A High-quality Draft Genome Assembly of *Sinella curviseta*: A Soil Model Organism (Collembola)

Feng Zhang[1,2,*], Yinhuan Ding[1], Qing-Song Zhou[2], Jun Wu[3], Arong Luo[2,*], and Chao-Dong Zhu[2,4]

[1]Department of Entomology, College of Plant Protection, Nanjing Agricultural University

[2]Key Laboratory of the Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing, China

[3]Nanjing Institute of Environmental Sciences under Ministry of Environmental Protection, Nanjing, China

[4]College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China

*Corresponding authors: E-mails: xtmtd.zf@gmail.com; luoar@ioz.ac.cn.

Accepted: January 16, 2019

Data deposition: This project has been deposited at the NCBI under the accessions RBVU00000000, GGYG00000000, MK014212 and SRR7948080–SRR7948082, and also at the Figshare under the link https://figshare.com/projects/A_high-quality_draft_genome_assembly_of_Sinella_curviseta_a_soil_model_organism_Collembola_/56291.

Abstract

Sinella curviseta, among the most widespread springtails (Collembola) in Northern Hemisphere, has often been treated as a model organism in soil ecology and environmental toxicology. However, little information on its genetic knowledge severely hinders our understanding of its adaptations to the soil habitat. We present the largest genome assembly within Collembola using ~44.86 Gb (118X) of single-molecule real-time Pacific Bioscience Sequel sequencing. The final assembly of 599 scaffolds was ~381.46 Mb with a N50 length of 3.28 Mb, which captured 95.3% complete and 1.5% partial arthropod Benchmarking Universal Single-Copy Orthologs ($n = 1066$). Transcripts and circularized mitochondrial genome were also assembled. We predicted 23,943 protein-coding genes, of which 83.88% were supported by transcriptome-based evidence and 82.49% matched protein records in UniProt. In addition, we also identified 222,501 repeats and 881 noncoding RNAs. Phylogenetic reconstructions for Collembola support Tomoceridae sistered to the remaining Entomobryomorpha with the position of Symphypleona not fully resolved. Gene family evolution analyses identified 9,898 gene families, of which 156 experienced significant expansions or contractions. Our high-quality reference genome of *S. curviseta* provides the genetic basis for future investigations in evolutionary biology, soil ecology, and ecotoxicology.

Key words: PacBio sequencing, Entomobryidae, comparative genomics, phylogenomics, gene family.

Introduction

Soil invertebrates, as well as soil microbes, contribute to essential soil functions and ecosystem services through trophic and nontrophic effects, such as organic matter degradation, nutrient cycling, pest control, human health (Wall et al. 2012, 2015). However, our knowledge on soil biodiversity and their function remains limited. As one of the oldest hexapod clade (Hirst and Maulik 1926), springtails (Collembola) are among the most dominant soil arthropods, predominantly living in almost all terrestrial ecosystems (Christiansen 1992). Besides the great values in evolutionary biology, collembolan species are often selected as model organisms for soil ecology especially ecotoxicology, such as "standard" parthenogenetic *Folsomia candida* Willem (ISO 1999; Organisation for Economic Co-operation and Development 2009). Bisexual reproducing species *Sinella curviseta* Brook (fig. 1) is an alternative model listed in OECD. It belongs to the largest collembolan family Entomobryidae (Entomobrymorpha) and is among the most widespread springtails in Northern Hemisphere (Bellinger et al. 1996–2018). Easy morphological identification and a high rate of reproduction make it suitable for various laboratory experiments. Reproduction, development and life history of *S. curviseta* have been well documented (Waldorf 1971; Nijima 1973; Gist et al. 1974; Zhang et al. 2011). However, limited genetic data, usually presented by sequences of barcoding and rRNA (Zhang

et al. 2014), severely hinders its further application in a wider scope. A high-quality reference genome will provide a solid genetic basis for the understanding of molecular mechanisms and physiological functions in adaptations to environmental change, as well as evolutionary biology. To date, only three genome assemblies have been published for Collembola: *Orchesella cincta* Linnaeus (Faddeeva-Vakhrusheva et al. 2016), *F. candida* (Faddeeva-Vakhrusheva et al. 2017), and *Holacanthella duospinosa* Salmon (Wu et al. 2017). Here, we present a de novo genome assembly of *S. curviseta* using single-molecule real-time (SMRT) Pacific Bioscience (PacBio) long reads. We annotated the essential genomic elements, repeats, protein-coding genes, and noncoding RNAs (ncRNAs), and further compared gene family evolution across main arthropod lineages. Phylogeny of Collembola was also investigated using genomic data for the first time.

Materials and Methods

Sample Collection and Sequencing

The culture of *S. curviseta* used in this study was collected from Purple Mountain (32.056°N, 118.83°E, Nanjing, China) in April 2015 and was maintained for three years in our laboratory. Animals were collected with aspirator, washed with ddH_2O, and crushed with liquid nitrogen. A total of 500, 10, 200 individuals were prepared for PacBio, Illumina whole genome, and Illumina transcriptome sequencing, respectively. Genomic DNA/RNA extraction, library preparation and sequencing were carried out at Novogene Co. Ltd. (Beijing, China). For long-read sequencing, a library was constructed with an insert size of 20 kb and sequenced using P6-C4 chemistry on the PacBio Sequel platform. For short-read sequencing, paired-end libraries were constructed with an insert size of 300 bp and sequenced (2×150 bp) on the Illumina HiSeq X Ten platform. Raw Illumina short reads were compressed into clumps and duplicates were removed with clumpify.sh (one of the BBTools suite v37.93, Bushnell). Quality control was performed with bbduk.sh (BBTools): Both sides were trimmed to Q20 using the Phred algorithm, reads shorter than 15 bp or with >5 Ns were discarded, poly-A or poly-T tails of at least 10 bp were trimmed, and overlapping paired reads were corrected. Assembly and annotation pipelines are shown in figure 2.

