



## Two high-quality de novo genomes from single ethanol-preserved specimens of tiny metazoans (Collembola)

Clément Schneider, Christian Woehle, Carola Greve, Cyrille d'Haese, Magnus Wolf, Michael Hiller, Axel Janke, Miklós Bálint, Bruno Huettel

### ► To cite this version:

Clément Schneider, Christian Woehle, Carola Greve, Cyrille d'Haese, Magnus Wolf, et al.. Two high-quality de novo genomes from single ethanol-preserved specimens of tiny metazoans (Collembola). GigaScience, 2021, 10 (5), 10.1093/gigascience/giab035 . mnhn-03365231

**HAL Id: mnhn-03365231**

**<https://mnhn.hal.science/mnhn-03365231>**










Submitted on 5 Oct 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

## DATA NOTE

# Two high-quality *de novo* genomes from single ethanol-preserved specimens of tiny metazoans (Collembola)

Clément Schneider <sup>1,2,\*</sup>, Christian Woehle <sup>3</sup>, Carola Greve <sup>1</sup>, Cyrille A. D’Haese <sup>4</sup>, Magnus Wolf <sup>1,5,6</sup>, Michael Hiller <sup>1,6,7</sup>, Axel Janke <sup>1,5,6</sup>, Miklós Bálint <sup>1,5,†</sup> and Bruno Huettel <sup>3,†</sup>

<sup>1</sup>LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), Senckenberganlage 25, 60325 Frankfurt am Main, Germany; <sup>2</sup>Senckenberg Gesellschaft für Naturforschung, Abteilung Bodenzooologie, Am Museum 1, 02826 Görlitz, Germany; <sup>3</sup>Max Planck Institute for Plant Breeding Research, Max Planck Genome-centre Cologne, Carl-von-Linné-Weg 10, 50829 Cologne, Germany; <sup>4</sup>Unité Mécanismes adaptatifs & Evolution (MECADEV), CNRS, Muséum national d’Histoire naturelle, 45 rue Buffon 75005 Paris, France; <sup>5</sup>Senckenberg Biodiversity and Climate Research Centre, Senckenberganlage 25, 60325 Frankfurt am Main, Germany; <sup>6</sup>Goethe University, Max-von-Laue-Str. 9, 60438 Frankfurt am Main, Germany and <sup>7</sup>Senckenberg Research Institute, Senckenberganlage 25, 60325 Frankfurt, Germany

\*Correspondence address. Clément Schneider, LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), Senckenberganlage 25, 60325 Frankfurt am Main, Germany. Email: [clement.schneider@senckenberg.de](mailto:clement.schneider@senckenberg.de)  <http://orcid.org/0000-0003-3743-9319>

<sup>†</sup>These authors contributed equally to the manuscript.

## Abstract

**Background:** Genome sequencing of all known eukaryotes on Earth promises unprecedented advances in biological sciences and in biodiversity-related applied fields such as environmental management and natural product research. Advances in long-read DNA sequencing make it feasible to generate high-quality genomes for many non-genetic model species. However, long-read sequencing today relies on sizable quantities of high-quality, high molecular weight DNA, which is mostly obtained from fresh tissues. This is a challenge for biodiversity genomics of most metazoan species, which are tiny and need to be preserved immediately after collection. Here we present *de novo* genomes of 2 species of submillimeter Collembola. For each, we prepared the sequencing library from high molecular weight DNA extracted from a single specimen and using a novel ultra-low input protocol from Pacific Biosciences. This protocol requires a DNA input of only 5 ng, permitted by a whole-genome amplification step. **Results:** The 2 assembled genomes have N50 values >5.5 and 8.5 Mb, respectively, and both contain ~96% of BUSCO genes. Thus, they are highly contiguous and complete. The genomes are supported by an integrative taxonomy approach including placement in a genome-based phylogeny of Collembola and designation of a neotype for 1 of the species. Higher heterozygosity values are recorded in the more mobile species. Both species are devoid of the biosynthetic pathway for  $\beta$ -lactam antibiotics known in several Collembola, confirming the tight correlation of antibiotic synthesis with the species way of life. **Conclusions:** It is now possible to generate high-quality

Received: 12 December 2020; Revised: 5 February 2021; Accepted: 27 April 2021

© The Author(s) 2021. Published by Oxford University Press GigaScience. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

genomes from single specimens of minute, field-preserved metazoans, exceeding the minimum contig N50 (1 Mb) required by the Earth BioGenome Project.

**Keywords:** long-read genome sequencing; PacBio; soil invertebrates; eukaryote biodiversity; low-input DNA; integrative taxonomy

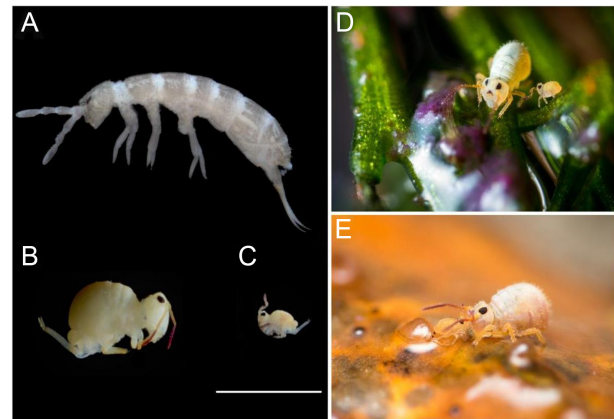
## Introduction

Biodiversity genomics uses genome-scale data to study the molecular basis of biodiversity. New genome data and their analyses are currently revolutionizing life sciences and environmental sciences by addressing scientific questions on evolution, phylogeny, ecology, medicine, and other fields of life sciences. One year after the start of the LOEWE Center for Translational Biodiversity Genome (LOEWE-TBG), the Earth BioGenome Project (EBP) announced plans to sequence reference genomes from all known ~1.5 M eukaryotic species [1]. High-quality (highly contiguous and complete, preferentially chromosome-level) genomes sequenced from accurately species-identified organisms are essential for these efforts. To achieve its goal, the biodiversity genomics faces a major challenge: most of the eukaryotic biodiversity belongs to highly diverse families of tiny species [2] that are (i) difficult to sequence and (ii) difficult to identify.

Advances in long-read sequencing technology changed the game for biodiversity genomics because this technology now allows high-quality genomes to be obtained for diverse taxa. However, minute metazoans pose a number of challenges to long-read sequencing. Standard protocols for long-read sequencing require a large input of high molecular weight (hmw) DNA—in the order of a microgram—which in turn requires larger amounts of fresh or well-preserved input tissue. Pooling individuals from field-collected specimens is often not possible and not desirable: many species cannot be captured in sufficiently large numbers, and pooling individuals complicates assembly by increasing genetic heterogeneity and bears the risk of mixing cryptic species. Small animals often need to be preserved as soon as they are removed from their natural habitats. Furthermore, to be precisely identified, individuals have to be sorted, prepared, and observed under a microscope. This results in delays between specimen collection and DNA extraction and cannot be done on living specimens. Therefore, most small metazoan species will have to be genome-sequenced from single, field-preserved specimens.

Recent progress has already decreased the amount of DNA needed for long-read sequencing. Kingan et al. genome-sequenced a single mosquito on the Pacific Biosciences (PacBio) platform [3]. Adams et al. obtained a chromosome-level assembly from a single, laboratory-bred, fruit fly based on a combination of Nanopore long reads, Illumina short reads, and low-input Hi-C sequencing [4]. However, most metazoans are even smaller than a single mosquito or fruit fly and would not yield the amount of DNA required by the applied protocols.

Here we present high-quality genomes of 2 non-model field-collected Collembola species (Arthropoda: Collembola): *Desoria tigrina* (length: 2 mm; Fig. 1A; NCBI:txid370036) and *Sminthurides aquaticus* (length: 1 mm; Fig. 1B–E; NCBI:txid281415). We extracted DNA from single specimens, preserved for 3–45 days in 96% ethanol, and used a recent whole-genome amplification-based ultra-low DNA input workflow for Single Molecule, Real-Time (SMRT) Sequencing (PacBio) [5] to produce libraries from as little as 5 ng DNA input. Using these libraries, we sequenced 1 SMRT cell for each species. To set the genomes as reliable ref-



**Figure 1:** (A) *Desoria tigrina*. (B–E) *Sminthurides aquaticus* (B) female, (C) male, (D) male and female on wet plant, (E) courtship on a floating dead twig: the male uses its clasping antennae to grab the antennae of the much bigger female. (A–C) Specimens preserved in 96% ethanol, scale bar = 1 mm.

erences, we followed a thorough taxonomic workflow leading to the designation of a needed neotype for *S. aquaticus*. We investigated the resulting genomes for the presence of a  $\beta$ -lactam antibiotic synthesis pathway, an exceptional trait in the metazoan kingdom known in some species of edaphic Collembola [6]. We placed the 2 species in a genome-based phylogeny of Collembola. The resulting genomes are highly contiguous and nearly complete. The *S. aquaticus* assembly even has the highest contiguity compared to the Collembola genomes sequenced so far from hundreds of cultured specimens [7, 8].

