**In-silico evaluation of DNA Metabarcoding as a universal surveillance tool for insect pests**

Alexander M. Piper1,2, John Paul Cunningham1, Noel O.I. Cogan1,2, Mark J. Blacket1

1 Agriculture Victoria Research, AgriBio, Bundoora, Victoria, Australia

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

**Corresponding author:**

**Alexander M. Piper**

Email: [alexander.piper@agriculture.vic.gov.au](mailto:alexander.piper@agriculture.vic.gov.au)

**Running title:** In-silico validation of metabarcoding diagnostics

**Target journal:**

* Methods in Ecology and Evolution
* eLIfe?
* Molecular Ecology
* Scientific Reports
* PeerJ

# Abstract

As the possibility of using metabarcoding as a ‘universal’ diagnostic has been raised, it is important that the appropriateness of mini barcode markers be evaluated, and the taxa for which this approach will be successful defined a-priori

Appropriate design and selection of PCR primers play a critical role in determining the sensitivity and specificity of a metabarcoding assay. Despite several recent studies applying metabarcoding within an insect biosecurity framework, the diagnostic ability of the smaller barcode regions for insect pests has not yet been systematically evaluated.. Here we curate a large reference database of insect DNA sequences and evaluate the diagnostic ability of a range of COI primers on this dataset, as well as just insect pests. We further predict the bias that each primer set will show for different taxa and the potential for off-target amplification. We find that well designed PCR primers can provide near comparable resolution to the barcode primers traditionally used for DNA barcoding-based insect diagnostics, even with amplicon sizes reduced to work with contemporary high throughput sequencing platforms. Of the taxa that could not be successfully identified, the majority came from classic problem groups that conventional DNA barcoding struggles with. We further found that many incorrect identifications were the results of misannotated taxonomy or taxonomic synonyms, highlighting the importance of using curated reference databases for DNA metabarcoding. We conclude that while not totally ‘universal’, DNA metabarcoding performs well across the large majority of insect taxa, and presents a valuable tool for future broad-spectrum surveillance programmes.

More broadly, we aim to evaluate the potential for DNA metabarcoding to act as a high throughput surveillance tool for insect pests

**Keywords:**

# Introduction

Early diagnosis and rapid response form a crucial component in reducing the spread and establishment of invasive pests and pathogens (Liebhold et al., 2016). Historically, quarantine legislation has centred around risk analysis and targeted inspections for predefined lists of regulated organisms (Schrader & Unger, 2003; Andersen, Adams, Hope, & Powell, 2004; S. Augustin et al., 2012; Sylvie Augustin et al., 2012). However, as global trade networks become increasingly interlinked (Hulme, 2009) and anthropogenic climate change alters species range distributions (Daniel P Bebber, Ramotowski, & Gurr, 2013), this list-based legislation can often lag behind the speed with which new pests can emerge and spread across borders (Daniel Patrick Bebber, 2015). The inadequacy of list-based inspection and diagnostics becomes particularly apparent when considering impacts beyond just agroecosystems, but also on biodiversity and ecosystem services where the size and complexity of the natural environment often makes assessing the potential impacts of a newly introduced species in advance challenging (Caley, Lonsdale, & Pheloung, 2006; Blackburn et al., 2014). In this modern biosecurity environment it is becoming increasingly obvious that a more holistic approach to surveillance that aims to detect and evaluate all pests and diseases, not just those that are regulated by national quarantine agencies is required to protect trade, industry and environment (Simberloff, 2006; Bishop & Hutchings, 2011). Ideally, a broad-spectrum surveillance programme would use a mixture of passive and active sampling strategies to capture, identify, and evaluate risk for all new organisms introduced to an environment through comparison to baseline knowledge of endemic biodiversity (Simmons et al., 2016; Scott et al., 2017; Trebitz et al., 2017). However, ongoing broad-spectrum surveillance on the scales required would capture a large amount of endemic diversity, and the sheer number of specimens that need to be sorted through and identified to a taxonomic rank that is informative for management will form a major diagnostic bottleneck (Davidovitch et al., 2009; Westfall, Therriault, & Abbott, 2020).

Plant pest and pathogen diagnostics currently rely on a mixture of morphological examination, biochemical techniques and both targeted and universal molecular assays such as diagnostic qPCR, and DNA barcoding (EPPO, 2019). While these methods can provide rapid and highly accurate identification of small numbers of specimens, the inherent restriction of analysing only on single-specimens at a time constrains their application to the large number of taxa captured in a broad-spectrum surveillance programme (Comtet, Sandionigi, Viard, & Casiraghi, 2015; Bulman, McDougal, Hill, & Lear, 2018; Piper et al., 2019). As an alternative to single specimen methods, recent high-throughput sequencing (HTS) platforms can comprehensively characterise mixed populations of genomic DNA (metagenomics), RNA (metatranscriptomics) or taxonomically informative marker genes (metabarcoding). While first emerging for exploring bio-diversity (Handelsman, 2004; Williamson, 2011; Taberlet, Coissac, Pompanon, Brochmann, & Willerslev, 2012; Deiner et al., 2017), these approaches have been recently co-opted by various disciplines of molecular diagnostics, where they been used to identify large reportable ranges of eukaryotic, prokaryotic and viral organisms (Brown, Chain, Zhan, MacIsaac, & Cristescu, 2016; Arulandhu et al., 2017; Blauwkamp et al., 2019; Jana Batovska, Mee, Lynch, Sawbridge, & Rodoni, 2019). High sensitivity with low specificity is a defining feature of non-targeted HTS assays, which lends themselves to use more akin to a surveillance tool to rapidly screen communities for a range of different invasive taxa. If the same assay can be used across diverse taxonomic or physiological groups, non-targeted diagnostics have the potential to broaden the diversity of organisms within the scope of a diagnostic laboratory, as well as decrease the costs of implementation (Massart, Olmos, Jijakli, & Candresse, 2014; Adams, Fox, Boonham, Massart, & De Jonghe, 2018), and may even fulfil the role of the ‘Invasive species chip’ originally speculated upon by (Darling & Blum, 2007).

Metabarcoding of the mitochondrial cytochrome c oxidase subunit 1 (COI) locus provides the most attainable method for integration of HTS diagnostics into insect pest surveillance. This is largely due to its cost effectiveness, access to substantial public reference database, and ability to leverage the already wide adoption of conventional DNA barcoding within diagnostic and regulatory frameworks (Piper et al., 2019). However while conventional single specimen DNA barcoding generally amplifies 658-bp region of this locus (Folmer, Black, Hoeh, Lutz, & Vrijenhoek, 1994) modern HTS platforms impose strict limitations on molecule length that can be sequenced and therefore smaller stretches of this conventional barcode region or ‘mini-barcodes’ must instead be used (Brandon-Mong et al., 2015). The diagnostic information contained within these mini barcodes is the primary determinant of the sensitivity and specificity for a metabarcoding assay, and thus its applicability within a broad-spectrum surveillance context. However despite several studies applying metabarcoding to insect pests (J Batovska et al., 2018; Bowser et al., 2019)( Batovska et al 2020), the diagnostic ability of these smaller barcode regions for a broad diversity of insect pests has not yet been systematically validated. This is particularly important, as due to the non-targeted nature of metabarcoding it is common that when incorporated into a surveillance programme reveals the presence of previously unrecorded or cryptic taxa that have been missed by former targeted surveys (Simmons et al., 2016; Westfall et al., 2020)(Batovska et al 2020). While this is advantageous to the overall concept of a broad-spectrum surveillance program, appropriate interpretation and response to these incidental detections has been identified as one of the major challenges for adoption into active surveillance programs (Massart et al., 2017; Adams et al., 2018; Piper et al., 2019), and this necessarily requires a-priori information of the differentiating power of the primers across a broad range of taxa.

Unlike single specimen diagnostics analysing entire communities brings further the risk of missing rare target individuals due to dilution within a mixture of other species (Elbrecht & Leese, 2017b) and this can be exacerbated by size differences between specimens (Elbrecht, Peinert, & Leese, 2017) and low sequencing depths or DNA concentrations (Leray & Knowlton, 2017; Kelly, Shelton, & Gallego, 2019; Elbrect et al., 2020). While this should be a simple matter to resolve if there was a simple quantitative relationship between input and output, biased detection of molecules from certain taxa over others throughout the laboratory process can large deviations from expected proportions and further compound this problem (Elbrecht & Leese, 2015). This bias is thought to primarily arise from mismatch between PCR primers and template sequence (Clarke, Soubrier, Weyrich, & Cooper, 2014; Piñol, Mir, Gomez-Polo, & Agustí, 2015), and is particularly problematic for protein-coding genes such as COI where variability in the third position of each codon leaves no strictly conserved nucleotide sites for placement of universal PCR primers (Deagle, Jarman, Coissac, Pompanon, & Taberlet, 2014). To overcome this, degenerate nucleotide bases are commonly incorporated into primers to ensure they match the target groups appropriately (Elbrecht et al., 2019). However, overuse of degenerate bases to resolve these issues can cause undesired amplification of non-target organisms and thus ‘wasted’ sequencing effort (Collins et al., 2019). While the challenges of non-specific amplification are not as problematic for DNA extracts from homogenised bulk specimens with high DNA concentrations (Macher et al., 2018; Elbrecht et al., 2019), as some studies begin to adopt non-destructive DNA extractions for metabarcoding the lower DNA concentrations may present similar issues (Batovska et al 2020, Martoni et al 2020).

