Computational Evaluation of DNA Metabarcoding for Universal Diagnostics of Invasive Insect Pests

Alexander M. Piper^{1,2*}, Noel O.I. Cogan^{1,2}, John Paul Cunningham^{1,2}, Mark J. Blacket¹

¹Agriculture Victoria Research, AgriBio, Centre for AgriBioscience, 5 Ring Road, Bundoora, VIC, 3038, Australia

²School of Applied Systems Biology, La Trobe University, Bundoora, VIC, 3086, Australia

*Correspondence should be addressed to A.M.P. (alexander.piper@agriculture.vic.gov.au)

Appropriate design and selection of PCR primers plays a critical role in determining the sensitivity and specificity of a metabarcoding assay. Despite several studies applying metabarcoding to insect pest surveillance, the diagnostic performance of the short “mini-barcodes” required by high-throughput sequencing platforms has not been established across the broader taxonomic diversity of invasive insects. We address this by computationally evaluating the diagnostic sensitivity and predicted amplification bias for 68 published and novel cytochrome c oxidase subunit 1 (COI) primers on a curated database of 110,676 insect species, including 2,625 registered on global invasive species lists. We find that mini-barcodes between 125-257 bp can provide comparable resolution to the full-length barcode for both invasive insect pests and the broader Insecta, conditional upon the subregion of COI targeted and the genetic similarity threshold used to identify species. Taxa that could not be identified by any barcode lengths were phylogenetically clustered within ‘problem groups’, many arising through taxonomic inconsistencies rather than insufficient diagnostic information within the barcode itself. Substantial variation in predicted PCR bias was seen across published primers, with those including 4-5 degenerate nucleotide bases showing almost no mismatch to major insect orders. While not completely universal, a single COI mini-barcode can successfully differentiate the majority of pest and non-pest insects from their congeners, even at the small amplicon size imposed by 2 × 150 bp sequencing. We provide a ranked summary of high-performing primers and discuss the bioinformatic steps required to curate reliable reference databases for metabarcoding studies.

Introduction

Early detection and rapid response are crucial for preventing the establishment and spread of invasive pests and pathogens (Liebhold et al., 2016; Reaser, Burgiel, et al., 2020). Historically, invasive species surveillance has relied upon targeted inspections for predefined lists of regulated taxa (Reaser, Frey, et al., 2020; Schrader & Unger, 2003). However, as global trade networks become increasingly interlinked and anthropogenic climate change alters species range distributions, this list-based framework often lags behind the speed at which new pests can emerge and spread across borders (Bebber, 2015; Hulme, 2009). This lag becomes particularly apparent when considering impacts beyond agroecosystems, where the size and complexity of the natural environment presents challenges for accurate risk prediction (Caley et al., 2006; Crooks, 2005). In light of this, it is becoming increasingly appreciated that modern biosecurity will need to adopt a more comprehensive approach to surveillance that aims to detect and evaluate all newly introduced species, not just those regulated by national quarantine agencies (Meyerson & Reaser, 2002; Reaser et al., 2008; Simberloff, 2006). In practice, however, adoption of this framework is bottlenecked by a lack of diagnostics capacity to sort and identify the large number of specimens collected by intensive surveillance efforts (Bishop & Hutchings, 2011; Piper et al., 2019).

Plant pest and pathogen diagnostics currently rely on a mixture of morphological examination, biochemical techniques, and molecular assays such as diagnostic qPCR, and DNA barcoding (EPPO, 2019a). While these methods provide highly accurate identification for small numbers of specimens (Armstrong & Ball, 2005; Darling & Blum, 2007), their inherent restriction to analysing single specimens per reaction limits their application to large mixed samples

collected in surveillance traps (Batovska et al., 2020; Carnegie & Nahrung, 2019). As an alternative, high-throughput sequencing (HTS) platforms can comprehensively characterise mixed populations of genomic DNA (metagenomics), RNA (metatranscriptomics) or taxonomically informative marker genes (metabarcoding), allowing whole communities to be identified without any prior isolation or specimen sorting (Piper et al., 2019; Tedersoo et al., 2019). While first emerging for exploring biodiversity (Handelsman, 2004; Taberlet et al., 2012), non-targeted HTS assays have recently been co-opted by various disciplines of molecular diagnostics, where their potential to act as a universal identification tool was quickly recognised (Adams et al., 2009; Comtet et al., 2015). By removing the requirement to separately develop and maintain hundreds of targeted diagnostic assays, universal HTS diagnostics could substantially expand the range of organisms within the scope of a diagnostic laboratory, as well as decrease the costs of implementation (Adams et al., 2018; Allcock et al., 2017).

Metabarcoding of the mitochondrial cytochrome c oxidase subunit 1 (COI) gene presents the most readily adoptable HTS approach for diagnostics of insect pests, due to its cost effectiveness, extensive public reference sequence databases, and ability to leverage widespread acceptance of DNA barcoding within regulatory frameworks (Andújar et al., 2018; Comtet et al., 2015; Piper et al., 2019). While conventional single-specimen DNA barcoding targets a 709 bp region of COI (Folmer et al., 1994), modern HTS platforms impose strict length limitations on sequenced molecules, and therefore “mini-barcodes” must instead be used (Brandon-Mong et al., 2015). The number of diagnostic nucleotides contained within these mini-barcodes largely determines the sensitivity and specificity for single specimens, however, mismatch between PCR primers and variable

template molecules can bias amplification and cause drop-outs of low-abundance taxa within mixed community samples (Elbrecht & Leese, 2015; D. W. Yu et al., 2012). Primer-template mismatch is a particular issue for protein-coding genes such as COI where variability in the third position of each codon leaves no strictly conserved regions for placement of universal PCR primers (Deagle et al., 2014). Therefore, degenerate nucleotide bases are commonly incorporated into COI metabarcoding primers to account for this inevitable mismatch (Elbrecht & Leese, 2017b), though overuse can result in undesired amplification of non-target organisms (Collins et al., 2019; Leese et al., 2021).

Historically, molecular diagnostic assays would have undergone stringent laboratory validation in order to resolve the aforementioned issues and establish performance parameters for every target designated in an unambiguously defined scope (EPP0, 2019b). However, when considering the sheer number of potential targets, hosts and matrices that would need to be evaluated for a universal assay, it is evident that many validation processes cannot be applied in their traditional sense and must be adapted to novel HTS based diagnostics (Maree et al., 2018; Roenhorst et al., 2018). In-silico methods pose a promising alternative that can leverage public reference sequence data to establish the diagnostic performance of a target barcode region and determine the best placement of degenerate PCR primers, all without requiring physical specimens (Elbrecht & Leese, 2017a; Ficetola et al., 2010). Nevertheless, the use of public reference data comes with some caveats, as issues of mislabelled taxonomic annotations, insufficiently identified specimens, and contamination are well documented (Garg et al., 2019; Locatelli et al., 2020; Pentinsaari et al., 2020; Siddall et al., 2009). It is therefore essential for public DNA barcode sequences to be appropriately curated before use for in-silico validation procedures or as reference databases for metabarcoding analysis (Piper et al., 2019).

In this article we develop a computational workflow for curating a large collection of Insect COI sequences from public sequence repositories. Using this curated database, in-silico methods are then used to evaluate the sensitivity, specificity, and predicted amplification bias for 68 published and novel metabarcoding primers on a globally relevant list of invasive insect pests and the wider insect diversity. We identify optimal subregions of the COI barcode for species differentiation and determine the amplicon lengths required to achieve comparable resolution to the full-length barcode. This study informs and offers recommendations for selection of metabarcoding primers and provides a robust workflow for assembling curated reference databases for DNA barcode-based insect identification.

Methods

Retrieval and curation of public reference data

COI records and mitochondrial genomes with the taxonomic annotation 'Insecta' were retrieved from NCBI GenBank and the Barcode of Life Data system (BOLD) (Ratnasingham & Hebert, 2007) using the *Rentrez* (Winter, David & Winter, 2017) and *bold* (Chamberlain, 2017) R packages. All retrieved sequences then went through a series of curation steps (Figure 1A). First, to resolve taxonomic synonyms between the two repositories,

sequence annotations were mapped into the Open Tree of life Taxonomy (OTT) (Hinchliff et al., 2015), and only those with complete binomial names and not flagged with uncertain taxonomic placement were retained (see supplementary notes 2 and 3 for relevant flags). All sequences were then aligned to a reference Profile Hidden Markov Model (PHMM) (Eddy, 1998) of the COI locus generated from a manually curated version of the Midori-longest V237 dataset (Machida et al., 2017) using the *aphid* R package (Wilkinson, 2019). All sequences that met a minimum odds-ratio alignment score of 100 without containing stop-codons or frameshift mutations were retained, and bases outside the bounds of the LCO1490-HCO2198 (Folmer et al., 1994) primer binding sites trimmed from the alignment. To identify putatively misannotated sequences, the alignment was hierarchically clustered using the *kmer* R package (Wilkinson, 2018) and sequences removed if their species annotation at 99% identity, genus annotation at 97% identity, or family annotation at 95% identity disagreed with more than 80% of other sequences within its respective cluster. A nucleotide BLAST search (Altschul et al., 1990) was then conducted against a local contaminants database and sequences with percentage coverage and identity >79% for *Wolbachia*, >98% for known pseudogene sequences, or >96% for human mitochondria were removed. To identify invasive insect species within the curated sequence database, taxonomic names were retrieved from 11 global and geographically focused pest and invasive species lists (Supplementary Note 1), hereafter referred to as the 'pest' dataset. All sequences <200 bp were removed, then the number of barcodes per pest taxon was compared to the remainder of the insect taxa using the Welch t-test. Finally, in order to accelerate downstream computations and minimize the effects of taxonomic sampling bias (Mutanen et al., 2016), the database was pruned to a maximum 5 representative sequences per species, discarding sequences sequentially from smallest to largest.

