SCIENTIFIC REPORTS
nature research

OPEN

Variations in terrestrial arthropod DNA metabarcoding methods recovers robust beta diversity but variable richness and site indicators

Teresita M. Porter[1,2*], Dave M. Morris[3], Nathan Basiliko[4], Mehrdad Hajibabaei[2], Daniel Doucet[1], Susan Bowman[1], Erik J. S. Emilson[1], Caroline E. Emilson[1], Derek Chartrand[1], Kerrie Wainio-Keizer[1], Armand Séguin[5] & Lisa Venier[1]

Terrestrial arthropod fauna have been suggested as a key indicator of ecological integrity in forest systems. Because phenotypic identification is expert-limited, a shift towards DNA metabarcoding could improve scalability and democratize the use of forest floor arthropods for biomonitoring applications. The objective of this study was to establish the level of field sampling and DNA extraction replication needed for arthropod biodiversity assessments from soil. Processing 15 individually collected soil samples recovered significantly higher median richness (488–614 sequence variants) than pooling the same number of samples (165–191 sequence variants) prior to DNA extraction, and we found no significant richness differences when using 1 or 3 pooled DNA extractions. Beta diversity was robust to changes in methodological regimes. Though our ability to identify taxa to species rank was limited, we were able to use arthropod COI metabarcodes from forest soil to assess richness, distinguish among sites, and recover site indicators based on unnamed exact sequence variants. Our results highlight the need to continue DNA barcoding local taxa during COI metabarcoding studies to help build reference databases. All together, these sampling considerations support the use of soil arthropod COI metabarcoding as a scalable method for biomonitoring.

Soil arthropod fauna have been suggested as a key indicator of faunal community structure[1–3]. These organisms are essential to ecological processes that include organic matter decomposition, nutrient cycling, and soil structural development (e.g. micropore formation that improves aeration porosity and water infiltration rates)[1,2]. Community shifts in soil arthropods in response to anthropogenic and natural disturbance have been documented in numerous studies[3–7].

Typically, soil arthropods are sampled by trapping (e.g. pitfall traps) or they are extracted directly from soil (e.g. Tullgren funnels). Because of the large numbers of individuals that are sampled in even small studies, and because of the relative difficulty of identifying soil fauna, phenotypic identification is often expert- and time-limited. There are also significant issues of low recovery efficiency and bias in the recovery of soil fauna for phenotypic identification. A shift towards DNA metabarcoding could improve scalability and facilitate the use of soil arthropods for biomonitoring applications. DNA metabarcoding is currently the method of choice for highly scalable biodiversity studies[8]. In the literature, specific arthropod taxa such as Acari (mites and ticks), Collembola (springtails), Coleoptera (beetles) or other predefined indicator taxa have been enriched from soil or leaf litter using light or pitfall traps, Winkler extractors, or protocols based on soil flotation with a Berlese apparatus, followed by COI metabarcoding or metagenomic sequencing[9–11]. With the development of highly scalable COI metabarcoding techniques, field sampling has now become a rate-limiting step for many studies[12]. To address this issue, community sampling from bulk environmental samples such as soil are easily collected and

[1]Great Lakes Forestry Centre, Natural Resources Canada, Sault Ste. Marie, ON, P6A 2E5, Canada. [2]Biodiversity Institute of Ontario, Centre for Biodiversity Genomics & Integrative Biology, University of Guelph, Guelph, ON, N1G 2W1, Canada. [3]Ministry of Natural Resources and Forestry, Centre for Northern Forest Ecosystem Research, Thunder Bay, ON, P7E 2V6, Canada. [4]Laurentian University, Department of Biology and the Vale Living with Lakes Centre, Sudbury, ON, P3E 2C6, Canada. [5]Laurentian Forestry Centre, Natural Resources Canada, Québec, QC, G1V 4C7, Canada. *email: terrimporter@gmail.com

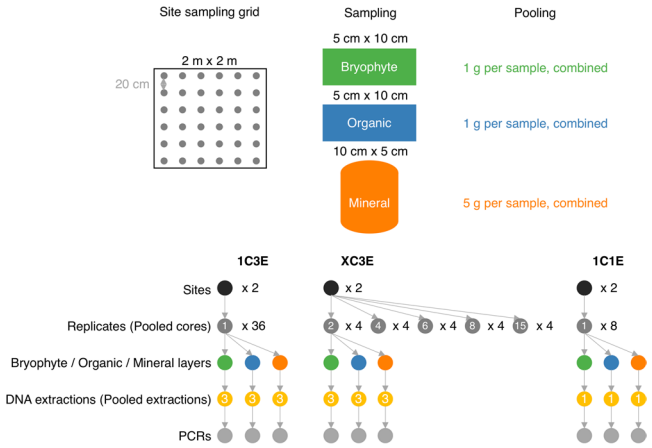

Figure 1. Overview of sampling methods. The 1C3E experiment was designed to look at the effect of increasing the volume of field soil sampled. The XC3E experiment was designed to look at the effect of the volume of field soil pooled before DNA extraction. The 1C1E and 1C3E samples were compared to look at the effect of processing 1 or 3 DNA extractions per sample. The 1C3E (1 core sample, 3 DNA extractions) experiment included 216 samples from 2 sites, 36 replicates, and 3 layers. The XC3E (2–15 pooled core samples, 3 DNA extractions) experiment included 120 samples from 2 sites, 5 pooling treatments, 4 replicates, and 3 layers. The 1C1E (1 core sample, 1 DNA extraction) experiment included 48 samples from 2 sites, 8 replicates, and 3 layers.

could facilitate repeated sampling for biomonitoring studies with a reasonable turnaround time. Previously, the mitochondrial 16S rRNA gene has been used to target Coleoptera from permafrost samples and the nuclear 18S rRNA gene regions has been used to target the metazoan community from soil[12,13]. In contrast, use of the highly variable mt COI marker for metabarcoding to survey the whole arthropod community from bulk soil is still relatively new[14,15].

For DNA metabarcoding, bulk samples of soil are homogenized, DNA from all resident organisms are extracted, and a marker gene of interest is amplified using mixed template PCR. Marker genes are chosen according to the target organism, such as the cytochrome c oxidase subunit I (COI) mitochondrial DNA (mtDNA) marker that is the official animal barcode marker and has the largest number of reference sequences for taxonomic identification[16,17]. This method produces exact sequence variants (ESVs) which are then compared to a reference sequence database[18]. The reference sequence database is built through DNA barcoding of individual specimens identified using phenotypic characters.

