Expedited Assessment of Terrestrial Arthropod Diversity by Coupling Malaise Traps with DNA Barcoding

Jeremy R. deWaard[1]*, Valerie Levesque-Beaudin[1], Stephanie L. deWaard[1], Natalia V. Ivanova[1], Jaclyn T.A. McKeown[1], Renee Miskie[1], Suresh Naik[1], Kate H.J. Perez[1], Sujeevan Ratnasingham[1], Crystal N. Sobel[1], Jayme E. Sones[1], Claudia Steinke[1], Angela C. Telfer[1], Andrew D. Young[1,2], Monica R. Young[1], Evgeny V. Zakharov[1], and Paul D.N. Hebert[1]

[1] Centre for Biodiversity Genomics, Biodiversity Institute of Ontario, University of Guelph, Guelph, Ontario, Canada

[2] Department of Biology, Carleton University, Ottawa, Ontario, Canada

* Corresponding author; email: dewaardj@uoguelph.ca

Abstract

Monitoring changes in terrestrial arthropod communities over space and time requires a dramatic increase in the speed and accuracy of processing samples that cannot be achieved with morphological approaches. The combination of DNA barcoding and Malaise traps allows expedited, comprehensive inventories of species abundance whose cost will rapidly decline as high-throughput sequencing technologies advance. Aside from detailing protocols from specimen sorting to data release, this paper describes their use in a survey of arthropod diversity in a national park that examined 21,194 specimens representing 2,255 species. These protocols can support arthropod monitoring programs at regional, national, and continental scales.

Keywords: malaise trap, DNA barcoding, biological inventory, biomonitoring, barcode index numbers

Introduction

Given unprecedented losses (Lawton and May 1995; Pimm et al. 1995, 2014), improved methods to quantify biodiversity at a massive scale and at low cost are essential, especially for small-bodied organisms such as arthropods. The melding of two technologies – DNA barcoding and passive, large-scale specimen collection – represents a potential solution. DNA barcoding simplifies and accelerates taxonomic identifications (Hebert et al. 2003; Packer et al. 2009; Cristescu 2014; Joly et al. 2014) by employing the 6.12 million reference sequences (July 2018) in the Barcode of Life Datasystems (BOLD; Ratnasingham and Hebert 2007). Coverage of the BOLD reference library varies for geographic regions and taxonomic groups, ranging from nearly complete for some continental faunas, e.g. beetles, spiders, moths and butterflies, (Hebert et al. 2013; Pentinsaari et al. 2014; Hendrich et al. 2014; Huemer et al. 2014; Rougerie et al. 2014; Zahiri et al. 2017; Blagoev et al. 2015; Gwiazdowski et al. 2015) to sparse for many taxa, e.g. nematodes, mites and molluscs (Ferri et al. 2009; Young et al. 2012; Layton et al. 2014). Because the latter groups include many undescribed species, operational taxonomic units (OTUs) must be employed to quantify their diversity. DNA barcoding represents a dramatic advance for such analysis because the Barcode Index Number (BIN) system (Ratnasingham and Hebert 2013) provides an objective approach for OTU delineation of animals that is coupled with a persistent registry. Since BINs correspond well with Linnaean species in many animal groups (Hausmann et al. 2013; Ratnasingham and Hebert 2013; Zahiri et al. 2014; Blagoev et al. 2015), BIN-based biodiversity assessments can be implemented for groups that lack well-developed taxonomy.

The Malaise trap (Malaise 1937) has gained popularity for assessing terrestrial arthropod communities (Karlsson et al. 2005) because it collects large samples with little effort (Marshall et al. 1994). However, the subsequent identification is a substantial challenge as a week-long collection often includes more than 1000 specimens representing several hundred species. Moreover, because many species are only represented by a few specimens, it is important to identify every individual. Conversely, very common species can consume considerable effort, particularly if they

belong to a closely allied group of taxa whose members are difficult to discriminate morphologically. DNA barcoding breaks this taxonomic barrier as it can rapidly assign individuals to OTUs, streamlining the identification process.

While the analysis of bulk samples through DNA metabarcoding (Hajibabaei et al. 2011; Taberlet et al. 2012; Yu et al. 2012; Ji et al. 2013; Gibson et al. 2014; Leray and Knowlton 2015) greatly reduces analytical costs, it has two limitations. It cannot maintain the link between each specimen and its cytochrome *c* oxidase subunit I (COI) sequence, which inhibits extending the DNA barcode reference library, and cannot determine species abundances.

This study describes a protocol for rapid biodiversity assessments which employs DNA barcoding and passive specimen trapping. Its effectiveness is demonstrated by describing a survey that examined more than 20,000 specimens representing over 2200 BINs from Point Pelee National Park. This protocol has already proven both efficient and effective in several studies (Bukowski et al. 2015; D'Souza et al. 2015; Kohn et al. 2015; Mazumdar et al. 2015; Perez et al. 2015; Zlotnick et al. 2015; Aagaard et al. 2017; Geiger et al. 2016; Hebert et al. 2016; Wirta et al. 2016; Steinke et al. 2017; Ashfaq et al. 2018; D'Souza and Hebert 2018) but is described in detail here for the first time.

Materials and Methods

Specimen Collection and Processing

A Townes-style Malaise trap was deployed for 20 weeks in a cedar-savannah habitat at Point Pelee National Park in southwestern Ontario, from May 2 until September 19, 2012. Each sample was collected in a 500 mL plastic Nalgene bottle that was filled with 375 mL of 95% ethanol and then attached to the trap head (Fig. 1 F2). The catch was harvested weekly and placed in 500 mL of fresh ethanol before storage at -20°C until it was analyzed at the Centre for Biodiversity Genomics (CBG; www.biodiversitygenomics.net).

Each weekly sample was accessioned and its collection data entered into an Access-based Collection Information Management System (CIMS; Fig. 1 P1). To reduce cost, samples collected in odd-numbered weeks (1, 3, 5…) were processed while the others were archived. The first stage in sample processing involved decanting excess ethanol and pouring the specimens into a sorting dish. Specimens were then partitioned by size (small, medium and large) and assigned to a taxonomic order. Most large specimens (>5mm) were pinned, with the exception of those taxa routinely stored in ethanol (e.g., Araneae, Gastropoda); all small and medium specimens were retained in ethanol. After sorting, specimens were arrayed in batches of 95 plus one control (Fig. 1 P2), mirroring the 8 x 12 format of 96 well microplates. Typically, each array included only one order to avoid mixing of taxa requiring different primers (Table 1). Specimens of different orders were only combined when necessary to complete an array. Pinned specimens were placed in Schmitt boxes with an 8 x 12 array grid marked on their foam base, while medium specimens (≈3-5mm) were placed in Matrix storage tubes (Thermo Fisher Scientific; Fig. 1 P2), and small specimens (<3mm) were placed directly in 96-well microplates (Eppendorf; Fig. 1 P2). The sample was also inspected to determine if an excessive number (>300 specimens) of a particular morphospecies was present, and if it could be distinguished morphologically. In these cases, to reduce cost yet still capture the genetic diversity of those morphospecies, at least 24 specimens, which amounted to two rows in a microplate, were selected for barcoding while the others were counted and archived. Each container was given a unique identifier (Root Plate ID, e.g. BIOUGXXXXX) and likewise, each specimen within the container was given a unique identifier reflecting its position in it (Sample ID, e.g. BIOUGXXXXX-A01 to BIOUGXXXXX-H11). The unique identifiers and collection data for each specimen were uploaded to BOLD (Ratnasingham and Hebert 2007; Fig. 1 P3) with records for each sample placed in a separate project to allow easier comparison among weeks. Once this was completed and BOLD Process IDs were generated, labels were printed and affixed to large and medium specimens while small specimens did not require individual labels (Fig. 1 P4). A small fragment of tissue was then removed from each large and medium specimen and placed into a microplate destined for DNA extraction (Fig. 1 P5). Small specimens did not require tissue sampling as they were already in

microplates. Each microplate was then submitted for molecular processing and its progress through the analytical chain was tracked with a Laboratory Information Management System (BOLD-LIMS).

DNA Barcode Analysis

Molecular analyses were conducted at the Canadian Centre for DNA Barcoding (CCDB; www.ccdb.ca). An automated, silica membrane-based DNA extraction protocol (Ivanova et al. 2006) was performed in 96-well microplate format using a 3 µm glass fibre over 0.2 µm Bio-Inert membrane filter plate (Pall Corporation). The extraction protocol, however, was modified following Porco et al. (2010; Fig. 1 P7) to allow recovery of vouchers for microplates containing whole specimens. To maximize DNA yield, tissue lysis was performed overnight at 56°C before DNA extraction (Fig. 1 S1 and S2). Subsequent PCR amplification of the COI barcode region was performed in 384-well plate format as this allowed a 50% reduction in reagent volumes from earlier methods (Hajibabaei et al. 2005; deWaard et al. 2008, Wilson 2012). This protocol involved consolidating aliquots of DNA extracts from four 96-well microplates into a 384-well PCR plate containing PCR master mix using a Biomek FX workstation (Beckman-Coulter; Fig. 1 S3) and ensured arthropod orders were processed with the same primer pair. The total PCR reaction volume was 6 µL: 3 µL of 10% D-(+)-trehalose dihydrate for microbiology (≥99.0%; Fluka Analytical), 0.92 µL of ultra-pure water (Hyclone, Thermo Scientific), 0.60 µL of 10× PlatinumTaq buffer (Invitrogen), 0.30 µL of 50 mM $MgCl_2$ (Invitrogen), 0.06 µL (0.1 uM) of each primer, 0.03 µL of 10 mM dNTP (KAPA Biosystems), 0.03 µL of 5 U/µL PlatinumTaq DNA Polymerase (Invitrogen), and 1 µL of DNA template. Table 1 details the primer pairs used on the first pass. All PCR reactions employed the same thermocycling parameters: 94°C for 1 min, 5 cycles at 94°C for 40 sec, 45°C for 40 sec, 72°C for 1 min, followed by 35 cycles at 94°C for 40 sec, 51°C for 40 sec, 72°C for 1 min, and a final extension at 72°C for 5 min (Fig. 1 S4).

