A Draft Genome Assembly of the Antarctic Springtail *Cryptopygus antarcticus* and Diversity of Glycoside Hydrolase Genes

Donggu Jeon [1], Chi-une Song [1], Hyeongwoo Choi [2], Junsang Youn [1], Hyungtaek Jung [3], Youn-Ho Lee [4], Sung-Hun Lee [5,*], Seong-il Eyun [1,*]

[1]Department of Life Science, Chung-Ang University, Seoul 06974, Korea

[2]Department of Earth Systems and Environmental Sciences, Chonnam National University, Gwangju 61186, Korea

[3]National Centre for Indigenous Genomics, John Curtin School of Medical Research, The Australian National University, Acton, ACT 2601, Australia

[4]Marine Environmental Research Center, Korea Institute of Ocean Science and Technology, Busan 49111, Korea

[5]Department of Fishery, Marine, Industry, Tourism, and Leisure, Chonnam National University, Yeosu 59626, Korea

*Corresponding authors: Emails: wahun@jnu.ac.kr; eyun@cau.ac.kr.

Accepted: November 19, 2025

Abstract

The Antarctic springtail *Cryptopygus antarcticus* Willem (family Isotomidae) is a representative arthropod species of the maritime Antarctic environment that could be used as an important organism for further study of animal evolution and adaptation. Despite the biological and ecological peculiarities that distinguish them from the majority of collembolan species, our understanding of the genetic background behind their success in an extreme habitat remains unclear. We present the first high-quality draft genome of *C. antarcticus* assembled from in-depth whole genome and transcriptome sequencing data obtained from both long-read and short-read sequencing platforms. The genome was 103.6 Mb in size with 81 scaffolds and a scaffold N50 of 3.4 Mb, appearing to have a smaller genome than that of hitherto known collembolan genomes. Following protein-coding gene prediction and annotation analyses, 19,808 non-redundant genes were identified, representing 97.0% Benchmarking Universal Single-Copy Ortholog (BUSCO) gene coverage. Subsequent Gene Ontology (GO) functional enrichment analyses revealed that significantly expanded gene families were mainly associated with cell cycle regulation and changes in cell states or activities, while contracted gene families were related to the inhibition of germ cell proliferation. Several glycoside hydrolase family genes were identified in *C. antarcticus*, some of which may have evolved to facilitate their survival in the extreme environment. These findings suggest that the evolution of these gene families is related to their adaptation to the habitat's extreme conditions.

Key words: adaptation, Antarctica, extreme environment, gene family evolution, glycoside hydrolase.

Significance

A high-quality draft genome of *Cryptopygus antarcticus* was assembled using comprehensive whole genome and transcriptome sequencing data. The assembled genome was found to be unusually small compared to other known Collembola species. Despite its small size, the overall gene content appears similar to that of other Collembola, with the reduction likely due to the loss of non-genic elements. Further analysis of gene family expansion and contraction, as well as the diversity of glycoside hydrolase genes, suggests that the evolutionary patterns of these genes reflect adaptations to extreme cold and arid conditions. This genome offers valuable opportunities for ongoing studies on animal adaptation and evolution.

Introduction

Springtails (Collembola) are diminutive terrestrial non-insect hexapods, representing one of the most successful arthropod groups in their immense abundance and diversity. They are known from various geographical regions covering a wide range of climatic regimes from arid deserts to snow-covered areas (Bellinger et al. 1996–2004; Bellini et al. 2023) and can even be encountered on the surface of aquatic habitats, although they are not really aquatic animals (Christiansen 1964; Deharveng et al. 2008; Olejniczak et al. 2021). In these diverse ecological niches, they play pivotal ecological roles by participating in organic matter processing, nutrient (re)cycling, regulation of microbial communities, and soil respiration (Block and Tilbrook 1975; Rusek 1998). Therefore, the recognition of Collembola as bioindicators for evaluating soil quality and its overall environmental health has been growing.

The Antarctic springtail *Cryptopygus antarcticus* is the most common and widespread arthropod in the maritime Antarctic region (Tilbrook 1970) and is an important species for the study of animal adaptations to extreme environments. As such, intensive investigation into their genetics is essential to shed light on their biological, physiological, and ecological significance. In this study, using long-read and short-read sequencing, we provide the first high-quality draft genome of *C. antarcticus* from King George Island.

The *C. antarcticus* genome was assembled into a smaller genome (~104 Mb). The analyses on the expanded and contracted gene families and survey on several glycoside hydrolase (GH) gene families implied that the gene evolution may be related with their adaptation to the extreme habitat conditions. The whole genome and transcriptome data acquired by this study will shed light on the biology of Collembola.

Results and Discussion

General Characteristics of the *C. antarcticus* Genome

We generated 201.9 Gb of whole genome sequencing (WGS) data, including 89.8 Gb of HiFi long-reads and 112.2 Gb of short-reads, and 15.6 Gb of whole transcriptome mRNA sequencing (WTS) short reads (Table S1). The final assembly yielded a genome size of 103.62 Mb with 81 scaffolds, which is consistent with estimates obtained using JellyFish (101 to 104 Mb; Fig. S1). The scaffold N50 value for the genome assembly was 3.37 Mb, and the largest and the shortest scaffold lengths were 11.52 Mb and 43.30 kb, respectively (Table S2). A combined 7.77% of the genome of *C. antarcticus* was identified as being repetitive content (Table S3). The HiFi read mapping rate to the final genome assembly was 99.67%. The Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis for the assessment of the genome assembly quality produced a completeness score of 96.5%, including 93.1% complete single-copy genes and 3.4% complete duplicated genes as well as 0.7% fragmented and 2.8% missing genes.

