Decisive Data Sets in Phylogenomics: Lessons from Studies on the Phylogenetic Relationships of Primarily Wingless Insects

Emiliano Dell'Ampio,[†,1] Karen Meusemann,[*,†,2,3] Nikolaus U. Szucsich,[†,1] Ralph S. Peters,[†,4] Benjamin Meyer,[5] Janus Borner,[6] Malte Petersen,[2] Andre J. Aberer,[7] Alexandros Stamatakis,[7,8] Manfred G. Walzl,[1] Bui Quang Minh,[9] Arndt von Haeseler,[10] Ingo Ebersberger,[11] Günther Pass,[1] and Bernhard Misof[*,2]

[1]Department of Integrative Zoology, University of Vienna, Vienna, Austria

[2]Zoologisches Forschungsmuseum Alexander Koenig, Zentrum für Molekulare Biodiversitätsforschung (zmb), Bonn, Germany

[3]CSIRO Ecosystem Sciences, Australian National Insect Collection, Acton, ACT, Australia

[4]Zoologisches Forschungsmuseum Alexander Koenig, Abteilung Arthropoda, Bonn, Germany

[5]Institut für Systemische Neurowissenschaften, Universitätsklinikum Hamburg-Eppendorf, Hamburg, Germany

[6]Biozentrum Grindel & Zoologisches Museum, Universität Hamburg, Hamburg, Germany

[7]Heidelberg Institute for Theoretical Studies (HITS), Scientific Computing Group, Heidelberg, Germany

[8]Karlsruher Institut für Technologie, Fakultät für Informatik, Karlsruhe, Germany

[9]Center for Integrative Bioinformatics Vienna (CIBIV), Max F Perutz Laboratories, University of Vienna, Medical University of Vienna, Vienna, Austria

[10]Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Vienna, Austria

[11]Institute for Cell Biology and Neuroscience, Goethe-Universität Frankfurt, Frankfurt am Main, Germany

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: mail@karen-meusemann.de; b.misof.zfmk@uni-bonn.de.

Associate editor: Nicolas Vidal

Abstract

Phylogenetic relationships of the primarily wingless insects are still considered unresolved. Even the most comprehensive phylogenomic studies that addressed this question did not yield congruent results. To get a grip on these problems, we here analyzed the sources of incongruence in these phylogenomic studies by using an extended transcriptome data set. Our analyses showed that unevenly distributed missing data can be severely misleading by inflating node support despite the absence of phylogenetic signal. In consequence, only decisive data sets should be used which exclusively comprise data blocks containing all taxa whose relationships are addressed. Additionally, we used Four-cluster Likelihood Mapping (FcLM) to measure the degree of congruence among genes of a data set, as a measure of support alternative to bootstrap. FcLM showed incongruent signal among genes, which in our case is correlated neither with functional class assignment of these genes nor with model misspecification due to unpartitioned analyses. The herein analyzed data set is the currently largest data set covering primarily wingless insects, but failed to elucidate their interordinal phylogenetic relationships. Although this is unsatisfying from a phylogenetic perspective, we try to show that the analyses of structure and signal within phylogenomic data can protect us from biased phylogenetic inferences due to analytical artifacts.

Key words: **phylogenomics, ESTs, likelihood quartet mapping, conflicting hypotheses, Entognatha, Nonoculata, Ellipura, Protura, Diplura, Collembola, missing data.**

Article

Introduction

Despite enormous efforts to resolve the tree of life, several deep nodes are still considered unresolved. A good example for such problems are the unresolved phylogenetic relationships of primarily wingless insects.

Most phylogenetic studies including multigene and phylogenomic analyses have recovered the monophyly of Hexapoda, the insect clade in a broad taxonomic sense (Regier et al. 2008, 2010; von Reumont et al. 2009, 2012; Meusemann et al. 2010; Trautwein et al. 2012). Furthermore, the monophyly of Ectognatha, which comprises insects in a strict taxonomic sense, namely jumping bristletails, silverfishes and firebrats, and winged insects, is well supported (reviewed in Grimaldi 2010; Trautwein et al. 2012). By contrast, phylogenetic relationships among the entognathous primarily wingless insects, the Protura (coneheads), Collembola (springtails), and Diplura (two-pronged bristletails), are unclear. Many authors consider these entognathous insects as being monophyletic, considering entognathy in which mouth parts are concealed in gnathal pouches (first discussed in detail by Hennig 1953) to have evolved in the last common ancestor of the three groups.

Within Entognatha, either a clade uniting Protura and Collembola, referred to as Ellipura (Börner 1910), or a clade uniting Protura and Diplura, referred to as Nonoculata (Luan et al. 2005), has been proposed (Ellipura [Hennig 1953; Kristensen 1981, 1997; Shao et al. 1999; Bitsch and Bitsch 2000, 2004; Carapelli et al. 2000; Zhang et al. 2001]; Nonoculata [Giribet and Wheeler 2001; Giribet et al. 2004; Luan et al. 2005; Kjer et al. 2006; Mallatt and Giribet 2006; Misof et al. 2007; Dell'Ampio et al. 2009; von Reumont et al. 2009; Mallatt et al. 2010]). Other authors consider a paraphyly of Entognatha to be more likely, with Diplura as closest relatives to Ectognatha. Possible arguments for this hypothesis include the evolutionary origin of paired pretarsal claws and paired cerci (Kukalová-Peck 1987; Koch 1997; Beutel and Gorb 2006), the ultrastructure of the sperm (Dallai et al. 2011), and the differentiation process of the embryonic amnion (Machida 2006) in the last common ancestor of Diplura and Ectognatha.

Meusemann et al. (2010) and von Reumont et al. (2012) published the most relevant data sets and analyses covering the phylogenetic relationships among primarily wingless insects by including expressed sequence tag (EST) data of representatives of Protura, Collembola, and Diplura. Although both studies recovered the monophyly of Entognatha, Meusemann et al. found strong evidence for Protura and Diplura as closest relatives (i.e., Nonoculata) and von Reumont et al. for Protura and Collembola as closest relatives (i.e., Ellipura). These incongruent results are puzzling because taxon sampling of the primarily wingless insects is comparable in both studies, as well as the strategies used for orthology assignment, alignment masking, matrix optimization, and tree inference.

These special circumstances put us into the exceptionally favorable position to analyze possible sources of incongruence among these two large phylogenomic data sets. Most phylogenomic studies are based on concatenated supermatrices with low gene data coverage. Focusing on relationships among specific groups, many data blocks within such supermatrices therefore may not contain data for all taxa under consideration. Consequently, our starting hypothesis was that extensive missing data may mislead proper tree reconstruction. To tackle this problem, we complement the publicly available EST data of primarily wingless insects with additional EST data from representatives of Japygidae (Diplura) and Zygentoma (silverfishes and firebrats). We took particular care to concatenate a data set that contains only gene data blocks for which entognathous hexapods and outgroups had gene data coverage. We call such a data set in the following a decisive data set. Note that the term decisiveness has been used before in the context of phylogenomic data sets (Steel and Sanderson 2010; Sanderson et al. 2010), albeit based on a distinct criterion. The concatenated data set is the largest known data set covering primarily wingless insects. It was this data set that allowed us to analyze the effect of the observed uneven distribution of missing data on the extent of bootstrap support (BS). Complementary to the application of BS measures, we applied a Four-cluster Likelihood Mapping (FcLM) approach (Strimmer and von Haeseler 1997), which has been shown to be effective in disentangling signal among four groups of species. The application of bootstrapping and FcLM helped to assess the effect of the uneven distribution of missing data in indecisive data sets. Complementary to the previously mentioned analyses, we addressed the problem of incongruent signal among genes in a multigene data set by comparing tree reconstructions based on the entire decisive data set with tree reconstructions based on subsets of genes that support incongruent hypotheses. Altogether, our approach provides potential explanations for contradictory results among phylogenomic studies by pointing out underestimated sources of error and incongruence.

