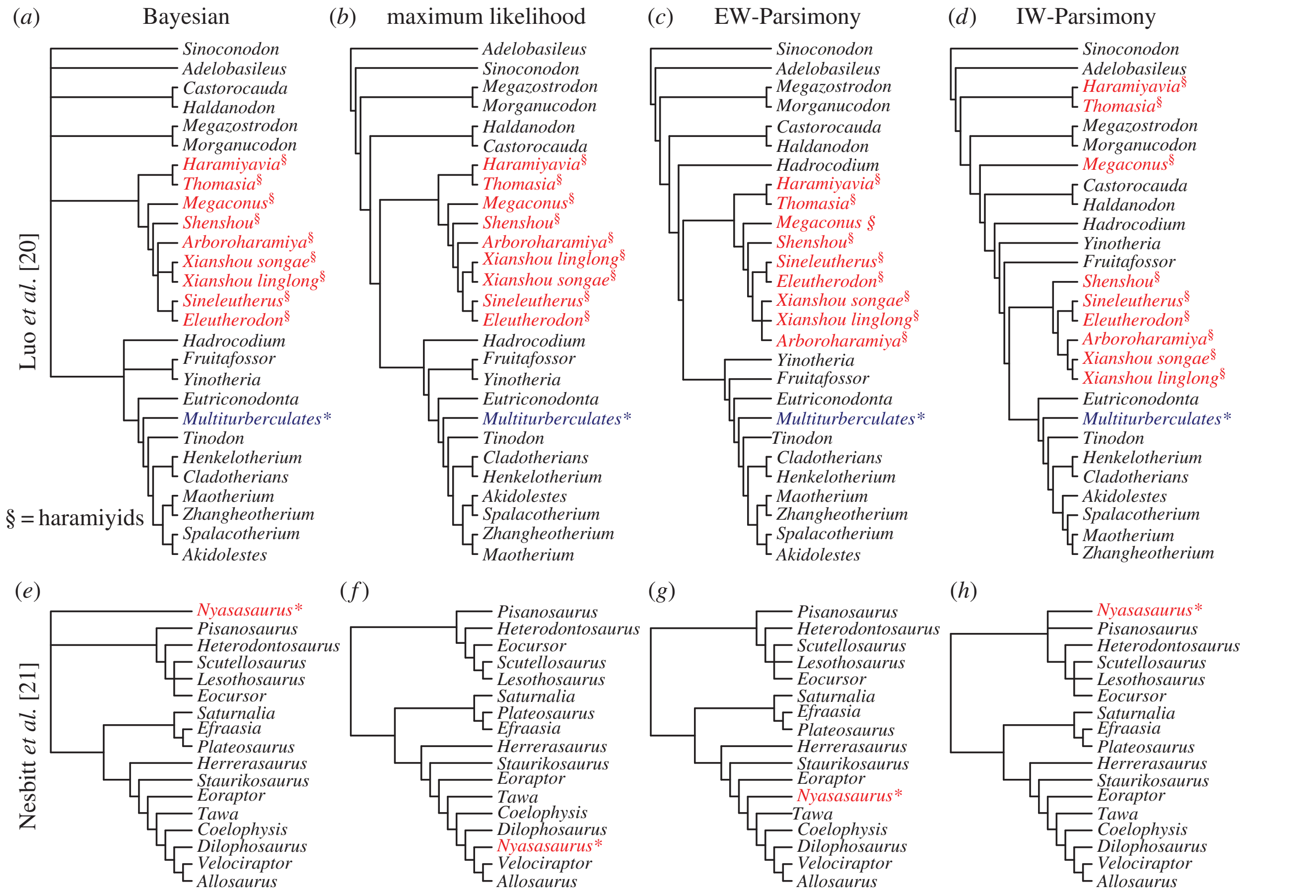


Figure 5. Alternative phylogenetic reconstruction methods produce generally congruent reconstructions of evolution with empirical matrices. For Luo *et al*. [20], the relationship between the haramiyids and multituberculates is largely unchanged across analyses (*a*–*d*). IW-Parsimony (*g*) and Bayesian analyses place *Nyasasaurus* as close to the earliest dinosaur (*e*) and IW-Parsimony places it close to the earliest diverging taxa (*g*), but EW-Parsimony and maximum likelihood place the taxa as a derived member of Dinosauria (*f,h*). (Online version in colour.)

Incompletely resolved trees can still be used as a basis for investigating interesting macroevolutionary questions, and methods exist for incorporating tree uncertainty in phylogenetic comparative methods (e.g. [29]).

5. Conclusion

A growing consensus shows that the Bayesian Mk model is the most accurate method of phylogenetic reconstruction, and here we show that this remains true across dramatically different tree shapes, when analysing datasets composed of both multistate and binary characters, and when compared with maximum-likelihood estimation using the Mk model. We recommend that Bayesian implementations of the Mk model should become the default method for phylogenetic analyses of cladistic morphological datasets, and we should expect low levels of resolution with small datasets. As parsimony methods appear to be less effective than probabilistic approaches, it may be necessary to alter data collection practices by moving away from choosing a selection of characters that undergo few changes, and moving towards scoring all possible characters from the available taxa irrespective of their expected homoplasy.

Data accessibility. Supplementary figures and the code used to simulate the data used in this publication can be accessed in the electronic supplementary material.

Authors' contributions. All authors contributed to the design of the study; M.N.P. and J.E.O.R. led the analyses; interpretation of results and writing was led by M.N.P., J.E.O.R., P.C.J.D. and D.P., though all authors contributed to the interpretation of results and the writing of the manuscript.

Competing interests. We have no competing interests.

Funding. This research was funded by NERC (NE/L501554/1 to J.E.O.R. and L.A.P.; NE/K500823/1 to M.N.P.; NE/L002434/1 to J.F.; NE/N003438/1 to P.C.J.D.), BBSRC (BB/N000919/1 to P.C.J.D.), the University of Bristol (STaR scholarship to A.R.T.), Royal Society Wolfson Research Merit Award (P.C.J.D.) and the John Templeton Foundation (43915 to D.P. and L.H.).

Acknowledgements. We thank the other members of the Bristol Palaeobiology research group for discussion; Rob Asher (Cambridge) and Thomas Guillerme for comments on the draft manuscript. We also thank April Wright and an anonymous reviewer for their help in improving the manuscript.

References

1. Harvey P, May R, Nee S. 1994 Phylogenies without fossils. *Evolution*. **48**, 523–529. (doi:10.2307/2410466)

2. Rabosky DL. 2010 Extinction rates should not be estimated from molecular phylogenies. *Evolution* **64**, 1816–1824. (doi:10.1111/j.1558-5646.2009.00926.x)

3. Losos JB *et al.* 2013 Evolutionary biology for the 21st century. *PLoS Biol*. **11**, e1001466. (doi:10.1371/journal.pbio.1001466)
4. dos Reis M, Donoghue PCJ, Yang Z. 2016 Bayesian molecular clock dating of species divergences in the genomics era. *Nat. Rev. Genet*. **17**, 1–10. (doi:10.1038/nrg.2015.8)
5. Goloboff PA, Carpenter JM, Arias JS, Miranda-Esquivel DR. 2008 Weighting against homoplasy improves phylogenetic analysis of morphological data sets. *Cladistics* **24**, 758–773. (doi:10.1111/j.1096-0031.2008.00209.x)
6. Lewis PO. 2001 A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol*. **50**, 913–925. (doi:10.1080/106351501753462876)
7. Wright AM, Hillis DM. 2014 Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. *PLoS ONE* **9**, e109210. (doi:10.1371/journal.pone.0109210)
8. O'Reilly JE, Puttick MN, Parry L, Tanner AR, Tarver JE, Fleming J, Pisani D, Donoghue PCJ. 2016 Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. *Biol. Lett*. **12**, 20160081. (doi:10.1098/rsbl.2016.0081)
9. Congreve CR, Lamsdell JC. 2016 Implied weighting and its utility in palaeontological datasets: a study using modelled phylogenetic matrices. *Palaeontology* **59**, 447–462. (doi:10.1111/pala.12236)
10. Holton TA, Wilkinson M, Pisani D. 2014 The shape of modern tree reconstruction methods. *Syst. Biol*. **63**, 436–441. (doi:10.1093/sysbio/syt103)
11. Guillerme T, Cooper N. 2016 Effects of missing data on topological inference using a total evidence approach. *Mol. Phylogenet. Evol*. **94**, 146–158. (doi:10.1016/j.ympev.2015.08.023)
12. Sanderson MJ, Donoghue M. 1996 *The relationship between homoplasy and the confidence in a phylogenetic tree*. San Diego, CA: Academic Press.
13. Goloboff PA, Farris S, Nixon K. 2008 TNT, a free programm for phylogenetic analysis. *Cladistics* **24**, 774–786. (doi:10.1111/j.1096-0031.2008.00217.x)
14. Stamatakis A. 2014 RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313. (doi:10.1093/bioinformatics/btu033)
15. Ronquist F *et al.* 2012 Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol*. **61**, 539–542. (doi:10.1093/sysbio/sys029)
16. Wright AM, Lyons KM, Brandley MC, Hillis DM. 2015 Which came first: the lizard or the egg? Robustness in phylogenetic reconstruction of ancestral states. *J. Exp. Zool. B. Mol. Dev. Evol*. **324**, 504–516. (doi:10.1002/jez.b.22642)
17. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015 IQ-TREE: a fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. *Mol. Biol. Evol*. **32**, 268–274. (doi:10.1093/molbev/msu300)
18. Robinson DF, Foulds LR. 1981 Comparison of phylogenetic trees. *Math. Biosci*. **53**, 131–147. (doi:10.1016/0025-5564(81)90043-2)
19. Hilton J, Bateman RM. 2006 Pteridosperms are the backbone of seed-plant phylogeny. *J. Torrey Bot. Soc*. **133**, 119–168. (doi:10.3159/1095-5674)
20. Luo ZX, Gatesy SM, Jenkins FA, Amaral WW, Shubin NH. 2015 Mandibular and dental characteristics of Late Triassic mammaliaform *Haramiyavia* and their ramifications for basal mammal evolution. *Proc. Natl Acad. Sci. USA* **112**, E7101–E7109. (doi:10.1073/pnas.1519387112)
21. Nesbitt SJ, Barrett PM, Werning S, Sidor CA, Charig AJ. 2013 The oldest dinosaur? A Middle Triassic dinosauriform from Tanzania. *Biol. Lett*. **9**, 20120949. (doi:10.1098/rsbl.2012.0949)
22. Sutton MD, Briggs DEG, Siveter DJ, Siveter DJ, Sigwart JD. 2012 A Silurian armoured aplacophoran and implications for molluscan phylogeny. *Nature* **490**, 94–97. (doi:10.1038/nature11328)
23. Mooers AO, Heard SB. 1997 Inferring evolutionary process from phylogenetics tree shape. *Q. Rev. Biol*. **72**, 31–54. (doi:10.1086/419657)
24. Shao KT, Sokal RR. 1990 Tree balance. *Syst. Zool*. **39**, 266–276. (doi:10.1007/s13398-014-0173-7.2)
25. Harcourt-Brown K, Pearson P, Wilkinson M. 2001 The imbalance of paleontological trees. *Paleobiology* **27**, 188–204. (doi:10.1666/0094-8373(2001)027<0188:TIOPT>2.0.CO;2)
26. Panchen A. 1982 The use of parsimony in testing phylogenetic hypotheses. *Zool. J. Linn. Soc*. **74**, 305–328. (doi:10.1111/j.1096-3642.1982.tb01154.x)
27. Scotland RW, Olmstead RG, Bennett JR. 2003 Phylogeny reconstruction: the role of morphology. *Syst. Biol*. **52**, 539–548. (doi:10.1080/10635150390223613)
28. Tarver JE *et al.* 2016 The interrelationships of placental mammals and the limits of phylogenetic inference. *Genome Biol. Evol*. **8**, 330–334. (doi:10.1093/gbe/evv261)
29. Healy K *et al.* 2014 Ecology and mode-of-life explain lifespan variation in birds and mammals. *Proc R. Soc. B* **281**, 20140298. (doi:10.1098/rspb.2014.0298)

PHILOSOPHICAL TRANSACTIONS B

rstb.royalsocietypublishing.org

Review

Cite this article: Lozano-Fernandez J *et al.* 2016 A molecular palaeobiological exploration of arthropod terrestrialization. *Phil. Trans. R. Soc. B* **371**: 20150133. http://dx.doi.org/10.1098/rstb.2015.0133

Accepted: 29 February 2016

One contribution of 15 to a discussion meeting issue 'Dating species divergences using rocks and clocks'.

Subject Areas:
molecular biology, palaeontology, evolution, taxonomy and systematics

Keywords:
terrestrialization, molecular palaeobiology, arthropod evolution, molecular clock, phylogeny

Author for correspondence:
Davide Pisani
e-mail: davide.pisani@bristol.ac.uk

Electronic supplementary material is available at http://dx.doi.org/10.1098/rstb.2015.0133 or via http://rstb.royalsocietypublishing.org.

THE ROYAL SOCIETY
PUBLISHING

A molecular palaeobiological exploration of arthropod terrestrialization

Jesus Lozano-Fernandez[1,2], Robert Carton[3], Alastair R. Tanner[2], Mark N. Puttick[1], Mark Blaxter[4], Jakob Vinther[1,2], Jørgen Olesen[5], Gonzalo Giribet[6], Gregory D. Edgecombe[7] and Davide Pisani[1,2]

[1]School of Earth Sciences, and [2]School of Biological Sciences, University of Bristol, Life Sciences Building, 24 Tyndall Avenue, Bristol BS8 1TQ, UK
[3]Department of Biology, The National University of Ireland Maynooth, Maynooth, Kildare, Ireland
[4]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3TF, UK
[5]Natural History Museum of Denmark, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen, Denmark
[6]Museum of Comparative Zoology, Department of Organismic and Evolutionary Biology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, USA
[7]Department of Earth Sciences, The Natural History Museum, Cromwell Road, London SW7 5BD, UK

JL-F, 0000-0003-3597-1221; ART, 0000-0001-8045-2856; MNP, 0000-0002-1011-3442; MB, 0000-0003-2861-949X; JV, 0000-0002-3584-9616; JO, 0000-0001-9582-7083; GG, 0000-0002-5467-8429; DP, 0000-0003-0949-6682

Understanding animal terrestrialization, the process through which animals colonized the land, is crucial to clarify extant biodiversity and biological adaptation. Arthropoda (insects, spiders, centipedes and their allies) represent the largest majority of terrestrial biodiversity. Here we implemented a molecular palaeobiological approach, merging molecular and fossil evidence, to elucidate the deepest history of the terrestrial arthropods. We focused on the three independent, Palaeozoic arthropod terrestrialization events (those of Myriapoda, Hexapoda and Arachnida) and showed that a marine route to the colonization of land is the most likely scenario. Molecular clock analyses confirmed an origin for the three terrestrial lineages bracketed between the Cambrian and the Silurian. While molecular divergence times for Arachnida are consistent with the fossil record, Myriapoda are inferred to have colonized land earlier, substantially predating trace or body fossil evidence. An estimated origin of myriapods by the Early Cambrian precedes the appearance of embryophytes and perhaps even terrestrial fungi, raising the possibility that terrestrialization had independent origins in crown-group myriapod lineages, consistent with morphological arguments for convergence in tracheal systems.

This article is part of the themed issue 'Dating species divergences using rocks and clocks'.

1. The long road to terrestrial life

Animals and life more broadly have marine origins, and the colonization of land started early in life's history. Possible evidence for subaerial prokaryotic life dates back to the Archaean [1,2], and terrestrial communities (either freshwater or subaerial) with a eukaryotic component are known from the Torridonian of Scotland approximately 1.2–1.0 billion years ago (Gya) [3]. These deposits include multicellular structures, cysts and thalli that can have a diameter of almost 1 mm [3]. While there is no evidence for land plants, animals and fungi, these deposits indicate that at approximately 1 Ga relatively complex terrestrial ecosystems already existed [4]. Definitive evidence for the existence of land plants is much more recent. The oldest embryophyte body fossils are from the Late Silurian [5]. The oldest spores of indisputable embryophyte origin (trilete spores) extend the history of plants only a little deeper, into the Ordovician (449 million years ago—Ma) [4,5], and the oldest embryophyte-like spores (which do not necessarily indicate the existence of embryophytes) barely reach the Late Cambrian [4]. Similarly, the fossil record of the terrestrial Fungi does not extend beyond the Ordovician, with the oldest known fungal fossils dating to approximately 460 Ma [6]. However, terrestrial rock sequences from the Cambrian and the Ediacaran are rare, and the late appearance of land plants and Fungi in the fossil record might represent preservational artefacts of the rock record [4].

Only few animal phyla include lineages that can complete every phase of their life cycle outside of water-saturated environments (from moisture films to the oceans) and are thus fully terrestrial. The most diverse and biologically important of the phyla with lineages that attained full terrestriality are the Vertebrata (with the reptiles, birds and mammals, i.e. Amniota); the Mollusca (with the land snails and the slugs); and the Arthropoda (e.g. insects, spiders, scorpions, centipedes) [7]. While the terrestrial vertebrates colonized the land only once even if some members (such as the cetaceans) secondarily reverted to life in water, molluscs and arthropods colonized the land multiple times independently and at different times in Earth history, constituting better model systems to study terrestrial adaptations at the genomic, physiological and morphological levels. In Arthropoda, there have been a minimum of three ancient (Palaeozoic) terrestrialization events: that of the Hexapoda, that of the Myriapoda and that of the Arachnida [8]. In addition, there have been multiple, more recent, land colonization events within malacostracans. These events correspond to the origin of terrestrial isopods (i.e. the woodlice) and amphipods (e.g. the landhoppers), and of a variety of semi-terrestrial species such as the coconut crab (*Birgus latro*), a decapod that lives its adult life on land but still retains marine larvae (see also [9]).

Previous studies [7,10–13] discussed at length the problems faced by animals crossing the water-to-land barrier, with [11] addressing them specifically in the case of the Arthropoda. These problems mostly relate to the different physical properties of air and water, and affect reproduction, sensory reception, locomotion, gas exchange, osmoregulation and protection from an increased exposure to ultraviolet radiation. A classic example of adaptation to terrestriality at the genomic level is observed, in both vertebrates and arthropods, when comparing the olfactory receptors of marine and terrestrial forms. Terrestrialization is associated with massive, independent, parallel changes in the olfactory receptor gene repertoires of both lineages probably because water-soluble and airborne odorants differ and cannot be efficiently bound by the same receptors [14–16].

Multiple independent terrestrialization events within the same lineage permit rigorous comparison of alternative solutions adopted by different (but genomically and morpho-physiologically comparable) groups to the same adaptive challenge, and represent a powerful tool for understanding evolution in a comparative framework [17]. To carry out meaningful comparative studies of animal terrestrialization, however, it is necessary to (i) clarify how many independent terrestrialization events happened in the lineage under scrutiny, (ii) estimate when these terrestrialization events happened and how long they took, and (iii) robustly identify the aquatic sister group of each terrestrial lineage. This information is, in turn, necessary to enable comparative analyses and to estimate the rate at which terrestrial adaptations emerged.

Here we explore the three deepest (Palaeozoic) arthropod terrestrialization events (those of the Hexapoda, Myriapoda and Arachnida), and summarize and expand current evidence about processes that led to their terrestrialization. We particularly focus on Hexapoda, because hexapod terrestrialization, an event that led to the origin of the majority of terrestrial animal biodiversity [18], is particularly poorly understood.

2. The phylogenetic perspective

Phylogenetic relationships among the major arthropod lineages have long been debated [19]. However, some consensus has emerged. Myriapoda, the first of the three major terrestrial arthropod groups we shall consider, is now generally accepted to represent the sister group of Pancrustacea (Hexapoda plus all the crustacean lineages). The Myriapoda–Pancrustacea clade is generally referred to as Mandibulata [20–23]. Alternative hypotheses of myriapod relationships have been previously proposed. Among these are the Atelocerata or Tracheata hypothesis, which suggested myriapods as the sister of hexapods, and the Myriochelata hypothesis, which saw the myriapods as the sister group of chelicerates. Atelocerata was based on morphological considerations (e.g. both myriapods and hexapods use tracheae to carry out gas exchange) and continues to have a few adherents among morphologists [24]. However, Atelocerata has only been recovered once in analyses combining molecular, morphological and fossil data [25]. The Myriochelata hypothesis was derived entirely from molecular analyses [26–30], and is now generally considered to have been the result of a long-branch attraction artefact caused by the faster-evolving pancrustaceans attracting to the outgroup and pushing Myriapoda and Chelicerata into an artefactual clade [20]. Both Myriochelata and Atelocerata are disfavoured by current available analyses, with strong molecular and morphological support favouring a placement of hexapods within ‘Crustacea’ (the Pancrustacea or Tetraconata concept—e.g. [20,23,26,31–35]), and a placement of Myriapoda as the sister group of Pancrustacea within Mandibulata (see references above and [19] for a recent review). Accordingly, there is now general agreement that the sister group of the terrestrial Myriapoda is the (primitively) marine Pancrustacea.

The sister group relationships of the Arachnida are quite well understood. This group includes all the terrestrial chelicerates and has two extant successively more distant marine sister taxa: Xiphosura (horseshoe crabs) and Pycnogonida (sea spiders) [23,36,37]. In contrast, the exact relationships of the Hexapoda within Pancrustacea are still unclear, and it is not obvious whether their sister taxon was a marine-, brackish- or freshwater-adapted organism.

Early analyses of eight molecular loci combined with morphological data provided some support for Hexapoda as the sister group of a monophyletic Crustacea, barring a long-branch clade [38], with Branchiopoda as the sister group of Remipedia plus Cephalocarida (the latter two taxa constituting Xenocarida *sensu* [23]). Subsequently, a taxonomically well-sampled molecular phylogeny of three protein coding genes [34] found support for Branchiopoda as the sister group of Hexapoda, and Remipedia as the sister group of those two taxa. While mitogenomic data have also been used in an attempt to resolve hexapod relationships, this type of data is notoriously difficult to analyse [39,40] and has frequently recovered misleading results (contrast [41,42]). With reference to the relationships of Pancrustacea, mitogenomic data were found to be unable to resolve hexapod relationships with confidence [43] and we shall not consider them further.

Based on a large dataset of 62 protein coding genes analysed as nucleotide sequences, support for a sister group relationship between Xenocarida (Remipedia + Cephalocarida—see also above) and Hexapoda was found [23,35]. This clade was called Miracrustacea [23]. In the same analysis, Branchiopoda grouped with Malacostraca, Copepoda and Thecostraca in a novel clade named Vericrustacea [23] rather than allying with Hexapoda. However, these findings were shown to be affected by an artefact of serine codon bias [37]. The close association between Remipedia and Hexapoda (to the exclusion of Cephalocarida) was the only high-level pancrustacean relationship proposed by [23] that was confirmed by [37], which reinstated Branchiopoda as a close relative of Hexapoda, finding Remipedia, Hexapoda, Branchiopoda and Copepoda to constitute an unresolved clade that was referred to as 'clade A' in [37]. Other recent studies found similar results, suggesting a Branchiopoda + Hexapoda + Remipedia [21,22,44] (and perhaps Cephalocarida [45]) clade, but with different internal resolutions. In particular, [21,44,45] found Remipedia as the closest relative of Hexapoda (as in [34]), whereas [22] found Branchiopoda as the sister taxon of Hexapoda. Oakley *et al.* [45] was the only one, among the studies mentioned above, that included Cephalocarida, and found Remipedia as the sister group of Hexapoda and Branchiopoda as the sister group of Cephalocarida. Overall, from the perspective of molecular phylogenetics, a strong case can be made that Hexapoda, Branchiopoda and Remipedia belong to the same clade. In addition, evidence exists that Cephalocarida might also be a member of this group of hexapod relatives, which was named Allotriocarida [45]. Yet, to date, molecular phylogenetics has not robustly resolved internal allotriocarid relationships.

A close association between Remipedia and Hexapoda had been suggested based on the presence of a duplication of the haemocyanin gene (haemocyanin being the respiratory pigment used by most arthropods) that is uniquely shared between Remipedia and Hexapoda [46]. This duplication could represent a rare genomic event indicative of a possible sister group relationship between Remipedia and Hexapoda. However, Branchiopoda use haemoglobin as a respiratory pigment rather than haemocyanin. Because haemoglobin is an autapomorphy of Branchiopoda, the presence of two haemocyanin genes in Remipedia and Hexapoda and one in Cephalocarida [46] would conclusively resolve the sister group relationship between these taxa only if the relationships between Cephalocarida and Branchiopoda delineated by [45] were correct. This is because if Cephalocarida (which has only one haemocyanin) is not closely related to Remipedia, Branchiopoda and Hexapoda, then the haemocyanin duplication could have happened in the stem lineage subtending Remipedia, Branchiopoda and Hexapoda, with Branchiopoda having lost both paralogues as it shifted to using haemoglobin as a respiratory pigment. To validate the haemocyanin evidence, it is thus of paramount importance that further studies be carried out to either reject or confirm the results of [45], as bootstrap support values for the monophyly of Allotriocarida and the deepest relationships within this clade were variable and never higher than 85% [45]. Similarities between Remipedia and Hexapoda were also previously suggested based on neurological characters [47,48]. However, more recent studies showed that while neuroanatomical similarities between Hexapoda and Remipedia exist, brain morphology suggests a closer association between Remipedia and Malacostraca [49]. Given that hexapods are generally not found to be close relatives to Malacostraca by other lines of evidence (see above for molecular analyses), similarities in the nervous systems of these three lineages might be subject to evolutionary convergence.

