IECD 2022

MDPI

Proceedings

A Genome-Scale Phylogeny of the Superfamily Entomobryoidea (Entomobryomorpha: Collembola) †

Nerivania Nunes Godeiro [1,2,*], Yinhuan Ding [2], Bruno Cavalcante Bellini [3], Nikolas Gioia Cipola [4], Sopark Jantarit [5] and Feng Zhang [2]

[1] Shanghai Natural History Museum, Shanghai Science & Technology Museum, Shanghai 200041, China
[2] Department of Entomology, College of Plant Protection, Nanjing Agricultural University, Nanjing 210095, China
[3] Biosciences Center, Department of Botany and Zoology, Federal University of Rio Grande do Norte, Natal, Rio Grande do Norte, Brazil
[4] Laboratório de Sistemática e Ecologia de Invertebrados do Solo, Instituto Nacional de Pesquisas da Amazônia—INPA, CPEN. Avenida André Araújo, 2936, Aleixo, Manaus, AM, Brazil
[5] Excellence Center for Biodiversity of Peninsular Thailand, Faculty of Science, Prince of Songkla University, Hat Yai, Songkhla 90110, Thailand
* Correspondence: nerivania@gmail.com
† Presented at the 2nd International Electronic Conference on Diversity (IECD 2022)—New Insights into the Biodiversity of Plants, Animals and Microbes, 1–15 March 2022; Available online: https://iecd2022.sciforum.net/.

Abstract: The superfamily Entomobryoidea has been the focus of molecular studies in recent years due to an intriguing divergence between morphological and genetic data. Recent studies based on mitogenomes have converged on a result that suggests the non-monophyly of Paronellidae and Entomobryidae. Here, we reanalyzed some of the raw published data and newly sequenced species of Entomobryoidea to create phylogenetic independent matrices containing single-copy nuclear genes (USCOs) and ultraconserved elements (UCEs). Our results corroborated with previous phylogenies and we recovered the Orchesellidae as an independent basal family; the Entomobryinae remained the most puzzling taxon gathering scaled and unscaled lineages of both traditional Entomobryidae and Paronellidae; and the Seirinae were reaffirmed as the sister-group of the Lepidocyrtinae. The sampled representatives of Paronellinae s. str. were recovered as the sister group of Seirinae+Lepidocyrtinae, supporting their reduction on the dorsal macrochaetotaxy and trunk sensillar pattern occurred independently from the Lepidocyrtinae.

Keywords: Whole-genome assembly; low-coverage data; soil fauna; Paronellidae; Entomobryidae

Academic Editor: Matthieu Chauva

Published: 14 March 2022

1. Introduction

Molecular studies related to the Entomobryoidea superfamily are still incipient, and no study in the literature analyzed its internal relationships with high-resolution molecular data, while just a few studies used mitogenomes or instead have focused on only a few species [1,2]. Currently, the superfamily is divided into three families: Entomobryidae, Paronellidae, and Orchesellidae [3], but during the past years many changes were made in its internal organization. Morphological phylogenies based on diagnostic characters which do not hold phylogenetic signal are the main reasons for the systematic errors [4]. While the relationships within some entomobryid clades are robust, like within the Seirinae, the resolution of the paraphyly between the families Paronellidae and Entomobryidae remains unsolved and have not been assessed genome-wide to determine the specific points of discordance suggested in previous studies.

The use of universal single copy orthologs (USCOs) and ultra-conserved elements (UCEs) have proven great efficiency for phylogenetic study [5,6], but some lineages

require probe sets specifically created to the analyzed lineage in order to increase the locus recovery. High-quality and nearly complete reference genome assemblies were fundamental for the initial steps of this study, and they were used to create a dataset of universal molecular markers for Entomobryoidea following the steps available on-line in the pipeline created by [7]. The full paper containing the details about the dataset elaboration is under preparation, but here we already tested its maker capture through De Novo genome assembling from low-coverage genome data (10x), employing a series of computationally efficient bioinformatic tools. We extracted thousands of genes used to create the first genome-scale phylogeny of the superfamily Entomobryoidea.

