Environmental Entomology, XX(X), 2017, 1–9
doi: 10.1093/ee/nvx166
Research

Metabarcoding Analyses Enable Differentiation of Both Interspecific Assemblages and Intraspecific Divergence in Habitats With Differing Management Practices

Pedro M. Pedro,[1,7] Ross Piper,[2] Pedro Bazilli Neto,[3] Laury Cullen, Jr.,[1] Milena Dropa,[4] Rogerio Lorencao,[1] Maria Helena Matté,[4] Tatiane Cristina Rech,[5] Mauro Osmar Rufato, Jr.,[1] Miriam Silva,[4] and Daniele Turganti Turati[6]

[1]IPE – Institute for Ecological Research, Nazaré Paulista, SP 12960-000, Brazil, [2]The Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, United Kingdom, [3]Fazenda São Pedro, Caconde, SP 13770-000, Brazil, [4]Faculdade de Saúde Pública – USP, São Paulo, SP 01246-904, Brazil, [5]AES Tietê, Promissão, SP 16370-000, Brazil, [6]Sítio São Luís, Caconde, SP 13770-000, Brazil, and [7]Corresponding author, e-mail: pedro@ipe.org.br

Received 21 March 2017; Editorial decision 26 September 2017

Abstract

Spatial and temporal collections provide important data on the distribution and dispersal of species. Regional-scale monitoring invariably involves hundreds of thousands of samples, the identification of which is costly in both time and money. In this respect, metabarcoding is increasingly seen as a viable alternative to traditional morphological identification, as it eliminates the taxonomic bottleneck previously impeding such work. Here, we assess whether terrestrial arthropods collected from 12 pitfall traps in two farms of a coffee (*Coffea arabica* L.) growing region of Sao Paulo State, Brazil could differentiate the two locations. We sequenced a portion of the cytochrome oxidase 1 region from minimally processed pools of samples and assessed inter- and intraspecific parameters across the two locations. Our sequencing was sufficient to circumscribe the overall diversity, which was characterized by few dominant taxa, principally small Coleoptera species and Collembola. Thirty-four operational taxonomic units were detected and of these, eight were present in significantly different quantities between the two farms. Analysis of community-wide Beta diversity grouped collections based on farm provenance. Moreover, haplotype-based analyses for a species of *Xyleborus* beetle showed that there is significant population genetic structuring between the two farms, suggesting limited dispersal. We conclude that metabarcoding can provide important management input and, considering the rapidly declining cost of sequencing, suggest that large-scale monitoring is now feasible and can identify both the taxa present as well as contribute information about genetic diversity of focal species.

Key words: next-generation sequencing, metabarcoding, Beta diversity, insect dispersal

So-called metagenetics initially involved the use of next-generation sequencing (NGS) platforms to reveal the previously hidden biodiversity of microscopic biomes (Buée et al. 2009, Petrosino et al. 2009). In such work, the study targets, generally fungi or bacteria, are so numerous and small that they cannot be taxonomically assigned in a financially efficient manner, hence the desirability of NGS. Moreover, such organisms are difficult to identify morphologically or cannot be cultured in vitro, which was traditionally done before the advent of NGS. This technology is now frequently used for organisms other than bacteria and fungi, including nematodes and protists (Creer 2010, Pawlowski et al. 2014).

More recent work includes macroscopic samples of pooled invertebrates, including arthropods (Shokralla et al. 2012, Yu et al. 2012). In combination with the growing databases that allow taxonomic diagnosis based on DNA sequences (e.g., www.boldsystems.org; GenBank), these metagenetic studies have been termed metabarcoding (Baird and Hajibabaei 2012, Taberlet et al. 2012). The most common metabarcoding pipeline involves the targeted PCR amplification of a taxonomically informative marker. In Metazoa, this is often the cytochrome oxidase 1 (*CO1*) mitochondrial gene.

Like the smaller taxa, macroinvertebrates are often taxonomically cryptic and can be sampled in enormous numbers, which make inventorying time- and resource-intensive. Initial attempts to explore this line of research included artificial assemblages created with known species compositions to test the output coverage from the NGS sequencers. Such sequencing normally yields a faithful representation of the original composition; generally, over 80% of a priori species are recovered by the sequencer (Hajibabaei et al. 2012, Gibson et al. 2014, Elbrecht and Leese 2015, Kekkonen et al. 2015).

While these results are encouraging, universal primer sets are not viable for *CO1* across broad taxonomic space (such as all Metazoa) because of this marker's highly polymorphic priming sites. However, some

© The Authors 2017. Published by Oxford University Press on behalf of Entomological Society of America.
All rights reserved. For permissions, please email: journals.permissions@oup.com.

sets have successfully used degenerate nucleotides to create near-universal priming. Of these, the primers designed by Leray et al. (2013) amplified a 313-bp fragment of the *CO1* across 14 animal phyla. Although some informative taxonomic data are lost in such a small sequence length, the relatively small amplicon is highly appropriate for field-collected samples, which are often degraded during sampling. Moreover, small amplicon size minimizes PCR bias, which can lead to better approximations of relative abundances (Huber et al. 2009).

Considering these encouraging results, there is a surprising lack of work where the pooled assemblage is not artificial, i.e., a blind study using wild-caught samples. Of the few that exist, most have dealt with the monitoring of aquatic biomes (e.g., Carew et al. 2013, Gibson et al. 2015) to both identify invasive species and estimate water quality indices.

Only recently have soil biomes been sampled from wild-caught (as opposed to artificially constructed) assemblages (e.g., Arribas et al. 2016, Beng et al. 2016). No studies, however, have tested explicitly the Alpha and Beta diversity of collections from pitfall traps. Given its proven capacity, metabarcoding should have the power to both describe and differentiate the biodiversity between sampling locations. In other words, Beta diversity indices should aggregate replicates within locations and reciprocally distinguish comparisons between locations.

