Debar: A sequence-by-sequence denoiser for COI-5P DNA barcode data

Cameron M. Nugent1,2 | Tyler A. Elliott2 | Sujeewan Ratnasingham2 | Paul D.N. Hebert2 | Sarah J. Adamowicz1

1Department of Integrative Biology, University of Guelph, Guelph, ON, Canada
2Centre for Biodiversity Genomics, University of Guelph, Guelph, ON, Canada

Correspondence
Cameron M. Nugent, Department of Integrative Biology, University of Guelph, Guelph, ON, Canada.
Email: nugentc@uoguelph.ca

Abstract
DNA barcoding and metabarcoding are now widely used to advance species discovery and biodiversity assessments. High-throughput sequencing (HTS) has expanded the volume and scope of these analyses, but elevated error rates introduce noise into sequence records that can inflate estimates of biodiversity. Denoising—the separation of biological signal from instrument (technical) noise—of barcode and metabarcoding data currently employs abundance-based methods which do not capitalize on the highly conserved structure of the cytochrome c oxidase subunit I (COI) region employed as the animal barcode. This manuscript introduces debar, an R package that utilizes a profile hidden Markov model to denoise indel errors in COI sequences introduced by instrument error. In silico studies demonstrated that debar recognized 95% of artificially introduced indels in COI sequences. When applied to real-world data, debar reduced indel errors in circular consensus sequences obtained with the Sequel platform by 75%, and those generated on the Ion Torrent S5 by 94%. The false correction rate was less than 0.1%, indicating that debar is receptive to the majority of true COI variation in the animal kingdom. In conclusion, the debar package improves DNA barcode and metabarcoding workflows by aiding the generation of more accurate sequences aiding the characterization of species diversity.

KEYWORDS
biodiversity, COI, denoising, DNA barcode, Markov model, metabarcoding

1 | INTRODUCTION

Motivated by global biodiversity decline, conservation policies and strategies are being implemented to mitigate extinction rates (Baynham-Herd et al., 2018; Driscoll et al., 2018). Accurate assessments of biodiversity and its change over time are critical to support conservation strategies, to remediate environmental damage, and to manage natural resources, but this information is lacking for most ecosystems (D’Souza & Hebert, 2018; Hajibabaei et al., 2016; Hebert et al., 2016; Sogin et al., 2006).

DNA barcoding provides a technological solution to the problem of identifying organisms and characterizing biodiversity (Hebert et al., 2003; Hubert & Hanner, 2015). Instead of identifying specimens through morphological study, standardized DNA regions—termed DNA barcodes—are used to identify specimens belonging to known species and to recognize new taxa. Reflecting advances in sequencing technology, DNA barcode studies are expanding in scale from analysing single specimens to characterizing bulk samples, an approach termed metabarcoding, as well as multimarker and metagenomics approaches (Cristescu, 2014; Hajibabaei et al., 2016;
Taberlet et al., 2012; Wilson et al., 2019). These advances are providing newly detailed information on species diversity in different geographic regions and habitats (Delabaye et al., 2019; Hajibabaei et al., 2012; Hebert et al., 2016; Lopez-Vaamonde et al., 2019) while also aiding the identification of invasive species (Brown et al., 2016), food web analysis (Kanuisto et al., 2017; Wirta et al., 2014; Zhang et al., 2021), and environmental monitoring (Cordier et al., 2019; Hajibabaei et al., 2016; Stat et al., 2017).

Despite the broad adoption of DNA barcoding and metabarcoding, a fundamental problem persists. Efforts to quantify biodiversity from barcode and metabarcoding data can be strongly affected by analytical methodology (Braukmann et al., 2019; Clare et al., 2016). For example, if high-throughput sequence (HTS) data are cleaned sub-optimally, the estimated number of taxa may be grossly inflated as variation introduced by sequencing (technical) errors are interpreted as biological variation (Hardge et al., 2018).

To reduce the impact of technical errors, sequence reads are often clustered into operational taxonomic units (OTUs) at specific identity thresholds (Elbrecht et al., 2018). Several software packages have attempted to increase the accuracy of this OTU method by separating biological signal from technical noise (Amir et al., 2017; Callahan et al., 2016; Edgar, 2016; Elbrecht et al., 2018; Kumar et al., 2019; Nearing et al., 2018; Rosen et al., 2008). Many standard denoisers, such as DADA2 (Callahan et al., 2016), Deblur (Amir et al., 2017), and UNOISE (Edgar, 2016), utilize cluster-based approaches, custom error models, or preclustering algorithms to account for and correct technical errors. Comparative studies have shown that all three of these methods outperform threshold-based OTU-clustering approaches (Nearing et al., 2018). It has also been shown that they produce similar estimates of species richness and relative abundance, but significantly different values for alpha diversity (intrahabitat diversity) and the number of unique exact sequence variants (ESVs) (Nearing et al., 2018). When a highly conserved protein-coding region, such as cytochrome c oxidase subunit I (COI), is employed as the barcode, structural information can be leveraged to improve denoising. The adoption of this approach can improve the accuracy of alpha-diversity estimates and the quality of identified barcode sequences by ensuring barcodes conform to biological reality. Additionally, rare sequences or important intraspecies variants need not be discarded based solely on their abundance and can be retained with higher confidence if they conform to the expected gene structure. This latter benefit will be particularly valuable for work on hyperdiverse communities, (e.g., tropical insects) and for analyses of metabarcoding data, where uneven sampling is often the norm and the resolution of intraspecies variation is challenging (Braukmann et al., 2019; Elbrecht et al., 2018; Nearing et al., 2018; Zizka et al., 2020).