PCR products were diluted 1:4 with molecular grade water and then unidirectionally sequenced using the appropriate reverse primer (Table 1). Unidirectional sequencing (3' to 5') was also

completed in 384-well format (Fig. 1 S5) to reduce costs. The total sequencing reaction volume was 5.5 µL: 0.14 µL of BigDye terminator v3.1 (Applied Biosystems), 1.04 µL of 5X sequencing buffer [400 mM Tris-HCl pH 9.0 + 10 mM $MgCl_2$ (Invitrogen)], 2.78 µL of 10% D-(+)-trehalose dihydrate from *Saccharomyces cerevisiae* (≥99%; Sigma-Aldrich), 0.48 µL of ultra-pure water (Hyclone, Thermo Scientific), 0.56 µL (0.1 uM) of primer; and 0.5 µL of diluted PCR template was added with a Biomek FX robot. All sequencing reactions employed the same thermocycling protocol: 96°C for 1 min followed by 15 cycles at 96°C for 10 sec, 55°C for 5 sec, 60°C for 1.25 min, followed by 5 cycles at 96°C for 10 sec, 55°C for 5 sec, 60°C for 1.75 min, then 60°C for 15 sec followed by 15 cycles at 96°C for 10 sec, 55°C for 5 sec, 60°C for 2 min and a final extension at 60°C for 1 min (Fig. 1 S6). An automated, magnetic bead-based sequencing cleanup method was employed in 384-well microplates using PureSEQ (ALINE Biosciences) on a separate Biomek FX robot before sequencing on an ABI 3730xL DNA Analyzer (Applied Biosystems; Fig. 1 S7).

Trace files were manually uploaded to BOLD and were automatically assessed for quality based on predefined parameters (Ratnasingham and Hebert 2007). Trace files that received medium and high-quality assessments were automatically trimmed and edited by the BOLD platform. Those deemed low quality or classified as failed reads were ignored. Trimming was performed using a sliding window approach, discarding leading and trailing segments of the sequence that had more than 4 bp with a quality value (QV) score lower than 20 in a window of 20 bp. All sequences with less than 500 bp in the barcode region (the threshold for BIN assignment; see below) were manually edited with CodonCode v. 3.0.1 (CodonCode Corporation) to see if additional sequence information could be recovered (Fig. 1 A1). In cases where multiple trace files were generated for a single individual (see below) they were manually inspected for chimeras.

The initial PCR failed to generate an amplicon from some DNA extracts, likely reflecting DNA degradation or low primer affinity. These failures were hitpicked to assemble new destination 96-well microplates of DNA extracts (Fig. 1 S8), which were subjected to another round of PCR

employing primers that generated two shorter, overlapping COI (307 bp, 407 bp) amplicons (Table 1; Fig. 1 S9). A Biomek NX Span 8 workstation (Beckman-Coulter) was used to hitpick DNA from the failed samples into new plates. This 'failure tracking' was supported by data generated by the BOLD-LIMS. The original DNA plates were scanned to identify all specimens that failed to generate a BIN compliant sequence. The well coordinates of these failures in the source and destination microplates were generated for input into a Biomek NX robot. The newly configured microplates were then processed through two PCR reactions followed by bidirectional sequencing and manual assembly as part of the failure tracking protocol (Fig. 1 S10, S11, S12 and A4). Failure-tracking PCR reactions were carried out in 96-well microplates. The total PCR reaction volume was 12.5 µL: 6.25 µL of 10% D-(+)-trehalose dihydrate for microbiology (≥99.0%; Fluka Analytical), 0.125 µL of ultra-pure water (Hyclone, Thermo Scientific), 2.5 µL of 5× KAPA Taq HotStart Buffer (KAPA Biosystems), 1.25 µL of 25 mM $MgCl_2$ (Invitrogen), 0.125 µL of each primer, 0.0625 µL of 10 mM dNTP (KAPA Biosystems), 0.0625 µL of 5 U/µL KAPA Taq HotStart DNA Polymerase (KAPA Biosystems), and 2 µL of DNA template. Failure-tracking sequencing reactions were also carried out in 96-well microplates. PCR products were diluted 1:5 and bidirectionally sequenced. The total sequencing reaction volume was 11 µL: 0.25 µL of BigDye terminator v3.1 (Applied Biosystems), 1.875 µL of 5X sequencing buffer [400 mM Tris-HCl pH 9.0 + 10 mM $MgCl_2$ (Invitrogen)], 5 µL of 10% D-(+)-trehalose dihydrate from *Saccharomyces cerevisiae* (≥99%; Sigma-Aldrich), 0.875 µL of ultra-pure water (Hyclone, Thermo Scientific), 1 µL of primer; and 2 µL of diluted PCR template.

The final step in barcode analysis involved a second round of 'BIN hitpicking' to ensure that each BIN was represented, whenever possible, by five individuals with bidirectional sequence coverage. BIN information on BOLD was utilized in conjunction with the BOLD-LIMS to select representatives of each BIN with <5 individuals with bidirectional coverage (Fig. 1 A5) and instructions were automatically generated for the Biomek NX Span 8 workstation. The hitpicked destination DNA microplates were then processed through the PCR to bidirectional sequencing steps (Fig. 1 S8 to S12), manually edited (Fig. 1 A4) and uploaded to BOLD (Fig. 1 A2).

Data Release and Barcode Index Numbers

Specimen and sequence data are available on BOLD (Fig. 1 A2) in the public dataset DS-PPNP12 entitled "Point Pelee National Park Malaise Trap Program 2012" (http://dx.doi.org/10.5883/DS-PPNP12). The record for each specimen includes its date and locality of collection, its taxonomic assignment (see *Taxonomic Assignment and Data Analysis*), and voucher specimen details. If its barcode was recovered, the specimen record also includes trace files, quality scores, its sequence, and corresponding GenBank accession. After final validation, the specimen data were also uploaded to the Global Biodiversity Information Facility (GBIF) as a Darwin Core Archive (Wieczorek et al. 2012) via the University of Guelph's Integrated Publishing Toolkit (Robertson et al. 2014) installation and are available at http://dx.doi.org/10.15468/mbwnw9. A condensed version of the data is available in Table S1[3].

The source specimen for each sequence that met quality checks was automatically designated a BIN by the Refined Single Linkage (RESL) algorithm implemented on BOLD (Ratnasingham and Hebert 2013; Fig. 1 A3). The requirements for BIN membership are >=500 bp coverage of the barcode region between positions 70 and 700 of the BOLD alignment (Ratnasingham and Hebert 2013), <1% ambiguous bases, and the absence of a stop codon or contamination flag. Alternatively, specimens can gain BIN assignment without formal membership if the sequence is 300–500 bp and unambiguously matches an existing BIN member (i.e. no conflicts among top matches at any hierarchy level), but will not create or split BINs. RESL runs monthly on all qualifying barcode sequences (see above) in BOLD which currently totals 6.12 million specimens and 0.56 million BINs (July 2018). The BIN designations generated through this approach are transparent, reproducible, and globally accessible through DOI-designated 'BIN pages' that collate the specimen and sequence information of its members (e.g., *Danaus plexippus* http://dx.doi.org/10.5883/BOLD:AAA9566).

Archiving and Imaging

All voucher specimens are archived in the natural history collection (institution code = BIOUG) at the CBG, University of Guelph, where they are available for taxonomic study (Fig. 1 P6). Large pinned specimens were assigned to an archive location using BIOUG's CIMS and transferred to a drawer in the dry collection. Each medium-sized specimen was retained in its storage tube in the Matrix box, assigned an archive location, and stored in BIOUG's fluid collection. Small specimens were returned from the CCDB after voucher recovery (Porco et al. 2010; Fig. 1 P7), retained in their microplates, and archived in BIOUG's fluid collection. All residual DNA extracts are stored in the DNA Archive at the CBG (Fig. 1 S13), where they are available for further sequence characterization.

Once sequence analysis was complete and specimens were designated BINs, up to three representatives of each BIN were photographed to aid taxonomic validation and build a digital image library (Fig. 1 I1) by employing a database query to recognize BINs lacking an image. Specimens were photographed at high resolution and the images were made accessible through both specimen and BIN pages under Creative Commons (BY-NC-CA) license.

Taxonomic Assignment and Data Analysis

Following BIN designation, every specimen received a taxonomic assignment based upon querying BOLD (Fig. 1 A6). If the specimen's BIN contained other specimens identified to a single family, genus or species by a taxonomic expert (i.e. denoted by the identifier and/or identification method field on BOLD), it received this identification. However, if a BIN contained specimens with multiple, conflicting identifications, the specimens gained the lowest level of taxonomy without discordance. Specimens assigned to a BIN lacking expert identification were queried through the BOLD Identification Engine (http://www.boldsystems.org/index.php/IDS_OpenIdEngine) If the result was a close match (<10% divergence for family, <5% for genus, e.g. Coddington et al 2016) and the query sequence fell within a cluster of BINs assigned to a particular genus or family in the taxon ID tree (see below), the record was assigned to this taxon. All assignments were further validated using the

taxon ID tree (Fig. S1)[3] along with matching specimen images (Fig. S2)[3]. Any anomalies in tree topology were investigated by retrieving the vouchered specimen and ensuring that all ancillary data on BOLD were correct (including the specimen image and preliminary identification). If the sequence was revealed as representing a contamination event, it was flagged, tagged on BOLD as a contamination, and removed from the analysis and its BIN page.

The final stage of the workflow involved report generation (Fig. 1 A7) aided by the varied functions on BOLD for calculating summary statistics. As well, supplementary analyses were performed to demonstrate the utility of the protocol for rapid biodiversity assessment. To explore the completeness of the inventory, sample- (with each weekly catch considered a sample) and individual-based BIN accumulation curves were computed using the software product R, version 3.1.1 (R Development Core Team) and the vegan package (Oksanen et al. 2013). The curves were computed as the mean of 1000 randomized BIN accumulation curves without replacement. As another measure of completeness, log-normal abundance plots were calculated using R and the package vegan. These software programs were also used to estimate total BIN richness for both sample- and individual-based data using the nonparametric incidence-based species richness estimator Chao 2 (Chao 1987). We summarized the number of specimens and BINs captured for each order and in each weekly sample, along with relative abundance, the incidence of unique and rare BINs, and the turnover of BINs among samples and across time. Finally, we compared our DNA barcode-based inventory to a 40-year (1970 – 2009) morphological inventory from Point Pelee National Park (Marshall et al. 2009), and combined these two inventories to generate a more comprehensive checklist for the park.