The objective of this study was to establish the level of replication needed for field sampling and DNA extraction procedures for COI metabarcoding of terrestrial arthropods from soil for biodiversity assessment among two similar jack pine stands of differing origins. We assessed the influence of (1) increasing spatial sampling by including more individual or pooled samples (biological replicates), (2) performing mixed template PCRs on single or pooled triplicate DNA extractions (technical replicates), and (3) sampling from bryophyte, organic, and mineral layers (Fig. 1) on observed richness, significance of sample clustering in beta diversity analyses, and the recovery of site indicators based on ESVs.

Results

Sequencing results. Raw sequence data was submitted to the NCBI SRA under BioProject PRJNA565010, BioSamples SAMN12257424–SAMN12257429. A total of ~41 million ×2 paired-end raw reads were sequenced (~110,000 reads per sample), of these ~35 million (86%) raw reads were successfully paired, and of these ~33 million (94%) paired reads were successfully primer-trimmed (Table S1). After primer trimming, the mode sequence length was ~325 bp and ~235 bp for the BE and F230R_modN markers, respectively. A total of 67,626 denoised ESVs were detected where 19,562,246 primer-trimmed reads were mapped representing ~47.5% of the original raw paired-end reads (Table S2). The phylum rank taxonomic distribution of the raw data is summarized in Fig. S1. Only the 3,598 (4.8% of all ESVs) (BE 775; F230R_modN 2,823) ESVs that were assigned to Arthropoda were retained for further analysis below (Table S3). This corresponds to ~2.7 million (6.5%) (BE 294,070; F230R_modN 2,398,638) of the original raw paired-end reads. Although we selected primers based on previous successful amplification of arthropods from freshwater benthic kicknet samples[19,20] and Malaise traps[21], the overall percentage of retained raw arthropod reads from our soil samples was low but consistent with previous work from bulk samples that are known to comprise a phylogenetically diverse mixture of taxa that are detected even when using primers originally developed to target arthropods[22,23]. Since only a proportion of arthropod ESVs could be identified with confidence, we present our results at the ESV rank wherever possible (Fig. S2). Rarefaction curves show that we saturated the sequencing of our arthropod COI PCR products (Fig. S3).

Effect of sampling method on richness. A total of 2,108 and 2,052 ESVs were detected from the Island Lake and Nimitz sites with some of the same ESVs detected across layers (Fig. S4). ESV richness increases rapidly as more individually collected samples are added to the dataset (bioinformatically pooled samples), especially for the bryophyte and organic layers (Fig. 2a). We also replicated this analysis using OTUs based on 97% sequence

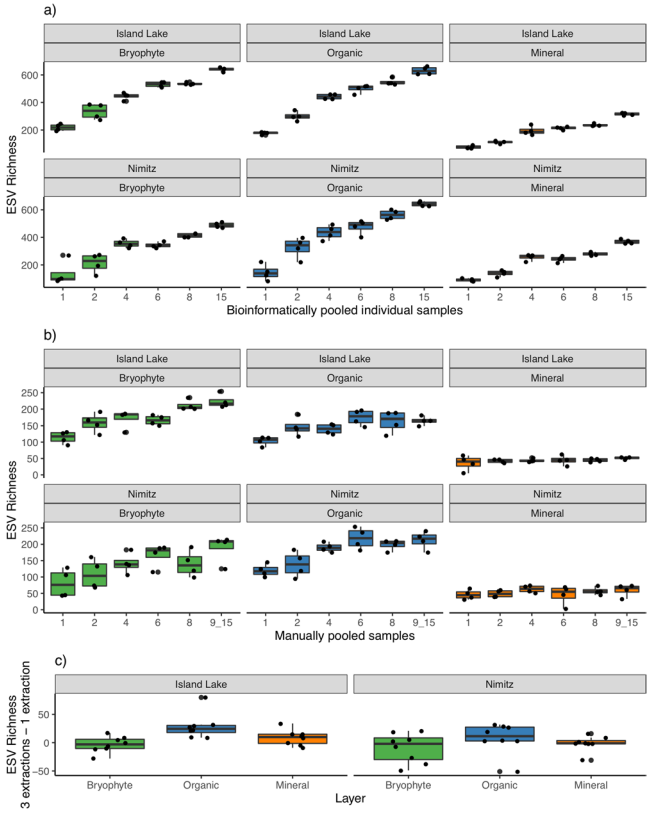

Figure 2. Arthropod ESV richness increases with increasing field sampling effort but varies little when more DNA extractions are performed. Richness is shown for (**a**) bioinformatically pooled, individually collected field samples, (**b**) manually pooled field samples, and (**c**) the difference between samples processed with 3 pooled DNA extractions and 1 DNA extraction (positive values indicate greater richness from 3 pooled DNA extractions; negative values indicate greater richness from 1 DNA extraction). '915' refers to the largest class of pooled samples that is 15 for all bioinformatically pooled samples but varies for manually pooled samples. At Island Lake the largest class is comprised of 15 pooled samples but at Nimitz, the largest class contains 15 pooled samples except for the bryophyte layer where 9–14 samples were pooled.

similarity showing similar trends as ESVs but with slightly lower richness values (not shown). Bryophyte and organic layer ESV richness also increases when more samples are manually pooled together, but at a lower rate than when individual samples are bioinformatically pooled (Fig. 2b). The median richness detected from 15 individually collected bioinformatically pooled samples ranges from 488–614 ESVs and from up to 15 manually pooled samples ranges from 165–191 ESVs across sites. Since we used rarefaction to normalize the number of sequence reads included in these comparisons, we determined that the greater richness detected from individually processed samples compared with composited samples is due to the overall difference in the amount of soil sampled not sequencing depth. For instance, a total of 33.75 g soil was extracted from 15 individually collected field samples (0.25 g × 3 layers × 3 DNA extraction replicates x 15 samples), compared with a total of 2.25 g soil from a composite of 15 pooled field samples (0.25 g × 3 layers × 3 DNA extraction replicates). We found that the ESV richness detected from a single individually collected field sample was not significantly different than processing a composite of up to 15 manually pooled samples (Pairwise Wilcox, p-value = 0.88). There was also no significant difference in the ESV richness recovered when 1 or 3 DNA extractions were performed (Pairwise Wilcox, p-value = 0.51) (Figs. 2c and S5).

