Unearthing soil arthropod diversity through DNA metabarcoding

Monica R. Young and Paul D. N. Hebert

Centre for Biodiversity Genomics, University of Guelph, Guelph, Ontario, Canada
Department of Integrative Biology, University of Guelph, Guelph, Ontario, Canada

ABSTRACT

DNA metabarcoding has the potential to greatly advance understanding of soil biodiversity, but this approach has seen limited application for the most abundant and species-rich group of soil fauna–the arthropods. This study begins to address this gap by comparing information on species composition recovered from metabarcoding two types of bulk samples (specimens, soil) from a temperate zone site and from bulk soil samples collected at eight sites in the Arctic. Analysis of 22 samples (3 specimen, 19 soil) revealed 410 arthropod OTUs belonging to 112 families, 25 orders, and nine classes. Studies at the temperate zone site revealed little overlap in species composition between soil and specimen samples, but more overlap at higher taxonomic levels (families, orders) and congruent patterns of α- and β-diversity. Expansion of soil analyses to the Arctic revealed locally rich, highly dissimilar, and spatially structured assemblages compatible with dispersal limited and environmentally driven assembly. The current study demonstrates that DNA metabarcoding of bulk soil enables rapid, large-scale assessments of soil arthropod diversity. However, deep sequence coverage is required to adequately capture the species present in these samples, and expansion of the DNA barcode reference library is necessary to improve taxonomic resolution of the sequences recovered through this approach.

Subjects Biodiversity, Bioinformatics, Ecology, Genomics, Molecular Biology
Keywords Metabarcoding, Arthropods, Soil, Biodiversity, NGS, Environmental DNA

INTRODUCTION

The "enigma of soil animal species diversity" (*Anderson, 1975*) will only be resolved through taxonomically comprehensive, large-scale surveys (*Cameron et al., 2018*; *Decaëns, 2008*). Although arthropods likely comprise the largest fraction of soil fauna diversity (*Geisen et al., 2019*), their taxonomy is so poorly understood that estimates of their species richness span two orders of magnitude (*André, Noti & Lebrun, 1994*; *Erwin, 1982*; *Stork, 1988*). Detailed assessments of soil biodiversity have been constrained by several factors, but inefficient methods for isolating arthropods from the soil matrix is an initial barrier (*André, Ducarme & Lebrun, 2002*). Once specimens are isolated, their high abundance combined with cryptic morphology, lack of taxonomic expertise, and the high incidence of undescribed species pose serious challenges to their identification (*André et al., 2001*; *Bickford et al., 2007*). Consequently, past studies of the soil fauna have typically examined a few taxonomic groups on a local scale or have employed coarse taxonomic resolution at a larger scale (*Donoso, Johnston & Kaspari, 2010*; *Marra & Edmonds, 2005*). However, these barriers

can be resolved by DNA-based approaches because most species can be distinguished by examining sequence variation in the 648 bp barcode region of the mitochondrial cytochrome *c* oxidase I (COI) gene (*Hebert et al., 2003*; *Hendrich et al., 2015*; *Zahiri et al., 2017*). Moreover, specimens can be assigned to operational taxonomic units (OTUs), an effective proxy for species, enabling large-scale, comprehensive assessments of diverse arthropod assemblages (*Hebert et al., 2016*).

Although DNA barcodes reliably discriminate species of soil organisms (*Hogg & Hebert, 2004*; *Raupach et al., 2016*; *Schäffer, Kerschbaumer & Koblmüller, 2019*), this approach is time consuming and expensive as it involves processing individual specimens (*Borisenko, Sones & Hebert, 2009*). Metabarcoding, the coupling of DNA barcoding with high-throughput sequencing, enables time- and cost-effective assessments of diversity by allowing the analysis of bulk samples of specimens or soil (*Taberlet et al., 2012*). Until recently, metabarcoding has seen limited application for the characterization of soil arthropod biodiversity (*Arribas et al., 2016*; *Dopheide et al., 2019a*; *Oliverio et al., 2018*). The results from such assessments are likely to differ from those based on morphological study because reference libraries are incomplete and PCR amplification introduces bias (*Schenekar et al., 2020*). Because methods used to extract arthropods from soil are inefficient, direct analysis of DNA extracted from soil may permit more comprehensive species recovery (*Deiner et al., 2017*), but 'relic' DNA in soils may overestimate biodiversity and dampen ecological signals by conflating contemporary and historic species assemblages (*Nagler et al., 2018*). Despite these complexities, a comparison of metabarcode data from soil and soil-isolated specimens revealed similar patterns of arthropod α- and β-diversity although the two sampling methods recovered complementary fractions of the biota (*Oliverio et al., 2018*). However, no other studies have previously evaluated congruence in β-diversity patterns of soil arthropod communities using these methods.

This study begins by comparing patterns of α- and β-diversity at a nature reserve in southwestern Ontario revealed by DNA metabarcoding paired samples, one based on DNA extracted from the soil and a second based on DNA from specimens isolated from the soil. The recovery of OTUs, families, and orders was compared between the two methods. As well, results were validated against family- and order-level identifications based on morphological examination of specimens recovered after DNA extraction. The study then considers diversity patterns at a larger geographic scale by extending analysis to include soil samples from eight sites in Arctic Canada. While additional invertebrate and microbial taxa were recovered at all locales, analyses focused on arthropods, especially mites.

METHODS

Sample collection

Soil samples were collected in July 2017 from two forested sites (OF1, OF2) and one meadow site (OM1) in *rare*, a research reserve in the Eastern Temperate Forests ecoregion (*EPA, 2015*) in Ontario, Canada (Fig. 1; Table S1). Sampling at each of these sites involved collecting approximately 100 g of soil and the surface leaf litter at 3 m intervals along a 30 m transect. The 10 subsamples from each site were homogenized and pooled using a
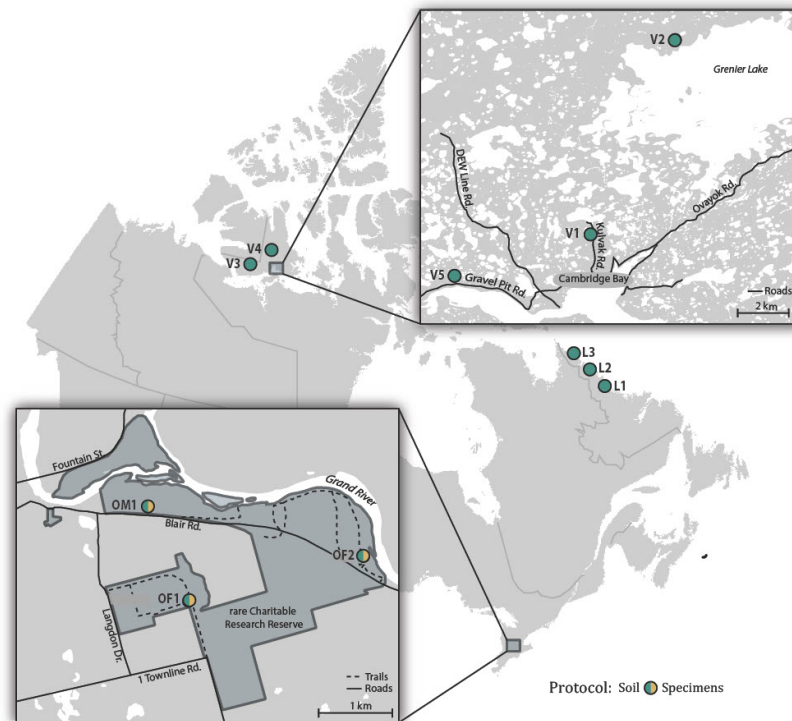
Figure 1 Map showing the 11 Canadian sites where soil samples were collected and the three sites where specimen samples were also obtained.

litter sifter. The sifter and tools used to collect each sample were sterilized between sites using H_2O to remove residual debris followed by soaking in 1:10 diluted bleach for 15 min, a final rinse with H_2O, and were then air dried. Approximately a tenth (100 g) of the total soil sample from each site was retained for bulk soil analysis, while the remaining $\sim$900 g of soil was placed in a modified Berlese-Tullgren funnel for one week with the isolated invertebrates preserved in 95% EtOH. We subsequently refer to these as ''soil'' and ''specimen'' samples. All samples were stored at $-20\,°C$ prior to DNA extraction.

To obtain comparative data from distant locales, samples were collected in July 2017 at three sites (L1-L3) spanning the transition from the Taiga to Arctic Cordillera ecoregions in coastal Labrador, and in July–August 2018 at five sites (V1-V5) in the Tundra ecoregion on southern Victoria Island near Ikaluktutiak (Cambridge Bay), Nunavut (Fig. 1; Table S1). At each site, soil samples were collected along two parallel 30 m transects (approximately 10 m apart) producing 16 samples that were processed using the same protocol as for the *rare* soil samples.

Samples were collected with permission at rare Charitable Research Reserve. Samples collected in the vicinity of Cambridge Bay fall under a Department of Environment Wildlife Research Permit (WL 2018-044). A permit was also obtained through the Canada C3 Expedition for sampling in National Parks in Labrador (ANP-2017-25145), while

collecting in other sites was approved by the Nunatsiavut Government Research Advisory Committee.

Positive control

A DNA extract prepared from a bulk arthropod sample was used as a positive control (CCDB positive control #07682). To gain a detailed understanding of its Barcode Index Number (BIN) diversity (*Ratnasingham & Hebert, 2013*), each of the component specimens (2,018 insects, 17 arachnids) were individually DNA barcoded. These specimens derived from an 8-day Malaise trap deployment at the Arkell Research Farm near Guelph, Ontario. DNA was extracted from a single leg of each specimen and the 658 bp barcode region was recovered using standard protocols (*Ivanova, de Waard & Hebert, 2006*; *Hebert et al., 2013*) utilizing the C_LepFolF/C_LepFolR primer pair (*Hernández-Triana et al., 2014*). When these primers failed to generate a sequence, the DNA extract was amplified using the MLepF1/C_LepFolR and C_LepR1/MLepR2 primer pairs (*Hebert et al., 2013*) followed by NGS failure tracking (*Prosser et al., 2016*) and SMRT sequencing (*Hebert et al., 2018*) when necessary. All sequences were uploaded to the public AGAKS project on BOLD, the Barcode of Life Dataystem (*Ratnasingham & Hebert, 2007*). Those meeting quality standards ($\geq$ 500 bp, <1% ambiguous bases, free of contamination and stop codons) were assigned a BIN. In total, 432 BINs from 11 orders and 90 families were represented among the 1,917 sequences that gained an assignment. The other 112 specimens failed to generate a quality sequence and consequently lacked a BIN assignment. Some of these specimens (1 Coleoptera, 36 Diptera, 5 Hemiptera, 75 Hymenoptera, 1 Lepidoptera) may represent BINs that were not characterized.

