

Short Note

Ecological monitoring using Collembola metabarcoding with extremely low bycatch amplification

Pedro M. Pedro^{a,b,*}, Laury Cullen Jr.^a, Fabiana Prado^a, Alexandre Uezu^a, Ross Piper^c, Christiana M.A. Faria^d, Christoph Knogge^b, Maria Tereza Pepe Razzolini^e, Marcela B. Paiva^b, Milena Dropa^e, Miriam Silva^e, Tatiane Cristina Rech^f, Thomas Püttker^g

^a IPE-Instituto de Pesquisas Ecológicas, Rod. Dom Pedro I, km 47 Caixa Postal 47 12960-000 Nazaré, Paulista, SP, Brazil

^b Squared Biomonitoring, Rod. Dom Pedro I, km 47 Nazaré Paulista, SP 12960-000, Brazil

^c The Faculty of Biological Sciences, University of Leeds, United Kingdom

^d Centro de Ciências, UFC, Av. Mister Hull, s/n, Fortaleza, CE 60440-900, Brazil

^e Faculdade de Saúde Pública – USP, Av. Dr. Arnaldo, 715, São Paulo, SP 01246-904, Brazil

^f Instituto Libio de Proteção à Natureza Rua Joaquim Floriano, 1052 São Paulo, SP 04534-004, Brazil

^g Departamento de Ciências Ambientais, Universidade Federal de São Paulo Rua, São Nicolau 210, 09913-030 Diadema, SP, Brazil

ARTICLE INFO

Keywords:

Mixed bulk samples
Soil ecology
Bioindicator
Pitfall trap

ABSTRACT

Collembola are used widely to monitor soil health and functional parameters. Recent developments in high throughput sequencing (especially metabarcoding) have substantially increased their potential for these ends. Collembola are especially amenable to metabarcoding because of their small size, high abundance, and ubiquity in most habitat types. However, most Collembola sampling protocols collect a substantial and highly varied bycatch that can be a considerable impediment to metabarcoding, especially because of data lost to non-target species. We designed a primer set amplifying the D2 expansion segment of ribosomal DNA that is highly conserved across Collembola and successfully excludes from amplification nearly all other invertebrate taxa. We tested the diagnostic power of the primer set by clearly distinguishing Collembola communities between forest sites with differing habitat qualities in São Paulo State, Brazil. The oligos amplified all Collembola orders previously encountered in the sampling locations, with no non-target amplification, and also excluded the closely related Protura and Diplura. Alpha diversity (OTU count) and phylogenetic diversity was significantly higher in high quality habitats. Moreover, the beta diversity indices successfully differentiated high and low-quality habitats. This new addition to the biomonitoring toolbox greatly increases the accessibility of Collembola metabarcoding for various types of habitat assessments.

1. Introduction

Collembola (springtails) are important invertebrate indicators of soil ecological parameters (Bispo et al., 2009; Potapov et al., 2020) and are organisms particularly amenable to field biomonitoring initiatives (Breure et al., 2003; Fiera, 2009; de Filho et al., 2016; Zeppelini et al., 2009). Collembola are a statistically appealing taxon because diversity calculations can be based on high sample sizes per collection, whilst often maintaining a low sampling effort. In tropical soils, for example, Collembola can comprise over 60,000 individuals per m² and are often represented by dozens of species (Basset et al., 2022; Culik et al., 2002).

The taxonomic bottleneck currently limits Collembolan's full

potential as an efficient soil biological indicator, particularly in understudied regions such as South America. Morphological identification is particularly difficult because of the small size of most taxa and a dearth of diagnostic morphological traits resulting in high occurrence of cryptic species (Porco et al., 2012; Sun et al., 2017). However, a growing DNA barcode reference library has largely made this “impediment” moot. Genetic markers can now provide effective identification of many species, regardless of their life stage and/or specimen integrity (Beng et al., 2016; Eaton et al., 2017).

Recently, metabarcoding pipelines have greatly broadened springtails' potential as biological indicators (Saitoh et al., 2016). Metazoan metabarcoding has built upon the advantages of single-specimen

* Corresponding author.

E-mail address: pedro@ipe.org.br (P.M. Pedro).

<https://doi.org/10.1016/j.ecolind.2023.111202>

Received 2 June 2023; Received in revised form 27 October 2023; Accepted 30 October 2023

Available online 6 November 2023

1470-160X/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

barcoding to simultaneously identify potentially thousands of individuals per sample (Yu et al., 2012; Ji et al., 2013). However, the technique has one significant drawback that can result in substantial analytical inefficiencies, particularly when using non-selective field sampling protocols: a significant portion of the DNA sequences resulting from bulk samples can comprise non-target DNA, which occupies space in sequencing flow cells and increases costs. To circumvent this issue, researchers can use primers targeting focal taxa, such that little or no *a priori* hand-sorting is required prior to DNA extraction (Brown et al., 2016; Ficetola et al., 2008; Pedro et al., 2020).

Here we present a primer set designed to amplify an approximately 440-bp fragment of the *Collembola* D2 expansion segment of the 28S operon without amplifying significant non-*Collembola* bycatch (validated *in silico* and *in vitro*). The primers' exclusivity to *Collembola* was benchmarked by sequencing the entire contents of field-set pitfall traps, with minimal or no sorting.

This primer set and the associated metabarcoding analytical pipeline were tested in conjunction with pitfall sampling to assess differences in *Collembola* species composition between mature forests and incipient reforestation plots. Considering many *Collembola* species are detritivores specialized on varying types of leaf litter (e.g., Fujii and Takeda, 2012), we hypothesised that these two successional stages would be reflected by metabarcoding. Furthermore, we tested biogeographic patterns in locations ~400-km apart to evaluate the regional distribution of *Collembola* within the Atlantic Forest, a geographical scale that has never been explicitly tested within a single Neotropical biome.