Construction of phylogenetic tree

The curated Insect reference sequences were supplemented with an outgroup of 15 COI sequences from Diplura, the sister taxa to Insecta (Misof et al., 2014), and all positions in the alignment that contained gaps in >95% of sequences were masked. A maximum likelihood phylogenetic tree was generated using FastTree (Price et al., 2009) following the General Time-Reversible (GTR) model (Tavaré, 1986) and gamma model of rate heterogeneity across sites, with taxonomic identities at the domain, phylum, class and order ranks used to constrain the deeper topology of the tree. The constructed phylogeny was rooted

Computational evaluation of insect pest metabarcoding

Figure 1: A) Overview of the computational pipeline used to curate public reference sequence data for primer evaluation. **B)** Number of sequences (upper) and species (lower) retained after each curation step, and their origin from BOLD, GenBank or duplicated across both repositories.

on the edge connecting the Diplura outgroup to the rest of the tree, and made ultrametric using the *geiger* R package (Pennell et al., 2014) and *PATHd8* (Britton et al., 2007). All phylogenetic trees were plotted using the *ggtree* (G. Yu et al., 2017, 2018) and *ggplot2* R packages (Wickham, 2016).

Diagnostic information within the COI barcode

The curated reference sequence database was split by taxonomic order, and Shannon's entropy (Scheider & Stephens, 1990) calculated for each alignment position, then visualised with structural motifs annotated as per Pentinsaari et al. (2016). To identify the most diagnostically informative subregions of the COI barcode, a sliding window approach was used to split the alignment into virtual amplicon sequences of 200 bp, 300 bp and 400 bp, in intervals of 3 bp. The sequences within each virtual amplicon window were clustered at 97% similarity using *UCLUST* (Edgar, 2010), and a species considered successfully identified if there were no other sequences with conflicting taxonomic annotations within its respective cluster.

Characteristics of published and novel primers

32 forward and 31 reverse primers (Figure 2B) overlapping the COI barcode region were identified from a literature search for the terms 'metabarcoding mini-barcode', and 'metabarcoding primer' and supplemented with 2 new forward and 3 new reverse primers designed using Primer3 (Untergasser et al., 2012). The number of reference sequences for which each primer had an appropriate binding site was determined via string matching using the *Biostrings* R package (Pagès et al., 2019), allowing for a hamming distance of 2. The frequency of each nucleotide base, as well as presence of homopolymers or GC clamps (2 or more contiguous G or C bases at the 5' end) were determined using the *Biostrings* and *DECIPHER* R packages (Wright, 2016). Primer melting temperatures were calculated using nearest neighbour thermodynamics (SantaLucia & Hicks, 2004) with the *TmCalculator* package (Li, 2019).

Diagnostic sensitivity of mini-barcodes

PCR amplification was simulated by truncating the reference sequences to the region amplified by each primer set as originally published, as well as the 1156 unique combinations of forward and reverse primers that could produce an amplicon >50 bp. To determine how the percentage identity threshold used to assign species impacted identification success, virtual amplicons were clustered at the commonly used identification threshold of 97% (Alberdi et al., 2018), as well as more stringent 98%, 99%, and 100% thresholds using *UCLUST* (Edgar, 2010). Again, each species was considered successfully identified if there were no sequences with conflicting taxonomic annotations within its respective cluster. As the relationship between identification success and amplicon length was not linear, a segmented regression model was fit separately to each distance interval using the *chngpt* R package (Fong et al., 2017). The change-point between the two regression segments, which can be considered the minimum amplicon length at which identification success using COI mini-barcodes becomes congruent with the full-length barcode region, was determined via maximum likelihood with confidence intervals obtained from 1000 bootstrap replicates (Fong, 2019). Primer specific identification performance was obtained by averaging the residuals from the regression model for all evaluated combinations that contained that respective primer. In order to determine how the taxa that could not be identified by any barcode length were distributed across the insect phylogeny, the consenTRAIT metric (Martiny et al., 2013) was calculated for the full-length barcode at the 97% identification threshold, with each clade weighted by the number of failed identifications. This metric measures the mean phylogenetic depth of clades for which a binary trait, in this case failed identification, is present in at least 50% of its tips, with significance assessed against 1000 permutations. The phylogenetic clades with the highest number of identification failures

Computational evaluation of insect pest metabarcoding

Figure 3: Identification success for all insects (upper) and Insect pests (lower) as a function of amplicon length for all possible combination of mini-barcode primers. Percentage identity threshold used to identify species increases from 97% (leftmost panel) to 100% (rightmost panel), with segmented regression model predictions overlaid. Each primer combination is coloured by its residual error from regression model predictions, with higher residuals indicating better than expected performance for its length.

scores from each metric were then weighted by their relative importance for overall primer performance; 1x for mismatch, 0.5x for pest insect identification, 0.5x for all insect identification, and 0.25x for each of the measured physical characteristics; fold-degeneracy, primer length, melting temperature, GC%, presence of GC clamps and longest homopolymer, then summed by primer to obtain a final ranking.

Results

Sequence database assembly

To assemble the reference database for primer evaluation, 4,491,128 COI sequences with taxonomy “Insecta” were retrieved from GenBank and BOLD, including 23,571 extracted from mitochondrial genomes. Of these sequences, 1,584,589 were exclusive to GenBank, 15,153 were exclusive to BOLD, and 2,891,386 shared across both repositories (Figure 1B). 2,745,595 sequences and 129,225 species could be mapped to valid binomial names within the OTT taxonomy, including 11,431 taxonomic synonyms resolved to currently accepted species names in the process. This step resulted in the largest reduction of both unique sequences and species (Figure 1B), with most unsuccessfully mapped sequences being flagged as *incertae sedis* (1,329,337 sequences) or not present in the taxonomy at all (346,299 sequences). Additional reasons for removal at this stage included infraspecific taxa (22,501 sequences), the taxon being extinct (4,493 sequences) or other more minor issues (supplementary Figure 1). In contrast, many of the later curation steps only marginally reduced the number of both sequences and species (Figure 1B), with homology and pseudogene filters removing 333,765, and 16,553 sequences respectively, and comparisons of taxonomy between highly similar sequences removing 3,130 putatively misannotated sequences. A BLAST search against a local

database of contaminants removed 30 *Wolbachia* and 56 human mitochondrial sequences and identified a further 1,667 sequences that were >98% similar to known COI pseudogenes but did not contain any characteristic stop codon or frameshift mutations. All sequences <200 bp in length were then removed, leaving 2,389,404 sequences from 110,676 distinct species remaining. When compared to globally relevant lists of invasive insects, a total of 2,625 species spanning 1,490 genera and 20 taxonomic orders were identified within the curated database (supplementary Figure 2). Each pest species was represented by a significantly higher number of sequences (mean 79.9 ± 5.60) than the average insect taxon (mean 20.2 ± 0.428) (Welch t-test: $t_{(2837)} = 11.1$, $p < .001$), however no reference data was available for 1,717 of the listed species. As the number of sequences per species was greatly skewed by a few highly sampled species (supplementary Figure 3), the database was pruned to a maximum of 5 sequences per taxa. This left a total of 315,754 sequences from 110,676 species remaining in the final curated reference database (Figure 1B).

Optimal diagnostic subregions within the COI barcode

The final curated reference database consisted of a 712 bp alignment (Figure 2), which included the 709 bp barcode region and a 3bp insertion at position 110-112 that occurred in the order Thysanoptera as well as some Hymenopteran species. Sequence coverage was relatively even across the COI barcode, with exception of the terminal regions where the standard practice of removing priming sequences before submission to public repositories resulted in low coverage (Figure 2D). The per-site Shannon entropy within regions coding for loop structures was significantly higher than within those coding for

Computational evaluation of insect pest metabarcoding

Figure 4: **A)** Phylogenetic tree of all insect genera contained within the curated reference database with major orders annotated. Clades are highlighted in red where $\geq 50\%$ of species could not be successfully differentiated from their congeners by the full-length barcode at the 97% identity threshold. **B)** Phylogenetic depth of clades where $\geq 50\%$ of specimens could not be identified, with those containing the highest number of failed identifications annotated. **C)** Number of failed identifications per taxonomic order with the full-length barcode region at the 97% identity threshold.

transmembrane helices (Welch t-test: $t_{(12306)} = 4.0459$, $p < .001$), however short segments of low entropy were distributed throughout both (Figure 2A). The large majority of the 58 primers retrieved from the literature were designed around these few low-entropic segments, with many overlapping around 180 bp, 255 bp, 370 bp, and 600 bp into the alignment (Figure 2B). As none of these segments of low entropy extended for the 18-24 bp length of a typical primer, many of the evaluated primers included multiple degenerate bases to account for variable positions (Supplementary Table 1). Sliding window analyses revealed that for a 200 bp or 300 bp amplicon, those positioned towards the 3' end of the barcode region could differentiate substantially more species than those towards the 5' end, however differences were much less pronounced for a 400 bp amplicon (Figure 2C).