We predicted protein-coding genes (PCGs) by integrating evidence from de novo, transcriptome, and protein-homology analyses. This produced and initial prediction of 24,363 PCGs. After removing isoforms, we obtained 19,808 non-redundant PCGs, which corresponded to 93.3% complete single-copy, 3.7% complete duplicated, 0.5% fragmented, and 2.5% missing BUSCOs (Table S4). Using public databases, we functionally annotated 88.58% of the gene models (Table S5).

We also assembled a transcriptome using the WTS data. After removal of redundant sequences, a BUSCO analysis revealed a completeness score of 95.8%, with 92.0% complete single-copy genes and 3.8% complete duplicated genes, and 1.8% fragmented and 2.4% missing BUSCOs (Table S4). The final BUSCO assessment for the transcriptome assembly (95.8%) showed relative lower values to that of the genome (96.5%), which may be caused by the stringent criteria applied for protein prediction (Table S4).

Cryptopygus antarcticus appears to have a significantly smaller genome compared to its collembolan relatives (~104 Mb vs. >220 Mb; Table 1) (Faddeeva-Vakhrusheva et al. 2016; Faddeeva-Vakhrusheva et al. 2017; Wu et al. 2017; Zhang et al. 2019; Pan et al. 2022; Jin et al. 2023). The BUSCO analysis, the mapping ratio of HiFi long reads to the genome, and genome size estimates based on different datasets all indicated that our assembly is complete, accurate, and of high quality. Notably, the high BUSCO gene coverage value and the generally conserved gene content imply that *C. antarcticus* has maintained the genetic information for encoding the essential core genes, albeit its small size.

Our findings suggest that the *C. antarcticus* has evolved a streamlined genome by reducing non-coding DNA regions. Repetitive elements in *C. antarcticus* make up only 7.8% (8,049,673 bp) of the genome, which is an unusually low portion compared to other Collembola species (9.8% to 43%; Table 1). Additional genome size reduction was also prominent in intronic regions (Table 1). The average intron number per gene in *C. antarcticus* was 4.9, whereas the values in other species ranged from 16.0 to 24.5—approximately four times higher. Interestingly, despite the lower intron counts, the number of exons per gene (5.9) in *C. antarcticus* remained comparable to that of other Collembola, which ranges from 5.9 to 8.0 per gene. These findings further explain a notably higher gene density within the compact genome of *C. antarcticus*. Similar patterns of adaptation have been reported previously. Kelley et al. (2014) observed a similarly small genome (~99 Mb) in the Antarctic midge *Belgica antarctica* (Diptera: Chironomidae). Like in our case, the smaller genome size of *B. antarctica* mainly results from a reduction in introns and repetitive elements, while maintaining a similar number of genes compared to other dipteran species.

Table 1 Statistical comparison of the genomes among the six Collembola species

Species	*C. antarcticus*	*F. candida*	*O. cincta*	*S. curviseta*	*T. qinae*	*H. duospinosa*
Assembly size (Mb)	103.62	221.70	286.77	381.46	334.44	327.57
No. of scaffolds	81	162	9,402	599	115	62,430
Scaffold N50	3.37 Mb	20.1 kb	65.9 kb	3.28 Mb	71.85 Mb	310.2 kb
GC%	35.2	37.5	36.8	37.5	34.4	33.5
Complete BUSCO%	96.5	84.0	96.9	95.3	96.8	95.3
Genes (no./Mb)[a]	19,808/52.65	28,734/132.62	20,249/60.56	23,943/96.75	20,451/-	9,911/56.66
Gene mean length (bp)[a]	2,658	4,615	2,991	4,041	6,083	5,717
Gene density (per Mb)	191.20	129.61	70.60	62.77	61.16	30.21
Exons (no./Mb)[a]	116,006/28.40	197,859/70.64	118,474/32.23	133,951/59.87	-/57.19	79,659/22.71
Exons per gene[a]	5.90	6.89	5.90	5.59	7.23	8.04
Exon mean length (bp)[a]	244	357	272	447	388	285
Introns (no./Mb)[a]	96,198/23.00	524,921/61.98	336,337/28.33	381,850/36.88	-/67.22	242,640/33.95
Introns per gene[a]	4.86	18.27	16.61	16.00	-	24.48
Intron mean length (bp)[a]	239	118	84	97	557	140
Repetitive elements (%)	7.8	23.3	15.0	9.8	26.1	43.0
Reference	This study	Faddeeva-Vakhrusheva et al. (2017)	Faddeeva-Vakhrusheva et al. (2016); Zhang et al. (2019)	Zhang et al. (2019)	Pan et al. (2022)	Wu et al. (2017); Zhang et al. (2019)

[a]Statistical values for *C. antarcticus* in the table are calculated based on the non-redundant genome dataset. The corresponding values including isoforms are as follows: number of genes (i.e. mRNAs), 24,363 (79.92 Mb); mean mRNA length, 3,280 bp; total number of exons in mRNAs, 161,103 (38.16 Mb); exons per mRNA, 6.61; mean exon length, 236 bp; total number of introns in mRNAs, 136,740 (41.76 Mb); introns per mRNA, 5.61; mean intron length, 305 bp. "-": data not available.