Historically, molecular diagnostic assays would have undergone a stringent laboratory based validation process in order to resolve the aforementioned issues and establish performance parameters for every individual target designated in an unambiguously defined scope (European and Mediterranean Plant Protection Organization, 2010). However, when considering the sheer number of potential pests, hosts and matrices that would need to be evaluated, it is evident that validation processes cannot be applied in their usual sense and must be adapted for non-targeted diagnostics (Maree, Fox, Al Rwahnih, Boonham, & Candresse, 2018; Roenhorst et al., 2018). Instead, In-silico methods present an ideal route for establishing performance criteria such as analytical specificity even if physical specimens of the target are unavailable (refs). While in-silico methods will not capture all of the factors involved (Elbrecht & Leese, 2017b)(refs), it will generally return meaningful conclusions if the target locus is well represented in a database. For COI metabarcoding the already wide availability of insect reference sequences (Andújar, Arribas, Yu, Vogler, & Emerson, 2018; Porter & Hajibabaei, 2018; Piper et al., 2019) provides a wealth of data to evaluate the specificity of the target region as well as determine the best placement of PCR primers to minimise taxon specific biases (Elbrecht & Leese, 2017a). Nevertheless, the use of public reference data comes with some caveats. Issues of mislabelled or outdated taxonomic annotations and contamination with non-homologous loci and pseudogenes in public reference databases are well documented (Bengtsson-Palme et al., 2016; Rees & Cranston, 2017; Mioduchowska, Jan, Gołdyn, Kur, & Sell, 2018), and can greatly confound any in-silico validation. It is therefore essential that DNA barcode reference sequences obtained from public repositories be appropriately curated before use in in-silico validation procedures or actual metabarcoding analysis (Piper et al., 2019).

In this article we use an automated workflow to curate a large collection of Insect COI sequences from public reference databases, then use in-silico methods to evaluate the sensitivity and specificity of published metabarcoding primers for both insect pests and insects more broadly. We locate the optimal diagnostic window within the target loci for placement of COI mini primers and evaluate the ability of published and novel primers overlapping these regions to differentiate a list of pest taxa from others in their group (sensitivity and specificity). We find that mini-barcode primers amplifying regions of COI of appropriate size for modern HTS platforms can provide comparable resolution to the full COI barcode region already widely adopted within insect pest diagnostic protocols. Finally, we provide recommendations on primers to use for diagnostics of pest insect taxa, and a workflow for assembling curated reference databases for metabarcoding studies

# Methods

*Retrieval and curation of public reference data*

Mitochondrial cytochrome c oxidase subunit 1 (COI) records as well as mitochondrial genomes with the taxonomic annotation ‘Insecta’ were retrieved from BOLD and GenBank using the bold (Chamberlain, 2017) and Rentrez (Winter, David & Winter, 2017) R packages respectively. As there is overlap in submitted sequences between both databases, sequences duplicated accession numbers were removed. Sequence annotations were then mapped into the Open Tree of life Taxonomy (OTT) (Hinchliff et al., 2015) through translation of NCBI taxonomic ids where present or string matching of species genus binomials. All taxonomic synonyms listed within the OTT synonyms list were resolved to the currently accepted name, and taxa with annotations containing terms that indicate insufficient identification or uncertain placement (i.e. sp., nr., cf., incertae sedis, not\_otu, unplaced) were removed (see Supplementary note 2 for full list of terms). All sequences were then aligned to a profile hidden markov model (PHMM) (Eddy, 1998) of the COI locus using the Viterbi algorithm and a minimum alignment score of 400 (HOLMES & DURBIN, 2009) which removed non-homologous sequences. Following alignment, all bases to the outside the bounds of the 712bp amplicon of the conventional ‘folmer’ COI barcode region where shaved from mitogenomes and other longer sequences, leaving only the target region. This reference PHMM was generated using a manually curated version of the Midori-unique dataset (Machida, Leray, Ho, & Knowlton, 2017) (Supplementary) and the aphid R package (Wilkinson, 2019). As the COI barcode is a protein coding region, underlying evolutionary constraints mean that any sequences containing stop codons or indels of lengths which are not a multiple of 3 commonly indicate pseudogenes (Roe & Sperling, 2007), and therefore any sequences with these features were removed. Next, all sequences were clustered at 97% using the Kmer R package (Wilkinson, 2018) and genus level taxonomies compared within clusters in order to identify putative taxonomic mis annotations. Any clusters containing sequences whose annotation at the genus level clashed with at least 80% of other sequences in the cluster were removed from the dataset. Finally, in order to accelerate downstream computations all large groups of taxa were pruned down to 5 representatives for each species, discarding sequences sequentially from smallest to largest.

Construction of phylogenetic tree

The reference sequences were further pruned by length down to only one representative per species and phylogenetic tree was generated using FastTree (Price, Dehal, & Arkin, 2009). As construction of deeper ancestral relationships from COI data alone is challenging (refs), the topology of the FastTree phylogeny was constrained by using the phylogenetic tree of (Chesters, 2017) as a guide tree. The tree was then made ultrametric and time scaled using congruification (Eastman, Harmon, & Tank, 2013) with the guide tree using the geiger R package (Pennell et al., 2014)

*Assembly of global pest list*

In order to assemble a globally relevant list of pest and invasive insects, all records from 11 global and geographically focused pest and invasive species databases (Supplementary note 1) were retrieved. Taxon names were filtered to retain only those taxa with complete genus species binomials and then mapped into the OTT consensus taxonomy as above. Higher taxonomic ranks for each species were then retrieved by recursion through the taxonomic tree, and any taxa from outside the class Insecta were removed from further analysis. Taxonomic intersections between the 11 different databases were visualised using the upsetR (Conway, Lex, & Gehlenborg, 2017) package and overall compositional differences were visualised using PCA.

*Identification of optimal diagnostic windows*

The curated sequence database was split by insect order, and the Shannon entropy *H* per site (Stephens, 1990) was calculated as where *fi* is the empirical frequencies of nucleotides at each position excluding gaps. The mean and percentiles of the entropy per nucleotide site within the 712bp conventional COI barcode region was visualised and structural motifs annotated as per (Pentinsaari, Salmela, Mutanen, & Roslin, 2016). To identify optimal diagnostic windows for placement of mini-barcodes, the overall entropy was divided using sliding windows of both 220bp (appropriate size for iSeq, NextSeq, HiSeq, NovaSeq 4000), and 420bp (appropriate size for MiSeq and NovaSeq SP flowcell).

*Selection and analysis of published primers*

N forward and N reversed overlapping the conventional COI barcode region were identified from a literature search, and n additional sets were designed using Primer3 implemented in Geneious Prime v2019.2 (Geneious, New Zealand). Presence within the reference sequences as well as nucleotide composition statistics were calculated for each primer using the Biostrings (Pagès, Aboyoun, Gentleman, & DebRoy, 2019) and DECIPHER R packages (Wright, 2016). Primer melting temperatures were calculated using the nearest neighbour thermodynamics (SantaLucia & Hicks, 2004) using the TmCalculator R package (Li, 2019). In addition to the original primer pairings, all possible combinations of forward and reverse primers that could produce an amplicon of at least 100bp were considered in further analysis, for a total of n combinations.

*Evaluation of primer specificity*

To evaluate the resolution of the above primer combinations for both pest taxa and the broader insect diversity, a PCR amplification was simulated by trimming the curated reference sequence database to the amplified region of each primer set. The pairwise distance between each sequence in the virtual amplicon reference sets were then generated using UCLUST (Edgar, 2010) and the number of times a pairwise comparison contained a sequence with differing taxonomic annotations was used as a metric for identification success. The relative performance of each mini-barcode set compared to that of the full length region was evaluated by taking the log ratio of identification success with each mini primer to that of the full length region. This metric was then summarised at pairwise similaritys between 95 and 100% for all insect taxa, and just those where one of the pairwise comparisons was contained in the pest list. In order to test whether barcode length is a significant predictor of congruence with the full length region, MANOVA tests were carried out in R (R Core Team, 2019) with the log ratio difference as the response variable primer as explanatory variables. The emmeans R package (ref) was then used to perform pairwise post-hoc Tukey tests between mini- and full-length barcodes in order to assess the best performing combinations.

*Evaluation of primer mismatch*

A mismatch score for each primer was calculated for each primer and sequence within the entire curated insecta dataset using the evaluate\_primer function in PrimerMiner (Elbrecht & Leese, 2017a). The default settings of “type-v1” mismatch (penalties as per Stadhouders et al., 2010) and “position-v1” (exponentially increasing mismatch penalty towards 3’ end of primer) were used to score types of mismatches, with penalty scores doubled for each contiguous mismatch. To determine how mismatch was differed across the phylogenetic tree, the phylogenetic tree generated above was subset to only those tips for which there was a matching sequence for each primer, and mismatch was modelled separately for each primer as a discrete trait on the phylogeny. Missing primer mismatch data for certain species and primer combinations were then imputed from related taxa and ancestral values in the tree inferred using phylogenetic independent contrasts (Felsenstein, 1985) within the Castor R package (Louca & Doebeli, 2018). The phylogenetic tree was then visualised alongside imputed species level primer mismatch data using ggtree (Yu, Smith, Zhu, Guan, & Lam, 2017; Yu, Lam, Zhu, & Guan, 2018). Finally, As COI is a protein coding gene, different functional constraints are imposed on different sections of the barcode due to the physical conformation of the COI within the mitochondrion membrane (Tsukihara et al., 1996; Pentinsaari et al., 2016). Therefore we also tested for positional effects by modelling mismatch score by position within the COI fragment.