Knowledge of the sister group of each terrestrial arthropod lineage is important not only to increase the power of comparative studies to test adaptive strategies to life on land (see above), but also to understand the route to terrestrialization taken by different lineages. While the sister groups of Myriapoda and Arachnida were undoubtedly marine, most branchiopods inhabit freshwater, and a freshwater route to hexapod terrestrialization was proposed based on this [50]. In contrast, Remipedia is exclusively found in coastal anchialine settings generally with some connection to the sea. Accordingly, a sister group relationship between Remipedia and Hexapoda would better support a direct, marine [10] route to terrestrialization [44].

3. The timescale of arthropod terrestrialization

The oldest arthropod fossils are undoubtedly marine. They include trilobites, the oldest representatives of which date back to the Early Cambrian (*ca* 521 Ma [51]); Trilobita is variably interpreted as either stem mandibulates [20] or as stem chelicerates [52]. Other Cambrian, marine fossils include chelicerates (pycnogonids [53]), and crustaceans; both cuticular fragments from Branchiopoda, and possibly also Ostracoda and Copepoda [54] and complete body fossils such as the allotriocarid (most likely stem branchiopod) *Rehbachiella kinnekullensis* [55].

The oldest subaerial arthropod traces (ichnofossils) are from the Mid- to Late Cambrian–Early Ordovician age. Examples include trackways impressed on eolian dune sands by an amphibious myriapod-like arthropod, perhaps a euthycarinoid [56]. Other Cambrian (Mid-Cambrian to Furongian) locomotory traces have been documented from subaerially exposed tidal flats in Wisconsin and Quebec [57]. A euthycarcinoid tracemaker has been confidently associated with these traces, further cementing the view that arthropod subaerial activities (if not terrestrial arthropods) were common on Cambrian shorelines. The oldest terrestrial myriapod body fossil (which is also the oldest undisputably terrestrial animal) is the *ca* 426 Ma millipede *Pneumodesmus newmani*, from the Silurian of Scotland [58]. The subaerial ecology of *P. newmani* is indisputable, because spiracles (segmental openings that allow air to enter the tracheal system) are present on the lateral part of its sternites. The Siluro-Devonian fossil record of Myriapoda consists only of taxa that can be assigned with confidence to the crown groups of extant classes (Diplopoda and Chilopoda), as well as the apparent diplopod-allied Kampecarida, and to date no well corroborated candidates for stem-group Myriapoda have been identified [59]. Critical reviews of the diagnostic/apomorphic characters of myriapods have outlined a search image for a stem-group myriapod that could potentially be recognized in Early Palaeozoic marine strata [60]. Arachnid fossils are just a little younger than those of the oldest Myriapoda, the earliest unequivocally terrestrial examples (trigonotarbids) being present in Silurian deposits dated at approximately 422 Ma [61]. Early Silurian arachnids are represented by the oldest scorpions, which have long been considered to be aquatic because of their associated biota and sediments, but phylogenomic evidence for Scorpiones being nested within terrestrial clades of Arachnida [36] is more compatible with terrestrial habits [62]. The stem group of Arachnida has an aquatic fossil record as far back as the Late Cambrian, the earliest fossils being resting traces of chasmataspidids [63], resolved as sister group to a eurypterid–arachnid clade [64]. Evidence for complex terrestrial ecosystems with land plants, fungi and a variety of arthropods is known from the Upper Silurian onward [65] and is confirmed in the beautifully preserved, and widely celebrated, Lower Devonian (approx. 411 Ma), Rhynie chert Konservat-Lagerstätte [66]. The latter includes the oldest examples of Hexapoda in the fossil record, including Collembola and Insecta.

Recent molecular clock analyses of the arthropod radiation (or of parts of it) generally corroborate the palaeontological evidence and suggest times of origin for Arachnida that are broadly consistent with the fossil evidence [8,21,67–70]. However, molecular divergence times for the origin of crown-group Hexapoda and Myriapoda substantially predate fossils, and this discrepancy is more pronounced in the case of Myriapoda, for which divergence estimates firmly place the modern representatives of this phylum deep in the Cambrian, despite the oldest known crown myriapod fossil being only 426 Ma [58]. This is problematic, because all crown myriapods are terrestrial, and all use tracheae for gas exchange. If tracheae have a single origin in Myriapoda, then current molecular clock results suggest a Cambrian terrestrialization for this lineage, which is not documented in the fossil record. Ephemeral, terrestrial ecosystems existed since approximately 1 Ga [3], and the fossil record of embryophyte-like spores suggests that some form of vegetation existed on land in the Cambrian [2,4,5]. Such limited terrestrial environments, as well as coastal environments [56,57], could have already been conducive to myriapod life on land in the Cambrian [2].

One recent molecular clock study of the arthropod radiation [71], despite being in agreement with other studies with reference to arthropod terrestrialization, is in disagreement with both the fossil record and other molecular clock studies with reference to the deepest divergences in the arthropod tree. However, this study was based on the gene set of [23], that was shown to be affected by strong codon-usage biases [37]. In the absence of correction, this dataset recovered a large number of otherwise unsupported pancrustacean clades (e.g. Vericrustacea and Miracrustacea, see [71]) and consequent erroneous estimation of branch lengths and divergence times. Indeed, subsequent analysis of the same data that attempted to correct for such biases [37] yielded results generally comparable to those obtained in other molecular clock studies.

4. A freshwater route to life on land?

An interesting question in the study of terrestrialization is whether land was invaded directly from the sea (the marine route [10,44]), or whether animals first colonized freshwater environments and only subsequently moved to the land (the freshwater route [50]). To address this question, we can look at the fossil record of stem terrestrial lineages when available, and to the sister group of these terrestrial lineages. A freshwater route would imply that the last common ancestor of the considered terrestrial taxa and its sister aquatic lineage separated in a freshwater habitat [50], whereas a marine route would imply that they separated either in a marine or brackish (estuarine) environment [44]. Myriapods and arachnids have marine sister groups. In the case of the Hexapoda, a freshwater route was suggested based on presumed sister-group relationships between Branchiopoda and Hexapoda [50]. While the freshwater origins hypothesis is challenged by the proposal that Remipedia are the sister group of Hexapoda [44], this is far from well established (see above), leaving space for the possibility that hexapod ancestors might have first colonized fresh water and only after that the land. Here we investigate whether hexapods took a marine or a freshwater route to the colonization of land.

5. Material and methods

(a) Dataset assembly

We expanded a published dataset [72] to include new arthropod taxa (see electronic supplementary material, table S1) mostly obtained from NCBI. Transcriptomes of the sea spider *Pycnogonus* sp. and of the horseshoe crab *Limulus polyphemus* were obtained as part of this study and sequenced, respectively, at Edinburgh Genomics and at the Geogenomic Center in Copenhagen. We also added other bilaterian taxa to increase the number of calibration points available for molecular clock analyses (electronic supplementary material, table S1 and figures S1–S5). The core dataset included 57 taxa and 246 genes. This dataset was then pruned of all non-panarthropod species, to avoid systematic biases that might have been induced by the presence of distant outgroups, and create a smaller dataset (including 30 species and 246 genes) used for phylogenetic analyses only. We developed a series of PERL scripts (available at github.com/jairly/MoSuMa_ tools) to add species to the existing dataset. BLASTp [73] was used, with an *E*-value cut-off of less than 10^{-20} to identify potential orthologues. The new potential orthologues were aligned with the existing orthologue set using MUSCLE [74], and a maximum-likelihood (ML) tree was generated using PhyML [75] under the LG + G model. Tree distances (branch length distances) were used to distinguish orthologues from paralogues using a few simple rules. (1) If only one putative orthologue existed and its average tree distance from all previously identified orthologues in the dataset was within 3 standard deviations of the average of the tree distances calculated across all previously identified orthologues, then the putative orthologue was retained. (2) If there was only one putative orthologue and its distance to other previously identified orthologues exceeded 3 standard deviations from the average of the tree distances calculated across all previously identified orthologues, then the tree and the alignment were visually inspected. (2a) If the sequence was misaligned, then the alignment was corrected and the procedure repeated. (2b) If the sequence was correctly aligned and the sequence clustered in a phylogenetically unexpected position (e.g. a new *Daphnia* sequence that clustered with a human sequence), then the sequence was deemed a possible paralog and not retained. Note that here ‘phylogenetically unexpected’ simply means obviously incorrect. A myriapod sequence clustering with a chelicerate, for example, was considered to cluster in an expected position, in contrast to *Daphnia* clustering with a human. (2c) If the sequence was correctly aligned and the sequence clustered in a phylogenetically plausible position (e.g. a new *Drosophila* sequence that clustered within insects) the sequence was retained but flagged to allow for directed exclusion (if necessary) in subsequent analyses. (3) If more than one putative orthologue was present in the dataset, then the tree was first visually inspected to evaluate whether all putative orthologues formed a monophyletic group (i.e. to make sure they constituted a set of in-paralogs). (3a) If they did and their average tree distance from other sequences was less than 3 standard deviations from the average distance across all previously identified orthologues, then the putative orthologue of minimal branch length was retained. (3b) If the putative orthologues did not cluster together and all but one had significant distance (in excess of 3 standard deviations) from the average distance across all previously identified orthologues, the putative orthologue of acceptable distance was retained if it also clustered in a phylogenetically plausible position. (3c) If all putative orthologues had excessively long branches (more than 3 standard deviations from the average), then they were all rejected. Each set of orthologues was realigned using MUSCLE [74] and trimmed using Gblocks [76] to exclude ambiguously aligned sections. Gblocks settings were: minimum number of sequences for a conserved position = 50% of the sequences in the protein family; minimum number of sequences for a flank position = 75% of the sequences in the protein family; minimum length of a block = 5; allowed gap positions = half. The final dataset of curated sequences was concatenated using FAScONCAT v. 1.0 [77]. It included 58 taxa across all Protostomia and Deuterostomia and 40 657 amino acid positions. Taxa were deleted from this dataset to generate the taxonomically reduced alignment used for phylogenetic reconstruction (see above). The latter included 30 panarthropod species and 40 657 amino acid positions.

(b) Phylogenetic reconstruction

Phylogenetic trees were inferred using PHYLOBAYES MPI v. 1.5 [78] under the site-heterogeneous CAT – GTR + G model of amino acid substitution [79]. Convergence was assessed by running two independent Markov chains and using the bpcomp and tracecomp tools from PHYLOBAYES to monitor the maximum discrepancy in clade support (maxdiff), the effective sample size (effsize) and the relative difference in posterior mean estimates (rel_diff) for several key parameters and summary statistics of the model. The appropriate number of samples to discard as 'burn in' was determined first by visual inspection of parameter trace plots, and then by optimizing convergence criteria.

(c) Molecular clock analyses

Divergence time estimation was performed using PHYLOBAYES 3.3f (serial version) [80] on a fixed topology (see electronic supplementary material, figures S1–S5). We used two alternative relaxed molecular clock models: the autocorrelated CIR model [81] and the uncorrelated gamma multipliers model (UGAMMA) [82], as in [83]. The tree was rooted on the Deuterostomia–Protostomia split. A set of 24 calibrations (see electronic supplementary material, table S2) was used, with a root prior defined using a Gamma distribution of mean 636 Ma and standard deviation of 30 Ma. However, previously we had also tested the effect of a much more relaxed root prior that used an exponential distribution of average 636 Ma (see electronic supplementary material, table S2 for justifications). The substitution model used to estimate branch lengths was the CAT – GTR + G model, as in the phylogenetic analysis. All analyses were conducted using soft bounds with 5% of the probability mass outside the calibration interval. A birth–death model was used to define prior node ages. Analyses were run under the priors to evaluate the effective joint priors induced by our choice of priors. Convergence was tested running the tracecomp tool as specified above.

(d) Ancestral environment reconstructions

Maximum-likelihood-based ancestral character state reconstruction was carried in R (www.R-project.org [84]) using maximum-likelihood estimation under the Mk model [85,86] to infer whether the last common ancestor of Branchiopoda was a freshwater-, marine- or brackish-adapted animal. The branchiopod phylogeny of [87] was modified to include key fossils from [88]: *Rehbachiella*, *Lepidocaris*, *Castracollis* and *Almatium*. *Rehbachiella kinnekullensis* (from the Upper Cambrian) is particularly important as it was initially described as a marine stem-group anostracan [55], and subsequently reassigned to a stem-group branchiopod [89]. This systematic placement has not been universally accepted, with some analyses instead allying *Rehbachiella* closer to cephalocarids than to branchiopods [45,90]. Whereas *Rehbachiella* is found in association with marine taxa [55], and the geological context of the bituminous limestones in which the fossils are preserved indicates dysoxic marine sediments, most extant branchiopods are found in fresh water or in continental brackish waters (vernal pools, saline lakes, etc.). *Lepidocaris rhyniensis* [91] and *Castracollis wilsonae* [92] are freshwater branchiopod fossils from the Early Devonian Rhynie chert. Kazacharthra (represented herein by *Almatium gusevi* [93]), are Triassic–Jurassic relatives of Notostraca limited to non-marine (lacustrine) deposits from Kazakhstan, Mongolia and China. A matrix representing ecological preferences for all considered taxa was assembled from the literature (see electronic supplementary material, table S3). The time-calibrated tree was generated by adding the fossils from [88] to the tree in [87] using 10 calibrations from [94] and setting tip taxa to their occurrence times. The time-calibrated topology was generated using the R package paleotree [95]. We calculated marginal likelihood under Mk for internal nodes in this time-calibrated tree and present the scaled marginal likelihoods of the three possible root states for total-group Branchiopoda.

6. Results

(a) Phylogeny

Our phylogenetic analyses are presented in figure 1. They clearly support monophyly of Arthropoda and of the three main arthropod lineages (Chelicerata, Myriapoda and Pancrustacea). While a few studies have suggested that Tardigrada, rather than Onychophora, might be the closest sister group of Arthropoda [96], evidence for this phylogenetic arrangement is limited to only a few morphological characters. Our choice of Tardigrada as outgroup is thus guided by results of previous phylogenomic studies [72,97,98]. The relationships among the arthropod lineages are resolved according to current convention and depict a Mandibulata clade (PP = 1) as the sister group of Chelicerata (PP = 1). Within Chelicerata, the sea spiders are recovered as the sister group of the other chelicerates, Euchelicerata (PP = 1), with xiphosurans as sister group to arachnids. Myriapods are likewise well resolved, dividing into Chilopoda and Diplopoda, and each group follows the currently well-accepted relationships [69,99]. Within Pancrustacea, we recovered an arrangement of taxa that is consistent with the monophyly of Allotriocarida. Of particular relevance to terrestrialization is the partial allotriocarid clade, including Branchiopoda, Remipedia and Hexapoda. Within this clade, we found Branchiopoda to be the sister group of Hexapoda (PP = 1), in agreement with [22,37] but contrasting with other studies (as summarized above [21,44,45]).

(b) Molecular divergence times

Molecular divergence times among arthropod major clades are presented in figure 2 and table 1 and in electronic

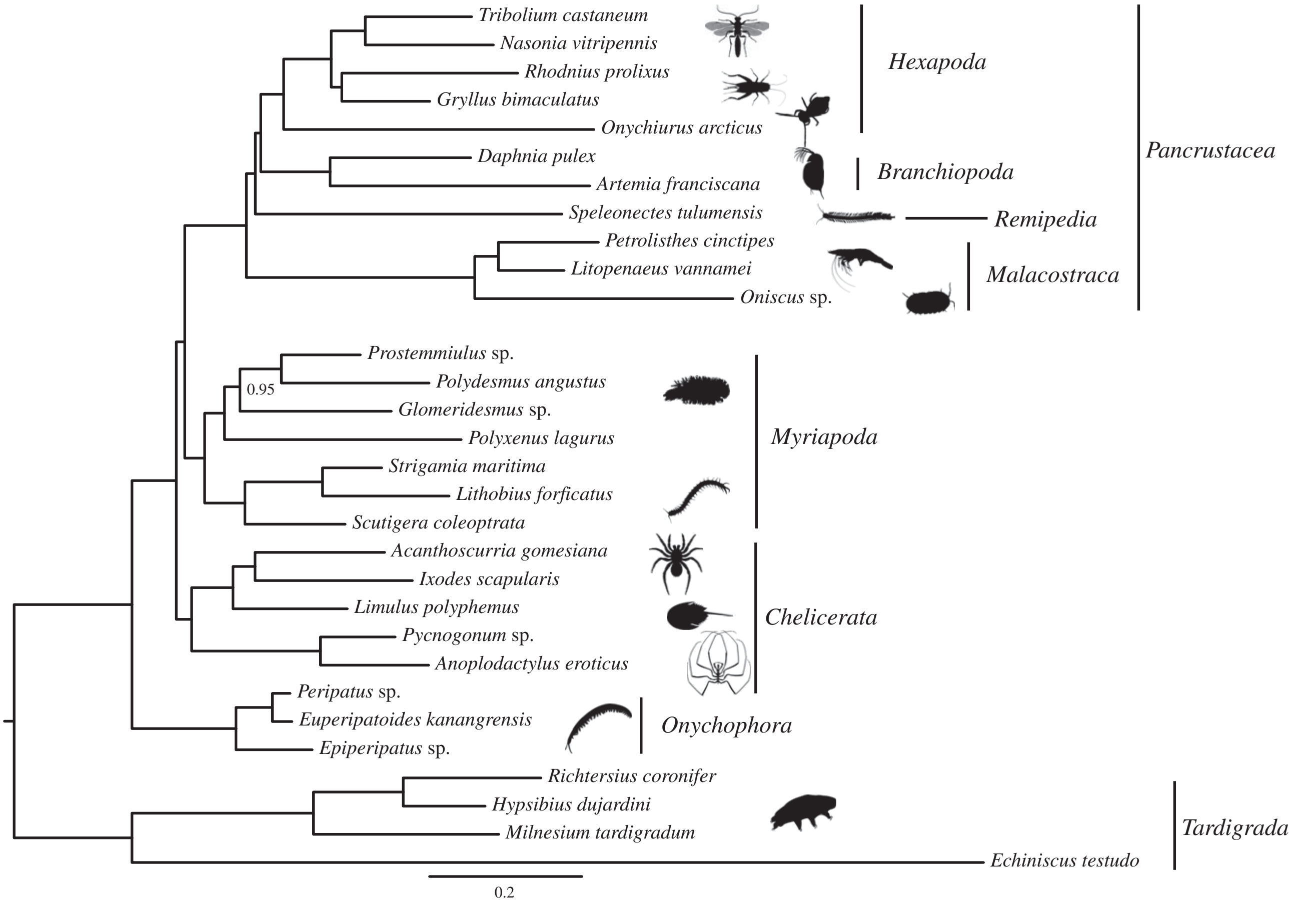


Figure 1. Bayesian phylogeny of Panarthropoda. This tree was obtained under the CAT − GTR + G model. All nodes but one had a posterior probability of 1. bpcomp maxdiff = 0; minimum effective size = 55; maximum rel_diff = 0.2. Most silhouettes from organisms are from Phylopic (phylopic.org/).

supplementary material, figures S1–S5. Results obtained using the UGAMMA model are shown in figure 2*a*, the autocorrelated CIR model in figure 2*b*. Results obtained using the UGAMMA model but with a more permissive exponential root prior are reported in figure 2*a*. Using UGAMMA, 95% credibility intervals surrounding the average divergence times were significantly larger than when the autocorrelated CIR model was used. However, it was evident that for the three nodes of interest (those representing Palaeozoic terrestrialization events) the values in the 95% credibility interval obtained under CIR always represented subsets of the values in the 95% credibility interval obtained using UGAMMA. While the two sets of results are thus statistically indistinguishable, they differ in their congruence with the fossil record. While the more permissive UGAMMA analyses did not reject a Late Cambrian to Silurian origin of the three terrestrial arthropod lineages (the upper limit consistent with the fossil evidence), the CIR model rejected an Ordovician origin for the Myriapoda, suggesting a Precambrian origin instead. Under UGAMMA, arachnid terrestrialization happened in the Silurian, whereas CIR suggests an Ordovician colonization of land. In the case of the Hexapoda, UGAMMA analysis suggested an Ordovician origin, whereas CIR suggested a Cambrian origin and statistically rejected an Early Ordovician origin for this group. Thus, in general, CIR results suggest deeper divergence times. The use of the exponential root, while affecting divergence times of the deepest nodes in our tree (e.g. the age of the Deuterostomia–Protostomia split which is not presented in figure 2, but see electronic supplementary material, figures S1–S5), did not have any effect on the divergence times of the nodes of interest (figure 2 and electronic supplementary material, figure S2).

(c) Ancestral environmental reconstruction

Our ancestral environmental reconstructions (figure 3) aimed to clarify whether the hexapods colonized the land through a freshwater route if their sister group is Branchiopoda rather than Remipedia (figure 1). We found that the last common ancestor of the stem-group Branchiopoda most likely inhabited a marine environment ($p = 0.84$; figure 3). A lower, but not negligible, probability is found for an ancestral freshwater habitat ($p = 0.15$), whereas a brackish ancestry for the total-group Branchiopoda can be confidently rejected ($p = 0.002$; figure 3). Note that these results used a topology where the marine *Rehbachiella* was considered the sister group of the extant branchiopods. As pointed out above, some studies suggested this fossil might instead be allied to cephalocarids [45,90]. If that were the case, given the sister group relationship between cephalocarids and branchiopods suggested in these studies, then a marine origin of Branchiopoda would be inevitable, thus not changing the results of our analyses.

7. Discussion

Terrestrialization is the process through which aquatic organisms adapt to a subaerial lifestyle [7], and abundant literature has addressed this process at the physiological level [9,10,12]. However, most of these studies were performed on isolated

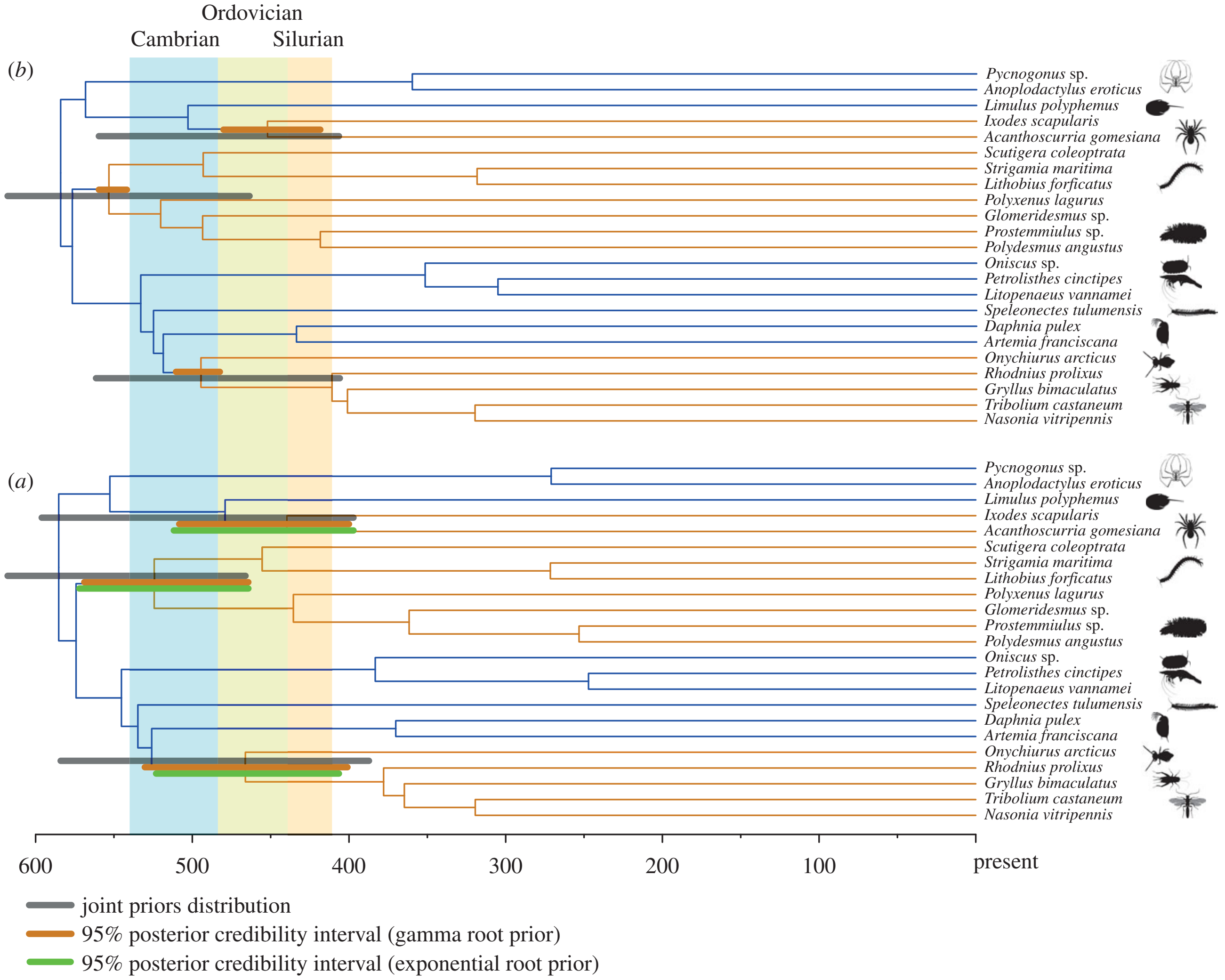


Figure 2. Results of molecular clock analyses. (*a*) Divergence times obtained under the CIR autocorrelated, relaxed, molecular clock model. (*b*) Divergence times obtained using the Uncorrelated Gamma Multipliers model. In both cases, nodes in the tree represent average divergence times estimated using the root prior with 636 Ma mean and 30 Ma SD. Brown bars represent 95% credibility intervals from the considered analysis. Grey bars represent the joint priors (for the considered nodes and analyses). Green bars in figure 2*b* indicate 95% credibility intervals obtained using the exponential prior of average 636 Ma. Blue branches indicate marine lineages. Brown branches terrestrial lineages. In the timescale, numbers represent Myr before the present.