2. Material and methods

2.1. Primer design

Ribosomal loci are used in metabarcoding initiatives because they are taxonomically informative and, because of their non-protein-coding nature, often allow for the design of primers fully conserved to the target region (i.e., with no degeneracy at third codon positions; Burki et al., 2021; Semmouri et al., 2021). There is expected to be minimal PCR-bias when using these types of primers and rDNA is therefore an attractive option when estimates of relative proportions of taxa are needed (e.g., Pedro et al., 2020). Here, we target the D2 domain of the 28S operon of nuclear ribosomal DNA (rDNA), which has been extensively used in species diagnosis, both pre- and post-metabarcoding (Campbell et al., 1994; Dodd et al., 2000; Pedro et al., 2021).

We initially downloaded all available GenBank accessions with keyword "28S" for *Collembola* and non-*Collembola* arthropods that contained either of the conserved D2 flanks using default MegaBLAST parameters (as per GenBank accession JX261730 for *Heteromurus* sp.). Sequence hits flanking the forward annealing region totalled 1,492 for *Collembola* and 42,129 for non-*Collembolan* arthropods. Reverse sequences were 562 and 38,200, respectively. Non-arthropods were not considered, as their D2 flanks were considerably diverged from the target ingroup and thus unlikely to compete with our target *Collembola* templates.

We evaluated only BLAST results with full binomial identification (i.e., genus and species) and filtered results to exclude hits where putative priming sites were less than 20 nucleotides from the sequence termini (in order to omit those unwittingly submitted to GenBank with the original amplification oligo sequences still included). The filtered *Collembola* D2 sequences comprised 13 families and represented all four currently recognized *Collembola* orders (Entomobryomorpha, Poduromorpha, Symphypleona and Neelipleona).

Forward and reverse primers were designed that amplified the polymorphic region of D2 and were conserved among all available filtered *Collembola* GenBank sequences (576 and 166, respectively) with little or no degeneracy. These were designed so that the non-*Collembola* sequences did not possess the matching 3' nucleotide in either primer. The result of these comparisons was *Collembola-F* (5'-

AGAGAGTTMAAWAGTACGTGAAACCT-3') and *Collembola-R* (5'-TGTTTTCAAGACGGGACAGGC-3').

We graphically confirmed the binding site variation and taxonomic resolution of the *Collembola-F* and *Collembola-R* primers designed above against all *Collembola* GenBank entries possessing both forward and reverse priming locations using the *ecopcr* module of Obitools (v1.2.13) (Boyer et al., 2016). We undertook an analogous comparative evaluation using two primer sets previously used in *Collembola* metabarcoding, 16S and CO1 (Saitoh et al., 2016), to assess the appropriateness of all currently available oligos. Our parameters for *ecopcr* allowed for a maximum of four mismatches in either primer and sequences had to possess priming sites at least 20 nucleotides from their termini. Expected *ecopcr* amplicon length in arthropods for D2 was 200–600 bp, for CO1 was 260–280 bp and for 16S 300–600 bp.

We also tested the fidelity of D2, 16S and CO1 primers to three non-*Collembola* taxa commonly found during pitfall and funnel sampling methods (based on our previous experience in the locations sampled): Acariformes, Coleoptera and Hymenoptera (principally ants). Here, we sought to assess each primer set's cross-amplification in these non-target groups.

The resulting *ecopcr* output was analysed and graphed in R (R Development Core Team 3.0.1, 2013) using the ROBITools package (<https://metabarcoding.org/obitools>).

2.2. Pitfall sampling

In order to test the applicability of the D2 primer sets in analyzing richness and composition of *Collembola* communities, we used them on invertebrate samples obtained by pitfall sampling in two locations in São Paulo State, Brazil. The first location, in Ubarana municipality (approximately 21°14'09" S, 49°43'12" W), consisted of a forest remnant adjacent to a hydroelectric reservoir, within which invertebrates were sampled at 13 sampling points. The second location was located near Nazaré Paulista (23°12'48" S, 46°21'58" W) where 15 points were sampled. In both locations, sampling points were distributed within forest classified as high-quality as well as low-quality (Supplemental Data 1). The latter category comprised areas of reforestation initiatives in initial stages of succession.

Sampling was undertaken with "mini" pitfall traps improvised from 50-ml Falcon tubes half-filled with absolute ethanol and protected from rain and detritus with a plastic plate. The lip of the tube was fitted with an improvised inverted plastic funnel to limit ethanol evaporation and fitted with a 5-mm screen to prevent entry of taxa larger than ~2 mm. Three such pitfalls were used per sampling location, each placed within a radius of approximately 3 m. Pitfall traps were deployed for approximately 30 days beginning in September 2015 in Ubarana) and July 2016 in Nazaré Paulista (Supplemental Data 1).

Although all pitfall traps contained much more bycatch DNA (from Diptera, Lepidoptera, Coleoptera and Hymenoptera (primarily ants)) than *Collembola* tissue, we nonetheless limited sorting effort to the rare cases of twigs, soil or very large beetles present amongst the catch.

2.3. DNA extraction, PCR and sequencing

The contents of pitfall traps were returned to the laboratory and stored at –20 °C for not more than 1 week prior to DNA extraction. The 50-ml tubes used as pitfalls contained substantial amounts of bycatch, dominated by soil mites, small beetles and ants. In some cases, the mesh used to prevent entry to larger organisms had been torn and larger items were occasionally sampled.

We decanted the contents of the 50-ml tubes into weighing boats and then removed bycatch that was cumbersome to the downstream protocol, such as very large-carapaced insects and sources of PCR inhibitors (twigs, leaves, soil), but no further sorting was done, as taxon-specific primers were used. Weighing boat contents (including targeted *Collembola* and bycatch) were transferred to 2-ml Eppendorf tubes, dried

overnight on silica gel, and macerated using a Savant FastPrep lysis mill at maximum speed for 20-s using 1-mm ceramic beads (when necessary, large samples were divided into multiple tubes that were re-pooled following maceration).

A subsample of the maceration product was then submitted to DNA extraction using a DNeasy Blood & Tissue Kit (Qiagen, Valencia, CA) following manufacturer's instructions for insects.

The working stock DNA from each sample extraction was diluted to 100 ng/μl. Nested PCR reactions were performed using the Collembola-specific D2 primer set described above. Samples from different collection sites were tagged with multiplex identifiers (MIDs) to allow combined 454 FLX sequencing. In summary: a first PCR was done using the forward primer *Collembola-F adF* (5'-GGCCACGCGCTCGACTAGTACAGAGAGTTMAAWAGTACGTGAAACCT-3'), where the underlined portion is an adaptor overhang used to decrease the cost of multiplexing PCR primers. The reverse primer for the first PCR was *Collembola-R* (5'-TGTTTCAAGACGGGACAGGC-3').

