Evaluating Topological Conflict in Centipede Phylogeny Using Transcriptomic Data Sets

Rosa Fernández,[1] Christopher E. Laumer,[1] Varpu Vahtera,[1,2] Silvia Libro,[3] Stefan Kaluziak,[3] Prashant P. Sharma,[4] Alicia R. Pérez-Porro,[1,5] Gregory D. Edgecombe,[6] and Gonzalo Giribet*[,1]

[1]Museum of Comparative Zoology & Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA

[2]Zoological Museum, Department of Biology, University of Turku, Turku, Finland

[3]Marine Science Center, Northeastern University, Nahant, MA

[4]Division of Invertebrate Zoology, American Museum of Natural History, New York, NY

[5]Centre d'Estudis Avançats de Blanes (CEAB-CSIC), Catalonia, Spain

[6]Department of Earth Sciences, The Natural History Museum, London, United Kingdom

***Corresponding author:** E-mail: ggiribet@g.harvard.edu.

Associate editor: Nicolas Vidal

Abstract

Relationships between the five extant orders of centipedes have been considered solved based on morphology. Phylogenies based on samples of up to a few dozen genes have largely been congruent with the morphological tree apart from an alternative placement of one order, the relictual Craterostigmomorpha, consisting of two species in Tasmania and New Zealand. To address this incongruence, novel transcriptomic data were generated to sample all five orders of centipedes and also used as a test case for studying gene-tree incongruence. Maximum likelihood and Bayesian mixture model analyses of a data set composed of 1,934 orthologs with 45% missing data, as well as the 389 orthologs in the least saturated, stationary quartile, retrieve strong support for a sister-group relationship between Craterostigmomorpha and all other pleurostigmophoran centipedes, of which the latter group is newly named Amalpighiata. The Amalpighiata hypothesis, which shows little gene-tree incongruence and is robust to the influence of among-taxon compositional heterogeneity, implies convergent evolution in several morphological and behavioral characters traditionally used in centipede phylogenetics, such as maternal brood care, but accords with patterns of first appearances in the fossil record.

Key words: **phylogenomics, incongruence, next-generation sequencing, Illumina, Myriapoda, Chilopoda, molecular dating.**

Article

Introduction

The myriapod Class Chilopoda—centipedes—consists of some 3,500 extant species classified in five extant orders and one extinct order (Edgecombe and Giribet 2007). They have also proven to be an interesting case study from a phylogenetic perspective because the interrelationships of the living orders have received a high level of consensus based on diverse kinds of morphological evidence. The morphological scheme has then been used as a benchmark for assessing the efficacy of different genes, kinds of molecular data, or methods of molecular data analysis in resolving the centipede tree (e.g., Regier et al. 2005; Giribet and Edgecombe 2006; Murienne et al. 2010). Over the past few years, molecular and morphological estimates of high-level centipede phylogeny have largely reached agreement apart from the placement of one lineage, the relictual order Craterostigmomorpha. This is composed of two species in the genus *Craterostigmus* Pocock, 1902 (Pocock 1902) with one species endemic to each of Tasmania and New Zealand (Edgecombe and Giribet 2008).

The morphological tree of centipedes reflects the groups established in classifications devised more than a century ago (Pocock 1902; Verhoeff 1902–1925) (fig. 1*A*). This hypothesis recognizes a fundamental division of Chilopoda into Notostigmophora (composed of the single order Scutigeromorpha, with ~100 species) and Pleurostigmophora, which groups the other four living orders. The names of the two groups reflect a difference in the position of the respiratory openings, either opening dorsally on the tergal plates (in Notostigmophora) or opening above the leg bases in the pleuron (in Pleurostigmophora). Pleurostigmophora in turn divides into Lithobiomorpha and a putative clade that groups the remaining three orders, named Phylactometria based on a shared behavior of maternal care of the eggs and hatchlings (Edgecombe and Giribet 2004). Phylactometria consists of *Craterostigmus*, Scolopendromorpha, and Geophilomorpha, with the latter two orders almost universally united in a group named Epimorpha based on their strictly epimorphic mode of development, that is, having a fixed number of segments throughout the course of postembryonic development.

These relationships were initially defended using nonquantitative methods applied to either diverse kinds of anatomical and behavioral data (Dohle 1985; Borucki 1996) or detailed studies of particular organ systems (Hilken 1997; Wirkner and

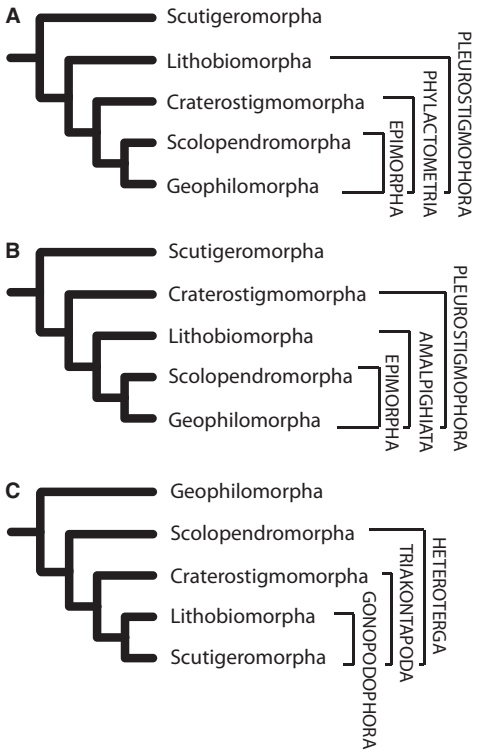

Fig. 1. Phylogenetic hypotheses of centipede relationships. (*A*) The morphologically supported tree (e.g., Pocock 1902). (*B*) Amalpighiata hypothesis as supported in early molecular analyses (Mallatt and Giribet 2006). (*C*) Ax's (2000) morphological hypothesis, with tree in (*A*) rerooted.

Pass 2002; Müller and Meyer-Rochow 2006). Numerical parsimony approaches supported this same cladogram using either a ground pattern coding for the five living orders (Shear and Bonamo 1988) or by coding a larger number of exemplar species (Edgecombe et al. 1999; Edgecombe and Giribet 2004; Giribet and Edgecombe 2006; Murienne et al. 2010). Other morphological hypotheses have also been introduced, including a rerooted version of the standard tree (fig. 1*C*) that infers a decrease rather than increase in segment numbers (Ax 2000).

The introduction of molecular data showed striking congruence with morphology with respect to most aspects of centipede phylogeny (Giribet et al. 1999), although in some cases the molecular data alone posed conflict with respect to the position of Craterostigmomorpha (e.g., Edgecombe et al. 1999, fig. 2). Early analyses of nuclear protein-encoding genes showed strong conflict with both ribosomal genes and morphology (Shultz and Regier 1997; Regier et al. 2005; Giribet and Edgecombe 2006). The most recent molecular phylogenies, using either a combination of nuclear ribosomal and mitochondrial loci (Murienne et al. 2010) or larger compilations of nuclear protein-coding genes (Regier et al. 2010), converge on trees that agree with morphology in most respects but present conflict on the *Craterostigmus* question, as was also found in an analysis based on nearly complete nuclear ribosomal genes across arthropods and other ecdysozoans (Mallatt and Giribet 2006) (fig. 1*B*). All these data sources agree on the basal division of Chilopoda into Notostigmophora and Pleurostigmophora, but conflict remains over whether Lithobiomorpha (supported by morphology) or Craterostigmomorpha (supported by molecules) is sister group to the other members of Pleurostigmophora. As well, the status of Epimorpha is unclear from existing molecular data: The group has been contradicted by a weakly supported alliance between Lithobiomorpha and Scolopendromorpha (Murienne et al. 2010) or has not been fully tested because Geophilomorpha was unsampled (Regier et al. 2010).

Here, we apply a phylogenomic approach to resolve these remaining controversies in centipede phylogeny. We draw upon a much larger number of genes (1,934 orthologs) than has been previously considered to re-evaluate the phylogenetic position of *Craterostigmus*, calibrate the centipede molecular clock using modern methods for dating phylogenies, and discuss the implications of the putative convergence of morphological and behavioral characters such as maternal care in centipedes. The availability of genome and transcriptome data brings novel challenges to tree reconstruction (e.g., compositional heterogeneity and gene-tree incongruence); we demonstrate that, though present in these data, such phenomena do not influence our major phylogenetic conclusions.