Table 1. Molecular divergence times for key terrestrial arthropod lineages.

	molecular clock model			
	UGAMMA		CIR	
taxon	mean age (Ma)	95% credibility interval	mean age (Ma)	95% credibility interval
Myriapoda	528	568–463	558	572–544
Chilopoda	457	526–408	490	511–452
Diplopoda	439	537–317	519	541–486
Hexapoda	468	512–407	499	431–394
Arachnida	440	518–397	460	493–413

lineages and did not take full advantage of the comparative approach [17], in part because the application of modern comparative methods [100] needs detailed phylogenetic information and divergence times for terrestrial lineages and their close relatives. Such information has only recently started to be available in sufficient detail.

Our phylogenetic analyses used an expanded multigene dataset of wide systematic scope. While our results are consistent with the monophyly of Allotriocarida, in contrast to [45] and other studies [21,23,35,44], we did not find support for a sister group relationship between Remipedia and Hexapoda. We instead recovered Branchiopoda as the sister group of Hexapoda, as has been proposed previously [22]. Our results cannot be taken as definitive, most importantly because, as with all previous relevant analyses we were able to include only one remipede species, and similar to all

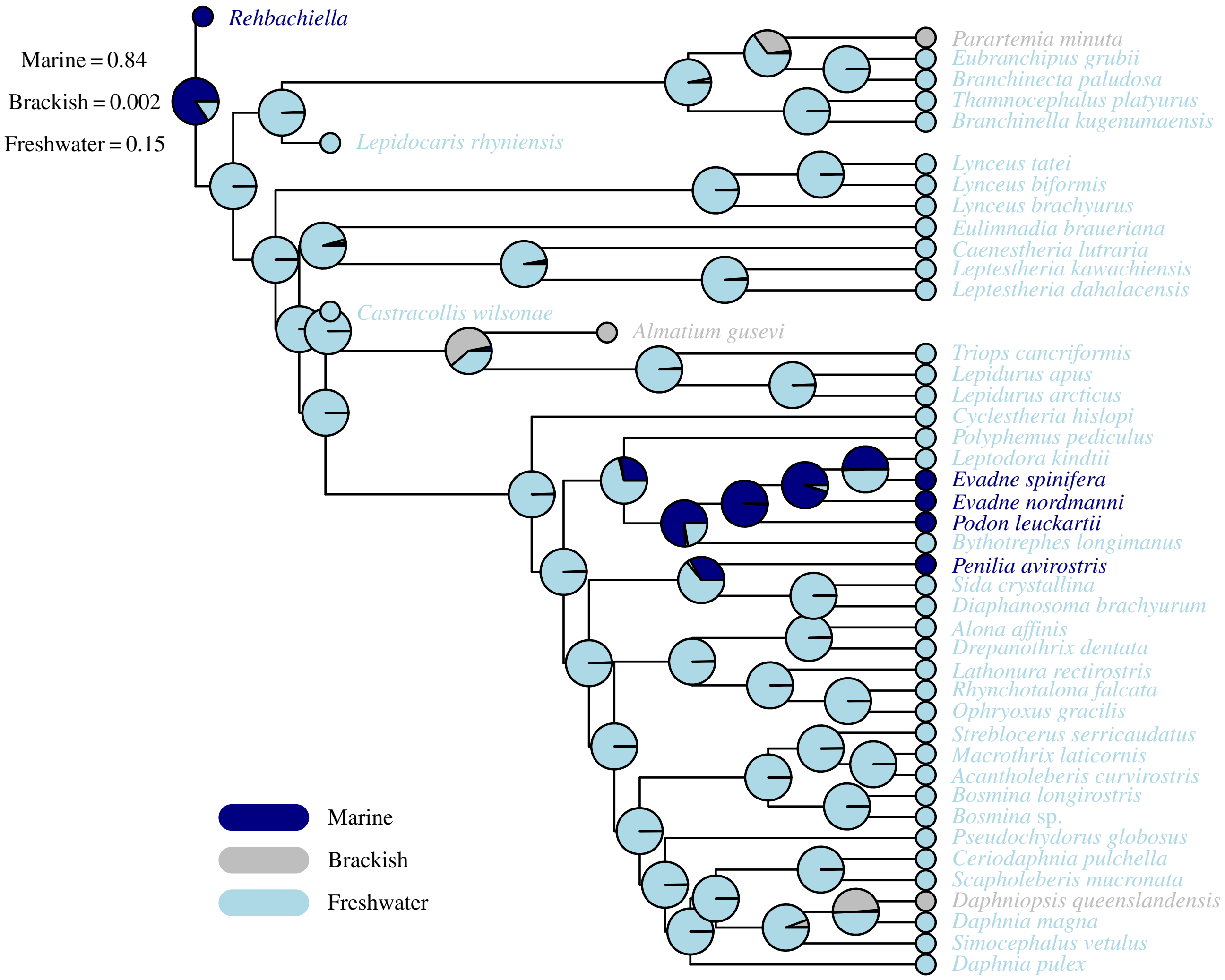


Figure 3. Results of the ancestral environment reconstruction analysis indicating that the last common total-group branchiopod ancestor was most likely a marine organism. The pie charts show the scaled marginal likelihoods of ancestral states for all nodes, with the scaled likelihoods of the total-group ancestor also shown in the text. Branch lengths are proportional to time.

previous studies except that of [45], we did not include cephalocarids. With reference to molecular divergence times, whereas [28] obtained the first set of estimates specifically aiming at clarifying terrestrialization in Arthropoda, their study used a dataset composed of only few genes and taxa and molecular clock methods and calibrations that are now obsolete [101]. The most relevant previous molecular clock study specifically addressing arthropod terrestrialization is that of [8], although divergence times among terrestrial lineages can be found in a variety of other studies [21,67–70,102]. Summarizing results from these previous studies indicates that crown (terrestrial) Myriapoda emerged at 554 Ma, crown (terrestrial) Arachnida emerged at 495 Ma, and crown terrestrial Hexapoda emerged at 495 Ma. These divergence times are broadly in line with the results of our analyses (figure 2 and table 1 and electronic supplementary material, figures S1–S5). In the case of Arachnida, this is broadly compatible with the fossil evidence, whereas in the cases of Hexapoda and particularly Myriapoda the molecular divergences are significantly older. Interpretation of the amphibious euthycarcinoids, which first appear in the Cambrian, as stem-group hexapods [103], goes some way to reconciling early estimates for the origin of Hexapoda and the substantially later appearance of crown-group fossils in the Early Devonian.

A recent fossil-independent attempt at dating the metazoan radiation [104] suggested that divergence times that are substantially in line with the fossil record, like all those reported above except [71], represent artefacts caused by over-constrained calibrations, and that the history of animals is much more in line with previous, outdated, findings that suggested the existence of metazoans approximately 1.5 Ga [105]. Indeed, Battistuzzi *et al.* [104] also suggested that the analyses of Wheat & Walberg [71], despite being in strong disagreement with the arthropod fossil record and with other molecular clock studies of the arthropod radiation, may be accurate. As discussed above, however, the results of [71] are based on a dataset affected by strong compositional biases, and used a pancrustacean topology that has now mostly been contradicted. In addition, it has now been shown that there is not enough information left in genomic datasets to correctly estimate rates of evolution in the deepest part of the animal tree without reference to fossils [102], as advocated by Battistuzzi *et al.* [104]. Tellingly, an analysis of the relative rates of substitution per branch inferred by Battistuzzi *et al.* [104] shows them to be identical (and set to the median rate across their entire tree) in 64.5% of the internal branches in their chronogram (electronic supplementary material, figure S6). Furthermore, these constant strict-clock rates are asymmetrically clustered in the root-ward part of their tree. In other words, the relative divergence time approach used in [104] did not relax the clock in the deepest part of their chronogram, and inferred that more than half of opisthokont history (the outgroup in their chronogram is Fungi) was strictly clocklike. The existence of a deep clock for Metazoa and Opisthokonta is clearly unrealistic and is rejected by the data [102], confirming Pisani & Liu's [101] suggestion that relative divergence times cannot meaningfully be applied in deep time. Given the results of [102], and the rate distribution in electronic supplementary material, figure S6, it is not unsurprising that [104] found results comparable to those found in outdated strict-clock studies [105] from two decades ago. From the point of view of arthropod evolution, the convergence of the results of [104] and [71] further suggests that deep divergence times for the origin of Arthropoda are likely to be artefactual.

Considering hexapod terrestrialization, both the freshwater [50] and the marine [44] routes should be considered valid alternatives. Key to distinguishing between the two is understanding whether the last common ancestor of the Hexapoda and either Remipedia or Branchiopoda inhabited a marine, brackish or freshwater habitat. If the last common ancestor of Hexapoda and its sister clade was a freshwater organism, then the colonization of land could have started from a freshwater habitat. If Remipedia (or Remipedia plus Cephalocarida—if Xenocarida were confirmed in future studies) is confirmed as the sister group of Hexapoda, then a marine route would be strongly favoured as there is no evidence that the anchialine–water dwelling remipedes might have ever been living away from the coasts, whereas cephalocarids are marine. If Branchiopoda is confirmed as the sister group of the hexapods, then the situation would be more ambiguous, as modern branchiopods are mostly found in continental waters, leaving the question of the environmental preferences of the last common branchiopod ancestor unresolved. To address this problem, we used ancestral character reconstruction which suggests that, when both extant and fossil taxa are considered, the last common ancestor of Branchiopoda and Hexapoda was most likely a marine organism. Thus, current evidence, when considering phylogenetic uncertainty of hexapod relationships and fossil evidence, seems to favour a marine route to land also for the Hexapoda. Future discoveries of additional Cambrian stem-group branchiopods could better clarify this problem.

8. Conclusion

Ephemeral, terrestrial habitats have long existed on the Earth, at the very least since approximately 1 Ga. However, animal terrestrialization was a much more recent process. This was first of all because animals originated in the Cryogenian and radiated close to the base of the Cambrian, in disagreement with [104], and in agreement with [83,102]. Our molecular clock results cannot reject fossil-based divergence times for Arachnida and Hexapoda, and we thus conclude that the most likely scenario, given the current evidence, is that these lineages colonized the land in the Ordovician or the Silurian (Arachnida) and the Ordovician (Hexapoda). Estimates that Myriapoda may have colonized land earlier are in disagreement with the myriapod fossil record, even allowing that terrestrial ecosystems already existed in the Cambrian. A mid-late Cambrian diversification of Diplopoda has, however, been predicted based on geographic distributions of extant millipedes and palaeogeography [106]. We do, however, note that our results for the origins of Chilopoda and Diplopoda are consistent with current fossil evidence (figure 2 and electronic supplementary material, figures S1–S5). One possible scenario that would partly resolve this clash between fossils and molecules would be that these two lineages independently colonized the land; but for that to be the case, tracheae should have evolved independently. This possibility has been suggested previously based on differences in structure of the tracheae and position of the spiracles [107] and should be subjected to critical testing. Irrespective of the precise time at which different arthropods colonized land, it seems currently more likely that the process of animal terrestrialization did not begin before the Late Cambrian and proceeded from the coastline towards the centre of the continents.

Data accessibility. Supplementary Information are available with the paper http://dx.doi.org/10.1098/rstb.2015.0133. The multiple-sequence alignment used for the analyses is available for download at https://bitbucket.org/bzxdp/terrestrialisation_arthropoda.

Authors' contributions. J.L.-F. carried out analyses and wrote the manuscript. R.C. assembled the dataset and helped writing the manuscript. A.T. provided scripts for dataset assembly and helped writing the manuscript. J.O., M.B. and J.V. provided data and helped writing the manuscript. G.G. and G.E. contributed to write the manuscript. M.P. carried out ancestral character state reconstruction and helped writing the manuscript. DP contributed to all steps of the analyses and to writing the manuscript.

Competing interests. We confirm that we have no competing interests.

Funding. This work was supported by a Marie Skłodowska-Curie Fellowship to J.L.-F. R.C. was supported by a Science Foundation Ireland grant to D.P. (11/RFP/EOB/3106), A.R.T. was supported by a University of Bristol (STAR) PhD studentship. M.P. was supported by a NERC PhD studentship. Edinburgh Genomics is partially supported by core grants from (NERC R8/H10/56), MRC (MR/K001744/1) and BBSRC (BB/J004243/1). J.O. was supported by a grant from the Danish Agency for Science, Technology and Innovation (0601-12345B).

Acknowledgements. We thank the editors for having invited us to contribute to this issue of Philosophical Transactions of the Royal Society and two anonymous reviewers for providing constructive criticism. Thanks to staff at Edinburgh Genomics and the Geogenetics center in Copenhagen for help with sequencing the pycnogonid and *Limulus* transcriptomes.

References

1. Labandeira CC. 2005 Invasion of the continents: cyanobacterial crusts to tree-inhabiting arthropods. *Trends Ecol. Evol.* **20**, 253–262. (doi:10.1016/j.tree.2005.03.002)
2. Shear WA. 1991 The early development of terrestrial ecosystems. *Nature* **351**, 283–289. (doi:10.1038/351283a0)
3. Strother PK, Battison L, Brasier MD, Wellman CH. 2011 Earth's earliest non-marine eukaryotes. *Nature* **473**, 505–509. (doi:10.1038/nature09943)
4. Clarke JT, Warnock R, Donoghue PCJ. 2011 Establishing a time-scale for plant evolution. *New Phytol.* **192**, 266–301. (doi:10.1111/j.1469-8137.2011.03794.x)
5. Kenrick P, Wellman CH, Schneider H, Edgecombe GD. 2012 A timeline for terrestrialization: consequences for the carbon cycle in the Palaeozoic. *Phil. Trans. R. Soc. B* **367**, 519–536. (doi:10.1098/rstb.2011.0271)
6. Redecker D, Kodner R, Graham LE. 2000 Glomalean fungi from the Ordovician. *Science*

289, 1920–1921. (doi:10.1126/science.289.5486.1920)

7. Little C. 1983 *The colonisation of land: origins and adaptations of terrestrial animals*, 300 p. Cambridge, UK: Cambridge University Press.
8. Rota-Stabelli O, Daley AC, Pisani D. 2013 Molecular timetrees reveal a Cambrian colonization of land and a new scenario for ecdysozoan evolution. *Curr. Biol*. **23**, 392–398. (doi:10.1016/j.cub.2013.01.026)
9. Richardson A, Araujo PB. 2015 Lifestyles of terrestrial crustaceans. In M Thiel, L Walting (eds), *The natural history of the Crustacea. Lifestyles and feeding biology*, pp. 299–336. New York, NY: Oxford University Press.
10. Little C. 1990 *The terrestrial invasion: an ecophysiological approach to the origins of land animals*, 304 p. Cambridge, UK: Cambridge University Press.
11. Dunlop JA, Scholtz G, Selden PA. 2013 Water-to-Land Transitions. In *Arthropod Biology and Evolution*, pp. 417–439. Berlin, Germany: Springer Berlin Heidelberg.
12. Gordon MS, Olson EC. 1995 *Invasions of the land: the transitions of organisms from aquatic to terrestrial life*. New York, NY: Columbia University Press.
13. Selden PA. 2001 Terrestrialization (Precambrian–Devonian). In *eLS*. Hoboken, NJ: John Wiley & Sons, Ltd. (doi:10.10.1038/npg.els.0001641)
14. Niimura Y. 2009 On the origin and evolution of vertebrate olfactory receptor genes: comparative genome analysis among 23 chordate species. *Genome Biol. Evol*. **1**, 34–44. (doi:10.1093/gbe/evp003)
15. Niimura Y, Nei M. 2005 Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods. *Proc. Natl Acad. Sci. USA* **102**, 6039–6044. (doi:10.1073/pnas.0501922102)
16. Vieira FG, Rozas J. 2011 Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: origin and evolutionary history of the chemosensory system. *Genome Biol. Evol*. **3**, 476–490. (doi:10.1093/gbe/evr033)
17. Felsenstein J. 1985 Phylogenies and the comparative method. *Am. Nat*. **125**, 1–15. (doi:10.1086/284325)
18. Stork NE, McBroom J, Gely C, Hamilton AJ. 2015 New approaches narrow global species estimates for beetles, insects, and terrestrial arthropods. *Proc. Natl Acad. Sci. USA* **112**, 7519–7523. (doi:10.1073/pnas.1502408112)
19. Giribet G, Edgecombe GD. 2012 Reevaluating the arthropod tree of life. *Annu. Rev. Entomol*. **57**, 167–186. (doi:10.1146/annurev-ento-120710-100659)
20. Rota-Stabelli O, Campbell L, Brinkmann H, Edgecombe GD, Longhorn SJ, Peterson KJ, Pisani D, Philippe H, Telford MJ. 2011 A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. *Proc. R. Soc. B* **278**, 298–306. (doi:10.1098/rspb.2010.0590)
21. Misof B *et al.* 2014 Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763–767. (doi:10.1126/science.1257570)
22. Borner J, Rehm P, Schill RO, Ebersberger I, Burmester T. 2014 A transcriptome approach to ecdysozoan phylogeny. *Mol. Phylogenet. Evol*. **80**, 79–87. (doi:10.1016/j.ympev.2014.08.001)
23. Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, Wetzer R, Martin JW, Cunningham CW. 2010 Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature* **463**, 1079–1083. (doi:10.1038/nature08742)
24. Wägele JW, Kück P. 2013 Arthropod phylogeny and the origin of Tracheata (=Atelocerata) from Remipedia-like ancestors. In *Deep metazoan phylogeny: the backbone of the tree of life*, pp. 285–341. Berlin, Germany: De Gruyter.
25. Wheeler WC, Giribet G, Edgecombe GD. 2004 Arthropod systematics. The comparative study of genomic, anatomical, and paleontological information. In *Assembling the tree of life*, pp. 281–295. New York, NY: Oxford University Press.
26. Friedrich M, Tautz D. 1995 Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods. *Nature* **376**, 165–167. (doi:10.1038/376165a0)
27. Cook CE, Smith ML, Telford MJ, 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)
28. Pisani D, Poling LL, Lyons-Weiler M, Hedges SB. 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)
29. 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. (doi:10.1016/j.ympev.2003.07.013)
30. Meusemann K *et al.* 2010 A phylogenomic approach to resolve the arthropod tree of life. *Mol. Biol. Evol*. **27**, 2451–2464. (doi:10.1093/molbev/msq130)
31. Boore JL, Lavrov DV, Brown WM. 1998 Gene translocation links insects and crustaceans. *Nature* **392**, 667–668. (doi:10.1038/33577)
32. Zrzavý J, Štys P. 1997 The basic body plan of arthropods: insights from evolutionary morphology and developmental biology. *J. Evol. Biol*. **10**, 353–367. (doi:10.1046/j.1420-9101.1997.10030353.x)
33. Richter S. 2002 The Tetraconata concept: hexapod-crustacean relationships and the phylogeny of Crustacea. *Org. Divers. Evol*. **2**, 217–237. (doi:10.1078/1439-6092-00048)
34. Regier JC, Shultz JW, Kambic RE. 2005 Pancrustacean phylogeny: hexapods are terrestrial crustaceans and maxillopods are not monophyletic. *Proc. R. Soc. B* **272**, 395–401. (doi:10.1098/rspb.2004.2917)
35. Regier JC *et al.* 2008 Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. *Syst. Biol*. **57**, 920–938. (doi:10.1080/106351432#50802570791)
36. Sharma PP, Kaluziak ST, Pérez-Porro AR, González VL, Hormiga G, Wheeler WC, Giribet G. 2014 Phylogenomic interrogation of Arachnida reveals systemic conflicts in phylogenetic signal. *Mol. Biol. Evol*. **31**, 2963–2984. (doi:10.1093/molbev/msu235)
37. Rota-Stabelli O, Lartillot N, Philippe H, Pisani D. 2013 Serine codon-usage bias in deep phylogenomics: pancrustacean relationships as a case study. *Syst. Biol*. **62**, 121–133. (doi:10.1093/sysbio/sys077)
38. Giribet G, Edgecombe GD, Wheeler WC. 2001 Arthropod phylogeny based on eight molecular loci and morphology. *Nature* **413**, 157–161. (doi:10.1038/35093097)
39. Bernt M, Braband A, Middendorf M, Misof B, Rota-Stabelli O, Stadler PF. 2013 Bioinformatics methods for the comparative analysis of metazoan mitochondrial genome sequences. *Mol. Phylogenet. Evol*. **69**, 320–327. (doi:10.1016/j.ympev.2012.09.019)
40. Rota-Stabelli O, Kayal E, Gleeson D, Daub J, Boore JL, Telford MJ, Pisani D, Blaxter M, Lavrov DV. 2010 Ecdysozoan mitogenomics: evidence for a common origin of the legged invertebrates, the Panarthropoda. *Genome Biol. Evol*. **2**, 425–440. (doi:10.1093/gbe/evq030)
41. Nardi F, Spinsanti G, Boore JL, Carapelli A, Dallai R, Frati F. 2003 Hexapod origins: monophyletic or paraphyletic? *Science* **299**, 1887–1889. (doi:10.1126/science.1078607)
42. Delsuc F, Phillips MJ, Penny D. 2003 Comment on 'Hexapod origins: monophyletic or paraphyletic?' *Science* **301**, 1482; author reply 1482. (doi:10.1126/science.1086558)
43. 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. *Mol. Phylogenet. Evol*. **38**, 100–116. (doi:10.1016/j.ympev.2005.09.012)
44. von Reumont BM *et al.* 2012 Pancrustacean phylogeny in the light of new phylogenomic data: support for Remipedia as the possible sister group of Hexapoda. *Mol. Biol. Evol*. **29**, 1031–1045. (doi:10.1093/molbev/msr270)
45. Oakley TH, Wolfe JM, Lindgren AR, Zaharoff AK. 2013 Phylotranscriptomics to bring the understudied into the fold: monophyletic Ostracoda, fossil placement, and pancrustacean phylogeny. *Mol. Biol. Evol*. **30**, 215–233. (doi:10.1093/molbev/mss216)
46. Ertas B, von Reumont BM, Wägele J-W, Misof B, Burmester T. 2009 Hemocyanin suggests a close relationship of Remipedia and Hexapoda. *Mol. Biol. Evol*. **26**, 2711–2718. (doi:10.1093/molbev/msp186)
47. Fanenbruck M, Harzsch S, Wägele JW. 2004 The brain of the Remipedia (Crustacea) and an alternative hypothesis on their phylogenetic relationships. *Proc. Natl Acad. Sci. USA* **101**, 3868–3873. (doi:10.1073/pnas.0306212101)
48. Fanenbruck M, Harzsch S. 2005 A brain atlas of *Godzilliognomus frondosus* Yager, 1989 (Remipedia, Godzilliidae) and comparison with the brain of *Speleonectes tulumensis* Yager, 1987 (Remipedia, Speleonectidae): implications for arthropod

relationships. *Arthropod Struct. Dev.* **34**, 343–378. (doi:10.1016/j.asd.2005.01.007)