Results

Orthology Assignment, Alignment, and Alignment Masking

Using the reference set of 1,886 1:1 orthologous genes (OGs), we identified between 52 and 682 putative 1:1 orthologous transcripts in the transcriptome assemblies of primarily wingless hexapods (table 1) and up to 1,886 for all taxa (supplementary table S1, Supplementary Material online). We excluded 20 OGs that were present in the five reference species but absent from all other species from subsequent analyses. After alignment masking (i.e., the exclusion of multiple sequence alignment sections in which sequence similarity cannot be distinguished from random similarity of sequences), the concatenated superalignment was composed of 73 taxa with a total alignment length of 881,235 amino acid sites, partitioned into 1,866 genes (supplementary fig. S1; for gene annotations, see supplementary table S2, Supplementary Material online).

Relationships among Entognathous Hexapod Lineages

The data set *M_Ento*, which is decisive for addressing relationships among the three entognathous groups, Protura, Collembola, and Diplura (73 taxa, 117 genes, 32,883 aligned aa sites), moderately supported a clade Protura + Diplura (Nonoculata) (fig. 1). This is compatible with the results of the FcLM approach (topology T_1 favored, fig. 2). Tree reconstruction supported Collembola as closest relatives to a clade comprising Nonoculata and Ectognatha. The clade Nonoculata + Ectognatha received moderate support (fig. 1; supplementary fig. S2, Supplementary Material online).

Our tree reconstructions based on a selected optimal subset (SOS) extracted from a complete data matrix by optimizing information content and data saturation in iterative steps of gene and/or taxon exclusion (see MARE manual; Meusemann et al. 2010; Meyer and Misof 2010) (62 taxa, 253 genes, alignment length 55,429 aa positions) yielded monophyletic Entognatha with moderate support and Nonoculata with low support (fig. 3*a* and table 2; supplementary fig. S3*a*, Supplementary Material online). It should be kept in mind that this SOS is indecisive for addressing the relationships of Entognatha, with only one-third of all genes (79) of this data set being covered by all three entognathous groups (supplementary table S3, Supplementary Material online). The tree based on the data set SOS_ω, in which

Table 1. Primarily Wingless Hexapod Species Included in This Study, and Their Number of OGs in the Original Supermatrix and in Three Data Subsets.

Order	Family	Species	Source	No. of Contigs	Total no. of OGs	No. of OGs in *M_Ento*	No. of OGs in SOS	No. of OGs in SOS_{ω}
Protura	Acerentomidae	*Acerentomon* sp.[a]	NCBI[a]	1,999	191	117	91	12
Diplura	Campodeidae	*Campodea fragilis*	NCBI	6,407	370	77	116	64
Diplura	Japygidae	*Megajapyx* sp.	this study	57,602	547	105	164	89
Collembola	Neanuridae	*Anurida maritima*	NCBI	3,504	328	55	105	60
Collembola	Onychiuridae	*Onychiurus arcticus*	NCBI	9,981	795	103	183	114
Collembola	Isotomidae	*Cryptopygus antarcticus*	NCBI	1,897	199	49	78	35
Collembola	Isotomidae	*Folsomia candida*	NCBI	5,967	442	60	122	78
Collembola	Entomobryidae	*Orchesella cincta*	NCBI	754	52	10	—	—
Archaeognatha	Machilidae	*Lepismachilis y-signata*	NCBI	2,288	270	60	107	54
Zygentoma	Lepismatidae	*Tricholepisma aurea*	NCBI	344	54	22	—	—
Zygentoma	Lepismatidae	*Thermobia domestica*	this study	45,358	682	96	194	124

NOTE.—*M_Ento* is the decisive data set in which all OGs are covered by Protura, Diplura, and Collembola; SOS and SOS_{ω} are indecisive to address the relationships of entognathous hexapod orders.

[a]*Acerentomon* sp.: erroneously assigned as *A. franzi* in Meusemann et al. (2010) and NCBI.

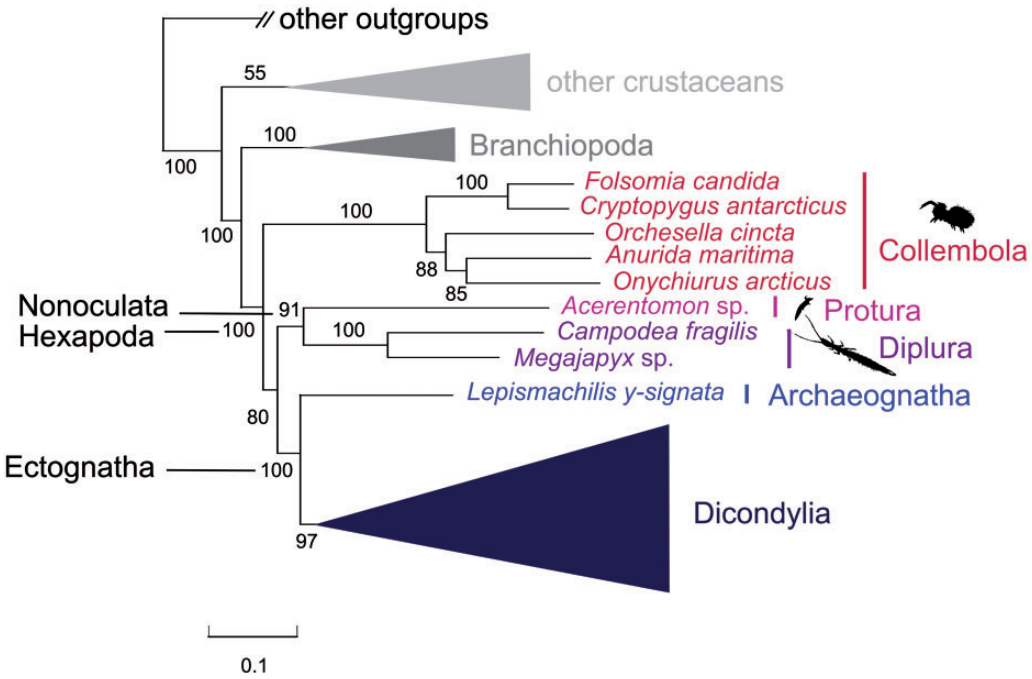

FIG. 1. Simplified phylogenetic tree of the decisive data set *M_Ento*. Best ML tree (RAxML v.7.2.8, PROTCAT, LG + GAMMA), based on 117 OGs that are covered by Protura, Diplura, and Collembola. BS is derived from 1,000 bootstrap replicates. Rogue taxa (supplementary material [section 4], Supplementary Material online) were pruned prior to tree inference. The tree was rooted with *Capitella* sp. For the full tree, see supplementary figure S2, Supplementary Material online.

these 79 genes were removed to artificially create a maximally indecisive data set, showed Entognatha with strong support (table 2) and additionally, diplurans were paraphyletic with respect to Protura (fig. 3*b*; supplementary fig. S3*b*, Supplementary Material online). Both SOS data sets (11 taxa from the supermatrix, which included the collembolan *Orchesella cincta* were removed in the optimization process) did not contain any rogue taxa, that is, taxa that assume incongruent phylogenetic positions in a set of bootstrap trees (Aberer and Stamatakis 2011) (supplementary material [section 3], Supplementary Material online). Tree reconstructions of all data sets strongly supported monophyletic Ectognatha and monophyletic Hexapoda (table 2).