Results

Transcriptome Assembly, Isoform Filtering, and Orthology Assignment

cDNA from the nine species used in this study was sequenced on the Illumina HiSeq 2500 platform (150 bp paired-end reads). A summary of the assembly statistics is shown in table 1. Following redundancy reduction, open reading frame (ORF) prediction, and selection of the longest ORF per putative unigene, 9,934–19,621 peptide sequences per taxon were retained. Application of the Orthologous MAtrix (OMA) stand-alone algorithm (Altenhoff et al. 2011, 2013) grouped these peptides into a total of 30,586 orthologs. Three different supermatrices were constructed for the phylogenetic analyses. The first supermatrix was constructed by selecting only the orthologs present in six or more taxa. The number of orthologs per taxon in this supermatrix ranged from 131 to 1,651, with 1,934 orthologs in total. The combined length of the aligned ortholog matrices (before probabilistic alignment masking of each individual ortholog with ZORRO; Wu et al. 2012) was 770,128 amino acid positions. Concatenation of individually ZORRO-masked ortholog alignments yielded a supermatrix of 526,145 amino acids. The number of orthologs represented per taxon varied from 530 to 10,070 (fig. 2). As expected, in all cases, the lowest values corresponded to *Archispirostreptus gigas*, as its transcriptome was pyrosequenced to relatively low throughput on the 454 Life Sciences platform (Meusemann et al. 2010). However, gene representation in the ingroup species of this study was relatively high, varying from 1,139 to 1,543, yielding less than 45% missing data. To discern if substitutional saturation and compositional heterogeneity

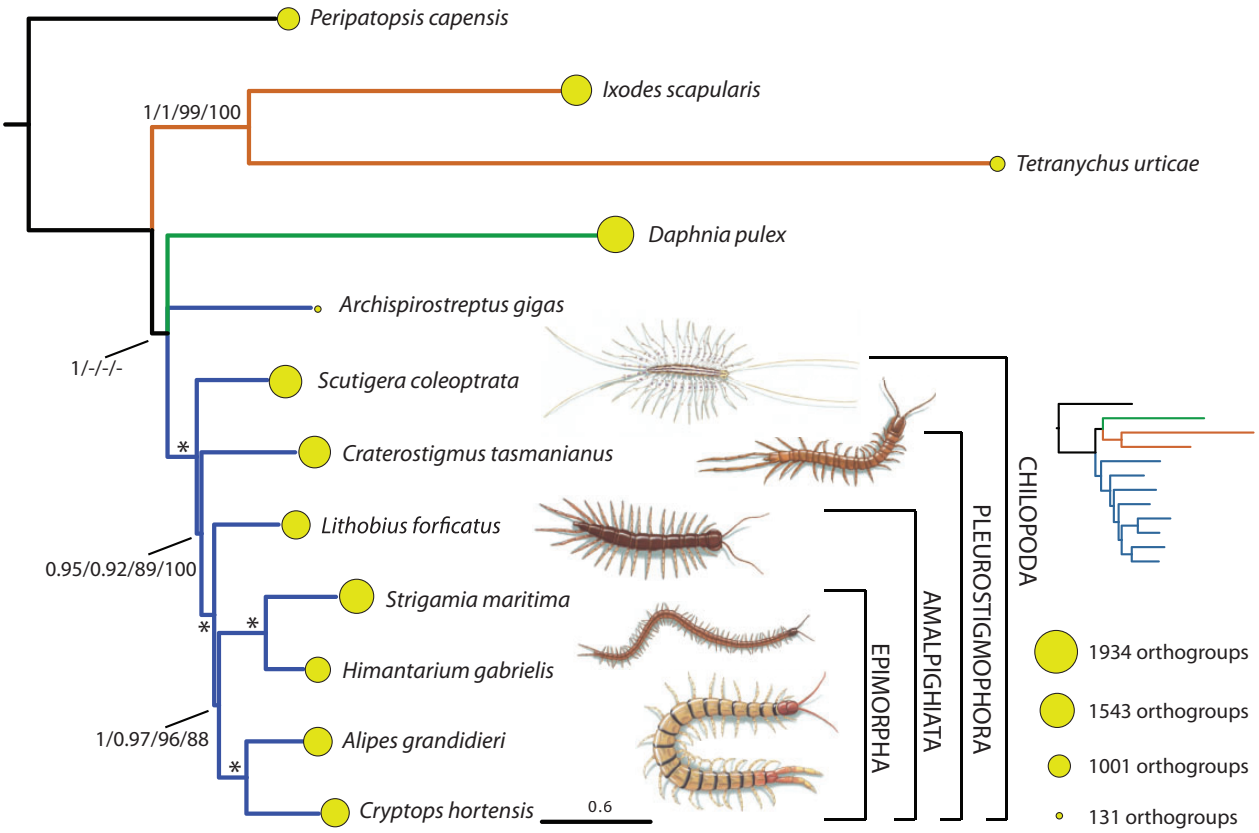

FIG. 2. Bayesian phylogenomic hypothesis of centipede interrelationships. Nodal support values (posterior probability/Shimodaira-Hasegawa-like support/bootstrap/bootstrap) of the four different analyses (Bayesian inference/ML with PhyML-PCMA/ML with RAxML/ML with RAxML for the reduced data set) are indicated in each case. Ln *L* PhyML-PCMA: −4,768,683.119. Ln *L* RAxML: −4,762,759.821600. Ln *L* RAxML: −799,809.006206. Asterisks indicate support values of 1/1/100/100. The size of the circles is proportional to the number of orthologs in each species for the large data set. Inset indicates the 389-gene-tree topology (reduced data set) with myriapod monophyly.

Table 1. Assembly Parameters.

	N Raw Reads	*N* Reads AT	% Discarded Reads	NRMC	*N* Contigs	*N* Bases (Mb)	Average Length of Contigs	Maximum Contig Length (bp)	N50	Total *N* Orthologs	*N* Selected Orthologs	Missing Data (%)
Scutigera coleoptrata	26,072,840	25,870,392	0.7	23,546,281	178,854	95.2	532.2	16,337	684	9,932	1,472	24.5
Craterostigmus tasmanianus	76,698,082	31,313,104	59.1	30,257,748	99,546	64.6	649.4	18,029	964	7,686	1,430	26
Lithobius forficatus	77,628,252	75,849,229	2.2	72,202,535	74,544	34.4	461.9	11,715	530	5,495	12,67	41.4
Strigamia maritima	—	—	—	—	24,079	176.2	11,955.5	1,344,237	24,745	7,583	15,43	17.4
Himantarium gabrielis	54,681,444	53,535,366	2.0	51,089,766	37,200	21.8	586.9	17,609	782	5,605	11,39	45.4
Alipes grandidieri	32,332,034	30,561,861	5.4	28,272,411	138,229	70.3	508.9	9,801	631	9,441	1,299	35.9
Cryptops hortensis	34,611,798	21,731,914	37.2	20,594,894	75,753	45.5	601.2	10,232	803	8,821	1,274	39.1
Peripatopsis capensis	40,774,042	40,774,042	0	—	92,516	35.6	384.8	10,252	386	5,148	1,001	53.9
Tetranychus urticae	—	—	—	—	18,313*	—	—	—	—	7,775	683	77.3
Ixodes scapularis	—	—	—	—	20,473*	—	—	—	—	6,891	1,368	27.4
Daphnia pulex	—	—	—	—	18,989	197.2	38,000.7	4,193,030	49,250	10,070	1,651	12.6
Archispirostreptus gigas	—	—	—	—	4,008	1.99	495.8	768	560	530	131	96.3

NOTE.—*N*, number; AT, after thinning and trimming; NRMC, number of reads matched to contigs. Percentage of missing data was calculated in relation to the final matrix (526,145 amino acids). Total number of orthologs corresponds to all the orthologs recovered per taxon. Number of selected orthologs is based on a minimum taxon occupancy of six (i.e., only the orthologs shared by at least six taxa were selected for the phylogenomic analyses). Asterisks for chelicerates indicate number of peptide sequences downloaded from annotated genome projects for analysis with OMA.

were affecting our phylogenetic results, we considered a second supermatrix that included the 389 genes that passed individual tests of compositional heterogeneity and which furthermore fell in the least saturated quartile of this stationary set (see Materials and Methods). Finally, to reduce the percentage of missing data, a smaller but more complete supermatrix was constructed by choosing a taxon occupancy of 11 (i.e., orthologs present in 11 or more taxa were selected) and removing *A. gigas* to obviate entirely the effects of the large amount of missing data in this taxon. This yielded a matrix composed of 61 orthologs with a gene occupancy per species ranging from 89% to 100%. The total number of amino acid sites after ZORRO-masked ortholog concatenation for this matrix was 13,106.

Centipede Phylogeny and Dating

All phylogenetic analyses of the three constructed supermatrices agree on the monophyly of Chilopoda, Pleurostigmophora, and a clade composed of Lithobiomorpha, Scolopendromorpha, and Geophilomorpha, thus placing Craterostigmomorpha as the sister group of the remaining pleurostigmophorans (fig. 2, supplementary fig. S1, Supplementary Material online). Resampling support for Chilopoda and for the clade containing Lithobiomorpha, Scolopendromorpha, and Geophilomorpha is absolute and is also high for Epimorpha and a clade composed of Craterostigmomorpha and the remaining pleurostigmophorans in most analyses (fig. 4).

Both the concatenated matrix overall and each taxon within it show evidence of substantial compositional heterogeneity in a χ^2 test based on a simulated homogeneous null distribution (Foster 2004) (P values for all taxa were $<10^{-2}$). However, a quartet supernetwork constructed from a more restricted subset of genes that individually pass this compositional heterogeneity test, and which additionally fall into the least saturated quartile of these stationary genes (Foster 2004), demonstrates a network topology broadly comparable (but see later) with the tree topology of our concatenated analyses (fig. 5), indicating that the compositional heterogeneity, though present, is likely not driving the recovered topology.

Furthermore, although missing data are distributed in a complex manner across the tree, the number of genes potentially decisive for each internal node is high (e.g., >800 for all nodes in Chilopoda in the 1,934 gene analysis), indicating that the high support recovered in analyses of our largest matrices is not likely an artifact of the pattern of missing data (Dell'Ampio et al. 2014; see also supplementary fig. S1, Supplementary Material online).

Our individual maximum likelihood (ML) gene-tree analyses display a complex pattern of topological discordance and missing data. Using quartet supernetworks, we visualized the predominant splits within these gene trees, both for the complete set considered in our largest supermatrix and for the reduced set of compositionally homogeneous, least-saturated genes (figs. 4 and 5). Network topologies are similar in both sets, although there is a greater degree of conflict displayed (particularly among outgroup species) in the reduced set, perhaps the result of increased sampling error in this more limited set of slowly evolving genes (Betancur-R et al. 2014). Although both supernetworks display a tree-like pattern of splits reminiscent in many respects to the concatenated topologies (fig. 2), they also both suggest the existence of considerable among-gene conflict in the relative position of *Craterostigmus*, with some genes suggesting a sister-group relationship with *Scutigera*, in contrast to the concatenated topology. We therefore quantified the number of genes congruent with either topology and also with a number of salient hypotheses on centipede interrelationships (fig. 6). These gene support frequencies largely favor the splits recovered in the concatenated topology, with, for example, 30.3% of genes grouping Craterostigmomorpha and the remaining pleurostigmophorans, and only 21.1% grouping Scutigeromorpha and Craterostigmomorpha. However, relationships within the clade formed by Lithobiomorpha and Epimorpha are somewhat more conflicted, with 24.6% of genes supporting Epimorpha versus 21.3% supporting Lithobiomorpha + Geophilomorpha.

