**In-silico validation of a non-targeted metabarcoding assay for insect pest surveillance.**

Alexander M. Piper1,2, 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:**

* Molecular Ecology Resources
* Methods in Ecology and Evolution
* Scientific Reports
* PeerJ

# Abstract

Appropriate design and selection of PCR primers play the most important role in determining the sensitivity and specificity of a metabarcoding assay. Despite several studies applying metabarcoding to detection of insect pests, and the possibility of using a universal molecular diagnostic assay for insect pests has been raised. The diagnostic ability of the smaller barcode regions for insect pests has not yet been systematically evaluated. With this in mind, here we use in-silico methods to test the hypothesis that DNA metabarcoding could act as a ‘universal’ diagnostic for insect pests. Firsly we curate a large reference database of insect sequences. We then evaluate the diagnostic ability of a range of COI primers on this dataset, as well as just insect pests. Next we evaluate the predicted bias for each primer across insecta, as well as predicted off-target amplification for all taxa in genbank. We provide recommendations for primers of suitable size for various short read sequencing platforms, and a list of taxa which could be successfully identified by these primers. Many taxa that could not be identified with the primers come from classic problem groups that also could not be identified by conventional DNA barcoding primers. Further were a problem of taxonomy. Therefore, 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.

# 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 on biodiversity and ecosystem services where the size and complexity of the natural environment means it is often difficult to assess the impact of a species in advance (Caley, Lonsdale, & Pheloung, 2006; Blackburn et al., 2014). In this rapidly changing biosecurity environment, it is obvious that a more holistic approach to surveillance that aims to detect and evaluate all pests and diseases, not just those that are regulated 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). Nevertheless, ongoing broad-spectrum surveillance 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 (i.e. species or strain) will form a major diagnostic bottleneck (Davidovitch et al., 2009; Westfall, Therriault, & Abbott, 2020).

Currently, plant pest and pathogen diagnostics rely on a mixture of morphological examination, biochemical techniques and both targeted and universal molecular approaches such as diagnostic qPCR, and DNA barcoding (EPPO, 2019). While these methods can be rapid and highly accurate for small numbers of specimens, the single-specimen nature of these methods restrict 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, recent high-throughput sequencing (HTS) platforms can comprehensively characterise mixed populations of genomic DNA (metagenomics), RNA (metatranscriptomics) or taxonomically informative marker genes (metabarcoding), allowing simultaneous identification of all organisms in a complex mixture (Allcock, Jennison, & Warrilow, 2017). 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 (E. A. Brown, Chain, Zhan, MacIsaac, & Cristescu, 2016; Arulandhu et al., 2017; Blauwkamp et al., 2019; Jana Batovska, Mee, Lynch, Sawbridge, & Rodoni, 2019). 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 (Darling & Blum, 2007; Massart, Olmos, Jijakli, & Candresse, 2014; Adams, Fox, Boonham, Massart, & De Jonghe, 2018). This raises the question, could a non-targetted HTS assay be used as a universal diagnostic for insect pests?

High-sensitivity with low specificity is a defining feature of non-targetted HTS assays, which lends themselves to use more akin to surveillance tool to rapidly detect a range of taxa. Due to these characteristics, it is common that when incorporated into a surveillance programme they reveal 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 beneficial for a broad-spectrum surveillance programme, appropriate interpretation and response to these incidental detections has been identified as one of the major challenges for adoption (Massart et al., 2017; Adams et al., 2018; Piper et al., 2019). The complexity of laboratory and bioinformatic pipelines, alongside the reliance on large reference databases introduce new methods of false positives being introduced (Zinger et al., 2019). Historically, molecular diagnostic assays would have undergone a stringent validation process in order to 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 pathogens/pests, hosts and matrices that would need to be evaluated, it is evident that a full validation for non-targeted diagnostics is unfeasible (Maree, Fox, Al Rwahnih, Boonham, & Candresse, 2018; Roenhorst et al., 2018).

However, 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). Consequently, under this systems-approach to diagnostic quality assurance, the incorporation of in-silico methods into HTS assay validation procedures has become increasingly important (Schlaberg, Chiu, Miller, Procop, & Weinstock, 2017; Blauwkamp et al., 2019). Because a large diversity of specimens can be generated in silico and critical data analysis steps can be performed with fewer resources, in-silico methods allow defining performance criteria such as specificity, sensitivity, and even if physical specimens of the target are unavailable (refs). Furthermore, in-silico methods allow in-depth testing of the robustness of algorithms and sequence databases and simulation of scenarios that are difficult for lab validation such as covering intraspecific diversity of target taxa (refs).

For HTS based diagnostics of insect pests, metabarcoding of the Mitochondrial cytochrome c oxidase subunit 1 (COI) locus provides the most cost effective approach due to large nuclear genome sizes and wide availability of reference data for this locus (Piper et al., 2019). While the 658-bp ‘folmer’ region of COI (Folmer, Black, Hoeh, Lutz, & Vrijenhoek, 1994) has been widely employed for single species diagnostics of insect pests via conventional DNA barcoding (Ashfaq, Hebert, & Naaum, 2016), modern HTS platforms impose strict limitations on molecule length that can be sequenced and therefore smaller stretches of the conventional barcode region or ‘mini-barcodes’ must instead be used (Brandon-Mong et al., 2015). Furthermore, unlike single specimen diagnostics, metabarcoding PCR bias can cause deviation from expected proportions, when bias is too high it can swamp out low abundance taxa causing false negatives. As COI is a protein-coding gene, the third position of codons can be variable, leaving no strictly conserved nucleotide sites for design of universal PCR primers, and therefore care should be taken to ensure they match the target groups appropriately. Ensuring the accuracy of detections is paramount as both false positive or negative detections of pest species can lead to severe environmental and economic consequences, 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 evaluated.