Incongruent Signal among Genes

Based on the *M_Ento* data set, the FcLM approach helped to identify a predominant signal for topology T_1 (Protura + Diplura) – (Collembola + remaining taxa) in 51 genes (12,548 aligned aa positions) (data set *M_Nono*, derived from Nonoculata), a predominant signal for topology T_2

(Protura + Collembola) – (Diplura + remaining taxa) in 35 genes (11,789 aligned aa positions) (data set *M_Elli*, derived from Ellipura), and a predominant signal for topology T_3 (Diplura + Collembola) – (Protura + remaining taxa) in 31 genes (8,546 aligned aa positions) (data set *M_DiCo*) (fig. 4*a* and *b*). Tree inferences from data sets *M_Nono*, *M_Elli*, and *M_DiCo* (rogue taxa pruned, see Materials and Methods section) yielded maximal BS support for Nonoculata, Ellipura, and Diplura + Collembola, respectively (table 2; supplementary fig. S4, Supplementary Material online). However, although tree reconstruction of our data subsets *M_Nono*, *M_Elli*, and *M_DiCo* showed maximal BS support for incongruent topologies among the entognathous insect orders, the results from the FcLM approach indicated that signal for alternative topologies was present in all data sets (fig. 4*a* and *b*; supplementary table S4, Supplementary Material online), which is not reflected by the trees. To identify possible reasons for incongruent signal among genes, we assessed the correlation between functional classes of genes and the different phylogenetic hypotheses that are supported by the data subsets. We found no correlation (supplementary material [section 4], table S5 and fig. S5, Supplementary Material online). Additionally, we tested whether model misspecification can explain the observed incongruence among genes and analyzed the data set *M_Ento* and data subsets *M_Nono*, *M_Elli*, and *M_Dico* using partitioned phylogenetic analyses (Minh et al. 2013) with the best model selected for each gene (partition) separately (supplementary material [section 5], table S6, and figs. S6–S9, Supplementary Material online). With respect to the phylogenetic relationships addressed in our study, resulting topologies did not differ from unpartitioned analyses, and BS only differed to a minor degree (table 2).

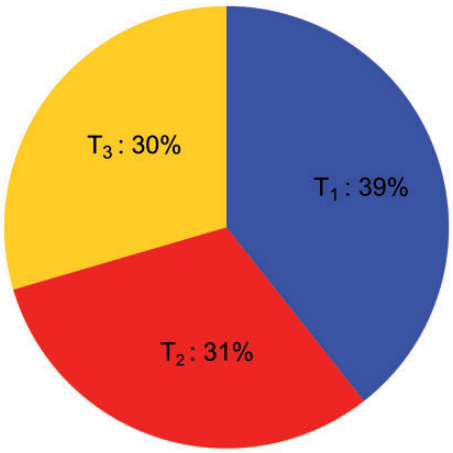

FIG. 2. Results of the FcLM for all OGs in data set *M_Ento*. The chart shows the proportion of quartets (summed up for 117 OGs) that show predominant support for T_1 ([Protura + Diplura] – [Collembola + remaining taxa], Nonoculata hypothesis, blue), T_2 ([Protura + Collembola] – [Diplura + remaining taxa], Ellipura hypothesis, red), and T_3 ([Diplura + Collembola] – [Protura + remaining taxa], yellow), see fig. 5. Quartets mapping in remaining Voronoi cells (gray) and T* (fig. 5) were not considered.

Discussion

The Importance of Data Set Decisiveness

Incongruences in proposed relationships among Protura, Collembola, and Diplura in the studies of Meusemann et al. (2010) and von Reumont et al. (2012), which both supported monophyly of Entognatha, motivated us to look for new approaches to uncover and analyze possible sources of incongruent signal in phylogenomic data sets.

Both SOS data sets in Meusemann et al. (2010) and von Reumont et al. (2012) were compiled with MARE (Meyer and Misof 2010) and were intended to address pancrustacean and arthropod relationships. Both data sets showed only low decisiveness for addressing the relationships of the three entognathous lineages: only 28 out of 128 genes in Meusemann et al. (2010) and 22 out of 316 genes in von Reumont et al. (2012) contained representatives of Protura, Diplura, and Collembola.

Despite low gene data coverage in both studies, the monophyly of Entognatha received high BS. By contrast, our decisive data set for addressing the relationships among these three insect orders lacks clear support for Entognatha

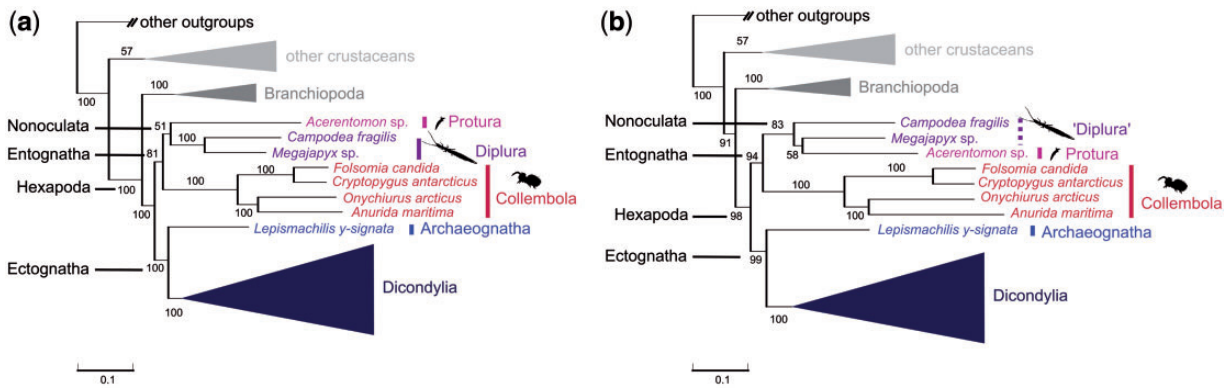

FIG. 3. Simplified phylogenetic trees of data sets SOS (*a*) and SOS_ω (*b*). Best ML tree (RAxML v.7.2.8, PROTCAT, LG + GAMMA) (*a*) based on 253 OGs, 79 of which are covered by Protura, Diplura, and Collembola (SOS) and (*b*) based on 174 OGs, none of which are covered by Protura, Diplura, and Collembola (SOS_ω). BS is derived from 1,000 bootstrap replicates. Trees were rooted with *Capitella* sp. For the full trees, see supplementary figure S3*a* and S3*b*, Supplementary Material online.

Table 2. BS (%) for Selected Clades in Tree Reconstructions with Various Data Sets.

Clade	Data Set					Data Subset of *M_Ento*		
	M_Ento	SOS	SOS_ω	Meusemann et al. (2010)	von Reumont et al. (2012)	*M_Nono*	*M_Elli*	*M_DiCo*
Hexapoda	100 (100)	100	98	100	99	72 (100)	100 (100)	100 (100)
Diplura	100 (100)	100	—[a]	N.A.	N.A.	100 (100)	100 (100)	100 (100)
Collembola	100 (100)	100	100	100	100	100 (100)	100 (100)	100 (100)
(Protura, Diplura)[b]	91 (96)	51	83[a]	100	—	100 (100)	— (—)	— (—)
(Protura, Collembola)[c]	— (—)	—	—	—	98	— (—)	100 (100)	— (—)
(Diplura, Collembola)	— (—)	—	—	—	—	— (—)	— (–)	99 (100)
Entognatha	— (—)	81	94	86	98	— (—)	— (—)	— (—)
((Protura, Diplura), Ectognatha)	80 (96)	—	—	—	—	98 (100)	— (—)	— (—)
((Collembola, Diplura), Ectognatha)	— (—)	—	—	—	—	— (—)	— (—)	60 (83)
(Diplura, Ectognatha)	— (—)	—	—	—	—	— (—)	66 (100)	— (—)
Ectognatha	100 (100)	100	99	100	100	100 (100)	100 (100)	95 (84)

NOTE.—BS was assessed with RAxML from 1,000 bootstrap replicates (see Materials and Method). BS printed in brackets was assessed from partitioned ML analyses of data sets *M_Ento*, and its subsets using the Ufboot algorithm of IQ-TREE with 5,000 bootstrap replicates (supplementary material [section 5], Supplementary Material online). *M_Ento* is the decisive data set in which all OGs are covered by Protura, Diplura, and Collembola; SOS, SOS_ω, and the data sets from Meusemann et al. (2010; data set SOS, ML tree) and von Reumont et al. (2012; data set SOS, ML tree Set 1_{red}) are indecisive to address the relationships of entognathous hexapod orders. *M_Nono*, *M_Elli*, and *M_DiCo* are subsets of *M_Ento* with predominant signal for different topologies and point out conflict of signal among genes.