Congruence in Beta diversity between pools sampled in the same habitat would be the desired outcome in any applied biomonitoring framework, as it would imply that a sufficient amount of effort has been spent sampling the local biodiversity. In many cases (such as megadiverse tropical forests), this may still be unrealistic, as a full representation of even macroinvertebrate biodiversity would likely include thousands of species. However, well-designed and dense sampling schemes combined with newer high-throughput sequencers, such as the Illumina platforms, may resolve this problem (Brandon-Mong et al. 2015, Aylagas and Rodríguez-Ezpeleta 2016).

NGS technology can also complement biomonitoring by efficiently characterizing the dispersal of focal taxa via intraspecific genetics. The dispersal capacity (either anthropogenic or natural) of such organisms is most easily estimated via population genetic studies that track lineages through space and time. Thus, in addition to the presence/absence of relevant taxa, pooled samples can also provide intraspecific population genetics metrics, assuming sufficient individuals from the same species are collected (Johnson and Slatkin 2006).

In practice, such intraspecific work is important for at least two reasons: identifying the geographical provenance of invasive species on a global scale and understanding dispersal routes of focal taxa through the landscape. Traditionally, such work has relied on the physical separation and PCR-based analyses of individual organisms, which is time- and resource-intensive (e.g., Guidolin et al. 2014). If organisms are captured in sufficient numbers (as many pests often are), NGS output can be considered to represent haplotype frequencies within the sampling point. These can then be used to track dispersal parameters and population size, among other parameters.

With growing concerns about the potential for climatic changes and exotic species to alter ecosystems (Walther et al. 2002), the potential for concurrent identification of relevant species and their population genetic parameters is imperative. Importantly, these two ends can be bundled into a single sampling and laboratory pipeline through metabarcoding. Although proof-of-concept studies are increasingly common for invertebrate pools, few have attempted to apply these protocols. Here, we show that NGS platforms can be used to both differentiate ecological assemblages and to describe intraspecific variation from pooled samples collected in pitfall traps on coffee farms. We show that even a small sampling effort yields informative data for both categories.

Materials and Methods

Sampling Locations

We sampled along linear gradients within two conventional full-sun coffee (*Coffea arabica* L.) farms inserted within a forested matrix. Traps were set approximately 50 m apart and placed at varying distances from adjacent forests. However, proximity to forests did not significantly impact pitfall contents (data not shown). A concurrent experiment sampled invertebrates within the forests, and thus operational taxonomic unit (OTU) identifications herein are shown in nonsequential order.

Farm I (approximately 21°35′15″S, 46°36′02″W) is situated on relatively flat terrain, where harvesting is undertaken by mechanized means. A second collection site (Farm II, approximately 21°26′56″S, 46°36′20″W) was chosen to include a hillside system, with manual harvest. However, because of topography and harvesting methods, Farm II contained significantly more understory and coffee trees were physically closer together, creating a more humid habitat harboring noticeably more diversity (personal field observation).

We placed pitfall traps in eight locations on Farm I (denoted Ia–Ih) and four on Farm II (IIa–IId). These traps were installed in the understory of the coffee trees, equidistant from tree trunks. They consisted of a 400-ml plastic cup, one-half filled with absolute ethanol. An improvised plastic funnel was inverted at the top of the cup to mitigate alcohol evaporation. The trap was covered with a plastic plate to prevent entry of rain and detritus. Pitfall traps were deployed from 23 January 2015 to 10 February 2015.

DNA Extraction and PCR

The collected invertebrates included a broad range of body sizes, although most were <0.5 cm. To diminish the proportion of DNA from larger-bodied individuals, we contributed only 4 mm of the legs from arthropods larger than 1 cm. The contents of pitfall traps were returned to the laboratory and stored at −20°C for not more than 1 wk prior to DNA extraction.

Samples that were initially processed as above were divided into 2 ml subsamples and macerated using a Savant FastPrep lysis mill at maximum speed for 20 s using 1-mm ceramic beads. Following lysis, the samples were reassembled into a single aggregate sample. A subsample of this product was then submitted to DNA extraction with a DNeasy Blood & Tissue Kit (Qiagen, Valencia, CA) following manufacturer's instructions.

Nested PCR reactions were performed using the Metazoa-specific primer set for *CO1* of Leray et al. (2013) adapted for use on the 454-FLX sequencing platform. In summary: a first PCR was done using the forward primer mlCOIintF_adF (*5′*-GGC CAC GCG TCG ACT AGT ACG GWA CWG GWT GAA CWG TWT AYC CYC C-3'), where the underlined portion is an adaptor overhang used in our laboratory to decrease the cost of multiplexing PCR primers. The reverse primer for the first PCR was jgHCO2198 (*5′*-TAI ACY TCI GGR TGI CCR AAR AAY CA-3′). The product from the initial reaction was diluted 10× and submitted to a second PCR using the forward primer 454A-MID-adF (*5′-CGT ATC GCC TCC CTC GCG CCA TCA G*NN NNN NGG CCA CGC GTC GAC TAG TAC-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 jgHCO2198_454R (*5′-CTA TGC GCC TTG CCA GCC CGC TCA G*TA IAC YTC IGG RTG ICC RAA RAA YCA-3′, where the 454 fusion reverse primer is italicized).

Both first- and second-round PCR reactions were performed in triplicate and pooled to minimize PCR variability and also included a negative control with no template added. For each, a 25-µl PCR reaction was carried out using the GoTaq PCR system (Promega, Madison, WI). First-round reactions included 1 µl of DNA template, 0.2 mM of forward and reverse primers, 1× PCR buffer, 1.5 mM $MgCl_2$, 0.1 mM dNTP mix, and 1 unit of Taq polymerase. Thermal cycler settings for the initial PCR included a 10-min denaturation step at 95°C followed by 15 cycles of 94°C for 20 s, 50°C for 30 s, and 72°C for 40 s. An elongation step at 72°C for 7 min was then performed. The second, nested, PCR utilized identical temperature profiles except that the initial denaturation was run for only 1 min. The efficiency of the reactions and the absence of contamination in negative controls were verified by electrophoresis of the PCR products on a 1% agarose gel, stained with ethidium bromide.