*Evaluation of off-target amplification.*

While inclusion of degenerate nucleotide bases in primers will reduce primer-template mismatch and therefore increase performance for the above metric, overly high primer degeneracy can cause undesired off target amplification (Mioduchowska et al., 2018). To evaluate potential off-target effects of each primer set, the primertree R package (Cannon et al., 2016) was used to pick 1000 sequences at random from the trimmed datasets for each primer set and conduct a primerBLAST (Ye et al., 2012) against the total GenBank nucleotide database (Benson et al., 2018). in order to summarise the expected taxonomic scope of each primer, neighbour joining trees were constructed from primerBLAST hits and annotated with higher taxonomic ranks.

# Results

## Sequence database assembly

To evaluate candidate primers, 1,770,587 and 1,750,185 sequences with taxonomy "Insecta" were retrieved from GenBank and BOLD respectively, as well as a further 23,571 mitochondrial genomes from GenBank. Duplicate sequence accessions across both databases were removed leaving a total of 3,256,663 unique sequences for which the taxonomic annotations were mapped to the OTT taxonomy. During taxonomic mapping, a total of n synonyms were resolved to current accepted species names, while a further n sequences were removed due to not mapping correctly or having non-binomial species names, leaving n sequences remaining. These sequences were then filtered using a PHMM of the full length COI barcode region, which removed n sequences that did not align sufficiently. Of those that did not align sufficiently and were deemed non-homologous, when sequence labels were investigated the majority were COII or COIII. A check for presence of frameshifts or stop codons removed a further n sequences that were sufficiently homologous to COI to not be removed by the PHMM, but contained features that commonly indicate pseudogenes. Next, in order to resolve misannotated taxonomy, all sequences were clustered at 97%, flagging n putative problem clusters with mixed annotations at the genus rank. All sequences within problem clusters that clashed with at least 80% of other sequences in the the cluster were removed, leaving n sequences remaining. As the number of sequences per taxon varied greatly (Fig Sn) the dataset was pruned to a maximum of 5 sequences per taxa, discarding from smallest to largest. Hierarchical taxonomy was then retrieved for each sequence by recursion through the OTT tree. Following all filtering stages a total of n sequences and n species remained in the curated reference database with n originating from GenBank and n from BOLD (supplementary figure n ).

*Pest insect datasets*

A total of 3914 insect pest species from 2108 Genera, 329 families and 20 orders were identified from public databases of invasive and pest insect species and successfully mapped into the OTT taxonomy (Figure 1). The more generally focused ‘invasive species’ inventories DAISIE and GRIIS contained the greatest total number of taxa, with 2064 and 2063 species respectively. This was followed by pest and biosecurity focused inventories of (Ashfaq, Hebert, & Naaum, 2016), EPPO, CABI and QBank with 397, 383, 366, and 354 species respectively. Finally, the smallest overall datasets were the regionally restricted Plant Health Australia, and DAWR top40 databases, and the taxonomically restricted VectorBase, with 112, 99 and 79 species respectively. Surprisingly the GISD dataset also had a low number of insect taxa at 79 species, reflecting different curation practices to DAISIE and GRIIS. In contrast, when considering proportion of species unique to each dataset the taxonomically restricted Vectorbase had the highest proportion of unique species at 88.5%. This was followed by the large DAISIE and GRIIS with 51% and 48% unique to that database, mostly overlapping with each other. The EPPO, Qbank, Ashfaq et al, and PHA datasets had 36%, 30%, 25% and 25% unique species respectively. The CABI and DAWR40 datasets had 16% and 17% unique. Finally, the GISD had the lowest proportion of unique species with only 8%. When sequences from taxa were extracted from the curated DNA reference database, a total of n were retrieved. the majority of species had more than one sequence for them, however n species did not have any sequence at all.

*Identifying optimal diagnostic windows*

*Evaluation of primer specificity*

When the log-ratio of identification success of each mini-primer set was taken to that of the full length primer general trend of lower success with lower length was seen. Despite this general trend, it is clear that some mini-primer sets performed nearly as effectively as the primers covering the full-length region. This pattern was consistent across all similarity thresholds tested, as long as the primers were above a minimum 200bp in length (Figure 2). The slope of the regression line suggested only minimal extra information and increases in identification success would be gained by using longer COI barcodes than the full length conventional region, predicting a gain of n per additional 100bp, suggesting that saturation of identification success is achieved quite early with COI barcoding. Notably, not even the full length region could identify some taxa, with a mean success rate of only n. When identification success with the full-length folmer primers was considered as a trait on the insect phylogeny, significant phylogenetic signal was found (pagels lamda = , p<0.05). When considering local autocorrelation across the phylogeny, branches with significant signal included x, y and z (LIPA = , p<0.05) (Figure 4B).

*Primer presence & evaluation mismatch*

The number of sequences containing matching primers differed greatly with those situated toward the centre of the folmer region matching a significantly higher number of sequences than those on the edges (Figure n). This reflects the common quality control practice of trimming primer binding sites from sequences before uploading to reference databases. Due to this missing data, evaluation of primers that overlap with the primer binding sites of the folmer primers was particularly constrained, and could only occur against mitochondrial genomes, which represented a much lower proportion of the dataset (fig 2x). When mismatch was treated as a trait on the phylogenetic tree, significant signal was found (Pagels lamda = , p<0.05), and therefore phylogenetic independent contrasts was used to impute this missing data. The reliability of data imputation is dependent on the level of missing data in the first place, and therefore much lower than for primers that bound within the folmer region (supplementary n). Nevertheless, here PIC imputation provides some insight into predicted performance for taxa where no reference data is available. Following imputation, the Forward primers with the lowest mean mismatch were x, y, and z, and reverse x,y, z however these primers still showed significant mismatch to the clades of the phylogenetic tree around the families x and y (Fig 4). On the other hand the Forward primers x,y,z, and reverse primers x,y,z while not showing the lowest mean mismatch, performed best in terms of evenness of mismatch to different taxa (Gini coefficient? ). In contrast, some primers such as the forward primers x, yz, and the reverse primers x,yz performed poorly in both mean mismatch and evenness of mismatch. Level of mismatch was found to be significantly driven by both the level of primer degeneracy (, p<0.05).

*Evaluation of off-target amplification.*

Unsurprisingly many of the primers with higher degeneracy and therefore higher performance in the evaluation performed the worst for off-target amplification, reflecting the balance required in selecting degenerate bases. In particular, primers x,y,z showed a high level of off-target amplification of non-insect taxa (Figure x). On the other hand, some primers that did show low level of mismatch also showed a low level of off-target amplification. In general, those with a level of degeneracy above 1000 showed significant problems with off-target amplification.

# Discussion

Scaling up surveillance for invasive pests and pathogens to match the rate of international movement of goods presents a significant biosecurity challenge in an increasingly globalised world. Darling and Blum speculated on future existence of a technology would utilize DNA to accurately identify all species in mixed samples, allowing simultaneous screening a wide variety of invasive species. Here we demonstrate that DNA Metabarcoding fulfils this criteria and is currently in the process of revolutionising the way in which invasive insect surveillance is conducted, just as conventional DNA barcoding did.

We demonstrate that effectively designed mini-barcode primers can achieve comparable accuracy to the conventional ‘folmer’ COI barcode region, which is already widely accepted within diagnostic protocols (refs). This finding is important as it indicates that there is no need to wait for development of long read sequence technologies for implementation of metabarcoding within active surveillance programmes. Furthermore, this finding is not just limited to insect pests, and in agreement with (Yeo et al., 2020) we see similar for the broader insect taxonomy. This is important as it means that use of highly cost effective short read platforms such as the illumina NovaSeq can be used for both DNA barcoding and metabarcoding (Piper et al., 2019). When combined with non-destructive DNA extractions (refs) or DNA extraction from ethanol, this approach could greatly alleviate the taxonomic impediment for invasive species surveillance and other biomonitoring programs, as well as hasten species description (Yeo et al., 2020).

We found the following primer sets to perform well globally, and therefore recommend them for general primer sets. For primer sets appropriate for 2x150bp sequencing: and for primers appropriate for 2x250bp sequencing.

