Fast-tracking bespoke DNA reference database generation from museum collections for biomonitoring and conservation

Andrew Dopheide[1], Talia Brav-Cubitt[1], Anastasija Podolyan[2], Rich Leschen[1], Darren Ward[1], Thomas Buckley[1], and Manpreet K. Dhami[2]

[1]Landcare Research New Zealand Auckland
[2]Landcare Research New Zealand

June 1, 2022

Abstract

Despite recent advances in high-throughput DNA sequencing technologies, a lack of locally relevant DNA reference databases may limit the potential for DNA-based monitoring of biodiversity for conservation and biosecurity applications. Museums and national collections represent a compelling source of authoritatively identified genetic material for DNA database development yet obtaining DNA barcodes from long-stored specimens may be difficult due to sample degradation. We demonstrate a sensitive and efficient laboratory and bioinformatic process for generating DNA barcodes from hundreds of invertebrate specimens simultaneously via the Illumina MiSeq system. Using this process, we recovered full-length (334) or partial (105) COI barcodes from 439 of 450 (98 %) national collection-held invertebrate specimens. This included full-length barcodes from 146 specimens which produced low-yield DNA and no visible PCR bands, and which produced as little as a single sequence per specimen, demonstrating high sensitivity of the process. In many cases, the identity of the most abundant sequences per specimen were not the correct barcodes, necessitating the development of a taxonomy-informed process for identifying correct sequences among the sequencing output. The recovery of only partial barcodes for some taxa indicates a need to refine certain PCR primers. Nonetheless, our approach represents a highly sensitive, accurate, and efficient method for targeted reference database generation, providing a foundation for DNA-based assessments and monitoring of biodiversity.

Fast-tracking bespoke DNA reference database generation from museum collections for biomonitoring and conservation

Andrew Dopheide[1], Talia Brav-Cubitt[1], Anastasija Podolyan[2], Richard A. B. Leschen[1], Darren Ward[1,3], Thomas R. Buckley[1,3], and Manpreet K Dhami[2]

[1] Manaaki Whenua Landcare Research, PO Box 92170, Auckland, New Zealand

[2] Manaaki Whenua Landcare Research, PO Box 69040, Lincoln, New Zealand

[3] School of Biological Sciences, The University of Auckland, PO Box 92019, Auckland, New Zealand

Corresponding author:

Manpreet K Dhami

dhamim@landcareresearch.co.nz

Abstract

Despite recent advances in high-throughput DNA sequencing technologies, a lack of locally relevant DNA reference databases may limit the potential for DNA-based monitoring of biodiversity for conservation and biosecurity applications. Museums and national collections represent a compelling source of authoritatively

identified genetic material for DNA database development yet obtaining DNA barcodes from long-stored specimens may be difficult due to sample degradation. We demonstrate a sensitive and efficient laboratory and bioinformatic process for generating DNA barcodes from hundreds of invertebrate specimens simultaneously via the Illumina MiSeq system. Using this process, we recovered full-length (334) or partial (105) COI barcodes from 439 of 450 (98 %) national collection-held invertebrate specimens. This included full-length barcodes from 146 specimens which produced low-yield DNA and no visible PCR bands, and which produced as little as a single sequence per specimen, demonstrating high sensitivity of the process. In many cases, the identity of the most abundant sequences per specimen were not the correct barcodes, necessitating the development of a taxonomy-informed process for identifying correct sequences among the sequencing output. The recovery of only partial barcodes for some taxa indicates a need to refine certain PCR primers. Nonetheless, our approach represents a highly sensitive, accurate, and efficient method for targeted reference database generation, providing a foundation for DNA-based assessments and monitoring of biodiversity.

Keywords

Molecular taxonomy, DNA-based monitoring, conservation, invertebrate barcoding, museum collection, taxonomy-informed bioinformatics pipeline

Introduction

Recent advances in high-throughput sequencing technology are enabling a shift towards environmental DNA (eDNA)-based methods for biodiversity assessment and biosecurity monitoring [1]. While still in their infancy, these tools offer great promise for rapid and accessible biodiversity monitoring applications in terrestrial, aquatic, and marine ecosystems [2, 3]. However, a scarcity of accurately identified reference DNA sequence data from local biota [4] remains a significant obstacle to the application of these eDNA tools to biomonitoring, preventing the confident identification and interpretation of detected organisms.

DNA-based identification methods, regardless of application, rely on the determination of similarity between newly detected sequences and existing reference sequence data [4, 5]. Depending on target taxa, this requires representative taxonomically validated data on established marker genes such as the ˜650 bp region of cytochrome*c* oxidase subunit I (COI) for metazoans [6], or combinations of plastid regions (*rbcl* , *matK* , *trnH–psbA*) and the ribosomal internal transcribed spacer region (ITS) for plants [7]. Large open data sources, such as the GenBank *nr* database or BOLD [8] are typically employed as reference databases, but rely on data submitters for fidelity of sequence to organism and suffer from geographic sampling biases [9]. Further, this reliance on large pre-existing databases also limits the emergence of new or taxon-specific markers, with individual studies utilising previously established markers even if they may be suboptimal for certain taxa [10]. Curated databases containing only sequences from a targeted ecosystem may result in improved accuracy of sequence identifications compared to a global database [11]. However, the current sparse database coverage of biodiversity from most ecosystems means that targeted reference databases typically must be populated with newly generated and locally relevant reference sequences.

Taxonomically validated reference sequences are difficult to generate. Not only do they require high levels of sequence accuracy (traditionally achieved via Sanger sequencing, more recently possible via PacBio hifi technology[12]), but also accurate taxonomic identification of specimens. The former may be time-consuming, contingent on sample quality, and expensive, especially when applied to large numbers of specimens, while the latter requires specialist taxonomic expertise across taxa. For example, generating a reliable reference database for a previously uncharacterized insect fauna may require taxonomic skills spanning 24 distinct insect orders. Natural history museums and national biological collections, however, are unparalleled repositories of both invaluable taxonomic knowledge [13] and authoritatively identified genetic source material [14], with the potential to allow the efficient generation of taxonomically comprehensive and locally relevant reference DNA sequence databases [15]. Generating full-length DNA barcodes via Sanger sequencing from dried or historical specimens stored over long periods may be difficult due to DNA degradation and low sensitivity of the sequencing approach, often resulting in only partial barcodes [15-18]. Furthermore, museum samples are often indispensable permanent records, and therefore unavailable for destructive DNA extrac-

tion. Non-destructive extraction [19, 20] and PCR [21] approaches can be effective, however, depending on the taxa being analysed, and multiplex PCR coupled with high-throughput DNA sequencing technologies has allowed the efficient recovery of barcodes from 50- to 100-year-old museum samples [12, 22], as well as recently collected specimens [23].