Here we provide a framework for in-silico evaluation of non-targeted metabarcoding assays for regulated organisms, using a diverse global list of invasive and pest insects as our case study. However, due to commonly occurring issues with public sequence reference databases, such as non-homologous loci, mislabelled taxonomy, pseudogenes, and taxonomic synonyms (Piper et al., 2019), it is essential that datasets be reliably curated before use in validation or metabarcoding analysis. Therefore firstly, we curate a large collection of reference sequences from public databases. Following curation. We locate the optimal diagnostic window within the target loci for placement of 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 then look at predicted mismatch and bias for different insect pest groups. Finally we provide recommendations on primers to use for diagnostics of pest insect taxa, and provide a comprehensive list of pest taxa for which a metabarcoding assay using these primers are likely to work and not, as well as a curated insect reference database for future studies. We believe that the most rapid way to get this technology into the hands of diagnostics labs is to propose a protocol that is as universal as possible, then the user validates the assay on their target pest.

# Methods

*Assembly of global pest list*

To assemble a global list of insect pests, all records from a range of global invasive species databases (Supplementary note 1) were retrieved and filtered to retain only those taxa with genus species binomials and remove duplicate entries. Species names were mapped to the Open Tree of life Taxonomy (OTT) (Hinchliff et al., 2015) and all species that could not be mapped were removed. Higher taxonomic were then retrieved by recursion through the taxonomic tree, and any taxa from outside the class Insecta were removed.

*Retrieval and curation of public reference data*

In order to evaluate the sensitivity and specificity of primers for the above pestlist as well as a wider range of insects, mitochondrial cytochrome c oxidase subunit 1 (COI) records and mitochondrial genomes with the taxonomic annotation ‘Insecta’ were retrieved from BOLD and GenBank using the bold (Chamberlain, 2017) and Rentrez (Winter, David, 2019) R packages. Following, download all sequences with duplicated accession numbers were removed and species names were mapped to the OTT taxonomy with all synonyms resolved to the currently accepted name. Taxa with annotations containing terms that indicate insufficient identification (i.e. sp., nr., cf.), or those mapped to brances in the OTT taxonomy that flagged with uncertain placement (i.e. incertae sedis, not\_otu, unplaced) – see Supplementary note 2 for full list of terms) were removed. Following resolution of taxonomy, remaining sequences went through a series of filtering stages to remove non-homologous or taxonomically misannotated sequences. Firstly, all sequences were 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), in order to remove non-homologous sequences and extract the 712bp amplicon of the conventional ‘folmer’ COI barcode region from mitogenomes. This PHMM was generated using a manually curated version of the Midori-unique dataset (Machida, Leray, Ho, & Knowlton, 2017) (Supplementary) and the aphid R package (S. Wilkinson, 2019). As COI is a protein coding region, evolutionary constraints mean that any sequences containing stop codons or frameshifts not in multiples of 3 commonly indicate pseudogenes (Roe & Sperling, 2007), and therefore any sequences with these features were removed. Next, in order to filter taxonomically mislabelled sequences that commonly plague public sequence databases (Bengtsson-Palme et al., 2016), all sequences were clustered at 97% and genus level taxonomies compared within clusters. Any clusters containing sequences with taxonomy that clashed with at least 80% of other sequences in the cluster were removed. Finally, to speed up computation and reduce reference bias due to incomplete sampling of the taxonomic tree, all large groups of taxa were pruned down to 5 representatives for each species, discarding sequences sequentially from smallest to largest.

*Identification of optimal diagnostic windows*

To identify optimal diagnostic windows for placement of mini-barcodes within the 712bp COI barcode region, the curated sequence database was split into every family containing pests, and the number of diagnostic nucleotides was summarised in sliding windows of both 220bp (appropriate size for NextSeq, HiSeq, NovaSeq), and 420bp (appropriate size for MiSeq, NovaSeq SP) using the SPIDER r package (S. D. J. Brown et al., 2012). A density plot of the 100 highest ranked window positions within each family was then used to identify globally optimal regions for primer placement for of all pest insect families on the list. Published primer sets overlapping these regions were retrieved, and several additional sets were designed using Primer3 implemented in Geneious Prime v2019.2 (Geneious, New Zealand).

*Evaluation of primer specificity*

To evaluate the resolution of published primer and novel primer sets across pest taxa, sequence alignments for each pest family were trimmed to the amplified region of each primer set using the virtualPCR function in the insect R package (S. P. Wilkinson, Davy, Bunce, & Stat, 2018), and pairwise genetic distances were generated using the raw model in SPIDER. In order to summarise the taxonomic groups likely to be misidentified or ambiguously identified when using each mini-barcode primer set, summary statistics of identification success were generated using the nearest neighbour, and reciprocal monophyly identification methods in SPIDER. All groups failing were summarised by the types of failures (insufficient resolution, synonyms etc).