Effect of sampling on beta diversity. The use of 70% ethanol to sterilize sampling tools and equipment in this study may not entirely remove residual free DNA that could result in cross-contamination and increased similarity among samples. Future studies should consider incorporating a step using 50% bleach[24] or a commercial solution such as DNA AWAY surface decontaminant or ELIMINase. Despite this, we recovered clear clusters of site and layer groups across each sampling method (Fig. 3). In our analysis of individually collected samples (Fig. 3a), site and layer groups are clearly distinguished (NMDS: stress = 0.13, linear fit $R^2 = 0.92$). We did not detect any significant beta dispersion (ANOVA: sites p = 0.67, layers p = 0.18) or interactions between site and layer groups (Table 1). In our analysis including samples derived from manually pooling increasing numbers of samples (Fig. 3b) (NMDS: stress = 0.11, linear fit $R^2 = 0.99$), we found significant beta dispersion among soil layer groups (ANOVA: manually pooled samples p-value = 0.86, sites p-value = 0.19, layers p-value = 0.01). We did not detect any significant interactions between site, layer, or manually pooled samples (Table 1). In our analysis including samples processed with 1 or 3 DNA extractions (Fig. 3c) (NMDS: stress = 0.12, linear fit

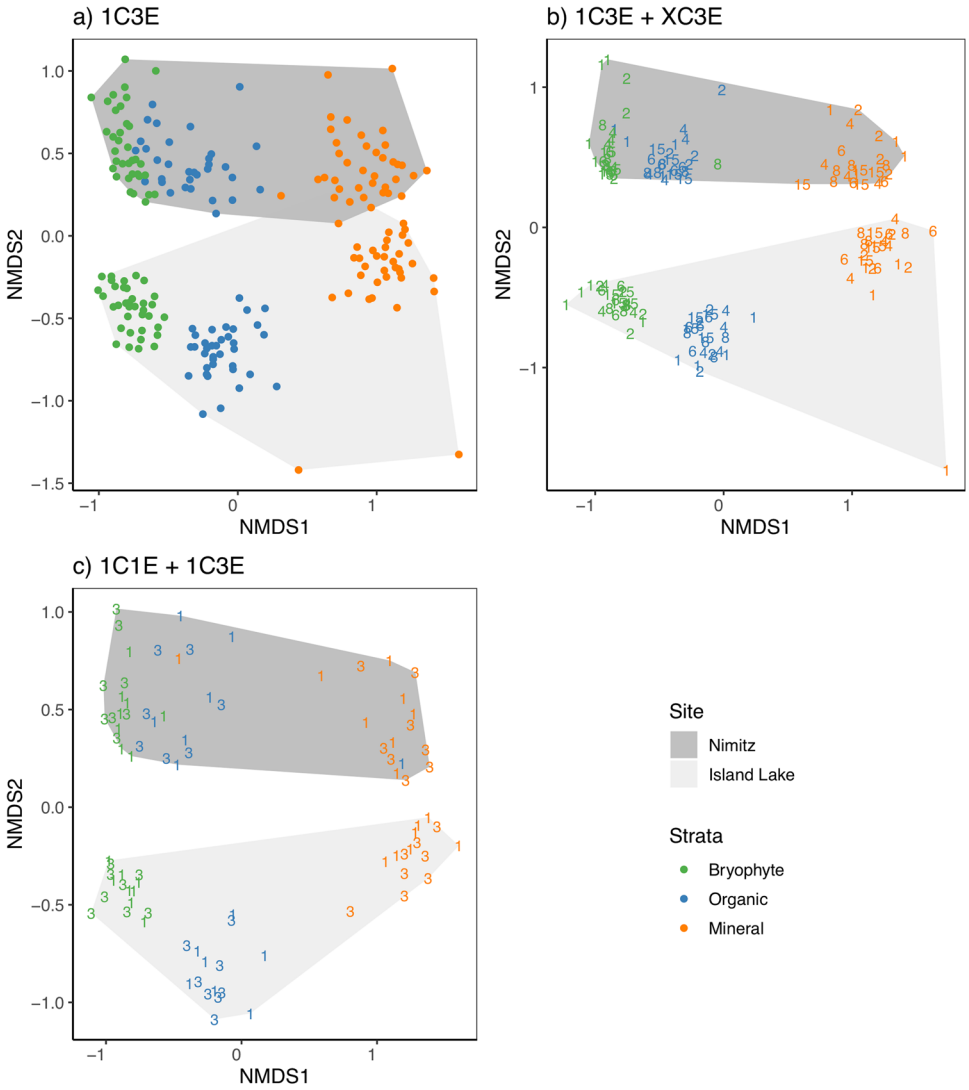

Figure 3. Clustering of samples across sites and soil layers are robust to intensity of field sampling and number of DNA extraction replicates. Clustered groups are shown based on (**a**), the collection of individually processed samples, (**b**) manual pooling of 1–15 samples with the number of pooled samples indicated by the plotted number, and (**c**) single samples processed with 1 or 3 pooled DNA extractions with the number of extractions indicated by the plotted number.

$R^2 = 0.93$), we found significant beta dispersion among layer groups (ANOVA: DNA extractions p-value = 0.88, sites p-value = 0.43, layers p-value = 0.03), and we did find a significant interaction among site, layer, and DNA extraction groups (Table 1). Although the NMDS plots show clear clustering across sites and layers, high residual R^2 values of 0.89, 0.72, and 0.79 from each PERMANOVA analysis suggests that there are additional environmental factors that explain our observed community dissimilarities (Table 1).

Assessing the stability of site indicator analyses. The higher level taxonomic composition of site indicator ESVs is similar across sites (Fig. 4). The fine level taxonomic composition of site indicator ESVs could not always be resolved to the species rank because of our inability to make high confidence taxonomic assignments. For improved readability, we plotted heat trees summarized to the species rank although site indicator analysis was conducted using ESVs. Where the same indicators appear to be detected from both sites, this is often due to our inability to confidently identify the ESVs. We found the variation of site indicators at the ESV level of resolution is quite variable across sampling methods (Fig. S6). Soil arthropods from both sites were comprised of mainly Arachnida (Scorpiones, Araneae, Sarcoptiformes), Insecta (Trichoptera, Hemiptera, Hymenoptera, Lepidoptera, Coleoptera, Diptera), Collembola (Entomobryomorpha), Malacostraca (Decapoda), and Diplopoda (Polydesmida). Many site indicator taxa are detected infrequently among samples. For the Island Lake site, site indicator ESVs from unknown Trombidiformes (plant parasitic mites), *Oppia nitens* (polyphagous fungiverous mite), *Eniochthonius crosbyi* (mite), unknown Plecoptera (stoneflies), Odonata (carnivorous dragonflies/damselflies), unknown Orthoptera (herbivorous grasshoppers/locus/crickets), Entomobryidae (omnivorous slender