Genome Size Estimation

We employ the strategy of short-read k-mer distributions to estimate the genome size. K-mer length and maximum k-mer coverage cutoffs may have impacts on estimated genome size. The histogram of k-mer frequencies was computed with 21-mers and 27-mers using Jellyfish v2.2.7 (Marçais and Kingsford 2011). Genome size was estimated with a maximum k-mer coverage of 1,000 and 10,000 using GenomeScope v1.0.0 (Vurture et al. 2017).

Fig. 1.—An adult of *Sinella curviseta*. It has a pale orange body and two longitudinally arranged eye spots on each side.

Genome, Mitochondrion, and Transcriptome Assembly

De novo genome assembly with long reads were performed using two pipelines, Canu and Minimap2/Miniasm. Because of high heterozygosity for *S. curviseta*, we used the parameters 'corOutCoverage = 200 "batOptions = -dg 3 -db 3 -dr 1 -ca 500 -cp 50"' with Canu v1.7.1 (Koren et al. 2017) to output more corrected reads and be more conservative at picking the error rate for the assembly to try to maintain haplotype separation. For Minimap2/Miniasm pipeline, overlaps between long reads were generated with Minimap2 v2.9 (Li 2018). We employed Miniasm v0.3 (Li 2016) to assemble contigs and three rounds of Racon v1.3.1 (Vaser et al. 2017) to generate consensus and correct errors. To improve genome contiguity, two assemblies generated from Canu and Minimap2/Miniasm pipelines were merged with three rounds of quickmerge (Chakraborty et al. 2016) following USAGE 2 (https://github.com/mahulchak/quickmerge/wiki; last accessed September 1, 2018). Redundant heterozygous sequences were removed from merged assembly with Purge Haplotigs v20180917 (Roach et al. 2018); percent cutoff for identifying a contig as a haplotig was set as 60 (-a 60) with other parameters as the default. The resulting contigs were polished with PacBio long reads using two rounds of Arrow mode in GenomicConsensus v2.3.2 (Chin et al. 2013). Furthermore, assembly was polished with Illumina short reads using two rounds of Pilon v1.22 (Walker et al. 2014). We again removed redundant sequences with Purge Haplotigs. Contaminants were examined using PhylOligo v0.9-alpha (Mallet et al. 2017). Potential untargeted sequences were identified by exploring compositional similarity on a tree (phyloselect.R) and hierarchical DBSCAN clustering (phyloselect.py), and inspected with BlastN v2.7.1 (Camacho et al. 2009) against the NCBI nucleotide database. Finally, vector contamination was checked using VecScreen against the UniVec database. The mitochondrial genome of *S. curviseta* was assembled based on Illumina short reads with NOVOPlasy v2.7.0 (Dierckxsens et al. 2017) using *COI* sequence

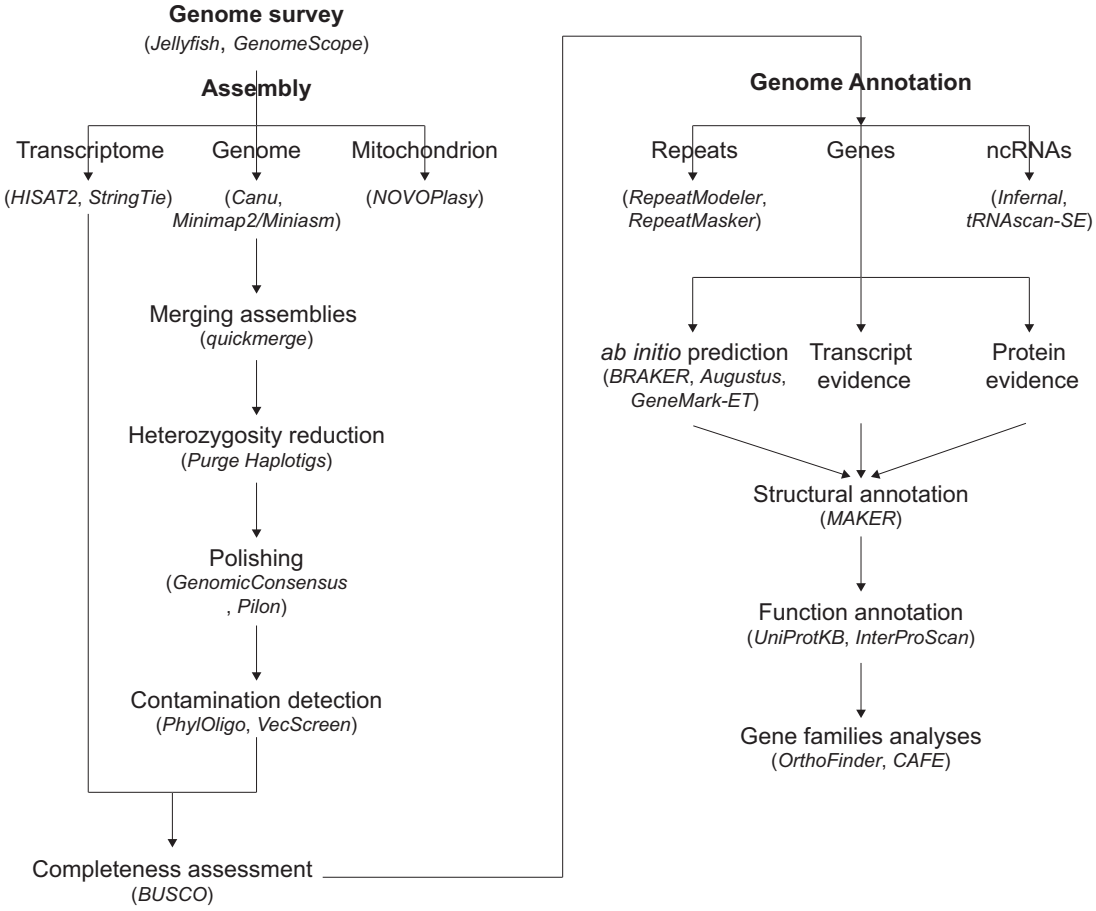

Fig. 2.—Flowchart of assembly and annotation pipelines. Bioinformatic tools used in each step are marked as italic.

