Syst. Biol. 71(5):1023–1031, 2022

https://doi.org/10.1093/sysbio/syac024
Advance Access publication March 15, 2022

Phylogenomics of Elongate-Bodied Springtails Reveals Independent Transitions from Aboveground to Belowground Habitats in Deep Time

DAOYUAN YU[1,2,3], YINHUAN DING[4], ERIK TIHELKA[5], CHENYANG CAI[5,6,7], FENG HU[1,2,3], MANQIANG LIU[1,2,3] AND FENG ZHANG[4,*]

[1]*Soil Ecology Laboratory, College of Resources and Environmental Sciences, Nanjing Agricultural University, 210095 Nanjing, China;* [2]*Jiangsu Collaborative Innovation Center for Solid Organic Waste Resource Utilization, 210095 Nanjing, China;* [3]*Jiangsu Key Laboratory for Solid Organic Waste Utilization, 210095 Nanjing, China;* [4]*Department of Entomology, College of Plant Protection, Nanjing Agricultural University, 210095 Nanjing, China;* [5]*School of Earth Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol BS8 1TQ, UK;* [6]*State Key Laboratory of Palaeobiology and Stratigraphy, Nanjing Institute of Geology and Palaeontology, 210008 Nanjing, China; and* [7]*Center for Excellence in Life and Paleoenvironment, Chinese Academy of Sciences, 210008 Nanjing, China Manqiang Liu and Feng Zhang contributed equally to this article.*
**Correspondence to be sent to: Department of Entomology, College of Plant Protection, Nanjing Agricultural University, 210095 Nanjing, China; E-mail: fzhang@njau.edu.cn.*

Received 24 February 2021; accepted 11 March 2022
Associate Editor: Vanessa González

Abstract.—Soil has become a major hotspot of biodiversity studies, yet the pattern and timing of the evolution of soil organisms are poorly known because of the scarcity of paleontological data. To overcome this limitation, we conducted a genome-based macroevolutionary study of an ancient, diversified, and widespread lineage of soil fauna, the elongate-bodied springtails (class Collembola, order Entomobryomorpha). To build the first robust backbone phylogeny of this previously refractory group, we sampled representatives of major higher taxa (6 out of 8 families, 11 out of 16 subfamilies) of the order with an emphasis on the most problematic superfamily Tomoceroidea, applied whole-genome sequencing methods, and compared the performance of different combinations of data sets (universal single-copy orthologs [USCO] vs. ultraconserved elements]) and modeling schemes. The fossil-calibrated timetree was used to reconstruct the evolution of body size, sensory organs, and pigmentation to establish a time frame of the ecomorphological divergences. The resultant trees based on different analyses were congruent in most nodes. Several discordant nodes were carefully evaluated by considering method fitness, morphological information, and topology test. The evaluation favored the well-resolved topology from analyses using USCO amino acid matrices and complex site-heterogeneous models (CAT+GTR and LG+PMSF (C60)). The preferred topology supports the monophyletic superfamily Tomoceroidea as an early-diverging lineage and a sister relationship between Entomobryoidea and Isotomoidea. The family Tomoceridae was recovered as monophyletic, whereas Oncopoduridae was recovered as paraphyletic, with *Harlomillsia* as a sister to Tomoceridae and hence deserving a separate family status as Harlomillsiidae Yu and Zhang **fam. n.** Ancestral Entomobryomorpha were reconstructed as surface-living, supporting independent origins of soil-living groups across the Paleozoic–Mesozoic, and highlighting the ancient evolutionary interaction between aboveground and belowground fauna. [Collembola; phylogenomics; soil-living adaptation; whole-genome sequencing.]

Soil habitats harbor a higher abundance and richness of organisms than most other terrestrial habitats on Earth (Decaëns et al. 2006; Coleman and Wall 2015; Orgiazzi et al. 2016). To a large extent, such tremendous diversity is contributed by a great number of specialized soil animals. Most prominently, animal traits related to perception, locomotion, and protection from ultraviolet radiation can become drastically modified to adapt to belowground environments, as exemplified in a number of invertebrate taxa (e.g., Rusek 2007; Kurth and Kier 2015). Reconstructing the pattern and timing of the differentiation between aboveground and belowground fauna is essential for a comprehensive understanding of the succession of terrestrial biodiversity, however, because soil animals generally have poor preservation potential and a sparse fossil record, only a few studies have attempted to addressed the history of their adaptive evolution (e.g., Schaefer and Caruso 2019; Yu et al. 2021).

Wingless hexapods of the class Collembola Lubbock, 1870, the springtails, are among the most abundant soil animals (Hopkin 1997). Up to 100,000 springtails can occur in only 1 m^2 of temperate forest topsoil (Beutel et al. 2014). They are also the first unequivocal hexapods to appear in the fossil record; the earliest in situ preserved terrestrial ecosystem, the 407 million years old Early Devonian Rhynie Chert in Scotland, already preserves exceptionally modern-looking springtails (Hirst and Maulik 1926; Edwards et al. 2017). Extant springtails have adapted to diverse habitats and can be generally classified as either atmobiotic (inhabiting vegetation or other aboveground habitats), epedaphic (inhabiting soil surface, upper litter, rocks, etc.), hemiedaphic (inhabit lower litter and humus), or euedaphic (true soil-dwellers; e.g., Stebaeva 1970; Rusek 2007). Notably, many euedaphic springtails possess extreme adaptations for life in soil such as blindness, depigmentation, minute size (<1 mm), and shortened appendages that facilitate easier movement between soil particles (Rusek 2007). By contrast, their atmobiotic and epedaphic relatives typically possess the opposite character states, whereas hemiedaphic taxa are usually intermediate.

Within Collembola, the elongate-bodied springtails (order Entomobryomorpha Börner, 1913) are the most widespread, species-rich, and ecologically diverse

group. Three main branches of this order, that is, superfamilies Entomobryoidea Womersley, 1934, Isotomoidea Szeptycki, 1979 and Tomoceroidea Szeptycki, 1979 (sensu Soto-Adames et al. 2008) are all cosmopolitan and inhabit almost all terrestrial niches imaginable, showing prominent ecological divergences between lineages. For example, within Tomoceroidea, the family Tomoceridae Schäffer, 1896 contains mostly large surface-living species (body length reaching ca. 10 mm in *Novacerus* Salmon, 1942), whereas Oncopoduridae Carl and Lebedinsky, 1905 contains minute (mostly around 0.5 mm) euedaphic genus *Oncopodura* Carl and Lebedinsky, 1905 and hemiedaphic *Harlomillsia* Bonet, 1944. Similarly, most members of Entomobryoidea are surface dwellers, whereas its sister Isotomoidea is more edaphic. Besides, Entomobryomorpha is probably the oldest extant lineage of Collembola, with the earliest unequivocal fossil record in the Early Permian (Riek 1976).

