Molecular phylogenetic analyses of Ecdysozoa and Haemosporida

Dissertation

submitted by

Janus Borner

to the

University of Hamburg

with the aim of achieving a
doctoral degree (Dr. rer.nat.)

at the

Faculty of Mathematics,
Informatics and Natural Sciences

Department of Biology

Hamburg, 2017

Gutachter der Dissertation:

Prof. Dr. Thorsten Burmester

Prof. Dr. Iris Bruchhaus

Prof. Dr. Arndt von Haeseler

Datum der mündlichen Prüfung:

07.09.2017

Danksagung

Mein Dank gilt Prof. Dr. Thorsten Burmester für die ausgezeichneten Forschungsbedingungen, für das Vertrauen und die Freiheit, die ich bei der Umsetzung von eigenen Projekten genossen habe, für die Unterstützung in allen Lebenslagen, und für die enorme Geduld, die ich am Ende reichlich in Anspruch genommen habe. Genauso danke ich Prof. Dr. Iris Bruchhaus, auf deren Hilfe und Engagement ich stets zählen konnte, und die viele wichtige Kollaborationen für mich in die Wege geleitet hat.

Ganz besonders möchte ich mich bei Maria Machola bedanken, der guten Seele des Labors, die mir über die Jahre immer wieder bei der Laborarbeit geholfen hat, und deren vorbildliche Organisation mir das ein oder andere Mal den Tag gerettet hat. Herzlich bedanken möchte ich mich auch bei Jenny Tiede, die im Rahmen ihrer Bachelorarbeit wichtige Daten zum Haemosporida-Projekt beigetragen hat.

Christian Pick danke ich für die viele Hilfe am Anfang meiner Doktorarbeit und dafür, dass er den Büro-Alltag nie hat langweilig werden lassen. Dafür danke ich auch meinen aktuellen Büro-Kollegen Andrej und Conny.

Ich möchte mich ganz herzlich bei allen Mitgliedern der Arbeitsgruppe bedanken für die angenehme Arbeitsatmosphäre, die lustigen Partys und die gegenseitige Unterstützung. Vielen Dank auch an die ehemaligen Kollegen, die ich manchmal vermisst habe, und zu deren Reihe ich mich nun geselle. Danke Sven, Mariana, Marco, Julia, Jessi, Sammy und Co.!

Ich danke meinen Eltern, die mich immer und in allen Dingen unterstützt haben. Vielen Dank an dieser Stelle auch an meine Großmutter!

Ganz besonders danke ich meiner Freundin, Annika Herwig, die mich mit viel Geduld (und Ungeduld) auf den langen letzten Metern begleitet hat. Mit Dir erscheinen mir alle Aufgaben etwas leichter (Doktorarbeiten ausgenommen), alle Horizonte etwas weiter, alle Tage etwas heller, und alle Morgen etwas früher. ;)

Table of Contents

1 Introduction

1.1 Molecular phylogenetics

Phylogenetics is the study of the evolutionary history and relationships among living or extinct organisms (Wägele, 2001; Brown, 2002; Storch & Welsch, 2003; Reece et al., 2011). In general, phylogenetic reconstruction is based on the comparison of homologous characters between organisms. Tree inference then aims to find the phylogeny that best explains the distribution of character states among taxa. Before large molecular datasets became available, phylogenetics relied on comparative morphology. While morphological data can be highly informative for answering phylogenetic questions, the amount of described characters is often insufficient for analysis via mathematical methods (Brown, 2002). Additionally, the definition of what constitutes a homologous morphological character is dependent on human interpretation and therefore subjective in nature (Graur & Li, 2000).

The advent of nucleotide sequencing techniques has enabled researchers to employ molecular sequence data for phylogenetic reconstruction and has challenged many traditional views on evolutionary relationships across the tree of life. The classification of all living organisms into three domains (Bacteria, Archaea, and Eukarya) was based on molecular data, which showed that Archaea, initially described as extremophile bacteria, represent an entirely new group of organisms that are genetically distinct from both bacteria and eukaryotes (Woese et al., 1990). Within Mammalia, a new superorder called Afrotheria was erected comprising elephants (Proboscidea), sea cows (Sirenia), hyraxes (Hyracoidea), aardvark (Tubulidentata), elephant shrews (Macroscelidea), and golden moles and tenrecs (Afrosoricida). These orders share few common morphological traits and were previously considered members of other established groups of mammals, including ungulates and insectivores. Nevertheless, molecular data unequivocally revealed that Afrotheria constitute an ancient group of mammals that evolved in Africa, presumably while the continent was isolated through plate tectonics (Springer et al., 1997; Madsen et al., 1997; Stanhope et al., 1998). Another example of the profound effect that molecular phylogenetics had on our view of evolutionary relationships are the protostomes. Based on sequence data, Protostomia have been divided into Ecdysozoa, including arthropods and nematodes, and Lophotrochozoa, including annelids and molluscs (Aguinaldo et al., 1997), thus refuting the sister group relationship of Arthropoda and Annelida, which morphologists considered to be one of the best-supported relationships among animal phyla (e.g., Westheide & Rieger, 1996; Brusca & Brusca, 2003).

Molecular sequence data can provide enormous amounts of phylogenetic information because each nucleotide or amino acid position can be considered as an independent character. Though individual positions only contain limited phylogenetic signal (or none at all), the combined information from hundreds or thousands of positions can be sufficient to reconstruct a well-resolved phylogeny. Due to the high costs and technical challenges initially associated with nucleotide sequencing, early molecular phylogenetic studies were usually limited to analyses of single genes (e.g., Woese et al., 1990; Irwin et al., 1991; Ruvolo et al., 1991). However, for the analysis of deep phylogenetic relationships, the amount of sequence data available from single genes may not be sufficient as nucleotide substitutions accumulate in the sequences over evolutionary time and stochastic noise may drown out the phylogenetic signal contained in data from a single gene (Saitou & Nei, 1986; Walsh et al., 1999). This issue can be overcome by employing supermatrices that contain concatenated data from multiple genes. It is important to note in this context that the phylogenetic history of an individual gene (gene tree) is not necessarily congruent with the branching pattern of the species (species tree) (Page & Holmes, 1998; Graur & Li, 2000). On an evolutionary time scale, gene duplications are common events (Lynch & Conery, 2000). When speciation occurs after gene duplication has produced multiple copies of a gene in the common ancestor, the resulting gene tree diverges from the species tree. Homologous genes that are related by duplication within a genome are called paralogs, whereas genes that evolved from a common ancestral gene by speciation are called orthologs (Jensen, 2001). However, the distinction between orthologs and paralogs can become complicated when different copies of a gene are subsequently lost during evolution. For this reason, studies aiming to reconstruct the evolutionary history of species or taxa have mostly employed datasets of single-copy orthologous genes (e.g., Baldauf et al., 2000; Dunn et al., 2008; Roeding et al., 2009).

The development of next-generation sequencing (NGS) techniques has allowed researchers to generate huge volumes of genetic data via massively parallel sequencing of small DNA fragments, which can subsequently be assembled into larger contigs. Because phylogenetic reconstruction of ancient evolutionary events requires highly conserved sequences, non-coding regions of the genome, which tend to be highly variable, are not well suited for this task. For this reason, transcriptome sequencing has become the method of choice for large-scale deep-level phylogenetic analysis (e.g., Roeding et al., 2009; Hittinger et al., 2010; Misof et al., 2014), as transcribed messenger RNA primarily consists of the coding sequence of a gene (as opposed to genomic DNA, which contains large amounts of non-coding regions). The expansion of phylogenetic data matrices to hundreds or even thousands of genes has eliminated stochastic noise as a source for erroneous phylogenies. However, increasing the amount of sequence data cannot solve systematic errors. One of the most serious issues for computational methods is long-branch attraction (LBA), which can occur when a tree includes a combination of long and short branches so that similarity due to convergent character substitutions (homoplasy) produces an artifactual grouping of distantly related lineages. This phenomenon was first described by Felsenstein (1978) for tree inference using maximum parsimony. While maximum likelihood analyses and Bayesian inference are more robust to the effect of LBA (Philippe et al., 2005a), they are not immune and long-branching taxa can lead to erroneous results with these methods too (Bergsten, 2005), as e.g., the high support for the now abandoned "Coelomata" concept based on poor taxon sampling has shown (see 1.2). Thus, both gene and taxon sampling may have profound effects on the outcome of phylogenetic analyses and have to be considered carefully.

In this thesis, I will present the results of my studies on the phylogeny of Ecdysozoa, with a special focus on Myriapoda and Chelicerata, and on the phylogeny of Apicomplexa, with a special focus on Haemosporida.

1.2 The phylogeny of Ecdysozoa

The superphylum Ecdysozoa was first proposed by Aguinaldo et al. (1997) based on phylogenetic analyses of 18S ribosomal RNA sequences. It comprises the two most species-rich animal phyla, Arthropoda and Nematoda, and six smaller phyla: Onychophora (velvet worms), Tardigrada (water bears), Nematomorpha (horsehair worms), Priapulida (penis worms), Kinorhyncha (mud dragons) and Loricifera. The eponymous shared character (synapomorphy) of Ecdysozoa is the periodic molting, or ecdysis, of the three-layered cuticle, which is controlled by ecdysteroid hormones (Westheide & Rieger, 2013). Apart from this, Ecdysozoa only have few morphological characters in common and are primarily characterized by the shared absence of common protostome traits, such as spiral cleavage or locomotory cilia.

Based on comparative morphology, the phyla that have now been united in Ecdysozoa were originally assigned to two major taxonomic groups: the segmented, limb-bearing panarthropods (Arthropoda plus Onychophora and Tardigrada) and the worm-like cycloneuralians (Nematoda, Nematomorpha, Priapulida, Kinorhyncha, and Loricifera). Before the advent of molecular phylogenetics, there was a strong consensus among taxonomists (e.g., Westheide & Rieger, 1996; Brusca & Brusca, 2003) for a sister group relationship of panarthropods and annelids (which meanwhile have been assigned to the superphylum Lophotrochozoa; see above). Based on the principal character uniting both taxa, a segmented body, this clade was called "Articulata". To the exclusion of the pseudocoelomate cycloneuralian phyla, "Articulata" were considered to be part of a larger assemblage of animal phyla called "Coelomata", which are linked by the possession of a coelomic body cavity, and which also include molluscs and vertebrates. The "Coelomata" concept also found support from several molecular analyses that employed large datasets derived from whole genomes (Blair et al., 2002; Wolf et al., 2004; Ciccarelli et al., 2006; Rogozin et al., 2007). However, the taxon sampling of these studies was limited and the basal position of the nematode *Caenorhabditis elegans* in the resulting phylogeny was probably an artifact caused by LBA due to the high substitution rate in the genome of *C. elegans* (Copley et al., 2004; Irimia et al., 2007). In fact, studies with improved taxon sampling, which have included more slowly evolving nematode species, consistently recovered Ecdysozoa (Philippe et al., 2005b; Webster et al., 2006; Roeding et al., 2007; Dunn et al., 2008; Meusemann et al., 2010).

While the Ecdysozoa concept has become widely accepted, the relationships within Ecdysozoa have remained poorly understood (Fig. 1). There is ample evidence for a close relationship between Nematoda and Nematomorpha (Nielsen, 1995; Schmidt-Rhaesa, 1996; Mallatt et al., 2004; Dunn et al., 2008), which together form the taxon Nematoida (Schmidt-Rhaesa, 1996). The remaining cycloneuralian taxa (Priapulida, Kinorhyncha, and Loricifera) have been united as Scalidophora on the basis of a shared spine-covered introvert (retractable and invertible proboscis) and the presence of two rings of retracting muscles on the introvert (Schmidt-Rhaesa, 1998). So far, only few molecular phylogenetic studies have included data from scalidophoran species. These studies found Scalidophora in a basal position within Ecdysozoa, thus rejecting monophyletic Cycloneuralia. While there is general agreement that Onychophora are closely associated with Arthropoda (e.g., Ballard et al., 1992; Boore et al., 1995; Kusche et al., 2002; Roeding et al., 2007), the phylogenetic position of the third panarthropod phylum, Tardigrada, is still matter of debate, with some studies favoring a nematode association (Giribet, 2003; Roeding et al., 2007; Lartillot & Philippe, 2008; Meusemann et al., 2010) and others a close relationship to arthropods (Gabriel & Goldstein, 2007; Rota-Stabelli et al., 2011; Campbell et al., 2011; Mayer et al., 2013).

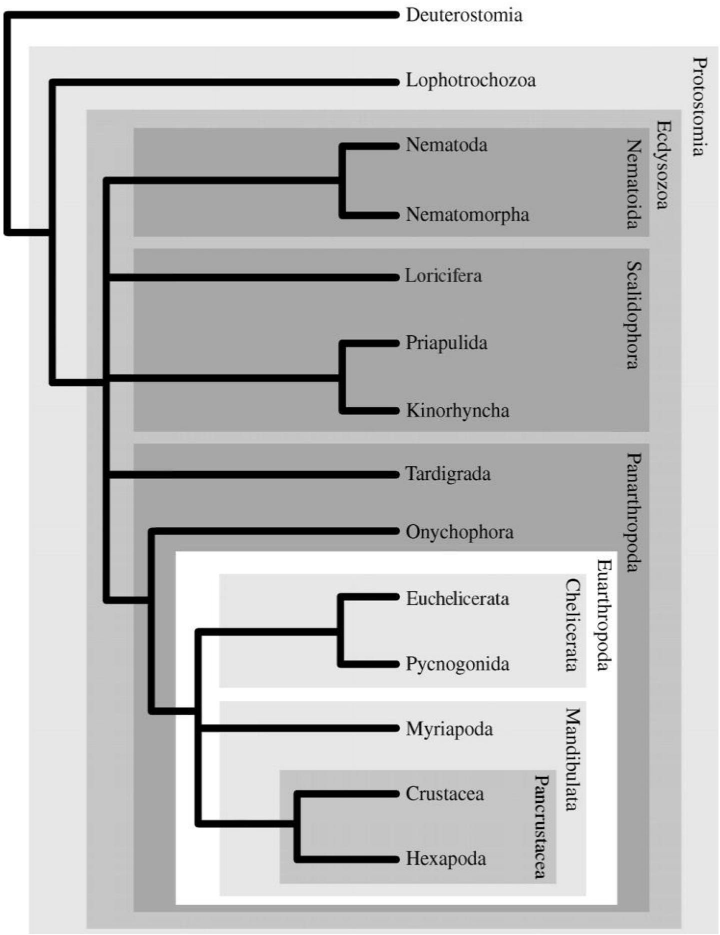


Fig. 1. Consensus phylogeny of Ecdysozoa. Contended nodes are shown as polytomies. Modified from Telford et al. (2008).

The relationships of the four eurthropod clades (Chelicerata, Myriapoda, Crustacea, and Hexapoda) have long been disputed. Chelicerates were traditionally placed at the base of the phylum as the sister group of Mandibulata, a taxon which comprises Crustacea, Hexapoda, and Myriapoda (Westheide & Rieger, 1996). Based on morphological data, Hexapoda and Myriapoda have been united in a taxon called "Tracheata" or "Atelocerata" (Fig. 2A). Molecular phylogenetic studies, however, have found Crustacea and Hexapoda to be more closely related (e.g., Friedrich & Tautz, 1995; Boore et al., 1998; Kusche & Burmester, 2001; Dunn et al., 2008), together forming the taxon Pancrustacea (Zrzavý & Štys, 1997) or Tetraconata (Dohle, 2001) and possibly rendering Crustacea paraphyletic with regard to Hexapoda (Nardi et al., 2003; Ertas et al., 2009). In most of these studies, Myriapoda were recovered as the sister group of Chelicerata, together referred to as "Myriochelata" (Pisani et al., 2004; Fig. 2B) or "Paradoxopoda" (Mallatt et al., 2004). While the Pancrustacea concept has found increasing support among morphologists (e.g., Duman-Scheel & Patel, 1999; Harzsch & Hafner, 2006), evidence in favor of Myriochelata is mostly limited to similarities in neurogenesis between myriapods and chelicerates (Dove & Stollewerk, 2003).

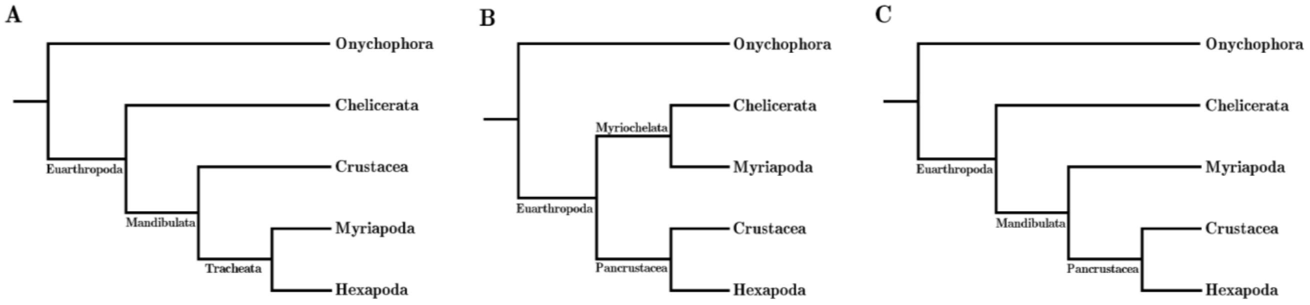


Fig. 2. Competing hypotheses of arthropod phylogeny. (A) Traditional "Tracheata" concept (Westheide & Rieger, 1996). (B) Myriochelata + Pancrustacea hypothesis (Friedrich & Tautz, 1995). (C) Pancrustacea as part of Mandibulata (Regier et al., 2010). Modified from Borner (2010).

1.2.1 Myriapoda

The subphylum Myriapoda comprises four extant classes: the predatory Chilopoda (centipedes), the mostly detritivore Diplopoda (millipedes), and the two lesser-known, soil-dwelling classes Symphyla and Pauropoda, which are minuscule, translucent animals often barely visible to the human eye. Following the "Tracheata" concept, myriapods were traditionally postulated to be paraphyletic in terms of the hexapods. However, considering the strong support for a close relationship between Crustacea and Hexapoda, this concept has been abandoned by most researchers. While the monophyly of the four myriapod classes is undisputed, almost every possible topology has been proposed for the internal relationships of Myriapoda (Edgecombe, 2011). Based on morphological characters, such as anterior placement of the genital openings, Symphyla, Pauropoda, and Diplopoda have been united in a clade named ''Progoneata'' (Dohle, 1980). Within "Progoneata", Pauropoda and Diplopoda were traditionally regarded as sister taxa ("Dignatha"; Fig. 3A). Molecular analyses, in contrast, have favored a sister group relationship of Symphyla and Pauropoda (together "Edafopoda";

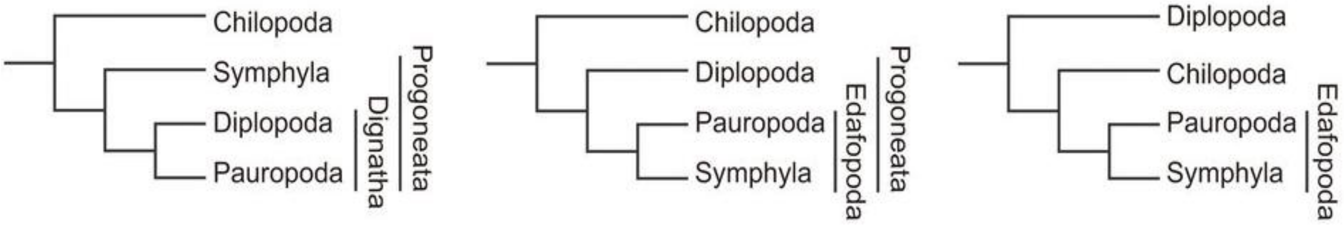


Fig. 3. Hypotheses of myriapod relations. (A) Traditional view based on morphology (Dohle, 1980). (B) Edafopoda as part of Progoneata (Regier et al., 2010). (C) Edafopoda as sister group of Chilopoda (Gai et al., 2006). Modified from Miyazawa et al. (2014).

Fig. 3B). Some of these studies support the monophyly of "Progoneata" (Regier et al., 2010; Dong et al., 2012; Zwick et al., 2012), while others found a sister group relationship of "Edafopoda" and Chilopoda (Gai et al., 2006; Fig. 3C).

1.2.2 Chelicerata

Chelicerates are characterized by the possession of claw-like head appendages, called chelicerae, which are used to grasp or pierce food (Westheide & Rieger, 2013). The inclusion of Pycnogonida (sea spiders) into Chelicerata at the base of the taxon has found strong support from molecular studies (Roeding et al., 2007; Dunn et al., 2008; Sanders & Lee, 2010; Meusemann et al., 2010; Regier et al., 2010) and studies on Hox genes (Jager et al., 2006) and neuroanatomy (Brenneis et al., 2008) have found evidence for the homology of the pycnogonid chelifores and the chelicerae of euchelicerates. The phylogenetic relationships among euchelicerate clades (all chelicerates excluding Pycnogonida; Weygold & Paulus, 1979) are poorly understood, and there is significant conflict between molecular and morphological data. While most morphological studies favor a sister group relationship between Xiphosura (horseshoe crabs) and the terrestrial Arachnida (Shultz, 1990; Wheeler & Hayashi, 1998), some palaeontological studies argue that there is fossil evidence for an independent aquatic origin of the taxon Scorpiones (Briggs, 1987; Jeram, 1998; Dunlop & Webster, 1999). Most molecular studies neither support a basal position of Scorpiones nor the taxon Arachnida *sensu stricto*, as Acari (mites and ticks) tend to group at the base of Euchelicerata (Dunn et al., 2008; Roeding et al., 2009; Meusemann et al., 2010). The best supported higher arachnid taxon is certainly Tetrapulmonata. This group comprising Araneae (spiders), Amblypygi (whip spiders), Thelyphonida (whip scorpions), and schizomids (Schizomida) has been consistently recovered in both morphological (e.g., Weygold & Paulus, 1979; Shear et al., 1987; Shultz, 1990) and molecular studies (Shultz & Regier, 2000; Jones et al., 2007; Pepato et al., 2010; Regier et al., 2010). However, the relationships of the remaining chelicerate orders have remained poorly resolved in molecular analyses, and the absence of NGS data for several key taxa has further exacerbated this problem.

1.3 The phylogeny of Apicomplexa

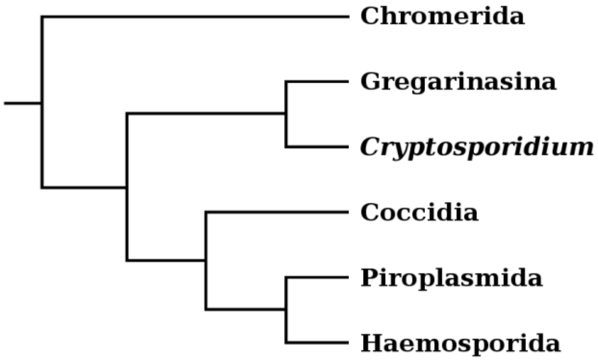


Fig. 4. Phylogenetic relations of major apicomplexan groups based on Templeton et al. (2010).

The protozoan phylum Apicomplexa comprises a diverse group of obligate intracellular parasites that may cause serious illnesses in humans and animals. For example, Apicomplexa include the causative agents of malaria (genus *Plasmodium*), toxoplasmosis (*Toxoplasma*), and babesiosis (*Babesia*). Despite the great diversity in their life cycles (Roos, 2005), involving a wide range of different hosts (both invertebrates and vertebrates), apicomplexans share several unique molecular and cellular features, i.e. an apical complex derived from elements of the flagellar apparatus (Francia et al., 2012; de Leon et al., 2013), a non-photosynthetic secondary plastid, called apicoplast (McFadden et al., 1996), and a conserved gliding motility and cell invasion machinery (Kappe et al., 1999; Baum et al., 2006). The closest relatives of Apicomplexa are the coral-endosymbiotic chromerid algae (Fig 4; Moore et al., 2008) and the parasite apicoplast is likely derived from the algal chloroplast (Janouškovec et al., 2010).

At the base of Apicomplexa, the gregarines (Gregarinasina), which exclusively parasitize invertebrates, form the sister group of *Cryptosporidium* (Fig. 4; Carreno et al., 1999; Zhu et al., 2000a; Templeton et al., 2010), a genus of vertebrate parasites that cause cryptosporidiosis in humans. Both parasite taxa appear to have lost their plastid genomes (Zhu et al., 2000b; Toso & Omoto, 2007). Originally, the genus *Cryptosporidium* was assigned to Coccidia, a diverse order of parasites that have been described from all major vertebrate groups including fish, reptiles, birds, and mammals. Various genera of coccidians infect livestock and poultry causing large economic costs for the agricultural industry (Williams, 1998; Trees et al., 1999). Toxoplasmosis, caused by the coccidian parasite *Toxoplasma gondii*, is the most prevalent infection of any kind in humans with an estimated prevalence of 30% to 50% of the world population. While the majority of individuals infected with *T. gondii* remain asymptomatic or only show minor symptoms (Montoya & Liesenfeld, 2004), primary infection in pregnant women can lead to spontaneous abortion or stillbirth (Havelaar et al., 2007) and, in immunosuppressed patients, infection can lead to life-threatening cerebral toxoplasmosis (Porter &

Sande, 1992). While coccidian parasites exclusively infect vertebrate hosts, Piroplasmida and Haemosporida rely on arthropod vectors for transmission. Piroplasmid parasites are transmitted via ixodid ticks, which are also the definite hosts. This order comprises two genera, *Babesia* and *Theileria*, which have a substantial economic impact on livestock and companion animals especially in the tropics and subtropics (Collett, 2000; Kivaria et al., 2007). Human babesiosis is an emerging disease in North America and parts of Europe and can, in severe cases, potentially be life threatening (Homer et al., 2000; Herwaldt et al., 2011). Parasites of the order Haemosporida are transmitted via dipteran vectors and include the agents of human malaria, which belong to the genus *Plasmodium*. With an estimated 438,000 casualties attributable to the disease in 2015 (WHO, 2015), malaria remains one of the greatest threats to human health.

1.3.1 Haemosporidian relationships

Several haemosporidian genome (e.g., Carlton et al.. 2002; Gardner et al.. 2002; Pain et al.. 2008; Tachibana et al.. 2012; Bensch et al., 2016) and transcriptome (e.g., Bozdech et al., 2003; Hall et al., 2005; Lauron et al., 2014; Videvall et al., 2015; Zhu et al., 2016a) sequencing projects have provided a wealth of data, which have been instrumental in gaining insights into the molecular basis of host–parasite interactions (e.g., Marti et al., 2004; Hiller et al., 2004; Hall et al., 2005) and have helped to identify potential drug targets (Yeh & Altman, 2006). Due to their enormous medical and economical importance, these sequencing efforts have mostly focused on a few members of the genus *Plasmodium* that infect mammalian hosts. However, they represent only a small fraction of the systematic and ecological diversity of haemosporidian parasites while other key taxa for the understanding of haemosporidian evolution have so far been neglected. For this reason, the deep-level phylogenetic relationships among major haemosporidian lineages have remained enigmatic. Yet, understanding the evolution of parasite life history traits and the emergence of new diseases depends on the knowledge of a solid phylogenetic backbone (Lefevre et al., 2007).

Before the advent of DNA sequencing techniques, the classification of haemosporidian parasites solely relied on their morphology, their life-history characteristics, and the taxonomy of the infected vertebrate hosts and insect vectors (e.g., Garnham, 1966). Based on these characters, 15 extant haemosporidian genera have been erected. However, several of these genera only contain a single

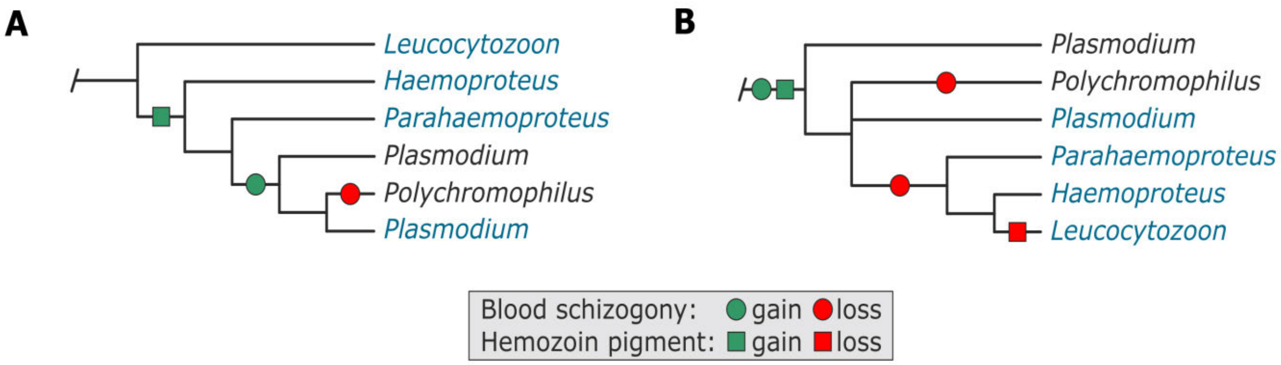


Fig. 5. Phylogenetic hypotheses on deep-level relationships among haemosporidian genera. (A) Traditional view of haemosporidian phylogeny with *Leucocytozoon* at the base of Haemosporida (based on Witsenburg et al., 2012). (B) Phylogeny based on an outgroup-free molecular clock based analysis with polyphyletic *Plasmodium* (Outlaw & Ricklefs, 2011). Parasites of sauropsid hosts are depicted in blue. Modified from Borner et al. (2014).

described species while the vast majority of the more than 500 described species have been assigned to the four genera *Plasmodium*, *Hepatocystis*, *Haemoproteus*, and *Leucocytozoon*. The latter has mostly been placed at the base of the haemosporidian tree for its lack of schizogony in the red blood cells and in its inability to produce hemozoin pigment (a metabolite of hemoglobin digestion), whereas *Plasmodium*, which exhibits both traits, has been considered to be the most derived lineage (Fig. 5A). Molecular phylogenetic studies have so far been limited to small numbers of gene fragments because genome or transcriptome data were only available for a small set of *Plasmodium* species. Most analyses relied on just four genes as the development of new phylogenetic markers has proven to be very challenging. While trees based on these datasets generally found good support on the level of genera and species (e.g., Martinsen et al., 2008; Schaer et al., 2013), the gene sampling is not well suited for uncovering the deepest phylogenetic relationships. A major factor contributing to this problem is that all potential outgroup taxa are too distantly related to be used with these datasets because their sequences are too divergent. For this reason, *Leucocytozoon* has been used as the outgroup in most analyses of haemosporidian phylogeny. This practice has been criticized by Outlaw & Ricklefs (2011) who employed an outgroup-free molecular clock approach to rooting, which resulted in a markedly different phylogeny, essentially dividing Haemosporida into a saurian and a mammalian clade. In this tree, *Leucocytozoon* is a derived lineage and *Plasmodium* is polyphyletic (Fig. 5B).

The bat-infecting genera *Hepatocystis* and *Polychromophilus* have been recovered nested within *Plasmodium* in all molecular analyses. While *Hepatocystis* has consistently been placed within the mammalian clade of *Plasmodium* parasites (Perkins & Schall, 2002; Martinsen et al., 2008; Outlaw &

Ricklews, 2011), the position of *Polychromophilus* is more ambiguous with some studies favoring a close relationship with sauropsid *Plasmodium* (Megali et al., 2011; Witsenburg et al., 2012) and others supporting an association with the mammalian parasites (Schaer et al., 2013).

While the datasets used for reconstructing the haemosporidian phylogeny have made steady progress in terms of taxon sampling, all studies have relied on similar sets of no more than four, rather short gene fragments mostly of mitochondrial or apicoplast origin, which are not well suited for deep-level phylogenetic analyses. The phylogenetic signal contained in these sequences might not be sufficient to resolve the deepest nodes of the tree. Another problem is that these genes are not well suited for the inclusion of distant outgroups because the sequences are too divergent (Martinsen et al., 2008). The inability to include outgroup taxa is especially problematic because the major point of contention regarding haemosporidian phylogeny relates to the position of the root, upon which basically all other deep-level relationships depend.

1.4 Publications in chronological order

In this thesis, I will present the main conclusions from the following publications:

Borner J, Burmester T (2017) Parasite infection of public databases: a data mining approach to identify apicomplexan contaminations in animal genome and transcriptome assemblies. *BMC Genomics* 18: 100.

Borner J, Pick C, Thiede J, Kolawole OM, Kingsley MT, Schulze J, Cottontail VM, Wellinghausen N, Schmidt-Chanasit J, Bruchhaus I, Burmester T (2016) Phylogeny of haemosporidian blood parasites revealed by a multi-gene approach. *Mol Phylogenet Evol* 94: 221-231.

Borner J, Rehm P, Schill RO, Ebersberger I, Burmester T (2014) A transcriptome approach to ecdysozoan phylogeny. *Mol Phylogenet Evol* 80: 79-87.

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.

Dunlop J, **Borner J**, Burmester T (2014) Phylogeny of the Chelicerates: Morphological and Molecular Evidence. In: Wägele JW, Bartholomaeus T (Eds.) *Deep metazoan phylogeny: the backbone of the tree of life. New insights from analyses of molecules, morphology, and theory of data analysis.* (pp. 395-408) Berlin: De Gruyter.

Hartig G, Peters RS, **Borner J**, Etzbauer C, Misof B, Niehuis O (2012) Oligonucleotide primers for targeted amplification of single-copy nuclear genes in apocritan Hymenoptera. *PLoS One* 7: e39826.

Rehm P, Pick C, **Borner J**, Markl J, Burmester T (2012) The diversity and evolution of chelicerate hemocyanins. *BMC Evol Biol* 12: 19.

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.

Peters RS, Meyer B, Krogmann L, **Borner J**, Meusemann K, Schütte K, Niehuis O, Misof B (2011) The taming of an impossible child: a standardized all-in approach to the phylogeny of Hymenoptera using public database sequences. *BMC Biol* 9: 55.

2 Discussion

2.1 Bioinformatic approaches in phylogenetics

2.1.1 Bioinformatic pipelines for the generation of phylogenetic datasets

(based on Peters et al., 2011)

The amount of molecular sequence data available in public databases has grown exponentially over the last decades (Cook et al., 2016). These databases represent an invaluable resource for phylogenetic studies. However, the annotation of sequences in uncurated databases is often highly inconsistent and, in some cases, even erroneous (e.g., Ben-Shitrit et al., 2012; Promponas et al., 2015). To generate multi-gene datasets suitable for phylogenomic analyses, many computational steps are required from sequence acquisition and curation, to orthology prediction, data selection, and sequence alignment. While a number of bioinformatic tools have been developed to perform these individual tasks, their execution on thousands of genes must be automated and parallelized, detailed records of all analyses need to be kept, and data files often have to be reformatted between analysis steps. In an automated bioinformatic approach, gene and taxon selection necessarily have to be based on clearly defined objective criteria. This is important because manual data selection may result in phylogenetic bias and, as the amount of publicly available data grows, it becomes unfeasible to simply include all available data from species belonging to the taxonomic group of interest. Several bioinformatic approaches to automate the generation of phylogenomic datasets from publicly available sequence data have been published (e.g., McMahon & Sanderson, 2006; Sanderson et al., 2008; Thomson & Shaffer, 2010; Robbertse et al., 2011). However, while these pioneering efforts were influential and innovative, they were either lacking in the degree of automation and detail of analysis or were limited to specific use cases. Furthermore, the problems of data scarcity, poor taxonomic overlap between datasets, non-stationary substitution processes, base compositional heterogeneity, and data quality deficits required new solutions (Peters et al., 2011).

To address the above mentioned issues, a novel bioinformatic pipeline (Fig. 6) was developed and employed to elucidate the phylogeny of the insect order Hymenoptera (Peters et al., 2011). This extremely diverse taxon was chosen to demonstrate the functionality of the pipeline and its ability to

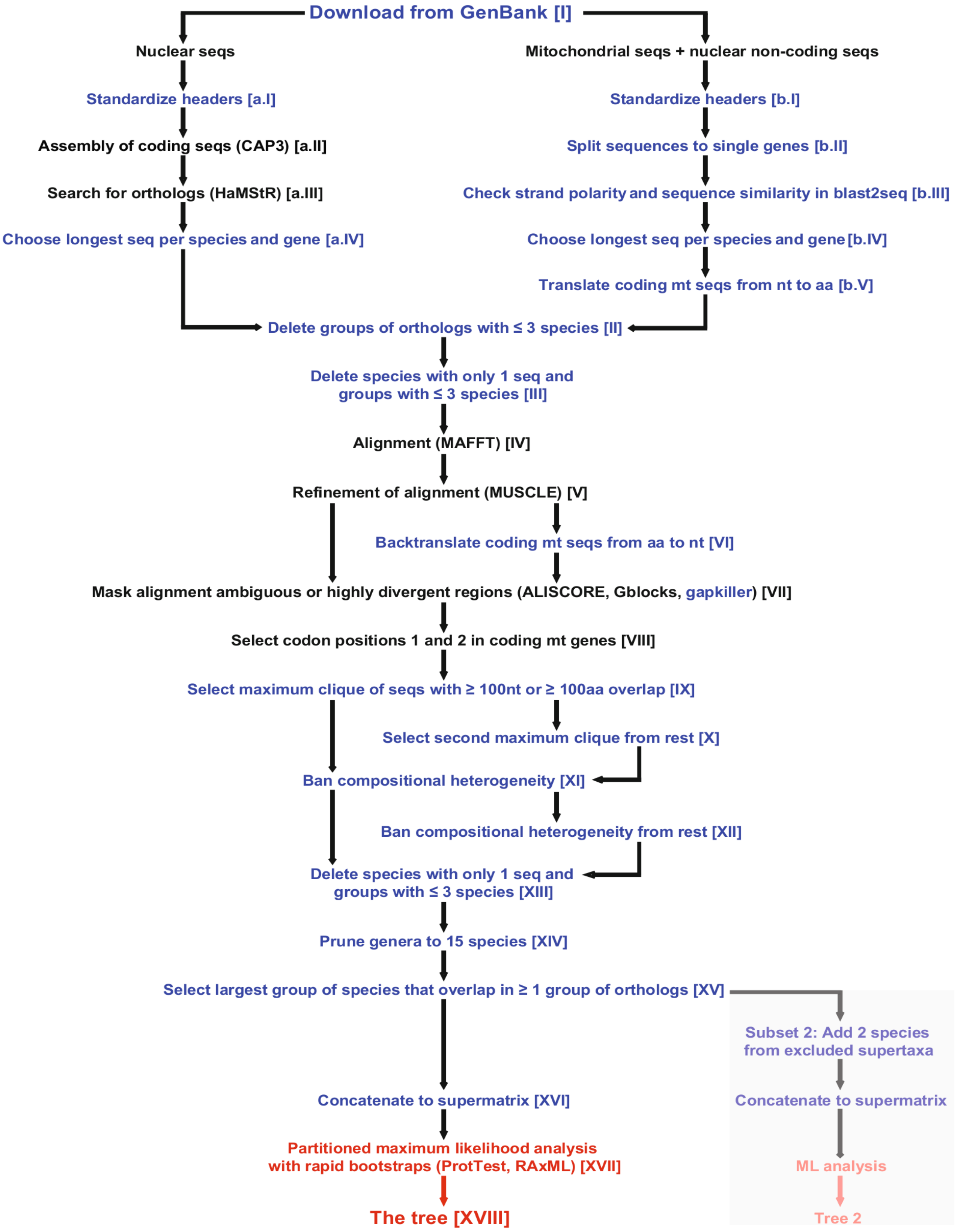

Download from GenBank [I]
Nuclear seqs
Mitochondrial seqs + nuclear non-coding seqs
Standardize headers [a.I]
Standardize headers [b.I]
Assembly of coding seqs (CAP3) [a.II]
Split sequences to single genes [b.II]
Search for orthologs (HaMStR) [a.III]
Check strand polarity and sequence similarity in blast2seq [b.III]
Choose longest seq per species and gene [a.IV]
Choose longest seq per species and gene [b.IV]
Translate coding mt seqs from nt to aa [b.V]
Delete groups of orthologs with ≤ 3 species [II]
Delete species with only 1 seq and groups with ≤ 3 species [III]
Alignment (MAFFT) [IV]
Refinement of alignment (MUSCLE) [V]
Backtranslate coding mt seqs from aa to nt [VI]
Mask alignment ambiguous or highly divergent regions (ALISCORE, Gblocks, gapkiller) [VII]
Select codon positions 1 and 2 in coding mt genes [VIII]
Select maximum clique of seqs with ≥ 100nt or ≥ 100aa overlap [IX]
Select second maximum clique from rest [X]
Ban compositional heterogeneity [XI]
Ban compositional heterogeneity from rest [XII]
Delete species with only 1 seq and groups with ≤ 3 species [XIII]
Prune genera to 15 species [XIV]
Select largest group of species that overlap in ≥ 1 group of orthologs [XV]
Subset 2: Add 2 species from excluded supertaxa
Concatenate to supermatrix [XVI]
Concatenate to supermatrix
Partitioned maximum likelihood analysis with rapid bootstraps (ProtTest, RAxML) [XVII]
ML analysis
The tree [XVIII]
Tree 2

Fig. 6. Schematic overview of the phylogenomic pipeline (Peters et al., 2011). Steps that are performed by newly developed scripts are highlighted in blue; steps that directly refer to the phylogenetic analysis are highlighted in red; external programs are written in parentheses after the step description.

deal with the above mentioned well-known phylogenetic challenges. More than 120,000 single gene sequences from ~4,500 hymenopteran species were downloaded and processed by the pipeline, resulting in a final supermatrix of ~80,000 sites from more than 1,100 species. Despite large amounts of missing data for most taxa, the inferred tree was generally consistent with previous studies, thus validating our approach.

Specifically, I implemented the parts of the pipeline that automate the sequence download from Genbank, the assembly into contigs, the orthology prediction by HaMStR (Ebersberger et al., 2009), and the alignment of individual genes (Fig. 6; steps I-IV and a.I-a.IV). These scripts were later expanded into a new software pipeline designed to automate all steps required for generating phylogenomic datasets based on NGS transcriptome data. The ability to reuse parts of the pipeline and adapt it to a slightly different use case highlights the strengths of a modular approach in which all tasks are performed by individual scripts that can be modified, rearranged, or replaced. The newly developed pipeline was successfully employed in several phylogenomic studies (Rehm et al., 2014; Borner et al., 2014; Borner et al., 2017).

In recent years, a number of novel phylogenomic pipelines have been published (Dunn et al., 2013; Grant & Katz, 2014; Kumar et al., 2015; Sahraeian et al., 2015), which mostly perform the same individual tasks but differ in the software they employ. For example, while Dunn et al. (2013) used TRIBE-MCL (Enright et al., 2002) for the orthology assignment, Grant & Katz (2014) employed OrthoMCL (Chen et al., 2006) for the same task, and Kumar et al. (2015) have implemented a new solution based on single gene phylogenetic analyses. The development and iterative improvement of new sequencing technologies continue to accelerate the growth of public databases, thereby shifting the bottle neck in biological research from experimental data acquisition to computational data management, processing, and knowledge extraction. For this reason, the development of new bioinformatic pipelines is crucial for our ability to leverage the full scientific potential from the vast amounts of sequence data.

2.1.2 Data mining of public databases for parasite contamination

(based on Borner & Burmester, 2017)

Contamination by DNA from external sources (e.g., cloning vectors or human DNA) is a common problem in NGS projects (Naccache et al., 2013; Laurence et al., 2014; Salter et al., 2014). If the contaminating sequences are not identified and remain in the datasets after sequence assembly and deposition into public databases, subsequent analyses may yield confusing results that can lead to false conclusions (Merchant et al., 2014; Tao et al., 2015). While several bioinformatic tools have been developed to identify and remove typical contaminants (e.g., Schmieder & Edwards, 2011; Jun et al., 2012), they are not suited for the identification of unexpected sources, such as pathogens infecting the sequenced organism. When working with wild animals, it is practically impossible to rule out infection by an unknown pathogen prior to sequencing. Moreover, the identification of parasite-derived contaminations may also enable the discovery of novel parasite lineages and shed light on previously unknown host-parasite associations. A number of studies have found evidence of endoparasite DNA in NGS data from humans (Strong et al., 2014) and animals (Orosz, 2015; Zhu et al., 2016b). However, these studies focused on small numbers of genes that are specific to the parasites of interest, while the majority of parasite-derived sequences remained unidentified. Therefore, the development of generalized bioinformatic approaches for the identification of parasite contaminations is of great importance.

In order to quantify the extent of contamination by apicomplexan parasites in the public genome and transcriptome databases and to extract as many parasite-derived contigs from the contaminated animal assemblies, I developed a software pipeline (ContamFinder) that uses a series of sequence similarity searches to identify contigs of parasite origin (Borner & Burmester, 2017). Due to the vast amounts of data generated by NGS projects and the enormous size of the public databases, a simple blastx all-vs-all search to identify contaminating sequences is not feasible for large numbers of genome and transcriptome assemblies, as the required computational resources would exceed even the limits of high-performance computer centers because blastx-style (translated nucleotide vs. protein) searches against large protein databases, such as Uniprot, are very computationally intensive, especially when using large genomic contigs as query. ContamFinder drastically reduces the computational complexity of this problem by first filtering out contigs with significant sequence similarity to known parasite proteins (Fig. 7A). Subsequent homology-based gene prediction further improves the performance of the search strategy by discarding non-coding regions (Fig. 7B) and allowing for protein vs. protein searches (Fig. 7C), which are significantly faster than using the full-length nucleotide contigs as query (Fig. 7D). Employing high-throughput local alignment tools (Suzuki et al., 2014) for the sequence similarity searches, ContamFinder achieved a more than 700-fold reduction in computation time compared to a simple blastx all-vs-all search. This massive improvement in performance allowed us to scan all publicly available genome and transcriptome assemblies from terrestrial animals. In total, 953 assemblies were analysed and, in 51 assemblies, a combined 20,907 contigs of apicomplexan origin were found. The contaminating parasite species were identified as members of the apicomplexan taxa Gregarinasina, Coccidia, Piroplasmida, and Haemosporida. Most contaminated assemblies contained only low to moderate numbers of parasite-derived sequences. From some assemblies, however, ContamFinder was able to extract several thousands of contigs, representing large amounts of the parasite's gene repertoire. For example, in the platypus genome assembly, we found a high number of contigs derived from a piroplasmid parasite (*Theileria*

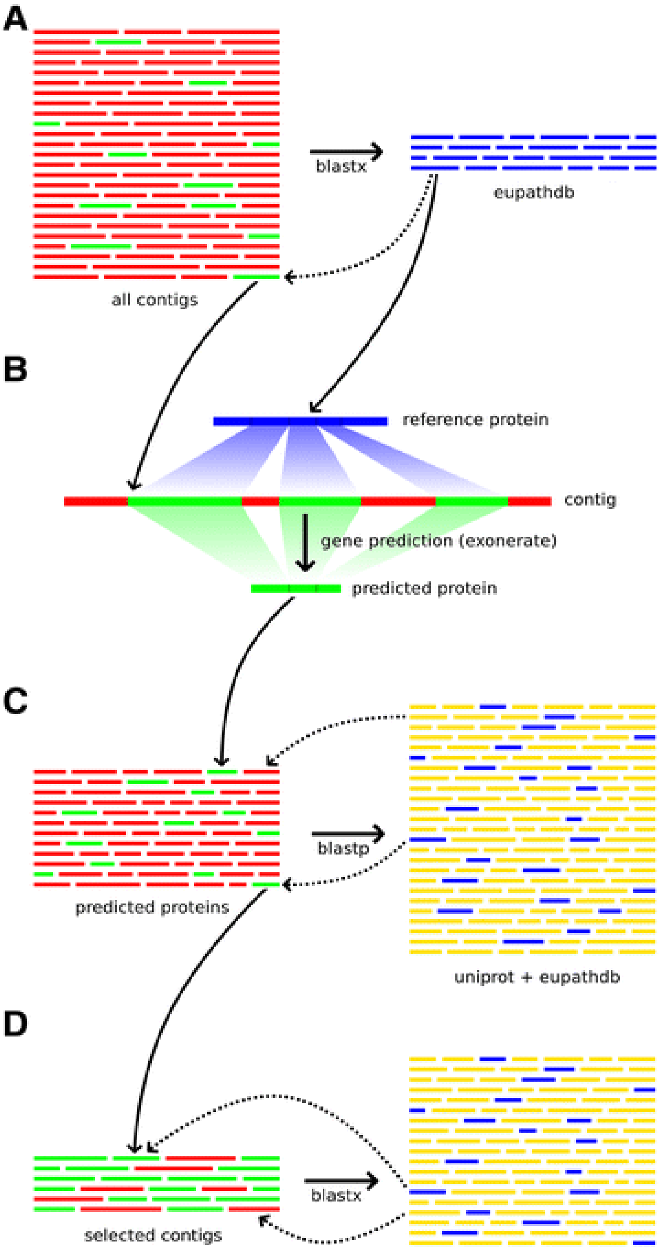


Fig. 7. Schematic overview of the ContamFinder pipeline (Borner & Burmester, 2017). (A) All contigs are searched against apicomplexan proteomes from the Eukaryotic Pathogen Database (EuPathDB; Aurrecoechea et al., 2011); contigs without significant hit are discarded. (B) Amino acid sequences are predicted using the best hitting apicomplexan protein; low complexity regions and repeats are masked. (C) Predicted amino acid sequences are searched against EuPathDB and UniProt; contigs with best hit outside of Apicomplexa are discarded. (D) Unprocessed contigs are searched against EuPathDB and UniProt; contigs with best hit outside of Apicomplexa are discarded. Contigs and sequence regions that were kept and used in the next step are shown in green, sequences that were discarded in red. Parasite-derived proteins in the search database are shown in blue, others in yellow.

ornithorhynchi). We also found massive amounts of sequences from gregarine parasites in multiple arthropod transcriptomes and from a coccidian parasite in the genome of the northern bobwhite (*Colinus virginianus*). For most of the infecting parasite species, no molecular data had been available previously. These results show that parasite-derived contaminations in genome and transcriptome data are not just a problem to be eliminated but also represent a valuable, cost-efficient source of information that can help to discover new parasites and provide information on previously unknown host-parasite interactions.

2.1.3 Automated primer design for phylogenetic datasets

(based on Hartig et al., 2012; Borner et al., 2016)

Despite the popularity of NGS techniques for phylogenomic approaches, targeted amplification of single-copy genes has remained a cornerstone of molecular phylogenetics (e.g., Schoch et al., 2011; Redmond et al., 2013; Schaer et al., 2013; Fuerst et al., 2015). While the cost per base is much lower for NGS projects, each individual sequencing run represents a substantial investment. Therefore, achieving a diverse taxon sampling can become cost prohibitive. Furthermore, the untargeted nature of shotgun sequencing approaches means that the majority of generated sequences will not be suitable for phylogenetic inference – though, once uploaded to the public databases, they constitute a valuable resource for a broad range of biological studies. These issues are especially true for samples from which RNA is not available (e.g., material from historical scientific collections), as whole genome sequencing is significantly more costly compared to transcriptome sequencing.

Regier et al. (2010) used a PCR-based approach to obtain data for 62 single-copy nuclear genes in a study on arthropod phylogeny. However, most studies relying on PCR amplification strategies have focused on small numbers of standard genes (mostly of mitochondrial or ribosomal origin), which are comparatively easy to amplify across a wide range of species but may not contain sufficient phylogenetic signal to resolve deep phylogenetic relationships (Springer et al., 2001). A major obstacle for the adoption of PCR-based approaches targeting large numbers of genes has been the development of oligonucleotide primers able to amplify nuclear genes from a diverse set of target species. To alleviate this problem, I have developed a bioinformatic pipeline that automates all steps of primer design for the amplification of nuclear coding sequences. The software searches for conserved regions in aligned protein-coding nucleotide sequences and scores potential oligonucleotide primer pairs based on parameters such as degree of degeneration, GC content, number of nucleotide repeats, melting temperature, and amplicon length. It also predicts the secondary structure of the oligonucleotides and calculates the hybridization energies of homo- and heterodimers. Optionally, multiple reference genomes can be searched for matches against the best scoring primer pairs. This allows estimating the actual length and intron content of each amplicon. To demonstrate the effectiveness of this approach, the primer design pipeline was run on 4,145 alignments of single copy genes from nine hymenopteran genomes (Hartig et al., 2012). Despite employing strict parameters for the quality of the oligonucleotide sequences, the software was able to infer 304 non-overlapping primer pairs for the amplification of sequence fragments from a total of 154 genes. To assess the viability of the primer sequences, ten pairs were randomly chosen and empirically tested on extracted DNA from six hymenopteran species. As expected, the success rate was significantly higher for species that were closely related to a reference species on which the primer design was based. For the five ingroup species, the primers were highly successful in amplifying the targeted DNA fragments (~80% success rate), whereas, for the single outgroup species, the success rate dropped to 30%. Extrapolating these results and considering that on average two primer pairs per gene were generated, ~150 genes of interest should be amplifiable in DNA samples from ingroup hymenopterans.

The application of the primer design pipeline to obtain nuclear sequence data from malaria parasites and related genera (Haemosporida) proved significantly more challenging. Since fully sequenced genomes were only available for mammalian species of the genus *Plasmodium*, the design of primers capable of amplifying gene fragments from the other haemosporidian genera had to be based on a severely restricted database. Furthermore, the pipeline had to be expanded to allow for the design of nested primer pairs to increase the specificity of the PCR, because birds and reptiles have nucleated red blood cells, which causes high levels of contamination by host DNA in the samples. Despite these challenges, the primer design yielded oligonucleotides capable of amplifying sequence fragments from 21 single copy genes across a wide range of haemosporidian lineages (Borner et al., 2016). Furthermore, the primer design pipeline has also been successfully employed to generate oligonucleotides for quantitative real-time PCR (Hoff et al., 2016; Fabrizius et al., 2016; Hoff et al., 2017), thus proving the versatility of the software.

2.2 Phylogeny of Ecdysozoa with focus on Arthropoda

2.2.1 The deep phylogeny of Ecdysozoa

(based on Borner et al., 2014)

The Ecdysozoa concept (Aguinaldo et al., 1997) was initially received with considerable skepticism and controversy (see Introduction) as it contradicted traditional animal systematics, which had grouped animal phyla according to similarities in their body plans. The monophyly of Ecdysozoa requires that basic aspects of animal body plans, such as segmentation or the presence of a body cavity with mesodermal epithelium (coelom), have either evolved convergently in multiple animal clades or were, to some extent, part of the original bilaterian body plan and had subsequently been lost several times in the course of evolution. Yet, the Ecdysozoa concept has found overwhelming support from recent morphological and molecular phylogenetic studies (see Introduction). It is now widely accepted in the scientific community and has found its way into major zoological textbooks as the standard view on protostome relationships (e.g., Burda et al., 2008; Reece et al., 2011; Westheide & Rieger, 2013).

Due to the high costs initially associated with obtaining NGS genome or transcriptome data, the taxon sampling of most phylogenomic studies has been strongly biased towards model species (e.g., *Drosophila melanogaster* or *Caenorhabditis elegans*) and species of medical (i.e., endo- and ectoparasites) or agricultural importance (i.e. pest species). The poor resolution of deep-level ecdysozoan relationships is most likely due to the lack of data from phylogenetically important taxa. While datasets based on mitochondrial sequences often had a more extensive taxon sampling, mitochondrial genes are not well suited for the inference of deep-level phylogeny (Sota & Vogler, 2001; Springer et al., 2001). To improve the taxon sampling of phylogenomic analyses, new transcriptome data from eight ecdysozoan species belonging to previously undersampled taxa were generated (Borner et al., 2014). Chelicerate transcriptomes were obtained from five specimens belonging to the previously neglected orders Solifugae (sun spiders), Uropygi (whip scorpions), Amblypygi (whip spiders), Opiliones (harvestmen), and Pseudoscorpiones (false scorpions). Additionally, three transcriptomes were sequenced from the ecdysozoan phyla Tardigrada, Priapulida, and Kinorhyncha. Data from 38 publicly available ecdysozoan genome and transcriptome sequencing projects were added, as well as data from 13 outgroup species. Phylogenetic analyses of the final dataset, which comprised 189 genes from 63 species, found strong support for the monophyly of Ecdysozoa (Fig. 8). All analyses recovered the scalidophoran taxa Priapulida (penis worms) and Kinorhyncha (mud dragons) in a sister group relationship at the base of Ecdysozoa. This topology is at odds with the "Cycloneuralia" hypothesis which postulates a common origin of Scalidophora and Nematoida (Nematoda and Nematomorpha) united by the possession of a circumpharyngeal nerve-ring (Ahlrichs, 1995; Schmidt-Rhaesa, 2012). However, support for "Cycloneuralia" from phylogenomic analyses is limited to a single study (Dunn et al., 2008). Other, more recent molecular studies have also favored a basal position of the included scalidophoran taxa (Campbell et al., 2011; Rota-Stabelli et al., 2013). It should be noted, however, that the third scalidophoran phylum, the Loricifera, has not been included in any phylogenomic studies. Until data from this group become available, the taxonomic status of Scalidophora must remain unclear, as phylogenetic analyses of 18S and 28S rRNA cast doubt on the monophyly of the taxon (Park et al., 2006; Yamasaki et al., 2015).

Another contentious issue is the position of Tardigrada (water bears). Based on several arthropod-like morphological characters, such as a segmented body, possession of limbs, and a ladder-like central nervous system, tardigrades have traditionally been united with Arthropoda and Onychophora (velvet worms) in a taxon called Panarthropoda (e.g., Westheide & Rieger, 1996; Brusca & Brusca, 2003). Yet, most molecular analyses recovered Tardigrada more closely related to Nematoda (Giribet, 2003; Roeding et al., 2007; Dunn et al., 2008; Lartillot & Philippe, 2008; Meusemann et al., 2010). This topology was also supported by all analyses of the full dataset of Borner et al. (2014). However, the results were not entirely conclusive, as tree inference based on a subset of only slowly evolving genes favored an arthropod association of tardigrades. The nematode affinity may, in fact, be attributed to LBA (Rota-Stabelli et al., 2011; Campbell et al., 2011). The monophyly of Panarthropoda with the inclusion of Tardigrada is supported by multiple lines of evidence, i.e. a unique shared microRNA (Campbell et al., 2011), shared structures of the nervous system (Mayer et al., 2013), and engrailed expression patterns (Gabriel & Goldstein, 2007). Some palaeontologists have even considered tardigrades as "stem-group arthropods" (Budd, 2001), and thus to be more closely related to the extant euarthropods than Onychophora are. However, a sister group relationship between

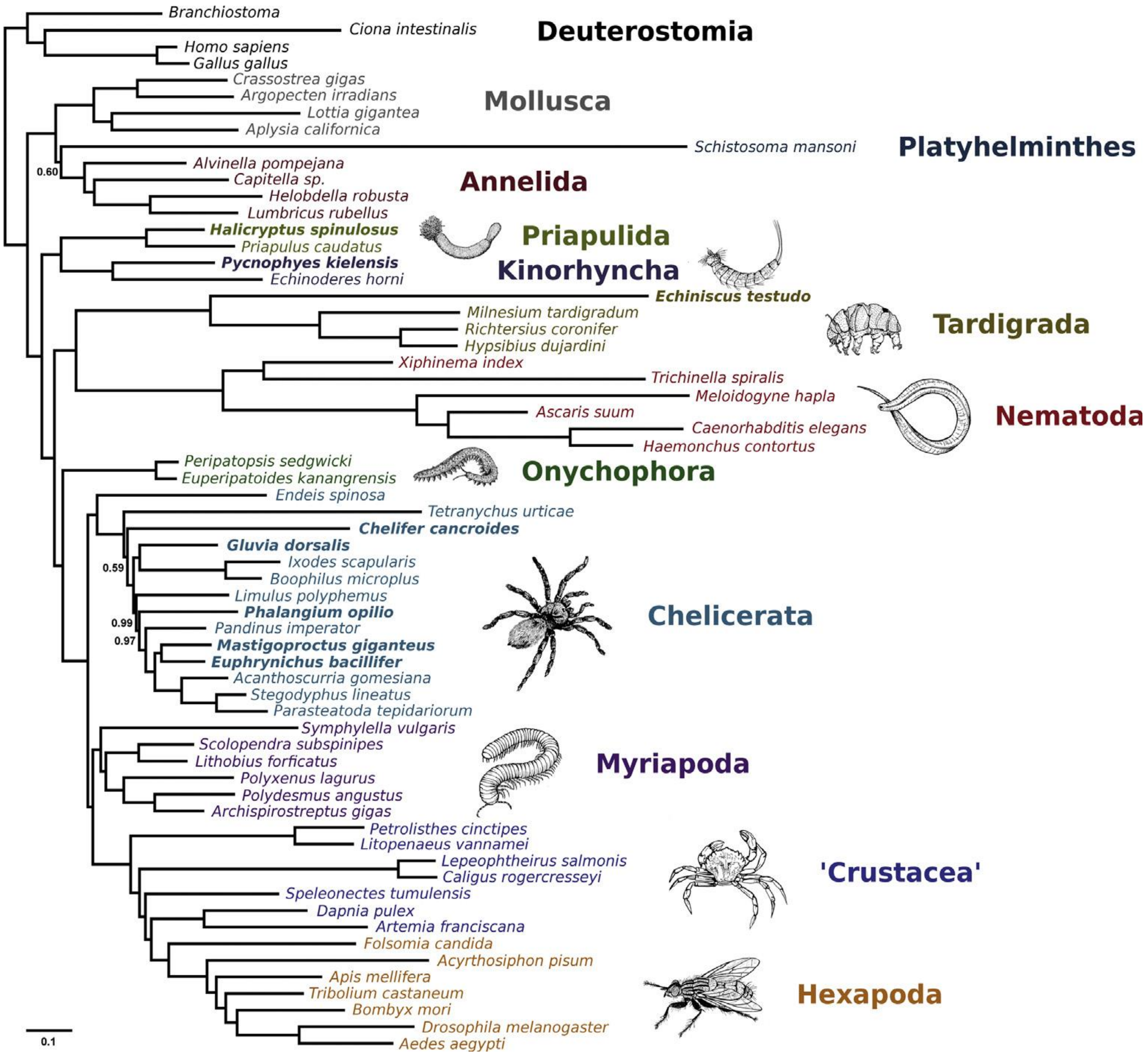


Fig. 8. Ecdysozoan phylogeny based on a Bayesian analysis of 189 genes from 63 taxa (Borner et al., 2014). Bayesian posterior probabilities <1.00 are given at the nodes; all other splits have a posterior probability of 1.00. Species that were sequenced specifically for this study are denoted in bold letters.

Tardigrada and Euarthropoda (together referred to as "Tactopoda") appears unlikely, as it has not been recovered in any phylogenomic studies, including those which supported monophyletic Panarthropoda (Rota-Stabelli et al., 2011; Campbell et al., 2011).

Within Euarthropoda, competing hypotheses have been suggested concerning the position of Myriapoda. While studies based on morphological evidence strongly favored a common origin of Myriapoda, Crustacea, and Hexapoda (Mandibulata hypothesis; see Westheide & Rieger, 1996; Brusca

& Brusca, 2003), several molecular phylogenetic studies initially found a sister group relationship between Myriapoda and Chelicerata ("Myriochelata" hypothesis; Hwang et al., 2001; Pisani et al., 2004; Mallatt et al., 2004; Dunn et al., 2008; Meusemann et al., 2010). In our analyses (Borner et al., 2014), Mandibulata were recovered as a valid (monophyletic) taxon (Fig. 8). This result is in line with other recent molecular studies (Regier et al., 2010; Rota-Stabelli et al., 2011; Giribet & Edgecombe, 2012; Chipman et al., 2014; Lozano-Fernandez et al., 2016) and may be attributed to improvements in taxon sampling and the application of phylogenetic methods that are more robust to the effects of LBA (Rota-Stabelli et al., 2011). Considering that Mandibulata also received support from recent studies on *Hox* gene expression (Janssen et al., 2014; Pace et al., 2016), neurogenesis (Stollewerk, 2016), and embryology (Chipman, 2015), it appears that a consensus in favor of the Mandibulata hypothesis has been reached in the scientific community.

Within Mandibulata, a close relationship of hexapods and crustaceans (together Pancrustacea or Tetraconata) has consistently been recovered in studies based on molecular data (e.g., Friedrich & Tautz, 1995; Boore et al., 1998; Kusche & Burmester, 2001; Dunn et al., 2008: Meusemann et al., 2010) and has found increasing support from morphological studies as well (Richter, 2002; Harzsch, 2004; Strausfeld, 2009; Strausfeld et al., 2011). Most molecular studies have placed Hexapoda nested within paraphyletic "Crustacea" (e.g., Wilson et al., 2000; Regier et al., 2005; Ertas et al., 2009; von Reumont et al., 2012), though the identity of the crustacean lineage that is most closely related to Hexapoda is still controversial. Our analyses (Borner et al., 2014) recovered Branchiopoda as the sister group of Hexapoda (Fig. 8). However, there is also strong evidence for a close relationship of Remipedia and Hexapoda (Ertas et al., 2009; Regier et al., 2010; von Reumont et al., 2012).

2.2.2 Myriapod relationships

(based on Rehm et al., 2014)

The taxonomic status of Myriapoda has long been subject of intense discussion. Based on molecular data, the traditional view of Myriapoda being paraphyletic with regard to Hexapoda has been rejected. However, some molecular studies have also failed to recover monophyletic Myriapoda (Negrisolo et al., 2004; von Reumont et al., 2009). While most studies in recent years have supported the monophyly of the taxon (e.g., Regier et al., 2010; Miyazawa et al., 2014; Lozano-Fernandez et al., 2016), the internal relationships among myriapod classes have remained poorly resolved. To improve our understanding of the evolutionary history of the taxon, transcriptomes from three diplopods, two chilopods, and a symphylan were sequenced. Phylogenetic analyses provided strong support for monophyletic Myriapoda as sister group of Pancrustacea (Rehm et al., 2014). Within Myriapoda, surprisingly, a sister group relationship between Chilopoda and Diplopoda was recovered and Symphyla were placed at the base of the taxon. This topology has not been proposed before (neither based on morphology nor based on molecular data). However, it has gained some support since publication and certain morphological characters fit such a grouping, as noted by Lozano-Fernandez et al. (2016). Both taxa possess a series of imbricated comb lamellae on the mandibles, a character that was proposed as a potential myriapod autapomorphy despite being absent in symphylans and pauropods (Edgecombe & Giribet, 2002). The analyses of Borner et al. (2014) and another study based on three nuclear genes (Miyazawa et al., 2014) independently recovered basal symphylans and a close relationship between chilopods and diplopods. More recently, Lozano-Fernandez et al. (2016) have significantly expanded the phylogenomic taxon sampling of Myriapoda – although data from pauropod species were still lacking. Using different datasets and phylogenetic methods, the authors

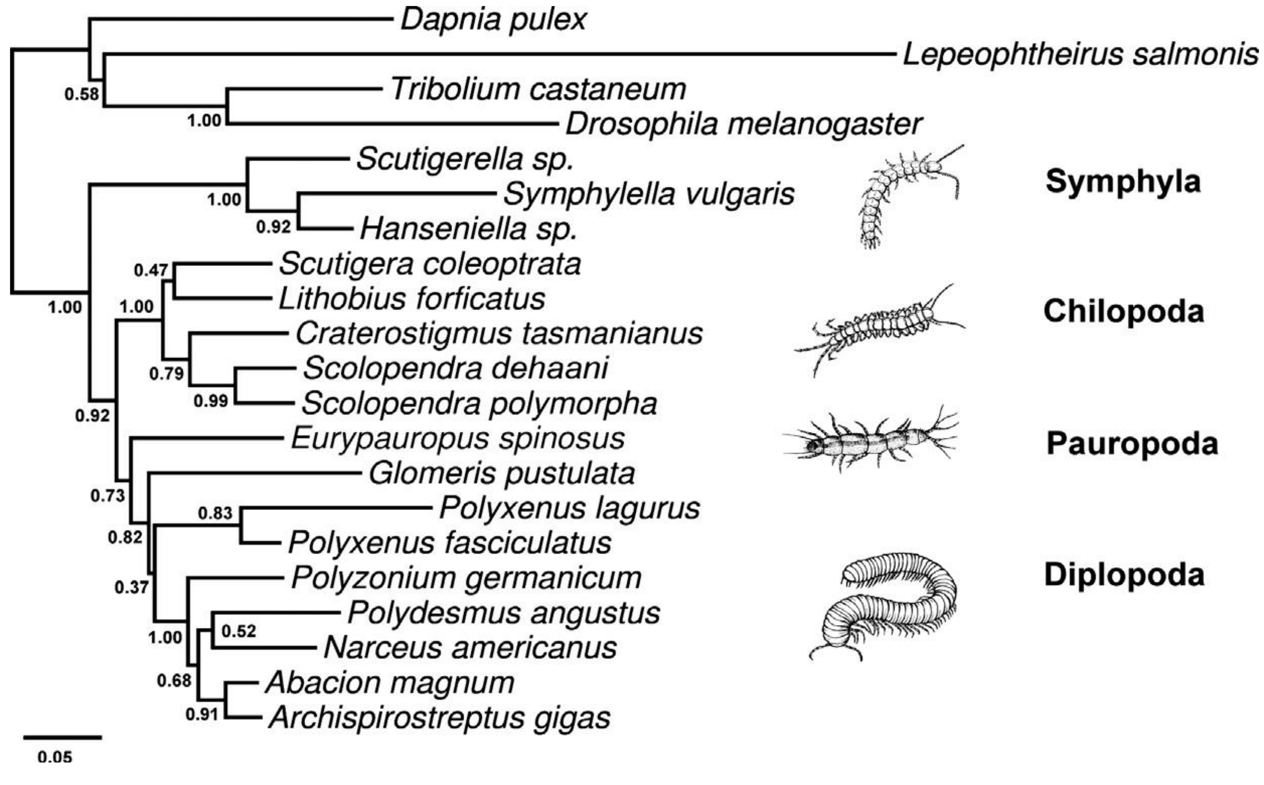


Fig. 9. Myriapod phylogeny based on a 22,339 amino acid alignment from 21 taxa (Rehm et al., 2014). Bayesian posterior probabilities are depicted at the nodes.

found two alternative topologies regarding the relationships among myriapod classes: Symphyla were either united with Diplopoda as predicted by the Progoneata hypothesis, or Chilopoda and Diplopoda formed a common clade to the exclusion of Symphyla, thus supporting the phylogeny of Rehm et al. (2014).

Due to the lack of NGS data, the position of Pauropoda has remained poorly resolved. Using a PCR-based approach, Regier et al. (2010), found a sister group relationship of Pauropoda and Symphyla (together "Edafopoda"). In a combined analysis that included the single gene data from Regier et al. (2010) in addition to the NGS data from Rehm et al. (2014), Pauropoda were found closely associated with Diplopoda (Fig. 9), thus supporting the Dignatha-hypothesis favored by most morphologists (Tiegs, 1947; Dohle, 1980). However, the deep-level relationships were poorly resolved in this tree due to the large amounts of missing data for the species from Regier et al. (2010).

2.2.3 Chelicerate relationships

(based on Rehm et al., 2012; Dunlop et al., 2014; Borner et al., 2014)

While it appears that a consensus is beginning to emerge for most aspects of Ecdysozoan phylogeny, there is surprisingly little agreement on the relationships among major chelicerate lineages. The majority of recent morphological and molecular studies have supported the inclusion Pycnogonida (sea spiders) in Chelicerata, placing them in a sister group relationship with Euchelicerata (Brenneis et al., 2008; Dunn et al., 2008; Meusemann et al., 2010). Within Euchelicerata, however, there is a high degree of discordance between studies based on molecular data and studies based on morphological evidence. The lack of nuclear sequence data from most chelicerate lineages did not allow for phylogenomic inference of chelicerate relations until recent years. While several genome and transcriptome sequencing projects had provided data from ticks (Parasitiformes) and mites (Acariformes), which are of medical and agricultural importance as vectors of human disease and pest species of plants, the other chelicerate orders had essentially been neglected. Transcriptomes from five of these orders were sequenced to enable phylogenomic analyses of Chelicerata (Dunlop et al., 2014; Borner et al., 2014). All analyses found strong support for monophyletic Chelicerata and a sister group relationship of Pycnogonida and Euchelicerata (Fig. 10). Within Euchelicerata, none of the analyses recovered monophyletic Arachnida, a taxon uniting all extant primarily terrestrial chelicerates to the exclusion of the marine Xiphosura (horseshoe crabs). Arachnida are considered as one of the best

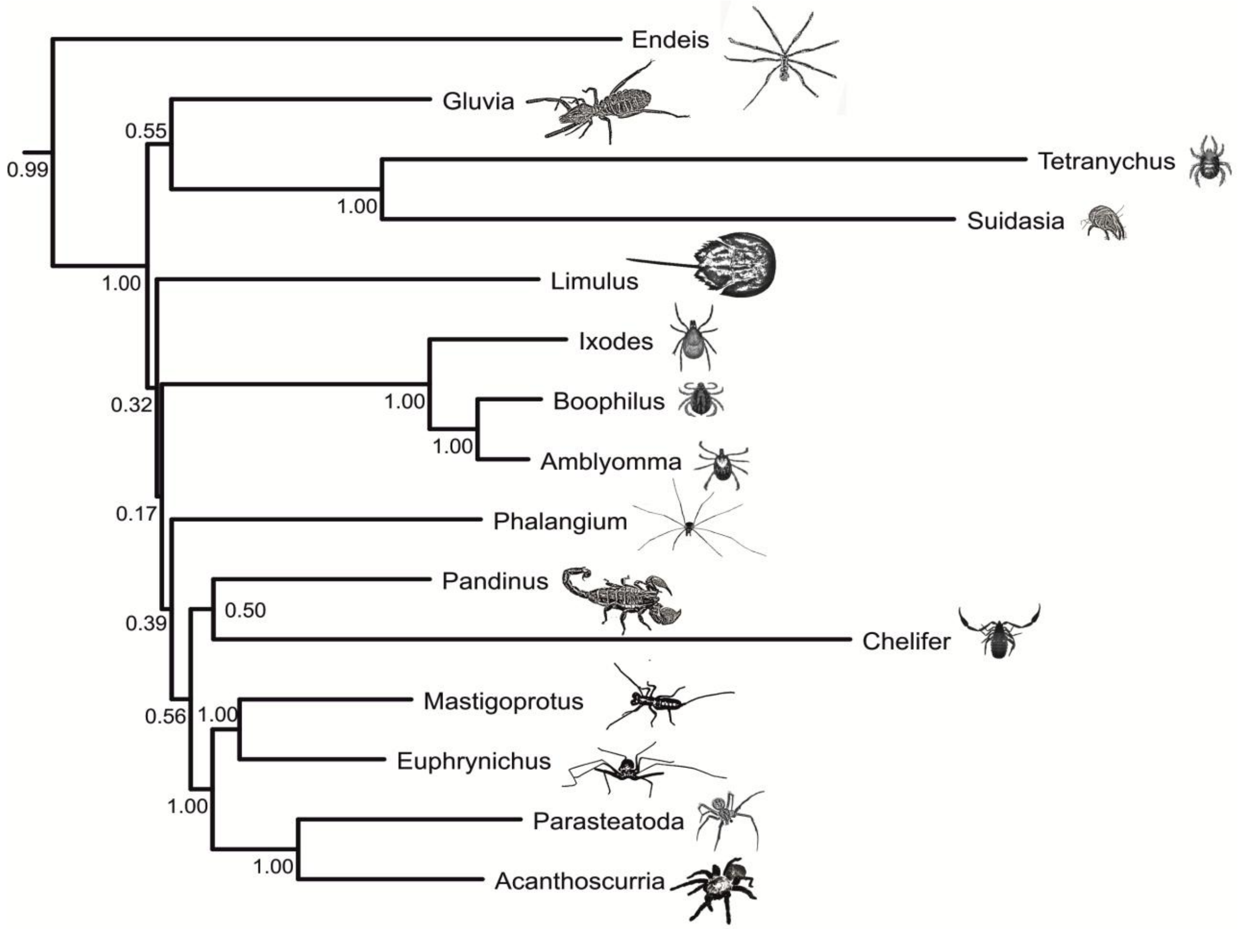


Fig. 10. Chelicerate phylogeny based on a Bayesian analysis of 197 genes from 15 chelicerate taxa (Dunlop et al., 2014). The numbers at the nodes represent the posterior probabilities.

supported chelicerate taxa by most morphologists. Yet, support for this taxon from molecular data is limited to a few studies and is hardly convincing. Most analyses of Regier et al. (2010) recovered Arachnida as the sister group of Xiphosura. However, support for this grouping was low – in fact, all deep-level relationships among euchelicerate orders were essentially unresolved.

Shortly after release of the data from Borner et al. (2014), another transcriptome-based study on chelicerate phylogeny was published (Sharma et al., 2014). In all analyses, the authors found a highly supported clade comprising Scorpiones, Pedipalpi (Amblypygi and Uropygi), and Araneae, while the positions of the remaining arachnid taxa were highly unstable and paraphyletic Acari were recovered at the base of the euchelicerate tree. These findings are all in line with the results of Dunlop et al. (2014) and Borner et al. (2014). This convergence of results is especially noteworthy because the datasets employed in these studies were generated independently and are not based on the same sequencing data due to the short succession of publication. In contrast to the results of Dunlop et al. (2014) and Borner et al. (2014), Sharma et al. (2014) found some support for Arachnida after reducing the phylogenetic dataset to the 500 slowest evolving genes. While it is interesting to note that a subset of genes supported Arachnida in maximum likelihood analyses, this result should be interpreted with caution, as Bayesian inference on the same subset failed to recover monophyletic Arachnida and, after further reduction to the 200 slowest-evolving genes, support for this topology disappeared also in the maximum likelihood analyses. Fast evolutionary changes in long branching taxa, such as Acariformes and Pseudoscorpiones, may in part explain the lack of resolution and instability of clades at the base of Euchelicerata. Alternatively, early divergence events may have occurred in quick succession within a relatively short time span of euchelicerate evolution.

Hemocyanins are the respiratory proteins of many arthropods and molluscs. Although arthropod and molluscan hemocyanins share some similarities in the structure of their active sites (both are large copper-proteins that are able to reversibly bind O_2), they are of independent evolutionary origin (Burmester, 2001; van Holde et al., 2001). In several studies, hemocyanin sequences have proven to be well suited for the inference of phylogenetic relationships within Arthropoda (e.g., Burmester, 2001; Kusche & Burmester, 2001; Ertas et al., 2009). To infer the evolutionary history of chelicerate hemocyanins, sequences from a sea spider, a scorpion, a whip scorpion, and a whip spider were sequenced (Rehm et al., 2012). Publicly available data from web spiders and xiphosurans were added to the dataset. While the sea spider has a simple hexameric hemocyanin, four distinct subunit types evolved before the divergence of Xiphosura and Arachnida. Phylogenetic analyses showed that the distinct subunits in each of the 8 × 6mer hemocyanin of Xiphosura and the 4 × 6mer of Arachnida evolved through subsequent independent gene duplication events. The phylogenetic relationships within the different subunit types support a basal position of Pycnogonida, a sister group relationship of Xiphosura and Arachnida, and monophyletic Pedipalpi (Amblypygi + Uropygi) closely related to Araneae. These results are fully congruent with the findings of Dunlop et al. (2014) and Borner et al. (2014). Unfortunately, hemocyanin has been (independently) lost in those chelicerate taxa that were unstable in the phylogenomic analyses of these studies, namely Opiliones, Pseudoscorpiones, Solifugae and Acari. This loss of a respiratory protein may be explained by the evolution of trachea or the miniscule size of some species.

2.2.4 Dating the arthropod tree

(based on Rehm et al., 2011; Rehm et al., 2012; Rehm et al., 2014)

Understanding the timeline of evolutionary events may allow researchers to address questions of general biological importance. To interpret the results of deep phylogenetic analyses in the correct geological and palaeontological context, it is necessary to obtain approximate dates for the divergence events of interest. Initially, our knowledge of evolutionary timescales relied entirely on the fossil record. Under the right geological conditions, radiometric or stratigraphic methods can confidently determine the age of a fossil with high accuracy (Martin et al., 2000). However, reliable taxonomic assignment of fossils is often difficult and the fossil record is far for from being complete (Benton & Donoghue, 2007). The development of molecular clock methods has allowed researchers to infer evolutionary timescales using genetic data. The original molecular clock concept (Zuckerkandl & Pauling, 1965) was based on the assumption of a constant rate of genetic change among lineages, such that it would be possible to determine the time elapsed since two taxa diverged based on the amount of accumulated nucleotide substitutions. This assumption has proven untenable, however, as ample evidence of rate variation among taxa has been described (see review by Lanfear et al. [2010]). These findings have motivated the development of relaxed molecular clock models which allow for variable rates of molecular evolution (e.g., Sanderson, 1997; Thorne et al., 1998; Drummond et al., 2006; Lepage et al., 2006).

To illuminate the timescale of arthropod evolution, molecular clock analyses (Rehm et al., 2011) were performed based on a superalignment of 37,476 amino acid positions, which had been derived from Expressed Sequence Tags (ESTs) in a previous study (Meusemann et al., 2010). At the time of publication, this was the largest dataset ever used in a molecular clock study – and the first based on EST data. Previous multi-gene analyses (e.g., Aris-Brosou & Yang, 2003; Douzery et al., 2004; Blair & Hedges, 2005; Peterson et al., 2008) were based on genome data, and, due to the small number of sequenced animal genomes, were severely limited in taxon sampling, resulting in date estimates based on artifactual phylogenetic relationships (i.e., polyphyletic Ecdysozoa). They produced a wide range of discordant dates for the earliest metazoan divergence events (e.g., estimates for the emergence of bilaterians ranged from 670 to 1,300 million years ago [mya]). In all cases, the estimated dates were significantly older than the earliest conclusive fossil evidence for crown group bilaterians, which dates ~550–530 mya (Benton & Donoghue, 2007). The dates obtained in the molecular clock analysis of the EST dataset (Rehm et al., 2011) are notably younger than the estimates of most previous studies based on sequence data. However, the molecular dates still significantly predate the earliest fossil evidence. For example, the divergence of Onychophora and Euarthropoda was dated ~589 mya in the molecular clock approach (Fig. 11), whereas the earliest unambiguous euarthropod fossils are ~521 million years old (Crimes, 1987; Chen, 2009). Thus, the results of Rehm et al. (2011) are still not compatible with a Cambrian origin of arthropods. It should be noted, however, that a significant uncertainty is associated with molecular time estimates and the 95% confidence intervals of most deep-level splits reach well into the Cambrian. Furthermore, fossils can only provide upper bounds for the timing of divergence events. For a fossil to be assigned to a certain taxon with confidence, it must already have evolved morphological features that are characteristic of the extant members of this taxon – a process that may take significant evolutionary time. The gap between a putative Pre-Cambrian emergence of metazoan phyla and the sudden appearance of crown group fossils from

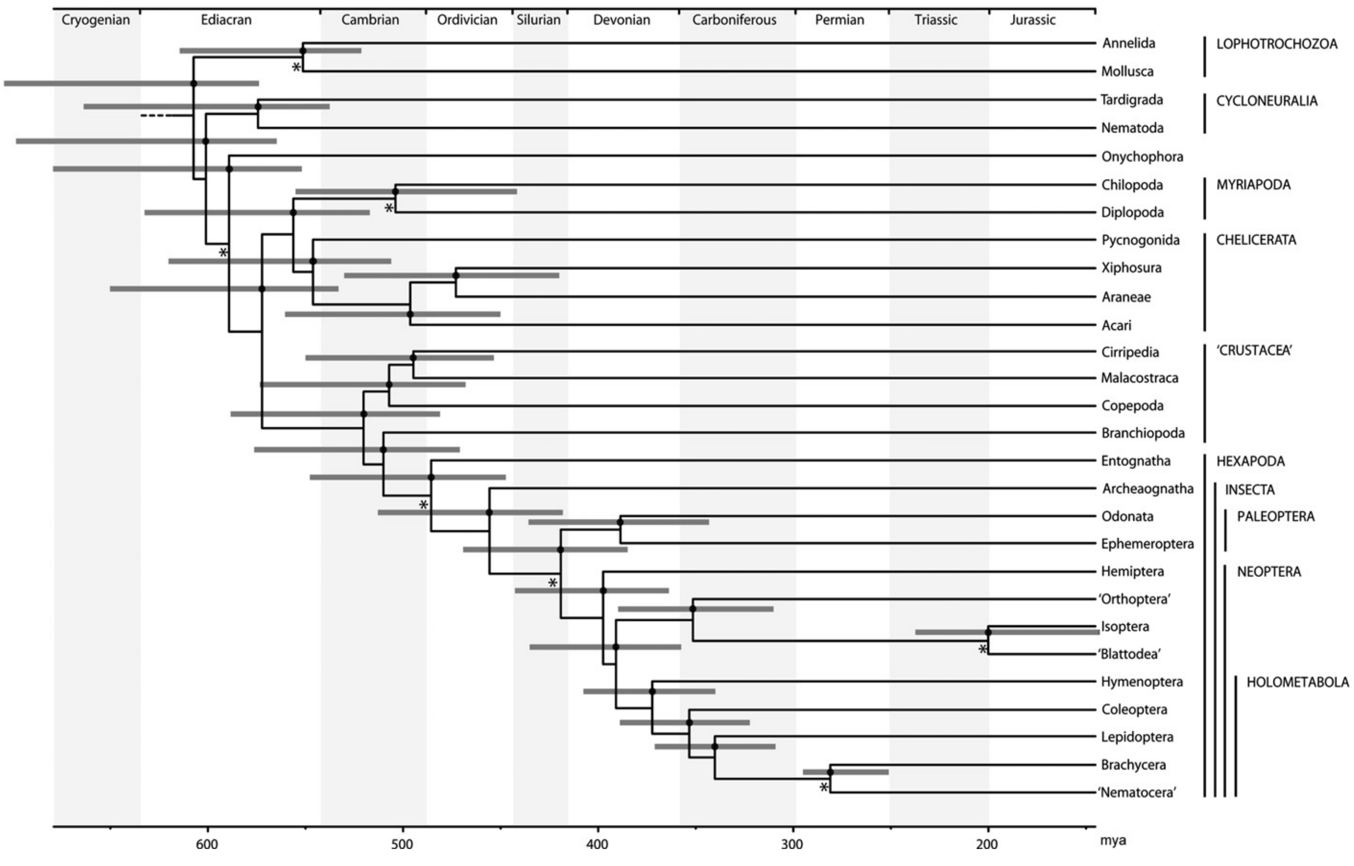


Fig. 11. Mean divergence times of major ecdysozoan taxa (Rehm et al., 2011). Mean divergence times were estimated under the log-normal autocorrelated clock model with PhyloBayes 3.3.f. Grey bars indicate 95% confidence intervals.

several metazoan lineages in the Cambrian may be explained by a period of cryptic evolution before an “explosive” radiation in the Cambrian. The ancestors of extant metazoans might have had limited geographic range and may therefore not be present in the small number of explored Pre-Cambrian Lagerstätten (Conway Morris, 1993; Fortey et al., 1996).

