Applicability and perspectives for DNA barcoding of soil invertebrates (#96785)

First submission

Guidance from your Editor

Please submit by **20 Mar 2024** for the benefit of the authors (and your token reward) .

Structure and Criteria
Please read the 'Structure and Criteria' page for general guidance.

Raw data check
Review the raw data.

Image check
Check that figures and images have not been inappropriately manipulated.

If this article is published your review will be made public. You can choose whether to sign your review. If uploading a PDF please remove any identifiable information (if you want to remain anonymous).

Files
Download and review all files from the materials page.

11 Figure file(s)
6 Table file(s)

Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

- Clear, unambiguous, professional English language used throughout.
- Intro & background to show context. Literature well referenced & relevant.
- Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- Figures are relevant, high quality, well labelled & described.
- Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

- Original primary research within Scope of the journal.
- Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- Rigorous investigation performed to a high technical & ethical standard.
- Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- Impact and novelty not assessed. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
- All underlying data have been provided; they are robust, statistically sound, & controlled.
- Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

The best reviewers use these techniques

Tip	*Example*
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult. I suggest you have a colleague who is proficient in English and familiar with the subject matter review your manuscript, or contact a professional editing service.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

Applicability and perspectives for DNA barcoding of soil invertebrates

Jéhan Le Cadre [Corresp., 1, 2], **Finn Luca Klemp** [1], **Miklós Bálint** [3, 4], **Stefan Scheu** [1, 5], **Ina Schaefer** [1, 3, 4]

[1] J. F. Blumenbach Institute of Zoology and Anthropology, University of Göttingen, Göttingen, Germany

[2] Biocenter, Ludwig-Maximilians-Universität München, Planegg-Martinsried, Germany

[3] Senckenberg Biodiversity Climate Research Center, Frankfurt Main, Germany

[4] Loewe Center for Translational Biodiversity Genomics (LOEWE-TBG), Frankfurt Main, Germany

[5] Centre of Biodiversity and Sustainable Land Use, University of Göttingen, Göttingen, Germany

Corresponding Author: Jéhan Le Cadre
Email address: jehanlecadre@gmail.com

Belowground invertebrate communities are dominated by species-rich and very small microarthropods that require long handling times and high taxonomic expertise for species determination. Molecular based methods like metabarcoding circumvent the morphological determination process by assigning taxa bioinformatically based on sequence information. The potential to analyse diverse and cryptic communities in short time at high taxonomic resolution is promising. However, metabarcoding studies revealed that taxonomic assignment below family-level in Collembola (Hexapoda) and Oribatida (Acariformes) is difficult and often fails. These are the most abundant and species-rich soil-living microarthropods, and the application of molecular-based, automated species determination would be most beneficial in these taxa. In this study, we analysed the presence of a barcoding gap in the standard barcoding gene *cytochrome oxidase I* (COI) in Collembola and Oribatida. The barcoding gap describes a significant difference between intra- and interspecific genetic distances among taxa and is essential for bioinformatic taxa assignment. We collected COI sequences of Collembola and Oribatida from BOLD and NCBI and focused on species with a wide geographic sampling to capture the range of their intraspecific variance. Our results show that intra- and interspecific genetic distances in COI overlapped in most species, impeding accurate assignment. When a barcoding gap was present, it exceeded the standard threshold of 3 % intraspecific distances and also differed between species. Automatic specimen assignments also showed that most species comprised of multiple genetic lineages that caused ambiguous taxon assignments in distance-based methods. Character-based taxonomic assignment using phylogenetic trees and monophyletic clades as ~~criterion~~ worked for some species of Oribatida but failed completely for Collembola. Notably, parthenogenetic species showed lower genetic

variance in COI and more accurate species assignment than sexual species. The different patterns in genetic diversity among species suggest that the different degrees of genetic variance result from deep evolutionary distances. This indicates that a single genetic threshold, or a single standard gene, will probably not be sufficient for the molecular species identification of many Collembola and Oribatida taxa. Our results also show that haplotype diversity in some of the investigated taxa was not even nearly covered, but coverage was better for Collembola than for Oribatida. Additional use of secondary barcoding genes and long-read sequencing of marker genes can improve metabarcoding studies. We also recommend the construction of pan-genomes and pan-barcodes of species lacking a barcoding gap. This will allow both to identify species boundaries, and to cover the full range of variability in the marker genes, making molecular identification also possible for species with highly diverse barcode sequences.

Title: **Applicability and perspectives for DNA barcoding of soil invertebrates**

Jéhan Le Cadre[1,2], Finn Luca Klemp[1], Miklós Bálint[3,4], Stefan Scheu[1,5], Ina Schaefer[1,3,4]

[1] J. F. Blumenbach Institute of Zoology and Anthropology, University of Göttingen, Göttingen, Germany
[2] Ludwig-Maximilians-Universität München, Biocenter, Großhaderner Str. 2, 82152 Planegg-Martinsried, Germany
[3] Senckenberg Biodiversity Climate Research Center, Frankfurt am Main, Germany
[4] Loewe Center for Translational Biodiversity Genomics (LOEWE-TBG), Frankfurt am Main, Germany
[5] Centre of Biodiversity and Sustainable Land Use, University of Göttingen, Göttingen, Germany

Corresponding Author: jehanlecadre@gmail.com
Jéhan Le Cadre[1,2]
[1] J. F. Blumenbach Institute of Zoology and Anthropology, University of Göttingen, Göttingen, Germany
[2] Ludwig-Maximilians-Universität München, Biocenter, Großhaderner Str. 2, 82152 Planegg-Martinsried, Germany
Email address: jehanlecadre@gmail.com

Background

Belowground invertebrate communities are dominated by species-rich and very small microarthropods that require long handling times and high taxonomic expertise for species determination. Molecular based methods like metabarcoding circumvent the morphological determination process by assigning taxa bioinformatically based on sequence information. The potential to analyse diverse and cryptic communities in short time at high taxonomic resolution is promising. However, metabarcoding studies revealed that taxonomic assignment below family-level in Collembola (Hexapoda) and Oribatida (Acariformes) is difficult and often fails. These are the most abundant and species-rich soil-living microarthropods, and the application of molecular-based, automated species determination would be most beneficial in these taxa. In this study, we analysed the presence of a barcoding gap in the standard barcoding gene *cytochrome oxidase I* (COI) in Collembola and Oribatida. The barcoding gap describes a significant difference between intra- and interspecific genetic distances among taxa and is essential for bioinformatic taxa assignment. We collected COI sequences of Collembola and Oribatida from BOLD and NCBI and focused on species with a wide geographic sampling to capture the range of their intraspecific variance. Our results show that intra- and interspecific genetic distances in COI overlapped in most species, impeding accurate assignment. When a barcoding gap was present, it exceeded the standard threshold of 3 % intraspecific distances and also differed between species. Automatic specimen assignments also showed that most species comprised of multiple genetic lineages that caused ambiguous taxon assignments in distance-based methods. Character-based taxonomic assignment using phylogenetic trees and monophyletic clades as criterion worked for some species of Oribatida but failed completely for Collembola. Notably, parthenogenetic species showed lower genetic variance in COI and more accurate species assignment than sexual species. The different patterns in genetic diversity among species suggest that the different degrees of genetic variance result from deep evolutionary distances. This indicates that a single genetic threshold, or a single standard gene, will probably not be sufficient for the molecular species identification of many Collembola and Oribatida taxa. Our results also show that haplotype diversity in some of the investigated taxa was not even nearly covered, but coverage was better for Collembola than for Oribatida. Additional use of secondary barcoding genes and long-read sequencing of marker genes can improve metabarcoding studies. We also recommend the construction of pan-genomes and pan-barcodes of species lacking a barcoding gap. This will allow both to identify species

boundaries, and to cover the full range of variability in the marker genes, making molecular identification also possible for species with highly diverse barcode sequences.

Introduction

Soils are among the most diverse habitats on earth, harbouring 25 % to 50 % of the biodiversity on Earth (Decaëns et al., 2006, Decaëns, 2010; Anthony, Bender & van der Heijden, 2023). This biodiversity drives essential processes for life on Earth and provides ecosystem services that impact human wellbeing, such as the decomposition of dead organic material, recycling of nutrients and carbon storage (Wardle et al., 2004; Lavelle et al., 2006; Bardgett & van der Putten, 2014). Characterizing and monitoring soil biodiversity therefore is of general interest to maintain and preserve soil functions (Orgiazzi et al., 2015). However, this is a challenging and time-consuming task due to the enormous taxonomic diversity and cryptic lifestyles of soil-organisms. Molecular methodologies offer great advantages for soil biodiversity assessment in terms of time and cost efficiency, and taxonomic resolution (Antil et al., 2012; Eisenhauer, Bonn & Guerra, 2019).

A large fraction of soil animal biodiversity is represented by microarthropods with body-sizes between 0.1 and 2 mm. Collembola (Hexapoda) and Oribatida (Acari: Sarcoptiformes) are dominant and omnipresent microarthropod taxa, and occur in all soil-related habitats where they reach high abundances of up to 50,000 - 100,000 individuals per square meter (Bardgett & van der Putten, 2014). Traditionally, Collembola and Oribatida have been described as decomposers, microbivorous and fungivores, but studies using stable isotopes showed that they actually cover several trophic levels, demonstrating trophic specialization and functional diversity within these taxa (Schneider et al., 2004; Pollierer et al., 2009; Potapov et al., 2016; Maraun et al., 2023). These microarthropods spend their entire life in the soil matrix or in the litter layer, which makes them interesting candidates as bioindicators of soil quality in monitoring programs (Gulvik, 2007). Collembola are typical r-strategists with fast reproduction cycles, whereas Oribatida are usually considered K-strategists with long-life spans of 1-3 years and low fecundity, but species with shorter life-cycles are also common (Maraun & Scheu, 2000; Pfingstl & Schatz, 2021). The general differences in life-history traits and trophic diversity between Collembola and Oribatida could be informative for monitoring programs. Collembola respond and recover more quickly to disturbances (Ponge et al., 2003; Santorufo et al., 2012) than Oribatida, which have long recovery times and therefore are more sensitive to environmental changes (Zaitsev et al., 2002; Gulvik, 2007; Pfingstl & Schatz, 2021). However, the wide range of functional and life-history traits

among different species necessitates species level determination in order to better understand their interactions in the soil system or to use them as bioindicators for changes in soil functions. About 9,000 species of Collembola and 11,000 species of Oribatida are described worldwide, but this likely represents only about 20 % of the expected species (Potapov et al., 2020; Behan-Pelletier & Lindo, 2023). Local species richness of these two taxa can be very high, reaching 60-100 species in forest soils (Rusek, 1998; Schatz & Behan-Pelletier, 2008). High species richness and abundance, and small body sizes of both, Collembola and Oribatida, pose a significant challenge for biodiversity assessments. Molecular applications, such as DNA barcoding and metabarcoding, have great potential to aid specimen identification and biodiversity assessment (Valentini, Pompanon & Taberlet, 2009). These methods utilize a standardized DNA fragment for taxonomic assignment of specimens by matching DNA sequences of undetermined individuals to a reference database (Hebert et al., 2003; Hebert & Gregory, 2005). This enables to automatically assign any taxonomic level and even species names to undetermined individuals. It is applicable to mixed samples of pooled specimens, which significantly reduces workload and costs. Further, molecular identification tools are equally applicable to juveniles that often lack taxonomic characters (Richard et al., 2010; Grzywacz et al., 2021). Automated handling of samples, simultaneous identification of multiple individuals in a single reaction, and the scalability of molecular data to any taxonomic level offers new opportunities for analysing spatial and temporal dynamics of soil-living animals (Arribas et al., 2021; Decaëns, 2021), and thereby provide new perspectives for monitoring of soil biodiversity. The method, however, relies on two preconditions: (1) a representative reference database and (2) a marker (barcoding) gene that reliably separates species. The most common databases are BOLD ("The Barcode of Life Data System", Ratnasingham & Hebert, 2007) and NCBI (https://ncbi.nlm.nih.gov/). The standard barcoding gene for Metazoa is a 658 bp region of the mitochondrial *cytochrome oxidase I* gene (COI; Herbert et al., 2003; Hubert et al., 2008). In general, a minimum of 500 bp of COI is required, but shorter fragments can also be used for specimen identification and species discovery (Hajibabaei et al., 2006; Collins & Cruickshank, 2012).