Gene Family Expansion and Contraction Analysis

We discovered an overall 430 expanded and 358 contracted gene families in the *C. antarcticus* genome. Among these, 60 and 34 gene families exhibited significant expansions and contractions, respectively ($P < 0.05$) (Fig. 1a). The significantly expanded gene families contain the GO terms related to cell cycle regulation, changes in cell states or activities (GO:0060544, GO:0060546, GO:0062098, GO:0062099, GO:1902230), and to behavioral response to starvation and cold acclimation (GO:0042595, GO:0009631) (Fig. 1b). The contracted gene families are associated with terms related to inhibition of germ cell proliferation (GO:2000254, GO:2000255, GO:1905937, GO:0002176) (Fig. 1c). The rapid evolution of the aforementioned gene families has likely been beneficial for their survival in the polar habitats. Due to the Antarctic environment limiting *C. antarcticus* to only a short period of biological activity, the genes within the expanded gene families may enable rapid reactivation of their physiological processes. Conversely, the reduction in the gene families involved in the suppression of gamete proliferation may improve reproductive efficiency during the short breeding season. Further studies are required to clarify the importance of these gene families in the context of Collembola evolution.

GH Family Genes in *C. antarcticus*

We initially analyzed six enzymes categorized into five GH gene families: endo-β-1,4-glucanase (EC 3.2.1.4) of GH45 and GH9; glucan 1,3-β-glucosidase (EC 3.2.1.58) and mannan endo-1,4-β-mannosidase (EC 3.2.1.78) of GH5; polygalacturonase (EC 3.2.1.15) of GH28; and α-N-acetylgalactosaminidase (EC 3.2.1.49) and α-galactosidase (EC 3.2.1.22) of GH27 (Table S6). Phylogenetic analyses revealed that two enzymes, endo-β-1,4-glucanase of GH45 and mannan endo-1,4-β-mannosidase of GH5, may have undergone distinct evolutionary trajectories.

In GH45, the collembolan gene subtree containing two putative *C. antarcticus* genes (g1646.t1 and g14392.t1) was split into two by the insertion of the Rotifera gene subtree (Fig. 2a). Consequently, g14392.t1 appeared to be separated from other collembolan GH45 genes, which include another novel gene, g1646.t1. Interestingly, g14392.t1 shows high sequence similarity to previously reported sequences ACV50414.1 and ACV50415.1, both known to exhibit unique cold-active and thermo-tolerant properties (Song et al. 2017).

Similarly, among four putative GH5 subfamily 10 (GH5_10) genes encoding mannan endo-1,4-β-mannosidase, two candidates (g6007.t1 and g14415.t1) exhibited unusual phylogenetic positions. Although all collembolan GH5_10 genes formed a lineage-specific clade related to those from other metazoans (Insecta and Mollusca), these two genes showed a closer affinity to those from the relatively distant collembolan species *Tomocerus qinae* (family Tomoceridae) (Fig. 2b). In contrast, the other two GH5_10 genes (g12285.t1 and g12801.t1) are more similar to their counterparts from the more phylogenetically related species *Folsomia candida* (Isotomidae, the same family as *C. antarcticus*). Notably, although the "unusual" positions of the two GH5_10 genes may represent only subtle deviations compared to the