Within Chelicerata, the earliest divergence event (Pycnogonida – Euchelicerata) was dated ~546 mya. While this estimate still predates the earliest chelicerate fossil from a larval sea spider by ~50 million years, it is much closer to the fossil record than previous calculations (e.g., Regier et al. [2005] dated this event 813–632 mya). By contrast, the date for the divergence of the myriapod classes Chilopoda and Diplopoda (~504 mya) is significantly older than the estimates of some previous studies (e.g., ~442 mya; Pisani et al., 2004) and the evidence from the fossil record (~420 mya; Edgecombe & Giribet, 2007). In fact, it predates the emergence of land plants in the Middle Ordovician ~490 mya (Steemans et al., 2009; Rubinstein et al., 2010). Thus the early evolution and divergence of myriapod classes may have taken place in the ocean. However, this would require trachea to have evolved independently in Diplopoda and Myriapoda. Considering that molecular phylogenetics have convincingly shown that “Tracheata” are paraphyletic and myriapod and hexapod trachea have separate evolutionary origins, this scenario cannot easily be dismissed. However, Lozano-Fernandez et al. (2016) argued that ephemeral, terrestrial ecosystems have existed since approximately one billion years ago (Strother et al., 2011) and could potentially have supported myriapod life on land already in the Cambrian.

Obviously, molecular clock analyses can only produce reliable time estimates if the tree on which they are based is correct. The sister group relationship of Myriapoda and Chelicerata in the phylogeny of Meusemann et al. (2010) has been suggested to be an artifact caused by LBA (Rota-Stabelli et al., 2011), which raises some concern over the effect of this topology on the time estimates of Rehm et al. (2011). To investigate this, a molecular clock analysis was performed on a dataset with improved myriapod taxon sampling (Rehm et al., 2014), which resulted in support for Mandibulata instead of “Myriochelata”. The divergence time estimates from this study are fully congruent with the previous analysis. In fact, the estimated time for the divergence of Diplopoda and Chilopoda is slightly older (~515 mya vs. ~504 mya) in the analysis of Rehm et al. (2014), though well within the 95% confidence interval. The results of Rehm et al. (2011) were also corroborated by a molecular clock analysis of chelicerate hemocyanin sequences (Rehm et al., 2012), which dated the age of the earliest divergence within Chelicerata (Pycnogonida – Euchelicerata) to be ~543 mya (vs. ~546 mya in the study of Rehm et al. [2011]). However, it should be noted that all of these molecular clock calculations were based on similar sets of calibration points. Rehm et al. (2011) employed seven carefully selected calibration points that were evenly distributed throughout the phylogenetic tree, while non-calibrated splits were used to compare calculated dates to the age of informative fossils. By contrast, a subsequent large-scale molecular clock study on the timetree of ecdysozoan evolution (Rota-Stabeli et al., 2013) included as many calibration points as possible (78 in total) to maximize the information from the fossil record, though at the risk of overparameterization of the analysis. Despite these differences in methodology, the time estimates are very close to the results of Rehm et al. (2011) (Table 1). On average, the calculated divergence times of Rota-Stabeli et al. (2013) are ~2% younger. By contrast, a more recent molecular clock study on arthropod terrestrialization (Lozano-Fernandez et al., 2016) found slightly older dates (~3%) compared to the results of Rehm et al. (2011) (Table 1). However, all three studies support an Ediacaran origin of Arthropoda followed by diversification of the extant arthropod subphyla in the Cambrian.

Table 1. Comparison of mean divergence time estimates (in mya) from three molecular clock studies.

Crown group	**Panarthropoda**	**Myriapoda**	**Chelicerata**	**Pancrustacea**	**Hexapoda**
Divergence of:	Onychophora–Euarthropoda	Diplopoda–Chilopoda	Pycnogonida–Euchelicerata	Malacostraca–Branchiopoda	Collembola–Insecta
Rehm et al. (2011)	562	504	546	520	488
Rota-Stabeli et al. (2013)	543	510	526	511	483
Lozano-Fernandez et al. (2016)	606	528	552	577	468

2.3 Phylogeny of Apicomplexa with focus on Haemosporida

2.3.1 Parasite contaminations help illuminate the deep phylogeny of Apicomplexa

(based on Borner & Burmester, 2017)

The need for phylogenomic approaches to resolve the deepest nodes of the apicomplexan tree has long been recognized (see review by Morrison [2008]). Over the last decade, complete nuclear genomes have been sequenced from representatives of all known major apicomplexan lineages. However, the taxon selection is heavily biased towards parasite taxa of medical or veterinary importance. For instance, complete genomes from 16 mammalian malaria parasites (genus *Plasmodium*) are currently available in Genbank (accessed 29.03.2017), whereas the extremely diverse gregarine parasites of invertebrates (Gregarinasina) are represented by only a single species.

By extracting parasite-derived contigs from contaminated animal genome and transcriptome assemblies (Borner et al., 2017), we were able to obtain nuclear sequence data from more than 50 apicomplexan parasites (see 2.1.2). From these contaminating species, 32 proved suitable for the inclusion into a large phylogenomic dataset comprising 1,420 genes from 35 publicly available apicomplexan and chromerid genomes. Phylogenetic inference yielded a well-resolved tree (Fig. 11) that is in good agreement with recent molecular studies (Templeton et al., 2010; Arisue & Hashimoto, 2015). Thus, a consensus on the deep phylogeny of Apicomplexa appears to be emerging.

At the base of Apicomplexa, we found a sister group relationship between *Cryptosporidium* and the gregarines (Fig. 11). Both taxa have apparently lost the apicoplast genome and, possibly, the whole organelle (Zhu et al., 2000b; Toso & Omoto, 2007). Originally, the genus *Cryptosporidium* was assigned to Coccidia based on the parasites' life cycle within the digestive tract of vertebrates (Levine, 1988), whereas gregarines exclusively infect invertebrate species. Nevertheless, a common origin of the gregarines and *Cryptosporidium* is widely accepted by now and is supported by numerous molecular and physiological similarities (see review by Thompson et al. [2005]). Because of their basal position within Apicomplexa, Gregarinasina constitute a key taxon for understanding the evolutionary history of the phylum. Yet, the gregarines have essentially been neglected in genome sequencing efforts due to their lack of medical or veterinary importance. Large amounts of contaminating contigs from gregarine parasites were identified in several arthropod transcriptomes and the extracted sequences significantly increase the amount of publicly available sequence data from this taxon.

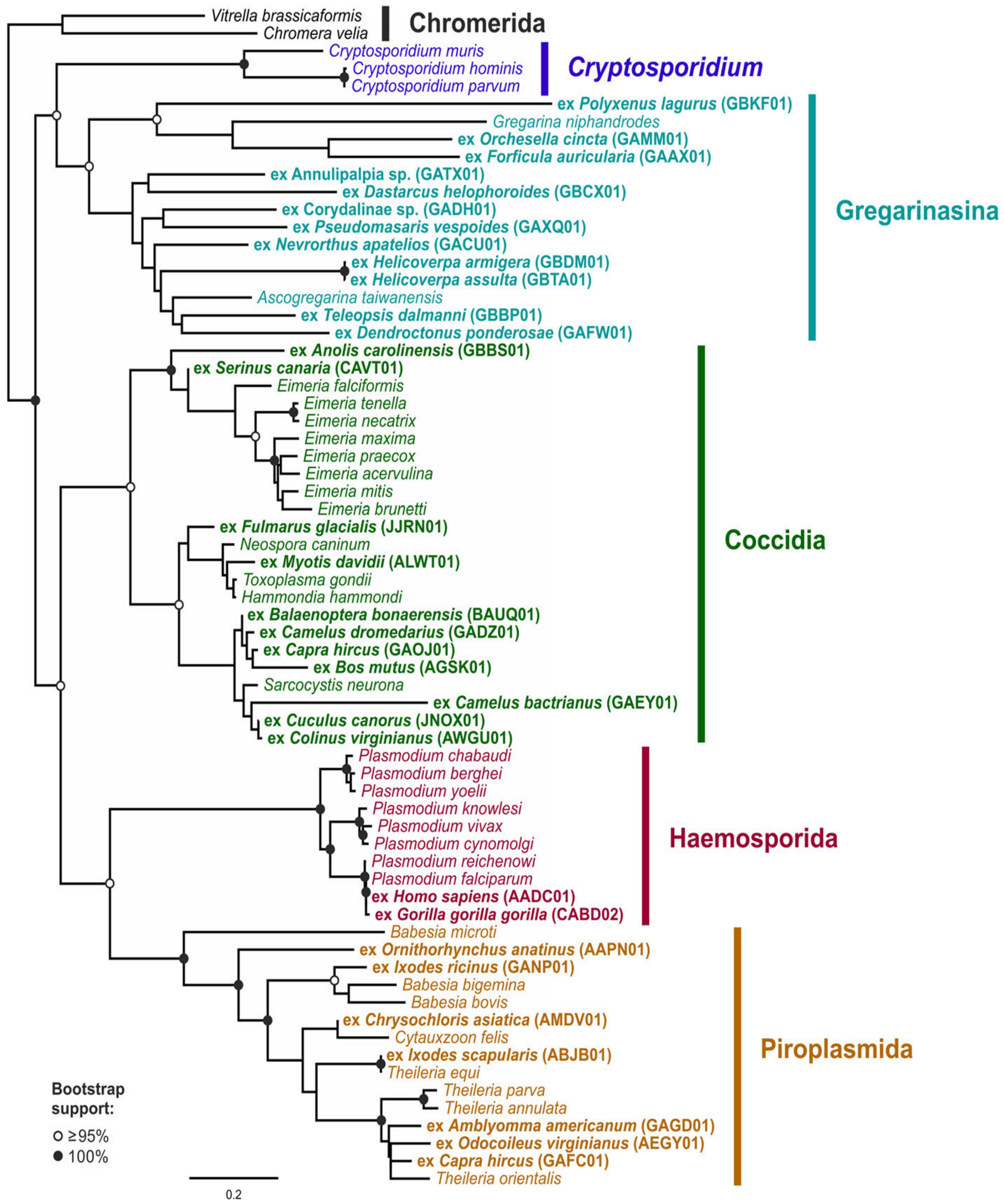

Vitrella brassicaformis
Chromera velia
Chromerida
Cryptosporidium muris
Cryptosporidium hominis
Cryptosporidium parvum
Cryptosporidium
ex Polyxenus lagurus (GBKF01)
Gregarina niphandrodes
ex Orchesella cincta (GAMM01)
ex Forficula auricularia (GAAX01)
ex Annulipalpia sp. (GATX01)
ex Dastarcus helophoroides (GBCX01)
ex Corydalinae sp. (GADH01)
ex Pseudomasaris vespoides (GAXQ01)
ex Nevrorthus apatelios (GACU01)
ex Helicoverpa armigera (GBDM01)
ex Helicoverpa assulta (GBTA01)
Ascogregarina taiwanensis
ex Teleopsis dalmanni (GBBP01)
ex Dendroctonus ponderosae (GAFW01)
Gregarinasina
ex Anolis carolinensis (GBBS01)
ex Serinus canaria (CAVT01)
Eimeria falciformis
Eimeria tenella
Eimeria necatrix
Eimeria maxima
Eimeria praecox
Eimeria acervulina
Eimeria mitis
Eimeria brunetti
ex Fulmarus glacialis (JJRN01)
Neospora caninum
ex Myotis davidii (ALWT01)
Toxoplasma gondii
Hammondia hammondi
ex Balaenoptera bonaerensis (BAUQ01)
ex Camelus dromedarius (GADZ01)
ex Capra hircus (GAOJ01)
ex Bos mutus (AGSK01)
Sarcocystis neurona
ex Camelus bactrianus (GAEY01)
ex Cuculus canorus (JNOX01)
ex Colinus virginianus (AWGU01)
Coccidia
Plasmodium chabaudi
Plasmodium berghei
Plasmodium yoelii
Plasmodium knowlesi
Plasmodium vivax
Plasmodium cynomolgi
Plasmodium reichenowi
Plasmodium falciparum
ex Homo sapiens (AADC01)
ex Gorilla gorilla gorilla (CABD02)
Haemosporida
Babesia microti
ex Ornithorhynchus anatinus (AAPN01)
ex Ixodes ricinus (GANP01)
Babesia bigemina
Babesia bovis
ex Chrysochloris asiatica (AMDV01)
Cytauxzoon felis
ex Ixodes scapularis (ABJB01)
Theileria equi
Theileria parva
Theileria annulata
ex Amblyomma americanum (GAGD01)
ex Odocoileus virginianus (AEGY01)
ex Capra hircus (GAFC01)
Theileria orientalis
Piroplasmida
Bootstrap support:
○ ≥95%
● 100%
0.2

Fig. 11. Apicomplexan phylogeny based on a maximum likelihood analysis of 1,420 genes from 35 complete apicomplexan and chromerid genomes and 32 contaminating parasite species (denoted in bold) extracted from animal genome and transcriptome assemblies (Borner et al., 2017).

Among the remaining apicomplexan taxa, Piroplasmida and Haemosporida were recovered in a sister group relationship to the exclusion of Coccidia (Fig. 11). Contigs derived from coccidian parasites were found in assemblies from a wide range of vertebrate hosts including birds, reptiles, a whale, and various other mammals. In the genome assemblies of the western lowland gorilla and a human sample, genetic sequences from the most malignant agent of human malaria, *Plasmodium falciparum*, were found. However, in the case of the gorilla genome, this is probably not the result of infection but rather contamination in the laboratory or sequencing center. Piroplasmid contaminations were found in the assemblies of both tick vectors and putative mammalian hosts. Especially noteworthy are the contaminations in the genome assemblies of the Cape golden mole (*Chrysochloris asiatica*) and the platypus (*Ornithorhynchus anatinus*). The sequences from *C. asiatica* constitute the first report of a piroplasmid infection in the order Afrosoricida. Unfortunately, due to the low number of extracted contigs from this species, the parasite's exact placement within Piroplasmida remained unresolved. The sequences extracted from the platypus allowed us to finally resolve the phylogenetic position of *Theileria ornithorhynchi* with high confidence. Its placement outside the clade of the theilerids and basal to all other piroplasms except *Babesia microti* is consistent with the tentative results of Paparini et al. (2015) based on 18S rRNA.

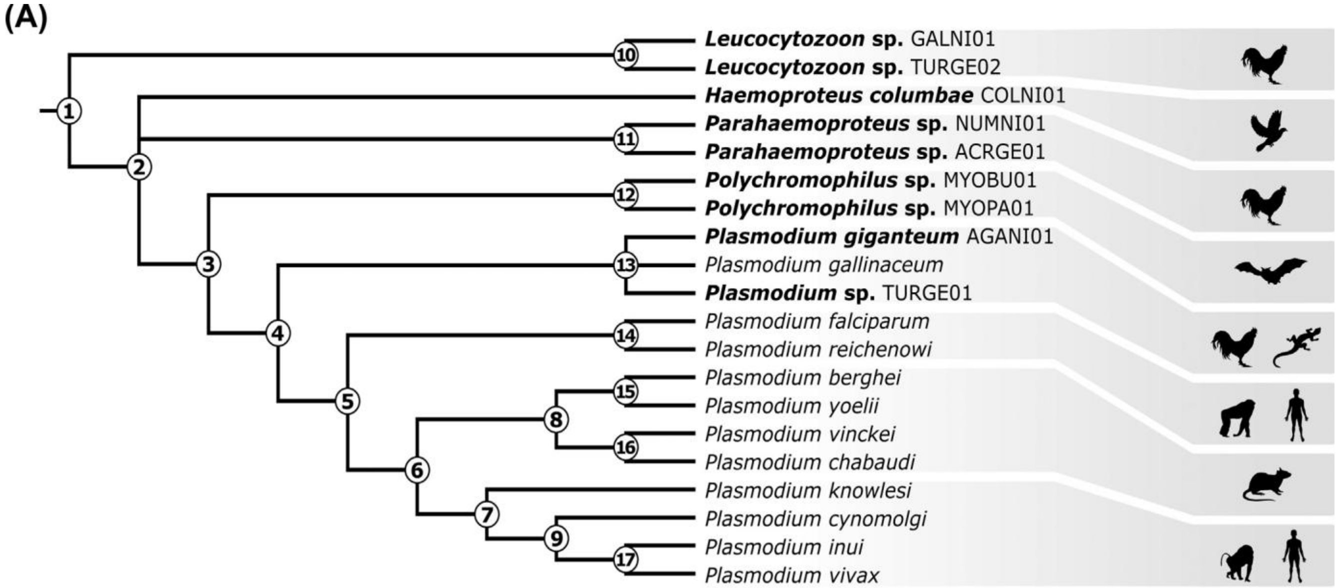
2.3.2 The phylogeny of haemosporidian parasites based on nuclear gene data

(based on Borner et al., 2016)

Despite being the focus of numerous studies, there is still no consensus on the deep phylogeny of Haemosporida. This can be attributed in large part to the unbalanced datasets used for phylogenetic inference. As genome sequencing projects have focused on mammalian parasites of the genus *Plasmodium*, sequence data from the other haemosporidian genera was only available for a small set of standard genes, which are not ideal for reconstructing the earliest events in haemosporidian evolution (see 1.3.1). This limited gene sampling is due to the challenges involved in developing nuclear markers for this diverse group of parasites (Perkins, 2014). By employing a newly developed bioinformatic pipeline (see 2.1.3) and by carefully optimizing primer design parameters and PCR protocols, we were able to overcome these challenges and obtain sequence data of 21 nuclear gene fragments from nine haemosporidian species belonging to the genera *Haemoproteus*, *Leucocytozoon*, *Polychromophilus*, and *Plasmodium* (Borner et al., 2016). This is still the only phylogenetic study to employ multiple nuclear genes from these parasite lineages.

Phylogenetic analyses were performed based on nucleotide, codon, and amino acid data from 20 haemosporidian species employing different apicomplexan outgroups and various tree inference methods. All analyses resulted in highly congruent topologies (Fig. 12). *Leucocytozoon* was consistently recovered at the base of Haemosporida (Fig. 12), thus rejecting the phylogeny of Outlaw & Ricklefs (2011), which essentially divided Haemosporida into two clades, one comprising all mammalian *Plasmodium* species, and the other comprising all remaining haemosporidian parasites of birds and reptiles. A basal position of the mammalian *Plasmodium* clade was also recently rejected by phylogenetic analyses based on the genome of *Haemoproteus tartakovskyi* (Bensch et al., 2016).

The question whether the avian haemoproteid parasites constitute a monophyletic group has



(B)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Bayes proteins WAG	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.88	1.00	1.00	1.00	1.00
Bayes codons F3X4	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Bayes nucleotides GTR	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ML proteins WAG	100	85	97	99	100	99	100	100	66	100	100	100	55	100	100	98	88
ML codons F3X4	100	93	100	100	100	100	100	100	60	100	100	100	100	100	100	100	100
ML nucleotides GTR	100	67	91	88	91	91	100	100	95	100	100	100	87	100	100	96	99
Parsimony proteins	100	99	100	100	100	95	100	100	67	100	100	100	86	100	100	100	88
Parsimony nucleotides	100	98	100	99	100	100	100	100	99	100	100	100	89	100	100	100	97

Fig. 12. (A) Strict consensus cladogram from all eight phylogenetic analyses of Borner et al. (2016). Nodes that were not recovered in all analyses are shown as polytomies. The outgroups are not displayed. Taxa sequenced in this study are depicted in bold letters. (B) Bootstrap support values and Bayesian posterior probabilities from the individual analyses for the splits depicted above.

remained contentious (see Perkins [2014] for a review of the history of haemosporidian systematics). Unfortunately, the results of Borner et al. (2016) are inconclusive on this matter. Some analyses supported a sister group relationship of *Haemoproteus* and *Parahaemoproteus*, whereas others favored a more basal position of the former taxon. In this context, it should be noted that the haemoproteid parasites of squamates and chelonians, which were not included in the dataset of Borner et al. (2016), have recently been found to be highly divergent from the avian parasites (Pineda-Catalan et al., 2013; Maia et al., 2016) leading to the erection of the new genus *Haemocystidium*. There is also evidence that *Haemoproteus antigonis*, a parasite species of whooping cranes, actually belongs to another independent clade that may rank at the genus level (Bertram et al., 2017).

Surprisingly, we found *Polychromophilus*, a genus of bat-infecting parasites, in a sister group relationship to all *Plasmodium* parasites (Fig. 12). Previous studies had recovered *Polychromophilus* either closely associated with the avian *Plasmodium* clade (Duval et al., 2007; Megali et al., 2011; Witsenburg et al., 2012) or with the mammalian clade (Schaer et al., 2013). Considering the differences in the life cycles (*Polychromophilus* undergoes schizogony in endothelial cells but not in red blood cells, while blood schizogony is the uniting character shared by all *Plasmodium* species), its placement outside of the *Plasmodium* clade appears plausible. The recently rediscovered of ungulate malaria parasites have shown a clear phylogenetic affinity to *Polychromophilus* based on mitochondrial data (Boundenga et al., 2016; Martinsen et al., 2016; Templeton et al., 2016). However, there is still considerable incongruence among the studies with regard to the deep phylogenetic relationships of these taxa.

In the analysis of the 21 gene dataset, we recovered *Plasmodium* as a monophyletic taxon (Fig. 12). However, the dataset did not include sequences from the other two bat-infecting genera, *Hepatocystis* and *Nycteria*, which have consistently been placed nested within or closely associated with the mammalian clade of *Plasmodium* parasites. Within *Plasmodium*, a sister group relationship between the avian and the mammalian parasites was found, thus rejecting the notion of a common origin of the most virulent agent of human malaria, *Plasmodium falciparum*, and the clade of avian *Plasmodium* parasites (Pick et al., 2011).

3 References

Aguinaldo AM, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA (1997) Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* 387: 489-493.

Ahlrichs W (1995) Ultrastruktur und Phylogenie von *Seison nebaliae* (Grube 1859) und *Seison annulatus* (Claus 1876). In: *Hypothesen zu phylogenetischen Verwandschaftsverhältnissen Innerhalb der Bilateria*. Göttingen: Cuvillier.

Aris-Brosou S, Yang Z (2003) Bayesian models of episodic evolution support a late precambrian explosive diversification of the Metazoa. *Mol Biol Evol* 20: 1947-1954.

Arisue N, Hashimoto T (2015) Phylogeny and evolution of apicoplasts and apicomplexan parasites. *Parasitol Int* 64: 254-259.

Aurrecoechea C, Barreto A, Brestelli J, Brunk BP, Caler EV, Fischer S, Gajria B, Gao X, Gingle A, Grant G, et al. (2011) AmoebaDB and MicrosporidiaDB: functional genomic resources for Amoebozoa and Microsporidia species. *Nucleic Acids Res* 39: 612-619.

Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* 290: 972-977.

Ballard JW, Olsen GJ, Faith DP, Odgers WA, Rowell DM, Atkinson PW (1992) Evidence from 12S ribosomal RNA sequences that onychophorans are modified arthropods. *Science* 258: 1345-1348.

Baum J, Richard D, Healer J, Rug M, Krnajski Z, Gilberger T, Green JL, Holder AA, Cowman AF (2006) A conserved molecular motor drives cell invasion and gliding motility across malaria life cycle stages and other apicomplexan parasites. *J Biol Chem* 281: 5197-5208.

Bensch S, Canbäck B, DeBarry JD, Johansson T, Hellgren O, Kissinger JC, Palinauskas V, Videvall E, Valkiūnas G (2016) The genome of *Haemoproteus tartakovskyi* and its relationship to human malaria parasites. *Genome Biol Evol* 8: 1361-1373.

Ben-Shitrit T, Yosef N, Shemesh K, Sharan R, Ruppin E, Kupiec M (2012) Systematic identification of gene annotation errors in the widely used yeast mutation collections. *Nat Methods* 9: 373-378.

Benton MJ, Donoghue PCJ (2007) Paleontological evidence to date the tree of life. *Mol Biol Evol* 24: 26-53.

Bergsten J (2005) A review of long-branch attraction. *Cladistics* 21: 163-193.

Bertram MR, Hamer SA, Hartup BK, Snowden KF, Medeiros MC, Outlaw DC, Hamer GL (2017) A novel Haemosporida clade at the rank of genus in North American cranes (Aves: Gruiformes). *Mol Phylogenet Evol* 109: 73-79.

Blair JE, Hedges SB (2005) Molecular phylogeny and divergence times of deuterostome animals. *Mol Biol Evol* 22: 2275-2284.

Blair JE, Ikeo K, Gojobori T, Hedges SB (2002) The evolutionary position of nematodes. *BMC Evol Biol* 2: 7.

Boore JL, Collins TM, Stanton D, Daehler LL, Brown WM (1995) Deducing the pattern of arthropod phylogeny from mitochondrial DNA rearrangements. *Nature* 376: 163-165.

Boore JL, Lavrov DV, Brown WM (1998) Gene translocation links insects and crustaceans. *Nature* 392: 667-668.

Borner J, Burmester T (2017) Parasite infection of public databases: a data mining approach to identify apicomplexan contaminations in animal genome and transcriptome assemblies. *BMC Genomics* 18: 100.

Borner J, Pick C, Thiede J, Kolawole OM, Kingsley MT, Schulze J, Cottontail VM, Wellinghausen N, Schmidt-Chanasit J, Bruchhaus I, Burmester T (2016) Phylogeny of haemosporidian blood parasites revealed by a multi-gene approach. *Mol Phylogenet Evol* 94: 221-231.

Borner J, Rehm P, Schill RO, Ebersberger I, Burmester T (2014) A transcriptome approach to ecdysozoan phylogeny. *Mol Phylogenet Evol* 80: 79-87.

Boundenga L, Makanga B, Ollomo B, Gilabert A, Rougeron V, Mve-Ondo B, Arnathau C, Durand P, Moukodoum ND, Okouga A, et al. (2016) Haemosporidian Parasites of Antelopes and Other Vertebrates from Gabon, Central Africa. *PLoS One* 11: e0148958.

Bozdech Z, Llinás M, Pulliam BL, Wong ED, Zhu J, DeRisi JL (2003) The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biol* 1: E5.

Brenneis G, Ungerer P, Scholtz G (2008) The chelifores of sea spiders (Arthropoda, Pycnogonida) are the appendages of the deutocerebral segment. *Evol Dev* 10: 717-724.

Briggs DEG (1987) Scorpions take to the water. *Nature* 326: 645-646.

Brown TA (2002) *Genomes.* (2nd ed.) Oxford: Bios.

Brusca RC, Brusca GJ (2003) *Invertebrates*. Sunderland: Sinauer.

Budd GE (2001) Tardigrades as 'Stem-Group Arthropods': The evidence from the Cambrian fauna. *Zool Anz* 240: 265-279.

Burda H, Hilken G, Zrzavý J (2008) *Systematische Zoologie*. Stuttgart: Ulmer.

Burmester T (2001) Molecular evolution of the arthropod hemocyanin superfamily. *Mol Biol Evol* 18: 184-195.

Campbell LI, Rota-Stabelli O, Edgecombe GD, Marchioro T, Longhorn SJ, Telford MJ, Philippe H, Rebecchi L, Peterson KJ, Pisani D (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: 15920-15924.

Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, Silva JC, Ermolaeva MD, Allen JE, Selengut JD, Koo HL, et al. (2002) Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature* 419: 512-519.

Carreno RA, Martin DS, Barta JR (1999) *Cryptosporidium* is more closely related to the gregarines than to coccidia as shown by phylogenetic analysis of apicomplexan parasites inferred using small-subunit ribosomal RNA gene sequences. *Parasitol Res* 85: 899-904.

Chen F, Mackey AJ, Stoeckert CJJ, Roos DS (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res 34: 363-368.

Chen J (2009) The sudden appearance of diverse animal body plansduring the Cambrian explosion. *Int J Dev Biol* 53: 733-751.

Chipman AD (2015) An embryological perspective on the early arthropod fossil record. *BMC Evol Biol* 15: 285.

Chipman AD, Ferrier DEK, Brena C, Qu J, Hughes DST, Schröder R, Torres-Oliva M, Znassi N, Jiang H, Almeida FC, et al. (2014) The first myriapod genome sequence reveals conservative arthropod gene content and genome organisation in the centipede *Strigamia maritima*. *PLoS Biol* 12: e1002005.

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* 311: 1283-1287.

Collett MG (2000) Survey of canine babesiosis in South Africa. *J S Afr Vet Assoc* 71: 180-186.

Conway Morris S (1993) The fossil record and the early evolution of the Metazoa. *Nature* 361: 219-225.

Cook CE, Bergman MT, Finn RD, Cochrane G, Birney E, Apweiler R (2016) The European Bioinformatics Institute in 2016: Data growth and integration. *Nucleic Acids Res* 44: 20-26.

Copley RR, Aloy P, Russell RB, Telford MJ (2004) Systematic searches for molecular synapomorphies in model metazoan genomes give some support for Ecdysozoa after accounting for the idiosyncrasies of *Caenorhabditis elegans*. *Evol Dev* 6: 164-169.

Crimes TP (1987) Trace fossils and correlation of late Precambrian and early Cambrian strata. *Geol Mag* 124: 97-119

de Leon JC, Scheumann N, Beatty W, Beck JR, Tran JQ, Yau C, Bradley PJ, Gull K, Wickstead B, Morrissette NS (2013) A SAS-6-like protein suggests that the *Toxoplasma* conoid complex evolved from flagellar components. *Eukaryot Cell* 12: 1009-1019.

Dohle W (1980) Sind die Myriapoden eine monophyletische Gruppe? Eine Diskussion der Verwandschaftsbeziehungen der Antennaten. *Abh naturwiss Ver Hamburg* 23: 45-104.

Dohle W (2001) Are the insects terrestrial crustaceans? A discussion of some new facts and arguments and the proposal of a proper name "Tetraconata" for the monophyletic unit Crustacea+Hexapoda. *Ann Soc Entomol France* 37: 85-103.

Dong Y, Sun H, Guo H, Pan D, Qian C, Hao S, Zhou K (2012) The complete mitochondrial genome of *Pauropus longiramus* (Myriapoda: Pauropoda): implications on early diversification of the myriapods revealed from comparative analysis. *Gene* 505: 57-65.

Douzery EJP, Snell EA, Bapteste E, Delsuc F, Philippe H (2004) The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils? *Proc Natl Acad Sci USA* 101: 15386-15391.

Dove H, Stollewerk A (2003) Comparative analysis of neurogenesis in the myriapod *Glomeris marginata* (Diplopoda) suggests more similarities to chelicerates than to insects. *Development* 130: 2161-2171.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol* 4: e88.

Duman-Scheel M, Patel NH (1999) Analysis of molecular marker expression reveals neuronal homology in distantly related arthropods. *Development* 126: 2327-2334.

Dunlop J, Borner J, Burmester T (2014) Phylogeny of the Chelicerates: Morphological and Molecular Evidence. In: Wägele JW, Bartholomaeus T (Eds.) *Deep metazoan phylogeny: the backbone of the tree of life. New insights from analyses of molecules, morphology, and theory of data analysis.* (pp. 395-408) Berlin: De Gruyter.

Dunlop JA, Webster M (1999) Fossil Evidence, Terrestrialization and Arachnid Phylogeny. *J Arachnol* 27: 86-93.

Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, et al. (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452: 745-749.

Dunn CW, Howison M, Zapata F (2013) Agalma: an automated phylogenomics workflow. *BMC Bioinformatics* 14: 330.

Duval L, Robert V, Csorba G, Hassanin A, Randrianarivelojosia M, Walston J, Nhim T, Goodman SM, Ariey F (2007) Multiple host-switching of Haemosporidia parasites in bats. *Malar J* 6: 157.

Ebersberger I, Strauss S, von Haeseler A (2009) HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evol Biol* 9: 157.

Edgecombe GD (2011) Phylogenetic relationships of Myriapoda. In: Minelli, A. (Ed.) *Treatise on Zoology – Anatomy, Taxonomy, Biology. The Myriapoda, Vol. 1.* (pp. 1-20) Leiden: Brill.

Edgecombe GD, Giribet G (2002) Myriapod phylogeny and the relationships of Chilopoda. In: Llorente-Bousquets JE, Morrone JJ (Eds.) *Biodiversidad, taxonomía y biogeografía de artrópodos de México: Hacia una síntesis de su conocimiento*. (pp. 143-168) Mexico City: Prensas de Ciencias, Universidad Nacional Autónoma de México.

Edgecombe GD, Giribet G (2007) Evolutionary biology of centipedes (Myriapoda: Chilopoda). *Annu Rev Entomol* 52: 151-170.

Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30: 1575-1584.

Ertas B, von Reumont BM, Wägele J, Misof B, Burmester T (2009) Hemocyanin suggests a close relationship of Remipedia and Hexapoda. *Mol Biol Evol* 26: 2711-2718.

Fabrizius A, Hoff MLM, Engler G, Folkow LP, Burmester T (2016) When the brain goes diving: transcriptome analysis reveals a reduced aerobic energy metabolism and increased stress proteins in the seal brain. *BMC Genomics* 17: 583.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool* 27: 401-410.

Fortey RA, Briggs DEG, Wills MA (1996) The Cambrian evolutionary 'explosion': decoupling cladogenesis from morphological disparity. *Biol J Linn Soc* 57: 13-33.

Francia ME, Jordan CN, Patel JD, Sheiner L, Demerly JL, Fellows JD, de Leon JC, Morrissette NS, Dubremetz J, Striepen B (2012) Cell division in Apicomplexan parasites is organized by a homolog of the striated rootlet fiber of algal flagella. *PLoS Biol* 10: e1001444.

Friedrich M, Tautz D (1995) Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods. *Nature* 376: 165-167.

Fuerst PA, Booton GC, Crary M (2015) Phylogenetic analysis and the evolution of the 18S rRNA gene typing system of Acanthamoeba. *J Eukaryot Microbiol* 62: 69-84.

Gabriel WN, Goldstein B (2007) Segmental expression of Pax3/7 and engrailed homologs in tardigrade development. *Dev Genes Evol* 217: 421-433.

Gai Y, Song D, Sun H, Zhou K (2006) Myriapod monophyly and relationships among myriapod classes based on nearly complete 28S and 18S rDNA sequences. *Zoolog Sci* 23: 1101-1108.

Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, et al. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419: 498-511.

Garnham PCC (1966) *Malaria Parasites and Other Haemosporidia*. Oxford: Blackwell Scientific.

Giribet G (2003) Molecules, development and fossils in the study of metazoan evolution; Articulata versus Ecdysozoa revisited. *Zoology (Jena)* 106: 303-326.

Giribet G, Edgecombe GD (2012) Reevaluating the arthropod tree of life. *Annu Rev Entomol* 57: 167-186.

Grant JR, Katz LA (2014) Building a phylogenomic pipeline for the eukaryotic tree of life - addressing deep phylogenies with genome-scale data. *PLoS Curr* 6.

Graur D, Li W (2000) *Fundamentals molecular evolution*. (2nd ed.) Sunderland: Sinauer.

Hall N, Karras M, Raine JD, Carlton JM, Kooij TWA, Berriman M, Florens L, Janssen CS, Pain A, Christophides GK, et al. (2005) A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses. *Science* 307: 82-86.

Hartig G, Peters RS, Borner J, Etzbauer C, Misof B, Niehuis O (2012) Oligonucleotide primers for targeted amplification of single-copy nuclear genes in apocritan Hymenoptera. *PLoS One* 7: e39826.

Harzsch S (2004) Phylogenetic comparison of serotonin-immunoreactive neurons in representatives of the Chilopoda, Diplopoda, and Chelicerata: implications for arthropod relationships. *J Morphol* 259: 198-213.

Harzsch S, Hafner G (2006) Evolution of eye development in arthropods: phylogenetic aspects. *Arthropod Struct Dev* 35: 319-340.

Havelaar AH, Kemmeren JM, Kortbeek LM (2007) Disease burden of congenital toxoplasmosis. *Clin Infect Dis* 44: 1467-1474.

Herwaldt BL, Linden JV, Bosserman E, Young C, Olkowska D, Wilson M (2011) Transfusion-associated babesiosis in the United States: a description of cases. *Ann Intern Med* 155: 509-519.

Hiller NL, Bhattacharjee S, van Ooij C, Liolios K, Harrison T, Lopez-Estraño C, Haldar K (2004) A host-targeting signal in virulence proteins reveals a secretome in malarial infection. *Science* 306: 1934-1937.

Hittinger CT, Johnston M, Tossberg JT, Rokas A (2010) Leveraging skewed transcript abundance by RNA-Seq to increase the genomic depth of the tree of life. *Proc Natl Acad Sci USA* 107: 1476-1481.

Hoff MLM, Fabrizius A, Czech-Damal NU, Folkow LP, Burmester T (2017) Transcriptome analysis identifies key metabolic changes in the hooded seal (*Cystophora cristata*) brain in response to hypoxia and reoxygenation. *PLoS One* 12: e0169366.

Hoff MLM, Fabrizius A, Folkow LP, Burmester T (2016) An atypical distribution of lactate dehydrogenase isoenzymes in the hooded seal (*Cystophora cristata*) brain may reflect a biochemical adaptation to diving. *J Comp Physiol B* 186: 373-386.

Homer MJ, Aguilar-Delfin I, Telford SR, Krause PJ, Persing DH (2000) Babesiosis. *Clin Microbiol Rev* 13: 451-469.

Hwang UW, Friedrich M, Tautz D, Park CJ, Kim W (2001) Mitochondrial protein phylogeny joins myriapods with chelicerates. *Nature* 413: 154-157.

Irimia M, Maeso I, Penny D, Garcia-Fernàndez J, Roy SW (2007) Rare coding sequence changes are consistent with Ecdysozoa, not Coelomata. *Mol Biol Evol* 24: 1604-1607.

Irwin DM, Kocher TD, Wilson AC (1991) Evolution of the cytochrome b gene of mammals. *J Mol Evol* 32: 128-144.

Jager M, Murienne J, Clabaut C, Deutsch J, Le Guyader H, Manuel M (2006) Homology of arthropod anterior appendages revealed by Hox gene expression in a sea spider. *Nature* 441: 506-508.

Janouskovec J, Horák A, Oborník M, Lukes J, Keeling PJ (2010) A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. *Proc Natl Acad Sci USA* 107: 10949-10954.

Janssen R, Eriksson BJ, Tait NN, Budd GE (2014) Onychophoran Hox genes and the evolution of arthropod Hox gene expression. *Front Zool* 11: 22.

Jensen RA (2001) Orthologs and paralogs - we need to get it right. *Genome Biol* 2: interactions1002.

Jeram AJ (1998) Phylogeny, classifications and evolution of Silurian and Devonian scorpions. In: Selden PA (Ed.) *Proceedings of the 17th European Colloquium of Arachnology.* (pp. 17-31) Burnham Beeches: British Arachnological Society.

Jones M, Gantenbein B, Fet V, Blaxter M (2007) The effect of model choice on phylogenetic inference using mitochondrial sequence data: lessons from the scorpions. *Mol Phylogenet Evol* 43: 583-595.

Kappe S, Bruderer T, Gantt S, Fujioka H, Nussenzweig V, Ménard R (1999) Conservation of a gliding motility and cell invasion machinery in Apicomplexan parasites. *J Cell Biol* 147: 937-944.

Kivaria FM, Ruheta MR, Mkonyi PA, Malamsha PC (2007) Epidemiological aspects and economic impact of bovine theileriosis (East Coast fever) and its control: a preliminary assessment with special reference to Kibaha district, Tanzania. *Vet J* 173: 384-390.

Kumar S, Krabberød AK, Neumann RS, Michalickova K, Zhao S, Zhang X, Shalchian-Tabrizi K (2015) BIR Pipeline for Preparation of Phylogenomic Data. *Evol Bioinform Online* 11: 79-83.

Kusche K, Burmester T (2001) Diplopod hemocyanin sequence and the phylogenetic position of the Myriapoda. *Mol Biol Evol* 18: 1566-1573.

Kusche K, Ruhberg H, Burmester T (2002) A hemocyanin from the Onychophora and the emergence of respiratory proteins. *Proc Natl Acad Sci USA* 99: 10545-10548.

Lanfear R, Welch JJ, Bromham L (2010) Watching the clock: studying variation in rates of molecular evolution between species. *Trends Ecol Evol* 25: 495-503.

Lartillot N, Philippe H (2008) Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. *Philos Trans R Soc Lond B Biol Sci* 363: 1463-1472.

Laurence M, Hatzis C, Brash DE (2014) Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. *PLoS One* 9: e97876.

Lauron EJ, Oakgrove KS, Tell LA, Biskar K, Roy SW, Sehgal RNM (2014) Transcriptome sequencing and analysis of *Plasmodium gallinaceum* reveals polymorphisms and selection on the apical membrane antigen-1. *Malar J* 13: 382.

Lefèvre T, Sanchez M, Ponton F, Hughes D, Thomas F (2007) Virulence and resistance in malaria: who drives the outcome of the infection? *Trends Parasitol* 23: 299-302.

Lepage T, Lawi S, Tupper P, Bryant D (2006) Continuous and tractable models for the variation of evolutionary rates. *Math Biosci* 199: 216-233.

Levine ND (1988) *The protozoan phylum apicomplexa*. Boca Raton: CRC Press.

Lozano-Fernandez J, Carton R, Tanner AR, Puttick MN, Blaxter M, Vinther J, Olesen J, Giribet G, Edgecombe GD, Pisani D (2016) A molecular palaeobiological exploration of arthropod terrestrialization. *Philos Trans R Soc Lond B Biol Sci* 371.

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290: 1151-1155.

Madsen O, Deen PM, Pesole G, Saccone C, de Jong WW (1997) Molecular evolution of mammalian aquaporin-2: further evidence that elephant shrew and aardvark join the paenungulate clade. *Mol Biol Evol* 14: 363-371.

Maia JP, Harris DJ, Carranza S (2016) Reconstruction of the evolutionary history of Haemosporida (Apicomplexa) based on the cyt b gene with characterization of *Haemocystidium* in geckos (Squamata: Gekkota) from Oman. Parasitol Int 65: 5-11.

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.

Marti M, Good RT, Rug M, Knuepfer E, Cowman AF (2004) Targeting malaria virulence and remodeling proteins to the host erythrocyte. *Science* 306: 1930-1933.

Martin MW, Grazhdankin DV, Bowring SA, Evans DA, Fedonkin MA, Kirschvink JL (2000) Age of Neoproterozoic bilatarian body and trace fossils, White Sea, Russia: implications for metazoan evolution. *Science* 288: 841-845.

Martinsen ES, McInerney N, Brightman H, Ferebee K, Walsh T, McShea WJ, Forrester TD, Ware L, Joyner PH, Perkins SL, et al. (2016) Hidden in plain sight: Cryptic and endemic malaria parasites in North American white-tailed deer (*Odocoileus virginianus*). *Sci Adv* 2: e1501486.

Martinsen ES, Perkins SL, Schall JJ (2008) A three-genome phylogeny of malaria parasites (*Plasmodium* and closely related genera): evolution of life-history traits and host switches. *Mol Phylogenet Evol* 47: 261-273.

Mayer G, Martin C, Rüdiger J, Kauschke S, Stevenson PA, Poprawa I, Hohberg K, Schill RO, Pflüger H, Schlegel M (2013) Selective neuronal staining in tardigrades and onychophorans provides insights into the evolution of segmental ganglia in panarthropods. *BMC Evol Biol* 13: 230.

McFadden GI, Reith ME, Munholland J, Lang-Unnasch N (1996) Plastid in human parasites. *Nature* 381: 482.

McMahon MM, Sanderson MJ (2006) Phylogenetic supermatrix analysis of GenBank sequences from 2228 papilionoid legumes. *Syst Biol* 55: 818-836.

Merchant S, Wood DE, Salzberg SL (2014) Unexpected cross-species contamination in genome sequencing projects. *PeerJ* 2: e675.

Meusemann K, von Reumont BM, Simon S, Roeding F, Strauss S, Kück P, Ebersberger I, Walzl M, Pass G, Breuers S, et al. (2010) A phylogenomic approach to resolve the arthropod tree of life. *Mol Biol Evol* 27: 2451-2464.

Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, et al. (2014) Phylogenomics resolves the timing and pattern of insect evolution. *Science* 346: 763-767.

Miyazawa H, Ueda C, Yahata K, Su Z (2014) Molecular phylogeny of Myriapoda provides insights into evolutionary patterns of the mode in post-embryonic development. *Sci Rep* 4: 4127.

Montoya JG, Liesenfeld O (2004) Toxoplasmosis. *Lancet* 363: 1965-1976.

Moore RB, Oborník M, Janouskovec J, Chrudimský T, Vancová M, Green DH, Wright SW, Davies NW, Bolch CJS, Heimann K, et al. (2008) A photosynthetic alveolate closely related to apicomplexan parasites. *Nature* 451: 959-963.

Morrison DA (2008) Prospects for elucidating the phylogeny of the Apicomplexa. *Parasite* 15: 191-196.

Naccache SN, Greninger AL, Lee D, Coffey LL, Phan T, Rein-Weston A, Aronsohn A, Hackett JJ, Delwart EL, Chiu CY (2013) The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. *J Virol* 87: 11966-11977.

Nardi F, Spinsanti G, Boore JL, Carapelli A, Dallai R, Frati F (2003) Hexapod origins: monophyletic or paraphyletic? *Science* 299: 1887-1889.

Negrisolo E, Minelli A, Valle G (2004) The mitochondrial genome of the house centipede scutigera and the monophyly versus paraphyly of myriapods. *Mol Biol Evol* 21: 770-780.

Nielsen C (1995) *Animal Evolution. Interrelationships of the Living Phyla.* Oxford: University Press.

Orosz F (2015) Two recently sequenced vertebrate genomes are contaminated with apicomplexan species of the Sarcocystidae family. *Int J Parasitol* 45: 871-878.

Outlaw DC, Ricklefs RE (2011) Rerooting the evolutionary tree of malaria parasites. *Proc Natl Acad Sci USA* 108: 13183-13187.

Pace RM, Grbić M, Nagy LM (2016) Composition and genomic organization of arthropod Hox clusters. *Evodevo* 7: 11.

Page RD, Holmes EC (1998). *Molecular evolution: a phylogenetic approach*. Oxford: Blackwell.

Pain A, Böhme U, Berry AE, Mungall K, Finn RD, Jackson AP, Mourier T, Mistry J, Pasini EM, Aslett MA, et al. (2008) The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature* 455: 799-803.

Paparini A, Macgregor J, Ryan UM, Irwin PJ (2015) First Molecular Characterization of Theileria ornithorhynchi Mackerras, 1959: yet Another Challenge to the Systematics of the Piroplasms. *Protist* 166: 609-620.

Park J, Rho HS, Kristensen RM, Kim W, Giribet G (2006) First molecular data on the phylum Loricifera: an investigation into the phylogeny of ecdysozoa with emphasis on the positions of Loricifera and Priapulida. *Zoolog Sci* 23: 943-954.

Pepato AR, da Rocha CEF, Dunlop JA (2010) Phylogenetic position of the acariform mites: sensitivity to homology assessment under total evidence. *BMC Evol Biol* 10: 235.

Perkins SL (2014) Malaria's many mates: past, present, and future of the systematics of the order Haemosporida. *J Parasitol* 100: 11-25.

Perkins SL, Schall JJ (2002) A molecular phylogeny of malarial parasites recovered from cytochrome b gene sequences. *J Parasitol* 88: 972-978.

Peters RS, Meyer B, Krogmann L, Borner J, Meusemann K, Schütte K, Niehuis O, Misof B (2011) The taming of an impossible child: a standardized all-in approach to the phylogeny of Hymenoptera using public database sequences. *BMC Biol* 9: 55.

Peterson KJ, Cotton JA, Gehling JG, Pisani D (2008) The Ediacaran emergence of bilaterians: congruence between the genetic and the geological fossil records. *Philos Trans R Soc Lond B Biol Sci* 363: 1435-1443.

Philippe H, Lartillot N, Brinkmann H (2005b) Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Mol Biol Evol* 22: 1246-1253.

Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F (2005a) Heterotachy and long-branch attraction in phylogenetics. *BMC Evol Biol* 5: 50.

Pick C, Ebersberger I, Spielmann T, Bruchhaus I, Burmester T (2011) Phylogenomic analyses of malaria parasites and evolution of their exported proteins. *BMC Evol Biol* 11: 167.

Pineda-Catalan O, Perkins SL, Peirce MA, Engstrand R, Garcia-Davila C, Pinedo-Vasquez M, Aguirre AA (2013) Revision of hemoproteid genera and description and redescription of two species of chelonian hemoproteid parasites. *J Parasitol* 99: 1089-1098.

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.

Porter SB, Sande MA (1992) Toxoplasmosis of the central nervous system in the acquired immunodeficiency syndrome. *N Engl J Med* 327: 1643-1648.

Promponas VJ, Iliopoulos I, Ouzounis CA (2015) Annotation inconsistencies beyond sequence similarity-based function prediction - phylogeny and genome structure. *Stand Genomic Sci* 10: 108.

Redmond NE, Morrow CC, Thacker RW, Diaz MC, Boury-Esnault N, Cárdenas P, Hajdu E, Lôbo-Hajdu G, Picton BE, Pomponi SA, et al. (2013) Phylogeny and systematics of demospongiae in light of new small-subunit ribosomal DNA (18S) sequences. *Integr Comp Biol* 53: 388-415.

Reece JB, Urry LA, Cain ML, Wasserman SA, Minorsky PV, Jackson R (2011) *Campbell Biology.* (9th ed.) New York: Pearson Education.

Regier JC, Shultz JW, Kambic RE (2005) Pancrustacean phylogeny: hexapods are terrestrial crustaceans and maxillopods are not monophyletic. *Proc Biol Sci* 272: 395-401.

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.

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.

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.

Rehm P, Pick C, Borner J, Markl J, Burmester T (2012) The diversity and evolution of chelicerate hemocyanins. *BMC Evol Biol* 12: 19.

Richter S (2002) The Tetraconata concept: hexapod-crustacean relationships and the phylogeny of Crustacea. *Org Divers Evol* 2: 217-237.

Robbertse B, Yoder RJ, Boyd A, Reeves J, Spatafora JW (2011) Hal: an automated pipeline for phylogenetic analyses of genomic data. *PLoS Curr* 3.

Roeding F, Borner J, Kube M, Klages S, Reinhardt R, Burmester T (2009) A 454 sequencing approach for large scale phylogenomic analysis of the common emperor scorpion (*Pandinus imperator*). *Mol Phylogenet Evol* 53: 826-834.

Roeding F, Hagner-Holler S, Ruhberg H, Ebersberger I, von Haeseler A, Kube M, Reinhardt R, Burmester T (2007) EST sequencing of Onychophora and phylogenomic analysis of Metazoa. *Mol Phylogenet Evol* 45: 942-951.

Rogozin IB, Wolf YI, Carmel L, Koonin EV (2007) Ecdysozoan clade rejected by genome-wide analysis of rare amino acid replacements. *Mol Biol Evol* 24: 1080-1090.

Roos DS (2005) Genetics. Themes and variations in apicomplexan parasite biology. *Science* 309: 72-73.

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 Biol Sci* 278: 298-306.

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.

Rubinstein CV, Gerrienne P, de la Puente GS, Astini RA, Steemans P (2010) Early Middle Ordovician evidence for land plants in Argentina (eastern Gondwana). *New Phytol* 188: 365-369.

Ruvolo M, Disotell TR, Allard MW, Brown WM, Honeycutt RL (1991) Resolution of the African hominoid trichotomy by use of a mitochondrial gene sequence. *Proc Natl Acad Sci USA* 88: 1570-1574.

Sahraeian SM, Luo KR, Brenner SE (2015) SIFTER search: a web server for accurate phylogeny-based protein function prediction. *Nucleic Acids Res* 43: 141-147.

Saitou N, Nei M (1986) The number of nucleotides required to determine the branching order of three species, with special reference to the human-chimpanzee-gorilla divergence. *J Mol Evol* 24: 189-204.

Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW (2014) Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* 12: 87.

Sanders KL, Lee MSY (2010) Arthropod molecular divergence times and the Cambrian origin of pentastomids. *Syst Biodivers* 8: 63-74.

Sanderson MJ (1997) A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol Biol Evol* 14: 1218-1231.

Sanderson MJ, Boss D, Chen D, Cranston KA, Wehe A (2008) The PhyLoTA Browser: processing GenBank for molecular phylogenetics research. *Syst Biol* 57: 335-346.

Schaer J, Perkins SL, Decher J, Leendertz FH, Fahr J, Weber N, Matuschewski K (2013) High diversity of West African bat malaria parasites and a tight link with rodent *Plasmodium* taxa. *Proc Natl Acad Sci USA* 110: 17415-17419.

Schmidt-Rhaesa A (1996) The nervous system of *Nectonema munidae* and *Gordius aquaticus,* with implications on the ground pattern of the Nematomorpha. *Zoomorphology* 116: 133-142.

Schmidt-Rhaesa A (1998) The position of the Arthropoda in the phylogenetic system. *J Morphol* 238: 263–285.

Schmidt-Rhaesa A (2012) *Nematomorpha, Priapulida, Kinorhyncha, Loricifera*. In: *Handbook of Zoology*. Berlin: de Gruyter.

Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W, (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc Natl Acad Sci USA* 109: 6241-6246.

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.

Shear WA, Selden PA, Rolfe WDI, Bonamo PM, Grierson JD (1987) New terrestrial arachnids from the Devonian of Gilboa, New York (Arachnida: Trigonotarbida). *Am Mus Novit* 2901: 1-74.

Shultz JW (1990) Evolutionary morphology and phylogeny of Arachnida. *Cladistics* 6: 1-38.

Shultz JW, Regier JC (2000) Phylogenetic analysis of arthropods using two nuclear protein-encoding genes supports a crustacean + hexapod clade. *Proc Biol Sci* 267: 1011-1019.

Sota T, Vogler AP (2001) Incongruence of mitochondrial and nuclear gene trees in the Carabid beetles *Ohomopterus*. Syst Biol 50: 39-59.

Springer MS, Cleven GC, Madsen O, de Jong WW, Waddell VG, Amrine HM, Stanhope MJ (1997) Endemic African mammals shake the phylogenetic tree. *Nature* 388: 61-64.

Springer MS, DeBry RW, Douady C, Amrine HM, Madsen O, de Jong WW, Stanhope MJ (2001) Mitochondrial versus nuclear gene sequences in deep-level mammalian phylogeny reconstruction. *Mol Biol Evol* 18: 132-143.

Stanhope MJ, Waddell VG, Madsen O, de Jong W, Hedges SB, Cleven GC, Kao D, Springer MS (1998) Molecular evidence for multiple origins of Insectivora and for a new order of endemic African insectivore mammals. *Proc Natl Acad Sci USA* 95: 9967-9972.

Steemans P, Hérissé AL, Melvin J, Miller MA, Paris F, Verniers J, Wellman CH (2009) Origin and radiation of the earliest vascular land plants. *Science* 324: 353.

Stollewerk A (2016) A flexible genetic toolkit for arthropod neurogenesis. P*hilos Trans R Soc Lond B Biol Sci* 371.

Storch V, Welsch U (2003) *Systematische Zoologie.* (6th ed.) Heidelberg: Spektrum.

Strong MJ, Xu G, Morici L, Splinter Bon-Durant S, Baddoo M, Lin Z, Fewell C, Taylor CM, Flemington EK (2014) Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples. *PLoS Pathog* 10: e1004437.

Strother PK, Battison L, Brasier MD, Wellman CH (2011) Earth's earliest non-marine eukaryotes. *Nature* 473: 505-509.

Suzuki S, Kakuta M, Ishida T, Akiyama Y (2014) GHOSTX: an improved sequence homology search algorithm using a query suffix array and a database suffix array. *PLoS One* 9: e103833.

Tachibana S, Sullivan SA, Kawai S, Nakamura S, Kim HR, Goto N, Arisue N, Palacpac NMQ, Honma H, Yagi M, et al. (2012) *Plasmodium cynomolgi* genome sequences provide insight into *Plasmodium vivax* and the monkey malaria clade. *Nat Genet* 44: 1051-1055.

Tao Z, Sui X, Jun C, Culleton R, Fang Q, Xia H, Gao Q (2015) Vector sequence contamination of the *Plasmodium vivax* sequence database in PlasmoDB and *in silico* correction of 26 parasite sequences. *Parasit Vectors* 8: 318.

Templeton TJ, Asada M, Jiratanh M, Ishikawa SA, Tiawsirisup S, Sivakumar T, Namangala B, Takeda M, Mohkaew K, Ngamjituea S, et al. (2016) Ungulate malaria parasites. *Sci Rep* 6: 23230.

Templeton TJ, Enomoto S, Chen W, Huang C, Lancto CA, Abrahamsen MS, Zhu G (2010) A genome-sequence survey for *Ascogregarina taiwanensis* supports evolutionary affiliation but metabolic diversity between a Gregarine and *Cryptosporidium*. *Mol Biol Evol* 27: 235-248.

Thompson RCA, Olson ME, Zhu G, Enomoto S, Abrahamsen MS, Hijjawi NS (2005) *Cryptosporidium* and cryptosporidiosis. *Adv Parasitol* 59: 77-158.

Thomson RC, Shaffer HB (2010) Sparse supermatrices for phylogenetic inference: taxonomy, alignment, rogue taxa, and the phylogeny of living turtles. *Syst Biol* 59: 42-58.

Thorne JL, Kishino H, Painter IS (1998) Estimating the rate of evolution of the rate of molecular evolution. *Mol Biol Evol* 15: 1647-1657.

Tiegs OW (1947) The development and affinities of the Pauropoda, based on a study of Pauropus silvaticus. *Q J Microsc Sci* 88: 165.

Toso MA, Omoto CK (2007) *Gregarina niphandrodes* may lack both a plastid genome and organelle. *J Eukaryot Microbiol* 54: 66-72.

Trees AJ, Davison HC, Innes EA, Wastling JM (1999) Towards evaluating the economic impact of bovine neosporosis. *Int J Parasitol* 29: 1195-1200.

Videvall E, Cornwallis CK, Palinauskas V, Valkiūnas G, Hellgren O (2015) The avian transcriptome response to malaria infection. *Mol Biol Evol* 32: 1255-1267.

von Reumont BM, Jenner RA, Wills MA, Dell'ampio E, Pass G, Ebersberger I, Meyer B, Koenemann S, Iliffe TM, Stamatakis A, 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.

von Reumont BM, Meusemann K, Szucsich NU, Dell'Ampio E, Gowri-Shankar V, Bartel D, Simon S, Letsch HO, Stocsits RR, Luan Y, et al. (2009) Can comprehensive background knowledge be incorporated into substitution models to improve phylogenetic analyses? A case study on major arthropod relationships. *BMC Evol Biol* 9: 119.

Wägele JW (2001) *Grundlagen der phylogenetischen Systematik.* (2nd ed.) München: Pfeil.

Walsh HE, Kidd MG, Moum T, Friesen VL (1999) Polytomies and the power of phylogenetic inference. *Evolution* 53: 932-937.

Webster BL, Copley RR, Jenner RA, Mackenzie-Dodds JA, Bourlat SJ, Rota-Stabelli O, Littlewood DTJ, Telford MJ (2006) Mitogenomics and phylogenomics reveal priapulid worms as extant models of the ancestral Ecdysozoan. *Evol Dev* 8: 502-510.

Westheide W, Rieger R (1996) *Spezielle Zoologie. Teil 1: Einzeller und Wirbellose Tiere*. (1st ed.) Stuttgart: Fischer.

Westheide W, Rieger R (2013) *Spezielle Zoologie. Teil 1: Einzeller und Wirbellose Tiere*. (3rd ed.) Heidelberg: Spektrum.

Weygold P, Paulus HF (1979) Untersuchungen zur Morphologie, Taxonomie und Phylogenie der Chelicerata II. Cladogramme und die Entfaltung der Chelicerata. *J Zool Syst and Evol Res* 17: 177-200.

Wheeler WC, Hayashi CY (1998) The Phylogeny of the Extant Chelicerate Orders. *Cladistics* 14: 173-192.

Williams RB (1999) A compartmentalised model for the estimation of the cost of coccidiosis to the world's chicken production industry. *Int J Parasitol* 29: 1209-1229.

Wilson K, Cahill V, Ballment E, Benzie J (2000) The complete sequence of the mitochondrial genome of the crustacean Penaeus monodon: are malacostracan crustaceans more closely related to insects than to branchiopods? *Mol Biol Evol* 17: 863-874.

Witsenburg F, Clément L, López-Baucells A, Palmeirim J, Pavlinić I, Scaravelli D, Ševčík M, Dutoit L, Salamin N, Goudet J, et al. (2015) How a haemosporidian parasite of bats gets around: the genetic structure of a parasite, vector and host compared. *Mol Ecol* 24: 926-940.

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA* 87: 4576-4579.

Wolf YI, Rogozin IB, Koonin EV (2004) Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. *Genome Res* 14: 29-36.

World Health Organization (2015) *World malaria report 2015*. Geneva: World Health Organisation.

Yamasaki H, Fujimoto S, Miyazaki K (2015) Phylogenetic position of Loricifera inferred from nearly complete 18S and 28S rRNA gene sequences. *Zoological Lett* 1: 18.

Yeh I, Altman RB (2006) Drug Targets for *Plasmodium falciparum*: a post-genomic review/survey. *Mini Rev Med Chem* 6: 177-202.

Zhu G, Keithly JS, Philippe H (2000a) What is the phylogenetic position of *Cryptosporidium*? *Int J Syst Evol Microbiol* 50: 1673-1681.

Zhu G, Marchewka MJ, Keithly JS (2000b) *Cryptosporidium parvum* appears to lack a plastid genome. *Microbiology* 146: 315-321.

Zhu L, Mok S, Imwong M, Jaidee A, Russell B, Nosten F, Day NP, White NJ, Preiser PR, Bozdech Z (2016a) New insights into the *Plasmodium vivax* transcriptome using RNA-Seq. *Sci Rep* 6: 20498.

Zhu J, Wang G, Pelosi P (2016b) Plant transcriptomes reveal hidden guests. *Biochem Biophys Res Commun* 474: 497-502.

Zrzavý J, Štys P (1997) The basic body plan of arthropods: insights from evolutionary morphology and developmental biology. *J Evol Biol* 10: 353-367.

Zuckerkandl E, Pauling L (1965) Molecules as documents of evolutionary history. *J Theor Biol* 8: 357-366.

Zwick A, Regier JC, Zwickl DJ (2012) Resolving discrepancy between nucleotides and amino acids in deep-level arthropod phylogenomics: differentiating serine codons in 21-amino-acid models. *PLoS One* 7: e47450.

4 Declaration of own contribution to the published manuscripts

Borner J, Burmester T (2017) Parasite infection of public databases: a data mining approach to identify apicomplexan contaminations in animal genome and transcriptome assemblies. *BMC Genomics* 18: 100.

- designed the study, implemented the software, analysed the data, written the manuscript

Borner J, Pick C, Thiede J, Kolawole OM, Kingsley MT, Schulze J, Cottontail VM, Wellinghausen N, Schmidt-Chanasit J, Bruchhaus I, Burmester T (2016) Phylogeny of haemosporidian blood parasites revealed by a multi-gene approach. *Mol Phylogenet Evol* 94: 221-231.

- designed the study, performed the majority of lab work, analysed the data, written the manuscript

Borner J, Rehm P, Schill RO, Ebersberger I, Burmester T (2014) A transcriptome approach to ecdysozoan phylogeny. *Mol Phylogenet Evol* 80: 79-87.

- performed parts of the lab work, analysed the data, written substantial parts of the manuscript

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.

- performed the phylogenetic analyses, written parts of the methods section

Dunlop J, **Borner J**, Burmester T (2014) Phylogeny of the Chelicerates: Morphological and Molecular Evidence. In: Wägele JW, Bartholomaeus T (Eds.) *Deep metazoan phylogeny: the backbone of the tree of life. New insights from analyses of molecules, morphology, and theory of data analysis.* (pp. 395-408) Berlin: De Gruyter.

- performed parts of the lab work, performed the phylogenetic analyses

Hartig G, Peters RS, **Borner J**, Etzbauer C, Misof B, Niehuis O (2012) Oligonucleotide primers for targeted amplification of single-copy nuclear genes in apocritan Hymenoptera. *PLoS One* 7: e39826.

- designed the primers, written parts of the methods section

Rehm P, Pick C, **Borner J**, Markl J, Burmester T (2012) The diversity and evolution of chelicerate hemocyanins. *BMC Evol Biol* 12: 19.

- performed some of the phylogenetic analyses

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.

- performed the thermodynamic integration and parts of the molecular clock analyses

Peters RS, Meyer B, Krogmann L, **Borner J**, Meusemann K, Schütte K, Niehuis O, Misof B (2011) The taming of an impossible child: a standardized all-in approach to the phylogeny of Hymenoptera using public database sequences. *BMC Biol* 9: 55.

- implemented parts of the software pipeline

5 Eidesstattliche Versicherung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Hamburg, den 11. April 2017

Janus Borner

Borner and Burmester *BMC Genomics* (2017) 18:100
DOI 10.1186/s12864-017-3504-1

BMC Genomics

RESEARCH ARTICLE Open Access

Parasite infection of public databases: a data mining approach to identify apicomplexan contaminations in animal genome and transcriptome assemblies

Janus Borner* and Thorsten Burmester*

Abstract

Background: Contaminations from various exogenous sources are a common problem in next-generation sequencing. Another possible source of contaminating DNA are endogenous parasites. On the one hand, undiscovered contaminations of animal sequence assemblies may lead to erroneous interpretation of data; on the other hand, when identified, parasite-derived sequences may provide a valuable source of information.

Results: Here we show that sequences deriving from apicomplexan parasites can be found in many animal genome and transcriptome projects, which in most cases derived from an infection of the sequenced host specimen. The apicomplexan sequences were extracted from the sequence assemblies using a newly developed bioinformatic pipeline (ContamFinder) and tentatively assigned to distinct taxa employing phylogenetic methods. We analysed 920 assemblies and found 20,907 contigs of apicomplexan origin in 51 of the datasets. The contaminating species were identified as members of the apicomplexan taxa Gregarinasina, Coccidia, Piroplasmida, and Haemosporida. For example, in the platypus genome assembly, we found a high number of contigs derived from a piroplasmid parasite (presumably *Theileria ornithorhynchi*). For most of the infecting parasite species, no molecular data had been available previously, and some of the datasets contain sequences representing large amounts of the parasite's gene repertoire.

Conclusion: Our study suggests that parasite-derived contaminations represent a valuable source of information that can help to discover and identify new parasites, and provide information on previously unknown host-parasite interactions. We, therefore, argue that uncurated assembly data should routinely be made available in addition to the final assemblies.

Keywords: Apicomplexa, Contamination, Database analysis, Phylogeny, Coccidia, Piroplasmida, Gregarinasina, Haemosporida, Malaria, Parasites

Background

Contaminations by DNA from non-target organisms are a common problem in next-generation sequencing projects [1–3]. If these contaminants are not flagged and remain in the datasets after sequence assembly and deposition into public databases, subsequent analyses of the datasets may yield confusing results and may lead to false conclusions [4, 5]. Various computational methods have been developed that are highly efficient at identifying and removing common contaminants, such as DNA from cloning vectors or human DNA, before sequence assembly [6, 7]. By contrast, contaminations by DNA from other sources, e.g. via aerosol contamination in the laboratory or at the sequencing center, are notoriously difficult to identify.

Another potential source of contamination are pathogens present in the source material [8–10]. In genome projects of wild animals, it is virtually impossible to rule

* Correspondence: janus.borner@uni-hamburg.de; thorsten.burmester@uni-hamburg.de
Institute of Zoology, Biocenter Grindel, University of Hamburg, Martin-Luther-King-Platz 3, D-20146 Hamburg, Germany

out infection by an unknown pathogen before sequencing. The development of bioinformatic approaches to identify contamination by pathogens is therefore of great importance. Most existing tools aim to assign individual reads to taxonomic groups without prior assembly. As the amount of read data in next-generation sequencing (NGS) projects is enormous and the reads are short and of low quality, the programs either rely on near exact matches at the nucleotide level [11], or employ smaller databases containing only selected marker genes [12] or genes that are specific to certain clades [13]. The former approach is not suited for the identification of contaminations by parasites for which only distantly related species are available in the public databases, whereas the latter approach is especially useful for quantitative estimates of genome abundance but can only find a small number of predefined genes. The program PathSeq [14], which was developed to identify microorganisms by deep sequencing of human tissue, uses a different approach by first subtracting all reads derived from the human host. However, this is obviously only feasible when high-quality genome data is already available for the host species.

While previous approaches have mostly focused on the removal of contaminating sequences, the identification of parasite-derived contaminations may also enable the discovery of novel parasite taxa and shed light on previously unknown host-parasite associations. For example, a recent study by Orosz [10] has highlighted that contaminations by parasite DNA may also represent a source of information. By searching published whole genome shotgun assemblies from various animal taxa for a protein (apicortin) that is characteristic for apicomplexan parasites but absent in animals (Eumetazoa), the author identified sequences from apicomplexan parasites in two animal genome assemblies from the northern bobwhite (*Colinus virginianus*) and the bat *Myotis davidii*. Data mining of genome assemblies from infected hosts may produce large amounts of genomic data from pathogens that are not yet represented in the public databases.

Members of the protozoan phylum Apicomplexa are obligate parasites that may cause serious illnesses in humans and animals. For example, five distinct species of the genus *Plasmodium* are the causative agents of human malaria and, as such, pose one of the greatest threats to public health [15]. While the gregarines (Gregarinasina) only infect invertebrates, members of the apicomplexan taxa Coccidia and Piroplasmida are responsible for numerous infectious diseases in wild and domesticated animals, such as coccidiosis and babesiosis, resulting in considerable animal health problems and economic losses [16].

Here we present a bioinformatic pipeline (ContamFinder) to identify parasite contamination in NGS assembly data and extract genetic sequences derived from the contaminating parasite. Phylogenetic methods were employed to assign the sequences to apicomplexan taxa. In total, we found contaminating sequences of apicomplexan origin in 51 genome and transcriptome assemblies. The amount of parasite-derived coding sequences varies greatly among the contaminated assemblies from just a few contigs to a significant amount of the parasite's gene repertoire.

Methods

Data selection

We downloaded all available metazoan genome and transcriptome assemblies from the Whole Genome Shotgun (WGS) [17] and Transcriptome Shotgun Assembly (TSA) [18] databases. As no gene predictions were available for the genome sequences from *Ascogregarina taiwanensis* (WGS prefix ABJQ01), the contigs were processed alongside the metazoan assemblies using the pipeline described below in order to obtain predicted protein sequences for this taxon.

Extraction of parasite-derived sequences

In the first step (Fig. 1a) of the ContamFinder pipeline, all contigs from each assembly were subjected to a search against all apicomplexan proteomes from the Eukaryotic Pathogen Database (EuPathDB) [19, 20]. All searches were performed employing GHOSTX [21] based on its high performance (Table 1) in a test run on the transcriptome assembly of the domestic goat, *Capra hircus* (TSA prefix GAOJ01), and the genome assembly of the white-tailed deer, *Odocoileus virginianus* (WGS prefix AEGY01), but ContamFinder also supports output from BLAST+ [22] and RAPSearch2 [23]. Sequences that showed significant sequence similarity (E-value cutoff: 1e-10; see below) to a parasite protein were analyzed further; the rest was discarded. By searching against a relatively small database (compared to UniProt) first, and by the subsequent removal of all contigs without sequence similarity, we massively reduced the amount of sequences that needed to be searched against the UniProt database. However, as highly conserved genes from a metazoan organism may have significant sequence similarity to parasite genes, this initial selection contained large amounts of false positives. Preliminary analyses showed that blastx-style searches of the remaining contigs against the UniProt database would still be too slow for large numbers of genome assemblies, which may contain very long contigs.

To further improve the performance, the amino acid sequence encoded in each of the potentially parasite-derived contigs was predicted in the second step (Fig. 1b). Gene prediction was performed by the program Exonerate [24] using the best hitting protein from EuPathDB as guide (with "full refinement" of the alignments, employing the

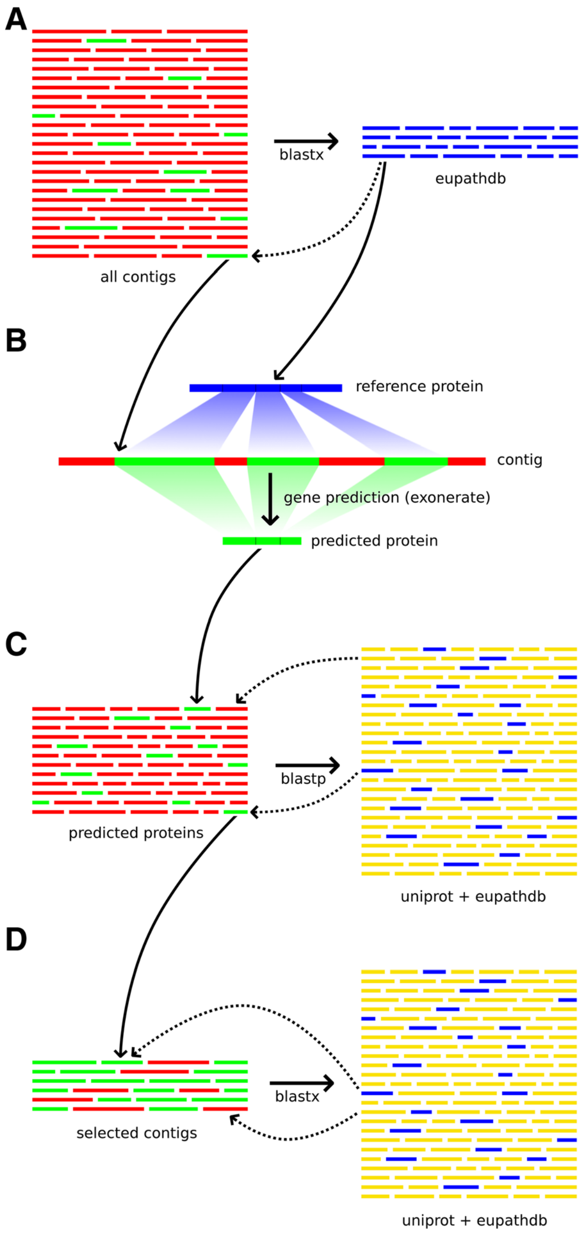


Fig. 1 Schematic overview of the ContamFinder pipeline. **a** All contigs from an assembly were searched against apicomplexan proteomes from the Eukaryotic Pathogen Database (EuPathDB [19, 20]). Sequences without significant hit were discarded. **b** Amino acid sequences were predicted using the best hitting apicomplexan protein. Low complexity regions and repeats in the sequence were masked. **c** The predicted amino acid sequences were searched against the EuPathDB and UniProt database. Sequences with the best hit outside of Apicomplexa were discarded. **d** Unprocessed contigs corresponding to the hits from the previous step were searched against the EuPathDB and UniProt databases. Sequences that had their best hit outside of Apicomplexa were discarded. Contigs and sequence regions that were kept and used in the next step are shown in green; sequences that were discarded are denoted in red. Parasite-derived proteins in the search database are shown in blue, others in yellow

protein2dna model for transcriptome data and the protein2genome model for genome data). Subsequently, regions of low complexity or repeats in the amino acid sequence were masked by the SEG filter from the BLAST + package.

In the third step (Fig. 1c), the predicted amino acid sequences were searched against all complete proteomes from the UniProt database. Sequences that had their best hit against a protein from an apicomplexan species were extracted for further analysis; the rest was discarded. In preliminary analyses, we found several false positive hits caused by falsely annotated proteins in the UniProt database that were in fact derived from the parasite's host. Therefore, we removed all protein sequences annotated as apicomplexan and replaced them with the genome-based proteome predictions available in the well-curated EuPathDB. Vice versa, undetected parasite contamination in a genome or transcriptome assembly may have led to parasite proteins being falsely assigned to the host species in the Uniprot database. This would cause similarity searches to produce false-negative results when analyzing the affected assembly. To avoid discarding such contaminants, hits against sequences from the source species were ignored.

Because the predicted amino acid sequences were obtained by using the best hitting parasite protein as a guide sequence, they may be biased towards showing a high similarity to this protein. Therefore, in the final step of the pipeline (Fig. 1d), we searched the unprocessed nucleotide contigs corresponding to the hits from the previous step against the same database (UniProt + EuPathDB). Again, sequences that had their best hits against proteins of non-apicomplexan origin were discarded.

For a few sequencing projects, the WGS and TSA databases contained multiple assemblies that were based on the same raw sequencing data. In these cases, we only kept the results from the assembly with the highest number of hits. All analyses were run on the high-performance computing cluster of the Regionales Rechenzentrum (RRZ), University of Hamburg, employing dual CPU compute nodes, each equipped with two Intel Xeon E5-2630v3 CPUs.

Orthology prediction and multiple sequence alignment

Predicted proteome data derived from all available apicomplexan and chromerid genomes (maximum one per species) were obtained from EuPathDB and assigned to ortholog groups based on their OrthoMCL [25] annotation available in EuPathDB. Ortholog groups were required to contain sequences from at least three of the six major taxonomic groups (Chromerida, Gregarinasina, *Cryptosporidium*, Coccidia, Piroplasmida, Haemosporida). To obtain a dataset of unambiguous one-to-one orthologs, groups that contained more than one sequence from the

Table 1 Performance of the ContamFinder pipeline employing three different sequence similarity search tools compared to an all-vs-all blastx search

	Assembly type	Assembly size	all-vs-all blastx search (BLAST+)	ContamFinder (BLAST+)	ContamFinder (RAPsearch2)	ContamFinder (GHOSTX)
Capra hircus (GAOJ01)	transcriptome	25.1 Mb	82 h 14 min 439 hits	15 h 57 min 418 hits	40 min 396 hits	25 min 405 hits
Odocoileus virginianus (AEGY01)	genome	14.3 Mb	36 h 9 min 127 hits	1 h 12 min 122 hits	8 min 104 hits	3 min 98 hits

same proteome were discarded. All predicted parasite proteins from the metazoan sequence assemblies were assigned to these orthologous groups by OrthoMCL. Genes with a taxon coverage of less than 30% were removed to reduce the amount of missing data in the final dataset, resulting in 1,420 genes from 67 taxa (dataset 1). As this dataset was too large for Bayesian tree inference, a reduced dataset was generated (minimum taxon coverage of 70% for each gene, minimum of 10 genes per taxon). This dataset comprises 301 genes from 49 taxa (dataset 2). Each group of orthologous proteins was aligned individually using MAFFT L-INS-i v7.013 [26]. Poorly aligned sections of the amino acid alignments were eliminated by Gblocks v0.91b [27] (settings: –b1 = [50% of the number of sequences + 1] -b2 = [85% of the number of sequences] -b3 = 8 -b4 = 10 –b5 = h). The final concatenated super alignment comprised 216,613 amino acid (aa) positions (57.0% missing data/gaps) for dataset 1 and 66,467 aa (31.3% missing data/gaps) for dataset 2.

Phylogenetic analyses

A maximum likelihood (ML) tree was calculated by RAxML 8.2.8 [28] based on dataset 1 using the LG amino acid substitution matrix [29] with empirical amino acid frequencies and assuming a gamma distribution of rates across sites. Bayesian tree inference was performed by PhyloBayes MPI 1.7b [30] based on dataset 2. Eight independent chains were run under the CAT model of sequence evolution [31] with four discrete gamma categories. Every 10th cycle was sampled, and the chains were stopped after 10,000 cycles. After 2500 cycles, all model parameters had entered the stationary phase. A majority rule consensus tree was calculated discarding the first 25% of samples as burn-in from all eight runs. The comparison of bipartitions showed minimal discrepancy among chains (maxdiff value = 0.11) indicating that all eight runs had converged in tree space. Additionally, the bootstrap support values from a ML analysis of dataset 2 (using the same parameters as described above) were mapped onto the Bayesian consensus tree. The resulting trees based on analyses of both datasets were rooted with the chromerid taxa *Chromera velia* and *Vitrella brassicaformis*.

Results and discussion

A data mining approach to identify parasite contamination

The goal of this study was (*i.*) to quantify the extent of contamination by apicomplexan parasites in animal genome and transcriptome assemblies and (*ii.*) to extract as much useful sequence information of parasite origin from these assemblies. A naive, brute force approach to the identification of contaminating sequences might employ a simple blastx query, i.e. searching all contigs of a genome project against a database containing the entire record of publicly available proteomes across all taxa. In a second step, contigs that show the highest similarity to sequences from parasite species could then be extracted as putative contaminants. While such an approach might be feasible for a small number of contigs, it is highly inefficient. The computational resources required to apply this procedure to all available animal genomes exceed even the limits of high-performance computer centers because blastx-style (translated nucleotide vs. protein) searches against large protein databases such as Uniprot are very computationally intensive, especially when using large genomic contigs as query.

In our approach, we drastically reduced the computational complexity of this problem by first filtering the genome data to extract only those contigs that show significant sequence similarity to proteins from apicomplexan parasites (Fig. 1a). By incorporating homology-based gene prediction into the process of contamination identification in the next step (Fig. 1b), we were able to further improve the performance of the search strategy. This allowed us to perform protein vs. protein searches against the UniProt database first (Fig. 1c), which is significantly faster than using the full-length nucleotide contigs as query. Additionally, this step provides high-quality amino acid data for all identified contaminating sequences, which can subsequently be used, e.g., for phylogenetic analyses. After removal of all contigs with a best hit outside of Apicomplexa, the final nucleotide vs. protein searches were performed on a minimal subset of suspect contigs to assess whether they were indeed of apicomplexan origin (Fig. 1d).

Comparison of sequence similarity search tools

To assess whether the performance gains achieved by the ContamFinder pipeline would be sufficient for large-scale analysis of all available genome and transcriptome assembly data, we compared the performance of ContamFinder (employing BLAST+ as search engine) to a naive all all-vs-all blastx search against the UniProt database. Analyses were performed on the transcriptome assembly of the domestic goat, *Capra hircus* (TSA prefix GAOJ01), which contains sequences of coccidian origin, and the comparatively small (14.3 Mb) genome assembly of the white-tailed deer, *Odocoileus virginianus* (WGS prefix AEGY01), infected with a piroplasmid parasite. In both analyses, ContamFinder was able to recover >95% of the hits identified in the all-vs-all blastx search (Table 1) while increasing the speed of the analysis 5-fold for the transcriptome assembly and 30-fold for the genome assembly. The difference in performance gain can be explained by the large amount of non-coding sequence regions in genome data which slow down the blastx search and which are discarded by ContamFinder during the gene prediction step (Fig. 1b). Considering that the total amount of sequence data available from genome assemblies far exceeds that from transcriptome assemblies, these performance metrics are highly favorable for the large scale application of ContamFinder on all available assembly data. However, as most genome assemblies contain much larger amounts of sequence data (in the order of Gb) than the small dataset that was used as a benchmark, we decided to investigate whether the use of alternative amino acid similarity search algorithms could further improve the speed of the analyses. We compared the performance of three local alignment tools (BLAST+ [22], RAPSearch2 [23], GHOSTX [21]). While BLAST+ identified slightly more parasite-derived contigs in both assemblies, GHOSTX and RAPSearch2 were able to speed up the search significantly with an acceptable impact on sensitivity (Table 1). As the amount of computational time required for BLAST+-based analyses of large genome assemblies becomes prohibitively large, we decided to perform all further analyses using GHOSTX, which reduced the run time of ContamFinder 24-fold compared to the BLAST+-based ContamFinder analysis and more than 700-fold compared to a simple blastx all-vs-all search (Table 1). Because in the last step of the pipeline ContamFinder basically performs a blastx all-vs-all search with a drastically reduced query pool (Fig. 1d), all hits from the BLAST+-based ContamFinder analysis were also found in the simple blastx all-vs-all search. When using GHOSTX or RAPSearch2 as the search tool, small numbers (three in each case) of additional hits were found (Fig. 2). Closer inspection of these hits showed that all of them constitute valid parasite-derived contaminations.