There is a pressing need to leverage museum collections for rapid and cost-effective generation of reference databases, in order to aid eDNA-based biodiversity monitoring [24]. Here, we present a fast, cost-effective, and efficient method for developing a reference COI database from a diverse selection of terrestrial invertebrates sourced from the New Zealand Arthropod Collection (NZAC). These taxonomically validated specimens exhibit a variety of field collected methods, specimen treatment and storage conditions, as well as variable accessibility for destructive sampling. We demonstrate the use of a dual indexing approach, in combination with a pair of overlapping short PCR amplicons suitable for sequencing on the Illumina MiSeq platform, for generating full length barcodes from hundreds of invertebrate specimens simultaneously. We provide a taxonomy-informed bioinformatics pipeline for processing and filtering the sequence data and the rapid assembly of successful barcodes. Together, our approach represents a highly sensitive, accurate, and efficient method for targeted reference database generation, providing a foundation for DNA-based assessments and monitoring of biodiversity.

Materials and Methods

Specimen sampling and DNA extraction

Ethanol-stored and pinned-dry invertebrate specimens were selected from a variety of taxa, primarily earthworms (132 specimens) and insects (315 specimens, mainly beetles and wasps), as well as individual millipede, spider and mite specimens; 450 specimens in total (Table 1). We randomly selected 1-4 individuals per species, depending on availability. Earthworm specimens were variously collected between 2004 and 2014, and arthropod samples from 2009 to 2019 using a range of techniques, including malaise traps, pitfall traps, sweeping and hand collection. Most specimens were stored in 95 % ethanol at -20 °C, but some were in 70 % ethanol, and some recent samples were stored at room temperature.

Total genomic DNA of specimens was extracted in a sterile environment with the following variations, depending on the size and availability of specimens for destructive sampling. Briefly, specimens were either soaked whole (112 specimens: 93 wasps, ten beetles, and nine moths) or crushed whole (48 small specimens: 25 beetles, 12 flies, nine other insects, one spider, and one mite) in lysis buffer, or a piece of tissue was sampled (290 specimens; all 132 earthworms, 125 beetles, 32 other insects, and one millipede) and added to the lysis buffer. DNA extraction was carried out using one of the following kits following the manufacturer's instructions: DX reagents kit on the X-tractor Gene (Qiagen, Germantown, Maryland, USA) (211 specimens; 115 earthworms, 96 insects, and one mite, spider, and millipede); QIAamp 96 DNA QIAcube HT Kit on the QIAcube HT (Qiagen, Germantown, Maryland, USA) (222 insects, including the 112 soaked specimens); or the AquaPure Genomic DNA Isolation kit (BioRad, Hercules, California, USA) (17 earthworms). Extracted DNA was stored at -20°C until amplicon library construction.

Library construction and sequencing

We amplified two short overlapping fragments, FC (235 bp) and BR (428 bp), that together form the standard 658 bp DNA barcode region, located at the 5' end of the cytochrome c oxidase subunit I (COI) gene [23]. These short fragments are expected to enhance PCR success rates and ensure compatibility with read length constraints of the Illumina MiSeq system. The FC and BR fragments were amplified using the primer pairs Ill_LCO1490 and Ill_C_R, and Ill_B_F and Ill_HCO2198, respectively [23]. Forward and reverse primers were tagged with Nextera XT forward or reverse adapters (Illumina, Inc., San Diego, CA, USA), respectively, padded with 0-4 mer nucleotide spacers to increase sequence heterogeneity, as described in [25]. PCRs for both amplicons were carried out using the FastStart Taq DNA Polymerase kit (Roche, Basel, Switzerland), with final concentrations of 1x PCR buffer with $MgCl_2$, 0.2 nM dNTP mix, 1 µg/ml Bovine Serum Albumin (BSA), 250 nM of primer Ill_LCO1490 or Ill_HCO2198 and 750 nM of primer Ill_C_R or Ill_B_F, 1 U Taq polymerase, and PCR grade water up to a total reaction volume of 20 µl. PCR cycle conditions were 95 °C

for 5 minutes; 40 cycles of 95 °C for 45 seconds, 50 °C for 45 seconds, and 72 °C for 45 seconds; and 72 °C for 5 minutes (Veriti^{TM} 96-Well Thermal Cycler; Applied Biosystems, Waltham, Massachusetts, USA). Following this, 4 µl of PCR product for each sample was run on a 1.5 % agarose gel to check amplification success. First-step PCR amplicons were cleaned, normalised for maximum concentration, and size selected using SerraMag SpeedBeads^{TM} magnetic carboxylate modified particles (Cytiva, Marlborough, Massachusetts, US) with 0.8x bead:buffer concentration [26] to remove primer dimers before the second-step PCR.

Second-step PCRs were performed using the KAPA3G Plant PCR kit (KAPABIOSYSTEMS, Cape Town, South Africa), with Fusion primers with custom indices synthesized with P5/P7 Illumina adapters (forward, 5'- AAT GAT ACG GCG ACC ACC GAG ATC TAC AC - [8-mer tag] - TCG TCG GCA GCG TC -3'; reverse, 5'- CAA GCA GAA GAC GGC ATA CGA GAT - [8-mer tag] - GTC TCG TGG GCT CGG -3')

[27], using 400 nM of each primer and 1.44 µL of first-step PCR product in 18 µL volume. The second-step PCR cycle was 95 °C for 2 min, then five cycles of 95 °C for 20 s, 50 °C for 20 s and 72 °C for 30 s, followed by 72 °C for 2 min (Bio-Rad T100TM Thermal Cycler; Hercules, California, US). Then 2.5 µL of second-step PCR products were run on a 2 % agarose gel to check amplification success.

The amplicons were again cleaned, normalised, and size selected using SerraMag SpeedBeads^{TM} magnetic carboxylate modified particles (Cytiva, Marlborough, US) with 1.5x bead:buffer concentration. Amplicons were eluted in 15 µL of 0.1x TE buffer, and pooled together at equal volumes to form the amplicon library, which was analysed on the Automated Bioanalysis System LabChip® GX Touch HT (Perkin-Elmer, Waltham, Massachusetts, USA). DNA concentration was estimated by Qubit® 2.0 Fluorometer (Invitrogen, Waltham, Massachusetts, USA), and adjusted to 4 nM. The final library was processed on an Illumina MiSeq 3000 sequencer with 2 x 300 paired-end mode at the Genomics Facility at the University of Auckland, with 10% PhiX spiked in.