The success of species delineation based on genetic markers depends on the presence of a barcoding gap, which implies that genetic distances within a species are smaller than genetic variances to congeneric and other species (Meyer & Paulay, 2005). A global threshold of 2 %-3 %

intraspecific sequence divergence, or 10x the mean intraspecific divergence, has been proposed to reliably separate species (Hebert et al., 2004). Such a universal threshold is extremely helpful for automated species assignment of genetic data in bioinformatic pipelines. This threshold seems to be valid for a range of taxa (Hebert, Ratnasingham & deWaard, 2003; Hebert et al., 2004; Barrett & Hebert, 2005), but its universal application has been questioned for other species (e.g., Burns et al., 2007; Chapple & Ritchie, 2013; Elias et al., 2007; Meier et al., 2006; Meyer & Paulay, 2005; Wiemers & Fiedler, 2007). In particular soil-living animals show high intraspecific divergences in the COI gene that commonly exceed the standard barcoding threshold. Examples cover different families of earthworms (King, Tibble & Symondson, 2008; Novo et al., 2009; Martinsson, Rhoden & Erséus, 2016), Collembola (Porco et al., 2012a, Porco et al., 2012b; von Saltzwedel, Scheu & Schaefer, 2016; Zhang et al., 2019) and Oribatida (Rosenberger et al., 2013; von Saltzwedel et al., 2014). These studies question the general effectiveness of COI for specimen identification in these taxa. Moreover, asexual reproduction occurs in 7-10 % of all species in several families of Collembola and 10 % of all species of Oribatida (Charhataghi, Scheu & Ruess, 2006; Cianciolo & Norton, 2006; Chernova et al. 2010, Bluhm, Scheu & Maraun, 2016), and asexual species can be dominant in temperate forests (Maraun & Scheu, 2000). According to theory, asexual organisms accumulate mutations over time, until they go extinct due to the accumulation of too many deleterious mutations (Muller's ratcher: Muller, 1964; Kondrashov's hatchet: Kondrashov, 1988). This suggests that present day populations of asexual species represent a range of COI haplotypes, while populations of sexual species should represent discrete clusters of similar COI haplotypes, standing for independently evolving lineages that interbreed (Barraclough, Birky & Burt, 2003). In consequence, a barcoding gap should not be present in asexual species but rather a continuum of slightly divergent individuals. Further, hybridization events are potential origins of asexual species, and if followed by mitochondrial introgression the detection of a barcoding gap is difficult (Mutanen et al., 2016; Dupont et al., 2016). Altogether, asexual reproduction could blur lines of species identification, and a hybrid species could be wrongly identified as its maternal species.

The aim of this study was to test, (i) if the standard barcoding marker gene COI meets the precondition to reliably assign species in Collembola and Oribatida and, to check (ii) the accuracy of separating species based on a barcoding gap. Many species of these taxa have wide distribution ranges, and European species often occur across palaearctic or holarctic regions. Geographic coverage of samples provided in DNA barcode reference libraries can affect species assignment

(Hebert et al., 2003). We therefore focused on species with a dense and broad geographic sampling to cover the potential range of intra-specific haplotype variation of COI (Philipps, Gillis & Hanner, 2022). To assess intra- and interspecific genetic variance of these species, we downloaded sequences of congeneric species that were represented in databases with a minimum of 3-5 sequences per species. We included parthenogenetic (asexual) species to test (iii) if the reproductive mode affects the barcoding gap, because parthenogenetic species likely carry a continuum of divergent haplotypes due to the accumulation of mutations and the absence of homogenizing effects of mixis. Datasets were obtained by checking literature and public databases (BOLD, NCBI). We selected five oribatid mite species (Rosenberger, 2011; Rosenberger et al., 2013, von Saltzwedel, 2014) and two Collembola species (von Saltzwedel, Scheu & Schaefer, 2016) that were collected across several countries in Europe. Three of the five oribatid mite species are parthenogenetic. We did not include the parthenogenetic Collembola *Parisotoma notabilis* in our analyses, which is also represented with a Europe-wide sampling, because multiple genetic lineages (cryptic species) have already been reported for this species (Porco et al., 2012a; von Saltzwedel, Scheu & Schaefer, 2017). *Isotomiella minor*, another parthenogenetic Collembola species, was excluded, because reference databases did not provide sequences of congeneric species. The Collembola *Lepidocyrtus* was omitted, because this genus has been reported to be a species complex (Cicconardi et al., 2010) with uncertain status of the species *L. cyaneus,* which appears to be polyphyletic within *L. lanuginosus* (Zhang et al., 2018; Zhang et al., 2019). We analysed the performance of COI for species delimitation using distance- and character-based methods. We estimated optimal barcoding thresholds with the *spider* package in R. Afterwards, we used the ASAP algorithm (Assemble Species by Automated Partitioning; Puillandre, Brouillet & Achaz, 2021) to check the distribution of genetic distances and size of the barcoding gap. This method is an improved version of the Automated Barcode Gap Discovery (ABGD; Puillandre et al., 2012), which partitions single-locus datasets into hypothetical species by re-iteratively finding the best partitions that separate nominal species in the dataset using genetic distances. Different from ABGD, this version does not require *a priori* values and provides scores for each partition, which helps users to identify the best partition. Additionally, we calculated a Maximum Likelihood tree for the complete Collembola and Oribatida datasets, to check if a character-based method accomplishes accurate species assignment. If the distance-based methods (threshold optimization and ASAP) ignore important diagnostic characters in the datasets, this would be a meaningful

alternative method (DeSalle, Egan & Siddal, 2005). We also performed a rarefaction analysis to quantify the representativeness of the sample sizes for species haplotype diversity.

Materials & Methods

Taxa collection:

For Oribatida, BOLD delivered 12,252 records (search term: "Sarcoptiformes") with species names, which represent 710 species; NCBI delivered 29,047 records (search term: "Oribatida COI") with a sequence length between 500 and 800 base pairs. For Collembola (search term: "Collembola"), BOLD delivered 62,681 records with species names, which represent 1,544 species, and NCBI (search term: "Collembola COI") had 51,684 sequences with a sequence length between 500 and 800 base pairs. Many records in NCBI do not have a geographic reference, and most sequences in BOLD are from various geographic regions, predominantly coming from North America (Centre for Biodiversity Genomics). Analysing specimens from different continents could generate confounding effects due to ancient geographic isolations. To obtain a comparable dataset for all investigated species we therefore decided to restrict our analyses to sequences published by Rosenberger, 2011; Rosenberger et al., 2013; von Saltzwedel et al., 2014; von Saltzwedel, Scheu & Schaefer, 2016 (Table 1), which have a comparable sampling across Europe. For both Collembola and one Oribatida species (*Oppiella nova*), sequences of the nuclear gene 28S rDNA of the same individuals were also available in NCBI (Supplemental Table S1) and used ~~for checking~~ if genetic divergences are congruent between the mitochondrial and nuclear genes. The dataset of *Oppiella nova* (Oppiidae) differs as it contains sequences from different habitats collected only in Germany. Further, only a single congeneric sequence was available for this genus (*O. subpectinata*), but several sequences from species of other genera in the family Oppiidae. We decided to include this species into our analysis, but checked if a barcoding gap is present at genus level. Oppiidae species are very small, their body size in general ranges from 130 to 300 µm (except *Oppia nitens*, with a body size of >400 µm), which makes species determination very laborious and explains why this family commonly is not resolved to lower taxonomic levels in community studies. Confirmation of a barcoding gap and accurate species delimitation at genus level for this family would be helpful for future DNA-based biodiversity assessments, because Oppiidae is a species rich and very common family across many habitats, reaching high

abundances and even being the dominant taxon in many oribatid mite communities (Zaitsev et al., 2002; Bluhm, Maraun & Scheu, 2016).

For Collembola, we selected geographically comparable datasets for two sexual species, *Folsomia quadrioculata* and *Ceratophysella denticulata*. We had to omit the parthenogenetic species *Isotomiella minor* because congeneric sequences were inadequate for this study (only two sequences of *Isotomiella* sp. and one sequence of *I. paraminor*).

First, congeneric taxa were downloaded from BOLD and NCBI. Second, species assignment and barcoding gap analyses were performed with two global datasets, including all Collembola and Oribatida species, respectively. Third, for a more detailed analysis, the global datasets were separated into local datasets, each comprising all sequences of a genus (family in Oppiidae).

Species delimitation, character- and distance-based

All sequences of a genus, and all sequences of Oppiidae were aligned separately in AliView v1.28 (Larssen, 2014) using default settings and trimmed to the approximately shortest sequence. For the global dataset, all alignments of Collembola and Oribatida were combined in two separate files and re-aligned using default settings. All alignments were gap-free and did not contain any stop-codons. In total, we separately analysed two global datasets that contained all Oribatida and all Collembola, respectively, and seven local datasets, one for each genus and one for the family Oppiidae.

Barcoding thresholds were estimated within a range from 1 % to 20 % distance, at intervals of 1 % for all datasets in R using the threshOpt( ) function in the *spider* package (Brown et al., 2012). The potential number of partitions and the corresponding barcoding gap were estimated for each local dataset using the ASAP web application (https://bioinfo.mnhn.fr/abi/public/asap/; Puillandre, Brouillet & Achaz, 2021), providing the sequence alignment, selecting the K2P parameter as model of sequence evolution and the remaining parameter as default settings. Intra- and interspecific genetic distances (corrected with K2P) were plotted with *ggplot2* (Wickham, 2016) and *gridExtra* (Auguie, 2017) to visualize the barcoding gap. Two plots were generated for all datasets, one using the species names (morphotype) for inter- and interspecific assignment and one in which species names were replaced by the number of subsets estimated by ASAP, which equal the number of hypothetical (cryptic) species. The two plots visualize the barcoding gap based on morphological and genetic partitions, respectively. Alternative visualizations for analysing intra-

and interspecific genetic distances are histograms (Supplemental Figs. S1-S2) and scatterplots (Phillips, Gillis & Hanner, 2022; Supplemental Figs. S3-S4) and are provided in the Supplementary.

For character-based analyses, all datasets were collapsed to haplotypes using FaBox Haplotype Collapser (Villesen, 2007) to exclude identical sequences and to reduce the number of sequences to informative taxa for the phylogenetic tree construction. Maximum Likelihood trees with 500 bootstrap replicates were calculated for the global Collembola and global Oribatida and the 28S rDNA datasets (*Oppiella*, *Ceratophysella, Folsomia*) using the optim.pml( ) function of the *phangorn* package (Schliep, 2011, Schliep et al., 2017)

Representativeness of haplotype diversity in datasets:

Rarefaction was conducted for all haplotypes for which we had geographic sampling information (Collembola: *C. denticulata, F. quadrioculata*; Oribatida: *A. coleoptrata, N. silvestris, O. nova, P. peltifer, S. magnus*). Analysis was performed with the *iNEXT* package (Chao et al., 2014; Hsie, Ma & Chao, 2022) in R with 1,000 bootstrap replicates, using species richness (q=0) and exponential Shannon entropy (q=1) as measures of diversity.

Results

Datasets included 970 Oribatida and 612 Collembola sequences of COI, and alignments were between 507 bp and 657 bp long (Table 2) and covered the standard barcoding region of COI. Only a few sequences were below 500 bp long, predominantly in the oribatid mite genus *Steganacarus* and the Collembola genus *Folsomia*.

Barcoding gap threshold detection for different genetic distances

The global datasets (all Oribatida, all Collembola) had relatively high cumulative errors (false positives and false negatives, Fig. 1; Table 3). The optimized local barcoding threshold for the local datasets differed among taxa (Fig. 1; Table 3). One dataset had a narrow threshold without species mismatches (*Achipteria*, 15 %), others had large threshold ranges without mismatches (*Nothrus, Platyntohrus* and *Ceratophysella*), and for the remaining datasets it was not possible to define a barcoding threshold without any mismatches (Oppiidae, *Steganacarus, Folsomia*). The optimal barcode thresholds at genus level exceeded the standard barcoding threshold of 2 %-3 %

by 1 % (*Platynothrus*) and up to 6 % (*Achipteria*), except in the two genera *Nothrus* and *Ceratophysella*.

Distance-based specimen assignment with ASAP

The ASAP algorithm provides scores for the ten most probable partitions. For all datasets, and in all partitions, ASAP found more subsets than nominal species, i.e. the datasets likely contained more (i.e., cryptic) species than were morphologically determined (Table 4). The ASAP partition with the smallest number of subsets increased the number of morphological species to hypothetical species (or genetic lineages) from five to seven (*Nothrus, Platynothrus*), from eleven to 21 (*Ceratophysella*), from three to 14 (*Achipteria*), from six to 25 (*Steganacarus*) and from nine to 43 (*Folsomia*). The highest numbers of hypothetical species, or additional genetic lineages, were detected in species with the densest sampling, i.e. *A. coleoptrata*, *S. magnus, C. denticulata* and *F. quadrioculata* (Table 1). Interestingly, in the two parthenogenetic genera *Nothrus* and *Platynothrus*, only two additional hypothetical species were detected by ASAP, which is little compared to the other genera.

Distance-based barcoding gap with ASAP

~~Ideal for~~ accurate specimen assignment ~~is~~ a gap between the largest genetic distance within and the smallest distance between species. We compared the distribution of intra- and interspecific distances of nominal species with that of the genetic lineages inferred by ASAP (Fig. 2), selecting the partitions with the least number of subsets. Our analysis consistently demonstrated that genetic distances of COI within and between morphologically assigned species overlap, which makes accurate species assignment impossible. The parthenogenetic oribatid mite genus *Nothrus* was a single exception, which showed a clear barcoding gap for morphologically assigned species. When genetic lineages (ASAP subsets) were considered, a barcoding gap between intra- and interspecific distances was present. Overall, the assignment of genetic lineages to morphospecies reduced the overlap of intra- and interspecific genetic distances considerably in all datasets, However, the effect was much more pronounced in Collembola than Oribatida and generated a barcoding gap that spanned a range of more than 10 % between intra- and interspecific genetic distances. Among Oribatida, the effect of splitting morphotypes into genetic lineages was not ~~that~~ strong. However, for two Oribatida datasets, *Achipteria* and *Platynothrus*, the choice of using the partition with the lowest number of subsets was too conservative because the resulting barcoding gap was very

narrow. The barcoding gap thresholds estimated for the partition with the lowest number of subsets ranged in Oribatida from 6.9 % (*Achipteria*), 12.2 % (*Nothrus*), 15.0 % (*Steganacarus*, *Platynothrus*, all Oribatida) and 15.8 % (Oppiidae); in Collembola from 8.0 % (*Folsomia*), 11.3 % (all Collembola) and 13.8 % (*Ceratophysella*). Notably, the intraspecific distances of the genetic lineages of *Platynothrus* show three clusters in distribution frequencies (< 3 %, at 3-8 %, 14-17 %) that likely represent three genetic lineages in *P. peltifer* (Table 1).