The bulk sample was then reconstituted by pooling all specimens and a bulk DNA extract was made following standard protocols (*Braukmann et al., 2019*) for use as a positive control in subsequent metabarcoding analyses.

Molecular analysis of specimen and soil samples

DNA was extracted from the bulk specimen samples using a non-destructive protocol so that the exoskeletons of the specimens were available for morphological study after lysis. Extraction began with the filtration of each sample through a sterile Microfunnel (0.2 μm Supor Membrane Filter, Pall Laboratory) followed by the addition of 10 mL of insect lysis buffer (*Ivanova, de Waard & Hebert, 2006*). The lysate was mixed gently and incubated overnight at 56 °C, and eight 50 μl aliquots were then transferred from each lysate into separate wells in a 96-well microplate to create technical replicates. Specimen exoskeletons were recovered from the remaining ~9.5 mL of lysate through filtration and stored at room temperature in 95% EtOH. 100 μl of binding mix was added to each well of the microplate and the resultant 150 μl solution was transferred to a 3.0 μm glass fibre (GF) plate (Pall Laboratory) followed by centrifugation at 5,000 g for 5 min. 180 μl of protein wash buffer was then added to each well in the plate and centrifuged at 5,000 g for 5 min. This step was followed by two additional rounds of wash buffer (600 μl each) and centrifugation for 5 min at 5,000 g. The GF plates were incubated for 30 min at 56 °C to evaporate residual EtOH before DNA was eluted into clean 96-well Eppendorf plates by adding 50 μl of 10 mM Tris–HCl (pH 8.0) and centrifugation at 5,000 g for 5 min.

DNA was extracted from the bulk soil samples using a slightly modified version of a standard protocol (*Ivanova, Fazekas & Hebert, 2008*). The ∼100 g aliquot of soil from each sample was mixed with 200 mL of insect lysis buffer which included 1% polyvinylpyrrolidone (PVP). The resulting lysates were mixed gently on a shaker and incubated overnight at 56 °C. Technical replicates were constructed by transferring eight 50 µl subsamples per lysate into a 96-well microplate. Negative extraction controls were also created by adding four 50 µl aliquots of ddH$_2$O to each 96-well plate. Each lysate (and control) was mixed with 100 µl of plant binding buffer, transferred to a 3.0 µm GF plate, and centrifuged at 5,000 g for 5 min. Each well in the GF plate was then washed with 150 µl of the plant binding buffer and centrifuged for 2 min at 5,000 g, followed by 150 µl of binding mix centrifuged at 5,000 g for 2 min, and two rounds of wash buffer (750 µl each) centrifuged for 5 min at 5,000 g. The GF plates were incubated for 30 min at 56 °C before DNA was eluted into clean 96-well Eppendorf plates with 50 µl of 10 mM Tris–HCl pH 8.0 centrifuged at 5,000 g for 5 min.

In total, three 96-well DNA plates were used to accommodate the 176 technical replicates (22 samples × 8 replicates each; Table S2). Each plate included up to 80 technical replicates (10 samples × 8 replicates each) as four wells were treated as negative extraction controls and 12 wells were left empty for PCR controls (8 positive, 4 negative) that were added later. Samples in each plate were grouped by ecoregion. Although DNA from the paired specimen/soil samples was extracted separately (as they required different protocols), they were transferred into the same plate following extraction. The first plate included all 48 replicates from the six Eastern Temperate Forests samples; the second included the six Taiga/Arctic Cordillera samples (48 replicates); and the third included all ten Tundra samples (80 replicates). Although the number of replicates differed by plate, there was no significant difference in their number of post-filtered reads (ANOVA $F_{2,173} = 2.2$, $p = 0.2$).

DNA extracts were stored at 4 °C for ≤4 days prior to PCR. Immediately before analysis, 50 µl of template DNA from the positive control sample was added to each of the eight wells in Row 11 while wells in Row 12 were employed as negative controls. Amplicon libraries were then constructed using a two-stage PCR protocol (*Braukmann et al., 2019*) that employed the AncientLepF3 (*Prosser et al., 2016*) and C_LepFolR primers to amplify a 463 bp fragment of COI and attach unique molecular identifiers (UMIs) with Ion Torrent S5 sequencing adaptors. The labelled products from each plate were pooled and normalized to 1 ng/µl prior to sequencing. Sequencing libraries were prepared on the Ion Chef^{TM} (Thermo Fisher Scientific) following manufacturer's instructions and were then analyzed on an Ion Torrent S5 using a 530 chip.

All molecular analysis was performed at the Canadian Centre for DNA Barcoding (CCDB; http://www.ccdb.ca). The DNA extracts were archived at −80 °C at the CCDB while the vouchers were deposited in the specimen archive at the Centre for Biodiversity Genomics.

Sequence analysis

The raw sequence data was analyzed with the 'JAMP' (*Elbrecht, 2018*) package in R which utilizes functions from USEARCH (*Edgar, 2013*), VSEARCH (*Rognes et al., 2016*), and

Cutadapt (*Martin, 2011*). The forward primer was first trimmed from each read and the resultant sequence was truncated to 463 bp. Reads shorter than 350 bp in length or with more than 0.5% expected errors (Emax = 2.3) were discarded. E is the sum of error probabilities based upon the quality (phred) score for each nucleotide position (*Edgar & Flyvbjerg, 2015*). Adopting an E max of 2.3 (= error rate of 0.5%) enabled a balance between read retention and error filtering since most reads (98%) were excluded by a strict error threshold (Emax = 1) while a higher threshold (Emax = 4.6; error rate = 1%) did not improve the recovery of known BINs in the positive control despite the recovery of more OTUs. The remaining reads were dereplicated with simultaneous chimera filtering and those with more than one replicate were clustered into OTUs at 3% nucleotide divergence using UPARSE (*Edgar, 2013*) as implemented by USEARCH in JAMP. Singleton reads were assigned to existing OTUs if their divergence was <3% to a cluster; otherwise they were discarded. OTUs with low abundance in a replicate (<0.05%) were also excluded (Table S3). The maximum read count for any OTU in the negative controls was subtracted from all other occurrences of that OTU to minimize the potential effects of tag switching and background contamination. Consequently, nine OTU occurrences were reduced to zero and one OTU was filtered entirely (Table S4). Although amino acid translation-based filtering to remove reads with stop codons and frameshift indels (1–2 bp insertions/deletions) can reduce the prevalence of spurious OTUs, this procedure is rarely adopted since it is not easily replicable using standard bioinformatics tools (*Creedy et al., 2021*). Its use is further complicated when samples contain a broad range of taxonomic groups as their analysis requires the use of several translation tables. Consequently, OTUs with stop codons and frameshift indels were not filtered from our dataset.

OTUs were assigned the taxonomy of their top hit in the reference library [June 2021] on BOLD (*Ratnasingham & Hebert, 2007*) using BOLDigger (*Buchner & Leese, 2020*) in Python v3.9 (Python Software Foundation). JAMP's default parameters for taxon assignment at the species (≤2% divergence), genus (≤5%), family (≤10%), and order (≤15%) levels were adopted for all OTUs except those with top hits to Acari (mite) orders where lineage-specific thresholds for family and order identification were employed (P₉₅; *Young, de Waard & Hebert, 2021*). To assess the extent of species-level library coverage, OTUs ≤2% divergent from a reference sequence in BOLD were categorized as 'known' while OTUs with higher divergence were categorized as 'novel'.

OTUs in the positive control were compared to the BINs detected through detailed DNA barcode analysis of its component specimens to evaluate recovery of the positive control. In cases where multiple OTUs matched to the same BIN, their consensus sequences were examined for stop codons and frameshifts. Such features may reflect amplification or sequencing errors but could also indicate the recovery of nuclear mitochondrial pseudogenes (NUMTs). OTUs that did not closely match a BIN known to be present in the positive control were also scanned for possible signs of NUMT recovery (*e.g.*, stop codons and frameshifts), and their read counts were compared with the experimental samples to assess possible cases of tag-switching.

Specimen vs soil analysis

Taxon recovery, composition, and diversity patterns were compared for the paired specimen and soil samples from Ontario. First, OTU accumulation curves (with 1,000 replicates) were generated using the BiodiversityR package (*Kindt & Coe, 2005*) in R version 4.0.2 (*R Core Team, 2020*) for all taxa combined and for arthropods separately to assess similarity among the technical replicates for each sample. The technical replicates for a sample were then pooled, and variation in taxon richness was evaluated at three levels (order, family, OTU) for all taxa combined, and for all arthropods and mites separately. One-tailed paired t-tests were used to determine if richness was lower for specimen than for soil samples. In cases where the assumption of normality was not met, one-sided Wilcoxon Sign Rank tests (SRT) were used. Since significant differences were found for all taxa combined but not for arthropods or mites, variation was also assessed for non-arthropod invertebrates and protists as potential explanatory factors. Bacteria, fungi, and chordates were excluded from these analyses since they comprised a very small fraction of the total richness.

Variation in the recovery of arthropod orders and families was compared with the morphologically identified specimens from the Berlese-Tullgren samples (Table S5). Patterns of arthropod and mite β-diversity were evaluated through the incidence-based Sorensen dissimilarity metric estimated for each pair of samples using the betapart (*Baselga & Orme, 2012*) package in R. Dissimilarity values range from 0 to 1, with 0 indicating complete OTU overlap, and 1 indicating no shared OTUs. Compositional differences between the two methods were visualized from Sorensen distances using complete-linkage hierarchical clustering in R, and tested for significance through permutational multivariate analysis of variance (perMANOVA; 1,000 randomizations) with the vegan package (*Oksanen et al., 2015*) in R.

Diversity patterns of soil arthropod assemblages

The soil samples from the eight Arctic sites were analyzed in a similar fashion to those from *rare*. Accumulation curves were first used to assess similarity among the technical replicates for each sample and replicates were then pooled by sample. Patterns of arthropod and mite β-diversity were then evaluated with Sorensen's dissimilarity metric. Mantel test correlations (1,000 randomizations) of Sorensen dissimilarity and geographic distance matrices were used to estimate the proportion of variation in β-diversity explained by spatial structure through the ade4 package in R (*Dray & Dufour, 2007*). Compositional differences between samples, sites, and ecoregions were visualized through complete-linkage hierarchical clustering of Sorensen dissimilarities, and tested for significance with spatial scale (Site, Ecoregion) as explanatory factors. The contributions to overall diversity made at each spatial scale were evaluated by hierarchically partitioning the total arthropod and mite OTU richness (γ-diversity) into its sample, site, and ecoregion components using PARTITION (*Veech & Crist, 2009*). Both additive ($\gamma = \alpha\text{Sample} + \beta\text{Sample} + \beta\text{Site} + \beta\text{Ecoregion}$) and multiplicative ($\gamma = \alpha\text{Sample} \times \beta\text{Sample} \times \beta\text{Site} \times \beta\text{Ecoregion}$) partitions were calculated to assess net (*i.e.*, $\beta\text{Site} = \alpha\text{Ecoregion} - \alpha\text{Site}$) and proportionate (*i.e.*, $\beta\text{Site} = \alpha\text{Ecoregion}/\alpha\text{Site}$) increases in OTU richness at successive spatial scales. To test whether the partitions differed significantly from chance they were compared to null

models using PARTITION's restricted individual-based randomization following *Crist et al. (2003)*.