Bioinformatics pipeline

We obtained DNA barcodes from raw sequence data with the following process, using bcl2fastq version 2.20 (Illumina), Claident version 2018.05.08 [28], VSEARCH 2.14 [30], and EMBOSS merger V.6.6.0 [31]. Steps 1-8 were executed in Bash scripts, and steps 9-10 in a Python script:

1. Raw sequence data in bcl format was converted into fastq format, without demultiplexing, using bcl2fastq.
2. The forward and reverse fastq sequences were each demultiplexed into separate FC and BR fastq files for each of the 450 specimens, using clsplitseq script from Claident. This resulted in a pair of R1 and R2 fastq files for the FC amplicon, and a pair of R1 and R2 fastq files for the BR amplicon, for each of the 450 specimens.
3. For each amplicon (FC and BR) and specimen, the R1 and R2 sequences contained in each pair of fastq files were merged using -fastq_mergepairs in VSEARCH with minimum overlap (-fastq_minovlen), minimum merged length (-fastq_minmergelen), and maximum merged length (-fastq_maxmergelen) options of 200, 250, and 400 respectively for FC, and 100, 350, and 500 respectively for BR, with simultaneous filtering of merged sequences for errors (-fastq_maxee 1) and ambiguous bases (-fastq_maxn 0).
4. The filtered sequences from each amplicon and specimen were then dereplicated using -derep_fulllength, denoised using -cluster_unoise with -minsize 1, and filtered for chimeras using -uchime3_denovo, all in VSEARCH. Most of the filtered and denoised sequence files still contained multiple sequences, necessitating additional filtering to recover the correct barcode sequences, based on taxonomy, sequence abundance, and attributes of pairwise FC-BR sequence alignments, using subsequent steps.
5. A taxonomic identification was obtained for each filtered and denoised FC and BR sequence using BLAST against GenBank nr, accepting the top match in each case.
6. Up to $n = 20$ most abundant denoised FC and BR sequences (if any) were output as fasta files for each specimen, using the VSEARCH command -derep_fulllength with option -topn n . The FC and BR sequence files were concatenated together into a single fasta file per specimen.

7. All possible pairwise sequence alignments with the expected overlap between FC and BR amplicons, namely an overlap of 85 bp with 100 % pairwise sequence identity, were identified between the sequences in the concatenated FC and BR fasta file for each specimen using the VSEARCH command -allpairs_global with options -id 1 and -mincols 85.
8. If no alignments were detected among the 20 most abundant sequences from a given specimen, steps 6 and 7 were repeated with all sequences (if > 20) from that specimen.
9. Putatively correct full-length or partial barcodes were then identified per specimen based on comparing FC and BR sequence identifications and FC-BR alignment characteristics. An optimal FC sequence, BR sequence, and aligned FC-BR sequence pair (if any), was selected for each specimen by matching their sequence identities to the *a-priori* specimen taxonomy, prioritizing the lowest taxonomic rank level. The sequences were also required to have lengths between 324 and 326 b.p. for FC and between 417 and 419 b.p. for BR, and BLAST identification bitscores [?] 200 for FC and [?] 250 for BR, to avoid spurious matches. If more than one FC sequence, BR sequence, or aligned FC-BR sequence pair per specimen met these criteria, the most abundant sequence or sequence pair was selected.
10. If an FC-BR sequence pair with expected taxonomy was detected, and the lowest taxonomic identification rank of this sequence pair was equal to or lower than that of the selected FC and/or BR sequences for that specimen, that aligned FC-BR sequence pair was accepted as the likely correct barcode components and combined by EMBOSS merger. This carries out a pairwise alignment according to the Needleman-Wunsch algorithm, outputting a merged barcode sequence, along with an alignment summary that was checked to ensure the alignment attributes were as expected (overlap of 85 bp and score of 425).
11. To avoid accepting a merged sequence from a non-target organism, if either of the selected FC or BR sequences had a lower taxonomic identification rank than that of the selected aligned FC-BR sequence pair, that FC or BR sequence was accepted as a likely correct partial barcode, instead of the aligned FC-BR sequence pair.
12. If an FC or BR sequence (but not an aligned FC-BR sequence pair) with expected taxonomy was detected, that sequence was accepted as a likely correct partial barcode. If both an FC and BR sequence (but no aligned FC-BR sequence pair) with expected taxonomy were detected, the FC or BR sequence with the lowest taxonomic rank was accepted as a likely correct partial barcode. If these FR and BR sequences had the same lowest taxonomic rank, the most abundant sequence was accepted as the likely correct partial barcode.

The merged FC and BR barcodes, plus any partial FC- or BR-only barcodes, were concatenated into a single fasta file, and aligned using MAFFT [32]. An approximately maximum-likelihood phylogeny was generated from the alignment using FastTree 2 [33], and visualised using the R package ggtree [34].

The resulting barcode dataset was examined for any effects of DNA extraction methodology, PCR primer, or taxonomy on successful barcode recovery. For this purpose, any orders and families represented by fewer than five specimens were pooled together as "Others".

To assess the accuracy of recovered Illumina barcodes, 96 of the 450 specimen DNA extracts (47 beetles, 21 flies, 14 wasps, 11 other insects, and individual mite, millipede, and spider specimens) were also subjected to PCR using primers LCO1490 and HCO2198 (Folmer et al., 1994), followed by Sanger sequencing of the amplicons using standard methods. The resulting Sanger barcode sequences were compared with Illumina-derived barcodes for the same specimens via generation of pairwise sequence alignments using MAFFT (Katoh & Standley, 2013) and determination of sequence identity between each sequence pair.

Results

Overall DNA barcoding success

Our barcoding process resulted in full-length COI barcodes for 334 specimens (74.2 %), plus FC-only and BR-only barcodes for a further 17 (3.78 %) and 88 (19.6 %) specimens, respectively. In total, full-length or partial COI barcodes were recovered for 439 of 450 specimens (97.6 %) (Table 1, Figure 1). This included

full-length / partial barcodes for 87.9 % / 9.85 % of earthworm specimens, 66.2 % / 31.9 % of Hymenoptera specimens, and 65.8 % / 30.7 % of Coleoptera specimens, respectively.