2. Materials and Methods

2.1. Taxon Sampling

From the current nine subfamilies of Entomobryoidea, seven were sampled here: Heteromurinae (four genera, four spp.), Orchesellinae (two genera, three spp.), Entomobryinae (seven genera, seven spp.), Lepidocyrtinae (four genera, six spp.), Seirinae (three genera, five spp.), Paronellinae s. str. (three genera, three spp.), and Salininae (four genera, five spp.). Twenty-one genomic data were newly generated for this study, while other data were previously published or are in the publication process.

2.2. DNA Extraction and Library Preparation

For all newly sequenced specimens, genomic DNA was extracted and amplified from one specimen/sample. All procedures followed manufacturer's protocols. BGI high-throughput sequencer MGIseq2000 platform was used for sequencing paired-end reads with 150 bp length. Approximately 10 G of low-coverage data were produced for each sample.

2.3. USCO and UCE Extraction

The initial input files were all the raw sequencing reads from 35 species. The pipeline PLWS v1.0.6 [7] was used to assembly clean sequenced reads. Single-copy genes were predicted using AUGUSTUS v3.3.2 [8] and then were assessed with lineage-specific BUSCO v3.1.0 [9] to assign them to one of the orthologous groups against the Entomobryoidea bait set (n = 3,406). USCO extraction of each species was made via custom scripts (see [7] under "script 2" for step-by-step commands). Genes were then aligned, trimmed and filtered based on sequence composition and based on Relative Composition Variability (RCV). Posteriorly, gene trees were generated using IQTree v2.0.7 [10]. The gene trees were filtered and only loci with more than 75% average bootstrap support (ABS) value of all internal branches on the gene tree were retained; these filtering steps have previously been shown to improve phylogenetic inference. UCE extraction of each species was made via custom scripts (see [7] under "script 4" for step-by-step commands). Phyluce v1.7.1 [11] was used for harvesting UCEs from genomes. The nucleotide sequences were aligned, trimmed and the same filtering procedures for the alignments and for gene trees described for the USCO dataset were made for the UCE dataset. PhyKIT v.1.9.0 [12] was used to generate the supermatrix, partition, and occupancy for loci alignments.

2.4. Phylogenetic Analyses

To avoid possible systematic errors in large genomic datasets, phylogenetic reconstructions were made using a diverse set of analytical methods. Phylogenetic relationships were inferred from the USCO protein and UCE nucleotides matrices based on 33 Entomobryoidea species and two outgroups (*Desoria trispinata* and *Folsomia candida*, both isotomids). Bayesian inference using PhyloBayes MPI Version 1.5a [13] and Maximum Likelihood inference performed using IQ-Tree v2.0.7 [10] were done. Individual gene trees from each gene alignment for both datasets (USCO and UCE data)

were estimated with IQTree v2.0.7 [10], species-tree was inferred with ASTRAL v5.7.1 [14]. Tree topology tests were conducted under the likelihood framework. Four alternative tree topology hypotheses on constraining monophyly were analyzed. Approximately unbiased (AU) tests, Shimodaira-Hasegawa (SH) and weighted Shimodaira-Hazegawa (WSH) tests calculated probability values (*p*-values) in IQ-Tree v2.0.7 [10], model LG+SSF+F+R4. Hypotheses with p-AU negative were rejected.

3. Results

3.1. Genome Assembly and Annotation

The 21 newly assembled genomes have coverage ranging from 18.91× to 48.92× and the approximate size ranged from 151 to 341 Mb (excluding *Entomobrya proxima* whose initial sequencing data was 30G). BUSCO completeness versus the Entomobryoidea reference set (n = 3,406) were 53.4–89%, 0.2–1.6% duplicated, 0.1–0.9% fragmented and 12.5–46.1% missing. The mean lengths of complete, single-copy BUSCO groups were 374–445 bp for amino acid sequences. For UCEs, the number of extracted loci ranged from 2,886 to 2,233 covering 72,58% (2,663 mean loci number) of the 3,669 targeted UCE loci. The length of each UCE locus ranged from 502 to 2,214 bp, with most around 950 bp.