Diagnostic sensitivity of mini-barcodes

PCR amplification was simulated for the 1156 primer combinations that could produce an amplicon > 50 bp, as well as the full-length barcode region. Identification success using mini-barcodes generally increased with amplicon length, but did not follow a simple linear relationship, and instead saw a sharp initial increase up to a certain length, followed by a second more gradual slope (Figure 3). A segmented regression model applied to the whole insect dataset inferred the change point between these trends to be 257 bp at the 97% identity threshold (95% CI: 242 bp –

266 bp), 237 bp at the 98% threshold (95% CI: 218 bp – 248 bp), 146 bp at the 99% threshold (95% CI: 137 bp – 206 bp) and 125 bp when only 100% matches were considered (95% CI: 110 bp – 131 bp). On the other hand, the inferred changepoints for the pest dataset were approximately 20 bp smaller at the 97%, 98% and 99% identity thresholds, but identical to the larger dataset at the 100% threshold. In spite of this general trend, certain amplicons deviated up to 10% above or below the regression line for both datasets (Figure 3), reinforcing that appropriate placement of mini-barcodes within the COI barcode region can be just as important as amplicon length for diagnostic performance.

The overall proportion of Insect species that could be successfully identified by any of the barcodes increased with the stringency of the percentage identity threshold used (Figure 3). For the complete Insect dataset, the proportion identified increased from 76.6% when using a 97% identification threshold up to 91.5% when only 100% matches were considered, and similarly increased from 77.6% to 89.4% in the pest dataset. Taxa that could not be identified even by the full-length barcode showed significant phylogenetic clustering (Figure 4A), with the mean phylogenetic depth of problem clades (those in which $\geq 50\%$ of species couldn't be differentiated from their congeners) ranging from 0.83% at a 97% identity

Computational evaluation of insect pest metabarcoding

Figure 5: **A**) Phylogenetic tree of all insect genera contained within the curated sequence database, coloured by their mean primer-template mismatch across all evaluated primers. **B**) Primer-template mismatch for each evaluated sequence and primer, summarised by insect genus. **C**) Dot-plot of mean mismatch per genus, with highly mismatched clades indicated. **D**) Summary of results by primer, from top to bottom: mean primer mismatch across all sequences, fold degeneracy on log2 scale, mean phylogenetic distance from each imputed tip to its nearest sequenced taxon.

threshold (consenTRAIT, $p < .001$) to 0.21% at 100% identity threshold ($p < .001$). These patterns indicate that identification failure with DNA barcoding can be considered

a phylogenetically conserved trait, but concentrated in smaller clades scattered throughout the tips of the phylogeny rather than being inherent to any broader

Computational evaluation of insect pest metabarcoding

lineage (Figure 4A). These problem clades mostly occurred within the order Lepidoptera, which contained the highest number of failed species identifications (Figure 4C). In particular, the families *Noctuidae*, *Nymphalidae*, and *Tortricidae* each contained several large problematic clades (Figure 4B). Other non-Lepidopteran taxa that posed problems for DNA barcode based identification included Hoverflies (Diptera: *Syrphidae*), sawflies (Hymenoptera: *Empriidae*) and the hemipteran families *Aphididae* and *Lachnidae* (Figure 4B), however to a substantially lesser degree than Lepidoptera (Figure 4C).

Predicted bias for metabarcoding primers

Primer-template mismatch scores were calculated between the 68 primers and all taxa that had reference sequences available for their respective binding sites, and scores for missing species phylogenetically imputed. Mismatch was found to be phylogenetically conserved across the majority of primers, with the autocorrelation function decaying moderately with phylogenetic divergence to reach a correlation of 0.5 around a phylogenetic distance of 5%, and falling to zero at a distance of 7-10% (supplementary Figure 4). This indicates that missing data imputation would be reliable for those taxa within 5% phylogenetic divergence from a sequenced clade, and close to random for species greater than 7% diverged. Almost all primers had a mean phylogenetic distance to their nearest sequenced taxon (NSTI) between 0.05% and 0.35%, well below this value. The exception was those primers situated at each terminus of the COI barcode where available sequence data was sparser (Figure 2), leaving the NSTI between 4% and 6% (Figure 5D). Following imputation, the forward primers with the lowest mean mismatch across all insect taxa were: C, BF1i, ARF5, and the reverse primers ArR5, E, EN and BR1 (Figure 5B). Nevertheless, these primers still showed significant mismatch to certain clades within the phylogenetic tree, most notably the families *Diaspididae*, *Pseudococcidae*, *Phlopteridae*, *Phlaeothripidae*, *Apidae*, *Gyrinidae*, *Leiodidae*, *Hesperiidae* and the Coleopteran genus *Exapion* (Figure 5C). At a higher level, the orders Hymenoptera and Hemiptera showed substantially more mismatching taxa than any of the other major insect groups (Figure 5A, B, C). Primer-template mismatch was found to be significantly related to primer degeneracy for both the imputed ($p < .001$, $R^2 = .187$) and unimputed datasets (linear regression, $p < .001$, $R^2 = .152$), yet those primers with the highest degeneracy were not necessarily the best performers (Figure 5D). The diminishing returns of adding degeneracy was particularly apparent for the ZBJ-ArtF1c-deg, mtCOIF-XT and MZPlankF2 forward primers and the reverse primer D, which despite extremely high degeneracy still showed moderate mismatch across the insect phylogeny (Figure 5D). The underperformance of highly degenerate primers such as ZBJ-ArtF1c-deg remained when only the unimputed data was viewed (Supplementary Figure 5), suggesting that these results were not confounded by difficulties imputing higher levels of missing data.

Final primer rankings

Many of the evaluated forward primers performed well across all metrics, with C, BF1, fwhF2, BF1i, SauronS878 and

MICOIintF all ranking highly (Figure 6). Nevertheless, there was no perfect forward primer, with many of those showing higher diagnostic sensitivity also having excessive mismatches, or physical characteristics outside of recommended design guidelines (Figure 6). In contrast, there was substantially more variability in the performance of reverse primers, with primers such as D, AgPestR1a and AgPestR1b showing exceptional diagnostic sensitivity, but having either too much degeneracy or a melting temperature well below the recommended guidelines. On the other hand, the reverse primers Ill-C-R, fwhR2n, and BR2 showed slightly less sensitivity, but had no critical issues for any of the measured physical characteristics (Figure 6). To comply with restrictions of 2×150 bp sequencing, the primer combinations fwhF2-fwhR2n, BF1-BR1, or SauronS878-fwhR2n present the best overall options, amplifying a ~250 bp subregion of COI that contains the most diagnostic nucleotides (Figure 2A, C), while showing little mismatch across all insects and physical characteristics within recommended guidelines. The novel AgPestF1-AgPestR1b primers designed in this study would also provide an appropriate 260 bp amplicon for 2×150 bp sequencing, however further laboratory evaluation would be required to ensure their lower than recommended melting temperatures does not introduce non-specific amplification. On the other hand, a much greater range of primer combinations are appropriate for sequencing technologies that can deliver read lengths of 2×300 bp. In particular, those which amplify a subregion from 250 bp into the full-length barcode, along to either the low entropy region around 600 bp, or onwards the 3' terminus will capture the most diagnostic nucleotides (Figure 2C). Published primer combinations that amplify these subregions and performed well in the evaluated metrics include HexCOIF4-HexCOIR4, BF2-BR2 and BF3-BR2, but many of the alternative forward or reverse primers that overlap the same positions would also prove suitable (Figure 2B). Nevertheless, once the amplicon has reached approximately 400 bp there is minimal difference in species discrimination across the COI barcode (Figure 2C), and therefore primer combinations that amplify from the 5' end of the barcode onwards would likely also perform adequately for 2×300 bp sequencing.