49. Stemme T, Iliffe TM, Bicker G, Harzsch S, Koenemann S. 2012 Serotonin immunoreactive interneurons in the brain of the Remipedia: new insights into the phylogenetic affinities of an enigmatic crustacean taxon. *BMC Evol. Biol.* **12**, 168. (doi:10.1186/1471-2148-12-168)
50. Glenner H, Thomsen PF, Hebsgaard MB, Sorensen MV, Willerslev E. 2006 Evolution: the origin of insects. *Science* **314**, 1883–1884. (doi:10.1126/science.1129844)
51. Maloof AC, Porter SM, Moore JL, Dudás FÖ, Bowring SA, Higgins JA, Fike DA, Eddy MP. 2010 The earliest Cambrian record of animals and ocean geochemical change. *GSA Bull.* **122**, 1731–1774. (doi:10.1130/B30346.1)
52. Legg DA. 2014 *Sanctacaris uncata*: the oldest chelicerate (Arthropoda). *Naturwissenschaften* **101**, 1065–1073. (doi:10.1007/s00114-014-1245-4)
53. Waloszek D, Dunlop JA. 2002 A larval sea spider (Arthropoda: Pycnogonida) from the Upper Cambrian 'Orsten' of Sweden, and the phylogenetic position of pycnogonids. *Palaeontology* **45**, 421–446. (doi:10.1111/1475-4983.00244)
54. Harvey THP, Vélez MI, Butterfield NJ. 2012 Exceptionally preserved crustaceans from western Canada reveal a cryptic Cambrian radiation. *Proc. Natl Acad. Sci. USA* **109**, 1589–1594. (doi:10.1073/pnas.1115244109)
55. Walossek D. 1993 *The Upper Cambrian Rehbachiella and the phylogeny of Branchiopoda and Crustacea.* Fossils and Strata no. 32, 202 p. Oslo, Norway: Scandinavian University Press.
56. MacNaughton RB, Cole JM, Dalrymple RW, Braddy SJ, Briggs DEG, Lukie TD. 2002 First steps on land: arthropod trackways in Cambrian–Ordovician eolian sandstone, southeastern Ontario, Canada. *Geology* **30**, 391–394. (doi:10.1130/0091-7613(2002)030<0391:FSOLAT>2.0.CO;2)
57. Collette JH, Gass KC, Hagadorn JW. 2012 *Protichnites eremita* unshelled? experimental model-based neoichnology and new evidence for a euthycarcinoid affinity for this ichnospecies. *J. Paleontol.* **86**, 442–454. (doi:10.1666/11-056.1)
58. Wilson HM, Anderson LI. 2004 Morphology and taxonomy of Paleozoic millipedes (Diplopoda: Chilognatha: Archipolypoda) from Scotland. *J. Paleontol.* **78**, 169–184. (doi:10.1666/0022-3360(2004)078<0169:MATOPM>2.0.CO;2)
59. Shear WA, Edgecombe GD. 2010 The geological record and phylogeny of the Myriapoda. *Arthropod Struct. Dev.* **39**, 174–190. (doi:10.1016/j.asd.2009.11.002)
60. Edgecombe GD. 2004 Morphological data, extant Myriapoda, and the myriapod stem-group. *Contrib. Zool.* **73**, 207–252.
61. Jeram AJ, Selden PA, Edwards D. 1990 Land animals in the Silurian: arachnids and myriapods from Shropshire, England. *Science* **250**, 658–661. (doi:10.1126/science.250.4981.658)
62. Scholtz G, Kamenz C. 2006 The book lungs of Scorpiones and Tetrapulmonata (Chelicerata, Arachnida): evidence for homology and a single terrestrialisation event of a common arachnid ancestor. *Zoology* **109**, 2–13. (doi:10.1016/j.zool.2005.06.003)
63. Dunlop JA, Anderson LI, Braddy SJ. 2003 A redescription of *Chasmataspis laurencii* Caster and Brooks, 1956 (Chelicerata: Chasmataspidida) from the Middle Ordovician of Tennessee, USA, with remarks on chasmataspid phylogeny. *Trans. R. Soc. Edinb. Earth Sci.* **94**, 207–225. (doi:10.1017/S0263593303000130)
64. Lamsdell JC. 2013 Revised systematics of Palaeozoic 'horseshoe crabs' and the myth of monophyletic Xiphosura. *Zool. J. Linn. Soc.* **167**, 1–27. (doi:10.1111/j.1096-3642.2012.00874.x)
65. Edwards D, Selden PA, Richardson JB, Axe L. 1995 Coprolites as evidence for plant–animal interaction in Siluro–Devonian terrestrial ecosystems. *Nature* **377**, 329–331. (doi:10.1038/377329a0)
66. Parry SF, Noble SR, Crowley QG, Wellman CH. 2011 A high-precision U–Pb age constraint on the Rhynie Chert Konservat-Lagerstätte: time scale and other implications. *J. Geol. Soc. Lond.* **168**, 863–872. (doi:10.1144/0016-76492010-043)
67. Rehm P, Borner J, Meusemann K, von Reumont BM, Simon S, Hadrys H, Misof B, Burmester T. 2011 Dating the arthropod tree based on large-scale transcriptome data. *Mol. Phylogenet. Evol.* **61**, 880–887. (doi:10.1016/j.ympev.2011.09.003)
68. Rehm P, Meusemann K, Borner J, Misof B, Burmester T. 2014 Phylogenetic position of Myriapoda revealed by 454 transcriptome sequencing. *Mol. Phylogenet. Evol.* **77**, 25–33. (doi:10.1016/j.ympev.2014.04.007)
69. Brewer MS, Bond JE. 2013 Ordinal-level phylogenomics of the arthropod class Diplopoda (millipedes) based on an analysis of 221 nuclear protein-coding loci generated using next-generation sequence analyses. *PLoS ONE* **8**, e79935. (doi:10.1371/journal.pone.0079935)
70. Tong KJ, Duchêne S, Ho SYW, Lo N. 2015 Insect phylogenomics. Comment on 'Phylogenomics resolves the timing and pattern of insect evolution'. *Science* **349**, 487. (doi:10.1126/science.aaa5460)
71. Wheat CW, Wahlberg N. 2013 Phylogenomic insights into the Cambrian explosion, the colonization of land and the evolution of flight in Arthropoda. *Syst. Biol.* **62**, 93–109. (doi:10.1093/sysbio/sys074)
72. Campbell LI *et al.* 2011 MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda. *Proc. Natl Acad. Sci. USA* **108**, 15 920–15 924. (doi:10.1073/pnas.1105499108)
73. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402. (doi:10.1093/nar/25.17.3389)
74. Edgar RC. 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797. (doi:10.1093/nar/gkh340)
75. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010 New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321. (doi:10.1093/sysbio/syq010)
76. Castresana J. 2000 Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552. (doi:10.1093/oxfordjournals.molbev.a026334)
77. Kück P, Meusemann K. 2010 FASconCAT: convenient handling of data matrices. *Mol. Phylogenet. Evol.* **56**, 1115–1118. (doi:10.1016/j.ympev.2010.04.024)
78. Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013 PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.* **62**, 611–615. (doi:10.1093/sysbio/syt022)
79. 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)
80. Lartillot N, Lepage T, Blanquart S. 2009 PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286–2288. (doi:10.1093/bioinformatics/btp368)
81. Lepage T, Bryant D, Philippe H, Lartillot N. 2007 A general comparison of relaxed molecular clock models. *Mol. Biol. Evol.* **24**, 2669–2680. (doi:10.1093/molbev/msm193)
82. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006 Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88. (doi:10.1371/journal.pbio.0040088)
83. Erwin DH, Laflamme M, Tweedt SM, Sperling EA, Pisani D, Peterson KJ. 2011 The Cambrian conundrum: early divergence and later ecological success in the early history of animals. *Science* **334**, 1091–1097. (doi:10.1126/science.1206375)
84. Paradis E, Claude J, Strimmer K. 2004 APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290. (doi:10.1093/bioinformatics/btg412)
85. Pagel M. 1994 Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R. Soc. B* **255**, 37–45. (doi:10.1098/rspb.1994.0006)
86. Lewis PO. 2001 A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* **50**, 913–925. (doi:10.1080/106351501753462876)
87. Stenderup JT, Olesen J, Glenner H. 2006 Molecular phylogeny of the Branchiopoda (Crustacea)–Multiple approaches suggest a 'diplostracan' ancestry of the Notostraca. *Mol. Phylogenet. Evol.* **41**, 182–194. (doi:10.1016/j.ympev.2006.06.006)
88. Olesen J. 2007 Monophyly and phylogeny of Branchiopoda, with focus on morphology and homologies of branchiopod phyllopodous limbs.

J. Crustacean Biol. **27**, 165–183. (doi:10.1651/S-2727.1)

89. Olesen J. 2009 Phylogeny of Branchiopoda (Crustacea)—character evolution and contribution of uniquely preserved fossils. *Arthropod Syst. Phylogeny* **67**, 3–39.

90. Wolfe JM, Hegna TA. 2014 Testing the phylogenetic position of Cambrian pancrustacean larval fossils by coding ontogenetic stages. *Cladistics* **30**, 366–390. (doi:10.1111/cla.12051)

91. Scourfield DJ. 1926 On a new type of crustacean from the Old Red Sandstone (Rhynie Chert Bed, Aberdeenshire)-*Lepidocaris rhyniensis*, gen. et sp. nov. *Phil. Trans. R. Soc. Lond. B* **214**, 153–187. (doi:10.1098/rstb.1926.0005)

92. Fayers SR, Trewin NH. 2002 A new crustacean from the Early Devonian Rhynie chert, Aberdeenshire, Scotland. *Earth Environ. Sci. Trans. R. Soc. Edinb.* **93**, 355–382. (doi:10.1017/S0263593302000196)

93. Novozhilov NI. 1957 Un nouvel ordre d'arthropodes particuliers: Kazacharthra du Lias des monts Ketmen: (Kazakhstan, SE., URSS). *Bull. Soc. Géol. Fr.* **7**, 171–184.

94. Mathers TC, Hammond RL, Jenner RA, Hänfling B, Gómez A. 2013 Multiple global radiations in tadpole shrimps challenge the concept of 'living fossils'. *PeerJ* **1**, e62. (doi:10.7717/peerj.62)

95. Bapst DW. 2012 paleotree: an R package for paleontological and phylogenetic analyses of evolution. *Methods Ecol. Evol.* **3**, 803–807. (doi:10.1111/j.2041-210X.2012.00223.x)

96. Smith MR, Ortega-Hernández J. 2014 *Hallucigenia*'s onychophoran-like claws and the case for Tactopoda. *Nature* **514**, 363–366. (doi:10.1038/nature13576)

97. Dunn CW *et al.* 2008 Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **452**, 745–749. (doi:10.1038/nature06614)

98. Laumer CE *et al.* 2015 Spiralian phylogeny informs the evolution of microscopic lineages. *Curr. Biol.* **25**, 2000–2006. (doi:10.1016/j.cub.2015.06.068)

99. Fernández R, Laumer CE, Vahtera V, Libro S, Kaluziak S, Sharma PP, Pérez-Porro AR, Edgecombe GD, Giribet G. 2014 Evaluating topological conflict in centipede phylogeny using transcriptomic data sets. *Mol. Biol. Evol.* **31**, 1500–1513. (doi:10.1093/molbev/msu108)

100. Paradis E. 2012 *Analysis of phylogenetics and evolution with R*, 2nd edn. 386 p. New York, NY: Springer.

101. Pisani D, Liu AG. 2015 Animal evolution: only rocks can set the clock. *Curr. Biol.* **25**, 1079–1081. (doi:10.1016/j.cub.2015.10.015)

102. dos Reis M, Thawornwattana Y, Angelis K, Telford MJ, Donoghue PCJ, Yang Z. 2015 Uncertainty in the timing of origin of animals and the limits of precision in molecular timescales. *Curr. Biol.* **25**, 2939–2950. (doi:10.1016/j.cub.2015.09.066)

103. Legg DA, Sutton MD, Edgecombe GD. 2013 Arthropod fossil data increase congruence of morphological and molecular phylogenies. *Nat. Commun.* **4**, 2485. (doi:10.1038/ncomms3485)

104. Battistuzzi FU, Billing-Ross P, Murillo O, Filipski A, Kumar S. 2015 A protocol for diagnosing the effect of calibration priors on posterior time estimates: a case study for the Cambrian explosion of animal phyla. *Mol. Biol. Evol.* **32**, 1907–1912. (doi:10.1093/molbev/msv075)

105. Wray GA, Levinton JS, Shapiro LH. 1996 Molecular evidence for deep Precambrian divergences among metazoan phyla. *Science* **274**, 568–573. (doi:10.1126/science.274.5287.568)

106. Shelley RM, Golavatch SI. 2011 Atlas of myriapod biogeography. I. Indigenous ordinal and supra-ordinal distributions in the Diplopoda: Perspectives on taxon origins and ages, and a hypothesis on the origin and early evolution of the class. *Insecta Mundi* **158**, 1–134.

107. Dohle W. 1998 Myriapod–insect relationships as opposed to an insect-crustacean sister group relationship. In *Arthropod relationships*, pp. 305–315. London, UK: Chapman & Hall.

PROCEEDINGS B

rspb.royalsocietypublishing.org

Research

Cite this article: Tanner AR *et al.* 2017 Molecular clocks indicate turnover and diversification of modern coleoid cephalopods during the Mesozoic Marine Revolution. *Proc. R. Soc. B* **284**: 20162818. http://dx.doi.org/10.1098/rspb.2016.2818

Received: 20 December 2016
Accepted: 8 February 2017

Subject Category:
Palaeobiology

Subject Areas:
evolution, palaeontology, molecular biology

Keywords:
Cephalopoda, molecular phylogenetics, phylogenomics, molecular clocks

Author for correspondence:
Jakob Vinther
e-mail: jakob.vinther@bristol.ac.uk

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3693223.

THE ROYAL SOCIETY
PUBLISHING

Molecular clocks indicate turnover and diversification of modern coleoid cephalopods during the Mesozoic Marine Revolution

Alastair R. Tanner[1], Dirk Fuchs[3], Inger E. Winkelmann[4], M. Thomas P. Gilbert[4,5,6], M. Sabrina Pankey[7], Ângela M. Ribeiro[4], Kevin M. Kocot[9], Kenneth M. Halanych[10], Todd H. Oakley[11], Rute R. da Fonseca[4,8], Davide Pisani[1,2] and Jakob Vinther[1,2]

[1]School of Biological Sciences, and [2]School of Earth Sciences, University of Bristol, Life Sciences Building, 24 Tyndall Avenue, Bristol BS8 1TQ, UK
[3]Earth and Planetary System Science, Department of Natural History Sciences, Hokkaido University, Sapporo, Japan
[4]Natural History Museum of Denmark, Øster Voldgade 5-7, 1350 Copenhagen, Denmark
[5]Trace and Environmental DNA Laboratory, Department of Environment and Agriculture, Curtin University, Perth, Western Australia, Australia
[6]NTNU University Museum, Norwegian University of Science and Technology, Trondheim, Norway
[7]Molecular, Cellular and Biomedical Sciences, University of New Hampshire, Durham, NH 03824, USA
[8]Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen N, Denmark
[9]Department of Biological Sciences, University of Alabama, Box 870344, Tuscaloosa, AL 35487, USA
[10]Department of Biological Sciences, Auburn University, Auburn, AL 36830, USA
[11]Department of Ecology, Evolution, and Marine Biology, University of California, Santa Barbara, CA 93106, USA

ART, 0000-0001-8045-2856; DP, 0000-0003-0949-6682; JV, 0000-0002-3584-9616

Coleoid cephalopod molluscs comprise squid, cuttlefish and octopuses, and represent nearly the entire diversity of modern cephalopods. Sophisticated adaptations such as the use of colour for camouflage and communication, jet propulsion and the ink sac highlight the unique nature of the group. Despite these striking adaptations, there are clear parallels in ecology between coleoids and bony fishes. The coleoid fossil record is limited, however, hindering confident analysis of the tempo and pattern of their evolution. Here we use a molecular dataset (180 genes, approx. 36 000 amino acids) of 26 cephalopod species to explore the phylogeny and timing of cephalopod evolution. We show that crown cephalopods diverged in the Silurian–Devonian, while crown coleoids had origins in the latest Palaeozoic. While the deep-sea vampire squid and dumbo octopuses have ancient origins extending to the Early Mesozoic Era, 242 ± 38 Ma, incirrate octopuses and the decabrachian coleoids (10-armed squid) diversified in the Jurassic Period. These divergence estimates highlight the modern diversity of coleoid cephalopods emerging in the Mesozoic Marine Revolution, a period that also witnessed the radiation of most ray-finned fish groups in addition to several other marine vertebrates. This suggests that that the origin of modern cephalopod biodiversity was contingent on ecological competition with marine vertebrates.

1. Introduction

Octopus, cuttlefish and squid showcase advanced intelligence, a wide range of body sizes, sophisticated camouflage and mimicry, unique jet-locomotion and ingenious decoy countermeasures in the ink sac [1–3]. Charismatic in these ways, and owing to their importance as fishing stocks, cephalopods have

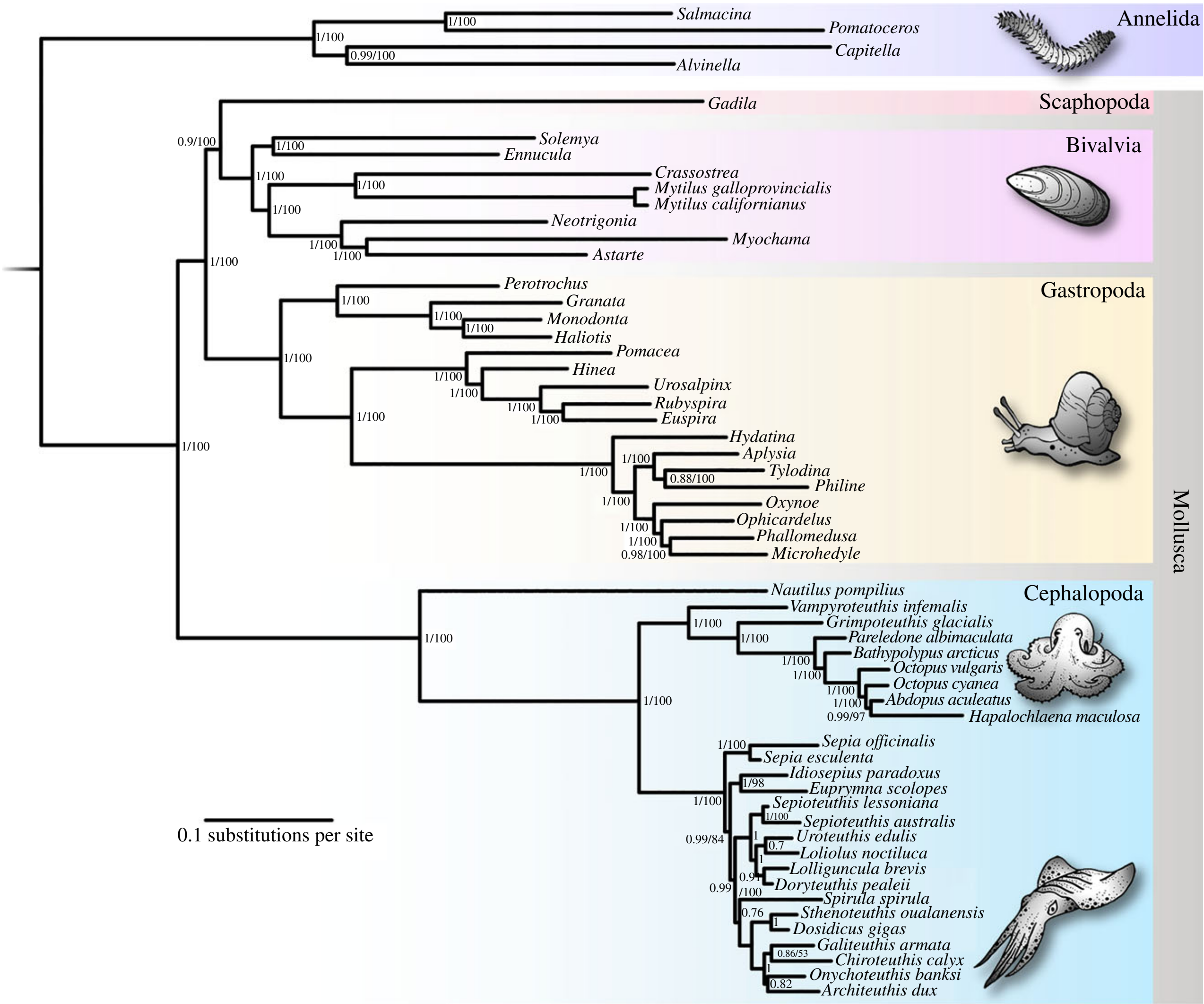


Figure 1. Molecular phylogeny of cephalopod, gastropod and bivalve molluscs (plus a scaphopod), with annelid outgroup; 180 genes, concatenated as 36 156 aligned amino acid positions with 26% missing data, modelled under CAT + GTR + Γ. Numbers at nodes denote Bayesian posterior probability/bootstrap support as returned by RAxML under the LG [33] substitution model. Scale bar is expected substitutions per site.

garnered great interest from ecologists and evolutionary biologists. However, cephalopod evolutionary relationships and divergence times have remained unclear, in part, owing to uncertainties in their fossil record. The past 540 Ma of cephalopod evolution can be viewed as having three ecologically distinct phases. Originally shelled, sea-floor-dwelling molluscs, cephalopods are descended from superficially limpet-like ancestors in the Cambrian [4,5]. The protective shell later became adapted as a chambered buoyancy organ [6], giving rise to free-swimming forms by the latest Cambrian that radiated into several Ordovician lineages [7]. Subsequently, internalization and reduction of the mineralized shell facilitated adaptation for alternative ecologies in the coleoids [8].

Anatomical evolution is in part shaped by the ecological relationships between predator– and prey species. Cephalopods (and in particular oceanic squid) fill a niche that largely overlaps with fishes as active mesopredators [9]. Considering the evolutionary trajectory of cephalopods from heavily shelled animals to rapid hunters, the question of how and when this development took place remains unresolved. Previously, coevolution between marine predators and prey has been hypothesized from the fossil record of the Jurassic and the Cretaceous, and this ecological shift has since become known as the Mesozoic Marine Revolution [10,11].

By contrast, the fossil record leaves limited insight on the providence of modern coleoid groups [12], despite their well-documented ancestors and relatives especially among the ammonites and belemnites. Their mineralized, chambered portion of the shell (phragmocone and rostrum) has a high potential for preservation, but as the phragmocone became internalized, reduced, and in many cases lost entirely, so too was a clear narrative through fossils. Soft tissue fossilization is rare, but cirrate and incirrate octopods are known from the Late Cretaceous (Cenomanian) Hâkel and Hâdjoula Lagerstätte, while cirrate forms and stem octobrachians are recorded in the Jurassic [13]; these are known to preserve the unmineralized gladius and soft tissues. Stem group decabrachians, such as belemnites and other belemnoids are known, preserving their phragmocones and, occasionally, soft tissues [14,15]. By contrast, the extant octopuses, cuttlefish and squid are characterized by shell reduction and loss [16], and are prone to major taphonomic biases in tissue preservation [14]. Consequently, clarifying evolution of coleoids from the Mid-Palaeozoic to the present must, therefore, rely on alternative palaeobiological approaches, such as the estimation of molecular divergence times.

The first molecular divergence times of cephalopod evolution recovered very ancient divergences for the coleoids [17],

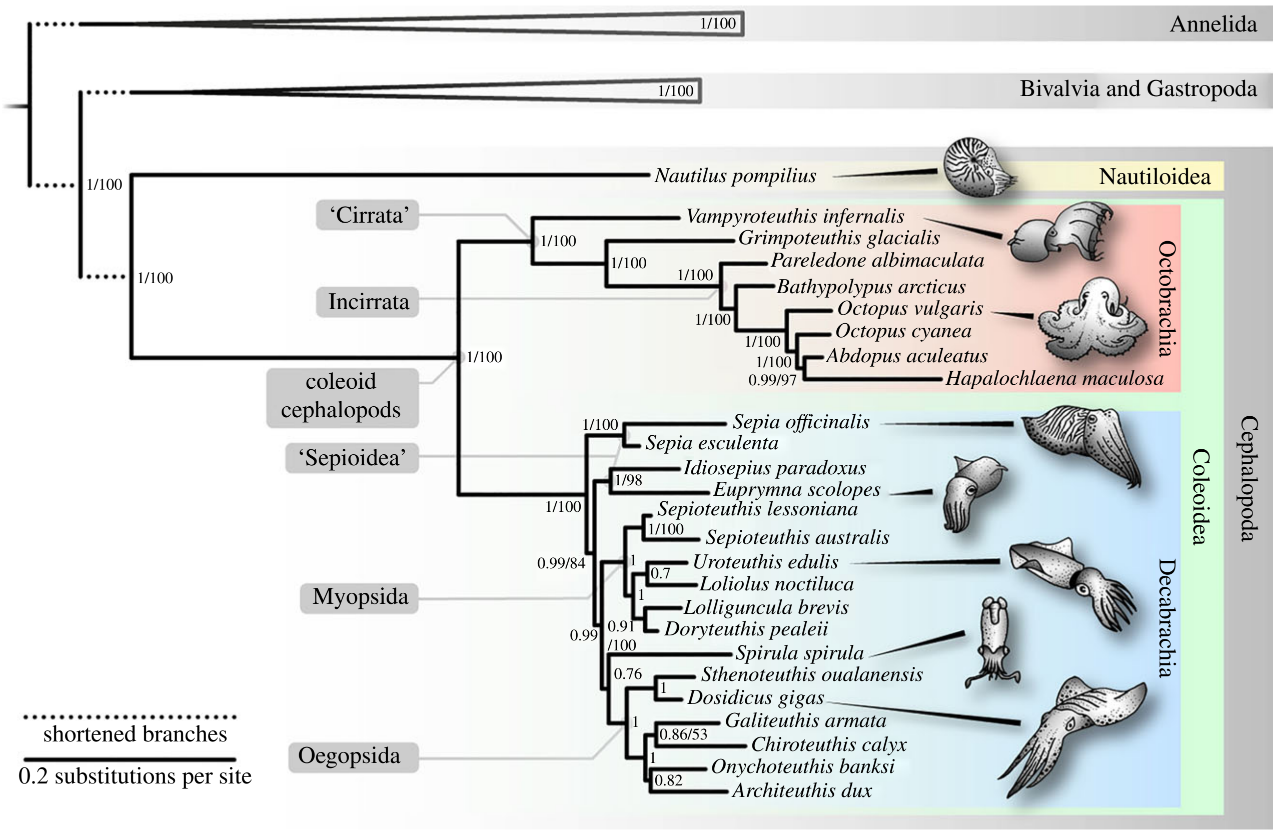


Figure 2. Phylogeny of 26 cephalopod species, plus outgroups (further details in figure 1); 180 genes, concatenated as 36 156 aligned amino acid positions with 26% missing data, modelled under CAT + GTR + Γ. Numbers at nodes denote Bayesian posterior probability/bootstrap support as returned by RAxML under the LG [33] substitution mode. Dotted branches at base of phylogeny are shortened for clarity, and outgroups (26 gastropods and bivalves, one scaphopod, four annelids) are collapsed for clarity (figure 1). Scale bar is expected substitutions per site.

suggesting extensive gaps in the fossil record. However, these studies used controversial calibrations from the Late Palaeozoic, such as *Shimanskya* [18] and *Pohlsepia* [19], for which the assignment to the coleoid crown group is dubious [20]. Subsequent studies attempted to estimate cephalopod divergences using calibrations from outgroups, such as bivalves and gastropods and recovered much younger divergence estimates, that were surprisingly congruent, irrespective of differences both in methodology and gene sampling [20,21]. These independent studies recovered a divergence between the nautilids and the coleoids around the Silurian–Devonian boundary, or the earliest Devonian (approx. 415 Ma), which is congruent with unequivocal evidence for fossil stem group coleoids (ammonoids and bactritids) [22,23] and stem group nautilids [24] in the Early Devonian. Cephalopod beaks also appear in the fossil record in the Devonian [25]. These observations suggest that the fossil record documents the origin of the crown group and that the concomitant evolution of the beak [20] coincides with a dramatic shift in predator–prey dynamics, termed the Devonian Nekton Revolution [26]. The jawed vertebrates radiated at this time, incident with a global shift in predatory style towards increased high-metabolism predation and durophagy [27]. The coincidence of jawed vertebrates and beaked cephalopods radiating at the Silurian–Devonian boundary may thus be interpreted as a response to the changes in the predator–prey landscape.

