Article title:

A quantitative protocol for DNA metabarcoding of springtails (Collembola)

Authors:

Seikoh Saitoh,† Hiroaki Aoyama,† Saori Fujii, Haruki Sunagawa, Hideki Nagahama, Masako Akutsu, Naoya Shinzato, Nobuhiro Kaneko, Taizo Nakamori

1 Tropical Biosphere Research Center, University of the Ryukyus, 1 Senbaru, Nishihara 903-0213, Japan

2 Department of Environment and Natural Sciences, Graduate School of Environment and Information Sciences, Yokohama National University, Yokohama 240-8501, Japan

3 Okinawa Prefectural Agricultural Research Center, 820 Makabe, Itoman 901-0336, Japan

4 Department of Electrical Engineering and Computer Science, School of Industrial and Welfare Engineering, Tokai University, Toroku 9-1-1, Higashi-ku, Kumamoto 862-8652, Japan

†These authors contributed equally to this work.

Correspondence to:

Seikoh Saitoh, PhD

Center of Molecular Biosciences, Tropical Biosphere Research Center, University of the Ryukyus

Senbaru 1, Nishihara, Okinawa 903-0213, Japan
Phone: +81 (0) 98-895-8943; Fax: +81 (0) 98-895-8944

E-mail: h109296@comb.u-ryukyu.ac.jp
Abstract

We developed a novel protocol with superior quantitative analysis results for DNA metabarcoding of Collembola, a major soil microarthropod order. Degenerate PCR primers were designed for conserved regions in the mitochondrial cytochrome c oxidase subunit I (mtCOI) and 16S ribosomal RNA (mt16S) genes based on published collembolan mitogenomes. The best primer pair was selected based on its ability to amplify each gene, irrespective of the species. DNA was extracted from 10 natural communities sampled in a temperate forest (with typically 25–30 collembolan species per 10 soil samples) and 10 mock communities (with seven cultured collembolan species). The two gene regions were then amplified using the selected primers, ligated with adapters for 454 technology, and sequenced. Examination of the natural community samples showed that 32 and 36 operational taxonomic units (defined at a 90% sequence similarity threshold) were recovered from the mtCOI and mt16S data, respectively, which were comparable to the results of the microscopic identification of 25 morphospecies. Further, sequence abundances for each collembolan species from the mtCOI and mt16S data of the mock communities, after normalization by using a species as the internal control, showed good correlation with the number of individuals in the samples ($R = 0.91–0.99$), although relative species abundances within a mock community sample estimated from sequences were skewed from community composition in terms of the number of individuals or biomass of the species. Thus, this protocol enables the comparison of collembolan communities in a quantitative manner by metabarcoding.

Keywords: metabarcoding; Collembola; 16S; COX1; quantification
Introduction

Springtails (Collembola) are a major group of soil microarthropods that are widely distributed in terrestrial ecosystems from tropical to polar regions and are abundant, especially in forests (Hopkin 1997). Typically, 60–80 species of Collembola and a total abundance of $10^4$–$10^5$ individuals m$^{-2}$ are found in a forest site (Petersen and Luxton 1982). To date, more than 7,000 collembolan species have been described (Deharveng 2004). Most are omnivorous and mainly feed on detritus and microbes growing on it (e.g., Anderson and Healey 1972) and are preyed upon by predatory mites and larger arthropods (Johnson and Wellington 1980; Usher 1985; Lawrence and Wise 2000), forming an intermediate link in food webs during the decomposition process (Filser 2002).

Collembola are conventionally identified to the species level by the microscopic examination of their morphological features; therefore, the community assessment of this group has been time-consuming. Additionally, DNA barcoding (Hebert et al. 2003) of Collembola, using a standardized portion of the mitochondrial cytochrome c oxidase subunit 1 (mtCOI) gene, revealed that cryptic species were frequently found among morphologically identical specimens from different regions (e.g., Porco et al. 2012b), suggesting that conventional microscopic examination has confounded different species and underestimated their biodiversity. However, due to the required time and cost, it has not been practical to sequence every specimen by Sanger sequencing in community assessments where it is required to examine numerous microarthropods extracted from soil samples; typically, a 100 mL of bulk sample contains 100–1000 total animal specimens, including 10–30 species of Collembola (e.g.,
Takeda 1987). These problems associated with taxonomic identification impede the comparison of collembolan communities between regions and thus the understanding of the global biodiversity of Collembola.

The application of next-generation sequencing (NGS) technology to eukaryote community assessment has emerged in recent years, called DNA metabarcoding (Taberlet et al. 2012), wherein DNA fragments are amplified and sequenced from environmental or bulk community samples for genes whose sequences could be used as diagnostic features for identifying specimens to the species level. Metabarcoding methods have been developed for various body sizes of animal assemblages from terrestrial and aquatic ecosystems, such as protists (Nolte et al. 2010), nematodes (Porazinska et al. 2009), aquatic meiofauna (Fonseca et al. 2010), river benthic macroinvertebrates (Hajibabaei et al. 2011), and insects (Yu et al. 2012), the assessment of whose communities had confronted a problem similar to that in Collembola.

For Collembola, Ramirez-Gonzalez et al. (2013) developed a metabarcoding protocol using the mtCOI gene and showed that metabarcoding is promising for diversity assessment in this taxon. However, they also reported that its quantitativity was poor. They metabarcoded mock DNA samples in which known and different concentrations of DNA from collembolan species were mixed and found no or only a weak correlation between the number of reads (output) and DNA amount across species within a sample, indicating that one could not determine the abundances of the species within the community sample using their protocol. In addition to their study, other investigators (e.g., Porazinska et al. 2010; Hajibabaei et al. 2011; Yu et al. 2012) have often reported poor quantitativity with respect to species composition.

Poor quantitativity of current metabarcoding methods has been considered to be associated with
species-dependent factors, such as copy number of the target gene per cell, body size (i.e., copy number per individual), and mismatches between PCR primers and binding sites (Deagle et al. 2014). Several studies found that number of reads differed by up to four orders of magnitude between species, even when a sample comprising species of similar body size was sequenced (Porazinska et al. 2010; Elbrecht and Leese 2015), indicating an effect of the degree of primer mismatch between species on PCR bias. The PCR protocol employed in NGS library preparation also may affect the extent of skewing in the final product ratio during PCR from mixed-species templates (Mathieu-Daudé et al. 1996; Polz and Cavanaugh 1998).

The genes used for metabarcoding may also be involved in the problem. The mtCOI gene region currently widely used for metazoans (Hebert et al. 2003) is a protein-coding sequence and thus harbors highly diversified nucleotides among species at the third positions of codons throughout the region owing to synonymous mutations, leading to difficulty in designing specific PCR primer sets that possess sufficient specificity and sensitivity to the target taxon (Deagle et al. 2014; see Leray et al. 2013 and Brandon-Mong et al. 2015 for efforts made to overcome this problem). This aspect of the mtCOI gene is associated with the problem of primer mismatches and thereby quantitativity in metabarcoding. Nuclear ribosomal RNA (rRNA) genes are also widely used for metabarcoding (e.g., Porazinska et al. 2009, 2010; Fonseca et al. 2010; Hirai et al. 2015), and conserved regions in a taxonomic group may be found more easily in these than in the mtCOI gene (e.g., Machida and Knowlton 2012; Clarke et al. 2014). However, with respect to reference sequences required for taxonomic identification of metabarcoding sequences, a far larger database is available for the mtCOI gene (Ratnasingham and Hebert 2007) than for (nuclear or mitochondrial) rRNA genes. Thus, both coding and non-coding genes have pros and cons at present.
To develop a more quantitative method of metabarcoding for Collembola, we designed novel PCR primer sets for mtCOI and mitochondrial 16S rRNA (mt16S) genes and tested their performances in sequencing natural and mock collembolan community samples, employing a carefully designed laboratory protocol to minimize bias during PCR. Based on the results, we discuss the specificity and sensitivity of our method and its quantitativity in community assessment. In quantitative terms, we examined 1) whether relative species abundances in sequences in a mock community sample reflect those in individuals and 2) the correlation between sequence and animal abundances per species among community samples and the effect of normalization methods (using relative abundance in individuals or biomass or an internal control).

**Materials and methods**

*Site description*

The collembolans used in the present study were collected mainly from a temperate coniferous forest in the Kamigamo Experimental Forest of Kyoto University (225 m ASL, 135.77°E and 35.07°N), 10-km north of Kyoto, Japan, where the collembolan community has been extensively studied and more than 50 collembolan species have been recorded (e.g., Takeda 1987; Fujii and Takeda 2012; Saitoh et al. 2014). The mean annual temperature and annual cumulative precipitation during 2012 were 14.8°C and 1,485 mm, respectively. The canopy layer was dominated by Japanese cypress, *Chamaecyparis obtusa*. The soil was a moder type with a ca. 4-cm thick organic layer.
Collection of reference sequences

The collembolan specimens were collected by hand-sorting from soil microarthropods extracted from soil samples using Tullgren funnels as well as from culture populations of some species that were established from collected individuals. Individuals of some species were collected from the surfaces of soil, tree trunks, and mushrooms with an aspirator. Specimens were stored in vials with 99.5% ethanol at −20°C until use.

To construct a reference database linking metabarcoding sequences to morphological species, mtCOI and mt16S gene sequences were collected as follows: total DNA was extracted from collembolan specimens by a non-destructive method (Aoyama et al. 2015). After DNA extraction, the specimens were mounted on glass slides and identified to the species level based on morphological traits observed under a microscope according to the literature (Suma 2009; Tanaka 2010; Niijima and Hasegawa 2011; Hasegawa and Niijima 2012; Ichisawa 2012; Itoh et al. 2012; Hasegawa and Tanaka 2013; Furuno et al. 2014; Nakamori et al. 2014). From the DNA, mtCOI and mt16S gene regions were amplified and directly sequenced in both directions on an ABI 3130xl sequencer (Thermo Fisher Scientific, Waltham, Massachusetts, USA). The PCR primers and conditions used for construction of the reference database are listed in Table S1. When a sample was not amplifiable with an ordinary primer pair (Folmer et al. 1994) and the degenerate primers for Collembola (Nakamori 2013), amplification using other combinations of several available primers were tried so that the sequence of the gene region that was targeted in metabarcoding (to be described later) was obtained. We also sequenced the targeted gene regions from single specimens by multiplexing in a sequencing run of a 454 GS Junior sequencer (Roche, Basel, Switzerland) using a method (Aoyama et al. unpublished) analogous to that of Shokralla et al. (2014); the former method used ligation in
library preparation to add adapters for 454 technology to amplified gene fragments, whereas the latter utilized fusion primers for the purpose.