Experiment	Permanova Formula	Group	F.model	R^2	P-value
Individual field samples	site * layer	site	18.9193	0.08243	0.001*
		layer	1.4108	0.01229	0.049
		site: layer	1.3864	0.01208	0.055
Manually pooled field samples	site * layer * expt	site	1.1129	0.00751	0.020*
		layer	5.2626	0.07103	0.001*
		expt	1.3034	0.04398	0.056
		site: layer	1.0515	0.01419	0.325
		site: expt	0.7989	0.02696	0.908
		layer: expt	0.9817	0.06625	0.512
		site: layer: expt	0.7113	0.04800	0.997
One versus three pooled DNA extractions	site * layer * expt	site	0.5036	0.00498	0.001*
		layer	1.5000	0.02969	0.047*
		expt	9.3499	0.09254	0.001*
		site: layer	1.0568	0.02092	0.371
		site: expt	0.4628	0.00458	0.998
		layer: expt	1.1707	0.02317	0.206
		site: layer: expt	1.6319	0.03230	0.030*

Table 1. Amount of variation of beta diversity explained by groups varies according to sampling method. The group 'site' refers to Island Lake or Nimitz field sites; 'layer' refers to bryophyte, organic, or mineral layers; 'expt' refers to variation in the sampling methods such number of individual or pooled samples or the number of pooled DNA extractions. The asterisk (*) indicates a p-value < 0.05.

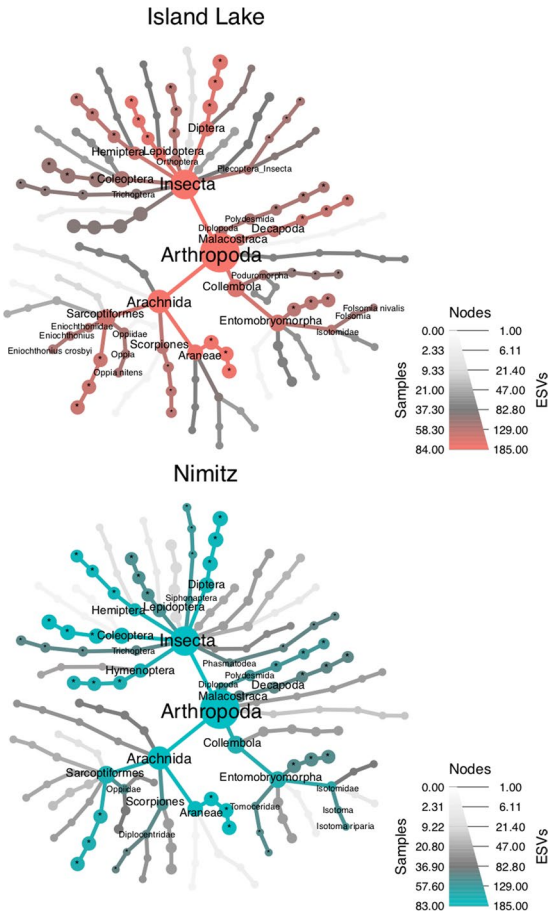

Figure 4. Taxonomic distribution of site indicator ESVs for each site. Heat trees comprised of all the site indicator ESVs, pooled across all sampling methods, are shown for each site. In each tree, color indicates the number of samples where each taxon was detected; text and node size indicate the number of site indicator ESVs in each taxon. To improve readability, labels have been added only to nodes present in at least half the samples. Taxa that could not be confidently identified are indicated by an asterisk (*).

springtails), *Folsomia nivalis* (elongate-bodied springtail), and unknown Poduromorpha (springtails) were found in more than half the samples. For the Nimitz site, indicator ESVs from unknown Siphonaptera (parasitic fleas), unknown Phasmatodea (herbivorous stick insects), and *Isotoma riparia* (springtail) were found in more than half the samples. We also illustrate how the taxonomic composition of site indicator ESVs varies slightly according to the number of manually pooled field samples, but in no consistent way (Fig. S7); and varies minimally according to the number of DNA extractions used (Fig. S8). However, we did find that taxonomic diversity of site indicator ESVs across soil layers was quite variable, with the majority of indicator ESVs recovered from the bryophyte and organic layers (Fig. S9).

Discussion

Our results highlight that the inclusion of replicate soil samples is critical to detect maximum richness of arthropods that have patchy spatial distributions. When we conducted richness calculations based on either ESVs or OTUs, results based on OTUs showed slightly lower richness but the trends were similar, i.e., increasing the amount of soil sampled resulted in a greater number of sequence clusters detected. For ease of bioinformatic reproducibility and comparability across studies, the use of exact sequence variants has been encouraged by others and was the method adopted in the current study[18]. Both ESVs and OTUs have been shown to perform similarly in biodiversity analyses when calculating richness and beta diversity[25]. When analyses at a certain taxonomic rank are needed, both ESVs and OTUs can be taxonomically assigned. Consistent with previous studies in bacteria and arthropods sampled from soil, sediments, and traps we also found that the comparison of beta diversity across sites is robust to variations in field sampling methods[26–28]. Changes in the number of pooled DNA extractions from the same sample also produced similar results with respect to richness and beta diversity. If resources are limited, a single DNA extraction per sample would be sufficient to process well-homogenized soil samples. Our results complement a previous study conducted across grassland, forest, and cropland sites where differences in sampling methods (conventional morphology, DNA metabarcodoing of bulk soil, and extracted arthropods) resulted in differences in the detection of individual taxa, but yielded similar site level diversity and composition[14]. Our results are also consistent with a previous simulation study that used an earthworm dataset to show how multiple samples from the same location sometimes recovered slightly different communities but multiple DNA extractions from the same sample accurately detected the target taxa[29]. With limited resources available, it would be more effective to put more effort into replicating sampling at the field site level, than it is to spend the time manually pooling field samples or performing replicate DNA extractions.

Richness, beta diversity, and indicator taxon analyses show differences across soil layers. The higher arthropod richness we observed in the bryophyte and organic layers is consistent with results from another Island Lake study that used phenotypic classification of Collembola (springtails) and Oribatida (mites)[3]. In the Rousseau *et al*. 2018 study, they showed higher density, biomass, and diversity of springtails and mites in moss and organic soils compared mineral soil. This has important implications with respect to sampling strategy and suggests that separating samples by soil horizon is a critical consideration for generating comparable samples between sites. In addition, this horizon separation supports the hypothesis that the moss layer is a critical resource for arthropods and its recovery after disturbance is likely necessary for a return of mature forest soil faunal communities[3]. In our study sites, minimal diversity would be missed if the mineral layer was not sampled, but future work should test this across a broader range of forest soils. These sampling considerations support the use of soil arthropod COI metabarcoding as a scalable method for biomonitoring.