Assemblies from aquatic metazoans contain high amounts of protozoan contaminants

For the analysis of apicomplexan parasite contaminations in public databases, we downloaded all available metazoan genome and transcriptome assemblies from the Whole Genome Shotgun (WGS; 658 assemblies) and Transcriptome Shotgun Assembly (TSA; 703 assemblies) databases. Preliminary analyses showed multiple putative apicomplexan species present in most genomes from aquatic species (with aquatic mammals being a notable exception). This may be caused either by infections with multiple parasite species or by contamination of the samples with free-living alveolates closely related to Apicomplexa (e.g. Chromerida). Because the goal of this study was to evaluate and reliably classify the contaminating parasites using multi-gene phylogenetic analyses, which require that each sample only contain a single species, we decided to discard all assemblies from non-mammalian aquatic species and to focus on terrestrial animals. Further analysis of parasite contamination in genomes and transcriptomes from aquatic animals might yield valuable insight into host-parasite associations in aquatic ecosystems.

Genome and transcriptome assemblies of terrestrial animals may contain large amounts of parasite-derived contigs

After removal of 349 assemblies from aquatic species and 59 assemblies from metazoan endoparasites, we performed analyses on the remaining 953 assemblies from terrestrial animals and aquatic mammals (583 Gb). We found contigs of putatively apicomplexan origin in 85 genome and transcriptome projects. The number of identified parasite-derived contigs varied greatly among the contaminated assemblies (Table 2). While most assemblies contained only low to moderate numbers of parasite-derived sequences, we found massive amounts of apicomplexan sequences in the genome assemblies of the northern bobwhite, *Colinus virginianus* (WGS prefix AWGU01; 4,081 contigs), and the duck-billed platypus, *Ornithorhynchus anatinus* (WGS prefix AAPN01; 1,397 contigs). We also found a large number of parasite-derived contigs in the transcriptome assemblies of the oriental tobacco budworm, *Helicoverpa assulta* (TSA prefix GBTA01; 8,347 contigs), the cotton bollworm, *Helicoverpa armigera* (TSA prefix GBDM01; 1,137 contigs) and the stalk-eyed fly, *Teleopsis dalmanni* (TSA prefix GBBP01; 919 contigs). These numbers show that our approach is valid for both genome and transcriptome data. As we were mostly interested in conserved genes for use in phylogenetic analyses, we performed all sequence similarity searches with a strict E-value cut-off of 1e-10. Lowering the E-value cut-off would certainly increase the amount of identified parasite sequences – though at the cost of an increased risk of false positives.

False-positive hits may be caused by low sequence complexity or high conservation

In 35 assemblies, only a single hit was found. Closer inspection revealed that 28 of the single hits were false

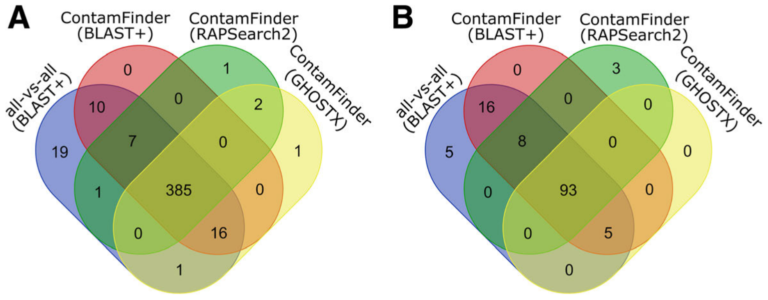


Fig. 2 Venn diagrams showing shared and unique hits from analyses using different search strategies on the assemblies of *Capra hircus* (**a**) and *Odocoileus virginianus* (**b**)

positives, which were either due to highly conserved proteins (20 hits), such as ubiquitin or tubulin, or caused by repetitive sequence patterns (8 hits) that had not been removed by the low complexity filtering step. The exclusion of these conserved proteins from the reference proteomes and the application of advanced filtering methods [32, 33] might alleviate this problem in the future. Among the 50 assemblies with more than one hit, another five were found to be based on small numbers of false positives (2–5 hits). However, the total number of hits identified as false-positive (43 contigs) pales in comparison to the total number of hits from assemblies that are indeed contaminated by parasite sequences (20,907 contigs). Of course, we cannot rule out that the extracted data from these assemblies also contain small numbers of erroneously identified contigs. Large fractions of the extracted contigs (between 20% and 80%, depending on fragmentation of the assembly) also had significant hits against proteins from non-Apicomplexan species. This is to be expected as the majority of apicomplexan genes have detectable homologs in other eukaryotes, especially in the closely related chromerids [34]. We inspected at least 20 (or as many as available) of these contigs for each assembly using single-gene phylogenetic analyses and sequence similarity searches and found no evidence of false-positive hits.

Unambiguous parasite contaminations were found in 51 assemblies

In total, 51 assemblies contained unambiguous contamination by apicomplexan parasites. However, six assemblies were based, at least in part, on the same raw sequencing data or source specimen as other assemblies in our dataset and were therefore removed. Of the remaining 45 assemblies, 11 did not contain sequences that could be assigned to any of the ortholog groups for the multi-gene phylogenetic analysis. In the transcriptome assemblies of *Dendroctonus frontalis* (TSA prefix GAFI01) and *Ixodes ricinus* (TSA prefix GADI01), we found multiple overlapping, yet clearly distinct, sequences of the same single-copy genes. As this indicates the presence of multiple parasite species in the sequenced sample, we also removed these assemblies from the phylogenetic analyses. In the following, we will focus on the 32 assemblies for which orthologous sequences were identified that putatively derived from a single parasite species. We also found overlapping sequences in some of the remaining assemblies. However, in these cases, the sequences were 100% identical in the overlapping regions but differed in length. We assume that poor sequence coverage of the parasite genes may have resulted in fragmented assemblies, though we cannot rule out haplotype variation or the presence of multiple, very closely related parasite species; neither of which should have an effect on the results of our phylogenetic analyses.

The efficiency of curation of publicly available assemblies

The extracted sequence data may prove useful for researchers working on various aspects of parasite biology. The number of parasite-derived contigs in an assembly may depend on several factors, such as source tissue, parasitaemia, sequencing depth or pre- and post-assembly filtering methods to remove low-quality contigs or sequences of unknown origin. In this context, it should be noted that earlier versions of the genome assemblies from the western lowland gorilla, *Gorilla gorilla gorilla* (WGS prefix CABD02), and the platypus, *Ornithorhynchus anatinus* (WGS prefix AAPN01), which were employed in this study, contained large numbers of sequences that originated from apicomplexan parasites. Meanwhile, however, the majority of these contaminating sequences have been removed from the current assembly versions that are available in the public databases (WGS prefix CABD03 for the gorilla; contaminating contigs flagged as ‘dead’ in the AAPN01 record for the platypus).

Our analyses showed that the measures that were taken to remove off-target contigs were reasonably effective (98.0% of contaminants removed from the gorilla assembly and 91.5% from the platypus assembly). It is, of course, desirable that the final genome and transcriptome assemblies contain only high-quality contigs originating

Table 2 Numbers of parasite-derived contigs in publicly available genome and transcriptome assemblies

Host species	WGS/TSA ID	Assembly type	# parasite-derived contigs	# sequences in dataset 1	# sequences in dataset 2
Helicoverpa assulta	GBTA01	transcriptome	8347	370	208
Colinus virginianus	AWGU01	genome	4013	793	244
Colinus virginianus[a]	AWGT01	genome	3098	-	-
Ornithorhynchus anatinus[c]	AAPN01	genome	1397 (119)	540	178
Helicoverpa armigera	GBDM01	transcriptome	1137	160	102
Teleopsis dalmanni	GBBP01	transcriptome	919	339	171
Capra hircus	GAOJ01	transcriptome	405	107	63
Annulipalpia sp.	GATX01	transcriptome	226	81	57
Gorilla gorilla gorilla[c]	CABD02 (CABD03)	genome	148 (3)	33	15
Camelus dromedarius	GADZ01	transcriptome	148	35	25
Anolis carolinensis	GBBS01	transcriptome	120	54	33
Anolis carolinensis[a]	GAFN01	transcriptome	119	-	-
Dendroctonus frontalis[b]	GAFI01	transcriptome	114	-	-
Dastarcus helophoroides	GBCX01	transcriptome	104	29	21
Odocoileus virginianus	AEGY01	genome	98	34	11
Odocoileus virginianus[a]	AEGZ01	genome	98	-	-
Motis davidii	ALWT01	genome	66	9	-
Anolis carolinensis[a]	GAFD01	transcriptome	62	-	-
Orchesella cincta	GAMM01	transcriptome	61	30	27
Ixodes ricinus[b]	GADI01	transcriptome	56	-	-
Corydalinae sp.	GADH01	transcriptome	41	18	-
Pseudomasaris vespoides	GAXQ01	transcriptome	39	18	17
Camelus dromedarius[a]	GADZ01	transcriptome	24	-	-
Ixodes scapularis	ABJB01	genome	26	7	-
Homo sapiens	AADC01	genome	24	6	-
Polyxenus lagurus	GBKF01	transcriptome	21	12	-
Dendroctonus ponderosae	GAFW01	transcriptome	15	6	-
Amblyomma americanum	GAGD01	transcriptome	10	4	-
Carduelis chloris	GBCG01	transcriptome	8	-	-
Capra hircus	GAOE01	transcriptome	8	-	-
Ixodes ricinus	GANP01	transcriptome	7	5	-
Camelus bactrianus	GAEY01	transcriptome	7	2	
Dendroctonus ponderosae[a]	GAFX01	transcriptome	6	-	-
Chrysochloris asiatica	AMDV01	genome	5	2	-
Cuculus canorus	JNOX01	genome	5	2	-
Bos mutus	AGSK01	transcriptome	5	1	-
Nevrorthus apatelios	GACU01	transcriptome	4	3	-
Fulmarus glacialis	JJRN01	genome	4	2	-
Forficula auricula	GAAX01	transcriptome	4	3	-
Serinus canaria	CAVT01	genome	3	2	-
Capra hircus	GAFC01	transcriptome	3	2	-
Balaenoptera bonaerensis	BAUQ01	genome	2	1	-
Blattela germanica	GBID01	transcriptome	2	-	-
Folsomia candida	GAMN01	transcriptome	2	-	-

Table 2 Numbers of parasite-derived contigs in publicly available genome and transcriptome assemblies *(Continued)*

Carabus granulatus	GACW01	transcriptome	1	-	-
Capra hircus	GAOG01	transcriptome	1	-	-
Nemurella pictetii	GAAV01	transcriptome	1	-	-
Anolis carolinensis	GADN01	transcriptome	1	-	-
Phaedon cochleariae	GAPU01	transcriptome	1	-	-
Gluvia dorsalis	GDAP01	transcriptome	1	-	-
Rhipicephalus microplus	ADMZ02	genome	1	-	-

[a] Assembly was not used in phylogenetic analyses because it is based on the same raw data as another assembly
[b] Assembly was not used in phylogenetic analyses because it contains sequences from multiple parasite species
[c] Data based on a superseded assembly version; the number of parasite-derived contigs in the current version is given in parentheses

exclusively from the target species. However, we argue that the uncurated assemblies should also be made available to the research community because they constitute a valuable resource for data mining approaches and may allow us to gain insights into the pathogens infecting the target species.

Phylogenetic classification of the contaminating parasites

To understand the phylogenetic origin of the contaminating parasites, the extracted amino acid sequences were assigned to ortholog groups and used in a multi-gene phylogenetic analysis. The final dataset comprised 1,420 genes from 32 parasite contaminations and 35 previously sequenced apicomplexan and chromerid genomes (dataset 1). The phylogenetic analysis identified the contaminating parasites in the metazoan genome and transcriptome assemblies as members of the apicomplexan taxa Gregarinasina, Coccidia, Piroplasmida and Haemosporida (Fig. 3).

Contaminations by gregarine parasites were found in 12 assemblies, all of which were derived from arthropod transcriptomes. This observation is in line with gregarine life history, as these parasites are only found in invertebrate hosts [35]. Due to the lack of medical or veterinary importance of Gregarinasina, this taxon has essentially been neglected in genome sequencing efforts. Only a single gregarine draft genome is available from *Gregarina niphandroides* and a highly fragmented assembly from *Ascogregarina taiwanensis* that was estimated to contain 25% of the parasite's genome. Yet, Gregarinasina constitute a key taxon for understanding the evolutionary history of Apicomplexa because of their basal position within the phylum. The extracted contaminating contigs significantly increase the amount of available sequence data from gregarine parasites and may prove to be a valuable resource for researchers studying the molecular evolution of these parasites.

In 11 assemblies from vertebrates, we identified contaminations by coccidian parasites, including the previously described contaminations in the genomes of *Myotis davidii* and *Colinus virginianus* [9]. In that study, the contaminations were identified by searching for a gene (apicortin) that is specific for apicomplexan parasites but absent from metazoan genomes. This method requires only few computational resources and is unlikely to produce false positives, as any significant hit is a clear indication of contamination. A similar methodology has recently been employed to identify sequences originating from insect pests in plant transcriptomes [10]. However, such an approach is bound to miss a large number of contaminations as it relies on a small, specific set of genes to be present in the (incomplete) assembly. Additionally, conserved genes which are suitable for deep-level phylogenetic analyses are rarely specific to a certain clade and often have homologs in extremely distantly related taxa. By targeting the whole parasite proteome, we are able to overcome these limitations for the identification and extraction of contaminating sequences.

In the assemblies of a human genome (WGS prefix AADC01) and the genome of the western lowland gorilla (WGS prefix CABD02), we found sequences that are ≥99.9% identical at the nucleotide level to sequences from the most virulent agent of human malaria, *Plasmodium falciparum*. The complete mitochondrial genome of the parasite is present in the superseded version of the gorilla genome assembly (EMBL/Genbank acc. nos. CABD02435943 and CABD02435942). The sequences are clearly more closely related to those from *P. falciparum* than to those from any known ape-infecting parasite (Additional file 1: Figure S1), including the *P. falciparum*-like parasites that have been reported from western lowland gorillas [36]. Additionally, exposure to parasites from wild gorillas seems implausible considering that the animal was born and raised in a North American zoo [37]. We, therefore, conclude that contamination with parasite DNA in the lab or at the sequencing center is the likely explanation in this case, though we cannot formally rule out an infection of the gorilla with *P. falciparum*. Taking into account that all other host-parasite associations that we found fit well with parasite biology (i.e. gregarines only in invertebrates, piroplasmids in tick vectors and vertebrate hosts), we consider infection of the sequenced

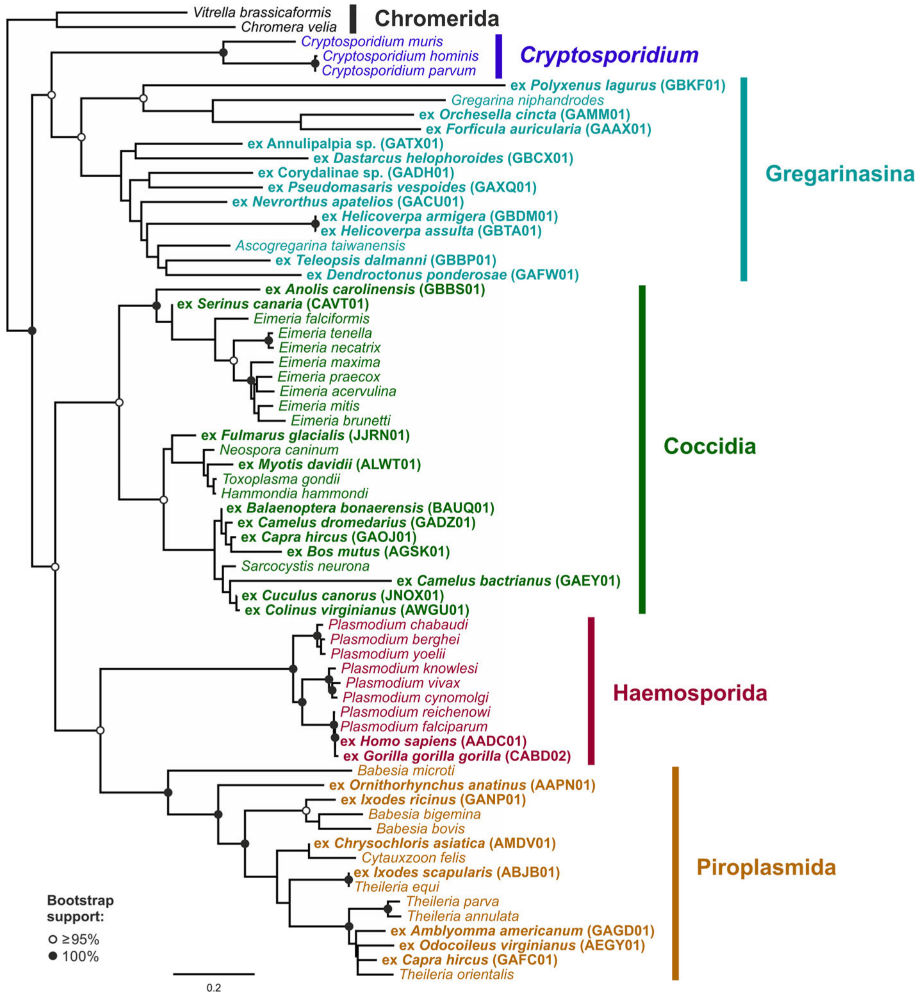

Vitrella brassicaformis
Chromera velia
Chromerida
Cryptosporidium muris
Cryptosporidium hominis
Cryptosporidium parvum
Cryptosporidium
ex Polyxenus lagurus (GBKF01)
Gregarina niphandrodes
ex Orchesella cincta (GAMM01)
ex Forficula auricularia (GAAX01)
ex Annulipalpia sp. (GATX01)
ex Dastarcus helophoroides (GBCX01)
ex Corydalinae sp. (GADH01)
ex Pseudomasaris vespoides (GAXQ01)
ex Nevrorthus apatelios (GACU01)
ex Helicoverpa armigera (GBDM01)
ex Helicoverpa assulta (GBTA01)
Ascogregarina taiwanensis
ex Teleopsis dalmanni (GBBP01)
ex Dendroctonus ponderosae (GAFW01)
Gregarinasina
ex Anolis carolinensis (GBBS01)
ex Serinus canaria (CAVT01)
Eimeria falciformis
Eimeria tenella
Eimeria necatrix
Eimeria maxima
Eimeria praecox
Eimeria acervulina
Eimeria mitis
Eimeria brunetti
ex Fulmarus glacialis (JJRN01)
Neospora caninum
ex Myotis davidii (ALWT01)
Toxoplasma gondii
Hammondia hammondi
ex Balaenoptera bonaerensis (BAUQ01)
ex Camelus dromedarius (GADZ01)
ex Capra hircus (GAOJ01)
ex Bos mutus (AGSK01)
Sarcocystis neurona
ex Camelus bactrianus (GAEY01)
ex Cuculus canorus (JNOX01)
ex Colinus virginianus (AWGU01)
Coccidia
Plasmodium chabaudi
Plasmodium berghei
Plasmodium yoelii
Plasmodium knowlesi
Plasmodium vivax
Plasmodium cynomolgi
Plasmodium reichenowi
Plasmodium falciparum
ex Homo sapiens (AADC01)
ex Gorilla gorilla gorilla (CABD02)
Haemosporida
Babesia microti
ex Ornithorhynchus anatinus (AAPN01)
ex Ixodes ricinus (GANP01)
Babesia bigemina
Babesia bovis
ex Chrysochloris asiatica (AMDV01)
Cytauxzoon felis
ex Ixodes scapularis (ABJB01)
Theileria equi
Theileria parva
Theileria annulata
ex Amblyomma americanum (GAGD01)
ex Odocoileus virginianus (AEGY01)
ex Capra hircus (GAFC01)
Theileria orientalis
Piroplasmida
Bootstrap support:
○ ≥95%
● 100%
0.2

Fig. 3 Maximum likelihood tree based on a RAxML analysis of dataset 1 (1,420 genes, 67 taxa). The tree was rooted with Chromerida

organism as the most likely source of parasite contamination in the other assemblies.

Contaminations with piroplasmid parasites were found in the assemblies of tick vectors (*Amblyomma americanum*, *Ixodes ricinus*, *Ixodes scapularis*), as well as in putative vertebrate hosts (*Chrysochloris asiatica*, *Capra hircus*, *Odocoileus virginianus*, *Ornithorhynchus anatinus*). A recent study by Paparini et al. [38] has provided the first molecular data from *Theileria ornithorhynchi*, a piroplasmid parasite of the platypus. In a blastn search of piroplasmid 18S rRNA sequences against the platypus genome assembly [39], we identified a contig of piroplasmid origin encoding a fragment of the parasite's 18S rRNA (EMBL/Genbank acc. nr. AAPN01188453). A phylogenetic analysis based on the dataset of Paparini et al. [38] indeed recovered this contig closely associated with the sequences from *T. ornithorhynchi* (Additional file 2: Figure S2). We also found a small number of sequences derived from a piroplasmid parasite in the genome assembly of the Cape golden mole (*Chrysochloris asiatica*; WGS prefix AMDV01). To the best of our knowledge, this is the first report of a piroplasmid infection in mammals belonging to the order Afrosoricida. The extracted sequences from the genome assembly of the blacklegged tick, *I. scapularis*, are identical to sequences from the equine parasite *Theileria equi*. While *I. scapularis* has not been described as a vector of this species, the sequenced ticks were fed on sheep [40], which may be natural hosts of *T. equi* [41]. However, a cautionary note is required: The presence of parasite DNA in the blood or tissue of a putative host indicates that the animal is naturally subjected to the parasite and that the parasite can develop in the host, but it does not prove that the parasite is able to complete its complex life cycle within the host and infect a new host.

Deep phylogeny of Apicomplexa

The advent of molecular phylogenetics has challenged several longstanding views on the relationships among apicomplexan taxa, such as the monophyly of *Plasmodium* parasites [42, 43] or the inclusion of *Cryptosopidium* in Coccidia [44, 45]. The deep-level phylogenetic relationships of our tree are in good agreement with the current view on apicomplexan phylogeny. Like previous molecular studies [44, 46], we found a sister group relationship between *Cryptosporidium* and the gregarines at the base of Apicomplexa. Both parasite taxa appear to have lost their plastid genomes [47, 48] and also share numerous molecular similarities [46]. Piroplasmida and Haemosporida were united in a clade to the exclusion of Coccidia. Within Piroplasmida, *Babesia* was found to be paraphyletic – a finding that is congruent with the results of Schnittger et al. [49], who inferred six major monophyletic piroplasmid lineages based on all available 18S rRNA data. The authors concluded that a robust phylogeny based on multi-gene data might be required before re-interpretation of traditional characters could reconcile morphological and molecular data. A recent study on the phylogenetic relationships of *Theileria ornithorhynchi*, a parasite of the monotreme platypus, placed this species outside the clade of the theilerids and basal to all other piroplasms [38]. However, the results were inconclusive as this relationship was only recovered in the analysis of 18S rRNA data, while tree inference using the heat shock protein 70 (Hsp70) resulted in a markedly different phylogeny. Dataset 1 contains data from the platypus parasite for 540 orthologous genes. The resulting tree supported the tentative placement of *Theileria ornithorhynchi* based on 18S rRNA with maximum support. We found good support (92% bootstrap support) for a placement of the afrosoricid parasite extracted from *Chrysochloris asiatica* within the clade comprising all other *Theileria* parasites and *Cytauxzoon*. However, due to the low amount of data available for this species (only two genes present in dataset 1), its exact phylogenetic position remains unresolved (Fig. 3).

Phylogenetic analyses based on a reduced dataset that only contains the genes and taxa with the highest coverage (dataset 2) yielded a tree that is fully congruent with the results from the first analysis but with maximum support for nearly all splits (Fig. 4). This indicates that the reduced support for some deep-level splits in the first analysis is not due to conflict in the phylogenetic signal but rather due to the unstable positioning of some taxa with very low gene coverage.

Conclusion

We were able to extract 20,907 parasite-derived contigs from 51 publicly available genome and transcriptome assemblies employing a new bioinformatic pipeline. Our results show that contaminations in sequencing data are

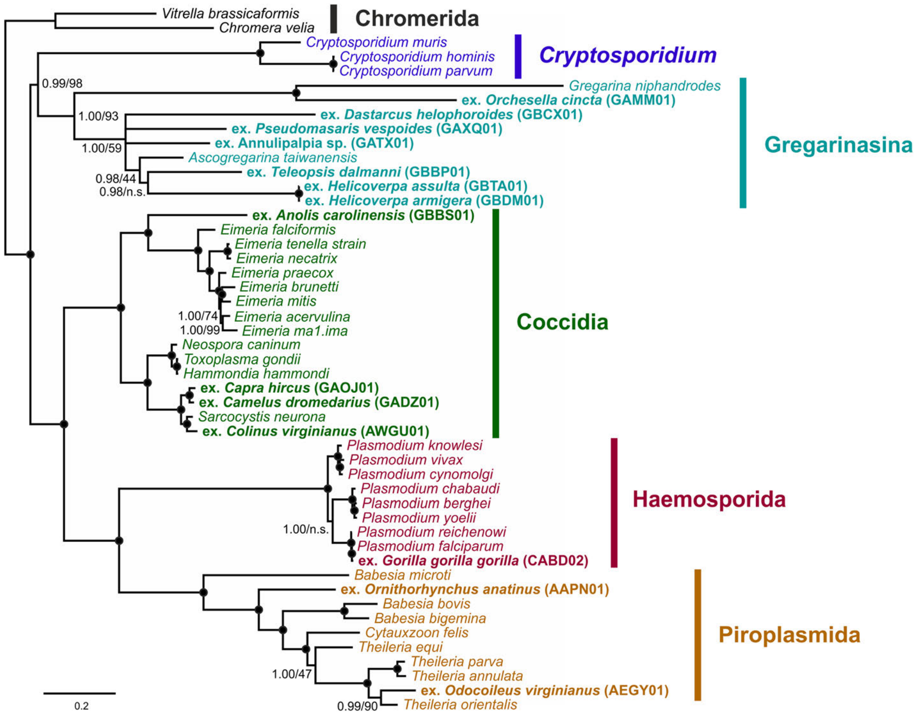


Fig. 4 Majority-rule consensus tree based on a PhyloBayes analysis of dataset 2 (301 genes, 49 taxa). Bootstrap support values from a RAxML analysis were mapped onto the tree topology. Bayesian posterior probabilities < 1.00 and bootstrap support values < 100% are given at the nodes, respectively; n.s.: split was not supported in the ML analysis; splits that have 1.00 posterior probability and 100% bootstrap support are denoted by a dark circle. The tree was rooted with Chromerida

not just a problem that needs to be eliminated but that they also constitute a valuable, cost-efficient source of information. Analysis of contaminations may enable the discovery and identification of novel parasite taxa and shed light on previously unknown host-parasite interactions. Our approach is not only valid for the identification of apicomplexan parasites but can also be used to study contaminations by other pathogens, such as bacteria or viruses. Most genomic and transcriptomic studies only make the raw sequencing data and the final curated and annotated assemblies available to the public. While these datasets are obviously most relevant to and useful for the subject of study, we argue that uncurated assemblies may contain valuable information from unexpected sources and should, therefore, routinely be made available.

Additional files

Additional file 1: Figure S1. Majority-rule consensus tree based on a PhyloBayes analysis of complete mitochondrial genomes from ape-infecting *Plasmodium* parasites. The alignment is based on the mitochondrial dataset from Liu et al. (2010) and only contains sequences from Clades C1 (from Chimpanzees) and G1 (from Gorillas; also contains human *P. falciparum*). Two contigs from the Gorilla genome assembly, which contain parasite-derived mitochondrial fragments, were added to the alignment. Bayesian posterior probabilities are given at the nodes. The tree was rooted with the C1 clade of Chimpanzee-infecting *Plasmodium* parasites. All EMBL/Genbank acc. nos. are given in parentheses. (PDF 236 kb)

Additional file 2: Figure S2. Majority-rule consensus tree based on a PhyloBayes analysis of 18 s rRNA sequences from Piroplasmida. The alignment is based on the 18 s dataset from Paparini et al. (2015). A single contig from the platypus genome assembly, which contains a parasite-derived 18 s rRNA fragment, was added to the alignment. Bayesian posterior probabilities are given at the nodes. The tree was rooted with *Cardiosporidium cionae*. All EMBL/Genbank acc. nos. are given in parentheses. (PDF 157 kb)

Abbreviations

Aa: Amino acid; EuPathDB: Eukaryotic Pathogen Database; Gb: Giga base pairs; Hsp70: Heat shock protein 70; Mb: Mega base pairs; ML: Maximum likelihood; NGS: Next-generation sequencing; TSA: Transcriptome Shotgun Assembly; WGS: Whole Genome Shotgun

Acknowledgments

We thank Christian Pick for his support during the initial phase of the project.

Funding

This work has been supported by the Deutsche Forschungsgemeinschaft (Bu956/16-1).

Availability of data and material

The software pipeline used to extract contigs of parasite origin is freely available from SourceForge: https://sourceforge.net/projects/contamfinder. All extracted contigs, predicted amino acid sequences, single gene alignments and concatenated super alignments are publicly available from the Dryad Digital Repository (http://datadryad.org) at http://dx.doi.org/10.5061/dryad.mn338.

Authors' contributions

Conception and design of the experiments: JB, TB. Performed research: JB. Analysis and interpretation of data: JB, TB. Wrote the paper: JB, TB. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Received: 22 September 2016 Accepted: 14 January 2017
Published online: 19 January 2017

References

1. Naccache SN, Greninger AL, Lee D, Coffey LL, Phan T, Rein-Weston A, Aronsohn A, Hackett JJ, Delwart EL, Chiu CY. The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. J Virol. 2013;87:11966–77.
2. Laurence M, Hatzis C, Brash DE. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. PLoS One. 2014;9:e97876.
3. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014;12:87.
4. Merchant S, Wood DE, Salzberg SL. Unexpected cross-species contamination in genome sequencing projects. PeerJ. 2014;2:e675.
5. Tao Z, Sui X, Jun C, Culleton R, Fang Q, Xia H, Gao Q. Vector sequence contamination of the *Plasmodium vivax* sequence database in PlasmoDB and *in silico* correction of 26 parasite sequences. Parasit Vectors. 2015;8:318.
6. Schmieder R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One. 2011;6:e17288.
7. Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, Boehnke M, Kang HM. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet. 2012;91:839–48.
8. Strong MJ, Xu G, Morici L, Splinter Bon-Durant S, Baddoo M, Lin Z, Fewell C, Taylor CM, Flemington EK. Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples. PLoS Pathog. 2014;10:e1004437.
9. Orosz F. Two recently sequenced vertebrate genomes are contaminated with apicomplexan species of the Sarcocystidae family. Int J Parasitol. 2015;45:871–8.
10. Zhu J, Wang G, Pelosi P. Plant transcriptomes reveal hidden guests. Biochem Biophys Res Commun. 2016;474:497–502.
11. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15:R46.
12. Liu B, Gibbons T, Ghodsi M, Treangen T, Pop M. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. BMC Genomics. 2011;12 Suppl 2:S4.
13. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods. 2012;9:811–4.
14. Kostic AD, Ojesina AI, Pedamallu CS, Jung J, Verhaak RGW, Getz G, Meyerson M. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nat Biotechnol. 2011;29:393–6.
15. World Health Organization. World malaria report 2015. Geneva, Switzerland: World Health Organisation; 2015.
16. Williams RB. A compartmentalised model for the estimation of the cost of coccidiosis to the world's chicken production industry. Int J Parasitol. 1999; 29:1209–29.
17. Whole Genome Shotgun Database. National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/genbank/wgs. Accessed on 22 Sept 2015.
18. Transcriptome Shotgun Assembly Database. National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/genbank/tsa. Accessed on 22 Sept 2015.
19. Eukaryotic Pathogen Database. http://eupathdb.org/eupathdb. Accessed on 1 Aug 2015.
20. Aurrecoechea C, Barreto A, Brestelli J, Brunk BP, Cade S, Doherty R, Fischer S, Gajria B, Gao X, Gingle A, et al. EuPathDB: the eukaryotic pathogen database. Nucleic Acids Res. 2013;41:D684–91.

21. Suzuki S, Kakuta M, Ishida T, Akiyama Y. GHOSTX: an improved sequence homology search algorithm using a query suffix array and a database suffix array. PLoS One. 2014;9:e103833.
22. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
23. Zhao Y, Tang H, Ye Y. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. Bioinformatics. 2012;28:125–6.
24. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005;6:31.
25. Chen F, Mackey AJ, Stoeckert CJJ, Roos DS. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res. 2006;34:D363–8.
26. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
27. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17:540–52.
28. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.
29. Le SQ, Gascuel O. An improved general amino acid replacement matrix. Mol Biol Evol. 2008;25:1307–20.
30. Lartillot N, Rodrigue N, Stubbs D, Richer J. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst Biol. 2013;62:611–5.
31. Lartillot N, Philippe H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol. 2004;21:1095–109.
32. Shin SW, Kim SM. A new algorithm for detecting low-complexity regions in protein sequences. Bioinformatics. 2005;21:160–70.
33. Li X, Kahveci T. A novel algorithm for identifying low-complexity regions in a protein sequence. Bioinformatics. 2006;22:2980–7.
34. Woo YH, Ansari H, Otto TD, Klinger CM, Kolisko M, Michalek J, Saxena A, Shanmugam D, Tayyrov A, Veluchamy A, et al. Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. Elife. 2015;4:e06974.
35. Desportes I. Systematics of Terrestrial and Fresh Water Gregarines. In: Desportes I, Schrével J, editors. Treatise on Zoology - Anatomy, Taxonomy, Biology. The Gregarines. Leiden: Brill NV; 2013. p. 377–710.
36. Liu W, Li Y, Learn GH, Rudicell RS, Robertson JD, Keele BF, Ndjango JN, Sanz CM, Morgan DB, Locatelli S, et al. Origin of the human malaria parasite *Plasmodium falciparum* in gorillas. Nature. 2010;467:420–5.
37. Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T, et al. Insights into hominid evolution from the gorilla genome sequence. Nature. 2012;483:169–75.
38. Paparini A, Macgregor J, Ryan UM, Irwin PJ. First molecular characterization of *Theileria ornithorhynchi* Mackerras, 1959: yet another challenge to the systematics of the Piroplasms. Protist. 2015;166:609–20.
39. Warren WC, Hillier LW, Marshall Graves JA, Birney E, Ponting CP, Grützner F, Belov K, Miller W, Clarke L, Chinwalla AT, et al. Genome analysis of the platypus reveals unique signatures of evolution. Nature. 2008;453:175–83.
40. Ayllon N, Villar M, Galindo RC, Kocan KM, Sima R, Lopez JA, Vazquez J, Alberdi P, Cabezas-Cruz A, Kopacek P, de la Fuente J. Systems Biology of Tissue-Specific Response to *Anaplasma phagocytophilum* Reveals Differentiated Apoptosis in the Tick Vector *Ixodes scapularis*. PLoS Genet. 2015;11:e1005120.
41. Zhang J, Kelly P, Li J, Xu C, Wang C. Molecular detection of *Theileria* spp. in livestock on five Caribbean islands. BioMed Res Int. 2015;2015:624728.
42. Outlaw DC, Ricklefs RE. Rerooting the evolutionary tree of malaria parasites. Proc Natl Acad Sci U S A. 2011;108:13183–7.
43. Schaer J, Perkins SL, Decher J, Leendertz FH, Fahr J, Weber N, Matuschewski K. High diversity of West African bat malaria parasites and a tight link with rodent *Plasmodium* taxa. Proc Natl Acad Sci U S A. 2013;110:17415–9.
44. Carreno RA, Martin DS, Barta JR. *Cryptosporidium* is more closely related to the gregarines than to coccidia as shown by phylogenetic analysis of apicomplexan parasites inferred using small-subunit ribosomal RNA gene sequences. Parasitol Res. 1999;85:899–904.
45. Zhu G, Keithly JS, Philippe H. What is the phylogenetic position of *Cryptosporidium*? Int J Syst Evol Microbiol. 2000;50(Pt 4):1673–81.
46. Templeton TJ, Enomoto S, Chen W, Huang C, Lancto CA, Abrahamsen MS, Zhu G. A genome-sequence survey for *Ascogregarina taiwanensis* supports evolutionary affiliation but metabolic diversity between a Gregarine and *Cryptosporidium*. Mol Biol Evol. 2010;27:235–48.
47. Zhu G, Marchewka MJ, Keithly JS. *Cryptosporidium parvum* appears to lack a plastid genome. Microbiology. 2000;146(Pt 2):315–21.
48. Toso MA, Omoto CK. *Gregarina niphandrodes* may lack both a plastid genome and organelle. J Eukaryot Microbiol. 2007;54:66–72.
49. Schnittger L, Rodriguez AE, Florin-Christensen M, Morrison DA. *Babesia*: a world emerging. Infect Genet Evol. 2012;12:1788–809.

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Phylogeny of haemosporidian blood parasites revealed by a multi-gene approach ☆

Janus Borner[a], Christian Pick[a], Jenny Thiede[a], Olatunji Matthew Kolawole[b], Manchang Tanyi Kingsley[c], Jana Schulze[a], Veronika M. Cottontail[d], Nele Wellinghausen[e], Jonas Schmidt-Chanasit[f], Iris Bruchhaus[f], Thorsten Burmester[a,]*

[a] *Institute of Zoology and Zoological Museum, University of Hamburg, Martin-Luther-King-Platz 3, D-20146 Hamburg, Germany*
[b] *Department of Microbiology, Faculty of Life Sciences, University of Ilorin, PMB 1515, Ilorin, Kwara State, Nigeria*
[c] *Institute of Agricultural Research for Development, Veterinary Research Laboratory, Wakwa Regional Center, PO Box 65, Ngaoundere, Cameroon*
[d] *Institute of Experimental Ecology, University of Ulm, Albert-Einstein Allee 11, D-89069 Ulm, Germany*
[e] *Gaertner & Colleagues Laboratory, Elisabethenstr. 11, D-88212 Ravensburg, Germany*
[f] *Bernhard Nocht Institute for Tropical Medicine, Bernhard-Nocht-Str. 74, D-20359 Hamburg, Germany*

ARTICLE INFO

Article history:
Received 28 February 2015
Revised 31 August 2015
Accepted 3 September 2015
Available online 11 September 2015

Keywords:
Alveolata
Apicomplexa
Haemosporida
Malaria
Parasitology

ABSTRACT

The apicomplexan order Haemosporida is a clade of unicellular blood parasites that infect a variety of reptilian, avian and mammalian hosts. Among them are the agents of human malaria, parasites of the genus *Plasmodium*, which pose a major threat to human health. Illuminating the evolutionary history of Haemosporida may help us in understanding their enormous biological diversity, as well as tracing the multiple host switches and associated acquisitions of novel life-history traits. However, the deep-level phylogenetic relationships among major haemosporidian clades have remained enigmatic because the datasets employed in phylogenetic analyses were severely limited in either gene coverage or taxon sampling. Using a PCR-based approach that employs a novel set of primers, we sequenced fragments of 21 nuclear genes from seven haemosporidian parasites of the genera *Leucocytozoon*, *Haemoproteus*, *Parahaemoproteus*, *Polychromophilus* and *Plasmodium*. After addition of genomic data from 25 apicomplexan species, the unreduced alignment comprised 20,580 bp from 32 species. Phylogenetic analyses were performed based on nucleotide, codon and amino acid data employing Bayesian inference, maximum likelihood and maximum parsimony. All analyses resulted in highly congruent topologies. We found consistent support for a basal position of *Leucocytozoon* within Haemosporida. In contrast to all previous studies, we recovered a sister group relationship between the genera *Polychromophilus* and *Plasmodium*. Within *Plasmodium*, the sauropsid and mammal-infecting lineages were recovered as sister clades. Support for these relationships was high in nearly all trees, revealing a novel phylogeny of Haemosporida, which is robust to the choice of the outgroup and the method of tree inference.

1. Introduction

Haemosporida are protozoan blood parasites with complex life cycles that infect a great variety of vertebrate hosts. Haemosporidians are member of the phylum Apicomplexa and include the genus *Plasmodium*. At least five *Plasmodium* species have independently acquired the ability to infect humans (Escalante et al., 1995; Cox-Singh et al., 2008). As agents of human malaria, *Plasmodium* parasites are one of the greatest threats to human health (WHO, 2013). Surveys of blood parasites in vertebrate wildlife have revealed a rich diversity of haemosporidian lineages infecting reptiles, birds and mammals (e.g., Ricklefs and Fallon, 2002; Duval et al., 2007; Valkiūnas et al., 2008; Bensch et al., 2009; Chasar et al., 2009). However, due to their importance in medical research, most studies have focused on *Plasmodium* species of primates and rodents. Large-scale genome sequencing efforts have produced several complete genomes of these parasites (e.g., Carlton et al., 2002, 2008; Gardner et al., 2002; Hall et al., 2005; Pain et al.,

Abbreviations: asl, adenylosuccinate lyase; *clpc*, caseinolytic protease; *col*, cytochrome oxidase I; *cytb*, cytochrome b; ML, maximum likelihood; MP, maximum parsimony.

☆ This paper was edited by the Associate Editor J.B. Dacks.

* Corresponding author.

E-mail address: thorsten.burmester@uni-hamburg.de (T. Burmester).

http://dx.doi.org/10.1016/j.ympev.2015.09.003

2008; Tachibana et al., 2012; Otto et al., 2014), while other key taxa for the understanding of haemosporidian evolution are only scarcely represented in public databases. Therefore, deep-level phylogenetic relationships among major haemosporidian lineages are still poorly resolved. Yet, our understanding of the emergence of new diseases and the acquisition of novel life-history traits by parasites depends on the knowledge of a solid phylogenetic backbone (Lefevre et al., 2007).

The order Haemosporida currently contains 15 extant genera, though the vast majority of the more than 500 described species have been assigned to the four genera *Plasmodium*, *Hepatocystis*, *Haemoproteus* and *Leucocytozoon*. Members of the genus *Plasmodium* infect a wide range of vertebrate hosts, whereas *Leucocytozoon* and *Haemoproteus* are limited to sauropsids and *Hepatocystis* is only found in mammals (predominantly bats and primates). Some of the other genera only contain a single described species for which molecular data is not available and the taxonomic status of some genera remains uncertain (see Perkins (2014) for a review of the history of haemosporidian systematics), e.g. some authors favored splitting both *Haemoproteus* and *Leucocytozoon* into two genera (Bennett et al., 1965; Martinsen et al., 2008) and several studies found *Plasmodium* to be paraphyletic (e.g. Perkins and Schall, 2002; Outlaw and Ricklefs, 2011). All haemosporidian parasites share similar life cycles. They use blood-feeding dipterans as vectors (Garnham, 1966; Valkiūnas, 2005). Sexual reproduction occurs in the gut of the vector and the infectious sporozoites develop in the salivary glands. When the vector feeds on a vertebrate host, the sporozoites enter the blood stream and invade hepatocytes or endothelial cells. In these cells, the parasites undergo the first cycle of schizogony. Once released, the merozoites infect new cells of various tissues where they undergo another cycle of schizogony (Garnham, 1966). In contrast to most other haemosporidians, *Plasmodium* parasites also undergo schizogony in erythrocytes (Garnham, 1966; Valkiūnas, 2005). Within erythrocytes or leucocytes, the merozoites develop into gametocytes, which can then infect a new vector. Except *Leucocytozoon*, most haemosporidians form a characteristic pigment in the red blood cells called hemozoin, which is a crystalline metabolite from hemoglobin digestion by the parasite (Goldberg et al., 1990).

Before the advent of DNA sequencing methodologies, the classification of haemosporidian parasites solely relied on their morphology, their life-history traits, and the taxonomy of the infected vertebrate hosts and insect vectors (e.g., Garnham, 1966). Based on these characters, early reconstructions of haemosporidian phylogeny concluded that the most parsimonious tree comprises a monophyletic group of *Plasmodium* parasites, which exhibit the most derived traits (i.e. schizogony in the red blood cells of the vertebrate host, formation of hemozoin pigment), whereas *Leucocytozoon*, which lacks these traits, was placed at the base of Haemosporida. However, the significance of these characters for use in phylogenetic analyses had been questioned long before the first genetic sequences became available (e.g., Manwell, 1957; Garnham, 1966). Morphological traits seen under the light microscope can be distorted by preservation and only give an approximate representation of the underlying three-dimensional structure of the parasites (Martinsen et al. (2008) compared it to "systematic study of insects based on remains seen on automobile windshields"). Life-history traits, such as the production of hemozoin pigment or the types of host cells used for schizogony, could have evolved convergently on the basis of similar ecological pressures. While host switches between distantly related hosts have long been regarded as major events in the evolution of Haemosporida (Garnham, 1966), this view has been challenged by recent evidence for multiple host switches between birds and bats (Duval et al., 2007; Witsenburg et al., 2012).

A major point of contention concerning the haemosporidian phylogeny is the position of the root. Early molecular analyses were limited to single gene fragments. In a study based on the mitochondrial gene cytochrome b (*cytb*) using the piroplasmid *Theileria annulata* as outgroup, Perkins and Schall (2002) supported a basal position of *Leucocytozoon*. Hagner et al. (2007), by contrast, employed fragments of three genes (including *cytb*) for independent phylogenetic reconstructions and concluded that none of the analyzed genes alone contained sufficient phylogenetic information to resolve deep-level relationships. A multigene analysis based on four genes (Martinsen et al., 2008) resulted in a topology with high support for most splits. However, the tree was rooted with *Leucocytozoon* and did not include any non-haemosporidian outgroup taxa because the outgroup sequences were considered too divergent. To address this issue, Outlaw and Ricklefs (2011) reevaluated the dataset of Martinsen et al. (2008) using an outgroup-free molecular clock approach for rooting. In the resulting tree, Haemosporida are split into two major clades, one comprising all mammalian *Plasmodium* lineages (plus *Hepatocystis*), the other uniting the sauropsid parasites.

Originally, all avian parasites that produce hemozoin pigment but do not undergo schizogony in the red blood cells were classified as members of the genus *Haemoproteus*. Bennett et al. (1965) proposed splitting *Haemoproteus* into two genera, *Haemoproteus* and *Parahaemoproteus*. *Haemoproteus sensu* Bennett et al. (1965) comprises the parasites that use hippoboscid flies as vectors while *Parahaemoproteus* relies on mosquitoes for transmission. Molecular analyses mostly recovered these two groups of parasites as distinct lineages. However, the taxonomic status of *Haemoproteus* remained uncertain, because some studies favored a sister group relationship between both clades, thereby supporting a single genus *Haemoproteus* divided into two subgenera (Iezhova et al., 2011; Pineda-Catalan et al., 2013), while other analyses found this taxon to be paraphyletic (Martinsen et al., 2008; Witsenburg et al., 2012).

The phylogenetic placement of the bat-infecting genera *Hepatocystis*, *Polychromophilus* and *Nycteria* has proven especially troublesome. In contrast to *Plasmodium* parasites, they lack the ability to reproduce asexually in erythrocytes (blood schizogony). However, studies based on molecular data have consistently recovered them nested within *Plasmodium*. *Hepatocystis* was found to be closely associated with mammalian *Plasmodium* in numerous analyses (e.g., Escalante et al., 1998; Perkins and Schall, 2002; Martinsen et al., 2008). Witsenburg et al. (2012) expanded the four-gene dataset (Martinsen et al., 2008) to include two species of *Polychromophilus* and recovered this taxon closely related to the clade of sauropsid-infecting *Plasmodium*, similar to the results of Duval et al. (2007) and Megali et al. (2011). Schaer et al. (2013) increased the taxon sampling of bat parasites by adding various species of the genera *Plasmodium*, *Hepatocystis*, *Nycteria* and *Polychromophilus* and found *Polychromophilus* to be most closely related to a clade comprising *Nycteria* and the mammalian lineage of *Plasmodium* and *Hepatocystis*.

The majority of recent studies found *Plasmodium* to be paraphyletic with regard to the chiropteran haemosporidians (see above), Outlaw and Ricklefs (2011) even recovered the genus *Plasmodium* as a polyphyletic group and placed the mammalian *Plasmodium* lineage at the base of Haemosporida. Despite these marked differences in topology, analyses based on single genes or on variations of the four-gene dataset of Martinsen et al. (2008) have generally recovered a monophyletic group comprising all mammalian Plasmodium species (also including *Hepatocystis*). By contrast, a phylogenetic analysis of the available genome data (Pick et al., 2011) found a close relationship between the avian parasite *P. gallinaceum* and the most malignant agent of human malaria, *P. falciparum*. The dataset employed in this analysis comprises 218 full length genes and thus provides much higher sequence coverage than previous phylogenetic analyses. However, the study is limited to taxa with fully sequenced nuclear genomes, which were available for only eight *Plasmodium* species, whereas all other haemosporidian genera are not included in the analysis.

Most molecular phylogenetic studies employed only small numbers of genes (one to four), and relied primarily on mitochondrial and apicoplast sequences, as the development of nuclear gene markers that are effective across the diverse lineages of Haemosporida has proven to be very challenging (Perkins, 2014). These datasets were not suited for the inclusion of distant outgroups and have therefore been unable to resolve the deep-level phylogenetic relationships among major haemosporidian clades robustly. We tackle the problem of rooting the evolutionary tree of malaria parasites by trying to strike a balance between gene coverage and taxon sampling. For the first time, we developed a large number of nuclear markers that amplify gene fragments from most major haemosporidian lineages. Using this PCR-based approach, we successfully generated a dataset of 21 genes that includes most major haemosporidian lineages. Phylogenetic analyses of a concatenated alignment resulted in a well-resolved phylogeny that was robust to the choice of the outgroup and the method of tree inference.

2. Materials and methods

2.1. Sample collection and parasite screening

Blood samples from birds, reptiles and bats were collected from a wide range of geographical locations (Table 1) and stored on Whatman FTA cards (Sigma–Aldrich, Munich, Germany) or as EDTA-blood in lysis buffer. Total DNA was extracted from each sample using the QIAamp DNA Investigator Kit (Qiagen, Hilden, Germany) according to manufacturer's instructions. All samples were screened for haemosporidian parasites by nested PCR using degenerate oligonucleotide primers that amplify a 317 bp fragment (after removal of primer sequences) of the cytochrome b gene. The outer PCR was performed using primers HaemoScrF1 (5′-AAH TAT GGA GYG GWT GGT G-3′) and HaemoScrR1 (5′-TTA RRY TTC TYT GTT CDG C-3′), 2 μl of genomic DNA were subjected to 30 cycles of 94 °C for 30 s, 42 °C for 30 s, and 66 °C for 45 s. A 0.5 μl aliquot of the product was used as template for a nested reaction with primers HaemoScrF2 (5′-TAA TAC GAC TCA CTA TAG GGA CCW TGG GGW CAA ATG AG-3′) and HaemoScrR2 (5′-ATT TAG GTG ACA CTA TAG AAG CAT TAT CWG GAT GWG MTA-3′) under 40 cycles of 94 °C for 30 s, 44 °C for 30 s, and 66 °C for 45 s. The nested screening primers add T7 and Sp6 adaptors, respectively, at the 5′-ends of the PCR product for direct sequencing. More than 50% of all positive samples had ambiguous base calls after direct sequencing of the PCR products and were discarded as potentially containing multiple infections. Additionally, coding sequences of cytochrome b (*cytb*), cytochrome oxidase I (*coI*), adenylosuccinate lyase (*asl*) and caseinolytic protease (*clpc*) were sequenced as described in Martinsen et al. (2008). The PCR products were ligated into the pGEM-T vector (Promega, Madison, USA) and amplified in *E. coli* (JM109). For each gene, three independent clones were sequenced. Despite the cautionary approach of removing all samples with ambiguous base calls, sequencing of independent clones of the four genes used in Martinsen et al. (2008) revealed the presence of two distinct parasite species in three of the remaining samples, which were discarded as well. A final set of nine samples was selected for amplification of nuclear genes (Table 1). The *cytb* sequence of each sample was used in a Blast search against the MalAvi database and the Genbank nucleotide database (http://www.ncbi.nlm.nih.gov/genbank/) to identify the genus and, if possible, species of the parasite.

Table 1
Host species, lineage and geographic origin of the parasite taxa used in the study.

Parasite species	Lineage	Host species	Sampling location	# Genes
Plasmodium sp.	TURGE01	*Turdus merula*	Germany	19
Plasmodium giganteum	AGANI01	*Agama agama*	Nigeria	16
Polychromophilus sp.	MYOBU01	*Myotis myotis*	Bulgaria	11
Polychromophilus sp.	MYOPA01	*Myotis nigricans*	Panama	5
Parahaemoproteus sp.	NUMNI01	*Numida meleagris*	Nigeria	18
Parahaemoproteus sp.	ACRGE01	*Acrocephalus scirpaceus*	Germany	12
Haemoproteus columbae	COLNI01	*Columba livia*	Nigeria	9
Leucocytozoon sp.	TURGE02	*Turdus merula*	Germany	11
Leucocytozoon sp.	GALNI01	*Gallus gallus*	Nigeria	4

2.2. Primer design, PCR and sequencing

Primer design was based on the dataset by Pick et al. (2011), which comprises 218 single-copy nuclear genes. All single gene, nucleotide alignments were searched for suitable primer binding sites using a custom-made Ruby script (Borner et al., unpublished) using only the *Plasmodium* sequences as template. The script searches for conserved regions in aligned protein-coding nucleotide sequences and evaluates potential primer pairs based on the degree of degeneration, GC content, product size, melting temperatures and hybridization energies of homo- and heterodimers. By searching for matches in all fully sequenced nuclear genomes of *Plasmodium* parasites, the script ensures that primers do not span exon/intron boundaries and only amplify a genomic region of a pre-defined maximum size (1500 bp was set as the limit for the outer pair). Finally, optimal quartets of primer oligonucleotides were selected comprising an inner and an outer pair to allow for nested PCR. This approach was successful for 21 nuclear genes, for which primer oligonucleotides were designed (Supplemental Table S1). The amplified gene fragments range in size from 609 bp to 1178 bp and the global ratio of non-synonymous to synonymous substitutions (dN/dS) is <0.1 for all alignments.

All PCR reactions on nuclear genes were performed with the AccuPrime Taq DNA polymerase according to the following protocol: First, a touchdown PCR was carried out using the outer primer pair on 2 μl of genomic DNA. The denaturing step was performed at 94 °C for 30 s. The initial annealing temperature was set to 47 °C (for 30 s). During the first ten cycles, the annealing temperature was decreased in 0.5 °C decrements before reaching a final annealing temperature of 42 °C for the remaining 25 cycles. Elongation was performed at 62 °C for 90 s. The same conditions were employed for the nested PCR, with the exception of performing ten additional cycles and using a 0.5 μl aliquot of the product from the outer PCR as template. Unsuccessful PCR reactions were repeated with reduced annealing temperatures. PCR fragments of the expected size were isolated by gel extraction, ligated into the pGEM-T vector (Promega, Madison, USA) and amplified in *E. coli*. All fragments were sequenced in both directions. Base calling, removal of vector and primer sequences, and consensus calculation were performed using Vector NTI (Life Technologies, Carlsbad, USA). All sequences were deposited in the European Nucleotide Archive (http://www.ebi.ac.uk/ena; Supplemental Tables S2 and S3).

2.3. Multiple sequence alignment

In addition to the PCR products, sequence data for the 21 gene fragments were obtained from Pick et al. (2011) and from gene predictions of all available apicomplexan genome projects (http://eupathdb.org; Aurrecoechea et al., 2009) resulting in a final taxon sampling of 20 haemosporidian species and 12 outgroup taxa. Orthology assignment was performed by reciprocal BLAST searches (*E*-value ⩽ 10E−10) requiring bidirectional best hits to all taxa included in the dataset of Pick et al. (2011). The sequences were translated and each group of orthologous proteins was aligned individually using MAFFT L-INS-i v7.013 (Katoh and Standley, 2013). Codon alignments were created using a custom Ruby script that uses the protein alignment to guide the corresponding alignment of nucleotide codons. Poorly aligned sections of the amino acid alignments were eliminated by Gblocks v0.91b (Castresana, 2000), allowing for smaller final blocks, gap positions within the final blocks, and less strict flanking positions. The corresponding codons were also removed from the nucleotide alignments (see Supplemental Table S4 for an overview of all datasets used in this study). All data associated with this paper are deposited in the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.72fg9.

To test the effect of missing data, a reduced dataset was created by removing the genes and taxa with the poorest coverage. After removal of four genes and two taxa, all genes had a taxon coverage of ⩾90% (>50% for the newly sequenced taxa) and all taxa had sequence data available for >50% of the genes. To test the effect of outgroup selection on tree topology, three datasets with reduced outgroups were created from the protein dataset. In the first dataset, the outgroup was limited to Coccidia (represented by *Toxoplasma*, *Neospora* and *Eimeria*), for the second dataset, the tree was rooted with Piroplasmida (*Theileria* and *Babesia*) and in the third dataset the outgroup comprised only the most distant apicomplexan genus *Cryptosporidium*. To test the effect of taxon sampling on the internal *Plasmodium* phylogeny, a superalignment was generated based on the 21 gene dataset but with the limited taxon sampling of Pick et al. (2011).

A combined dataset was created by adding the haemosporidian sequences from the studies of Martinsen et al. (2008) (*cytb*, *coI*, *clpc*, *asl*) and Schaer et al. (2013) (*cytb*, *coI*, *clpc*, elongation factor 2A) to the sequences from this study resulting in a dataset that comprises data of 26 genes from 103 haemosporidian parasites. For this dataset, all sequences from both studies were downloaded from the EMBL/Genbank nucleotide database and added to the nucleotide dataset. Additionally, the sequences for *cytb*, *coI*, *asl* and *clpc* from the nine new taxa of the current study were added to the combined dataset. Sequences of the 21 nuclear genes were not available for the added taxa and were coded as missing data.

2.4. Phylogenetic analysis

Phylogenetic trees were calculated by GARLI 2.01 (Zwickl, 2006) for ML analyses, MrBayes 3.2 (Ronquist and Huelsenbeck, 2003) for Bayesian inference and PHYLIP (Felsenstein, 2005) for MP analyses (see Supplemental Table S5 for a detailed overview of all phylogenetic analyses). All Bayesian analyses assumed four Gamma categories of rate heterogeneity and were run for 15,000,000 generations, sampling every 1000th generation and employing two runs with four chains each. Phylogenetic analyses of the nucleotide alignment containing only the first two codon positions were run under the GTR + Γ model of nucleotide substitution for all ML and Bayesian inferences. Based on fitting estimates by ProtTest (Abascal et al., 2005), the WAG amino acid substitution matrix (Whelan and Goldman, 2001) was specified in the MrBayes and GARLI analyses of the protein dataset assuming a Gamma distribution of rate heterogeneity (four categories) and empirical amino acid frequencies. A cross-validation analysis in PhyloBayes 4.1 (Lartillot et al., 2009) showed that for the amino acid dataset the WAG model provides a significantly better fit than the CAT model (cross-validation score of 91.5 ± 32.8 in favor of WAG). Phylogenetic reconstructions based on the codon dataset were performed by MrBayes and GARLI. A codon model (F3X4) was applied in GARLI using different base frequencies for each codon position, a GTR-like model of nucleotide substitution and a uniform rate of non-synonymous to synonymous substitutions (*dN*/*dS*). Bayesian inference was performed under the M3 model on nonsynonymous to synonymous substitution and a GTR model of nucleotide substitution. Additionally, MP analyses were performed on the protein and the nucleotide alignments using the programs Protpars and Dnapars from the PHYLIP package. For ML and MP analyses, bootstrap support was calculated from 100 replicates and mapped on the best tree. Convergence of the independent MrBayes runs was checked using the program AWTY (Nylander et al., 2008). Based on the topological variation between runs, a burnin of 5,000,000 generations was chosen for the calculation of consensus trees.

The phylogenetic analyses of the full 32 taxa dataset were performed employing all tree inference methods described above and a strict consensus cladogram was generated from the resulting trees. Tree inference based on the combined nucleotide dataset was performed by MrBayes. The amino acid datasets with reduced outgroups or minimized missing data were analyzed with GARLI and the dataset with the limited taxon sampling of Pick et al. (2011) was analyzed with MrBayes. Additionally, an outgroup-free phylogenetic reconstruction based on the nucleotide dataset was performed using BEAST v1.80 (Drummond et al., 2012) after removal of all non-haemosporidian taxa. For this analysis, the same parameters were used as described in Outlaw and Ricklefs (2011). Two BEAST runs were performed in parallel under a relaxed molecular clock using the GTR + I model with four discrete Gamma categories and estimated base frequencies. The chains were run employing the Yule tree prior for 10,000,000 generations, sampling every 1000th tree and discarding the first 10% of samples as burnin. All phylogenetic analyses were performed on the CIPRES Science Gateway (Miller et al., 2010).

3. Results

3.1. Parasite identification

Based on Blast searches against the MalAvi database (Bensch et al., 2009) and the Genbank nucleotide database (http://www.ncbi.nlm.nih.gov/genbank/), the parasites were identified as species of the genera *Leucocytozoon*, *Haemoproteus*, *Parahaemoproteus*, *Polychromophilus* and *Plasmodium* (Table 1). Two parasite samples showed >99% sequence identity in the *cytb* gene to previously described species. As the host record for both species matches the source of our samples, we have assigned them to the species *Haemoproteus columbae* and *Plasmodium giganteum*.

3.2. Sequencing of nuclear genes and alignment processing

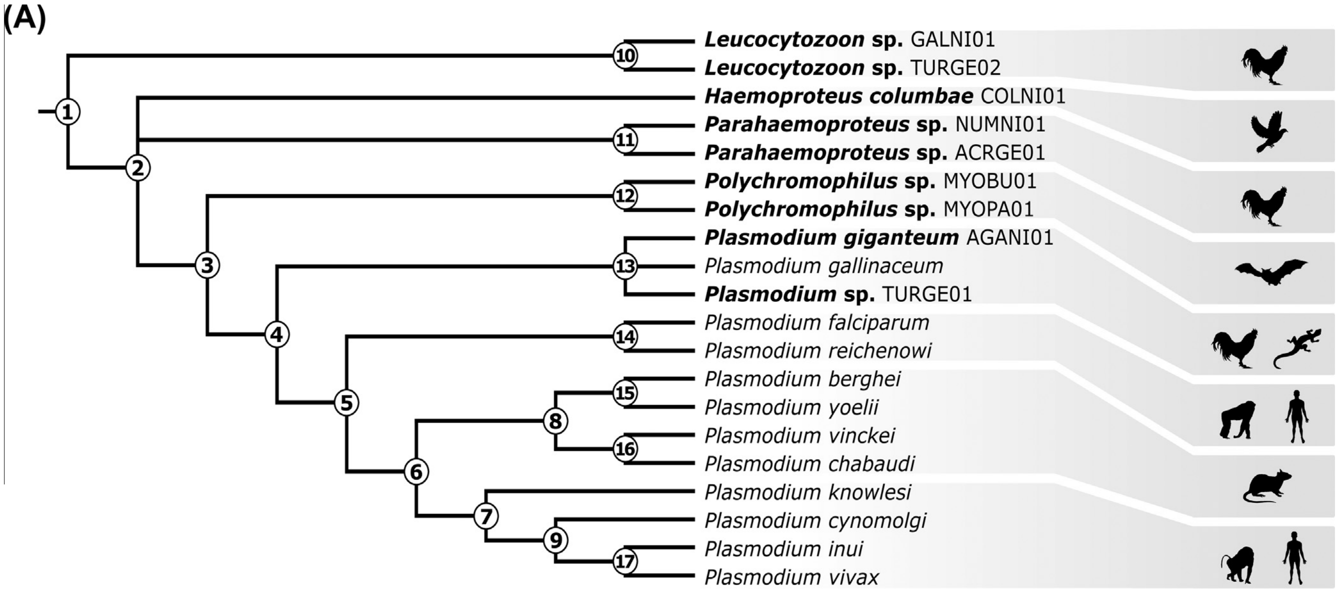
A set of primers for the amplification of 21 orthologous nuclear genes that are conserved among Haemosporida was developed on the basis of the dataset of Pick et al. (2011) (Supplemental Table S1). We obtained between four and 19 gene fragments from each sample (Table 1). Amplification from samples of *Plasmodium* and *Parahaemoproteus* parasites yielded better results (twelve – 19 genes) than amplification from the other samples (four – eleven genes). This 21-gene dataset was complemented with orthologs from Pick et al. (2011) and the available apicomplexan genome projects. The resulting superalignment included 32 species and 6860 amino acid positions (20,580 bp, 34.5% missing data/gaps). After removal of ambiguously aligned positions using Gblocks, the dataset covered 4699 amino acid positions (14.6% missing data/gaps). The corresponding nucleotide alignment contained

14,097 bp, removal of the third codon position resulted in a dataset of 9398 bp. To test the effect of missing data, a reduced amino acid dataset was created comprising data from 17 genes and 30 taxa. The resulting alignment contains 3999 amino acid positions (9.4% missing data/gaps). In addition, a combined dataset was created by adding the haemosporidian sequences from the studies of Martinsen et al. (2008) and Schaer et al. (2013) to the data from this study. This alignment comprised 103 species and 16,898 bp (74.5% missing data/gaps).

3.3. Molecular phylogeny of Haemosporida

Ten phylogenetic analyses were performed on the 21-gene dataset with a variety of methods (Bayesian inference, maximum likelihood [ML] and maximum parsimony [MP]) based on nucleotide, codon and amino acid data (see Supplementary Table S5 for an overview of phylogenetic analyses). All analyses resulted – with the exception of two nodes – in identical topologies (Fig. 1A) with high support values (Figs. 1B and 2, Supplementary Figs. S1–S7). Only the tree derived from the ML analysis of the nucleotide dataset showed slightly reduced support values (Supplementary Fig. S5), though its topology is still highly congruent with the results of the other analyses. In all trees, Haemosporida were recovered as a monophyletic taxon, which is the sister group of Piroplasmida (represented by *Theileria* and *Babesia*). Within Haemosporida, *Leucocytozoon* was consistently recovered as the earliest branching clade. While Bayesian tree inference employing the amino acid dataset yielded high support for a clade uniting *Haemoproteus columbae* and *Parahaemoproteus* (Fig. 2), some of the other analyses found a close relationship of either *Haemoproteus columbae* or *Parahaemoproteus* with the clade comprising *Polychromophilus* and *Plasmodium*. Though, resolution of this split was poor in most trees (Supplemental Figs. S1–S7). There is strong support from all analyses for a sister group relationship of *Polychromophilus* and *Plasmodium* (Fig. 1). *Plasmodium* was recovered as a monophyletic group in all analyses. Within *Plasmodium*, the mammal-infecting lineages form a common clade to the exclusion of the *Plasmodium* parasites of sauropsids. The Laverania, a subgenus containing *P. falciparum* and the ape-infecting *Plasmodium* species (represented by *P. reichenowi*), were recovered as the sister group of all other the mammalian *Plasmodium* species, while the malaria agents of rodents (*P. berghei*, *P. chabaudi*, *P. vinckei* and *P. yoelii*) were found to be most closely related to a clade comprising the other primate *Plasmodium* species (*P. cynomlgi*, *P. knowlesi*, *P. inui* and *P. vivax*). By contrast, phylogenetic analysis of the amino acid dataset under the limited taxon sampling of Pick et al. (2011) found a close relationship between the Laverania and the avian parasite *P. gallinaceum* (Supplemental Fig. S8), as was originally reported in that study.

To evaluate the effect of outgroup selection on the topology of the trees, alternative analyses were performed, each with only one of the three major outgroup taxa (Supplemental Figs. S9–



(B)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Bayes proteins WAG	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.88	1.00	1.00	1.00	1.00
Bayes codons F3X4	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Bayes nucleotides GTR	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ML proteins WAG	100	85	97	99	100	99	100	100	66	100	100	100	55	100	100	98	88
ML codons F3X4	100	93	100	100	100	100	100	100	60	100	100	100	100	100	100	100	100
ML nucleotides GTR	100	67	91	88	91	91	100	100	95	100	100	100	87	100	100	96	99
Parsimony proteins	100	99	100	100	100	95	100	100	67	100	100	100	86	100	100	100	88
Parsimony nucleotides	100	98	100	99	100	100	100	100	99	100	100	100	89	100	100	100	97

Fig. 1. (A) Strict consensus cladogram from all eight phylogenetic analyses of the full dataset. Nodes that were not recovered in all analyses are shown as polytomies. The outgroups are not displayed; phylogenetic relationships among the outgroup taxa are identical to Fig. 2 and received maximum support in all analyses. Taxa sequenced in this study are depicted in bold letters. (B) Bootstrap support values and Bayesian posterior probabilities from the individual analyses (Fig. 2; Supplemental Figs. S1–S7) for the splits depicted above.

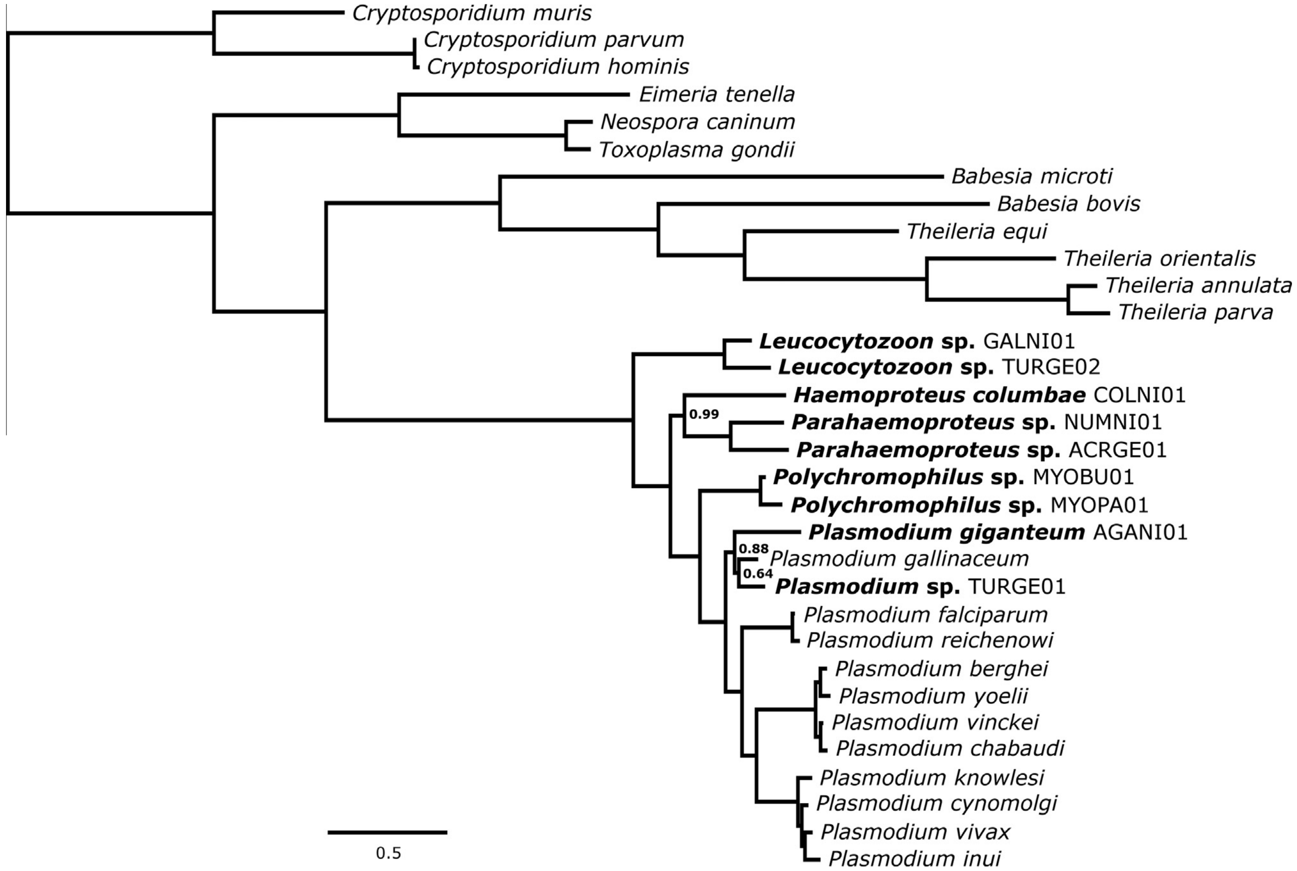


Fig. 2. Majority-rule consensus tree based on a MrBayes analysis of the amino acid dataset (21 genes, 32 taxa). Posterior probabilities < 1.00 are given at the nodes, all other splits have 1.00 posterior probability. The tree was rooted with *Cryptosporidium*.

S11). Within Haemosporida, the topology described above was recovered with all three datasets. To further address the effect of outgroup rooting, an outgroup-free molecular clock based approach to rooting was employed after the removal of all outgroup taxa. This analysis also strongly supported a sister group relationship of monophyletic *Plasmodium* and *Polychromophilus* (Supplemental Fig. S12), and further resulted in the same internal *Plasmodium* phylogeny as described above. However, the phylogenetic relationships at the base of Haemosporida were essentially unresolved in this analysis, because the placement of *Haemoproteus* and *Parahaemoproteus* as two independent lineages on the branch leading up to the clade comprising *Plasmodium* and *Polychromophilus* received only weak support. To minimize the effect of missing data, a reduced dataset was created by removing the genes and taxa with the highest amount of missing data. Phylogenetic analysis of this dataset resulted in the same deep-level relationships as obtained from the full dataset. However, the sauropsid *Plasmodium* species were recovered as a paraphyletic assemblage at the base of *Plasmodium* (Supplemental Fig. S13).

The analysis of a combined dataset (Fig. 3), which included the nuclear sequences from this study as well as those of Martinsen et al. (2008) and Schaer et al. (2013), confirmed the species identification based on Blast searches. The resulting tree (Fig. 3) is fully congruent with the topology obtained from the 21-gene dataset (Fig. 1). It was rooted with *Leucocytozoon* and supports a basal position of *H. columbae* among the remaining haemosporidians (1.0 posterior probability). The genus *Polychromophilus* was recovered as the sister group of a clade comprising all *Plasmodium* species and the other bat-infecting genera, i.e. *Hepatocystis* and *Nycteria* (1.0 posterior probability), which were not included in the 21-gene dataset. *Hepatocystis* and *Nycteria* were found nested within the genus *Plasmodium*. However, their positions were poorly resolved, like most deep-level relationships among major *Plasmodium* lineages, which is in contrast to the trees obtained from the nuclear datasets. Bayesian tree inference with the same taxon sampling but employing a dataset that is limited to the data from the five genes of Martinsen et al. (2008) and Schaer et al. (2013) recovered both *Polychromophilus* and *Nycteria* as closely related to the clade of mammalian parasites (Supplemental Fig. S14, which is consistent with the results of Schaer et al. (2013)).

4. Discussion

Although numerous studies have focused on the phylogeny of Haemosporida, there is still no consensus on the deep-level relationships within this order. By contrast, most of the shallower nodes have received high support (i.e. monophyly of most commonly accepted haemosporidian genera and subgenera) (e.g., Martinsen et al., 2008; Outlaw and Ricklefs, 2011; Pineda-Catalan et al., 2013). Perkins (2014) suggested that this pattern is due to founder effects caused by shifts in vector use as Martinsen et al. (2008) found each major haemosporidian clade to be associated with a unique vector family. While this explanation is plausible, it is also quite obvious that the datasets used so far were not ideal for reconstructing the earliest events in haemosporidian evolution. Due to the challenges involved in developing nuclear markers for this diverse group of parasites (discussed below), all studies have relied on similar sets of no more than four genes mostly of mitochondrial or apicoplast origin. The phylogenetic signal contained in these sequences might

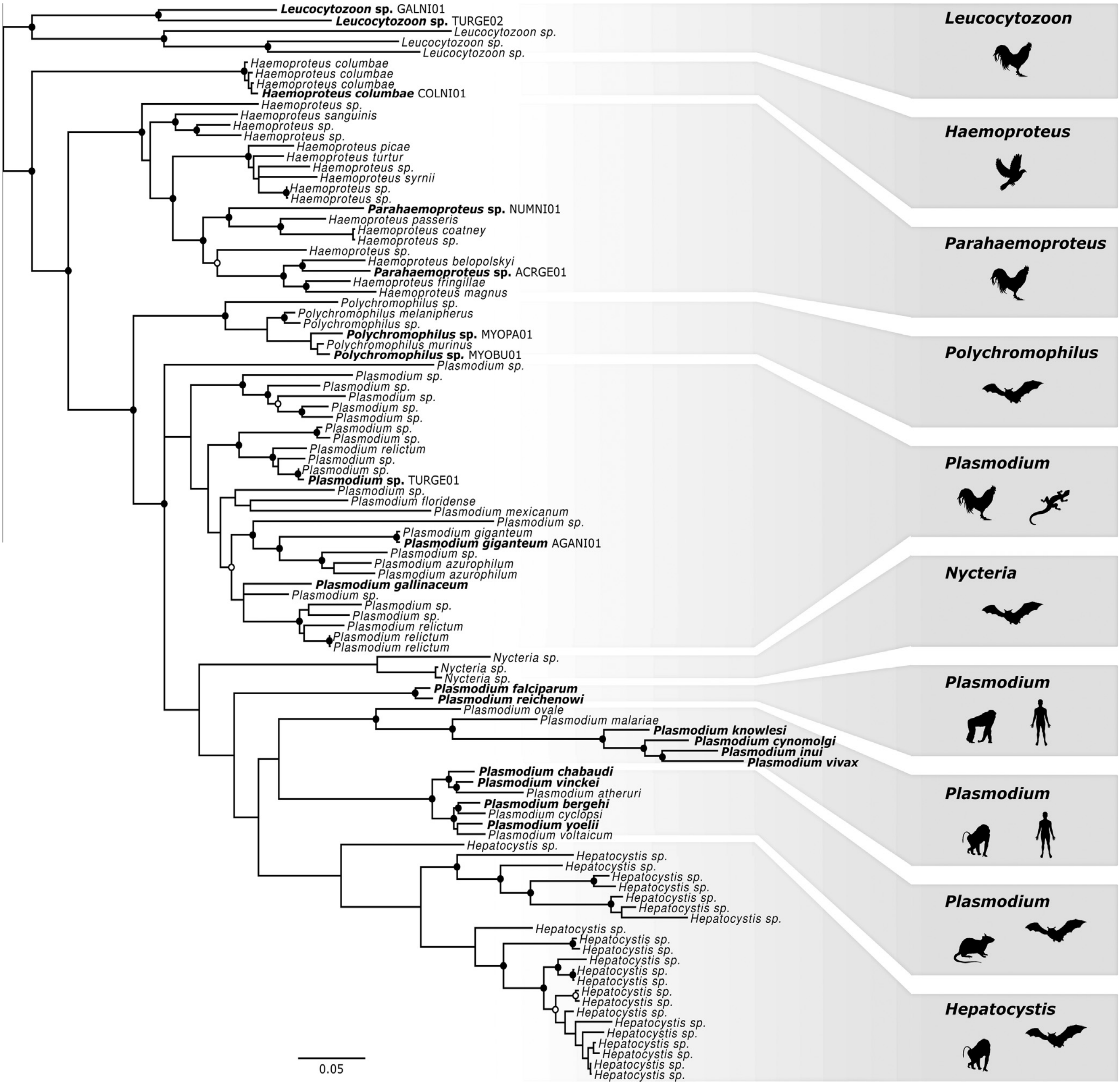


Fig. 3. Majority-rule consensus tree based on a MrBayes analysis of the combined nucleotide dataset comprising the sequence data from Martinsen et al. (2008) and Schaer et al. (2013) and the current study (26 genes, 103 taxa). Splits with a posterior probability ⩾ 0.95 are indicated by a hollow dot, splits with 1.00 posterior probability are indicated by a filled dot. Taxa included in the nuclear dataset are highlighted in bold. The tree was rooted with Leucocytozoon as outgroup.

not be sufficient to resolve the deepest nodes of the tree, especially as some of the gene fragments are rather short. In addition, there is tentative evidence for a loss of mitochondrial DNA sequences or their transfer into the nuclear genome in some parasites (Zehtindjiev et al., 2012; Schaer et al., 2015). Another problem is that the genes regularly used in phylogenetic studies are not well suited for the inclusion of distant outgroups because the sequences are too divergent (Martinsen et al., 2008) and at least for one of the genes (*clpc*), there is no clear ortholog in the other apicomplexans. The inability to include outgroup taxa is especially problematic because the major point of disagreement on haemosporidian phylogeny relates to the position of the root, upon which basically all other deep-level relationships depend.

We have tackled the problem of rooting the evolutionary tree of malaria parasites using a PCR-based approach by sequencing 21 nuclear gene fragments, which were selected on the basis of a dataset of orthologous genes derived from fully sequenced *Plasmodium* genomes (Pick et al., 2011). With this approach, we were able to produce nuclear sequence data from avian parasites of the genera *Haemoproteus* and *Leucocytozoon*, the bat-infecting genus *Polychromophilus*, and two additional sauropsid *Plasmodium* species. To the best of our knowledge, this is the first phylogenetic analysis that employs multiple nuclear gene sequences of these parasite lineages.

4.1. Leucocytozoon is the deepest branching taxon within Haemosporida

Based on a dataset of four genes, Martinsen et al. (2008) obtained the first phylogenetic tree with high support for most major haemosporidian lineages (Fig. 4A), which is generally in line with the traditional view of haemosporidian evolution. Using *Leucocytozoon* as outgroup, *Plasmodium* was found as the most derived clade, while *Haemoproteus* and *Parahaemoproteus* were recovered as two distinct lineages as already proposed by Bennett et al. (1965). Outlaw and Ricklefs (2011), however, argued that there was insufficient evidence for a basal position of *Leucocytozoon* and criticized its *a priori* use as outgroup in most studies on haemosporidian phylogeny. The authors suggested an alternative rooting of the tree, based on an outgroup-free rooting method assuming a molecular clock. This tree inference resulted in a markedly different phylogeny, essentially dividing Haemosporida into two major clades. The first clade comprised all mammalian *Plasmodium* species, while the second clade included all other haemosporidian parasites infecting birds or reptiles (Fig. 4B). Thus, the genus *Plasmodium* was considered polyphyletic. Such a topology would require both taxonomic revision of the genus *Plasmodium* and reevaluation of the evolution of life history traits within Haemosporida. A clade comprising all sauropsid parasites would mean that the ability to reproduce asexually in red blood cells (blood schizogony), which is exclusive to *Plasmodium*, has either evolved independently in mammalian and sauropsid *Plasmodium* species or has been lost multiple times (Outlaw and Ricklefs, 2011). Our results based on the 21-gene dataset reject this notion and strongly support monophyletic *Plasmodium* and a basal position of *Leucocytozoon* within Haemosporida (Fig. 4C). This topology was consistently recovered in all analyses employing amino acid, nucleotide or codon data and was also robust to the choice of outgroup and phylogenetic reconstruction method (ML, MP or Bayesian inference) (Fig. 1). Notably, the application of the same outgroup-free rooting method as used by Outlaw and Ricklefs (2011) to our dataset also recovered a sister group relationship of monophyletic *Plasmodium* and *Polychromophilus* (Supplemental Fig. S12). However, the relationships at the base of Haemosporida were essentially unresolved. In a comparative analysis of gene evolution, Outlaw and Ricklefs (2010) found a significantly elevated ratio of nonsynonymous to synonymous substitutions along the branch linking mammal and bird parasites in *cytb*, suggesting that this gene has undergone episodic adaptive evolution associated with the transition from avian to mammalian hosts, thereby resulting in a long branch between mammalian and sauropsid parasites. However, this phenomenon was not found in the other two genes the authors analyzed (*coI* and *clpc*), and was also not detected in any of the nuclear genes employed in our study. Such strong, putatively non-phylogenetic signal in *cytb* might be the cause for the separation into a mammalian and a sauropsid clade in the subsequent outgroup-free analysis (Outlaw and Ricklefs, 2011).