Discussion

High sensitivity across a broad taxonomic scope is a defining feature of DNA metabarcoding assays, facilitating their use as a universal diagnostic assay to rapidly screen mixed samples for a range of target pests or pathogens. Despite several studies applying metabarcoding to certain pest taxa (Batovska et al., 2018, 2020; Bowser et al., 2019; Young et al., 2021), the diagnostic performance of the required mini-barcodes has not until now been systematically evaluated across the broader diversity of invasive insect species. Using a curated reference database covering 110,676 insect species, including 2,625 species registered on global invasive species lists, we here demonstrate that mini-barcodes can achieve comparable resolution to the full-length COI barcode region already widely accepted within insect diagnostic protocols. Our findings are largely in agreement with previous

Computational evaluation of insect pest metabarcoding

Figure 6 Summary of primer performance across all metrics measured within this study, with primers ranked from best overall performance (top) to worst (bottom). Metrics from left to right: Average performance of primer compared to regression model predictions on all insects and pest insects, mean mismatch score for each primer, and primer characteristics. Error bars represent 2 standard deviations.

investigations showing congruence between mini-barcodes and the full-length barcode (Hajibabaei, Smith, et al., 2006; Meusnier et al., 2008), as well as morphospecies (Yeo et al., 2020). However, our study expands these predictions to a more than five-fold larger sample of insect taxa, with an additional focus on invasive insect pests.

While 97% identity is considered the default threshold for delineating taxonomic units from DNA barcodes (Alberdi et al., 2018; Porter & Hajibabaei, 2020), our analyses demonstrate that a more stringent 98% or 99% identification threshold not only increases the number of insects that can successfully be differentiated, but also

reduces the amplicon length required to do so. This is particularly notable for implementing metabarcoding on production scale HTS platforms such as the Illumina NovaSeq, which offer the lowest cost per sample but require much shorter amplicons due to their typical read lengths of only 2 × 150 bp (Piper et al., 2019). While use of more stringent identification thresholds has been constrained by the common practice of clustering metabarcoding reads to resolve sequencing errors (Porter & Hajibabaei, 2020), recent denoising algorithms provide single nucleotide resolution that can be leveraged for more accurate and reproducible taxonomic assignment (Callahan et al., 2017; Porter & Hajibabaei, 2020). Nevertheless, the metric of identification success used within our study (proportion of clusters containing only a single species) does not consider the actual availability of reference sequences to match unknown taxa against, and false negatives may be introduced when using these more stringent thresholds if intraspecific diversity isn't sufficiently represented in the reference database. While this will not be an issue for most invasive insect pests due to their general overrepresentation in

public sequence repositories, some taxa were represented by only single sequences, and a further 1,717 had no publicly available COI data whatsoever. Moreover, the 110,676 insect species represented in our curated database barely accounts for 10% of described insect diversity (Stork, 2018), highlighting the considerable efforts still required to increase the taxonomic coverage of reference databases before metabarcoding assays can operate in a truly universal manner.

While many of the evaluated mini-barcodes performed comparably to the full-length barcode region, 10.6% of

Computational evaluation of insect pest metabarcoding

insect pests, and 8.5% of all insect species could not be differentiated at all, even when solely considering 100% matches. While this may at first glance seem high, it is largely consistent with previous research that suggests between 12.3% and 26.5% of described Arthropod species are non-monophyletic for the COI barcode region (Funk & Omland, 2003; Mutanen et al., 2016; Ross, 2014). In our study, these misidentified taxa were phylogenetically clustered in “problem clades” towards the tips of the phylogeny, which predominantly occurred within the order Lepidoptera. Issues of DNA barcode failure for Lepidoptera and other speciose taxonomic groups has long been noted (Meier et al., 2006; Wiemers & Fiedler, 2007), however only recently has it been appreciated how much of this can be attributed to underlying misidentifications, databasing errors, or flawed taxonomic delimitation (Locatelli et al., 2020; Mutanen et al., 2016). While our study has followed current best practices in reference database curation, we notably did not go to the extra length of verifying the original source for the taxonomic identities applied to each sequence, a task which would prove insurmountable for a dataset of this size. While computational curation presents an extremely scalable approach for resolving annotation errors and contamination within public datasets, the quality of any identification ultimately depends on the quality of the systematics and taxonomy that originally delimited and described the species (Clarke & Schutze, 2014). Therefore, additional taxonomic consideration using more comprehensive genomic (Leaché et al., 2014) or integrative methods (Padial et al., 2010) may be required to determine if the problem clades identified in our study are actually due to insufficient resolution within the COI barcode, or rather, over-splitting in the underlying taxonomy (Mutanen et al., 2016).

The 11,431 defunct or synonymous species names identified and corrected in our study clearly demonstrates that taxonomic synonyms remain one of the largest and seldom discussed issues within public sequence repositories (Leray et al., 2019). While taxonomic names must be free to change to reflect revised species concepts, this becomes a problem when historically defunct species names are retained in reference databases, propagating errors through later studies and the management decisions made from them (Clarke et al., 2019). For insect metabarcoding, issues arising from taxonomic synonyms will become most apparent as hierarchical taxonomic classifiers already widely adopted by microbiome researchers become more prevalent (Porter et al., 2014; Porter & Hajibabaei, 2018). These methods require a query sequence to reach a certain bootstrap support to be assigned to lower ranks in the taxonomy, but conflicts in taxonomic annotation between genetically similar reference sequences could result in a failure to reach the required confidence, and thus false negative detections. Therefore, we recommend that resolving taxonomic synonyms to their currently accepted name become a default step in all metabarcoding database curation efforts, alongside the more common practices of removing non-homologous sequences, contaminants, and misannotated taxonomy (Kozlov et al., 2016; Richardson et al., 2020). While some curation of taxonomic synonyms already occurs

within both GenBank and BOLD (Schoch et al., 2020), determining the currently valid taxonomic name from a diverse and constantly evolving primary literature is by no means a trivial task (Schoch et al., 2017). Continued investment into digital infrastructure for distribution of taxonomic information will therefore prove critical for ensuring identifications obtained through metabarcoding remain robust to the inevitable future description and renaming of taxa (Miralles et al., 2020).

In addition to its ability to differentiate target species, the amount of PCR amplification bias towards or against certain taxonomic groups plays an important role in the selection of primers for metabarcoding studies. While the lack of evolutionarily conserved primer binding sites led to early studies questioning the suitability of COI for metabarcoding (Deagle et al., 2014), our analyses demonstrate that incorporating a moderate 216 to 512-fold degeneracy into primers (4-5 degenerate bases) can adequately resolve primer-template mismatch across the large majority of insect taxa, with diminishing returns beyond this. While not explicitly evaluated in this study, previous research has shown that primers with over 2000-fold degeneracy are likely to cause undesired amplification of non-target taxa (Collins et al., 2019), a particular problem for samples with low DNA concentrations (Leese et al., 2021; Macher et al., 2018). With this in mind, we advise against the use of extremely degenerate primers such as ArR5, ZBJ-ArtF1c-deg, or D for insect metabarcoding, in favour of other primers that overlap similar regions of COI and show similar performance despite substantially less degeneracy. Of the taxonomic groups that still showed a high level of mismatch despite inclusion of degenerate bases, the most concerning for an invasive species surveillance programme would be the Armoured Scales (Hemiptera: *Diaspididae*), Mealybugs (Hemiptera: *Pseudococcidae*), and the thrips family *Phlaeothripidae*, for which 120, 90, and 28 species respectively were registered on global invasive species lists. While *Apidae* also showed substantial mismatch across many primers, this was only towards non-pest taxa within the family, and most well-designed primers matched the 26 invasive *Bombus* and *Apis* species well. In all these cases, researchers and diagnosticians should be aware that primer-template mismatches could cause false negatives when these taxa are at a low relative abundance within mixed trap samples. Therefore, the sequencing effort applied to each sample may need be adjusted in proportion to both the overall biomass and expected taxonomic composition of the communities under study, if known in advance.

Despite this study being a purely in-silico evaluation and the many limitations that this entails (Alberdi et al., 2018; Corse et al., 2019; Zhang et al., 2020), it represents a comprehensive first step in applying big data principles to inform development of HTS based diagnostics for invasive insect pests. As well as providing a starting point for diagnosticians and researchers selecting mini-barcode primers for insect identification, our results offer a degree of confidence to managers and regulators grappling with the consequences of non-targeted HTS assays and how to appropriately respond the incidental detection of regulated species (Darling et al., 2020). Whilst we still recommend

Computational evaluation of insect pest metabarcoding

additional laboratory validation be conducted on high-priority targets before adoption in active surveillance, promisingly, many of the outstanding primers highlighted in our rankings have also shown similarly high performance in a recent in-vitro evaluation on a diverse insect mock community (Elbrecht et al., 2019). In summary, appropriately chosen COI mini-barcode primers perform effectively across the majority of pest and non-pest insects, opening for the adoption of universal metabarcoding assays within diagnostic laboratories. While the computational curation pipeline presented here can resolve many issues inherent to public reference sequence data, regardless of whether the diagnostic tool is a microscope or a HTS assay, the accuracy of results will ultimately depend on the underlying quality and completeness of the taxonomy for the target groups.

Acknowledgements

The authors would like to thank Francesco Martoni, Conrad Trollip and Brendan Rodoni for valuable discussions on metabarcoding and taxonomy. Financial support for this study was provided by Agriculture Victoria's Improved Market Access for Horticulture programme (CMI105584) and the Horticulture Innovation Australia iMapPESTS project (ST16010) through funding from the Australian Government Department of Agriculture as part of its Rural R&D for Profit program and Grains Research and Development Corporation. A.M.P. is supported by Australian Government Research Training Program Scholarship.