The dated phylogenies for the 389 gene matrix under either an autocorrelated log normal (fig. 3, top) or uncorrelated gamma model (fig. 3, bottom) support a diversification of Chilopoda between the Early Ordovician and the Middle Devonian, diversification of Pleurostigmophora in the Middle Ordovician to Early Carboniferous, and the diversification of the clade uniting Lithobiomorpha, Scolopendromorpha, and Geophilomorpha in the Silurian-Carboniferous. The diversification of Epimorpha ranges from the Early Devonian to the early Permian (fig. 3).

Discussion

Phylogenomic Hypothesis of Centipede Interrelationships

Resolving the Tree of Life has been prioritized as one of the 125 most important unsolved scientific questions in 2005 by *Science* (Kennedy 2005), and the advent of phylogenomics has aided in resolving many contentious aspects in animal phylogeny (e.g., Philippe and Telford 2006; Dunn et al. 2008; Bleidorn et al. 2009; Hejnol et al. 2009; Meusemann et al. 2010; Kocot et al. 2011; Rehm et al. 2011; Smith et al. 2011; Struck et al. 2011; Hartmann et al. 2012; von Reumont et al. 2012). This is not without controversy, and several phenomena unique to genome-scale data have been identified as negatively impacting tree reconstruction in this paradigm, perhaps foremost among these being gene occupancy (missing data) (Roure et al. 2013; Dell'Ampio et al. 2014), taxon sampling (Pick et al. 2010), and quality of data (e.g., proper ortholog assignment and controls on exogenous contamination) (Philippe et al. 2011; Salichos and Rokas 2011). In addition, assumptions underlying concatenation (Salichos and Rokas 2013) and model (mis-)specification (Lartillot and Philippe 2008) have also been identified as possible pitfalls for tree reconstruction in a phylogenomic framework. Finally, it has been shown that conflict may exist between classes of genes, with some advocating exclusive usage of slowly

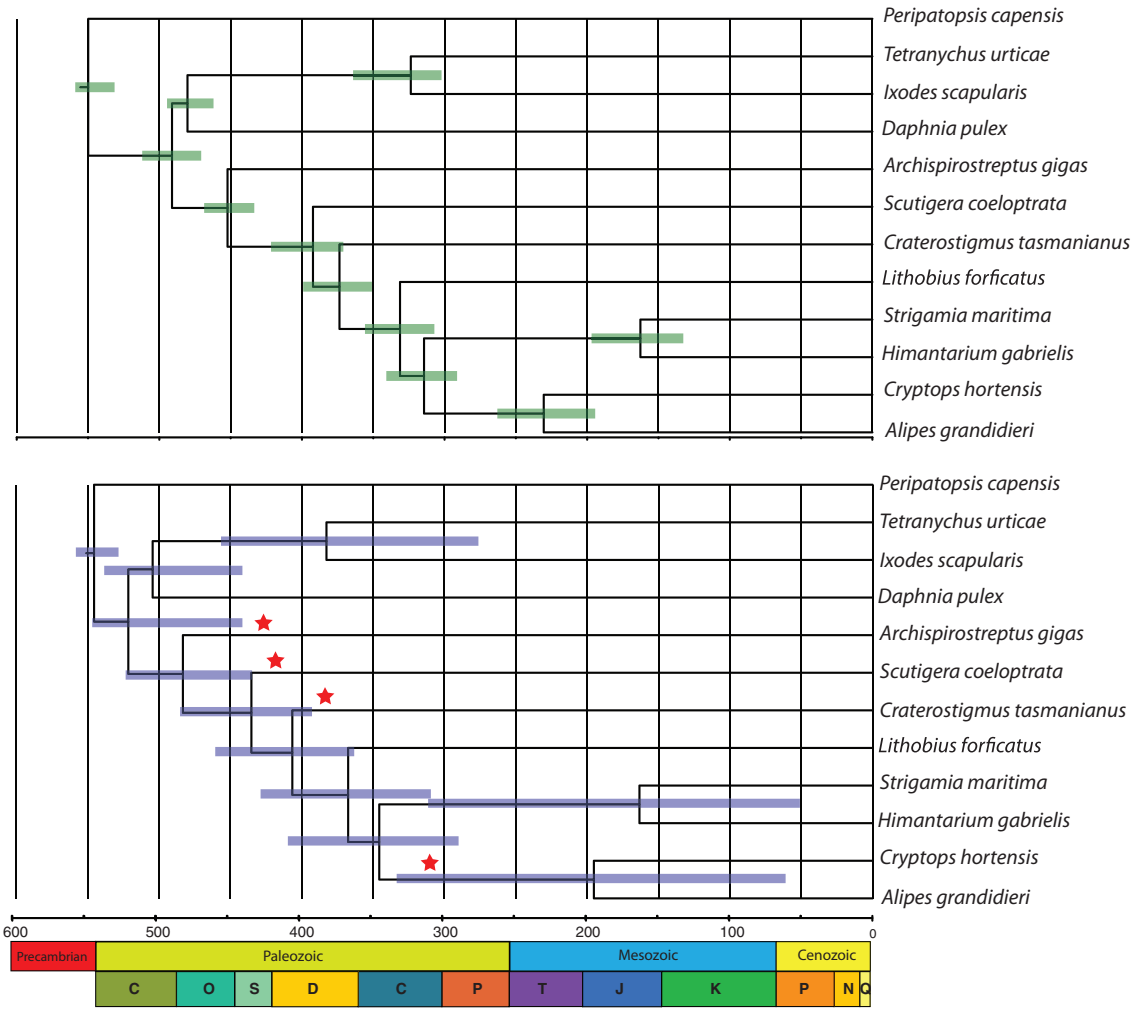

Fig. 3. Chronogram of centipede evolution for the 389-gene data set with 95% highest posterior density (HPD) bar for the dating under autocorrelated log normal (top) and uncorrelated gamma model (bottom). Nodes that were calibrated with fossils are indicated with a star.

evolving genes to resolve deep metazoan splits (e.g., Nosenko et al. 2013), although others believe that slow-evolving genes may not be able to resolve deep splits due to the lack of sufficient signal.

We have worked carefully to try to address all these issues in our analyses. First, we have minimized the amount of missing data and worked with one of the most complete phylogenomic matrices for nonmodel invertebrates, based on genomes and deeply sequenced transcriptomes, with the exception of the millipede outgroup transcriptome (Meusemann et al. 2010). Although debate exists as to whether or not large amounts of missing data negatively affect phylogeny reconstruction (Hejnol et al. 2009; Lemmon et al. 2009; Dell'Ampio et al. 2014), Roure et al. (2013) concluded that although the effects of missing data can have a negative impact in certain situations, other issues have a more direct effect, including modeling among-site substitutional heterogeneity; they recommended graphically displaying the amount of missing data in phylogenetic trees, as done for example by Hejnol et al. (2009) or as presented here (fig. 2). Given our low levels of missing data and high matrix completeness (supplementary table S2, Supplementary Material online), it is unlikely that our results are affected by gene occupancy or missing data. Likewise, taxon sampling has been optimized to represent all major centipede lineages. No major hypothesis on deep centipede phylogenetics remains untested with our sampling, although it may be advisable to add additional species for each order in future studies, especially of Scutigeromorpha and Lithobiomorpha, whose positions show incomplete support or substantial among-gene conflict, perhaps because they are represented by a single species each (figs. 4–6).

It has been argued that in the case of strong gene-tree incongruence (e.g., from incomplete lineage sorting, horizontal gene transfer, or cryptic duplication/loss), concatenated analysis may be misleading, causing an erroneous species tree to be inferred, and furthermore with inflated resampling support, giving no indication of intergene conflict (e.g., Jeffroy et al. 2006; Kubatko and Degnan 2007; Nosenko et al. 2013; Salichos and Rokas 2013). Many therefore question exclusive reliance on concatenation, and instead argue for the application of a model of the source of incongruence (e.g., the emerging species tree-methods based on the multispecies coalescent; Edwards 2009), or for discarding data by selecting

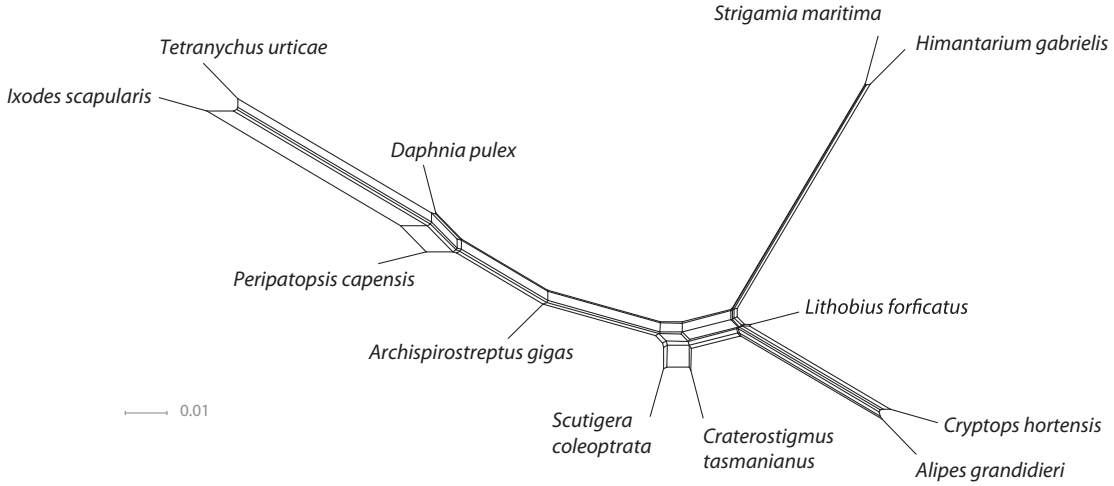

Fig. 4. Supernetwork representation of quartets derived from individual ML gene trees, for all 1,934 genes concatenated in the supermatrix presented in figure 2. Phylogenetic conflict is represented by nontree-like splits (e.g., Scutigeromorpha–Craterostigmomorpha).