Results

DNA Barcode Analysis

[3] Supplementary data are available with the article through the journal Web site

All specimens in the ten weekly samples were processed except for three abundant morphospeices, each from a different sample (week 3: 8,595 specimens of a chironomid; week 5: 313 specimens of a chironomid; week 9: 334 specimens of a trombidiform mite), which were excluded from the analysis. In total, 21,194 specimens were processed from the ten samples with first pass analysis generating successful sequences (i.e. > 0 bp) from 81.6% of them (17,300; Fig. 2). The second pass analysis recovered another 1885 sequences, bringing the success rate to 90.5% (19,185; Fig. 2). Aside from these records, 144 sequences were found to be contaminants and another eight possessed stop codons (Fig. 2). Sequence recovery varied among taxa with Acari displaying the lowest success (chi-square test, p< 0.0001) with just 48.0% of specimens generating a BIN compliant sequence. There was also evidence of a taxonomic bias (chi-square test, p<0.0001) in the 309 (1.6%) specimens that were either destroyed or unrecoverable after analysis, with most being small, soft-bodied Hemiptera (104 specimens, 33.7%), Diptera (75 specimens, 24.3%) and Acari (67 specimens, 21.7%).

Specimen and BIN Analyses

Among the specimens that generated a sequence, most (99.4%) received a BIN designation (n = 19,071) (Fig. 2). From these specimens with BINs, 2,043 specimens represented new BINs on BOLD (at the time of analysis) and were 'BIN hitpicked' to acquire a bidirectional sequence and 3,662 specimens were imaged (mean = 1.6 images/BIN).The 114 sequences that failed to meet the criteria for BIN designation were run through the stand-alone version of the RESL algorithm (using the function 'Cluster sequences' on BOLD) to estimate the number of additional OTUs (or species) represented; this analysis revealed 65 OTUs. One representative of each OTU was queried against the BOLD ID Engine: 49 were highly similar (p-distance > 97.8%) and matched to known BINs while 16 appeared to be new to BOLD, as they were less similar to known BINs (p-distance < 97.8%).

All subsequent analyses considered the 19,071 specimens with a BIN designation. They included taxa belonging to four classes and 25 orders (Fig. 3, Table S2). Diptera were dominant comprising

57.0% of the specimens (Fig. 3a) and 49.7% of the BINs (Fig. 3b). Hymenoptera was also very diverse with the third highest percentage of specimens (11.3%) and the second highest proportion of BINs (25.3%).

In total, 2,255 BINs were present in the ten samples with an average of 458 BINs and 1,907 specimens per sample (BIN range = 253–640, specimen range: 814–3,795) (Fig. 4, Table S3). Most BINs were uncommon; 47.6% (1,074) were represented by a single specimen while only 36 (1.6%) had >100 specimens (Fig. 5). There was a positive correlation between the number of individuals in a sample and the number of BINs unique to it ($R^2 = 0.69$, $p = 0.003$, Fig. 6), reinforcing the prevalence of rare BINs and the effort required to discover them.

Species Richness and Turnover

Species richness extrapolation based on the (Preston) log-normal species distribution indicated that complete sampling of the Malaise-trappable arthropod fauna at this site in Point Pelee would reveal about 5,700 BINs, roughly double the observed number (Fig. 7). A similar result (6,161 BINs) was obtained when the analysis was repeated with the specimen totals for the three excluded BINs (see above). BIN accumulation curves based on Chao 2 suggested a lower count with an estimate of 3,836 (SE ± 133) BINs based on specimens (Fig. 8a) and 3,889 (SE ± 125) based on samples (Fig. 8b). These three estimators suggest the site inventory is roughly 36.6–58.8% complete.

Individual samples contained an average of 458 BINs, but their similarity was low (mean shared BINs = 0.33; mean Jaccard index = 0.16) (Table S4). The proportion of shared BINs (for adjacent and non-adjacent weekly samples) increased as the season progressed (Fig. 9a) and decreased with the interval between samples ($R^2 = 0.52$, $p << 0.001$, Fig. 9b) with similarity values (Jaccard index) halved in 81.1 days. For example, only 99 BINs were shared between weeks 1 and 19, samples that contained 461 and 486 BINs, respectively. By comparison, samples from weeks 7 and 9 (containing 641 and 619 BINs respectively) shared 266 BINs.

Taxon Diversity and Abundance

The Cecidomyiidae (351), Ichneumonidae (127) and Chironomidae (113) included the most BINs while the Chironomidae (10,827), Cicadellidae (3,070), and Cecidomyiidae (1,919) were represented by the most specimens. The most abundant BINs were BOLD:AAG2868 (Cicadellidae: *Empoasca fabae*), BOLD:AAB7030 (Chironomidae: *Chironomus* sp.) and BOLD:AAV0161 (Cicadellidae: *Erythroneura bakeri*) with 555, 446 and 431 specimens respectively. Each of these species and many of the abundant taxa had closely-related allies, often morphologically indistinguishable and in low frequency, making oversampling unavoidable without risking the oversight of some species.

New and Existing Inventories

Three quarters of the specimens (n = 14,313/19,071) with a sequence gained a genus- or species-level taxonomic assignment following their comparison with records on BOLD. They represented 58.6% of all BINs (n = 1320); the other BINs were assigned to a subfamily or family. A few species were represented by more than one BIN [e.g., Araneae: Thomisidae: *Xysticus pellax* was represented by BINs BOLD:ACE4932 and BOLD:ACE4935], but most species (95.5%) showed perfect correspondence between a single taxon name and a single BIN.

By comparison, a 40-year (1970 – 2009) inventory using morphology (Marshall et al. 2009) revealed 2,423 taxa identified to a genus- or species-level among 30,000 specimens collected from Point Pelee and vicinity. After merging the two inventories, there were 3,217 genera/species combinations in the checklist with just 7.8% overlap (Table S5, doi:10.5883/DS-PPNP12). The overall taxonomic coverage includes 343 families, 597 subfamilies, 1,783 genera, 2,290 species and another 118 interim or uncertain species. While the study by Marshall et al. (2009) only examined insects, the present study examined four classes of arthropods. Only considering insects, the present inventory revealed more species of Trichoptera, Thysanoptera, and Psocodea. When all BINs are considered, the present inventory was biased toward Diptera and Hymenoptera where it collected 19.8% and

13.0% more respectively. By contrast, the diverse collecting methods employed by Marshall et al. (2009) yielded more Coleoptera, Hemiptera and Lepidoptera (64.6%, 31.6%, 22.2%). In total, the present effort added 780 taxonomic records to the checklist (Table S5; doi:10.5883/DS-PPNP12) which included 523 new species, 396 new genera, 91 new subfamilies, and 86 new families.

Discussion

This paper describes the steps involved in moving from specimen collection through DNA barcode analysis to a summary of species, their abundances and associated diversity metrics. Aside from enabling a rapid, inexpensive assessment of terrestrial arthropod diversity, this approach aids extension of the DNA barcode reference library.

Capturing Presence and Abundance

The current pipeline overcomes several barriers that usually constrain Malaise trap surveys of arthropod diversity. Most importantly, DNA barcoding minimizes the time demand on taxonomic experts by automating the identification of specimens that belong to species in the reference library (deWaard et al. 2009, Telfer et al. 2015). As a consequence, taxonomic advice is only required when a new BIN is encountered or when a BIN contains conflicting information. The use of BINs also streamlines barcode workflows. For example, imaging representatives of each BIN facilitates the detection of contamination and mis-identification, but also the assignment of taxonomy at higher levels (e.g. Order, Family). Similarly, a carefully edited bidirectional sequence is required for each new BIN, but a unidirectional sequence is perfectly adequate for BIN assignment since intraspecific variation within a population is low (Bergsten et al. 2012). Sequencing error rates are also expected to be lower than intraspecific variability, making the unidirectional BIN assignment a great option in the vast majority of cases. For instance, two BINs of *Empoasca* (Hemiptera: Cicadellidae) were represented in the Point Pelee collection by 555 (BOLD:AAG2868) and one (BOLD:ACZ4093) specimens respectively. Just a few representatives of the abundant BIN were imaged and bidirectionally sequenced, but every specimen could be identified by unidirectional analysis. Aside

from allowing the strategic deployment of analytical effort, the key advantage of DNA barcoding lies in its capacity to allow technicians with no taxonomic training to generate the species abundance data needed for most diversity indices (Magurran 2004). As well, abundance data are valuable to employ functional traits to quantify ecosystem processes and services (e.g., Devictor et al. 2010). In addition, abundance data coupled with sequence information on each specimen allows genetic diversity to be quantified (Miraldo et al. 2016), which enables follow-up examinations such as probing the correlation between species richness and genetic diversity (Vellend 2005).

Assembly of Resources

As evidenced by our study at Point Pelee, this approach generates a taxonomic inventory, an image library, a DNA archive, sequence data and specimens with associated collection data; information that can be shared through diverse online portals (e.g. Telfer et al. 2015). It also expands the DNA barcode reference library with an alternate approach that complements the analysis of legacy specimens that is complicated by degraded DNA (Hebert et al. 2013; Prosser et al. 2016). As well, the analysis of newly collected specimens permits supplemental investigations, such as genome size determination (Hanner and Gregory 2007) and stable isotope analysis (Dittrich et al. 2017). The barcode library has utility beyond species identification, including the reconstruction of community phylogenies (e.g. Boyle and Adamowicz 2015) for studying the structure and assembly of biological communities, as well as for flagging new species (e.g. van Nieukerken et al. 2015) and new occurrence records (Fernandez-Triana et al. 2014).

Protocol Use and Refinements

The present method has gained wide adoption (Perez et al. 2015, www.globalMalaise.org; Zlotnick et al. 2015; Steinke et al. 2017) and has been employed in several studies (Bukowski et al. 2015; D'Souza et al. 2015; Kohn et al. 2015; Mazumdar et al. 2015; Aagaard et al. 2017; Geiger et al. 2016; Hebert et al. 2016; Wirta et al. 2016; Ashfaq et al. 2018; D'Souza and Hebert 2018). As of July 2018, 3.1 million specimens have now been processed using this method. Large core facilities are

best-suited for the high-throughput execution of this method — where the front-end processing, laboratory analysis and informatics workflows are supported under one roof. However, this detailed protocol can also guide smaller-scale projects and facilities that can partition the workflow into sections that can be done 'in-house', and those contracted out, such as the sequencing component.