Author contributions

A.M.P. conceptualised the study, performed all analyses, and drafted the manuscript with input and supervision from N.O.I.C., J.P.C., and M.J.B. All authors read and approved the final version of the manuscript.

Availability of data and materials:

Functions and a tutorial for curating COI reference databases are provided in the 'taxreturn' R package, available on GitHub: <https://github.com/alexpiper/taxreturn>. All code required to reproduce the statistical analyses and figures presented in this paper are contained within the following GitHub repository: https://github.com/alexpiper/primer_evaluation

References

Abd-El Salam, K. A. (2003). Bioinformatic tools and guideline for PCR primer design. *African Journal of Biotechnology*, 2(5), 91–95. <https://doi.org/10.4161/cc.8.22.9956>

Adams, I. P., Fox, A., Boonham, N., Massart, S., & De Jonghe, K. (2018). The impact of high throughput sequencing on plant health diagnostics. *European Journal of Plant Pathology*, 152(4), 909–919. <https://doi.org/10.1007/s10658-018-1570-0>

Adams, I. P., Glover, R. H., Monger, W. A., Mumford, R., Jackeviciene, E., Navalinskiene, M., Samuitiene, M., & Boonham, N. (2009). Next-generation sequencing and metagenomic analysis: A universal diagnostic tool in plant virology. *Molecular Plant Pathology*, 10(4), 537–545. <https://doi.org/10.1111/j.1364-3703.2009.00545.x>

Alberdi, A., Aizpurua, O., Gilbert, M. T. P., & Bohmann, K. (2018). Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods in Ecology and Evolution*, 9(1), 134–147. <https://doi.org/10.1111/2041-210X.12849>

Allcock, R. J. N., Jennison, A. V., & Warrilow, D. (2017). Towards a Universal Molecular Microbiological Test. *Journal of Clinical Microbiology*, 55(11), 3175–3182. <https://doi.org/10.1128/JCM.01155-17>

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)

Andújar, C., Arribas, P., Yu, D. W., Vogler, A. P., & Emerson, B. C. (2018). Why the COI barcode should be the community DNA metabarcode for the metazoa. *Molecular Ecology*, 27(20), 3968–3975. <https://doi.org/10.1111/mec.14844>

Armstrong, K. F., & Ball, S. L. (2005). DNA Barcodes for Biosecurity: Invasive Species Identification. *Philosophical Transactions: Biological Sciences*, 360(1462), 1813–1823. <https://doi.org/10.1098/rstb.2005.1713>

Batovska, J., Lynch, S. E., Cogan, N. O. I., Brown, K., Darbro, J. M., Kho, E. A., & Blacket, M. J. (2018). Effective mosquito and arbovirus surveillance using metabarcoding. *Molecular Ecology Resources*, 18(1), 32–40. <https://doi.org/10.1111/1755-0998.12682>

Batovska, J., Piper, A. M., Valenzuela, I., Cunningham, J. P., & Blacket, M. J. (2020). Developing a Non-destructive Metabarcoding Protocol for Detection of Pest Insects in Bulk Trap Catches. *Research Square*. <https://doi.org/10.21203/rs.3.rs-125070/v1>

Bebber, D. P. (2015). Range-Expanding Pests and Pathogens in a Warming World. *Annual Review of Phytopathology*, 53(1), 335–356. <https://doi.org/10.1146/annurev-phyto-080614-120207>

Bishop, M. J., & Hutchings, P. A. (2011). How useful are port surveys focused on target pest identification for exotic species management? *Marine Pollution Bulletin*, 62(1), 36–42. <https://doi.org/10.1016/j.marpolbul.2010.09.014>

Bowser, M., Burr, S., Davis, I., Dubois, G., Graham, E., Moan, J., & Swenson, S. (2019). A test of metabarcoding for Early Detection and Rapid Response monitoring for non-native forest pest beetles (Coleoptera). *Research Ideas and Outcomes*, 5, e48536. <https://doi.org/10.3897/rio.5.e48536>

Brandon-Mong, G.-J. J., Gan, H.-M. M., Sing, K.-W. W., Lee, P.-S. S., Lim, P.-E. E., & Wilson, J.-J. J. (2015). DNA metabarcoding of insects and allies: An evaluation of primers and pipelines. *Bulletin of Entomological Research*, 105(6), 717–727. <https://doi.org/10.1017/S0007485315000681>

Britton, T., Anderson, C. L., Jacquet, D., Lundqvist, S., & Bremer, K. (2007). Estimating divergence times in large phylogenetic trees. *Systematic Biology*, 56(5), 741–752. <https://doi.org/10.1080/10635150701613783>

Caley, P., Lonsdale, W. M., & Pheloung, P. C. (2006). Quantifying uncertainty in predictions of invasiveness. *Biological Invasions*, 8(2), 277–286. <https://doi.org/10.1007/s10530-004-6703-z>

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME Journal*, 11(12), 2639–2643. <https://doi.org/10.1038/ismej.2017.119>

Carnegie, A. J., & Nahrung, H. F. (2019). Post-Border Forest Biosecurity in Australia: Response to recent Exotic Detections, Current Surveillance and Ongoing Needs. *Forests*, 10(4), 336. <https://doi.org/10.3390/f10040336>

Chamberlain, S. (2017). bold: Interface to Bold Systems API. <https://cran.r-project.org/package=bold>

Clare, E. L., Fazekas, A. J., Ivanova, N. V., Floyd, R. M., Hebert, P. D. N., Adams, A. M., Nagel, J., Girton, R., Newmaster, S. G., & Fenton, M. B. (2019). Approaches to integrating genetic data into ecological networks. *Molecular Ecology*, 28(2), 503–519. <https://doi.org/10.1111/mec.14941>

Clarke, A. R., Li, Z., Qin, Y., Zhao, Z., Liu, L., & Schutze, M. K. (2019). *Bactrocera dorsalis* (Hendel) (Diptera: Tephritidae) is not invasive through Asia: It's been there all along. *Journal of Applied Entomology*, 143(8), 797–801. <https://doi.org/10.1111/jen.12649>

Clarke, A. R., & Schutze, M. K. (2014). The Complexities of Knowing What It Is You Are Trapping. In T. Shelly, N. Epsky, E. B. Jang, J. Reyes-Flores, & R. Vargas (Eds.), *Trapping and the Detection, Control, and Regulation of Tephritid Fruit Flies: Lures, Area-Wide Programs, and Trade Implications* (pp. 611–632). Springer Netherlands. https://doi.org/10.1007/978-94-017-9193-9_18

Collins, R. A., Bakker, J., Wangenstein, O. S., Soto, A. Z., Corrigan, L., Sims, D. W., Genner, M. J., & Mariani, S. (2019). Non-specific amplification compromises environmental DNA metabarcoding with COI. *Methods in Ecology and Evolution*, 10(11), 1985–2001. <https://doi.org/10.1111/2041-210x.13276>

Comtet, T., Sandionigi, A., Viard, F., & Casiraghi, M. (2015). DNA (meta)barcoding of biological invasions: a powerful tool to elucidate invasion processes and help managing aliens. *Biological Invasions*, 17(3), 905–922. <https://doi.org/10.1007/s10530-015-0854-y>

Corse, E., Tougaard, C., Archambaud-Suard, G., Agnès, J. F., Messu-Mandeng, F. D., Bilong Bilong, C. F., Duneau, D., Zinger, L., Chappaz, R., Xu, C. C. Y., Meglécz, E., & Dubut, V. (2019). One-locus-several-primers: A strategy to improve the taxonomic and haplotypic coverage in diet metabarcoding