Amplicon concentrations were measured using a Qubit fluorometer (Thermo Fisher Scientific, Waltham, MA) and equimolar concentrations of each sample were pooled and purified using the QIAquick PCR Purification Kit (Qiagen). Clean products 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). The sequencing run was shared with 36 additional samples from other experiments, thus using a combined 48 MID tags within the 1/8th run. Because the pitfall samples analyzed herein contained on average fewer than ~200 individual invertebrates, we calibrated the concentration of the 12 samples to occupy only ~1/500th of the gasket lane.

Sequence Data Processing

Sequence quality filtering was undertaken using MOTHUR v.1.36.1 (Schloss et al. 2009). Quality criteria included a minimum sequence length of 200 bp for OTU-based analyses and 250 bp for haplotype analyses, all with a minimum average quality score of 25. We allowed for no nucleotide differences in the barcode region and four differences in the priming region. Sequence reads were clustered into OTUs as described in the USEARCH SOP (http://drive5.com/usearch/manual/upp_454.html; Edgar 2010). This pipeline uses the USEARCH algorithm to remove form downstream analyses chimeras based on de novo detection from the supplied sequences. The longest representative from each cluster was that adopted as the representative OTU. OTUs represented by fewer than five reads were removed from all downstream analyses. All phylogenetic trees calculated from sequence data were created using MEGA v.7.0.9 with the Jukes-Cantor genetic distance model (Kumar et al. 2016).

As described below, we utilized two clustering thresholds. To assess species assemblage-scale parameters, we used a USEARCH 97% threshold, which removes potential intraspecific distinctions (i.e., *CO1* haplotypes). In the second analysis, which sought to extract exact (unclustered) haplotypes, we used a USEARCH threshold of 100%.

For both analyses, we removed OTUs that were not assigned to Arthropoda with bootstrap support of 70% or higher as estimated using the RDP classifier (Wang et al. 2007). Here, we identified the database matches to OTUs using a 50% bootstrap cutoff, as recommended by authors, using the MIDORI UNIQUE Metazoa dataset (Machida et al. 2017). In all diversity calculations, we used the MOTHUR command *subsample=T* to correct for disparate sequencing depth among the 12 collections.

OTU-Based Analyses

Beta Diversity

Community ecological parameters were calculated in MOTHUR using the Jaccard diversity estimate. We used two methods to test whether Beta values indicated more similar intersite values within farms than between farms: MOTHUR's *parsimony* module based on tree topology (Schloss and Handelsman 2006) and permutational multivariate analysis of variance (PERMANOVA; Anderson 2001). PERMANOVA was implemented using the *adonis* function of the vegan package in R with 1,000 permutations (Oksanen 2013). Differences in community composition were evaluated graphically using nonmetric multidimensional scaling (NMDS) ordination based on rank scores, also using the Jaccard index.

Alpha Diversity

Alpha diversity was estimated using the MOTHUR command *summary.single* to calculate Shannon and Inverse Simpson diversity. The command *phylo.diversity* was used to calculate scaled phylogenetic diversity. Rarefaction curves were generated separately for each sample with the *rarefaction.single* command.

MOTHUR was also utilized to diagnose OTUs differentially detected between the two farms, using an implementation of the Metastats program of White et al. (2009). Here, statistical correction for multiple comparisons was done using *q* values.

Haplotype-Based Analyses

We used USEARCH to cluster quality-filtered sequences as described above, except we utilized a 100% threshold and sequence length of 250 bp in order to capture as much haplotypic diversity as possible. In other words, all sequence variants from the quality-filtered reads were derived with no clustering except when they were identical replicates. These were considered *CO1* haplotypes.

We then clustered these unique haplotypes to a 97% threshold under the assumption that this cutoff does not capture interspecific variation (intraspecific variation is often 3% or less for most species [Hajibabaei et al. 2006]). We selected species haplotypes that occurred at more than one location on both farms in order to calculate analysis of molecular variance (AMOVA) using Arlequin v.3.5. AMOVA was based on F_{ST} values for haplotype frequencies with significance tested using 1,000 permutations (Excoffier et al. 2005).

Results

Prior to quality filtering, the 12 pitfall samples collected from the two coffee farms accounted for 2,345 sequence reads. As discussed previously, *CO1* amplicon reads from the samples analyzed herein were part of a larger sequencing run that included an additional 36 samples collected from other experiments within forested areas in the same region. All OTUs and haplotypes derived from this experiment were deposited in GenBank under accession numbers MF987844-MF987877.

OTU-Based Analyses

Thirty-four individual OTUs (clustered at a threshold of 97%), derived from 2027 quality-filtered sequences, were characterized following sample processing (Table 1). The highest proportion of sequences in both farm locations were assigned to Coleoptera (54%) and Entomobryomorpha (Collembola; 21%) (see Supp Data online only). Within the Coleoptera, the families most represented were Nitidulidae (34% of Coleoptera reads), Curculionidae (35%), and Staphylinidae (26%). These data agree with visual observations prior to DNA extraction that indicated most pitfall samples were composed of smaller Coleoptera (<0.5 cm) and Collembola.

The individual Coleoptera OTUs most commonly sampled from the two farms were OTU 8, identified as *Urophorus* sp. with 100% RDP bootstrap support (comprising 18% of total sequence reads), and OTU 3, identified to *Xyleborus* at 98% bootstrap support (15%

Table 1. Taxonomic assignment of 34 OTUs based on RDP analysis with assignment bootstrap values above 50% listed for each taxonomic level