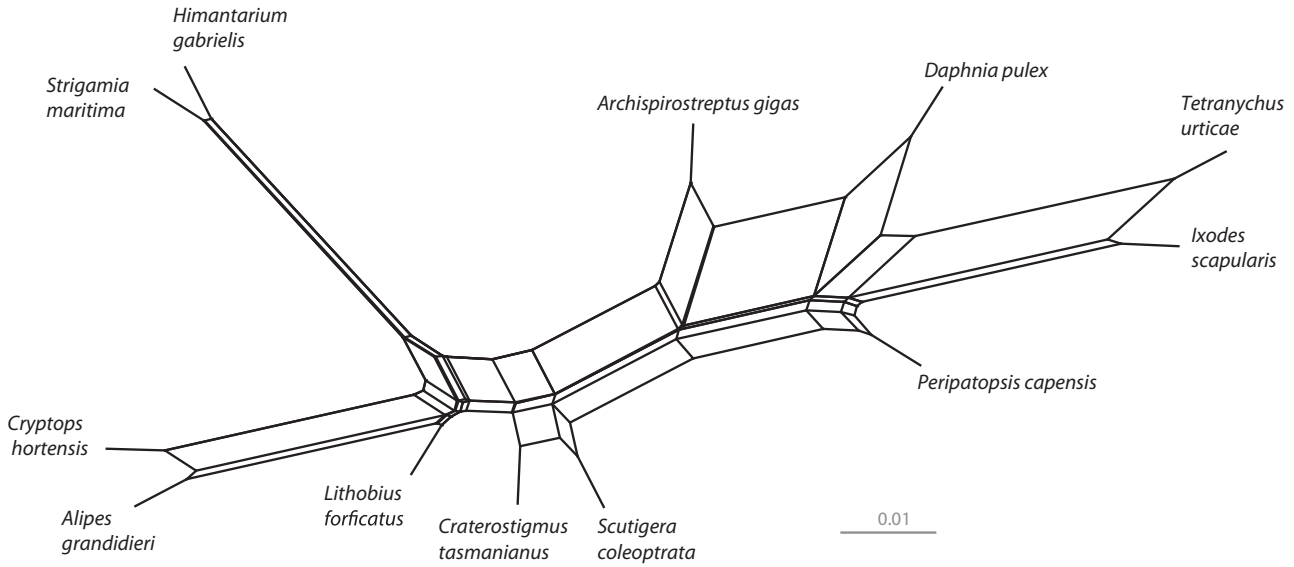

Fig. 5. Supernetwork representation of quartets derived from individual ML gene trees, considering only 389 genes showing evidence of compositional homogeneity and relatively low substitutional saturation.

genes with strong phylogenetic signal and demonstrated absence of significant incongruence when reconstructing ancient divergences. To address this possibility, we also adopted a gene-tree perspective, inferring individual ML trees for each gene included in our concatenated matrix. However, many available methods of investigating gene-tree incongruence (including recent metrics such as internode certainty and its derivatives; Salichos and Rokas 2013; Salichos et al. 2014) can only be calculated for genes that contain the same set of taxa, an assumption that our data set (and many similar data sets generated by highly parallel sequencing) assuredly violates.

We therefore visualized the dominant bipartitions among these gene trees by constructing a supernetwork (a generalization of supertrees permitting reticulation where intergene conflicts exist) using the SuperQ method (Grünewald et al. 2013), which decomposes all gene trees into quartets and then infers a supernetwork from these quartets, assigning edge lengths by examining quartet frequencies. We inferred supernetworks from both ML gene trees (figs. 4 and 5) and the majority rule consensus (MRC) trees from the bootstrap replicates of each gene (figure not shown). These supernetworks display a predominantly tree-like structure, topologically resembling our concatenated species trees; however, they indicate the presence of intergene conflict in the relative positions of Craterostigmomorpha and Scutigeromorpha, with some genes uniting these taxa as a clade, in contrast to our species tree (figs. 4 and 5). Interestingly, the same overarching network structure is present in both the ML-tree-based and bootstrap MRC-based networks, indicating that these conflicts are not merely the result of stochastic sampling error. To examine how these phylogenetic conflicts are distributed among our input genes, we employed Conclustador, which automatically assigns genes into

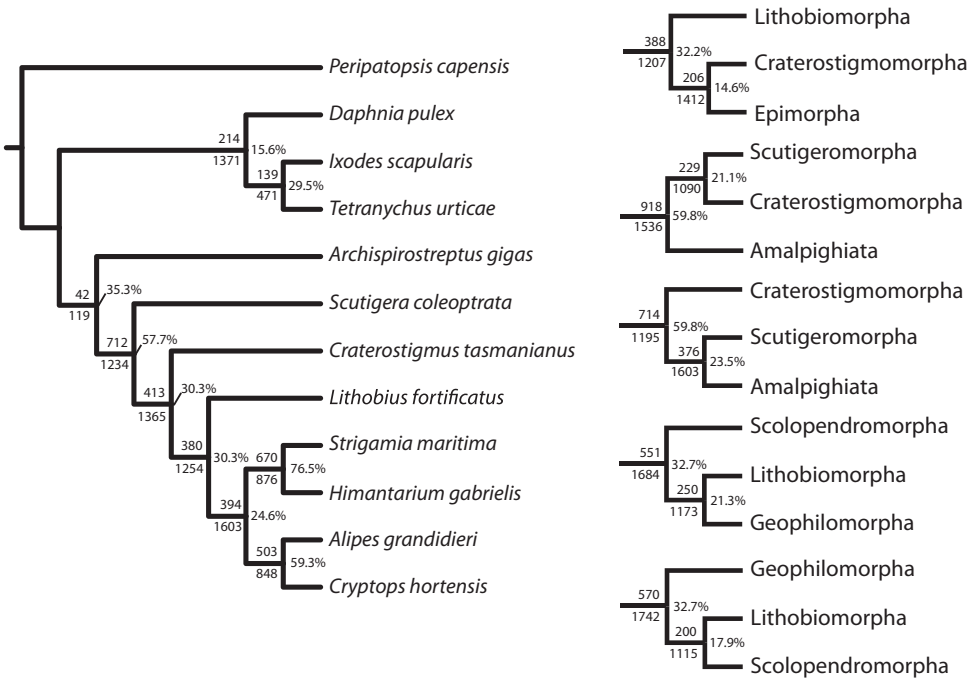

Fig. 6. Counts of potentially decisive genes and (within this set) counts of actually congruent genes computed for each node in the concatenated topology (left) as well as for select alternative hypotheses of centipede interrelationships (right). Potentially decisive genes printed below node, congruent genes in this set printed above node, and gene support percentage printed to side.

exclusive clusters if they display significant (and well supported) areas of local topological incongruence by examining bipartition distances among and within pseudoreplicate trees for each gene (Leigh et al. 2011). Intriguingly, this method did not find evidence for any well-supported local incongruence between genes in our data set (i.e., only one cluster containing all genes was selected). Although at first glance this result is at odds with the conflicts observed in our supernetwork analyses, this seeming paradox could be reconciled if most or all genes display the same patterns of topological conflict within their bootstrap replicates.

To address gene-tree conflict and the distribution of missing data in a more quantitative manner, we counted the potentially decisive genes and the frequency of genes within this set that are congruent with both the nodes in our concatenated topology and with several alternative hypotheses (fig. 6). Large numbers of genes (from 1,115 to 1,742; fig. 6) are available to test interordinal relationships within Chilopoda. The gene support frequencies within these potentially decisive sets also clearly favor the nodes within our concatenated topology over competing hypotheses, although we note that the potentially decisive gene sets are not directly comparable across nodes, as they contain different populations of genes depending on the particular taxonomic distribution of missing data. We also emphasize that although the gene support frequencies appear relatively low (i.e., usually <50%) even for nodes in our concatenated topology, these frequencies were derived from counts of fully bifurcating ML trees, many of which contain splits that are poorly supported due to limited-length alignments and/or inappropriate rates of evolution for the splits in question; we hypothesize that these gene support frequencies would appear even more decisively in favor of the concatenated topology if considering only well-supported splits.

