Chromosomal-Level Genome Assembly of the Springtail *Tomocerus qinae* (Collembola: Tomoceridae)

Zhixiang Pan[1], Jianfeng Jin[2], Cong Xu[2], and Daoyuan Yu[3,*]

[1]School of Life Sciences, Taizhou University, Zhejiang 318000, China

[2]Department of Entomology, College of Plant Protection, Nanjing Agricultural University, Nanjing 210095, China

[3]Soil Ecology Lab, College of Resources and Environmental Sciences, Nanjing Agricultural University, Nanjing 210095, China

*Corresponding author: E-mail: yudy@njau.edu.cn.

Accepted: 10 March 2022

Abstract

The family Tomoceridae is among the earliest derived collembolan lineages, thus is of key importance in understanding the evolution of Collembola. Here, we assembled a chromosome-level genome of one tomocerid species *Tomocerus qinae* by combining Nanopore long reads and Hi-C data. The final genome size was 334.44 Mb with the scaffold/contig N50 length of 71.85/13.94 Mb. BUSCO assessment indicated that 96.80% of complete arthropod universal single-copy orthologs ($n = 1,013$) were present in the assembly. The repeat elements accounted for 26.11% (87.26 Mb) and 494 noncoding RNAs were identified in the genome. A total of 20,451 protein-coding genes were predicted, which captured 96.0% (973) BUSCO genes. Gene family evolution analyses identified 4,825 expanded gene families of *T. qinae*, among them, 47 experienced significant expansions, and these significantly expanded gene families mainly involved in proliferation and growth. This study provides an important genomic resource for future evolution and comparative genomics analyses of Collembola.

Key words: Nanopore, Hi-C, gene family evolution, comparative genomics, Tomocerinae.

Significance

Collembolans are particularly significant members of the soil communities and are of key importance in understanding arthropod evolution. A total of 39 genomes of Collembola have been reported at the NCBI database, whereas only seven species have been sequenced based on long sequencing reads (PacBio). The lack of high-quality genome has limited the study of Collembola evolution. In the present study, we generated a chromosome-level genome assembly of *Tomocerus qinae*, which had a size of 334.44 Mb forming five chromosomes. Our high-quality genome provides a valuable genomics resource for comparative genomic studies and a better understanding of the evolution of Collembola.

Introduction

Collembola (springtails) are basal hexapods renowned for their ancient origin. Fossil of the Devonian (ca. 400 million years ago) Collembola species *Rhyniella praecursor* is the oldest unequivocal records of hexapods. Being one of the most successful arthropod lineages, springtails are ubiquitous in terrestrial ecosystems and particularly important members of the soil communities, occurring primarily in humid habitats of forests and grasslands, and even colonizing deep soil layers (Handschin 1955; Orgiazzi 2016). Some collembolan species are sensitive to environmental factors, thus their appearance and life history traits are ideal indicators for soil ecotoxicology and environmental changes (Hopkin 1997). Despite its remarkable evolutionary and ecological significance, many aspects of Collembola have not been sufficiently understood. Notably, its systematic position and the relationship between its families are still controversial; the exact ecological roles of each groups in soil ecosystems have not

been determined; the functional divergence of different taxa is poorly known; the gene-level regulation of its metabolism, growth, reproduction, and body organization has not been addressed. This situation is highly attributed to the deficiency of high-quality genetic references data of Collembola. To date, a total of 39 genomes of Collembola have been available in the NCBI database (accessed January 10, 2022), including the genome of *Tomocerus qinae* (264.85 Mb) which was based on the Illumina sequencing (Sun et al. 2020). Among them, only seven species have been sequenced based on PacBio long reads.

To increase the knowledge on the evolution and ecology of Collembola, here we reported a collembolan species, *T. qinae*, belonging to the family Tomoceridae (Yu et al. 2016). According to recent molecular phylogeny, this family is probably the earliest derived collembolan lineage (Sun et al. 2020), thus may provide information about the ancestral states of collembolan genome. Besides, this family has distinctly larger body size (3–8 mm) than most other Collembola families (0.2–2 mm), thus comparing its genome to other collembolan genomes may help to understand the mechanism of body size regulation in this ancient arthropod lineage (Yu et al. 2021). In this study, we applied Oxford Nanopore (ONT) and Hi-C sequencing to obtain a chromosome-level genome assembly for *T. qinae*, and further analyzed gene annotation, gene family evolution, and significantly expanded gene families in this species.