PCR and sequencing outcomes

Initial PCR success rates differed between the FC and BR amplicons and among different taxa (Table 1). Visible PCR products were amplified from 233 of 450 specimens (51.7 %) in FC PCRs, compared to 384 specimens (85.3 %) in BR PCRs. Both FC and BR PCRs visibly succeeded for 205 specimens (45.6 %). Visible FC PCR success rates were lowest for Coleoptera (25.6 %) followed by Diptera (38.1 %), and highest for Annelida (90.1 %). In contrast, visible BR PCR success rates exceeded 70 % for all Orders except for Psocoptera (with only two specimens), including 90 % for Coleoptera, 72.6 % for Diptera, and 86.2 % for Annelida.

The numbers of sequences per amplicon and specimen after filtering and denoising varied widely, from zero (for 67 FC PCRs and five BR PCRs) to several thousand, with means of 72 in FC PCRs and 161 in BR PCRs. High numbers of denoised FC sequences per specimen were strongly correlated with high numbers of denoised BR sequences per specimen (Pearson correlation coefficient = 0.85, $p < 0.001$). The numbers of pairwise alignments between FC and BR amplicons with expected characteristics (overlap of 85 bp and 100 % identity) per specimen ranged between zero (for 92 specimens) to 180 (for a Hymenoptera specimen), with a mean of 14. Numbers of pairwise alignments were not obviously correlated with numbers of denoised FC and BR sequences. Low to moderate numbers of alignments were detected for 140 specimens from which FC PCRs did not produce visible products, and 39 specimens from which BR PCRs did not produce visible products.

After taxonomic filtering of sequences and pairwise alignments to identify optimal barcodes, full-length COI barcodes were recovered for 334 of 450 specimens (74 %). This included full-length barcodes for 146 specimens from which FC PCRs (109), BR PCRs (22), or both (15) did not result in a visible PCR product. Partial barcodes in the form of BR sequences were only recovered for a further 88 specimens, and FC sequences only for another 17 specimens.

No evidence of DNA extraction methodology effects on barcoding outcomes was observed. Rather, the most obvious factor affecting successful barcode detection was a combination of PCR amplicon and taxonomy (Table 2). FC sequences (either full-length or FC-only barcodes) were successfully detected for 75 % of specimens on average across 19 different orders and families considered, compared to > 94 % for BR sequences. Rates of FC sequence detection were lower than rates of BR sequence detection in 13 groups, the same in five, and higher in only one (Annelida, by 7 %). Among insect taxa, FC sequence detection rates were the same as BR sequence detection rates for two orders (Hemiptera and Lepidoptera) and one family (Ichneumonidae), and lower for all other insect groups. The largest difference between successful FC and BR sequence detection rates was observed for Staphylinidae (10 specimens, -100 %), with disparities [?] -20 % observed for a further five insect groups including Chrysomelidae (107 specimens, -30 %), and Braconidae (87 specimens, -23 %).

PCR amplification specificity

Sequences identified as the most likely partial barcodes by taxonomic filtering were the maximally abundant sequence per specimen in 345 cases for the FC amplicon and in 365 cases for the BR amplicon. Combined, the selected FC and BR sequences were both the maximally abundant sequences per specimen in only 286 cases. A further 29 selected FC sequences and 50 selected BR sequences each had abundance ranks between two and ten. Two selected FC sequences had abundance ranks of 18 and 30, respectively, and 10 selected BR sequences each had abundance ranks between 11 and 101. In 37 cases where the maximally abundant FC sequence was not selected, the maximally abundant sequences were identified as deriving from insects (25), annelids (8), arachnids (2), algae (1), and amoebae (1). Similarly, in 74 cases where the maximally abundant BR sequence was not selected, these were identified as deriving from insects (44), other hexapods (2), annelids (2), or gastropods (1); and as *Homo sapiens* (5), and eukaryote (2) or prokaryote (18) micro-organisms, including 10 cases of *Wolbachia* .

To investigate the origins of maximally abundant but non-target sequences (i.e. those with unexpected taxonomic identifications), the allpairs_global function in VSEARCH was used to identify any identical sequences among the maximally abundant sequences per specimen, plus the selected (presumed correct) sequences (if not maximally abundant) from each specimen. Out of 37 specimens with maximally abundant but non-target FC sequences, only three of those sequences were identical to another maximally abundant but non-target FC sequence, in two cases from adjacent PCR wells and in all three cases from within the same PCR plates. Similarly, out of 74 specimens with maximally abundant but non-target BR sequences, ten of those sequences were each identical to one or more other maximally abundant but non-target BR sequence, in only four cases from adjacent PCR wells but in all cases from within the same PCR plates.

Comparison of MiSeq barcodes with Sanger barcodes

Full-length barcodes were successfully recovered by both MiSeq and Sanger barcoding approaches for 68 of 96 specimens subjected to both methods. Full length Illumina barcodes were obtained from seven specimens, and partial Illumina barcodes (one LCO-only and eight BR-only) from a further nine specimens, that each failed to produce Sanger barcodes; 15 of these 16 specimens were beetles. On the other hand, only partial barcodes (one LCO-only and nine BR-only) were recovered using the Illumina method from ten specimens from which full-length Sanger barcodes were obtained. No barcodes were obtained using either method from just two specimens (*Ephutomorpha bivulnerata* , a wasp; and the single mite specimen).

Among 68 specimens from which both Sanger and Illumina barcodes were recovered, pairwise sequence identities between these barcodes were from 75.5 % to 85.7 % for three specimens, 97.6 % to 98.6 % for six specimens, 99.1 % to 99.9 % for 18 specimens, and 100 % for 40 specimens.

Discussion

A lack of high quality and location-specific reference sequence data appears to limit the potential for DNA-based monitoring of terrestrial biodiversity, despite the great promise of these techniques. A scarcity of invertebrate taxonomic expertise and an associated lack of authoritatively identified specimens, and the high costs of Sanger sequencing, pose significant barriers to reference database generation. We present an efficient strategy that helps to overcome these barriers, with the potential to lessen the need for reliance on publicly available databases with inadequate local relevance for biodiversity monitoring. By leveraging a set of taxonomically identified specimens from a national collection coupled with a sensitive high-throughput sequencing approach, we rapidly generated a reference COI sequence database consisting of full-length barcodes for 334 specimens and partial (FC or BR) barcodes for a further 105 specimens, representing a wide range of invertebrate taxa from a diverse range of locations and with varied storage conditions. We observed no obvious effects of sampling or DNA extraction methods on barcoding success, indicating that a variety of protocols and specimen types can provide acceptable outcomes using this process. Furthermore, BR sequences were recovered from nearly all specimens, highlighting the potential for this process to achieve exceptionally high rates of DNA barcoding success, apparent deficiencies with the FC PCR primers notwithstanding.