*Evaluation of primer mismatch*

In addition to the number of diagnostic nucleotides between species, mismatched nucleotides between the primers and template DNA can also influence detection success and accurate quantification by biasing amplification towards better matching targets (Piñol, Mir, Gomez-Polo, & Agustí, 2015; Piñol, Senar, & Symondson, 2019). To evaluate each primer set for mismatching nucleotides, the primers were aligned against the entire Insecta dataset and a mismatch score was calculated for each primer and sequence using the evaluate\_primer function in PrimerMiner (Elbrecht & Leese, 2017a) using the default “type-v1” mismatches (penalties as per Stadhouders et al., 2010) and “position-v1” settings (exponentially increasing mismatch penalty towards 3’ end of primer), with penalty scores doubled for each contiguous mismatch.

*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, Jan, Gołdyn, Kur, & Sell, 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

*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 et al., 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 databasets, and the taxonomicly restricted vectorbase, with 112, 99 and 79 species respectively. . Supprisingly the GISD dataset also had a low number of taxa at 79 species, reflecting different curation practices to DAISIE and GRIIS.

In contrast, when considering proportion of species unique to that database the taxonomically restricted Vectorbase had the highest proportion of unique species at 88.5%. This was followed by the large DAISIE adnd 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%.

## 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. When duplicate sequences that were present across both databases were removed, a total of 3,256,663 unique sequences remained, 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. After mapping to the OTT taxonomy a total of n sequences remained, which were then filtered with a PHMM of COI in order to remove non-homologous regions, retaining a total of n sequences that aligned sufficiently. A further 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 problem groups with mixed genus annotations, which were resolved by removing all sequences that clashed with at least 80% of other sequences in the group. As numbers of sequences per species varied from singletons to n sequences, in order to improve computation, all large groups were then pruned down to 5 sequences, discarding from smallest to largest. Following all stages of filtering, a total of n sequences and n sequences were retained, with n originating from GenBank and n from BOLD. See (supplementary figure n ) for a summary of numbers of unique sequences, unique species, and length distributions throughout the filtering process. Finally, the dataset was combined with Arachnida and other non-Arthropod outgroups and reformatted to hierarchical taxonomy by recursively getting nodes in the tree. The final curated insecta dataset has been output In formats suitable for a range of taxonomic classifiers for further use in metabarcoding analysis (supplementary data ).

*Identifying optimal diagnostic windows*

220bp sliding window analysis of diagnostic nucleotides proportion of monophyletic trees identified 2 regions located 255-475bp and 340-560bp into the folmer DNA barcode region for which the highest density best diagnostic windows for all families on the pest list occurred (fig 2). In contrast, 420bp windows identified the highest density of best diagnostic windows at. A range of published primers overlapping these regions were retrieved from the literature, and supplemented with novel primers (Table 1).

*Evaluation of primer specificity*

All primer pairs showed comparable resolution across the, except for the BF1-AgpestR2 primer (Figure 2). For all those that failed, generally the result of synonyms and indicate the need for curation of synonyms in reference sequence databases.

*Evaluation of primer mismatch*

Most primers performed well, those in certain sections of COI did not, largely due to the difficulty of matching the poly AT’s

*Evaluation of off-target amplification.*

Unsurprisingly the primers with the best performance in primer mismatch evaluation performed the worst for off-target amplification, revealing the trade off required in selecting appropriate primer degeneracy

# Discussion

*A universal diagnostic assay for insect pests?*

As interest continues in using non-targeted HTS assays as universal detection tools for pests and pathogens, it is essential to define in advance the selectivity. For DNA metabarcoding, appropriate selection of PCR primers presents a critical component for determining the sensitivity and selectivity of a diagnostic assay. In order to determine the potential for universal insect pest detection and recommend primers for future users we conducted an in-silico validation of published COI metabarcoding primers on insect pests.

In general, metabarcoding primers of appropriate size for both 2x250bp sequencing (common illumina miseq) as well as the 2x150bp sequencing (common to other illumina as well as MGI platforms) perform effectively for identification of insect pests. 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, Jin, Liang, Zhang, & Li, 2014; Krosch, Schutze, Strutt, Clarke, & Cameron, 2017). Those that could be identified by the conventional Folmer primers, but not mini primers included.

In many 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 symonyms, 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.

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: 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). While long-read sequencing will provide access to 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 important for Fungi, bacteria etc.

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, Soubrier, Weyrich, & Cooper, 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.

Short primers such as fwhF2 + fwhR2n provide flexibility. They are short enough to be used with short-read sequencers, and when used with longer reads (ie miseq)

Our study demonstrates the potential for use of metabarcoding as a diagnostic tool in a broad-spectrum surveillance tool. We recommend fwhF2 + fwhR2n for use with sequencers such as the illumina HiSeq, NovaSeq, NextSeq, and BF3 + BR2 for use with the 2x250bp illumina MiSeq. We provide a list of taxa for which these primers contain or do not contain sufficient resolution.