(KM978373) as the initial seed. Transcriptome assembly was performed with a genome-guided method. RNA-seq reads were mapped to assembled genome with HISAT2 v2.1.0 (Kim et al. 2015) and assembled with StringTie v1.3.4 (Pertea et al. 2015). Redundant isoforms were removed with Redundans v0.13c (Pryszcz and Gabaldón 2016) with the defaults. To assess the completeness of assemblies, we applied Benchmarking Universal Single-Copy Orthologs (BUSCO, Waterhouse et al. 2018) analyses against arthropod data set ($n = 1066$). In addition, we also mapped PacBio long reads and Illumina short reads to the final genome assembly with Minimap2.

Genome Annotation

A de novo species specific repeat library was constructed using RepeatModeler v1.0.11 (Smit and Hubley 2008–2015), and was then combined with Dfam_2.0, Dfam_Consensus-20170127 (Hubley et al. 2016) and RepBase-20170127 databases (Bao et al. 2015) to generate a custom library. We then used RepeatMasker v4.0.7 (Smit et al. 2013–2015) with the custom library to identify and mask repeats in the genome assembly.

Gene prediction was conducted with the MAKER v2.31.10 pipeline (Holt and Yandell 2011) by integrating ab initio, transcriptome-based and protein homology-based evidence. Ab initio gene predictions were performed with Augustus v3.3 (Stanke et al. 2004) and GeneMark-ET v4.33 (Lomsadze et al. 2005). Two predictors were trained using BRAKER v2.1.0 (Hoff et al. 2016) with RNA-seq data. Previously assembled genome-guided transcripts were used as transcriptome-based evidence. Protein sequences of *Daphnia pulex*, *Acyrthosiphon pisum*, and *Drosophila melanogaster* were downloaded from Ensembl (Flicek et al. 2014) as protein homology-based evidence.

Homology-based gene functions were assigned using Diamond v0.9.18 (Buchfink et al. 2015) against UniProtKB (SwissProt + TrEMBL) database with a sensitive mode and an e-value threshold of 1e–5 (–sensitive -e 1e–5). Protein domains, as well as Gene Ontology (GO) and pathway annotation, were searched with InterProScan 5.30-69.0 (Finn et al. 2017) against Pfam (Finn et al. 2014), PANTHER (Mi et al. 2017), Gene3D (Lewis et al. 2018), Superfamily (Wilson et al. 2009), and CDD (Marchler-Bauer et al. 2017) databases (-dp -f TSV, GFF3 -goterms -iprlookup -pa -t p -appl Pfam, PANTHER, Gene3D, Superfamily, CDD).

ncRNAs were identified with Infernal v1.1.2 (Nawrocki and Eddy 2013) against Rfam v14.0 (Kalvari et al. 2018) database. Transfer RNAs were further refined with tRNAscan-SE v2.0 (Lowe and Eddy 1997).

Phylogenomic Analyses

We conducted a phylogeny of Collembola using public genomic data (three genomes and four transcriptome assemblies): *O. cincta* (GCA_001718145.1), *F. candida* (GCA_002217175.1), *H. duospinosa* (GCA_002738285.1), *Anurida maritima* (GAUE00000000.2), *Tetrodontophora bielanensis* (GAXI00000000.2), *Pogonognathellus* sp. (GATD00000000.2), *Sminthurus viridis* (GATZ00000000.2). One Protura (*Acerentomon* sp., GAXE00000000.2) and one Diplura (*Catajapyx aquilonaris*, GCA_000934665.2) species were selected as the outgroup. Transcriptomic assemblies were reported in Misof et al. (2014) and Diplura genome was from Thomas et al. (2018). Complete single-copy genes were generated with BUSCO assessments against arthropod data set. Gene training set constructed by BRAKER was used for Augustus species gene-finding parameters. Shared single-copy genes were aligned using MAFFT v7.394 (Katoh and Standley 2013) with the L-INS-I strategy, trimmed using trimAl v1.4.1 (Capella-Gutiérrez et al. 2009) with the heuristic method automated1, and concatenated using FASconCAT-G v1.04 (Kück and Longo 2014). We constructed the phylogenetic trees using maximum likelihood (ML) and coalescent-based species tree (ASTRAL) methods. ML reconstructions were performed using IQ-TREE v1.6.3 (Nguyen et al. 2015) with 1,000 ultrafast bootstrap (UFBoot, Hoang et al. 2018) and 1,000 SH-aLRT replicates (Guindon et al. 2010) estimated. Partitioning schemes and substitution models were estimated with ModelFinder (Kalyaanamoorthy et al. 2017) built-in in IQ-TREE. We used a subset of substitution models with the options "-mset" (HKY and GTR for nucleotides, WAG and LG for proteins), and implemented the relaxed hierarchical clustering algorithm (Lanfear et al. 2014) with the setting "-rcluster 10." Species trees were estimated using ASTRAL-III v5.6.1 (Zhang et al. 2018) based on gene trees generated with IQ-TREE on individual gene alignments. Local branch supports were estimated from quartet frequencies (Sayyari and Mirarab 2016).