RESULTS

Controls

In total, 142 reads representing 109 OTUs were recovered from the 24 negative controls. The maximum read abundance ($\leq$6) for each of these OTUs was subtracted from its occurrence in each replicate during the final filtering step, causing one OTU to be eliminated. All other OTUs had higher read counts in the positive control or specimen/soil samples (29–300K reads) compared to the negative controls (1–6 reads), and most (96.2%) had >100 reads in another sample indicating that their presence in the negative controls likely reflected tag-switching.

After filtering, 0.5M reads representing 161 OTUs were recovered from the 24 positive controls (Table 1) with an average of 141 OTUs (SD = 9.5) in the 8 replicates included in each run. Five of these OTUs (4 arthropods, 1 possible heterokont) reflected tag switching as they had low read counts in the positive controls (7–71 reads) versus >100K reads in the specimen/soil samples. Three quarters of the other OTUs (115/156) were close matches ($\leq$2% divergence) to a BIN known to be present in the positive control, meaning that 27% of the 432 BINs in the positive control were recovered across the three runs. Within single runs, the recovery of BINs in the positive control averaged 24% (= $\bar{x}$ = 104 OTUs, SD = 7.6). Some of the other 41 OTUs were clearly NUMTs derived from a species in the control sample. For example, one OTU was a perfect match to a dipteran BIN (BOLD:AAU6656; *Neolimnophila pacida*) in the positive control, while a second OTU with 6.8% divergence possessed a stop codon and 5 bp deletion indicating that it was a NUMT. Six BINs abundant in the positive control were linked to several closely matching ($\geq$95% similarity) OTUs that were clearly NUMTs. For example one OTU lacking stop codons and frameshifts matched 100% to a fly BIN (BOLD:AAH3914; *Sciara humeralis*) that was represented by 75 specimens in the positive control. Two more OTUs matched this BIN but both possessed a frameshift indel; one was a perfect match bar the indel while the other was 2.2% divergent. Three other OTUs showed a close match ($\geq$98.9% similarity) to the sole orthopteran BIN (BOLD:AAA4555; *Melanoplus*; 5 specimens) that represented the largest biomass of any species in the sample, but just one possessed a frameshift indel. A similar scenario was observed for six OTUs matching ($\geq$92% similarity) an abundant fly BIN (BOLD:AAG6745; *Pollenia*; 32 specimens) where just one of the six OTUs possessed a frameshift indel. Such cases likely indicate the detection of young NUMTs, but might also reflect BINs present in the positive control that failed to sequence in the initial characterization of the sample. For example, 14 OTUs lacking stop codons and frameshift indels were unique to the positive controls and consistently recovered in each replicate. They represented a bacterium (*Wolbachia pipientis*) and 13 arthropods that might derive from gut contents or specimens in the positive control that lacked a BIN assignment.

Table 1 **Summary of post-filtered reads and OTUs generated from metabarcoding 19 soil samples and three soil invertebrate samples from 11 Canadian sites.** Data is summarized separately for the positive and negative controls, all experimental samples combined, the subset of paired specimen and soil samples, and the soil samples alone.

Source	Samples/ Replicates	Total Reads	Total OTUs	Mean Reads/ Replicate (SD)	Mean OTUs/ Replicate (SD)	Mean Reads/ Sample (SD)	Mean OTUs/ Sample	Unidentified reads	Unidentified OTUs	Identified Reads	Identified OTUs
Negative controls	- /24	142	109	6.5 (8.8)	5.9 (8.0)	–	–	75	65	67	44
Positive controls	- /24	544,897	161	22,704 (14,836)	141 (4.3)	–	–	7	1	544,890	160
All samples	22/176	3,726,640	7,487	21,295 (17,879)	106 (67)	169,393 (126,444)	514 (297)	2,214,161	6,336	1,512,479	1,151
Specimens + Soil	6/48	1,214,492	2,126	25,840 (20,422)	109 (97)	202,415 (173,734)	439 (456)	286,257	1,776	928,235	350
Specimens	3/24	812,090	120	33,837 (26,086)	28 (10)	270,697 (247,020)	47 (22)	798	37	811,292	83
Soil	3/24	402,402	2,052	17,496 (4,246)	193 (71)	134,134 (21,271)	831 (229)	285,459	1,764	116,943	288
All Soil	19/152	2,914,550	7,414	19,302 (15,404)	119 (64)	153,397 (99,461)	587 (246)	2,213,363	6,324	701,187	1,090

Sequences from specimen and soil samples

A total of 27.75 million reads were recovered from the 22 experimental samples. Approximately 13% (3,726,640) of these reads were retained post-filtering with an average of 21.4K reads (SD = 14.8K) per replicate and 169.4K reads (SD = 126K) per sample (Table 1). These reads were assigned to 7,487 OTUs with an average of 106 OTUs (SD = 67) per replicate and 514 OTUs (SD = 297) per sample. Many OTUs (84.6%) recovered from the 22 samples showed low similarity to any sequence in the BOLD reference library (<64.3% to a Trombidiformes, <74.9% to a Sarcoptiformes, <75.3% to a Mesostigmata, <85.0% to all other taxa). As a result, 6,366 of the 7,487 OTUs failed to gain a taxonomic assignment (Table 1). Examination of the top hits for these unidentified OTUs revealed that nearly half (38.6%) were protists with heterokonts comprising more than half their total (20.1%). Many other unassigned OTUs were likely arthropods (19.0%), other invertebrates (22.3%), or bacteria (14.5%). However, 1,151 OTUs did gain a phylum or better assignment. They included representatives of 20 phyla, 30 classes, and 60 orders (Fig. 2). More than a third of the identified OTUs (35.6%) were arthropods; these 410 OTUs included representatives of 112 families, 25 orders, and nine classes. However, other invertebrate lineages were also well represented (39.1% of identified OTUs), while protists (22.8%), fungi (1.5%), and bacteria (1.0%) were also recovered. Another OTU showed highest similarity (87.5%) to a fish sequence but BLAST analysis indicated that it was a human COI NUMT. Two other identified OTUs (Cnidaria, Porifera) appeared to be spurious since they were detected in soil >2 km from water and their top matches on BOLD and GenBank included protists and fungi in addition to cnidarians and poriferans. Stop codons and frameshift indels were noted in about 20% of the identified OTUs, but most involved a single nucleotide indel in a poly-T tract or near the 3′-end, potentially reflecting NUMT recovery or PCR/sequencing errors.

Just 28.6% of the identified OTUs were represented in the reference library on BOLD, and only 15.1% were assigned to a species (Fig. 3, Table 2). Because taxonomic coverage and parameterization of the barcode reference library varied among groups, identification success differed among taxa. For example, 67.1% of the arthropod OTUs matched (≥98% similarity) a reference sequence on BOLD and nearly a third (32.9%) were assigned a species. By contrast, just 8.9% of the OTUs in other invertebrate phyla matched a reference sequence and only 6.7% were identified to a species. Among arthropods, OTU matches on BOLD were considerably higher for insects (87.7%) and for other arthropods (60.1%) than for arachnids (54.3%). The low success for arachnids reflected the fact that just 47.9% of acarine OTUs found a match versus 83.3% for spiders (Fig. 3). While OTUs from some other invertebrate phyla (e.g., Rotifera) also had very poor coverage in BOLD, all eight gastropod (Mollusca) OTUs and 56.2% of the Clitellata (Annelida) OTUs matched a reference sequence. By contrast, just 1 of 356 rotifer (Bdelloidea) OTUs matched a reference sequence and coverage was only slightly better for nematodes (Chromadoria, Enoplea; 8.0%). Parameterization improved at higher taxonomic levels, with 23.8% of all OTUs assigned to a genus, 44.0% to a family, 80.0% to an order, and 98.0% to a class (Table 2). Across all taxonomic levels, the proportion of OTUs gaining an identification was consistently higher for arthropods than for other phyla.

Figure 2 Taxonomic breakdown of the classified reads and OTUs generated from (A) 22 samples including three specimen and 19 soil samples, and (B) the subset of six paired specimen and soil samples.
Full-size ⊠ DOI: 10.7717/peerj.12845/fig-2

Specimens vs soil

The three pairs of samples included 2,126 OTUs and 1,214,492 reads, but the OTU count was 17x higher for soil than specimen samples (Table 1). Many of these OTUs (83.5%) could not be assigned taxonomy, even at a kingdom level. Among these unidentified OTUs, 99.3% derived from the soil versus 2.1% from the specimen samples (Table 1). Among the 350 OTUs that gained a taxonomic assignment, there were representatives of 76 families and 40 orders (Fig. 2). Nearly half (150) the identified OTUs were arthropods belonging to 62 families and 21 orders. OTU accumulation curves for the eight technical replicates
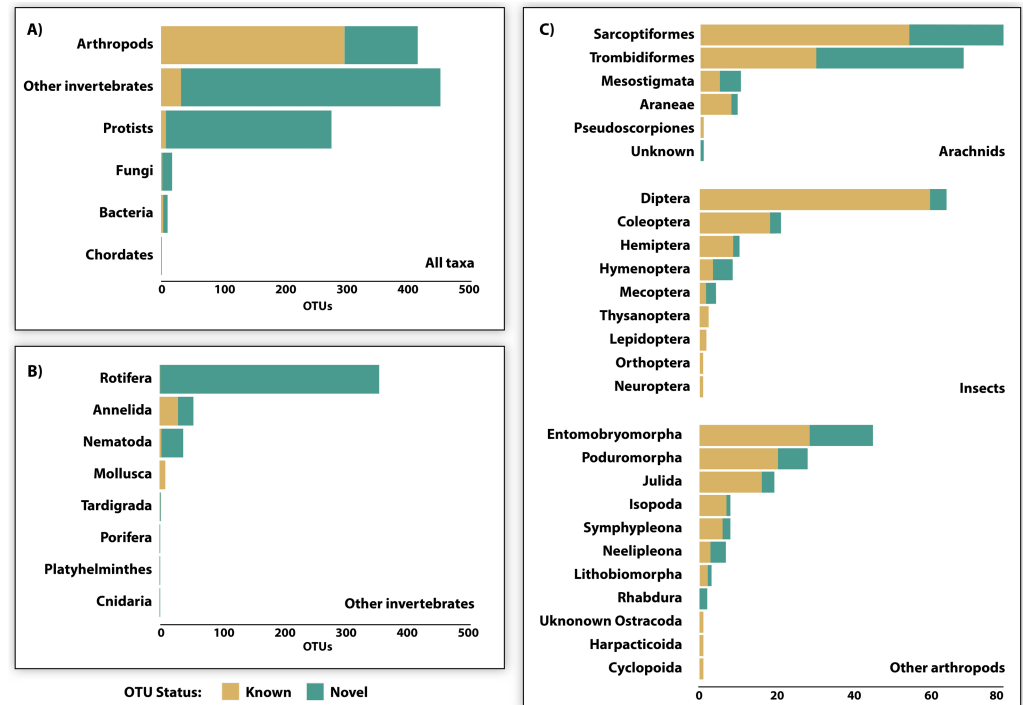
Figure 3 OTU diversity recovered by DNA metabarcoding 22 soil samples (3 specimen, 19 soil) and the proportion of known and novel OTUs based on comparison with the BOLD reference library. Comparisons are provided for (A) 6 major taxonomic groups, (B) 8 invertebrate phyla, and (C) 23 arthropod orders.