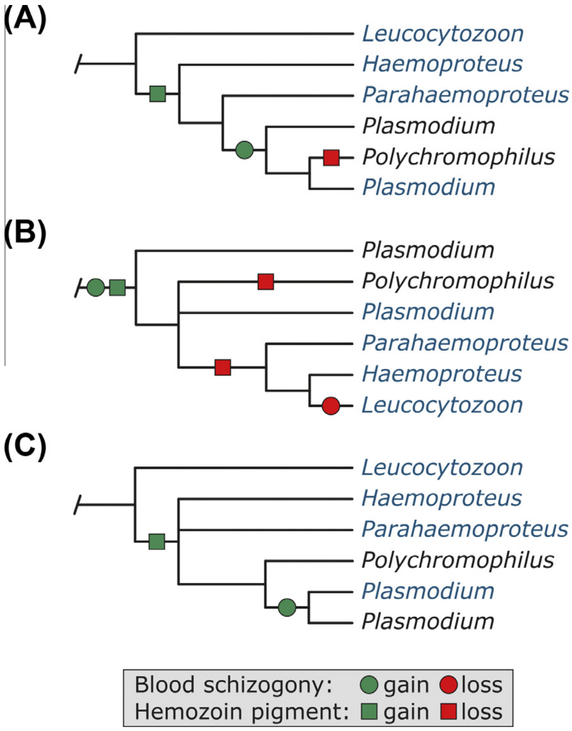


Fig. 4. Comparison of phylogenetic hypotheses on the deep-level relationships of Haemosporida. (A) Topology from Megali et al. (2011) and Witsenburg et al. (2012). (B) Phylogeny from Outlaw and Ricklefs (2011). (C) Phylogenetic relationships based on the current analysis of the 21-gene dataset. In all phylogenetic analyses presented here, Hepatocystis was recovered nested within the mammalian clade of Plasmodium parasites. Parasites of sauropsid hosts are depicted in blue.

4.2. The relationships of Haemoproteus and Parahaemoproteus remain unresolved

The taxonomic status of the genus *Haemoproteus* has been questioned for a long time. According to Bennett et al. (1965) the avian haemoproteid species form two distinct clades. The authors designated the parasites vectored by hippoboscid flies to the genus *Haemoproteus* (including the type species *H. columbae*) and erected a new genus *Parahaemoproteus* for the species, which use mosquitoes as vectors. Levine and Campbell (1971) argued that the differences between both groups were not sufficient to warrant a division into two genera and relegated them to the status of subgenera, a view later echoed by Valkiūnas (2005) in his extensive review of avian blood parasites. Both hypotheses have found support from molecular data. Martinsen et al. (2008) recovered the haemoproteid parasites of birds as paraphyletic and favored the division of *Haemoproteus* into two genera as proposed by Bennett et al. (1965). Other studies with a broader taxonomic sampling of this group recovered both clades as sister subgenera (Iezhova et al., 2011; Pineda-Catalan et al., 2013). Unfortunately, we are not able to resolve this issue as our results mirror the uncertainty of previous analyses, with support for either hypothesis (Figs. 2 and 3; Supplemental Figs. S1–S7).

4.3. Polychromophilus is the sister group of the genus Plasmodium

The genus *Polychromophilus* has proven difficult to classify based on life history data and morphology. As these parasites use endothelial cells for reproduction, Mattingly (1983) already speculated whether they constituted a secondary invasion into mammals. Molecular phylogenetic studies consistently recovered *Polychromophilus* nested within the genus *Plasmodium*. While most analyses based on similar datasets have found *Polychromophilus* to be closely associated with the sauropsid *Plasmodium* species (Fig. 4A) (Duval et al., 2007; Megali et al., 2011; Witsenburg et al., 2012), a recent study by Schaer et al. (2013) resolved *Polychromophilus* more closely related to mammalian *Plasmodium*, thereby rejecting the notion of multiple across-clade switches of the parasite host. In contrast to all previous studies, our analyses of 21 nuclear genes consistently recovered *Plasmodium* as a monophyletic taxon to the exclusion of *Polychromophilus*, which forms the sister group. Notably, this position of *Polychromophilus* received high support in all of our analyses and was also recovered using a combined dataset, which includes the genes and taxa from Martinsen et al. (2008) and Schaer et al. (2013) and the present study. Considering the different life cycle of *Polychromophilus*, which undergoes schizogony in endothelial cells but not in infected red blood cells, its placement outside of the *Plasmodium* clade appears plausible. This relationship further indicates that blood schizogony was not lost during the evolution of *Polychromophilus*, but rather evolved in the lineage leading to *Plasmodium*.

Parasites of the other bat-infecting genera, i.e. *Hepatocystis* and *Nycteria*, also lack the ability for blood schizogony, but like *Plasmodium* species use liver cells for schizogony instead of endothelial cells. While *Hepatocystis* and *Nycteria* are missing from the nuclear datasets, there is support from the analysis of the combined dataset for a placement of both taxa within *Plasmodium* (Fig. 3). Similar to the results of Schaer et al. (2013), *Hepatocystis* was found nested within the clade of mammal-infecting *Plasmodium* species, and *Nycteria* was recovered as the sister group of this clade. However, the support values for these positions are poor, which is most likely due to the large amount of missing data for both taxa.

An additional analysis of the combined dataset without the nuclear data of the current study recovered *Polychromophilus* as the sister group of a clade comprising *Nycteria*, *Hepatocystis* and the mammalian *Plasmodium* parasites (Supplemental Fig. S14) as reported by Schaer et al. (2013). The fact that the addition of the 21 nuclear genes resulted in a phylogeny that is congruent with the results from the nuclear-only datasets shows that the nuclear markers provide a solid backbone for combined phylogenetic analyses, even when nuclear data is only available for a limited subset of taxa.

Even though previous phylogenetic studies based on molecular data have consistently placed *Hepatocystis*, *Polychromophilus* and *Nycteria* within the clade of *Plasmodium* species (Escalante et al., 1998; Perkins and Schall, 2002; Schaer et al., 2013, 2015), the authors were reluctant to call for a revision of the genus *Plasmodium* to include these parasites. Considering the striking differences in the life cycle and morphology, we agree that it is preferable to proceed with caution until more nuclear sequence data become available for these haemosporidian groups.

4.4. Relationships among Plasmodium lineages

Within *Plasmodium*, we found that the sauropsid parasites diverged first, thus uniting all mammalian *Plasmodium* species in a single clade. This topology is in contrast to the study of Pick et al. (2011), which was based on all available fully sequenced *Plasmodium* genomes and which found the Laverania (represented by *P. falciparum* and *P. reichenowi*) to be most closely related to the avian parasite *P. gallinaceum*, rendering the mammalian *Plasmodium* species paraphyletic. In a recent review of the history of haemosporidian systematics, Perkins (2014) argued that this topology might be a result of the incomplete taxon sampling of the study, which was limited to one avian and seven mammalian *Plasmodium* species because no nuclear genomes of other haemosporidian lineages were available. To test whether the difference in topology between Pick et al. (2011) and the present study is due to the improved taxon sampling or the reduced number of genes used in our analysis (21 gene fragments vs. 218 full length genes), a phylogenetic analysis of the 21-gene dataset was conducted under the limited taxon sampling of Pick et al. (2011). This analysis recovered the original phylogeny with a sister group relationship of Laverania and *P. gallinaceum* (Supplemental Fig. S14), indicating that the difference in tree topology is indeed due to the inclusion of additional haemosporidian lineages.

All other human malaria parasites (*P. ovale*, *P. malariae*, *P. knowlesi* and *P. vivax*) were recovered in a common clade with the primate *Plasmodium* species *P. inui* and *P. cynomolgi*. A clade comprising all primate *Plasmodium* species except *P. falciparum* and its closely related ape-infecting relatives (i.e. Laverania) was already proposed based on morphology (Garnham, 1964), and has subsequently been corroborated by phylogenetic analyses of genetic data (e.g., Escalante et al., 1995; Perkins and Schall, 2002).

4.5. The effect of data and model selection

Several studies have shown that the choice of outgroup can have a profound impact on tree topology (e.g. Milinkovitch and Lyons-Weiler, 1998; Holland et al., 2003). Martinsen et al. (2008) did not include non-haemosporidian taxa in their dataset and decided to use *Leucocytozoon* as outgroup instead because they considered the gene sequences of the other apicomplexan parasites to be too divergent. In our analyses, the choice of outgroup did not have an effect on the phylogeny of the ingroup (Supplemental Figs. S9–S11). An outgroup-free analysis also resulted in the same topology (Supplemental Fig. S12). These findings indicate that – at least for our dataset of 21 nuclear genes – the apicomplexan relatives of Haemosporida represent viable outgroup taxa.

As the analysis of the dataset with reduced missing data yielded congruent results, but was also unable to resolve the positions of *Haemoproteus* and *Parahaemoproteus*, we conclude that our inability to resolve this issue is most likely due to insufficient taxon sampling. Davalos and Perkins (2008) evaluated the performance of different models of sequence evolution in recovering phylogenetic relationships among *Plasmodium* parasites employing genomic data. The authors found that the removal of the third codon position improved tree resolution in analyses based on nucleotide models, but also concluded that nucleotide models were generally outperformed by codon and protein models, which were less sensitive to taxon sampling. This conclusion is in line with the results of our ML analyses, in which the tree based on nucleotide data had the lowest overall support, and the inclusion of the third codon position reduced the resolution of the tree even further (data not shown). Davalos and Perkins (2008) advocated including genes from as many species as possible because incomplete taxon sampling can have a profound effect on tree topology. In our analyses, the improved taxon sampling compared to Pick et al. (2011) indeed strongly affected the internal phylogeny of *Plasmodium*, while the dataset was robust to the application of different models of evolution (Fig. 1). However, despite the significant improvements in taxon sampling, the present dataset represents only a fraction of the true diversity of Haemosporida. The addition of more species from the already included genera as well as from the lineages missing in this study, such as the bat parasites *Hepatocystis* and *Nycteria* or the hemoproteid parasites of non-avian sauropsids, will be required to improve our understanding of the evolutionary history of Haemosporida.

4.6. Perspectives of haemosporidian phylogeny

Advancement of our understanding of the haemosporidian phylogeny can only be achieved by improving both taxon and gene sampling of the datasets used for phylogenetic inference. Considering the huge number of haemosporidian lineages and taking into account the challenges involved in sequencing nuclear genomes of sauropsid blood parasites, we expect PCR based approaches to be used in order to solve questions on haemosporidian phylogeny for some time to come. Perkins (2014) called the development of a larger number of molecular markers "the greatest challenge confronting workers who are interested in the systematics of the Haemosporida". This is primarily due to two factors. As full genomes are only available for (mostly mammalian) *Plasmodium* species, the design of primers that are capable of amplifying gene fragments from other haemosporidian lineages relies on a restricted database. Further complicating the matter for PCR-based approaches – but even more so for full genome sequencing projects – is the fact that birds and reptiles have nucleated red blood cells, causing high amounts of contamination by host DNA in the samples. By carefully optimizing primer design parameters and PCR protocols, we were able to obtain a much higher number of genes than used in previous studies. However, due to the challenges described above, some of our samples have a relatively high amount of missing data in the final alignment and our inability to reliably resolve the phylogenetic position of *Haemoproteus columbae* might be attributed to that. Once full genomes of *Haemoproteus* and other haemosporidian genera become available, it will be possible to significantly increase the gene coverage of this dataset and expand the amount of target genes by designing genus-specific primers.

Acknowledgments

This work is supported by a grant of the Deutsche Forschungsgemeinschaft (BR 1744/12-1, BU 956/16-1 and PI 1087/1-1). Field work in Panama was supported by the German National Academic Foundation (VMC) and the German Federal Ministry for Education and Research coordinated under the umbrella of the National Research Platform for Zoonoses (pilot Project Grant P3141077; VMC). We thank Jan Felix Drexler and Christian Drosten (Institute of Virology, University of Bonn) for providing access to their collection of DNA samples from bats. We dedicate this paper to the memory of Elisabeth Kalko and Björn Siemers.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2015.09.003.

References

Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21, 2104–2105.

Aurrecoechea, C., Brestelli, J., Brunk, B.P., Dommer, J., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O.S., Heiges, M., Innamorato, F., Iodice, J., Kissinger, J. C., Kraemer, E., Li, W., Miller, J.A., Nayak, V., Pennington, C., Pinney, D.F., Roos, D. S., Ross, C., Stoeckert Jr., C.J., Treatman, C., Wang, H., 2009. PlasmoDB: a functional genomic database for malaria parasites. Nucleic Acids Res. 37, D539–D543.

Bennett, G.F., Garnham, P.C., Fallis, A.M., 1965. On the status of the genera *Leucocytozoon* Ziemann, 1893 and *Haemoproteus* Kruse, 1890 (Haemosporidiida: Leucocytozoidae and Haemoproteidae). Can. J. Zool. 43, 927–932.

Bensch, S., Hellgren, O., Perez-Tris, J., 2009. MalAvi: a public database of malaria parasites and related haemosporidians in avian hosts based on mitochondrial cytochrome b lineages. Mol. Ecol. Resour. 9, 1353–1358.

Carlton, J.M., Adams, J.H., Silva, J.C., Bidwell, S.L., Lorenzi, H., Caler, E., Crabtree, J., Angiuoli, S.V., Merino, E.F., Amedeo, P., Cheng, Q., Coulson, R.M., Crabb, B.S., Del Portillo, H.A., Essien, K., Feldblyum, T.V., Fernandez-Becerra, C., Gilson, P.R., Gueye, A.H., Guo, X., Kang'a, S., Kooij, T.W., Korsinczky, M., Meyer, E.V., Nene, V., Paulsen, I., White, O., Ralph, S.A., Ren, Q., Sargeant, T.J., Salzberg, S.L., Stoeckert, C.J., Sullivan, S.A., Yamamoto, M.M., Hoffman, S.L., Wortman, J.R., Gardner, M.J., Galinski, M.R., Barnwell, J.W., Fraser-Liggett, C.M., 2008. Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. Nature 455, 757–763.

Carlton, J.M., Angiuoli, S.V., Suh, B.B., Kooij, T.W., Pertea, M., Silva, J.C., Ermolaeva, M. D., Allen, J.E., Selengut, J.D., Koo, H.L., Peterson, J.D., Pop, M., Kosack, D.S., Shumway, M.F., Bidwell, S.L., Shallom, S.J., van Aken, S.E., Riedmuller, S.B., Feldblyum, T.V., Cho, J.K., Quackenbush, J., Sedegah, M., Shoaibi, A., Cummings, L.M., Florens, L., Yates, J.R., Raine, J.D., Sinden, R.E., Harris, M.A., Cunningham, D. A., Preiser, P.R., Bergman, L.W., Vaidya, A.B., van Lin, L.H., Janse, C.J., Waters, A.P., Smith, H.O., White, O.R., Salzberg, S.L., Venter, J.C., Fraser, C.M., Hoffman, S.L., Gardner, M.J., Carucci, D.J., 2002. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. Nature 419, 512–519.

Castresana, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17, 540–552.

Chasar, A., Loiseau, C., Valkiunas, G., Iezhova, T., Smith, T.B., Sehgal, R.N., 2009. Prevalence and diversity patterns of avian blood parasites in degraded African rainforest habitats. Mol. Ecol. 18, 4121–4133.

Cox-Singh, J., Davis, T.M., Lee, K.S., Shamsul, S.S., Matusop, A., Ratnam, S., Rahman, H.A., Conway, D.J., Singh, B., 2008. *Plasmodium knowlesi* malaria in humans is widely distributed and potentially life threatening. Clin. Infect. Dis. 46, 165–171.

Davalos, L.M., Perkins, S.L., 2008. Saturation and base composition bias explain phylogenomic conflict in *Plasmodium*. Genomics 91, 433–442.

Drummond, A.J., Suchard, M.A., Xie, D., Rambaut, A., 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29, 1969–1973.

Duval, L., Robert, V., Csorba, G., Hassanin, A., Randrianarivelojosia, M., Walston, J., Nhim, T., Goodman, S.M., Ariey, F., 2007. Multiple host-switching of Haemosporidia parasites in bats. Malar. J. 6, 157.

Escalante, A.A., Barrio, E., Ayala, F.J., 1995. Evolutionary origin of human and primate malarias: evidence from the circumsporozoite protein gene. Mol. Biol. Evol. 12, 616–626.

Escalante, A.A., Freeland, D.E., Collins, W.E., Lal, A.A., 1998. The evolution of primate malaria parasites based on the gene encoding cytochrome b from the linear mitochondrial genome. Proc. Natl. Acad. Sci. USA 95, 8124–8129.

Felsenstein, J., 2005. PHYLIP (Phylogeny Inference Package) Version 3.695.

Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., Paulsen, I.T., James, K., Eisen, J.A., Rutherford, K., Salzberg, S.L., Craig, A., Kyes, S., Chan, M.S., Nene, V., Shallom, S.J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M.W., Vaidya, A.B., Martin, D.M., Fairlamb, A.H., Fraunholz, M.J., Roos, D.S., Ralph, S.A., McFadden, G.I., Cummings, L.M., Subramanian, G.M., Mungall, C., Venter, J.C., Carucci, D.J., Hoffman, S.L., Newbold, C., Davis, R.W., Fraser, C.M., Barrell, B., 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419, 498–511.

Garnham, P.C., 1964. The subgenera of *Plasmodium* in mammals. Ann. Soc. Belges Med. Trop., Parasitol. Mycol. 44, 267–271.

Garnham, P.C.C., 1966. Malaria Parasites and Other Haemosporidia. Blackwell Scientific, Oxford.

Goldberg, D.E., Slater, A.F., Cerami, A., Henderson, G.B., 1990. Hemoglobin degradation in the malaria parasite *Plasmodium falciparum*: an ordered process in a unique organelle. Proc. Natl. Acad. Sci. USA 87, 2931–2935.

Hagner, S.C., Misof, B., Maier, W.A., Kampen, H., 2007. Bayesian analysis of new and old malaria parasite DNA sequence data demonstrates the need for more phylogenetic signal to clarify the descent of *Plasmodium falciparum*. Parasitol. Res. 101, 493–503.

Hall, N., Karras, M., Raine, J.D., Carlton, J.M., Kooij, T.W., Berriman, M., Florens, L., Janssen, C.S., Pain, A., Christophides, G.K., James, K., Rutherford, K., Harris, B., Harris, D., Churcher, C., Quail, M.A., Ormond, D., Doggett, J., Trueman, H.E., Mendoza, J., Bidwell, S.L., Rajandream, M.A., Carucci, D.J., Yates 3rd, J.R., Kafatos, F.C., Janse, C.J., Barrell, B., Turner, C.M., Waters, A.P., Sinden, R.E., 2005. A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses. Science 307, 82–86.

Holland, B.R., Penny, D., Hendy, M.D., 2003. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock – a simulation study. Syst. Biol. 52, 229–238.

Iezhova, T.A., Dodge, M., Sehgal, R.N., Smith, T.B., Valkiunas, G., 2011. New avian *Haemoproteus* species (Haemosporida: Haemoproteidae) from African birds, with a critique of the use of host taxonomic information in hemoproteid classification. J. Parasitol. 97, 682–694.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780.

Lartillot, N., Lepage, T., Blanquart, S., 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25, 2286–2288.

Lefevre, T., Sanchez, M., Ponton, F., Hughes, D., Thomas, F., 2007. Virulence and resistance in malaria: who drives the outcome of the infection? Trends Parasitol. 23, 299–302.

Levine, N.D., Campbell, G.R., 1971. A check-list of the species of the genus *Haemoproteus* (Apicomplexa, Plasmodiidae). J. Protozool. 18, 475–484.

Manwell, R.D., 1957. Intraspecific variation in parasitic protozoa. Syst. Zool. 6, 2–6.

Martinsen, E.S., Perkins, S.L., Schall, J.J., 2008. A three-genome phylogeny of malaria parasites (*Plasmodium* and closely related genera): evolution of life-history traits and host switches. Mol. Phylogenet. Evol. 47, 261–273.

Mattingly, P.F., 1983. The palaeogeography of mosquito-borne disease. Biol. J. Linn. Soc. 19, 185–210.

Megali, A., Yannic, G., Christe, P., 2011. Disease in the dark: molecular characterization of *Polychromophilus murinus* in temperate zone bats revealed a worldwide distribution of this malaria-like disease. Mol. Ecol. 20, 1039–1048.

Milinkovitch, M.C., Lyons-Weiler, J., 1998. Finding optimal ingroup topologies and convexities when the choice of outgroups is not obvious. Mol. Phylogenet. Evol. 9, 348–357.

Miller, M.A., Pfeiffer, W., Schwartz, T., 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Proceeding of the Gateway Computing Environments Workshop (GCE), New Orleans, LA, pp. 1–8.

Nylander, J.A., Wilgenbusch, J.C., Warren, D.L., Swofford, D.L., 2008. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics 24, 581–583.

Otto, T.D., Rayner, J.C., Böhme, U., Pain, A., Spottiswoode, N., Sanders, M., Quail, M., Ollomo, B., Renaud, F., Thomas, A.W., Prugnolle, F., Conway, D.J., Newbold, C., Berriman, M., 2014. Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. Nat. Commun. 5, 4754.

Outlaw, D.C., Ricklefs, R.E., 2010. Comparative gene evolution in haemosporidian (apicomplexa) parasites of birds and mammals. Mol. Biol. Evol. 27, 537–542.

Outlaw, D.C., Ricklefs, R.E., 2011. Rerooting the evolutionary tree of malaria parasites. Proc. Natl. Acad. Sci. USA 108, 13183–13187.

Pain, A., Bohme, U., Berry, A.E., Mungall, K., Finn, R.D., Jackson, A.P., Mourier, T., Mistry, J., Pasini, E.M., Aslett, M.A., Balasubrammaniam, S., Borgwardt, K., Brooks, K., Carret, C., Carver, T.J., Cherevach, I., Chillingworth, T., Clark, T.G., Galinski, Hall, N., Harper, D., Harris, D., Hauser, H., Ivens, A., Janssen, C.S., Keane, T., Larke, N., Lapp, S., Marti, M., Moule, S., Meyer, I.M., Ormond, D., Peters, N., Sanders, M., Sanders, S., Sargeant, T.J., Simmonds, M., Smith, F., Squares, R., Thurston, S., Tivey, A.R., Walker, D., White, B., Zuiderwijk, E., Churcher, C., Quail, M.A., Cowman, A.F., Turner, C.M., Rajandream, M.A., Kocken, C.H., Thomas, A.W., Newbold, C.I., Barrell, B.G., Berriman, M., . The genome of the simian and human malaria parasite *Plasmodium knowlesi*. Nature 455, 799–803.

Perkins, S.L., 2014. Malaria's many mates: past, present, and future of the systematics of the order Haemosporida. J. Parasitol. 100, 11–25.

Perkins, S.L., Schall, J.J., 2002. A molecular phylogeny of malarial parasites recovered from cytochrome b gene sequences. J. Parasitol. 88, 972–978.

Pick, C., Ebersberger, I., Spielmann, T., Bruchhaus, I., Burmester, T., 2011. Phylogenomic analyses of malaria parasites and evolution of their exported proteins. BMC Evol. Biol. 11, 167.

Pineda-Catalan, O., Perkins, S.L., Peirce, M.A., Engstrand, R., Garcia-Davila, C., Pinedo-Vasquez, M., Aguirre, A.A., 2013. Revision of hemoproteid genera and description and redescription of two species of chelonian hemoproteid parasites. J. Parasitol. 99, 1089–1098.

Ricklefs, R.E., Fallon, S.M., 2002. Diversification and host switching in avian malaria parasites. Proc. R. Soc. Lond. B: Biol. Sci. 269, 885–892.

Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.

Schaer, J., Perkins, S.L., Decher, J., Leendertz, F.H., Fahr, J., Weber, N., Matuschewski, K., 2013. High diversity of West African bat malaria parasites and a tight link with rodent *Plasmodium* taxa. Proc. Natl. Acad. Sci. USA 110, 17415–17419.

Schaer, J., Reeder, D.M., Vodzak, M.E., Olival, K.J., Weber, N., Mayer, F., Matuschewski, K., Perkins, S.L., 2015. *Nycteria* parasites of Afrotropical insectivorous bats. Int. J. Parasitol. 45, 375–384.

Tachibana, S., Sullivan, S.A., Kawai, S., Nakamura, S., Kim, H.R., Goto, N., Arisue, N., Palacpac, N.M., Honma, H., Yagi, M., Tougan, T., Katakai, Y., Kaneko, O., Mita, T., Kita, K., Yasutomi, Y., Sutton, P.L., Shakhbatyan, R., Horii, T., Yasunaga, T., Barnwell, J.W., Escalante, A.A., Carlton, J.M., Tanabe, K., 2012. *Plasmodium cynomolgi* genome sequences provide insight into *Plasmodium vivax* and the monkey malaria clade. Nat. Genet. 44, 1051–1055.

Valkiūnas, G., Iezhova, T.A., Loiseau, C., Chasar, A., Smith, T.B., Sehgal, R.N., 2008. New species of haemosporidian parasites (Haemosporida) from African rainforest birds, with remarks on their classification. Parasitol. Res. 103, 1213–1228.

Valkiūnas, G., 2005. Avian Malaria Parasites and Other Haemosporidia. CRC Press, Boca Raton, Florida.

Whelan, S., Goldman, N., 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18, 691–699.

WHO, 2013. World Malaria Report 2013. World Health Organization, Geneva (Switzerland).

Witsenburg, F., Salamin, N., Christe, P., 2012. The evolutionary host switches of *Polychromophilus*: a multi-gene phylogeny of the bat malaria genus suggests a second invasion of mammals by a haemosporidian parasite. Malar. J. 11, 53.

Zehtindjiev, P., Krizanauskiene, A., Bensch, S., Palinauskas, V., Asghar, M., Dimitrov, D., Scebba, S., Valkiunas, G.G., 2012. A new morphologically distinct avian malaria parasite that fails detection by established polymerase chain reaction based protocols for amplification of the cytochrome B gene. J. Parasitol. 98, 657–665.

Zwickl, D.J., 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. University of Texas, Austin.

Molecular Phylogenetics and Evolution 80 (2014) 79–87

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

A transcriptome approach to ecdysozoan phylogeny

Janus Borner[a], Peter Rehm[a], Ralph O. Schill[b], Ingo Ebersberger[c], Thorsten Burmester[a,*]

[a] *Institute of Zoology and Zoological Museum, University of Hamburg, D-20146 Hamburg, Germany*
[b] *Zoology, Biological Institute, University of Stuttgart, Germany*
[c] *Department for Applied Bioinformatics, University of Frankfurt, Institute for Cell Biology and Neuroscience, Germany*

ARTICLE INFO

Article history:
Received 18 March 2014
Revised 15 July 2014
Accepted 1 August 2014
Available online 12 August 2014

Keywords:
Bayesian analyses
Ecdysozoa
Indel
Next generation sequencing
Phylogenomics

ABSTRACT

The monophyly of Ecdysozoa, which comprise molting phyla, has received strong support from several lines of evidence. However, the internal relationships of Ecdysozoa are still contended. We generated expressed sequence tags from a priapulid (penis worm), a kinorhynch (mud dragon), a tardigrade (water bear) and five chelicerate taxa by 454 transcriptome sequencing. A multigene alignment was assembled from 63 taxa, which comprised after matrix optimization 24,249 amino acid positions with high data density (2.6% gaps, 19.1% missing data). Phylogenetic analyses employing various models support the monophyly of Ecdysozoa. A clade combining Priapulida and Kinorhyncha (i.e. Scalidophora) was recovered as the earliest branch among Ecdysozoa. We conclude that Cycloneuralia, a taxon erected to combine Priapulida, Kinorhyncha and Nematoda (and others), are paraphyletic. Rather Arthropoda (including Onychophora) are allied with Nematoda and Tardigrada. Within Arthropoda, we found strong support for most clades, including monophyletic Mandibulata and Pancrustacea. The phylogeny within the Euchelicerata remained largely unresolved. There is conflicting evidence on the position of tardigrades: While Bayesian and maximum likelihood analyses of only slowly evolving genes recovered Tardigrada as a sister group to Arthropoda, analyses of the full data set, and of subsets containing genes evolving at fast and intermediate rates identified a clade of Tardigrada and Nematoda. Notably, the latter topology is also supported by the analyses of indel patterns.

1. Introduction

The phylogenetic relationships of animal phyla are one of the most hotly debated topics of zoology. Resolving early evolutionary events also has fundamental impact on the understanding of animal biology. Based on phylogenetic analyses of rRNA sequences, Aguinaldo and colleagues (Aguinaldo et al., 1997) defined the superphylum "Ecdysozoa", which comprises the molting phyla Arthropoda, Onychophora (velvet worms), Tardigrada (water bears), Nematoda (roundworms), Nematomorpha (horsehair worms), Priapulida (penis worms), Kinorhyncha (mud dragons) and Loricifera. Ecdysozoa include the most species-rich animal phylum (Arthropoda) and thus outnumber the other protostome superphylum (Lophotrochozoa) and the Deuterostomia (Telford et al., 2008). In addition to the process of molting of the three-layered cuticle, which is controlled by ecdysteroid hormones, Ecdysozoa share only few other morphological characters ("synapomorphies"), including the lack of ciliated epithelia and the absence of spiral cleavage (Giribet and Ribera, 1998; Schmidt-Rhaesa et al., 1998; Telford et al., 2008, 2009). The "Ecdysozoa" hypothesis is at odds with the more "traditional" animal systematics, which holds the view of a close relationship of panarthropods (Arthropoda plus Onychophora and Tardigrada) and annelids (which are now regarded as members of the superphylum "Lophotrochozoa"), and a common origin of animals with a coelomate body cavity (Westheide and Rieger, 1996; Brusca and Brusca, 2003).

The monophyly of Ecdysozoa has received support from molecular phylogenetic studies using selected genes (Mallatt et al., 2004; Webster et al., 2006; Bourlat et al., 2008; Dunn et al., 2008; Telford et al., 2008; Hejnol et al., 2009). Still, several approaches that applied large datasets deriving from whole genomes suggested that *Drosophila melanogaster* (Arthropoda) is closer related to humans than to *Caenorhabditis elegans* (Nematoda), thereby supporting the Coelomata concept (Blair et al., 2002; Wolf et al., 2004; Philip et al., 2005; Ciccarelli et al., 2006; Rogozin et al., 2007). However, others have argued that this topology was the result of long branch

Abbreviations: EST, expressed sequence tag; ML, maximum likelihood; mya, million years ago; SIC, simple indel coding.

[*] Corresponding author. Address: Institute of Zoology and Zoological Museum, Biocenter Grindel, University of Hamburg, Martin-Luther-King-Platz 3, D-20146 Hamburg, Germany. Fax: +49 40 42838 3937.

E-mail address: thorsten.burmester@uni-hamburg.de (T. Burmester).

http://dx.doi.org/10.1016/j.ympev.2014.08.001
1055-7903/© 2014 Elsevier Inc. All rights reserved.

attraction (LBA), which positions e.g. the nematode *C. elegans* close to the root (Copley et al., 2004; Irimia et al., 2007). In fact, inclusion of additional taxa, a procedure that tends to reduce the effect of LBA on phylogenetic tree reconstruction, consistently recovered Ecdysozoa (Philippe et al., 2005; Webster et al., 2006; Roeding et al., 2007; Dunn et al., 2008; Lartillot and Philippe, 2008; Meusemann et al., 2010; Campbell et al., 2011).

While the Ecdysozoa concept has become widely accepted, the relationships within the Ecdysozoa are not well resolved (for review, see: Telford et al., 2008, 2009; Schmidt-Rhaesa, 2013). There is general agreement that Arthropoda and Onychophora are closely related phyla (Westheide and Rieger, 1996; Brusca and Brusca, 2003) and that Nematomorpha are associated with Nematoda (Nielsen, 1995; Dunn et al., 2008; Telford et al., 2008; Schmidt-Rhaesa, 2013). Otherwise, ecdysozoan relationships are disputed. For example, tardigrades have been traditionally considered to be allied with Arthropoda (Westheide and Rieger, 1996; Brusca and Brusca, 2003), a topology that is tentatively supported by a shared microRNA (Campbell et al., 2011), shared structures of the nervous system (Mayer et al., 2013) and *engrailed* expression patterns (Gabriel and Goldstein, 2007). Molecular studies using large-scale sequence alignments suggested that tardigrades may be more closely related to Nematoda (Giribet, 2003; Roeding et al., 2007; Lartillot and Philippe, 2008; Meusemann et al., 2010), although this topology may also be attributed to long-branch attraction (Rota-Stabelli et al., 2011). The worm-like ecdysozoan phyla (i.e., Nematoda, Nematomorpha, Priapulida, Kinorhyncha and Loricifera) have been referred to as "Cycloneuralia" (Schmidt-Rhaesa, 2013). This classification is at odds with studies that e.g. found the priapulids as sister taxon of all other Ecdysozoa (Webster et al., 2006; Lartillot and Philippe, 2008).

The poor resolution of ecdysozoan relationships is most likely due to the lack of data from important taxa. Because of their enormous biological, ecological and biomedical importance, a huge amount of sequences has been generated from Arthropoda and Nematoda, whereas the other ecdysozoan phyla are considerably undersampled. While the sequencing of specifically selected genes for molecular phylogenetic purposes is a tedious procedure that usually leads to short multiple sequence alignments, more recent molecular phylogenetic studies mostly rely on expressed sequence tags (ESTs). We approach to resolve the relationships among Ecdysozoa by obtaining transcriptomes of key taxa employing next generation sequencing. In addition to the phylogenetic approach based on multigene alignments, we traced the evolution of Ecdysozoa by analyzing indel patterns.

2. Materials and methods

2.1. Species collection and RNA isolation

New transcriptome data from eight ecdysozoan species were generated (see also Supplemental Table S1). Specimens of five chelicerates were used in this study: *Gluvia dorsalis* (Solifugae), *Mastigoproctus giganteus* (Uropygi), *Euphrynichus bacillifer* (Amblypygi), *Phalangium opilio* (Opiliones), *Chelifer cancroides* (Pseudoscorpiones). Additionally, transcriptomes of the tardigrade *Echiniscus testudo* (Echiniscoidea), the priapulid *Halicryptus spinulosus* (Halicryptomorphida) and the kinorhynch *Pycnophyes kielensis* (Homalorhagida) were sequenced. Total RNA of each species was extracted according to Holmes and Bonner (Holmes and Bonner, 1973).

2.2. Transcriptome sequencing

cDNA libraries were constructed using a modified template-switching (SMART) procedure (Mint-Universal cDNA synthesis kit, Evrogen, Russia) and sequenced with the 454 GS FLX Titanium chemistry (Roche). Each cDNA library was sequenced in a half PicoTiterPlate (Roche) according to the manufacturer's protocol. Transcriptome sequencing of *G. dorsalis*, *M. giganteus*, *E bacillifer*, *P. opilio, C. cancroides* and *H. spinulosus* was carried out at the Max Planck Institute for Molecular Genetics, Berlin, Germany. The transcriptomes of *E. testudo* and *P. kielensis* were sequenced by LGC Genomics GmbH (Berlin, Germany). Vector-clipping, trimming and quality checking of raw sequence reads and assembly into contigs were performed at the Center for Integrative Bioinformatics (CIBIV), Vienna, Austria. The transcriptomes were checked for possible contaminations with various BLAST-based approaches by cross-comparisons and searches with known protein sequences. Raw data have been deposited in the NCBI Sequence Read Archive (SRA) and assembled contig sequences are available from the Transcriptome Sequences Database (TSA) (BioProject IDs PRJNA236247, PRJNA236248, PRJNA236250, PRJNA236252, PRJNA236253, PRJNA236410, PRJNA236410, PRJNA258412).

2.3. Taxon sampling and orthology assignment

In addition to the assemblies of the six ecdysozoan species, gene predictions of all ecdysozoan genome projects were added to the dataset and transcriptome data of all ecdysozoan species which contained more than 1000 contigs were obtained from the Deep Metazoan Phylogeny (DMP) database (http://www.deep-phylogeny.org/). If more than two species from the same order fulfilled these criteria, only the top two species were selected. The resulting dataset comprises 63 species: 50 ecdysozoans, nine other protostomes and four deuterostomes.

Orthology assignment was performed employing the HaMStR pipeline (Ebersberger et al., 2009); http://sourceforge.net/projects/hamstr). A reference set of 1253 orthologous sequence clusters was used in the analysis, which is based on the proteomes of seven primer taxa: *Apis melifera*, *Caenorhabditis elegans*, *Capitella capitata*, *Daphnia pulex*, *Helobdella robusta*, *Lottia gigantea* and *Schistosoma mansoni* (http://www.deep-phylogeny.org). HaMStR was run with the -strict option using sequentially all seven primer taxa as reference species for the reverse BLAST search. A candidate sequence was only then accepted as an ortholog if it obtained the corresponding reference protein as best hit in all seven BLAST searches.

2.4. Multiple sequence alignments and generation of datasets

Each group of orthologous proteins was aligned individually using MAFFT L-INS-i v7.013 (Katoh and Standley, 2013). Trailing and leading gaps were coded as missing data in each gene alignment. Poorly aligned sections were eliminated by Gblocks v0.91b (settings: −b2 = 41 [65% of the number of sequences] −b3 = 10 −b4 = 5 −b5 = a; Talavera and Castresana, 2007). Alignments of orthologous proteins with less than 50% taxon coverage were removed from the dataset. Finally, the individual alignments were concatenated into a single supermatrix.

In an additional approach, the dataset was divided into three partitions based on the average substitution rate of each gene. To avoid skewed results due to missing data, only the taxa for which sequence data from all genes was available were used for the assessment of substitution rates (*A. melifera*, *C. elegans*, *C. capitata*, *D. pulex*, *H. robusta*, *L. gigantea* and *S. mansoni*). The substitution rates were calculated as the average sum of pairwise scores of all positions in the alignment according to a PAM150 matrix. Positions with gaps were ignored. The individual alignments were concatenated into three subsets (slow, intermediate and fast), processed with Gblocks and used for tree reconstruction as described.

We further created three datasets by successively removing the alignment positions with the highest evolutionary rates. For this approach, the cumulative Parsimony score (P-score) was calculated for each site using PAUP (Swofford, 2003). To exclude biased results due to missing data, again only the taxa for which sequence data from all genes was available were used for the calculation of P-scores and positions that contain missing data in these taxa were excluded from the analysis.

To assess the effect of long-branch taxa on the tree topology, the LB (long branch) scores for all taxa were calculated based on the full multiple sequence alignment using TreSpEx (Struck, 2014). An additional dataset was created after the deletion of the taxa with the highest LB scores (*Schistosoma mansoni*, *Caenorhabditis elegans*, *Meloidogyne hapla*, *Echiniscus testudo*, *Trichinella spiralis* and *Haemonchus contortus*).

2.5. Character coding of indels

The unreduced dataset was used to create a presence/absence matrix based on simple indel coding (SIC; Simmons and Ochoterena, 2000). In SIC, all gaps that have different 5′ and/or 3′ termini are coded as separate presence/absence characters. If a gap from one sequence completely overlaps a gap in another sequence (i.e., extending to or beyond both the 5‘ and 3‘ termini of the gap), it is coded as missing data for the smaller gap. To remove artificial indels in unreliably aligned regions, the degree of conservation in the positions flanking each indel was analyzed. Taxa that were coded as missing data for a particular indel were ignored in the scoring of flanking positions. The sum of pairs scores for the ten amino acid positions leading up to and trailing each indel were calculated according to a BLOSUM62 matrix (gaps were scored as −4). Indels with a negative score in at least one of the flanking regions were removed from the dataset. The implementation of SIC and the analysis of flanking regions were performed using a custom Ruby script (Supplemental Information S1).

2.6. Phylogenetic analyses

Phylogenetic trees based on the amino acid dataset were calculated using RAxML 7.2.8 (Stamatakis, 2006) for Maximum Likelihood (ML) analyses and PhyloBayes3.3f (Lartillot et al., 2009) for Bayesian inference. All phylogenetic analyses were performed on the HPC Linux cluster of the Regionales Rechenzentrum (RRZ), University of Hamburg. The trees were rooted with four deuterostome taxa as outgroup (*Ciona intestinalis*, *Branchiostoma florida*e, *Gallus gallus*, *Homo sapiens*). According to fitting estimates derived from ProtTest (Abascal et al., 2005), ML trees were inferred using the LG amino acid substitution matrix (Le and Gascuel, 2008) modeling rates across sites with a Γ distribution. Bootstrap support was calculated from 1000 replicates applying the rapid hill-climbing algorithm.

Bayesian tree inference was performed assuming the CAT and CAT–GTR mixture models (Lartillot and Philippe, 2004). For both models, 16 chains were run again modeling substitution rate heterogeneity across sites with four discrete Γ categories. Sampling was performed every 10th cycle and the chains were stopped after 20,000 cycles for the CAT model and 10,000 cycles for the CAT–GTR model due to the high computational complexity of these calculations. The first 50% of samples were discarded as burn-in and the discrepancy among chains was determined by comparison of bipartitions from all possible multiple chain combinations with the *bpcomp* tool. The harmonic mean of the likelihood values was calculated for each chain. A majority rule consensus tree was inferred from the best combination of at least 3 chains (*maxdiff* value $\leqslant$0.2).

The indel based dataset was analyzed employing maximum parsimony and Bayesian inference. Parsimony analysis was based on the Wagner parsimony criterion using the program TNT (available at http://www.zmuc.dk/public/phylogeny/tnt; Goloboff et al., 2008). A heuristic search on 10,000 replicates was performed with the tree bisection and reconnection (TBR) branch-swapping algorithm. Bayesian inference was performed following the same protocol as described for the amino acid dataset. For both CAT and CAT–GTR, 20,000 cycles were computed by Phylobayes3.3f.

3. Results

3.1. Transcriptome sequencing and assembly

To increase the sequence data coverage of ecdysozoan taxa, we generated novel transcriptomes from five chelicerate species (*G. dorsalis*, *M. giganteus*, *E. bacillifer*, *P. opilio, C. cancroides*), a priapulid (*H. spinulosus*), a kinorhynch (*P. kielensis*) and a tardigrade (*E. testudo*) (Table 1). 454 sequencing runs produced 422,797–481,905 reads (between 119 and 198 Mb per species). Sequence assembly produced 20,769–59,029 contigs, which were used in subsequent analyses.

Additional sequences from selected metazoan genomes and transcriptomes were obtained from the public databases, resulting in a dataset of 63 species, which included 50 ecdysozoan species, nine other protostomes and four deuterostomes (Supplemental Table S2). This initial dataset encompassed 1253 genes, resulting in a super alignment of 951,716 amino acid positions, which had 6.2% gaps and 71.3% missing data. To minimize the possible effect of missing data, we first removed genes with less than 50% taxon coverage and masked the alignment with Gblocks. These procedures reduced the amino acid dataset to 189 genes with 24,249 positions, which markedly increased data density (only 2.6% gaps, 19.1% missing data).

3.2. Molecular phylogenomic analyses of Ecdysozoa

Phylogenetic analyses for the amino acid dataset were conducted using maximum likelihood (ML) (Supplemental Fig. S1) and Bayesian inference. Two independent Bayesian analyses were run under the CAT + Γ (Supplemental Fig. S2) and the CAT–GTR + Γ (Fig. 1) mixture models. In all analyses, we received strong support for the monophyly of Ecdysozoa, which were consistently recovered as sister group of monophyletic Lophotrochozoa (including the trematode *Schistosoma mansoni*). The clade comprising Priapulida and Kinorhyncha branched off first within Ecdysozoa. Nematoda and Tardigrada formed a common clade and were sister group of Arthropoda. Within the arthropods, we found support for monophyletic Mandibulata (i.e., Myriapoda, Crustacea and Hexapoda) and Pancrustacea (i.e., Hexapoda nested within paraphyletic Crustacea). Branchiopoda (*Daphnia pulex* and *Artemia franciscana*) were the crustacean taxon closest to the monophyletic Hexapoda.

The topology within the subphylum Chelicerata was only poorly resolved, although the monophyly of Euchelicerata, Araneae (true spiders) and Tetrapulmonata (Araneae + (Amblypygi [whip spiders] + Uropygi [whip scorpions])) was recovered with high support in all trees. Otherwise, the relationships remained ambiguous. Notably, in none of our analyses we recovered monophyletic Arachnida. Within the myriapods, Symphyla (garden centipedes) were the sister taxon of Chilopoda (centipedes) and Diplopoda (millipedes).

3.3. Effect of the evolution rate on tree topologies

We assessed the effect of evolutionary rates on tree topology by subdividing the full dataset into subsets based on the average

Table 1
Sequencing and assembly of ESTs.

	Gluvia dorsalis	*Mastigoproctus giganteus*	*Euphrynichus bacillifer*	*Phalangium opilio*	*Chelifer cancroides*	*Echiniscus testudo*	*Halicryptus spinulosus*	*Pycnophyes kielensis*
Reads	426,219	481,905	433,348	474,081	443,697	439,708	422,797	454,711
Mean read length	353	293	275	319	321	421	355	435
Unigene contigs	41,872	44,605	31,233	40,817	28,707	20,769	26,335	59,029
HaMStr orthologs	466	447	364	556	313	606	405	349

substitution rate of each gene. Each subset comprised 63 genes. Dataset 1, which included the slowly evolving proteins, covered 8017 amino acid positions, dataset 2 comprised the proteins that evolved at an intermediate rate (8530 positions), and dataset 3 with the fast evolving genes (7702 positions). In Bayesian phylogenetic analyses, all three datasets recovered monophyletic Ecdysozoa, monophyletic Arthropoda and support a sister group relationship of Priapulida and Kinorhyncha (Fig. 2). The ML trees agree with this (Supplemental Fig. S3), with the exception that in the dataset of proteins evolving at an intermediate speed, *S. mansoni* (Platyhelminthes) was found associated with the nematodes. Conflicting results were obtained for the positions of the Tardigrada and the monophyly of Mandibulata. The tree resulting from the slowly evolving proteins (Fig. 2B) had several poorly supported nodes. It showed the Tardigrada as sister group of Arthropoda + Onychophora, albeit with only 0.80 posterior probability. In all other trees, a sister group relationship of Tardigrada and Nematoda was recovered with maximum support. The position of the Myriapoda was affected as: While the fast evolving genes recovered Myriochelata (Myriapoda + Chelicerata; 0.91 posterior probability), the other analyses found Mandibulata (Myriapoda + Pancrustacea) with maximum support.

In an additional approach to assess the effect of evolutionary rates, we successively removed the alignment positions with the highest substitution rates. After removal of all sites with a P-score > 4, the dataset comprised 21,923 amino-acid positions, further removal of sites with a P-score >3 and >2 resulted in datasets of 20,282 and 17,707 positions, respectively. ML analyses of all three reduced datasets generally resulted in the same topology that was recovered in the analysis of the full datasets (Supplemental Fig. S4). However, the tree that based on the most conserved positions favored Myriochelata over Mandibulata and found Platyhelminthes associated with Nematoda (Supplemental Fig. S4C). On the other hand, removal of the sequence positions 100% conserved among the reference taxa, resulted in a tree identical with that based on all positions (Supplemental Fig. S4D).

Deletion of the six taxa with the highest LB score, which is indicative for long branches, did not have an effect on the overall tree topology (Supplemental Fig. S5). However, we observed significantly increased bootstrap support values for several deep-level splits, i.e. high support for a basal position of Scalidophora (Priapulida + Kinorhyncha) within Ecdysozoa, maximum support for monophyletic Lophotrochozoa and for a sister group relationship between Nematoda and Tardigrada. Furthermore, specific removal of the comparably fast evolving tardigrade *E. testudo* did not change tree topology (Supplemental Fig. S6).

3.4. Application of character-coding for phylogenetic tree reconstructions

Additional analyses employing a different type of data were performed using a presence/absence matrix based on simple indel coding (SIC), which was generated from the unreduced super alignment. The dataset comprised 29,829 indel characters (69.4% missing data). The results of the maximum parsimony analysis of the characters were discarded because ten trees were found to be equally parsimonious and a strict consensus of these trees produced a phylogeny that was unresolved in nearly all deep-level splits. Bayesian analyses were performed assuming the CAT–GTR + Γ model (Supplemental Fig. S7). Because an initial analysis showed that the priapulids destabilized the tree inference, *Priapulus caudatus* and *H. spinulosus* were identified as “rogue taxa” and were therefore removed from the indel dataset. The Bayesian support values of this analysis were projected onto a tree derived from Bayesian inference of the amino acid dataset after removal of the priapulids (Fig. 3). While the phylogenetic relationships within the four euarthropod subphyla remain poorly resolved, most deep-level phylogenetic splits agree with the results of the amino-acid dataset. With the exception of Kinorhyncha, which associated weakly with Lophotrochozoa, the backbone of the Bayesian tree derived from the indel dataset resembles that calculated from the amino acid sequences. We found strong support for monophyletic Arthropoda, Euarthropoda, Chelicerata, Myriapoda, Pancrustacea and Insecta (Ectognatha). Hexapods were not recovered as a monophyletic clade because the collembolan *Folsomia candida* associated with the branchiopod crustaceans. Notably, the clade comprising of Tardigrada and Nematoda received maximum support.

4. Discussion

The Ecdysozoa concept (Aguinaldo et al., 1997) has revolutionized our understanding of animal evolution. For example, the rejection of the monophyly of coelomate animals accompanied by the alternative hypothesis of a close relationship of the model organisms *D. melanogaster* and *C. elegans* has weakened many biological hypotheses, e.g. on animal development, that rely on an outgroup position of Nematoda. Although there is still no universal acceptance of the Ecdysozoa concept (see introduction), most recent morphological and molecular phylogenetic studies support the monophyly of this taxon.

4.1. Data matter

Our phylogenomic analyses provide strong support for the monophyly of Ecdysozoa and also help to resolve the evolution within this taxon. We paid particular attention to the reduction of the amount of missing data, which might skew the results. Although our dataset included only 19.1% missing data (plus 2.6% gaps) it covers 24,249 well-defined amino acid positions. To the best of our knowledge, previous phylogenomic studies either included fewer positions and/or had more missing data (Dunn et al., 2008; Meusemann et al., 2010; Regier et al., 2010; Campbell et al., 2011; Rota-Stabelli et al., 2013).

There is some inconsistency in the interpretation of the relationships within Ecdysozoa. The two most prominent conflicts concern the phylogenetic positions of Tardigrada (see Section 4.3. for a detailed discussion) and Myriapoda (Hwang et al., 2001; Mallatt et al., 2004; Pisani et al., 2004; Dunn et al., 2008; Meusemann et al., 2010; Regier et al., 2010; Brewer and Bond, 2013; Rehm

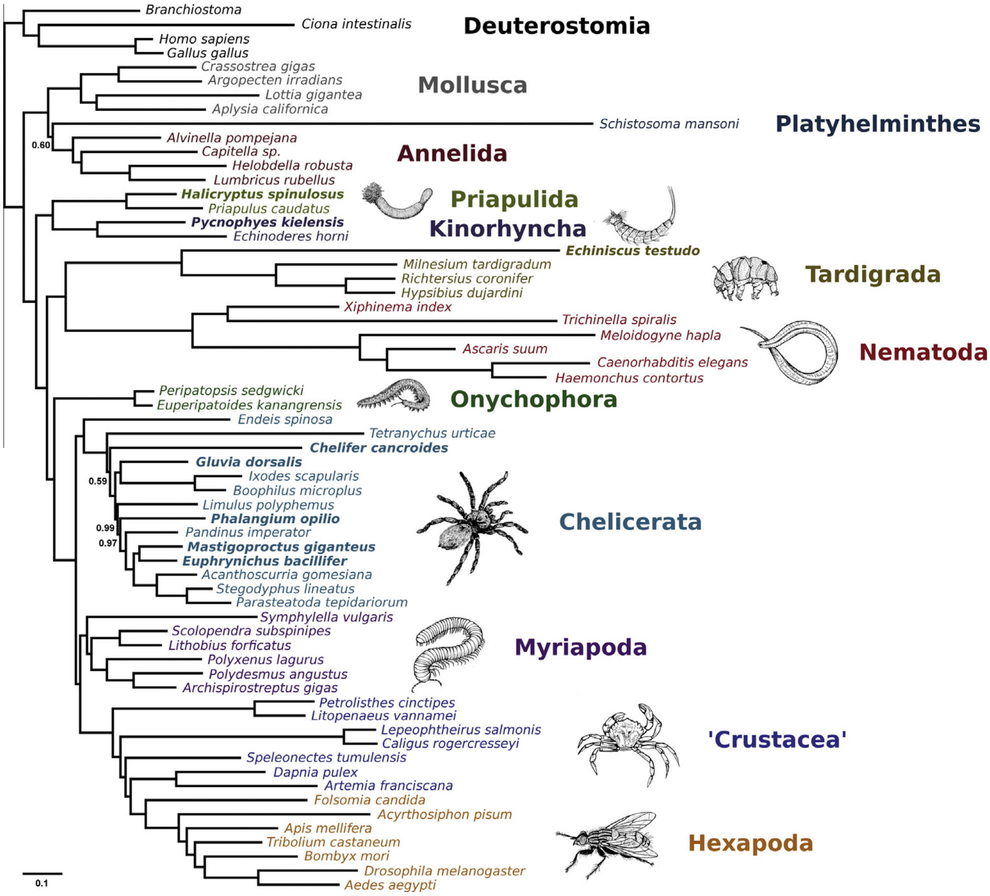


Fig. 1. Ecdysozoan phylogeny based on a Bayesian analysis of 63 taxa and 24,249 amino acid positions using the CAT–GTR Γ model. Bayesian posterior probabilities <1.00 are given at the nodes; all other splits have maximum support. Taxa for which we collected new data are depicted in bold letters.

et al., 2014). This uncertainty is perfectly mirrored in our results and can be explained by the application of datasets with different evolution rates (Fig. 2). We are not in the position to finally resolve these two issues, but certainly the inclusion of a large amount of data should increase reliability of the phylogenetic reconstructions by enhancing the signal to noise ratio (Hillis and Huelsenbeck, 1992). Moreover, given the consistency of morphological considerations and recent phylogenetic analyses (Regier et al., 2010; Rota-Stabelli et al., 2011; Giribet and Edgecombe, 2012; Rehm et al., 2014), including this study, the validity of Mandibulata should be considered more likely than alternative hypotheses. The relationships of the tardigrades are less clear (see below).

4.2. Monophyly of Ecdysozoa but paraphyly of Cycloneuralia

The relationships among the pseudocoelomate phyla, initially referred to as Nemathelminthes, have been disputed for more than 50 years (for review, see Schmidt-Rhaesa, 2013). In recent studies (Dunn et al., 2008), the worm-like nemathelminth phyla Nematoda, Nematomorpha, Priapulida, Kinorhyncha and Loricifera have been united as “Cycloneuralia”, a designation that refers to the circumpharyngeal nerve-ring (Ahlrichs, 1995; Schmidt-Rhaesa, 2013). Another alleged synapomorphy of the cycloneuralian taxa is a retractable head (introvert), which led to the proposal of the alternative name “Introverta” (Nielsen, 1995). However, this character is disputed, not present in all taxa, and in many cases only in larval stages (Schmidt-Rhaesa, 2013). Other phyla (Acanthocephala, Gastrotricha and Rotifera) that had initially been included in Nemathelminthes were actually identified as members of the lophotrochozoans (Witek et al., 2009; Rota-Stabelli et al., 2010; Schmidt-Rhaesa, 2013; Wey-Fabrizius et al., 2013). We obtained phylogenomic data from three of the five cycloneuralian phyla. ESTs are in fact available from the nematomorph *Spinochordodes tellinii* (Dunn et al., 2008), but the sequences had 98.6% missing data in the initial dataset and this taxon was thus excluded (Note that other cycloneuralians used by Dunn et al. (2008) had similarly low coverage in our dataset). However, the sister group relationship of Nematomorpha and Nematoda is undisputed and has received significant support from both morphological and molecular studies (Nielsen, 1995; Dunn et al., 2008; Telford et al., 2008; Schmidt-Rhaesa, 2013). These two phyla are also referred to as “Nematoida” (Nielsen, 1995). Loriciferans, which are minute marine sediment-dwelling animals, could not be obtained for transcriptome analyses.

The inclusion of the transcriptome data from a priapulid, a kinorhynch and a heterotardigrade (previous sequence data were from eutardigrades) significantly increases data and taxon coverage of Ecdysozoa. The phylogenetic trees agree that Cycloneuralia are associated with Arthropoda, Onychophora and Tardigrada, thereby supporting monophyletic Ecdysozoa. While e.g. Dunn et al. (2008) recovered monophyletic Cycloneuralia (including Tardigrada; see below) as sister group of a common clade of Arthropoda + Onychophora, our analyses suggest that the cycloneuralians are a paraphyletic assemblage (Figs. 1–3). A common clade of Priapulida and Kinorhyncha, which received consistent support in all of our analyses based on amino acid data, was recovered in a basal position within Ecdysozoa. A close relationship of Priapulida and Kinorhyncha has repeatedly been proposed and the name Scalidophora (a taxon that also includes Loricifera) has been proposed, which refers to the spines (scalids) covering the introvert.

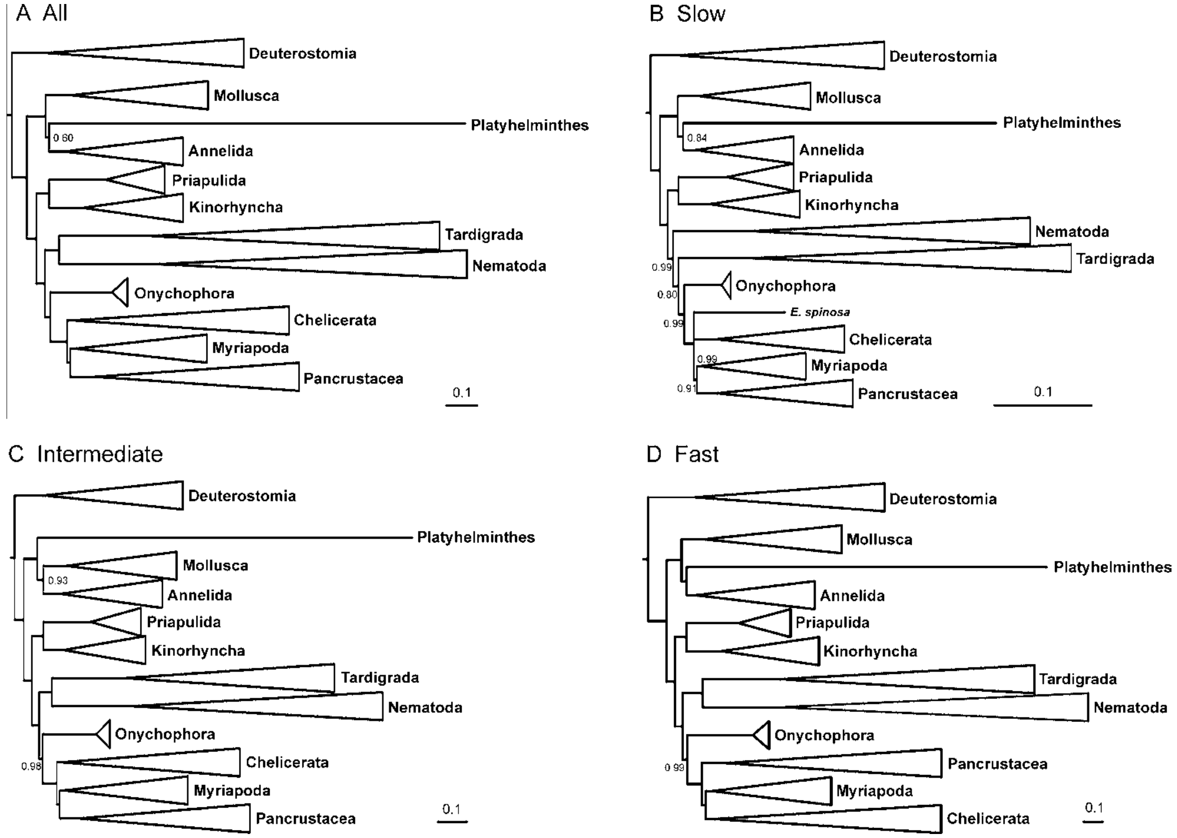


Fig. 2. Phylogenetic relationships among Ecdysozoa inferred using three subsets of proteins with different evolutionary rates. The proteins used for the initial phylogenetic inference (A) were categorized into three subsets comprising 63 proteins each: (B) slow-evolving proteins, (C) proteins evolving with an intermediate rate, and (D) fast-evolving-proteins. Trees were inferred by Bayesian analyses with the CAT–GTR Γ model. Bayesian posterior probabilities <1.00 are given at the nodes; all other splits have maximum support. Scale bars are equal to 0.1 expected substitutions per site.

Notably, the basal position of Scalidophora and thus the paraphyly of Cycloneuralia has already been suggested on the basis of rRNA (Garey, 2001; Mallatt et al., 2012) and by recent multigene approaches (Campbell et al., 2011; Rota-Stabelli et al., 2013).

4.3. Nematoda and the conflicting evidence for the relationships of water bears

In most early studies, the nematodes were exclusively represented by *C. elegans*, for which genomic data was already available. However, *C. elegans* (along with several other nematodes) is a long branching taxon, which has accumulated many substitutions in its genome (Dopazo and Dopazo, 2005). Therefore, its position tended to be close to the root of Metazoa. The inclusion of more slowly evolving taxa robustly resolved the nematodes (possibly together with the tardigrades; see below) as sister taxon of Panarthropoda (Figs. 1–3) (see also Campbell et al., 2011; Rota-Stabelli et al., 2013).

There is surprisingly little agreement in the literature on the position of Tardigrada (water bears). Based on arthropod-like characters such as the segmented body, the presence of a peritrophic membrane, limbs, and a ladder-like central nervous system, Tardigrada have been considered by most textbooks as closely related to Arthropoda (Westheide and Rieger, 1996; Brusca and Brusca, 2003). The arthropod affinities of tardigrades received support from comparative developmental studies (Gabriel and Goldstein, 2007). However, tardigrades also share characters with Cycloneuralia, including similarities of the mouth, pharynx, cuticle and some sensory organs (Giribet, 2003). Molecular studies gave conflicting results either supporting an arthropod affiliation of tardigrades (Garey, 2001; Mallatt et al., 2004; Campbell et al., 2011; Rota-Stabelli et al., 2011) or suggesting their association with nematodes (Giribet, 2003; Roeding et al., 2007; Lartillot and Philippe, 2008; Meusemann et al., 2010). This conflict is also mirrored by our results (Fig. 2). Employing only slowly evolving proteins, we recovered the tardigrades as sister clade to the arthropods (including Onychophora), but the support was comparably low (0.80 Bayesian posterior probability). In fact, Campbell and colleagues (2011) suggested that the nematode affinity of tardigrades may be the result of LBA, an effect that can be minimized by using slowly evolving sequences. However, we recovered a close relationship of tardigrades and nematodes with maximum support in all other analyses employing either the full dataset (Fig. 1) or the datasets including proteins that evolve at an intermediate or fast rate (Fig. 2). The topology also remained unaffected when fast evolving positions (Supplemental Fig. S4) or taxa (Supplemental Fig. S5) were excluded. Thus there is no clear indication that LBA plays a major role in our analyses. The indel dataset, which provides a largely independent approach, also recovered a clade of Nematoda and Tardigrada (Fig. 3). Notably, neither this nor any other phylogenomic study (Campbell et al., 2011; Rota-Stabelli et al., 2011) found a sister group relationship of Tardigrada and Euarthopoda (Arthropoda excluding Onychophora). Thus, the "Tactopoda" hypothesis (Budd, 2001) was rejected.

Although none of our datasets or analysis methods significantly support a close relationship of Tardigrada and Arthropoda to the exclusion of Nematoda, we must point out that neither the multiple sequence alignment (Campbell et al., 2011) nor indel dataset (Telford and Copley, 2011) are safe from disturbing effects such as LBA, which may not be identified with our methods. Given the conflicting evidence in the literature, the phylogenetic affinities of Tardigrada must still be considered as unresolved. If a sister group relationship of Tardigrada and Nematoda (or, probably, Tar-

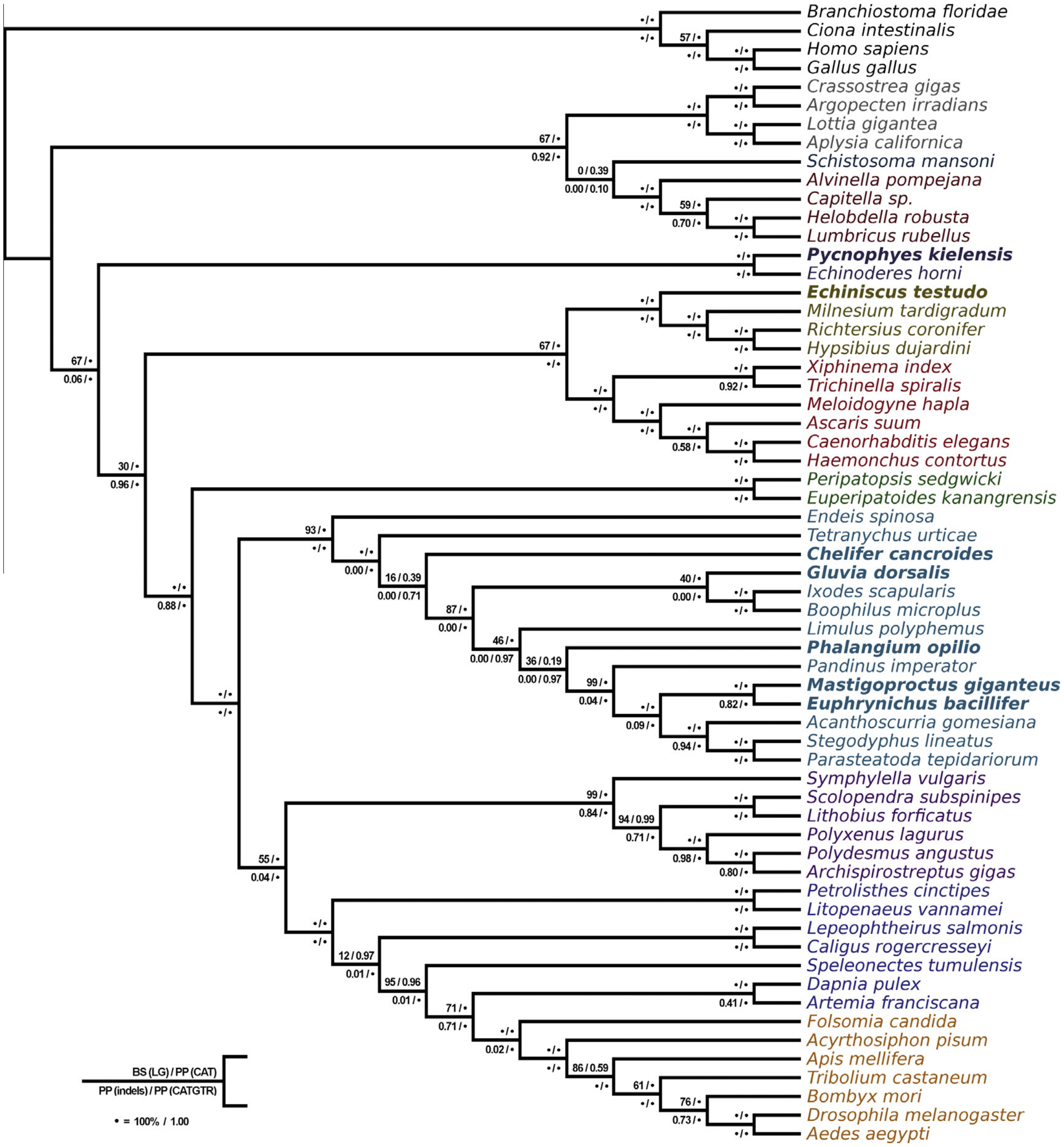


Fig. 3. Cladogram of Ecdysozoan relationships after removal of Priapulida. The topology is based on a Bayesian analysis of 61 taxa and 24,249 amino acid positions with the CAT–GTR Γ model. In addition to the posterior probabilities of this analysis, support values from a ML analysis of the amino-acid dataset, a Bayesian analysis employing the CAT model and a Bayesian analyses of the indel dataset were mapped onto the cladogram with SumTrees from the DendroPy package (Sukumaran and Holder, 2010). Taxa for which we collected new data are depicted in bold letters.

digrada + "Nematoida" sensu Nielsen) is eventually confirmed, there are two possible explanations for the morphology of tardigrades: either the arthropod-like characters (see above) are the result of parallel evolution or they may actually be part of the common ground-pattern of the clade of Panarthropoda + Nematoida, and have subsequently been lost in Nematoida.

4.4. Phylogenomics supports Mandibulata but failed to solve chelicerate relationships

The relationships within Arthropoda have received much attention in recent years (Webster et al., 2006; Roeding et al., 2007, 2009; Dunn et al., 2008; Telford et al., 2008; Meusemann et al., 2010; Rota-Stabelli et al., 2011). Our phylogenetic analyses agree with current taxonomic interpretations and recovered the onychophorans as sister group of the euarthropods. Notably, the suspect placement of Myriapoda as sister group of Chelicerata was found in the tree based on fast evolving proteins (Fig. 2D) and, surprisingly, slowly evolving positions (Supplemental Fig. S4C). This finding emphasizes the importance of the amount and the completeness of the data and might explain previous results from phylogenomic analyses using ESTs (Dunn et al., 2008; Meusemann et al., 2010). In fact, the addition of more taxa to the datasets resulted in monophyletic Mandibulata (i.e. Myriapoda, Crustacea, Hexapoda) (see also Regier et al., 2010; Rota-Stabelli et al., 2011; Rehm et al., 2014).

We observed a surprisingly poor resolution of the tree within the euchelicerates. Notably, this problem applies to both the analyses of sequence data (Fig. 1) and to tree reconstructions based on indel patterns (Fig. 3). Initially, we hoped that the addition of five key chelicerate taxa would improve our understanding of chelicerate evolution (see also Dunlop et al., 2014). However, only the relative topology (Pycnogonida, (Xiphosura, (Opiliones, (Scorpiones, ((Amblypygi, Uropygi), Araneae))))), which is in line with current interpretations of chelicerate taxonomy (Dunlop et al., 2014), was recovered with high support. Acari (mites and ticks), Pseudoscorpiones and Solifugae were found essentially unresolved at the base of the euchelicerate tree. Most trees show the Acari paraphyletic because Acariformes and Parasitiformes do not form sister taxa in our analyses. In none of our analyses, Arachnida were found monophyletic because the generally accepted sister taxon (Xiphosura; horseshoe crabs) assumes an ingroup position. This may be in part explained by fast evolutionary changes in Acariformes and Solifugae that led to LBA effects (Fig. 1). Alternatively, early euchelicerate evolution, probably accompanied by independent terrestrialization events within this taxon, occurred in a relatively short period of time and was thus more complex than commonly assumed. Clearly additional sequences, particularly from poorly covered chelicerate taxa, are required.

4.5. Applicability of indel patterns for phylogenetic reconstructions

Rare genomic changes (RGC) are considered as an alternative and powerful tool to resolve early evolutionary events within a phylogenetic tree, particularly in cases when sequence based studies fail (Rokas and Holland, 2000). The particular advantage of RGC is the comparably low rate of homoplasy. RGC may refer to the insertion of introns or retrotransposons (SINEs, LINEs), the presence of microRNA families, or the pattern of orthologous indels, which represent information that is mostly independent from the sequences themselves. Our analyses of the indels show modest resolution of the ecdysozoan tree (Fig. 3). It must be considered that the indel dataset had a high amount of missing data (69.4%), firstly because it derived from the large, unreduced multiple sequence alignment and secondly because completely overlapping gaps were treated as missing data for the smaller gap (see Materials and Methods). Large amounts of missing data can increase the number of equally parsimonious trees (Wilkinson, 1995), as observed in our parsimony trees. Bayesian analysis confirmed that the indel dataset does not contain sufficient phylogenetic signal to infer a fully resolved tree. However, focusing on those splits which received high support in this analysis, such as monophyletic Arthropoda, Chelicerata, Myriapoda and Pancrustacea, there is clearly a high level of congruence between the indel-based tree and the trees derived from amino acid sequences. The sister group relationship of Tardigrada and Nematoda, which is contended (see above), is also supported in analyses of both types of data. In this context, it should be pointed out that analyses of indel patterns were not possible with the three subsets (slow, intermediate, fast evolving proteins), as the subsets of the indel dataset do not contain sufficient phylogenetic information, resulting in essentially unresolved trees (data not shown). Thus, additional sequences from taxa with poor coverage may be required.

4.6. Conclusions: Ecdysozoan origins

There is conclusive evidence from this as well as from several other studies (Webster et al., 2006; Roeding et al., 2007; Dunn et al., 2008; Telford et al., 2008; Meusemann et al., 2010; Campbell et al., 2011; Rota-Stabelli et al., 2011, 2013) that Ecdysozoa, Scalidophora, Arthropoda (including Onychophora, but possibly not Tardigrada; see above) are monophyletic clades and thus valid taxa. This may allow some conclusions on the appearance of early stemline-ecdysozoans and the last common ancestor of Ecdysozoa. Regardless of the uncertainty of the position of the tardigrades, the basal position of Scalidophora and the sister group relationship of Nematoida and Panarthropoda suggest that the proposed cycloneuralian synapomorphies, such as the terminal mouth, the circumpharyngeal central nervous system and (possibly) the introvert may actually be plesiomorphies of Ecdysozoa (see also Campbell et al., 2011), which have been secondarily lost in Arthropoda. In fact, it is not difficult to imagine that the higher degree of cephalization and the ventral position of the mouth in Arthropoda account for the modification of the circumpharyngeal nerve-ring and the loss of the introvert. On the other hand, characters such as segmentation, paired appendages and jointed limbs, are arthropod autapomorphies.

Putative plesiomorphies (terminal mouth, circumpharyngeal nerve-ring and introvert) should be considered in the search for fossils of stemline ecdysozoans. Candidates for close relatives of stemline ecdysozoans are the palaeoscolecids from the early Cambrian to late Silurian period. These worm-like animals with distinctive cuticle ornamentation resemble articulated priapulids (Budd and Jensen, 2000; Budd, 2003; Harvey et al., 2010). It has been argued that the interpretation of these fossils as ancestral ecdysozoans is logical because of the lack of alternative scenarios: kinorhynchs and loriciferans are adapted to meiofaunal ecology, nematodes and nematomorphs are formed by their parasitic lifestyle and panarthropods are shaped by articulation and the paired appendages (Harvey et al., 2010). Our results are fully compatible with this interpretation and suggest that morphological and molecular interpretations of animal evolution may eventually converge.

Acknowledgments

We thank Andreas Schmidt-Rhaesa for his help with the collection of *Pycnophyes kielensis* and for critical reading of the manuscript. We are deeply saddened of the passing of Heye Rumohr in September last year and would like to express our thanks for his aid in the sampling of *Halicryptus spinulosus*. We are also grateful to his colleagues from GEOMAR Kiel and to the crew of RV *Meteor* for their support during two sampling excursions to the Baltic Sea near Kiel. The TNT program is provided by the Willi Hennig Society. We thank Monika Hänel for her help with the drawings. This work has been supported by grants of the Deutsche Forschungsgemeinschaft within the Priority Project 1174 “Deep Metazoan Phylogeny” to TB (Bu 956/8). IE acknowledges financial support by the Biodiversity and Climate Research Center Frankfurt (BIK-F), Frankfurt, Germany.

Appendix A. Supplementary matrial

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2014.08.001.

References

Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21, 2104–2105.

Aguinaldo, A.M.A., Turbeville, J.M., Linford, L.S., Rivera, M.C., Garey, J.R., Raff, R.A., Lake, J.A., 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 387, 489–493.

Ahlrichs, W., 1995. Ultrastruktur und Phylogenie von *Seison nebaliae* (Grube 1859) und *Seison annulatus* (Claus 1876): Hypothesen zu phylogenetischen Verwandschaftsverhältnissen Innerhalb der Bilateria. Cuvillier Verlag, Göttingen.

Blair, J.E., Ikeo, K., Gojobori, T., Hedges, S.B., 2002. The evolutionary position of nematodes. BMC Evol. Biol. 2, 7.

Bourlat, S.J., Nielsen, C., Economou, A.D., Telford, M.J., 2008. Testing the new animal phylogeny: a phylum level molecular analysis of the animal kingdom. Mol. Phylogenet. Evol. 49, 23–31.

Brewer, M.S., Bond, J.E., 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.

Brusca, R.C., Brusca, G.J., 2003. Invertebrates. Sunderland.

Budd, G.E., 2001. Tardigrades as ‘Stem-Group Arthropods’: The Evidence from the Cambrian Fauna. Zool. Anz. 240, 265–279.

Budd, G.E., 2003. Arthropods as Ecdysozoans: the Fossil Evidence. In: Legakis, A., Sfenthourakis, S., Polymeni, R., Thessalou-Leggaki, M. (Eds.), The New Panorama of Animal Evolution. Proceedings of the XVIII International Congress of Zoology, Athens, Greece, September 2000. Pensoft, Sofia, pp. 479–487.

Budd, G.E., Jensen, S., 2000. A critical reappraisal of the fossil record of the bilaterian phyla. Biol. Rev. (Camb) 75, 253–295.

Campbell, L.I., Rota-Stabelli, O., Edgecombe, G.D., Marchioro, T., Longhorn, S.J., Telford, M.J., Philippe, H., Rebecchi, L., Peterson, K.J., Pisani, D., 2011. MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda. Proc. Natil. Acad. Sci. USA 108, 15920–15924.

Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., Bork, P., 2006. Toward automatic reconstruction of a highly resolved tree of life. Science 311, 1283–1287.

Copley, R.R., Aloy, P., Russell, R.B., Telford, M.J., 2004. Systematic searches for molecular synapomorphies in model metazoan genomes give some support for Ecdysozoa after accounting for the idiosyncrasies of *Caenorhabditis elegans*. Evol. Dev. 6, 164–169.

Dopazo, H., Dopazo, J., 2005. Genome-scale evidence of the nematode-arthropod clade. Genome Biol. 6, R41.

Dunlop, J.A., Borner, J., Burmester, T., 2014. Phylogeny of the Chelicerates: Morphological and Molecular Evidence. In: Waegele, J.W., Bartolomaeus, T. (Eds.), Deep Metazoan Phylogeny: The Backbone of the Tree of Life: New insights from analyses of molecules, morphology, and theory of data analysis. De Gruyter, Berlin, pp. 395–408.

Dunn, C.W., Hejnol, A., Matus, D.Q., Pang, K., Browne, W.E., Smith, S.A., Seaver, E., Rouse, G.W., Obst, M., Edgecombe, G.D., Sorensen, M.V., Haddock, S.H., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R.M., Wheeler, W.C., Martindale, M.Q., Giribet, G., 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452, 745–749.

Ebersberger, I., Strauss, S., Haeseler, A., 2009. HaMStR: profile hidden markov model based search for orthologs in ESTs. BMC Evol. Biol. 9, 157.

Gabriel, W.N., Goldstein, B., 2007. Segmental expression of Pax3/7 and engrailed homologs in tardigrade development. Dev. Genes Evol. 217, 421–433.

Garey, J.R., 2001. Ecdysozoa: The relationship between Cycloneuralia and Panarthropoda. Zool. Anz. 240, 321–330.

Giribet, G., 2003. Molecules, development and fossils in the study of metazoan evolution; Articulata versus Ecdysozoa revisited. Zoology (Jena) 106, 303–326.

Giribet, G., Edgecombe, G.D., 2012. Reevaluating the arthropod tree of life. Annu. Rev. Entomol. 57, 167–186.

Giribet, G., Ribera, C., 1998. The position of arthropods in the animal kingdom: a search for a reliable outgroup for internal arthropod phylogeny. Mol. Phylogenet. Evol. 9, 481–488.

Goloboff, P.A., Farris, J.S., Nixon, K.C., 2008. TNT, a free program for phylogenetic analysis. Cladistics 24, 774–786.

Harvey, T.H.P., Dong, X., Donoghue, P.C.J., 2010. Are palaeoscolecids ancestral ecdysozoans? Evol. Dev. 12, 177–200.

Hejnol, A., Obst, M., Stamatakis, A., Ott, M., Rouse, G.W., Edgecombe, G.D., Martinez, P., Baguna, J., Bailly, X., Jondelius, U., Wiens, M., Müller, W.E., Seaver, E., Wheeler, W.C., Martindale, M.Q., Giribet, G., Dunn, C.W., 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc. R. Soc. Lond. B Biol. Sci. 276, 4261–4270.

Hillis, D.M., Huelsenbeck, J.P., 1992. Signal, noise, and reliability in molecular phylogenetic analyses. J. Hered. 83, 189–195.

Holmes, D.S., Bonner, J., 1973. Preparation, molecular weight, base composition, and secondary structure of giant nuclear ribonucleic acid. Biochemistry 12, 2330–2338.

Hwang, U.W., Friedrich, M., Tautz, D., Park, C.J., Kim, W., 2001. Mitochondrial protein phylogeny joins myriapods with chelicerates. Nature 413, 154–157.

Irimia, M., Maeso, I., Penny, D., Garcia-Fernandez, J., Roy, S.W., 2007. Rare coding sequence changes are consistent with Ecdysozoa, not Coelomata. Mol. Biol. Evol. 24, 1604–1607.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780.

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.

Lartillot, N., Philippe, H., 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Philos. Trans. R. Soc. Lond. B Biol. Sci. 363, 1463–1472.

Lartillot, N., Lepage, T., Blanquart, S., 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25, 2286–2288.

Le, S.Q., Gascuel, O., 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25, 1307–1320.

Mallatt, J.M., Garey, J.R., Shultz, J.W., 2004. Ecdysozoan phylogeny and Bayesian inference: first use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin. Mol. Phylogenet. Evol. 31, 178–191.

Mallatt, J., Craig, C.W., Yoder, M.J., 2012. Nearly complete rRNA genes from 371 Animalia: updated structure-based alignment and detailed phylogenetic analysis. Mol. Phylogenet. Evol. 64, 603–617.

Mayer, G., Martin, C., Rudiger, J., Kauschke, S., Stevenson, P.A., Poprawa, I., Hohberg, K., Schill, R.O., Pfluger, H.J., Schlegel, M., 2013. Selective neuronal staining in tardigrades and onychophorans provides insights into the evolution of segmental ganglia in panarthropods. BMC Evol. Biol. 13, 230.

Meusemann, K., Reumont, B.M., Simon, S., Roeding, F., Strauss, S., Kück, P., Ebersberger, I., Walzl, M., Pass, G., Breuers, S., Achter, V., Haeseler, A., Burmester, T., Hadrys, H., Wägele, J.W., Misof, B., 2010. A phylogenomic approach to resolve the arthropod tree of life. Mol. Biol. Evol. 27, 2451–2464.

Nielsen, C., 1995. Animal Evolution. Interrelationships of the Living Phyla. Oxford University Press, Oxford, UK.

Philip, G.K., Creevey, C.J., McInerney, J.O., 2005. The Opisthokonta and the Ecdysozoa may not be clades: stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa. Mol. Biol. Evol. 22, 1175–1184.

Philippe, H., Lartillot, N., Brinkmann, H., 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol. Biol. Evol. 22, 1246–1253.

Pisani, D., Poling, L.L., Lyons-Weiler, M., Hedges, S.B., 2004. The colonization of land by animals: molecular phylogeny and divergence times among arthropods. BMC Biol. 2, 1.

Regier, J.C., Shultz, J.W., Zwick, A., Hussey, A., Ball, B., Wetzer, R., Martin, J.W., Cunningham, C.W., 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature 463, 1079–1083.

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.

Roeding, F., Hagner-Holler, S., Ruhberg, H., Ebersberger, I., Haeseler, A., Kube, M., Reinhardt, R., Burmester, T., 2007. EST sequencing of Onychophora and phylogenomic analysis of Metazoa. Mol. Phylogenet. Evol. 45, 942–951.

Roeding, F., Borner, J., Kube, M., Klages, S., Reinhardt, R., Burmester, T., 2009. A 454 sequencing approach for large scale phylogenomic analysis of the common emperor scorpion (*Pandinus imperator*). Mol. Phylogenet. Evol. 53, 826–834.

Rogozin, I.B., Wolf, Y.I., Carmel, L., Koonin, E.V., 2007. Ecdysozoan clade rejected by genome-wide analysis of rare amino acid replacements. Mol. Biol. Evol. 24, 1080–1090.

Rokas, A., Holland, P.W., 2000. Rare genomic changes as a tool for phylogenetics. Trends Ecol. Evol. 15, 454–459.

Rota-Stabelli, O., Kayal, E., Gleeson, D., Daub, J., Boore, J.L., Telford, M.J., Pisani, D., Blaxter, M., Lavrov, D.V., 2010. Ecdysozoan mitogenomics: evidence for a common origin of the legged invertebrates, the Panarthropoda. Genome Biol. Evol. 2, 425–440.

Rota-Stabelli, O., Campbell, L., Brinkmann, H., Edgecombe, G.D., Longhorn, S.J., Peterson, K.J., Pisani, D., Philippe, H., Telford, M.J., 2011. A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. Proc. R. Soc. Lond. B. Biol. Sci. 278, 298–306.

Rota-Stabelli, O., Daley, A.C., Pisani, D., 2013. Molecular timetrees reveal a cambrian colonization of land and a new scenario for ecdysozoan evolution. Curr. Biol. 23, 392–398.

Schmidt-Rhaesa, A., 2013. Nematomorpha, Priapulida, Kinorhyncha. Loricifera. de Gruyter, Berlin/Boston.