While there were certain taxonomic groups which could that could not be reliably differentiated by mini-barcodes, the majority of these resided in problem species complexes that are also unreliable for conventional DNA barcoding primers. These included the family Tephritidae, for which the full DNA barcode region also struggles with (Jiang et al., 2014; Krosch, Schutze, Strutt, Clarke, & Cameron, 2017). Those that could be identified by the conventional Folmer primers, but not mini primers included. In other cases, the inability to appropriately identify a taxon was due to issues with reference data rather than the metabarcoding primers themselves. The initial analysis was greatly confounded by taxonomic synonyms, indicating problems of taxonomy in public reference databases. While there Is some curation, there is currently no systematic correction of taxonomic annotations in public reference databases such as BOLD or GenBank. This produces particular problems for the automated hierarchical and probabilistic classifiers commonly used in metabarcoding analysis (Wang, Garrity, Tiedje, & Cole, 2007). These classifiers rely on passing a specific confidence threshold to descend to the next rank in the taxonomic hierarchy, however conflicts in taxonomic annotations due to synonyms could cause a lack of confidence to reach species level taxonomy. While this inherent conservatism is important for many applications of metabarcoding, it could cause problems for metabarcoding diagnostic where species or strain level information is required. Therefore we recommend that as well as the common filtering for non-homologous regions (Richardson, Sponsler, McMinn-Sauder, & Johnson, 2020), pseudogenes, and misannotated taxonomy (Kozlov, Zhang, Yilmaz, Glöckner, & Stamatakis, 2016), reference database curation should further involve correcting of taxonomic synonyms where possible.

However, we expect that as uptake of long-read sequencing increases for taxon identification, new primers will have to be designed to extend barcode regions (Callahan et al., 2018).

However, we show that additional length for COI will not add much. While use of long read sequencing will enable the use of new nuclear barcodes, we expect that COI will remain important well into the future due to the availability of public reference databases for this locus, and therefore perhaps full mitochondrial barcoding will be the first long read sequencing application for insect metabarcoding. Whatever new locus is chosen to take advantage of the inevitable rise in use of long-read sequencing, we have now defined a process for in-silico validation method would be applicable. Similarly, this process could be used for other taxa. Biosecurity diagnostics is becoming increasingly important for Fungi, bacteria etc.

Nevertheless, COI still has limitations in regards to bias where the degenerate third position in the codon leaves no conserved regions to place primers. Recent analyses have shown that other mitochondrial genes provide sections with comparable resolution and highly stable structures for placing primers. While it is likely that for now we are probably stuck with COI due to the availability of reference sequences available, with the dramatic fall in sequencing costs over the past decadeit is important that future efforts begin to describe larger sections of the genome, such as whole mitochondria or whole genome sequences.

*Importance of curated reference databases*

Here we considered the impacts of bias solely related to its effect on detection, however the impacts of bias must be resolved if Metabarcoding is to be used quantitatively (Piper et al., 2019).

*Integrating in-silico methods into assay validation*

Due to the non-targeted nature of HTS, performance characteristics of traditional means of validation, such as exclusivity, cannot be applied in their usual sense (Roenhorst *et al*., 2018) and should be adapted. It is for example not possible to develop and validate protocols for sample analysis of any possible combination of pest, host or sample matrix and a representative set of pests should be included in the validation. Given the broad range of targets an HTS test may detect, the validation of the HTS test should focus on the method of analysis using key representative of the targets mimicking the concentration and composition of real samples expected to be tested routinely by the HTS test.

In addition to use of in-silico methods, use of comparative phylogenetic methods could expand predictions from those species for which there is available data to those which there isn’t. Here we used these methods to expand prediction of primer mismatch from those species for which there was readily available mitochondrial genomes across the phylogeny to those for which there was no readily available sequences to test primers against. We found strong phylogenetic signal in both primer mismatch. Multiple studies have postulated that inclusion of mock communities in the form of calibration controls could be used to correct for taxon specific bias, and (McLaren, Willis, & Callahan, 2019) have proposed the possibility of a phylogenetic extension to overcome the inability to assemble mock communities for all species in the environment. Promisingly, here we found mismatch to have a high level of phylogenetic signal, suggesting that this approach could work for correcting PCR bias. Nevertheless, the bias seen in metabarcoding reflects the multiplicative contribution of many aspects of the workflow, of which PCR may not even play a majority role (Martoni et al 2020).

We also used phylogenetic comparative methods to look at the patterns of identifiability using conventional primers and thus locate the clades where DNA barcoding generally performs poorly.

Nevertheless, this in-silico validation process is designed to supplement rather than replace laboratory validation on targets. Indeed additional untested effects and complexities means it can be difficult to predict the full performance in silico (Clarke et al., 2014; Elbrecht & Leese, 2017b; Corse et al., 2019). For example, PCR can be further biased by polymerase used (Nichols et al., 2018), cycle number (Krehenwinkel et al., 2017) , GC content (Braukmann et al., 2019), annealing temperature (Elbrecht et al., 2019), and inhibitors from the matrix (Demeke & Jenkins, 2010). Furthermore, non-specific amplification can increase when target DNA Is scarce (refs). However, in-silico it allows you to be more selective of what to conduct a full laboratory validation on. Luckily, the optimal primers we have selected have been previously tested on insects and other arthropods (Elbrecht et al., 2019), which provides further confidence to our results. However, we still recommend that this process be followed up by validation to be conducted on a phylogenetically and physiologically diverse group of organisms as well as high priority targets (ref).

Nevertheless perhaps an unexplored expansion of our approach is using phylogenetic methods to expand the predictions of more conventional validation to taxa that could not be covered, and this should be explored in future and contrasted against the use of purely in-silico evaluations such as ours.

*Integration into surveillance program / Benefits of unambiguous molecular identification*

Many of the benfits of integration have been covered in Piper et al 2019

Problems fo parataxonomy missing things etc.

While individual laws may require, if instead the non-targeted assay is used as a first pass only with a positive detection followed by high specificity confirmatory testing using a complementary method, validation could instead focus on minimizing the risk of false negative results (Roenhorst et al., 2018).

Approaches such as robust in-silico validation, alongside other advances such as non-destructive DNA extraction should provide confidence in the technique.

*Conclusions*

We here demonstrate that appropriately designed mini-barcode primers provide comparable resolution to already widely adopted DNA barcode primers for insect pests, proving that DNA metabarcoding will provide a effective approach for identification of insect pests. Furthermore, these primers show good resolution across the broader insect phylogenetic tree, demonstrating an further advantage of metabarcoding for use in a larger bio surveillance concept. DNA Metabarcoding has the potential to become the default first pass surveillance tool for insect pests in order for effective response to incursions of invasive pests and the pathogens they can harbour.

# Acknowledgments

The authors wish to acknowledge Conrad Trollip and other members of the iMapPESTS project team for fruitful discussions on taxonomy and metabarcoding. We wish to acknowledge the wurundjeri people as traditional owners of the land on which this was conducted and pay respects to elders past, present and emerging. Financial support for this study was provided by Horticulture Innovation Australia (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 and Agriculture Victoria's Improved Market Access for Horticulture programme (CMI105584). A.M.P. is further supported by Australian Government Research Training Program Scholarship.

# Availability of data and materials:

Functions and a tutorial for curating and formatting COI reference databases for metabarcoding analyses and in-silico validation are provided in the ‘taxreturn’ R package, available on GitHub <https://github.com/alexpiper/taxreturn>. All further code required to reproduce the statistical analyses and generate figures in this manuscript is contained in a GitHub repository <https://github.com/alexpiper/Drosophila_metabarcoding>

# Author contributions

A.M.P conceptualised the study, performed all analyses, and wrote the manuscript with input and supervision from J.P.C, NC, and M.J.B. All authors read and reviewed the final version of the manuscript.

# References

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), 1–11. doi:10.1007/s10658-018-1570-0

Andersen, M. C., Adams, H., Hope, B., & Powell, M. (2004). Risk Assessment for Invasive Species. *Risk Analysis*, *24*(4), 787–793.

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. doi:10.1111/mec.14844

Arulandhu, A. J., Staats, M., Hagelaar, R., Voorhuijzen, M. M., Prins, T. W., Scholtens, I., … Kok, E. (2017). Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples. *GigaScience*, *6*(10), gix080. doi:10.1093/gigascience/gix080

Ashfaq, M., Hebert, P. D. N. P. D. N., & Naaum, A. (2016). DNA barcodes for bio-surveillance: Regulated and economically important arthropod plant pests. *Genome*, *59*(11), 933–945. doi:10.1139/gen-2016-0024

Augustin, S., De Kogel, W. J., Donner, P., Faccoli, M., Lees, D. C., Marini, L., … Battisti, A. (2012). A list of methods to detect arthropod quarantine pests in Europe\*. *EPPO Bulletin*, *42*(1), 93–94. doi:10.1111/j.1365-2338.2011.02523.x

Augustin, Sylvie, Boonham, N., De Kogel, W. J., Donner, P., Faccoli, M., Lees, D. C., … Battisti, A. (2012). A review of pest surveillance techniques for detecting quarantine pests in Europe. *EPPO Bulletin*, *42*(3), 515–551. doi:10.1111/epp.2600

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. *Mol Ecol Resour*, *18*(1), 32–40. doi:10.1111/1755-0998.12682

Batovska, Jana, Mee, P. T., Lynch, S. E., Sawbridge, T. I., & Rodoni, B. C. (2019). Sensitivity and specificity of metatranscriptomics as an arbovirus surveillance tool. *Scientific Reports*, *9*(1), 1–13. doi:10.1038/s41598-019-55741-3

Bebber, Daniel P, Ramotowski, M. A. T., & Gurr, S. J. (2013). Crop pests and pathogens move polewards in a warming world. *Nature Climate Change*, *3*(11), 985–988. doi:10.1038/nclimate1990