Hidden Markov models (HMMs) are probabilistic representations of sequences that allow unobserved (hidden) states to be inferred through the observation of a series of nonhidden states (Durbin et al., 1998; Wilkinson, 2019). HMMs have been applied widely in the analysis of biological sequences, in areas such as sequence alignment, annotation, and species identification (Durbin et al., 1998; Eddy, 1998; Shi et al., 2018). Profile Hidden Markov models (PHMMs) are a variant well suited for the representation of biological sequences with a shared evolutionary origin (Durbin et al., 1998; Eddy, 1998, 2009). They are probabilistic models that contain position-specific information about the likelihood of potential characters (base pairs [bp] or amino acid residues) at the given position in the sequence (emission probabilities) and the likelihood of the observed character given the previously observed character in the sequence (transition probabilities). Once a PHMM is trained on a set of sequences, the Viterbi algorithm can be used to obtain the path of hidden states that align the novel sequences to the PHMM (Durbin et al., 1998). The Viterbi path is comprised of hidden match states (indicating the observed character matches to a position in the PHMM) and nonmatch states: either inserts or deletions. In the context of error correction, hidden nonmatch states identify the most likely positions at which novel sequences deviate from the PHMM’s statistical profile. In this manner, individual sequences can be queried for evidence of insertion or deletion (indel) errors and adjusted in a statistically informed manner. The conserved protein-coding structure of the most common animal barcode gene, COI, and the wealth of available training sequences (Ratnasingham & Hebert, 2013) for this region have allowed PHMMs to be successfully applied in the detection of technical errors in novel barcode sequences (Nugent et al., 2020). Correction of technical indel errors in data from protein-coding barcode sequences is an important development as it maximizes the likelihood that both the nucleotide and amino acid sequences correspond to the true biological sequence. Mitigation of indels arising due to technical errors also makes sequence reads from a given specimen more directly comparable, allowing low-frequency point mutations to be eliminated when multiple reads are available for a given biological sequence. Here, we aim to extend the use of PHMMs in COI data processing to allow for the sequence-by-sequence correction (denoising) of technical errors.

This study had four primary goals: (i) design a denoising tool for COI barcode data that utilizes PHMMs to identify and correct insertion and deletion errors resulting from technical error, (ii) test the tool’s performance and optimize its default parameters by denoising a set of 10,000 barcode sequences with artificially introduced indel errors, (iii) develop, implement, and evaluate a workflow for denoising DNA barcode data produced through single-molecule, real-time (SMRT) sequencing of 29,525 specimens on the Sequel platform (Pacific Biosciences), and (iv) denoise a DNA metabarcoding mock community data set using debar and evaluate the improvement in quality of consensus sequences and the ability to resolve intra-OTU haplotype variation. The denoiser resulting from this work, debar (DEnoising BARcodes), is a free, publicly available package written in R that is available through CRAN (https://CRAN.R-project.org/package=debar) and GitHub (https://github.com/CNuge/debar).

2 | MATERIALS AND METHODS

2.1 | Implementation

The debar utility includes several customizable steps which denoise DNA barcode and metabarcoding data (Figure 1; Data S1). Corrections with debar are based upon the comparison of input
sequences with a nucleotide-based profile hidden Markov model (PHMM) (model training detailed in Nugent et al., 2020) using the Viterbi algorithm (Durbin et al., 1998). Briefly, debar's PHMM was trained using a curated set of 11,387 COI-5P barcode sequences obtained from the Barcode of Life Data Systems (BOLD: www.boldsystems.org) public database that were checked to ensure: (i)
the sequence was >600 bp in length, (ii) taxonomy was known to a genus level, (iii) there were no missing base pairs, (iv) the amino acid sequence did not contain stop codons, and (v) BOLD’s internal check for contaminants was negative (Nugent et al., 2020). The Viterbi path produced through alignment of the sequence to the PHMMs is used to match the input sequence to the PHMM (by finding the first set of 10 consecutive match states which indicate the absence of indels for the given 10 base pairs). The read is then adjusted to account for detected insertions or deletions (Figure 1). Three consecutive nucleotide insertions or deletions are permitted (not adjusted) as sequences of this kind are more likely to reflect true biological variants than technical errors (they do not result in reading frame shifts and may reflect an insertion or deletion of an amino acid in a functional protein-coding gene). The probability of such changes through sequencing error is relatively low (i.e., for the Pacific Biosciences Sequel platform the baseline probability of three consecutive deletions would be 0.05% (baseline delete probability) cubed, or 0.000125%).

The denoising of sequences with debar is controlled using a suite of parameters (Figure 1). The censorship parameter is most important as it controls the size of the masks (substitution of nucleotides for placeholder N characters) applied around sequence adjustments. This option is designed to prevent the introduction of errors that would be caused if the denoising process deleted the wrong base pair or inserted a placeholder in the incorrect position. Derivation of the default value for the censorship parameter is detailed in the Methods and Results sections. The package also enables the translation of denoised sequences to amino acids to confirm that denoised outputs conform to the expected properties of the protein-coding gene region. Because debar can interface directly with fasta and fastq files, it enables file-to-file denoising in addition to denoising within an R programming environment. The default PHMM used for denoising by debar represents the complete 657 bp barcode region of COI. The package also permits the use of customized PHMMs provided by a user, which allows the denoiser to be applied to data from other gene regions or for the denoiser to be targeted to a specific user-defined subsection of the COI barcode. Training of a PHMM for a new barcode or gene is supported by the R package aphid (Wilkinson, 2019), while subsetting of debar’s default PHMM is enabled by the R package coil (Nugent et al., 2020). Details of the package’s components together with a demonstration of its implementation is available in the package’s vignette (Data S1).

2.2 | Quantification of package performance

2.2.1 | Simulated error data

The debar package was tested using a phylogenetically stratified random sample of publicly available COI-5P sequences with artificially introduced indels. This test was designed to assess the accuracy of sequence corrections and to obtain a quantitatively informed set of default parameters for the denoising process. A random sample of 10,000 animal COI-5P sequences (excluding those used in PHMM model training) were obtained from BOLD and cleaned using the steps described in Nugent et al., (2020) (methods section – BOLD data acquisition). Errors were introduced into each sequence in accordance with the statistical error profile of the Pacific Biosciences Sequel based upon the error profile for COI barcode region in Hebert et al., (2018). This profile indicated a baseline indel rate of 0.1% (insertions and deletions equally likely), a baseline substitution rate of 0.5%, and an elevated indel rate for long homopolymers (repeat length of 6.7, and 8+ with indel probabilities of 0.75%, 1.2%, and 3.8%, respectively) (Hebert et al., 2018). The location of all errors was recorded so that accuracy of subsequent corrections could be evaluated. Sequences were iteratively processed, and errors were limited to a single insertion or deletion error of one base pair in length (with the error introduction process being repeated for the original sequence when more than one indel occurred), which allowed for the accuracy of corrections to be assessed without the need to consider interaction effects.