# Acknowledgments

The authors wish to acknowledge the iMapPESTS project team for fruitful discussions on taxonomy and metabarcoding, as well as the diverse open-source software community for which our work owes a great debt to. 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 reference databases for validation and metabarcoding analyses are provided in the ‘taxreturn’ R package, available on GitHub <https://github.com/alexpiper/taxreturn>. All further code required to reproduce all statistical analyses and generate all figures is contained in the manuscript’s 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*, 1–11. doi:10.1007/s10658-018-1570-0

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

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

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

Brown, S. D. J., Collins, R. A., Boyer, S., Lefort, M.-C., Malumbres-Olarte, J., Vink, C. J., & Cruickshank, R. H. (2012). SPIDER: an R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. *Molecular Ecology Resources*, *12*, 562–565.

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

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

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

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

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

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

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. (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

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

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

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

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

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., Candresse, T., Gil, J., Lacomme, C., Predajna, L., Ravnikar, M., … Wetzel, T. (2017). A framework for the evaluation of biosecurity, commercial, regulatory, and scientific impacts of plant viruses and viroids identified by NGS technologies. *Frontiers in Microbiology*, *8*, 45. doi:10.3389/fmicb.2017.00045

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

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

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

Piñol, J., Senar, M. A., & Symondson, W. O. C. (2019). The choice of universal primers and the characteristics of the species mixture determines when DNA metabarcoding can be quantitative. *Molecular Ecology*, *28*(May), 407– 419. doi:10.1111/mec.14776

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

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

Schlaberg, R., Chiu, C. Y., Miller, S., Procop, G. W., & Weinstock, G. (2017). Validation of metagenomic next-generation sequencing tests for universal pathogen detection. *Archives of Pathology and Laboratory Medicine*, *141*(6), 776–786. doi:10.5858/arpa.2016-0539-RA

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

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

Wilkinson, S. P., Davy, S. K., Bunce, M., & Stat, M. (2018). Taxonomic identification of environmental DNA with informatic sequence classification trees. *PeerJ Preprints*, *6*, e26812v1. doi:10.7287/peerj.preprints.26812v1

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. (2019). rentrez: An R package for the NCBI eUtils API. *The R Journal*, *9*(2), 520. doi:10.32614/rj-2017-058

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

Zinger, L., Bonin, A., Alsos, I. G., Bálint, M., Bik, H., Boyer, F., … Taberlet, P. (2019). DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. *Molecular Ecology*, *28*(8), 1857–1862. doi:10.1111/mec.15060

# Tables

Table 1: Published and novel primers evaluated in this study

1. Upset plot of species overlap between pestlist datasets – Could also include PCA of differences between datasets
2. Summary of sequences / Unique species kept through dataset filtering. Need to write out at each stage? (Supplementary) – Have some icons for the different stages (ie phmm icon, etc)
3. Plot of primer positions/entropy within COI, as well as summary of identification success for different primers – for pest taxa only
4. - Figure summarising failed classifications. Ie exact match different taxonomy, 99% match different taxonomy. Whether they were in a big group or just a couple of sequences (ie could majority rules taxonomy removal solve this) etc
5. Plot of predicted mismatch between different primers and entire insect pes tgroups with dendogram on left to show that mismatch is concentrated on some clades
6. In-silico predicted off target effects of primers – trees
7. Summary of recommended primers?