In total, mitochondrial COI and 16S RNA gene sequences were collected from 302 collembolan specimens assigned to 34 morphological species. A total of 99 and 134 unique sequences of mtCOI and mt16S were found, respectively. Of these, mtCOI sequences from nine species were already published as part of the study by Aoyama et al. (2015). We found that the 454 sequencer frequently yielded multiple sequence types from a single specimen (76 and 103 specimens out of 190 total specimens sequenced for mtCOI and mt16S genes, respectively), which were considered to be unlikely as PCR or sequencing errors because they were found in both directions of reads and from multiple conspecific specimens whose PCR reactions were constructed independently. Although these sequences presumably include potential pseudogenes termed as nuclear mitochondrial genes (numts) and true haplotypes, we could not yet completely determine which sequences are truly mitochondrial genes or numts; only a small fraction (7%) of the sequences contained symptom(s) for numt (i.e., insertion/deletion or stop codon) and were thereby considered pseudogenes. Translations were possible for most mtCOI sequences without a premature stop codon, and their sequences are often identical within a specimen. We considered that including numts into the reference database is useful for identification of their source organism when they are identified from metabarcoding data. Therefore, all unique sequences were included in our reference database.

Intra- and interspecific variation of the gene regions used for the metabarcoding method were evaluated using our reference database (Fig. S1). Dissimilarities between sequences were calculated by pairwise alignment (p-distance); insertions and deletions were considered as differences. As the size of our database was limited, published sequence
data were combined with our data to conduct this and the following analyses. We used mtCOI sequence data from Porco et al. (2012a), who revealed sequence divergences between cryptic species within cosmopolitan morphospecies of Collembola; reproductively isolated lineages found in their study were regarded as different species (in total 188 sequences from 28 (crypto)species from 16 morphospecies). For mt16S data, sequences from published complete collembolan mitogenomes (10 sequences from 9 species; see later subsection for the accession numbers) were combined because no other sequence deposited in a public database completely overlapped the region utilized in our metabarcoding method.

The percentage of gaps in the distributions of sequence divergences were observed to be approximately 10% in both mtCOI and mt16S genes (Fig. S1). Sequences of the mtCOI gene from different species differed by at least 15% (Fig. S1a,b), whereas those of mt16S differed by more than 17% (Fig. S1c,d). At most, 32.5% and 25.2% intraspecific distances were observed in the mtCOI and mt16S genes, respectively (Fig. S1b,d). Because considerable intraspecific distances were found in Onychiurus flavescens (17.0% in mtCOI and 18.6% in mt16S) and Neanuridae sp.1 (18.1% in mtCOL and 20.9% in mt16S), these morphospecies were suspected to include cryptic species (see below consideration on sequence divergence between collembolan species), although we found that in the other species, intraspecific variation >10% were in most cases due to potential numt sequences with large insertions and/or deletions.

Based on our reference database, threshold values for each gene were used for clustering sequences into operational taxonomic units (OTUs) in the data processing pipeline for metabarcoding data of the collembolan community (to be described later). The threshOpt function from the R package SPIDER (Brown et al. 2012) was
used to evaluate the effect of the threshold value for delimitation of species on false positive (over-splitting) and false negative (over-lumping) rates for optimization of the threshold (Meyer and Paulay 2005). Over-splitting decreased with increasingly different threshold values, but occurred until thresholds of 31% and 20% in mtCOI (Fig. S2a, b) and mt16S (Fig. S2c, d), respectively. Over-lumping of species occurred at a threshold value of ≥14% in mtCOI (Fig. S2a, b) and ≥15% in mt16S (Fig. S2c, d), respectively.

The success rate of taxonomic identification based on our database by the best close match method (Meier et al. 2006; using the bestCloseMatch function from the SPIDER package) was examined with a range of threshold values. A large proportion of sequences in the database were successfully identified by the other sequences in the same database at any threshold value. Misidentification of singleton sequences occurred at a threshold value of ≥15% difference in mtCOI (Fig. S3a, b) and ≥17% in mt16S (Fig. S3c, d), respectively. The same test was performed for identification to the family level, and it was revealed that no error occurred at ≤17% and ≤20% threshold values in mtCOI and mt16S, respectively (data not shown).

Based on these observations, the decision was taken to use a threshold of 90% sequence similarity to delineate collembolan sequences to OTUs so that sequences from different species might not be clustered into an OTU, although this threshold might divide sequences from the same currently recognized species into separate OTUs. We also decided to use a 90% sequence similarity for taxonomic assignment, as this value was considered safe from misidentification.

Design and selection of PCR primers for DNA metabarcoding of Collembola
DNA sequences of mtCOI and mt16S genes of 10 complete mitogenomes of Collembola were retrieved from the NCBI GenBank database (accessions AF272824, AY191995, Nardi et al. 2001, 2003; NC_006074, NC_006075, Cook et al. 2005; EU016192, EU016194-EU016196, EU084034, Carapelli et al. 2007, 2014; EU124719, Torricelli et al. 2010), consisting of nine species from seven families and including the most distantly related families (Xiong et al. 2008). The sequences were aligned with ClustalW (Larkin et al. 2007) to identify conserved regions in Collembola to design candidate PCR primers for metabarcoding. Degenerate primers for the mtCOI gene were designed within the standard barcode region for animals so that public sequence databases can be used as the reference database in the taxonomic identification of the metabarcoding data, whereas primers for the mt16S genes were not allowed to completely overlap regions that had been used in previous studies (e.g. Schneider et al. 2011).

Primer nucleotides were determined according to Boyce et al. (2009), who suggested that only small parts of oligonucleotide primers should be degenerate (typically up to three of the nine bases at the 3’ ends become ambiguous) and that the remaining part should be determined based on the most abundant nucleotide at each position in the alignment. In total, 9 forward and 12 reverse primers for mtCOI and five forward and six reverse primers for mt16S were designed (data not shown).

All possible combinations of the primers that were expected to amplify 250–500-bp gene fragments were tested to select the best primer pair for each gene based on their ability to amplify the gene, irrespective of collembolan species. For this screening, we used seven species from five families: Desoria sensibilis, Folsomia octoculata, Tetracanthella sylvatica (Isotomidae), Megalothorax minimus (Neelidae), Oncopodura yosiiana (Oncopoduridae), Onychiurus flavescens (Onychiuridae), and Sinella dubiosa (Entomobryidae). Based on the results, the primer pairs
Col_mtCOI_382F (5′-acaggatacaagtttayccnccB3′) and Col_mtCOI_695R (5′-gggtetctctecnngg-3′), and of
Col_mt16S_39F (5′-accgtgcaaaggtagcB3′) and Col_mt16S_523R (5′-aagatagaaccaacctggcB3′), were determined to
be the best pairs for the mtCOI and mt16S genes, respectively. The expected sizes of the amplicons were 314 bp for
mtCOI and ca. 400–490 bp for mt16S (depending on species).

Preparation of metabarcoding library for 454 technology

Bulk microarthropod community samples (Fig. 1.1) to be analysed by this protocol were composed of specimens
extracted from 100 mL of core-sampled soil using a Tullgren funnel, each of which typically consisted of 100–
1,000 animals comprising primarily Collembola and mites. Natural community samples that were sampled in this
manner and artificially comprised mock community samples were analysed (see later subsections); samples were
stored in vials with ethanol as a preservative at −20°C until use. Samples were transferred to 1.5-mL PCR tubes,
and ethanol was removed prior to DNA extraction. Total DNA was extracted from the community samples using
DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions except that
microarthropods were milled using a micro homogenizer (BioMasher II; NIP, Tokyo, Japan) in lysis buffer. The
extracted DNA samples were quantified using a Qubit 2.0 Fluorometer with a Qubit dsDNA HS assay kit (Thermo
Fisher Scientific). The target gene region (mtCOI or mt16S) was amplified for each community sample using the
above primer pair tailed with four-base multiplex identifier (MID) tags at the 5′-end of each primer (listed in Table
S2). PCR was performed in a 20-µl reaction volume using a KAPA HiFi HotStart ReadyMix kit (2-ng DNA
template, 10 µl of 2× KAPA HiFi HotStart ReadyMix, 0.3 µM of each primer; KAPA Biosystems, Wilmington,
Massachusetts, USA). The PCR conditions were as follows: 95°C for 2 min; 15–25 cycles at 98°C for 20 s, 53°C for 30 s, and 72°C for 30 s; followed by a final extension step at 72°C for 1 min. To keep PCR amplification from reaching a plateau, the PCR cycles to be used were determined for each sample beforehand by a preliminary examination in which the band intensities of the PCR products from several distinct cycles (e.g., 15, 18, 21, and 24) were visually compared using electrophoresis to identify the range of cycles where PCR was yet to reach a plateau for the sample (Fig. 1.2.1; we recommend fewer than three cycles of difference in the PCR cycles between the samples to be analysed simultaneously). The PCR products were purified by size selection for the expected amplicon size using an SPRIselect reagent kit (200–500 bp for mtCOI; 300–700 bp for mt16S; Beckman Coulter, Brea, California, USA). The purified amplicons were checked with an Agilent 2100 Bioanalyzer with High Sensitivity DNA kit (Agilent Technologies, Santa Clara, California, USA) for evaluation of their size distributions and thereby quantification of their molecular concentration by Agilent 2100 Expert Software (Fig. 1.3.1). The amplicons to be sequenced simultaneously in a sequencing run were pooled into molar concentrations equal between samples. The pooled amplicons were ligated with adapters for 454 technology using a GS Rapid Library Preparation Kit (Roche) to prepare a single library (Fig. 1.4.1). After further purification by size selection using the SPRIselect reagent kit, the library was quantified with KAPA Library Quantification Kit (Kapa Biosystems). Sequencing of the library was performed with a 454 GS Junior sequencer (Roche).