The resultant sequences, each with one indel, were then denoised with debar (“denoise” function, using the parameter censor_length = 0). The outputs of the denoise function were queried to determine the number and location of indel corrections applied by debar. This information was compared to the recorded ground truth error locations to quantify the following: (i) the frequency with which debar located and exactly corrected indels, (ii) the miss distance (number of nucleotide positions) between introduced errors and corrections applied in instances where debar did not correct the indel errors in exactly the correct position, and (iii) the frequency at which debar applied an incorrect number of sequence corrections (i.e., 0 correction or 2+ corrections). If one correction was made and the distance between the correction and true indel position was 0, then the correction was considered accurate. Corrections were also considered accurate if all base pairs between the correction location and the true indel position were the same (i.e., if base pair 2 in the homopolymer “TTTTT” was an insertion, but the fifth T in the sequence was removed by debar, this is functionally an exact correction as the true sequence is restored). All other corrections at inexact positions were considered inaccurate, and the distance (number of positions) between the correction and true indel location was recorded. The mean and standard deviation of the miss distance were determined and used to select the default censor_length parameter for the debar package, equal to the mean miss distance plus 2 standard deviations (censor_length = ceiling (miss_distance + 2 x σmiss_distance). This value was selected as it would be expected to avoid the introduction of an error for >95% of inexact corrections. Sequences where no corrections or multiple corrections were made had their outputs inspected further to determine if other parts of the denoising pipeline (e.g., the check for stop codons in the translated amino acid sequence or trimming of sequence edges in the framing process) removed the error or led to the complete rejection of the sequence.
2.2.2 False correction rate

The performance of debar on sequences with no indel errors was also quantified to determine the frequency and cause of erroneous corrections applied to cleaned, publicly available COI-5P barcode sequences with no known technical errors. A random sample of 10,000 sequences from all the animal COI-5P barcode sequences available on BOLD was obtained (Data S2) meeting the following criteria was obtained: (i) the barcode was publicly available on the BOLD database, (ii) the barcode was >600 bp in length, (iii) the barcode did not contain missing characters ("N") in the Folmer region, (iv) the corresponding amino sequence did not contain stop codons, (v) the result of BOLD's internal check for contaminants was negative, and (vi) the sequence was not used in PHMM training and the simulated error data set. Sequences were processed using debar's denoise function (censor_length = 0). All sequences that had corrections applied, or that were flagged for rejection, were counted and examined in detail to search for evidence of the proximal cause of the false correction. To search for evidence of taxonomic bias, the taxonomy associated with all falsely corrected sequences were tallied at the order level, and manually examined for evidence of bias.

2.2.3 Denoising PacBio sequel data

We quantified the performance of debar on raw DNA barcode sequence data by interfacing with the existing mBRAVE workflow (http://www.mbrave.net) used to process DNA barcode circular consensus sequences (CCS) obtained with the Sequel platform. A
custom analysis pipeline (Data S3) was constructed to analyse and denoise the final set of CCS barcodes produced by the mBRAVE workflow (one CCS per OTU) (Figure 2). The pipeline was designed to search the final barcodes produced by mBRAVE for evidence of indel errors (by considering the translated amino acid sequence with the R package coil (Nugent et al., 2020)), denoise all the associated CCS with detected errors using the debar package, and then regenerate a consensus barcode sequence using the denoised data to produce a final, denoised barcode sequence for each specimen (Figure 2).

The outputs of this analysis were examined to determine if the debar pipeline decreased the number of technical errors in the barcode sequences and that those barcode sequences resulted in likely amino acid sequences when translated. Initial quantification of the improvement was conducted by comparing the number of barcode sequences whose amino acid sequences were flagged by the R package coil (Nugent et al., 2020, default parameters) before and after denoising. Barcodes are flagged by coil when they possess a stop codon when translated to amino acids or when the resultant amino acid sequence is improbable, both indicating that the sequence probably possesses an indel error.

Since the coil and debar packages both employ the same nucleotide profile hidden Markov model (coil also utilizes an amino acid PHMM), an independent test of pipeline effectiveness was also conducted. The effectiveness of the denoising pipeline was quantified by submitting both the original and denoised barcode sequences to BOLD. It was used to determine the number of original barcodes and denoised barcodes with evidence of stop codons after aligning the sequences using the BOLD’s hidden Markov model (a model developed independently of the debar PHMM) and translating the sequence using the appropriate translation table corresponding to the taxonomic information accompanying the sequence record. Comparison of these numbers made it possible to quantify the increase in barcode-compliant sequences (i.e., those with no stop codon) produced by debar. Additionally, the Sequence Quality Report on BOLD was examined to determine the number of unknown nucleotides (“N”) in the barcode sequences after denoising. The report categorizes barcode quality as: high (<1% Ns), medium (<2% Ns), low (<4% Ns), or unreliable (>4% Ns), and the number of barcodes in these different categories was recorded.

2.2.4 | Denoising metabarcode data

To characterize debar’s performance on metabarcode data, we analysed a metabarcode data set for a mock arthropod community (Braukmann et al., 2019). These data derived from a single sequencing run on an Ion Torrent S5 on COI amplicons generated by pooled DNA extracts from abdomens from single specimens of 369 arthropod species (methods described in detail in Braukmann et al., 2019). Sequences were from a 407 bp fragment of the COI barcode region targeted using the primers MlepF1 and LepR1 (Hebert et al., 2004; Braukmann et al., 2019). Following amplification and sequencing on the Ion S5, quality control, sequence dereplication, chimeric read filtering, matching to reference sequences, and clustering were performed on mBRAVE (Braukman et al., 2019). Two sets of data resulted from this process, a set of 123,926 unique sequences that were assigned to 398 different Barcode Index Numbers (BINs) (Ratnasingham & Hebert, 2013) through the comparison to reference sequences (matched at >98% similarity), and a set of 2,199 unique sequences not matching to available references that were clustered into an additional 1,255 OTUs at a 97% similarity threshold (using clustering algorithm described in Braukmann et al., 2019). All sequences were denoised using debar’s denoise_list function and a custom nucleotide PHMM. The custom PHMM was a 398 bp subset of the complete COI PHMM (PHMM profile positions 250–648), corresponding to a segment of the Folmer (Folmer et al., 1994) region targeted by the metabarcoding primers. The PHMM was
created using coil’s “subsetPHMM” function (Nugent et al., 2020). After denoising, two tests were conducted to determine if denoising improved the quality of the metabarcode pipeline’s output data.

First, for each BIN and OTU consensus sequences were generated using denoised sequences and the debar function ”consensus_sequence”. These consensus sequences were assessed for evidence of stop codons using coil and the same custom PHMMs used in denoising (function coilSp.pipe with the additional parameter: trans_table = 5). This test revealed the number of denoised consensus sequences which contained a stop codon when translated to amino acids, indicating an indel error persisted in the nucleotide sequence. The centroid sequences for the BINs and OTUs were used as a baseline metric for the number of barcode-compliant sequences. For each BIN, centroid sequences were obtained by clustering the sequences in the group using the R package kmer’s ”ut” function (parameters: k = 4, threshold = 0.95) (Wilkinson, 2018, Version 1.0.0). For the OTUs, centroids were obtained from data generated by mBRAVE. All centroids were assessed with coil (Nugent et al., 2020, Version 1.0.0), and the number of barcode-compliant representative sequences for the original centroids and the final consensus sequences was compared.