We know that current COI reference databases such as BOLD and GenBank are not complete, but database representation has been shown to be improving year after year[30,31]. This limitation does have implications for studies working to benchmark DNA metabarcoding protocols against previous work based on commonly used bioindicator species. False negatives, taxa missed by DNA metabarcoding, can occur when local species have not yet been DNA barcoded and are missing from the reference sequence databases[14,20]. For example, when we compared the species list from the Rousseau *et al*. 2018 study also conducted at Island lake with the taxa present in the COI classifier v3, we found that 70% of their fully identified springtail and mite species (36% of genera) were missing from the reference database. This further highlights the importance of supplementing COI metabarcoding studies with local DNA barcoding to improve taxonomic assignment rates[32].

Site comparisons and the detection of site indicators using soil arthropod metabarcodes, however, can still be conducted whether or not the sequence clusters have been taxonomically assigned. In this study, we showed beta diversity comparisons using ordination and PERMANOVA that successfully distinguished samples among sites without using any of our taxonomic annotation data except for some upfront filtering of the dataset for arthropoda sequences. We also showed how site indicators based on exact sequence variants were successfully recovered even though taxonomic assignments to more inclusive levels of resolution appeared similar across sites. Our results are consistent with a previous study that showed how COI metabarcoding may actually recover a greater taxonomic diversity of site indicators in addition to the usual expected bioindicators[22]. Future studies should attempt to pair soil arthropod metabarcoding with local DNA barcoding to improve taxonomic assignment rates. Despite this, the use of soil arthropod metarbarcodes as site bioindicators was successful and samples from two similar jack pine stands with different origins were distinguished from each other based on beta diversity and the presence of site indicators.

Methods

Study area and field sample collection. Moss and soil samples were collected from 2 boreal forest stands in north-central Ontario that differ in origin on July 1, 2016. The first site is a 51-year old jack pine (*Pinus banksiana*) stand that was previously clearcut and located at the Island Lake Biomass Harvest Research and Demonstration area approximately 20 km from Chapleau, Ontario, Canada (47°42′N, 83°36′W)[33]. The second site was a 92-year old jack pine stand of wildfire origin (47°38′N, 83°15′W). Mean annual temperature

and precipitation for the area is 1.7 °C and 797 mm (532 mm of rainfall and 277 cm of snowfall), respectively (Environment Canada 2013). These two jack pine-dominated stands (>90% jack pine, based on live tree basal area) were established on glaciofluvial, coarse-textured, glacial outwash deposits characterized by sandy (medium sand) parent material overtopped with a variable depth loess (windblown) cap of finer textured soil (silty fine sand to silt loam)[34]. They both have a moderately dry soil moisture regime with rapid drainage. Forest floor depth (i.e. LFH – Litter, Fermented, Humic) was approximately 9–10 cm.

At each site we chose a 2 m × 2 m area of continuous moss cover (Fig. 1). Starting in the northwest corner, we used a bread knife to cut a 5 cm × 10 cm × full depth volume of moss and placed it in a labeled zip top bag. We then used a spoon to sample a 5 cm × 10 cm × full depth volume of the organic horizon (LFH) and placed it in a labeled zip top bag. Sampling tools were wiped to remove any visible soil, sterilized with 70% ethanol, then wiped with a clean cloth. The top 10 cm of the mineral horizon was sampled by hammering a fresh piece of 5 cm diameter × 10 cm long piece of polyvinyl chloride (PVC) pipe into the mineral horizon, extracting it, and placing it in a labeled zip- top bag. We repeated this procedure in a 6 × 6 grid (36 samples in total) with 20 cm spacing between each sample. In total we had 36 samples, with a subsample from each layer, at each site, for a total of 216 samples. Samples were immediately placed in a cooler with ice packs and were frozen at −20C within several hours of collection.

Sample preparation. The wet weight was obtained for each sample. Bryophyte and organic samples were separately homogenized using a knife mill and mineral samples were homogenized by forcing them through a 0.2 mm sieve (Fig. 1). The knife mill and sieve were both rinsed with water and then cleaned with 70% ethanol between samples. Samples from different soil layers were always kept separate. To thoroughly sample the soil arthropod community, samples were processed by subsampling 0.25 g of soil 3 times, extracting DNA from each replicate, and then pooling the DNA prior to PCR (1C3E method, 1 core, 3 DNA extractions). To assess the influence of using samples drawn from increased spatial sampling of soil, we subsampled 1 g of soil from each homogenized bryophyte and organic sample and 5 g of mineral soil to create composites drawn from each of 2, 4, 6, 8, and 15 samples (keeping layers separate). Each composite sample was put into a zip-top bag and shaken by hand to mix. This pooling was replicated 4 times with different samples represented in each pool (XC3E method, 2–15 pooled samples, each with 3 DNA extractions). For the Nimitz site, there was not always enough bryophyte sample so the largest pool had anywhere from 9 to 14 samples included. We used 0.25 g from each pooled soil sample for triplicate DNA extractions that were pooled prior to PCR. To assess the value of pooling multiple DNA extractions per sample, 0.25 g from each of 8 un-pooled soil samples ×3 soil layers were extracted one time only prior to PCR (1C1E method, 1 core, 1 DNA extraction).

Molecular biology methods. DNA extraction was carried out using the DNeasy PowerSoil Kit (Qiagen Cat# 12888-100) modified with the Braid *et al.* (2003) protocol that uses a chemical flocculant to help remove soil-derived PCR inhibitors[35]. We extracted DNA from 0.25 g of soil per sample following the manufacturer's protocol except that 200 µl of 100 mM aluminum ammonium sulfate dodecahydrate was added to the tube with 60 µl of solution C1 followed by a 10 minute incubation at 70 °C to help lyse difficult samples.