Thus, we show that high-quality, *de novo* genomes can be sequenced following a typical taxonomic workflow, even from sub-millimeter species that have been preserved for several days in 96% ethanol. This novel approach will add to the aim of biodiversity genomics to sequence all life on Earth, and make closer the day when whole-genome sequencing will be a routine component of integrative taxonomy.

## Materials and Methods

### Sequenced species

The collembolan *D. tigrina* (Entomobryomorpha, Isotomidae) is a hemiedaphic species: it is found in the upper layer of soil and litter. It is mostly found in anthropized environments [9]. It can be very abundant in vegetal compost, is found in crop fields [10], and can occur in caves as a troglophile [11]. In Western Europe it remains active in winter. The collembolan *S. aquaticus* (Symphypleona, Sminthuridae) is an epigeous, hygrophilous species that is widely spread in the Holarctic region [12]. Specimens often gather on plants, wood, and rocks emerging from the water surface. The animals can walk and jump on water surfaces thanks to elongated claws and a strong furca (jump appendage) with a tip that functions as a paddle on the surface tension. The species has a remarkably pronounced sexual dimorphism: the male is significantly smaller than the female and its modified

antennae in the form of a prehensile organ allow it to clasp the female antennae in a courtship dance preceding external fecundation (Fig. 1B–E).

### Specimen collection and preparation

*Desoria tigrina* was collected in a garden compost bin 8 31.278 E, 50 8.358 N, 14 December 2019). Specimens were extracted from the compost with a Berlese funnel directly into 96% ethanol. DNA extraction was performed within ~72 h. *Sminthurides aquaticus* was collected from a pond in a public garden 2 23.994 E, 48 51.534 N, 27 October 2019). Specimens were caught manually by eye using a small net and mouth-aspirator. They were preserved in 96% ethanol, kept at ambient temperature for 1 day until they could be stored at  $-20^{\circ}\text{C}$ . They remained in cold storage for 1.5 months until we could proceed with DNA extraction. For each species, we gathered a pool of specimens collected simultaneously and pre-identified them all using a stereomicroscope ( $\leq 60\times$  magnification). For *D. tigrina*, 4 specimens were used for DNA extraction (involving their destruction) and 30 were used for precise morphological identification. For *S. aquaticus*, 8 specimens were used for DNA extraction and 17 for morphological identification. Specimens used for morphological identification were cleared in lactic acid and potassium hydroxide, and they were mounted in permanent slides using Marc-André II mounting medium. Observations were made using a Leitz Wetzlar Diaplan with phase contrast, at 400–1,000 $\times$  magnification.

### Ultra-low input PacBio sequencing

Our workflow for DNA extraction and ultra-low DNA input follows the flow chart shown in Fig. 2. Extraction was performed from a single specimen for both species. Specimens were rinsed in  $1\times$  phosphate-buffered saline (PBS) to remove residual ethanol. The solution was replaced 4 times with fresh PBS. Specimens were crushed by 1-way pistons, then DNA was extracted using the Qiagen MagAttract kit (Hilden, Germany). DNA was eluted once in 40  $\mu\text{L}$  AE buffer. We performed 8 individual extractions from *S. aquaticus* and 4 individual extractions from *D. tigrina* specimens. Each DNA extract was quantified with the Quantus dsDNA system (Promega) and DNA quality was assessed with FEMTOpulse (Agilent). Randomly 1 DNA extract was selected for each species. The selected extract contained 59.24 ng hmw DNA (*D. tigrina*) and 16.16 ng hmw DNA (*S. aquaticus*), respectively (FEMTOpulse measurements are provided in Supplementary File S1).

Libraries were prepared using an early access kit for the Ultra-Low DNA Input Workflow for SMRT Sequencing (PacBio) [5]. Of the genomic hmw DNA extracts, 5 ng was fragmented with g-Tubes (Covaris) in an Eppendorf 5424 R centrifuge for 2 min at 1,902g. The resulting fragment sizes were again inspected with FEMTOpulse (Agilent). Next, single-stranded overhangs were enzymatically removed, followed by DNA damage repair, repair of DNA ends, and an A-tailing step. Double-stranded DNA adapter with a T-overhang was ligated for 1 h at  $20^{\circ}\text{C}$  and the resulting products were bead purified (ProNex, Promega, Madison, Wisconsin, USA), eluted, and split into 2 identical aliquots. DNA fragments with adapters were amplified in 2 different PCR reactions (Reaction 1:  $98^{\circ}\text{C}$  for 45 s, 14 cycles:  $98^{\circ}\text{C}$  for 10 s,  $62^{\circ}\text{C}$  for 15 s,  $72^{\circ}\text{C}$  for 7 min, final elongation  $72^{\circ}\text{C}$  5 min; Reaction 2:  $98^{\circ}\text{C}$  for 30 s, 14 cycles:  $98^{\circ}\text{C}$  for 10 s,  $60^{\circ}\text{C}$  for 15 s,  $68^{\circ}\text{C}$  for 10 min, final elongation  $68^{\circ}\text{C}$  5 min). PCR reactions were again bead purified and eluted in EB. Library fragments

were assessed for quantity (Quantus, Promega) and quality (FEMTOpulse, Santa Clara, California, USA). PCR fragments from both reactions were pooled in equal concentrations to achieve a total of 500 ng input for library preparation. Libraries were prepared following the low DNA input workflow for SMRT Sequencing (PacBio, Menlo Park, California, USA). Libraries were annealed to a sequencing primer (V4), bound to Sequel II DNA polymerase 2.0 with Binding kit 2.0, and sequenced in a Sequel II 8M SMRT cell for 30 h.