[a]Diplurans are paraphyletic: *Campodea* + (*Acerentomon*,*Megajapyx*).

[b]Nonoculata hypothesis.

[c]Ellipura hypothesis.

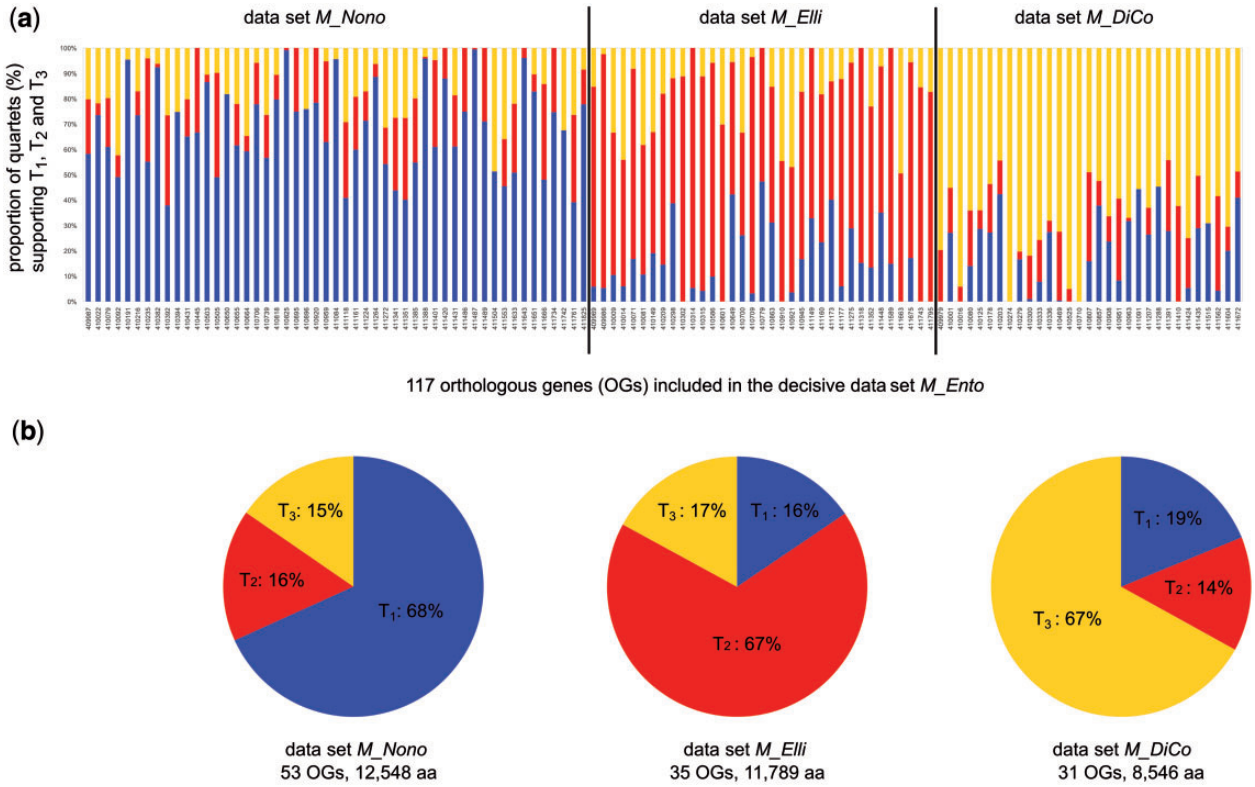

FIG. 4. Detailed results of the FcLM *Mapping* for all OGs included in data set *M_Ento* and data subsets *M_Nono*, *M_Elli*, *M_DiCo*. (*a*) Histogram of FcLM results. Each bar refers to an OG (for OG-IDs, see supplementary table S2, Supplementary Material online). *Y* axis: amount of quartets (in %), that predominantly support T_1 ([Protura + Diplura] – [Collembola + remaining taxa], blue), T_2 ([Protura + Collembola] – [Diplura + remaining taxa], red), and T_3 ([Diplura + Collembola] – [Protura + remaining taxa], yellow), quartets that show ambiguous support are not considered (fig. 5). OGs with predominant support for T_1 are classified into data set *M_Nono* (51 genes, 12,548 aligned aa positions); OGs with predominant support for T_2 are classified into data set *M_Elli*; (35 genes, 11,789 aligned aa positions); OGs with predominant support for T_3 are classified into data set *M_DiCo* (31 genes, 8,546 aligned aa positions). (*b*) FcLM results for data set *M_Nono* (left), *M_Elli* (middle), and *M_DiCo* (right). Each chart shows the proportion of quartets (summed up for the OGs included in the data sets) that show predominant support for T_1, T_2, and T_3 (see above and fig. 5). Quartets that show ambiguous support (fig. 1) are not considered.

(fig. 1). This puzzling result might be explained by the presence of an uneven distribution of missing data. We gained indirect evidence for this hypothesis with the analyses of the worst case data set SOS_{ω}. This data set is maximally indecisive for testing the monophyly of Entognatha, that is, none of the included genes were common to all three entognathous insect groups. Any inferred support for this clade in the SOS_{ω} analysis can be considered an artifact. Remarkably, bootstrapping delivered high, clearly artificial support for monophyletic Entognatha in the SOS_{ω} tree (fig. 3*b*).

We conclude from this indirect evidence that the support for Entognatha in Meusemann et al. (2010), von Reumont et al. (2012) and in our data set indecisive concerning this question (fig. 3*a*) probably results from an artificial signal due to uneven distribution of missing data (Philippe et al. 2011) among Protura, Diplura, and Collembola.

Based on the analyses of the decisive and indecisive data sets, we reject the hypothesis that missing data are unproblematic as long as many characters have been sampled overall (Wiens 2006). Missing data can be misleading as shown by the worst case SOS_{ω} data set analysis, in which relationships received high BS although the data set was maximally indecisive. Therefore, we strongly advocate the exclusive use of decisive data sets in phylogenomic studies.

Incongruent Signal between Genes in a Multigene Data Set

Even decisive data sets can contain incongruent signal (Degnan and Rosenberg 2009; Knowles 2009; Philippe et al. 2011). Using FcLM, we identified groups of genes that support different relationships of Protura, Collembola, and Diplura in the decisive data set *M_Ento* (fig. 4*a* and *b*). Additionally, we assessed conflict within the data with split analyses relying on NeighborNetworks (supplementary material [section 6] and figs. S10–S13, Supplementary Material online). This analysis corroborates the results of FcLM that all analyzed data sets did contain incongruent signal. Additional to the problem of indecisiveness discussed earlier, this incongruent signal among genes may partly be responsible for the contradictory results of Meusemann et al. (2010) and von Reumont et al. (2012). However, incongruent signal among genes is difficult to address and rectify. We analyzed two potential sources of conflict and can conclude that both can be excluded. First, we tested for homoplasy due to analogous selection regimes in functional complexes but found no correlation between predicted gene function and phylogenetic signal (supplementary material [section 4], fig. S5, and table S5, Supplementary Material online). Second, we were able to indirectly exclude model misspecifications as sources of incongruent signal because unpartitioned and partitioned maximum likelihood (ML) analyses yielded topologically congruent results and almost identical BS (table 2; supplementary material [section 5], table S6, and figs. S6–S9, Supplementary Material online). With respect to the FcLM, it may well be that this likelihood mapping approach selects sets of genes with congruent substitution processes. A possible solution, but certainly not a fully satisfying one, would be to increase the number of genes to minimize noise and confounding signal.