Gene Family Identification and Evolution

We identified gene families among 12 arthropod species, including four collembolans, four insects (*A. pisum*, *D. melanogaster*, *Tribolium castaneum*, *Zootermopsis nevadensis*), three other nonhexapods (*Ixodes scapularis*, *Strigamia maritima*, *D. pulex*). OrthoFinder v2.2.7 (Emms and Kelly 2015) was used to infer orthogroups with Diamond as the sequence aligner. Gene family evolution (gain and loss) was analyzed using CAFE v4.2 (Han et al. 2013) with lambda parameter to calculate birth and death rates. Species tree and divergence time were generated from TimeTree database (Kumar et al. 2017).

Results and Discussion

Genome Sequencing and Assembly

We generated 4,196,991 subreads of 44.86 Gb (118X) on the PacBio Sequel platform. The mean and N50 length of long subreads reached 9.80 kb and 14.67 kb. A total of 37.95 Gb (99X) and 7.86 Gb clean data were produced on the Illumina HiSeq X Ten platform for whole genome and transcriptome sequencing, respectively.

We estimated the genome size with GenomeScope under the four parameter combinations, ranging from 327.12 Mb to 340.37 Mb (table 1). Genome repetitive length estimates increased with the larger maximum k-mer coverage cutoff, ranging from 19.96 Mb to 32.65 Mb. Unique (nonrepetitive) length estimates were more consistent among analyses, ranging from 304.71 Mb to 311.28 Mb. Overall rate of heterozygosity (0.746–0.886) and the distinct first peak at a mean kmer coverage of 30.7–32.5 in the k-mer plots (supplementary fig. S1, Supplementary Material online) indicated that this genome may have a high rate of heterozygous regions, which should be carefully considered in the subsequent assembly processes.

A Canu assembly of 633.80 Mb and 5,606 contigs (table 2) was generated with most haplotypes kept for diploid populations. Minimap2/Miniasm pipeline resulted in an assembly of 549.01 Mb and 3,288 contigs. Size of both assemblies was much larger than estimated due to the presence of a great number of heterozygous sequences. Both assemblies were merged into 5,437 contigs (L50 708 kb) with quickmerge. A total of 4,779 (246.11 Mb) and 58 (1.37 Mb) redundant heterozygous sequences were respectively removed with two rounds of Purge Haplotigs, resulting a great improvement in N50 length (3.28 Mb). No evident contaminants were found using PhyIOligo (supplementary fig. S2, Supplementary Material online). A vector sequence was excluded. Final draft assembly of *S. curviseta* has 599 contigs/scaffolds (no gaps), total length of 381.46 Mb, N50 length of 3.28 Mb, a maximum scaffold length of 12.99 Mb, and 37.51% GC content. Our assembly has a largest genome size among four collembolan species, a much higher assembly quality than *O. cincta* and *H. duospinosa*, but is slightly more fragmented than *F. candida* (table 3). With the genome-guided strategy, a total of 27,976 transcripts were assembled with a mean and N50 length of 2.26 kb and 3.50 kb.

We generated a circularized mitochondrial genome of 14,840 bp. It consists of 13 protein-coding genes, 2 rRNA genes and 22 tRNA genes. The mitochondrial gene number and order are similar to most collembolan species. The A + T

Table 1
GenomeScope Genome Size Estimates for *Sinella curviseta*

K-mer	Max K-mer Coverage	Heterozygosity (%)		Repeat Length (Mb)		Unique Length (Mb)		Genome Size (Mb)	
		Mix	Max	Mix	Max	Mix	Max	Mix	Max
21	1,000	0.851	0.886	22.42	22.53	304.71	306.22	327.12	328.85
21	10,000	0.863	0.874	32.60	32.65	305.22	305.70	337.82	338.36
27	1,000	0.746	0.772	19.96	20.05	309.83	311.28	329.79	331.33
27	10,000	0.755	0.763	29.54	29.59	310.32	310.79	339.87	340.37

Table 2
Summary of Each Assembly at Each Step for *Sinella curviseta*

Assembly	Total Length (Mb)	No. Scaffolds	N50 Length (kb)	Longest Scaffold (Mb)	GC (%)	BUSCO ($n = 1066$) (%)			
						C	D	F	M
Canu	633.80	5,606	521	8.89	37.55	95.4	40.5	1.7	2.9
Minimap2/Miniasm	549.01	3,288	309	7.249	37.60	92.8	15.8	3.5	3.7
quickmerge	628.88	5,437	708	12.98	37.55	95.0	39.5	1.8	3.2
purge_haplotigs_1	382.77	658	3288	12.98	37.53	94.2	5.2	2.4	3.4
Arrow	383.93	658	3290	12.99	37.53	95.3	4.9	1.7	3.0
Pilon	382.83	658	3284	12.90	37.52	95.3	5.6	1.6	3.1
purge_haplotigs_2	381.46	600	3284	12.99	37.51	95.3	5.4	1.5	3.2
Final genome assembly	**381.46**	**599**	**3284**	**12.99**	**37.51**	**95.3**	**5.4**	**1.5**	**3.2**
Transcript assembly	**63.34**	**27,976**	**3.50**	**0.056**	**40.68**	**94.6**	**8.4**	**2.3**	**3.1**

Note.—Reduction of heterozygous regions was carried out twice (_1 and _2) with purge_haplotigs. Values of final assemblies are bold. C, complete BUSCOs; D, complete and duplicated BUSCOs; F, fragmented BUSCOs; M, missing BUSCOs.

content (69.8%) in *S. curviseta* is slightly smaller than those in known Entomobryomorpha species.