Results and Discussion

Genome Assembly

In total, we generated 41.26 Gb (~123×) ONT reads, 39.65 Gb (~119×) Illumina short reads, 41.26 Gb (~123×) Hi-C data, and 9.67 Gb transcriptomic data. The mean and N50 length of the genomic ONT reads were 22.45 and 33.62 kb, respectively.

Based on the Illumina reads, the genome size of *T. qinae* was estimated to be 313.14–313.67 Mb through *k*-mer analysis ($k = 21$). The estimated heterozygosity and duplication were about 0.57% and 17.97%, respectively. Genome assembly was conducted using NextDenovo, which had a size of 370.38 Mb with 177 contigs. After polishing, removing heterozygous sequences, Hi-C scaffolding and contaminant detection, the final assembly length of *T. qinae* was 334.44 Mb, consisting of 115 contigs. The scaffold/contig N50 length was 71.91/13.94 Mb, and 96.16% of assembled sequences anchored on five chromosomes and GC content was 34.42% (supplementary fig. S1, Supplementary Material online). A total of 64,468 transcripts were assembled with N50 length of 4.88 kb based on the genome-guided strategy.

The mapping rates of the ONT and Illumina were 98.09% and 94.78%, respectively. BUSCO analysis found 96.80% (single-copied genes: 94.10%, duplicated genes: 2.70%), 0.80%, and 2.40% of the 1,013 expected genes, which were identified as complete, fragmented, and missing sequences, respectively. Compared with the *Sinella curviseta* (Zhang et al. 2019) and *Folsomia candida* (Faddeeva-Vakhrusheva et al. 2017) assembly, the genome of *T. qinae* had greater contig contiguity and completeness, indicating that the assembled genome in this study was of high quality (table 1).

Gene Annotation

A total of 26.11% (87.26 Mb) repetitive elements were identified in the genome of *T. qinae*. The top five abundant repeat categories were unclassified (15.92%), long terminal repeat (LTR) (3.88%), DNA elements (2.77%), long interspersed element (1.43%), and rolling circle (0.91%). The most abundant LTR and DNA transposons found in the *T. qinae* assembly were Pao and Maverick, accounting for 1.23% and 0.46%, respectively (supplementary table S1, Supplementary Material online).

We predicted 20,451 protein-coding genes (PSGs) of the *T. qinae* genome. The average exon number per gene was 7.23, with mean exon and intron lengths of 387.54 and 556.85 bp, respectively. According to the BUSCO assessment, 96% complete BUSCO genes were successfully identified in the 1,013 assessed genes. Compared with *S. curviseta* (90.8%) and *F. candida* (77.6%), gene predictions of *T. qinae* had higher BUSCO completeness (table 1), which indicated the high quality of the genome annotation in this study. Furthermore, a total of 16,537 (80.86%) of the predicted genes were functionally annotated in the UniProt database. There were 12,932, 9,558, 3,953, 6,037, 4,006, and 13,882 genes annotated with Gene Ontology (GO) terms, Kyoto Encyclopedia of Genes and Genomes (KEGG) KEGG Orthology (KO) terms, Enzyme Codes, KEGG pathways, Reactome pathways, and Clusters of Orthologous Gene categories, respectively. A total of 494 noncoding RNAs (ncRNAs) were predicted using Infernal and tRNAscan, with 119 rRNAs, 38 miRNAs, 61 small nuclear RNAs, 1 long ncRNA, 131 tRNAs (22 isotypes), 6 ribozymes, and 138 other ncRNAs (supplementary table S2, Supplementary Material online).