Relationships of Protura, Collembola, and Diplura

Monophyly of Entognatha

The monophyly of Entognatha has never been maximally supported and this has not changed in our analyses (table 2). Studies encompassing representatives of Protura, Collembola, and Diplura are limited to only a few analyses (Colgan et al. 1998; Carapelli et al. 2000; Edgecombe et al. 2000; Giribet et al. 2001, 2005). Monophyletic Entognatha were recovered in all recent studies based on nuclear rRNA genes (Gao et al. 2008; Dell'Ampio et al. 2009; von Reumont et al. 2009; Mallatt et al. 2010). However, BS was low, which was either explained by character choice (Dell'Ampio et al. 2009) or the influence of nonstationary processes across taxa (von Reumont et al. 2009). From the morphological point of view, most apomorphies suggesting the monophyly of Entognatha represent reductions (malpighian papillae vs. tubules; reduction to loss of compound eyes). The only exception is the evolution of mouthparts that are concealed in gnathal pouches (Beutel and Gorb 2006). Diplura as closest relatives to Ectognatha is the only relation that contradicts monophyletic Entognatha, and for which morphological evidence has been published (Kukalová-Peck 1991; Koch 1997; Beutel and Gorb 2006; Dallai et al. 2011). In general, morphological support for any clade encompassing more than one of the entognathous lineages Protura, Diplura, and Collembola is weak, largely because character polarization is problematic. This is due to the lack of applicability of characters and/or missing comparative studies in the crustacean groups that are discussed to be most closely related to Hexapoda (Szucsich and Pass 2008).

Ellipura versus Nonoculata

Molecular analyses mostly support Nonoculata (Protura + Diplura) (Giribet et al. 2004; Luan et al. 2005; Kjer et al. 2006; Mallatt and Giribet 2006; Misof et al. 2007; Dell'Ampio et al. 2009; von Reumont et al. 2009; Mallatt et al. 2010; see Dell'Ampio et al. 2011 for a review) while most morphologists merge Protura and Collembola into Ellipura (Börner 1910; Hennig 1953; Kristensen 1981, 1997; Kukalová-Peck 1987; Bitsch and Bitsch 2000, 2004; Beutel and Gorb 2006). Molecular evidence for Ellipura is weak and limited to three mitochondrial single-gene analyses (Shao et al. 1999; Carapelli et al. 2000; Zhang et al. 2001), and morphological support for Nonoculata is nearly missing (Szucsich and Pass 2008). These controversies call for phylogenomic approaches.

The majority of the 117 genes that compose the decisive data set *M_Ento* contain predominant signal for Nonoculata (fig. 4*a*). Also, the FcLM analysis of *M_Ento* (fig. 2) and the phylogenetic tree of *M_Ento* (fig. 1) yielded monophyletic Nonoculata, albeit not being well supported. In summary, Nonoculata is slightly favored over Ellipura in our study, but the question of the phylogenetic relationships of the three entognathous hexapod orders remains unsettled.

Conclusions

Clades may be incorrect, even if receiving high BS support (e.g., monophyly of Entognatha in Meusemann et al. [2010], von Reumont et al. [2012], and in data sets SOS and SOS_{ω} of

this study). This is a trivial conclusion and different reasons are mentioned in the literature (Lehtonen 2011, Simmons and Freudenstein 2011). We show that an uneven distribution of missing data (i.e., the use of indecisive data sets) can lead to strongly supported, yet incorrect, clades. To avoid misleading phylogenetic conclusions from seemingly robust trees based on phylogenomic data sets, we advise 1) using only data sets that are decisive for the phylogenetic question of interest, 2) including an alternative measure of support (Salichos and Rokas 2013); our method of choice was the FcLM approach, and 3) analyzing and documenting the inferred incongruence of signal between genes.

In our decisive data set, we found strong incongruence among genes that is neither correlated with functional classes of genes nor with model misspecifications in unpartitioned analyses. Based upon these notes of caution, we found no signal for the monophyly of Entognatha, and we found no strong signal for Ellipura or Nonoculata despite extending our data set with additional data from key taxa. In other words, the phylogeny and evolution of early hexapods remains enigmatic. Despite this, we show that there are valuable lessons to be learned from the analyses of phylogenomic data of primarily wingless insects, particularly in terms of incongruence among genes and data decisiveness.

Materials and Methods

Taxon Sampling and New Transcriptome Data

Our taxon sampling included 73 species: 46 hexapods, and, as outgroup species, 25 crustaceans, the chelicerate *Ixodes scapularis*, and the polychaete worm *Capitella* sp., both present in the reference set of taxa used for orthology assignment (discussed later). Transcriptome assemblies of 71 species were obtained from the Deep Metazoan Phylogeny database (http://www.deep-phylogeny.org/, last accessed November 4, 2013). We only used species for which more than 1,000 contigs were available (status: December 2011), with two exceptions: the springtail *Orchesella cincta* (Collembola, Entomobryidae, 754 contigs) and the silverfish *Tricholepisma aurea* (Zygentoma, Lepismatidae, 344 contigs), the only publicly available zygentoman transcriptome assembly (supplementary table S1, Supplementary Material online).

We generated new transcriptome data for *Megajapyx* sp. (Diplura, Japygidae) and the firebrat *Thermobia domestica* (Packard 1837) (Zygentoma, Lepismatidae) (table 1). Extraction of RNA, complementary deoxyribonucleic acid (cDNA) library construction, library normalization, and 454 pyrosequencing of ~1,000,000 ESTs per species using the GS-FLX Titanium System, ROCHE were carried out at the Max Planck Institute for Molecular Genetics (MPIMG), Berlin, Germany. Vector clipping, trimming, and soft masking of raw reads and assembly into contigs was conducted at the Center for Integrative Bioinformatics (CIBIV), Vienna, Austria. Steps at the MPIMG and the CIBIV were done as described in von Reumont et al. (2012) and Simon et al. (2012), for details see supplementary material (section 1; Supplementary Material online). Raw sequence reads were deposited at the National Center for Biotechnology Information (NCBI), Sequence Read Archive (accession numbers *Megajapyx* sp.: SRR400673; *T. domestica*: SRR400672). Transcriptome assemblies of *Megajapyx* sp. (accession numbers JT047774–JT094274) and *T. domestic*a (accession numbers T494145–JT533227) were deposited at the Transcriptome Shotgun Assembly (TSA) Database, NCBI Bioproject ID PRJNA81579 and PRJNA81581 (http://www.ncbi.nlm.nih.gov/bioproject, last accessed November 4, 2013). For submission, we excluded contigs shorter than 200 bp, according to the submission guidelines; the full transcriptome assemblies are available at http://zfmk.de/bioinformatics/Full_Transcriptome_Assemblies.zip (last accessed November 4, 2013).