While this analysis represents only a small portion of the source collection, the number of specimens included was arbitrarily limited, and there is considerable scope to greatly increase the throughput of this process. We recovered an average of over 10,000 sequences per amplicon and specimen (albeit with considerable variance). Given that the theoretical capacity of the MiSeq system exceeds 20 million sequence reads, and that only one correct FC and BR sequence is required to form a complete barcode, this suggests, conservatively, that at least one (or perhaps several) thousand specimens could be sequenced in one MiSeq run using this process. Typically, as the number of samples pooled together increases, the read depth per sample decreases, which may influence the detection of sequences from specimens that were difficult to amplify.

Previous attempts to obtain DNA barcodes from multiple invertebrate specimens have used a variety of sequencing approaches. In one example, Sanger sequencing was used to obtain DNA barcodes from 86 % of over 40,000 museum-held Lepidoptera specimens, demonstrating a profound effect of specimen age on barcoding success using this method [15]. However, this effort required six months of molecular work by five people, illustrating the inefficiencies/impracticality of Sanger sequencing applied to large numbers of

specimens. Invertebrate DNA barcoding efforts utilizing high-throughput sequencing technologies typically report greater efficiency, lower costs, and higher barcoding success rates than equivalent Sanger sequencing-based efforts [12, 22, 35]. The two-amplicon PCR approach used in this study was previously used to obtain barcodes from 97 % of > 1000 freshly trapped arthropod specimens [23]. We achieved comparable success rates from older specimens, from a wide range of locations, including a diverse selection of earthworms, confirming the utility of this MiSeq approach for efficiently barcoding numerous specimens from diverse lineages and sources. On the other hand, the same approach applied to barcoding of dried saproxylic beetle specimens achieved a lower success rate of 55 %, perhaps due to specimen collection methods being suboptimal for DNA preservation [36]. The MiSeq system has also been used in a multi-locus metabarcoding approach for detecting insect pests in bulk trap catches, which confirmed the importance of taxonomic information for confirming metabarcoding outcomes [19].

Single molecule real-time (SMRT) sequencing on the PacBio Sequel platform has recently been used to recover DNA barcodes from 20,000 insect specimens [35], and to recover barcodes from hundreds of ˜50 year-old butterfly specimens [12]. This system is argued to offer the most economic high-throughput barcoding system [35], but this relies upon very large numbers (tens of thousands) of input specimens. The MiSeq system is arguably more accessible (in terms of platform availability and sequencing run costs) than the PacBio Sequel system, and is suited to more modest numbers of specimens (hundreds to low thousands), which may be more compatible with biomonitoring requirements. Below we discuss some of the salient features and limitations of our approach.

Sensitivity, specificity and accuracy of Illumina barcode generation

The balance between sensitivity and specificity of high-throughput sequencing is often difficult to maintain [37], and is in our case tilted in favour of sensitivity to increase the rate of barcode recovery. This high sensitivity of Illumina sequencing enabled the recovery of numerous complete DNA barcodes from PCRs that work insufficiently well to produce a product visible by gel electrophoresis. Such specimens, accounting for almost half of our samples for the FC fragment, would likely fail to produce Sanger sequences, and be relegated to the 'difficult to sequence' set of preserved specimens [38]. Furthermore, full-length barcodes were recovered from various specimens with clearly suboptimal sequencing outcomes, including eight specimens with a single FC sequence (after error filtering), another with a single BR sequence, and a further 14 specimens with between two and five FC or BR sequences. On the other hand, this sensitivity also allowed the amplification of non-target sequences from well-known sources of contamination, such as extraction buffers ('kitome') [39], human specimen handling or bacterial symbionts [40], as well as apparent cross-contaminant sequences from other samples. Indeed, many of the maximally abundant sequences that were not selected as part of correct barcodes were similar or identical to those from other specimens included in the analysis, suggesting that these may variously result from PCR errors [41], co-amplification of numts [42], cross-contamination during library preparation [43], or index switching during sequencing [44]. Similar issues were observed in a study using the same sequencing approach to barcode a saproxylic beetle collection [36], indicating such contaminants may be an inevitable consequence of applying a highly sensitive method to specimens that may not have been collected with DNA analyses in mind. These issues might be mitigated in future analyses by stringent laboratory protocols to limit contamination [45], alternative library generation workflows [46], and fewer PCR cycles [47], although the latter might be at the expense of sensitivity. Additionally, bioinformatic tools can be used to identify correct barcodes among sequencing output. In this case, for example, the presence of contaminants necessitated the development of a semi-automated barcode assembly pipeline to accurately resolve barcode sequences.

To assess recovered barcode sequence accuracy, 96 of the specimens included in this analysis were also subjected to DNA barcoding via Sanger sequencing. Full-length barcodes were recovered from most of these specimens (68) by both methods. However, only Illumina barcodes (seven full-length and nine partial) were obtained from 16 specimens that failed to produce Sanger barcodes. This included 15 beetles, which can be challenging DNA barcoding subjects due to their tough exoskeletons, further illustrating the sensitivity of the Illumina approach. On the other hand, only partial barcode sequences were obtained via Illumina sequencing

from ten specimens from which full Sanger barcodes were obtained. Some of these required multiple PCR optimization attempts to obtain Sanger barcodes, however, along with examination and manual editing of sequence chromatograms. High levels of pairwise sequence identity were observed between Sanger and Illumina barcodes from most specimens (99 to 100 % for 59 of 68 specimens), indicating generally high levels of accuracy for both sequencing approaches. Three specimens (all beetles) had pairwise sequence identities between 75.5 % and 85.7 %, suggesting that either the Illumina approach or the Sanger approach recovered a sequence from a non-target organism in these cases. These FC-BR sequence pairs were each selected due to their constituent FC and/or BR fragments being the most abundant sequences identified to the correct taxonomic families, with no lower rank taxonomic information available among the BLAST results to further guide correct barcode selection.