3.2. Phylogenetic Analyses

The USCO data matrix used for ML analyses had 85% completeness and contained 259.307 sites and 800 loci. For the Bayesian inference, the matrix was reduced to 90% completeness and the size was 141.262 sites and 433 loci. The UCE data matrix had 75% completeness and contained 419.188 sites and 650 loci. Most of the datasets used in this study recovered the topology presented in Figure 1. All tested hypotheses on the monophyly of Paronellidae and Entomobryidae of the tree topology tests were rejected, as well as the sister relation of Seirinae and Entomobryinae. The monophyly of Orchesellidae was confirmed as well as its basal position within the Entomobryoidea. Seirinae and Lepidocyrtinae are both monophyletic and sister-groups. Our results confirmed the paraphyly of Entomobryinae and Paronellidae, and the close relationship of the sampled representatives of Paronellinae s. str. (*Cyphoderus* and *Troglopedetes*), with the clade Seirinae + Lepidocyrtinae with high support in all analyses.

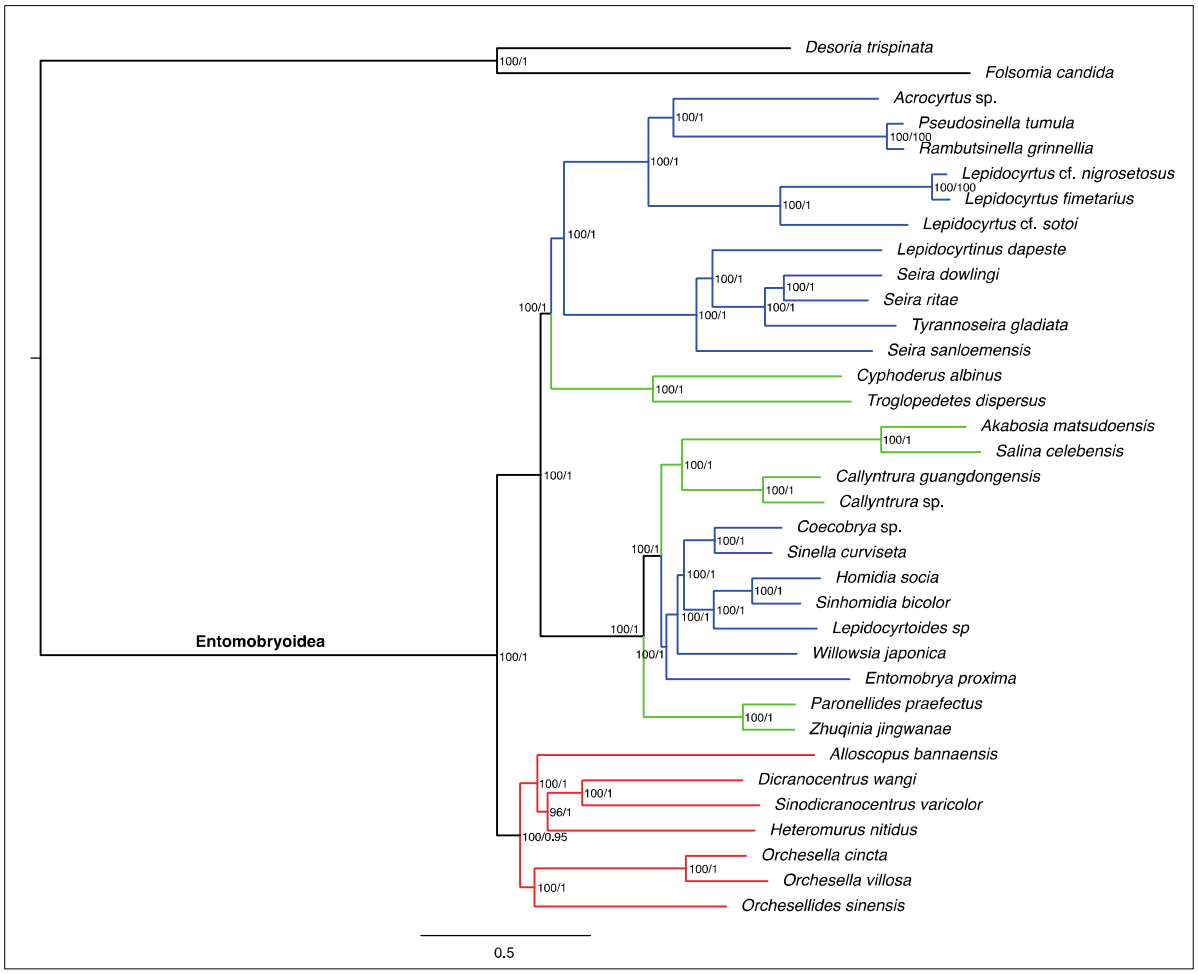

Figure 1. Phylogenetic relationships of the three families of Entomobryoidea and two outgroup taxa. Orchesellidae is marked in red. Species traditionally classified as Entomobryidae are marked in blue and as Paronellidae are marked in green. Numbers at the nodes represent ML bootstrap values and Bayesian posterior probabilities, respectively.

4. Discussion

Our main results are on pair with recent molecular phylogenies of the Entomobryoidea, as well as the detailed ontogenetic study of Szeptycki [15]. Although the systematic view of the Orchesellidae as a family, gathering scaled and scaleless taxa, is not a consensus, recent studies combining morphological and molecular data [3] and mitogenomes [1] support the separation of the Orchesellidae from other Entomobryoidea. Also, our tree found the scaled Heteromurinae as an independent taxon from the unscaled Orchesellinae. Although such data is highly supported by morphology and other molecular studies [3,16], it was not observed in mitogenome based analysis [1].