The product from the initial reaction was diluted 10x, purified and submitted to a second PCR using the forward primer *454A-MID-adF* (5'-CGTATCGCCTCCCTCGCGCATCAG NNNNNN GGCCACGCGCTCGACTAGTAC-3'; where Ns represent a 6-bp barcode, the forward 454 fusion primer is italicized, and the adaptor overhang sequence is underlined) and the reverse *Collembola-R_454B* (5'-CTATGCGCCTTGCCAGCCCGCTCAG TGTTTCAAGACGGGACAGGC-3'), where the 454 fusion reverse primer is italicized.

Clean products (QIAquick PCR Purification Kit) were sequenced in the forward direction on 1/8 plate of a 454 Life Sciences Genome Sequencer FLX machine (Roche, Branford, CT) using the MacroGen facilities (South Korea).

2.4. Sequence processing

We used MOTHUR v.1.36.1 (Schloss et al., 2009) to filter NGS sequences with a minimum average quality score of 25 and minimum length, after trimming of primer sequences, of 150-bp. We allowed for no nucleotide differences in the barcode region of the oligo and four differences in the priming region. Clustering of reads into OTUs was done as described in the USEARCH 454 SOP (http://drive5.com/usearch/manual8.1/upp_454.html; Edgar, 2010), which also removes chimeras based on de novo detection. OTUs represented by nine or fewer reads were removed from all subsequent analyses. A read-clustering threshold of 3 % was adopted to bin OTUs. We assigned OTUs to taxonomy using a database created from all available Collembola GenBank entries using the RDP classifier (Wang et al., 2007).

2.5. Alpha and beta diversity estimates

In order to evaluate the influence of habitat quality on Collembola richness, we estimated the number of operational taxonomic units (OTUs; MOTHUR command *summary.single*) and the scaled phylogenetic diversity (command *phylo.diversity*) at each sampling location. For each of the two response variables, we ran a model selection based on a candidate model set including "forest quality" and "sampling month" as fixed factors, and "sampling region" as a random factor.

To test for differences in community composition, we calculated pairwise dissimilarity between all sampling points relying on i) the Jaccard index (occurrence-based index) and ii) the Bray-Curtis dissimilarity index (abundance-based), and used Non-metric multidimensional scaling (NMDS), as well as Permutational multivariate analysis of variance (PERMANOVA; Anderson, 2006) (see Supplemental Data 2 for details on statistical analyses). All analyses were carried out in R using lme4 (Bates et al., 2015) and vegan package (Oksanen et al. 2022).

3. Results

3.1. Primer design

The GenBank sequences employed in the initial primer design were used to create the *ecopcr* libraries. For Collembola, 278 D2 sequences (representing all four Collembola orders) conformed to the *ecopcr* parameters, i.e., possessed both forward and reverse primers that were at least 20-bp away from sequence end. Another 26,468 non-Collembola arthropod D2 sequences were retrieved that matched the parameters above.

The priming sites were nearly universally conserved in Collembola except for two polymorphic nucleotides in the forward primer (Fig. 1). However, the two degenerate positions in primer *Collembola-F* were 15 and 18 nucleotides away from the 3'-end, locations that generally result in little PCR bias (Kwok et al., 1990).

Our primer design strategy for Collembola sought to maximally exclude non-Collembola amplification by having 3' nucleotides that mismatched non-Collembola. The 3'-T in *Collembola-F* and 3'-C in *Collembola-R* were rarely found in any non-Collembola arthropod sequences (in 0.06 % and 4.6 % of those conforming to the *ecopcr* parameters, respectively), and in these instances the two primers never occurred together in an individual sequence, thus precluding PCR.

Conversely to the D2 primers, substantial primer conservation (both generally and at the 3' nucleotide) among all Metazoa was seen in the two primer sets previously used in Collembola metabarcoding (Saitoh et al., 2016). For example, the 16S priming positions were relatively well conserved amongst the Collembola sequences (thereby minimizing primer bias), but also were not substantially different at their 3' end from the other taxa evaluated (Supplemental Data 2, Fig. S1). Likewise, the CO1 primers were equally similar to Collembola and to non-Collembola targets (Supplemental Data 2, Fig. S2). Moreover, the CO1 primers were polymorphic at silent substitutions at each third codon positions, even those near the sensitive 3' end. This may increase primer bias (Arnheim and Erlich, 1992) and, consequently, limit comprehensive estimates of relative abundances of PCR targets in bulk samples.

The taxonomic resolution of the D2 primers was relatively similar to both the COI and 16S sets previously used by Saitoh (2016b). In comparisons of only those species that were shared between the two markers, D2 (36 sequences) and 16S (42 sequences) both distinguished 100 % of species. The COI primer set has marginally better resolution in 718 GenBank sequences than D2 (89 sequences): 96.9 % versus 95.4 %.

3.2. PCR and sequencing results

The first of the nested PCRs did not produce primer dimers or non-specific bands (Supplemental Data 2, Fig. S3). This substantially simplified the laboratory pipeline, as a clean-up step was not needed before the second, indexing PCR.

Following sequence filtering, 35 unique OTUs were derived from the two sampling locations, averaging ~460-bp. RDP results assigned all of these to class Collembola (100 % bootstrap; see Supplemental Data 3), indicating no cross-amplification of bycatch DNA. All OTUs were assigned to genera at RDP 80 % threshold or above, except OTUs 36, 27, and 8, which had support values of 72, 72, and 57 % respectively. The RDP results spanned ten of the 19 Collembola families currently represented in Brazil (Abrantes et al., 2010). Of the families previously recorded in Sao Paulo State, only Arrhopalitidae, Brachystomellidae and Tullbergiidae were undetected in our samples. Resulting sequences were assigned to all Collembola orders except Neelipleona, which has not previously been documented in Sao Paulo State.