Sequence data analysis

A scheme of the strategy to cluster read sequences from metabarcoding data in the present study is shown in
Figure S4. An outline of the current data processing pipeline is as follows: As the present protocol added NGS adapters to amplicons by ligation, the target gene was sequenced in forward and reverse directions. After quality filtering of raw reads, calling of OTUs was separately conducted for forward and reverse reads using the open-reference method (QIIME pipeline version 1.8, Caporaso et al. 2010), in which reads were clustered with reference sequences using a threshold (97%), and novel OTUs were selected from remaining reads that were not clustered with the reference sequences. The following step merged forward and reverse reads belonging to identical OTUs. For OTUs clustered with reference sequences, this was conducted by sequence mapping of the representative sequences onto reference sequences; newly selected forward and reverse OTUs were merged by exploring similar sequences between them. As we observed that the first clustering by a 97% threshold often clustered identical sequences into separate OTUs due to sequencing errors, sequence variation generated by degenerate primers and MID-tag sequences on 3′-ends, a less conservative threshold (i.e., 94%) was applied in the above-mentioned merging steps to eliminate the artifacts. Finally, the merged OTUs were further clustered into 90% OTUs which we currently assumed as an analogous level to species, as described below.

Reads sequenced in the forward and reverse directions were separated according to the primer sequences. Using the QIIME pipeline, reads were processed (with script split_libraries.py) prior to the calling of OTUs: the primer sequences were removed; the low-quality bases were trimmed from the 3′-ends (using a sliding window of 50 bases, truncated from first base of window when the mean QV was <25); the numbers of allowed mismatches with primer sequence and allowed homopolymer lengths were two and 12 bases, respectively. After trimming, reads of length <200 bases were discarded.
OTU calling was first performed separately for the forward and reverse reads. Open-reference OTU picking (for 97% of OTUs) was performed with reference sequences (see above; sequences were trimmed to insertion length used in metabarcoding) using the pick_open_reference_otus.py script.

The 97% OTUs were secondarily grouped into OTU groups that were analogous to 90% OTUs (Fig. S4): (1) representative sequences of OTUs were mapped onto reference sequences using GMAP software (Wu and Watanabe 2005). An OTU whose representative sequence was aligned to a reference sequence (95% of sequence similarity, ≥150 bp of alignment, >70% of query coverage, and <60 bp of clipped length of query) was considered to be a member of a group represented by the reference sequence. Note that during this step, it was intended to limit sequence variation between the furthest members of the cluster to <6%. To achieve this empirically, a slightly stricter threshold (i.e., 95% similarity) was required when clustering was conducted by sequence mapping. (2) For unmapped OTUs, similarities between their representative sequences were evaluated by pairwise alignment by ClustalW. Based on the result, these OTUs were clustered using the furthest neighbour method (Sørensen 1948) and grouped by threshold similarity of 94%. This grouping was made separately for forward and reverse OTUs at first, and forward and reverse groups were later merged when a reciprocal best alignment with 94% was found between the two groups. For each group, a representative sequence of the OTU with the most abundant member reads was selected as the representative sequence of the group. (3) Finally, representative sequences of these groups (constructed with GMAP or ClustalW) were pairwise aligned and clustered by the furthest-neighbour method. They were thereby further grouped into OTU groups by a threshold similarity of 90%. When an OTU group included a reference sequence, it was determined as the representative sequence. Otherwise, a representative sequence of the
OTU with most abundant member reads among the group was selected. These OTU groups by 90% similarity are here referred to as 90% OTUs.

Taxonomy of the OTUs was inferred via BLAST searches \((\text{E} < 10^{-12}, \text{Zhang et al. 2000})\) against the NCBI nucleotide database as well as reference sequences collected in the present study. For OTUs with the best match to a collembolan sequence, we determined their taxonomy using the following (tentative) thresholds: a hit with query coverage >90% and sequence identity >90% was considered to belong to the same species; with query coverage >80% and sequence identity >85% as indicators of belonging to the same family. For OTUs with the best hit to a taxon other than Collembola, the orders were inferred when query coverage >80% and sequence identity >75%.

See earlier subsection “Collection of reference sequences” for origins of these threshold values.

**Quality assessment using natural community samples**

The metabarcoding method presented in this study was tested with respect to its selectivity of PCR primers and sensitivity to comprehend collembolan biodiversity. Topsoil was sampled in a plot \((10 \times 25 \text{ m})\) consisting of 10 adjacent subplots of \(5 \times 5 \text{ m}\), established in the Kamigamo Experimental Forest. Using cylindrical cores \((4 \text{ cm in depth, } 25 \text{ cm}^2 \text{ in area, and } 100 \text{ cm}^3 \text{ in volume})\), a pair of soil samples was collected from a randomly determined point in each subplot \(\text{i.e.}, \text{the two soil samples had been adjacent in the forest floor}\). Microarthropods were extracted from the soil samples via Tullgren funnels. Half of the community samples \(\text{one sample per subplot; in total } 10 \text{ samples; samples N}_01 \text{ to N}_10\) were analysed by our metabarcoding method, and the remaining half \(\text{C}_01 \text{ to C}_10\) were mounted on glass slides and examined under a microscope to identify Collembola by species
and count.

Species accumulation curves with increasing number of samples were drawn based on the number of
morphospecies (C_01 to C_10) or 90% OTUs (N_01 to N_10) using specaccum function in the R package vegan
(version 2.3-4; Oksanen et al. 2016). The correlation of relative species abundances (mean values of 10 samples)
between the three methods were evaluated using the Pearson product–moment correlation coefficient. A total of 13
species found by all the three methods were included in this correlation analysis.

Quality assessment using mock community samples

Mock community samples were analysed by the metabarcoding method to test its quantitativity. A total of seven
species of Collembola: Pseudosinella pseudolanuginosa, S. dubiosa, Sinella umesaoi (Entomobryidae),
Hypogastrura reticulata (Hypogastruridae), O. flavescens (Onychiuridae), Arrhopalites japonicus (Katiannidae),
and Folsomia candida (Isotomidae), were used to prepare mock community samples. The first six species were
included in different numbers among the mock samples, while F. candida was included in the same number as an
internal control (see below).

Individuals of these species were obtained from laboratory cultures maintained in Yokohama National University
except for H. reticulata, of which individuals were collected from a wild population swarming on a mushroom in
the Kamigamo Experimental Forest. Laboratory cultures of the other collembolan species were established from
individuals collected from the forest, with the exception of F. candida, which were from Fuji-Yoshida (central
Japan; Itoh et al. 1995). The mean dry weights of the collembolan specimens were indirectly estimated based on
body length (from measurement of at least 20 specimens per species) according to Tanaka (1970) and Petersen (1975). Information for the specimens is presented in Table S3.

The species compositions of mock community samples were determined as follows: collembolan species (see above) were randomly permuted to be placed in an order (from 1 to 6) according to which their numbers of individuals were determined to be included in a sample (32, 16, 8, 4, 2, or 1 individuals). To construct 1000 community matrices (each consisting of six species by 10 samples), this permutation was performed 1000 times per sample (thus, 10,000 times in total). Community matrices not satisfying the conditions that a species must be ranked any place (from 1 to 6) in at least one sample and ranked in the same place in at most three samples were discarded. One of the remaining matrices was randomly selected to examine the species compositions of the mock community samples by metabarcoding. According to this matrix (consisting of six species by 10 samples), each sample received different numbers of individuals from the six species and three individuals of *F. candida* as an internal standard. In total, each of the 10 mock community samples (S_01 to S_10) was composed of 66 individuals of Collembola (Table S4).

The correlation between the read counts and numbers of individuals across samples was calculated for each species, by three separate normalization methods: (1) both the read counts and number of individuals were normalized to relative abundance per sample; (2) the read counts were normalized to the relative abundance, while the number of individuals was converted to relative biomass; (3) the read counts were normalized to the internal control (i.e., divided by read counts per individual of the control species), while the number of individuals was left unchanged. The Pearson product-moment correlation coefficient was used to evaluate the correlations.
Mock community samples were ordinated using principal component analysis (PCA) based on the number of individuals or normalized number of reads (by the internal control) from metabarcoding datasets. To ordinate samples based on the relative difference of abundance between samples, the data were standardized to a mean of 0 and standard deviation of 1 across samples for each species and input to this analysis.

The sequence data obtained in this study have been submitted to the DDBJ Sequence Read Archive (DRA) as accession no. DRA004529.
Results

Sequencing throughput from metabarcoding libraries

Sequencing throughput using mtCOI or mt16S library from natural or mock community samples (in total four sequencing runs; 10 samples per sequencing run) is described in Table S5. From 107,043 to 187,573 reads were sequenced per library. The number of quality-filtered reads varied at most 7.7 times between samples in a sequencing run, and at least 1,690 quality-filtered reads were obtained per sample (Fig. S5). Of these, from 574 to 26,964 per sample (24,262 to 133,522 reads per library) were aligned to 97% OTUs (and finally to 90% OTUs).

Quality assessment using natural community samples

The soil microarthropod community in the study site was dominated by Collembola and Acari (Prostigmata, Mesostigmata, and Oribatida) (300 to 1000 animal individuals per 100 mL soil sample, identified under a microscope, Fig. 2a). The taxonomic composition of sequences obtained from the metabarcoding libraries (inferred via BLAST searches, Fig. 2b, c) showed that the largest part of the output sequences were derived from Collembola (47% and 65% reads from mtCOI and mt16S libraries were aligned to collemboulan OTUs, respectively; Fig. S6a, c). Nonetheless, animal taxa other than Collembola accounted for a considerable portion of reads in several samples (e.g., Oribatida in sample N_01 in mtCOI, Hymenoptera in N_01 and N_06 in mt16S, Coleoptera in N_08 sample in both libraries). From the 10 natural community samples, 32 and 36 collemboulan OTUs (10–33 and 9–36 per sample) were found in the mtCOI and mt16S libraries, respectively, whereas a total of 25 species (6–13 per sample) were found by morphological identification (Fig. 3).
Taxonomic assignment of these collembolan OTUs linked 18 mtCOI and 22 mt16S OTUs to 15 and 16 morphological species, respectively. Several of these OTUs were considered to be derived from the same morphological species. For example, three mtCOI and four mt16S OTUs were assigned as *F. octoculata* (see Table 1). When such OTUs were combined under their assigned species, the difference between the numbers of morphological species and OTUs became smaller (25 morphological species vs. 29 and 30 combined OTUs under assigned species).