The Entomobryinae remains a puzzling group based on our data and previously published papers [1,3,4]. It gathers scaled and unscaled taxa with different furca morphologies, Paronellidae and Entomobryidae-like, following the traditional systematics of the Entomobryoidea [1,3,4]. Such findings in independent studies based in different molecular markers and/or morphological evidence strongly support that the emergence of scales within the Entomobryoidea occurred multiples times, as well as the modifications on the furca morphology (e.g., loss of dental crenulations and changes in the mucronal morphology) [3,4,15].

The finding of the Seirinae as the sister-group of the Lepidocyrtinae was already expected, since it was observed in more recent molecular studies, and this relationship is

supported by morphological evidence as well [1,3]. On the other hand, the finding of Paronellinae s. str. as the sister-group of the Seirinae+Lepidocyrtinae clade was not expected, since both morphology and some previous phylogenies support the Paronellinae s. str. as a closer related taxon to the Lepidocyrtinae, sometimes even as an ingroup of the later [3,4]. In this case, the use of a larger set of Paronellinae s. str. species from different genera should provide a more solid understanding of this lineage position among the more derived Entomobryoidea.

5. Conclusions

Our results provide further support to the recent advances on the Systematics of the Entomobryoidea. Based on an expanded set of molecular markers we were able to reaffirm the Orchesellidae as a family, holding at least two independent subfamilies: Heteromurinae and Orchesellinae; the Entomobryinae as group gathering unscaled and scaled taxa with different furcal morphologies; and the Seirinae as the sister-group of the Lepidocyrtinae, with the Paronellinae s. str. as a closely related group to both.