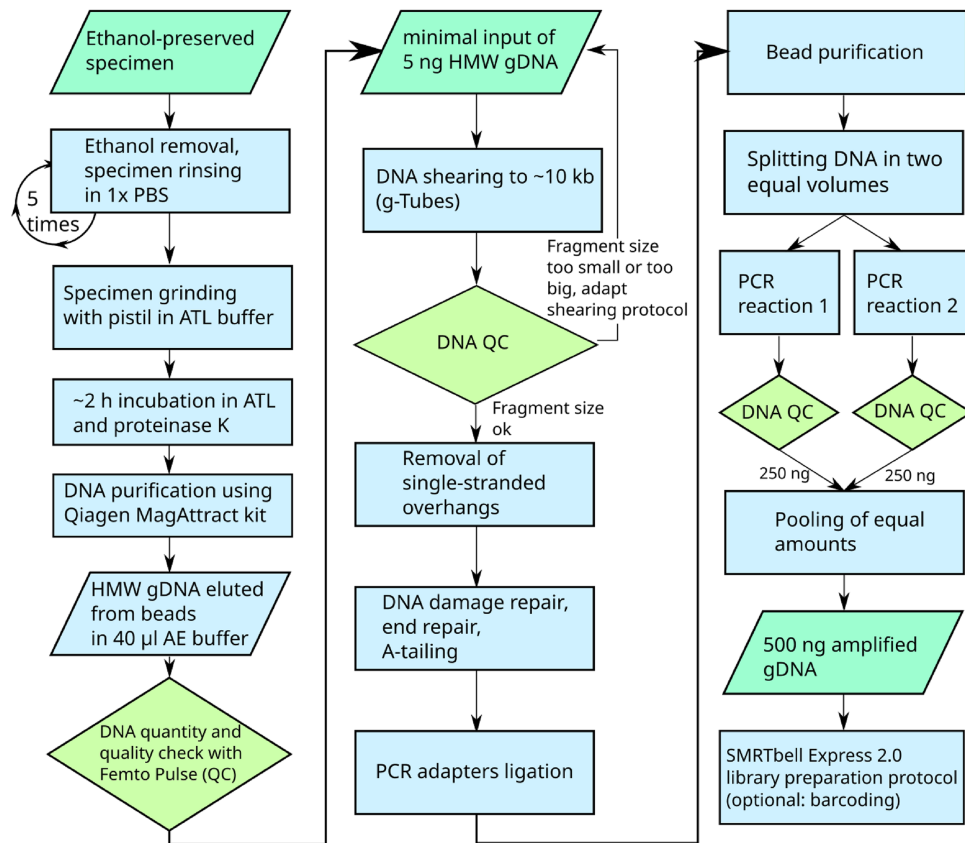
### Genome assembly

Generation of circular consensus sequencing (CCS) reads and adapter trimming was done in PacBio SMRTLink 8 with default parameters followed by deduplication of reads via pbmarkdup (v0.2.0 [13]) as recommended by PacBio. HiFi reads containing complete PCR adapter sequences were discarded. Genome properties were estimated with *k*-mer statistics prior to assembly. This is possible owing to the low error rates of HiFi reads. The *k*-mers were counted and aggregated using jellyfish 2.2.10 [14] (“jellyfish count -C -m 21 -t 20 -s 1000000000 -o jelly.k21.jf CCS.fasta” and “jellyfish histo -t 10 jelly.k21.jf > kmer.histo”) (Jellyfish, [RRID:SCR.005491](#)). We used GenomeScope 1.0 (GenomeScope, [RRID:SCR.017014](#)) [15] to estimate genome length, level of duplication, and heterozygosity through the web application [16].

Several long-read assembly tools were compared: FALCON (falcon-kit v1.8.0) (Falcon, [RRID:SCR.016089](#)) [17], Flye (v2.9.1-b1676) (Flye, [RRID:SCR.017016](#)) [18], HiCanu (v2.1) [19], Hifiasm (v0.12-r304) [20], IPA (v1.1.2) [21], and wtdbg2 v2.5 (WTDBG, [RRID:SCR.017225](#)) [22]. The command lines and statistics of preliminary assemblies are provided in Supplementary File S1. We retained Hifiasm, which produced the assemblies with significantly higher N50 for both species. We repeated the assembly process with Hifiasm after removing 5% and 10% of the shortest reads. Further separation of haplotigs was performed using alternatively purge\_dups [23] (v1.0.1) or purge\_haplotigs [24] (v1.1.0), and for each species we retained the method that achieved better BUSCO deduplication. The effect of purging on genome completeness was assessed with BUSCO v4.1.4 (BUSCO, [RRID:SCR.015008](#)) [25], in genome mode and with the “long” option, with the arthropoda\_odb10 dataset [26]. For each species, we selected the method that led to the highest N50 and optimal purging (lowest amount of duplicated BUSCOs without significant loss of complete BUSCOs). For additional polishing of the resulting assemblies, we followed PacBio guidelines [17]. We used racon (Racon, [RRID:SCR.017642](#)) [27] (v1.4.10, parameter: “-u”) in combination with samtools (SAMTOOLS, [RRID:SCR.002105](#)) [28] (V1.9, parameters: “view -F 1796 -q 20”) and pbmm2 [29] (v1.1.0, parameters: “-preset CCS -sort”), a wrapper of minimap2 (Minimap2, [RRID:SCR.018550](#)) [30].

To assemble the mitochondrial genomes, we gathered CCS containing exclusively mitochondrial sequences, used blastn (blastn+ suite v2.10.0; [RRID:SCR.001598](#)) [31], and assembled them using Geneious 2020.1.2 (Geneious, [RRID:SCR.010519](#)) [32]. Circularity was validated manually, and nucleotide bases were called with a 75% threshold consensus. The mitochondrial genomes were annotated with the MITOS2 web server [33]. Coding DNA sequences were checked and corrected using Geneious to ensure that the presence of uncommon start codons and incomplete stop codons did not mislead the automatic annotation algorithms. Boundaries of the recombinant DNAs were slightly adjusted to make them contiguous with the tRNA(val) gene.





**Figure 2:** Flow chart of DNA extraction and ultra-low input workflow for SMRTbell Express 2.0 library preparation, for a single ethanol-preserved specimen. gDNA: genomic DNA; QC: quality control.

We used *blastn* to identify insertions of the mitochondrial genome in the nuclear genome (NUMTs). For this query, we used a 2× duplicated sequence of the mitochondrial genome to handle circularity. We recognized the presence of almost complete copies of the mitochondrial genome in the nuclear genomes of both species. We investigated the mapping of the CCS to the assembly in those locations using IGV [34] (v2.8.13) and recognized that in 1 instance, a misassembly occurred through the soldering of 2 NUMTs with CCS of mitochondrial origin. All CCS aligning with those 2 NUMTs were gathered with *blastn* and reassembled using Geneious. We could not find unambiguous NUMTs CCS (i.e., CCS carrying both nuclear and mitochondrial sequence) that would support the original assembly connection, and therefore we split the contig.

### Contamination control

We checked the assemblies for potential contamination from other organisms by querying the contigs against the NCBI database using protein-based (DIAMOND [35]) and nucleotide-based (*blastn*) alignments. Results were merged with Blobtools2 [36] (v2.3.3) using the “bestsum” algorithm. Contigs explicitly assigned to another lineage than metazoan were excluded from the assembly. Contigs assigned to Chordata were checked for presence of Arthropoda BUSCO. If Arthropoda BUSCOs were confirmed on such contigs, we retained them for the assembly.

### Assembly assessment

Curated assemblies were again evaluated with BUSCO (same parameters as before). We mapped the CCS on the assemblies using *backmap* [37] (v0.3), a perl wrapper of *minimap2* and *QualiMap2* [38]. *Minimap2* was run with “-H -ax asm10” to map CCS on the assembly. We then performed another estimation of the genome size by dividing the number of mapped nucleotides by mode of the coverage distribution [37].

### Comparison with previous long-read assemblies

We compared our new genomes sequenced to previous Collembola assemblies that were generated with long-read and sometimes additional short-read data [7, 8, 45]. We also compared our Collembola assemblies to the draft genomes of 2 larger insects [3, 39] (4 and 20 mm), which were also sequenced from single specimens but with the PacBio low-input workflow [40] (amplification-free).

### Alternative haplotig assembly

For both species, we obtained the alternative haplotig assembly by concatenating the alternate haplotig produced by Hifiasm with the duplicated contigs identified in the primary assembly during the purging step. We then further curated the alternative haplotig assembly by using sequentially *Purge.dups* and the decontamination strategy described above; and finally evaluated the BUSCOs completeness.

## Genome annotation

The primary assemblies were annotated with *ab initio* gene prediction. Repetitive regions were masked with RepeatModeler (RepeatModeler, [RRID:SCR.015027](#)) [41] (v2.0.1) with the options: “-LTRStruct -engine ncbi” using RepeatMasker (RepeatMasker, [RRID:SCR.012954](#)) [42] (open-4.0.9, options: “-xsmall -gff -nolow”). Protein sequences were predicted with AUGUSTUS (Augustus, [RRID:SCR.008417](#)) [43] (v3.3.3, option: “-softmasking = on”) reusing the BUSCO training results. Functional annotations were obtained by a local installation of eggNOG-mapper [44] (v2.0.1, option: “-m diamond”). If emapper recovered no annotations, we denoted sequences as “hypothetical protein” (for proteins without hits in emapper) or “uncharacterized protein” (for proteins with hits without annotations). To determine whether *D. tigrina* and *S. aquaticus* share the  $\beta$ -lactam synthesis gene found in some other Collembola, we searched the genomes for genes homologous of the isopenicillin N synthase (IPNS) and  $\delta$ -(L- $\alpha$ -aminoadipoyl)-L-cysteine-D-valine synthetase (ACVS) genes of *Folsomia candida*. Those 2 genes belong to the same gene cluster. We used blastn and megablast to query the DNA sequences and tblastn to query the protein sequences against the combined primary and alternative haplotig assemblies. We also used blastp to query the protein sequences against the predicted protein sequences from the primary assemblies. The NCBI accession numbers of the searched sequences are IPNS—JX270832.1, ACVS—OXA60265.1.

## Phylogenetic analysis

We gathered 13 Collembola genome assemblies [7, 8, 45, 46] from NCBI. For the outgroup, we selected a Diplura [47] and a Diptera [3] genome assemblies. The species list and the genome accession numbers are provided in Table 1. We used BUSCO v4.0.6 in short mode to search for orthologs, restricting the search to the arthropoda.odb10 dataset. We screened the obtained BUSCO sets to identify genes shared among the species, allowing only genes found for  $\geq 75\%$  of the species. We aligned single protein sequences with MAFFT (MAFFT, [RRID:SCR.011811](#)) [48] (v7.450), concatenated the alignments with FASconCAT-G [49] (v1.04), and trimmed the final alignment with trimAl (trimAl, [RRID:SCR.017334](#)) [50] (v1.2). We calculated a maximum likelihood tree with IQtree [51] (v1.6.12) with 1,000 non-parametric bootstrap replications.