OTU-ID	Class	Bootstrap support	Order	Bootstrap support	Family	Bootstrap support	Genus	Bootstrap support
OTU_165	Arachnida	1.00	Araneae	1.00	Agelenidae	0.53	*Tegenaria*	0.53
OTU_87	Collembola	0.88	Entomobryomorpha	0.88	Entomobryidae	0.88	*Entomobrya*	0.88
OTU_75	Collembola	0.50	Entomobryomorpha	0.50	Entomobryidae	0.50	*Entomobrya*	0.50
OTU_9	Collembola	1.00	Entomobryomorpha	1.00	Entomobryidae	1.00	*Lepidocyrtus*	1.00
OTU_13	Collembola	1.00	Entomobryomorpha	1.00	Isotomidae	1.00	*Weberacantha*	1.00
OTU_55	Diplopoda	1.00	Polydesmida	1.00	Xystodesmidae	1.00	*Appalachioria*	1.00
OTU_61	Insecta	1.00	Blattodea	1.00	Blaberidae	1.00	*Pycnoscelus*	1.00
OTU_56	Insecta	1.00	Blattodea	0.82	Blattidae	0.50		
OTU_28	Insecta	1.00	Blattodea	1.00	Blattidae	1.00	*Hebardina*	0.95
OTU_17	Insecta	1.00	Coleoptera	1.00	Carabidae	1.00	*Calosoma*	1.00
OTU_39	Insecta	1.00	Coleoptera	1.00	Chrysomelidae	1.00	*Bruchidius*	0.97
OTU_3	Insecta	1.00	Coleoptera	1.00	Curculionidae	1.00	*Xyleborus*	0.98
OTU_31	Insecta	1.00	Coleoptera	1.00	Curculionidae	1.00	*Xyleborus*	0.93
OTU_29	Insecta	1.00	Coleoptera	1.00	Languriidae	1.00	*Cryptophilus*	1.00
OTU_8	Insecta	1.00	Coleoptera	1.00	Nitidulidae	1.00	*Urophorus*	1.00
OTU_59	Insecta	1.00	Coleoptera	1.00	Staphylinidae	1.00	*Atheta*	1.00
OTU_43	Insecta	1.00	Coleoptera	1.00	Staphylinidae	0.78	*Atheta*	0.78
OTU_47	Insecta	1.00	Coleoptera	1.00	Staphylinidae	1.00	*Lordithon*	1.00
OTU_11	Insecta	1.00	Coleoptera	1.00	Staphylinidae	1.00	*Oxypoda*	1.00
OTU_16	Insecta	0.99	Coleoptera	0.82	Staphylinidae	0.69	*Phyllodrepoidea*	0.69
OTU_105	Insecta	1.00	Diptera	1.00	Drosophilidae	1.00	*Drosophila*	1.00
OTU_38	Insecta	1.00	Diptera	1.00	Drosophilidae	1.00	*Drosophila*	1.00
OTU_23	Insecta	1.00	Diptera	1.00	Sarcophagidae	1.00	*Helicobia*	1.00
OTU_26	Insecta	1.00	Diptera	0.99	Syrphidae	0.99	*Didea*	0.99
OTU_20	Insecta	1.00	Hymenoptera	1.00	Formicidae	1.00	*Dorymyrmex*	1.00
OTU_15	Insecta	1.00	Hymenoptera	1.00	Formicidae	1.00	*Labidus*	1.00
OTU_53	Insecta	1.00	Lepidoptera	1.00	Cossidae	1.00	*Catopta*	1.00
OTU_33	Insecta	1.00	Lepidoptera	1.00	Geometridae	1.00	*Melinodes*	1.00
OTU_22	Insecta	1.00	Mantodea	0.99	Mantidae	0.99	*Amantis*	0.99
OTU_5	Insecta	1.00	Mecoptera	0.99	Panorpidae	0.99	*Panorpa*	0.99
OTU_100	Insecta	1.00	Mecoptera	1.00	Panorpidae	1.00	*Panorpa*	1.00
OTU_24	Insecta	1.00	Orthoptera	1.00	Gryllidae	1.00	*Anaxipha*	1.00
OTU_18	Insecta	1.00	Orthoptera	1.00	Gryllidae	1.00	*Gryllus*	1.00
OTU_57	Insecta	1.00	Orthoptera	1.00	Gryllotalpidae	1.00	*Gryllotalpa*	1.00

of total reads). Among the Collembola, OTU 9 (100% support to the genus *Lepidocyrtus*) comprised 17% of total reads and was detected only on Farm I.

Beta Diversity

The Jaccard diversity matrix yielded groupings where within-farm collections were reciprocally monophyletic; i.e., they were more similar to each other than to samples from the other farm (Fig. 1). The parsimony test confirmed that the pairwise differences between collections always lead to reciprocal groupings of samples collected in either farm (P < 0.01), as did PERMANOVA based on the Jaccard dissimilarity index (F= 2.70; df=1, 10; R^2 =0.21; P <0.005). The results of NMDS ordination in two dimensions (stress = 0.08) are described in Fig. 2.

Metastats analysis identified eight OTUs that were present in significantly different quantities between Farm I and Farm II (Table 2). These included representatives from Araneae (OTU 165), Coleoptera (8, 11, 47), Orthoptera (18), Diptera (26), and Entomobryomorpha (9, 87). Five of these OTUs were exclusively detected on only one of the farms.

Alpha Diversity

Although a relatively small number of reads were sequenced, collection efforts at most sampling stations within both farms were generally sufficient, as indicated by rarefaction curves. Other than samples If, Ig, IIa, and IId, most samples' curves reached a plateau, suggesting a sufficient sampling effort was undertaken to measure biodiversity (Fig. 3).

Alpha diversity indices indicate a generally more diverse community on Farm II (Table 3), although more sequences were recovered from Farm I (note that sample sizes were standardized for these calculations). The twofold difference in phylogenetic diversity between the farms was likely due to the fact that orders Orthoptera and Blattodea were detected primarily on Farm II.

Haplotype-Based Analyses

Within the two farms sampled, we identified 706 *unique haplotypes* of sequence length above 250 bp based on the clustering of sequence reads at a 100% threshold. We were able to identify several taxonomic units (defined as a cluster of *unique haplotypes* that shared no more than 3% divergence) that occurred on both farms at frequencies sufficient to calculate population genetic parameters. Of these, haplotypes belonging to OTU 3 (genus *Xyleborus*; Table 1) were found in collection sites Ia, Ib, Ic, Id, Ig, Ih, IIa, IIb, IIc, and IId (Fig. 4).

The five OTU 3 haplotypes identified did not contain mutations leading to frameshifts or stop codons and three were detected in more than one location (Fig. 4). These two results, along with the fact that only sequence groupings containing five or more representatives were retained for the analysis, suggest that the haplotype sequences were not spurious PCR artifacts.