Full-size 🖾 DOI: 10.7717/peerj.12845/fig-3

Table 2 The number of OTUs (and percent of total OTUs) assigned taxonomy at five major ranks based on comparison with BOLDs reference library. Data is summarised for all taxa combined and each major group detected from metabarcoding 19 soil samples and three soil invertebrate samples from 11 Canadian sites.

Taxon	Phylum	Class	Order	Family	Genus	Species
Arthropods	410 (100%)	408 (99.5%)	407 (99.3%)	382 (93.2%)	217 (52.9%)	135 (32.9%)
Other invertebrates	450 (100%)	448 (99.6%)	364 (80.9%)	96 (21.3%)	42 (9.3%)	30 (6.7%)
Protists	262 (100%)	257 (98.1%)	136 (51.9%)	26 (9.9%)	13 (5.0%)	7 (2.7%)
Fungi	17 (100%)	6 (35.3%)	6 (35.3%)	1 (5.9%)	1 (5.9%)	1 (5.9%)
Bacteria	11 (100%)	9 (81.8%)	7 (63.6%)	1 (9.1%)	1 (9.1%)	1 (9.1%)
Chordates	1 (100%)	1 (100%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)
Total	1,151 (100%)	1,129 (98.0%)	921 (80.0%)	506 (44.0%)	274 (23.8%)	174 (15.1%)

indicated that only one sample reached an asymptote (OF1) and that the slope of the curves was consistently steeper for soil than specimen samples (Fig. 4). While a similar pattern was observed for the accumulation of arthropod OTUs, the slope of these curves was less than those for all taxa.

Few OTUs were shared between the specimen and soil samples from a particular site (Fig. S1). On average, 146 identified OTUs (SD = 12) were recovered from a site but just 7 (SD = 303) were shared between paired samples. The same trend was observed for
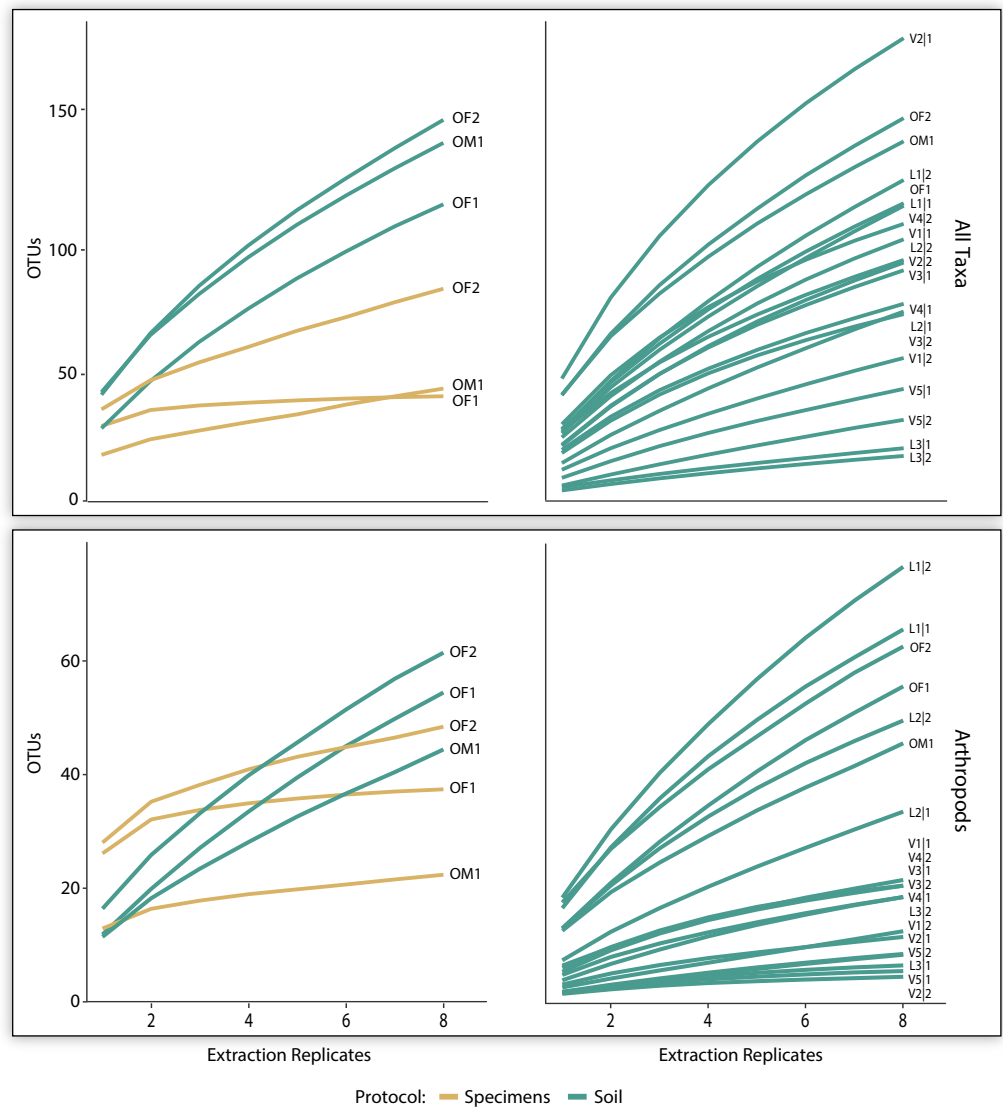
Figure 4 OTU accumulation curves (with 1,000 randomizations) across eight extraction replicates summarized for all taxa combined and all arthropods detected for the three paired specimen and soil samples, and all 19 soil samples.

Full-size ⬚ DOI: 10.7717/peerj.12845/fig-4

arthropods and mites as mean site totals were 60 (SD = 9.5) and 23 (SD = 14) OTUs respectively with an average of just 5.7 (SD = 4.6) and 2.7 (SD = 2.9) OTUs shared. Many more orders, families, and OTUs were detected in the soil than specimen samples (Table 3) reflecting their much higher recovery of protozoan and non-arthropod invertebrate OTUs. Although soil samples recovered slightly more arthropods than specimen samples at all three levels in the taxonomic hierarchy, these differences were not significant (Table 3). There was also no significant difference between mite family and OTU richness with the two sampling approaches.

Table 3 Comparison of mean taxon richness detected in six paired specimen and soil samples for three levels of taxonomic resolution (order, family, OTU). Tests of significance are provided for differences in mean richness values by paired t-tests or, in cases with non-normal residual distributions, Wilcoxon Sign Rank tests (SRT).

Taxon	Rank	Specimens		Soil		Test of significance
		Mean	(SD)	Mean	(SD)	
All taxa	Order	9.3	(4.0)	27.3	(0.6)	$2 = -8.6, p = 0.007$
	Family	15.3	(5.5)	31.7	(5.1)	$2 = -4.6, p = 0.02$
	OTU	34.3	(14.2)	118.3	(18.9)	$2 = -4.9, p = 0.02$
Arthropods	Order	7.0	(2.6)	12.3	(0.6)	SRT $p = 0.09$
	Family	14.3	(5.5)	24.3	(3.7)	$2 = -2.8, p = 0.05$
	OTU	31.3	(12.4)	36.7	(6.5)	$2 = -0.6, p = 0.3$
Mites	Order	2.3	(0.6)	2.3	(0.6)	*NA*
	Family	7.3	(3.2)	9.0	(4.6)	$2 = -1.8, p = 0.1$
	OTU	13.0	(7.0)	13.0	(9.6)	$2 = 0, p = 0.5$
Other invertebrates	Order	1.5	(0.7)	9.0	(0.0)	$2 = -3.6, p = 0.03$
	Family	1.5	(0.7)	7.3	(1.5)	SRT $p = 0.08$
	OTU	1.5	(0.7)	40.7	(9.1)	$2 = -7.3, p = 0.009$
Protists	Order	0.3	(0.6)	5.7	(0.6)	$2 = -3.6, p = 0.03$
	Family	0.3	(0.6)	2.3	(0.6)	$2 = -3.5, p = 0.04$
	OTU	0.3	(0.6)	38.0	(6.9)	$2 = -4.9, p = 0.02$

Morphological inspection of specimens recovered after DNA extraction revealed that orders and families corresponded well with those detected by DNA analysis (Fig. 5). Just two (Araneae, Hymenoptera) of 14 orders and seven (Eupththiracaridae, Formicidae, Hypochthoniidae, Lycosidae, Nanorchestidae, Oppiidae, Parasitidae) of 35 families present among the morphologically identified taxa were not recovered by metabarcoding. Conversely, just two mite families (Trombidiformes: Scutacaridae, Tydeidae) detected by metabarcoding were not identified morphologically. These families were not present, even as immatures, as all 24 juvenile mites were Mesostigmata or Sarcoptiformes. Their detection could also not be explained by tag-switching or misidentification as they were unique to the sample and their sequence was deeply embedded within their family's clade in the reference library. Compositional differences between the specimen and soil samples were also prominent as two orders and 16 families were detected by specimen but not by soil analysis, while seven orders and 32 families were recovered by soil but not by specimen analysis (Fig. 5).