Because of their diversified niches and long evolutionary history, Entomobryomorpha is an ideal model taxon for studying the evolutionary patterns of soil biodiversity. However, despite exhaustive morphological and molecular studies over the past two decades, several pressing problems still exist in the entomobryomorph tree of life (e.g., Xiong et al. 2008; Yu et al. 2016; Leo et al. 2019; Sun, Yu et al. 2020). Although Entomobryoidea and Isotomoidea constantly clustered together, the position of Tomoceroidea has not been determined; the monophyly of Tomoceroidea, Tomoceridae, and Oncopoduridae has been questioned (Szeptycki 1977; Yu et al. 2016; Cucini et al. 2021). The lack of a well-resolved entomobryomorph phylogeny severely hampers attempts at elucidating the pattern and timing of their diversification.

To build a robust backbone phylogeny of Entomobryomorpha and to better understand their ecomorphological diversification, we used a phylogenomic approach based on whole-genome sequencing. We sampled representatives of major suprageneric taxa covering a wide range of ecomorphological types within Entomobryomorpha, with a special emphasis on the problematic Tomoceroidea. We experimented with various data sets and modeling approaches to overcome common sources of phylogenomic error. Divergence time estimation and ancestral character state reconstruction (ACSR) were performed to constrain the timing of entomobryomorph ecomorphological divergences. Specifically, we focused on testing the monophyly of Tomoceroidea, Tomoceridae, and Oncopoduridae, and finding the pattern of divergence between aboveground and belowground groups.

Materials and Methods

Detailed materials and methods, including sampling, sequencing, and analytical methods, are available in Supplementary Appendix S1 available on Zenodo at https://doi.org/10.5281/zenodo.5910398. Twenty ingroup species were sampled to cover major subordinate taxa of three main branches of Entomobryomorpha, that is, all subfamilies of Isotomidae and Tomoceridae, six major subfamilies of Entomobryoidea, and two distinct lineages of Oncopoduridae. The sampled species represented a wide range of trait states related to above/belowground habitats (e.g., eyed vs. eyeless, dark-pigmented vs. pale) within each main branch. One Poduromorpha Börner, 1913 and one Symphypleona Börner, 1901 species were used as outgroups based on previous phylogenomic analyses (e.g., Sun, Ding et al. 2020). Genome assemblies of 7 species were downloaded from NCBI, whereas those of 15 species were newly sequenced on the Illumina Novaseq 6000 platform. Species names, taxonomic ranks, raw sequencing data, and assembly accessions are provided in Supplementary Appendix S2: Table S1 available on Zenodo.

Genomes were assembled by using the rapid pipeline PLWS v1.0.5 (https://github.com/xtmtd/PLWS/, Zhang, Ding, Zhu et al. 2019). Contaminants were detected by using HS-BLASTN (Chen et al. 2015) and BLAST+ (i.e., blastn) v2.7.1 (Camacho et al. 2009) against the NCBI nt and UniVec databases. The basic statistics for genome assemblies are provided in Supplementary Appendix S2: Table S1 available on Zenodo).

For matrices generation, universal single-copy orthologs (USCOs) and ultraconserved elements (UCEs) were extracted from genomes with BUSCO v3.0.2 (Waterhouse et al. 2018) against a collembolan reference gene set ($n = 1997$) and with PHYLUCE v1.6.6 (Faircloth 2016) against a probe set customized for Collembola, respectively (Sun, Ding et al. 2020). The basic statistics for the captured USCOs and UCEs are provided in Supplementary Appendix S2: Table S1 available on Zenodo. USCOs were translated into amino acid sequences. Both USCOs and UCEs were aligned with MAFFT v7.450 (Katoh and Standley 2013), trimmed with BMGE v1.12 (Criscuolo and Gribaldo 2010), concatenated with FASConCAT-g v1.04 (Kück and Longo 2014), and filtered SRH (stationary, reversible, and homogeneous, Naser-Khdour et al. 2019) model violations with IQ-TREE v2.0-rc1 (Minh, Schmidt et al. 2020). Primarily, we generated three USCO matrices (USCO75, USCO90, and USCO100) of 75%, 90%, and 100% completeness, and three UCE matrices (UCE50, UCE75, and UCE90) of 50%, 75%, and 90% completeness, which represents the lowest ratio of taxa for all partitions. Furthermore, to overcome gene tree topological incongruence (Salichos and Rokas 2013), we inferred individual gene trees with IQ-TREE and selected USCOs of average UFBoot2 (Hoang et al. 2018) values greater than 75 and UCEs of average UFBoot2 values greater than 70 to generate the new matrices, that is, USCO75_abs75, USCO90_abs75, USCO100_abs75, UCE50_abs70, UCE75_abs70, and UCE90_abs70. A total of 12 matrices (Supplementary Appendix S3 available on Dryad at https://doi.org/10.5061/dryad.cjsxksn76.) were generated for subsequent analyses, their properties are summarized in Supplementary Appendix S2: Table S2 available on Zenodo.

We conducted phylogenetic inference using a diverse set of analytical methods to account for biological and methodological sources of systematic error (Kumar et al. 2012; Young and Gillung 2020). For both the USCO amino acid and UCE nucleotide matrices, phylogenetic trees were inferred with partitioned maximum likelihood (ML, with partitioning and models selected in MODELFINDER, Kalyaanamoorthy et al. 2017) and heterotachy model (General Heterogeneous evolution On a Single Topology, GHOST, Crotty et al. 2020) based methods implemented in IQ-TREE, and with multispecies coalescent model (MSCM) based method implemented in ASTRAL-III v5.6.1 (Zhang et al. 2018). We quantified genealogical concordance with the gene concordance factor (gCF) and the site concordance factor (sCF) given the reference tree and gene trees (Minh, Hahn et al. 2020) using IQ-TREE. We also applied site-heterogeneous models to mitigate possible long branch attraction (LBA) artifacts. For USCO matrices, the posterior mean site frequency (PMSF, Wang, Minh et al. 2019) model was performed in IQ-TREE; Bayesian inference (only for data set USCO90_abs75 due to the computational burden) was performed in PhyloBayes MPI v1.8b (Lartillot et al. 2013). The resultant three alternative topological hypotheses for deep relationships (H1, H2, H3, see Results section) were tested with the matrix USCO90_abs75 in IQ-TREE. Because introgression may cause gene tree discordance (e.g., Vanderpool et al. 2020), we tested introgression by calculating the statistic Δ (Supplementary Appendix S1 available on Zenodo).