Although our cutoff of 97% sequence identity to describe biological species is arguably error-prone because of the nuances of intraspecific diversity, we note that these five haplotypes consistently clustered together at thresholds of 92–96%. Given that the average intraspecific divergence for closely related *Xyleborus* species shown in Fig. 4 is 3.4% (range from 0.9 to 6.4%) and that the interspecific divergence within genus *Xyleborus* averages 16.4% (range from 10.5 to 19.7%), we are satisfied that the five OTU 3 haplotypes described by the phylogenetic tree in Fig. 4 accurately encompassed a single species.

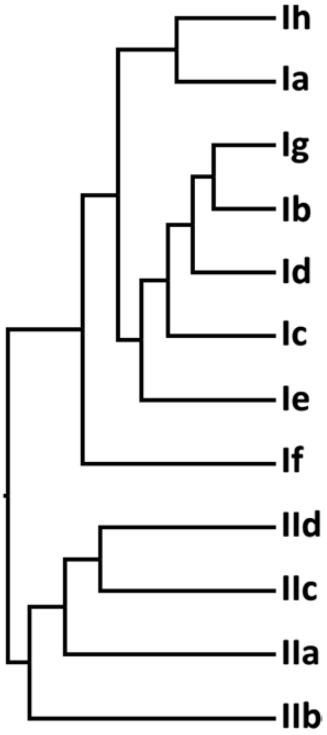

Fig. 1. Dendrograms showing the averaged relationship of Beta diversity values between each of the collection locations within the two farms sampled (using the Jaccard Beta diversity index).

The AMOVA results indicate significant population structuring at all levels evaluated: among collection sites, among collection sites within farms and between farms (F_{ST}, F_{SC}, and F_{CT}, respectively; Table 4). A relatively high percentage of the total variance is explained between the farms (26%; $P = 0.04$), which indicates that populations on either farm can be considered genetically isolated.

Discussion

A main concern of environmental managers is assessing the impact of agricultural practices upon cropping ecosystems (McLaughlin and Mineau 1995). Although this has traditionally focused on describing pest and pathogen taxa, other species are now commonly targeted, either as a requirement for socio-environmental certifications (Perfecto et al. 2005) or to assess the sustainability of such practices, e.g., promotion of biocontrol (Gurr et al. 2003, Scherr and McNeely 2008). Although traditional, molecular biology-based monitoring is well established in this sense, it has generally relied on 'single-target' protocols, where only the presence/absence of DNA from the target species in a pool is sought (e.g., Kikkert et al. 2006, Nagoshi et al. 2011). However, many crops typically have a variety of relevant interactions; a reality that NGS technology can address by bundling the search for expected (and unexpected) taxa into a single pipeline.

The current work is part of a nascent regional-scale monitoring initiative in the coffee-growing centre of Sao Paulo State. We are establishing a network of collecting stations throughout this

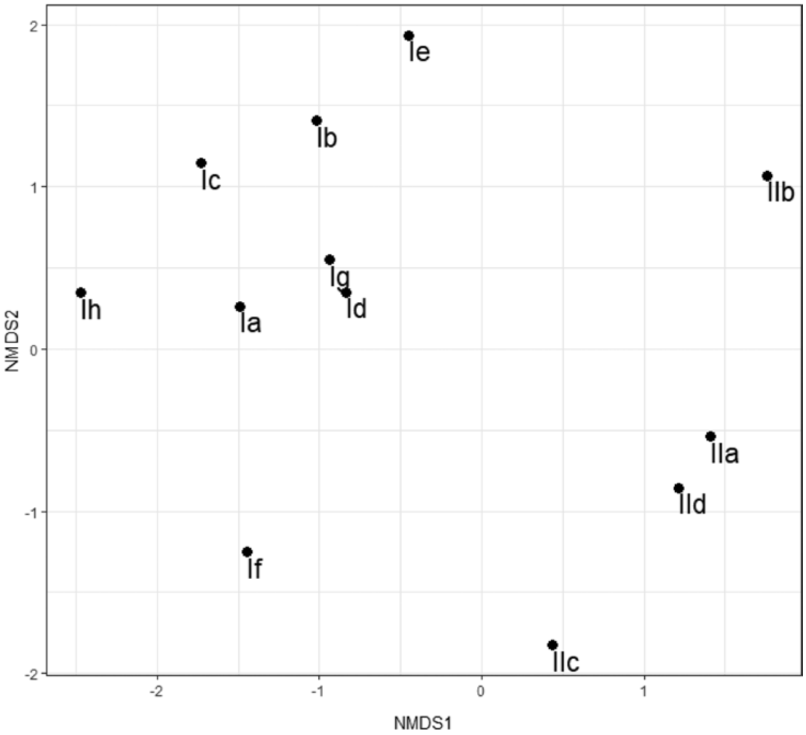

Fig. 2. NMDS ordination of invertebrate assemblages from two coffee farms in Sao Paulo State, based on the Jaccard Beta diversity index. Farm provenance is indicated by either I or II and sampling point within farms is represented by letters.

Table 2. Results of Metastats analysis identifying OTUs represented differentially between coffee farms

Name	Mean (Farm I)	Variance (Farm I)	SE (Farm I)	Mean (Farm II)	Variance (Farm II)	SE (Farm II)	*P* value	*q* value
OTU 11	12.6	1.6	4.4	0	0	0	0.002	0.005
OTU 165	0	0	0	1.2	0.001	0.2	0	0
OTU 18	0	0	0	9.8	0.8	4.4	0.017	0.029
OTU 26	0.5	0.01	0.4	5.5	0.1	1.7	0.002	0.005
OTU 47	0.1	0.0007	0.1	1.8	0.03	0.8	0.026	0.038
OTU 8	24.7	4.2	7.3	0.8	0.03	0.8	0.001	0.003
OTU 87	0	0	0	1.2	0.02	0.7	0.035	0.045
OTU 9	26.7	3.5	6.6	0	0	0	0.0002	0.001