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

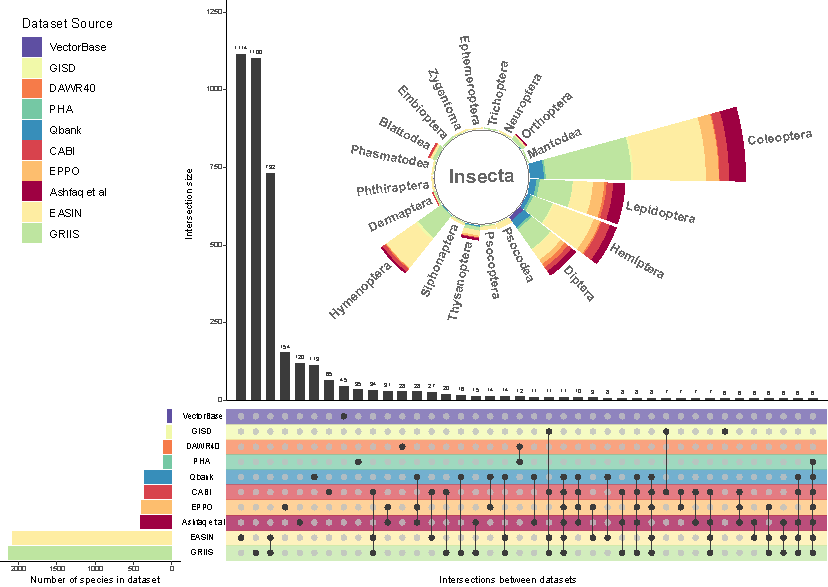
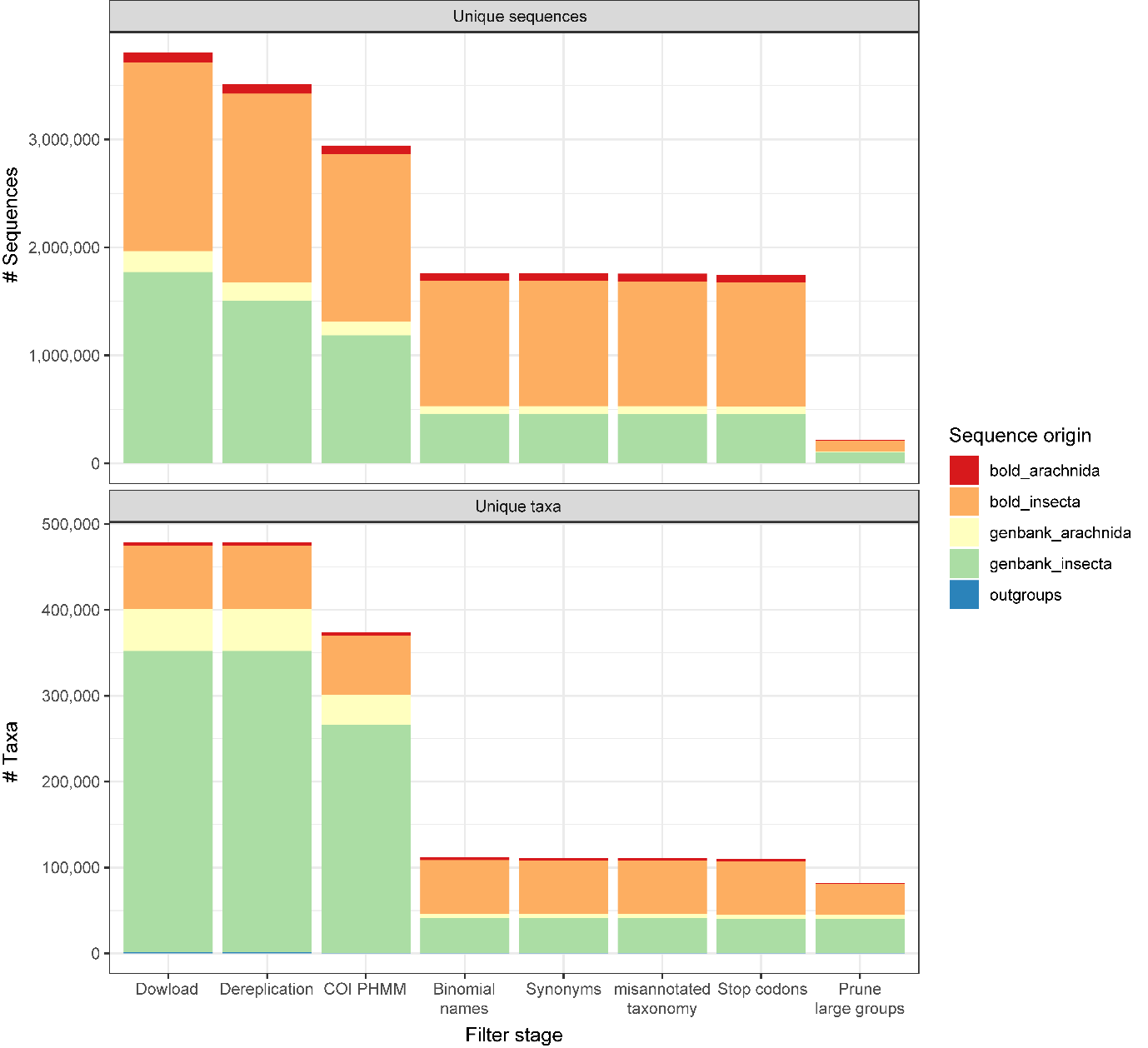


Figure n : Summary of seuqences retained throughout reference sequence cleaning steps, and their databse origins.

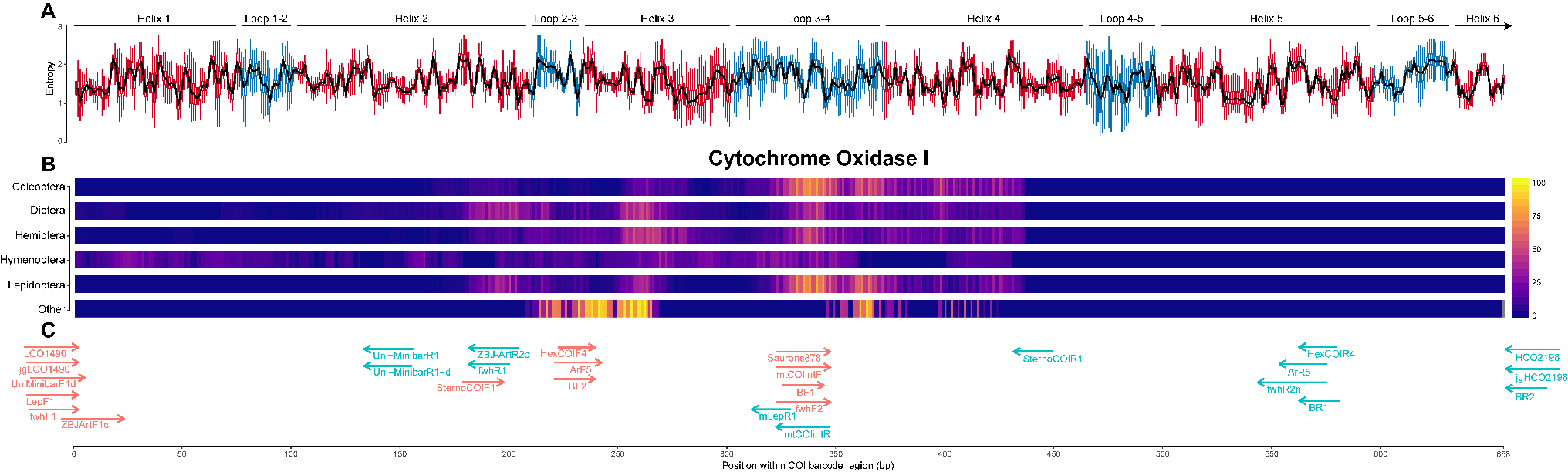
Figure n - Summary of cytochrome oxidase 1 entropy, highest ranked diagnostic windows, and placement of all primers tested