We estimated divergence times using MCMCTree in PAML v4.9j (Yang 2007) based on three matrices (USCO75, USCO90_abs75, and USCO100_abs75) to account for potential influences of data size levels on the estimation. Loci were merged into larger partitions based on schemes from partitioned ML reconstructions. Preferred topology estimated from PMSF, ASTRAL, and Phylobayes (see Discussion section) was selected as the input tree. Divergence time analyses applied approximate likelihood calculation and ML estimation of branch lengths to reduce computational burden. Hessian matrices were calculated by using the LG substitution model and the independent rates clock model. Six nodes (the root and five internodes) were selected for calibration (Supplementary Appendix S2: Table S3 available on Zenodo). Details for parameter setting, calibration points, and MCMC runs were provided in Supplementary Appendix S1 available on Zenodo. Prior and posterior times were compared. The quality of MCMC runs was assessed based on convergence and infinite-sites plots following the package manual.

We performed Bayesian character state reconstructions for five ecomorphological traits related to collembolan spatial niche: body length, number of eyes, degree of pigmentation, presence/absence of bothriotricha, and presence/absence of sticky chaetae on legs (Supplementary Appendix S2: Table S4 available on Zenodo, Rusek 2007; Salmon et al. 2014; Yu et al. 2017). The analyses were conducted in BayesTraits V3.0.2 (Pagel et al. 2006) on the PhyloBayes consensus topology and 1000 posterior trees.

Codes for the aforementioned analyses are available in Supplementary Appendix S3 available on Dryad.

Results

Phylogenetic Inference

The 12 matrices for phylogenetic analyses included 91–1698 USCO and 192–697 UCE loci across 47,887–683,871 amino acid and 103,751–373,885 nucleotide sites, respectively (Supplementary Appendix S2: Table S2 available on Zenodo). Trees based on different matrices and inference models were congruent in most nodes, but the position of *Oncopodura* was not stable, resulting in three topological hypotheses of deep entomobryomorph relationships (Supplementary Appendix S2: Table S5 and Appendix S5 available on Zenodo). Under the first hypothesis (H1), *Oncopodura* was the sister group of the remaining entomobryomorphs. Under the second hypothesis (H2), *Oncopodura* was sister to a clade comprising *Harlomillsia* + Tomoceridae. The third hypothesis (H3) supported an *Oncopodura* + *Harlomillsia* clade sister to Tomoceridae. Most reconstructions with five USCO matrices (USCO100 excluded) under the partitioning and GHOST model generated topology H1. Analyses under the site-heterogeneous models (CAT+GTR and LG+PMSF(C60)) and USCO matrices lent absolute support to H2 (Bayesian Posterior Probabilities [BPP] = 1 and SH-aLRT/UFBoot2 $\geq$ 99; Fig. 1). All ASTRAL analyses based on both USCO and UCE matrices under the multispecies coalescent model recovered H2 but with many nodes poorly supported. Matrix USCO100 under the partitioning/GHOST model and three UCE matrices excluding loci of low phylogenetic signals (UCE50_abs70, UCE75_abs70, and UCE90_abs70) under the partitioning model also supported H2, but usually with weak supports. The remaining UCE analyses not mentioned above supported H3. Topology tests rejected hypotheses H1 and H3 with strong confidence (Supplementary Appendix S2: Table S6 available on Zenodo). In addition, UCE data sets consistently recovered *Lepidocyrtus fimetarius* Gisin, 1964 as the earliest-diverging member of Entomobryoidea, disagreeing with the USCO results and other sources of evidence (e.g., Zhang, Bellini et al. 2019). All tree files are available in Supplementary Appendix S3 available on Dryad.

Gene tree conflicts quantified by concordance factors showed very strong genealogical incongruence (gCF < 30, sCF < 40) for the deep Tomoceroidea and Entomobryoidea nodes. Matrices excluding loci with low phylogenetic signal (labeled as "_abs") had slightly higher gCF/sCF and bootstrap values (Supplementary Appendix S5 available on Zenodo). No evidence of introgression was detected at all internal branches (Supplementary Appendix S1 available on Zenodo).