Gene Family Evolution

A total of 143,573 (88.8%) genes were clustered into 18,476 gene families (orthogroups) using OrthoFinder. The gene family clusters were divided into five categories with the following counts of orthogroups genes: 3,292 all species present; 1,110 universal single-copy; 312 unique to five Collembola species, and 18,078 other unassigned, respectively. In the *T. qinae* genome, 17,606 (86.1%) genes were clustered into 10,030 orthogroups, including 721 orthogroups that contained 3,286 species-specific genes.

Table 1
Genome assembly and annotation statistics of three Collembola species

	T. qinae	*S. curviseta*	*F. candida*
Genome assembly			
Assembly size (Mb)	334.44	381.46	221.70
Number of scaffolds/contigs	115/272	599/599	162/228
Longest scaffold/contig (Mb)	140.45/25.68	12.99/12.99	28.53/20.23
N50 scaffold/contig length (Mb)	71.85/13.94	3.28/3.28	6.52/4.89
GC (%)	34.42	37.51	37.52
Gaps (%)	0.05	0	0.11
BUSCO completeness (%)	96.8	95.3	97.0
Gene annotation			
Protein-coding genes	20,451	23,943	28,734
Mean protein length (aa)	518.16	524.60	461.54
Mean gene length (bp)	6,083.26	4,040.85	4,615.44
Exons per gene	7.23	5.59	6.89
Exon (%)	17.13	15.69	31.90
Mean exon length	387.54	446.95	357.02
Intron (%)	20.07	9.67	28.00
Mean intron length	556.85	364.58	414.36
BUSCO completeness (%)	96.0	90.8	77.6

The phylogenetic tree was structured using IQ-TREE based on 1,110 single-copy orthologous genes. In the tree obtained, we confirmed that Collembola and Insecta were found to be reciprocally monophyletic once more. Collembola was sister-group to Insecta (fig. 1*a*), this result is in agreement with that of Timmermans et al. (2008), who used ribosomal protein genes, and that of Faddeeva et al. (2015), who used transcriptomes to reconstruct the phylogeny of Pancrustacea.

A total of 2,189 (11.85%) and 4,825 (21.15%) gene families experienced expansions and contractions in *T. qinae*, respectively; among which 49 (47 expansions and 2 contractions) were significantly evolving gene families. The significant expanded families were involved in detoxification, digestion, growth, and taste (fig. 1*b*). Significantly expanded gene families suggest that advancing proliferation and growth abilities of *T. qinae* are probably responsible for its larger body size than other Collembola species. However, further research is needed to investigate the importance of these expanded gene families in the evolution of Collembola.

Materials and Methods

Sample Collection and Sequencing

The strains of *T. qinae* used for genome sequencing were collected from the Purple Mountain (China, Nanjing) in July of 2016, which were continuously cultured on peat-soil based microcosms for 5 years. In order to reduce heterozygosity, samples were bred for more than ten generations in our lab. A hundred, 30, 20, and 50 adult individuals were prepared for ONT, genome survey, transcriptome, and Hi-C sequencing, respectively.

Genomic DNA was extracted using Qiagen Blood and Cell Culture DNA Mini Kit for Illumina sequencing and using 1D DNA Ligation Sequencing kit SQK-LSK109 for Nanopore sequencing. Genomic RNA was extracted using (DP441) RNAprep Pure Plant Plus Kit with default handing protocols, and library was constructed by ONT PromethION platform. The paired-end libraries of a 350 bp-insert size were generated using Truseq DNA PCR-free kit based on the Illumina NovaSeq 6,000 platform. The library with an insert size of 30 kb was constructed using the ONT PromethION platform. Illumina sequencing was carried out by Berry Genomics (Beijing, China), ONT and transcriptome sequencing was performed by BENAGEN (Wuhan, China). The Hi-C library was prepared and sequenced by BGI MGISEQ-2000 with paired-end reads of 150 bp by Frasergen (Wuhan, China).