Collembola OTUs were highly specific to either of the two Atlantic Forest localities. Out of the 35 OTUs, 24 were registered in Ubarana and 16 in Nazaré Paulista. Only OTU1, OTU2, OTU7, OTU10 and OTU20 occurred in both Nazaré Paulista and Ubarana (Supplemental Data 3). This high level of species turnover over ~500-km has been shown in

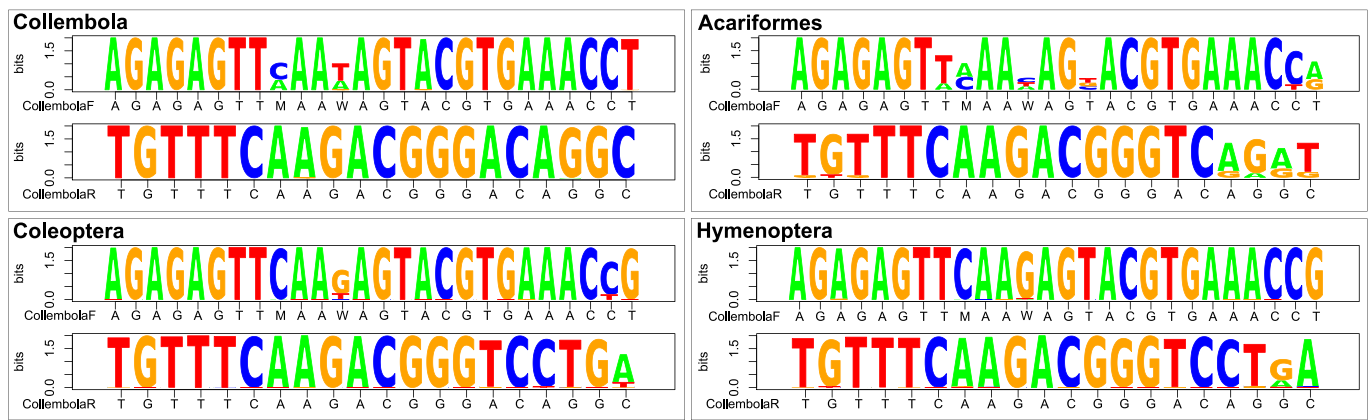


Fig. 1. Sequence logos of the D2 primers designed herein (*Collembola-F* and *Collembola-R*) assessed for their complementarity to Collembola and three additional taxa commonly found during pitfall sampling protocols: Acariformes, Coleoptera and Hymenoptera.

biomes of similar latitudes (South Africa; Janion-Scheepers, et al., 2020). This indicates that biogeography must be a consideration when using Collembola as bioindicators across large spatial areas.

Average number of OTUs registered in Ubarana high-quality sites was 6.6 (ranging from 2 to 14 OTUs) and 6.25 for Ubarana low quality sites (ranging from 3 to 8 OTUs). While for Nazaré Paulista, average number of OTUs registered in high quality sites was 3.75 (ranging from 2 to 5 OTUs) and 1.8 for low quality sites (ranging from 1 to 4 OTUs).

Demultiplexed 454 sequence data for each of the 35 samples analysed herein are in NCBI's Sequence Read Archive under project PRJNA645344 (<https://www.ncbi.nlm.nih.gov/sra/>).

3.3. Alpha and beta diversity

The results of the model selection explaining richness were similar for both response variables (number of OTUs and phylogenetic diversity). In both cases, the model including only forest quality as fixed factor was selected as the best model, and the additive model including sampling month as second-best model (Supplemental Data 2 Table S1). While the additive model was equally plausible to the best model explaining the variation in number of OTUs (i.e., $\Delta AICc < 2$) it was

not selected for phylogenetic diversity. The two models including forest quality thereby accumulated an AICc-weight of 94 % (number of OTUs) and 95 % (phylogenetic diversity). Model estimates indicated that sampling sites within high-quality habitat had higher number of OTUs observed as well as higher phylogenetic diversity (Fig. 2).

Most OTUs were highly specific to either high- or low-quality habitat, with only twelve OTUs occurring in both habitat classes (Supplemental Data 3), leading to a significant effect of habitat quality on community dissimilarity independently of the distance metric used (Supplemental Data 2 Table S2). Accordingly, high- and low-quality sampling sites were clearly separated by NMDS (Figs. 3 and S4). Communities were also significantly dissimilar among geographic sampling locations (Supplemental Data 2).

4. Discussion

Metabarcoding is a useful technique to assess soil species- and community-level information, thereby providing useful insight on terrestrial and overall ecosystem health. For example, Yang et al. (2014) found that leaf litter samples were considerably more informative in differentiating between habitat types than above-ground/aerial

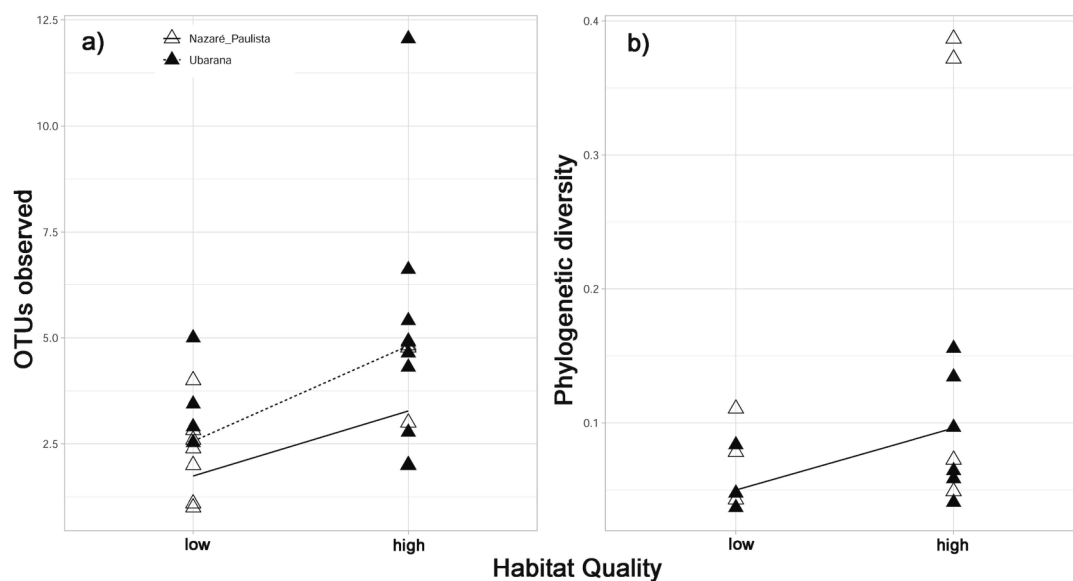


Fig. 2. Predictions of best model explaining variation in number of observed OTUs (a) and phylogenetic diversity (b). Open triangles and smooth line: samples from Nazaré Paulista, filled triangles and dashed line: samples from Ubarana. Note that predictions for the increase in phylogenetic diversity are similar between the two regions, which led to superimposition of the respective lines in b).