This work has led to one important modification — a standard primer cocktail, C_LepFolF and C_LepFolR (Folmer et al. 1994; Hebert et al. 2004) that can be used for all arthropods, simplifying consolidation and sequencing. The present protocol generates high quality barcode records for approximately $5 a specimen with about two thirds of the cost derived from Sanger sequencing. A substantial reduction in analytical costs can be achieved by shifting to a high-throughput sequencing (HTS) platform that allows samples to be individually tagged and subsequently multiplexed; the CBG has recently integrated the PacBio Sequel System for this purpose (Hebert et al. 2018). The Illumina MiSeq and Ion S5 platforms reduce sequencing costs four-fold (e.g. Shokralla et al. 2014, 2015, Meier et al. 2016, Morinière et al. 2016) while the Sequel System reduces them 40-fold (Hebert et al. 2018). Although HTS platforms are frequently associated with increased error rates compared to Sanger technology (Kircher and Kelson 2010), these rates can be reduced to a comparable level given sufficient read depth per specimen (see Hebert et al. 2018).

A Global Terrestrial Arthropod Monitoring Network?

The deployment of an extensive network of Malaise traps is relatively inexpensive, as evidenced by past deployments in national parks (Perez et al. 2015), schoolyards (Steinke et al. 2017), and backyards (Zlotnick et al. 2015). Once the present approach has been integrated with HTS, the mass samples resulting from a broad trap network will deliver accurate occurrence data while extending the barcode reference library. By monitoring biodiversity on a massive scale, this activity would advance each country's capacity to deliver factually-based reports on the status of biodiversity as required to meet the Convention on Biological Diversity's Aichi Targets of the Strategic Plan for Biodiversity 2011–2020 (https://www.cbd.int/sp/targets/).

Acknowledgements

The Ontario Ministry of Research and Innovation enabled this study through grants in support of the International Barcode of Life project while the Canada Foundation for Innovation provided essential infrastructure to the Centre for Biodiversity Genomics (CBG, www.biodiversitygenomics.net). We particularly thank Anne McCain Evans and Chris Evans for generously supporting our research program. Our work depended heavily on analytical support provided by the Barcode of Life Data Systems (BOLD, www.boldsystems.org). We thank John Waithaka, Heidi Brown, Tammy Dobbie and other Parks Canada staff who facilitated both permit acquisition and specimen collections. We also thank colleagues at the CBG including S. Bateson, G. Blagoev, A. Borisenko, V. Campbell, C. Christopoulos, J. Gleason, K. Hough, L. Lu, R. Manjunath, M. Milton, S. Pedersen, S. Prosser, J. Robertson, D. Roes, D. Steinke, A. Stoneham, J. Topan, C. Warne, and C. Wei.

References

Aagaard, K., Berggren, K., Hebert, P.D.N., Sones, J., McClenaghan, B., and Ekrem, T. 2017. Investigating suburban micromoth diversity using DNA barcoding of Malaise trap samples. Urban Ecosyst. **20**(2)**:** 353-361. doi:10.1007/s11252-016-0597-2.

Ashfaq, M., Sabir, J.S.M., El-Ansary, H.O., Perez, K., Levesque-Beaudin, V., Khan, A.M., et al. 2018. Insect diversity in the Saharo-Arabian region: Revealing a little-studied fauna by DNA barcoding. PLoS ONE **13**(7): e0199965. doi:10.1371/journal.pone.0199965

Bergsten, J., Bilton, D.T., Fujisawa, T., Elliott, M., Monaghan, M.T., Balke, M., et al. 2012. The effect of geographical scale of sampling on DNA barcoding. Syst. Biol. **61**(5)**:** 851-869. doi:10.1093/sysbio/sys037.

Blagoev, G.A., deWaard, J.R., Ratnasingham, S., deWaard, S.L., Lu, L.Q., Robertson, J., et al. 2016. Untangling taxonomy: a DNA barcode reference library for Canadian spiders. Mol. Ecol. Resour. **16**(1)**:** 325-341. doi:10.1111/1755-0998.12444.

Boyle, E.E., and Adamowicz, S.J. 2015. Community phylogenetics: Assessing tree reconstruction methods and the utility of DNA barcodes. PLoS One, **10**(6)**:** 18. doi:10.1371/journal.pone.0126662.

Bukowski, B., Hanisch, P.E., Tubaro, P.L., and Lijtmaer, D.A. 2015. First results of the global Malaise trap program in Argentina: strikingly high biodiversity in the southern extreme of the Atlantic forest. Genome, **58**(5)**:** 202.

Chao, A. 1987. Estimating the population-size for capture recapture data with unequal catchability. Biometrics, **43**(4)**:** 783-791. doi:10.2307/2531532.

Colwell, R.K. 2013. EstimateS: Statistical estimation of species richness and shared species from samples. Version 9.1. Available from http://viceroy.eeb.uconn.edu/estimates.

Cristescu, M.E. 2014. From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. Trends Ecol. Evol. **29**(10)**:** 566-571. doi:10.1016/j.tree.2014.08.001.

Devictor, V., Mouillot, D., Meynard, C., Jiguet, F., Thuiller, W., and Mouquet, N. 2010. Spatial mismatch and congruence between taxonomic, phylogenetic and functional diversity: the need for integrative conservation strategies in a changing world. Ecol. Lett. **13**(8)**:** 1030-1040. doi:10.1111/j.1461-0248.2010.01493.x.

D'Souza M.L. and Hebert, P.D.N. 2018. Stable baselines of temporal turnover underlie high beta diversity in tropical arthropod communities. Mol. Ecol. **27**(10): 2447-2460. doi:10.1111/mec.14693.

deWaard, J.R., Ivanova, N.V., Hajibabaei, M., and Hebert, P.D.N. 2008. Assembling DNA barcodes. Analytical protocols. Methods Mol. Biol. **410:** 275-293.

deWaard, J.R., Landry, J.F., Schmidt, B.C., Derhousoff, J., McLean, J.A., and Humble, L.M. 2009. In the dark in a large urban park: DNA barcodes illuminate cryptic and introduced moth species. Biodivers. Conserv. **18**(14)**:** 3825-3839. doi:10.1007/s10531-009-9682-7.

Coddington, J.A., Agnarsson, I., Cheng, R.C., Candek, K., Driskell, A., Frick, H., Gregoric, M., Kostanjsek, R., Kropf, C., Kweskin, M., Lokovsek, T., Pipan, M., Vidergar, N., Kuntner, M. 2016. DNA barcode data accurately assign higher spider taxa. PeerJ, **4**: e2201. doi:10.7717/peerj.2201.

Dittrich, C., Struck, U., and Rodel, M.O. 2017. Stable isotope analyses-A method to distinguish intensively farmed from wild frogs. Ecol. Evol. **7**(8)**:** 2525-2534. doi:10.1002/ece3.2878.

D'Souza, M. 2015. Investigating terrestrial arthropod biodiversity in a tropical ecosystem using barcode index numbers and phylogenetic community structure. Genome, **58**(5)**:** 208-209.

Fernandez-Triana, J.L., Penev, L., Ratnasingham, S., Smith, M.A., Sones, J., Telfer, A., et al. 2014. Streamlining the use of BOLD specimen data to record species distributions: a case study with ten Nearctic species of Microgastrinae (Hymenoptera: Braconidae). Biodivers. Data J. **2:** e4153. doi:10.3897/BDJ.2.e4153.

Ferri, E., Barbuto, M., Bain, O., Galimberti, A., Uni, S., Guerrero, R., et al. 2009. Integrated taxonomy: traditional approach and DNA barcoding for the identification of filarioid worms and related parasites (Nematoda). Front. Zool. **6:** 12. doi:10.1186/1742-9994-6-1.

Folmer, O., Black, M., Hoeh, W., Lutz, R., and Vrijenhoek, R. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Mol. Marine Biol. Biotechnol**. 3**(5)**:** 294-299.

Geiger, M.F., Moriniere, J., Hausmann, A., Haszprunar, G., Wagele, W., Hebert, P.D.N., et al. 2016. Testing the global Malaise trap program: How well does the current barcode reference library identify flying insects in Germany? Biodivers. Data J. **4:** e10671. doi: 10.3897/BDJ.4.e10671.

Gibson, J., Shokralla, S., Porter, T.M., King, I., van Konynenburg, S., Janzen, D.H., et al. 2014. Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. Proc. Natl. Acad. Sci. U. S. A. **111**(22)**:** 8007-8012. doi:10.1073/pnas.1406468111.

Gwiazdowski, R.A., Foottit, R.G., Maw, H.E.L., and Hebert, P.D.N. 2015. The Hemiptera (Insecta) of Canada: Constructing a reference library of DNA barcodes. PLoS One, **10**(4)**:** 20. doi:10.1371/journal.pone.0125635.

Hajibabaei, M., deWaard, J.R., Ivanova, N.V., Ratnasingham, S., Dooh, R.T., Kirk, S.L., et al. 2005. Critical factors for assembling a high volume of DNA barcodes. Philos. Trans. R. Soc. B-Biol. Sci. **360**(1462)**:** 1959-1967. doi:10.1098/rstb.2005.1727.

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G.A.C., and Baird, D.J. 2011. Environmental barcoding: A next-generation sequencing approach for biomonitoring applications using river benthos. PLoS One, **6**(4)**:** 7. doi:10.1371/journal.pone.0017497.

Hanner, R.H., and Gregory, T.R. 2007. Genomic diversity research and the role of biorepositories. Cell Preserv. Technol. **5**(2)**:** 93-103. doi:10.1089/cpt.2007.9993.

Hausmann, A., Godfray, H.C.J., Huemer, P., Mutanen, M., Rougerie, R., van Nieukerken, E.J., et al. 2013. Genetic patterns in european Geometrid moths revealed by the barcode index number (BIN) system. PLoS One, **8**(12)**:** 11. doi:10.1371/journal.pone.0084518.

Hebert, P.D.N., Braukmann, T.W.A., Prosser, S.W.J., Ratnasingham, S., deWaard, J.R., Ivanova, N.V., et al. 2018. A Sequel to Sanger: amplicon sequencing that scales. BMC Genomics, **19:** 14. doi:10.1186/s12864-018-4611-3.

Hebert, P.D.N., Cywinska, A., Ball, S.L., and deWaard, J.R. 2003. Biological identifications through DNA barcodes. Proc. R. Soc. B-Biol. Sci. **270**(1512)**:** 313-321. doi:10.1098/rspb.2002.2218.

Hebert, P.D.N., deWaard, J.R., Zakharov, E.V., Prosser, S.W.J., Sones, J.E., McKeown, J.T.A., et al. 2013. A DNA 'barcode blitz': Rapid digitization and sequencing of a natural history collection. PLoS One, **8**(7)**:** 14. doi:10.1371/journal.pone.0068535.