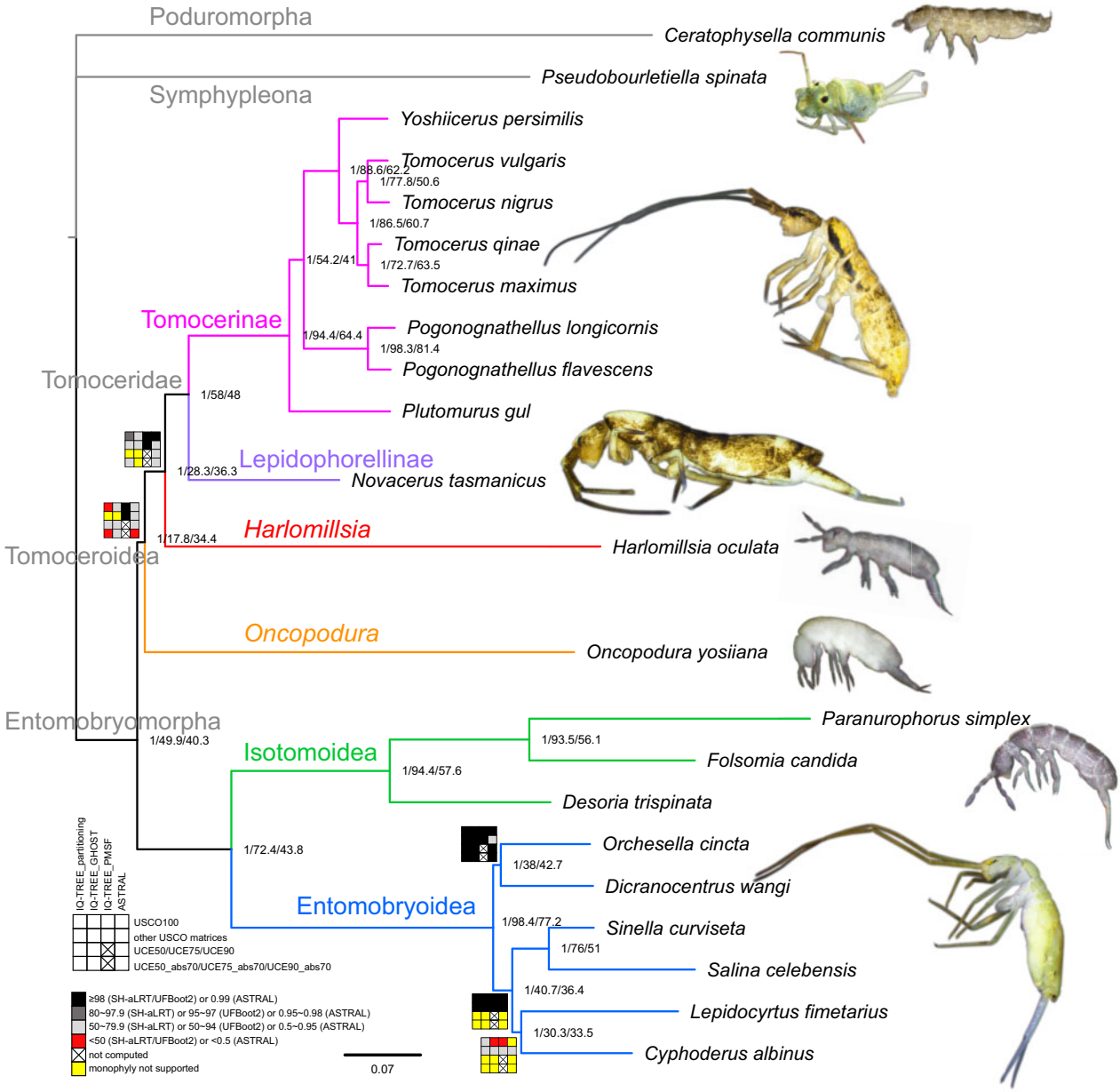
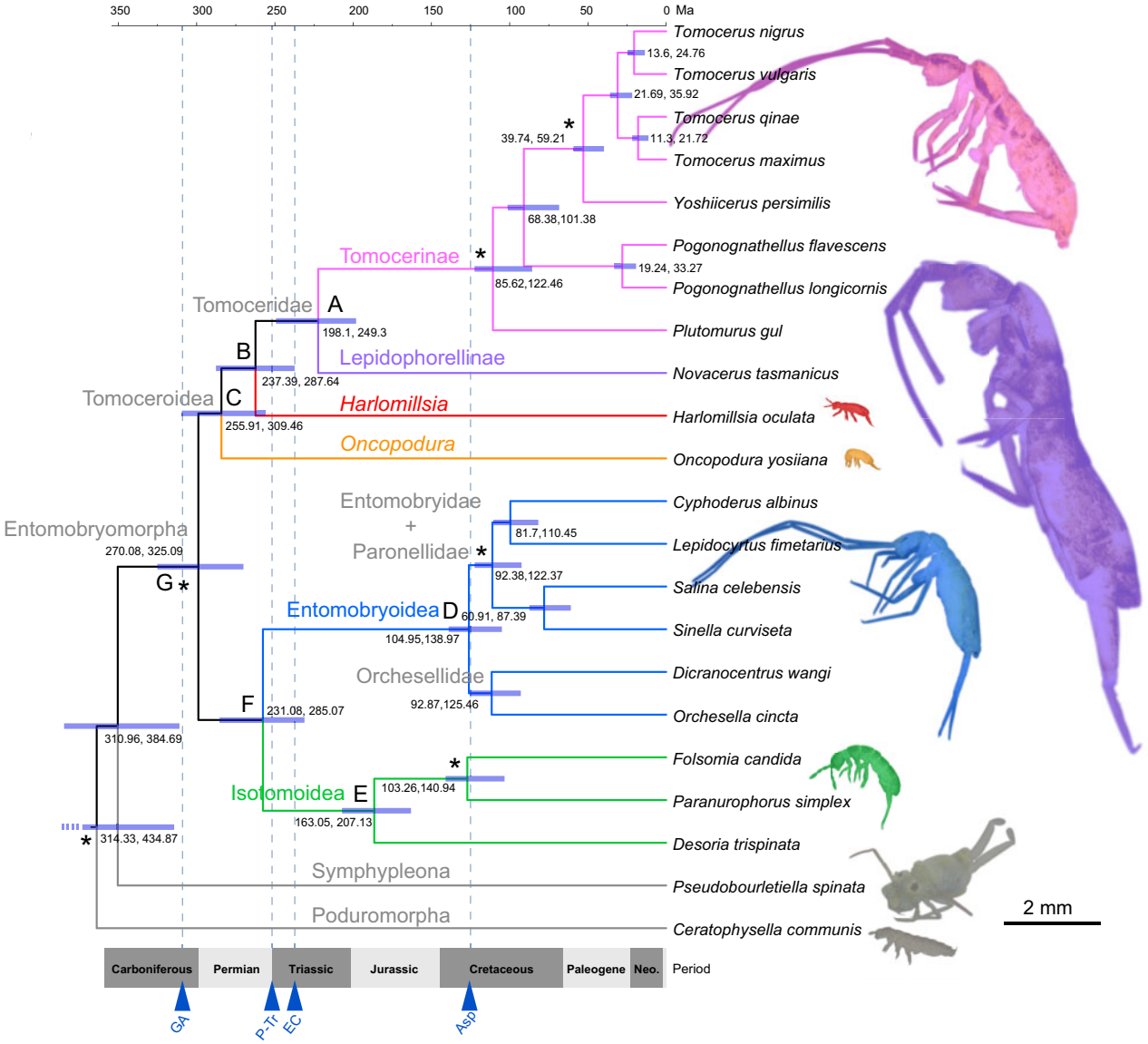


FIGURE 1. Phylogeny of Entomobryomorpha inferred from matrix USCO90_abs75 using the site-heterogeneous Bayesian GTR+CAT model implemented in PhyloBayes. Node labels show BPP/gCFs/sCFs. Node supports from other analyses are also indicated by the colored squares. Squares are not shown for the nodes congruent with the PhyloBayes tree. Only the lowest category is shown when different matrices or different supporting measures of the same matrix produced conflict results.

Divergence Time Estimation

Based on the MCMCTree results (Supplementary Table S7 available on Zenodo), the convergence plots (Supplementary Appendix S4: Fig. S2a available on Zenodo) fitted perfectly a straight line ($R^2 = 1$), indicating good convergence between parallel runs; the infinite-sites plots (Supplementary Fig. S2b,c available on Zenodo) also tended to a straight line ($R^2 = 0.88$–0.99), suggesting sufficient data sizes for the estimation; the posterior times were close to but not entirely congruent with priors (Supplementary Fig. S2d,e available on Zenodo), suggesting the estimation was informed by both priors and molecular data.

Integrating the MCMCTree results (Fig. 2), Entomobryomorpha and Tomoceroidea originated during the Carboniferous–Permian (270.08–325.09 and 255.91–309.46 Ma, respectively, 95% highest posterior density (HPD) confidence interval [CI]). The divergences between *Harlomillsia* and Tomoceridae, and between Isotomoidea and Entomobryoidea occurred during the Permian–Triassic (237.39–287.64 and 231.08–285.07



	A	B	C	D	E	F	G
Body length (mm)	3.91	3.39	3.11	2.55	1.88	2.58	3.03
Bothriotricha	Present	Present	Present	Present	Absent	Equivocal	Present
Tenent hair	Equivocal	Equivocal	Equivocal	Equivocal	Equivocal	Equivocal	Equivocal
Eye number	6–8	6–8	6–8	6–8	Equivocal	Equivocal	6–8
Pigmentation	Well pigmented	Well pigmented	Well pigmented	Well pigmented	Equivocal	Equivocal	Well pigmented

FIGURE 2. Entomobryomorpha divergence time estimation performed by using MCMCTree on the PhyloBayes topology. Node numbers and node bars represent 95% CIs of the estimated divergence times integrated from all MCMC runs. Asterisks mark nodes constrained with the fossils. The arrows and dashed lines mark major palaeoenvironmental change events. GA, the beginning of global aridification in the Pennsylvanian; P–Tr, Permian–Triassic extinction; EC, the end of "coal gap"; Asp, the first unequivocal fossil of angiosperms. Profiles on the right show true scales of representative species. Bottom table shows the results of ACSR for five selected ecomorphological traits on nodes A–G.