The LG4X + F and CAT + GTR models we employed are both able to model site-heterogeneous patterns of molecular evolution to a degree and are therefore better able to detect multiple substitutions, an important issue for this data set, as the gene selection procedure we employed was agnostic to the rate of molecular evolution (Lartillot and Philippe 2004; Philippe et al. 2011). Both, however, also assume stationarity of amino acid frequencies through time, an assumption rejected by a sensitive statistical test for the 1,934 gene supermatrix. To discern whether substitutional saturation and compositional heterogeneity were influencing our phylogenetic results, we considered a subset of 389 genes that pass individual tests of compositional heterogeneity and which furthermore fall in the least saturated quartile of this stationary set. A quartet supernetwork built from ML trees drawn only within this subset of genes (fig. 5) resembles the ingroup tree topology inferred from our concatenated analyses (fig. 2) in supporting the monophyly of Lithobiomorpha + Epimorpha. In the 389-gene quartet supernetwork, however, we observe essentially the same gene-tree conflict as in the 1,934 gene quartet supernetwork, that is, disagreement on the relative position of *Craterostigmus* with respect to Scutigeromorpha and the clade comprising the remaining pleurostigmophoran orders. This conflict could be attributable in part to exacerbation of gene-tree incongruence reported for slowly evolving partitions in analyses of genomic data sets, due to the confluent effect on reducing real internode length and the theoretical expectation of greater incongruence for short internodes (Salichos and Rokas 2013; but see Betancur-R et al. 2014).

Finally, our amino acid level, transcriptome-based phylogeny is highly similar to previous hypotheses on centipede phylogenetics using much smaller sets of genes (Edgecombe et al. 1999; Giribet et al. 1999; Edgecombe and Giribet 2004; Mallatt and Giribet 2006; Murienne et al. 2010) and finds strong support for such accepted groups as Pleurostigmophora, Epimorpha, Scolopendromorpha, and Geophilomorpha, thus supporting the idea that our morphologically surprising result for Lithobiomorpha + Epimorpha is not artifactual.

Evolutionary Implications

Two salient evolutionary implications of this study are 1) the basal position of Craterostigmomorpha with respect to Lithobiomorpha, a result anticipated in prior phylogenetic analyses based on molecular data (Mallatt and Giribet 2006; Regier et al. 2010), and 2) the superior reconciliation of the new topology with the fossil record of centipedes (Edgecombe 2011b), thus improving upon prior dating studies (Murienne et al. 2010). The former has consequences in our interpretation of an important number of morphological traits thought to be shared by the clade Phylactometria (see Implications for Morphological and Behavioral Characters section) and requires a taxonomic proposal for a clade including Lithobiomorpha, Scolopendromorpha, and Geophilomorpha. We name this group Amalpighiata, on the basis of the three constituent orders lacking the supernumerary Malpighian tubules found in Scutigeromorpha and Craterostigmomorpha. Amalpighiata appears well supported and shows no strong conflict among the analyzed genes. Among-gene conflict, and also relatively lower nodal support in the concatenated analysis, exists for the relative positions of *Craterostigmus* and *Scutigera*, although the monophyly of Pleurostigmophora (Amalpighiata + Craterostigmomorpha) receives support in several analyses

The second implication of the Amalpighiata hypothesis accounts for one of the inconsistencies in the centipede fossil record under the Phylactometria hypothesis. Stem-group Lithobiomorpha are predicted to have originated by at least the Middle Devonian (under either of the two competing hypotheses for the relationships of the Devonian *Devonobius*) (e.g., Murienne et al. 2010), yet no Paleozoic or even Mesozoic fossils are known for Lithobiomorpha. Their fossil record does not extend back further than the Eocene (when they are represented by a few species of Lithobiidae in Baltic amber). This is no doubt a gross underestimate of their real age; even the alternative Amalpighiata tree dates the split between Lithobiomorpha and Epimorpha to at least the Upper Carboniferous (constrained by the earliest known Epimorpha: Scolopendromorpha) (fig. 3). Still, the Siluro-Devonian fossil record of Chilopoda consists only of Scutigeromorpha and the extinct Devonobiomorpha, which are allied to either Craterostigmomorpha (Borucki 1996) or Epimorpha (Shear and Bonamo 1988; Murienne et al. 2010). The former hypothesis is consistent with an earlier divergence of Craterostigmomorpha than Lithobiomorpha and could thus account for the lack of mid-Paleozoic fossil lithobiomorphs.

That said, the scattered records of Chilopoda in Mesozoic shales, lithographic limestones, and ambers, which include exemplars of Scutigeromorpha, Scolopendromorpha, and Geophilomorpha but not Lithobiomorpha (reviewed by Edgecombe 2011b) suggests that lithobiomorphs are underrepresented in the fossil record as a whole, and their apparent absence in the Paleozoic is likely a taphonomic artifact. The Amalpighiata tree dates the origins of centipede orders with better fit to their first known fossil occurrences, and divergence dates are concordant with the fossil record, perhaps indicating that the Paleozoic fossil record of centipedes is not as incomplete as previously thought. An additional interesting point of our data set is that the diversification of myriapods is Silurian-Devonian for the analysis under the autocorrelated log normal model, and only our uncorrelated gamma model supports a Cambrian terrestrialization event, as suggested recently for this lineage (Rota-Stabelli et al. 2013). Additional transcriptomes should allow for a better estimation of the diversification dates of each order but should not affect their origin. Should additional taxon sampling continue to yield substantial among-gene incongruence and a succession of interordinal divergences closely spaced together in time, consistent with the scenario of an rapid radiation, we note that biological (as opposed to methodological/artifactual) explanations for the incongruence, such as incomplete lineage sorting, may need to be considered.

Implications for Morphological and Behavioral Characters

A substantial list of characters has been tabled in defense of a sister-group relationship between *Craterostigmus* and Epimorpha. They include the following putative synapomorphies (Edgecombe 2011a):

1) Brooding involving the mother guarding the egg cluster by wrapping her body around it.
2) Extraordinarily high cell numbers in the lateral ocelli (e.g., more than 1,000 retinula cells).
3) Proximal retinula cells partly with monodirectional rhabdomeres.
4) Maxillary nephridia lacking in postembryonic stadia.
5) Rigidity of the forcipules, a character complex including the forcipular pleurite arching over the coxosternite, the hinge of the coxosternite being sclerotized and inflexible, the coxosternite deeply embedded into the cuticle above the first pedigerous trunk segment, and a truncation of sternal muscles in the forcipular segment rather than extending into the head.
6) Presternites distinct.
7) Separate sternal and lateral longitudinal muscles.
8) Lateral testicular vesicles linked by a central, posteriorly extended deferens duct.
9) Coxal organs confined to the last leg pair.
10) Internal valves formed by lips of the ostia projecting deeply into the heart lumen.

Under the alternative Amalpighiata hypothesis in which Lithobiomorpha rather than *Craterostigmus* is sister group of Epimorpha, these characters are implied to be homoplastic. They are either general characters of Pleurostigmophora as a whole that have been secondarily modified (lost, subject to reversal, or otherwise transformed) in Lithobiomorpha or they have been convergently acquired by *Craterostigmus* and Epimorpha.

Craterostigmus has a single anamorphic stage in its life cycle; it hatches from the egg as a 12-legged instar and acquires the adult number of 15 leg-pairs—the plesiomorphic number of legs in centipedes—in the succeeding instar. This contrasts with numerous anamorphic stages in Scutigeromorpha and Lithobiomorpha, which hatch with 4 or 6–8 leg pairs, respectively, and keep adding segments until reaching the definitive 15 leg pairs. The Phylactometria hypothesis has generally viewed the "reduced hemianamorphosis" (Borucki 1996) of *Craterostigmus* as a transitional stage in the evolution of complete epimorphosis in Epimorpha. The alternative hypothesis in which *Craterostigmus* is sister group of all other Pleurostigmophora rejects this interpretation in favor of two independent reductions of anamorphic instars (or secondary reacquisition of them in Lithobiomorpha). Apropos, posterior segmentation in arthropods has been shown to be evolutionarily labile, as demonstrated by the extreme case of posterior segment reduction during mite embryogenesis, with concomitant loss of certain gene expression domains, or entire genes altogether, in the genome of the mite *Tetranychus urticae* (Grbic et al. 2011).

At present, few characters have been identified that could be reinterpreted as shared derived characters of Lithobiomorpha and Epimorpha. One noteworthy example is the absence of supernumerary Malpighian tubules. In Chilopoda, these occur only in Scutigeromorpha and in *Craterostigmus*, and their presence in these two orders alone has been regarded as a plesiomorphic feature (Prunescu and Prunescu 1996; Prunescu and Prunescu 2006). Their inferred loss in Lithobiomorpha and Epimorpha would imply a single step under the favored topology in this study and gives the name to the clade Amalpighiata.

Other potential apomorphies of Lithobiomorpha and Epimorpha involve characters that are shared by Lithobiomorpha and Scolopendromorpha alone and are thus required to be lost or otherwise modified in Geophilomorpha.

For example, *Craterostigmus* shares three multicusped teeth in the mandible with Scutigeromorpha, in contrast to four and five such teeth on opposing mandibles in Lithobiomorpha and Scolopendromorpha. Geophilomorpha primitively lack teeth (Koch and Edgecombe 2012), so if a transformation from three teeth (common ancestor of Chilopoda) to four/five teeth (common ancestor of Lithobiomorpha + Epimorpha) were envisioned, the loss of mandibular teeth in Geophilomorpha is a necessary part of the transformation series. A more homoplastic character that could be postulated in support of Amalpighiata as a clade is the presence of marginated tergites. These are lacking in *Craterostigmus*, are ubiquitous in Lithobiomorpha, and are variably present in Scolopendromorpha (present in Scolopocryptopinae and Scolopendridae). Their absence in Geophilomorpha and several blind scolopendromorph clades (Newportiinae, Cryptopidae, and Plutoniumidae), however, means that their status as a character for Amalpighiata would demand a few instances of reversal/loss. As such, though it must be conceded that this scheme introduces considerable homoplasy from the perspective of morphology, a few characters can be suggested as potential synapomorphies for Amalpighiata.