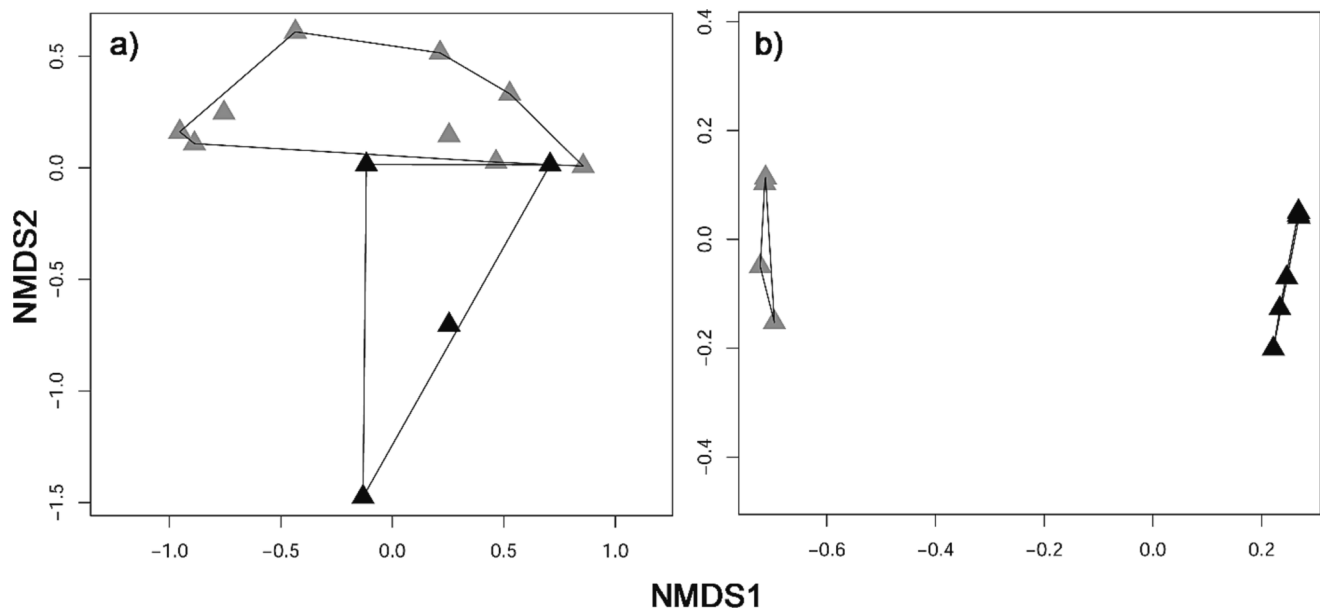


Fig. 3. Non-metric multi-dimensional scaling (NMDS) of pairwise comparison of Collembola communities in Ubarana (a, 13 communities) and Nazaré Paulista (b, 15 communities) based on Bray-Curtis-dissimilarity. Grey symbols: sampling sites in high forest quality areas; black symbols: sampling sites in low forest quality areas; Polygons enclose sites within same forest quality.

measures (Malaise trapping or fogging).

Among soil invertebrates, Collembola are of special interest given their proven potential as ecological and biological indicators (e.g., Arenhardt et al., 2021; Cassagne et al., 2006). The D2 primers herein allow for the exclusive amplification of Collembola from bulk samples with minimal PCR bias. The pipeline obviates time-consuming sorting and does not require that samples are morphologically analysed. It is especially appropriate when the sampling protocols used yield large and diverse amounts of bycatch (such as pitfall trapping), which, in previous Collembola metabarcoding studies, competed for large portions of the flow-cell yield (Saitoh, et al., 2016).

Our protocol differentiated several high- and low-quality samples that were defined principally based on the size of trees at pitfall locations. High-quality sites were composed of mature trees with closed canopies and consequently hosted fewer herbaceous ground species. Thus, leaf litter and surface sunlight/heat may be largely responsible for differences seen between sites among the epidaphic Collembola trapped by the pitfalls (Argote et al., 2023; Machado, et al., 2019).

The locations used to benchmark our protocol were approximately 400-km apart and although Collembola species distributions have been estimated across hundreds of kilometers (Fiera & Ulrich, 2012; Zeppelini Filho, et al., 2013), taxonomic turnover within the same Neotropical bioregion is little studied. Herein, Ubarana and Nazare Paulista (both characterized as Atlantic Forest) shared 14 % of unique species. This is similar to previous comparisons between samples from the Atlantic and Amazon forests (Zeppelini Filho, et al., 2013). These results confirm that a few dominant species may be relatively cosmopolitan across different biomes, but a majority are substantially limited in their dispersal and/or in their micro-environmental tolerances (Culik and Filho, 2003).

In comparisons with previously used markers, the D2 amplicon's species diagnostic performance is similar to both 16S and COI. Moreover, the low degeneracy of the D2 primers minimizes the PCR bias from synonymous mutations inherent in polymorphic protein-coding loci such as COI (Clarke et al., 2014; Deagle et al., 2014; Pedro et al., 2020). Although Saitoh et al. (2016) found a strong correlation between Collembola biomass and normalized sequence reads ($R = 0.91\text{--}0.99$) (suggesting the 16S primers also have little primer bias), their results with bulk samples included 35 % non-target sequences. This is a substantial loss of flow-cell capacity.