Ma, respectively). The crown Tomoceridae and Isotomoidea originated during the Triassic–Jurassic (198.1–249.3 and 163.05–207.13 Ma, respectively); the origins of Tomocerinae, Entomobryoidea, Orchesellidae, and the clade containing Entomobryidae and Paronellidae occurred during the Cretaceous (85.62–122.46, 104.95–138.97, 92.87–125.46, and 92.38–122.37 Ma, respectively).

Evolution of Ecomorphological Traits

According to the reconstruction (Fig. 2), the most recent common ancestor (MRCA) of elongate-bodied springtails was moderately large (ca. 3.03 mm) and well-pigmented, with bothriotricha, and a set of functional eyes. On the deep internodes of Tomoceroidea, most trait states were the same as for the MRCA

of Entomobryomorpha, whereas the estimated body lengths showed a continuous increase, being 3.11 mm for Tomoceroidea, 3.39 mm for Tomoceridae + *Harlomillsia*, and 3.91 mm for Tomoceridae. The reduction of body length, pigmentation, and eye number appeared independently in *Oncopodura* and *Harlomillsia*. The absence of bothriotricha is an apomorphy of *Harlomillsia*. In the other two major branches of the order, the estimated body lengths decreased, being 2.58 mm for Isotomoidea + Entomobryoidea, 2.55 mm for Entomobryoidea, and 1.88 mm for Isotomoidea. The other trait states of the MRCA of Entomobryoidea were reconstructed as the same as those of the MRCA of Entomobryomorpha. The states of most traits of the MRCAs of Isotomoidea and Entomobryoidea + Isotomoidea remained equivocal.

Discussion

Which Phylogenetic Hypothesis Is Best Supported?

To account for common sources of error in phylogenomic reconstruction (Kumar et al. 2012; Brown and Thomson 2016), we employed a variety of analytical strategies, combining different data sets and modeling schemes. Results from these analyses are generally congruent in recovering the monophyly of most higher taxa and their relationships within Entomobryomorpha, but are still incongruent in several deep nodes, most crucially the deep relationships in Tomoceroidea.

As indicated by the low values of concordance factors (gCF and sCF) for the problematic nodes, systematic error induced by gene tree conflict is a major source of incongruence between our analyses. Notably, gene tree conflicts are often more severe for deep and short internodes on the tree (Salichos and Rokas 2013), which is the case for the deep nodes in Tomoceroidea. Moreover, the long branch lengths of the recalcitrant taxa in our trees indicate a possible LBA artifact. Faced with LBA, methods using MSCM (e.g., by ASTRAL) or site-heterogeneous models (e.g., LG+PMSF(C60), CAT-GTR), which incorporate a higher degree of biological realism than the conventional concatenation-based and site-homogeneous models (e.g., Rannala and Yang 2003; Lartillot and Philippe 2004; Zhang et al. 2018; Wang, Minh et al. 2019), are expected to produce more plausible reconstructions (e.g., Feuda et al. 2017; Marlétaz et al. 2019; Williams et al. 2020). Besides, selecting phylogenetically informative genes can also reduce the incongruence between gene trees and improve the robustness of inference (Salichos and Rokas 2013). In the present study, most of the analyses accounting for these sources of incongruence (i.e., all analyses using LG+PMSF(C60) and CAT-GTR models, ASTRAL, UCE_abs70) support H2. In addition, new morphological evidence (Supplementary Appendix S6 available on Zenodo) and topology tests also show strong preference for H2, suggesting a paraphyletic Oncopoduridae and a sister relationship between *Harlomillsia* and Tomoceridae.

In comparison, most analyses based on partitioning and GHOST model accounting for protein-wise heterotachy favor H1 and H3, both rejected by the topology test. H1 (*Oncopodura* is sister to the other Entomobryomorpha) has never been proposed before (e.g., Bonet 1943; Szeptycki 1977, 1979; D'Haese 2003; Sun, Yu et al. 2020) and is also rejected by our morphological examination; H3 (monophyletic Oncopoduridae) represents the long questioned traditional classification (Szeptycki 1977) and is rejected by the lack of morphological apomorphy (see next section). Therefore, our results are in line with a recent evaluation showing that site-wise heterogeneity is usually a more important source of bias than protein-wise heterotachy to be modeled in phylogenomic inference (Wang, Susko et al. 2019).

Interestingly, H1 and H3 are recovered by analyses based on the USCO (amino acid) and UCE (nucleotide) matrices, respectively, reflecting discordance between the two data capture strategies. As frequently discussed in phylogenomic studies, nucleotide-based analyses often generate inaccurate or less robust deep relationships due to compositional biases or model violation (e.g., Rota-Stabelli et al. 2013; Cox et al. 2014; Shin et al. 2018; but see Gillung et al. 2018; Baker et al. 2021), for which reason we used only amino acid sequences for the USCO data sets as a common practice to maintain accuracy and reduce computational burden. The use of site-heterogeneous models and exclusion of uninformative loci have partially mitigated this discordance via generating topology H2, however, no matter what approaches have been applied, all UCE-based analyses consistently place *Lepidocyrtus* incorrectly at the base of Entomobryoidea, suggesting the appropriate use of UCE markers in such ancient lineages should be further explored.

Phylogeny and Classification of Tomoceroidea

The best-supported topology suggests a monophyletic Tomoceroidea as the sister group of Isotomoidea + Entomobryoidea. This result supports the idea that Tomoceroidea is an early-diverging branch of Entomobryomorpha, as being previously hypothesized based on analyses using 18S/28S ribosomal RNA gene markers and mitochondrial genomes (Yu et al. 2016; Sun, Yu et al. 2020). The position of Tomoceroidea is supported by characters shared with the other orders of Collembola (Poduromorpha, Symphypleona, and Neelipleona Massoud, 1971), notably the presence of compound postantennal organs (*Novacerus, Oncopodura*, and *Harlomillsia*), a filamentous extension on the unguiculus (in some species of *Novacerus* and *Pogonognathellus* Börner, 1908), subsegmented dens, and elongate mucro.