Computational evaluation of insect pest metabarcoding

- studies. *Ecology and Evolution*, 9(8), 4603–4620. <https://doi.org/10.1002/ece3.5063>
- Crooks, J. A. (2005). Lag times and exotic species: The ecology and management of biological invasions in slow-motion. *Ecoscience*, 12(3), 316–329. <https://doi.org/10.2980/i1195-6860-12-3-316.1>
- Darling, J. A., & Blum, M. J. (2007). DNA-based methods for monitoring invasive species: A review and prospectus. *Biological Invasions*, 9(7), 751–765. <https://doi.org/10.1007/s10530-006-9079-4>
- Darling, J. A., Pochon, X., Abbott, C. L., Inglis, G. J., & Zaiko, A. (2020). The risks of using molecular biodiversity data for incidental detection of species of concern. *Diversity and Distributions*, 26(9), 1116–1121. <https://doi.org/10.1111/ddi.13108>
- Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., & Taberlet, P. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: Not a perfect match. *Biology Letters*, 10(9), 20140562. <https://doi.org/10.1098/rsbl.2014.0562>
- Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, 14(9), 755–763. <https://doi.org/10.1093/bioinformatics/14.9.755>
- Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19), 2460–2461. <https://doi.org/10.1093/bioinformatics/btq461>
- Elbrecht, V., Braukmann, T. W. A., Ivanova, N. V., Prosser, S. W. J., Hajibabaei, M., Wright, M., Zakharov, E. V., Hebert, P. D. N., & Steinke, D. (2019). Validation of COI metabarcoding primers for terrestrial arthropods. *PeerJ*, 7, e7745. <https://doi.org/10.7717/peerj.7745>
- Elbrecht, V., & Leese, F. (2015). Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass-sequence relationships with an innovative metabarcoding protocol. *PLoS ONE*, 10(7), e0130324. <https://doi.org/10.1371/journal.pone.0130324>
- Elbrecht, V., & Leese, F. (2017a). PrimerMiner: an R package for development and in silico validation of DNA metabarcoding primers. *Methods in Ecology and Evolution*, 8(5), 622–626. <https://doi.org/10.1111/2041-210X.12687>
- Elbrecht, V., & Leese, F. (2017b). Validation and Development of COI Metabarcoding Primers for Freshwater Macroinvertebrate Bioassessment. *Frontiers in Environmental Science*, 5, 11. <https://doi.org/10.3389/fenvs.2017.00011>
- EPPO. (2019a). EPPO Standards - Diagnostics. *EPPO Bulletin*, 49(2), 170–174. <https://doi.org/10.1111/epp.12588>
- EPPO. (2019b). PM 7/98 (4) Specific requirements for laboratories preparing accreditation for a plant pest diagnostic activity. *EPPO Bulletin*, 49(3), 530–563. <https://doi.org/10.1111/epp.12629>
- Felsenstein, J. (1985). Phylogenies and the comparative method. *American Naturalist*, 125(1), 1–15. <https://doi.org/10.1086/284325>
- Ficetola, G. F., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessière, J., Taberlet, P., & Pompanon, F. (2010). An In silico approach for the evaluation of DNA barcodes. *BMC Genomics*, 11, 434. <https://doi.org/10.1186/1471-2164-11-434>
- 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.
- Fong, Y. (2019). Fast Bootstrap Confidence Intervals for Continuous Threshold Linear Regression. *Journal of Computational and Graphical Statistics*, 28(2), 466–470. <https://doi.org/10.1080/10618600.2018.1537927>
- Fong, Y., Huang, Y., Gilbert, P. B., & Permar, S. R. (2017). chngpt: Threshold regression model estimation and inference. *BMC Bioinformatics*, 18(1), 1–7. <https://doi.org/10.1186/s12859-017-1863-x>
- Funk, D. J., & Omland, K. E. (2003). Species-Level Paraphyly and Polyphyly: Frequency, Causes, and Consequences, with Insights from Animal Mitochondrial DNA. *Annual Review of Ecology, Evolution, and Systematics*, 34(1), 397–423. <https://doi.org/10.1146/annurev.ecol-sys.34.011802.132421>
- Galan, M., Pons, J. B., Tournayre, O., Pierre, É., Leuchtmann, M., Pontier, D., & Charbonnel, N. (2018). Metabarcoding for the parallel identification of several hundred predators and their prey: Application to bat species diet analysis. *Molecular Ecology Resources*, 18(3), 474–489. <https://doi.org/10.1111/1755-0998.12749>
- Garg, A., Leipe, D., & Uetz, P. (2019). The disconnect between DNA and species names: Lessons from reptile species in the NCBI taxonomy database. *Zootaxa*, 4706(3), 401–407. <https://doi.org/10.11646/zootaxa.4706.3.1>
- Geller, J., Meyer, C., Parker, M., & Hawk, H. (2013). Redesign of PCR primers for mitochondrial cytochrome c oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. *Molecular Ecology Resources*, 13(5), 851–861. <https://doi.org/10.1111/1755-0998.12138>
- Gibson, J. F., Shokralla, S., Curry, C., Baird, D. J., Monk, W. A., King, I., & Hajibabaei, M. (2015). Large-scale biomonitoring of remote and threatened ecosystems via high-throughput sequencing. *PLoS ONE*, 10(10), e0138432. <https://doi.org/10.1371/journal.pone.0138432>
- Gibson, J. F., Shokralla, S., Porter, T. M., King, I., van Konynenburg, S., Janzen, D. H., Hallwachs, W., & Hajibabaei, M. (2014). Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasyntematics. *Proceedings of the National Academy of Sciences of the United States of America*, 111(22), 8007–8012. <https://doi.org/10.1073/pnas.1406468111>
- Hajibabaei, M., Janzen, D. H., Burns, J. M., Hallwachs, W., & Hebert, P. D. N. (2006). DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences of the United States of America*, 103(4), 968–971. <https://doi.org/10.1073/pnas.0510466103>
- Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G. A. C., & Baird, D. J. (2011). Environmental barcoding: A next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE*, 6(4), e17497. <https://doi.org/10.1371/journal.pone.0017497>
- Hajibabaei, M., Smith, M. A., Janzen, D. H., Rodriguez, J. J., Whitfield, J. B., & Hebert, P. D. N. (2006). A minimalist barcode can identify a specimen whose DNA is degraded. *Molecular Ecology Notes*, 6(4), 959–964. <https://doi.org/10.1111/j.1471-8286.2006.01470.x>
- Hajibabaei, M., Spall, J. L., Shokralla, S., & van Konynenburg, 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. <https://doi.org/10.1186/1472-6785-12-28>
- Handelsman, J. (2004). Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews*, 68(4), 669–685. <https://doi.org/10.1128/MBR.68.4.669>
- Hebert, P. D. N., 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 of the United States of America*, 101(41), 14812–14817. <https://doi.org/10.1073/pnas.0406166101>
- Hernández-Triana, L. M., Prosser, S. W., Rodríguez-Perez, M. A., Chaverri, L. G., Hebert, P. D. N., & Ryan Gregory, T. (2014). Recovery of DNA barcodes from blackfly museum specimens (Diptera: Simuliidae) using primer sets that target a variety of sequence lengths. *Molecular Ecology Resources*, 14(3), 508–518. <https://doi.org/10.1111/1755-0998.12208>
- Hinchliff, C. E., Smith, S. A., Allman, J. F., Burleigh, J. G., Chaudhary, R., Coghill, L. M., Crandall, K. A., Deng, J., Drew, B. T., Gazis, R., Gude, K., Hibbett, D. S., Katz, L. A., Dail Laughinghouse, H., McTavish, E. J., Midford, P. E., Owen, C. L., Ree, R. H., Rees, J. A., ... Cranston, K. A. (2015). Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences of the United States of America*, 112(41), 12764–12769. <https://doi.org/10.1073/pnas.1423041112>
- Hulme, P. E. (2009). Trade, transport and trouble: Managing invasive species pathways in an era of globalization. *Journal of Applied Ecology*, 46(1), 10–18. <https://doi.org/10.1111/j.1365-2664.2008.01600.x>
- Jordaens, K., Sonet, G., Richet, R., Dupont, E., Braet, Y., & Desmyter, S. (2013). Identification of forensically important *Sarcophaga* species (Diptera: Sarcophagidae) using the mitochondrial COI gene. *International Journal of Legal Medicine*, 127(2), 491–504. <https://doi.org/10.1007/s00414-012-0767-6>
- Kozlov, A. M., Zhang, J., Yilmaz, P., Glöckner, F. O., & Stamatakis, A. (2016). Phylogeny-aware identification and correction of taxonomically mislabeled sequences. *Nucleic Acids Research*, 44(11), 5022–5033. <https://doi.org/10.1093/nar/gkw396>