Secondly, the individual sequences within each BIN and OTU were analysed with coil to determine the number that were probably error free, as evidenced by the absence of stop codons after translation. This assessment was repeated on the denoised reads to determine the effectiveness of debar in correcting errors in individual sequences and to reveal if the denoising process improved the resolution of ESVs for subsequent analysis of intraspecies genetic variation by placing the ESVs in reading frame and reducing the frequency of identified indel errors.

3 | RESULTS

3.1 | Quantification of package performance

3.1.1 | Simulated error data

Debar was used to correct 10,000 barcodes, each with a single indel error (Data S1). The denoised sequences and associated data were compared to the ground truth error locations to determine the accuracy of corrections applied by debar (Figure 3). For 9,459 sequences (95.59%), a single correction was applied by debar, indicating that the package correctly identified the type of error in these sequences. However, debar either failed to recognize an indel or made too many corrections (2+) in the other 541 sequences. No correction was made for most (426) of these sequences, meaning that debar’s PHMM did not identify the indel error. The overlooked indels were largely restricted to the terminal regions of the sequence; 75% (329/426) of them were positioned within 20 base pairs of the read termini (Figure 4), regions that only comprised 5% (40 bp/650 bp) of the sequences. The cause of this is that the debar denoising algorithm uses the first observation of 10 consecutive bp matching to the PHMM to establish the corrective window. Errors on the periphery of sequences therefore lead to trimming of the sequence (via the keep_flanks function) instead of indel correction. A substantial fraction of the remaining uncorrected indel errors (43) occurred between positions 452 to 465 (Figure 4), a region associated with a 3 bp indel present in some animal groups and absent in others. Its presence reduced the PHMM’s indel detection ability in this region due to greater true variability. Not all unidentified indels were retained in the final output sequences as double checks of debar (employing the keep_flanks and aa_check parameters) identified many (266/426 – 62%) of the uncorrected sequences and either omit the problem region or flag the sequence as likely to contain an error. Therefore, debar’s double checks allow many false negatives to be trimmed or flagged as problematic.

For 119 sequences (1.2%), two or more corrections were applied by debar when only a single indel existed (Figure 3). In contrast to the false negatives, debar’s double checks only captured three of the false positives. Many of the false corrections appeared to be the presence of indels near codons that are not present in all animals. Due to true biological variation in the training data, these regions of the PHMM have higher probabilities of transitioning from a match state to an insert or delete state, and therefore indels in these locations are sometimes handled incorrectly (i.e., the sequence is characterized as having two deleted base pairs, when there was a 1 bp insertion). Because false corrections of this type result in sequences that conform to the structure of the protein-coding gene region (i.e., a lack of stop codons in the amino acid sequence), they are not identified by debar’s aa_check function.

The 9,459 sequences for which the presence of a single indel was correctly identified were further analysed to determine how accurately they were located (Figure 3). The analysis showed that debar was able to exactly locate and correct 5,847 (61.81% of sequences in single correction category) of the indel errors in the data set. For the other 3,612 sequences (38.19% of the single corrections category), the indel corrections were not placed in exactly the correct position (Figure 5). For these sequences, the average distance between the true indel location and the applied correction was 2.31 base pairs (standard deviation = 1.9767).

These results were used to select a default censorship value for debar to ensure that inexact indel errors are masked in most sequences (Figure 1). A default censorship length of 7 (the average miss distance plus two times the standard deviation, rounded up) was selected in order to mask the true error in >95% of instances where inexact corrections were applied, thereby successfully denoising sequences, albeit with some associated loss of information in the sequences, which can be overcome by building a consensus sequence when multiple reads are available for an individual.

Overall, denoising of the 10,000 barcodes with the default censorship parameter (censor_length = 7) resulted in 9,309/10,000 (93.09%) of sequences with errors being successfully denoised. The additional double check parameters (aa_check = True, keep_flanks = False) captured, but did not correct, 269 (2.69%) errors. The debar package thereby corrected or removed 95.74% of sequences with indel errors (Figure 3).
3.1.2 | False correction rate

A set of 10,000 barcode sequences with no known indel errors was analysed with debar to determine the incidence of erroneous corrections. Nearly all sequences (99.91%) were not altered nor flagged as erroneous. Nine sequences were erroneously corrected, and none were flagged for rejection. These sequences included a single sequence from each of five orders and four sequences from the order Diptera (flies). Interestingly, the four Diptera sequences that were incorrectly altered all belonged to the same genus: Culicoides. They represented 4/58 of all sequences from the family Ceratopogonidae that were in data set, indicating that the performance issue was isolated to this single genus.

These results indicate that debar deals well with variation in COI sequences across most of the animal kingdom, but that it displays some taxonomic bias in performance. This is a limitation of debar, as any genus with a COI profile that systematically deviates from the COI PHMM used in debar will be erroneously denoised. The benefit of the conservative censorship approach used in the package is that although these reads are erroneously adjusted, the corrections made are masked by Ns, and the entire sequence is not rejected. Rather, only a small section of the sequences is lost, as if it were to contain an indel error. Most of any falsely corrected sequences can thereby be recovered, and in most instances, this would be sufficient to identify associated taxonomy and inform biological conclusions.

3.1.3 | Denoising PacBio sequel data

We applied debar in the analysis of real DNA barcode data by developing a processing pipeline (Figure 2 – hereafter “the debar pipeline”) and compared the amount of technical noise in the barcodes before and after processing. A set of 29,525 consensus barcode sequences derived from processing data from four Sequel runs were obtained from mBRAVE and were re-processed with the debar pipeline (Table 1).

Analysis of the consensus barcodes with coil (step ii. of the debar pipeline) flagged 3,495 (11.8% of total) of consensus sequences due to the detection of a stop codon in the translated sequence or due to the presence of an unexpected amino acid (log likelihood score below the default threshold). The large number of flagged sequences is probably reflective of false positives (sequences flagged by coil that lack indel errors due to the incorrect establishment of reading frame). In fact, 2,418 sequences (8.1% of total, 69.2% of flagged sequences) were flagged due to the presence of a stop codon, and 1,282 of them (4.3% of total, 36.7% of flagged sequences) contained a stop codon in all three forward reading frames, providing extremely strong evidence of an indel error (i.e. a low likelihood of being a false positive).