## Results

### Species biology and taxonomy

In terms of biomass and number *D. tigrina* was by far the dominant Collembola found in the compost bin during the winter season. Morphological observations placed the collected specimens unambiguously within the *D. tigrina* group [52]. Within this group, outer maxillary palp chaetotaxy was used to distinguish *D. tigrina* from its sibling species *Desoria grisea* following Fjellberg [52]. Identification was further validated following Potapow [9]. Six females, 2 males, and 2 juveniles on 4 slides numbered EA013940–3 were deposited in the Apterygota collection of the National Museum of Natural History, Paris. Seventeen females and 3 males on 12 slides labelled CSCH-1326–37 were deposited in the Apterygota collection at Senckenberg, G rlitz.

*Sminthurides aquaticus* was the only Collembola forming a population on the pond at the site of collection (i.e., no accidental fall on water surface). Abundantly found in October 2019,

it was observed again in June 2020 in large numbers and with courtship behavior undergoing (Fig. 1D and E). The specimen identification was unambiguous following [12, 52, 53]. One male on a slide numbered EA060001 is designated as the neotype for *S. aquaticus* (see discussion) and was deposited in the Apterygota collection at the National Museum of Natural History, Paris, along with 3 females and 2 males on 5 slides (EA060002–6) and 20 individuals in 96% ethanol (CS.371, leg. C. Schneider). Two males, 3 females, and 1 juvenile on 5 slides numbered CSCH-1344–8 were deposited in the Apterygota collection at Senckenberg, G rlitz.

## DNA sequencing

For *D. tigrina* a total of 20.22 Gb HiFi data ( $Q \geq 20$ ) were generated, with mean read length of 12,155 bp, median read length of 11,792 bp, and maximum read length of 37,982 bp. The distribution of read length is reported in Fig. 3. From the *k*-mer content of the reads, the genome haploid length was estimated to be  $\sim 168$  Mb with 1.43% of heterozygosity and 3% duplications.

For *S. aquaticus* a total of 12.4 Gb HiFi data ( $Q \geq 20$ ) were generated with mean length of 12,308 bp, median read length of 11,893 bp, and maximum read length of 29,073 bp. The distribution of read length is reported in Fig. 3. From the *k*-mer content of the reads, the genome haploid length was estimated to be  $\sim 152$  Mb with 0.96% of heterozygosity and 0.78% duplications.

## Genome assembly

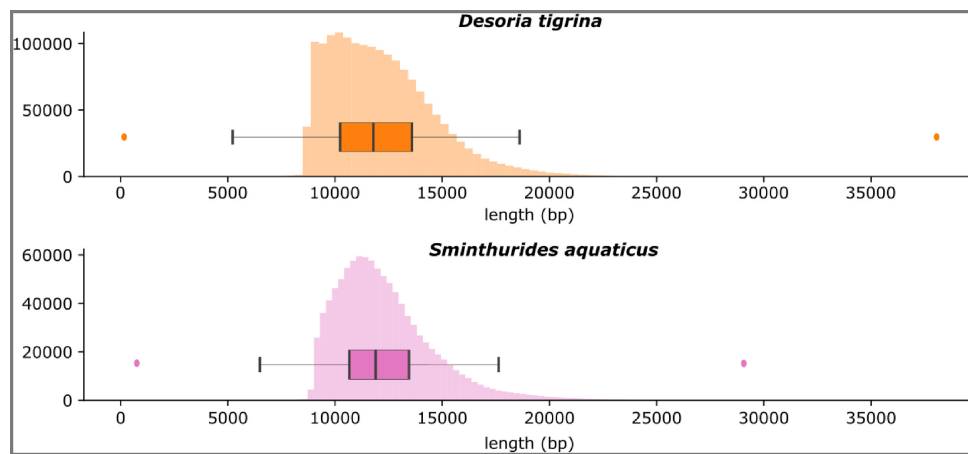
Overall, Hifiasm produced the best assemblies for both species (Supplementary File S1).

For *D. tigrina*, the most contiguous assembly (Table 2, Fig. 4) was obtained by selecting 95% of the reads excluding the shortest one. Purging haplotigs with Purge\_dups resulted in fewer duplicated BUSCOs than with Purge\_haplotigs. No contigs were found to be of non-metazoan origin. While some contigs were assigned to Chordata taxa, they all carried Arthropoda-specific BUSCOs and were therefore kept. The curated primary assembly of *D. tigrina* is composed of 142 contigs and has a size of 211,462,971 bp and an N50 value of 5.63 Mb (Table 2, Fig. 4). Mean coverage is 95.40 $\times$ , with a coverage distribution mode of 103 $\times$ . The genome size is 196 Mb, estimated from mapped reads and coverage. BUSCO search on the whole assembly yielded 96% complete (C) BUSCOs (including 1.7% duplicated [D]), 0.9% fragmented (F) BUSCOs, and 3.1% missing (M) BUSCOs. The mitochondrial genome assembly was complete, for a size of 15,139 bp. Two large NUMTs were found, each on a different contig. One was 18,113 bp (120% of the mitochondrial genome size) and the other was 28,173 bp (186% of the mitochondrial genome size). Examination of the mapped reads revealed no obvious misassembly for the smaller NUMT (spanned by reads that contained mitochondrial and genomic sequence), but the larger NUMT was bridged in the middle by reads containing exclusively mitochondrial sequence. Therefore, we split the contig carrying the larger NUMT, keeping on each side a partial NUMT sequence supported by reads containing mitochondrial and genomic sequence. The alternative haplotig assembly of *D. tigrina* is composed of 1,611 contigs and has a size of 189,752,789 bp and an N50 value of 0.27 Mb; BUSCO search on the alternative haplotig assembly yielded 89% complete (including 3.8% duplicated), 0.9% fragmented, and 10.1% missing BUSCOs.

For *S. aquaticus* the best assembly was obtained by using all the reads (Table 2, Fig. 4). Purging haplotigs with Purge\_haplotigs

**Table 1:** Species included in the phylogenetic analysis (taxonomic dataset expanded from [44])

Species	Order	Family	Repository	Accession	Source
<i>Anopheles coluzzii</i>	Diptera	Culicidae	NCBI	ASM413651v2	[3]
<i>Catajapyx aquilonaris</i>	Dicellurata	Japygidae	NCBI	GCA.000934665.2	[47]
<i>Ceratophysella communis</i>	Poduromorpha	Hypogastruridae	NCBI	GCA.009869905.1	[44]
<i>Desoria tigrina</i>	Entomobryomorpha	Isotomidae	EMBL-ENA	ERZ1473261	This study
<i>Folsomia candida</i>	Entomobryomorpha	Isotomidae	NCBI	GCA.002217175.1	[7]
<i>Lipothrix lubbocki</i>	Symphyleona	Sminthuridae	NCBI	GCA.009872335.1	[44]
<i>Mesaphorura yosii</i>	Poduromorpha	Tullbergiidae	NCBI	GCA.009869945.1	[44]
<i>Neelides</i> sp.	Neelipleona	Neelidae	NCBI	GCA.009869795.1	[44]
<i>Oncopodura yosiiana</i>	Entomobryomorpha	Oncopoduridae	NCBI	GCA.009869805.1	[44]
<i>Orchesella cincta</i>	Entomobryomorpha	Entomobryidae	NCBI	GCA.001718145.1	[45]
<i>Pseudachorutes palmiensis</i>	Poduromorpha	Neanuridae	NCBI	GCA.009869845.1	[44]
<i>Pseudobourletiella spinata</i>	Symphyleona	Bourletiellidae	NCBI	GCA.009870155.1	[44]
<i>Pygmarrhopalites habei</i>	Symphyleona	Arrhopalitidae	NCBI	GCA.009870185.1	[44]
<i>Sinella curviseta</i>	Entomobryomorpha	Entomobryidae	NCBI	GCA.004115045.1	[8]
<i>Sminthurides aquaticus</i>	Symphyleona	Sminthurididae	EMBL-ENA	ERZ1473260	This study
<i>Sminthurides bifidus</i>	Symphyleona	Sminthurididae	NCBI	GCA.009872375.1	[44]
<i>Thalassaphorura encarpata</i>	Poduromorpha	Onychiuridae	NCBI	GCA.009869925.1	[44]
<i>Tomocerus qinae</i>	Entomobryomorpha	Tomoceridae	NCBI	GCA.009869885.1	[44]