The outliers, i.e. single datapoints scattered within the range of the barcoding gap, likely belong to sequences that were considerably shorter than the average sequences. Both Collembola datasets were more heterogeneous in sequence lengths than the Oribatida datasets. Only the datasets of *Steganacarus* and Oppiidae also had very short sequences compared to the median sequence lengths, and both also had outliers after splitting morphotypes into genetic lineages. A few outliers remained for the genus *Nothrus*, which likely belonged to the species *N. palustris* and *N. pratensis.* After splitting both species into two genetic lineages as proposed by ASAP, the outliers disappeared except for one, which likely belonged to one very short sequence (399 bp) of *N. anauniensis*.

Character-based specimen assignment with Maximum Likelihood

Reliability of specimen assignment based on phylogenetic inference and therefore on molecular characters was very poor in Collembola (Fig. 3). The two genera *Ceratophysella* and *Folsomia* and the species within each genus were not monophyletic. Species clustered within clades of ~~different~~ species several times. The topology of Oribatida supported monopohyly for most genera (Fig. 4). The species in the genus *Nothrus* and *Platynothrus* were also monophyletic. Only *Nothrus pratensis* separated into two highly supported, non-monophyletic clades. Within the genus *Steganacarus* all species were monophyletic, except *S. magnus* which formed five clades, two with 100 % bootstrap support. The species *S. carinatus* was monophyletic, but separated into two highly supported clades. One sequences of *S. applicatus* clustered within *S. carinatus* while the remaining sequences were monophyletic with very high support. Possibly this single sequence represents a misidentified individual. Most genera of Oppiidae were monophyletic, except for *Disorrhina* and *Oppia*. The sequences assigned to *Oppia* sp. were sister to one clade of *O. nitens* with very high bootstrap support. It is possible that these sequences belong to the species *O. nitens*. All remaining species were monopohyletic with 100 % bootstrap support. The genus *Achipteria* was represented

by only three species. The two species *A. howardi* and *A. catskillensis* were monophyletic, the sequences of *A. coleoptrata* were non-monophyletic.

Nuclear gene

The uncorrected p-distances among 28S rDNA in *C. denticulata* were relatively high across all sequences the maximum genetic distances were 5.6 %, but the mean distances were only 0.16 % (median 3.2 %). The different haplotypes corresponded very well with the seven genetic lineages suggested by ASAP (Supplemental Table S2), i.e. each 28S rDNA haplotype included a single ASAP lineage. However, the two datasets were not entirely congruent, i.e. seven specimens of the 28S rDNA dataset were not represented as COI sequences, and four specimens in the COI dataset were not present in the 28S rDNA dataset. In *F. quadrioculata*, the 24 genetic lineages did not reflect at all the 28S rDNA sequences. The nuclear gene represented only three haplotypes with uncorrected p-distances below 1 % (max: 0.35 %, mean: 0.16 %, 0.18 %). These 28S rDNA haplotypes comprised nine, three and one COI lineages that were identified by ASAP, respectively. Notably, only 56 specimens of 28S rDNA were represented from the 166 specimens of the COI nucleotide dataset. Among *O. nova* p-distances of 28S rDNA were also small, below 2 % (maximum 1.98 %, mean 0.43 %, median 0.29 %). In contrast to the two species above, each of the nine genetic lineages of *O. nova* supported by ASAP carried different 28S rDNA haplotypes, e.g. in one common COI lineage that comprised 30 specimens (lineage_1; Supplemental Table S2), individuals represented twelve (slightly) different 28S rDNA haplotypes.

Representativeness of sampling effort

Rarefaction curves (Fig. 5) showed that Oribatida species had more haplotypes than Collembola and that sexual Oribatida species (*A. coleoptrata, S. magnus*) had more haplotypes than parthenogenetic Oribatida (*P. peltifer, N. silvestris*). Further, Collembola reached saturation in species diversity at a sampling size of less than 200 individuals (for COI and 28S rDNA), the pattern was similar for the parthenogenetic Oribatida *N. silvestris*. However, the parthenogenetic Oribatida *P. peltifer* and both sexual Oribatida species did not reach saturation at a sampling size of more than 600 individuals and the expected diversity exceeded that of 200 haplotypes.

Discussion

This study tested the validity of a barcoding gap and the applicability of the standard barcoding gene COI for species assignment in two of the most species rich and abundant taxa of soil-living invertebrates, Collembola and Oribatida. The analysed datasets comprised two genera of Collembola with eleven and nine species, respectively. Oribatida datasets comprised four genera with three to six species per genus, and one family-level dataset with ten species in six genera.

Our results showed that correct species assignment was possible within some genera, but not all. However, both distance- and character-based methods were not able to assign species without mismatches when all Collembola or all Oribatida were analysed together. This is likely due to the different ranges of intraspecific genetic distances, demonstrating the absence of a general (global) barcoding gap for COI in these taxa. The genetic divergence separating intra- and interspecific distances differed among taxa and exceeded the standard species threshold of 3 % intraspecific genetic distance in all but one species, indicating that taxa-specific thresholds should be applied for correct specimen assignment (Phillips, Gillis & Hanner, 2022). Here, the application of algorithms that dynamically adjust thresholds for sequence clusters, and therefore apply flexible thresholds, could improve species assignment in soil invertebrates (James, Luczak & Girgis, 2018; Chiu & Ong 2022).

Absence of a global barcoding gap in the COI gene seems to be particularly relevant for soil-living animals and hampers the application of automated specimen assignments in DNA-based biodiversity surveys such as metabarcoding. Absence of a global barcoding gap had also been demonstrated for Annelida, among which earthworm taxa accounted for one third of interspecific comparisons with 0 % genetic divergence (Kvist, 2016). In metabarcoding studies, Collembola had a high failure rate and high numbers of false positives for species assignments based on public databases and COI (Recuero, Etzler & Caterino, 2023). Among mites, specimen assignment is in general correct at least to family and order level (Oliverio et al., 2018; Ustinova et al., 2021; Young, deWaard & Hebert, 2021; Young & Hebert, 2022; Recuero, Etzler & Caterino 2023). General explanations for failures in species assignments include the lack of completeness and misidentified individuals in reference databases, geographic underrepresentation of species and a neglect of assigning genetic lineage identities to sequences in reference databases (Kvist, 2016; Martinsson, Rhoden & Erseus 2016; Young et al., 2019; Young, deWaard & Hebert, 2021; Phillips, Gillis & Hanner, 2022; Recuero, Etzler & Caterino, 2023). The rarefaction analysis demonstrated that genetic diversity is exceptionally high within morphospecies of soil-living

invertebrates, and more genetic diversity is to be expected in additional samples. In particular, rarefaction curves for Oribatida did not reach saturation without sampling hundreds of additional individuals. For Collembola the number of expected COI haplotypes is lower as curves reached saturation at an expected sampling size of about 200 individuals, indicating that required sampling effort can be reached sooner than in Oribatida.

Results of this study provide an additional explanation why molecular species assignment often fails in Collembola and Oribatida. The more detailed analysis of the individual datasets at genus level showed that intra- and interspecific distances of taxa greatly overlapped, demonstrating the absence of a barcoding gap between species for all taxa, except for the parthenogenetic Oribatida genus *Nothrus*. The automated partitioning of datasets based on genetic distances (ASAP) suggested that each morphospecies (except most species within *Nothrus* and *Platynothrus*) consists of several genetic lineages, indicating the presence of putative or cryptic species. After assigning individuals according to genetic lineages, a barcoding gap between intra- and interspecific distances became apparent, but it still exceeded the standard threshold of 3 %. Alternative partitions in the species assignment analyses also opted for smaller thresholds, but resulted in even more genetic lineages. From a conservative approach, the two sexual Oribatida species *A. coleoptrata* and *S. magnus* comprised twelve and 18 genetic lineages, respectively, with the relatively high barcoding threshold estimates of 6.9 % (*A. coleoptrata*) and 15.0 % (*S. magnus*). The Collembola species *C. denticulata* consisted of seven genetic lineages (barcoding gap of 13.8 %) and *F. quadrioculata* of 24 genetic lineages (barcoding threshold of 8.0 %). Notably, the parthenogenetic Oribatida species *O. nova* separated only into nine genetic lineages, *P. peltifer* into three and *N. silvestris* remained a single species which was consistent with morphological assignments.

In contrast to our hypothesis, the detection of a barcoding gap and thus species delimitation worked well for the parthenogenetic, but not for the sexual taxa. Species boundaries of *Nothrus* were clear and unequivocal. However, intra- and interspecific distances among *Platynothrus* overlapped, likely due the presence of three genetic lineages in *P. peltifer*. This is consistent with previous studies that identified seven genetic lineages in *P. peltifer* based on a transcontinental sampling, and demonstrated that lineages are consistent with species based on the 4X rule of parthenogenetic speciation (Heethoff et al., 2007; Birky & Barraclough, 2009; Birky et al., 2010).

Detection of deeply divergent genetic lineages in morphological consistent species is a common phenomenon and detection rate of cryptic species accelerated with the application of molecular identification tools (Bickford et al., 2007; Pfenninger & Schwenk, 2007; Skoracka et al., 2015; Struck et al., 2018). However, it remains important to consider these putative species carefully based on barcoding approaches, as delimitation is based only on a single genetic marker. The putative genetic lineages were highly congruent with nuclear haplotype diversity in *C. denticulata*, but not in *F. quadrioculata* and *O. nova*. Interestingly, genetic variance of nuclear and mitochondrial genes was opposite in the two latter species. In *F. quadrioculata* a single nuclear haplotype comprised many COI lineages, but in *O. nova* a single COI lineage comprised several nuclear haplotypes. This suggests that different selective forces might act on mitochondrial and nuclear genes in the two species. The higher mutation rate of mitochondrial compared to nuclear genes explains the higher diversity in COI in *F. quadrioculata*, indicating relatively recent divergence of lineages that had not yet been accompanied by variation in the nuclear gene. By contrast, in *O. nova* the mitochondrial gene shows relatively little variation, which likely is related to stronger purifying selection in parthenogenetic Oribatida (Brandt et al., 2017, Brandt et al., 2021). The two other parthenogenetic species (*N. silvestris* and *P. peltifer*) also show very little genetic variation, unfortunately, no additional genes were available for these taxa.

Our results demonstrate, that soil-living microarthropods comprise deeply divergent genetic lineages. Barcoding or metabarcoding studies based on single genes will therefore likely result in high numbers of unassigned reads or overestimate species numbers and consequently misrepresent species richness in communities. Potential species status should therefore be corroborated with an integrative taxonomic approach using multiple genetic markers and, if possible, re-examination of morphotypes (Schäffer, Kerschbaumer & Koblmüller 2019; Lienhard & Krisper, 2021). However, morphological differences are often subtle, making traditional determination of soil microarthropods even more challenging. The nuclear 28S rDNA gene has been proposed as secondary barcoding marker for Oribatida (Lehmitz & Decker, 2017), but its applicability in a wider geographic range and different habitats has not been tested. Alternatively, metagenomic studies provide multiple genes per specimen which likely improves accuracy in specimen assignment. However, similar to metabarcoding based on single genes such as COI and/or 28S rDNA, successful application of metagenomics depends on representative reference databases. Notably, single reference genomes, or one/few barcodes per species will not cover intraspecific

variation. Species with high intraspecific genetic variance would require "pan-barcodes" i.e., multiple barcodes from individuals that were sequenced across the range of a species to cover the extent of its intraspecific genetic variance.

The limited taxon sampling in this study demonstrates that even for the relatively intensive sequenced COI gene, databases do not provide taxonomic breadth for reliable species delimitation of Collembola and Oribatida. It is possible that species assignment will improve with a better reference database, but it is also important to understand the mechanisms that explain the barcoding gap, i.e., the substantial genetic divergence of COI sequences between closely related Collembola and Oribatida taxa. It is unknown if genetic variance is neutral or adaptive, or if mitonuclear or environmental interactions (Hill, 2020) generate the genetic structure in soil-living microarthropods. Fixation of neutral variance is one likely mechanism in the investigated taxa. The high numbers of haplotypes and nucleotide diversity suggest that COI is already highly saturated in these species. Many Collembola and Oribatida species are very abundant in local communities, suggesting high effective population sizes. This could enable the maintenance of neutral allelic variation and blur a barcoding gap in order to maintain the highly conserved protein sequence of COI. Repeated episodes of extreme population bottlenecks can also generate a barcoding gap between species. However, this is unlikely because high genetic variance in general argues against repeated population bottlenecks. However, the Oribatida species *N. silvestris* shows exceptionally low genetic variance compared to the other taxa, and consists of a single genetic lineage. It is possible that the low genetic variance resulted from a bottleneck this species experienced during Quaternary glaciations (<2.6 mya). Molecular divergence times among genetic lineages in the other species are several million years old, most date back to the Miocene (23-5 mya) and support the accumulation of neutral variance by genetic drift in Oribatida and founder events in Collembola (Rosenberger et al., 2013; von Saltzwedel, Scheu & Schaefer, 2016). Directional selection on mitochondrial genotypes and disrupted gene flow can lead to rapid divergence among populations. Collembola and in particular Oribatida are poor active dispersers due to their small body size, which reduces gene flow among populations and is a possible explanation for mitochondrial lineages corresponding with nuclear 28S rDNA haplotypes and sampling locations in *C. denticulata* (Porco et al., 2012b; von Saltzwedel, Scheu & Schaefer, 2016). However, reduced gene flow seems unlikely in *F. quadrioculata* due to the low genetic variance in the nuclear 28S rDNA gene compared to the highly variable mitochondrial COI gene.