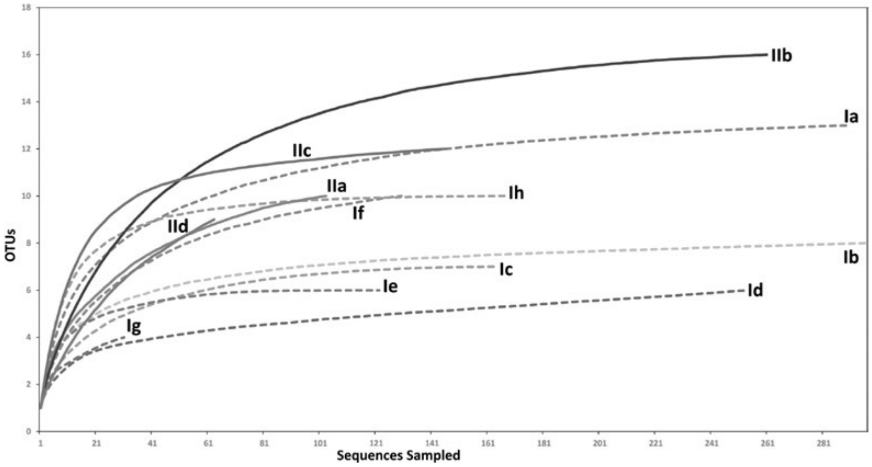

Fig. 3. Rarefaction curves for the 12 points sampled. Samples collected on Farm I are indicated as dashed lines, on Farm II as solid lines.

Table 3. Diversity values for sampling locations on both farms

Location	OTUs observed	Number of sequences	Shannon diversity	Inverse Simpson diversity	Scaled phylogenetic diversity
Farm Ia	13	290	1.91	4.87	0.0095
Farm Ib	8	296	1.51	3.75	0.0033
Farm Ic	7	164	1.1	2.13	0.0215
Farm Id	6	254	0.86	1.77	0.0061
Farm Ie	6	122	1.45	3.73	0.0024
Farm If	10	130	1.43	2.74	0.0239
Farm Ig	4	31	1	2.53	0.0120
Farm Ih	10	167	2.07	7.33	0.0068
Farm I mean (SD)	8.00 ± 2.69	181.75 ± 86.24	1.42 ± 0.40	3.61 ± 1.69	0.01 ± 0.008
Farm IIa	10	103	1.63	3.97	0.01
Farm IIb	16	261	1.69	2.91	0.03
Farm IIc	12	146	2.17	7.29	0.02
Farm IId	9	63	1.1	1.86	0.03
Farm II mean (SD)	11.75 ± 2.68	143.25 ± 74.05	1.65 ± 0.38	4.01 ± 2.04	0.02 ± 0.008

Values were calculated after correction for sample size.

agricultural landscape to track the distribution of invertebrate taxa. We are particularly interested in how these dynamics are impacted by the considerable amount of remnant forests. Here, we show that metabarcoding can attend to this demand by identifying the distribution of dozens of taxa across the landscape, robustly differentiate areas, and yield population genetic information.

Our results suggest that, although very few sequences were generated in this study, they were nonetheless sufficient to capture the biodiversity of the sampling sites. This may not be surprising given that high levels of anthropic intervention on conventional farms (as evaluated herein) generally reduce biodiversity in comparison to organic and shade coffee properties (Ibarra-Nuflez and Garcia 2001, Armbrecht and Gallego 2007).

OTU-Based Analyses

We found that within each of the two coffee farms, pitfall samples were generally homogeneous in their community structures and could be described by a few dominant sequences. Almost 70% of DNA detected from pitfall traps was represented by only five OTUs (3, 5, 8, 9, 11; Table 1). Of these, OTUs 8 (*Urophorus* sp.),

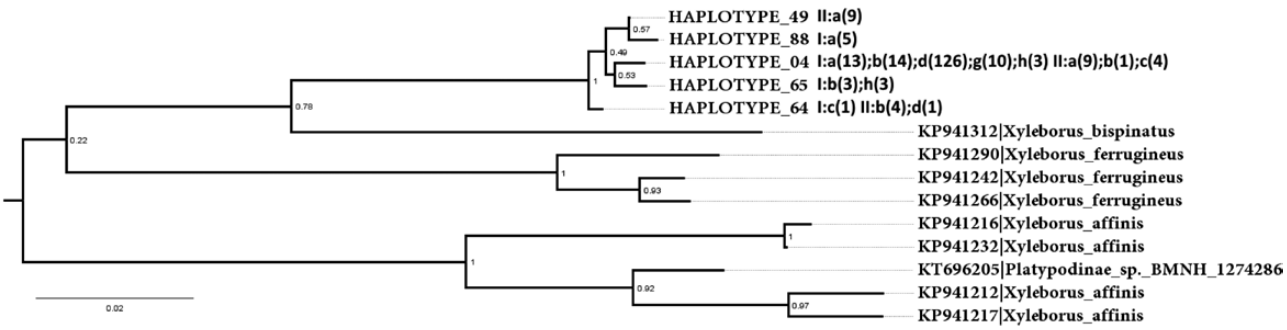
HAPLOTYPE_49 II:a(9)
HAPLOTYPE_88 I:a(5)
HAPLOTYPE_04 I:a(13);b(14);d(126);g(10);h(3) II:a(9);b(1);c(4)
HAPLOTYPE_65 I:b(3);h(3)
HAPLOTYPE_64 I:c(1) II:b(4);d(1)
KP941312|Xyleborus_bispinatus
KP941290|Xyleborus_ferrugineus
KP941242|Xyleborus_ferrugineus
KP941266|Xyleborus_ferrugineus
KP941216|Xyleborus_affinis
KP941232|Xyleborus_affinis
KT696205|Platypodinae_sp._BMNH_1274286
KP941212|Xyleborus_affinis
KP941217|Xyleborus_affinis
0.57
0.49
0.53
1
0.78
0.22
1
0.93
1
1
0.92
0.97
0.02

Fig. 4. Neighbour-joining tree of the five haplotypes identified herein for OTU 3 along with the most similar mega BLAST (Altschul et al. 1990) hits from GenBank. Bootstrap values are listed on inside nodes. Accession numbers are followed by GenBank taxonomic identifiers. For the five haplotypes detected (top clade), we list their abundance in each farm collection.