Computational evaluation of insect pest metabarcoding

- Kwok, S., Chang, S. Y., Sninsky, J. J., & Wang, A. (1994). A guide to the design and use of mismatched and degenerate primers. *Genome Research*, 3, S39–S47. <https://doi.org/10.1101/gr.3.4.S39>
- Langille, M. G. I., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., Clemente, J. C., Burkepile, D. E., Vega Thurber, R. L., Knight, R., Beiko, R. G., & Huttenhower, C. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, 31(9), 814–821. <https://doi.org/10.1038/nbt.2676>
- Leaché, A. D., Fujita, M. K., Minin, V. N., & Bouckaert, R. R. (2014). Species delimitation using genome-wide SNP Data. *Systematic Biology*, 63(4), 534–542. <https://doi.org/10.1093/sysbio/syu018>
- Leese, F., Sander, M., Buchner, D., Elbrecht, V., Haase, P., & Zizka, V. M. A. (2021). Improved freshwater macroinvertebrate detection from environmental DNA through minimized nontarget amplification. *Environmental DNA*, 3(1), 261–276. <https://doi.org/10.1002/edn3.177>
- Leray, M., Knowlton, N., Ho, S. L., Nguyen, B. N., & Machida, R. J. (2019). GenBank is a reliable resource for 21st century biodiversity research. *Proceedings of the National Academy of Sciences of the United States of America*, 116(45), 22651–22656. <https://doi.org/10.1073/pnas.1911714116>
- Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., Boehm, J. T., & Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: Application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, 10(1), 34. <https://doi.org/10.1186/1742-9994-10-34>
- Li, J. (2019). TmCalculator: Melting Temperature of Nucleic Acid Sequences. <https://cran.r-project.org/package=TmCalculator>
- Liebholt, A. M., Beres, L., Brockerhoff, E. G., Epanchin-Niell, R. S., Hastings, A., Herms, D. A., Kean, J. M., McCullough, D. G., Suckling, D. M., Tobin, P. C., & Yamanaka, T. (2016). Eradication of Invading Insect Populations: From Concepts to Applications. *Annual Review of Entomology*, 61(1), 335–352. <https://doi.org/10.1146/annurev-ento-010715-023809>
- Locatelli, N. S., McIntyre, P. B., Therkildsen, N. O., & Baetscher, D. S. (2020). GenBank’s reliability is uncertain for biodiversity researchers seeking species-level assignment for eDNA. *Proceedings of the National Academy of Sciences of the United States of America*, 117(51), 32211–32212. <https://doi.org/10.1073/pnas.2007421117>
- Louca, S., & Doebeli, M. (2018). Efficient comparative phylogenetics on large trees. *Bioinformatics*, 34(6), 1053–1055. <https://doi.org/10.1093/bioinformatics/btx701>
- Macher, J. N., Vivancos, A., Piggott, J. J., Centeno, F. C., Mattheae, C. D., & Leese, F. (2018). Comparison of environmental DNA and bulk-sample metabarcoding using highly degenerate cytochrome c oxidase I primers. *Molecular Ecology Resources*, 18(6), 1456–1468. <https://doi.org/10.1111/1755-0998.12940>
- Machida, R. J., Leray, M., Ho, S. L., & Knowlton, N. (2017). Data Descriptor: Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. *Scientific Data*, 4, 170027. <https://doi.org/10.1038/sdata.2017.27>
- Maree, H. J., Fox, A., Al Rwhanih, M., Boonham, N., & Candresse, T. (2018). Application of HTS for routine plant virus diagnostics: state of the art and challenges. *Frontiers in Plant Science*, 9, 1082. <https://doi.org/10.3389/fpls.2018.01082>
- Marquina, D., Andersson, A. F., & Ronquist, F. (2019). New mitochondrial primers for metabarcoding of insects, designed and evaluated using in silico methods. *Molecular Ecology Resources*, 19(1), 90–104. <https://doi.org/10.1111/1755-0998.12942>
- Martiny, A. C., Treseder, K., & Pusch, G. (2013). Phylogenetic conservatism of functional traits in microorganisms. *ISME Journal*, 7(4), 830–838. <https://doi.org/10.1038/ismej.2012.160>
- Meier, R., Shiyang, K., Vaidya, G., & Ng, P. K. L. (2006). DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Systematic Biology*, 55(5), 715–728. <https://doi.org/10.1080/10635150600969864>
- Meusnier, I., Singer, G. A. C., Landry, J. F., Hickey, D. A., Hebert, P. D. N., & Hajibabaei, M. (2008). A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics*, 9(1), 1–4. <https://doi.org/10.1186/1471-2164-9-214>
- Meyer, C. P. (2003). Molecular systematics of cowries (Gastropoda: Cypraeaidae) and diversification patterns in the tropics. *Biological Journal of the Linnean Society*, 79(3), 401–459. <https://doi.org/10.1046/j.1095-8312.2003.00197.x>
- Meyerson, L. A., & Reaser, J. K. (2002). Biosecurity: Moving toward a comprehensive approach. *BioScience*, 52(7), 593–600. [https://doi.org/10.1641/0006-3568\(2002\)052\[0593:BMTACA\]2.0.CO;2](https://doi.org/10.1641/0006-3568(2002)052[0593:BMTACA]2.0.CO;2)
- Miralles, A., Bruy, T., Wolcott, K., Scherz, M. D., Begerow, D., Beszteri, B., Bonkowski, M., Felden, J., Gemeinholzer, B., Glaw, F., Glöckner, F. O., Hawlitschek, O., Kostadinov, I., Nattkemper, T. W., Printzen, C., Renz, J., Rybalka, N., Stadler, M., Weibulat, T., ... Vences, M. (2020). Repositories for taxonomic data: Where we are and what is missing. *Systematic Biology*, 69(6), 1231–1253. <https://doi.org/10.1093/sysbio/syaa026>
- Misof, B., Liu, S., Meusemann, K., Peters, R. S., Donath, A., Mayer, C., Frandsen, P. B., Ware, J., Flouri, T., Beutel, R. G., Niehuis, O., Petersen, M., Izquierdo-Carrasco, F., Wappler, T., Rust, J., Aberer, A. J., Aspöck, U., Aspöck, H., Bartel, D., ... Zhou, X. (2014). Phylogenomics resolves the timing and pattern of insect evolution. *Science*, 346(6210), 763–767. <https://doi.org/10.1126/science.1257570>
- Mutanen, M., Kivelä, S. M., Vos, R. A., Doorenweerd, C., Ratnasingham, S., Hausmann, A., Huemer, P., Dinča, V., Van Nieuwerkerken, E. J., Lopez-Vaamonde, C., Vila, R., Aarvik, L., Decaens, T., Efetov, K. A., Hebert, P. D. N., Johnsen, A., Karsholt, O., Pentinsaari, M., Rougerie, R., ... Godfray, H. C. J. (2016). Species-level para- and polyphyly in DNA barcode gene trees: Strong operational bias in European Lepidoptera. *Systematic Biology*, 65(6), 1024–1040. <https://doi.org/10.1093/sysbio/syw044>
- Padial, J. M., Miralles, A., De la Riva, I., & Vences, M. (2010). The integrative future of taxonomy. *Frontiers in Zoology*, 7, 16. <https://doi.org/10.1186/1742-9994-7-16>
- Page, H., Aboyou, P., Gentleman, R., & DebRoy, S. (2019). Biostrings: Efficient manipulation of biological strings.
- Park, D. S., Footitt, R., Maw, E., & Hebert, P. D. N. (2011). Barcoding bugs: DNA-based identification of the true bugs (insecta: Hemiptera: Heteroptera). *PLoS ONE*, 6(4), e18749. <https://doi.org/10.1371/journal.pone.0018749>
- Pennell, M. W., Eastman, J. M., Slater, G. J., Brown, J. W., Uyeda, J. C., Fitzjohn, R. G., Alfaro, M. E., & Harmon, L. J. (2014). Geiger v2.0: An expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics*, 30(15), 2216–2218. <https://doi.org/10.1093/bioinformatics/btu181>
- Pentinsaari, M., Ratnasingham, S., Miller, S. E., & Hebert, P. D. N. (2020). BOLD and GenBank revisited - Do identification errors arise in the lab or in the sequence libraries? *PLoS ONE*, 15(4), e0231814. <https://doi.org/10.1371/journal.pone.0231814>
- Pentinsaari, M., Salmela, H., Mutanen, M., & Roslin, T. (2016). Molecular evolution of a widely-adopted taxonomic marker (COI) across the animal tree of life. *Scientific Reports*, 6, 35275. <https://doi.org/10.1038/srep35275>
- Piper, A. M., Batovska, J., Cogan, N. O. I., Weiss, J., Cunningham, J. P., Rodoni, B. C., & Blacket, M. J. (2019). Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance. *GigaScience*, 8(8), giz092. <https://doi.org/10.1093/gigascience/giz092>
- Porter, T. M., Gibson, J. F., Shokralla, S., & Baird, D. J. (2014). Rapid and accurate taxonomic classification of insect (class Insecta) cytochrome c oxidase subunit 1 (COI) DNA barcode sequences using a naive Bayesian classifier. *Molecular Ecology Resources*, 14, 929–942. <https://doi.org/10.5061/dryad.bc8pc>
- Porter, T. M., & Hajibabaei, M. (2018). Automated high throughput animal COI metabarcoding classification. *Scientific Reports*, 8, 4226. <https://doi.org/10.1038/s41598-018-22505-4>
- Porter, T. M., & Hajibabaei, M. (2020). Putting COI Metabarcoding in Context: The Utility of Exact Sequence Variants (ESVs) in Biodiversity Analysis. *Frontiers in Ecology and Evolution*, 8, 248. <https://doi.org/10.3389/fevo.2020.00248>
- Price, M. N., Dehal, P. S., & Arkin, A. P. (2009). Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, 26(7), 1641–1650. <https://doi.org/10.1093/molbev/msp077>