After denoising, the output sequences were again assessed with coil (step viii. of the debar pipeline) and this analysis revealed that debar had corrected many indel errors (Table 1, Table 2). Only
1,123 (3.8%) of the final barcode sequences were flagged by coil's coi5p_pipe function, suggesting that 66.8% (2,335) of the flagged sequences were successfully denoised. When comparison was restricted to the 2,418 sequences with stop codons, only 176 were still flagged as containing stop codons, indicating that 92.7% (2,242/2,418) of the sequences in this subcategory were effectively denoised. A more conservative estimate of correction success was provided by the subset of flagged sequences with stop codons in all reading frames. Of these sequences, 1,106/1,282 (86.27%) passed the coil check following denoising, suggesting the successful correction of an indel error and improved representation of the true sequence.

External quantification of the debar pipeline's denoising ability was obtained by the submission of pre- and post-pipeline barcode sequences to BOLD (http://www.boldsystems.org). The sample size for this test was smaller as BOLD requires taxonomic designations and this information was only provided by mBRAVE for 27,041 sequences. The total number of original sequences flagged
TABLE 2 Assessment of the correction ability of the debar pipeline for the subset of sequences in the high-confidence error set

| PacBio Sequel run | Run 1 | Run 2 | Run 3 | Run 4 | Total |
|-------------------|-------|-------|-------|-------|-------|
| Original flagged  | 551   | 547   | 609   | 610   | 2,317 |
| Flagged post-     | 254   | 280   | 300   | 271   | 1,105 |
| denoising         |       |       |       |       |       |
| Corrected         | 53.9% | 48.8% | 50.7% | 55.6% | 52.3% |
| Original stop     | 319   | 295   | 318   | 350   | 1,212 |
| codon             |       |       |       |       |       |
| Stop codon        | 43    | 42    | 36    | 55    | 176   |
| post-denoising    |       |       |       |       |       |
| Corrected stop    | 86.5% | 85.7% | 88.7% | 84.2% | 86.3% |
| codons            |       |       |       |       |       |

Note: This set of sequences was flagged by coil and produced a stop codon when translated within all reading frames. The top half of the table indicates the number of sequences flagged by coil as likely to be erroneous, based on the log likelihood values of the sequences. Results are shown for sequences both before and after the denoising process. The bottom half of the table contains the number of sequences flagged by coil as likely to be erroneous, based on the presence of a stop codon in the amino acid sequence resulting from the censored translation of the framed nucleotide sequence. This high success for the stop-codon metric (86.3% of errors removed) indicates that the pipeline is an effective means of correcting frameshift-causing insertion or deletion errors. The relatively lower success at correcting sequences with low log likelihood values suggests that frameshift-causing errors are not the only set of errors being flagged by coil, and that nonframeshift errors are not effectively corrected by the debar pipeline.

by BOLD due to its detection of a stop codon was 1,515 (6.3%), a considerably lower frequency than reported by coil on the initial pipeline inputs. Of the 1,515 sequences with initial evidence of stop codons, 14 were rejected outright by the debar pipeline, 223 were flagged but not successfully corrected, 147 were unflagged and not corrected, and 1,131 had no evidence of errors following denoising (Table 3). Based on this assessment with BOLD, the debar pipeline produced a 75% reduction in the number of errors in the data set from 6.3% (1,515) to 1.6% (384). Of the remaining 384 errors, the majority (223) were detected as problematic and flagged as erroneous by debar. As a consequence, the debar pipeline reduced the number of unidentified errors by >90% (from 1,515 to 147) in the barcode data set (Table 3).

The denoising of the barcodes with the debar pipeline did not result in sequences with large amounts of missing information. Of the 29,525 output barcodes, 28,802 were high quality (<1% Ns), 11 were medium quality (<2% Ns), 498 were low quality (<4% Ns), and 214 were unreliable (>4% Ns). There was a strong negative relationship between the number of CCS available for a sample and the amount of missing information in the final barcode sequence (Figure 6).

3.2 | Denoising metabarcode data

3.2.1 | Consensus sequence quality

Metabarcoding data from a mock arthropod community were also denoised followed by comparison of original sequences to the denoised consensus sequences to determine if the debar improved sequence quality (Table 4). Of the original centroid sequences for the 398 BINs, 125/398 (31.4%) contained evidence of indel errors when analysed with coil. Following denoising and consensus sequence generation via debar, the number of barcode-compliant outputs was considerably higher with only 7/394 (1.8%) displaying evidence of indel errors. Four BINs had all their component sequences rejected by debar so no consensus sequences were generated. The rate of apparent indel errors was higher in the centroids of the 1255 OTUs; 681 (54%) displayed evidence of a stop codon when analysed with coil, suggesting the presence of indels in more than half of the sequences representing each OTU. The consensus sequences produced through denoising and consensus sequence generation with debar were of apparent higher quality as only 134 (10.6%) displayed evidence of a stop codon when analysed with coil. An additional 31 OTUs (2.5%) failed to produce a valid consensus sequence after denoising because all their component sequences were rejected by debar.

The corrections did cause some loss of information; 46/394 (11.7%) of the consensus sequences for the BIN groups contained at least one “N” due to ambiguous or censored base pairs in their component reads, and 861/1255 (68.6%) of the OTU consensus sequences contained at least one “N”. The number of “Ns” per sequence was generally low for the BINs (median = 0; 12 sequences with 14 or more “Ns”) but was higher for the OTUs (median number of “Ns” = 15), indicating there was on average one correction per OTU (correction of an indel, plus the seven bp mask in either direction result in 14 (insertion) or 15 (deletion) consecutive “Ns”). There was a positive relationship between the number of sequences within an OTU and the completeness of information in the final consensus sequence.

3.2.2 | ESV data quality

Data analysis on mBRAVE revealed 398 BINs represented by 123,926 unique dereplicated reads as well as 1255 OTUs lacking taxonomic assignment that were represented by 2,199 unique sequence reads. When original sequences were checked with coil, it indicated that 61,351/123,926 (49.5%) of BIN sequences and 1310/2199 (59.97%) of the OTU sequences displayed strong evidence of an indel error as they contained a stop codon when translated. By contrast, following denoising with debar the incidence of stop codons was far lower as just 2858/122,349 (2.3%) of the BIN sequences and 418/2,145 (19.49%) of the OTU sequences had evidence of indels. This result indicated that denoising of individual sequences reduced the incidence of apparent indel errors by over 95% for the BINs (58,593
fewer indel errors) and by 68% for the OTUs (892 fewer indel errors). Most sequences were subjected to at least one indel correction by debar, with 85,298/122,349 (69.7%) of the final BIN sequences and 1387/2145 (64.7%) of final OTU sequences containing at least one "N" character. Low abundance OTUs in the data set represented by biologically valid sequences need not be discarded solely due to their low abundance and could be further inspected for putative evidence of rare community members.