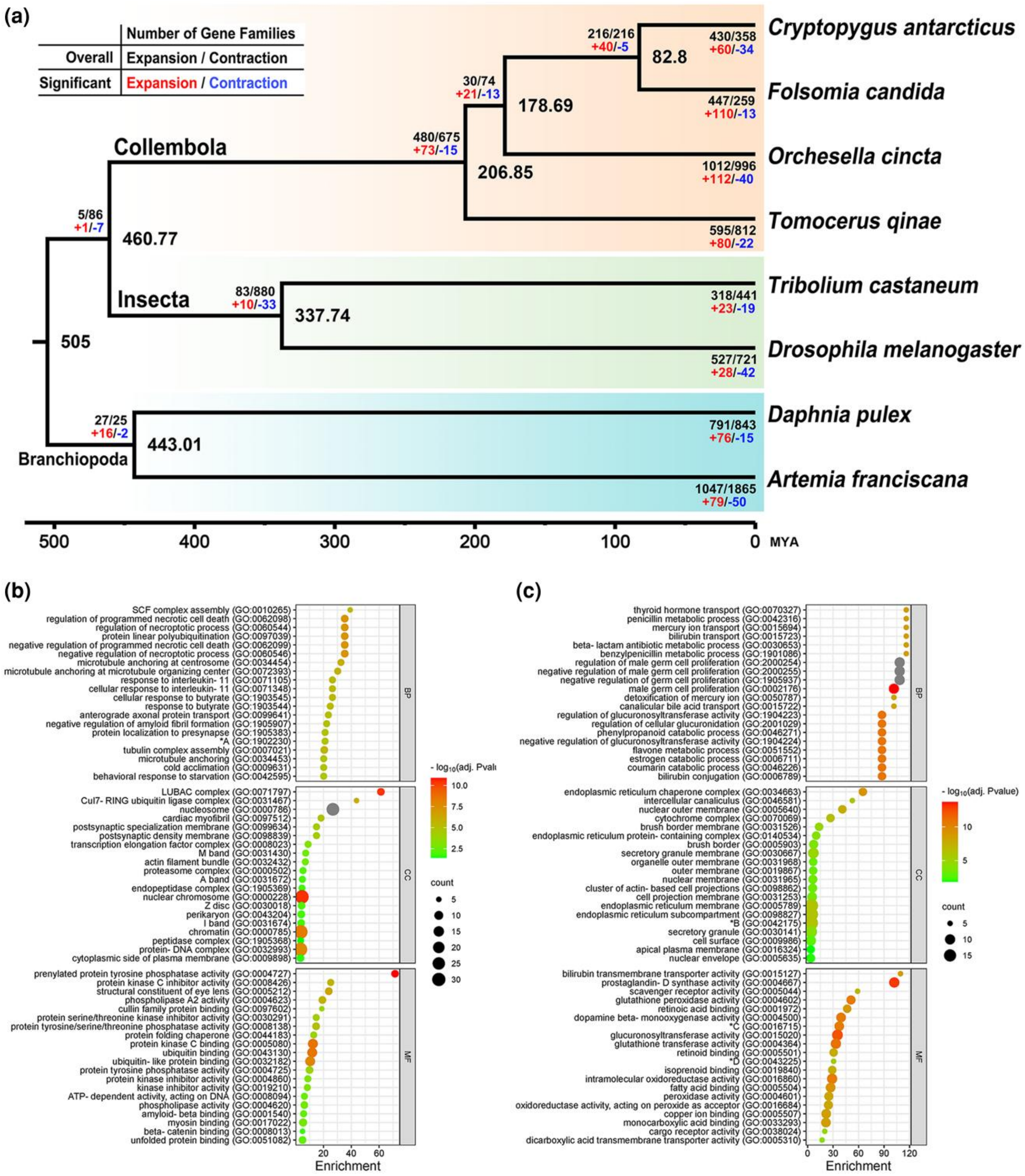

Fig. 1. Gene family expansion and contraction analysis. a) A phylogenetic tree showing gene family expansion and contraction in each evolutionary branch. The overall numbers of expanded and contracted gene families (shown in black) are displayed near the branches along with those of significantly expanded (red) and contracted (blue) families. Estimated divergence times are presented at each branching point in millions of years ago (MYA). GO function enrichment in the significantly expanded b) and significantly contracted c) gene families. Only the top 20 terms for each category, molecular function, cellular component, and biological process, are shown. GO functions that are too long to fit in legends in b) and c) are marked with *A to *D: *A, "negative regulation of intrinsic apoptotic signaling pathway in response to DNA damage"; *B, "nuclear outer membrane-endoplasmic reticulum membrane network"; *C, "oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced ascorbate as one donor, and incorporation of one atom of oxygen"; and *D, "ATPase-coupled inorganic anion transmembrane transporter activity."

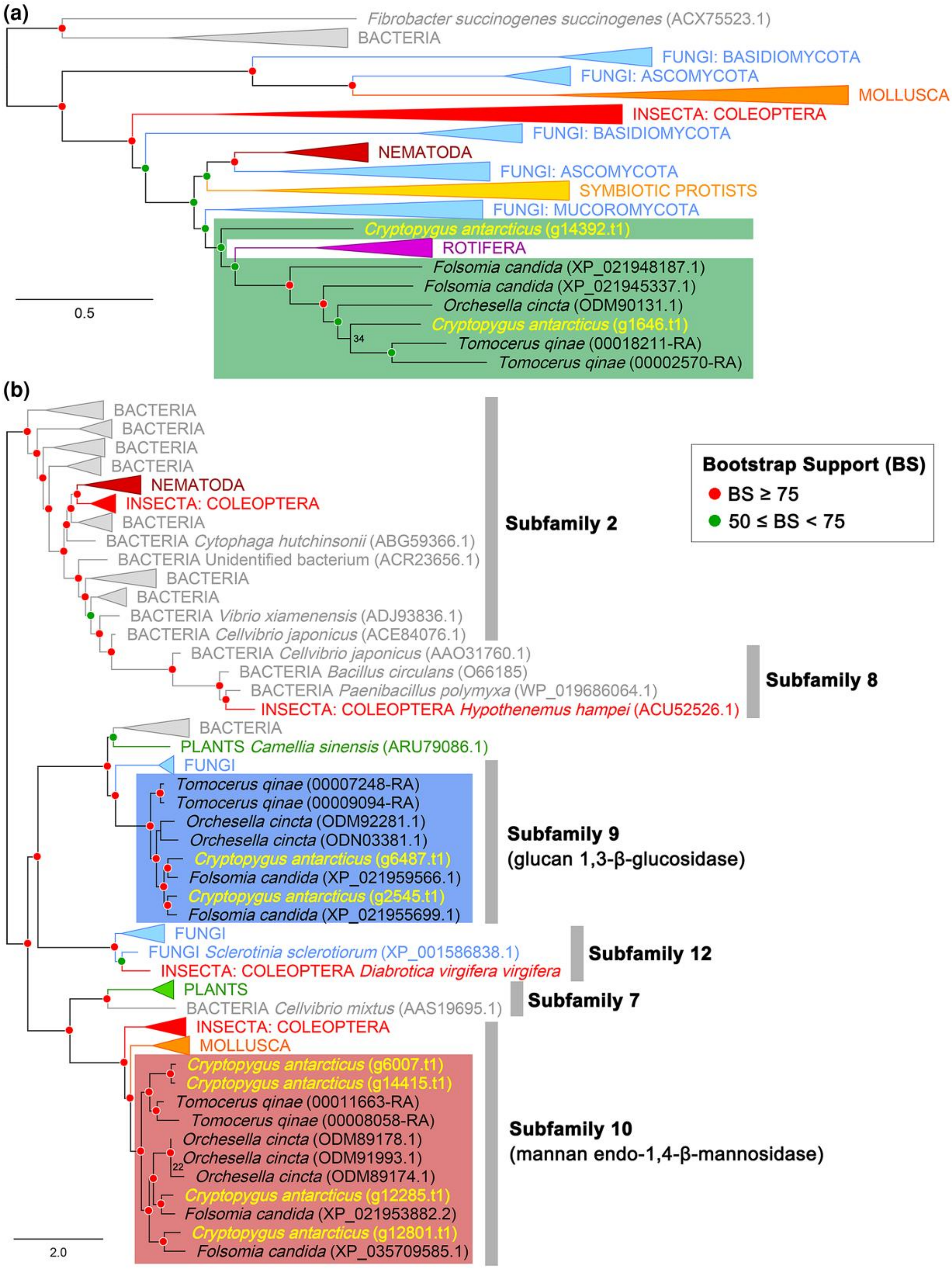
(a)
Fibrobacter succinogenes succinogenes (ACX75523.1)
BACTERIA
FUNGI: BASIDIOMYCOTA
FUNGI: ASCOMYCOTA
MOLLUSCA
INSECTA: COLEOPTERA
FUNGI: BASIDIOMYCOTA
NEMATODA
FUNGI: ASCOMYCOTA
SYMBIOTIC PROTISTS
FUNGI: MUCOROMYCOTA
Cryptopygus antarcticus (g14392.t1)
ROTIFERA
Folsomia candida (XP_021948187.1)
Folsomia candida (XP_021945337.1)
Orchesella cincta (ODM90131.1)
Cryptopygus antarcticus (g1646.t1)
34
Tomocerus qinae (00018211-RA)
Tomocerus qinae (00002570-RA)
0.5
(b)
BACTERIA
BACTERIA
BACTERIA
BACTERIA
NEMATODA
INSECTA: COLEOPTERA
BACTERIA
BACTERIA Cytophaga hutchinsonii (ABG59366.1)
BACTERIA Unidentified bacterium (ACR23656.1)
BACTERIA
BACTERIA
BACTERIA Vibrio xiamenensis (ADJ93836.1)
BACTERIA Cellvibrio japonicus (ACE84076.1)
Subfamily 2
Bootstrap Support (BS)
● BS ≥ 75
● 50 ≤ BS < 75
BACTERIA Cellvibrio japonicus (AAO31760.1)
BACTERIA Bacillus circulans (O66185)
BACTERIA Paenibacillus polymyxa (WP_019686064.1)
INSECTA: COLEOPTERA Hypothenemus hampei (ACU52526.1)
Subfamily 8
BACTERIA
PLANTS Camellia sinensis (ARU79086.1)
FUNGI
Tomocerus qinae (00007248-RA)
Tomocerus qinae (00009094-RA)
Orchesella cincta (ODM92281.1)
Orchesella cincta (ODN03381.1)
Cryptopygus antarcticus (g6487.t1)
Folsomia candida (XP_021959566.1)
Cryptopygus antarcticus (g2545.t1)
Folsomia candida (XP_021955699.1)
Subfamily 9
(glucan 1,3-β-glucosidase)
FUNGI
FUNGI Sclerotinia sclerotiorum (XP_001586838.1)
INSECTA: COLEOPTERA Diabrotica virgifera virgifera
Subfamily 12
PLANTS
BACTERIA Cellvibrio mixtus (AAS19695.1)
Subfamily 7
INSECTA: COLEOPTERA
MOLLUSCA
Cryptopygus antarcticus (g6007.t1)
Cryptopygus antarcticus (g14415.t1)
Tomocerus qinae (00011663-RA)
Tomocerus qinae (00008058-RA)
Orchesella cincta (ODM89178.1)
22
Orchesella cincta (ODM91993.1)
Orchesella cincta (ODM89174.1)
Cryptopygus antarcticus (g12285.t1)
Folsomia candida (XP_021953882.2)
Cryptopygus antarcticus (g12801.t1)
Folsomia candida (XP_035709585.1)
Subfamily 10
(mannan endo-1,4-β-mannosidase)
2.0