Genome Size Estimation and Genome Assembly

Raw Illumina data were quality filtered using BBTools suite v.38.67 (Bushnell 2014). Duplicated reads were removed using "clumpify.sh." The "bbduk.sh" was used to trim sequences with quality scores <20; filter out sequences with >5 Ns, and reads shorter than 10 bp; trim the poly-A/G/C tails longer than 10 bp; and correct overlapping paired reads. The frequency distribution of 21-mer was performed using GenomeScope v.1.0.0 with the maximum k-mer coverage of 10,000 (Vurture et al. 2017).

Raw ONT reads longer than 4.3 kb were assembled using Nextdenovo v.2.3.1 (https://github.com/Nextomics/NextDenovo). Preliminary assembly was polished with two rounds of Illumina short reads using NextPolish v.1.1.0 (Hu et al. 2020). Redundant heterozygous regions were

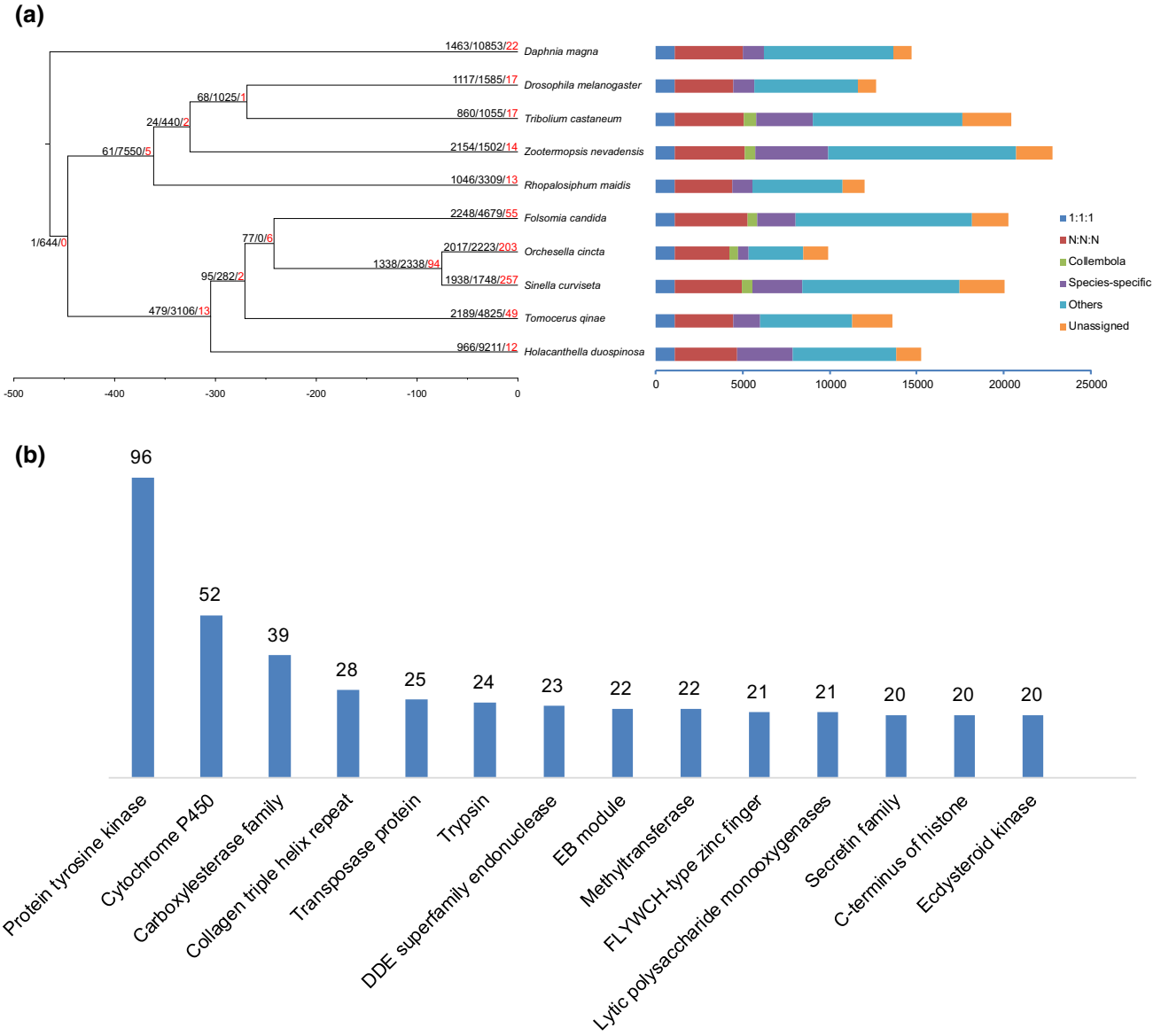

Fig. 1.—(*a*) Phylogenetic and gene family evolution analyses of *T. qinae* and another nine arthropod species. Node values indicate the number of gene families showing expansion, contraction, and rapid evolution. "1:1:1" represents shared single-copy genes, "N:N:N" as multicopy genes shared by all species, "Collembola" as shared orthologs unique to Collembola, "Others" as unclassified orthologs, "Unassigned" as orthologs which cannot be assigned into any gene families (orthogroups). (*b*) The bars show top 20 significantly expanded families.

removed using Purge Haplotigs v.1.1.0 (Roach et al. 2018) based on the cutoff set as 60 for identifying each contig as a haplotig ("-s 60"). During the redundancy removal and short-read polishing steps, Minimap2 v.2.17 (Li 2018) was used as the read mapper.

The Hi-C sequencing reads were aligned to the assembly using Juicer v.1.6.2 (Durand et al. 2016). The primary contigs were further assembled into chromosomes using the 3D-DNA v.180922 with default parameters (Dudchenko et al. 2017). We used Juicebox v.1.11.08 (Durand et al. 2016) to manually correct the errors of the assembly by visualizing the Hi-C heatmaps. We used BLAST+ (blastn) v.2.7.1 (Camacho et al. 2009) against the NCBI nucleotide and UniVec databases to detect potential contaminant sequences.