The D2 marker may currently be inappropriate in contexts where species diagnosis is required (such as in pest detection or toxicology assays) because of the relatively small taxonomic reference databases. However, many soil biomonitoring studies, rely on sample comparisons with reference sites, rather than molecular taxonomy *per se* (e.g., comparisons between pristine forests and perturbed sampling sites (Russell and Alberti, 1998; Arbolález et al., 2023)). Moreover, although other markers, notably COI, have substantially more complete database, they are nonetheless probably still underrepresented because of the sheer number of species thought to still be described, especially in tropical soils (Bernard and Felderhoff, 2007; Cicconardi et al., 2013).

CRediT authorship contribution statement

Pedro M. Pedro: Conceptualization, Methodology, Writing – review & editing, Writing – original draft, Investigation, Data analysis. **Laury Cullen Jr.:** Conceptualization, Funding acquisition. **Fabiana Prado:** Conceptualization, Funding acquisition. **Alexandre Uezu:** Investigation, Data analysis. **Ross Piper:** Investigation, Writing – review & editing. **Christiana M.A. Faria:** Investigation, Data analysis. **Christoph Knogge:** Writing – review & editing. **Maria Tereza Pepe Razzolini:** Investigation, Data analysis. **Marcela B. Paiva:** Investigation, Data analysis. **Milena Dropa:** Project administration, Investigation, Data analysis. **Miriam Silva:** Project administration, Investigation, Data analysis. **Tatiane Cristina Rech:** Project administration, Funding acquisition. **Thomas Püttker:** Methodology, Writing – review & editing, Writing – original draft, Investigation, Data analysis.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Pedro M. Pedro reports financial support was provided by Agência Nacional de Energia Elétrica. Pedro M. Pedro reports financial support was provided by Projeto LIRA - Legado Integrado da Região Amazônica. Pedro M. Pedro reports financial support was provided by Fundo Amazônia. Pedro M. Pedro reports financial support was provided by Gordon and Betty Moore Foundation.

Data availability

Demultiplexed 454 sequence data for each of the 35 samples analysed herein are in NCBI's Sequence Read Archive under project PRJNA645344 (<https://www.ncbi.nlm.nih.gov/sra/>).

Acknowledgments

This research received R&D funding from ANEEL (*Agência Nacional de Energia Elétrica*; grant number 0064-1035/2014) and also from *Projeto LIRA - Legado Integrado da Região Amazônica*, with funding partners *Fundo Amazônia/Banco Nacional de Desenvolvimento Econômico e Social* and the *Gordon and Betty Moore Foundation*. Sample collections were undertaken with permit number 54835 administered by the Brazilian Federal agency SISBIO. We would like to thank IdeaWild (<https://ideawild.org/>) for important equipment contributions to this project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ecolind.2023.111202>. These data include Google maps of the most important areas described in this article.