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

Bengtsson-Palme, J., Boulund, F., Edström, R., Feizi, A., Johnning, A., Jonsson, V. A., … Thorell, K. (2016). Strategies to improve usability and preserve accuracy in biological sequence databases. *Proteomics*, *16*(18), 2454–2460. doi:10.1002/pmic.201600034

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K. D., & Sayers, E. W. (2018). GenBank. *Nucleic Acids Research*, *46*(D1), D41–D47. doi:10.1093/nar/gkx1094

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. doi:10.1016/j.marpolbul.2010.09.014

Blackburn, T. M., Essl, F., Evans, T., Hulme, P. E., Jeschke, J. M., Kühn, I., … Bacher, S. (2014). A Unified Classification of Alien Species Based on the Magnitude of their Environmental Impacts. *PLoS Biology*, *12*(5), e1001850. doi:10.1371/journal.pbio.1001850

Blauwkamp, T. A., Thair, S., Rosen, M. J., Blair, L., Lindner, M. S., Vilfan, I. D., … Yang, S. (2019). Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease. *Nature Microbiology*, *4*(4), 663–674. doi:10.1038/s41564-018-0349-6

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*. doi: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. doi:10.1017/S0007485315000681

Braukmann, T. W. A., Ivanova, N. V., Prosser, S. W. J., Elbrecht, V., Steinke, D., Ratnasingham, S., … Hebert, P. D. N. (2019). Metabarcoding a Diverse Arthropod Mock Community. *Molecular Ecology Resources*, *19*, 711–727. doi:10.1111/1755-0998.13008

Brown, E. A., Chain, F. J. J., Zhan, A., MacIsaac, H. J., & Cristescu, 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. doi:10.1111/ddi.12465

Bulman, S. R., McDougal, R. L., Hill, K., & Lear, G. (2018). Opportunities and limitations for DNA metabarcoding in Australasian plant-pathogen biosecurity. *Australasian Plant Pathology*, *47*(5), 467–474. doi:10.1007/s13313-018-0579-3

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

Callahan, B. J., Wong, J., Heiner, C., Oh, S., Theriot, C. M., Gulati, A. S., … Dougherty, M. K. (2018). High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *BioRxiv*, 392332. doi:10.1101/392332

Cannon, M. V., Hester, J., Shalkhauser, A., Chan, E. R., Logue, K., Small, S. T., & Serre, D. (2016). In silico assessment of primers for eDNA studies using PrimerTree and application to characterize the biodiversity surrounding the Cuyahoga River. *Scientific Reports*, *6*(December 2015), 1–11. doi:10.1038/srep22908

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

Chesters, D. (2017). Construction of a species-level tree of life for the insects and utility in taxonomic profiling. *Systematic Biology*, *66*(3), 426–439. doi:10.1093/sysbio/syw099

Clarke, L. J., Soubrier, J., Weyrich, L. S., & Cooper, A. (2014). Environmental metabarcodes for insects: In silico PCR reveals potential for taxonomic bias. *Molecular Ecology Resources*, *14*(6), 1160–1170. doi:10.1111/1755-0998.12265

Collins, R. A., Bakker, J., Wangensteen, O. S., Soto, A. Z., Corrigan, L., Sims, D. W., … Mariani, S. (2019). Non‐specific amplification compromises environmental DNA metabarcoding with COI. *Methods in Ecology and Evolution*, *2019*(July), 1–17. doi: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. doi:10.1007/s10530-015-0854-y

Conway, J. R., Lex, A., & Gehlenborg, N. (2017). UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics*, *33*(18), 2938–2940. doi:10.1093/bioinformatics/btx364

Corse, E., Tougard, C., Archambaud-Suard, G., Agnèse, J. F., Messu Mandeng, F. D., Bilong Bilong, C. F., … Dubut, V. (2019). One-locus-several-primers: A strategy to improve the taxonomic and haplotypic coverage in diet metabarcoding studies. *Ecology and Evolution*, *9*(8), 4603–4620. doi:10.1002/ece3.5063

Darling, J. A., & Blum, M. J. (2007). DNA-based methods for monitoring invasive species: A review and prospectus. *Biological Invasions*, *9*(7), 751–765. doi:10.1007/s10530-006-9079-4

Davidovitch, L., Stoklosa, R., Majer, J., Nietrzeba, A., Whittle, P., Mengersen, K., & Ben-Haim, Y. (2009). Info-gap theory and robust design of surveillance for invasive species: The case study of Barrow Island. *Journal of Environmental Management*, *90*(8), 2785–2793. doi:10.1016/j.jenvman.2009.03.011

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*, 20140562. doi:10.1098/rsbl.2014.0562

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., … Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, *26*(21), 5872–5895. doi:10.1111/mec.14350

Demeke, T., & Jenkins, G. R. (2010). Influence of DNA extraction methods, PCR inhibitors and quantification methods on real-time PCR assay of biotechnology-derived traits. *Analytical and Bioanalytical Chemistry*, *396*(6), 1977–1990. doi:10.1007/s00216-009-3150-9

Eastman, J. M., Harmon, L. J., & Tank, D. C. (2013). Congruification: Support for time scaling large phylogenetic trees. *Methods in Ecology and Evolution*, *4*(7), 688–691. doi:10.1111/2041-210X.12051

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*. doi:10.1093/bioinformatics/14.9.755

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460–2461. doi:10.1093/bioinformatics/btq461

Elbrecht, V., Braukmann, T. W. A., Ivanova, N. V, Prosser, S. W. J., Hajibabaei, M., Wright, M., … Steinke, D. (2019). Validation of COI metabarcoding primers for terrestrial arthropods. *PeerJ Preprints*, *7*, e27801v1. doi:10.7287/peerj.preprints.27801v1

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. doi: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. doi: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. doi:10.3389/fenvs.2017.00011

Elbrecht, V., Peinert, B., & Leese, F. (2017). Sorting things out: Assessing effects of unequal specimen biomass on DNA metabarcoding. *Ecology and Evolution*, *7*(17), 6918–6926. doi:10.1002/ece3.3192

Elbrect, V., Bourlat, S. J., Lindner, A., Noll, N. W., Sorg, M., Zizka, V. M. A., & Alexander, M. (2020). Pooling size sorted malaise trap fractions to maximise taxon recovery with metabarcoding Abstract :

EPPO. (2019). EPPO Standards - Diagnostics. *EPPO Bulletin*, *49*(2), 170–174. doi:10.1111/epp.12588

European and Mediterranean Plant Protection Organization. (2010). PM 7/98 (2) Specific requirements for laboratories preparing accreditation for a plant pest diagnostic activity. *EPPO Bulletin*, *44*, 117–147. doi:10.1111/j.1365-2338.2009.02347.x

Felsenstein, J. (1985). Phylogenies and the comparative method. *American Naturalist*. doi:10.1086/284325

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. doi:10.1371/journal.pone.0013102

Handelsman, J. (2004). Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews*, *68*(4), 669–685. doi:10.1128/MBR.68.4.669

Hinchliff, C. E., Smith, S. A., Allman, J. F., Burleigh, J. G., Chaudhary, R., Coghill, L. M., … 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. doi:10.1073/pnas.1423041112

HOLMES, I., & DURBIN, R. (2009). Dynamic Programming Alignment Accuracy. *Journal of Computational Biology*. doi:10.1089/cmb.1998.5.493

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. doi:10.1111/j.1365-2664.2008.01600.x

Jiang, F., Jin, Q., Liang, L., Zhang, A. B., & Li, Z. H. (2014). Existence of species complex largely reduced barcoding success for invasive species of Tephritidae: A case study in Bactrocera spp. *Molecular Ecology Resources*, *14*(6), 1114–1128. doi:10.1111/1755-0998.12259

Kelly, R. P., Shelton, A. O., & Gallego, R. (2019). Understanding PCR Processes to Draw Meaningful Conclusions from Environmental DNA Studies. *BioRxiv*, 660530. doi:10.1101/660530

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. doi:10.1093/nar/gkw396

Krehenwinkel, H., Wolf, M., Lim, J. Y., Rominger, A. J., Simison, W. B., & Gillespie, R. G. (2017). Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Scientific Reports*, *7*, 17668. doi:10.1038/s41598-017-17333-x

Krosch, M. N., Schutze, M. K., Strutt, F., Clarke, A. R., & Cameron, S. L. (2017). A transcriptome-based analytical workflow for identifying loci for species diagnosis: A case study with Bactrocera fruit flies (Diptera: Tephritidae). *Austral Entomology*, *58*, 395– 408. doi:10.1111/aen.12321

Leray, M., & Knowlton, N. (2017). Random sampling causes the low reproducibility of rare eukaryotic OTUs in Illumina COI metabarcoding. *PeerJ*, *5*, e3006. doi:10.7717/peerj.3006

Li, J. (2019). TmCalculator: Melting Temperature of Nucleic Acid Sequences. Retrieved from https://cran.r-project.org/package=TmCalculator

Liebhold, A. M., Berec, L., Brockerhoff, E. G., Epanchin-Niell, R. S., Hastings, A., Herms, D. A., … Yamanaka, T. (2016). Eradication of Invading Insect Populations: From Concepts to Applications. *Annual Review of Entomology*, *61*(1), 335–352. doi:10.1146/annurev-ento-010715-023809