Table 4. Results from AMOVA identifying the distribution of genetic diversity among the five haplotypes of OTU 3

Source of variation	df	Sum of squares	Percentage of variation	Fixation indices	*P* value
F_{CT}	1	3.899	25.9	0.26	0.04
F_{SC}	8	7.147	37.27	0.5	<0.001
F_{ST}	196	12.882	36.82	0.63	<0.001

Note that the F_{CT} result indicates significant difference between the two farms.

9 (*Lepidocyrtus* sp.), and 11 (*Oxypoda* sp.) were differentially detected between the two farms. In light of environmental monitoring, this is a relevant distinction because these three taxa are relatively small (<0.5 cm) and would thus require extensive microscope work hours to identify morphologically (as is often the norm in biomonitoring). Rarer taxa that also occurred differentially between the farms were OTU 26 (*Didea* sp.), OTU 47 (*Lordithon* sp.), and 165 (Araneae).

Because we sampled only two properties, the differences between them cannot be attributed specifically to management practices. However, the consistency of Beta diversity within farms does suggest the variation is not random and may be a result of exogenous influences. Ecological or management parameters may also explain the consistently higher Alpha diversity in Farm II. The denser canopy and less-frequent removal of undergrowth in this hillside area may provide greater diversity of host plants and more cover for dispersal.

Haplotype-Based Analyses

The fact that metabarcoding allows the detection of distinct haplotypes in the sampled populations has important implications for agriculture and environmental management. In populations of pest species, for example, there will be variation in dispersal ability and population size, both parameters that can be estimated from DNA data. Effective characterization of individual haplotypes can also identify the provenance of introduced species, especially those that already have substantial geographic annotations in genetic databases.

An important consideration in population genetic analyses is that sufficient individuals be collected from each sample to make comparisons of relative frequencies meaningful. Although our protocol did not measure directly the number of individuals from each trap, visual inspection of contents prior to homogenization identified elevated quantities of Curculionidae beetles similar to the *Xyleborus* sp. identified herein as OTU 3. Moreover, previous trapping in Brazil has shown that hundreds of *Xyleborus* individuals are commonly caught in pitfall traps (Abreu et al. 2012). We thus feel confident that the sequencing output are indeed representative of the relative frequencies from a large number of *Xyleborus*, and thus amenable to population genetic analyses.

A second consideration in the haplotype-based analysis is accurate species delimitation for OTU 3; the potential exists that the haplotypes detected are derived from more than one species. We believe this likelihood is minimal because clustering thresholds from 97 to 92% consistently grouped the five haplotypes together. Previous data have suggested that the intraspecific variation of this genus ranges from 0 to 6.5% (Chang et al. 2014). Conversely, interspecific divergence within the *Xyleborus* averaged 23.6%. We thus concluded that the five haplotypes sampled as OTU 3 were likely from the same species.

Conclusions

The principal impediment to current ecological assemblage-scale research is the taxonomic bottleneck, which often necessitates heavy investments in the identification of collections. Herein, we show that a research project that was essentially morphologically blind (i.e., no morphological information was used in the analyses) could nonetheless provide important ecological conclusions for terrestrial invertebrates collected from pitfall traps. Moreover, the intraspecific haplotype-based analysis highlights NGS potential to track invertebrate dispersal through a landscape.

Different habitats, such as tropical forests, will undoubtedly require higher sequencing depth than was provided by the 454-FLX pyrosequencing used herein. Currently, the popular MiSeq platform produces one order of magnitude more metabarcoding data at a comparable price. Given this rapidly diminishing cost, field collection protocols are now a limiting factor in biomonitoring, particularly when such efforts require dense sampling. The pitfall traps used herein are an inexpensive means of capturing terrestrial arthropods and can be installed during several weeks, which increases the sampling window and decreases the variation that arises from punctual collections, especially in agricultural landscapes.

Acknowledgments

This research received R&D funding from ANEEL (Agência Nacional de Energia Elétrica; grant number 0064-1035/2014). Sample collections were undertaken with permit number 54835 administered by the Brazilian Federal agency SISBIO. We thank the contributions from two anonymous reviewers, who provided important considerations to this work.

Supplementary Data

Supplementary data are available at *Environmental Entomology* online.

References Cited

Abreu, R. L. S., G. A. de Ribeiro, B. F. Vianez, C. Sales-Campos, and C. Sales-Campos. 2012. Insects of the subfamily Scolytinae (Insecta: Coleoptera, Curculionidae) collected with pitfall and ethanol traps in primary forests of Central Amazonia. Psyche **2012**: 1–8.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. 215: 403–410.

Anderson, M. 2001. A new method for non parametric multivariate analysis of variance. Austral. Ecol. 26: 32–46.

Armbrecht, I., and M. C. Gallego. 2007. Testing ant predation on the coffee berry borer in shaded and sun coffee plantations in Colombia. Entomol. Exp. Appl. 124: 261–267.

Arribas, P., C. Andújar, K. Hopkins, M. Shepherd, and A. P. Vogler. 2016. Metabarcoding and mitochondrial metagenomics of endogean arthropods to unveil the mesofauna of the soil. Methods Ecol. Evol. 7: 1071–1081.

Aylagas, E., and N. Rodríguez-Ezpeleta. 2016. Analysis of Illumina MiSeq metabarcoding data: application to benthic indices for environmental monitoring. Marine Genom. 1452: 237–249.

Baird, D. J., and M. Hajibabaei. 2012. Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. Mol. Ecol. 21: 2039–2044.

Beng, K. C., K. W. Tomlinson, X. H. Shen, Y. Surget-Groba, A. C. Hughes, R. T. Corlett, and J. W. F. Slik. 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.

Brandon-Mong, G., H. Gan, and K. Sing. 2015. DNA metabarcoding of insects and allies: an evaluation of primers and pipelines. Bull. Entom. Res. 105: 717–727.

Buée, M., M. Reich, C. Murat, E. Morin, R. H. Nilsson, S. Uroz, and F. Martin. 2009. 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. New Phytol. 184: 449–456.

Carew, M. E., V. J. Pettigrove, L. Metzeling, and A. A. Hoffmann. 2013. Environmental monitoring using next generation sequencing: rapid identification of macroinvertebrate bioindicator species. Front. Zool. 10: 45.