With all families and subfamilies of Tomoceroidea included, the preferred analyses have also generated the first robust family-level tree of Tomoceroidea. Significantly, Oncopoduridae has been recovered as paraphyletic, with *Harlomillsia* as sister to Tomoceridae. This relationship is strongly supported

by morphological evidence. First, there is no reliable apomorphy between *Harlomillsia* and *Oncopodura*. Current diagnostic characters of Oncopoduridae, that is, the absence of cephalic bothriotricha, the presence of scales, undeveloped trochanteral organs, subsegmented furca, and elongate mucro, can also be treated as plesiomorphies shared with Tomoceridae. Second, *Harlomillsia* differs from *Oncopodura* in several key characters (Supplementary Appendix S6 available on Zenodo), including the substantial differences in chaetotaxy that usually lead to familial level divisions (Szeptycki 1977). Third, despite its minute body size, *Harlomillsia* resembles *Novacerus*, the early-diverging genus of Tomoceridae, in the presence of additional labral chaetae, cephalic macrochaetae, multilobed postantennal organs, and dental spines on both the inner and outer edges of the dens (Supplementary Appendix S6 available on Zenodo). Therefore, integrating the phylogenetic inference and morphological evidence, here, we propose a new family Harlomillsiidae Yu and Zhang **fam. n.** (http://zoobank.org/urn:lsid:zoobank.org:act:D112B4E1-80A8-4826-8EAC-121F5A86E6E6), named after the type genus *Harlomillsia* (described in Supplementary Appendix S6 available on Zenodo).

The monophyly of Tomoceridae recovered by us has been supported by previous phylogenetic analyses (Yu et al. 2016; Sun, Yu et al. 2020). Drastic morphological differences between its two constituent subfamilies (Tomocerinae and Lepidophorellinae) have historically led to a proposal to split the family in two separate families (Absolon 1942), which is supported by the deep genetic divergence revealed by our analyses and deserves a future assessment using a more specific taxon sampling.

Ecological Divergences of Elongate-Bodied Springtails in Deep Time

Results of the ACSR show that the MRCAs of Entomobryomorpha, Tomoceroidea, and Entomobryoidea were most likely surface-living. Although two rare families Actaletidae (littoral, atmobiotic) and Coenaletidae (strictly commensal with hermit crabs) were not included in the analyses, their functional traits suggest both of them are most probably derived from epigeic ancestors (Soto-Adames 1988; Palacios-Vargas et al. 2000). In other words, the adaptations to life in soil evolved independently across entomobryomorph lineages, resulting in the convergent reduction of eye number, body size, and pigmentation. According to the fossil-calibrated evolutionary time frame, the first transition from aboveground to belowground habitat (the divergence between *Oncopodura* and the stem Tomoceroidea) occurred during the Carboniferous–Permian, followed by a second transition (the divergence between *Harlomillsia* and Tomoceridae) during the Permian–Triassic and multiple later transitions (within Entomobryoidea and Isotomoidea) during the Mesozoic. Therefore, the results suggest that the stratification structure of terrestrial ecosystem, consisting of aboveground–interface–belowground subsystems, had already formed by the Late Paleozoic, and the independent transitions may reflect multiple ecological successions in the geological history. Our finding partially resembles that about oribatid mites, also inferring a Palaeozoic establishment of aboveground–belowground ecological interactions (Schaefer and Caruso 2019). However, unlike the primarily surface-living elongate-bodied springtails, the crown oribatids had an edaphic origin, while most epigeic taxa only emerged later in the Mesozoic (Maraun et al. 2009; Schäffer et al. 2020), suggesting that the two major soil microarthropod lineages have different ecology and evolutionary trajectory, and may have responded differently to environmental changes. Interestingly, the estimated times (CIs) of the ecological divergences of Entomobryomorpha roughly overlap several key palaeoenvironmental changes, notably the Late Pennsylvanian–Permian global aridification (DiMichele and Aronson 1992; Gulbranson et al. 2015), the end-Permian extinction and subsequent "Coal Gap" (Erwin 1994; Retallack et al. 1996), and the Cretaceous diversification of angiosperms (Herendeen et al. 2017). Exploring the correlation between the ecological divergences and the palaeoenvironmental events (e.g., by implementing diversification rate analyses with denser taxon sampling and more fossil evidences) will be an interesting topic for future phylogenomic studies of springtails and other soil animals.

Conclusion

Overall, our phylogenomic analyses based on different combinations of data matrices and models produced generally congruent trees of elongate-bodied springtails. However, several deep nodes show incongruence between analyses, highlighting the necessity of data set and model selection for phylogenetics using big data. Integrating method fitness and morphological evidence, the topology recovered by using amino acid matrices of USCOs and MSCM/site-heterogeneous models is finally preferred. The novel, well-resolved phylogeny supports Tomoceroidea as an early-diverging lineage of Entomobryomorpha and clarifying the problematic relationship between *Oncopodura* and *Harlomillsia*. Ancestral entomobryomorphs were reconstructed as surface-living, whereas soil-living groups evolved several times independently across the Palaeozoic–Mesozoic, suggesting ancient evolutionary interaction between aboveground and belowground ecosystems since ~300 Ma. More recent ecological divergences at shallow nodes may be revealed by future studies focusing on lower taxa using denser taxon sampling. Applying phylogenomics to other major soil invertebrate lineages (e.g., mites, earthworms, nematodes) can be used to test whether a general evolutionary pattern or multiple lineage-specific patterns should be introduced to elucidate the formation of soil biodiversity.

Supplementary Material

Data available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.cjsxksn76 and Zenodo: https://doi.org/10.5281/zenodo.5910398

Acknowledgments

We are grateful to our colleagues who kindly provided specimens for this study, including but not limited to Dr. Xiaodong Yang (Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences), Dr. Donghui Wu (Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences), Dr. Penelope Greenslade (Federation University, Australia), Dr. Louis Deharveng and Anne Bedos (Muséum national d'Histoire naturelle, France), and Dr. Arne Fjellberg (Tjøme, Norway).

Funding

This work was supported by the National Natural Science Foundation of China [grant numbers 41971063, 31970434, 31772491] and the National Science and Technology Fundamental Resources Investigation Program of China [grant number 2018FY100300].

Data Availability Statement

The genome data underlying this article are available in the GenBank Nucleotide Database and can be accessed with the accession number listed in Supplementary Appendix S2: Table S1 available on Zenodo. All data matrices, codes, and resultant tree files are available in Supplementary Appendix S3 available on Dryad. All appendices have been uploaded to Dryad and Zenodo. The PhyloBayes consensus tree is also available on TreeBASE (Study Accession URL: http://purl.org/phylo/treebase/phylows/study/TB2:S28823).