Fig. 2. Maximum-likelihood phylogenies of a) the endo-β-1,4-glucanase of the GH45 gene family and b) the glucan 1,3-β-glucosidase (subfamily 9) and mannan endo-1,4-β-mannosidase (subfamily 10) of the GH5 gene family. Genes from *C. antarcticus* are indicated in yellow font. Bootstrap support values are represented by red or green dots at nodes according to their value ranges; values outside these ranges are displayed numerically next to the corresponding nodes. Collembolan groups are highlighted in green, blue, and red shadings by the respective genes.

GH45 case, the phylogenetic relationships of other GH genes, including GH5 subfamily 9 (GH5_9, glucan 1,3-β-glucosidase), were generally consistent with their species-specific evolutionary patterns (Fig. 2b; Figs. S2 to S4). In addition to the phylogenetic implications, the high sequence similarity (especially for g14415.t1) to the previously reported β-1,4-mannanase sequence of *C. antarcticus* (ABV68808.1), which is also known to exhibit cold-active properties (Song et al. 2008), may further support its potential enzymatic function under cold conditions.

Both GH45 and GH5_10 gene families in this study contained two types of genes: phylogenetically conserved genes (g1646.t1 for GH45; g12285.t1 and g12801.t1 for GH5_10) and more divergent genes (g14392.t1 for GH45; g6007.t1 and g14415.t1 for GH5_10). This may suggest that possessing both general and specialized genes enhances not only the organism's survival in extreme environments but also its ability to cope with severe seasonal fluctuations such as drastic changes in temperature, water availability, and food resources.

Hypothetical Insights into Small Genome Evolution

The Mutational Hazard Hypothesis posits that the effective population size (N_e) reflects the intensity of random genetic drift and that lineages with high N_e tend to have smaller genomes due to more efficient selection against the accumulation of slightly deleterious non-genic elements such as transposable elements (TEs) (Lynch and Conery 2003; Lynch 2010). In this context, the present compact genome of *C. antarcticus* may have originated from its initially large N_e. The Antarctic springtail is numerically dominant throughout the maritime Antarctic, with population densities estimated between 2.5×10^5 and 1.5×10^6 individuals·m^{-2} (Block 1982; Convey and Smith 1997; Hayward et al. 2004). In comparison, another Antarctic Collembola, *Friesea grisea*, has densities between 20 and 4.6×10^4 individuals·m^{-2} (Block 1982; Convey and Smith 1997; Hayward et al. 2004). Thus, although we do not know the historical demographic changes in the N_e of *C. antarcticus*, its overall population size may be sufficiently large to drive genome size reduction.