Hebert, P.D.N., Penton, E.H., Burns, J.M., Janzen, D.H., and Hallwachs, W. 2004. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. Proc. Natl. Acad. Sci. U. S. A. **101**(41)**:** 14812-14817. doi:10.1073/pnas.0406166101.

Hebert, P.D.N., Ratnasingham, S., Zakharov, E.V., Telfer, A.C., Levesque-Beaudin, V., Milton, M.A., et al. 2016. Counting animal species with DNA barcodes: Canadian insects. Philos. Trans. R. Soc. B-Biol. Sci. **371**(1702)**:** 10. doi:10.1098/rstb.2015.0333.

Hendrich, L., Moriniere, J., Haszprunar, G., Hebert, P.D.N., Hausmann, A., Kohler, F., et al. 2015. A comprehensive DNA barcode database for central European beetles with a focus on Germany: Adding more than 3500 identified species to BOLD. Mol. Ecol. Resour. **15**(4)**:** 795-818. doi:10.1111/1755-0998.12354.

Huemer, P., Mutanen, M., Sefc, K.M., and Hebert, P.D.N. 2014. Testing DNA Barcode Performance in 1000 species of european Lepidoptera: Large geographic distances have small genetic impacts. PLoS One, **9**(12)**:** 21. doi:10.1371/journal.pone.0115774.

Ivanova, N.V., deWaard, J.R., and Hebert, P.D.N. 2006. An inexpensive, automation-friendly protocol for recovering high-quality DNA. Mol. Ecol. Notes, **6**(4)**:** 998-1002. doi:10.1111/j.1471-8286.2006.01428.x.

Ji, Y.Q., Ashton, L., Pedley, S.M., Edwards, D.P., Tang, Y., Nakamura, A., et al. 2013. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Ecol. Lett. **16**(10)**:** 1245-1257. doi:10.1111/ele.12162.

Joly, S., Davies, T.J., Archambault, A., Bruneau, A., Derry, A., Kembel, S.W., et al. 2014. Ecology in the age of DNA barcoding: The resource, the promise and the challenges ahead. Mol. Ecol. Resour. **14**(2)**:** 221-232. doi:10.1111/1755-0998.12173.

Karlsson, D., Pape, T., Johanson, K.A., Liljeblad, J., and Ronquist, F. 2005. The Swedish Malaise trap project, or how many species of Hymenoptera and Diptera are there in Sweden? Entomologisk Tidskrift, **126**(1-2)**:** 43-53.

Kircher, M., Kelso, J. 2010. High-throughput DNA sequencing - concepts and limitations. Bioessays, **32**, 524-536.

Kohn, J., O'Connor, D., Danoff-Burg, J., Henter, H., and Zlotnick, B. 2015. Barcoding a biodiversity hotspot: Malaise-trapped insects of Southern California. Genome, **58**(5)**:** 238.

Lawton, J.H., and May, R.M (*Editors*). 1995. Extinction rates. Oxford University Press, Oxford, N.Y. xii, pp 1-233.

Layton, K.K.S., Martel, A.L., and Hebert, P.D.N. 2014. Patterns of DNA barcode variation in Canadian marine Molluscs. PLoS One, **9**(4)**:** 9. doi:10.1371/journal.pone.0095003.

Leray, M., and Knowlton, N. 2015. DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. Proc. Natl. Acad. Sci. U. S. A. **112**(7)**:** 2076-2081. doi:10.1073/pnas.1424997112.

Magurran, A.E. 2004. Measuring biological diversity. Blackwell Publishing, Malden, MA.

Malaise, R. 1937. A new insect-trap. Entomologisk Tidskrift, **58:** 148-160.

Marshall, S.A., Anderson, R.S., Roughley, R.E., Behan-Pelletier, V., and Danks, H.V. 1994. Terrestrial arthropod biodiversity: planning a study and recommended sampling techniques. A brief prepared by the Biological Survey of Canada (Terrestrial Arthropods). Bull. Ent. Soc. Can. **26**(1): Supplement: 1–33.

Marshall, S., Paiero, S., and Buck, M. 2009. Point Pelee national park species list [online]. Available from http://www.uoguelph.ca/debu/pelee_specieslist.htm. [accessed 1 June 2017].

Mazumdar, S., Hebert, P.D.N., and Bhuiya, B.A. 2015. Biodiversity study of Bangladeshi parasitoid wasps (Insecta: Hymenoptera) of Malaise trap collections using DNA barcoding techniques. Genome, **58**(5)**:** 254.

Meier, R., Wong, W.H., Srivathsan, A., and Foo, M.S. 2016. $1 DNA barcodes for reconstructing complex phenomes and finding rare species in specimen-rich samples. Cladistics, **32**(1)**:** 100-110. doi:10.1111/cla.12115.

Miraldo, A., Li, S., Borregaard, M.K., Florez-Rodriguez, A., Gopalakrishnan, S., Rizvanovic, M., et al. 2016. An Anthropocene map of genetic diversity. Science, **353**(6307)**:** 1532-1535. doi:10.1126/science.aaf4381.

Moriniere, J., de Araujo, B.C., Lam, A.W., Hausmann, A., Balke, M., Schmidt, S., et al. 2016. Species identification in Malaise trap samples by DNA Barcoding based on NGS technologies and a scoring matrix. PLoS One, **11**(5)**:** 14. doi:10.1371/journal.pone.0155497.

Nieukerken, E.J., Doorenweerd, C., Mutanen, M., Landry, J.F., Miller, J., and deWaard, J.R. 2015. A great inventory of the small: Combining BOLD datamining and focused sampling hugely increases knowledge of taxonomy, biology, and distribution of leafmining pygmy moths (Lepidoptera: Nepticulidae). Genome, **58**(5)**:** 262.

Oksanen, J., Blanchet, F., Kindt, R., Legendre, P., Minchin, P., O'Hara, R., et al. 2013. Vegan: Community ecology package. R package version 2.0-10. Available from http://CRAN.R-project.org/package=vegan.

Packer, L., Gibbs, J., Sheffield, C., and Hanner, R. 2009. DNA barcoding and the mediocrity of morphology. Mol. Ecol. Resour. **9:** 42-50. doi:10.1111/j.1755-0998.2009.02631.x.

Pentinsaari, M., Hebert, P.D.N., and Mutanen, M. 2014. Barcoding beetles: A regional survey of 1872 species reveals high identification success and unusually deep interspecific divergences. PLoS One, **9**(9)**:** 8. doi:10.1371/journal.pone.0108651.

Perez, K.H.J., Sones, J.E., deWaard, J.R., and Hebert, P.D.N. 2015. The global Malaise program: Assessing global biodiversity using mass sampling and DNA barcoding. Genome, **58**(5)**:** 266.

Pimm, S.L., Russell, G.J., Gittleman, J.L., and Brooks, T.M. 1995. The future of biodiversity. Science, **269**(5222)**:** 347-350. doi:10.1126/science.269.5222.347.

Pimm, S.L., Jenkins, C.N., Abell, R., Brooks, T.M., Gittleman, J.L., Joppa, L.N., et al. 2014. The biodiversity of species and their rates of extinction, distribution, and protection. Science, **344**(6187): 987,1246752-1-10. doi:10.1126/science.1246752.

Porco, D., Rougerie, R., Deharveng, L., and Hebert, P. 2010. Coupling non-destructive DNA extraction and voucher retrieval for small soft-bodied Arthropods in a high-throughput context: The example of Collembola. Mol. Ecol. Resour. **10**(6)**:** 942-945. doi:10.1111/j.1755-0998.2010.02839.x.

Prosser, S.W.J., deWaard, J.R., Miller, S.E., and Hebert, P.D.N. 2016. DNA barcodes from century-old type specimens using next-generation sequencing. Mol. Ecol. Resour. **16**(2)**:** 487-497. doi:10.1111/1755-0998.12474.

Ratnasingham, S., and Hebert, P.D.N. 2007. BOLD: The barcode of life data system (www.barcodinglife.org). Mol. Ecol. Notes, **7**(3)**:** 355-364. doi:10.1111/j.1471-8286.2006.01678.x.

Ratnasingham, S., and Hebert, P.D.N. 2013. A DNA-based registry for all animal species: The barcode index number (BIN) system. PLoS One, **8**(7)**:** 16. doi:10.1371/journal.pone.0066213.

Robertson, T., Doring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., et al. 2014. The GBIF integrated publishing toolkit: Facilitating the efficient publishing of biodiversity data on the internet. PLoS One, **9**(8)**:** 7. doi:10.1371/journal.pone.0102623.

Rougerie, R., Kitching, I.J., Haxaire, J., Miller, S.E., Hausmann, A., and Hebert, P.D.N. 2014. Australian Sphingidae - DNA barcodes challenge current species boundaries and distributions. PLoS One, **9**(7)**:** 12. doi:10.1371/journal.pone.0101108.

Shokralla, S., Gibson, J.F., Nikbakht, H., Janzen, D.H., Hallwachs, W., and Hajibabaei, M. 2014. Next-generation DNA barcoding: Using next-generation sequencing to enhance

and accelerate DNA barcode capture from single specimens. Mol. Ecol. Resour. **14**(5)**:** 892-901. doi:10.1111/1755-0998.12236.

Shokralla, S., Porter, T.M., Gibson, J.F., Dobosz, R., Janzen, D.H., Hallwachs, W., et al. 2015. Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. Sci. Rep. **5:** 7. doi:10.1038/srep09687.

Steinke, D., Breton, V., Berzitis, E., and Hebert, P.D.N. 2017. The school Malaise trap program: Coupling educational outreach with scientific discovery. PLoS Biol. **15**(4)**:** e2001829. doi:10.1371/journal.pbio.2001829.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., and Willerslev, E. 2012. Towards next-generation biodiversity assessment using DNA metabarcoding. Mol. Ecol. **21**(8)**:** 2045-2050. doi:10.1111/j.1365-294X.2012.05470.x.

Telfer, A.C., Young, M.R., Quinn, J., Perez, K., Sobel, C.N., Sones, J.E., et al. 2015. Biodiversity inventories in high gear: DNA barcoding facilitates a rapid biotic survey of a temperate nature reserve. Biodiv. Data, J. **3:** e6313. doi:10.3897/BDJ.3.e6313.

Vellend, M. 2005. Species diversity and genetic diversity: Parallel processes and correlated patterns. Am. Nat. **166**(2)**:** 199-215. doi:10.1086/431318.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Doring, M., Giovanni, R., et al. 2012. Darwin core: An evolving community-developed biodiversity data standard. PLoS One, **7**(1)**:** 8. doi:10.1371/journal.pone.0029715.