Genetic distances among lineages suggest maintenance of relatively ancient divergences, which argues against rapid divergence and disrupted gene flow. Further, this explanation does not apply for parthenogenetic species. Apparently, different mechanisms seem to account for the genetic variance in COI within species of Collembola and Oribatida. This is not surprising, considering that the species in this study likely are separated by tens to hundreds of millions of years, each having its own evolutionary trajectory (Schaefer et al., 2010; Schaefer & Caruso, 2019; Leo et al., 2019; Katz, 2020; van Straalen, 2021).

This study showed that metabarcoding using the standard gene COI is problematic when investigating biodiversity of soil invertebrates. Advances in second- and third-generation sequencing technologies can significantly contribute to improve the reliability of barcodes for genetically diverse and potentially cryptic species. Proposed as an alternative to small barcoding fragments, low coverage shotgun sequencing and genome skimming offer increased species discrimination by covering entire organellar genomes and ribosomal sequences (Coissac et al., 2016). PacBio sequencing technology generates reads of approximately 3 kb with very low error rates. This enables sequencing of nearly full-length marker genes and their flanking regions, which improves taxonomic resolution and reduces spurious Operational Taxonomic Units (OTUs) (Terdersoo & Anslan, 2019). Notably, genomes of Collembola and Oribatida typically range between 350 and 500 Mb, enabling to obtain reasonable sequencing read depth at moderate prices. Further, wet-lab protocols for genome sequencing of small, non-model invertebrates have been developed (Collins et al., 2023) and the results underscore the importance of taking intragenomic variance into account in order to integrate genetic and morphological species boundaries. We propose that characterizing pan-genomes is crucial for identifying species in soil invertebrates. (Tettelin et al., 2005). This approach will also contribute to develop informative barcoding genes (pan-barcodes) in soil invertebrates that lack a distinct barcoding gap. A pan-genome includes the complete set of genes shared by all individuals within a species and consists of conserved (core) and variable (accessory) gene regions (Golicz et al., 2020). The core genome covers all genes that are present in all individuals and the accessory genome includes the genomic regions that are variable among species. This variance is often due to ecological, geographical or reproductive boundaries (Reno et al., 2005). Accordingly, pan-genomes offer a holistic view of a species' genome, allowing to identify both conserved and variable regions that are suitable for designing robust barcoding markers, in particular in taxonomically challenging organisms.

512

513 *Conclusions*

514 This study demonstrated that intra- and interspecific genetic divergences in the standard barcoding
515 gene COI overlap in several species of Collembola and Oribatida. This is violating the assumption
516 of a barcoding gap, which is a precondition for molecular species assignment and questions the
517 applicability of the standard barcoding gene COI for soil-living microarthropods. Further, the
518 presence of deeply divergent genetic lineages within morphologically consistent species
519 emphasizes that (meta-)barcoding results solely based on a single genetic marker should be
520 interpreted carefully. Based on COI, morphologically consistent species comprised numerous
521 cryptic species. Without additional genetic and morphological data, the taxonomic status of these
522 cryptic species is questionable. The assignment of genetic lineages to sequences in references
523 databases and application of flexible or species-specific thresholds could improve specimen
524 assignment. However, the strong discrepancy between morphological conservativeness and
525 genetic variance of many soil invertebrates calls for a more general approach. We are promoting
526 to develop barcoding approaches with alternative sequencing technologies that generate more
527 genetic data than metabarcoding, like low-coverage shotgun sequencing of genomes (e.g., genome
528 skimming and metagenomics) or long-read sequencing of marker genes using third generation
529 sequencing technologies. Further, we advocate the construction and analysis of pan-genomes to
530 understand genetic species boundaries and to develop reliable barcoding markers that cover the
531 whole range of genomic variance of species (pan-barcodes). Regardless of the approach taken, it
532 is essential for reference databases to cover the intraspecific variability of a species throughout its
533 geographic range.

534

References

Anthony MM, Bender SF, van der Heijden MAG. 2023. Enumarating soil biodiversity. *Proceedings of the National Academy Sciences of the United States of America* 120:e2304663120 DOI: 10.1073/pnas.2304663120.

Antil S, Abraham J, Sripoona S, Maurya S, Daga J, Makhija S, Bhagat P, Gupta R, Sood U Lal R, Toteja R. 2022. DNA barcoding, an effective tool for species identification: a review. *Molecular Biology Reports* 50:761-778 DOI: 10.1007/s11033-022-08015-7.

Arribas P, Andújar C, Salces-Castellano A, Emerson B, Vogler A. 2021. The limited spatial scale of dispersal in soil arthropods revealed with whole-community haplotype-level metabarcoding. *Molecular Ecology* 30:48-61 DOI: 10.1111/mec.15591.

Auguie B. 2017. gridExtra: miscellaneous functions for "grid" graphics. https://CRAN.R-project.org/package=gridExtra.

Bardgett R, van der Putten W. 2014. Belowground biodiversity and ecosystem functioning. *Nature* 515:505-511 DOI: 10.1038/nature13855.

Barraclough TG, Birky CW jr, Burt A. 2003. Diversification in sexual and asexual organisms. *Evolution* 57:2166-2172.

Barrett RDH, Hebert PDN. 2005. Identifying spiders through DNA barcodes. *Canadian Journal of Zoology* 83:481-491 DOI: 10.1139/z05-024.

Behan-Pelletier V, Lindo Z. 2023. Oribatid Mites. Biodiversity, Taxonomy and Ecology. CRC Press DOI: 10.1201/9781003214649.

Bickford D, Lohman D, Sodhi N, Ng P, Meier R, Winker K, Infram K, Das I. 2007. Cryptic species as a window on diversity and conservation. *Trends Ecology and Evolution* 22:148-155 DOI: 10.1016/j.tree.2006.11.004.

Birky WC, Barraclough TG. 2009. Asexual speciation. In: Schoen I, Martens K, Dijk P, ed. *Lost Sex*. The Evolutionary Biology of Parthenogenesis. Berlin: Springer Verlag, 201-216.

Birky WC, Adams J, Gemmel M, Perry J. 2010. Using population genetic theory and DNA sequences for species detection and identification in asexual organisms. *PLoS One* 5.5:e10609 DOI: 10.1371/journal.pone.0010609.

Bluhm C, Scheu S, Maraun M. 2016. Temporal fluctuations in oribatid mites indicate that density-independent factors favour parthenogenetic reproduction. *Experimental and Applied Acarology* 68:387-407 DOI: 10.1007/s10493-015-0001-6.

Brandt A, Tran Van P, Bluhm C, Anselmetti Y, Dumas Z, Figuet E, Francois C, Galtier N, Heimburger B, Jaron K, Labédan M, Maraun M, Parker D, Robinson-Rechavi M, Schaefer I, Simion P, Scheu S, Schwander T, Bast J. 2021. Haplotype divergence supports long-term asexuality in the oribatid mite *Oppiella nova*. *Proceedings of the National Academy of Sciences of the United States of America* 118:e2101485118 DOI: 10.1073/pnas.2101485118

Brandt A, Schaefer I, Glanz J, Schwander T, Maraun M, Scheu S, Bast J. 2017. Effective purifying selection in ancnient asexual oribatid mites. *Nature Communications* 8: 873 DOI: 10.1038/s41467-017-01002-8

Brown SDJ, Collins RA, Boyer S, Lefort M-C, Malumbres-Olarte J, Vink CJ, Cruickshank RH 2012. SPIDER: an R package for the analysis of species identify and evolution, with particular reference to DNA barcoding. *Molecular Ecology Resources* 12:562-565 DOI: 10.1111/j.1755-0998.2011.03108.x.

Burns JM, Janzen DH, Hajibabaei M, Hallwachs W, Hebert PDN. 2007. DNA barcodes of closely related (but morphologically and ecologically distinct) species of skipper butterflies

(Hesperiidae) can differ by only one to three nucleotides. *Journal of the Lepidopterists' Society* 61:138-153.

Chao A, Gotelli NJ, Hsie TC, Sander EL, Ma KH, Colwell RK, Ellison AM .2014. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs* 84:45-67 DOI: 10.1890/13-0133.1.

Chao A, Gotelli NJ, Hsie TC, Sander EL, Ma KH, Colwell RK, Ellison AM .2014. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs* 84:45-67 DOI: 10.1890/13-0133.1.

Chahartaghi M, Scheu S, Ruess L. 2006. Sex ratio and mode of reproduction in Collembola of an oak-beech forest. *Pedobiologia* 50:331-340 DOI: 10.1016/j.pedobi.2006.06.001.

Chernova N, Potapov M, Savenkova Y, Bokova A. 2010. Ecological significance of parthenogenesis in Collembola. *Entomological Review* 90:23-38 DOI: 10.1134/S0013873810010033.

Chiu J, Ong R. 2022. Clustering biological sequences with dynamic sequence similarity threshold. *BMC Bioinformatics* 23:108 DOI: 10.1186/s12859-022-04643-9.

Ciancolo J, Norton R. 2006. The ecological distribution of reproductive mode in oribatid mites, as related to biological complexity. *Experimental and Applied Acarology* 40:1-25 DOI: 10.1007/s10493-006-9016-3.

Cicconardi F, Nardi F, Emerson B, Frati F, Fanciulli P. 2010. Deep phylogeographic divisions and long-term persistence of forest invertebrates (Hexapoda: Collembola) in the north-western Mediterranean basin. *Molecular Ecology* 19:386-400 DOI: 10.1111/j.1365-294X.2009.04457.x.

Coissac E, Hollingsworth P, Lavergne S, Taberlet P. 2016. From barcodes to genomes: extending the concept of DNA barcoding. *Molecular Ecology* 25:1423-1428 DOI: 10.1111/mec.13549.

Collins G, Schneider C, Bostjancic LL, Ulrich B, Axel C, Decker P, Ebersberger I, Hohberg K, Lecompte O, Merges D, Muelbaier H, Juliane R, Römbke J, Rutz C, Schmelz R, Schmidt A, Theissinger K, Veres R, Lehmitz R, Pfenninger M, Bálint M. 2023. The MetaInvert soil invertebrate genome resource provides insights into below-ground biodiversity and evolution. *Communications Biology* 6:1241 DOI: 10.1038/s42003-023-05621-4.

Collins R, Cruickshank R. 2013. The seven deadly sins of DNA barcoding. *Molecular Ecology Resources* 13:969-975 DOI: 10.1111/1755-0998.12046

Decaëns T. 2021. DNA metabarcoding illuminates the black box of soil animal biodiversity. *Molecular Ecology* 30:33-36 DOI: 10.1111/mec.15761

Decaëns T. 2010. Macroecological patterns in soil communities. *Global Ecology and Biogeography* 19:287-302 DOI: 10.1111/j.1466-8238.2009.00517.x.

Decaëns T, Jiménez J, Goia C, Measeay G, Lavelle P. 2006. The values of soil animals for conservation biology. *European Journal of Soil Biology* 42:S23-S38 DOI: 10.1016/j.ejsobi.2006.07.001.

DeSalle R, Egan MG, Siddal M. 2005. The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philosophical Transactions of the Royal Society London B* 360:1905-1916 DOI: 10.1098/rstb.2005.1722.

Dupont L, Porco D, Symondson WOC, Roy V. 2016. Hybridization relics complicate barcode-based identification of species in earthworms. Molecular Ecology Resources 16:883-894 DOI: 10.1111/1755-0998.12517.

Eisenhauer N, Bonn AA, Guerra C. 2019. Recognizing the quiet extinction of invertebrates. *Nature Communications* 50:10 DOI: 10.1038/s41467-018-07916-1.

Elias M, Hill RI, Willmott KR, Dasmahapatra KK, Brower V, Mallet J, Jiggins CD. 2007. Limited performance of DNA barcoding in a diverse community of tropical butterflies. *Proceedings of the Royal Society London B* 274:2881-2889 DOI: 10.1098/rspb.2007.1035.

Golicz AA, Bayer PE, Bhalla PL, Batley J, David E. 2020. Pangenomics comes of age: from bacteria to plant and animal applications. *Trends in Genetics* 36:132-145 DOI: 10.1016/j.tig.2019.11.006.

Grzywacz A, Jarmusz M, Walczak K, Skowronek R, Johnston NP, Szpila K. 2021. DNA barcoding identifies unknown females and larvae of Fannia R.-D. (Diptera: Fanniidae) from carrion succession experiment and case report. *Insects* 12:381 DOI: 10.3390/insects12050381.

Gulvik ME. 2007. Mites (Acari) as indicators of soil biodiversity and land use monitoring: a review. *Polish Journal of Ecology* 55:415-440

Hajibabaei M, Janzen DH, Burns JM, Hallwachs W, Hebert PDN. 2006. DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences of the United States of America* 103:968–971 DOI: 10.1073/pnas.0510466103.

Hebert PDN, Gregory TR. 2005. The promise of DNA barcoding for taxonomy. *Systematic Biology* 54:852–859 DOI: 10.1080/10635150500354886.