Schmidt-Rhaesa, A., Bartolomaeus, T., Lemburg, C., Ehlers, U., Garey, J.R., 1998. The position of the Arthropoda in the phylogenetic system. J. Morphol. 238, 263–285.

Simmons, M.P., Ochoterena, H., 2000. Gaps as characters in sequence-based phylogenetic analyses. Syst. Biol. 49, 369–381.

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

Struck, T.H., 2014. TreSpEx-detection of misleading signal in phylogenetic reconstructions based on tree information. Evol. Bioinform. Online 10, 51–67.

Sukumaran, J., Holder, M.T., 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26, 1569–1571.

Swofford, D.L., 2003. PAUP∗. Phylogenetic Analysis Using Parsimony (∗and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts, USA.

Talavera, G., Castresana, J., 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56, 564–577.

Telford, M.J., Copley, R.R., 2011. Improving animal phylogenies with genomic data. Trends Genet. 27, 186–195.

Telford, M.J., Bourlat, S.J., Economou, A., Papillon, D., Rota-Stabelli, O., 2008. The evolution of the Ecdysozoa. Philos. Trans. R. Soc. Lond. B Biol. Sci. 363, 1529–1537.

Telford, M.J., Bourlat, S.J., Economou, A., Papillon, D., Rota-Stabelli, O., 2009. The Origins and Evolution of the Ecdysozoa. In: Telford, M.J. (Ed.), Animal Evolution. Genomes, Fossils, and Trees, pp. 71–79.

Webster, B.L., Copley, R.R., Jenner, R.A., Mackenzie-Dodds, J.A., Bourlat, S.J., Rota-Stabelli, O., Littlewood, D.T., Telford, M.J., 2006. Mitogenomics and phylogenomics reveal priapulid worms as extant models of the ancestral Ecdysozoan. Evol. Dev. 8, 502–510.

Westheide, W., Rieger, R. (Eds.), 1996. Spezielle Zoologie. Gustav Fischer, Stuttgard, Jena, New York.

Wey-Fabrizius, A.R., Podsiadlowski, L., Herlyn, H., Hankeln, T., 2013. Platyzoan mitochondrial genomes. Mol. Phylogenet. Evol. 69, 365–375.

Wilkinson, M., 1995. Coping with abundant missing entries in phylogenetic inference using parsimony. Syst. Biol. 44, 501–514.

Witek, A., Herlyn, H., Ebersberger, I., Mark Welch, D.B., Hankeln, T., 2009. Support for the monophyletic origin of Gnathifera from phylogenomics. Mol. Phylogenet. Evol. 53, 1037–1041.

Wolf, Y.I., Rogozin, I.B., Koonin, E., 2004. Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. Genome Res. 14, 29–36.

Molecular Phylogenetics and Evolution 77 (2014) 25–33

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Phylogenetic position of Myriapoda revealed by 454 transcriptome sequencing

Peter Rehm[a], Karen Meusemann[b,c], Janus Borner[a], Bernhard Misof[b], Thorsten Burmester[a,*]

[a] *Zoologisches Institut & Museum, Biozentrum Grindel, Martin-Luther-King Platz 3, D-20146 Hamburg, Germany*
[b] *Zoologisches Forschungsmuseum Alexander Koenig, Zentrum für Molekulare Biodiversitätsforschung (zmb), Adenauerallee 160, D-53113 Bonn, Germany*
[c] *CSIRO Ecosystem Sciences, Australian National Insect Collection, Clunies Ross Street, Acton, ACT 2601, Australia*

ARTICLE INFO

Article history:
Received 16 December 2013
Revised 31 March 2014
Accepted 3 April 2014
Available online 13 April 2014

Keywords:
Arthropoda
Bayesian analyses
Mandibulata
Maximum Likelihood
Myriapoda
Myriochelata

ABSTRACT

Myriapods had been considered closely allied to hexapods (insects and relatives). However, analyses of molecular sequence data have consistently placed Myriapoda either as a sister group of Pancrustacea, comprising crustaceans and hexapods, and thereby supporting the monophyly of Mandibulata, or retrieved Myriapoda as a sister group of Chelicerata (spiders, ticks, mites and allies). In addition, the relationships among the four myriapod groups (Pauropoda, Symphyla, Diplopoda, Chilopoda) are unclear. To resolve the phylogeny of myriapods and their relationship to other main arthropod groups, we collected transcriptome data from the symphylan *Symphylella vulgaris*, the centipedes *Lithobius forficatus* and *Scolopendra dehaani*, and the millipedes *Polyxenus lagurus*, *Glomeris pustulata* and *Polydesmus angustus* by 454 sequencing. We concatenated a multiple sequence alignment that contained 1550 orthologous single copy genes (1,109,847 amino acid positions) from 55 euarthropod and 14 outgroup taxa. The final selected alignment included 181 genes and 37,425 amino acid positions from 55 taxa, with eight myriapods and 33 other euarthropods. Bayesian analyses robustly recovered monophyletic Mandibulata, Pancrustacea and Myriapoda. Most analyses support a sister group relationship of Symphyla in respect to a clade comprising Chilopoda and Diplopoda. Inclusion of additional sequence data from nine myriapod species resulted in an alignment with poor data density, but broader taxon average. With this dataset we inferred Diplopoda + Pauropoda as closest relatives (*i.e.*, Dignatha) and recovered monophyletic Helminthomorpha. Molecular clock calculations suggest an early Cambrian emergence of Myriapoda ~513 million years ago and a late Cambrian divergence of myriapod classes. This implies a marine origin of the myriapods and independent terrestrialization events during myriapod evolution.

1. Introduction

Arthropods are by far the most successful and species-rich multicellular animal group, but still the relationships among major arthropod taxa are hotly debated (for review, see *e.g.* Edgecombe, 2011; Giribet and Edgecombe, 2012; Giribet and Ribera, 2000; Richter, 2002; Telford et al., 2008). Traditionally, arthropods comprise four subphyla: Chelicerata (spiders, ticks, mites and allies), Myriapoda (centipedes, millipedes, symphylans and pauropods), "Crustacea" (shrimps, crabs and others) and Hexapoda ("true" insects and relatives). Onychophora (velvet worms) and Tardigrada (water bears) have either been included in the arthropod phylum or are considered as related taxa.

Based on a number of supposedly derived morphological characters, Hexapoda were considered being closely related to Myriapoda forming the taxon "Tracheata" or "Atelocerata" (Heymonds, 1901; Pocock, 1893). Alleged synapomorphies that support Tracheata include the elaborate tracheal system, Malpighian tubules as excretory organs, or the loss of the second antennae (Bitsch and Bitsch, 2004). However, in the last ~15 years, a number of molecular studies have accumulated strong evidence that Hexapoda are actually closely related to crustaceans (Boore et al., 1995; Friedrich and Tautz, 1995; Hwang et al., 2001; Kusche and Burmester, 2001; Meusemann et al., 2010; Regier et al., 2010; von Reumont et al., 2012). This view has gained support from comparative morphological studies (Harzsch et al., 2005; Ungerer and Scholtz, 2008; Zrzavý and Štys, 1997). Several lines of evidence also suggested that

Abbreviations: BS, bootstrap support; CIR, Cox–Ingersoll–Ross model; EST, Expressed Sequence Tag; ML, Maximum Likelihood; mya, million years ago; PP, Bayesian posterior probability.

[*] Corresponding author. Address: Institute of Zoology and Zoological Museum, Biocenter Grindel, University of Hamburg, Martin-Luther-King-Platz 3, D-20146 Hamburg, Germany. Fax: +49 40 42838 3937.

E-mail address: thorsten.burmester@uni-hamburg.de (T. Burmester).

http://dx.doi.org/10.1016/j.ympev.2014.04.007
1055-7903/© 2014 Elsevier Inc. All rights reserved.

"Crustacea" may actually be paraphyletic with respect to Hexapoda. Initially, it had been suggested that either the malacostracan (García-Machado et al., 1999; Wilson et al., 2000) or branchiopod crustaceans (Cook et al., 2001; Nardi et al., 2003; Regier et al., 2005) were sister group of Hexapoda. Recently, however, the enigmatic crustacean taxon Remipedia (possibly along with Cephalocarida) was found being closely related to Hexapoda (Ertas et al., 2009; Regier et al., 2010; von Reumont et al., 2012).

Yet, the placement of Myriapoda within the arthropod tree has remained ambiguous. Some molecular phylogenetic studies suggested the inclusion of Myriapoda along with Pancrustacea in the clade Mandibulata (*e.g.*, Boore et al., 1998; Giribet and Ribera, 2000; Kusche et al., 2002; Regier et al., 2010; Rota-Stabelli and Telford, 2008), a taxon which has strong support by morphological evidence. However, other molecular analyses, either applying single- or multi-gene approaches, obtained a common clade of Myriapoda and Chelicerata (*i.e.* "Myriochelata" or "Paradoxopoda" hypothesis; *e.g.*, Dunn et al., 2008; Hwang et al., 2001; Mallatt et al., 2004; Pisani et al., 2004). Morphological evidence for such a hypothesis is poor and restricted to similarities of neurogenesis (Kadner and Stollewerk, 2004).

Almost every possible topology has been proposed for the relationships among myriapods (Edgecombe, 2011). Morphological characters such as anteriorly placed genital openings unite Symphyla, Pauropoda, and Diplopoda in a clade named "Progoneata" (Dohle, 1980). Some morphological studies, as well as molecular analyses of mitochondrial and rRNA sequences, support a sister group relationship of Symphyla and Pauropoda (Dong et al., 2012; Gai et al., 2008; Podsiadlowski et al., 2007). In fact, the phylogenomic study by Regier et al. (2010) recovered monophyletic Progoneata, as well as a sister group relationship of Symphyla and Pauropoda (Edafopoda; Zwick et al., 2012). However, studies based on rRNA have suggested a close relationship of Symphyla + Pauropoda and Chilopoda (Gai et al., 2006) or placed the symphylans as a long branching taxon outside the myriapods (Mallatt and Giribet, 2006; Mallatt et al., 2004).

Arthropod relationships have been mostly investigated on the basis of single-gene analyses (*e.g.*, Aguinaldo et al., 1997; de Rosa et al., 1999; Mallatt et al., 2004; Ruiz-Trillo et al., 2002) or by analyses of a few specifically selected genes (*e.g.*, Philippe et al., 2004). Recently, Regier and colleagues (Regier et al., 2008; Regier et al., 2010) employed 62 protein coding genes (~41 kbp per species) that were amplified by RT–PCR to study arthropod relationships. An alternative method is a "phylogenomic" approach, *i.e.*, using a large number of sequences for phylogenetic purposes (Telford, 2007) that have been derived either from completely sequenced genomes or from transcriptome data ("Expressed Sequence Tags"; ESTs). An increasing number of studies offer promising perspectives demonstrating that ESTs are well suited for phylogenetic analyses at different taxonomic levels (Bourlat et al., 2006; Brewer and Bond, 2013; Dunn et al., 2008; Philippe and Telford, 2006; Roeding et al., 2007).

Although these studies demonstrated that EST-derived multigene matrices harbor sufficient phylogenetic information that may help to resolve deep metazoan phylogeny, many questions concerning important phylogenetic relationships within the deep arthropod phylogeny are still unanswered. One reason may be the lack of sequence data from crucial taxa. The present approach is aimed at closing this gap and allowing for a better resolution among major arthropod taxa, with particular emphasis on myriapods.

2. Materials and methods

2.1. Species collection, preservation and RNA isolation

Specimens of six myriapod species were used in this study to generate transcriptome data: *Glomeris pustulata* Latreille, 1804 (Diplopoda), *Polyxenus lagurus* L., 1758 (Diplopoda), *Polydesmus angustus* Latzel, 1884 (Diplopoda), *Scolopendra dehaani* Brandt, 1840 (Chilopoda), *Lithobius forficatus* L., 1758 (Chilopoda), and *Symphylella vulgaris* Hansen, 1903 (Symphyla) (Supplemental Table 1). Total RNA of *L. forficatus*, *S. dehaani*, *P. angustus* and *P. lagurus* was extracted according to the method of Holmes and Bonner (1973). Total RNA of *G. pustulata* and *S. vulgaris* was extracted with the Absolutely RNA Microprep Kit (Stratagene/Agilent, Waldbronn, Germany).

2.2. Transcriptome sequencing

The cDNA libraries of *G. pustulata*, *P. lagurus*, *S. dehaani* and *S. vulgaris* were constructed at the Max Planck Institute for Molecular Genetics, Berlin, Germany, using CloneMiner (Invitrogen, Darmstadt, Germany) or the SMART approach (Mint-Universal cDNA synthesis kit, Evrogen, Russia). Libraries were normalized using duplex-specific nuclease (Trimmer kit, Evrogen) according to manufacturer's instructions. They were directionally ligated to self-made 454 adaptors with molecular identifier (MID) tags following Roche's technical bulletin TCB 09004 introducing SfiI-sites. The 454 libraries were then immobilized on beads and clonally amplified using the GS FLX Titanium LV emPCR Kit (Roche, Mannheim, Germany), and sequenced using the GS FLX Titanium Sequencing Kit XLR70 (Roche) and GS FLX Titanium PicoTiterPlate Kit (Roche) according to the manufacturer's protocol. 454 sequencing of transcriptomes of *L. forficatus* and *P. angustus* were carried out by LGC Genomics GmbH (Berlin, Germany) according to the same protocol, apart from omitting the cDNA library normalization step. Raw sequence reads were processed and quality checked at the Center for Integrative Bioinformatics (CIBIV), Vienna, Austria (see Table 1). The transcriptomes were screened for possible contaminations with various BLAST-based approaches by cross-comparisons and searches with known protein sequences. *E.g.*, we performed an all-vs.-all BLAST across transcriptomes to identify possible cross-contaminants. We also performed BLAST searches of the transcriptome assemblies with the *Drosophila melanogaster* ribosomal proteins as query. These proteins are expected to represent single copy genes and thus should give single hits in the contigs. No contaminations were detected. The HaMStR procedure described below is further expected to remove remaining contaminants. Finally, we identified the top hit of the myriapod sequences included in the phylogenetic studies within the public databases by TBLASTN. Further we tested the single gene trees employing RAxML (see below). No suspicious sequences were detected.

Raw data have been deposited in the NCBI Sequence Read Archive (SRA), Bioproject IDs PRJNA188160 (*G. pustulata* and *S. vurlgaris*), PRJNA222647 (*P. angustus*), PRJNA222648 (*S. dehaani*), PRJNA222654 (*P. lagurus*) and PRJNA198080 (*Lithobius forficatus*). Assembled contigs are available at the Transcriptome Shotgun Assembly (TSA) database.

2.3. Taxon sampling and orthology assignment

In addition to the transcriptome assemblies of six myriapod species, we selected 63 other species for which transcriptome or genome data were available from public databases (Supplemental Table 2). Species with EST data were selected on the basis of the following criteria: *i.* minimum 1000 contigs per species; *ii.* maximally two species per higher taxonomic unit; *iii.* three species per outgroup taxon; *iv.* within winged insects, only species for which an Official Gene Set from a full genome is available. Exceptions here are dipterans (flies and relatives), because the calibration of the molecular clock analyses required at least two representatives of this order). To increase the taxonomic coverage,

Table 1
Sequencing and assembly of myriapod ESTs.

	Scolopendra dehaani	*Polyxenus lagurus*	*Lithobius forficatus*	*Polydesmus angustus*	*Glomeris pustulata*	*Symphylella vulgaris*
Reads	530,968	296,808	426,071	427,639	853,317	921,312
Mean read length	368	217	399	463	245	404
Unigene contigs	25,577	21,937	13,235	8116	55,149	79,396
Contigs ⩾ 1 kb	184	82	897	1441	662	2037
HaMStr orthologs	584	516	183	238	1264	1140

we added the gene fragments from the taxa sequenced by Regier et al. (2010).

Orthology of transcripts was assigned on amino acid level with the HaMStR pipeline (Ebersberger et al., 2009; http://www.deep-phylogeny.org/hamstr/), version 3.v4, using ortholog set *insecta-hmmer*3–2 (http://www.deep-phylogeny.org/hamstr/download/datasets/hmmer3/). This set includes 1579 orthologous sequence groups based on the official gene sets of six species: *Apis mellifera*, *Bombyx mori*, *Capitella capitata*, *Daphnia pulex*, *Ixodes scapularis*, *Tribolium castaneum*. The *-strict* option was applied for the best reciprocal hit BLAST search against all reference species, with the exception of *T. castaneum*.

2.4. Multiple sequence alignments

Each group of orthologous sequences was aligned individually using the MAFFT L-INS-i algorithm v. 6.850 (Katoh and Toh, 2008) on amino acid level. Randomly similar aligned sections of each multiple sequence alignment were identified with ALISCORE (Kück et al., 2010; Misof and Misof, 2009). We chose the maximal number of pairwise sequence comparisons, default sliding window size, and a special scoring for gappy amino acid data (see: Meusemann et al., 2010; von Reumont et al., 2012). We discarded poorly aligned sections with ALICUT 2.0 (http://utilities.zfmk.de). Masked gene alignments were finally concatenated into a supermatrix with FASconCAT v.1.0 (Kück and Meusemann, 2010).

To increase matrix saturation in terms of gene coverage and information content, we applied the software *mare* v. 0.1.2-rc (http://mare.zfmk.de; Misof et al., 2013). The selected optimal subset (dataset 1) included 181 genes (37,425 amino acid positions) from 55 taxa (Supplemental Table 2). Six lophotrochozoan taxa (three mollusks and three annelids) were used as outgroup. To elucidate the phylogenetic relationships within Myriapoda, a second dataset was created (dataset 2), which included sequence data of nine additional myriapod species from Regier et al. (2010) and only four outgroup taxa. Dataset 2 covers an alignment length of 22,339 amino acid positions and includes 21 taxa.

2.5. Phylogenetic analyses

Phylogenetic trees were calculated using Maximum Likelihood (ML) and Bayesian approaches. ML trees were inferred with RAxML 7.2.8-ALPHA using the CAT model of rate heterogeneity (Stamatakis, 2006). The LG substitution matrix (Le and Gascuel, 2008) + Γ model was applied based on fitting estimates derived from ModelGenerator v0.85 (Keane et al., 2006). 1000 bootstrap replicates were performed applying the rapid hill-climbing method. We checked *a posteriori* bootstopping criteria (majority rule and majority rule extended, $B = 0.03$ and 0.01) to ensure a statistical convergence of bootstrap replicates. ML analyses were computed on HPC Linux clusters at the SuGI (Sustainable Grid Infrastructures) platform and using Cheops (Regionales Rechenzentrum Cologne, RRZK).

Bayesian tree inference was performed with PhyloBayes3.3f (Lartillot et al., 2009 689) assuming the CAT mixture model (Lartillot and Philippe, 2004). We ran the discrete Γ model (four categories) and the Dirichlet process on site-specific rates each with 16 chains for 20,000 cycles. Sampling was performed every 10th cycle. Bayesian analyses were performed on the HPC Linux cluster of the Regionales Rechenzentrum (RRZ), University of Hamburg. Based on the convergence of all parameters, the first 50% of samples were discarded as burn-in. The discrepancy across all bipartitions (*maxdiff* value) among all chains was derived from comparison of multiple chain combinations with the *bpcomp* tool. The harmonic mean of the likelihood values was calculated for each chain excluding the burn-in. To infer a majority rule consensus tree, the combination of chains was selected which showed the best harmonic mean from all combinations of at least three chains (*maxdiff* value ⩽ 0.2). In addition, 32 chains were run under the CAT GTR model (Lartillot and Philippe, 2004), following the procedure described for the CAT model. Due to the time consuming calculations, these calculations were stopped after 10,000 cycles. The CAT Γ and CAT–GTR Γ models were compared by cross-validation tests as implemented in PhyloBayes. Comparison of models incorporating the Dirichlet process is not implemented in PhyloBayes.

To visualize conflicts in the dataset between different chains and model assumptions, we computed a consensus network (Holland and Moulton, 2003) of all trees sampled from the 64 PhyloBayes chains with SplitsTree 4.8 (Huson and Bryant, 2006), choosing a threshold of 0.01 and averaged edge weights.

2.6. Molecular clock analyses

The Bayesian majority rule consensus tree derived from the CAT mixture and discrete Γ model was used as input for the molecular clock estimates. We applied PhyloBayes 3.3f to calculate divergence times and 95% confidence intervals (Lartillot et al., 2009). Three relaxed clock models were applied: *i.* the uncorrelated gamma multipliers (Drummond et al., 2006), *ii.* the autocorrelated log-normal model (Thorne et al., 1998) and *iii.* the CIR process (Cox et al., 1985; Lepage et al., 2006). The models were compared by calculating the Bayes factors against the unconstrained model employing thermodynamic integration (Lartillot et al., 2009), as implemented in PhyloBayes. Divergence times were computed with four discrete Γ rate categories and a Dirichlet process. In addition, we varied the priors on divergence times using uniform priors with hard bounds and the birth–death prior with both, hard and soft bounds.

Stratigraphic ages were obtained from the International Stratigraphic Chart 2012 (http://www.stratigraphy.org). Eleven calibration points were set essentially as described in Rehm et al. (2011) (see Supplemental Table 3). Further, we defined the divergence time of Lophotrochozoa and Ecdysozoa at 581 Ma (Benton and Donoghue, 2007). The minimum age of the myriapod – pancrustacean split was set 514 mya based on the eucrustacean fossil *Yicaris* (Zhang et al., 2007) from the Atdabanian. The minimum age of the first pancrustacean split 510 mya based on early Cambrian crown-group crustacean fossils (Harvey et al., 2012; Harvey and Butterfield, 2008). We chose a minimal age of the split between Pycnogonida and Euchelicerata based on the first pycnogonid fossil of *Cambropycnogon klausmuelleri* (Waloszek and Dunlop, 2002). The maximum age of the Hexapoda (calibration point 6) was set to the mid-Ordovician (Llanvirn, ~475 mya) according to Regier et al. (2004). Estimations of divergence times were run with two alternative sets, the first including all eleven calibration points, the second omitting the calibration point 8, because it referred to indirect evidence only (see below).

All molecular clock analyses were run for 50,000 cycles sampling every 10th cycle with a burn-in of 2000 samples. Bayes factors between each relaxed clock model and the deconstrained model were estimated using thermodynamic integration as implemented in PhyloBayes (Lepage et al., 2007) with 100,000 generations and a burn-in of 10,000.

3. Results

3.1. Myriapod transcriptomes

Individual cDNA libraries were constructed from mRNA of six myriapod species (*G. pustulata*, *L. forficatus*, *P. angustus*, *P. lagurus*, *S. dehaani*, *S. vulgaris*) and submitted to 454 pyrosequencing. The runs produced 296,808 to 921,312 reads (Table 1). Clustering of the sequences resulted in 8116 to 79,396 contigs based on at least two reads. Between 82 and 2037 contigs were longer than 1 kb. These properties, as well as random BLAST searches of individual contigs for every species, suggest sufficient quality of the sequencing procedure and the transcriptome assemblies for phylogenetic purposes.

3.2. Concatenating datasets

In addition to the myriapod ESTs obtained for this study, we further assembled a broad range of taxa. The initial taxon sampling encompassed 69 taxa (Supplemental Table 2), including three annelids, three mollusks, one priapulid, three nematodes, two tardigrades and two onychophorans. For twelve species, Official Gene Sets derived from full genomes were available and the deduced proteomes were included. Transcripts were assigned to 1550 orthologous genes. Aligning and concatenating all orthologous sequence clusters resulted in an initial super-alignment of 1,109,847 amino acid positions. 36.12% of the sites were identified as randomly similar aligned by ALISCORE (Kück et al., 2010; Misof and Misof, 2009) and excluded from further analyses. Thus, the first supermatrix encompassed 69 taxa with 1550 genes and 708,956 aligned amino acid sites. It displayed a poor information content of 0.086 and matrix coverage in terms of presence/absence of genes of 34.7%. Matrix optimization using MARE, which deletes genes and/or taxa with low information content and low coverage, resulted in a super-alignment of 40,130 amino acids with 181 orthologous genes and 55 taxa. This first selected optimal subset (dataset 1) displayed an about fourfold increase of information content (0.316) and a matrix coverage of 66.1%.

A second dataset (dataset 2) was generated from the first super-alignment to specifically study myriapod evolution. We kept all myriapod taxa used by Regier et al. (2010), which had been deleted from dataset 1 because of poor coverage. However, we excluded the sequence data of *S. coleoptrata* and *L. forficatus* used in Regier et al. (2010) and used available EST data instead (Meusemann et al., 2010). As dataset 2 was created to study the internal myriapod relationships, the outgroup was reduced to four arthropod species resulting in a super-alignment of 22,339 aa positions comprising 21 taxa. Since only about half of the seemingly orthologous genes used in Regier et al. (2010) survived orthology prediction, gene coverage among these taxa was very poor (89.9% missing data).

3.3. Molecular phylogeny of arthropods

Phylogenetic trees were reconstructed by ML and Bayesian methods. In Bayesian analyses, cross-validation showed that the CAT–GTR Γ outperforms the CAT Γ model. With dataset 1 we received high support for the monophyly of Ecdysozoa (Fig. 1; Supplemental Figs. 1–3). Within this taxon, the priapulid *P. caudatus* diverged first, rendering Cycloneuralia paraphyletic. The monophyly of Euarthropoda + Onychophora was consistently recovered in all analyses, although this relationship did not always receive maximum support. The monophyly of Arthropoda, Chelicerata and Myriapoda was maximally supported (each with 100% bootstrap support [BS]/1.00 Bayesian posterior probability [PP] in all analyses). Hexapods and crustaceans formed a common clade, as well with maximal support (*i.e.* Pancrustacea, comprising paraphyletic crustaceans). However, the trees resulting from the four Bayesian and the ML analyses differed in several important aspects: *i.* While the ML tree weakly supported a sister group relationship of Myriapoda and Chelicerata (63% BS), the monophyly of Mandibulata received maximal support in the Bayesian tree (1.00 PP); *ii.* Tardigrada were sister taxon of Nematoda in the ML topology (100% BS, Supplemental Fig. 3), while they were sister taxon of Arthropoda + Onychophora in the Bayesian approaches (1.0 PP); *iii.* within the myriapods, *S. vulgaris* (Symphyla) was either found as sister group of Chilopoda and Diplopoda (ML: 99% BS; Bayes: CAT–GTR Γ and CAT–GTR Dirichlet; 0.99 and 1.00 PP, respectively) or as sister group of Diplopoda (CAT–Γ and CAT-Dirichlet; 0.98 and 0.92 PP, respectively). Additionally, the trees differed with regard to the monophyly or paraphyly of "Acari" (mites and ticks), "Entognatha" (primarily wingless and enthognatous groups: Protura, Diplura and Collembola), "Paleoptera" (mayflies and dragonflies being closest relatives), and the position of copepods within crustaceans.

The consensus network, which provides an indication of the congruence or incongruence among the chains of all Bayesian analyses, supported the monophyly of Ecdysozoa and Arthropoda (Fig. 2). Within Ecdysozoa, the data harbored incongruent signal on the exact position of higher positioned taxa: Tardigrades were sister group to monophyletic Arthropoda (including Onychophora) in all trees derived from Bayesian analyses. However, the network showed that the dataset also contained phylogenetic signal for a close relationship of tardigrades and nematodes. There was strong signal, which did not depend on the phylogenetic model, for the monophyly of Mandibulata. Chelicerata, Myriapoda, Pancrustacea and Hexapoda were also resolved as monophyletic clades. Within these taxa, relationships were less well resolved, with differences under various models, which was compatible with our findings in ML and Bayesian trees. The consensus network also highlights the ambiguity of the mono- or paraphyly of Acari (Acariformes and Parasitiformes); the positions of Symphyla within Myriapoda and Copepoda within crustaceans were not consistent. Likewise, the relationships of entognathous hexapods were poorly resolved, and the position of Paleoptera received only poor support.

3.4. Deducing myriapod phylogeny

As mentioned, the position of Symphyla within Myriapoda remained unsettled. In dataset 1, Bayesian analyses employing CAT-Γ or CAT-Dirichlet found Symphyla as sister group of Diplopoda, while analyses under the GTR model as well as the ML approach recovered Symphyla as sister group of Diplopoda + Chilopoda. Within Chilopoda, the trees were congruent: *S. dehaani* (Scolopendromorpha) and *L. forficatus* (Lithobiomorpha) were more closely related, and *S. coleoptrata* (Scutigeromorpha) was recovered in basal position. This topology received high support in the tree derived from the Bayesian analyses with the CAT–GTR

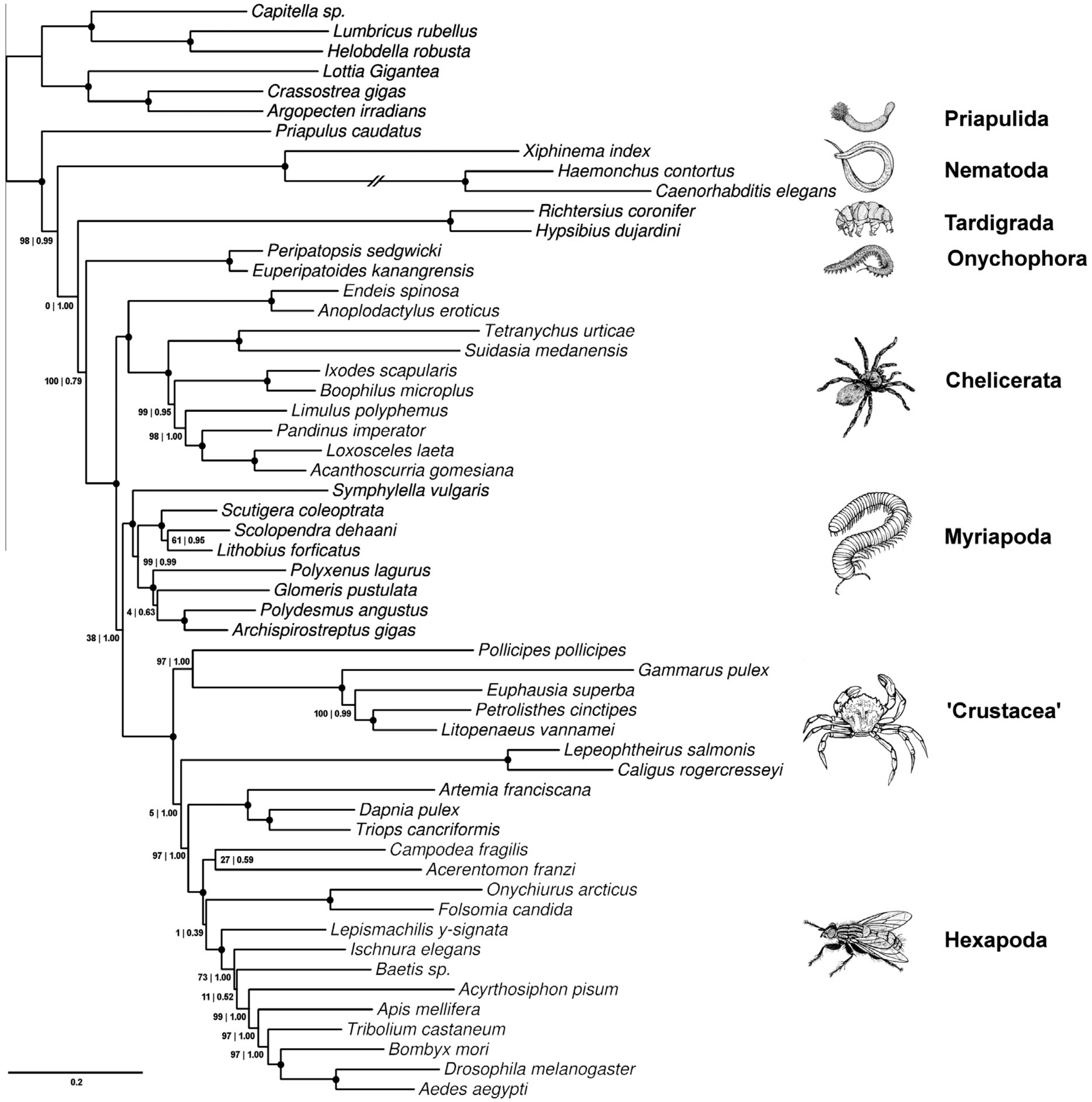


Fig. 1. Arthropod phylogeny based on 181 genes (40,130 amino acid positions) and 55 taxa. The topology was derived from Bayesian analysis with the CAT–GTR Γ model as implemented in PhyloBayes version 3.3.f. Bayesian posterior probabilities (left) and ML bootstrap support values (right) are depicted at the nodes.

model (Γ: 0.95 PP, Dirichlet: 0.97 PP), but was poorly supported in ML analyses (61% BS) and Bayesian analyses with the CAT model (CAT + Γ and CAT-Dirichlet; 0.51 and 0.61 PP, respectively). Within Diplopoda, a sister group relationship of *G. pustulata* (Pentazonia) and *P. lagurus* (Penicillata) received weak support from Bayesian analyses, while *G. pustulata* was basally positioned within the ML trees. A common clade of *A. gigas* (Spirostreptida) and *P. angustus* (Polydesmida) was consistently recovered with high support.

We therefore aimed to extend our taxon sampling within Myriapoda by generating a second dataset (dataset 2), which included additional myriapod taxa from Regier et al. (2010), and fewer outgroup taxa to focus on internal myriapod relationships. Here we found Symphyla consistently being placed in a basal position within Myriapoda (0.92 PP in CAT–GTR + Γ) (Fig. 3). The monophyly of Dignatha (*i.e.* Diplopoda + Pauropoda) was recovered in all analyses. Within Diplopoda and Chilopoda, most clades received poor support in dataset 2. The reasonably well supported clades included Scolopendromorpha, represented by two species of the genus *Scolopendra*, and a clade consisting of *Craterostigmus tasmanianus* (Craterostigmidae) and Scolopendromorpha. Within Diplopoda, a clade consisting of *Abacion magnum* (Callipodida) and *A. gigas* (Spirostreptida), and a clade comprising these two species and *Narceus americanus* (Spirobolida) and *P. angustus* (Polydesmida) were recovered. The two *Polyxenus* species also formed a well supported clade.

3.5. Dating arthropod evolution

Divergence times were estimated based on the Bayesian tree deriving from the CAT–GTR mixture and discrete gamma model as described in Rehm et al. (2011). Comparison of the relaxed clock models employing thermodynamic integration found that the log-normal autocorrelated clock model fits the data best. The logarithms of the Bayes factors were 41.6 for the uncorrelated gamma model, 54.3 for the CIR process and 59.8 for the auto-correlated log-normal model. Divergence times and 95% confidence intervals were estimated according to the log-normal autocorrelated model and are given in Supplemental Table 4. Eleven calibration points

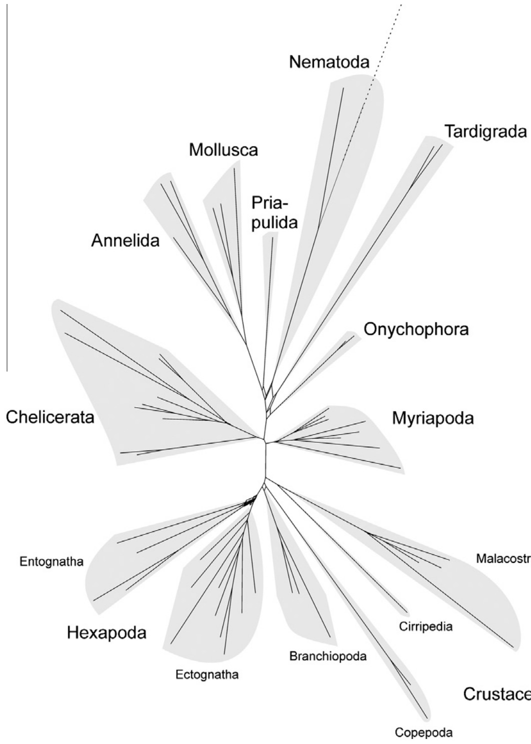


Fig. 2. Consensus network of all trees from the 64 PhyloBayes chains (burn-in excluded). The consensus network was calculated with SplitsTree 4.8 to visualize incongruence between topologies derived from the individual Bayesian chains (threshold = 0.01, averaged weights).

were applied (Supplemental Table 3). Omission of the hexapod maximum age, which only relies on indirect evidence from paleoecological considerations on arthropod terrestrialization (Regier et al., 2004), resulted in less than 1% difference in mean divergence time estimates (Supplemental Table 4). We therefore only give the results based on ten calibration points here.

According to our molecular clock calculations, Ecdysozoa and Lophotrochozoa diverged 582–574 mya (mean 579 mya) (Fig. 4). The origin of Ecdysozoa (split Priapulida vs. others) dated to 581–567 mya (mean 576 mya) in the Precambrian, the origin of Arthropoda (including Tardigrada and Onychophora) was at the border between the Precambrian and the Cambrian 566–548 mya (mean 557 mya). The emergence of Euarthropoda dated to 561–543 mya (mean 552 mya) and this taxon diversified during the early Cambrian: Chelicerata and Mandibulata diverged 543–526 mya (mean 535 mya) and Myriapoda and Pancrustacea diverged 539–522 mya (mean 531 mya). Within Myriapoda, the two splits of Symphyla vs. other myriapods and divergence of Chilopoda and Diplopoda occurred within a short period around 515 mya (528–511 mya mean 520 mya, and 524–506 mya, mean 515 mya, respectively). Modern pancrustaceans commenced to diversify 506–494 mya (mean 499 mya) and the lineage leading to Hexapoda split from Branchiopoda 498–482 mya (mean 490 mya).

4. Discussion

4.1. The phylogenetic position of the Myriapoda

Within the past more than 120 years, the monophyly of Mandibulata, comprising myriapods, hexapods and crustaceans, has received substantial support from morphological analyses (Brusca and Brusca, 2003). However, molecular data have been far less convincing, providing ambiguous results, often along with low support. While – to the best of our knowledge – not a single molecular phylogenetic study supports the traditional view of monophyletic Tracheata (*i.e.* Hexapoda + Myriapoda), several approaches recovered a sister group relationship of Chelicerata and Myriapoda (Paradoxopoda) with more or less strong support (Boore et al., 1995; Dunn et al., 2008; Friedrich and Tautz, 1995; Hwang et al., 2001; Kusche and Burmester, 2001; Mallatt et al., 2004; Meusemann et al., 2010; Pisani et al., 2004), while others suggest Myriapoda to be sister group of Pancrustacea, *i.e.* thereby supporting monophyletic Mandibulata (Boore et al., 1998; Brusca and Brusca, 2003; Giribet and Ribera, 2000; Kusche et al., 2002; Regier et al., 2010; Rota-Stabelli et al., 2011; Rota-Stabelli and Telford, 2008). Recent studies have demonstrated the sensitivity of the position of Myriapoda and its support with respect to data choice, taxon sampling, outgroup selection, and analysis method (Rota-Stabelli et al., 2011). This sensitivity is also highlighted in our study: While the ML approach recovered Myriapoda and Pancrustacea as sister groups (Supplemental Fig. 3), albeit with poor bootstrap support, all Bayesian analyses strongly supported the monophyly of Mandibulata (Figs. 1 and 2; Supplemental Figs. 1 and 2). The poor resolution of the trichotomy of Chelicerata–

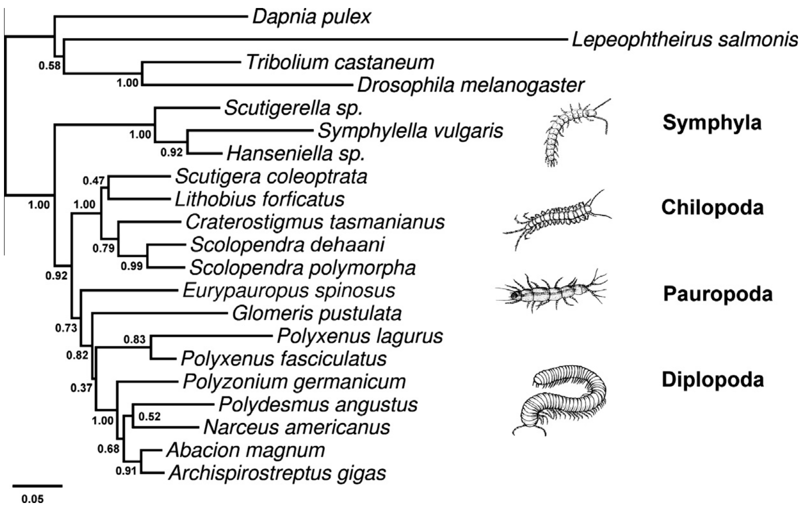


Fig. 3. Myriapod phylogeny based on a 22,339 amino acid alignment from 21 taxa. The topology was derived from Bayesian analysis with the CAT–GTR Γ model as implemented in PhyloBayes version 3.3.f. Bayesian posterior probabilities are depicted at the nodes.

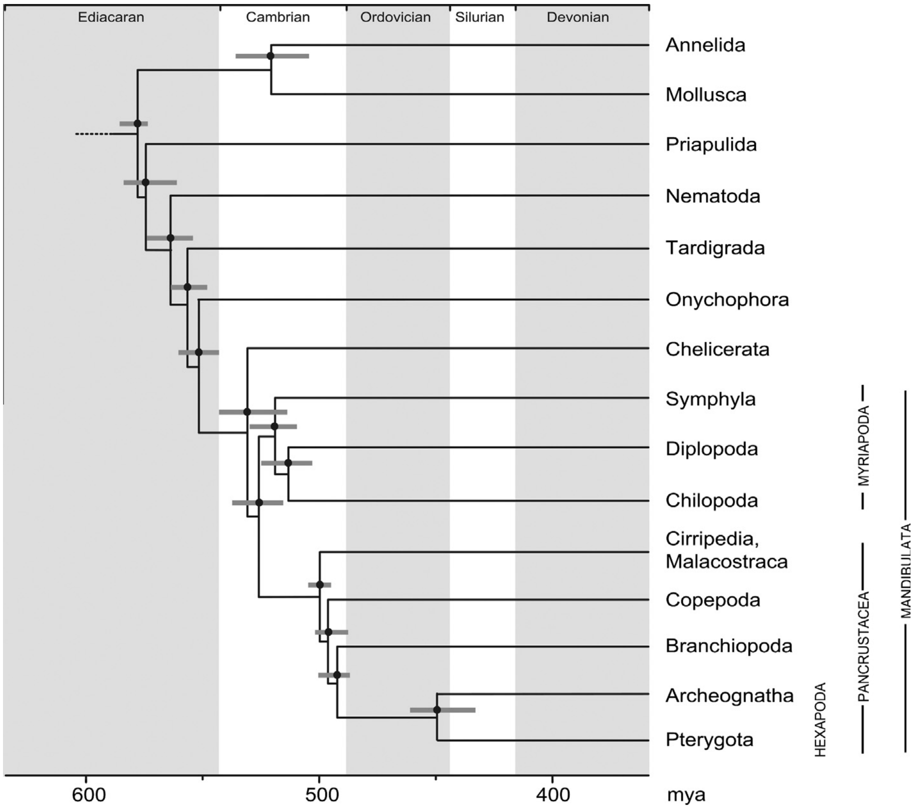


Fig. 4. Molecular clock analyses of arthropod evolution. Mean divergence times were estimated under the log-normal autocorrelated clock model with PhyloBayes version 3.3.f. Gray bars indicate 95% mean confidence intervals (see Supplemental Table 4). The tree was calibrated using the calibrated nodes marked with an asterisk (see Supplemental Table 3 for calibration points), mya, million years ago.

Myriapoda–Pancrustacea, which was also observed in most of the above mentioned studies, can be interpreted in terms of the paucity of phylogenetic signal that may result from a rapid and early divergence of main arthropod subphyla during the “Cambrian explosion” (Fig. 4).

4.2. Relationships within the myriapods

Some authors have doubted myriapod monophyly and suggested that myriapods are either paraphyletic with regard to Hexapoda (Bitsch and Bitsch, 2004; Kraus, 1998), Chelicerata (Negrisolo et al., 2004), or may even form a paraphyletic assemblage at the base of the arthropod tree (Loesel et al., 2002). Our results agree with more recent morphological and molecular analyses (Dunn et al., 2008; Edgecombe, 2011; Gai et al., 2008; Gai et al., 2006; Giribet et al., 2001) and consistently support monophyletic Myriapoda (classes Diplopoda, Chilopoda, Symphyla, Pauropoda).

There is surprisingly little agreement on the relative relationships among the four myriapod classes (Edgecombe, 2011). Our results are ambiguous, too. In the EST-based approach (dataset 1), we found in most of the analyses the symphylans as sister group to diplopods and chilopods (Figs. 1, 3 and 4). Notably, none of the earlier molecular or morphological studies (see introduction) recovered this topology. Only in the Bayesian approach with CAT-Γ and CAT-Dirichlet models the symphylans were sister group of the diplopods.

To further evaluate the relationships within the myriapods, we included myriapod sequence data from Regier et al. (2010). Gene coverage was poor (89.9% missing data), which is the most likely reason for the lack of resolution of the tree (Fig. 3). Therefore, results gained with these analyses should be considered with caution. Some conclusions can be drawn nevertheless: First, the position of the symphylans, which we also found in most trees deriving from dataset 1 as sistergroup to the other myriapods was consistently supported; second, pauropods and diplopods were sister taxa, thereby supporting monophyletic Dignatha. Third, *C. tasmanianus* (Craterostigmomorpha) and Scolopendromorpha are closely related, supporting monophyletic Phylactometria (Murienne et al., 2010). Fourth, *G. pustulata* (Pentazonia) were sister group of the remaining diplopods, which agrees with a recent phylogenomic analysis of diplopod ESTs (Brewer and Bond, 2013). Fifth, the monophyly of Helminthomorpha (*i.e.*, *A. gigas* [Spirostreptida], *A. magnum* [Callipodida], *N. americanus* [Spirobolida] and *P. angustus* [Polydesmida]) was strongly supported, also agreeing with the results from Brewer and Bond (2013). These taxa also share a derived morphological character (autapomorphy), which consists of the transformation of at least one pair of legs of the seventh trunk segment into a copulatory organ (Ax, 1999).

4.3. The timeline of myriapod evolution

The early fossil record of Myriapoda is notoriously poor (Shear and Edgecombe, 2010). In fact, Silurian diplopods (*Albadesmus almondi*, *Pneumodesmus newmani*, and *Cowiedesmus eroticopodus*) from the Cowie Formation are the oldest unambiguous myriapod fossils (Wilson and Anderson, 2004). They date back to the base of the Ludfordian, ~418.7 mya. *C. eroticopodus* provides evidence that Chilopoda and Diplopoda had already separated in the Silurian, strongly pointing to a much earlier origin of the myriapod subphylum. There is no conclusive fossil evidence for the presence of myriapods in the Ordovician or an earlier period, and attempts to identify the myriapod stem-group have not yet been successful (Edgecombe, 2004). However, the unambiguous presence of fossils from the putative sister taxa (Pancrustacea, Chelicerata) in the Cambrian strongly suggests an origin of Myriapoda already at that time (Shear and Edgecombe, 2010). Our molecular clock calculations are in line with this view, suggesting that Myriapoda split from Pancrustacea ~513 mya and that myriapod classes commenced to diversify ~500 mya. These dates correspond well with other recent molecular clock-based estimates (Brewer and Bond, 2013; Rehm et al., 2011; Rota-Stabelli et al., 2013; Wheat and Wahlberg, 2013), with the exception that Wheat and Wahlberg (2013) proposed a much earlier origin of the myriapod subphylum 639 mya. The Cambrian origin of Myriapoda is noteworthy because it significantly predates the emergence of land plants in the Middle Ordovician (Rubinstein et al., 2010; Steemans et al., 2009) and the first myriapod fossil from the Silurian (Wilson and Anderson, 2004). Thus we may speculate that the early evolution and divergence of Myriapoda took place in the ocean, and terrestrialization occurred several times independently.

Acknowledgments

We acknowledge Michael Kube and Richard Reinhardt (MPIMG, Berlin, Germany) for extraction of RNA, generating cDNA libraries and ESTs. We thank Arndt v. Haeseler, Ingo Ebersberger, Sascha Strauss, and Peter Schmitzberger for EST data processing and assembly, and help with HaMStR, Nikola Szucsich, Guenther Pass, Daniela Bartel, Manfred Walzl (all University of Vienna), Johannes Dambach (ZFMK Bonn), Michael Raupach (Wilhelmshafen) and Falko Roeding (University of Hamburg) for help with species collections, and A. Donath (ZFMK Bonn) for submitting the Data to NCBI. We acknowledge Emiliano Dell'Ampio (University of Vienna) for his help with contamination check, Malte Petersen (ZFMK ZMB Bonn) for providing Perl scripts and Monika Hänel (University of Hamburg) for her help with the drawings. This work has been supported by Grants of the Deutsche Forschungsgemeinschaft within the Priority Project 1174 "Deep Metazoan Phylogeny" (Bu 956/8; Mi 649/6).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2014.04.007.

References

Aguinaldo, A.M.A., Turbeville, J.M., Linford, L.S., Rivera, M.C., Garey, J.R., Raff, R.A., Lake, J.A., 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 387, 489–493.

Ax, P., 1999. Das system der Metazoa II. Ein Lehrbuch der phylogenetischen Systematik.

Benton, M.J., Donoghue, P.C., 2007. Paleontological evidence to date the tree of life. Mol. Biol. Evol. 24, 26–53.

Bitsch, C., Bitsch, J., 2004. Phylogenetic relationships of basal hexapods among the mandibulate arthropods: a cladistic analysis based on comparative morphological characters. Zool. Scr. 33, 511–550.

Boore, J.L., Collins, T.M., Stanton, D., Daehler, L.L., Brown, W.M., 1995. Deducing the pattern of arthropod phylogeny from mitochondrial DNA rearrangements. Nature 376, 163–165.

Boore, J.L., Lavrov, D.V., Brown, W.M., 1998. Gene translocation links insects and crustaceans. Nature 392, 667–668.

Bourlat, S.J., Juliusdottir, T., Lowe, C.J., Freeman, R., Aronowicz, J., Kirschner, M., Lander, E.S., Thorndyke, M., Nakano, H., Kohn, A.B., Heyland, A., Moroz, L.L., Copley, R.R., Telford, M.J., 2006. Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. Nature 444, 85–88.

Brewer, M.S., Bond, J.E., 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.

Brusca, R.C., Brusca, G.J., 2003. Invertebrates. Sunderland.

Cook, C.E., Smith, M.L., Telford, M.J., Bastianello, A., Akam, M., 2001. *Hox* genes and the phylogeny of the arthropods. Curr. Biol. 11, 759–763.

Cox, J.C., Ingersoll, J.E., Ross, S.A., 1985. A theory of the term structure of interest rates. Econometrica 53, 385–407.

de Rosa, R., Grenier, J.K., Andreeva, T., Cookk, C.E., Adoutte, A., Akamk, M., Carroll, S.B., Balavoinek, G., 1999. Hox genes in brachiopods and priapulids and protostome evolution. Nature 399, 772–776.

Dohle, W., 1980. Sind die Myriapoden eine monophyletische Gruppe? Eine Diskussion der Verwandschaftsbeziehungen der Antennaten. Abh. naturwiss. Ver. Hamburg 23, 45–104.

Dong, Y., Sun, H., Guo, H., Pan, D., Qian, C., Hao, S., Zhou, K., 2012. The complete mitochondrial genome of *Pauropus longiramus* (Myriapoda: Pauropoda): implications on early diversification of the myriapods revealed from comparative analysis. Gene 505, 57–65.

Drummond, A.J., Ho, S.Y.W., Phillips, M.J., Rambaut, A., 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4, 699–710.

Dunn, C.W., Hejnol, A., Matus, D.Q., Pang, K., Browne, W.E., Smith, S.A., Seaver, E., Rouse, G.W., Obst, M., Edgecombe, G.D., Sorensen, M.V., Haddock, S.H., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R.M., Wheeler, W.C., Martindale, M.Q., Giribet, G., 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452, 745–749.

Ebersberger, I., Strauss, S., Haeseler, A., 2009. HaMStR: profile hidden markov model based search for orthologs in ESTs. BMC Evol. Biol. 9, 157.

Edgecombe, G.D., 2004. Morphological data, extant Myriapoda, and the myriapod stem-group. Contrib. Zool. 73, 207–252.

Edgecombe, G.D., 2011. Phylogenetic relationships of Myriapoda. In: Minelli, A. (Ed.), Treatise on Zoology – Anatomy, Taxonomy, Biology. Myriapoda, vol. 1. Brill, Leiden, The Netherlands, pp. 1–20.

Ertas, B., von Reumont, B., Wägele, J.W., Misof, B., Burmester, T., 2009. Hemocyanin suggests a close relationship of Remipedia and Hexapoda. Mol. Biol. Evol. 26, 2711–2718.

Friedrich, M., Tautz, D., 1995. Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods. Nature 376, 165–167.

Gai, Y.-H., Song, D.-X., Sun, H.-Y., Zhou, K.-Y., 2006. Myriapod monophyly and relationships among myriapod classes based on nearly complete 28S and 18S rDNA sequences. Zool. Sci. 23, 1101–1108.

Gai, Y.-H., Song, D.-X., Sun, H.-Y., Yang, Q., Zhou, K.-Y., 2008. The complete mitochondrial genome of *Symphylella* sp. (Myriapoda: Symphyla): extensive gene order rearrangement and evidence in favor of Progoneata. Mol. Phylogenet. Evol. 49, 574–585.

García-Machado, E., Pempera, M., Dennebouy, N., Oliva-Suarez, M., Mounolou, J.C., Monnerot, M., 1999. Mitochondrial genes collectively suggest the paraphyly of Crustacea with respect to Insecta. J. Mol. Evol. 49, 142–149.

Giribet, G., Edgecombe, G.D., 2012. Reevaluating the arthropod tree of life. Annu. Rev. Entomol. 57, 167–186.

Giribet, G., Ribera, C., 2000. A review of arthropod phylogeny: new data based on ribosomal DNA sequences and direct character optimization. Cladistics 16, 204–231.

Giribet, G., Edgecombe, G.D., Wheeler, W.C., 2001. Arthropod phylogeny based on eight molecular loci and morphology. Nature 413, 157–161.

Harvey, T.H.P., Butterfield, N.J., 2008. Sophisticated particle-feeding in a large early Cambrian crustacean. Nature 452, 868–871.

Harvey, T.H., Velez, M.I., Butterfield, N.J., 2012. Exceptionally preserved crustaceans from western Canada reveal a cryptic Cambrian radiation. Proc. Natl. Acad. Sci. USA 109, 1589–1594.

Harzsch, S., Müller, C.H.G., Wolf, H., 2005. From variable to constant cell numbers: cellular characteristics of the arthropod nervous system argue against a sister-group relationship of Chelicerata and "Myriapoda" but favour the Mandibulata concept. Dev. Genes Evol. 215, 53–68.

Heymonds, R., 1901. Die Entwicklungsgeschichte derr Scolopender. Zool. Stud. 33, 1–244.

Holland, B., Moulton, V., 2003. Consensus networks: a method for visualising incompatibilities in collections of trees. In: Benson, G., Page, R. (Eds.), Algorithms in Bioinformatics. Springer-Verlag, Berlin, Germany.

Holmes, D.S., Bonner, J., 1973. Preparation, molecular weight, base composition, and secondary structure of giant nuclear ribonucleic acid. Biochemistry 12, 2330–2338.

Huson, D.H., Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23, 254–267.

Hwang, U.W., Friedrich, M., Tautz, D., Park, C.J., Kim, W., 2001. Mitochondrial protein phylogeny joins myriapods with chelicerates. Nature 413, 154–157.

Kadner, D., Stollewerk, A., 2004. Neurogenesis in the chilopod *Lithobius forficatus* suggests more similarities to chelicerates than to insects. Dev. Genes Evol. 214, 367–379.

Katoh, K., Toh, H., 2008. Recent developments in the MAFFT multiple sequence alignment program. Brief. Bioinform. 9, 286–298.

Keane, T.M., Creevey, C.J., Pentony, M.M., Naughton, T.J., McLnerney, J.O., 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evol. Biol. 6, 29.

Kraus, O., 1998. Phylogenetic relationships between higher taxa of tracheate arthropods. In: Fortey, R.A., Thomas, R.H. (Eds.), Arthropod Relationships. Chapman & Hall, London, pp. 295–303.

Kück, P., Meusemann, K., 2010. FASconCAT: convenient handling of data matrices. Mol. Phylogenet. Evol. 56, 1115–1118.

Kück, P., Meusemann, K., Dambach, J., Thormann, B., Reumont, B., Wägele, J.W., Misof, B., 2010. Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. Front. Zool. 7, 10.

Kusche, K., Burmester, T., 2001. Diplopod hemocyanin sequence and the phylogenetic position of the Myriapoda. Mol. Biol. Evol. 18, 1566–1573.

Kusche, K., Ruhberg, H., Burmester, T., 2002. A hemocyanin from the Onychophora and the emergence of respiratory proteins. Proc. Natil. Acad. Sci. USA 99, 10545–10548.

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.

Lartillot, N., Lepage, T., Blanquart, S., 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25, 2286–2288.

Le, S.Q., Gascuel, O., 2008. An improved general amino acid replacement matrix. Mol. Biol. Evol. 25, 1307–1320.

Lepage, T., Lawi, S., Tupper, P., Bryant, D., 2006. Continuous and tractable models for the variation of evolutionary rates. Math. Biosci. 199, 216–233.

Lepage, T., Bryant, D., Philippe, H., Lartillot, N., 2007. A general comparison of relaxed molecular clock models. Mol. Biol. Evol. 24, 2669–2680.

Loesel, R., Nässel, D.R., Strausfeld, N.J., 2002. Common design in a unique midline neuropil in the brains of arthropods. Arthropod. Struct. Dev. 31, 77–91.

Mallatt, J., Giribet, G., 2006. Further use of nearly complete, 28S and 18S rRNA genes to classify Ecdysozoa: 37 more arthropods and a kinorhynch. Mol. Phylogenet. Evol. 40, 772–794.

Mallatt, J.M., Garey, J.R., Shultz, J.W., 2004. Ecdysozoan phylogeny and Bayesian inference. first use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin. Mol. Phylogenet. Evol. 31, 178–191.

Meusemann, K., Reumont, B.M., Simon, S., Roeding, F., Strauss, S., Kück, P., Ebersberger, I., Walzl, M., Pass, G., Breuers, S., Achter, V., Haeseler, A., Burmester, T., Hadrys, H., Wägele, J.W., Misof, B., 2010. A phylogenomic approach to resolve the arthropod tree of life. Mol. Biol. Evol. 27, 2451–2464.

Misof, B., Misof, K., 2009. A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. Syst. Biol. 58, 21–34.

Misof, B., Meyer, B., von Reumont, B.M., Kuck, P., Misof, K., Meusemann, K., 2013. Selecting informative subsets of sparse supermatrices increases the chance to find correct trees. BMC Bioinformatics 14, 348.

Murienne, J., Edgecombe, G.D., Giribet, G., 2010. Including secondary structure, fossils and molecular dating in the centipede tree of life. Mol. Phylogenet. Evol. 57, 301–313.

Nardi, F., Spinsanti, G., Boore, J.L., Carapelli, A., Dallai, R., Frati, F., 2003. Hexapod origins: monophyletic or paraphyletic? Science 299, 1887–1889.

Negrisolo, E., Minelli, A., Valle, G., 2004. The mitochondrial genome of the house centipede scutigera and the monophyly versus paraphyly of myriapods. Mol. Biol. Evol. 21, 770–780.

Philippe, H., Telford, M.J., 2006. Large-scale sequencing and the new animal phylogeny. Trends Ecol. Evol. 21, 614–620.

Philippe, H., Snell, E.A., Bapteste, E., Lopez, P., Holland, P.W.H., Casane, D., 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol. Biol. Evol. 21, 1740–1752.

Pisani, D., Poling, L.L., Lyons-Weiler, M., Hedges, S.B., 2004. The colonization of land by animals: molecular phylogeny and divergence times among arthropods. BMC Biol. 2, 1.

Pocock, R.I., 1893. On the classification of the tracheate Arthropoda. Zool. Anz. 16, 271–275.

Podsiadlowski, L., Kohlhagen, H., Koch, M., 2007. The complete mitochondrial genome of *Scutigerella causeyae* (Myriapoda: Symphyla) and the phylogenetic position of Symphyla. Mol. Phylogenet. Evol. 45, 251–256.

Regier, J.C., Shultz, J.W., Kambic, R.E., 2004. Phylogeny of basal hexapod lineages and estimates of divergence times. Ann. Entomol. Soc. Am. 97, 411–419.

Regier, J.C., Wilson, H.M., Shultz, J.W., 2005. Phylogenetic analysis of Myriapoda using three nuclear protein-coding genes. Mol. Phylogenet. Evol. 34, 147–158.

Regier, J.C., Shultz, J.W., Ganley, A.R., Hussey, A., Shi, D., Ball, B., Zwick, A., Stajich, J.E., Cummings, M.P., Martin, J.W., Cunningham, C.W., 2008. Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. Syst. Biol. 57, 920–938.

Regier, J.C., Shultz, J.W., Zwick, A., Hussey, A., Ball, B., Wetzer, R., Martin, J.W., Cunningham, C.W., 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature 463, 1079–1083.

Rehm, P., Borner, J., Meusemann, K., von Reumont, B.M., 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.

Richter, S., 2002. The Tetraconata concept: hexapod-crustacean relationships and the phylogeny of Crustacea. Org. Divers. Evol. 2, 217–237.

Roeding, F., Hagner-Holler, S., Ruhberg, H., Ebersberger, I., Haeseler, A., Kube, M., Reinhardt, R., Burmester, T., 2007. EST sequencing of Onychophora and phylogenomic analysis of Metazoa. Mol. Phylogenet. Evol. 45, 942–951.

Rota-Stabelli, O., Telford, M.J., 2008. A multi criterion approach for the selection of optimal outgroups in phylogeny: recovering some support for Mandibulata over Myriochelata using mitogenomics. Mol. Phylogenet. Evol. 48, 103–111.

Rota-Stabelli, O., Campbell, L., Brinkmann, H., Edgecombe, G.D., Longhorn, S.J., Peterson, K.J., Pisani, D., Philippe, H., Telford, M.J., 2011. A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. Proc. R. Soc. Lond. B. Biol. Sci. 278, 298–306.

Rota-Stabelli, O., Daley, A.C., Pisani, D., 2013. Molecular timetrees reveal a cambrian colonization of land and a new scenario for ecdysozoan evolution. Curr. Biol. 23, 392–398.

Rubinstein, C.V., Gerrienne, P., de la Puente, G.S., Astini, R.A., Steemans, P., 2010. Early Middle Ordovician evidence for land plants in Argentina (eastern Gondwana). New Phytol. 188, 365–369.

Ruiz-Trillo, I., Paps, J., Loukota, M., Ribera, C., Jondelius, U., Bagun, J., Riutort, M., 2002. A phylogenetic analysis of myosin heavy chain type II sequences corroborates that Acoela and Nemertodermatida are basal bilaterians. Proc. Natl. Acad. Sci. USA 99, 11246–11251.

Shear, W.A., Edgecombe, G.D., 2010. The geological record and phylogeny of the Myriapoda. Arthropod Struct. Dev. 39, 174–190.

Stamatakis, A., 2006. RAxML-VI-HPC: Maximum Likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690.

Steemans, P., Herisse, A.L., Melvin, J., Miller, M.A., Paris, F., Verniers, J., Wellman, C.H., 2009. Origin and radiation of the earliest vascular land plants. Science 324, 353.

Telford, M.J., 2007. Phylogenomics. Curr. Biol. 17, R945–R946.

Telford, M.J., Bourlat, S.J., Economou, A., Papillon, D., Rota-Stabelli, O., 2008. The evolution of the Ecdysozoa. Philos. Trans. R. Soc. Lond. B Biol. Sci. 363, 1529–1537.

Thorne, J.L., Kishino, H., Painter, I.S., 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15, 1647–1657.

Ungerer, P., Scholtz, G., 2008. The ancestral cleavage pattern of arthropods was not spiral-early development of sea spiders (Arthropoda, Pycnogonidal). J. Morphol. 269, pp. 1469–1469.

von Reumont, B.M., Jenner, R.A., Wills, M.A., Dell'ampio, E., Pass, G., Ebersberger, I., Meyer, B., Koenemann, S., Iliffe, T.M., Stamatakis, A., Niehuis, O., Meusemann, K., Misof, B., 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.

Waloszek, D., Dunlop, J.A., 2002. A larval sea spider (Arthropoda: Pycnogonida) from the Upper Cambrian 'Orsten' of Sweden, and the phylogenetic position of pycnogonids. Paleontology 45, 421–446.

Wheat, C.W., 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.

Wilson, H.M., Anderson, L.I., 2004. Morphology and taxonomy of Paleozoic millipedes (Diplopoda: Chilognatha: Archipolypoda) from Scotland. J. Paleontol. 78, 169–184.

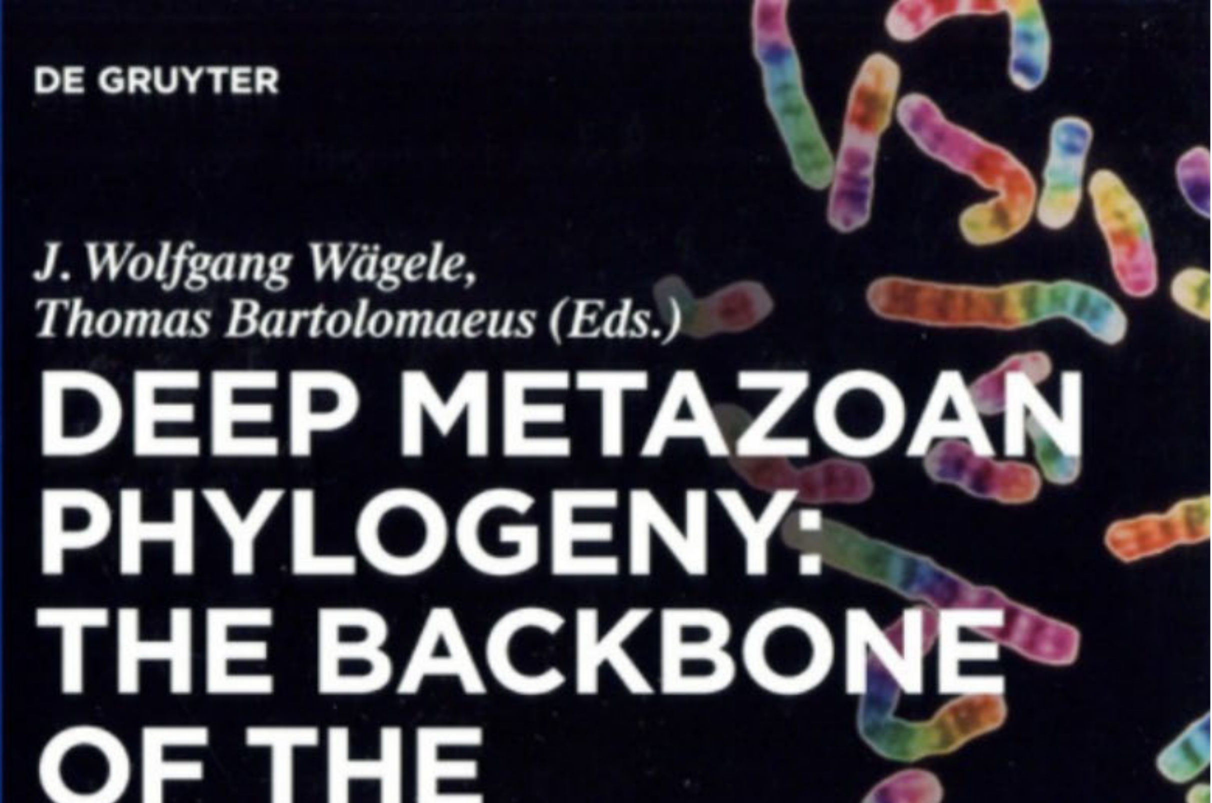
Wilson, K., Cahill, V., Ballment, E., Benzie, J., 2000. The complete sequence of the mitochondrial genome of the crustacean *Penaeus monodon*: Are malacostracan crustaceans more closely related to insects than to branchiopods? Mol. Biol. Evol. 17, 863–874.

Zhang, X.G., Siveter, D.J., Waloszek, D., Maas, A., 2007. An epipodite-bearing crown-group crustacean from the Lower Cambrian. Nature 449, 595–598.

Zrzavý, J., Štys, P., 1997. The basic body plan of arthropods: insights from evolutionary morphology and developmental biology. J. Evol. Biol. 10, 353–367.

Zwick, A., Regier, J.C., Zwickl, D.J., 2012. Resolving discrepancy between nucleotides and amino acids in deep-level arthropod phylogenomics: differentiating serine codons in 21-amino-acid models. PLoS ONE 7, e47450.

DE GRUYTER



J. Wolfgang Wägele,
Thomas Bartolomaeus (Eds.)

DEEP METAZOAN PHYLOGENY: THE BACKBONE OF THE TREE OF LIFE

NEW INSIGHTS FROM ANALYSES OF MOLECULES, MORPHOLOGY, AND THEORY OF DATA ANALYSIS

DE G

Jason Dunlop, Janus Borner, and Thorsten Burmester

16 Phylogeny of the Chelicerates: Morphological and molecular evidence

Abstract: The arthropod subphylum Chelicerata encompasses ~110,000 described living species. Here we review the present state of knowledge on chelicerate phylogeny, thereby including morphological, paleontological and molecular evidence. We must conclude that chelicerate still are largely unresolved. The minimal consensus tree supported by various methods is (Pycnogonida, (Xiphosura, (Scorpiones, ((Amblypygi, (Thelyphonida, Schizomida)), Araneae)))). In addition, the Acari – combining Acariformes and Parasitiformes – are probably a valid taxon. The positions of Opiliones, Palpigradi, Pseudoscorpiones, Ricinulei, and Solifugae are, however, essentially unresolved. Even a novel large multi-gene data set deriving from expressed sequence tags from various chelicerate taxa did not help improving the tree. The lack of phylogenetic signal may be explained by rapid adaptive radiation associated with the terrestrialization of the arachnids.

16.1 Introduction

Chelicerata is one of the major branches of the arthropods, encompassing arachnids and their relatives. With nearly 110,000 described living species, they are second only to insects in diversity. Arachnids obviously dominate, with spiders (more than 43,000 species) and mites (more than 55,000) representing the ‘megadiverse’ orders. Spiders are ubiquitous, and a key group of predators in most terrestrial ecosystems. Mites are ecologically more diverse and range from free-living predators to parasites, to detritivores and even herbivores. Some are aquatic and others are of economic significance as crop pests or disease vectors. Further species-rich arachnid groups include the harvestmen (ca. 6,500 species), pseudoscorpions (3,400), scorpions (2,000), and camel spiders (1,100). The remaining arachnids are largely restricted to the tropics and are known from, at most, only a few hundred species. In addition to the arachnids, Chelicerata also includes the marine xiphosurans (horseshoe crabs) known only from four living species. Finally pycnogonids (sea spiders) are another marine group, comprising nearly 1,500 species.

Relationships between these animals have long been debated (e.g. Pocock, 1893; Börner, 1904). A significant step forward was the introduction of cladistic methods in a seminal paper by Weygoldt and Paulus (1979), followed by the first computer-assisted cladograms (Shultz, 1990) and the first applications of molecular data (Wheeler and Hayashi, 1998) (Figure 16.1). In recent years, additional molecular phylogenetic studies using a variety of markers and more sophisticated techniques have attempted to enhance our understanding of chelicerate relationships (Shultz and

Regier, 2000; Hassanin, 2006; Podsiadlowski and Braband, 2006; Jones et al., 2007; Shultz, 2007; Dunn et al., 2008; Jeyaprakash and Hoy, 2009; Roeding et al., 2009; Pepato et al., 2010; Regier et al., 2010; Rehm et al., 2011; Rota-Stabelli et al., 2011; Rehm et al., 2012). Yet despite this wealth of modern data there is, at present, no single accepted phylogeny for Chelicerata and discrepancies remain between trees derived from morphology and molecules. This striking lack of resolution – in spite of the range of analytical techniques applied – is surely the key issue for contemporary work on chelicerate relationships.

16.2 Chelicerate origins: Mandibulata or Myriochelata?

Traditionally, chelicerates were placed within the Euarthopoda as the sister-group of the Mandibulata (Brusca and Brusca, 2003), a taxon that comprises the subphyla Myriapoda, Crustacea, and Hexapoda. Synapomorphies that unite the Mandibulata are the mandibles, which are mouthparts used for biting and chewing, the possession of antennae and the division of the body into three tagmata: head, thorax and abdomen. The monophyly of Mandibulata receives support from some molecular phylogenetic studies, which include analyses of selected single genes (Giribet and Ribera, 2000; Kusche et al., 2003), concatenated alignments (Regier et al., 2010), total evidence (Giribet et al., 2001), and mircoRNA (Rota-Stabelli et al., 2011).

However, other molecular studies have challenged the monophyly of Mandibulata and suggested a common clade of chelicerates and myriapods (the “Myriochelata” or “Paradoxopoda” hypothesis; e.g., Friedrich and Tautz, 1995; Hwang et al., 2001; Mallatt et al., 2004; Pisani et al., 2004; Dunn et al., 2008). Morphological evidence in favor of Myriochelata is poor, and essentially restricted to some similarities of neurogenesis (Kadner and Stollewerk, 2004) and embryonic development (Mayer and Whitington, 2009). These characters, however, may also be plesiomorphic and reflect the ancient state in stem-line arthropods. An analogous explanation may explain similarities in the organization of the onychophoran and chelicerate brain, which even led to the proposal of a sister-group relationship between these two taxa (Strausfeld et al., 2006). According to Rota-Stabelli and Telford (2008), the choice of outgroup in analyses of mitochondrial sequences influences whether Myriochelata is recovered.

16.2.1 Evidence from the fossil record of chelicerates

Chelicerate origins in deep time are uncertain. Historically they were often linked to the extinct Trilobita; a hypothesis largely based on superficial similarities between trilobites and horseshoe crabs (Xiphosura). Newly hatched xiphosurans retain vestiges of segmentation and are still commonly referred to as ‘trilobite larvae’. The names

Arachnomorpha or Arachnata can often be found in the paleontological literature and encompass a broad, but not necessarily well-defined, group of chelicerates, trilobites and various early Paleozoic arthropods. The concept of Arachnomorpha was explicitly challenged by Scholtz and Edgecombe (2005), who pointed out that most of the potential synapomorphies shared by trilobites and xiphosurans – like the broad head shield – are not seen in pycnogonids or arachnids. They preferred to group trilobites with other antennae-bearing arthropods, i.e. the Mandibulata.

Where does this leave Chelicerata? One emerging hypothesis (e.g., Chen et al., 2004) places a number of Cambrian fossils – sometimes called 'great appendage' arthropods or megacherians – on the chelicerate stem-lineage. These extinct arthropods share with chelicerates a modification of the first pair of head appendages into increasingly raptorial and claw-like structures; see also Haug et al. (2012). In this scenario, these raptorial head limbs would represent precursors of the chelate chelicerae (see below). Yet claws are functional adaptations, easy to evolve, and thus potentially prone to homoplasy. Homologizing head appendages between early fossil arthropods and their living relatives remains controversial and other interpretations of the 'great appendage' (e.g., Budd, 2002) have been published. Resolving the chelicerate stem-lineage is nevertheless important with a view towards selecting appropriate outgroup taxa for polarizing morphological character states.

16.3 Chelicerate phylogeny

The name Chelicerata literally means 'claw-bearer' and was coined by Heymons (1901) for arthropods in which the first pair of head appendages – the chelicerae – are used for grasping and/or tearing up prey. These chelicerae are either claw-like or else modified, in groups like spiders into fangs which take the form of a pocket knife. In its original conception, Chelicerata comprised arachnids and xiphosurans only.

16.3.1 Position of the sea spiders (Pycnogonida)

Sea spiders (Pycnogonida) were often added to the chelicerates, although this placement does not have universal support. It is fair to say that pycnogonids are strange-looking creatures with many unusual features. For example, they have a large proboscis for sucking up prey and displace many organ systems from their remarkably narrow body into the legs. Some authors simply accumulated these autapomorphies as evidence for the 'uniqueness' of pycnogonids, and suggested that they evolved independently from the other arthropods (for a review, see Dunlop and Arango, 2005). The name Cormogonida (Zrzavý et al., 1998) has been suggested for a clade of arthropods excluding the pycnogonids. In support of this hypothesis is the fact that in cormogonid arthropods (i.e. euchelicerates plus mandibulates) the genital opening is on the body, while in pycnogonids the genital openings are on the proximal articles of the legs. Additional support for the Cormogonida hypothesis comes from the combined analysis of molecular sequences and morphological characters (Giribet et al., 2001) and from comparative neuroanatomy, which suggests a unique innervation of the chelifores (Maxmen et al., 2005) (but see below).

Morphologically, the key issue for a sister-group relationship of Pycnogonida and Euchelicerata is equating the pycnogonid chelifores (or cheliphores) with the euchelicerate chelicerae. Both structures are fundamentally chelate, but as noted above claws are functional elements which could potentially evolve in parallel. As mentioned, there have also been proposals that the chelifores and chelicerae are not serially homologous elements, i.e. they are not innervated from the same part of the brain (Maxmen et al., 2005). However, Hox gene (Jager et al., 2006) and neuroanatomical studies (Brenneis et al., 2008) have argued that this is the same appendage – and in this context chelate chelifores/chelicerae remain the best autapomorphy of a monophyletic Chelicerata.

Most of the modern literature favors the traditional textbook concept of Chelicerata, namely a sister-group relationship of Pycnogonida and Euchelicerata. Molecular phylogenetic studies using hemocyanin sequence and structure (Rehm et al., 2012), selected genes (Regier et al., 2010; Sanders and Lee, 2010) or expressed sequence tags (ESTs) (Dunn et al., 2008; Meusemann et al., 2010) strongly support this topology. An ingroup position of the Pycnogonida with the Arachnida – as deduced from mitochondrial DNA sequences (Hassanin, 2006; Podsiadlowski and Braband, 2006; Jones et al., 2007; Park et al., 2007; Jeyaprakash and Hoy, 2009) – is not supported by other data and is best considered an artifact.

16.3.2 Euchelicerata

The term Euchelicerata was introduced by Weygoldt and Paulus (1979) for all chelicerates, excluding pycnogonids. In their original definition it included an extinct fossil group called Aglaspidida, but these animals – which superficially resemble xiphosurans – are now usually placed on the mandibulate stem-lineage instead; see Ortega Hernández et al. (2013; and references therein). Euchelicerata is an uncontroversial group, easily defined by the presence of plate-like appendages on the underside of the opisthosoma. These are clearly visible in xiphosurans as the movable, flap-like opercula covering the paired genital openings and the five subsequent pairs of gills. The presence of lamellate respiratory organs (i.e. book gills or book lungs) is also characteristic for euchelicerates; although numerous arachnids have lost the lungs and replaced them either partially or wholly with trachea. In general, it can also be argued that Euchelicerata have the typical prosoma–opisthosoma division of the body; the front half focused on feeding, locomotion and sensory systems and the back half focused on digestion, respiration and reproduction. A caveat to this is that in xiphosurans the dividing line is less clear cut. Comparative morphology and embryology reveals that parts of the anterior dorsal opisthosoma contribute to the prosomal dorsal shield, while the first opisthosmal appendages (the small, stubby chilaria) are functionally integrated into the ventral prosoma (Shultz, 2001). The textbook prosoma–opisthosoma division is also questionable for acariform mites, in which the principal division of the body actually runs between the second and third pair of legs (Dunlop et al., 2012), rather than behind the fourth pair as in, e.g., spiders.

Within euchelicerates, two classes were traditionally recognized. Merostomata include xiphosurans and the extinct Paleozoic Eurypterida (sea scorpions). The other class is the (largely) terrestrial Arachnida. As critiqued by Kraus (1976), this is primarily an ecological division rather than an explicitly phylogenetic one. Note that Lamsdell (2013) recently challenged the monophyly of Xiphosura, arguing that certain Paleozoic fossils traditionally placed here are not stem-group horseshoe crabs, but represent stem-group euchelicerates instead.

Weygoldt and Paulus (1979) placed eurypterids closer to arachnids in a clade they called Metastomata. This was ostensibly defined by a plate-like metastoma immediately behind the leg coxae. This metastoma is a fundamental part of the eurypterid body plan, but it could represent the modified first opisthosomal appendages; thus making them potentially homologous to the chilaria of xiphosurans. A direct equivalent in Arachnida is harder to demonstrate, and in general arachnids are thought to lack appendages on the first opisthosomal segment (Shultz, 1990). As an alternative diagnostic character, in exceptionally preserved eurypterids, Kamenz et al. (2011) identified precursor sclerites within internal parts of the genitalia which go on, eventually, to form spermatophores. Since xiphosurans simply release their sperm during mating, spermatophore-mediated sperm transfer offers a good potential synapomorphy for (Eurypterida + Arachnida); a clade for which the new name Sclerophorata was proposed.

16.4 Arachnids: Conquerors of the land

The name Arachnida can be traced back to Lamarck (1801) and emerged, historically, from among the ‘wingless insects’ as recognized by eighteenth century zoologists. Current schemes usually accept sixteen arachnid orders: twelve extant (Acariformes [mites], Amblypygi [whip spiders], Araneae [true spiders], Opiliones [harvestmen], Palpigradi [micro-whip scorpions], Parasitiformes [predatory mites and ticks], Pseudoscorpiones [false scorpions], Ricinulei [hooded tickspiders], Schizomida [schizomids], Scorpiones [scorpions], Solifugae [sun spiders], and Thelyphonida [whip scorpions]), and four extinct (Haptopoda, Phalangiotarbida, Trigonotarbida, Uraraneida). Cladistic studies employing morphological characters have almost invariably recovered Arachnida as a monophyletic group (Shultz, 1990; Wheeler and Hayashi, 1998) (Figure 16.1). Putative synapomorphies supporting Arachnida include a reduction in the width of the prosomal dorsal shield – usually just called the carapace in the taxonomic literature – and the possession of "terrestrial" respiratory organs like book lungs or trachea. Other proposed arachnid characters include the presence of tiny 'strain gauges' in the cuticle called slit sense organs, although these are not seen in all arachnid orders. Another potential synapomorphy is the loss of appendages from the first opisthosomal segment (see above). This is problematic as scorpions retain limb buds here, at least during their embryological development.

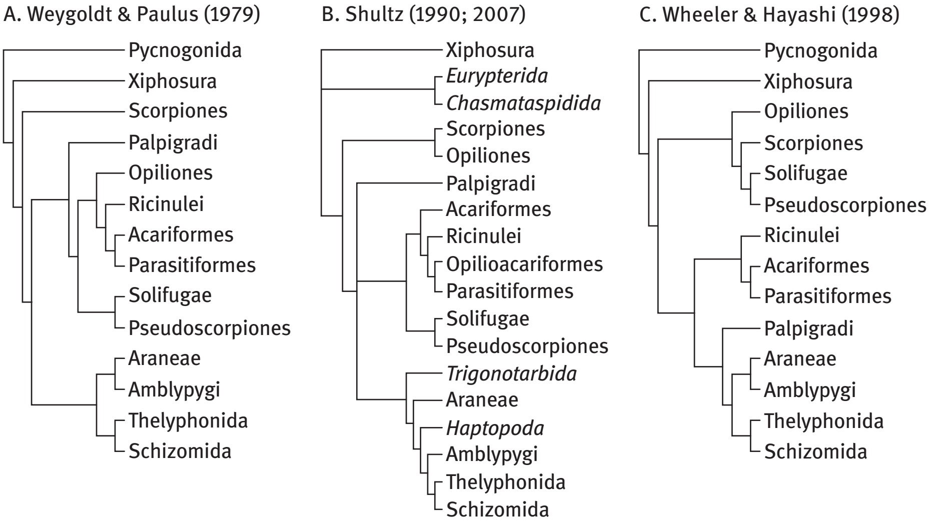


Figure 16.1: Chelicerate phylogeny by **A.** Weygoldt and Paulus (1979), **B.** Shultz (1990; 2007) and **C.** Wheeler and Hayashi (1998). The names in italics in (B) denote extinct taxa.

16.4.1 Are arachnids monophyletic?

Some paleontological studies have challenged arachnid monophyly, specifically with respect to scorpions and eurypterids (reviewed in Dunlop and Braddy, 2001). Both express a similar body plan in which the last five segments of the opisthosoma are somewhat narrow and ring-like. The question is whether this is synapomorphic or homoplastic? Although it is tempting to envisage scorpions evolving directly from 'sea scorpions', it has to be conceded that the most scorpion-like eurypterids – the mixopteroids, complete with a telson shaped like a sting – actually resolve in quite a derived position among the eurypterids (e.g. Tetlie, 2007).

Several molecular phylogenetic studies also failed to recover arachnids as a monophyletic clade (Roeding et al., 2009; Meusemann et al., 2010; Sanders and Lee, 2010). Sanders and Lee (2010) used concatenated sequences of 18S rRNA, 28S rRNA, elongation factors 1α and 2, and RNA polymerase II subunit in Bayesian analyses and obtained Xiphosura nested within the arachnids. Phylogenomic studies employing large multi-gene alignments of 11,168 (Roeding et al., 2009) or 37,476 (Meusemann et al., 2010) amino acid positions, which had been inferred from EST-derived orthologous genes, consistently obtained the Acari as the sister-group of all other euchelicerates. However, these studies only included representatives of the taxa Xiphosura, Araneae, Scorpiones and Acari.