More recent studies, however, have reported no statistical support for a relationship between increased genetic drift (expected to be stronger under low N_e) and TE expansions (Bast et al. 2016; Yang et al. 2024; Marino et al. 2025). In light of these findings, no universal explanation for genome size evolution has yet been established. In this regard, we suggest an alternative mechanism driven by "extrinsic environmental selection" which might be equally or even more important in shaping the *C. antarcticus* genome. Prolonged cold and dry Antarctic seasons force organisms into extended inactive states, thereby limiting reproduction and resulting in low genetic exchange within populations. These conditions may further act as a strong selective constraint, favoring individuals with greater fitness to withstand the environmental stresses. Subsequent reproduction among these "survivors" during the short summer period, after the "screening process" (see Block 1982 for the annual fluctuation in the population size), may lead to a rapid fixation of strongly selected mutations (e.g. the GH genes and the expanded/contracted gene families in this study) that confer higher tolerance and resistance to extreme conditions. Such annually repeated processes may effectively amplify the dominance of strong selection. In addition, the potential (dis)advantages and efficiency of maintaining relatively large genomes under extremely low temperatures and water scarcity represent another issue to consider, though this lies beyond the scope of our present study.

Admittedly, our alternative hypothesis remains largely theoretical and speculative. Nevertheless, providing genomic information for *C. antarcticus* offers a critical starting point for investigating adaptation and evolution in polar organisms. Continued focus on these animals will thus significantly enhance our understanding of interactions between organisms and their environments as well.

Materials and Methods

Sample Collection and Sequencing

Approximately 4,000 Antarctic springtails, collected near King Sejong Station (King George Island) in December 2022 and 2023, were used for NGS analysis using a PacBio Revio system for HiFi long reads and an Illumina NovaSeq 6000 platform for short-read WGS and WTS data.

Data Filtering and Assembly

Illumina reads were filtered with fastp v0.23.4 (Chen et al. 2018), and potential contaminants were removed with KRAKEN2 v2.1.3 (Wood et al. 2019). HiFi reads were de novo assembled with NextDenovo v2.5.2 (Hu et al. 2024), followed by polishing (NextPolish v1.4.1; Hu et al. 2020), haplotig removal (Purge_haplotigs v1.1.2; Roach et al. 2018), scaffolding (P_RNA_scaffolder; Zhu et al. 2018), and gap closing (LR_Gapcloser; Xu et al. 2019). A second round of the polishing steps was conducted. Assembly quality was assessed with BUSCO v5.6.1 (Manni et al. 2021) using the arthropoda_odb10 dataset ($n = 1{,}013$). HiFi read mapping was performed with minimap2 v2.26 (Li 2018) and Samtools v1.9 (Danecek et al. 2021). Transcriptome assembly was conducted from cleaned WTS data using SPAdes v3.15.5 (Bushmanova et al. 2019).

Genome Size Estimation

Genome size was estimated using Jellyfish v2.3.0 (Marçais and Kingsford 2011), and the *k*-mer histograms were inspected via GenomeScope2.0 (Ranallo-Benavidez et al. 2020).

Gene Prediction and Annotation

Repetitive elements were identified de novo using RepeatModeler v2.0.5 (Smit and Hubley 2008–2015) and masked by RepeatMasker v4.1.5 (Smit and Hubley 2013–2015). Gene prediction was performed using BRAKER3 v3.0.6 (Gabriel et al. 2024), incorporating RNA-seq data and reference proteins from Arthropoda and three other Collembola (Supplementary Material). Predicted PCGs were annotated using Diamond BLASTp v2.1.9.163 (Buchfink et al. 2021) search against the SWISS-PROT (2024 January 22; Bairoch and Apweiler 2000) and NCBI non-redundant databases (2024 January 10; https://www.ncbi.nlm.nih.gov/protein). Functional annotation was carried out with eggNOG-mapper v2.1.12 (Cantalapiedra et al. 2021) based on GO, KEGG, KOG, and Pfam databases.

Gene Family Evolution

Orthologs were identified among eight arthropods, including four Collembola, two Insecta, and two Branchiopoda, using OrthoFinder v2.5.5 (Emms and Kelly 2019). The results of OrthoFinder run can be found in Supplementary Material. A species phylogenetic tree based on single-copy genes was reconstructed using FastTree v2.1.11 (Price et al. 2010) by applying the ultrasensitive Diamond BLAST search strategy and maximum-likelihood tree inference options. The phylogeny tree was converted into an ultrametric tree topology with a root age of 505 MYA, representing the split of Branchiopoda and Hexapoda (Kumar et al. 2022). Gene family expansions and contractions were inferred using CAFÉ5 v5.1.0 (Mendes et al. 2021) with a single birth-death parameter lambda (λ). Functional enrichment of expanded and contracted gene families was analyzed with TBtools-II v2.102 (Chen et al. 2023), incorporating the GO ontology database (2024 June 20; https://geneontology.org/docs/download-ontology).

Identification of GH Family Genes

The GH gene sequences from various organisms were retrieved from GenBank using accession numbers listed in the CAZy database (2024 September 9; https://www.cazy.org/Glycoside-Hydrolases) and previous studies (Watanabe and Tokuda 2010; Eyun et al. 2014). These sequences were used as queries in BLAST searches against the predicted protein sequences of *C. antarcticus* obtained from BRAKER analysis (*e*-value < 1e-10). Additional protein data were sourced from the transcriptome assembly, and coding regions were identified using TransDecoder v5.7.1 (https://github.com/TransDecoder/TransDecoder). To reduce redundancy, CD-HIT v4.8.1 (Li and Godzik 2006) was employed with a 95% sequence similarity threshold. The final transcriptome-derived proteins were used to supplement any missing genes from the BRAKER proteins. The same BLAST search was repeated for the protein data from three other Collembola.