The quality assessment of the final genome assembly was performed using the following two methods. Firstly, ONT long reads and Illumina short reads were mapped to the assembly using Minimap2 and the mapping rate was calculated using SAMtools v.1.10 with "flagstat" parameter (Li et al. 2009). Secondly, the completeness of the assembly was evaluated using BUSCO v.3.0.2 (Waterhouse et al. 2018) based on the arthropoda_odb 10 database ($n = 1{,}013$).

Genome Annotation

We annotated the repetitive elements in the genome of *T. qinae* by homology searching and de novo predictions. RepeatModeler v.2.0.1 (Flynn et al. 2020) was used to construct a de novo repeat library, which combined Dfam 3.1 (Hubley et al. 2016) and RepBase-20181026 databases (Bao et al. 2015) to build the custom library. Finally, RepeatMasker v.4.0.9 (Smit et al. 2013–2015) was used to identify repetitive elements in the assembled genome based on the repeat library. Noncoding RNAs were predicted using Infernal v.1.1.2 (Nawrocki and Eddy 2013) and tRNAscan-SE v.2.0.6 (Chan and Lowe 2019), and tRNAs of high confidence were confirmed using tRNAscan-SE with "EukHighConfidenceFilter."

PSGs were predicted under three lines of evidence including ab initio, RNA-seq, and protein homology. Ab initio gene prediction and gene model training were performed with BRAKER v.2.1.5 pipeline (Hoff et al. 2016), which trained Augustus v.3.3.2 (Stanke et al. 2004) and GeneMark-ES/ET/EP 4.48_3.60_lic (Brůna et al. 2020) integrating evidence from OrthoDB10 v1 database (Kriventseva et al. 2019) and transcriptomic data. Then gene structure annotations in the genome were automatically generated by BRAKER. RNA-seq alignments were produced using HISAT2 v.2.2.0 (Kim et al. 2015). RNA-seq data were further assembled into transcripts with genome-guided assembler Stringtie v.2.1.3 (Kovaka et al. 2019). The protein sequence of *Drosophila melanogaster*, *Tribolium castaneum*, *Apis mellifera*, *Bombyx mori*, and *Daphnia magna* were downloaded from the NCBI database as protein homology evidence (supplementary table S3, Supplementary Material online). All the above evidence was integrated by MAKER v.3.01.03 (Holt and Yandell 2011) genome annotation pipeline.