Assembly completeness were assessed with BUSCO analyses against arthropod data set ($n = 1066$). We identified 92.8–95.4% complete, 1.5–3.5% fragmented, and 2.9–3.7 missing BUSCOs for all versions of genome assemblies (table 2). Comparable results to other collembolan genomes indicated the high completeness of our assembly. Genome-guided transcriptome assembly also showed similar completeness to the genome assembly. The tremendous decreasing of duplicated BUSCOs indicated that Purge Haplotigs and Redundans could be highly efficient for reducing heterozygous regions. In addition, we mapped 93.22% and 96.07% of PacBio long reads and Illumina short reads to the final genome assembly. Also, 27,956 (99.93%) assembled transcripts were aligned to the genome using BlastN with an identity value of 0.99.

Genome Annotation

RepeatMasker identified 222,501 repeats which masked 9.79% of the genome assembly. The top five abundant repeat types were simple repeats, unclassified repeats, low complexity repeats, Helitron transposable elements, and Gypsy LTR retrotransposons (supplementary table S1, Supplementary Material online). We compared repeat components among four collembolan species (table 3). Three species (*S. curviseta*, *O. cincta*, *F. candida*) belonging to Entomobryomorpha have similar repeat compositions, but sharply differ from *H. duospinosa* in DNA and unclassified repeats. High similarity between *S. curviseta* and *O. cincta* is also consistent with their systematic positions (within the same family Entomobryidae).

MAKER pipeline identified 23,943 protein-coding genes with a mean of 5.59 exons and 15.95 introns per gene. Exons and introns have a mean length of 446.95 bp and 96.58 bp, respectively (table 3). Gene density of the *S. curviseta* genome was 62.77 genes per Mbp, which is smaller than the density in the genomes of *O. cincta* (70.60) and *F. candia* (129.6) but higher than that in *H. duospinosa* (30.21). Among predicted genes, 20,083 (83.88%) were supported by transcriptome-based evidence. BUSCO analysis identified 987 (92.6%) complete 92 (8.6%) duplicated, 15 (1.4%) fragmented, and 64 (6%) missing BUSCOs. A total of 19,750 (82.49%) genes hit at least one record in SwissProt or TrEMBL databases. InterProScan identified protein domains for 18,870 (78.71%) genes; among them, 11,581 were assigned with GO terms, and 977, 745 and 3,922 genes matched KEGG, MetaCyc, and Reactome pathway databases, respectively.

For ncRNAs, we identified 147 rRNAs (5S, 5.8S, LSU, SSU), 394 tRNAs, 33 miRNAs (19 families), 71 small nuclear RNAs (snRNAs), 17 ribozymes (3 families) and 235 cis-regulatory

Table 3
Genome Comparison among Four Collembolan Species

Elements	*S. curviseta*	*O. cincta*	*F. candida*	*H. duospinosa*
Genome assembly				
Total length (Mb)	381.46	286.77	221.70	327.57
No. scaffolds	599	9402	162	62430
L50 length (kb)	3284	66	6519	328
Longest scaffold (Mb)	12.99	0.81	28.53	2.81
GC (%)	37.51	36.81	37.52	33.35
C	95.3	92.5	9.6	95.3
D	5.4	4.4	1.4	8.0
F	1.5	2.2	1.2	1.2
M	3.2	5.3	2.8	3.5
Repeats (bp/P%)				
Total	37,343,105/9.79	43,148,530/15.04	51,606,299/23.28	161,336,129/42.96
DNA	2,660,344/0.70	3,351,265/1.17	9,462,341/4.27	31,620,408/8.42
LINE	2,751,749/0.72	3,528,411/1.23	1,939,306/0.87	5,971,075/1.59
LTR	5,931,748/1.56	5,016,293/1.75	2,557,166/1.15	10,439,992/2.78
SINE	212,035/0.06	214,353/0.07	22,158/0.01	110,785/0.00
Unclassified	17,125,523/4.49	26,087,080/9.10	30,170,966/13.60	106,352,725/28.32
Simple repeats	4,424,627/1.16	3,838,732/1.34	5,080,872/2.29	6,196,398/1.65
Others	4,237,034/1.11	1,112,396/0.39	2,373,490/1.07	640,294/0.17
Gene annotations (number/length)				
Genes	23,943 (96.75M)	20,249(60.56M)	28734(132.62M)	9,911(56.66M)
Gene mean length (bp)	4040.85	2990.75	4615.44	5716.88
Exons	133,951 (59.87M)	118,474(32.23M)	197,859(70.64M)	79,659(22.71M)
Exon mean length (bp)	446.95	272.04	357.02	285.09
Introns	381,850 (36.88M)	336,337(28.33M)	524,921(61.98M)	242,640(33.95M)
Intron mean length (bp)	96.58	84.23	118.07	139.92

NOTE.—C, complete BUSCOs; D, complete and duplicated BUSCOs; F, fragmented BUSCOs; M, missing BUSCOs; P%, percentage of the genome.

elements (3 families), and 1 other (Metazoa_SRP) ncRNAs (supplementary table S2, Supplementary Material online). snRNAs were classified into six spliceosomal RNAs (U1, U2, U4, U5, U6, U11), three minor spliceosomal RNAs (U12, U4atac, U6atac), two H/ACA box and 12 C/D box snoRNAs (small nucleolar RNA), and one (SCARNA8) scaRNA (small Cajal body-specific RNA). A total of 21 tRNA isotypes were identified except for Supres-isotype missing.

Phylogenomic Analyses

Nucleotide and protein matrices comprising 229 shared single-copy genes had 250,783 and 81,557 sites, and were divided with ModelFinder into 32 and 39 partitions, respectively. ML trees from nucleotide and protein matrices generated similar topologies except for the position of Symphypleona *S. viridis*. Species trees generated with ASTRAL-III had the same performance as the ML trees. All support values were absolutely high for most nodes (fig. 3). The phylogeny of Collembola at high levels were far from resolved in previous studies (D'Haese 2002, 2003; Xiong et al. 2008; Yu et al. 2016). Topological hypothesis of Symphypleona sistered to Entomobryomorpha based on nucleotide matrix agreed with morphologically cladistic analyses (D'Haese 2003), but alternative hypothesis of Symphypleona sistered to the remaining collembolan taxa were usually consistent with molecular phylogenies, as well as our results from the nucleotide matrix. Interestingly, our trees provided robust evidence supporting the sister relationship between Tomoceridae (*Pogonognathellus* sp.) and the remaining Entomobryomorpha, as indicated by evidence from the first instar larvae (Yu et al. 2016). Wider taxa sampling with more families included may help to achieve the ultimate phylogeny of Collembola.