In a more recent approach, additional taxa, namely Amblypygi, Opiliones, Pseudoscorpiones, Solifugae, and Thelyphonida, were included (Borner, Rehm and Burmester, unpublished). 454 pyrosequencing was used to obtain ESTs, resulting in a concatenated multi-gene alignment consisting of 197 orthologous genes (18,163 aa positions, 32.4 % missing data). However, employing Maximum Likelihood and Bayesian methods for tree reconstructions (similar to the methods described in Meusemann et al., 2010), even this impressive data set was not sufficient to identify the relative relationships among the arachnid taxa, with the exception of support for Euchelicerata, Tetrapulmonata and Pedipalpi (Figure 16.2).

Connected to this debate is the question of whether arachnids all share a common, terrestrial ancestor, or whether multiple lineages moved onto land independently. Terrestrialization was achieved in some (if not all) groups by at least the late Silurian period (Jeram et al., 1990). Note that some Paleozoic fossil scorpions have been interpreted as aquatic animals which, if correct, would imply at least two separate terrestrialization events: one for the scorpions themselves and at least one for the remaining arachnids. However, aquatic scorpions have not been universally accepted (for a recent critique see Kühl et al., 2012) and the trend is now to see most, if not all, fossil scorpions as terrestrial. Scholtz and Kamenz (2006) compared the book lungs of scorpions with those of tetrapulmonate arachnids, identifying numerous common features in their fine structure which imply a single origin for the book lungs in a terrestrial arachnid ancestor.

16.4.2 Tangled relationships: The arachnid groups

The monophyly of each of the twelve arachnid orders is undisputed. However, the relationships between them remain controversial. Traditionally, scorpions were often envisaged as ‘primitive’ arachnids – perhaps again influenced by their resemblance to eurypterids – and placed basal to the remaining Arachnida, a clade for which Pocock’s (1893) name Lipoctena has been adopted. In Weygoldt and Paulus (1979), lipoctenid arachnids were characterized by, for example, the fine structure of both their sperm cells and the lateral eyes, and a reduction in the number of respiratory openings. Alternatively, Shultz (2007, and references therein) developed a novel hypothesis, largely based on skeletomuscular characters, in which scorpions are the sister-group of harvestmen (Opiliones). The name Stomothecata was proposed for this

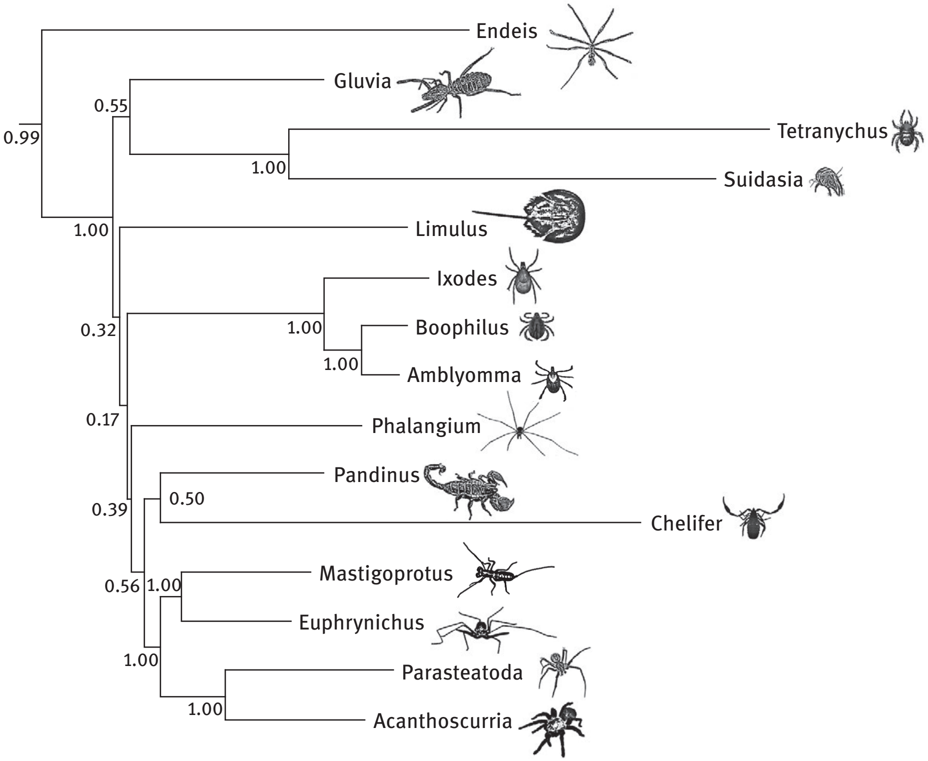


Figure 16.2: Bayesian phylogenetic analysis of 15 chelicerate taxa based on a multi-gene alignment selected from EST data sets (Borner, Rehm and Burmester, unpublished). 197 orthologous genes (88,044 amino acid positions) were selected by the aid of HaMStR (Ebersberger et al., 2009) and aligned with MAFFT (Katoh and Toh, 2008). Poorly aligned sections were removed by Gblocks employing the "less stringent" criteria (Castresana, 2000; http://molevol.cmima.csic.es/castresana/Gblocks_server.html). The final alignment covered 18,163 positions with 32.4 % missing data. Tree reconstructions were carried out with PhyloBayes3.3e (Lartillot et al., 2008) assuming the CAT mixture model with discrete gamma model (four categories) and the Dirichlet process. 20 chains were run for 20,000 cycles each. The numbers at the nodes represent the Bayesian posterior probabilities.

clade (Shultz, 2007), and refers to the shared presence of a distinct preoral chamber – the stomotheca – which in scorpions and harvestmen is explicitly constructed from projections (apophyses) derived from the first two pairs of leg coxae. A critique of this hypothesis would be that some early fossil scorpions have a much simpler coxosternal arrangement and appear to lack these so-called coxapophyses. With regard to harvestman feeding ecology, a further interesting point to note is that they have an omnivorous diet and are one of the few arachnid orders which are (still) capable of ingesting solid food. This is presumably the plesiomorphic condition, and is seen in xiphosurans (as an outgroup) and also some mites (Walter and Proctor, 1998). Most arachnids liquefy their food preorally.

Börner (1904) introduced the name Haplocnemata for a group comprising camel spiders (Solifugae) and pseudoscorpions (Pseudoscorpiones). Apart from their large pedipalpal claws, basal pseudoscorpions do indeed resemble camel spiders. Haplocnemata can be defined on a range of morphological characters, such as two pairs of tracheal openings on the third and fourth opisthosomal segments and a very short femur such that the leg bends principally from the patella–tibia articulation (hence the confusion about the 'missing' patella). In fact, cladistic studies (Weygoldt and Paulus, 1979; Shultz, 1990) based on morphological characters have recovered Haplocnemata, although we should note that male genital characters do not support this relationship (Alberti and Peretti, 2002).

16.4.3 Are Acari monophyletic and arachnids at all?

Mites and ticks were traditionally grouped together as a single order named Acari. Despite this, acarologists have long recognized (e.g., Grandjean 1935) that there are at least two fundamental lineages, which can be differentiated on whether or not the cuticle of their setae shows birefringence; as well as a swathe of other internal and external characters (reviewed by Dunlop and Alberti, 2008). The parasitiform (or anactinotrichid) mites include opilioacarids, holothyrids, mesostigmatids and ticks. The acariform (or actinotrichid) mites include all the rest. The possibility that Acari is not monophyletic has long been discussed; championed primarily in the non-cladistic work of Zachvatkin (1952) and van der Hammen (1989). All mites share one very good synapomorphy: the gnathosoma. This is a specific functional unit at the front of the body encompassing the chelicerae, mouth lips and pedipalps, which can move independently of the rest of the body (or idiosoma). Any scenario in which mites are not monophyletic must contend with the idea that the gnathosoma is homoplastic. For a further discussion of this key character see Alberti et al. (2011). Some authors have speculated that opilioacariform mites are – as their name implies – related to harvestmen, or have compared palpigrades to the acariform mites (van der Hammen, 1989). However, these hypotheses tend to rely on superficial resemblances and are rarely backed by robust sets of apomorphies.

Perhaps the strongest recurring model is that all mites (or at least the Parasitiformes) are related to the rare ricinuleids. This is the Acaromorpha concept. Acari and Ricinulei both uniquely have hexapodal larvae, i.e. the hatching instar has only three pairs of legs and acquires the fourth pair later in ontogeny. Other authors have argued that ricinuleids also have a gnathosoma, or at least the fused pedipalpal coxae which contribute towards a gnathosoma (Shultz, 1990). However, this interpretation is open to question and other authors treat the gnathosoma as a mite feature only. A further challenge to Acaromorpha is the fact that ricinuleids also share a number of putative synapomorphies (reviewed by Dunlop et al., 2009) with the extinct arachnid order Trigonotarbida; a lung-bearing taxon which is evidently close to the Tetrapulmonata.

Alberti and Peretti (2002) identified potential synapomorphies in sperm and testis structure shared between camel spiders (Solifugae) and the acariform branch of the mites only. Both groups also share a so-called sejugal furrow, a sulcus dividing the body between the second and third pair of legs (e.g. Dunlop et al., 2012). Taking up this theme, the molecular study of Dabert et al. (2010) formally recognized (Solifugae + Acariformes), and the same result was found independently in a combined molecular and morphological analysis published shortly afterwards by Pepato et al. (2010); who proposed using the historical name Poecilophysidea for this clade. Where, in this scenario, the pseudoscorpions and the other mite clade (Parasitiformes) belong is uncertain, although Dabert et al. (2010) recovered these two groups together.

Other molecular phylogenetic studies also gave conflicting results on both the monophyly and the position of the Acari, respectively. Wheeler and Hayashi (1998) treated the Acari as monophyletic. By using total evidence, combining rRNA and morphological characters, they recovered Acari as sister-group of Ricinulei within the monophyletic Arachnida. A large body of literature exists in which mitochondrial amino acid or nucleotide sequence data have been employed to resolve the phylogenetic position of Acari. Fahrein et al. (2007) derived a paraphyletic Acari, with Acariformes + Ricinulei being sister-group of the Araneae. Diphyletic Acari were also found by Ovchinnikov and Masta (2012), who recovered the clades Acariformes + Pseudoscorpiones and Parasitiformes + Ricinulei (cf. the Acaromorpha concept, as discussed above). By contrast Park et al. (2007) derived a monophyletic Acari as sister-group of the Araneae. In analyses employing orthologous genes selected from the ESTs, we either found the Acari as sister-group to the other Euchelicerates (Roeding et al., 2007; Roeding et al., 2009; Meusemann et al., 2010) or, upon the inclusion of additional arachnid taxa, Parasitiformes and Acariformes in an unresolved position within the euchelicerates (Figure 16.2) (Borner, Rehm and Burmester, unpublished).

16.4.4 Tetrapulmonata

Perhaps the least controversial higher arachnid taxon in terms of morphology and molecular data is Tetrapulmonata. This group has been consistently recovered – in one form or another – since the early work of Pocock (1893). It essentially comprises the spiders (Araneae) and their closest relatives: namely whip spiders (Amblypygi), whip scorpions (Thelyphonida), schizomids (Schizomida) plus three extinct orders: Trigonotarbida, Uraraneida and Haptopoda. What these arachnids all share in common is a ground pattern of two pairs of book lungs, opening on the second and third opisthosomal segments. They also have similar chelicerae, which are less chelate and shaped instead more like a pocket knife with a fang articulating against a basal region. Another typical feature is a degree of constriction between the prosoma and opisthosoma, although this is also seen in other arachnids such as palpigrades and camel spiders. The name Tetrapulmonata was introduced by Shultz (1990) for spiders, whip spiders, whip scorpions and schizomids (Figure 16.1B). Study of well-preserved trigonotarbid fossils indicates that they belong to this assemblage too (Shear et al., 1987). Tetrapulmonata was thus subsequently expanded to a clade Pantetrapulmonata by Shultz (2007) specifically to encompass this extinct order. The monophyly of extant Tetrapulmonata has been consistently recovered by the molecular phylogenetic studies using various sets of genes (Wheeler and Hayashi, 1998; Shultz and Regier, 2000; Jones et al., 2007; Pepato et al., 2010; Rehm et al., 2012), as well as by a large concatenated multi-gene alignment (Borner, Rehm and Burmester, unpublished).

Within the pantetrapulmonates there remains some debate about the precise phylogenetic relationships of the individual orders. Integrating the recently identified (and probably quite spider-like, Selden et al., 2008) fossil uraranenids into the scheme of Shultz (2007), we would recover a hypothesis along the lines of (Trigonotarbida, ((Uraranenida, Araneae), (Haptopoda, (Amblypygi, (Thelyphonida, Schizomida)))))). There seems little doubt that the similar-looking whip scorpions and schizomids are closely related – together forming the Uropygi in some nomenclature schemes – and historically they were further combined with the whip spiders (Amblypygi) into a single arachnid order: Pedipalpi. The Pedipalpi concept remains a robust hypothesis on morphological grounds (Shear et al., 1987; Shultz, 1990, 2007) and is justified by synapomorphies such as modification of the pedipalps into subchelate limbs for prey capture, and the first pair of legs becoming long and slender such that they are used more like tactile organs. Further skeletomuscular details can be added to this list (Shultz 1999).

The alternative hypothesis, the Labellata concept (Petrunkevitch, 1949; Weygoldt and Paulus, 1979) groups spiders (Araneae) with whip spiders (Amblypygi) and is justified morphologically by the presence of a muscular sucking stomach within the prosoma which aids the ingestion process, and an especially narrow junction between the prosoma and opisthosoma which is often referred to explicitly here as a pedicel or petiolous (Figure 16.1A). Wheeler and Hayashi (1998) recovered Labellata based on the phylogenetic analysis of ribosomal RNA sequences and total evidence, and more recently, Ovchinnikov and Masta (2012) found this topology with mitochondrial amino acid sequences. The unusual topology of a clade comprising Araneae + Uropygi was identified by analyses of 18S + 28S ribosomal RNA (Pepato et al., 2010). Most other molecular analyses, however, recover monophyletic Pedipalpi (Amblypygi and Uropygi) as one of the best-supported taxa within the arachnids. The data sets applied in these studies included two nuclear genes (elongation factor 1-a and the large subunit of RNA polymerase II) (Shultz and Regier, 2000), selected mitochondrial data (Jones et al., 2007), single and concatenated hemocyanin sequences (Rehm et al., 2012), and the above-mentioned multi-gene alignment derived from ESTs (Borner, Rehm and Burmester, unpublished).

16.4.5 Araneae: The true spiders

Spiders are the most prominent representatives of the chelicerates, and are characterized by the possession of opisthosomal silk glands opening via spinnerets (Selden et al., 2008), cheliceral venom glands and male pedipalps modified as a copulatory organ. The monophyly of Araneae has never been disputed among morphologists and has also been consistently recovered in molecular studies (Wheeler and Hayashi, 1998; Roeding et al., 2009; Rehm et al., 2012). Two suborders of spiders are recognized; the Mesothelae are considered as the most basal clade, being sister-group to the Opisthothele. The latter is divided into the infraorders Mygalomorphae (tarantulas and relatives) and Araneomorphae (orb weaving spiders, wandering spiders, and others).

16.5 Dating chelicerate evolution

The fossil record of chelicerates is fragmentary, but still helps to date their evolutionary history (Dunlop and Selden, 2009; Dunlop, 2010). The origin of the subphylum Chelicerata dates back at least to the Cambrian period. The earliest fossil evidence for the split of Pycnogonida and Euchelicerata is a larval sea spider from the upper Cambrian ~ 500 million years ago (MYA) (Waloszek and Dunlop, 2002). A molecular clock approach, which compares the DNA or amino acid sequences, is an additional and alternative approach to date chelicerate origin and evolution. Initial clock analyses suggested that the earliest divergence within the chelicerates took place already 813–632 MYA (Regier et al., 2005). This estimate is most likely too old and predates the first unequivocal evidence for metazoan life. We employed ESTs and calculated that the split between Pycnogonida and Euchelicerata occurred ~ 546 MYA (Rehm et al., 2011). Other calculations employing hemocyanin sequences came to a similar conclusion and suggest a divergence of 543 MYA (Rehm et al., 2012). A similar date was found in the molecular clock study of Rota-Stabelli et al. (2013). These estimates only slightly predate the earliest putative (stem) chelicerate fossils, which derive from the Lower Cambrian ~ 530 MYA (Chen, 2009).

Likewise, the fossil and molecular clock dates of euchelicerate evolution essentially agree. The earliest true euchelicerates are xiphosuran fossils from the Early Ordovician ~ 480 MYA of Morocco (Van Roy et al., 2010). Studies employing hemocyanin sequences derived ~ 463 MYA (444 to 489 MYA) as the time of divergence of arachnids and xiphosurans (Rehm et al., 2012). The first unambiguous arachnid is a ~ 428 MYA old Silurian scorpion (Dunlop and Selden, 2009; Dunlop, 2010); we calculated that scorpions and Tetrapulmonata split ~ 419 MYA (405 to 440 MYA). The Tetrapulmonata likely emerged in the Devonian period, but the first fossils (i.e. recognizable mesothele spiders, whip spiders and whip scorpions) are from the Carboniferous (Dunlop, 2010). An early (i.e. Devonian) origin was supported by the molecular clock calculations (~ 369 MYA). According to the hemocyanin data set Araneae and Pedipalpi diverged ~ 369 MYA (357 to 414 MYA), Amblypygi and Uropygi 334 MYA (316 to 344 MYA). The oldest known non-mesothele spider is a mygalomorph dating 240 MYA (Triassic), while Araneomorphae are also Triassic in age (Selden et al., 1999). From the hemocyanin data set we derived that Mygalomorphae and Araneomorphae diverged ~ 271 MYA (254 to 288 MYA), which is somewhat older and implies a Permian origin of today's two principal spider clades (Rehm et al., 2012).

The most successful subgroup within the Araneomorphae are the Entelegynae. These spiders are characterized by their complex genitalia, and are further subdivided into the Orbicularidae (i.e. orb weavers and their relatives) and the spiders of the RTA clade which includes both web-builders and hunting or ambushing spiders which have secondarily given up their web-building behavior. A recent fossil suggests that the Orbicularidae already occurred in the Jurassic 165 MYA (Selden et al., 2011). Our calculations derived a date ~ 239 MYA for the divergence of Orbicularidae and the RTA clade (Rehm et al., 2012), which suggests that the origin of orb-weaving spiders may even be earlier.

16.6 Perspectives: Resolving the chelicerate tree

Despite more than 100 years of morphological research and the recent advance of molecular techniques, relationships among the chelicerate orders are still largely unresolved. The minimal consensus, which is well supported by various methods, would be Pycnogonida, (Xiphosura, (Scorpiones, ((Amblypygi, (Thelyphonida, Schizomida)), Araneae)))). There is also evidence, albeit disputed, that the Acari – combining the Acariformes and Parasitiformes – are a valid taxon. However, the positions of the remaining arachnid orders remain elusive and different results have emerged from different data sets. Even our very large multi-gene data set that also includes Opiliones, Pseudoscorpiones and Solifugae does not improve the tree. Alternative data sets in which either slow or fast evolving genes, or both, or long branching taxa have been excluded, did not result in any improvement of the tree (not shown). Likewise, alternative methods, such as tree reconstructions using rare genomic changes such as indels failed at the same nodes.

This observation is remarkable, and in sharp contrast to the other arthropod subphyla, in which the generation of large sequence alignments and the application of similar methods led to a significant advance in the resolution of the trees (Roeding et al., 2007; Roeding et al., 2009; Meusemann et al., 2010; Regier et al., 2010; von Reumont et al., 2012). A possible explanation for the lack of phylogenetic signal within the arachnid data sets may be a rapid diversification in the early Paleozoic, which may in turn be evidence of an adaptive radiation associated with the terrestrialization of these animals (see also: Pisani et al., 2005; Rota-Stabelli et al. 2013). Such an interpretation is tentatively supported by the very short internal branches at the relevant nodes (cf. Figure 16.2) and is in line with the notorious difficulties encountered when trying to resolve the chelicerate tree with the help of morphological characters.

Acknowledgments

We thank Peter Rehm for his help with data acquisition, Paul Selden and Gerd Alberti for helpful discussions, and the comments of a reviewer.

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

The diversity and evolution of chelicerate hemocyanins

Peter Rehm[1], Christian Pick[1], Janus Borner[1], Jürgen Markl[2] and Thorsten Burmester[1*]

Abstract

Background: Oxygen transport in the hemolymph of many arthropod species is facilitated by large copper-proteins referred to as hemocyanins. Arthropod hemocyanins are hexamers or oligomers of hexamers, which are characterized by a high O_2 transport capacity and a high cooperativity, thereby enhancing O_2 supply. Hemocyanin subunit sequences had been available from horseshoe crabs (Xiphosura) and various spiders (Araneae), but not from any other chelicerate taxon. To trace the evolution of hemocyanins and the emergence of the large hemocyanin oligomers, hemocyanin cDNA sequences were obtained from representatives of selected chelicerate classes.

Results: Hemocyanin subunits from a sea spider, a scorpion, a whip scorpion and a whip spider were sequenced. Hemocyanin has been lost in Opiliones, Pseudoscorpiones, Solifugae and Acari, which may be explained by the evolution of trachea (i.e., taxon Apulmonata). Bayesian phylogenetic analysis was used to reconstruct the evolution of hemocyanin subunits and a relaxed molecular clock approach was applied to date the major events. While the sea spider has a simple hexameric hemocyanin, four distinct subunit types evolved before Xiphosura and Arachnida diverged around 470 Ma ago, suggesting the existence of a 4 × 6mer at that time. Subsequently, independent gene duplication events gave rise to the other distinct subunits in each of the 8 × 6mer hemocyanin of Xiphosura and the 4 × 6mer of Arachnida. The hemocyanin sequences were used to infer the evolutionary history of chelicerates. The phylogenetic trees support a basal position of Pycnogonida, a sister group relationship of Xiphosura and Arachnida, and a sister group relationship of the whip scorpions and the whip spiders.

Conclusion: Formation of a complex hemocyanin oligomer commenced early in the evolution of euchelicerates. A 4 × 6mer hemocyanin consisting of seven subunit types is conserved in most arachnids since more than 400 Ma, although some entelegyne spiders display selective subunit loss and independent oligomerization. Hemocyanins also turned out to be a good marker to trace chelicerate evolution, which is, however, limited by the loss of hemocyanin in some taxa. The molecular clock calculations were in excellent agreement with the fossil record, also demonstrating the applicability of hemocyanins for such approach.

Background

Hemocyanins are large copper-proteins that transport O_2 in the hemolymph of many arthropods and mollusks [1-3]. However, the hemocyanins of these two phyla are structurally different and emerged independently [4]. Hemocyanins evolved early in the arthropod stem lineage from the phenoloxidases, which are O_2-consuming enzymes involved in the melanin pathway [3]. Other members of the arthropod hemocyanin superfamily have lost the ability to bind copper and thus O_2, and gave rise to the non-respiratory pseudo-hemocyanins (cryptocyanins) in decapod crustaceans and the hexamerins in hexapods, which serve as storage proteins [3,5,6].

Arthropod hemocyanins form hexamers or oligo-hexamers of identical or related subunits with a molecular mass of about 75 kDa [1,2]. In each subunit, O_2-binding is mediated by two Cu^+ ions, which are coordinated by six histidine residues ("type III" copper binding site). Based on biochemical, immunochemical and molecular phylogenetic analyses, distinct hemocyanin subunit types have been identified in Chelicerata, Myriapoda, Crustacea and Hexapoda [3,7-12]. These subunits experienced

* Correspondence: thorsten.burmester@uni-hamburg.de
[1]Institute of Zoology and Zoological Museum, University of Hamburg, D-20146 Hamburg, Germany
Full list of author information is available at the end of the article

an independent evolution within each of these taxa, with the exception of a more complex pattern within the Pancrustacea (Crustacea and Hexapoda) [13].

Within the chelicerates, biochemical analyses have demonstrated the presence of hemocyanins in Xiphosura (horseshoe crabs), Scorpiones, Uropygi (whip scorpions), Amblypygi (whip spiders), and Araneae (true spiders), but failed to identify these respiratory proteins in Pycnogonida (sea spiders; Pantopoda), Solifugae (sunspiders) and Acari (mites and ticks) [8,10,11]. Complete sets of hemocyanin subunit sequences are available from the tarantula *Eurypelma californicum* (= *Aphonopelma hentzi*) [14], the hunting spider *Cupiennius salei* [15], the golden orb web spider *Nephila inaurata* [16] and the horse shoe crab *Carcinoscorpius rotundicauda* (GenBank acc. nos. DQ090484-DQ090469). Chelicerate hemocyanins are composed of up to eight distinct subunit types and form either 1 × 6, 2 × 6, 4 × 6 or 8 × 6mers [10,11]. Each subunit type occupies a distinct position within the native oligomer [7,17-21]. The subunits have similar oxygen binding properties, but different physico-chemical characteristics [22,23] and evolutionary origins [3,14-16].

Xiphosurans (horseshoe crabs) have the largest hemocyanin molecules known, consisting of 48 (8 × 6) subunits and up to eight distinct subunits types in *Limulus polyphemus* (6 × subunit type I, 8 × II, 2 × IIA, 8 × IIIA, 8 × IIIB, 8 × IV, 4 × V, 4 × VI) [1,10,19,21]. Scorpiones, Amblypygi, Uropygi, and some Araneae have 4 × 6mer hemocyanins. The 4 × 6mer hemocyanin of the tarantula *E. californicum* is the best studied example and comprises seven distinct subunit types (4 × a, 2 × b, 2 × c, 4 × d, 4 × e, 4 × f, and 4 × g-type subunits) [7,10,11,23]. A similar subunit composition was found in many other Araneae, the Amblypygi and the Uropygi [8,10]. Among the Araneae, variations from this “standard” scheme have been found in the entelegyne spiders of the RTA-clade (RTA = “retrolateral tibial apophysis”). In this large taxon, 1 × 6mer and 2 × 6mer hemocyanins occur that consist only of g-type subunits, but have lost the other six subunit types (a through f) present in other Araneae [10,11,15]. Scorpion hemocyanins are composed of eight subunit types named 2, 3A, 3B, 3C, 4, 5A and 5B [24]. Immunological and structural studies have suggested the correspondence between the distinct araneaen, scorpion and *Limulus* hemocyanin subunits [10,11,17,25-27]. This indicates that the evolution of distinct subunit types preceded the separation of Xiphosura and Arachnida.

To understand the evolution of the complex oligomeric structures of chelicerate hemocyanins, we have obtained 16 novel cDNA sequences of hemocyanin subunits from a sea spider, a horseshoe crab, a whip scorpion and a whip spider. Together with the previously sequenced hemocyanins from horseshoe crabs, scorpions and spiders, and those assembled from expressed sequence tags (ESTs), 67 full length chelicerate subunit sequences are available. These data allow us *i.* to trace the hemocyanin subunit evolution, *ii.* to reconstruct the emergence of the hemocyanin oligomers, and *iii.* to infer chelicerate phylogeny and divergence times.

Methods

Sequencing of chelicerate hemocyanin cDNA

Full length coding sequences were available for the hemocyanins of the tarantula *E. californicum* [14], the mangrove horseshoe crab *C. rotundicauda*, the golden orb web spider *N. inaurata* [16] and the hunting spider *C. salei* [15] (Additional file 1). In addition, hemocyanin primary structures had been obtained by conventional amino acid sequencing from the horseshoe crab *Tachypleus tridentatus* (TtrHcA; see Additional file 1 for the abbreviations of the proteins) [28] and the scorpion *Androctonus australis* (AauHc6) [29].

4,062 ESTs were generated from total RNA of the sea spider *Endeis spinosa* (Pycnogonida) as described before [30]. Gene ontology assessment and BLAST searches identified six ESTs with significant similarities to arthropod hemocyanins. Assembly of the ESTs resulted in a single cDNA sequence. Two cDNA clones from the original library were selected for sequencing by a commercial service (GATC, Konstanz, Germany), which both yielded identical sequences (acc. no. FR865911).

A cDNA library form the horseshoe crab *L. polyphemus* (Xiphosura) total RNA was prepared and screened with specific anti-*Limulus*-hemocyanin antibodies [21]. Four complete hemocyanin cDNAs sequences were obtained by primer walking. The cDNAs were assigned after translation to distinct subunits on the basis of known N-termini [31], identifying subunits II, IIIa, IV and VI (acc. nos. AM260213-AM260216). An additional hemocyanin cDNA (coding for subunit IIIB) was identified from a set of ESTs [30] and the complete coding sequence was obtained by primer walking (acc. no. FR865912).

A CloneMiner (Invitrogen) cDNA library was constructed from total RNA of a female emperor scorpion *Pandinus imperator* (Scorpiones) and submitted to 454 pyrosequencing [32], resulting in 428,844 high-quality reads. Hemocyanin subunit sequences were deduced from the assembled contigs (acc. nos. FN424079-FN424086).

A single whip spider *Euphrynichus bacillifer* (Amblypygi) was purchased from a commercial pet supplier. A cDNA library was constructed employing the Mint Universal kit (Evrogen). 433,348 reads were obtained from the cDNA by 454 pyrosequencing and the hemocyanin sequences were deduced from the assembled contigs (acc. nos. FR865913-FR865920).

A single whip scorpion *Mastigoproctus giganteus* (Uropygi) was purchased from a commercial pet supplier. Total RNA was extracted and converted into a Mint Universal cDNA library, from which 481,905 reads were obtained by 454 pyrosequencing. Full length coding sequences of hemocyanin subunits a, d, e, f, and g, as well as partial sequences from subunits b and c were deduced from the assembled contigs. The missing fragments of subunits b and c were obtained by RT-RCR employing gene specific primers. The cDNA fragments were cloned into pGEM and sequenced. The final subunit sequences have been submitted to the databases under the accession numbers FR865920-FR865926.

Sequence assembly and analyses

The web-based tools provided by the ExPASy Molecular Biology Server of the Swiss Institute of Bioinformatics (http://www.expasy.org) were used for cDNA translation and the analyses of amino acid sequences. A multiple sequence alignment of the amino acid sequences of all available arthropod hemocyanins and selected arthropod phenoloxidases was constructed employing MAFFT 6 [33] with the G-INS-i routine and the BLOSUM 45 matrix at http://mafft.cbrc.jp/alignment/server/. A complete list of sequences used in this study is provided in Additional file 1. For the phylogenetic inferences, signal peptides as well as the N- and C-terminal extensions of some phenoloxidases and hemocyanins were excluded from the multiple sequence alignment. The final alignment comprised 143 sequences and 912 positions (Additional file 2).

Phylogenetic analyses

The most appropriate model of amino acid sequence evolution (WAG + Γ model; [34]) was selected with ProtTest [35] using the Akaike Information Criterion. Bayesian phylogenetic analysis was performed using MrBayes 3.1.2 [36]. We assumed the WAG model with a gamma distribution of substitution rates. Metropolis-coupled Markov chain Monte Carlo sampling was performed with one cold and three heated chains. Two independent runs were performed in parallel for 4.5 million generations until the average standard deviation of split frequencies was < 0.01. Starting trees were random and the trees were sampled every 100th generation. The program Tracer 1.4 (http://tree.bio.ed.ac.uk/software/tracer/) was used to examine log-likelihood plots and Markov chain Monte Carlo summaries for all parameters. Posterior probabilities were estimated on the final 35,000 trees (burnin = 10,000). Trees were displayed using the arthropod phenoloxidases as an outgroup [4].

Molecular clock calculations

The program PhyloBayes 3.3 was used for molecular clock estimates [37], employing the MrBayes consensus tree as input. First three relaxed clock models, the lognormal autocorrelated clock model (LOG) [38], the Cox-Ingersoll-Ross process (CIR) [39] and uncorrelated gamma multipliers (UGM) [40] were compared by tenfold cross-validation with eight replicates, as specified in PhyloBayes 3.3. Rates across sites were modeled assuming a discrete gamma distribution with four categories. Divergence time priors were either uniform or modeled with a birth death process. Node ages were calculated using either hard constrains, which do not allow calibrated nodes to fall outside the calibration dates or soft bounds, which allows for divergence times outside the calibration interval [37]. For each of these settings rates across sites were modeled assuming a discrete gamma distribution with four categories and a Dirichlet process. All calculations were run for 50,000 (burnin 20,000) cycles.

The tree was calibrated with fossil constraints [41-43]. The maximum age for the separation of arthropod subphyla and thus the maximum age of the origin of the Chelicerata was set to the base of the Cambrian period 543 Ma ago (Ma). Stratigraphic information was obtained from http://www.fossilrecord.net[44]. Numerical ages derive from the "International Stratigraphic Chart" 2009 (http://www.stratigraphy.org) (Table 1).

Results

Hemocyanin sequences

A putative hemocyanin of the sea spider *E. spinosa* was identified in the ESTs of this species [30]. The EspHc1 cDNA measures 2,104 bp and translates into a protein of 631 amino acids with a predicted molecular mass of 72.2 kDa (Additional file 3). Full length cDNAs of five hemocyanin subunits of the Atlantic horseshoe crab *L. polyphemus* were obtained (LpoHcII, LpoHcIIIa, LpoHcIIIb, LpoHcIV, LpoHcVI). The predicted *L. polyphemus* hemocyanin subunits measure between 624 and 638 amino acids, with molecular masses of 72.3-73.4 kDa (Additional file 3). Subunits I, IIa and V could not identified in the cDNA library [21] or in the ESTs [30]. Seven hemocyanin subunit are available from the mangrove horseshoe crab *C. rotundicauda*, which corresponds to the subunits I, II, IIIa, IIIb, IV, V, and VI (acc. nos. DQ090484-DQ090490; Additional file 3). Thus seven of the eight hemocyanin subunit types of Xiphosura could be included in our analyses. The nature of an eighth type identified exclusively in *L. polyphemus* (IIa) and its exact topological position within the 4 × 6mer remains unclear. Its N-terminal sequence resembles that of subunit IIIa and it may occupy a homologous position in some hexamers [21,27,31].

454 pyrosequencing was employed to obtain ESTs from the scorpion *P. imperator* [32], the whip spider *E. bacillifer* (unpublished), the whip scorpion *M. giganteus* (unpublished), the pseudoscorpion *Chelifer cancroides*

Table 1 Calibration points used for relaxed Bayesian molecular clock analyses

Split	Bounds		Strata	Fossils	Reference
	Max	Min			
Euchelicerata-Pycnogonida	543	501	lower bound: first arthropods in the Early Cambrian upper bound: base of Upper Cambrian	*Rusophycus*; *Cambropycnogon klausmuelleri*	Benton 1993 [41], Crimes 1987 [45], Waloszek and Dunlop 2002 [46]
origin of Xiphosura		445	Late Ordovician	*Lunataspis aurora*	Rudkin et al. 2008 [47]
origin of Scorpiones		428	Silurian	*Allopalaeophonus caledonicus*	Dunlop 2010 [43]
Mygalomorphae-Araneaomorphae	382.7	240	lower bound: Grès à meules, upper Buntsandstein, Trias; upper bound: Givetian, middle Devonian	*Rosamygale grauvogeli;* *Attercopus fimbriunguis*	Selden and Gall 1992 [48], Selden et al. 2008 [49]

(unpublished), the sun spider *Gluvia dorsalis* (unpublished) and the harvestman *Phalangium opilio* (unpublished). Hemocyanin sequences were identified in *E. bacillifer* and *M. giganteus*, but not in *P. opilio* (474,081 reads), *G. dorsalis* (425,934 reads) or *C. cancroides* (443,697 reads). In both, *E. bacillifer* and *M. giganteus* seven hemocyanin sequences were found, which are orthologous to *E. californicum* hemocyanin subunits a-g (Figure 1). The predicted proteins measure between 621 and 639 amino acids, with molecular masses of 71.2-73.7 kDa (Additional file 3).

The publicly available chelicerate ESTs were searched for hemocyanin sequences using the tblastn algorithm (4 November 2011). We identified hemocyanin sequences in the ESTs from the tarantula *Acanthoscurria gomesiana* (see below), the hunting spider *C. salei* (two ESTs), the orb web spider *Nephila antipodiana* (two ESTs) and the tarantula *Aphonopelma* sp. (one EST). A total of 173 ESTs from the house spider *Parasteatoda tepidariorum* display significant similarities to hemocyanin and represent all seven subunits (a-g) found in other Araneae. However, none of the ESTs from *C. salei, N. antipodiana, Aphonopelma* sp. or *P. tepidariorum* could be assembled to a complete coding sequence and therefore these sequences were not included in our analyses. Lorenzini and colleagues obtained 6,790 ESTs from a hemocyte library of *A. gomesiana* [50]. A total of 463 ESTs display significant similarities with hemocyanin, as identified by BLAST searches. These sequences were assembled into eight contigs. On the basis of similarity searches, seven subunits were assigned to arachnid hemocyanin subunits a-g (Additional file 3). The eighth putative hemocyanin sequence (tentatively named AgoHcX) has no obvious ortholog among the chelicerate hemocyanin subunits.

Phylogeny of chelicerate hemocyanin subunits

Bayesian phylogenetic analyses show that the hemocyanins tree basically follows the accepted arthropod relationships on the level of the subphyla (Figure 1). The euarthropods split in the two sister groups Chelicerata and Mandibulata; though, the clade comprising the hemocyanins of the Mandibulata (Myriapoda + Pancrustacea) is poorly supported (Bayesian posterior probability 0.58). Myriapod and pancrustacean hemocyanins each are monophyletic, whereas the crustacean hemocyanins are not because the remipede hemocyanins were found more closely related to those of the insects than to the malacostracan hemocyanins [13].

Chelicerate hemocyanins are monophyletic (Bayesian posterior probability = 1.0). The lineage leading to the hemocyanin subunit from the sea spider *E. spinosa* (EspHc1) diverges first. Within the euchelicerate hemocyanins, four well-supported clades of distinct subunit types were identified (clades 1-4; Bayesian posterior probabilities ≥ 0.99). Clade 1 is the first branch within the euchelicerate hemocyanins and is formed by the arachnid b/c-type and xiphosuran V/VI-type subunits. These subunits facilitate the inter-hexamer contacts within the 4 × 6mer, as well as between the 2 × 6mer half-structures [1,21]. While the xiphosuran subunits V and VI form a common monophyletic clade, the arachnid hemocyanin subunits c do not. Here, the sequence of the putative c-subunit of the whip spider *E. bacillifer* (EbaHc-c) was found more diverged, mirrored by its basal position within the b/c clade.

Clade 2 comprises the arachnid a-type and the xiphosuran type II subunits. In agreement with immunological studies with the scorpion *A. australis* [17], the scorpion *P. imperator* has two a-type subunits, of which PimHc3A groups with a basal position to the other arachnid a-type subunits, while the position of PimHc3B is not well resolved. Note that subunit 3B represents a unique feature of scorpions that does not occur in the other arachnid hemocyanins.

The common clade that includes the remaining euchelicerate subunits received 0.99 Bayesian support. This clade splits into two sub-clades, leading to arachnid

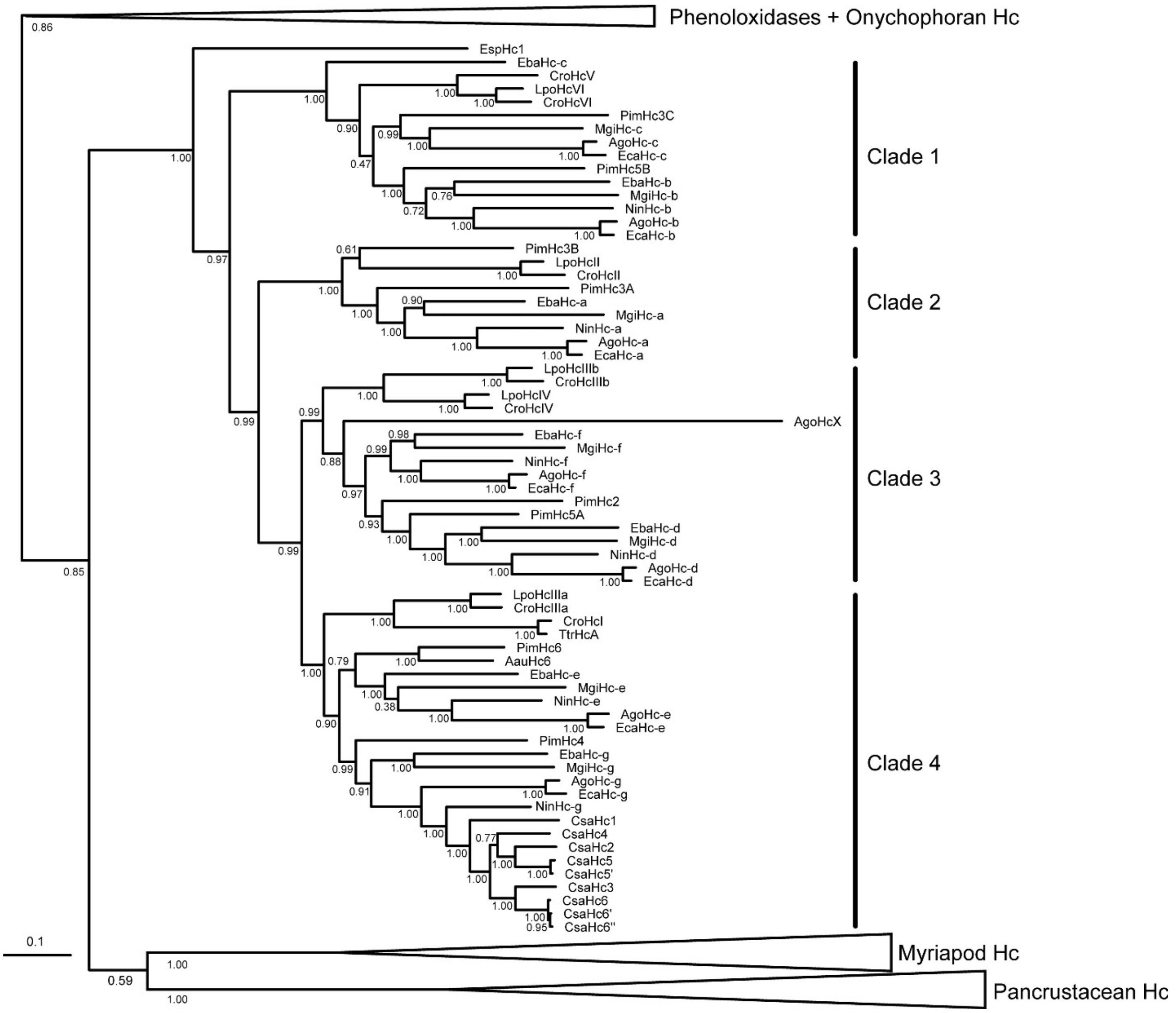


Figure 1 Phylogenetic tree of the chelicerate hemocyanin subunits. The numbers at the nodes represent Bayesian posterior probabilities estimated with the WAG model of amino acid substitution. The species abbreviations are: Aau, *Androctonus australis*; Ago, *Acanthoscurria gomesiana*; Cro, *Carcinoscorpius rotundicauda*; Csa, *Cupiennius salei*; Eba, *Euphrynichus bacillifer*; Eca, *Eurypelma californicum*; Esp, *Endeis spinosa*; Lpo, *Limulus polyphemus*; Mgi, *Mastigoproctus giganteus*; Nin, *Nephila inaurata*; Pim, *Pandinus imperator*; Ttr, *Tachypleus tridentatus*. The bar represents 0.1 expected substitutions per site. See Additional file 1 for abbreviations of the proteins

subunits d and f, and xiphosuran subunits IIIb and IV, on the one hand (clade 3), and arachnid subunits e and g, and xiphosuran subunits I and IIIa, on the other (clade 4). Within the arachnid d/f and e/g subtrees, respectively, the observed subunit relationships essentially mirror the expected phylogeny of the species, although the positions of the scorpion subunits are somewhat ambiguous (notably PimHc2). The HcX sequence, which had only been identified in *A. gomesiana*, is included in the d/f-clade, a position that is confirmed by pairwise comparisons, revealing that AgoHcX displays the highest sequence similarity to the arachnid f-subunits (not shown). Exclusion of AgoHcX from the phylogenetic analyses results in the same topology as described here, but Bayesian support slightly increased throughout the tree (Additional file 4).

Entelegyne spiders of the RTA-clade, such as *C. salei*, are known to diverge from the arachnid standard scheme of 4 × 6mer hemocyanins and to possess a mixture of 1 × 6mers and 2 × 6mers [10,11,18]. In agreement with previous studies [15], we found that the *C. salei* hemocyanin subunits all belong to the arachnid g-type, whereas the other six types (a-f) appear to have been lost during the evolution of this taxon. The *C. salei* hemocyanin subunits consistently group with the subunit g of *N. inaurata*, an entelegyne spider with a 4 × 6mer hemocyanin.

Molecular clock analyses of chelicerate hemocyanins

A timescale of chelicerate hemocyanin evolution was inferred on the basis of the chelicerate hemocyanins described above. Due to the divergent evolutionary rates (cf. Figure 1), we excluded the hemocyanin sequences of Onychophora, Myriapoda, Crustacea and Hexapoda, as well as the phenoloxidases. Cross-validation shows that that the uncorrelated gamma clock model (UGM) fits better than the log-normal model or the CIR process. The cross-validation score of UGM vs. log-normal was 6.4625 +/- 18.8999 and of UGM vs. CIR 5.8125 +/- 21.1925. Further support comes from a comparison of the calculated divergence times of different orthologous subunit pairs (e.g., EcaHc-a vs. NinHc-a compared to EcaHc-d vs. NinHc-d), which are most similar under UGM. Thus UGM was applied in our calculations. The time estimates with soft bounds and hard bounds were similar, with estimates being on average ~5% older when using hard bounds (Additional file 5 and Additional file 6). No notable difference was observed between the results obtained with rates modeled with the Dirichlet process or with a discrete gamma distribution (< 0.1% mean difference).

The divergence times resulting from the estimates obtained with soft bounds, a birth death process and rates modeled with the Gamma distribution are displayed in Figure 2 (see Additional file 5). We applied five calibration points, which correspond to four upper (minimum) and two lower (maximum) bounds derived from the fossil record (Table 1). Subunits with uncertain orthology were ignored. We first calculated the divergence times of the distinct subunit types in Arachnida and Xiphosura (Figure 2A). The earliest split of euchelicerate hemocyanins is formed by the clade of arachnid b/c and xiphosuran V/VI-type subunits, which occurred ~540 Ma. Within this clade, the exact arrangement of the three clades consisting of *i.* arachnid b-type subunits, *ii.* arachnid c-type subunits and xiphosuran V/VI-type subunits is not well resolved (Figure 1), which is reflected by a rapid diversification 437-453 Ma. Xiphosuran subunits V and VI separated ~169 Ma. The a-type subunits, which include the xiphosuran subunit II, split ~536 Ma, followed by the four clades defined above, consisting of arachnid subunits d/f, arachnid e/g, xiphosuran I/IIIa, and xiphosuran IIIb/IV subunits, which commenced to diversify ~509 Ma. Arachnid d- and f-type subunits split ~441 Ma, arachnid e- and g-type subunits ~467 Ma. Separation events within the xiphosuran subunits (I vs. IIIa and IIIb vs. IV) occurred 304 and 309 Ma, respectively. In subsequent analyses, the hemocyanin sequences were employed to estimate chelicerate divergence times (Figure 2B, see below).

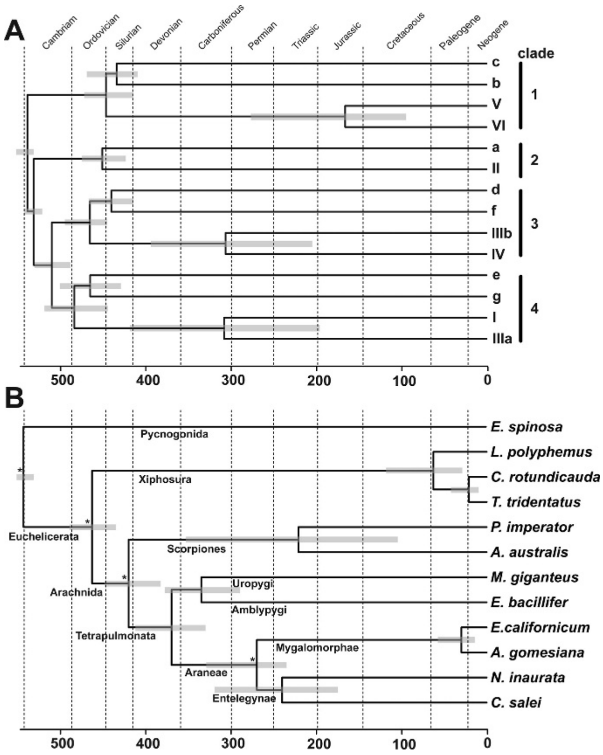


Figure 2 Timescale of hemocyanin evolution. **A**. Evolution in the chelicerate hemocyanin subunits. **B**. Hemocyanin-derived timescale of chelicerate evolution. The divergence times are means resulting from the estimates obtained with a birth-death process and soft bounds. Rates across sites were modeled assuming a gamma distribution. The grey bars correspond to the 95% confidence intervals. Ma, million years ago; asterisks denote the nodes used for calibration (see Table 1).

Hemocyanin-derived phylogeny of Chelicerata

We specifically studied the relationships among chelicerate taxa by concatenating the seven orthologous hemocyanin subunit sequences, representing spider subunits a-g, into a single sequence alignment. Clear orthologs were available for *E. californicum, A. gomesiana, N. inaurata, E. bacillifer*, and *M. giganteus*, respectively. There is no c-type subunit in *N. inaurata*, which was coded as missing data. *P. imperator* has two a-type subunits (PimHcIIIA and PimHcIIIB), of which we selected PimHcIIIA on the basis of its closer relationship to araneaen a-type subunits. In case of *C. rotundicauda*, clear orthology assessment was only possible for subunit a (see above). We assigned CroHcV and VI to b and c, CroHcIIIb and IV to d and f, and CroHcI and IIIa to e and g-subunits. In a second approach, we exchanged these sequences. Because the single *E. spinosa* hemocyanin subunit constitutes a conclusive outgroup for all seven subunit types (Figure 1), seven copies of this sequence were concatenated.

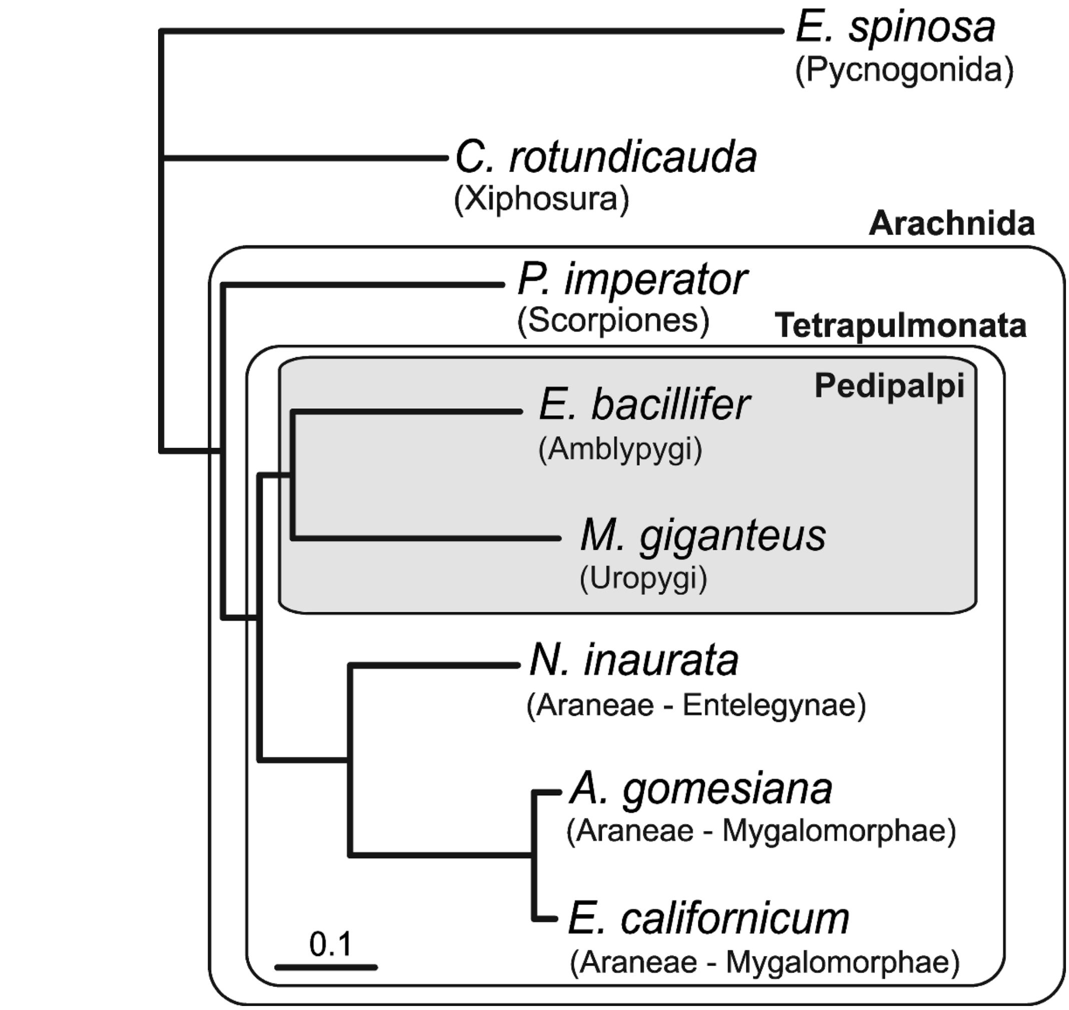


Figure 3 Hemocyanin-based phylogeny of chelicerates. The tree was derived from the concatenated hemocyanin alignment. The Pedipalpi are shaded. All nodes are supported with 1.0 Bayesian probabilities. The bar represents 0.1 substitutions per site.

Bayesian phylogenetic analyses were applied and the sea spider *E. spinosa* was used as outgroup (Figure 3.3). All nodes displayed a Bayesian support of 1.0 and there was no effect of an exchange of the ambiguous subunits from *C. rotundicauda*. The Xiphosura (*C. rotundicauda*) are the sister taxon of the Arachnida. Within the Arachnida, the scorpion *P. imperator* diverged first. Whip spiders (Amblypygi) and whip scorpions (Uropygi) form a common clade (Pedipalpi). The Pedipalpi form the sister taxon of the Araneae, represented by the mygalomorph spiders *E. californicum* and *A. gomesiana* on the one hand, and the entelegyne spider *N. inaurata* on the other.

According to the molecular clock calculations (see above), the hemocyanins of Pycnogonida and Euchelicerata diverged ~543 Ma (Figure 2B). Considering the different orthologous subunits, xiphosuran and arachnid hemocyanins separated between 444 to 489 Ma (mean 462 Ma). We calculated that the orthologous hemocyanin subunits of Scorpiones and Tetrapulmonata (i.e. Araneae + Pedipalpi) split ~419 Ma (405-440 Ma). Within the scorpions, *P. imperator* (Iurida) and *A. australis* (Buthida) separated ~221 Ma. The hemocyanins of Pedipalpi and Araneae diverged ~369 Ma (357-414 Ma), those of Amblypygi and Uropygi 334 Ma (316-344 Ma). Within the Araneae, hemocyanins of *N. inaurata* (Entelegynae) and the Mygalomorphae (*E. californicum* + *A. gomesiana*) diverged ~271 Ma (254-288 Ma). The hemocyanins of *E. californicum* and *A. gomesiana* split 30 Ma (21-35 Ma). *N. inaurata* subunit g and the *C. salei* hemocyanins separated ~239 Ma; the xiphosurans *L. polyphemus* and *C. rotundicauda* diverged ~62 Ma (56-67 Ma).

Discussion

Hemocyanin subunits assemble into hexamers, which may form quaternary structures comprising up to 8 ×

6mers [1]. The evolutionary advantage of large oligomers presumably lies in a higher O_2-carrying capacity per mol and higher cooperativity, which also enhances the O_2 transport, combined with a low viscosity and a low colloid-osmotic pressure of the hemolymph. The phylogenetic tree permits inferring the origins and modifications of these complex protein structures in the chelicerates.

The presence of only a single subunit in *E. spinosa* along with its basal position in the tree suggests that early chelicerate hemocyanins had a simple, homo-hexameric structure (Figure 4). This hypothesis is supported by the independent emergence of hemocyanin oligohexamers in the other arthropod subphyla, which hints to more simple hemocyanins in the last common arthropod ancestor [1,3].

Early emergence of euchelicerate hemocyanin oligomers

The phylogenetic analyses demonstrate that the early euchelicerate hemocyanin, which was already used for O_2 supply in the last common ancestor of the arachnids and the xiphosurans more than 445 Ma, was composed of at least four distinct subunit types. These subunits were the ancestors of the subunits represented in clades 1-4, respectively (Figure 1 and 2A). According to our phylogenetic reconstruction, clade 1, which includes arachnid b/c and xiphosuran V/VI subunits, diverged first. Both b/c and V/VI-subunits form heterodimers, which are responsible for the contacts between the hexamers [1,7,17,21,31]. Thus the emergence of clade 1-subunits was most likely associated with the organization of the first oligo-hexameric hemocyanin (Figure 4). This event must have taken place very early in the evolution of euchelicerates, and may be associated with significant morphological and physiological changes. We calculated that the corresponding gene duplication occurred ~540 Ma (Figure 2A). It may be speculated that first a 2 × 6mer consisting of two distinct subunit types evolved. Such dodecamers occur today in various crustaceans [1,10] and in the spiders of the RTA-clade [15,18]. Alternatively, this gene duplication has already resulted in a typical chelicerate 4 × 6mer, which is stabilized exclusively by a central tetrameric ring of clade 1 subunits. Notably, stable 4 × 6mers were obtained in hybrid reassembly experiments from mixtures of a clade 1 heterodimer as “linker” and another subunit type as “hexamer former” [51]. These experiments convincingly worked with scorpion heterodimer 5B-3C plus *L. polyphemus* subunit II, tarantula heterodimer b-c plus scorpion subunit 4, and *Limulus* heterodimer V-VI plus scorpion subunit 4.

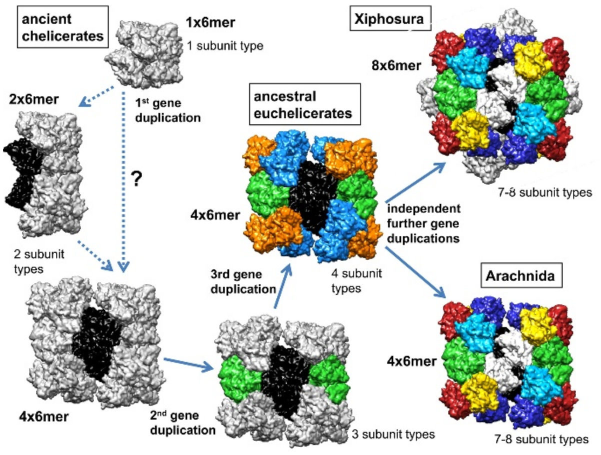


Figure 4 Scheme of hemocyanin evolution in Chelicerata. Color code: black/white, subunit clade 1 (b/c/V/VI); green, subunit clade 2 (a/II); medium blue, subunit clade 3 (d/f/IIIb/IV); orange, subunit clade 4 (e/g/I/IIIa); light blue, d/IV; dark blue f/IIIb^{Q3}; yellow, g/IIIa; red, e/I. See text for further details and explanations.

The next step in evolution was the separation of clade 2 from the remaining subunits ~536 Ma. Clade 2 includes arachnid subunit a and xiphosuran subunit II. With the exception of PimHc3B (which exclusively occurs in scorpions), the a-subunits follow the expected phylogeny of the euchelicerates. In the hemocyanin quaternary structure, subunit a/II is located at the interhexamer interface of the basic 2 × 6mer [21]. It may therefore be speculated that this step was required not only for stabilization of the 4 × 6mer hemocyanin, but also for improving cooperativity (Figure 4).

An additional gene duplication event, which may have taken place ~510 Ma, long before the arachnid-xiphosuran split, gave rise to both, clade 3, consisting of arachnid subunits d and f, and xiphosuran subunits IIIb and IV, and clade 4, comprising arachnid subunits e and g, and xiphosuran subunits I and IIIa. Even though this largely is in accordance with previous comparative immunochemical studies [10,11,25,27], a common origin of subunits e and I, d and IV, f and IIIb, and IIIa and g, respectively, as proposed before [10,26], can now be excluded.

The separation of the arachnid subunits d and f, and e and g, respectively, occurred before scorpions and spiders diverged. Thus the last common ancestor of all arachnids had a 4 × 6mer of seven subunits that were similar to subunit types a-g, demonstrating the evolutionary success of this conserved structure, which has remained essentially unchanged for more than 450 Ma.

Independent but parallel evolution of hemocyanin oligomers in Xiphosura and Arachnida

In horseshoe crabs, a large 8 × 6mer evolved. Although cooperativity is not further enhanced by this step, it might have been required for reducing the osmotic pressure and the viscosity of the hemolymph [1]. Notably, the duplication of the subunits corresponding to arachnid b and c (i.e., V and VI), d and f (IIIb and IV) and e and g (I and IIIa) occurred independently and the separation of xiphosuran subunits occurred more than 150 Ma later. Nevertheless, a comparable subunit diversity evolved: Six distinct subunit types for each topological position within the hexamer, plus a heterogeneity of the central linker unit to form an asymmetric 2 × 6mer. Consequently, today both arachnid and xiphosuran hemocyanins consist of seven distinct subunits (plus an independently evolved eighth subunit in *L. polyphemus* [IIa] and the scorpions [3B]), but only subunits a and II (clade 2) are one-to-one-orthologs. Thus, there was an evolutionary pressure to maximize the distinctiveness of subunits within the hemocyanin hexamer, which may be explained by better regulatory properties.

Evolution of the arachnid hemocyanins

As outlined above, the principal structure of the early arachnid hemocyanin was most likely a 4 × 6mer (Figure 4). Previous immunological and structural investigations identified orthologs between scorpions and spiders (Araneae) [10,11,17,24,25,27]. The phylogenetic tree (Figure 1) shows that these studies were essentially correct. We should also note that in our previous phylogenetic analyses, the subunit AauHc6 was assigned to the araneaen g-type subunits [16], while on the basis of the structural and immunological similarities AauHc6 was homologized with e-type subunits [29]. The new tree, which includes more sequences, suggests that the protein-based studies were correct and AauHc6 is indeed an e-type subunit.

The 4 × 6mer structure is present in the mygalomorph spiders (*E. californicum* and *A. gomesiana*), and is found also in many Entelegynae (eight-eyed spiders; e.g., *N. inaurata*) [10,14,16]. The entelegyne spiders of the RTA-clade, however, diverge from this standard scheme and have a mixture of 1 × 6mer and 2 × 6mer hemocyanins [10,15,18,25]. This hemocyanin type is built by six distinct g-type subunits, with subunit CsaHc1 (see Figure 1) forming the inter-hexamer bridge within the 2 × 6mer molecules [15]. This suggests a loss of the other six subunit types (a-f) during evolution. Thus the ancestor of RTA-clade spiders most likely had a simple hexameric hemocyanin, exclusively built by g-type subunits. Some 170 Ma, the reconstruction of a more complex hemocyanin type commenced. This might be explained by physiological and behavioral changes that e.g. required a higher oxygen capacity in the hemolymph and/or a hemocyanin with a higher cooperativity.

There is little information about the subunit AgoHcX, which is only known from the ESTs of *A. gomesiana* [50]. Phylogenetic analyses place the protein with a long branch at the base of the arachnid d/f-subunit clade. The fact that the AgoHcX sequence was found in ESTs and does not display any nonsense mutation suggests that this unique hemocyanin-like protein is translated into a functional protein. However, it is neither known whether HcX is restricted to certain taxa (e.g., the mygalomorph spiders) nor whether it is component of the hemocyanin oligomer. The latter seems to be unlikely because of its derived sequence. In addition, no evidence for HcX was found in the hemocyanin of the closely related tarantula *E. californicum*, despite more than 30 years of research.

Absence of hemocyanin in some chelicerate taxa

Notably, some arachnids do not have hemocyanin or any other O_2-transport protein in their hemolymph. Despite the large number of ESTs obtained, no hemocyanin sequences were detected in the harvestman *P. opilio* (Opiliones), the pseudoscorpion *C. cancroides* (Pseudoscorpiones) and the sun spider *G. dorsalis* (Solifugae). This observation essentially agrees with previous findings [10,11,26], with the exception that Kempter et al. [26] suggested the presence of a dodecameric hemocyanin in the harvestman *Leiobunum limbatum*. However, re-evaluation of the original data and new experiments suggest that the protein in question may actually be a vitellogenin-like, di-tetrameric protein (not shown), similar to those found in other arachnids [8]. In this context, we would also like to note that in contrast to previous suggestions [10], the haplogyne spider *Dysdera* does not possess a 1 × 6mer hemocyanin. A recent reinvestigation of a number of individuals demonstrated that *Dysdera* lacks any hemocyanin, but express a tetrameric non-respiratory protein (not shown). We formally cannot exclude that hemocyanins are present in other species of these taxa or are expressed only under certain environmental conditions. However, we consider such scenario unlikely because such specific expression is not observed in other chelicerates. In addition, no evidence for hemocyanin was found in the 394,960 ESTs or the available genomic sequences of Acari. Thus mites and ticks most likely lack hemocyanin as well.

Opiliones, Pseudoscorpiones and Solifugae are apulmonate arachnids. The absence of hemocyanin may be a synapomorphic character and an indication for a close relationship of these taxa, which agrees with some character-based phylogenetic studies [52,53] (see below). Morphological and/or physiological characteristics may have rendered a respiratory protein unnecessary. Notably, apulmonate arachnids do not breathe through book lungs, but possess trachea, which may be sufficient to support the aerobic metabolism. The hemocyanin-less Acari (mites and ticks) are usually small and also have trachea [54]. By contrast, spiders, scorpions, whip spiders, whip scorpions have book lungs, which are filled with hemolymph and which may limit O_2 consumption. Here, hemocyanin may be required for efficient O_2 uptake and distribution.

Implications for chelicerate phylogeny

Hemocyanin sequences have been successfully used to infer arthropod phylogeny [3,4,12,15,16,55-57]. Traditionally, Chelicerata were considered as the sister group of the Mandibulata, a taxon that comprises Myriapoda, Crustacea and Hexapoda [58]. However, several molecular phylogenetic studies have provided evidence for a common clade of Myriapoda and Chelicerata ("Myriochelata" or "Paradoxopoda" hypothesis; e.g., [59-62]). Morphological evidence is poor and restricted to similarities of neurogenesis [63], which may, however, also represent a plesiomorphic state. In our study, we received some support for the monophyly of Mandibulata, which agrees with previous studies employing hemocyanin sequences [9], as well as other molecular approaches [64,65].

The relationship among the major chelicerate lineages is controversial. Phylogenetic trees derived from the hemocyanin subunit sequences (Figure 1) or the concatenated alignment (Figure 3) can also be used to deduce the relative position of some chelicerate taxa, while others cannot be considered due to the lack of hemocyanin (Acari, Opiliones, Pseudoscorpiones, Solifugae). Notably, Weygoldt and Paulus [52] suggested a taxon "Apulmonata", which joins Solifugae, Opiliones, Pseudoscorpiones, Acari, Ricinulei (hooded tickspiders), and Palpigradi (microwhip scorpions). Palpigradi and Ricinulei were not available for our studies. However, the absence of hemocyanin in Acari, Opiliones, Pseudoscorpiones and Solifugae tentatively supports monophyletic "Apulmonata".

Morphological and molecular studies have placed the pycnogonids (sea spiders) either as sister group of the Euchelicerata [59,66], nested within the Chelicerata [67], or considered them as the sister group of all other Euarthropoda ("Cormogonida" hypothesis; [68]). Our phylogenetic tree (Figure 1) strongly supports the inclusion of the Pycnogonida in the Chelicerata as sister group of the Euchelicerata. This position is also tentatively supported by the hemocyanin mono-hexamer and is in line with recent neuroanatomical studies, which demonstrated the homology of deuterocerebral appendages of Pycnogonida and Euchelicerata [69]. An ingroup position of the Pycnogonida with the Arachnida, as deduced from complete mitochondrial DNA sequences [70,71] is not supported by our data and should be considered unlikely.

In agreement with morphological considerations and most previous molecular phylogenetic studies, the Xiphosura form the sister group of the Arachnida. Within the arachnids, the relative positions of Scorpiones, Araneae, Uropygi and Amblypygi are controversial [66]. We found monophyletic Tetrapulmonata (Araneae, Uropygi, and Amblypygi), which is the sister group of the scorpions. In previous phylogenetic analyses, the relative position of the taxa Araneae, Amblypygi and Uropygi has been controversial. While some morphological studies favor a sister group relationship between Araneae and Amblypygi, forming the taxon Labellata [66,72], others support a common taxon referred to as Pedipalpi, which comprises the Uropygi and Amblypygi [73]. The latter view is supported by our molecular phylogenetic trees. This finding holds for the tree derived from the concatenated alignment (Figure 3) as well as for most analyses of single subunits (Figure 1), with the exception of an unusual position of *E. bacillifer* subunit c and unresolved relationships among the subunits e.

The fossil record of chelicerates is far from being complete, but still allows the estimation of the evolutionary history of this taxon [42,43]. The true origin of the chelicerates is currently uncertain, but dates back at least to the early Cambrian period [43]. We calculated that the first split within the chelicerates occurred 542 Ma (Figure 2B). This slightly predates the earliest stemline chelicerates, which derive from the Lower Cambrian Maotianshan Shale some 530 Ma [74]. The first putative pycnogonid derives from the Orsten fauna ~500 Ma [46], the oldest xiphosuran fossil was found in a Late Ordovician Lagerstätte and dates ~445 Ma [47] and the first unambiguous arachnid is a ~428 Ma old Silurian scorpion [42,43]. These dates are actually close to our molecular clock estimates (463 and 420 Ma, respectively; Figure 2B). The first fossils of true spiders, whip spiders and whip scorpions were found in Carboniferous strata, which are 310-320 Ma old, although the Tetrapulmonata are probably of Devonian origin [43]. Our calculations confirm this notion and date the origin of the clade leading to Tetrapulmonata 369 Ma. The oldest opisthothele fossil (modern spiders) is a mygalomorph spider dating 240 Ma, while the oldest representative of the sistergroup Araneomorphae (web-building spiders) is of early Cretaceous origin [43]. We calculated the earliest divergence within the Araneae 271 Ma, which is somewhat older.

The most successful subgroup within the Araneomorphae are the Entelegynae, which are subdivided into the Orbicularidae and the spiders of the RTA clade. The lower bound of divergence of the Orbicularidae (e.g., *N. inaurata*) and the RTA-clade (*C. salei*) is a net from an orbicularian spider from the early Cretaceous period, some 140 Ma. We calculated that the formation of the *Cupiennius*-type hemocyanin commenced about 171 Ma, which should be considered as the lower bound for the time of emergence of the RTA-clade.

Conclusions

Our results clearly demonstrate that chelicerate hemocyanin structure is conservative, but also allows innovations. There is little doubt that hemocyanin evolution commenced as a hexamer with a single subunit type, as present today in the sea spider. The first hemocyanin oligo-hexamer emerged early in euchelicerate evolution, probably associated with the demand for better oxygen supply. Gene duplications led to the formation of a 4 × 6mer hemocyanin in early euchelicerates, which was structurally retained in the arachnids. In xiphosurans, however, an 8 × 6mer hemocyanin built from two identical 4 × 6mers emerged. Although in both arachnids and xiphosurans at least two additional but independent subunit duplications occurred, the architecture of the 4 × 6mer has remained conserved in most taxa for more than 450 Ma. Only in the spiders of the RTA-clade, gene losses and independent duplications gave rise to a novel hemocyanin version, as exemplified by the 2 × 6mer hemocyanin of *C. salei*. Again changing physiological demands may have been the cause for these events. The conservative structure of hemocyanins makes them an excellent marker to trace chelicerate evolution, which is only limited by the absence of hemocyanin in some taxa.

Additional material

Additional file 1: List of sequences used in this study. The accession numbers of the cDNA sequences are given, except (*), which has been derived by conventional protein sequencing. SU = subunit.

Additional file 2: Multiple sequence alignment of chelicerate, crustacean, myriapod and insect hemocyanins, and selected arthropod phenoloxidases.

Additional file 3: Molecular properties of chelicerate hemocyanin cDNA and the deduced amino acid. The asterisks (*) denote incomplete N-terminal sequences of *P. imperator* hemocyanins subunits, with 8 and 9 amino acids missing.

Additional file 4: Phylogenetic tree of the chelicerate hemocyanin subunits excluding AgoHcX. The numbers at the nodes represent Bayesian posterior probabilities estimated with the WAG model of amino acid substitution. The species abbreviations are: Aau, *Androctonus australis*; Ago, *Acanthoscurria gomesiana*; Cro, *Carcinoscorpius rotundicauda*; Csa, *Cupiennius salei*; Eba, *Euphrynichus bacillifer*; Eca, *Eurypelma californicum*; Esp, *Endeis spinosa*; Lpo, *Limulus polyphemus*; Mgi, *Mastigoproctus giganteus*; Nin, *Nephila inaurata*; Pin, *Pandinus imperator*; Ttr, *Tachypleus tridentatus*. See Additional file 1 for abbreviations of the proteins.

Additional file 5: Divergence times of chelicerate hemocyanin subunit types (see Figure 2A). Rates across sites were modeled assuming a gamma distribution (Γ) or with a Dirichlet process (D). Divergence time priors were either uniform or modeled with a birth death process. Hard or soft bounds were applied. Divergence times are given in Ma.

Additional file 6: Divergence times of chelicerate taxa, as estimated from the hemocyanin sequences (see Figure 2B). Rates across sites were modeled assuming a gamma distribution (Γ) or with a Dirichlet process (D). Divergence time priors were either uniform or modeled with a birth death process. Hard or soft bounds were applied. Divergence times are given in Ma.

Abbreviations

CIR: Cox-Ingersoll-Ross process; ESTs: Expressed sequence tags; LOG: Lognormal autocorrelated clock model; Ma: Million years ago; RTA: Retrolateral tibial apophysis; UGM: Uncorrelated gamma multipliers.

Acknowledgements

We thank Falko Roeding, Carina Edel and Natalija Winokurowa for their help with cloning and sequencing, and Even Solem for his support with the experiments with *Dysdera*. We are grateful to Jason Dunlop for checking the chelicerate calibration points. The work was supported by grants of the Deutsche Forschungsgemeinschaft (Bu956/8 and Bu956/9).

Author details

[1]Institute of Zoology and Zoological Museum, University of Hamburg, D-20146 Hamburg, Germany. [2]Institute of Zoology, Johannes Gutenberg University Mainz, D-55099 Mainz, Germany.

Authors' contributions

TB conceived the study and carried out the phylogenetic analyses. PR, CP and JB provided and analyzed sequence data. PR performed the molecular clock analyses. JM drafted the model of hemocyanin evolution. PR, JM and TB drafted the manuscript. All authors read and approved the final version of the manuscript.

Received: 4 November 2011 Accepted: 14 February 2012
Published: 14 February 2012

References

1. Markl J, Decker H: **Molecular structure of the arthropod hemocyanins.** *Adv Comp Environm Physiol* 1992, **13**:325-376.
2. Van-Holde KE, Miller KI: **Hemocyanins.** *Adv Protein Chem* 1995, **47**:1-81.
3. Burmester T: **Origin and evolution of arthropod hemocyanins and related proteins.** *J Comp Physiol B* 2002, **172**:95-107.
4. Burmester T: **Molecular evolution of the arthropod hemocyanin superfamily.** *Mol Biol Evol* 2001, **18**:184-195.
5. Markl J, Burmester T, Decker H, Savel-Niemann A, Harris JR, Süling M, Naumann U, Scheller K: **Quaternary and subunit structure of *Calliphora* arylphorin as deduced from electron microscopy, electrophoresis, and sequence similarities with arthropod hemocyanin.** *J Comp Physiol B* 1992, **162**:665-680.
6. Pick C, Burmester T: **A putative hexamerin from a Campodea sp. suggests an independent origin of haemocyanin-related storage proteins in Hexapoda.** *Insect Mol Biol* 679, **18**:673-679.
7. Markl J, Kempter B, Linzen B, Bijlholt MMC, Van-Bruggen EFJ: **Hemocyanins in spiders, XVI. Subunit topography and a model of the quaternary structure of *Eurypelma* hemocyanin.** *Hoppe Seylers Z Physiol Chem* 1981, **362**:1631-1641.
8. Markl J, Markl A, Schartau W, Linzen B: **Subunit heterogeneity in arthropod hemocyanins: I. Chelicerata.** *J Comp Physiol B* 1979, **130**:283-292.
9. Kusche K, Hembach A, Hagner-Holler S, Gebauer W, Burmester T: **Complete subunit sequences, structure and evolution of the 6 × 6-mer hemocyanin from the common house centipede, *Scutigera coleoptrata*.** *Eur J Biochem* 2003, **270**:2860-2868.
10. Markl J: **Evolution and function of structurally diverse subunits in the respiratory protein hemocyanin from arthropods.** *Biol Bull (Woods Hole, MA)* 1986, **171**:90-115.
11. Markl J, Stöcker W, Runzler R, Precht E: **Immunological correspondences between the hemocyanin subunits of 86 arthropods: evolution of a multigene protein family.** In *Invertebrate oxygen carriers.* Edited by: Linzen B. Heidelberg: Springer; 1986:281-292.
12. Pick C, Schneuer M, Burmester T: **The occurrence of hemocyanin in Hexapoda.** *FEBS J* 2009, **276**:1930-1941.
13. Ertas B, Von-Reumont BM, Wagele JW, Misof B, Burmester T: **Hemocyanin suggests a close relationship of Remipedia and Hexapoda.** *Mol Biol Evol* 2009, **26**:2711-2718.
14. Voit R, Feldmaier-Fuchs G, Schweikardt T, Decker H, Burmester T: **Complete sequence of the 24-mer hemocyanin of the tarantula *Eurypelma californicum*. Structure and intramolecular evolution of the subunits.** *J Biol Chem* 2000, **275**:39339-39344.

15. Ballweber P, Markl J, Burmester T: **Complete hemocyanin subunit sequences of the hunting spider *Cupiennius salei*: recent hemocyanin remodeling in entelegyne spiders.** *J Biol Chem* 2002, **277**:14451-14457.
16. Averdam A, Markl J, Burmester T: **Subunit sequences of the 4 × 6-mer hemocyanin from the golden orb-web spider, *Nephila inaurata*.** *Eur J Biochem* 2003, **270**:3432-3439.
17. Lamy J, Bijlholt MC, Sizaret PY, Van-Bruggen EF: **Quaternary structure of scorpion (*Androctonus australis*) hemocyanin. Localization of subunits with immunological methods and electron microscopy.** *Biochemistry* 1981, **20**:1849-1856.
18. Markl J: **Hemocyanins in spiders, XI. The quaternary structure of *Cupiennius* hemocyanin.** *J Comp Physiol B* 1980, **140**:199-207.
19. Lamy J, Sizaret PY, Frank J, Verschoor A, Feldmann R, Bonaventura J: **Architecture of *Limulus polyphemus* hemocyanin.** *Biochemistry* 1982, **21**:6825-6833.
20. Taveau JC, Boisset N, Lamy J, Lambert O, Lamy JN: **Three-dimensional reconstruction of *Limulus polyphemus* hemocyanin from cryoelectron microscopy.** *J Mol Biol* 1997, **266**:1002-1015.
21. Martin AG, Depoix F, Stohr M, Meissner U, Hagner-Holler S, Hammouti K, Burmester T, Heyd J, Wriggers W, Markl J: ***Limulus polyphemus* hemocyanin: 10Å cryo-EM structure, sequence analysis, molecular modelling and rigid-body fitting reveal the interfaces between the eight hexamers.** *J Mol Biol* 2007, **366**:1332-1350.
22. Decker H, Markl J, Loewe R, Linzen B: **Hemocyanins in spiders VIII. Oxygen affinity of the individual subunits isolated from *Eurypelma californicum* hemocyanin.** *Hoppe Seylers Z Physiol Chem* 1979, **360**:1505-1507.
23. Markl J, Savel A, Linzen B: **Hemocyanins in spiders XIV. Subunit composition of dissociation intermediates and its bearing on quarternary structure of Eurypelma hemocyanin.** *Hoppe Seylers Z Physiol Chem* 1981, **362**:1255-1262.
24. Lamy J, Billiald P, Sizaret PY, Cave G, Frank J, Motta G: **Approach to the direct intramolecular localization of antigenic determinants in *Androctonus australis* hemocyanin with monoclonal antibodies by molecular immunoelectron microscopy.** *Biochemistry* 1985, **24**:5532-5542.
25. Markl J, Gebauer W, Runzler R, Avissar I: **Immunological correspondence between arthropod hemocyanin subunits. I. Scorpion (*Leiurus, Androctonus*) and spider (*Eurypelma, Cupiennius*) hemocyanin.** *Hoppe Seylers Z Physiol Chem* 1984, **365**:619-631.
26. Kempter B, Markl J, Brenowitz M, Bonaventura C, Bonaventura J: **Immunological correspondence between arthropod hemocyanin subunits. II. Xiphosuran (*Limulus*) and spider (*Eurypelma, Cupiennius*) hemocyanin.** *Biol Chem Hoppe Seyler* 1985, **366**:77-86.
27. Lamy J, Compin S, Lamy JN: **Immunological correlates between multiple isolated subunits of *Androctonus australis* and *Limulus polyphemus* hemocyanins: an evolutionary approach.** *Arch Biochem Biophys* 1983, **223**:584-603.
28. Linzen B, Soeter NM, Riggs AF, Schneider HJ, Schartau W, Moore MD, Yokota E, Behrens PQ, Nakashima H, Takagi T, *et al*: **The structure of arthropod hemocyanins.** *Science* 1985, **229**:519-524.
29. Buzy A, Gagnon J, Lamy J, Thibault P, Forest E, Hudry-Clergeon G: **Complete amino acid sequence of the Aa6 subunit of the scorpion *Androctonus australis* hemocyanin determined by Edman degradation and mass spectrometry.** *Eur J Biochem* 1995, **233**:93-101.
30. Meusemann K, Von-Reumont BM, Simon S, Roeding F, Strauss S, Kuck P, Ebersberger I, Walzl M, Pass G, Breuers S, *et al*: **A phylogenomic approach to resolve the arthropod tree of life.** *Mol Biol Evol* 2010, **27**:2451-2464.
31. Lamy J, Lamy J, Sizaret PY, Billiald P, Jolles P, Jolles J, Feldmann RJ, Bonaventura J: **Quaternary structure of *Limulus polyphemus* hemocyanin.** *Biochemistry* 1983, **22**:5573-5583.
32. Roeding F, Borner J, Kube M, Klages S, Reinhardt R, Burmester T: **A 454 sequencing approach for large scale phylogenomic analysis of the common emperor scorpion (*Pandinus imperator*).** *Mol Phylogenet Evol* 2009, **53**:826-834.
33. Katoh K, Kuma K, Toh H, Miyata T: **MAFFT version 5: improvement in accuracy of multiple sequence alignment.** *Nucleic Acids Res* 2005, **33**:511-518.
34. Whelan S, Goldman N: **A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach.** *Mol Biol Evol* 2001, **18**:691-699.
35. Abascal F, Zardoya R, Posada D: **ProtTest: selection of best-fit models of protein evolution.** *Bioinformatics* 2005, **21**:2104-2105.
36. Huelsenbeck JP, Ronquist F: **MRBAYES: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17**:754-755.
37. Lartillot N, Lepage T, Blanquart S: **PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating.** *Bioinformatics* 2009, **25**:2286-2288.
38. Thorne JL, Kishino H, Painter IS: **Estimating the rate of evolution of the rate of molecular evolution.** *Mol Biol Evol* 1998, **15**:1647-1657.
39. Cox JC, Ingersoll JE, Ross SA: **A theory of the term structure of interest rates.** *Econometrica* 1985, **53**:385-407.
40. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A: **Relaxed phylogenetics and dating with confidence.** *PLoS Biol* 2006, **4**:699-710.
41. Benton MJ: *The fossil record 2* London: Chapman & Hall; 1993.
42. Dunlop JA, Selden PA: **Calibrating the chelicerate clock: a paleontological reply to Jeyaprakash and Hoy.** *Exp Appl Acarol* 2009, **48**:183-197.
43. Dunlop JA: **Geological history and phylogeny of Chelicerata.** *Arthropod Struct Dev* 2010, **39**:124-142.
44. Benton MJ, Donoghue PC: **Paleontological evidence to date the tree of life.** *Mol Biol Evol* 2007, **24**:26-53.
45. Crimes TP: **Trace fossils and correlation of late Precambrian and early Cambrian strata.** *Geol Mag* 1987, **124**:97-119.
46. Waloszek D, Dunlop J: **A larval sea spider (Arthropoda: Pycnogonida) from the Upper Cambrian 'Orsten' of Sweden, and the phylogenetic position of pycnogonids.** *Palaeontology* 2002, **45**:421-446.
47. Rudkin DM, Young GA, Nowlan GS: **The oldest horseshoe crab: a new xiphosurid from the Late Ordovician Konservat-Lagerstätten deposits, Manitoba, Canada.** *Palaeontology* 2008, **51**:1-9.
48. Selden PA, Gall JC: **A Triassic mygalomorph spider from the northern Vosges, France.** *Palaeontology* 1992, **35**:211-223.
49. Selden PA, Shear WA, Sutton MD: **Fossil evidence for the origin of spider spinnerets, and a proposed arachnid order.** *Proc Natl Acad Sci USA* 2008, **105**:20781-20785.
50. Lorenzini DM, Da-Silva PI Jr, Soares MB, Arruda P, Setubal J, Daffre S: **Discovery of immune-related genes expressed in hemocytes of the tarantula spider *Acanthoscurria gomesiana*.** *Dev Comp Immunol* 2006, **30**:545-556.
51. Van Bruggen EFJ, Bijlholt M, Schutter W, Wichertjes T, Bonaventura J, Bonaventura C, Lamy J, Lamy J, Leclerc M, Schneider H-J, *et al*: **The role of structurally diverse subunits in the assembly of three cheliceratan hemocyanins.** *FEBS Lett* 1980, **116**:207-210.
52. Weygoldt P, Paulus HF: **Untersuchungen zur Morphologie, Taxonomie und Phylogenie der Chelicerata.** *Zeitschrift für Zoologische Systematik und Evolutionsforschung* 1979, **17**:85-116.
53. Giribet G, Edgecombe GD, Wheeler WC, Babbitt C: **Phylogeny and systematic position of Opiliones: a combined analysis of chelicerate relationships using morphological and molecular data.** *Cladistics* 2002, **18**:5-70.
54. Obenchain FD, Oliver JH Jr: **The heart and arterial circulatory system of ticks (Acari: Ixodioidea).** *J Arachnol* 1976, **3**:57-74.
55. Immesberger A, Burmester T: **Putative phenoloxidases in the tunicate *Ciona intestinalis* and the origin of the arthropod hemocyanin superfamily.** *J Comp Physiol B* 2004, **174**:169-180.
56. Kusche K, Burmester T: **Diplopod hemocyanin sequence and the phylogenetic position of the Myriapoda.** *Mol Biol Evol* 2001, **18**:1566-1573.
57. Kusche K, Ruhberg H, Burmester T: **A hemocyanin from the Onychophora and the emergence of respiratory proteins.** *Proc Natl Acad Sci USA* 2002, **99**:10545-10548.
58. Brusca RC, Brusca GJ: *Invertebrates* Sunderland Mass: Sinauer Associates; 2003.
59. Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, *et al*: **Broad phylogenomic sampling improves resolution of the animal tree of life.** *Nature* 2008, **452**:745-749.
60. Hwang UW, Friedrich M, Tautz D, Park CJ, Kim W: **Mitochondrial protein phylogeny joins myriapods with chelicerates.** *Nature* 2001, **413**:154-157.
61. Mallatt JM, Garey JR, Shultz JW: **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* 2004, **31**:178-191.
62. Pisani D, Poling LL, Lyons-Weiler M, Hedges SB: **The colonization of land by animals: molecular phylogeny and divergence times among arthropods.** *BMC Biol* 2004, **2**:1.