Hierarchical clustering of pairwise β-diversities suggested that compositional differences in arthropod OTU assemblages were greater between the paired soil/specimen samples than between sites although this difference was not significant (perMANOVA $R^2 = 0.31, p = 0.1$, Fig. 6). A similar, non-significant pattern was also noted for mite OTUs (perMANOVA $R^2 = 0.44, p = 0.1$). Regardless of sampling protocol, compositional differences between the two forest sites (OF1, OF2) were consistently less than those between the forest sites and the meadow (OM1).
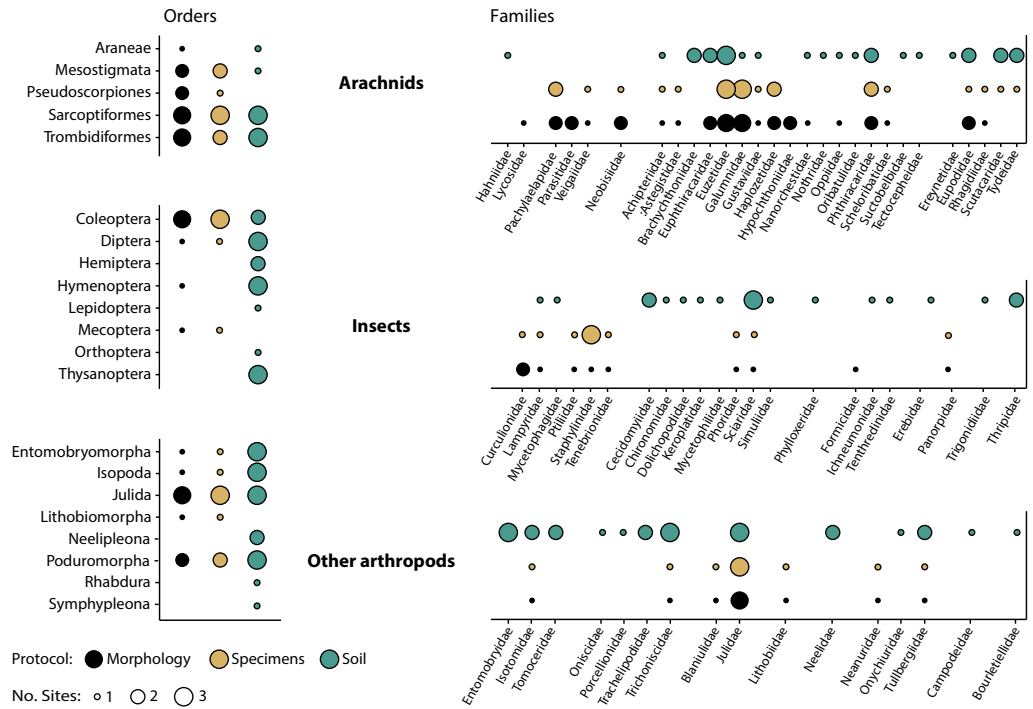
Figure 5 Taxonomic recovery of arthropod orders and families from three paired samples of specimens and soil analysed by DNA metabarcoding compared to taxa detected through morphological analysis of vouchers from the specimen samples.

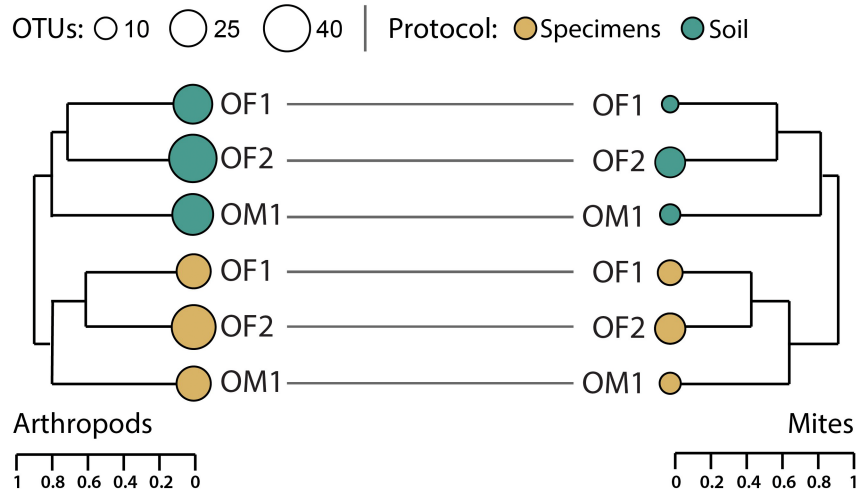Full-size ⬚ DOI: 10.7717/peerj.12845/fig-5

Figure 6 Hierarchical clustering of pairwise Sorensen dissimilarity values for arthropod and mite OTUs detected from three paired specimen and soil samples.

Full-size ⬚ DOI: 10.7717/peerj.12845/fig-6

Diversity patterns of soil arthropod assemblages

The 700K post-filtered reads from the 19 soil samples included 1,090 identified OTUs representing 123 families and 57 orders (Table 1). Most of these OTUs derived from non-arthropod invertebrates (449) followed by arthropods (356), protists (262), fungi (17), bacteria (8), and chordates (1). Although slopes varied among the samples, the accumulation curves for all identified OTUs did not reach an asymptote for any sample based on the eight technical replicates. A similar pattern was observed for the arthropods alone (Fig. 4). On average, 86 (SD = 42) OTUs including 25 (SD = 21) arthropod and 11 (SD = 12) mite OTUs were recovered from a sample.

Compositional differences between samples increased significantly with geographic distance for both the arthropod (Mantel statistic $r = 0.51$, $p = 0.001$) and mite (Mantel statistic $r = 0.41$, $p = 0.001$) assemblages. The arthropod composition was significantly different between sites and between ecoregions, but differences among sites (perMANOVA $R^2 = 0.44$, $p = 0.007$) explained more variation in β-diversity than differences among ecoregions (perMANOVA $R^2 = 0.25$, $p < 0.001$). However, hierarchical clustering of pairwise β-diversities revealed that most samples clustered by ecoregion while paired samples from just five sites (L1, L2, V1, V3, V4) formed sister groups (Fig. 6). More variation in mite β-diversity was also explained by differences among sites (perMANOVA $R^2 = 0.42$, $p = 0.049$) than among ecoregions (perMANOVA $R^2 = 0.28$, $p < 0.01$). The hierarchical clustering of pairwise mite β-diversities also revealed that most samples clustered by ecoregion, but three samples were outliers (L3|2, V1|1, V1|2). No mites were detected in V5|2, and samples from just two sites (L1, L2) formed sister groups (Fig. 7).

Hierarchical partitions of diversity were similar for arthropods and mites; β-diversity at the finest scale (sample) was low and increased with each hierarchical level with ecoregions contributing the greatest proportion of β-diversity to γ-diversity (Fig. 7). Most partitions were greater for all arthropods than for mites alone, but multiplicative partitions of βSite and βEcoregion were higher for mites (Fig. 7). Compared to null models, partitions of αSample were higher than expected by chance ($p < 0.0001$) for all arthropods (3.9%) and for mites (5.3%). Partitions of β-diversity were also higher than expected ($p < 0.0001$) for arthropods (additive = 12.3%, multiplicative = 20.8%) and mites (additive = 10.7%, multiplicative = 18.9%) at the ecoregion level, but not at the sample or site levels (Fig. S2).

DISCUSSION

The present study compared the efficacy of two bulk sampling approaches (specimens, soil) for the assessment of soil arthropod diversity by DNA metabarcoding. Interpretation of the resulting sequence data was constrained because just 41% of reads and 15% of OTUs were identified at any rank. Despite this constraint, a broad range of invertebrate and microbial taxa were detected from the analysis of just three specimen and 19 soil samples. The overall data revealed 410 arthropod OTUs belonging to 112 families, 25 orders, and nine classes. The level of taxonomic resolution differed among groups reflecting variation in the completeness of their reference library on BOLD. Resolution was typically highest for arthropods owing to extensive surveys of the Canadian fauna (*de Waard et al., 2019*;
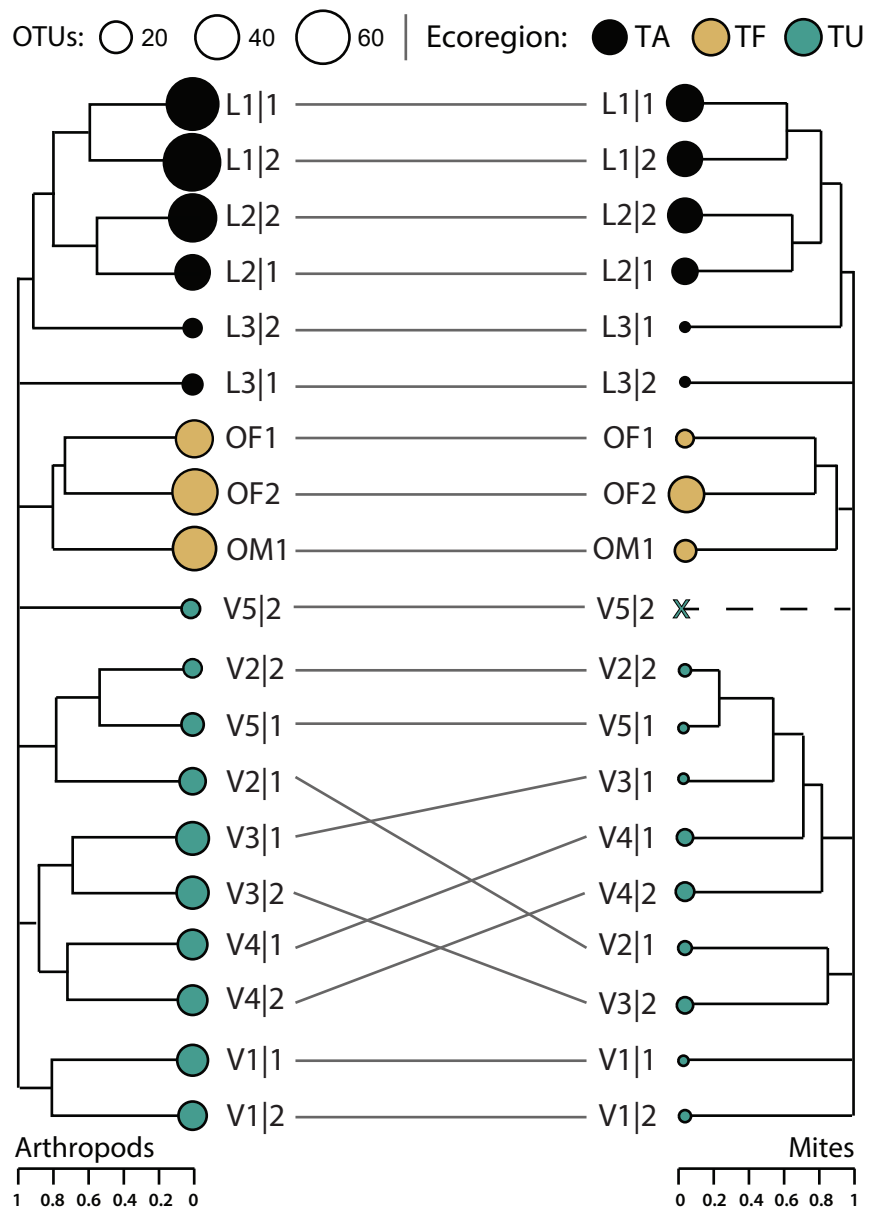
Figure 7 Hierarchical clustering of pairwise Sorensen dissimilarity values for arthropod and mite OTUs detected from 19 soil samples from 11 sites and three ecoregions. (TA, Taiga/Arctic Cordillera transition; TF, Temperate Forest; TU, Tundra) within Canada.