Gene Family Evolution

Gene families were identified among 12 arthropod species with OrthoFinder. A total 67.03% (166,850) genes were assigned into 14,387 gene families with a mean orthogroups size of 11.6. Among 2,396 families shared by all species, 207 are single-copy orthogroups. In *S. curviseta*, 18,590 (77.64%) genes were clustered into 9,898 gene families, and 66 families and 487 genes were species-specific (supplementary table S3, Supplementary Material online). We analyzed gene family evolution (gain and loss) using CAFE. Estimated gene birth rate (lambda) was 0.00165, accounting for

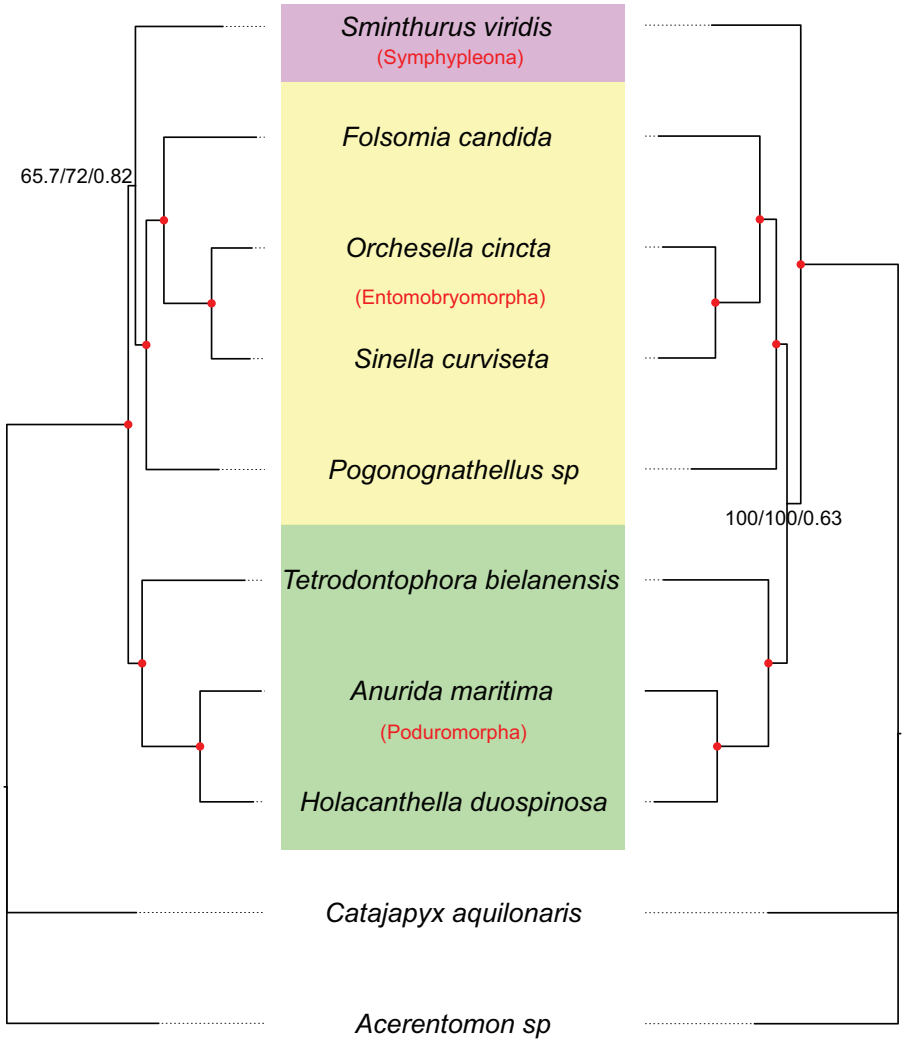

Fig. 3.—Phylogeny of Collembola based on protein (left) and nucleotide (right) matrices comprising 229 single-copy genes. Node support values below 100 (SH-aLRT, UFBoot) or 1 (quartet frequencies) are given on nodes. Red circles represent the best support values of 100/100/1.

duplications/gene/Mya. Expansions and contractions of gene families for 12 species are shown in figure 4. A total of 445 gene families experienced significant expansion or contraction events across the tree with a family-wide P-value < 0.05 (supplementary table S4, Supplementary Material online). *Sinella curviseta* showed 156 (131 expansions, 25 contractions) rapidly evolving families among 445 ones previously detected. The top five of the largest expanded families included Zinc-finger proteins (392), Ribonuclease H-like proteins (139), C-type lectin proteins (104), Carboxylesterase proteins (83), and F-box proteins (81) (supplementary table S5, Supplementary Material online). Zinc-finger proteins are transcription factors serving a wide variety of biological functions by binding DNA, RNA, proteins, or small molecules (Laity et al. 2001). Ribonuclease H-like superfamily is involved with nucleic acid metabolism, including DNA replication and repair, homologous recombination, transposition and RNA interference. C-type lectins have functions in innate and adaptive antimicrobial immune responses (Brown et al. 2018). Carboxylesterase enzymes participate in phase I xenobiotic metabolism. F-box proteins are associated with cellular functions such as signal transduction and regulation of the cell cycle (Craig and Tyers 1999). The top expanded gene families of three Entomobryomorpha species (*S. curviseta*, *O. cincta*, *F. candida*) have great similarities, participating in detoxication and xenobiotic metabolism, nucleic acid metabolism, immune system progress, signaling, etc. Expansion of these families are essential for adaptations to the complicated soil environment. This partly explains the reasons why the three species can be widespread in the Northern Hemisphere. It is not clear that whether these expansions are Entomobryomorpha-specific because genomic data are still lacking for the indigenous species.