63. Kadner D, Stollewerk A: **Neurogenesis in the chilopod *Lithobius forficatus* suggests more similarities to chelicerates than to insects.** *Dev Genes Evol* 2004, **214**:367-379.
64. Boore JL, Lavrov DV, Brown WM: **Gene translocation links insects and crustaceans.** *Nature* 1998, **392**:667-668.
65. Rota-Stabelli O, Telford MJ: **A multi criterion approach for the selection of optimal outgroups in phylogeny: recovering some support for Mandibulata over Myriochelata using mitogenomics.** *Mol Phylogenet Evol* 2008, **48**:103-111.
66. Wheeler WC, Hayashi CY: **The phylogeny of the extant chelicerate orders.** *Cladistics* 1998, **14**:173-192.
67. Podsiadlowski L, Braband A: **The mitochondrial genome of the sea spider *Nymphon gracile* (Arthropoda: Pycnogonida).** *BMC Genomics* 2006, **7**:284.
68. Zrzavy J, Hypsa V, Vlaskova M: **Arthropod phylogeny: taxonomic congruence, total evidence and conditional combination approaches to morphological and molecular data sets.** In *Arthropod Relationships.* Edited by: Fortey R, Thomas R. London: Chapman 1998:97-107.
69. Brenneis G, Ungerer P, Scholtz G: **The chelifores of sea spiders (Arthropoda, Pycnogonida) are the appendages of the deutocerebral segment.** *Evol Dev* 2008, **10**:717-724.
70. Hassanin A: **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* 2006, **38**:100-116.
71. Jeyaprakash A, Hoy MA: **First divergence time estimate of spiders, scorpions, mites and ticks (subphylum: Chelicerata) inferred from mitochondrial phylogeny.** *Exp Appl Acarol* 2009, **47**:1-18.
72. Weygoldt P: **Evolution and systematics of the Chelicerata.** *Exp Appl Acarol* 1998, **22**:63-79.
73. Shultz JW, Regier JC: **Phylogenetic analysis of arthropods using two nuclear protein-encoding genes supports a crustacean + hexapod clade.** *Proceedings of the Royal Society B: Biological Sciences* 2000, **267**:1011-1019.
74. Chen JY: **The sudden appearance of diverse animal body plans during the Cambrian explosion.** *Int J Dev Biol* 2009, **53**:733-751.

doi:10.1186/1471-2148-12-19
Cite this article as: Rehm *et al.*: The diversity and evolution of chelicerate hemocyanins. *BMC Evolutionary Biology* 2012 **12**:19.

Oligonucleotide Primers for Targeted Amplification of Single-Copy Nuclear Genes in Apocritan Hymenoptera

Gerrit Hartig[1,2], **Ralph S. Peters**[3], **Janus Borner**[4], **Claudia Etzbauer**[1], **Bernhard Misof**[1], **Oliver Niehuis**[1]*

1 Zoologisches Forschungsmuseum Alexander Koenig, Zentrum für Molekulare Biodiversitätsforschung, Bonn, Germany, **2** Universität Münster, Institut für Bioinformatik, Münster, Germany, **3** Zoologisches Forschungsmuseum Alexander Koenig, Abteilung Arthropoda, Bonn, Germany, **4** Universität Hamburg, Biozentrum Grindel und Zoologisches Museum, Hamburg, Germany

Abstract

Background: Published nucleotide sequence data from the mega-diverse insect order Hymenoptera (sawflies, bees, wasps, and ants) are taxonomically scattered and still inadequate for reconstructing a well-supported phylogenetic tree for the order. The analysis of comprehensive multiple gene data sets obtained via targeted PCR could provide a cost-effective solution to this problem. However, oligonucleotide primers for PCR amplification of nuclear genes across a wide range of hymenopteran species are still scarce.

Findings: Here we present a suite of degenerate oligonucleotide primer pairs for PCR amplification of 154 single-copy nuclear protein-coding genes from Hymenoptera. These primers were inferred from genome sequence data from nine Hymenoptera (seven species of ants, the honeybee, and the parasitoid wasp *Nasonia vitripennis*). We empirically tested a randomly chosen subset of these primer pairs for amplifying target genes from six Hymenoptera, representing the families Chrysididae, Crabronidae, Gasteruptiidae, Leucospidae, Pompilidae, and Stephanidae. Based on our results, we estimate that these primers are suitable for studying a large number of nuclear genes across a wide range of apocritan Hymenoptera (i.e., all hymenopterans with a wasp-waist) and of aculeate Hymenoptera in particular (i.e., apocritan wasps with stingers).

Conclusions: The amplified nucleotide sequences are (a) with high probability from single-copy genes, (b) easily generated at low financial costs, especially when compared to phylogenomic approaches, (c) easily sequenced by means of an additionally provided set of sequencing primers, and (d) suitable to address a wide range of phylogenetic questions and to aid rapid species identification via barcoding, as many amplicons contain both exonic and fast-evolving intronic nucleotides.

Citation: Hartig G, Peters RS, Borner J, Etzbauer C, Misof B, et al. (2012) Oligonucleotide Primers for Targeted Amplification of Single-Copy Nuclear Genes in Apocritan Hymenoptera. PLoS ONE 7(6): e39826. doi:10.1371/journal.pone.0039826

Editor: Keith A. Crandall, Brigham Young University, United States of America

Received March 22, 2012; **Accepted** May 27, 2012; **Published** June 29, 2012

Funding: This research was supported by a grant of the German Research Foundation to ON (DFG NI 1387/1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: o.niehuis.zfmk@uni-bonn.de

These authors contributed equally to this work.

Introduction

Targeted amplification of single-copy genes is still a cornerstone of molecular phylogenetics despite the emergence of phylogenomic approaches analyzing transcriptome data and entire genomes. PCR approaches have focused primarily on mitochondrial genes, rRNA, and a restricted number of nuclear genes [1–4]. The phylogenetic analysis of a set of these standard genes and the study of phylogenomic data both have their pros and cons: the few standard genes are comparatively easy to amplify across a wide range of species, but their phylogenetic signal may be insufficient to answer the research question(s) of interest. In contrast, phylogenomic approaches provide a plethora of nucleotide sequence data and facilitate addressing difficult phylogenetic questions. However, phylogenomic approaches are (still) expensive and may require specially treated sample material (e.g., for preservation of RNA), which means that material from most scientific collections cannot be used. Degenerate oligonucleotide PCR primers designed to amplify a large set of single-copy nuclear genes in species of interest could close the gap between the two approaches and could be a viable alternative to both of them. Here, we present such a suite of PCR primers for amplifying single-copy nuclear genes from Hymenoptera (sawflies, bees, wasps, and ants).

Hymenoptera are one of the mega-diverse insect orders and encompass more than 125,000 described species, many of which have key functions in ecosystems and are of fundamental economical, agricultural, and medical importance [5]. Given this importance, it is surprising how few molecular markers are currently in use for phylogenetic and evolutionary studies of Hymenoptera (e.g., [6–9]). Even the most recent comprehensive phylogenetic investigation of Hymenoptera used a PCR approach that targeted only four genes (18S, 28S, EF1α, COX1) [4]. Many important nodes in the resulting phylogeny are not robust, indicating that more nucleotide sequence data are required to answer these and other fundamental phylogenetic questions involving Hymenoptera. Additionally, only two phylogenomic studies have been published that analyze EST data from Hymenoptera, both with very limited taxon samples [10,11]. Peters and colleagues [12] combined all published sequence data of Hymenoptera for a comprehensive phylogenetic analysis. This study revealed that only about ten molecular markers are frequently used to tackle phylogenetic questions in the Hymenoptera. These markers have undoubtedly given important insights into the evolutionary history of this group. Nonetheless, their limited phylogenetic signal has also left many difficult and longstanding phylogenetic questions unresolved.

Genome sequence data offer new opportunities to establish markers for phylogenetic and evolutionary studies. This strategy has already successfully been pursued for fungi [13]. In Hymenoptera, nine genomes have been published (seven ants [14–19]; honeybee [20]; parasitoid wasp [21]). These genomes offer a rich and unexploited library of molecular markers for phylogenetic analyses.

There are three major advantages of establishing molecular markers for phylogenetic analyses from sequenced genomes compared to traditional approaches and to the exploration of EST data: (a) the ability to reliably assess the orthology of genes; (b) the ability to assess the probability of obtaining undesired secondary PCR products; and (c) the availability of gene models that inform about the position and length of introns and exons. One-to-one orthologous (single-copy) protein-coding genes can be identified with high confidence using orthology assessment software such as OrthoMCL [22]. When restricting oligonucleotide primer design to single-copy genes, the risk of accidentally sequencing pseudogenes and other paralogous genes is greatly reduced. If the main interest of a study lies in amplifying fast evolving sites, for example to address relationships within species or among closely related species, it is possible to focus on PCR primer pairs that maximize the amount of intronic sites in the PCR product. This kind of information cannot be inferred from EST data.

We present a suite of new degenerate oligonucleotide primers that are expected to amplify single-copy nuclear protein-coding genes in a taxonomically wide array of apocritan Hymenoptera (i.e., Hymenoptera with a wasp-waist). This lineage of Hymenoptera comprises the vast majority (>95%) of hymenopteran species [5]. We provide detailed primer statistics and a PCR protocol for rapidly assessing the functionality of primer pairs, and we show results from empirically testing ten randomly selected primer pairs on DNA from six Hymenoptera species, representing the families Chrysididae, Crabronidae, Gasteruptiidae, Leucospidae, Pompilidae, and Stephanidae. The targeted molecular markers can be used to address a wide range of phylogenetic and/or comparative evolutionary questions, may prove valuable for rapid species identification via barcoding, and can be easily generated at low financial costs.

Methods

Search for and Annotation of 1:1 Orthologous Genes

We searched for orthologous genes in the genomes of nine Hymenoptera: a parasitoid wasp (*Nasonia vitripennis*; Pteromalidae; assembly 1.0; OGS 1.2) [21], the honeybee (*Apis mellifera*; Apidae; assembly 2.0; OGS pre-release 2) [20], Jerdon's jumping ant (*Harpegnathos saltator*; Formicidae: Ponerinae; assembly 3.3; OGS 3.3) [14], the Argentine ant (*Linepithema humile*; Dolichoderinae; assembly 1.0; OGS 1.1) [15], the Florida carpenter ant (*Camponotus floridanus*; Formicinae; assembly 3.3; OGS 3.3) [14], the red harvester ant (*Pogonomyrmex barbatus*; Myrmicinae; assembly 3.0; OGS 1.1) [16], the red fire ant (*Solenopsis invicta*; Myrmicinae; assembly 1.0; OGS 2.2) [17], and two leaf-cutter ants (*Atta cephalotes*; Myrmicinae; assembly 4.0; OGS 1.1; *Acromyrmex echinatior*; Myrmicinae; assembly 1.0; OGS 1.0) [18,19].

Orthology of proteins between the nine genomes was inferred using a graph-based approach as implemented in OrthoMCL 2.0 [22]. This approach has been shown to have reasonably low false positive and false negative rates among the available methods to estimate gene orthology [23]. We only used sequence pairs from the 'orthologs.txt' output file for Markov clustering. The inflation value was set to 1.5. Finally, we extracted sets of 1:1 orthologs from the final OrthoMCL output file with the aid of a custom-made Perl script. The amino acids in each set of 1:1 orthologous proteins were aligned with MAFFT 6.833b [24,25] using the 'L-INS-I' alignment strategy. Note that we replaced the amino acid code 'U', which stands for selenocysteine and is not recognized by MAFFT, with the ambiguity code 'X' prior to alignment. The alignment was subsequently refined with MUSCLE 3.7 [26] using the refinement option. Each amino acid alignment was then used as a blueprint to align the nucleotides of the corresponding coding sequences with a custom-made Perl script and the BioPerl tool kit [27]. All sets of 1:1 orthologs were annotated by generating profile hidden Markov models (pHMMs) from the protein alignments. The pHMMs were used to search the official gene set (OGS) of the fruit fly *Drosophila melanogaster* (FlyBase release 5.22) [28] for the most similar sequence (E value $<10^{-10}$) with the HMMER 3.0 [29,30] software package. We also estimated the average nucleotide sequence divergence among the nine reference genomes for each amplified region by calculating Hamming distances (= uncorrected p-distances) using a custom-made Perl script.

Oligonucleotide Primer Design

All 4,145 multiple nucleotide alignments of 1:1 orthologous genes were searched for suitable primer binding sites using a custom-made Ruby script (Janus Borner, Christian Pick, Thorsten Burmester, unpublished). The script designs degenerate primers for PCR-amplification of coding sequences from the nuclear genome. It searches for conserved regions in aligned protein-coding nucleotide sequences and checks whether or not possible oligonucleotide primers that would bind at these conserved regions do not exceed a certain degree of degeneration, exhibit a GC content within a given range, and do not possess more than a given number of nucleotide repeats (Table 1). All primer pairs consistent with these criteria were searched for matches in the genomic nucleotide sequences of the nine reference species. This allowed estimating the actual length and the relative intron content of each amplicon. Primers that did not match because they bind at an exon/intron boundary or because they would amplify a region exceeding a pre-defined size (Table 1), were discarded. Approximate genomic matches were also considered to assess the probability of obtaining undesired secondary amplification products. To allow for direct sequencing of the PCR products using specific oligonucleotide sequencing primers, pre-designed oligonucleotides were added to the 5′ end of each primer sequence (Table 2). Finally, we evaluated the melting temperatures and hybridization energies of homo- and heterodimers for each pair of primers with the aid of UNAFold 3.8 [31]. All primer design parameters are summarized in Table 1.

Empirical Evaluation of Oligonucleotide Primer Pairs

Ten randomly chosen PCR primer pairs, each with the forward and reverse oligonucleotide primers of sequencing primer set HOG-Seq A (Table 2) attached to their 5′ ends, were tested for amplifying the target genes in six apocritan Hymenoptera:

Table 1. Oligonucleotide PCR primer design parameters.

Parameter	Minimum Value	Maximum Value
Amplicon length (bp)	300	1000
Primer length (bp)	20	25
Degree of degeneration	–	256
GC content (%)	20	80
Repeats of single nucleotide (bp)	–	4
Melting temperature (°C)	45	66
Difference of melting temperatures (°C)	–	10
dG of homodimer (kcal/mole)	−11.0	–
dG of heterodimer (kcal/mole)	−11.0	–
Degree of degeneration at 3′ end*	–	4
GC content (%) at 3′ end*	20	80
Repeats of single nucleotide (bp) at 3′ end*	–	3

*Terminal six nucleotides.
doi:10.1371/journal.pone.0039826.t001

Stephanus serrator (Stephanidae), *Leucospis dorsigera* (Leucospidae), *Gasteruption tournieri* (Gasteruptiidae), *Chrysis mediata* (Chrysididae), *Lestica alata* (Crabronidae), and *Episyron albonotatum* (Pompilidae). With Stephanidae, the possible sister group of all remaining Apocrita, and with representatives of the superfamilies Chalcidoidea (*Leucospis*), Evanioidea (*Gasteruption*), Chrysidoidea (*Chrysis*), Apoidea (*Lestica*), and Vespoidea (*Episyron*), our taxon sampling includes representatives of several deeply-divergent major lineages of the mega-diverse Hymenoptera (Figure 1). All taxa were collected by ON in Rhineland-Palatinate, Germany, in 2011 and were preserved in 96% ethanol.

DNA was extracted from thoracic muscle tissue using the QIAGEN DNeasy Blood & Tissue Kit and following the protocol for insects (QIAGEN GmbH, Hilden, Germany). DNA quality and quantity were assessed by running the extracted DNA on a 1.5% agarose gel and by analyzing the DNA with a NanoDrop 1000 Spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA). Polymerase chain reactions (PCRs) were run in 20 µl volumes consisting of 0.5× QIAGEN Q-Solution, 1× QIAGEN Multiplex PCR Master Mix (QIAGEN GmbH, Hilden, Germany), 0.8 µM of each oligonucleotide primer, and 50 ng DNA. The touch-down PCR temperature profile started with an initial denaturation and QIAGEN HotStarTaq DNA polymerase activation step at 95°C for 15 min., followed by 16 cycles of 95°C for 0.5 min., 60–45°C for 0.5 min., and 72°C for 1.5 min, followed by 20 cycles of 95°C for 0.5 min., 65 for 0.5 min., and 72°C for 1.5 min, followed by 10 min. at 72°C. Note that the annealing temperature (T_a) was decreased during the first 16 cycles by 1°C each cycle. All PCRs were run on a Biometra Whatman T3000 Thermocycler (Biometra GmbH, Göttingen, Germany). PCR products were separated on a 1.5% agarose gel, together with a Fermentas 100 bp Plus DNA Ladder (Fermentas GmbH, Sankt Leon-Rot, Germany). PCR products chosen for sequencing (i.e., the amplicons of the five best-performing PCR primer pairs) were purified with the QIAquick PCR Purification Kit (QIAGEN GmbH, Hilden, Germany) and sent to Macrogen Inc. (Amsterdam, Netherlands) for direct Sanger sequencing with the sequencing primers HOG-Seq-A-F and HOG-Seq-A-R (Table 2). Forward and reverse DNA strands were assembled to contigs, trimmed (to exclude the binding sites of the PCR and sequencing oligonucleotide primers), and aligned with Geneious Pro 5.4.6 [32] to the sequences of the nine Hymenoptera, from which the primer pairs were inferred.

Table 2. Oligonucleotide sequencing primer pairs.

Primer pair	Forward (5′ → 3′)	T_m	Reverse (5′ → 3′)	T_m
HOG-Seq-A	CAGTAGGTGCGTATGTCA	49.9	TGGTCAGTGGCTATTCGT	50.9
HOG-Seq-B	CGCTCATACACTTGGTTC	49.7	TCAGTCATCCTCACTTCG	50.3
HOG-Seq-C	ATACTAACTGGTGGAGCGAG	52.6	TCACTACATTACCGTATGAC	48.6
HOG-Seq-D	TCGGTCACATTGGGCTACT	54.5	CCTTGGGTCTTCGGCTTGA	56.5

The nucleotide sequences of the sequencing primers were attached as a binding site to the 5′ end of the degenerate oligonucleotide polymerase chain reaction (PCR) primers. Each of the oligonucleotide primers in Table S1 is compatible with at least one of the sequencing primers added to the 5′ end of the PCR primer. T_m = approximate melting temperature [°C].
doi:10.1371/journal.pone.0039826.t002

All new sequences generated in this study have been submitted to the European Nucleotide Archive (accession numbers HE612159–HE612181).

Results

Gene Orthology and Oligonucleotide Primer Design

Analyzing the official gene sets of the nine hymenopterans with sequenced genomes, we identified a total of 4,145 single-copy orthologous genes that were present in every species. Studying the multiple nucleotide sequence alignments of these 4,145 orthologous genes, we inferred 304 oligonucleotide primer pairs for amplifying a total of 154 single-copy nuclear protein-coding genes. The length of the inferred primers ranges between 20 and 25 nucleotides (avg. 21), their estimated T_m (approximate melting temperature) ranges between 44.6° and 65.7°C (avg. 53.5°C), and their degree of degeneration ranges between 1 and 192 (avg. 31). For each of the 154 genes, we inferred between 1 and 11 (avg. 2) primer pairs with a maximum overlap between amplicons of 50%. The total degree of degeneration of the primer pairs ranges

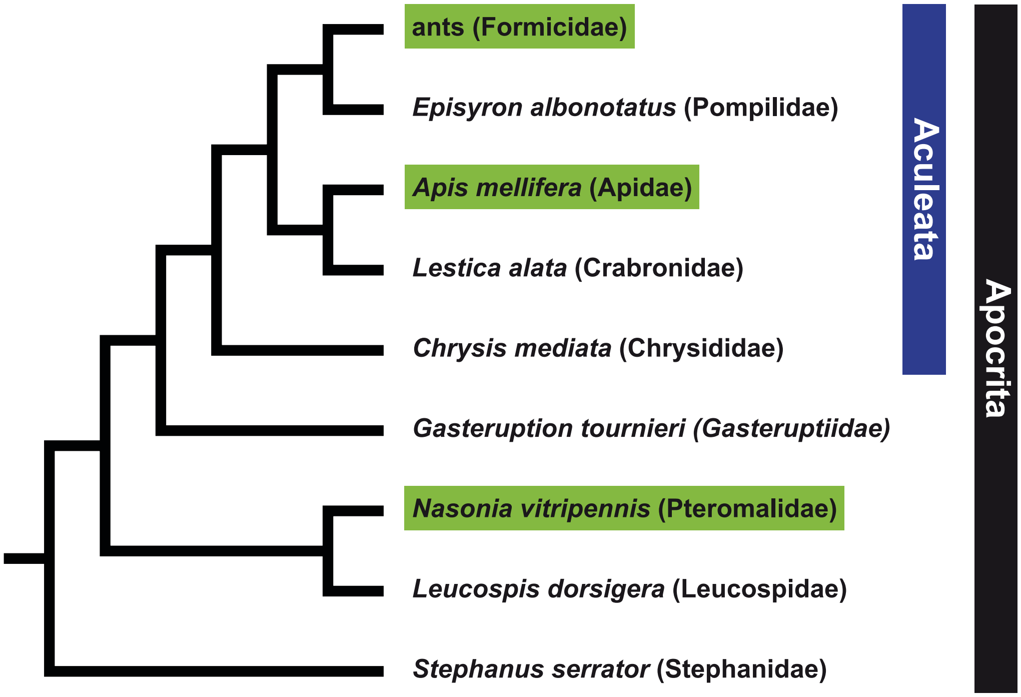

ants (Formicidae)
Episyron albonotatus (Pompilidae)
Apis mellifera (Apidae)
Lestica alata (Crabronidae)
Chrysis mediata (Chrysididae)
Gasteruption tournieri (Gasteruptiidae)
Nasonia vitripennis (Pteromalidae)
Leucospis dorsigera (Leucospidae)
Stephanus serrator (Stephanidae)
Aculeata
Apocrita

Figure 1. Hypothesized phylogenetic relationships of apocritan Hymenoptera studied in this investigation [4,12]. Taxa with sequenced genomes are highlighted in green; their genome sequences were analyzed to identify single-copy genes and to design degenerate oligonucleotide PCR primers. DNA of non-highlighted species was used to assess the functionality of the inferred PCR and sequencing primers.
doi:10.1371/journal.pone.0039826.g001

between 2 and 12,288 (avg. 860) (Table S1). The expected sizes of the PCR products range between 378 and 1,074 bp (avg. 683 bp; N = 2,736), the expected sizes of the amplified target regions range between 301 and 996 bp (avg. 604 bp; N = 2,736), and the average uncorrected (p) distance among the amplified target regions in the nine reference genomes ranges between 7.9% and 35.3% (avg. 17.3%; N = 304).

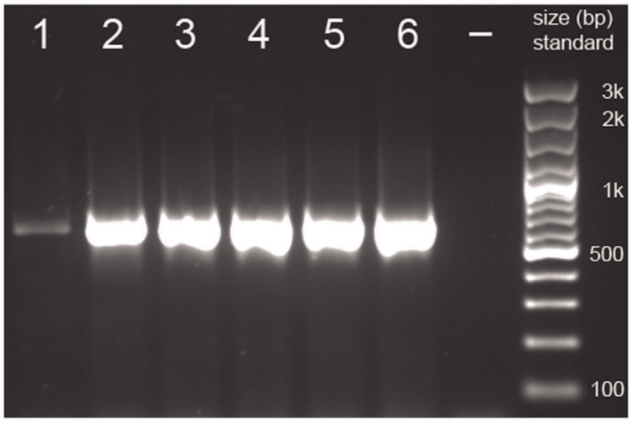


Figure 2. Polymerase chain reaction (PCR) products separated on 1.5% agarose gel. The depicted gel shows the PCR products obtained from using the inferred oligonucleotide primer pair 7229_02_A (Table 3) to PCR amplify DNA of *Stephanus serrator* (Stephanidae, 1), *Leucospis dorsigera* (Leucospidae, 2), *Gasteruption tournieri* (Gasteruptiidae, 3), *Chrysis mediata* (Chrysididae, 4), *Lestica alata* (Crabronidae, 5), and *Episyron albonotatum* (Pompilidae, 6). All PCR products were suitable for direct sequencing with the sequencing oligonucleotide primers HOG-Seq-A-F/−R (Table 2). – = negative control.
doi:10.1371/journal.pone.0039826.g002

Of the 304 inferred oligonucleotide primer pairs, 233 amplify genomic regions that according to the available gene models include at least one predicted intron in the reference genomes. These primer pairs amplify 112 (~73%) of the 154 genes. In contrast, only 28 inferred oligonucleotide primer pairs amplify genomic regions that do not include introns in the nine reference species. These primer pairs amplify 17 different (~11% of the here covered 154) genes. All remaining primer pairs amplify genomic regions that may or may not include introns. The number of exonic nucleotides in those 233 genomic target regions that include at least one predicted intron ranges from 154 to 814 (avg. 429; N = 2,097). The corresponding number of intronic nucleotides ranges from 45 to 653 (avg. 195; N = 2,097). The percentage of exonic nucleotides in the above mentioned 233 genomic regions ranges from 22.4 to 93.4 (avg. 69.8; N = 2,097). The number of exonic nucleotides in those 28 genomic regions that do not contain introns in any of the reference species ranges from 304 to 816 (avg. 454; N = 252).

Of the 304 inferred oligonucleotide primer pairs, we found 80 (referring to 71 different genes) to be compatible with sequencing primer pair HOG-Seq-A, 130 (referring to 107 different genes) to be compatible with sequencing primer pair HOG-Seq-B, 46 (referring to 46 different genes) to be compatible with sequencing primer pair HOG-Seq-C, and 73 (referring to 62 different genes) to be compatible with sequencing primer pair HOG-Seq-D (Table 2).

The complete list of inferred degenerate oligonucleotide primers, along with complementary information (e.g., annealing temperature, degree of degeneration, expected length of amplicons, compatibility with sequencing primers attached to the 5′ end of the PCR primers) is given in Table S1. Additional supplemen-

Table 3. Empirically evaluated degenerate oligonucleotide PCR primer pairs.

ID	Forward (5′ → 3′)	*d*	T_m min	T_m max	Reverse (5′ → 3′)	*d*	T_m min	T_m max	Total *d*
3683_01_A	GCYATYTTCGAYTTYGAYAG	32	46.0	56.8	AAVGTRAAKGATTCGTTGTA	12	47.5	54.4	384
4652_02_A	ATGATGTDGARTTTATMATACARAC	24	46.9	53.8	CWACRCTWATTTCTCTWTCAAC	16	47.1	51.9	384
4747_02_A	TTCTACGGBATGATCTTYAG	6	47.1	53.0	ACCTBGACATRATCTTVGGC	18	49.8	57.2	108
5119_01_A	GGDATYGTMGARGAGAGYGT	48	48.7	60.8	TYTTCATYTTRTCCATGTGYTC	16	48.9	56.5	768
5257_01_A	MACVAATAARTAYGGHTGYAGA	144	46.7	58.9	TAATTGGTCTARRTTGAARCT	8	47.0	52.7	1,152
5592_01_A	AAYTRAATAAAGACTGGAAAGAAGA	4	50.3	53.8	GTYARATCCATYCCRTGATC	16	47.6	55.7	64
5768_01_A	ACDGTHAARGTDTGGAATGC	54	48.6	58.2	GCWACCCAAATRCWAGWTTG	16	48.8	55.0	864
6917_01_A	ATGCCVTTCTACACRGTCTA	6	52.8	58.0	CYTCGCTYTTCTTCTGCATRTC	8	53.5	58.9	48
7036_02_A	TTTGTCWGYGKGTGCCTTGT	8	55.4	60.1	TTCATRGTWGCTTCRGTATCNGT	32	51.2	59.2	256
7229_02_A	TGCYTGATHCTSTTCTTCGT	12	51.1	55.8	TRTGRAAYCTRTGRAAGATGCA	32	49.6	58.6	384

The ten degenerate oligonucleotide primers were tested with the respective binding sites for sequencing primer HOG-Seq-A (see Table 2) attached to the 5′ end and used to amplify ten target genes in six apocritan Hymenoptera. *d* = degree of degeneration. T_m = approximate melting temperature [°C].
doi:10.1371/journal.pone.0039826.t003

tary material has been deposited in the Dryad data repository (http://datadryad.org/; doi:10.5061/dryad.d73k0).

Empirical Evaluation of Oligonucleotide Primer Pairs

All tested PCR primers had the oligonucleotides of the sequencing primer pair set HOG-Seq A attached to their 5′ ends (Table 2 and Table 3). One pair of tested primers produced amplicons suitable for direct sequencing in all six species (Tables 4 and 5, Figure 2). An additional five primer pairs produced amplicons suitable for direct sequencing in at least four of the six studied species, with a tendency to less reliably produce a PCR product suitable for direct sequencing with increasing evolutionary distance from ants (Tables 4 and 5, Figure S1). Overall, the PCR success rate when using DNA from species of Aculeata (i.e., apocritan wasps with stingers) was ~80%. When considering all Apocrita, the PCR success rate was still ~60%.

Discussion

We inferred 304 oligonucleotide primer pairs that can be used for PCR amplification of up to 154 different genes in apocritan Hymenoptera. The ten primer pairs that were empirically tested proved to be highly successful in amplifying the desired target DNA of Aculeata and showed a reasonable success-rate when applied to DNA of other Apocrita. Extrapolating these results and considering that we provide on average two primer pairs for a given gene, we expect up to 148 genes of interest to be amplifiable in aculeate Hymenoptera and roughly 110 to be amplifiable in many other groups of Apocrita. The high success-rate of our new PCR primers is most likely the result of the strict selection criteria that we applied during primer design (e.g., low potential for self-priming and the formation of hairpin loops, no alternative binding sites in the reference genomes). However, given that seven of the nine analyzed reference genomes are from ants, we expect fewer primers to amplify the desired product when they are applied to DNA of species that are distantly related to ants (e.g., non-aculeate Apocrita).

Table 4. Rating of obtained polymerase chain reaction (PCR) products.

Marker	*S. serrator*	*L. dorsigera*	*G. tournieri*	*C. mediata*	*L. alata*	*E. albonotatum*
3683_01_A	++*	+/−	++	++	++	++
4652_02_A	–	–	–	–	–	+
4747_02_A	–	++	+	++	+*	++
5119_01_A	–	++* (?)	–	++* (?)	++*	++*
5257_01_A	–	–	–	–	++	++
5592_01_A	–	++	+/−	++	++	++
5768_01_A	–	–	–	+/−*	+	–
6917_01_A	–	++	+	++	++	++
7036_02_A	+/−	++	++	++	++	++
7229_02_A	+	++	++	++	++	++

Rating of the PCR products obtained from using the degenerate oligonucleotide primers shown in Table 3 to amplify ten target genes in six apocritan Hymenoptera. ++ = target PCR product in excess. + = target PCR product sufficient for direct sequencing. +/− = target PCR product insufficient for direct sequencing. – = no target PCR product. (?) = unclear whether or not PCR products include amplicon of target gene.
*Secondary PCR amplification product likely hampering direct sequencing.
doi:10.1371/journal.pone.0039826.t004

Table 5. European Nucleotide Archive accession numbers for all sequences generated from primer testing.

Marker	*S. serrator*	*L. dorsigera*	*G. tournieri*	*C. mediata*	*L. alata*	*E. albonotatum*
3683_01_A	failed	NA	HE612159	HE612169	HE612174	HE612181
5592_01_A	NA	HE612164	NA	HE612170	HE612175	HE612180
6917_01_A	NA	HE612165	HE612167	HE612168	HE612177	failed
7036_02_A	NA	HE612166	HE612160	HE612171	HE612176	HE612179
7229_02_A	HE612162	HE612163	HE612161	HE612172	HE612173	HE612178

doi:10.1371/journal.pone.0039826.t005

There are several options for improving the PCR success-rate of the primers reported here. For example, while we used a touch-down temperature profile to rapidly assess the functionality of the ten evaluated primer pairs, one could instead use a PCR temperature profile with a constant annealing temperature that is close to the optimal annealing temperature of the specific primer pair (Table S1). Such a temperature profile could reduce the risk of obtaining secondary amplification products. Since we did not apply primer-specific PCR temperature profiles when empirically testing primer pairs, we expect their success-rate to be slightly underestimated. Researchers using these new primers should also consider increasing the concentration of oligonucleotides in the PCR mix to counterbalance the high degree of degeneration of some of the oligonucleotides (Table S1).

We calculated the average nucleotide sequence divergence among the nine reference genomes for the amplified region plus the absolute number of intronic and exonic nucleotides in the expected amplicon for each primer pair (Table S1). Consequently, users are able to search for markers that are more- or less-conserved than others, and users are additionally able to select for primers that specifically amplify genes with or without introns. Intronic DNA could prove highly valuable for resolving genealogical relationships of recently diverged lineages. These nuclear markers may also prove to be very useful for DNA barcoding. Overall, the ability to select genes that seem particularly suitable to address a specific research question makes the plethora of PCR primers presented here a highly valuable toolbox for research in apocritan Hymenoptera. Finally, the inferred primers are compatible with pre-designed oligonucleotides (Table 2) attached to their 5′ end. This allows users to select a single oligonucleotide sequencing primer pair from a set of four for sequencing all PCR products.

Our approach for designing oligonucleotides for PCR-amplification of orthologous genes in a wide range of species requires the availability of sequenced genomes. One group of insects, besides Hymenoptera, for which genomes of several taxa have been sequenced, and for which such an approach might prove fruitful, is Diptera. Genome sequences from more than 15 species of Diptera are currently available and those of many more are already in progress. As in Hymenoptera, however, there is a strong taxonomic bias: only genomes of fruit flies (*Drosophila* spp.) and of mosquitos (Culicidae) have been published. As these two taxa belong to two distantly related lineages that split early in the evolution of Diptera, the available genomes might nonetheless already reflect a significant proportion of the molecular diversity in Diptera. With the i5K initiative [33], we expect the number of sequenced insect genomes to explode in the very near future. This will likely allow the inference of large numbers of phylogenetic markers for many more insect orders.

Supporting Information

Figure S1 Polymerase chain reaction (PCR) products separated on 1.5% agarose gels. The depicted gels show the PCR products obtained from using the inferred oligonucleotide primer pairs **A.** 3683_01_A, **B.** 4652_02_A, **C.** 4747_02_A, **D.** 5119_01_A, **E.** 5257_01_A, **F.** 5592_01_A, **G.** 5768_01_A, **H.** 6917_01_A, and **I.** 7036_02_A (see Table 3) to PCR amplify DNA of **1.** Stephanus serrator (Stephanidae), **2.** Leucospis dorsigera (Leucospidae), **3.** Gasteruption tournieri (Gasteruptiidae), **4.** Chrysis mediata (Chrysididae), **5.** Lestica alata (Crabronidae), and **6.** Episyron albonotatum (Pompilidae). – = negative control. L = 100 bp ladder (see also Figure 2).
(TIF)

Table S1 Inferred degenerate oligonucleotide primers for studying single-copy nuclear genes in apocritan Hymenoptera.
(XLS)

Acknowledgments

We are grateful to Sina Herbertz for her initial tests of oligonucleotide primers that did not have tails for binding of oligonucleotide sequencing primers. Duane D. McKenna provided helpful comments on the manuscript.

Author Contributions

Conceived and designed the experiments: BM JB ON. Performed the experiments: CE GH JB. Analyzed the data: GH JB ON. Contributed reagents/materials/analysis tools: JB ON. Wrote the paper: BM GH JB ON RSP.

References

1. Brower AVZ, DeSalle R (1994) Practical and theoretical considerations for choice of a DNA sequence region in insect molecular systematics, with a short review of published studies using nuclear gene regions. Ann. Entomol. Soc. Am. 87: 702–716.
2. Simon C, Buckley TR, Frati F, Stewart JB, Beckenbach AT (2006) Incorporating molecular evolution into phylogenetic analysis, and a new compilation of conserved polymerase chain reaction primers for animal mitochondrial DNA. Annu. Rev. Ecol. Evol. Syst. 37: 545–579. doi:10.1146/annurev.ecolsys.37.091305.110018.
3. Wiegmann BM, Trautwein MD, Kim J-W, Cassel BK, Bertone MA, et al. (2009) Single-copy nuclear genes resolve the phylogeny of the holometabolous insects. BMC Biol. 7: 34. doi:10.1186/1741-7007-7-34.
4. Heraty J, Ronquist F, Carpenter JM, Hawks D, Schulmeister S, et al. (2011) Evolution of the hymenopteran megaradiation. Mol. Phylogenet. Evol. 60: 73–88. doi:10.1016/j.ympev.2011.04.003.
5. Grimaldi D, Engel MS (2005) Evolution of the insects. New York, NY: Cambridge University Press. 755 pp.

6. Dowton M, Austin AD (1994) Molecular phylogeny of the insect order Hymenoptera: apocritan relationships. Proc. Natl. Acad. Sci. U.S.A. 91: 9911–9915.
7. Dowton M, Austin AD (2001) Simultaneous analysis of 16S, 28S, COI and morphology in the Hymenoptera: Apocrita – evolutionary transitions among parasitic wasps. Biol. J. Linnean Soc. 74: 87–111. doi:l0.1006/bij1.2001.0577.
8. Desjardins CA, Regier JC, Mitter C (2007) Phylogeny of pteromalid parasitic wasps (Hymenoptera: Pteromalidae): initial evidence from four protein-coding nuclear genes. Mol. Phylogenet. Evol. 45: 454–469. doi:10.1016/j.ympev.2007.08.004.
9. Pilgrim EM, Von Dohlen CD, Pitts JP (2008) Molecular phylogenetics of Vespoidea indicate paraphyly of the superfamily and novel relationships of its component families and subfamilies. Zool. Scripta 37: 539–560. doi:10.1111/j.1463-6409.2008.00340.x
10. Sharanowski BJ, Robbertse B, Walker J, Voss SR, Yoder R, et al. (2010) Expressed sequence tags reveal Proctotrupomorpha (minus Chalcidoidea) as sister to Aculeata (Hymenoptera: Insecta). Mol. Phylogenet. Evol. 57: 101–112. doi:10.1016/j.ympev.2010.07.006.
11. Woodard SH, Fischman BJ, Venkat A, Hudson ME, Varala K, et al. (2011) Genes involved in convergent evolution of eusociality in bees. Proc. Natl. Acad. Sci. U.S.A. 108: 7472–7477. doi:10.1073/pnas.1103457108.
12. Peters RS, Meyer B, Krogmann L, Borner J, Meusemann K, et al. (2011) The taming of an impossible child: a standardized all-in approach to the phylogeny of Hymenoptera using public database sequences. BMC Biol. 9: 55. doi:10.1186/1741-7007-9-55.
13. Feau N, Decourcelle T, Husson C, Desprez-Loustau M-L, Dutech C (2011) Finding single copy genes out of sequenced genomes for multilocus phylogenetics in non-model fungi. PLoS ONE 6: e18803. doi:10.1371/journal.pone.0018803.
14. Bonasio R, Zhang G, Ye C, Mutti NS, Fang X, et al. (2010) Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. Science 329: 1068–1071. doi:10.1126/science.1192428.
15. Smith CD, Zimin A, Holt C, Abouheif E, Benton R, et al. (2011) Draft genome of the globally widespread and invasive Argentine ant (*Linepithema humile*). Proc. Natl. Acad. Sci. U.S.A. 108: 5673–5678. doi:10.1073/pnas.1008617108.
16. Smith CR, Smith CD, Robertson HM, Helmkampf M, Zimin A, et al. (2011) Draft genome of the red harvester ant *Pogonomyrmex barbatus*. Proc. Natl. Acad. Sci. U.S.A. 108: 5667–5672. doi:10.1073/pnas.1007901108.
17. Wurm Y, Wang J, Riba-Grognuz O, Corona M, Nygaard S, et al. (2011) The genome of the fire ant *Solenopsis invicta*. Proc. Natl. Acad. Sci. U.S.A. 108: 5679–5684. doi:10.1073/pnas.1009690108.
18. Nygaard S, Zhang G, Schiøtt M, Li C, Wurm Y, et al. (2011) The genome of the leaf-cutting ant *Acromyrmex echinatior* suggests key adaptations to advanced social life and fungus farming. Genome Res. 21: 1339–1348. doi:10.1101/gr.121392.111.
19. Suen G, Teiling C, Li L, Holt C, Abouheif E, et al. (2011) The genome sequence of the leaf-cutter ant *Atta cephalotes* reveals insights into its obligate symbiotic lifestyle. PLoS Genet. 7: e1002007. doi:10.1371/journal.pgen.1002007.
20. Weinstock GM, Robinson GE, Gibbs RA, Worley KC, Evans JD, et al. (2006) Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature 443: 931–949. doi:10.1038/nature05260.
21. Werren JH, Richards S, Desjardins CA, Niehuis O, Gadau J, et al. (2010) Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. Science 327: 343–348. doi:10.1126/science.1178028.
22. Li L, Stoeckert CJ Jr, Roos DS (2003) OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res. 13: 2178–2189. doi:10.1101/gr.1224503.
23. Chen F, Mackey AJ, Vermunt JK, Roos DS (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS ONE 2: e383. doi:10.1371/journal.pone.0000383.
24. Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33: 511–518. doi:10.1093/nar/gki198.
25. Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. Brief. Bioinformatics 9: 286–298. doi:10.1093/bib/bbn013.
26. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113. doi:10.1186/1471-2105-5-113.
27. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Res. 12: 1611–1618. doi:10.1101/gr.361602.
28. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, et al. (2000) The genome sequence of *Drosophila melanogaster*. Science 287: 2185–2195. doi:10.1126/science.287.5461.2185.
29. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14: 755–763. doi:10.1093/bioinformatics/14.9.755.
30. Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39: W29–37. doi:10.1093/nar/gkr367.
31. Markham NR, Zuker M (2008) UNAFold: software for nucleic acid folding and hybridization. Methods Mol. Biol. 453: 3–31. doi:10.1007/978-1-60327-429-6_1.
32. Drummond AJ, Ashton B, Buxton S, Cheung M, Cooper A, et al. (2011) Geneious. Auckland (New Zealand): Biomatters Ltd. p. Available: http://www.geneious.com.
33. Robinson GE, Hackett KJ, Purcell-Miramontes M, Brown SJ, Evans JD, et al. (2011) Creating a buzz about insect genomes. Science 331: 1386. doi:10.1126/science.331.6023.1386.

Molecular Phylogenetics and Evolution 61 (2011) 880–887

Contents lists available at SciVerse ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Dating the arthropod tree based on large-scale transcriptome data

Peter Rehm [a], Janus Borner [a], Karen Meusemann [b], Björn M. von Reumont [b], Sabrina Simon [c], Heike Hadrys [c], Bernhard Misof [b], Thorsten Burmester [a,*]

[a] *Biozentrum Grindel & Zoologisches Museum, Martin-Luther-King Platz 3, D-20146 Hamburg, Germany*
[b] *Zoologisches Forschungsmuseum Alexander Koenig, Adenauerallee 160, D-53113 Bonn, Germany*
[c] *Stiftung Tierärztliche Hochschule Hannover, ITZ, Ecology & Evolution, Bünteweg 17d, D-30559 Hannover, Germany*

ARTICLE INFO

Article history:
Received 18 March 2011
Revised 1 September 2011
Accepted 5 September 2011
Available online 17 September 2011

Keywords:
Arthropoda
Bayesian methods
Cambrian explosion
CIR process
Molecular clock

ABSTRACT

Molecular sequences do not only allow the reconstruction of phylogenetic relationships among species, but also provide information on the approximate divergence times. Whereas the fossil record dates the origin of most multicellular animal phyla during the Cambrian explosion less than 540 million years ago (mya), molecular clock calculations usually suggest much older dates. Here we used a large multiple sequence alignment derived from Expressed Sequence Tags and genomes comprising 129 genes (37,476 amino acid positions) and 117 taxa, including 101 arthropods. We obtained consistent divergence time estimates applying relaxed Bayesian clock models with different priors and multiple calibration points. While the influence of substitution rates, missing data, and model priors were negligible, the clock model had significant effect. A log–normal autocorrelated model was selected on basis of cross-validation. We calculated that arthropods emerged ~600 mya. Onychophorans (velvet worms) and euarthropods split ~590 mya, Pancrustacea and Myriochelata ~560 mya, Myriapoda and Chelicerata ~555 mya, and 'Crustacea' and Hexapoda ~510 mya. Endopterygote insects appeared ~390 mya. These dates are considerably younger than most previous molecular clock estimates and in better agreement with the fossil record. Nevertheless, a Precambrian origin of arthropods and other metazoan phyla is still supported. Our results also demonstrate the applicability of large datasets of random nuclear sequences for approximating the timing of multicellular animal evolution.

1. Introduction

Dating of diversification and speciation events is a major aim of evolutionary studies. For a long time, fossil remains were the prime source of such time estimates. The fossil record, however, is far from complete and in many cases the taxonomic assignment of fossil specimens is uncertain (Benton and Donoghue, 2007). DNA and protein sequences provide a complementary source of information for the inference of life history. Although there is an ongoing debate whether such a molecular clock approach is actually valid (Graur and Martin, 2004), many studies have obtained reasonable time estimates for a broad range of taxa (for review, see: Benton and Ayala, 2003; Hedges and Kumar, 2003).

In theory, molecular clock calculations have the power to be more precise than fossil dates because latter usually are underestimates. At best fossils provide an approximation to the oldest member of the taxon in question (cf. Benton and Ayala, 2003). In fact, sequence-derived dates tend to be older than the fossil dates (Hedges and Kumar, 2003). This is particularly true for deep divergence times. For example, the first conclusive fossil evidence for crown group bilaterians dates ~550–530 mya (Benton and Donoghue, 2007), but molecular estimates suggest an emergence of bilaterians between 1300 and 670 mya (e.g., Blair and Hedges, 2005; Lynch, 1999; Otsuka and Sugaya, 2003; Peterson et al., 2008). The discrepancy of molecular and fossil dates, and among different molecular clock approaches can be attributed to insufficient data, wrong taxonomic assignment or dating of fossils, and, most importantly, to rate heterogeneity among lineages over time and between genes (e.g., Adachi and Hasegawa, 1995; Benton and Donoghue, 2007; Bromham et al., 1998; Graur and Martin, 2004).

The undisputed fossil record of the phylum Arthropoda dates back to the early Cambrian period (Budd and Jensen, 2000; Budd and Telford, 2009; Edgecombe, 2010). The identity of possible representatives of arthropods from the earlier Ediacaran period is questionable (Nielsen, 2001). Based on fossils and geological

Abbreviations: mya, million years ago; myr, million years; LOG, log–normal autocorrelated clock model; CIR, Cox–Ingersoll–Ross model; UGM, uncorrelated gamma multipliers.

* Corresponding author. Address: Institute of Zoology and Zoological Museum, Biocenter Grindelm University of Hamburg, Martin-Luther-King-Platz 3, D-20146 Hamburg, Germany. Fax: +49 40 42838 3937.

E-mail address: thorsten.burmester@uni-hamburg.de (T. Burmester).

1055-7903/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2011.09.003

considerations, Benton and Donoghue (2007) assumed an earliest date of 581 mya for the divergence of Arthropoda and Nematoda. Molecular clock analyses, however, usually support much older time estimates that range from 1200 to 625 mya for the origin of Arthropoda (Blair, 2009; Blair et al., 2005; Douzery et al., 2004; Hausdorf, 2000; Lee, 1999; Sanders and Lee, 2010; Wang et al., 1999). Due to the lack of sequence data from important taxa, calculations of internal divergence times within arthropods are sparse, with the exception of the insects (e.g., Gaunt and Miles, 2002; Regier et al., 2004, 2005).

Here we analyze the divergence times of major arthropod taxa based on a superalignment spanning 37,476 amino acid positions, which had been derived from Expressed Sequence Tags (ESTs) (Meusemann et al., 2010). This is – to the best of our knowledge – the largest dataset that has ever been used for molecular clock studies.

2. Materials and methods

2.1. Sequence data and phylogenetic tree

In a previous study (Meusemann et al., 2010), 775 orthologous genes from 214 euarthropods, three onychophorans, two tardigrades, eight nematodes, three annelids and three mollusks, derived from EST data and selected genomes were identified with the HaMStR approach (Ebersberger et al., 2009). Single multiple protein alignments were generated with MAFFT L-INSI (Katoh and Toh, 2008). Alignment masking was conducted with ALISCORE and ALICUT (Kück et al., 2010; Misof and Misof, 2009). An optimal subset of data was selected by MARE 01-alpha (MAtrix REduction; http://mare.zfmk.de). The finally selected superalignment spans 37,476 amino acid positions, comprised 129 genes and 117 taxa, including 101 arthropods (available at TreeBase, http://www.treebase.org, under study accession no. S10507). A Bayesian phylogenetic tree was inferred with PhyloBayes (Lartillot et al., 2009). For details, refer to Meusemann et al. (2010).

2.2. Bayesian estimates of divergence times

The Bayesian phylogenetic majority rule consensus tree (Meusemann et al., 2010) was used as input for molecular clock estimates. The program PhyloBayes 3.2d was applied to calculate divergence times and 95% confidence intervals within a Bayesian framework (Lartillot et al., 2009). Three relaxed clock models, the log–normal autocorrelated clock model (LOG) (Thorne et al., 1998), the 'CIR' process (CIR) (Cox et al., 1985; see also Lepage et al., 2006) and uncorrelated gamma multipliers (UGM) (Drummond et al., 2006), were used under a uniform prior on divergence times for 50,000 cycles with a burn-in of 20,000 cycles. The CIR process is similar to the Ornstein–Uhlenbeck model, but with superior properties (Aris-Brosou and Yang, 2003; Lepage et al., 2006, 2007). Rates across sites were modeled assuming a discrete gamma distribution with four categories and with a Dirichlet process. Bayes factors were estimated using thermodynamic integration as implemented in PhyloBayes (Lepage et al., 2007) with 100,000 generations and a burn-in of 10,000. The three relaxed clock models were compared with the unconstrained model. Cross-validation of the models was performed by dividing the alignment into eight subsets (seven learning sets and one test set). Ten repetitions were run, as specified in PhyloBayses.

In a first approach, we tested the effects of gamma distributed priors for the root node on our results. To evaluate the impact of the priors, we defined different means and standard deviations (SD) of the prior distribution: mean 1000 mya (SD 1000 myr/500 myr) and 750 mya (SD 750 myr/325 myr), respectively. In addition, a uniform root prior was assumed with a maximum limit of 5000 mya imposed by PhyloBayes. All analyses were also run under the priors (*i.e.* with no data) to assess whether the prior distribution was sufficiently wide (*i.e.* non-informative). The results were compared with those obtained when the data were analyzed.

To assess the impact of missing data, all amino acid positions in the concatenated alignment were sorted according to their taxon coverage. Only the 50% of positions with the highest taxon coverage were used in a separate molecular clock analysis with the same settings as described above. The effect of substitution rates was tested by dividing the complete superalignment (129 genes) into three subsets, each containing 43 genes with *i* lowest, *ii* intermediate, and *iii* highest substitution rates. Genes were assigned to these categories according to the mean PAM distance of all possible sequence pairs within each alignment. To avoid artifacts due to missing data, only taxa for which sequences of all genes are present were selected for the assessment of pairwise distances: *Apis mellifera*, *Bombyx mori*, *Daphnia pulex*, *Drosophila melanogaster*, and *Tribolium castaneum*. Positions with gaps were ignored. All three subsets were analyzed in separate runs according to the procedure described above.

2.3. Calibration of the molecular clock

Seven calibration points were evenly distributed throughout the phylogenetic tree, including one calibration point within the outgroup (Table 1; Supplemental Table S1). We aimed to cover different regions of the tree and to include calibrations for deep nodes as well as for shallow nodes. To avoid a distortion of the time estimates by systematic misplacement of fossil calibration points, we used fossils with reliable systematic placement. Numerical ages were obtained from the International Stratigraphic Chart 2009 (http://www.stratigraphy.org), assuming the minimum age of the respective stage interval for calibration points 1–4 and 7 (Supplemental Table S1). The minimum age of calibration point 5 was dated according to the minimal age of the Namurian A/E1 (Dusar, 2006) and calibration point 6, for which a minimum and a maximum age was obtained from Benton and Donoghue (2007). Each of the settings described above was run with seven calibration points for each dataset. In addition, the complete dataset was analyzed with six calibration points (omitting calibration point 1 within the outgroup).

Calibration point 1: the minimum age of the divergence of Mollusca and Annelida is defined by small helcionelloids of the genus *Oelandiella* from the pre-Tommotian (Cambrian, Purella Biozone, Nemakit-Daldynian) period 528 mya (Gubanov and Peel, 1999; Khomentovsky and Karlova, 1993). Calibration point 2: evidence for euarthropods is provided by *Rusophycus*-like trace fossils from the early Tommotian 521 mya (Crimes, 1987). This calibration point is confirmed by recently described Lower Cambrian euarthropods fossils (Chen, 2009), which derives from the Maotianshan Shale (Qiongzhusian) dating 521–515 mya. Calibration point 3: the oldest unambiguous myriapod fossil is the millipede *Cowiedesmus eroticopodus* (Wilson and Anderson, 2004) from the Cowie Formation, Silurian. At that time, millipedes and centipedes had separated, thus *C. eroticopodus* provides a minimal age for the divergence of Diplopoda (millipedes) and Chilopoda (centipedes) at the transition from Wenlock to Ludlow 418.7 mya (base of Ludfordian, Ludlow). Calibration point 4: the split between Entognatha and Ectognatha (true insects) dates to the early Devonian (Pragian) period 404.2 mya, delimited by the first entognathan fossil of the springtail *Rhyniella precursor* (Whalley and Jarzembowski, 1981). Calibration point 5: the minimum date for the split between paleopteran and neopteran lineages is provided by an insect wing from the Upper Silesian Basin, Czech Republic (Béthoux and Nel, 2005), which dates to the Lower Carboniferous 324.8 mya and has been

Table 1
Evolutionary history of Arthropoda. Mean divergence times averaged over settings of the log–normal autocorrelated clock model (see Supplemental Table S2). The asterisks denote calibration points (for additional information, see Supplemental Table S1).

Split	Mean divergence time (mya)	Fossil record (upper bound) (mya)	References
Ecdysozoa–Lophotrochozoa	607	528 (Mollusca)	Khomentovsky and Karlova (1993)
Cycloneuralia–Arthropoda	601	521 (Arthropoda)	Chen (2009) and Crimes (1987)
Nematoda–Tardigrada	574	503 (Tardigrada)	Müller et al. (1995)
Onychophora–Euarthropoda*	589	521 (Euarthropoda)	Chen (2009) and Crimes (1987)
Myriochelata–Pancrustacea	562	Early Cambrian	Shear and Edgecombe (2010)
Myriapoda–Chelicerata	556	Early Cambrian	Shear and Edgecombe (2010)
Diplopoda–Chilopoda*	504	419 (Diplopoda)	Wilson and Anderson (2004)
Pycnogonida–Euchelicerata	546	501 (Pycnogonida)	Waloszek and Dunlop (2002)
Xiphosura/Araneae–Acari	496	445 (Xiphosura)	Rudkin et al. (2008)
Xiphosura–Araneae	473	445 (Xiphosura)	Rudkin et al. (2008)
Malacostraca/‘Maxillipoda’–Branchiopoda/Hexapoda	520	510 (Crustacea)	Harvey and Butterfield (2008)
Copepoda–Cirripedia/Malacostraca	507	505 (Cirripedia)	Collins and Rudkin (1981)
Cirripedia–Malacostraca	495	505 (Cirripedia)	Collins and Rudkin (1981)
Branchiopoda–Hexapoda	510	404 (Collembola)	Kukalová-Peck (1991)
Entognatha–Ectognatha*	485	404 (Collembola)	Kukalová-Peck (1991)
Archeognatha–Pterygota	455	390 (Archeognatha)	Labandeira et al. (1988)
Paleoptera–Neoptera*	419	325 (Archaeorthoptera)	Béthoux and Nel (2005)
Odonata–Ephemeroptera	388	318 (Odonata)	(Brauckmann and Schneider (1996)
Hemiptera–other neopterans	397	284 (Paleorrhyncha)	Shcherbakov (2000)
Orthoptera/’Blattodea’/Isoptera–Endopterygota (Holometabola)	391	307 (Coleoptera)	Béthoux (2009)
Orthoptera–‘Blattodea’/Isoptera	351	311(Blattodea)	Jarzembowski and Schneider (2007)
Isoptera–‘Blattodea’*	200	137 (Isoptera)	Engel et al. (2007)
Hymenoptera–Coleoptera/Lepidoptera/Diptera	372	307 (Coleoptera)	Béthoux (2009)
Coleoptera–Lepidoptera/Diptera	353	307 (Coleoptera)	Béthoux (2009)
Lepidoptera–Diptera	342	270 (Panorpida)	Minet et al. (2010)
Brachycera–Culicomorpha*	281	239 (Psychodomorpha)	Krzeminski et al. (1994)

assigned to Archaeorthoptera based on its venation (Prokop et al., 2005). Calibration point 6: the oldest fossil of a clade including Culicomorpha and Brachycera (Diptera) is *Grauvogelia arzvilleriana* from the middle Triassic Grés-a-Voltzia Formation of France, thereby providing a minimum age of 238.5 mya. The maximum age (295.4 mya) is defined by the insect fauna of the Boskovice Furrow, Czech Republic (Krzeminski et al., 1994). This deposit harbors a wide range of insects, but no representative of the brachyceran + culicomorphan clade has been described from this or from any older deposit (Benton and Donoghue, 2007). Dating of both, minimum and maximum constraints have been explained in Benton and Donoghue (2007). Calibration point 7: first evidence for the split of isopterans (termites) from other blattodeans dates to the Berriasian (lower Cretaceous) period 137.2 mya and is defined by the isopteran fossil *Baissatermes lapideus* (Engel et al., 2007).

3. Results

3.1. Molecular clock models

Molecular clock estimates were performed based on the Bayesian topology of the arthropod tree presented by Meusemann et al. (2010) (cf. Supplemental Fig. S1). Calculations of divergence times within the Bayesian framework were consistent within the different relaxed clock models (LOG, CIR and UGM; Supplemental Tables S2–S5). Estimates under the autocorrelated models (CIR and LOG) were similar, with CIR resulting in on average ~1% older dates (Table 2). The uncorrelated clock model (UGM) gave ~10% older dates than CIR or LOG. The choice of model for the rates across sites (gamma distribution or Dirichlet process) only had a minor effect on divergence time estimates, with mean absolute differences lower than 0.5% (Table 2).

The models were further compared by calculating the Bayes factors against the unconstrained model employing thermodynamic integration (Lartillot and Philippe, 2006), as implemented in PhyloBayes. The logarithms of the Bayes factors were 36.75 for UGM, 31.34 for CIR and 39.09 for LOG, suggesting that the log–normal autocorrelated clock model fits the data best. Cross-validation confirmed this conclusion and showed that the LOG model outperforms CIR and UGM. The cross-validation score of LOG vs. CIR was 2.25 ± 11.8506 and LOG vs. UGM was 7.75 ± 36.2138 (CIR vs. UGM: 5.5 ± 37.5167). Therefore, LOG was applied in the following analyses.

3.2. Bayesian inference of arthropod divergence times

The divergence times over the means of all calculations resulting from the LOG model (Table 1; Supplemental Table S2) and 95% confidence intervals (Supplemental Table S3) were displayed in a linearized tree (Fig. 1). We estimated that the divergence of Ecdysozoa and Lophotrochozoa occurred 629–590 mya (mean 607 mya). Separation of the clade leading to Nematoda and Tardigrada on the one hand, and Panarthropoda on the other, dated to 621–584 mya (mean 601 mya). The branch leading to Onychophora diverged from Euarthropoda 607–573 mya (mean 589 mya). Myriochelata (Myriapoda + Chelicerata) and Pancrustacea diverged during the Ediacaran 580–546 mya (mean 562 mya). The split between Myriapoda and Chelicerata occurred 573–539 mya (mean 556 mya) in the late Precambrian, the radiation of Pancrustacea commenced during the Cambrian period (533–500 mya, mean 520 mya). The split between hexapods and branchiopod crustaceans occurred 521–489 mya (mean 510 mya). Ectognathan insects split from Entognatha 498–465 mya (mean 485 mya). Winged insects (Pterygota) emerged about 455 mya (468–436 mya) and about 64 myr later holometabolous (endopterygote) insects appeared (400–379 mya, mean 391 mya). Running the analyses under the priors showed that the calculated divergence times were not biased by the selected priors (data not shown).

Table 2
Pairwise differences of divergence time estimates with different molecular clock models (LOG, log–normal autocorrelated clock model; CIR, Cox–Ingersoll–Ross model; UGM, uncorrelated gamma multipliers). Rates across sites were modeled according to a gamma distribution with four categories (Γ) or the Dirichlet process (D). Above the diagonal, absolute differences are given, below the diagonal are the mean differences. For the mean differences, positive values show older time estimates for the models in the horizontal row, negative values indicate younger dates.

	LOG Γ	LOG D	CIR Γ	CIR D	UGM Γ	UGM D
LOG Γ		0.004984	0.027751	0.028321	0.109461	0.109718
LOG D	0.003757		0.028923	0.029956	0.112200	0.112359
CIR Γ	−0.007211	−0.010987		0.003706	0.092873	0.090782
CIR D	−0.009793	−0.013702	0.002589		0.089166	0.089328
UGM Γ	−0.090723	−0.094126	−0.088827	−0.082692		0.001537
UGM D	−0.090971	−0.094367	−0.085352	−0.082959	−0.000293	

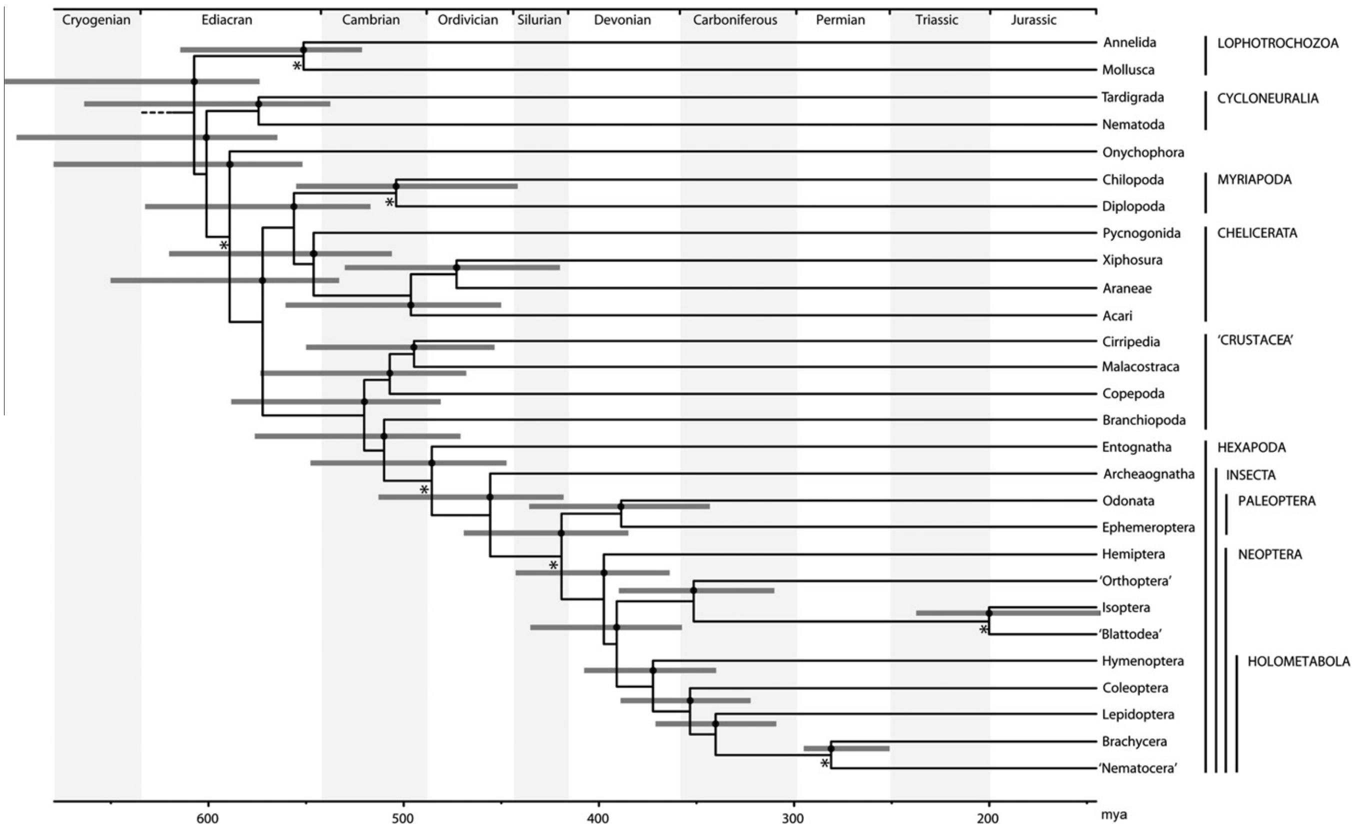


Fig. 1. Mean divergence times of major ecdysozoan taxa averaged over all estimations under the log–normal autocorrelated clock model shown in Supplemental Table S2. Gray bars indicate 95% mean confidence intervals (see Supplemental Table S3). Calibrated nodes are marked with an asterisk (see Supplemental Table S1 for calibration points). mya, million years ago.

3.3. Effect of root priors and calibration on time estimates

Bayesian time estimates were robust to changes in root priors, which determine age distributions and corresponding standard deviations for the root node. In a first approach, a uniformly distributed prior with an upper limit of the root of 5000 mya was assumed, as imposed by PhyloBayes (lane "–" in Supplemental Tables S2 and S3). To evaluate the impact of specific root priors, we defined different age means and standard deviations of the prior distribution: (a) 1000 mya and 1000 myr standard deviation, (b) 1000/500, (c) 750/750, and (d) 750/325. For all nodes, time estimates show low variation and differed – on average – less than 2% from the mean age with different root priors (standard deviation averaged over all nodes).

Standard deviations of the mean time estimates were lower with modeling the Dirichlet process (1.1%; root 1.8%) than with gamma distributed rates across sites (2.1%; root 2.2%). The exclusion of the outgroup calibration point, *i.e.* the minimum age of 521 mya for the split between Mollusca and Polychaeta, resulted in slightly (4.1%) younger divergence time estimates (not shown). The mean age of the root, *i.e.* the origin of Ecdysozoa, was about 5.7% younger using the reduced set of calibration points. Averaged over all calculations, the ecdysozoan divergence dated to 573 mya.

3.4. Effect of substitution rates on time estimates

To assess the effect of variations of substitution rates among genes, the full 129 gene dataset was subdivided into three data subsets with 43 genes each. The mean PAM distances of the genes in subset *i* ranged from 0.01 to 0.17 (slowest substitution rate), in subset *ii* from 0.17 to 0.30 (intermediate rate) and in subset *iii* from 0.31 to 0.80 (fastest rate). Averaged over data subsets of the autocorrelated LOG model, molecular clock analyses resulted in dates that were only ~0.1% younger compared to the complete dataset (Supplemental Table S6). In subset *i* the estimated dates were on average 2.9% younger (standard deviation 8.7%) as compared to the complete dataset. Subsets *ii* and *iii* showed on average slightly older dates (1.8% and 1.5%, respectively; standard deviation 6.5% and 6.9%, respectively). Most divergence times in the data subsets agree very well with the values derived from the complete dataset. Only two splits showed notable deviations. The split between Odonata and Ephemeroptera displayed high variation between subsets under all molecular clock settings, with subset *i* resulting in 22% younger dates and subsets *ii* and *iii* having 4% older dates. The results of the intermediate subset *ii* were similar with that of the entire dataset. The variations of divergence times of Isoptera and 'Blattodea' were even higher, with the intermediate subset *ii* resulting in 29% younger dates, whereas the slowly and fast evolving subsets showed older dates (subset *i*: 32%; subset *iii*: 35%).

3.5. Effect of missing data on time estimates

By removing 50% of the amino acid positions with the lowest taxon coverage, the relative amount of missing data in the alignment decreased from 51.0% to 38.5%. The effect of data reduction on the resulting divergence time estimates was minimal (Supplemental Table S7). Mean divergence times were 1.5% older than the results obtained from the complete dataset. Similar to the results from the subdivided datasets, the splits between Odonata–Ephemeroptera and Isoptera–'Blattodea' showed high variance. Under the autocorrelated clock models, divergence times of Odonata and Ephemeroptera were 7.7% older, whereas divergence time estimates of Isoptera and 'Blattodea' were 5.5% younger compared to the complete dataset.

4. Discussion

Molecular clock analysis has become a powerful tool based on a data source largely independent from the fossil record for the inference of divergence times of organisms. Still, there is much discrepancy between time estimates of different studies (e.g., Douzery et al., 2004; Peterson et al., 2008; Pisani et al., 2004). Factors that influence the outcome of molecular clock calculations are the sampling size, the selection of taxa and genes, rate heterogeneity, the suitability of the dating method, and the accuracy of calibration points. Simultaneous analyses of a large number of orthologous genes and application of multiple fossil calibration points provide more reliable estimates of divergence times if rate heterogeneity is considered (Thorne and Kishino, 2002; Yang and Yoder, 2003).

Due to the limited availability of orthologous genes, studies on large multi-gene datasets were usually restricted to only few taxa (e.g., Aris-Brosou and Yang, 2003; Blair and Hedges, 2005; Blair et al., 2005; Douzery et al., 2004; Gu, 1998; Wang et al., 1999). An alternative source of sequence data, which had not been applied to a molecular clock approach so far, is provided by EST data.

4.1. Applicability of EST data for molecular clock analyses

Our supermatrix with 129 orthologous genes and 117 taxa (Meusemann et al., 2010) is – to the best of our knowledge – the largest dataset that has ever been used for molecular clock studies (cf. Aris-Brosou and Yang, 2003; Blair and Hedges, 2005; Blair et al., 2005; Douzery et al., 2004; Gaunt and Miles, 2002; Gu, 1998; Lynch, 1999; Peterson et al., 2008; Regier et al., 2005; Wang et al., 1999). Our analyses are expected to provide more reliable estimates than the inference from few genes due to rate homogenization (Battistuzzi et al., 2010; Thorne and Kishino, 2002; Yang and Yoder, 2003). In addition, the essentially stochastic nature of ESTs further is expected to reduce sampling bias caused by the selection of specific genes.

A potential drawback of our approach may be the fragmentary nature of ESTs, which is reflected by 51% missing data in the concatenated superalignment (Meusemann et al., 2010). It has been demonstrated that large datasets are less sensitive to missing data (Philippe et al., 2004). In fact, when we reduced the amount of missing data by removing positions with low coverage, thereby increasing data density, only minimal changes in divergence time calculations were observed (Supplemental Table S7). Only two splits showed notable variation (Odonata–Ephemeroptera and Isoptera–'Blattodea'). Therefore, molecular clock analyses of our EST-dataset were essentially robust to the effect of missing data.

Another factor that may influence the outcome of a molecular clock approach are differences in substitution rates between genes. The *a priori* stochastic approach of obtaining ESTs is expected to result in large variations of rates. This is reflected by the PAM distances of the individual proteins ranging from 0.01 to 0.80. However, splitting the dataset into three subsets with different evolutionary rates shows little variation in divergence times between the three estimates (Supplemental Table S6). Only for the two splits mentioned above (*i.e.* Odonata–Ephemeroptera and Isoptera–'Blattodea'), the variance was notably large, which may be due to a bias introduced by gene selection. Therefore, we can conclude that – at least if the datasets are large enough – the effect of substitution rate differences in the EST dataset is low.

4.2. Molecular clock models

The Bayesian relaxed clock model approach to the arthropod EST-derived tree was also robust to the choice of priors and parameter settings. Neither the root priors, which specify the age of the root and its standard deviation, nor the model for the site-specific rates (discrete gamma distribution with four rate categories or the Dirichlet process) had significant effects (Table 2; Supplemental Tables S2, S4 and S5). The main factor that actually influenced the time estimates was the applied clock model.

In recent years, a whole range of different molecular clock methods that either rely on a maximum likelihood approach or on Bayesian methods have been proposed (for review, see Lepage et al., 2007). In initial tests, we applied a maximum likelihood, local clock approach, as implemented in the program r8s (Sanderson, 2002) to our data. However, the resulting time estimates were highly dependent on the age of the root and resulted in unreasonable divergence times (data not shown). The dates that derived from the Bayesian models were more consistent and not mutually exclusive, but still showed large differences in the calculated divergence times. We applied three different relaxed Bayesian clock models, two autocorrelated (LOG and CIR) and the uncorrelated UGM model. But which model is correct, *i.e.* fits best to the data and is thus expected to provide the best time estimate?

Both cross-validation and Bayes factors showed that the autocorrelated clock models were significantly better than UGM. Autocorrelation assumes that adjacent branches in a phylogenetic tree evolve with a similar rate, while in an uncorrelated model the individual rates cluster around the mean. Thus the assumption of autocorrelation of rates in related species appears to be more realistic than averaging rates across the branches. Our results also agree with the study by Lepage et al. (2007), who demonstrated that autocorrelated models outperform uncorrelated models, particularly when the dataset is large. Cross-validation also found that the LOG model was slightly better than the CIR process. Therefore, we discuss below only divergence times averaged over all LOG clock model settings (Table 1), as displayed in Fig. 1.

4.3. Arthropod origins and age of major arthropod taxa

While the fossil record suggests the emergence of the Metazoa during the Cambrian period 542–488 mya (e.g., Chen et al., 2004; Chen, 2009; Conway Morris, 1993; Crimes, 1987; Harvey and Butterfield, 2008; Shu et al., 1996), most molecular clock studies estimated much older dates of up to 1200 mya (e.g., Blair et al., 2005; Feng et al., 1997; Hausdorf, 2000; Lee, 1999; Nei et al., 2001; Peterson et al., 2008; Wang et al., 1999). We obtained notably younger estimates in our studies, which, however, still do not agree with a Cambrian origin of metazoan phyla. For example, we dated the earliest divergence time within the Arthropoda 589 mya, while the first unambiguous arthropod fossils are 521 myr old (Chen, 2009; Crimes, 1987). The gap between a possible Precambrian emergence and the Cambrian metazoan fossils may be explained by a period of cryptic evolution or detection bias, e.g., due to largely unexplored Early Cambrian and Pre-Cambrian Lagerstätten (Benton and Ayala, 2003; Conway Morris, 1993; Fortey et al., 1996; Valentine et al., 1991).

Traditionally, tardigrades have been joined with the arthropods (Brusca and Brusca, 2003). Recent molecular studies suggested a sister group relationship of tardigrades with Cycloneuralia (nematodes and allies) (Bleidorn et al., 2009; Lartillot and Philippe, 2008; Meusemann et al., 2010; Roeding et al., 2007). However, this topology has been discussed as an artifact due to long branch attraction (Rota-Stabelli et al., 2011). Previous molecular clock studies estimated an early origin of tardigrades 813–670 mya (Regier et al., 2005; Sanders and Lee, 2010), while we calculated that the divergence of Nematoda and Tardigrada took place during the Ediacaran (~575 mya). Although our estimate is in better agreement with the Cambrian fossils of crown-group tardigrades (Müller et al., 1995), it should be considered with caution because of the uncertain tardigrade relationships.

The closest arthropod relative of the myriapods is uncertain. While some molecular studies either suggested a sister group relationship of Myriapoda and Pancrustacea ("Mandibulata"; e.g., Giribet and Ribera, 2000; Regier et al., 2010; Rota-Stabelli et al., 2011), others provided evidence for a common clade of Myriapoda and Chelicerata ("Myriochelata"; Pisani et al., 2004; Roeding et al., 2009). Because Meusemann et al. (2010) recovered Myriochelata in their Bayesian approach, this topology was assumed here although it may be an artifact (Rota-Stabelli et al., 2011). We inferred that Myriochelata and Pancrustacea diverged 562 mya. This is notably younger than previous calculations based on topologies supporting Myriochelata, which ranged from 672–642 mya (Pisani et al., 2004; Regier et al., 2005). The branch that joins Myriapoda and Chelicerata is comparatively short, corresponding to ~15 million years (myr) with a large confidence interval (Fig. 1; Supplemental Table S3). Thus, a rapid divergence of the three clades Myriapoda, Chelicerata and Pancrustacea may explain at least in part the problems associated with the relationships among these taxa.

There is still no conclusive myriapod record from the Cambrian, but presence of fossils from putative sister group taxa (Crustacea, Chelicerata) strongly suggests a Cambrian or earlier origin of Myriapoda (Shear and Edgecombe, 2010). While previous studies date the emergence of Myriapoda more than 600 mya (Otsuka and Sugaya, 2003; Pisani et al., 2004; Regier et al., 2005), our estimates of myriapod origin are comparatively young (~556 mya). Within the Myriapoda, our results showed an age for the split of Diplopoda and Chilopoda of ~504 mya, which is slightly older than previous molecular studies (e.g., 442 mya; Pisani et al., 2004) and the fossil record (~420 mya; Edgecombe and Giribet, 2007).

Most studies agree that Pycnogonida (sea spiders) represent the earliest branch within Chelicerata (Meusemann et al., 2010; Regier et al., 2010; Roeding et al., 2009). A larval sea spider from the upper Cambrian (~500 mya) is the oldest fossil evidence for the split of Pycnogonida and Euchelicerata (Waloszek and Dunlop, 2002). Regier et al. (2005) suggested that this event took place 813–632 mya, but our calculation (~546 mya) is closer to the fossil record. Our estimate for the origin of Xiphosura (horseshoe crabs; ~473 mya) agrees well with the fossil dating, ~445 mya (Rudkin et al., 2008).

Traditionally, the divergence of Arachnida and Xiphosura has been considered the first split within the Euchelicerata (Regier et al., 2010; Weygoldt, 1998). However, several molecular studies did not recover monophyletic Arachnida, but suggest a basal position of the Acari (Meusemann et al., 2010; Roeding et al., 2007, 2009; Sanders and Lee, 2010). Given the low taxon sampling within the arachnids, it must remain uncertain which topology may reflect a true relationship or a long branch attraction phenomenon. This unresolved topology is the most likely explanation for the younger divergence times of Acari (~424 mya) and Araneae in other studies (~401–390 mya) (Aris-Brosou and Yang, 2002; Jeyaprakash and Hoy, 2009; Sanders and Lee, 2010).

The origin of the clade leading to Pancrustacea ('Crustacea' and Hexapoda) was previously estimated between 725 and 565 mya (Burmester, 2001; Otsuka and Sugaya, 2003; Pisani et al., 2004; Regier et al., 2005). Our results (~562 mya) are on the younger side. We estimated the divergence of the clade leading to the crustacean taxa Malacostraca, 'Maxillopoda' (Copepoda and Cirripedia) and Branchiopoda, and to the subphylum Hexapoda at ~520 mya in the early Cambrian. This timing is in line with recent findings of a crown group crustacean from the Mount Cap Formation 515–510 mya (Harvey and Butterfield, 2008).

The oldest known hexapods are collembolans (springtails) from the Lower Devonian ~400 mya (Kukalová-Peck, 1991). Previous molecular analyses estimated the split between crustaceans and hexapods (either Malacostraca–Hexapoda or Branchiopoda–Hexapoda) from 492–420 mya, but these studies relied on single or a limited number of genes (Burmester, 2001; Gaunt and Miles, 2002; Otsuka and Sugaya, 2003; Regier et al., 2005; Sanders and Lee, 2010). Based on an alignment of multiple genes, Pisani et al. (2004) proposed that the divergence of Hexapoda and Crustacea took place ~666 mya. Although our estimate (divergence of Branchiopoda and Hexapoda 510 mya) is closer to the hexapod fossil record, there is still a gap of ~100 myr. It must be considered that the true crustacean sister group of the Hexapoda is ambiguous. Recent studies have suggested that the enigmatic Remipedia may represent the closest living crustacean relatives of Hexapoda (Ertas et al., 2009; Regier et al., 2010). Unfortunately, fossil Remipedia are rare and ambiguous, and ESTs are currently not available for dating of the divergence from hexapods.

Within the insects, there is a general discrepancy between molecular time estimates and fossils. For example, we dated the divergence of Archeognatha and Pterygota ~455 mya, while first archeognath fossils derive from ~390 myr old (Labandeira et al., 1988) and first pterygotes from ~325 myr old strata (Prokop et al., 2005). Likewise, the time of the origin of Holometabola was estimated ~391 mya, while the first unambiguous fossils are ~307 mya (Béthoux, 2009). Our time estimates are actually close to previous calculations by (Gaunt and Miles, 2002), but the relatively large gap between molecular and fossil dating requires further investigations.

5. Conclusions

Although any molecular clock calculation for the inference of divergence times embraces problems beyond experimental control, we undertook measures to reduce potential errors to a minimum. The large amount of orthologous sequences from many arthropod species, and the application of a relaxed Bayesian clock model using evenly distributed calibration points yielded consistent molecular divergence time estimates. Missing data had only minor effect on the estimation of divergence times highlighting the suitability of ESTs for molecular clock analyses. Likewise, selection of three data subsets (from fast, intermediate or slow evolving genes) and different model priors had only negligible influence. The application of different models (uncorrelated vs. autocorrelated models) had notable effects on divergence time calculations. Along with errors in calibration points, inappropriate data and similar problems, unsuitable models may explain in part the unreasonably early divergence times obtained in some previous molecular clock studies. Our approach resulted in divergence time estimates of the arthropods that are generally in much better agreement with the fossil record.