References

Absolon K., Ksenemann M. 1942. Troglopedetini. Vergleichende studie über eine altertümliche höhlenbewohnende Kollembolengruppe aus den dinarischen Karstgebieten. Brünn: Barvič & Novotný.

Baker C.M., Buckman-Young R.S., Costa C.S., Giribet G. 2021. Phylogenomic analysis of velvet worms (Onychophora) uncovers an evolutionary radiation in the Neotropics. Mol. Biol. Evol. 38:5391–5404.

Beutel R.G., Friedrich F., Ge S.Q., Yang X.K. 2014. Insect morphology and phylogeny: a textbook for students of entomology. Berlin: De Gruyter.

Bonet F. 1943. Sobre la clasificación de los Oncopoduridae (Collembola), con descriptión de especies nuevas. An. Esc. Nacion. Cienc. Biol. 3:127–153.

Brown J.M., Thomson R.C. 2016. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. Syst. Biol. 66:517–530.

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden T.L. 2009. BLAST+: architecture and applications. BMC Bioinform. 10:421.

Chen Y., Ye W., Zhang Y., Xu Y. 2015. High speed BLASTN: an accelerated MegaBLAST search tool. Nucleic Acids Res. 43:7762–7768.

Coleman D.C., Wall D.H. 2015. Soil fauna: occurrence, biodiversity, and roles in ecosystem function. In: Paul E.A., editor. Soil microbiology, ecology and biochemistry. 4th ed. Cambridge: Elsevier Academic Press. p. 111–149.

Cox C.J., Li B., Foster P.G., Embley T.M., Civáň P. 2014. Conflicting phylogenies for early land plants are caused by composition biases among synonymous substitutions. Syst. Biol. 63:272–279.

Criscuolo A., Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol. Biol. 10:210.

Crotty S.M., Minh B.Q., Bean N.G., Holland B.R., Tuke J., Jermiin L.S., von Haeseler A. 2020. GHOST: recovering historical signal from heterotachously-evolved sequence alignments. Syst. Biol. 69:249–264.

Cucini C., Fanciulli P.P., Frati F., Convey P., Nardi F., Carapelli A. 2021. Re-evaluating the internal phylogenetic relationships of Collembola by means of mitogenome data. Genes. 12:44.

Decaëns T., Jimenez J.J., Gioia C., Measey G.J., Lavelle P. 2006. The values of soil animals for conservation biology. Eur. J. Soil Biol. 42:S23–S38.

D'Haese C.A. 2003. Morphological appraisal of Collembola phylogeny with special emphasis on Poduromorpha and a test of the aquatic origin hypothesis. Zool. Scr. 32:563–586.

DiMichele W.A., Aronson R.B. 1992. The Pennsylvanian-Permian vegetational transition: a terrestrial analogue to the onshore-offshore hypothesis. Evolution. 46:807–824.

Edwards D., Kenrick P., Dolan L. 2017. History and contemporary significance of the Rhynie cherts—our earliest preserved terrestrial ecosystem. Phil. Trans. R. Soc. B 373:20160489.

Erwin D.H. 1994. The Permo–Triassic extinction. Nature. 367:231–236.

Faircloth B.C. 2016. PHYLUCE is a software package for the analysis of conserved genomic loci. Bioinformatics. 32:786–788.

Feuda R., Dohrmann M., Pett W., Philippe H., Rota-Stabelli O., Lartillot N., Wörheide G., Pisani D. 2017. Improved modeling of compositional heterogeneity supports sponges as sister to all other animals. Curr. Biol. 27:3864–3870.

Gillung J.P., Winterton S.L., Bayless K.M., Khouri Z., Borowiec M.L., Yeates D., Kimsey L. S., Misof B., Shin S., Zhou X., Mayer C., Petersen M., Wiegmann B.M. 2018. Anchored phylogenomics unravels the evolution of spider flies (Diptera, Acroceridae) and reveals discordance between nucleotides and amino acids. Mol. Phylogenet. Evol. 128:233–245.

Gulbranson E.L., Montañez I.P., Tabor N.J., Limarino C.O. 2015. Late Pennsylvanian aridification on the southwestern margin of Gondwana (Paganzo Basin, NW Argentina): a regional expression of a global climate perturbation. Palaeogeogr. Palaeoclimatol. Palaeoecol. 417:220–235.

Herendeen P.S., Friis E.M., Pedersen K.R., Crane P.R. 2017. Palaeobotanical redux: revisiting the age of the angiosperms. Nat. Plants. 3:17015.

Hirst S., Maulik S. 1926. On some arthropod remains from the Rhynie Chert (Old Red Sandstone). Geol. Mag. 63:69–71.

Hoang D.T., Chernomor O., von Haeseler A., Minh B.Q., Vinh L.S. 2018. UFBoot2: improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 35:518–522.

Hopkin S.P. 1997. Biology of the springtails (Insecta: Collembola). New York: Oxford University Press.

Kalyaanamoorthy S., Minh B.Q., Wong T.K.F., von Haeseler A., Jermiin L.S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods. 14:587–589.

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

Kück P., Longo G.C. 2014. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. Front. Zool. 11:81.

Kumar S., Filipski A.J., Battistuzzi F.U., Pond S.L.K., Tamura K. 2012. Statistics and truth in phylogenomics. Mol. Biol. Evol. 29:457–472.

Kurth J.A., Kier W.M. 2015. Differences in scaling and morphology between lumbricid earthworm ecotypes. J. Exp. Biol. 218:2970–2978.