The collembolan community assessed by microscopic examination as well as metabarcoding of paired samples is summarized by species in Table 1. Of 25 species found in the 10 natural community samples by microscopic examination, a total of 15 species had been barcoded in our current reference database (Table 1a). These barcoded species represented 97% of the collembolan individuals found. In the metabarcoding data, the OTUs identified to the species level (see above) accounted for 93 and 97% of collembolan reads from the mtCOI and mt16S libraries, respectively (Table 1b, c). A total of 13 collembolan species were found by both microscopic identification and metabarcoding using mtCOI and mt16S libraries. The relative abundances (mean values across samples) of these 13 species, which represented >90% of collembolan individuals or reads of any datasets, were positively correlated (microscope vs. mtCOI, $R = 0.87, P < 0.001$; microscope vs. mt16S, $R = 0.91, P < 0.001$; mtCOI vs. mt16S, $R = 0.79, P < 0.01$; Fig. S7).

Quality assessment using mock community samples

Almost all the reads were derived from target gene regions of Collembola when metabarcoding libraries of mock
community samples consisting of seven collembolan species, although a larger number of OTUs than species were found in both libraries (Fig. S6b, d). Thus, in addition to the seven major OTUs that had abundant reads and were considered to represent individual species, extra OTUs were found that were also considered to be derived from some of these species (92.6 to 95.9% sequence identity with the representative sequences) but accounted for only minor proportions of reads from the species (mtCOI data: 12, 7, and 0.5% in *P. pseudolanuginosa*, *S. umesaoi*, and *H. reticulata*, respectively; mt16S data: 4 and 0.7% in *P. pseudolanuginosa* and *H. reticulata*, respectively).

The relative species composition of each mock community sample is depicted in Figure 4 by the number of individuals, biomass and read counts in the mtCOI and mt16S datasets; the relative species compositions based on read counts reflected for the most part those based on the number of individuals and biomass (Fig. 4). The relative abundance of each species based on read count correlated to that of the number of individuals across samples, irrespective of genes or species (Fig. S8a, mtCOI: $R = 0.81–0.98$, $P < 0.01$ for each; Fig. S8c, mt16S: $R = 0.84–0.98$, $P < 0.01$ for each), except for *F. candida* (the species included in the same number in the samples). Correlations of the relative abundance of reads to the relative portion of biomass were almost identical to those with the relative portion of individuals described above (Fig. S8b, mtCOI: $R = 0.81–0.99$, $P < 0.01$ for each; Fig. S8d, mt16S: $R = 0.82–0.98$, $P < 0.01$ for each). When all the species were pooled, correlations were lower (mtCOI, $R = 0.66$ and 0.71, and mt16S, $R = 0.74$ and 0.70, to relative abundance and to relative portion of biomass, respectively) than those separately by species examined above.

Next, another method of normalization using a species as internal standard was assessed. We note that, although *F. candida* was originally intended to be used as the internal standard, this species was less abundantly sequenced
from the mtCOI library (0–0.19% of total reads per sample) than from the mt16S library (0.79–3.5%). However, no read from *F. candida* was found in one of the samples (S_06) in the mtCOI data, rendering this species unsuitable for use as internal standard. Accordingly, *P. pseudolanuginosa* was selected as the internal standard in place of *F. candida*, after identification of the species yielding the highest correlation coefficients between the numbers of individuals and the reads when used as internal standard (Table S6). After this normalization, every species showed strong positive correlation between the number of individuals and the normalized read counts ($R = 0.91–0.99$ and $R = 0.92–0.99$ in mtCOI and mt16S libraries, respectively; $P < 0.001$ for each), whereas the slope of the regression model varied depending on both the species and gene ($\beta = 0.04–0.6$ and $\beta = 0.03–1.0$ in the mtCOI and mt16S datasets, respectively) (Fig. 5). When data from all the samples and species (excluding internal control species) were pooled, the correlation between normalized count and number of individuals was moderate ($R = 0.62$, $P < 0.001$), as shown above for the other normalization methods. The PCA ordination analysis of mock community samples based on the number of individuals or reads showed that four species differentiated community samples along axis 1, whereas two other species differentiated community samples along axis 2, resulting in very similar ordinations of samples between the three datasets (Fig. 6).

**Discussion**

We designed two novel PCR primer pairs for metabarcoding so that target gene regions (mtCOI and mt16S) could be amplified from Collembola irrespective of species and employed a library preparation protocol to achieve improved quantitativity in sequencing output. The results from the assessment of the collembolan community of a
forest site using the present metabarcoding method were comparable, in terms of the species composition and biodiversity, to those obtained from paired soil samples by the conventional method (taxonomy identification of each animal under a microscope). Further, sequencing using mock community samples by the metabarcoding method revealed that the abundance of sequence reads from a species showed a strong linear relationship ($R = 0.91–0.99$) to the number of the individuals included in samples, especially when normalized to an internal control.

These results indicated that this method is applicable to surveys of Collembola biodiversity.

**Selectivity and sensitivity assessed using natural community samples**

The present results from metabarcoding of natural community samples could be evaluated in terms of specificity for the target taxon (Collembola) as well as capability to comprehend the species within the taxonomic group. Specificity of the metabarcoding primers is important in the sequencing of a bulk soil microarthropod community sample, extracted via Tullgren funnel, that includes a broad range of taxa of animals such as mites, small insects, and other microarthropods as well as Collembola (see Fig. 2a). If the selectivity of metabarcoding PCR primers for Collembola is poor, amplicons derived from non-target taxa included in such a sample may consume the capacity of a sequencer and reduce the space for Collembola. The present result indicated that the largest part of the sequences from metabarcoding libraries was successfully derived from Collembola. Nonetheless, sequences from other taxa (Oribatida, Coleoptera, or Hymenoptera) were dominant in a few samples, indicating that the primer pairs designed in the present study could amplify the target genes occasionally from these undesired taxa. We speculated, based on the sequences, that the oribatid sequences were from species abundant at the site.
(Tectocepheidae and Camisiidae; Hishi and Takeda 2008), whereas coleopteran and hymenopteran sequences were from (likely a larva of) Tenebrionidae and ants (Formicidae), respectively. Considering that, apart from oribatid mites, these taxa of insects were rarely found under a microscope (only two larvae of Coleoptera and three ants were found in the 10 samples by microscopic examination), their influence on sequencing output per individual was far larger than that of Collembola, likely because of their large body masses compared with mesofaunal taxa (1–20 µg vs. several mg per individual; Swift et al. 1979). If a specific organism is a frequent contaminant then a blocking primer (Vestheim and Jarman 2008) may be used to prevent it from being amplified, although this is not useful for most soil microarthropod community samples in which specific contaminant organisms cannot always be expected. Thus, if the capacity of a sequencer is limited, it is recommended to remove animals of these taxa (easily distinguishable from the others) from microarthropod community samples before DNA extraction.

The assessed collembolan community at the study site was comparable between microscopic identification and metabarcoding of paired samples: the numbers of morphological species and OTUs after taxonomy assignment were in agreement (25 vs. 29 or 30, Table 1). Although the current reference sequence database still did not fully cover the collembolan species in the study site, it covered most of the individuals in the community. Similarly, only approximately half of the collembolan OTUs were identified to the species level, but these OTUs accounted for most of the collembolan reads in both the mtCOI and mt16S datasets. These barcoded species and identified OTUs were largely consistent in terms of relative abundance (Fig. S7), although metabarcoding datasets are somewhat biased (see below); heterogeneity between paired samples (which could not be isolated in this study) also may have caused some differences between microscopic and molecular methods. The high species-dependent variation in the
abundance of Collembola in the natural community (the exponential decrease in proportion to total individuals explained by a species with increasing rank order of the species based on abundance) may have masked the potential effect of PCR bias. These results nonetheless indicate that, comparable to conventional microscopic identification, the metabarcoding method in the present study could comprehend the biodiversity of a collembolan community in a survey of a natural community.

Our current sequence data-processing pipeline employs a threshold value of 90% sequence identity to delineate OTUs, which were assumed to be comparable to species in present study. While >3% of difference in COI sequence (i.e., <97% of identity) has been usually considered to indicate distinct species in various taxa (Hebert et al. 2003), greater than 3% difference within morphological species has been previously reported in Collembola (e.g., Porco et al. 2012a, 2014; Anslan and Tedersoo 2015) and was found in the present study. As distributions of inter- and intraspecific sequence similarities overlapped (Fig. S1), observed sequence variations within a species may have comprised those by cryptic species as well as true intraspecific diversity, suggesting a need to determine a threshold to distinguish the two. Although Porco et al. (2014) suggested a threshold of 14% difference in the mtCOI gene for this purpose, we used a stricter value (10%) to cluster metabarcoding sequences, considering that it would be best to avoid including sequences from different species into an OTU, and afterwards to combine separate OTUs if they were assigned to the same species according to a reference database. Nonetheless, this threshold should be reconsidered when different morphological or cryptic collembolan species with a smaller genetic distance than the current threshold are found in future studies.

The current reference sequence database (of 34 species that were locally collected) and public databases (NCBI
NT and BOLD databases) could identify only a few more than half of the collembolan OTU groups from metabarcoding of the natural community. This outcome was likely due to the poor coverage for rare species, each of which accounts for less than 1% of the total collembolan individuals in the community, as well as rare types not yet barcoded from dominant species. Further, of the collembolan OTUs that were identified to species, only one mtCOI was identified to a species barcoded elsewhere (*Sminthurinus trinotatus*, KJ155516), whereas the other identified OTUs were derived from those barcoded at the study site. Even for species that had been barcoded at other sites, their sequences were different from those collected from identical morphological species at the study site (i.e., <90% of sequence similarity). These results support the huge and hidden diversity in Collembola shown in previous studies (e.g., Porco et al. 2012a,b; Cicconi et al. 2013) and suggest the need for generating more reference sequences of this group, more extensively at our study site as well as in various other geographic regions.

At present, the construction of a locally collected database for mtCOI and mt16S genes will be needed to identify collembolan species from metabarcoding data. Nevertheless, even in the absence of associated Linnaean names, our analysis pipeline would permit the quantification of biodiversity and the study of compositional heterogeneity among collembolan samples.