References

- Abrantes, E.A., Bellini, B.C., Bernardo, A.N., Fernandes, L.H., Mendonça, M.C., Oliveira, E.P., Queiroz, G.C., Sautter, K.D., Silveira, T.C., Zeppelini, D., 2010. Synthesis of Brazilian Collembola: An update to the species list. *Zootaxa* 2388, 1–22. <https://doi.org/10.11646/zootaxa.2388.1.1>.
- Anderson, M.J., 2006. Distance-based tests for homogeneity of multivariate dispersions. *Biometrics* 62, 245–253. <https://doi.org/10.1111/j.1541-0420.2005.00440.x>.
- Arboláez, H.P., Hu, J., Orozco, Y.N., Gebremikael, M.T., Alcantara, E.A., Sleutel, S., Höfte, M., De Neve, S., 2023. Mesofauna as effective indicators of soil quality differences in the agricultural systems of central Cuba. *Appl. Soil Ecol.* 182, 104688. <https://doi.org/10.1016/j.apsoil.2022.104688>.
- Arenhardt, T.C.P., Vitorino, M.D., Martins, S.V., 2021. Insecta and Collembola as bioindicators of ecological restoration in the Ombrophilous Dense Forest in Southern Brazil. *Floresta e Ambiente* 28, 2–11. <https://doi.org/10.1590/2179-8087-FLORAM-2021-0008>.
- Argote, K., Albert, C.H., Geslin, B., Biryol, C., Santonja, M., 2023. Effects of litter quality on foraging behaviour and demographic parameters in *Folsomia candida* (Collembola). *Ecol. Evol.* 13, e10240.
- Arnheim, N., Erlich, H., 1992. Polymerase Chain Reaction Strategy. *Annu. Rev. Biochem.* 61, 131–156. <https://doi.org/10.1146/annurev.bi.61.070192.001023>.
- Basset, Y., Hajibabaei, M., Wright, M.T.G., Castillo, A.M., Donoso, D.A., Segar, S.T., Souto-Vilarós, D., Soliman, D.Y., Roslin, T., Smith, M.A., Lamarre, G.P.A., de León, L. F., Decaëns, T., Palacios-Vargas, J.G., Castaño-Meneses, G., Scheffrahn, R.H., Rivera, M., Perez, F., Bobadilla, R., Lopez, Y., Alejandro, J., Silva, R., Cruz, M.M., Galván, A. A., Barrios, H., 2022. Comparison of traditional and DNA metabarcoding samples for monitoring tropical soil arthropods (Formicidae, Collembola and Isoptera). *Scientific Reports* 2022 12:1 12, 1–16. <https://doi.org/10.1038/s41598-022-14915-2>.
- Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R.H.B., Singmann, H., Dai, B., et al., 2015. *lme4: Linear mixed-effects models using Eigen and S4*. R Package Version 1 (1–7), 2014.
- Beng, K.C., Tomlinson, K.W., Shen, X.H., Surget-Groba, Y., Hughes, A.C., Corlett, R.T., Slik, J.W.F., 2016. The utility of DNA metabarcoding for studying the response of arthropod diversity and composition to land-use change in the tropics. *Sci. Rep.* 6, 24965. <https://doi.org/10.1038/srep24965>.
- Bernard, E.C., Felderhoff, K.L., 2007. Biodiversity explosion: Collembola (springtails) of Great Smoky Mountains National Park. *Southeast. Nat.* [https://doi.org/10.1656/1528-7092\(2007\)6\[175:BECSOG\]2.0.CO;2](https://doi.org/10.1656/1528-7092(2007)6[175:BECSOG]2.0.CO;2).
- Bispo, A., Cluzeau, D., Creamer, R., Dombos, M., Graefe, U., Krogh, P.H., Sousa, J.P., Peres, G., Rutgers, M., Winding, A., Römbke, J., 2009. Indicators for monitoring soil biodiversity. *Integr. Environ. Assess. Manag.* 5, 717–719. <https://setac.onlinelibrary.wiley.com/doi/10.1897/IEAM.2009-064.1>.
- Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., Coissac, E., 2016. obitools: A unix-inspired software package for DNA metabarcoding. *Mol. Ecol. Resour.* 16, 176–182. <https://doi.org/10.1111/1755-0998.12428>.
- Breure, A.M., Mulder, C., Rutgers, M., Schouten, T., de Zwart, D., Bloem, J., 2003. A biological indicator for soil quality. *Proceedings from an OECD Expert Meeting Rome, Italy* 485–494.
- Brown, E.A., Chain, F.J.J., Zhan, A., MacIsaac, H.J., Cristescu, M.E., 2016. Early detection of aquatic invaders using metabarcoding reveals a high number of non-indigenous species in Canadian ports. *Divers. Distrib.* 22, 1045–1059. <https://doi.org/10.1111/ddi.12465>.
- Burki, F., Sandin, M.M., Jamy, M., 2021. Diversity and ecology of protists revealed by metabarcoding. *Curr. Biol.* 31, R1267–R1280. <https://doi.org/10.1016/j.cub.2021.07.066>.
- Campbell, B.C., Steffen-Campbell, J.D., Werren, J.H., 1994. Phylogeny of the *Nasonia* species complex (Hymenoptera: Pteromalidae) inferred from an internal transcribed spacer (ITS2) and 28S rDNA sequences. *Insect Mol. Biol.* 2, 225–237. <https://doi.org/10.1111/j.1365-2583.1994.tb00142.x>.
- Cassagne, N., Gauquelin, T., Bal-Serin, M.C., Gers, C., 2006. Endemic Collembola, privileged bioindicators of forest management. *Pedobiologia (jena)* 50, 127–134. <https://doi.org/10.1016/j.pedobi.2005.10.002>.
- Cicconardi, F., Fanciulli, P.P., Emerson, B.C., 2013. Collembola, the biological species concept and the underestimation of global species richness. *Mol. Ecol.* 22, 5382–5396. <https://doi.org/10.1111/mec.12472>.
- Clarke, L.J., Soubrier, J., Weyrich, L.S., Cooper, A., 2014. Environmental metabarcodes for insects: In silico PCR reveals potential for taxonomic bias. *Mol. Ecol. Resour.* 14, 1160–1170. <https://doi.org/10.1111/1755-0998.12265>.
- Culik, M.P., de Souza, J.L., Ventura, J.A., 2002. Biodiversity of Collembola in tropical agricultural environments of Espírito Santo, Brazil. *Appl. Soil Ecol.* 21, 49–58. [https://doi.org/10.1016/S0929-1393\(02\)00048-3](https://doi.org/10.1016/S0929-1393(02)00048-3).
- Culik, M.P., Filho, D.Z., 2003. Diversity and distribution of collembola (arthropoda: hexapoda) of Brazil. *Biod. Conser.* 12, 1119–1143.
- de Filho, L.C.I.O., Klauberg Filho, O., Baretta, D., Tanaka, C.A.S., Sousa, J.P., 2016. Collembola community structure as a tool to assess land use effects on soil quality. *Rev. Bras. Cienc Solo* 40, 1–18. <https://doi.org/10.1590/18069657rbc20150432>.
- Deagle, B.E., Jarman, S.N., Coissac, E., Pompanon, F., Taberlet, P., 2014. DNA metabarcoding and the cytochrome c oxidase subunit I marker: Not a perfect match. *Biol. Lett.* 10, 20140562. <https://doi.org/10.1098/rsbl.2014.0562>.
- Dodd, S.L., Crowhurst, R.N., Rodrigo, A.G., Samuels, G.J., Hill, R.A., Stewart, A., 2000. Examination of Trichoderma phylogenies derived from ribosomal DNA sequence data. *Mycol. Res.* 104, 23–34. <https://doi.org/10.1017/S0953756299001100>.
- Eaton, W.D., Shokralla, S., McGee, K.M., Hajibabaei, M., 2017. Using metagenomics to show the efficacy of forest restoration in the New Jersey Pine Barrens. *Genome* 60, 825–836. <https://doi.org/10.1139/gen-2015-0199>.
- Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. <https://doi.org/10.1093/bioinformatics/btq461>.
- Ficetola, G.F., Miaud, C., Pompanon, F., Taberlet, P., 2008. Species detection using environmental DNA from water samples. *Biol. Lett.* 4, 423–425. <https://doi.org/10.1098/rsbl.2008.0118>.
- Fiera, C., 2009. Biodiversity of Collembola in urban soils and their use as bioindicators for pollution. *Pesqui Agropecu Bras.* 44, 868–873. <https://doi.org/10.1590/S0100-204X2009000800010>.
- Fiera, C., Ulrich, W., 2012. Spatial patterns in the distribution of European springtails (Hexapoda: Collembola). *Biol. J. Linn. Soc.* 105, 498–506. <https://doi.org/10.1111/j.1095-8312.2011.01816.x>.
- Fujii, S., Takeda, H., 2012. Succession of collembolan communities during decomposition of leaf and root litter: effects of litter type and position. *Soil Biol. Biochem.* 54, 77–85. <https://doi.org/10.1016/j.soilbio.2012.04.021>.
- Janion-Scheepers, C., Bengtsson, J., Duffy, G.A., Deharveng, L., Leinaas, H.P., Chown, S. L., 2020. High spatial turnover in springtails of the Cape Floristic Region. *J. Biogeogr.* 47, 1007–1018. <https://doi.org/10.1111/jbi.13801>.
- Ji, Y., Ashton, L., Pedley, S.M., Edwards, D.P., Tang, Y., Nakamura, A., Kitching, R., Dolman, P.M., Woodcock, P., Edwards, F.A., Larsen, T.H., Hsu, W.W., Benedick, S., Hamer, K.C., Wilcove, D.S., Bruce, C., Wang, X., Levi, T., Lott, M., Emerson, B.C., Yu, D.W., 2013. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecol. Lett.* 16, 1245–1257. <https://doi.org/10.1111/ele.12162>.
- Kwok, S., Kellogg, D.E., McKinney, N., Spasic, D., Goda, L., Levenson, C., Sninsky, J.J., 1990. Effects of primer-template mismatches on the polymerase chain reaction: Human immunodeficiency virus type 1 model studies [WWW Document]. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/18.4.999>.
- Machado, J.D.S., Oliveira, L.C.I., Santos, J.C.P., Paulino, A.T., Baretta, D., 2019. Morphological diversity of springtails (Hexapoda: Collembola) as soil quality bioindicators in land use systems. *Biota Neotrop* 19. <https://doi.org/10.1590/1676-0611-BN-2018-0618>.
- Oksanen, J., Simpson, G., Blanchet, F., Kindt, R., Legendre, P., Minchin, P., O'Hara, R., Solymos, P., Stevens, M., Szöcs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Carvalho, G., Chirico, M., De Caceres, M., Durand, S., Evangelista, H., FitzJohn, R., Friendly, M., Furneaux, B., Hannigan, G., Hill, M., Lahti, L., McGlenn, D., Ouellette, M., Ribeiro Cunha, E., Smith, T., Stier, A., Ter Braak, C., Weedon, J., 2022. *Vegan: Community Ecology Package*. R package version 2.6-2, <https://CRAN.R-project.org/package=vegan>.
- Pedro, P.M., Amorim, J., Rojas, M.V.R., Sá, L.L., Galardo, A.K.R., Santos Neto, N.F., Pires de Carvalho, D., Nabas Ribeiro, K.A., Razzolini, M.T.P., Sallum, M.A.M., 2020. *Culicidae-centric metabarcoding through targeted use of D2 ribosomal DNA primers*. *PeerJ* 8, e9057.
- Pedro, P.M., de Sá, L.L.R., Rojas, M.V.R., Amorim, J.A., Galardo, A.K.R., Santos Neto, N. F., Furtado, N.V.R., de Carvalho, D.P., Ribeiro, K.A.N., de Paiva, M., Razzolini, M.T.P., Sallum, M.A.M., 2021. Efficient monitoring of adult and immature mosquitoes through metabarcoding of bulk samples: a case study for non-model culicids with unique ecologies. *J. Med. Entomol.* 58, 1210–1218. <https://doi.org/10.1093/jme/tjaa267>.
- Porco, D., Bedos, A., Greenslade, P., Janion, C., Skarżyński, D., Stevens, M.L., Jansen Van Vuuren, B., Deharveng, L., 2012. Challenging species delimitation in Collembola: Cryptic diversity among common springtails unveiled by DNA barcoding. *Invertebr. Syst.* 26, 470–477. <https://doi.org/10.1071/IS12026>.
- Potapov, A., Bellini, B., Chown, S., Deharveng, L., Janssens, F., Kováč, L., Kuznetsova, N., Ponge, J.F., Potapov, M., Querner, P., Russell, D., Sun, X., Zhang, F., Berg, M., 2020. Towards a global synthesis of Collembola knowledge - challenges and potential solutions. *Soil Org.* 92: 161-188. <https://doi.org/10.25674/so92iss3pp161>.