Hebert PDN, Cywinska A, Ball SL, deWaard JR, Jeremy R. 2003. Biological identifications through DNA barcodes. *Proceedings of the Royal Society London* B 270:313–321 DOI: 10.1098/rspb.2002.2218.

Hebert PDN, Ratnasingham S, deWaard JR. 2003. Barcoding animal life: *cytochrome c oxidase subunit 1* divergences among closely related species. *Proceedings of the Royal Society London* B 270:S96–S99 DOI: 10.1098/rsbl.2003.0025.

Hebert PDN, Stoeckle MY, Zemlak TS, and Francis CM. 2004. Identification of birds through DNA barcodes. *PLoS Biology* 2:e312 DOI: 10.1371/journal.pbio.0020312.

Heethoff M, Domes K, Laumann M, Maraun M, Norton RA, Scheu S. 2007. High genetic divergences indicate ancient separation of parthenogenetic lineages of the oribatid mite *Platynothrus peltifer* (Acari, Oribatida). *Journal of Evolutionary Biology* 20:392-402 DOI: 10.1111/j.1420-9101.2006.01183.x.

Hill G. 2020. Genetic hitchhiking, mitonuclear coadaptation, and the origins of mt DNA barcode gaps. *Ecology and Evolution* 10:9048-9059 DOI: 10.1002/ece3.6640.

Hsie TC, Ma KH, Chao A. 2022. iNEXT: iNterpolation and EXTrapolation for species diversity. http://chao.stat.nthu.edu.tw/wordpress/software-download.

Hubert N, Hanner R, Holm E, Mandrak E, Taylor E, Burridge M, Watkinson D, Dumont P, Curry A, Bentzen P, Zhang J, April J, Bernatchez L. 2008. Identifying Canadian freshwater fishes through DNA barcodes. *PLoSOne* 3:e2490 DOI: 10.1371/journal.pone.0002490.

James B, Luczak B, Girgis H. 2018. MeShClust: an intelligent tool for clustering DNA sequences. *Nucleic Acids Research* 46:e83 DOI: 10.1093/nar/gky315.

Katz AD. 2020. Inferring evolutionary timescales without independent timing information: an assessment of "universal" insect rates to calibrate a Collembola (Hexapoda) molecular clock. *Genes* 11:1172 DOI: :10.3390/genes11101172.

King RA, Tibble AL, Symondson WOC. 2008. Opening a can of worms: unprecedented sympatric cryptic diversity within British lumbricid earthworms. *Molecular Ecology* 17:4684-4698 DOI: 10.1111/j.1365-294X.2008.03931.x.

Kondrashov AS. 1988. Deleterious mutations and the evolution of sexual reproduction. *Nature* 336: 435-440.

Kvist S. 2016. Does a global DNA barcoding gap exist in Annelida? *Mitochondrial DNA* 27:2241-2252 DOI: 10.3109/19401736.2014.984166.

Larsson A. 2014. AliView: a fast and lightweigth alignment viewer and editor for large datasets. *Bioinformatics* 30:3276-3278 DOI: 10.1093/bioinformatics/btu531.

Lavelle P, Decaëns T, Aubert M, Barot S, Blouin M, Bureaeu F, Margerie P, Mora P, Rossi J. 2006. Soil invertebrates and ecosystem services. *European Journal of Soil Biology* 42:S3-S15 DOI: 10.1016/j.ejsobi.2006.10.002.

Lehmitz R, Decker P. 2017. The nuclear 28S gene fragment D3 as species marker in oribatid mites (Acari, Oribatida) from German peatlands. *Experimental and Applied Acarology* 71:259-276 DOI: 10.1007/s10493-017-0126-x.

Leo C, Carapelli A, Cicconardi F, Frati F, Nardi F. 2019. Mitochondrial genome diversity in Collembola: phylogeny, dating and gene order. *Diversity* 11:169 DOI: 10.3390/d11090169.

Lienhard A, Krisper G. 2021. Hidden biodiversity in microarthropods (Acari, Oribatida, Eremaeoidea, Caleremaeus). *Scientific Reports* 11:23123 DOI: 10.1038/s41598-021-02602-7.

Maraun M, Scheu S. 2000. The structure of oribatid mite communities (Acari, Oribatida): patterns, mechanisms and implications for future research. *Ecography* 23:374-382 DOI: 10.1038/s41598-021-02602-7.

Maraun M, Thomas T, Fast E, Treibert N, Caruso T, Schaefer I, Lu J, Scheu S. 2023. New perspectives on soil animal trophic ecology through the lens of C and N stable isotope ratios of oribatid mites. *Soil Biology and Biochemistry* 177:108890 DOI: 10.1016/j.soilbio.2022.108890.

Martinsson S, Rhoden C, Erseus C. 2016. Barcoding gap, but no support for cryptic speciation in the earthworm *Aporrectodea longa* (Clitellata: Lumbricidae). *Mitochondrial DNA 28:147-155* DOI: 10.3109/19401736.2015.1115487.

Meier R, Shiyang K, Vaidya G, Ng P. 2006. DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology* 55: 715–728 DOI: 10.1080/10635150600969864.

Meyer C, Paulay G. 2005. DNA barcoding: error rates based on comprehensive sampling. PLoS Biology 3:e422 DOI: 10.1371/journal.pbio.0030422.

Muller HJ. 1964. The relation of recombination to mutational advance. *Mutation Research* 1:2-9 DOI: 10.1016/0027-5107(64)90047-8.

Mutanen M, Kivelä SM, Vos RA, Doorenweerd C, Ratnasingham S, Hausmann A, Huemer P, Dincă V, van Nieukerken EJ, Lopez-Vaamonde C, Vila R, Aarvik L, Decaëns T, Efetov KA, Hebert PDN, Johnson A, Karsholt O, Pentinsaari M, Rougerie R, Segerer A, Tarmann G, Zahiri R, Godfray HCJ. 2016. Species-level para- and polyphyly in DNA barcode gene trees: strong operational bias in European Lepidoptera. *Systematic Biology* 65:1024-1040 DOI: 10.1093/sysbio/syw044.

Novo M, Almodóvar A, Fernández R, Trigo D, Díaz Cosín D. 2010. Cryptic speciation of homogastrid earthworms reveald by mitochondrial and nuclear data. *Molecular Phylogeny and Evolution* 56:507-512 DOI: 10.1016/j.ympev.2010.04.010.

Oliverio A, Gan H, Wixkings K, Fierer N. 2018. A DNA metabarcoding approach to characterize soil arthropod communities. Soil Biology and Biochemistry 125:37-43 DOI: 10.1016/j.soilbio.2018.06.026.

Orgiazzi A, Dunbar M, Panagos P, de Groot G, Lemanceau P. 2018. Soil biodiversity and DNA barcodes: opportunities and challenges. *Soil Biology and Biochemistry* 80:244-250 DOI: 10.1016/j.soilbio.2014.10.014.

Pfenninger M, Schwenk K. 2007. Cryptic animal species are homogeneously distributed among taxa and biogeographical regions. *BMC Evolutionary Biology* 71:121 DOI: 10.1186/1471-2148-7-121.

Pfingstl T, Schatz H. 2021. A survey of lifespans in Oribatida excluding Astigmata (Acari). *Zoosymposia* 20:007-027 DOI: 10.11646/zoosymposia.20.1.4.

Phillips JD, Gillis DJ, Hanner RH. 2022. Lack of statistical rigor in DNA barcoding likely invalidates the presence of a true species' barcode gap. *Frontiers in Ecology and Evolution* 10:859099 DOI: 10.3389/fevo.2022.859099.

Pollierer M. Langel R, Scheu S, Maraun M 2009. Compartmentalization of the soil animal food web as indicated by dual analysis of stable isotope ratios (15N/14N and 13C/12C). *Soil Biology and Biochemistry* 41:1221-1226 DOI: 10.1016/j.soilbio.2009.03.002.

Ponge J, Gillet S, Dubs F, Fedoroff E, Haese L, Sousa J, Lavelle P. 2003. Collembolan communities as bioindicators of land use intensification. *Soil Biology and Biochemistry* 35:813-826 DOI: 10.1016/S0038-0717(03)00108-1.

Porco D, Potapov M, Bedos A, Busmachiu G, Weiner W, Hamra-Kroua S, Deharveng L. 2012a. Cryptic diversity in the ubiquist species *Parisotoma notabilis* (Collembola, Isotomidae): a long-used chimeric species? *PLoS One* 7:e46056 DOI: 10.1371/journal.pone.0046056.

Porco D, Bedos A, Greenslade P, Janion C, Skarzynski D, Stevens M, Jansen van Vuuren B, Deharveng L. 2012b. Challenging species delimitation in Collembola-cryptic diversity among common springtails unveiled by DNA barcoding. *Invertebrate Systematics* 26:470 DOI: 10.1071/IS12026.

Potapov A, Semenina E, Korotkevich A, Kuznetsova N, Tiunov A. 2016. Connecting taxonomy and ecology: trophic niches of collembolans as related to taxonomic identity and life forms. *Soil Biology and Biochemistry* 101:20-31 DOI: 10.1016/j.soilbio.2016.07.002.

Potapov A, Bellini BC, Chown SL, Deharveng L, Janssens F, Kovac L, Kuznetsova N, Ponge J-F, Potapov M, Querner P, Russel D, Sun X, Zhang F, Berg MP. 2020. Towards a global synthesis of Collembola knowledge - challenges and potential solutions. *Soil Organisms* 92:161-188 DOI: 10.25674/so92iss3pp161.

Puillandre N, Lambert A, Brouillet S, Achaz G. 2012. ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology* 21:1864-1877 DOI: 10.1111/j.1365-294X.2011.05239.x.

Puillandre N, Brouillet S, Achaz G 2021. ASAP: assemble species by automatic partitioning. *Molecular Ecology Resource* 21:609-620 DOI: 10.1111/1755-0998.13281.

Ratnasingham S, Hebert PD. 2013. A DNA-based registry for all animal species: The barcode index number (BIN) system. *PLoS ONE* 8:e66213. DOI: 10.1371/journal.pone.0066213.

Recuero E, Etzler F, Caterino M. 2023. Most soil and litter arthropods are unidentifiable based on current DNA barcode reference libraries. *Current Zoology* DOI: 10.1093/cz/zoad051.

Reno M, Held N, Fields C, Burke P, Whitaker R. 2009. Biogeography of the *Sulfolobus islandicus* pan-genome. *Proceedings of the National Academy of Sciences* 106:8605-8610 DOI: 10.1073/pnas.0808945106.

Richard B, Decaëns T, Rougerie R, James S, Porco D, Hebert DP 2010. Re- integrating earthworm juveniles into soil biodiversity studies: species identification through DNA barcoding. *Molecular Ecology Resources* 10:606–14 DOI: 10.1111/j.1755-0998.2009.02822.x 10.1111/j.1755-0998.2009.02822.x.

Rosenberger M. 2011. Phylogeography in sexual and parthenogenetic European Oribatida. *Doctoral thesis.* DOI: http://dx.doi.org/10.53846/goediss-3325.

Rosenberger M, Maraun M, Scheu S, Schaefer I. 2013. Pre- and post-glacial diversifications shape genetic complexity of soil-living microarthropod species. *Pedobiologia* 56:79-87 DOI: 10.1016/j.pedobi.2012.11.003.

Rusek J. 1998. Biodiversity of Collembola and their functional role in the ecosystem. *Biodiversity and Conservation* 7:1207-1219

Santorufo L, Van Gestel C, Rocco A, Maisto G. 2012. Soil invertebrates as bioindicators of urban soil quality. *Environmtal Pollution* 161:57-63 DOI: 10.1016/j.envpol.2011.09.042.

Schaefer I, Caruso T. 2019. Oribatid mites show that soil food web complexity and close aboveground-belowground linkages emerged in the early Paleozoic. *Communications Biology* 2:387 DOI: 10.1038/s42003-019-0628-7.

Schaefer I, Norton RA, Scheu S, Maraun M. 2010. Arthropod colonization of land - linking molecules and fossils in oribatid mites (Acari, Oribatida). *Molecular Phylogeny and Evolution* 57:113-121 DOI: 10.1016/j.ympev.2010.04.015.

Schäffer S, Kerschbaumer M, Koblmüller S. 2019. Multiple new species: cryptic diversity in the widespread mite species Cymbaeremaeus cymba (Oribatida, Cymbaeremaeidae). Mol Phyl Evol 135:185-192 DOI: 10.1016/j.ympev.2019.03.008.

Schatz H, Behan-Pelletier V. 2008. Global diversity of oribatids (Oribatida: Acari: Arachnida). *Hydrobiologia* 595:323-328 DOI: 10.1007/s10750-007-9027-z.

Schliep KP. 2011. Phangorn: phylogenetic analysis in R. *Bioinformatics* 27:592-593 DOI: 10.1093/bioinformatics/btq706.

Schliep KP, Potts AJ, Morrison DA, Grumm GW. 2017. Intertwining phylogenetic trees and networks. *Methods in Ecology and Evolution* 8:1212-1220 DOI: 10.1111/2041-210X.12760.

Schneider K, Migge S, Norton R, Scheu S, Langel R, Reineking A, Maraun M. 2004. Trophic niche differentiation in soil microarthropods (Oribatida, Acari): evidence from stable isotope ratios (15N/14N). *Soil Biology and Biochemistry* 36:1769-1774 DOI: 10.1016/j.soilbio.2004.04.033.