Orthology Assignment

To identify 1:1 OGs in our transcriptome assemblies, we used the Hidden Markov Model based Search for Orthologs using Reciprocity (HaMStR) pipeline (Ebersberger et al. 2009; http://www.deep-phylogeny.org/hamstr/, last accessed November 4, 2013), version 4. As reference set for clusters of OGs, we used a set of 1,886 1:1 OGs (represented by amino acid sequences) based on five reference species (supplementary material [section 2] and table S2, Supplementary Material online). We defined orthology being present if bi-directional best hits were found between our transcript sequences and the reference species *Daphnia pulex*, *Ixodes scapularis*, *Apis mellifera*, and *Capitella* sp.

Alignment, Alignment Masking, and Concatenation

We aligned amino acid sequences using MAFFT L-INS-i (Katoh and Toh 2008) v.6.850 for each gene separately. Afterwards, randomly similar aligned sections were identified with a modified version of ALISCORE (Misof B and Misof K 2009; Kück et al. 2010; Meusemann et al. 2010; for modifications, see Meusemann et al. 2010) using the following options: default sliding window size; *-r*: maximum number of pairwise sequence comparisons; *-e*: special scoring for gappy amino acid data. Identified randomly similar aligned sections were masked with ALICUT v.2.0 (Kück 2009; www.utilities.zfmk.de, last accessed November 4, 2013). Masked alignments were concatenated into supermatrices with FASconCAT v.1.0 (Kück and Meusemann 2010).

Design of Decisive and Indecisive Data Sets

We extracted all genes from the supermatrix that contain at least one representative of each 1) Protura, 2) Diplura, 3) Collembola, and 4) remaining species to generate a decisive data set among entognathous lineages. The resulting data set is called *M_Ento*.

We generated two additional data subsets from the original supermatrix: 1) A so-called selected optimal subset (SOS), generated with MARE v.0.1.2-rc (Meyer and Misof 2010; http://mare.zfmk.de, last accessed November 4, 2013), applying taxon weighting *-t 1.5*. This approach is analogous to Meusemann et al. (2010) and von Reumont et al. (2012). 2) From this SOS data set, we compiled a data set called SOS_{ω} by removing all genes that were covered by all three entognathous lineages to receive a maximally indecisive

"worst case" data set in which each gene contained maximally two entognathous lineages.

Four-Cluster Likelihood Mapping

Additional to tree reconstruction with BS, we applied the FcLM approach using the *M_Ento* data set (Strimmer and von Haeseler 1997). We binned sequenced species into four clusters: 1) Protura (1 species), 2) Diplura (2 species), 3) Collembola (5 species), and 4) remaining species (65 species) (supplementary table S1, Supplementary Material online). Next, we 1) estimated the tree-likeness of each gene, that is the amount of quartets that showed support for one out of the three possible topologies and 2) evaluated which of the three possible topologies was supported by the majority of those quartets (predominant support): T_1 (Protura + Diplura) and (Collembola + remaining taxa), T_2 (Protura + Collembola), and (Diplura + remaining taxa), or T_3 (Diplura + Collembola) and (Protura + remaining taxa) (fig. 5). The competing hypotheses of Meusemann et al. (2010) and von Reumont et al. (2012) are represented by either T_1 (Nonoculata hypothesis) or T_2 (Ellipura hypothesis); the third topology T_3 does not represent a currently debated hypothesis. FcLM was conducted using TREE-PUZZLE v.5.2 (Schmidt et al. 2002; http://www.tree-puzzle.de, last accessed November 4, 2013), applying the BLOSUM62 substitution matrix (Henikoff S and Henikoff JG 1992) as the BLOSUM62 substitution matrix is implemented in the software MARE (Meyer and Misof, 2010; http://mare.zfmk.de, last accessed November 4, 2013).

For each gene in the data set *M_Ento*, we calculated the proportions of quartets that predominantly supported either topology T_1, T_2, or T_3. According to the topology that was supported by the majority of quartets, we classified each gene into one of three groups, supporting Nonoculata, Ellipura, or Diplura + Collembola (fig. 5 and supplementary table S4, Supplementary Material online). Quartets for which the support remained ambiguous (T_{12}, T_{23}, T_{13}, and T*; fig. 5) were not used for classification (see supplementary fig. S14 [Supplementary Material online] for the results with all quartets). All classified genes (supplementary table S4, Supplementary Material online) were subsequently concatenated into three submatrices called *M_Nono* (genes supporting Nonoculata), *M_Elli* (genes supporting Ellipura), and *M_DiCo* (genes supporting Diplura + Collembola).

Phylogenetic Tree Inference

ML tree reconstruction was done from all data sets: *M_Ento*, *M_Nono*, *M_Elli*, and *M_DiCo*, SOS, and SOS_ω (discussed earlier). We estimated evolutionary models for each data set with ModelGenerator v.0.85 (Keane et al. 2006). The

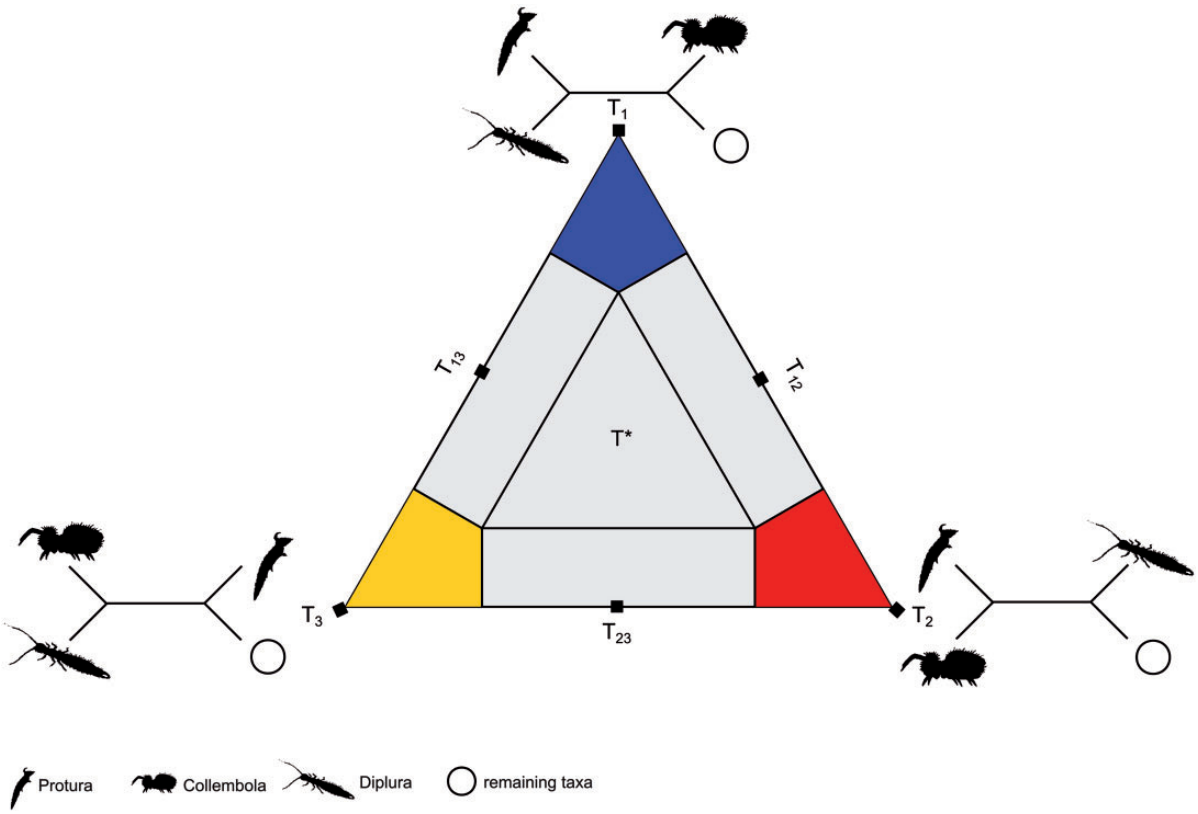

FIG. 5. 2D simplex graph. Voronoi cells are areas, in which quartets show predominant or maximal support for either of the three topologies T_1, T_2, T_3, or in which quartets show ambiguous support T_{12}, T_{13}, T_{23}, and T*. For further explanations, refer to Strimmer and von Haeseler (1997, fig. 3). Voronoi cell corresponding to T_1 (blue): quartets show support for (Protura + Diplura) – (Collembola + remaining taxa); Voronoi cell corresponding to T_2 (red): quartets show support for (Protura + Collembola) – (Diplura + remaining taxa); Voronoi cell corresponding to T_3 (yellow): quartets show support for (Diplura + Collembola) – (Protura + remaining taxa); Voronoi cells corresponding to T_{12}, T_{13}, T_{23} (gray) do not show clear support for T_1, T_2, and T_3; in T* all topologies are equally likely.