Louca, S., & Doebeli, M. (2018). Efficient comparative phylogenetics on large trees. *Bioinformatics*, *34*(6), 1053–1055. doi:10.1093/bioinformatics/btx701

Macher, J. N., Vivancos, A., Piggott, J. J., Centeno, F. C., Matthaei, 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. doi: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. doi:10.1038/sdata.2017.27

Maree, H. J., Fox, A., Al Rwahnih, 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*(August), 1–4. doi:10.3389/fpls.2018.01082

Massart, S., Olmos, A., Jijakli, H., & Candresse, T. (2014). Current impact and future directions of high throughput sequencing in plant virus diagnostics. *Virus Research*, *188*, 90–96. doi:10.1016/j.virusres.2014.03.029

McLaren, M. R., Willis, A. D., & Callahan, B. J. (2019). Consistent and correctable bias in metagenomic sequencing measurements. *BioRxiv*, 559831. doi:10.1101/559831

Mioduchowska, M., Jan, M., Gołdyn, B., Kur, J., & Sell, J. (2018). Instances of erroneous DNA barcoding of metazoan invertebrates: Are universal cox1 gene primers too “universal”? *Plos One*, *13*(6), e0199609. doi:10.1371/journal.pone.0199609

Nichols, R. V, Vollmers, C., Newsom, L. A., Wang, Y., Heintzman, P. D., Leighton, M., … Shapiro, B. (2018). Minimizing polymerase biases in metabarcoding. *Molecular Ecology Resources*, *18*, 927– 939. doi:10.1111/1755-0998.12895

Pagès, H., Aboyoun, P., Gentleman, R., & DebRoy, S. (2019). Biostrings: Efficient manipulation of biological strings.

Pennell, M. W., Eastman, J. M., Slater, G. J., Brown, J. W., Uyeda, J. C., Fitzjohn, R. G., … Harmon, L. J. (2014). Geiger v2.0: An expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics*, *30*(15), 2216–2218. doi:10.1093/bioinformatics/btu181

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*(October), 1–12. doi:10.1038/srep35275

Piñol, J., Mir, G., Gomez-Polo, P., & Agustí, N. (2015). Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Molecular Ecology Resources*, *15*(4), 819–830. doi:10.1111/1755-0998.12355

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), 1–22. doi:10.1093/gigascience/giz092

Porter, T. M., & Hajibabaei, M. (2018). Over 2.5 million sequences in GenBank and Growing. *Plos ONE*, *13*(9), e0200177. doi:https://doi.org/ 10.1371/journal.pone.0200177

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. doi:10.1093/molbev/msp077

R Core Team. (2019). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from https://www.r-project.org/

Rees, J. A., & Cranston, K. (2017). Automated assembly of a reference taxonomy for phylogenetic data synthesis. *Biodiversity Data Journal*, *5*(1). doi:10.3897/BDJ.5.e12581

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. doi:10.1111/2041-210X.13314

Roe, A. D., & Sperling, F. A. H. (2007). Patterns of evolution of mitochondrial cytochrome c oxidase I and II DNA and implications for DNA barcoding. *Molecular Phylogenetics and Evolution*, *44*(1), 325–345. doi:10.1016/j.ympev.2006.12.005

Roenhorst, J. W., de Krom, C., Fox, A., Mehle, N., Ravnikar, 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. doi:10.1111/epp.12445

SantaLucia, J., & Hicks, D. (2004). The thermodynamics of DNA structural motifs. *Annual Review of Biophysics and Biomolecular Structure*, *33*, 415–440. doi:10.1146/annurev.biophys.32.110601.141800

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. doi:10.1023/B:BINV.0000005567.58234.b9

Scott, J. K., McKirdy, S. J., Merwe, J. Van Der, Green, R., Burbidge, A. A., Pickles, G., … Mengersen, K. (2017). Zero-tolerance biosecurity protects high-conservation-value island nature reserve. *Scientific Reports*, *7*(1), 1–9. doi:10.1038/s41598-017-00450-y

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. doi:10.1017/S1068280500010005

Simmons, M., Tucker, A., Chadderton, W. L., Jerde, C. L., Mahon, A. R., & Taylor, E. (2016). Active and passive environmental DNA surveillance of aquatic invasive species. *Canadian Journal of Fisheries and Aquatic Sciences*, *73*(1), 76–83. doi:10.1139/cjfas-2015-0262

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. doi:10.2353/jmoldx.2010.090035

Stephens, R. M. (1990). Sequence logos:, *18*(20), 6097–6100.

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. doi:10.1111/j.1365-294X.2012.05470.x

Trebitz, A. S., Hoffman, J. C., Darling, J. A., Pilgrim, E. M., Kelly, J. R., Brown, E. A., … Schardt, J. C. (2017). Early detection monitoring for aquatic non-indigenous species: Optimizing surveillance, incorporating advanced technologies, and identifying research needs. *Journal of Environmental Management*, *202*, 299–310. doi:10.1016/j.jenvman.2017.07.045

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, *73*(16), 5261–5267. doi:10.1128/AEM.00062-07

Westfall, K. M., Therriault, T. W., & Abbott, C. L. (2020). A new approach to molecular biosurveillance of invasive species using DNA metabarcoding. *Global Change Biology*, *26*(2), 1012–1022. doi:10.1111/gcb.14886

Wilkinson, S. (2018). kmer: an R package for fast alignment-free clustering of biological sequences. doi:10.5281/zenodo.1227690

Wilkinson, S. (2019). aphid: an R package for analysis with profile hidden Markov models. *Bioinformatics*. doi:10.1093/bioinformatics/btz159

Williamson, S. J. (2011). Viral Metagenomics. *Handbook of Molecular Microbial Ecology II: Metagenomics in Different Habitats*, *3*(June), 3–13. doi:10.1002/9781118010549.ch2

Winter, David, J., & Winter, D. J. (2017). rentrez: An R package for the NCBI eUtils API. *The R Journal*, *9*(2), 520–526. doi: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.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., & Madden, T. L. (2012). Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. doi:10.1186/1471-2105-13-134

Yeo, D., Srivathsan, A., & Meier, R. (2020). Longer is Not Always Better: Optimizing Barcode Length for Large-Scale Species Discovery and Identification. *Systematic Biology*, 1–78. doi:10.1093/sysbio/syaa014

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. doi: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. doi:10.1111/2041-210X.12628

# Tables

Table 1: Published and novel primers evaluated in this study

**Figures**

Figure n– Summary of data sources, and taxonomic orders of all species on public pest and invasive insect datasets