The observed intraspecific sequence diversity could be partly explained by potential pseudogenes co-amplified from the nuclear DNA (nuclear mitochondrial pseudogenes, numts; Song et al. 2008) along with the target genes. By manual examination of representative sequences of OTUs, we found some sequences similar to known species but with substitutions, insertions, and/or deletions, resulting in >10% of difference from the known species, and accordingly, clustered with separate OTUs. Such OTUs had only small numbers of member reads. For example,
one of three mtCOI OTUs from *F. octoculata*, which had a deletion and stop codon and was thus considered to be a numt, explained < 1% of reads derived from this species. A similar observation was reported by Ramírez-Gonzalez et al. (2013), who previously metabarcoded a collembolan community with 454 technology. We also often found multiple types of mtCOI and/or mt16S sequences from a single collembolan specimen when sequencing the amplicons of these genes by 454 (data not shown), supporting the possibility of co-amplification of pseudogenes by PCR. Indeed, it should be difficult (especially for mt16S, a non-coding gene) to determine based only on sequencing data from NGS whether a suspected sequence is numt or merely due to errors during PCR and/or sequencing. It is, of course, also necessary to consider carefully other possibilities: heteroplasmmy, female sperm storage, gut content (Collembola often consume conspecific corpses and exuviae of others), and contamination. A patch for this problem could be to include (suspected) numts as well as definitely identified target genes in the reference database so that an analyst can determine the taxonomy of the source organism and thereby avoid overestimation of biodiversity.

*Quantitativity assessed using mock community samples*

As shown in the results from the metabarcoding of mock community samples, very high linearity between numbers of collembolan individuals and sequences per species (i.e., $R = 0.91–0.99$) was achieved by the present method when the read counts were normalized by using a species as the internal control. On the other hand, the slope of the regression line varied greatly depending on the both species and the gene region (mtCOI and mt16S; Fig. 5) used in metabarcoding. This coefficient was considered (from the method of normalization) to represent the ratio of the
number of reads obtained per individual of the species to that of reads from the standard species, and thus, represent the biases between the species (including those by PCR and others) under the PCR conditions used in library preparation for NGS. The observed linearity between normalized read count and number of individuals of each species across samples suggests what may have occurred during PCR. That is, ratios of amplicons between species were, likely owing to different PCR efficiencies, skewed from those of their templates with increasing PCR cycles, though the extent was nonetheless similar among samples. Further, the observation that the slope varied with both species and gene (thus, primers) indicates the effect of different extents of mismatches between genomic sequences and the primers at the binding sites (Deagle et al. 2014) as well as that of copy number of the mitogenome in each individual.

These observations are, in some respects, consistent with previous studies in which metabarcoding methods were developed. Within a sample, different species yielded highly different numbers of sequences even between specimens with similar body weight (Porazinska et al. 2009; Hajibabaei et al. 2011; Yu et al. 2012; Elbrecht and Leese 2015); Elbrecht and Leese (2015) showed a high correlation between the numbers of reads and the biomass by metabarcoding a mock sample consisting of specimens of the same species but of different haplotypes and body weights. Similarly to the present study, Diaz-Real et al. (2015) assessed the correlation between the relative species abundances of read counts and the numbers of individuals using mock community samples and found strong correlations in two focal species but with different slopes of the regression lines.

Some of the authors of these previous reports concluded that metabarcoding might not be quantitative, from the observation that PCR bias strongly skewed the relative abundances between species and made it impossible to
recover the composition of individuals (or biomass) within a sample from sequencing data (Porazinska et al. 2009; Yu et al. 2012; Elbrecht and Leese 2015). Accordingly, it is considered inappropriate to use the relative species abundances in read counts, in place of the numbers of individuals, to calculate alpha diversity indices of a community sample. Some authors, accordingly, recommended transformation of metabarcoding data into presence/absence prior to statistical analyses (Yu et al. 2012; Elbrecht and Leese 2015). From this point of view, the present method was also shown still to be non-quantitative, given that it could not recover relative species abundance in a sample, owing to observed bias between species. Nonetheless, compared with the previous study by Ramirez-Gonzalez et al. (2013) who had metabarcoded Collembola and reported no or weak correlation between relative species abundances in mock community samples with read counts from metabarcoding data ($R = 0.03–0.47$ when pooling all the species), the present method considerably improved the ability to recover relative species abundance ($R = 0.66–0.74$, Fig. S8). Furthermore, the present method may become possible in the future when slope values (ratio to a standard species) under a standard protocol are accumulated for various species. This idea is similar to that described in a recently published article by Thomas et al. (2016) who suggested the use of a coefficient for correction of the read abundance of focal species to standard species in metabarcoding a tissue mixture of multiple species.

Even though the relative species abundance within a sample could not be determined by metabarcoding, the linearity of normalized read counts to individuals across samples in the present method allows us to discuss the relative abundance of species among samples in a quantitative manner. Typically, the abundance of focal species can be compared between samples using normalized read counts, much as one does by the number of individuals.
This comparison is important for ecologists who wish to evaluate the variation of population densities of the focal species among environmental factors, between treatments, and/or with time. Furthermore, by making use of linearity between individuals and read counts, ordination of community samples via principal component analysis (PCA) can be applied to a community matrix based on sequence data. As an example, the mock community samples were ordinated based on individuals and OTUs from read counts from metabarcoding data (Fig. 6). Prior to the analysis, the abundance (individuals or normalized read count) was standardized (to a mean value of 0 and standard deviation of 1) between samples per species in every dataset (that is, only information about relative difference between samples was conserved for each species). As a result, the PCAs showed almost identical ordinations of communities by respective principal components 1 and 2. Note that, since the species included in these mock community samples were fully overlapped, no ordination method was applicable when transformed into presence/absence data.

Selection of the species to be used as internal control is likely to be important. We had originally intended to use *F. candida* for the purpose, and included three individuals of this species per sample. However, this species was very infrequently sequenced, especially from the mtCOI library, and unfit for use as the standard. Accordingly, in place of *F. candida, P. pseudolanuginosa* was used as an internal control as this species showed the most preferable result in terms of the correlation coefficient per species across the samples. From the observation that *P. pseudolanuginosa*, the best internal control species, was the most frequently sequenced species in the mock community samples (see Fig. 4c, d), we speculate that the abundant templates and amplicons of this species minimized the effect of stochasticity during PCR (Polz and Cavanaugh 1998) and sequencing on the read count.
(Amend et al. 2010; Elbrecht and Leese 2015), resulting in the observed good performance as an internal control species. This observation suggests a recommendation for the use of sufficient numbers of individuals (tentatively, >10 per sample) of sufficiently amplifiable (and cultivable) species such as *P. pseudolanuginosa* as the internal standard.

The proportion of read counts of a species to the number of total reads (i.e., relative read abundance) is an often-used normalization standard for metabarcoding data (e.g., Hajibabaei et al. 2011; Hirai et al. 2015). Although, in the present results, this normalization method also showed high or moderate correlation (Fig. S8), its performance was inferior to that by internal control. This is likely because the relative read abundance of a species in a community sample was affected by other species included in the sample (Fig. 4). As pointed out by Amend et al. (2010), when a focal species with high PCR efficiency is sequenced together with another species with low efficiency, the abundance of the focal species may be overestimated, to an extent that may differ depending on the number of individuals of the less efficient species are included in the sample. Accordingly, even when the same number of individuals of a species is included in multiple samples, its relative abundance may vary between these samples.

The satisfactory results of the present study were most likely due to the laboratory protocol used, which was designed with care to minimize the effects of PCR amplification. We did not use fusion primers, which have gene-specific sequences as well as adapters for the NGS platform and could bias amplification in PCR (Berry et al. 2011). In addition, we had previously designed and examined fusion primers but observed frequent primer dimers and unstable amplification depending on incorporated MIDs in another project (Sunagawa et al. unpublished data).
Accordingly, we instead used ligation to add NGS adapters to amplicons, similar to the practices of De Barba et al. (2014) and Hirai et al. (2015), the latter of whom reported good linearity between relative species abundances in read counts and those in biomass. Nonetheless, the effect of methods of adding NGS adapters on linearity is unclear, given that Diaz-Real et al. (2015) and Elbrecht and Leese (2015) also reported high linearity using fusion primers.

The present study employed a restricted number of PCR cycles (15–25, determined by preliminary examination for each sample; Fig. 1.2.1) to avoid a plateau during PCR, which could affect final proportions of amplicons from different templates (Mathieu-Daudé et al. 1996). Hirai et al. (2015) also employed small numbers of PCR cycles (22), whereas Diaz-Real et al. (2015) and Elbrecht and Leese (2015) used 40 (with touching down) and 30 cycles, respectively, but all of these authors reported good results. Thus, the effect of PCR cycles is also unclear. To make the present protocol more robust, PCR replication (Porazinska et al. 2010; Diaz-Real et al. 2015; Elbrecht and Leese et al. 2015) and longer extension time (30 s in the present study vs. 1 min or longer in other studies) could be incorporated to minimize the effect of stochasticity on rare templates (Polz and Cavanaugh 1998) and formation of chimeras (Lahr and Katz 2009), respectively, although these did not apparently interfere with the sequencing results in the present study.

Costs and prospective

The cost per sample (for reagents required for DNA extraction, library preparation and sequencing) using the present method was estimated to be approximately $100 for sequencing 10 collembolan community samples at once using the 454 GS Junior sequencer, a cost presumably acceptable for use in place of conventional microscopic
identification. As NGS adapters were ligated in the final step of library preparation in the present method, the migration of our metabarcoding method from an NGS platform (i.e. 454 technology) to another would be possible by the substitution of the adapters of the target NGS platform for those used in the present protocol. Use of the present method as well as its adaptation to more powerful sequencing platforms such as the Illumina and Ion Torrent platforms may contribute to the understanding of the biodiversity of Collembola. In particular, 300 + 300 bp paired-end sequencing using Illumina MiSeq is presumably promising for both markers examined in the present study because it can fully sequence the gene fragments.

Results in the present study indicate that mtCOI and mt16S genes were comparable in terms of the ability of surveying collembolan biodiversity. Considering the availability of reference databases and online identification systems (i.e., the BOLD system), the use of mtCOI rather than mt16S has an advantage. Nonetheless, as these two markers were revealed to have different PCR biases depending on species, the combined use of both markers may help in understanding the collembolan community more comprehensively when sufficient reference sequences have been accumulated for both genes.