To explore the tempo and mode of coleoid evolution, we assembled a dataset of 180 nuclear genes of consistent rate of molecular evolution, representing crown diversity across Coleoidea. Phylogenetic and molecular divergence time analyses were carried out in a Bayesian framework, applying a molecular evolution model accommodating rate and compositional heterogeneity.

2. Experimental procedures

For full details of experimental procedures, see the electronic supplementary material. We compiled a supermatrix with data from 56 species (electronic supplementary material, table S2) for 180 genes. Phylogeny was inferred from this superalignment using the software package PhyloBayes MPI v. 1.5a [28] under CAT + GTR + Γ. The maximum-likelihood software RAxML MPI v. 8.1.15 [29] was applied to the same dataset as used in Bayesian inference, applying LG + I + Γ.

PhyloBayes 3.3f was used to infer molecular divergence times under the CIR [30] clock model, soft-bounds of 0.05 and a Yule-process birth–death model, with topology fixed to that inferred by PhyloBayes MPI v. 1.5a. A prior was applied to the root of 565 ± 10 Ma, representing the root of lophotrochozoa. Eleven fossil calibration points were applied to the analysis, as shown in table (electronic supplementary material, table S1).

3. Results

Our phylogenetic results confirm *Nautilus* as sister group to coleoids [20,31]. In turn, coleoids comprise two monophyletic groups: Octobrachia (Vampire squid, dumbo octopuses and

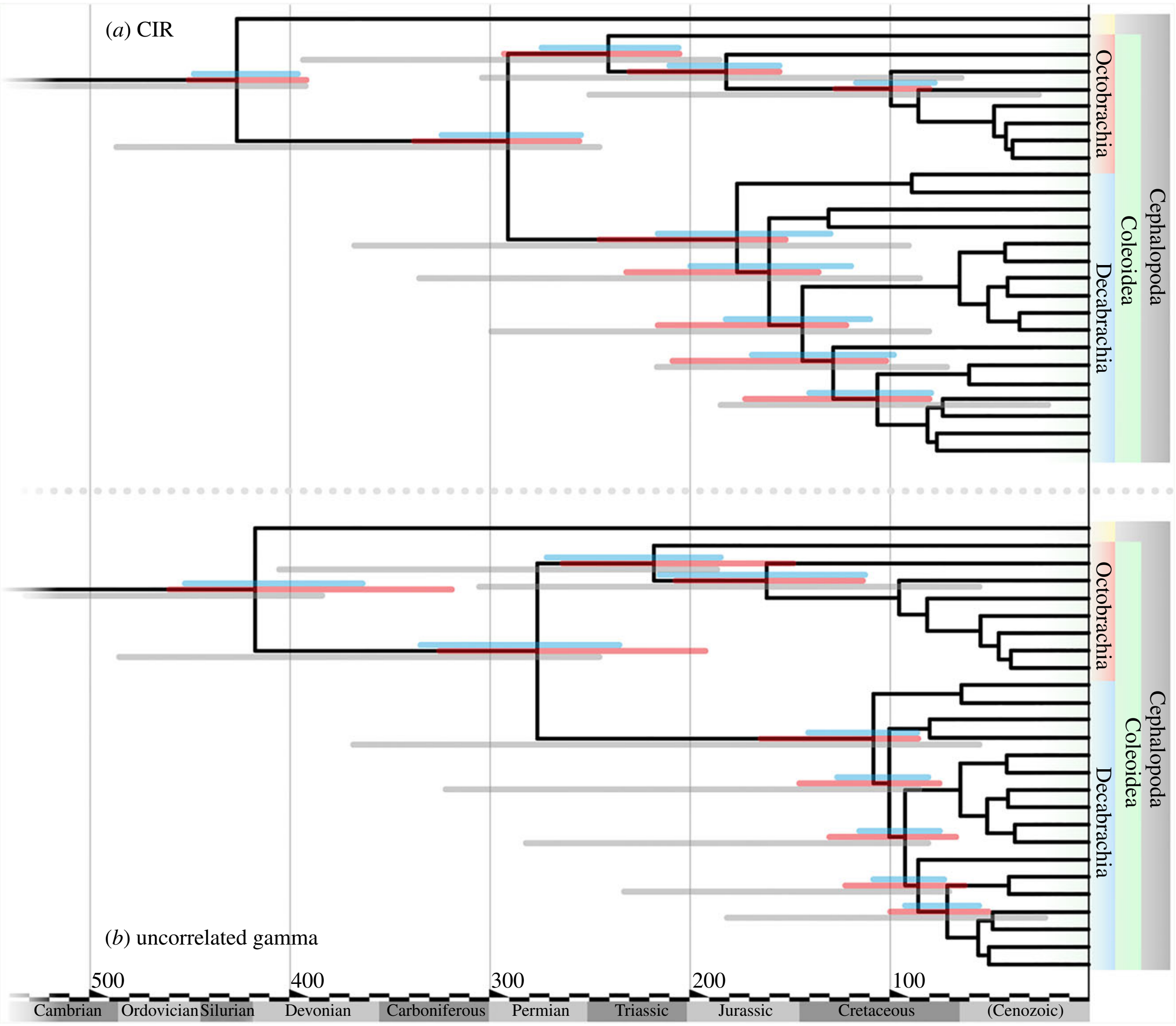


Figure 3. Comparison of molecular clock model and calibration scheme on confidence intervals for node timing inference. (*a*) Applying CIR clock model, (*b*) applying uncorrelated gamma multipier model. Red bars at nodes are confidence intervals with only calibrations external to cephalopods applied. Blue bars are confidence intervals with the full calibration applied. Grey bars are the joint prior distribution at nodes. Not all nodes are labelled to aid clarity, full details in the electronic supplementary material.

incirrate octopuses) and Decabrachia (cuttlefish and squid, including *Spirula*), in agreement with morphology and previous molecular studies [16,17,32] (figure 1). The vampire squid *Vampyroteuthis* and the cirroctopod *Grimpoteuthis* represent cirrate octopuses, branching deep as successive sister groups to the incirrate octopuses (figure 1). Within Decabrachia, we recover a monophyletic Myopsida assemblage, along with support for Teuthoidea with the inclusion of *Spirula*, similar to previous studies [16,20]. However, the relationships between the orders comprising the Sepioidea (Sepiida, Idiosepiidae, Sepiolidae) are recovered as paraphyletic. Oegopsid monophyly is supported, with *Spirula* sister to this clade, in agreement with previous studies [16], but the posterior probability values for many decabrachian basal nodes are generally lower than in other parts of the phylogeny. Sepioid and myopsid relationships have proved difficult to resolve [16], and further phylogenetic work remains to clarify these.

Molecular divergence times were estimated, from the same matrix used for phylogenetic inference, applying an autocorrelated relaxed clock model (CIR process, figures 2 and 3; electronic supplementary material for further details and additional analyses). Alternative treatments, model applications and comparison of the joint priors induced by our calibrations and models and the posterior divergence times supported the data as informative, and resulted in consistency in divergence time inference (figure 3; electronic supplementary material, table S3 and figure S3). Notably, our molecular divergence times are highly congruent with previous molecular divergence estimates [20,34] that used comparable calibration schemes. These studies, however, had insufficient taxonomic spread and sample required for more comprehensive investigation of the evolutionary tempo of coleoids. Furthermore, our wide sample represents crown diversity.

The oldest unequivocal crown group coleoids appear in the latest Triassic, with belemnites representing stem group decabrachians, and phragmoteuthidids (Early Triassic or latest Permian) proposed to represent stem group Octobrachia [35]. Our divergence times suggest that the coleoid crown diverged in the Late Carboniferous or Permian. Fossil consilience is shown by stem group vampire squid (loligosepiids) fossils of the earliest Jurassic (approx. 195 Ma) [13,36]. Octopus-like forms that are lacking the mantle fins and with reduced gladius appear in the latest Cretaceous (Cenomanian, 94–100 Ma) Lagerstätte of Hâkel and Hâdjoula, Lebanon [37]. Our

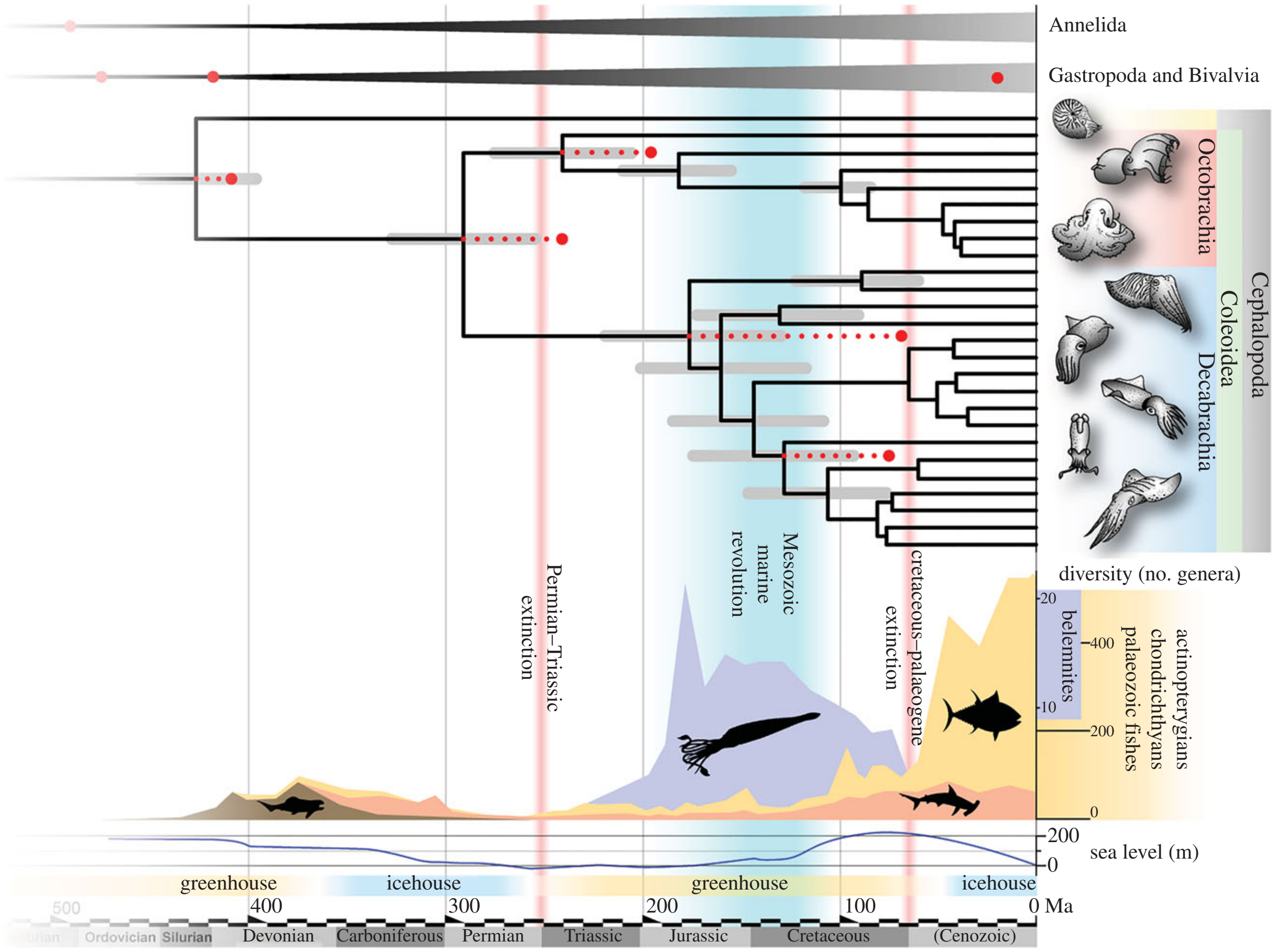


Figure 4. Chronogram of cephalopods, plus 26 bivalve and gastropod molluscs, one scaphopod and four annelids as outgroups and calibration nodes; 36 156 amino acid positions analysed under CAT-GTR substitution model, CIR clock model, Yule birth–death process, soft bound of 0.05, and a root prior of 565 Ma with a standard deviation of ±10 Ma. Bars at nodes represent 95% confidence intervals (recent nodes not labelled with bars to aid clarity). Red dots indicated calibrated nodes (electronic supplementary material, table S1 and figure S3); red dotted lines represent extent of calibration minima. Environmental conditions and sea-level curve simplified from Miller *et al.* [45]. Curves for belemnite, actinopterygian, chondrichthyan and Palaeozoic fish diversity are based on fossil observations on diversity, data from Palaeobiology Database (pbdb.org), electronic supplementary material, table S5. Red vertical lines represent major extinction events. Aqua-blue vertical bar signifies the extent of the Mesozoic Marine Revolution [10].

divergence estimate for the incirrate octopods is in the Late Cretaceous (approx. 100 Ma). Decabrachians have a near non-existent fossil record, except for members of their stem group (e.g. belemnites) and some forms that retain remnants of the phragmocone—*Spirula* and cuttlefish. Stem group spirulids appear in the latest Cretaceous (approx. 66–72 Ma) of West Greenland [38]. Molecular estimates here suggest that spirulids diverged from the Oegopsids at approximately 128 Ma. Sepiid cuttlebones appear in the fossil record in the latest Cretaceous (approx. 75 Ma [37]) and we estimate the sepiids represented in our analysis to have diverged approximately 88 Ma.

4. Discussion

Our molecular divergence estimates show that the coleoid fossil record [13,39] belies not only an earlier origin for key cephalopod groups, but also significant differences in their rate of diversification. Together with the molecular clock estimates for coleoids that are lacking a fossil record, it is possible to investigate events that shaped the diversity of the group. Decabrachians diversify rapidly in the middle Mesozoic (Jurassic), while incirrate octopuses arose in the Cretaceous. Since this time documents an escalation—the evolution of novel predation strategies—it prompts a consideration of what anatomical changes took place in coleoids, particularly decabrachians, at this time.

The iconic shell has had a shifting functional role through cephalopod evolution, and is informative as to lifestyle and ecology. Subsequent to ancestral internalization of the phragmocone through the Carboniferous and Devonian, the decabrachian and octobrachian lineages independently evolved towards shell reduction [13,16], allowing enhanced manoeuvrability and speed [15]. These groups would have been in ecological competition with belemnites: stem group decabrachians [39,40] with an elaborate internal shell, diversifying in the Mid-Jurassic [41]. Our analysis suggests that in the Late Jurassic and at the onset of the Cretaceous, belemnites became marginalized and replaced by modern groups of decabrachians and finned octobrachians (figure 2) [13]. By retaining an elaborate internal phragmocone, belemnites could not compress their mantle cavity for jet propulsion to the same extent as the coleoid forms with a much more reduced internal shell. Similar patterns have been inferred from the Pacific fossil record in Japan [42], suggesting a dramatic turnover in particular approximately 100 Ma (figure 3).

Decabrachian coleoids are nektonic predators with streamlined morphology, high metabolic rates and shoaling behaviour; adaptations in common with teleost fishes [43]. The majority of modern teleost groups radiated during the Jurassic and Cretaceous [44], concomitantly with the origin of most modern coleoids as revealed by our molecular estimates and the fossil record. The scenario in which Mesozoic ecological shifts are exhibited in teleost fishes, chondrichthyans (sharks and rays), and shelled invertebrates as investigated by Vermeij [10] can be extended to cephalopods (figure 4). In the face of high-metabolism, robust predators and niche-competitors, the cephalopods may have responded in kind to these evolutionary pressures. We hypothesize that the cephalopods evolved into the forms we are familiar with today, while shelled groups fell into extinction owing to the shifts in predation in this time period. The Mesozoic Marine Revolution can thus be viewed as the final stage in the shift from Palaeozoic ecologies into the modern structure of marine ecosystems, where (at least in the nektonic realm), agility superseded passive defence.

Ammonoids are stem group coleoids, which were common throughout the Late Palaeozoic until the end of the Mesozoic. Evidence from their radula morphology [23,46] suggests that ammonoids primitively had stout teeth, similar to macrophagous predatory cephalopods. In the Jurassic, the group evolved an enlarged calcareous lower jaw (aptychus) and longer, multicuspidate radula teeth, which has been attributed to a shift into microphagous suspension feeding [23,47]. As such, the group ‘stepped out’ of the arms race and ecological competition with the macrophagous predatory coleoids, fishes and marine reptiles during the Jurassic and Cretaceous. The group evolve increasingly ornamented shells in response to increased predation, as revealed from shell repair scar frequency [48], but eventually became extinct at the end of the Cretaceous.

5. Conclusion

Taken together, molecular divergence times and the cephalopod fossil record are consistent with a scenario in which predator–prey arms races shaped the coleoid body plan, biodiversity and ecology. The coincidence with the evolution of jawed vertebrates and teleost fishes during the Devonian Nekton Revolution and the Mesozoic Marine Revolution, suggests that nektonic marine vertebrates have been key antagonists towards cephalopods throughout most of their evolution.

Data accessibility. All sequence information archived on NCBI, see the electronic supplementary material for accessions. All matrices available on Dryad, doi:10.5061/dryad.180nh [49].

Authors’ contributions. J.V. carried out all RNA extractions and arranged sequencing procedures in Bristol. A.R.T. carried out all bioinformatics analysis, authored the text, and produced figures. J.V. and D.P. contributed to experimental design and authored the text. I.E.W., A.M.R., K.M.K., R.R.d.F., K.M.H., T.H.O., M.T.P.G., F.D. and M.S.P. contributed sequence data and content on the text.

Competing interests. The authors declare no competing interests.

Funding. J.V. and D.P. are supported by NERC BETR grant NE/P013643/1. Sequencing was supported by a start-up funds from the University of Bristol to J.V. A.R.T. was funded by the University of Bristol Scholarship in Teaching and Research programme.

Acknowledgements. A.R.T. thanks Callum Wright of UoB HPC facility, and Luke Parry, Joseph Keating and James Fleming for valuable discussion. We also thank the late Hans G. Hansson for identification of *Bathypolypus arcticus*.

References

1. Mather JA. 2008 Cephalopod consciousness: behavioural evidence. *Conscious. Cogn.* **17**, 37–48. (doi:10.1016/j.concog.2006.11.006)
2. Wells MJ, O’Dor RK. 1991 Jet propulsion and the evolution of the cephalopods. *Bull. Mar. Sci.* **49**, 419–432.
3. Mather JA, Kuba MJ. 2013 The cephalopod specialties: complex nervous system, learning, and cognition 1. *Can. J. Zool.* **91**, 431–449. (doi:10.1139/cjz-2013-0009)
4. Yochelson EL, Flower RH, Webers GF. 1973 The bearing of the new Late Cambrian monoplacophoran genus *Knightoconus* upon the origin of the Cephalopoda. *Lethaia* **6**, 275–309. (doi:10.1111/j.1502-3931.1973.tb01199.x)
5. Vinther J, Sperling EA, Briggs DEG, Peterson KJ. 2012 A molecular palaeobiological hypothesis for the origin of aplacophoran molluscs and their derivation from chiton-like ancestors. *Proc. R. Soc. B* **279**, 1259–1268. (doi:10.1098/rspb.2011.1773)
6. Mutvei H, Zhang Y-B, Dunca E. 2007 Late Cambrian plectronocerid nautiloids and their role in cephalopod evolution. *Palaeontology* **50**, 1327–1333. (doi:10.1111/j.1475-4983.2007.00708.x)
7. Kröger B. 2005 Adaptive evolution in Paleozoic coiled cephalopods. *Paleobiology* **31**, 253–268. (doi:10.1666/0094-8373(2005)031[0253:AEIPCC]2.0.CO;2)
8. Boyle P, Rodhouse P. 2008 *Cephalopods: ecology and fisheries*. New York, NY: Wiley.
9. O’Dor RK, Webber DM. 1986 The constraints on cephalopods: why squid aren’t fish. *Can. J. Zool.* **64**, 1591–1605. (doi:10.1139/z86-241)
10. Vermeij GJ. 1977 The Mesozoic marine revolution: evidence from snails, predators and grazers. *Paleobiology* **3**, 245–258. (doi:10.1017/S0094837300005352)
11. Vermeij GJ. 1987 *Evolution and escalation: an ecological history of life*. Princeton, NJ: Princeton University Press.
12. Strugnell J, Nishiguchi MK. 2007 Molecular phylogeny of coleoid cephalopods (Mollusca: Cephalopoda) inferred from three mitochondrial and six nuclear loci: a comparison of alignment, implied alignment and analysis methods. *J. Mollusc. Stud.* **73**, 399–410. (doi:10.1093/mollus/eym038)
13. Fuchs D, Iba Y, Tischlinger H, Keupp H, Klug C. 2015 The locomotion system of Mesozoic Coleoidea (Cephalopoda) and its phylogenetic significance. *Lethaia* **49**, 433–454. (doi:10.1111/let.12155)
14. Clements T, Colleary C, De Baets K, Vinther J. 2016 Buoyancy mechanisms limit preservation of coleoid cephalopod soft tissues in Mesozoic Lagerstätten. *Palaeontology* **60**, 1–14. (doi:10.1111/pala.12267)
15. Klug C, Schweigert G, Fuchs D, Kruta I, Tischlinger H. 2016 Adaptations to squid-style high-speed swimming in Jurassic belemnitids. *Biol. Lett.* **12**, 20150877. (doi:10.1098/rsbl.2015.0877)
16. Lindgren AR, Pankey MS, Hochberg FG, Oakley TH. 2012 A multi-gene phylogeny of Cephalopoda supports convergent morphological evolution in association with multiple habitat shifts in the marine environment. *BMC Evol. Biol.* **12**, 129. (doi:10.1186/1471-2148-12-129)
17. Strugnell J, Jackson J, Drummond AJ, Cooper A. 2006 Divergence time estimates for major cephalopod groups: evidence from multiple genes. *Cladistics* **22**, 89–96. (doi:10.1111/j.1096-0031.2006.00086.x)
18. Doguzhaeva LA, Mapes RH, Mutvei H. 1999 A Late Carboniferous spirulid coleoid from the Southern Mid-Continent (USA). In *Advancing research on living and fossil cephalopods* (eds F Olóriz, FJ Rodríguez-Tovar), pp. 47–57. New York, NY: Springer.
19. Kluessendorf J, Doyle P. 2000 *Pohlsepia mazonensis*, an early ‘octopus’ from the carboniferous of Illinois, USA. *Paleontology* **43**, 919–926. (doi:10.1111/1475-4983.00155)
20. Kröger B, Vinther J, Fuchs D. 2011 Cephalopod origin and evolution: a congruent picture emerging from fossils, development and molecules. *Bioessays* **33**, 602–613. (doi:10.1002/bies.201100001)
21. Warnke KM, Meyer A, Ebner B, Lieb B. 2011 Assessing divergence time of Spirulida and Sepiida (Cephalopoda) based on hemocyanin sequences. *Mol. Phylogenet. Evol.* **58**, 390–394. (doi:10.1016/j.ympev.2010.11.024)