Acknowledgments

We are grateful to Nicolas Lartillot for his help with PhyloBayes and J.-Wolfgang Wägele for his support. We are grateful to Falko Roeding for his help with the data and initial analyses, and Manfred Menning, Alexander Gubanov, Jacques Verniers and Chales H. Wellman for their support on stratigraphical issues. We thank Jakub Prokop for his advice on insect evolution. We acknowledge the associate editor and two anonymous reviewers for their helpful comments which improved the manuscript. This work has been supported by grants of the Deutsche Forschungsgemeinschaft within the Priority Project 1174 “Deep Metazoan Phylogeny” (Bu 956/8; Mi 649/6; Ha 1947/5; Wa 530/34).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2011.09.003.

References

Adachi, J., Hasegawa, M., 1995. Improved dating of the human/chimpanzee separation in the mitochondrial DNA tree: heterogeneity among amino acid sites. J. Mol. Evol. 40, 622–628.

Aris-Brosou, S., Yang, Z., 2002. Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny. Syst. Biol. 51, 703–714.

Aris-Brosou, S., Yang, Z., 2003. Bayesian models of episodic evolution support a late precambrian explosive diversification of the Metazoa. Mol. Biol. Evol. 20, 1947–1954.

Battistuzzi, F.U., Filipski, A., Hedges, S.B., Kumar, S., 2010. Performance of relaxed-clock methods in estimating evolutionary divergence times and their credibility intervals. Mol. Biol. Evol. 27, 1289–1300.

Benton, M.J., Ayala, F.J., 2003. Dating the tree of life. Science 300, 1698–1700.

Benton, M.J., Donoghue, P.C.J., 2007. Paleontological evidence to date the tree of life. Trends Ecol. Evol. 24, 26–53.

Béthoux, O., 2009. The earliest beetle identified. J. Paleont. 83, 931–937.

Béthoux, O., Nel, A., 2005. Some palaeozoic ‘protorthoptera’ are ‘ancestral’ orthopteroids: major wing braces as clues to a new split among the ‘protorthoptera’ (Insecta). J. Syst. Palaeontol. 2, 285–309.

Blair, J.E., 2009. Animals (Metazoa). In: Hedges, S.B., Kumar, S. (Eds.), The Timetree of Life. Oxford University Press, New York, pp. 223–230.

Blair, J.E., Hedges, S.B., 2005. Molecular phylogeny and divergence times of deuterostome animals. Mol. Biol. Evol. 22, 2275–2284.

Blair, J.E., Shah, P., Hedges, S.B., 2005. Evolutionary sequence analysis of complete eukaryote genomes. BMC Bioinform. 6, 53.

Bleidorn, C., Podsiadlowski, L., Zhong, M., Eeckhaut, I., Hartmann, S., Halanych, K.M., Tiedemann, R., 2009. On the phylogenetic position of Myzostomida: can 77 genes get it wrong? BMC Evol. Biol. 9, 150.

Brauckmann, C., Schneider, J., 1996. Ein unter-karbonisches Insekt aus dem Raum Bitterfeld-Delitzsch (Pterygota, Arnsbergium, Deutschland). N. Jb. Geol. Paläont. Mh. 1996, 17–30.

Bromham, L., Rambaut, A., Fortey, R., Cooper, A., Penny, D., 1998. Testing the Cambrian explosion hypothesis by using a molecular dating technique. Proc. Natl. Acad. Sci. USA 95, 12386–12389.

Brusca, R.C., Brusca, G.J., 2003. Invertebrates. Second ed. Sinauer Associates, Sunderland, MA.

Budd, G.E., Jensen, S., 2000. A critical reappraisal of the fossil record of the bilaterian phyla. Biol. Rev. (Camb) 75, 253–295.

Budd, G.E., Telford, M.J., 2009. The origin and evolution of arthropods. Nature 457, 812–817.

Burmester, T., 2001. Molecular evolution of the arthropod hemocyanin superfamily. Mol. Biol. Evol. 18, 184–195.

Chen, J.Y., 2009. The sudden appearance of diverse animal body plans during the Cambrian explosion. Int. J. Dev. Biol. 53, 733–751.

Chen, J.-Y., Bottjer, D.J., Oliveri, P., Dornbos, S.Q., Gao, F., Ruffins, S., Chi, H., Li, C.-W., Davidson, E.H., 2004. Small bilaterian fossils from 40 to 55 million years before the Cambrian. Science 305, 218–222.

Collins, D., Rudkin, D.M., 1981. *Priscansermarinus barnetti*, a probable lepadomorph barnacle from the Middle Cambrian Burgess Shale of British Columbia. J. Paleontol. 55, 1006–1015.

Conway Morris, S., 1993. The fossil record and the early evolution of the Metazoa. Nature 361, 219–225.

Cox, J.C., Ingersoll, J.E., Ross, S.A., 1985. A theory of the term structure of interest rates. Econometrica 53, 385–407.

Crimes, T.P., 1987. Trace fossils and correlation of late Precambrian and early Cambrian strata. Geol. Mag. 12, 97–119.

Drummond, A.J., Ho, S.Y.W., Phillips, M.J., Rambaut, A., 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4, 699–710.

Douzery, E.J.P., Snell, E.A., Bapteste, E., Delsuc, F., Philippe, H., 2004. The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils? Mol. Biol. Evol. 101, 15386–15391.

Dusar, M., 2006. Namurian. Geol. Belg. 9, 163–175.

Ebersberger, I., Strauss, S., Haeseler, A., 2009. HaMStR: profile hidden markov model based search for orthologs in ESTs. BMC Evol. Biol. 9, 157.

Edgecombe, G.D., 2010. Arthropod phylogeny: an overview from the perspectives of morphology, molecular data and the fossil record. Arthropod Struct. Dev. 39, 74–87.

Edgecombe, G.D., Giribet, G., 2007. Evolutionary biology of centipedes (Myriapoda: Chilopoda). Annu. Rev. Entomol. 52, 151–170.

Engel, M.S., Grimaldi, D.A., Krishna, K., 2007. Primitive termites from the Early Cretaceous of Asia (Isoptera). Stuttg. Beitr. Naturk. Ser. B (Geol. Paläontol.) 371, 1–32.

Ertas, B., von Reumont, B., Wägele, J.W., Misof, B., Burmester, T., 2009. Hemocyanin suggests a close relationship of Remipedia and Hexapoda. Mol. Biol. Evol. 26, 2711–2718.

Feng, D.F., Cho, G., Doolittle, R.F., 1997. Determining divergence times with a protein clock: update and reevaluation. Proc. Natl. Acad. Sci. USA 94, 13028–13033.

Fortey, R.A., Briggs, D.E.G., Wills, M.A., 1996. The Cambrian evolutionary ‘explosion’: decoupling cladogenesis from morphological disparity. Biol. J. Linn. Soc. 57, 13–33.

Gaunt, M.W., Miles, M.A., 2002. An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks. Mol. Biol. Evol. 19, 748–761.

Giribet, G., Ribera, C.A., 2000. A review of arthropod phylogeny: new data based on ribosomal DNA sequences and direct character optimization. Cladistics 16, 204–231.

Graur, D., Martin, W., 2004. Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. Trends Genet. 20, 80–86.

Gu, X., 1998. Early metazoan divergence was about 830 million years ago. J. Mol. Evol. 47, 369–371.

Gubanov, A.P., Peel, J.S., 1999. Oelandiella, the earlyest Cambrian helcionelloid mollusc from Siberia. Paleontology 42, 211–222.

Harvey, T.H.P., Butterfield, N.J., 2008. Sophisticated particle-feeding in a large Early Cambrian crustacean. Nature 452, 868–871.

Hausdorf, B., 2000. Early evolution of the bilateria. Syst. Biol. 49, 130–142.

Hedges, S.B., Kumar, S., 2003. Genomic clocks and evolutionary timescales. Trends Genet. 19, 200–206.

Jarzembowski, E.A., Schneider, J.W., 2007. The stratigraphical potential of blattodean insects from the late Carboniferous of southern Britain. Geol. Mag. 144, 449–456.

Jeyaprakash, A., Hoy, M.A., 2009. First divergence time estimate of spiders, scorpions, mites and ticks (subphylum: Chelicerata) inferred from mitochondrial phylogeny. Exp. Appl. Acarol. 47, 1–18.

Katoh, K., Toh, H., 2008. Recent developments in the MAFFT multiple sequence alignment program. Brief. Bioinform. 9, 286–298.

Khomentovsky, V.V., Karlova, G.A., 1993. Biostratigraphy of the Vendian–Cambrian beds and the lower Cambrian boundary in Siberia. Geol. Mag. 10, 29–45.

Krzeminski, W., Krzeminski, E., Papier, F., 1994. The oldest Polyneura (Diptera) and their importance to the phylogeny of the group. Acta Zool. Cracov. 37, 95–99.

Kück, P., Meusemann, K., Dambach, J., Thormann, B., Reumont, B., Wägele, J.W., Misof, B., 2010. Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. Front. Zool. 7, 10.

Kukalová-Peck, J., 1991. Fossil history and the evolution of hexapod structures. In: Naumann, I.D. (Ed.), The Insects of Australia, vol. 1, second ed. Melbourne University Press, Carlton, Australia, pp. 125–140.

Labandeira, C.C., Beall, B.S., Hueber, F.M., 1988. Early insect diversification: evidence from a Lower Devonian bristletail Quebéc. Science 242, 913–916.

Lartillot, N., Philippe, H., 2006. Computing Bayes factors using thermodynamic integration. Syst. Biol. 55, 195–207.

Lartillot, N., Philippe, H., 2008. Improvement of molecular phylogenetic inference and the phylogeny of bilateria. Philos. Trans. R. Soc. Lond. B Biol. Sci. 363, 1463–1472.

Lartillot, N., Lepage, T., Blanquart, S., 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25, 2286–2288.

Lee, M.S., 1999. Molecular clock calibrations and metazoan divergence dates. J. Mol. Evol. 49, 385–391.

Lepage, T., Lawi, S., Tupper, P., Bryant, D., 2006. Continuous and tractable models for the variation of evolutionary rates. Math. Biosci. 199, 216–233.

Lepage, T., Bryant, D., Philippe, H., Lartillot, N., 2007. A general comparison of relaxed molecular clock models. Mol. Biol. Evol. 24, 2669–2680.

Lynch, M., 1999. The age and relationships of the major animal phyla. Evolution 53, 319–325.

Meusemann, K., Reumont, B.M., Simon, S., Roeding, F., Strauss, S., Kück, P., Ebersberger, I., Walzl, M., Pass, G., Breuers, S., Achter, V., Haeseler, A., Burmester, T., Hadrys, H., Wägele, J.W., Misof, B., 2010. A phylogenomic approach to resolve the arthropod tree of life. Mol. Biol. Evol. 27, 2451–2464.

Minet, J., Huang, D.Y., Wu, H., Nel, A., 2010. Early Mecopterida and the systematic position of the Microptysmatidae (Insecta: Endopterygota). Ann. Soc. Entomol. Fr. 46, 262–270.

Misof, B., Misof, K., 2009. A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. Syst. Biol. 58, 21–34.

Müller, K.J., Walossek, D., Zakharov, A., 1995. 'Orsten' type phosphatized soft-integument preservation and a new record from the Middle Cambrian Kuonamka formation in Siberia. N. Jb. Geol. Paläont. Abh. 197, 101–118.

Nei, M., Xu, P., Glazko, G., 2001. Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. Proc. Natl. Acad. Sci. USA 98, 2497–2502.

Nielsen, C., 2001. Animal Evolution: Interrelationships of the Living Phyla. Oxford University Press, Oxford.

Otsuka, J., Sugaya, N., 2003. Advanced formulation of base pair changes in the stem regions of ribosomal RNAs; its application to mitochondrial rRNAs for resolving the phylogeny of animals. J. Theor. Biol. 222, 447–460.

Peterson, K.J., Cotton, J.A., Gehling, J.G., Pisani, D., 2008. The Ediacaran emergence of bilaterians: congruence between the genetic and the geological fossil records. Philos. Trans. R. Soc. Lond. B Biol. Sci. 363, 1435–1443.

Philippe, H., Snell, E.A., Bapteste, E., Lopez, P., Holland, P.W.H., Casane, D., 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol. Biol. Evol. 21, 1740–1752.

Pisani, D., Poling, L.L., Lyons-Weiler, M., Hedges, S.B., 2004. The colonization of land by animals: molecular phylogeny and divergence times among arthropods. BMC Biol. 2, 1.

Prokop, J., Nel, A., Hoch, I., 2005. Discovery of the oldest known Pterygota in the Lower Carboniferous of the Upper Silesian Basin in the Czech Republic (Insecta: Archeoptera). Geobios 38, 383–387.

Regier, J.C., Shultz, J.W., Kambic, R.E., 2004. Phylogeny of basal hexapod lineages and estimates of divergence times. Ann. Entomol. Soc. Am. 97, 411–419.

Regier, J.C., Shultz, J.W., Kambic, R.E., 2005. Pancrustacean phylogeny: hexapods are terrestrial crustaceans and maxillopods are not monophyletic. Proc. R. Soc. Lond. B. Biol. Sci. 272, 395–401.

Regier, J.C., Shultz, J.W., Zwick, A., Hussey, A., Ball, B., Wetzer, R., Martin, J.W., Cunningham, C.W., 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature 463, 1079–1083.

Roeding, F., Hagner-Holler, S., Ruhberg, H., Ebersberger, I., Haeseler, A., Kube, M., Reinhardt, R., Burmester, T., 2007. EST sequencing of Onychophora and phylogenomic analysis of Metazoa. Mol. Phylogenet. Evol. 45, 942–951.

Roeding, F., Borner, J., Kube, M., Klages, S., Reinhardt, R., Burmester, T., 2009. A 454 sequencing approach for large scale phylogenomic analysis of the common emperor scorpion (*Pandinus imperator*). Mol. Phylogenet. Evol. 53, 826–834.

Rota-Stabelli, O., Campbell, L., Brinkmann, H., Edgecombe, G.D., Longhorn, S.J., Peterson, K.J., Pisani, D., Philippe, H., Telford, M.J., 2011. A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. Proc. R. Soc. Lond. B. Biol. Sci. 278, 298–306.

Rudkin, D.M., Young, G.A., Nowlan, G.S., 2008. The oldest horseshoe crab: A new xiphosurid from Late Ordovician Konservat-Lagerstatten deposits, Manitoba, Canada. Paleontology 51, 1–9.

Sanders, K.L., Lee, M.S.Y., 2010. Arthropod molecular divergence times and the Cambrian origin of pentastomids. Syst. Biodivers. 8, 63–74.

Sanderson, M.J., 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol. Biol. Evol. 19, 101–109.

Shcherbakov, D.E., 2000. Permian Faunas of Homoptera (Hemiptera) in Relation to Phytogeography and the Permo-Triassic crisis. Paleontol. J. 34 (Suppl. 3), S251–S267.

Shear, W.A., Edgecombe, G.D., 2010. The geological record and phylogeny of the Myriapoda. Arthropod Struct. Dev. 39, 174–190.

Shu, D.G., Conway Morris, S., Zhang, X.L., 1996. A Pikaia-like chordate from the Lower Cambrian of China. Nature 157, 158.

Thorne, J.L., Kishino, H., 2002. Divergence time and evolutionary rate estimation with multilocus data. Syst. Biol. 51, 689–702.

Thorne, J.L., Kishino, H., Painter, I.S., 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15, 1647–1657.

Valentine, J.W., Awramik, S.M., Signor, P.W., Sadler, P.M., 1991. The Biological explosion at the Precambrian Cambrian boundary. Evol. Biol. 25, 279–356.

Waloszek, D., Dunlop, J.A., 2002. A larval sea spider (Arthropoda: Pycnogonida) from the Upper Cambrian 'Orsten' of Sweden, and the phylogenetic position of pycnogonids. Paleontology 45, 421–446.

Wang, D.Y., Kumar, S., Hedges, S.B., 1999. Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. Proc. R. Soc. Lond. B. Biol. Sci. 266, 163–171.

Weygoldt, P., 1998. Evolution and systematics of the Chelicerata. Exp. Appl. Acarol. 22, 63–79.

Whalley, P., Jarzembowski, E.A., 1981. A new assessment of Rhyniella, the earliest known insect, from the Devonian of Rhynie, Scotland. Nature 291, 317.

Wilson, H.M., Anderson, L.I., 2004. Morphology and taxonomy of Paleozoic millipedes (Diplopoda: Chilognatha: Archipolypoda) from Scotland. J. Paleontol. 78, 169–184.

Yang, Z., Yoder, A.D., 2003. Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene Loci and calibration points, with application to a radiation of cute-looking mouse lemur species. Syst. Biol. 52, 705–716.

BMC Biology
incorporating Journal of Biology

RESEARCH ARTICLE Open Access

The taming of an impossible child: a standardized all-in approach to the phylogeny of Hymenoptera using public database sequences

Ralph S Peters[1*], Benjamin Meyer[2], Lars Krogmann[3], Janus Borner[4], Karen Meusemann[1], Kai Schütte[5], Oliver Niehuis[1] and Bernhard Misof[1]

Abstract

Background: Enormous molecular sequence data have been accumulated over the past several years and are still exponentially growing with the use of faster and cheaper sequencing techniques. There is high and widespread interest in using these data for phylogenetic analyses. However, the amount of data that one can retrieve from public sequence repositories is virtually impossible to tame without dedicated software that automates processes. Here we present a novel bioinformatics pipeline for downloading, formatting, filtering and analyzing public sequence data deposited in GenBank. It combines some well-established programs with numerous newly developed software tools (available at http://software.zfmk.de/).

Results: We used the bioinformatics pipeline to investigate the phylogeny of the megadiverse insect order Hymenoptera (sawflies, bees, wasps and ants) by retrieving and processing more than 120,000 sequences and by selecting subsets under the criteria of compositional homogeneity and defined levels of density and overlap. Tree reconstruction was done with a partitioned maximum likelihood analysis from a supermatrix with more than 80,000 sites and more than 1,100 species. In the inferred tree, consistent with previous studies, "Symphyta" is paraphyletic. Within Apocrita, our analysis suggests a topology of Stephanoidea + (Ichneumonoidea + (Proctotrupomorpha + (Evanioidea + Aculeata))). Despite the huge amount of data, we identified several persistent problems in the Hymenoptera tree. Data coverage is still extremely low, and additional data have to be collected to reliably infer the phylogeny of Hymenoptera.

Conclusions: While we applied our bioinformatics pipeline to Hymenoptera, we designed the approach to be as general as possible. With this pipeline, it is possible to produce phylogenetic trees for any taxonomic group and to monitor new data and tree robustness in a taxon of interest. It therefore has great potential to meet the challenges of the phylogenomic era and to deepen our understanding of the tree of life.

Background

Reconstructing the phylogeny of organisms, the tree of life, is one of the major goals in biology and is essential for research in other biological disciplines ranging from evolutionary biology and systematics to biological control and conservation. In phylogenetics, molecular characters have become an indispensable tool, since they can be collected in a standardized and automated way. This is indicated by the exponential growth of published data, with a current doubling time of approximately 30 months [1] and expected massively accelerated data generation over the next several years. The sequencing of expressed sequence tags (ESTs), complete genomes and countless single-gene fragments has resulted in enormous, yet highly incomplete and unbalanced, data sets accessible via public databases such as the National Center for Biotechnology Information (NCBI) GenBank, the European Molecular Biology Laboratory (EMBL) and the DNA Database of Japan (DDBJ).

The accumulation of new data is, of course, important, but the potential of the currently available data for phylogenetic analysis has not yet been sufficiently explored.

* Correspondence: r.peters@zfmk.de
[1]Zoologisches Forschungsmuseum Alexander Koenig, Adenauerallee 160, D-53113 Bonn, Germany
Full list of author information is available at the end of the article

McMahon and Sanderson [2], Sanderson *et al.* [3] and Thomson and Shaffer [4] have published their attempts to use molecular data from public databases and to process them for phylogenetic analysis. However, these approaches, while valuable and trend-setting, did not offer thorough solutions and call for extension, improvements and updates in terms of generalization, detail, analysis and degree of automation. Any new approach must offer solutions to deal with data scarcity, poor data overlap, nonstationary substitution processes, base compositional heterogeneity and data quality deficits. In this study, we address these problems with a newly developed bioinformatics pipeline. We use a large exemplar taxon for which far more than 100,000 sequences have been published and show that comprehensive analyses can potentially deliver new results which were not available from each included data set separately.

As an exemplary taxon, we chose the insect order Hymenoptera, which comprises prominent groups such as bees, ants and wasps, the latter including the overwhelming armada of parasitoid species [5]. The Hymenoptera seem well-suited to demonstrate the power of our approach, since the taxon is megadiverse and offers a number of phylogenetic challenges, including many unresolved relationships and well-known problems that are associated with so-called long-branch taxa and rapid radiations (see, for example, [6-8]). Over a long period, comparatively few authors tried to resolve the phylogenetic relationships of the major lineages of Hymenoptera (see, for example, [9-16]). In recent years, however, interest and effort in solving higher-level relationships within the Hymenoptera have notably increased and led to the publication of an extensive analysis based exclusively on morphological characters [17], a study using complete mitochondrial genomes [18], a supertree approach using previously published trees [19], a phylogenetic estimate based on EST data [20] and a taxon-rich four-gene study [21]. In the past five years, complete nuclear genomes of several Hymenoptera species have been sequenced. Most noteworthy in this context are the genomes of the honey bee *Apis mellifera* [22] and the jewel wasp *Nasonia vitripennis*, with its sibling species *N. giraulti* and *N. longicornis* [23]. These genomes contributed significantly to the amount of sequence data available for Hymenoptera. However, their number is still too small to profitably augment phylogenetic analyses.

Overall, there are only few phylogenetic hypotheses on major lineages within Hymenoptera that are generally accepted. These are as follows: (1) “Symphyta” (sawflies) are paraphyletic, with the absence of the constriction between the first and second abdominal segments (that is, the wasp waist) as a symplesiomorphic character; (2) Apocrita (wasp-waisted wasps) are monophyletic (see, for example, [24]); (3) Xyelidae are sister group to all other Hymenoptera (see, for example, [25-27]); (4) Orussidae are sister group to Apocrita (see, for example, [17,18,27]) and (5) Aculeata (stinging wasps; Apoidea, Chrysidoidea and Vespoidea) are monophyletic (see, for example, [28]). In addition, most of the 22 currently recognized superfamilies are presumed to be monophyletic (see [29] for a synopsis). Numerous relationships within Hymenoptera are still unresolved. Among them, the most intriguing ones are the phylogeny of the major lineages within Apocrita, and in particular what the sister group of Aculeata is, and the monophyly and phylogeny of Proctotrupomorpha *sensu* Rasnitsyn 1988 [13] (Chalcidoidea, Cynipoidea, Diaprioidea, Mymarommatoidea, Platygastroidea and Proctotrupoidea).

In this study, we present a standardized, fast and transparent bioinformatics pipeline to collect, filter and analyze public sequence data deposited in GenBank. The pipeline is designed to be generally applicable in terms of taxa, genes and the variety of potential users. We apply this pipeline to sequences of Hymenoptera and discuss our results against the background of current hypotheses on two selected questions: the phylogeny of the major lineages within Apocrita and the monophyly and phylogeny of Proctotrupomorpha. Additionally, we use the results to diagnose persistent problems in the hymenopteran tree. Finally, we illustrate the merit of being able to easily generate trees from available sequence data at a time when data sets are accumulating at an ever-increasing speed.

Methods

We developed a bioinformatics pipeline that includes automated data retrieval, processing, filtering and analysis of sequence data using available programs in combination with newly developed scripts. The individual steps of the pipeline are illustrated in Figure 1. Those steps that are executed by new scripts are highlighted in blue. These scripts can be downloaded from http://software.zfmk.de/ or accessed as part of Additional file 1. They have been written in the Ruby or Perl programming language and will run on any Linux operating system. Each of our scripts comes with a manual that provides more detailed information on what it does and how to use it (manuals are available at http://software.zfmk.de/ and also are located in Additional file 1). Table 1 summarizes all new scripts and their respective tasks. To maneuver through the pipeline, each script has to be manually started with the output from the preceding step. This allows the user to manually interfere at each step or to modify the pipeline to adapt it to new demands. In the following paragraphs, we explain the individual steps of the pipeline using the example of the analysis of Hymenoptera sequences deposited in GenBank.

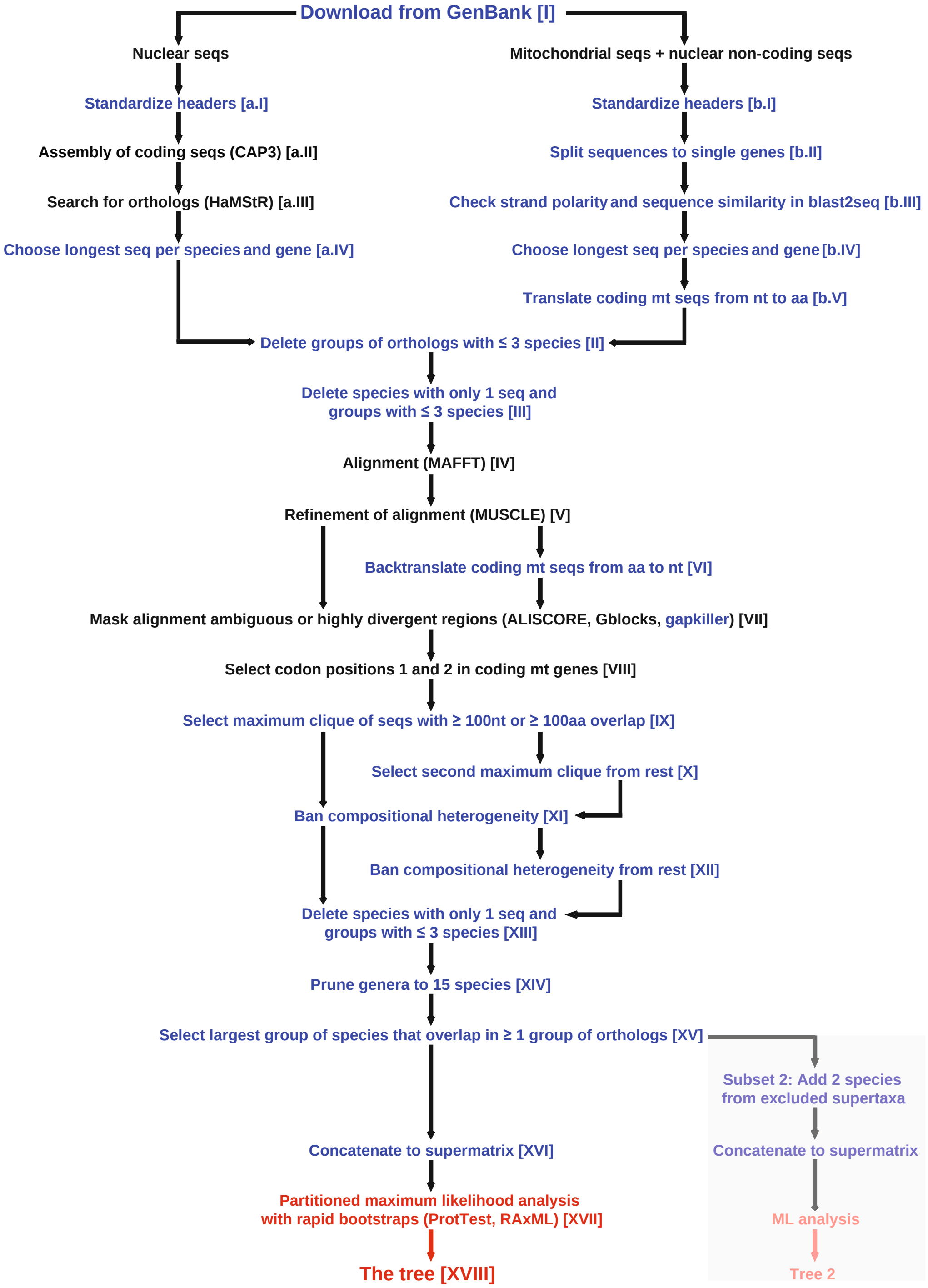

Download from GenBank [I]
Nuclear seqs
Mitochondrial seqs + nuclear non-coding seqs
Standardize headers [a.I]
Standardize headers [b.I]
Assembly of coding seqs (CAP3) [a.II]
Split sequences to single genes [b.II]
Search for orthologs (HaMStR) [a.III]
Check strand polarity and sequence similarity in blast2seq [b.III]
Choose longest seq per species and gene [a.IV]
Choose longest seq per species and gene [b.IV]
Translate coding mt seqs from nt to aa [b.V]
Delete groups of orthologs with ≤ 3 species [II]
Delete species with only 1 seq and groups with ≤ 3 species [III]
Alignment (MAFFT) [IV]
Refinement of alignment (MUSCLE) [V]
Backtranslate coding mt seqs from aa to nt [VI]
Mask alignment ambiguous or highly divergent regions (ALISCORE, Gblocks, gapkiller) [VII]
Select codon positions 1 and 2 in coding mt genes [VIII]
Select maximum clique of seqs with ≥ 100nt or ≥ 100aa overlap [IX]
Select second maximum clique from rest [X]
Ban compositional heterogeneity [XI]
Ban compositional heterogeneity from rest [XII]
Delete species with only 1 seq and groups with ≤ 3 species [XIII]
Prune genera to 15 species [XIV]
Select largest group of species that overlap in ≥ 1 group of orthologs [XV]
Subset 2: Add 2 species from excluded supertaxa
Concatenate to supermatrix [XVI]
Concatenate to supermatrix
Partitioned maximum likelihood analysis with rapid bootstraps (ProtTest, RAxML) [XVII]
ML analysis
The tree [XVIII]
Tree 2

Figure 1 Outline of our pipeline that processes GenBank sequence data for phylogenetic analysis. Steps that are executed by newly developed scripts are highlighted in blue, and external programs are written in parentheses after step description. Steps that directly refer to the phylogenetic analysis are highlighted in red. The additional procedure to infer subset 2 is shaded in gray.

Sequence data retrieval and data processing

We downloaded all sequences of Hymenoptera deposited in GenBank 172.0 (as of 18 August 2009) with the aid of the script *proseqco* [I] (Roman numerals in square brackets correspond to those in Figure 1). The script searched for the query taxon (Hymenoptera) in the nucleotide and in the EST database of GenBank (NCBI) and stored the sequences of each species in a separate

Table 1 New scripts used in our pipeline[a]

Step	Number	Script
Download from GenBank	[I]	*proseqco*
Standardize headers	[a.I], [b.I]	*header_standardizer*
Split sequences to single genes	[b.II]	*multiple_sequence_splitter*
Check strand polarity and sequence similarity	[b.III]	*checking_seq*
Choose longest sequence per species and gene	[a.IV], [b.IV]	*choose_longest_seq*
Translate coding mitochondrial sequences from nucleotides to amino acids	[b.V]	*dna2aa*
Delete groups of orthologs with three or fewer species	[II], [III], [XIII]	*small_groups_deleter*
Delete species with only one sequence	[III], [XIII]	*taxon_deleter*
Backtranslate coding mitochondrial sequences from amino acids to nucleotides	[VI]	*aa2dna*
Mask gappy regions in alignment	[VII]	*gap_killer*
Select maximum clique of overlapping sequences	[IX], [X]	*minimum_sequence_overlap*
Ban compositional heterogeneity	[XI], [XII]	*nucleotide_chi*
Prune genera to best represented species	[XIV]	*prune_genera*
Select largest group of species that overlap in at least one group of orthologs	[XV]	*reduce2leading_gene*
Concatenate alignments	[XVI]	*concatenator*

[a] Available at http://software.zfmk.de/ and in Additional file 1. All scripts were written in Ruby, except for *checking_seq*, which was written in Perl. Numerals (column "Number") correspond to those in Figure 1.

Fasta file. Mitochondrial sequences plus nuclear noncoding sequences (ITS1, ITS2 and nuclear rRNA) (right path b) and all other nuclear sequences (left path a) were retrieved in two separate downloads. For outgroup comparison, we additionally retrieved sequence data of the transcriptome, the nuclear noncoding genes and the complete mitochondrial genome of *Bombyx mori* (Lepidoptera), *Aedes aegypti* (Diptera) and *Tribolium castaneum* (Coleoptera). The gi numbers of all downloaded sequences are listed in Additional file 2.

Left path a

The nuclear sequences were assembled into contigs for each species using the sequence assembly program CAP3 [30] [a.II]. Orthologous sequences were identified using HaMStR 1.3 [31] [a.III]. We used the Insecta core set (available at http://www.deep-phylogeny.org/hamstr/download/datasets/hmmer2/) to build hidden Markov models (default settings). The genome of *A. mellifera* was chosen for the reciprocal BLAST search [31]. (If sequences of other taxa are processed, a different core set and a different species for the reciprocal BLAST search will have to be selected.) We chose HaMStR as the currently most practicable tool to automatically assign orthology among nucleotide and EST sequence data. During the HaMStR orthology prediction, all nucleotide sequences are translated into the corresponding amino acid sequences.

Right path b

The mitochondrial sequences and the nuclear noncoding sequences deposited in GenBank often include regions that span more than just one gene. In these instances, the script *multiple_sequence_splitter* uses information from the corresponding GenBank file to split sequences into fragments that correspond to single genes; that is, it creates multiple sequence files of single genes [b.II]. This step was serially applied to each file that we obtained from the previous step by means of a shell script. (See the *multiple_sequence_splitter* manual for a description of how to do this. Any step of the pipeline that had to be serially applied to a set of files was executed by means of a similar shell script [a.I, a.IV, b.I, b.II, b.IV, b.V, IV, V, VI, VII, IX, X, XI and XII].) In each of the obtained files, we used the script *checking_seq* to check for consistent strand polarity and overall similarity between sequences [b.III]. This was done to revert sequences with deviating strand polarity, to exclude wrongly annotated sequences and to ensure that all sequences in a single-gene file were orthologous. The script *checking_seq* compares a template of a gene with all the sequences of the single-gene files that were created in step [b.II] in blast2seq [32]. The identity (blast2seq results) between template and target sequence had to be more than 15 nucleotides. Otherwise, the reverse complement of the target sequence was checked, and hits were reverted. If identities were still below the match threshold, the target sequences were compared with a second, third or fourth template. Primary templates were taken from *A. mellifera*. (If sequences of other taxa are processed, other templates will have to be selected.) We randomly selected sequences from previously successfully checked sequences as subsequent templates. A maximum of four templates were used before we finally discarded a sequence. Then, to prepare the remaining sequences for the subsequent alignment, all coding mitochondrial sequences were translated from nucleotide to corresponding amino acid sequences with the aid of the script *dna2aa*, which uses the respective

GenBank information for this task [b.V]. Steps b.IV and b.V of our pipeline are automatically consecutively executed when using the script batch1_bIVtobV.sh. (See manual of batch scripts for details.)

Both paths

Sequence headers of all sequences were standardized to ">species,family,gi no." with the aid of the script *header_standardizer*, which uses the data included in the GenBank entries [a.I and b.I]. If multiple sequences were available for a given species and gene after respective steps [a.I to a.III] and [b.I to b.III], we chose the longest sequence from the unaligned multiple sequence files [a.IV and b.IV]. This was done by using the script *choose_longest_seq*.

Converged paths

We obtained numerous groups of orthologous sequences from path a and path b. Groups of orthologs that comprised three or fewer species were deleted by the script *small_groups_deleter* [II]. To increase data density, we discarded all species with only a single sequence in the data set by using the script *taxon_deleter* and again deleted groups of orthologs with three or fewer species by using *small_groups_deleter* [III].

Multiple sequence alignment and alignment masking

Orthologous sequences were aligned with MAFFT v6.712b using the auto option [IV]. Depending on the size of an alignment, MAFFT automatically chooses a suitable alignment option, such as L-INS-i for < 200 sequences and FFT-NS-2 for > 2,000 sequences [33,34]. All alignments were subsequently refined with the refinement option in MUSCLE version 3.7 [35] [V]. These are powerful alignment tools that allow processing very large data sets in reasonable time. Steps II through VI of our pipeline are automatically consecutively executed when using the script batch2_IItoVI.sh. (See the manual of batch scripts for details.) Aligned and refined mitochondrial amino acid sequences were then translated back into nucleotide sequences with the aid of the script *aa2dna*, which uses the corresponding reading frame information from the GenBank file [VI]. From this point on, we proceeded with nucleotide sequences for all mitochondrial sequences and nuclear noncoding sequences, as well as with amino acid sequences for the nuclear coding sequences (available since step [a.III]).

Ambiguously aligned or highly diverged regions of the alignment were masked with three different algorithms [VII]. We applied ALISCORE [36,37] and ALICUT [38] for noncoding nucleotide sequences and for nuclear amino acid sequences (default settings). Since the multiple sequence alignment of 28S rRNA was too big to be processed with ALISCORE, we used Gblocks 0.91b [39,40] for 28S instead (block parameter settings: (1) number of included seq/2 = 1020, (2) 1020, (3) 5, (4) 10, and (5) all). Finally, we used the script *gapkiller* to identify and delete sites with more than 70% gaps in coding mitochondrial sequences. Then we masked all third codon positions of mitochondrial coding sequences [VIII] and concatenated all tRNA alignments to one single alignment.

Species and sequence subset selection

In each group of orthologous sequences, we selected the largest group of species in which the sequences of all species overlap in at least 100 nucleotide or amino acid positions [IX]. This was done with the aid of the script *minimum_sequence_overlap*. The script applies a maximum clique algorithm. Generally, a maximum clique search is a way to find the largest group of items that fulfill a certain pairwise criterion. (See Additional file 3 for a short introduction to maximum cliques.) This approach is the formal solution to guarantee that our overlap criterion is fulfilled. Species that were not included in this first maximum clique were considered again in a search for a second maximum clique using the same criteria and the same script as before [X]. So, for each gene, we retained two separate files with groups of orthologous sequences: the first and the second maximum clique, respectively. Sequences that were not included in either of the maximum cliques were discarded.

To identify sequences that showed compositional heterogeneity in each group of orthologous nucleotide sequences, we used the script *nucleotide_chi*. The script applies a χ^2 test (test procedure identical to the χ^2 test implemented in TREE-PUZZLE [41]) and proceeds with excluding sequences with a base composition that significantly deviates until all sequences show compositional homogeneity [XI]. Since excluded sequences could comprise another set of homogeneous sequences, they were again tested with the same procedure as before to obtain a second group of sequences with compositional homogeneity [XII]. Sequences that did not end up in either of the two groups with compositional homogeneity were discarded. After discarding numerous sequences in steps IX through XII, we again excluded species with only one sequence in the data set by using the script *taxon_deleter* and groups of orthologs with three or fewer species by using the script *small_groups_deleter* [XIII]. Next, we pruned species-rich genera to the 15 species that were best represented in the data set by using the script *prune_-genera*. The representation criteria were, in this order, (1) the number of sequences in the data set and (2) the overall length of the sequence in the data set [XIV].

In a final subset selection step, we ensured that all species to be included in this subset overlap in at least one gene fragment of at least 100 nucleotide or amino acid positions [XV]. With the aid of the script *reduce2-leading_gene*, we pruned the data set to those species that were present in the most sequence-rich group of orthologs. This was the largest group of species that fulfilled the overlap criterion. In case of Hymenoptera, this group was a group of COX1 sequences. All corresponding sequences were concatenated with the script *concatenator* to one supermatrix. This supermatrix is referred to as "subset 1" [XVI]. Steps IX through XVI of our pipeline are automatically consecutively executed when using the script batch3_IXtoXVI.sh. (See manual of batch scripts for details.) In addition to subset 1, we selected a second subset. To accomplish this, we made concessions to systematic considerations and added to subset 1 representatives of Hymenoptera families that were excluded by any of the previous filtering steps. If more than two species of the respective families were available, we selected the two best-represented species using criteria identical to those in step [XIV]. With those sequences reincluded in the respective groups of orthologs, the tests for compositional heterogeneity (as described in step [XI]) were repeated and all sequences were finally concatenated to a supermatrix. This supermatrix is referred to as "subset 2."

Tree reconstruction

Phylogenetic inference of subset 1 and of subset 2 was done under the maximum likelihood (ML) optimality criterion in partitioned analyses with RAxML 7.2.8 [42,43] under the GTRCAT model. Analyses were computed on HPC Linux clusters, 8 nodes with 12 cores each, at the Regionales Rechenzentrum Köln (RRZK) using Cologne High Efficient Operating Platform for Science (CHEOPS); input was done in phylip format; and conversion of Fasta to phylip was done using Readseq [44] [XVII]. Nuclear coding genes were treated as one partition (PROTCAT model, substitution matrix LG + F, taken from ProtTest [45]). All other groups of orthologs were treated as separate partitions (32 partitions in total). (See Additional file 4 for the character partitions of subset 1 and 2.) We applied the rapid bootstrap algorithm [46] with a subsequent tree search. The numbers of bootstrap replicates were estimated on the fly by the "bootstopping" criteria implemented in RAxML 7.2.8 (default settings) [47]. The analyses yielded two trees. These trees are referred to as "tree 1" (corresponding to subset 1) and "tree 2" (corresponding to subset 2). Trees were edited in Dendroscope [48] [XVIII].

Hymenoptera systematics

We follow the terminology of [29] for supraspecific taxa of Hymenoptera.

Results

We downloaded 122,723 Hymenoptera sequences from GenBank 172.0 (as of 18 August 2009), including those of the nuclear genome of *N. vitripennis* (9,254 contigs). The annotation of the nuclear genome of *A. mellifera* was used as a reference when searching for orthologs (see Methods, step [a.III]), and corresponding sequences of this species were added during this step. After the first processing steps [a.I/b.I to II], including a search for orthologs, a sequence check with *checking_seq*, filtering for longest sequence per species and gene, and excluding groups of orthologs with fewer than four species, the data set included a total of 13,573 sequences from 4,536 species and 375 genes. Step [III], the exclusion of species with only one sequence in the data set, led to the exclusion of 1,074 species and subsequently of 68 groups of orthologs. Accordingly, sequences of 3,462 species in 307 groups of orthologs were aligned in step [IV]. The selection of the first and second maximum cliques of species with an overlap of at least 100 nucleotides or amino acids [steps IX and X] and the subsequent tests for compositional heterogeneity [steps XI and XII] led to the exclusion of 669 species and reduced the data set to 2,793 species. The pruning of species-rich genera to 15 species led to the exclusion of another 549 species [step XIV]. Pruned genera were *Camponotus, Cardiocondyla, Dorylus, Lasius, Myrmecocystus, Pheidole, Pogonomyrmex, Polyrhachis, Pseudomyrmex* (Formicidae), *Bombus, Diadasia, Euglossa, Xylocopa* (Apidae), *Colletes, Hylaeus* (Colletidae), *Aleiodes, Cotesia* (Braconidae), *Ceratosolen* (Agaonidae), *Andricus* (Cynipidae), *Neodiprion* (Diprionidae), *Pontania* (Tenthredinidae), *Megastigmus* (Torymidae) and *Polistes* (Vespidae).

After selecting the largest group of species that overlap in at least one group of orthologs [step XV], the final concatenated data set (subset 1) included 1,146 species (46 families), 222 groups of orthologs, 3,951 sequences and 88,626 aligned sites. Data coverage in subset 1 (number of sequences ÷ number of groups of orthologs × number of species) was 1.55%. Tree reconstruction and 560 rapid bootstrap replicates took 8.3 days. Tree 1 obtained from subset 1 is shown in Figures 2 and 3 and Additional file 5.

Subset 2 included an additional 115 sequences of 51 species from 31 families. Overall, the concatenated subset 2 consisted of 1,207 species (77 families), 222 groups of orthologs, 4,005 sequences and 88,807 aligned sites. The number of species is > 1,146 plus 51 due to repeated tests for compositional heterogeneity with slightly different results. (Both subsets are available at http://www.zfmk.de/web/Forschung/Molekularlabor/Datenstze/index.en.html). Data coverage (number of sequences ÷ number of groups of orthologs × number of species) in subset 2 was 1.49%. Tree reconstruction and 512 rapid bootstrap replicates took 8.9 days. Tree 2 obtained from subset 2 is shown in Figure 4 and Additional file 6. All species and all groups of orthologs included in subsets 1 and 2 are listed in Additional files 7, 8, 9 and 10.

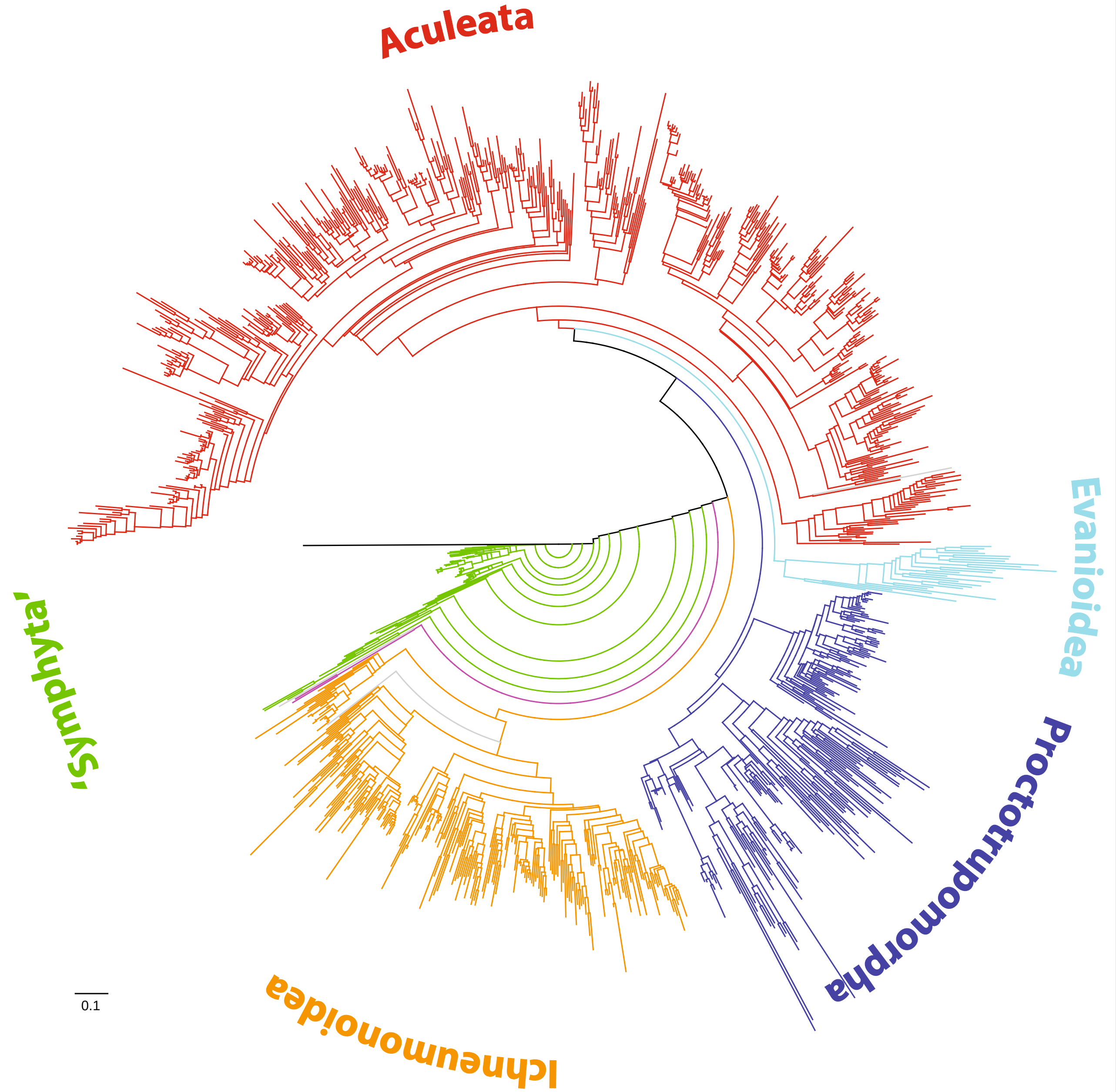


Figure 2 Simplified phylogenetic tree of Hymenoptera inferred from GenBank sequences (tree 1 obtained from subset 1). The tree includes 1,142 species. The applied color code indicates major lineages.

Discussion

The aim of the present investigation was to develop a bioinformatics pipeline for retrieving, processing, filtering, editing and analyzing large amounts of sequence data from GenBank in a phylogenetic context. Instead of using supertree approaches to explore existing data (see, for example, [19,49]), we relied on a direct reanalysis of the sequence data. Smith *et al.* [50] presented an alternative approach that they called a "mega-phylogeny approach", which also directly uses sequence data. It includes an *a priori* selection of gene regions of interest and an *a priori* separation of sequences into alleged monophyla with the aims of reducing the size of the supermatrix and improving alignment quality. A number of taxon-specific studies have also made use of GenBank sequence data, but those studies focused on specific genes (see, for example, [51,52]). We intended to avoid *a priori* decisions. In our pipeline, we suggest solutions for almost any obstacle that may appear along the way from sequence retrieval to tree reconstruction under the ML

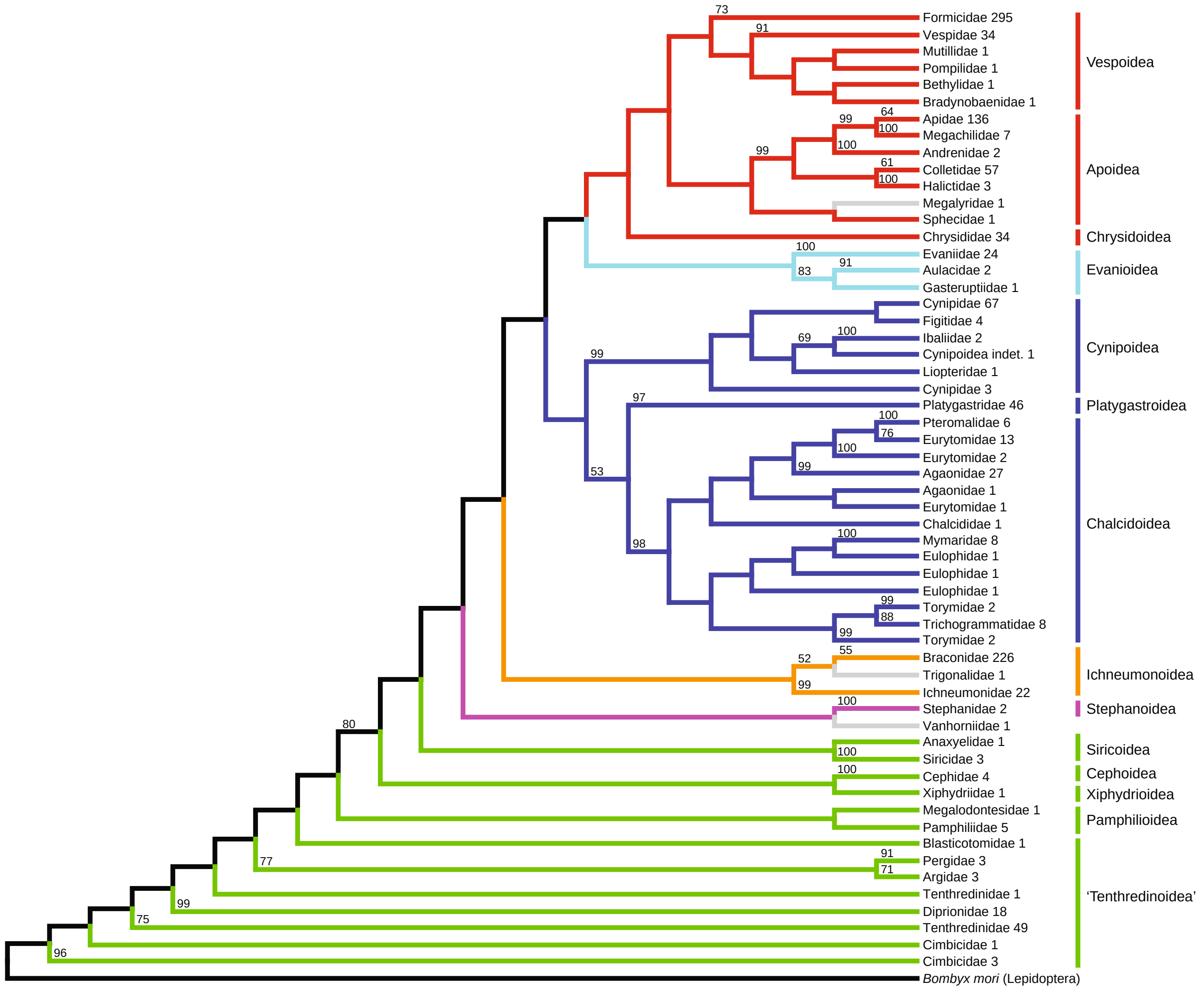


Figure 3 Phylogenetic tree of Hymenoptera inferred from GenBank sequences (tree 1), reduced to family level. Numbers that follow the family names indicate the number of analyzed species. Numbers above branches indicate bootstrap support values (%). Values < 50% are omitted. The applied color code corresponds to that of Figure 2. Single species whose position in the inferred phylogenetic tree we consider erroneous are shown in gray.

optimality criterion. In various regards, our approach is an extension and improvement of earlier efforts [2,4]. It offers an extended degree of automation in steps such as downloading from GenBank, sorting of sequences and translating and backtranslating sequences [steps I, b.II, b. V and VI] (Figure 1). Also, our approach includes improved quality management, such as by automatically checking the GenBank sequences for strand polarity and annotation, by masking problematic alignment regions and by handling compositional heterogeneity [steps b.III, VII and XI] (Figure 1). Our data selection steps [for example, steps III, IX and XV] (Figure 1) guarantee standardized levels of the density of the data set and of sequence overlap between included species. By choosing a minimum sequence overlap of 100 positions, we attempted to find a reasonable compromise between sequence overlap and number of species in the analysis. A larger overlap would have led to a significant decrease of the number of species in our phylogenetic tree. Furthermore, the present study is an update in terms of tree reconstruction facilities. We have, for the first time, applied a ML algorithm to such a large amount of GenBank data [step XVII] (Figure 1). Our approach is more general and independent of the taxonomic group. Finally, our bioinformatics solution is transparent and user-friendly. We provide all new scripts with respective comments and detailed manuals as part of this publication so that the pipeline is ready for use by anybody interested. In the following paragraphs, we discuss the results of our exemplary pipeline run with Hymenoptera data.

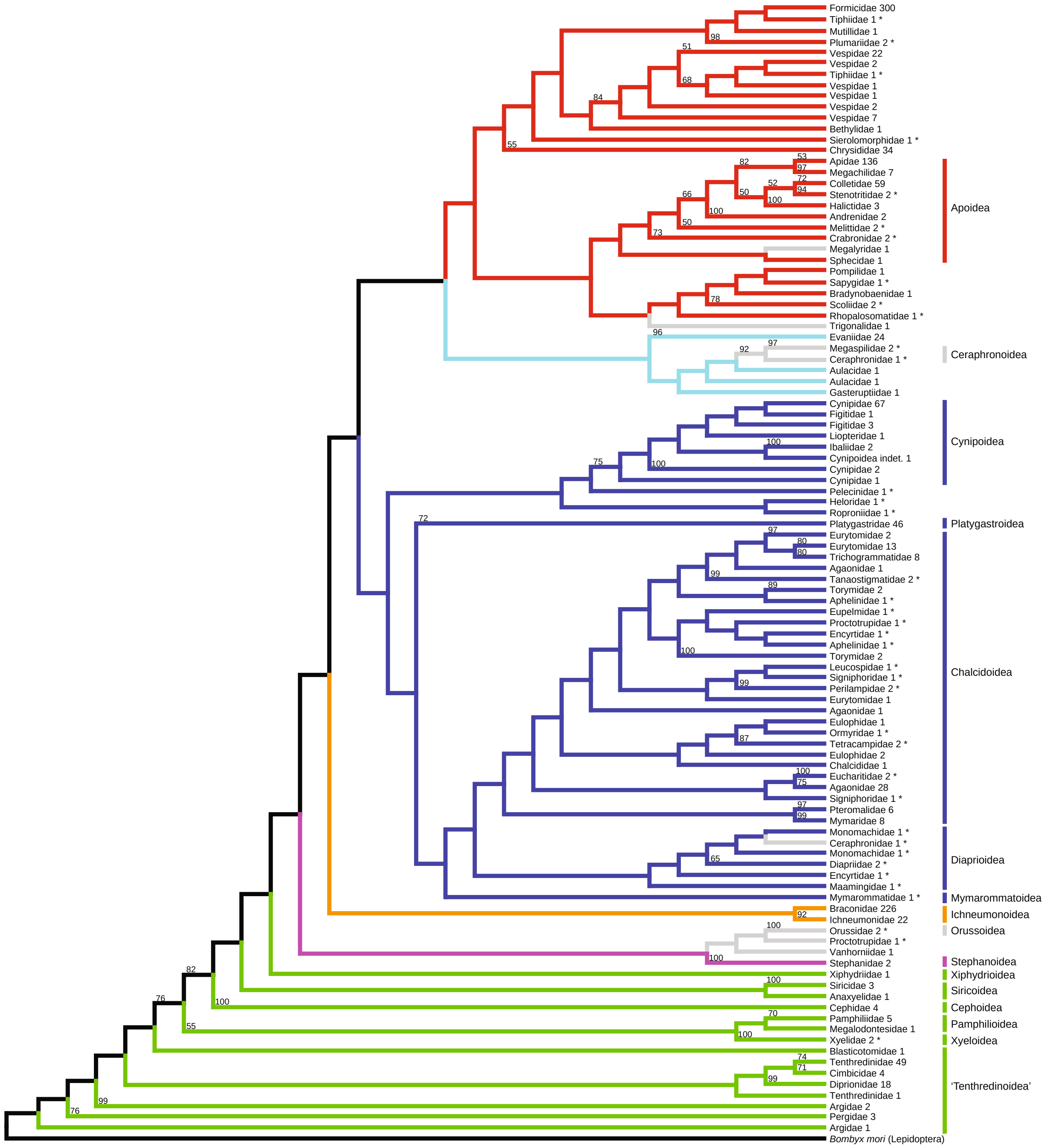


Figure 4 Phylogenetic tree of Hymenoptera inferred from GenBank sequences (tree 2 obtained from subset 2), reduced to family level. In this tree, species that were excluded by our pipeline in the course of generating subset 1 are reincluded. These taxa are marked with asterisks. The meaning of numbers and the applied color code correspond to those in Figure 3.

Data set and analysis

One of the main characteristics of data sets when combining sequence data from independently conducted investigations is data scarcity; that is, the lack of data overlap. Data distribution in supermatrices is unbalanced, and, as a consequence, there is a huge amount of missing data. However, data sets do not necessarily have to be complete to provide phylogenetic information. In fact, there is evidence that even with very low coverage, reliable phylogenetic estimates can be obtained (see, for example, [53]). The sheer proportion of missing data is not decisive as long as the number of characters scored is sufficient to correctly place the taxa in the tree [54]. Accordingly, we tried to cope with the problem of data scarcity by ensuring a minimum sequence overlap between taxa and a standardized data set density [steps III, IX, XIII, XIV and XV] (Figure 1). Still, our Hymenoptera data matrix is very large and exhibits very low coverage (1.5%). This is a direct consequence of the characteristics of the original sequence information present in GenBank. A large number of species for which only few sequences are available contrasts with a small number of species for which the transcriptome, the mitochondrial genome or even the entire nuclear genome have been sequenced. By combining all of these data in a single analysis, this data set will inevitably become large and unbalanced and will suffer from low overlap between taxa. Irrespective of the fact that sequencing is getting cheaper and faster and that phylogenomic data will rapidly increase the size of data sets, the data characteristics described herein are still expected to prevail in the near future. The challenge is to find optimal subsets for phylogenetic analysis in order to explore available information and to subsequently identify and fill the most severe gaps via target-specific sequencing. Accordingly, one of the goals of our approach has been to identify unstable nodes and to suggest future foci of molecular phylogenetic studies, in Hymenoptera, for an effective, economical and time-saving process.

For tree reconstruction, we performed supermatrix ML analyses. To the best of our knowledge, this is the largest set of eukaryotic real data studied using ML analysis. Past studies that utilized very large data sets applied supertrees or parsimony analyses. For example, McMahon and Sanderson [2] and Thomson and Shaffer [4] applied maximum parsimony analyses with supermatrices in their pipelines, but stated that they based this decision mainly on speed and computational capacity. However, with the latest program version of RAxML implementing partitioned analysis, rapid bootstrap functions, and the ability of parallel analyses, even very large data sets, can be analyzed in a reasonable amount of time. In the next few years, systematic biologists' access to multicore computers will get easier and broader, and high-performance computing (HPC) will become routine. At the moment, subsets should be constrained in size to allow ML analysis. During our work, we set an approximate maximum of 1,500 taxa and 100,000 sites. Phylogenetic analyses of subsets of this size take a maximum of two weeks on a fully parallelized HPC unit such as the one that we used. Unless one wants to analyze data sets that are significantly larger than ours, there is no computational or speed argument left to perform supertree or parsimony methods in favor of ML analyses. Accordingly, our approach was designed to prepare data for ML analysis. However, if a user wants to apply other algorithms for tree reconstruction (for example, maximum parsimony) or to adjust parameters (for example, to seek an extension of exploration of tree space or a comparison between inferred trees), the supermatrix produced by our pipeline can be used just as well (after step XVI) (Figure 1).

The phylogeny of Hymenoptera

We have restricted our results and discussion to (1) new contributions to the phylogeny of major lineages within Apocrita and to the monophyly and phylogeny of Proctotrupomorpha, (2) the recovery of some noncontroversial relationships and (3) the diagnosis of persistent problems and possible solutions. Phylogenetic relations within Hymenoptera are far too numerous and complex to be exhaustively discussed. The complete trees in Additional files 5 and 6 can be consulted for lower systematic level relationships.

In the following subsections, we repeatedly refer to single species as "misplaced". This means that their position as inferred in our trees clearly contradicts previous results from taxonomic as well as morphological and molecular phylogenetic studies. Accordingly, the phylogenetic positions of these taxa were considered artefacts and were excluded from discussion of topologies.

Major lineages within Apocrita

Within Apocrita, our analysis suggests a topology of Stephanoidea + (Ichneumonoidea + (Proctotrupomorpha + (Evanioidea + Aculeata))) (with misplacement of a single Vanhorniidae as sister to Stephanoidea being ignored) (Figure 3). Stephanoidea was inferred to be sister group to all other Apocrita in the morphological analyses of Vilhelmsen *et al.* [17]. Our analysis gives additional support for this relationship. The Ichneumonoidea are monophyletic in our trees. (Misplacement of a single Trigonalidae as sister to Braconidae is ignored.) Ichneumonoidea has been suggested as sister group to Aculeata by Rasnitsyn [13], a relationship that found only moderate support from Vilhelmsen *et al.* [17] and was not retrieved by most recent analyses (see, for example, [16,21,24,55,56]). Our trees corroborate the results of most analyses cited above and suggest a rejection of the clade Aculeata + Ichneumonoidea. Instead, we found Evanioidea to be sister group to Aculeata in our trees. A sister group relationship of Evanioidea and Aculeata has been suggested only by the combined morphological and molecular analysis by Sharkey *et al.* [57], and there are currently no convincing morphological synapomorphies that would support this clade. However, despite low branch support, we consider it quite possible that the Evanioidea are the long-sought sister group to the Aculeata and suggest further investigation of this particular clade. Rasnitsyn [13] introduced the supertaxon

Evaniomorpha, which includes Evanioidea, Ceraphronoidea, Megalyroidea, Trigonaloidea and Stephanoidea. We argue against the monophyly of Evaniomorpha, as our data support Stephanoidea as sister taxon of the remaining Apocrita (corroborating Vilhelmsen *et al.* [17]). We cannot provide substantial information on the position of the superfamilies Ceraphronoidea, Megalyroidea and Trigonaloidea, because their representatives are either included solely in the extended, possibly less reliable tree 2 (Ceraphronoidea) or obviously misplaced (Megalyroidea and Trigonaloidea).

Proctotrupomorpha

In our analyses, Proctotrupomorpha *s.l.* (that is, sensu Rasnitsyn 1988 [13]) was retrieved when again ignoring a few misplaced taxa. In tree 1, Proctotrupomorpha comprises Chalcidoidea, Platygastroidea and Cynipoidea (all of which are monophyletic, forming Cynipoidea + (Platygastroidea + Chalcidoidea)). In tree 2, more representatives of Proctotrupomorpha *s.l.* are present, and the inferred topology suggests the following relationships: Cynipoidea + (Platygastroidea + (Mymarommatoidea + (Diaprioidea + Chalcidoidea))). This contradicts the often proposed sister group relationship between Mymarommatoidea and Chalcidoidea (see, for example, [24,57,58]; but see the ambiguity in [17]). A sister group relationship between Diaprioidea and Chalcidoidea was retrieved in the molecular analyses of Castro and Dowton [56], but their taxon sampling lacked Mymarommatoidea, and was retrieved by Heraty *et al.* [21]. Our study is one of the first to include Mymarommatoidea in a molecular phylogenetic analysis, but the position of Mymarommatoidea in our analysis is not well supported and the group is represented only in the less reliable tree 2. A position of Chalcidoidea outside Proctotrupomorpha was recently proposed by Sharanowski *et al.* [20] based on the analysis of 24 putative orthologous genes (derived from ESTs) from a small number of taxa. We regard this position as unlikely based on our own results and those of previous molecular studies that provided respective parts of our data set [16,21,56]. The most recent morphological or combined morphological and molecular analyses also contradict an origin of Chalcidoidea outside Proctotrupomorpha [17,57].

Recovery of noncontroversial relationships

We evaluated the reliability of the inferred phylogenetic trees by the recovery of phylogenetic relationships that are largely considered noncontroversial. We found positive indications in tree 1. Specifically, our results are consistent with the generally accepted paraphyly of "Symphyta" (see, for example, [24]) and with the generally accepted monophyly of Apocrita and Aculeata (see, for example, [24,28]) (with misplacement of one Megalyridae within Aculeata being ignored). Also, we retrieved the noncontroversially monophyletic superfamilies Apoidea, Chalcidoidea, Cynipoidea, Evanioidea, Ichneumonoidea and Siricoidea. However, some crucial taxa were not represented in tree 1: Xyelidae and Orussidae. If we add them to the data set to infer tree 2, they are misplaced. The Xyelidae are found as a sister group to Pamphilioidea (Figure 4). This position is not very likely, as the sister group relationship of Xyelidae and the remaining Hymenoptera is well supported [25-27]. The Orussidae, which have a key position within Hymenoptera evolution as sister group of Apocrita, are placed at the base of Apocrita along with some Proctotrupoidea taxa (Figure 4). However, the clade Orussidae + Apocrita is well established and supported by morphological and molecular data (see, for example, [13,17,18,57]). This demonstrates the necessity of sequence overlap definitions and shows that the positions of reincluded taxa (indicated by asterisks in Figure 4 and Additional file 6) have to be discussed with caution. The backbone of the tree, with its major splits, however, remains largely unaffected by adding taxa that do not fulfill our overlap criteria.

Diagnosis of persistent problems and possible solutions

With the aid of our trees, we identified several persistent problems in the Hymenoptera tree. While the available sequence data already cover all major lineages of Hymenoptera, they are unequally distributed and there is poor overlap among taxa. This contradiction between taxonomic breadth and genomic depth in the data of Hymenoptera is in accordance with the conclusions of Sanderson [59] in his evaluation of the phylogenetic signal in Eukaryota. The large amount of missing data and the low taxonomic overlap between mitochondrial and nuclear data in our sets call for a solution. To get more independent markers and to close the taxonomic gap between mitochondrial and nuclear data, we suggest EST studies (nuclear genes) for taxa with completely sequenced mitochondrial genomes and sequencing of mitochondrial genomes of those taxa for which we already have a large number of nuclear sequence data available.

An obvious problem for solving higher-level relationships within Hymenoptera is the underrepresentation of the small superfamilies Megalyroidea, Trigonaloidea, Ceraphronoidea and Mymarommatoidea. Another highly problematic issue is those families of Proctotrupoidea that we currently cannot map on the phylogenetic tree. Any additional data regarding these taxa in terms of species and genes will be of great value.

As extensive EST studies are still expensive, we also recommend target-specific amplification of nuclear coding genes. With the prospect of new primer design tools (J. Borner, C. Pick, T. Burmester, unpublished data), amplification and sequencing of a data set of, for example, 22 taxa (all superfamilies) and 50 nuclear coding genes can be accomplished in a reasonable amount of time and at reasonable cost. Taxon sampling should again be based on taxa with completely sequenced mitochondrial genomes.

Conclusions

Exemplarily for Hymenoptera, we have demonstrated that the tree reconstructed from our pipeline output can make a substantial contribution to the phylogeny of the taxon and that comprehensive results can complement the discrete inferences from the single studies that have produced the data that were reanalyzed. Inspired by McMahon and Sanderson [2] and Sanderson *et al.* [3], we found an adequate approach to analyze all currently available molecular data in a single phylogenetic study in a standardized and efficient way. The impossible child of the scientific community, a sequence data monster, can be tamed. Every systematic biologist, even without advanced programming and bioinformatics skills, is given the capability to produce a tree of his taxon of interest. Our approach offers the possibility of relatively simple and reliable monitoring of new data and tree robustness, that is, the possibility to keep track of the phylogenetic signal in a taxonomic group. This also enables researchers to monitor how phylogenetic trees change over time with an increase of data size and density. This might promote a better understanding of more theoretical issues related to the analyses of molecular data, such as the information content of genes or the suitability and selection of genes to answer phylogenetic questions. Our approach therefore has great potential to meet the challenges of the phylogenomic era, to improve our ideas on phylogenetic affinities and to contribute to a better understanding of the evolution of organisms.

Additional material

Additional file 1: Software tools and manuals. All newly developed software tools and corresponding manuals.

Additional file 2: gi numbers of sequences from GenBank used in our pipeline. List of the GenBank gi numbers of all Hymenoptera sequences that were initially inputted in our pipeline run.

Additional file 3: On maximum cliques. A short introduction to maximum cliques and how we used them in our analysis.

Additional file 4: Character partitions of subset 1 and 2. The character partitions of the two subsets that were used in the phylogenetic analyses (subset 1 and subset 2).

Additional file 5: Tree 1, complete. Phylogenetic tree of Hymenoptera inferred from GenBank sequences (tree 1). Numbers on branches indicate bootstrap support values (%). The applied color code corresponds to that of Figures 2 and 3. Single species whose position in the inferred phylogenetic tree we consider erroneous are shown in gray.

Additional file 6: Tree 2, complete. Phylogenetic tree of Hymenoptera inferred from GenBank sequences (tree 2). In this tree, species that were excluded by our pipeline in the course of generating subset 1 are reincluded. These taxa are marked with asterisks. The meaning of numbers and the applied color code correspond to those in Additional file 5.

Additional file 7: Species included in subset 1. All species included in subset 1, sorted by family.

Additional file 8: Groups of orthologs included in subset 1. All groups of orthologs included in subset 1, plus coverage of each group.

Additional file 9: Species included in subset 2. All species included in subset 2, sorted by family.

Additional file 10: Groups of orthologs included in subset 2. All groups of orthologs included in subset 2, plus coverage of each group.

Acknowledgements

We thank Alexandros Stamatakis (Heidelberg, Germany) for help with the phylogenetic analyses, Ingo Ebersberger (Vienna, Austria) for help with HaMStR, Nikola Szucsich (Hamburg, Germany) for valuable discussions and three anonymous referees for their helpful reviews. We acknowledge Viktor Achter, Volker Winkelmann and Sebastian Breuers for help with likelihood analyses on the Cologne High Efficiency Operating Platform for Sciences (CHEOPS, a DFG-granted HPC cluster at the RRZK, University of Cologne, Cologne, Germany).

Author details

[1]Zoologisches Forschungsmuseum Alexander Koenig, Adenauerallee 160, D-53113 Bonn, Germany. [2]Institut für Systemische Neurowissenschaften, Universitätsklinikum Hamburg-Eppendorf, Martinistrasse 52, D-20246 Hamburg, Germany. [3]Staatliches Museum für Naturkunde Stuttgart, Rosenstein 1, D-70191 Stuttgart, Germany. [4]Zoologisches Institut der Universität Hamburg, Martin-Luther-King-Platz 3, D-20146 Hamburg, Germany. [5]Zoologisches Museum Hamburg, Martin-Luther-King-Platz 3, D-20146 Hamburg, Germany.

Authors' contributions

BMi and RSP conceived of the study. BMe, BMi and RSP designed the study. RSP coordinated the study. BMe, JB, KM and RSP carried out the analyses. BMe wrote the major part of the bioinformatics tools, and JB and BMi wrote minor parts of the bioinformatics tools. BMe, JB and RSP wrote the manuals with comments and revisions from KM. BMe, BMi, ON and RSP wrote the manuscript. JB, KM, KS and LK provided comments on and made revisions to the manuscript. All authors read and approved the final manuscript.

Received: 27 May 2011 Accepted: 18 August 2011
Published: 18 August 2011

References

1. Benson D, Karsch-Mizrachi I, Lipman D, Ostell J, Sayers E: **GenBank.** *Nucleic Acids Res* 2009, **37**:D26-D31.
2. McMahon MM, Sanderson MJ: **Phylogenetic supermatrix analysis of GenBank sequences from 2228 papilionoid legumes.** *Syst Biol* 2006, **55**:818-836.
3. Sanderson MJ, Boss D, Chen D, Cranston KA, Wehe A: **The PhyLoTA Browser: processing GenBank for molecular phylogenetics research.** *Syst Biol* 2008, **57**:335-346.
4. Thomson RC, Shaffer HB: **Sparse supermatrices for phylogenetic inference: taxonomy, alignment, rogue taxa and the phylogeny of living turtles.** *Syst Biol* 2010, **59**:42-58.
5. LaSalle J, Gould ID: **Hymenoptera: their diversity and their impact on diversity of other organisms.** In *Hymenoptera and Biodiversity.* Edited by: LaSalle J, Gauld ID. Washington DC: CAB International; 1993:1-26.
6. Quicke DLJ: **Parasitic Wasps.** New York: Kluwer Academic Publishers; 1997.
7. Whitfield JB, Lockhart PJ: **Deciphering ancient rapid radiations.** *Trends Ecol Evol* 2007, **22**:258-265.
8. Murphy NP, Carey D, Castro LR, Dowton M, Austin AD: **Phylogeny of the platygastroid wasps (Hymenoptera) based on sequences from the 18S rRNA, 28S rRNA and cytochrome oxidase I genes: implications for the evolution of the ovipositor system and host relationships.** *Biol J Linnean Soc* 2007, **91**:653-669.

9. Königsmann E: **Das phylogenetische System der Hymenoptera. Teil 1: Einführung, Grundplanmerkmale, Schwestergruppe und Fossilfunde.** *D Entomol Z (NF)* 1976, **23**:253-279.

10. Königsmann E: **Das phylogenetische System der Hymenoptera. Teil 2: Symphyta.** *D Entomol Z (NF)* 1977, **24**:1-40.

11. Königsmann E: **Das phylogenetische System der Hymenoptera. Teil 3: Terebrantes (Unterordnung Apocrita).** *D Entomol Z (NF)* 1978, **25**:1-55.

12. Königsmann E: **Das phylogenetische System der Hymenoptera. Teil 4: Aculeata (Unterordnung Apocrita).** *D Entomol Z (NF)* 1978, **25**:365-435.

13. Rasnitsyn AP: **An outline of the evolution of the hymenopterous insects (order Vespida).** *Orient Insects* 1988, **22**:115-145.

14. Dowton M, Austin AD: **Molecular phylogeny of the insect order Hymenoptera: apocritan relationships.** *Proc Natl Acad Sci USA* 1994, **91**:9911-9915.

15. Carpenter JM, Wheeler WC: **Towards simultaneous analysis of morphological and molecular data in Hymenoptera.** *Zool Scripta* 1999, **28**:251-260.

16. Dowton M, Austin AD: **Simultaneous analysis of 16S, 28S, COI and morphology in the Hymenoptera: Apocrita evolutionary transitions among parasitic wasps.** *Biol J Linnean Soc* 2001, **74**:87-111.

17. Vilhelmsen L, Mikó I, Krogmann L: **Beyond the wasp-waist: structural diversity and phylogenetic significance of the mesosoma in apocritan wasps (Insecta: Hymenoptera).** *Zool J Linnean Soc* 2010, **159**:22-194.

18. Dowton M, Cameron SL, Austin AD, Whiting MF: **Phylogenetic approaches for the analysis of mitochondrial genome sequence data in the Hymenoptera: a lineage with both rapidly and slowly evolving mitochondrial sequences.** *Mol Phylogenet Evol* 2009, **52**:512-519.

19. Davis RB, Baldauf SL, Mayhew PJ: **The origins of species richness in the Hymenoptera: insights from a family-level supertree.** *BMC Evol Biol* 2010, **10**:109.

20. Sharanowski BJ, Robbertse B, Walker J, Voss SR, Yoder R, Spatafora J, Sharkey MJ: **Expressed sequence tags reveal Proctotrupomorpha (minus Chalcidoidea) as sister to Aculeata (Hymenoptera: Insecta).** *Mol Phylogenet Evol* 2010, **57**:101-112.

21. Heraty J, Ronquist F, Carpenter JM, Hawks D, Schulmeister S, Dowling AP, Murray D, Munro J, Wheeler WC, Schiff N, Sharkey M: **Evolution of the hymenopteran megaradiation.** *Mol Phylogenet Evol* 2011, **60**:73-88.

22. Weinstock GM, Robinson GE, Gibbs RA, Worley KC, Evans JD, Maleszka R, Robertson HM, Weaver DB, Beye M, Bork P, Elsik CG, Hartfelder K, Hunt GJ, Zdobnov EM, Amdam GV, Bitondi MM, Collins AM, Cristino AS, Lattorff MG, Lobo CH, Moritz RFA, Nunes FMF, Page RE, Simoes ZLP, Wheeler D, Carninci P, Fukuda S, Hayashizaki Y, Kai C, Kawai J, *et al*: **Insights into social insects from the genome of the honeybee *Apis mellifera*.** *Nature* 2006, **443**:931-949, A published erratum appears in Nature 2006, 444:512.

23. Werren JH, Richards S, Desjardins CA, Niehuis O, Gadau J, Colbourne JK, Beukeboom LW, Desplan C, Elsik CG, Grimmelikhuijzen CJ, Kitts P, Lynch JA, Murphy T, Oliveira DC, Smith CD, van de Zande L, Worley KC, Zdobnov EM, Aerts M, Albert S, Anaya VH, Anzola JM, Barchuk AR, Behura SK, Bera AN, Berenbaum MR, Bertossa RC, Bitondi MMG, Bordenstein SR, Bork P, *et al*: **Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species.** *Science* 2010, **327**:343-348, A published erratum appears in Science 2010, 327:1577.

24. Ronquist F, Rasnitsyn AP, Roy A, Eriksson K, Lindgren M: **Phylogeny of the Hymenoptera: a cladistic reanalysis of Rasnitsyn's (1988) data.** *Zool Scripta* 1999, **28**:13-50.

25. Vilhelmsen L: **Phylogeny and classification of the extant basal lineages of the Hymenoptera (Insecta).** *Zool J Linnean Soc* 2001, **131**:393-442.

26. Rasnitsyn AP: **Superorder Vespidea Laicharting, 1781. Order Hymenoptera Linn, 1758.** In *History of Insects.* Edited by: Rasnitsyn AP, Quicke DLJ. Dordrecht: Kluwer Academic Publishers; 2002:242-254.

27. Schulmeister S: **Simultaneous analysis of basal Hymenoptera (Insecta), introducing robust-choice sensitivity analysis.** *Biol J Linnean Soc* 2003, **79**:245-275.

28. Brothers DJ: **Phylogeny and classification of the aculeate Hymenoptera, with special reference to Mutillidae.** *Univ Kansas Sci Bull* 1975, **50**:483-648.

29. Sharkey MJ: **Phylogeny and classification of Hymenoptera.** *Zootaxa* 2007, **1668**:521-548.

30. Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome Res* 1999, **9**:868-877.

31. Ebersberger I, Strauss S, von Haeseler A: **HaMStR: Profile hidden Markov model based search for orthologs in ESTs.** *BMC Evol Biol* 2009, **9**:157.

32. Tatusova TA, Madden TL: **BLAST 2 sequences, a new tool for comparing protein and nucleotide sequences.** *FEMS Microbiol Lett* 1999, **174**:247-250.

33. Katoh K, Misawa K, Kuma K, Miyata T: **MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.** *Nucleic Acids Res* 2002, **30**:3059-3066.

34. Katoh K, Toh H: **Recent developments in the MAFFT multiple sequence alignment program.** *Brief Bioinform* 2008, **9**:286-298.

35. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32**:1792-1797.

36. Misof B, Misof K: **A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion.** *Syst Biol* 2009, **58**:21-34.

37. Kück P, Meusemann K, Dambach J, Thormann B, von Reumont BM, Waegele JW, Misof B: **Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees.** *Front Zool* 2010, **7**:10.

38. Kück P: *ALICUT: a PerlScript which cuts ALISCORE identified RSS* Department of Bioinformatics, Zoologisches Forschungsmuseum A. Koenig (ZFMK), Bonn, Germany, version 2.0 edition; 2009.

39. Castresana J: **Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.** *Mol Biol Evol* 2000, **17**:540-552.

40. Talavera G, Castresana J: **Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments.** *Syst Biol* 2007, **56**:564-577.

41. Schmidt HA, Strimmer K, Vingron M, von Haeseler A: **TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing.** *Bioinformatics* 2002, **18**:502-504.

42. Stamatakis A: **RAxML-VI-HPC: Maximum Likelihood-based Phylogenetic Analyses with Thousands of Taxa and Mixed Models.** *Bioinformatics* 2006, **22**:2688-2690.

43. Ott M, Zola J, Stamatakis A, Aluru S: **Large-scale maximum likelihood-based phylogenetic analysis on the IBM BlueGene/L.** *Proceedings of the 2007 ACM/IEEE Conference on Supercomputing: 2007 Reno, NV, USA* Berlin: VDE Verlag; 2007, 1-11.

44. Gilbert D: *Readseq* Indiana University, Bloomington, Indiana; 2001 [http://iubio.bio.indiana.edu/soft/molbio/readseq/java/].

45. Abascal F, Zardoya R, Posada D: **ProtTest: Selection of best-fit models of protein evolution.** *Bioinformatics* 2005, **21**:2104-2105.

46. Stamatakis A, Hoover P, Rougemont J: **A rapid bootstrap algorithm for the RAxML web servers.** *Syst Biol* 2008, **57**:758-771.

47. Pattengale ND, Alipour M, Bininda-Emonds ORP, Moret BME, Stamatakis A: **How many bootstrap replicates are necessary?** *J Comput Biol* 2010, **17**:337-354.

48. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R: **Dendroscope: an interactive viewer for large phylogenetic trees.** *BMC Bioinformatics* 2007, **8**:460.

49. Davis RB, Baldauf SL, Mayhew PJ: **Many hexapod groups originated earlier and withstood extinction events better than previously realized: inferences from supertrees.** *Proc Royal Soc London B* 2010, **277**:1597-1606.

50. Smith SA, Beaulieu J, Donoghue MJ: **Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches.** *BMC Evol Biol* 2009, **9**:37.

51. Hunt T, Vogler AP: **A protocol for large-scale rRNA sequence analysis: towards a detailed phylogeny of Coleoptera.** *Mol Phylogenet Evol* 2008, **47**:289-301.

52. Pyron RA, Wiens JJ: **A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians.** *Mol Phylogenet Evol* .

53. Driskell AC, Ané C, Burleigh JG, McMahon MM, O'Meara B, Sanderson MJ: **Prospects for building the tree of life from large sequence databases.** *Science* 2004, **306**:1172-1174.

54. Wiens JJ: **Missing data, incomplete taxa, and phylogenetic accuracy.** *Syst Biol* 2003, **52**:528-538.

55. Sharkey MJ, Roy A: **Phylogeny of the Hymenoptera: a reanalysis of the Ronquist et al. (1999) reanalysis, with an emphasis on wing venation and apocritan relationships.** *Zool Scripta* 2002, **31**:57-66.

56. Castro LR, Dowton M: **Molecular analyses of the Apocrita (Insecta: Hymenoptera) suggest that the Chalcidoidea are sister to the diaprioid complex.** *Invert Syst* 2006, **20**:603-614.

57. Sharkey MJ, Carpenter JM, Vilhelmsen L, Heraty J, Liljeblad J, Dowling APG, Schulmeister S, Murray D, Deans AR, Ronquist F, Krogmann L, Wheeler WC:

Phylogenetic relationships among superfamilies of Hymenoptera. *Cladistics* 2011, **27**:1-33.

58. Gibson GAP: Evidence for monophyly and relationships of Chalcidoidea, Mymaridae, and Mymarommatidae (Hymenoptera: Terebrantes). *Can Entomol* 1986, **118**:205-240.
59. Sanderson M: Phylogenetic signal in the eukaryotic tree of life. *Science* 2008, **321**:121-123.

doi:10.1186/1741-7007-9-55
Cite this article as: Peters *et al.*: The taming of an impossible child: a standardized all-in approach to the phylogeny of Hymenoptera using public database sequences. *BMC Biology* 2011 **9**:55.