Acknowledgements:

We thank the Field Science Education and Research Center, Kyoto University, Japan, for granting permission for this study. We thank the guest editor of Genome Dr. S. J. Adamowicz, an anonymous associate editor and two reviewers for carefully reading our manuscript and providing constructive comments. The authors would like to thank Enago (www.enago.jp) for the English language review. This research was partly supported by Grants-in-Aid
for Scientific Research, MEXT of Japan (nos. 50443081 and 15K16137), and Research Grants for Young Researchers from University of the Ryukyus (nos. 21 in 2012 and 16 in 2013).
References:

Amend, A.S., Seifert, K.A., and Bruns, T.D. 2010. Quantifying microbial communities with 454 pyrosequencing: does read abundance count?. Mol. Ecol. 19(24): 5555–5565.

Anderson, J.M., and Healey, I.N. 1972. Seasonal and inter-specific variation in major components of the gut contents of some woodland Collembola. J. Anim. Ecol., 41(2): 359–368.

Anslan, S., and Tedersoo, L. 2015. Performance of cytochrome c oxidase subunit I (COI), ribosomal DNA Large Subunit (LSU) and Internal Transcribed Spacer 2 (ITS2) in DNA barcoding of Collembola. Eur. J. Soil Biol., 69: 1–7.

Aoyama, H., Saitoh, S., Fujii, S., Nagahama, H., Shinzato, N., Kaneko, N., and Nakamori, T. 2015. A rapid method of non-destructive DNA extraction from individual springtails (Collembola). Appl. Entomol. Zool., 50(3): 419–425.

Berry, D., Mahfoudh, K.B., Wagner, M., and Loy, A. 2011. Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. Appl. Environ. Microbiol., 77(21): 7846–7849.

Boyce, R., Chilana, P., and Rose, T.M. 2009. iCODEHOP: a new interactive program for designing Consensus-Degenerate Hybrid Oligonucleotide Primers from multiply aligned protein sequences. Nucl. Acids Res., 37: W222–W228.

Brandon-Mong, G.J., Gan, H.M., Sing, K.W., Lee, P.S., Lim, P.E., and Wilson, J.J. 2015. DNA metabarcoding of insects and allies: an evaluation of primers and pipelines. Bull. Entomol. Res., 105(06): 717–727.
Brown, S.D., Collins, R.A., Boyer, S., Lefort, M.C., Malumbres-Olarte, J., Vink, C.J., and Cruickshank, R.H. 2012. SPIDER: an R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. Mol. Ecol. Resour., 12(3):562–565.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., Huttenhower, C.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pinto, R., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J., and Huttenhower, C.A. 2010. QIIME allows analysis of high-throughput community sequencing data. Nature Methods, 7(5): 335–336.

Carapelli, A., Liò, P., Nardi, F., Van der Wath, E., and Frati, F. 2007. Phylogenetic analysis of mitochondrial protein coding genes confirms the reciprocal paraphyly of Hexapoda and Crustacea. BMC Evol. Biol., 7(Suppl 2): S8.

Carapelli, A., Convey, P., Nardi, F., and Frati, F. 2014. The mitochondrial genome of the antarctic springtail *Folsomotoma octooculata* (Hexapoda; Collembola), and an update on the phylogeny of collembolan lineages based on mitogenomic data. Entomologia, 2(1): 190.

Clarke, L.J., Soubrier, J., Weyrich, L.S., and Cooper, A. 2014. Environmental metabarcodes for insects: in silico PCR reveals potential for taxonomic bias. Mol. Ecol. Resour., 14(6): 1160–1170.

Cicconardi, F., Fanciulli, P.P., and Emerson, B.C. 2013. Collembola, the biological species concept and the underestimation of global species richness. Mol. Ecol., 22(21): 5382–5396.
Cook, C.E., Yue, Q., and Akam, M. 2005. Mitochondrial genomes suggest that hexapods and crustaceans are mutually paraphyletic. Proc. R. Soc. Lond. B Biol. Sci., 272(1569): 1295–1304.

Deagle, B.E., Jarman, S.N., Coissac, E., Pompanon, F., and Taberlet, P. 2014. DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. Biol. Lett., 10(9): 20140562.

De Barba, M., Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E., and Taberlet, P. 2014. DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet. Mol. Ecol. Resour., 14(2): 306–323.

Deharveng, L. 2004. Recent advances in Collembola systematics. Pedobiologia, 48(5): 415–433.

Diaz-Real, J., Serrano, D., Piriz, A., and Jovani, R. 2015. NGS metabarcoding proves successful for quantitative assessment of symbiont abundance: the case of feather mites on birds. Exp. Appl. Acarol., 67(2): 209–218.

Filser, J. 2002. The role of Collembola in carbon and nitrogen cycling in soil: Proceedings of the Xth international Colloquium on Apterygota, České Budějovice 2000: Apterygota at the Beginning of the Third Millennium. Pedobiologia, 46(3): 234–245.

Folmer, O., Black, M., Hoeh, W. Lutz, R., and Vrijenhoek, R. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Mol. Marine Biol. Biotechnol., 3(5): 294–299.

Fonseca, V.G., Carvalho, G.R., Sung, W., Johnson, H.F., Power, D.M., Neill, S.P., Packer, M., Blaxter, M.L., Lambshead, P.J.D., Thomas, W.K., and Creer, S. 2010. Second-generation environmental sequencing un masks marine metazoan biodiversity. Nature Communications, 1: 98.
Fujii, S., and Takeda, H. 2012. Succession of collembolan communities during decomposition of leaf and root litter: effects of litter type and position. Soil Biol. Biochem., 54: 77–85.

Furuno, K., Suma, Y., and Niijima, K. 2014. Classification of the Family Onychiuridae Börner, 1913 (Hexapoda: Entognatha: Collembola) from Japan. Edaphologia, 95: 15–42.

Elbrecht, V., and Leese, F. 2015. Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass-sequence relationships with an innovative metabarcoding protocol. PLoS ONE, 10(7): e0130324.

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G. A., and Baird, D.J. 2011. Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. PLoS ONE, 6(4): e17497.

Hasegawa, M., and Niijima, K. 2012. Classification of the Family Isotomidae Börner, 1913 (Apterygota: Collembola) from Japan: 2. Isotominae Schaffer, 1896. Edaphologia, 90: 31–59.

Hasegawa, M., and Tanaka, S. 2013. Classification of the family Neanuridae (Hexapoda: Entognatha: Collembola) from Japan: 2. Subfamilies Brachystomellinae Massoud, 1967, Odontellinae Massoud, 1967, Frieseinae Massoud, 1967 and Pseudachorutinae Massoud, 1967. Edaphologia 92: 37–73.

Hebert, P.D., Cywinska, A., and Ball, S.L. 2003. Biological identifications through DNA barcodes. Proceedings of the Royal Society of London B: Biol. Sci., 270(1512): 313–321.

Hirai, J., Kuriyama, M., Ichikawa, T., Hidaka, K., and Tsuda, A. 2015. A metagenetic approach for revealing community structure of marine planktonic copepods. Mol. Ecol. Res., 15(1): 68–80.
Hishi, T., and Takeda, H. 2008. Soil microarthropods alter the growth and morphology of fungi and fine roots of *Chamaecyparis obtusa*. Pedobiologia, 52(2): 97–110.

Hopkin, S.P. 1997. Biology of the Springtails (Insecta: Collembola). Oxford University Press, Oxford.

Ichisawa, K. 2012. Classification of the Family Entomobryidae Schaffer, 1896 and related families (Hexapoda: Entognatha: Collembola) from Japan. Including the Family Orchesellidae Börner, 1906, Paronellidae Börner, 1913, Cyphoderidae Börner, 1913 and Oncopoduridae Carl and Lebedinsky, 1905. Edaphologia, 91: 31–97.

Itoh R, Hisamatsu M, Matsunaga M, and Hishida F. 1995. Growth and reproduction of a collembolan species, *Folsomia candida* (Willem), under experimental conditions. J. Coll. Arts. Sci. Showa Univ., 26: 23–30.

Itoh, R., Hasegawa, M., Ichisawa, K., Furuno, K., Suma, Y., Tanaka, S., Hasegawa, M. and Niijima, K. 2012. Classification of the Suborder Neelipleona Massoud, 1971 and Symphypleona Börner, 1901 (Hexapoda: Entognatha: Collembola) from Japan. Edaphologia, 91: 99–156.

Johnson, D.L., and Wellington, W.G. 1980. Predation of *Apochthonius minimus* (Pseudoscorpionida: Chthoniidae) on *Folsomia candida* (Collembola: Isotomidae) I. Predation rate and size-selection. Res. Popul. Ecol., 22(2): 339–352.

Lahr, D.J., and Katz, L.A. 2009. Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. Biotechniques, 47(4): 857–866.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. and Higgins, D.G. 2007. ClustalW and ClustalX version 2. Bioinformatics 23(21): 2947–2948.
Lawrence, K.L., and Wise, D.H. 2000. Spider predation on forest-floor Collembola and evidence for indirect effects on decomposition. Pedobiologia, 44(1): 33–39.

Leray, M., Yang, J.Y., Meyer, C.P., Mills, S.C., Agudelo, N., Ranwez, V., Boehm, J.T., and Machida, R.J. 2013. A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. Front. Zool., 10(1), 34.

Machida, R.J., and Knowlton, N. 2012. PCR Primers for Metazoan Nuclear 18S and 28S Ribosomal DNA Sequences. PLoS ONE, 7(9): e46180.

Mathieu-Daudé, F., Welsh, J., Vogt, T., and McClelland, M. 1996. DNA rehybridization during PCR: the ‘Cot effect’ and its consequences. Nucl. Acids Res., 24(11): 2080–2086.

Meier, R., Shiyang, K., Vaidya, G., and Ng, P. 2006. DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. Systemat. Biol., 55(5): 715–728.

Meyer, C.P., and Paulay, G. 2005. DNA barcoding: error rates based on comprehensive sampling. PLoS Biol. 3(12): 2229–2238.

Nakamori, T. 2013. A new species of Ceratophysella (Collembola: Hypogastruridae) from Japan, with notes on its DNA barcode and a key to Japanese species in the genus. Zootaxa, 3641(4): 371–378.