Skoracka A, Magalhaes S, Rector B, Kuczynski L. 2018. Cryptic speciation in the Acari: a function of species lifestyles or our ability to separate species? *Experimental and Applied Acarology* 67:165-182 DOI: 10.1007/s10493-015-9954-8.

Struck T, Feder J, Bendiksby M, Birkeland S, Cerca J, Gusarov V, Kistenich S, Larsson K, Liow L, Nowak M, Stedje B, Bachmann L, Dimitrov D. 2018. *Trends in Ecology and Evolution* 33:153-163 DOI: 10.1016/j.tree.2017.11.007.

Tedersoo L, Anslan S. 2019. Towards PacBio-based pan-eukaryote metabarcoding using full-length ITS sequences. *Environmental Microbiology Reports* 11:659-668.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, DeBoy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R,

Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM. 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences* 102:13950-13955 DOI: 10.1073/pnas.0506758102.

Ustinova E, Schepetov D, Lysenkov S, Tiunov A. 2021. Soil arthropod communities are not affected by invasive *Solidago gigantea* Aiton (Asteraceae), based on morphology and metabarcoding analyses. *Soil Biology and Biochemistry* 159:108288 DOI: 10.1016/j.soilbio.2021.108288.

Valentini A, Pompanon F, Taberlet P. 2009. DNA barcoding for ecologists. *Trends in Ecology and Evolution* 24:110-117 DOI: 10.1016/j.tree.2008.09.011.

van Straalen NM. 2021. Evolutionary terrestrialization scenarios for soil invertebrates. *Pedobiologia* 87-88:150753 DOI: 10.1016/j.pedobi.2021.150753.

Villensen P. 2007. FaBox: an online toolbox for FASTA sequences. *Molecular and Ecology Notes* 7:965-968 DOI: 10.1111/j.1471-8286.2007.01821.x.

von Saltzwedel H, Maraun M, Scheu S, Schaefer I. 2014. Evidence for frozen-niche variation in a cosmopolitan parthenogenetic soil mite species (Acari, Oribatida). *PLoS One* 9:e113268 DOI: 10.1371/journal.pone.0113268.

von Saltzwedel H, Scheu S, Schaefer I. 2016. Founder events and pre-glacial divergences shape the genetic structure of European Collembola species. *BMC Evolutionary Biology* 16:148 DOI: 10.1186/s12862-016-0719-8.

von Saltzwedel H, Scheu S, Schaefer I. 2017. Genetic structure and distribution of *Parisotoma notabilis* (Collembola) in Europe: cryptic diversity, split of lineages and colonization patterns. *PLoS One* 12:e0170909 DOI: 10.1371/journal.pone.0170909.

Wardle D, Bardgett R, Klironomos J, Setälä H, van der Putten W, Wall D. 2004. Ecological linkages between aboveground and belowground biota. *Science* 304:1629-1633 DOI: 10.1126/science.1094875.

Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wiemers M, Fiedler K. 2007. Does the DNA barcoding gap exist? – a case study in blue butterflies (Lepidoptera: Lycaenidae). *Frontiers in Zoology* 4:8 DOI: 10.1186/1742-9994-4-8.

Young M, Proctor H, deWaard J, Hebert P. 2019. DNA barcodes expose unexpected diversity in Canadian mites. *Molecular Ecology* 28:5347-5359 DOI: 10.1111/mec.15292.

Young M, deWaard J, Hebert P. 2021. DNA barcodes enable higher taxonomic assignments in the Acari. *Scientific Reports* 11:15922 DOI: 10.1038/s41598-021-95147-8.

Young M, Hebert P. 2022. Unerathing soil arthropod diversity through DNA metabarcoding. *PeerJ* 10:e12845 DOI: 10.7717/peerj.12845.

Zaitsev AS, Chauvat M, Pflug A, Wolters V. 2002. Oribatid mite diversity and community dynamics in a spruce chronosequence. *Soil Biology and Biochemistry* 34:1919-1927.

Zhang B, Chen T, Mateos E, Scheu S, Schaefer I .2018. Cryptic species in *Lepidocyrtus lanuginosus* (Collembola: Entomobryidae) are sorted by habitat type. *Pedobiologia* 68:12-19 DOI: 10.1016/j.pedobi.2018.03.001.

Zhang B, Chen T, Mateos E, Scheu S, Schaefer I. 2019. DNA-based approaches uncover cryptic diversity in the European *Lepidocyrtus lanuginosus* species group (Collembola: Entomobryidae). *Invertebrate Systematics* 33:661-670 DOI: 10.1071/IS18068.

855 **Figure legends**

856 **Figure 1. Summary of the barcoding threshold optimization of the global and local datasets.**
857 Threshold between 1 % and 20 % genetic distances were analysed at intervals of 1 %. Light grey
858 bars indicate the number of false positive (no conspecific matches within threshold of query), dark
859 grey bars are false negatives (non-conspecific species match within threshold distance of query)
860 of the species assignments for the respective threshold (x-axis). Note the different scales of the y-
861 axis.
862

863 **Figure 2. Distribution of intra- (red violins) and interspecific (yellow violins) genetic**
864 **distances in morphological and genetic entities in Collembola and oribatid mites.** Distances
865 were calculated for each dataset based on the nominal species names (Morphotypes) and using the
866 same dataset but assigning sequences to genetic lineages (ASAP). The ASAP partition with the
867 smallest number of subsets was used to assign genetic lineages. Specimens that overlap in intra-
868 and interspecific distances cannot be assigned accurately to species based on COI. The splitting of
869 the dataset into genetic lineages created a barcoding gap that improved the accuracy of specimen
870 assignment. Solid blue lines indicate the 3 % genetic distances threshold, dashed lines represent
871 the genetic distances of the barcoding gap calculated with ASAP for the respective dataset
872 (Collembola: 11.3 %, *Ceratophysella*: 13.8 %, *Folsomia*: 8 %, Oribatida: 15 %, *Achipteria*: 6.9 %,
873 *Steganacarus*: 15 %, Oppiidae: 15.8 %, *Nothrus*: 12 %, *Platynothrus*: 15 %). Notice the different
874 scales of the y-axis.

875

876 **Figure 3. Phylogenetic tree of all Collembola for character-based species assignment.**
877 Likelihood tree based on 313 haplotypes of 612 COI sequences and 500 bootstrap replicates.
878 Monophyletic nodes were collapsed, bootstrap values >50 % are shown on nodes. The two genera
879 *Ceratophysella* and *Folsomia* are not monophyletic, and species within genera are also not
880 monophyletic.

881

882 **Figure 4. Phylogenetic tree of all Oribatida for character-based species assignment.**
883 Likelihood tree based on 514 haplotypes of 970 COI sequences and 500 bootstrap replicates.
884 Monophyletic nodes were collapsed, bootstrap values >50 % are shown on nodes. Grey circles on

branches highlight monophyletic lineages. Red circles highlight non-monophyletic lineages, indicating species for which character-based species assignment is problematic or equivocal. Grey circles with red outlines indicate species that are monophyletic but split into at least two clades. All genera but *Achipteria, Dissorhina*, *Oppia* and *Oppiella* are monophyletic. The single sequence of *Oppiella subpectinata* was sister to *Berniniella* and potentially represents a misidentified individual. A phylogenetic tree of 28S rDNA haplotypes is provided in Supplemental Figure S5.

Figure 5. Rarefaction of Collembola and Oribatida species. Only species with sampling site information are included for quantifying the representatives of genetic diversity in the different datasets. Collembola species reach soon saturation in haplotype diversity, while sexual Oribatida species (*A. coleoptrata, S. magnus*) do not reach saturation. The parthenogenetic Oribatida species *P. peltifer* also reaches saturation close to a sampling size of 600 individuals, but expected diversity is lower with less than 200 haplotypes (note the different scales of the y-axis). The parthenogenetic Oribatida species *N. silvestris* shows the lowest diversity and reaches soon saturation, indicating that sampling size was almost representative for the expected haplotype diversity in this species. Solid lines indicate the rarefaction, dotted lines the extrapolation. The tested diversity measures using *iNEXT* were species richness (q=0, red lines) and Shannon diversity (q=1, blue lines). Notably, the two indices are more similar in Collembola than in Oribatida. Rarefaction plots of 28S rDNA haplotypes are provided in Supplemental Figure S6.

Figure 1

Summary of the barcoding threshold optimization of the global and local datasets.

Threshold between 1 % and 20 % genetic distances were analysed at intervals of 1%. Light grey bars indicate the number of false positive (no conspecific matches within threshold of query), dark grey bars are false negatives (non-conspecific species match within threshold distance of query) of the species assignments for the respective threshold (x-axis). Note the different scales of the y-axis.

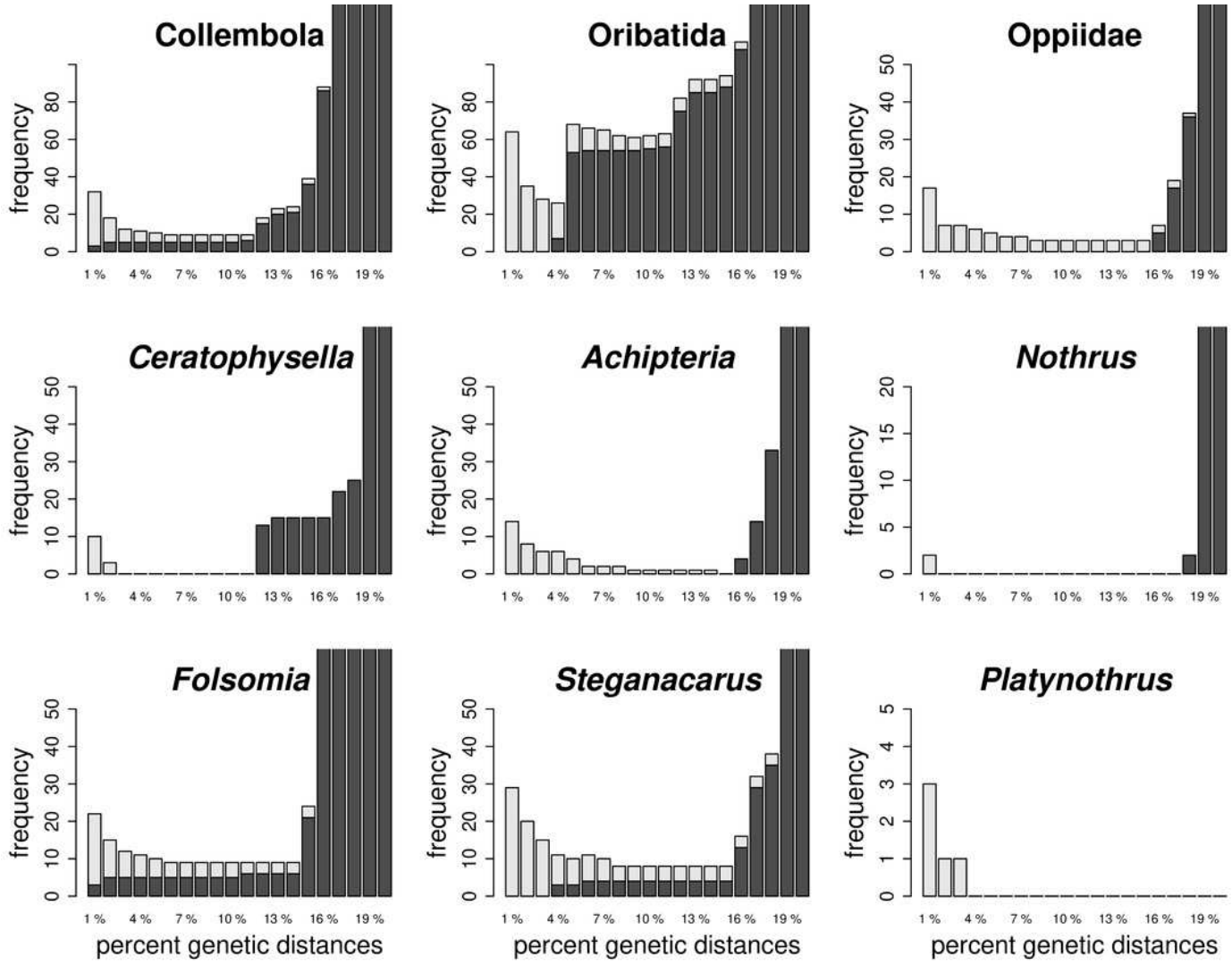

Figure 2

Figure 2. Distribution of intra- (red violins) and interspecific (yellow violins) genetic distances in morphological and genetic entities in Collembola and oribatid mites.

Distances were calculated for each dataset based on the nominal species names (Morphotypes) and using the same dataset but assigning sequences to genetic lineages (ASAP). The ASAP partition with the smallest number of subsets was used to assign genetic lineages. Specimens that overlap in intra- and interspecific distances cannot be assigned accurately to species based on COI. The splitting of the dataset into genetic lineages created a barcoding gap that improved the accuracy of specimen assignment. Solid blue lines indicate the 3 % genetic distances threshold, dashed lines represent the genetic distances of the barcoding gap calculated with ASAP for the respective dataset (Collembola: 11.3 %, *Ceratophysella*: 13.8 %, *Folsomia*: 8 %, Oribatida: 15 %, *Achipteria*: 6.9 %, *Steganacarus*: 15 %, Oppiidae: 15.8 %, *Nothrus*: 12 %, *Platynothrus*: 15 %). Notice the different scales of the y-axis.