Wilson, J.J. 2012. DNA barcodes for insects. Methods Mol. Biol. **858:** 17-46. doi:10.1007/978-1-61779-591-6_3.

Wirta, H., Varkonyi, G., Rasmussen, C., Kaartinen, R., Schmidt, N.M., Hebert, P.D.N., et al. 2016. Establishing a community-wide DNA barcode library as a new tool for arctic research. Mol. Ecol. Resour. **16**(3)**:** 809-822. doi:10.1111/1755-0998.12489.

Young, M.R., Behan-Pelletier, V.M., and Hebert, P.D.N. 2012. Revealing the hyperdiverse mite fauna of subarctic Canada through DNA barcoding. PLoS One, **7**(11)**:** 11. doi:10.1371/journal.pone.0048755.

Yu, D.W., Ji, Y.Q., Emerson, B.C., Wang, X.Y., Ye, C.X., Yang, C.Y., et al. 2012. Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods Ecol. Evol. **3**(4)**:** 613-623. doi:10.1111/j.2041-210X.2012.00198.x.

Zahiri, R., Lafontaine, D., Schmidt, B.C., deWaard, J.R., Zakharov, E.V., and Hebert, P.D.N. 2014. A transcontinental challenge: A test of DNA barcode performance for 1,541 species of Canadian Noctuoidea (Lepidoptera). PLoS One, **9**(3)**:** 12. doi:10.1371/journal.pone.0092797.

Zahiri, R., Lafontaine, J.D., Schmidt, B.C., deWaard, J.R., Zakharov, E.V., and Hebert, P.D.N. 2017. Probing planetary biodiversity with DNA barcodes: The Noctuoidea of North America. PLoS One, **12**(6)**:** 18. doi:10.1371/journal.pone.0178548.

Zlotnick, B., Kohn, J., Dannecker, D., and Levesque-Beaudin, V. 2015. "Barcoding our backyard" at ResMed, Inc.: 52-consecutive weeks Malaise trap project at a corporate headquarters in a Global biodiversity hotspot. Genome, **58**(5)**:** 303.

Figure Captions

Figure 1. Workflow for biodiversity monitoring through DNA barcoding.

Figure 2. Flowchart showing the success in sequence recovery from 21,194 specimens of arthropods in ten Malaise trap samples.

Figure 3. Taxonomic breakdown of the Malaise trap samples by (a) specimens and (b) BINs.

Figure 4. The number of specimens and BINs in ten Malaise trap samples from Point Pelee National Park. Unique BINs are those found in only one of the ten weekly samples.

Figure 5. Relative species abundance plot for the ten Malaise trap samples.

Figure 6. Relationship between the number of specimens in each of ten samples from Point Pelee National Park and the number of BINs unique to it ($R^2 = 0.69$, $p = 0.003$).

Figure 7. Preston plot with veil line and extrapolation based upon the abundance data for the taxa represented among the 19,071 arthropods that generated a sequence.

Figure 8. BIN accumulation curves for the 19,071 arthropods from Point Pelee National Park estimated with (a) specimens and (b) weekly samples. Grey shading indicates the 95% confidence interval.

Figure 9. Species overlap between the ten Malaise trap samples, shown (a) in chronological order with the size of each node proportional to the number of BINs in a sample while the width of each

arcs reflects BIN overlap between samples. (b) as a comparison of BIN overlap with time between samples ($R^2 = 0.52$, $p << 0.001$).

Tables

Table 1. Primers used for DNA barcode analysis. For each taxonomic group, there was a single first pass primer pair (listed first) used to amplify the 658 bp barcode region of COI and two second pass PCR primer pairs (list second and third) used to amplify two smaller, overlapping COI fragments (307 bp and 407 bp). The listed primer was used for sequencing unless indicated by a symbol.

Taxonomy	PCR Primer Pair	Forward Primer(s)	Sequence (5'-->3')	Reverse Primer(s)	Sequence (5'-->3')	Fragment Length (bp)
Diptera,	C_LepFolF/C_LepFolR	LepF1	ATTCAACCAATCATAAAGATATTGG	LepR1	TAAACTTCTGGATGTCCAAAAAATCA	658
Coleoptera,		LCO1490	GGTCAACAAATCATAAAGATATTGG	HCO2198	TAAACTTCAGGGTGACCAAAAAATCA	
Arachnida,	C_LepFolF/MLepR2	LepF1	ATTCAACCAATCATAAAGATATTGG	MLepR2	GTTCAWCCWGTWCCWGCYCCATTTTC	307
Collembola		LCO1490	GGTCAACAAATCATAAAGATATTGG	-	-	
and small	MLepF1/C_LepFolR	MLepF1	GCTTTCCCACGAATAAATAATA	LepR1	TAAACTTCTGGATGTCCAAAAAATCA	407
Orders		-	-	HCO2198	TAAACTTCAGGGTGACCAAAAAATCA	
Lepidoptera	LepF1/LepR1	LepF1	ATTCAACCAATCATAAAGATATTGG	LepR1	TAAACTTCTGGATGTCCAAAAAATCA	658
	LepF1/MLepR2	LepF1	ATTCAACCAATCATAAAGATATTGG	MLepR2	GTTCAWCCWGTWCCWGCYCCATTTTC	307
	MLepF1/LepR1	MLepF1	GCTTTCCCACGAATAAATAATA	LepR1	TAAACTTCTGGATGTCCAAAAAATCA	407
Hymenoptera	LepF1/LepR1	LepF1	ATTCAACCAATCATAAAGATATTGG	LepR1	TAAACTTCTGGATGTCCAAAAAATCA	658
	LepF1/C_ANTMR1D †	LepF1	ATTCAACCAATCATAAAGATATTGG	N/A		~307
	RonMWASPdeg_t1/LepR1	RonMWASPdeg_t1*	TGTAAAACGACGGCCAGTGGWTCWCCWGATATAKCWTTTCC	LepR1	TAAACTTCTGGATGTCCAAAAAATCA	407
Hemiptera	LepF2_t1/LepR1	LepF2_t1*	TGTAAAACGACGGCCAGTAATCATAARGATATYGG	LepR1	TAAACTTCTGGATGTCCAAAAAATCA	658
	LepF2_t1/MHemR	LepF2_t1*	TGTAAAACGACGGCCAGTAATCATAARGATATYGG	MHemR	GGTGGATAAACTGTTCAWCC	307
	MHemF/LepR1	MHemF	GCATTYCCACGAATAAATAAYATAAG	LepR1	TAAACTTCTGGATGTCCAAAAAATCA	407

* M13 tailed forward primers sequenced with M13F

† C_ANTMR1D cocktail not used in sequencing reaction

Supplementary material

Fig. S1 Neighbor-Joining tree based on sequence divergences at COI (K2P distance model) for one representative of all 2,255 BINs.

Fig. S2 Image library matching the COI Neighbor-Joining tree of BIN representatives. In a few instances, an image for the BIN representative was unavailable because the specimen was not recovered after DNA extraction. In these cases, an image of a different representative of the same BIN from another site was chosen, or in rare cases, from the nearest neighbor BIN (as marked below the image).

Table S1 BOLD and GenBank accessions, as well as BIN assignments and collection details for the 19,185 arthropods from Point Pelee National Park.

Table S2 Summary of specimens and BINs by taxonomic order.

Table S3 Summary of specimens, BINs, and BINs unique to each weekly sample.

Table S4 Jaccard similarity index and temporal distance in days between each pair of weekly samples.

Table S5 Combined checklist of genera and species recorded at Point Pelee National Park by morphological (Marshall et al. 2009) and DNA barcode inventories.

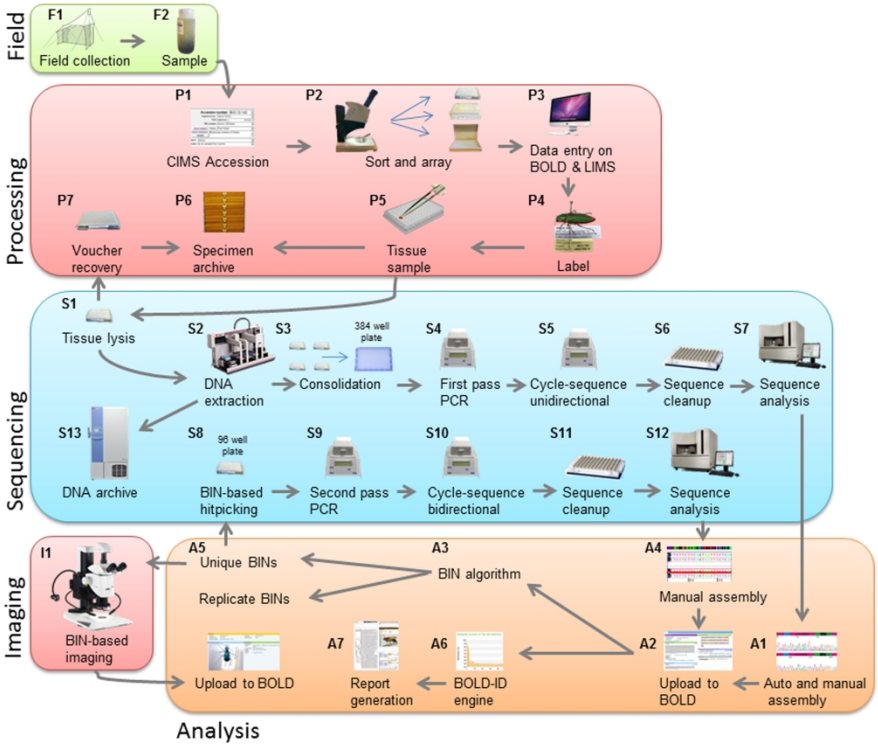

Figure 1. Workflow for biodiversity monitoring through DNA barcoding.