best fitting model was selected based upon the Akaike Information Criterion (AIC; Akaike 1974). ML trees were inferred with RAxML (Stamatakis 2006), v.7.2.8-ALPHA, HYBRID (Ott et al. 2007; Pfeiffer and Stamatakis 2010) using the CAT model of rate heterogeneity (Stamatakis 2006) and the LG protein substitution matrix (Le and Gascuel 2008). Final tree searches were conducted under the GAMMA model of rate heterogeneity (Yang 1996). Bootstrap analyses were performed with the rapid algorithm (Stamatakis 2006), which also included subsequent searches for the best scoring ML tree. We obtained BS for each node from 1,000 rapid bootstrap replicates, and checked a posteriori if sufficient bootstrap trees were computed using the bootstopping criteria (Pattengale et al. 2010, default settings). ML analyses were conducted on a Linux cluster at the Cologne High Efficient Operating Platform for Science (CHEOPS), Regionales Rechenzentrum Köln (RRZK), using eight nodes with 12 cores each.

After tree inference, we scrutinized our trees for rogue taxa (Aberer et al. 2013; Aberer and Stamatakis 2011, see supplementary material [section 3], figs. S2 and S4, table S7, Supplementary Material online, for details and results). We removed sequences corresponding to taxa that were identified as rogues from the concatenated alignments and repeated the tree inferences. All trees were edited with Treegraph v.2.0 (Stöver and Müller 2010), and rooted with *Capitella* sp. Data sets are deposited at Dryad: http://doi.org/10.5061/dryad.mk8p7 (last accessed November 4, 2013).

Supplementary Material

Supplementary material (sections 1–6), tables S1–S7, and figures S1–S14 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

K.M. and B.Mi. provided EST data for *Thermobia domestica*, and G.P., E.D.A. and N.U.S. for *Megajapyx* sp. J.B. and M.P. wrote Perl scripts for the analysis pipeline and B.Me. adopted the FcLM approach. Processing and sequence assembly of EST data were performed by I.E. and A.v.H. The study was conceived by B.Mi., K.M., and E.D.A. Orthology assignment and subsequent analyses were conducted by K.M. Rogue taxa analyses were performed by A.J.A. and A.S. Partitioned ML analyses were provided by B.Q.M. The manuscript was written by K.M., N.U.S., R.S.P., E.D.A., and B.Mi with useful comments and revisions from A.J.A., B.Q.M., M.G.W., A.v.H., I.E., A.S., and G.P. All authors read and approved the final manuscript. The authors thank Martin Streinzer for help in collecting *Megajapyx* sp. They acknowledge Michael Kube and Richard Reinhardt (MPIMG, Berlin, Germany) for extraction of RNA, generating cDNA libraries, and ESTs. They thank Sascha Strauss for help with processing and assembling the EST data and John Plant (University of Vienna, Austria) for examining the English. They acknowledge the Cologne High Efficient Operating Platform for Science (CHEOPS, HPC cluster at the RRZK, University of Cologne, Cologne, Germany; available from: http://rrzk.uni-koeln.de/cheops.html) for the opportunity to perform analyses. Finally, the authors thank two anonymous reviewers for helpful comments that considerably improved the manuscript. This work was supported by the Austrian Science Foundation (FWF) grant P 20497-B17 to E.D.A., N.U.S. and G.P., the German Science Foundation (DFG): priority program SPP 1174 "Deep Metazoan Phylogeny" (http://www.deep-phylogeny.org), and by institutional funding of the Heidelberg Institute for Theoretical Studies. A.v.H. and I.E. were funded by the German Science Foundation (DFG) grant HA1628/9. A.v.H. and B.Q.M. were supported by the Austrian Science Foundation (FWF) grant I760. K.M. and B.M. were funded by the German Science Foundation (DFG) grant MI 649/6.

References

Aberer AJ, Komprass D, Stamatakis A. 2013. Pruning Rogue Taxa improves phylogenetic accuracy: an efficient algorithm and web service. *Syst Biol.* 62(1):162–166.

Aberer AJ, Stamatakis A. 2011. A simple and accurate method for rogue taxon identification. *IEEE BIBM* 2011:118–122.

Akaike H. 1974. A new look at the statistical model identification. *IEEE Trans Aut Control.* 19:716–723.

Bitsch C, Bitsch J. 2000. The phylogenetic interrelationships of the higher taxa of apterygote hexapods. *Zool Scr.* 29:131–156.

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.

Beutel RG, Gorb SN. 2006. A revised interpretation of attachment structures in Hexapoda with special emphasis on Mantophasmatodea. *Arthropod Syst Phylogeny.* 64:3–25.

Börner C. 1910. Die phylogenetische Bedeutung der Protura. *Sdr Biolog Centralbl.* 30:633–641.

Carapelli A, Frati F, Nardi F, Dallai R, Simon C. 2000. Molecular phylogeny of the apterygote insects based on nuclear and mitochondrial genes. *Pedobiologia* 44:361–373.

Colgan DJ, McLauchlan A, Wilson GDF, Livingston S, Macaranas J, Edgecombe D, Cassis G, Gray MR. 1998. Histone H3 and U2 snRNA DNA sequences and arthropod molecular evolution. *Aust J Zool.* 46:419–437.

Dallai R, Mercati D, Carapelli A, Nardi F, Machida R, Sekiya K, Frati F. 2011. Sperm accessory microtubules suggest the placement of Diplura as the sister-group of Insecta s.s. *Arthropod Struct Dev.* 40: 77–92.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference, and the multispecies coalescent. *Trends Ecol Evol.* 24: 332–340.

Dell'Ampio E, Szucsich NU, Carapelli A, Frati F, Steiner G, Steinacher A, Pass G. 2009. Testing for misleading effects in the phylogenetic reconstruction of ancient lineages of hexapods: influence of character dependence and character choice in analyses of 28S rRNA sequences. *Zool Scr.* 38:155–170.

Dell'Ampio E, Szucsich NU, Pass G. 2011. Protura and molecular phylogenetics: status quo of a young love. *Soil Organisms* 83:347–358.

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, Wilson GDF, Colgan DJ, Gray MR, Cassis G. 2000. Arthropod cladistics: combined analysis of Histone H3 and U2 snRNA sequences and morphology. *Cladistics* 16:155–203.

Gao Y, Bu Y, Luan Y. 2008. Phylogenetic relationships of basal hexapods reconstructed from nearly complete 18S and 28S rRNA gene sequences. *Zool Sci.* 25:1139–1145.

Giribet G, Edgecombe GD, Carpenter JM, D'Haese CA, Wheeler WC. 2004. Is Ellipura monophyletic? A combined analysis of basal hexapod relationships with emphasis on the origin of insects. *Org Div Evol.* 4:319–340.