Using taxonomy to assemble barcodes

Because the maximally abundant sequences from 20-30 % of specimens were not from the expected taxa according to BLAST, we developed a taxonomy-weighted barcode-assembly approach. For each specimen, we considered the most abundant FC and/or BR sequences with the expected taxonomic identifications—prioritising the lowest identifiable taxonomic rank in each case—to be correct and considered merged FC and BR sequences to be correct barcodes only if the contributing sequences had 100 % identity across the expected overlap length. This approach typically identified correct FC and BR sequences among the 20 most abundant sequences per specimen, but in a small number of cases, the correct sequences were identified at abundance ranks between 21 and 101. The ability to examine multiple sequences for correct identity is a key advantage of this process over Sanger sequencing, in which only a single sequence per specimen can typically be examined. This simple yet effective filtering approach greatly enhanced successful barcode recovery and provided evidence against relying solely on sequence abundances to select barcode sequences. Directly leveraging*a-priori* taxonomic data from validated specimens allowed accurate identification of non-target contaminant sequences, further stressing the value that taxonomically validated specimens can confer towards barcode generation. Similarly, taxonomic information was considered important for confirming the identity of insect pests detected in bulk trap catches by multi-locus metabarcoding [19].

Limitations and potential improvements

The lower rate of FC sequence recovery compared to BR sequence recovery implies that factors associated with FC PCRs, rather than sampling or DNA extraction, were the main cause of failures to obtain complete barcode sequences. The most likely explanation for this is that one or both primers used in FC PCRs have suboptimal matches with the specimens in question [48]. While it is unclear which of the FC primers (Ill_LCO1490 and Ill_C_R) might cause this problem, deficiencies of the LCO1490 / HCO2198 primer pair have been noted previously [49, 50], pointing to LCO1490 as problematic. These failures were concentrated in certain Coleoptera and Hymenoptera families, suggesting that the primer sequences may need adjustment to improve outcomes for these groups.

Improvements to our bioinformatic process may be possible. We separately identified FC and BR amplicons before attempting to align and merge those with expected taxonomic identities. It may seem intuitively simpler to merge all detected FC and BR sequences into putative barcodes, and then to identify the correct barcode among those based on taxonomy. However, there were unexpectedly high numbers of FC and BR sequences for many specimens after filtering and denoising, which would result in exceedingly high numbers of pairwise combinations of sequences requiring examination for correct taxonomic identity.

Conclusion

As the need for DNA-based biodiversity assessment continues to grow [12, 51], the development of fit-for-purpose and reliable reference databases follows. Further, the development of methods for non-destructive DNA extraction from museum/stored samples has created an opportunity to develop rapid barcoding technology. Along with recent attempts at using Illumina and PacBio technologies [12, 23], we provide another approach for rapid and high-throughput barcode generation. Our taxonomy-informed pipeline leverages the benefits and value of keeping specimens in taxonomic collections in the long-term and helps develop targeted

reference databases to support regional and national biodiversity surveys.

Acknowledgements

This work is supported via the Strategic Science Investment Fund from the Ministry of Business Innovation and Employment and supported via the B3 (Better Border Biosecurity) Science Collaboration Project #D17.22. The authors would like to acknowledge the use of the New Zealand eScience Infrastructure (NeSI) for the data analysis.

Author Contributions

MKD, AD & TRB conceived and designed the study, MKD & AD wrote the first draft of the manuscript with input from all the authors. TBC & AP performed the laboratory experiments and AD performed the analysis. Taxonomic identification of voucher specimens was undertaken by RABL (Coleoptera), DFW (Hymenoptera) and TRB (Annelida).

Authors declare no conflict of interest

Data & Code accessibility statement

Raw sequence reads and Sample metadata are deposited in the SRA (BioProject PRNJAXXXXX [will be

Bioinformatics pipeline is available on github: https://github.com/manaakiwhenua/Fast_DNA_barcoding

Benefit sharing statement:

Benefits Generated: Benefits from this research accrue from the sharing of our data and results on public databases as described above. The New Zealand Arthropod Collection continually engages with relevant iwi and tangata whenua representatives (indigenous Māori peoples of New Zealand) to ensure collection materials are sourced respectfully and with permission.

References:

1. Bohmann, K., et al., *Environmental DNA for wildlife biology and biodiversity monitoring.* Trends in ecology & evolution, 2014.**29** (6): p. 358-367.

2. Deiner, K., H. Yamanaka, and L. Bernatchez, *The future of biodiversity monitoring and conservation utilizing environmental DNA.*Environmental DNA, 2021. **3** (1): p. 3-7.

3. Harper, L.R., et al., *Prospects and challenges of environmental DNA (eDNA) monitoring in freshwater ponds.* Hydrobiologia, 2019.**826** (1): p. 25-41.

4. McGee, K.M., C.V. Robinson, and M. Hajibabaei, *Gaps in DNA-Based Biomonitoring Across the Globe.* Frontiers in Ecology and Evolution, 2019. **7** (337).

5. Schenekar, T., et al., *Reference databases, primer choice, and assay sensitivity for environmental metabarcoding: Lessons learnt from a re-evaluation of an eDNA fish assessment in the Volga headwaters.* River Research and Applications, 2020. **36** (7): p. 1004-1013.

6. Folmer, O., et al., *DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates.* Molecular Marine Biology and Biotechnology, 1994.**3** (5): p. 294–299.

7. China Plant BOL Group, *Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants.* Proceedings of the National Academy of Sciences, 2011. **108** (49): p. 19641-19646.

8. Ratnasingham, S. and P.D.N. Hebert, *BOLD : The barcode of life data system.* Molecular Ecology Notes, 2007. **7** : p. 355–364.

9. Marques, V., et al., *GAPeDNA: Assessing and mapping global species gaps in genetic databases for eDNA metabarcoding.* Diversity and Distributions.

10. Luo, A., et al., *Potential efficacy of mitochondrial genes for animal DNA barcoding: a case study using eutherian mammals.* BMC Genomics, 2011. **12** : p. 84.

11. Gold, Z., et al., *Improving metabarcoding taxonomic assignment: A case study of fishes in a large marine ecosystem.*Molecular Ecology Resources, 2021. **21** (7): p. 2546-2564.

12. D'Ercole, J., S.W.J. Prosser, and P.D.N. Hebert, *A SMRT approach for targeted amplicon sequencing of museum specimens (Lepidoptera)—patterns of nucleotide misincorporation.* PeerJ, 2021.**9** : p. e10420.