Materials and Methods

Sample Collection

We sampled six specimens representing the main groups of centipedes (MCZ voucher numbers in brackets): *Scutigera coleoptrata* (IZ-20415) (Scutigeromorpha), *Lithobius forficatus* (IZ-131534) (Lithobiomorpha), *Craterostigmus tasmanianus* (IZ-128299) (Craterostigmomorpha), *Alipes grandidieri* (IZ-130616), *Cryptops hortensis* (IZ-130583) (Scolopendromorpha), and *Himantarium gabrielis* (IZ-131564) (Geophilomorpha). Information about the sampling localities collection details can be found in the MCZ online collections database (http://mczbase.mcz.harvard.edu, last accessed April 1, 2014). The following taxa were sampled as outgroups: *Peripatopsis capensis* (Onychophora; a transcriptome) and *Ixodes scapularis* and *T. urticae* (Arthropoda, Arachnida, and Acari; complete genomes). Samples were preserved in RNAlater (Ambion) or flash frozen in liquid nitrogen immediately after collection. A portion of the central part of each animal (including the body and legs) or the anterior part (in *S. coleoptrata*, including the head and a part of the trunk) was dissected and stored at −80 °C until RNA extraction. Three more taxa retrieved from external sources were also included. The genome of *Strigamia maritima* (Geophilomorpha) was downloaded from http://metazoa.ensembl.org/Strigamia_maritima/Info/Index (last accessed April 1, 2014). The genome of *Daphnia pulex* was retrieved from http://metazoa.ensembl.org/Daphnia_pulex/Info/Index (last accessed April 1, 2014). The transcriptome of *A. gigas* generated by 454 pyrosequencing was retrieved from Meusemann et al. (2010).

mRNA Extraction

Total RNA was extracted with a standard trizol-based method using TRIzol (Life Sciences) following the manufacturer's protocol. Tissue fragments were disrupted with a drill in 500 µl of TRIzol using an RNAse-free plastic pestle for grinding. TRIzol was added up to a total volume of 1 ml. After 5 min incubating at room temperature (RT), 100 ml of bromochloropropane was mixed by vortexing. After incubation at RT for 10 min, the samples were centrifuged at 16,000 rpm for 15 min at 4 °C. Afterward, the upper aqueous layer was recovered, mixed with 500 ml of isopropanol, and incubated at −20 °C overnight. Total RNA precipitation was made as follows: samples were centrifuged for 15 min at 16,000 rpm at 4 °C, the pellet was washed twice with 1 ml of 75% isopropanol and centrifuged at 16,000 rpm at 4 °C for

15 and 5 min in the first and second washing step, respectively, air dried and eluted in 100 µl of RNA Storage solution (Ambion). mRNA purification was done with the Dynabeads mRNA Purification Kit (Invitrogen) following manufacturer's instructions. After incubation of total RNA at 65 °C for 5 min, the samples were incubated for 30 min with 200 µl of magnetic beads in a rocker and washed twice with washing buffer. This step was repeated again with an incubation time of 5 min. mRNA was eluted in 15 µl of Tris-HCl buffer. Quality of mRNA was measured with a pico RNA assay in an Agilent 2100 Bioanalyzer (Agilent Technologies). Quantity was measured with a RNA assay in Qubit fluorometer (Life Technologies).

cDNA Library Construction and Next-Generation Sequencing

cDNA libraries were constructed with the TruSeq for RNA Sample Preparation kit (Illumina) following the manufacturer's instructions. Libraries for *C. tasmanianus* and *S. coleoptrata* were constructed in the Apollo 324 automated system using the PrepX mRNA kit (IntegenX). The samples were marked with a different index to allow pooling for sequencing. Concentration of the cDNA libraries was measured through a dsDNA High Sensitivity (HS) assay in a Qubit fluorometer (Invitrogen). Library quality and size selection were checked in an Agilent 2100 Bioanalyzer (Agilent Technologies) with the HS DNA assay. The samples were run using the Illumina HiSeq 2500 platform with paired-end reads of 150 bp at the FAS Center for Systems Biology at Harvard University.

Raw Data Sanitation and Sequence Assembly

Demultiplexed Illumina HiSeq 2500 sequencing results, in FASTQ format, were retrieved from the sequencing facility via FTP. The files were decompressed and imported into CLC Genomics Workbench 5.5.1, retaining read header information, as paired read files. Each sample was quality filtered according to a threshold average quality score of 30 based on a Phred scale and adaptor trimmed using a list of known adaptors provided by Illumina. All reads shorter than 25 bp were discarded and the resulting files were exported as a single FASTQ file.

To simplify the assembly process, ribosomal RNA (rRNA) was filtered out. All known metazoan rRNA sequences were downloaded from GenBank and formatted into bowtie index using "bowtie-build." Each sample was sequentially aligned to the index allowing up to two mismatches via Bowtie 1.0.0 (Langmead et al. 2009), retaining all unaligned reads in FASTQ format. Unaligned results, stored as a single file, were imported into Geneious 6.1.6 (Kearse et al. 2012), paired using read header information, and exported as separate files for left and right mate pairs. This process was repeated for each sample.

De novo assemblies (strand specific in *C. tasmanianus* and *S. coleoptrata*) were done individually per organism in Trinity (Grabherr et al. 2011; Haas et al. 2013) using paired read files and default parameters. Raw reads and assembled sequences have been deposited in the National Center for Biotechnology Information Sequence Read Archive and Transcriptome Shotgun Assembly databases, respectively (table 2).

Identification of Candidate Coding Regions

Redundancy reduction was done with CD-HIT-EST (Fu et al. 2012) in the raw assemblies (95% global similarity). Resulting assemblies were processed in TransDecoder (Haas et al. 2013) to identify candidate ORFs within the transcripts. Predicted peptides were then processed with a further filter to select only one peptide per putative unigene, by choosing the longest ORF per Trinity subcomponent with a Python script, thus removing the variation in the coding regions of our assemblies due to by alternative splicing, closely related paralogs, and allelic diversity. Peptide sequences with all final candidate ORFs were retained as multifasta files. Sequence data for *Strigamia maritima* and *D. pulex* were obtained from genome assemblies, as opposed to the de novo assembled transcriptomes. Annotated predicted peptide sequences for *D. pulex* were downloaded from "wFleaBase," and sequences were clustered using CD-HIT at a similarity of 98% for redundancy reduction. Peptide sequences for *Strigamia maritima* were downloaded from Ensembl and also clustered at 98% similarity. Results for both *D. pulex* and *Strigamia maritima* were converted to a single line multifasta file.

Orthology Assignment

We assigned predicted ORFs into orthologous groups across all samples using OMA stand-alone v0.99t (Altenhoff et al. 2011; Altenhoff et al. 2013). The advantages of the algorithm of OMA over standard bidirectional best-hit approaches rely on the use of evolutionary distances instead of scores, consideration of distance inference uncertainty and differential gene losses, and inclusion of many-to-many orthologous relations. The ortholog matrix is constructed from all-against-all

Table 2. Accession Numbers for the Raw Transcriptomes and Assemblies from the Taxa Newly Sequenced in This Study.

	BioProject	BioSample	Experiment	Run
Scutigera coleoptrata	PRJNA237135	SAMN02614634	SRX462011	SRR1158078
Craterostigmus tasmanianus	PRJNA237134	SAMN02614633	SRX461877	SRR1157986
Lithobius forficatus	PRJNA237133	SAMN02614632	SRX462145	SRR1159752
Himantarium gabrielis	PRJNA237131	SAMN02614631	SRX461787	SRR1159787
Alipes grandidieri	PRJNA179374	SAMN01816532	SRX205685	SRR619311
Cryptops hortensis	PRJNA237130	SAMN02614630	SRX457664	SRR1153457
Peripatopsis capensis	PRJNA236598	SAMN02598680	SRX451023	SRR1145776

Smith-Waterman protein alignments. The program identifies the so called "stable pairs," verifies them, and checks against potential paralogous genes. In a last step, cliques of stable pairs are clustered as groups of orthologs. All input files were single-line multifasta files, and the parameters.drw file specified retained all default settings with the exception of "NP," which was set at 300. We parallelized all-by-all local alignments across 256 cores of a single compute node, implementing a custom Bash script allowing for execution of independent threads with at least 3 s between each instance of OMA to minimize risk of collisions.

Phylogenomic Analyses and Congruence Assessment

We constructed three different amino acid supermatrices. First, a large matrix was constructed by concatenating the set of OMA groups containing six or more taxa. To increase gene occupancy and to reduce the percentage of missing data, a second matrix was created by selecting the orthologs contained in 11 or more taxa. Because of the high percentage of missing data in *A gigas*, this taxon was eliminated from this supermatrix. This supermatrix reconstruction approach was selected over other available software (e.g., MARE) because it guarantees a minimum taxon occupancy per orthogroup, therefore resulting in a supermatrix with a good compromise between gene and taxon occupancy and missing data.

The selected orthogroups (1,934 and 61 for both matrixes, respectively) were aligned individually using MUSCLE version 3.6 (Edgar 2004). To increase the signal-to-noise ratio and improve the discriminatory power of phylogenetic methods, we applied a probabilistic character masking with ZORRO (Wu et al. 2012) to account for alignment uncertainty, run using default parameters (and using FastTree 2.1.4 [Price et al. 2010] to produce guide trees). This program first calculates the probability of all alignments that pass through a specified matched pair of residues using a pair hidden Markov model framework, and it then compares this value with the full probability of all alignments of the pair of sequences. The posterior probability value indicates how reliable the match is (i.e., highly reliable if it is close to 1, ambiguous if it is close to 0). It then sums up the probability that a particular column would appear over the alignment landscape and assigns a confidence score between 0 and 10 to each column, providing an objective measure that has an explicit evolutionary model and is mathematically rigorous (Wu et al. 2012). We discarded the positions assigned a confidence score below a threshold of 5 with a custom Python script prior to concatenation (using Phyutility 2.6; Smith and Dunn 2008) and subsequent phylogenomic analyses.