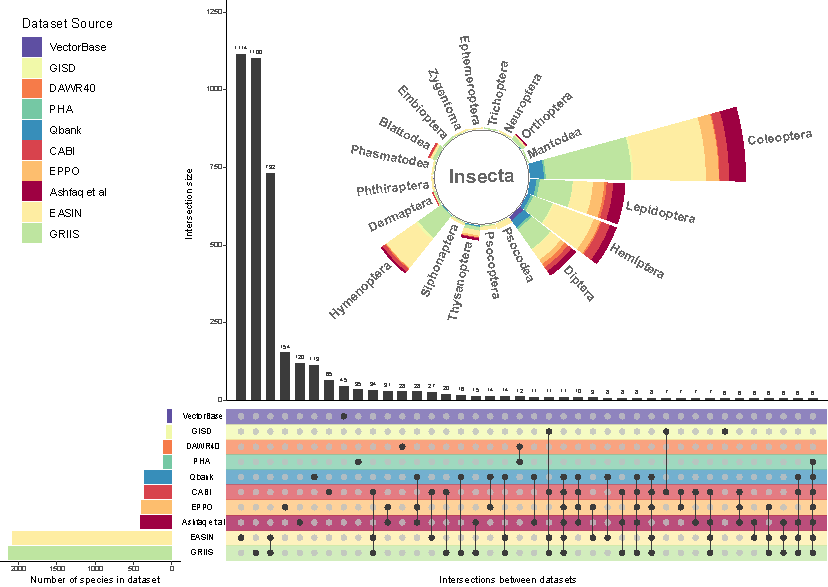
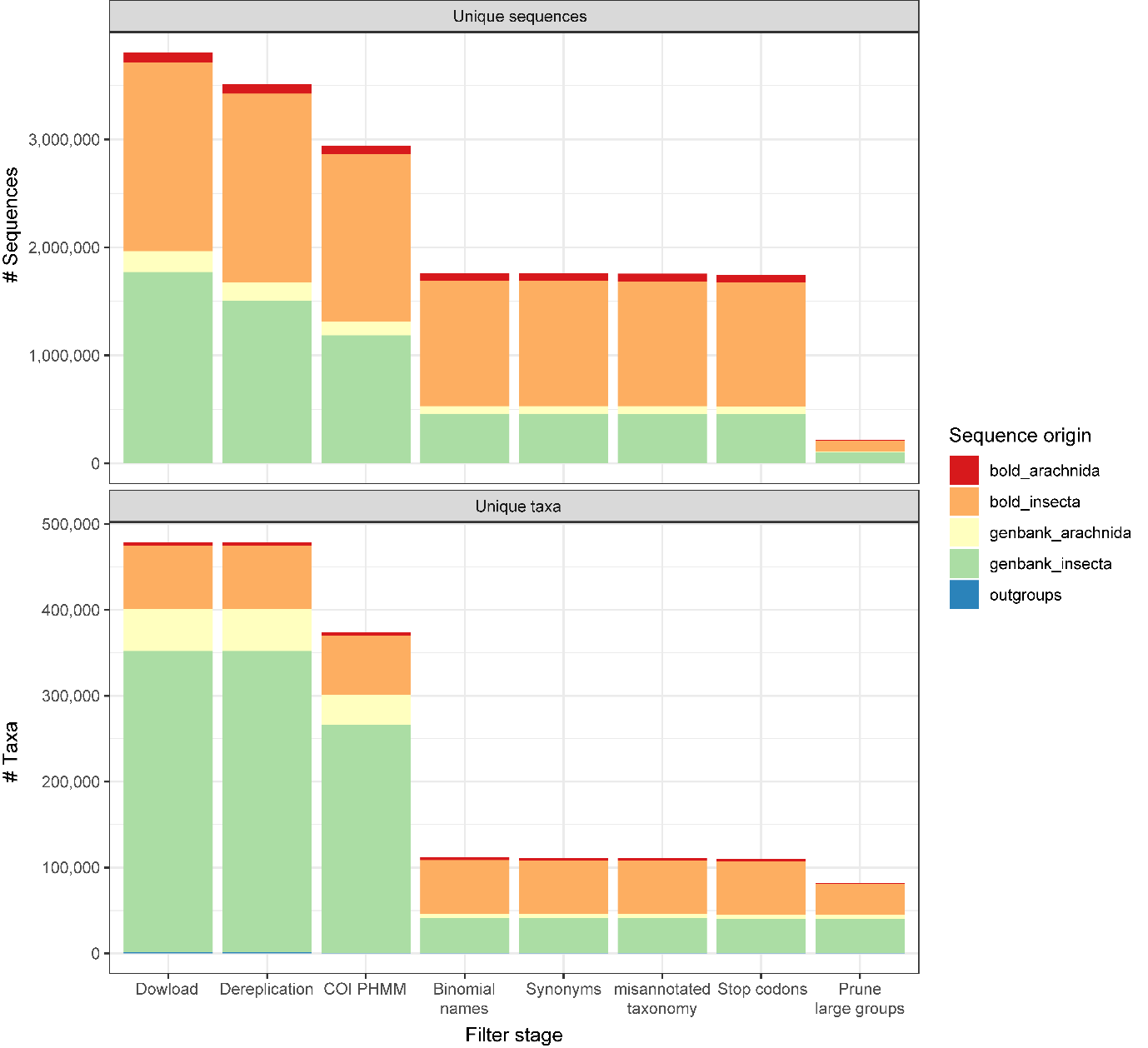
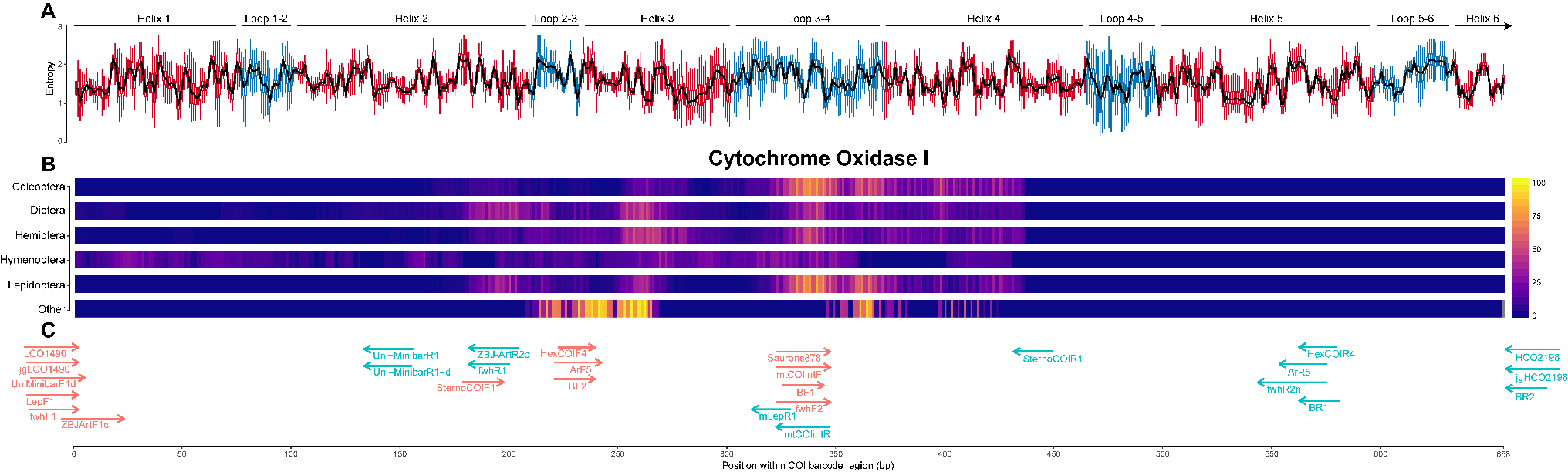
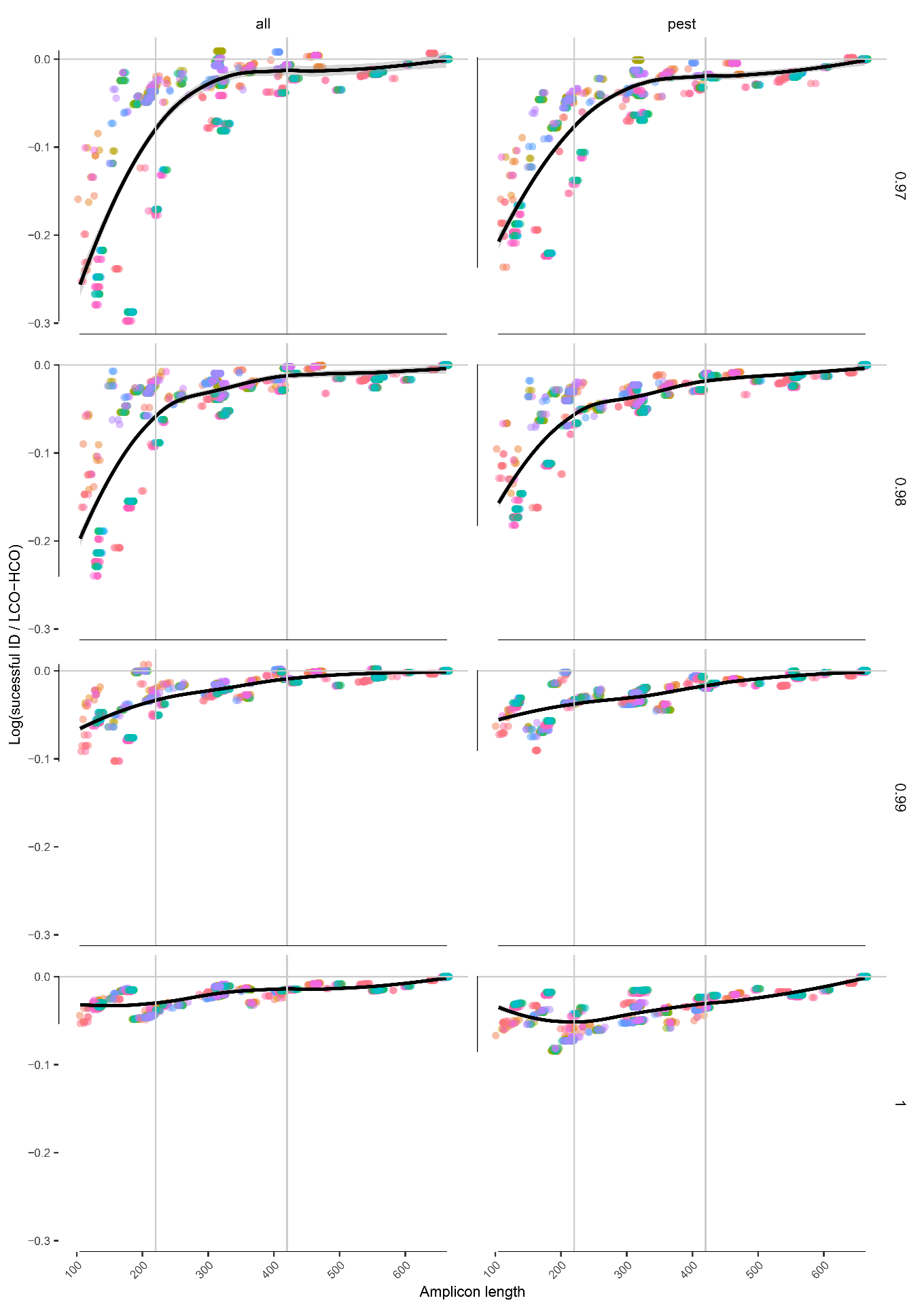
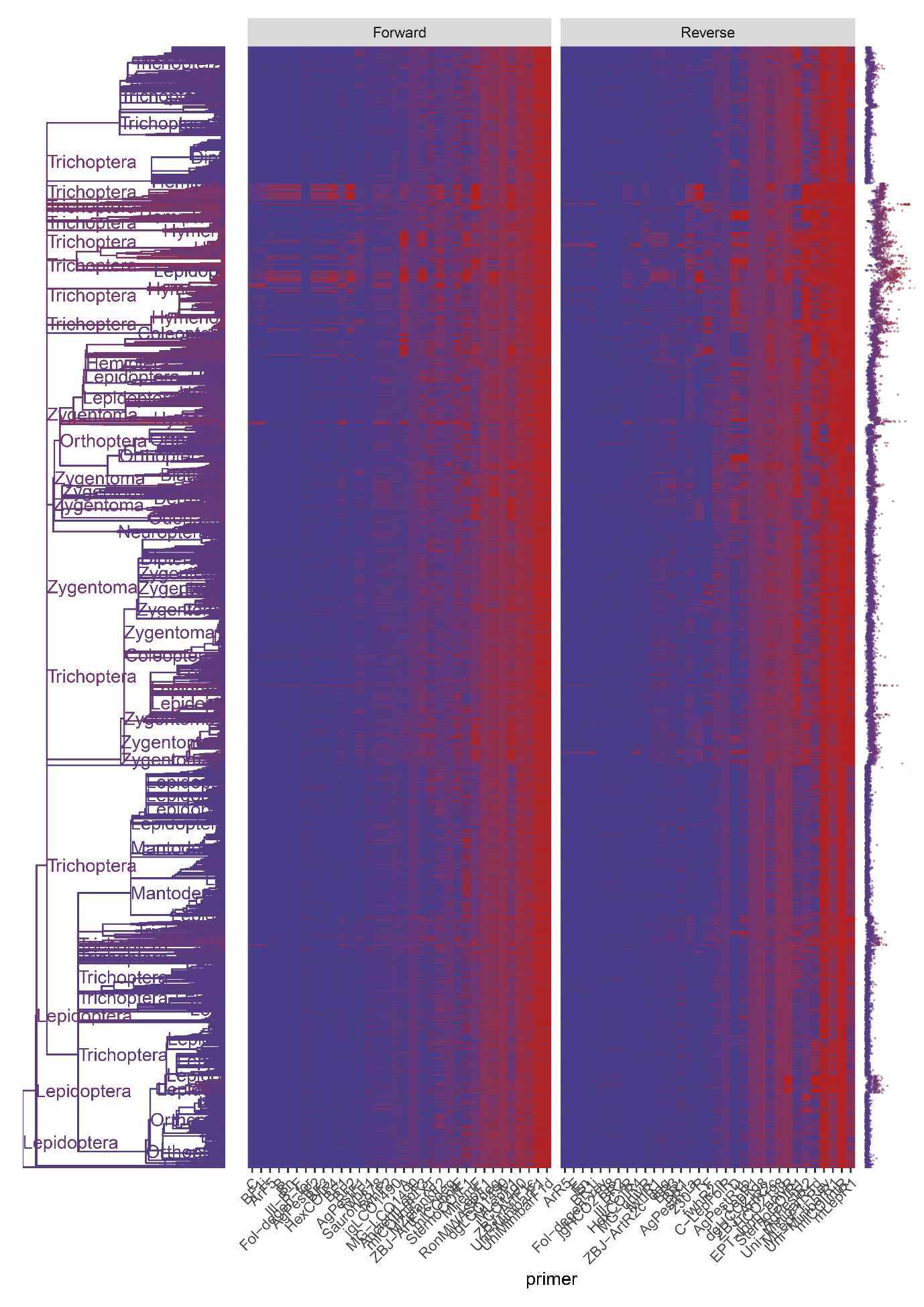