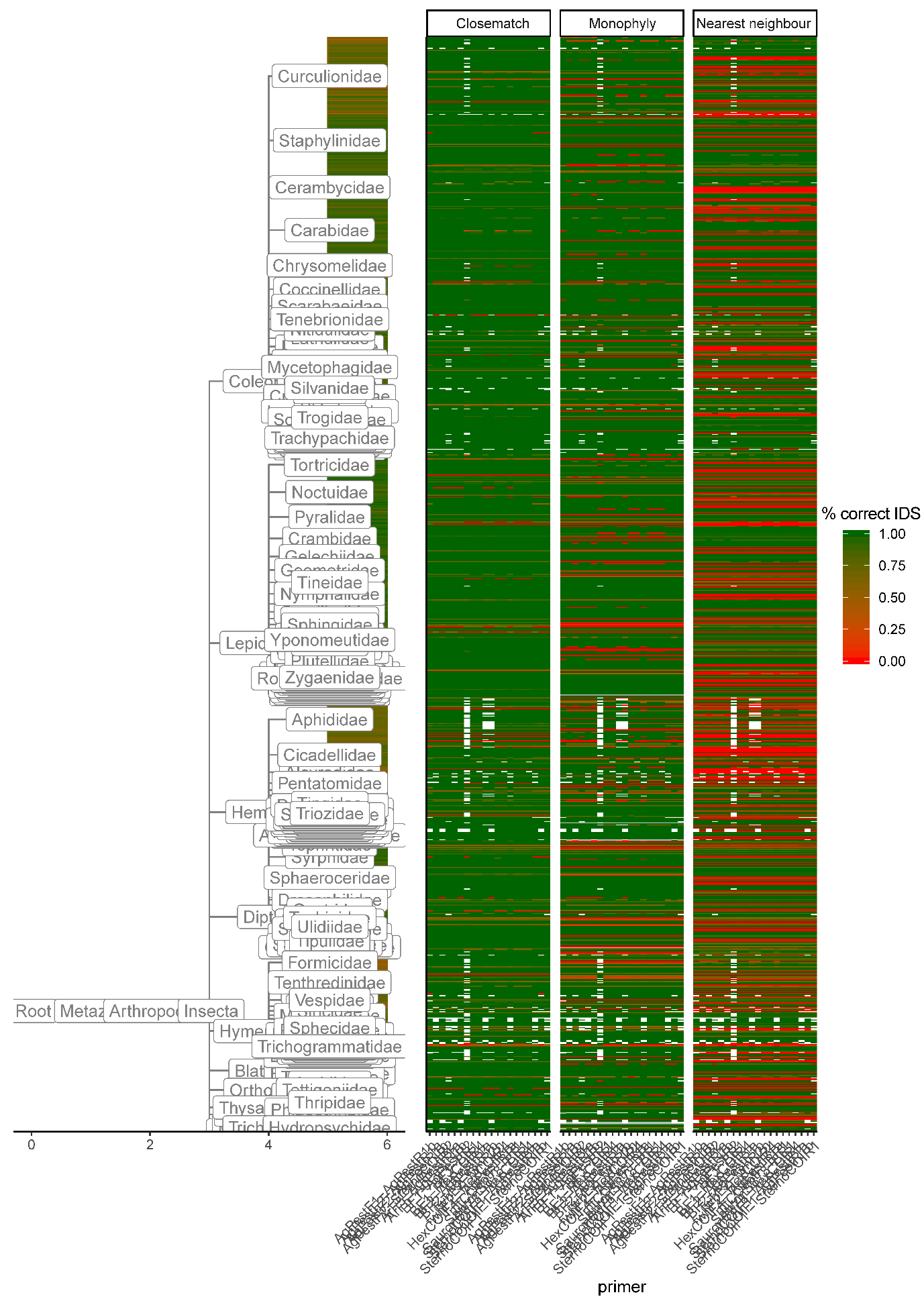
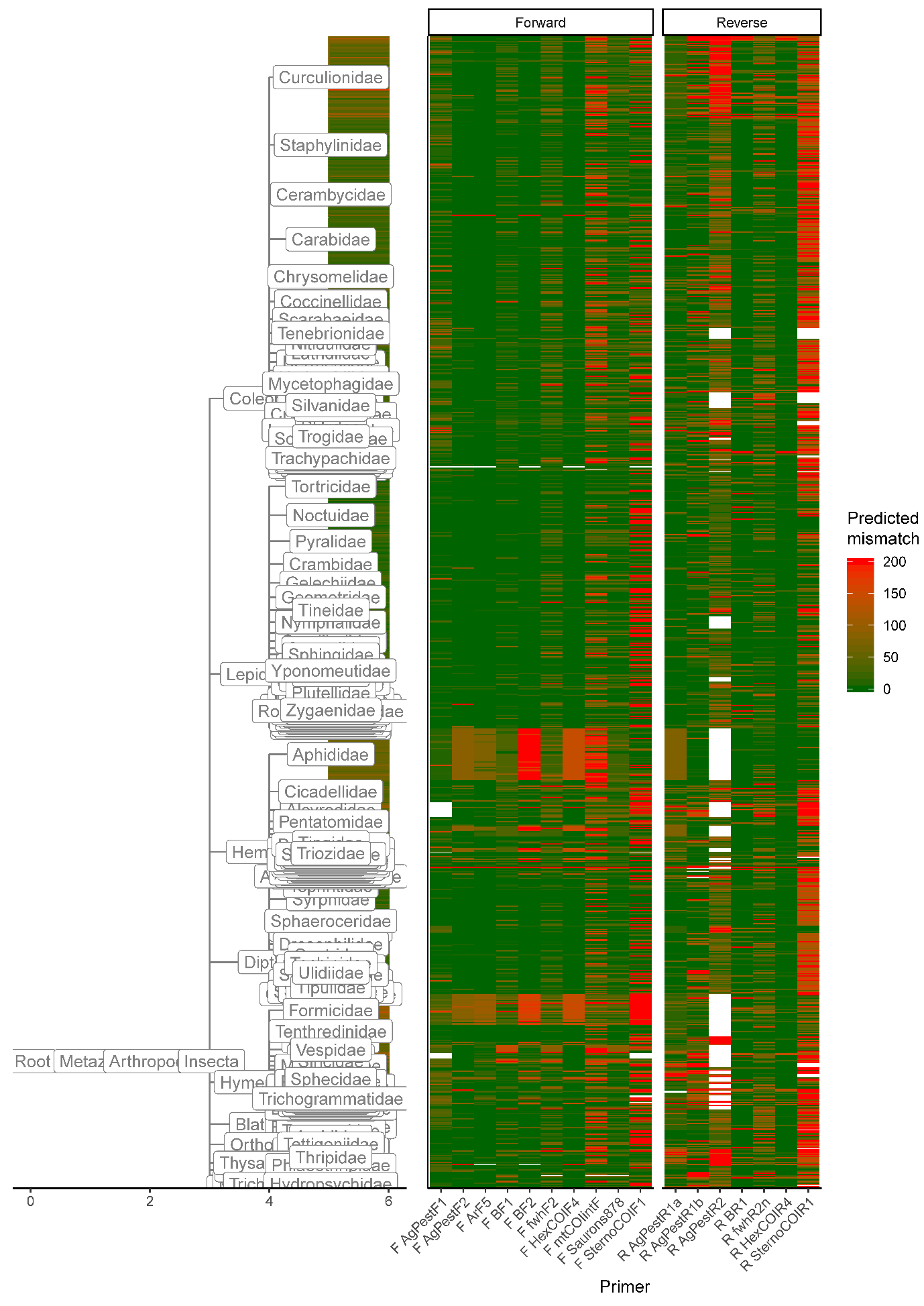
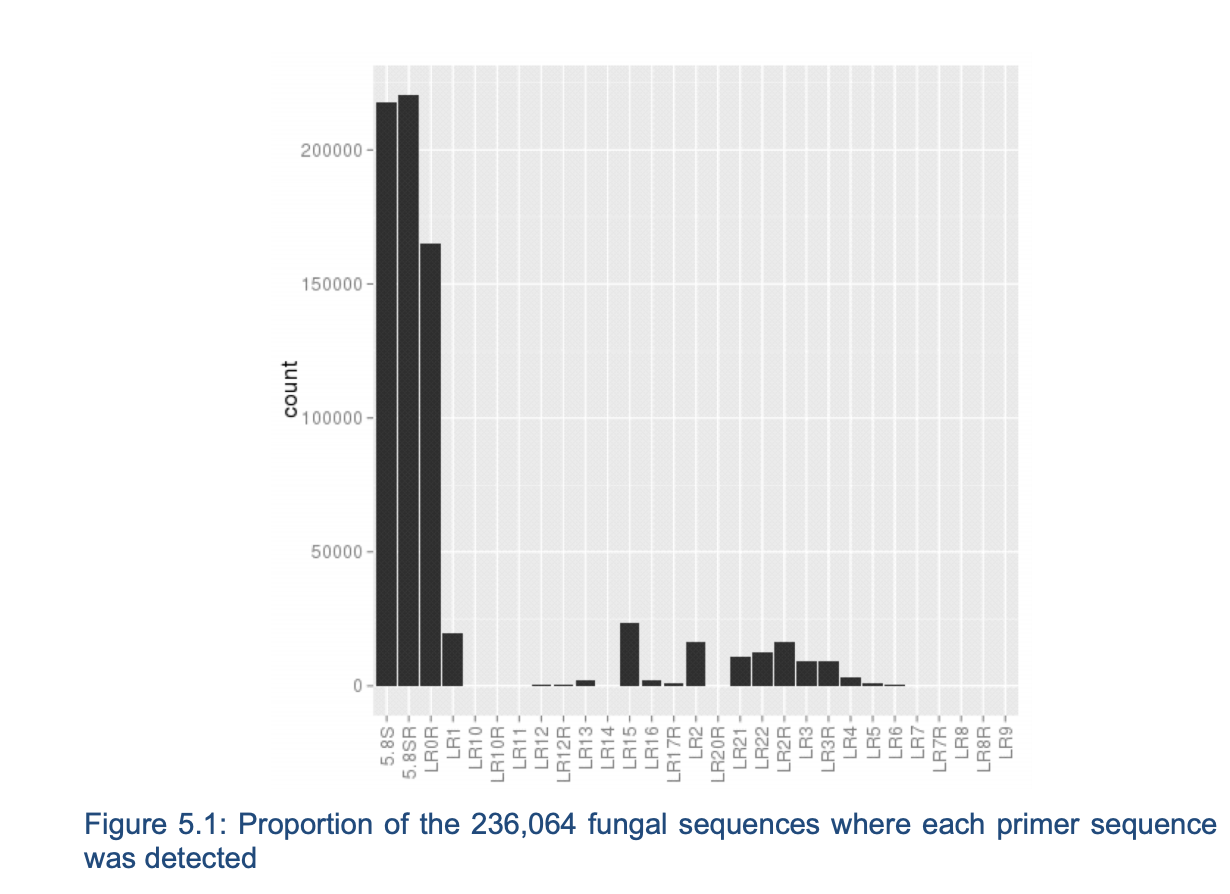
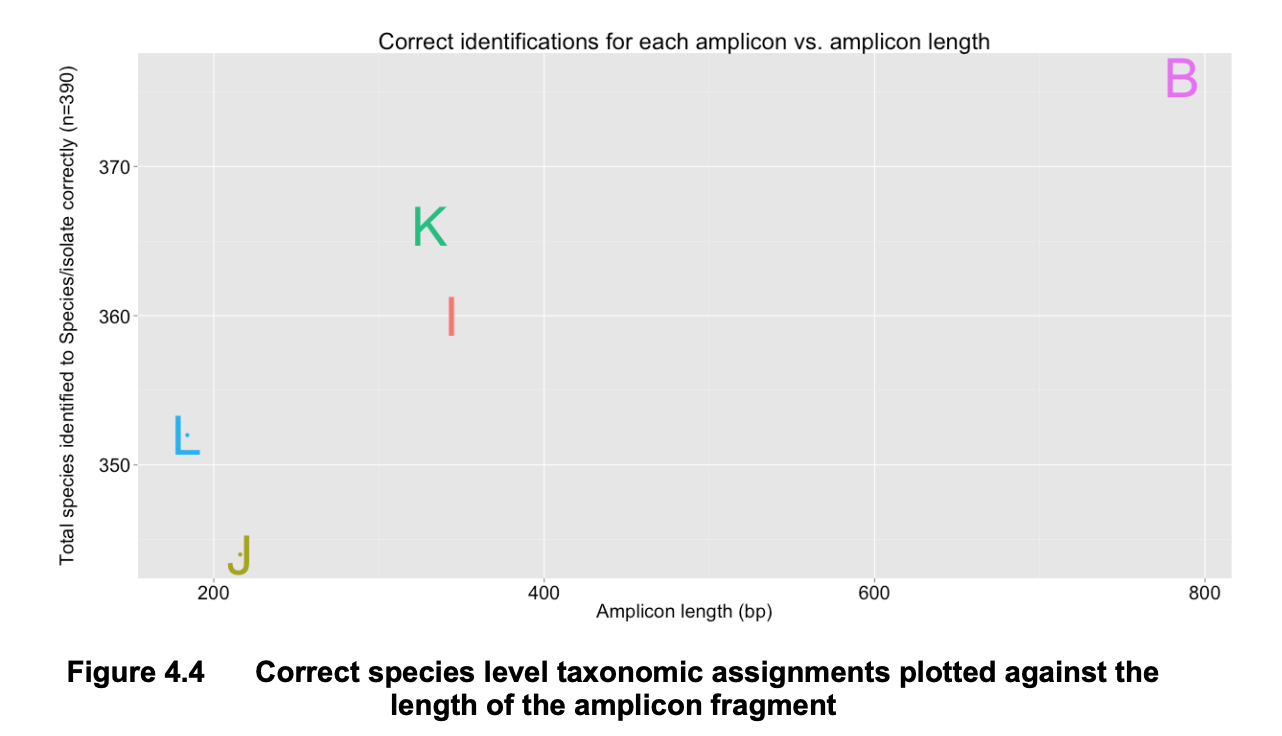
Figure n - Evaluation of novel and existing primers for diagnostic ability across pest insect families, for different methods of taxonomic assignment. 

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

**Supplementary Figures**



Proportion of sequences in which each individual primer was detected



Summary of correct identification by amplicon length

**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)