13. Winker, K., *Natural History Museums in a Postbiodiversity Era.* BioScience, 2004. **54** (5): p. 455.

14. Wandeler, P., P.E. Hoeck, and L.F. Keller, *Back to the future: museum specimens in population genetics.* Trends Ecol Evol, 2007.**22** (12): p. 634-42.

15. Hebert, P.D.N., et al., *A DNA 'Barcode Blitz': Rapid Digitization and Sequencing of a Natural History Collection.* PLOS ONE, 2013. **8** (7): p. e68535.

16. Shokralla, S., et al., *Pyrosequencing for Mini-Barcoding of Fresh and Old Museum Specimens.* PLoS ONE, 2011. **6** (7): p. e21252.

17. Boyer, S., et al., *Sliding Window Analyses for Optimal Selection of Mini-Barcodes, and Application to 454-Pyrosequencing for Specimen Identification from Degraded DNA.* PLoS ONE, 2012.**7** (5): p. e38215.

18. Lindahl, T., *Instability and decay of the primary structure of DNA.* Nature, 1993. **362** (6422): p. 709-715.

19. Batovska, J., et al., *Developing a non-destructive metabarcoding protocol for detection of pest insects in bulk trap catches.* Scientific Reports, 2021. **11** (1): p. 7946.

20. Carew, M.E., R.A. Coleman, and A.A. Hoffmann, *Can non-destructive DNA extraction of bulk invertebrate samples be used for metabarcoding?* PeerJ, 2018. **6** : p. e4980.

21. Wong, W.H., et al., *'Direct PCR' optimization yields a rapid, cost-effective, nondestructive and efficient method for obtaining DNA barcodes without DNA extraction.* Molecular Ecology Resources, 2014.**14** (6): p. 1271-1280.

22. Prosser, S.W.J., et al., *DNA barcodes from century-old type specimens using next-generation sequencing.* Molecular Ecology Resources, 2016. **16** (2): p. 487-497.

23. Shokralla, S., et al., *Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform.* Scientific Reports, 2015. **5** (1): p. 9687.

24. de Santana, C.D., et al., *The critical role of natural history museums in advancing eDNA for biodiversity studies: a case study with Amazonian fishes.* Sci Rep, 2021. **11** (1): p. 18159.

25. Moinet, G.Y.K., et al., *Soil microbial sensitivity to temperature remains unchanged despite community compositional shifts along geothermal gradients.* Global Change Biology, 2021.

26. Toju, H., et al., *Priority effects can persist across floral generations in nectar microbial metacommunities.* Oikos, 2018.**127** (3): p. 345-352.

27. Hamady, M., et al., *Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex.* Nature Methods, 2008.**5** (3): p. 235-237.

28. Tanabe, A.S. and H. Toju, *Two New Computational Methods for Universal DNA Barcoding: A Benchmark Using Barcode Sequences of Bacteria, Archaea, Animals, Fungi, and Land Plants.* PLoS ONE, 2013.**8** (10): p. e76910.

29. Zhang, J., et al., *PEAR: a fast and accurate Illumina Paired-End reAd mergeR.* Bioinformatics, 2014. **30** (5): p. 614-620.

30. Rognes, T., et al., *VSEARCH: a versatile open source tool for metagenomics.* PeerJ, 2016. **4** : p. e2584.

31. Rice, P., I. Longden, and A. Bleasby, *EMBOSS: The European Molecular Biology Open Software Suite.* Trends in Genetics, 2000.**16** (6): p. 276-277.

32. Katoh, K. and D.M. Standley, *MAFFT multiple sequence alignment software version 7: Improvements in performance and usability.*Molecular Biology and Evolution, 2013. **30** (4): p. 772–780.

33. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastTree 2 - approximately maximum-likelihood trees for large alignments.* PLoS ONE, 2010. **5** (3): p. e9490.

34. Yu, G., et al., *ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data.* Methods in Ecology and Evolution, 2017. **8** (1): p. 28-36.

35. Hebert, P.D.N., et al., *A Sequel to Sanger: amplicon sequencing that scales.* BMC Genomics, 2018. **19** .

36. Sire, L., et al., *The Challenge of DNA Barcoding Saproxylic Beetles in Natural History Collections—Exploring the Potential of Parallel Multiplex Sequencing With Illumina MiSeq.* Frontiers in Ecology and Evolution, 2019. **7** : p. 495.

37. Sint, D., L. Raso, and M. Traugott, *Advances in multiplex PCR: balancing primer efficiencies and improving detection success.* Methods in Ecology and Evolution, 2012. **3** (5): p. 898-905.

38. Hajibabaei, M., et al., *A minimalist barcode can identify a specimen whose DNA is degraded.* Molecular Ecology Notes, 2006.**6** (4): p. 959-964.

39. Paniagua Voirol, L.R., et al., *How the 'kitome' influences the characterization of bacterial communities in lepidopteran samples with low bacterial biomass.* Journal of Applied Microbiology, 2021.**130** (6): p. 1780-1793.

40. Sicard, M., M. Bonneau, and M. Weill, *Wolbachia prevalence, diversity, and ability to induce cytoplasmic incompatibility in mosquitoes.* Current Opinion in Insect Science, 2019. **34** : p. 12-20.

41. Potapov, V. and J.L. Ong, *Examining Sources of Error in PCR by Single-Molecule Sequencing.* PloS one, 2017. **12** (1): p. e0169774-e0169774.

42. Song, H., et al., *Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified.* Proc Natl Acad Sci U S A, 2008.**105** (36): p. 13486-91.

43. Minich, J.J., et al., *Quantifying and Understanding Well-to-Well Contamination in Microbiome Research.* mSystems, 2019.**4** (4): p. e00186-19.

44. Schnell, I.B., K. Bohmann, and M.T.P. Gilbert, *Tag jumps illuminated – reducing sequence-to-sample misidentifications in metabarcoding studies.* Molecular Ecology Resources, 2015.**15** (6): p. 1289-1303.

45. Eisenhofer, R., et al., *Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations.* Trends Microbiol, 2019.**27** (2): p. 105-117.

46. Bohmann, K., et al., *Strategies for sample labelling and library preparation in DNA metabarcoding studies.* Molecular Ecology Resources. **n/a** (n/a).

47. Sze, M.A. and P.D. Schloss, *The Impact of DNA Polymerase and Number of Rounds of Amplification in PCR on 16S rRNA Gene Sequence Data.* mSphere, 2019. **4** (3): p. e00163-19.