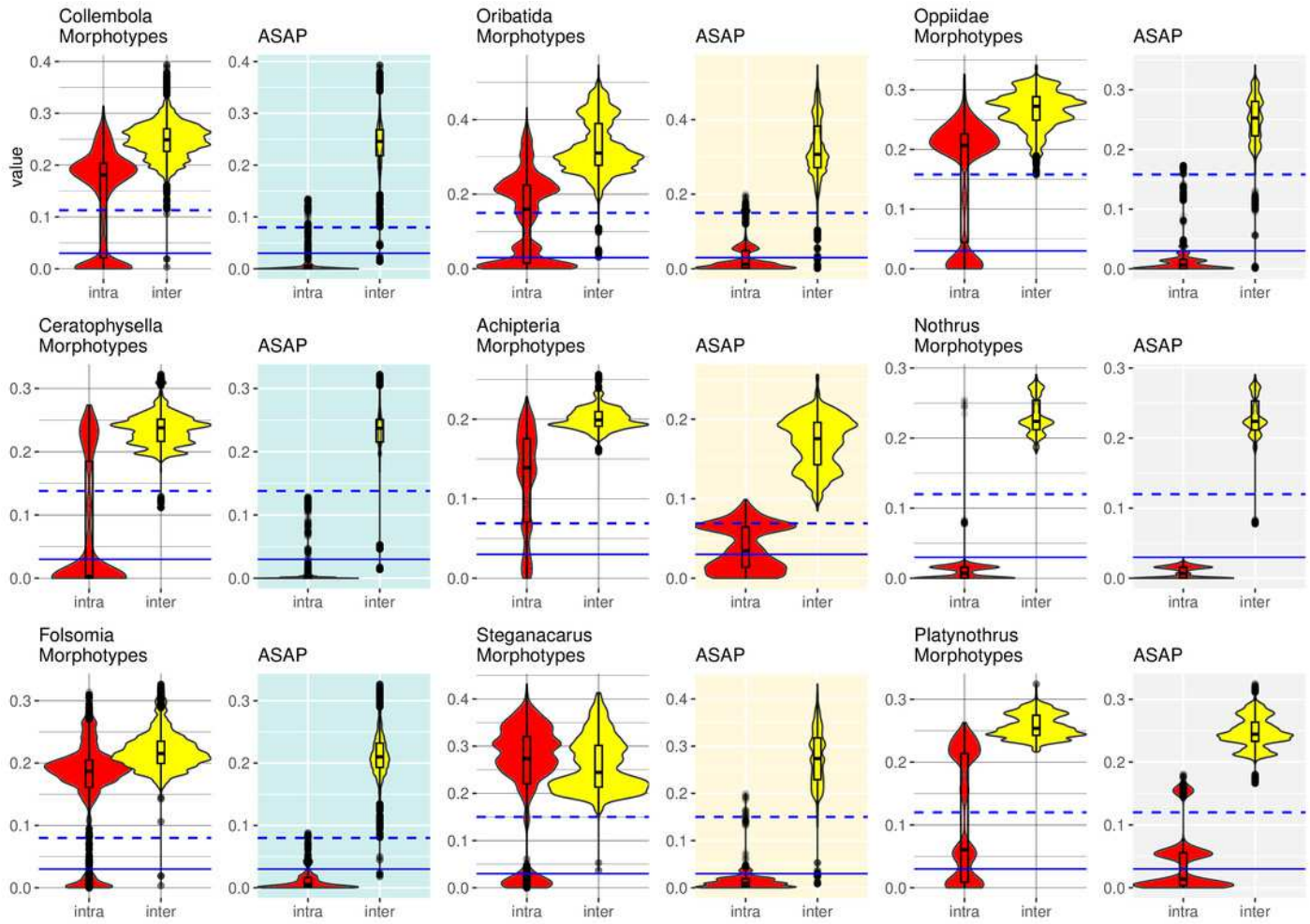
Collembola
Morphotypes
ASAP
Oribatida
Morphotypes
ASAP
Oppiidae
Morphotypes
ASAP
Ceratophysella
Morphotypes
ASAP
Achipteria
Morphotypes
ASAP
Nothrus
Morphotypes
ASAP
Folsomia
Morphotypes
ASAP
Steganacarus
Morphotypes
ASAP
Platynothrus
Morphotypes
ASAP
value
intra
inter

Figure 3

Phylogenetic tree of all Collembola for character-based species assignment.

Likelihood tree based on 313 haplotypes of 612 COI sequences and 500 bootstrap replicates. Monophyletic nodes were collapsed, bootstrap values > 50 % are shown on nodes. The two genera *Ceratophysella* and *Folsomia* are not monophyletic, and species within genera are also not monophyletic.

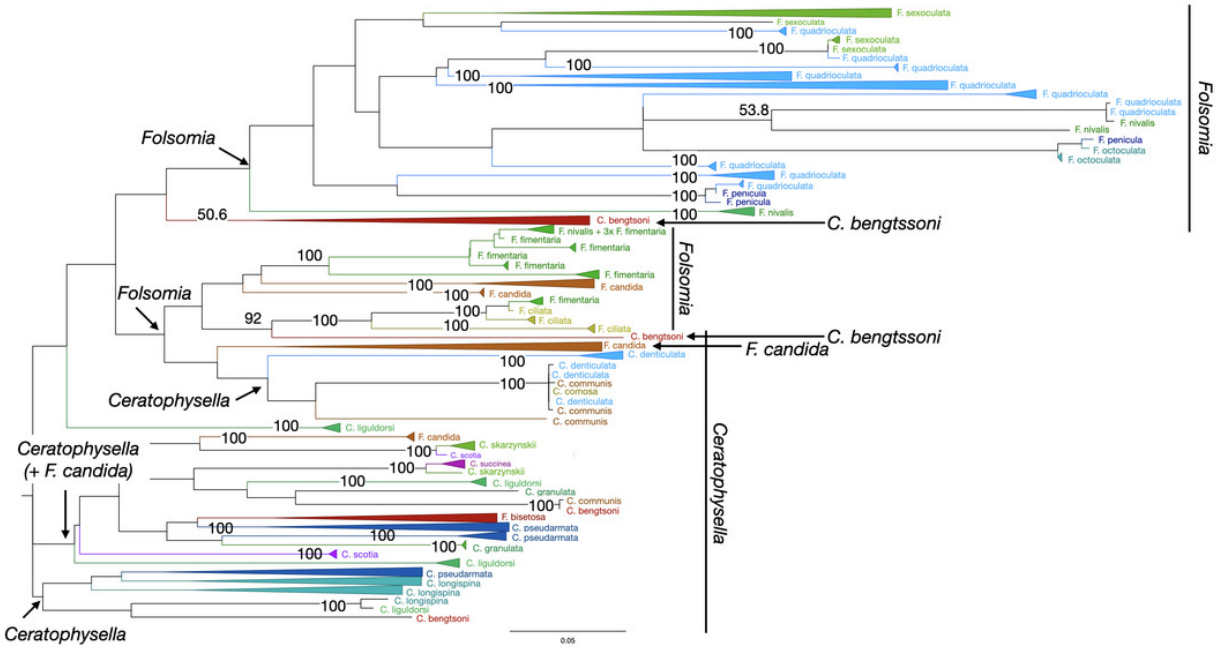
Folsomia
Folsomia
Ceratophysella
Ceratophysella
(+ F. candida)
Ceratophysella
Folsomia
Folsomia
Ceratophysella
C. bengtssoni
C. bengtssoni
F. candida
F. sexoculata
F. quadrioculata
F. nivalis
F. penicula
F. octoculata
F. fimentaria
F. candida
F. ciliata
C. bengtsoni
C. denticulata
C. communis
C. comosa
C. liguldorsi
C. skarzynskii
C. scotia
C. succinea
C. granulata
C. bengtsoni
F. bisetosa
C. pseudarmata
C. longispina
C. liguldorsi
C. bengtsoni
0.05

Figure 4

Phylogenetic tree of all Oribatida for character-based species assignment.

Likelihood tree based on 514 haplotypes of 970 COI sequences and 500 bootstrap replicates. Monophyletic nodes were collapsed, bootstrap values > 50 % are shown on nodes. Grey circles on branches highlight monophyletic lineages. Red circles highlight non-monophyletic lineages, indicating species for which character-based species assignment is problematic or equivocal. Grey circles with red outlines indicate species that are monophyletic but split into at least two clades. All genera but *Achipteria, Dissorhina*, *Oppia* and *Oppiella* are monophyletic. The single sequence of *Oppiella subpectinata* was sister to *Berniniella* and potentially represents a misidentified individual. A phylogenetic tree of 28S rDNA haplotypes is provided in Supplemental Figure S5.

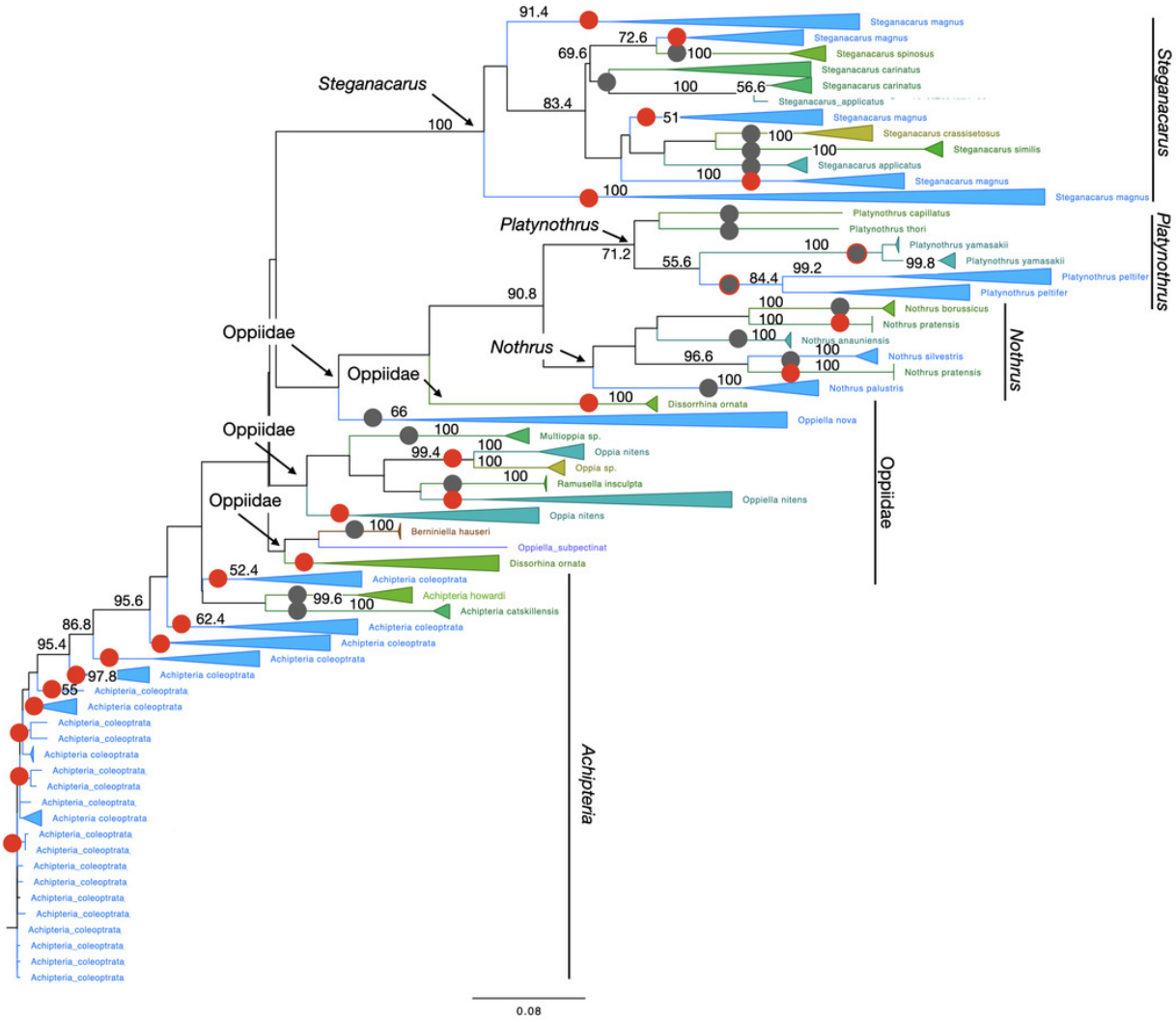
Steganacarus
Platynothrus
Nothrus
Oppiidae
Oppiidae
Oppiidae
Oppiidae
Steganacarus magnus
Steganacarus magnus
Steganacarus spinosus
Steganacarus carinatus
Steganacarus carinatus
Steganacarus_applicatus
Steganacarus magnus
Steganacarus crassisetosus
Steganacarus similis
Steganacarus applicatus
Steganacarus magnus
Steganacarus magnus
Platynothrus capillatus
Platynothrus thori
Platynothrus yamasakii
Platynothrus yamasakii
Platynothrus peltifer
Platynothrus peltifer
Nothrus borussicus
Nothrus pratensis
Nothrus anauniensis
Nothrus silvestris
Nothrus pratensis
Nothrus palustris
Dissorrhina ornata
Oppiella nova
Multioppia sp.
Oppia nitens
Oppia sp.
Ramusella insculpta
Oppiella nitens
Oppia nitens
Berniniella hauseri
Oppiella_subpectinat
Dissorhina ornata
Achipteria coleoptrata
Achipteria howardi
Achipteria catskillensis
Achipteria coleoptrata
Achipteria coleoptrata
Achipteria coleoptrata
Achipteria_coleoptrata
Steganacarus
Platynothrus
Nothrus
Oppiidae
Achipteria
0.08

Figure 5

Rarefaction of Collembola and Oribatida species.