Full-size ⊡ DOI: 10.7717/peerj.12845/fig-7

Hebert et al., 2016), including studies near **rare** (*Telfer et al., 2015*) and near Cambridge Bay (*Pentinsaari et al., 2020*). Identification success was highest for insects (*Hebert et al., 2016*) and spiders (*Blagoev et al., 2016*). By comparison, identification success for mites was low despite an extensive survey of the Canada fauna, confirming the prediction that many species in this group await barcode characterization (*Young et al., 2019*). Success was also low for highly diverse non-arthropod groups that are rarely surveyed in soil such as the bdelloid rotifers (*Robeson et al., 2011*) and protists, especially heterokonts (*Pawlowski*

et al., 2012). Proteobacteria were likely also hyperdiverse but largely undetected due to a lack of close matches (*e.g.*, >85% similarity) in the reference library. Expanding reference library coverage for these groups would substantially increase the proportion of identified OTUs, raising the taxonomic resolution of soil metabarcoding studies. Despite biased and incomplete library coverage, our analyses demonstrate that DNA metabarcoding of bulk specimen and soil samples recovered many of the arthropod taxa isolated by Berlese-Tullgren funnels and revealed generally congruent patterns of α- and β-diversity.

Taxon recovery

Bulk soil analysis recovered significantly more orders, families, and OTUs than specimen analysis reflecting the higher recovery of non-target taxa from bulk soil DNA. The amplification of non-target taxa from soil DNA extracts is inescapable with degenerate primers (*Dopheide et al., 2019a*; *Marquina et al., 2019*; *Saitoh et al., 2016*) like those employed in this study (*Prosser et al., 2016*). Degenerate primers will also amplify NUMTs, a problem that can be reduced by discarding OTUs with stop codons, or those lacking a close match in the reference library (*Creedy et al., 2021*). However, such filtering would also reduce the sensitivity of DNA metabarcoding studies generated by platforms prone to indel errors and from highly diverse assemblages that are poorly represented in the reference library. Despite our adoption of a modest quality filter, just one apparent NUMT was detected in our post-filtered dataset. However, discriminating NUMTs from sequencing artefacts is difficult, especially for those lacking stop codons or large indels. For example, nearly 20% of the OTUs in our dataset possessed a single bp indel in a homopolymer tract consistent with IonTorrent's sequencing error profile (*Piñol et al., 2015*), but similar indels were observed in many suspected NUMTs from the positive control. Ultimately, degenerate primers are necessary to minimize amplification bias and to characterize a breadth of taxa as wide as the soil arthropods (*Krehenwinkel et al., 2017*) but the recovery of NUMTs will complicate data interpretation.

While most (59%) false positives encountered in the positive control were likely NUMTs, almost a third (31%) potentially derived from gut contents or other trace DNA associated with the specimens. The two unconfirmed families (Trombidiformes: Scutacaridae, Tydeidae) detected by bulk specimen analysis may also represent trace DNA but were most likely present in the sample and lost during voucher recovery, because of their small body size (<200 µm). Some tag-switching was observed in the negative controls although all cases had low abundances (<6 reads). As a result, abundance-based filtering was applied to the remaining samples but a few cases of tag-switching persisted in the positive control despite this intervention and may have also been present in the specimen/soil samples. In the positive control, tag-switching occurrences always represented less than 0.05% of the total read count for an OTU. As a consequence, they could have been eliminated by the use of dynamic abundance filtering based on total OTU abundance (in addition to the sample abundance-based filtering that was employed).

Despite the prevalence of non-target recovery, most (80%) of the arthropod families present in the specimen samples were recovered by metabarcode analysis. However, the present analysis undoubtedly recovered just a fraction of the species in each sample since

sequences were recovered from less than a quarter of the BINs in the positive control in each run. These undetected taxa included species overlooked by the relatively low sequencing depth employed in this study because of their low biomass (*Elbrecht, Peinert & Leese, 2017*) or low affinity with the primers (*Elbrecht et al., 2019*). A previous analysis utilizing the same primer pair and sequencing platform as this study revealed that 100K reads were necessary to recover 95% of the species in a bulk sample that represented a simple challenge as it included a single individual of 374 arthropod species (*Braukmann et al., 2019*). Because the mean sequencing depth in our study averaged just 21K reads per technical replicate (169K per sample), sequencing coverage was clearly inadequate to reveal all taxa, as demonstrated by the fact that accumulation curves for OTUs across the 8 technical replicates reached an asymptote for just one of the 22 samples. The slopes of these curves were consistently higher for bulk soil than specimen samples, indicating that a much larger proportion of bulk soil richness was undetected. While optimizing extraction and PCR protocols will certainly improve OTU recovery from bulk soil samples (*Dopheide et al., 2019b*), future work should also evaluate the ideal level of replication and sequence coverage required to adequately capture bulk soil diversity (*Lanzén et al., 2017*). However, it is near certain that a 25x–50x increase in read depth will be required to reach asymptotic richness of a sample.

Although the composition of arthropod OTUs recovered by the paired specimen and soil samples were not statistically different, this result likely reflects a Type II error due to low sample size ($n = 3$) since a large proportion of the variation in β-diversity ($R^2 < 0.3$) was explained by compositional differences between the two sampling methods. While improving the recovery of OTUs may reduce compositional differences, they will more likely be magnified since the slopes of the accumulation curves suggest that OTU richness from bulk soil will greatly exceed that from specimens given sufficient sampling. Some compositional differences between the specimen and soil samples undoubtedly reflect the physical presence/absence of taxa in each subsample. However, most differences likely stem from systematic biases between the methods, since some taxa are infrequently isolated by Berlese-Tullgren funnels (*André, Ducarme & Lebrun, 2002*) while bulk soil will also detect rare or transient taxa through environmental traces of their DNA (*Deiner et al., 2017*). Relic DNA is also abundant in soils and may inflate estimates of arthropod richness (*Nagler et al., 2018*). However, the magnitude of this effect depends on the persistence of relic DNA in the environment (*Nielsen et al., 2007*) which remains uncertain for arthropods in the soil. Regardless of the cause, it is clear that bulk specimen and soil analyses capture complementary portions of the soil fauna.

Patterns of arthropod α- and β-diversity

Despite low read depth, limited sample size, and compositional differences, results from the three paired specimen and soil samples indicated similar patterns of arthropod α- and β-diversity. These results confirm prior studies that revealed similar estimates of arthropod order, family, and OTU richness obtained from metabarcode analyses of soil and soil-isolated specimens (*Dopheide et al., 2019a*; *Oliverio et al., 2018*), demonstrating their ability to recover ecological signal despite the incomplete recovery of OTUs. For example, compositional differences between the meadow and forested sites were consistently greater

than differences between the pair of forested sites regardless of the taxon or protocol examined. These results support prior studies revealing strong differentiation between grassland and forest soil communities through morphological (*Caruso, Taormina & Migliorini, 2012*) and molecular analyses (*Arribas et al., 2021*). Diversity patterns revealed by easily collected soil samples may parallel those seen in above-ground communities (*i.e.*, leaf litter and Malaise trap samples; *Yang et al., 2014*). If so, bulk soil analyses could provide a rapid, cost-effective solution for terrestrial biomonitoring. Although sufficient replication and read depth will be necessary to develop comprehensive baselines for the soil fauna using this method, adopting standardized methods should also limit analytical biases and allow robust assessments of diversity patterns despite undersampling (*Mathieu et al., 2020*).

Our comparison of soil samples across the three ecoregions confirmed previously observed patterns of distance-decay congruent with dispersal-limited, neutral community assembly for mites and other soil arthropod assemblages (*Arribas et al., 2021*; *Young et al., 2019*; *Zinger et al., 2019*). The importance of neutral processes in shaping soil arthropod assembly was supported by partitions of β-diversity at the sample and site scales which did not differ from null models. However, high turnover was observed even between close sites (*e.g.*, between transects just 10 m apart). This pattern could result from undersampling (*Beck, Holloway & Schwanghart, 2013*), but may also reflect the highly patchy distributions of soil arthropods (*Bardgett, 2002*). For example, environmental factors such as soil type and ground cover have been linked to high variability in local soil mite assemblages (*Lindo & Winchester, 2009*; *Meehan, Song & Proctor, 2018*), while moisture, temperature, and organic matter are known to shape the assembly of soil arthropod communities at very fine scales (*Ghiglieno et al., 2020*). Environmentally driven assembly at a micro scale was further supported by higher than expected partitions of αSample, indicating that mechanisms promoting species coexistence support locally rich assemblages of soil arthropods (*Leibold & Chase, 2018*). Partitions of β Ecoregion were also higher than expected by chance, and samples typically clustered by ecoregion, mirroring patterns of non-neutral assembly previously observed for soil mite assemblages in Canadian protected areas (*Young et al., 2019*). Most communities are structured by both neutral and niche processes (*Cottenie, 2005*), and our results affirm the importance of both in shaping soil arthropod diversity across the Canadian landscape.

CONCLUSION

This study has demonstrated that DNA metabarcoding can greatly extend knowledge of soil arthropod communities. Although there was little species overlap, patterns of α- and β-diversity for soil arthropods were congruent between bulk soil and specimen samples. Because of its simplicity, the analysis of soil samples is conducive to large-scale surveys. Increased sequencing depth will improve taxon recovery, while expansion of the DNA barcode reference library will enhance the resolution of taxonomic assignments generated from metabarcode data.

ACKNOWLEDGEMENTS

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Grant Disclosures

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Monica R. Young conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored and reviewed drafts of the paper, and approved the final draft.
- Paul D. N. Hebert conceived and designed the experiments, reviewed drafts of the paper, and approved the final draft.

Field Study Permissions

The following information was supplied relating to field study approvals (i.e., approving body and any reference numbers):

Samples were collected with permission at the *rare* Charitable Research Reserve. Samples collected in the vicinity of the Canadian High Arctic Research Station were contained with the Department of Environment Wildlife Research Permit (WL 2018-044). A permit was also obtained through the Canada C3 Initiative for sampling in National Parks in Labrador (ANP-2017-25145) and collecting in other sites was approved by the Nunatsiavut Government Research Advisory Committee.

Data Availability

The following information was supplied regarding data availability:

The raw sequence data is available at NCBI's Sequence Read Archive (BioProject: PRJNA755277).

The R scripts for bioinformatics and statistical analysis are available in the Supplementary Files.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.12845#supplemental-information.

REFERENCES

Anderson JM. 1975. The enigma of soil animal species diversity. In: Vaněk J, ed. *Progress in soil zoology*. Dordrecht: Springer Netherlands, 51–58 DOI 10.1007/978-94-010-1933-0_5.

André H, Ducarme X, Anderson J, Jr DCrossley, Koehler H, Paoletti M, Walter D, Lebrun P. 2001. Skilled eyes are needed to go on studying the richness of the soil. *Nature* **409(6822)**:761 DOI 10.1038/35057493.