4 | DISCUSSION

This manuscript introduces debar, a PHMM-based denoiser, and demonstrates how it can improve the quality of sequence data used for both DNA barcode library construction and for metabarcoding studies by correcting indels introduced by sequencing error. We first evaluated its effectiveness through an in silico study that tested its capacity to recognize and repair reference barcodes with artificially introduced indels. Debar was shown to be effective, as it corrected >95.7% of the errors and applied erroneous adjustments to less than 0.01% of correct sequences. This strong performance extended to real-world data sets. Debar reduced the rate of frameshift indels by 75% in sequence records generated by the long-read Sequel platform, generating more barcode-compliant sequences, most with little or no missing information.

Debar also improved the quality of metabarcoding data generated by the ION S5 allowing for ESVs to be considered with higher confidence and for the recovery of higher-quality representative sequences for OTUs.

Denoising sequences with artificial errors and known ground truths showed that the corrections performed by debar were imperfect, with the exact indel location being identified only 61.8% of the time. The application of a default 7 bp censorship on both sides of putative indel corrections proved to be an effective means of masking most errors, improving the denoiser’s error removal rate to >95.75%. This high error removal rate involves a tradeoff, as sequence adjustments are accompanied with a loss of 14 base pairs of information. This information loss is an acceptable cost, as it ensures that all remaining base pairs can be considered with high confidence. The nature of high-throughput sequence data, namely that there are usually multiple sequencing reads for a given specimen available, can help mitigate the loss of information. Corrected sequences from a specimen or OTU can be used in conjunction with one another, filling in the different censored locations and overcoming the loss of information. The censorship of bases adjacent to indel corrections is an optional parameter that users may alter to suit their needs. Smaller censorship values, or no censorship at all, would result in less loss of information per sequence, but would come at the cost of more errors remaining in the final data.

Denoising of real DNA barcode data obtained from sequencing of specimens on the Pacific Biosciences Sequel platform resulted in higher-quality output sequences. An exact metric quantifying the improvement is, however, difficult to state with certainty, as the ground truth of the sequences is not known. The independent tests of the sequences through submission of consensus sequences to BOLD before and after denoising provided a conservative estimate of the debar package’s effectiveness. Conservatively, this test showed a 75% reduction in the number of barcode sequences with technical indel errors after application of the debar pipeline and a low false negative rate (147 unidentified errors out of 1,515 total putative errors). This is an important improvement because the Pacific Biosciences Sequel platform is used at the Centre for Biodiversity...
Genomics to produce high-quality reference barcodes for the barcoding research community (Hebert et al., 2018). Accuracy of these sequences is therefore important; the debar package is shown to improve sequence quality, yielding more biologically likely and therefore reliable outputs. The generation of barcode sequences is also made more efficient. By increasing the rate of barcode-compliant outputs from 93.7% to 98%, fewer samples require reprocessing or resequencing.

Understanding within-species patterns of genetic diversity is an essential metric for characterizing community health. High intra-species genetic diversity is assumed to indicate healthy ecosystems, comprised of large and stable populations with the standing genetic variation needed to survive environmental stressors (Zizka et al., 2020). The characterization of ESVs within OTUs can provide intra-species diversity measures for member species of a community (Frøslev et al., 2017). The initial check of the sub-OTU sequence data from the mock community sequenced with IonTorrent revealed a high rate of putative indel errors (54% of sequences), which would lead to a gross over estimation of the number of ESVs within the OTUs. The reduction of the error rate
after denoising with debar allows for a more accurate examination of intra-OTU ESVs and therefore allows for more accurate assessments of intraspecies diversity and community health, despite the fact that debar is not capable of eliminating nonindel errors from sequences. Even with the improvements to ESV quality by debar, intraspecies diversity estimates will probably remain inflated to some extent, as the sequence-by-sequence corrections applied by debar exclusively account for indel errors while substitution errors could persist within the data.

We have demonstrated that debar is an effective means of reducing technical errors in DNA barcode and metabarcode data, but the package is not without limitations. The package is designed to correct insertion and deletion errors, but these are not the only technical issues that can lead to inflated biodiversity estimates. The program is not an effective means of identifying or correcting chimeric sequences or nonanimal COI biological contaminants and should exist within an input data set they are likely to go uncorrected. Additionally, debar does not have the ability to correct substitution errors on a sequence-by-sequence basis. Because of indel correction, denoised sequences are aligned, and nucleotide positions become directly comparable across different sequences from a given specimen or OTU. Random point substitution errors can thereby be corrected in consensus sequence generation, through the “majority rule” approach debar uses in base calling. However, if systematic errors exist (i.e., most sequences possess the same substitution), few sequences are available for consensus sequence generation, or ESVs are being examined, then substitution errors may persist in the data. An additional source of error unaccounted for by debar is contaminant sequences. It has been demonstrated previously that the PHMM utilized in debar is not an effective means of separating animal barcode sequences from off-target barcodes derived from bacteria, plant, fungi, or other origins (Nugent et al., 2020). Taken together, these limitations show that debar cannot single handedly address the technical challenges associated with DNA barcoding. The tool is probably most effective when applied in conjunction with existing barcode and metabarcode workflows and improves the quality of final sequences if the inputs have been filtered based on quality, had primers removed, and been cleaned of chimeric and contaminant sequences. The sequence-by-sequence denoising approach of debar means that it is a flexible tool capable of integrating into analysis pipelines for sequencing data from various sources. Application of debar in tandem with conventional, clustering-based denoising tools would probably lead to the highest quality assessment of biodiversity. Following OTU generation with other tools, using debar to denoise all reads within a given OTU prior to consensus sequence generation would maximize accuracy of the consensus sequence while conforming to the conserved structure of the COI barcode region. The removal of intra-OTU noise can also improve the accuracy of alpha-diversity estimates. Additionally, application of debar in the denoising of rare, low-abundance sequences not present in the OTUs would allow these data to be further examined with higher confidence, revealing biological insights that would be overlooked in conventional workflows.