We used Diamond v.0.9.24 (Buchfink et al. 2015) with the sensitive mode "--more-sensitive -e 1e-5" to generate gene functional annotations against the UniProtKB version 2020_01 database. Furthermore, we annotated protein domains, GO, and KEGG pathways using eggNOGmapper v.2.0.1 (Huerta-Cepas et al. 2017) based on the eggnog v.5.0 database (Huerta-Cepas et al. 2019) and InterProScan 5.41–78.0 (Finn et al. 2017) against six databases, including Pfam (El-Gebali et al. 2019), Panther (Mi et al. 2019), Gene3D (Lewis et al. 2018), Superfamily (Wilson et al. 2009), SMART (Letunic and Bork 2018), and Conserved Domain Database (Marchler-Bauer et al. 2017).

Gene Family Evolution

OrthoFinder v.2.3.8 (Emms and Kelly 2019) were used to infer orthologous from ten arthropod species, including one Crustea species (*D. magna*), one Diptera species (*D. melanogaster*), five Collembola species (*F. candida*, *Holacanthella duospinosa*, *S. curviseta*, *Orchesella cincta*, and *T. qinae*), one Hemiptera species (*Rhopalosiphum maidis*), one Coleoptera species (*T. castaneum*), and one Isoptera species (*Zootermopsis nevadensis*) (supplementary table S3, Supplementary Material online).

Protein sequences of single-copy orthologs were used to construct phylogenetic tree. MAFFT v.7.450 (Katoh and Standley 2013) was employed to align sequences in each single-copy ortholog with the L-INS-I strategy. We used trimAl v.1.4.1 (Capella-Gutierrez et al. 2009) with the "automated1" heuristic method to remove alignment gaps and unreliable regions. The aligned sequences were concatenated using FASconCAT-G v.1.04 (Kück and Longo 2014). IQ-TREE v.2.07 (Minh et al. 2020) was used to construct the tree with the partitioning strategy ("-m MFP --mset LG --msub nuclear --rclusterf 10 -B 1,000 --alrt 1,000").

MCMCTree of PAML v.4.9j (Yang 2007) was used to estimate divergence time. Four standard divergence time points from the paleobiodb database (https://paleobiodb.org/) were used to calibration: root (PanCrustacea $<$ 541 Ma), Collembola (407.6–410.8 Ma), Holometabola (314.6–318.1 Ma), and Entomobryoidea (272.3–279.3 Ma).

The expansion or contraction of gene families across the phylogenetic tree was calculated using CAFÉ v.4.2.1 (Han et al. 2013) with the single birth–death parameter lambda (λ) and the significance level of 0.01.

Supplementary Material

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

Acknowledgments

This work was supported by Zhejiang Provincial Natural Science Foundation of China (LTY20C030001) and the National Natural Science Foundation of China (36101880 and 41971063).

Data Availability

Genome assembly (JAECZT000000000) and raw sequencing data (SRR13342218–SRR13342221) of *Tomocerus qinae* have been deposited in NCBI, under the Bioproject PRJNA682767. Genome annotations are available at Figshare and can be accessed at https://figshare.com/projects/Chromosomal-level_genome_assembly_of_the_springtail_Tomocerus_qinae_Collembola_Tomoceridae_/126584.

Literature Cited

Bao W, Kojima KK, Kohany O. 2015. Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA 6(11):1–6.

Brůna T, Lomsadze A, Borodovsky M. 2020. GeneMark-EP^{+}: eukaryotic gene prediction with self-training in the space of genes and proteins. NAR Genom Bioinform. 2(2):lqaa026.