Only species with sampling site information are included for quantifying the representatives of genetic diversity in the different datasets. Collembola species reach soon saturation in haplotype diversity, while sexual Oribatida species (*A. coleoptrata, S. magnus*) do not reach saturation. The parthenogenetic Oribatida species *P. peltifer* also reaches saturation close to a sampling size of 600 individuals, but expected diversity is lower with less than 200 haplotypes (note the different scales of the y-axis). The parthenogenetic Oribatida species *N. silvestris* shows the lowest diversity and reaches soon saturation, indicating that sampling size was almost representative for the expected haplotype diversity in this species. Solid lines indicate the rarefaction, dotted lines the extrapolation. The tested diversity measures using *iNEXT* were species richness (q=0, red lines) and Shannon diversity (q=1, blue lines). Notably, the two indices are more similar in Collembola than in Oribatida. Rarefaction plots of 28S rDNA haplotypes are provided in Supplemental Figure S6.

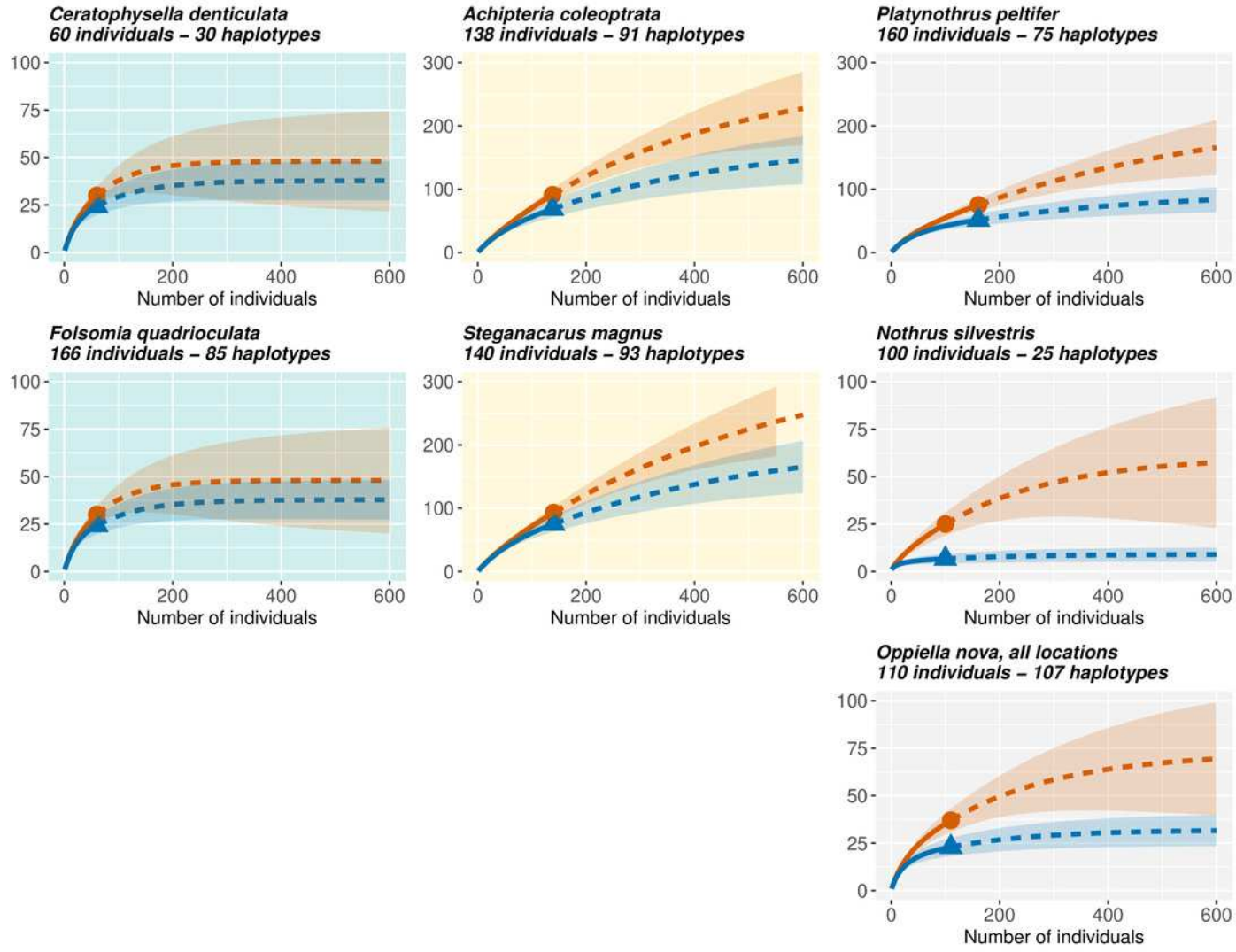
Ceratophysella denticulata
60 individuals – 30 haplotypes
Achipteria coleoptrata
138 individuals – 91 haplotypes
Platynothrus peltifer
160 individuals – 75 haplotypes
Folsomia quadrioculata
166 individuals – 85 haplotypes
Steganacarus magnus
140 individuals – 93 haplotypes
Nothrus silvestris
100 individuals – 25 haplotypes
Oppiella nova, all locations
110 individuals – 107 haplotypes
Number of individuals

Table 1(on next page)

Summary of oribatid mites and Collembola used in this study for identifying a barcoding gap in soil-living invertebrates.

Bold taxa have the broadest and densest geographic sampling within the investigated genus and sampling range is comparable among all genera, except for Oppiidae, which covered a smaller sampling area. Accession numbers of specimens are provided in the alignments in the supplementary material. The column ASAP refers to the number of genetic lineages (subsets) for each species detected by the ASAP analysis (see Table 4). One or more individuals of species marked with asterisk (*) were assigned to the same genetic lineage (ASAP subset).

Table 1. Summary of oribatid mites and Collembola used in this study for identifying a barcoding gap in soil-living invertebrates. Bold taxa have the broadest and densest geographic sampling within the investigated genus and sampling range is comparable among all genera, except for Oppiidae, which covered a smaller sampling area. Accession numbers of specimens are provided in the alignments in the supplementary material. The column ASAP refers to the number of genetic lineages (subsets) for each species detected by the ASAP analysis (see Table 4). One or more individuals of species marked with asterisk (*) were assigned to the same genetic lineage (ASAP subset).

taxon	congeneric	no. inds.	ASAP
all Oribatida		853	
Achipteria	*A. catskillensis*	11	1
	A. coleoptrata	**138**	12
	A. howardi	4	1
	total	153	14
Nothrus	*N. anauniensis*	20	1
	N. borussicus	8	1
	N. palustris	10	2
	N. pratensis	5	2
	N. silvestris	**100**	1
	total	143	7
Platynothrus	*P. capillatus*	4	1
	P. peltifer	**160**	3
	P. thori	4	1
	P. yamasakii	81	1
	total	249	6
Oppiidae	*Aeroppia* sp.	6	1
	Berniniella hauseri	2	1
	Dissorhina ornata	11	2
	Multioppia sp	6	1
	Oppia nitens	92	*3
	Oppia sp.	3	*2
	Oppiella nova	**110**	9
	Oppiella subpectinata	1	1
	Oppiella uliginosa	3	1
	Ramusella insculpta	3	1
	total	237	24
Steganacarus	*S. applicatus*	14	*2
	S. carinatus	8	*2
	S. crassisetosus	6	1
	S. magnus	**140**	18

taxon	congeneric	no. inds.	ASAP
all Collembola		612	
Ceratophysella	*C. bengtssonii*	20	1
	C. communis	31	2
	C. comosa	7	1
	C. denticulata	**60**	7
	C. granulata	7	2
	C. liguldorsi	12	2
	C. longispina	59	1
	C. pseudarmata	44	2
	C. scotica	4	1
	C. skarzynskii	17	1
	C. succinea	5	1
	total	266	21
Folsomia	*F. bisetosa*	15	*3
	F. candida	47	3
	F. ciliata	6	1
	F. fimentaria (incl. L1-L3)	28	5
	F. nivalis	39	1
	F. octoculata	7	*3
	F. peniculata	15	2
	F. quadrioculata	**166**	24
	F. sexoculata	23	2
	total	346	43

	S. similis	5	1
	S. spinosus	15	1
	total	188	25

1

Table 2(on next page)

Summary statistics of datasets.

(A) Information on number of genera and species per taxon, the minimum, maximum, median and mean number of sequences used. Oppiidae were analysed on a higher taxonomic level, i.e. at genus instead of species level. (B) Sequence information of datasets, giving the number of sequences, the length of the alignment, the minimum, maximum, median mean and median length of sequences and the number of sequences that were below 500 bp per taxon.

Table 2. Summary statistics of datasets. (A) Information on number of genera and species per taxon, the minimum, maximum, median and mean number of sequences used. Oppiidae were analysed on a higher taxonomic level, i.e. at genus instead of species level. (B) Sequence information of datasets, giving the number of sequences, the length of the alignment, the minimum, maximum, median mean and median length of sequences and the number of sequences that were below 500 bp per taxon.

(A)	**taxa information**					
	genera	**species**	**min**	**max**	**median**	**mean**
all Oribatida	10	28	1	160	8	35
Achipteria	1	3	4	138	11	51
Nothrus	1	5	5	100	10	29
Platynothrus	1	4	4	160	42	62
Steganacarus	1	6	5	140	11	31
all Collembola	2	20	4	166	18	31
Ceratophysella	1	11	4	60	17	24
Folsomia	1	9	6	166	23	38
	families	**genera**				
Oppiidae	1	7	2	114	6	34

(B)	**alignment information (bp)**						
	no. sequences	**length**	**min**	**max**	**mean**	**median**	**no. sequences <500 bp**
all Oribatida	970	657	371	657	565	558	32
Achipteria	153	507	507	507	507	507	0
Nothrus	143	580	371	580	572	580	2
Platynothrus	249	558	417	558	544	558	8
Steganacarus	188	591	476	591	551	526	21
all Collembola	612	583	310	583	564	583	48
Ceratophysella	266	651	485	651	642	651	1
Folsomia	346	583	310	583	552	583	47
Oppiidae	237	657	459	657	632	657	1

1
2

Table 3(on next page)

Range of barcoding gap thresholds and cumulative errors for all datasets.

The cumulative error is the sum of false positives and false negatives. Except for *Nothrus* and *Ceratophysella* the barcoding gap is not present in the investigated species, or exceeds the standard threshold of 2 % - 3 %.

1

Table 3. Range of barcoding gap thresholds and cumulative errors for all datasets. The cumulative error is the sum of false positives and false negatives. Except for *Nothrus* and *Ceratophysella* the barcoding gap is not present in the investigated species, or exceeds the standard threshold of 2%-3%**.**

	optimal barcoding gap threshold	
	cumulative error = 0	**smallest cumulative error**
all Oribatida	-	4% (error: 26)
Achipteria	15%	9%-14 % (error: 1)
Steganacarus	-	8%-15% (error: 8)
Nothrus	2%-17%	1%, 18% (error: 2)
Platynothrus	> 4%	2%-3% (error: 1)
Oppiidae	-	8%-15 % (error: 3)
all Collembola	-	6%-11% (error: 9)
Ceratophysella	3%-11%	2% (error: 3)
Folsomia	-	6%-14% (error: 9)

2

3

Table 4(on next page)

Summary of the estimated number of genetic lineages for each local dataset.

The number of morphologically determined species (No. of species) is given for each dataset together with the number of genetic lineages (No. of subsets) estimated by ASAP. For each dataset, the partition with the lowest number of subsets and the highest ranks was selected. The respective scores (including ranks) and statistic support are provided, along with the estimated genetic distance threshold (Threshold distance) that separates the individual subsets. For a detailed list of subsets per species refer to Table 1.

Table 4. Summary of the estimated number of genetic lineages for each local dataset. The number of morphologically determined species (No. of species) is given for each dataset together with the number of genetic lineages (No. of subsets) estimated by ASAP. For each dataset, the partition with the lowest number of subsets and the highest ranks was selected. The respective scores (including ranks) and statistic support are provided, along with the estimated genetic distance threshold (Threshold distance) that separates the individual subsets. For a detailed list of subsets per species refer to Table 1.

	No. of species	No. of subsets	ASAP-score	P-value (rank)	W (rank)	Threshold distance [%]
all Oribatida	28	69	25.0	1.20e-04 (4)	8.12e-06 (46)	15.0
Achipteria	3	14	5.5	2.99e-02 (3)	7.13e-04 (8)	6.9
Nothrus	5	7	2.0	1.00e-05 (2)	5.53e-04	12.2
Platynothrus	4	6	4.5	1.00e-05 (1)	3.39e-05 (3)	15.0
Steganacarus	6	25	3.5	1.60e-04 (1)	1.79e-04 (6)	15.0
Oppiidae	10	24	3.00	3.00e-04 (1)	3.17e-04 (5)	15.8
all Collembola	20	64	12.5	1.00e-05 (2)	6.28e-05 (23)	11.3
Ceratophysella	11	21	5.5	2.02e-03 (4)	3.93e-04 (7)	13.8
Folsomia	9	43	4.5	1.00e-05 (1)	1.70e-04 (8)	8.02