Mixed template PCR and Illumina library preparation was carried out at the Canadian Forest Service's Laurentian Forestry Centre. DNA was quantified using the Qubit dsDNA HS Assay Kit (Life Technologies, Burlington, ON, Canada). DNA concentrations were standardized to 5 ng/µl for all samples and each sample was amplified in triplicates to ensure reproducibility[36,37]. Invertebrate communities were targeted using two sets of primers targeting the COI gene (5′->3′): the F230R_modN marker with the forward primer LCO1490 GGTCAACAAATCATAAAGATATTGG and the reverse primer 230R_modN CTTATRTTRTTTATNCGNGGRAANGC adapted from the Gibson *et al.* 2014 230R primer to include N's instead of inosines[21,38]; and the BE marker with the B forward primer CCIGAYATRGCITTYCCICG and the E reverse primer GTRATIGCICCIGCIARIAC[19]. These primers were combined with the required Illumina adaptors at the 5′ end of the primer sequences, TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG for the forward primer and GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG for the reverse primer. PCR reactions were set up by creating a master mix of 37.5 µl of HotStarTaq Plus Master Mix (QIAGEN Inc., Germantown, MD, USA), 1.5 µl of each 10 µM primer, 27 µl of UltraPure DNase/RNase-Free Distilled Water (GIBCO, Life Technologies) and 7.5 µl of gDNA at 5 ng/µl. The final volume of 75 µL was then distributed in three 96-well plates placed in separate thermocyclers. Thermal cycling conditions were as follows: initial denaturation at 95 °C for 5 min; 40 cycles at 94 °C for 30 s, 50 °C for 30 s, 72 °C for 1 min; and a final elongation at 72 °C for 10 min. Triplicates PCR products were pooled and visualized on GelRed-stained 1% agarose gels using the ChemiGenius Bioimaging System (Syngene, Cambridge, UK). PCR products were purified using 81 µl of magnetic bead solution (Agencourt AMPure XP, Beckman Coulter Life Science, Indianapolis, IN, USA) according to Illumina's protocol[39]. Indexes were added to each sample by amplifying 5 µl of the purified PCR product with 25 µl of KAPA HIFI HotStart Ready Mix, 5 µl of each Nextera XT Index Primer (Illumina Inc., San Diego, CA, USA) and 10 µl of UltraPure DNase/RNase-Free Distilled Water for a total volume of 50 µl. Thermal cycling conditions were as follows: 3 min at 98 °C, 8 cycles of 30 sec at 98 °C, 30 sec at 55 °C, 30 sec at 72 °C, and a final elongation step of 5 min at 72 °C. Indexed amplicons were purified with the magnetic beads as previously described, quantified using a Qubit dsDNA BR Assay Kit (Life Technologies) and combined at equimolar concentration. Paired-end sequencing (2 × 250 bp) of the pools was carried out on an Illumina MiSeq at the National Research Council Canada, Saskatoon. 15% PhiX was added to help compensate for low sequence heterogeneity on the plate.

Bioinformatic methods. Reads were processed using the SCVUC v2.0 bioinformatic pipeline available from GitHub at https://github.com/Hajibabaei-Lab/SCVUC_COI_metabarcode_pipeline. SCVUC is an acronym that stands for the major programs/algorithms used in the pipeline: "S" SEQPREP, "C" CUTADAPT, "V"

VSEARCH, "U" USEARCH-unoise, "C" COI Classifier. At certain points commands were run in parallel using GNU parallel[40]. First, the compressed fastq raw reads were paired with SEQPREP using the default parameters except that we required a minimum Phred score of 20 in the overlap region and a minimum overlap of at least 25 bp. Primer sequences were trimmed with CUTADAPT v1.10 with the default settings (maximum of 10% mismatched bases between the matching primer and sequence region) except that we required a minimum length (after trimming) of at least 150 bp, a minimum Phred score of 20 at the ends, allowing a maximum of 3 Ns. CUTADAPT was also used to convert the compressed fastq files to compressed FASTA files. We added the sample name to the FASTA headers and concatenated all the sequences into a single file to permit the generation of global ESVs below. Sequences were dereplicated with VSEARCH v2.5.0 with the–derep_fulllength command, sequences comprised of identical substrings are retained as unique sequences, and the number of reads in each cluster were tracked with the–sizein–sizeout commands. Unique sequences were denoised and a set of ESVs were generated with USEARCH v10.0.240 with the unoise3 algorithm. With this method, sequences with predicted errors are removed, putative PhiX contamination is removed, putative chimeric sequences are removed, and rare ESVs are removed. We defined rare ESVs as clusters containing only one or two reads (singletons and doubletons) because it has been shown that rare clusters tend to be predominantly comprised of reads with sequence errors[41,42]. In total, 41% of primer-trimmed reads belonging to rare ESVs were removed after denoising. To compensate for a known bug in this version of the program, we changed the 'Zotu' prefix in the FASTA file headers to 'Otu'. At each major step of bioinformatic processing above, statistics including read/cluster number and read length (min, max, mean, median, mode) were calculated. We used VSEARCH to construct the ESV x sample table that tracks read numbers in the ESVs. This was done by mapping good quality primer-trimmed reads to the denoised ESVs with 100% sequence similarity. At this step, shorter sequence substrings may be mapped to longer ESVs. The denoised ESVs were taxonomically assigned with the COI Classifier v3 available from https://github.com/terrimporter/CO1Classifier. Read number and samples were mapped to the taxonomic assignment table. We identified high confidence taxonomic assignments using the recommended minimum bootstrap cutoff values for 200 bp fragments (species $>=0.70$, genus $>=0.30$, family $>=0.20$). Assuming that our taxa are in the reference database, then taxonomic assignments should be at least 99% correct (95% correct for species).

To assess the stability of results at varying levels of resolution, we compared results based on ESVs and OTUs (operational taxonomic units). Denoised ESVs from above were fed into the '–cluster_smallmem' command in VSEARCH and OTU clusters based on 97% sequence similarity were generated. These results were then processed as described above for ESVs, except that '–id 0.97' was used to map reads to the sample x OTU table.

Data analysis. The BE and F230R_modN taxonomy tables were prepared at the command line, with Perl, and analyzed in R v3.4.3 with scripts available from GitHub at https://github.com/terrimporter/PorterEtAl2019grid[43]. The 'vegan' v2.4–6 package in R was used to plot rarefaction curves using the 'rarecurve' function[44]. Rarefaction to the 15th percentile was performed in vegan with the 'rrarefy' function. This was done to minimize library size bias in diversity comparisons[45]. We then transformed read abundances to presence-absence data. We did this because PCR primer bias may distort template to PCR product ratios making read number unsuitable for inferring quantitative differences in biomass, density, or community composition[46–48].

We assessed the effect of increasing spatial sampling by including more individual samples. We simulated sampling increasing numbers of individual samples by randomly bioinformatically pooling data from 1–15 individually collected samples and replicated this sampling 4 times. We calculated richness for each level of sampling effort using the 'specnumber' function in vegan. We calculated venn diagrams using the 'vennCounts' function in the limma Bioconductor package and plotted this using the ggforce package to draw circles[49,50]. We assessed the effect of sampling effort on beta diversity using non-metric multi-dimensional scaling (NMDS) ordination ($n=211$). The NMDS plot was created with the 'metaMDS' function in vegan with 2 dimensions using Bray-Curtis dissimilarity with binary data (Sorensen dissimilarity). The number of dimensions was chosen by calculating a scree plot using the 'dimcheckMDS' function in the goeveg package (not shown)[51]. A Shephard diagram and goodness of fit calculations were created using the 'stressplot' and 'goodness' functions in vegan. Beta dispersion was assessed using the 'betadisper' function in vegan. We tested for significant interacting factors with permutational multivariate analysis of variance (PERMANOVA) using the 'adonis' function in vegan with the strata option so randomizations occur within sites.