To discern if substitutional saturation and among-taxon compositional heterogeneity were affecting our phylogenetic results, we considered a third supermatrix. Using the p4 Python library (Foster 2004), we performed a test of compositional homogeneity using χ^2 statistics, simulating a null distribution using the Tree.compoTestUsingSimulations() method (nSims = 100). We used the best-scoring ML tree from our RAxML analyses as the phylogram on which to simulate data, using optimized parameters of a homogeneous, unpartitioned LG + I + Γ4 model. We also conducted compositional homogeneity tests on individual ortholog alignments, modeling a homogeneous null distribution on the ML tree for each gene, using the best-fitting model available in p4 for that gene. We characterized genewise substitutional saturation by regressing uncorrected amino acid distances on the patristic distance implied by each ML tree in a p4 script and taking the slope of this linear regression as a measure of substitutional saturation (Jeffroy et al. 2006). We calculated the third quartile of the slope from among the set of compositionally homogeneous genes with a slope between 0 and 1 and then selected the set of ML trees from genes with a slope above this value (0.465) for inclusion in a quartet supernetwork, as described earlier. The 389 genes that passed individual tests of compositional heterogeneity and which furthermore fell in the least saturated quartile of this stationary set were concatenated into a third supermatrix.

ML inference was conducted with PhyML-PCMA (Zoller and Schneider 2013) and RAxML 7.7.5 (Berger et al. 2011). Although RAxML uses a predetermined empirical model of amino acid substitution, PhyML-PCMA estimates a model through the use of a principal component (PC) analysis. The obtained PCs describe the substitution rates that covary the most among different protein families and therefore define a semiempirically determined parameterization for an amino acid substitution model specific to each data set (Zoller and Schneider 2013). We selected 20 PCs in the PhyML-PCMA analyses and empirical amino acid frequencies. PROTGAMMALG4XF was selected as the best model of amino acid substitution for the unpartitioned RAxML analyses using a modification of the ProteinModelSelection perl wrapper to RAxML produced by A. Stamatakis (http://sco.h-its.org/exelixis/software.html, last accessed April 1, 2014). Bootstrap values were estimated with 1,000 replicates under the rapid bootstrapping algorithm.

Bayesian analysis was conducted with PhyloBayes MPI 1.4e (Lartillot et al. 2013) using the site-heterogeneous CAT-GTR model of evolution (Lartillot and Philippe 2004). Three independent Markov chain Monte Carlo (MCMC) chains were run for 5,818–7,471 cycles. The initial 2,000 trees (27–34%) sampled in each MCMC run prior to convergence (judged when maximum bipartition discrepancies across chains < 0.1) were discarded as the burn-in period. A 50% majority-rule consensus tree was then computed from the remaining 1,537 trees (sampled every 10 cycles) combined from the three independent runs.

To investigate potential incongruence between individual gene trees, we also inferred gene trees for each OMA group included in our supermatrix. For each aligned, ZORRO-masked OMA group, we selected a best-fitting model using a ProteinModelSelection script modified to permit testing of the recent LG4M(+ F) and LG4X(+ F) models (Le et al. 2012). Best-scoring ML trees were inferred for each gene under the selected model (with the gamma model of rate variation, but no invariant term) from 20 replicates of parsimony starting trees. One hundred traditional (nonrapid) bootstrap replicates for each gene were also inferred. A majority rule consensus for each gene was constructed using the SumTrees

script from the DendroPy Python library (Sukumaran and Holder 2010). To visualize predominant intergene conflicts for both ML and bootstrap MRC trees, we employed SuperQ v.1.1 (Grünewald et al. 2013), selecting the "balanced" edge-weight optimization function, and applying no filter; SplitsTree v.4.13.1 (Huson and Bryant 2006) was then used to visualize the resulting NEXUS files. To investigate the distribution of strongly supported conflict among our gene trees, we employed Conclustador v.0.4a (Leigh et al. 2011), presently the only automated incongruence-detection algorithm readily applicable to data sets of this size. We applied the spectral clustering algorithm on the bootstrap replicates for each OMA group, limiting the maximum number of clusters to 15.

To quantify the number of genes potentially decisive for a given split and also the number within this set that actually support the split in question (fig. 6), we employed a custom Python script to parse the set of individual gene trees. If a tree contained at least one species in either descendant group for a given node, plus at least two distinct species basal to the node in question, it was considered potentially decisive (forming minimally a quartet; Dell'Ampio et al. 2014); if the descendants were monophyletic with respect to their relative outgroups, that gene tree was considered congruent with the node in question (although this count is agnostic to the topology within the node in question). To quantify the relative support for different hypotheses within our individual gene-tree analysis, we displayed the above counts for all nodes in the 1,934 gene ML topology (fig. 6), and also for select alternative scenarios of centipede interrelationships, both including historical hypotheses (see Introduction) and alternative topologies for Scutigeromorpha and Craterostigmomorpha suggested by our quartet supernetworks. All Python custom scripts can be downloaded from https://github.com/claumer (last accessed April 1, 2014).

Fossil Calibrations

A few Paleozoic fossil occurrences constrain the timing of divergences between chilopod orders (for a review of the myriapod fossil record see Shear and Edgecombe 2010). The Siluro–Devonian genus *Crussolum* (Shear et al. 1998) is identified as a stem-group scutigeromorph, displaying derived characters of total-group Scutigeromorpha but lacking some apomorphies that are shared by all members of the scutigeromorph crown group (Edgecombe 2011b). The best preserved fossils of *Crussolum* are from the Early Devonian (Pragian) and Middle Devonian (Givetian) (Shear et al. 1998; Anderson and Trewin 2003), but the oldest confidently assigned record is in the Late Silurian (early Přidolí) Ludlow Bone Bed in western England (at least 419.2 My, dating the end of the Přidolí). This dating constrains the split of Scutigeromorpha from Pleurostigmophora.

The occurrence of *Devonobius delta* in the Middle Devonian of New York, at least 382.7 My (date for the end of the Givetian) constrains the divergence between Craterostigmomorpha and the remaining pleurostigmophorans. This minimum applies irrespective of whether *Devonobius* is sister group of Craterostigmomorpha or of Epimorpha, which were equally parsimonious in previous analyses (Edgecombe and Giribet 2004). The basal divergence in Epimorpha, that is, the split between Scolopendromorpha and Geophilomorpha is constrained by the oldest scolopendromorph, *Mazoscolopendra richardsoni* in Upper Carboniferous (Westphalian D) deposits of Mazon Creek, IL (at least 306 Ma). Available character data for *Mazoscolopendra* are insufficient to establish whether it is a stem- or crown-group scolopendromorph, and it only provides a minimum age for total-group Scolopendromorpha.

Mid-Silurian body fossils of Diplopoda provide constraints on the divergence of Chilopoda and Diplopoda. The divergence between these two groups is dated by the occurrence of Archipolypoda (stem-group helminthomorphs) in the late Wenlock or early Ludlow (Wilson and Anderson 2004). Based on these data (ages based on spores), the split between Chilopoda and Diplopoda is at least as old as the end of the Gorstian Stage (early Ludlow), 425.6 My.

The split between Onychophora and Arthropoda was dated between 528 My (the minimum age for Arthropoda used by Lee et al. [2013] on the basis of the earliest *Rusophycus* traces) and 558 My, used as the root of Panarthropoda (Lee et al. 2013).

Divergence Time Estimates

Divergence dates were estimated using the Bayesian relaxed molecular clock approach as implemented in PhyloBayes v.3.3f (Lartillot et al. 2013). To compare the performance of the autocorrelated log normal and uncorrelated gamma models, we estimated the divergence dates under both models in the 389-gene data set. A series of calibration constraints specified above in "Fossil calibrations" (see also Murienne et al. 2010; Edgecombe 2011b) were used with soft bounds (Yang and Rannala 2006) under a birth–death prior in PhyloBayes, this strategy having been found to provide the best compromise for dating estimates (Inoue et al. 2010). Two independent MCMC chains were run for 5,000–6,000 cycles, sampling posterior rates and dates every 10 cycles. The initial 20% of the cycles were discarded as burn-in. Posterior estimates of divergence dates were then computed from approximately the last 4,000 samples of each chain.

Supplementary Material

Supplementary figure S1 and tables S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

Mark Harvey kindly provided the sample of *Craterostigmus tasmanianus* and Savel Daniels the sample of *Peripatopsis capensis*. The Research Computing group at the Harvard Faculty of Arts and Sciences was instrumental in facilitating the computational resources required. Sarah Bastkowski is kindly thanked for her advice on the usage of SuperQ v.1.1. This work was supported by internal funds from the Museum of Comparative Zoology and by a postdoctoral fellowship

from the *Fundación Ramón Areces* to R.F. Five anonymous reviewers and the editor provided comments that helped to improve this article.

References

Altenhoff AM, Gil M, Gonnet GH, Dessimoz C. 2013. Inferring hierarchical orthologous groups from orthologous gene pairs. *PLoS One* 8: e53786.

Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C. 2011. OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res.* 39:D289–D294.

Anderson LI, Trewin NH. 2003. An Early Devonian arthropod fauna from the Windyfield Cherts, Aberdeenshire, Scotland. *Palaeontology* 46: 457–509.

Ax P. 2000. Multicellular animals, Volume II. The phylogenetic system of the Metazoa. Berlin (Germany): Springer.

Berger SA, Krompass D, Stamatakis A. 2011. Performance, accuracy, and Web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst Biol.* 60:291–302.

Betancur-R R, Naylor G, Ortí G. 2014. Conserved genes, sampling error, and phylogenomic inference. *Syst Biol.* 63:257–262.