250x199mm (300 x 300 DPI)

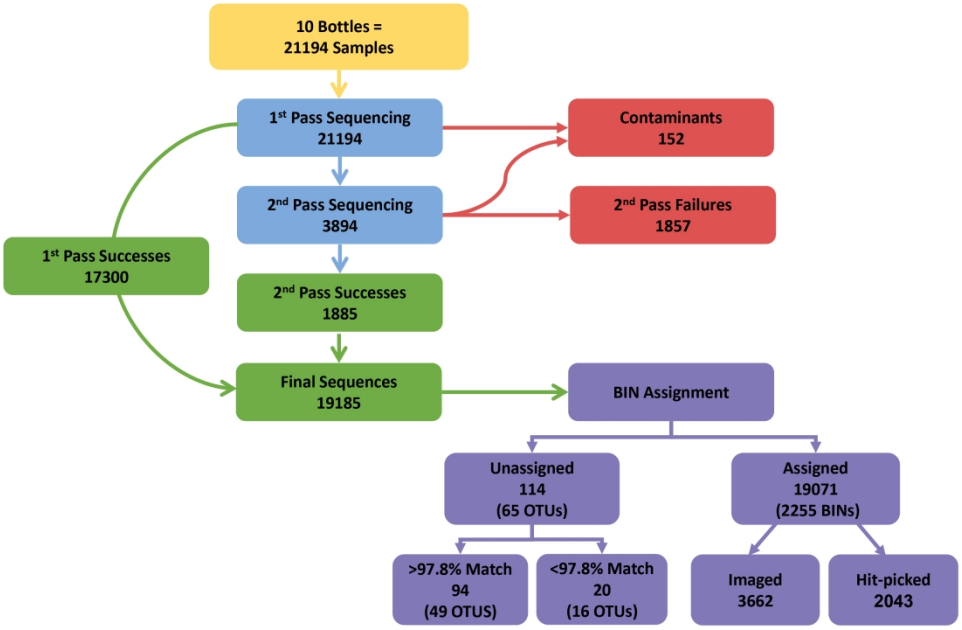

Figure 2. Flowchart showing the success in sequence recovery from 21,194 specimens of arthropods in ten Malaise trap samples.

280x184mm (300 x 300 DPI)

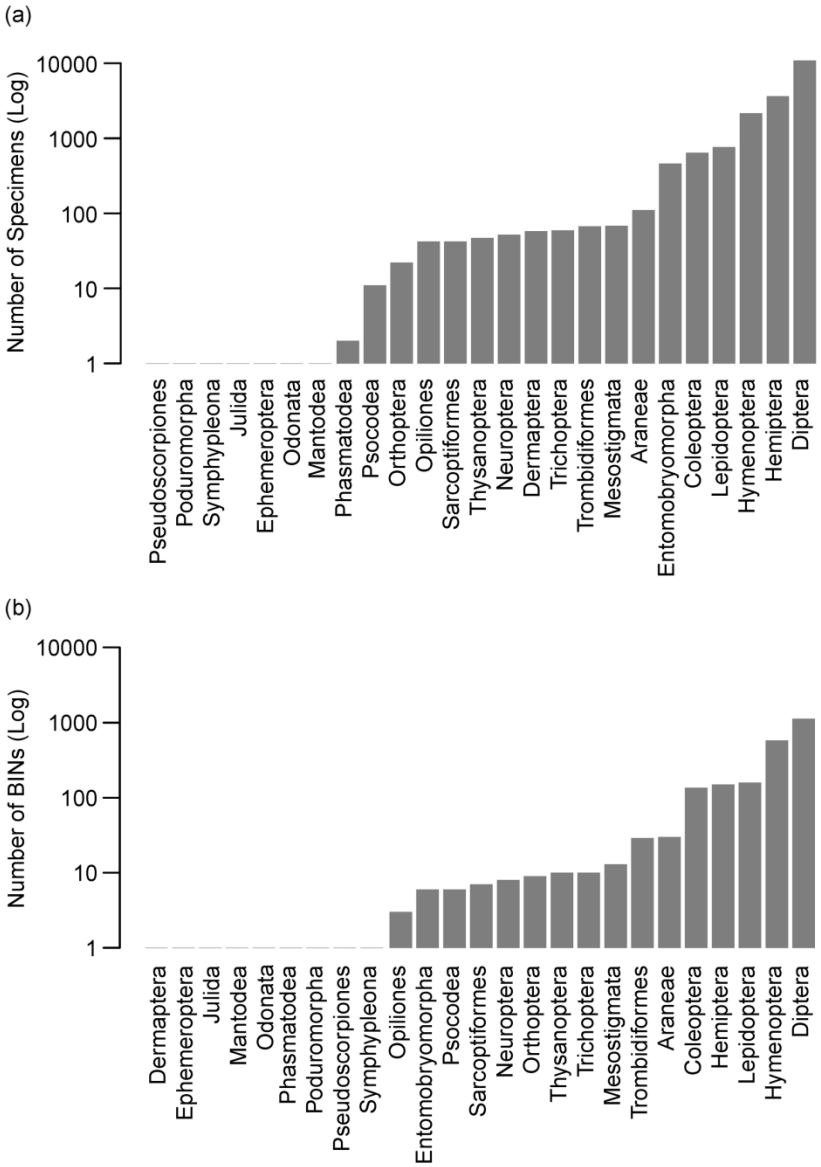

Figure 3. Taxonomic breakdown of the Malaise trap samples by (a) specimens and (b) BINs.

187x262mm (300 x 300 DPI)

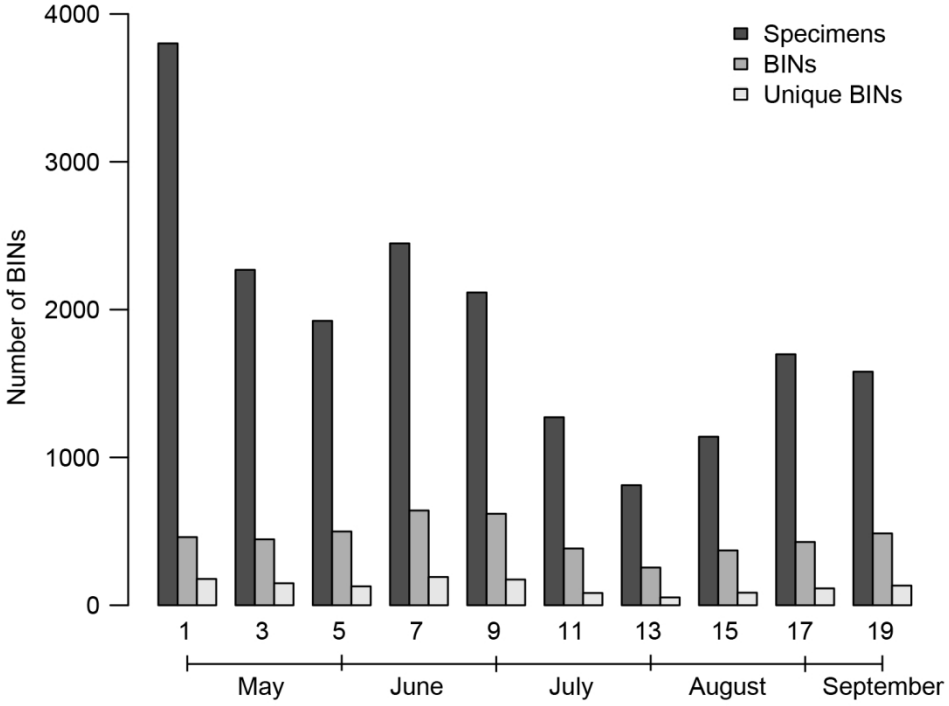

Figure 4. The number of specimens and BINs in ten Malaise trap samples from Point Pelee National Park. Unique BINs are those found in only one of the ten weekly samples.

142x101mm (300 x 300 DPI)

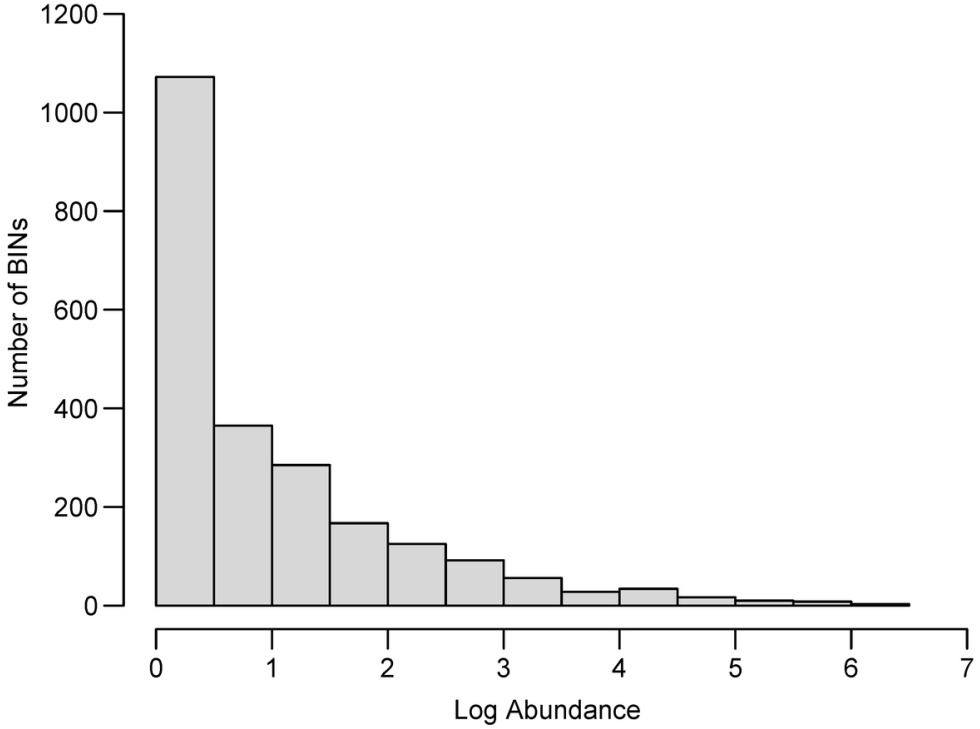

Figure 5. Relative species abundance plot for the ten Malaise trap samples.