André HM, Ducarme X, Lebrun P. 2002. Soil biodiversity: myth, reality or conning?. *Oikos* **96(1)**:3–24 DOI 10.1034/j.1600-0706.2002.11216.x.

André HM, Noti M-I, Lebrun P. 1994. The soil fauna: the other last biotic frontier. *Biodiversity and Conservation* **3(1)**:45–56 DOI 10.1007/BF00115332.

Arribas P, Andújar C, Hopkins K, Shepherd M, Vogler AP. 2016. Metabarcoding and mitochondrial metagenomics of endogean arthropods to unveil the mesofauna of the soil. *Methods in Ecology and Evolution* **7(9)**:1071–1081 DOI 10.1111/2041-210X.12557.

Arribas P, Andújar C, Salces-Castellano A, Emerson BC, Vogler AP. 2021. The limited spatial scale of dispersal in soil arthropods revealed with whole-community haplotype-level metabarcoding. *Molecular Ecology* **30(1)**:48–61 DOI 10.1111/mec.15591.

Bardgett R. 2002. Causes and consequences of biological diversity in soil. *Zoology* **105(4)**:367–375 DOI 10.1078/0944-2006-00072.

Baselga A, Orme CDL. 2012. betapart: an R package for the study of beta diversity: *Betapart package*. *Methods in Ecology and Evolution* **3(5)**:808–812 DOI 10.1111/j.2041-210X.2012.00224.x.

Beck J, Holloway JD, Schwanghart W. 2013. Undersampling and the measurement of beta diversity. *Methods in Ecology and Evolution* **4(4)**:370–382 DOI 10.1111/2041-210x.12023.

Bickford D, Lohman DJ, Sodhi NS, Ng PKL, Meier R, Winker K, Ingram KK, Das I. 2007. Cryptic species as a window on diversity and conservation. *Trends in Ecology & Evolution* **22(3)**:148–155 DOI 10.1016/j.tree.2006.11.004.

Borisenko AV, Sones JE, Hebert PDN. 2009. The front-end logistics of DNA barcoding: challenges and prospects. *Molecular Ecology Resources* **9(s1)**:27–34 DOI 10.1111/j.1755-0998.2009.02629.x.

Blagoev GA, Ratnasingham S, de Waard JR, Lu L, de Waard SL, Robertson J, Telfer AC, Hebert PDN. 2016. Untangling taxonomy: a DNA barcode reference library for Canadian spiders. *Molecular Ecology Resources* **16(1)**:325–341 DOI 10.1111/1755-0998.12444.

Braukmann TWA, Ivanova NV, Prosser SWJ, Elbrecht V, Steinke D, Ratnasingham S, de Waard JR, Sones JE, Zakharov EV, Hebert PDN. 2019. Metabarcoding a diverse arthropod mock community. *Molecular Ecology Resources* **19(3)**:711–727 DOI 10.1111/1755-0998.13008.

Buchner D, Leese F. 2020. BOLDigger –a Python package to identify and organise sequences with the Barcode of Life Data systems. *Metabarcoding and Metagenomics* **4**:e53535 DOI 10.3897/mbmg.4.53535.

Cameron EK, Martins IS, Lavelle P, Mathieu J, Tedersoo L, Gottschall F, Guerra CA, Hines J, Patoine G, Siebert J, Winter M, Cesarz S, Delgado-Baquerizo M, Ferlian O, Fierer N, Kreft H, Lovejoy TE, Montanarella L, Orgiazzi A, Pereira HM, Phillips HRP, Settele J, Wall DH, Eisenhauer N. 2018. Global gaps in soil biodiversity data. *Nature Ecology & Evolution* **2(7)**:1042–1043 DOI 10.1038/s41559-018-0573-8.

Caruso T, Taormina M, Migliorini M. 2012. Relative role of deterministic and stochastic determinants of soil animal community: a spatially explicit analysis of oribatid mites: deterministic and stochastic determinants of soil animal community. *Journal of Animal Ecology* **81(1)**:214–221 DOI 10.1111/j.1365-2656.2011.01886.x.

Cottenie K. 2005. Integrating environmental and spatial processes in ecological community dynamics: meta-analysis of metacommunities. *Ecology Letters* **8(11)**:1175–1182 DOI 10.1111/j.1461-0248.2005.00820.x.

Creedy T, Andujar C, Meramveliotakis E, Noguerales V, Overcast I, Papadopoulou A, Morlon H, Vogler A, Emerson B, Arribas P. 2021. Coming of age for COI metabarcoding of whole organism community DNA: towards bioinformatic harmonisation [Preprint]. *Authorea* DOI 10.22541/au.162141276.61766048/v2.

Crist TO, Veech JA, Gering JC, Summerville KS. 2003. Partitioning species diversity across landscapes and regions: a hierarchical analysis of α, β, and γ diversity. *The American Naturalist* **162(6)**:734–743 DOI 10.1086/378901.

Decaëns T. 2008. Priorities for conservation of soil animals. *CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources* **3(014)**:18 DOI 10.1079/PAVSNNR20083014.

Deiner K, Bik HM, Mächler E, Seymour M, Lacoursière-Roussel A, Altermatt F, Creer S, Bista I, Lodge DM, de Vere N, Pfrender ME, Bernatchez L. 2017. Environmental DNA metabarcoding: transforming how we survey animal and plant communities. *Molecular Ecology* **26(21)**:5872–5895 DOI 10.1111/mec.14350.

de Waard JR, Ratnasingham S, Zakharov EV, Borisenko AV, Steinke D, Telfer AC, Hebert PDN. 2019. A reference library for Canadian invertebrates with 1.5 million barcodes, voucher specimens, and DNA samples. *Scientific Data*. 308 DOI 10.1038/s41597-019-0320-2.

Donoso DA, Johnston MK, Kaspari M. 2010. Trees as templates for tropical litter arthropod diversity. *Oecologia* **164(1)**:201–211 DOI 10.1007/s00442-010-1607-3.

Dopheide A, Tooman LK, Grosser S, Agabiti B, Rhode B, Xie D, Stevens MI, Nelson N, Buckley TR, Drummond AJ, Newcomb RD, Newcomb RD. 2019a. Estimating the biodiversity of terrestrial invertebrates on a forested island using DNA barcodes and metabarcoding data. *Ecological Applications* **29(4)**:e01877 DOI 10.1002/eap.1877.

Dopheide A, Xie D, Buckley TR, Drummond AJ, Newcomb RD. 2019b. Impacts of DNA extraction and PCR on DNA metabarcoding estimates of soil biodiversity. *Methods in Ecology and Evolution* **10(1)**:120–133 DOI 10.1111/2041-210X.13086.

Dray S, Dufour A-B. 2007. The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software* **22(4)**:1–20 DOI 10.18637/jss.v022.i04.

Edgar RC. 2013. UPARSE: Highly accurate OTU sequences from microbial amplicon reads. *Nature Methods* **10(10)**:996–998 DOI 10.1038/nmeth.2604.

Edgar RC, Flyvbjerg H. 2015. Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics* **31(21)**:3476–3482 DOI 10.1093/bioinformatics/btv401.

Elbrecht V. 2018. JAMP: just another metabarcoding pipeline. R Package Version 0.67.

Elbrecht V, Braukmann TWA, Ivanova NV, Prosser SWJ, Hajibabaei M, Wright M, Zakharov EV, Steinke D. 2019. Validation of COI metabarcoding primers for terrestrial arthropods. *PeerJ* **7**:e7745 DOI 10.7717/peerj.7745.

Elbrecht V, Peinert B, Leese F. 2017. Sorting things out: assessing effects of unequal specimen biomass on DNA metabarcoding. *Ecology and Evolution* **7(17)**:6918–6926 DOI 10.1002/ece3.3192.

EPA. 2015. Ecoregions of North America [Data and Tools]. *Available at https://www.epa.gov/eco-research/ecoregions-north-america*.

Erwin TL. 1982. Tropical forests: their richness in Coleoptera and other arthropod species. *The Coleopterists Bulletin* **36(1)**:74–75.

Geisen S, Briones MJI, Gan H, Behan-Pelletier VM, FrimanVM V-P, de Groot GA, Hannulaa SE, Lindo Z, Philippot L, Tiunov AV, Wall DH. 2019. A methodological framework to embrace soil biodiversity. *Soil Biology and Biochemistry* **136**:107536 DOI 10.1016/j.soilbio.2019.107536.

Ghiglieno I, Simonetto A, Orlando F, Donna P, Tonni M, Valenti L, Gilioli G. 2020. Response of the arthropod community to soil characteristics and management in the Franciacorta viticultural area (Lombardy, Italy). *Agronomy* **10(5)**:740 DOI 10.3390/agronomy10050740.

Hebert PDN, Braukmann TWA, Prosser SWJ, Ratnasingham S, de Waard JR, Ivanova NV, Janzen DH, Hallwachs W, Naik S, Sones JE, Zakharov EV. 2018. A sequel to sanger: amplicon sequencing that scales. *BMC Genomics* **19**(1):14 DOI 10.1186/s12864-018-4611-3.

Hebert PDN, Cywinska A, Ball SL, de Waard JR. 2003. Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences* **270**(1512):313–321 DOI 10.1098/rspb.2002.2218.

Hebert PDN, de Waard JR, Zakharov EV, Prosser SWJ, Sones JE, McKeown JTA, Mantle B, La Salle J. 2013. A DNA 'Barcode Blitz': rapid digitization and sequencing of a natural history collection. *PLOS ONE* **8**(7):e68535 DOI 10.1371/journal.pone.0068535.

Hebert PDN, Ratnasingham S, Levesque-Beaudin V, Zakharov EV, Telfer AC, Milton MA, Pedersen S, Janetta P, de Waard JR. 2016. Counting animal species with DNA barcodes: Canadian insects. *Philosophical Transactions of the Royal Society B: Biological Sciences* **371**(1702):20150333 DOI 10.1098/rstb.2015.0333.

Hendrich L, Morinière J, Haszprunar G, Hebert PDN, Hausmann A, Köhler F, Balke M. 2015. A comprehensive DNA barcode database for Central European beetles with a focus on Germany: adding more than 3500 identified species to BOLD. *Molecular Ecology Resources* **15**(4):795–818 DOI 10.1111/1755-0998.12354.

Hernández-Triana LM, Prosser SW, Rodríguez-Perez MA, Chaverri LG, Hebert PDN, Gregory TR. 2014. Recovery of DNA barcodes from blackfly museum specimens (Diptera: Simuliidae) using primer sets that target a variety of sequence lengths. *Molecular Ecology Resources* **14**(3):508–518 DOI 10.1111/1755-0998.12208.

Hogg ID, Hebert PDN. 2004. Biological identification of springtails (Hexapoda: Collembola) from the Canadian Arctic, using mitochondrial DNA barcodes. *Canadian Journal of Zoology* **82**(5):749–754 DOI 10.1139/z04-041.