48. Elbrecht, V., et al., *Validation of COI metabarcoding primers for terrestrial arthropods.* PeerJ, 2019. **7** : p. e7745.

49. Lobo, J., et al., *Enhanced primers for amplification of DNA barcodes from a broad range of marine metazoans.* BMC Ecology, 2013.**13** (1): p. 34.

50. Geller, J., et al., *Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys.* Mol Ecol Resour, 2013. **13** (5): p. 851-61.

51. Gibson, J.F., et al., *Large-Scale Biomonitoring of Remote and Threatened Ecosystems via High-Throughput Sequencing.* PLOS ONE, 2015.**10** (10): p. e0138432.

Figures

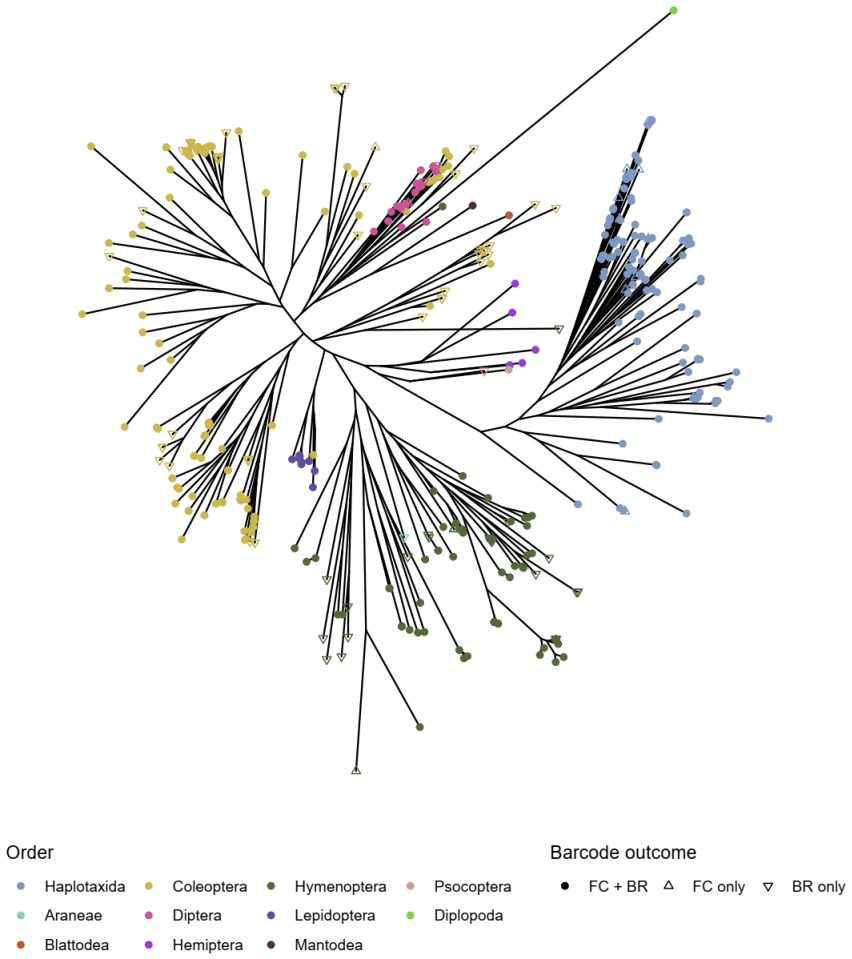

Figure 1. An approximately maximum-likelihood phylogeny of 334 full-length (FC + BR) and 105 partial (FC only or BR only) COI barcodes recovered from 450 invertebrate specimens using Illumina sequencing, generated using FastTree 2 [33].

Tables

Table 1. PCR outcomes and DNA barcode recovery from 450 invertebrate specimens. PCR values are the numbers of FC and BR PCRS that resulted in a visible PCR product. Alignment values are the mean numbers of pairwise FC-BR alignments with 85 bp overlap and 100 % sequence identity with ranges in parentheses, among [?] 20 most abundant filtered and denoised FC and BR sequences per specimen (or all sequences for 12 specimens).

Phylum	Class	Order	N	PCRs FC	PCRs BR	Alignments	Barcodes Full	Barcodes FC only	Barcodes BR only

Phylum	Class	Order	N	PCRs	PCRs	Alignments	Barcodes	Barcodes	Barcodes
Annelida	Clitellata	Haplotaxida	132	120	109	11.7 (0-105)	116	11	2
Arthropoda	Arachnida	Arachnida	2	0	2	1 (0-2)	0	0	1
	Insecta	Coleoptera	160	41	144	15.1 (0-133)	106	1	50
		Diptera	21	8	16	9.8 (0-42)	17	1	3
		Hemiptera	5	4	5	66.8 (10-122)	5	0	0
		Hymenoptera	114	46	94	10.6 (0-180)	75	4	31
		Lepidoptera	11	11	10	17.1 (1-125)	11	0	0
		Others	4	3	3	48.2 (0-102)	3	0	1
	Myriapoda	Diplopoda	1	0	1	10	1	0	0
		Totals	450	233	384	-	334	17	88

Table 2. Disparities between FC and BR sequence detection rates among different taxonomic families. FC and BR recovery rates represent the proportions of specimens for which full-length, plus either FC-only or BR-only barcodes, respectively, were detected.

Phylum	Class	Order	Family	Specimens	FC recovery	BR recovery	Disparity (%
Annelida	Clitellata	Haplotaxida		132	96	89	7
Arthropoda	Arachnida	Arachnida		2	0	50	-50
	Insecta	Coleoptera	Carabidae	12	83	100	-17
			Cerambycidae	6	83	100	-17
			Chrysomelidae	107	67	97	-30
			Coccinellidae	7	86	100	-14
			Curculionidae	5	80	100	-20
			Staphylinidae	10	0	100	-100
			Others	13	77	92	-15
		Diptera	Ephydridae	5	80	80	0
			Others	16	88	100	-12
		Hemiptera		5	100	100	0
		Hymenoptera	Bethylidae	6	83	100	-17
			Braconidae	87	69	92	-23
			Ichneumonidae	7	100	100	0
			Others	14	50	93	-43
		Lepidoptera		11	100	100	0
		Others		4	75	100	-25
	Myriapoda	Diplopoda		1	100	100	0
			Mean	24	75	94	-20