Bleidorn C, Podsiadlowski L, Zhong M, Eeckhaut I, Hartmann S, Halanych KM, Tiedemann R. 2009. On the phylogenetic position of Myzostomida: can 77 genes get it wrong? *BMC Evol Biol.* 9:150.

Borucki H. 1996. Evolution und phylogenetisches system der Chilopoda (Mandibulata, Tracheata). *Vehr naturwiss Ver Hamburg.* 35:95–226.

Dell'Ampio E, Meusemann K, Szucsich NU, Peters RS, Meyer B, Borner J, Petersen M, Aberer AJ, Stamatakis A, Walzl MG, et al. 2014. Decisive data sets in phylogenomics: lessons from studies on the phylogenetic relationships of primarily wingless insects. *Mol Biol Evol.* 31: 239–249.

Dohle W. 1985. Phylogenetic pathways in the Chilopoda. *Bijdr Dierkunde.* 55:55–66.

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.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.

Edgecombe GD. 2011a. Chilopoda—phylogeny. In: Minelli A, editor. Treatise on zoology—anatomy, taxonomy, biology. The Myriapoda, Vol. 1. Leiden and Boston: Brill. p. 339–354.

Edgecombe GD. 2011b. Chilopoda—the fossil history. In: Minelli A, editor. Treatise on zoology—anatomy, taxonomy, biology. The Myriapoda, Vol. 1. Leiden and Boston: Brill. p. 355–361.

Edgecombe GD, Giribet G. 2004. Adding mitochondrial sequence data (16S rRNA and cytochrome *c* oxidase subunit I) to the phylogeny of centipedes (Myriapoda, Chilopoda): an analysis of morphology and four molecular loci. *J Zool Syst Evol Res.* 42:89–134.

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

Edgecombe GD, Giribet G. 2008. A New Zealand species of the trans-Tasman centipede order Craterostigmomorpha (Arthropoda:Chilopoda) corroborated by molecular evidence. *Invert Syst.* 22:1–15.

Edgecombe GD, Giribet G, Wheeler WC. 1999. Phylogeny of Chilopoda: combining 18S and 28S rRNA sequences and morphology. *Boletín Sociedad Entomol Aragonesa* 26:293–331.

Edwards SV. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* 63:1–19.

Foster P. 2004. Modeling compositional heterogeneity. *Syst Biol.* 53: 485–495.

Fu LM, Niu BF, Zhu ZW, Wu ST, Li WZ. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28: 3150–3152.

Giribet G, Carranza S, Riutort M, Baguñà J, Ribera C. 1999. Internal phylogeny of the Chilopoda (Myriapoda, Arthropoda) using complete 18S rDNA and partial 28S rDNA sequences. *Philos Trans R Soc Lond B Biol Sci.* 354:215–222.

Giribet G, Edgecombe GD. 2006. Conflict between data sets and phylogeny of centipedes: an analysis based on seven genes and morphology. *Proc R Soc B Biol Sci.* 273:531–538.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* 29:644–652.

Grbic M, Van Leeuwen T, Clark RM, Rombauts S, Rouzé P, Grbić V, Osborne EJ, Dermauw W, Ngoc PC, Ortego F, et al. 2011. The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. *Nature* 479:487–492.

Grünewald S, Spillner A, Bastkowski S, Bogershausen A, Moulton V. 2013. SuperQ: computing supernetworks from quartets. *IEEE/ACM Trans Comput Biol Bioinform.* 10:151–160.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc.* 8: 1494–1512.

Hartmann S, Helm C, Nickel B, Meyer M, Struck TH, Tiedemann R, Selbig J, Bleidorn C. 2012. Exploiting gene families for phylogenomic analysis of myzostomid transcriptome data. *PLoS One* 7:e29843.

Hejnol A, Obst M, Stamatakis A, Ott M, Rouse GW, Edgecombe GD, Martinez P, Baguñà J, Bailly X, Jondelius U, et al. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc R Soc B Biol Sci.* 276:4261–4270.

Hilken G. 1997. Tracheal systems in Chilopoda: a comparison under phylogenetic aspects. *Entomol Scand Suppl.* 51:49–60.

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

Inoue J, Donoghue PCJ, Yang ZH. 2010. The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. *Syst Biol.* 59:74–89.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? *Trends Genet.* 22:225–231.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28: 1647–1649.

Kennedy D. 2005. 125 [Editorial]. *Science* 309:19.

Koch M, Edgecombe GD. 2012. The preoral chamber in geophilomorph centipedes: comparative morphology, phylogeny, and the evolution of centipede feeding structures. *Zool J Linn Soc.* 165:1–62.

Kocot KM, Cannon JT, Todt C, Citarella MR, Kohn AB, Meyer A, Santos SR, Schander C, Moroz LL, Lieb B, et al. 2011. Phylogenomics reveals deep molluscan relationships. *Nature* 477:452–456.

Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst Biol.* 56:17–24.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10:R25.

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, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst Biol.* 62:611–615.

Le SQ, Dang CC, Gascuel O. 2012. Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Mol Biol Evol.* 29:2921–2936.

Lee MSY, Soubrier J, Edgecombe GD. 2013. Rates of phenotypic and genomic evolution during the Cambrian Explosion. *Curr Biol.* 23:1–7.

Leigh JW, Schliep K, Lopez P, Bapteste E. 2011. Let them fall where they may: congruence analysis in massive phylogenetically messy data sets. *Mol Biol Evol.* 28:2773–2785.

Lemmon AR, Brown JM, Stanger-Hall K, Lemmon EM. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. *Syst Biol.* 58: 130–145.

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.

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.

Müller CHG, Meyer-Rochow VB. 2006. Fine structural organization of the lateral ocelli in two species of *Scolopendra* (Chilopoda: Pleurostigmophora): an evolutionary evaluation. *Zoomorphology* 125:13–26.

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

Nosenko T, Schreiber F, Adamska M, Adamski M, Eitel M, Hammel J, Maldonado M, Müller WE, Nickel M, Schierwater B, et al. 2013. Deep metazoan phylogeny: when different genes tell different stories. *Mol Phylogenet Evol.* 67:223–233.

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol.* 9: e1000602.

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

Pick KS, Philippe H, Schreiber F, Erpenbeck D, Jackson DJ, Wrede P, Wiens M, Alié A, Morgenstern B, Manuel M, et al. 2010. Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. *Mol Biol Evol.* 27: 1983–1987.

Pocock RI. 1902. A new and annectant type of chilopod. *Quart J Microsc Sci.* 45:417–448.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2-approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490.

Prunescu C-C, Prunescu P. 1996. Supernumerary Malpighian tubules in chilopods. *Acta Myriapodol.* 169:437–440.

Prunescu C-C, Prunescu P. 2006. Rudimentary supernumerary Malpighian tubules in the order Craterostigmomorpha Pocock 1902. *Norwegian J Entomol.* 53:113–118.

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.

Regier JC, Wilson HM, Shultz JW. 2005. Phylogenetic analysis of Myriapoda using three nuclear protein-coding genes. *Mol Phylogenet Evol.* 34:147–158.

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.

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.

Roure B, Baurain D, Philippe H. 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Mol Biol Evol.* 30:197–214.

Salichos L, Rokas A. 2011. Evaluating ortholog prediction algorithms in a yeast model clade. *PLoS One* 6:e18755.

Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497:327–331.

Salichos L, Stamatakis A, Rokas A. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Mol Biol Evol.* 31(5):1261–1271.

Shear WA, Bonamo PM. 1988. Devonobiomorpha, a new order of centipeds (Chilopoda) from the middle Devonian of Gilboa, New York State, USA, and the phylogeny of centiped orders. *Am Mus Novitates.* 2927:1–30.

Shear WA, Edgecombe GD. 2010. The geological record and phylogeny of Myriapoda. *Arthropod Struct Dev.* 39:174–190.

Shear WA, Jeram AJ, Selden PA. 1998. Centiped legs (Arthropoda, Chilopoda, Scutigeromorpha) from the Silurian and Devonian of Britain and the Devonian of North America. *Am Mus Novitates.* 3231:1–16.

Shultz JW, Regier JC. 1997. Progress toward a molecular phylogeny of the centipede orders (Chilopoda). *Entomol Scand Suppl.* 51:25–32.

Smith S, Wilson NG, Goetz F, Feehery C, Andrade SCS, Rouse GW, Giribet G, Dunn CW. 2011. Resolving the evolutionary relationships of molluscs with phylogenomic tools. *Nature* 480:364–367.

Smith SA, Dunn CW. 2008. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics* 24:715–716.

Struck TH, Paul C, Hill N, Hartmann S, Hösel C, Kube M, Lieb B, Meyer A, Tiedemann R, Purschke G, et al. 2011. Phylogenomic analyses unravel annelid evolution. *Nature* 471:95–98.

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

Verhoeff KW. 1902–1925. Gliederfüssler: Arthropoda: Klasse Chilopoda. In: Klassen und Ordnungen des Tierreichs, 5, abt. 2, Buch 1. Leipzig (Germany): Akademische Verlagsgesellschaft. p. 725.

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.

Wilson HM, Anderson LI. 2004. Morphology and taxonomy of Paleozoic millipedes (Diplopoda: Chilognatha: Archipolypoda) from Scotland. *J Paleontol.* 78:169–184.

Wirkner CS, Pass G. 2002. The circulatory system in Chilopoda: functional morphology and phylogenetic aspects. *Acta Zool.* 83: 193–202.

Wu M, Chatterji S, Eisen JA. 2012. Accounting for alignment uncertainty in phylogenomics. *PLoS One* 7:e30288.

Yang Z, Rannala B. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol Biol Evol.* 23:212–226.

Zoller S, Schneider A. 2013. Improving phylogenetic inference with a semiempirical amino acid substitution model. *Mol Biol Evol.* 30: 469–479.