Nakamori, T., Ichisawa, K. and Tamura, H. 2014. Classification of the families Poduridae Latreille, 1804 and Hypogastruridae Börner, 1913 (Hexapoda: Entognatha: Collembola) from Japan. Edaphologia, 95: 43–82.
Nardi, F., Carapelli, A., Fanciulli, P.P., Dallai, R., and Frati, F. 2001. The complete mitochondrial DNA sequence of the basal hexapod Tetrodontophora bielanensis: evidence for heteroplasmy and tRNA translocations. Mol. Biol. Evol., 18(7): 1293–1304.

Nardi, F., Spinsanti, G., Boore, J.L., Carapelli, A., Dallai, R., and Frati, F. 2003. Hexapod origins: monophyletic or paraphyletic?. Science, 299(5614): 1887–1889.

Niijima, K., and Hasegawa, M. 2011. Classification of the Family Isotomidae Börner, 1913 (Apterygota: Collembola) from Japan: 1. Anurophorinae Börner, 1906 and Proisotominae Stach, 1947. Edaphologia (89): 29–69.

Nolte, V., Pandey, R.V., Jost, S., Medinger, R., Ottenwälder, B., Boenigk, J., and Schlotterer, C. 2010. Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. Mol. Ecol., 19: 2908–2915.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., and Wagner, H. 2016. vegan: Community Ecology Package. R package version 2.3-4. https://CRAN.R-project.org/package=vegan

Petersen, H. 1975. Estimation of dry weight, fresh weight, and calorific content of various Collembolan species. Pedobiologia, 15: 222–243.

Petersen, H., and Luxton, M. 1982. A comparative analysis of soil fauna populations and their role in decomposition processes. Oikos, 39(3): 288–388.

Polz, M.F., and Cavanaugh, C.M. 1998. Bias in template-to-product ratios in multitemplate PCR. Appl. Environ.
Microbiol., 64(10): 3724–3730.

Porazinska, D.L., Giblin-Davis, R.M., Faller, L., Farmerie, W., Kanzaki, N., Morris, K., Powers, T.O., Tucker, A.E., Sung, W., and Thomas, W.K. 2009. Evaluating high-throughput sequencing as a method for metagenomic analysis of nematode diversity. Mol. Ecol. Resour. 9(6): 1439–1450.

Porazinska, D.L., Sung, W.A.Y., Giblin-Davis, R.M., and Thomas, W.K. 2010. Reproducibility of read numbers in high-throughput sequencing analysis of nematode community composition and structure. Mol. Ecol. Resour., 10(4): 666–676.

Porco, D., Bedos, A., Greenslade, P., Janion, C., Skarżyński, D., Stevens, M.I., Vuuren, J.V., and Deharveng, L. 2012a. Challenging species delimitation in Collembola: cryptic diversity among common springtails unveiled by DNA barcoding. Invertebr. Systemat., 26(6): 470–477.

Porco, D., Potapov, M., Bedos, A., Busmachiu, G., Weiner, W.M., Hamra-Kroua, S., and Deharveng, L. 2012b. Cryptic diversity in the ubiquitous species Parisotoma notabilis (Collembola, Isotomidae): a long used chimeric species. PLoS ONE, 7(9): e46056.

Porco, D., Skarżyński, D., Decaëns, T., Hebert, P.D., and Deharveng, L. 2014. Barcoding the Collembola of Churchill: a molecular taxonomic reassessment of species diversity in a sub-Arctic area. Mol. Ecol. Resour., 14(2): 249–261.

Ramirez-Gonzalez, R., Douglas, W.Y., Bruce, C., Heavens, D., Caccamo, M., and Emerson, B.C. 2013. PyroClean: denoising pyrosequences from protein-coding amplicons for the recovery of interspecific and intraspecific genetic variation. PLoS ONE 8(3): e57615.
Ratnasingham, S., and Hebert, P.D. 2007. BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). Mol. Ecol. Notes, 7(3): 355–364.

Saitoh, S., Fujii, S., and Takeda, H. 2014. Effect of habitat structural complexity on collembolan communities. Ecol. Res., 29(1): 81–90.

Schneider, C., Cruaud, C., and D’Haese, C.A. 2011. Unexpected diversity in Neelipleona revealed by molecular phylogeny approach (Hexapoda, Collembola). Soil Organisms, 83(3): 383–398.

Shokralla, S., Gibson, J.F., Nikbakht, H., Janzen, D.H., Hallwachs, W., and Hajibabaei, M. 2014. Next-generation DNA barcoding: using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. Mol. Ecol. Resour., 14(5): 892–901.

Song, H., Buhay, J.E., Whiting, M.F., and Crandall, K.A. 2008. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. Proc. Natl. Acad. Sci., 105(36): 13486–13491.

Sørensen, T. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. Biol. Skr., 5, 1–34.

Suma, Y. 2009. Classification of the Family Tomoceridae (Collembola) from Japan. Edaphologia, 84: 25–56.

Swift, M.J., Heal, O.W., and Anderson, J.M. 1979. Decomposition in Terrestrial Ecosystems. University of California Press, Berkeley.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., and Willerslev, E. 2012. Towards next-generation biodiversity assessment using DNA metabarcoding. Mol. Ecol., 21(8): 2045–2050.
Takeda, H. 1987. Dynamics and maintenance of collembolan community structure in a forest soil system. Res. Popul. Ecol., 29(2): 291–346.

Tanaka, M. 1970. Ecological studies on communities of soil Collembola in Mt. Sobo, southwest Japan. Jpn. J. Ecol., 20(3): 102–110.

Tanaka, S. 2010. Classification of the family Neanuridae (Collembola) from Japan. Edaphologia, 86: 27–79.

Thomas, A.C., Deagle, B.E., Paige Eveson, J., Harsch, C.H., and Trites, A.W. 2016. Quantitative DNA metabarcoding: improved estimates of species proportional biomass using correction factors derived from control material. Mol. Ecol. Resour. 16(3): 714–726.

Torricelli, G., Carapelli, A., Convey, P., Nardi, F., Boore, J.L., and Frati, F. 2010. High divergence across the whole mitochondrial genome in the “pan-Antarctic” springtail Friesea grisea: Evidence for cryptic species?. Gene, 449(1): 30–40.

Usher, M.B. 1985. Population and community dynamics in the soil ecosystem. In Ecological Interactions in Soil: Plants, Microbes and Animals. Edited by A.H. Fitter, D. Atkinson, D.J. Read, and M.B. Usher. Blackwell, Palo Alto, pp. 243–265.

Vestheim, H., and Jarman, S.N. 2008. Blocking primers to enhance PCR amplification of rare sequences in mixed samples—a case study on prey DNA in Antarctic krill stomachs. Front. Zool., 5: 12.

Wu, T.D., and Watanabe, C.K. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics, 21(9): 1859–1875.

Xiong, Y., Gao, Y., Yin, W.Y., and Luan, Y.X. 2008. Molecular phylogeny of Collembola inferred from ribosomal
RNA genes. Mol. Phylogenet. Evol., 49(3): 728–735.

Yu, D.W., Ji, Y., Emerson, B.C., Wang, X., Ye, C., Yang, C., and Ding, Z. 2012. Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods Ecol. Evol., 3(4): 613–623.

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. 2000. A greedy algorithm for aligning DNA sequences. J. Comput. Biol., 7(1–2): 203–214.
Table footnotes and figure legends:

Table 1. Community assessment of Collembola at the study site by microscopic examination and metabarcoding using mitochondrial cytochrome c oxidase subunit 1 (mtCOI) and 16S ribosomal RNA (mt16S) genes

* Species whose mtCOI and/or mt16S genes were barcoded at the study site.

† Number of individuals found per bulk microarthropod sample extracted from 100 ml of soil

‡ Relative species (or OTU) abundance per sample based on number of individuals (by microscopic examination) or reads (by metabarcoding)

§ Number of samples in which at least one individual or read of the species or OTU was found.

|| Number of OTUs assigned to species after taxonomic identification of OTUs via reference database (see text).

Shading indicates species found by all of three methods (microscopic examination and metabarcoding by mtCOI and mt16S genes).

Note: Families of unidentified collembolan OTUs were inferred based on BLAST searches against the NCBI Nucleotide Collection database (see text for detail).

Table S1. PCR primers used for Sanger sequencing in the present study

* PCR conditions were as follows: for mtCOI, hot start at 94°C for 2 min, 40 cycles at 98°C for 10 s, 48°C for 30 s and 68°C for 1 min; for mt16S, hot start at 94°C for 2 min, 35 cycles at 98°C for 10 s, 53°C for 30 s and 68°C for 1 min.
Table S1. PCR primer sets for metabarcoding used in the present study.

Table S2. Taxonomy, body length, and estimated biomass of collembolan specimens used in mock community samples

Values are mean ± SD ($n > 20$). †, Collected from a single cohort. ‡, Collected from a field population.

Table S3. Species composition in mock community samples

Values are numbers of collembolan individuals.

Table S4. Sequencing throughput using metabarcoding libraries and 454 GS Junior

*, Accession number for sequence dataset deposited in the DDBJ Read Archive (DRA) under submission no. DRA004529; †, Number of reads after trimming and filtering by split_libraries.py script (QIIME version 1.8.0, Caporaso et al. 2010; see text for detailed settings); ‡, Number of reads aligned to 97% OTUs output by pick_open_reference_otas.py script (QIIME) with reference sequences.

Table S5. Correlation coefficients between numbers of individuals and read counts across mock community samples after normalization using different species as internal control

*, Species originally intended as internal control that was included in the same number of individuals per
sample; this species could not be used as an internal control in mtCOI data because no read was found from this species in one of the samples. For species abbreviations, see Fig. 4.
Figure 1. A schematic of laboratory protocol of the metabarcoding method employed in the present study.

Figure 2. Taxonomic composition based on (a) proportion of individuals and (b) sequence read counts from mtCOI and (c) mt16S data in natural community samples. Note that two soil samples were sampled from each of 10 subplots, one of which was analysed by morphological identification (samples C_01 to C_10), whereas the other was analysed by metabarcoding (N_01 to N_10).

Figure 3. Cumulative curves of collembolan (a) morphological species and (b) 90% OTUs in mtCOI and (c) mt16S data against number of samples. The number of OTUs includes those from same morphological species (see text and Table 1). Vertical bars represent 95% confidence intervals.

Figure 4. Relative species composition in mock community samples based on (a) individual animals, (b) biomass, and (c) sequence read counts in metabarcoding datasets from the mtCOI and (d) mt16S libraries.