22. Kröger B, Mapes RH. 2007 On the origin of bactritoids (Cephalopoda). *Paläontol. Z.* **81**, 316–327. (doi:10.1007/BF02990181)
23. Klug C, Korn D, De Baets K, Kruta I, Mapes RH. 2015 *Ammonoid paleobiology: from macroevolution to paleogeography*. Berlin, Germany: Springer.
24. Dzik J, Korn D. 1992 Devonian ancestors of *Nautilus*. *Paläont. Z.* **66**, 81–98. (doi:10.1007/BF02989479)
25. Klug C, Frey L, Korn D, Jattiot R, Rücklin M. 2016 The oldest Gondwanan cephalopod mandibles (Hangenberg Black Shale, Late Devonian) and the mid-Palaeozoic rise of jaws. *Palaeontology* **59**, 611–629. (doi:10.1111/pala.12248)
26. Klug C, Kröger B, Kiessling W, Mullins GL, Servais T, Frýda J, Korn D, Turner S. 2010 The Devonian nekton revolution. *Lethaia* **43**, 465–477. (doi:10.1111/j.1502-3931.2009.00206.x)
27. Bush AM, Bambach RK. 2011 Paleoecologic megatrends in marine Metazoa. *Annu. Rev. Earth Planet. Sci*. **39**, 241–269. (doi:10.1146/annurev-earth-040809-152556)
28. Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013 PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol*. **62**, 611–615. (doi:10.1093/sysbio/syt022)
29. Stamatakis A. 2014 RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313. (doi:10.1093/bioinformatics/btu033)
30. Lepage T, Bryant D, Philippe H, Lartillot N. 2007 A general comparison of relaxed molecular clock models. *Mol. Biol. Evol*. **24**, 2669–2680. (doi:10.1093/molbev/msm193)
31. Kocot KM *et al.* 2011 Phylogenomics reveals deep molluscan relationships. *Nature* **477**, 452–456. (doi:10.1038/nature10382)
32. Lindgren AR. 2010 Molecular inference of phylogenetic relationships among Decapodiformes (Mollusca: Cephalopoda) with special focus on the squid Order Oegopsida. *Mol. Phylogenet. Evol*. **56**, 77–90. (doi:10.1016/j.ympev.2010.03.025)
33. Le SQ, Gascuel O. 2008 An improved general amino acid replacement matrix. *Mol. Biol. Evol*. **25**, 1307–1320. (doi:10.1093/molbev/msn067)
34. Bergmann S, Lieb B, Ruth P, Markl J. 2006 The hemocyanin from a living fossil, the cephalopod *Nautilus pompilius*: protein structure, gene organization, and evolution. *J. Mol. Evol*. **62**, 362–374. (doi:10.1007/s00239-005-0160-x)
35. Fuchs D, Keupp H, Schweigert G. 2013 First record of a complete arm crown of the Early Jurassic coleoid *Loligosepia* (Cephalopoda). *Paläontol. Z.* **87**, 431–435. (doi:10.1007/s12542-013-0182-4)
36. Fuchs D, Weis R. 2008 Taxonomy, morphology and phylogeny of Lower Jurassic loligosepiid coleoids (Cephalopoda). *Neues Jahrb. Geol. Palaeontol. Abhandlungen* **249**, 93–112. (doi:10.1127/0077-7749/2008/0249-0093)
37. Fuchs D, Bracchi G, Weis R. 2009 New octopods (cephalopoda: Coleoidea) from the Late Cretaceous (upper Cenomanian) of Hâkel and Hâdjoula, Lebanon. *Palaeontology* **52**, 65–81. (doi:10.1111/j.1475-4983.2008.00828.x)
38. Fuchs D, Keupp H, Trask P, Tanabe K. 2012 Taxonomy, morphology and phylogeny of Late Cretaceous spirulid coleoids (Cephalopoda) from Greenland and Canada. *Palaeontology* **55**, 285–303. (doi:10.1111/j.1475-4983.2011.01125.x)
39. Schweigert G, Fuchs D. 2012 First record of a true coleoid cephalopod from the Germanic Triassic (Ladinian). *Neues Jahrb. Geol. Palaeontol. Abhandlungen* **266**, 19–30. (doi:10.1127/0077-7749/2012/0258)
40. Iba Y, Sano S-I, Mutterlose J, Kondo Y. 2012 Belemnites originated in the Triassic: a new look at an old group. *Geology* **40**, 911–914. (doi:10.1130/G33402.1)
41. Dera G, Toumoulin A, De Baets K. 2016 Diversity and morphological evolution of Jurassic belemnites from South Germany. *Palaeogeogr. Palaeoclimatol. Palaeoecol*. **457**, 80–97. (doi:10.1016/j.palaeo.2016.05.029)
42. Iba Y, Mutterlose J, Tanabe K, Sano S-I, Misaki A, Terabe K. 2011 Belemnite extinction and the origin of modern cephalopods 35 m.y. prior to the Cretaceous–Paleogene event. *Geology* **39**, 483–486. (doi:10.1130/G31724.1)
43. Packard A. 1972 Cephalopods and fish: the limits of convergence. *Biol. Rev. Camb. Philos. Soc.* **47**, 241–307. (doi:10.1111/j.1469-185X.1972.tb00975.x)
44. Near TJ, Eytan RI, Dornburg A, Kuhn KL, Moore JA, Davis MP, Wainwright PC, Friedman M, Smith WL. 2012 Resolution of ray-finned fish phylogeny and timing of diversification. *Proc. Natl Acad. Sci. USA* **109**, 13 698–13 703. (doi:10.1073/pnas.1206625109)
45. Miller KG *et al.* 2005 The Phanerozoic record of global sea-level change. *Science* **310**, 1293–1298. (doi:10.1126/science.1116412)
46. Kruta I, Landman NH, Mapes R, Pradel A. 2014 New insights into the buccal apparatus of the Goniatitina: palaeobiological and phylogenetic implications. *Lethaia* **47**, 38–48. (doi:10.1111/let.12036)
47. Kruta I, Landman N, Rouget I, Cecca F, Tafforeau P. 2011 The role of ammonites in the Mesozoic marine food web revealed by jaw preservation. *Science* **331**, 70–72. (doi:10.1126/science.1198793)
48. Kerr JP, Kelley PH. 2015 Assessing the influence of escalation during the Mesozoic Marine Revolution: shell breakage and adaptation against enemies in Mesozoic ammonites. *Palaeogeogr. Palaeoclimatol. Palaeoecol*. **440**, 632–646. (doi:10.1016/j.palaeo.2015.08.047)
49. Tanner AR *et al.* 2017 Data from: Molecular clocks indicate turnover and diversification of modern coleoid cephalopods during the Mesozoic Marine Revolution. Dryad Digital Repository. (http://dx.doi.org/10.5061/dryad.180nh)

PROCEEDINGS B

rspb.royalsocietypublishing.org

Invited reply

Cite this article: Puttick MN *et al*. 2017 Parsimony and maximum-likelihood phylogenetic analyses of morphology do not generally integrate uncertainty in inferring evolutionary history: a response to Brown *et al*. *Proc. R. Soc. B* **284**: 20171636.
http://dx.doi.org/10.1098/rspb.2017.1636

Received: 21 July 2017
Accepted: 11 September 2017

Subject Category:
Palaeobiology

Subject Areas:
evolution, palaeontology, taxonomy and systematics

Authors for correspondence:
Davide Pisani
e-mail: davide.pisani@bristol.ac.uk
Philip C. J. Donoghue
e-mail: phil.donoghue@bristol.ac.uk

The accompanying comment can be viewed at http://dx.doi.org/10.1098/rspb.2017.0986.

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3888106.

THE ROYAL SOCIETY
PUBLISHING

Parsimony and maximum-likelihood phylogenetic analyses of morphology do not generally integrate uncertainty in inferring evolutionary history: a response to Brown *et al.*

Mark N. Puttick[1,3], Joseph E. O'Reilly[1], Derek Oakley[1], Alistair R. Tanner[2], James F. Fleming[1], James Clark[1], Lucy Holloway[1], Jesus Lozano-Fernandez[1], Luke A. Parry[1], James E. Tarver[1], Davide Pisani[1,2] and Philip C. J. Donoghue[1]

[1]School of Earth Sciences, and [2]School of Biological Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol BS8 1TQ, UK
[3]Department of Life Sciences, The Natural History Museum, Cromwell Road, South Kensington, London SW7 5BD, UK

MNP, 0000-0002-1011-3442; JEO, 0000-0001-9775-253X; JC, 0000-0003-2896-1631; LH, 0000-0003-1603-2296; JL-F, 0000-0003-3597-1221; LAP, 0000-0002-3910-0346; PCJD, 0000-0003-3116-7463

Our recent study evaluated the performance of parsimony and probabilistic models of phylogenetic inference based on categorical data [1]. We found that a Bayesian implementation of a probabilistic Markov model produced more accurate results than either of the competing parsimony approaches (the main method currently employed), and the maximum-likelihood implementation of the same model. This occurs principally because the results of Bayesian analyses are less resolved (less precise) as a measure of topological uncertainty is intrinsically recovered in this MCMC-based approach and can be used to construct a majority-rule consensus tree that reflects this. Of the three main methods, maximum likelihood performed the worst of all as a single exclusively bifurcating tree is estimated in this framework which does not integrate an intrinsic measure of support.

In their comment on our article, Brown *et al*. [2] argue that our experiments are invalid because we did not employ uncertainty measures after obtaining a maximum-likelihood estimate of the topology. When bootstrapping is employed, a 50% consensus tree constructed from the bootstrap distribution is often indistinguishable from the majority-rule consensus tree constructed from the posterior sample obtained in a Bayesian analysis. This result is not entirely unexpected, as the maximum-likelihood and Bayesian statistical frameworks share many statistical similarities, including a dependence on a likelihood function that incorporates the Mk model in this context. On this basis, Brown and colleagues conclude that they cannot advocate one method of phylogenetic inference over another: Bayesian, maximum-likelihood and parsimony methods differ, and thoughtful consideration is required in order to choose among these methods. Unfortunately, their analyses do not wholly support this conclusion because they exclusively focus on the performance of the two implementations of the same probabilistic model, without considering their performance relative to parsimony. This was a key aspect of our study comparing the primary methods of phylogenetic reconstruction as they are commonly implemented. Our and other previous studies [1,3,4] reject parsimony in favour of a Bayesian MCMC framework in which uncertainty is incorporated, further drawing into question the veracity of Brown and colleagues' assertion that there is equivalent performance among methods.

The principle thrust of the argument presented by Brown *et al*. [2] is that the experiments performed by us [1] did not allow for a fair comparison between phylogenetic methods: the Bayesian implementation intrinsically integrates uncertainty, while it is common practice to evaluate uncertainty *post hoc* for maximum-likelihood and parsimony inferences using bootstrap methodology. In our study [1], we explicitly addressed this issue in two ways. The first argument was that bootstrapping is not an intrinsic aspect of maximum-likelihood estimation or parsimony phylogenetic analysis. Thus, we did not need to consider support values in our analyses. Using Bayesian estimation, it is intractable to analytically estimate topology using the Mk model and so it is necessary to use an MCMC sampling procedure to produce a posterior sample of trees. From this approximation of the posterior distribution, it is straightforward to interpret a 50% majority-rule consensus tree and clade support measures (posterior probabilities), unlike analogous measures produced from bootstrapping [5]. Our second argument was that bootstrapping is arguably unsuited to analysis of morphological data because its statistical expectations are not met, *viz*. that the phylogenetic signal is not independently and identically distributed through the data, which is a view common to phylogenetic textbooks (e.g. [6–8]). Brown *et al*. [2] correctly highlight that this is an issue shared by both Bayesian and maximum-likelihood implementations of the Mk model, as independence is assumed when summing the log-likelihood of individual characters. However, the interpretation of posterior probabilities as the probability of observing a clade given the morphological data is straightforward, whereas the exact meaning of a bootstrap proportion is still equivocal, with numerous proposed interpretations [9], all of which are contingent on the maximum-likelihood estimate of topology.

We agree that bootstrapping has been used commonly in phylogenetic reconstruction, including analyses based on morphological traits, to assign a level of support to the constituent nodes of a most parsimonious or maximum-likelihood topology estimate. In this sense, our experiments could be viewed as failing to faithfully simulate common practice. However, while it is common practice to measure support for the clades through bootstrapping in maximum-likelihood and parsimony phylogenetic analyses of morphological traits, most studies present these support measures on fully resolved topology estimates that include nodes with negligible support, rather than collapsing nodes that exhibit less than 50% support into soft polytomies, as Brown *et al*. suggest [2]. To underline the prevalence of this approach, we reviewed studies citing Lewis [10], the originator of the Mk model, published since the start of this year, as recorded in Web of Science (census date 14 June 2017). Of the 48 citing articles (see the electronic supplementary material), 31 phylogenetic studies were based on morphological traits, in whole or in part. Of the 11 studies that employed maximum likelihood, 10 evaluated bootstrap support, all of which resolved nodes with less than 50% support. The same pattern is seen in parsimony analyses where, among 18 studies, only 12 evaluated bootstrap support, of which eight resolved nodes with less than 50% support—though these nodes were usually supported by other metrics like Bremer support. Resolution of unsupported nodes is less prevalent in Bayesian analyses where, among the 29 studies examined (27 of which presented posterior probabilities), only 12 resolved unsupported nodes; many of these were in maximum clade credibility trees. Unsupported nodes were present in Bayesian trees in only two of the nine studies that employed both maximum-likelihood and Bayesian analyses. Thus, while many of these studies present maximum-likelihood- and parsimony-based trees that are more fully resolved than their support measures should perhaps permit, when they are associated with parallel Bayesian analyses, these are invariably summarized by majority rule consensus.

Hence, the experiments presented in our paper [1] followed common practice, as demonstrated by the literature. Brown *et al*. [2] are correct in their view that measures of support are widely employed in phylogenetics, and poorly supported clades should be collapsed in maximum-likelihood or maximum parsimony topologies. However, most maximum-likelihood- and parsimony-based studies effectively ignore *post hoc* topological support measures in their inferences of evolutionary history, which are most often based on more fully resolved, maximum-likelihood and parsimony trees. Practice shows that the same is not true of Bayesian analyses, which are usually summarized by the majority rule consensus (though some studies also seek further resolution using other methods for summarizing a distribution of trees, such as maximum clade credibility). Therefore, based on current use of phylogenetic models, our support for Bayesian inference is validated based on the current practice used by phylogeneticists.

In effect, Brown *et al*. [2] have not addressed the core questions of our study. Rather, they have extended the experiments we undertook, with a different aim, and they have extended the conclusions. They observe that when clade support is considered, maximum-likelihood and Bayesian implementations of the Mk model perform equally well. This is an important observation that will provide some confidence in maximum-likelihood-based analyses of morphological trait data—just as soon as common practice catches up with the need to control for topological uncertainty when inferring evolutionary history.

Brown *et al*. [2] close out their manuscript without advocating a method of phylogenetic inference and, indeed, argue that there is no superior method. Suitable methods, they argue, should be identified in each instance given the biological question at hand. In so doing, they explicitly draw parsimony back into consideration—despite the fact that their analyses do not address this method. This declaration ignores previous studies that highlight the inaccuracy of parsimony [1,3,4], to which they present no counter-evidence. The focus of our study was an objective comparison of the efficacy of the primary methods of phylogenetic reconstruction, including parsimony, as commonly implemented by practitioners. Our experimental design, focused on such common practices, is valid, as are the results, interpretations and conclusions that we derived from our experiments.

References

1. Puttick MN *et al.* 2017 Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data. *Proc. R. Soc. B* **284**, 20162290. (doi:10.1098/rspb.2016.2290)

2. Brown JW, Parins-Fukuchi C, Stull GW, Vargas OM, Smith SA. 2017 Bayesian and likelihood phylogenetic

reconstructions of morphological traits are not discordant when taking uncertainty into consideration: a comment on Puttick *et al*. *Proc. R. Soc. B* **284**, 20170986. (doi:10.1098/rspb.2017.0986)

3. O'Reilly JE, Puttick MN, Parry LA, Tanner AR, Tarver JE, Fleming J, Pisani D, Donoghue PCJ. 2016 Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. *Biol. Lett.* **12**, 20160081. (doi:10.1098/rsbl.2016.0081)
4. Wright AM, Hillis DM. 2014 Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. *PLoS ONE* **9**, e109210. (doi:10.1371/journal.pone.0109210)
5. Yang Z, Rannala B. 2005 Branch-length prior influences Bayesian posterior probability of phylogeny. *Syst. Biol.* **54**, 455–470. (doi:10.1080/10635150590945313)
6. Felsenstein J. 2004 *Inferring phylogenies*. Sunderland, MA: Sinauer.
7. Kitching IJ, Forey PL, Humphries CJ, Williams DM. 1998 *Cladistics: the theory and practice of parsimony analysis*, 2nd edn. Oxford, UK: Oxford University Press.
8. Schuh RT. 2000 *Biological systematics: principles and applications*. Ithaca, NY: Constock.
9. Yang Z. 2006 *Computational molecular evolution*. Oxford, UK: Ocford University Press.
10. Lewis PO. 2001 A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* **50**, 913–925.

BioEssays
www.bioessays-journal.com

Soft-Bodied Fossils Are Not Simply Rotten Carcasses – Toward a Holistic Understanding of Exceptional Fossil Preservation

Exceptional Fossil Preservation Is Complex and Involves the Interplay of Numerous Biological and Geological Processes

*Luke A. Parry, Fiann Smithwick, Klara K. Nordén, Evan T. Saitta, Jesus Lozano-Fernandez, Alastair R. Tanner, Jean-Bernard Caron, Gregory D. Edgecombe, Derek E. G. Briggs, and Jakob Vinther**

Exceptionally preserved fossils are the product of complex interplays of biological and geological processes including burial, autolysis and microbial decay, authigenic mineralization, diagenesis, metamorphism, and finally weathering and exhumation. Determining which tissues are preserved and how biases affect their preservation pathways is important for interpreting fossils in phylogenetic, ecological, and evolutionary frameworks. Although laboratory decay experiments reveal important aspects of fossilization, applying the results directly to the interpretation of exceptionally preserved fossils may overlook the impact of other key processes that remove or preserve morphological information. Investigations of fossils preserving non-biomineralized tissues suggest that certain structures that are decay resistant (e.g., the notochord) are rarely preserved (even where carbonaceous components survive), and decay-prone structures (e.g., nervous systems) can fossilize, albeit rarely. As we review here, decay resistance is an imperfect indicator of fossilization potential, and a suite of biological and geological processes account for the features preserved in exceptional fossils.

1. Introduction

Most of the species that ever existed are extinct, and the vast majority will never be known as fossils. This is because fossilization, even of organisms with mineralized skeletons, is a rare event and few taxa enter the sedimentary record; likewise few sedimentary sequences survive subduction, or uplift and erosion, to be sampled for fossils.[1] The bulk of the fossil record consists of those parts of organisms that are most resistant to degradation – shells, bones, and teeth. In some cases, shelly fossil remains are so abundant that thick accumulations form entire rock units – chalk, for example, is composed of the calcium carbonate plates of unicellular eukaryotes called coccolithophores. Soft parts, in contrast, are usually lost through scavenging and decay.

L. A. Parry, F. Smithwick, K. K. Nordén, E. T. Saitta, J. Vinther
School of Earth Sciences
University of Bristol
Wills Memorial Building
Queen's Road, Bristol BS8 1RJ, UK
E-mail: jakob.vinther@bristol.ac.uk

L. A. Parry, J.-B. Caron
Royal Ontario Museum
100 Queen's Park
Toronto, ON M5S 2C6, Canada

J.-B. Caron
Departments of Ecology and Evolutionary Biology and Earth Sciences
University of Toronto
Toronto, ON M5S 3B2, Canada

DOI: 10.1002/bies.201700167

L. A. Parry, G. D. Edgecombe
Department of Earth Sciences
The Natural History Museum
Cromwell Road, London SW7 5BD, UK

J. Lozano-Fernandez, A. Tanner, J. Vinther
School of Biological Sciences
University of Bristol
Life Sciences Building
24 Tyndall Avenue, Bristol BS8 1TQ, UK

D. E. G. Briggs
Department of Geology and Geophysics
Yale University
210 Whitney Avenue, New Haven, CT 06511, USA

D. E. G. Briggs
Yale Peabody Museum of Natural History
170 Whitney Avenue, New Haven, CT 06520, USA

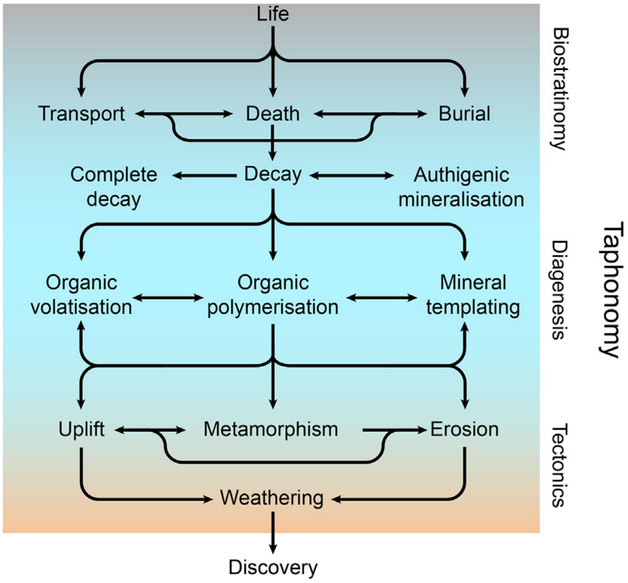


Figure 1. The long journey from live organism to fossil.

In rare cases the soft (i.e., non-biomineralized) parts of animals survive and are fossilized alongside the hard skeleton, and even wholly soft-bodied organisms (those without biomineralized tissues) can be preserved. The journey of these fossils from death to discovery involves a complex interplay of geological and biological processes (**Figure 1**) and although they are rare, they offer unique insights into the anatomy and biology of extinct life (**Figure 2**). Such "exceptional" deposits are commonly referred to as "Konservat-Lagerstätten"[2] – a German term that is now common currency among paleontologists (Lagerstätte is borrowed from the mining industry where it means an ore deposit). Konservat-Lagerstätten occur throughout the geological record in a diversity of paleoenvironmental settings and sedimentary rock types.[3] Soft parts of organisms can be preserved in a variety of ways: as carbonaceous compressions (**Figure** 3A and E); via early (authigenic) mineralization in iron sulfide (pyrite) (Figure 2F and 3B) and apatite (calcium phosphate) (Figure 2C); and by early cementation or entombment, such as in concretions (Figure 3D) or within amber (Figure 2D). Within a single specimen, a combination of these preservational pathways can account for the preservation of the whole organism and different tissues follow particular preservational pathways. For example, Figure 3E shows scanning electron microscope energy dispersive x-ray (SEM-EDX) maps of a specimen of *Marrella splendens* from the Cambrian Burgess Shale of British Columbia, which preserves certain anatomical features as carbon films, pyrite, or calcium phosphate.

The Burgess Shale is one of a number of well-known examples of exceptional preservation (Figure 3A and E) which reveal diverse assemblages of early animals.[4,5] Other examples of exceptionally preserved biotas include the plants and animals found in the Carboniferous Mazon Creek concretions of Illinois (Figure 3D),[6] the fishes that preserve phosphatized subcellular details of muscle tissue in the Cretaceous Santana Formation concretions from Brazil,[7] and the feathered dinosaurs that reveal evidence of plumage color and flight capability from the Cretaceous Jehol sequences of north-eastern China (Figure 2A).[8]

Despite the diversity of settings that yield exceptionally preserved fossils, many Konservat-Lagerstätten share biological and geological processes such as rapid burial, limited or no bioturbation, decay suppression through anoxia or euxinia, and sealing of sedimentary laminae by microbial mats and early diagenetic cements (Figure 1). These factors contribute to the survival of organic macromolecules[9,10] and create the necessary microenvironments for the replication of soft tissues through authigenesis, the early precipitation of minerals.[11] Understanding preservation (the field of taphonomy) is critical to interpreting the morphology of fossils and, in turn, their place in the tree of life and consequent significance for organismal evolution. A first step is determining which characters were originally present and which have been lost or modified by taphonomic processes.[12] A second step involves recognizing possible homologies between features of the fossil organism and those of living taxa.[12] The identification of homologies is essential for determining the affinity of fossils, but it is particularly challenging in cases where there is no obvious close living relative.

Rather than representing perfect snapshots of extinct organisms, soft-bodied fossils have passed through numerous filters prior to discovery that remove, modify, or preserve anatomical characters (Figure 1). Such processes include autolysis (self-digestion through enzymes) and microbial decay, precipitation of authigenic minerals, diagenesis (plus metamorphism in some cases), and finally weathering (Figure 1). The pathways travelled by exceptional fossils prior to discovery are complex, and understanding preservation is an active field of research based on investigations of fossil specimens and taphonomic experiments on extant organisms.[13] Following discovery, further biological information can be lost or modified during excavation and preparation of a fossil; the method used to remove surrounding matrix may create artifacts and should be taken into account when analyzing important features.[14]

A key hurdle to interpreting fossils correctly is determining which characters are missing because they were originally absent in vivo and which characters have failed to survive all of the processes involved in fossilization. Decay experiments have played a central role in interpretations of soft-bodied fossils for many years, illuminating the relative preservation potential and likely identity of different soft parts in fossils, determining the conditions required for the replication of tissues in authigenic minerals, and documenting how the molecular components of an organism are impacted by decay.[15] More recently, however, there has emerged a tendency to apply the results of decay experiments more literally to the interpretation of soft-bodied fossils, using the relative susceptibility of morphological characters to decay as a measure of whether or not they could be preserved at all.[16–18] While an experimental approach is important to determining how exceptional fossils are formed[19] microbial decay is just one of many processes that can distort the original morphology of an organism. A variety of interlinked processes play a role in the preservation of different anatomical features.

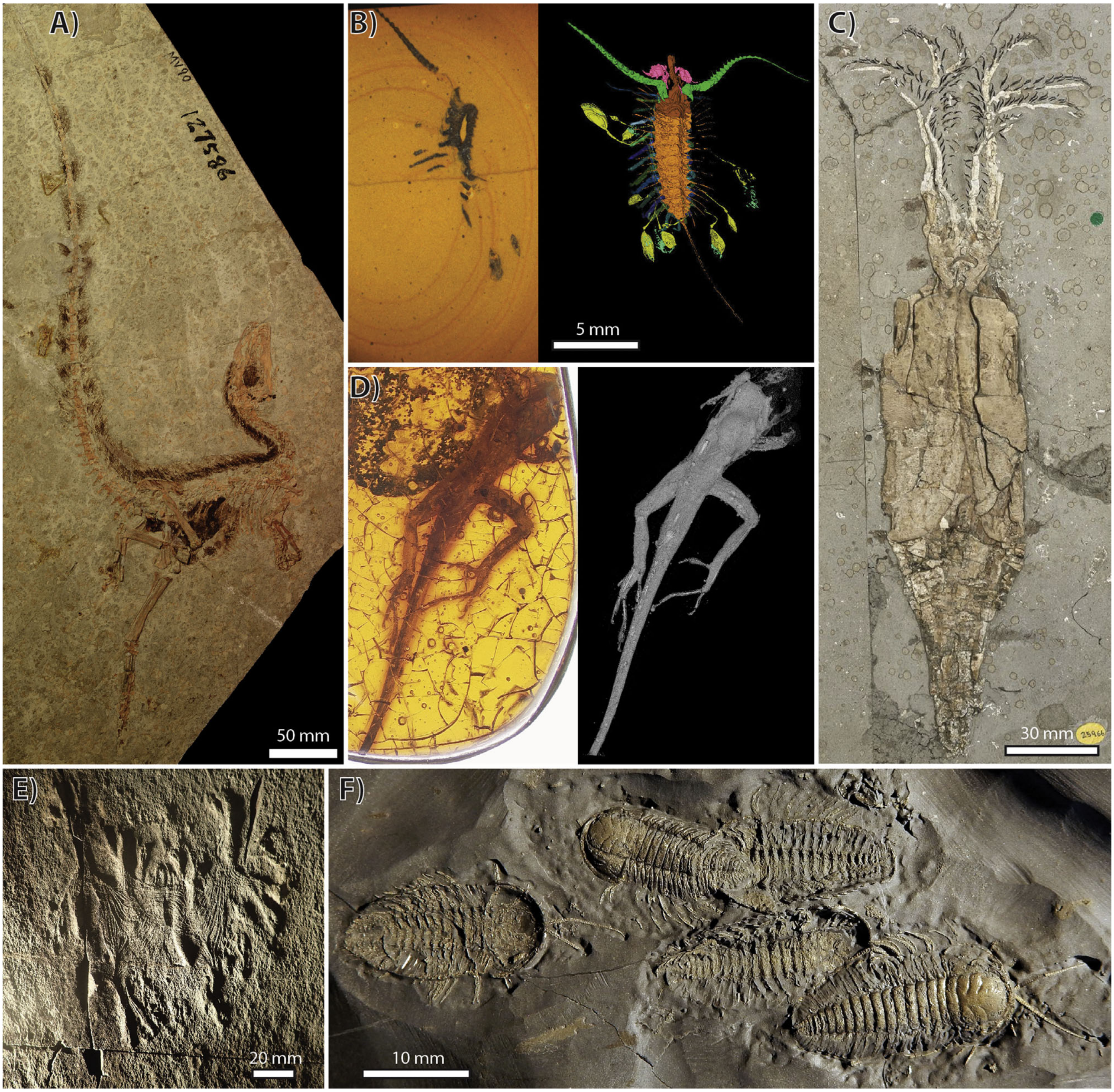


Figure 2. Exceptionally preserved fossils. A) *Sinosauropteryx prima* (NIGP 127586), a feathered dinosaur from the Cretaceous Jehol Biota preserving melanized tissues (feathers, eyes, and abdominal organs). B) *Aquilonifer spinosus* (OUMNH C.29695), a Silurian arthropod preserved in three dimensions in volcanic ash-hosted carbonate concretions from Herefordshire. Image at left shows a surface captured during serial grinding, image at right shows a three dimensional reconstruction from serial photographs.[112] C) *Belemnotheutis antiquus* (NHMUK 25966), a Jurassic stem group decabrachian (belemnoid) from Christian Malford, Wiltshire, UK, preserving creamy colored musculature replaced by calcium phosphate and organic arm hooks. D) Fossil *Anolis* lizard preserved in Miocene Dominican amber.[113] Image at left is a photograph of specimen, image at right shows 3D reconstruction using micro CT. E) *Haootia quadriformis,* a possible medusozoan from the Ediacaran of Newfoundland. F) Pyritised specimens of the trilobite *Triarthrus eatoni* (ROM 62891), with preserved limbs from the Late Ordovician Beecher's Trilobite Bed, New York, State. Image credits to the authors, except C (Jonathan Jackson, NHM) D (Russell Garwood and Emma Sherratt) E (Alex Liu), F (David Rudkin).