Computational evaluation of insect pest metabarcoding

- Prosser, S. W. J., Dewaard, J. R., Miller, S. E., & Hebert, P. D. N. (2016). DNA barcoding from century-old type specimens using next-generation sequencing. *Molecular Ecology Resources*, 16(2), 487–497. <https://doi.org/10.1111/1755-0998.12474>
- Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD : The Barcode of Life Data System (www.barcodinglife.org). *Molecular Ecology Notes*, 7(3), 355–364. <https://doi.org/10.1111/j.1471-8286.2006.01678.x>
- Reaser, J. K., Burgiel, S. W., Kirkey, J., Brantley, K. A., Veatch, S. D., & Burgos-Rodríguez, J. (2020). The early detection of and rapid response (EDRR) to invasive species: a conceptual framework and federal capacities assessment. *Biological Invasions*, 22(1), 1–19. <https://doi.org/10.1007/s10530-019-02156-w>
- Reaser, J. K., Frey, M., & Meyers, N. M. (2020). Invasive species watch lists: guidance for development, communication, and application. *Biological Invasions*, 22(1), 47–51. <https://doi.org/10.1007/s10530-019-02176-6>
- Reaser, J. K., Meyerson, L. A., & von Holle, B. (2008). Saving camels from straws: How propagule pressure-based prevention policies can reduce the risk of biological invasion. *Biological Invasions*, 10(7), 1085–1098. <https://doi.org/10.1007/s10530-007-9186-x>
- Rennstam Rubbmark, O., Sint, D., Horngacher, N., & Traugott, M. (2018). A broadly applicable COI primer pair and an efficient single-tube amplicon library preparation protocol for metabarcoding. *Ecology and Evolution*, 8(24), 12335–12350. <https://doi.org/10.1002/ece3.4520>
- Richardson, R. T., Sponsler, D. B., McMinn-Sauder, H., & Johnson, R. M. (2020). MetaCurator: A hidden Markov model-based toolkit for extracting and curating sequences from taxonomically-informative genetic markers. *Methods in Ecology and Evolution*, 11(1), 181–186. <https://doi.org/10.1111/2041-210X.13314>
- Roehorst, J. W., de Krom, C., Fox, A., Mehle, N., Ravnika, M., & Werkman, A. W. (2018). Ensuring validation in diagnostic testing is fit for purpose: a view from the plant virology laboratory. *EPPO Bulletin*, 48(1), 105–115. <https://doi.org/10.1111/epp.12445>
- Ross, H. A. (2014). The incidence of species-level paraphyly in animals: A re-assessment. *Molecular Phylogenetics and Evolution*, 76(1), 10–17. <https://doi.org/10.1016/j.ympev.2014.02.021>
- SantaLucia, J., & Hicks, D. (2004). The thermodynamics of DNA structural motifs. *Annual Review of Biophysics and Biomolecular Structure*, 33, 415–440. <https://doi.org/10.1146/annurev.biophys.32.110601.141800>
- Scheider, T. D., & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18(20), 6097–6100.
- Schoch, C. L., Aime, M. C., de Beer, W., Crous, P. W., Hyde, K. D., Penev, L., Seifert, K. A., Stadler, M., Zhang, N., & Miller, A. N. (2017). Using standard keywords in publications to facilitate updates of new fungal taxonomic names. *IMA Fungus*, 8(2), 70–73. <https://doi.org/10.1007/BF03449466>
- Schoch, C. L., Ciuffo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovan-skaya, R., Leipe, D., McVeigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., & Karsch-Mizrachi, I. (2020). NCBI Taxonomy: A comprehensive update on curation, resources and tools. *Database*, baaa062. <https://doi.org/10.1093/database/baaa062>
- Schrader, G., & Unger, J. G. (2003). Plant quarantine as a measure against invasive alien species: The framework of the International Plant Protection Convention and the plant health regulations in the European Union. *Biological Invasions*, 5(4), 357–364. <https://doi.org/10.1023/B:BINV.00000005567.58234.b9>
- Shen, Z., Qu, W., Wang, W., Lu, Y., Wu, Y., Li, Z., Hang, X., Wang, X., Zhao, D., & Zhang, C. (2010). MPprimer: A program for reliable multiplex PCR primer design. *BMC Bioinformatics*, 11, 143. <https://doi.org/10.1186/1471-2105-11-143>
- Shokralla, S., Porter, T. M., Gibson, J. F., Dobosz, R., Janzen, D. H., Hallwachs, W., Golding, G. B., & Hajibabaei, M. (2015). Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific Reports*, 5, 9687. <https://doi.org/10.1038/srep09687>
- Siddall, M. E., Fontanella, F. M., Watson, S. C., Kvist, S., & Erséus, C. (2009). Barcoding bamboozled by bacteria: Convergence to metazoan mitochondrial primer targets by marine microbes. *Systematic Biology*, 58(4), 445–451. <https://doi.org/10.1093/sysbio/syp033>
- Simberloff, D. (2006). Risk assessments, blacklists, and white lists for introduced species: Are predictions good enough to be useful? *Agricultural and Resource Economics Review*, 35(1), 1–10. <https://doi.org/10.1017/S1068280500010005>
- Stadhouders, R., Pas, S. D., Anber, J., Voermans, J., Mes, T. H. M., & Schutten, M. (2010). The effect of primer-template mismatches on the detection and quantification of nucleic acids using the 5' nuclease assay. *Journal of Molecular Diagnostics*, 12(1), 109–117. <https://doi.org/10.2353/jmoldx.2010.090035>
- Stork, N. E. (2018). How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth? *Annual Review of Entomology*, 63, 31–45. <https://doi.org/10.1146/annurev-ento-020117-043348>
- Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, 21(8), 2045–2050. <https://doi.org/10.1111/j.1365-294X.2012.05470.x>
- Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *American Mathematical Society: Lectures on Mathematics in the Life Sciences*, 17(2), 57–86.
- Tedersoo, L., Drenkhan, R., Anslan, S., Morales-Rodriguez, C., & Cleary, M. (2019). High-throughput identification and diagnostics of pathogens and pests: overview and practical recommendations. *Molecular Ecology Resources*, 19, 47–76. <https://doi.org/10.1111/1755-0998.12959>
- Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gks596>
- Vamos, E. E., Elbrecht, V., & Leese, F. (2017). Short COI markers for freshwater macroinvertebrate metabarcoding. *Metabarcoding and Metagenomics*, 1, e14625. <https://doi.org/10.3897/mbmg.1.14625>
- Wangenstein, O. S., Palacín, C., Guardiola, M., & Turon, X. (2018). DNA metabarcoding of littoral hardbottom communities: High diversity and database gaps revealed by two molecular markers. *PeerJ*, 6, e4705. <https://doi.org/10.7717/peerj.4705>
- Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. <http://ggplot2.org>
- Wiemers, M., & Fiedler, K. (2007). Does the DNA barcoding gap exist? - A case study in blue butterflies (Lepidoptera: Lycaenidae). *Frontiers in Zoology*, 4, 8. <https://doi.org/10.1186/1742-9994-4-8>
- Wilkinson, S. (2018). kmer: an R package for fast alignment-free clustering of biological sequences. <https://doi.org/10.5281/zenodo.1227690>
- Wilkinson, S. (2019). aphid: an R package for analysis with profile hidden Markov models. *Bioinformatics*, 35(19), 3829–3830. <https://doi.org/10.1093/bioinformatics/btz159>
- Winter, David, J., & Winter, D. J. (2017). rentrez: An R package for the NCBI eUtils API. *The R Journal*, 9(2), 520–526. <https://doi.org/10.7287/peerj.preprints.3179v2>
- Wright, E. S. (2016). Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R. *The R Journal*, 8(1), 352–359. <https://doi.org/10.32614/RJ-2016-025>
- Yeo, D., Srivathsan, A., & Meier, R. (2020). Longer is Not Always Better: Optimizing Barcode Length for Large-Scale Species Discovery and Identification. *Systematic Biology*, 69(5), 999–1015. <https://doi.org/10.1093/sysbio/syaa014>
- Young, R. G., Milián-García, Y., Yu, J., Bullas-Appleton, E., & Hanner, R. H. (2021). Biosurveillance for invasive insect pest species using an environmental DNA metabarcoding approach and a high salt trap collection fluid. *Ecology and Evolution*, 11(4), 1558–1569. <https://doi.org/10.1002/ece3.7113>
- Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., & Ding, Z. (2012). Biodiversity soup : metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, 3, 613–623. <https://doi.org/10.1111/j.2041-210X.2012.00198.x>
- Yu, G., Lam, T. T. Y., Zhu, H., & Guan, Y. (2018). Two methods for mapping and visualizing associated data on phylogeny using ggtree. *Molecular Biology and Evolution*, 35(12), 3041–3043. <https://doi.org/10.1093/molbev/msy194>
- Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T. Y. (2017). ggtree: an R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data. *Methods in Ecology and Evolution*, 8(1), 28–36. <https://doi.org/10.1111/2041-210X.12628>

Computational evaluation of insect pest metabarcoding

Zaneveld, J. R. R., & Thurber, R. L. V. (2014). Hidden state prediction: A modification of classic ancestral state reconstruction algorithms helps unravel complex symbioses. *Frontiers in Microbiology*, 5, 431. <https://doi.org/10.3389/fmicb.2014.00431>

Zeale, M. R. K., Butlin, R. K., Barker, G. L. A., Lees, D. C., & Jones, G. (2011). Taxon-specific PCR for DNA barcoding arthropod prey in bat faeces.

Molecular Ecology Resources, 11(2), 236–244. <https://doi.org/10.1111/j.1755-0998.2010.02920.x>

Zhang, S., Zhao, J., & Yao, M. (2020). A comprehensive and comparative evaluation of primers for metabarcoding eDNA from fish. *Methods in Ecology and Evolution*, 11(12), 1609–1625. <https://doi.org/10.1111/2041-210X.13485>