**Figure 3:** Distribution of CCS length. Outliers are not shown on the boxplot, except minimum and maximum length values each represented by a dot.

resulted in fewer duplicated BUSCOs than with Purge.dups. Two contigs (totaling 243,436 bp) were found to be from a fungus and a cyanobacterium, respectively, and were removed. Some contigs were assigned to Chordata taxa but all of those carried Arthropoda-specific BUSCOs and were kept. The curated primary assembly of *S. aquaticus* is composed of 79 contigs and has a size of 165,915,169 bp and an N50 value of 8.78 Mb (Table 2, Fig. 4). Mean coverage is 72.67 $\times$ , and coverage distribution mode is 77 $\times$ . The genome size is 157 Mb, estimated from mapped reads and coverage. BUSCO search on the whole assembly yielded 96.1% complete BUSCOs (including 1.6% duplicated), 1.3% fragmented BUSCOs, and 2.6% missing BUSCOs. The mitochondrial genome assembly was complete, for a size of 16,099 bp. A large NUMT was detected in 1 of the purged contigs (haplotigs), but none were found in the primary contigs, so we decided not to investigate further. Several small contigs were found to be assembled from mitochondrial reads and were removed. The alternative haplotig assembly of *S. aquaticus* is composed of 459 contigs and has a size of 150,171,336 bp and an N50 value of 1.00 Mb; BUSCO search on the alternative haplotig assembly yielded 87.5% complete (including 2.9% duplicated), 1.4% fragmented, and 2.6% missing BUSCOs.

### Comparison with previous long-read assemblies

In terms of BUSCO completeness scores, our assemblies are comparable to previous high-quality Collembola genomes assembled from a large pool of specimens (95.8% and 96.1% vs 94.5–97.1% complete; Table 2). In terms of assembly contiguity, our *S. aquaticus* has the highest and the *D. tigrina* assembly has the third-highest contig N50 value (Table 2, Fig. 4). The insect genomes obtained sequencing from single specimen using the PacBio low-input workflow have higher BUSCO scores (96.5% and 99.6%) but lower contiguity (Table 2, Fig. 4). Together, this shows that assemblies generated with the ultra-low input workflow and long-read sequencing can reach or surpass the level of quality of assemblies obtained with the standard or low-input workflow.

### Genome annotation

In the mitochondrial genome of both species, we identified the complete set of 37 mitochondrial genes (13 proteins, 22 transfer RNA, and 2 ribosomal RNA coding genes) typically found in Hexapoda. In the nuclear genome, we predicted 24,423 pro-

Table 2: Statistics of several assemblies generated from long-read sequencing (with or without additional short reads) and/or low-input approach

Parameter	<i>Desoria tigrina</i>	<i>Sminthurides aquaticus</i>	<i>Folsomia candida</i>	<i>Orchesella cincta</i>	<i>Sinella curviseta</i>	<i>Anopheles coluzzii</i>	<i>Lycorma delicatula</i>
Class	Collembola	Collembola	Collembola	Collembola	Collembola	Insecta	Insecta
Body size class	2 mm	1 mm	2 mm	2 mm	2 mm	4 mm	20 mm
No. specimens in input	1 (PacBio)	1 (PacBio)	1,600 (PacBio) + 100 (Illumina)	40 (PacBio) + 1 (Illumina)	500 (PacBio) + 10 (Illumina)	1 (PacBio)	1 (PacBio)
WGA	Yes	Yes	No	No	No	No	No
No. contigs	142	79	162	9,398	599	1,034	2,927
Largest contig (bp)	14,592,742	19,603,089	28,534,321	807,113	12,986,801	11,911,669	9,998,986
Total length (bp)	211,462,971	165,915,169	221,702,752	286,764,906	381,458,724	340,555,854	2,252,044,789
N50 (bp)	5,628,779	8,776,828	6,519,406	65,879	3,284,409	2,625,112	1,519,606
N75 (bp)	2,264,114	4,641,270	2,726,164	23,461	1,147,450	440,054	811,306
L50 (bp)	11	7	8	925	32	36	434
L75 (bp)	28	13	21	2,812	74	126	935
BUSCO % C (D), E, M	96 (1.7), 0.9, 3.1	96.1 (1.6), 1.3, 2.6	97.1 (0.9), 0.5, 2.4	94.5 (3.2), 1.8, 3.7	95.6 (4.4), 1.3, 3.1	99.6 (2.9), 0.0, 0.4	96.5 (1.9), 2.0, 1.5
Source	This study	This study	[45]	[7]	[8]	[3]	[39]
Assembly	EMBL-ENA: PRJEB39696	EMBL-ENA: PRJEB39696	NCBI: ASM221717v1	NCBI: ASM171814v	NCBI: ASM411504v1	NCBI: ASM413651v2	doi:10.15482/USDA.ADC
accession							/1503745



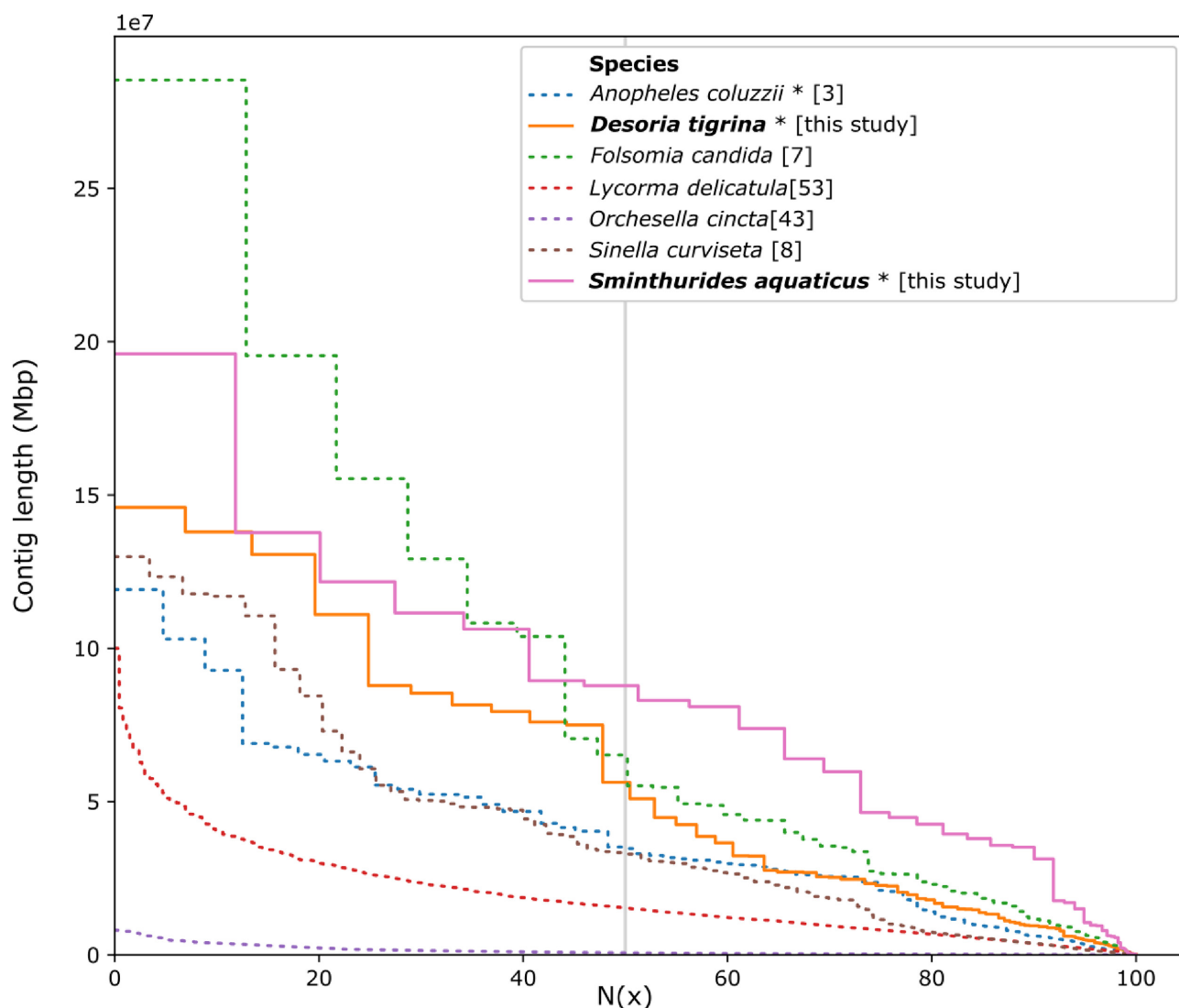


Figure 4: N(x) plot of recent high-quality genomes assembled with long reads, including the assemblies presented in this study.

teins for *D. tigrina*, 15,546 (63.65%) of which had homologs in other organisms and 8,877 were labeled as “hypothetical protein.” BUSCO search on the predicted proteins yielded 96.2% complete BUSCO including 2.6% duplicated, 1.2% fragmented, and 2.6% missing. For *S. aquaticus*, we predicted 17,624 proteins in the nuclear genome, 11,989 (68.03%) of which had homologs in other organisms and 5,635 were labeled “hypothetical protein.” BUSCO search on the predicted proteins yielded 95.3% complete BUSCO including 2.1% duplicated, 1.6% fragmented, and 3.1% missing.