The PHMM denoising technique used by debar is an effective barcode-focused framework that can be extended to fit a variety of needs. Data from only two sequencing platforms were tested in this study: the Pacific Biosciences Sequel and Thermo IonTorrent S5. Since the PHMM used in debar is barcode specific and not sequence specific, debar can be effectively applied in denoising of barcode data obtained from any sequencing platform. However, the effectiveness of the denoiser will depend on the types and rates of technical errors associated with a given platform. When applied to data from sequencers such as the Illumina MiSeq, the rate of technical errors corrected by debar will be lower, as this platform is more prone to introduction of substitution, as opposed to indel errors (Schirmer et al., 2015). Although the debar package contains a PHMM for only the common animal barcode COI, the denoising algorithm can in the future be extended and applied in the correction of data for other DNA barcodes with conserved structures.

5 | CONCLUSION

This study has described debar, an R package for denoising DNA barcode data, and demonstrated its ability to correct indels in both barcode and metabarcode sequences due to instrument error. In each data set, debar improved sequence quality. It reduced the apparent number of indels by 75% in data generated by Sequel, increasing the proportion of sequences that met the quality standards required to qualify as a reference barcode. The merits of debar for metabarcode analysis were twofold, allowing more likely consensus sequences to be obtained for OTUs, and for intra-OTU variation to be quantified with higher confidence. Overall, debar is a robust utility for identifying deviations from the highly conserved protein-coding sequence of the COI barcode region. Corrections informed by its use improve the separation of true biological variation from technical noise, with low frequencies of false corrections. Integration of debar into the workflows for processing barcode and metabarcode data will allow biological variation to be characterized with higher accuracy.

ACKNOWLEDGEMENTS

This research was supported by grants from Genome Canada through Ontario Genomics and from the Ontario Ministry of Economic Development, Job Creation and Trade. The funders played no role in study design or decision to publish. This research was enabled in part by resources provided by Compute Canada (www.computecanada.ca). We thank Tony Kuo and Thomas Braukmann for aid with data acquisition and interpretation and Tony for helpful comments on the manuscript.

AUTHOR CONTRIBUTIONS

The study was conceived and designed by Sarah J. Adamowicz, Paul D. N. Hebert, Sujeevan Ratnasingham, and Cameron M. Nugent. The programming of the debar package was performed by Cameron M. Nugent. Analyses of package performance were performed by...
Cameron M. Nugent with resources, design, and other assistance provided by Tyler A. Elliott, Sujeewan Ratnasingham, and Sarah J. Adamowicz. The initial draft of the manuscript was written by Cameron M. Nugent and Sarah J. Adamowicz. All authors contributed to the editing of the manuscript.

DATA AVAILABILITY STATEMENT
DNA barcode sequences used in training of the Profile Hidden Markov Models are available in the Supplementary data of the following paper: https://doi.org/10.1139/gen-2019-0206. DNA barcode sequences used in model testing are available in this manuscript’s Supporting Information files. The R source code for the debar package is available on GitHub: https://github.com/CNuge/debar. Additional data and code available on request from the authors.

ORCID
Cameron M. Nugent https://orcid.org/0000-0002-1135-2605

REFERENCES
Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Xu, Z. Z., Kightley, E. P., Thompson, L. R., Hyde, E. R., Gonzalez, A., & Knight, R. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems, 2(2), e00191–16.

Baynhm-Herd, Z., Amano, T., Sutherland, W. J., & Donald, P. F. (2018). Governance explains variation in national responses to the biodiversity crisis. Environmental Conservation, 45(4), 407–418.

Braukmann, T. W. A., Ivanova, N. V., Prosser, S. W. J., Elbrecht, V., Steinke, D., Ratnasingham, S., de Waard, J. R., Sones, J. E., Zakharov, E. V., & Hebert, P. D. N. (2019). Metabarcoding a diverse arthropod mock community. Molecular Ecology Resources, 19(3), 711–727.

Brown, E. A., Chain, F. J., Chiu, Z., Hebert, P. D. N., & Crisescu, M. E. (2016). Early detection of aquatic invaders using metabarcoding reveals a high number of non-indigenous species in Canadian ports. Diversity and Distributions, 22(10), 1045–1059.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. Nature Methods, 13(7), 581.

Clare, E. L., Chain, F. J., Littlefair, J. E., & Crisescu, M. E. (2016). The effects of parameter choice on defining molecular operational taxonomic units and resulting ecological analyses of metabarcoding data. Genome, 59(11), 981–990.

Corder, T., Lanzén, A., Apotheloz-Perrat-Gentil, L., Stoeck, T., & Pawlowski, J. (2019). Embracing environmental genomics and machine learning for routine biomonitoring. Trends in Microbiology, 27(5), 387–397.

Crisescu, M. E. (2014). From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. Trends in Ecology & Evolution, 29(10), 566–571.

Delabye, S., Rougerie, R., Bayendi, S., Andeime-Eyene, M., Zakharov, E. V., deWaard, J. R., Hebert, P. D. N., Kamgang, R., Le Gall, P., Lopez-Vaamonde, C., Mavoungou, J.-F., Moussavou, G., Moulin, N., Osilidy, R., Rahola, N., Sebag, D., & Decaëns, T. (2019). Characterization and comparison of poorly known moth communities through DNA barcoding in two Afrotropical environments in Gabon. Genome, 62(3), 96–107.

Driscoll, D. A., Bland, L. M., Bryan, B. A., Newsome, T. M., Nicholson, E., Ritchie, E. G., & Doherty, T. S. (2018). A biodiversity-crisis hierarchy to evaluate and refine conservation indicators. Nature Ecology & Evolution, 2(5), 775–781.

D’Souza, M. L., & Hebert, P. D. (2018). Stable baselines of temporal turnover underlie high beta diversity in tropical arthropod communities. Molecular ecology, 27(10), 2447–60. https://doi.org/10.1111/mec.14693.

Durbin, R., Eddy, S. R., Krogh, A., & Mitchison, G. (1998). Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press.

Eddy, S. R. (1998). Profile hidden Markov models. Bioinformatics, 14(9), 755–763.

Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. Genome Informatics 2009: Genome Informatics Series, 23, 205–211.

Edgar, R. C. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. BioRxiv, 081257.

Elbrecht, V., Vamos, E. E., Steinke, D., & Leese, F. (2018). Estimating intraspecific genetic diversity from community DNA Metabarcoding Data. PeerJ, 6, e4644.

Folmer, O., Black, M., Hoeh, W., Lutz, R., & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Molecular Marine Biology and Biotechnology, 3(5), 294–299.

Frosløv, T. G., Kjeller, R., Bruun, H. H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., & Hansen, A. J. (2017). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. Nature Communications, 8(1), 1–11.

Hajibabaei, M., Baird, D. J., Fahner, N. A., Beiko, R., & Golding, G. B. (2016). A new way to contemplate Darwin’s tangled bank: how DNA barcodes are reconnecting biodiversity science and biomonitoring. Philosophical Transactions of the Royal Society B: Biological Sciences, 371(1702), 20150330.