Cambrian fossils from Burgess Shale-type localities have featured most prominently in discussions of how decay determines the information preserved in exceptional fossils, because many Cambrian animals are difficult to place with confidence in a phylogeny with modern groups. The phylogenetic position of early chordate-like fossils, for example, has attracted particular attention following the proposal of "stemward slippage."[20] As chordates decay, characters are lost in the opposite order to their stepwise acquisition during the evolutionary transition from the chordate stem lineage to the

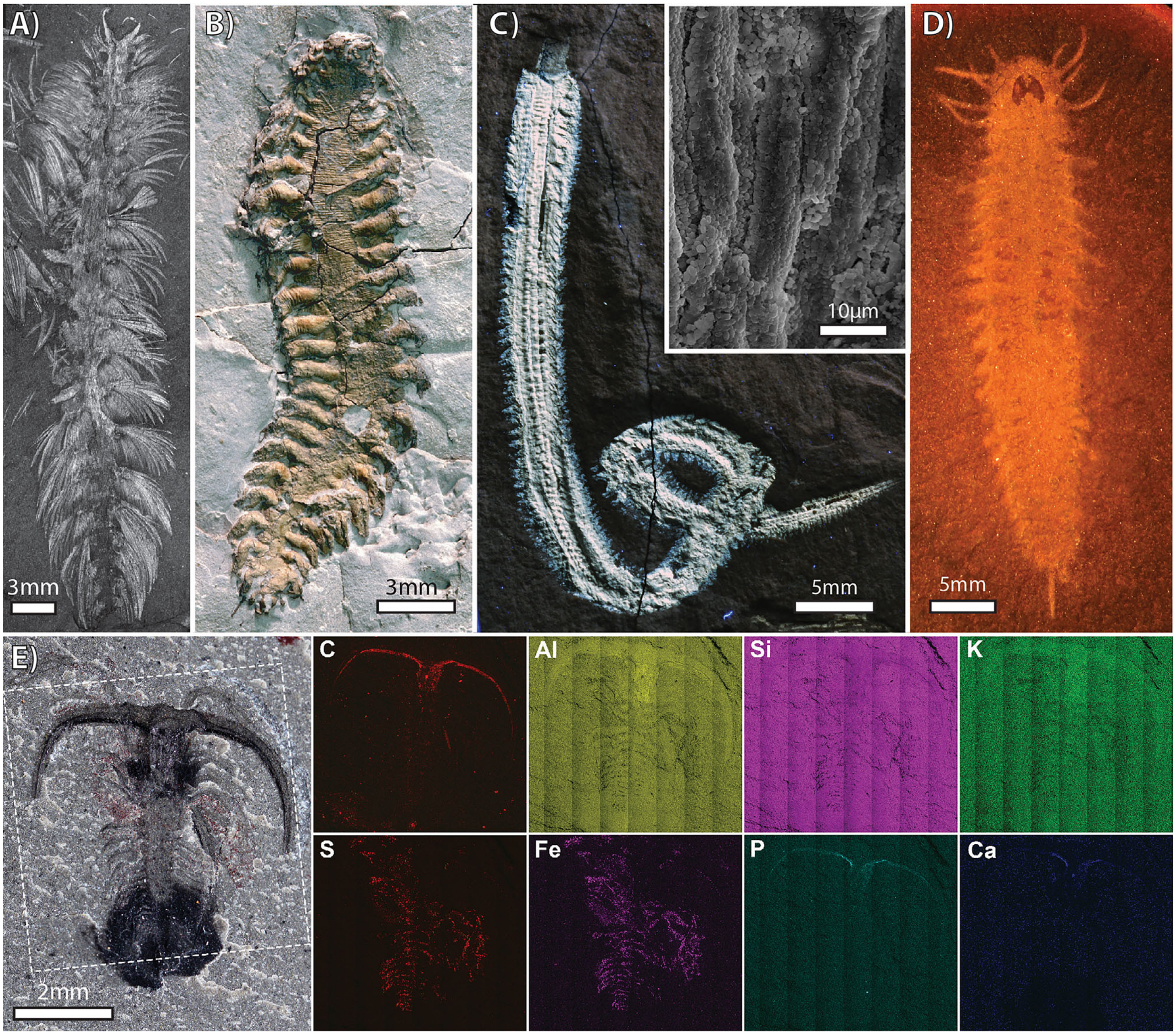


Figure 3. Same organism, different pathways of preservation. A–D show epibenthic polychaete worms preserved through different key preservational pathways. A) Preservation as a carbonaceous compression, *Canadia spinosa* (USNM 83929c), Middle Cambrian Burgess Shale of British Columbia. B) Three dimensional preservation in pyrite, *Arkonips topororum* (UMMP 73795), Devonian of Ontario. C) Three dimensional preservation of mainly muscle tissue in calcium phosphate, *Rollinschaeta myoplena* (AN 15078), Late Cretaceous, Lebanon. Inset image shows SEM photomicrograph of preserved muscle fibres. D) Entombment in an ironstone concretion, *Fossundecima konecniorum* (ROMIP 47990), Mazon Creek, Late Carboniferous, Illinois. E) Tissue specificity of taphonomic pathways in *Marrella splendens* from the Burgess Shale. Image at left shows photograph of specimen ROMIP60748. Images at right show SED-EDX elemental maps of region encompassed by the white box in the photograph where the intensity of the color indicates elemental abundance. Structures preserved as carbon films are highlighted in the C map, structures preserved by clay minerals are highlighted in the Al, Si, and K maps, pyritized structures are highlighted in the Fe and S maps and structures preserved as apatite (calcium phosphate) are highlighted in the Ca and P maps.

vertebrate crown: the farther decay progresses the more “primitive” the resultant fossil supposedly appears. Reports of organically preserved neural and circulatory[21] tissues in Cambrian panarthropods have proved particularly controversial as an interpretation based on stages of decay[16–18] implies that such decay-prone features should not persist and fossilize.

Here we review the diversity of processes that occur during fossilization and identify circumstances where the sequence of character loss and modification in fossils may deviate from the null model provided by the decay of related extant animals in seawater.[22] Clearly it is important to avoid overinterpretation of features in a soft-bodied fossil based on a simplistic comparison with the anatomy of its nearest living relative, but equally, the evidence of the fossils themselves should not be dismissed without good cause. In some cases, features that are decay-resistant do not survive diagenesis, while others that are decay-prone preserve readily. Such considerations challenge the assumption that the relative decay resistance of morphological

Box 1

The late Ediacaran (~580–541 Ma) is a unique time in earth history, predating the major radiation of the animal phyla in the Cambrian, when assemblages of macroscopic, soft-bodied organisms were preserved as high relief casts and molds, sometimes with hundreds of individuals on a bedding plane.[29,115,116] Although most common in the Ediacaran, this taphonomic window persisted until the Devonian.[29] Such fossils occur in a range of depositional environments, including deep marine basins, marginal marine settings, storm influenced shore faces, and shelf carbonates.[116] Specimens may retain sub-millimetric details of mostly external, but sometimes internal,[117] anatomy, and are sometimes three dimensionally preserved within beds.[118] These Ediacaran organisms were buried rapidly in event beds, either by storm deposits, turbidites, volcaniclastic events, or ash falls, depending on locality.[116] Ediacaran deposits were interpreted as census "snapshots,"[119] but it is now recognized that they can include partially decayed individuals that died prior to the event that smothered the sea floor.[23] The preservation of abundant in situ carcasses reflects limited or absent macrophagous scavenging during the Ediacaran.[23] Although the mechanism that led to the preservation of these organisms remains controversial, and a single explanation may not apply to all localities, most models involve sealing the sediment. Candidates include rapidly forming pyrite crusts referred to as "death masks,"[28,116] microbial mats,[28] clay mineral templating[120] and, most recently, early silicate cementation.[29]

characters alone can be used to interpret the morphology of soft-bodied fossils.[16]

2. The Advantages of Being Buried Alive

In order to survive the test of time, organismal remains need to be shielded from the natural processes that degrade them. Burial is common to nearly all fossils, although remains may survive on a geologically short timescale in caves or bogs, for example. The impact of burial depends on factors such as rate and type of sedimentation, availability of oxygen, and subsequent cementation and compaction. Deep burial by a single event, such as a storm, enhances the chances of exceptional preservation particularly where low levels of oxygen inhibit scavenging and destruction by macro- and micro-organisms. Carcasses typically survive on the seabed only where scavengers are absent, as in the famous Ediacaran biotas,[23] which predate the major radiation of scavenging and macrophagous animals in the Cambrian (Box 1). Rapid burial creates a microenvironment around a carcass where bacterial activity rapidly consumes available oxygen. The anaerobic processes that follow may generate conditions that favor the precipitation of authigenic minerals.[24,25] Anaerobic conditions also protect organic substances from oxidation, and reactive substances, such as hydrogen sulfide, may be generated which can stabilize organic materials further (see below). Generally, the more fine grained the sediment the better the preservation of soft tissues because clay and silt limit the rate of diffusion and promote the establishment of chemical gradients around a carcass.[26,27] Such gradients also form where a microbial mat and early diagenetic cement seal in the buried organism (**Box 1**): this may allow preservation in coarser sediment – even in sandstones in the case of Ediacaran assemblages.[28,29] Sediment mineralogy, particularly of clays, may also play a role in tissue stabilization.[13,26,30,31]

Early cementation of the surrounding sediment promotes exceptional preservation by eliminating pore space and may create a cast of soft tissue anatomy. Early precipitation of carbonate at the sediment surface[32] or the presence of microbial mats[33] may have promoted preservation in Burgess Shale-type deposits, for example, and microbial mats are a common feature of deposits preserving muscle tissue.[34] In other cases, a concretion may form around a carcass, preventing collapse and promoting mineralization. The three-dimensional fossils of the Silurian Herefordshire Konservat-Lagerstätte, for example, preserve remarkable details in carbonate nodules within a volcanic ash (bentonite) which was deposited on the seafloor.[35] Silica precipitates as chert in other settings, providing a medium for preserving carbonaceous fossils: notable examples include early prokaryotes and eukaryotes of Precambrian age,[36] and one of the oldest terrestrial freshwater ecosystems associated with a hot siliceous spring in the Devonian Rhynie Chert of Scotland.[37]

Flattening during and following burial is not equivalent to the squashing that characterizes road-kill, although fossils are often said to look like one. Fossils collapse as a result of decay but their outline is maintained by the confining sediment – lateral expansion due to pressure from above is not the norm. Even highly compacted vertebrate fossils which preserve soft tissue outlines show little evidence of lateral expansion.[27,38] Flattening a fossil on a bedding plane is more like projecting a three-dimensional object onto two dimensions, as in a photograph.[39] Specimens of the same animal buried in different orientations, such as the fossils from the Cambrian Burgess Shale (which were transported in turbulent flows), can be used to inform a three-dimensional reconstruction.[39]

3. Decay Experiments in Sea Water Show That Information Loss is the Norm

Although fossilized muscle tissue was first recognized in a Jurassic coleoid cephalopod over 170 years ago,[19] systematic investigation of the role of decay in the preservation of exceptional fossils has only been a major topic of research in the last few decades (for a summary of decay experiments in the literature, see Supporting Information). Earlier studies involved observations on vertebrates in natural or laboratory conditions, with little control on variables, and often took advantage of natural deaths in marine settings.[40,41] One focus was the effect of a decaying organism on the surrounding micro-environment, as in concretion formation.[42] Observations of a decaying priapulid were used to interpret Burgess Shale specimens of the Cambrian priapulid *Ottoia*[43] but it was not until the late 1980s that experiments started to explore the impact of various controls on decay.[44,45] These early laboratory experiments showed that decay can proceed rapidly even under anoxic conditions, leading to the realization that authigenic mineralization is necessary to retain the morphology of certain decay-prone soft tissues[45] (see Table S1, Supporting Information).

A series of decay experiments carried out in the 1990s attempted to monitor and control the complex variables involved, as well as exploring the impact of different experimental conditions on morphological decay.[46–52] Annelids and arthropods decaying under different conditions of oxygen and temperature, for example, showed consistent patterns of morphological decay, reflecting the nature of their tissues.[49–51,53] Interpretations of soft-bodied fossils were informed by which features were more likely to survive decay versus those that degraded rapidly.[46,54] Observations of decay of the lancelet *Branchiostoma lanceolatum*, for example, were used to argue that the axial lines preserved along the trunk of conodonts represent the notochord, and that the apparent offset position of the conodont elements below the head reflects the decay of the supporting tissue.[51] The same decay experiments allowed the chevron-shaped structures in *Conopiscius*, a Carboniferous chordate, to be interpreted as myomeres rather than external scales, and also indicated that a decay-resistant cuticle was not necessarily present in *Pikaia* from the Burgess Shale.[51,55]

Decay in seawater has now been monitored in a range of taxa in laboratory experiments (see Table S1, Supporting Information): anthozoans,[56] annelids,[48] chaetognaths,[57] priapulids,[18] onychophorans,[17] pterobranchs,[58] enteropneusts,[59] non-vertebrate chordates,[20] and cyclostomes.[60] Thus the sequence of character loss has been determined for taxa representing most clades of eumetazoans. Despite the diversity of body plans analyzed in these experiments, collectively they show that different tissues decay at different rates, with some common patterns of susceptibility to decay across different organisms, and that different character systems are lost at different stages in the decay process. Gut, muscle, and nervous tissue, for example, are among the first to decay in a broad range of taxa in decay experiments.[17,18,48]

The majority of recent experiments were carried out in the absence of sediment in order to facilitate observations of the sequence of decay stages and to reduce the number of variables involved in the experiments. The sedimentary environment in which a carcass is buried is an important control on decay. The chemical gradients that form may stabilize organic substances or induce mineral precipitation, and the sediment supports decaying tissues and prevents the organism from disarticulating. Decay experiments that incorporate sediment reveal a role for sediment chemistry in soft-tissue preservation, where different clays, for example, may promote the preservation of some tissues but not others.[30]

During decay experiments, certain structures persist for weeks or even months. Notable examples are the jaws and chaetae of nereid polychaetes,[48] the notochord and myomeres of chordates,[20] and the chitinous parts of non-arthropod ecdysozoans such as the claws of onychophorans[17] and scalids of priapulids.[18] Despite the apparent decay resistance of these structures, however, they are not always preserved in fossils. The jaws of nereid polychaetes, for example, do not survive diagenesis despite being heavily sclerotized: they only survive in recent sediments,[61] whereas the jaws of other polychaetes occur abundantly as fossils.[62] Somewhat counterintuitively, polychaetes that mineralize their jaws are absent or rare as fossils as they are more weakly sclerotized, allowing their mineral components to disaggregate.[61] Similarly, the notochord is absent in fossils of some members of the vertebrate crown group[63] despite its decay resistance.

4. The Molecular Composition of Tissues and Their Decay Environment Influence Preservation Potential

Structural tissues, such as the exoskeleton of arthropods and the non-biomineralized jaws of polychaetes, are often fossilized even though when, unlike shells, they do not contain biominerals. Fossils of structural tissues encompass a broad range of taxa from across the tree of life, ranging from the cuticles of plants to the plethora of early Paleozoic "small carbonaceous fossils," which reveal a hidden diversity of early animals, including meiofauna.[64] These carbonaceous fossils are composed of recalcitrant biomolecules, i.e., their molecular composition protects them from decaying or breaking down rapidly and allows them to survive elevated temperatures and pressures. The collagen in notochords and the keratin in claws, feathers, and hair are decay-resistant, but do not survive geological maturation.[65] In some cases, biomolecules may remain as biomarkers in the rock when all morphology is lost.[10] Bond strengths, functional groups, and steric effects influence the susceptibility of different biomolecules to degradation.[9] Nucleic acids are the least stable, followed by proteins, carbohydrates, lipids, pigments, and structural macromolecules.[9,10] Under certain conditions, it is possible to recover more resistant biomolecules associated with fossils in a nearly intact state. Recently, for example, sterols have been reported in a 380 million-year-old Devonian crustacean preserved in a concretion[66] and nearly intact melanin in a 200 million-year-old coleoid cephalopod.[67] But, just as decay resistance is an incomplete guide to the preservation potential of soft tissues, in most cases carbonaceous material must undergo diagenetic modification to survive.[68]

Labile molecules may be stabilized by reactions that occur during fossilization, including processes equivalent to tanning, caramelization, and sulfurization (vulcanization). Tannins are polyphenolic compounds with multiple hydroxyl and carboxyl groups that react with proteins and their constituent amino acids in a process similar to tanning, as in the leather industry. Tanning was invoked as an explanation of the survival of polychaete and shrimp carcasses in experiments with clays.[30] Caramelization, well known in cooking, involves anhydrous reactions between sugars and amino acids in Maillard-type condensations to form melanoidin compounds. Melanoidins have been reported in fossil molluscs and brachiopods[69,70] and are important in the formation of humic acids and kerogens.[71] The reaction of proteins with saccharides to form melanoidin complexes may also explain the preservation of skin in human bog bodies.[72]

Sulfurized molecules are a significant component of kerogens and asphaltenes.[73] Sulfurization involves the formation of sulfide and disulfide bridges in a manner reminiscent of the vulcanization of rubber. The preservation of bone marrow and muscles in amphibians from Miocene sulfur-rich lake deposits in Spain has been attributed to this process.[74–76] Analyses of older fossils, complemented by maturation experiments, have shown that over time the composition of animal and plant cuticles, for example, is transformed by cross-linking reactions into more stable longer chain hydrocarbons (in situ polymerization), which incorporates lipids,[77] a process enhanced in the presence of sulfide. This diagenetic change is time dependent, but accelerated by the elevated temperatures experienced by rocks at depth, and although it modifies the original chemical composition and internal structure of tissues, their external morphology remains largely intact.[10]

We have a general understanding of the chemical processes involved in the fossilization of soft tissues, but the details of how preservation is affected by the composition of specific tissues and the nature of the microenvironments that develop within a buried carcass are largely unknown. Such an understanding is hampered by the need to deconstruct the extensive chemical alteration that fossilized soft tissues have undergone in order to determine the processes involved. It has been clear for some time, however, that the resistance of molecular components to microbial degradation (selective preservation) is an inadequate explanation of the survival of organic matter in sedimentary rocks and, consequently, of the fossilization of soft tissues.[68,77]

5. Authigenic Mineralization Saves Tissues Apparently Doomed to Decay

Authigenic mineralization provides a mechanism for fossilizing decay-prone tissues before they are lost. The key pathways are (1) phosphatization, which preserve soft tissues at high fidelity, (2) pyritization, which retains less fine detail but played a critical role in a number of famous fossil Konservat-Lagerstätten, and (3) templating by clay minerals.

Features known to be preserved through phosphatization include microbes,[78] cells and embryos with possible nuclei,[79] guts,[80] epidermis,[75] and muscles.[46,78,81,82] Experiments have revealed the importance of microbial activity in releasing phosphate and generating the necessary geochemical gradients to induce phosphatization in a decaying carcass. Sufficient calcium and phosphate ions must be available and pH must drop in order for calcium phosphate to precipitate instead of calcium carbonate (i.e., the calcium carbonate/phosphate switch).[24] Such a decrease is a normal result of bacterial decay,[26,31,48] but phosphatization tends to favor the preservation of particular tissues and taxa.[81,83]

Decay experiments have shown that phosphatization occurs on a laboratory time scale and is not necessarily restricted to a few unusual settings.[53] Microbial activity promotes decay, destroying morphological information in soft tissues, but it is also essential to establishing the conditions that lead to the replication of soft tissues in authigenic minerals.[11,84–86] The nature of microbial controls is subtle and poorly understood. For example, different species of the same genus of bacteria have been shown to degrade soft tissue on the one hand and replicate cellular organization and morphology on the other, providing a potential pathway for mineral replication of soft tissue features.[19,87]

Authigenic mineralization varies with conditions and between taxa. The fidelity of preservation differs in different muscle tissue types, for example,[81] mineralization of soft tissue is rare or absent in some taxa even where they occur in association with others that are heavily phosphatized,[81] and some taxonomic groups are not represented in the fossil record due to taxon-specific effects during decay.[88] The longitudinal and parapodial muscles of the Cretaceous amphinomid polychaete *Rollinschaeta myoplena* (Figure 3C) are preserved with greater fidelity than other muscle groups although muscle tissue is rarely preserved in associated polychaetes, and only with low fidelity.[81] These differences may reflect specific properties of amphinomid muscle, such as greater density or availability of phosphate compared to other polychaetes. Circular muscle may be preserved with less fidelity than other muscle types, based on the evidence in fossil annelids,[81] or the presence of these muscles may be uncertain due to poor preservation, as in the gilled lobopod *Pambdelurion* from Sirius Passet.[89] The absence of phosphatized soft tissue in fossil decabrachian cephalopods has been shown experimentally to be due to the presence of ammonia for buoyancy regulation, which prevents the drop in pH necessary to allow phosphatization.[88] An understanding of the controls on phosphatization is therefore important for constraining interpretations of authigenically mineralized soft tissues.

Authigenic mineralization can upend the sequence of character loss observed in decay experiments. In polychaetes, for example, the cuticle and chaetae persist in decay experiments for many weeks,[48] while muscle tissue and digestive organs are readily lost. In contrast, fossil polychaetes show that complete myoanatomy may survive when conditions favor extensive phosphatization[82] while decay-resistant cuticular features such as chaetae may be absent or poorly preserved.[81] In extreme cases, characters that decay rapidly are preserved to the exclusion of characters that undergo little degradation on a laboratory timescale.[81]

Pyritization, like phosphatization, although relatively rare, can preserve the original three-dimensional morphology of structures that normally decay. Examples include the appendages and eggs of trilobites and ostracods in Beecher's Trilobite Bed in the Ordovician of New York State (Figure 2F),[25,90] the soft parts of a diversity of marine animals in the Devonian Hunsrück Slate of Germany[91] and of the polychaete *Arkonips* from the Devonian of Ontario (Figure 3B).[92] Pyritization of soft tissues occurs in fine-grained siliciclastic sediments that are otherwise poor in organic matter but enriched in iron.[91] In such settings, decaying carcasses provide a locus for anaerobic sulfate reduction, resulting in the production of sulfide and formation of pyrite.[91,93] Iron-enriched pore water is a prerequisite for pyritization, and may explain why pyrite framboids commonly occur in association with soft-bodied fossils from the Cambrian Chengjiang biota but are rare in similar Burgess Shale-type assemblages elsewhere in the world.[93,94]

Templating by clay minerals has also been invoked as a tissue specific mineralization process responsible for preservation of organisms in the Burgess Shale.[95] Such clay mineral templates are common in organic walled fossils (such as graptolites) in

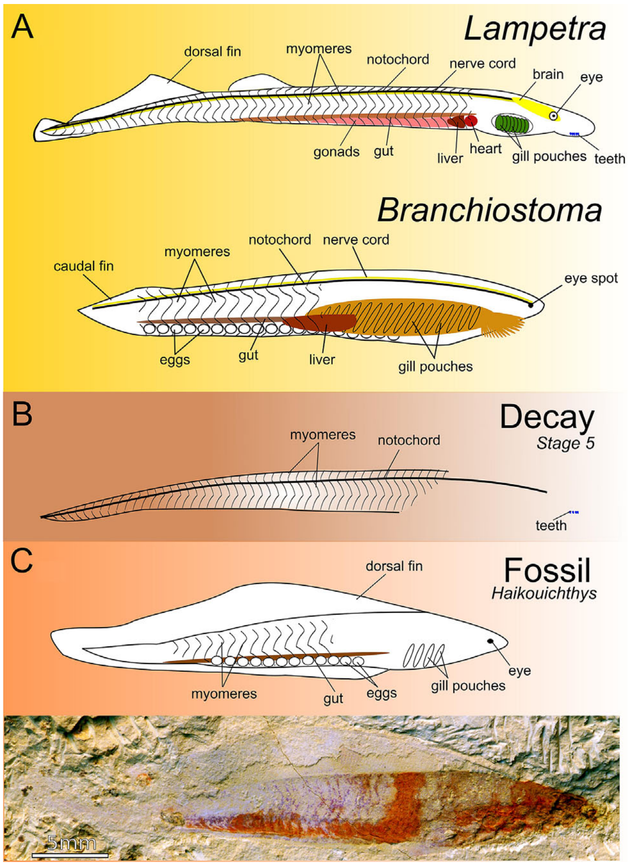


Figure 4. Characters resistant to experimental decay do not closely match characters preserved in fossils. Instead, fossils preserve a combination of decay-resistant and decay-prone characters. A) Schematic anatomy of extant lamprey (*Lampetra*) (top) and lancelet (*Branchiostoma*) (bottom). B) Reconstruction of lamprey in an advanced state of decay (decay stage 5, sensu Sansom et al.[20]). C) Drawing (top; after Zhang and Hou[114]) and photograph (bottom) of an exceptionally preserved fossil chordate, *Haikouichthys* Yunnan Key Laboratory of Palaeontology YKLP00195. Photograph by Peiyun Cong, Yunnan University and NHM, London.

metamorphosed fine grained siliciclastic sediments[96] and a broad survey of Burgess Shale-type localities suggest that conservation of organic tissues is the primary mode of preservation that unites these Lagerstätten.[94] Clay minerals have long been implicated in having a role in the processes that suppressed the breakdown of tissues in Burgess Shale type localities[97] and experimental evidence suggests that clay mineralogy has a profound impact on tissue decay.[13]

6. Using Decay as a Guide to Preservation Can Compromise the Interpretation of Fossils

An overemphasis on the sequence of decay observed in experiments in interpreting soft-bodied fossils assumes that the anatomy preserved is a reflection of original morphology tempered by decay loss (the "rotting away" of characters).[16–18,57] Decay experiments on a diversity of taxa (Table S1, Supporting Information) have shown that "stemward slippage"[20] appears to be a peculiarity of chordates. This is perhaps not surprising as there is no a priori reason why derived characters should be more or less decay-prone than others – in arthropods, for example, morphological characters sheathed in cuticle have a high preservation potential, and cuticular characters are subject to evolutionary change at all levels in the systematic hierarchy of Arthropoda.