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

Bushnell B. 2014. BBtools. Available from: https://sourceforge.net/projects/bbmap/.
Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinform. 10(1):421.
Capella-Gutierrez S, Silla-Martınez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25(15):1972–1973.
Chan PP, Lowe TM. 2019. tRNAscan-SE: searching for tRNA genes in genomic sequences. Methods Mol Biol. 1962:1–14.
Dudchenko O, et al. 2017. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science 356(6333):92–95.
Durand NC, et al. 2016. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 3(1):95–98.
EI-Gebali S, et al. 2019. The Pfam protein families database in 2019. Nucleic Acids Res. 47:D427–D432.
Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20(1):238.
Faddeeva-Vakhrusheva A, et al. 2017. Coping with living in the soil: the genome of the parthenogenetic springtail *Folsomia candida*. BMC Genom. 18(1):493.
Faddeeva A, et al. 2015. Collembolan transcriptomes highlight molecular evolution of hexapods and provide clues on the adaptation to terrestrial life. PLoS One 10(6):e0130600.
Finn RD, et al. 2017. InterPro in 2017—beyond protein family and domain annotations. Nucleic Acids Res. 45(D1):D190–D199.
Flynn JM, et al. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci USA 117(17):9451–9457.
Han MV, Thomas G, Lugo-Martinez J, Hah 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.
Handschin E. 1955. Considérations sur la position systématique des Collemboles. Mémoires de la Société Royale d'Entomologie de Belgique, Tome Vingt-Septième, Volume Jublaire, p. 40-53.
Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS: Table 1. 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 Bioinform. 12(1):491.
Hopkin SP. 1997. Biology of the springtails. Oxford: Oxford University Press.
Hu J, Fan J, Sun ZY, Liu SL, Berger B. 2020. NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics 36(7):2253–2255.
Hubley R, et al. 2016. The Dfam database of repetitive DNA families. Nucleic Acids Res. 44(D1):D81–D89.
Huerta-Cepas J, et al. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol Biol Evol. 34(8):2115–2122.
Huerta-Cepas J, et al. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 47(D1):D309–D314.
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.
Kovaka S, et al. 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 20(1):278.
Kriventseva EV, et al. 2019. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res. 47(D1): D807–D811.
Kück P, Longo GC. 2014. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. Front Zool. 11(1):81.
Letunic I, Bork P. 2018. 20 years of the SMART protein domain annotation resource. Nucleic Acids Res. 46(D1):D493–D496.
Lewis T, et al. 2018. Gene3D: extensive prediction of globular domains in proteins. Nucleic Acids Res. 46:D435–D439.
Li H, et al. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25(16):2078–2079.
Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34(18):3094–3100.
Marchler-Bauer A, et al. 2017. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. 45:D200–D203.
Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. 2019. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. 47(D1):D419–D426.
Minh BQ, et al. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 37(5): 1530–1534.
Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29(22):2933–2935.
Orgiazzi A. 2016. Global soil biodiversity atlas. Europaische Kommission, Joint Research Centre.
Roach MJ, Schmidt SA, Borneman AR. 2018. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinform. 19:460.
Smit AFA, Hubley R, Green P. 2013–2015. RepeatMasker Open-4.0. 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(Suppl. 2):W309–W312.
Sun X, et al. 2020. Phylomitogenomic analyses on collembolan higher taxa with enhanced taxon sampling and discussion on method selection. PLoS One 5(4):e0230827.
Timmermans M, Roelofs D, Mariën J, van Straalen NM. 2008. Revealing pancrustacean relationships: phylogenetic analysis of ribosomal protein genes places Collembola (springtails) in a monophyletic Hexapoda and reinforces the discrepancy between mitochondrial and nuclear DNA markers. J Pain Symptom Manage. 8:83.
Vurture GW, et al. 2017. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics 33(14):2202–2204.
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.
Yang ZH. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24(8):1586–1591.
Yu DY, et al. 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 DY, Yao J, Hu F. 2016. Two new species of Tomocerus ocreatus complex (Collembola, Tomoceridae) from Nanjing, China. Zootaxa 4084(1):125–134.
Zhang F, et al. 2019. A high-quality draft genome assembly of *Sinella curviseta*: a soil model organism (Collembola). Genome Biol Evol. 11(2):521–530.

Associate editor: Christopher Wheat