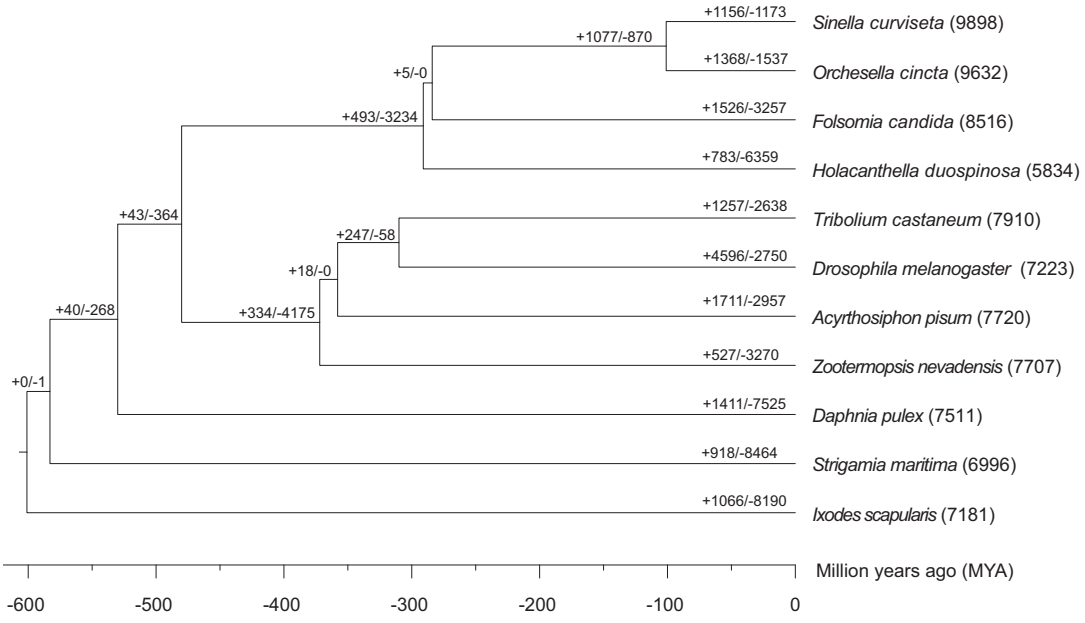

Fig. 4.—Expansions (gain) and contractions (loss) of 14,387 gene families on a species tree. Gain and loss are indicated with symbol + and –. Numbers of gene families for each species are shown following species name. Tree topology and divergence time were generated from TimeTree database.

Conclusion

With PacBio long reads, we report the largest collembolan draft assembly of *S. curviseta*, a soil model organism. Our high-quality genome assembly comprises 599 contigs of 381.46 Mb (N50 length of 3.28 Mb), covering over 95% the arthropod universal BUSCO sets. We also predict 23,943 protein-coding genes. Phylogenomic analysis support Tomoceridae closer to Entomobryomorpha than Poduromorpha. The genomic data produced in this study will provide a valuable resource for future studies in evolutionary biology, soil ecology, and ecotoxicology.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Science (XDB310304) and the National Key R&D Program of China (2016YFC1200705). F.Z. was also supported by the National Natural Science Foundation of China (31772491), and a grant from the Key Laboratory of the Zoological Systematics and Evolution of the Chinese Academy of Sciences (Y229YX5105). C.Z. acknowledges funding supports by the National Science Fund for Distinguished Young Scholars (31625024).

Author Contributions

F.Z., A.L., and C.Z. designed the study. F.Z., Y.D., J.W., and Q.Z. collected the samples and performed the analyses. F.Z. and A.L. wrote the paper. All authors edited and approved the final manuscript.

Literature Cited

Bao W, Kojima KK, Kohany O. 2015. Repbase update, a database of repetitive elements in eukaryotic genomes. Mobile DNA. 6:11.

Bellinger PF, Christiansen KA, Janssens F. 1996–2018. Checklist of the Collembola of the World [cited 2018 Sep 1]. Available from: http://www.collembola.org.

Brown GD, Willment JA, Whitehead L. 2018. C-type lectins in immunity and homeostasis. Nat Rev Immunol. 18(6):374–389.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 12(1):59–60.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25(15):1972–1973.

Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. 2016. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. 44(19):e147.

Chin CS, et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 10(6):563–569.

Christiansen KA. 1992. Springtails. Kansas Sch Nat. 39:1–16.

Craig KL, Tyers M. 1999. The F-box: a new motif for ubiquitin dependent proteolysis in cell cycle regulation and signal transduction. Prog Biophys Mol Biol. 72(3):299–328.

Dierckxsens N, Mardulyn P, Smits G. 2017. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. 45(4):e18.