Figure 5. Relationships between number of animal individuals and read count per species in metabarcoding data of mock community samples using (a) mtCOI and (b) mt16S genes, after normalization using internal standard species. *P. pseudolanuginosa* was used as internal standard. For species abbreviations, see Fig. 4.

Figure 6. Principle component analyses of mock community samples based on (a) number of animal individuals
and (b) read counts in the mtCOI and (c) mt16S data. Read counts were normalized using *P. pseudolanuginosa* as internal control and further standardized for each species (to a mean value of zero and standard deviation of 1 among samples). Note that axis of the component 1 for mt16S library was inverted. Abbreviations used for the collembolan species are as follows: Aj, *Arrhopalites japonicus*; Hr, *Hypogastrura reticulata*; Of, *Onychiurus flavescens*; Pp, *Pseudosinella pseudolanuginosa*; Sd, *Sinella dubiosa* and Su, *Sinella umesaoi*.

Figure S1. Inter- and intra-specific variation in gene regions used in the metabarcoding method in present study. Filled parts of the columns (in red) indicate within-species comparisons.

Figure S2. Effect of the threshold value on false positive (over-splitting) and negative (over-lumping) rates in species delimitation by gene markers based on reference sequences.

Figure S3. Effect of threshold value on the success rate of taxonomic identification using gene markers based on reference sequences.

Figure S4. A schematic of the OTU clustering pipeline for metabarcoding data used in present study.

Figure S5. Portion of quality-filtered reads per sample in multiplexing sequencing runs in present study. Labels around pie charts are 4-base MID tags.
Figure S6. Taxonomy composition in metabarcoding data per sequencing run in present study.

Figure S7. Relationships of relative species abundances in the natural collembolan community between morphological identification and metabarcoding using mtCOI and mt16S genes. Mean relative species abundances (in percentage; \( n = 10 \)) were plotted for 13 species commonly identified by the three methods.

Figure S8. Relationships between relative abundance of reads (a-b, mtCOI library; c-d, mt16S library) and relative proportion of individuals (a, c) or biomass (b, d; estimated dry weight) in the mock community samples. For species abbreviations, see Fig. 4. Note that *F. candida* (Fc) was included in the same number in each sample.
Table 1. Community assessment of Collembola at the study site by microscopic examination and metabarcoding using mitochondrial cytochrome c oxidase subunit 1 (mtCOI) and 16S ribosomal RNA (mt16S) genes

| Family           | Species or OTU ID | (a) Microscopic examination | (b) mtCOI data | (c) mt16S data |
|------------------|------------------|----------------------------|----------------|---------------|
|                  |                  | Relative abundance² | # OTUs | Relative abundance² | # OTUs | Relative abundance² | # OTUs |
|                  |                  | # individual³ | # samples found |                  |                  |                  |                  |
|                  |                  | (mean ± SD) |                  |                  |                  |                  |                  |
| Isotomidae       | Folsomia excludens * | 50.3 ± 40.5 | 30.0 ± 20.6 | 10 | 1.72 ± 2.16 | 10 | 0.457 ± 0.52 | 10 |
| Isotomidae       | Terecrathella sylvestris * | 55.2 ± 56.8 | 27.2 ± 23.5 | 10 | 1.72 ± 0.4 | 10 | 0.457 ± 0.6 | 10 |
| Isotomidae       | Isotomus s lasseni ² | 16.1 ± 9.8 | 12.9 ± 13.1 | 10 | 1.8 ± 0.8 | 10 | 1.3 ± 0.8 | 10 |
| Isotomidae       | Pseudosminthurus sp. | 14.5 ± 16.8 | 9.3 ± 14.3 | 9 | 1.8 ± 0.2 | 9 | 2.0 ± 0.3 | 9 |
| Isotomidae       | Isotomus minor ² | 22.8 ± 59.3 | 6.3 ± 15.8 | 4 | 8.2 ± 0.6 | 4 | 2.10 ± 0.22 | 2 |
| Onychiuridae     | Onychiurus flavescens   | 6.3 ± 10.3 | 3.9 ± 5.4 | 7 | 6.0 ± 0.5 | 4 | 0.85 ± 0.06 | 5 |
| Nectidae         | Megalogyrus minimum ² | 3.5 ± 56.8 | 2.0 ± 2.7 | 7 | 0.01 ± 0.043 | 1 | 1.073 ± 0.09 | 7 |
| Entomobryidae    | Sinella dehiscens ² | 2.0 ± 1.6 | 1.4 ± 1.5 | 8 | 0.67 ± 0.1 | 6 | 1.63 ± 0.13 | 5 |
| Talitridae       | Talitrus yosi ² | 2.2 ± 3.0 | 1.1 ± 2.4 | 5 | 2.3 ± 0.8 | 9 | 1.7 ± 12.3 | 10 |
| Tomoceridae      | Tomocerus surrana ² | 1.3 ± 2.8 | 0.89 ± 1.7 | 4 | 0.23 ± 0.72 | 1 | 1.059 ± 0.19 | 1 |
| Onycopodidae     | Onycopoda yosi ² | 1.8 ± 5.3 | 0.85 ± 1.3 | 4 | 0.02 ± 0.038 | 2 | 1.005 ± 0.02 | 1 |
| Neunatidae       | Paratenuis sp. ² | 3.5 ± 5.2 | 2.7 ± 1.9 | 2 | NA | NA | NA | NA |
| Neᡠnuatidae      | Neunatidae sp. ² | 0.6 ± 1.1 | 0.72 ± 2.1 | 2 | NA | NA | NA | NA |
| Smithridae       | Secomithrum merulicolis ² | 1.1 ± 1.5 | 0.76 ± 1.2 | 6 | 0 | 0 | 0 | 0 |
| Isotomidae       | Microtions avrichmensis ² | 0.8 ± 2.2 | 0.71 ± 2.1 | 2 | 0 | 0 | 0 | 0 |
| Neunatidae       | Microtium sp. ² | 0.2 ± 0.42 | 0.18 ± 0.42 | 2 | NA | NA | NA | NA |
| Hygropoctridae   | Ceratophysella pilosa ² | 0.3 ± 0.95 | 0.16 ± 0.49 | 1 | NA | NA | 1.038 ± 0.10 | 4 |
| Neunatidae       | Tenuatella sp. ² | 0.4 ± 0.70 | 0.14 ± 0.23 | 3 | NA | NA | NA | NA |
| Neunatidae       | Ceratophysella sp. ² | 0.1 ± 0.32 | 0.09 ± 0.30 | 1 | 2.0 ± 0.0 | 2 | 1.70 ± 0.2 | 2 |
| Odontellidae     | Superodontella sp. ² | 0.2 ± 0.42 | 0.09 ± 0.20 | 2 | NA | NA | NA | NA |
| Neunatidae       | Dicranotellidae sp. ² | 0.2 ± 0.42 | 0.08 ± 0.18 | 2 | NA | NA | NA | NA |
| Entomobryidae    | Entomobrya sp. ² | 0.1 ± 0.32 | 0.06 ± 0.19 | 1 | 0.07 ± 0.22 | 1 | 1.03 ± 0.18 | 8 |
| Isotomidae       | Isotoma carpenters ² | 0.1 ± 0.32 | 0.06 ± 0.19 | 1 | NA | NA | NA | NA |
| Neunatidae       | Neunatidae sp. ² | 0.1 ± 0.32 | 0.06 ± 0.19 | 1 | 0 | 0 | 0 | 0 |
| Neunatidae       | Neunatidae sp. ² | 0.1 ± 0.32 | 0.05 ± 0.17 | 1 | NA | NA | NA | NA |
| Entomobryidae    | Entomobrya sp. ² | 0.1 ± 0.32 | 0.05 ± 0.17 | 1 | NA | NA | NA | NA |
| Hygropoctridae   | Hygropocta sp. ² | 0 | 0 | 0 | 1.76 ± 2.4 | 1 | 2 | 1.5 | 9 |
| Neunatidae       | Neunatidae sp. ² | 0 | 0 | 0 | 0 | 0.004 ± 0.013 | 1 | 0 | 0 |
| * Species whose mtCOI and/or mt16S genes were barcoded at the study site. 
* Number of individuals found per bulk microcrayopod sample extracted from 100 ml of soil 
‡ Relative species (or OTU) abundance per sample based on number of individuals (by microscopic examination) or reads (by metabarcoding) 
† Number of samples in which at least one individual or read of the species or OTU was found. 
§ Number of OTUs assigned to species after taxonomic assignment of OTUs via reference database (see text). 
Shading indicates species found by all of three methods (microscopic examination and metabarcoding by mtCOI and mt16S genes). 
Note: Families of unidentified collombolan OTUs were inferred based on BLAST searches against the NCBI Nucleotide Collection database (see text for detail).
1. DNA extraction from community samples

- Community sample A
- Sample B
- Sample C
- Sample D
- Animals
- DNA

2. PCR amplification

2-1. Optimization of PCR cycles

2-2. PCR with MID tag-tailed primers

- DNA
- 5′ MID F-primer
- 5′ MID R-primer
- PCR
- 5′ Tagged amplicon
- 5′

3. Pooling amplicons for multiplexing

3-1. Measurement of molar concentration

Molar concentration determined based on size distribution

- DNA quantity
- Amplicon size

3-2. Equimolar mixture of amplicons

- Sample A
- Sample B
- Sample C
- Amplicon pool

4. Library preparation for NGS and sequencing

4-1. Ligation with Y-adapters for 454 technology

Both strands are to be sequenced

- Number of reads
- Read length
(a) Morphological species

(b) mtCOI data

(c) mt16S data
(a) Number of individuals

(b) Biomass (estimated dry weight)

(c) mtCOI data

(d) mt16S data

Abbrev. | Species (Family)
---|---
Fc | Folsomia candida (Isotomidae)
Pp | Pseudosinella pseudolanuginosa (Entomobryidae)
Sd | Sinella dubiosa (Entomobryidae)
Su | Sinella unesoi (Entomobryidae)
Aj | Arrhopalites japonicus (Katiannidae)
Hr | Hypogastrura reticulata (Hypogastruridae)
Olf | Onychiurus flavescens (Onychiuridae)
(a) Number of animal individuals

(b) mtCOI data

(c) mt16S data