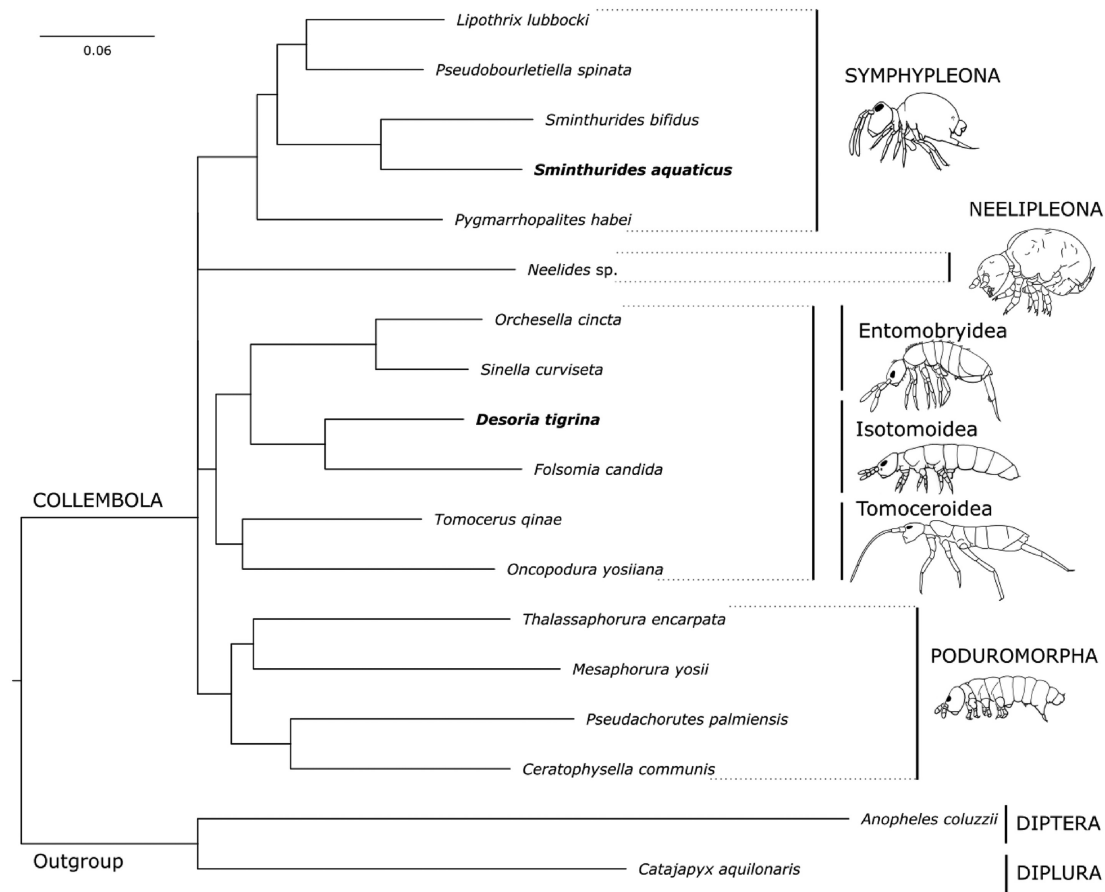
### $\beta$ -lactam biosynthetic pathway

Collembola exhibit a diversity in the presence of a  $\beta$ -lactam antibiotic synthesis pathway, which is secondarily lost in some species. Therefore, we analyzed our genomes for the presence of a  $\beta$ -lactam antibiotic synthesis pathway. No homologs of the genes *IPNS* and *ACVS* could be identified in the 2 genomes, indicating the absence of the  $\beta$ -lactam antibiotic synthesis pathway in *D. tigrina* and *S. aquaticus*. However, by screening the functional annotation of the predicted genes, we identified 4 genes

related to aminopenicillanic-acid-acyltransferase (penDE) in the genome of *S. aquaticus* and 5 penDE-like genes in the genome of *D. tigrina*.

### Phylogeny

To place our 2 species in a molecular phylogeny of Collembola, we used BUSCO genes as conserved phylogenetic markers. Allowing for a maximum of 25% missing sequence for each ortholog, we retained 545 complete BUSCOs to align. The total length of the trimmed alignments is 171,703 sites. We used IQTree to infer a phylogenetic tree, shown in Fig. 5. Our 2 newly sequenced species find their expected placement on the Collembola phylogeny with *Sminthurides aquaticus* as a sister species to *S. bifidus* (both are representatives of the genus *Sminthurides*, family Sminthurididae) and *D. tigrina* as a sister species to *F. candida* (both are representatives of the family Isotomidae). Our tree also recovered the monophyly of orders Symphypleona, Poduromorpha, and Entomobryomorpha with 100% bootstrap support. However, the basal relationships between the 4 orders of Collembola receive negligible bootstrap support (=73%), indicating phy-



**Figure 5:** Phylogeny of Collembola based on the alignment of 545 protein sequences. Bootstrap support of shown nodes is 100%; nodes with bootstrap supports of 73% were collapsed.

logenetic irresolution. The rest of the tree is consistent with the to-date most detailed genome-based phylogeny of Sun et al. [46]. The >400 million years old basal relationships of Collembola have long been debated. They are sensitive to data sampling, and phylogenetic artifacts such as long branch attraction and random root occur [54]. Additional genomes of key Collembola representatives and more informative phylogenetic markers [46] are needed to properly address the problem of basal relationships within Collembola.

## Discussion

### Value of the ultra-low input workflow

Long-read sequencing as the future for *de novo* genome assembly has normally required larger amounts of input tissue, which limits its application to larger organisms. However, a substantial portion of biodiversity is represented by tiny species. Here, we address this important challenge in biodiversity genomics and provide a proof of concept that it is now possible to sequence high-quality reference genomes from field-collected individual tiny Collembola species. The 5 ng input of the PacBio ultra-low input workflow is a significant decrease from the 150 ng input required by the PacBio low-input workflow (whole-genome amplification [WGA]-free). And yet the ultra-low input still allows high-quality genomic data to be captured: our final assemblies were of high contiguity and completeness on par with recent genomes from larger insects sequenced using the low-input pro-

tol [3, 39]. Our new genomes are also on par with the previously best reference genomes for Collembola: *F. candida* and *S. curviseta*, which were DNA sequenced from hundreds of specimens maintained in culture [7, 8]. *Sminthurides aquaticus* even achieved the highest N50 and N75 among the compared assemblies. The quality of the new assemblies makes us consider that there are even additional benefits in the ultra-low input protocol besides sequencing organisms too small for WGA-free approaches. For species that are not too small, it can be used to generate long-read data from a fraction of the total DNA. This could be levered to implement approaches combining long-read and Hi-C for even smaller species than a fruit fly [4]. This can also allow the sequenced specimen to be retained to serve as a voucher, by removing the need to crush the specimen to maximize hmw DNA recovery.

### Ensuring taxonomic quality

It is essential that a reliable reference genome be supported by a solid and revisable taxonomy, to be useful for any meaningful downstream analysis. Taxonomy quality has always been an issue of sequence databases [55, 56]. This is especially true for field-collected specimens from taxonomically poorly known groups that are often riddled with cryptic diversity and difficulty of species identification based on a few subtle characters. Therefore, we documented species collection and identification by morphological characters, provided macro photographs, and preserved co-captured specimens of the same species in the col-

lections of 2 European museums. This way, we ensure the taxonomic traceability of the reference genome, which should be a prerequisite for any meaningful biodiversity genomics where species identification is not straightforward.