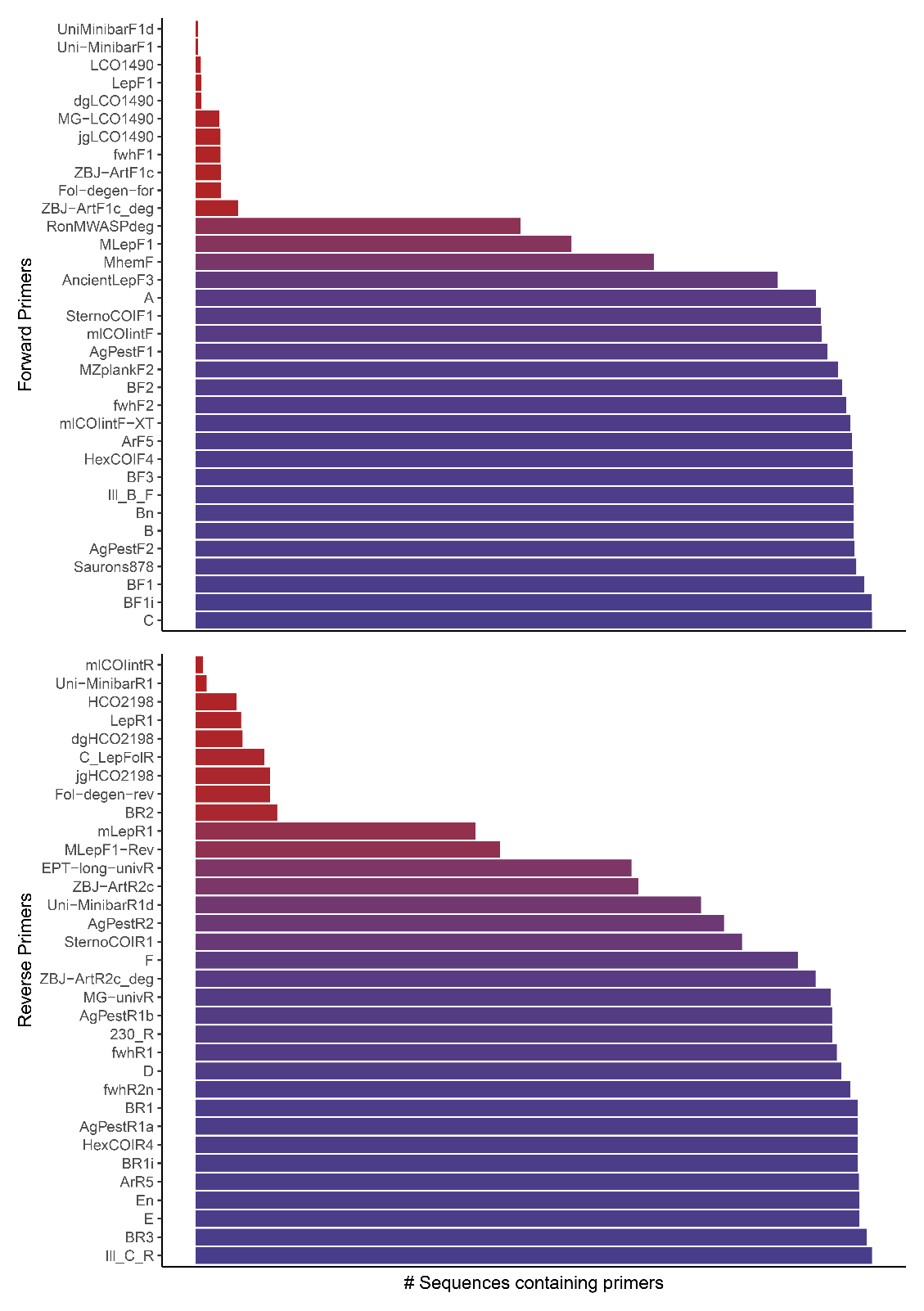
Figure n : Summary of sequences retained throughout reference sequence cleaning steps, and their databse origins.

Figure n - Summary of cytochrome oxidase 1 entropy, highest ranked diagnostic windows, and placement of all primers tested

Figure n - Evaluation of novel and existing primers for diagnostic ability across insect families, and pest insects only.

Figure n - Evaluation of individual forward and reverse primers for mismatch against all insect sequences

**Supplementary Figures**



Proportion of sequences in which each individual primer was detected

**Supplementary note 1: Sources of taxonomic data**

* EPPO global database https://gd.eppo.int/
* US APHIS - https://www.aphis.usda.gov/aphis/home/
* QBank - https://qbank.eppo.int/arthropods/organisms
* Global invasive species database - http://www.iucngisd.org/gisd/search.php
* Global register of introduced or invasive species http://www.griis.org/
* VectorBase: https://www.vectorbase.org/organisms
* DAWR top 40 - http://www.agriculture.gov.au/pests-diseases-weeds/plant
* PHA National biosecurity status report - http://www.planthealthaustralia.com.au/national-programs/national-plant-biosecurity-status-report/
* Ashfaq & Herbert 2016 - DNA barcodes for bio-surveillance: regulated and economically important arthropod plant pests
* CABI - https://t.co/LGjlFoOazd
* http://www.europe-aliens.org

**Supplementary note 2: OTT Flags removed during sequence filtering**

* incertae\_sedis
* major\_rank\_conflict
* unplaced
* environmental
* inconsistent
* extinct
* hidden
* hybrid
* not\_otu
* viral
* barren

**Supplementary note 2: Sequence terms removed during sequence filtering**

* sp.
* spp.
* aff.
* nr.
* bv.
* cf.
* nom
* nud
* environment
* undescribed
* unverified
* unclassified
* uncultured
* unidentified
* [0-9] (all numeric)
* [:punct:] (Punctuation and symbols)