Chang, H., Q. Liu, D. Hao, Y. Liu, Y. An, L. Qian, and X. Yang. 2014. DNA barcodes and molecular diagnostics for distinguishing introduced *Xyleborus* (Coleoptera: Scolytinae) species in China. Mitochondrial DNA 25: 63–69.

Creer, S. 2010. Second-generation sequencing derived insights into the temporal biodiversity dynamics of freshwater protists. Mol. Ecol. 19: 2829–2831.

Edgar, R. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26: 2460–2461.

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

Excoffier, L., G. Laval, and S. Schneider. 2005. Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol. Bioinforma. 1: 47–50.

Gibson, J. F., S. Shokralla, C. Curry, D. J. Baird, W. A. Monk, I. King, and M. Hajibabaei. 2015. Large-scale biomonitoring of remote and threatened ecosystems via high-throughput sequencing. PLoS ONE 10: e0138432.

Gibson, J., S. Shokralla, T. M. Porter, I. King, S. van Konynenburg, D. H. Janzen, W. Hallwachs, and M. Hajibabaei. 2014. Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. Proc. Natl Acad. Sci. USA 111: 8007–8012.

Guidolin, A. S., P. Fresia, and F. L. Cônsoli. 2014. The genetic structure of an invasive pest, the Asian citrus psyllid *Diaphorina citri* (Hemiptera: Liviidae). PLoS ONE 9: e115749.

Gurr, G. M., S. D. Wratten, and J. M. Luna. 2003. Multi-function agricultural biodiversity: pest management and other benefits. Basic Appl. Ecol. 4: 107–116.

Hajibabaei, M., D. H. Janzen, J. M. Burns, W. Hallwachs, and P. D. N. Hebert. 2006. DNA barcodes distinguish species of tropical Lepidoptera. Proc. Natl Acad. Sci. USA 103: 968–971.

Hajibabaei, M., J. L. Spall, S. Shokralla, and S. van Konynenburg. 2012. Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. BMC Ecol. 12: 28.

Huber, J. A., H. G. Morrison, S. M. Huse, P. R. Neal, M. L. Sogin, and D. B. Mark Welch. 2009. Effect of PCR amplicon size on assessments of clone library microbial diversity and community structure. Environ. Microbiol. 11: 1292–1302.

Ibarra-Nuflez, G., and J. Garcia. 2001. Prey analysis in the diet of some ponerine ants (Hymenoptera: Formicidae) and web-building spiders (Araneae) in coffee plantations in Chiapas, Mexico. Sociobiology 37: 723–755.

Johnson, P. L. F., and M. Slatkin. 2006. Inference of population genetic parameters in metagenomics: a clean look at messy data. Genome Res. 16: 1320–1327.

Kekkonen, M., M. Mutanen, L. Kaila, M. Nieminen, and P. D. N. Hebert. 2015. Delineating species with DNA barcodes: a case of taxon dependent method performance in moths. PLoS ONE 10: e0122481.

Kikkert, J. R., C. A. Hoepting, Q. Wu, P. Wang, R. Baur, and A. M. Shelton. 2006. Detection of *Contarinia nasturtii* (Diptera: Cecidomyiidae) in New York, a new pest of cruciferous plants in the United States. J. Econ. Entomol. 99: 1310–1315.

Kumar, S., G. Stecher, and K. Tamura. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33: 1870–1874.

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

Machida, R. J., M. Leray, S.-L. Ho, and N. Knowlton. 2017. Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. Sci. Data. 4: 170027.

McLaughlin, A., and P. Mineau. 1995. The impact of agricultural practices on biodiversity. Agric. Ecosyst. Environ. 55: 201–212.

Nagoshi, R. N., J. Brambila, and R. L. Meagher. 2011. Use of DNA Barcodes to identify invasive armyworm *Spodoptera* species in Florida. J. Insect Sci. 11: 1–11.

Oksanen, J. 2013. Multivariate analysis of ecological communities in R: vegan tutorial. (http://cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf), accessed 15 December 2016.

Pawlowski, J., P. Esling, F. Lejzerowicz, T. Cedhagen, and T. A. Wilding. 2014. Environmental monitoring through protist next-generation sequencing metabarcoding: assessing the impact of fish farming on benthic foraminifera communities. Mol. Ecol. Resour. 14: 1129–1140.

Perfecto, I., J. Vandermeer, A. Mas, and L. S. Pinto. 2005. Biodiversity, yield, and shade coffee certification. Ecol. Econ. 54: 435–446.

Petrosino, J. F., S. Highlander, R. A. Luna, and R. A. Gibbs. 2009. Metagenomic pyrosequencing and microbial identification. Clin. Chem. 55: 856–866.

Scherr, S., and J. McNeely. 2008. Biodiversity conservation and agricultural sustainability: towards a new paradigm of 'ecoagriculture' landscapes. Philos. Trans. R. Soc. Lond. B Biol. Sci. 363:477–494.

Schloss, P., and J. Handelsman. 2006. Introducing TreeClimber, a test to compare microbial community structures. Appl. Environ. Microbiol. 72: 2379–2384.

Schloss, P., S. Westcott, and T. Ryabin. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75: 7537–7541.

Shokralla, S., J. L. Spall, J. F. Gibson, and M. Hajibabaei. 2012. Next-generation sequencing technologies for environmental DNA research. Mol. Ecol. 21: 1794–1805.

Taberlet, P., E. Coissac, F. Pompanon, C. Brochmann, and E. Willerslev. 2012. Towards next-generation biodiversity assessment using DNA metabarcoding. Mol. Ecol. 21: 2045–2050.

Walther, G.-R., E. Post, P. Convey, A. Menzel, C. Parmesan, T. J. C. Beebee, J.-M. Fromentin, O. Hoegh-Guldberg, and F. Bairlein. 2002. Ecological responses to recent climate change. Nature 416: 389–395.

Wang, Q., G. M. Garrity, J. M. Tiedje, and J. R. Cole. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. 73: 5261–5267.

White, J. R., N. Nagarajan, and M. Pop. 2009. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput. Biol. 5: e1000352.

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