We also assessed the effect of increasing spatial sampling by manually pooling increasing numbers of samples. For a balanced design, we randomly subsampled samples from the 1C3E method down to 4 replicates to match the number of replicates available for the XC3E method. We calculated richness for each level of sampling effort as described above. Beta diversity was assessed as described above ($n=144$), a single outlier was identified, removed, then the analysis was re-run.

We assessed the effect of performing mixed template PCRs on single or pooled triplicate DNA extractions. For a balanced design, we subsampled from the 1C3E experiment to match the same 8 grid coordinates as used in the 1C1E experiment. We calculated richness and beta diversity as described above. We calculated the difference in richness from the same samples processed using one or three DNA extractions, checked for normality using Shapiro-Wilk's test for normality and a quantile-quantile plot, and tested for significant differences in richness across samples using pairwise Wilcox tests and adjusting p-values for multiple comparisons using the Benjamini & Hochberg (1995) method[52,53]. Beta diversity was assessed as described above ($n=94$), two outliers were identified, removed, then the analysis was re-run.

Site indicator ESVS were determined using the 'indicspecies' v1.7.6 package in R using the multipatt command[54]. Briefly, the indicator species concept describes the species associated with a certain site or condition based on their fidelity to those conditions and absence from others. This concept can be extended to the ESV or OTU rank when current reference databases do not allow us to identify all sequences to the species rank. To create a balanced design, a subsample from the same 4 grid coordinates from the 1C1E and 1C3E methods was used

to compare with the 4 replicates available from the XC3E method. For each sampling method, site indicators were retained if they had a p-value $<= 0.05$. To illustrate the taxonomic distribution of site indicators and their prevalence across samples, the 'metacoder' v0.3.0 package in R was used to create heat trees[55]. An ESV × sample matrices enumerating the reads recovered for site indicator ESVs using each method were formatted in R to resemble QIIME output. From this, a sample matrix was constructed in R. To improve clarity, we reduced the number of edges in the heat trees by summarizing ESV taxonomic assignments to the species rank (instead of the ESV rank). The taxonomic information was parsed using the parse_tax_data command from the 'tax' v0.3.1 package in R[56]. Taxon abundance at all ranks was calculated with the calc_taxon_abund command. Taxon occurrence per sample group was calculated with the calc_n_samples command.

Data availability

Raw reads are available from the NCBI Short Read Archive SRA under BioProject PRJNA565010, BioSamples SAMN12257424–SAMN12257429. The SCVUC v2.0 bioinformatic pipeline is available from GitHub at https://github.com/Hajibabaei-Lab/SCVUC_COI_metabarcode_pipeline. A FASTA file of final ESVs, taxonomy table, and scripts used to produce figures are available from GitHub at https://github.com/terrimporter/PorterEtAl2019grid.

Received: 9 July 2019; Accepted: 13 November 2019;
Published online: 03 December 2019

References

1. Neher, D. A., Weicht, T. R. & Barbercheck, M. E. Linking invertebrate communities to decomposition rate and nitrogen availability in pine forest soils. *Applied Soil Ecology* **54**, 14–23 (2012).
2. Maab, S., Caruso, T. & Rillig, M. C. Functional role of microarthropods in soil aggregation. *Pedobiologia* **58**, 59–63 (2015).
3. Rousseau, L. *et al*. Forest floor mesofauna communities respond to a gradient of biomass removal and soil disturbance in a boreal jack pine (Pinus banksiana) stand of northeastern Ontario (Canada). *Forest Ecology and Management* **407**, 155–165 (2018).
4. Bird, G. A. & Chatarpaul, L. Effect of whole-tree and conventional forest harvest on soil microarthropods. *Canadian Journal of Zoology* **64**, 1986–1993 (1986).
5. Battigelli, J. P., Spence, J. R., Langor, D. W. & Berch, S. M. Short-term impact of forest soil compaction and organic matter removal on soil mesofauna density and oribatid mite diversity. *Canadian Journal of Forest Research* **34**, 1136–1149 (2004).
6. Addison, J. A. & Barber, K. N. *Response of Soil Invertebrates to Clear-cutting and Partial Cutting in a Boreal Mixedwood Forest in Northern Ontario* (1997).
7. Rousseau, L. *et al*. Long-term effects of biomass removal on soil mesofaunal communities in northeastern Ontario (Canada) jack pine (Pinus banksiana) stands. *Forest Ecology and Management* **421**, 72–83 (2018).
8. Porter, T. M. & Hajibabaei, M. Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Molecular Ecology* **27**, 313–338 (2018).
9. Ji, Y. *et al*. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters* **16**, 1245–1257 (2013).
10. Arribas, P., Andújar, C., Hopkins, K., Shepherd, M. & Vogler, A. P. Metabarcoding and mitochondrial metagenomics of endogean arthropods to unveil the mesofauna of the soil. *Methods Ecol Evol* **7**, 1071–1081 (2016).
11. Andújar, C. *et al*. Phylogenetic community ecology of soil biodiversity using mitochondrial metagenomics. *Mol Ecol* **24**, 3603–3617 (2015).
12. Yang, C. *et al*. 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 (2014).
13. Epp, L. S. *et al*. New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems: Metabarcodes to Analyse Soil DNA. *Molecular Ecology* **21**, 1821–1833 (2012).
14. Oliverio, A. M., Gan, H., Wickings, K. & Fierer, N. A DNA metabarcoding approach to characterize soil arthropod communities. *Soil Biology and Biochemistry* **125**, 37–43 (2018).
15. Watts, C. *et al*. DNA metabarcoding as a tool for invertebrate community monitoring: a case study comparison with conventional techniques: Monitoring invertebrates using DNA metabarcoding. *Austral Entomology*, https://doi.org/10.1111/aen.12384 (2019).
16. Hebert, P. D. N., Cywinska, A., Ball, S. L. & deWaard, J. R. Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences* **270**, 313–321 (2003).
17. Hajibabaei, M., Baird, D. J., Fahner, N. A., Beiko, R. & Golding, G. B. A new way to contemplate Darwin's tangled bank: how DNA barcodes are reconnecting biodiversity science and biomonitoring. *Phil. Trans. R. Soc. B* **371**, 20150330 (2016).
18. Callahan, B. J., McMurdie, P. J. & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal* **11**, 2639–2643 (2017).
19. Hajibabaei, M., Spall, J. L., Shokralla, S. & van Konynenburg, S. Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. *BMC Ecology* **12**, 28 (2012).
20. Emilson, C. E. *et al*. DNA metabarcoding and morphological macroinvertebrate metrics reveal the same changes in boreal watersheds across an environmental gradient. *Scientific Reports* **7** (2017).
21. Gibson, J. *et al*. Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. *PNAS* **111**, 8007–8012 (2014).
22. Hajibabaei, M., Porter, T. M., Wright, M. & Rudar, J. COI metabarcoding primer choice affects richness and recovery of indicator taxa in freshwater systems. *PLoS One* **14**, e0220953 (2019).
23. Hajibabaei, M. *et al. Watered-down biodiversity? A comparison of metabarcoding results from DNA extracted from matched water and bulk tissue biomonitoring samples*, https://doi.org/10.1101/575928 (2019).
24. Goldberg, C. S. *et al*. Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods Ecol Evol* **7**, 1299–1307 (2016).
25. Glassman, S. I. & Martiny, J. B. Ecological patterns are robust to use of exact sequence variants versus operational taxonomic units. *mSphere* **3**, e00148–18 (2018).
26. Haegeman, B. *et al*. Robust estimation of microbial diversity in theory and in practice. *ISME J* **7**, 1092–1101 (2013).
27. Drummond, A. J. *et al*. Evaluating a multigene environmental DNA approach for biodiversity assessment. *GigaSci* **4**, 46 (2015).
28. Grey, E. K. *et al*. Effects of sampling effort on biodiversity patterns estimated from environmental DNA metabarcoding surveys. *Scientific Reports* **8** (2018).
29. Ficetola, G. F. *et al*. Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources* **15**, 543–556 (2015).
30. Porter, T. M. & Hajibabaei, M. Over 2.5 million COI sequences in GenBank and growing. *PLoS One* **13**, e0200177 (2018).
31. Ratnasingham, S. & Hebert, P. D. BOLD: The Barcode of Life Data System. http://www.barcodinglife.org. *Molecular ecology notes* **7**, 355–364 (2007).