Ivanova NV, Fazekas AJ, Hebert PDN. 2008. Semi-automated, membrane-based protocol for DNA isolation from plants. *Plant Molecular Biology Reporter* **26**(3):186–198 DOI 10.1007/s11105-008-0029-4.

Ivanova NV, de Waard JR, Hebert PDN. 2006. An inexpensive, automation-friendly protocol for recovering high-quality DNA. *Molecular Ecology Notes* **6**(4):998–1002 DOI 10.1111/j.1471-8286.2006.01428.x.

Kindt R, Coe R. 2005. Tree diversity analysis. In: *A manual and software for common statistical methods for ecological and biodiversity studies.* Nairobi: World Agroforestry Centre (ICRAF).

Krehenwinkel H, Wolf M, Lim JY, Rominger AJ, Simison WB, Gillespie RG. 2017. Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Scientific Reports* **7**(1):17668 DOI 10.1038/s41598-017-17333-x.

Lanzén A, Lekang K, Jonassen I, Thompson EM, Troedsson C. 2017. DNA extraction replicates improve diversity and compositional dissimilarity in metabarcoding of eukaryotes in marine sediments. *PLOS ONE* **12**(6):e0179443 DOI 10.1371/journal.pone.0179443.

Leibold MA, Chase JM. 2018. *Metacommunity ecology.* Volume 59. Princeton: Princeton University Press.

Lindo Z, Winchester NN. 2009. Spatial and environmental factors contributing to patterns in arboreal and terrestrial oribatid mite diversity across spatial scales. *Oecologia* **160(4)**:817–825 DOI 10.1007/s00442-009-1348-3.

Marquina D, Esparza-Salas R, Roslin T, Ronquist F. 2019. Establishing arthropod community composition using metabarcoding: surprising inconsistencies between soil samples and preservative ethanol and homogenate from Malaise trap catches. *Molecular Ecology Resources* **19(6)**:1516–1530 DOI 10.1111/1755-0998.13071.

Marra JL, Edmonds RL. 2005. Soil arthropod responses to different patch types in a mixed-conifer forest of the Sierra Nevada. *Forest Science* **51(3)**:255–265 DOI 10.1093/forestscience/51.3.255.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal* **17(1)**:10–12.

Mathieu C, Hermans SM, Lear G, Buckley TR, Lee KC, Buckley HL. 2020. A systematic review of sources of variability and uncertainty in eDNA data for environmental monitoring. *Frontiers in Ecology and Evolution* **8**:135 DOI 10.3389/fevo.2020.00135.

Meehan ML, Song Z, Proctor H. 2018. Roles of environmental and spatial factors in structuring assemblages of forest-floor Mesostigmata in the boreal region of Northern Alberta, Canada. *International Journal of Acarology* **44(7)**:300–309 DOI 10.1080/01647954.2018.1520297.

Nagler M, Insam H, Pietramellara G, Ascher-Jenull J. 2018. Extracellular DNA in natural environments: Features, relevance and applications. *Applied Microbiology and Biotechnology* **102(15)**:6343–6356 DOI 10.1007/s00253-018-9120-4.

Oksanen Blanchet JFG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H. 2015. vegan: Community ecology package. R Package Version 2.3.

Nielsen KM, Johnsen PJ, Bensasson D, Daffonchio D. 2007. Release and persistence of extracellular DNA in the environment. *Environmental Biosafety Research* **6(1)**:37–53 DOI 10.1051/ebr:2007031.

Oliverio AM, Gan H, Wickings 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.

Pawlowski J, Audic S, Adl S, Bass D, Belbahri L, Berney C, de Vargas C. 2012. CBOL protist working group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLOS Biology* **10(11)**:e1001419 DOI 10.1371/journal.pbio.1001419.

Pentinsaari M, Levesque-Beaudin V, Blagoev GA, Hogg ID, Perez K, Sobel CN, Borisenko A, Vandenbrink B. 2020. A DNA barcoding survey of an Arctic arthropod community: implications for future monitoring. *Insects* **11(1)**:46 DOI 10.3390/insects11010046.

Piñol J, Mir G, Gomez-Polo P, Agustí N. 2015. Universal and blocking primer mis-matches limit the use of high-throughput DNA sequencing for the quantita-tive metabarcoding of arthropods. *Molecular Ecology Resources* **15(4)**:819–830 DOI 10.1111/1755-0998.12355.

Prosser SWJ, de Waard SE, Miller JR, Hebert PDN. 2016. DNA barcodes from century-old type specimens using next-generation sequencing. *Molecular Ecology Resources* **16(2)**:487–497 DOI 10.1111/1755-0998.12474.

Ratnasingham S, Hebert PDN. 2007. BOLD: The Barcode of Life Data system (http://www.barcodinglife.org). *Molecular Ecology Resources* **7(3)**:355–364.

Ratnasingham S, Hebert PDN. 2013. A DNA-based registry for all animal species: The Barcode Index Number (BIN) System. *PLOS ONE* **8(7)**:e66213 DOI 10.1371/journal.pone.0066213.

Raupach M, Hannig K, Moriniere J, Hendrich L. 2016. A DNA barcode library for ground beetles (Insecta, Coleoptera, Carabidae) of Germany: The genus *Bembidion* Latreille, 1802 and allied taxa. *ZooKeys* **592**:121–141 DOI 10.3897/zookeys.592.8316.

R Core Team. 2020. *R: a language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing.

Robeson MS, King AJ, Freeman KR, Birky CW, Martin AP, Schmidt SK. 2011. Soil rotifer communities are extremely diverse globally but spatially autocorrelated locally. *Proceedings of the National Academy of Sciences of the United States of America* **108(11)**:4406–4410 DOI 10.1073/pnas.1012678108.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**:e2584 DOI 10.7717/peerj.2584.

Saitoh S, Aoyama H, Fujii S, Sunagawa H, Nagahama H, Akutsu M, Shinzato N, Kaneko N, Nakamori T. 2016. A quantitative protocol for DNA metabarcoding of springtails (Collembola). *Genome* **59(9)**:705–723 DOI 10.1139/gen-2015-0228.

Schäffer S, Kerschbaumer M, Koblmüller S. 2019. Multiple new species: Cryptic diversity in the widespread mite species Cymbaeremaeus cymba (Oribatida, Cymbaeremaeidae). *Molecular Phylogenetics and Evolution* **135**:185–192 DOI 10.1016/j.ympev.2019.03.008.

Schenekar T, Schletterer M, Lecaudey LA, Weiss SJ. 2020. Reference databases, primer choice, and assay sensitivity for environmental metabarcoding: Lessons learnt from a re-evaluation of an eDNA fish assessment in the Volga headwaters. *River Research and Applications* **36(7)**:1004–1013 DOI 10.1002/rra.3610.

Stork NE. 1988. Insect diversity: facts, fiction and speculation. *Biological Journal of the Linnean Society* **35(4)**:321–337 DOI 10.1111/j.1095-8312.1988.tb00474.x.

Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH. 2012. Environmental DNA. *Molecular Ecology* **21(8)**:1789–1793 DOI 10.1111/j.1365-294X.2012.05542.x.

Telfer A, Young M, Quinn J, Perez K, Sobel C, Sones JE, Levesque-Beaudin V, Derbyshire R, Fernandez-Triana J, Rougerie R, Thevanayagam A, Boskovic A, Borisenko AV, Cadel A, Brown A, Pages A, Castillo AH, Nicolai A, Mockford B, Mockford G, Bukowski B, Wilson B, Trojahn B, Lacroix CA, Brimblecombe C, Hay C, Ho C, Steinke C, Warne CP, Cortes CG, Engelking D, Wright D, Lijtmaer DA, Gascoigne D, Martich DH, Morningstar D, Neumann D, Steinke D, DeBruin D, DeBruin M, Dobias D, Sears E, Richard E, Damstra E, Zakharov EV, Laberge F, Collins GE, Blagoev GA, Grainge G, Ansell G, Meredith G, Hogg I, McKeown J, Topan J, Bracey J, Guenther J, Sills-Gilligan J, Addesi J, Persi J, Layton KKS,

D'Souza K, Dorji K, Grundy K, Nghidinwa K, Ronnenberg K, Lee KM, Gonzalez M, Rosati ME, Xie L, Lu L, Penev L, Gonzalez M, Rosati ME, Kekkonen M, Kuzmina M, Iskandar M, Mutanen M, Fatahi M, Pentinsaari M, Bauman M, Nikolova N, Ivanova NV, Jones N, Weerasuriya N, Monkhouse N, Lavinia PD, Jannetta P, Hanisch PE, McMullin RT, Flores RO, Mouttet R, Vender R, Labbee RN, Forsyth R, Lauder R, Dickson R, Kroft R, Miller SE, MacDonald S, Panthi S, Pedersen S, Sobek-Swant S, Naik S, Lipinskaya T, Eagalle T, Decaëns T, Kosuth T, Braukmann T, Woodcock T, Roslin T, Zammit T, Campbell V, Dinca V, Peneva V, Hebert PDN, de Waard J. 2015. Biodiversity inventories in high gear: DNA barcoding facilitates a rapid biotic survey of a temperate nature reserve. *Biodiversity Data Journal* **3**:e6313 DOI 10.3897/BDJ.3.e6313.

Yang C, Wang X, Miller JA, de Blécourt M, Ji Y, Yang C, Harrison RD, Yu DW. 2014. Using metabarcoding to ask if easily collected soil and leaf-litter samples can be used as a general biodiversity indicator. *Ecological Indicators* **46**:379–389 DOI 10.1016/j.ecolind.2014.06.028.

Veech JA, Crist TO. 2009. PARTITION: software for hierarchical partitioning of species diversity. Version 3.0. *Available at http://www.users.miamioh.edu/cristto/partition.html*.

Young MR, Proctor HC, de Waard JR, Hebert PDN. 2019. DNA barcodes expose unexpected diversity in Canadian mites. *Molecular Ecology* **28(24)**:5347–5359 DOI 10.1111/mec.15292.

Young MR, de Waard JR, Hebert PDN. 2021. DNA barcodes enable higher taxonomic assignments in the Acari. *Scientific Reports* **11(1)**:15922 DOI 10.1038/s41598-021-95147-8.

Zahiri R, Lafontaine JD, Schmidt BC, de Waard JR, Zakharov EV, Hebert PDN. 2017. Probing planetary biodiversity with DNA barcodes: the Noctuoidea of North America. *PLOS ONE* **12(6)**:e0178548 DOI 10.1371/journal.pone.0178548.

Zinger L, Taberlet P, Schimann H, Bonin A, Boyer F, De Barba M, Gaucher P, Gielly L, Giguet-Covex C, Iribar A, Réjou-Méchain M, Rayé G, Rioux D, Schilling V, Tymen B, Viers J, Zouiten C, Thuiller W, Coissac E, Chave J. 2019. Body size determines soil community assembly in a tropical forest. *Molecular Ecology* **28(3)**:528–543 DOI 10.1111/mec.14919.