Funding: This research was financed by National Natural Science Foundation of China (31970434) and National Science & Technology Fundamental Resources Investigation Program of China (2018FY100300). NNG is currently funded by the Research Foundation of Shanghai Science and Technology Museum and Postdoctoral fund of Haibo Program of Pudong New Area in 2021. BCB is currently funded by CNPQ/PQ2018, Process #305426/2018-4 and NGC by CNPq (PCI-DB, Process #300925/2019-0).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Godeiro, N.N.; Bellini, B.C.; Ding, N.; Xu, C.; Ding, Y.; Zhang, F. A Mitogenomic Phylogeny of the Entomobryoidea (Collembola): A Comparative Perspective. *Zool. Scr.* **2021**, *50*, 658–666. https://doi.org/10.1111/zsc.12487.
2. Sun, X.; Ding, Y.; Orr, M.C.; Zhang, F. Streamlining Universal Single-Copy Orthologue and Ultraconserved Element Design: A Case Study in Collembola. *Mol. Ecol. Resour.* **2020**, 1–12. https://doi.org/10.1111/1755-0998.13146.
3. Zhang, F.; Bellini, B.C.; Soto-adames, F.N. New Insights into the Systematics of Entomobryoidea (Collembola: Entomobryomorpha): First Instar Chaetotaxy, Homology and Classification. *Zool. Syst.* **2019**, *44*, 249–278. https://doi.org/10.11865/zs.201926.
4. Zhang, F.; Sun, D.-D.; Yu, D.-Y.; Wang, B.-X. Molecular Phylogeny Supports S-Chaetae as a Key Character Better than Jumping Organs and Body Scales in Classification of Entomobryoidea (Collembola). *Sci. Rep.* **2015**, *5*, 12471. https://doi.org/10.1038/srep12471.
5. Baca, S.M.; Alexander, A.; Gustafson, G.T.; Short, A.E.Z. Ultraconserved Elements Show Utility in Phylogenetic Inference of Adephaga (Coleoptera) and Suggest Paraphyly of "Hydradephaga." *Syst. Entom.* **2017**, *42*, 786–795.
6. Starrett, J.; Derkarabetian, S.; Hedin, M.; Bryson, R.W., Jr.; McCormack, J.E.; Faircloth, B.C. High Phylogenetic Utility of an Ultraconserved Element Probe Set Designed for Arachnida. *Mol. Ecol. Resour.* **2017**, *17*, 812–823. https://doi.org/10.1111/1755-0998.12621.
7. Zhang, F.; Ding, Y.; Zhu, C.-D.; Zhou, X.; Orr, M.; Scheu, S.; Luan, Y.-X. Phylogenomics from Low-Coverage Whole-Genome Sequencing. *Methods Ecol. Evol.* **2019**, *10*, 507–517. https://doi.org/10.1111/2041-210X.13145.
8. Stanke, M.; Morgenstern, B. AUGUSTUS: A Web Server for Gene Prediction in Eukaryotes That Allows User-Defined Constraints. *Nucl. Ac. Res.* **2005**, *33*, W465–W467.
9. Waterhouse, R.M.; Seppey, M.; Simão, F.A.; Manni, M.; Ioannidis, P.; Klioutchnikov, G.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol. Biol. Evol.* **2018**, *35*, 543–548. https://doi.org/10.1093/molbev/msx319.
10. Minh, B.Q.; Schmidt, H.A.; Chernomor, O.; Schrempf, D.; Woodhams, M.D.; von Haeseler, A.; Lanfear, R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **2020**, *37*, 1530–1534. https://doi.org/10.1093/molbev/msaa015.
11. Faircloth, B.C. PHYLUCE Is a Software Package for the Analysis of Conserved Genomic Loci. *Bioinformatics* **2016**, *32*, 786–788. https://doi.org/10.1093/bioinformatics/btv646.

12. Steenwyk, J.L.; Buida, T.J.; Labella, A.L.; Li, Y.; Shen, X.-X.; Rokas, A. PhyKIT: A Broadly Applicable UNIX Shell Toolkit for Processing and Analyzing Phylogenomic Data. *Bioinformatics* **2021**, *37*, 2325–2331. https://doi.org/10.1093/bioinformatics/btab096.
13. Lartillot, N.; Rodrigue, N.; Stubbs, D.; Richer, J. PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. *Syst. Biol.* **2013**, *62*, 611–615. https://doi.org/10.1093/sysbio/syt022.
14. Mirarab, S.; Warnow, T. ASTRAL-II: Coalescent-Based Species Tree Estimation with Many Hundreds of Taxa and Thousands of Genes. *Bioinformatics* **2015**, *31*, i44–i52. https://doi.org/10.1093/bioinformatics/btv234.
15. Szeptycki, A. *Chaetotaxy of the Entomobryidae and Its Phylogenetical Significance. Morpho-Systematic Studies on Collembola*; Polska Akademia Nauk: Kraków, Poland, 1979; Volume IV, pp. 1–219.
16. Zhang, F.; Cipola, N.G.; Pan, Z.-X.; Ding, Y. New insight into the systematics of Heteromurini (Collembola: Entomobryidae: Heteromurinae) with special reference to *Alloscopus* and *Sinodicranocentrus* gen. n. *Arth. Syst. Phylogeny* **2020**, *78*, 1–16.