- Russell, D.J., Alberti, G., 1998. Effects of long-term, geogenic heavy metal contamination on soil organic matter and microarthropod communities, in particular Collembola. *Appl. Soil Ecol.* 9, 483–488. [https://doi.org/10.1016/S0929-1393\(98\)00109-7](https://doi.org/10.1016/S0929-1393(98)00109-7).
- 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, 705–723. <https://doi.org/10.1139/gen-2015-0228>.
- Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., Van Horn, D.J., Weber, C.F., 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541. <https://doi.org/10.1128/AEM.01541-09>.
- Semmouri, I., de Schampelaere, K.A.C., Willemsse, S., Vandegheuchte, M.B., Janssen, C. R., Asselman, J., 2021. Metabarcoding reveals hidden species and improves identification of marine zooplankton communities in the North Sea. *ICES J. Mar. Sci.* 78, 3411–3427. <https://doi.org/10.1093/icesjms/fsaa256>.
- Sun, X., Zhang, F., Ding, Y., Davies, T.W., Li, Y., Wu, D., 2017. Delimiting species of Protaphorura (Collembola: Onychiuridae): integrative evidence based on morphology, DNA sequences and geography. *Sci Rep* 7. <https://doi.org/10.1038/s41598-017-08381-4>.
- Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R., 2007. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267. <https://doi.org/10.1128/AEM.00062-07>.
- Yang, C., Wang, X., Miller, J.A., De Blécourt, M., Ji, Y., Yang, C., Harrison, R.D., Yu, D. W., 2014. Using metabarcoding to ask if easily collected soil and leaf-litter samples can be used as a general biodiversity indicator. *Ecol. Ind.* 46, 379–389. <https://doi.org/10.1016/j.ecolind.2014.06.028>.
- Yu, D.W., Ji, Y., Emerson, B.C., Wang, X., Ye, C., Yang, C., Ding, Z., 2012. Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods Ecol. Evol.* 613–623 <https://doi.org/10.1111/j.2041-210X.2012.00198.x>.
- Zeppelini, D., Bellini, B.C., Creão-Duarte, A.J., Hernández, M.I.M., 2009. Collembola as bioindicators of restoration in mined sand dunes of Northeastern Brazil. *Biodivers. Conserv.* 18, 1161–1170. <https://doi.org/10.1007/s10531-008-9505-2>.
- Zeppelini Filho, D., Costa, Q.G., Assis, A.E., Bruno, B., Cavalcante, M.E.S.F., Pereira, O.E., Liugo, M., 2013. Diversity of Collembola (Arthropoda: Hexapoda) across different types of vegetation in Brazil. *Int. J. Biodivers. Conserv.* 5, 176–184.