32. Hoage, J. F. J. *Metabarcoding soil microarthropods for soil quality assessment: Importance of integrated taxonomy, phylogenetic marker selection and sampling design*. (Laurentian University, 2018).
33. Kwiaton, M. *et al*. Island Lake Biomass Harvest Research and Demonstra on Area: Establishment Report. **82** (2014).
34. Venier, L. A. *et al*. Ground-dwelling arthropod response to fire and clearcutting in jack pine: implications for ecosystem management. *Canadian Journal of Forest Research* **47**, 1614–1631 (2017).
35. Braid, M. D., Daniels, L. M. & Kitts, C. L. Removal of PCR inhibitors from soil DNA by chemical flocculation. *Journal of Microbiological Methods* **52**, 389–393 (2003).
36. Schmidt, P.-A. *et al*. Illumina metabarcoding of a soil fungal community. *Soil Biology and Biochemistry* **65**, 128–132 (2013).
37. Kennedy, K., Hall, M. W., Lynch, M. D. J., Moreno-Hagelsieb, G. & Neufeld, J. D. Evaluating Bias of Illumina-Based Bacterial 16S rRNA Gene Profiles. *Applied and Environmental Microbiology* **80**, 5717–5722 (2014).
38. Folmer, O., Black, M., Hoeh, W., Lutz, R. & Vrijenhoek, R. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular marine biology and biotechnology* **3**, 294–299 (1994).
39. Illumina. 16S metagenomic sequencing library preparation - Preparing 16S ribosomal RNA gene amplicons for the Illumina MiSeq System, https://support.illumina.com/downloads/16s_metagenomic_sequencing_library_preparation.html (2013).
40. Tange, O. GNU Parallel - The Command-Line Power Tool.; *login: The USENIX Magazine* February, 42–47 (2011).
41. Tedersoo, L. *et al*. 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New Phytologist* **188**, 291–301 (2010).
42. Kunin, V., Engelbrektson, A., Ochman, H. & Hugenholtz, P. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology* **12**, 118–123 (2010).
43. R Core Team. R: A Language and Environment for Statistical Computing, https://www.R-project.org/ (2017).
44. Oksanen, J. *et al*. Vegan: Community Ecology Package. R package version 2.5–2, https://CRAN.R-project.org/package=vegan (2018).
45. Weiss, S. *et al*. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**, 27 (2017).
46. Polz, M. F. & Cavanaugh, C. M. Bias in template-to-product ratios in multitemplate PCR. *Applied and environmental Microbiology* **64**, 3724–3730 (1998).
47. Suzuki, M. T. & Giovannoni, S. J. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and environmental microbiology* **62**, 625–630 (1996).
48. Elbrecht, V. & Leese, F. Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer Bias and Biomass—Sequence Relationships with an Innovative Metabarcoding Protocol. *PLOS One* **10**, e0130324 (2015).
49. Ritchie, M. E. *et al*. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43**, e47–e47 (2015).
50. Pedersen, T. L. Ggforce: Accelerating 'ggplot2', https://CRAN.R-project.org/package=ggforce (2019).
51. Goral, F. & Schellenberg, J. Goeveg: Functions for Community Data and Ordinations, https://CRAN.R-project.org/package=goeveg (2018).
52. Shapiro, S. S. & Wilk, M. B. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* **52**, 591–611 (1965).
53. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* **57**, 289–300 (1995).
54. De Cáceres, M. & Legendre, P. Associations between species and groups of sites: indices and statistical inference. *Ecology* **90**, 3566–3574 (2009).
55. Foster, Z. S. L., Sharpton, T. J. & Grünwald, N. J. Metacoder: An R package for visualization and manipulation of community taxonomic diversity data. *Plos Computational Biology* **13**, e1005404 (2017).
56. Chamberlain, S. & Foster, Z. *Taxa: Taxonomic Classes* (2018).

Acknowledgements

T.P. and L.V. would like to acknowledge funding from the Canadian government through the Genomics Research and Development Initiative (GRDI), Metagenomics-based Ecosystem Biomonitoring (Ecobiomics) project. We would like to acknowledge Rob Fleming and Paul Hazlett at the Canadian Forest Service's Great Lakes Forestry Centre for the development of the study area. We also appreciate the contribution by Christine Martineau and Marie-Josée Morency at the Canadian Forest Service's Laurentian Forestry Centre for the metabarcoding and Illumina library preparation.

Author contributions

L.V. conceived of the idea, designed the experiment, and helped collect samples. K.-W.K. collected and curated samples. D.D. provided laboratory resources and advice for DNA extractions and PCRs. D.C. and S.B. did the DNA extractions and PCRs. A.S. provided library preparation resources. M.H. provided bioinformatic resources. T.P. conducted the bioinformatic processing, data analyses, and wrote the manuscript. L.V. and C.E. provided critical guidance on figure generation and the manuscript. N.B., D.M.M. and E.E. provided input on developing objectives and study design. All authors edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-54532-0.

Correspondence and requests for materials should be addressed to T.M.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.