Giribet G, Edgecombe GD, Wheeler WC. 2001. Arthropod phylogeny based on eight molecular loci and morphology. *Nature* 413:157–161.
Giribet G, Richter S, Edgecombe GD, Wheeler WC. 2005. The position of crustaceans within Arthropoda—evidence from nine molecular loci and morphology. In: Koenemann S, Jenner RA, editors. Crustacea and arthropod relationships. Crustacean issues 16: Festschrift for F. R. Schram. Boca Raton (FL): Taylor & Francis. p. 307–352.
Giribet G, Wheeler WC. 2001. Some unusual small-subunit ribosomal DNA sequences of metazoans. *Am Mus Novit.* 3337:1–14.
Grimaldi DA. 2010. 400 million years on six legs: on the origin and early evolution of Hexapoda. *Arthropod Struct Dev.* 39:191–203.
Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A.* 89:10915–10919.
Hennig W. 1953. Kritische Bemerkungen zum phylogenetischen System der Insekten. *Beitr Entomol Sonderheft.* 3:1–85.
Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform.* 9:286–298.
Keane TM, Creevey CJ, Pentony MM, Naughton TJ, McInerney JO. 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.
Kjer KM, Carle FL, Litman J, Ware J. 2006. A molecular phylogeny of Hexapoda. *Arthropod Syst Phylogeny.* 64:3–44.
Knowles LL. 2009. Estimating species trees: methods of phylogenetic analysis when there is incongruence across genes. *Syst Biol.* 58: 463–467.
Koch M. 1997. Monophyly and phylogenetic position of the Diplura (Hexapoda). *Pedobiologia* 41:9–12.
Kristensen NP. 1981. Phylogeny of insect orders. *Annu Rev Entomol.* 26: 135–157.
Kristensen NP. 1997. The ground plan and basal diversification of the hexapods. In: Fortey RA, Thomas RH, editors. Arthropod relationships, systematic association. Special volume series 55. London: Chapman & Hall. p. 281–293.
Kück P. 2009. ALICUT: a Perlscript which cuts ALISCORE identified RSS. Version 2.0 ed. Bonn (Germany): Department of Bioinformatics, Zoologisches Forschungs Museum A. Koenig (ZFMK). [cited 2013 Oct 30]. Available from: http://www.zfmk.utilities.de.
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, von Reumont BM, Waegele JW, 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. 1987. New Carboniferous Diplura, Monura and Thysanura, the hexapod ground plan, and the role of thoracic side lobes in the origin of wings (Insecta). *Can J Zool.* 65:2327–2345.
Kukalová-Peck J. 1991. Fossil history and the evolution of hexapod structures. In: Naumann ID, editor. Insects of Australia: a textbook for students and research workers. Melbourne (Australia): CSIRO, Melbourne University Press. p. 141–179.
Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. *Mol Biol Evol.* 25:1307–1320.
Lehtonen S. 2011. Can sensitivity analysis help to detect long-branch attraction? *Mol Phylogenet Evol.* 61:899–903.
Luan Y, Mallatt JM, Xie R, Yang Y, Yin W. 2005. The phylogenetic positions of three basal-hexapod groups (Protura, Diplura, and Collembola) based on ribosomal RNA gene sequences. *Mol Biol Evol.* 22:1579–1592.
Machida R. 2006. Evidence from embryology for reconstructing the relationships of hexapod basal clades. *Arthropod Syst Phylogeny.* 64:95–104.
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 JM, Craig CW, Yoder MJ. 2010. Nearly complete rRNA genes assembled from across the metazoan animals: effects of more taxa, a structure-based alignment, and paired-sites evolutionary models on phylogeny reconstruction. *Mol Phylogenet Evol.* 55:1–17.
Meusemann K, von Reumont BM, Simon S, et al. (16 co-authors). 2010. A phylogenomic approach to resolve the arthropod tree of life. *Mol Biol Evol.* 27:2451–2464.
Meyer B, Misof B. 2010. MARE: matrix reduction—a tool to select optimized data subsets from supermatrices for phylogenetic inference. Bonn (Germany): Zentrum für Molekulare Biodiversitätsforschung (zmb) am ZFMK. [cited 2013 Oct 30]. Available from: http://mare.zfmk.de (current version).
Minh BQ, Nguyen MAT, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. *Mol Biol Evol.* 30:1188–1195.
Misof B, Niehuis O, Bischoff I, Rickert A, Erpenbeck D, Staniczek A. 2007. Towards an 18S phylogeny of hexapods: accounting for group-specific character covariance in optimized mixed nucleotide/doublet models. *Zoology* 110:409–429.
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.
Ott M, Zola J, Stamatakis A, Aluru S. 2007. Large-scale maximum likelihood-based phylogenetic analysis on the IBM BlueGene/L. Proceedings of the 2007 ACM/IEEE conference on Supercomputing. IEEE/ACM Supercomputing conference 2007 (SC2007); November 2007. Reno (NV): ACM.
Pattengale ND, Alipour M, Bininda-Emonds OR, Moret BM, Stamatakis A. 2010. How many bootstrap replicates are necessary? *J Comput Biol.* 17:337–354.
Pfeiffer W, Stamatakis A. 2010. Hybrid MPI/Pthreads parallelization of the RAxML phylogenetics code. Paper presented at HICOMB workshop, held in conjunction with IPDPS 2010; April 2010; Atlanta, GA.
Philippe H, Brinkmann H, Lavrov DV, Littlewood DT, Manuel M, Worheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol.* 9: e1000602.
Regier JC, Shultz JW, Ganley AR, et al. (11 co-authors). 2008. Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. *Syst Biol.* 57:920–938.
Regier J, Shultz J, 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.
Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497: 327–331.
Sanderson MJ, McMahon MM, Steel M. 2010. Phylogenomics with incomplete taxon coverage: the limits to inference. *BMC Evol Biol.* 10:155.
Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18:502–504.
Shao HG, Zhang YP, Xie RD, Yin WY. 1999. Mitochondria cytochrome *b* sequences variation of Protura and molecular systematics of Apterygota. *Chin Sci Bull.* 44:2031–2036.
Simmons MP, Freudenstein JV. 2011. Spurious 99% bootstrap and jackknife support for unsupported clades. *Mol Phylogenet Evol.* 61: 177–191.
Simon S, Narechania A, DeSalle R, Hadrys H. 2012. Insect phylogenomics: exploring the source of incongruence using new transcriptomic data. *Genome Biol Evol.* 4:1295–1309.
Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
Steel M, Sanderson MJ. 2010. Characterizing phylogenetically decisive taxon coverage. *Appl Math Lett.* 23:82–86.
Strimmer K, von Haeseler A. 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proc Natl Acad Sci U S A.* 94:6815–6819.
Stöver BC, Müller KF. 2010. TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinformatics* 11:7.

Szucsich NU, Pass G. 2008. Incongruent phylogenetic hypotheses and character conflicts in morphology: the root and early branches of the hexapodan tree. *Mitt Dtsch Ges Allg Angew Ent.* 16:415–430.

Trautwein MD, Wiegmann BM, Beutel R, Kjer KM, Yeates DK. 2012. Advances in insect phylogeny at the dawn of the postgenomic era. *Annu Rev Entomol.* 57:449–468.

von Reumont BM, Meusemann K, Szucsich N, et al. (14 co-authors). 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.

von Reumont MB, Jenner RA, Wills MA, et al. (13 co-authors). 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.

Wiens JJ. 2006. Missing data and the design of phylogenetic analyses. *J Biomed Inform.* 39:34–42.

Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol.* 11(9):367–372.

Zhang Y, Zhang Y, Luan Y, Chen Y, Yin W. 2001. Phylogeny of higher taxa of Hexapoda according to 12sRNA sequences. *Chin Sci Bull.* 46: 840–842.