Hajibabaei, M., Spall, J. L., Shokralla, S., & van Konyenburg, S. (2012). Assessing biodiversity of a freshwater benthic macroinvertebrate community through non-destructive environmental barcoding of DNA from preservative ethanol. BMC Ecology, 12(1), 28.

Hardge, K., Neuhaus, S., Kilias, E. S., Wolf, C., Metfies, K., & Frickenhous, S. (2018). Impact of sequence processing and taxonomic classification approaches on eukaryotic community structure from environmental samples with emphasis on diatoms. Molecular ecology resources, 18(2), 204–216. https://doi.org/10.1111/1755-0998.12726.

Hebert, P. D. N., Braukmann, T. W. A., Prosser, S. W. J., Ratnasingham, S., deWaard, J. R., Ivanova, N. V., Janzen, D. H., Hallwachs, W., Naik, S., Sones, J. E., & Zakharov, E. V. (2018). A Sequel to Sanger: Amplicon sequencing that scales. BMC Genomics, 19(1), 219.

Hebert, P. D. N., Cywinska, A., Ball, S. L., & DeWard, J. R. (2003). Biological identifications through DNA barcodes. Proceedings of the Royal Society of London. Series B: Biological Sciences, 270(1512), 313–321.

Hebert, P. D., Penton, E. H., Burns, J. M., Janzen, D. H., & Hallwachs, W. (2004). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. Proceedings of the National Academy of Sciences, 101(41), 14812–7. https://doi.org/10.1073/pnas.0406166101.

Hebert, P. D. N., Ratnasingham, S., Zakharov, E. V., Telfer, A. C., Lesveques-Beaudin, V., Milton, M. A., & DeWard, J. R. (2016). Counting animal species with DNA barcodes: Canadian insects. Philosophical Transactions of the Royal Society B: Biological Sciences, 371(1702), 20150333.

Hubert, N., & Hanner, R. (2015). DNA barcoding, species delineation and taxonomy: a historical perspective. DNA Barcodes, 3(1), 44–58.

Kaustio, K. M., Roslin, T., Säkijärvi, I. E., & Vesterinen, E. J. (2017). Pellets of proof: First glimpse of the dietary composition of adult odonates as revealed by metabarcoding of feces. Ecology and Evolution, 7(20), 8588–8598.

Kumar, V., Vollbrecht, T., Chernyshev, M., Mohan, S., Hanst, B., Bavafa, N., Lorenzo, A., Kumar, N., Ketteringham, R., Eren, K., Golden, M.,
Oliveira, M. F., & Murrell, B. (2019). Long-read amplicon denoising. *Nucleic Acids Research*, 47(18), e104–e104.

Lopez-Vaamonde, C., Sire, L., Rasmussen, B., Rougerie, R., Wieser, C., Alloufi, A. A., Minet, J., deWaard, J. R., Decaëns, T., & Lees, D. C. (2019). DNA barcodes reveal deeply neglected diversity and numerous invasions of micromoths in Madagascar. *Genome*, 62(3), 108–121.

Nearing, J. T., Douglas, G. M., Comeau, A. M., & Langille, M. G. (2018). Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ*, 6, e5364.

Nugent, C. M., Elliott, T. A., Ratnasingham, S., & Adamowicz, S. J. (2020). CoIL: an R package for cytochrome C oxidase I (COI) DNA barcode data cleaning, translation, and error evaluation. *Genome*, 63(6), 291–305.

Ratnasingham, S., & Hebert, P. D. N. (2013). A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. *PLoS One*, 8(7), e66213.

Rosen, G., Garbarine, E., Caseiro, D., Políkar, R., & Sokhansanj, B. (2008). Metagenome fragment classification using k-mer frequency profiles. *Advances in Bioinformatics*, 2008, 1–12. https://doi.org/10.1155/2008/205969.

Schirmer, M., Ijaz, U. Z., D’Amore, R., Hall, N., Sloan, W. T., & Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research*, 43(6), e37.

Shi, Z. Y., Yang, C. Q., Hao, M. D., Wang, X. Y., Ward, R. D., & Zhang, A. B. (2018). FuzzyID2: A software package for large dataset species identification via barcoding and metabarcoding using Hidden Markov Models and fuzzy set methods. *Molecular Ecology Resources*, 18(3), 666–675.

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Arrieta, J. M., & Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored “rare biosphere”. *Proceedings of the National Academy of Sciences USA*, 103(32), 12115–12120.

Stat, M., Huggett, M. J., Bernasconi, R., DiBattista, J. D., Berry, T. E., Newman, S. J., Harvey, E. S., & Bunce, M. (2017). Ecosystem biomonitoring with eDNA: metabarcoding across the tree of life in a tropical marine environment. *Scientific Reports*, 7(1), 1–11.

Taberlet, P., Coissac, E., Hajibabaei, M., & Rieseberg, L. H. (2012). Environmental DNA. *Molecular Ecology*, 21(8), 1789–1793.

Wilkinson, S. P. (2018). kmer: an R package for fast alignment-free clustering of biological sequences. R package version 1.0.0. https://cran.r-project.org/package=kmer.

Wilkinson, S. P. (2019). aphid: an R package for analysis with profile hidden Markov models. *Bioinformatics*, 35(19), 3829–3830.

Wilson, J. J., Brandon-Mong, G. J., Gan, H. M., & Sing, K. W. (2019). High-throughput terrestrial biodiversity assessments: mitochondrial metabarcoding, metagenomics or metatranscriptomics? *Mitochondrial DNA Part A*, 30(1), 60–67.

Wirta, H. K., Hebert, P. D. N., Kaartinen, R., Prosser, S. W., Várkonyi, G., & Roslin, T. (2014). Complementary molecular information changes our perception of food web structure. *Proceedings of the National Academy of Sciences*, 111(5), 1885–1890.

Zhang, X. M., Shi, Z. Y., Zhang, S. Q., Zhang, P., Wilson, J. J., Shih, C. K., Li, J., Li, X. D., Yu, G. Y., & Zhang, A. B. (2021). Plant-herbivorous insect networks: who is eating what revealed by long-barcodes using high-throughput sequencing and Trinity assembly. *Insect Science*, 28(1), 127–143.

Zizka, V. M., Weiss, M., & Leese, F. (2020). Can metabarcoding resolve intraspecific genetic diversity changes to environmental stressors? A test case using river macrozoobenthos. *BioRxiv*. https://doi.org/10.1101/2020.03.08.982561.

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section.