Multiple Sequence Alignment and Phylogenetic Analysis

Protein sequences of GH family genes were aligned using MAFFT v7.520 (Katoh and Standley 2013) with L-INS-i algorithm. Phylogenetic trees were generated using the maximum-likelihood method in RAxML-NG v1.2.2 (Kozlov et al. 2019) under best-fit substitution model inferred by IQ-TREE2 v2.3.6 (Minh et al. 2020) (Table S7). Node support was calculated with 3,000 bootstrap replicates using an "autoMRE" option and the transfer bootstrap expectation method (Lemoine et al. 2018).

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

Author Contributions

S.E. conceived and supervised the project. C.S. collected the samples. C.S. and H.C. conducted the experiments. D.J. and J.Y. performed the bioinformatic analyses. D.J. and S.E. wrote the draft manuscript. All authors contributed to the interpretation, discussion, and revision of the manuscript and approved the final submission.

Funding

This study was supported by the Korea Polar Research Institute (PE22900) and the Korea Institute of Marine Science & Technology Promotion (RS-2022-KS221653 and RS-2025-02215227) funded by the Ministry of Oceans and Fisheries to S.E. and the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education (RS-2023-00243296) to D.J.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability

All NGS data are deposited in the NCBI under BioProject ID PRJNA1133627: WGS, WTS, and HiFi data are available under the Sequencing Read Archive (SRA) IDs, SRR29749513 to SRR29749518. The datasets are also available at http://eyunlab.cau.ac.kr/antarctic_springtail.

Ethics Approval and Consent to Participate

No ethical approval is required for the materials or analyses.

Literature Cited

Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 2000:28: 45–48. https://doi.org/10.1093/nar/28.1.45.

Bast J, et al. No accumulation of transposable elements in asexual arthropods. Mol Biol Evol. 2016:33:697–706. https://doi.org/10.1093/molbev/msv261.

Bellinger PF, Christiansen KA, Janssens F. Checklist of the Collembola of the world. 1996–2024. http://www.collembola.org. Accessed 21 June 2024.

Bellini BC, Weiner WM, Winck BR. Systematics, ecology and taxonomy of Collembola: introduction to the Special Issue. Diversity. 2023:15:221. https://doi.org/10.3390/d15020221.

Block W. The Signy Island terrestrial reference sites: XIV. Population studies on the Collembola. Br. Antarct. Surv. Bull. 1982:55: 33–49. https://nora.nerc.ac.uk/id/eprint/524264/1/bulletin55_05.pdf.

Block W, Tilbrook PJ. Respiration studies on the Antarctic collembolan *Cryptopygus antarcticus*. Oikos. 1975:26:15–25. https://doi.org/10.2307/3543271.

Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021:18: 366–368. https://doi.org/10.1038/s41592-021-01101-x.

Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. rnaSPAdes: a *de novo* transcriptome assembler and its application to RNA-Seq data. GigaScience. 2019:8:giz100. https://doi.org/10.1093/gigascience/giz100.

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol. 2021:38:5825–5829. https://doi.org/10.1093/molbev/msab293.

Chen C, et al. TBtools-II: a "one for all, all for one" bioinformatics platform for biological big-data mining. Mol Plant. 2023:16: 1733–1742. https://doi.org/10.1016/j.molp.2023.09.010.

Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018:34:i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

Christiansen K. Bionomics of Collembola. Ann Rev Entomol. 1964:9: 147–178. https://doi.org/10.1146/annurev.en.09.010164.001051.

Convey P, Smith RIL. The terrestrial arthropod fauna and its habitats in northern Marguerite Bay and Alexander Island, maritime Antarctic. Atarct Sci. 1997:9:12–26. https://doi.org/10.1017/S0954102097000035.

Danecek P, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021:10:1–4. https://doi.org/10.1093/gigascience/giab008.

Deharveng L, D'Haese CA, Bedos A. Global diversity of springtails (Collembola; Hexapoda) in freshwater. Hydrobiologia. 2008:595: 329–338. https://doi.org/10.1007/s10750-007-9116-z.

Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019:20:238. https://doi.org/10.1186/s13059-019-1832-y.

Eyun S, et al. Molecular evolution of glycoside hydrolase genes in the western corn rootworm (*Diabrotica virgifera virgifera*). PLoS One. 2014:9:e94052. https://doi.org/10.1371/journal.pone.0094052.

Faddeeva-Vakhrusheva A, et al. Gene family evolution reflects adaptation to soil environmental stressors in the genome of the collembolan *Orchesella cincta*. Genome Biol Evol. 2016:8:2106–2117. https://doi.org/10.1093/gbe/evw134.

Faddeeva-Vakhrusheva A, et al. Coping with living in the soil: the genome of the parthenogenetic springtail *Folsomia candida*. BMC Genomics. 2017:18:493. https://doi.org/10.1186/s12864-017-3852-x.

Gabriel L, et al. BRAKER3: fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. Genome Res. 2024:34:769–777. https://doi.org/10.1101/gr.278090.123.

Hayward SAL, Worland MR, Convey P, Bale JS. Habitat moisture availability and the local distribution of the Antarctic Collembola *Cryptopygus antarcticus* and *Friesea grisea*. Soil Biol Biochem. 2004:36:927–934. https://doi.org/10.1016/j.soilbio.2004.02.007.