101x78mm (300 x 300 DPI)

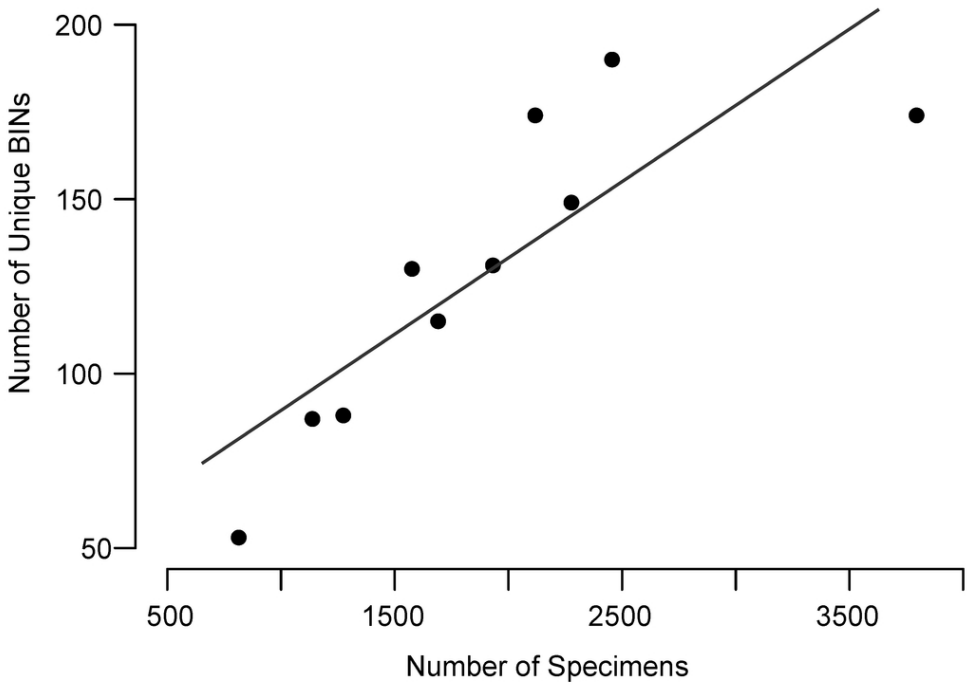

Figure 6. Relationship between the number of specimens in each of ten samples from Point Pelee National Park and the number of BINs unique to it (R^2 = 0.69, p = 0.003).

89x64mm (300 x 300 DPI)

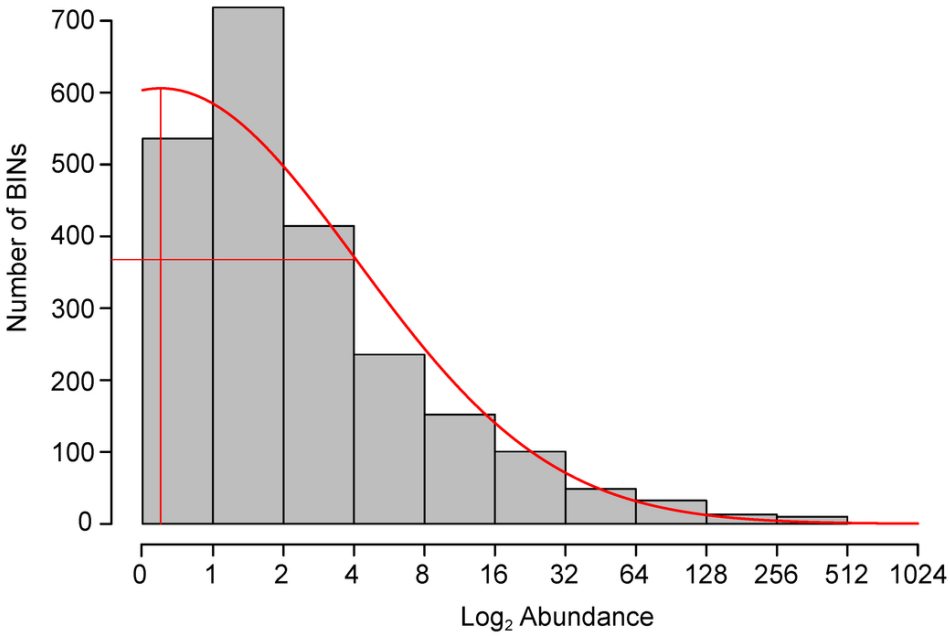

Figure 7. Preston plot with veil line and extrapolation based upon the abundance data for the taxa represented among the 19,071 arthropods that generated a sequence.

92x64mm (300 x 300 DPI)

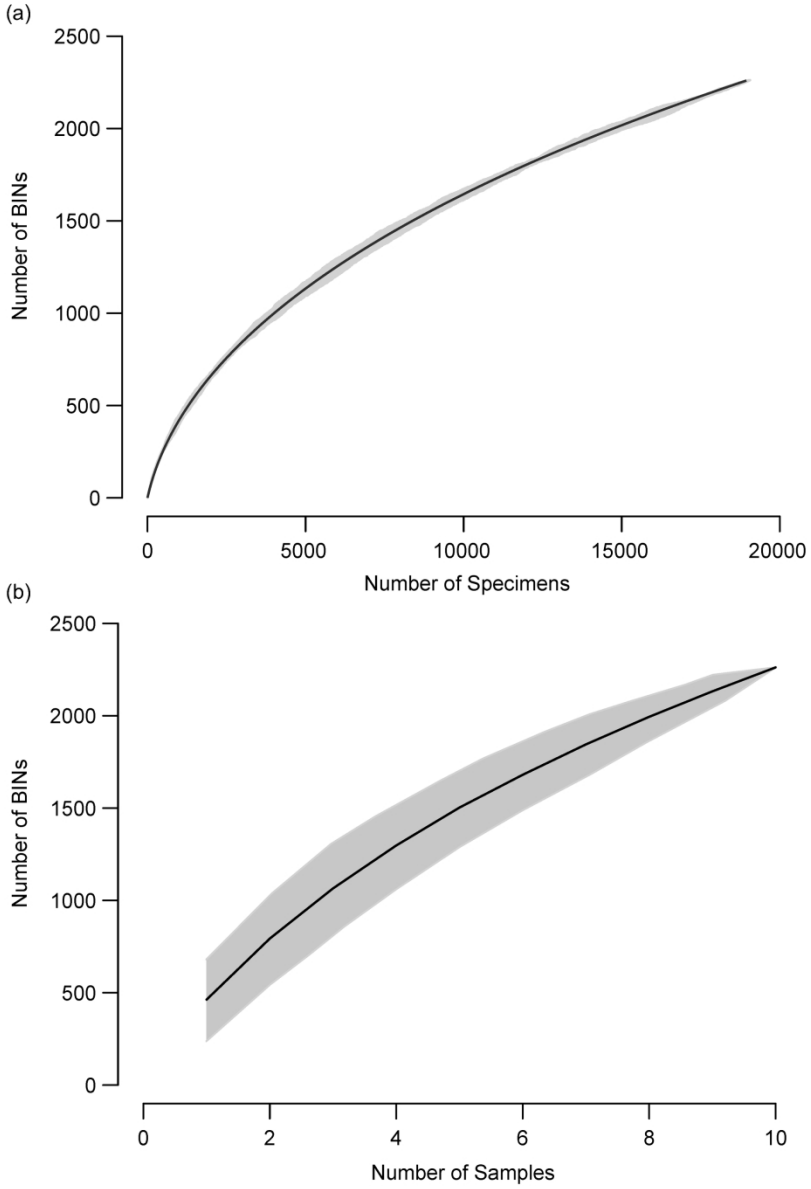

Figure 8. BIN accumulation curves for the 19,071 arthropods from Point Pelee National Park estimated with (a) specimens and (b) weekly samples. Grey shading indicates the 95% confidence interval.

202x288mm (300 x 300 DPI)

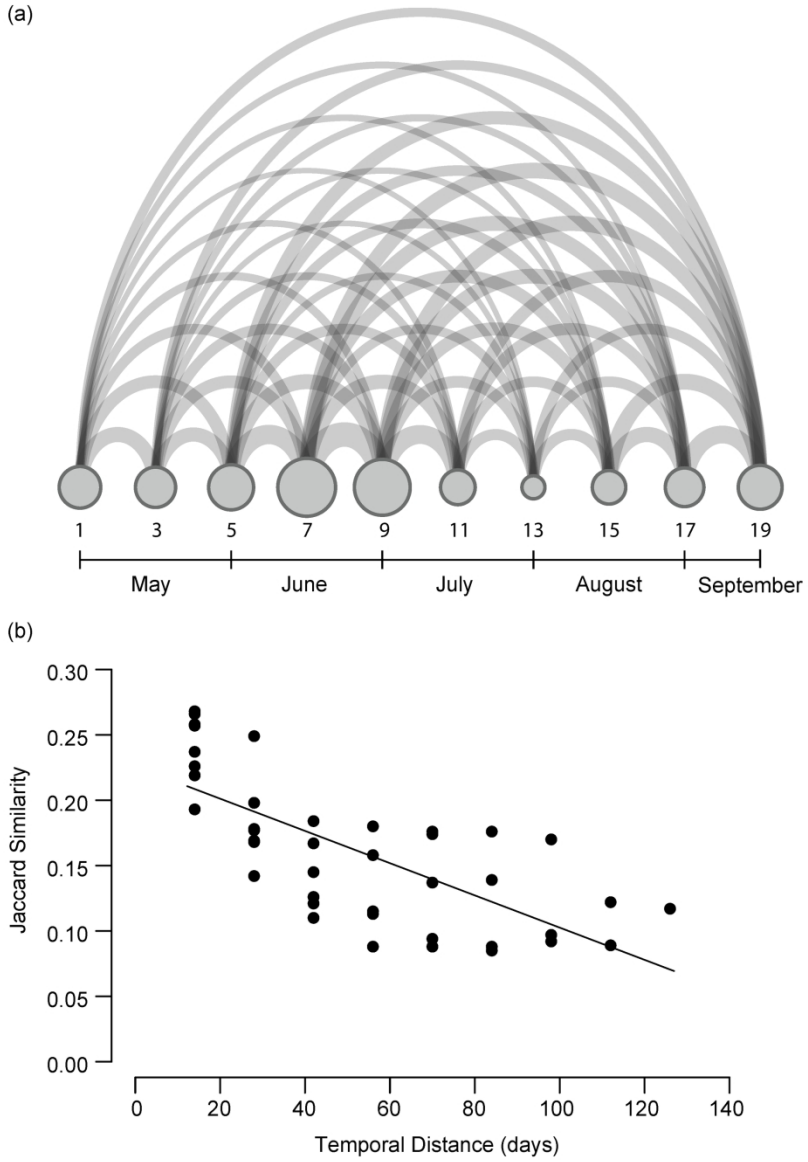

Figure 9. Species overlap between the ten Malaise trap samples, shown (a) in chronological order with the size of each node proportional to the number of BINs in a sample while the width of each arcs reflects BIN overlap between samples. (b) as a comparison of BIN overlap with time between samples (R^2 = 0.52, p << 0.001).

197x275mm (300 x 300 DPI)