A too-literal interpretation of fossils as representing a stage of decay in the laboratory risks ignoring other factors that affect the loss or preservation of morphological features. Although we need to be careful not to overinterpret the anatomy of soft-bodied fossils, we cannot assume that because features decay rapidly in experiments, they can never be fossilized, particularly if the fossil evidence itself is compelling. Animals that lack an extracellular cuticle, such as the soft-bodied mollusc *Odontogriphus*,[98] enteropneusts[59] and the chordate *Pikaia*[55] are preserved in the Burgess Shale, and chaetognaths are preserved in both the Burgess Shale[99] and Chengjiang biotas.[100] Although the body outlines of fossil chaetognaths are poorly defined,[100] those of *Odontogriphus*, *Spartobranchus*, *Oesia*, and *Pikaia* are clearly preserved, indicating that structures that lack the extracellular materials in cuticles nonetheless survive in Burgess Shale-type deposits (*contra*[101]). Other decay-prone characters, such as features of the digestive system, are preserved as reflective films (representing carbon) in both Sirius Passet and Burgess Shale fossils. The identification of features of the digestive system is relatively straightforward based on their position and morphology (e.g., often highly detailed anatomy preserved in midgut glands) and has caused little controversy, even though decay studies suggest that they should have a very low preservation potential.[18] Early authigenic mineralization often confers a greater degree of three-dimensionality to fossilized guts than to more decay-resistant features, including cuticle.

6.1 Decay Induced Distortions Are Not Characteristic of Exceptional Fossils

Yang et al.[102] identified well organized segmental ganglia in a total group euarthropod from the Chengjiang biota. Sansom[18] argued that this interpretation was implausible based on the rapid loss of nervous system morphology in his decay experiments on priapulids. However, it is difficult to conceive how shrinkage of other anatomical features could generate the well-organized features[103–105] and serially repeated structures[102] interpreted as fossil nervous systems. Shrinking a cuticle would not be expected to generate a rope-ladder morphology that was the primary basis for identification as a nerve cord.

Decay experiments on priapulid worms have shown that carcasses develop pronounced asymmetrical bulges as they decay in seawater, presumably as a result of fluid and gas build up (e.g., Sansom,[18] Figure 3 and 4). It does not necessarily follow, however, that the relative body dimensions of fossil priapulids are likewise distorted and should be excluded from phylogenetic analysis. Priapulid specimens from the Burgess Shale are approximately symmetrical even where separation of the body wall from the cuticle indicates that some decay has taken place.[43] The familiar dark stains at the anterior and posterior of Burgess Shale fossils are not due to compaction, but reflect the escape of decay fluids; distortion of the body shape was limited by the confining effect of the sediment. Similar considerations apply to the decay of onychophorans. Asymmetrical bulges and distortion of the body observed in experiments[17] have not been observed in fossil lobopodians, even where the internal anatomical features have separated from the cuticle indicating that decay has taken place, such as in *Antennacanthopodia*.[106] Lobopodian fossils typically show no evidence of distortion, suggesting that build-up of decay fluids (sometimes evidenced by dark stains) is sometimes accommodated by leakage rather than deformation of the body.

6.2 Some Decay Resistant Features Do Not Preserve in Exceptional Fossils

The claws and jaws of onychophorans are decay resistant and, on that basis, their absence in *Helenodora* from the Carboniferous Mazon Creek deposit has been argued to be primary.[16] Likewise *Helenodora* is thought to have lacked slime papillae; they too are absent, and their preservation potential should be similar to other cuticular structures such as dermal papillae and limbs. The presence or absence of slime papillae is significant, as their presence in *Helenodora* would indicate a phylogenetic position close to the crown group of Onychophora. A recently described onychophoran from Montceau-les-Mines, France, a similar late Carboniferous assemblage preserved in concretions, preserves slime papillae and crown group-like antennal annuli, papillae, and trunk plicae but not claws.[107] Onychophoran claws have a deep evolutionary origin evidenced by their presence in stem onychophorans (lobopodians) such as *Hallucigenia* from the Cambrian Burgess Shale.[108] The presence of an otherwise crown onychophoran-like suite of characters without claws suggests that other mechanisms may explain their absence in both Carboniferous taxa, such as rapid shedding from the body soon after death, as observed in fossils in amber.[109] Furthermore, the highly retractile nature of slime papillae renders them difficult to observe, even with near pristine preservation of external cuticular anatomy and the use of synchrotron tomography,[109] so they too may also have been present in *Helenodora*, but are not preserved.

Experiments on cyclostomes and invertebrate chordates[20,60] showed that the notochord persists until the latest stages of decay (**Figure 4**). Nonetheless, the notochord is apparently absent in several taxa from Mazon Creek[63] even though other characters indicate that they belong to the vertebrate crown group, and therefore, possessed a notochord. The notochord is also absent in *Haikouichthys*, a total group vertebrate from the early Cambrian (Figure 4B), despite the preservation of characters such as eyes, gill pouches, and a dorsal fin, which disappear more rapidly in decay experiments, but clearly indicate a phylogenetic position consistent with the presence of a notochord.[20,60,110] *Haikouichthys* preserves a chimaeric assemblage of decay-prone and decay-resistant characters rather than corresponding to a particular decay stage (Figure 4). Likewise, the notochord is poorly preserved or equivocal in *Pikaia* and *Haikouichthys*, whereas other decay-prone characters including the eyes and nasal capsules are preserved in both taxa as well as the liver and heart in *Metaspriggina*, a vertebrate from the Burgess Shale.[111] Explaining the characters preserved in these fossils requires an appeal to more than just simply decay resistance. Furthermore, the quality of preservation varies among individuals of the same taxon, between taxa preserved in the same bed and between fossil assemblages, demonstrating that variations in environmental parameters influence the quality of preservation at different temporal and spatial scales.[33,81]

7. Conclusions

The fossilization of a carcass involves the interplay of rapid burial, decay, precipitation of minerals such as phosphate or pyrite, and subsequent diagenetic changes that occur on a geological time scale. Although decay experiments provide an important model for understanding the processes that impact soft-tissue preservation,[19,22] fossils do not simply represent a stage of decay. Decay-prone tissues (e.g., muscle tissue) can be preserved by authigenic mineralization even when more decay-resistant tissues are lost. Conversely decay-resistant structures (e.g., the notochord) often do not survive longer-term alteration. The assumption that decay-resistance determines which features fossilize[17,18] does not apply to every soft-bodied fossil. Factors other than decay can result in counterintuitive results (such as the preservation of muscle and not cuticle). Understanding and interpreting fossils requires the consideration of geological as well as biological processes; the preservational context is as critical as the evidence of the fossils themselves.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

Luke Parry and Fiann Smithwick were supported by NERC grant no.s NE/L501554/1and NE/L002434/1.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

decay experiments, exceptional preservation, Lagerstätten, taphonomy

Received: September 12, 2017
Revised: October 21, 2017
Published online:

[1] D. M. Raup, *Science* **1972**, *177*, 1065.
[2] A. Seilacher, *Neues Jahrbuch für Geologie und Paläontologie, Monatshefte* **1970**, *1970*, 34.
[3] A. Muscente, J. D. Schiffbauer, J. Broce, M. Laflamme, K. O'Donnell, T. H. Boag, M. Meyer, A. D. Hawkins, J. W. Huntley, M. McNamara,

L. A. Mackenzie, G. D. Stanley Jr, N. W. Hinman, M. H. Hofmann, S. Xiao, *Gondwana Res*. **2017**, *48*, 164.
[4] J.-B. Caron, R. R. Gaines, C. Aria, M. G. Mángano, M. Streng, *Nat. Commun*. **2014**, *5*, 3210.
[5] P. Van Roy, P. J. Orr, J. P. Botting, L. A. Muir, J. Vinther, B. Lefebvre, K. el Hariri, D. E. G. Briggs, *Nature* **2010**, *465*, 215.
[6] C. W. Shabica, A. Hay, *Richardson's Guide to the Fossil Fauna of Mazon Creek*. Northeastern Illinois University, Chicago **1997**.
[7] D. M. Martill, *Nature* **1990**, *346*, 171.
[8] Q. Li, K.-Q. Gao, J. Vinther, M. D. Shawkey, J. A. Clarke, L. D'alba, Q. Meng, D. E. G. Briggs, R. O. Prum, *Science* **2010**, *327*, 1369.
[9] G. Eglinton, G. A. Logan, R. Ambler, J. Boon, W. Perizonius, *Philos. Transac. Royal Soc. B: Biol. Sci*. **1991**, *333*, 315.
[10] D. E. G. Briggs, R. E. Summons, *BioEssays* **2014**, *36*, 482.
[11] D. E. G. Briggs, *Ann. Rev. Earth Planet. Sci*. **2003**, *31*, 275.
[12] P. C. Donoghue, M. A. Purnell, *BioEssays* **2009**, *31*, 178.
[13] S. McMahon, R. P. Anderson, E. E. Saupe, D. E. G. Briggs, *Geology* **2016**, *44*, 867.
[14] F. M. Smithwick, G. Mayr, E. T. Saitta, M. J. Benton, J. Vinther, *Palaeontology* **2017**, *60*, 409.
[15] D. E. G. Briggs, *Palaios* **1995**, *10*, 539.
[16] D. J. Murdock, S. E. Gabbott, M. A. Purnell, *BMC Evol. Biol*. **2016**, *16*, 1.
[17] D. J. Murdock, S. E. Gabbott, G. Mayer, M. A. Purnell, *BMC Evol. Biol*. **2014**, *14*, 222.
[18] R. S. Sansom, *Scientific Rep*. **2016**, *6*, 32817.
[19] D. E. G. Briggs, S. McMahon, *Palaeontology* **2016**, *59*, 1.
[20] R. S. Sansom, S. E. Gabbott, M. A. Purnell, *Nature* **2010**, *463*, 797.
[21] X. Ma, P. Cong, X. Hou, G. D. Edgecombe, N. J. Strausfeld, *Nat. Commun*. **2014**, *5*, 3560.
[22] R. S. Sansom, *Paleontol. Soc. Paper* **2014**, *20*, 259.
[23] A. G. Liu, D. Mcilroy, J. B. Antcliffe, M. D. Brasier, *Palaeontology* **2011**, *54*, 607.
[24] D. E. G. Briggs, P. R. Wilby, *J. Geol. Soc*. **1996**, *153*, 665.
[25] D. E. G. Briggs, S. H. Bottrell, R. Raiswell, *Geology* **1991**, *19*, 1221.
[26] E. B. Naimark, M. A. Kalinina, A. V. Shokurov, A. V. Markov, N. M. Boeva, *J. Paleontol*. **2016**, *90*, 472.
[27] J. Vinther, R. Nicholls, S. Lautenschlager, M. Pittman, T. G. Kaye, E. Rayfield, G. Mayr, I. C. Cuthill, *Curr. Biol*. **2016**, *26*, 2456.
[28] J. G. Gehling, *Palaios* **1999**, *14*, 40.
[29] L. G. Tarhan, A. vS. Hood, M. L. Droser, J. G. Gehling, D. E. G. Briggs, *Geology* **2016**, *44*, 951.
[30] L. A. Wilson, N. J. Butterfield, *Palaios* **2014**, *29*, 145.
[31] E. Naimark, M. Kalinina, A. Shokurov, N. Boeva, A. Markov, L. Zaytseva, *Palaeontology* **2016**, *59*, 583.
[32] R. R. Gaines, E. U. Hammarlund, X. Hou, C. Qi, S. E. Gabbott, Y. Zhao, J. Peng, D. E. Canfield, *Proc. Natl. Acad. Sci*. **2012**, *109*, 5180.
[33] J.-B. Caron, D. A. Jackson, *Palaios* **2006**, *21*, 451.
[34] P. R. Wilby, D. E. G. Briggs, P. Bernier, C. Gaillard, *Geology* **1996**, *24*, 787.
[35] D. E. G. Briggs, D. J. Siveter, D. J. Siveter, M. D. Sutton, *Am. Scientist* **2008**, *96*, 474.
[36] E. S. Barghoorn, S. A. Tyler, *Science* **1965**, *147*, 563.
[37] C. Rice, N. Trewin, L. Anderson, *J. Geol. Soc*. **2002**, *159*, 203.
[38] D. M. Martill, *Palaeontology* *38*, 897.
[39] D. E. G. Briggs, S. H. Williams, *Lethaia* **1981**, *14*, 157.
[40] W. Schäfer, *Senckenbergiana lethaea* **1955**, *36*, 1.
[41] C. Breder, *Copeia* **1957**, *1957*, 132.
[42] R. A. Berner, *Science* **1968**, *159*, 195.
[43] S. Conway Morris, *Special Papers in Palaeontology* **1977**, *20*, 1.
[44] R. E. Plotnick, *Palaios* **1986**, 286.
[45] P. A. Allison, *Paleobiology* **1988**, *14*, 139.
[46] A. J. Kear, D. E. G. Briggs, D. T. Donovan, *Palaeontology* **1995**, *38*, 105.
[47] D. E. G. Briggs, A. J. Kear, M. Baas, J. W. Leeuw, S. Rigby, *Lethaia* **1995**, *28*, 15.
[48] D. E. G. Briggs, A. J. Kear, *Paleobiology* **1993**, *19*, 107.
[49] D. E. G. Briggs, A. J. Kear, *Palaios* **1994**, 431.
[50] D. E. G. Briggs, A. J. Kear, D. Martill, P. Wilby, *J. Geol. Soc*. **1993**, *150*, 1035.
[51] D. E. G. Briggs, A. J. Kear, *Lethaia* **1993**, *26*, 275.
[52] B. J. Greenstein, *Palaios* **1991**, 519.
[53] D. E. G. Briggs, A. J. Kear, *Science* **1993**, *150*, 1035.
[54] C. H. Hof, D. E. G. Briggs, *Palaios* **1997**, *12*, 420.
[55] S. Conway Morris, J.-B. Caron, *Biol. Rev*. **2012**, *87*, 480.
[56] S. McMahon, L. G. Tarhan, D. E. G. Briggs, *Palaios* **2017**, *32*, 388.
[57] D. Casenove, T. Oji, T. Goto, *Paleontol. Res*. **2011**, *15*, 146.
[58] E. Beli, S. Piraino, C. B. Cameron, *Palaeontology* **2017**, *60*, 389.
[59] K. Nanglu, J.-B. Caron, C. B. Cameron, *Paleobiology* **2015**, *41*, 460.
[60] R. S. Sansom, S. E. Gabbott, M. A. Purnell, *Proc. Royal Soc. Lond. B: Biol. Sci*. **2011**, *278*, 1150.
[61] G. K. Colbath, *Micropaleontology* **1988**, *34*, 83.
[62] O. Hints, M. Eriksson, *Palaeogeo. Palaeoclimatol. Palaeoecol*. **2007**, *245*, 95.
[63] S. E. Gabbott, P. C. Donoghue, R. S. Sansom, J. Vinther, A. Dolocan, M. A. Purnell, *Proc. Royal Soc. Lond. B: Biol. Sci*. **2016**, *283*, 20161151.
[64] T. H. Harvey, N. J. Butterfield, *Nat. Ecol. Evol*. **2017**, *1*, 0022.
[65] E. T. Saitta, C. Rogers, R. A. Brooker, G. D. Abbott, S. Kumar, S. S. O'Reilly, P. Donohoe, S. Dutta, R. E. Summons, J. Vinther, *Palaeontology* **2017**, *60*, 547.
[66] I. Melendez, K. Grice, L. Schwark, *Scientific Rep*. **2013**, *3*, Article number 2768.
[67] K. Glass, S. Ito, P. Wilby, T. Sota, A. Nakamura, C. Bowers, J. Vinther, S. Dutta, R. Summons, D. E. G. Briggs, K. Wakamatsu, J. Simon, *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 10218.
[68] D. E. G. Briggs, *Philos. Transac. Royal Soc. Lond. B: Biol. Sci*. **1999**, *354*, 7.
[69] M. J. Collins, G. Muyzer, G. B. Curry, P. Sandberg, P. Westbroek, *Lethaia* **1991**, *24*, 387.
[70] M. Collins, P. Westbroek, G. Muyzer, J. De Leeuw, *Geochimica et Cosmochimica Acta* **1992**, *56*, 1539.
[71] S. Larter, A. Douglas, *Geochimica et Cosmochimica Acta* **1980**, *44*, 2087.
[72] B. A. Stankiewicz, J. C. Hutchins, R. Thomson, D. E. G. Briggs, R. P. Evershed, *Rapid Commun. Mass Spectrom*. **1997**, *11*, 1884.
[73] J. S. Sinninghe Damsté, T. I. Eglinton, J. W. De Leeuw, P. Schenck, *Geochimica et Cosmochimica Acta* **1989**, *53*, 873.
[74] M. E. McNamara, P. J. Orr, S. L. Kearns, L. Alcalá, P. Anadón, E. Peñalver-Mollá, *Geology* **2006**, *34*, 641.
[75] M. E. McNamara, P. J. Orr, S. L. Kearns, L. Alcalá, P. Anadón, E. Penalver Molla, *Palaios* **2009**, *24*, 104.
[76] J. S. Sinninghe Damsté, W. I. C. Rijpstra, A. Kock-van Dalen, J. W. De Leeuw, P. Schenck, *Geochimica et Cosmochimica Acta* **1989**, *53*, 1343.
[77] B. A. Stankiewicz, D. E. G. Briggs, R. Michels, M. E. Collinson, M. B. Flannery, R. P. Evershed, *Geology* **2000**, *28*, 559.
[78] D. E. G. Briggs, R. A. Moore, J. W. Shultz, G. Schweigert, *Proc. Royal Soc. Lond. B: Biol. Sci*. **2005**, *272*, 627.
[79] Z. Yin, J. A. Cunningham, K. Vargas, S. Bengtson, M. Zhu, P. C. Donoghue, *Precambrian Res*. **2017**, *301*, 145.
[80] N. J. Butterfield, *Paleobiology* **2002**, *28*, 155.
[81] P. Wilson, L. A. Parry, J. Vinther, G. D. Edgecombe, *Palaeontology* **2016**, *59*, 463.
[82] L. A. Parry, P. Wilson, D. Sykes, G. D. Edgecombe, J. Vinther, *BMC Evol. Biol*. **2015**, *15*, 256.

[83] P. R. Wilby, D. E. G. Briggs, *Geobios* **1997**, *30*, 493.
[84] J. Sagemann, S. J. Bale, D. E. G. Briggs, R. J. Parkes, *Geochimica et Cosmochimica Acta* **1999**, *63*, 1083.
[85] D. Martin, D. E. G. Briggs, R. J. Parkes, *J. Geol. Soc.* **2004**, *161*, 735.
[86] A. D. Butler, J. A. Cunningham, G. E. Budd, P. C. Donoghue, *Proc. R Soc. B* **2015**, *282*, 20150476.
[87] E. C. Raff, M. E. Andrews, F. R. Turner, E. Toh, D. E. Nelson, R. A. Raff, *Evol. Dev.* **2013**, *15*, 243.
[88] T. Clements, C. Colleary, K. De Baets, J. Vinther, *Palaeontology* **2017**, *60*, 1.
[89] F. J. Young, J. Vinther, *Palaeontology* **2017**, *60*, 27.
[90] T. A. Hegna, M. J. Martin, S. A. Darroch, *Geology* **2017**, *45*, 199.
[91] D. E. G. Briggs, R. Raiswell, S. Bottrell, D. T. Hatfield, C. Bartels, *Am. J. Sci.* **1996**, *296*, 633.
[92] U. Farrell, D. E. G. Briggs, *Proc. Royal Soc. Lond. B: Biol. Sci.* **2007**, *274*, 499.
[93] S. E. Gabbott, H. Xian-Guang, M. J. Norry, D. J. Siveter, *Geology* **2004**, *32*, 901.
[94] R. R. Gaines, D. E. G. Briggs, Y. Zhao, *Geology* **2008**, *36*, 755.
[95] P. J. Orr, D. E. G. Briggs, S. L. Kearns, *Science* **1998**, *281*, 1173.
[96] A. Page, S. E. Gabbott, P. R. Wilby, J. A. Zalasiewicz, *Geology* **2008**, *36*, 855.
[97] N. J. Butterfield, *Lethaia* **1995**, *28*, 1.
[98] J.-B. Caron, A. Scheltema, C. Schander, D. Rudkin, *Nature* **2006**, *442*, 159.
[99] D. E. G. Briggs, J.-B. Caron, *Curr. Biol.* **2017**, *27*, 2536.
[100] J. Vannier, M. Steiner, E. Renvoisé, S.-X. Hu, J.-P. Casanova, *Proc. Royal Soc. Lond. B: Biol. Sci.* **2007**, *274*, 627.
[101] N. J. Butterfield, *Integr. Compar. Biol.* **2003**, *43*, 166.
[102] J. Yang, J. Ortega-Hernández, N. J. Butterfield, Y. Liu, G. S. Boyan, J.-B. Hou, T. Lan, X.-G. Zhang, *Proc. Natl. Acad. Sci.* **2016**, *113*, 2988.
[103] X. Ma, X. Hou, G. D. Edgecombe, N. J. Strausfeld, *Nature* **2012**, *490*, 258.
[104] G. Tanaka, X. Hou, X. Ma, G. D. Edgecombe, N. J. Strausfeld, *Nature* **2013**, *502*, 364.
[105] J. Ortega-Hernández, *Curr. Biol.* **2015**, *25*, 1625.
[106] Q. Ou, J. Liu, D. Shu, J. Han, Z. Zhang, X. Wan, Q. Lei, *J. Paleontol.* **2011**, *85*, 587.
[107] R. J. Garwood, G. D. Edgecombe, S. Charbonnier, D. Chabard, D. Sotty, G. Giribet, *Invertebr. Biol.* **2016**, *135*, 179.
[108] M. R. Smith, J. Ortega-Hernández, *Nature* **2014**, *514*, 363.
[109] W. Zhang, G. Mayer, *Curr. Biol.* **2016**, *26*, 1.
[110] R. S. Sansom, S. E. Gabbott, M. A. Purnell, *Palaeontology* **2013**, *56*, 457.
[111] S. Conway Morris, J.-B. Caron, *Nature* **2014**, *512*, 419.
[112] D. E. G. Briggs, D. J. Siveter, D. J. Siveter, M. D. Sutton, D. Legg, *Proc. Natl. Acad. Sci.* **2016**, *113*, 4410.
[113] E. Sherratt, M. del Rosario Castañeda, R. J. Garwood, D. L. Mahler, T. J. Sanger, A. Herrel, K. de Queiroz, J. B. Losos, *Proc. Natl. Acad. Sci.* **2015**, *112*, 9961.
[114] X. G. Zhang, X. G. Hou, *J. Evol. Biol.* **2004**, *17*, 1162.
[115] M. E. Clapham, G. M. Narbonne, *Geology* **2002**, *30*, 627.
[116] A. G. Liu, *Palaios* **2016**, *31*, 259.
[117] G. M. Narbonne, *Science* **2004**, *305*, 1141.
[118] D. Grazhdankin, A. Seilacher, *Palaeontology* **2002**, *45*, 57.
[119] M. E. Clapham, G. M. Narbonne, J. G. Gehling, *Paleobiology* **2003**, *29*, 527.
[120] S. A. Darroch, M. Laflamme, J. D. Schiffbauer, D. E. G. Briggs, *Palaios* **2012**, *27*, 293.