Hu J, et al. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. Genome Biol. 2024:25:107. https://doi.org/10.1186/s13059-024-03252-4.

Hu J, Fan J, Sun Z, Liu S. NestPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics. 2020:36: 2253–2255. https://doi.org/10.1093/bioinformatics/btz891.

Jin J, Zhao Y, Zhang G, Pan Z, Zhang F. The first chromosome-level genome assembly of *Entomobrya proxima* Folsom, 1924 (Collembola: Entomobryidae). Sci Data. 2023:10:541. https://doi.org/10.1038/s41597-023-02456-w.

Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013:30:772–780. https://doi.org/10.1093/molbev/mst010.

Kelley JL, et al. Compact genome of the Antarctic midge is likely an adaptation to an extreme environment. Nat Commun. 2014:5: 4611. https://doi.org/10.1038/ncomms5611.

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 2019:35:4453–4455. https://doi.org/10.1093/bioinformatics/btz305.

Kumar S, et al. TimeTree 5: an expanded resource for species divergence times. Mol Biol Evol. 2022:39:msac174. https://doi.org/10.1093/molbev/msac174.

Lemoine F, et al. Renewing Felsenstein's phylogenetic bootstrap in the era of big data. Nature. 2018:556:452–456. https://doi.org/10.1038/s41586-018-0043-0.

Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018:34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006:22: 1658–1659. https://doi.org/10.1093/bioinformatics/btl158.

Lynch M. Evolution of the mutation rate. Trends Genet. 2010:26: 345–352. https://doi.org/10.1016/j.tig.2010.05.003.

Lynch M, Conery JS. The origins of genome complexity. Science. 2003:302:1401–1404. https://doi.org/10.1126/science.1089370.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021:38:4647–4654. https://doi.org/10.1093/molbev/msab199.

Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. Bioinformatics. 2011:27: 764–770. https://doi.org/10.1093/bioinformatics/btr011.

Marino A, Debaecker G, Fiston-Lavier A-S, Haudry A, Nabholz B. Effective population size does not explain long-term variation in genome size and transposable element content in animals. eLife. 2025:13:RP100574. https://doi.org/10.7554/eLife.100574.3.

Mendes FK, Vanderpool D, Fulton B, Hahn MW. CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics. 2021:36: 5516–5518. https://doi.org/10.1093/bioinformatics/btaa1022.

Minh BQ, et al. 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.

Olejniczak I, Sterzyńska M, Boniecki P, Kaliszewicz A, Panteleeva N. Collembola (Hexapoda) as biological drivers between land and sea. Biology. 2021:10:568. https://doi.org/10.3390/biology10070568.

Pan Z, Jin J, Xu C, Yu D. Chromosomal-level genome assembly of the springtail *Tomocerus qinae* (Collembola: Tomoceridae). Genome Biol Evol. 2022:14:evac039. https://doi.org/10.1093/gbe/evac039.

Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One. 2010:5:e9490. https://doi.org/10.1371/journal.pone.0009490.

Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 2020:11:1432. https://doi.org/10.1038/s41467-020-14998-3.

Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 2018:19:460. https://doi.org/10.1186/s12859-018-2485-7.

Rusek J. Biodiversity of Collembola and their functional role in the ecosystem. Biodivers Conserv. 1998:7:1207–1219. https://doi.org/10.1023/A:1008887817883.

Smit AFA, Hubley R. RepeatModeler Open-1.0. 2008–2015. http://www.repeatmasker.org.

Smit AFA, Hubley R. RepeatMasker Open-4.0. 2013–2015. http://www.repeatmasker.org.

Song JM, et al. Molecular cloning and characterization of a novel cold-active β-1,4-D-mannanase from the Antarctic springtail, *Cryptopygus antarcticus*. Comp. Biochem. Physiol. B. 2008:151:32–40. https://doi.org/10.1016/j.cbpb.2008.05.005.

Song JM, et al. Genetic and structural characterization of a thermotolerant, cold-active, and acidic endo-*β*-1,4-glucanase from Antarctic springtail, *Cryptopygus antarcticus*. J Agric Food Chem. 2017:65:1630–1640. https://doi.org/10.1021/acs.jafc.6b05037.

Tilbrook PJ. The biology of *Cryptopygus antarcticus*. In: Holdgate MW, editors. Antarctic ecology. Academic Press; 1970. p. 908–918.

Watanabe H, Tokuda G. Cellulolytic systems in insects. Ann. Rev. Entomol. 2010:55:609–632. https://doi.org/10.1146/annurev-ento-112408-085319.

Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019:20:257. https://doi.org/10.1186/s13059-019-1891-0.

Wu C, et al. Analysis of the genome of the New Zealand giant collembolan (*Holacanthella duospinosa*) sheds light on hexapod evolution. BMC Genomics. 2017:18:795. https://doi.org/10.1186/s12864-017-4197-1.

Xu G-C, et al. LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. GigaScience. 2019:8: giy157. https://doi.org/10.1093/gigascience/giy157.

Yang H, et al. Consistent accumulation of transposable elements in species of the Hawaiian Tetragnatha spiny-leg adaptive radiation across the archipelago chronosequence. Evol J Linn Soc. 2024:3: kzae005. https://doi.org/10.1093/evolinnean/kzae005.

Zhang F, et al. A high-quality draft genome assembly of *Sinella curviseta*: a soil model organism (Collembola). Genome Biol Evol. 2019:11:521–530. https://doi.org/10.1093/gbe/evz013.

Zhu B-H, et al. P_RNA_scaffolder: a fast and accurate genome scaffolder using paired-end RNA-sequencing reads. BMC Genomics. 2018:19:175. https://doi.org/10.1186/s12864-018-4567-3.

Associate editor: John Wang