The genus *Desoria* has a complex taxonomy. Within the *D. tigrina* group sensu Fjellberg 2007 [52], *D. tigrina* and *D. grisea* are 2 sibling species, described in the early times of modern Collembola systematics. *Desoria grisea* was redescribed by Fjellberg [52] from its type locality. Fjellberg reported that the 2 species, while extremely similar, could be consistently distinguished by the organization of the labial palp chaetae. We examined 30 specimens from our collection spot and each of them were identified as *D. tigrina*, supporting the identity of the specimen used for sequencing.

*Sminthurides aquaticus* was originally described from France and has been recognized to be widely spread throughout the Holarctic region. We confirmed that all our collected specimens are identical to the accepted descriptions of *S. aquaticus*. The species was originally described by Bourlet in 1841 probably from the north of France. However, Bourlet did not make any reference to a type series and, to our knowledge, did not preserve any specimens. We consider that the population we sampled in Paris is suitable to provide a neotype for this species: the population is abundant, settled, and easily accessible for further studies. This also offers the uncommon opportunity to have a neotype closely related to the reference genome for the species.

### Heterozygosity

The higher level of heterozygosity in *D. tigrina* compared to *S. aquaticus* seems consistent with the expected level of isolation of the populations. *Desoria tigrina* invaded the compost that was set up 1 year before the collection. The species is very mobile, being rather large and equipped with a long furca, and gene flow must be active across the nearby surrounding fields and gardens. On the other hand, the sampled population of *S. aquaticus* seems rather isolated in a small area (artificial pond in a public garden).

### $\beta$ -lactam synthesis in Collembola

Recent results from transcriptomes show that several edaphic species from the orders Poduromorpha and Entomobryomorpha can synthesize  $\beta$ -lactam antibiotics [6]. Two essential genes of the  $\beta$ -lactam synthesis pathway, ACVS and IPNS, are consistently found in 4 euedaphic species ("true" soil dweller) but missing in 2 of 4 hemiedaphic species (living in upper layer of soil, litter, and dead wood) and always missing in 7 atmobioc species (species living on vegetation, freshwater surface, or tidal zone). The genes are absent from soil dwellers from the class Diplura and Protura, 2 close relatives of Collembola. The antibiotic biosynthesis likely resulted from a single horizontal gene transfer event with subsequent loss of antibiotic synthesis ability in some of the investigated species [6].

We report the absence of ACVS and IPNS in the genomes of *S. aquaticus* and *D. tigrina*. *Sminthurides aquaticus* belongs to a family of Symphypleona that was not investigated by Suring et al. [6]. So far, no Symphypleona are known to carry those genes, but it must be noted that none of the tested species are soil-dwelling species. The Symphypleona species in the Suring et al. [6] dataset are vegetation dwellers. Because *S. aquaticus* dwells on freshwater surfaces, our results support the lack of antibiotic production in semi-aquatic species. The absence of the genes in *D. tigrina* is rather unexpected because the species lives in

organic-rich litter with potentially high microbial contents. After *F. candida*, *D. tigrina* is the second member of the large Isotomidae family to be investigated for antibiotic production. *Desoria tigrina* is in the same class size as *F. candida* but is expected to be more mobile owing to its more developed legs, furca, and eye-patch (*F. candida* is eyeless). This suggests that antibiotic synthesis is specific to a true soil-dwelling (euedaphic) lifestyle, and it might be lost by more mobile species.

### Antibiotic synthesis in Collembola

Both *D. tigrina* and *S. aquaticus* possess *penDE*-like genes. Such genes were also reported in *F. candida* [6]. The *penDE* is the last enzyme in the penicillin biosynthetic pathway of the fungi *Emericella nidulans*, and converts isopenicillin N (product of IPNS activity) to penicillin G. In *F. candida*, the *penDE*-like gene does not belong to the  $\beta$ -lactam synthesis gene cluster. Homologs of *penDE* are also known in fungi that do not produce antibiotics. Suring et al. [6] suggest that *penDE*-like genes may have been co-opted for the completion of the penicillin synthesis in *F. candida* after the acquisition of the  $\beta$ -lactam synthesis gene cluster. Consequently, the presence of *penDE* in *S. aquaticus* and *D. tigrina* is not a solid indicator of a lost antibiotic production trait in these species. Altogether, the assumption that the horizontal gene transfer is an ancestral acquisition to Collembola should be taken with caution because the basal relationships between Collembola orders are still unresolved. For further elucidation, edaphic species of orders Symphypleona and Neelipleona should be investigated for the antibiotic production trait.

### Conclusions

The LOEWE-TBG excellence cluster supports the goal of the EBP, which aims to sequence all eukaryotic species. Although the first high-quality genomes were generated for species with easy access to abundant and fresh samples, similar high-quality genomes can now be generated for tiny taxa or taxa that are otherwise difficult to sequence. Most known eukaryotic biodiversity belongs to very small metazoans that in addition need to be preserved for some time before genome sequencing. Access to their genomes provides insights into the formation, maintenance, and functioning of eukaryotic biodiversity and presents new opportunities for natural resource management and bio-prospecting. The ability to genome-sequence these species is essential for the success of biodiversity genomics initiatives. Our genomes sequenced from 5 ng DNA actually exceed the 1Mb N50 contig continuity required by the EBP project when >100 ng DNA are available. We are convinced that integrating high-quality genomics with the typical workflow of small, field-collected metazoans is an essential approach toward the creation of a solid reference genome database for millions of minute non-model species belonging to taxonomically challenging groups.

### Data Availability

The data underlying this article are available in the EMBL-ENA database and can be accessed with accession No. PRJEB39696 including: *S. aquaticus* CCS, curated primary assembly, and annotation with accessions Nos. ERR4407379, GCA\_906901655; and *D. tigrina* CCS, curated primary assembly, and annotation with accession Nos. ERR4407422, GCA\_906901685.

Supporting data, including primary and alternative haplotig assemblies, and annotation files, are deposited in the GigaScience

database, GigaDB, for both *Sminthurides aquaticus* [57] and *Desoria tigrine* [58].

## Additional Files

**Supplementary File S1:** Report on preliminary assemblies, including assembly statistics and details of assembly tools and command lines.

## Abbreviations

BLAST: Basic Local Alignment Search Tool; bp: base pairs; BUSCO: Benchmarking Universal Single-Copy Orthologs; CCS: circular consensus sequencing; EBP: Earth BioGenome Project; Gb: gigabase pairs; hmw: high molecular weight; IGV: Integrative Genomics Viewer; kb: kilobase pairs; LOEWE-TBG: LOEWE Center for Translational Biodiversity Genomics; MAFFT: Multiple Alignment using Fast Fourier Transform; Mb: megabase pairs; NCBI: National Center for Biotechnology Information; NUMT: nuclear mitochondrial DNA; PacBio: Pacific Biosciences; SMRT: single molecule, real-time; WGA: whole-genome amplification.

## Competing Interests

The authors declare that they have no competing interests.

## Authors' Contributions

C.S. conceived the project; C.S. and C.A.D'H. collected, identified, and photographed the specimens; B.H. performed the DNA extraction, the library preparation, and the sequencing; C.W. and C.S. assembled and analyzed the genomes; M.W. and A.J. contributed the phylogenomic analysis; C.S. and M.B. led the writing of the manuscript; C.G. performed experiments (not presented here) that helped steer the project and further advised on the study; and M.H. revised the manuscript. All authors read and approved the final manuscript for submission.

## Acknowledgements

The genomes will contribute to the European Reference Genome Atlas and the Earth BioGenome Project. The present study is a collaboration between the LOEWE-TBG and the Max Planck Genome-centre Cologne. It was supported through the programme "LOEWE—Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz" of Hesse's Ministry of Higher Education, Research, and the Arts. We highly appreciate the generous support by Pacific Biosciences with respect to the ultra-low amplification kit and library preparation kit, as well as SMRT cells and sequencing chemistry during the course of the beta test. The Max-Planck Genome Center Cologne acknowledges the support from the Max-Planck Society. We give our warm thanks to Tilman Schell for his advice on genome assembly. We thank Dr. Arong Luo and Dr. Mahul Chakraborty for the reviewing our work; their suggestions and corrections improved the quality of the manuscript.