D'Haese CA. 2002. Were the first springtails semi-aquatic? A phylogenetic approach by means of 28S rDNA and optimization alignment. Proc Biol Sci. 269:1143–1151.
D'Haese CA. 2003. Morphological appraisal of Collembola phylogeny with special emphasis on Poduromorpha and a test of the aquatic origin hypothesis. Zool Scr. 32:563–586.
Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16:157.
Faddeeva-Vakhrusheva A, et al. 2016. Gene family evolution reflects adaptation to soil environmental stressors in the genome of the Collembolan *Orchesella cincta*. Genome Biol Evol. 8(7):2106–2017.
Faddeeva-Vakhrusheva A, et al. 2017. Coping with living in the soil: the genome of the parthenogenetic springtail *Folsomia candida*. BMC Genomics 18(1):493.
Finn RD, et al. 2017. InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res. 45(D1):D190–D199.
Finn RD, et al. 2014. Pfam: the protein families database. Nucleic Acids Res. 42(D1):D222–D230.
Flicek P, et al. 2014. Ensembl 2014. Nucleic Acids Res. 42(Database issue):D749–D755.
Gist CS, Crossley DA, Merchant VA. 1974. An analysis of life tables for *Sinella curviseta* (Collembola). Environ Entomol. 3(5):840–845.
Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59(3):307–321.
Han MV, Thomas GW, Lugo-Martinez J, Hahn MW. 2013. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol Biol Evol. 30(8):1987–1997.
Hirst S, Maulik S. 1926. On some arthropod remains from the Rhynie Chert (Old Red Sandstone). Geol Mag. 63(02):69–71.
Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1: unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32(5):767–769.
Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 12:491.
Hubley R, et al. 2016. The Dfam database of repetitive DNA families. Nucleic Acids Res. 44(D1):D81–D89.
ISO 1999. 11267: soil quality. Inhibition of reproduction of Collembola (Folsomia candida) by soil pollutants. Geneva (Switzerland): International Organization for Standardization.
Kalvari I, et al. 2018. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res. 46(D1):D335–D342.
Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 14(6):587–589.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30(4):772–780.
Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 12(4):357–360.
Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol. 35(2):518–522.
Koren S, et al. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27(5):722–736.
Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: a resource for timelines, timetrees, and divergence times. Mol Biol Evol. 34 (7):1812–1819.
Kück P, Longo GC. 2014. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. Front Zool. 11(1):81.
Lanfear R, Calcott B, Kainer D, Mayer C, Stamatakis A. 2014. Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol Biol. 14:82.
Laity JH, Lee BM, Wright PE. 2001. Zinc finger proteins: new insights into structural and functional diversity. Curr Opin Struct Biol. 11(1):39–46.
Lewis TE, et al. 2018. Gene3D: extensive prediction of globular domains in proteins. Nucleic Acids Res. 46(D1):D435–D439.
Li H. 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics 32(14):2103–2110.
Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34(18):3094–3100.
Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 33(20):6494–6506.
Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25(5):955–964.
Mallet L, Bitard-Feildel T, Cerutti F, Chiapello H. 2017. PhylOligo: a package to identify contaminant or untargeted organism sequences in genome assemblies. Bioinformatics 33(20):3283–3285.
Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27(6):764–770.
Marchler-Bauer A, et al. 2017. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. 45(D1):D200–D203.
Mi H, et al. 2017. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. 45(D1):D183–D189.
Misof B, et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science 346(6210):763–767.
Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29(22):2933–2935.
Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. Mol Biol Evol. 32(1):268–274.
Nijima K. 1973. Experimental studies on the life history, fecundity and growth of *Sinella curviseta* (Apterygota, Collembola). Pedobiologia 13:186–204.
Organisation for Economic Co-operation and Development. 2009. Test no. 232: Collembolan reproduction test in soil. Paris (France): OECD Publishing.
Pertea M, et al. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 33(3):290–295.
Pryszcz LP, Gabaldón T. 2016. Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Res. 44(12):e113.
Roach MJ, Schmidt SA, Borneman AR. 2018. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics 19:460.
Sayyari E, Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Mol Biol Evol. 33(7):1654–1668.
Smit AFA, Hubley R. 2008–2015. RepeatModeler Open-1.0 [cited 2018 Apr 1]. Available from: http://www.repeatmasker.org.
Smit AFA, Hubley R, Green P. 2013–2015. RepeatMasker Open-4.0 [cited 2018 Apr 1]. Available from: http://www.repeatmasker.org.
Stanke M, Steinkamp R, Waack S, Morgenstern B. 2004. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res. 32(Web Server issue):W309–W312.
Thomas GWC, et al. 2018. The genomic basis of arthropod diversity. bioRxiv 382945.
Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27(5):737–746.

Vurture GW, et al. 2017. GenomeScope: fast reference-free genome profling from short reads. Bioinformatics 33(14):2202–2204.
Walker BJ, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9(11):e112963.
Waldorf ES. 1971. Reproductive biology of *Sinella curviseta* (Collembola: Entomobryidae) in laboratory culture. Rev Ecol Biol Sol. 8(3):451–463.
Wall DH, et al. 2012. Soil ecology and ecosystem services. Oxford: Oxford University Press.
Wall DH, Nielsen UN, Six J. 2015. Soil biodiversity and human health. Nature 528(7580):69–76.
Waterhouse RM, et al. 2018. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 35(3):543–548.
Wilson D, et al. 2009. SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res. 37(Suppl 1):D380–D386.
Wu C, et al. 2017. Analysis of the genome of the New Zealand giant Collembolan (*Holacanthella duospinosa*) sheds light on hexapod evolution. BMC Genomics 18(1):795.
Xiong Y, Gao Y, Yin WY, Luan YX. 2008. Molecular phylogeny of Collembola inferred from ribosomal RNA genes. Mol Phylogenet Evol. 49(3):728–735.
Yu D, et al. 2016. New insight into the systematics of Tomoceridae (Hexapoda, Collembola) by integrating molecular and morphological evidence. Zool Scr. 45(3):286–299.
Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19(Suppl 6):153.
Zhang F, et al. 2014. Molecular phylogeny reveals independent origins of body scales in Entomobryidae (Hexapoda: Collembola). Mol Phylogenet Evol. 70:231–239.
Zhang F, Yu D, Xu G. 2011. Transformational homology of the tergal setae during postembryonic development in the *Sinella-Coecobrya* group (Collembola: entomobryidae). Contrib Zool. 80(4):213–230.

Associate editor: Maria Costantini