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., 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.
Leo C., Carapelli A., Cicconardi F., Frati F., Nardi F. 2019. Mitochondrial genome diversity in Collembola: phylogeny, dating and gene order. Diversity. 11:169.
Maraun M., Erdmann G., Schulz G., Norton R.A., Scheu S., Domes K. 2009. Multiple convergent evolution of arboreal life in oribatid mites indicates the primacy of ecology. P. R. Soc. B - Biol. Sci. 276:3219–3227.
Marlétaz F., Peijnenburg K.T.C.A., Goto T., Satoh N., Rokhsar D.S. 2019. A new spiralian phylogeny places the enigmatic arrow worms among gnathiferans. Curr. Biol. 29:312–318.
Minh B.Q., Hahn M.W., Lanfear R. 2020. New methods to calculate concordance factors for phylogenomic datasets. Mol. Biol. Evol. 37:2727–2733.
Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D., Woodhams M.D., von Haeseler A., Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37:1530–1534.
Naser-Khdour S., Minh B.Q., Zhang W., Stone E.A., Lanfear R. 2019. The prevalence and impact of model violations in phylogenetic analysis. Genome Biol. Evol. 11:3341–3352.
Orgiazzi A., Bardgett R.D., Barrios E., Behan-Pelletier V., Briones M.J.I., Chotte J.-L., De Deyn G.B., Eggleton P., Fierer N., Fraser T., Hedlund K., Jeffery S., Johnson N.C., Jones A., Kandeler E., Kaneko N., Lavelle P., Lemanceau P., Miko L., Montanarella L., Moreira F.M.S., Ramirez K.S., Scheu S., Singh B.K., Six J., van der Putten W.H., Wall D.H. 2016. Global soil biodiversity atlas. Luxembourg: Publication Office of the European Union.
Pagel M., Meade A. 2006. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. Am. Nat. 167:808–825.
Palacios-Vargas J.G., Cutz L.Q., Maldonado C. 2000. Redescription of the male of *Coenaletes caribaeus* (Collembola: Coenaletidae) associated with hermit crabs (Decapoda: Coenobitidae). Ann. Entomol. Soc. Am. 93:194–197.
Rannala B., Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645–1656.
Retallack G.J., Veevers J.J., Morante R. 1996. Global coal gap between Permian–Triassic extinctions and middle Triassic recovery of peat forming plants. Geol. Soc. Am. Bull. 108:195–207.
Riek E.F. 1976. An entomobryid collembolan (Hexapoda: Collembola) from the Lower Permian of Southern Africa. Paleontol. Afr. 19:141–143.
Rota-Stabelli O., Lartillot N., Philippe H., Pisani D. 2013. Serine codon-usage bias in deep phylogenomics: Pancrustacean relationships as a case study. Syst. Biol. 62:121–133.
Rusek J. 2007. A new classification of Collembola and Protura life forms. In: Tajovský K., Schlaghamerský J., Pižl V., editors. Contributions to soil zoology in central Europe II. České Budějovice: ISB BC ASCR. p. 109–115.
Salichos L., Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature. 497:327–331.
Salmon S., Ponge J.F., Gachet S., Deharveng L., Lefebvre N., Delabrosse F. 2014. Linking species, traits and habitat characteristics of Collembola at European scale. Soil Biol. Biochem. 75:73–85.
Schaefer I., Caruso T. 2019. Oribatid mites show that soil food web complexity and close aboveground-belowground linkages emerged in the early Paleozoic. Commun. Biol. 2:387.
Schäffer S., Koblmüller S., Krisper G. 2020. Revisiting the evolution of arboreal life in oribatid mites. Diversity. 12:255.
Shin S., Clarke D.J., Lemmon A.R., Lemmon E.M., Aitken A.L., Haddad S., Farrell B.D., Marvaldi A.E., Oberprieler R.G., McKenna D.D. 2018. Phylogenomic data yield new and robust insights into the phylogeny and evolution of weevils. Mol. Biol. Evol. 35:823–836.
Soto-Adames F.N. 1988. Revisión de la familia Actaletidae Borner, 1902 (Insects: Collembola). Caribb. J. Sci. 24:161–196.
Soto-Adames F.N., Barra J.-A., Christiansen K., Jordana R. 2008. Suprageneric classification of Collembola Entomobryomorpha. Ann. Entomol. Soc. Am. 101: 501–513.
Stebaeva S.K. 1970. Life forms of springtails (Collembola). Zool. Zh. 49:1437–1455.
Sun X., Ding Y., Orr M.C., Zhang F. 2020. Streamlining universal single-copy orthologue and ultraconserved element design: a case study in Collembola. Mol. Ecol. Res. 20:706–717.
Sun X., Yu D., Xie Z., Dong J., Ding Y., Yao H., Greenslade P. 2020. Phylomitogenomic analyses on collembolan higher taxa with enhanced taxon sampling and discussion on method selection. PLoS One. 15:e0230827.
Szeptycki A. 1977. Morpho-systematic studies on Collembola. V. The body chaetotaxy of the genera Oncopodura Carl et Lebedinsky, 1905 and Harlomillsia Bonet, 1944 (Oncopoduridae). Rev. Ecol. Biol. Sol. 14:199–209.
Szeptycki A. 1979. Morpho-systematic studies on Collembola. IV. Chaetotaxy of the Entomobryidae and its phylogenetical significance. Kraków: Polska Akademia Nauk.
Vanderpool D., Minh B.Q., Lanfear R., Hughes D., Murali S., Harris R.A., Raveendran M., Muzny D.M., Hibbins M.S., Williamson R.J., Gibbs R.A., Worley K.C., Rogers J., Hahn M.W. 2020. Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression. PLoS Biol. 18:e3000954.
Wang H.C., Minh B.Q., Susko E., Roger A.J. 2019. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. Syst. Biol. 67:216–235.
Wang H.C., Susko E., Roger A.J. 2019. The relative importance of modeling site pattern heterogeneity versus partition-wise heterotachy in phylogenomic inference. Syst. Biol. 68:1003–1019.
Waterhouse R.M., Seppey M., Simão F.A., Manni M., Ioannidis P., Klioutchnikov G., Kriventseva E.V., Zdobnov E.M. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol. Biol. Evol. 35:543–548.
Williams T.A., Cox C.J., Foster P.G., Szöllösi G.J., Embley T.M. 2020. Phylogenomics provides robust support for a two-domains tree of life. Nat. Ecol. Evol. 4:138–147.
Xiong Y., Gao Y., Yin W.Y., Luan Y.X. 2008. Molecular phylogeny of Collembola inferred from ribosomal RNA genes. Mol. Phylogenet. Evol. 49:728–735.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.
Young A.D., Gillung J.P. 2020. Phylogenomics—principles, opportunities and pitfalls of big-phylogenetics. Syst. Entomol. 45:225–247.
Yu D., Deharveng L., Lukić M., Wei Y., Hu F., Liu M. 2021. Molecular phylogeny and trait evolution in an ancient terrestrial arthropod lineage: systematic revision and implications for ecological divergence (Collembola, Tomocerinae). Mol. Phylogenet. Evol. 154:106995.
Yu D., Ding Y., Ma Y. 2017. Revision of *Tomocerus similis* Chen & Ma, with discussion of the *kinoshitai* complex and the distal tibiotarsal chaetae in Tomocerinae (Collembola, Tomoceridae). Zootaxa. 4268:395–410.
Yu D., Zhang F., Stevens M.I., Yan Q., Liu M., Hu F. 2016. New insight into the systematics of Tomoceridae (Hexapoda, Collembola) by integrating molecular and morphological evidence. Zool. Scr. 45:286–299.
Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinform. 19:153.
Zhang F., Bellini B.C., Soto-Adames F.N. 2019. New insights into the systematics of Entomobryoidea (Collembola: Entomobryomorpha): first instar chaetotaxy, homology and classification. Zool. Syst. 44:249–278.
Zhang F., Ding Y., Zhu C., Zhou X., Orr M.C., Scheu S., Luan Y. 2019. Phylogenomics from low-coverage whole-genome sequencing. Methods Ecol. Evol. 10:507–517.