## References

- Lewin HA, Robinson GE, Kress WJ, et al. Earth BioGenome Project: Sequencing life for the future of life. *Proc Natl Acad Sci U S A* 2018;**115**(17):4325–33.
- Stork NE, McBroom J, Gely C, et al. New approaches narrow global species estimates for beetles, insects, and terrestrial arthropods. *Proc Natl Acad Sci U S A* 2015;**112**(24):7519–23.
- Kingan SB, Heaton H, Cudini J, et al. A high-quality de novo genome assembly from a single mosquito using PacBio sequencing. *Genes* 2019;**10**(1):62.
- Adams M, McBroome J, Maurer N, et al. One fly—one genome: chromosome-scale genome assembly of a single outbred *Drosophila melanogaster*. *Nucleic Acids Res* 2020;**48**(13):e75.
- Pacific Biosciences. Now available: Ultra-low DNA input workflow for SMRT sequencing. <https://www.pacb.com/blog/introducing-the-ultra-low-input-protocol-for-smrt-sequencing/>. 2020. Accessed 4 December 2020.
- Suring W, Meusemann K, Blanke A, et al. Evolutionary ecology of beta-lactam gene clusters in animals. *Mol Ecol* 2017;**26**(12):3217–29.
- Faddeeva-Vakhrusheva A, Kraaijeveld K, Derks MFL, et al. Coping with living in the soil: The genome of the parthenogenetic springtail *Folsomia candida*. *BMC Genomics* 2017;**18**(1):493.
- Zhang F, Ding Y, Zhou Q-S, et al. A high-quality draft genome assembly of *Sinella curviseta*: A soil model organism (Collembola). *Genome Biol Evol* 2019;**11**(2):521–30.
- Potapow M. Synopses on Palaearctic Collembola, Volume 3, Isotomidae. Staatliches Museum für Naturkunde Görlitz; 2001.
- Gruss I, Twardowski J. The assemblages of soil-dwelling springtails (Collembola) in winter rye under long-term monoculture and crop rotation. *Zemdirbyste* 2016;**103**(2):159–66.
- Dányi L. Cave dwelling springtails (Collembola) of Hungary: a review. *Soil Org* 2011;**83**:419–32.
- Bretfeld G. Synopses on Palaearctic Collembola : Symphypleona. Staatliches Museum für Naturkunde Görlitz; 1999.
- Pacific Biosciences. pbmarkdup. <https://github.com/PacificBiosciences/pbmarkdup>. 2020. Accessed 1 March 2020.
- Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 2011;**27**(6):764–70.
- Vurtture GW, Sedlazeck FJ, Nattestad M, et al. GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics* 2017;**33**(14):2202–4.
- Cold Spring Harbor Laboratory. GenomeScope. <http://qb.csh.edu/genomescope>. Accessed 15 April 2020.
- Pacific Biosciences. pbbioconda. <https://github.com/PacificBiosciences/pbbioconda>. Accessed 1 March 2020.
- Kolmogorov M, Yuan J, Lin Y, et al. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 2019;**37**(5):540–6.
- Nurk S, Walenz BP, Rhie A, et al. HiCanu: Accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res* 2020;**30**(9):1291–305.
- Cheng H, Concepcion GT, Feng X, et al. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* 2021;**18**:170–5.
- Pacific Biosciences. pbipa. <https://github.com/PacificBiosciences/pbipa>. 2020. Accessed 12 September 2020.
- Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods* 2020;**17**(2):155–8.
- Guan D, McCarthy SA, Wood J, et al. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* 2020;**36**(9):2896–8.



24. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: Allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* 2018;**19**(1):460.
25. Simão FA, Waterhouse RM, Ioannidis P, et al. BUSCO: Assessing genome assembly and annotation completeness with singlecopy orthologs. *Bioinformatics* 2015;**31**(19):3210–2.
26. Kriventseva EV, Kuznetsov D, Tegenfeldt F, et al. OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res* 2019;**47**(D1):D807–11.
27. Sovic I: isovic/racon. <https://github.com/isovic/racon>. 2020. Accessed 2 March 2020.
28. Samtools. <http://www.htslib.org/>. Accessed 2 March 2020.
29. Pacific Biosciences. pbmm2. <https://github.com/PacificBiosciences/pbmm2>. 2020. Accessed 12 March 2020.
30. Li H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**(18):3094–100.
31. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: Architecture and applications. *BMC Bioinformatics* 2009;**10**(1) 421.
32. Geneious: Geneious | Bioinformatics Software for Sequence Data Analysis. <https://www.geneious.com>. Accessed 2 December 2020.
33. Bernt M, Donath A, Jühling F, et al. MITOS: Improved *de novo* metazoan mitochondrial genome annotation. *Mol Phylogenet Evol* 2013;**69**(2):313–9.
34. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol* 2011;**29**(1):24–6.
35. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;**12**(1):59–60.
36. Challis R, Richards E, Rajan J, et al. BlobToolKit – Interactive Quality Assessment of Genome Assemblies. *G3 (Bethesda)* 2020;**10**(4):1361–74.
37. Schell T, Feldmeyer B, Schmidt H, et al. An annotated draft genome for *Radix auricularia* (Gastropoda, Mollusca). *Genome Biol Evol* 2017;**9**(3):585–92.
38. Okonechnikov K, Conesa A, García-Alcalde F. Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* 2016;**32**(2):292–4.
39. Kingan SB, Urban J, Lambert CC, et al. A high-quality genome assembly from a single, field-collected spotted lanternfly (*Lycorma delicatula*) using the PacBio Sequel II system. *GigaScience* 2019;**8**(10), doi:10.1093/gigascience/giz122.
40. Duncan T, Kingan S B, Lambert CC, et al. A low DNA input protocol for high-quality PacBio *de novo* genome assemblies. *J Biomol Tech* 2019;**30**:S1–2.
41. Flynn JM, Hubley R, Goubert C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A* 2020;**117**(17):9451–7.
42. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. <http://www.repeatmasker.org>. Accessed 12 September 2020.
43. Stanke M, Keller O, Gunduz I, et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* 2006;**34**(Web Server):W435–9.
44. Huerta-Cepas J, Szklarczyk D, Heller D, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 2019;**47**(D1):D309–14.
45. Faddeeva-Vakhrusheva A, Derks MFL, Anvar SY, et al. Gene family evolution reflects adaptation to soil environmental stressors in the genome of the Collembolan *Orchesella cincta*. *Genome Biol Evol* 2016;**8**(7):2106–17.
46. Sun X, Ding Y, Orr MC, et al. Streamlining universal single-copy orthologue and ultraconserved element design: A case study in Collembola. *Mol Ecol Resour* 2020;**20**(3):706–17.
47. i5K Consortium. The i5K Initiative: Advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J Hered* 2013;**104**(5):595–600.
48. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in performance and usability. *Mol Biol Evol* 2013;**30**(4):772–80.
49. Kück P, Longo GC. FASconCAT-G: Extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Front Zool* 2014;**11**(1):81.
50. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009;**25**(15):1972–3.
51. Nguyen L-T, Schmidt HA, von Haeseler A, et al. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015;**32**(1):268–74.
52. Fjellberg A. The Collembola of Fennoscandia and Denmark Part II : Entomobryomorpha and Symphypleona. Brill; 2007.
53. Stach J. The Apterygotan Fauna of Poland in Relation to the World-Fauna of This Group of Insects. Family: Sminthuridae. Krakow: Institute of Systematics and Evolution of Animals, Polish Academy of Sciences; 1954.
54. Schneider C, Cruaud C, D’Haese CA. Unexpected diversity in Neelipleona revealed by molecular phylogeny approach (Hexapoda, Collembola). *Soil Org* 2011;**83**:383–98.
55. Bridge PD, Roberts PJ, Spooner BM, et al. On the unreliability of published DNA sequences. *New Phytol* 2003;**160**(1):43–8.
56. Seah YG, Ariffin AF, Jaafar T. Levels of COI divergence in Family Leiognathidae using sequences available in GenBank and BOLD Systems: A review on the accuracy of public databases. *Aquac Aquar Conserv Legis Int J Bioflux Soc* 2017;**10**:391–401.
57. Schneider C, Woehle C, Greve C, et al. Supporting data for “High-quality *de novo* genome from an ethanol-preserved specimen of *Sminthurides aquaticus*.” *GigaScience Database* 2021. <http://dx.doi.org/10.5524/100871>.
58. Schneider C, Woehle C, Greve C, et al. Supporting data for “High-quality *de novo* genome from an ethanol-preserved specimen of *Desoria tigrine*.” *GigaScience Database* 2021. <http://dx.doi.org/10.5524/100897>.